# Exploiting an Amaranth MAGIC population and diversity set for improvement of leafy vegetable traits.

Maud A.J. de Kinderen

**PhD**

University of York
Department of Biology
September 2025

# Abstract

Amaranths are traditional nutritious underutilised crops with potential to contribute to sustainable, healthy food systems, although unfortunately research and breeding is often neglected. Hence improvement of vegetable amaranth through understanding genomics underlaying agronomic traits and selecting preferred lines for cultivation could greatly benefit African smallholders.

Most of this work focussed on a Multi-parent Advanced Generation Inter-Cross (MAGIC) population created for linking phenotypic with genetic variation and developing new amaranth lines. Using MAGIC founder plus existing reference genomes, an inter-specific *Amaranthus* pan-genome has been made as a novel genetic resource to be used by scientists. Additionally, a 90-accession diversity set has been exploited both to aid species demarcation and to conduct the first metabolite-based genome-wide association study (mGWAS) with amaranths. Lastly, genomic and transcriptomic data of both sets (MAGIC and diversity) were used to identify putative biosynthetic gene clusters (BGC) aiding prediction of nutritional metabolite profiles.

The *Amaranthus* pan-genome showed the MAGIC founders to be phenotypically metabolically and genomically variable, with a graphical pan-genome based on structural variants developed for read alignment. 51 offspring accessions are selected as best due to consistent performance. Phenotypic variation of the founders segregates in the MAGIC offspring indicating a sufficient quality mapping population. This was confirmed by the ability to map genome associations with leaf yield as a relevant agronomic trait at two loci (chromosome 9 and 10). Phylogenetic analysis and mGWAS using the diversity set was successful, classifying 38 of 90 accessions, questioning existence of the *A. hybridus* species and revealing 81 metabolite-genome associations. 43 putative BGC are identified with 11 selected as of interest due to gene co-expression.

Both exploited amaranth sets contain lines useful for direct farming or further breeding practices. Additional trials are suggested incorporating more phenotypes and to identify candidate genes, however, current association results can already benefit amaranth improvement.

# Table of Contents

Abstract ..... ii
Author's declaration ..... 1
Acknowledgements ..... 1
Chapter 1: Introduction and literature review. ..... 2
1.1 Project relevance and gaps in research ..... 2
1.2 Current status of amaranth in scientific literature ..... 4
History and background ..... 4
Species evolution and domestication ..... 5
Amaranth genomics ..... 8
*Amaranthus* genomic references and resources used in this thesis ..... 13
Amaranth phenotyping and metabolite detection ..... 16
Biosynthetic gene clusters (BGC) relevance and identification ..... 22
1.3 Genomic variation and its usage for trait mapping ..... 25
DNA sequencing techniques and genomic variation responsible for traits ..... 25
Pan-genomes in plant breeding ..... 29
Multi-parent advanced generation inter-cross (MAGIC) populations for trait mapping ..... 35
1.4 Project description and research aims ..... 41
Chapter 2: MAGIC founders and creation of an inter-specific *Amaranthus* pan-genome. 43
Statement regarding chapter 2 and submitted manuscript ..... 43
Introduction ..... 43
Results ..... 43
Work in manuscript ..... 43
Additional work ..... 52
Discussion ..... 65
Methods ..... 67
Additional metabolite profiling and analysis ..... 67
Initial genome alignment with BWA and coverage assessment ..... 67
Additional MAGIC founder genome investigation and GO term enrichment ..... 68
Chapter 3: It's a kind of MAGIC, revealing the mysteries of vegetable amaranth and mapping the genomic regions impacting leaf yield. ..... 69
Abstract ..... 69
Introduction ..... 70
Quantitative trait loci (QTL) mapping in various types of populations ..... 70
Multi-parent advanced generation inter-cross (MAGIC) populations ..... 71

An *Amaranthus* MAGIC population in South Africa ..... 72
Analysis of the *Amaranthus* MAGIC offspring RILs ..... 73
Results ..... 76
Relevant agronomic traits are segregating in the MAGIC population ..... 76
Quality control for genotyping of the MAGIC RIL sequences ..... 83
QTLs could be mapped for pleiotropic leafy yield traits ..... 91
Discussion ..... 102
Phenotypic variation among the MAGIC offspring ..... 102
Improved genotyping of the MAGIC RIL sequences ..... 104
Loci for relevant agronomic traits identified ..... 105
Methods ..... 107
Field trial ..... 107
Genomics ..... 110
Chapter 4: Domestication confusion in 'hybridus' species complex: Questioning the existence of *A. hybridus* based on exploiting a 90 accession diversity set ..... 115
Abstract ..... 115
Introduction ..... 116
Amaranth taxonomy ..... 116
Genome-wide association studies (GWAS) and metabolite-based GWAS (mGWAS) 118
Analysis of the *Amaranthus* diversity set ..... 119
Results ..... 120
Obtaining highly variable *Amaranthus* diversity sets from various sources ..... 120
Correcting errors in species definitions ..... 120
Alignment and variant calling against two available *Amaranthus* reference genomes reveals read quality and accession selection for analysis ..... 121
Assigning species groups based on novel phylogenetic analysis ..... 131
Varying degrees of polymorphism and genetic differentiation observed across newly defined subpopulations. ..... 147
*Amaranthus* species show continent specific origins, especially in their native environment the America's ..... 150
Metabolite clustering roughly confirms phylogenetic species assignment ..... 156
Nutritionally relevant metabolites could be identified and occasionally appear species specific ..... 159
mGWAS identify SNPs that are significantly associated with 81 metabolites. ..... 163
Discussion ..... 174
Successful species assignment and questioning of *A. hybridus* existence ..... 174

Nutritionally relevant metabolites detected and further validation of species assignment ..... 179
A hint to genes that may be involved in relevant metabolite production ..... 181
Methods ..... 184
Sequences obtained and read mapping ..... 184
Assessing mapping coverage and selecting high quality reads ..... 184
Variant calling ..... 184
Phylogenetic analysis for species identification ..... 185
Genetic diversity, differentiation and origin of the novel assigned species groups ..... 186
Metabolite profiling ..... 186
mGWAS ..... 188
Heritability of metabolite traits ..... 189
Chapter 5: Identification and analysis of putative biosynthetic gene clusters (BGC). ... 191
Abstract ..... 191
Introduction ..... 192
Biosynthetic gene clusters (BGC) in plants and amaranths. ..... 192
Identification of putative BGC in our dataset ..... 193
Results ..... 194
Putative biosynthetic gene cluster (BGC) identification ..... 194
Evaluation of gene expression data and species assignment ..... 195
Identifying putative BGC of interest based on gene expression data ..... 199
*DODAα1* and *CYP76ADα1* gene expression involved in betalain biosynthesis ..... 203
Discussion ..... 205
Methods ..... 210
Data generation and RNA-sequencing ..... 210
Biosynthetic gene cluster (BGC) identification ..... 210
Quantification and quality control of RNA-seq data ..... 210
Plotting of expression data and synteny analysis ..... 211
Chapter 6: General discussion ..... 213
6.1 An inter-specific *Amaranthus* pan-genome ..... 213
6.2 QTL mapping for leafy traits, their usage in breeding and prospects of the *Amaranthus* MAGIC population ..... 214
6.3 The selection of promising individuals for farmers and onward breeding practices .. 216
6.4 Future outlook and continued analysis ..... 218
Concluding remarks ..... 220
References ..... 221

Appendix ..... 243
Manuscript ..... 243
Introduction ..... 244
Materials and Methods ..... 246
Field trial ..... 246
Metabolite profiling ..... 246
RNA sequencing and analysis ..... 246
Genome sequencing and assembly ..... 247
Pseudochromosome construction and genome annotation ..... 247
Gene-based pangenome ..... 248
Identification of variants and generation of graphical pangenome ..... 248
Results ..... 248
The eight parent accessions of a *Amaranthus* MAGIC population show variation in agronomic traits ..... 248
The eight amaranth accessions vary in metabolite profiles relevant to human nutrition ..... 251
Generation of high-quality annotated genome assemblies for the eight Amaranth accessions ..... 253
The genetic repertoire of an inter-specific Amaranth pangenome ..... 255
Sequence and structural variation in the amaranth pangenome ..... 258
Discussion ..... 263
Acknowledgments ..... 265
Author Contributions ..... 265
Data availability statement ..... 266
Conflict of interests ..... 266
Supplementary information ..... 266
References ..... 267
Figure legends ..... 271
Figures ..... 273

# List of Figures

**Figure 1.1** Trees of the genus *Amaranthus* made conducting SNAPP (SNP and AFLP Package for Phylogenetic analysis) with genotyping by sequencing (GBS) of 94 accessions presenting 35 species. **A)** Consensus tree including genome size estimates of interior nodes obtained with a Maximum Likelihood method mapped onto it, Mbp displayed by branch colour, groups are subgenera, and asterisks (∗) indicating deviation from random genome size evolution at 95% confidence based on a two-tailed test. **B)** Species tree based on the multispecies coalescent calculated with SNAPP, consisting of consensus tree (blue) and cloudogram (green lines), the later representing 3980 individual trees (Stetter and Schmid, 2017). ..... 8
**Figure 1.2** Synonymous nucleotide substitutions per synonymous site (Ks) distributions for the paranome and orthologues of the *A. cruentus* genome and related species (Ma *et al.*, 2021). ..... 14
**Figure 1.3** Chromosome evolution models. **A)** Phylogenetic tree and chromosome numbers of related Amaranthaceae family species; *A. hypochondriacus*, quinoa, and beet. Haploid chromosomes in each species represented by orange, grey, and blue boxes, orthologous chromosomes between the species represented by orange and blue boxes only. Black stars represent lineage-specific whole genome duplications. (Lightfoot *et al.*, 2017). **B)** *A. hypochondriacus* (Lightfoot *et al.*, 2017) and **C)** (a) *A. cruentus* (Ma *et al.*, 2021) synteny blocks between their homoeologous chromosomes. (b) Pseudomolecules (with length in Mb), (c) gene density and (d) transposable element density based on the *A. cruentus* genome assembly landscape and annotation (Ma *et al.*, 2021). ..... 15
**Figure 1.4** Synteny blocks of relationship between *A. hypochondriacus* and *A. cruentus* genomes. **A)** Synteny with colouring based on chromosomes of *A. hypochondriacus*. **B)** Fission of chromosome 2 in *A. cruentus* with separate colouring of the two halves of *A. hypochondriacus* chromosome 2 (Ma *et al.*, 2021). ..... 15
**Figure 1.5** Three types of BGC organisation involved in secondary metabolite biosynthetic pathways. **A)** Compact gene clusters with gene order roughly colinearly with pathway reactions. **B)** Loose gene clusters with closely adjacent core gene cluster components with distributed pathway enzymes which are indicative of a fragmented metabolic pathway. **C)** Super gene clusters which are a combination of different metabolic gene clusters (Zhan *et al.*, 2022). ..... 24
**Figure 1.6** Types of structural variations (SVs). First two boxes presenting unbalanced SVs (deletion, insertion and duplication), third box balanced SVs (inversion and translocation) (Della Coletta *et al.*, 2021) ..... 27
**Figure 1.7** Increasing numbers of pan-genomics related publications from **A)** the first pan-genome in 2005 and **B)** first plant pan-genome in 2014 onwards. Terms searched for on 12-06-2025 are **A)** ‘pan-genome’, ‘pangenome’ and ‘pan genome’ and **B)** ‘pan-genome’, ‘pangenome’, ‘pan genome’, ‘crop’ and ‘plant’ in NCBI Pubmed (https://pubmed.ncbi.nlm.nih.gov/). ..... 30
**Figure 1.8** Models of pan-genomes displaying the enhanced alignment towards ‘graph-based’ versus linear pan-genomes. **A)** Linear pan-genome constructed by aligning genomes of multiple accessions displaying sequence variation. **B)** Graph-based pan-genome constructed from the same aligned sequences, but now with variation between genotypes displayed with red arrows (edges) which are additional paths for alignment against the conserved regions (nodes). **C)** Alignment of the unknown genotypes towards the linear pan-genome leading to mismatched and sequence gaps as variant information is not fully

captured. **D)** Alignment of the unknown individuals towards the 'graph-based' pan-genome can reveal all genetic variation and properly map to matching sequences through edges (red arrows) (Hameed *et al.*, 2022)..... 33
**Figure 1.9** Optimized read alignment by combining single linear genome as well as graph-based pan-genome references. Top: First reads (red) are roughly mapped to the graph-based pan-genome genome (nodes are blue; edges are black arrows). Middle: Then based on where most reads map a path is a predicted through the graph. Bottom: reads are aligned towards the linear path derived from the graph for optimal mapping (Grytten *et al.*, 2020).. 35
**Figure 1.10** Potential approaches for a four-way MAGIC population creation consisting of four founder parent accessions; A, B, C, and D. **A)** the 'funnel' approach, **B)** the 'diallel' approach and **C)** inbreeding and homozygosity achievement by either single-seed descent (SSD) or double haploid (DH) production (Arrones *et al.*, 2020)..... 39
**Figure 1.11** Pedigree representation of two eight parent accession MAGIC populations using either a single **A)** or multiple **B)** funnel crossing design. **A)** Parents of the single funnel design were paired to create four F1s which in their turn were followed by two four-way crosses and one eight-way cross. **B)** Parent accessions of the multiple funnel design with a half diallel approach are intercrossed to create a 28 two-way F1, which in its turn was crossed to 15 complementary F1, creating 210 four-way hybrids. Then 315 eight-way individuals were produced by crossing each four-way hybrid to three complementary four-way hybrids (Ladejobi *et al.*, 2016)..... 40
**Figure 2.1** Raw metabolite value heatmap displaying 85 most abundant presence absence values (PAV) ordered vertically by median abundance over eight parent accessions, and horizontally by species. Absence of metabolite abundance is displayed as white. Low median abundance levels of 0-0.001 are displayed light pink to dark purple. Higher median abundance levels of >0.0001 grey to black. Annotated metabolites and significance between species by multiple testing corrected t test indicated. '*' behind accession ID's indicates lines with admixture. Three arrows indicated are metabolites with their raw values displayed in **Figure 2.2**..... 54
**Figure 2.2** Varying raw metabolite abundance values of three example features highlighted in (**Figure 2.1**). These are all presence absence value (PAV) metabolites and raw values are given per three repeat sample of each founder accession. Here sample 1 is leaf tissue A and samples 2 and 3 repeated leaf tissues B1 and B2..... 55
**Figure 2.3** Boxplots with histograms at the right side displaying coverage per MAGIC parent accession aligned to an *A. cruentus* reference genome (Ma *et al.*, 2021) using BWA v.0.7.17-foss-2018b (Li and Durbin, 2010). Coverage is per exon or non-exon region and ordered by median of the exon coverage. Species are given behind accession ID's with '*' indicating admixture. ..... 57
**Figure 2.4** Density plots displaying density of mean coverage (x-axis) per 100 base pair windows of eight MAGIC parent accessions per 17 main chromosomes of the *A. cruentus* reference genome (Ma *et al.*, 2021), used to align the parent reads to..... 58
**Figure 2.5** Mean coverage (y-axis) of MAGIC parent reads per 100 base pairs plotted per main *A. cruentus* reference chromosome. Coverage is ranging 0-10000. ..... 59
**Figure 2.6** Density plots displaying density of certain coverages (x-axis) of eight MAGIC parent accessions per 100 base pair windows per 17 main chromosomes of the *A. cruentus* reference genome (Ma *et al.*, 2021) used to align the parent reads to. '*' behind accession ID indicates presence of species admixture..... 60

**Figure 2.7** Coverage (y-axis) of MAGIC parent reads per 100 base pairs plotted per main *A. cruentus* reference chromosome. Coverage is ranges 0-10000 and '*' behind accession ID indicates presence of species admixture. ..... 61
**Figure 2.8 A)** Nucleotide diversity (π) per species groups considered as separate populations of MAGIC founders. A t-test was used for significance between species populations and median π per species group are provided. **B)** Midpoint rooted phylogenetic tree of the revealing relationships between the MAGIC founders based on SNP calling towards *A. cruentus* Arusha (Ma *et al.*, 2021). '*' behind accession ID indicates admixture. **C)** Presence (red)/absence (white) of gene families identified using OrthoFinder v.2.3.3 (Emms and Kelly, 2019) with accessions arranged according to species. ..... 62
**Figure 2.9** Original pan- components (top) and those ignoring identification including *A. h.* reference genes used for GO-enrichment analysis (bottom). Soft core genes are those absent in one accession, MAGIC private those unique for one founder accession. Number of (No) genes (orthogroups or single copy) is provided per bar graph. ..... 63
**Figure 2.10** Top ten most significantly (top to bottom) enriched GO-terms in the **A)** core, **B)** dispensable and **C)** MAGIC private (expected gene count of >1) components of the *Amaranthus* pan-genome. Component proportions are displayed as percentage and the total (100 %) number of (No) genes are provided at right side next to the proportional bars. *A. c.* and *A. h.* species specific are pulled out from the dispensable component. ..... 64
**Figure 2.11** Top 10 most significantly (top to bottom) enriched GO-terms in the **A)** *A. c.* or **B)** *A. h.* species specific parts pulled out of the dispensable *Amaranthus* pan-genome and **C)** the private genes with an expected gene count of > 1. Pan- component proportions are displayed as percentage. Total (100 %) number of (No) genes are provided at right side next to the pan-genome component proportional bars. *A. c.* and *A. h.* species specific are pulled out from the dispensable component. ..... 65
**Figure 3.1** *Amaranthus* MAGIC population crossing scheme representing which founders were crossed in which order to create an F4 generation of MAGIC offspring RILs used for mapping traits. Properties of the founders are provided as mentioned in the manuscript in the appendix with updated species information after correction. The map of South Africa is visualized including the four different field trials locations the MAGIC RILs are planted in and with photos of the four MAGIC population growth sites. ..... 75
**Figure 3.2** Proportional bar graphs showing how categorical traits are divided among the eight MAGIC parents (top) and further segregate among the MAGIC offspring (bottom). The assignment of a certain trait for a single accession was based on what was most commonly assigned over all data of the different field trial locations and repeats. In case a different assignment per accession was in an equal proportion with each other the descriptor of this trait was then described as 'other'. ..... 77
**Figure 3.3** Fresh and dry leaf weight and chlorophyll data collected at four field trial locations. **A)** MAGIC offspring collected data plotted as boxplots and violin plots and MAGIC parent or accessions Anna and Arusha data grown in two locations plotted as separate datapoints. Wilcoxon signed-rank test was used for non-normally distributed fresh and dry leaf weight data with significance indicated as **** = $p \leq 0.0001$ and *** = $p \leq 0.001$. Tukey's HSD test was used for post-hoc comparisons for the normal distributed chlorophyll data with letter codes for significance. **B)** Histograms and quantile-quantile (Q-Q) plots displaying distributions of raw, and in case not normal, log transformed values, of the numerical data for the three recorded traits for significance..... 78
**Figure 3.4** Agronomic trait values observed in the MAGIC offspring accessions are all significantly ($p < 0.05$, Pearson's correlation test, detailed statistics are in **Table 3.1**) and

highly positively correlated with each other which suggests pleiotropy. 2013 data points were used for each trait category because these are all three recorded for 322 different offspring accessions in certain test locations or replicate. The maximum measurements per individual accession was 10, minimum 1 with a median of 6. The darker blue and bigger the circle the higher correlations between traits are. ..... 79
**Figure 3.5** 'Best' selected MAGIC offspring accessions having the highest 60% fresh (left) or dry (right) leaf yield weight in all four or at least three field trial locations (top), or in all three field trial locations excluding the limited data of the single repeat Nelspruit trial (bottom). Number of accessions plotted in parallel coordinates plots is provided (N=) and accessions ID's are given at the left side of each plot. Y-axis values represent a variable corresponding to observations but do not have any further inherent meaning. ..... 82
**Figure 3.6** Histograms showing the frequency of mean coverage towards exonic and non-exonic regions of the complete Arusha reference genome that MAGIC offspring reads were mapped to. ..... 85
**Figure 3.7** Boxplots displaying coverage per MAGIC offspring accession mapped to the complete *A. cruentus* reference genome (Ma *et al.*, 2021). Coverage is per exon or non-exon region and ordered by increasing median exon coverage from bottom to top and from left to right. ..... 86
**Figure 3.8** Genetic quality control metrics after genotyping the MAGIC offspring RILs with two different methods, i.e., GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies *et al.*, 2016). **A)** Improving call rates due to SNP imputation method (GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies *et al.*, 2016)) prior (grey) and after (back) filtering (info score <0.4, minor allele frequency <5% and missingness >10%). Here call rates are proportions (1=100%) of the total available SNPs called for each of the 303 offspring accessions, i.e. 303 black or grey dots. **B)** Box- and violin plots of call rates, inbreeding coefficients (F) and homozygosity (%) of all 303 offspring accessions per genotyping method, i.e., GATK haplotype caller or STITCH after filtering. Minimum, mean and maximum values of these plots are also provided in **Table 3.8**. ..... 88
**Figure 3.9** Principal component analysis (PCA) of MAGIC founders and offspring in our analysed *Amaranthus* (top) and an example of a *Brassica juncea* (Yan *et al.*, 2020: Fig. 2d) (bottom) population. PCA of the *Amaranthus* (top) MAGIC population is based on SNPs imputed with STITCH (Davies *et al.*, 2016) after filtering (info score<0.4, MAF<0.05 and MISS>0.01) while the *Brassica j.* PCA is based on 346 intron length polymorphism (ILP) polymorphic markers. The *Brassica j.* PCA provides a clear example in which the MAGIC founders (blue) are evenly distributed among the MAGIC offspring (red) with no clustering. Our *Amaranthus* MAGIC founders contribute relatively equally to the MAGIC offspring, however, some exceptions are founders VI044371 and VI044437-A (PC1-2). *Amaranthus* MAGIC founders cluster per species group (faded circles), with MAGIC offspring being hybrid mosaics of the founders of which some are more related to a certain species. ..... 90
**Figure 3.10** QTL mapped on chromosomes 9 and 10 for leaf yield traits, i.e., dry and fresh weight. At chromosome 9 significant SNPs are found for both traits. At chromosome 10, SNPs were only found to be significant for dry weight, while for fresh weight SNPs were approaching significance. Dashed lines indicate significance thresholds: a nominal per-marker threshold (-logP=6.15 for dry and 6.04 for fresh weight) and a more stringent genome-wide threshold based on phenotype permutations (-logP=6.88 for dry and 6.96 for fresh weight). Proportion of variance explained for QTL on chromosome 9 and 10: 14%, 6% (dry) or 19%, 10% (fresh). ..... 91

**Figure 3.11 A)** Fresh (left) and dry weight (right) phenotypes of the MAGIC offspring, plotted per genotype (reference (Ma *et al.*, 2021) and alternative) of the three most significant SNP locations (**Table 3.9**) at chromosome 10. Significance (p<0.05) was tested with a Wilcoxon rank-sum test. Number of total 303 RILs is provided behind the genotype, e.g., AA:240..... 93
**Figure 3.11 B)** Fresh (left) and dry weight (right) phenotypes of the MAGIC offspring, plotted per genotype (reference (Ma *et al.*, 2021) and alternative) of the three most significant SNP locations (**Table 3.9**) at chromosome 9. Significance (p<0.05) was tested with a Wilcoxon rank-sum test. Number of total 303 RILs is provided behind the genotype, e.g., CC:61. ..... 94
**Figure 3.12** Fresh (left) and dry (right) weight leaf yield performance of MAGIC parental accessions and Arusha during field trials in south Africa. Here the same data of **Figure 3.3 A** is visualized but only of the MAGIC founders and Arusha. The measurements of three repeats at two locations are shown as six individual observations per accession, with boxplots to illustrate the data distribution. V1062433 performs lower than the other accessions and V1044371 performs higher for fresh leaf weight only. ..... 95
**Figure 3.13 A)** Gene expression profiles of genes located within the two QTL at chromosome 9 identified for leaf yield. The *A. cruentus* reference (Ma *et al.*, 2021) gene annotations are given, with the expression of orthologous genes identified with Orthofinder (Emms and Kelly, 2019) of the MAGIC founders towards Arusha. Accessions are ordered from left to right with respectively lowest to highest median dry weight as shown in boxplots of **Figure 3.12**. Most genes located at the QTL are (non-)expressed simultaneously in all MAGIC founder accessions and leaf tissue of Arusha. Gene Amacr09g05810 is highlighted due to its absence in accession V1062433 for which lower leaf yields than the other accessions were observed (**Figure 3.12**). ..... 97
**Figure 3.13 B)** Gene expression profiles of genes located within the two QTL at chromosome 10 identified for leaf yield. The *A. cruentus* reference (Ma *et al.*, 2021) gene annotations are given, with the expression of orthologous genes identified with Orthofinder (Emms and Kelly, 2019) of the MAGIC founders towards Arusha. Accessions are ordered from left to right with respectively lowest to highest median dry weight as shown in boxplots of **Figure 3.12**. Most genes located at the QTL are (non-)expressed simultaneously in all MAGIC founder accessions and leaf tissue of Arusha. ..... 98
**Figure 3.14** Variation of the MAGIC founders in an *Amaranthus* graphical pan-genome (manuscript in appendix) towards the 1000 reference genome base pairs surrounding the most significantly associated SNPs at two identified leafy trait QTL (**Table 3.9**). **A)** Chromosome 9 and **B)** chromosome 10 QTL with the visualized *A. cruentus* reference (Ma *et al.*, 2021) base pair window indicated..... 101
**Figure 4.1** Frequency of base pair lengths of exons and non-exons of the main chromosomes of the two reference genomes mapped to. ..... 125
**Figure 4.2** Histograms showing the frequency of mean coverage values for exon and non-exon regions of all accessions of a diversity set combined. Diversity sets consisting of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Reads were mapped to *A. hypochondriacus* (Lightfoot *et al.*, 2017) or *A. cruentus* (Ma *et al.*, 2021) reference genomes. ..... 128
**Figure 4.3 A)** Boxplots with coverage per sequenced amaranth accession in two diversity sets mapped to an *A. hypochondriacus* (Lightfoot *et al.*, 2017) reference genome. Coverage is per exon or non-exon region and ordered by median of the exon coverage. Diversity sets consist of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Species are given behind

accession ID's with in case re-classified by (Stetter *et al.*, 2020), original gene bank annotation species followed by reclassified genomic group (gene bank – reclassified). .... 129
**Figure 4.3 B)** Boxplots with coverage per sequenced amaranth accession in two diversity sets mapped to an *A. cruentus* (Ma *et al.*, 2021) reference genome. Coverage is per exon or non-exon region and ordered by median of the exon coverage. Diversity sets consist of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Species are given behind accession ID's with in case re-classified by (Stetter *et al.*, 2020), original gene bank annotation species followed by reclassified genomic group (gene bank – reclassified)..... 130
**Figure 4.4** Cross validation errors from genetic structure analysis of 206 *Amaranthus* accessions mapped to the *A. cruentus* reference genome, shown across different numbers of hypothetical subpopulation clusters (k)..... 133
**Figure 4.5** *Amaranthus* accessions clustering together in 9 subpopulation clusters (K=9) which are named after the most common Gene bank species annotation per cluster. Potential *A. hybridus* species clusters are split into two (k=5 and k=6) and marked with *. The Gene bank IDs with their original species annotation (**Supp table diversity set**, without brackets) is given at the x-axis, with numbers of accessions per species mentioned behind the species name. Accessions originally assigned as *A. hybridus* are marked with black arrows displaying their dispersal among the different population clusters..... 134
**Figure 4.6** Population clusters k=8 and k=10 of Admixture analysis after SNP calling in 206 *Amaranthus* accessions mapped to an *A. cruentus* reference genome. The Gene bank IDs with their original species annotation (**Supp table diversity set**, without brackets) is given at the x-axes, with numbers of accessions per species mentioned behind the species name. ..... 135
**Figure 4.7** Principal components 1-2, 2-3 and 3-4 plotted for 206 *Amaranthus* accessions based on mapping to an *A. cruentus* reference genome. Individuals are coloured and shaped by the Gene bank original species annotation as provided in **Supp table diversity set** (without brackets) with number of accessions per species given. Observed distinctive species clustering are visualized by outlined circles and numbers of accessions mentioned behind the species name. Less distinctive species clustering is indicated by faded circles with no numbers of accessions mentioned, since those could not be clearly distinguished. The number prior the species circles in the legend and plots is the cluster number of the k=9 species subpopulations in **Figure 4.5**. ..... 137
**Figure 4.8** Principal components 1-2 plotted for *Amaranthus* accessions mapped to an *A. cruentus* reference genome which distinctively cluster into two groups; **A)** *A. hybridus* and **B)** *A. dubius* here visualized in close up. The *A. hybridus* group is based on only two accessions of the available 17 *A. hybridus* annotations in the complete diversity set and therefore less reliable than the *A. dubius* group which contains all three available accessions. Colours and shapes are based on the Gene bank original species annotation as provided in **Supp table diversity set** *(*without brackets) with number of accessions per species in the total set given. ..... 138
**Figure 4.9** Principal components 1-2 of **Figure 4.7** plotted with accession annotation according to k=9 Admixture results (**Figure 4.5**, including admixed accessions annotated with double species or as *A. sp.*). Accessions clustering inconsistently with expectations based on k=9 Admixture annotation are highlighted. The enlarged part (right) mostly contains accessions of the 'hybridus' species complex. ..... 139
**Figure 4.10 A)** Maximum likelihood trees of *Amaranthus* accessions annotated with their Gene bank species (17 *A. hybridus* accessions annotated all around the tree indicated with

green arrows). Clades are numbered and colored with k=9 subpopulations from **Figure 4.5** (*A. powellii* = 6*) and which clades group to the 'hybridus' species complex is indicated in grey. Bootstrap results based on 1000 replicates indicate a well predicted tree apart from some uncertainties within the *A. hypochondriacus* clade..... 140
**Figure 4.10 B)** Maximum likelihood trees of *Amaranthus* accessions annotated with species according to Admixture analysis with k=9 subpopulations derived from **Figure 4.5** (including admixed accessions annotated with double species or as *A. sp.* hybrids with red asterisks). Clades are numbered and colored with k=9 subpopulations from **Figure 4.5** (*A. powellii* = 6*) and which clades group to the 'hybridus' species complex is indicated in grey. It is given what is believed to be the three grain amaranths and their two wild related species within the complex. Bootstrap results based on 1000 replicates indicate a well predicted tree apart from some uncertainties within the *A. hypochondriacus* clade..... 141
**Figure 4.11** Nucleotide diversity (π) per newly assigned species groups considered as subpopulations based on phylogenetic analysis (**Table 4.5**) calculated as a distribution over 10kb SNP windows. Tukey's honest significance test letter codes for pairwise significance between species populations and median π per species group are provided. The high π value for the *A. dubius* subpopulation could be an indication of a large long-term population size, while a higher π for the *A. hybridus* subpopulation can be explained by the fact that this group likely should not be considered a species. ..... 148
**Figure 4.12** Pairwise fixation index ($F_{ST}$) values between novel defined species groups (**Table 4.5**) showing less genetic differentiation among the 'hybridus' species complex subpopulations (grey triangle) while more differentiation is observed between other more distant subpopulations (black triangle) confirming our phylogenetic analysis. ..... 149
**Figure 4.13** World map displaying the location 188 diversity set accessions of which the origin is known (**Supp table diversity set**). Species and related groups are based on confirmation by phylogenetic analysis (**Table 4.5**). Count of species and groups behind the name in the legend is how often it occurs in the set of 188..... 152
**Figure 4.14** World map displaying the location of diversity set accessions of which the origin is known **A)** Species annotation according to the original gene bank information (**Supp table diversity set**, without brackets). **B)** Species annotation and hybrids (double species) according to k=9 Admixture results (**Figure 4.5**, double species annotation noted for admixed accessions sharing <85% of genome to a single subpopulation or *A. sp.* for admixed accessions with >15% of their genome assigned to more than two subpopulations). Count of species and groups behind the name in the legend is how often it occurs in the set. This is 188 accessions displayed as starting point species in **A)** and 183 accessions used for the phylogenetic analysis of which the origin is known. ..... 153
**Figure 4.15** Maximum likelihood tree (of **Figure 4.10**) coloured by origin (black is unknown). Accessions annotated with new species (red *, **Table 4.5**) and country code with each respective occurrence. Clades with ≥ 4 accessions of same continent origin are highlighted (margin: ≤ 2 other origins). ..... 154
**Figure 4.16** Pairwise fixation index ($F_{ST}$) values between *Amaranthus* subpopulations of various continental origins (**Figure 4.15**) showing a trend of less genetic differentiation of North American accessions to European, African and Asian subpopulations in contrast with the seemingly more distantly related South American accessions. All $F_{ST}$ comparisons with the European subpopulation are negative, likely due to only two accessions belonging to this subpopulation. This causes unstable estimates of heterozygosity with $F_{ST}=(Ht-Hs)/Ht$. Here Ht and Hs are the expected heterozygosity in the total and subpopulation respectively. ... 155

**Figure 4.17** Principal component analysis (PCA) displaying clustering of two individual metabolite repeated runs per 90 diversity set accessions based on metabolite leaf data collected in a field trial. This reveals sufficient quality metabolite data with clustering of the two repeats per accession. Overall clustering of samples might indicate higher relatedness of accessions based on similar metabolite profiles. ..... 156
**Figure 4.18** Principal component analysis (PCA) displaying clustering of 90 diversity set accessions based on metabolite data collected from leaves using the average values of two extract run replicates per accession. Observed distinctive species clustering are visualized by outlined circles and less distinctive clustering is indicated by faded circles. Subpopulation species are novel assignments as by **Table 4.5**, with the numbers of accessions mentioned behind the species name. ..... 157
**Figure 4.19** Tangle gram linking a dendrogram based on repeat average metabolic data (top) with a phylogenetic tree based on SNP calling (bottom) of which 90 diversity set accessions were extracted from the tree in **Figure 4.10**. Species clustering blocks were added manually for which species annotation was based on genomic analysis performed (**Table 4.5**). * Indicates individual accessions or species blocks which do not cluster fully in the dendrogram based on metabolomic data. ..... 158
**Figure 4.20 A)** Total sum of squared errors (SSE) between metabolite means within a cell of the matrix and data values within cells measuring the within cell homogeneity, decreasing per bi-clustering analysis iteration. **B)** Bi-clustering analysis results plot for the optimized total within-cluster sum of squares, with the 90 diversity set accession assigned to nine clusters on the y-axis and 486 annotated metabolites divided over 22 clusters on the x-axis based on normalized metabolite data. Value is the normalized metabolite data and x-axis cluster numbers 10,16, 18 and 20 are not written due to space. Species of accessions per y-axis cluster are in **Table 4.8** and numbers of metabolites assigned per x-axis cluster in **Table 4.9**..... 160
**Figure 4.21** Histograms showing the distribution of base pair (bp) distance between significant SNPs for the same metabolite located on the same chromosome or scaffold. On the left side all bp distances are plotted while on the right it is only a close-up of the total plot (x-axis 0 to 5000 bp) displaying the picked cut-off of 1000 bp distance between SNPs for identifying SNPs located close enough to each other to potentially be involved in the same putative biosynthetic gene cluster and hence significant for different metabolite traits..... 164
**Figure 4.22** Histograms showing the distributions of heritability (top) and their standard deviation (SD) (middle), as well as a scatter plot showing their correlation (bottom), visualised for all 1605 metabolite traits (left) or 81 for which a SNP was found to be significantly associated with that metabolite (right) by performing mGWAS using 90 diversity set accessions. ..... 167
**Figure 4.23** Scatterplots (left) and binned histograms (right) showing the increased number of significant SNPs per 81 metabolite traits associated with either low or high estimated heritabilities (top) or with a high estimated genomic inflation factor (λ) (bottom). ..... 168
**Figure 4.24 A)** Manhattan and Q-Q plots of four metabolite traits which could be annotated (bold Masstag ID's and more information in **Table 4.12**) with considerate low p-values, high heritabilities and non-extreme SD's while having good quality data according to the Q-Q plots. λ estimates with p-values used to plot the Q-Q plots is given..... 172
**Figure 4.24 B)** Manhattan and Q-Q plots of four metabolite traits which could be annotated (bold Masstag ID's and more information in **Table 4.12**) with considerate low p-values, high heritabilities and non-extreme SD's while having good quality data according to the Q-Q plots. λ estimates with p-values used to plot the Q-Q plots is given..... 173

**Figure 5.1** Co-localisation of genes (red) of cluster 13 which is species specific to *A. hypochondriacus* accessions and assigned to be a nutritionally relevance saccharide type BGC. ..... 194
**Figure 5.2** PCA based on expression data showing clustering of RNA-seq samples for **A)** MAGIC parents and Arusha (*A. cruentus* reference (Ma *et al.*, 2021)), **B)** 21 diversity set ARC accessions, **C)** PC1-PC2 and **D)** PC2-PC3 of both sets combined. Clustering mostly occurred per species (A, B and D) and per sample tissue type (A, C and D) with only PC1-PC2 of all data combined separating for the transcriptome used for RNA-seq quantification (C). ..... 196
**Figure 5.3** Gene expression heatmaps of putative BGC in Arusha (an *A. cruentus* reference genome) and four *A. c.* MAGIC parents. In Arusha RNA sequencing data was taken from leaf, root and seedling tissue, while for the *A. c.* MAGIC parents it was taken from leaf only (A and B) with a repeated sample for B (B1 and B2). ..... 197
**Figure 5.4** Gene expression heatmaps of putative BGC in four *A. h.* MAGIC parents and BGC identified in Arusha with expression of those genes in 21 diversity set accessions. All RNA data was taken from leaf tissue with samples A and B for the *A. h.* MAGIC parents with a repeated sample for B (B1 and B2). Of the diversity set accessions only a single leaf tissue sample was taken and those are labelled with the assumed *A. h.* or *A. c.* species but ordered by their new species classification. Expression patterns are less consistent towards the right side of the heatmap with four non *A. h.* or *A. c.* diversity set accessions. ..... 198
**Figure 5.5 A)** Selected BGC of interest based on their consistency in gene expression patterns across multiple accessions. These 11 BGC were selected from a total of 43 identified putative BGC (**Figure Appendix**), and for the MAGIC parent accessions the average of the three leaf tissue samples are displayed in the current heatmaps. ..... 200
**Figure 5.5 B)** Selected BGC of interest based on their consistency in gene expression patterns across multiple accessions. These 11 BGC were selected from a total of 43 identified putative BGC (**Figure Appendix**), and for the MAGIC parent accessions the average of the three leaf tissue samples are displayed in the current heatmaps.. ..... 201
**Figure 5.6 A)** Co-localisation of genes (red) identified as putative BGC with an RNA sequencing expression profile of interest, i.e., all genes are either expressed or non-expressed simultaneously (). More information regarding the function of these putative BGC is available in **Table 5.2** and the synteny analysis of an *A. hypochondriacus* specific cluster 13 is in **Figure 0.1**. ..... 202
**Figure 5.6 B)** Co-localisation of genes (red) identified as putative BGC with an RNA sequencing expression profile of interest, i.e., all genes are either expressed or non-expressed simultaneously (). More information regarding the function of these putative BGC is available in **Table 5.2** and the synteny analysis of an *A. hypochondriacus* specific cluster 13 is in **Figure 0.2**. ..... 203
**Figure 5.7** Co-localisation of betalain biosynthesis involved genes CYP76ADα1 (blue) and DODAα1 (red) on chromosome 16 of the eight MAGIC parents and the *A. cruentus* reference genome. ..... 204
**Figure 5.8** Gene expression heatmaps of *DODAα1* and *CYP76ADα1* enzyme encoding genes known to be involved in betalain biosynthesis. The genes are located in eight MAGIC parent accessions for which the average of three leaf tissue samples are displayed, three different tissue samples were taken for Arusha and only a single leaf tissue sample for 21 diversity set accessions. Observed leaf pigmentation recorded during different field trials (MAGIC founders grown as described in **Chapter 2:** in Arusha, Tanzania and **Chapter 3:**

four locations in South Africa, and diversity set accessions grown in three replicates at Roodeplaat in South Africa) is indicated. For ARC-IV-53 only one rep was recorded. ..... 205
**Figure Appendix** RNA expression heatmaps of genes in 43 putative biosynthetic gene clusters (BGC) (**Supp table BGC**) identified in one to 30 different accessions. Samples were taken from leaf tissue only with the exception of Arusha, i.e., *A. cruentus* reference (Ma *et al.*, 2021), for which it was also taken at root and seedling tissue. The three leaf samples of the MAGIC parents are physically different samples (A and B) and a repeated sample for B (B1 and B2). The 21 diversity set accessions are labelled with the assumed *A. h.* or *A. c.* species but ordered by their new species classification as indicated in (**Figure 5.4**). Expression patterns are therefore less consistent towards the right side of the heatmaps with four non *A. h.* or *A. c.* diversity set accessions and two non-leaf tissue samples of Arusha. Clusters of interest with similar expression patterns are highlighted in **Figure 5.5**. 276

# List of Tables

**Table 1.1** Characteristics of *A. hypochondriacus*, *A. cruentus* and *A. tricolor* reference genomes used for analysis in this thesis (Lightfoot *et al.*, 2017; Ma *et al.*, 2021; Wang *et al.*, 2022). *A. cruentus* V1 refers to the previous publication of (Ma *et al.*, 2021) while V2 is an updated version of V1 with improved genome annotation. ..... 13
**Table 2.1** Combined ionization mode non-redundant recoded metabolite peak dataset, split up in bins according to and ordered by median abundance of raw data over the eight parent accessions. Number of peaks increasing per bin with decreasing median abundance. The range of median abundance the raw data of the peaks was recorded at, the difference of the range of that median abundance (Δ abund.), the number of peaks in the bin, those of which significantly differentiating between parents (sig. par.), those significantly differentiating between species (sig. spec.), raw data presence absence values (PAV) including those significantly differentiating between MAGIC parents (sig. PAV), annotated metabolites significantly differentiating among parents (ann.) and percentage of those divided by peaks per bin are given. Annotations performed with SIRIUS 4 (Dührkop *et al.*, 2019) only. ..... 53
**Table 2.2** Median coverage of eight MAGIC parent accessions aligned to exonic and non-exonic regions of the complete genome as well as main reference chromosomes only (Excl. scaffold) of the *A. cruentus* reference (Ma *et al.*, 2021). The mapping percentage is also provided. Accessions are grouped by species and '*' behind the accession ID indicates admixture. ..... 56
**Table 2.3** File sizes in gigabytes of zipped FASTA files used for, and Binary Alignment Map (BAM) files created after, alignment to an *A. cruentus* reference genome (Ma *et al.*, 2021) with BWA v.0.7.17-foss-2018b (Li and Durbin, 2010). ..... 56
**Table 3.1** Highly significant p-values of agronomic traits observed to be correlated in the MAGIC offspring accessions by Pearson's correlation test. All traits were shown to be positively correlated (**Figure 3.4**). ..... 79
**Table 3.2** Best fitting ANOVA fixed and random model effects of three relevant agronomic traits; fresh, dry leaf weight and chlorophyll. Data was log transformed (**Figure 3.3 B**) in case of no normal distribution. Fixed effect are genotypes (MAGIC offspring accessions), environment (field trial locations) and their interaction (GxE). Repeat (Rep) of the experiment was taken as a random model effect. P-values are given for the fixed effects and displayed in bold when significant ($P < 0.05$). ..... 80
**Table 3.3** ANOVA effect of accession on data which was separated per different field trial location. Significance of accession was tested for three different relevant agronomic traits (dry (DW) and fresh (FW) leaf yield and chlorophyll) and displayed bold in case of significance. Data was log transformed (log()) (**Figure 3.3 B**) in case the data was non-normally distributed. Number of (No) accessions calculations were based on are provided. 80
**Table 3.4** 51 accessions selected as 'best' due to consistent performance over several field trial locations for the most relevant agronomic leaf traits (fresh and dry weight). These are MAGIC offspring accession ID's starting with ' AVAM 22 –' followed by the number in the table. ..... 81
**Table 3.5** Lowest (Min.), median of mean per accession, and highest (Max.) coverage of 303 MAGIC offspring reads mapped to exonic and non-exonic regions of an *A. cruentus* reference genome (Ma *et al.*, 2021). Mapping performed to complete genome and main reference chromosomes only (Excl. scaffold). ..... 83
**Table 3.6** Number of MAGIC offspring lines mapping in bands of <6x, 6-7x or >7x coverage to an *A. cruentus* reference genome (Ma *et al.*, 2021) including percentage of accessions

compared to the entire group of 303 MAGIC offspring. Mapping was performed to exonic and non-exonic regions and to the complete genome and main chromosomes only excluding scaffolds..... 84
**Table 3.7** Outputs for running STITCH on a single chromosome (7 of *A. cruentus* (Ma *et al.*, 2021)) without a reference panel but with a different value of nGen, i.e., 4 or 7, controlling recombination rates. These are call rates of the total available SNPs, percentage of imputed SNPs that are homozygous, and inbreeding coefficients per individual offspring RIL with minimum, mean and maximum values provided..... 87
**Table 3.8** Differences in call rates, homozygosity and inbreeding coefficients of the MAGIC offspring RILs using two different genotyping methods, i.e., GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies *et al.*, 2016). Values for STITCH were based on the VCF after filtering (info score <0.4, minor allele frequency <5% and missingness >10%). All values are also plotted in **Figure 3.8 B**. ..... 89
**Table 3.9** Three most significant SNP locations with the highest p-values are given for each of the identified QTL. The leafy trait they are significant for (fresh (FW) and dry weight (DW)), p-values, Arusha reference (Ma *et al.*, 2021) allele and alternative allele in the offspring haplotypes are indicated. Whether it is the preferred or non-preferred allele resulting in higher or lower leaf yields according to **Figure 3.11** is indicated with green and orange colouring respectively..... 92
**Table 3.10** Three most significant, i.e. highest p-values, SNP locations and alleles are given for the two identified QTL for Arusha (Ma *et al.*, 2021) (reference and alternative allele) and each of the MAGIC founder haplotypes. Preferred or non-preferred allele resulting in higher or lower leaf yields in the offspring (**Figure 3.11**) is indicated with green and orange colouring respectively. If a SNP is located in an annotated gene of the Arusha reference (In gene), and the number of genes between the SNP location on the row below (No. genes) is provided. ..... 96
**Table 3.11** Structural variants (SVs) (<100 kb) found within the two QTL identified for leaf yield traits on chromosome 9 and 10. Presence absence variations (PAV) are insertions (INS) and deletions (DEL), for which numbers slightly increase with a more distant relatedness to the *A. cruentus* reference (Ma *et al.*, 2021) the SVs were identified towards (accessions in grey shading are *A. hypochondriacus* species). Other SVs are inversions (INV), copy number variations (CNV) and translocations with for the later the length being unknown. The number of unique SVs for each accession, i.e. not found in other MAGIC founders, is indicated which was notably high for V1063769. * indicates an accession showing admixture (Lin *et al.*, 2022). The total number of SVs of all accessions are provided per type of SV and per individual QTL..... 100
**Table 3.12** Dates of data collection of MAGIC population field trials conducted at four South African locations..... 108
**Table 3.13** Coordinates and long-term climatic conditions in the provinces of the four field trial locations the MAGIC RILs were planted. For each parameters the overall annual (A) as well as summer (S) means are provided since our field trial was during the summer months. Values are obtained from (Benhin, 2006). ..... 108
**Table 3.14** Number of MAGIC offspring accessions for which data is available; the total of data points (total), which of those are unique to a single accession ID (unique accession) and therefore a rough estimation of repeats the data is based on (~# reps) per three types of phenotypic traits (dry or fresh leaf weight with the same amount of data available, or chlorophyll) and per four field trial locations (Roodeplaat, Mafikeng, Bloemfontein, and Nelspruit). The total of all data points of all four locations per trait, those unique to one

accession ID, and those unique accessions overlapping between all four locations or only three locations excluding single repeat Nelspruit data is provided. ..... 109
**Table 4.1** Diversity set information regarding the 90 accessions grown by the ARC and 121 imported sequences of (Stetter *et al.* 2020). The assumed species in the diversity sets at the start of the experiment prior performing any new phylogenetic analysis, and hence based on phenotypes or gene bank information only for the ARC set and the published phylogenetic analysis for the 121 set (Stetter *et al.* 2020). Accessions of whose species were not defined are *A. sp.* and hybrids are *A. hybr.* ..... 120
**Table 4.2** Median per diversity set of mean exon or non-exon coverage per accession mapped to either the *A. hypochondriacus* or *A. cruentus* reference genome. Mapping performed to complete genome and main reference chromosomes only (Excl. scaffold). Diversity sets consisting of raw and trimmed sequences of the same 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). ..... 122
**Table 4.3** Number and proportion of accessions in each diversity set showing mapping rates of <10x, 10-15x or >15x coverage to the two reference genomes. Diversity sets consist of raw and trimmed sequences of the same 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Mapping was performed to exonic and non-exonic regions and to the complete genome and main chromosomes only (No scaffolds). Species column describes number and proportion of diversity set accessions mapped to a reference genome of the same species as the accession is assigned to (either *A. hypochondriacus* or *A. cruentus*) per investigated mapping rate (<10x, 10-15x or >15x). If these numbers of accessions changed due to updated species re-classifications, novel numbers are in brackets. If the number of accessions did not match comparing within mapping to (non-) exonic regions or (in-) complete genome, values are coloured. ..... 123
**Table 4.4** Number of accessions (Acc.) and SNPs called per main chromosomes (On xx chr.) and scaffolds (On Scaff.) including the number of scaffolds (No. Scaff.) variants were called towards of two reference genomes (*A. hypochondriacus* (Lightfoot *et al.*, 2017) or *A. cruentus* (Ma *et al.*, 2021)). These are results derived from final output VCF files per individual diversity set or sets combined. Diversity sets are a total of 90 accessions grown by the Agricultural Research Council (ARC), sufficient quality externally obtained sequences (Stetter *et al.*, 2020). ..... 131
**Table 4.5** Certain re-assignment of species and related groups based on combined phylogenetic analysis (Admixture k=9 **Figure 4.5**, PCA **Figure 4.7** and maximum likelihood tree **Figure 4.10**). Three accessions imported from Stetter *et al.* (2020) could be re-assigned while 38 from the ARC set of which 17 previously unknown species (*A. sp.*) could be re-named. ..... 144
**Table 4.6** Twelve diversity set accessions obtained from the WVC (AVRDC) for which the ARC ID, external name, starting point species and their verification by the AVRDC website (https://genebank.worldveg.org/#/) is given. These first four columns are the same as in **Supp table diversity set**. In case found, the AVRDC ID, but also species annotation as analysed in Lin *et al.* (2022) and lastly our own species annotation results (**Table 4.5**) are provided in the last three columns. The cells in the last columns coloured green if our own annotation results match the AVRDC gene bank and publication results and coloured orange in case of no match. An '-' in the 'Species in Lin *et al.* (2022)' column means the accession was not incorporated in the publication analysis. '-' in the last column means we choose to keep the starting point species for this accession. ..... 146

**Table 4.7** Metabolite classes of 486 metabolites annotated by SIRIUS4 and/or GNPS and numbers of those occurring per class in the dataset. If metabolites were annotated with a name only by SIRIUS4, PhytoHub, KNApSAcK or ChemSpider but no class could be annotated those are assigned as 'other'..... 159
**Table 4.8** Y-axis metabolite based accession cluster, number of accessions in that cluster (No acc. cluster tot.), species of those accessions, number of accessions of that species in that cluster (No acc. cluster spec.) and number of accessions of that species in total (No acc. species tot.) of bi-clustering analysis in **Figure 4.20 B**. Fisher exact test results indicating if species are over-/underrepresented in a cluster are provided with significant values ($p < 0.05$) marked green, just as the odds ratio of how much more (or less) a species is to occur in this cluster compared to others..... 161
**Table 4.9** X-axis metabolite clusters (1-22) and their numbers of metabolites assigned to them and the total (TOT) number of metabolites in bi-clustering analysis of **Figure 4.20 B**. ..... 162
**Table 4.10** Number of significant SNPs per metabolite (top row) and number of metabolite traits (bottom row) which had this number of significant SNPs associated with them. ..... 163
**Table 4.11** Chromosome (Chr), metabolite ID's, base pair locations (bp loc.), significance by Wald p-value (Sig.), metabolite annotations (Anno.) and bp distance (Dist.) between two SNPs significantly associated with a different metabolite trait but located within a 1000 bp distance on the same chromosome. These two sets of 14 SNPs are ordered by bd distance (increasing top to bottom) in therefore decreasing relevance of likeliness of involvement in the same biosynthetic pathway. For metabolite M225.1132T225 many different annotations could be given by PhytoHub hence no further name is given. The 'Gene' column indicates gene ID's (excluding AmacrXXg, with XX as chromosome number) of *A. cruentus* reference genes both sig. SNPs in that row were located in..... 166
**Table 4.12** 26 Metabolite ID's (Masstag ID), as a subset of 81 metabolites for which a significant SNP could be found. The metabolites are ordered with highest significance top to bottom and provided with their annotated name by SIRIUS4 (*most reliable), PhytoHub, KNApSAcK or ChemSpider databases, subclass and class according to SIRIUS4 or GNPS, lowest Wald-P value of the most significant SNP found for this metabolite according to mGWAS, chromosome this most significant SNP is located at (Chr.), estimated heritability (Her.) plus standard deviation (SD) and mean normalised metabolite data value over 90 accessions plus it's SD. Masstag ID's in bold are selected for having low p-values, considerate high heritabilities with non-extreme SD's while having good quality data according to Q-Q plots and therefore Manhattan plots are visualised in **Figure 4.24**. ..... 170
**Table 5.1** The 43 identified putative biosynthetic gene cluster (BGC) ID's ordered by number of (No) accessions they occur in. The No of BGC found per No accession group is provided. It is mentioned between brackets which BGC ID is specific to *A. h.*, *A. c.* or MAGIC parent accessions only..... 194
**Table 5.2** Selected putative BGC of interest based on their consistency in gene expression patterns across multiple accessions (**Figure 5.5**). The putative BGC ID, type, the number of (No) genes, protein types or core domains in enzymes encoded by the putative BGC, and No accessions the putative BGC is detected in with plantiSMASH (Kautsar *et al.*, 2017), is provided. ..... 199

# Author's declaration

I declare that this thesis is a presentation of original work, and I am the sole author (unless stated otherwise, see **Statement regarding chapter 2 and submitted manuscript**). This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

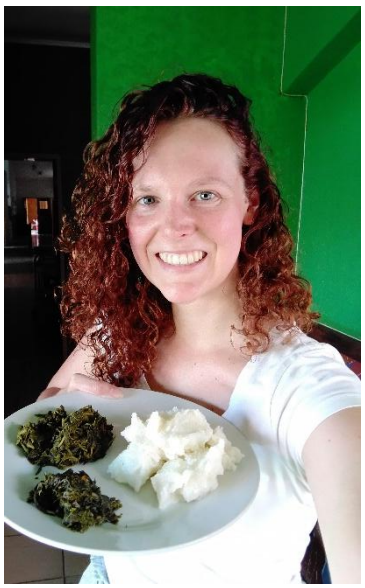
# Acknowledgements

Firstly, I would like to thank my main supervisor Prof. Katherine Denby for her great support and contributions to successfully guide me through the PhD. Just as all the Denby lab members, some of them who were there at the start (James, Harry, Fatih and Jun), some at the end (Maria, Katherine W., Luisa, Feixue, Salman and Sam) and Fabian who was always there. You've all contributed both scientifically and socially, by answering questions and providing research suggestions as well as by the much appreciated Friday lab drinks.

Additionally, the lab of my second supervisor, Dan Jeffares, has been of great assistance, especially the help of Joao Cunha benefitted the genomics analysis of the diversity set chapter significantly.

All the visits and meetings I made to external institutions and researchers like the VIB in Belgium and ARC in South Africa have been a great contribution to the PhD journey and have aided the research outcomes of this thesis. Like Xiao Ma who conducted a lot of genome analysis and created the *Amaranthus* pan-genome, while Willem S. Jansen van Rensburg was kind enough to drive me around South Africa to visit and collect data at all the field trial sites.

I'm grateful for all the friends I made and met during the PhD, everyone part of the Bio-Goth Inner Circle with whom we started off together and who were there until the end. Especially bus stop friend Lydia Barber, Charlay Wood for road tripping and my housemates Gemma Banister and Isaac Reynolds. The house parties at Frances Street were a blast.



Of course, all the events of the Dutch society (NWS) have been an important part of my PhD life, and it was a wonderful experience to be president of this group for three years. I truly appreciated not having to leave my heritage fully behind and being able to connect and build friendships with fellow Dutchies during this adventure, especially with Amy Nieuwenstein and Lianne Lansink as close friends.

Lastly, I would like to thank my family, who all at some point came to visit me in York and each time welcomed me back again when I stayed at home during Christmas or for other longer periods in the Netherlands.

# Chapter 1: Introduction and literature review.

## 1.1 Project relevance and gaps in research

Multiple factors, like population growth or climate change could have adverse effects for safeguarding food security in developing countries. The global population is expected to increase from 7.3 billion today to about 9.8 billion by 2050 (FAO, 2017b). The largest increases are expected in developing countries, most particularly sub-Saharan Africa and Asia. It is those regions where small-scale and family farmers produce 80% of the food supply (FAO, 2017b). In South Africa specifically, smallholder farmers are potential key drivers for achieving food security (Oluwatayo, 2019). But unfortunately, the prevalence of undernourishment is highest in Africa (FAO, 2017a).

Amaranth cultivation by small-scale farmers, a widely grown and eaten leafy vegetable in Southern Africa known for having positive nutritional characteristics, could offer solutions for food sovereignty issues (Mabhaudhi *et al.*, 2016). Currently, amaranth is commonly harvested from the wild, and traded amongst locals and smallholder farmers in South Africa and elsewhere on the continent (Grubben, 2004; Mabhaudhi *et al.*, 2016; Emmanuel and Babalola, 2021). However, there is potential for amaranth breeding and crop improvement as broad genetic diversity can contribute to nutritional enhancement (Rastogi and Shukla, 2013; Shukla *et al.*, 2018; Schafleitner *et al.*, 2022). For example, the consumption of leafy amaranth is widely promoted to tackle problems like under-nutrition and micronutrient deficiencies in sub-Saharan Africa (Achigan-Dako, Sogbohossou and Maundu, 2014). Additionally, amaranth is selected as a forgotten food crop which could be a candidate to not only contribute to healthy diets but also climate-resilient food systems in sub-Saharan Africa (van Zonneveld *et al.*, 2023). Leafy amaranth is known to grow with little inputs or labour and is easy to cultivate (Achigan-Dako, Sogbohossou and Maundu, 2014). Conducting $C_4$ photosynthesis makes amaranth highly beneficial due to adaptivity to drought stress (Kadereit *et al.*, 2003; Valdayskikh *et al.*, 2019; Motyleva *et al.*, 2021). $C_4$ photosynthesis is typically more efficient under warm temperatures or water stress due to reduced oxygenation activity in of the enzyme Rubisco (Edwards, Franceschi and Voznesenskaya, 2004) which can be relevant considering prospects of climate change.

Amaranths are not only known to be heat, drought and pest resistant, additionally leaves contain high levels of antioxidants, micronutrients (iron, calcium, zinc and essential vitamins) as well as macro components (protein, lipids and carbohydrates) (Venskutonis and Kraujalis, 2013; Achigan-Dako, Sogbohossou and Maundu, 2014). In Africa and other developing countries, micronutrient deficiencies among children and woman are known to commonly cause malnutrition and lower birth weights (Rastogi and Shukla, 2013; Achigan-Dako, Sogbohossou and Maundu, 2014). Major staple food crops like maize, wheat or rice might be calories rich but are not sufficient to meet the requirements for essential nutrients and protein content (Joshi *et al.*, 2018). Amaranth is a known underutilized crop which has a high potential for meeting this demand for diverse micronutrient profiles, as a pseudocereal, but also definitely as a leafy vegetable in Africa (Rastogi and Shukla, 2013; Achigan-Dako, Sogbohossou and Maundu, 2014; Mabhaudhi *et al.*, 2016; Joshi *et al.*, 2018). Protein, calcium, iron and many other nutrient levels in amaranth are higher than in other leafy vegetables like spinach, basella, chard, cabbage and Chinese cabbage (Ebert, Wu and

Wang, 2011; Rastogi and Shukla, 2013). Amaranth can be a cheap source of protein as composition studies show leaves have a well-balanced amino acid profile (Rodriguez *et al.*, 2011; Rastogi and Shukla, 2013). The essential amino acid lysine in protein even ranges 40–50 g/kg (Rastogi and Shukla, 2013).

When comparing amaranth to other forgotten food crops in sub-Saharan Africa various types of leafy amaranths were among the highest reported concentrations for iron, zinc and folate (van Zonneveld et al., 2023). Spider plant (*Cleome gynandra*), Ethiopian mustard or kale (*Brassica carinata*) and jute mallow (*Corchorus olitorius*) are examples of such forgotten but popular leafy vegetables which could support climate-resilient and health diets in sub-Saharan Africa (van Zonneveld et al., 2021, 2023). Amaranth competes with such crops due to a similar position in this African region as a traditional leafy vegetable. While amaranth together with spider plant contains the highest reported iron concentration, this cleome spider plant, just as leafy vegetables Arrowleaf elephant ear (*Xanthosoma sagittifolium*) and Moringa (*Moringa oleifera*) contain higher vitamin A contents (van Zonneveld et al., 2023). While amaranth is one of the most-mentioned traditional African vegetables it appears to have the poorest conservation status, raising concerns for sustaining crop improvement programmes (van Zonneveld et al., 2021).

Despite the promising qualities of amaranth, there are still many remaining challenges to overcome. Currently there is no large-scale production of amaranth in South Africa, causing a lack of data collection, inadequate agronomic knowledge, non-structured marketing and neglect in research (Emmanuel and Babalola, 2021). Absence of accurate breeding policies hampers amaranth's potential for development of a sustainable and healthy food system in South Africa (Mabhaudhi *et al.*, 2019). Moreover, it occurs that popularity and usage of amaranth in South Africa seems to decrease due to younger people preferring other food over these traditional vegetables (Nyonje *et al.*, 2022). Amaranth is often considered an underutilised crop, present across various agro–ecologies in Southern Africa, with varying performance depending on these diverse environmental conditions (Mabhaudhi *et al.*, 2016). Despite the availability of resources like germplasm collections and genomic references, amaranth is often considered an orphan crop, and genetic improvement of amaranths has been limited with only a few traits being characterized at the molecular and genetic level (Riggins and Mumm, 2021). Those traits which have been targeted are usually relevant to grain amaranths and not leafy varieties (Anuradha *et al.*, 2023).

Amaranth is considered incompletely domesticated and therefore lacks full adaptation to agro-ecological environments as well as progress in their genetic gain limiting yield potential (Stetter, Vidal-Villarejo and Schmid, 2020; Anuradha *et al.*, 2023). Regarding early investigation of amaranth speciation and evolution, there have been major defects while recording and recognizing the variety of amaranth species by taxonomists (Sauer, 1957). Even nowadays characteristics like hybridization, plasticity, gene flow and only a small number of useful phenotypic traits, make taxonomy of amaranth difficult (Stetter and Schmid, 2017). Additionally, the scientific investigation of metabolites in leafy amaranth, and identifying their nutritional value is in its infancy (Sarker and Oba, 2020; Ma *et al.*, 2021). Linking nutritionally relevant metabolites to genetic data hasn't been performed yet, while this might give a clearer scientific understanding of certain potentially important amaranth gene families (Ma *et al.*, 2021). For all these previously mentioned challenges like Amaranth breeding, cultivation, taxonomy, genomics, and exploration of nutritional values, further

elaborate investigation of amaranth varieties is highly needed and is promising to contribute to fill up important gaps in knowledge and science.

## 1.2 Current status of amaranth in scientific literature

### History and background

Amaranth originates from the Americas, where it was mainly grown for grain purposes by ancient native American civilizations (Sauer, 1950; Stetter, Vidal-Villarejo and Schmid, 2020). In South America, archaeological findings revealed amaranth grain collection roughly 8000 years ago (Arreguez, Martínez and Ponessa, 2013). In North America, amaranth seed findings date back to about 6000 years ago (MacNeish, 1965). During these pre-Columbian times grain amaranth was one of the basic foods and it was grown on thousands of hectares of Aztec, Inca, and other native American farmland (Schultes, 1984). With the arrival of the Spanish conquistadors who were shocked by pagan rituals involving amaranth, the crop fell in disuse and was only limitedly cultivated in mountain areas of Mexico and the Andes (Schultes, 1984). However, it is likely the Spanish took Amaranth seed to Europe (possibly inadvertently) where it first spread as an ornamental garden plant in the 16$^{th}$ and 17$^{th}$ century but also as a minor grain crop in central Europe and Russia around 1700 (Schultes, 1984). How Amaranth further spread to Asia and Africa remains unclear, but it had been taken there by the early 19th century and is a popular grain crop in many widespread regions (Schultes, 1984). Prior a more recent scientific resurgence of interest in Amaranth as a highly nutritional crop, grain amaranth was mostly cultivated small scale by farmers in Mexico, Guatemala, Peru, northern India, and Nepal, and vegetable amaranth was grown in China, Southeast Asia, southern India, West Africa, and the Caribbean basin (Tucker, 1986).

Nowadays, amaranth is widely consumed as a leafy vegetable in sub-Saharan Africa, with consumption concentrated in rural areas (Grubben, 2004; Emmanuel and Babalola, 2021). In Kenya and Tanzania leafy amaranth consumption is on the rise, while the contrary is reported for South Africa with a decline in consumption and usage of amaranth (Nyonje *et al.*, 2022). Consumption of amaranth is concentrated in rural areas where it is harvested from the wild, with no large-scale production causing a lack of data, knowledge of usage, research efforts and organized markets making amaranth an underutilized crop (Emmanuel and Babalola, 2021). Underutilized can mean undervalued or unrecognised as well as under-researched or unimproved, with limited formal research and breeding classifying amaranth as an underutilized crop in South Africa (Mabhaudhi *et al.*, 2016). Typically, amaranth species appear in great variability, with hybridization and introgression being possible (Sauer, 1957; Stetter, Vidal-Villarejo and Schmid, 2020). Hence, past attempts for clear taxonomy of species of the *Amaranthus* genus have been challenging (Sauer, 1957). This difficulty in morphological taxonomy makes the genus challenging for scientific understanding as well as for performing adequate breeding strategies by identification of varieties. For all previously mentioned reasons, more research regarding amaranth genetics and its breeding potential in South Africa, is urgent.

## Species evolution and domestication

### Taxonomy and domestication

Taxonomy of the *Amaranthus* genus is difficult, due to historical lack of recording, not many suitable traits, high phenotypic plasticity, gene flow and hybridization (Sauer, 1957; Stetter and Schmid, 2017). Many species do exist, with about 60 species native to the Americas and about 15 to Europe, Asia, Africa and Australia (Sauer, 1967). More recent publications state that the genus *Amaranthus* contains of 50–70, or even ~74 species, consisting of cultivated as well as weedy kinds with great economic importance (Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). The genus *Amaranthus* is generally recognized to be subdivided into three subgenera based on morphology and phenotypic characteristics; *Amaranthus Albersia, Amaranthus Amaranthus* and *Amaranthus Acnida* (Mosyakin and Robertson, 1996; Costea and DeMason, 2001; Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). However, more recent studies show that *Amaranthus Acnida* can be subdivided into two highly differentiated clades (**Figure 1.1**) (Stetter and Schmid, 2017) and that there are small monophyletic groups of *Amaranthus* species outside of the three recognized subgenera (Waselkov, Boleda and Olsen, 2018).

Despite selection due to human cultivation for predominant white grain over dark brown seeds in wild amaranth, grain amaranth is considered an incompletely domesticated crop (Stetter, Müller and Schmid, 2017). Three cultivated grain amaranth species are *A. caudatus*, *A. cruentus* and *A. hypochondriacus*, with *A. hybridus*, *A. quitensis* and *A. powellii* being originally identified as three wild relatives and putative progenitors of the three cultivated species (Sauer, 1967; Stetter, Vidal-Villarejo and Schmid, 2020). *A. caudatus* and *A. quitensis* originate from South America, *A. cruentus* and *A. hypochondriacus* from Central America, while *A. hybridus* covers both regions (Stetter et al., 2017, 2020). Most later genetic research supports that *A. hybridus* is the wild ancestor of the three closely related grain amaranth species (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*) (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Clouse *et al.*, 2016; Stetter and Schmid, 2017; Graf *et al.*, 2025). Previous uncertainties regarding separate or a single domestication from *A. hybridus* were cleared, as three separate domestication events with independent selection for white seed colour is supported by analysis (Stetter et al., 2020). Another older recognized phenotype associated with domestication is altered regulation of the waxy gene, responsible for amylose synthesis in perisperm starch which was proven to be derived from a single mutational event (Park *et al.*, 2012). Appearance of this domestication trait by mutation also occurred independently in each of the three grain amaranths, with a different frameshift or base substitution per species causing complete loss of gene function (Park *et al.*, 2012). A more recent finding is that the overall amount of accessible chromatin, which is relevant for regulating gene expression, is mostly highly conserved between two wild (*A. hybridus* and *A. quitensis*) and the three grain amaranths (Graf *et al.*, 2025). However, about 2.5% of chromatin switches states, and these differentially accessible chromatin regions are species specific and significantly associated with selective sweeps which reflects the repeated independent domestication of *A. caudatus*, *A. cruentus* and *A. hypochondriacus* amaranth species from wild *A. hybridus* (Graf *et al.*, 2025).

### ‘hybridus’ species complex

Originally, what was called the ‘hybridus’ species complex, consisted of the six previously mentioned amaranth species (three grain species and their three wild relatives) (Costea and DeMason, 2001; Xu and Sun, 2001). The species have relatively uniform stem epidermal features and were impossible to separate based on stomata or trichome characteristics, excepting types incorporating *A. powellii* (Costea and DeMason, 2001). Studies using genetic markers and DNA clones encoding starch branching enzymes (SBEs), show that *A. powellii* did not seem to be closely related to any of the three grain species (Xu and Sun, 2001; Mallory *et al.*, 2008; Park *et al.*, 2014; Stetter and Schmid, 2017). Therefore, nowadays, the ‘hybridus’ species complex consists of the five remaining species (three grain amaranths, *A. hybridus* and *A. quitensis*) (Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017). The exon and intron structures of the SBE genes in the three grain amaranths (*A. caudatus*, *A. cruentus*, and *A. hypochondriacus*) seems the same, with similar gene lengths and high homology, which is an example of the similarity among the ‘hybridus’ species complex grain varieties (Park *et al.*, 2014).

Plants part of the ‘hybridus’ species complex are considered difficult to identify (Adhikary and Pratt, 2015). A study shows that weedy *A. quitensis* appears to be its own species, it has crossed with *A. caudatus* and that *A. caudatus* and *A. hypochondriacus* seem to be closely related (Kietlinski *et al.*, 2014). Based on SBE sequence data, some close relationships were identified between *A. caudatus* and *A. hypochondriacus*, but also *A. caudatus* and *A. quitensis* (Park *et al.*, 2014). Additionally, a study shows a weak genetic differentiation of *A. caudatus* from *A. quitensis*, however *A. hybridus* is more closely related to *A. caudatus* and *A. quitensis* than *A. caudatus* is to *A. quitensis* (Stetter et al., 2020). *A. hybridus* remains identified as the common ancestor of all three grain species (Chan and Sun, 1997; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020). However, *A. quitensis* might additionally be involved in the evolution of *A. caudatus* (Xu and Sun, 2001; Park *et al.*, 2014; Stetter and Schmid, 2017).

### Phylogenetics

A consensus phylogenetic tree including the three subgenera of the genus *Amaranthus,* based on a very complete analysis of 35 *Amaranthus* species, is provided in **Figure 1.1 A** (Stetter and Schmid, 2017). A species tree (blue) including cloudogram (green), identifying the certainty of clades in the tree, in **Figure 1.1 B** (Stetter and Schmid, 2017). Both trees in **Figure 1.1** are based on thousands of genome-wide markers derived by genotyping by sequencing (GBS), and for phylogenetic analysis SNAPP (SNP and AFLP Package for Phylogenetic analysis) was used which infers species trees directly from these biallelic markers (Stetter and Schmid, 2017). *A. Acnida* consists of two highly separated clades which is not only due to strong variety in molecular phylogeny but they also show different average genome sizes, either largest (658.2-824.2 Mega base pairs (Mbp)) or smallest (421.8-471.6 Mbp) within the genus *Amaranthus* (**Figure 1.1 A**: bottom versus top clades respectively) (Stetter and Schmid, 2017).

The three grain species and their three wild relatives all belong to the subgenus *A. Amaranthus* (**Figure 1.1 A**) (Mosyakin and Robertson, 1996; Costea and DeMason, 2001; Stetter and Schmid, 2017). According to this analysis, the five species of the ‘hybridus’

species complex can generally be well separated from other amaranth species (**Figure 1.1**) (Stetter and Schmid, 2017). Within the complex uncertainty was high among *A. caudatus* with its related wild putative ancestors *A. quitensis* and *A. hybridus*, but with a clear split for these three species compared to the Central American species (*A. cruentus* and *A. hypochondriacus*) (cloudogram, **Figure 1.1 B**) (Stetter and Schmid, 2017). Another uncertainty regards positioning of the *A. deflexus*, *A. tricolor* and *A. crispus* group of species in the tree (cloudogram, **Figure 1.1 B**) (Stetter and Schmid, 2017). Clustering of the genus *Amaranthus* into four clades is highly supported by a Maximum likelihood tree based on genome size evolution (**Figure 1.1 A**) as well as a species inferred tree based on a multispecies coalescent approach (**Figure 1.1 B**). (Stetter and Schmid, 2017). This provides complementary evidence of *Amaranthus* clades based on genome size evolution as a trait with the statistical maximum likelihood estimating a probable tree (**Figure 1.1 A**) as well as a tree which estimates relationships among species according to the coalescent theory that the sampled markers originate from a common ancestor (**Figure 1.1 B**).

A recent study states that the phylogenetic tree based on concatenated loci from the nuclear genome (Waselkov et al., 2018: Figure 1A) is comparable to the SNAPP consensus tree of (Stetter and Schmid, 2017) (**Figure 1.1 A**). Trees in **Figure 1.1 A** and **Figure 1.1 B** generally overlap with previously published Amaranthaceae phylogenetic trees based on SBE sequences, molecular sequences and complete chloroplast genomes (Park *et al.*, 2014; Berger *et al.*, 2016; Xu *et al.*, 2020). However, small differences are *A. powellii* and *A. wrightii* being more related to each other than *A. retroflexus* (Berger *et al.*, 2016) and different placement within the 'hybridus' species complex lacking the clear split of *A. cruentus* and *A. hypochondriacus* from other species (Park *et al.*, 2014; Berger *et al.*, 2016). Additional recent research using complete chloroplast genomes, placed *A. tricolor* and *A. viridis* similarly in a separate clade, but inconsistently indicated that *A. dubius* is sister to a clade of *A. hypochondriacus* and *A. caudatus* in a phylogenetic tree (Xu *et al.*, 2021). Most contradictory recent results based on genome-wide microsatellites clustered three clades: 1) *A. viridis* and *A. edulis*, 2) *A. retroflexus, A. cruentus* and *A. hypochondriacus* and 3) *A. lividus, A. dubius*, *A. caudatus*, *A. hybridus* and *A. tricolor* with the latter two species having highest similarity (Tiwari *et al.*, 2021). Inconsistency in literature and the cloudogram (**Figure 1.1 B**) indicate that there is still high uncertainty about the evolutionary relationships between *Amaranth* species, and further investigation is needed. A common way to look into these phylogenetic relationships within *Amaranthus* is by GBS and genomic markers in multiple individuals expected to be of different species in a diversity set, with any novel sequences being beneficial to further understanding of the genus (Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Lin *et al.*, 2022; Blair *et al.*, 2023).

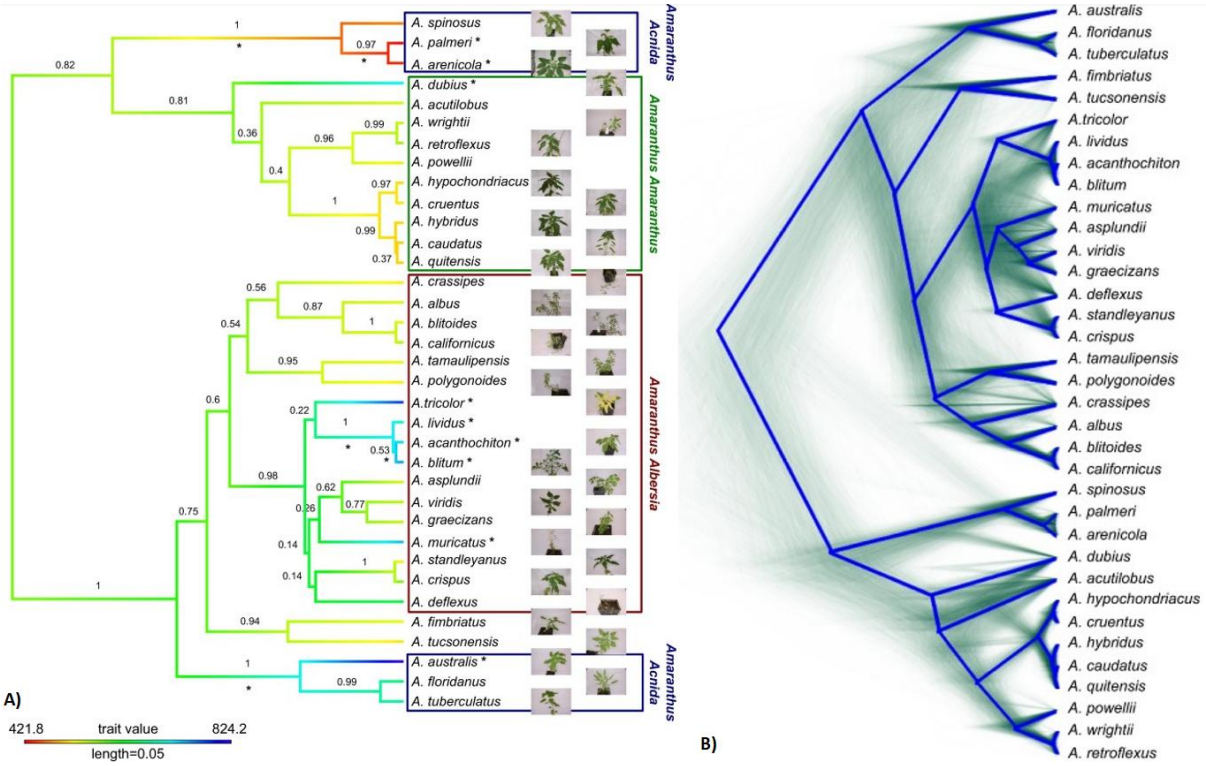



**Figure 1.1** Trees of the genus *Amaranthus* made conducting SNAPP (SNP and AFLP Package for Phylogenetic analysis) with genotyping by sequencing (GBS) of 94 accessions presenting 35 species. **A)** Consensus tree including genome size estimates of interior nodes obtained with a Maximum Likelihood method mapped onto it, Mbp displayed by branch colour, groups are subgenera, and asterisks (∗) indicating deviation from random genome size evolution at 95% confidence based on a two-tailed test. **B)** Species tree based on the multispecies coalescent calculated with SNAPP, consisting of consensus tree (blue) and cloudogram (green lines), the later representing 3980 individual trees (Stetter and Schmid, 2017).

## Amaranth genomics

### Genome size and polyploidization

Genome size evolution of amaranth species was mapped onto the consensus tree in **Figure 1.1 A**. Significantly larger genome size was observed in *A. tricolor* (782.7 Mbp) and *A. australis* (824.2 Mbp), originally suggesting recent polyploidization in these species (Stetter and Schmid, 2017). However, with a more recently estimated genome size of only 520 Mb by sequencing an *A. tricolor* accession for high quality reference genome creation, this previous observation is no longer supported (Wang *et al.*, 2022). *A. australis* and its relatives (*A. florianus* and *A. tuberculatus*) were predicted to have larger genome sizes than most species (658.2-824.2 Mbp) while *A. palmeri* and its relatives (*A. spinosus* and *A. arenicola*) are expected to have the smallest genome sizes (421.8-471.6 Mbp) of the genus *Amaranthus* (Stetter and Schmid, 2017). According to this study the separation of the *A. Acnida* clades was therefore likely based on independent occurrences of polyploidization versus sequence loss (**Figure 1.1 A**) (Stetter and Schmid, 2017). This separation of the *A. Acnida* clade is backed up by an older study using nuclear and chloroplast PCR-amplified loci (Lanoue *et al.*, 1996), as well as proof that controlled crossing of *A. palmeri* and *A. tuberculatus* accessions largely results in either nonhybrid progeny or nonviable hybrids (Trucco *et al.*, 2007). While polyploidization appears to explain the observed genome size differences within the *A. Acnida* clade, allopolyploidization via interspecific hybridization may also have contributed. Lastly, the genome size of *A. dubius* was also observed to be high

(711.9 Mbp) (**Figure 1.1 A**), and thus albeit the indicated low variety of genome sizes within the genus, *A. australis* and *A. dubius* are exceptions being tetraploid species (Stetter and Schmid, 2017). This is consistent with the existing literature that *A. dubius* is tetraploid (Grant, 1959; Behera and Patnaik, 1982; Lin *et al.*, 2022). Despite *A. dubius* having a high chromosome number of 2*n*=64, it can produce spontaneous hybrids with diploid amaranths, although the triploid often sterile offspring have low agricultural relevance (Grant, 1959; Behera and Patnaik, 1982).

Generally, little variation in genome sizes is observed, suggesting cytological stability of the genus *Amaranthus*, indicating that polyploidization does not play a major role in recent speciation within Amaranth (**Figure 1.1 A**) (Stetter and Schmid, 2017), especially since polyploidization of a larger *A. tricolor* genome is no longer supported (Wang *et al.*, 2022). This builds on the existing knowledge that the last polyploidization event was a single whole genome duplication (WGD) which took place in a common ancestor of Amaranths prior to speciation of the genus (Lightfoot *et al.*, 2017; Ma *et al.*, 2021; Wang *et al.*, 2022). This WGD was previously estimated 18-34 million years ago (MYA) based on *A. cruentus* and *A. hypochondriacus* references (Ma *et al.*, 2021), with a more recent estimation of 26.56 MYA including an *A. tricolor* genome (Wang *et al.*, 2022).

### Genome-trait associations

Regarding genomic variation responsible for traits, two major genomic regions controlling seed colour adaptation were identified on chromosome 3 and 9 in three amaranth species (*A. caudatus*, *A. hypochondriacus* and *A. cruentus*) (Stetter et al., 2020). On chromosome 3 five candidate genes were identified, while on chromosome 9 a larger region with 50 genes of interest was identified. Within that region on chromosome 9 a MYB-like transcription factor gene was found, which is a known regulator for seed colour in plants (Stetter et al., 2020). Additionally, a soft selective sweep is detected in that genomic region during the rise of white seed colour, but only in the grain species *A. hypochondriacus* (Stetter et al., 2020). A possible explanation for the larger (~600 kb) region on chromosome 9 significantly associated with seed colour is a potential structural variant in *A. hypochondriacus*. This was discovered by alignment of a white seeded *A. hypochondriacus* reference genome (Lightfoot *et al.*, 2017) and pseudo scaffolded wild relative *A. tuberculatus* (Kreiner *et al.*, 2019b), displaying an inverted segment between the two genomes on the specific region of chromosome 9 (Stetter et al., 2020). Ongoing gene flow from wild into grain amaranths might have prevented the fixation of white seed colour, which as a key domestication trait suggesting incomplete domestication of the grain amaranths (Stetter et al., 2017, 2020).

Being able to map a genomic variant determinant for seed colour in Amaranths is just an example of how genomic data can be relevant for the scientific understanding and breeding strategies. Despite amaranth being an underutilized crop, genomic and biotechnological tools are becoming available that allow for rapid improvement of it (Joshi *et al.*, 2018). The first genetic map of *Amaranthus* was ever created was by SNP assays based on KBiosciences Allele Specific PCR (KASP)ar genotyping chemistry detected on the Fluidigm dynamic array platform, which spans 1288 cM with an average marker density of 3.1 cM per marker (Maughan *et al.*, 2011). It was 41 accessions which were used to create the map, of which three cultivated species; 11 *A. caudatus* accessions, 10 *A. cruentus* accessions and 10 *A. hypochondriacus* accessions, and one putative ancestor species; 10 *A. hybridus*

accessions, all part of the *hybridus* species complex (Maughan *et al.*, 2011). The map was of potential relevance towards genetic dissection of agronomically important traits in amaranth (Maughan *et al.*, 2011). An example of KASP marker usage is the ability to classify grain amaranth species and differentiating them from weedy species by revealing the phylogenetic relationships between accessions (Thapa, Edwards and Blair, 2021).

### Amaranth genomic resources

Just like the genetic map, many other relevant genomic resources of amaranths have been created in recent years. Several *Amaranthus* chloroplast genomes are available (Chaney *et al.*, 2016; Hong *et al.*, 2019; Xu *et al.*, 2020, 2021), with the recent most complete version being that in (Xu *et al.*, 2022). Chloroplast genomes are known to have a compact size, little recombination, and mostly uniparental inheritance (Birky Jr., 2001). The most complete *Amaranthus* chloroplast genome made of 22 samples from 17 species of three subgenera confirmed stabilization and limited recombination of *Amaranthus* chloroplast genes by observing a regular distribution of repeats (Xu *et al.*, 2022). Generally, this chloroplast genome is of relevance for the identification of similar *Amaranthus* species, improving evidence and clarification of evolution and phylogenetic relationships within the genus (Xu *et al.*, 2022). Additionally, a variety of transcriptomes have been created for different *Amaranthus* species like *A. hypochondriacus* (Délano-Frier *et al.*, 2011), *A. tuberculatus* (Riggins *et al.*, 2010) and *A. palmeri* (Salas-Perez *et al.*, 2018). These transcriptomes are relevant genetic resources for varying purposes. Like understanding (a)biotic stress-resistance and its underlying mechanisms in *A. hypochondriacus* and other nutritionally important crops (Délano-Frier *et al.*, 2011). While the other two transcriptomes are used for amaranths which are considered weeds, and investigate their stress response and resistance to herbicides (Riggins *et al.*, 2010; Salas-Perez *et al.*, 2018).

Apart from chloroplast genomes or transcriptomes, attempts have been made to generate more complete reference genomes of several different amaranth species. A pseudo scaffolded reference genome created by Pacific Biosciences (PacBio) long-reads was made for *A. tuberculatus* with a total genome size of 663 Mb (Kreiner *et al.*, 2019b, 2019a), just as initial assemblies were created for *A. palmeri* (Molin, Patterson and Saski, 2020). The most recent draft genomes of both these species and *A. hybridus*, which are all considered weeds, is published by (Montgomery *et al.*, 2020). Here PacBio long-read sequencing was performed for all three amaranth accessions. However, assembly strategies differed as for *A. palmeri* additional high-throughput chromatin conformation capturing (Hi-C) was used to scaffold sequences to near-chromosome-level resolution, while for *A. tuberculatus* and *A. hybridus* a trio binning approach was used to resolve heterozygosity issues to generate haplotype assemblies (Montgomery *et al.*, 2020). The current more contiguous *A. tuberculatus* genome size was reduced to 573 Mb (Montgomery *et al.*, 2020) compared to the previous 663 Mb (Kreiner *et al.*, 2019b). This could be due to genetic differences responsible for dioecy in this species as the draft genomes were of male and female *A. tuberculatus* individuals respectively (Kreiner *et al.*, 2019b; Montgomery *et al.*, 2020). Both draft genome sizes (663 Mb and 573 Mb) roughly comply with the previously estimated genome size of *A. tuberculatus* (675.6 Mbp) which was considered bigger on average compared to other *Amaranthus* species, a difference originally attributed to polyploidization (Stetter and Schmid, 2017). The more recently generated draft genomes, however, demonstrate that *A. tuberculatus* is diploid rather than polyploid, contradicting previous

assumptions (Kreiner *et al.*, 2019b; Montgomery *et al.*, 2020). While draft genome sizes of *A. palmeri* and *A. hybridus* (408 Mb and 403 Mb) approximately match the previously estimated genome sizes of 421.8 Mbp and 503.8 Mbp respectively, which is average for amaranths (Stetter and Schmid, 2017). An example of how these draft genomes can be used is investigating the evolution and origin of glyphosate (herbicide) resistant versus susceptible *A. palmeri* accessions (Molin, Patterson and Saski, 2020). More recently, after assembling extrachromosomal circular DNA from shotgun sequencing and PacBio HiFi reads, a specific large genomic rearranged variant could be identified which causes resistance to both glyphosate and glufosinate herbicides in *A. palmeri* (Carvalho-Moore *et al.*, 2025). Apart from evolution and identifying genomic variation responsible for herbicide resistance, many other relevant biological questions could be answered using the draft reference genomes potentially easing management of these three weedy species (Montgomery *et al.*, 2020).

### Reference genomes of cultivated amaranths

Apart from weedy species multiple complete reference genomes of agriculturally relevant and cultivated amaranths have been created. For the *A. hypochondriacus* grain amaranth species multiple reference genomes have been created over time. The first draft genome was of a land race (Rajgira) collected locally from farmers in northern Karnataka, India, and was created by Illumina short-read sequencing of leaf tissue, followed by assembly using SOAPdenovo (Sunil *et al.*, 2014). A higher quality reference genome was needed since this initial draft assembly consisted of many (367,441) scaffolds suggesting high fragmentation, and the total assembly length was 40% greater than the expected genome size (466 Mb) suggesting contamination and redundant sequences (Sunil *et al.*, 2014; Clouse *et al.*, 2016). A higher quality draft genome sequence of the *A. hypochondriacus* grain amaranth was created by whole-genome sequencing (WGS) seeds of the cultivar Plainsman (PI558499) originating from Nebraska in the USA (Clouse *et al.*, 2016). It was again short-read sequencing with Illumina which was done, but performing genome assembly with ALLPATHS-LG and additionally GapCloser resulted in an assembly containing 377 Mb, in only 3518 scaffolds (Clouse *et al.*, 2016).

It was suggested that PacBio long-read sequencing could improve this assembly by potentially reducing the issue of appearing smaller scaffolds as well as reduce the number of gaps within the larger scaffolds, just as Hi-C technology could help organizing scaffolds into chromosome-level assemblies and reduce fragmentation (Clouse *et al.*, 2016). Exactly this approach was applied when improving the original *A. hypochondriacus* assembly combining PacBio long-reads for gap filling and polishing, and Hi-C for enhanced contig ordering and orientation (Lightfoot *et al.*, 2017). This improved version of the *A. hypochondriacus* reference genome now has a sequence length of 403.9 Mb, which is 93.5% of the estimated genome size with the number of scaffolds even more reduced to only 892 mostly small sized scaffolds (Lightfoot *et al.*, 2017). This reference genome has been used for much relevant scientific work. For example a population genetic genome browser, made to represent the genetic variation of the five ‘hybridus’ complex species towards this version of the *A. hypochondriacus* genome (Gonçalves-Dias and Stetter, 2021), comparisons made while creating new high quality *Amaranthus* reference genomes like *A. cruentus* (Ma *et al.*, 2021) and *A. tricolor* (Wang *et al.*, 2022), or usage as a reference for read mapping in many

*Amaranthus* phylogenetic studies (Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Lin *et al.*, 2022; Blair *et al.*, 2023).

As just mentioned, apart from this *A. hypochondriacus* grain species, additional high quality reference genomes for agriculturally important amaranths have been made. For example, *A. tricolor* which can be a vegetable as well as ornamental amaranth (Wang *et al.*, 2022) and *A. cruentus* which can be dual usage for grain as well as leaf purposes (Ma *et al.*, 2021). The latter sequenced genome of *A. cruentus*, is from a line called 'Arusha', and was a black-seeded semi-domesticated accession (Ma *et al.*, 2021). It was a combination of Oxford Nanopore long-read, Illumina short-read (Hi-C, 10× Genomics, and short inserts), and phased sequencing technologies which were used to generate a 371 Mb *A. cruentus* genome assembly (Ma *et al.*, 2021). This is slightly shorter than the previous genome of *A. hypochondriacus* (403.9 Mb) (Lightfoot *et al.*, 2017). For the *A. tricolor* genome, it was also a combination of sequencing technologies like PacBio HiFi long-reads, Nanopore ultra-long-reads and Illumina short-reads (Hi–C) used to create a 520 Mb assembly (Wang *et al.*, 2022). Many of the previously mentioned *Amaranthus* references, drafts as well as high quality genomes, known for weedy as well as agricultural signatures, have been used in relevant literature for amaranth researchers. Like an Amaranth Genomic Resource Database (AGRDB) (Singh *et al.*, 2023) and an extensive literature review regarding the amaranth genome (Adhikary, Deyholos and Délano-Frier, 2021).

Using the previous sequences of the *A. hypochondriacus* genome (Lightfoot *et al.*, 2017), the reference has even been more enhanced by correcting indel errors in the assembly with NextPolish (Winkler *et al.*, 2024). Additionally, long-read Iso-Seq circular consensus sequencing (CCS) was conducted to generate full-length Iso-Seq transcripts which improved the completeness of the genome annotation by adding precise isoform-level details (Winkler *et al.*, 2024). Not much later, a completely novel version of the *A. hypochondriacus* Plainsman (PI558499) genome was created by combining high-coverage PacBio HiFi and Oxford Nanopore long-read sequencing, scaffolded with Hi-C data, and annotated using a combination of BRAKER3 ab initio predictions, RNA-seq evidence, and full-length PacBio Iso-Seq transcriptome data (Graf *et al.*, 2025). These new versions of *A. hypochondriacus* genomes are very valuable resources since the number of annotated genes increased due to improved detection (i.e., 23,883 genes in Lightfoot et al. (2017) versus 24,583 in Graf et al. (2025)), with a drastically improved gene annotation completeness and overall improved assembly contiguity. Benchmarking Universal Single-Copy Orthologs (BUSCO) scores are commonly used to assess the completeness of genomic data by comparing it to a set of universal conserved single-copy orthologs from a related species (Manni *et al.*, 2021). The BUSCO scores improved from 80.3% (Lightfoot *et al.*, 2017), to 98.0% (Winkler *et al.*, 2024) and eventually even 98.8% which is the highest annotation completeness for amaranth to date (Graf *et al.*, 2025).

Unfortunately, these two new versions of *A. hypochondriacus* genomes (Winkler *et al.*, 2024; Graf *et al.*, 2025) became available after most of the work outlined in this thesis was carried out. Hence all thesis work is performed using the *A. hypochondriacus* reference genome of Lightfoot et al. (2017). The other two available high quality chromosome-level *Amaranthus* references used in this thesis are *A. cruentus* (Ma *et al.*, 2021) and *A. tricolor* (Wang *et al.*, 2022). A more elaborate overview about these three reference genomes created of agriculturally relevant amaranths is provided below.

## *Amaranthus* genomic references and resources used in this thesis

### Reference genome characteristics

Amaranth is typically a diploid organism. Haploid chromosome-level genomes of three amaranth species, all part of the genus *Amaranthus*, have been used in this thesis. Reference genomes of the closely related species *A. hypochondriacus* (Lightfoot *et al.*, 2017) and *A. cruentus* (Ma *et al.*, 2021) have often been used in this thesis for most relevant work like read mapping as well as making genomic comparisons. The *A. tricolor* reference genome (Wang *et al.*, 2022) has only been used once for identifying single copy genes to create a phylogenetic tree. The *A. cruentus* and *A. tricolor* reference genomes were de novo assemblies (Ma *et al.*, 2021; Wang *et al.*, 2022). Some key genome characteristics of each amaranth species reference genome are provided in **Table 1.1**. Here two versions of the *A. cruentus* reference are provided since the original genome of Ma et al. (2021) (V1) was updated with manual curation to improve the genome annotation creating V2. Only V2 was used in this thesis which from now will be referred to as the reference called 'Arusha'.

**Table 1.1** Characteristics of *A. hypochondriacus*, *A. cruentus* and *A. tricolor* reference genomes used for analysis in this thesis (Lightfoot et al., 2017; Ma et al., 2021; Wang et al., 2022). *A. cruentus* V1 refers to the previous publication of (Ma et al., 2021) while V2 is an updated version of V1 with improved genome annotation.

| | | ***A. h.*** | ***A. c.* V1** | ***A. c.* V2** | ***A. t.*** |
|---|---|---|---|---|---|
| Genome assembly | Chromosome number | 16 | 17 | 17 | 17 |
| | Genome size (Mb) | 403.9 | 370.9 | 370.9 | 520 |
| | Scaffold number | 892 | 625 | 625 | 48 |
| | Contig N50s size (Mb) | 1.25 | 1.02 | 1.02 | 0.91 |
| Gene annotation | Protein coding genes | 23,883 | 25,477 | 23,814 | 27,813 |
| | Mean gene length (bp) | 4,102 | 4,337 | 4606 | 4,893 |
| | Mean CDS length (bp) | 1,066 | 1,141 | 1,240 | 1,246 |
| | Mean exon per gene | 4.87 | 4.86 | 5.19 | 5.30 |
| | Busco score (complete) | 80.3% | 90.5% | 94.8% | 97.5% |
| Repeats | Repetitive genome | 48% | 58% | 56% | 68% |
| | Transposable elements (TEs) | 24% | 35% | 22% | 37% |
| | Long terminal repeats (LTRs) retrotransposons | 12% | 25% | 12% | 28% |
| | Tandem repeats | 5% | 4% | 2% | 23% |

The manual curation reduced the number of annotated genes from 25,477 to 23,814 but increased the average gene length from 4,337 to 4,606 base pairs (bp) (**Table 1.1**). The total percentage of repetitive genome stayed roughly the same (58% to 56%), although the percentage of transposable elements (TEs) and long terminal repeats (LTRs) reduced from respectively 35% and 25% to 22% and 12% (**Table 1.1**). The BUSCO score increased quite significantly by manual curation of *A. cruentus* from 90.5% for V1 to 94.8% for V2, which confirms improvement by generating a more complete genome (**Table 1.1**). Generally, both *A. cruentus* genome versions have a higher BUSCO score, contain higher proportions of repetitive sequence and longer predicted protein and CDS sequences than *A. hypochondriacus* (**Table 1.1**). This indicates improved annotation and quality of the *A. cruentus* genome compared with *A. hypochondriacus*, due to higher genome completeness as well as improved assembly contiguity and annotation of repetitive elements (Ma *et al.*, 2021). A BUSCO score of 97.5% for the complete gene set of the *A. tricolor* genome, being higher than BUSCO scores of other *Amaranthus* genomes, indicates this new reference was very complete and therefore likely of excellent quality (**Table 1.1**) (Wang *et al.*, 2022).

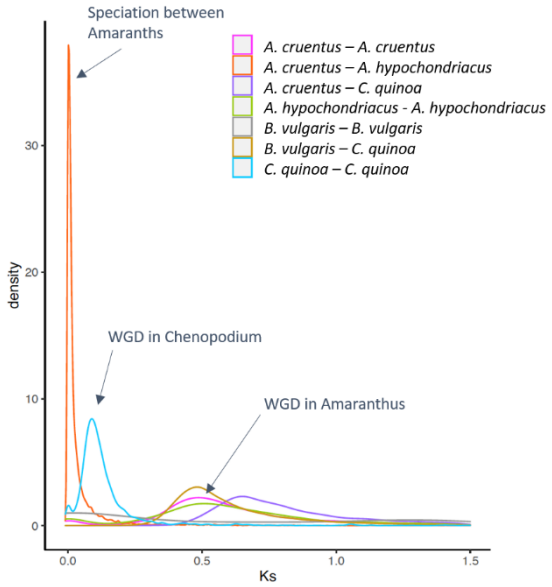



**Figure 1.2** Synonymous nucleotide substitutions per synonymous site (Ks) distributions for the paranome and orthologues of the *A. cruentus* genome and related species (Ma et al., 2021).

## Reference genome evolution and synteny

Occurrence of polyploidization events was determined by synonymous nucleotide substitutions per synonymous site ($K_s$) in duplicate gene pairs in the *A. hypochondriacus* and *A. cruentus* assemblies. Both *A. hypochondriacus* and *A. cruentus* species share the same whole genome duplication (WGD) estimated to be 18-34 million years ago (MYA)) (Ma *et al.*, 2021). This was first indicated by a $K_s$ distribution clearly peaking at $K_s$ = 0.55 in the draft genome of *A. hypochondriacus* (Clouse *et al.*, 2016), followed by $K_s$ = 0.53 in improved *A. hypochondriacus* (Lightfoot *et al.*, 2017) and another consistent result of $K_s$ = 0.54 in the anchor pairs (homologues) of *A. cruentus* and *A. hypochondriacus* (Ma *et al.*, 2021). This WGD took place before the divergence of the two species approximately 1.45 (0.61–2.41) MYA. This was revealed by a cross-species (*A. cruentus – A. hypochondriacus*) *Ks* comparison based on orthologues showing much more recent genome divergence (**Figure 1.2**) (Ma *et al.*, 2021). With the development of the *A. tricolor* genome the individual $K_s$ peaks for *A. tricolor* and *A. cruentus* were both approximately $K_s$ = 0.51 and the WGD was therefore estimated to be 26.56 MYA in both these amaranths (Wang *et al.*, 2022). The divergence between *A. tricolor* and *A. cruentus* took place 5.73 MYA, so before the divergence of *A. hypochondriacus* and *A. cruentus* (Wang *et al.*, 2022). Hence it is expected that the WGD took place in the last common ancestor of all three of these amaranth species and results indicate that it occurred in all three subgenera of the genus *Amaranthus* (Wang *et al.*, 2022).

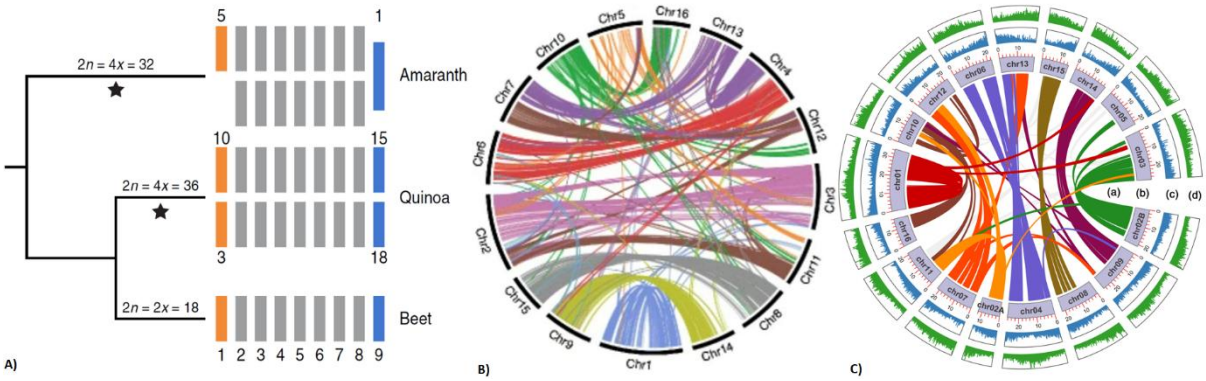



**Figure 1.3** Chromosome evolution models. **A)** Phylogenetic tree and chromosome numbers of related *Amaranthaceae* family species; *A. hypochondriacus*, quinoa, and beet. Haploid chromosomes in each species represented by orange, grey, and blue boxes, orthologous chromosomes between the species represented by orange and blue boxes only. Black stars represent lineage-specific whole genome duplications. (Lightfoot *et al.*, 2017). **B)** *A. hypochondriacus* (Lightfoot *et al.*, 2017) and **C)** (a) *A. cruentus* (Ma *et al.*, 2021) synteny blocks between their homoeologous chromosomes. (b) Pseudomolecules (with length in Mb), (c) gene density and (d) transposable element density based on the *A. cruentus* genome assembly landscape and annotation (Ma *et al.*, 2021).

The genus *Amaranthus*, quinoa (*Chenopodium quinoa* Willd.) and beet (*Beta vulgaris*) are all part of the *Amaranthaceae* family (Lightfoot *et al.*, 2017). It is believed that the common ancestor of *A. hypochondriacus*, quinoa and beet, had a haploid chromosome number of n=9 (**Figure 1.3 A**) (Lightfoot *et al.*, 2017). No recent WGD event took place in *B. vulgaris*, hence remaining a haploid chromosome number of n=9 (**Figure 1.3 A**) (Lightfoot *et al.*, 2017). However, the genus *Amaranthus* and quinoa have each undergone an independent WGD, with the later for quinoa lineages happening more recently than the WGD of Amaranths (**Figure 1.3 A**) (Lightfoot *et al.*, 2017). *C. quinoa* has retained its haploid chromosome number of n=18 (**Figure 1.3 A**), while *A. hypochondriacus* has a haploid chromosome number of n=16 due to the fusion of the two copies of chromosome 1 and loss of one copy of chromosome 5 (**Figure 1.3 B**) (Lightfoot *et al.*, 2017). Similar events happened for the *A. cruentus* genome (**Figure 1.3 C**) (Ma *et al.*, 2021). The difference between the two species is a fission of chromosome 2 of *A. hypochondriacus*, into two for the *A. cruentus* genome, creating a haploid genome of n=17 chromosomes (**Figure 1.4**) (Ma *et al.*, 2021).

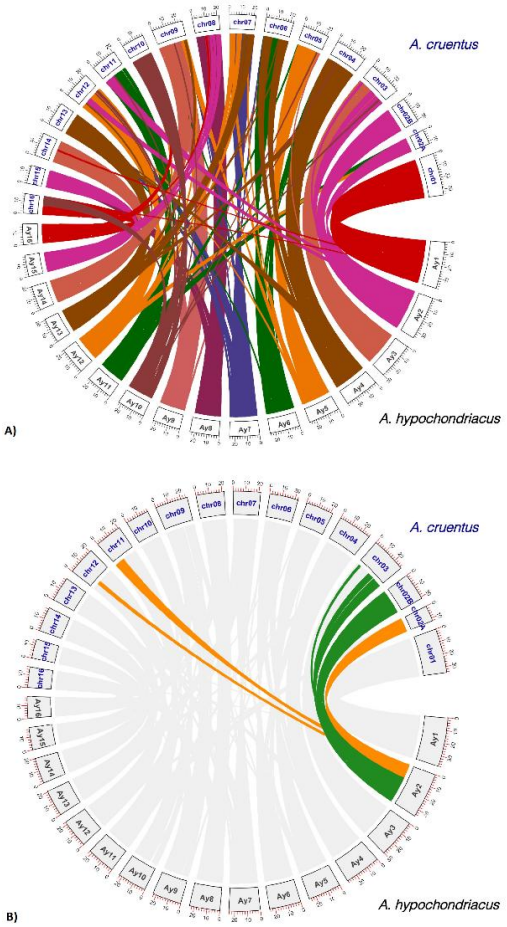



**Figure 1.4** Synteny blocks of relationship between *A. hypochondriacus* and *A. cruentus* genomes. **A)** Synteny with colouring based on chromosomes of *A. hypochondriacus*. **B)** Fission of chromosome 2 in *A. cruentus* with separate colouring of the two halves of *A. hypochondriacus* chromosome 2 (Ma et al., 2021).

The *A. tricolor* genome consists of n=17 haploid chromosomes and, as previously explained, likely underwent the same WGD event as other species in the *Amaranthus* lineage (Wang *et al.*, 2022). This was additionally identified by a 1:2 synteny relationship between *B. vulgaris* and *A. tricolor*, while for all other comparisons between amaranths (*A. tricolor* versus *A. cruentus*, *A. palmeri* or *A. hypochondriacus*) a 2:2 relationship was observed, suggesting the WGD occurred after the divergence of *Amaranthus* from *B. vulgaris*, (Wang *et al.*, 2022). However, after this shared WGD among amaranths many homologous chromosomes underwent fission and fusion events causing rearrangements resulting in significantly diverged genome structures (Wang *et al.*, 2022). Some examples of macro-synteny due to the WGD can still be observed in *A. tricolor*, like; Chr03:Chr08, Chr16:Chr17 and Chr06:Chr09, although, in general the *A. tricolor* genome underwent substantial chromosome rearrangements after the WGD (Wang *et al.*, 2022).

### Other *Amaranthus* genomic resources

Apart from existing reference genomes, other relevant *Amaranthus* genomic resources have been used in this thesis. A big contributor is the World Vegetable Center (WVC) gene bank which conserves approximately 1,000 amaranth accessions known for various agronomic traits (Lin *et al.*, 2022). The WVC genotyped more than 700 accessions of their collection with a GBS approach using Diversity Arrays Technology Pty Ltd. (DArT)Seq, which successfully clarified their taxonomy (Lin *et al.*, 2022). The WVC, also known as the Asian Vegetable Research and Development Center (AVRDC) (https://avrdc.org/) and United State Department of Agriculture (USDA) Germplasm Resources Information Network (GRIN) (https://www.ars-grin.gov/) gene banks are large providers of diverse *Amaranthus* accessions, with lines being obtained for a variety of relevant amaranth studies (Wu and Blair, 2017; Jamalluddin *et al.*, 2022; Hoshikawa *et al.*, 2023). Additionally, there are local African gene banks like the Agricultural Research Council (ARC) (https://www.arc.agric.za/arc-vopi/Pages/Plant%20Breeding/Genebank%20Collections.aspx) which conserve many other amaranth accessions used for this current and other research in South Africa (Gerrano, van Rensburg and Adebola, 2015; Mncwango *et al.*, 2021). Lastly, 121 accessions obtained from the USDA, belonging to the *hybridus* species complex, were sequenced with Illumina generating short-reads which were valuable resources to investigate (incomplete) domestication in literature and to conduct phylogenetic analysis in this thesis (Stetter, Vidal-Villarejo and Schmid, 2020).

## Amaranth phenotyping and metabolite detection

### Phenotyping of agronomic traits or nutritional metabolites

Phenotyping, like characterization of agronomic traits or metabolites in amaranths can be of relevance for general deeper understanding as well as genetic improvement of the crop. Relatively few traits have been characterized at the molecular or gene level in amaranths, and traits which are desirable for cultivated types can be undesirable in weedy amaranths (Riggins and Mumm, 2021). Up to now mostly grain yielding traits have been assessed while attempting to genetically improve amaranths, however, while selecting individuals for hybridization, vegetative features, growth and yield are considered (Anuradha *et al.*, 2023). Obviously, the environment can have an effect on many relevant traits such as plant height, days to maturity, and plant habit (Anuradha *et al.*, 2023). Some morphological descriptors or traits which are typically assessed at the WVC can be of seedling data regarding

germination, vegetative characteristics, inflorescence and seed data, this to note down the pigmentation, appearance and general growth habits of different varieties (Lin *et al.*, 2022). Another key trait which is commonly associated with the (incomplete) domestication of Amaranth is the shift from dark to white seed colour (Stetter, Vidal-Villarejo and Schmid, 2020; Riggins and Mumm, 2021). These are all examples of agronomically relevant traits, with phenotypes characterized measurable by appearance, while in current literature there is mostly a high emphasis on yield traits and grain amaranths.

Although yield is obviously a very relevant trait, also for vegetable amaranth, getting insight into the composition of leaves and their relevance for human nutrition is another important way of phenotyping. Some interesting studies on different amaranth species and their nutritional composition, with results obviously depending on growth conditions and individual variety, reveal the following findings. In leaves of *A. dubius* (*A. spp.*) a balanced essential amino acid composition was found with high levels of albumins, globulins, prolamins and glutelins, only the presence of essential leucine was limited (Rodriguez *et al.*, 2011). The amino acid, proline was in high quantities in cultivated *A. cruentus*, whereas, both proline and leucine were prominent in *A. hybridus*, and it was found that maltose and sucrose increased in both in cultivated compared to wild leafy varieties (Nkobole and Prinsloo, 2021). In *A. tricolor L.* leaves contain high levels of protein and carotenoids (vitamin A), but also considerable amounts of fibre and ascorbic acid (vitamin C) (Shukla et al., 2006). Generally amaranth contains higher levels of protein, calcium and iron than other leafy vegetables like spinach and cabbage (Ebert, Wu and Wang, 2011; Rastogi and Shukla, 2013). Amaranths contain among the highest reported concentrations of zinc and folate when compared to other forgotten African leafy crops, while sharing the position of highest iron concentration with spider plant (van Zonneveld et al., 2023). In contrast, vitamin A contents appear to be higher in forgotten leafy vegetables like spider plant, Arrowleaf elephant ear and Moringa, while vitamin C has highest concentrations in three forgotten fruits; guava, marula and carissa (van Zonneveld et al., 2023).

Ascorbic acid (vitamin C) is important as it is commonly known to be a main enhancer of iron absorption (Teucher, Olivares and Cori, 2004; Basrowi and Dilantika, 2021). Interestingly it was found that varying iron content in leaves is not always associated with higher iron bioavailability or impact on the body. When feeding a single line from three species, *A. tricolor* caused the highest haemoglobin gain in anaemic rats, despite having the lowest relative iron bioavailability compared to *A. hypochondriacus* and *A. cruentus* (Rangarajan *et al.*, 1998). Two out of four selected drought-tolerant leafy vegetable amaranth accessions had a high antioxidant content potentiality due to abundant phenols, vitamins, and flavonoids significantly positively correlated with antioxidant capacities (Sarker and Oba, 2020). Despite amaranth being known for its highly nutritious properties being rich in protein, calcium, iron and vitamins, antinutrients can be present, like oxalate in *A. cruentus* or nitrate and oxalate in *A. tricolor* (Ebert, Wu and Wang, 2011). These antinutrients, nitrate and oxalate, are found in leafy type amaranths and are known to inhibit the absorption of calcium and zinc which can cause kidney problems (Rastogi and Shukla, 2013). However, moderate consumption of antinutrients does not cause adverse nutritional effects and boiling the seeds or leaves in sufficient water removes the toxic components (Ebert, Wu and Wang, 2011; Rastogi and Shukla, 2013).

The previous studies are just an example of what (anti-)nutritional components can be present or absent in various amaranths. A common method of detecting these components is by metabolomics, which is a large-scale analysis of metabolites present in a cell, tissue or the entire organism. Metabolites are small molecules present in an organism at a certain point in time, which can be involved in cellular reactions (Dührkop *et al.*, 2015; Perez de Souza *et al.*, 2019). Plants can synthesize metabolites, and these small organic molecules can be categorized into two groups; primary and specialized (historically called secondary) metabolites (Lacchini and Goossens, 2020; Zhan *et al.*, 2022). Primary metabolites are typically found in all plant species, as they are involved in essential metabolic processes and therefore fundamental for homeostasis and maintenance of basic cellular functions (Lacchini and Goossens, 2020). Some examples are amino acids, nucleic acids and sugars (Zhan *et al.*, 2022). Specialized metabolites are recognized for having specialized functions in specific groups of cells, organisms, species or taxa. Hence, they are renamed as the historical name 'secondary metabolites' refers to previous beliefs of dispensable functions, which was later shown is not the case (Lacchini and Goossens, 2020). It is estimated that there are more than 200,000 of these unique natural products among plants, all due to the high diversity of secondary metabolism in the plant kingdom (Dixon and Strack, 2003; Fernie and Tohge, 2017). Understanding plant metabolomics, the participation of metabolites in plant biosynthetic pathways, and how those pathways can be elucidated, is highly relevant to explore the potential benefits of the existing high variety of natural products (Yang *et al.*, 2014; Sumner *et al.*, 2015; Li *et al.*, 2021; Zhan *et al.*, 2022). Moreover, it known that plant metabolic pathways can be of high relevance for creating specific nutrient components relevant for human nutrition, and that more research and innovations are needed in order to meet requirements for nutrient-enriched crops (Zhu *et al.*, 2020).

### Metabolomic analysis

Metabolome analysis emerged in science in the late 1990's (Perez de Souza *et al.*, 2017, 2019; Alseekh and Fernie, 2018). The term metabolome got coined in a review of Oliver et al. (1998), explaining that relative concentrations of metabolites can be a result of deletions or overexpression of a gene. The metabolome represents the metabolite profile of a cell or tissue, in which metabolites are the end products of biochemical pathways and therefore closest to being a phenotype (Bhattacharyya *et al.*, 2021). It is not only full gene deletions or overexpression which shape the metabolome, also single point mutations within a gene or other sequence changes can result in a different metabolite profile (Bhattacharyya *et al.*, 2021; Aharoni, Goodacre and Fernie, 2023). For a long time, an advantage of metabolome analysis was that it did not depend on the genome sequence like for example transcriptomics and proteomics technologies (Schauer and Fernie, 2006; Fernie and Schauer, 2009; Perez de Souza *et al.*, 2017). However, the powerful development of next-generation sequencing made sequences readily available and the earlier advantage of metabolomics as independent of genomics is now lost. At the same time, the lack of a direct link between genome and metabolome contributes to identification issues of metabolite phenotypes which are not shaped solely by the genome (Fernie and Schauer, 2009; Perez de Souza *et al.*, 2017, 2021; Alseekh *et al.*, 2021). The study of using multiple techniques for phenotyping, like combining two or more of genomics, transcriptomics, proteomics, epigenomics, phenomics or metabolomics, is called multi-omics, and can be useful for discovery and further exploration of biosynthetic pathways due to a more comprehensive understanding by integrating different levels of molecular information (Zhan *et al.*, 2022).

Metabolomic studies can be untargeted or targeted, the latter focussing on detecting known compounds expected to be in the sample (Perez de Souza *et al.*, 2019). First measurements of metabolites were usually highly targeted to specific metabolites, but with the development of spectrophotometric approaches a broader range of metabolites could be analysed without prior selection, which led to the emergence of untargeted metabolomics (Aharoni, Goodacre and Fernie, 2023). Untargeted metabolomic studies focus on all metabolites and are usually reliant on either nuclear magnetic resonance (NMR) spectroscopy or mass spectrometry (MS) for detection (Fernie *et al.*, 2011; Dührkop *et al.*, 2015; Alseekh and Fernie, 2018; Perez de Souza *et al.*, 2019). The following steps are typically performed during metabolite analysis: 1) sample preparation and metabolite extraction, 2) data acquisition by spectrometry analysis, 3) data processing, 4) data and statistical analysis, 5) metabolite annotation and further interpretation (Perez de Souza *et al.*, 2017, 2019). Regarding the first step, metabolite extraction, the extraction protocols are selective to the classes of metabolites of interest to reduce sample complexity, and those protocols should be as simple as possible to avoid errors (Perez de Souza *et al.*, 2019).

Metabolome analysis can be challenging because it is hard to link chemical compounds with data obtained with mass spectrometers and therefore a lot of metabolites remain unknown (da Silva, Dorrestein and Quinn, 2015; Dührkop *et al.*, 2015, 2019; Alseekh *et al.*, 2021; Perez de Souza *et al.*, 2021). On average only 1.8% of spectra of untargeted metabolomics can be annotated in biological samples, while for genome analysis the number of identified genes can even reach 80% (da Silva, Dorrestein and Quinn, 2015). Therefore, you could argue that we know less about the chemical composition of organisms compared to their genetics (da Silva, Dorrestein and Quinn, 2015). Usually, metabolomics data is generated in full-scan mode with first stage mass spectrometry (MS1), scanning through a broad *m/z* (mass-to-charge ratio) range with all ions obtained within this range being directly analysed (Perez de Souza *et al.*, 2021). Often the majority of detected ion features obtained from a source are non-biological contaminants or redundant components (Perez de Souza *et al.*, 2021). In addition to MS1, second stage or tandem mass spectrometry (MS2 or MS/MS) can be conducted, which is a strong analytical tool, using two mass analysers, adding a step to MS1 by selecting a precursor ion of interest to fragment and then analyse the resulting ions in the second mass analysation (MS2) (Dongré, Eng and Yates, 1997; Perez de Souza *et al.*, 2021; Yu, Dolios and Petrick, 2022).

Several metabolite databases are available for annotation of metabolites, like PhytoHub for dietary phytochemicals (*PhytoHub*, no date), KNApSAcK for plant metabolites (*KNApSAcK*, no date) and ChemSpider for broad chemical structures (*ChemSpider | Search and share chemistry*, no date) which are all based on MS1. GNPS (Global Natural Products Social Molecular Networking) and Sirius both rely on MS2 and are commonly used in untargeted metabolomics (Wang *et al.*, 2016; Dührkop *et al.*, 2019). GNPS is a community-wide continuously updated reference database for molecular networking and matching spectral libraries (Wang *et al.*, 2016), while Sirius tries to identify molecular structures achieving identification rates even more than 70% on challenging metabolomics datasets (Dührkop *et al.*, 2019). In addition to these existing databases, there are many suggestions in recent literature to potentially improve the typically low coverage of the metabolome, like enhanced MS techniques (Perez de Souza *et al.*, 2021), or updated extraction, quantification and annotation methods complementary to existing ones (Alseekh *et al.*, 2021). Even

artificial intelligence is suggested, and it is believed that with a better link between metabolome and genome data plant and microbial sciences can be key drivers in the development of metabolomics research (Aharoni, Goodacre and Fernie, 2023). Despite these promising developments and existing datasets, and the now already 25th anniversary of the term metabolome, mayor challenges remain, as the confidence and coverage for mass feature annotation stays low (Aharoni, Goodacre and Fernie, 2023).

### Metabolomics in Amaranths

Regardless of the difficulty of detection and annotation, many metabolomic studies have been conducted with various types of Amaranths, trying to identify nutritionally relevant compounds by performing spectrophotometric approaches. Of course, metabolites detected are dependent on the growth conditions and the individual accession, as was shown with untargeted NMR based feature detection revealing different metabolite profiles of red and green amaranths grown in an open field versus greenhouse in South Africa (Nemadodzi and Managa, 2024). Moreover, both cultivated and wild *A. cruentus* and *A. hybridus* were found to be rich and variable in macro- and micronutrients, such as increased maltose and sucrose identified in both cultivated varieties using NMR spectroscopy, and MS2 revealing rutin, 2-phenylethenamine and amaranthussaponin I present in both wild *A. cruentus* and *A. hybridus,* with chlorogenic acid only identified in cultivated *A. hybridus* (Nkobole and Prinsloo, 2021). Untargeted metabolite profiling using MS1 identified mainly polyphenols and lipids but also betalain, in *Amaranthus* sp. and beetroot (*B. vulgaris*) microgreens, with in vitro digestion strongly altering the phytochemical profile, increasing the total phenolic content and antioxidant capacity (Rocchetti *et al.*, 2020). In three edible vegetable *Amaranthus* species, *A. hybridus*, *A. blitum*, and *A. caudatus*, untargeted liquid chromatography MS2 revealed mostly amino acids and flavonoid glycosides, but also many other relevant metabolites with classes mostly known for their pharmacological activities (Abdel-Moez *et al.*, 2023). Additionally, targeted metabolite analysis using UV-visible spectrophotometry investigating flavonoids, cinnamic acid derivatives and benzoic acids in *A. cruentus* leaves, highlights the nutritional potential of this crop suggesting antidiabetic and antihypertensive properties (Araujo-León *et al.*, 2024).

These are only some very recent examples that targeted as well as untargeted metabolite detection in amaranth is a possible and a highly important way of phenotyping to reveal its nutritional relevance. However, exploration of metabolite abundance and linking it to genomic information started only recently in Amaranths (Sarker and Oba, 2020; Ma *et al.*, 2021; Winkler *et al.*, 2024). As previously mentioned, by combining various ways of phenotyping, i.e., by multi-omics techniques, genes responsible for metabolites or entire plant metabolic pathways can be revealed (Zhan *et al.*, 2022). Functional annotation of genes potentially associated with metabolites is typically done with three different methods; 1) directly testing candidate genes with (co-)expression patterns or knock-out mutants, 2) metabolite-based quantitative trait loci (mQTLs) mapping and 3) metabolite-based genome-wide association studies (mGWAS) (Luo, 2015; Alseekh and Fernie, 2018). Detecting mQTL by mapping genes determining abundance of specific metabolites is an older technique than mGWAS (Alseekh and Fernie, 2018). It is laborious, time-consuming and with limited usage of the available germplasm diversity in comparison to mGWAS due to only including variation captured of only two or more parental accessions (Luo, 2015). Hence, mGWAS has been performed in many recent plant studies, like for metabolites in seeds of Arabidopsis

(Naake *et al.*, 2024) and buckwheat (Zargar *et al.*, 2023), in fruit tissue of wolfberry (Zhao *et al.*, 2024) and jujube trees (Zhang *et al.*, 2022) and leaves of millet (Wei *et al.*, 2021), maize (Riedelsheimer *et al.*, 2012) and apples (Cao *et al.*, 2021).

No mGWAS has been performed in amaranths yet, and only limited studies investigating genes responsible for metabolites exist. Looking into metabolome-genome associations is scientifically very relevant. The mGWAS plant studies being examples revealing information regarding complex metabolites for economic traits (Zargar *et al.*, 2023; Naake *et al.*, 2024), nutritional and medicinal traits (Zhao *et al.*, 2024), domestication influencing edible traits (Zhang *et al.*, 2022), physiological plant development traits (Riedelsheimer *et al.*, 2012) and stress resistance traits for drought, cold and diseases (Cao *et al.*, 2021). Some other good multi-omics approaches were performed investigating the tomato metabolome. Using metabolomic combined with transcriptomic data in tomatoes during their growth cycle, clustering of metabolites and their co-expressed genes could be revealed based on their biological functions (Li *et al.*, 2020). Another extensive study incorporating genomes, transcriptomes, and metabolomes from hundreds of tomato varieties, performing mGWAS as well as expression QTL, revealed new metabolic genes, pathways and showed changes in the tomato fruit metabolome due to selection by breeding (Zhu *et al.*, 2018). Combining metabolomes with phylogenetic studies is another way to dig into correlations between metabolites and the genome. It helped to verify that stronger plant-enemy interactions lead to greater metabolomic variation in tropical than temperate forest plants in Panama and Maryland (Sedio *et al.*, 2018). Additionally, it highlighted chemo-evolutionary relationships in Australian *Eremophila*, which is useful for understanding complex interactions in biological systems across the continent (Gericke *et al.*, 2021).

Some of the recent but limited studies which do show insights into genes responsible for metabolites or metabolic pathways in amaranths are the following. A study in *A. palmeri* showed that in varieties which were known to have a higher EPSPS gene copy number, a phytochemical stress response can be induced resulting in disruption of the shikimate (phenylpropanoid) pathway (Sandhu *et al.*, 2023). The shikimate/phenylpropanoid pathway is an example of a known biosynthetic pathway in plants which produces various phenolic compounds like phenolic acids, flavonoids, stilbenes, and lignans which are involved in a wide range of physiological activities (Naikoo *et al.*, 2019, p. 9; Sharma *et al.*, 2019; Rocchetti *et al.*, 2020). It is known that EPSPS genes can upregulate the phenylpropanoid metabolism in plants and this study highlights the response of different *Amaranthus* varieties to abiotic stresses by looking at metabolite content only (Sandhu *et al.*, 2023). In the most recent version of the *A. hypochondriacus* reference genome, updated gene annotations identified multiple conserved genes of the phenylpropanoid and flavonoid biosynthetic pathway, with phenylpropanoid pathway genes being strongly expressed in most tissues (Winkler *et al.*, 2024).

Betalain production for pigmentation is likely the least recent and most commonly known metabolite finding in amaranths for which the responsible genes are known (Brockington *et al.*, 2011). Betalain is a relevant compound, since it is a natural food colorant and a betalain intermediate (*L*-DOPA) is used to treat Parkinson's disease (Wang *et al.*, 2022). *Amaranthus* and other genera within the *Amaranthaceae* family produce tyrosine-derived betalain pigments rather than the ubiquitous anthocyanins as colouring pigments (Brockington *et al.*, 2011; Ma *et al.*, 2021; Wang *et al.*, 2022; Winkler *et al.*, 2024). Betalain is produced by a

conserved biosynthetic gene cluster (BGC). BGC are closely located linked genes, which together form gene clusters, that encode for enzymes which are involved in the biosynthesis of the same secondary metabolites or biosynthetic pathways (Boycheva *et al.*, 2014). The *Amaranthus* betalain BGC consists of two orthologues; *DODAα1* and *CYP76AD1*, which are co-located on chromosome 16 of *A. hypochondriacus*, *A. cruentus* and *A. tricolor*, but also on different chromosomes in *B. vulgaris*, *C. quinoa*, *S. oleracea* (Ma *et al.*, 2021; Wang *et al.*, 2022). It is found that expression of *DODAα1*, *CYP76ADα1*, but also *MYB2* genes is profoundly lower in green *A. tricolor* leaf tissue compared to red or mixed colour tissue (Wang *et al.*, 2022). The *MYB2* gene is not co-located with *DODAα1* and *CYP76ADα1*, but a homologous transcription factor R2R3-MYB is known to control the expression of these two genes in *B. vulgaris* (Wang *et al.*, 2022). In the updated version of the *A. hypochondriacus* reference genome betalain and anthocyanin pathway genes were annotated, with strong *MYB2* gene expression observed in red flowers, while *MYB2*, *DODA* and *CYP76AD* genes were also expressed in root and stem tissue (Winkler *et al.*, 2024). These examples highlight how BGC, such as *DODAα1* and *CYP76AD1* genes responsible for betalain production, are crucial in specialized metabolite biosynthesis, suggesting broader investigation into BGC.

## Biosynthetic gene clusters (BGC) relevance and identification

### BGC relevance and evolvement

The betalain biosynthetic pathway is an example of a commonly known BGC, and knowledge of such gene clusters can be useful for varying purposes. Many BGC pathway derived metabolites can be beneficial for plants, by for example defending against herbivores or conferring pathogen and disease resistance (Zhan *et al.*, 2022). Additionally, metabolic pathways initiated by BGC in plants are obviously also highly relevant to human health. Some BGC derived metabolites are important nutrient components, are anti-inflammatory, anti-cancerous, could be food additives or are associated with other relevant agronomic, pharmaceutical, medicinal or therapeutic traits (Guo *et al.*, 2018; Zhu *et al.*, 2020; Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). More scientific investigation is therefore needed for identifying putative BGC responsible for those metabolites relevant to humans.

The first mentioning of a term for genes clumped together or a gene cluster, whose expression is co-ordinated was ’operon’ in prokaryotes (Jacob *et al.*, 1960). In plants BGC are often referred to as operon-like gene clusters (Zhan *et al.*, 2022). However, unlike true prokaryote operons, plant gene clusters are not transcribed as a single mRNA transcript but transcribed as individual genes. BGC are found in genomes of all plant species, although they are more commonly described for association with metabolic pathways in microbes than in plants (Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). The first discovery of clustered genes in plants responsible for the production of secondary metabolites was in 1997 (Frey *et al.*, 1997). Since then more than 30 metabolic gene clusters have been discovered across the plant kingdom, present in distant phylogenetic clades (Zhan *et al.*, 2022). A potential theory for genes clustering together in plants could be avoiding the accumulation of toxic intermediates (Nützmann, Huang and Osbourn, 2016). A rise of plant phytotoxicity of intermediates can be a potential cause for negative selection against partial pathway inheritance (Zhang and Peters, 2020). This is hence an explanation for the first hypothesis regarding BGC formation in plants which is the co-inheritance hypothesis. By maximalising inheritance of a whole intact pathway, cluster function is optimized (Smit and Lichman,

2022). Another hypothesis for cluster formation and functioning is co-regulation. Clustering could enhance the functional advantages of genes in a pathway, since functional metabolic pathways typically share expression patterns among different tissues which could be induced by similar triggers (Smit and Lichman, 2022).

### Identification and types of BGC

To identify and characterise genes in BGC, reverse genetic (from gene to phenotype) and forward genetic (from phenotype to gene) strategies can be used, the later most powerful to identify BGC responsible for secondary metabolic pathways (Zhan *et al.*, 2022). Examples of such forward genetic analysis include mGWAS and mQTL mapping, aiming to identify genes responsible for individual metabolite production. Apart from mGWAS and mQTL mapping, a variety of bioinformatics computational algorithms or other multi-omics techniques, but most importantly genome mining can be applied to identify BGC (Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). Genome mining starts with exploiting the genome sequences directly to search for presence of BGC. A variety of gene cluster prediction tools available to predict plant BGC directly from genomic data are PhytoClust, C-Hunter and plantiSMASH and many more (Chavali and Rhee, 2018). A good example of BGC detection using the plantiSMASH genome mining algorithm was the discovery of 15 clustered genes responsible for morphinan benzylisoquinoline alkaloid metabolism in opium poppy (Guo *et al.*, 2018). Additionally the co-expression of these 15 genes was revealed across various tissue types through RNA sequencing data (Guo *et al.*, 2018). This is a common way to provide more evidence that the detected BGC could be real, which was described for the betalain BGC in amaranths (Wang *et al.*, 2022; Winkler *et al.*, 2024). Another way to give more evidence is through the conservation of BGC between species, since BGC are expected to be co-inherited (Smit and Lichman, 2022). This was implemented after detecting BGC in the *A. cruentus* reference genome with plantiSMASH, by showing the conversation of the betalain as well as a different BGC among five *Amaranthaceae* species by synteny (Ma *et al.*, 2021).

Co-inherited BGC in the *A. cruentus* genome included three different enzyme domains: epimerase, terpene synthase and cytochrome P450 (Ma *et al.*, 2021). Knowledge of BGC can be useful in identifying metabolic enzymes in the same pathway and their causal function. Plant gene clusters typically contain various tailoring enzymes which modify the core structure of the metabolite produced in the first catalytic step. These enzymes are CYP450s, acyl transferases, UDP-glycosyl transferases (UGTs), short-chain alcohol dehydrogenases, transaminases, and decarboxylases (Bharadwaj *et al.*, 2021). Genes of the first pathway step are followed by downstream tailoring enzymes which can range in numbers from 3 to 12 depending on the complexity of different metabolic pathways (Bharadwaj *et al.*, 2021). Depending on location and distance of the genes and downstream tailoring enzymes in the genome, plant BGC can be compact, loose, or super gene clusters, (**Figure 1.5**) (Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). Compact gene clusters show gene order collinearity roughly with the biosynthetic pathway reactions (**Figure 1.5 A**), loose gene clusters are closely adjacent core gene cluster components with distantly distributed pathway enzymes for initiation, modification or regulation, indicating a fragmented pathway (**Figure 1.5 B**), while super gene clusters are formed by combination of several different metabolic gene clusters coming together as hotspots in the genome (**Figure 1.5 C**) (Zhan *et al.*, 2022). Based on present research, it currently is hard to draw conclusions related to general organisation of plant gene clusters and their behaviour (Bharadwaj *et al.*, 2021).

What recent studies have shown is that gene clusters have evolved through obtaining new functions or duplications of primary metabolic pathway genes, and that BGC are usually well regulated at the nucleosome level (Bharadwaj *et al.*, 2021).

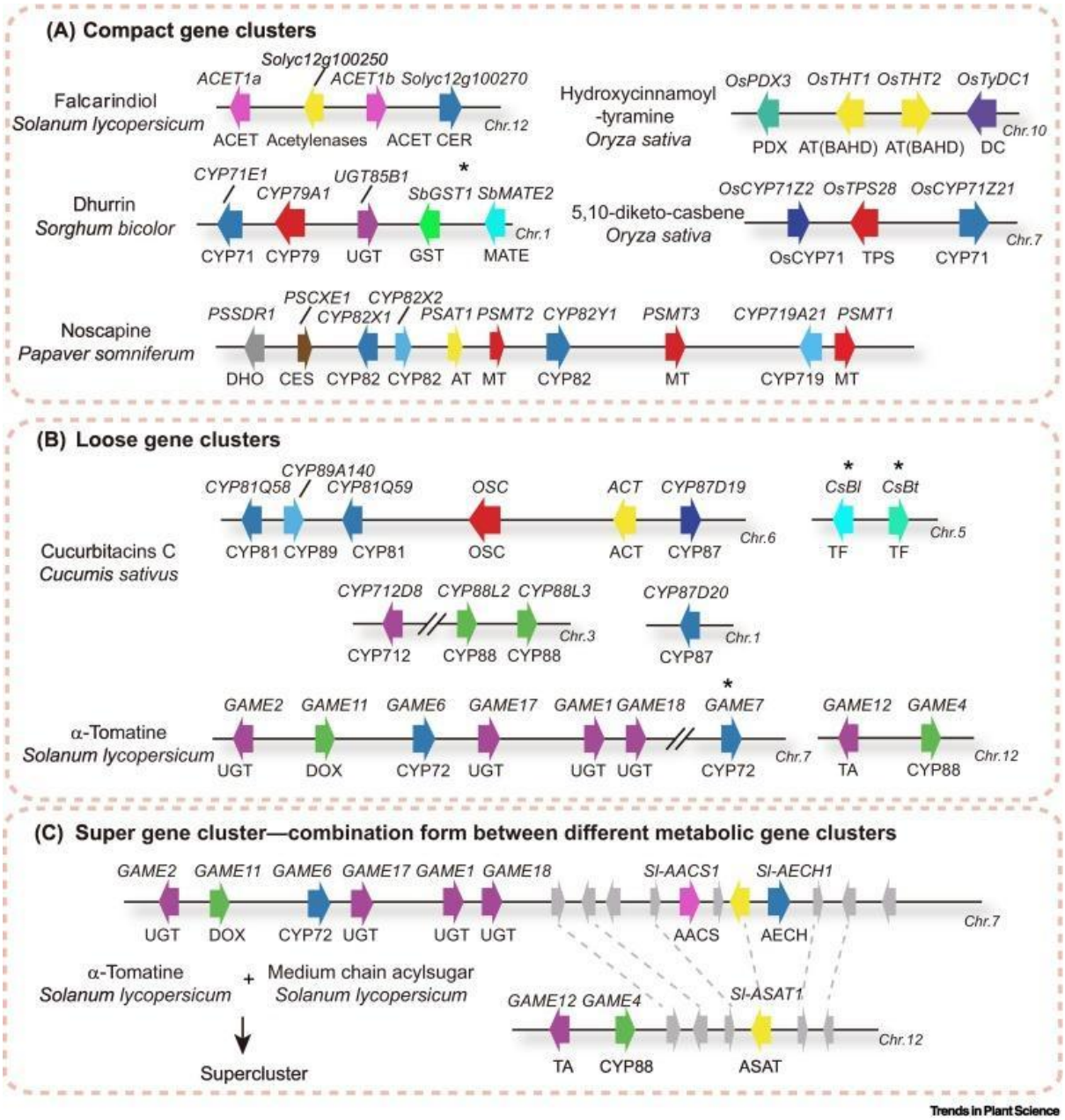



**Figure 1.5** Three types of BGC organisation involved in secondary metabolite biosynthetic pathways. **A)** Compact gene clusters with gene order roughly colinearly with pathway reactions. **B)** Loose gene clusters with closely adjacent core gene cluster components with distributed pathway enzymes which are indicative of a fragmented metabolic pathway. **C)** Super gene clusters which are a combination of different metabolic gene clusters (Zhan *et al*., 2022).

## Metabolite classifications, functions and BGC in amaranths

Specialized plant metabolites are often assigned to classes, which are classified by their chemical backbone and functional groups (Bharadwaj *et al.*, 2021). Some previously mentioned classes of metabolites detected in amaranths include polyphenols, lipids, flavonoid glycosides, flavonoids, cinnamic acid derivatives and benzoic acids (Rocchetti *et al.*, 2020; Abdel-Moez *et al.*, 2023; Araujo-León *et al.*, 2024). Biosynthetic pathways on the

other hand are elucidated by the identification and characterization of biosynthetic pathways genes, regulatory elements, and their gene products (Bharadwaj *et al.*, 2021). Apart from 'alkaloids' in the opium poppy BGC example (Guo *et al.*, 2018), other important metabolite classes which are known to be synthesized by plant BGC are oxylipins, terpenoids, phenylpropanoids, benzoxazinoids, cyanogenic glycosides and others like polyketides (Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). Terpenoids are the biggest and most structurally diverse metabolite class with more than half of the known BGC associated with it (Zhan *et al.*, 2022). Some examples of terpenoids are large and structurally diverse phenylpropanoid compounds or phenolamines and hydroxycinnamoyl-tyramine, the latter both shown to enhance disease resistance (Zhan *et al.*, 2022). Glucosinolates are also mentioned as a group of molecules derived from a plant specialised biosynthetic pathways (Smit and Lichman, 2022). However, cyanogenic glycosides are stated as possible precursors of glucosinolate metabolism (Barco and Clay, 2019), hence they likely are associated to the same class. It is known that cyanogenic glucoside biosynthesis occurs through BGC in some plants, which derives compounds beneficial for herbivore and other defensive properties (Takos *et al.*, 2011; Gleadow and Møller, 2014).

Like disease resistant terpenoids and defence enhancing cyanogenic glycosides, each of the metabolite classes identified in biosynthetic pathways derived by BGC have their own functions and relevance. Often agronomic as well as therapeutic traits are associated with metabolites produced by BGC, and more in-depth studies for improving bioproduction of target metabolites is needed (Bharadwaj *et al.*, 2021). Examples are pharmacologically active alkaloids, anti-inflammatory oxylipins and pathogen defensive benzoxazinoids and cyanogenic glycosides (Zhan *et al.*, 2022). The BGC in opium poppy being a perfect example producing pharmacologically active alkaloids since it is essential for morphinan biosynthesis (Guo *et al.*, 2018). During an attempt to identify BGC in amaranths, 10 shared clusters could be found between *A. hypochondriacus* and *A. cruentus* likely synthesizing lignans, terpenes, alkaloids, polyketides as well as BGC with UGT genes for glycosylation and saccharide production (Ma *et al.*, 2021). Six of these ten BGC showed conserved co-location of biosynthetic enzymes across *A. cruentus* and *A. hypochondriacus*, but also *B. vulgaris*, *C. quinoa* and *S. oleracea* genomes, providing additional proof that they are likely real BGC (Ma *et al.*, 2021). The main aim of BGC detection in amaranths was due to the relevance of secondary metabolites to the nutritional content of this underutilized crop (Ma *et al.*, 2021). Generation of novel genomic as well as metabolomic datasets is needed to largely enhance knowledge regarding amaranth BGC and provide insights in beneficial nutrients present in this leafy vegetable and its relevance to human diets.

## 1.3 Genomic variation and its usage for trait mapping

### DNA sequencing techniques and genomic variation responsible for traits

#### SNP detection by next generation sequencing

Next generation sequencing (NGS) was a revolutionizing technique in genomics research which enabled the human genome to be sequenced within a single day and deliver accurate genomic data to obtain insight into DNA variation (Behjati and Tarpey, 2013). With the ability to generate whole genome sequences by NGS approaches, single nucleotide polymorphisms (SNPs) could be detected (Nielsen *et al.*, 2011). A SNP is an individual

nucleotide base difference between two DNA sequences, and hence is an observation of a mutation evolved between two individuals over time (Morin *et al.*, 2004; Edwards *et al.*, 2007). SNPs are identified by aligning sequencing reads to a reference genome and calling variants at positions where nucleotides differ among individuals (Nielsen *et al.*, 2011). SNPs are the smallest units which can possibly be inherited and have a low mutation rate, hence SNPs are an excellent form of molecular genetic markers which can be used for mapping traits associations in genomic data (Edwards *et al.*, 2007). Genotyping with SNPs is generally cost-effective, time-saving, has a genome wide coverage and high resolution, all positively contributing to mapping traits (Kumar *et al.*, 2021). Like identifying metabolite-genome associations, quantitative trait loci (QTLs) and genome-wide association studies (GWAS) are used for mapping various phenotypic traits, with SNPs being a common genetic variant used to perform these techniques (Korte and Farlow, 2013). There are various good plant examples available of trait-genome associations identified with SNPs, such as in rice where GWAS with SNPs revealed associations for 14 agronomic traits (Huang *et al.*, 2010), or resistance for bacterial blight being mapped initially with GWAS and later further confirmed by QTL (Kim and Reinke, 2019).

### Larger genomic structural variants (SVs)

Despite the success of SNP based trait mapping, often follow up studies are suggested using structural variants (SVs) to further reveal functional genomic variation (Huang *et al.*, 2010). SNPs and small insertions or deletions (indels) are the easiest for fast genotyping in mapping populations, and hence are effective at identifying genomic regions of interest, however, in the search for candidate genes, larger SVs need to be taken into account (Schreiber *et al.*, 2024). Due to the limitations of SNPs covering only a small fraction of the genome, a shift from SNPs towards bigger SVs occurred within the field of crop genomics and improvement (Della Coletta *et al.*, 2021). SVs used to be largely unexplored in plant genomes (Zhou *et al.*, 2019), although, with more crop genomes being sequenced more exploration of SVs and their links towards crop evolution will emerge (Schreiber *et al.*, 2024). SVs, just like SNPs is DNA differing between individuals responsible for heritable phenotypic diversity (Yuan *et al.*, 2021). SVs are larger in size with a greater total number of nucleotide changes and hence less numerous than SNPs (De Coster, Weissensteiner and Sedlazeck, 2021; Della Coletta *et al.*, 2021). SVs are usually at least 50 bp long, while indels are shorter insertions or deletions of less than 50 bp (Zhao *et al.*, 2016; Jayakodi *et al.*, 2021; Yildiz *et al.*, 2023). SVs are large enough that they can cover entire genes present or lacking in certain accessions (Schreiber *et al.*, 2024). SVs can lead to gene loss, gene duplication and novel gene generation, and can therefore affect phenotypic variation in plants (Zhou *et al.*, 2019; Yuan *et al.*, 2021). A good example of how identified SVs can contribute to crop breeding and improvement is evidence of SVs changing gene dosages and expression modifying fruit flavour, size, and production of tomatoes (Alonge *et al.*, 2020).

SVs can be distinguished in several categories: deletions and insertions (presence-absence variation (PAVs)), duplications (copy number variations (CNVs)), inversions (changed orientation) and translocations (different chromosomal location), the latter two both being chromosomal rearrangements (**Figure 1.6**) (Escaramís, Docampo and Rabionet, 2015; Tattini, D'Aurizio and Magi, 2015; Zhao *et al.*, 2016; Sedlazeck *et al.*, 2018; Della Coletta *et al.*, 2021; Yuan *et al.*, 2021; Yildiz *et al.*, 2023). Deletions are loss of DNA while insertions are new sequences (Escaramís, Docampo and Rabionet, 2015; Tattini, D'Aurizio and Magi,

2015). Both deletions and insertions are typically the most abundant SVs in plants and crops (Yildiz *et al.*, 2023). CNVs are duplications of DNA sequence which can be inserted contiguously (in tandem) or elsewhere in the genome (segmental) (Escaramís, Docampo and Rabionet, 2015; Tattini, D'Aurizio and Magi, 2015). CNVs means that a DNA segment is present in a variable number of copies compared to a reference genome or different individual (Escaramís, Docampo and Rabionet, 2015; Tattini, D'Aurizio and Magi, 2015). CNVs are a relevant type of SV, since they are vastly present in plants and have influence on crop phenotypes and adaptability (Silaiyiman *et al.*, 2025). Examples include additional copies of the *KLUH* CNV in tomato, which result in larger fruit sizes (Alonge *et al.*, 2020), or duplication of a CNV in maize contributing to antiherbivore defence and resistance (Yuan *et al.*, 2024). Deletions, insertions and duplications (CNVs) are all examples of unbalanced SVs, which result in gain or loss of genetic material, while balanced SVs like inversions and translocations remain the same amount of genome (Escaramís, Docampo and Rabionet, 2015; Jayakodi *et al.*, 2021). Inversions are segments of DNA that are reversed in orientation from the rest of the chromosome, which can be pericentric or paracentric, i.e., located in the centromere or elsewhere, respectively (Escaramís, Docampo and Rabionet, 2015). Translocations are DNA segments which change its position, which can be inter- and intrachromosomally (Escaramís, Docampo and Rabionet, 2015; Tattini, D'Aurizio and Magi, 2015). SVs can be chained or nested together and hence occur into larger regions of variation (Sedlazeck *et al.*, 2018).

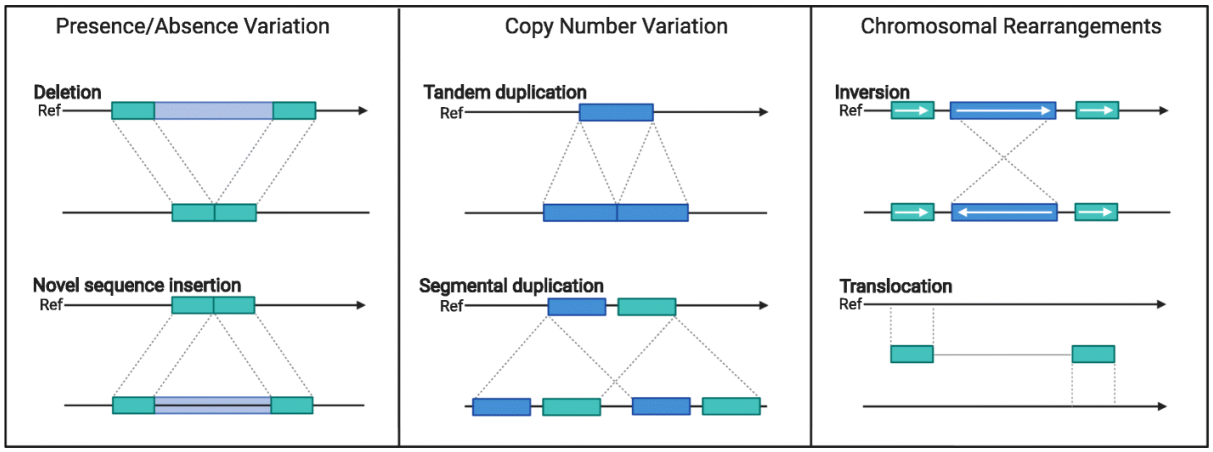



**Figure 1.6** Types of structural variations (SVs). First two boxes presenting unbalanced SVs (deletion, insertion and duplication), third box balanced SVs (inversion and translocation) (Della Coletta et al., 2021).

There are many ways or mechanisms that the cause SVs to be generated among individuals. For example, transposable elements (TEs) which replicate themselves in the genome and can capture and carry gene sequences to new locations in the genome, potentially causing significant disruption of coding regions (Della Coletta *et al.*, 2021). TEs are very prevalent in plant genomes and underlie many SVs, like in the tomato SV study the majority of the 200,000 detected SVs were TEs and related repeated sequences (Alonge *et al.*, 2020). Although TEs were largely ignored in crop improvement in the past, they are highly relevant for agronomic phenotypic variation, creating the majority of insertions and deletions in many crops (Della Coletta *et al.*, 2021). It is even shown that SVs caused by TEs insertions in non-coding regions like promotors or upstream to gene sequence, can influence expression of nearby genes affecting traits like flowering time in maize (Yang *et al.*, 2013; Huang *et al.*, 2018). Another common way SVs are introduced, especially in plants, is the rise of PAVs by differential genome fractionation across individuals following whole-

genome duplication, particularly in polyploids formed through hybridisation (Della Coletta *et al.*, 2021). SVs can also be created by errors during meiotic recombination, such as unequal or non-allelic homologous recombination when recombination occurs between similar sequences but at different genomic locations (Della Coletta *et al.*, 2021). In some highly heterozygous plants, like grapevines, it is shown that SVs increase due to clonal propagation and accumulation of recessive somatic mutations, while negative selection seems to act against SVs, particularly against chromosomal rearrangements (Zhou *et al.*, 2019).

### Techniques for variant detection

In the past limited techniques like microscopic observation, array comparative genomic hybridization (array-CGH) and SNP arrays were used to identify SVs (Yuan *et al.*, 2021). Array-CGH identifies CNVs by comparing sample DNA to a reference using fluorescent fragments, detecting gains or losses of genomic regions, while SNP arrays genotype known SNPs and can infer SVs based on the observed pattern of polymorphisms. Array-CGH remains a robust technique for efficient CNVs detection at multiple genomic loci, however array-CGH cannot detect balanced SVs (Escaramís, Docampo and Rabionet, 2015; Yuan *et al.*, 2021). SNP arrays are helpful to detect large-scale or allele-specific CNVs, although, they have a poor signal-to-noise ratio due to the smaller target size (Yuan *et al.*, 2021). SNP arrays as well as array-CGH both cannot detect insertions and have difficulties recognizing breakpoints of SVs (Yuan *et al.*, 2021). Nowadays, with NGS enabling rapid DNA sequencing and whole genome analysis, SVs detection at nucleotide level is more efficient, including detection of inversions, translocations and recombinant breakpoints of SVs (Behjati and Tarpey, 2013; Yuan *et al.*, 2021). Many recent advances in sequence technologies have drastically improved plant genome contiguity, so how continuous or complete a genome is. Some of these techniques which emerged in the past 20 years are; Sanger and Illumina sequencing for short-reads having a high quality per base, and PacBio and NanoPore sequencing for long-reads (can exceed 1 Mb) which originally had a much lower per base quality (Michael and VanBuren, 2020; Jayakodi *et al.*, 2021). Although, in the last 5 years, long-read sequencing has drastically improved regarding genome contiguity being more detailed with higher resolutions but also faster with higher throughput (Sedlazeck *et al.*, 2018; Michael and VanBuren, 2020).

Both short as well as long-read sequencing techniques can be used for SV detection. Some methods to detect SVs using short-reads are; 1) mapping reads to a reference by read pair (RP) (paired end mapping), split read (SR) mapping or read depth (RD), 2) direct comparison with *de novo* assembly or lastly 3) combination of these strategies (Alkan, Coe and Eichler, 2011; Tattini, D'Aurizio and Magi, 2015; Fuentes *et al.*, 2019; Della Coletta *et al.*, 2021; Yuan *et al.*, 2021). Combining multiple methods, approach 3), can improve the accuracy of SV discovery and using a high quality *de novo* assembly is recommended (Alkan, Coe and Eichler, 2011). Assemblytics is a common tool for detecting variant of *de novo* assemblies aligned to a reference genome (Nattestad and Schatz, 2016). Despite the ability to detect SVs with short-read sequencing it is commonly known that genomic regions can be missed, which limits the accuracy of SV detection, resulting in many false positives (Tattini, D'Aurizio and Magi, 2015; Sedlazeck *et al.*, 2018; Michael and VanBuren, 2020; Jayakodi *et al.*, 2021; Yuan *et al.*, 2021). It is hard to identify full SVs with short-read, because SVs can be large and go across the break points of short-read sequences (Chen *et al.*, 2014). Hence, the best way of detecting SVs while mapping to a reference genome

remains long-read (>10 kb) alignment, since those cover more than the typically few hundred base pair (~150 bp) length short-reads (Jayakodi *et al.*, 2021).

It is therefore commonly recognized that long-read is superior to short-read sequencing or other methods like arrays, providing better opportunities for comprehensive SV identification in crop plants (De Coster, Weissensteiner and Sedlazeck, 2021; Yildiz *et al.*, 2023). Some studies verifying this is long-read SV calls being used as a “gold standard” in grapevines, while short-read sequencing only detected short insertions (Zhou *et al.*, 2019). In humans long-reads also exceeded the ability of short-reads to detect small variants and SVs (Wenger *et al.*, 2019). While in tomato it was additionally confirmed that the majority of detected SVs would not have been revealed solely using short-read data (Alonge *et al.*, 2020). PacBio HiFi long-read sequencing is improving rapidly with increasing sequencing output and decreasing errors, this reduces assembly costs and makes long-read assemblies more attractive for detecting SVs, especially in complex polyploid plants (Della Coletta *et al.*, 2021). All three examples which showed long-reads to be superior to short-reads for SV detection, generated long-reads with PacBio sequencing (Wenger *et al.*, 2019; Zhou *et al.*, 2019; Alonge *et al.*, 2020).

Apart from the type of sequencing, the genome coverage after alignment, i.e., the number of sequenced reads mapped to a position of the reference genome, should be considered since coverage is an insightful indicator of irregularities within sequencing (Lindner *et al.*, 2013). Sequence quality can highly influence the accuracy of variant calling and thus a uniform coverage is preferred (Sims *et al.*, 2014). With the current advanced technological innovations like PacBio HiFi long-read sequencing as well as genomes increasingly being sequenced, an extensive breath of traits affected by SVs is expected to be characterized (Della Coletta *et al.*, 2021; Schreiber *et al.*, 2024). This is highly important in plant science and crop improvement due to potentially finding candidate genes, although, SNP detection definitely remains a quick, easy and cost-effective way to effectively identifying genomic regions of interest underlaying agronomic traits (Kumar *et al.*, 2021; Schreiber *et al.*, 2024).

## Pan-genomes in plant breeding

### Emerge of first reference and pan-genomes

Traditionally genomic variation, like SNPs and SVs, is detected by aligning lower quality sequences against a single high quality reference genome which is derived from one individual (Golicz, Batley and Edwards, 2016; Tao *et al.*, 2019; Bayer *et al.*, 2020; Sherman and Salzberg, 2020; Jayakodi *et al.*, 2021; Shi *et al.*, 2023; Schreiber *et al.*, 2024). The first creation of such a single reference genome in plants was of the flowering model plant *Arabidopsis thaliana*, which was fully sequenced and published in the year 2000 (The Arabidopsis Genome Initiative, 2000). The first publication of a crop genome sequence was of rice (*Oryza sativa* L.) in 2005 (Sasaki, 2005). Nowadays with the emerge of pan-genomes (**Figure 1.7**) the opportunity is created to detect genomic variation present in high quality genomes among multiple related individuals (Golicz, Batley and Edwards, 2016; Tao *et al.*, 2019; Bayer *et al.*, 2020; Sherman and Salzberg, 2020; Della Coletta *et al.*, 2021; Jayakodi *et al.*, 2021; Shi *et al.*, 2023; Wang *et al.*, 2023; Matthews *et al.*, 2024).

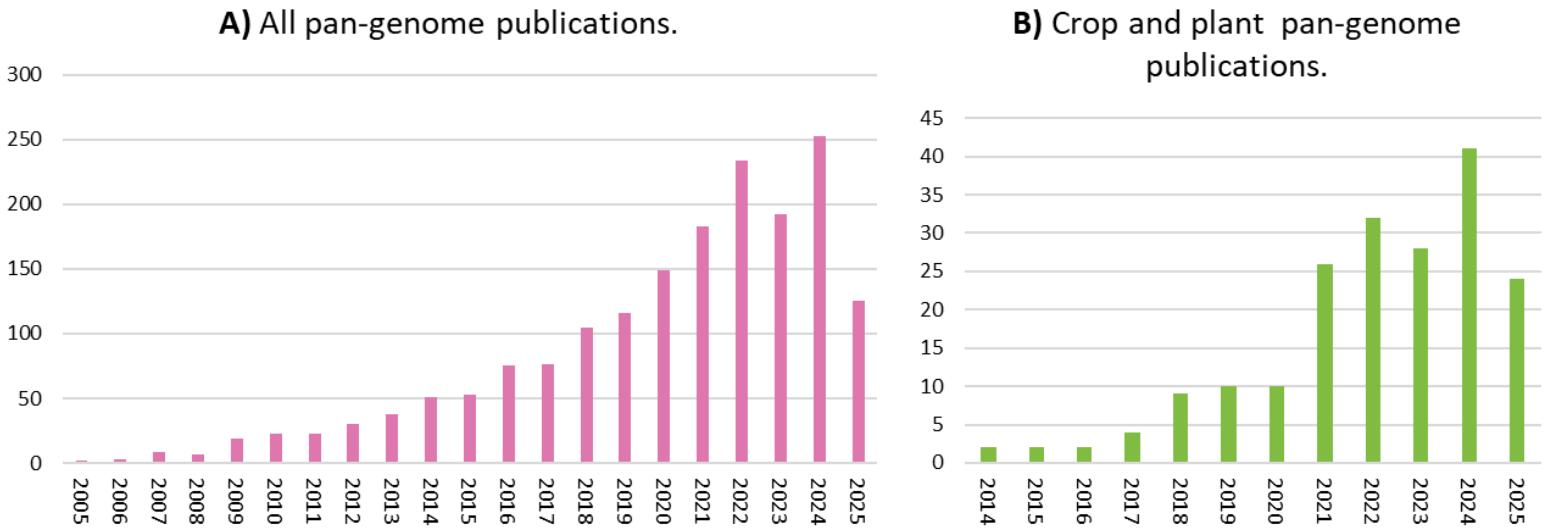



**Figure 1.7** Increasing numbers of pan-genomics related publications from **A)** the first pan-genome in 2005 and **B)** first plant pan-genome in 2014 onwards. Terms searched for on 12-06-2025 are **A)** 'pan-genome', 'pangenome' and 'pan genome' and **B)** 'pan-genome', 'pangenome', 'pan genome', 'crop' and 'plant' in NCBI Pubmed (https://pubmed.ncbi.nlm.nih.gov/).

The first pan-genome was created in *Streptococcus agalactiae* bacteria, with 'pan' coming from Greek meaning 'whole' (Tettelin *et al.*, 2005). The idea of a pan-genome is that with inclusion of multiple high quality genomes obtained from various individuals into a combined consensus a more complete reference resource can be created. In the first pan-genome eight *S. agalactiae* genomes were incorporated. It was built-up by two components; 1) a core genome containing genes present in all individuals and 2) a dispensable genome composed of genes present or absent in some or unique to individuals (Tettelin *et al.*, 2005). The majority of genes (~80%) present in this first pan-genome were core and those core genes were responsible for housekeeping, regulatory, transport and building functions (Tettelin *et al.*, 2005). This separation into core and variable genome components is still commonly performed, with core genes presumably having critical functions and therefore being conserved in a species, while dispensable genes contribute to species diversity, hence present in some but not all individuals (Golicz, Batley and Edwards, 2016; Tao *et al.*, 2019; Sherman and Salzberg, 2020; Della Coletta *et al.*, 2021; Jayakodi *et al.*, 2021; Hameed *et al.*, 2022; Matthews *et al.*, 2024).

The first ever created plant pan-genome was of seven wild soybean varieties (*Glycine soja*), in which 80% of the pan-genome was core, so shared among all seven accessions, and the rest was dispensable likely reflecting a role in adaptation to diverse environments (Li *et al.*, 2014). From an evolutionary perspective it is expected that dispensable genes typically evolve faster than core genes to enable species to adapt to different environments (Tao *et al.*, 2019). Collective evidence for this was found in the first plant pan-genome (Li *et al.*, 2014) and other early developed crop pan-genomes like rice (Wang *et al.*, 2018) and rapeseed (*Brassica napus*) (Hurgobin *et al.*, 2018), showing a more variable dispensable genome with higher SNP and indel densities, higher non-synonymous to synonymous mutation ratios, less functionally conserved genes and more environmental and defence functionally enriched genes (Tao *et al.*, 2019). Since their first creation there has been a rapid growth of plant pan-genome studies and this trend expected to continue (Bayer *et al.*, 2020). The number of publications has mostly been increasing since the first bacterial pan-genome (**Figure 1.7 A**) and the first plant pan-genome of soybean (**Figure 1.7 B**).

## Purpose of pan-genomes

As previously mentioned the advantage of pan-genomes over a single reference genome is a more complete landscape of genetic diversity within a species, making pan-genomes more useful to detect SNPs and SVs present among related individuals, typically within a species (Golicz, Batley and Edwards, 2016; Tao *et al.*, 2019; Bayer *et al.*, 2020; Sherman and Salzberg, 2020; Della Coletta *et al.*, 2021; Jayakodi *et al.*, 2021; Shi *et al.*, 2023; Wang *et al.*, 2023; Matthews *et al.*, 2024). Therefore, while constructing a pan-genome, the goal is to capture broad genetic diversity from the population with a limited number of individuals, ideally containing domesticated as well as wild genotypes (Jayakodi *et al.*, 2021; Matthews *et al.*, 2024). There are many plant pan-genomes which aim for capturing such a broad diversity. A most extensive example is the *Solanum* pan-genome capturing 22 wild, consumed, ornamental and domesticated species of the *Solanum* genus which includes a large variety of major crops like eggplant (*S. melongena*), potato (*S. tuberosum*) and tomato (*S. lycopersicum*) (Benoit *et al.*, 2025). The tea (*Camellia sinensis*) pan-genome of 22 elite cultivars obtained from a population cultivated in eight major tea-growing countries in which population structure analysis showed that the accessions can be classified into three groups (Chen *et al.*, 2023). A pearl millet (*Pennisetum glaucum* (L.) R. Br., syn. *Cenchrus americanus* (L.) Morrone) pan-genome used 11 wild as well as cultivated accessions originating from geographically varying regions around the world (Yan *et al.*, 2023). Generally, pan-genomes can be divided into open or closed. In open pan-genomes each newly added individual contributes with new genes and hence incorporates diversity, while in closed (restricted) pan-genomes adding new individuals does not result in pan-genome expansion due to newly added genes already present (Golicz, Batley and Edwards, 2016; Azimuddin and Aslam, 2023; Matthews *et al.*, 2024).

Often the goal of creating pan-genomes for crops is to identify genomic variation that can be linked to phenotypes potentially affecting agricultural traits (Sherman and Salzberg, 2020; Hameed *et al.*, 2022). Domesticated crops have been selected and bred by humans overtime but the genes or variants underlaying preferred agronomic phenotypes often remain unknown (Sherman and Salzberg, 2020). Comparing the genomes of wild with cultivated species in a pan-genome, can reveal the impact of domestication, and identify lost or desirable traits for which genetics responsible can be reintroduced into accessions by breeding (Matthews *et al.*, 2024). The first ever created plant pan-genome of soybean (*G. soja*), also aided the identification of genes which could be linked to agronomic traits like biotic resistance, seed composition, flowering and maturity time, organ size and biomass (Li *et al.*, 2014). The pearl millet pan-genome was used to identify structural variation associated with heat tolerance (Yan *et al.*, 2023). The selection of individuals in a crop pan-genome can also include founder accessions of an experimental population used for breeding purposes (Jayakodi *et al.*, 2021). This method can be highly relevant while creating a mapping population, with the genomic variation captured in a pan-genome potentially being responsible for various traits of interest, such as in the more recent pan-genome example of 22 *Solanum* species, the African eggplant (*S. aethiopicum*) accessions were used to create a mapping population enabling to detect genomic regions responsible for fruit size (Benoit *et al.*, 2025).

## Genomics variants in pan-genomes

Pan-genomes as a genomic reference resource are created by detecting genomic variation, like SNPs and SVs, present between closely related individuals (Golicz, Batley and Edwards, 2016; Tao *et al.*, 2019; Bayer *et al.*, 2020; Sherman and Salzberg, 2020; Della Coletta *et al.*, 2021; Jayakodi *et al.*, 2021; Shi *et al.*, 2023; Wang *et al.*, 2023; Matthews *et al.*, 2024). In order to do this, the individuals to be included in the pan-genome should be adequately sequenced to obtain high quality genomic data. Due to improved sequencing technologies the shift towards detection of larger SVs could take place which is complementary to the previously most common and quick SNP detection (Sherman and Salzberg, 2020; Bharadwaj *et al.*, 2021; Della Coletta *et al.*, 2021; Kumar *et al.*, 2021). For genome wide SVs detection, short-read sequencing is known to be error-prone, often not covering full large variants, easily missing inversions, trans-locations or other SVs (Chen *et al.*, 2014; Tattini, D'Aurizio and Magi, 2015; Golicz, Batley and Edwards, 2016; Sedlazeck *et al.*, 2018; Michael and VanBuren, 2020; Jayakodi *et al.*, 2021; Yuan *et al.*, 2021; Shi *et al.*, 2023; Matthews *et al.*, 2024). Hence, high quality long-read sequencing is a breakthrough technology used and better suited for constructing large and complex plant pan-genomes (Golicz, Batley and Edwards, 2016; Jayakodi *et al.*, 2021; Hameed *et al.*, 2022; Shi *et al.*, 2023; Matthews *et al.*, 2024). The only downside is that long-reads are costlier to generate and less common in public datasets than short-reads limiting their current availability for good quality pan-genome construction (Matthews *et al.*, 2024). Nevertheless long-reads are preferred for creating pan-genomes, with PacBio HiFi long-read sequencing improving fast along with algorithms and computing power, making it attractive for detecting SVs in plant genomes (Della Coletta *et al.*, 2021; Hameed *et al.*, 2022). Hence, the use of PacBio HiFi sequences is highly recommended while constructing plant pan-genomes since it ensures high quality of both primary reads and assembled sequences (Shi *et al.*, 2023).

The first created bacteria pan-genome was very different from most complex plant pan-genomes, as it was small and could be based on gene sequences only (Tettelin *et al.*, 2005; Jayakodi *et al.*, 2021). Another difference is that transposable elements (TEs) are highly present in the variable component of plant pan-genomes (Morgante, De Paoli and Radovic, 2007; Della Coletta *et al.*, 2021). TEs activity is one of the major forces involved in generating SVs creating the majority of insertions and deletions in plant genomes (Della Coletta *et al.*, 2021; Shi *et al.*, 2023). Hence it is expected that attention regarding TEs will increase in complex crop genomes (Morgante, De Paoli and Radovic, 2007; Della Coletta *et al.*, 2021). An alternative way to deal with the high prevalence of TEs as well as large sizes of plant genomes is by building pan-transcriptomes based on RNA-sequencing data which reveal the transcript-level differences between individuals (Morgante, De Paoli and Radovic, 2007; Sherman and Salzberg, 2020). This is often a more cost effective and targeted approach to investigate the genetics of plants, since focusing on genes only is more affordable and accessible (Sherman and Salzberg, 2020; Wang *et al.*, 2023). The development of plant pan-transcriptomes has been shown to be beneficial, since they enabled to detect RNA-seq transcripts which could not be aligned to the standard maize reference genome and aided finding disease resistance genes in wild versus cultivated barley (Sherman and Salzberg, 2020).

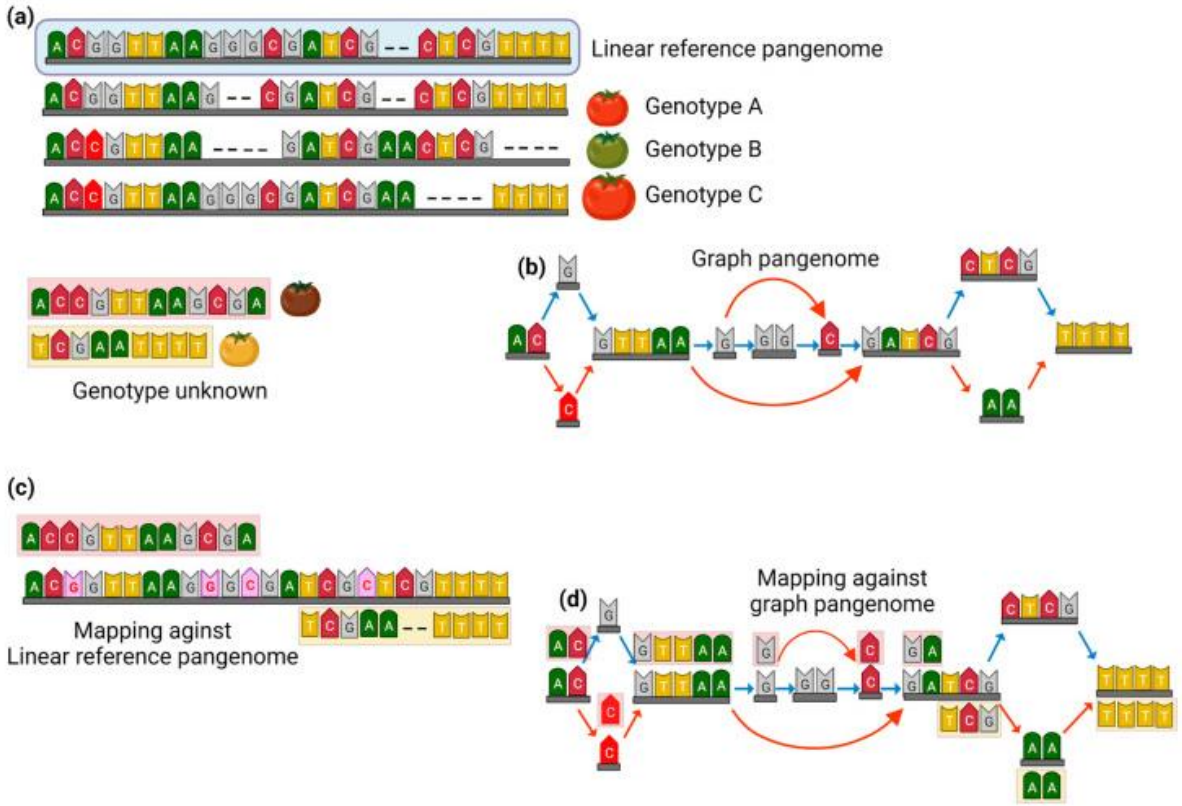



**Figure 1.8** Models of pan-genomes displaying the enhanced alignment towards 'graph-based' versus linear pan-genomes. **A)** Linear pan-genome constructed by aligning genomes of multiple accessions displaying sequence variation. **B)** Graph-based pan-genome constructed from the same aligned sequences, but now with variation between genotypes displayed with red arrows (edges) which are additional paths for alignment against the conserved regions (nodes). **C)** Alignment of the unknown genotypes towards the linear pan-genome leading to mismatched and sequence gaps as variant information is not fully captured. **D)** Alignment of the unknown individuals towards the 'graph-based' pan-genome can reveal all genetic variation and properly map to matching sequences through edges (red arrows) (Hameed et al., 2022).

## Pan-genome creation

Generally, there are two methods or types to construct pan-genomes; 1) a linear pan-genome (**Figure 1.8 A**), which includes homolog-based strategies and map-to-pan strategies, or 2) a 'graph-based' (**Figure 1.8 B**) pan-genome (Della Coletta *et al.*, 2021; Hameed *et al.*, 2022; Wang *et al.*, 2023; Matthews *et al.*, 2024). Homolog-based strategies were first developed in bacterial pan-genomes, which is a method of constructing a linear pan-genome by gene family clustering (Wang *et al.*, 2023; Matthews *et al.*, 2024). Homologous relationships between coding sequences can therefore be identified by orthogroups of gene families (Emms and Kelly, 2015). In the pan-genome context it enables identification of functionally similar genes across the individuals and hence identification of core and variable gene families (Golicz, Batley and Edwards, 2016; Glick and Mayrose, 2023). Several tools can be used to perform this orthology clustering analysis in plant pan-genomes like OrthoMCL, INPARANOID, reciprocal best blast hit or the more recent OrthoFinder (Golicz, Batley and Edwards, 2016; Glick and Mayrose, 2023). The latter, OrthoFinder, was tested in comparison to other older orthogroup inference methods and is claimed to be more accurate by between 8% and 33% (Emms and Kelly, 2015). In plant pan-transcriptomes also homolog-based strategies have been applied, since their creation is based on using genes only to be compared between individuals (Wang *et al.*, 2023).

With the map-to-pan approach reads are first mapped to a reference genome, followed by de novo assembly of the unmapped reads into contigs, which are then added to the reference assembly to form a linear pan-genome (Della Coletta *et al.*, 2021; Shi *et al.*, 2023; Wang *et al.*, 2023; Matthews *et al.*, 2024). This map-to-pan approach is a very common method which is used successfully for building pan-genome in numerous crops in earlier studies based on short-reads, but can also similarly be applied to more high quality genome assemblies based on long-reads (Shi *et al.*, 2023). It is a simplified method compared to the time consuming and computationally intensive de Bruijn graph-based framework, which was originally used for short-read assemblies and created by overlapping sets of *k*-mers, which are short sequences (Golicz, Batley and Edwards, 2016; Khan *et al.*, 2020; Shi *et al.*, 2023; Matthews *et al.*, 2024). The map-to-pan approach is a good strategy to minimize errors by using information of a high quality reference genome, however, genomic locations of newly assembled contigs remain unknown without further analysis (Della Coletta *et al.*, 2021). Due to the unknown locations of sequences the linear pan-genome does not reveal the locations of mutations, which makes this type of pan-genome unsuitable for usage as a reference genome (Wang *et al.*, 2023). When hypothetically trying to map towards a linear pan-genome (**Figure 1.8 C**) it results in mismatches and sequence gaps due to not all variants being captured, while aligning towards a 'graph-based' pan-genome (**Figure 1.8 D**) avoids mismatches due to the possibility to map with maximum coverage (Hameed *et al.*, 2022).

The 'graph-based' method for pan-genome construction recently gained much popularity (Hameed *et al.*, 2022) It is a newer, more effective and a more information-rich pan-genome construction approach than the linear pan-genome (Della Coletta *et al.*, 2021; Shi *et al.*, 2023; Matthews *et al.*, 2024). It is created by storing any variant (SNP or SV) along a reference genome as a node at the genomic location where it is discovered with the connections of the variants towards the nodes referred to as edges (Della Coletta *et al.*, 2021; Hameed *et al.*, 2022; Shi *et al.*, 2023; Matthews *et al.*, 2024) (**Figure 1.8 B & D**). Building and indexing the graph can be limiting steps, but once the graph has been set up reads can usually be aligned fast to this pan-genome reference, and graph aligners are expected to better utilize graph-based genomes for calling SVs (Shi *et al.*, 2023). There are multiple tools available for creating graph-based plant pan-genomes with long-reads. The construction is typically performed in several steps within workflows with tools such as NGMLR, Pbmm2 and Minimap2 for read alignment, or Sniffles, SVIM, CuteSV and PBSV for PAV and SV calling (Wang *et al.*, 2023).

### Onward analysis after construction

Once the graph-based pan-genome is created, visualization remains difficult due to multiple different paths through a graph combing different lengths of genomic sequences (Matthews *et al.*, 2024). Graph visualization tools typically focus either on the larger graph variation versus nucleotide base variants, with moving between these scales still a problem, especially for larger eukaryotic genomes (Matthews *et al.*, 2024). Some large variation or macro-level visualization tools are Bandage, GfaViz and Panache while some small variation or base-level visualization tools are vg viz, Sequence Tube Map and ODGI (Wang *et al.*, 2023). To build on the strength of using a single linear genome as well as graph-based pan-genomes as references for read alignment a hybrid approach has recently been developed (Della Coletta *et al.*, 2021). In this approach reads are first mapped to a graph-based pan-

genome, then the path reads are most associated with are used to build a linear reference genome towards which all reads are aligned again using a single linear read mapper (Grytten *et al.*, 2020) (**Figure 1.9**). This increases mapping accuracy compared to aligning towards a graph-based pan-genome alone (Grytten *et al.*, 2020; Della Coletta *et al.*, 2021).

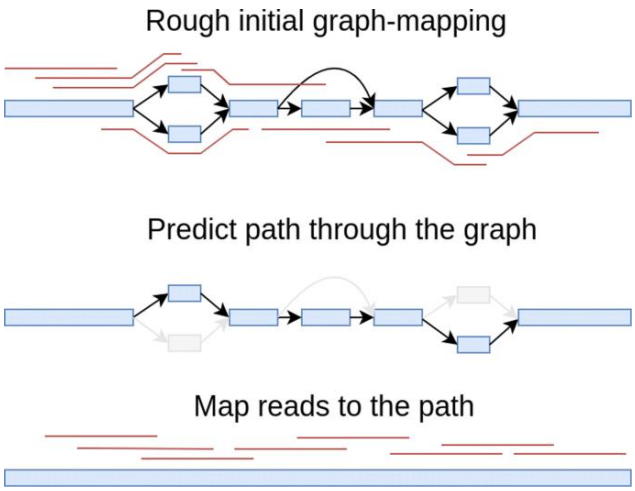



**Figure 1.9** Optimized read alignment by combining single linear genome as well as graph-based pan-genome references. Top: First reads (red) are roughly mapped to the graph-based pan-genome genome (nodes are blue; edges are black arrows). Middle: Then based on where most reads map a path is a predicted through the graph. Bottom: reads are aligned towards the linear path derived from the graph for optimal mapping (Grytten et al., 2020).

In modern plant pan-genomic studies both linear and graph-based construction methods are commonly used simultaneously and respectively for identifying functionally similar orthologous genes, as well as genomic variation. Some of the many recent literature examples available with linear as well as graph-based construction methods combined into one publication are soybean (Liu *et al.*, 2020), radish (*Raphanus*) (Zhang *et al.*, 2021), cucumber (*Cucumis sativus* L.) (Li *et al.*, 2022), pearl millet (Yan *et al.*, 2023) and tea (Chen *et al.*, 2023) pan-genomes. Building a graph-based pan-genome with genomic variation detected by SNPs and SVs, can be useful for mapping relevant traits. Especially when creating experimental mapping populations of which the founders can be constructed into a pan-genome to capture their genomic variation potentially underlaying traits.

## Multi-parent advanced generation inter-cross (MAGIC) populations for trait mapping

### Genetics underlaying traits

Selective breeding based on solely phenotypes to improve crops and livestock has been performed for centuries, although, the genomics responsible for favourable traits have often been discovered later. Some famous examples of successful breeding were during the green revolution in the 1960s, where wheat and rice plants were bred to be shorter to improve lodging resistance and obtain higher yields. The dwarf genes responsible for these traits were discovered decades later, and interestingly the rice ‘green evolution’ dwarf gene was shown to already been selected during early rice domestication by ancient humans (Asano *et al.*, 2011; Wang *et al.*, 2023). As previously described, sequencing techniques have recently drastically improved, with NGS revolutionizing the speed of generating accurate full genome sequences (Behjati and Tarpey, 2013). Additionally, the later improvement of long-read sequencing for generating high quality genomes enables SV detection and construction of pan-genomes in complex plant genomics (Golicz, Batley and Edwards, 2016; De Coster, Weissensteiner and Sedlazeck, 2021; Jayakodi *et al.*, 2021; Hameed *et al.*, 2022; Shi *et al.*, 2023; Yildiz *et al.*, 2023; Matthews *et al.*, 2024). These are all promising developments easing capturing genetic variation among individuals. However, the main challenge in molecular based plant breeding remains predicting how genetic polymorphism underlie desirable crop phenotypes like growth rate and yield (Korte and Farlow, 2013).

The longest-standing question in genetics is to understand how genetic variation contributes to phenotypic variation (Boyle, Li and Pritchard, 2017). The earliest way of trying to explain genes underlaying traits was by Gregor Mendel in 1866 through his experiments with pea plant hybridization (Hofer and Ellis, 2025). According to Mendelian genetics, traits which segregate in the progeny of hybrids are controlled by a single gene with a mutant allele (Hofer and Ellis, 2025). Mendel investigated the inheritance and segregation of seven traits, with for example pea varieties having either dominant or recessive alleles causing their mature cotyledons to be respectively yellow or green (Hofer and Ellis, 2025). Mendel's inheritance theory is still believed to be true, with some of the investigated pea genes being confirmed with modern techniques (Hofer and Ellis, 2025). However, already in the early 1900s Mendelian genetics was debated since this explanation of inheritance could only be applied to discrete and monogenic phenotypes, while it could not explain the inheritance of traits with a continuous distribution such as height (Boyle, Li and Pritchard, 2017). This debate was largely resolved with the 'infinitesimal model' in the 1918 paper by RA Fisher discussing the genetic architecture of complex traits (Boyle, Li and Pritchard, 2017; Abdellaoui *et al.*, 2023). It explains that many genes can affect a trait resulting in a continuous, normally distributed phenotype in a population, and the larger the number of genes affecting the trait, the smaller the contribution of each gene becomes (Boyle, Li and Pritchard, 2017).

### Techniques for discovering genetics responsible for trait variation

To identify multiple underlying causative genomic regions for a complex phenotypic trait is obviously more challenging compared to if a single gene is responsible. Up to 1980 studies attempting to investigate quantitative traits involved statistical methods based on means, variances and covariances of relatives, lacking actual knowledge on gene number and location underlying the traits (Kearsey and Farquhar, 1998). To incorporate genomic information and investigate genetics underlaying complex traits various techniques have been designed, including QTL mapping or Genome-Wide Association Studies (GWAS), with both techniques screening individuals differing in genotypes for phenotypes of interest (Korte and Farlow, 2013). QTL mapping is the oldest method of the two. By introduction of the term Quantitative Trait Locus (QTL) it was meant that a genomic region is associated with a continuously varying character (Geldermann, 1975). QTL can be one or multiple regions (loci) in the genome, which contain the causal genes for the quantitative trait, and can be mapped by using markers (Geldermann, 1975). The basic technique for marker mapping had been available since the 1920s, however, this focussed on a single responsible marker, and hence had to be updated to handle multiple markers simultaneously (Kearsey and Farquhar, 1998). In the 1980s the first QTL studies using multiple molecular markers emerged, greatly aided by developments in molecular biology like improved sequencing and DNA marker techniques (Kearsey and Farquhar, 1998).

Once identified by QTL mapping, markers associated with traits can be highly relevant for breeding practices. For example, by conducting marker-assisted selection (MAS) aiming to integrate the identified markers which are tightly linked to agronomically relevant genes in accessions to obtain improved cultivars (Collard *et al.*, 2005; Cockram and Mackay, 2018; Arrones *et al.*, 2020). There are various types of mapping populations to map QTLs for relevant traits in plants and other organisms. Traditionally, bi-parental populations were used

as a simple to develop method for high power QTL detection and mapping in plants (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020; Ponce *et al.*, 2020). For this method two parental accessions differing in the trait of interest are crossed in order to construct populations like $F_2$ and first backcross (BC1) generations (Xu *et al.*, 2017; Arrones *et al.*, 2020). These bi-parental populations can be directly studied, but are ideally fixed by selfing to generate homozygous populations consisting of double haploids (DH), recombinant inbred lines (RILs) or near isogenic lines (NILs) (Xu *et al.*, 2017; Arrones *et al.*, 2020). These homozygous individuals with fixed genotypes are then useful for high-resolution genetic mapping and reproducible phenotyping across multiple generations. Bi-parental methodologies have successfully been used for QTL mapping in many plant and crop species. Some examples are mapping grain yield in eight third generation maize bi-parental populations (ranging 148 to 300 progenies) under drought stress (Beyene *et al.*, 2015), a 451 individual second generation rice bi-parental population enabling mapping bacterial blight resistance (Kim and Reinke, 2019), or mapping stem rust resistance using two (145 and 171 RILs) fifth generation wheat bi-parental populations (Randhawa *et al.*, 2018). However, a commonly known remaining issue is the narrow genetic base due to using only two parents, which limits genetic diversity, reduces the number of recombination events leading to low mapping resolution, and decreases statistical power for detecting trait associations (Jannink, 2007; Huang *et al.*, 2015).

Genome-wide association studies (GWAS) using a large number of individuals with unknown kinship for obtaining high-genetic-diversity panels, were proposed to overcome these limitations of QTL detection in bi-parental populations (Korte and Farlow, 2013; Pascual *et al.*, 2016; Arrones *et al.*, 2020; Ponce *et al.*, 2020). The advantage of GWAS is that it investigates a large and assumed to be natural population, rather than controlled experimental populations, capturing broader genetic diversity and more historical recombination improving mapping resolution (Korte and Farlow, 2013; Visscher *et al.*, 2017; Abdellaoui *et al.*, 2023). In order to do this a minimum sample size of at least 100 individuals is suggested enabling sufficient statistical power to perform GWAS with valid results (Visscher *et al.*, 2017). This is due to this statistical power in GWAS being determined by sample size, allele frequency, and effect size of genetic variants. Larger sample sizes up to (tens of) thousands of individuals are needed to detect small effect size variants, while large effect size variants require smaller sample sizes, and 100 individuals is considered sufficient for exploratory studies (Visscher *et al.*, 2017). The concept of GWAS is only 20 years old, with the first GWAS conducted in 2005 to investigate the disease trait age-related macular degeneration (AMD) in a human population (Southam and Zeggini, 2025). Some successful plant GWAS examples are SNP markers in rice being associated with 14 agronomic traits including grain yield (Huang *et al.*, 2010), or with bacterial blight pathogen resistance (Kim and Reinke, 2019). In the latter study the mapped genomic region of interest was confirmed by QTL mapping using a bi-parental population (Kim and Reinke, 2019). GWAS is a powerful approach with many advantages over experimental mapping populations. However, GWAS success can be limited if related individuals are included sharing both causal and non-causal alleles resulting in confounded results due to linkage disequilibrium (LD) hindering identification of true causal variants (Korte and Farlow, 2013).

### Multi-parent populations capturing diversity

GWAS requires a large sample of individuals in order to capture the broad spectrum of diversity, requiring more original resources compared to experimental populations created by only a few individuals as a starting point. To enable capturing more genetic diversity into experimental populations, multi-parent populations were introduced, overcoming limitations of traditional QTL mapping like bi-parental populations (Ladejobi *et al.*, 2016). For example, it is proven more QTL are detected in founder accessions of an eight accession multi-parental tomato population, compared to in a bi-parental population, which encourages multi-parental QTL mapping approaches (Pascual *et al.*, 2016). The two most common types of multi-parent populations are nested association mapping (NAM) and multi-parent advanced generation inter-cross (MAGIC) (Ladejobi *et al.*, 2016). NAM populations are made by crossing several diverse lines into a single inbred parent of which the F1 are selfed for several generations to produce a panel of RILs after several generations of selfing (Ladejobi *et al.*, 2016). The concept of MAGIC populations was firstly proposed by Mackay and Powell (2007). Like other mapping populations, a MAGIC population is used for two purposes: 1) obtaining a permanent population consisting of RILs for precise QTL mapping and 2) development of newly bred materials which can either be used directly, or used for further breeding (Arrones *et al.*, 2020). Achieving genome fixation in the offspring is highly relevant for accurate QTL mapping and the possibility to generate reproducible accessions. Hence additional MAGIC-like multi-parent populations are suggested for crops which cannot be selfed due to self-incompatibility, limiting generation of pure lines (Arrones *et al.*, 2020). The degree of fixation can still be increased by conducting repeated generations of inbreeding through sibling mating. This idea is based on what was originally applied in animals, for which their difficulty to induce inbreeding is overcome by generating heterogeneous stocks in mice enabling robust QTL fine mapping (Mott *et al.*, 2000).

A standard MAGIC population is created by crossing multiple founder parents to produce high quantities of RILs as offspring, which could be considered genome mosaics of their parent accessions (Dell'Acqua *et al.*, 2015; Huang *et al.*, 2015; Ladejobi *et al.*, 2016; Arrones *et al.*, 2020). Therefore, all founder parents theoretically contribute equally to the haplotypes of the RIL genome mosaics (Dell'Acqua *et al.*, 2015; Ladejobi *et al.*, 2016; Arrones *et al.*, 2020). In this way, high genetic diversity, a balanced allele frequency and dense recombination due to multiple intercrossing cycles is obtained within RILs of MAGIC populations, allowing more precise QTL mapping than in bi-parental (Mackay *et al.*, 2014; Dell'Acqua *et al.*, 2015; Huang *et al.*, 2015; Cockram and Mackay, 2018; Ravelombola, Shi and Huynh, 2021). Crops with larger genomes contain more genetic information with more potential QTL. Hence MAGIC population sizes with sufficient offspring RILs are essential to generate more recombination events, and population sizes are required to increase with genome size of the investigated crop (Arrones *et al.*, 2020). As an example, for a cowpea eight parent MAGIC population having a diploid 620 Mb genome, a population size of 305 $F_8$ RILs was shown to be sufficient for QTL mapping, however, a crop like barley or diploid wheat with a 5 Gb genome range ideally requires a population size of about 2500 RILs (Huynh *et al.*, 2018).

In order to capture broad diversity, which is the advantage of MAGIC over bi-parental populations, founders with variable genetics, phenotypes and geographic backgrounds should be selected for obtaining adequate results (Arrones *et al.*, 2020). However, by broadening the variation, for example by including a wild species among the founders,

additional issues might occur while creating MAGIC populations such as lack of recombination in some genomic regions or unintentional selection of certain phenotypic traits such as lack of seed dormancy or early flowering time causing lower numbers of offspring (Arrones *et al.*, 2020). Due to these known difficulties, most current MAGIC populations capture only intraspecific diversity (Arrones *et al.*, 2020). However, there are some exceptions like existing tomato MAGIC populations which are interspecific incorporating wild founders. For example seven *Solanum lycopersicum* lines which are crossed with one wild *Solanum cheesmaniae* accession were used to map disease resistance and fruit shape (Campanelli *et al.*, 2019). For developing ToMAGIC, four weedy cherry-type *S. lycopersicum var. cerasiforme* accessions are crossed with four wild ancestral *Solanum pimpinellifolium* tomato species, enabling mapping of fruit size (Arrones *et al.*, 2024).

### Crossing schemes and proven success of MAGIC populations

Two main approaches to create MAGIC populations are either a 'funnel' or 'diallel' approach (Arrones *et al.*, 2020). The 'funnel' approach starts with a funnel breeding scheme by intercrossing a number (n) of founder parent accessions a single time to create n/2 F1 hybrids. These hybrids are then intercrossed in a mating design to combine the genomes of all founder parent accessions in the offspring (**Figure 1.10 A**). In the 'diallel' approach the founder parent accessions are intercrossed similar to a funnel design, but now in multiple funnels incorporating all possible crossing combinations between the founders for creating F1 (Arrones *et al.*, 2020). The created F1 offspring is then intercrossed too, until all founder parent genomes are present in the lastly created generation (**Figure 1.10 B**). As a last step for funnel as well as diallel approaches, either single-seed descent (SSD) or DH production is performed in order to achieve high homozygosity among the offspring individuals (**Figure 1.10 C**). SSD means inbreeding of the individuals by several generations of selfing, while just selecting each accession from a single parent, therefore creating progeny representing the genetic diversity of the $F_2$ generation (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020; Singh, Singh and Singh, 2021).

The 'diallel' is similar to the 'funnel' approach, but it covers all possible ways of cross combinations in multiple funnels (Arrones *et al.*, 2020). So, for an eight founder MAGIC population, the first generation would be all 28 two-way crosses, followed by all possible four-way, and eight-way crosses. Due to a higher diversity of allele combinations, the 'diallel' approach is preferred for MAGIC population creation. However, reducing the number of hybridizations using the single 'funnel' approach, is useful for faster advancement of development of MAGIC populations and in case of using founders with hybridization difficulties (Arrones *et al.*, 2020). Often a 'half-diallel' approach is used to create MAGIC populations, in which multiple funnel crossings are used but not all

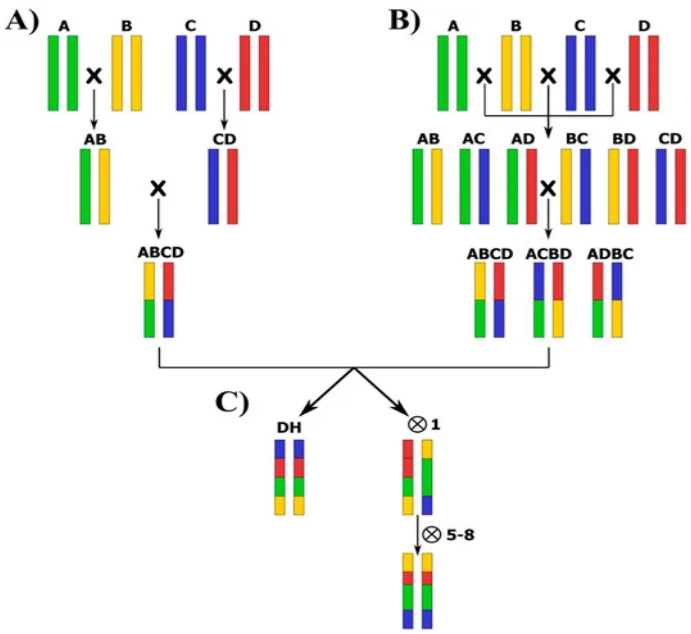



**Figure 1.10** Potential approaches for a four-way MAGIC population creation consisting of four founder parent accessions; A, B, C, and D. **A)** the 'funnel' approach, **B)** the 'diallel' approach and **C)** inbreeding and homozygosity achievement by either single-seed descent (SSD) or double haploid (DH) production (Arrones et al., 2020).

possible cross combinations are included, due to impracticality of hybridization difficulties or labour intensity. Here two examples of eight founder parent accession MAGIC population pedigrees are displayed, using either a single (**Figure 1.11 A**) or multiple funnel (**Figure 1.11 B**) MAGIC crossing design (Ladejobi *et al.*, 2016). The multiple funnel design is a 'half-diallel' approach, only including all possible crossing ways to obtain the F1 generation (28 two-way crosses), but not to generate the following generations with four-way and eight-way crosses (Ladejobi *et al.*, 2016).

Since its first proposal in 2007 (Mackay and Powell, 2007), many crop MAGIC populations have been successfully been created, enabling mapping of various agronomically relevant traits. Some recent examples are cowpea (Huynh *et al.*, 2018), maize (Zea mays) (Dell'Acqua *et al.*, 2015), common bean (Phaseolus vulgaris L.) (Diaz *et al.*, 2020), Sorghum (Sorghum bicolor (L.) Moench) (Ongom and Ejeta, 2017), rice (Ponce *et al.*, 2020), several tomato (Campanelli *et al.*, 2019; Diouf *et al.*, 2020; Arrones *et al.*, 2024) and many different wheat (Huang *et al.*, 2012; Mackay *et al.*, 2014; Gardner, Wittern and Mackay, 2016; Sannemann *et al.*, 2018; Scott *et al.*, 2021) MAGIC populations. The number of parental lines in these MAGIC populations varied from four (Huang *et al.*, 2012; Ponce *et al.*, 2020) to eight (Mackay *et al.*, 2014; Dell'Acqua *et al.*, 2015; Gardner, Wittern and Mackay, 2016; Huynh *et al.*, 2018; Sannemann *et al.*, 2018; Campanelli *et al.*, 2019; Diaz *et al.*, 2020; Diouf *et al.*, 2020; Arrones *et al.*, 2024) to 16 (Scott *et al.*, 2021) to even 19 founders (Ongom and Ejeta, 2017).

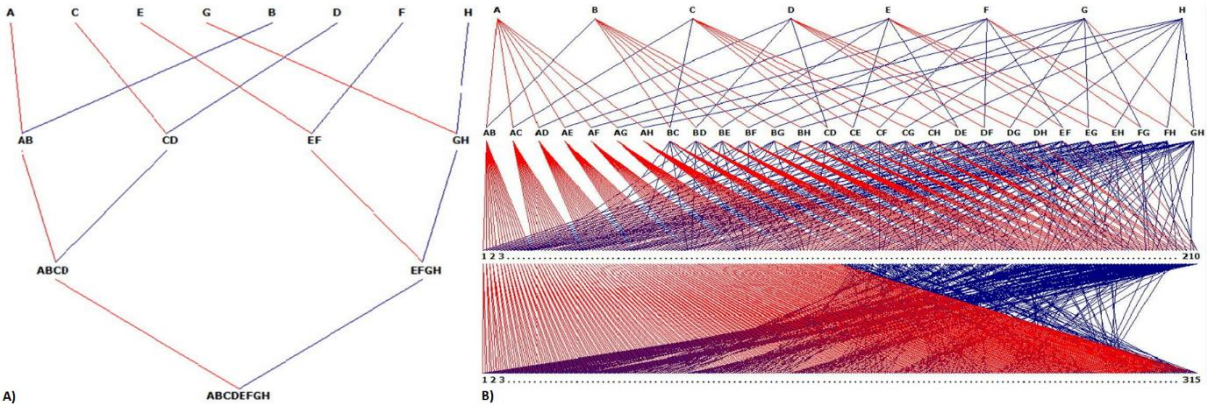



**Figure 1.11** Pedigree representation of two eight parent accession MAGIC populations using either a single **A)** or multiple **B)** funnel crossing design. **A)** Parents of the single funnel design were paired to create four F1s which in their turn were followed by two four-way crosses and one eight-way cross. **B)** Parent accessions of the multiple funnel design with a half diallel approach are intercrossed to create a 28 two-way F1, which in its turn was crossed to 15 complementary F1, creating 210 four-way hybrids. Then 315 eight-way individuals were produced by crossing each four-way hybrid to three complementary four-way hybrids (Ladejobi et al., 2016).

Single (Huang *et al.*, 2012; Sannemann *et al.*, 2018; Campanelli *et al.*, 2019; Diouf *et al.*, 2020) as well as multiple funnels (Mackay *et al.*, 2014; Dell'Acqua *et al.*, 2015; Gardner, Wittern and Mackay, 2016; Ongom and Ejeta, 2017; Huynh *et al.*, 2018; Diaz *et al.*, 2020; Ponce *et al.*, 2020; Scott *et al.*, 2021; Arrones *et al.*, 2024) were used as crossing designs. A broad variety of traits were enabled to be mapped with these populations, like plant height with hectolitre weight and dwarf genes (Huang *et al.*, 2012; Ongom and Ejeta, 2017; Sannemann *et al.*, 2018), cowpea flowering time, growth habit, seed size and maturity (Huynh *et al.*, 2018), maize grain yield and flowering time (Dell'Acqua *et al.*, 2015), bean phenology traits and yield (Diaz *et al.*, 2020), rice grain size (Ponce *et al.*, 2020), wheat awn presence (Mackay *et al.*, 2014) and another large variety of 47 agronomic traits (Scott *et al.*,

2021). In tomato fruit shape (Campanelli *et al.*, 2019) and fruit size (Arrones *et al.*, 2024) could be mapped, just as various other traits measured in multi-environment trials revealing Genotype by Environment interaction (GxE) interactions (Diouf *et al.*, 2020). In all mentioned populations SNPs were used as markers to detect the QTL which can lead to the identification of candidate genes underlaying traits (Huang *et al.*, 2012; Mackay *et al.*, 2014; Dell'Acqua *et al.*, 2015; Gardner, Wittern and Mackay, 2016; Ongom and Ejeta, 2017; Huynh *et al.*, 2018; Sannemann *et al.*, 2018; Campanelli *et al.*, 2019; Diaz *et al.*, 2020; Diouf *et al.*, 2020; Ponce *et al.*, 2020; Scott *et al.*, 2021; Arrones *et al.*, 2024). Generally, these are a broad range of examples that the creation of MAGIC populations is successful for mapping agronomic traits, with SNPs being a useful type of marker to detect QTL. No MAGIC population has yet been created for amaranths, neither for grain nor leafy varieties. By capturing much variation using diverse amaranth accessions in a MAGIC population, a large variety of relevant agronomic traits could potentially be identified. This would hugely benefit the *Amaranthus* breeding programme by conducting MAS while making breeding efforts for obtaining improved varieties.

## 1.4 Project description and research aims

Leafy amaranth is commonly grown, traded and eaten by locals in South Africa, has large potential for genetic improvement and is known for its great nutritional benefits (Grubben, 2004; Rastogi and Shukla, 2013; Achigan-Dako, Sogbohossou and Maundu, 2014; Mabhaudhi *et al.*, 2016; Emmanuel and Babalola, 2021). However, there is still a lack of agronomic knowledge regarding leafy amaranth and how it could contribute to sustainable food systems in South Africa (Mabhaudhi *et al.*, 2019; Emmanuel and Babalola, 2021). By exploiting leafy *Amaranthus* diversity with multi-omics approaches, gaps in knowledge regarding the crop can be addressed which is beneficial for breeding and other scientific purposes. The main goal of this PhD thesis is therefore using novel generated molecular data of amaranths, including genomics, transcriptomics and metabolomics datasets, to expand the current scientific knowledge regarding leafy *Amaranthus*. Analysis of this newly obtained information could offer support and enhance efforts to develop improved amaranth lines for various agronomic traits to contribute to food and nutrition security in Africa.

Several scientific approaches of integrative omics data types have been performed in this thesis, aided by collaborations, to contribute to expanding current knowledge on leafy amaranth. At first an *Amaranthus* pan-genome has been created using founder accessions of a mapping population. This novel *Amaranthus* pan-genome was developed with help of colleagues of VIB Ghent, resulting in a valuable manuscript which is incorporated in the appendix of this thesis. The work was partially conducted by shared first author Maud de Kinderen with her contributions and additional analysis conducted by her included in **Chapter 2**. The aim of generating this *Amaranthus* pan-genome is to acquire a highly relevant and complete genomic reference source not only for the current study but also for other scientific purposes and breeders. By creating this pan-genome, the genomic variation captured in it can be useful for QTL mapping, which potentially can reveal genetics underlaying relevant agronomic traits of interest. The *Amaranthus* mapping population of which the founders are constructed into a pan-genome is a Multi-parent Advanced Generation Inter-Cross (MAGIC) population, created by the World Vegetable Center (WVC). The goal of making this *Amaranthus* MAGIC population was to create a new and highly

valuable type of mapping population with founder lines capturing much phenotypic as well as genetic diversity. The second chapter, including the pan-genome manuscript, attempts to investigate that captured diversity in the founder lines by analysing phenotypic growth, metabolomics, transcriptomics and genomic data.

The captured variation among the founders is expected to segregate in the MAGIC offspring. Onward analysis of the MAGIC population, focussing on the progeny lines is addressed in chapter three. Here the aim is to assess the growth performance of the offspring, identify favourable lines and eventually map QTL for relevant agronomic traits in the population. The offspring RILs are grown in climatically different locations across South Africa, necessitating environment-specific considerations when selecting the best performing accessions for breeding as well as growing purposes. The creation of a MAGIC population and performing QTL mapping contributes to generation of new cultivars as well as identification of trait-associated genes. This is useful as information for the scientific community as well as for breeding material to obtain improved high yielding leafy amaranths. Hence in chapter three growth, leaf yield and other phenotypic trait data of the MAGIC progeny RILs is investigated incorporating genomics to enable QTL mapping.

In the fourth chapter a diversity set consisting of 90 accessions of various *Amaranthus* species has been assessed. Here the goal is to get a broader insight into the genetic diversity within the genus *Amaranthus*. Novel *Amaranthus* sequences are highly beneficial for further understanding of the phylogenetic relationships within the genus, since there is still much uncertainty and inconsistency in literature (Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Lin *et al.*, 2022; Blair *et al.*, 2023). While the MAGIC population project will be relevant regarding revealing leafy trait associations or creating novel breeding material the aim of chapter four is molecular characterisation of the diversity set. Genomic as well as metabolomic data is investigated to understand the relationships between the lines and profile metabolite content. This can be of great contribution to understanding speciation events, phylogeny, clarifying origins and diversification of the genus *Amaranthus*. Especially the elaborate metabolite profiles detected in this diversity set can provide relevant information regarding nutritionally important compounds.

The genomic as well as metabolomic data available for the MAGIC founders as well as diversity set enables the potential to identify putative BGC for amaranths. Hence a fifth chapter focusses on the attempt for putative BGC detection incorporating genomics and metabolomics of these two amaranth collections. Additional evidence for the identified putative BGC is then provided by analysing transcriptomics data to reveal co-expression of genes within the same BGC. The aim of this fifth chapter, focussing on BGC, is to further understand the biochemical pathways present in amaranths, elucidating their nutritional as well as medicinal properties.

In short, this thesis aims to investigate the following research objectives;

- Variation captured in a novel created *Amaranthus* pan-genome of MAGIC founders.
- Segregation of that variation in the MAGIC offspring potentially enabling to map QTL.
- Investigating taxonomy and nutrition in the genus *Amaranthus* by a diversity set.
- Identifying putative BGC in amaranths and their beneficial properties.

# Chapter 2: MAGIC founders and creation of an inter-specific *Amaranthus* pan-genome.

## Statement regarding chapter 2 and submitted manuscript

Using the MAGIC founders, an *Amaranthus* pan-genome has been created which has been submitted for publication. The manuscript which was first submitted is provided in the Appendix. A more recent version after peer review has been submitted which incorporates the new versions of *A. hypochondriacus* (Winkler *et al.*, 2024; Graf *et al.*, 2025) and *A. tricolor* (Wang *et al.*, 2022) reference genomes into the *Amaranthus* pan-genome. This of course will become available once accepted. Since the rest of this thesis is written based this older manuscript, with pan-components including a lower quality *A. hypochondriacus* (Lightfoot et al., 2017) in the pan-genome, it is this version of the manuscript which is presented here. The work in the manuscript was partially completed by external parties such as VIB Ghent who were of great help while building the pan-genome. About half of the work in this manuscript is completed by Maud de Kinderen who also wrote the first draft of the manuscript incorporating text from Xiao Ma of VIB. Hence this manuscript has shared first co-authorship between Maud and Xiao. Maud conducted the following data analysis regarding phenotypic field trials (**Figure 1**, **Supp Figure 1**, **Supp Table 1**), metabolomics (**Figure 2**, **Supp Figure 2**, **Supp Figure 3**) and genomics or GO-enrichment analysis (**Supp Figure 6B**, **Table 2**, **Figure 4A**, **Supp Dataset 6**, **Supp Dataset 7**). These and more figures and analysis regarding the MAGIC founders and pan-genome created by Maud which did not fit into the submitted manuscript is provided below in **chapter 2**.

## Introduction

All work conducted by Maud de Kinderen in the first submitted manuscript, including additional work with the available data of the MAGIC founders which didn't fit into the limitations of the manuscript, are presented in this chapter. All additional analysis regards further investigation in available metabolomics as well as genomics.

## Results

### Work in manuscript

#### The eight parent accessions of a *Amaranthus* MAGIC population show variation in agronomic traits

The World Vegetable Center (WorldVeg) conserves 1,455 *Amaranthus* accessions, with more than half of them classified as vegetable or leafy amaranth (Lin *et al.* 2022). From this large diversity panel, eight founder parent accessions were selected to generate a MAGIC population. These parents were classified in the WorldVeg genebank as four *Amaranthus hypochondriacus* and four *Amaranthus cruentus* species, although two of the lines were reclassified following *de novo* SNP calling (Lin *et al.* 2022)(Supplementary Dataset 1).

Agronomic characteristics of the MAGIC population parent accessions were assessed in an irrigated randomised complete block design field trial at WorldVeg in Arusha, Tanzania during the warm season (November 2019 – February 2020, average temperature 21.6°C) (Supp Dataset 2). Leaves were harvested five times during the trial with regrowth of the plants in between; time between harvests varied from 11 to 30 days. The significance of the variation in traits was assessed using linear mixed-effects models including harvest and accession as fixed variables along with their interaction (Supp Table 1). Leaf yield, length, width and number all varied significantly by harvest, with leaf yield peaking at the third harvest and the first and fifth harvests having the lowest yield (Figure 1, Supp Figure 1a). Leaf number followed a similar pattern (Supp Figure 1b) whereas leaf length and width decreased from the first harvest (Supp Figure 1c, d). Crucially there was significant variation between accessions for leaf yield (Supp Table 1), with a significant difference between accessions VI044371 and VI050446 for third harvest yield (Figure 1a) and total yield (i.e. the sum of all five harvests) (Figure 1b). VI044371 and VI050446 also showed a significant difference in leaf length (but not width) across the whole trial (Supp Figure 1e, f), likely underlying the difference in yield as leaf number did not differ significantly between accessions (Supp Table 1).

**Supplementary Table 1: Statistical significance of trait variation in the eight MAGIC population founder lines from irrigated field trials.** Results of the best fitting linear model with variables: founder accessions (Accession, n=8); Harvest (n=5); and Accession*Harvest interaction. P-values (P) are given for the fixed effects and displayed in bold when less than 0.01.

| **Model** | **Fixed** | | |
|---|---|---|---|
| | **Accession** | **Harvest** | **Accession*Harvest** |
| Leaf yield | **P = 0.0032** | **P < 2e-16** | P = 0.7301 |
| Leaf length | **P = 0.0001** | **P < 2e-16** | **P = 0.0035** |
| Leaf width | **P < 3e-6** | **P < 2e-16** | **P < 4e-6** |
| Number of leaves | P = 0.8562 | **P < 3e-10** | P = 0.7532 |

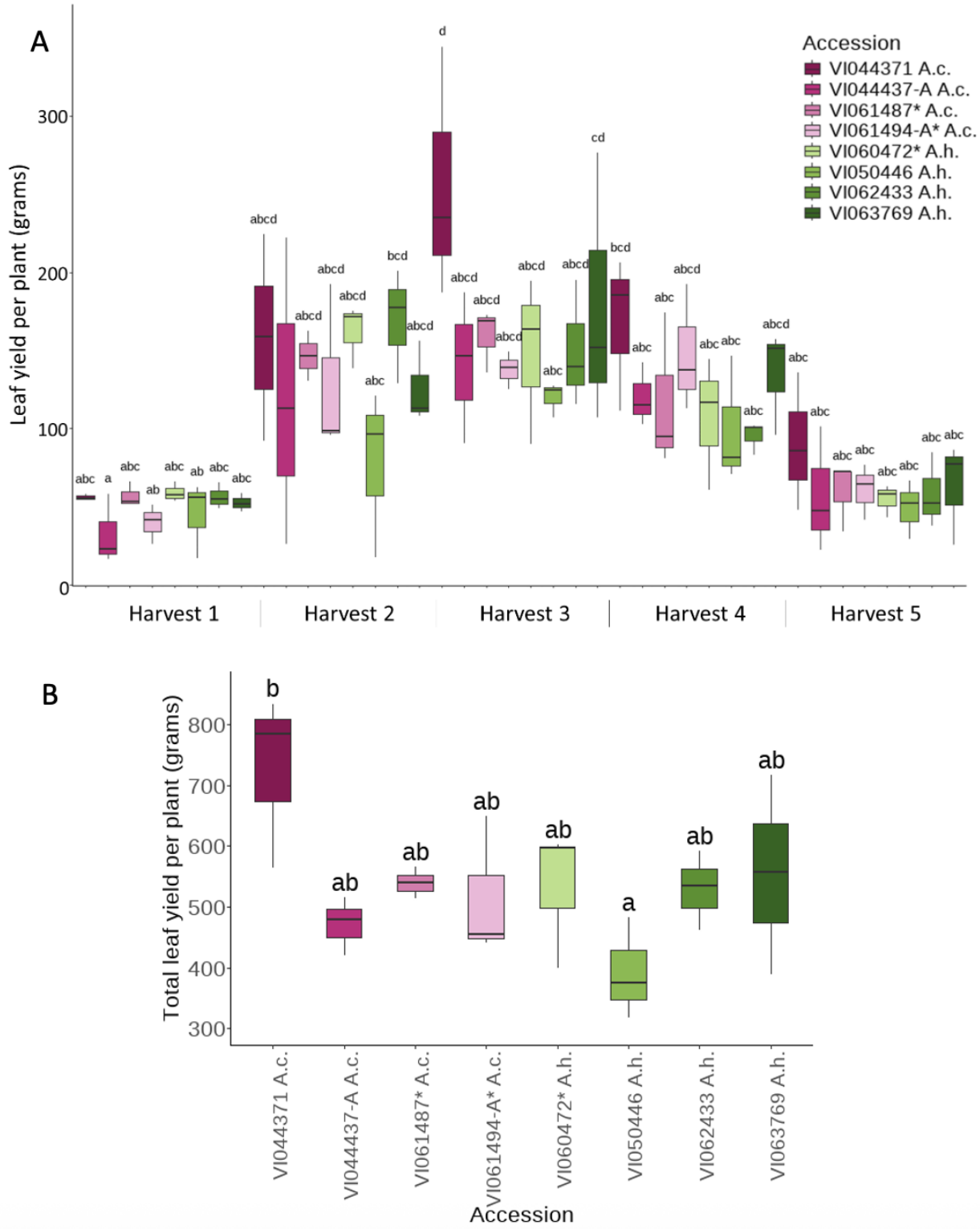
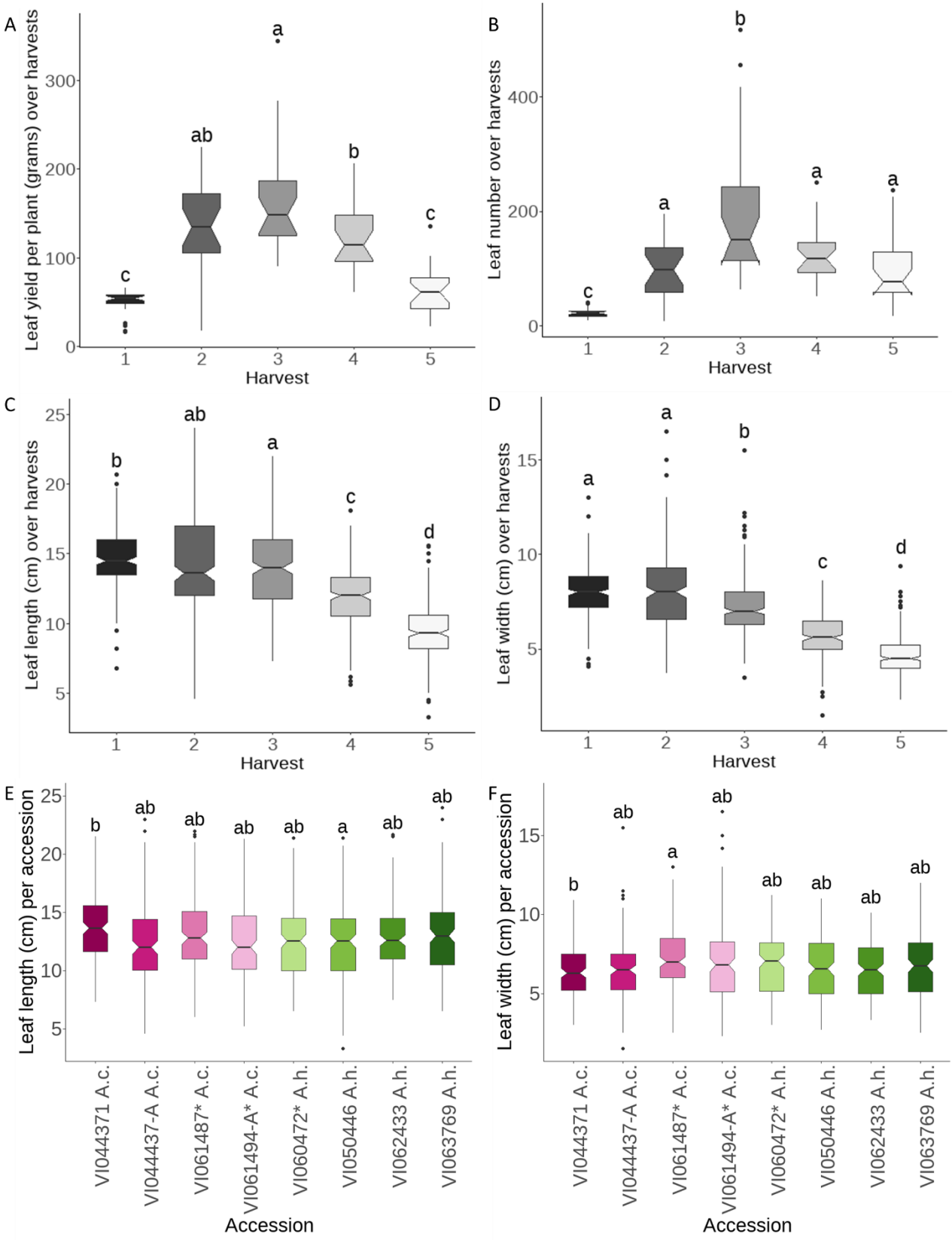



**Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A)** Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters indicate significant differences within each harvest as assessed by Tukey's honest significance test. Initials A.c. and A.h. indicate the species of each line with '*' indicating lines with admixture.



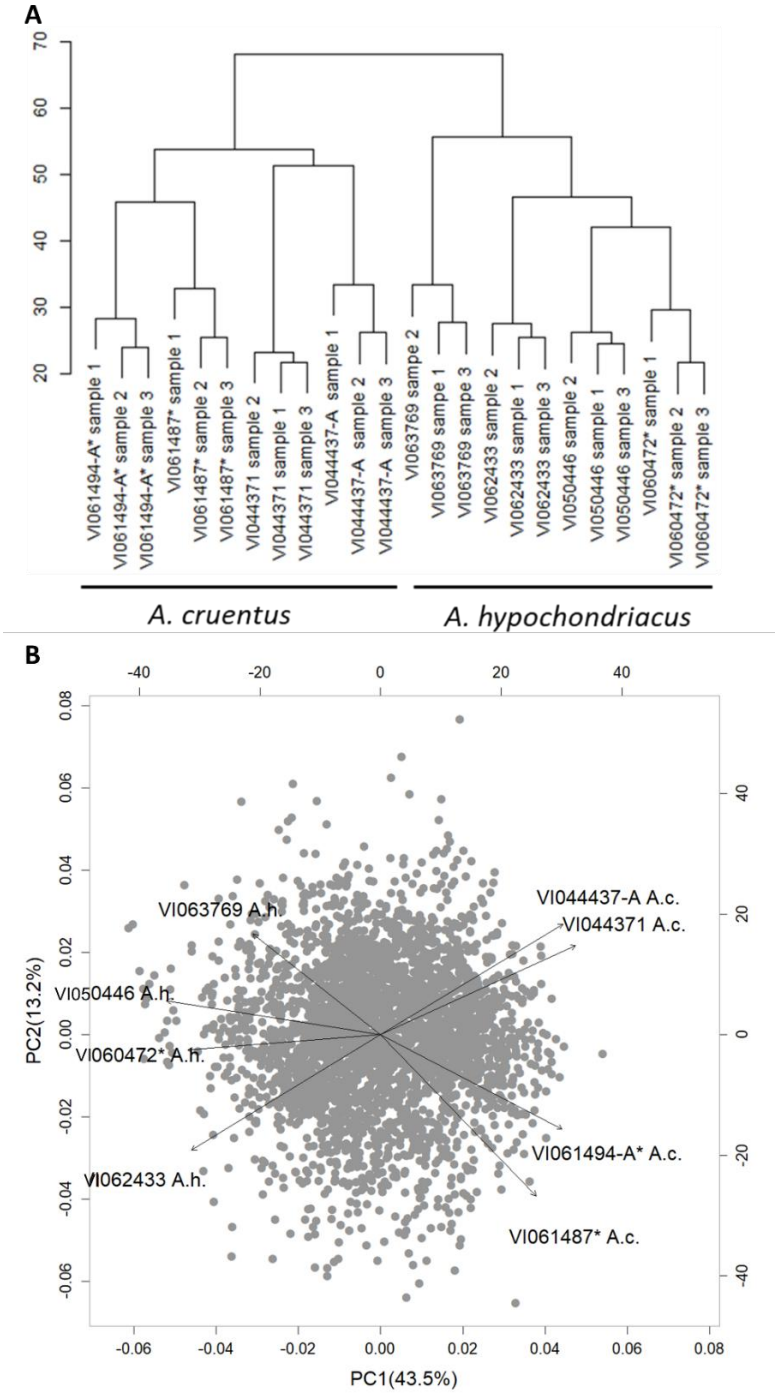

**Supplementary Figure 1. Leaf trait data from the irrigated field trial of the eight MAGIC parent accessions.** Box plots showing the median and interquartile range of **A)** leaf yield per plant at five different harvests (N = 120), **B)** number of leaves per harvest (N = 120), **C)** leaf length and **D)** leaf width at each harvest (N = 1200), **E)** leaf length and **F)** leaf width per accession over all harvests (N = 1200). Tukey's test letter codes indicate significance ($p < 0.05$) between the boxplots. '*' by accession IDs indicates the accessions showing genomic admixture.

## The eight amaranth accessions vary in metabolite profiles relevant to human nutrition

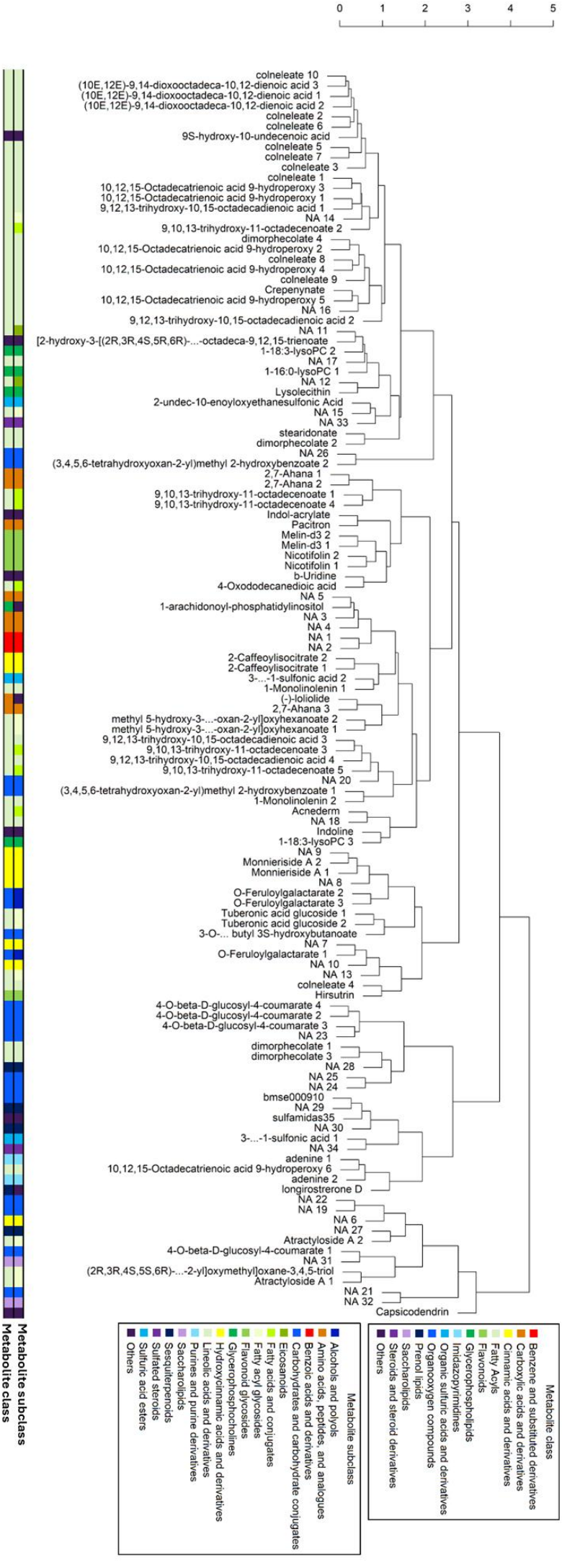
The nutritional content of traditional crops, such as amaranth, is an important factor in the value of these crops to local food systems. We assessed the eight founder lines for variation in metabolite content of greenhouse-grown leaves using untargeted metabolite profiling of methanolic extracts with three replicates.

Liquid chromatography–mass spectrometry (LC-MS) analysis resulted in 3256 features from the positive and negative ionization modes after correcting for redundancy (Supp Dataset 3). Hierarchical clustering using normalised data for these 3256 non-redundant features demonstrates the similarity of replicates as well as clear separation of the two species of amaranth (Supp Figure 2a). Principal component analysis using the abundance of the 3256 non-redundant features again demonstrated clear variation between the two amaranth species (Supp Fig 2b).

After removing very low abundance features, 1421 non-redundant peaks remained of which 1227 varied significantly in abundance between the 8 parent accessions and 847 showed a significant difference in abundance between the two species (*A. cruentus* and *A. hypochondriacus*) (Supp Dataset 3a). Of the 1227, 122 peaks could be annotated and represented 18 major metabolite classes. After clustering of these 122 features on the basis of their abundance across the samples, features with the same annotation cluster together (Supp Figure 3) providing support for the accuracy of annotations. The abundance of the 122 annotated metabolites was also able to clearly distinguish between the two species groups, consistent with their revised species attribution (Lin et al. 2022) (Figure 2a). Figure 2b highlights the variation in abundance of these 122 metabolites both across and within species. Of interest to our aim of developing improved amaranth lines, many of these annotated metabolites are of relevance to human nutrition.




**Supplementary Figure 2. Accession and metabolite sample clustering based on normalised data for 3259 non-redundant detected features. A)** Hierarchical clustering dendrogram revealing similarity between biological replicate samples as well as differences between the two species. **B)** A bi-plot demonstrating clear separation of the two species groups based on principal component analysis of the metabolites.




**Supplementary Figure 3:** Clustering of 122 annotated metabolites on their abundance across samples from the 8 Amaranth MAGIC population founder lines.

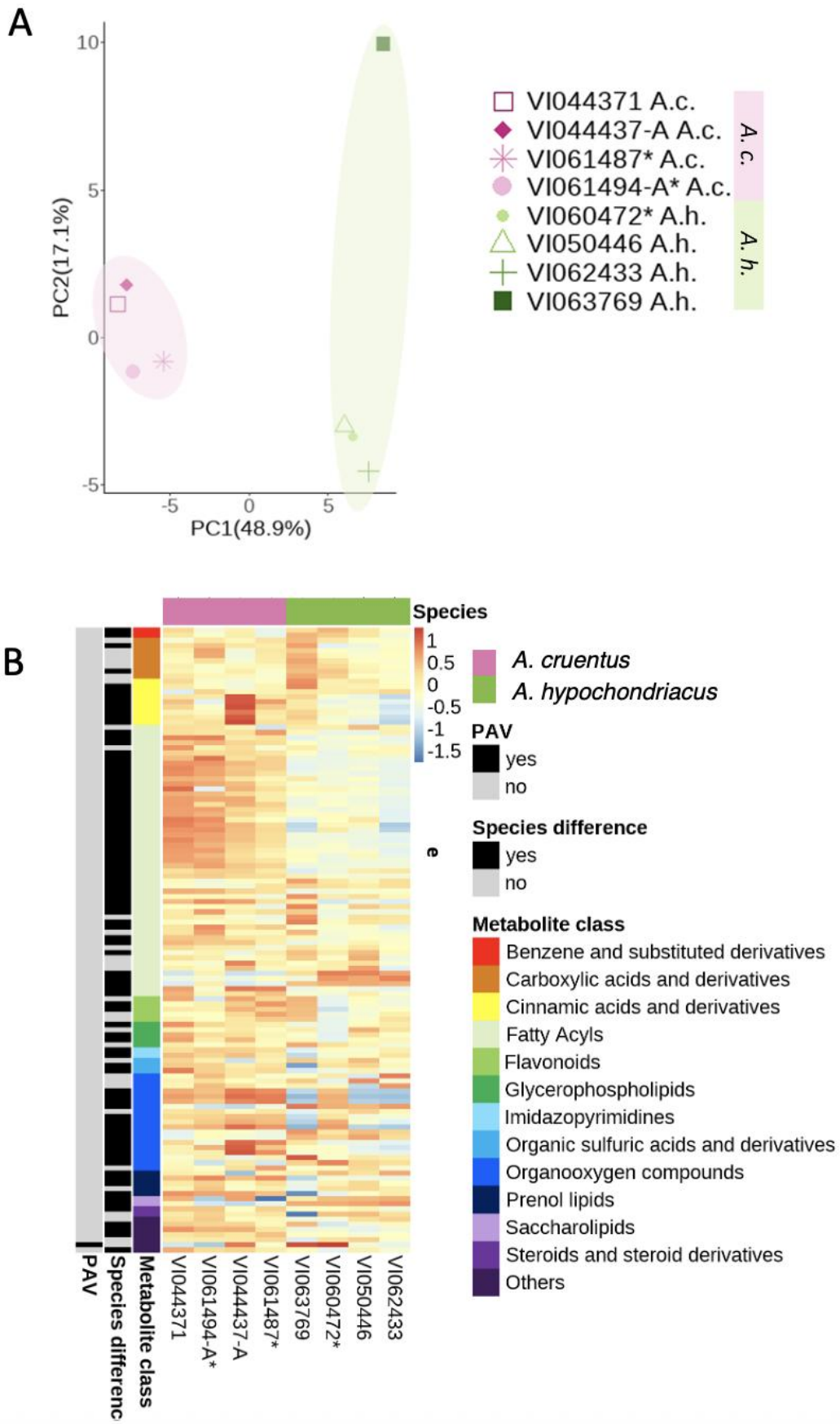


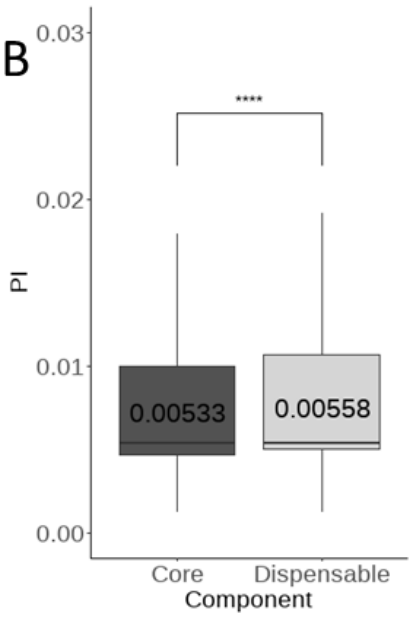

**Figure 2. Variation in metabolite content between the eight *Amaranthus* MAGIC population parent accessions. A)** Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

The largest group of annotated metabolites varying across the 8 accessions are lipids, including fatty acyls, glycerolipids, glycerophospholipids, prenols, saccharolipids and steroids. Lipids are obviously an important nutritional component, and many of these (particularly fatty acyls) appear to differ between the *A. cruentus* and *A. hypochondriacus* accessions (Figure 2b) with higher levels in the *A. cruentus* accessions. Polyphenolic compounds including flavonoids, benzenoids, sesquiterpenes, tannin and hydroxycinnamic acid esters vary across the 8 MAGIC accessions (Figure 2b). These compounds have been associated with multiple health benefits, most likely through their anti-oxidant properties (Monjotin *et al.* 2022).

## The genetic repertoire of an inter-specific Amaranth pangenome

Using Orthofinder (Emms and Kelly 2019), genes from the ten amaranth genomes were classified into 23,796 assigned families and 5,042 unassigned single-copy genes unique to one accession (Supplementary Dataset 6). 14,805 orthogroups were detected in all ten genomes and defined as the core gene families (51%), 8,979 were identified in 2 - 9 accessions, termed dispensable gene families (31%), and 12 gene families (41 genes) were detected in only one accession, which along with the single-copy genes unique to a single accession, resulted in 5,054 private orthogroups (18%). Evaluating conservation of genes within each amaranth species demonstrated a higher proportion of core genes in each species sub-pangenome compared to the full pangenome (Table 2).

Core genes showed reduced nucleotide diversity compared to dispensable genes indicating a greater level of conservation (Supp Figure 6b). The species-specific subpangenome of just the four *A. hypochondriacus* founder lines had a much higher proportion of core gene families than an *A. hypochondriacus* subpangenome also including the reference genome (Table 2). The annotation of the *A. hypochondriacus* reference genome appears to be of lower quality than the other accessions, with shorter gene lengths compared to the *A. cruentus* reference (Ma *et al.* 2021), which is likely the reason for the high proportion of private genes in this genome. In contrast, the subpangenome of just the *A. cruentus* MAGIC founders (not including the reference genome) had a similar proportion of private gene families to the full *A. cruentus* subpangenome Table 2).




**Supplementary Figure 6. Analysis of pangenome components. B,** nucleotide diversity (π) within core and dispensable genes.

We analysed GO term enrichment among the genes assigned to different pangenome components. Single-copy genes and gene families were newly assigned to pangenome components based on their presence in the 9 amaranth accessions ignoring the *A. hypochondriacus* reference (Supp Dataset 6a) with the key change being a significant reduction in the number of private genes (Table 2). GO term enrichment analysis did not reveal striking differences in categories of gene function between the genes in the pangenome components, although the most significantly enriched term in the core gene families was DNA-binding transcription factors with over 80% of the pangenome transcription factor genes being part of core gene families (Supp Dataset 7). Other central biological processes were also enriched such as RNA binding, chromatin binding, protein

deubiquitination, microtubule binding and nitrogen metabolism. Dispensable gene families included genes with a role in response to auxin, wounding and oxidative stress, as well as cell wall biogenesis, potentially underlying plant adaptation to changing environmental conditions (Supp Dataset 7).

**Table 2**: Composition of species-specific subpangenomes. The proportion of core, dispensable and private gene families (and single copy private genes) in the species-specific sub pangenomes, and subpangenomes of the *A. cruentus* and *A. hypochondriacus* MAGIC founder accessions only (i.e. not including the *A. cruentus* or *A. hypochondriacus* reference genomes of Ma et al. 2021 and Lightfoot et al. 2017, respectively).

| | **Full pangenome** | ***A. cruentus* accessions only** | ***A. hypochondriacus* accessions only** | ***A. cruentus* MAGIC founder lines only** | ***A. hypochondriacus* MAGIC founder lines only** |
|---|---|---|---|---|---|
| **Core gene families** | 51% | 72% | 64% | 80% | 86% |
| **Dispensable gene families** | 31% | 18% | 20% | 11% | 10% |
| **Private gene families and single copy private genes** | 18% | 10% | 16% | 9% | 4% |

## Sequence and structural variation in the amaranth pangenome

The HiFi sequencing reads of the eight MAGIC population parent accessions were aligned to the *A. cruentus* "Arusha" reference genome (Ma *et al.* 2021) to identify both sequence and structural variants. A total of 8,378,957 single nucleotide polymorphisms (SNPs) were identified: 1,767,497 were found in the four *A. cruentus* parent accessions, while 6,611,460 were detected in the four *A. hypochondriacus* parent accessions, indicating these genomes were less similar to the *A. cruentus* reference. Furthermore, the two *A. cruentus* accessions showing genome admixture (VI061487 and VI061494, Li et al. 2022) contained a higher number of SNPs compared to the two other *A. cruentus* accessions, and conversely, the *A. hypochondriacus* accession with genome admixture (VI060472) had the lowest number of variants amongst the *A. hypochondriacus* parent genomes, although the difference was less pronounced. Despite the genome admixture, PCA analysis using the SNPs clearly split the MAGIC parent lines into two species-defined groups (Figure 4a) as seen with the metabolite profiling data (Figure 2a).

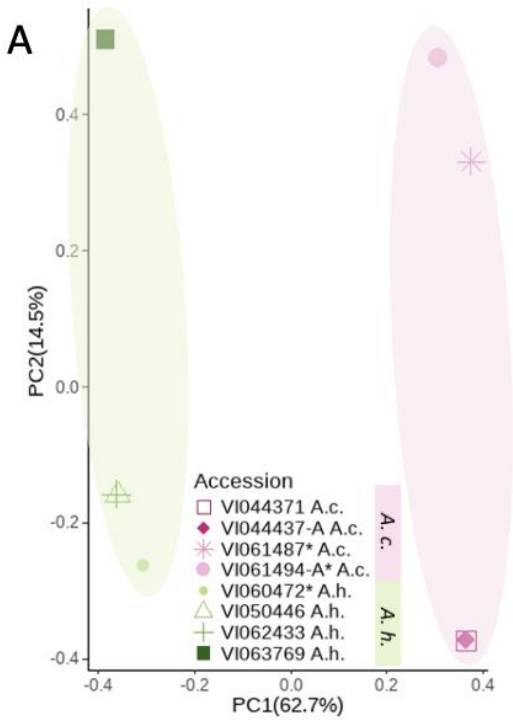



**Figure 4**. **Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome. A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups.

## Additional work

### Metabolite abundance and potential for further annotation and gene discovery

In order to get insight into the distribution of abundance, the metabolite features have been assigned into 7 bins classified per median abundance (**Table 2.1**). This was based on the total amount of features prior checking for redundancy or discarding low abundance peaks (**Supp Dataset 3b**). What could be observed is that the number of features assigned to lower median abundance bins is much higher than the high median abundance bins. While the range of median abundance (Δ abund.) decreased from high to low abundance bins the number of metabolite peaks drastically increased (**Table 2.1**). These proportions were deliberately chosen since low abundance metabolites are typically of less relevance. This is reflected by the ease to annotate high abundance metabolites, i.e., higher 'ann. / peaks %' for high abundance bins (**Table 2.1**). The threshold used to discard low abundance features in the manuscript was based on **Table 2.1**. By excluding bins 6 and 7 from the complete dataset only metabolites with a median abundance of >0.000015 were included.

After performing this cutt-off of low abundance metabolites, as well as only retaining features significantly differentiating between founders, peaks with a high potential for further annotation were investigated. In order to do this, raw data presence absence value (PAV) was plotted for 85 features with the highest median abundances (**Figure 2.1**). If metabolite abundance is entirely lacking in one or several of the founders, it might be a specific gene or gene cluster lacking in this accession. Hence, if genomic information can be linked to absolute metabolite absence, there are enhanced annotation chances for these 85 features. Only one of these 85 PAV features could be annotated with SIRIUS 4 (Dührkop *et al.*, 2019), which was a tannin, absent in founder VI061494-A (**Figure 2.1**). Capsicodendrin is just an example of a very beneficial metabolite due its anticancer properties discovered when extracted from Cinnamosma species, which are endemic and used as traditional medicinal plant in Madagascar (Pan *et al.*, 2016; Karmahapatra *et al.*, 2018). Absence of capsicodendrin in VI061494-A suggests this founder lacks these cancer-inhibiting benefits as well as genes underlaying capsicodendrin production.

To further investigate differences of PAV metabolites between accessions, metabolite abundance has been plotted of three example features. Capsicodendrin is present in all three sample repeats of accessions VI044437-A, VI060472 and VI063769 which are hence expected to have tumour suppressive properties (**Figure 2.2**). The other two highlighted examples are one metabolite (1: M161.0457T88) having extremely high abundance in VI063769, while the other (3: M745.5029T479) is highly abundant in all accessions apart from VI044437-A (**Figure 2.2**). These are nice additional visualizations to get a grasp of the high variability between abundance of PAV metabolites. In order to examine potential species specificity of PAV metabolites 'species difference' as statistically tested in the manuscript is indicated in **Figure 2.1**. Here 63 of the 85 PAV features were significantly different between species, while only 12 were entirely PAV between species, i.e., two absent in *A. hypochondriacus* and ten absent in *A. cruentus* (**Figure 2.1**). This indicates that a large amount of PAV metabolites was statistically differentiating in abundance per species group, while only 12 of those were likely due to absence of genes. This knowledge could lead to identification of species specific genes or biosynthetic gene clusters responsible for the PAV metabolites lacking in certain *Amaranthus* species.

**Table 2.1.** Combined ionization mode non-redundant recoded metabolite peak dataset, split up in bins according to and ordered by median abundance of raw data over the eight parent accessions. Number of peaks increasing per bin with decreasing median abundance. The range of median abundance the raw data of the peaks was recorded at, the difference of the range of that median abundance (Δ abund.), the number of peaks in the bin, those of which significantly differentiating between parents (sig. par.), those significantly differentiating between species (sig. spec.), raw data presence absence values (PAV) including those significantly differentiating between MAGIC parents (sig. PAV), annotated metabolites significantly differentiating among parents (ann.) and percentage of those divided by peaks per bin are given. Annotations performed with SIRIUS 4 (Dührkop et al., 2019) only.

| bin | range median abundance | Δ abund. | peaks | sig. par. | sig. spec. | PAV | sig. PAV | ann. | ann. / peaks % |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00500000 - 0.13999999 | 0.135 | 44 | 40 | 27 | 0 | 0 | 17 | 38.64 |
| 2 | 0.00150000 - 0.00499999 | 0.0035 | 61 | 54 | 39 | 2 | 2 | 16 | 26.23 |
| 3 | 0.00050000 - 0.00149999 | 0.001 | 148 | 133 | 86 | 6 | 6 | 33 | 22.30 |
| 4 | 0.000100000 - 0.00049999 | 0.0004 | 409 | 358 | 243 | 19 | 19 | 41 | 10.02 |
| 5 | 0.000015000 - 0.000099999 | 0.000085 | 762 | 655 | 453 | 63 | 58 | 29 | 3.81 |
| 6 | 0.000002500 - 0.000014999 | 0.0000125 | 849 | 712 | 470 | 193 | 186 | 24 | 2.83 |
| 7 | 0.000000000 - 0.000002499 | 0.0000025 | 986 | 866 | 557 | 477 | 459 | 12 | 1.22 |
| **Tot** | 0.000000000 - 0.13999999 | 0.1399999 | 3259 | 2818 | 1875 | 760 | 730 | 172 | 5.28 |

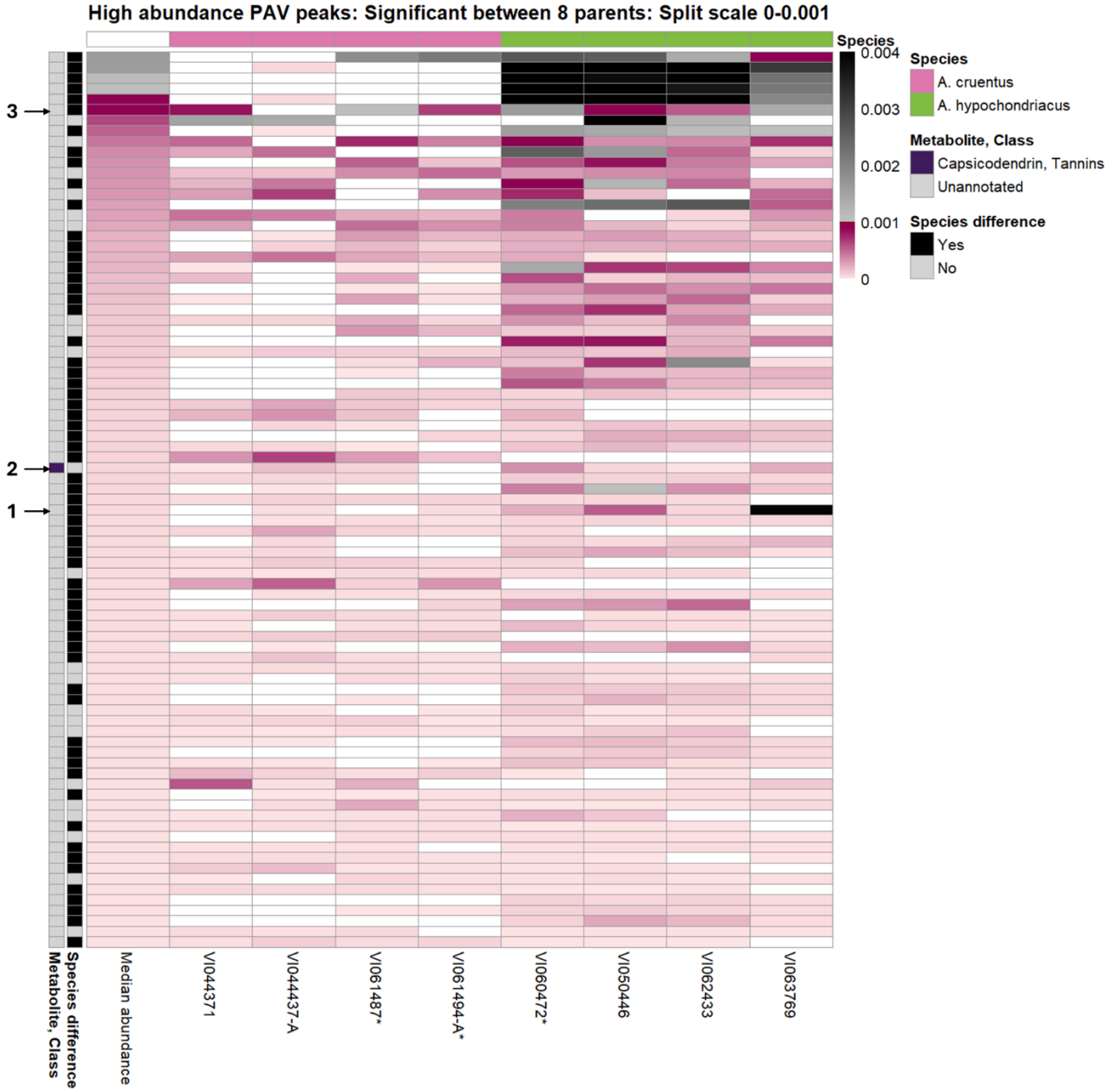



**Figure 2.1** Raw metabolite value heatmap displaying 85 most abundant presence absence values (PAV) ordered vertically by median abundance over eight parent accessions, and horizontally by species. Absence of metabolite abundance is displayed as white. Low median abundance levels of 0-0.001 are displayed light pink to dark purple. Higher median abundance levels of >0.0001 grey to black. Annotated metabolites and significance between species by multiple testing corrected t test indicated. '*' behind accession ID's indicates lines with admixture. Three arrows indicated are metabolites with their raw values displayed in **Figure 2.2**.

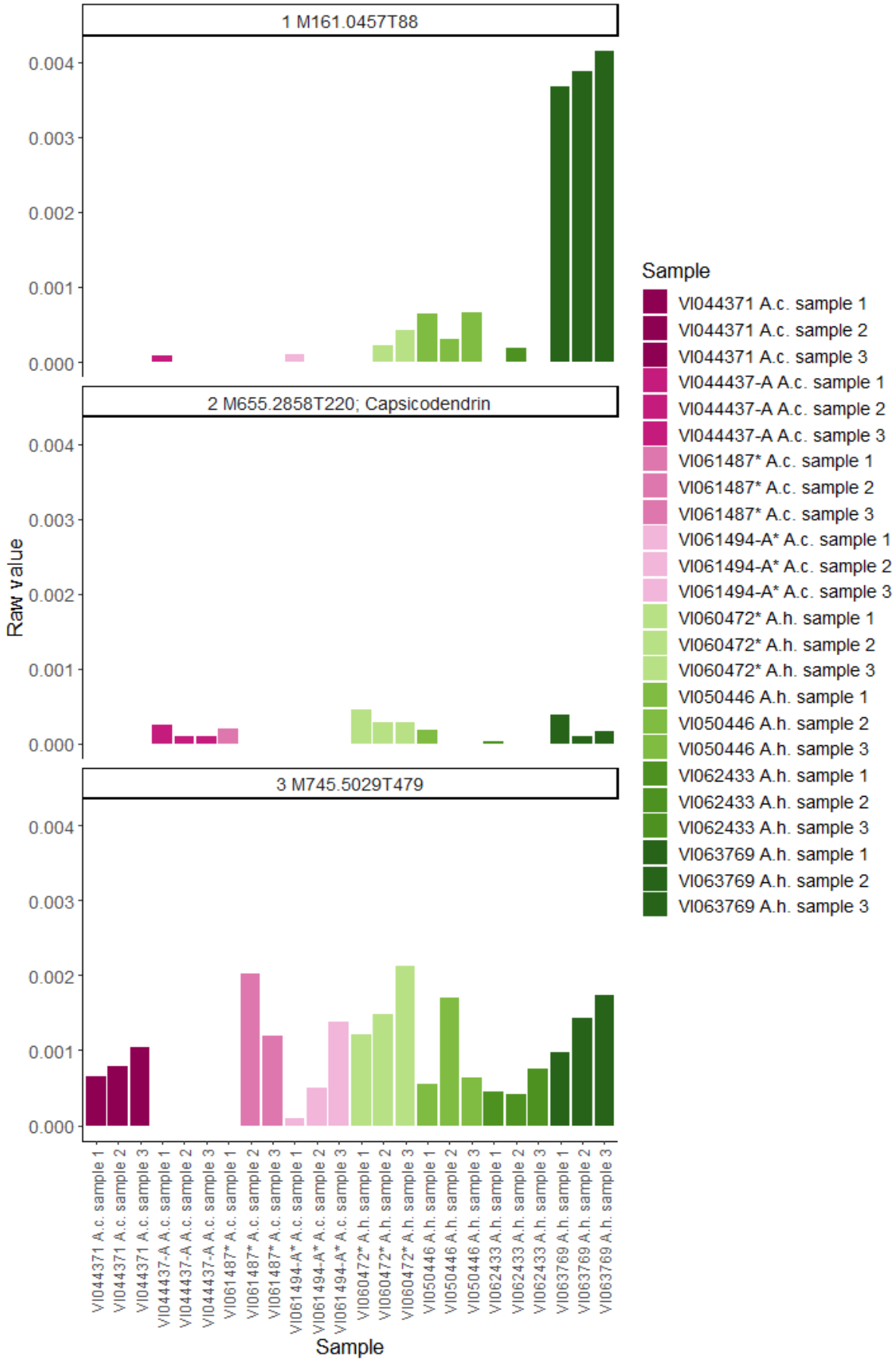



**Figure 2.2** Varying raw metabolite abundance values of three example features highlighted in (**Figure 2.1**). These are all presence absence value (PAV) metabolites and raw values are given per three repeat sample of each founder accession. Here sample 1 is leaf tissue A and samples 2 and 3 repeated leaf tissues B1 and B2.

## Initial genome alignment reveals species specific mapping regions

**Table 2.2** Median coverage of eight MAGIC parent accessions aligned to exonic and non-exonic regions of the complete genome as well as main reference chromosomes only (Excl. scaffold) of the *A. cruentus* reference (Ma et al., 2021). The mapping percentage is also provided. Accessions are grouped by species and '*' behind the accession ID indicates admixture.

| Accessions | | Median coverage | | | | % Mapped to reference |
|---|---|---|---|---|---|---|
| | | Complete reference | | Excl. scaffold | | |
| Species assumed | ID | Exon | Non-exon | Exon | Non-exon | |
| *A. cruentus* | VI044371 | 25.49 | 25.61 | 25.50 | 25.64 | 99.85 |
| | VI044437-A | 23.74 | 23.80 | 23.74 | 23.81 | 99.82 |
| | VI061487* | 27.69 | 27.73 | 27.69 | 27.75 | 99.83 |
| | VI061494-A* | 20.77 | 20.79 | 20.77 | 20.80 | 99.87 |
| *A. hypochondriacus* | VI060472* | 24.53 | 24.51 | 24.55 | 24.57 | 99.79 |
| | VI050446 | 24.44 | 24.41 | 24.48 | 24.47 | 99.81 |
| | VI062433 | 31.81 | 31.86 | 31.82 | 31.90 | 99.72 |
| | VI063769 | 25.63 | 25.66 | 25.64 | 25.68 | 99.85 |
| **Mean:** | | 25.51 | 25.55 | 25.52 | 25.58 | 99.82 |

An initial attempt to align the founder genomes to the *A. cruentus* reference genome (Ma *et al.*, 2021) was performed using Burrows-Wheeler Aligner (BWA) v.0.7.17-foss-2018b (Li and Durbin, 2010). To investigate if alignment was performed adequately the mapping coverage was assessed. The mean mapping coverage over all accessions ranged 25.51 to 25.58 and the mean mapping percentage was 99.82% with at least >99.70% for all accessions (**Table 2.2**, **Figure 2.3**). This indicates sufficient alignment of all founder accessions which were sequenced with ~12-17x coverage. The (non-)exonic coverage of accession VI062433 is high (>31) in comparison to the other accessions (**Table 2.2**, **Figure 2.3**). This is likely due to the higher sequencing depth of this founder (**Table 2.3**). Generally lower or higher (non-)exonic coverages of accessions coincided with either lower or higher sequencing depths or lower or higher relatedness to the *A. cruentus* reference (**Table 2.2**, **Figure 2.3**, **Table 2.3**). This is as expected and another initial indication for adequate alignment. The differences between (non-)exonic mapping coverages is marginal, with often a higher coverage recorded for mapping to non-exonic versus exonic regions of the reference genome (**Table 2.2**, **Figure 2.3**). This is against expectations since usually more coding genes are present in exons typically easing genome alignment (Bourque *et al.*, 2018; Colonna Romano and Fanti, 2022). This could suggest that something is off while performing genome alignment with BWA, although the differences between coverage towards non- versus exonic regions are minimal.

**Table 2.3** File sizes in gigabytes of zipped FASTA files used for, and Binary Alignment Map (BAM) files created after, alignment to an *A. cruentus* reference genome (Ma *et al.*, 2021) with BWA v.0.7.17-foss-2018b (Li and Durbin, 2010).

| Accessions | | File size (Gb) | |
|---|---|---|---|
| Species | ID | FASTA | BAM |
| *A. c.* | VI044371 | 8.3 | 13.2 |
| | VI044437-A | 7.7 | 12.3 |
| | VI061487* | 9.2 | 16.2 |
| | VI061494-A* | 7.3 | 14.1 |
| *A. h.* | VI060472* | 8.1 | 17.7 |
| | VI050446 | 8.3 | 19.1 |
| | VI062433 | 11 | 24.6 |
| | VI063769 | 8.5 | 19.1 |

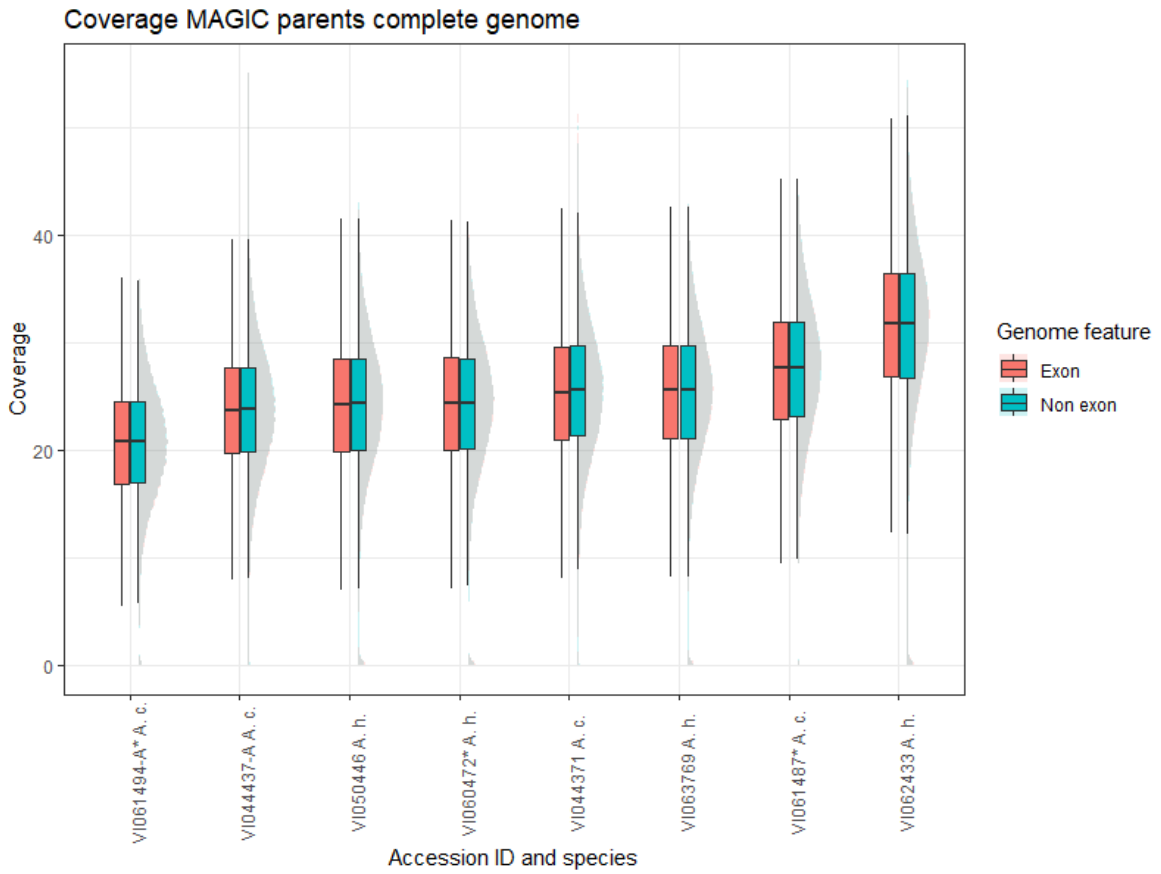



**Figure 2.3** Boxplots with histograms at the right side displaying coverage per MAGIC parent accession aligned to an *A. cruentus* reference genome (Ma *et al.*, 2021) using BWA v.0.7.17-foss-2018b (Li and Durbin, 2010). Coverage is per exon or non-exon region and ordered by median of the exon coverage. Species are given behind accession ID's with '*' indicating admixture.

The mapping coverage towards the individual chromosomes of the reference genome has also been examined. A similar to slightly lower coverage is observed when aligning to the complete genome compared to main chromosomes of the reference only (**Table 2.2**). This is expected since scaffolds often do not contain much coding genes but repetitive, non-complex regions instead, resulting in fewer uniquely mapping reads causing low coverage. The median coverage over all eight accessions had normally distributed density plots per chromosome (**Figure 2.4**) as well as always some coverage along all reference chromosomes (**Figure 2.5**). This indicates alignment occurred with equally distributed coverage and was performed along all reference chromosomes. The coverage density peaks plotted per founder showed that the *A. hypochondriacus* species accessions have lower and wider distributions compared to *A. cruentus* accessions (**Figure 2.6**). This is likely due to less related species or accessions covering the *A. cruentus* reference less consistently compared to more related individuals. When visualizing the coverage per accession over the reference chromosomes it could indeed be observed that parts of the chromosome we not or only marginally covered by *A. hypochondriacus* accessions (**Figure 2.7**). This was a first indication that some parts of the reference genome are conserved among *A. cruentus* species while not present in *A. hypochondriacus* genomes.

Ideally alignment is performed towards the complete reference genome to include all available information. This initial genome alignment performed with BWA v.0.7.17-foss-2018b (Li and Durbin, 2010) showed that mapping with sufficient and stable coverage towards the *A. cruentus* reference (Ma *et al.*, 2021) is possible, and that conserved species specific genomic regions can be identified. However, with the development of more

advanced genome alignment tools, MUMmer v.4.0.0 (Marçais *et al.*, 2018) has eventually be chosen for application prior variant calling to create the *Amaranthus* pan-genome described in the manuscript. This because MUMmer is developed more recently, the alignment speed is comparable and sometimes faster than BWA and because BWA is designed primarily for short Illumina sequences while we are dealing with PacBio HiFi long-reads (Marçais *et al.*, 2018).

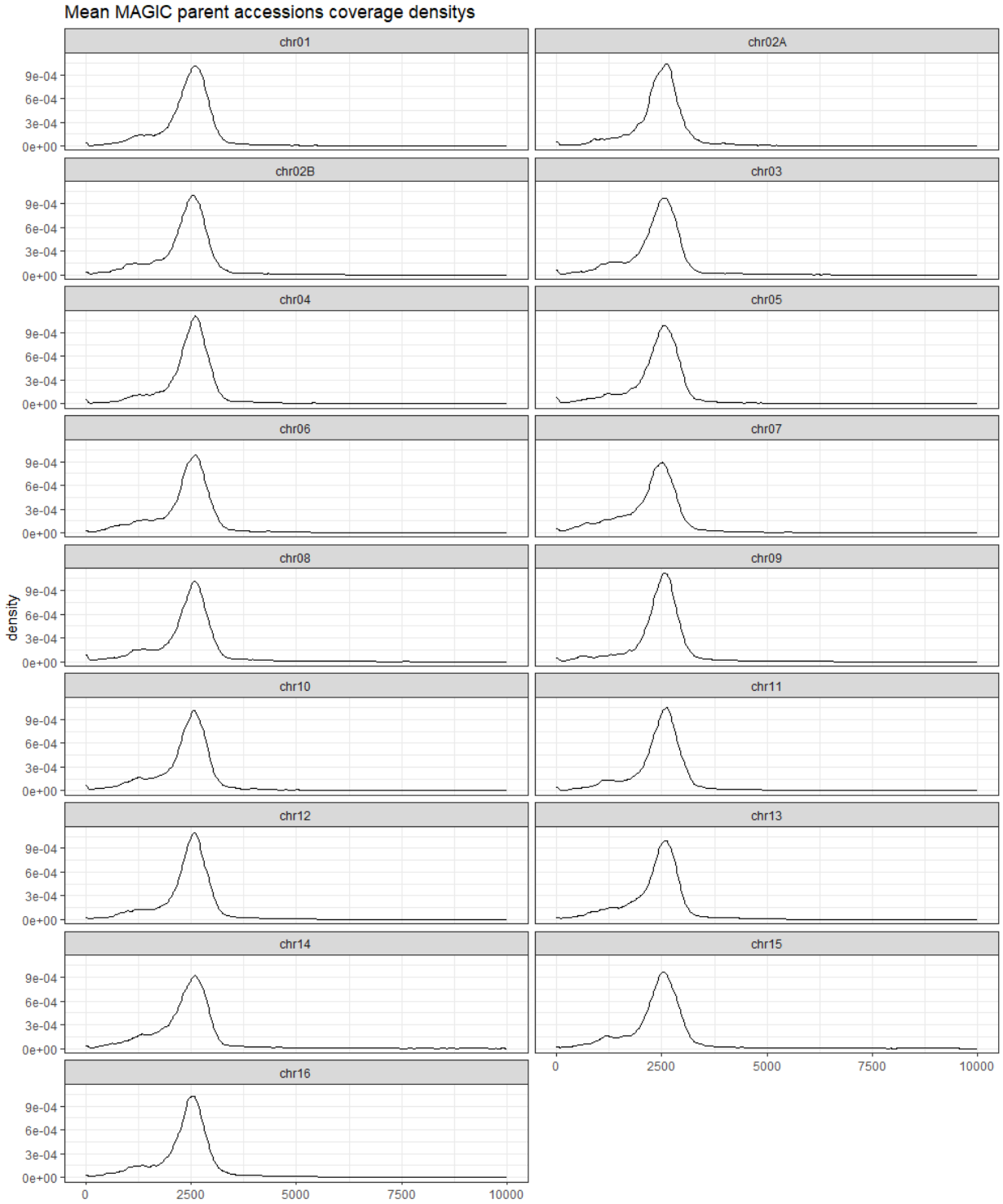



**Figure 2.4** Density plots displaying density of mean coverage (x-axis) per 100 base pair windows of eight MAGIC parent accessions per 17 main chromosomes of the *A. cruentus* reference genome (Ma *et al.*, 2021), used to align the parent reads to.

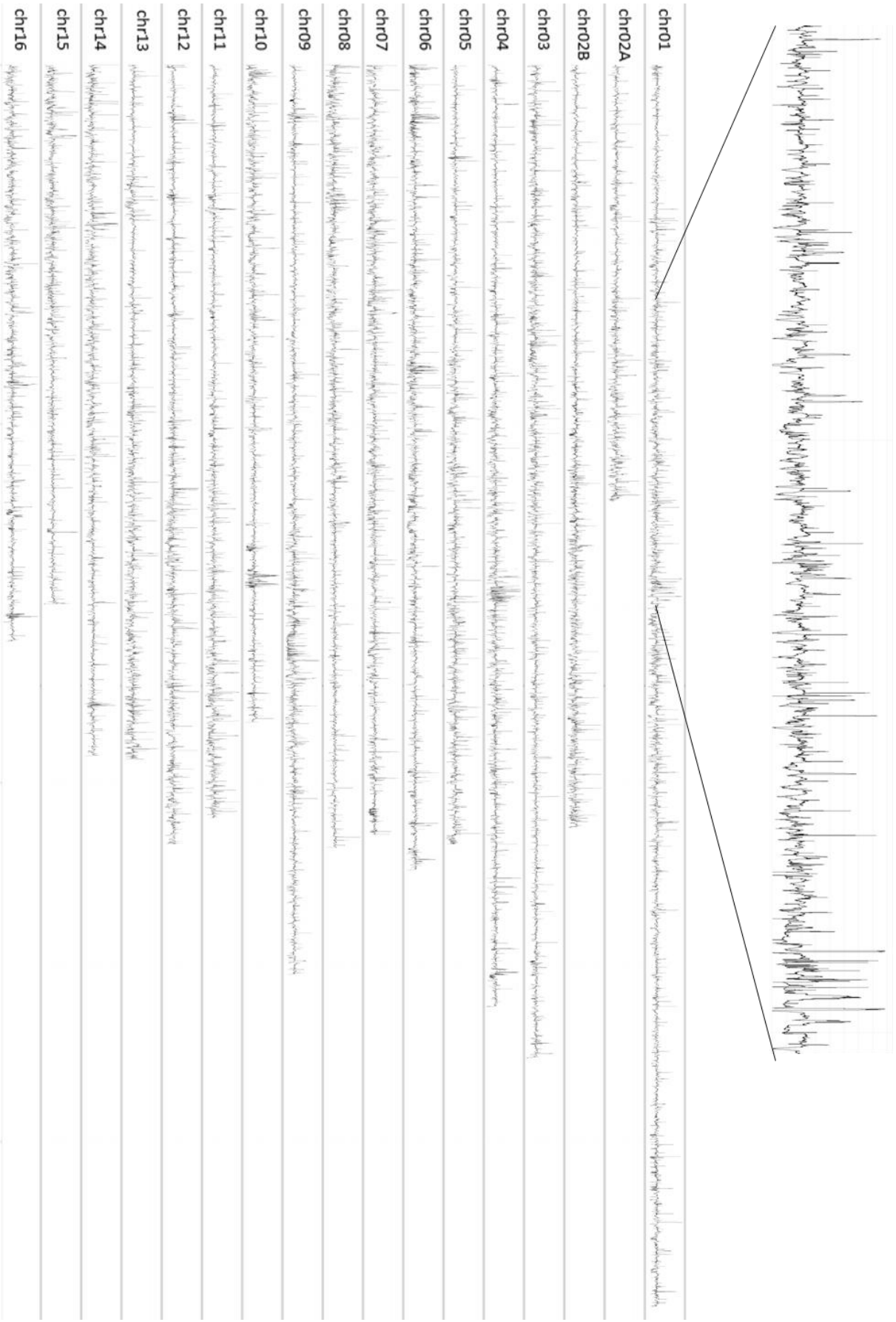



**Figure 2.5.** Mean coverage (y-axis) of MAGIC parent reads per 100 base pairs plotted per main *A. cruentus* reference chromosome. Coverage is ranging 0-10000.

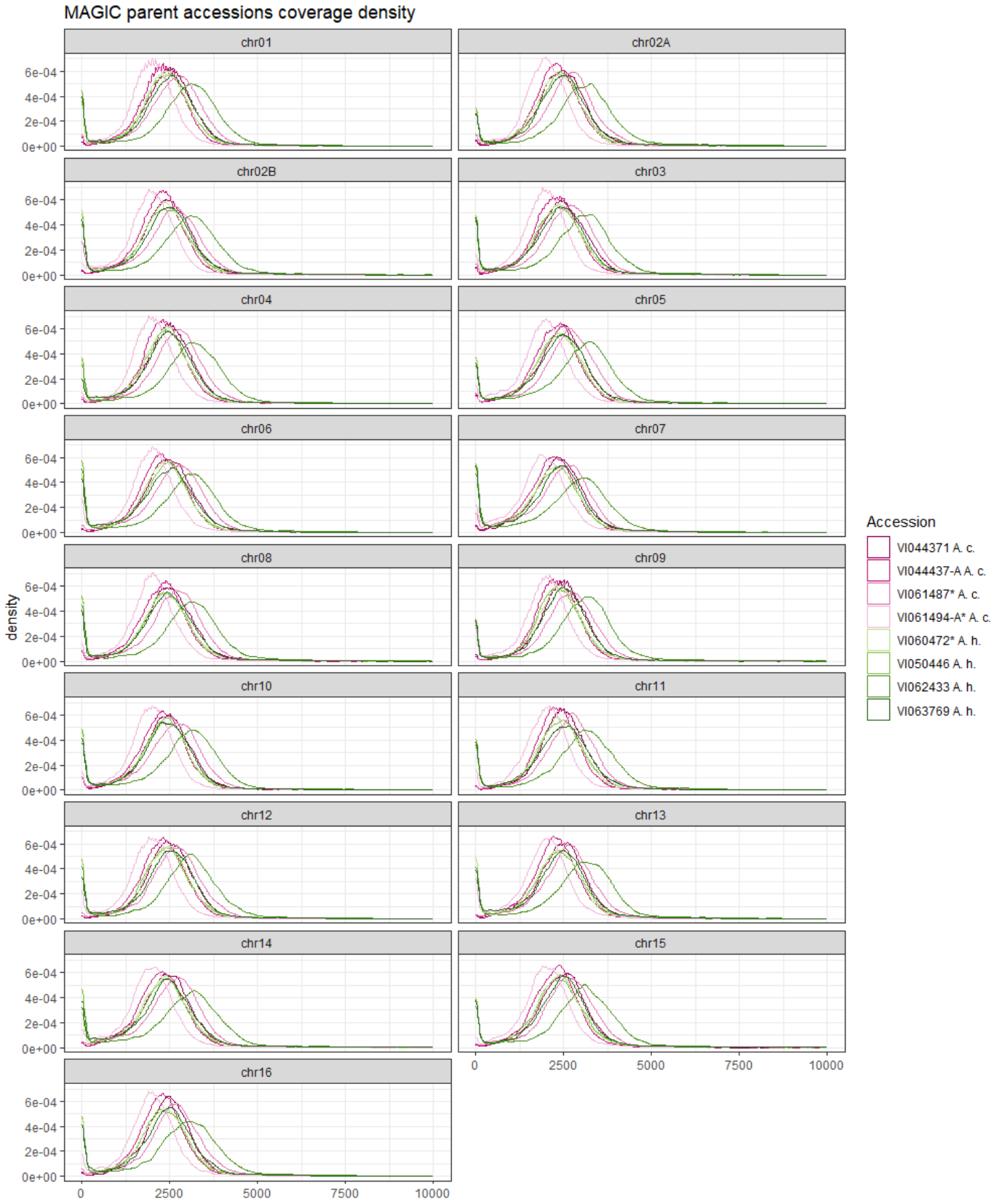



**Figure 2.6** Density plots displaying density of certain coverages (x-axis) of eight MAGIC parent accessions per 100 base pair windows per 17 main chromosomes of the *A. cruentus* reference genome (Ma et al., 2021) used to align the parent reads to. '*' behind accession ID indicates presence of species admixture.

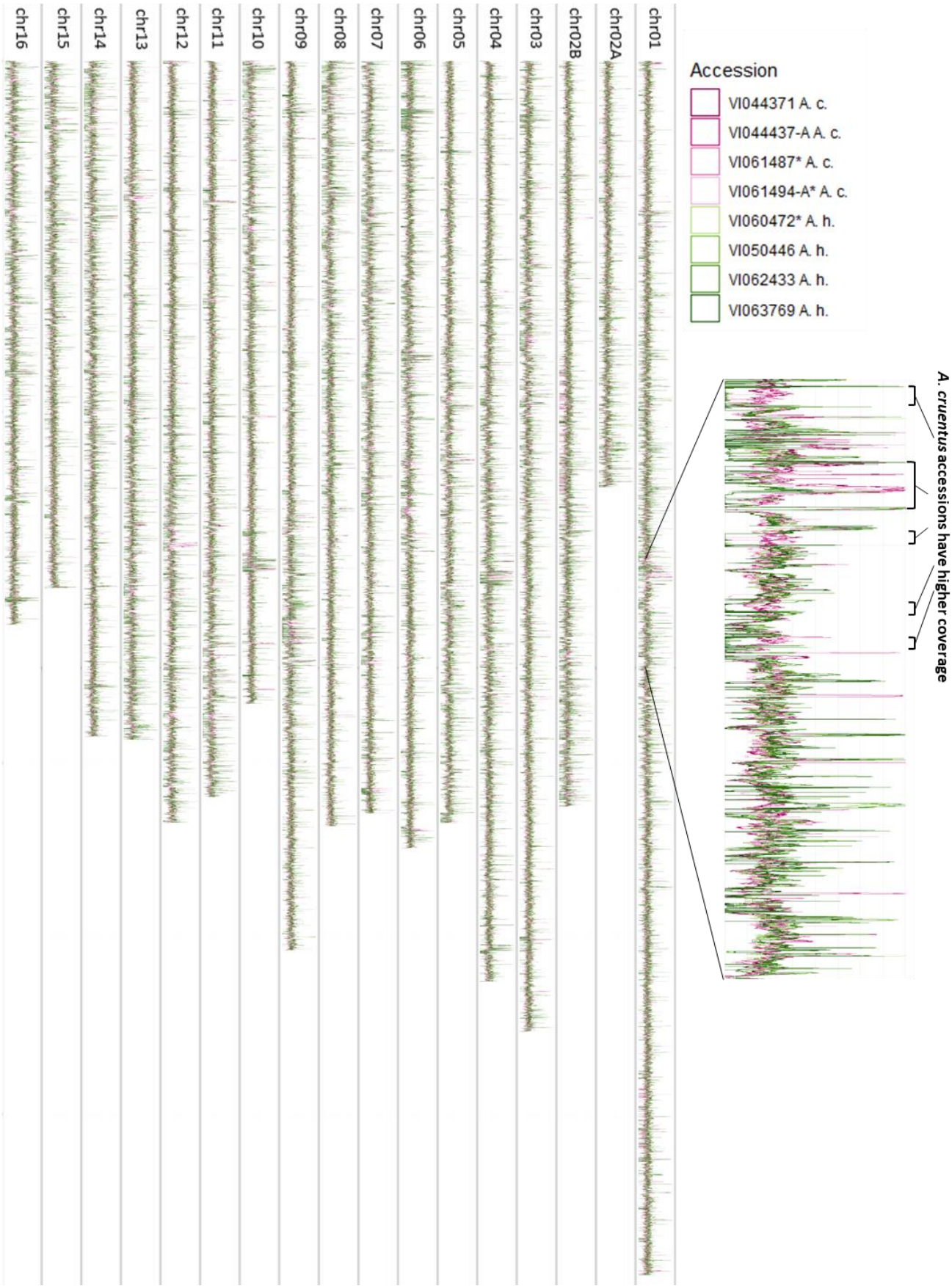



**Figure 2.7** Coverage (y-axis) of MAGIC parent reads per 100 base pairs plotted per main *A. cruentus* reference chromosome. Coverage is ranges 0-10000 and '*' behind accession ID indicates presence of species admixture.

## Further investigation of the MAGIC founder genomes and GO term enrichment

After performing genome alignment with MUMmer v.4.0.0 (Marçais *et al.*, 2018) and SNP calling with Longshot (Edge and Bansal, 2019) more phylogenetic analysis was performed with the MAGIC founder genomes. In order to examine the degree of polymorphism among the two subsets of species of the MAGIC founders the nucleotide diversity (π) has been investigated. The *A. hypochondriacus* has significantly (p-value < 2e-16) more genetic diversity (median π = 0.00185) than *A. cruentus* accessions (median π = 0.00175) (**Figure 2.8 A**). This higher level of genetic variation within the subset of *A. hypochondriacus* accessions indicates that these founders are likely evolutionary less related to each other than the selected *A. cruentus* accessions are. Another way to assess the genetic relationships between the founder accessions is creation of a phylogenetic tree. A midpoint rooted tree revealed a clear separation between species, and species admixture accessions clustering together in the middle (**Figure 2.8 B**). This tree illustrates the phylogenetic relationships and evolutionary distances among the founders, as well as the *Amaranthus* species groups to which they belong. To visualize which genes are shared versus varying among the pan-genome genomes, a presence absence (PAV) heatmap has been created with the gene families identified with OrthoFinder v.2.3.3 (Emms and Kelly, 2019). Many orthogroups are conserved among species and a large chunk of genes identified in nine of the genomes is missing in the *A. hypochondriacus* reference (**Figure 2.8 C**). This shows that some gene families are species-specific and provides additional evidence that the reference genome of (Lightfoot *et al.*, 2017) is likely of lower quality.

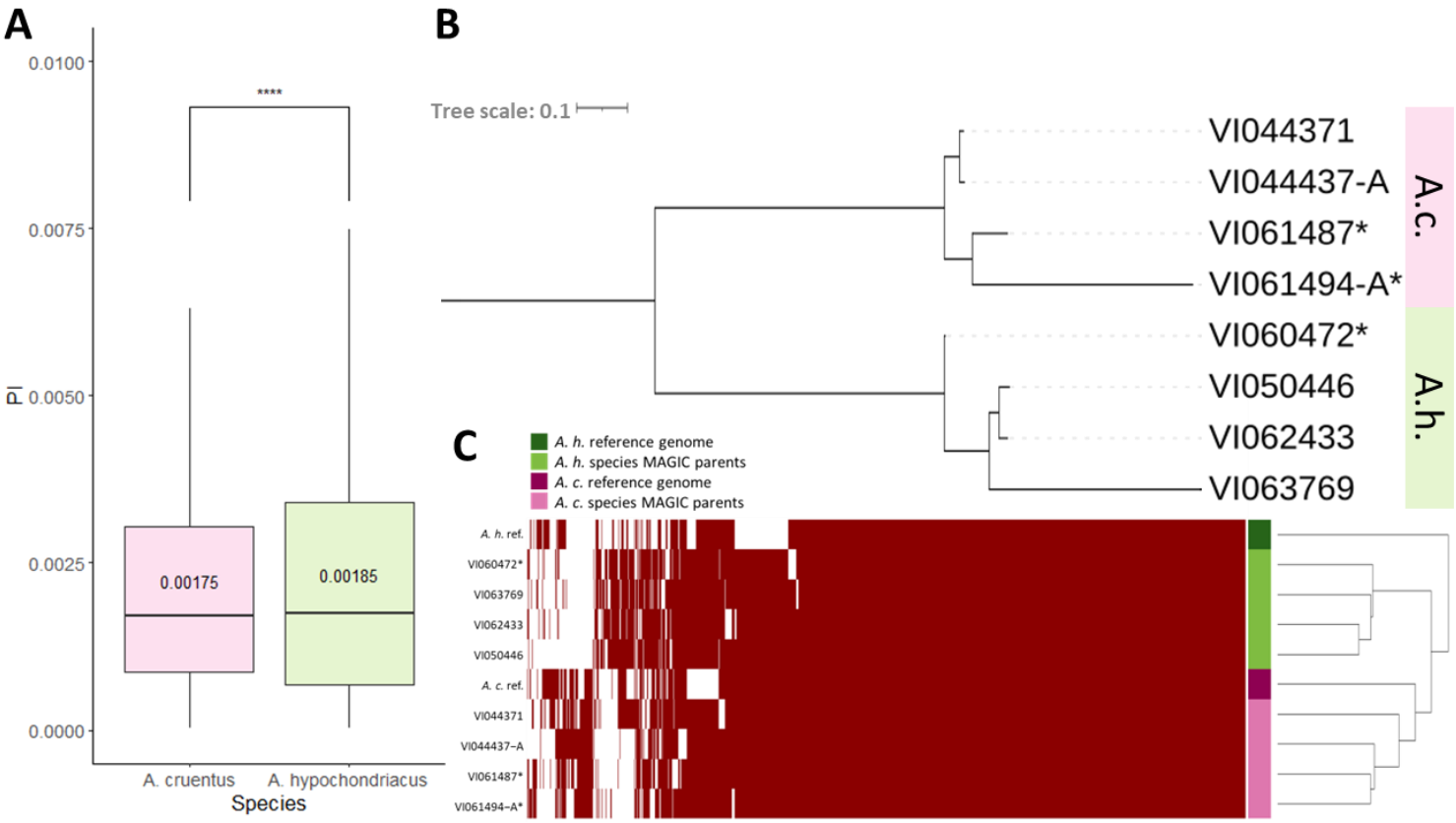



**Figure 2.8 A)** Nucleotide diversity (π) per species groups considered as separate populations of MAGIC founders. A t-test was used for significance between species populations and median π per species group are provided. **B)** Midpoint rooted phylogenetic tree of the revealing relationships between the MAGIC founders based on SNP calling towards *A. cruentus* Arusha (Ma et al., 2021). '*' behind accession ID indicates admixture. **C)** Presence (red)/absence (white) of gene families identified using OrthoFinder v.2.3.3 (Emms and Kelly, 2019) with accessions arranged according to species.

In order to investigate differences between functional annotation of genes of certain pan-genome components, the GO term enrichment results with pan- components ignoring the *A. hypochondriacus* reference (**Figure 2.9**) has been assessed more in depth. About half (55.40%) of the orthogroups and genes could be annotated with a GO-term, with clearly more GO-term annotations for core (68.28%) than by dispensable orthogroups (46.11%) with the least for private genes (21.75%). This is additional proof that there is higher functional conservation of core genes among amaranths. The top ten most significantly (i.e., lowest p-value) enriched GO-terms for the core and dispensable genome are visualised in **Figure 2.10 A** and **B**, respectively. Results have been discussed in the manuscript with all data in **Supp Dataset 7**, showing that there were no pronounced differences in gene function between the two pan- components. However, just as for many other recent plant pan-genomes (Liu et al., 2020; Song et al., 2020; Li et al., 2022; Yan et al., 2023), genes with more essential biological functions were enriched in the core genome, while more general biological processes like plant adaptation functions were enriched in dispensable genes (**Figure 2.10**). To investigate if some MAGIC founders have unique properties, GO term enrichment was performed with genes unique for one founder accession considered as a MAGIC private genome component (**Figure 2.9**). The top ten most significantly enriched MAGIC private GO-terms were filtered for an expected count of >1 (**Supp table GO enrichment**), with heme binding, response to oxidative stress and to auxin being the most evident (**Figure 2.10 C**). Most of these overlap with enrichment of plant adaptation roles like in dispensable genes. It is expected that these biological processes will further segregate in the MAGIC offspring likely making these novel created accessions developmentally adaptive to changing climatic conditions. Especially the response to auxin being among the top three most enriched MAGIC private GO-terms can be beneficial, since auxin is an important plant hormone involved in growth and plant development being crucial for survival in dynamic environments (Teale, Paponov and Palme, 2006; Mishra, Sharma and Laxmi, 2022).

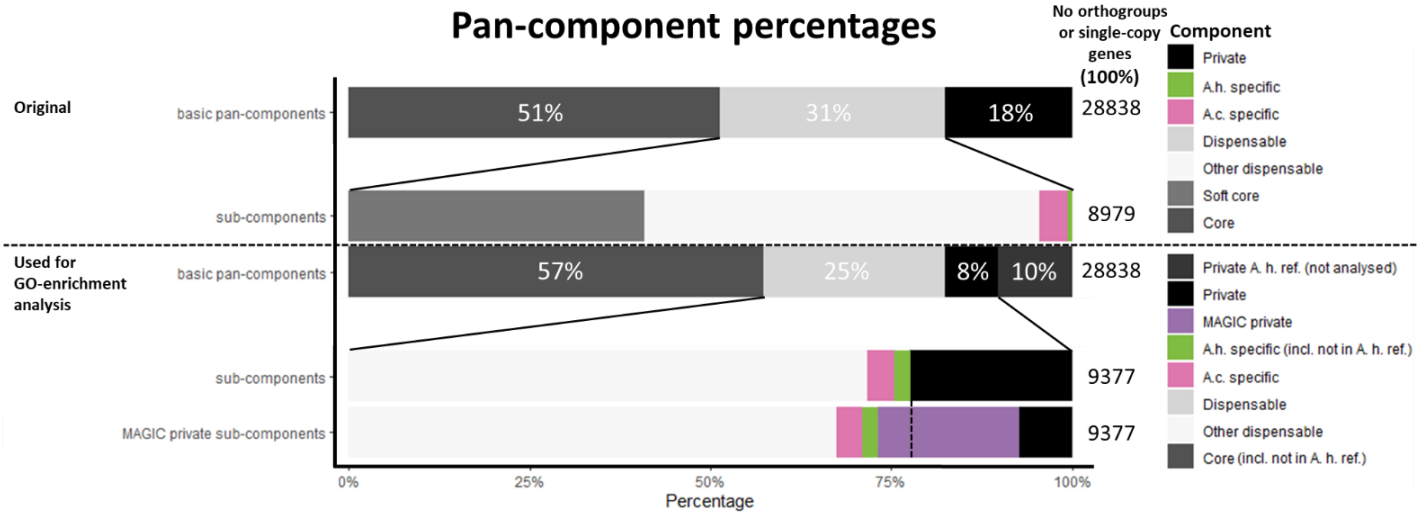



**Figure 2.9** Original pan- components (top) and those ignoring identification including *A. h.* reference genes used for GO-enrichment analysis (bottom). Soft core genes are those absent in one accession, MAGIC private those unique for one founder accession. Number of (No) genes (orthogroups or single copy) is provided per bar graph.

In an attempt to highlight gene functions unique to species, GO term enrichment was performed with species specific subsets of the dispensable genome. Most of the *A. c.* (**Figure 2.11 A**) and *A. h.* (**Figure 2.11 B**) specific pan-genome components have unique enriched biological functions, however, as expected, some overlap with the top ten enrichment of the dispensable genome the species specific is pulled out of, like nucleic acid and ADP binding, or DNA integration (**Figure 2.10 B**). Note that apart from two *A. c.* specific, all species specific enriched GO-terms have a low expected gene count (<1) (**Supp table**

**GO enrichment**). This indicates that enrichment results for the species specific pan-genome are less obvious, with likely only the nucleic acid and guanyl nucleotide binding functional genes being notably enriched in *A. c.* accessions. GO-term enrichment of the private genome has also been performed in order to investigate functional genes unique to accessions (**Supp Dataset 7**). Nine out of the top ten GO-terms found to be significantly enriched in the private genes (**Figure 2.11 C**) overlapped with the top ten MAGIC private enriched GO-terms (**Figure 2.10 C**), both filtered for an expected gene count of >1. This is likely due to many genes in these components being the same since only the *A. c.* Arusha reference private genes are added to the MAGIC private genome in order to create the private genome subset.

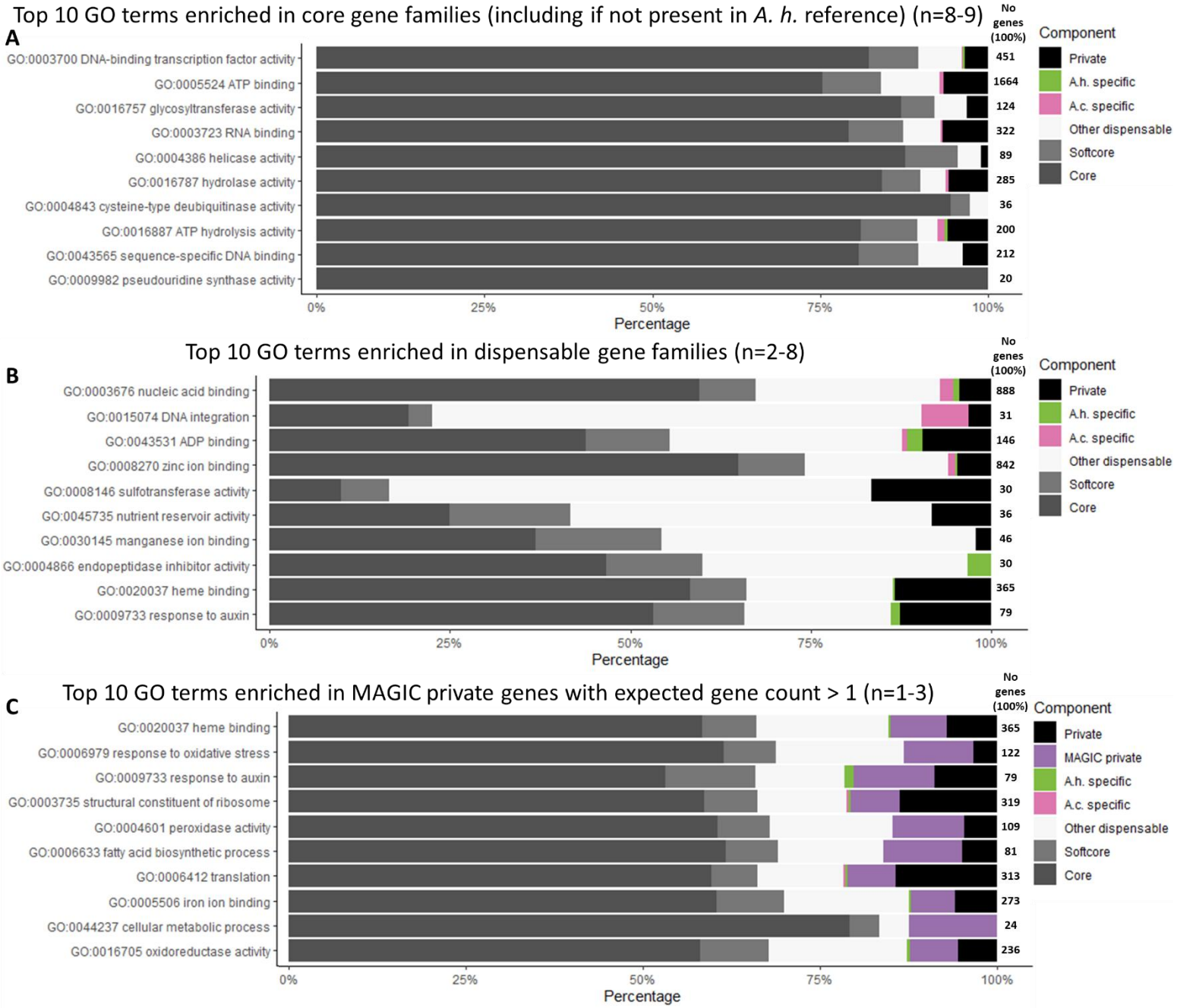



**Figure 2.10** Top ten most significantly (top to bottom) enriched GO-terms in the **A)** core, **B)** dispensable and **C)** MAGIC private (expected gene count of >1) components of the *Amaranthus* pan-genome. Component proportions are displayed as percentage and the total (100 %) number of (No) genes are provided at right side next to the proportional bars. *A. c.* and *A. h.* species specific are pulled out from the dispensable component.

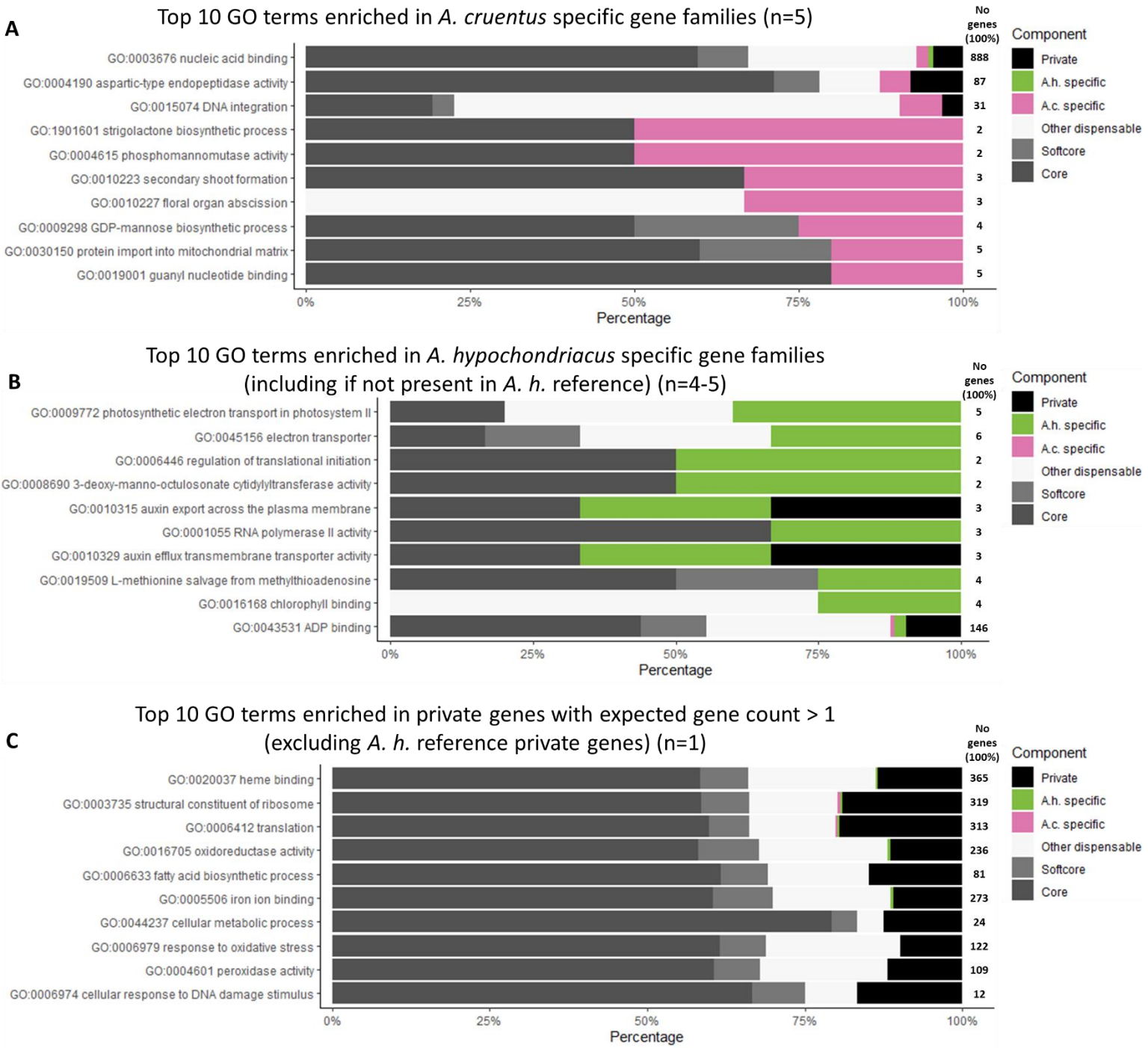



**Figure 2.11** Top 10 most significantly (top to bottom) enriched GO-terms in the **A)** A. c. or **B)** A. h. species specific parts pulled out of the dispensable *Amaranthus* pan-genome and **C)** the private genes with an expected gene count of > 1. Pan- component proportions are displayed as percentage. Total (100 %) number of (No) genes are provided at right side next to the pan-genome component proportional bars. *A. c.* and *A. h.* species specific are pulled out from the dispensable component.

# Discussion

For the discussion regarding the work in the manuscript please refer to the appendix. This section discusses additional work reflecting examination of available metabolomics and genomics data which didn't fit into the limitations of the submitted manuscript described in the appendix. It reveals an insight in the distribution of abundance of all recorded metabolite features in the MAGIC founders (**Table 2.1**). Moreover, it highlights metabolites with high

abundance but absent in one or more of the parental lines, which therefore have a high potential for further function annotation and underlaying gene or gene cluster discovery (**Figure 2.1**). An example of an annotated presence absence metabolite is capsicodendrin, which has beneficial cancer-inhibiting properties (Pan *et al.*, 2016; Karmahapatra *et al.*, 2018). Capsicodendrin present in all three sample repeats of accessions VI044437-A, VI060472 and VI063769 which are hence expected to have these cancer suppressive benefits, while founder VI061494-A likely lacks it (**Figure 2.2**). A next step would be looking into which metabolites in **Figure 2.1** can be annotated and what their relevance is to human nutrition is so founder specific dietary properties can be identified. Ideally genes or gene clusters responsible for production of certain metabolites or entire metabolic pathways could be revealed which might be fully lacking in some of the founder genomes.

That the founder genomes are unique, with differing coverages when aligned to the *A. cruentus* reference genome (Ma *et al.*, 2021) was shown after initial genome alignment using BWA v.0.7.17-foss-2018b (Li and Durbin, 2010). Results showed sufficient alignment of all founder accessions is possible towards to the *A. cruentus* reference (Ma *et al.*, 2021), and that lower or higher coverages of accessions coincided with either lower or higher sequencing depths or lower or higher relatedness to the reference (**Table 2.2**, **Figure 2.3**, **Table 2.3**). Most notably was the species specific alignment, showing that parts of the *A. cruentus* chromosomes were not or only marginally covered by *A. hypochondriacus* accessions, indicating some chromosome parts are conserved among *A. cruentus* accessions only (**Figure 2.7**). A controversial finding was the often higher coverage at non-exonic versus exonic reference regions (**Table 2.2**, **Figure 2.3**), while the gene information is exons is expected to result in higher coverage results. This is an indication that these genome alignment results are not optimal despite the sufficient mapping coverages. Hence, further analysis in the manuscript was performed with the more advanced and recently developed genome alignment tool MUMmer v.4.0.0, which is more suitable for aligning PacBio HiFi long-reads (Marçais *et al.*, 2018).

After SNP calling with Longshot (Edge and Bansal, 2019) and performing phylogenetic analysis, additional results further confirmed findings of the manuscript. Like correct species assignment and relatedness among the MAGIC founders (**Figure 2.8**). A new finding was that the *A. hypochondriacus* parental accessions are likely less related to each other than the *A. cruentus* accessions (**Figure 2.8 A**), while the lower quality of the *A. hypochondriacus* reference was only further confirmed (**Figure 2.8 C**). **Figure 2.9**, **Figure 2.10** and **Figure 2.11** provide additional visualization of the GO term enrichment results described in the manuscript with data provided in **Supp Dataset 7**. The MAGIC private enriched GO-terms were not previously mentioned (**Figure 2.10 C**, **Supp table GO enrichment**). It reveals that some MAGIC founders contain specific developmental characteristics, which likely will segregate in the MAGIC offspring making these accessions potentially climatically adaptive. The attempt to identify species specific enriched GO terms was also novel (**Figure 2.11 A**, **B, Supp table GO enrichment**). However, due to low expected gene counts these results are less promising for adequate interpretation which suggest that almost no gene functions are unique for a certain species.

# Methods

## Additional metabolite profiling and analysis

The methods to generate and analyse the metabolite features are elaborately described in the manuscript in the appendix, especially in the Supplementary Methods: Detailed Metabolite Profiling methodology. Additional analysis performed which were not previously explained are described here. All recorded metabolites (**Supp Dataset 3b**) prior removing redundant features or discarding low abundance peaks were binned into 7 categories defined by median abundance over raw data values. Thresholds used to discard low abundance features in the manuscript was based on this binned metabolite table by excluding the $6^{th}$ and $7^{th}$ bin. This resulted in a cutt-off of low abundance (>0.000015) metabolites. Presence absence values (PAV) are defined by the raw data replicate average being zero for at least one or more accessions. For the 85 most highly abundant PAV recorded metabolite peaks, the raw data was plotted in a heatmap, ordered by median abundance, using the pheatmap (Kolde, 2019) R package. This was done after performing the >0.000015 median abundance cutt-off and while only keeping significantly differentiating (ANOVA Tukey's family-wise error correction) features between accessions. In the heatmap the 'zero' values are made white, low median abundance levels pink to purple and median abundance higher than 0.001 visualised by colouring grey to black. Of these 85 highly abundant PAV metabolites, three example features were isolated and raw data values of all accessions and their repeats were plotted with bar graphs using R package ggplot2 (Wickham, 2016). All metabolite graphs were made with R v.4.2.3 (R Core Team, 2021).

## Initial genome alignment with BWA and coverage assessment

FASTA files of the eight parent accessions were zipped using the 'gzip' linux command. BAM files were created using Burrows-Wheeler Aligner (BWA) BWA v.0.7.17-foss-2018b (Li and Durbin, 2010) for genome aligning to the *A. cruentus* reference genome (Ma *et al.*, 2021). Percentage of reads in the BAM file that mapped the *A. cruentus* reference was checked with the command 'samtools flagstat' of SAMtools v.1.15-GCC-11.2.0 (Danecek *et al.*, 2021). From the *A. cruentus* reference GFF file (https://bioinformatics.psb.ugent.be/gdb/amaranthus/), a BED file containing exonic regions were created using the Linux awk command searching for 'exon'. A BED file containing non-exonic regions was made using the command 'bedtools complement' of BEDTools v.2.27.1-foss-2018b (Quinlan and Hall, 2010). These (non-)exonic BED files were produced for the complete reference genome as well as for the genome excluding scaffolds (chromosomes only). Additionally, a BED file was made using a file with chromosome sizes excluding the scaffolds, with windows of 100 base pairs (bp), using the 'bedtools makewindows' command of BEDTools v.2.27.1-foss-2018b (Quinlan and Hall, 2010). Sequence coverage of the BAM files of the eight accessions was calculated for all five types of BED files (2x exonic, 2x non-exonic and 100 bp window) using SAMtools v.1.15-GCC-11.2.0 (Danecek *et al.*, 2021) command 'samtools bedcov'.

Median coverage of all accessions was calculated per exon or non-exon length, per complete reference genome or excluding scaffolds. Of these four calculated options the mean coverage is taken over all eight accessions. It was checked whether sequencing depth approximately did match coverage results, i.e., FASTA or BAM file sizes in- or decreasing

with coverage or with allocated species. Boxplots with median coverage per length per accession, with distributions at the side, were made for the complete reference genome using R version 4.1.1. (R Core Team, 2021) package ggdist (Kay, 2022). Density plots of the coverage per 100 bp length is plotted per 17 main *A. cruentus* reference chromosomes as well as coverage over the length of each of the 17 chromosomes using R package ggplot2 (Wickham, 2016). This was done for the mean coverage over the eight MAGIC founders as well as for the coverage of the eight parent accessions separately.

## Additional MAGIC founder genome investigation and GO term enrichment

Methods for genome alignment with MUMmer v.4.0.0 (Marçais *et al.*, 2018), SNP calling with Longshot (Edge and Bansal, 2019), orthologous gene family identification with OrthoFinder v.2.3.3 (Emms and Kelly, 2019) and GO-enrichment analysis with topGO (Alexa and Rahnenfuhrer, 2023) have been previously described in the manuscript. The Longshot SNP VCF was filtered for quality, LD and $R^2$ like while creating a PCA in the manuscript and then split up into two VCF files to have one VCF per species with the vcftools command '—keep'. Then π was calculated for each species in windows of 10kb with '--window-pi 10000' using VCFtools v.0.1.16-GCC-9.3.0 (Danecek *et al.*, 2011). π values were visualised as boxplots and significance between them was calculated by t-test using R package ggplot2 (Wickham, 2016). For phylogenetic analysis, the individual eight SNP calling VCF files were used to create a PHYLIP matrix format with vcf2phylip v2.0 (Ortiz, 2019). Tree files were created with the command: 'iqtree2 -s file.phy --threads-max 1 -nt AUTO' of IQ-TREE v.2 (Minh *et al.*, 2020). Tree files were visualised as a midpoint rooted tree with Interactive Tree Of Life (iTOL) v.5 (Letunic and Bork, 2021). Of the orthogroups a presence absence variation (PAV) heatmap including clustering dendrograms was made using function heatmap.2 of R package gplots (Warnes *et al.*, 2022). Proportional bar-graphs of percentages of total pan-genome components as well as pan- components per most significant GO-terms were plotted with R package ggplot2 (Wickham, 2016). Here the soft core, MAGIC private, *A. c.* and *A. h.* specific pan- components are pulled out of the complete dispensable component with the 'other dispensable' left for visualization. For all R packages R v.4.2.3 (R Core Team, 2023) was used.

# Chapter 3: It's a kind of MAGIC, revealing the mysteries of vegetable amaranth and mapping the genomic regions impacting leaf yield.


## Abstract

Leafy amaranths are known to be a highly nutritious underutilised crop commonly consumed in Africa, with great potential for contribute to healthy diets and food systems. However, unfortunately research and breeding of amaranth is often neglected and not much is known about amaranth genomics underlaying key agronomic traits. Hence an *Amaranthus* multi-parent advanced generation inter-cross (MAGIC) population has been generated to enable mapping of quantitative trait loci (QTL) responsible for relevant traits including leaf yield. The parental lines were shown to be genomically, metabolically as well as phenotypically varying (**Chapter 2**) which indicates a sufficient quality mapping population. This onward study therefore focuses on analysing the MAGIC offspring accessions while aiming to map QTL.

342 offspring accessions were grown during the South African summer in four climatically different locations, i.e., Mafikeng, Bloemfontein, Roodeplaat and Nelspruit. Eight MAGIC founders were grown in Mafikeng and Roodeplaat only. Most of the categorical descriptor traits recorded in the founders were further segregating in the offspring lines, suggesting an adequate MAGIC population. Numerical growth data, i.e., dry as well as fresh leaf weight and chlorophyll, was shows to be correlated and significantly different between all four locations. Significant variation between the offspring accessions was only observed for dry and fresh leaf yield. An attempt to identify 'best' lines due to consistent performance in most different field trial locations for both leaf weight traits resulted in 51 selected offspring lines.

Investigation of two genotyping methods showed STITCH to be preferred over GATK haplotype caller for SNP imputation in 303 MAGIC offspring accessions with available sequences. These SNP markers genotyped with STITCH enabled successful QTL mapping for leaf yield. SNPs at chromosome 10 were found significantly associated with dry weight only and approached significance with fresh weight, while SNPs at chromosome 9 were significantly associated with both yield traits. The later chromosome 9 QTL has a larger proportion of variance explained with more genetic variation compared to the chromosome 10 QTL, marking high relevance of the chromosome 9 QTL. Genotypes of the three most significant SNP locations per QTL were clearly reflected in phenotypes of the offspring. This revealed beneficial allelic variants which can directly be used in amaranth breeding practices through marker-assisted breeding or selection. Identification of specific causal candidate genes or structural variation has been attempted, although, this remained challenging due to data constraints.

For future work we suggest conducting more generations of crossing as well as selfing to increase recombination rates and homozygosity in the offspring. Additionally, complete haplotype blocks or structural variation information from the founders should be genotyped in the offspring to aid identification of causal genes or sequence. Lastly, additional field trials recording more traits would potentially enable mapping diverse agriculturally relevant QTL.

# Introduction

Traditional breeding practices initially aimed for creation of high yielding individuals by improvement of agronomic traits based on phenotypic selection (Sun *et al.*, 2024). With the development of modern technologies and the ability to investigate and incorporate genomic information, major improvements and opportunities have risen in plant breeding (Sun *et al.*, 2024). One such successful technique is quantitative trait loci (QTL) mapping. This is an approach to mine gene function by investigating the relationship between genomic and complex phenotypic trait variation and identify marker-trait associations (Xu *et al.*, 2017; Cockram and Mackay, 2018; Ibrahim *et al.*, 2020; Sun *et al.*, 2024). Identified markers can in their turn be used in marker-assisted breeding or selection. This overcomes various problems of conventional breeding by direct selection of preferable genes, which can drastically improve cultivars by enhancing yields and other relevant traits (Mackay, Stone and Ayroles, 2009; Cockram and Mackay, 2018; Arrones *et al.*, 2020; Hasan *et al.*, 2021).

## Quantitative trait loci (QTL) mapping in various types of populations

Despite its success, identification of QTLs underlying quantitative relevant traits remains challenging for plant breeders. Typically four components are needed for enabling QTL detection; 1) a segregating population with a high genetic as well as phenotypic variability, 2) reliable marker systems that enable genotyping of the population, 3) reproducible phenotyping techniques for measuring the quantitative traits and 4) a well-structured experimental design accounting for environmental effects with suitable statistical methods to detect QTLs (Arrones *et al.*, 2020). Traditionally bi-parental populations were created for QTL mapping, since those are simple to develop by crossing two parental lines which differ for the trait of interest (Ladejobi *et al.*, 2016; Xu *et al.*, 2017; Arrones *et al.*, 2020; Sun *et al.*, 2024). However, by crossing only two founder lines narrow genetic diversity is captured, which limits detection of rare alleles (Ladejobi *et al.*, 2016; Xu *et al.*, 2017; Arrones *et al.*, 2020; Sun *et al.*, 2024). Later genome-wide association studies (GWAS) have been developed which overcome the limitation of narrow bi-parental marker density by leveraging natural populations of hundreds to tens of thousands of individuals. GWAS therefore investigates broader genetic diversity which achieves higher mapping resolution of multiple traits simultaneously (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020; Hasan *et al.*, 2021; Sun *et al.*, 2024). GWAS is a powerful tool to detect QTL, although, some constraints remain. It requires very large sample sizes with many markers for sufficient detection power, and missing parental information can lead to inaccurate results (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020; Ibrahim *et al.*, 2020).

Missing pedigree information in GWAS, as well as lack of broad genetic diversity in bi-parental populations, can be overcome by creation of multi-parental mapping populations (Ladejobi *et al.*, 2016; Xu *et al.*, 2017; Cockram and Mackay, 2018; Arrones *et al.*, 2020; Ibrahim *et al.*, 2020; Hasan *et al.*, 2021; Sun *et al.*, 2024). Nested association mapping (NAM) or multi-parent advanced generation inter-cross (MAGIC) are examples of such multi-parental populations. NAM populations are created by crossing a single inbred parent to a successive collection of diverse inbred lines which combines the advantages of association mapping and bi-parental populations by capturing rare alleles allowing new loci to be seen (Ladejobi *et al.*, 2016). NAM populations have been successfully created with a very commonly mentioned example of QTL mapping in maize for diverse traits like flowering time,

leaf blight resistance and leaf architecture (Xu *et al.*, 2017; Sun *et al.*, 2024). MAGIC populations are created by inter-crossing multiple founders (typically four, eight or sixteen) over various generations resulting in recombinant inbred line (RIL) offspring which are highly recombined mosaics of the founder genomes (Ladejobi *et al.*, 2016; Xu *et al.*, 2017; Arrones *et al.*, 2020; Sun *et al.*, 2024). The big advantage of MAGIC populations over other multi-parent populations is the crossing of multiple inbred founders for multiple generations prior to the construction of the RILs. This considerably increases accuracy and precision of QTL detection which means MAGIC populations provide excellent opportunities for dissecting complex traits and enhancing breeding populations (Xu *et al.*, 2017; Sun *et al.*, 2024).

The first statistical approaches used to map QTL which were originally applied in bi-parental populations, are single-marker analysis and several types of interval mapping (Xu *et al.*, 2017). For the first method, only a single marker is identified as a QTL. Interval mapping, on the other hand, scans the genome for genomic positions within pairs of two flanking markers estimating the likelihood of a QTL at each position. In this basic form it identifies a single QTL within each interval defined by flanking markers, but in more advanced forms, it can even identify multiple QTL positions in the genome. This is particularly useful for complex traits which are typically influenced by several loci responsible for varying phenotypes. With the development of multiparent mapping populations statistical analysis for detecting QTL had to be adjusted, since offspring marker information no longer directly reveal the origin of alleles from multiple parents. Hence, nowadays, multiple tools have been designed for QTL mapping in multiparent populations adjusting for their complex cross-design (Xu *et al.*, 2017; Arrones *et al.*, 2020). These methods usually apply interval mapping strategies for QTL detection like that implemented in the wheat MAGIC population of Scott et al., (2021).

An example of a successful QTL study in amaranth is the identification of a large QTL on chromosome 16 of *Amaranthus hypochondriacus* associated with tissue colour (red versus green) (Winkler *et al.*, 2024). Two well-known genes involved in the betalain pathway (*AhCYP76AD2* and *AhDODAα1*) were located in this QTL region. Another example is identification of QTL on chromosomes 3 and 9 of the *A. hypochondriacus* genome associated with seed colour, which is a key domestication trait in amaranths (Stetter, Vidal-Villarejo and Schmid, 2020). Here 116 accessions of five different *Amaranthus* species were used to conduct GWAS, with a candidate *MYB-like* transcription factor gene located in the chromosome 9 QTL region likely controlling seed colour. Both these examples are performed in grain amaranths, with QTL results for pigmentation related traits. No QTL analysis has been performed yet in vegetable amaranths or for yield traits. In case such analysis is successful, it would reveal highly valuable information for *Amaranthus* breeding programmes potentially generating increased yields.

## Multi-parent advanced generation inter-cross (MAGIC) populations

Development of MAGIC populations provide an excellent way to accurately map QTL, due to broad genetic diversity captured and multiple generations of crossing the founder lines (Xu *et al.*, 2017; Sun *et al.*, 2024). This highlights the advantage of MAGIC populations to create novel phenotypic diversity that may not be available in natural panels as they can be rapidly propagated and subjected to selection. There are various ways to create MAGIC populations as founders can be crossed in funnel or diallel designs (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020). In a funnel design founders are intercrossed a single time resulting in n/2 F1 hybrids.

A diallel design is similar to a funnel design, although now all founders are intercrossed in multiple funnels incorporating all possible crossing combinations between the founders for creating F1. For both designs, the F1 generations are subsequently intercrossed, followed by several generations of further intercrossing until all founder genomes are represented in a single generation of progeny RILs (Ladejobi *et al.*, 2016; Arrones *et al.*, 2020). The diallel crossing design would be preferred due to a higher variety of allele combinations, however a funnel design MAGIC population can be generated faster. A diallel crossing design is more labour intensive and the multiple crossing process can be problematic if the founders are difficult to hybridize (Arrones *et al.*, 2020).

Two examples of MAGIC populations with diallel crossing designs are an eight founder cowpea (Huynh *et al.*, 2018) and a 16 founder wheat (Scott *et al.*, 2021) MAGIC population created with six and 15 funnels respectively. The number of offspring RILs varies with 305 F8 cowpea RILs versus 504 F7 wheat RILs. It could be argued that ideally a wheat MAGIC population should require about 2500 RILs to obtain a similar subcentimorgan resolution to the 305 RIL cowpea population. This is due to the large genome size of wheat (within a 5 Gb range), compared to the small 620 Mb diploid genome of cowpea (Huynh *et al.*, 2018). However, it is obviously not always practical to achieve optimal large MAGIC population sizes for plants with larger genomes (Arrones *et al.*, 2020). Sizes of other developed wheat MAGIC populations have offspring RIL numbers ranging from 643 to 1091 (Huang *et al.*, 2012; Mackay *et al.*, 2014; Gardner, Wittern and Mackay, 2016; Sannemann *et al.*, 2018). Once MAGIC populations are developed there are various ways to generate genomic data and genotype the RILs. For example, the cowpea MAGIC population was genotyped using the existing Illumina Cowpea Consortium SNP array (Huynh *et al.*, 2018). While the wheat MAGIC population RILs were genotyped using low-coverage (~0.304×) whole-genome sequencing with Illumina paired end reads (2×150 bp) (Scott *et al.*, 2021). Two genotyping methods were tested for comparison in the wheat RILs, i.e., GATK HaplotypeCaller versus STITCH software. Using STITCH only marginally seemed to improve genotyping accuracy but increased the call rate threefold, which made it a preferable method for genotype imputation in this study (Scott *et al.*, 2021). Traits for which QTLs could be mapped in the cowpea population were all relevant to optimize growth and farming practices, like flowering time, growth habit, seed size and maturity, however, these traits are not directly testing for yield (Huynh *et al.*, 2018). In the wheat MAGIC population a large variety of traits (almost 40) was tested for QTLs, but most notably loci for grain yield could be mapped, which is agronomically highly relevant (Scott *et al.*, 2021).

## An *Amaranthus* MAGIC population in South Africa

No MAGIC population has been published yet for amaranths, which would be highly valuable to map QTL responsible for leaf yield and other related vegetable traits. Hence this PhD work attempted to characterise an *Amaranthus* MAGIC population developed by the world vegetable center, with the variation and characteristics of the founder lines tested in the manuscript in the appendix. By creating a pan-genome as well as performing field trials, it has been shown that the MAGIC founders vary for genomics as well as agriculturally relevant leafy traits. It is expected that if the founders differ, these characteristics will further segregate in the MAGIC offspring RILs, which is one of the key necessities to develop an adequate MAGIC population to enable QTL mapping (Arrones *et al.*, 2020). If QTL can be mapped for leafy traits this could highly benefit amaranth breeding programmes and

eventually contribute to food and nutrition security. Especially in regions where leafy amaranth is very commonly consumed, like sub-Saharan Africa, it could greatly contribute to climate-resilient agriculture and healthy diets (Achigan-Dako, Sogbohossou and Maundu, 2014; van Zonneveld *et al.*, 2023). More specifically in South Africa, leafy amaranths are commonly eaten and traded by locals, however, food systems could be greatly optimized if the lack of agronomic knowledge regarding leafy amaranth would be overcome (Mabhaudhi *et al.*, 2016, 2019; Emmanuel and Babalola, 2021). Hence the aim of this chapter is to contribute to knowledge of amaranth leaf yield and potentially map QTL for this most important trait. In order to do this, the developed MAGIC RILs have been grown in four different locations in South Africa (**Figure 3.1**). These field trial locations are climatically diverse, with the largest difference being the very wet and humid climate of Nelspruit in the Mpumalanga province (61.13 mm/month annual mean precipitation), versus Mafikeng in the dry North West province (44.31 mm/month annual mean precipitation) (Benhin, 2006). These different growth conditions can have large effects on the performance of the MAGIC offspring accessions.

If MAGIC offspring RILs or other breeding material is grown in different environmental conditions potential GxE effects might occur. Genotype by Environment interaction (GxE) can be defined as the relative change in phenotypic performance of two or more genotypes due to measurement in different environments (Falconer, 1952; Falconer and Mackay, 1996). Presence of an GxE effect is an urgent matter, since accounting for it might be needed using different breeding programs for different environments. This happens when the magnitude of the GxE is so big that environment in which the breeding program takes place and where genotypes are selected differs significantly from the environment the selected genotype eventually has to produce in (Falconer, 1952). Previous GxE research has proven its relevance and contribution to leafy amaranth breeding, by ability to select stable varieties and identifying environments suitable for breeding (Dewan *et al.*, 2017; Dinssa *et al.*, 2019; Lakshmidevamma *et al.*, 2022). In plant breeding, GxE investigation is of high relevance and there are many available methods to assess it (Zakir, 2018; Begna, 2020). Linear fixed effects models identifying significance with Analysis of Variance (ANOVA) can be used to reveal GxE effects, for example to detect GxE in rural dual-purpose chicken genotypes in Ethiopia (de Kinderen *et al.*, 2020, 2023). It is known that different amaranth varieties perform differently in various agro–ecologies in Southern Africa, hence looking into environmental suitability of varieties is important (Mabhaudhi *et al.*, 2016).

## Analysis of the *Amaranthus* MAGIC offspring RILs

The overall aim of this chapter is to investigate the performance of the MAGIC population RILs and to potentially map QTLs relevant for leafy traits. In order to do this, an *Amaranthus* MAGIC population has been newly developed, using the eight founder lines outlined in **Chapter 2** and a single funnel design for crossing the founders (**Figure 3.1**). Originally all crosses to generate F1 would have been intra-species, however, with corrected species of accessions VI061494-A and VI060472, this is no longer the case. The eight founder lines have been sequenced to generate high quality PacBio HiFi long-reads, which are assembled and annotated, with SNPs and structural variation identified and captured in an *Amaranthus* pan-genome as described in the manuscript in the appendix. The MAGIC RILs have been whole genome sequenced with low coverage (10x) to generate lower quality short-reads. Genotyping the MAGIC offspring and identifying QTL responsible for important agronomic

traits can be hugely beneficial for the *Amaranthus* breeding programme. Hence this current chapter focuses on evaluating the offspring of the developed *Amaranthus* MAGIC population, by analysing the phenotypic data of the field trials in South Africa and trying to map QTL.

A first research aim is to examine the phenotypic characteristics of the MAGIC population when grown at four different locations in South Africa (**Figure 3.1**). To assess the quality of the *Amaranthus* MAGIC population and its potential ability to map QTL, its variation for relevant agronomic traits will be investigated and if those traits further segregate in the offspring RILs. Additionally, it will be examined if the climatically different environments influence performance of the MAGIC population. This will be done by analysing if the accessions perform differently in the various field trial locations and if GxE effects can be observed. Despite the potential environmental effects on phenotypes, an attempt to select robust and best performing novel generated offspring accessions will be done.

After phenotypic characterisation of the MAGIC population, the next step is to perform genomic analysis and genotype the offspring RIL sequences. Hence a second goal of this chapter is to see if the RILs can be successfully genotyped and investigate what is the best genotyping approach.

Lastly, once genotyped and with varying phenotypes, it can be investigated if QTL can be mapped using the *Amaranthus* MAGIC population. Therefore, a last research aim is to investigate if QTL can be identified for leaf related traits and if genes responsible for leaf yield can be identified.

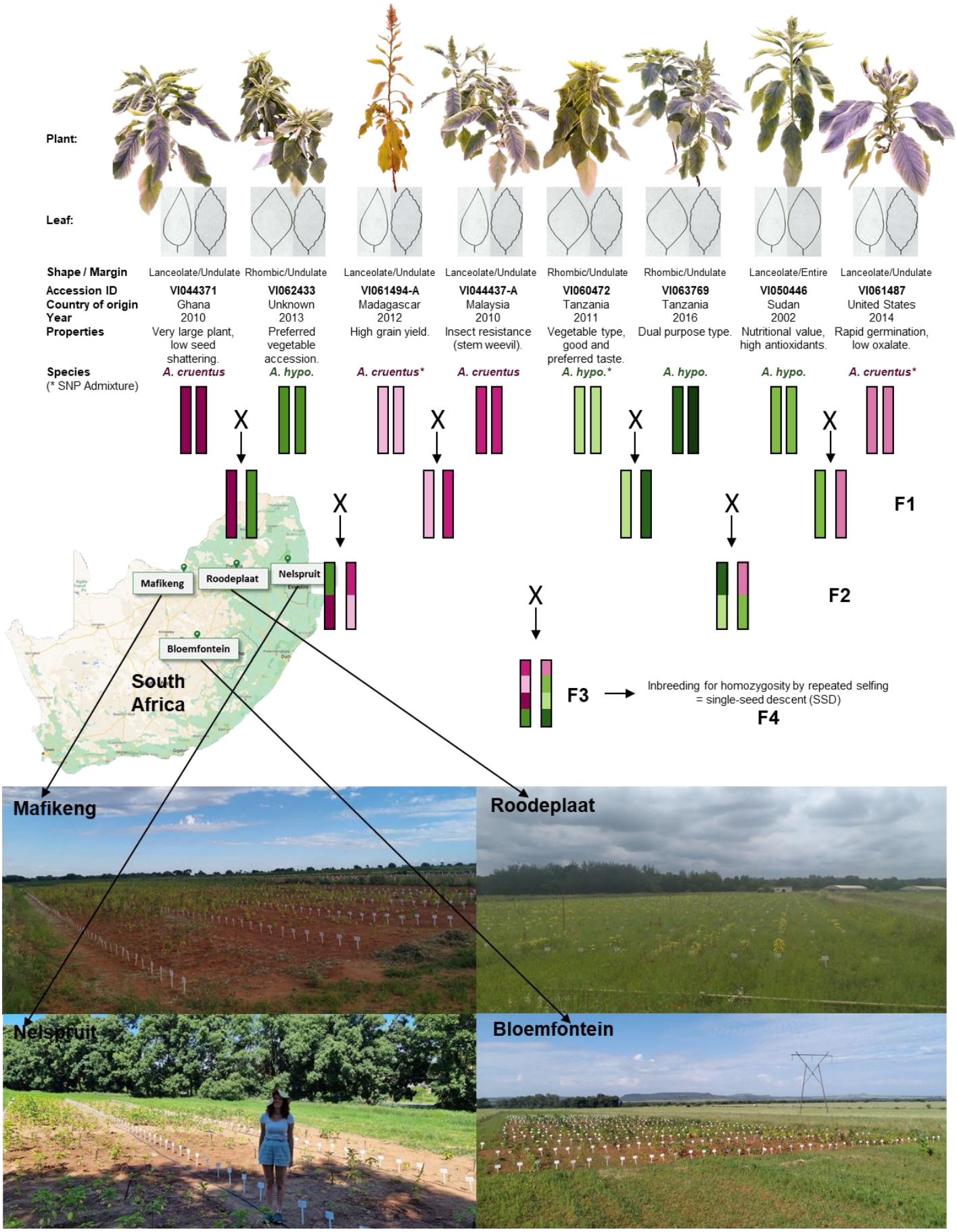



**Figure 3.1** *Amaranthus* MAGIC population crossing scheme representing which founders were crossed in which order to create an F4 generation of MAGIC offspring RILs used for mapping traits. Properties of the founders are provided as mentioned in the manuscript in the appendix with updated species information after correction. The map of South Africa is visualized including the four different field trial locations the MAGIC RILs are planted in and with photos of the four MAGIC population growth sites.

# Results

## Relevant agronomic traits are segregating in the MAGIC population

The MAGIC offspring accessions were grown in four locations in South Africa (**Figure 3.1**), while the MAGIC founders were only grown in two locations, i.e., Mafikeng and Roodeplaat. At three field trial locations, the plants were grown in a completely randomized design with three repetitions per location and ten plants of the same accession per row per plot. Nelspruit was an exception with only one replicate plot due to lacking seed availability. Categorical nominal data for descriptor traits was collected for MAGIC founders as well as offspring at all the different field trial locations for all available replicates. In most cases, phenotypic data could be collected for totals of eight parental lines and 342 offspring accessions.

In order to get an insight into the distribution of descriptor traits and their segregation among the MAGIC offspring, the most commonly assigned phenotype of categorical traits recorded in the individual MAGIC parents as well as offspring is visualised in **Figure 3.2**. Often further segregation of traits among the MAGIC offspring was observed in comparison to the parents. Some clear examples are leaf shape and inflorescence shape, density or colour, with many different descriptors identified per trait in the offspring compared to the MAGIC parents. Often 'other' is assigned for these four traits in the offspring because different descriptors were assigned to the same accession in equal proportion. This indicates difficulty to identify consistent descriptors among offspring and therefore likely continuing segregation among these accessions. For growth habit it might seem that only one descriptor is present, i.e., erect. However, when looking at single data points, some MAGIC offspring plants grown in various locations or replicates were assigned as prostate growth habit (23 out of 2497 data points). This could either be and additional example of further segregation among MAGIC offspring accessions and that they are not all homogenous, or that the lines could be misassigned for phenotypes. The most notable example of segregation observed in the MAGIC population is for inflorescence density, which is predominantly intermediate among the offspring, contrasting with the more extreme either lax or dense phenotypes seen in the founders (**Figure 3.2**). This suggests inflorescence density is probably a quantitative trait, likely influenced by multiple genetic loci.

Numerical growth-related data, i.e. dry as well as fresh leaf weight and chlorophyll, was collected from the MAGIC offspring in four distinct agro-ecological locations in South Africa (**Figure 3.1**). These measurements were total kilograms (KG) of dry or fresh leaf weight per plot (usually ten plants) or the average of three chlorophyll measurements per plot, with the (usually three) replicate plots and four locations per accession taken as individual data points (**Table 3.14**) used for analysis. In order to identify differentiation of these growth-related traits due to climatic conditions, their significance between the locations was tested with a Wilcoxon signed-rank test for non-normally distributed data and with Tukey's letter codes for normal distributed data. For all traits, the data collected was significantly different in each location (**Figure 3.3 A**). Segregation of the growth-related traits can be assessed by comparing MAGIC offspring with parental data in Mafikeng and Roodeplaat. This showed that especially for chlorophyll, segregation is very clear, with offspring data scattered around the MAGIC parent line data (**Figure 3.3 A**). For conducting proper analysis, the normality of

the numerical data of the three growth-related traits was investigated. In case of non-normality, the data was log transformed prior applying any further statistics (**Figure 3.3 B**).

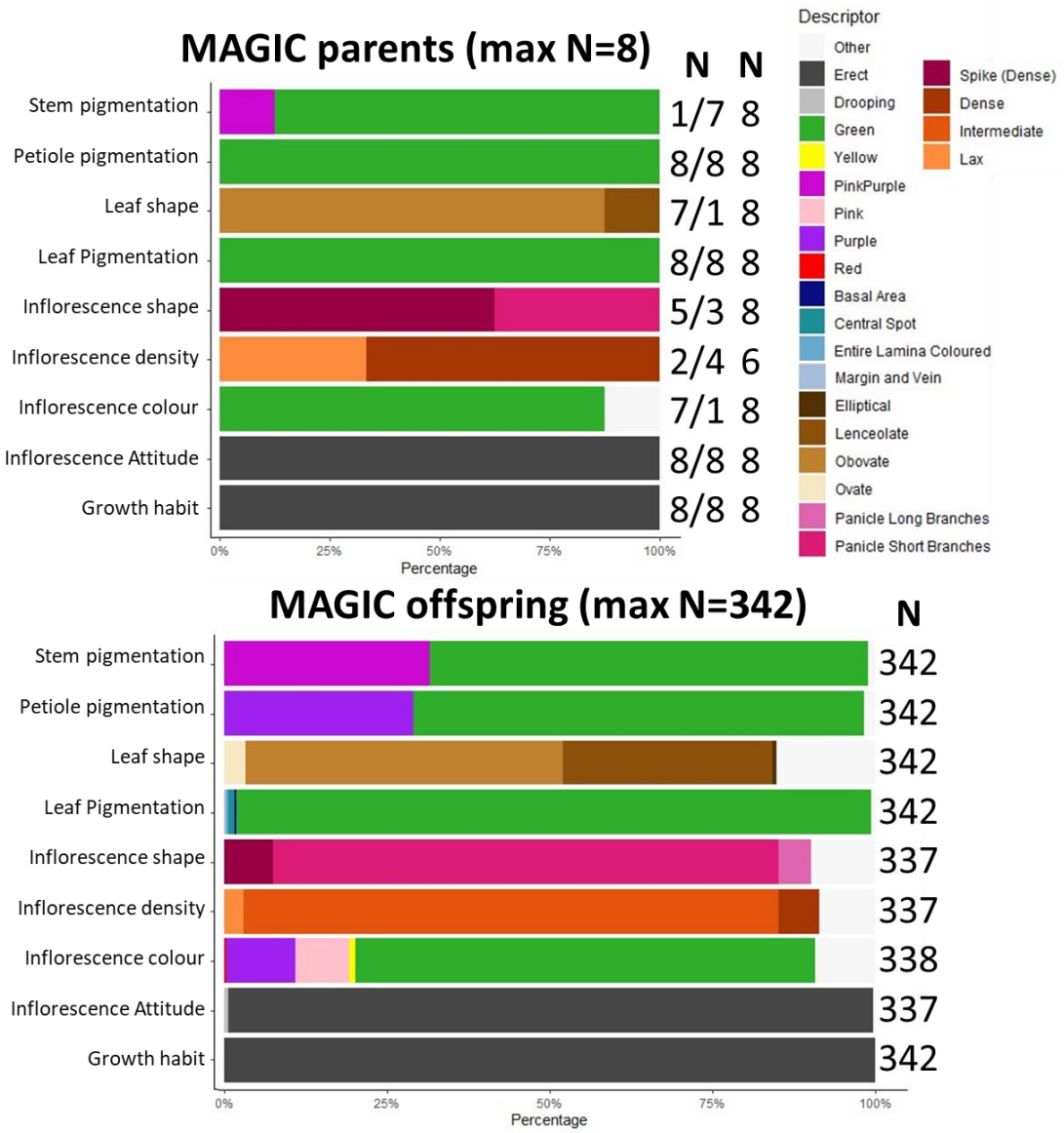



**Figure 3.2.** Proportional bar graphs showing how categorical traits are divided among the eight MAGIC parents (top) and further segregate among the MAGIC offspring (bottom). The assignment of a certain trait for a single accession was based on what was most commonly assigned over all data of the different field trial locations and repeats. In case a different assignment per accession was in an equal proportion with each other the descriptor of this trait was then described as 'other'.

Regarding leaf yield, MAGIC offspring seemed to have lower fresh and dry leaf yield weights compared to the MAGIC parents (**Figure 3.3 A**). This could indicate these are multigenic traits with multiple loci responsible for the observed phenotypic variation, however, such differences may also arise from dominance effects, epistasis, or environmental influences rather than multigenic inheritance alone. Additionally, it might be due to the absence of selection during the creation of the MAGIC population offspring, allowing recombination to occur, whereas the parental lines are well selected cultivars. This recombination is likely to break up beneficial allelic combinations which were selected in the parents, rather than creating new preferable combinations contributing to leaf yield. Pleiotropy occurs if one or multiple loci have an effect on more than one phenotype, hence affecting multiple

quantitative traits simultaneously (Mackay, Stone and Ayroles, 2009). The potential for pleiotropy was investigated by identifying positive correlations among the non-categorical traits. As expected, all three growth related traits are all positively correlated with each other likely due to their biological relatedness (**Figure 3.4**). The trait correlations are all highly significant ($p < 0.0001$), especially between the leaf weight traits (**Table 3.1**). Therefore, existence of pleiotropy is suggested between these three traits, especially between dry and fresh leaf weight which are known to have a shared biological basis.

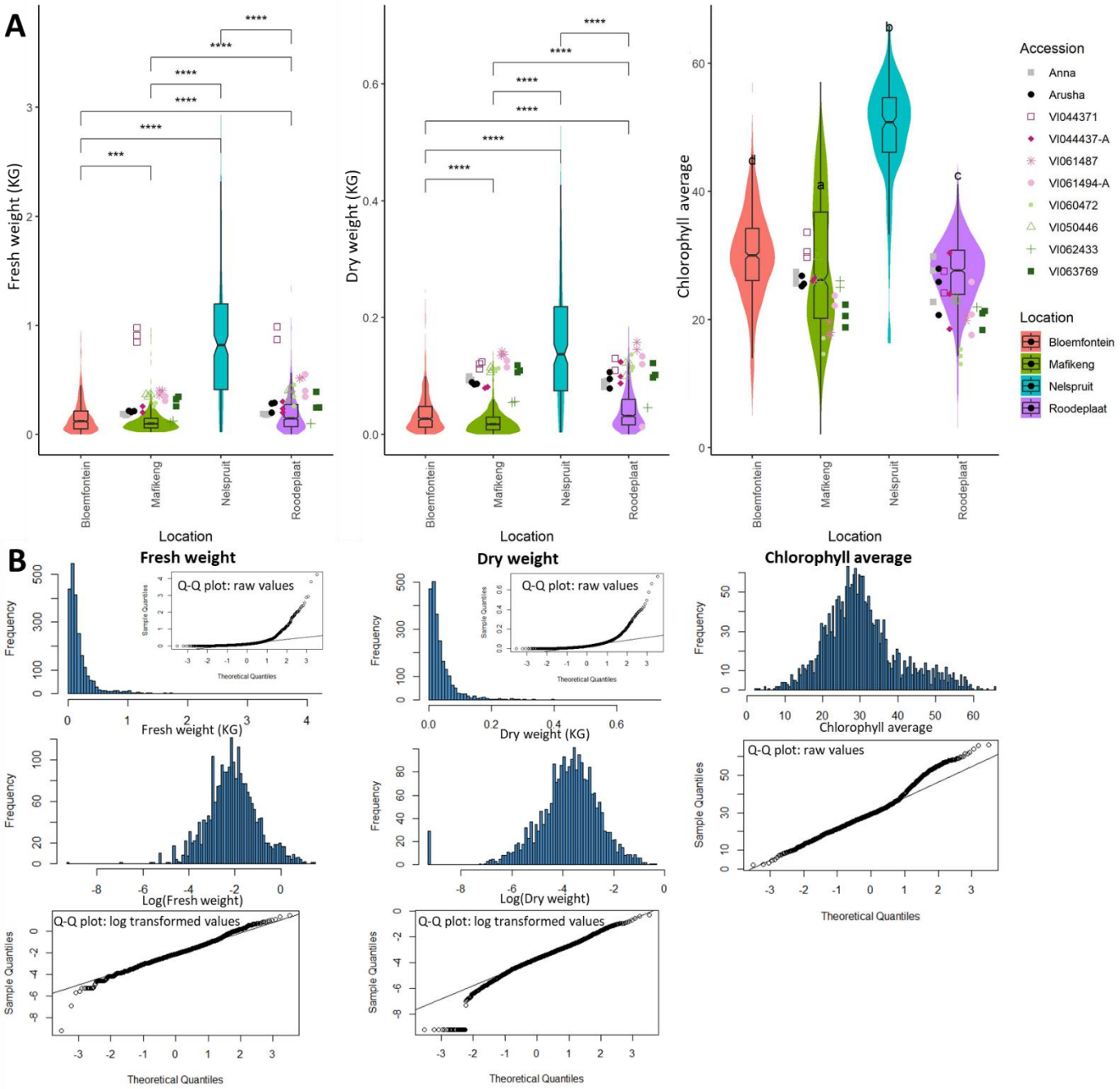



**Figure 3.3.** Fresh and dry leaf weight and chlorophyll data collected at four field trial locations. **A)** MAGIC offspring collected data plotted as boxplots and violin plots and MAGIC parent or accessions Anna and Arusha data grown in two locations plotted as separate datapoints. Wilcoxon signed-rank test was used for non-normally distributed fresh and dry leaf weight data with significance indicated as **** = $p \leq 0.0001$ and *** = $p \leq 0.001$. Tukey's HSD test was used for post-hoc comparisons for the normal distributed chlorophyll data with letter codes for significance. **B)** Histograms and quantile-quantile (Q-Q) plots displaying distributions of raw, and in case not normal, log transformed values, of the numerical data for the three recorded traits for significance.

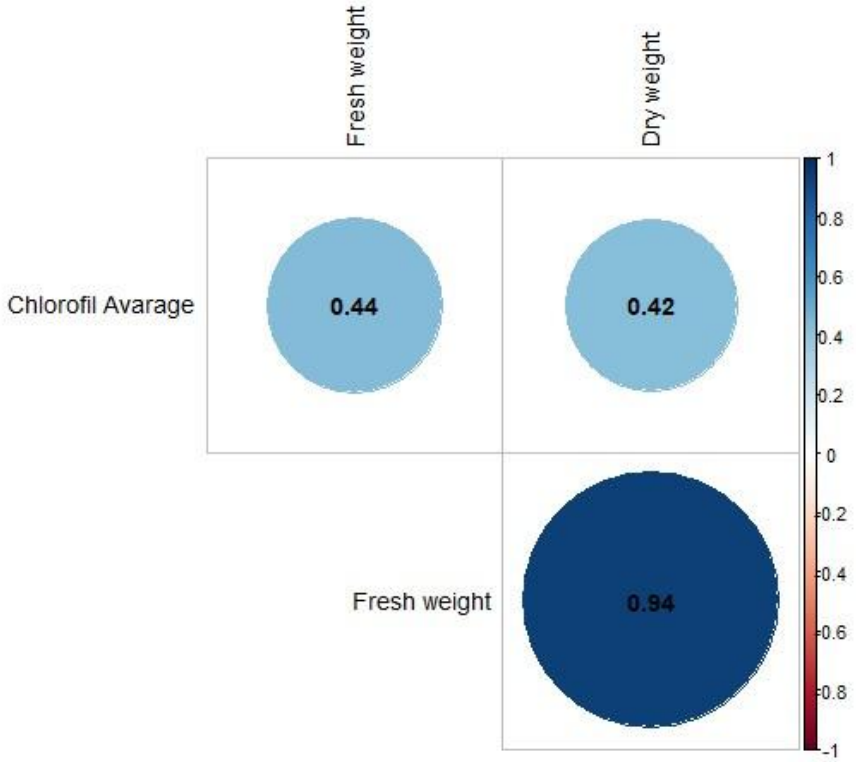



**Figure 3.4.** Agronomic trait values observed in the MAGIC offspring accessions are all significantly ($p < 0.05$, Pearson's correlation test, detailed statistics are in **Table 3.1**) and highly positively correlated with each other which suggests pleiotropy. 2013 data points were used for each trait category because these are all three recorded for 322 different offspring accessions in certain test locations or replicates. The maximum measurements per individual accession was 10, minimum 1 with a median of 6. The darker blue and bigger the circle the higher correlations between traits are.

**Table 3.1** Highly significant p-values of agronomic traits observed to be correlated in the MAGIC offspring accessions by Pearson's correlation test. All traits were shown to be positively correlated (**Figure 3.4**).

| **Trait combination** | **p-value** |
|---|---|
| Fresh weight - Dry weight | ~ 0 |
| Fresh weight - Chlorophyll | 2.75E-94 |
| Dry weight - Chlorophyll | 3.15E-87 |

To examine if there is variation between the MAGIC offspring accessions and the different growth locations, linear model effects were calculated for leaf weight and chlorophyll traits. For fresh as well as dry leaf yield, there were significant effects of genotypes and field trial locations, with the latter environmental effect only observed for chlorophyll (**Table 3.2**). This shows there is significant variation between the MAGIC offspring accessions for both leaf yield traits, but not for chlorophyll, while the environmental locations impact all three traits. To investigate if a change in phenotypic performance occurs when the various offspring accessions are grown in different environments, genotype by environment interaction (GxE) effects were assessed. Only for the most important agronomic trait, fresh leaf yield, was a G*x*E effect observed (**Table 3.2**). This means that fresh leaf yield performance was not consistent across genotypes in all environments which should be considered while selecting the offspring genotypes for certain environmental conditions.

In order to investigate if the MAGIC offspring vary at each separate field trial location, genotype effects were estimated for data separated by environment. No accession effect could be calculated within Nelspruit as a location since there was only one repeat performed. For leaf yield traits a significant effect of accession was observed in Roodeplaat and Bloemfontein (**Table 3.3**). For chlorophyll a significant accession effect was only observed at

Roodeplaat, however, significance was approached at Bloemfontein. No significant genotype effect was observed at Mafikeng for any of the traits (**Table 3.3**). This means that the MAGIC offspring accessions performance for leaf yield differs within the Roodeplaat and Bloemfontein locations, while no phenotypic variation could be found at Mafikeng. The latter is likely explained by Mafikeng being the driest environment with the lowest yields, restricting the variation in yield that could be observed (**Figure 3.3 A**).

**Table 3.2.** Best fitting ANOVA fixed and random model effects of three relevant agronomic traits; fresh, dry leaf weight and chlorophyll. Data was log transformed (**Figure 3.3 B**) in case of no normal distribution. Fixed effect are genotypes (MAGIC offspring accessions), environment (field trial locations) and their interaction (GxE). Repeat (Rep) of the experiment was taken as a random model effect. P-values are given for the fixed effects and displayed in bold when significant (P < 0.05).

| **Model effects:** | **Fixed** | | | **Random** |
|---|---|---|---|---|
| Traits: | **Genotypes (n=342 offspring)** | **Environment (n=4 locations)** | **GxE** | **Rep (n=3)** |
| log (Fresh weight) (N=2335) | **P = 2.602e-07** | **P < 2e-16** | **P = 0.0012** | x |
| log (Dry weight) (N=2306) | **P = 3.32e-07** | **P < 2e-16** | P = 0.3244 | x |
| Chlorophyll (N=2126) | P = 0.3418 | **P < 2e-16** | P = 0.9931 | x |

**Table 3.3.** ANOVA effect of accession on data which was separated per different field trial location. Significance of accession was tested for three different relevant agronomic traits (dry (DW) and fresh (FW) leaf yield and chlorophyll) and displayed bold in case of significance. Data was log transformed (log()) (**Figure 3.3 B**) in case the data was non-normally distributed. Number of (No) accessions calculations were based on are provided.

| **Traits:** | **Yield** | | | **Chlorophyll average** | |
|---|---|---|---|---|---|
| **Location** | **log(DW)** | **log(FW)** | No accessions | **Chlorophyll** | No accessions |
| Roodeplaat | **P = 9.056e-05** | **P = 9.53e-06** | 305 | **P = 0.01257** | 310 |
| Mafikeng | P = 0.5346 | P = 0.203 | 312 | P = 0.3253 | 316 |
| Bloemfontein | **P = 0.02198** | **P = 0.01148** | 296 | P =0.05229 | 297 |

Although significant GxE and environmental effects could be observed for leaf yield, we wanted to identify accessions that may not provide the highest yield in a single location but would provide the best overall yield across all four environments therefore being robust. An attempt to identify which offspring accessions are the 'best' regarding leaf yield across the four environments was done by plotting their yield performance in parallel coordinates plots. Two versions were made per leaf yield trait, one plot with all accessions present in all four locations and one plot with all accessions present in three locations by excluding the limited data of the single plot repeat in Nelspruit (**Figure 3.5**). Accessions which are among the best 60% of leaf yield performance within each location are highlighted with black instead of grey lines. For the first plot version visualizing performance in four locations (N=161), 82 and 79 accessions provided the best 60% of fresh and dry leaf weight performance respectively, in all four or at least three locations of the four locations. These

thresholds were implemented since if accessions among the best 60% in all four locations were selected, only 28 or 19 lines would be best for dry and fresh weight respectively. These numbers would be too stringent to eventually select about 50 overlapping best ones.

For the second plot version visualizing performance in three locations (N=273), 68 and 69 accessions provided the best 60% for fresh and dry leaf weight respectively for all three locations. Overlapping all these selected lines resulted in a total of 51 MAGIC offspring accessions which were amongst the best 60% performance in all parallel coordinates plots (**Table 3.4**). These 51 accessions can therefore be considered 'best' due to their consistent performance in most different field trial locations for both most relevant leaf yield weight traits. To further quantify these results estimated breeding values (EBV) have been calculated for each leaf yield trait. These are presented and ordered by accessions with highest EBV for fresh weight in **Supp table EBV offspring**. Many of the accessions found with high EBV are also present in **Table 3.4**, further supporting our analysis. Identification of these well performing accessions is of high relevance due to their yield potential despite environmental conditions. Therefore, these novel generated accessions can directly be grown by farmers located in various agro-ecological zones as well as being used as material for continuing breeding purposes.

**Table 3.4.** 51 accessions selected as 'best' due to consistent performance over several field trial locations for the most relevant agronomic leaf traits (fresh and dry weight). These are MAGIC offspring accession ID's starting with ' AVAM 22 –' followed by the number in the table.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 199 | 202 | 205 | 209 | 216 | 217 | 222 | 225 | 229 | 232 |
| 237 | 238 | 243 | 253 | 257 | 263 | 269 | 276 | 277 | 282 |
| 289 | 293 | 301 | 308 | 320 | 334 | 339 | 344 | 354 | 356 |
| 364 | 376 | 378 | 389 | 390 | 409 | 411 | 415 | 418 | 435 |
| 445 | 450 | 451 | 452 | 455 | 456 | 457 | 461 | 469 | 480 |
| 500 | | | | | | | | | |

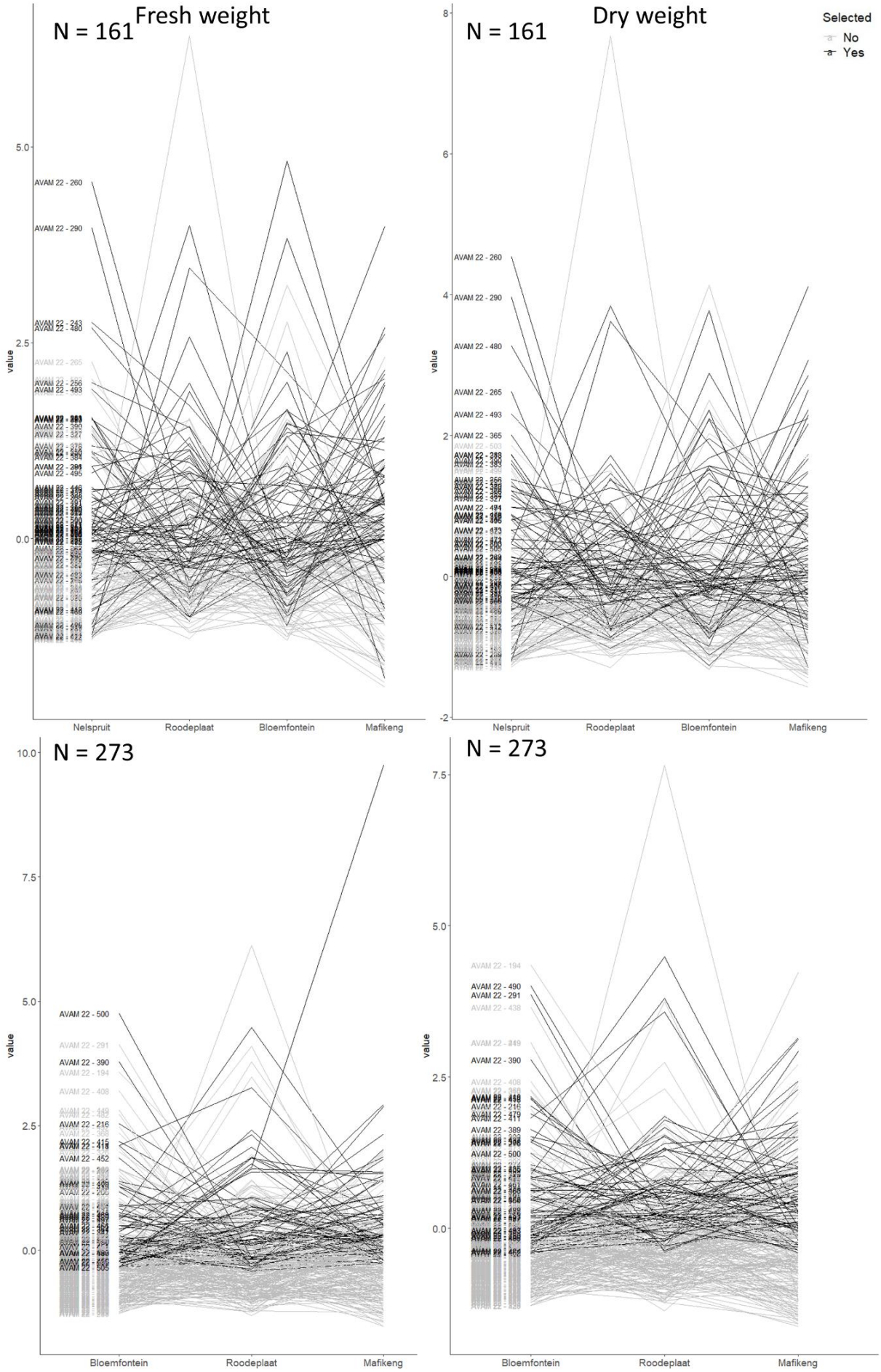



**Figure 3.5.** 'Best' selected MAGIC offspring accessions having the highest 60% fresh (left) or dry (right) leaf yield weight in all four or at least three field trial locations (top), or in all three field trial locations excluding the limited data of the single repeat Nelspruit trial (bottom). Number of accessions plotted in parallel coordinates plots is provided (N=) and accessions ID's are given at the left side of each plot. Y-axis values represent a variable corresponding to observations but do not have any further inherent meaning.

## Quality control for genotyping of the MAGIC RIL sequences

Now we know that the MAGIC population varies for leaf traits, the next step is to try and determine genomic regions responsible for this variation, and ultimately candidate genes (and variants) that control the varying traits. In order to do this, the MAGIC offspring RIL sequences should be genotyped. Whole genome sequences (10x coverage) of the MAGIC offspring were generated which were simultaneously trimmed and checked for quality with Fastp. Trimming slightly, but not drastically, improved sequence quality (**Supp table sequencing stats RILs**), hence trimmed reads were used for further analysis.

### Mapping coverage of MAGIC offspring sequences to investigate read quality

It was previously shown that the *A. cruentus* reference genome (Ma *et al.*, 2021) with updated gene annotation is of high quality while using it for pan-genome construction in **Chapter 2**. Trimmed reads of the MAGIC offspring have been mapped against this good quality reference genome using Burrows-Wheeler Aligner (BWA) BWA v.0.7.17-GCCcore-11.3.0 (Li and Durbin, 2010). To look into the offspring read quality as well as to make sure that information is detected while mapping, the mapping coverage of the reads towards the reference genome has been investigated. Overall, the median coverage means per accession was approximately 6.5 with no mean coverage lower than 5.09 or higher than 8.25 (**Table 3.5**). Most (57% of non-exons and 83% of exons) of the offspring reads mapped against the reference with coverage of >6 (**Table 3.6**). This all indicates good mapping coverage results for only 10x coverage sequences.

**Table 3.5.** Lowest (Min.), median of mean per accession, and highest (Max.) coverage of 303 MAGIC offspring reads mapped to exonic and non-exonic regions of an *A. cruentus* reference genome (Ma et al., 2021). Mapping performed to complete genome and main reference chromosomes only (Excl. scaffold).

| Genome region: | Exon | | | Non-Exon | | |
|---|---|---|---|---|---|---|
| | Min. | Median | Max. | Min. | Median | Max. |
| Complete | 5.22 | 6.58 | 8.25 | 5.09 | 6.40 | 7.00 |
| Excl. scaffold | 5.22 | 6.58 | 8.25 | 5.09 | 6.41 | 7.02 |

A lower coverage was found for the non-exonic versus the exonic regions of the reference genome (**Table 3.5**, **Table 3.6**, **Figure 3.6**, **Figure 3.7**). A potential reason for higher mapping to exons compared to non-exons is the presence of coding genes and therefore less transposable elements due to their damaging effects in exons (Bourque *et al.*, 2018; Colonna Romano and Fanti, 2022). This results in more stable exonic regions, which improve read alignment. What can be observed is a slightly lower coverage if reads are mapped to the complete reference genome compared to the main chromosomes of the reference only (**Table 3.5**, **Table 3.6**). This is likely due to scaffolds not containing much information, such as repetitive sequences that often lack diversity or annotated genes and therefore offering less information or unique sequence to map to. Since there was no major difference between using the complete reference or main chromosomes only it was chosen to continue analysis with the complete reference genome to provide the most optimal data availability for genotyping and QTL mapping.

**Table 3.6.** Number of MAGIC offspring lines mapping in bands of <6x, 6-7x or >7x coverage to an *A. cruentus* reference genome (Ma et al., 2021) including percentage of accessions compared to the entire group of 303 MAGIC offspring. Mapping was performed to exonic and non-exonic regions and to the complete genome and main chromosomes only excluding scaffolds.

| | <6 | 6-7 | >7 | <6 | 6-7 | >7 |
|---|---|---|---|---|---|---|
| | Complete genome: | | | No scaffolds: | | |
| Exon | 51: 16.8% | 230: 71.9% | 22: 7.3% | 51: 16.8% | 229: 75.6% | 23: 7.6% |
| Non-exon | 131: 43.2% | 171: 56.4% | 1: 0.3% | 129: 42.6% | 173: 57.1% | 1: 0.3% |

In order to investigate the consistency of read mapping towards (non-)exonic reference genome regions, histograms were plotted with the frequencies of all mean coverages across these genomic regions. The unimodal distribution in **Figure 3.6** indicates a good and stable coverage distribution with a high (>5) mean coverage for the 10x sequence reads. The only exception to the unimodal distribution is a peak for very low coverage at value zero (**Figure 3.6**). This can be explained by either variation in the sequenced reads (like mutations or errors) which becomes visible while comparing reads to a reference, or by no coverage by change due to statistical fluctuations in coverage (Lindner *et al.*, 2013). Another way of showing the consistency while performing read mapping are boxplots of coverage to (non-)exonic regions per offspring accession. **Figure 3.7** shows that there is little variation of coverage per accession with a continual range from lowest exon coverage accession to the highest, i.e. 5.22 to 8.25. This indicates good quality and consistent sequence data of all offspring accessions which could map well and without concerning amounts of variation between accessions towards the reference genome.

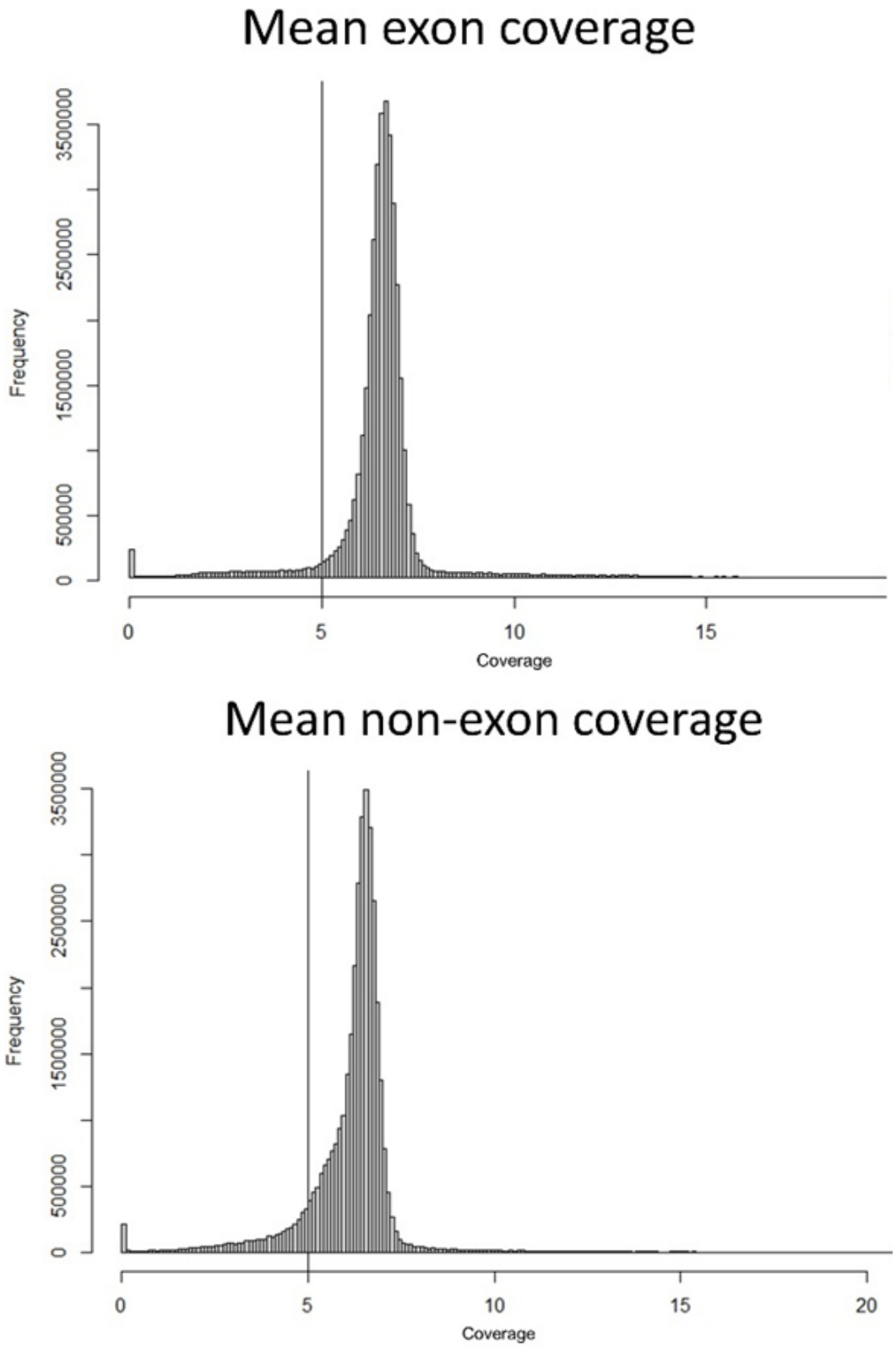



**Figure 3.6.** Histograms showing the frequency of mean coverage towards exonic and non-exonic regions of the complete Arusha reference genome that MAGIC offspring reads were mapped to.

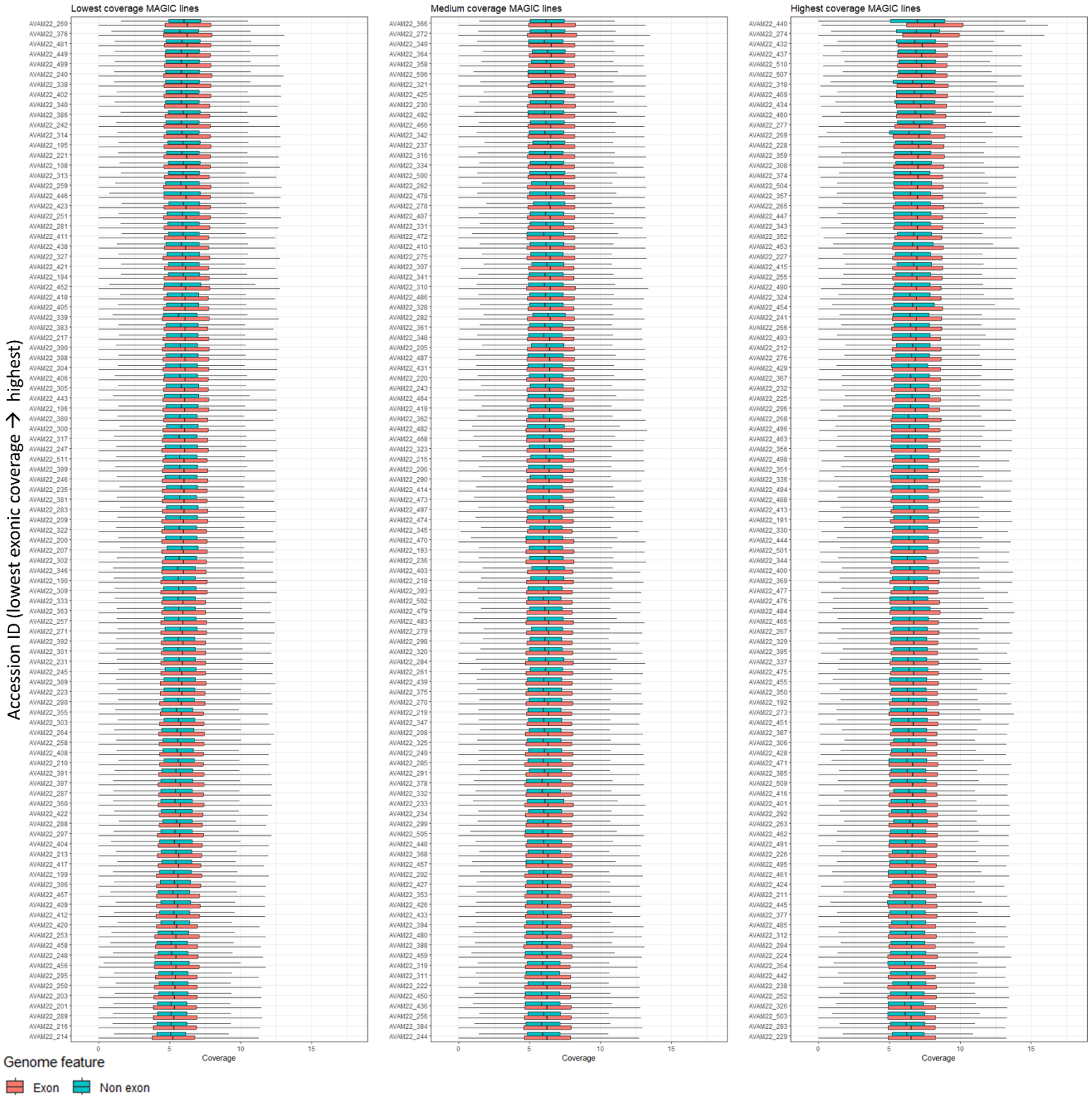



**Figure 3.7.** Boxplots displaying coverage per MAGIC offspring accession mapped to the complete *A. cruentus* reference genome (Ma *et al.*, 2021). Coverage is per exon or non-exon region and ordered by increasing median exon coverage from bottom to top and from left to right.

## MAGIC offspring genotyping with two methods reveals benefits of using STITCH

Two SNP imputation methods (STITCH and GATK haplotype caller) were tested to identify which one was best suitable for genotyping the MAGIC offspring RILs. Here the MAGIC founder SNPs called with Longshot (Edge and Bansal, 2019) against the *A. cruentus* reference genome (Ma *et al.*, 2021) and filtered for "PASS", QUAL > 50 and DP > 10 (like in the manuscript in appendix) were used for imputation. The first genotyping method investigated was STITCH (Davies *et al.*, 2016), which conducts haplotype-based SNP

imputation by modelling each chromosome as a mosaic of the population founder haplotypes. Since it is known that SNP imputation accuracy only marginally improves when including founder information (Scott *et al.*, 2021), it was chosen to run STITCH without a founder reference panel in our analysis. The number of founders (*k*) was set to eight while running STITCH. In order to investigate various levels of recombination rates between the offspring and the founder haplotypes, different values for the nGen parameter were chosen while running STITCH. To explore model performance and to make sure that we are not underfitting recombination, we tested both nGen = 4 (close to the true number of crossing generations in the MAGIC population) and nGen = 7 (representing higher recombination) on a single chromosome. Both nGen parameters resulted in all 293,876 SNPs called against reference chromosome 7 being imputed at least once in one or more offspring accessions. There were no major differences in relevant parameters such as call rates, homozygosity and inbreeding after SNP imputation with different recombination rates (**Table 3.7**). As different recombination rates did not seem to have much influence, it was chosen to continue with nGen=4 which is the true number of generations for our *Amaranthus* MAGIC population.

**Table 3.7.** Outputs for running STITCH on a single chromosome (7 of *A. cruentus* (Ma et al., 2021)) without a reference panel but with a different value of nGen, i.e., 4 or 7, controlling recombination rates. These are call rates of the total available SNPs, percentage of imputed SNPs that are homozygous, and inbreeding coefficients per individual offspring RIL with minimum, mean and maximum values provided.

| | **Min** | | **Mean** | | **Max** | |
|---|---|---|---|---|---|---|
| **nGen=** | **4** | **7** | **4** | **7** | **4** | **7** |
| **Call rate** | 0.878 | 0.874 | 0.949 | 0.947 | 0.982 | 0.982 |
| **Homozygosity (%)** | 32.7 | 32.9 | 75.1 | 75.3 | 99.1 | 99.1 |
| **Inbreeding (F)** | -0.921 | -0.921 | 0.286 | 0.290 | 0.974 | 0.975 |

The second imputation method investigated, GATK haplotype caller (Auwera and O'Connor, 2020), is a reference-based method which performs a pairwise alignment of each read against haplotypes of a reference genome. The call rate improved when using STITCH instead of GATK haplotype caller (206 grey dots above the red dotted diagonal of the total 303 offspring) (**Figure 3.8 A**). This indicates improvement of genotyping using STITCH compared to GATK haplotype caller. Ideally accuracy of the calls should be addressed to verify whether their presence is true or not, but this was not possible with the current data. In order to improve the quality of the STITCH imputed SNPs, the output VCF file was filtered for an info score of <0.4, minor allele frequency of <5% and missingness of >10%. After filtering the number of SNPs in the STITCH VCF file was reduced from 4,774,390 to 2,750,618 SNPs. Call rates further improved after filtering with 221 black dots now above the red diagonal instead of the 206 grey dots (both proportions of the total of 303 offspring accessions) and many accessions having much higher call rates (**Figure 3.8 A**). Mean call rates increased from 0.933 with GATK haplotype caller to 0.949 in unfiltered and 0.964 in filtered SNPs imputed with STITCH (**Table 3.7**, **Table 3.8**). Just as the minimum values shift from 0.854 to 0.878 to 0.887, and the maximum values from 0.957 to 0.982 to 0.996 (**Table 3.7**, **Table 3.8**), respectively, showing the overall distribution shifts towards higher call rates (**Figure 3.8 B**). This shows more SNPs remain after filtering and this STITCH dataset is the completest, however, ideally accuracy should be assed to determine if this set is the best quality.

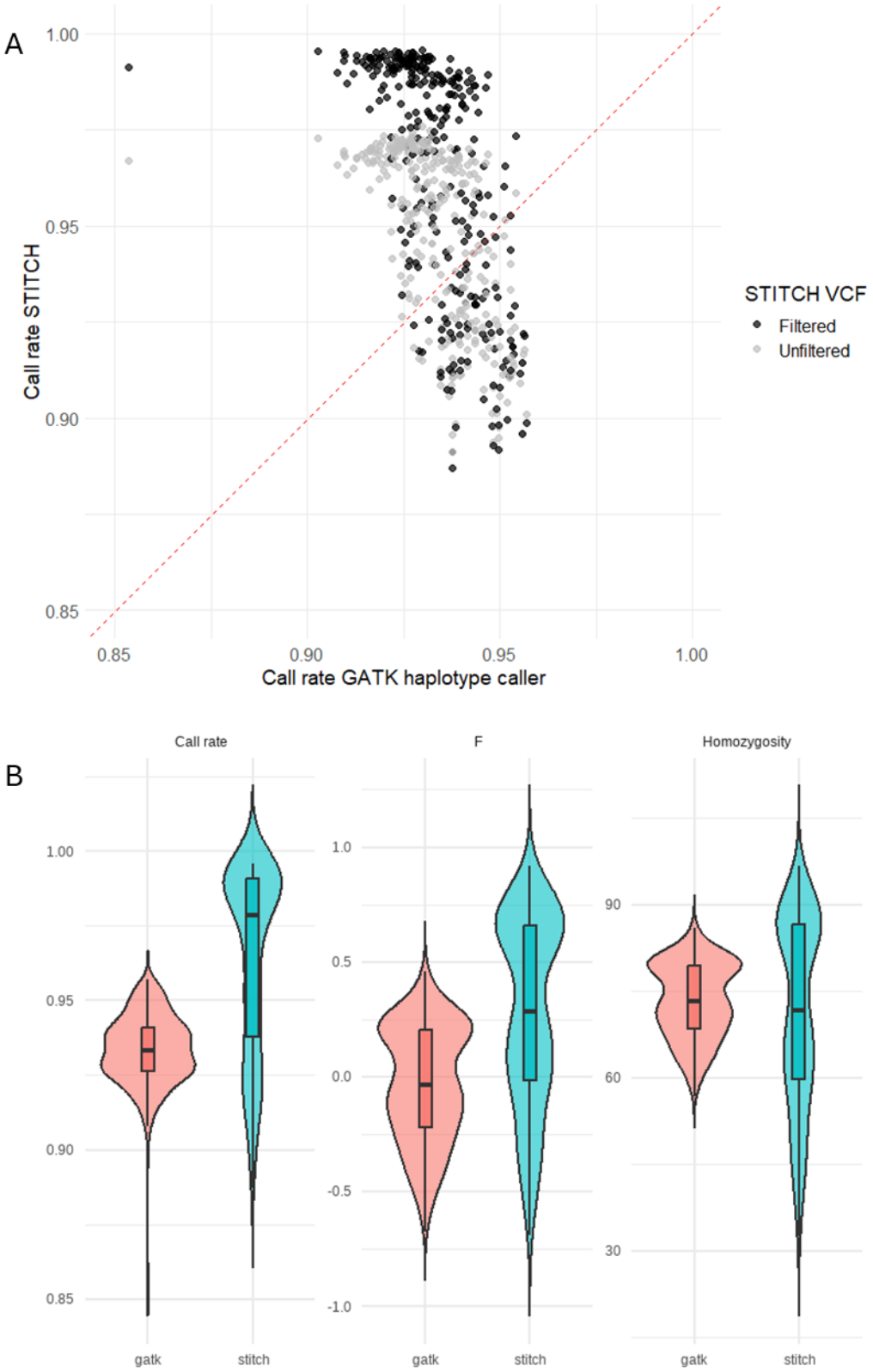



**Figure 3.8**. Genetic quality control metrics after genotyping the MAGIC offspring RILs with two different methods, i.e., GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies et al., 2016). **A)** Improving call rates due to SNP imputation method (GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies *et al.*, 2016)) prior (grey) and after (back) filtering (info score <0.4, minor allele frequency <5% and missingness >10%). Here call rates are proportions (1=100%) of the total available SNPs called for each of the 303 offspring accessions, i.e. 303 black or grey dots. **B)** Box- and violin plots of call rates, inbreeding coefficients (F) and homozygosity (%) of all 303 offspring accessions per genotyping method, i.e., GATK haplotype caller or STITCH after filtering. Minimum, mean and maximum values of these plots are also provided in **Table 3.8**.

To investigate the level of inbreeding of the RILs the percentage of homozygosity and inbreeding coefficients (F) were calculated. These were notably variable, with F ranging from negative values to almost 1 when genotyping with STITCH (**Figure 3.8 B**, **Table 3.8**). This is not unsurprising since the parental accessions had a mean F of 0.54 and mean homozygosity of 88%. This indicates that the selected founder accessions were fairly inbred, and therefore, after crossing them notably more outbred RILs (mean ~0.1 F and mean ~72% homozygosity) were generated. Generally, no extreme differences could be observed comparing the mean homozygosity and inbreeding coefficient of the two genotyping methods (**Figure 3.8 B**, **Table 3.8**). This indicates that the SNP imputation approach did not substantially matter for investigating inbreeding and neither method is preferred.

**Table 3.8** Differences in call rates, homozygosity and inbreeding coefficients of the MAGIC offspring RILs using two different genotyping methods, i.e., GATK haplotype caller (Auwera and O'Connor, 2020) versus STITCH (Davies et al., 2016). Values for STITCH were based on the VCF after filtering (info score <0.4, minor allele frequency <5% and missingness >10%). All values are also plotted in **Figure 3.8 B**.

| | **Min** | | **Mean** | | **Max** | |
|---|---|---|---|---|---|---|
| | **GATK** | **STITCH** | **GATK** | **STITCH** | **GATK** | **STITCH** |
| **Call rate** | 0.854 | 0.887 | 0.933 | 0.964 | 0.957 | 0.996 |
| **Homozygosity (%)** | 56.7 | 32.6 | 73.3 | 71.4 | 86.0 | 96.8 |
| **Inbreeding (F)** | -0.673 | -0.691 | -0.033 | 0.280 | 0.458 | 0.919 |

To examine the contribution of the MAGIC founders to the MAGIC offspring, principal component analysis (PCA) was conducted after genotyping with STITCH (Davies *et al.*, 2016) using the filtered SNPs. Principal component 1-3 explained about equal amounts of the data (23.3%-15.3%) slightly decreasing with increasing components (**Figure 3.9** top). This indicates that the data is relatively evenly distributed across multiple axes, which makes all axes relevant for interpretation, however, PC1 explaining most of the data is the most relevant for interpretation. MAGIC founders clustered again per species group (**Figure 3.9** top; faded circles) similarly as observed in PCA based metabolic as well as genomic data in **Chapter 2**. This further confirms species allocation of the founders. All offspring accessions are expected to be mosaics of the founders and hybrids, although, some cluster closer to either *A. hypochondriacus* or *A. cruentus* founders and hence are likely more related to these species (**Figure 3.9** top). Most MAGIC founders have offspring accessions clustering close to them, which indicates that these founders seem to contribute relatively equally to the offspring. The exceptions are founders VI044371 and VI044437-A which cluster further from the offspring especially in PC1-2 (**Figure 3.9** top).

For comparison, a PCA of an eight founder *Brassica juncea* MAGIC population is provided including the clustering of 170 RILs based on intron length polymorphism (ILP) polymorphic markers (**Figure 3.9** bottom). This PCA analysis was part of the verification of this newly created MAGIC population and its potential application for mapping QTLs in *Brassica j.* (Yan *et al.*, 2020). It can be observed that the distribution of the founders is almost even among the offspring and not clustered (**Figure 3.9** bottom). This dispersion of the MAGIC lines is what you ideally would want to observe in a MAGIC population PCA graph, and is also used for validating the even genomic distribution of a six founder strawberry MAGIC population (Wada *et al.*, 2017). Our *Amaranthus* PCA slightly contradict these ideal expectations with especially well performing founder VI044371 (**Chapter 2: Figure 1**, **Figure 3.3 A**) likely

contributing less to the offspring (**Figure 3.9** top). This is not directly concerning, since in a Maize (*Zea mays*) MAGIC population QTL could be successfully mapped, despite PCA results not always showing this ideal dispersion when plotted using centromeric SNPs (Dell'Acqua *et al.*, 2015). Data structure was lacking for this MAGIC population when plotting PCA with telomeric SNPs (Dell'Acqua *et al.*, 2015). Since most of our MAGIC founders have offspring accessions distributed among them (**Figure 3.9** top), it is still close to what would be an ideal dispersion of MAGIC lines. This suggests a relatively even genomic distribution of most founders (except VI044371) among the population and therefore sufficient quality genotyping by STITCH for QTL mapping.

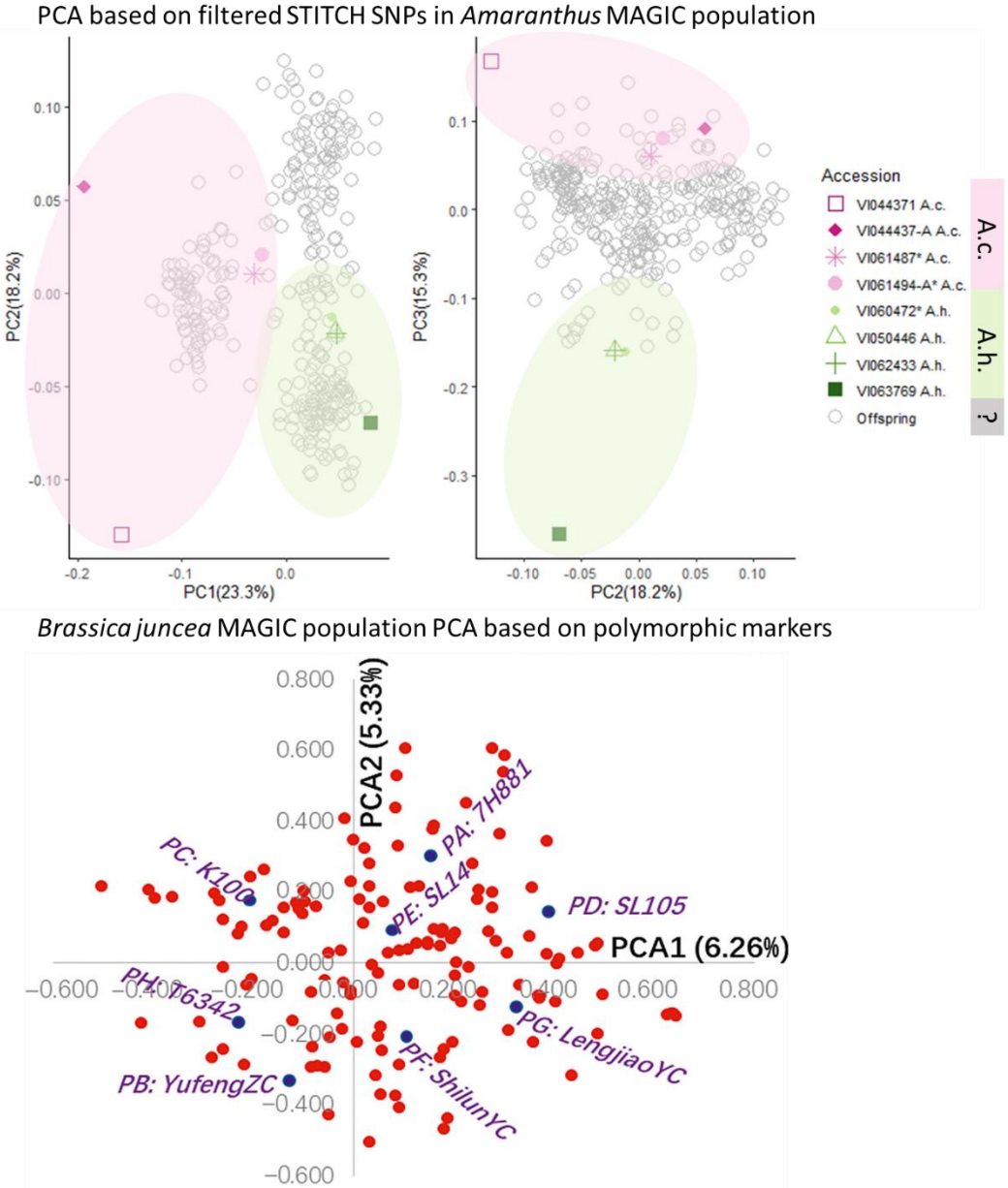



**Figure 3.9.** Principal component analysis (PCA) of MAGIC founders and offspring in our analysed *Amaranthus* (top) and an example of a *Brassica juncea* (Yan et al., 2020: Fig. 2d) (bottom) population. PCA of the *Amaranthus* (top) MAGIC population is based on SNPs imputed with STITCH (Davies *et al.*, 2016) after filtering (info score<0.4, MAF<0.05 and MISS>0.01) while the *Brassica j.* PCA is based on 346 intron length polymorphism (ILP) polymorphic markers. The *Brassica j.* PCA provides a clear example in which the MAGIC founders (blue) are evenly distributed among the MAGIC offspring (red) with no clustering. Our *Amaranthus* MAGIC founders contribute relatively equally to the MAGIC offspring, however, some exceptions are founders VI044371 and VI044437-A (PC1-2). *Amaranthus* MAGIC founders cluster per species group (faded circles), with MAGIC offspring being hybrid mosaics of the founders of which some are more related to a certain species.

## QTLs could be mapped for pleiotropic leafy yield traits

In order to identify which genomic regions could be responsible for the recorded numerical growth-related data, an attempt to map QTL for the yield traits was performed. QTLs were identified using the SNPs called in the MAGIC founders using Longshot (Edge and Bansal, 2019) filtered for "PASS", QUAL > 50 and DP > 10 (manuscript in appendix) which were then imputed in the offspring using STITCH (Davies *et al.*, 2016) and further filtered for enhanced call rates (**Figure 3.8**). For dry and fresh leaf yield traits QTL could be identified on chromosomes 9 and 10 (**Figure 3.10**) using QTL approaches described in Scott et al., (2021). At the chromosome 9 QTL, four SNPs each for dry and fresh leaf weight were found to have a significant association with the trait (-logP>6.15 and -logP>6.04 respectively). For the chromosome 10 QTL, 13 SNPs were significant (-logP>6.15) for dry weight, and only a peak of SNPs approaching significance for fresh weight. Identifying matching QTL for the two leaf yield traits confirms the previously suggested pleiotropy of dry with fresh leaf weight. This is as expected due to the leaf weight traits being based on the same biological parameter and being highly correlated (**Figure 3.4**; 0.94 correlation coefficient with a P=~0 significance (**Table 3.1**)). No QTLs could be mapped for chlorophyll content reflecting the lack of significant variation among the MAGIC offspring for this trait (**Table 3.2**). Heritability of the traits was estimated at 0.23 for dry weight, 0.15 for fresh leaf weight and 0.14 for chlorophyll. This indicates that dry leaf weight is more strongly influenced by genetic factors, while fresh leaf weight and chlorophyll content are influenced to a lesser extent. Findings comply with other leaf amaranth studies. Heritabilities of 0.56 (dry) and 0.08 (fresh), show the environmental effects on fresh biomass (Gerrano, van Rensburg and Adebola, 2015).

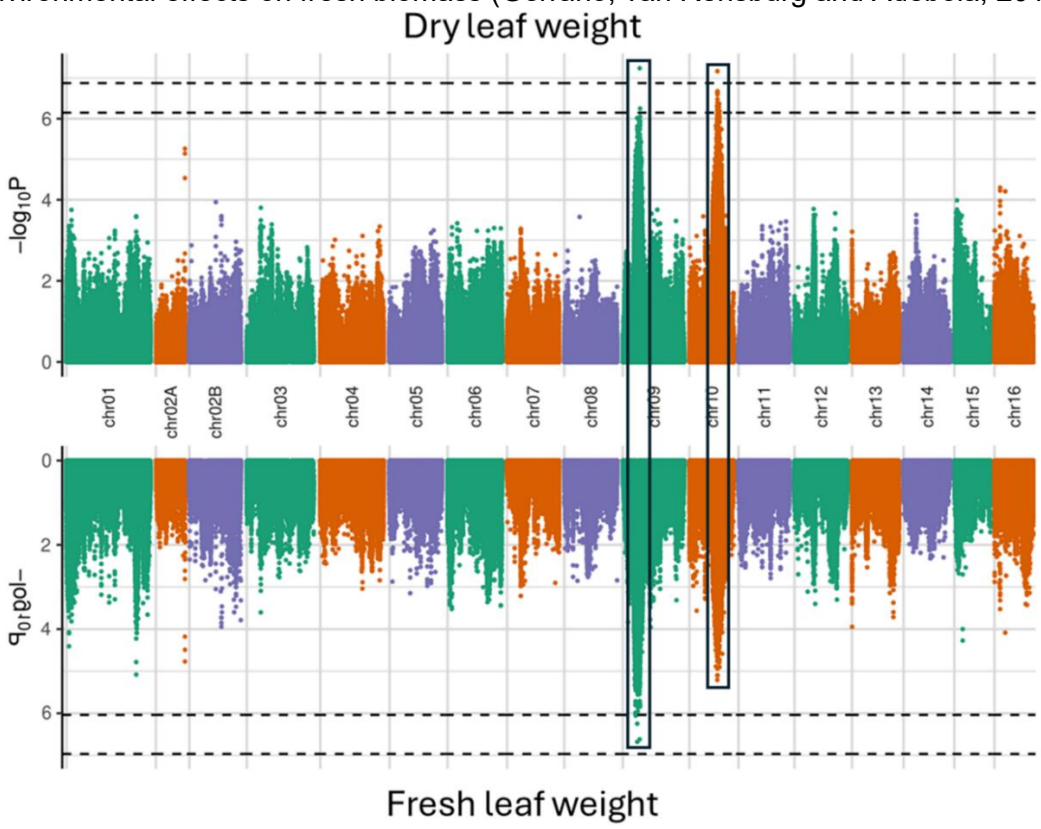



**Figure 3.10**. QTL mapped on chromosomes 9 and 10 for leaf yield traits, i.e., dry and fresh weight. At chromosome 9 significant SNPs are found for both traits. At chromosome 10, SNPs were only found to be significant for dry weight, while for fresh weight SNPs were approaching significance. Dashed lines indicate significance thresholds: a nominal per-marker threshold (-logP=6.15 for dry and 6.04 for fresh weight) and a more stringent genome-wide threshold based on phenotype permutations (-logP=6.88 for dry and 6.96 for fresh weight). Proportion of variance explained for QTL on chromosome 9 and 10: 14%, 6% (dry) or 19%, 10% (fresh).

In order to investigate which SNP allele would be preferable for generating higher leaf yield (i.e. the beneficial allele), leaf yield performance is plotted per genotype for the most significant SNPs at the two identified QTL. The three SNPs which were found to be most strongly significantly associated (i.e. highest p-value) with either one or both leafy traits at each QTL were selected (**Table 3.9**). Using the dosage outputs from STITCH (Davies *et al.*, 2016), the reference and alternative alleles could be identified for all 303 RILs. The leaf yield performance, i.e. dry and fresh weight, is plotted for each of the six most significant SNPs, per genotype. It was found that leaf weight was always significantly different between lines homozygous for the reference allele versus those homozygous for the alternative allele (**Figure 3.11**). This means that the favourable allele which generates higher leaf yields could be identified for all six SNPs. For most SNPs (chr09:5961287, chr09:6952972, chr10:11555550 and chr10:11556036) the Arusha reference allele (Ma *et al.*, 2021) is favourable, resulting in significantly higher dry as well as fresh leaf yield. For locations chr09:7123246 and chr10:11566886 presence of the alternative haplotype allele would be beneficial for leaf performance (**Figure 3.11**, **Table 3.9**).

**Table 3.9.** Three most significant SNP locations with the highest p-values are given for each of the identified QTL. The leafy trait they are significant for (fresh (FW) and dry weight (DW)), p-values, Arusha reference (Ma et al., 2021) allele and alternative allele in the offspring haplotypes are indicated. Whether it is the preferred or non-preferred allele resulting in higher or lower leaf yields according to **Figure 3.11** is indicated with green and orange colouring respectively.

| **Chromosome and bp location** | **Significant for trait** | **Log P-value** | **Allele** | |
|---|---|---|---|---|
| | | | **Counted (reference)** | **Alternative in haplotypes** |
| chr09:5961287 | FW | 6.68 | C | T |
| chr09:6952972 | FW<br>DW | 6.62<br>7.25 | A | G |
| chr09:7123246 | DW | 6.25 | T | C |
| chr10:11555550 | DW | 7.17 | A | T |
| chr10:11556036 | DW | 6.65 | T | C |
| chr10:11566886 | DW | 6.68 | A | C |

To identify potential candidate genes responsible for differences in dry as well as fresh leaf weight, which are expected to be the same genes according to biology, QTL sizes should be set. To narrow down QTL size boundaries for candidate gene identification, the 100 most significant SNPs, i.e., lowest p-value, for each leafy trait were overlapped with each other at each QTL. This resulted in 51 overlapping SNPs at chromosome 9 with those being located between base pairs 5336338 to 7138794. At the chromosome 10 QTL, 72 SNP locations overlapped between dry and fresh weight located between 10925239 to 12393262 base pairs. Now these QTL base pair boundaries are identified, it is interesting to investigate genes, significant SNP alleles and structural variants (SVs) located between these boundaries due to their putative causal effects on leaf yield. In order to do this, the MAGIC founders of which genes, SNPs and SVs are known, could also differ for leaf weight traits.

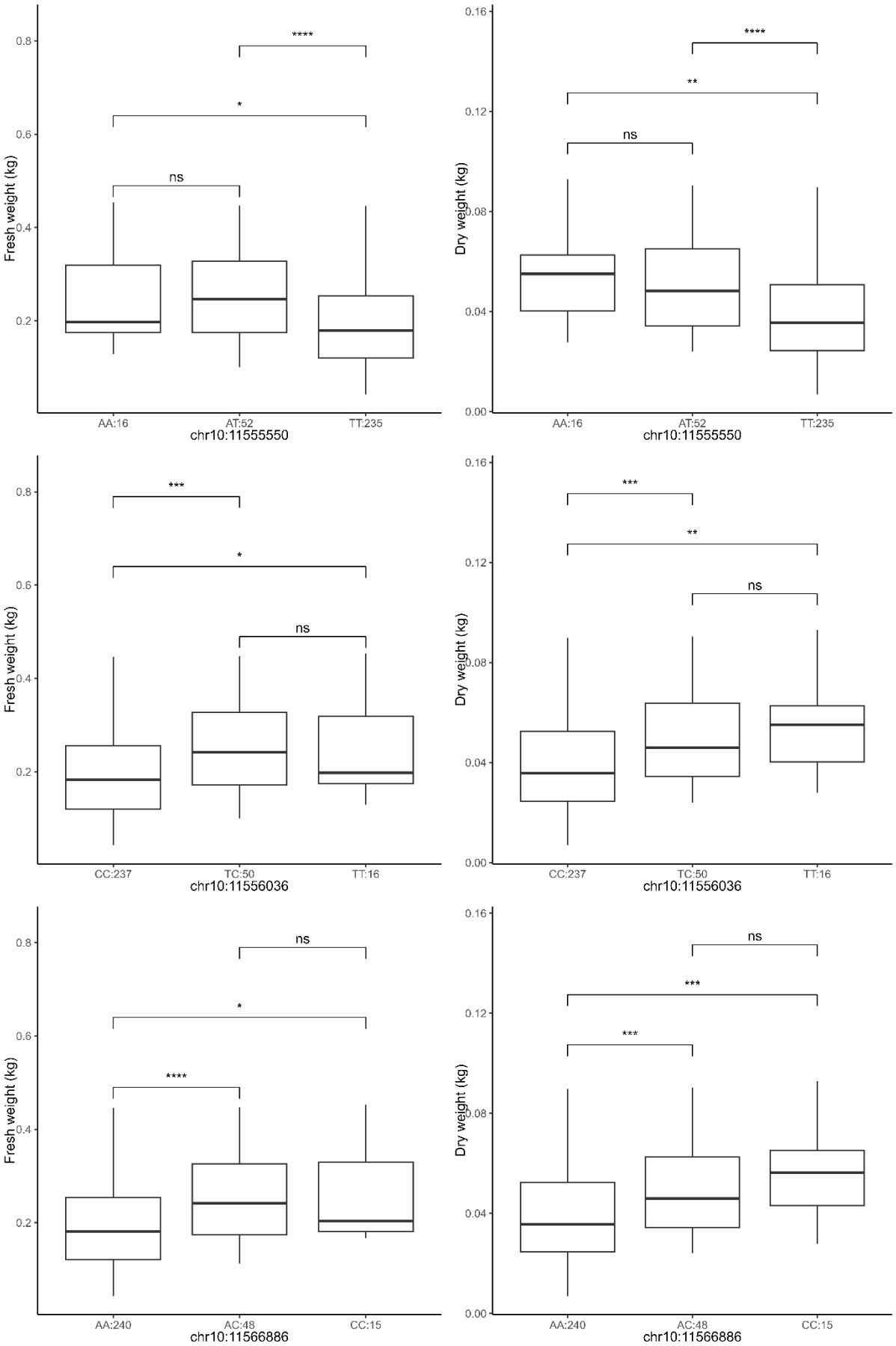



**Figure 3.11. A)** Fresh (left) and dry weight (right) phenotypes of the MAGIC offspring, plotted per genotype (reference (Ma *et al.*, 2021) and alternative) of the three most significant SNP locations (**Table 3.9**) at chromosome 10. Significance ($p<0.05$) was tested with a Wilcoxon rank-sum test. Number of total 303 RILs is provided behind the genotype, e.g., AA:240.

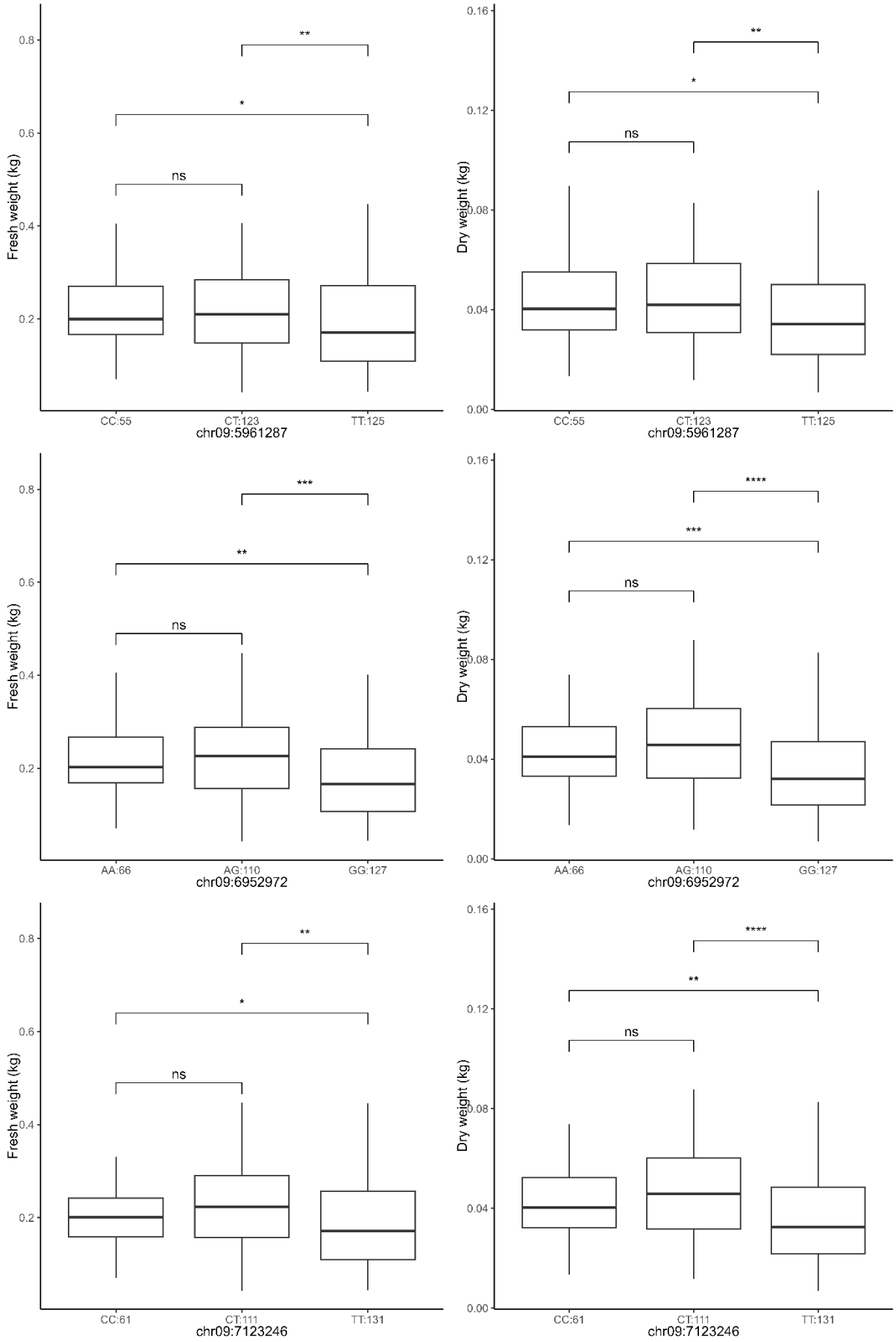



**Figure 3.11. B)** Fresh (left) and dry weight (right) phenotypes of the MAGIC offspring, plotted per genotype (reference (Ma *et al.*, 2021) and alternative) of the three most significant SNP locations (**Table 3.9**) at chromosome 9. Significance ($p<0.05$) was tested with a Wilcoxon rank-sum test. Number of total 303 RILs is provided behind the genotype, e.g., CC:61.

To examine if there is a link between leaf performance and the genes, SNPs and SVs of interest, fresh and dry leaf weights of only the MAGIC parents and Arusha in all field trial locations is plotted. V1062433 performs lower in comparison to the other accessions for both leafy traits, while V1044371 performs higher for fresh leaf weight only (**Figure 3.12**). It is therefore possible that certain preferred genes, SNP alleles or SVs are either present or not in these two accessions in comparison to the other founders, with different alleles in genes potentially making different proteins. Additionally, there might be varying gene activity or altered regulation among the founders, and this change in regulation can be observed by that genes are higher or lower expressed measured on the regulatory transcript sequence.

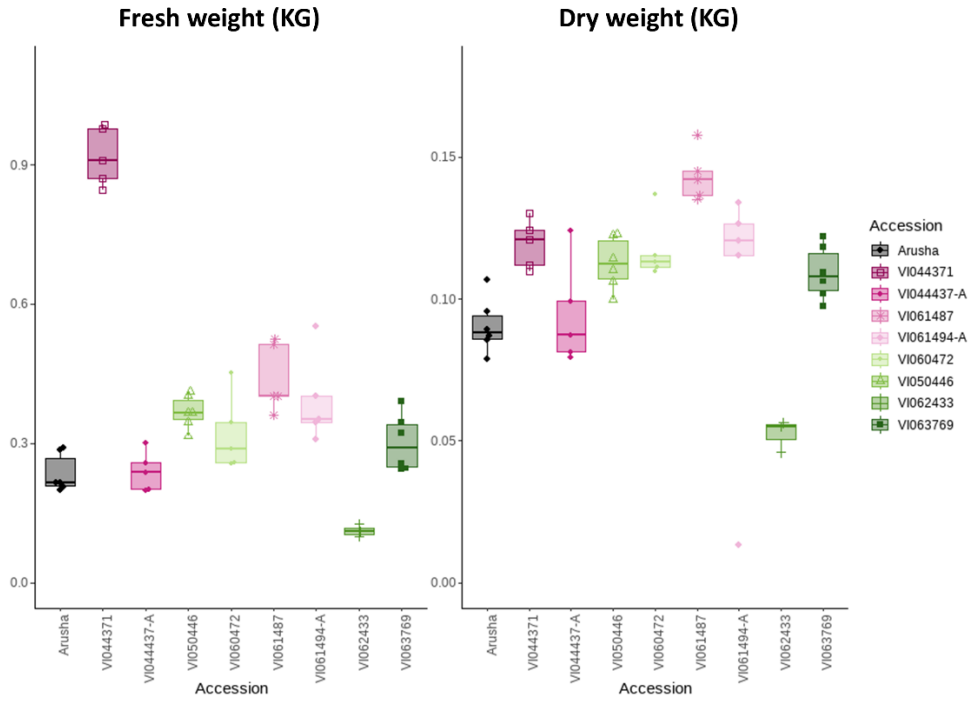



**Figure 3.12.** Fresh (left) and dry (right) weight leaf yield performance of MAGIC parental accessions and Arusha during field trials in south Africa. Here the same data of **Figure 3.3 A** is visualized but only of the MAGIC founders and Arusha. The measurements of three repeats at two locations are shown as six individual observations per accession, with boxplots to illustrate the data distribution. V1062433 performs lower than the other accessions and V1044371 performs higher for fresh leaf weight only.

To investigate if the MAGIC founders contain the preferred alleles at the three most significant SNP locations (**Table 3.9**), their presence was addressed in the VCF file created by SNP calling with longshot after filtering for "PASS", QUAL > 50 and DP > 10 (manuscript in appendix). All accessions had only four or five times the preferred allele, just like the reference Arusha, with the exception of founder V1063769 counting the preferred allele only twice out of six SNP locations (**Table 3.10**). Of the three SNPs located in annotated genes, seven accessions contained the preferred allele twice and Arusha and V1063769 contained it once (**Table 3.10**). This suggests that these results cannot very directly be correlated with the observed leaf yields in which V1062433 performs lower and V1044371 higher while V1063769 as well as Arusha perform average in comparison to the other accessions (**Figure 3.12**). What should be kept in mind is that **Table 3.10** focusses on six SNP locations only, out of many which are significantly associated or potentially involved in enhanced leaf yields. Mutations over generations or sequencing errors could have occurred while generating data. Additionally, there can be many other entire candidate genes, or varying alleles among these genes, impacting leaf yields.

**Table 3.10** Three most significant, i.e. highest p-values, SNP locations and alleles are given for the two identified QTL for Arusha (Ma et al., 2021) (reference and alternative allele) and each of the MAGIC founder haplotypes. Preferred or non-preferred allele resulting in higher or lower leaf yields in the offspring (**Figure 3.11**) is indicated with green and orange colouring respectively. If a SNP is located in an annotated gene of the Arusha reference (In gene), and the number of genes between the SNP location on the row below (No. genes) is provided.

| Chromosome and bp location | Arusha ref. | Arusha alt. | In gene | No. genes | V1044371 | V1044437 | V1061487* | V1061494* | V1060472* | V1050446 | V1062433 | V1063769 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr09:5961287 | C | T | Yes | 96 | C/C | C/C | T/T | T/T | C/C | C/C | C/C | C/C |
| chr09:6952972 | A | G | No | 12 | A/A | A/A | G/G | G/G | A/A | A/A | A/A | A/A |
| chr09:7123246 | T | C | Yes | - | T/T | T/T | C/C | C/C | T/T | T/T | T/T | T/T |
| chr10:11555550 | A | T | No | 0 | A/A | A/A | A/A | A/A | A/A | A/A | A/A | T/T |
| chr10:11556036 | T | C | No | 2 | T/T | T/T | T/T | T/T | T/T | T/T | T/T | C/C |
| chr10:11566886 | A | C | Yes | - | C/C | C/C | C/C | C/C | C/C | C/C | C/C | A/A |

To broaden our insight into the genetic differentiation among the nine accessions beyond just six SNPs and attempt to identify potential candidate genes or structural variation responsible for varying leaf yield, annotated genes and SVs at the QTL were investigated. The numbers of genes (**Figure 3.13**) as well as SVs (**Table 3.11**) located between the QTL boundaries were similar when comparing V1062433 and V1044371 to the other founders. The only exception was Arusha gene 'Amacr09g05810' which was absent in VI062433 but highly expressed in the other lines. This gene is annotated as part of the panther (PTHR27001) family 'OS01G0253100 PROTEIN', which is typically involved in protein phosphorylation and located in the plasma membrane (**Figure 3.13**). Hence, it could play a role in plant development. Nevertheless, it remains uncertain whether the absence of this gene alone is responsible for the observed variation in leaf yield.

Another way to potentially allocate candidate genes is by observing if there is a difference in gene regulation between the founders, especially if that variation in regulation is correlated with the trait of interest. Hence, expression profiles of genes located at QTL based on transcript sequences from leaves of the MAGIC founders and Arusha were plotted and ordered by the median value of the trait for which significant associations were found for both QTL, i.e. median dry weight (**Figure 3.12**). Most genes located at the QTL were (non-)expressed at a similar level across the founder lines and ordering the accessions by dry leaf weight observations did not result in clear correlation patterns between gene expression and the trait (**Figure 3.13**). No variation in expression profiles, just as no correlation with dry leaf yield observed, means that solely gene regulation likely did not contribute to the differences in leaf yield observed among the founders and Arusha (**Figure 3.12**). Therefore, all genes at the identified QTL could potentially be causal for the generally high leaf performance of all MAGIC founders (**Figure 3.3 A**). What could be a causal factor underlaying differences in leaf yield among the nine accessions (**Figure 3.12**) not based on gene presence/absence or regulation is allelic variation in the gene sequence, leading to variation in protein sequence which can then impact activity. Such SNP alleles or mutations within a gene cannot be observed in **Figure 3.13** but can underpin varying leaf performance.

**Figure 3.13. A)** Gene expression profiles of genes located within the two QTL at chromosome 9 identified for leaf yield. The *A. cruentus* reference (Ma *et al.*, 2021) gene annotations are given, with the expression of orthologous genes identified with Orthofinder (Emms and Kelly, 2019) of the MAGIC founders towards Arusha. Accessions are ordered from left to right with respectively lowest to highest median dry weight as shown in boxplots of **Figure 3.12**. Most genes located at the QTL are (non-)expressed simultaneously in all MAGIC founder accessions and leaf tissue of Arusha. Gene Amacr09g05810 is highlighted due to its absence in accession V1062433 for which lower leaf yields than the other accessions were observed (**Figure 3.12**).

**Figure 3.13. B)** Gene expression profiles of genes located within the two QTL at chromosome 10 identified for leaf yield. The *A. cruentus* reference (Ma *et al.*, 2021) gene annotations are given, with the expression of orthologous genes identified with Orthofinder (Emms and Kelly, 2019) of the MAGIC founders towards Arusha. Accessions are ordered from left to right with respectively lowest to highest median dry weight as shown in boxplots of **Figure 3.12**. Most genes located at the QTL are (non-)expressed simultaneously in all MAGIC founder accessions and leaf tissue of Arusha.

To see if genes located at QTL are of relevance for *Amaranthus* leaf tissue, their conservation as well as regulation is addressed. What could be observed is that if a gene is non-existent in certain accessions, it is often non or lowly expressed other accessions. Some examples at the chromosome 9 QTL are the heavy metal associated gene Amacr09g05950, most germin like protein genes (Amacr09g06600 - Amacr09g06640), cupin Amacr09g06690 and DNA-directed RNA polymerase Amacr09g06790 genes. For the QTL at chromosome 10 this pattern of non-existence coinciding with low expression occurred for gene Amacr10g06010 (**Figure 3.13**). This reveals that the relevance of these genes in leaf tissue is potentially low, due to low expression as well as low pressure for conservation amongst individuals. However, what should be kept in mind is that we are looking at leaf gene expression only at one time point, so these genes might be higher expressed in different tissues or under different environmental conditions. Germin like protein gene Amacr09g06790 is the only exception of a highly expressed gene in leaf tissue but not conserved in one third of the accessions (**Figure 3.13**). This is an example of a gene likely being of relevance to leaf tissue in some amaranths although not conserved in others. Notably, gene sequence variation could lead to different phenotypes and genes might be missed in accession sequences during annotation which is both not detected in this analysis. Hence it remains difficult to identify which genes are important to leaf tissue of amaranths based on conservation and regulation in leaf tissue at a single time point only.

To look into species specific regulation, expression patterns differing between species were investigated. Some gene expression profiles are subtly species specific, like a slightly higher expression of the chitin recognition protein gene Amacr09g05820 for *A. hypochondriacus*, and of the cysteine rich secretory protein related gene Amacr10g06640 for *A. cruentus* accessions (**Figure 3.13**). This indicates that there are minor gene expression differences which may result in species-specific adaptations or regulatory mechanisms. However, to confirm this, phenotypes should be evaluated, and conclusions cannot be made on solely gene regulation patterns.

Investigating the SVs at the QTL is important since they could play a role in obtaining higher leaf yields due to the inheritance of favourable sequences present in SVs in some accessions but not in others. Many insertions, deletions and copy number variations could be found between the QTL boundaries for the MAGIC founders (**Table 3.11**). Only limited inversions and translocations were present, with the latter always towards a different chromosome apart from one in V1063769 which translocated within the 9$^{th}$ chromosome (chr09:13721509-chr09:6830578) (**Table 3.11**). It was previously observed that V1062433 performed lower for leaf yield traits than the other founders and Arusha (**Figure 3.12**), but this accession has SVs counts and lengths comparable with the other accessions (**Table 3.11**). Similarly for accession V1044371 which performed higher for fresh leaf weight. V1044371 has similar numbers of SVs with comparable lengths to the other founders, especially similar to the closely related V1044437. Looking at the SVs unique to these accessions, three copy number variations, two translocations, one insertion and one deletion were found for V1062433, and one inversion plus and insertion for V1044371 at the chromosome 9 QTL (**Table 3.11**). One 52 bp length deletion for V1062433 and three deletions ranging 69 to 340 bp for V1044371 were at the chromosome 10 QTL (**Table 3.11**). Sequence differences in these SVs at QTL specific to these MAGIC founders are of interest since they could be causal to the lower or higher leaf yield observed in accessions V1062433 and V1044371 respectively.

**Table 3.11.** Structural variants (SVs) (<100 kb) found within the two QTL identified for leaf yield traits on chromosome 9 and 10. Presence absence variations (PAV) are insertions (INS) and deletions (DEL), for which numbers slightly increase with a more distant relatedness to the *A. cruentus* reference (Ma et al., 2021) the SVs were identified towards (accessions in grey shading are *A. hypochondriacus* species). Other SVs are inversions (INV), copy number variations (CNV) and translocations with for the later the length being unknown. The number of unique SVs for each accession, i.e. not found in other MAGIC founders, is indicated which was notably high for V1063769. * indicates an accession showing admixture (Lin et al., 2022). The total number of SVs of all accessions are provided per type of SV and per individual QTL.

| | | chr09 | | | | | | chr10 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PAV | | | | | | PAV | | | | | |
| | | INS | DEL | INV | CNV | TRANS | Unique | INS | DEL | INV | CNV | TRANS | Unique |
| V1044371 | No | 41 | | 1 | 6 | | 2 | 77 | 29 | | 5 | | 3 |
| | avg length | 170 | | 526 | 122 | | | 558 | -1162 | | 52 | | |
| V1044437 | No | 42 | | 1 | 6 | | 2 | 77 | 28 | | 6 | | 1 |
| | avg length | 187 | | 531 | 122 | | | 596 | -1202 | | 46 | | |
| V1061487* | No | 72 | 35 | | 13 | 1 | 7 | 82 | 28 | | 7 | | 4 |
| | avg length | 979 | -736 | | 55 | | | 555 | -1201 | | 44 | | |
| V1061494* | No | 72 | 35 | | 17 | | 9 | 96 | 51 | | 10 | | 40 |
| | avg length | 827 | -745 | | 48 | | | 623 | -914 | | 178 | | |
| V1050446 | No | 127 | 125 | 1 | 20 | 6 | 3 | 106 | 73 | | 19 | | 2 |
| | avg length | 426 | -609 | 267 | 41 | | | 459 | -755 | | 34 | | |
| V1060472* | No | 126 | 126 | 1 | 21 | 7 | 9 | 110 | 72 | | 19 | | 2 |
| | avg length | 426 | -604 | 267 | 43 | | | 446 | -767 | | 39 | | |
| V1062433 | No | 129 | 126 | 1 | 22 | 5 | 7 | 109 | 72 | | 17 | | 1 |
| | avg length | 419 | -630 | 267 | 68 | | | 447 | -760 | | 36 | | |
| V1063769 | No | 130 | 123 | 2 | 20 | 7 | 47 | 126 | 88 | | 18 | 1 | 61 |
| | avg length | 627 | -618 | 376 | 39 | | | 449 | -666 | | 116 | | |
| Total | No per SV | 739 | 570 | 7 | 125 | 26 | | 783 | 441 | 0 | 101 | 1 | |
| Total | No overall | | | 1467 | | | | | | 1326 | | | |

To get an insight into the genomic variation at the identified QTL the parental pan-genome (manuscript in appendix), i.e. the Arusha reference highlighting where the founders are different, was visualized. To obtain clear visualizations, the QTL base pair boundaries were drastically narrowed down. A 1000 base pair window around the most significant SNP of each QTL (**Table 3.9**) was visualized. For the chromosome 9 QTL this resulted in 30 variants (28 SNPs and 2 SVs) (**Figure 3.14 A**) of the MAGIC founder sequences to be present in the pan-genome. For the chromosome 10 QTL 16 parental variants (15 SNPs and one SV) (**Figure 3.14 B**) are within this region. The SVs were six and four base pairs long at the chromosome 9 QTL, and 11 base pairs at the chromosome 10 QTL. Due to variation captured in the pan-genome, the Arusha reference genome was split into 60 parts (28 SNPs and 32 larger blocks of base pairs) at the chromosome 9 QTL (**Figure 3.14 A**) and into 28 parts (15 SNPs and 14 larger blocks) at the chromosome 10 QTL (**Figure 3.14 B**). The larger blocks have likely no biological meaning since these are just reference sequence separated by the SNPs or SVs of the MAGIC founders which is the variation of interest. These founder variants at QTL are nicely shown in **Figure 3.14**, however, we do realise that only a small region to demonstrate the pan-genome is represented and that the causal variation underlying the QTL may well be outside this region.

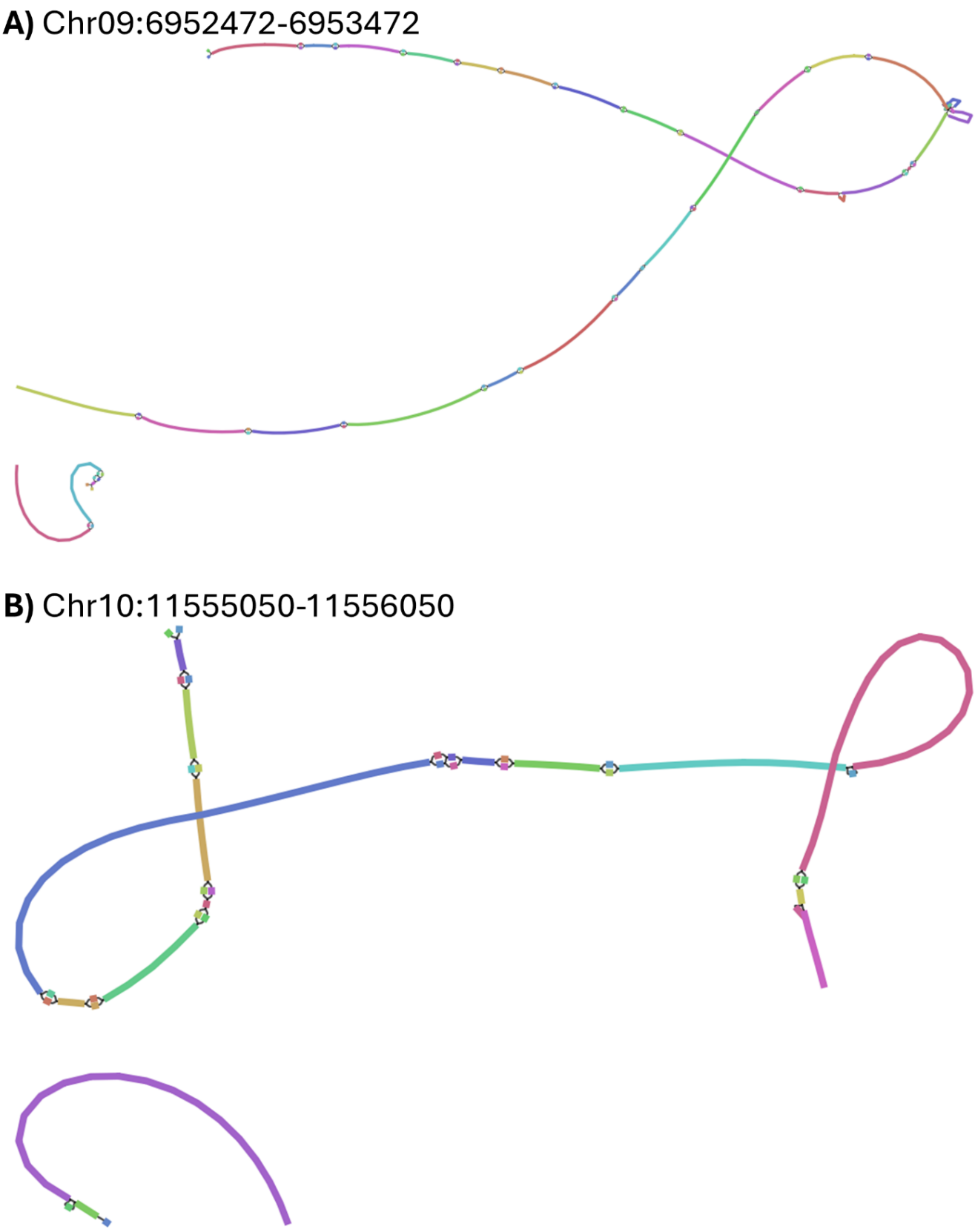



**Figure 3.14** Variation of the MAGIC founders in an *Amaranthus* graphical pan-genome (manuscript in appendix) towards the 1000 reference genome base pairs surrounding the most significantly associated SNPs at two identified leafy trait QTL (**Table 3.9**). **A)** Chromosome 9 and **B)** chromosome 10 QTL with the visualized *A. cruentus* reference (Ma et al., 2021) base pair window indicated.

It can be observed that more genomic variation is present at the chromosome 9 QTL (**Figure 3.14**), which is the QTL for which SNPs were significantly associated with both leaf weight traits. This means that the chromosome 9 QTL region may harbour a higher diversity of functional variants that contribute to variation in leaf yield, potentially making it a more relevant and influential locus. This is further supported by the proportion of variance explained (PVE) at this locus being higher in comparison to PVE values for the chromosome 10 QTL. For dry and fresh leaf weight respectively, 14% and 19% PVE were calculated at chromosome 9 versus 6% and 10% PVE at the chromosome 10 QTL (**Figure 3.10**). Additionally, also more SVs are found at the complete chromosome 9 QTL compared to at the chromosome 10 QTL, i.e. 1467 instead of 1326 (**Table 3.11**). QTL with more genomic variation, i.e., higher genetic diversity, can potentially explain more phenotypic variance resulting in higher PVE. Although, what should not be forgotten is that the chromosome 9 QTL is of longer length (1802456 base pairs) compared to the chromosome 10 QTL (1468023 base pairs), which could additionally increase numbers of SVs per reference sequence length. In general, all genomic variation captured at the QTL, here partially visualized with the parental pan-genome (**Figure 3.14**), could be responsible for variation in leaf yield and, if so, this variation is highly relevant for trait improvement in breeding programs.

# Discussion

## Phenotypic variation among the MAGIC offspring

The MAGIC founders vary for various descriptor traits, with these observed traits notably further segregating in the offspring RILs (**Figure 3.2**). This is a first indication that the newly developed *Amaranthus* MAGIC population has captured phenotypic diversity and is segregating for this diversity, containing key characteristics needed to be able to map QTL (Arrones *et al.*, 2020). There is variation in performance of the offspring as well as founder accessions when assessing numerical growth data (i.e. dry or fresh leaf weight and chlorophyll) in different field trial locations (**Figure 3.3 A**). Again, segregation was clearly observed for chlorophyll, but not for dry and fresh leaf yield, with MAGIC offspring often performing worse than the founders (**Figure 3.3 A**). This is likely due to recombination which took place while creating the MAGIC population. Beneficial allelic combinations for relevant traits present in the selected founder lines likely have been broken up, which is a common phenomenon observed in MAGIC populations (Huang *et al.*, 2015; Scott *et al.*, 2021).

It was observed that the offspring accessions do perform significantly differently in the various field trial locations when assessing numerical growth data (**Figure 3.3 A**). Most notably is the high performance of accessions grown in Nelspruit versus the low performance of accessions grown in Mafikeng. This is likely due to the climatic conditions being most divergent between these two regions. Nelspruit is cooler, wetter and more humid with mean temperatures of 17.75 °C annually or 21.02 °C over summer and mean rainfall of 61.13 mm/month annually or 103.47 mm/month over summer, while Mafikeng is warmer and drier with mean temperatures of 18.35 °C annually or 22.25 °C over summer and mean rainfall of 44.31 mm/month annually or 76.67 mm/month over summer (Benhin, 2006) (**Table 3.13**). It could be argued that Nelspruit is the preferred growth climate for most amaranth accessions due to it being located in the province with the highest precipitation compared to

all the other field trial locations (Benhin, 2006) (**Table 3.13**). Mafikeng is clearly the least preferred growth location for most accessions (**Figure 3.3 A**), with not even any significant variation observed within the field trial location (**Table 3.3**) likely due to the low yield. Again precipitation is the most outstanding difference with Mafikeng being the driest in comparison to the other field trial locations with the lowest rainfall (Benhin, 2006) (**Table 3.13**). Sufficient precipitation and a wet and humid climate seem to benefit leafy amaranths in the MAGIC population. Hence, additional research including irrigated field trials on top of the current summer drought trials would be useful to generate knowledge of performance of the RILs in more advanced growth conditions.

The highly positive correlation of non-categorical traits suggests pleiotropy (**Figure 3.4**, **Table 3.1**). This could also be observed in the QTL we managed to map for dry and fresh leaf yield traits (**Figure 3.10**). These findings were very much expected since the two leaf weight traits share a biological basis. The observed correlations between dry and fresh leaf yield are therefore anticipated, unless there were drastically different water contents between the accessions, which our results suggest was not the case. The expected leaf correlations are further confirmed in literature, as it is commonly observed that leaf related traits are (potentially) pleiotropic in crops. Like leaf length being correlated to leaf width in spinach (*Spinacia oleracea L.*) (Liu *et al.*, 2021) which is a species closely related to Amaranth. But also, dry leaf weight of pepper is pleiotropic with leaf area ($cm^2$) (Alimi *et al.*, 2013) and many common leaf-related traits together with chlorophyll content could be mapped in party shared QTL in soybean (Yu *et al.*, 2020). We could not map any QTL for chlorophyll probably because no variation was observed in the MAGIC offspring for this trait (**Table 3.2**). However, the anticipated strong ($p =\sim 0$) correlation between dry and fresh leaf weight (**Table 3.1**) has definitely resulted in evident pleiotropy with QTL identified at two genomic regions for both traits.

Fresh leaf yield, which is the most relevant trait for our work, was the only trait for which a GxE effect was observed. This GxE effect was less significant compared to other model effects, i.e., $p = 0.0012$ for GxE compared to $p < 0.0001$ for the other fixed model effects (**Table 3.2**). GxE is therefore present but limited, and ideally should be accounted for by using different breeding programmes for varying climates as is suggested in literature (Falconer, 1952). However, the performance of all three non-categorical traits is highly significantly ($p < 0.0001$) correlated (**Table 3.1**, **Figure 3.4**) and no GxE effect was found for dry leaf weight and chlorophyll. Hence when attempting to select the best performing offspring RILs the overall performance was assessed of the accessions over the multiple trial locations. The identification of well performing accessions was successful, with 51 accessions being selected for substantial dry as well as fresh leaf yield despite their climatic growth environment (**Table 3.4**, **Figure 3.5**). EBV per accession estimated individually for dry and fresh leaf yield could further aid decisions regarding selection of well performing offspring RILs (**Supp table EBV offspring**). These selected and newly developed inbred lines can directly be used by small scale farmers in diverse agro-ecologies but also as novel accessions for further breeding. Additional irrigated field trials could further aid the selection of well performing RILs. If more evident GxE effects occur in such trials, accessions with higher leaf yield under wet versus dry growth conditions can be distinguished. Appropriate lines can then be selected for farmers as well as the *Amaranthus* breeding programme.

## Improved genotyping of the MAGIC RIL sequences

It was shown that the MAGIC RIL sequences were of adequate quality, with a median mapping coverage of around 6.5 (**Table 3.5**) and stable mapping with not much variation in coverage between the accessions (**Figure 3.7**). This provided a good foundation for genotyping of the RILs. Genotyping the RILs using two different approaches showed that call rates improved when using STITCH compared to GATK haplotype caller, especially once further filtering the STITCH imputed SNPs (**Figure 3.8**, **Table 3.7**, **Table 3.8**). Improved call rates in addition to clear clustering of accessions in the PCA based on STITCH genotyping (**Figure 3.9** top) suggested STITCH to be the preferred method to continue analysis. This is complementary to previous findings of STITCH causing improved call rates and therefore being more reliable for genotype imputation even in very low-coverage (mean 0.304x) sequencing data of wheat RILs (Scott *et al.*, 2021). Ideally the call accuracy should additionally be investigated to verify if the calls are truly present, however, unfortunately this was not possible due to limitations in the current data. Conducting PCA after genotyping with STITCH aids our conclusion that most of the MAGIC founders contribute to the offspring, possibly excepting founders VI044371 and VI044437-A (**Figure 3.9** top). Unfortunately, this reveals the high yielding founder VI044371 (**Chapter 2: Figure 1**, **Figure 3.12**) potentially less contributing, with beneficial properties possibly lost in the offspring. Overall, the PCA is a valuable verification of a good quality newly created *Amaranthus* MAGIC population with potential for mapping QTL, like implemented in *Brassica j.* (Yan *et al.*, 2020) (**Figure 3.9** bottom) and strawberry (Wada *et al.*, 2017) MAGIC populations. Lastly, the PCA reveals that the population clusters per species groups, and hybrid offspring accessions of which some are more related to either of these species can likely be assigned (**Figure 3.9** top).

When investigating inbreeding of the RILs after genotyping we discovered they are quite heterozygous with even negative inbreeding coefficients (F) (**Figure 3.8 B**, **Table 3.8**). Therefore, results suggest outbreeding of the offspring compared to the fairly inbred parental lines (0.54 mean F and 88% mean homozygosity). An example of a MAGIC population in which founders are less heterozygous compared to the offspring RILs after six generations of crossing is a maize (*Zea mays*) population (Dell'Acqua *et al.*, 2015). In contrast to the maize MAGIC population, where increased heterozygosity in offspring results from homozygous parents and the crossing design, the higher level of heterozygosity in offspring of a Sorghum (*Sorghum bicolor* (L.) Moench) MAGIC population can potentially be explained by contamination and poor genotyping (Ongom and Ejeta, 2017). For both examples, the MAGIC founders were already highly inbred and homozygous at the start, with average proportions of heterozygotes below 5% (Dell'Acqua *et al.*, 2015; Ongom and Ejeta, 2017). Our own *Amaranthus* MAGIC founders were also very homozygous (ranging 81% to 95%) which could have contributed to outbreeding of the MAGIC RILs. A mean homozygosity of ~72% still indicates the RILs are inbred (**Table 3.8**) and we generated a sufficient quality population to map QTL (**Figure 3.10**). However, ideally the offspring accessions would highly inbred, with for example RILs being 99.74% homozygous in a cowpea (Huynh *et al.*, 2018) and 0.67% heterozygotes in a wheat (Scott *et al.*, 2021) MAGIC population. One way of potentially achieving this would be to conduct more generations of selfing with the offspring lines after their final crossing generation. Since our MAGIC population was only selfed for one generation (F4) after three generations of crossing (F1-F3), a suggestion for future work would be to try to increase homozygosity in the offspring by an advanced number of generations of selfing, e.g. to F8.

Generally, we can conclude that the RILs successfully have been genotyped, and that STITCH could be selected as a preferred SNP imputation method due to improved call rates (**Figure 3.8**, **Table 3.8**), ignoring investigation of accuracy. Additional analysis which could potentially improve our current SNP imputation would be to include the founder genotypes as a reference panel in STITCH. However, in the wheat RILs using a reference panel only marginally improved the imputation accuracy (Scott *et al.*, 2021). Hence it is expected to not hugely enhance current genotyping results. The only benefit of using a reference panel is that it enables exact identification of complete founder haplotype blocks inherited in each offspring accession. This could be beneficial because it could improve the resolution of QTL mapping and enables the identification of beneficial genetic variants contributed by specific founders. Hence, we would suggest it for additional analysis. Another way to strengthen our QTL analysis would be by genotyping the RILs with the structural variation captured in the *Amaranthus* pan-genome described in the manuscript in appendix. If a pipeline is created allowing imputation of the structural variants on top of the current SNP imputation, larger genomic regions responsible for the mapped QTL can potentially be identified. This would benefit resolution and biological interpretation of the QTL by recognizing genetic variation beyond SNPs which can be relevant in amaranth breeding programmes.

## Loci for relevant agronomic traits identified

We managed to successfully identify QTL for the most relevant traits investigated, i.e., dry and fresh leaf weight (**Figure 3.10**). This is a major finding aiding breeding of vegetable amaranth, which is an underutilized crop often lacking adequate agronomic knowledge (Mabhaudhi *et al.*, 2016, 2019; Emmanuel and Babalola, 2021). Previous QTL studies in amaranths were only conducted in grain varieties, mostly focussing on pigmentation related traits (Stetter, Vidal-Villarejo and Schmid, 2020; Winkler *et al.*, 2024). Hence we've obtained novel results which can contribute to the great potentials of leafy amaranth since it is already commonly consumed and provides healthy nutrition in sub-Saharan Africa (Achigan-Dako, Sogbohossou and Maundu, 2014; van Zonneveld *et al.*, 2023).

The suspicion of pleiotropy between the leafy traits (**Figure 3.4**, **Table 3.1**) was confirmed with two identical genomic regions (almost) significantly associated with both traits (**Figure 3.10**). This provides some validity of the identified QTL since both traits map to the same loci, although, this was likely and expected since the traits are correlated. The fact that both leafy traits are associated with the same QTL eases interpretation as well as implementation of this knowledge in the leafy *Amaranthus* breeding programme. The genotypes of the three most significant SNP locations per QTL region (**Table 3.9**) were clearly reflected in the phenotypic variation of the offspring (**Figure 3.11**). Hence, we now have identified which allele is preferable to obtain higher leaf yields. This correlation was less evidentially observed when comparing the founder leaf phenotypes (**Figure 3.12**) with the genotypes of these six SNPs (**Table 3.10**). This could be due to the limited availability of phenotypic as well as genotypic data, i.e. for both only six datapoints, just as data generation issues for the founder genotypes due to mutations or sequencing artifacts. With the clear pattern of certain alleles (**Table 3.9**) likely contributing to improved leaf yields observed in the many offspring RILs (**Figure 3.11**), valuable information regarding beneficial allelic variation is obtained.

This information can directly be used in amaranth breeding practices through marker-assisted breeding or selection, with lines carrying the beneficial alleles being selected for crossing or directly used to accelerate the development of high-yielding cultivars.

Identifying candidate causal genes or structural variants for leaf yield variation was more challenging. Gene annotations as well as expression and structural variation, is only known for the parental lines and not for the offspring. Looking at gene presence/absence, difference in gene regulation and accessions with unique SVs correlated with the observed phenotypic variation could reveal candidate causal genes and SVs. Hence comparisons of this data to phenotypic leaf variation in the parental lines and Arusha was made (**Figure 3.12**). Only one gene (Amacr09g05810: typically involved in protein phosphorylation) was absent in the low performing VI062433 accession, while highly expressed in the other founders (**Figure 3.13**). The potential roles it could play in plant development could be causal for improved leaf yields, however the absence of solely one gene is only a marginal difference which could also be due to gene annotation errors. Ordering gene expression heatmaps by dry leaf weight did not result in clear correlation patterns (**Figure 3.13**), suggesting that gene regulation underlying leaf yield variation could not be observed from our data. Hence, we conclude that all genes located at the QTL could potentially be causal for high MAGIC founder leaf yields (**Figure 3.3 A**). It should not be forgotten that differing of SNP alleles or mutations within a gene can alter the eventual proteins synthesized or yields obtained, which is not investigated when looking at gene presence/absence or gene regulation only. This is a major limitation of the current data, constraining our analysis and outcomes.

Similarly for the structural variation located at the QTL, which could all be relevant and can potentially be responsible for differentiating leaf yields in the parental lines (**Table 3.11**). It is the favourable alleles part of the complete inherited SV causal to obtaining higher leaf yields. This information is hard to investigate when SVs are detected in some of the eight MAGIC founders but not in others or could be missed due to sequencing errors. Generally, the high as well as low performing accessions, i.e. V1044371 and V1062433 (**Figure 3.12**) have similar numbers of SVs with comparable lengths to the other founders, although five and eight unique SVs could be found at the QTLs for these founders respectively (**Table 3.11**). These unique SVs are of interest due to their genomic variants potentially playing a role in the observed varying leaf yields, however, the other SVs at QTLs of which many vary between accessions can all potentially contribute which is hard to derive from our data.

Ideally genetic markers such as SNPs or larger structural variation identification would be performed in the MAGIC offspring genomes directly, but performing higher quality sequencing on many lines is not cost effective. To overcome this problem, we genotyped the offspring with SNPs only using STITCH without a founder reference panel. This enabled us to successfully map QTL for leaf yield, although, identifying candidate genes or specific sequence variation responsible remains challenging. One way to overcome this would be to run STITCH using a reference panel so we know which haplotype block of the founders is inherited in each offspring accession, revealing which putative causal genes are inherited. Additionally, the SVs of the MAGIC founder could be genotyped in the RIL 10x sequences. However, therefore an adequate pipeline should be created which allows imputation as well as QTL mapping with structural variants. This pipeline is under development but was not yet finished by the end of this PhD project. Hence this approach of identifying structural variation in the offspring could yet not be implemented. Investigating the information of annotated

genes and SVs versus phenotypic leaf differences in many offspring accessions could aid positive outcomes for identifying causal genes or genomic variation for leaf yield. However, underlaying differences in sequence within genes or SVs, even when more accurately predicted in the offspring genomes, might constrain results.

The chromosome 9 QTL shows a larger proportion of variance explained (PVE) compared to the QTL at chromosome 10, i.e. 14% and 19% versus 6% and 10% PVE for dry and fresh leaf weight respectively (**Figure 3.10**). This marks the QTL at chromosome 9 as most relevant due to a stronger genetic influence and potentially being more informative for trait selection in breeding programs. Among the founder genomes, notably more genetic variation was observed around the chromosome 9 QTL which had SNP associations with two leafy traits, compared to the chromosome 10 QTL which was associated with dry leaf weight only (**Figure 3.14**). It is expected that genetic variation is present at the QTL since this is needed for identifying QTL (Arrones *et al.*, 2020). In order to map QTL recombination events need to occur in the mapping population, and many loci with small effects are often responsible for most quantitative trait variation (Mackay, Stone and Ayroles, 2009). One way to increase recombination is through additional generations of intercrossing before selfing, which helps break up linkage blocks and improves mapping resolution enabling fine mapping of QTL in smaller genomic regions (Cockram and Mackay, 2018). More crossing generations would therefore be a valuable next step for improving analysis. Additionally, genotyping structural variants in the offspring genomes and mapping QTL with those could also be useful to potentially identify larger genomic variants underlaying enhanced leaf yields. We are fortunate to have mapped QTL for leafy traits and to have identified clear SNP positions with the responsible allele for increased yields. This is valuable information for amaranth breeders who want to speed up the process of selective breeding of vegetable varieties. Due to successful QTL localization we now know that our novel created MAGIC population is of sufficient quality for association mapping. Additional field trials, recording more traits would be a great addition to enable mapping a higher variety of QTL involved in more diverse phenotypes.

# Methods

## Field trial

### Data generation

Of ~200 F4 offspring accessions 1 gram of seed per accession was directly transported from the WVC to the ARC. For another 186 offspring accessions at generation 4 (F4) only ~200 seed per accession were received at the ARC and were therefore bulked there creating F5 used for the field trials. Eventually phenotypic data of 342 accessions could be recorded in the field trials and paired end short-read (~150 bp) sequence data (10x coverage) of 303 MAGIC offspring accessions was generated.

**Table 3.12.** Dates of data collection of MAGIC population field trials conducted at four South African locations.

| **Location:** | Mafikeng | Bloemfontein | Roodeplaat | Nelspruit |
|---|---|---|---|---|
| **Data collection and harvesting** | 13-17 Feb. | 20-24 Feb. | 27 Feb-3 Mar. | 8-10 Feb. |
| **Final data collection** | 15-19 May | 22-26 May | 5-10 June | 29-31 May |

The MAGIC offspring accessions were planted in four climatically diverse environments (Mafikeng, Bloemfontein, Roodeplaat and Nelspruit) during the South African summer of the 2022/2023 growing season. Trials roughly took place from December to March with the exact data collection dates provided in **Table 3.12**. Location of the field trials is provided on a map at the left and with photos at the bottom of **Figure 3.1**. The plants were challenged for drought tolerance by providing irrigation during the first two weeks of growth only and afterwards just being challenged by the hot South Africa summer climate. The set-up of the experiment was a completely randomized design with three repetitions per location and ten plants of the same accession per row per plot. Only for Nelspruit there was less seed available and hence only one replicate could be conducted. The MAGIC parents, Arusha (*A. cruentus* reference) and Anna (another leafy *A. cruentus* strain commonly grown by the ARC) accessions were also planted in Mafikeng and Roodeplaat with three repetitions for each location. More information regarding the four trial locations and their climatic conditions are provided in **Table 3.13**. These values are the long-term weather conditions in the provinces the field trials are located in, which all four belong to an arid (steppe) agro-ecological zone (Benhin, 2006).

**Table 3.13.** Coordinates and long-term climatic conditions in the provinces of the four field trial locations the MAGIC RILs were planted. For each parameters the overall annual (A) as well as summer (S) means are provided since our field trial was during the summer months. Values are obtained from (Benhin, 2006).

| **Location** | **Coordinates** | **Province** | **Min. (°C)** | **Mean (°C)** | **Max. (°C)** | **Rainfall (mm/month)** |
|---|---|---|---|---|---|---|
| Mafikeng | -25.819678805272243, 25.613790395926507 | North West | A: 18.35<br>S: 22.25 | A: 18.35<br>S: 22.25 | A: 18.35<br>S: 22.25 | A: 44.31<br>S: 76.67 |
| Bloemfontein | -28.929177185056666, 26.325469442522294 | Free State | A: 13.81<br>S: 17.56 | A: 14.59<br>S: 18.42 | A: 15.42<br>S: 19.33 | A: 49.68<br>S: 79.85 |
| Roodeplaat | -25.59835695987395, 28.359295475594585 | Gauteng | A: 16.22<br>S: 19.65 | A: 16.22<br>S: 19.65 | A: 16.22<br>S: 19.65 | A: 59.02<br>S: 101.76 |
| Nelspruit | -25.45358302918083, 30.993149508638304 | Mpumalanga | A: 17.65<br>S: 20.93 | A: 17.75<br>S: 21.02 | A: 17.87<br>S: 21.10 | A: 61.13<br>S: 103.47 |

### Data availability and analysis

Categorical nominal data for descriptor traits like growth habit, stem pigmentation, leaf pigmentation, leaf shape, petiole pigmentation, inflorescence attitude, inflorescence shape, inflorescence density and inflorescence colour were collected and presented in a proportional bar plot R v.4.2.3 (R Core Team, 2021). The number of plants per row and those of which flowering was recorded. Numerical data was collected for three highly relevant agricultural traits; fresh as well as dry leaf weight and chlorophyll (average of three measurements) of which the available data points and numbers of MAGIC offspring accessions is provided in **Table 3.14**.

The numerical data of the three relevant agronomic traits was checked for quality by making a histogram for a normal distribution as well as Q-Q plots (quantile-quantile plots) for comparing the distribution of data to the standard normal distribution. MAGIC offspring data for the three traits was plotted as boxplots and violin plots per location with MAGIC parent, Arusha and Anna accession repeats plotted as individual data points. If data was normally distributed Tukey's HSD test was used for post-hoc comparisons with letter codes for pairwise significance between locations were calculated and if not normally distributed Wilcoxon signed-rank test comparison bars were plotted above the boxplots. In case of not normally distributed data the trait data was log-transformed to create normally distributed data. Variables of this normally distributed trait data were fitted into linear mixed-effects models producing restricted maximum likelihood (REML) estimates determined by the lmer function in the lme4 package (Bates *et al.*, 2015). The most complete version of the model used to test significance of various variables on the different agronomic traits recorded is;
$y_{ijk} = \mu + Accession_i + Location_j + Acc.* Location_{ij} + Rep_k + e_{ijk}$

**Table 3.14.** Number of MAGIC offspring accessions for which data is available; the total of data points (total), which of those are unique to a single accession ID (unique accession) and therefore a rough estimation of repeats the data is based on (~# reps) per three types of phenotypic traits (dry or fresh leaf weight with the same amount of data available, or chlorophyll) and per four field trial locations (Roodeplaat, Mafikeng, Bloemfontein, and Nelspruit). The total of all data points of all four locations per trait, those unique to one accession ID, and those unique accessions overlapping between all four locations or only three locations excluding single repeat Nelspruit data is provided.

| **Traits:** | **Location** | **Total** | **Unique accessions** | **~# reps** |
|---|---|---|---|---|
| Dry / Fresh weight | Roodeplaat | 780 | 305 | 2.6 |
| | Mafikeng | 770 | 312 | 2.5 |
| | Bloemfontein | 593 | 296 | 2.0 |
| | Nelspruit | 192 | 192 | 1.0 |
| | Total | 2335 | 322 (overlap: 161 in 4 and 273 in 3 locations) | 1.8 |
| Chlorophyll | Roodeplaat | 745 | 310 | 2.4 |
| | Mafikeng | 575 | 316 | 1.8 |
| | Bloemfontein | 604 | 297 | 2.0 |
| | Nelspruit | 202 | 202 | 1.0 |
| | Total | 2126 | 339 (overlap: 168 in 4 locations) | 1.6 |
| Total | | 3113 | 342 | 2.3 |

$y_{ijk}$ is the trait (dry or fresh leaf weight or chlorophyll) prediction, $\mu$ is the mean, $Accession_i$ is the fixed effect of the MAGIC offspring (n=322 or n=339), $Location_j$ is the fixed effect of location (n=4), $Acc.*Locatoin_{ij}$ is the fixed effect of accession by location pairwise interaction, also

considered Genotype by Environment, $Rep_k$ is the random effect of repeat (n= ~3) and $e_{ijk}$ is the random residual assumed to be $\sim N(0, I\sigma^2{}_e)$, with I being an identity matrix and $\sigma^2{}_e$ the residual variance. Significance of the fixed main as well as interaction variables was tested using Analysis of Variance (ANOVA). If not significant, the model was adjusted in a backward stepwise manner and the model with the lowest akaike information criterion was considered best fitting. $Rep_k$ was considered a random model effect since potential differences between replicates can still relate to a hierarchy that the analysed data is drawn from in which repeats can be considered as different populations. To test whether MAGIC offspring accessions differ within different testing environments the datasets were split up per four locations and the following linear mixed-effects model was run for the three locations for which multiple repeat data was available:

$$y_{ij} = \mu + Accession_i + Rep_j + e_{ij}$$

Here $y_{ij}$ is trait (leaf weight or chlorophyll) prediction, $\mu$ is mean, $Accession_i$ fixed effect of the MAGIC offspring accessions, $Rep_j$ is the random effect of repeat (n= ~3) and $e_{ij}$ is the random residual assumed to be $\sim N(0, I\sigma^2{}_e)$ with significance of the fixed accession effect again tested with ANOVA. All data transforming and statistics was performed with R v.4.2.3 (R Core Team, 2023) and plotting with package ggplot2 (Wickham, 2016). For the most relevant agronomic traits (fresh and dry leaf weight) parallel coordinates plots were made using the GGally (Schloerke *et al.*, 2021) R package, considering either the four or three field trial locations (single replicate Nelspruit excluded) as parallel lines with the leaf yield data as coordinates. These parallel coordinates plots were used to identify the best performing accessions despite climatic conditions. To further quantify these results, estimated breeding values (EBV) have been calculated for each accession regarding dry or fresh leaf yield traits using R package sommer (Covarrubias-Pazaran, 2016). Correlation of phenotypes was investigated by plotting raw agronomic trait values as variables in a correlation plot with Pearson correlation coefficients using R v.4.2.3 (R Core Team, 2023). This was done for only those datapoints which had dry as well as fresh leaf weight as well as chlorophyll values available for an individual MAGIC offspring accession in a certain trial location or replicate with hence the number of datapoints lower than the total N=2335 and N=2126 in **Table 3.14**.

## Genomics

### MAGIC offspring reads and quality.

Low coverage (10x) paired end MAGIC offspring short-reads (~150 bp) were simultaneously trimmed and checked for quality with Fastp v.0.23.2-GCC-10.3.0 (Chen, 2023). The combining of the individual FastQC results of the trimmed sequences was performed with MultiQC v.1.13-intel-2021b (Ewels *et al.*, 2016). The sequence quality was slightly improved after trimming although there were no major differences (**Supp table sequencing stats RILs**). For security of using slightly better data, we decided to continue analysis with the trimmed reads.

### MAGIC offspring mapping to *A. cruentus* reference genome and coverage calculations

The trimmed fastq sequence reads were used to create Sequence Alignment Map (SAM) tab-delimited text files, followed by their binary version; Binary Alignment Map (BAM) files,

per offspring accession mapped to the *A. cruentus* reference genome (Ma *et al.*, 2021). This was done using Burrows-Wheeler Aligner (BWA) BWA v.0.7.17-GCCcore-11.3.0 (Li and Durbin, 2010) for genome aligning, followed by SAMtools v.1.17-GCC-12.2.0 (Danecek *et al.*, 2021) commands. After mapping to the reference genome, sequence coverage on exonic and non-exonic regions was calculated using BAM files of each accession compared to the complete reference genome as well as the genome excluding scaffolds (pseudo-chromosomes only). For the *A. cruentus* reference genome a GFF file was published (Ma *et al.*, 2021) (https://bioinformatics.psb.ugent.be/gdb/amaranthus/). Browser Extensible Data (BED) files containing exonic regions were created using the Linux awk command searching for ‘exon’ in the GFF file. Non-exonic BED files were created using the GFF file, a file containing chromosome sizes of the reference genome, and the tool BEDTools v.2.27.1-foss-2018b (Quinlan and Hall, 2010), command ‘bedtools complement’. SAMtools v.1.15-GCC-11.2.0 (Danecek *et al.*, 2021) command ‘samtools bedcov’, was used to calculate coverage for each accession sequence. Mean coverage of all accessions was calculated per exon or non-exon length, of which the median is calculated of the entire MAGIC offspring set. Coverage per exon or non-exon length were also calculated per accession individually. Boxplots including median coverage per length per accession were made using R v.4.2.3 (R Core Team, 2021) package ggplot2 (Wickham, 2016).

### MAGIC offspring longshot SNP imputation with GATK haplotype caller

Picard v.2.25.5-Java-13 (‘Picard Toolkit’, 2019) with tools in JavaScript was used to add read groups to the BAM files and outputs were again indexed using SAMtools v.1.17-GCC-12.2.0 (Danecek *et al.*, 2021). In this way BAM files were prepared for reading with Genome Analysis Toolkit (GATK). GATK Haplotype Caller v.4.3.0.0-GCCcore-11.3.0-Java-11 (Auwera and O’Connor, 2020) was used to impute the SNPs and indels previously found in the eight MAGIC parents by Longshot (Edge and Bansal, 2019) filtered for "PASS", QUAL > 50 and DP > 10 (manuscript in appendix), in the MAGIC offspring read alignments by using the following command: ‘gatk --java-options "-Xmx4g" Haplotype Caller -R Amaranthus_cruentus_genome.fa -I $1.bam -O $1.output.g.vcf.gz --ERC GVCF --alleles concat.longshot.vcf.gz’. All VCF files per offspring accession were then merged into one VCF file for which regenotyping was performed. This means identified but empty SNP or indel locations were filled with allele information using the originally created accession BAM files and the reference genome. The output is VCF files per reference chromosomes and scaffolds, created with freebayes-parallel v.1.3.2-foss-2018b (Garrison and Marth, 2012). Concatenating was performed with BCFtools v.1.9-foss-2018b (Danecek *et al.*, 2021) to producing one final output VCF file.

### MAGIC offspring longshot SNP imputation with STITCH

STITCH (Davies *et al.*, 2016) was run without a founder reference panel with parameters K=8, nCores = 40, and first testing nGen to be either 4 or 7 on chromosome 7 for comparison. Position text files, used as an input for STITCH, were converted from the SNPs called in the MAGIC founders with Longshot (Edge and Bansal, 2019) towards the *A. cruentus* reference genome (Ma *et al.*, 2021) filtered for "PASS", QUAL > 50 and DP > 10 (manuscript in appendix). This was done by first making the VCF file biallelic with BCFtools v.1.19-GCC-13.2.0 (Danecek *et al.*, 2021); ‘bcftools view -m2 -M2 -v snps concat.longshot.vcf.gz -O z -o biallelic.concat.longshot.vcf.gz’. Then isolating the positions with; ‘bcftools query -f '%CHROM\t%POS\t%REF\t%ALT\n' biallelic.concat.longshot.vcf.gz >

pos.txt' followed by 'awk '$4 ~ /^[ACGT]$/ {print}' pos.txt > single.pos.txt'. MAGIC offspring bam files were prepared for STITCH by first reheadering them with SAMtools v.1.17-GCC-12.2.0 (Danecek *et al.*, 2021); 'samtools reheader ../AVAM22_190.header.sam ../$1\.sort.bam > $1\.sort.reheadered.bam' followed by adding read groups with picard v.2.25.5-Java-13 ('Picard Toolkit', 2019); 'java -jar $EBROOTPICARD/picard.jar AddOrReplaceReadGroups I=$1\.sort.reheadered.bam O=$1\.reheadered_read_group.bam RGID=$1 RGLB=lib_$1 RGPL=illumina RGPU=unit1 RGSM=$1'. Intermediate bam files were indexed with SAMtools v.1.17-GCC-12.2.0 (Danecek *et al.*, 2021). The output VCF file from STITCH with the correct nGen value run on all chromosomes was filtered for all sites with an info score below 0.4, minor allele frequency below 5%, or missingness above 10%, inspired by analysis of the wheat MAGIC population in Scott et al., (2021). This was done with; 'bcftools filter -i 'INFO_SCORE > 0.4' input.vcf.gz -Oz -o concat_correct_filtered.vcf.gz' using BCFtools v.1.15.1-GCC-11.3.0 (Danecek *et al.*, 2021) followed by 'vcftools --gzvcf concat_correct_filtered.vcf.gz --maf 0.05 --max-missing 0.9 --recode --out filtered_diffMAF_miss' using VCFtools v.0.1.16-GCC-11.2.0 (Danecek *et al.*, 2011).

### Genotyping method comparisons

The number of SNPs were counted in all final output VCF's, prior and after filtering, with the following command: 'zcat file.vcf.gz | grep -v "^#" | wc -l > count_snps'. The call rate per individual accession in a VCF file was detected with VCFtools v.0.1.16-GCC-11.2.0 (Danecek *et al.*, 2011) 'vcftools --vcf input.vcf --missing-indv', in which the call rate was 1 minus the fraction of missingness. Observed homozygosity and the inbreeding coefficient (F) per individual was found with 'vcftools --vcf your_data.vcf --het --out het_results'. This was done for both genotyping methods, i.e., GATK Haplotype Caller (Auwera and O'Connor, 2020) and STITCH (Davies *et al.*, 2016), prior as well as after SNP filtering ("PASS", QUAL > 50 and DP > 10 (manuscript in appendix). STITCH after SNP filtering was chosen as the preferred SNP imputation method. Principal component analysis (PCA) was conducted after genotyping the offspring with STITCH while including the parental lines as samples. The individual MAGIC founder VCF files were merged with the offspring genotyping VCF's using 'bcftools merge' of BCFtools v.1.19-GCC-13.2.0 (Danecek *et al.*, 2021). The VCF files was made biallelic using 'bcftools view' followed by annotation using 'bcftools annotate' with BCFtools v.1.19-GCC-13.2.0 (Danecek *et al.*, 2021). The VCF was further filtered for LD with 'plink --indep-pairwise 10 1 0.5' of PLINK v.2.0.0-a.6.5-gfbf-2024a (Chang *et al.*, 2015), so with pairwise comparisons calculated in windows of 10kb, a step size of 1, and an R Squared ($R^2$) threshold of 0.5 like for PCA in the manuscript in the appendix. A binary file set was created using 'plink --make-bed', followed by producing eigenvec and eigenval files for PCA with '–pca' of PLINK v.2.0.0-a.6.5-gfbf-2024a (Chang *et al.*, 2015). Plots for method comparisons as well as the PCA were made with ggplot2 (Wickham, 2016) in R v.4.2.3 (R Core Team, 2023).

### QTL mapping using STITCH SNPs

An attempt to map QTLs has been done using the numerical data which was available for three traits: chlorophyll, fresh and dry leaf weight. Data of the latter two was log-transformed due to non-normal distribution which was discovered with making histograms and Q-Q plots (M&M: Data availability and analysis). A few datapoints (n=29) which were zero after log-transforming dry weight were excluded. To correct for the effects of data collected in different agro-ecologies, predicted values of a linear mixed-effects model were used instead of raw

datapoints. This was done by producing restricted maximum likelihood (REML) estimates with the lmer function in the lme4 R package (Bates *et al.*, 2015). After this correction, average phenotypes per accession were calculated which were used for QTL mapping. SNPs imputed with STITCH (Davies *et al.*, 2016) which were filtered (M&M: MAGIC offspring longshot SNP imputation with STITCH) were used as genotypes. QTL mapping was performed using custom R scripts derived from those described in Scott et al., (2021). This statistical method for QTL mapping is designed for MAGIC populations with models accounting for multiple founders and defining QTL mapping intervals by SNPs flanking the peak SNP signal. Here, QTL mapping intervals including SNPs surrounding the peak SNP are defined with $\log_{10}(p)$ values within $d$ units ($d = \max\{2, 0.1x\}$) of $x$ ($x$ = peak $\log_{10}(p)$ value), which can include a range of sites that have $\log_{10}(p)$ values within $d$ units of $x$. Associations were tested using a linear mixed model incorporating a genomic relationship matrix (GRM) to control for relatedness. This GRM was generated similarly like in Scott et al., (2021) as (GRM) $K = GG'/p$, with $p$ being a subset of SNPs after filtering and $G$ a genotype dosage matrix based on standardized genotype dosages over all RILs. Significance thresholds were determined in two ways: a nominal per-marker threshold, based on raw p-values from the association test, and a more stringent genome-wide threshold derived from permutation testing, in which phenotypes were randomly shuffled. In the latter, the maximum -logP observed in each permutation was recorded to estimate the 95th percentile threshold, thereby correcting for multiple testing across the genome.

Once QTLs were identified, phenotypes were plotted with boxplots per genotype (reference (Ma *et al.*, 2021) and alternative) with ggplot2 (Wickham, 2016) using R v.4.2.3 (R Core Team, 2021). Those genotypes were derived from dosages output traw files from STITCH (Davies *et al.*, 2016), in which dosages <0.5 were considered homozygous reference alleles, >1.5 is homozygous alternative alleles and everything between 0.5 to 1.5 a hypothetical heterozygous of those two allele possibilities. Significance ($p < 0.05$) between these genotypes was found with a Wilcoxon signed-rank test for which comparison bars were plotted above the boxplots. Exact base pair locations for QTL of pleiotropic traits were identified by matching the 100 most significant SNP locations of each trait with each other, and of those overlapping the most distant SNP locations are set as QTL boundaries.

### Genes and structural variants (SVs) at QTL

To further investigate the identified QTLs and link it to performance, the numerical trait data with found QTLs was plotted as single datapoints as well as boxplots per MAGIC founder, combining data from two field trial locations and three replicates with R package ggplot2 (Wickham, 2016) using R v.4.2.3 (R Core Team, 2021). Genes of the *A. cruentus* reference (Ma *et al.*, 2021) located within the selected QTL boundaries and MAGIC founder genes which are their orthologous (found with OrthoFinder v.2.3.3) (Emms and Kelly, 2015)) were identified to visualize the available gene expression data (manuscript in appendix). First this RNAseq data was checked for quality with FastQC v.0.11.7-Java-1.8.0_181 (Andrews, 2010) and then all quality data was combined with MultiQC v.1.13-intel-2021b (Ewels *et al.*, 2016). Individual transcriptomes were created with the available genome fasta and GFF file using gffread v.0.12.7-GCCcore-11.2.0 (Pertea and Pertea, 2020). From this transcriptome an index has been created using Salmon v.1.10.0-GCC-11.3.0 (Patro *et al.*, 2017) which was also used for the quantification of the paired end reads with the following command: ‘salmon quant --index “index folder” -p 12 -l A -1 “*_1.fq.gz” -2 “*_2.fq.gz” -o "output file"/ --

seqBias --posBias --gcBias --numBootstraps 100 –validateMappings'. Output of Salmon provided transcripts per kilobase million (TPM) values which were further processed with the online webtool 3D RNA-seq (Guo *et al.*, 2021). Here parameters set were 'lengthScaledTPM' as a tximport method and 'Filter & Mean-variance trend plot' was pressed prior downloading results. The 'counts' matrix of the txi_trans results file were imported and all values below 1 were put to 0. All data was then log 2 transformed with 'log2()' in R v.4.2.3 (R Core Team, 2023) in order to normalize it to making it suitable for visualisation. Using the pheatmap (Kolde, 2019) R package, gene expression heatmaps were made with Arusha reference genes and founder orthologs located at QTL.

Of the SVs <100kb (found with PBSV ('PacificBiosciences/pbsv', 2024) and cuteSV (Jiang *et al.*, 2020): manuscript in appendix) which were identified of the MAGIC founder genomes against *A. cruentus*, those located within the QTL are highlighted. Genomic variation captured in the *Amaranthus* pan-genome (manuscript in appendix) at the QTL was visualized with Bandage (Wick *et al.*, 2015). The most significantly associated SNP at each leafy trait QTL was identified and a 1000 base pair window was taken around these SNP locations at the *A. cruentus* reference (Ma *et al.*, 2021) that the MAGIC founder sequences were aligned to while constructing the pan-genome. The genomic variation of the MAGIC founders towards this 1000 base pair reference genome window captured in the pan-genome is visualized for each QTL.

# Chapter 4: Domestication confusion in 'hybridus' species complex: Questioning the existence of *A. hybridus* based on exploiting a 90 accession diversity set.


## Abstract

Amaranth species have a large potential to contribute to sustainable, healthy food systems, although unfortunately this underutilised crop is often neglected in research and breeding. *Amaranthus* is native to the America's, and attempts for species demarcation have been performed in the genus by exploiting diversity sets. However, this remains difficult due to frequent hybridization amongst accessions, complicating taxonomy. To explore the metabolome and its nutritional relevance for humans, many studies have conducted metabolite profiling in Amaranth diversity sets. Although, investigating associations of metabolites with amaranth genomics is less common, with the first metabolite-based genome-wide association study (mGWAS) yet to be performed.

Here we present investigation of two diversity sets for species demarcation and the first ever conducted mGWAS in amaranths. Explored were a 90 accession set, newly generated and sequenced by the Agricultural Research Council (ARC), and a set previously analysed by Stetter et al. (2020) containing 121 accessions of which most presumably belong to the 'hybridus' species complex. A species complex contains closely related subpopulations, lacking clear boundaries for species distinction. The 'hybridus' complex is believed to consist of three grain amaranths (*A. caudatus*, *A. cruentus*, *A. hypochondriacus*) and two putative wild ancestors (*A. hybridus* and *A. quitensis*). Sequences of both sets were aligned to two references, i.e., *A. cruentus* (Ma *et al.*, 2021) and *A. hypochondriacus* (Lightfoot *et al.*, 2017), using SNP markers for phylogenetic analysis as well as conducting mGWAS.

Current phylogenetic analysis successfully contributed to species demarcation, classifying 38 of 90 ARC accessions, notably including 17 of 22 accessions with previously unknown species. Only three of the 121 external accessions were re-assigned, indicating that inclusion of this set was a useful guideline and that our results comply with previous phylogenetic analysis. No clear separation of *A. hybridus* as a group in our as well as previous research did occur, this raises a strong suspicion of its existence as a species. Species assignments of ARC diversity set were further confirmed by clustering of the 90 accessions based on leaf metabolomics. 486 of the 1605 detected metabolite features could be annotated, among which many compounds relevant to human nutrition. For 81 metabolite traits, SNP associations could be found, of which 22 could be annotated and with eight being selected as of interest due to high significance, heritabilities and preferable data quality. This shows performing mGWAS in leafy amaranth is useful, revealing many hints towards genomic regions responsible for nutritionally relevant metabolites despite a low number of accessions used. For further investigation, regarding *Amaranthus* species demarcation, the existence of *A. hybridus* as a species, as well as optimizing mGWAS results, we recommend the exploration of more and novel sequenced amaranth accessions as part of diversity sets.

# Introduction

*Amaranth*, a widely consumed leafy vegetable in sub-Saharan Africa, is underutilized in research, breeding and food systems (Mabhaudhi *et al.*, 2016, 2019; Emmanuel and Babalola, 2021). *Amaranth* originates from the America's where it was first cultivated as a grain crop in South, and later in North America, with Spanish conquerors responsible for distribution to Europe followed by eventual spread to the rest of the world (MacNeish, 1965; Schultes, 1984; Arreguez, Martínez and Ponessa, 2013).

## Amaranth taxonomy

As amaranth appears to have high levels of phenotypic variation, with hybridization possible and difficult taxonomy in the past, clear knowledge regarding genetic diversity and species identification might be lacking (Sauer, 1957; Stetter et al., 2020). Various studies investigating diversity sets of amaranths, consisting of sequenced accessions of different species have been performed to further explore the relationships within the genus (Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Jamalluddin *et al.*, 2022; Lin *et al.*, 2022; Blair *et al.*, 2023; Graf *et al.*, 2025). These studies revealed a variety of findings. They aided species demarcation within the taxonomically challenging genus *Amaranthus* (Stetter and Schmid, 2017; Jamalluddin *et al.*, 2022; Lin *et al.*, 2022), with some studies specifically focussing on highly related grain amaranths and their suggested domestication from wild species *A. hybridus* (Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023; Graf *et al.*, 2025). Other diversity sets were used to map trait associations in genomes for relevant agronomic as well as domestication characteristics (Stetter, Vidal-Villarejo and Schmid, 2020; Jamalluddin *et al.*, 2022; Lin *et al.*, 2022). Assessing diversity sets is therefore commonly practiced and useful for studying phylogenetic relationships among accessions, their historic domestication and potential genome trait associations. However, within the genus *Amaranthus*, there is still uncertainty among some phylogenetic clades and domestication, hence further examination of the evolutionary history and relations between amaranths is often requested (Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Waselkov, Boleda and Olsen, 2018; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Lin *et al.*, 2022; Blair *et al.*, 2023).

Generally, three subgenera are recognised within the genus *Amaranthus* based on phenotypic, morphological and genomic data; *Amaranthus Albersia, Amaranthus Amaranthus* and *Amaranthus Acnida* (Mosyakin and Robertson, 1996; Costea and DeMason, 2001; Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). These subgenera represent taxonomic groupings based on species clustering into so called clades. Nonetheless, in recent phylogenetic studies, inconsistency can be observed regarding the positioning of clades, with still much uncertainty of affinity among highly related species within clades (Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). A typical example of *Amaranthus* species among which relationships are known to be uncertain are those assigned to the 'hybridus' species complex, currently defined as part of the *Amaranthus Amaranthus* subgenus (Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). A species complex is a group of closely related organisms which are morphologically similar and genetically somewhat distinct but tend to hybridise with each

other blurring the boundaries for clearly separated species. There is considerable confusion about this complex. Some analyses suggest that five species are contained within this 'hybridus' species complex; three grain amaranths (*A. caudatus*, *A. cruentus*, *A. hypochondriacus*) and two wild relatives or putative ancestors (*A. hybridus* and *A. quitensis*) (Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017). Previously *A. powellii* was also proposed as a putative wild progenitor belonging to the 'hybridus' species complex (Sauer, 1967; Stetter, Vidal-Villarejo and Schmid, 2020), however, later it was discovered that *A. powellii* was not closely related to the three grain amaranths (Xu and Sun, 2001; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017). These findings that *A. powellii* is unrelated to the 'hybridus' complex species was mostly based on genomic data with DNA markers (Xu and Sun, 2001; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014) like SNPs (Stetter and Schmid, 2017), but also full length sequences of starch branching enzyme genes (Park *et al.*, 2014) and additional morphological phenotypic data (Adhikary and Pratt, 2015). The three grain amaranths of the 'hybridus' species complex are still considered incompletely domesticated crops. For example, they retain the brown grain colour rather than white grains which is a known key domestication trait (Stetter, Müller and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020). The lack of fixation of seed colour could be explained by ongoing gene flow from wild into the domesticated amaranths, while gene flow is also known to occur among those three grain amaranth species as they can hybridize within the species complex (Stetter, Vidal-Villarejo and Schmid, 2020). This makes clear identification and separation of species within the complex challenging without analysis of genome-scale data.

Generally, it is believed that *A. caudatus* and *A. quitensis* are Andean species originating from South America, *A. cruentus* and *A. hypochondriacus* are Mesoamerican species from North and Central America, while accessions thought to be *A. hybridus* appear to be found in both regions (Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). However, the taxonomic classification of *A. hybridus* has a long history marking uncertainty, largely due to its apparent capacity to hybridize with other species, making clear identification challenging for experts (Sauer, 1950, 1967; Brenner *et al.*, 2010). This is often observed in research investigating the relations among the 'hybridus' complex species, with various *A. hybridus* accessions grouping among clearer clustering of accessions assigned to other species (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023). These analyses concluded that *A. hybridus* was the wild ancestor of the three closely related grain amaranth species in the 'hybridus' species complex (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Clouse *et al.*, 2016; Stetter and Schmid, 2017). At first it was assumed that at least two separate domestication events occurred from *A. hybridus* to the three grain amaranth species based on inferred microsatellite simple sequence repeat markers (Kietlinski *et al.*, 2014). Later it was suggested that three separate domestication events for each individual grain species may have taken place, since the fixation index ($F_{ST}$, a measure of genetic distance between populations) between species showed the three domesticated amaranths more strongly differentiated from each other than each grain amaranth was from their assumed wild ancestor *A. hybridus* (Stetter et al., 2020). Due to the uncertainty of taxonomy of *A. hybridus* and how this wild species led to domestication of cultivated grain amaranths, as well as the general ambiguity of clades within the genus *Amaranthus*, further investigation of the genus from whole genome data

appears to be valuable. This need for more sequenced amaranth individuals is commonly requested in recent studies, since by capturing more genetic diversity, phylogenetic resolution can be enhanced, providing stronger evidence for species relationships and domestication history (Stetter and Schmid, 2017; Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Waselkov, Boleda and Olsen, 2018; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Lin *et al.*, 2022; Blair *et al.*, 2023).

## Genome-wide association studies (GWAS) and metabolite-based GWAS (mGWAS)

Many *Amaranthus* diversity set studies conduct (Stetter, Vidal-Villarejo and Schmid, 2020; Jamalluddin *et al.*, 2022; Lin *et al.*, 2022) or suggest conducting (Wu and Blair, 2017; Blair *et al.*, 2023), trait association studies, as diversity sets may be a valuable tool to identify genomic regions responsible for phenotypic traits. To examine this possibility, we performed genome-wide association studies (GWAS) with amaranth diversity sets to associate genetic variants with agronomic and domestication related traits (Stetter, Vidal-Villarejo and Schmid, 2020; Jamalluddin *et al.*, 2022; Lin *et al.*, 2022). GWAS commonly uses SNPs as genetic variants to be linked with diverse phenotypic traits (Korte and Farlow, 2013), helping identify genes for crop improvement and understanding the biology of agronomic traits. Detecting metabolites is an important way of collecting phenotypes for GWAS. Such metabolite-based genome-wide association study (mGWAS) are commonly used as a technique to explore associations of plant metabolism with crop genomes (Luo, 2015; Alseekh and Fernie, 2018).

Metabolites are molecular compounds present in tissue at a certain point in time, which can be involved in cellular reactions or metabolic pathways with a variety of compounds being highly relevant for human nutrition (Dührkop *et al.*, 2015; Perez de Souza *et al.*, 2019; Zhu *et al.*, 2020). There has been an emergence of untargeted metabolomics studies, which aims to detect unknown compounds by focussing on a broad range of metabolites without prior selection (Perez de Souza *et al.*, 2019; Aharoni, Goodacre and Fernie, 2023). Many recent studies have successfully identified nutritionally and pharmacologically beneficial metabolites in *Amaranthus* leaves by untargeted detection (Rocchetti *et al.*, 2020; Nkobole and Prinsloo, 2021; Abdel-Moez *et al.*, 2023; Nemadodzi and Managa, 2024). Some examples of nutritionally relevant metabolites detected in vegetable amaranths are amino acids and flavonoid glycosides in *A. hybridus*, *A. blitum*, and *A. caudatus* varieties (Abdel-Moez *et al.*, 2023), or flavonoids, cinnamic acid derivatives and benzoic acids in *A. cruentus* leaves suggesting antidiabetic features (Araujo-León *et al.*, 2024). Despite the success of untargeted metabolite detection in leafy amaranth, and acknowledgement of mGWAS for discovering metabolome-genome associations, no mGWAS has yet been conducted in leafy amaranth. We expect that with the inclusion of more accessions and the application of mGWAS, nutritionally relevant metabolites and amaranth lines would be identified, and genes responsible for these nutritional compounds might be detectable.

Most genes responsible for plant metabolite synthesis are located independently in the genome. However, recent science shows that genes responsible for metabolite synthesis occasionally can be located (clustered) together for coordinated expression and involvement in biosynthetic pathways (Boycheva *et al.*, 2014; Bharadwaj *et al.*, 2021). mGWAS can also be used as technique to explore such plant metabolic gene clusters, by identifying closely located genomic regions associated similar metabolites potentially involved in the same

biosynthetic pathway (Zhan *et al.*, 2022). Here mGWAS is a forward genetic approach to detect putative biosynthetic gene clusters (BGC) from phenotype (metabolite) to gene (Zhan *et al.*, 2022). This is different from genome mining to directly investigate genome sequences for potential BGC, as performed in the *A. cruentus* reference genome (Ma *et al.*, 2021). For genome mining, a high quality sequence of a single individual is required, while for (m)GWAS many individuals from a natural population are required (Korte and Farlow, 2013; Pascual *et al.*, 2016; Visscher *et al.*, 2017; Arrones *et al.*, 2020; Ponce *et al.*, 2020; Abdellaoui *et al.*, 2023). The power of GWAS studies is difficult to assess precisely, but depends upon population structure, the complexity of traits (how many genes cause the traits), and the penetrance of the phenotypes examined.

## Analysis of the *Amaranthus* diversity set

*A. cruentus* (Ma *et al.*, 2021) is one of the three reference genomes which were available at the start of this thesis, the other two being *A. hypochondriacus* (Lightfoot *et al.*, 2017) and *A. tricolor* (Wang *et al.*, 2022). All three are used in this chapter for genome alignment and phylogenetic analysis of a novel *Amaranthus* diversity set containing unanalysed sequences.

This newly generated diversity set consists of 90 different amaranth accessions which have been obtained and grown by the Agricultural Research Council (ARC) science institute in South Africa (https://arc.agric.za/Pages/Home.aspx). Genome sequences and leaf samples (analysed for metabolite content) were collected for these 90 accessions. Additional genome sequences of a 121 amaranth diversity set are available from a previously published article (Stetter et al., 2020). The genetic diversity captured in both diversity sets with sequenced genomes (**Table 4.1**) can largely aid species demarcation within the taxonomically challenging genus *Amaranthus*. Hence, a first main research aim of this chapter is to exploit the genetic diversity present in these sets to investigate if species can be identified and if phylogenetic relationships between species can be revealed. Especially regarding accessions with for now unknown species assignment in a not previously analysed diversity set, or for the challenging to assess ‘hybridus’ complex species with various domestication and evolutionary theories in literature.

The metabolomic data available in the 90 accession ARC diversity set could reveal nutritionally relevant metabolite profiles in novel assessed amaranth lines. Hence a second aim of this chapter is to attempt to detect and annotate metabolites for their dietary properties and investigate which nutritional compounds are present in the diversity set. Additionally, can be investigated if these detected metabolite compounds or specific classes differ or cluster per amaranth species or across accessions.

By combing the genomic as well as metabolomic data of the 90 accessions, a final research aim is to potentially link genetic variants to metabolite profiles. Conducting mGWAS with the presence or absence of metabolites could alleviate mapping associations using SNPs as genomic variants in the diversity set genomes. Furthermore, an attempt to predict heritabilities for identified genome-metabolite associations can reflect the genetic architecture of the metabolite trait, which are the joint effect of SNPs and other variants responsible for the trait. Finally, looking into closely located SNPs or genes associated with similar metabolites can reveal putative biosynthetic gene clusters (BGC) with metabolites which are hypothetically involved in the same biosynthetic pathway.

# Results

## Obtaining highly variable *Amaranthus* diversity sets from various sources

The 90 accession diversity set was grown in an irrigated field trial by the ARC in South Africa during the 2018/2019 growing season. What were expected to be the species of the ARC diversity set, as well as of the 121 externally sequenced accessions (Stetter et al., 2020) prior performing novel phylogenetic analysis is provided in **Table 4.1**. For the ARC set, there were many accessions whose species were not defined (*A. sp.*). The additional genome sequences of the external publication were mostly obtained from 'hybridus' complex species, containing the three grain crop amaranths (*A. caudatus*, *A. cruentus*, and *A. hypochondriacus*), the two believed to be wild ancestors (*A. hybridus* and *A. quitensis*), and lastly *A. powellii* and a hybrid species (*A. hybr.*) (Stetter *et al.*, 2020) (**Table 4.1**).

**Table 4.1.** Diversity set information regarding the 90 accessions grown by the ARC and 121 imported sequences of (Stetter *et al.* 2020). The assumed species in the diversity sets at the start of the experiment prior performing any new phylogenetic analysis, and hence based on phenotypes or gene bank information only for the ARC set and the published phylogenetic analysis for the 121 set (Stetter *et al.* 2020). Accessions of whose species were not defined are *A. sp.* and hybrids are *A. hybr.*

| **Species** | **ARC set** | **External sequences** (Stetter *et al.*, 2020) |
|---|---|---|
| *A. hypochondriacus* | 15 | 24 |
| *A. tricolor* | 16 | |
| *A. cruentus* | 11 | 25 |
| *A. caudatus* | 12 | 32 |
| *A. hybridus* | 7 | 19 |
| *A. dubius* | 3 | |
| *A. graecizans* | 3 | |
| *A. quitensis* | 1 | 19 |
| *A. sp.* | 22 | |
| *A. powellii* | | 1 |
| *A. hybr.* | | 1 |
| **Total** | 90 | 121 |

## Correcting errors in species definitions

Of the 90 accession diversity set plants grown by the ARC, the variety of gene bank sources these are obtained from (with the number of accessions) are as follows; ACAT seedfair (3), ARC (7), the world vegetable center (AVRDC) (12), Mayford (1), North Central Plant introduction station (2), NPGRC (Tanzania National Plant Genetic Resources Centre) (7), USDA (30), mutation breeding (10) or unknown (18). All 121 accessions in Stetter *et al.* (2020), were obtained from the USDA ARS gene bank. There appeared to be some errors in the species records, since for six accessions phylogenetic analysis in the publication did not match species classification of their USDA gene bank passport (three *A. hypochondriacus* classified as *A. cruentus*, and three vice versa) (Stetter *et al.*, 2020; Table S4). Addressing the published Admixture Figure S1 where the species re-assignment in Table S4 was based on, we noted that accessions PI604587 and PI604595 were accidently mistaken for accessions PI652432 and PI649587. These later two should have been re-assigned to *A. cruentus,* while PI604587 and PI604595 should have remained to have their original assignment as *A. hypochondriacus* like shown in Figure S1.

In order to update the expected species information prior to any novel genome analysis (**Table 4.1**) and include this new species annotation from (Stetter *et al.*, 2020) as well as existing gene bank annotations of both diversity set accessions, verification of accession information was performed. The new species annotation from (Stetter *et al.*, 2020) has been combined with more recent available accession information obtained from the USDA-ARS Germplasm Resources Information (GRIN) (https://www.ars-grin.gov/) accessed on 07-02-2022, the world vegetable center (AVRDC) (https://genebank.worldveg.org/#/) on 12-12-2023 and National Plant Genetic Resources Centre (https://www.nbri.org.na/sections/npgrc) on 20-06-2022 and all updated annotations compared to previous ones (**Table 4.1**) are provided for both diversity sets in **Supp table diversity set**. Here if the updated gene bank annotations could be found for an accession, a ✓ is provided to confirm verification. This resulted in 28 accessions which could be verified for the 90 accession ARC set, while 109 accessions could be verified for the 121 accession external set (**Supp table diversity set**). In the ARC set, two accessions (ARC-IV-577 and ARC-IV-578) of which the source was previously not known could be found in the USDA Grin database. Only two accessions, i.e., PI511713 and PI642741, (**Supp table diversity set**, in bold) overlap between the two diversity sets, showing that the diversity sets contain largely different sets of accessions.

To obtain the starting point annotations for our new phylogenetic analysis, we examined the species assignments (**Supp table diversity set**) and attempted to reconcile conflicting species assignments. Regarding the 121 accession diversity set, if not re-identified by Admixture in Figure S1 (Stetter *et al.*, 2020), the most recent USDA Grin species annotation is taken as a starting point species. Regarding the 90 accession set species starting point annotations, we chose to stay in line with the information which was initially provided by the ARC (**Table 4.1**), however, annotation verifications of this diversity set were useful for identifying final origin information data. While assessing this data, the species without brackets (see **Supp table diversity set**) are used as a starting point annotation for the current novel phylogenetic analysis in this chapter.

## Alignment and variant calling against two available *Amaranthus* reference genomes reveals read quality and accession selection for analysis

### Mapping coverage aids assessing read quality and accession selection

In order to perform phylogenetic and population analysis, genetic variation data was obtained. Previously, 121 accessions were sequenced as paired end short-reads with Illumina HiSeq4000 obtaining 2 × 100 bp reads (Stetter *et al.*, 2020). The sequencing statistics of the 121 accessions is available from the European Nucleotide Archive (ENA) under the project number PRJEB30531 (Stetter *et al.*, 2020). This describes the reads after trimming and includes fastq results that reveal sequencing quality. The 90 accessions were sequenced by short-read Illumina to a 23-35x genome coverage. Reads were then trimmed, and FastQC was run on the individual raw as well as trimmed reads to obtain the sequencing statistics and quality results (**Supp table sequencing stats div set**). The 90 accession ARC set was paired end sequenced obtaining 2 × 151 bp reads, with average sequence lengths decreasing after trimming but remaining at least 150 bp (**Supp table sequencing stats div set**).

**Table 4.2.** Median per diversity set of mean exon or non-exon coverage per accession mapped to either the *A. hypochondriacus* or *A. cruentus* reference genome. Mapping performed to complete genome and main reference chromosomes only (Excl. scaffold). Diversity sets consisting of raw and trimmed sequences of the same 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020).

| **Reference genome:** | | *A. hypochondriacus* | | *A. cruentus* | |
|---|---|---|---|---|---|
| Diversity set sequences | Genome | **Exon** | **Non-Exon** | **Exon** | **Non-Exon** |
| Raw 90 ARC | Excl. scaffold | 12.42 | 11.35 | 13.10 | 11.83 |
| | Complete | 12.33 | 11.25 | 13.10 | 11.81 |
| Trimmed 90 ARC | Excl. scaffold | 12.41 | 11.34 | 13.09 | 11.82 |
| | Complete | 12.31 | 11.25 | 13.09 | 11.80 |
| 121 externals | Excl. scaffold | 13.85 | 13.74 | 14.54 | 14.44 |
| | Complete | 13.75 | 13.61 | 14.53 | 14.42 |

To call variants, all 211 read data sets, were aligned to two available reference genomes of *A. cruentus* (Ma *et al.*, 2021) and *A. hypochondriacus* (Lightfoot *et al.*, 2017) species using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010). Since the ARC set is novel unanalysed data, the 90 raw as well as trimmed reads were aligned in order to investigate differences in mapping efficiency, coverage and quality before and after trimming. It was the previously published *A. cruentus* (Ma *et al.*, 2021) reference which was used for alignment since the updated annotation (manuscript in appendix) was not yet available. To get insight in mapping coverage, the median of mean coverages over all accessions in each diversity set is given in **Table 4.2**. Medians do not vary much between trimmed and raw reads (90 ARC diversity set) or between complete reference genome versus main chromosomes only (**Table 4.2**). Coverage is slightly lower for mapping to the *A. hypochondriacus* compared to the *A. cruentus* reference genome, similarly for non-exonic regions compared to exons. Mapping coverage differences per reference genome is likely due to the *A. hypochondriacus* reference being older and of lesser quality (**Chapter 2**). Potential reasons for exons having higher coverage compared to non-exonic regions is presence of coding genes in exons making them likely more conserved among accessions and present in reference genomes, just as less transposable element activity in exons which is typically selected against and hence not conserved (Bourque *et al.*, 2018; Colonna Romano and Fanti, 2022). Furthermore, more reliable mapping in exons might be aided by a GC content sequencing bias, as exons often have moderate GC levels compared to either GC rich or poor non-exonic regions, which results in efficient and accurate sequencing of exons (Benjamini and Speed, 2012). That the medians do not vary much between mapping trimmed versus raw reads (90 accessions) or between using complete genomes versus main chromosomes only, indicates no major differences between using these methods, justifying the use of both data sources in one analysis.

**Table 4.3.** Number and proportion of accessions in each diversity set showing mapping rates of <10x, 10-15x or >15x coverage to the two reference genomes. Diversity sets consist of raw and trimmed sequences of the same 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Mapping was performed to exonic and non-exonic regions and to the complete genome and main chromosomes only (No scaffolds). Species column describes number and proportion of diversity set accessions mapped to a reference genome of the same species as the accession is assigned to (either *A. hypochondriacus* or *A. cruentus*) per investigated mapping rate (<10x, 10-15x or >15x). If these numbers of accessions changed due to updated species re-classifications, novel numbers are in brackets. If the number of accessions did not match comparing within mapping to (non-) exonic regions or (in-) complete genome, values are coloured.

| | <10x | species | 10-15x | species | >15x | species |
|---|---|---|---|---|---|---|
| ***A. hypochondriacus* reference genome** | | | | | | |
| 90 ARC accessions | | | | | | |
| Complete genome: raw reads | | | | | | |
| Exon | 36: 40% | 2: 5.6% | 32: 35.6% | 9: 28.1% | 22: 24.4% | 4: 18.2% |
| Non-exon | 43: 47.7% | 2: 4.7% | 29: 32.2% | 9: 31.0% | 18: 20% | 4: 22.2% |
| Complete genome: trimmed reads | | | | | | |
| Exon | 36: 40% | 2: 5.6% | 33: 36.6% | 9: 27.3% | 21: 23.3% | 4: 19.0% |
| Non-exon | 43: 47.7% | 2: 4.7% | 29: 32.2% | 9: 31.0% | 18: 20% | 4: 22.2% |
| No scaffolds: raw reads | | | | | | |
| Exon | 36: 40% | 2: 5.6% | 31: 34.4% | 8: 25.8% | 23: 25.5% | 5: 21.7% |
| Non-exon | 42: 46.7% | 2: 4.8% | 30: 33.3% | 9: 30% | 18: 20% | 4: 22.2% |
| No scaffolds: trimmed reads | | | | | | |
| Exon | 36: 40% | 2: 5.6% | 31: 34.4% | 8: 25.8% | 23: 25.5% | 5: 21.7% |
| Non-exon | 42: 46.7% | 2: 4.8% | 30: 33.3% | 9: 30% | 18: 20% | 4: 22.2% |
| External 121 sequences | | | | | | |
| Complete genome: | | | | | | |
| Exon | 12: 10% | 1(-2):8.3% | 101: 83.5% | 19(-23):18.8% | 8: 6.6% | 1(-2):12.5% |
| Non-exon | 9: 7.4% | 1:11.1% | 99: 81.8% | 18(-23):18.2% | 13: 10.7% | 2(-3):15.3% |
| No scaffolds: | | | | | | |
| Exon | 10: 8.3% | 1(-2):8.3% | 103: 85.1% | 19(-23):18.8% | 8: 6.6% | 1(-2):12.5% |
| Non-exon | 9: 7.4% | 1:11.1% | 99: 81.8% | 18(-23):18.2% | 13: 10.7% | 2(-3):15.3% |

| | <10x | species | 10-15x | species | >15x | species |
|---|---|---|---|---|---|---|
| ***A. cruentus* reference genome** | | | | | | |
| 90 ARC accessions | | | | | | |
| Complete genome: raw reads | | | | | | |
| Exon | 25: 27.8% | 3: 12.0% | 26: 28.8% | 2: 7.7% | 39: 43.3% | 6: 15.4% |
| Non-exon | 38: 42.2% | 4: 10.5% | 27: 30% | 2: 7.4% | 25: 27.8% | 5: 20% |
| Complete genome: trimmed reads | | | | | | |
| Exon | 25: 27.8% | 3: 12.0% | 26: 28.8% | 2: 7.7% | 39: 43.3% | 6: 15.4% |
| Non-exon | 38: 42.2% | 4: 10.5% | 29: 32.2% | 2:6.9% | 23: 25.6% | 5: 21.7% |
| No scaffolds: raw reads | | | | | | |
| Exon | 25: 27.8% | 3: 12.0% | 26: 28.8% | 2: 7.7% | 39: 43.3% | 6: 15.4% |
| Non-exon | 38: 42.2% | 4: 10.5% | 27: 30% | 2: 7.4% | 25: 27.8% | 5: 20% |
| No scaffolds: trimmed reads | | | | | | |
| Exon | 25: 27.8% | 3: 12.0% | 26: 28.8% | 2: 7.7% | 39: 43.3% | 6: 15.4% |
| Non-exon | 38: 42.2% | 4: 10.5% | 27: 30% | 2: 7.4% | 25: 27.8% | 5: 20% |
| External 121 sequences | | | | | | |
| Complete genome: | | | | | | |
| Exon | 8: 6.6% | 2: 25% | 87: 71.9% | 17(-22): 19.5% | 26: 21.5% | 3(-4): 11.5% |
| Non-exon | 8: 6.6% | 2: 25% | 92: 76.0% | 17(-22): 18.5% | 21: 17.4% | 3(-4): 14.3% |
| No scaffolds: | | | | | | |
| Exon | 8: 6.6% | 2: 25% | 87: 71.9% | 17(-22): 19.5% | 26: 21.5% | 3(-4): 11.5% |
| Non-exon | 8: 6.6% | 2: 25% | 91: 75.2% | 17(-22): 18.7% | 22: 18.2% | 3(-4): 13.6% |

In order to assess a wider spread of coverage results the number and proportion of diversity set accessions are provided per mapping rates of <10x, 10-15x or >15x coverage in **Table 4.3**. More lower coverage accessions (<10x) were mapped towards non exonic regions while more high coverage accessions (>15x) were aligned to exons, and overall, more accessions mapped better with higher coverages for to the *A. cruentus* compared to the *A. hypochondriacus* reference genome. The number of accessions per a certain coverage section does not seem to vary extremely between using (in-) complete reference genomes or the raw versus trimmed 90 reads (**Table 4.3**). However, occasionally differences in numbers of accessions mapping towards the complete versus main chromosomes only of the reference genome occurred per investigated mapping rate. The observations of differences in alignment to exons versus non-exons or complete versus only chromosomes

of references is emphasized with coloured cells in **Table 4.3**. For example, all cells showing 10-15x mapping rates towards the *A. hypochondriacus* reference genome are coloured since the number of accessions did not match across aligning to (non-) exonic regions or (in-) complete genomes. Additionally addressed in **Table 4.3** is the number and proportion of accessions mapping with the same species assignment as the reference it is aligned to. The number of accessions being the same species as the reference it is mapped to is mostly lower for the <10x compared to the 10-15x or >15x coverage rates for both diversity sets (**Table 4.3**). This shows that accessions of the same species as the reference often achieve higher mapping coverage, likely due to greater sequence similarity with the reference genome, which is an overall indication that alignment occurred correctly.

Generally, all observations in **Table 4.3** are similar as in **Table 4.2**, with potential reasons for results previously explained. The difference in coverage between complete versus main reference chromosomes, but also between raw and trimmed 90 ARC reads were minimal (**Table 4.2**, **Table 4.3**). Trimming is performed to improve read quality, while complete references typically provide more comprehensive and complete genomic data. Hence continued analysis is performed with the trimmed reads mapped to the complete reference genomes, since it is expected to be the best quality data to go forward with.

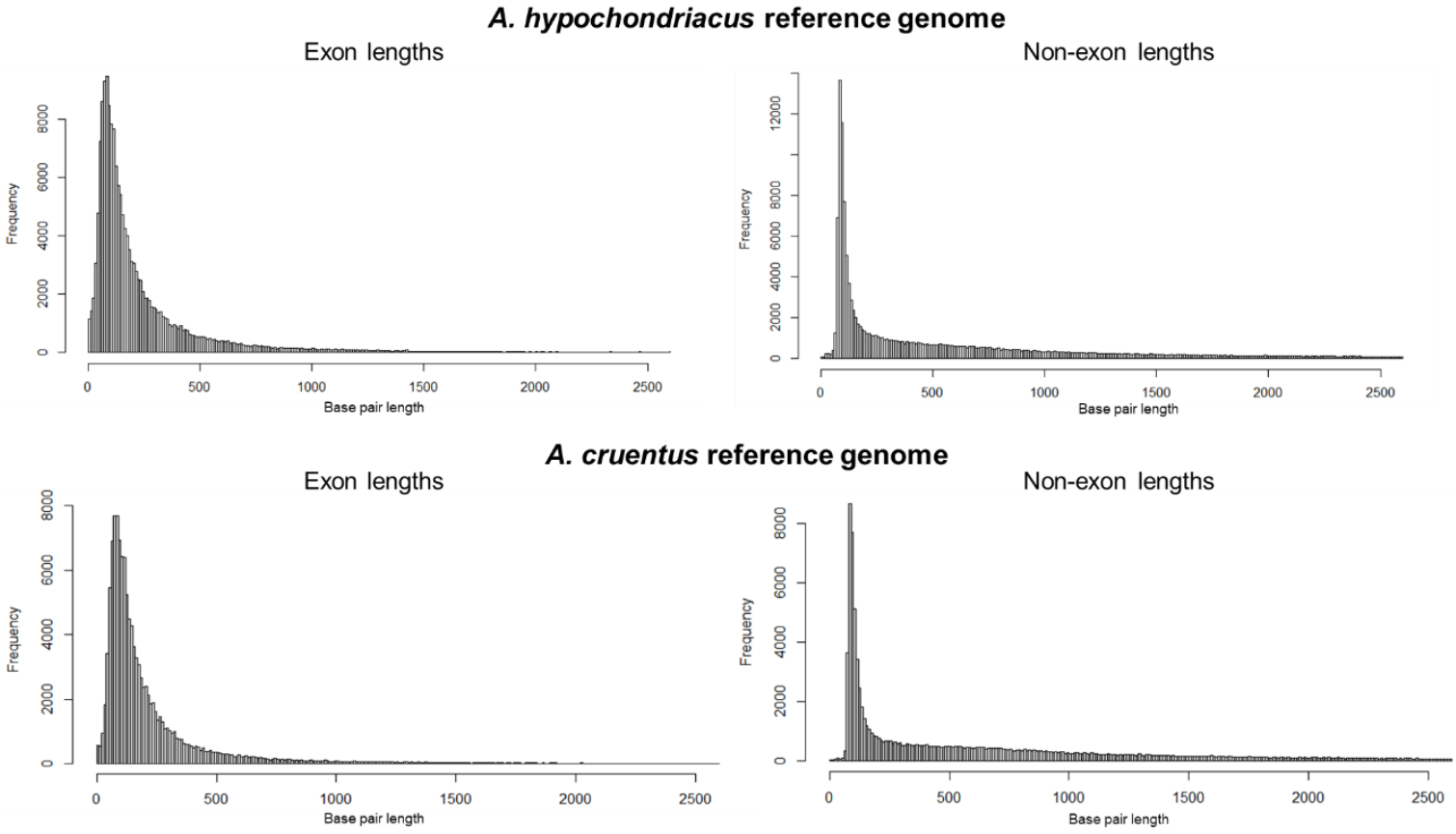



**Figure 4.1**. Frequency of base pair lengths of exons and non-exons of the main chromosomes of the two reference genomes mapped to.

In order to further assess the (non-)exonic regions and quality of the reference genomes that the reads were aligned to, base pair lengths of (non-)exons are plotted in histograms. **Figure 4.1** visualizes the frequency of base pair lengths of exonic and non-exonic regions. More long (>1000 bp) non-exonic regions are recorded than long exons, with especially exon base

pair lengths nicely normally distributed (**Figure 4.1**). This indicates that exons are more uniform in length likely due to stronger evolutionary constraints, whereas non-exonic regions are more variable in size. Generally, for both types of genomic regions base pair lengths are normally distributed for both the reference genomes (**Figure 4.1**). This is an additional indication that the reference genomes we used are of sufficient quality for read alignment, complementary to the original quality assessment provided in reference genome papers, i.e., *A. cruentus* (Ma *et al.*, 2021) and *A. hypochondriacus* (Lightfoot *et al.*, 2017).

To examine mapping of the reads towards the (non-)exonic regions, mean coverage values of all diversity set accessions towards exon or non-exon regions of both reference genomes is plotted in histograms. **Figure 4.2** shows that mean coverage is mostly higher than 10x, which is considered a sufficient coverage to identify heterozygous sites for standard SNP calling procedures (Song, Li and Zhang, 2016). Such high mapping coverage is important for ensuring that variants are called accurately and it can be influenced by sequence quality (Sims *et al.*, 2014). Hence this is a first indication of sufficient mapping towards most (non)exonic regions of the reference genomes with adequate quality sequences. For mapping to the *A. hypochondriacus* reference genome more genomic regions appeared to have 0x coverage compared to mapping to *A. cruentus* (**Figure 4.2**). This suggests *A. hypochondriacus* to be a lower quality reference resource, which is further confirmation of findings in **Chapter 2**. The histogram distribution with mean (non-)exon coverages of the 121 external sequences is narrower and peaks at a higher x coverage compared to the 90 ARC grown accession distributions (**Figure 4.2**). This indicates that that the coverage of the external reads is higher and more uniform than the coverage of the 90 accessions, potentially indicating that the 121 accessions are more similar to the reference genome.

To assess the mapping of individual accessions, rather than comparing overall means, their individual coverage distribution towards (non-) exonic regions have been plotted as boxplots. **Figure 4.3** reveals the 121 imported sequences to have a more stable distribution of and higher exon and non-exon region coverage compared to the 90 ARC grown accessions which have more variable coverages. This could be due to either better and more consistent read quality or sequencing depth of the imported reads (Sims *et al.*, 2014). Additionally, biology might play a role, since the 90 ARC diversity set is likely to be more genetically variable covering a higher variety of distantly related species which do not only belong to the 'hybridus' species complex like the 121 set (**Table 4.1**, **Supp table diversity set**). Of the 90 ARC accessions, those with a low mapping coverage are typically accessions with a low sequencing depth or species less related to the *A. c.* and *A. h.* references (**Figure 4.3**). E.g., *A. tricolor* or unknown *A. sp* accessions mostly map with lower coverage despite average to high sequencing depth (3.86 and 4.12 GB respectively) while 'hybridus' species complex accessions map well with a relatively lower sequencing depth (ranging 3.3 to 4.04 GB). For comparison, the average sequencing depth of all 90 trimmed reads was 3.91 GB. All three *A. dubius* accessions were among the 15 reads with the lowest median exon coverage (**Figure 4.3**) but also had the lowest average sequencing depth of only 2.63 GB. Their lower mapping coverage cannot be attributed solely to lower relatedness to the *A. cruentus* and *A. hypochondriacus* references. The allotetraploid (2*n*=64) genome of *A. dubius* (Grant, 1959; Behera and Patnaik, 1982; Stetter and Schmid, 2017; Lin *et al.*, 2022) likely contributes to this issue. Allotetraploidy often causes multiple similar genomic regions, with structural rearrangements and mismatches relative to the diploid reference genomes. This can all lead to ambiguous read mapping, which are often discarded by read aligners.

The previous alignment in the paper (Stetter, Vidal-Villarejo and Schmid, 2020) where the 121 sequences were obtained from indicated that their average sequence coverage per accession was 9x, and that 94% of reads could be mapped to the *A. hypochondriacus* reference genome (Lightfoot *et al.*, 2017). Our current alignment results of these reads seem to exceed that, with a median of mean accession coverage of at least 13.61 (**Table 4.2**) and the majority of the 121 accessions mapping with >10x coverage (**Table 4.3**, **Figure 4.3**). This suggests adequately performed alignment and likely even improvement over the previous analysis.

To investigate whether accessions should be excluded from analysis, read quality was assessed by examining whether mapping coverage matched sequencing depth. In the 90 ARC set accessions, low mapping coverage typically reflected low sequencing depth or distant relatedness to reference genomes, whereas the 121 external sequences (Stetter, Vidal-Villarejo and Schmid, 2020) sometimes showed low coverage despite large FASTQ or BAM file sizes. Analysis with 'samtools flagstat' (Danecek *et al.*, 2021) identified five accessions with less than 80% of reads mapped to both references: PI643042 (~16.5%), PI604582 (~19.5%), PI511731 (~20%), PI490609 (~41%), and PI649514 (~55.5%). FastQC reports for these five accessions indicated poor 'Per tile sequence quality', suggesting poor read quality rather than low read quantity as the cause of low mapping. The only exception was three FastQC reports of accession PI490609. Based on thresholds of <80% mapped reads and poor 'Per tile sequence quality', these five accessions (PI643042, PI604582, PI511731, PI490609, and PI649514) were excluded from further analysis.

Stetter *et al.* (2020) reported only using high quality reads, however, assessing their Admixture (Figure S1) and phylogenetic tree (Figure S3) analysis, all 121 accessions seem to be used, including the five accessions (PI643042, PI604582, PI511731, PI490609 and PI649514) which do not meet our criteria for read quality or continued analysis.

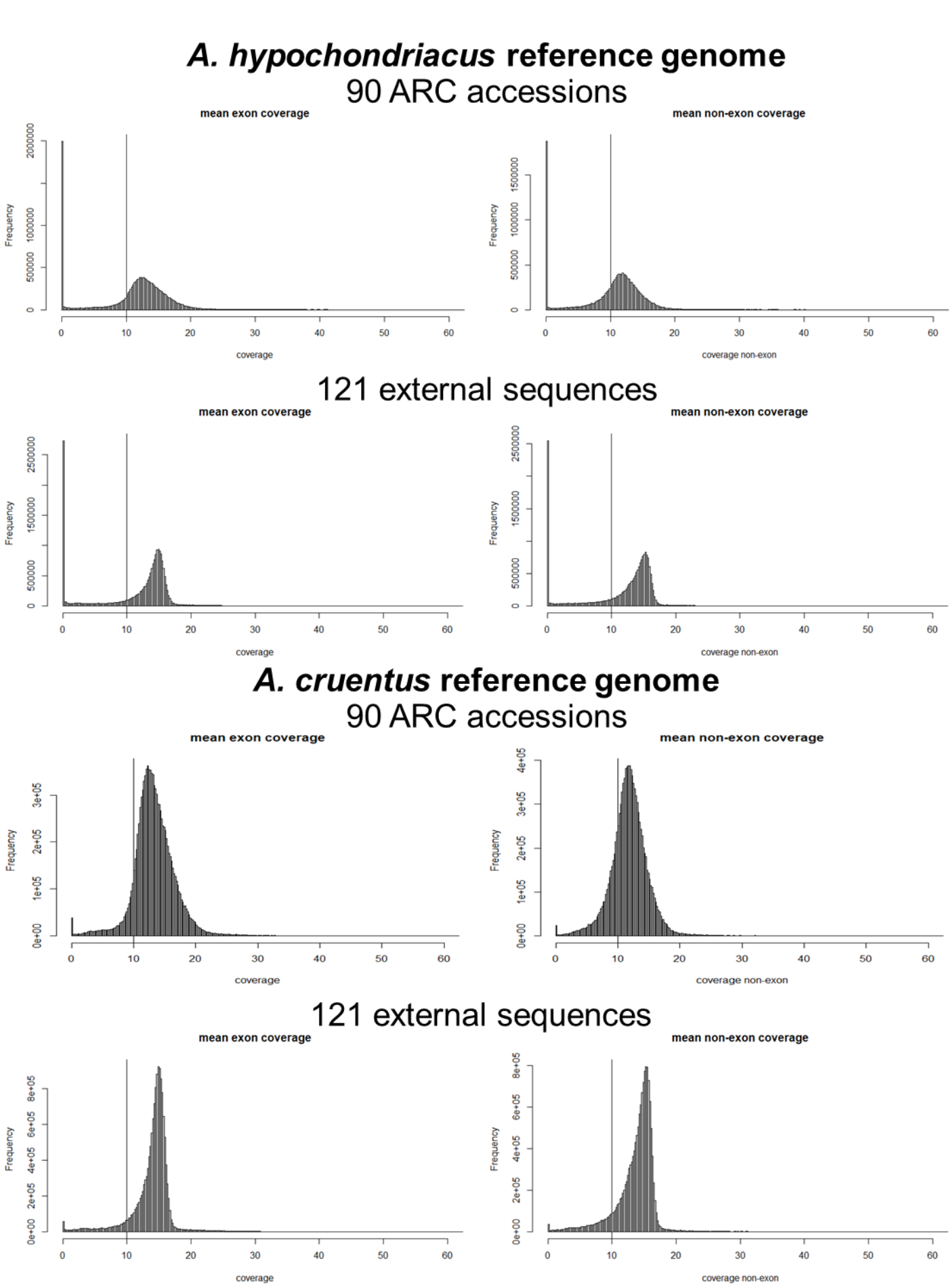



**Figure 4.2.** Histograms showing the frequency of mean coverage values for exon and non-exon regions of all accessions of a diversity set combined. Diversity sets consisting of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Reads were mapped to *A. hypochondriacus* (Lightfoot *et al.*, 2017) or *A. cruentus* (Ma *et al.*, 2021) reference genomes.

# *A. hypochondriacus* reference genome

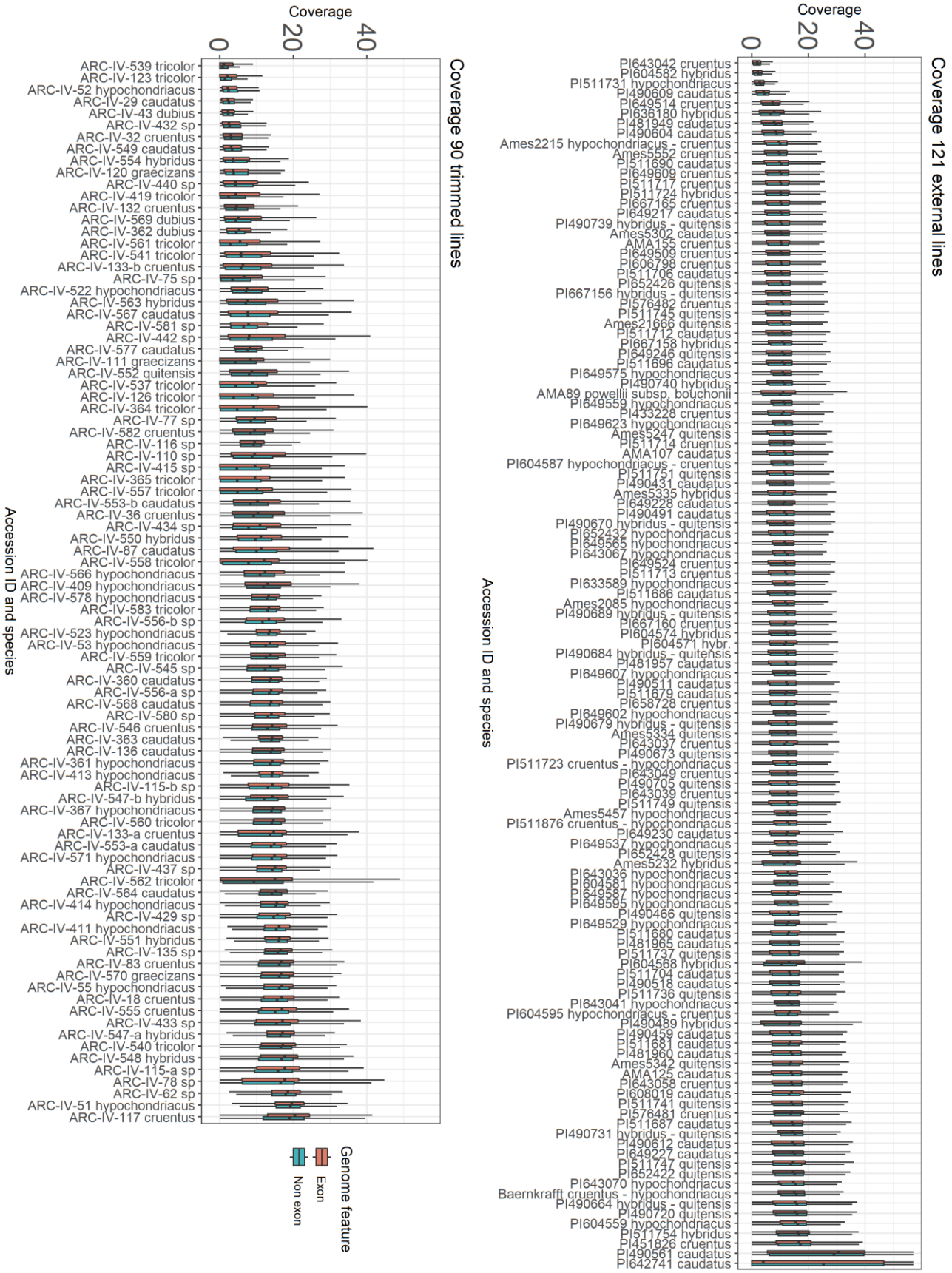



**Figure 4.3. A)** Boxplots with coverage per sequenced amaranth accession in two diversity sets mapped to an *A. hypochondriacus* (Lightfoot *et al.*, 2017) reference genome. Coverage is per exon or non-exon region and ordered by median of the exon coverage. Diversity sets consist of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Species are given behind accession ID's with in case re-classified by (Stetter *et al.*, 2020), original gene bank annotation species followed by reclassified genomic group (gene bank – reclassified).

# *A. cruentus* reference genome

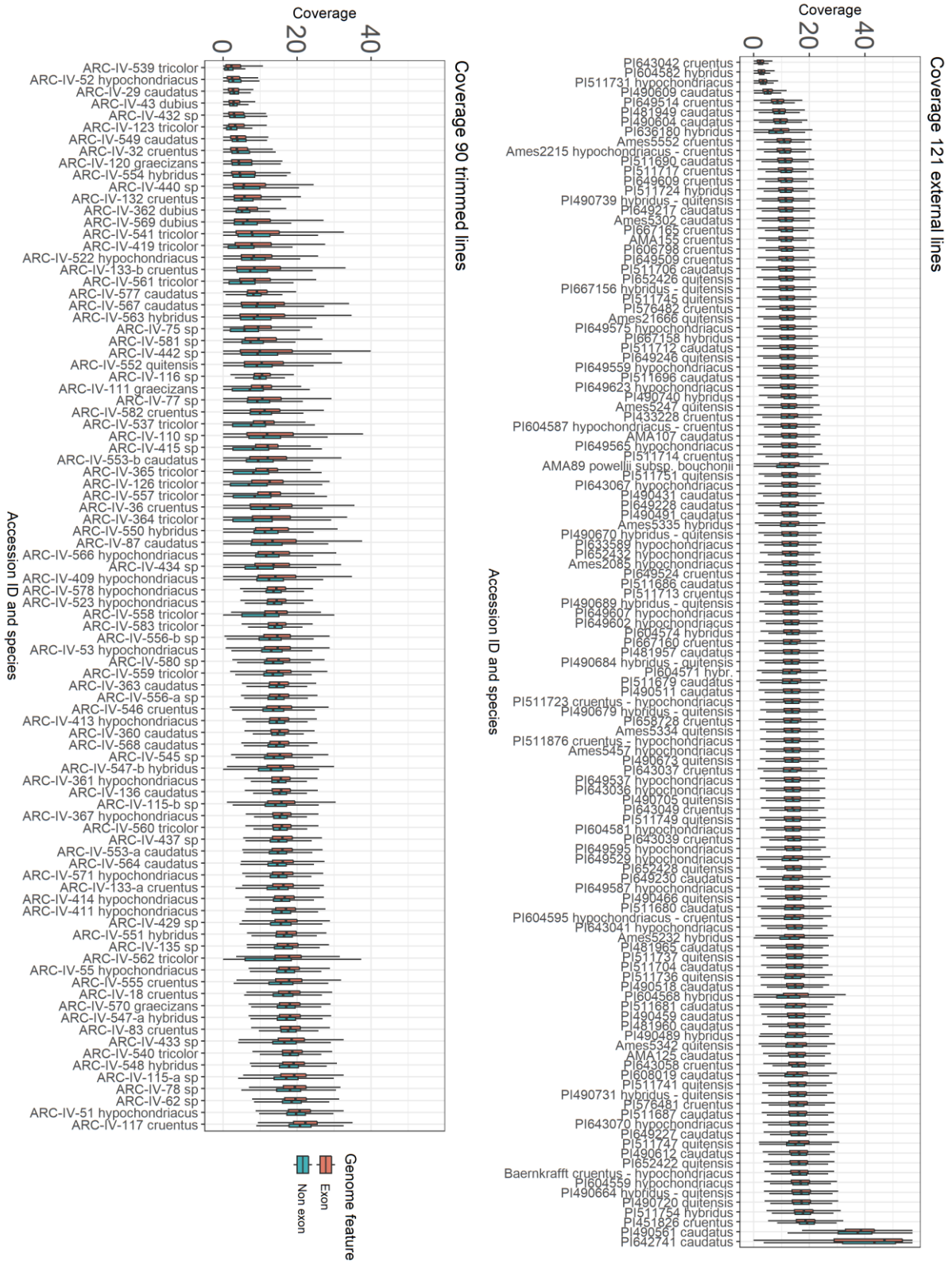



**Figure 4.3. B)** Boxplots with coverage per sequenced amaranth accession in two diversity sets mapped to an *A. cruentus* (Ma *et al.*, 2021) reference genome. Coverage is per exon or non-exon region and ordered by median of the exon coverage. Diversity sets consist of trimmed sequences of 90 accessions grown by the Agricultural Research Council (ARC), and 121 external sequences of (Stetter *et al.*, 2020). Species are given behind accession ID's with in case re-classified by (Stetter *et al.*, 2020), original gene bank annotation species followed by reclassified genomic group (gene bank – reclassified).

### High quality SNP calling against two references with decisions on using the *A. cruentus* genome for continued analysis

After mapping the 90 trimmed ARC read data sets and 116 sufficiently high quality external read sets against two reference genomes, SNPs and indels were called using two variant calling tools; GATK (Auwera and O'Connor, 2020) and freebayes (Garrison and Marth, 2012), accepting the intersection (variants called by both tools). Results were combined into one final output VCF file. To get a first insight in SNP calling results, the number of SNPs called in each separate diversity set and both sets combined, i.e., 90, 116 or 206 accessions, are provided per main chromosomes and scaffolds of the reference genomes in **Table 4.4** and **Supp table SNP count**. From all samples combined, approximately 33 million SNPs were identified on the main chromosomes of the *A. cruentus* reference genome and 29 million from the *A. hypochondriacus* reference. As the *A. hypochondriacus* assembly contains more scaffolds that were not assigned to chromosomes, it contains a larger proportion of the SNPs discovered, i.e., 1.6 instead of 0.4 million SNPs for *A. cruentus* scaffolds (**Table 4.4**).

**Table 4.4.** Number of accessions (Acc.) and SNPs called per main chromosomes (On xx chr.) and scaffolds (On Scaff.) including the number of scaffolds (No. Scaff.) variants were called towards of two reference genomes (*A. hypochondriacus* (Lightfoot *et al.*, 2017) or *A. cruentus* (Ma *et al.*, 2021)). These are results derived from final output VCF files per individual diversity set or sets combined. Diversity sets are a total of 90 accessions grown by the Agricultural Research Council (ARC), sufficient quality externally obtained sequences (Stetter *et al.*, 2020).

| **Diversity sets** | | ***A. hypochondriacus* reference** | | | ***A. cruentus* reference** | | |
|---|---|---|---|---|---|---|---|
| **VCF file** | **Acc.** | **On 16 chr.** | **On Scaff.** | **No. Scaff.** | **On 17 chr.** | **On Scaff.** | **No. Scaff.** |
| **ARC** | 90 | 24,154,474 | 1,294,473 | 1505 | 27,476,609 | 309,370 | 477 |
| **(Stetter *et al.*, 2020)** | 116 | 20,285,264 | 1,195,281 | 1504 | 22,967,913 | 286,097 | 478 |
| **Both sets** | 206 | 29,318,456 | 1,648,744 | 1516 | 33,405,503 | 386,671 | 489 |

Additionally, it can be observed that the 90 accension diversity set has more SNPs called than the 121 accession diversity set (**Table 4.4**, **Supp table SNP count**). This further suggests that this ARC diversity set is more genetically diverse. The pan-genome analysis in **Chapter 2** and mapping coverage results (**Table 4.2**, **Table 4.3**, **Figure 4.2**) concluded that the *A. hypochondriacus* assembly (Lightfoot *et al.*, 2017) is lower quality than the *A. cruentus* assembly (Ma *et al.*, 2021). Current SNP calling results are consistent with this, since more SNPs called towards *A. cruentus* main chromosomes (**Table 4.4**). Hence from here on all continued analysis is performed using variants called against the *A. cruentus* (Ma *et al.*, 2021) reference genome only.

## Assigning species groups based on novel phylogenetic analysis

To perform species demarcation using the diversity sets, phylogenetic, population clustering and principal component analysis (PCA) was conducted with the SNPs called from the full set of accessions (90 trimmed ARC reads and the 116 external) against the *A. cruentus* reference genome (Ma *et al.*, 2021). SNPs were filtered for call quality, and linked SNPs were removed to ensure that phylogenetic information in each SNPs was independent (see methods for details). After filtering 11,696,020 SNPs remained. This data set was used for phylogenetic and population structure analysis.

Genetic population clustering analysis is commonly done to understand the genetic diversity among a population and identify hypothetical subpopulations. These findings reveal the proportion of genetic material inherited from different ancestral groups and hence show potential shared genetic ancestry among accessions. Population structure analysis has been performed in *Amaranthus* populations helping to clarify species boundaries and identify shared ancestry within diversity sets (Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Lin *et al.*, 2022). Here Admixture (Alexander, Novembre and Lange, 2009) (or STRUCTURE (Falush, Stephens and Pritchard, 2003)) analyses are an example in which the proportion of genetic ancestry an individual has from a number of hypothetical subpopulations (k) is visualised, (Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Lin *et al.*, 2022). The number of subpopulations (k), is empirically determined, aided by calculating cross validation (CV) errors which are a way of measuring how well the chosen k models the data without over fitting. High CV errors suggest underfitting of data with too little clusters, while increasing k typically leads to lower CV errors which suggest that the genetic structure is well explained using this k.

To choose the hypothetical population species clusters (k) for our own data we used Admixture (Alexander, Novembre and Lange, 2009), which is considerably faster than STRUCTURE (Falush, Stephens and Pritchard, 2003), defining k values from 1 to 20. CV errors are plotted against different values of k. As expected, CV errors decreased if k increases, which was mostly a gradual decrease with the first increase from k=8 to k=9 (**Figure 4.4**). After k=9 the CV errors drop again and flatten out. Additionally, the choice of best k value can be aided by trying to assign species to the different k clusters based on the already existing species information of accessions (**Supp table diversity set**, without brackets). For the k=9 Admixture plot (**Figure 4.5**) species groups can easily be assigned per cluster. The k=8 plot showed no clear separation between *A. tricolor* and *A. dubius* species (both in cluster 5), which was similar for the k=10 plot (cluster 2) while cluster 3 in k=10 seems impossible to assign into one potential species (**Figure 4.6**). This suggests that k=8 and k=10 are not suiting the data well since for k=8 multiple species are grouped together in clusters, while for k=10 or higher k values population clusters appear within species. Hence, for biological interpretation, k=9 is the most sensible species groupings (**Figure 4.5**). This was further verified with principal component analysis (PCA) (**Figure 4.7**) and a phylogenetic tree (**Figure 4.10**) with accession grouping based on k=9 subpopulations. Hence for further phylogenetic analysis k=9 is chosen to be the most reliable number of hypothetical species present in our investigated population.

Now the pragmatic decision has been made that k=9 is best suited to our population, the hybridization between accessions in our diversity sets can be investigated. In Admixture analysis, a hybrid is an individual whose genome shows significant mixed ancestry from multiple clusters, visible as a combination of colours in the Admixture plot rather than dominance of a single one. In the k=9 Admixture analysis hybridization can be observed, particularly between subpopulations 8 (*A. caudatus*) and 9 (*A. quitensis*), but also 4 (*A. hypochondriacus)*, 5 (*A. hybridus 1*) and 7 (*A. cruentus)* (**Figure 4.5**). Notably 12 of the 17 accessions originally assigned as *A. hybridus* species are placed within four of the nine subpopulation clusters which are not annotated as *A. hybridus* (**Figure 4.5**, black arrows). Only two small subpopulations, comprising some *A. hybridus* accessions, could be distinguished, and no evidence of admixture between these two was shown (**Figure 4.5**, see *). This is a first indication that *A. hybridus* cannot be justified as a distinguished species.

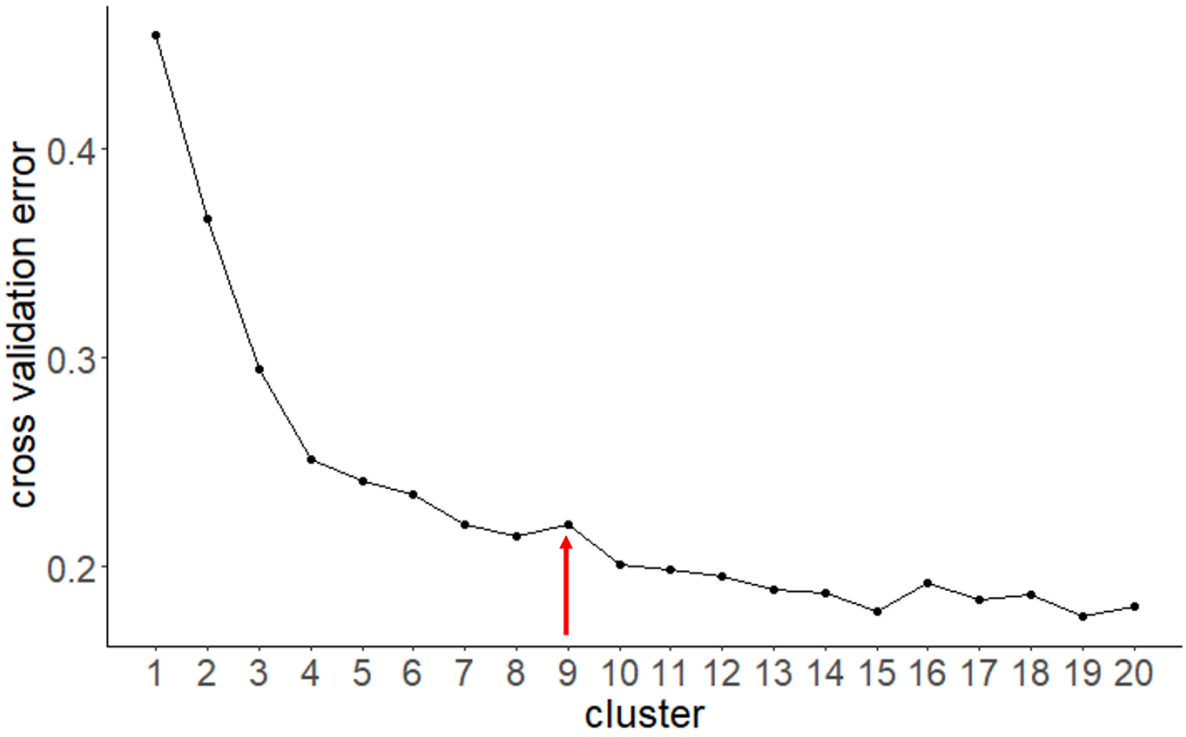



**Figure 4.4.** Cross validation errors from genetic structure analysis of 206 *Amaranthus* accessions mapped to the *A. cruentus* reference genome, shown across different numbers of hypothetical subpopulation clusters (k).

In order to assign species groups to diversity set accessions, the starting point species (**Supp table diversity set**, without brackets) were combined together with the k=9 Admixture results to be plotted over the other population analyses results (PCA (**Figure 4.9**) and phylogenetic tree (**Figure 4.10 B**)) for eventual species proposals. To include the k=9 Admixture information, the species definition of clusters 1 (*A. dubius*) to 9 (*A. quitensis*) were assigned based on the most common accession in their respective cluster. In this assignment we exclude accessions that show extensive admixture (**Figure 4.5**). Here admixture, between species was defined such that accessions with >85% of their genome assigned to a single subpopulation were considered not admixed; otherwise, accessions were classified as admixed. In the cases of admixed accessions, with two subpopulations assigned to their genome, a double species annotation was recorded, e.g. ARC-IV-434 *A. dubius – A. hybridus* 2 and ARC-IV-442 *A. dubius - A. graecizans* (**Figure 4.5**). Some admixed accessions contained more than two subpopulations assigned, with each >15% to their genome. These highly admixed accessions were assigned to be unknown (*A. sp.*), like PI604595 *A. sp* and ARC-IV-77 *A. sp* (**Figure 4.5**). The only exception to this later *A. sp.* assignment was accession AMA89, for which the starting point annotation *A. powellii-subsp.bouchonii* was kept since other phylogenetic results (**Figure 4.7**, **Figure 4.10**) showed this accession to be separating on its own. This suggests that *A. powellii-subsp. bouchonii* assignment for AMA89 is likely true as a single species present in the diversity set which cannot be grouped with the other 9 subpopulations.

To interpret the reliability of the k=9 Admixture results, the proportion of re-assigned accessions based on this analysis is assessed. If results of k=9 with a <85% admixture threshold was to be the absolute guide for species assignment, a high proportion (27%) of the accessions (excluding *A. sp.* accessions) should be re-assigned from the known starting point species (**Supp table diversity set**, without brackets). This is notably higher among the

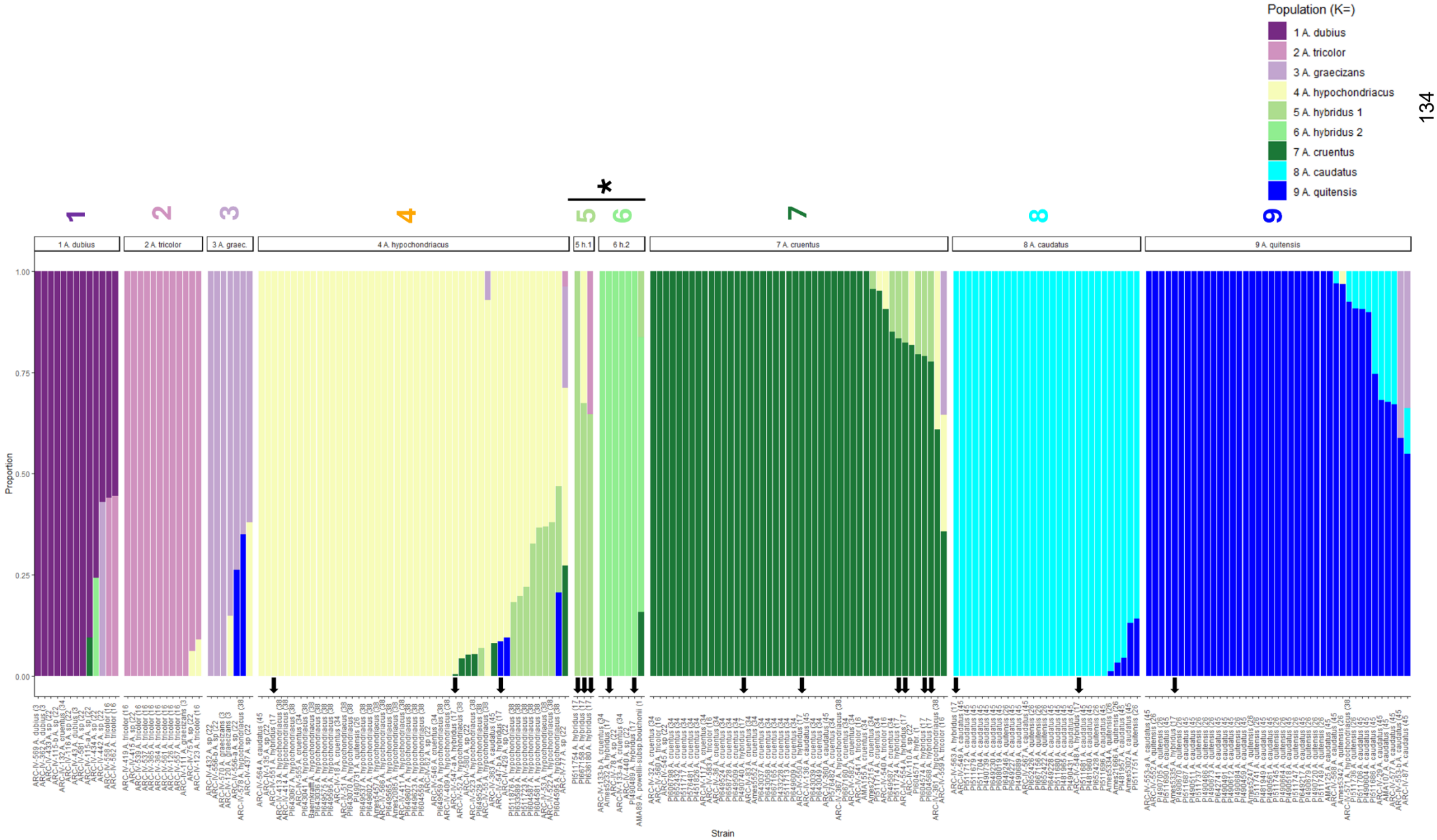



**Figure 4.5** *Amaranthus* accessions clustering together in 9 subpopulation clusters (K=9) which are named after the most common Gene bank species annotation per cluster. Potential *A. hybridus* species clusters are split into two (k=5 and k=6) and marked with *. The Gene bank IDs with their original species annotation (**Supp table diversity set**, without brackets) is given at the x-axis, with numbers of accessions per species mentioned behind the species name. Accessions originally assigned as *A. hybridus* are marked with black arrows displaying their dispersal among the different population clusters.

**Figure 4.6** Population clusters k=8 and k=10 of Admixture analysis after SNP calling in 206 *Amaranthus* accessions mapped to an *A. cruentus* reference genome. The Gene bank IDs with their original species annotation (**Supp table diversity set**, without brackets) is given at the x-axes, with numbers of accessions per species mentioned behind the species name.

ARC accessions (25 of 68: 37%) than among the imported accession diversity set (25 of 116: 22%). This suggests that many of the ARC accessions have incorrect species allocation. This is unsurprising since species definitions were originally based on morphology and gene bank information only, with no previous genome-scale phylogenetic analysis. Additionally, using k=9 for definitive species assignment may be too sensitive, splitting the genetic variation too finely and causing many accessions to be re-assigned. Regarding the external diversity set accessions, all of the re-assignments (25 of 116: 22%) would be from *A. caudatus* to *A. quitensis*, and *A. quitensis* - *A. caudatus* admixture or vice versa. This indicates major difficulty of distinguishing between *A. quitensis* and *A. caudatus* accessions, which was also observed during previous phylogenetic analysis of this diversity set (Stetter, Vidal-Villarejo and Schmid, 2020).

If we accepted the species assignment for admixed accessions (i.e. two species or *A. sp.*, where <85% of their genome defined as one cluster), 42% of accessions would be re-assigned in total, with 55% among the ARC accessions and 35% among the Stetter diversity set. While more accessions could be re-assigned in this way, this approach would likely reassign hybrids with a small portion of genetic material from another cluster. Applying this way of k=9 re-assignment to the 22 ARC *A. sp.* accessions in **Table 4.1**, 16 could be assigned to a new species, 5 to a double subpopulation species and one remaining *A. sp.* This suggests that k=9 Admixture analysis provides useful first insights into potential species allocation for the 22 unknown species accessions, although, results should be interpetated cautiously considering potential hybridization and shared genetic material among the accessions. While implementing k=9 re-assignment there appear to be distinct species or populations, however, this analysis indicates that there has been interbreeding between accessions with many identified as hybrids.

To incorporate further phylogenetic analysis for species demarcation in the diversity sets, PCA results were assessed to complement Admixture results by providing an independent visualization of genetic structure to confirm groupings. PCA was conducted with the same set of 11,696,020 SNPs used for the Admixture analysis, called against the *A. cruentus* reference (**Figure 4.7**). The amount of variation explained by each principal components (PC) decreased from 19.9% to 8.7% for PC 1 to 4 respectively. This means that the genetic structure with variation is spread across several axes, with PC1 capturing the largest proportion of variation (**Figure 4.7**). When plotting PC 1-2, 2-3 and 3-4 with the starting point species annotations (**Supp table diversity set**, without brackets), clear separation of *A. tricolor*, *A. hybridus*, *A. powellii subsp. bouchonii*, *A. dubius* and *A. graecizans* species clusters can be observed (**Figure 4.7**, outlined circles 1, 2, 3, 6 and 6*). This means that these accessions are highly related within their cluster likely belonging to the same species, which is reflected by no overlaps of these clusters with other subpopulation clusters.

Clusters of *A. cruentus*, *A. hypochondriacus*, *A. quitensis* and *A. caudatus* species were present but in a less distinctive way than the five previously mentioned species clusters (**Figure 4.7**, faded circles 4, 7, 8 and 9). This is reflected by cluster overlaps with many accessions likely belonging to another species, or that the entire cluster overlaps with a different subpopulation cluster. Especially the *A. quitensis* and *A. caudatus* groups seem to overlap in all plotted PC's suggesting it is hard to distinguish between them (**Figure 4.7**). These four less distinctive clustering subpopulations clustered closely together in what is visualized as enlarged parts of PC1-2 and PC2-3. This reveals higher relatedness between

these four subpopulations, which can be logically explained since all four subpopulations are identified as hybridus complex species (*A. cruentus*, *A. hypochondriacus*, *A. quitensis*, *A. caudatus* and *A. hybridus*).

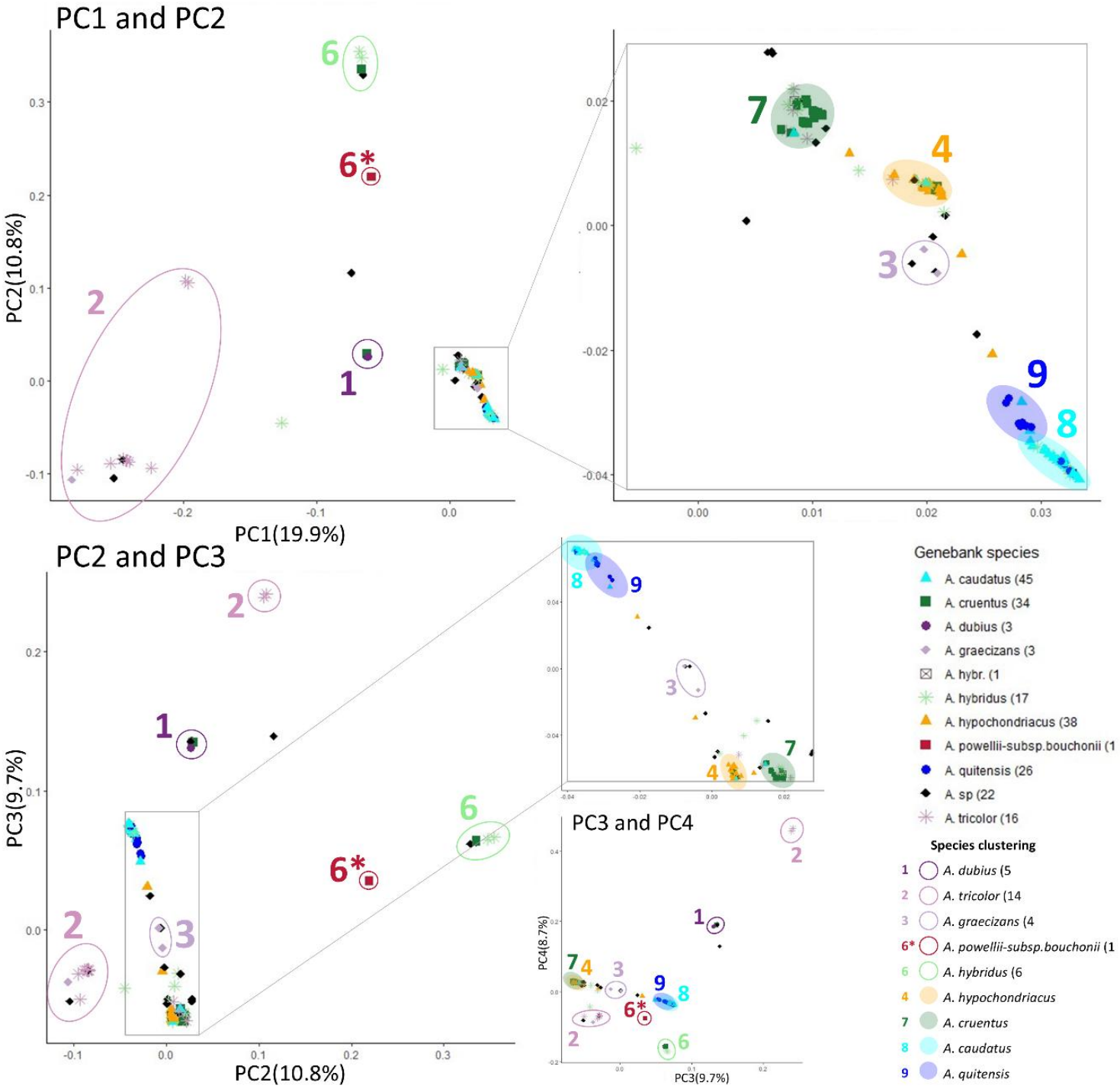



**Figure 4.7** Principal components 1-2, 2-3 and 3-4 plotted for 206 *Amaranthus* accessions based on mapping to an *A. cruentus* reference genome. Individuals are coloured and shaped by the Gene bank original species annotation as provided in **Supp table diversity set** (without brackets) with number of accessions per species given. Observed distinctive species clustering are visualized by outlined circles and numbers of accessions mentioned behind the species name. Less distinctive species clustering is indicated by faded circles with no numbers of accessions mentioned, since those could not be clearly distinguished. The number prior the species circles in the legend and plots is the cluster number of the k=9 species subpopulations in **Figure 4.5**.

These PCA results are complementary with the k=9 Admixture results, showing close genetic relationships between *A. cruentus* with *A. hypochondriacus* and between *A. quitensis* with *A. caudatus* subpopulations, by clustering close together (**Figure 4.7**) and many admixed accessions between these subpopulation combinations (**Figure 4.5**). This could potentially be explained due to geographic origin of these subpopulations, since it is assumed that *A. caudatus* and *A. quitensis* are of South, while *A. cruentus* and *A. hypochondriacus* are of North and Central American origin (Stetter, Müller and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023).

In the enlarged parts of PC1-2 and PC2-3, the *A. graecizans* species are still distinct between the less obvious clustering hybridus complex species groups, however, this *A. graecizans* cluster contains only four accessions and is based on two original annotations only (**Figure 4.7**). This suggests that these four accessions are likely much related with the species complex accessions, although, conclusions regarding separation of this cluster should be drawn cautiously, due to the limited number of accessions as well as initial species annotations. Subpopulations 1 (*A. dubius*), 2 (*A. tricolor*), 6* (*A. powellii subsp. bouchonii*) and 6 (*A. hybridus*) in the PCA (**Figure 4.7**) were distinctive clusters grouping far away from the hybridus complex species clusters. This indicates that these subpopulations are likely less related to the species complex. The *A. powellii subsp. bouchonii* 6* accession clusters on its own (**Figure 4.7**), suggesting that accession AMA89 should be assigned as this single species present in the diversity set. Subpopulation 2 is based on many (14) accessions with most of those being originally annotated as *A. tricolor*, which suggests correct species assignment to this cluster. The many *A. tricolor* accessions group far away from each within their cluster making them seemingly more genetically diverse from another within the subpopulation.

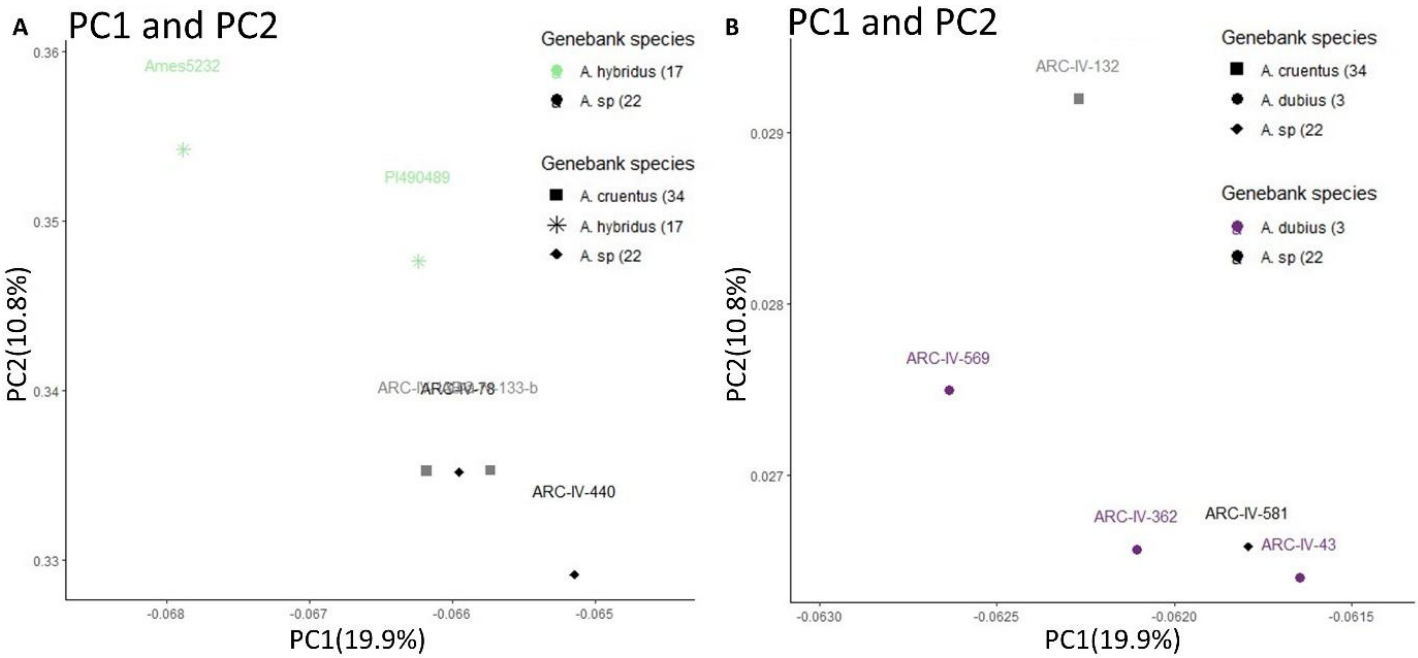



**Figure 4.8.** Principal components 1-2 plotted for *Amaranthus* accessions mapped to an *A. cruentus* reference genome which distinctively cluster into two groups; **A)** *A. hybridus* and **B)** *A. dubius* here visualized in close up. The *A. hybridus* group is based on only two accessions of the available 17 *A. hybridus* annotations in the complete diversity set and therefore less reliable than the *A. dubius* group which contains all three available accessions. Colours and shapes are based on the Gene bank original species annotation as provided in **Supp table diversity set** (without brackets) with number of accessions per species in the total set given.

To investigate the reliability of annotations of the smaller distinctive subpopulations clustering far away from the hybridus species complex, i.e. 1 (*A. dubius*) with five and 6 (*A. hybridus*) with six accessions, a close up of the PCA around these subpopulations have been made (**Figure 4.8**). This shows that the annotation of cluster 6 is based on only two of the 17 available *A. hybridus* gene bank assigned accessions (**Figure 4.8 A**) while for cluster 1 the annotation is based on all three available *A. dubius* accessions in the set (**Figure 4.8 B**). This suggests that the assignment of *A. dubius* to cluster 1 is likely correct, while cluster 6 being annotated as *A. hybridus* is debatable. In case *A. hybridus* were to be the correct assignment for subpopulation 6, it is not expected that these *A. hybridus* accessions cluster far out from other 'hybridus' species complex accessions in the PCA (**Figure 4.7**). Most of the remaining 15 *A. hybridus* gene bank accessions cluster close to the species complex accessions (**Figure 4.7**), therefore *A. hybridus* as a correct assignment to this distinctive cluster is questionable. Despite the species annotation being debatable for subpopulation 6, with *A. dubius* likely being assigned to correctly to cluster 1, accessions in both these subpopulations group close to each other, far away from other clusters (**Figure 4.7**). The geographic background is varying between accessions within both clusters of **Figure 4.8**, i.e., American as well as African accessions are highly related (**Supp table diversity set**). This shows that regardless their different origin, accessions within these subpopulation clusters are likely highly related towards each other and belong to the same species.

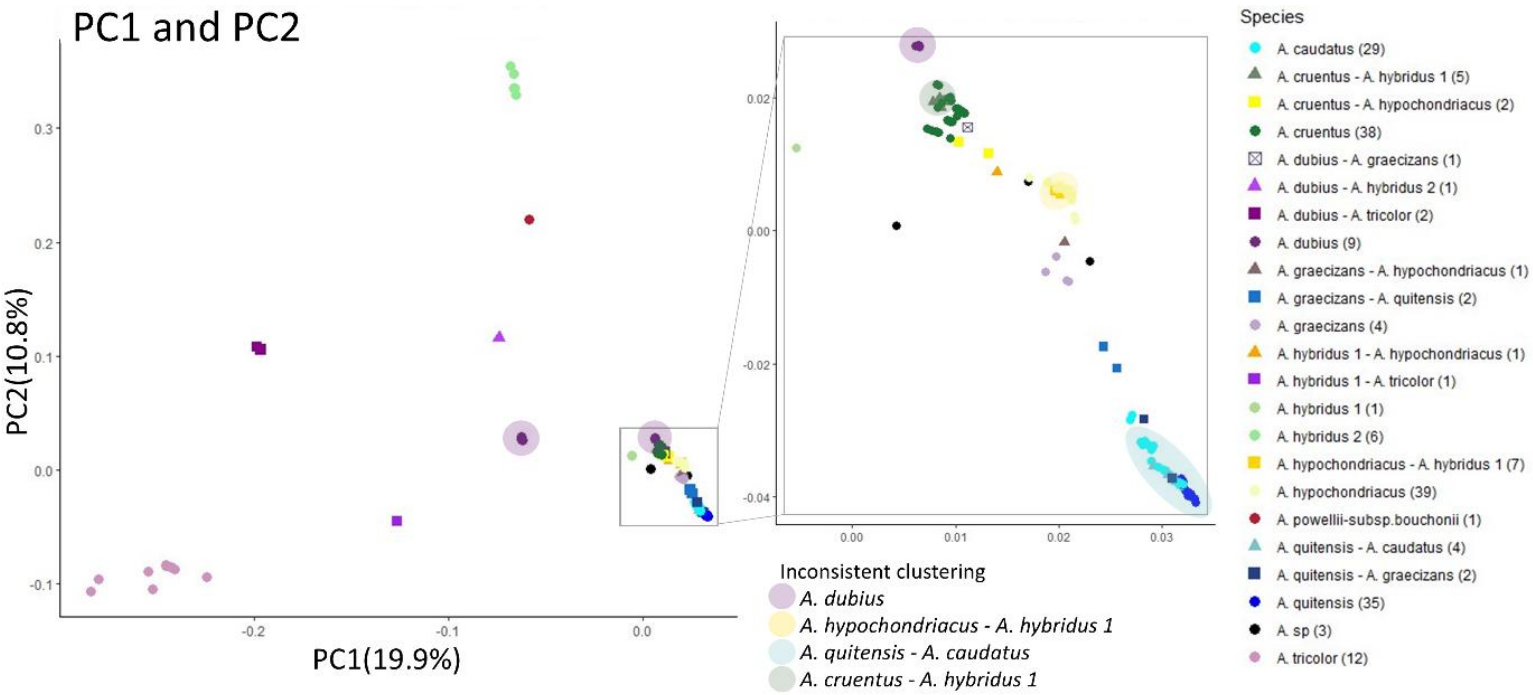



**Figure 4.9.** Principal components 1-2 of **Figure 4.7** plotted with accession annotation according to k=9 Admixture results (**Figure 4.5**, including admixed accessions annotated with double species or as *A. sp.*). Accessions clustering inconsistently with expectations based on k=9 Admixture annotation are highlighted. The enlarged part (right) mostly contains accessions of the 'hybridus' species complex.

To further validate the Admixture results with a <85% threshold to hypothetically be considered real for species assignment, k=9 annotations are plotted in PC1-2 (**Figure 4.9**) and a phylogenetic tree (**Figure 4.10 B**). The Admixture, PCA and phylogenetic results were mostly complimentary. There were a few exceptions of inconsistent clustering which are indicated in the PCA (**Figure 4.9**, faded circled). Some *A. dubius* accessions group separately from the other distinctive Gene bank group identified in **Figure 4.7** and **Figure 4.10 A**, which indicates that these accessions likely belong to a different subpopulation

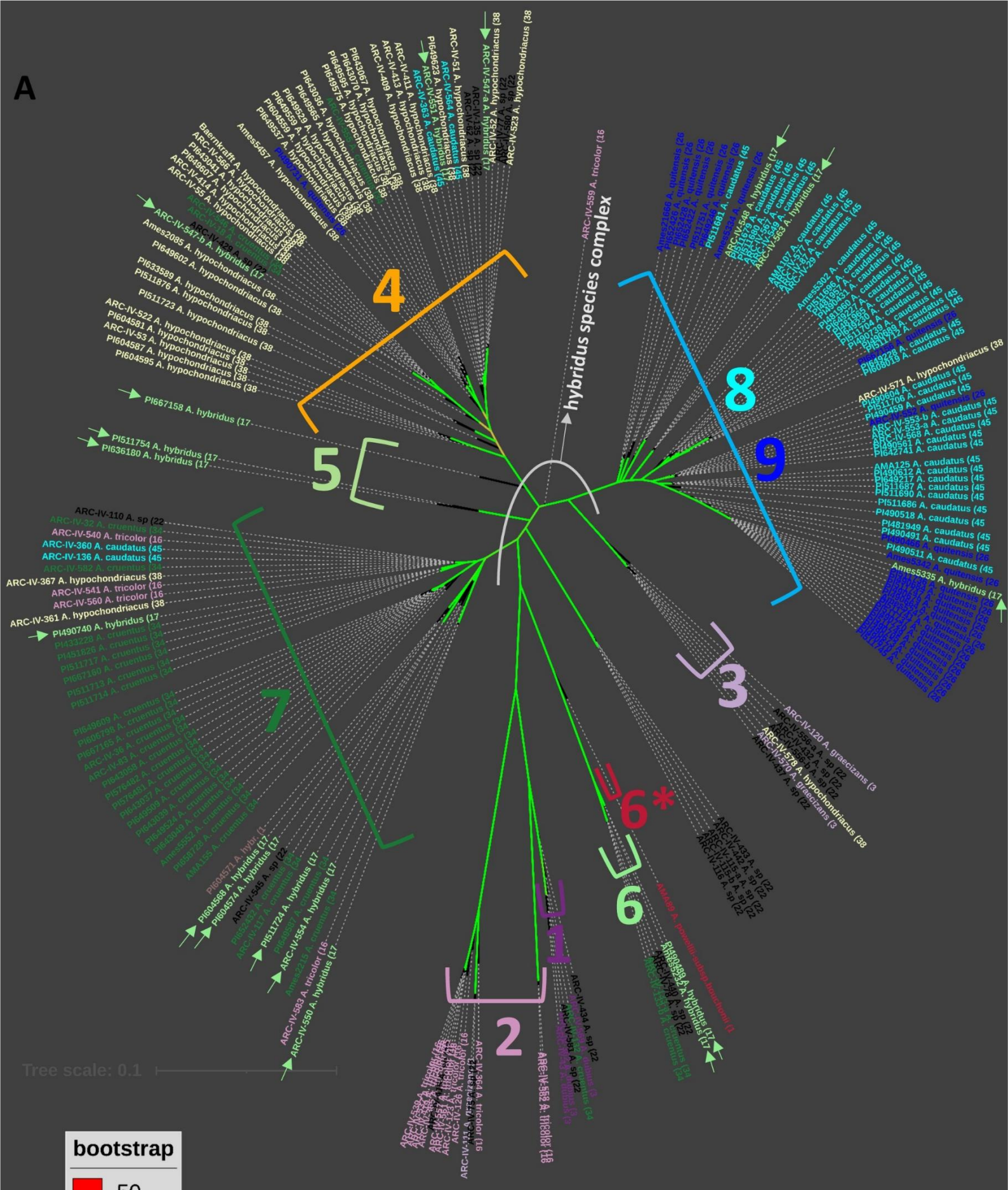



**Figure 4.10. A)** Maximum likelihood trees of *Amaranthus* accessions annotated with their Gene bank species (17 *A. hybridus* accessions annotated all around the tree indicated with green arrows). Clades are numbered and colored with k=9 subpopulations from **Figure 4.5** (*A. powellii* = 6*) and which clades group to the 'hybridus' species complex is indicated in grey. Bootstrap results based on 1000 replicates indicate a well predicted tree apart from some uncertainties within the *A. hypochondriacus* clade.

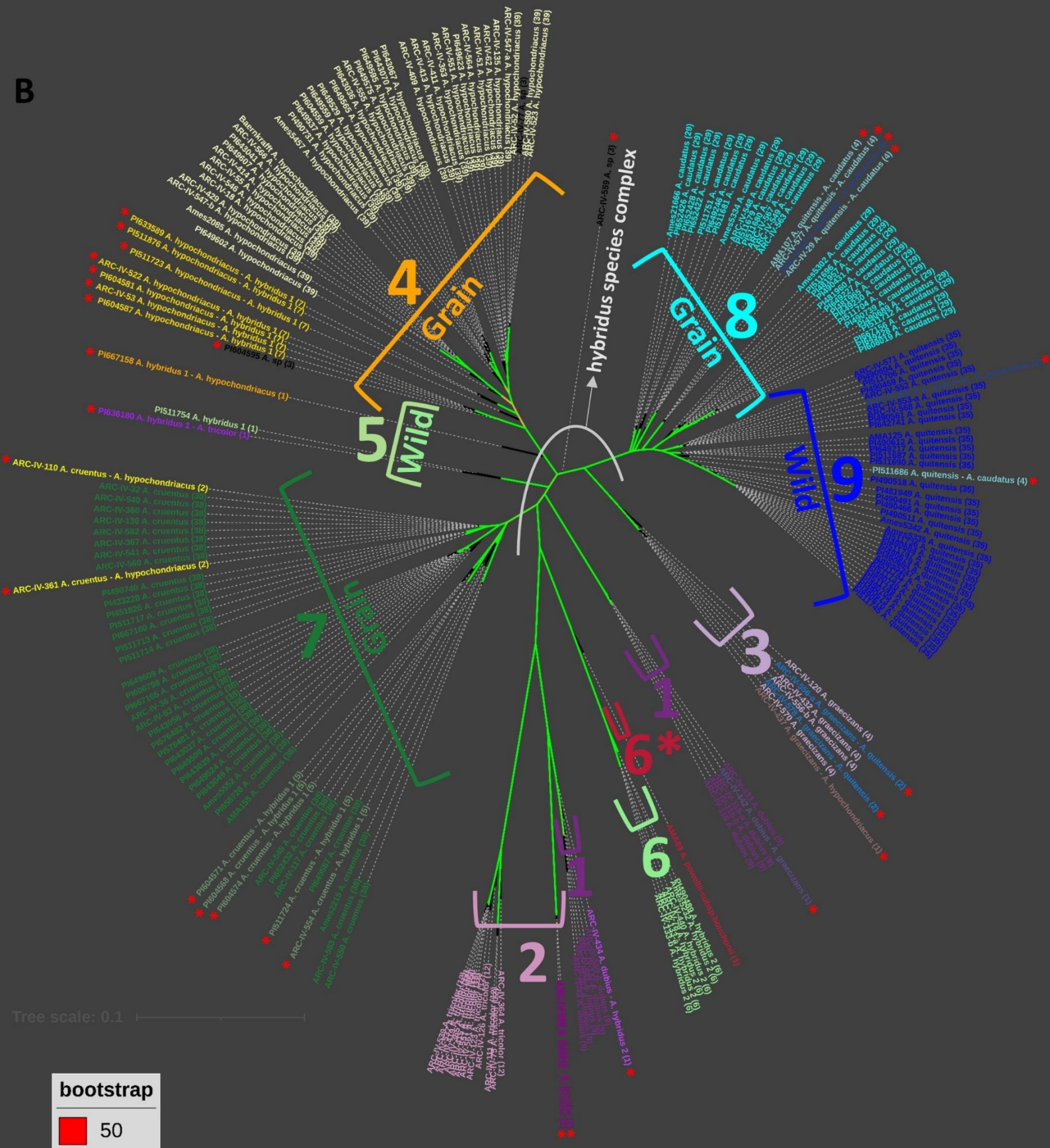



**Figure 4.10. B)** Maximum likelihood trees of *Amaranthus* accessions annotated with species according to Admixture analysis with k=9 subpopulations derived from **Figure 4.5** (including admixed accessions annotated with double species or as *A. sp.* hybrids with red asterisks). Clades are numbered and colored with k=9 subpopulations from **Figure 4.5** (*A. powellii* = 6*) and which clades group to the 'hybridus' species complex is indicated in grey. It is given what is believed to be the three grain amaranths and their two wild related species within the complex. Bootstrap results based on 1000 replicates indicate a well predicted tree apart from some uncertainties within the *A. hypochondriacu*s clade.

which is not *A. dubius*. Some *A. hypochondriacus – A. hybridus 1* clearly cluster with *A. hypochondriacus* accessions, while *A. cruentus – A. hybridus 1* cluster with *A. cruentus* accessions in the PCA (**Figure 4.9**) as well as phylogenetic tree (**Figure 4.10 B**). This suggests that these admixed accessions with double species annotations based on k=9 assignment are likely belonging to the individual subpopulations of *A. hypochondriacus* and *A. cruentus* species. Notably, the bottom of 7 clade in **Figure 4.10 B**, containing admixed accessions (PI604571 to ARC-IV-550) are all present in cluster 4 while all the other *A. cruentus* accessions of clade 7 are in cluster 6 of the k=8 admixture results (**Figure 4.6**). This could suggest that these accessions are present in cluster 4 of k=8 are hybrids separating from other *A. cruentus* accessions, although, a similar phenomenon is not observed for the admixed accessions (PI633589 to PI604595) in the *A. hypochondriacus* clade 4 of **Figure 4.10 B**, with all these accessions present in cluster 7 of the k=8 admixture results (**Figure 4.6**). This further confirms that k=9 is likely the best suiting to our investigated population and another number of k clusters does not clearly reveal hybrids for *A. hypochondriacus* and *A. cruentus* species. Despite some separation in the phylogenetic tree **Figure 4.10 B**, there is still much overlap between *A. quitensis* and *A. caudatus* in the PCA (**Figure 4.9**). This remarks the difficulty of distinguishing between these subpopulations and high relatedness among accessions of these two species.

To conduct species demarcation within our diversity sets, phylogenetic analysis was performed using the same set of SNPs used for Admixture and PCA analysis (Ma *et al.*, 2021). If results matched across different types of population genetic analysis, i.e. Admixture (**Figure 4.5**), PCA (**Figure 4.7**) and a phylogenetic tree (**Figure 4.10**), new species or related groups were assigned (**Table 4.5**). Accession groups, rather than species names, were assigned in cases where relatedness among accessions was shown, but no clear species name could be assigned based on previous knowledge. Here accessions belonging to such subpopulations are indicated as ‘group’ followed by the species name of Admixture k=9 cluster results. In total, 39 accessions in our diversity set could be confidently re-assigned into different species or subpopulation groups, which marks successful analysis validated by various types of populations genetics results (**Table 4.5**). Only two accessions could not be identified based on the current results, but these also did not fit anymore with their previous species assignment. Hence these two accessions are now assigned as unknown (*A. sp.*), revealing that for only a few accessions the available genetic data was insufficient to determine their species identity (**Table 4.5**).

A main aim of conducting this phylogenetic analysis was for species demarcation within the previously not genetically analysed 90 accessions ARC diversity set, especially regarding the 22 accessions with unknown species assignment. Of these 90, 38 ARC accessions could be classified into a species or related group (**Table 4.5**), and about half (17) of these accessions were previously assigned as unknown species (*A. sp.*). Enabling to assign species to 17 of the previously 22 unknown accessions, provides further evidence that the current phylogenetic analysis was successful. The two accession pairs overlapping between the two diversity sets, i.e. PI642741 and PI511713 (**Supp table diversity set**, in bold) were investigated in order to see if these indeed are related and could be derived from the same gene bank germplasm. Accessions PI642741 and ARC-IV-568 which both have the same gene bank ID (**Table 4.5**; PI642741) show high relatedness (**Figure 4.10**) and would both be re-assignment from *A. caudatus* to *A. quitensis* if Admixture k=9 were to be followed (**Figure 4.5**). Hence these two accessions likely originate from the same gene bank germplasm.

ARC accession ARC-IV-555 was re-classified from *A. cruentus* to *A. hypochondriacus* while it shares the same gene bank ID with *A. cruentus* accession PI511713 (**Table 4.5**). Shared gene bank germplasm for these two accessions is therefore less likely, unless ARC-IV-555 has segregated much over many generations of reproduction.

To see if our analysis does comply with the previous phylogenetic analysis of the 121 external accessions, results were compared. Only three of the external accessions were re-classified (**Table 4.5**), indicating the high similarity of our results with the previous analysis of Stetter *et al.* (2020). Ames5232 and PI490489 show high relatedness in the Admixture plot and phylogenetic tree of Stetter *et al.* (2020) which further complements our results. In Stetter *et al.* (2020) all these three accessions were previously annotated as *A. hybridus* (**Supp table diversity set**), and in their analysis none of the *A. hybridus* accessions were re-assigned (Table S4), while most of these accessions showed Admixture (Figure S1) or clustering with accessions of different species annotations (Figure 1 and S3).

Our analysis shows very similar results with scattering of the accessions defined as the *A. hybridus* species among clusters and clades of the three grain species (**Figure 4.5** black arrows, **Figure 4.7**, **Figure 4.10 A** green arrows). Results of both phylogenetic studies are hence complementary, and these do not support the identification of *A. hybridus* as a separate subpopulation or species. *A. hybridus* is considered challenging to identify, notoriously hybridizing, and the common ancestor of the three related grain amaranth species *A. caudatus*, *A. cruentus* and *A. hypochondriacus* (Sauer, 1950, 1967; Chan and Sun, 1997; Mallory *et al.*, 2008; Brenner *et al.*, 2010; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Clouse *et al.*, 2016; Stetter and Schmid, 2017). If *A. hybridus* were to be the ancestral species, it was expected that it would be placed as early branches on the phylogenetic tree, close to the root as early diverging lineages. Even if the created phylogeny is not rooted and multiple domestication events occurred, it is expected that the wild ancestors should cluster outside clusters of domesticated accessions that they gave rise to and not within. This is not the case for our (**Figure 4.10**) as well as the results of Stetter *et al.* (2020) (Figure S3). There is an apparent separation of two *A. hybridus* subpopulations, however, these consist of three accessions in clade 5 (*A. hybridus* 1) of which two are hybrid according to k=9 (**Figure 4.10**), while those three accessions do not cluster (**Figure 4.7**) or hybridize (**Figure 4.5**, see *) with the other subpopulation of clade 6 (*A. hybridus* 2) accessions. Hence, our results together with analysis of Stetter *et al.* (2020) suggest that is it not feasible to distinguish *A. hybridus* as a subpopulation and that it is not the common ancestor of grain amaranths. The three accessions of clade 5 (*A. hybridus* 1) could be potential wild ancestors according to the tree topology (**Figure 4.10**), however, the limited numbers and lacking species annotations are insufficient information for definitive conclusions.

Since there was no clear scientific data to identify *A. hybridus* as a species, we initially choose to not re-assign the accessions previously annotated as *A. hybridus*. That means that if accessions were previously called *A. hybridus* we kept it as such, although many of those are likely another species based on phylogenetic analysis and should ideally be reassigned. The only exception to this are the two *A. hybridus* accessions Ames5232 and PI490489, which are clearly separating and now assigned as Group 1 – *A. hybridus* 2 subpopulation (or species) (**Figure 4.5**, **Figure 4.7**, **Figure 4.8 A**, **Figure 4.10** cluster 6, **Table 4.5**). Our results show scattering of the 15 *A. hybridus* Gene bank accessions species among clusters and clades of the ‘hybridus’ complex species (**Figure 4.5** black arrows,

**Table 4.5.** Certain re-assignment of species and related groups based on combined phylogenetic analysis (Admixture k=9 **Figure 4.5**, PCA **Figure 4.7** and maximum likelihood tree **Figure 4.10**). Three accessions imported from Stetter *et al.* (2020) could be re-assigned while 38 from the ARC set of which 17 previously unknown species (*A. sp.*) could be re-named.

| ID | Gene bank annotation | new species |
|---|---|---|
| **Stetter *et al.* (2020)** | | |
| PI490731 | *A. quitensis* | *A. hypochondriacus* |
| Ames5232 | *A. hybridus* | Group 1 – *A. hybridus* 2 |
| PI490489 | *A. hybridus* | Group 1 – *A. hybridus* 2 |
| **ARC collection** | | |
| ARC-IV-136 | *A. caudatus* | *A. cruentus* |
| ARC-IV-360 | *A. caudatus* | *A. cruentus* |
| ARC-IV-367 | *A. hypochondriacus* | *A. cruentus* |
| ARC-IV-110 | *A. sp.* | *A. cruentus* |
| ARC-IV-545 | *A. sp.* | *A. cruentus* |
| ARC-IV-540 | *A. tricolor* | *A. cruentus* |
| ARC-IV-541 | *A. tricolor* | *A. cruentus* |
| ARC-IV-560 | *A. tricolor* | *A. cruentus* |
| ARC-IV-583 | *A. tricolor* | *A. cruentus* |
| ARC-IV-132 | *A. cruentus* | *A. dubius* |
| ARC-IV-581 | *A. sp.* | *A. dubius* |
| ARC-IV-363 | *A. caudatus* | *A. hypochondriacus* |
| ARC-IV-564 | *A. caudatus* | *A. hypochondriacus* |
| ARC-IV-18 | *A. cruentus* | *A. hypochondriacus* |
| ARC-IV-546 | *A. cruentus* | *A. hypochondriacus* |
| ARC-IV-555 | *A. cruentus* | *A. hypochondriacus* |
| ARC-IV-135 | *A. sp.* | *A. hypochondriacus* |
| ARC-IV-429 | *A. sp.* | *A. hypochondriacus* |
| ARC-IV-580 | *A. sp.* | *A. hypochondriacus* |
| ARC-IV-62 | *A. sp.* | *A. hypochondriacus* |
| ARC-IV-571 | *A. hypochondriacus* | *A. quitensis* |
| ARC-IV-578 | *A. hypochondriacus* | *A. sp* |
| ARC-IV-559 | *A. tricolor* | *A. sp* |
| ARC-IV-111 | *A. graecizans* | *A. tricolor* |
| ARC-IV-415 | *A. sp.* | *A. tricolor* |
| ARC-IV-75 | *A. sp.* | *A. tricolor* |
| ARC-IV-133 a | *A. cruentus* | Group 1 – *A. hybridus* 2 |
| ARC-IV-133 b | *A. cruentus* | Group 1 – *A. hybridus* 2 |
| ARC-IV-440 | *A. sp.* | Group 1 – *A. hybridus* 2 |
| ARC-IV-78 | *A. sp.* | Group 1 – *A. hybridus* 2 |
| ARC-IV-432 | *A. sp* | Group 2 – *A. graecizans* |
| ARC-IV-556 b | *A. sp* | Group 2 – *A. graecizans* |
| ARC-IV-570 | *A. graecizans* | Group 2 – *A. graecizans* |
| ARC-IV-120 | *A. graecizans* | Group 2 – *A. graecizans* |
| ARC-IV-115 a | *A. sp.* | Group 3 – *A. dubius* |
| ARC-IV-115 b | *A. sp.* | Group 3 – *A. dubius* |
| ARC-IV-116 | *A. sp.* | Group 3 – *A. dubius* |
| ARC-IV-433 | *A. sp.* | Group 3 – *A. dubius* |

**Figure 4.7**, **Figure 4.10 A** green arrows). Hence, all these 15 *A. hybridus* Gene bank accessions could likely be re-classified to any of the three different grain species, although, due to the close relatedness of these subpopulations within the species complex (**Figure 4.7** enlarged parts, **Figure 4.10** grey), we opted not to pursue additional reclassification efforts.

Just like in Stetter *et al.* (2020), our results show *A. caudatus* and *A. quitensis* populations to be admixed (**Figure 4.5**) and clustering together (**Figure 4.7**) with close phylogenetic clades (**Figure 4.10**). This further confirms that our analysis complies with Stetter *et al.* (2020), but also supports previous studies recognizing the close relationships between these species, suggesting that *A. quitensis* might be involved in the evolution of *A. caudatus* (Xu and Sun, 2001; Park *et al.*, 2014; Stetter and Schmid, 2017). Due to the difficulty in distinguishing these two species in our datasets, we choose to not re-assign *A. caudatus* to *A. quitensis* and vice versa (**Table 4.5**). In case a hard line was attempted to be drawn based on combined phylogenetic analysis (**Figure 4.5**, **Figure 4.7**, **Figure 4.10**) many (30) accessions (24 of the 121 imported and 6 of the 90 ARC set) would have been re-assigned from *A. caudatus* to *A. quitensis* or from *A. quitensis* to *A. caudatus*. However, making such harsh distinctions between highly related species causes uncertainty of validity and was hence not performed.

To further support our suggestion for updated species annotation (**Table 4.5**), our results were compared with a recent publication of the WVC as well as existing gene bank species annotations. In this study species were updated of 151 out of 731 *Amaranthus* accessions based on double-digest restriction site-associated DNA (ddRAD) sequencing used to perform de novo SNP calling (Lin *et al.*, 2022). Of our 90 accession ARC diversity set, twelve accessions were obtained from the WVC (AVRDC) (**Supp table diversity set**, without brackets), and their external accession ID was matched to information available at https://genebank.worldveg.org/#/ on 12-12-2023. Eight of the twelve accessions could be traced back in the WVC gene bank, of which five accessions match species annotation across WVC gene bank information, Lin *et al.* (2022) (if included in analysis) and our current analysis (**Table 4.6**, green cells). Four of those five accessions were included in analysis of Lin *et al.* (2022), which shows there is great additional support for the updated species annotation of these accessions based on our current analysis (**Table 4.6**).

To investigate the species annotations according to our analysis which were not supported by either the WVC gene bank or Lin *et al.* (2022) (**Table 4.6**, orange cells), our phylogenetic results of these accessions were re-addressed. Our annotation for accession ARC-IV-133(a-b) was into subpopulation group ‘Group 1 – *A. hybridus* 2’, based on the clearly separating cluster 6 (**Figure 4.5**, **Figure 4.7**, **Figure 4.8 A**, **Figure 4.10**). Hence, it can be logically explained that this species assignment does not match with previous external annotations since we assigned a provisional species name to this subpopulation that is not based on existing taxonomic names. ARC-IV-133(a-b) was annotated as *A. dubius* in the WVC gene bank (**Table 4.6**), and our phylogenetic tree results suggest some relatedness of ARC-IV-133(a-b) to *A. dubius* accessions due to the cluster 6 clade not being far from cluster 1 (**Figure 4.10 A**). Hence, results are somewhat complementary suggesting relatedness, although, the clear separation of cluster 6 Group 1 – *A. hybridus* 2 subpopulation in the PCA (**Figure 4.7**, **Figure 4.8 A**), with no admixture shown for these accessions (**Figure 4.5**) suggests that our analysis does not comply with assigning this accession as *A. dubius*.

In our analysis ARC-IV-361 is an *A. cruentus* - *A. hypochondriacus* hybrid according to Admixture analysis (**Figure 4.5**), it is in the *A. cruentus* clade of the phylogenetic tree (**Figure 4.10**), while it is not clustering with either *A. cruentus* or *A. hypochondriacus* accessions in the PCA (orange triangle in between both clusters in **Figure 4.7** PC1-PC2 , or most bottom right yellow square of the two *A. cruentus* - *A. hypochondriacus* accessions in **Figure 4.9**). Hence, we choose to not re-assign ARC-IV-361 and keep the original *A. hypochondriacus* annotation, however, our results clearly reveal ARC-IV-361 might be a *A. cruentus* - *A. hypochondriacus* hybrid. Therefore, the WVC gene bank reclassification of *A. hypochondriacus* to unknown *A. sp.* is complementary to our phylogeny results for ARC-IV-361 showing unclarity of assignment to a single subpopulation for this accession. Lastly accession ARC-IV-132 is annotated as *A. hypochondriacus* in the WVC gene bank, which does not support our results at all, since our analysis shows strong evidence that ARC-IV-132 groups with other *A. dubius* accessions and should be assigned as such (**Figure 4.5**, **Figure 4.8 B**, **Figure 4.10**).

**Table 4.6.** Twelve diversity set accessions obtained from the WVC (AVRDC) for which the ARC ID, external name, starting point species and their verification by the AVRDC website (https://genebank.worldveg.org/#/) is given. These first four columns are the same as in **Supp table diversity set**. In case found, the AVRDC ID, but also species annotation as analysed in Lin *et al.* (2022) and lastly our own species annotation results (**Table 4.5**) are provided in the last three columns. The cells in the last columns coloured green if our own annotation results match the AVRDC gene bank and publication results and coloured orange in case of no match. An '-' in the 'Species in Lin *et al.* (2022)' column means the accession was not incorporated in the publication analysis. '-' in the last column means we choose to keep the starting point species for this accession.

| **ARC-IV-ID** | **External ID** | **Starting species** | **Verified by AVRDC** | **AVRDC ID** | **Species in Lin *et al.* (2022).** | **Own species results** |
|---|---|---|---|---|---|---|
| 18 | TL | cruentus | (hypochondriacus) | RVI000007 | - | hypo. |
| 32 | Arusha | cruentus | | | | |
| 132 | AM-Fune | cruentus | (hypochondriacus) | VI059412 | hypo. | dubius |
| 133a 133b | AM-Gare | cruentus | (dubius) | RVI000005 | - | Group 1 – A. hybridus 2 |
| 135 | IP5 | sp | | | | |
| 546 | Arusha Grain | cruentus | | | | |
| 361 | TOT2355 / PI 604796 / RRC 78S-147 | hypochondriacus | (sp.) | VI044466 | sp. | - |
| 362 | TOT2266 | dubius | (dubius) | VI044377 | dubius | - |
| 363 | TOT2295 / RRC 90 | caudatus | (hypochondriacus) | VI044406 | hypo. | hypo. |
| 364 | TOT4238 | tricolor | | VI047668 | tricolor | - |
| 365 | TOT8789 | tricolor | | | | |
| 367 | TOT4151 / DEN DO | hypochondriacus | (cruentus) | VI047551 | cruentus | cruentus |

Most (5) of the eight accessions present in the WVC database do support our analysis results by WVC gene bank information and analysis of Lin *et al.* (2022) (**Table 4.6**). For only one of the remaining three accessions there was no logical explanation why external annotations do not match our species assignment results (**Table 4.5**). Hence, it can be concluded that our phylogenetic analysis is likely well performed which is strongly backed up by external support of WVC gene bank annotations and Lin *et al.* (2022).

## Varying degrees of polymorphism and genetic differentiation observed across newly defined subpopulations.

To investigate the genetic diversity within the newly assigned species groups as subpopulations, π, i.e., nucleotide diversity, was calculated using the called SNPs as molecular markers in 10kb windows. Here π is the measurement of the degree of polymorphism within a population, which can be useful in crop breeding to examine the genetic variation in crops or to determine evolutionary relationships among species (Kanaka *et al.*, 2023). To assess if the values in π differed per species subpopulations, Tukey's letter codes for pairwise significance were calculated. This shows that π was significantly different for all species groups apart from *A. tricolor* and Group 1 – *A. hybridus* 2 (**Figure 4.11**), revealing these two populations to be genetically equally diverse. Generally, π was low for most groups, with medians ranging 0.0025 to 0.0037, suggesting most of the investigated subpopulations are not very genetically diverse (**Figure 4.11**). However, exceptions with a higher median π were the *A. hybridus* and *A. dubius* species groups, nucleotide diversity was 0.0068 and 0.01 respectively (**Figure 4.11**). Hence these two subpopulations are more polymorphic than the other *Amaranthus* groups, with the latter 0.01 π showing the *A. dubius* species to be 10 times more diverse as humans (π ≈ 0.001).

The highest value of π for *A. dubius*, could be an indication of a large long-term population size, since π is an estimator of the population genetic parameter θ in $\theta = 4Ne\mu$. Here *Ne* is the effective population size and μ the mutation rate. Biologically, this suggests that *A. dubius* has retained more genetic variation than the other subpopulations, due to a historically larger breeding population with limited bottlenecks. That these five accessions do belong to the *A. dubius* species is clearly suggested from our analysis, since they very clearly separate from other accessions in the PCA (**Figure 4.7**), k=9 Admixture plot (**Figure 4.5**) and phylogenetic tree (**Figure 4.10 A**). The higher level of genetic diversity (median π = 0.0068) within the *A. hybridus* subpopulation can be explained by our decision to not re-classify most of these accessions while they could not be clearly distinguished as a species in our phylogenetic results (**Figure 4.5** black arrows, **Figure 4.7**, **Figure 4.10 A** green arrows). This further suggests that calculating π for *A. hybridus* as a species group might not be meaningful in case these accessions actually belong to other 'hybridus' species complex subpopulations and it justifies that *A. hybridus* likely should not be considered a species.

In order to find support for our analysis, the π results of Stetter *et al.* (2020) were compared. π was also the highest (0.0049) for the *A. hybridus* group in analysis of Stetter *et al.* (2020) (Fig. 1A) while π ranged 0.0015 to 0.0023 for the other 'hybridus' complex species. This does comply with our current results, only slightly increasing median π for the 'hybridus' complex species now ranging 0.0025 to 0.0037 for the non *A. hybridus* subpopulations (**Figure 4.11**). This suggests that after inclusion of the 90 accession diversity set, the newly defined species subpopulations are little but not notably more genetically diverse, providing

strong support of Stetter *et al.* (2020) complementing our analysis. Notably *A. quitensis*, the only wild species in the species complex (not considering *A. hybridus*), does have similar π results as the three cultivated species in the complex, i.e. 0.0027 versus 0.0025 to 0.0037 (**Figure 4.11**). This phenomenon was also observed in Stetter *et al.* (2020) and hence the expected higher π for this wild species is consistently not observed.

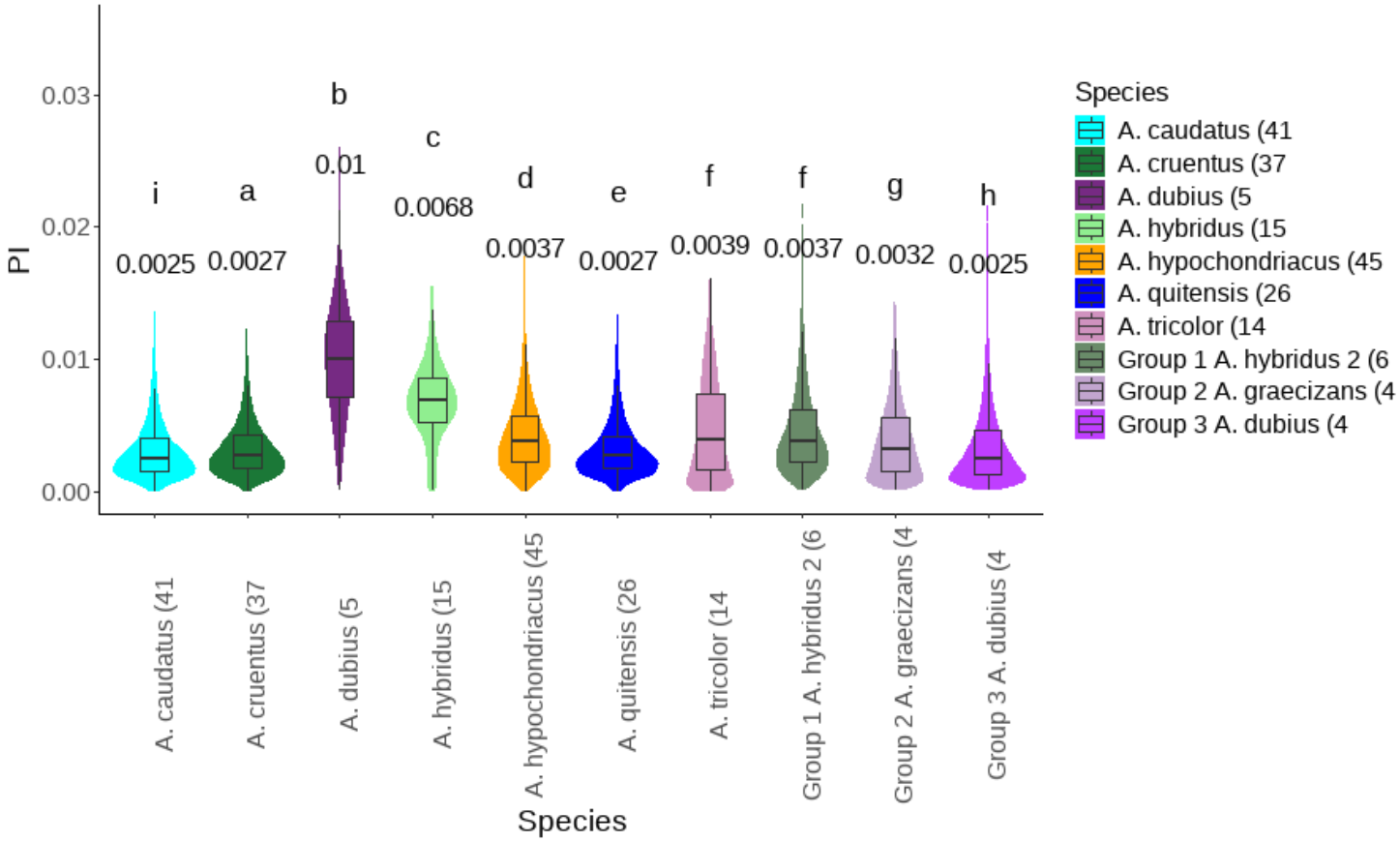



**Figure 4.11.** Nucleotide diversity (π) per newly assigned species groups considered as subpopulations based on phylogenetic analysis (**Table 4.5**) calculated as a distribution over 10kb SNP windows. Tukey's honest significance test letter codes for pairwise significance between species populations and median π per species group are provided. The high π value for the *A. dubius* subpopulation could be an indication of a large long-term population size, while a higher π for the *A. hybridus* subpopulation can be explained by the fact that this group likely should not be considered a species.

To measure the genetic differentiation between the re-defined species groups pairwise fixation index ($F_{ST}$) values are calculated. This was done in order to quantify the genetic distances between subpopulations, providing a more robust numeric estimate compared to solely performing phylogenetic analysis. Here, $F_{ST} \approx 1/(1 + 4Nm)$, where $N$ is the effective population size of each population and $m$ is the migration rate between populations (Whitlock and McCauley, 1999). Hence, $Nm$ is the rate of migration per generation, with low $F_{ST}$ values indicating more migration between those two populations and thus less time apart from each other. $F_{ST}$ will also increase with longer separation times between populations. Low $F_{ST}$ values are found between species groups belonging to the 'hybridus' species complex especially for the *A. caudatus* - *A. quitenis* comparison (**Figure 4.12**, grey triangle, Fst 0.04-0.23). This highlights the expected high relatedness between these subpopulations belonging to the species complex, even between wild *A. quitensis* versus the three cultivated species due to shared evolutionary history and potential gene flow.

On the contrary, higher $F_{ST}$ values are found among the non 'hybridus' complex species subpopulations (**Figure 4.12**, black triangle, $F_{ST}$ 0.27-0.56). This shows that these subpopulations are less related to each other and complements our phylogenetic analysis in which these species groups separate more distinctively. Note that some of these $F_{ST}$ values are very high, e.g. 0.5 (**Figure 4.12**, black triangle), meaning that only half (0.5) of their alleles are shared which indicates that while these species potentially can hybridise, they are not extremely closely related.

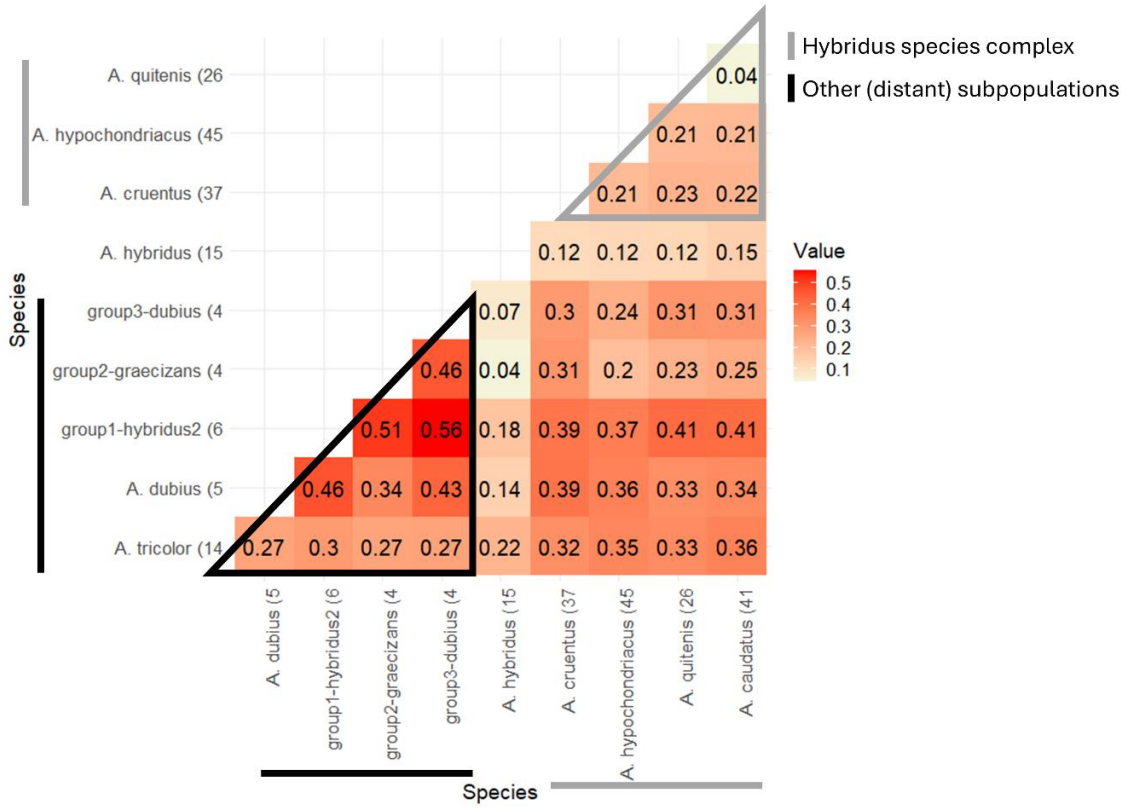



**Figure 4.12.** Pairwise fixation index ($F_{ST}$) values between novel defined species groups (**Table 4.5**) showing less genetic differentiation among the 'hybridus' species complex subpopulations (grey triangle) while more differentiation is observed between other more distant subpopulations (black triangle) confirming our phylogenetic analysis.

A slight exception to the trend of high $F_{ST}$ when comparing distinctive separating subpopulations was observed for the *A. tricolor* species with $F_{ST}$ 0.27-0.3 (**Figure 4.12**, black triangle). This suggests that they are genetically less distant than the other subpopulations are from each other. However, *A. tricolor* does not cluster as distinctively as the other distinctive subpopulations do (**Figure 4.5**, **Figure 4.7**, **Figure 4.10** cluster 2), and consist of many accessions (n=14) which might compromise $F_{ST}$ results. For *A. hybridus* mostly low $F_{ST}$ values are found when paired with all other species groups (**Figure 4.12**, no triangle, $F_{ST}$ 0.07-0.22). This implies that this subpopulation is highly related to all other subpopulations. Although, these results are likely cofounded and provide additional reason that *A. hybridus* should be considered a separate species and that accessions annotated as such should be reclassified to different subpopulations.

From our $F_{ST}$ analysis can be concluded that 'hybridus' complex species subpopulations are highly related while non-complex species are less related towards each other, with results of

*A. hybridus* likely arbitrary due to it not being a species. The implications for plant breeding are that more $F_{ST}$ distant groups are expected to harbour divergent alleles, reflecting population differentiation. These distinct alleles may underlie agronomically relevant traits and can be exploited by breeders as valuable genetic resources to be introduced into improved cultivars.

## *Amaranthus* species show continent specific origins, especially in their native environment the America's

To examine whether the geographic origin of species varies across continents, accessions with known origins are plotted on a world map. For the imported 121 accessions, it is expected that *A. caudatus* and *A. quitensis* originate from South America, *A. cruentus* and *A. hypochondriacus* from Central America, while *A. hybridus* accessions originate from both regions (Stetter *et al.*, 2017, 2020). For 188 of the 206 analysed accessions the origin was known, and this pattern of continent specific origins for the 'hybridus' complex species is still largely present when including the ARC collection (**Figure 4.13**). The only exceptions are three *A. hypochondriacus* accessions originating from South America, one *A. caudatus* accession originating from the US, and four *A. cruentus* accessions originating from Brazil and Peru. Some of these accessions were of Stetter *et al.* (2020), like the four South American *A. cruentus* accessions, and *A. hypochondriacus* PI652432, while the other two *A. hypochondriacus* and one *A. caudatus* accession belong to the ARC set. It can be concluded that our re-definition of subpopulations while adding a 90 accession set complements the expected continent specific origins of the 'hybridus' complex species in their native habitats across the Americas (**Figure 4.13**). Opposite results consider only a few accessions, not solely contributed by inclusion of the ARC diversity set.

To investigate if the species specific geographic distribution of the 188 accessions was influenced by species re-assignment, their starting Gene bank species (**Supp table diversity set**, without brackets) or k=9 subpopulations annotations (**Figure 4.5**) were plotted on the world map. The same distribution of the species complex, with *A. caudatus* and *A. quitensis* assigned accessions originating from South America, *A. cruentus* and *A. hypochondriacus* accessions from Central America and *A. hybridus* from both America's was also observed in the starting gene bank annotations (**Figure 4.14 A**). For k=9 assignments a similar pattern was observed, with notably all admixed accessions of *A. cruentus* or *A. hypochondriacus* with the *A. hybridus 1* subpopulation found in Mexico, which suggests much interbreeding might have taken place between subpopulations of the 'hybridus' species complex in Central America (**Figure 4.14 B**). However, not many accessions could be assigned as *A. hybridus* using the k=9 subpopulations annotations questioning the existence of this species (**Figure 4.14 B**). Generally, can be observed that subpopulations belonging to the 'hybridus' species complex show a clear continent specific origin in their native environment the America's. This pattern was visible with our final subpopulations annotations after species re-assignment (**Table 4.5**, **Figure 4.13**), as well as with intermediate annotations (**Figure 4.14**).This reveals the consistency of results, further complemented by existing literature (Stetter *et al.*, 2017, 2020) as well as PCA where the isolation by distance of the species complex is clearly visible (**Figure 4.7**, **Figure 4.9** enlarged part).

In order to visualise if geographic origin also played a role amongst accessions belonging to the same phylogenetic clades, a per continent colour coded tree was generated (**Figure 4.15**). Many accessions originating from the same continent group together within species clades (**Figure 4.15**). This suggests that continent specific origin contributes to development and interbreeding of closely located accessions and provides additional evidence for the observed clades in the tree. Furthermore, often *A. caudatus* and *A. quitensis* originate from South America, while *A. cruentus* and *A. hypochondriacus* accessions originate from North America, with only a few exceptions of some distinguishable clusters of Asian *A. caudatus* or *A. quitensis* accessions (clade 8-9) and African *A. hypochondriacus* accessions (clade 4) (**Figure 4.15**). This complements the previously notified pattern of continent specific origins for the 'hybridus' complex species.

To see if our novel species annotations complemented continental origin, the maps with final (**Figure 4.13**) and intermediate subpopulation (**Figure 4.14**) assignments were compared with our phylogenetic tree coloured by continent (**Figure 4.15**). The three accessions which were known to be *A. dubius* from the start all came from the Americas (**Figure 4.14 A**, **Figure 4.15**; bottom clade 1), while notably, the other accessions which were newly assigned to be *A. dubius* or *Group 3 – A. dubius* were collected from Africa (**Figure 4.13**, **Figure 4.15**; clades 1). This indicates that our revised species assignments partially reflect geographic patterns, suggesting that *A. dubius* accessions originating from the Americas were initially more easily to be identified as such prior contribution of our phylogenetic analysis. Additionally, can be observed that all four newly annotated *Group 3 – A. dubius* accessions originate from South Africa (**Figure 4.13**, **Figure 4.15**; top clade 1). This could indicate that location plays a role in their subgroup development. However, since we deal with a random collection, accessions belonging to the *Group 3 – A. dubius* subpopulation of different continental origins can simply not be included in our analysis. All of the *A. dubius* or *Group 3 – A. dubius* accessions (previously annotated as well as re-assigned) belong to the ARC diversity set, just as the unknown species (*A. sp.*) accessions of which many originate from South Africa (**Figure 4.14 A**). This shows that the collection of this diversity set, aided by efforts of the ARC, might not be equally represented around the world, with a centralized sampling in (South) Africa. Additionally, it is expected that many accessions in the ARC set were previously mis-identified because their annotation relied on morphological characterisation and amaranth can be quite plastic in their phenotype. Hence, it is no guarantee that the examined accessions represent the natural distribution of the identified species or subgroups and therefore the observed geographic patterns should be cautiously interpreted, especially for the ARC diversity set.

Many of the ARC diversity set accessions tested in South Africa appeared to be assigned to groups for which no clear species name could be assigned based on previous information (**Figure 4.15**; group 1, 2 and 3 with ZAF tags). These groups are distinctive subpopulations with two of them clustering far away from the hybridus species complex (**Figure 4.7**, **Figure 4.8**). After re-assignment seven different subpopulations with a South African Gene bank origin were classified (**Figure 4.13**). This means that these highly genetically different accessions located in South Africa are not derived from a limited species or geographic range and in a far history are likely all originating from different parts of the Americas. Therefore, all these accessions tested in South Africa by the ARC should be further explored since they can provide useful perspectives for breeding, potentially containing relevant and highly varying traits due to their genomic distances.

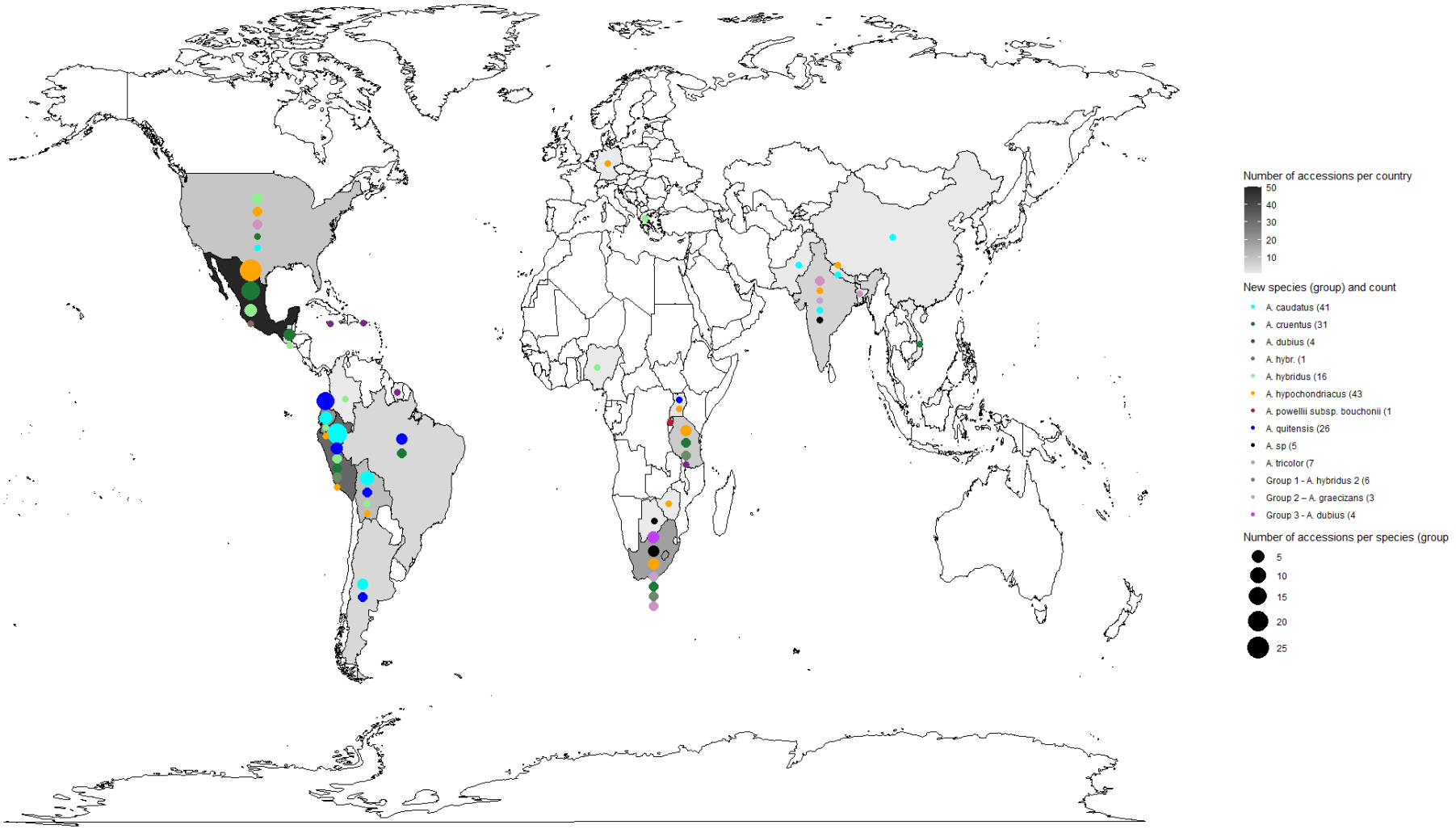



**Figure 4.13** World map displaying the location 188 diversity set accessions of which the origin is known (**Supp table diversity set**). Species and related groups are based on confirmation by phylogenetic analysis (**Table 4.5**). Count of species and groups behind the name in the legend is how often it occurs in the set of 188.

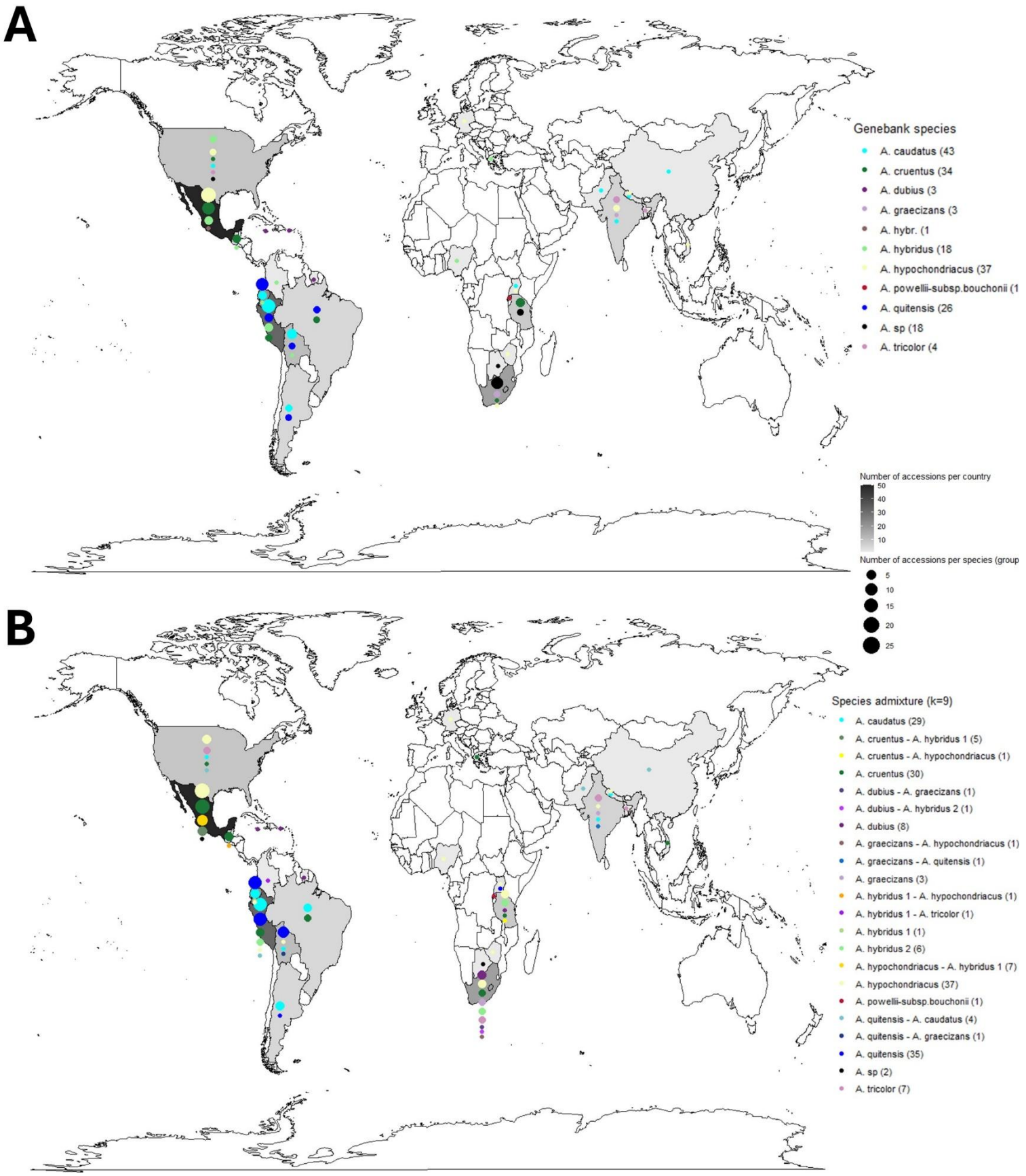



**Figure 4.14** World map displaying the location of diversity set accessions of which the origin is known **A)** Species annotation according to the original gene bank information (**Supp table diversity set**, without brackets). **B)** Species annotation and hybrids (double species) according to k=9 Admixture results (**Figure 4.5**, double species annotation noted for admixed accessions sharing <85% of genome to a single subpopulation or *A. sp.* for admixed accessions with >15% of their genome assigned to more than two subpopulations). Count of species and groups behind the name in the legend is how often it occurs in the set. This is 188 accessions displayed as starting point species in **A)** and 183 accessions used for the phylogenetic analysis of which the origin is known.

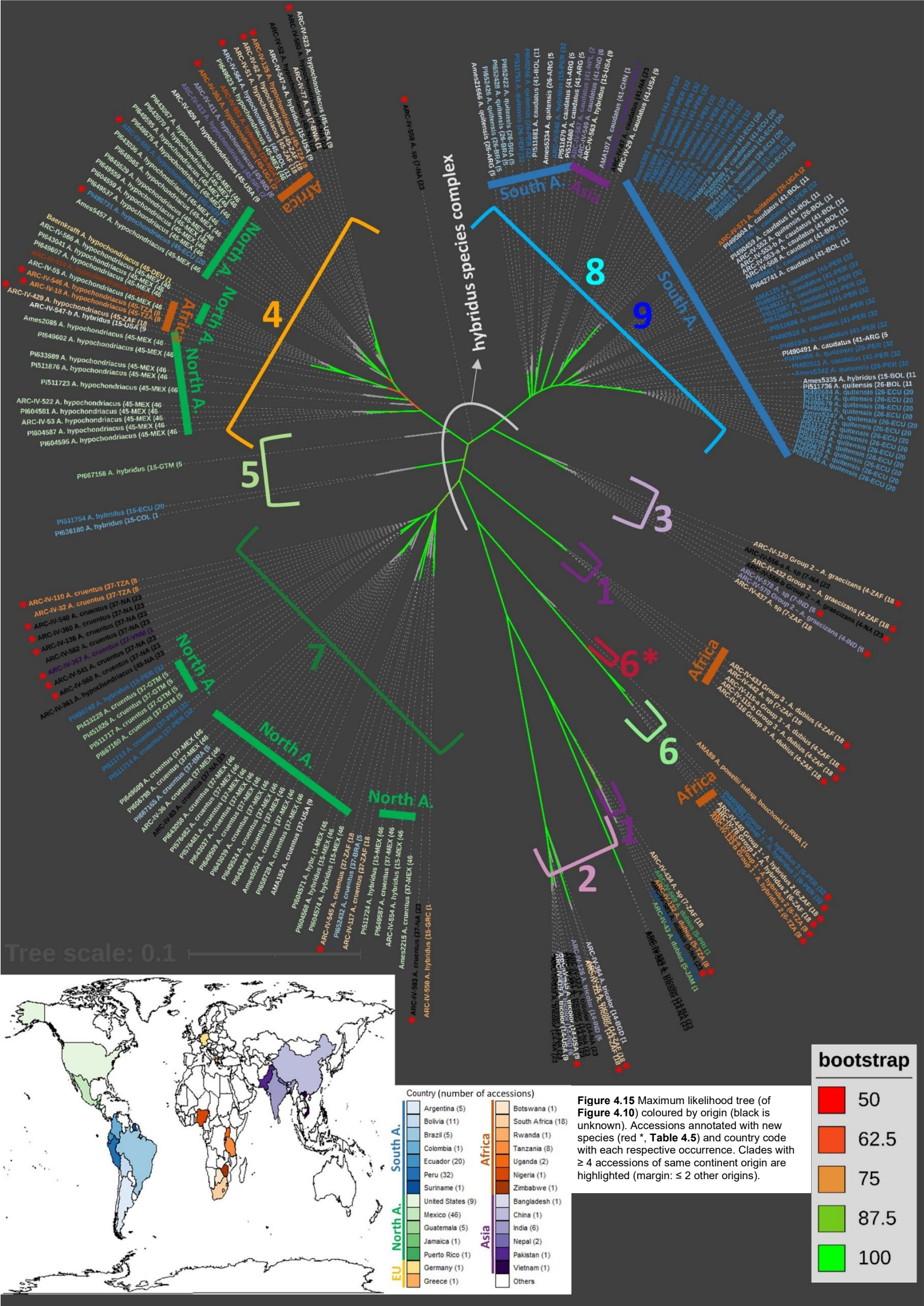



**Figure 4.15** Maximum likelihood tree (of **Figure 4.10**) coloured by origin (black is unknown). Accessions annotated with new species (red *, **Table 4.5**) and country code with each respective occurrence. Clades with ≥ 4 accessions of same continent origin are highlighted (margin: ≤ 2 other origins).

To measure the genetic differentiation between subpopulations with the same origin pairwise fixation index ($F_{ST}$) values were calculated. Lower $F_{ST}$ values were found compared to calculating $F_{ST}$ for novel species groups assignments (**Figure 4.16**: $F_{ST}$ -0.17-0.11 versus **Figure 4.12**: 0.04-0.56). This can be explained due to dividing our dataset in these arbitrary subpopulations by continental origin, which are not subpopulations based on actual species assignment. $F_{ST}$ values between subpopulations by continental origin, revealed a trend of South American accessions having higher $F_{ST}$ values when paired with respectively European, African and Asian subpopulations compared to North American accessions (**Figure 4.16**: $F_{ST}$ -0.04, 0.08, 0.11 versus $F_{ST}$ -0.17, 0.04, 0.09, respectively). This indicates higher relatedness of North American accessions with regions outside the Americas compared to South American accessions, and roughly reveals a path of transportation and migration in the history of *Amaranthus*; i.e., South to North America, to Europe, to Africa to Asia. All $F_{ST}$ values for comparisons involving Europe are negative, likely due to the presence of only two European accessions in our dataset (**Figure 4.16**). This limited sample size may render the results less reliable and should therefore be interpreted cautiously. We observe origin plays an important role in interbreeding and relatedness between accessions even within species groups (**Figure 4.15**, **Figure 4.16**). However, there might be sampling bias, or the native information of the accession might be untrue or lacking due to sharing overtime of the accession between countries. Although this analysis of origin provided further support for the groupings in our diversity set, further studies with broader data sampling are necessary to accurately investigate the natural distribution of *Amaranthus* subpopulations.

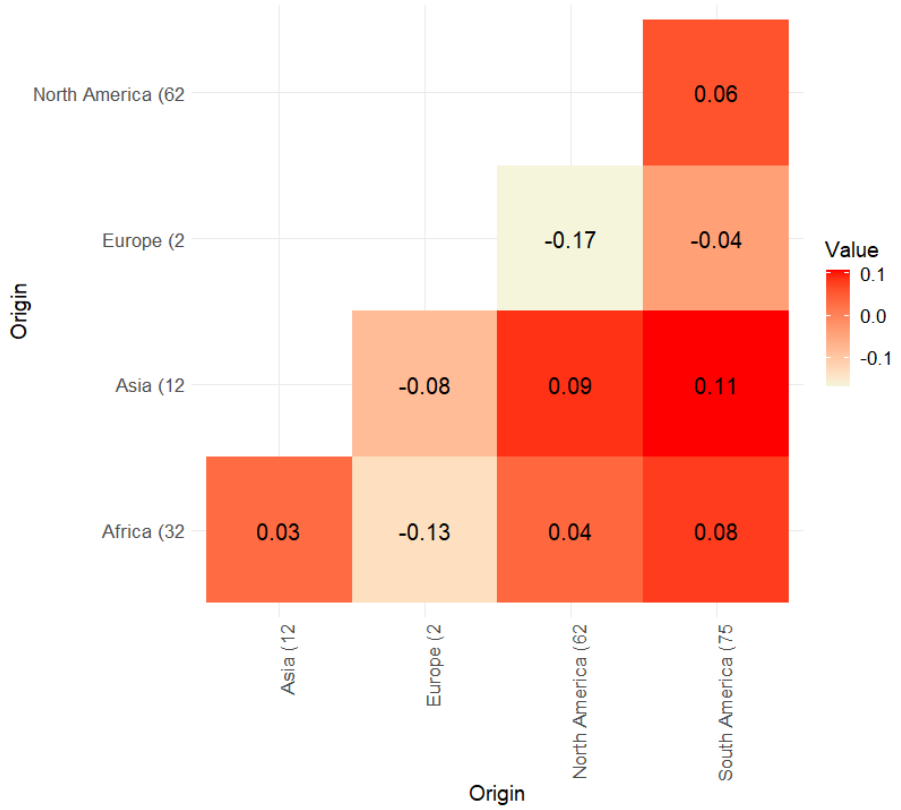



**Figure 4.16** Pairwise fixation index ($F_{ST}$) values between *Amaranthus* subpopulations of various continental origins (**Figure 4.15**) showing a trend of less genetic differentiation of North American accessions to European, African and Asian subpopulations in contrast with the seemingly more distantly related South American accessions. All $F_{ST}$ comparisons with the European subpopulation are negative, likely due to only two accessions belonging to this subpopulation. This causes unstable estimates of heterozygosity with $F_{ST}=(Ht-Hs)/Ht$. Here Ht and Hs are the expected heterozygosity in the total and subpopulation respectively.

## Metabolite clustering roughly confirms phylogenetic species assignment

In order to examine the composition of molecular compounds in the leaves of the 90 accessions, untargeted metabolite profiling was performed in leaf samples taken during a field trial on which untargeted mass spectrometry was performed in two ionization modes. This resulted in 1630 negative and 646 positive detected peaks, with after correcting for redundancy, 1605 features left in the positive and negative peak combined dataset (**Supp table all metabolites**). To check if the normalized metabolite data is of sufficient quality, the two replicate runs of methanolic extracts per accession are plotted in a PCA. The repeats mostly cluster together per accession (**Figure 4.17**). This reveals adequate data for further analysis.

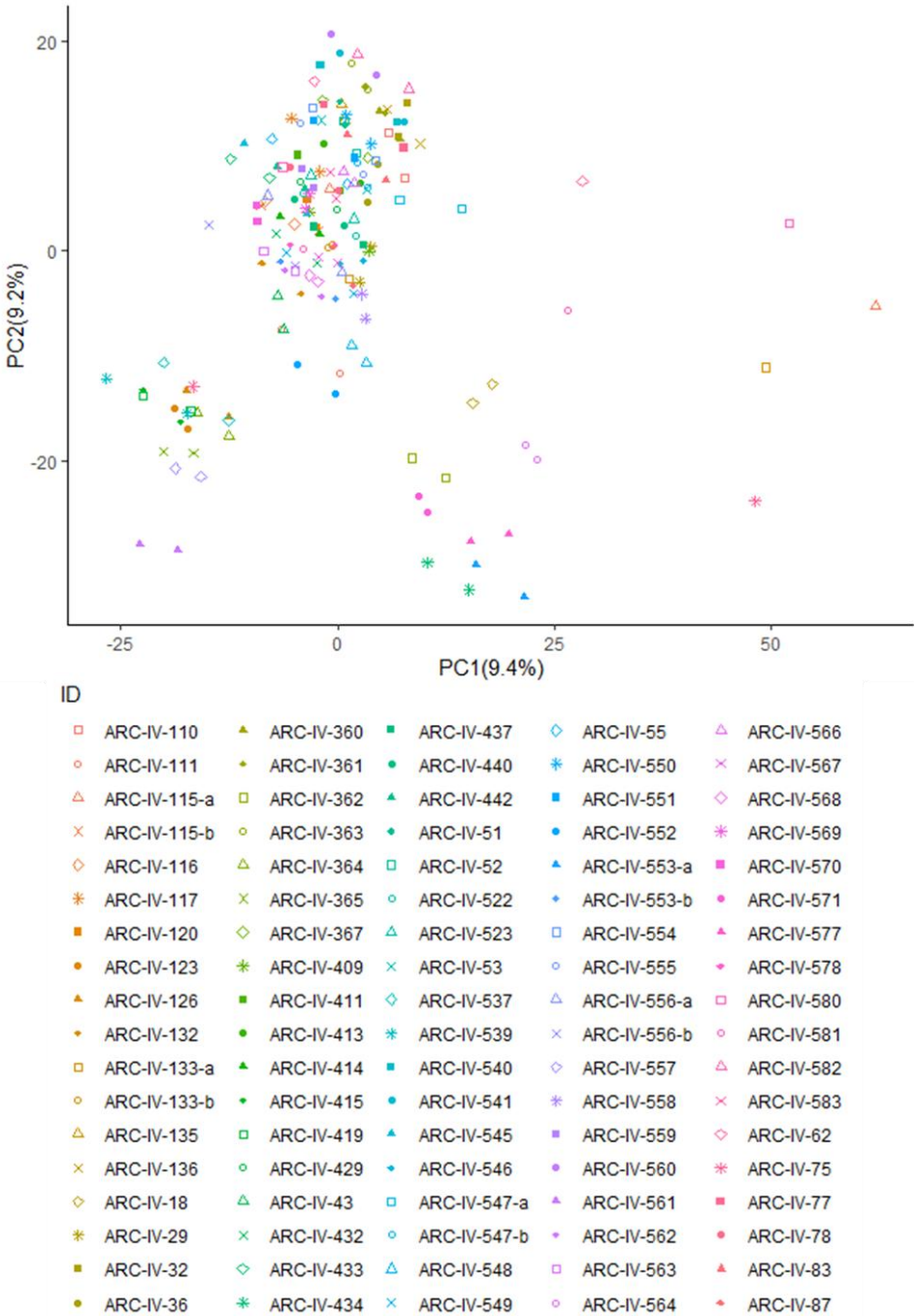



**Figure 4.17** Principal component analysis (PCA) displaying clustering of two individual metabolite repeated runs per 90 diversity set accessions based on metabolite leaf data collected in a field trial. This reveals sufficient quality metabolite data with clustering of the two repeats per accession. Overall clustering of samples might indicate higher relatedness of accessions based on similar metabolite profiles.

In **Chapter 2**, **Figure 2A** the metabolite abundance can clearly be used to distinguish between the two species of the MAGIC parent accessions by PCA. Hence, to investigate the species clustering of the diversity set accessions based on metabolomics, PCA was conducted with the average of the two replicated normalized metabolite data per accession. This showed accessions clustering roughly according to the same as species groups if coloured with the new assignment (**Table 4.5**) after genomic analysis conducted in this chapter (**Figure 4.18**). In particular, ten *A. tricolor* accessions were clustering very distinctively from the other accessions just as they did in the PCA based on SNP calling (**Figure 4.7**, **Figure 4.9**). This accession clustering based on metabolites further confirms our novel subpopulation identifications (**Table 4.5**), with *A. tricolor* accessions likely being the least related to the other accessions of the diversity set.

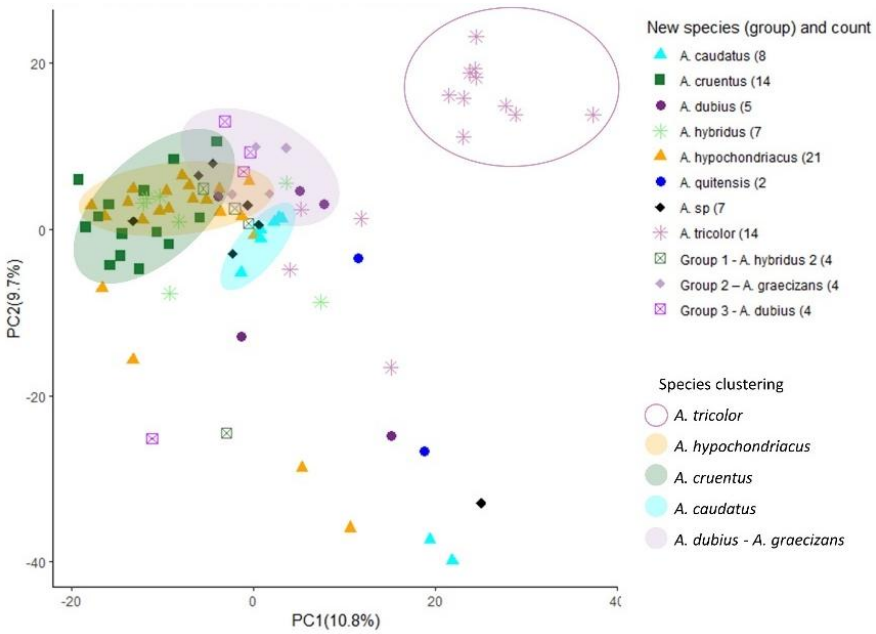



**Figure 4.18** Principal component analysis (PCA) displaying clustering of 90 diversity set accessions based on metabolite data collected from leaves using the average values of two extract run replicates per accession. Observed distinctive species clustering are visualized by outlined circles and less distinctive clustering is indicated by faded circles. Subpopulation species are novel assignments as by **Table 4.5**, with the numbers of accessions mentioned behind the species name.

All the species belonging to the hybridus complex are clustering less distinctively (**Figure 4.18**, faded circles), with *A. hypochondriacus* and *A. cruentus* groups overlapping and together clustering further away from the *A. caudatus* clusters. This indicates that all three species within the '*hybridus*' complex are highly related, with *A. hypochondriacus* and *A. cruentus* being more closely related to each other than to *A. caudatus*, as also confirmed by our PCA based on genomic data (**Figure 4.7**, **Figure 4.9**). Some non hybridus complex species like *A. dubius,* Group 3 *A. dubius* and *A. graecizans* seem to be clustering together, however this is less obvious (**Figure 4.18**, faded circles). This might suggest relatedness between these subpopulations. However, clustering is based on only a few accessions per species, and the subpopulations might separate independently in PC3, PC4. The variation explained by each principal components is low, i.e. ~10% for PC1 as well as PC2 (**Figure 4.18**). This suggests that no PC axis strongly separates the accessions, with the remaining ~80% of the variation spread across other components and PC3 and PC4 explaining even less variation. Hence results should be interpreted cautiously.

To further investigate if subpopulation clustering based on metabolites agrees with our phylogenetic results a tangle gram is created. This tangle gram combines the phylogenetic tree based on SNP calling (**Figure 4.10**) with a dendrogram made using the metabolite data with the average replicates per accession. It shows that most accessions cluster according to the newly assigned subpopulations (**Table 4.5**) in both trees, indicated with coloured blocks (**Figure 4.19**). This additionally provides rough confirmation of correct species

assignment by combined analysis based on genomics as well as metabolomics. The only exception are some the individual non-clustering accessions (**Figure 4.19**, black *), the separated *A. dubius* clade (**Figure 4.19**, purple *), and some separate groups of *A. hypochondriacus* or *A. caudatus* and *A. quitensis* accessions not clustering with the main block based on metabolomic data (**Figure 4.19**, yellow and blue *). The separation of the *A. dubius* clade could be due to two groups of accessions originating from the same geographic regions, i.e. America versus Africa (**Figure 4.13**, **Figure 4.14**), which could cause interbreeding and higher relatedness within these two groups. The separate small blocks of some *A. hypochondriacus* or *A. caudatus* and *A. quitensis* accessions could be explained by these accessions being part of the 'hybridus' species complex which is known to hybridize among subpopulations. Subpopulation clustering results based on metabolites should be interpreted cautiously, since the metabolome of plants is variable. Time of data collection, environment and activity of pathways can influence the metabolome, while the genome is relatively static. Hence, it is expected that the clustering of individuals based on genomic and metabolomic data does not always overlap between these two methods.

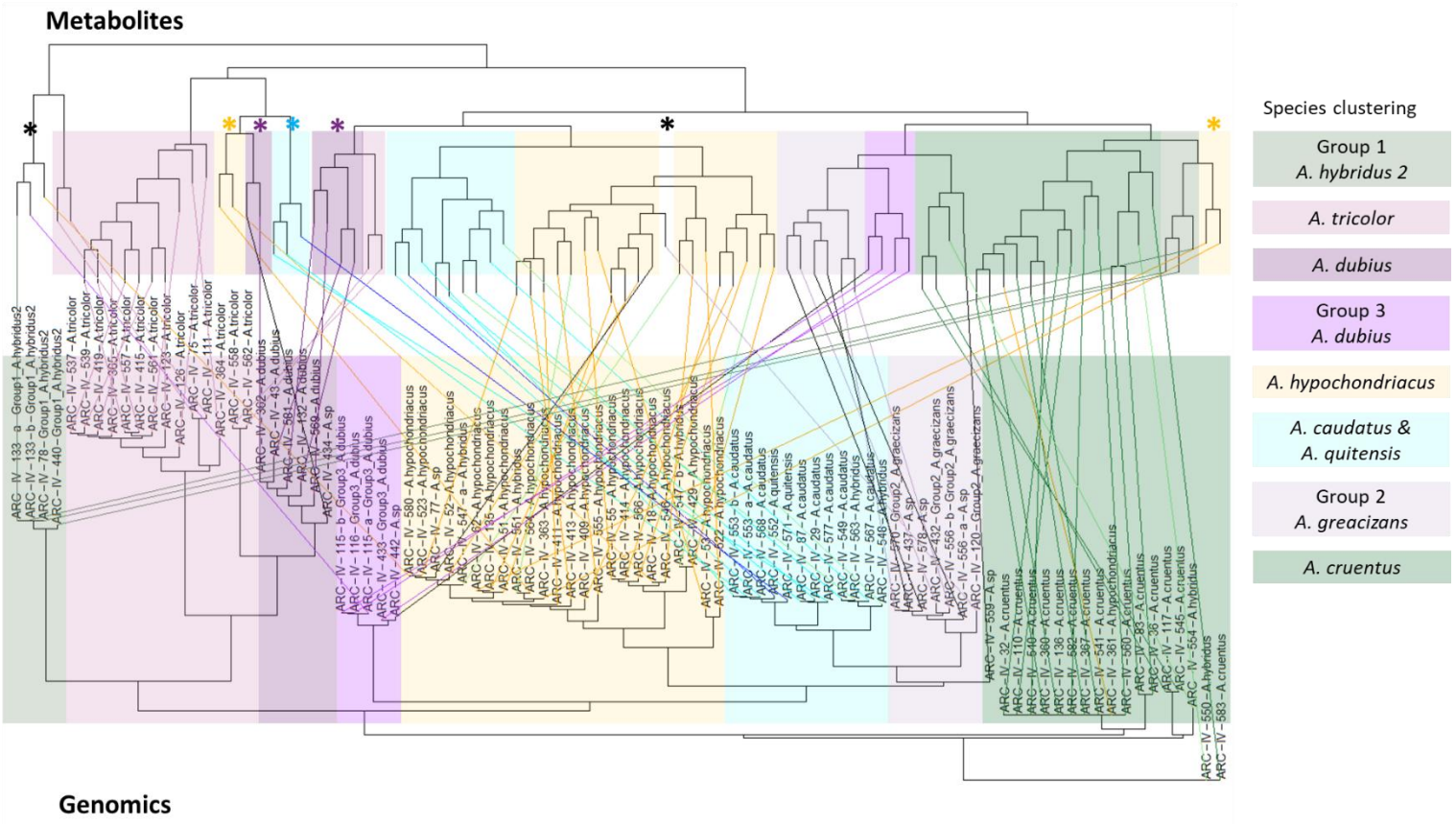



**Figure 4.19** Tangle gram linking a dendrogram based on repeat average metabolic data (top) with a phylogenetic tree based on SNP calling (bottom) of which 90 diversity set accessions were extracted from the tree in **Figure 4.10**. Species clustering blocks were added manually for which species annotation was based on genomic analysis performed (**Table 4.5**). * Indicates individual accessions or species blocks which do not cluster fully in the dendrogram based on metabolomic data.

## Nutritionally relevant metabolites could be identified and occasionally appear species specific

In order to see what the detected features in the combined positive and negative peak dataset are, an attempt to annotate them was performed using SIRIUS4 (Dührkop *et al.*, 2019), GNPS (Wang *et al.*, 2016), PhytoHub (*PhytoHub*, no date) KNApSAcK Version 1.200.03 (*KNApSAcK*, no date) and ChemSpider (*ChemSpider | Search and share chemistry*, no date). 486 of the 1605 detected features could be annotated (**Table 4.7**). 272 of the 486 annotated features could be assigned to a metabolite class, which was identified with either SIRIUS4 or GNPS, while the 214 remaining features could only be annotated with a metabolite name and are put in the 'other' section (**Table 4.7**). Most features belong to lipid metabolite classes like Fatty Acyls and Prenol lipids, indicating that many of these nutritionally relevant components are present in accessions of the diversity set.

**Table 4.7.** Metabolite classes of 486 metabolites annotated by SIRIUS4 and/or GNPS and numbers of those occurring per class in the dataset. If metabolites were annotated with a name only by SIRIUS4, PhytoHub, KNApSAcK or ChemSpider but no class could be annotated those are assigned as 'other'.

| Metabolite class | Amount |
|---|---|
| Allyl sulfur compounds | 1 |
| Alpha-methyldeoxybenzoin flavonoids | 1 |
| Benzene and substituted derivatives | 1 |
| Benzofurans | 1 |
| Benzopyrans | 1 |
| Carboxylic acids and derivatives | 7 |
| Cinnamic acids and derivatives | 1 |
| Coumarins and derivatives | 2 |
| Diazines | 1 |
| Fatty Acyls | 200 |
| Flavonoids | 7 |
| Glycerolipids | 1 |
| Glycerophospholipids | 4 |
| Indoles and derivatives | 1 |
| Keto acids and derivatives | 1 |
| Naphthalenes | 2 |
| Organooxygen compounds | 8 |
| Other | 214 |
| Peptidomimetics | 5 |
| Phenol ethers | 1 |
| Phenols | 2 |
| Prenol lipids | 18 |
| Purine nucleosides | 2 |
| Pyrimidine nucleosides | 2 |
| Steroids and steroid derivatives | 1 |
| Tannins | 1 |
| Total | 486 |

To investigate if certain metabolite classes are more or less prevalent in certain species groups of the diversity set, bi-clustering analysis was performed using the normalized metabolite data. Here we attempt to examine whether certain groups of metabolites cluster more or less with certain amaranth accessions. In order to generate optimal bi-clustering results, the total within-cluster sum of squares is optimized by the bi-clustering algorithm to ensure that each bi-cluster is as homogeneous as possible. The total sum of squared errors (SSE), between metabolite means within a clustering cell and data values within cells, therefore the measure of within cell homogeneity, is plotted against each iteration performed in the bi-clustering analysis. A clear SSE decrease per iteration increase can be observed (**Figure 4.20 A**). This indicates that the bi-clustering algorithm successfully organizes the data into groups (cells) where metabolites have more similar values, which reveals meaningful patterns for biological interpretation.

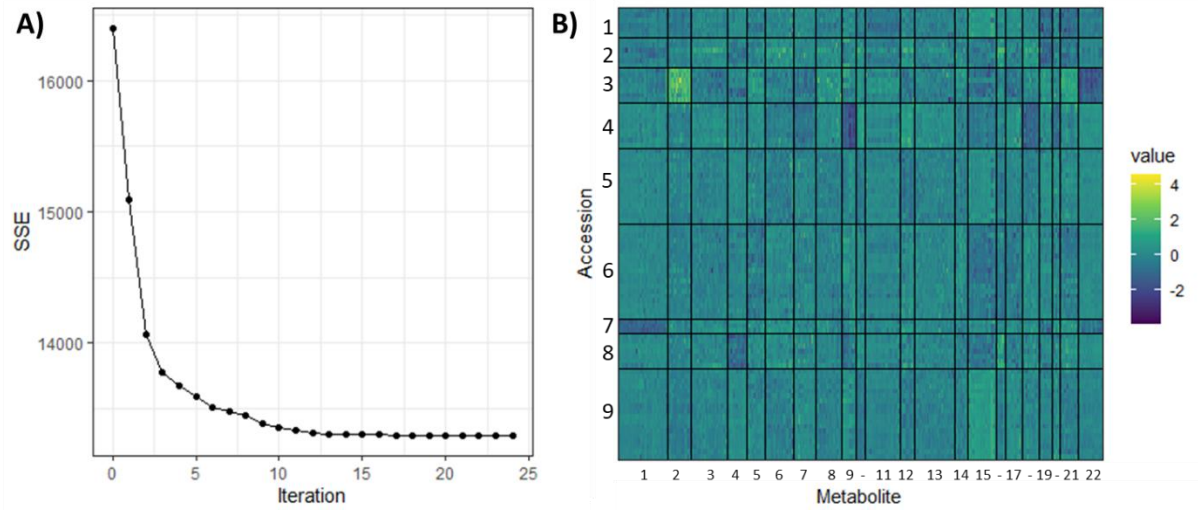



**Figure 4.20. A)** Total sum of squared errors (SSE) between metabolite means within a cell of the matrix and data values within cells measuring the within cell homogeneity, decreasing per bi-clustering analysis iteration. **B)** Bi-clustering analysis results plot for the optimized total within-cluster sum of squares, with the 90 diversity set accession assigned to nine clusters on the y-axis and 486 annotated metabolites divided over 22 clusters on the x-axis based on normalized metabolite data. Value is the normalized metabolite data and x-axis cluster numbers 10,16, 18 and 20 are not written due to space. Species of accessions per y-axis cluster are in **Table 4.8** and numbers of metabolites assigned per x-axis cluster in **Table 4.9**.

To see if certain metabolites cluster with certain species, the obtained optimal bi-clustering results are plotted. Nine accession clusters and 22 metabolite clusters could be identified (**Figure 4.20 B**). Some of the accession clusters were species specific, with information regarding which species are assigned to which cluster in **Table 4.8**. Nine *A. tricolor* accessions were assigned to a single cluster 4, and in cluster 8, four *A. tricolor* accessions cluster together with other accessions which do not belong to the 'hybridus' species complex. This suggests that especially *A. tricolor* accessions but also accessions of other species groups which are not closely related to the species complex, have a distinctive pattern in metabolite abundance. On the other hand, *A. cruentus* and *A. hypochondriacus* accessions are often grouping together (clusters 1, 5 and 9) as opposed to the other accessions (clusters 2, 3, 6 and 7) (**Table 4.8**). This suggests that the *A. cruentus* and *A. hypochondriacus* accessions in the diversity set, which are also closely related based on phylogenetic analysis, have similar metabolite profiles. To quantify these observations, a Fisher's exact test shows indeed that *A. tricolor*, *A. cruentus* and *A. hypochondriacus* accessions are significantly overrepresented in clusters 4, 5, 8 and 9 (**Table 4.8**, green

**Table 4.8.** Y-axis metabolite based accession cluster, number of accessions in that cluster (No acc. cluster tot.), species of those accessions, number of accessions of that species in that cluster (No acc. cluster spec.) and number of accessions of that species in total (No acc. species tot.) of bi-clustering analysis in **Figure 4.20 B**. Fisher exact test results indicating if species are over-/underrepresented in a cluster are provided with significant values ($p < 0.05$) marked green, just as the odds ratio of how much more (or less) a species is to occur in this cluster compared to others.

| Accession cluster | No acc. cluster tot. | Species | No acc. cluster spec. | No acc. species tot. | P-value | Odds ratio |
|---|---|---|---|---|---|---|
| 1 | 6 | A. cruentus | 3 | 14 | 0.07 | 5.28 |
| | | A. hypochondriacus | 3 | 21 | 0.34 | 2.90 |
| 2 | 6 | A. caudatus | 2 | 8 | 0.12 | 5.22 |
| | | A. dubius | 1 | 5 | 0.34 | 3.27 |
| | | A. tricolor | 1 | 14 | 1.00 | 0.89 |
| | | Group 1 - *A. hybridus* 2 | 2 | 4 | 0.03 | 15.71 |
| 3 | 7 | A. caudatus | 2 | 8 | 0.15 | 4.14 |
| | | A. dubius | 1 | 5 | 0.39 | 2.70 |
| | | A. hypochondriacus | 2 | 21 | 1.00 | 1.07 |
| | | A. quitensis | 1 | 2 | 0.17 | 10.72 |
| | | A. sp | 1 | 7 | 0.50 | 1.76 |
| 4 | 9 | A. tricolor | 9 | 14 | 0.00 | Inf |
| 5 | 15 | A. cruentus | 4 | 14 | 0.45 | 1.87 |
| | | A. hybridus | 1 | 7 | 1.00 | 0.67 |
| | | A. hypochondriacus | 9 | 21 | 0.00 | 6.06 |
| | | A. sp | 1 | 7 | 1.00 | 0.67 |
| 6 | 19 | A. caudatus | 4 | 8 | 0.10 | 3.53 |
| | | A. dubius | 1 | 5 | 1.00 | 0.75 |
| | | A. hybridus | 2 | 7 | 1.00 | 1.24 |
| | | A. quitensis | 1 | 2 | 0.44 | 3.11 |
| | | A. sp | 4 | 7 | 0.06 | 4.76 |
| | | Group 2 – *A. graecizans* | 4 | 4 | 0.00 | Inf |
| | | Group 3 - *A. dubius* | 3 | 4 | 0.04 | 10.27 |
| 7 | 3 | A. hypochondriacus | 1 | 21 | 1.00 | 1.34 |
| | | Group 1 - *A. hybridus* 2 | 1 | 4 | 0.15 | 10.88 |
| | | Group 3 - *A. dubius* | 1 | 4 | 0.15 | 10.88 |
| 8 | 7 | A. dubius | 2 | 5 | 0.06 | 8.44 |
| | | A. tricolor | 4 | 14 | 0.02 | 7.67 |
| | | Group 1 - *A. hybridus* 2 | 1 | 4 | 0.32 | 3.62 |
| 9 | 18 | A. cruentus | 7 | 14 | 0.02 | 4.61 |
| | | A. hybridus | 4 | 7 | 0.05 | 5.18 |
| | | A. hypochondriacus | 6 | 21 | 0.55 | 1.46 |
| | | A. sp | 1 | 7 | 1.00 | 0.52 |

cells). However, also Group 1 - *A. hybridus* 2, Group 2 – *A. graecizans* and Group 3 - *A. dubius* are significantly overrepresented in clusters 2 and 6, likely due to the low number of total accessions (n=4) of for these species groups (**Table 4.8**). Overall, the bi-clustering results (**Figure 4.20 B**, **Table 4.8**) further confirm correct species re-assignment based on metabolites complementing our phylogenetic analysis.

In order to investigate the distribution of metabolites over the clusters of bi-clustering analysis, the numbers of metabolites per cluster (x-axis, in **Figure 4.20 B**) are counted. These range 8 to 49 (**Table 4.9**), suggesting that the 486 annotated metabolites are relatively equally distributed over the 22 clusters. This eases interpretation with no extremely large or small metabolite clusters. It was the normalised metabolite values which were used for bi-clustering analysis, which was metabolite data normalised by sample medians, glog10 transformation and pareto scaling. Values in **Figure 4.20 B** are therefore not the absolute metabolite recordings but are an interpretable pattern of higher versus lower values corresponding to respectively more or less of the metabolite recorded for an accession.

To investigate the clusters with notably high or low abundance, the annotations of metabolites which are exclusively present in these clusters are revealed. Some metabolites which are exclusively present in x-axis cluster 2 and therefore extremely highly abundant in the accessions of y-axis cluster 3 are; 1-(3',4'-Dihydroxyphenyl)-3-(2'',4'',6''-trihydroxyphenyl)propan-2-ol, 2-hydroxy-2-(3'hydroxyphenyl) acetic acid, 3',4',5'-Trihydroxyphenylvaleric acid, 3-hydroxycitronelic acid, 7-Hydroxysecoisolariciresinol, cadinol, cinnamoyl glucose, dihydrozeatin, lotaustralin, phenylalanine-betaxanthin, a pyrimidone, a salicylic acid and taxifolin. Many of these compounds are biochemically connected in the phenylpropanoid pathway, which generates a large array of secondary metabolites responsible for plant defence, stress tolerance and other relevant developmental mechanisms, although some compounds belong to other specialized metabolic pathways. Metabolites which have very low abundance for these y-axis cluster 3 accessions are those in x-axis cluster 22, of which the exclusive compounds are 1-18:3-lysoPC, 2-18:2(9Z,12Z)-lysoPC, alpha-Chaconine, lysolecithin, a phosphatidylcholine and sucrose. Half of these are phospholipids and derivatives, used in cell membranes, signalling and lipid metabolism, which are therefore likely less present in the y-axis cluster 3 accessions. The nine *A. tricolor* accessions in y-axis cluster 4 appeared to have a low abundance of the following metabolites which were exclusively present in x-axis cluster 9; 4-p-Coumaroylquinic acid, kaempferol 3-O-glucosyl-rhamnosyl-glucoside and quinic acid. All these three compounds originate from the shikimate and phenylpropanoid pathways, which produce aromatic amino acids and many secondary metabolites essential for plant defense and adaptation, which therefore seems less evident among *A. tricolor* accessions. The only metabolite which was exclusively present in x-axis cluster 16 was an anisole which is therefore more abundant among the accessions of y-axis cluster 8. These are the four *A. tricolor* and other non 'hybridus' species complex accessions, which likely have more of this aromatic compound which might enhance flavour, plant signalling and defence.

**Table 4.9.** X-axis metabolite clusters (1-22) and their numbers of metabolites assigned to them and the total (TOT) number of metabolites in bi-clustering analysis of **Figure 4.20 B**.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | TOT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | 24 | 37 | 19 | 18 | 28 | 23 | 26 | 14 | 9 | 36 | 14 | 41 | 12 | 29 | 9 | 17 | 18 | 13 | 8 | 18 | 24 | 486 |

## mGWAS identify SNPs that are significantly associated with 81 metabolites.

To identify metabolite-genome associations within the ARC diversity set, metabolite-based genome-wide association study (mGWAS) was performed, using the 11,696,020 SNPs called against the 90 ARC genomes as genotypes (predictors) and the 1605 metabolites as traits. mGWAS is a powerful and commonly used technique in crops (Luo, 2015; Alseekh and Fernie, 2018), however, it has not yet been conducted in amaranth. GWAS can work with few individuals if it considers an obvious highly penetrant trait controlled by just one locus with no environmental influence, however, a subtle complex trait which is affected by many loci and the environment can fail to be detected by GWAS even if many samples are included. Typically, a minimum of at least 100 individuals is suggested for sufficient statistical power while performing GWAS (Visscher *et al.*, 2017). Therefore, the availability of only 90 accessions with a single metabolite measurement might reduce the power to detect possible associations. Nevertheless, performing mGWAS in the ARC diversity set is still useful, as there may be enough detection power if the difference in metabolite abundance is large. It can provide initial insights into the genomic regions or genes in amaranth that are involved in metabolites relevant to human nutrition.

**Table 4.10.** Number of significant SNPs per metabolite (top row) and number of metabolite traits (bottom row) which had this number of significant SNPs associated with them.

| 1 | 2 | 3 | 4 | 5 | 6 | 9 | 11 | 12 | 13 | 14 | 17 | 22 | 28 | 37 | 42 | 55 | 65 | 69 | 111 | 197 | 230 | 658 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 37 | 11 | 5 | 4 | 3 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

To perform mGWAS with the ARC diversity set, the average of the two replicate metabolites values after normalization was used together with SNPs called in the 90 genomes against *A. cruentus* (Ma *et al.*, 2021). These SNPs were filtered for quality by excluding those in the bottom first quantile of quality scores, with read depths between 283 and 4533, mean mapping quality > 30, minor allele count ≥ 1, and minor allele frequency ≥ 0.05. mGWAS was performed using a mixed model that accounts for unequal relatedness between strains using LDAK software v.5.2 (Zhang *et al.*, 2021), using trait permutation to generate an empirical P-value threshold with a false positive rate of 5% (see Methods: mGWAS).

Of the 1605 metabolite positive and negative peak combined dataset, 81 had at least one SNP significantly associated with that metabolite (**Supp table metabolites with SNP associations**). In most cases (37 of 81) only one SNP was significantly associated for the trait while for four metabolites more than 100 SNPs were significantly associated with the metabolite trait (**Table 4.10**). This shows that finding metabolite-genome associations in the ARC diversity set is possible for about one-twentieth of the metabolites, despite the set only containing 90 accessions. However, many of these associations rely on only one significant SNP, which suggests that these results should be interpreted cautiously, since single-marker effects could be sensitive to sampling or population structure. Since our analysis incorporates many reads of a diverse population of amaranth accessions, the single SNP associations are likely not very strong support of any real genome metabolite associations.

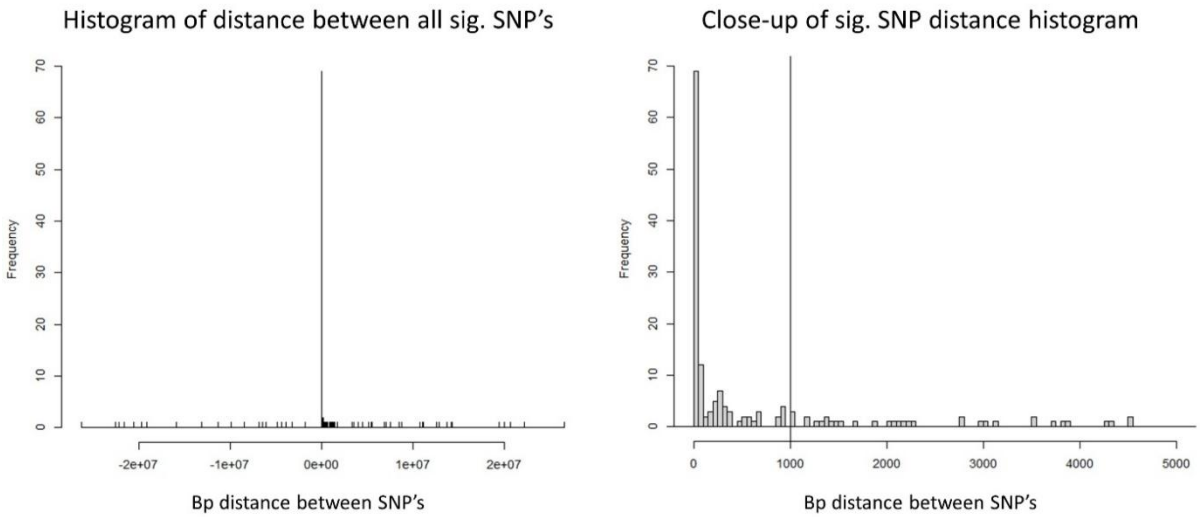



**Figure 4.21.** Histograms showing the distribution of base pair (bp) distance between significant SNPs for the same metabolite located on the same chromosome or scaffold. On the left side all bp distances are plotted while on the right it is only a close-up of the total plot (x-axis 0 to 5000 bp) displaying the picked cut-off of 1000 bp distance between SNPs for identifying SNPs located close enough to each other to potentially be involved in the same putative biosynthetic gene cluster and hence significant for different metabolite traits.

To investigate if some of the detected metabolites are potentially involved in the same putative biosynthetic gene clusters (BGC), identical or closely located SNPs were identified among all SNPs which were found associated with different metabolites. In total 1738 SNPs were significantly associated with 81 metabolite traits, and none of these SNPs were identical, i.e. associated with more than one metabolite. To identify closely located SNPs, a window was picked based on the distribution of the distance between significant SNPs for a metabolite but located on the same chromosome or scaffold. A close up of a histogram with this distribution shows the curve of high frequencies with bp distances between SNPs flattens off when the distance is bigger than 1000 bp (**Figure 4.21**, right). This means that using a bigger window would not result in drastically more closely located SNPs, and since 1000 bp is sufficient length for BGC, this window threshold was used for detection.

Of the 1738 SNPs with significant metabolite associations, 14 were within a 1000 bp distance located on the same chromosome with a SNP that was significant for a different metabolite trait (**Table 4.11**). These 14 SNPs located closely to each other might suggest that the different metabolites they are significant for are different compounds but potentially involved in the same biosynthetic pathways derived by a BGC. However, the lack of identical SNPs across the metabolites makes this conclusion less certain. If BGC were present and were influencing multiple metabolites, it is expected to increase the frequency of overlapping SNPs, i.e. SNPs that affected one enzyme in the pathway would impact all the metabolites downstream of that. Nonetheless, it is also possible that different SNPs within the same region tag the same underlying genetic variation due to linkage disequilibrium, and hence further investigation in the closely located SNPs is relevant.

To see if the closely located SNPs that are significantly associated with different metabolites could potentially be involved in the same BGC, their locations within genes and gene functional annotations were investigated, as was the annotation of the metabolites the SNPs were significantly associated with. For 7 of the 14 closely located SNP pairs, both SNPs were located in the same gene for which the gene ID is provided in **Table 4.11**. If SNPs are

located in the same gene, it increases the chance that the metabolites related to the SNPs are in the same biosynthetic pathway. Checking the functional annotations of these 7 genes reveals the following: Amacr05g11100 as coil, Amacr08g06810 as AAA ATPase, Amacr04g14190 as Histone-lysine N-methyltransferase, Amacr10g09050 as beta-sandwich domain of Sec23/24 and Amacr16g08200 as proline-rich receptor-like protein kinase PERK. Amacr13g00942 was not annotated and was generated only as a gene by manual curation. The five genes which could be annotated are likely not directly responsible for catalysing biosynthetic reactions that produce metabolites, however, all could influence metabolite levels indirectly through regulation and signalling. This suggests that these genes could be part of a BGC responsible for, or alternatively may regulate, pathways underlying the biosynthesis of the detected metabolites.

Looking at the annotations of metabolites associated with the SNP pairs, only SNPs closely located on chromosome 6 are both significant for different Fatty Acyl metabolites (**Table 4.11**). This suggest that these two metabolites might be related or could simply be precursors or intermediate compounds to each other in a metabolic pathway. However, a large proportion (200/486; **Table 4.7**) of the total annotated metabolites belong to the Fatty Acyls class, and these two metabolites were the only ones in **Table 4.11** annotated by GNPS rather than with the more reliable SIRIUS4. Additionally, M355.0664T113 also could be annotated as caffeic acid-3'-O-glucuronide or similar compounds by PhytoHub. This suggests that the small hint that these SNPs might be part of the same BGC is likely untrue, since there is no strong evidence that the metabolites they are associated with are related.

To obtain insight into how much of the phenotypic variation of the metabolite traits is explained by additive genetic effects, heritabilities and their standard deviations (SD) were calculated for each metabolite. Here, a high heritability means that much of the observed metabolite variation in the diversity set is attributable to the additive effects of the common SNPs used while performing GWAS. Plotting the distributions of estimated heritability per metabolite trait and their SD, for all 1605 metabolites and the 81 metabolites with a significantly associated SNP, showed a peak of heritability often being just a bit bigger than 0.2 or often having a SD of 0 (**Figure 4.22**). This means that for many of the metabolite traits the additive genetic effects contributing to the phenotypic variation is low, and that these values of heritability could mostly be estimated accurately, i.e. low SD.

The heritability data is not normally distributed but plotting it reveals that some metabolites have a high heritability of very close to 1 while SD are never bigger than 0.2 (**Figure 4.22**). This suggests that there are metabolites which have a high attribution of additive effects of SNPs, and that none of the estimated heritabilities is predicted with high uncertainty. Metabolites with a lower estimated heritability have a lower SD, which increases with an increasing estimated heritability, but gradually decreases again after reaching an estimated heritability of 0.5 or higher (**Figure 4.22**, bottom). This shows that either very low or high heritabilities of metabolite traits are estimated more accurately compared to heritabilities of around 0.5, which indicates that the latter should be interpreted more carefully. These observations could be explained by the fact that, among the 81 metabolite traits, those with either high or low estimated heritabilities tend to have a greater number of significant SNPs (**Figure 4.23**, top). This could result in a stronger genetic signal containing more evidence and a smaller standard error, causing a lower SD of estimated heritabilities close to either 0 or 1. No correlation was found when plotting the heritability against the mean of a single

metabolite trait over the different accessions, or their SD. Hence those plots are not displayed, and this reveals that overall metabolite abundance is not causal or influential while estimating heritabilities.

**Table 4.11.** Chromosome (Chr), metabolite ID's, base pair locations (bp loc.), significance by Wald p-value (Sig.), metabolite annotations (Anno.) and bp distance (Dist.) between two SNPs significantly associated with a different metabolite trait but located within a 1000 bp distance on the same chromosome. These two sets of 14 SNPs are ordered by bd distance (increasing top to bottom) in therefore decreasing relevance of likeliness of involvement in the same biosynthetic pathway. For metabolite M225.1132T225 many different annotations could be given by PhytoHub hence no further name is given. The 'Gene' column indicates gene ID's (excluding AmacrXXg, with XX as chromosome number) of *A. cruentus* reference genes both sig. SNPs in that row were located in.

| Chr | Metabolite 1 | bp loc. 1 | Sig. 1 | Anno. 1 | Metabolite 2 | bp loc. 2 | Sig. 2 | Anno. 2 | Dist. | Gene |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | M245.1394T202 | 6848605 | 2.28E-13 | - | M671.4042T347 | 6848590 | 3.01E-12 | - | 15 | - |
| 6 | M271.2278T431 | 5724868 | 1.40E-13 | Fatty Acyl | M355.0664T113 | 5724815 | 3.74E-14 | Fatty Acyl | 53 | - |
| 5 | M188.8947T37 | 17808658 | 0 | - | M212.8847T37 | 17808353 | 2.6E-16 | - | 305 | 11100 |
| 13 | M671.4042T347 | 19223048 | 5.31E-14 | - | M245.1394T202 | 19222700 | 2.28E-13 | Menthane | 348 | - |
| 8 | M225.1132T225 | 12882841 | 0 | Many | M212.8847T37 | 12882489 | 2.18E-12 | - | 352 | 06810 |
| 4 | M225.1132T225 | 21325817 | 0 | Many | M671.4042T347 | 21325436 | 2.18E-12 | - | 381 | 14190 |
| 8 | M212.8847T37 | 12882435 | 2.18E-12 | - | M225.1132T225 | 12881998 | 2.25E-12 | Many | 437 | 06810 |
| 1 | M188.8947T37 | 9944203 | 4.29E-14 | - | M294.9628T38 | 9943748 | 3.8E-14 | - | 455 | - |
| 10 | M294.9628T38 | 15283855 | 2.18E-12 | - | M225.1132T225 | 15283357 | 0 | Many | 498 | 09050 |
| 16 | M188.8947T37 | 14050883 | 1.47E-12 | - | M813.4829T496 | 14050310 | 0 | - | 573 | 08200 |
| 16 | M241.8989T37 | 2228565 | 4.88E-15 | - | M188.8947T37 | 2227782 | 2E-17 | - | 783 | - |
| 14 | M671.4042T347 | 12712472 | 0 | - | M240.9248T38 | 12711623 | 2.28E-13 | - | 849 | - |
| 10 | M212.8847T37 | 14345388 | 2.46E-13 | - | M331.1288T174 | 14344460 | 3.3E-16 | Tyramine-betaxanthin | 928 | - |
| 13 | M225.1132T225 | 892606 | 0 | Many | M188.8947T37 | 891628 | 2.63E-12 | - | 978 | 00942 |

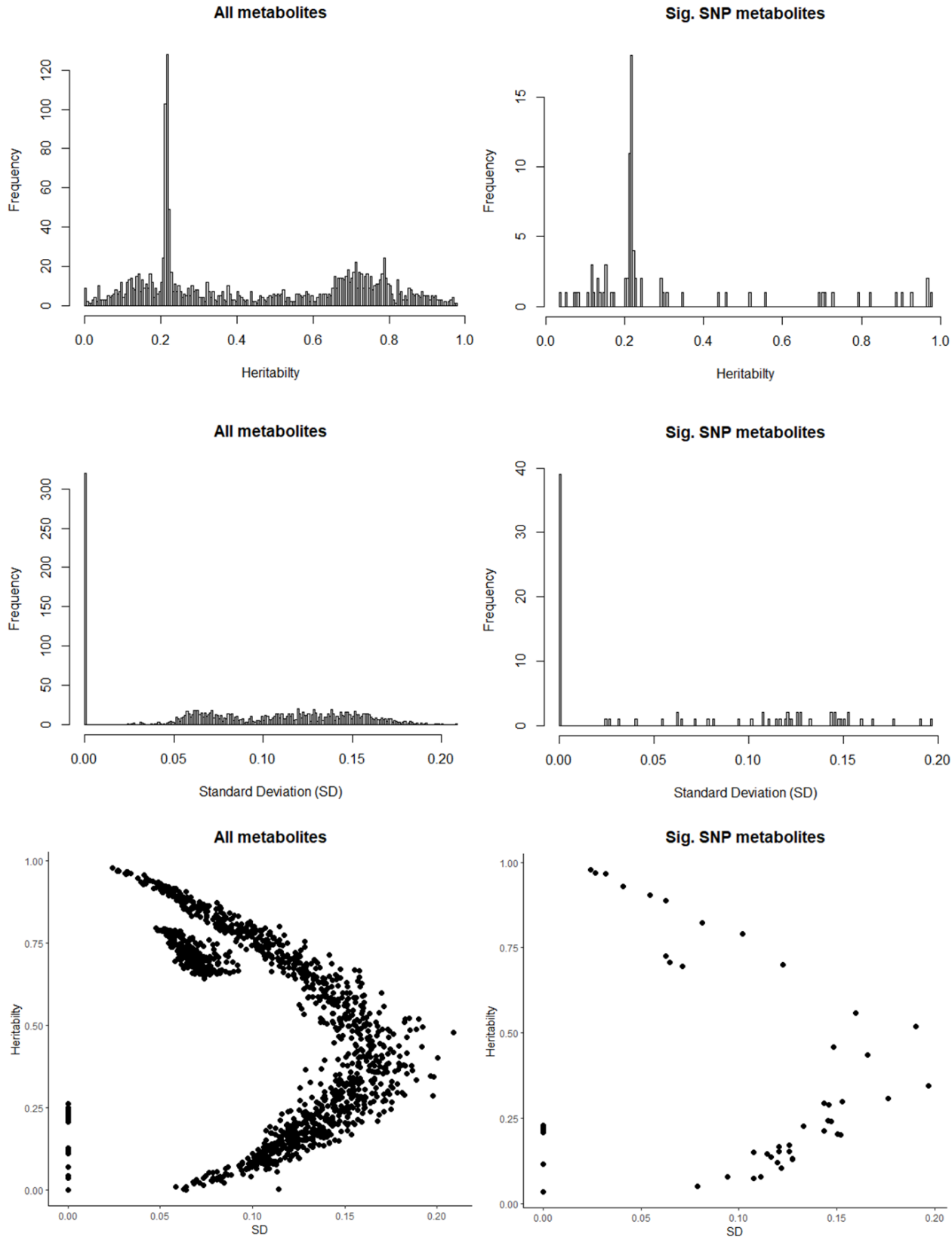



**Figure 4.22.** Histograms showing the distributions of heritability (top) and their standard deviation (SD) (middle), as well as a scatter plot showing their correlation (bottom), visualised for all 1605 metabolite traits (left) or 81 for which a SNP was found to be significantly associated with that metabolite (right) by performing mGWAS using 90 diversity set accessions.

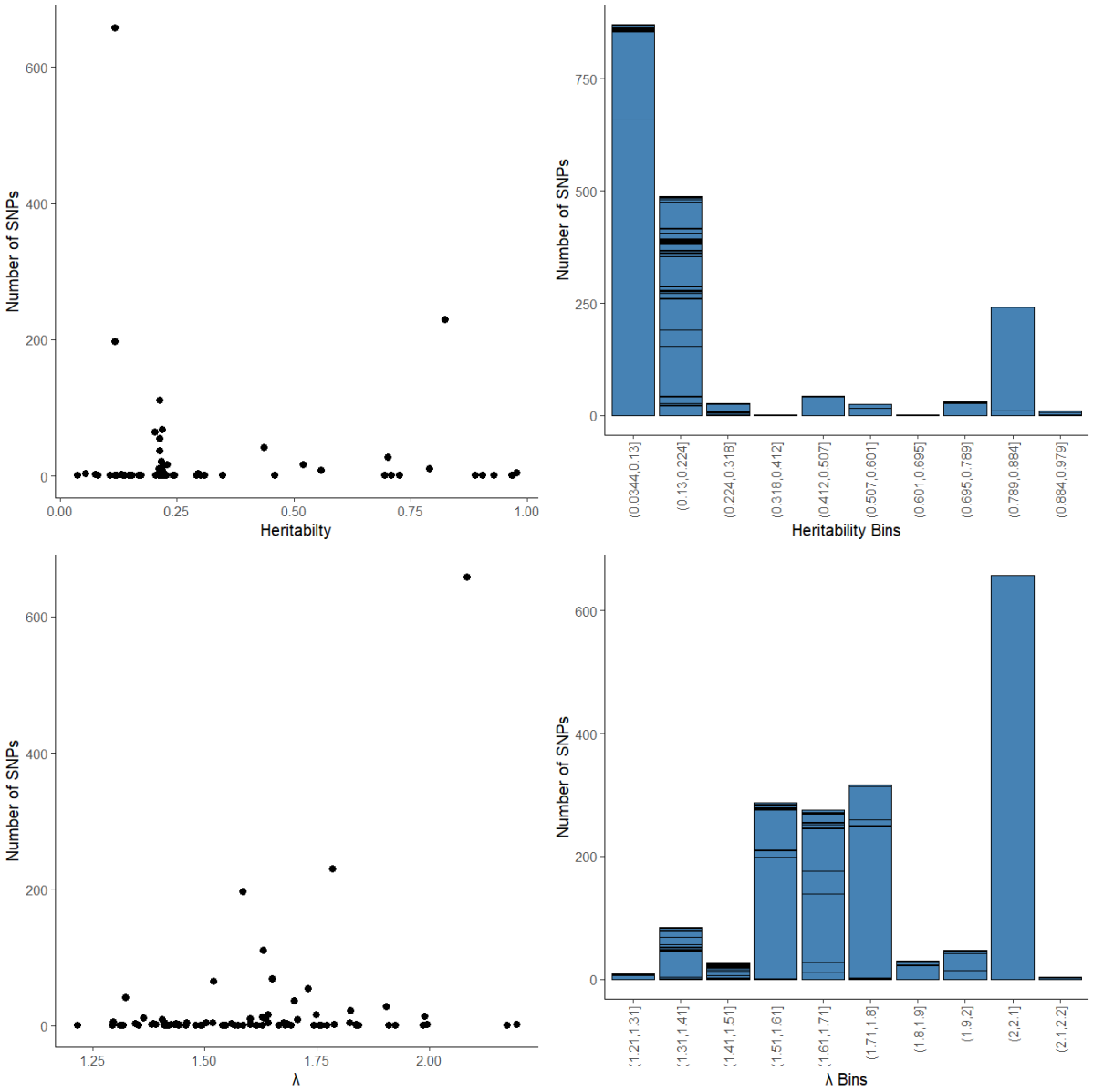



**Figure 4.23.** Scatterplots (left) and binned histograms (right) showing the increased number of significant SNPs per 81 metabolite traits associated with either low or high estimated heritabilities (top) or with a high estimated genomic inflation factor (λ) (bottom).

To investigate the 81 metabolites with SNP associations and select some of interest, their annotations were addressed together with the significance of SNPs, heritabilities, and overall data quality. Of the 81 metabolites with genome associations according to mGWAS, 26 could be annotated (**Table 4.12**, **Supp table annotated metabolites with SNP associations**). Quantile-quantile (Q-Q) plots are used to identify the data as of good quality to examine it in more detail. Q-Q plots demonstrate the distribution of all SNPs against a standard normal distribution, with more SNPs on the diagonal in Q-Q plots meaning that expectations are met and data is of sufficient quality. To quantify the Q-Q plot p-value results, the genomic inflation factor (λ), i.e. λ = median(observed test statistics) / median(expected test statistics), was calculated for each metabolite trait. λ close to 1 indicates a distribution of test statistics consistent with a null hypothesis of no association expected. λ is significantly greater than 1 (e.g., > 1.05) or less than 1 suggests inflation or

deflation in test statistics, respectively. λ was 1.60 on average over all the 81 traits, with a SD of 0.21, maximum of 2.19 and minimum of 1.22. With λ > 1.05 the null hypothesis of no association can be rejected. Hence λ results show that associations are found for all 81 metabolites with a significant SNP. This is anticipated as SNPs which are significantly associated with the metabolite trait do not meet the expected test statistics and therefore are not on the diagonal in Q-Q plots which causes λ to increase. This concept was further confirmed when plotting numbers of significant SNPs against λ per metabolite trait. λ typically increases as the number of significant SNPs rises (**Figure 4.23**), indicating that their increased presence contributes to λ inflation in some cases.

Of the 26 annotated metabolites with SNP associations (**Table 4.12**), eight were selected as of interest due to a low p-value (≤2.35E-13) for their most significant SNP, a considerable high heritability (>0.21) with not too extreme SD (<0.15), while Q-Q plots and λ identify the data as of good quality to examine it in more detail (**Figure 4.24**). Overall, the estimated λ for the eight selected metabolites with provided Manhattan plots was low, with six of them being below average (λ < 1.60), one exactly the mean λ value and one just slightly higher than the average (**Figure 4.24**). This indicates that despite SNPs being associated with high significance, their overall distribution is closer towards a standard normal distribution compared to SNP distributions of other metabolite traits, indicating good quality data. There are two, not mutually exclusive, explanations for the abundance of SNPs above the x=y Q-Q plots diagonal determining λ. Either genomic inflation is caused by the mixed model not correcting for population structure (relatedness) well enough, or the investigated trait is influenced by many SNPs with several potentially in regions of high LD. From the Manhattan plots high LD could not clearly be observed (**Figure 4.24**). Hence, no correction for population structure as well as many SNPs influencing the trait could both play a role for current λ results, with curves in the Q-Q plots sharply upward at the end typically indicating that some SNPs that are associated.

The eight metabolites selected as of interest (**Figure 4.24**), are biochemically quite diverse and do not belong to the same metabolite classes. However, they are all components of plant specialized metabolism, involved in stress response, defense, or signaling. High heritabilities (>0.69) were estimated for the metabolites annotated as Caffeic acid-3′-O-glucuronide (M355.0664T113 and M379.0636T127), and an extremely high heritability of 0.98 was estimated for M387.3112T361, annotated as Hydroxyphytanic acid (alpha-) (**Table 4.12**). These high heritabilities make these metabolites particularly interesting, as their abundance appears to be strongly influenced by genetic variation, rather than environmental factors. This makes these metabolites genetically trackable, with a large proportion of the metabolite variation attributable to genetic effects captured by SNPs, suggesting higher potential for predicting these metabolite levels from genomic data, and for breeding plants that produce more (or less) of these metabolites. The very heritable Hydroxyphytanic acid (alpha-) metabolite is a fatty acid derivative which is not very commonly detected in plants and hence their role in crops is not well understood. Caffeic acid-3′-O-glucuronide is a phenolic acid conjugate known for its antioxidant properties. Hence this metabolite is beneficial human nutrition and health but also for plant defence and resilience. The large SD of the estimated heritability of lineolic acid metabolite M277.2170T427 marks it as the least of interest of all eight metabolites with provided Manhattan plots (**Table 4.12**).

Overall, it is shown that one-twentieth (81/1605) of the metabolites are associated with genomics, and about one third (26/81) of those are annotated with some being relevant to human nutrition. Therefore, can be concluded that this mGWAS analysis was successful, providing a novel contribution to science revealing metabolite-genome associations in amaranths.

**Table 4.12.** 26 Metabolite ID's (Masstag ID), as a subset of 81 metabolites for which a significant SNP could be found. The metabolites are ordered with highest significance top to bottom and provided with their annotated name by SIRIUS4 (*most reliable), PhytoHub, KNApSAcK or ChemSpider databases, subclass and class according to SIRIUS4 or GNPS, lowest Wald-P value of the most significant SNP found for this metabolite according to mGWAS, chromosome this most significant SNP is located at (Chr.), estimated heritability (Her.) plus standard deviation (SD) and mean normalised metabolite data value over 90 accessions plus it's SD. Masstag ID's in bold are selected for having low p-values, considerate high heritabilities with non-extreme SD's while having good quality data according to Q-Q plots and therefore Manhattan plots are visualised in **Figure 4.24**.

| Masstag ID | Name *SIRIUS4 annotation | Subclass | Class | Wald P | Chr. | Her. (SD) | Mean (SD) |
|---|---|---|---|---|---|---|---|
| M209.0456T156 | 2-(4-Hydroxymethyl-3-hydroxyphenyl)propan-1-ol | - | - | 0 | 4 | 0.52 (0.19) | -2.72E-16(0.81) |
| M245.1394T202 | 3-hydroxycitronelic acid | - | - | 0 | 12 | 0.22 (0) | -5.93E-17(0.75) |
| M253.0830T97 | Ethanolamine-betaxanthin | - | - | 0 | 7 | 0.44 (0.17) | -1.05E-16(0.62) |
| M267.1965T426 | Cadinol | Lineolic acids and derivatives | Fatty Acyls | 0 | 8 | 0.22 (0) | 4.10E-16(0.69) |
| M331.1288T174 | Tyramine-betaxanthin | - | - | 0 | 2B | 0.82 (0.08) | -7.28E-17(0.59) |
| M455.3165T324 | delta-Tocotrienol | - | - | 0 | 1 | 0.13 (0.13) | -9.44E-17(0.52) |
| **M277.2170T427** | - | Lineolic acids and derivatives | Fatty Acyls | 1.16E-15 | 14 | 0.21 (0.14) | 1.46E-16(0.35) |
| **M261.1343T172** | 3,8-dihydroxy-3,7-dimethyl-6-octenoic acid | - | - | 2.30E-15 | 10 | 0.21 (0) | -1.50E-16(0.31) |
| **M357.1086T475** | Phenylalanine-betaxanthin | - | - | 4.88E-15 | 6 | 0.22 (0) | 1.51E-16(0.74) |
| M279.2318T365 | alpha-LA* | Lineolic acids and derivatives | Fatty Acyls | 5.28E-15 | 3 | 0.70 (0.12) | -4.78E-17(0.52) |
| **M387.3112T361** | Hydroxyphytanic acid (alpha-) | - | - | 6.44E-15 | 1 | 0.98 (0.02) | 3.76E-17(0.65) |
| **M691.3549T261** | - | Lineolic acids and derivatives | Fatty Acyls | 1.81E-14 | 3 | 0.22 (0) | 4.44E-17(0.69) |
| M271.2278T431 | - | Lineolic acids and derivatives | Fatty Acyls | 2.05E-14 | 15 | 0.15 (0.11) | 5.24E-17(0.41) |
| M518.3240T326 | 1-18:3-lysoPC* | 1-acyl-sn-glycero-3-phosphocholines | Glycero-phospholipids | 5.63E-14 | 6 | 0.15 (0.12) | -3.42E-17(0.61) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M279.2327T445 | - | Lineolic acids and derivatives | Fatty Acyls | 5.95E-14 | 16 | 0.20 (0.15) | -2.34E-16(0.40) |
| **M341.1082T40** | Sucrose | Lineolic acids and derivatives | Fatty Acyls | 7.00E-14 | 12 | 0.22 (0) | -1.11E-16(0.74) |
| **M355.0664T113** | Caffeic acid-3'-O-glucuronide | Lineolic acids and derivatives | Fatty Acyls | 1.40E-13 | 6 | 0.71 (0.06) | 7.40E-17(0.59) |
| **M379.0636T127** | Caffeic acid-3'-O-glucuronide | - | - | 2.35E-13 | 10 | 0.69 (0.07) | -1.48E-17(0.71) |
| M231.1237T180 | 3-hydroxycitronelic acid | - | - | 3.71E-13 | 6 | 0.21 (0) | 3.47E-16(0.73) |
| M581.1502T137 | Cyanidin 3-O-sambubioside | - | - | 6.71E-13 | 14 | 0.93 (0.04) | -4.44E-17(0.39) |
| M369.0821T137 | Isoferulic acid-3'-O-glucuronide | Lineolic acids and derivatives | Fatty Acyls | 1.12E-12 | 8 | 0.73 (0.06) | 1.18E-16(0.54) |
| M573.3282T346 | - | Lineolic acids and derivatives | Fatty Acyls | 1.29E-12 | 4 | 0.14 (0.12) | -1.75E-16(0.57) |
| M293.2121T331 | - | Lineolic acids and derivatives | Fatty Acyls | 1.47E-12 | 4 | 0.08 (0.11) | -6.35E-17(0.45) |
| M225.1132T225 | 3,7-Dimethyl-2,7-decadien-1,10-dioic acid | - | - | 2.18E-12 | 15 | 0.22 (0) | -1.28E-16(0.68) |
| M937.5857T483 | - | Triterpene saponins | Prenol lipids | 2.24E-12 | 2B | 0.21 (0) | 9.56E-18(0.44) |
| M843.5294T457 | - | Lineolic acids and derivatives | Fatty Acyls | 2.95E-12 | 5 | 0.11 (0.12) | 1.62E-16(0.70) |

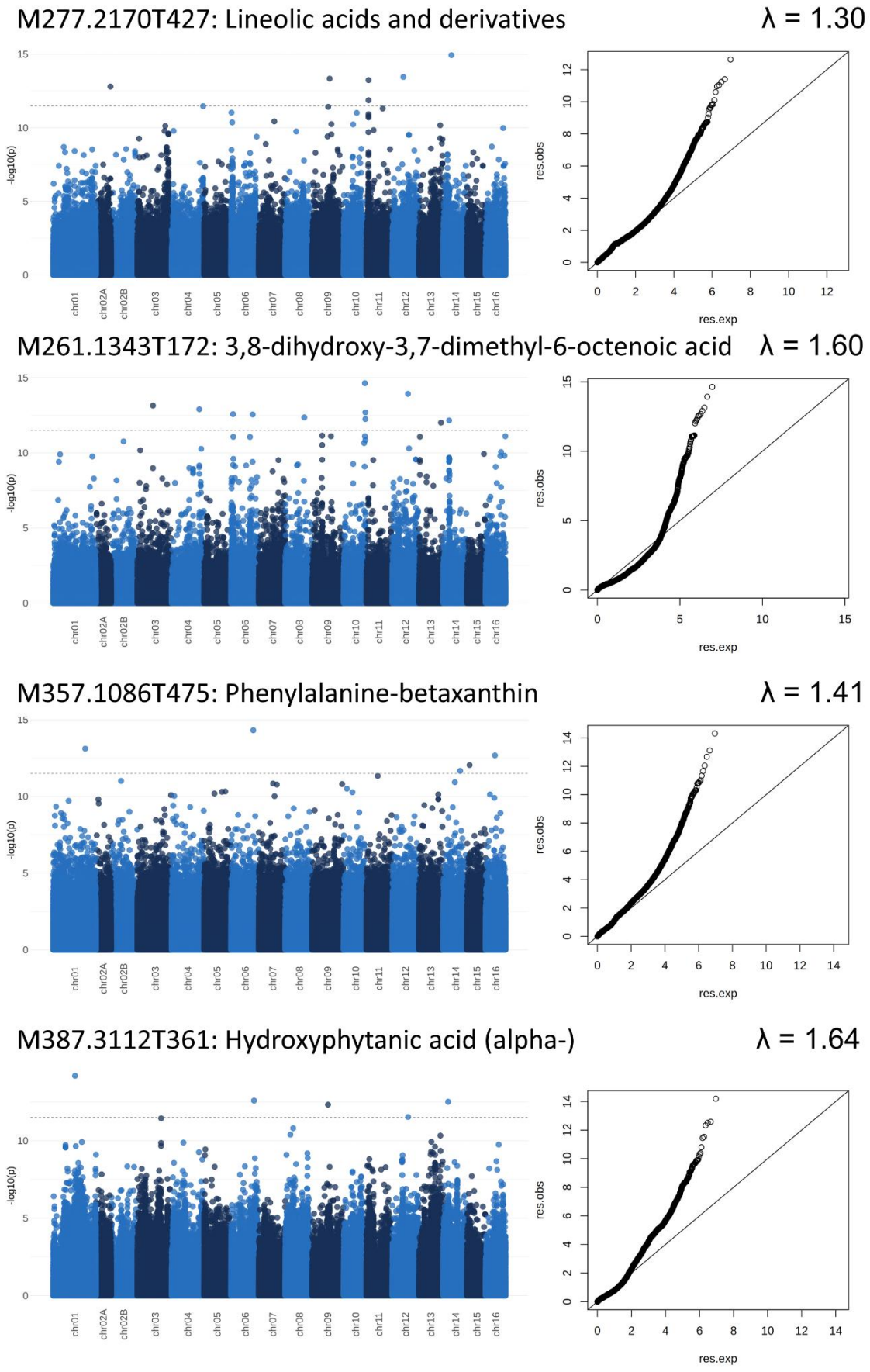



**Figure 4.24. A)** Manhattan and Q-Q plots of four metabolite traits which could be annotated (bold Masstag ID's and more information in **Table 4.12**) with considerate low p-values, high heritabilities and non-extreme SD's while having good quality data according to the Q-Q plots. λ estimates with p-values used to plot the Q-Q plots is given.

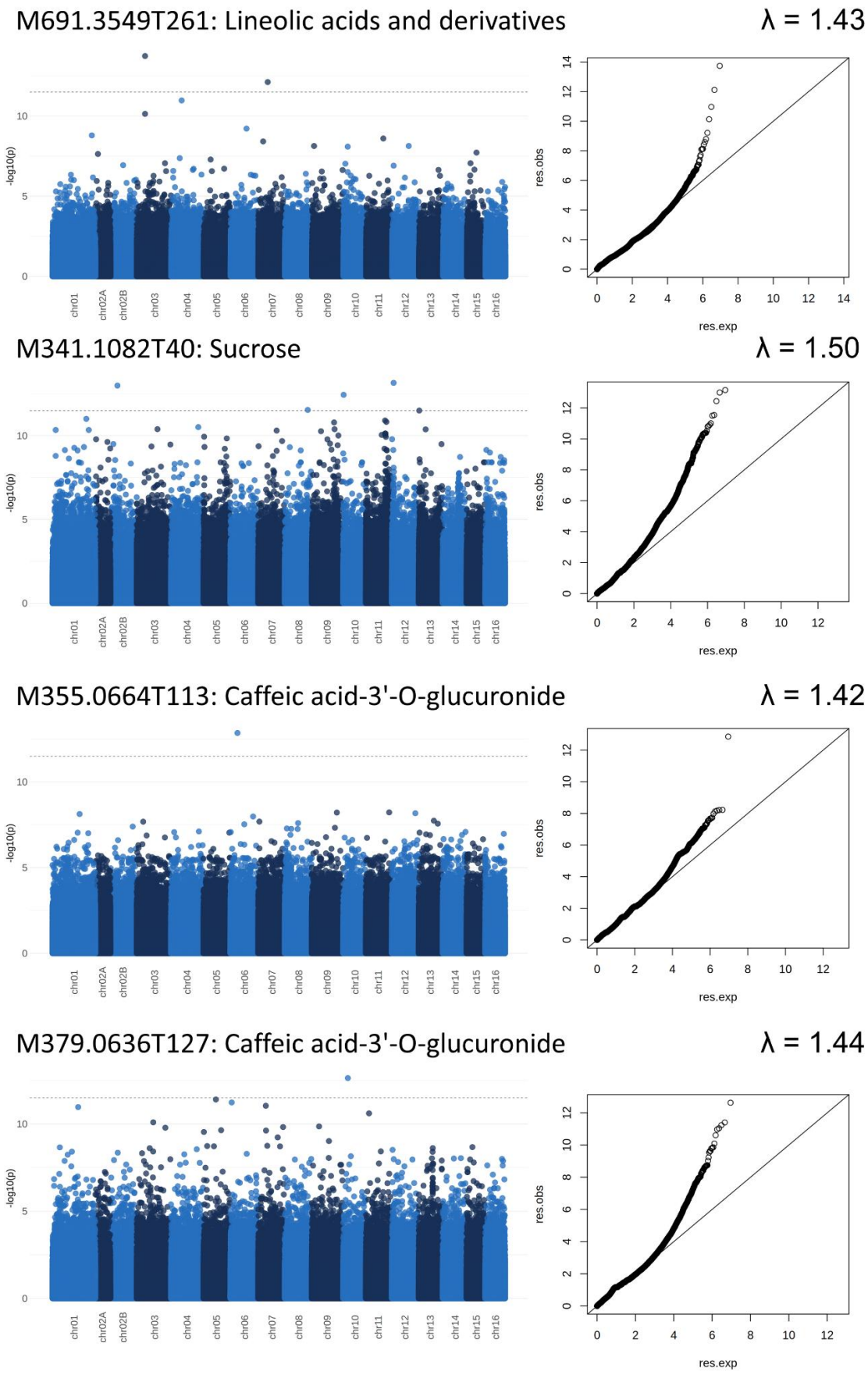



**Figure 4.24. B)** Manhattan and Q-Q plots of four metabolite traits which could be annotated (bold Masstag ID's and more information in **Table 4.12**) with considerate low p-values, high heritabilities and non-extreme SD's while having good quality data according to the Q-Q plots. λ estimates with p-values used to plot the Q-Q plots is given.

# Discussion

## Successful species assignment and questioning of *A. hybridus* existence

The current study presents a successful analysis in which 17 of the 22 unknown (*A. sp.*) species in an ARC diversity set consisting of 90 genome sequences could be assigned (**Table 4.5**). This was confirmed by complementary evidence using metabolomic as well as genomic data for this ARC set and an external set of 116 sufficient quality sequences (Stetter, Vidal-Villarejo and Schmid, 2020). An initial attempt to obtain species allocation of all available accessions did show that information acquired from different sources did not always match (**Supp table diversity set**). This highlights the complexity of the datasets and the known phenomenon of difficulty in clear taxonomy of the *Amaranthus* genus (Sauer, 1957). We show that previous classifications mis-identified many accessions, which is expected for those accessions of which no genome-wide data was available. Without their genomic information, annotations were likely based on insufficient power. Our results confirm this assumption since only three of the 116 accessions for which sequences were imported from Stetter *et al.* (2020) were re-assigned (**Table 4.5**), and accessions from our ARC set which were assessed in the previous genomic study of Lin *et al.* (2022) did mostly match our current annotations (**Table 4.6**). Therefore, those accessions for which no genomic data previously was available (most accessions of the 90 accessions ARC set) are the highlight in this current re-classification study, in which the other pre-sequenced accessions were a relevant guideline in the annotation process.

Our combined genomic dataset consisted of species allocated to two of the three commonly recognized subgenera according to recent very complete phylogenetic studies of the *Amaranthus* genus (Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). The two species (*A. tricolor* and *A. graecizans*) in our set belonging to the *Amaranthus Albersia* subgenus indeed showed a clear separation from other accessions in our analysis (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**, **Figure 4.19**). This separation was less distinct for *A. graecizans* compared to *A. tricolor* (especially in the metabolite data; **Figure 4.18**), likely due to having less accessions as a starting point (N=3). Hence our choice to re-classify it as 'Group 2 – *A. graecizans*' due to high uncertainty regarding species name. Note that one of the *A. graecizans* accessions has been re-classified into an *A. tricolor* accession highlighting difficulty in morphological identification (**Table 4.5**). All other available species allocations prior to reclassification belong to the *Amaranthus Amaranthus* subgenus (*A. hypochondriacus*, *A. cruentus, A. caudatus, A. hybridus, A. quitensis, A. dubius, A. powellii*) (**Table 4.1**) (Stetter and Schmid, 2017; Waselkov, Boleda and Olsen, 2018). This includes the five species which are known to be part of the 'hybridus' species complex (three grain amaranths; *A. caudatus*, *A. cruentus*, *A. hypochondriacus*, and two wild relatives; *A. hybridus* and *A. quitensis*) (Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017). Most of the accessions allocated as 'hybridus' complex species clustered together in our analysis and the single accession allocated as *A. powellii* separated clearly on its own (**Figure 4.7**, **Figure 4.10 A**). This matches the assumption that *A. powellii* should not be included in the 'hybridus' species complex like it historically was (Xu and Sun, 2001; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Park *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017).

Four of the five 'hybridus' complex species separated semi-distinctively in our phylogenetic analysis (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**). *A. cruentus* and *A. hypochondriacus* accessions clustered closely to each other just as *A. caudatus* and *A. quitensis* accessions did. This phenomenon is commonly observed and usually explained by the geographic origin of each species, i.e., *A. cruentus* and *A. hypochondriacus* originating from North or Central America being Mesoamerican species and *A. caudatus* and *A. quitensis* originating from South America being Andean species (Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). Inclusion of the 90 accession ARC diversity set in the phylogenetic analysis of Stetter *et al.* (2020) did change that demographic background per species since it included *A. cruentus* and *A. hypochondriacus* accessions originating from Tanzania, Peru, Mexico, South Africa, India, Uganda, the US, Vietnam, Nepal and Zimbabwe, but also *A. caudatus* and *A. quitensis* accessions originating from Pakistan, Bolivia, the US, Nepal, India and Uganda (**Supp table diversity set**). After species re-classification, a pattern of more *A. cruentus* and *A. hypochondriacus* in Central versus more *A. caudatus* and *A. quitensis* accessions in South America is still evident (**Figure 4.13**, **Figure 4.15**). However, with three *A. hypochondriacus* and four *A. cruentus* accessions originating from South America, one *A. caudatus* accession originating from the US, and many accessions of the four less distinctively clustering 'hybridus' complex species not originating from the America's, geographic origin does not seem to be the only explanation for clustering of *A. cruentus* with *A. hypochondriacus* and *A. caudatus* with *A. quitensis* accessions. What should not be forgotten is that the stated origin of our accessions might not be the natural origin, since accessions are often moved between countries or regions causing bias in the analysis. Even if the available origin information is correct, the question remains when did the accessions go there, since likely all amaranth at some point originated from Central and South America.

Correct species allocation as well as relatedness likely both play a role for current clustering results. A phylogenetic tree based on thousands of genome-wide markers derived by genotyping by sequencing (GBS) does indeed reveal a higher relatedness of *A. cruentus* with *A. hypochondriacus* and *A. caudatus* with *A. quitensis*, with a strongly supported split between the two groups (Stetter and Schmid, 2017; Figure 5, 4). This pattern is not always observed in other published phylogenetic trees, but this may be due to different data types and inclusion of many very distant species masking the relationships among the closely related 'hybridus' complex species (Park *et al.*, 2014; Berger *et al.*, 2016; Waselkov, Boleda and Olsen, 2018). In a neighbour joining phylogenetic tree based on these same genome-wide markers there is no strong separation of the 'hybridus' complex species into different clusters which confirms conflicting phylogenetic signals among accessions that may reflect gene flow or hybridization (Stetter and Schmid, 2017; Figure 3). Essentially, widespread misidentification of accessions and species complicates the classification process in the current but also previous studies.

Strong support of a close relationship between *A. quitensis* and *A. caudatus* has previously been found (Chan and Sun, 1997; Xu and Sun, 2001) as well as evidence for the hypothesized role that *A. quitensis* played in the evolution of *A. caudatus* through introgressive hybridization (Park *et al.*, 2014). *A. quitensis* and *A. caudatus* accessions group together based on morphology when assessing pistillate and staminate floral characteristics in the 'hybridus' complex species (Adhikary and Pratt, 2015). The support for the split between *A. caudatus* and *A. quitensis* was low in the tree based on GBS markers

(Stetter and Schmid, 2017; Figure 4). This reflects the recent gene flow between the two species and identification of *A. caudatus* being an incompletely domesticated grain crop (Stetter, Müller and Schmid, 2017). Comparing our phylogenetic with our metabolomic results we even choose to not separate between *A. quitensis* and *A. caudatus* since many accessions of both species were assigned to the same clade likely due to high relatedness or misidentification (**Figure 4.19**). Our studies do not comply with previous findings of *A. caudatus* and *A. hypochondriacus* seemingly being closely related (Kietlinski *et al.*, 2014), but do comply with recent phylogeny confirming relations between *A. cruentus* and *A. hypochondriacus* and both being distant from *A. caudatus* and *A. quitensis* (Wu and Blair, 2017; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023).

The most commonly mentioned reason for clustering of *A. cruentus* with *A. hypochondriacus* in other phylogenetic analysis is their Central American origin (Stetter and Schmid, 2017; Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023). Our consistent separation but also close clustering of these species (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**, **Figure 4.19**), despite different demographic backgrounds (**Figure 4.13**, **Figure 4.15**), suggests high relatedness could additionally be caused due to different factors. Domestication of the species could play a role in relatedness and separation between subpopulations. Of the three 'hybridus' species complex cultivated grain amaranths, only *A. caudatus* is identified as incompletely domesticated with key domestication traits such as the shape of inflorescences, seed shattering and seed size similar to the wild relatives *A. hybridus* and *A. quitensis* (Stetter, Müller and Schmid, 2017). If *A. caudatus* is indeed less domesticated than *A. cruentus* and *A. hypochondriacus*, we would expect its clade to be closer to the root of the phylogeny (**Figure 4.10**) and higher values of π (**Figure 4.11**), which was both not observed. What could be observed is that *A. caudatus* clustered very closely with the subpopulation of wild *A. quitensis* (**Figure 4.7**) being present in the same phylogenetic clade (**Figure 4.10**). This reveals the close relationship of *A. caudatus* as a less domesticated grain amaranth with similar phenotypic features to wild *A. quitensis*. Wild *Amaranthus* species are known to have seed dormancy, but for *A. hypochondriacus* grain yield heterosis has been proven while it also seems promising in *A. cruentus* (Brenner *et al.*, 2010). Interspecific crosses of *A. hypochondriacus* and *A. cruentus* have been showing high yielding biomass heterosis results (Lehmann, Clark and Frey, 1991).

Clustering the 'hybridus' species complex subpopulations based on morphology has previously been attempted but remains difficult while for example addressing pistillate and staminate floral characters (Adhikary and Pratt, 2015). *A. caudatus* being the only 'hybridus' species complex grain amaranth identified as incompletely domesticated due to similar seed traits to wild *Amaranthus* species (Stetter, Müller and Schmid, 2017), while *A. hypochondriacus* and *A. cruentus* have high yielding properties (Lehmann, Clark and Frey, 1991; Brenner *et al.*, 2010), could be an additional underlaying morphological difference resulting in more distant clustering of *A. caudatus* towards *A. hypochondriacus* and *A. cruentus* (**Figure 4.7**, **Figure 4.10**). However, fixation of the domestication trait white seed colour did not occur in all three 'hybridus' complex grain amaranths, and signals of a soft selective sweep for seed colour has only been identified in *A. hypochondriacus* (Stetter, Vidal-Villarejo and Schmid, 2020). According to this study, the lack of fixation of white seed colour could be explained by ongoing gene flow from wild into domesticated amaranths, while gene flow also occurs among the three 'hybridus' complex grain amaranths with *A. hypochondriacus* and *A. cruentus* being most related to each other (Stetter *et al.*, 2020;

Figure 4). There are additional phylogenetic studies supporting higher relatedness of *A. hypochondriacus* and *A. cruentus* within the 'hybridus' species complex (Kirkpatrick, 1995; Sun, Chen and Leung, 1999; Xu and Sun, 2001; Wu and Blair, 2017; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). These combined external analyses support our current findings suggesting *A. hypochondriacus* and *A. cruentus* to be closely related, clustering close (**Figure 4.7**, **Figure 4.10**) with likely much gene flow, hybridization and admixture (**Figure 4.5**) taking place between these subpopulations.

Delimitation of *A. hybridus* is uncertain amongst taxonomists (Brenner *et al.*, 2010). In early classification studies *A. hybridus* is described as a weed which can hybridize with many species and which has similar morphology to *A. hypochondriacus* and *A. cruentus* (Sauer, 1967). A scattering of *A. hybridus* accessions among the other 'hybridus' complex species is commonly observed in other phylogenetic (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023) and morphologic studies (Adhikary and Pratt, 2015) and higher heterozygosity is observed in *A. hybridus* compared to the grain 'hybridus' complex species (Clouse *et al.*, 2016). This complements our own results in which *A. hybridus* accessions scattered around the other species groups (**Figure 4.5** black arrows, **Figure 4.7**, **Figure 4.10 A** green arrows) and high π was found in case *A. hybridus* was considered a subpopulation (**Figure 4.11**). Based on previous literature, it is a frequent statement that *A. hybridus* is the common ancestor or progenitor of the three cultivated grain amaranth species. Previous research stated that at least two separate domestication events occurred from *A. hybridus* to the three grain amaranth species (Kietlinski *et al.*, 2014). Later investigation claims it must be three separate domestication events for each separate grain species due to those being more strongly differentiated from each other than from *A. hybridus*, and subsets of *A. hybridus* clustering with each of the three grain species individually (Stetter, Vidal-Villarejo and Schmid, 2020).

The three event domestication hypothesis was based on pairwise Fst values which revealed genetic differentiation between species. However, we now suggest this hypothesis is inaccurate, as we argue that *A. hybridus* should not be considered a distinct species. Typically, Fst values of *A. hybridus* accessions are found low or intermediate when compared with other species in our work (**Figure 4.12**) but also other phylogenetic *Amaranthus* studies (Stetter, Müller and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). This is likely an artefact since currently assigned *A. hybridus* accessions actually belong to another subpopulation. What could have existed in the past is a gene pool of wild ancestral amaranth accessions containing broad genetic diversity the three cultivated species have been derived from. However, we cannot certainly state that the accessions currently annotated as *A. hybridus* do belong to such a gene pool, and since they all cluster with other amaranth species instead of with each other we conclude that *A. hybridus* is not a distinct species.

A low pairwise Fst between *A. quitensis* and *A. caudatus* species is commonly observed (**Figure 4.12**) (Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023), indicating close relationships and low genetic differentiation between these two species. Redefining artificial subpopulations by continental origin caused Fst values to drop (**Figure 4.16** versus **Figure 4.12**). However, it revealed that accessions collected from Asia, Africa and Europe seemed to be closer related to North American than

to South American accessions (**Figure 4.16**). This pattern, previously observed by Thapa *et al.* (2021), provides valuable insights into the migration of *Amaranthus* likely with human globalisation. The oldest archaeological Amaranth seed collection findings were in South America dating roughly 8000 years ago (Arreguez, Martínez and Ponessa, 2013), followed by North American findings dating back about 6000 years ago (MacNeish, 1965). Spanish conquistadors made the commonly cultivated grain crop fall into disuse in the Americas but were likely causal for *Amaranth* seed to migrate Europe (Schultes, 1984). Amaranth spread first as an ornamental plant in the 16$^{th}$ and 17$^{th}$ century and later (around 1700) as a minor grain crop in central Europa and Russia (Schultes, 1984). Amaranth has been taken to Asia and Africa by the early 19th century, but how this dispersal took place remains unclear (Schultes, 1984). What should be kept in mind is that pairwise Fst values between subpopulations defined by origin are slightly artificial since *Amaranthus* accession tend to separate strongest by species rather than geographic origin (**Figure 4.15**). Additionally, the geographic information of accessions might be incorrect due to sampling bias or sharing of accessions between locations over time.

Some studies suggest that the geographic origin of the species contribute to the broad distribution of *A. hybridus* and its ambiguous separation from other members of the 'hybridus' complex species (Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023). However, current results show that four 'hybridus' complex species can clearly be separated when using accessions of diverse global origin (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**, **Figure 4.13**, **Figure 4.15**, **Figure 4.19**). We do observe that close origin likely causes more interbreeding, hybridization and higher relatedness among accessions within species groups, but separated clades are identified per species and not geography (**Figure 4.13**, **Figure 4.15**). This is expected since species are usually defined based on genetics, and we can additionally state that the presumption of geographic background paying a role in separation of *A. hybridus* (and other 'hybridus' complex species) cannot be the solely causal factor. A similar global study using competitive KBiosciences Allele Specific PCR (KASP) technology for phylogenetic analysis of accessions originating from 35 countries suggested that *A. hybridus* clusters to a lesser extent with *A. caudatus* and *A. quitensis* (Thapa, Edwards and Blair, 2021). However, authors do recognise that more *A. hybridus* accessions should be included in future analysis to investigate the species and its potential role of domestication on multiple continents (Thapa, Edwards and Blair, 2021). The other four 'hybridus' complex species clustered similarly to our analysis (Thapa, Edwards and Blair, 2021).

We want to clarify the confusion regarding considering *A. hybridus* as a species and its relationship to other species in the 'hybridus' complex. First of all, many accessions annotated as *A. hybridus* in our analysis are likely actually different species according to our phylogenetic results since they do not seem to cluster or form a separate clade while the other species of the 'hybridus' complex do (**Figure 4.5** black arrows, **Figure 4.7**, **Figure 4.10 A** green arrows). There is only a small cluster annotated as *A. hybridus* accessions in the k=9 admixture plot marked with * (**Figure 4.5**). However, the two subpopulations (Q 5 and 6) do not seem to hybridize and have a well separated phylogeny (**Figure 4.5**, **Figure 4.9**, **Figure 4.10 B**). Q=6 accessions are now reclassified to be 'Group 1 – *A. hybridus* 2', since only two previous *A. hybridus* annotations is not sufficient evidence for the entire subpopulation to be classified as such (**Table 4.5**, **Figure 4.8**). If this Q=6 'Group 1 – *A. hybridus* 2' subpopulation was to be what is described as an *A. hybridus* species in previous

literature, it is expected to cluster with the other four 'hybridus' complex species. This is not observed in our analysis indicating Q=6 'Group 1 – *A. hybridus* 2' should be considered a different (non 'hybridus' complex) subpopulation (**Figure 4.5**, **Figure 4.7**, **Figure 4.10**). Q=5 is too small to be classified a separate species with only three accessions which are hybridizing with other subpopulation clusters (**Figure 4.5**). In short, our results do not identify *A. hybridus* as a separate species, nor does it provide evidence that *A. hybridus* is an ancestor of the other 'hybridus' complex species. Difficulties and discrepancies in investigating how the three grain amaranths domesticated from *A. hybridus* in literature are therefore reconciled, since we state that *A. hybridus* cannot be identified as an individual species.

Our analysis builds on the recent request for more *Amaranthus* sequences including more *A. hybridus* accessions for phylogenetic studies (Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). Our conclusion is that we question the existence of *A. hybridus* as a species due to unclear identification based on genomics and metabolites in our current study (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**, **Figure 4.18**, **Figure 4.19**), but also many failed separation attempts in the past (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023). Additional evidence is the high genetic diversity observed among *A. hybridus* being 2-3 times higher than the other 'hybridus' complex species (**Figure 4.11**), and low genetic differentiation when making pairwise Fst comparisons of *A. hybridus* with other species subpopulations (**Figure 4.12**). These are both confounded observations due to accessions now assigned as *A. hybridus* which should actually be allocated to different species and *A. hybridus* likely being non-existent. The existence of *A. quitensis* as a separate species has previously been discussed and sources state it should be considered the same species as *A. hybridus* (Brenner *et al.*, 2010; Stetter, Müller and Schmid, 2017). Our current results are contradictory and show clear identification of *A. quitensis*, but not for *A. hybridus* as a species (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**, **Figure 4.18**, **Figure 4.19**), which is also observed in previous studies (Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023). The difficulty of identifying *A. hybridus* as a separate species is not unexpected, since the historic definition of the species is simple, it has the notorious ability to hybridize and different races are described (Sauer, 1950). There might have been an ancestral gene pool of wild amaranth accessions with genetic variation the current domesticated species are obtained from, but *A. hybridus* cannot be identified as a distinct species. Hence, we suggest that the status of *A. hybridus* as the common ancestor of the 'hybridus' complex species likely was a myth, and that *A. hybridus* should not be considered a species or subpopulation in any future studies assessing the genus *Amaranthus*.

## Nutritionally relevant metabolites detected and further validation of species assignment

487 of 1605 detected metabolite features (30%) could be annotated (**Table 4.7**) which is way more that the average of only 1.8% of untargeted metabolomics spectra which usually is annotated in biological samples (da Silva, Dorrestein and Quinn, 2015). Hence our current annotation results indicate successful untargeted metabolite profiling of the 90 ARC accession diversity set. A contribution to this annotation success might have been combined usage of annotation databases, like SIRIUS4 (Dührkop *et al.*, 2019), GNPS (Wang *et al.*,

2016), PhytoHub (*PhytoHub*, no date) KNApSAcK Version 1.200.03 (*KNApSAcK*, no date) and ChemSpider (*ChemSpider | Search and share chemistry*, no date). It is commonly known that metabolite annotation is hard, with low feature annotation coverage as well as low confidence (da Silva, Dorrestein and Quinn, 2015; Aharoni, Goodacre and Fernie, 2023). The high 30% annotation coverage is therefore uncommon, especially when comparing it to the 5% coverage of the MAGIC founder metabolites in the manuscript in the appendix using SIRIUS4 (Dührkop *et al.*, 2019) annotations only. Our combined annotation database and dataset analysis might result in higher annotation coverages, but difficulty in interpretation and confidentiality remains due to varying annotation methods and mismatches per metabolite. We therefore want to emphasize that annotations performed with SIRIUS4 (Dührkop *et al.*, 2019) are considered most reliable in our current analysis due to its sophisticated computational approach identifying molecular structures resulting in high-confidence metabolite identification.

Many of the annotated metabolites are known to be of relevance to human nutrition, which makes accessions of the ARC diversity set potentially good lines for breeding and cultivation. Just like with MAGIC founder metabolites analysis in **Chapter 2**, the largest metabolite class was lipids, such as Prenol lipids and Fatty Acyls (**Table 4.7**). The latter, Fatty Acyls, was the largest annotated metabolite class (**Table 4.7**), which were observed in higher concentrations in the *A. cruentus* compared to *A. hypochondriacus* MAGIC founder plants in **Chapter 2**. Lipids are nutritionally important, have a wide range of biological functions, an enormous diversity in structures, and despite this diversity, lipid annotation is known to become more accurate due to improved mass spectrometry techniques (Koelmel *et al.*, 2017). However, metabolites with higher masses are usually harder to fully characterize, such as high mass spectra lipids which do not have a complete structure like an unresolved location of double bonds (de Jonge *et al.*, 2022). Another large group of annotated metabolites were organooxygen compounds (**Table 4.7**), which are often part of a carbohydrates subclass, which are sugars created by plants typically being the main source of energy for the body (Bowen and Orvig, 2008; Gerwig, 2021). Carbohydrates are involved in many biological interactions, with most cells of the human body being covered with carbohydrates linked to proteins and fats present in the cell membrane (Gerwig, 2021). Hence, carbohydrates are essential as the primary source of energy for humans. However, they are less relevant in terms of breeding efforts aimed at improving nutrition where it is more practical to focus on specialised metabolites such as flavonoids, polyphenols, and minerals. Such compounds were also detected in the ARC diversity set, with flavonoids forming a largely annotated class, along with cinnamic acids, coumarins, phenol (ether)s and tannins, which are all polyphenolic compounds present in the set (**Table 4.7**). Polyphenolic compounds have antioxidant properties, which likely results in association with various health benefits (Monjotin *et al.*, 2022). Additional benefits of polyphenolics are their contribution to flavour and to defence against pathogens and herbivory (Stiller *et al.*, 2021). The metabolites found in the ARC diversity set therefore containing promising benefits, highlighting the potential of contribution of some of these accessions for human nutrition and food systems.

Not only the class of metabolite (**Table 4.7**), but also their abundance is varying per accession (**Figure 4.20 B**). Novel species assignment based on phylogenetics (**Table 4.5**), did roughly match clusters based on this varying metabolite abundance (**Table 4.8**). This further validates our re-assignment attempt and likely non-existence of *A. hybridus* as a

species with most for now assigned *A. hybridus* accessions scattered among the various identified metabolite clusters in bi-clustering analysis and *A. hybridus* never significantly overrepresented in a cluster (**Figure 4.20 B**, **Table 4.8**). It is commonly observed that while conducting untargeted metabolite profiling of amaranths, the detected metabolites vary in class as well as abundance between species (Nkobole and Prinsloo, 2021; Abdel-Moez *et al.*, 2023). These are examples of studies relevant for medicinal (Nkobole and Prinsloo, 2021) as well as nutritional purposes (Abdel-Moez *et al.*, 2023).

Another application of untargeted metabolomics is investigating metabolic responses to abiotic stressors, such as herbicide exposure and drought in weedy species like *A. palmeri*, in which SIRIUS4 was also used as a main annotation tool (Sandhu *et al.*, 2023). However, metabolite profiling for nutrition and health benefits remains the most commonly performed in leafy amaranths (Rocchetti *et al.*, 2020; Araujo-León *et al.*, 2024; Nemadodzi and Managa, 2024). Our findings greatly contribute to the current knowledge regarding nutritionally relevant metabolites in vegetable accessions, by characterizing the presence and abundance of various metabolites across a wide range of diverse amaranth lines. Although further investigation of genomics responsible for the detected metabolite profiles would greatly benefit further understanding of underlaying mechanisms for production of relevant metabolites.

## A hint to genes that may be involved in relevant metabolite production

An attempt to get insight into potential genomic regions being causal to our identified metabolites varying between lines was done by performing mGWAS, considering metabolites as phenotypes. This gave positive results with 81 of the 1605 detected metabolite features having at least one SNP significantly associated with that metabolite (**Table 4.10**). mGWAS has been performed in many recent plant studies, like for the seed metabolome of Arabidopsis (Naake *et al.*, 2024) and buckwheat (Zargar *et al.*, 2023) to understand complex metabolomic or economic traits, or to investigate nutritional and medicinal metabolites in fruit tissue of wolfberry (Zhao *et al.*, 2024). Some examples of leaf mGWAS studies are in millet (Wei *et al.*, 2021) and maize (Riedelsheimer *et al.*, 2012) to investigate complex physiological and agronomic traits for plant growth and development, or to research waxy components on apple leaves which play an important role in stress resistance (Cao *et al.*, 2021). No mGWAS has been performed in amaranths yet, and all these previously mentioned studies used at least 100 accessions (ranging 123-3015) to conduct GWAS with.

It is commonly known that a sufficient sample size is required in order to have enough statistical power to perform GWAS, with typically a minimum of at least 100 individuals suggested for obtaining valid results (Visscher *et al.*, 2017). Although, GWAS can also work with a small number of individuals depending on the population structure, i.e. high population divergence decreases statistical power and the effect size of SNPs, making smaller effects harder to detect. Since our GWAS was performed with 90 accessions, the significant SNPs we found for metabolites can only be considered a hint to what genomic regions could be involved in the leafy amaranth metabolome. Significant SNPs are therefore just leads towards what might be a SNP located in a gene responsible for identified and annotated metabolite (**Figure 4.24**). Similar for the 55 unannotated metabolite traits for which SNPs were found to be significantly associated but which are not further visualised (**Supp table**

**metabolites with SNP associations**). Another limitation of the current study is that we only analysed metabolite data of a single field trial year while data of three consecutive years is available just as of the MAGIC founders in **Chapter 2**. Overlapping consensus metabolites of these additional datasets were used to annotate the current single year dataset in case annotations were lacking, but the feature peaks itself of the different datasets were not incorporated. Despite the likely enhanced annotation rates by including consensus of these additional sets, not using their actual peak data might make our study less complete. Including the additional metabolite data could uncover more robust changes and clarify which metabolites are likely genetically determined, i.e. consistent across time, versus environmentally influenced, i.e. present only in one of the single-trial datasets. This could enhance finding metabolome-genome associations like we are trying to achieve with mGWAS.

Quantifying SNP heritabilities provides insight into the genetic architecture of the trait, which is the joint contribution of effects and number of SNPs or variants responsible for the trait (Visscher *et al.*, 2017). Since complex quantitative metabolite traits were analysed, variants identified in GWAS generally only capture a few percent of the estimated heritability for these traits, with remaining heritability potentially being explained by other variables like rare variants, epistasis, epigenetics, environment and genotype–environment interactions (Manolio *et al.*, 2009; Eichler *et al.*, 2010; Gibson, 2010). The high heritability of metabolites M355.0664T113, M379.0636T127 and M387.3112T361 was calculated based on a heritability model describing how much each SNP is expected to influence the phenotype (Speed, Holmes and Balding, 2020). Since it was found that a major part of the phenotypic variation for these three metabolites is explained by genetics, i.e. high heritability, the reliability and relevance of the GWAS results is higher for these metabolites (**Table 4.12**). Moreover, metabolites with a higher estimated heritability (> 0.5) are likely more reliable since they additionally appear to have a lower standard deviation for their estimation compared to metabolites with an estimated heritability of around 0.5 (**Figure 4.22**). What should not be forgotten is that data of these metabolite traits are obtained from accessions grown in the same environment during a single field trial, likely causing the trait variance to be low and largely stochastic. Including data of more environments would enhance the knowledge regarding how much of the phenotypic metabolite variation is explained by genetics beyond solely heritability estimations. Just as investigation of larger genomic variants responsible for metabolites would by conducting mGWAS with SVs rather than SNPs only.

Generally, there is variation in numbers of SNPs identified as significant per metabolite trait with either high or low estimated heritabilities (**Figure 4.23**, top). Traits with a high heritability can be mendelian traits, but also more complex traits which typically have more SNPs associated with them. Although other factors like the effect sizes of individual SNPs and the sample size of the GWAS can additionally play a role while estimating heritabilities (Uffelmann *et al.*, 2021). Heritability estimations are contributory to understanding the genetic architecture of traits, with λ being additionally informative since it usually increases with higher numbers of significant SNPs per trait (**Figure 4.23**, bottom). λ greater than 1 (e.g., > 1.05) suggests inflation in test statistics and therefore not meeting the null hypothesis of no association of the 81 metabolites with significant SNPs. This makes sense since significant SNPs do cause λ to be inflated. The more SNPs are not on the diagonal in Q-Q plots means that there are more responsible SNPs for a metabolite which causes λ to

increase (**Figure 4.24**). Overfitting of the linear mixed model used for mGWAS could also cause a higher λ, although this is usually not the primary reason and only happens under specific conditions like too many fixed effects relative to the sample size, misspecification of the effect structure, or a small sample size with many phenotypes. Additional causes for inflated λ values might be unknown familial relationships, systematic technical bias, or population stratification (Clayton *et al.*, 2005; Bakker *et al.*, 2008). Population stratification is known to cofound genetic association studies (Bouaziz *et al.*, 2021). Since we know that our 90 accessions do consist of genetically heterogeneous populations by our phylogenetic analysis clustering results (**Figure 4.5**, **Figure 4.7**, **Figure 4.10 A**), population stratification might play a role.

Despite the limitations in this GWAS study like population stratification, limited sample size, and a high but potentially unreliable metabolite annotation coverage, we still mark our analysis as relevant investigation of the leafy *Amaranthus* metabolome. SNP associations were found for 81 the 1605 available metabolites, of which 26 could be annotated (**Table 4.12**). Of those, eight could be selected as of high relevance with good quality data, low p-values, high heritabilities and non-extreme SD's (**Figure 4.24**). Particularly a Hydroxyphytanic acid (alpha-) metabolite (M387.3112T361), but also two Caffeic acid-3'-O-glucuronide metabolites (M355.0664T113 and M379.0636T127) can be selected as of interest due to their high estimated heritabilities (0.98, 0.71 and 0.69, respectively). An attempt has been made to identify SNPs significant for different metabolites but co-located on the genome to address their potential involvement in the same putative biosynthetic metabolic pathway. 14 of the 1738 available significant SNPs were indeed co-located, of which seven SNP pairs are located in the same gene and another pair having similarly annotated metabolites, suggesting their involvement in functionally comparable properties (**Table 4.11**). This may indicate that multiple co-located genes contribute to the same metabolite trait and could represent a putative BGC or pathway, however, it is also very likely that these SNPs are co-inherited due to linkage disequilibrium. Our mGWAS results are a valid starting point for *Amaranthus* metabolome investigation with an attempt to identify SNPs involved in metabolite biosynthetic pathways (**Table 4.11**). However, mapping quantitative trait loci (QTL) as well as genome mining tools such as plantiSMASH could be of great contribution to identify plant gene clusters, especially if high quality genome data is available (Chavali and Rhee, 2018; Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022).

In addition to our current mGWAS, we suggest further investigation into leafy amaranth metabolites using commonly used techniques like knock-out mutants for testing of candidate genes or trying to map QTL (Alseekh and Fernie, 2018). Although, mGWAS is probably already an adequate technique to detect metabolome-genome associations having advantages over QTL studies which rely on recombination and creating mapping populations. In order to strengthen our current mGWAS study we would recommend using more repeats of metabolites data over years for robustness, as well as a higher number of accessions to increase the statistical power of the current analysis.

# Methods

## Sequences obtained and read mapping

The 90 accession diversity set were grown during the 2018/2019 growing season genotyped by short-read Illumina sequencing with 23-35x coverage. Reads were then trimmed using cutadapt v.2.3 (Martin, 2011) with a custom script in Python v.3.7.0. In order to address the sequencing statistics and quality FastQC v.0.11.7-Java-1.8.0_181 (Andrews, 2010) was run and results were combined with MultiQC v.1.13-intel-2021b (Ewels *et al.*, 2016) (**Supp table sequencing stats div set**). Fastq files were imported for all reads of all accessions of both diversity sets. Binary Alignment Map (BAM) files, were created per accession mapped to the *A. cruentus* (Ma *et al.*, 2021) and of *A. hypochondriacus* (Lightfoot *et al.*, 2017) references individually, using Burrows-Wheeler Aligner (BWA) BWA v.0.7.17-foss-2018b (Li and Durbin, 2010) for genome aligning.

## Assessing mapping coverage and selecting high quality reads

After mapping to the references, sequence coverage on exonic and non-exonic regions was calculated using BAM files of each accession compared to the complete as well as chromosomes only of the reference genomes. This was done for trimmed and raw reads of the 90 ARC accessions, and for the 121 imported reads of (Stetter *et al.*, 2020). SAMtools v.1.15-GCC-11.2.0 (Danecek *et al.*, 2021) command 'samtools bedcov', was used to calculate coverage for each accession sequence. Mean coverage of all accessions was calculated per exon or non-exon length, of which the median is calculated per diversity set (90 or 121 sequences). Coverage per exon or non-exon length were also calculated per accession individually. Boxplots including median coverage per length per accession were made using R v.4.1.1. (R Core Team, 2021) package ggplot2 (Wickham, 2016). It was checked whether coverage did match sequencing depth, i.e., fastq or BAM file size approximately in- or decreasing with coverage but also with initial species assignment (**Table 4.1**, **Supp table diversity set**). For precision, actual percentage of reads that mapped the reference genomes was checked with SAMtools v.1.15-GCC-11.2.0 (Danecek *et al.*, 2021) command 'samtools flagstat'. Quality of reads was checked with FastQC v.0.11.7-Java-1.8.0_181 (Andrews, 2010) and results combined with MultiQC v.1.13-intel-2021b (Ewels *et al.*, 2016). Accessions with reads mapping below 80% having poor read quality were excluded for further analysis (accession ID's: PI643042, PI604582, PI511731, PI490609 and PI649514).

## Variant calling

Picard v.2.21.6-Java-1.8.0_212 ('Picard Toolkit', 2019) with tools in JavaScript was used to add read groups to the BAM files and outputs were indexed using SAMtools v.1.10-foss-2018b (Danecek *et al.*, 2021). In this way BAM files were prepared for reading with Genome Analysis Toolkit (GATK). To increase accuracy, both GATK v.4.1.0.0-foss-2018b-Python-3.6.6 (Auwera and O'Connor, 2020) and freebayes v.1.3.2-foss-2018b (Garrison and Marth, 2012) were used for SNP calling. The intersection of SNPs and indels called by both tools are kept in one Variant Call Format (VCF) per accession, created with VCFtools v.0.1.15-foss-2018b-Perl-5.26.1 (Danecek *et al.*, 2011). With BCFtools v.1.9-foss-2018b (Danecek *et al.*, 2021) all VCF files were then merged into one VCF file. For this VCF file, regenotyping

was performed, i.e., identified but empty SNP or indel locations were filled with allele information using the originally created accession BAM files and the reference genome. Outputs were VCF files per reference chromosomes and scaffolds, created with freebayes-parallel v.1.3.2-foss-2018b (Garrison and Marth, 2012). Lastly, concatenating was performed with BCFtools v.1.9-foss-2018b (Danecek *et al.*, 2021), producing output VCF files for each diversity set separately. Final steps were repeated to create one final output VCF file containing both diversity sets. Number of SNPs per chromosome and scaffold were then counted in the separate and combined diversity set VCF files with the following command: grep -v '##' final.vcf | awk '{print $1}' | sort | uniq -c > count_snps.vcf & (**Table 4.4**, **Supp table SNP count**).

## Phylogenetic analysis for species identification

The final VCF file was filtered for biallelic SNPs using the 'bcftools view' command of BCFtools v.1.10.2-foss-2018b (Danecek *et al.*, 2021). The SNP quality (QUAL) column was extracted to filter by excluding the first quantile using VCFtools v.0.1.16-foss-2018b-Perl-5.28.0 (Danecek *et al.*, 2011). The VCF file was then zipped and indexed with 'bcftools index' of of BCFtools v.1.10.2-foss-2018b (Danecek *et al.*, 2021). For conducting population phylogenetic analysis the VCF was filtered by LD using the 'plink --indep-pairwise' command of PLINK v.1.9_beta-6.16 (Chang *et al.*, 2015), with pairwise comparisons calculated in windows of 10kb, a step size of 1, and an R Squared ($R^2$) threshold of 0.5. A VCF file and binary file sets were loaded using the 'plink –vcf' and '--make-bed' commands, followed by referencing with '--bfile' and producing eigenvec and eigenval files for principal components with '–pca' of PLINK v.1.9_beta-6.16 (Chang *et al.*, 2015). The eventual PCA graph was then plotted with R package ggplot2 (Wickham, 2016) using R v.4.1.1. (R Core Team, 2021).

The VCF file which was filtered for quality as well as linkage as just described for making the PCA was used to for Admixture analysis. This vcf which was further filtered with the following command: 'vcftools --vcf phylo_chr.vcf.recode.vcf --mac 1 --maf 0.01 --recode --recode-INFO-all --out out_new_phylo_chr.vcf.recode.vcf'. A bed and additional files were created with PLINK v.1.9_beta-6.16 (Chang *et al.*, 2015), commands: 'plink --vcf file.vcf.gz --make-bed --out output --allow-extra-chr'. ADMIXTURE v.1.3.0 (Alexander, Novembre and Lange, 2009) was run on the bed file for k=20 clusters. Graphs derived from the Q output files were then plotted with R package ggplot2 (Wickham, 2016) using R v.4.2.3 (R Core Team, 2021).

The VCF file filtered for quality as well as linkage for conducting PCA was used for making a phylogenetic tree. This vcf was further filtered with the following command: 'vcftools --vcf old.vcf --mac 1 --maf 0.01 --min-alleles 2 --max-alleles 2 --bed single_copy.bed --max-missing 0.9 --remove-indels --recode --recode-INFO-all --out new'. This made the SNPs in the vcf drop from 11,696,020 prior to 393,611 after filtering (counted with; cat file.vcf | grep -v "^#" | wc -l > count_snps). Here 'single_copy.bed' contains conserved single copy genes identified by OrthoFinder of the *A. hypochondriacus*, *A. cruentus* and *A. tricolor* reference genomes (Lightfoot *et al.*, 2017; Ma *et al.*, 2021; Wang *et al.*, 2022). A PHYLIP matrix format for phylogenetic analysis was created with vcf2phylip v2.0 (Ortiz, 2019) using Python v.3.10.4-GCCcore-11.3.0. IQ-TREE 2 (Minh *et al.*, 2020) was then used with 1000 fast bootstrap replicates to create a tree file as input for visualisation of an unrooted phylogenetic tree by maximum likelihood with Interactive Tree Of Life (iTOL) v.5 (Letunic and Bork, 2021).

## Genetic diversity, differentiation and origin of the novel assigned species groups

Based on the previously performed phylogenetic analysis (PCA, Admixture plot and phylogenetic tree) new species groups have been assigned to the diversity set accessions. The SNP VCF file was split up into separate VCF's per species group and π was calculated in 10kb windows for each VCF set respectively with VCFtools v.0.1.16-foss-2018b-Perl-5.28.0 (Danecek *et al.*, 2011). The VCF file which was filtered for quality as well as linkage prior performing PCA was used to calculate pairwise fixation index ($F_{ST}$) statistics between subpopulations according to Weir and Cockerham (1984). This was done with VCFtools v.0.1.16-GCC-11.2.0 using the following command; 'vcftools --vcf "$vcf_file" --weir-fst-pop "$pop1" --weir-fst-pop "$pop2"' (Danecek *et al.*, 2011). Box- and violin plots visualizing π with test letter codes for significance (Tukey's HSD test for post-hoc comparisons) between subpopulations, heatmaps displaying the pariwise $F_{ST}$ values, and the world maps with the origin information of accessions were all made with R package ggplot2 (Wickham, 2016) in R v.4.2.3 (R Core Team, 2021).

## Metabolite profiling

The 90 different diversity set amaranth accessions were grown by the ARC in an irrigated field trial during the 2018/2019 growing season and leaf samples were collected and freeze-dried for metabolite content. This field trial was repeated twice in two consecutive growing seasons, with metabolite data on 93 and 75 accessions for years 2019/2020 and 2020/2021, respectively. Overlapping with the 90 accessions of the first trial this resulted in approximately 78 and 60 consensus accessions for the two consecutive growing trials respectively. All leaf samples were used for untargeted metabolome analysis with methanolic extracts run twice for repeatability for data of the first field trial and three times for data of the second and third field trial. Samples are run in positive and negative ionization modes for obtaining the broadest metabolome coverage (Perez de Souza *et al.*, 2019).

Obtained feature area data were background-corrected, internal standard and dry-weight (DW) normalised area values. Using R v.4.1.1. (R Core Team, 2021), the data were filtered to keep all features with a non-zero abundance for all repeats of at least one accession, and with a relative standard deviation (RSD) of =<30% of the mean for at least one accession. The internal standard (ISTD) values for reserpine were removed. All (un)annotated values were kept, and zeros were imputed (using the 'imp()' function in R v.4.1.1. (R Core Team, 2021)) with 1/5 of the non-zero minimum in order to calculate the max fold-changes. Positive and negative ionization mode peaks were combined, and potentially redundant features identified by a squared Pearson's correlation coefficient ($R^2$) > 0.8 for abundance, features within a two second retention time, and mass-to-charge ratio (m/z) differences matching differences of potential sensible adducts. In case of multiple matches, the best correlated pair was selected and from this pair, the single feature with the lowest RSD was kept. Then the data was additionally normalised by sample medians with R v.4.1.1. (R Core Team, 2021), followed by normalisation by glog10 transformation (LogNorm() function), and lastly pareto scaling for normalisation (ParetoNormM() function).

Metabolite annotation was performed with SIRIUS4 (Dührkop *et al.*, 2019) and Global Natural Products Social Molecular Networking (GNPS) (Wang *et al.*, 2016). Both Sirius and

GNPS annotation methods rely on tandem mass spectrometry (MS/MS or MS2), commonly used in untargeted metabolomics (Wang *et al.*, 2016; Dührkop *et al.*, 2019). For GNPS both collision induced dissociation (CID) and higher-energy collisional dissociation (HCD) molecule fragmentation was used. Metabolites were additionally annotated with PhytoHub v.1.4 (*PhytoHub*, no date), KNApSAcK v.1.200.03 (*KNApSAcK*, no date) and ChemSpider (*ChemSpider | Search and share chemistry*, no date) which are based on MS1. All SIRIUS4 annotations were taken as most reliable annotation since SIRIUS4 uses high resolution fragmentation data (Dührkop *et al.*, 2019). Names and classes according to SIRIUS4 were therefore taken as annotation first. If not annotated by SIRIUS4, metabolite names were filled in by PhytoHub, KNApSAcK and ChemSpider hits. If classes were not annotated by SIRIUS4 gaps were filled in by GNPS identified classes and subclasses. For the metabolite names annotated by PhytoHub, KNApSAcK and ChemSpider it were always the first field trial data hit, followed by the consecutive years hits and lastly an additional *Amaranthus* dataset hit of **Chapter 2** (MAGIC founders) if not annotated by any of the other datasets in this order.

A principal component analysis (PCA) plot was made based on the normalised data of all recorded peaks with positive versus negative ionization mode redundancy removed for the two sample run repeats of the first field trial data (2018/2019 growing season) using the GGally (Schloerke *et al.*, 2021) R package. Once this replicate data was validated for similarity the mean of two repeat samples was taken and another PCA was made to reveal species clustering of the 90 diversity set accessions using GGally (Schloerke *et al.*, 2021). Species clustering based on metabolites as well as genomics was performed by isolating the branches of the 90 ARC set accessions out of the phylogenetic tree previously made in **Figure 4.10** using ape (Paradis and Schliep, 2019), then making a dendrogram based on the average sample metabolite data with ggtree (Yu, 2022) and plot both trees connected as a tangle gram in R v.4.2.3 using ggplot2 (Wickham, 2016).

For the annotated metabolites, normalized data was used to perform bi-clustering in order to generate a rearranged checkerboard structure data matrix (X) with each cell of the matrix defined by a as homogeneous as possible Q and P. Here X, Q and P are:

- X is a data matrix with I rows and J columns. Xij is a response measure of row i in column j for $i \in \{1, 2, \ldots, I\}$ and $j \in \{1, 2, \ldots, J\}$.
- Row index set $I = \{1, 2, \ldots, I\}$ is partitioned into r mutually exclusive and exhaustive sets T1 , T2, . . . , Tr. Q ≡ partition of the row index set.
- Column index set $J = \{1, 2, \ldots, J\}$ is partitioned into c mutually exclusive and exhaustive sets S1 , S2, . . . , Sc. P ≡ partition of the column index set.

Therefore, there are r row groups (or row clusters, here annotated metabolites) and c column groups (or column clusters, here accessions). The algorithm has the goal of optimizing a total within-cluster sum of squares given both the r row groups in Q and c column groups in P. This SSE can hence be referred to as the total sum of squared errors between metabolite means per cell and data values within cells. Biclustering was performed with r package ‘biclustermd’ (Li *et al.*, 2020) with R v.4.2.3 (R Core Team, 2021). Accessions per species after re-classification present in each cluster were counted, and a Fisher’s exact test was conducted with R v.4.5.2 to determine if species are significantly over-/underrepresented and how much more or less (odds ratio) a species is to occur in a cluster compared to others.

## mGWAS

The final VCF file after concatenation as described in **Chapter 4: Variant calling**, of 90 ARC accession trimmed reads mapped towards the *A. cruentus* reference genome contained all available SNPs. This VCF was further filtered to remove the lowest-quality SNPs (the first quantile) and retain only variants with sufficient information scores for mGWAS. The first quantile of quality SNPs was calculated as follows:
'vcftools –gzvcf merge.crom_regeno_freebayes.gatk.union.vcf.gz --site-quality --out site_quality' for which 'wc -l site_quality.lqual' gave: 27785980 site_quality.lqual which divided by four is 6946495. Then 'cat site_quality.lqual | grep -v 'CHROM' | sort -k3,3n | head -n 6946495| tail -n1' gave a quality cutoff of 'chr08 21075054 356.97'
The read depth (DP) coverage of each SNP position was checked with:
'bcftools query -f '%DP\n' merge.crom_regeno_freebayes.gatk.union.vcf.gz | awk '{ total += $1; count++ } END { print total/count }' which gave 1133.47.
The higher coverage cutoff is therefore 1133.47*4 = 4533.88 ~ 4533 while the lower coverage cut-off was set to 1133.47/4 = 283.3675 ~ 283.
Filtering thresholds were determined empirically using site-quality and info metrics, with DP limits of 283–4533 and mapping quality metrics (MQM and MQMR >30) to ensure high-confidence SNPs. Filtering was eventually applied using the following bcftools command:
'bcftools view -m 2 -M 2 -i ' TYPE="snp" & QUAL > 356 & INFO/DP > 283 & INFO/DP < 4533 & INFO/MQM >30 & INFO/MQMR >30'
merge.crom_regeno_freebayes.gatk.union.vcf.gz > Maud_Filtered1.vcf'

Then the VCF file was further filtered for Minor Allele Count (MAC) ≥ 1 to remove monomorphic sites, ensuring that only positions with at least one allele count of the least common allele were retained. Additionally, the Minor Allele Frequency (MAF) threshold of 0.05 was applied to exclude rare variants, retaining only variants with sufficient frequency in the population to provide statistical power for mGWAS. Lastly only positions where the missing data was at max 30% were included in order to minimize biases from variant calling coverage issues and to ensure robust downstream analyses. Hence these filtering commands were implemented:
'vcftools --vcf Maud_Filtered1.vcf --mac 1 --maf 0.05 --max-missing 0.7 --recode --recode-INFO-all --out Maud_Filtered2_maf0.05'.

In order to run ldak for mGWAS additional files were created with the following commands:
'./vcftools --plink --vcf Maud_Filtered2_maf0.05.recode.vcf --out div_set'
'./plink --file div_set --make-bed --autosome-num 36 --out div_set'
In the bim file larger indels were replaced with 'A' and 'B' because ldak can only deal with single character SNPs:
'mv div_set.bim div_set_OLD.bim' and then 'awk '{print $1"\t"$2"\t"$3"\t"$4"\t""A""\t""B"}' div_set_OLD.bim > div_set.bim'.
A kinship matrix was generated to account for population structure and relatedness using ldak v.5.2 (Zhang *et al.*, 2021) commands:
'./ldak --calc-kins-direct div_set.kinship --bfile div_set --ignore-weights YES --power -0.25'.

It was checked if the average of two sample repeat metabolite data which was normalized as previously described was suitable for analysis by making a histogram which showed a normal distribution. mGWAS analysis was performed using a linear mixed model

considering the metabolites as 1605 different traits with the following command: ‘for ((i=1;i<=1605;i++)); do ../ldak5.2.linux --linear metabolite.trait.${i} --bfile ../div_set --grm ../div_set.kinship --pheno ../div_set_complete --mpheno ${i} --max-iter 500 --tolerance 0.01; done’.
Then permutations of the model were run with up to 500 iterations and a convergence tolerance of 0.01 to ensure robust fitting:
‘for ((i=1;i<=1605;i++)); do ../ldak5.2.linux --linear metabolite.trait_1_permutation.${i} --bfile ../div_set --grm ../div_set.kinship --pheno ../div_set_complete --mpheno ${i} --max-iter 500 --tolerance 0.01 --permute YES; done’.
Significance thresholds for the mGWAS were determined empirically by repeated phenotype permutation analysis to generate a null distribution of association statistics: ‘sort -g -k7,7 metabolite.trait_1_permutation.*.assoc | awk '$7!="NA"{print $7}' | head -n 2 | tail -n 1 >> ZZZ_assoc_all_lowest’ and then ‘sort -g ZZZ_assoc_all_lowest | head -n 81 | tail -n1’ gave 3.2011e-12 as a genome-wide cutoff.

Data of significant SNPs was extracted. For metabolites with significant SNPs Q-Q (quantile-quantile) plots were made to check the quality of data by comparing the distribution of SNPs to the standard normal distribution, and Manhattan plots were made to plot the significant SNPs all while using the ggplot2 (Wickham, 2016) package in R v.4.2.2 (R Core Team, 2021) in Linux. Metabolites for which at least one SNP was found significantly associated were isolated. Locations of significant SNPs from different metabolites were compared and those located within a 1000bp window identified as being potentially involved in the same biosynthetic pathway. If the co-located SNPs were physically located in annotated genes was checked with GenomeView (Abeel *et al.*, 2012). The genomic inflation factor (λ) was estimated for p-values of each Q-Q plot for each separate metabolite trait for which at least one SNP was found to be associated. This was done in linux with R v.4.2.2 (R Core Team, 2021) with λ = median(observed test statistics) / median(expected test statistics).

## Heritability of metabolite traits

The concatenated VCF file made by the 90 trimmed reads mapped to the *A. cruentus* reference was filtered for quality based on the previously detected threshold of 356.97 with: ‘vcftools --gzvcf merge.crom_regeno_freebayes.gatk.union.vcf --minQ 357 --recode --recode-INFO-all --out merge.crom_biallelic.QUAL.filter.357’. Then the file was filtered for biallelic SNPs with ‘vcftools --vcf merge.crom_biallelic.QUAL.filter.357.recode.vcf --min-alleles 2 --max-alleles 2 --recode --recode-INFO-all --out biallelic.QUAL.filter.357’. PLINK v.2.00a3.6-GCC-11.3.0 (Chang *et al.*, 2015) was run to make files needed for running LDAK: ‘plink2 --vcf biallelic.QUAL.filter.357.recode.vcf --make-bed --autosome-num 35 --allow-extra-chr --out output.357’. Chromosome and scaffold names were made to a positive integer and indels were replaced with ‘A’ and ‘B’ in the bim files so it could be read by LDAK. The kinship matrix was generated with the command: ‘./ldak5.2.linux --calc-kins-direct kinship.output --bfile output.357 --ignore-weights YES --power -0.25’ with ldak v.5.2 (Speed, Holmes and Balding, 2020). The same normalized averages of two metabolite sample repeats were used for the mGWAS and the heritability of all 1605 traits which were calculated with LDAK as follows: ‘for ((i=1;i<=1605;i++)); do ../ldak5.2.linux --reml metabolite.trait.${i} --grm ../kinship.output --pheno ../div_set_complete --mpheno ${i} --max-iter 500 --constrain YES; done’. The heritability and their standard deviation (SD) of all metabolites, but also only those for which a significant SNP was detected by GWAS, were

plotted as histograms using R v.4.2.3 (R Core Team, 2021). The correlations between heritability and the standard deviation of this estimated heritability were also plotted for all metabolites as well as those for which at least one significant SNP was identified using ggplot2 (Wickham, 2016) in R v.4.2.3 (R Core Team, 2021). Lastly, using these same R tools the heritability as well as λ have been plotted against the number of significantly associated SNPs per metabolite.

# Chapter 5: Identification and analysis of putative biosynthetic gene clusters (BGC).


## Abstract

Leafy amaranth is an underutilized crop commonly consumed in Africa, containing a rich and diverse micronutrient profile. Therefore, it has a great potential to contribute to healthy diets and nutrition beyond simply meeting calory demands. Studies have been conducted exploring the metabolome of amaranths, however, linking metabolite content to Amaranth genomics is still in its infancy. One way to explore genes involved in metabolites is by elucidating biosynthetic pathways, which are stepwise enzymatic processes producing these nutritional compounds. If the genes encoding for these enzymes involved in metabolite biosynthetic pathways are located closely or clustered on the genome, these are called biosynthetic gene clusters (BGC). Knowledge of BGC existence in amaranths can aid the prediction of nutritional metabolite profiles in this vegetable crop.

Previous attempts for BGC identification have been performed in a high quality *Amaranthus cruentus* reference genome, i.e. Arusha. Additionally, the well-known betalain pigmentation BGC genes, *DODAα1* and *CYP76AD1*, have been investigated in *A. hypochondriacus*, *A. cruentus* and *A. tricolor* genomes and other species belonging to the *Amaranthaceae* family. Ways to identify BGC are by genome mining with plant BGC prediction tools like plantiSMASH and further evidence of BGC existence can be provided by revealing BGC gene co-expression indicating co-regulation or co-inheritance suggesting conservation.

With the generation of eight novel high quality sequences of the MAGIC founder lines and updated gene annotation of Arusha, the current study attempted to reveal putative BGC with plantiSMASH in these *Amaranthus* genomes. We have successfully identified 43 putative BGC in the nine available genomes, with 11 BGC selected as of interest due to co-expressed genes. Conservation of all BGC was observed to be either high or low, with especially high conservation observed among those selected BGC of interest marking their functional relevance. Not many BGC were species specific, which indicates little selective pressure of BGC for certain species and their biological purpose being useful for all amaranth species. The co-localized genes of the betalain BGC were found in all novel or re-annotated genomes with their expression either high or low per sample which often could be correlated with tissue pigmentation of the collected sample. Pigmentation matching gene expression patterns shows analysis and gene identification likely being correct, while the high conservation of *DODAα1* and *CYP76AD1* genes among all accessions highlights their relevance among *Amaranthaceae* species.

We are happy with the successful identification of relevant BGC in our nine high quality genomic resources, of which many are likely involved in nutritionally relevant metabolite synthesis. However, we have a strong request for additional identification of biosynthetic pathways and genes involved in amaranth dietary metabolites, which could be investigated by using multi-omics approaches and expanded genomic datasets.

# Introduction

To overcome food security issues, improvement of staple food crops like maize, wheat or rice is often performed, and although these crops are typically calorie rich, they are usually not sufficient to meet the requirements for essential nutrients (Joshi *et al.*, 2018). Hence, vegetable amaranth as an underutilized crop with a diverse micronutrient profile, has a high potential for contributing to the nutritional demand in Africa (Rastogi and Shukla, 2013; Achigan-Dako, Sogbohossou and Maundu, 2014; Mabhaudhi *et al.*, 2016; Joshi *et al.*, 2018). In order to investigate the broad spectrum of nutritional as well as medical purposes of amaranths, many recent studies have been conducted exploring the metabolome and metabolite content of various varieties under diverse growing conditions (Rocchetti *et al.*, 2020; Nkobole and Prinsloo, 2021; Abdel-Moez *et al.*, 2023; Sandhu *et al.*, 2023; Araujo-León *et al.*, 2024; Nemadodzi and Managa, 2024). However, linking the metabolite content of Amaranth to its genomics and investigating which genes are responsible for the *Amaranthus* metabolome only started recently (Sarker and Oba, 2020; Ma *et al.*, 2021).

## Biosynthetic gene clusters (BGC) in plants and amaranths.

Identifying the genes involved is crucial for investigating the large variety of plant metabolites and their diverse functions, this requires elucidating biosynthetic pathways which is the stepwise enzymatic processes by which these compounds are produced. (Yang *et al.*, 2014; Sumner *et al.*, 2015; Li *et al.*, 2021; Zhan *et al.*, 2022). Some examples of known biosynthetic pathways present in *Amaranthus* are the shikimate/phenylpropanoid pathway which produces physiologically important phenolic compounds like phenolic acids, flavonoids, and lignans, but also the betalain biosynthetic pathway (Brockington *et al.*, 2011; Naikoo *et al.*, 2019, p. 9; Sharma *et al.*, 2019; Rocchetti *et al.*, 2020; Ma *et al.*, 2021; Wang *et al.*, 2022; Sandhu *et al.*, 2023; Winkler *et al.*, 2024). The later, betalain, was one of the first metabolites detected in the *Amaranthaceae* family which is a well-studied pathway producing pigmentation (Brockington *et al.*, 2011). This reveals the significance of betalain as a natural food colorant, while additionally the betalain intermediate *L*-DOPA is highly relevant due to its usage to treat Parkinson's disease (Wang *et al.*, 2022). The expression of *DODAα1*, *CYP76AD1* and *MYB2* genes is often observed higher in red or mixed colour leaf, flower, root and stem tissue of amaranths (Wang *et al.*, 2016; Winkler *et al.*, 2024). It is shown that the two betalain-specific orthologues genes, *DODAα1* and *CYP76AD1* are co-located in the genomes of six *Amaranthaceae* species: on chromosome 16 in *A. hypochondriacus*, *A. cruentus* and *A. tricolor*, and on different chromosomes in *Beta vulgaris*, *Chenopodium quinoa*, and *Spinacia oleracea* (Ma *et al.*, 2021; Wang *et al.*, 2022). Hence, this betalain metabolomic pathways is a famous example of a so called biosynthetic gene cluster (BGC), since BGC are closely located clustering genes which encode for enzymes involved in metabolite biosynthetic pathways (Boycheva *et al.*, 2014). *MYB2* is not co-located with *DODAα1* and *CYP76ADα1* and thus not part of the BGC (Wang *et al.*, 2022).

One way to identify BGC responsible for biosynthetic pathways using metabolomic plus genomic datasets directly is by performing metabolite-based genome-wide association studies (mGWAS) or metabolite-based quantitative trait loci (mQTLs) (Alseekh and Fernie, 2018; Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). Although both techniques can be used for identifying plenty of metabolite-genome associations along the genome, they can reveal putative BGC in case multiple closely located genomic regions are found associated for

similar but not the same metabolite. An example of attempting to identify such putative BGC by mGWAS was performed with the diversity set metabolite data in **Chapter 4**. Here 14 closely located (1000 bp distance) SNPs were significantly associated with different metabolite compounds (**Chapter 4: Table 4.11**). This is only a hint to what could be SNPs located in multiple genes which could be a putative BGC. However, since 7 of the 14 SNP pairs lie within the same gene, it is more likely that the associated metabolites belong to the same biosynthetic pathway, rather than being products of multiple genes in a BGC. Another way of identifying potential BGC responsible for metabolomic biosynthetic pathways is by using plant gene cluster prediction tools, like plantiSMASH, PhytoClust, C-Hunter and many others, which predict candidate BGC directly from high quality genomic data (Chavali and Rhee, 2018). PlantiSMASH was used in previously published versions of the *A. cruentus* and *A. hypochondriacus* references, which detected 22 and 23 BGC in each genome respectively (Ma *et al.*, 2021). Looking if BGC are conserved between species is another common way to identify or provide more evidence for putative BGC, since BGC are expected to be co-inherited (Smit and Lichman, 2022). Ten of the BGC detected in *A. cruentus* and *A. hypochondriacus* were conserved among both genomes, of which six also showed conserved co-location of genes across the *Amaranthaceae* family species *B. vulgaris*, *C. quinoa* and *Spinacia oleracea* genomes (Ma *et al.*, 2021). One of these six clusters was highlighted since genes encoding three different types of biosynthetic enzymes were conserved across all five *Amaranthaceae* genomes (Ma *et al.*, 2021; Figure 6).

## Identification of putative BGC in our dataset

With the creation of a MAGIC population and *Amaranthus* pan-genome, eight novel high quality *Amaranthus* genome sequences have been generated and the gene annotation of the *A. cruentus* reference genome (Arusha) has been updated (manuscript in appendix). It would be of great interest to reveal putative BGC located in these nine valuable genomic resources, due to the potentially nutritionally beneficial metabolic compounds they could produce. Hence, a first research aim of this chapter is to investigate if putative BGC can be identified in the nine high quality genomic resources by using plant BGC prediction tools.

With the inclusion of *A. cruentus* as well as *A. hypochondriacus* accessions, conservation of putative BGC among species can be investigated. A second aim of this chapter is to investigate if the identified candidate BGC are species specific or conserved among accessions, which provides further support of their existence and co-inheritance.

Gene expression data has been collected for various tissues and samples of the nine accessions with available genomes, and 21 ARC diversity set accessions which were expected to be *A. cruentus* or *A. hypochondriacus* before re-annotation. With the existence of RNA sequencing data for *Amaranthus* accessions investigated in this thesis, further evidence can be provided for the identified potential BGC. This can be done by showing if the identified genes of a putative BGC are found to be expressed simultaneously in different samples. Therefore, a third research aim is to further investigate the existence of identified putative BGC by exploring co-expression patterns of their clustered genes.

Lastly, the well-known betalain BGC can be thoroughly investigated due to its role in tissue pigmentation. An important goal is to examine if the betalain BGC is present in all newly obtained genomes and if its gene expression coincides with pigmentation as expected.

# Results

## Putative biosynthetic gene cluster (BGC) identification

To investigate presence of putative BGC in the nine available genomes, i.e. eight novel MAGIC founder sequences and updated Arusha, BGC prediction tool plantiSMASH (Kautsar *et al.*, 2017) was used. A total of 226 BGC were identified (23-29 per accession), of which 43 unique ones either overlapping between accessions being conserved or being unique to a single accession (**Supp table BGC**). According to plantiSMASH, this total of 43 BGC consisted of the following types: 8 alkaloids, 3 lignans, 1 lignan-terpene, 3 polyketides, 11 saccharides, 1 saccharide-terpene, 2 terpene and 14 putative or unknown clusters. This reveals that identification of putative BGCs in the nine *Amaranthus* genomes has been successful and that many of the BGCs are likely nutritionally relevant, such as those for lignans, saccharides, and terpenes, or pharmaceutically relevant, like those for alkaloids.

**Table 5.1** The 43 identified putative biosynthetic gene cluster (BGC) ID's ordered by number of (No) accessions they occur in. The No of BGC found per No accession group is provided. It is mentioned between brackets which BGC ID is specific to *A. h.*, *A. c.* or MAGIC parent accessions only.

| No accessions with this BGC | No BGC | BGC ID's (1 to 43) |
|---|---|---|
| 1 | 12 | 2, 6, 14, 15, 16, 20, 22, 33, 34, 38, 40, 43 |
| 2 | 4 | 10, 19, 29, 32 |
| 4 | 2 | 13 (*A. h.*), 31 |
| 5 | 3 | 9 (*A. c.*), 21, 37 |
| 6 | 1 | 41 |
| 7 | 4 | 1, 4, 5, 8 |
| 8 | 4 | 23 (MAGIC), 28 (MAGIC), 30, 42 |
| 9 | 13 | 3, 7, 11, 12, 17, 18, 24, 25, 26, 27, 35, 36, 39 |

To examine the conservation of BGC, their presence among accessions and species was addressed, by manually comparing the cluster type, type of genes and if applicable the chromosomal location. Of the 43 BGC, twelve putative BGC are unique to a single accession and thirteen were conserved among all nine accessions, with only a few BGC being identified in 2 to 8 accessions with a maximum of four per group (**Table 5.1**). This indicates that BGC generally are not conserved at all or very well conserved among accessions, the latter providing additional proof that the BGC are real. One BGC was specific to the five *A. cruentus* accessions only (putative uncategorized BGC type with cytochrome 450 and methyltransferase genes), one was specific to the four *A. hypochondriacus* accessions only (saccharide BGC type with cytochrome 450 and glycosyltransferase genes) (**Figure 5.1**) and two BGC were MAGIC parent specific (absent in Arusha) (**Table 5.1**). This shows that not many BGC were species specific,

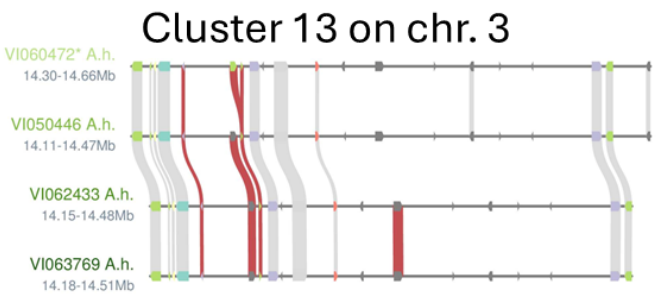



**Figure 5.1** Co-localisation of genes (red) of cluster 13 which is species specific to *A. hypochondriacus* accessions and assigned to be a nutritionally relevance saccharide type BGC.

and those that are may be of nutritional relevance. Especially the saccharide BGC found only in *A. hypochondriacus* species, for which their synteny of genes across accessions is shown in **Figure 5.1**.

## Evaluation of gene expression data and species assignment

To evaluating the reproducibility of the available gene expression data of leaf tissue of 21 ARC diversity set accessions, the MAGIC parents and leaf, root and seedling tissue of Arusha principal component analysis (PCA) was conducted. PCA was conducted for each data set separately, i.e., MAGIC parents with three leaf repeat samples plus three tissue samples of Arusha or single leaf samples of the ARC diversity set accessions (**Figure 5.2 A**, **B**, respectively) or combined data sets (**Figure 5.2 C**, **D**). RNA-seq quantification was performed using different transcriptomes for different datasets, i.e., each individual transcriptome for the eight MAGIC founders and the Arusha transcriptome for Arusha and the diversity set accessions. Hence PCA comparisons between gene expression quantified with different transcriptomes were made by using core genes only as identified by OrthoFinder v.2.3.3 (Emms and Kelly, 2015) (**Figure 5.2 A, C**, **D**). The repeated leaf samples (B1 and B2) for the MAGIC parents cluster more nicely together than the physically different leaf sample A, while overall the RNA samples cluster per species group and tissue type (**Figure 5.2 A**). Additional clustering per species group could be observed in PC1-PC2 of samples of the 21 ARC set accessions with species annotation according to **Chapter 4** (**Figure 5.2 B**). PC1-PC2 of both datasets combined, separates the RNA-seq samples by transcriptome used for quantifying and by tissue type, with leaf Arusha clustering with diversity set accessions (**Figure 5.2 C**). PC2-PC3 of the combined datasets then separates by species group and sample tissue type, with the exception of one *A. quitensis* accession clustering with the *A. hypochondriacus* accessions (**Figure 5.2 D**). Overall, this shows that we have adequate and reproducible gene expression data, clustering per tissue type, sample repeat, species groups and transcriptome used for RNA-seq quantification. Additionally, these results provide more evidence that the assigned species groups of the diversity accessions are correct (**Chapter 4: Table 4.5**).

Now we have shown that our RNA-seq data is of sufficient quality, heatmaps of gene expression of the putative BGC genes are made to explore if these BGC are real. First expression heatmaps of each gene in the BGC are generated per accession and per sample, to further evaluate quality of the RNA-seq data. These heatmaps display similar expression patterns per leaf sample type (A and especially B1 and B2 being similar) while the different tissue samples taken from Arusha clearly show different patterns (**Figure 5.3**, **Figure 5.4**). This shows reproducibility of gene expression patters over sample repeats, which indicates a good quality of RNA-seq data. The heatmap based on single leaf sample RNA expression of the diversity set accessions, clusters at the top according to the assigned species, with expression patterns becoming more divergent towards the right for more distantly related species (**Figure 5.4**). This also indicates good quality RNA-seq data further confirming species assignments of the diversity accessions (**Chapter 4: Table 4.5**).

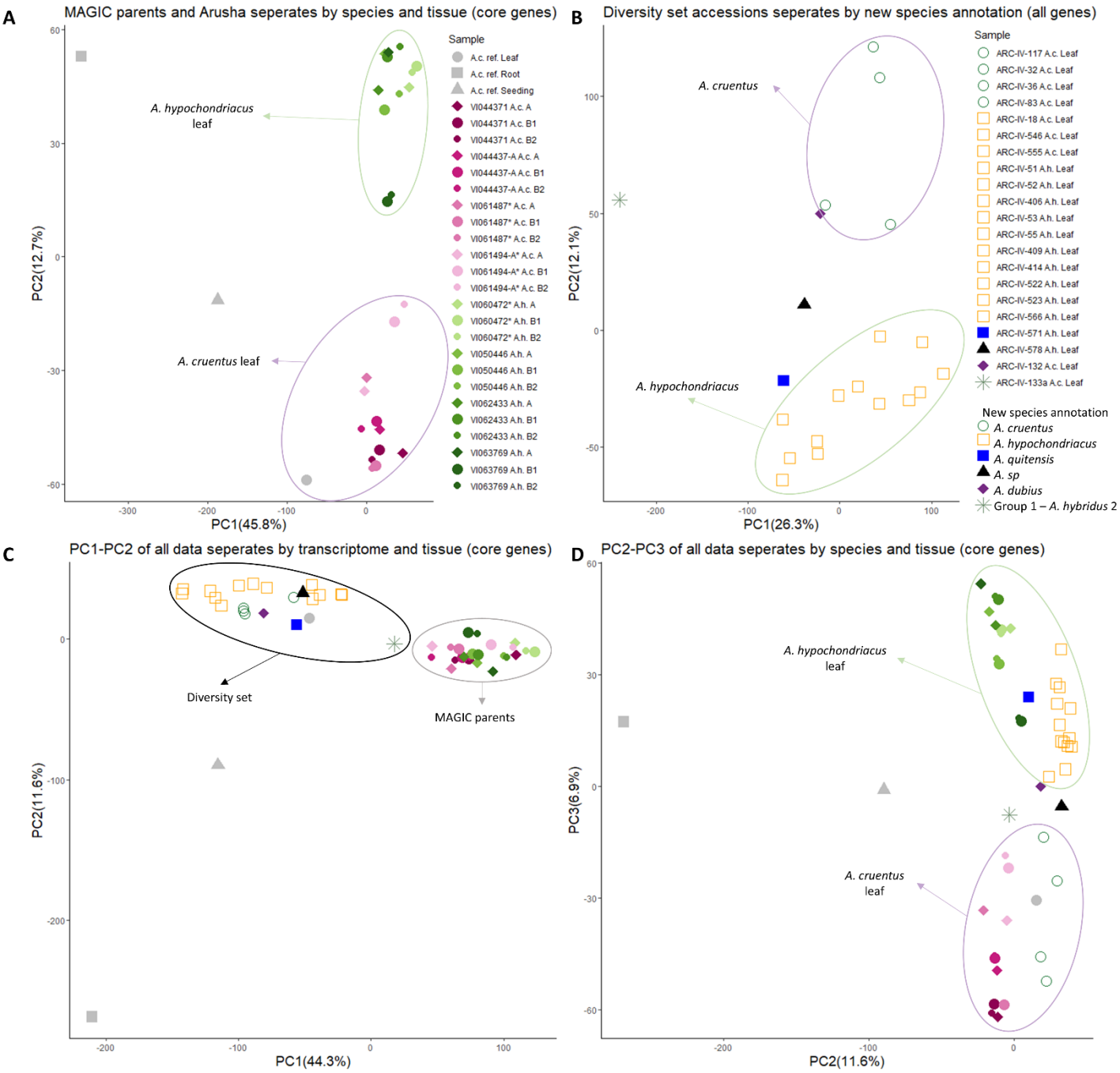



**Figure 5.2** PCA based on expression data showing clustering of RNA-seq samples for **A)** MAGIC parents and Arusha (*A. cruentus* reference (Ma *et al.*, 2021)), **B)** 21 diversity set ARC accessions, **C)** PC1-PC2 and **D)** PC2-PC3 of both sets combined. Clustering mostly occurred per species (A, B and D) and per sample tissue type (A, C and D) with only PC1-PC2 of all data combined separating for the transcriptome used for RNA-seq quantification (C).

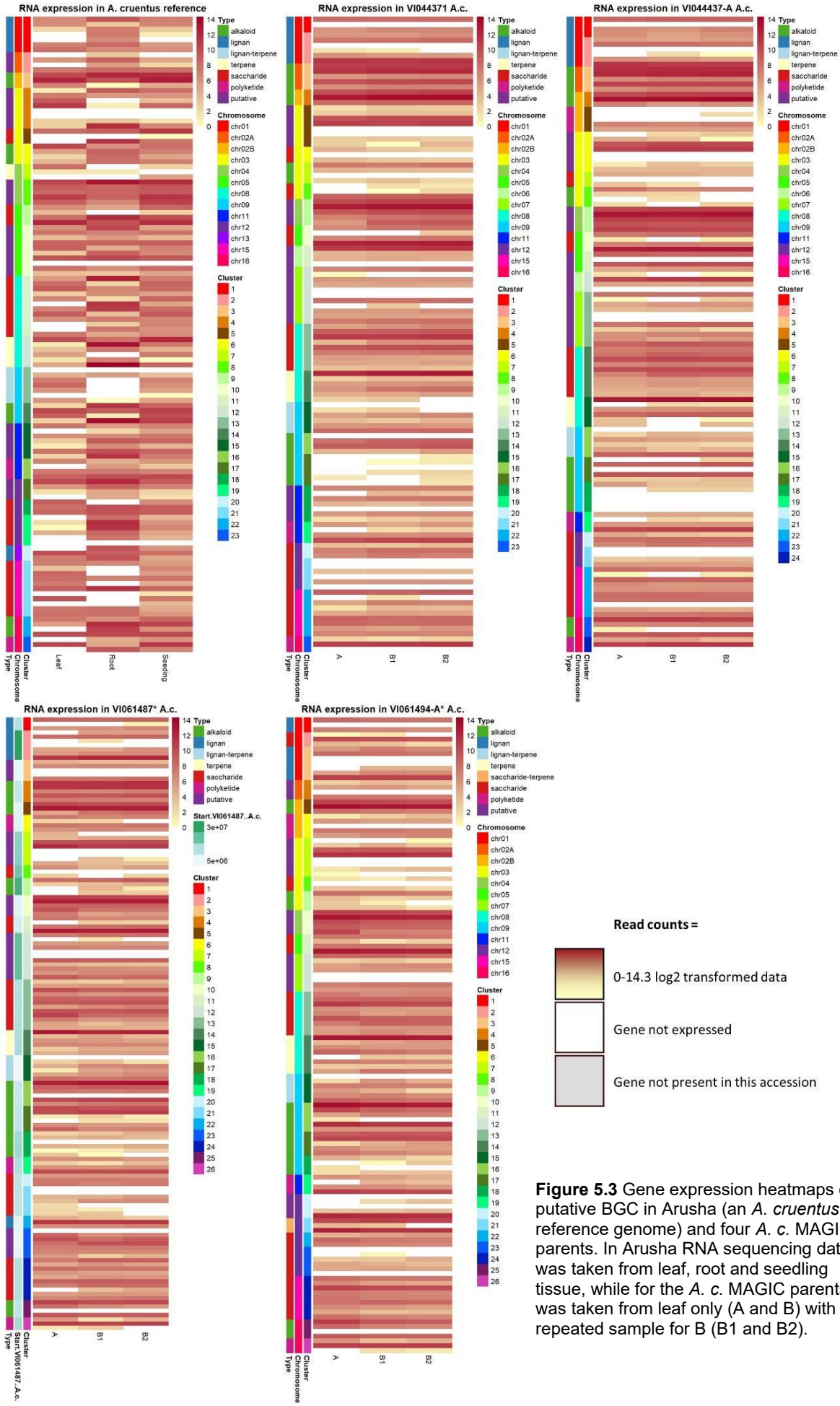



**Figure 5.3** Gene expression heatmaps of putative BGC in Arusha (an *A. cruentus* reference genome) and four *A. c.* MAGIC parents. In Arusha RNA sequencing data was taken from leaf, root and seedling tissue, while for the *A. c.* MAGIC parents it was taken from leaf only (A and B) with a repeated sample for B (B1 and B2).

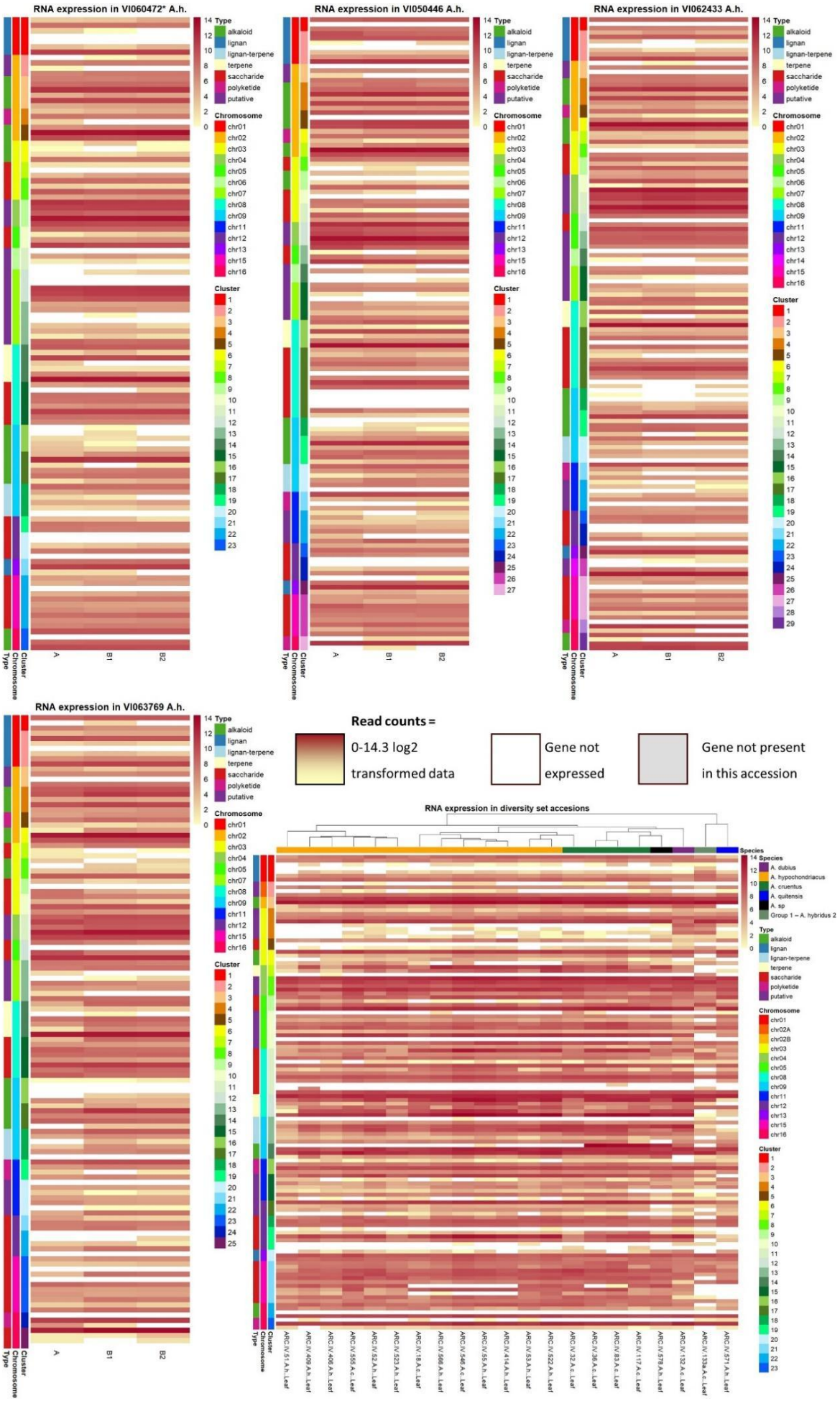



**Figure 5.4** Gene expression heatmaps of putative BGC in four *A. h.* MAGIC parents and BGC identified in Arusha with expression of those genes in 21 diversity set accessions. All RNA data was taken from leaf tissue with samples A and B for the *A. h.* MAGIC parents with a repeated sample for B (B1 and B2). Of the diversity set accessions only a single leaf tissue sample was taken and those are labelled with the assumed *A. h.* or *A. c.* species but ordered by their new species classification. Expression patterns are less consistent towards the right side of the heatmap with four non *A. h.* or *A. c.* diversity set accessions.

## Identifying putative BGC of interest based on gene expression data

To investigate if the putative BGC are real and to identify some of interest, BGC gene expression heatmaps are created for each of the 43 individual identified BGC. Those with the most homogenous expression patterns among all the genes in the BGC in various accessions have been pulled out as BGCs of interest, which are 11 BGC (**Figure 5.5**) out of the total 43 (**Figure Appendix**). In particular, clusters 5, 7, 17 and 35 show a very distinctive similarly high expression pattern for all of the genes in these BGCs across multiple accessions (**Figure 5.5**). This provides additional evidence that these putative BGC are real, since potential genes within a functional BGC would have similar expression patterns. To assess the relevance of the 11 selected BGC of interest, information about them is provided in **Table 5.2**. This reveals that there are four alkaloid BGCs likely of pharmaceutical importance, but it especially highlights nutritionally relevant clusters, such as the three saccharide BGCs containing glycosyltransferase and BAHD acyltransferase protein types, as well as the lignan BGC featuring dirigent enzymes. This suggests that accessions in which these nutritionally relevant clusters 13, 35, 37 and 39 are identified, like contain dietary relevant compounds in their leaves.

**Table 5.2** Selected putative BGC of interest based on their consistency in gene expression patterns across multiple accessions (**Figure 5.5**). The putative BGC ID, type, the number of (No) genes, protein types or core domains in enzymes encoded by the putative BGC, and No accessions the putative BGC is detected in with plantiSMASH (Kautsar *et al.*, 2017), is provided.

| ID | BGC Type | No genes | Protein types | No accessions |
|---|---|---|---|---|
| 5 | Alkaloid | 5-6 | Cytochrome 450, Copper amine oxidase, Methyltransferase | 7 |
| 7 | Alkaloid | 3 | Cytochrome 450, Copper amine oxidase, Aminotransferase | 30 (all) |
| 13 | Saccharide | 4 | Cytochrome 450, Glycosyltransferase | 4 (*A. h.* only) |
| 17 | Putative | 5 | Cytochrome 450, Epimerase, Polyprenyl synthetase | 30 (all) |
| 20 | Putative | 4 | COesterase, Dioxygenase | 1 |
| 28 | Alkaloid | 4-6 | Cytochrome 450, Pictet-Spengler enzyme (Bet v1) | 8 (MAGIC only) |
| 29 | Alkaloid | 4-5 | Cytochrome 450, Pictet-Spengler enzyme (Bet v1), Dioxygenase | 2 |
| 30 | Polyketide | 4 | Ketosynthase, Scl acyltransferase | 29 |
| 35 | Saccharide | 3 | Glycosyltransferase, BAHD acyltransferase, Epimerase | 30 (all) |
| 37 | Lignan | 3 | Methyltransferase, Dirigent enzymes | 26 |
| 39 | Saccharide | 9-11 | Cytochrome 450, Glycosyltransferase | 30 (all) |

To additionally provide evidence for the BGC of interest, their evolutionary conservation among the various accessions was addressed. Four of the 11 BGC are conserved among all 30 accessions, i.e. eight MAGIC founders, Arusha and 21 diversity set accessions, but others also in 29 and 26 accessions or in 8 and 7 accessions if not detected in Arusha and the diversity set accessions (**Table 5.2**, **Figure 5.5**). Co-localisation of orthologous genes which were identified as part of putative BGC conserved among accessions are

visualized by synteny analysis (**Figure 5.6**). Only three of the 11 BGC of interest were conserved across less than five accessions, among which the *A. hypochondriacus* specific saccharide cluster 13 (**Figure 5.1**). This shows that most (eight) of the 11 BGC of interest are very well conserved among accessions, additionally proving that these BGC are likely real. The MAGIC parent specific cluster 28, shows low co-expression of genes in the available leaf tissue samples (**Figure 5.5**). This putative BGC therefore doesn't appear to be expressed in leaves but could possibly be expressed in other tissues.

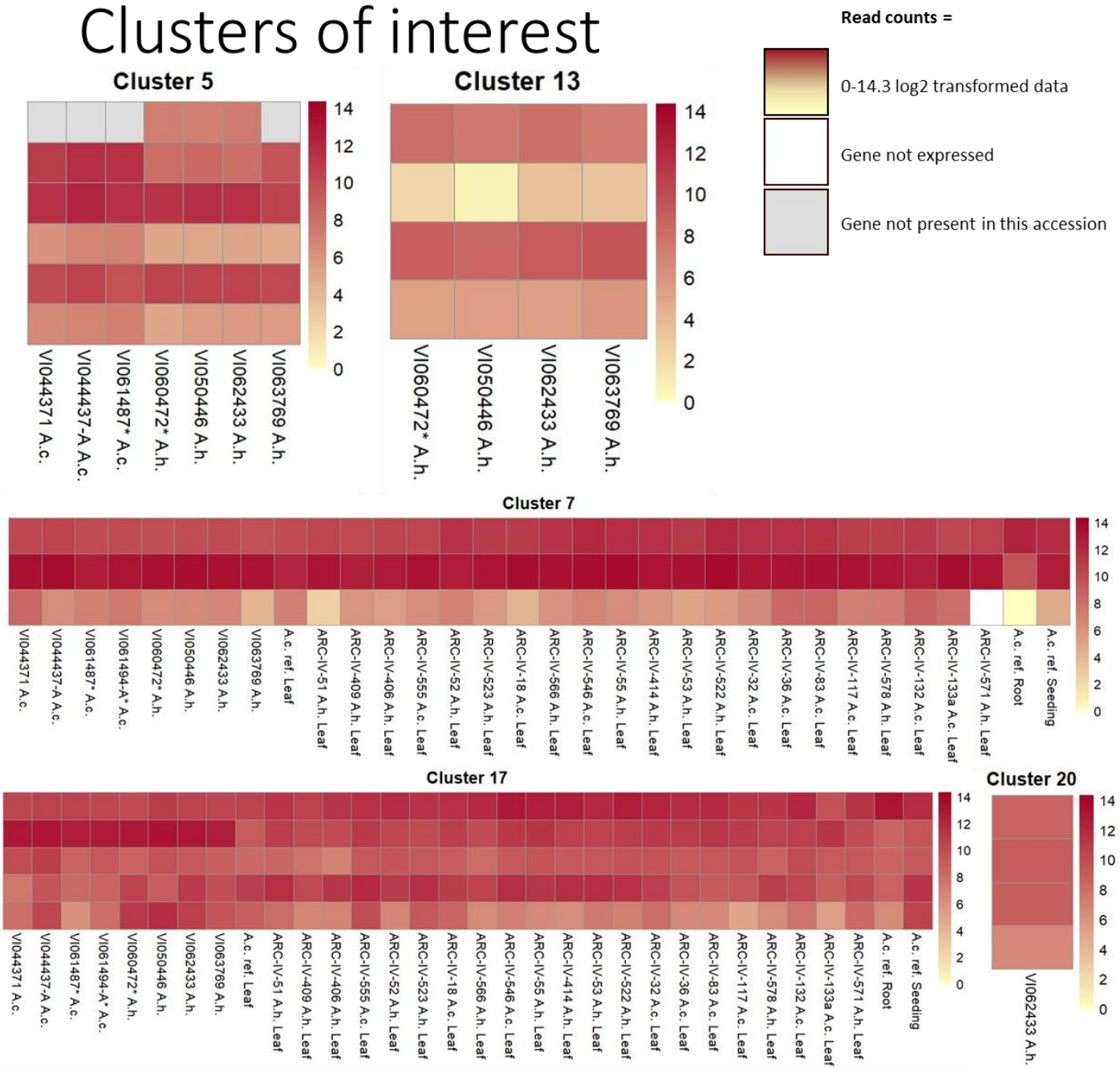



**Figure 5.5. A)** Selected BGC of interest based on their consistency in gene expression patterns across multiple accessions. These 11 BGC were selected from a total of 43 identified putative BGC (**Figure Appendix**), and for the MAGIC parent accessions the average of the three leaf tissue samples are displayed in the current heatmaps.

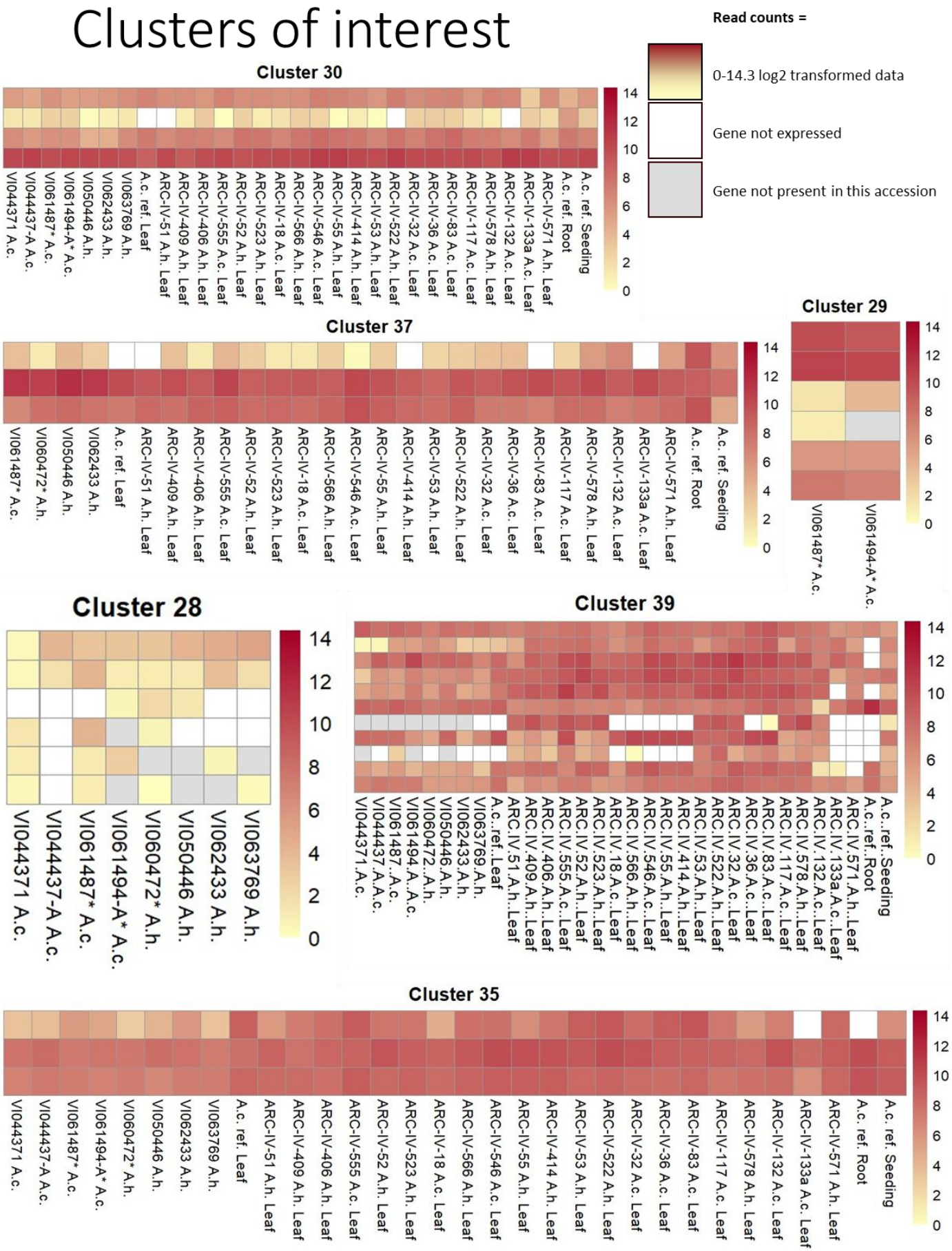



**Figure 5.5. B)** Selected BGC of interest based on their consistency in gene expression patterns across multiple accessions. These 11 BGC were selected from a total of 43 identified putative BGC (**Figure Appendix**), and for the MAGIC parent accessions the average of the three leaf tissue samples are displayed in the current heatmaps.

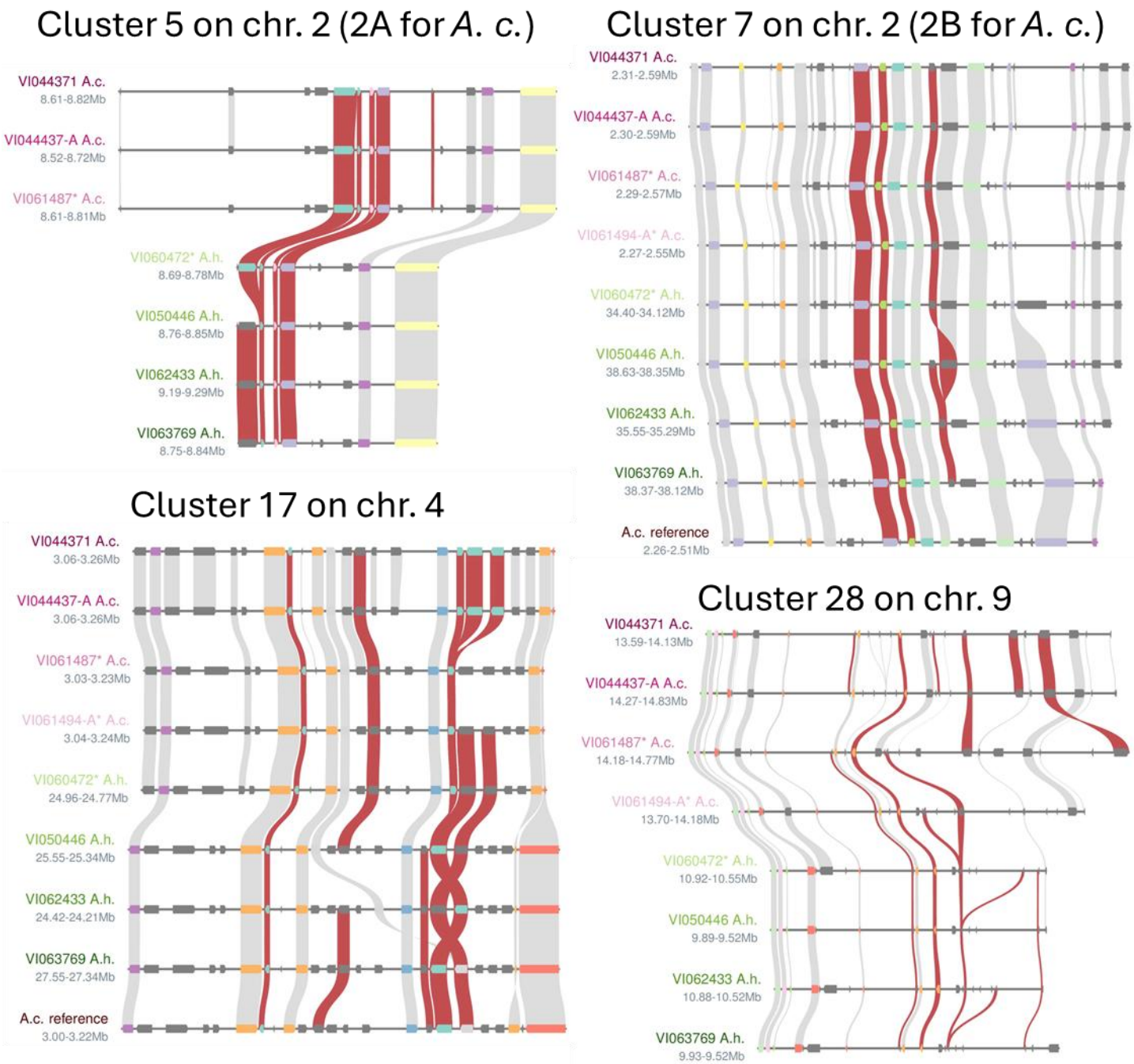



**Figure 5.6. A)** Co-localisation of genes (red) identified as putative BGC with an RNA sequencing expression profile of interest, i.e., all genes are either expressed or non-expressed simultaneously (). More information regarding the function of these putative BGC is available in **Table 5.2** and the synteny analysis of an *A. hypochondriacus* specific cluster 13 is in **Figure 5.49**.

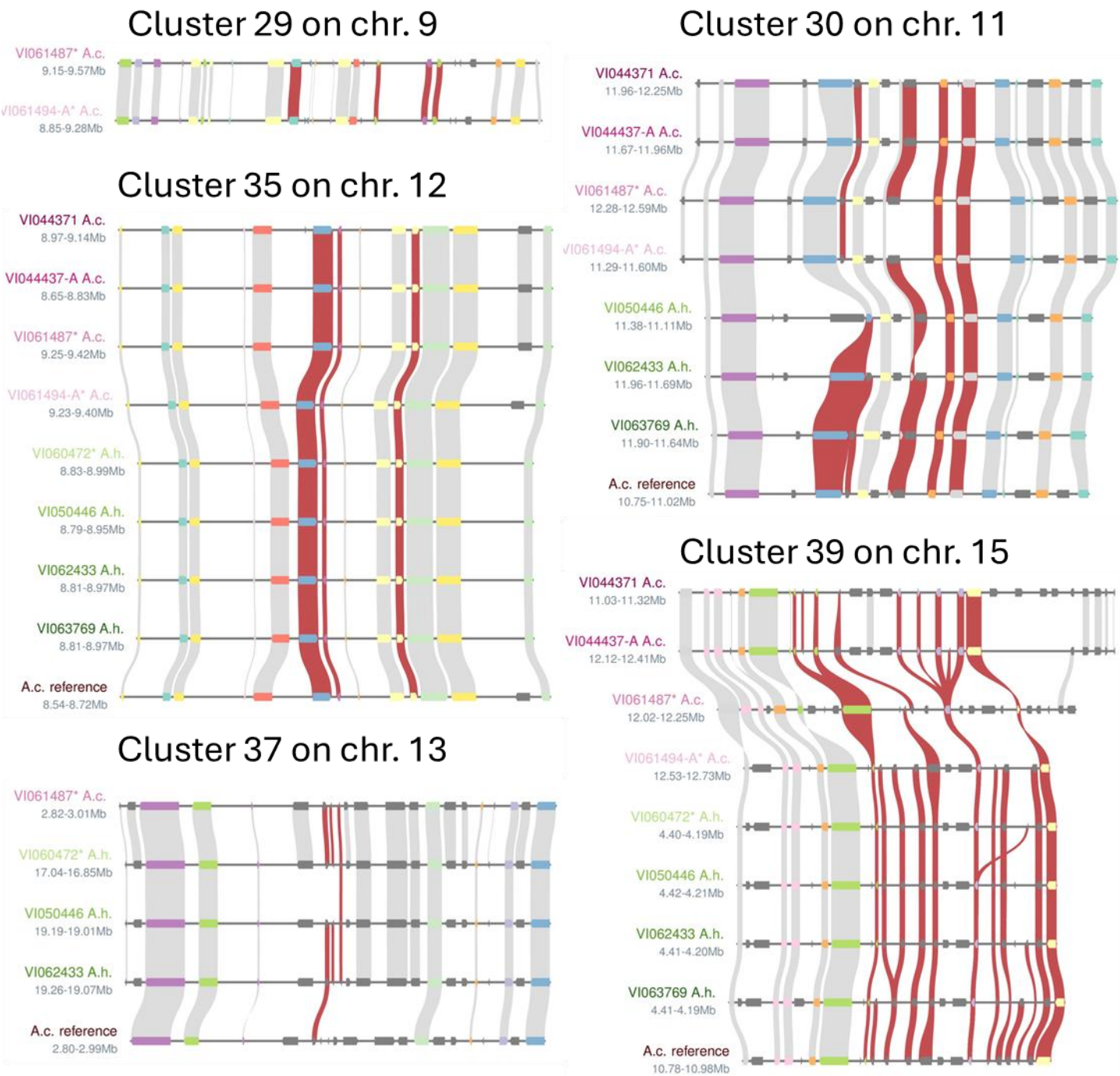



**Figure 5.6. B)** Co-localisation of genes (red) identified as putative BGC with an RNA sequencing expression profile of interest, i.e., all genes are either expressed or non-expressed simultaneously (). More information regarding the function of these putative BGC is available in **Table 5.2** and the synteny analysis of an *A. hypochondriacus* specific cluster 13 is in **Figure 5.80**.

## *DODAα1* and *CYP76ADα1* gene expression involved in betalain biosynthesis

To look into the well-known betalain BGC, its presence was investigated in all nine available high quality genomes. BLAST searches against *A. tricolor* (NCBI taxid:29722), confirmed the two co-localized *DODAα1* and *CYP76ADα1* enzyme encoding genes for tyrosine-derived betalain pigment production on chromosome 16 of the re-annotated Arusha genome. With OrthoFinder v.2.3.3 (Emms and Kelly, 2015), orthologues of these betalain genes were

identified which were also co-localised on chromosome 16 in all the MAGIC parent genomes. There were only two exceptions of a different gene located in between the two betalain involved genes for accessions VI044371 and VI061487 (**Figure 5.7**). This shows that the betalain BGC is well conserved among the Amaranth accessions, being present in all nine genomes at the same chromosome.

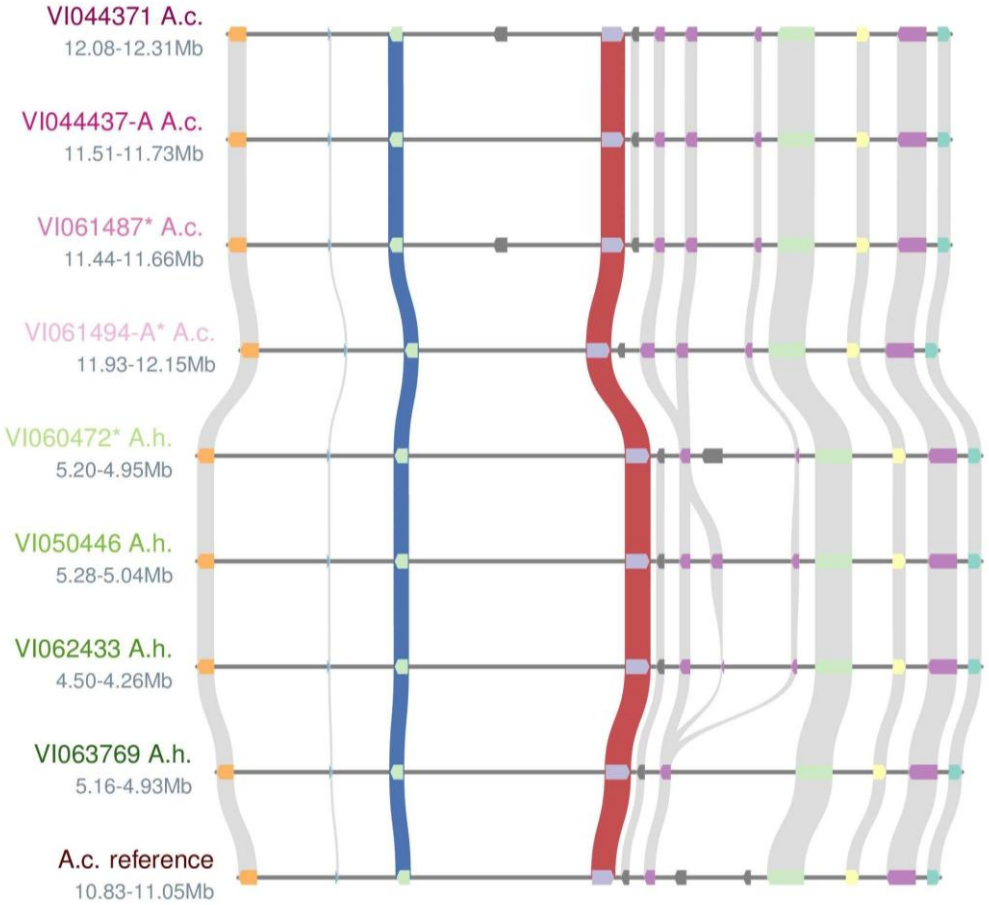



**Figure 5.7** Co-localisation of betalain biosynthesis involved genes *CYP76ADα1* (blue) and *DODAα1* (red) on chromosome 16 of the eight MAGIC parents and the *A. cruentus* reference genome.

Now to further evaluate the betalain BGC, its gene expression is examined to see if it coincides with pigmentation as to be expected. Gene expression for the two betalain biosynthetic genes was shown to be either very high or low per individual, especially among the diversity set accessions (**Figure 5.8**). This suggests that the genes were either actively expressed or silenced per accession, likely reflecting the presence or absence of betalain pigmentation. In the *A. c.* reference accession, expression is higher in root and especially seeding tissue with intermediate expression found in leaf (**Figure 5.8**). This indicates likely more betalain is produced in roots and seeding than in leaves of the Arusha accession. Variation in expression is not species specific (**Figure 5.8**), suggesting there is no difference in betalain production between *Amaranthus* species.

Leaf pigmentation was recorded, for the MAGIC founders and *A. c.* reference during the irrigated trial in Arusha, Tanzania (**Chapter 2: Supp Dataset 1**) as well as during the drought trials in South Africa (**Chapter 3: Figure 3.2**) while for the diversity set accessions it was recorded during an irrigated field trial at Roodeplaat in South Africa with three different repeats. Variation in the quantified expression data of the betalain genes can partially be

correlated with the abundance of red and/or purple versus green tissue in the plant, i.e., higher expression in observed pigmented leaf tissue (**Figure 5.8**). Some clear examples of accessions for which this is observed are ARC-IV-406, ARC-IV-555, ARC-IV-53, ARC-IV-522, ARC-IV-117. Only clear exceptions are accessions ARC-IV-51 and ARC-IV-133a which had a high betalain gene expression despite only green leaf observations and vice versa for accessions ARC-IV-414 and ARC-IV-571. This suggests that observed tissue pigmentation is mostly correlated with betalain gene expression as is expected, providing additional evidence for this well conserved BGC in amaranths. However, some clear mismatches are found between leaf pigmentation and betalain gene expressions patterns, which indicates uncertainty among observations.

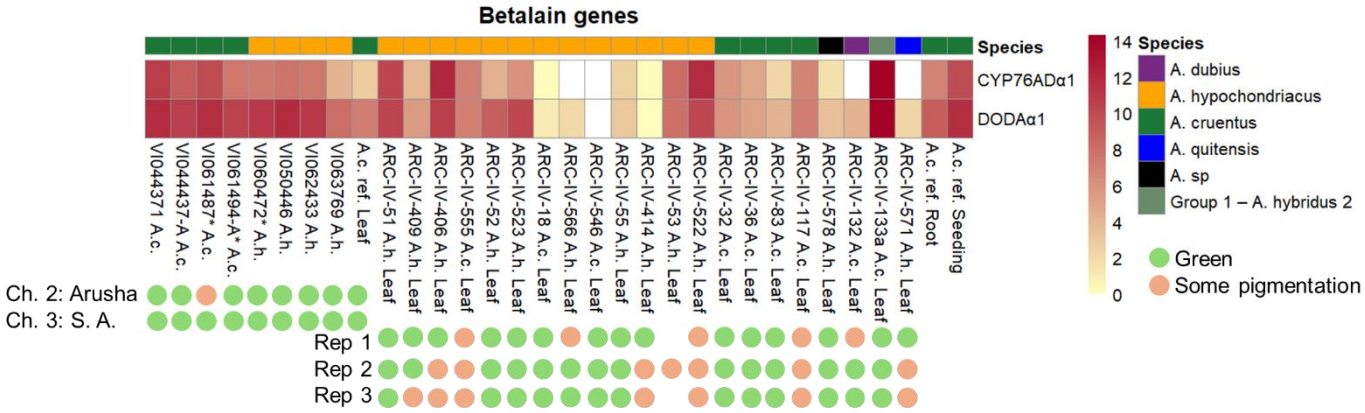



**Figure 5.8** Gene expression heatmaps of *DODAα1* and *CYP76ADα1* enzyme encoding genes known to be involved in betalain biosynthesis. The genes are located in eight MAGIC parent accessions for which the average of three leaf tissue samples are displayed, three different tissue samples were taken for Arusha and only a single leaf tissue sample for 21 diversity set accessions. Observed leaf pigmentation recorded during different field trials (MAGIC founders grown as described in **Chapter 2:** in Arusha, Tanzania and **Chapter 3:** four locations in South Africa, and diversity set accessions grown in three replicates at Roodeplaat in South Africa) is indicated. For ARC-IV-53 only one rep was recorded.

To quantify these results a multivariate regression model was run taking the expression data of each of the genes as fixed effects modelling pigmentation as a phenotypic trait. Here pigmentation was considered a binary value, either being green or red based on having more or less than 50% of the records noted down as 'Some pigmentation' in **Figure 5.8**, e.g. VI061487 is considered red while ARC-IV-409 is considered green. This regression model revealed that each of the genes had a significant ($p<0.05$) effect on leaf pigmentation, with $p=0.0247$ for *CYP76ADα1* and $p=0.0332$ for *DODAα1*. This provides statistical evidence that the measured gene expression patterns are associated with observed leaf pigmentation.

## Discussion

We can state that BGC are successfully identified in the eight novel generated genomes and the updated version of Arusha. 43 BGC have been identified (**Table 5.1**, **Supp table BGC**) of which 11 are selected as of interest due to co-expression patterns of genes (**Table 5.2**, **Figure 5.5**). About 23 to 29 BGC were found per accession (**Supp table BGC**, **Figure 5.3**, **Figure 5.4**), which is consistent with literature results of running plantiSMASH. In 47 other high quality chromosome-level assemblies of plant genomes, plantiSMASH detected about 10-50 BGC per genome, with slightly more (30–50 BGC) in dicots than monocots, and exceptions of many BGC in large genomes with amplification events like 88 BGC in *Camelina sativa*, 76 in *G. max*. and 68 in *Brassica napus* (Kautsar *et al.*, 2017). Hence, the

23 to 29 BGCs identified in our *Amaranthus* genomes may be on the low side for a dicot species. This could be due to several factors, such as the relatively small genome sizes (370.9–420.3 Mb), lower biosynthetic diversity resulting in fewer specialized metabolites produced, or the possibility that genes underlying metabolic pathways are not clustered in the genome. The number of downstream tailoring enzymes in a BGC typically ranges from 3 to 12, depending on the complexity of the metabolic pathways (Bharadwaj *et al.*, 2021). This general pattern is consistent with our results, where BGC contained between 3 and 11 genes (**Supp table BGC**), encoding for diverse enzyme types like cytochrome P450s, glycosyltransferases, methyltransferases, and others among the BGC of interest (**Table 5.2**).

Often BGC are either conserved among all or unique to a single accession, with only two being identified as either *A. cruentus* or *A. hypochondriacus* species specific (**Table 5.1**). This indicates that especially these well conserved putative BGC are more likely to be real, since BGC are expected to be co-inherited to ensure the complete pathway is passed on intact, reflecting their role as functional clusters (Smit and Lichman, 2022). This co-inheritance hypothesis is supported by that intact BGC would prevent the accumulation of toxic intermediates in plants (Nützmann, Huang and Osbourn, 2016) and that increasing plant phytotoxins would negatively select against partial inheritance of BGC (Zhang and Peters, 2020). Not many (two) BGC are species specific, with cluster 13 specific to *A. hypochondriacus* species likely of most relevance (**Figure 5.1**). This due to its promising nutritional benefits being classified as a saccharide BGC (**Table 5.2**), as well as shared co-expression of genes in this BGC (**Figure 5.5**). The latter, co-expression of genes, provides additional evidence for existence of this BGC, since often similar triggers induce co-regulation of genes underlaying metabolic pathways to be simultaneously expressed (Smit and Lichman, 2022).

Updated genome annotation by manual curation resulted in 23 (**Figure 5.3**) instead of the previously detected 22 putative BGC in *A. cruentus* (Ma *et al.*, 2021). Most of these (16) were overlapping while only six (Ma *et al.*, 2021; *A. cruentus* clusters 1, 2, 3, 6, 17 and 20) of the 22 previously published BCG were now not detected by plantiSMASH. This indicates that the updated gene annotation likely improved the quality of the previous *A. cruentus* reference (Ma *et al.*, 2021), with more BGC being identified. Most identified BGC are still being consistent with previous results, however, six previously identified BGC are lost in the new annotation. This indicates that there are limitations while attempting to detect BGC with tools like plantiSMASH. Generally computational algorithms and tools can have arbitrary effects confounding analysis and results. For example, MCScan jcvi (Tang *et al.*, 2008) has different similarity thresholds for identifying orthologous genes compared to Orthofinder (Emms and Kelly, 2015). The fact that we manually had to combine the orthologous genes in the made expression heatmaps (**Figure 5.5**) or synteny plots (**Figure 5.6**) makes our analysis and plots error prone. Especially since we based our orthologous genes in the BGC heatmaps on visual outputs generated by plantiSMASH, which gives different BGC identification results when using the re-annotated genome. Or since MCscan uses specific thresholds for detecting orthologs which vary from other identification tools.

The highlighted previously detected cluster (Ma *et al.*, 2021; Figure 6) is equivalent to cluster 25, which was also conserved across all MAGIC genomes, Arusha and the diversity set accessions (**Table 5.1**). This suggests evidence of this BGC by co-inheritance, however, the number of genes is varying much per accession (6-8 genes) and the expression pattern

does not seem too homogeneous (**Figure Appendix**). Hence the high interest in this cluster might have been unjustified, unless co-expression of this BGC occurs in other tissue than leaf. Along with cluster 25, it was one of nine BCG which were currently identified in the re-annotated Arusha genome of the ten previously detected BCG which were conserved among Arusha and *A. hypochondriacus* (Ma *et al.*, 2021; Table S6c). This pattern for high conservation of six of these nine BGC which were previously identified as conserved in five different *Amaranthaceae* species was still present. These six BGC are clusters 12, 25, 26, 30, 39 and 41, almost all being conserved in all nine currently assessed genomes with the exceptions of cluster 30 (8 genomes) and 41 (6 genomes) (**Table 5.1**, **Supp table BGC**). This provides strong evidence that these six BGC are real due to their evolutionary co-inheritance among accessions and species. Two of these six BGC have been identified as of interest due to their consistent gene co-expression (**Figure 5.5**, cluster 30 and 39), additionally supporting their existence. Three of the nine BCG detected in the newly annotated *A. cruentus* reference and also present only in the published *A. c.* and *A. h.* genomes, again showed notable patterns of conservation, suggesting potential evolutionary significance. Clusters 18 and 36 were conserved across all nine genomes, while cluster 31 appeared in only four (**Table 5.1**), highlighting retention or loss across accessions.

Only two of the identified BGC were found to be species specific and another two to be MAGIC founder specific (**Table 5.1**), which indicates little difference between species or data sets for BGC detection. However, when looking at PCA results of RNA expression, clear separations occurred between species (**Figure 5.2 A, B**, **D**), but also between datasets likely due to different transcriptomes used for RNA-seq quantification (**Figure 5.2 C**). Especially the species separation of the diversity set accessions based on leaf transcriptome data (**Figure 5.2 B**, **D**) is important further validation of correct novel assignment performed in **Chapter 4: Table 4.5**, since it matches the genome and metabolome analysis of that chapter. The diversity set accessions were all grown in the same greenhouse under the same conditions while RNA-seq was quantified against the same Arusha reference. Hence the observed species separation (**Figure 5.2 B**, **D**) is based on genetic expression variation by different regulation of the same genes. This is also visible in the clustered heatmap for BGC genes in the diversity set accessions, showing differentiating expression profiles for newly assigned species which are more distantly related to *A. cruentus* or *A. hypochondriacus* (**Figure 5.4**; RNA expression in diversity set accessions). Such gene regulatory differences between species are quite common. Like between the stress response transcriptomes of distantly related rice (*Oryza sativa*) and *Arabidopsis thaliana* species (Narsai, Castleden and Whelan, 2010), but also within the *Arabidopsis* genus, with many (2,205) genes showing allele specific expression in either leaf or floral tissue of closely related *A. thaliana* versus *A. lyrata* (He *et al.*, 2012). The clustering per sample tissue type (leaf, root and seeding in Arusha but also leaf A, B1 and B2 in the MAGIC founders) in the RNA expression PCA (**Figure 5.2 A, C**, **D**), but also visible as differentiating expression patterns in the heatmaps for all BGC genes per accession (**Figure 5.3**, **Figure 5.4**) is important evidence for good quality of RNA sequencing data. Additional validation is the clustering observed per species (**Figure 5.2 A, B**, **D**, **Figure 5.4**; RNA expression in diversity set accessions) and per transcriptome used for certain datasets (**Figure 5.2 C**).

With this good quality RNA expression data, it was successfully possible to select 11 BGC as of interest due to co-expressed BGC genes (**Figure 5.5**). Four of the 11 BGC of interest are predicted to synthesize alkaloids (**Table 5.2**), which are relevant due to their known

medicinal and pharmacological activities (Guo *et al.*, 2018; Zhan *et al.*, 2022). Only cluster 13 seemed to be *A. hypochondriacus* specific (**Figure 5.1**), but generally a high conservation of BGC of interest was observed with eight of the 11 clusters being present in >7 accessions (**Table 5.2**, **Figure 5.5**). This highlights the need for conservation of these putative BGC among amaranths suggesting potential importance of these clusters. Low numbers of non-BGC genes between co-localized BGC genes were observed in the synteny plots of orthologous BGC genes (**Figure 5.1**, **Figure 5.6**). This suggests that the 11 BGCs of interest are likely compact gene clusters, with biosynthetic genes positioned closely together on the genome, which typically facilitates coordinated gene expression in a biosynthetic pathway (Zhan *et al.*, 2022). This compact organization provides additional support for the integrity of these BGCs, in contrast to loose gene clusters, where distantly located pathway genes are more likely to be independently regulated, or modified leading to fragmented biosynthetic pathways (Zhan *et al.*, 2022).

The well-known betalain biosynthetic pathway genes, *DODAα1* and *CYP76AD1* were previously found to be co-located on chromosome 16 in *A. hypochondriacus* and *A. cruentus* species reference genomes (Ma *et al.*, 2021). This co-localization was confirmed in the newly obtained high quality genomes of accessions of these two species (**Figure 5.7**). Since only two genes are clustered, the betalain genes would not fit the definition of a BGC very well. However, their presence has been clearly conserved across species (**Figure 5.7**) even wider than amaranths in other lineages in the *Amaranthaceae* family (Ma *et al.*, 2021; Wang *et al.*, 2022). Only accessions VI044371 and VI061487 have a different gene located in between the two betalain involved genes (**Figure 5.7**). Coincidently it was also those two accessions observing the strongest expression of the two co-localized betalain genes in comparison to the other MAGIC founders, with VI063769 showing lowest expression (**Figure 5.8**). According to the MAGIC founder evaluation data from the WVC during trial in Arusha (**Chapter 2: Supp Dataset 1**), accession VI061487 has a pigmented leaf margin and vein, while the leaf pigmentation of the other seven MAGIC founder accessions was registered as normal green (**Figure 5.8**). During the drought field trials in South Africa all eight MAGIC founders are noted having green leaf pigmentation (**Figure 5.8**, **Chapter 3: Figure 3.2**). It is expected that higher expression of *DODAα1* and *CYP76AD1* genes is observed in red, purple or mixed leaf tissue due to betacyanins production (Wang *et al.*, 2016; Winkler *et al.*, 2024), which seems to be partially confirmed for the VI061487 MAGIC parent.

When assessing the diversity set accessions a pattern of strong betalain gene expression with more pigmented leaves was observed for five of the 21 accessions, i.e. ARC-IV-406, ARC-IV-555, ARC-IV-53, ARC-IV-522, ARC-IV-117 (**Figure 5.8**). However, also some clear opposite exceptions were observed for four accessions, i.e. ARC-IV-51, ARC-IV-133a, ARC-IV-414 and ARC-IV-571 (**Figure 5.8**). Hence, there is some correlation of betalain gene expression with leaf pigmentation meeting our hypothesis, although the clear mismatches between pigmentation and expressions patterns for some accessions reveal uncertainty of results. This could be due to several reasons. The observations of leaf pigmentation can obviously vary due to personal opinions as well as variation of leaf colour within a line over generations. Additionally, pigmentation may only be present in a small part of the leaf, and if RNA sequencing is then performed in a different part of the leaf or vice versa, discrepancies can occur between expression signals and observations. What should be kept in mind is that the leaf colour was not recorded for these accessions while taking the RNA sequencing samples, and hence betalain gene expression can mismatch with the observed leaf

pigmentation of other field trials. When running a multivariate regression model both betalain genes had a significant ($p<0.05$) effect on leaf pigmentation, this quantifies the observed pattens of gene expression being associated with leaf pigmentation, despite the mismatching discrepancies. Lastly, a strong differentiation between gene expression among the diversity set accessions versus less evident differentiation among the MAGIC founders is evident (**Figure 5.8**). This might be because of the MAGIC founder gene expression being displayed as an average of three leaf tissue samples opposed to a single leaf sample for the diversity set.

The betalain genes are generally co-expressed simultaneously per accession or tissue type (**Figure 5.8**). This validates the data showing that identification of the betalain genes either by BLAST searches or OrthoFinder is likely correct. Looking at the co-localized betalain gene expression for the different tissue types collected from the Arusha accession, it was observed to be higher in root and especially seeding tissue with intermediate expression in leaf (**Figure 5.8**). In an accession of the optimised version of the *A. hypochondriacus* genome, expression of these *DODAα1* and *CYP76AD1* genes was found very strongly in flower but also in root and stem tissue, however not in leaves, cotyledons and developing as well as mature seed (Winkler *et al.*, 2024). As expected, these expression profiles are largely, though not entirely, consistent, due to the different amaranth varieties investigated. This highlights how gene expression levels can differ across tissues and genotypes, illustrating tissue and variety specific regulation.

Using plantiSMASH (Kautsar *et al.*, 2017), we were able to identify putative BGC in the eight novel genomes, which were subsequently analysed for gene co-expression or co-inheritance patterns. These are several ways of obtaining experimental evidence for BGC identification. Gene cluster prediction tools predict putative BGC in newly generated genomic resources (Chavali and Rhee, 2018), while the hypotheses of co-regulation of BGC genes or BGC conservation by co-inheritance, additionally support the existence of identified BGC (Smit and Lichman, 2022). In addition to the techniques outlined above, various other methods can be used for the identification or investigation of new plant biosynthetic gene clusters. These include genome mining, bioinformatics and multi-omics approaches to identify putative BGC, as well as knock-out mutants, GWAS and QTL mapping to investigate gene-metabolite associations (Alseekh and Fernie, 2018; Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). The latter two, GWAS and QTL, typically being performed in bigger mapping populations to map genomic regions associated with metabolites (Alseekh and Fernie, 2018; Bharadwaj *et al.*, 2021; Zhan *et al.*, 2022). This does not directly provide evidence for BGC, however, these techniques are a useful first indication of putative BGC or single genes responsible for metabolite production. A good example is the attempt of GWAS with the *Amaranthus* diversity set metabolites for which closely located significant SNPs were identified as potentially part of genes which could be putative BGC (**Chapter 4: Table 4.11**). However, there are some remaining challenges when attempting metabolite detection needed for these techniques which should be kept in mind. Like low metabolome coverages as well as issues while identifying their metabolomic function (Alseekh and Fernie, 2018).

Overall, we conclude to have successfully identified 43 putative BGC in the eight novel generated genomes and the updated Arusha genome, with further evidence of existence provided by investigation of co-regulation and co-inheritance of the identified BGC genes. Although these results are promising new information and analysis of BGC in amaranths and

their nutritional relevance, we strongly pledge for further investigation to expand current knowledge. Future work could involve multi-omics approaches to generate more high quality data for genome mining or metabolite profiling. Additionally larger sample datasets could be generated and exploited to perform GWAS and QTL mapping studies for metabolome-genome associations. Despite these promising suggestions, we are aware of the known limitations, difficulties and high data requirements for metabolic biosynthetic pathway discovery including identification of the genes involved.

# Methods

## Data generation and RNA-sequencing

High quality reference genomes for the MAGIC parents were acquired, assembled and annotated as described in the manuscript in the appendix. A reference genome for *A. cruentus* called Arusha was previously published (Ma *et al.*, 2021) for which the gene annotation is updated. For all these nine accessions, plus 21 ARC diversity set accessions which were expected to be *A. cruentus* or *A. hypochondriacus* before re-annotation (**Chapter 4: Supp table diversity set**, except accession *A. h.* ARC-IV-406 for which no genome data was available), RNAseq data has been collected from plants grown in the greenhouse at the University of York. Leaf, root and seedling tissue was harvested for Arusha. For the MAGIC parents leaf tissue only was obtained, for which two different pools of leaf samples were harvested (A and B) and two independent extractions of pool B carried out (B1 and B2). For the 21 diversity set accessions a single pooled leaf tissue sample was taken. Information regarding sample collection and age of the plants is provided in the m&m of the manuscript in the appendix, section 'RNA sequencing and analysis'.

## Biosynthetic gene cluster (BGC) identification

Sequences of two co-located genes on chromosome 16, annotated as 'DODA-type' and 'Cytochrome P450' in the *A. cruentus* genome were blasted against the *A. tricolor* genome (NCBI taxid:29722) using BLASTN 2.14.1+ (Zhang *et al.*, 2000). This was done to verify that these genes are orthologues likely involved in the betalain biosynthetic pathway. The sequences matched having '4,5-dopa dioxygenase extradiol 1-like' and 'cytochrome p450 76ad1-like' features of the *A. tricolor* cultivar Red isolate AtriRed21 chromosome 16 (ASM2621246v1) and are therefore confirmed to be the key genes (*DODAα1* and *CYP76ADα1*) of betalain biosynthesis (Wang *et al.*, 2022). Additionally, the nine available *Amaranthus* genomes (eight MAGIC parents plus Arusha) have been scanned for presence of potential BGC on 10-03-2023 using plantiSMASH (Kautsar *et al.*, 2017). Clusters which were conserved among accessions were identified manually by comparison of cluster type, type of genes present and chromosomal location. Conservation of certain BGC per species group or per MAGIC parents only was noted.

## Quantification and quality control of RNA-seq data

Obtained expression data of Arusha and the MAGIC parents was prepared as described in m&m of **Chapter 3:** 'Genes and structural variants (SVs) at QTL'. For the 21 diversity set accessions, individual transcriptome data was checked for quality with FastQC v.0.11.9-Java-11 (Andrews, 2010), and then combined into a single file with MultiQC v.1.13-foss-

2021b (Ewels *et al.*, 2016). High quality scores indicated that all RNA sequences were valuable for further analysis. For the diversity set accessions the Arusha transcriptome and index were used for quantification with Salmon v.1.10.0-GCC-11.3.0 (Patro *et al.*, 2017) since no individual reference genome is available. Further data processing with 3D RNA-seq (Guo *et al.*, 2021), and log 2 transformation were also performed as described in the m&m of **Chapter 3:** 'Genes and structural variants (SVs) at QTL'.

## Plotting of expression data and synteny analysis

For data validation, principal component analysis (PCA) plots were made based on the quantified log 2 transformed RNA sequence data outputs of webtool 3D RNA-seq, using ggplot2 (Wickham, 2016) in R v.4.2.3 (R Core Team, 2023). This processed expression data of all genes was used for plotting the PCA of the diversity set accessions only. For PCA of the MAGIC parents and Arusha, but also for PCA including these nine accessions with the diversity set accessions combined, the processed expression data of core genes (identified as such by OrthoFinder v.2.3.3 (Emms and Kelly, 2015) but ignored if not present in the *A. h.* reference genome) was used. Locations of the *DODAα1* and *CYP76ADα1* genes in Arusha, their orthologues in the MAGIC parents as identified by OrthoFinder v.2.3.3 (Emms and Kelly, 2015), or genes identified as part of putative BGC by plantiSMASH, were manually extracted using R v.4.2.3 (R Core Team, 2023). Using the pheatmap (Kolde, 2019) R package gene expression was visualized. Data validation heatmaps were made of all the putative clusters per accession, which are nine three sample heatmaps of the MAGIC parents and Arusha and one heatmap for all the diversity set accessions. The latter includes a clustered dendrogram based on the expression profiles to further validate species annotations of the diversity set accessions by revealing species specific gene regulation.

43 heatmaps per individual cluster were made combining the expression patterns of genes in all accessions the BGC was identified in (**Figure Appendix**). Here orthologous genes between accessions were combined manually based on visual outputs of plantiSMASH. Putative BGC of interest were identified by similar expression of all genes in the BGC among all the accessions and simplified heatmaps of those BGC were re-made by displaying the average of the three MAGIC parent sample repeats. Lastly, microsynteny plots were made of the betalain clustered genes, and the other identified BGC of interest using MCscan jcvi (Tang *et al.*, 2008). Orthologous genes were identified with MCscan pairwise between two accessions and then these sets were manually combined to generate plots including all accessions with the BGC. A heatmap of the *DODAα1* and *CYP76ADα1* betalain biosynthesis involved genes and their MAGIC parent orthologues, was made using the average of the three MAGIC parent leaf sample repeats, Arusha and diversity set expression patterns. This betalain heatmap included leaf pigmentation data observed during different field trials for each different dataset. This was recorded for the MAGIC founders grown in Arusha, Tanzania (**Chapter 2**) or in four locations in South Africa (**Chapter 3**) and for the diversity set accessions grown in three replicates at Roodeplaat in South Africa with for accession ARC-IV-53 only one rep recorded.

The following multivariate logistic regression model was fitted to test the association between gene expression and observed leaf pigmentation status (0 = green, 1 = red):
$logit(P(Y_i = 1)) = \beta 0 + \beta 1 \cdot CYP76AD\alpha 1_i + \beta 2 \cdot DODA\alpha 1_i$

Here $Y_i$ is the observed leaf pigmentation (green or red), $CYP76ADα1_i$ and $DODAα1_i$ the fixed effects of log-transformed gene expression values of the genes. Model coefficients were estimated using maximum likelihood, and statistical significance of effects was assessed using Wald *z*-tests using R v.4.5.3 (R Core Team, 2023).

# Chapter 6: General discussion

## 6.1 An inter-specific *Amaranthus* pan-genome

In the appendix a manuscript is provided containing a novel genomic reference resource for amaranths incorporating broad genetic as well as phenotypic diversity captured in a ten accession *Amaranthus* pan-genome. This pan-genome consisted of eight novel not yet published sequences, i.e. eight MAGIC population founders, and two already available sequences, i.e. *A. cruentus* (Ma *et al.*, 2021) and *A. hypochondriacus* (Lightfoot *et al.*, 2017). Our combined analysis showed that the later genome, *A. hypochondriacus* (Lightfoot *et al.*, 2017), was of substantial lower quality than the other nine reference genomes, likely due to its earlier development. With the contribution of other scientists, the *A. hypochondriacus* reference genome has been radically improved, now containing a 98.8% gene annotation completeness BUSCO score (Graf *et al.*, 2025) compared to the previous 80.3% (Lightfoot *et al.*, 2017). Hence, next steps in development of this resource will be to incorporate the new *A. hypochondriacus* reference version (Graf *et al.*, 2025), as well as the available genome of the distantly related *A. tricolor* species (Wang *et al.*, 2022). This will eventually result in the best possible version of an inter-specific *Amaranthus* pan-genome to be used by other scientists.

This novel pan-genomic *Amaranthus* resource is highly valuable for many purposes but also has constraints and limitations. First of all, the genomic variants, SNPs as well as SVs, captured in the pan-genome, can be used as molecular genetic markers for mapping trait associations with genomic data (Edwards *et al.*, 2007; Schreiber *et al.*, 2024). Markers for relevant traits can be identified, which can be highly valuable information. Now breeders can perform MAS to integrate the relevant markers and genes in accessions to obtain improved cultivars (Collard *et al.*, 2005; Cockram and Mackay, 2018; Arrones *et al.*, 2020). However, to get to this point can be difficult and bioinformatics can have drawbacks. Building and indexing a graph-based pan-genome can be challenging (Shi *et al.*, 2023). In our case pan-genome creation took so long that newer and better amaranth genomes were created meanwhile. Once created, many suitable and quick graph aligners are available for graph-based plant pan-genomes consisting of long-read sequencing data (Shi *et al.*, 2023; Wang *et al.*, 2023). However, adequate pipelines to genotype SVs in our MAGIC offspring RIL sequences specifically and eventually map QTL with larger genomic variation were lacking and still under development as discussed in **Chapter 3**, therefore limiting our analysis.

Lastly, although graph pan-genomes can be created and rapidly be aligned to, visualization remains problematic due to multiple paths with different sequence lengths through a graph (Matthews *et al.*, 2024). Attempts to focus on either small or large variation, i.e. base- or macro-level visualization respectively, and moving between these scales remains problematic (Wang *et al.*, 2023; Matthews *et al.*, 2024). Once we mapped QTL for leaf yield in **Chapter 3**, we visualized a 1000 base pair window around the most significant SNP of each QTL (**Figure 3.14**) using Bandage (Wick *et al.*, 2015), which is typically used for macro-level pan-genome visualization (Wang *et al.*, 2023). 1000 base pairs is a very small window, needed to obtain clear visualizations, but drastically smaller than the span of the actual QTL. Hence causal variation underlying the QTL may well not be visualised in **Figure 3.14**. Additionally, Bandage does not clearly output what are SNPs or SVs versus the

reference sequence, and which genomic variant belongs to which single or multiple founder accessions. This is relevant information when interested in which genomic variants underly traits and cannot be directly observed from **Figure 3.14**.

Despite the known limitations and difficulties while creating pan-genomes of complex plant and crop species, we have a highly valuable reference resource for the amaranth breeding community. Bioinformatics, data generation and analysis have limitations, resulting in long duration, lacking pipelines, detection power or data quality constraining QTL mapping incorporating all available pan-genomic variation. With the rapid development in bioinformatics, we hope issues can be overcome. For future work we suggest that SVs available in the *Amaranthus* pan-genome should be genotyped in the MAGIC offspring and used for QTL mapping while visualization of genomic variation at QTL could be optimized if improved tools become available.

## 6.2 QTL mapping for leafy traits, their usage in breeding and prospects of the *Amaranthus* MAGIC population

In **Chapter 3** we mapped QTL for leafy traits and identified two QTL both (almost) significant for dry and fresh leaf weight (**Figure 3.10**). This is a significant achievement and very valuable information for amaranth breeders. It provides proof that our newly created segregating *Amaranthus* MAGIC population (**Figure 3.2**) is of sufficient quality to map QTL. Some beneficial SNP alleles could be identified (**Table 3.9**) which were clearly reflected in the MAGIC offspring phenotypes (**Figure 3.11**). Now we have identified which allelic variants are preferable as markers, this knowledge can be directly used for marker-assisted breeding or MAS. By conducting MAS the identified preferable markers are followed through crossing pipelines aiming to include the beneficial alleles in accessions to obtain improved cultivars (Collard *et al.*, 2005; Cockram and Mackay, 2018; Arrones *et al.*, 2020). Selection for Amaranth accessions containing these beneficial SNP alleles is likely to result in higher leaf yields, which is one of the main goals of this PhD project.

Conducting MAS or marker-assisted breeding overcomes various problems of conventional breeding since now the preferable markers can be directly selected for and their presence often drastically improve cultivars with enhanced yields and other traits (Mackay, Stone and Ayroles, 2009; Cockram and Mackay, 2018; Arrones *et al.*, 2020; Hasan *et al.*, 2021). One modern, fast and efficient way to perform molecular based breeding is by gene editing technologies, such as ZFN, TALEN or CRISPR/Cas (Sun *et al.*, 2024). Here preferred genes identified after QTL, where variants (markers) linked to causal genes are discovered, are introduced or recreated by precisely modifying native DNA sequence of plants. Gene editing is successfully performed in model plants as well as economically relevant crops and has played a revolutionary role in plant breeding (Hasan *et al.*, 2021; Sun *et al.*, 2024). Performing gene editing in amaranth lines, modifying the DNA such that the beneficial genes for leaf weight (**Table 3.9**) are present, could potentially obtain improved accessions with higher leaf yields. However, in order to do this the preferred genes must first be identified, which can be accomplished either by fine mapping to narrow the current QTL interval or by testing candidate genes within the identified QTL regions (Cockram and Mackay, 2018).

Not only SNPs but also larger SVs or epigenetic differences are responsible for heritable phenotypic diversity among individuals affecting relevant agronomic traits (Yuan *et al.*, 2021). The current thesis focuses on DNA sequence variation, but future work could investigate epigenetic regulation to determine whether loci carrying preferable variants are transcriptionally accessible. Epigenetic modifications are often induced by environmental factors but can persist across multiple generations (Herman and Sultan, 2011). Epigenetic mechanisms that can be measured include DNA methylation, histone modifications, chromatin structure, and small RNAs, all of which can potentially hinder DNA transcription even when preferable DNA sequence is present (Herman and Sultan, 2011). Measuring these parameters in our population could reveal epigenetic differences among accessions potentially influencing leaf yield observations. Epialleles can be considered new genetic variation such as alternative patterns due to DNA methylation, which can be stably inherited, selected and causal for phenotypes (Herman and Sultan, 2011). Hence exploring epialleles, their agronomic relevance, trait associations, inheritance and potential selection can be a huge addition to our current analysis.

The identified beneficial SNP alleles (**Table 3.9**) were not very well reflected in the MAGIC founders (**Table 3.10**, **Figure 3.12**). This could be due to unknown epigenetic effects, although limited data availability, mutations, sequencing errors and more causal variants could all play a role. Ideally candidate genes at QTL would be identified for relevant traits, which then could be integrated into accessions to obtain improved varieties by MAS or breeding practices such as gene editing (Mackay, Stone and Ayroles, 2009; Cockram and Mackay, 2018; Arrones *et al.*, 2020; Hasan *et al.*, 2021; Sun *et al.*, 2024). Unfortunately, identifying candidate genes or larger SVs responsible for leaf yield variation remains challenging, as information regarding annotated genes and large SVs is available only for the MAGIC founders (**Figure 3.12**, **Figure 3.13, Table 3.11**). Hence, despite the successful optimized SNP genotyping (**Figure 3.8**) enabling QTL mapping (**Figure 3.10**) and identifying beneficial SNP alleles in the offspring (**Table 3.9**, **Figure 3.11**), we would suggest additional analysis. Like running STITCH using a founder reference panel to identify inherited founder haplotype blocks in the offspring containing gene variant information of each specific parent. Additionally, SVs of the *Amaranthus* pan-genome could be genotyped in the offspring and investigated at QTL. Investigating gene and SV information in the offspring, for which more phenotypic data is available, would be a more complete way to exploit a mapping population in order to identify candidate genes or larger SVs responsible for leaf yield, beyond small SNP variants. Additionally fine mapping or testing candidate genes could be performed to reveal candidate genes in the population. Eventually, all those larger genomic variants or genes underlaying relevant traits can in their turn be used for generating improved amaranth accessions by breeding and selection.

Generating more phenotypic data with the offspring could potentially map more QTL for a large variety of agronomic traits beyond leaf yield. What could be observed after the current SNP genotyping was that the RILs were quite heterozygous, suggesting outbreeding of the offspring compared to relatively inbred parents (**Table 3.8**). Hence prior to performing future trials we would suggest more generations of selfing with the offspring lines after their final crossing generation to aim for higher homozygosity. Additionally, more intercrossing generations before selfing could be conducted to increase recombination events to cause more genetic variation which is needed for fine mapping QTL (Mackay, Stone and Ayroles, 2009; Cockram and Mackay, 2018; Arrones *et al.*, 2020). This especially enables the

mapping of complex quantitative traits in smaller regions narrowing down the number of candidate genes. The successful mapping of QTL for leaf yield as a complex trait is a big achievement (**Figure 3.10**), however, many additional traits recorded as numerical data rather than the available descriptor data (**Figure 3.2**) could be investigated, such as leaf width and leaf length which are expected to be pleiotropic with leaf weight (**Figure 3.4**, **Table 3.1**) and could hence potentially further confirm our mapped QTL (**Figure 3.10**). Other parameters such as branching habit could be investigated, since a higher number of branches from the base of the stem rather than branches up the stem makes manual leaf harvesting easier and quicker. Potentially QTL could be mapped for number of days to flowering, since a longer time to flowering lengthens the vegetative phase in which leaves can be produced.

In addition to our current drought trials during the South African summer (**Table 3.12**, **Table 3.13**), irrigated field trials could provide insights into performance of the RILs in more advanced growth conditions or climates. It is generally known that complex quantitative agronomic traits such as leaf yield are often controlled by multiple loci and are typically largely influenced by environmental factors (Arrones *et al.*, 2020). This was confirmed in our current field trial analysis with the offspring accessions performing significantly different in four climatically different locations, with fresh leaf weight GxE effects observed (**Figure 3.3 A**, **Table 3.2**). We attempted to select the best performing offspring individuals despite environmental influences (**Figure 3.5**, **Table 3.4**, **Supp table EBV offspring**), however, further exploring their growth differences over climate conditions could aid the selection of well performing RILs in specific contexts. Additional irrigated field trials could greatly contribute to this, and if more evident GxE effects are found in such trials, well performing offspring accessions under wet versus dry growth conditions can be appropriately distinguished.

## 6.3 The selection of promising individuals for farmers and onward breeding practices

Despite the lack of irrigated field trials or GxE effects for all leafy traits a first attempt to identify robust and well performing offspring RILs regardless of differing climates was successful (**Figure 3.5**, **Supp table EBV offspring**). These 51 selected newly created inbred lines (**Table 3.4**) can directly be used by farmers but also by breeders for continued breeding practices. What should be kept in mind is that the MAGIC founders often outperformed the MAGIC offspring for leaf yield (**Figure 3.3 A**). This is likely due to the founders being selected cultivars, with the beneficial allelic combinations for leaf weight potentially being broken up by recombination while generating the MAGIC offspring. This phenomenon often occurs while creating MAGIC populations, due to absence of selection for well performing offspring while founders are well selected elites (Huang *et al.*, 2015; Scott *et al.*, 2021).

This indicates that the eight MAGIC founder lines are also highly valuable accessions regarding leaf yield. Irrigated field trials in Arusha, Tanzania, showed that there was variation in growth performance between the founder accessions but also at different times of harvesting (**Chapter 2: Figure 1A**, **Supp Table 1**). Addressing the total leaf yield over harvests from these irrigated trials suggests that founder VI044371 performs significantly

better than VI050446 (**Chapter 2: Figure 1B**). Looking at the drought trials performed in South Africa, VI044371 also performs higher in comparison to the other founders for fresh leaf weight only, while VI062433 performs lower for fresh as well as dry leaf weight (**Figure 3.12**). This information regarding leaf yield performance can be used to select the best individuals of the MAGIC founders to be used for farming or breeding practices. Although, as data suggests, the climatic and environmental conditions, as well as harvest time, should be incorporated while isolating a certain founder line for usage.

The 90 accession diversity set analysed in **Chapter 4**, contains a large pool of amaranth diversity, of which some lines potentially could be selected as promising varieties. The elaborate phylogenetic and evolutionary analysis of the diversity set greatly contributes to understanding the genetic relationships between these newly sequenced accessions which was previously unexplored. It concludes that the existence of *A. hybridus* as a species can be questioned due to accessions assigned as such not clustering separately in our own (**Figure 4.5** black arrows, **Figure 4.7**, **Figure 4.10 A** green arrows, **Figure 4.18**, **Figure 4.19**) as well as in external analysis (Chan and Sun, 1997; Mallory *et al.*, 2008; Kietlinski *et al.*, 2014; Adhikary and Pratt, 2015; Stetter and Schmid, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Blair *et al.*, 2023). This helps clear up the previous uncertainty regarding *A. hybridus* among taxologists (Brenner *et al.*, 2010) and is a key finding in *Amaranthus* species demarcation. A geographic distribution of other 'hybridus' complex species, i.e. more *A. cruentus* and *A. hypochondriacus* in Central versus more *A. caudatus* and *A. quitensis* accessions in South America, is visible in our analysis (**Figure 4.13**, **Figure 4.15**). This is distribution is commonly explained by the geographic origin of amaranths, i.e., Mesoamerican species with North or Central American origin or Andean species originating from South America (Wu and Blair, 2017; Stetter, Vidal-Villarejo and Schmid, 2020; Thapa, Edwards and Blair, 2021; Blair *et al.*, 2023).

While likely all varieties at some point originated from the Americas, the 90 accession diversity set contained lines obtained from other parts of the world (**Supp table diversity set**), though recorded origins might possibly be unreliable due to historical sharing between countries. Many of the ARC accessions originate from Africa, specifically many originally unknown (*A. sp*) accessions were isolated and grown in South Africa by the ARC (**Figure 4.14 A**). This shows that these amaranth lineages can adapt and grow in African climates and are potentially of interest as vegetable varieties. Phenotypic leaf growth data is available for some of these accessions grown by the ARC. Analysing these trials would greatly aid selection of the best leafy varieties amongst the 90 accessions, based on phenotypic performance only. Additionally, since genomes are available and SNPs are called against *A. cruentus* (Ma *et al.*, 2021), presence of the favourable alleles for leaf weight (**Table 3.9**) could be investigated in the 90 accessions. However, since this only considers six small variants while likely many others at the two identified QTL are responsible for leaf yield, further investigation of candidate genes or SVs underlaying this trait is needed by using the MAGIC population. In case such information can be revealed, knowing which genomic region from which founder is beneficial, presence of these larger SVs could likely be investigated in the 90 accessions by mapping their genomes to the pan-genome developed in the manuscript in the appendix. Accessions with more beneficial sequence for leaf yield could then be selected based on genomic information of the diversity set, expected to obtain higher yields for farmers or useful for onwards breeding.

Leaf yield and enhanced growth phenotypes are important but should not be the only target to blindly select amaranth accessions for. Although rapid growth and many leaves are beneficial, the nutritional composition of amaranth leaves should not be forgotten due to the relevance it has for healthy diets (Achigan-Dako, Sogbohossou and Maundu, 2014; van Zonneveld *et al.*, 2023). Assessing the MAGIC founders, 122 metabolites could be annotated which are significantly different between the eight parents (**Chapter 2: Figure 2B**, **Supp Dataset 3**). Most of these are lipids while likely many are of relevance to human nutrition. MAGIC founder lines could be selected based on these known metabolite profiles to isolate the healthiest variety to be consumed. The detected metabolites in the MAGIC founders are expected to segregate in the MAGIC offspring, and additional leaf metabolite profiling in these accessions could ease selection of the most nutritional varieties among the RILs.

What should be considered is it that the revealed metabolite profiles are only an overview of nutritional compounds in the founders at a certain point in time, which could vary under different environmental conditions. Hence the more static and high quality genomes of the MAGIC founders were investigated to predict presence of BGC potentially underlaying production of nutritional compounds. 43 BGC could be detected in the available genomes (**Chapter 5**: **Supp table BGC**) of which 11 were successfully identified as of interest due to co-expression patterns of clustered genes (**Figure 5.5**, **Table 5.2**). This information further reveals which MAGIC founder likely produces, or is able to produce, which nutritionally relevant metabolite, which could be used as an indicator for selection.

Lastly, in the 90 accession ARC diversity set 486 metabolite features could be annotated (**Table 4.7**), among which many relevant for nutrition and of which some clustered significantly according to species (**Figure 4.20 B**, **Table 4.8**). This reveals which of the diversity set accessions have a highly nutritionally profile under certain growing conditions which might be selected to contribute to healthy diets. Despite the low number of individuals, i.e. 90, an attempt for mGWAS revealed metabolite-genome associations for 81 metabolite traits of which 26 could be annotated (**Table 4.12**). Ideally increasing the size of the diversity set would be needed to enable sufficient statistical power to perform GWAS, although, current mGWAS did detect SNP associations revealing hints to genes involved in metabolite production. Knowing which SNP markers are likely responsible for these metabolite traits could be used for identifying or selecting beneficial alleles or performing genome editing practices while breeding other accessions. This could all contribute to selecting individuals with potential for highly nutritional profiles by containing genomes responsible for dietary relevant metabolites.

## 6.4 Future outlook and continued analysis

To summarize what work needs to be done to build on everything performed in this thesis an outlook on future analysis is provided. First, the inter-specific *Amaranthus* pan-genome should be expanded by incorporating the new *A. hypochondriacus* (Graf *et al.*, 2025) and the *A. tricolor* (Wang *et al.*, 2022) reference genomes. The MAGIC offspring RILs as well as the diversity set accessions could then be aligned towards this most complete version of the pan-genome and genetic variation can be identified. Once available, a pipeline to genotype SVs in the MAGIC offspring RILs should be used to map QTL based on larger genomic

variants beyond the current QTL identification with SNPs. Additionally, running STITCH using a founder reference panel will identify inherited founder haplotype blocks containing gene information of each parent in the offspring. Accessions of the diversity set aligned to the pan-genome can then be selected for containing beneficial sequence based on presence of variants or genes associated with a certain trait. In order to do this QTL mapping should be optimized, by more generations of intercrossing to increase recombination for QTL fine mapping, just as more generations of selfing to obtain higher homozygosity. Investigation of additional phenotypic traits, such as numerical rather than descriptor data (**Figure 3.2**), or irrigated field trials exploring evident GxE effects (**Table 3.2**), could potentially reveal a broader spectrum of QTL. QTL fine mapping or testing candidate genes could then identify the exact variant or gene responsible for a larger variety of phenotypes.

Once detailed QTL information is obtained for more agronomically relevant traits, visualising this important genomic variation should be performed when better tools for pan-genome visualization are available. Now, lines can be selected and improved based on their sequence information beyond solely exploring phenotypes (**Table 3.4**, **Supp table EBV offspring**). This can be performed by conducting molecular based or marker-assisted breeding (MAS) or by gene editing technologies. On top of that the epigenome could be explored to see in the preferable genetic variants or candidate genes can be transcribed. Lastly, while looking at leaf yield and other agronomic traits, likely involving growth related phenotypes, the metabolomic profile including nutritionally relevant compounds should not be forgotten while selecting individuals. The current mGWAS did detect SNPs associated with nutritionally relevant metabolites, however, ideally the size of the diversity set should be increased for obtaining valid results. The MAGIC founders can be selected based on their recorded metabolites (**Chapter 2: Figure 2B**) and presence of putative BGC in their genomes. Although more elaborate metabolite profiling including records under different growth conditions or including the MAGIC offspring accessions could expand the current knowledge regarding metabolite prevalence in the MAGIC population as a whole.

# Concluding remarks

The generation of an inter-specific *Amaranthus* pan-genome (**Chapter 2**, manuscript in appendix) as the first one of its kind is a major achievement and will greatly contribute to the scientific community by providing a novel genomic reference resource. It shows that the MAGIC founders are genomically but also phenotypically variable, with segregation of a diverse set of traits observed in the MAGIC offspring. This indicates that the created *Amaranthus* MAGIC population is likely of sufficient quality, which was further confirmed by the ability to map QTL for leaf yield traits (**Chapter 3**). Assessing a 90 accession diversity set greatly aided species demarcation of amaranths and revealed that *A. hybridus* of the 'hybridus' species complex might not be a species at all (**Chapter 4**). Accessions belonging to the MAGIC population as well as the diversity set could be selected as relevant cultivars for varying purposes. It has been revealed which of the MAGIC population accessions are the best performing for leaf yield or contain nutritionally relevant compounds (**Chapter 2**, **Chapter 3**) with the latter also investigated in the 90 accession diversity set (**Chapter 4**). In the MAGIC founder genomes BGC of interest could be identified, which can aid prediction the biosynthesis of dietary and other relevant metabolites by these accessions (**Chapter 5**).

This combined novel generated information of leaf performance as well as metabolite profiles in accessions could aid the selection of individuals for direct usage by farmers as well as further breeding. Additional field trials are recommended to generate data regarding phenotypic variation under different environmental conditions and potentially map QTL for a broader range of traits. Identifying specific causal variants or candidate genes remains difficult, additional crosses for increased recombination as well as functional testing will be needed.

# References


Abdellaoui, A. et al. (2023) ‘15 years of GWAS discovery: Realizing the promise’, The American Journal of Human Genetics, 110(2), pp. 179–194. Available at: https://doi.org/10.1016/j.ajhg.2022.12.011.

Abdel-Moez, G. et al. (2023) ‘Phytochemical profiling of three Amaranthus species using LC-MS/MS metabolomic approach and chemometric tools’, Journal of Pharmaceutical and Biomedical Analysis, 236, p. 115722. Available at: https://doi.org/10.1016/j.jpba.2023.115722.

Abeel, T. et al. (2012) ‘GenomeView: a next-generation genome browser’, Nucleic Acids Research, 40(2), p. e12. Available at: https://doi.org/10.1093/nar/gkr995.

Achigan-Dako, E.G., Sogbohossou, O.E.D. and Maundu, P. (2014) ‘Current knowledge on Amaranthus spp.: research avenues for improved nutritional value and yield in leafy amaranths in sub-Saharan Africa’, Euphytica, 197(3), pp. 303–317. Available at: https://doi.org/10.1007/s10681-014-1081-9.

Adhikary, D., Deyholos, M.K. and Délano-Frier, J.P. (eds) (2021) The Amaranth Genome. Cham: Springer International Publishing (Compendium of Plant Genomes). Available at: https://doi.org/10.1007/978-3-030-72365-1.

Adhikary, D. and Pratt, D.B. (2015) ‘Morphologic and Taxonomic Analysis of the Weedy and Cultivated Amaranthus hybridus Species Complex’, Systematic Botany, 40(2), pp. 604–610. Available at: https://doi.org/10.1600/036364415X688376.

Aharoni, A., Goodacre, R. and Fernie, A.R. (2023) ‘Plant and microbial sciences as key drivers in the development of metabolomics research’, Proceedings of the National Academy of Sciences, 120(12), p. e2217383120. Available at: https://doi.org/10.1073/pnas.2217383120.

Alexa, A. and Rahnenfuhrer, J. (2023) ‘topGO: Enrichment Analysis for Gene Ontology’. Bioconductor version: Release (3.16). Available at: https://doi.org/10.18129/B9.bioc.topGO.

Alexander, D.H., Novembre, J. and Lange, K. (2009) ‘Fast model-based estimation of ancestry in unrelated individuals’, Genome Research, 19(9), pp. 1655–1664. Available at: https://doi.org/10.1101/gr.094052.109.

Alimi, N.A. et al. (2013) ‘Genetic and QTL analyses of yield and a set of physiological traits in pepper’, Euphytica, 190(2), pp. 181–201. Available at: https://doi.org/10.1007/s10681-012-0767-0.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) ‘Genome structural variation discovery and genotyping’, Nature reviews. Genetics, 12(5), pp. 363–376. Available at: https://doi.org/10.1038/nrg2958.

Alonge, M. et al. (2020) ‘Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato’, Cell, 182(1), pp. 145-161.e23. Available at: https://doi.org/10.1016/j.cell.2020.05.021.

Alseekh, S. et al. (2021) ‘Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices’, Nature Methods, 18(7), pp. 747–756. Available at: https://doi.org/10.1038/s41592-021-01197-1.

Alseekh, S. and Fernie, A.R. (2018) ‘Metabolomics 20 years on: what have we learned and what hurdles remain?’, The Plant Journal, 94(6), pp. 933–942. Available at: https://doi.org/10.1111/tpj.13950.

Andrews, S. (2010) ‘FastQC: a quality control tool for high throughput sequence data.’ Available at: https://cir.nii.ac.jp/crid/1370584340724053142 (Accessed: 24 February 2025).
Anuradha et al. (2023) ‘Genetic resources and breeding approaches for improvement of amaranth (Amaranthus spp.) and quinoa (Chenopodium quinoa)’, Frontiers in Nutrition, 10, p. 1129723. Available at: https://doi.org/10.3389/fnut.2023.1129723.
Araujo-León, J.A. et al. (2024) ‘HPLC-Based Metabolomic Analysis and Characterization of Amaranthus cruentus Leaf and Inflorescence Extracts for Their Antidiabetic and Antihypertensive Potential’, Molecules, 29(9), p. 2003. Available at: https://doi.org/10.3390/molecules29092003.
Arreguez, G.A., Martínez, J.G. and Ponessa, G. (2013) ‘Amaranthus hybridus L. ssp. hybridus in an archaeological site from the initial mid-Holocene in the Southern Argentinian Puna’, Quaternary International, 307, pp. 81–85. Available at: https://doi.org/10.1016/j.quaint.2013.02.035.
Arrones, A. et al. (2020) ‘The Dawn of the Age of Multi-Parent MAGIC Populations in Plant Breeding: Novel Powerful Next-Generation Resources for Genetic Analysis and Selection of Recombinant Elite Material’, Biology, 9(8), p. 229. Available at: https://doi.org/10.3390/biology9080229.
Arrones, A. et al. (2024) ‘A novel tomato interspecific (Solanum lycopersicum var. cerasiforme and Solanum pimpinellifolium) MAGIC population facilitates trait association and candidate gene discovery in untapped exotic germplasm’, Horticulture Research, 11(7), p. uhae154. Available at: https://doi.org/10.1093/hr/uhae154.
Asano, K. et al. (2011) ‘Artificial selection for a green revolution gene during japonica rice domestication’, Proceedings of the National Academy of Sciences, 108(27), pp. 11034–11039. Available at: https://doi.org/10.1073/pnas.1019490108.
Auwera, G.A.V. der and O’Connor, B.D. (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O’Reilly Media, Inc.
Azimuddin, S.M.I. and Aslam, H. (2023) ‘Pan-Genome Analysis of Multiple Strains of Mycoplasma pneumoniae’. Available at: https://doi.org/10.21203/rs.3.rs-2495118/v1.
Bakker, P.I.W. de et al. (2008) ‘Practical aspects of imputation-driven meta-analysis of genome-wide association studies’, Human Molecular Genetics, 17(R2), p. R122. Available at: https://doi.org/10.1093/hmg/ddn288.
Barco, B. and Clay, N.K. (2019) ‘Evolution of Glucosinolate Diversity via Whole-Genome Duplications, Gene Rearrangements, and Substrate Promiscuity’, Annual Review of Plant Biology, 70(1), pp. 585–604. Available at: https://doi.org/10.1146/annurev-arplant-050718-100152.
Basrowi, R.W. and Dilantika, C. (2021) ‘Optimizing iron adequacy and absorption to prevent iron deficiency anemia: The role of combination of fortified iron and vitamin C’, World Nutrition Journal, 5(S1), pp. 33–39. Available at: https://doi.org/10.25220/WNJ.V05.S1.0005.
Bates, D. et al. (2015) ‘Fitting Linear Mixed-Effects Models Using lme4’, Journal of Statistical Software, 67, pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.
Bayer, P.E. et al. (2020) ‘Plant pan-genomes are the new reference’, Nature Plants, 6(8), pp. 914–920. Available at: https://doi.org/10.1038/s41477-020-0733-0.
Begna, T. (2020) ‘The Role of Genotype by Environmental Interaction in Plant Breeding’, Journal of Natural Sciences Research [Preprint]. Available at: https://doi.org/10.7176/JNSR/11-20-02.

Behera, B. and Patnaik, S.N. (1982) ‘Genome Analysis of Amaranthus dubius Mart. ex Thell. through the Study of Amaranthus spinosus×A. dubius Hybrids’, Cytologia, 47(2), pp. 379–389. Available at: https://doi.org/10.1508/cytologia.47.379.

Behjati, S. and Tarpey, P.S. (2013) ‘What is next generation sequencing?’, Archives of Disease in Childhood. Education and Practice Edition, 98(6), pp. 236–238. Available at: https://doi.org/10.1136/archdischild-2013-304340.

Benhin, J.K. (2006) ‘Climate change and South African agriculture: impacts and adaptation options.’, CEEPA Discussion Paper No. 21 Special Series on Climate Change and Agriculture in Africa ISBN 1-920160-01-09, pp. 1–114.

Benjamini, Y. and Speed, T.P. (2012) ‘Summarizing and correcting the GC content bias in high-throughput sequencing’, Nucleic Acids Research, 40(10), p. e72. Available at: https://doi.org/10.1093/nar/gks001.

Benoit, M. et al. (2025) ‘Solanum pan-genetics reveals paralogues as contingencies in crop engineering’, Nature, 640(8057), pp. 135–145. Available at: https://doi.org/10.1038/s41586-025-08619-6.

Berger, S. et al. (2016) ‘Palmer Amaranth ( Amaranthus palmeri ) Identification and Documentation of ALS-Resistance in Argentina’, Weed Science, 64(2), pp. 312–320. Available at: https://doi.org/10.1614/WS-D-15-00125.1.

Beyene, Y. et al. (2015) ‘Genetic Gains in Grain Yield Through Genomic Selection in Eight Bi-parental Maize Populations under Drought Stress’, Crop Science, 55(1), pp. 154–163. Available at: https://doi.org/10.2135/cropsci2014.07.0460.

Bharadwaj, R. et al. (2021) ‘Plant Metabolic Gene Clusters: Evolution, Organization, and Their Applications in Synthetic Biology’, Frontiers in Plant Science, 12, p. 697318. Available at: https://doi.org/10.3389/fpls.2021.697318.

Bhattacharyya, S. et al. (2021) ‘Metabolic response to point mutations reveals principles of modulation of in vivo enzyme activity and phenotype’, Molecular Systems Biology, 17(6), p. e10200. Available at: https://doi.org/10.15252/msb.202110200.

Birky Jr., C.W. (2001) ‘The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models’, Annual Review of Genetics, 35(Volume 35, 2001), pp. 125–148. Available at: https://doi.org/10.1146/annurev.genet.35.102401.090231.

Blair, M.W. et al. (2023) ‘Differentiation of Andean and Mesoamerican accessions in a proposed core collection of grain amaranths’, Frontiers in Plant Science, 14. Available at: https://doi.org/10.3389/fpls.2023.1144681.

Bouaziz, M. et al. (2021) ‘Controlling for human population stratification in rare variant association studies’, Scientific Reports, 11(1), p. 19015. Available at: https://doi.org/10.1038/s41598-021-98370-5.

Bourque, G. et al. (2018) ‘Ten things you should know about transposable elements’, Genome Biology, 19, p. 199. Available at: https://doi.org/10.1186/s13059-018-1577-z.

Bowen, M.L. and Orvig, C. (2008) ‘99m-Technetium carbohydrate conjugates as potential agents in molecular imaging’, Chemical Communications, 0(41), pp. 5077–5091. Available at: https://doi.org/10.1039/B809365B.

Boycheva, S. et al. (2014) ‘The rise of operon-like gene clusters in plants’, Trends in Plant Science, 19(7), pp. 447–459. Available at: https://doi.org/10.1016/j.tplants.2014.01.013.

Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) ‘An expanded view of complex traits: from polygenic to omnigenic’, Cell, 169(7), pp. 1177–1186. Available at: https://doi.org/10.1016/j.cell.2017.05.038.

Brenner, D.M. et al. (2010) ‘Genetic resources and breeding of Amaranthus’, in Plant Breeding Reviews, Volume 19. John Wiley & Sons.
Brockington, S.F. et al. (2011) ‘Complex pigment evolution in the Caryophyllales’, New Phytologist, 190(4), pp. 854–864. Available at: https://doi.org/10.1111/j.1469-8137.2011.03687.x.
Campanelli, G. et al. (2019) ‘Multi-Parental Advances Generation Inter-Cross Population, to Develop Organic Tomato Genotypes by Participatory Plant Breeding’, Agronomy, 9(3), p. 119. Available at: https://doi.org/10.3390/agronomy9030119.
Cao, F. et al. (2021) ‘Genome-wide association study (GWAS) of leaf wax components of apple’, Stress Biology, 1(1), p. 13. Available at: https://doi.org/10.1007/s44154-021-00012-3.
Carvalho-Moore, P. et al. (2025) ‘A rearranged Amaranthus palmeri extrachromosomal circular DNA confers resistance to glyphosate and glufosinate’, The Plant Cell, 37(4), p. koaf069. Available at: https://doi.org/10.1093/plcell/koaf069.
Chan, K.F. and Sun, M. (1997) ‘Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of Amaranthus’:, Theoretical and Applied Genetics, 95(5–6), pp. 865–873. Available at: https://doi.org/10.1007/s001220050637.
Chaney, L. et al. (2016) ‘The complete chloroplast genome sequences for four Amaranthus species (Amaranthaceae)’, Applications in Plant Sciences, 4(9), p. 1600063. Available at: https://doi.org/10.3732/apps.1600063.
Chang, C.C. et al. (2015) ‘Second-generation PLINK: rising to the challenge of larger and richer datasets’, GigaScience, 4(1), p. 7. Available at: https://doi.org/10.1186/s13742-015-0047-8.
Chavali, A.K. and Rhee, S.Y. (2018) ‘Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites’, Briefings in Bioinformatics, 19(5), pp. 1022–1034. Available at: https://doi.org/10.1093/bib/bbx020.
ChemSpider | Search and share chemistry (no date). Available at: http://www.chemspider.com/Default.aspx (Accessed: 28 October 2022).
Chen, K. et al. (2014) ‘TIGRA: A targeted iterative graph routing assembler for breakpoint assembly’, Genome Research, 24(2), pp. 310–317. Available at: https://doi.org/10.1101/gr.162883.113.
Chen, S. et al. (2023) ‘Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis’, Nature Plants, 9(12), pp. 1986–1999. Available at: https://doi.org/10.1038/s41477-023-01565-z.
Chen, S. (2023) ‘Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp’, iMeta, 2(2), p. e107. Available at: https://doi.org/10.1002/imt2.107.
Clayton, D.G. et al. (2005) ‘Population structure, differential bias and genomic control in a large-scale, case-control association study’, Nature Genetics, 37(11), pp. 1243–1246. Available at: https://doi.org/10.1038/ng1653.
Clouse, J.W. et al. (2016) ‘The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly’, The Plant Genome, 9(1), p. plantgenome2015.07.0062. Available at: https://doi.org/10.3835/plantgenome2015.07.0062.
Cockram, J. and Mackay, I. (2018) ‘Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants’, in R.K. Varshney, M.K. Pandey, and A. Chitikineni (eds) Plant Genetics and Molecular Biology. Cham: Springer International

Publishing (Advances in Biochemical Engineering/Biotechnology), pp. 109–138. Available at: https://doi.org/10.1007/10_2017_48.
Collard, B.C.Y. et al. (2005) ‘An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts’, Euphytica, 142(1–2), pp. 169–196. Available at: https://doi.org/10.1007/s10681-005-1681-5.
Colonna Romano, N. and Fanti, L. (2022) ‘Transposable Elements: Major Players in Shaping Genomic and Evolutionary Patterns’, Cells, 11(6), p. 1048. Available at: https://doi.org/10.3390/cells11061048.
Costea, M. and DeMason, D.A. (2001) ‘Stem Morphology and Anatomy in Amaranthus L. (Amaranthaceae), Taxonomic Significance’, The Journal of the Torrey Botanical Society, 128(3), pp. 254–281. Available at: https://doi.org/10.2307/3088717.
Covarrubias-Pazaran, Giovanny. ‘Genome-Assisted Prediction of Quantitative Traits Using the R Package Sommer’. *PLOS ONE* 11, no. 6 (2016): e0156744. https://doi.org/10.1371/journal.pone.0156744.
Danecek, P. et al. (2011) ‘The variant call format and VCFtools’, Bioinformatics, 27(15), pp. 2156–2158. Available at: https://doi.org/10.1093/bioinformatics/btr330.
Danecek, P. et al. (2021) ‘Twelve years of SAMtools and BCFtools’, GigaScience, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.
Davies, R.W. et al. (2016) ‘Rapid genotype imputation from sequence without reference panels’, Nature Genetics, 48(8), pp. 965–969. Available at: https://doi.org/10.1038/ng.3594.
De Coster, W., Weissensteiner, M.H. and Sedlazeck, F.J. (2021) ‘Towards population-scale long-read sequencing’, Nature Reviews Genetics, 22(9), pp. 572–587. Available at: https://doi.org/10.1038/s41576-021-00367-3.
Délano-Frier, J.P. et al. (2011) ‘Transcriptomic analysis of grain amaranth (Amaranthus hypochondriacus) using 454 pyrosequencing: comparison with A. tuberculatus, expression profiling in stems and in response to biotic and abiotic stress’, BMC Genomics, 12(1), p. 363. Available at: https://doi.org/10.1186/1471-2164-12-363.
Della Coletta, R. et al. (2021) ‘How the pan-genome is changing crop genomics and improvement’, Genome Biology, 22(1), p. 3. Available at: https://doi.org/10.1186/s13059-020-02224-8.
Dell’Acqua, M. et al. (2015) ‘Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays’, Genome Biology, 16(1), p. 167. Available at: https://doi.org/10.1186/s13059-015-0716-z.
Dewan, M. et al. (2017) ‘GENOTYPE X ENVIRONMENT INTERACTION EFFECTS ON THE FIELD PERFORMANCE OF STEM AMARANTH (Amaranthus tricolor L.)’, Bangladesh Journal of Plant Breeding and Genetics, 30, p. 21. Available at: https://doi.org/10.3329/bjpbg.v30i1.36530.
Diaz, S. et al. (2020) ‘Genetic mapping for agronomic traits in a MAGIC population of common bean (Phaseolus vulgaris L.) under drought conditions’, BMC Genomics, 21(1), p. 799. Available at: https://doi.org/10.1186/s12864-020-07213-6.
Dinssa, F.F. et al. (2019) ‘Yield of Vegetable Amaranth in Diverse Tanzanian Production Environments’. Available at: https://doi.org/10.21273/HORTTECH04374-19.
Diouf, I. et al. (2020) ‘Genetic basis of phenotypic plasticity and genotype × environment interactions in a multi-parental tomato population’, Journal of Experimental Botany, 71(18), pp. 5365–5376. Available at: https://doi.org/10.1093/jxb/eraa265.

Dixon, R.A. and Strack, D. (2003) ‘Phytochemistry meets genome analysis, and beyond.....’, Phytochemistry, 62(6), pp. 815–816. Available at: https://doi.org/10.1016/S0031-9422(02)00712-4.
Dongré, A.R., Eng, J.K. and Yates, J.R. (1997) ‘Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins’, Trends in Biotechnology, 15(10), pp. 418–425. Available at: https://doi.org/10.1016/S0167-7799(97)01110-4.
Dührkop, K. et al. (2015) ‘Searching molecular structure databases with tandem mass spectra using CSI:FingerID’, Proceedings of the National Academy of Sciences, 112(41), pp. 12580–12585. Available at: https://doi.org/10.1073/pnas.1509788112.
Dührkop, K. et al. (2019) ‘SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information’, Nature Methods, 16(4), pp. 299–302. Available at: https://doi.org/10.1038/s41592-019-0344-8.
Ebert, A.W., Wu, T. and Wang, S. (2011) ‘Vegetable amaranth (Amaranthus L.)’, AVRDC – The World Vegetable Center, p. 9.
Edge, P. and Bansal, V. (2019) ‘Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing’, Nature Communications, 10(1), p. 4660. Available at: https://doi.org/10.1038/s41467-019-12493-y.
Edwards, D. et al. (2007) ‘What Are SNPs?’, in N.C. Oraguzie et al. (eds) Association Mapping in Plants. New York, NY: Springer, pp. 41–52. Available at: https://doi.org/10.1007/978-0-387-36011-9_3.
Edwards, G.E., Franceschi, V.R. and Voznesenskaya, E.V. (2004) ‘SINGLE-CELL C4 PHOTOSYNTHESIS VERSUS THE DUAL-CELL (KRANZ) PARADIGM’, Annual Review of Plant Biology, 55(Volume 55, 2004), pp. 173–196. Available at: https://doi.org/10.1146/annurev.arplant.55.031903.141725.
Eichler, E.E. et al. (2010) ‘Missing heritability and strategies for finding the underlying causes of complex disease’, Nature Reviews Genetics, 11(6), pp. 446–450. Available at: https://doi.org/10.1038/nrg2809.
Emmanuel, O.C. and Babalola, O.O. (2021) ‘Amaranth production and consumption in South Africa: the challenges of sustainability for food and nutrition security’, International Journal of Agricultural Sustainability, 0(0), pp. 1–12. Available at: https://doi.org/10.1080/14735903.2021.1940729.
Emms, D.M. and Kelly, S. (2015) ‘OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy’, Genome Biology, 16(1), p. 157. Available at: https://doi.org/10.1186/s13059-015-0721-2.
Emms, D.M. and Kelly, S. (2019) ‘OrthoFinder: phylogenetic orthology inference for comparative genomics’, Genome Biology, 20(1), p. 238. Available at: https://doi.org/10.1186/s13059-019-1832-y.
Escaramís, G., Docampo, E. and Rabionet, R. (2015) ‘A decade of structural variants: description, history and methods to detect structural variation’, Briefings in Functional Genomics, 14(5), pp. 305–314. Available at: https://doi.org/10.1093/bfgp/elv014.
Ewels, P. et al. (2016) ‘MultiQC: summarize analysis results for multiple tools and samples in a single report’, Bioinformatics, 32(19), pp. 3047–3048. Available at: https://doi.org/10.1093/bioinformatics/btw354.
Falconer, D.S. (1952) ‘The Problem of Environment and Selection’, The American Naturalist, 86(830), pp. 293–298. Available at: https://doi.org/10.1086/281736.
Falconer, D.S. and Mackay, T.F.C. (1996) Introduction to quantitative genetics. 4th edn. Harlow, United Kindom: Prentice Hall.

Falush, D., Stephens, M. and Pritchard, J.K. (2003) ‘Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies’, Genetics, 164(4), pp. 1567–1587. Available at: https://doi.org/10.1093/genetics/164.4.1567.
FAO (ed.) (2017a) Building resilience for food and food security. Rome: FAO (The state of food security and nutrition in the world, 2017).
FAO (ed.) (2017b) Leveraging food systems for inclusive rural transformation. Rome: Food and Agriculture Organization of the United Nations (The state of food and agriculture, 2017).
Fernie, A.R. et al. (2011) ‘Recommendations for Reporting Metabolite Data’, The Plant Cell, 23(7), pp. 2477–2482. Available at: https://doi.org/10.1105/tpc.111.086272.
Fernie, A.R. and Schauer, N. (2009) ‘Metabolomics-assisted breeding: a viable option for crop improvement?’, Trends in Genetics, 25(1), pp. 39–48. Available at: https://doi.org/10.1016/j.tig.2008.10.010.
Fernie, A.R. and Tohge, T. (2017) ‘The Genetics of Plant Metabolism’, Annual Review of Genetics, 51(1), pp. 287–310. Available at: https://doi.org/10.1146/annurev-genet-120116-024640.
Frey, M. et al. (1997) ‘Analysis of a Chemical Plant Defense Mechanism in Grasses’, Science, 277(5326), pp. 696–699. Available at: https://doi.org/10.1126/science.277.5326.696.
Fuentes, R.R. et al. (2019) ‘Structural variants in 3000 rice genomes’, Genome Research, 29(5), pp. 870–880. Available at: https://doi.org/10.1101/gr.241240.118.
Gardner, K.A., Wittern, L.M. and Mackay, I.J. (2016) ‘A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments’, Plant Biotechnology Journal, 14(6), pp. 1406–1417. Available at: https://doi.org/10.1111/pbi.12504.
Garrison, E. and Marth, G. (2012) ‘Haplotype-based variant detection from short-read sequencing’. arXiv. Available at: https://doi.org/10.48550/arXiv.1207.3907.
Geldermann, H. (1975) ‘Investigations on inheritance of quantitative characters in animals by gene markers I. Methods’, Theoretical and Applied Genetics, 46(7), pp. 319–330. Available at: https://doi.org/10.1007/BF00281673.
Gericke, O. et al. (2021) ‘Navigating through chemical space and evolutionary time across the Australian continent in plant genus Eremophila’, The Plant Journal, 108(2), pp. 555–578. Available at: https://doi.org/10.1111/tpj.15448.
Gerrano, A.S., van Rensburg, W.S.J. and Adebola, P.O. (2015) ‘Genetic diversity of Amaranthus species in South Africa’, South African Journal of Plant and Soil, 32(1), pp. 39–46. Available at: https://doi.org/10.1080/02571862.2014.973069.
Gerwig, G.J. (2021) ‘The World of Carbohydrates’, in G.J. Gerwig (ed.) The Art of Carbohydrate Analysis. Cham: Springer International Publishing (Techniques in Life Science and Biomedicine for the Non-Expert), pp. 1–10. Available at: https://doi.org/10.1007/978-3-030-77791-3_1.
Gibson, G. (2010) ‘Hints of hidden heritability in GWAS’, Nature Genetics, 42(7), pp. 558–560. Available at: https://doi.org/10.1038/ng0710-558.
Gleadow, R.M. and Møller, B.L. (2014) ‘Cyanogenic Glycosides: Synthesis, Physiology, and Phenotypic Plasticity’, Annual Review of Plant Biology, 65(1), pp. 155–185. Available at: https://doi.org/10.1146/annurev-arplant-050213-040027.
Glick, L. and Mayrose, I. (2023) ‘The Effect of Methodological Considerations on the Construction of Gene-Based Plant Pan-genomes’, Genome Biology and Evolution, 15(7), p. evad121. Available at: https://doi.org/10.1093/gbe/evad121.

Golicz, A.A., Batley, J. and Edwards, D. (2016) ‘Towards plant pangenomics’, Plant Biotechnology Journal, 14(4), pp. 1099–1105. Available at: https://doi.org/10.1111/pbi.12499.
Gonçalves-Dias, J. and Stetter, M.G. (2021) ‘PopAmaranth: a population genetic genome browser for grain amaranths and their wild relatives’, G3 Genes|Genomes|Genetics, 11(7), p. jkab103. Available at: https://doi.org/10.1093/g3journal/jkab103.
Graf, C. et al. (2025) ‘Domestication shaped the chromatin landscape of grain amaranth’. bioRxiv, p. 2025.04.17.649361. Available at: https://doi.org/10.1101/2025.04.17.649361.
Grant, W.F. (1959) ‘Cytogenetic studies in amaranthus: ii. natural interspecific hybridization between amaranthus dubius and a. spinosus’, Canadian Journal of Botany, 37(5), pp. 1063–1070. Available at: https://doi.org/10.1139/b59-084.
Grubben, G.J.H. (ed.) (2004) Legumés. Wageningen: PROTA (Ressources végétales de l’Afrique tropicale, 2).
Grytten, I. et al. (2020) ‘Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods’, BMC Genomics, 21(1), p. 282. Available at: https://doi.org/10.1186/s12864-020-6685-y.
Guo, L. et al. (2018) ‘The opium poppy genome and morphinan production’, Science, 362(6412), pp. 343–347. Available at: https://doi.org/10.1126/science.aat4096.
Guo, W. et al. (2021) ‘3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists’, RNA Biology, 18(11), pp. 1574–1587. Available at: https://doi.org/10.1080/15476286.2020.1858253.
Hameed, A. et al. (2022) ‘Graph Pangenomes Track Genetic Variants for Crop Improvement’, International Journal of Molecular Sciences, 23(21), p. 13420. Available at: https://doi.org/10.3390/ijms232113420.
Hasan, N. et al. (2021) ‘Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes’, Journal of Genetic Engineering and Biotechnology, 19(1), p. 128. Available at: https://doi.org/10.1186/s43141-021-00231-1.
He, F. et al. (2012) ‘Genome-wide Analysis of Cis-regulatory Divergence between Species in the Arabidopsis Genus’, Molecular Biology and Evolution, 29(11), pp. 3385–3395. Available at: https://doi.org/10.1093/molbev/mss146.
Herman, J. and Sultan, S. (2011) ‘Adaptive Transgenerational Plasticity in Plants: Case Studies, Mechanisms, and Implications for Natural Populations’, Frontiers in Plant Science, 2. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2011.00102 (Accessed: 20 April 2023).
Hofer, J.M.I. and Ellis, N. (2025) ‘Mendel’s Pea Genes’, Cold Spring Harbor Perspectives in Biology, p. a041872. Available at: https://doi.org/10.1101/cshperspect.a041872.
Hong, S.-Y. et al. (2019) ‘Comparative analysis of the complete chloroplast genome sequences of three Amaranthus species’, Plant Genetic Resources, 17(3), pp. 245–254. Available at: https://doi.org/10.1017/S1479262118000485.
Hoshikawa, K. et al. (2023) ‘Genetic diversity analysis and core collection construction for Amaranthus tricolor germplasm based on genome-wide single-nucleotide polymorphisms’, Scientia Horticulturae, 307, p. 111428. Available at: https://doi.org/10.1016/j.scienta.2022.111428.

Huang, B.E. et al. (2012) ‘A multiparent advanced generation inter-cross population for genetic analysis in wheat’, Plant Biotechnology Journal, 10(7), pp. 826–839. Available at: https://doi.org/10.1111/j.1467-7652.2012.00702.x.
Huang, B.E. et al. (2015) ‘MAGIC populations in crops: current status and future prospects’, Theoretical and Applied Genetics, 128(6), pp. 999–1017. Available at: https://doi.org/10.1007/s00122-015-2506-0.
Huang, C. et al. (2018) ‘ZmCCT9 enhances maize adaptation to higher latitudes’, Proceedings of the National Academy of Sciences, 115(2), pp. E334–E341. Available at: https://doi.org/10.1073/pnas.1718058115.
Huang, X. et al. (2010) ‘Genome-wide association studies of 14 agronomic traits in rice landraces’, Nature Genetics, 42(11), pp. 961–967. Available at: https://doi.org/10.1038/ng.695.
Hurgobin, B. et al. (2018) ‘Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus’, Plant Biotechnology Journal, 16(7), pp. 1265–1274. Available at: https://doi.org/10.1111/pbi.12867.
Huynh, B.-L. et al. (2018) ‘A multi-parent advanced generation inter-cross (MAGIC) population for genetic analysis and improvement of cowpea (Vigna unguiculata L. Walp.)’, The Plant Journal, 93(6), pp. 1129–1142. Available at: https://doi.org/10.1111/tpj.13827.
Ibrahim, A.K. et al. (2020) ‘Principles and approaches of association mapping in plant breeding’, Tropical Plant Biology, 13(3), pp. 212–224. Available at: https://doi.org/10.1007/s12042-020-09261-4.
Jacob, F. et al. (1960) ‘The operon: a group of genes whose expression is co-ordinated by an operator.’, Compte Rendu de l’Academie des Sciences, 250, pp. 1727–1729.
Jamalluddin, N. et al. (2022) ‘Genetic diversity analysis and marker-trait associations in Amaranthus species’, PLOS ONE, 17(5), p. e0267752. Available at: https://doi.org/10.1371/journal.pone.0267752.
Jannink, J.-L. (2007) ‘Identifying Quantitative Trait Locus by Genetic Background Interactions in Association Studies’, Genetics, 176(1), pp. 553–561. Available at: https://doi.org/10.1534/genetics.106.062992.
Jayakodi, M. et al. (2021) ‘Building pan-genome infrastructures for crop plants and their use in association genetics’, DNA Research, 28(1), p. dsaa030. Available at: https://doi.org/10.1093/dnares/dsaa030.
Jiang, T. et al. (2020) ‘Long-read-based human genomic structural variation detection with cuteSV’, Genome Biology, 21(1), p. 189. Available at: https://doi.org/10.1186/s13059-020-02107-y.
de Jonge, N.F. et al. (2022) Good Practices and Recommendations for Using and Benchmarking Computational Metabolomics Metabolite Annotation Tools. preprint. In Review. Available at: https://doi.org/10.21203/rs.3.rs-1662223/v1.
Joshi, D.C. et al. (2018) ‘From zero to hero: the past, present and future of grain amaranth breeding’, Theoretical and Applied Genetics, 131(9), pp. 1807–1823. Available at: https://doi.org/10.1007/s00122-018-3138-y.
Kadereit, G. et al. (2003) ‘Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C4 Photosynthesis’, International Journal of Plant Sciences, 164(6), pp. 959–986. Available at: https://doi.org/10.1086/378649.
Kanaka, K.K. et al. (2023) ‘On the concepts and measures of diversity in the genomics era’, Current Plant Biology, 33, p. 100278. Available at: https://doi.org/10.1016/j.cpb.2023.100278.

Karmahapatra, S. et al. (2018) ‘Capsicodendrin from Cinnamosma fragrans Exhibits Antiproliferative and Cytotoxic Activity in Human Leukemia Cells: Modulation by Glutathione’, Journal of Natural Products, 81(3), pp. 625–629. Available at: https://doi.org/10.1021/acs.jnatprod.7b00887.

Kautsar, S.A. et al. (2017) ‘plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters’, Nucleic Acids Research, 45(W1), pp. W55–W63. Available at: https://doi.org/10.1093/nar/gkx305.

Kay, M. (2022) ‘ggdist: Visualizations of distributions and uncertainty’. Zenodo. Available at: https://doi.org/10.5281/zenodo.6862765.

Kearsey, M.J. and Farquhar, A.G.L. (1998) ‘QTL analysis in plants; where are we now?’, Heredity, 80(2), pp. 137–142. Available at: https://doi.org/10.1046/j.1365-2540.1998.00500.x.

Khan, A.W. et al. (2020) ‘Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement’, Trends in Plant Science, 25(2), pp. 148–158. Available at: https://doi.org/10.1016/j.tplants.2019.10.012.

Kietlinski, K.D. et al. (2014) ‘Relationships between the Weedy Amaranthus hybridus (Amaranthaceae) and the Grain Amaranths’, Crop Science, 54(1), pp. 220–228. Available at: https://doi.org/10.2135/cropsci2013.03.0173.

Kim, S.-M. and Reinke, R.F. (2019) ‘A novel resistance gene for bacterial blight in rice, Xa43(t) identified by GWAS, confirmed by QTL mapping using a bi-parental population’, PLOS ONE, 14(2), p. e0211775. Available at: https://doi.org/10.1371/journal.pone.0211775.

de Kinderen, M.A.J. et al. (2020) ‘Genotype by Environment interactions for egg number and egg weight of five dual-purpose chicken breeds in different zones of Oromia region in Ethiopia.’, Acta Fytotechnica et Zootechnica, 23(5), pp. 205–213. Available at: https://doi.org/10.15414/afz.2020.23.mi-fpap.205-213.

de Kinderen, M.A.J. et al. (2023) ‘Genotype by Environment Interactions (G*E) of Chickens Tested in Ethiopia Using Body Weight as a Performance Trait’, Animals, 13(19), p. 3121. Available at: https://doi.org/10.3390/ani13193121.

Kirkpatrick, B.A. (1995) ‘Interspecific relationships within the genus Amaranthus (Amaranthaceae) - ProQuest’. Available at: https://www.proquest.com/openview/f33125858fb96a414483b38ec66102e8/1?pq-origsite=gscholar&cbl=18750&diss=y (Accessed: 31 July 2024).

KNApSAcK (no date). Available at: http://www.knapsackfamily.com/KNApSAcK/ (Accessed: 28 October 2022).

Koelmel, J.P. et al. (2017) ‘Common cases of improper lipid annotation using high-resolution tandem mass spectrometry data and corresponding limitations in biological interpretation’, Biochimica et biophysica acta, 1862(8), pp. 766–770. Available at: https://doi.org/10.1016/j.bbalip.2017.02.016.

Kolde, R. (2019) ‘pheatmap: Pretty Heatmaps’. Available at: https://CRAN.R-project.org/package=pheatmap (Accessed: 24 May 2022).

Korte, A. and Farlow, A. (2013) ‘The advantages and limitations of trait analysis with GWAS: a review’, Plant Methods, 9(1), p. 29. Available at: https://doi.org/10.1186/1746-4811-9-29.

Kreiner, J.M. et al. (2019a) ‘Correction for Kreiner et al., Multiple modes of convergent adaptation in the spread of glyphosate-resistant Amaranthus tuberculatus’, Proceedings of the National Academy of Sciences, 116(46), pp. 23363–23363. Available at: https://doi.org/10.1073/pnas.1917644116.

Kreiner, J.M. et al. (2019b) ‘Multiple modes of convergent adaptation in the spread of glyphosate-resistant Amaranthus tuberculatus’, Proceedings of the National Academy of Sciences, 116(42), pp. 21076–21084. Available at: https://doi.org/10.1073/pnas.1900870116.
Kumar, P. et al. (2021) ‘Skim sequencing: an advanced NGS technology for crop improvement’, Journal of Genetics, 100(2), p. 38. Available at: https://doi.org/10.1007/s12041-021-01285-3.
Lacchini, E. and Goossens, A. (2020) ‘Combinatorial Control of Plant Specialized Metabolism: Mechanisms, Functions, and Consequences’, Annual Review of Cell and Developmental Biology, 36(1), pp. 291–313. Available at: https://doi.org/10.1146/annurev-cellbio-011620-031429.
Ladejobi, O. et al. (2016) ‘Maximizing the potential of multi-parental crop populations’, Applied & Translational Genomics, 11, pp. 9–17. Available at: https://doi.org/10.1016/j.atg.2016.10.002.
Lakshmidevamma, T.N. et al. (2022) ‘Assessment of Promising Diverse Germplasm Accessions for Stability with Respect to Yield and Its Attributing Traits in Vegetable Amaranth’.
Lanoue, K.Z. et al. (1996) ‘Phylogenetic analysis of restriction-site variation in wild and cultivated Amaranthus species (Amaranthaceae)’, Theoretical and Applied Genetics, 93(5), pp. 722–732. Available at: https://doi.org/10.1007/BF00224068.
Lehmann, J.W., Clark, R.L. and Frey, K.J. (1991) ‘Biomass Heterosis and Combining Ability in Interspecific and Intraspecific Matings of Grain Amaranths’, Crop Science, 31(5), p. cropsci1991.0011183X003100050004x. Available at: https://doi.org/10.2135/cropsci1991.0011183X003100050004x.
Letunic, I. and Bork, P. (2021) ‘Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation’, Nucleic Acids Research, 49(W1), pp. W293–W296. Available at: https://doi.org/10.1093/nar/gkab301.
Li, H. et al. (2022) ‘Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber’, Nature Communications, 13(1), p. 682. Available at: https://doi.org/10.1038/s41467-022-28362-0.
Li, H. and Durbin, R. (2010) ‘Fast and accurate long-read alignment with Burrows-Wheeler transform’, Bioinformatics (Oxford, England), 26(5), pp. 589–595. Available at: https://doi.org/10.1093/bioinformatics/btp698.
Li, J. et al. (2020) ‘Biclustering with missing data’, Information Sciences, 510, pp. 304–316. Available at: https://doi.org/10.1016/j.ins.2019.09.047.
Li, Y. et al. (2020) ‘MicroTom Metabolic Network: Rewiring Tomato Metabolic Regulatory Network throughout the Growth Cycle’, Molecular Plant, 13(8), pp. 1203–1218. Available at: https://doi.org/10.1016/j.molp.2020.06.005.
Li, Y. et al. (2021) ‘Benefiting others and self: Production of vitamins in plants’, Journal of Integrative Plant Biology, 63(1), pp. 210–227. Available at: https://doi.org/10.1111/jipb.13047.
Li, Ying-hui et al. (2014) ‘De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits’, Nature Biotechnology, 32(10), pp. 1045–1052. Available at: https://doi.org/10.1038/nbt.2979.
Lightfoot, D.J. et al. (2017) ‘Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution’, BMC Biology, 15(1), p. 74. Available at: https://doi.org/10.1186/s12915-017-0412-4.

Lin, Y.-P. et al. (2022) ‘De novo SNP calling reveals the genetic differentiation and morphological divergence in genus Amaranthus’, The Plant Genome, 15(2), p. e20206. Available at: https://doi.org/10.1002/tpg2.20206.
Lindner, M.S. et al. (2013) ‘Analyzing genome coverage profiles with applications to quality control in metagenomics’, Bioinformatics, 29(10), pp. 1260–1267. Available at: https://doi.org/10.1093/bioinformatics/btt147.
Liu, Y. et al. (2020) ‘Pan-Genome of Wild and Cultivated Soybeans’, Cell, 182(1), pp. 162-176.e13. Available at: https://doi.org/10.1016/j.cell.2020.05.023.
Liu, Z. et al. (2021) ‘Quantitative trait loci (QTL) analysis of leaf related traits in spinach (Spinacia oleracea L.)’, BMC Plant Biology, 21(1), p. 290. Available at: https://doi.org/10.1186/s12870-021-03092-5.
Luo, J. (2015) ‘Metabolite-based genome-wide association studies in plants’, Current Opinion in Plant Biology, 24, pp. 31–38. Available at: https://doi.org/10.1016/j.pbi.2015.01.006.
Ma, X. et al. (2021) ‘A chromosome-level Amaranthus cruentus genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop’, The Plant Journal, 107(2), pp. 613–628. Available at: https://doi.org/10.1111/tpj.15298.
Mabhaudhi, T. et al. (2016) ‘Opportunities for Underutilised Crops in Southern Africa’s Post–2015 Development Agenda’, Sustainability, 8(4), p. 302. Available at: https://doi.org/10.3390/su8040302.
Mabhaudhi, T. et al. (2019) ‘Mainstreaming Underutilized Indigenous and Traditional Crops into Food Systems: A South African Perspective’, Sustainability, 11(1), p. 172. Available at: https://doi.org/10.3390/su11010172.
Mackay, I. and Powell, W. (2007) ‘Methods for linkage disequilibrium mapping in crops’, Trends in Plant Science, 12(2), pp. 57–63. Available at: https://doi.org/10.1016/j.tplants.2006.12.001.
Mackay, I.J. et al. (2014) ‘An Eight-Parent Multiparent Advanced Generation Inter-Cross Population for Winter-Sown Wheat: Creation, Properties, and Validation’, G3: Genes|Genomes|Genetics, 4(9), pp. 1603–1610. Available at: https://doi.org/10.1534/g3.114.012963.
Mackay, T.F.C., Stone, E.A. and Ayroles, J.F. (2009) ‘The genetics of quantitative traits: challenges and prospects’, Nature Reviews Genetics, 10(8), pp. 565–577. Available at: https://doi.org/10.1038/nrg2612.
MacNeish, R.S. (1965) ‘The Origins of American Agriculture’, Antiquity, 39(154), pp. 87–94. Available at: https://doi.org/10.1017/S0003598X00031598.
Mallory, M.A. et al. (2008) ‘Development and Characterization of Microsatellite Markers for the Grain Amaranths’, Crop Science, 48(3), pp. 1098–1106. Available at: https://doi.org/10.2135/cropsci2007.08.0457.
Manni, M. et al. (2021) ‘BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes’, Molecular Biology and Evolution, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.
Manolio, T.A. et al. (2009) ‘Finding the missing heritability of complex diseases’, Nature, 461(7265), pp. 747–753. Available at: https://doi.org/10.1038/nature08494.
Marçais, G. et al. (2018) ‘MUMmer4: A fast and versatile genome alignment system’, PLoS computational biology, 14(1), p. e1005944. Available at: https://doi.org/10.1371/journal.pcbi.1005944.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, EMBnet.journal, 17(1), pp. 10–12. Available at: https://doi.org/10.14806/ej.17.1.200.
Matthews, C.A. et al. (2024) ‘A gentle introduction to pangenomics’, Briefings in Bioinformatics, 25(6), p. bbae588. Available at: https://doi.org/10.1093/bib/bbae588.
Maughan, P. j. et al. (2011) ‘Development, Characterization, and Linkage Mapping of Single Nucleotide Polymorphisms in the Grain Amaranths (Amaranthus sp.)’, The Plant Genome, 4(1). Available at: https://doi.org/10.3835/plantgenome2010.12.0027.
Michael, T.P. and VanBuren, R. (2020) ‘Building near-complete plant genomes’, Current Opinion in Plant Biology, 54, pp. 26–33. Available at: https://doi.org/10.1016/j.pbi.2019.12.009.
Minh, B.Q. et al. (2020) ‘IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era’, Molecular Biology and Evolution, 37(5), pp. 1530–1534. Available at: https://doi.org/10.1093/molbev/msaa015.
Mishra, B.S., Sharma, M. and Laxmi, A. (2022) ‘Role of sugar and auxin crosstalk in plant growth and development’, Physiologia Plantarum, 174(1), p. e13546. Available at: https://doi.org/10.1111/ppl.13546.
Mncwango, N.C. et al. (2021) ‘Participatory Selection of Amaranthus Genotypes in the KwaMbonambi Area, KwaZulu-Natal, South Africa’, Sustainability, 13(21), p. 11962. Available at: https://doi.org/10.3390/su132111962.
Molin, W.T., Patterson, E.L. and Saski, C.A. (2020) ‘Homogeneity among glyphosate-resistant Amaranthus palmeri in geographically distant locations’, PLOS ONE, 15(9), p. e0233813. Available at: https://doi.org/10.1371/journal.pone.0233813.
Monjotin, N. et al. (2022) ‘Clinical Evidence of the Benefits of Phytonutrients in Human Healthcare’, Nutrients, 14(9), p. 1712. Available at: https://doi.org/10.3390/nu14091712.
Montgomery, J.S. et al. (2020) ‘Draft Genomes of Amaranthus tuberculatus, Amaranthus hybridus, and Amaranthus palmeri’, Genome Biology and Evolution, 12(11), pp. 1988–1993. Available at: https://doi.org/10.1093/gbe/evaa177.
Morgante, M., De Paoli, E. and Radovic, S. (2007) ‘Transposable elements and the plant pan-genomes’, Current Opinion in Plant Biology, 10(2), pp. 149–155. Available at: https://doi.org/10.1016/j.pbi.2007.02.001.
Morin, P.A. et al. (2004) ‘SNPs in ecology, evolution and conservation’, Trends in Ecology & Evolution, 19(4), pp. 208–216. Available at: https://doi.org/10.1016/j.tree.2004.01.009.
Mosyakin, S.L. and Robertson, K.R. (1996) ‘New infrageneric taxa and combinations in Amaranthus (Amaranthaceae)’, Annales Botanici Fennici, 33(4), pp. 275–281.
Mott, R. et al. (2000) ‘A method for fine mapping quantitative trait loci in outbred animal stocks’, Proceedings of the National Academy of Sciences, 97(23), pp. 12649–12654. Available at: https://doi.org/10.1073/pnas.230304397.
Motyleva, S.M. et al. (2021) ‘DROUGHT EFFECTS ON THE PHYSIOLOGICAL AND BIOCHEMICAL PARAMETERS OF AMARANTH (C-3) AND ACTINIDIA (C-4) PLANTS’, p. 15.
Naake, T. et al. (2024) ‘Genome-wide association studies identify loci controlling specialized seed metabolites in Arabidopsis’, Plant Physiology, 194(3), pp. 1705–1721. Available at: https://doi.org/10.1093/plphys/kiad511.
Naikoo, M.I. et al. (2019) ‘Chapter 9 - Role and Regulation of Plants Phenolics in Abiotic Stress Tolerance: An Overview’, in M.I.R. Khan et al. (eds) Plant Signaling

Molecules. Woodhead Publishing, pp. 157–168. Available at: https://doi.org/10.1016/B978-0-12-816451-8.00009-5.
Narsai, R., Castleden, I. and Whelan, J. (2010) 'Common and distinct organ and stress responsive transcriptomic patterns in Oryza sativa and Arabidopsis thaliana', BMC Plant Biology, 10(1), p. 262. Available at: https://doi.org/10.1186/1471-2229-10-262.
Nattestad, M. and Schatz, M.C. (2016) 'Assemblytics: a web analytics tool for the detection of variants from an assembly', Bioinformatics, 32(19), pp. 3021–3023. Available at: https://doi.org/10.1093/bioinformatics/btw369.
Nemadodzi, L.E. and Managa, G.M. (2024) '1H NMR-Based Metabolomics Profile of Green and Red Amaranthus Grown in Open Field versus Greenhouse Cultivation System', Metabolites, 14(1), p. 21. Available at: https://doi.org/10.3390/metabo14010021.
Nielsen, R. et al. (2011) 'Genotype and SNP calling from next-generation sequencing data', Nature Reviews Genetics, 12(6), pp. 443–451. Available at: https://doi.org/10.1038/nrg2986.
Nkobole, N. and Prinsloo, G. (2021) '1H-NMR and LC-MS Based Metabolomics Analysis of Wild and Cultivated Amaranthus spp.', Molecules, 26(4), p. 795. Available at: https://doi.org/10.3390/molecules26040795.
Nützmann, H.-W., Huang, A. and Osbourn, A. (2016) 'Plant metabolic clusters – from genetics to genomics', New Phytologist, 211(3), pp. 771–789. Available at: https://doi.org/10.1111/nph.13981.
Nyonje, W.A. et al. (2022) 'Exploring the Status of Preference, Utilization Practices, and Challenges to Consumption of Amaranth in Kenya and Tanzania', Journal of Nutrition and Metabolism, 2022(1), p. 2240724. Available at: https://doi.org/10.1155/2022/2240724.
Oliver, S.G. et al. (1998) 'Systematic functional analysis of the yeast genome', Trends in Biotechnology, 16(9), pp. 373–378. Available at: https://doi.org/10.1016/S0167-7799(98)01214-1.
Oluwatayo, I.B. (2019) 'Towards assuring food security in South Africa: Smallholder farmers as drivers', AIMS Agriculture and Food, 4(2), pp. 485–500. Available at: https://doi.org/10.3934/agrfood.2019.2.485.
Ongom, P.O. and Ejeta, G. (2017) 'Mating Design and Genetic Structure of a Multi-Parent Advanced Generation Intercross (MAGIC) Population of Sorghum (Sorghum bicolor (L.) Moench)', G3: Genes|Genomes|Genetics, 8(1), pp. 331–341. Available at: https://doi.org/10.1534/g3.117.300248.
Ortiz, E.M. (2019) 'vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis.' Zenodo. Available at: https://doi.org/10.5281/zenodo.2540861.
'PacificBiosciences/pbsv' (2024). PacBio. Available at: https://github.com/PacificBiosciences/pbsv (Accessed: 13 May 2024).
Pan, C.C. et al. (2016) 'Angiostatic actions of capsicodendrin through selective inhibition of VEGFR2-mediated AKT signaling and disregulated autophagy', Oncotarget, 8(8), pp. 12675–12685. Available at: https://doi.org/10.18632/oncotarget.9307.
Paradis, E. and Schliep, K. (2019) 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', Bioinformatics, 35(3), pp. 526–528. Available at: https://doi.org/10.1093/bioinformatics/bty633.
Park, Y.-J. et al. (2012) 'The molecular basis of mutations at the Waxy locus from Amaranthus caudatus L.: evolution of the waxy phenotype in three species of grain

amaranth', Molecular Breeding, 30(1), pp. 511–520. Available at: https://doi.org/10.1007/s11032-011-9640-4.
Park, Y.-J. et al. (2014) 'Molecular characterization and genetic diversity of the starch branching enzyme (SBE) gene from Amaranthus: the evolutionary origin of grain amaranths', Molecular Breeding, 34(4), pp. 1975–1985. Available at: https://doi.org/10.1007/s11032-014-0156-6.
Pascual, L. et al. (2016) 'Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels', Plant Science, 242, pp. 120–130. Available at: https://doi.org/10.1016/j.plantsci.2015.06.017.
Patro, R. et al. (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', Nature Methods, 14(4), pp. 417–419. Available at: https://doi.org/10.1038/nmeth.4197.
Perez de Souza, L. et al. (2017) 'From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web resources for mass spectral plant metabolomics', GigaScience, 6(7), pp. 1–20. Available at: https://doi.org/10.1093/gigascience/gix037.
Perez de Souza, L. et al. (2019) 'Mass Spectrometry-Based Untargeted Plant Metabolomics', Current Protocols in Plant Biology, 4(4), p. e20100. Available at: https://doi.org/10.1002/cppb.20100.
Perez de Souza, L. et al. (2021) 'Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research', Nature Methods, 18(7), pp. 733–746. Available at: https://doi.org/10.1038/s41592-021-01116-4.
Pertea, G. and Pertea, M. (2020) 'GFF Utilities: GffRead and GffCompare'. F1000Research. Available at: https://doi.org/10.12688/f1000research.23297.1.
PhytoHub (no date). Available at: https://phytohub.eu/ (Accessed: 28 October 2022).
'Picard Toolkit' (2019). Broad Institute. Available at: https://broadinstitute.github.io/picard/ (Accessed: 16 June 2022).
Ponce, K. et al. (2020) 'Genome-Wide Association Study of Grain Size Traits in Indica Rice Multiparent Advanced Generation Intercross (MAGIC) Population', Frontiers in Plant Science, 11. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.00395 (Accessed: 20 April 2023).
Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', Bioinformatics (Oxford, England), 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.
R Core Team (2021) 'R: A language and environment for statistical computing.' Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.
Randhawa, M.S. et al. (2018) 'Identification and Validation of a Common Stem Rust Resistance Locus in Two Bi-parental Populations', Frontiers in Plant Science, 9. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01788 (Accessed: 17 October 2022).
Rangarajan, A. et al. (1998) 'Iron bioavailability from Amaranthus species: 2—Evaluation using haemoglobin repletion in anaemic rats', Journal of the Science of Food and Agriculture, 78(2), pp. 274–280. Available at: https://doi.org/10.1002/(SICI)1097-0010(199810)78:2<274::AID-JSFA115>3.0.CO;2-Z.

Rastogi, A. and Shukla, S. (2013) ‘Amaranth: A New Millennium Crop of Nutraceutical Values’, Critical Reviews in Food Science and Nutrition, 53(2), pp. 109–125. Available at: https://doi.org/10.1080/10408398.2010.517876.

Ravelombola, W., Shi, A. and Huynh, B.-L. (2021) ‘Loci discovery, network-guided approach, and genomic prediction for drought tolerance index in a multi-parent advanced generation intercross (MAGIC) cowpea population’, Horticulture Research, 8, p. 24. Available at: https://doi.org/10.1038/s41438-021-00462-w.

Riedelsheimer, C. et al. (2012) ‘Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize’, Proceedings of the National Academy of Sciences, 109(23), pp. 8872–8877. Available at: https://doi.org/10.1073/pnas.1120813109.

Riggins, C.W. et al. (2010) ‘Characterization of de novo transcriptome for waterhemp (Amaranthus tuberculatus) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes’, Pest Management Science, 66(10), pp. 1042–1052. Available at: https://doi.org/10.1002/ps.2006.

Riggins, C.W. and Mumm, R.H. (2021) ‘Amaranths’, Current Biology, 31(13), pp. R834–R835. Available at: https://doi.org/10.1016/j.cub.2021.05.058.

Rocchetti, G. et al. (2020) ‘Red beet (Beta vulgaris) and amaranth (Amaranthus sp.) microgreens: Effect of storage and in vitro gastrointestinal digestion on the untargeted metabolomic profile’, Food Chemistry, 332, p. 127415. Available at: https://doi.org/10.1016/j.foodchem.2020.127415.

Rodriguez, P. et al. (2011) ‘Characterization of the proteins fractions extracted from leaves of Amaranthus dubius (Amaranthus spp.)’, African Journal of Food Science, 5(7), pp. 417–424. Available at: https://doi.org/10.5897/AJFS.9000040.

Salas-Perez, R.A. et al. (2018) ‘RNA-Seq transcriptome analysis of Amaranthus palmeri with differential tolerance to glufosinate herbicide’, PLOS ONE, 13(4), p. e0195488. Available at: https://doi.org/10.1371/journal.pone.0195488.

Sandhu, P.K. et al. (2023) ‘Global Metabolome of Palmer Amaranth (Amaranthus palmeri) Populations Highlights the Specificity and Inducibility of Phytochemical Responses to Abiotic Stress’, Journal of Agricultural and Food Chemistry, 71(7), pp. 3518–3530. Available at: https://doi.org/10.1021/acs.jafc.2c07162.

Sannemann, W. et al. (2018) ‘Adaptive selection of founder segments and epistatic control of plant height in the MAGIC winter wheat population WM-800’, BMC Genomics, 19(1), p. 559. Available at: https://doi.org/10.1186/s12864-018-4915-3.

Sarker, U. and Oba, S. (2020) ‘Phenolic profiles and antioxidant activities in selected drought-tolerant leafy vegetable amaranth’, Scientific Reports, 10(1), p. 18287. Available at: https://doi.org/10.1038/s41598-020-71727-y.

Sasaki, T. (2005) ‘The map-based sequence of the rice genome’, Nature, 436(7052), pp. 793–800. Available at: https://doi.org/10.1038/nature03895.

Sauer, J. (1957) ‘Recent Migration and Evolution of the Dioecious Amaranths’, Evolution, 11(1), pp. 11–31. Available at: https://doi.org/10.1111/j.1558-5646.1957.tb02872.x.

Sauer, J.D. (1950) ‘The Grain Amaranths: A Survey of Their History and Classification’, Annals of the Missouri Botanical Garden, 37(4), pp. 561–632. Available at: https://doi.org/10.2307/2394403.

Sauer, J.D. (1967) ‘The Grain Amaranths and Their Relatives: A Revised Taxonomic and Geographic Survey’, Annals of the Missouri Botanical Garden, 54(2), pp. 103–137. Available at: https://doi.org/10.2307/2394998.

Schafleitner, R. et al. (2022) ‘The world vegetable center amaranthus germplasm collection: Core collection development and evaluation of agronomic and nutritional traits’, Crop Science, n/a(n/a). Available at: https://doi.org/10.1002/csc2.20715.
Schauer, N. and Fernie, A.R. (2006) ‘Plant metabolomics: towards biological function and mechanism’, Trends in Plant Science, 11(10), pp. 508–516. Available at: https://doi.org/10.1016/j.tplants.2006.08.007.
Schloerke, B. et al. (2021) ‘GGally: Extension to “ggplot2”’. Available at: https://CRAN.R-project.org/package=GGally (Accessed: 31 October 2022).
Schreiber, M. et al. (2024) ‘Plant pangenomes for crop improvement, biodiversity and evolution’, Nature Reviews Genetics, 25(8), pp. 563–577. Available at: https://doi.org/10.1038/s41576-024-00691-4.
Schultes, R.E. (1984) Amaranth: Modern Prospects for an Ancient Crop. Washington, D.C.: National Academies Press. Available at: https://doi.org/10.17226/19381.
Scott, M.F. et al. (2021) ‘Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding’, Genome Biology, 22(1), p. 137. Available at: https://doi.org/10.1186/s13059-021-02354-7.
Sedio, B.E. et al. (2018) ‘Comparative foliar metabolomics of a tropical and a temperate forest community’, Ecology, 99(12), pp. 2647–2653. Available at: https://doi.org/10.1002/ecy.2533.
Sedlazeck, F.J. et al. (2018) ‘Piercing the dark matter: bioinformatics of long-range sequencing and mapping’, Nature Reviews Genetics, 19(6), pp. 329–346. Available at: https://doi.org/10.1038/s41576-018-0003-4.
Sharma, A. et al. (2019) ‘Response of Phenylpropanoid Pathway and the Role of Polyphenols in Plants under Abiotic Stress’, Molecules, 24(13), p. 2452. Available at: https://doi.org/10.3390/molecules24132452.
Sherman, R.M. and Salzberg, S.L. (2020) ‘Pan-genomics in the human genome era’, Nature Reviews Genetics, 21(4), pp. 243–254. Available at: https://doi.org/10.1038/s41576-020-0210-7.
Shi, J. et al. (2023) ‘Plant pan-genomics and its applications’, Molecular Plant, 16(1), pp. 168–186. Available at: https://doi.org/10.1016/j.molp.2022.12.009.
Shukla, A. et al. (2018) ‘Untapped amaranth (Amaranthus spp.) genetic diversity with potential for nutritional enhancement’, Genetic Resources and Crop Evolution, 65(1), pp. 243–253. Available at: https://doi.org/10.1007/s10722-017-0526-0.
Shukla, S. et al. (2006) ‘Genotypic variability in vegetable amaranth (Amaranthus tricolor L for foliage yield and its contributing traits over successive cuttings and years’, Euphytica, 151(1), pp. 103–110. Available at: https://doi.org/10.1007/s10681-006-9134-3.
Silaiyiman, S. et al. (2025) ‘A Systematic Review of the Advances and New Insights into Copy Number Variations in Plant Genomes’, Plants, 14(9), p. 1399. Available at: https://doi.org/10.3390/plants14091399.
da Silva, R.R., Dorrestein, P.C. and Quinn, R.A. (2015) ‘Illuminating the dark matter in metabolomics’, Proceedings of the National Academy of Sciences, 112(41), pp. 12549–12550. Available at: https://doi.org/10.1073/pnas.1516878112.
Sims, D. et al. (2014) ‘Sequencing depth and coverage: key considerations in genomic analyses’, Nature Reviews Genetics, 15(2), pp. 121–132. Available at: https://doi.org/10.1038/nrg3642.

Singh, A. et al. (2023) ‘Amaranth Genomic Resource Database: an integrated database resource of Amaranth genes and genomics’, Frontiers in Plant Science, 14. Available at: https://doi.org/10.3389/fpls.2023.1203855.
Singh, D.P., Singh, A.K. and Singh, A. (2021) ‘Chapter 12 - Single seed descent method’, in D.P. Singh, A.K. Singh, and A. Singh (eds) Plant Breeding and Cultivar Development. Academic Press, pp. 255–264. Available at: https://doi.org/10.1016/B978-0-12-817563-7.00024-6.
Smit, S.J. and Lichman, B.R. (2022) ‘Plant biosynthetic gene clusters in the context of metabolic evolution’, Natural Product Reports, 39(7), pp. 1465–1482. Available at: https://doi.org/10.1039/d2np00005a.
Song, J.-M. et al. (2020) ‘Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus’, Nature Plants, 6(1), pp. 34–45. Available at: https://doi.org/10.1038/s41477-019-0577-7.
Song, K., Li, L. and Zhang, G. (2016) ‘Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology’, Scientific Reports, 6(1), p. 35736. Available at: https://doi.org/10.1038/srep35736.
Southam, L. and Zeggini, E. (2025) ‘Twenty years of genome-wide association studies’, Nature, 641(8061), pp. 47–49. Available at: https://doi.org/10.1038/d41586-025-01128-6.
Speed, D., Holmes, J. and Balding, D.J. (2020) ‘Evaluating and improving heritability models using summary statistics’, Nature Genetics, 52(4), pp. 458–462. Available at: https://doi.org/10.1038/s41588-020-0600-y.
Stetter, M.G., Müller, T. and Schmid, K.J. (2017) ‘Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (Amaranthus caudatus)’, Molecular Ecology, 26(3), pp. 871–886. Available at: https://doi.org/10.1111/mec.13974.
Stetter, M.G. and Schmid, K.J. (2017) ‘Analysis of phylogenetic relationships and genome size evolution of the Amaranthus genus using GBS indicates the ancestors of an ancient crop’, Molecular Phylogenetics and Evolution, 109, pp. 80–92. Available at: https://doi.org/10.1016/j.ympev.2016.12.029.
Stetter, M.G., Vidal-Villarejo, M. and Schmid, K.J. (2020) ‘Parallel Seed Color Adaptation during Multiple Domestication Attempts of an Ancient New World Grain’, Molecular Biology and Evolution, 37(5), pp. 1407–1419. Available at: https://doi.org/10.1093/molbev/msz304.
Stiller, A. et al. (2021) ‘From Fighting Critters to Saving Lives: Polyphenols in Plant Defense and Human Health’, International Journal of Molecular Sciences, 22(16), p. 8995. Available at: https://doi.org/10.3390/ijms22168995.
Sumner, L.W. et al. (2015) ‘Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects’, Natural Product Reports, 32(2), pp. 212–229. Available at: https://doi.org/10.1039/C4NP00072B.
Sun, L. et al. (2024) ‘Modern Plant Breeding Techniques in Crop Improvement and Genetic Diversity: From Molecular Markers and Gene Editing to Artificial Intelligence—A Critical Review’, Plants, 13(19), p. 2676. Available at: https://doi.org/10.3390/plants13192676.
Sun, M., Chen, H. and Leung, F.C. (1999) ‘Low-Cot DNA sequences for fingerprinting analysis of germplasm diversity and relationships in Amaranthus’, Theoretical and Applied Genetics, 99(3), pp. 464–472. Available at: https://doi.org/10.1007/s001220051258.

Sunil, M. et al. (2014) ‘The Draft Genome and Transcriptome of Amaranthus hypochondriacus: A C4 Dicot Producing High-Lysine Edible Pseudo-Cereal’, DNA Research, 21(6), pp. 585–602. Available at: https://doi.org/10.1093/dnares/dsu021.
Takos, A.M. et al. (2011) ‘Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway’, The Plant Journal, 68(2), pp. 273–286. Available at: https://doi.org/10.1111/j.1365-313X.2011.04685.x.
Tang, H. et al. (2008) ‘Synteny and Collinearity in Plant Genomes’, Science, 320(5875), pp. 486–488. Available at: https://doi.org/10.1126/science.1153917.
Tao, Y. et al. (2019) ‘Exploring and Exploiting Pan-genomics for Crop Improvement’, Molecular Plant, 12(2), pp. 156–169. Available at: https://doi.org/10.1016/j.molp.2018.12.016.
Tattini, L., D’Aurizio, R. and Magi, A. (2015) ‘Detection of Genomic Structural Variants from Next-Generation Sequencing Data’, Frontiers in Bioengineering and Biotechnology, 3. Available at: https://www.frontiersin.org/article/10.3389/fbioe.2015.00092 (Accessed: 23 May 2022).
Teale, W.D., Paponov, I.A. and Palme, K. (2006) ‘Auxin in action: signalling, transport and the control of plant growth and development’, Nature Reviews Molecular Cell Biology, 7(11), pp. 847–859. Available at: https://doi.org/10.1038/nrm2020.
Tettelin, H. et al. (2005) ‘Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”’, Proceedings of the National Academy of Sciences, 102(39), pp. 13950–13955. Available at: https://doi.org/10.1073/pnas.0506758102.
Teucher, Olivares and Cori (2004) ‘Enhancers of Iron Absorption: Ascorbic Acid and other Organic Acids’, International Journal for Vitamin and Nutrition Research, 74(6), pp. 403–419. Available at: https://doi.org/10.1024/0300-9831.74.6.403.
Thapa, R., Edwards, M. and Blair, M.W. (2021) ‘Relationship of Cultivated Grain Amaranth Species and Wild Relative Accessions’, Genes, 12(12), p. 1849. Available at: https://doi.org/10.3390/genes12121849.
The Arabidopsis Genome Initiative (2000) ‘Analysis of the genome sequence of the flowering plant Arabidopsis thaliana’, Nature, 408(6814), pp. 796–815. Available at: https://doi.org/10.1038/35048692.
Tiwari, K.K. et al. (2021) ‘Genome-wide microsatellites in amaranth: development, characterization, and cross-species transferability’, 3 Biotech, 11(9), p. 395. Available at: https://doi.org/10.1007/s13205-021-02930-5.
Trucco, F. et al. (2007) ‘Nonhybrid Progeny from Crosses of Dioecious Amaranths: Implications for Gene-Flow Research’, Weed Science, 55(2), pp. 119–122. Available at: https://doi.org/10.1614/WS-06-132.1.
Tucker, J.B. (1986) ‘Amaranth: The Once and Future Crop’, BioScience, 36(1), pp. 9–13. Available at: https://doi.org/10.2307/1309789.
Uffelmann, E. et al. (2021) ‘Genome-wide association studies’, Nature Reviews Methods Primers, 1(1), pp. 1–21. Available at: https://doi.org/10.1038/s43586-021-00056-9.
Valdayskikh, V.V. et al. (2019) ‘Amaranth responses to experimental soil drought’, in. PROCEEDINGS OF THE 3RD INTERNATIONAL CONFERENCE ON AUTOMOTIVE INNOVATION GREEN ENERGY VEHICLE: AIGEV 2018, Kuantan, Malaysia, p. 030023. Available at: https://doi.org/10.1063/1.5087331.
Venskutonis, P.R. and Kraujalis, P. (2013) ‘Nutritional Components of Amaranth Seeds and Vegetables: A Review on Composition, Properties, and Uses’, Comprehensive

Reviews in Food Science and Food Safety, 12(4), pp. 381–412. Available at: https://doi.org/10.1111/1541-4337.12021.
Visscher, P.M. et al. (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation', American Journal of Human Genetics, 101(1), pp. 5–22. Available at: https://doi.org/10.1016/j.ajhg.2017.06.005.
Wada, T. et al. (2017) 'Development and characterization of a strawberry MAGIC population derived from crosses with six strawberry cultivars', Breeding Science, 67(4), pp. 370–381. Available at: https://doi.org/10.1270/jsbbs.17009.
Wang, H. et al. (2022) 'Chromosome-scale Amaranthus tricolor genome provides insights into the evolution of the genus Amaranthus and the mechanism of betalain biosynthesis', DNA Research, p. dsac050. Available at: https://doi.org/10.1093/dnares/dsac050.
Wang, M. et al. (2016) 'Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking', Nature Biotechnology, 34(8), pp. 828–837. Available at: https://doi.org/10.1038/nbt.3597.
Wang, Q. et al. (2023) 'Genetic analysis and mapping of dwarf gene without yield penalty in a γ-ray-induced wheat mutant', Frontiers in Plant Science, 14, p. 1133024. Available at: https://doi.org/10.3389/fpls.2023.1133024.
Wang, S. et al. (2023) 'Graph-based pan-genomes: increased opportunities in plant genomics', Journal of Experimental Botany, 74(1), pp. 24–39. Available at: https://doi.org/10.1093/jxb/erac412.
Wang, W. et al. (2018) 'Genomic variation in 3,010 diverse accessions of Asian cultivated rice', Nature, 557(7703), pp. 43–49. Available at: https://doi.org/10.1038/s41586-018-0063-9.
Warnes, G.R. et al. (2022) 'gplots: Various R Programming Tools for Plotting Data'. Available at: https://cran.r-project.org/web/packages/gplots/index.html (Accessed: 24 July 2023).
Waselkov, K.E., Boleda, A.S. and Olsen, K.M. (2018) 'A Phylogeny of the Genus Amaranthus (Amaranthaceae) Based on Several Low-Copy Nuclear Loci and Chloroplast Regions', Systematic Botany, 43(2), pp. 439–458. Available at: https://doi.org/10.1600/036364418X697193.
Wei, W. et al. (2021) 'Metabolome-Based Genome-Wide Association Study Provides Genetic Insights Into the Natural Variation of Foxtail Millet', Frontiers in Plant Science, 12, p. 665530. Available at: https://doi.org/10.3389/fpls.2021.665530.
Weir, B.S. and Cockerham, C.C. (1984) 'Estimating F-Statistics for the Analysis of Population Structure', Evolution, 38(6), pp. 1358–1370. Available at: https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.
Wenger, A.M. et al. (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', Nature Biotechnology, 37(10), pp. 1155–1162. Available at: https://doi.org/10.1038/s41587-019-0217-9.
Whitlock, M.C. and McCauley, D.E. (1999) 'Indirect measures of gene flow and migration: FST≠1/(4Nm+1)', Heredity, 82(2), pp. 117–125. Available at: https://doi.org/10.1038/sj.hdy.6884960.
Wick, R.R. et al. (2015) 'Bandage: interactive visualization of de novo genome assemblies', Bioinformatics, 31(20), pp. 3350–3352. Available at: https://doi.org/10.1093/bioinformatics/btv383.
Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Winkler, T.S. et al. (2024) ‘Isoform-resolved genome annotation enables mapping of tissue-specific betalain regulation in amaranth’, New Phytologist, 243(3), pp. 1082–1100. Available at: https://doi.org/10.1111/nph.19736.
Wu, X. and Blair, M.W. (2017) ‘Diversity in Grain Amaranths and Relatives Distinguished by Genotyping by Sequencing (GBS)’, Frontiers in Plant Science, 8, p. 1960. Available at: https://doi.org/10.3389/fpls.2017.01960.
Xu, F. and Sun, M. (2001) ‘Comparative Analysis of Phylogenetic Relationships of Grain Amaranths and Their Wild Relatives (Amaranthus; Amaranthaceae) Using Internal Transcribed Spacer, Amplified Fragment Length Polymorphism, and Double-Primer Fluorescent Intersimple Sequence Repeat Markers’, Molecular Phylogenetics and Evolution, 21(3), pp. 372–387. Available at: https://doi.org/10.1006/mpev.2001.1016.
Xu, H. et al. (2022) ‘Genetic variation and structure of complete chloroplast genome in alien monoecious and dioecious Amaranthus weeds’, Scientific Reports, 12(1), p. 8255. Available at: https://doi.org/10.1038/s41598-022-11983-2.
Xu, X.-Y. et al. (2020) ‘The complete chloroplast genome sequence of Palmer Amaranth ( Amaranthus palmeri S. Watson)’, Mitochondrial DNA Part B, 5(2), pp. 1991–1992. Available at: https://doi.org/10.1080/23802359.2020.1756969.
Xu, X.-Y. et al. (2021) ‘The complete chloroplast genome sequence of spleen amaranth (Amaranthus dubius Mart. ex Thell., Amaranthaceae)’, Mitochondrial DNA Part B, 6(11), pp. 3267–3268. Available at: https://doi.org/10.1080/23802359.2021.1992318.
Xu, Y. et al. (2017) ‘Genetic mapping of quantitative trait loci in crops’, The Crop Journal, 5(2), pp. 175–184. Available at: https://doi.org/10.1016/j.cj.2016.06.003.
Yan, H. et al. (2023) ‘Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet’, Nature Genetics, 55(3), pp. 507–518. Available at: https://doi.org/10.1038/s41588-023-01302-4.
Yan, W. et al. (2020) ‘Development of a multiparent advanced generation intercross (MAGIC) population for genetic exploitation of complex traits in Brassica juncea: Glucosinolate content as an example’, Plant Breeding, 139(4), pp. 779–789. Available at: https://doi.org/10.1111/pbr.12820.
Yang, D. et al. (2014) ‘Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism’, Engineering in Life Sciences, 14(5), pp. 456–466. Available at: https://doi.org/10.1002/elsc.201300075.
Yang, Q. et al. (2013) ‘CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize’, Proceedings of the National Academy of Sciences, 110(42), pp. 16969–16974. Available at: https://doi.org/10.1073/pnas.1310949110.
Yildiz, G. et al. (2023) ‘Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes’, The Plant Genome, 16(2), p. e20314. Available at: https://doi.org/10.1002/tpg2.20314.
Yu, G. (2022) Data Integration, Manipulation and Visualization of Phylogenetic Trees. New York: Chapman and Hall/CRC. Available at: https://doi.org/10.1201/9781003279242.
Yu, K. et al. (2020) ‘High-density QTL mapping of leaf-related traits and chlorophyll content in three soybean RIL populations’, BMC Plant Biology, 20(1), p. 470. Available at: https://doi.org/10.1186/s12870-020-02684-x.
Yu, M., Dolios, G. and Petrick, L. (2022) ‘Reproducible untargeted metabolomics workflow for exhaustive MS2 data acquisition of MS1 features’, Journal of Cheminformatics, 14(1), p. 6. Available at: https://doi.org/10.1186/s13321-022-00586-8.

Yuan, P. et al. (2024) ‘Duplicated Copy Number Variant of the Maize 9-Lipoxygenase ZmLOX5 Improves 9,10-KODA-Mediated Resistance to Fall Armyworms’, Genes, 15(4), p. 401. Available at: https://doi.org/10.3390/genes15040401.
Yuan, Y. et al. (2021) ‘Current status of structural variation studies in plants’, Plant Biotechnology Journal [Preprint]. Available at: https://doi.org/10.1111/pbi.13646.
Zakir, M. (2018) ‘Review on Genotype X Environment Interaction in Plant Breeding and Agronomic Stability of Crops’, Journal of Biology [Preprint].
Zargar, S.M. et al. (2023) ‘Metabolic-GWAS provides insights into genetic architecture of seed metabolome in buckwheat’, BMC Plant Biology, 23(1), p. 373. Available at: https://doi.org/10.1186/s12870-023-04381-x.
Zhan, C. et al. (2022) ‘Plant metabolic gene clusters in the multi-omics era’, Trends in Plant Science, 27(10), pp. 981–1001. Available at: https://doi.org/10.1016/j.tplants.2022.03.002.
Zhang, J. and Peters, R.J. (2020) ‘Why are momilactones always associated with biosynthetic gene clusters in plants?’, Proceedings of the National Academy of Sciences, 117(25), pp. 13867–13869. Available at: https://doi.org/10.1073/pnas.2007934117.
Zhang, Q. et al. (2021) ‘Improved genetic prediction of complex traits from individual-level data or summary statistics’, Nature Communications, 12(1), p. 4192. Available at: https://doi.org/10.1038/s41467-021-24485-y.
Zhang, X. et al. (2021) ‘Pan-genome of Raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes’, Molecular Plant, 14(12), pp. 2032–2055. Available at: https://doi.org/10.1016/j.molp.2021.08.005.
Zhang, Z. et al. (2000) ‘A greedy algorithm for aligning DNA sequences’, Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 7(1–2), pp. 203–214. Available at: https://doi.org/10.1089/10665270050081478.
Zhang, Z. et al. (2022) ‘Jujube metabolome selection determined the edible properties acquired during domestication’, The Plant Journal, 109(5), pp. 1116–1133. Available at: https://doi.org/10.1111/tpj.15617.
Zhao, J. et al. (2024) ‘Metabolite-based genome-wide association studies enable the dissection of the genetic bases of flavonoids, betaine and spermidine in wolfberry (Lycium)’, Plant Biotechnology Journal, 22(6), pp. 1435–1452. Available at: https://doi.org/10.1111/pbi.14278.
Zhao, P. et al. (2016) ‘Structural Variant Detection by Large-scale Sequencing Reveals New Evolutionary Evidence on Breed Divergence between Chinese and European Pigs’, Scientific Reports, 6(1), p. 18501. Available at: https://doi.org/10.1038/srep18501.
Zhou, Y. et al. (2019) Structural variants, hemizygosity and clonal propagation in grapevines, p. 508119. Available at: https://doi.org/10.1101/508119.
Zhu, G. et al. (2018) ‘Rewiring of the Fruit Metabolome in Tomato Breeding’, Cell, 172(1), pp. 249-261.e12. Available at: https://doi.org/10.1016/j.cell.2017.12.019.
Zhu, Q. et al. (2020) ‘Plant Synthetic Metabolic Engineering for Enhancing Crop Nutritional Quality’, Plant Communications, 1(1), p. 100017. Available at: https://doi.org/10.1016/j.xplc.2019.100017.
van Zonneveld, M. et al. (2023) ‘Forgotten food crops in sub-Saharan Africa for healthy diets in a changing climate’, Proceedings of the National Academy of Sciences, 120(14), p. e2205794120. Available at: https://doi.org/10.1073/pnas.2205794120.

# Appendix

## Manuscript

### An inter-specific *Amaranthus* pangenome captures genetic variation potentially underlying key leafy vegetable traits in this underutilised crop.

Maud A.J. de Kinderen[1*], Xiao Ma[2,3*], Fabián E. Vaistij[1], Jiyang Chang[2,3], Tony Larson[1], Yi Li[1], Fekadu Dinssa[4], Lin Chen-yu[4], Tulsi Suchak[1], Samuel J. Smit[1], Daniel C. Jeffares[5], Allen Van Deynze[6], Willem S. Jansen van Rensburg[7], Abe S. Gerrano[7,8], Michael W. Bairu[7], Sonja L. Venter[7], Roland Schafleitner[9], Yves Van de Peer[2,3,10,11] and Katherine Denby[1]

[1] Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9054 Ghent, Belgium

[3] Center for Plant Systems Biology, VIB, 9054 Ghent, Belgium

[4] World Vegetable Center, Eastern and Southern Africa, Duluti, PO Box 10, Arusha, Tanzania

[5] Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK

[6] Seed Biotechnology Center, Department of Plant Sciences, University of California, Davis, USA, 95616; African Orphan Crops Consortium, CIFOR-ICRAF, 1 UN Lane, Nairobi, Kenya

[7] Agricultural Research Council, Vegetable, Industrial and Medicinal Plants Research Campus, Private Bag X293 Pretoria 0001, South Africa

[8] Food Security and Safety Focus Area, Faculty of Natural and Agricultural Science, North-West University, Mmabatho 2735, South Africa

[9] World Vegetable Center, Shanhua, Tainan, 74155, Taiwan

[10] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[11] College of Horticulture, Nanjing Agricultural University, Nanjing, 210095, China

[*] Joint first authors

Author email addresses: maud.de.kinderen@york.ac.uk, xiao.ma@psb.vib-ugent.be, fabian.vaistij@york.ac.uk, jiyang.chang@psb.vib-ugent.be, tony.larson@york.ac.uk, yi.li@york.ac.uk, fekadu.dinssa@worldveg.org, jean.lin@worldveg.org, t.suchak@surrey.ac.uk, cobus.smit@york.ac.uk, daniel.jeffares@york.ac.uk, avandeynze@ucdavis.edu, WjvRensburg@arc.agric.za, agerrano@arc.agric.za, BairuM@arc.agric.za, SVenter@arc.agric.za, roland.schafleitner@worldveg.org, yves.vandepeer@ugent.be, katherine.denby@york.ac.uk.

**ORCID IDs:**

Maud A.J. de Kinderen 0000-0002-9439-7314
Xiao Ma 0000-0003-4787-9318
Fabián E. Vaistij 0000-0002-1469-0589
Jiyang Chang 0000-0002-3646-2878
Tony Larson 0000-0003-1337-3482
Yi Li 0000-0002-4526-7185
Fekadu Dinssa 0000-0002-9277-1413

Lin Chen-yu
Samuel Smit 0000-0002-7382-5113
Dan Jeffares 0000-0001-7320-0706
Allen van Deynze 0000-0002-2093-0577
Willem S. Jansen van Rensburg 0000-0002-5684-6174
Abe S. Gerrano 0000-0001-7472-8246
Michael W. Bairu 0000-0002-8904-1663
Sonja L. Venter 0000-0003-0513-1029
Roland Schafleitner 0000-0003-3637-7603
Yves van de Peer 0000-0003-4327-3730
Katherine Denby 0000-0002-7857-6814

**Running Head:** *Amaranthus* inter-specific pangenome

**Key Words:** *Amaranthus*, pangenome, MAGIC population, structural variants, field trial, metabolomics

## Summary

*Amaranthus* species (particularly *Amaranthus cruentus*, *Amaranthus hypochondriacus*, *Amaranthus tricolor* and *Amaranthus caudatus*) are traditional underutilised crops with the potential to contribute to sustainable, healthy food systems. We focus on amaranth as a leafy vegetable with the aim of linking phenotypic and genetic variation and developing improved lines of amaranth for cultivation by smallholder farmers in Sub-Saharan Africa. We demonstrate differences in leaf yield and metabolites relevant to human nutrition across eight amaranth accessions: four *A. cruentus* and four *A. hypochondriacus*. These accessions are founders of an inter-specific Multi-parent Advanced Generation InterCross (MAGIC) population. We generated high-quality genome assemblies and annotations for these founder lines and identified sequence and structural variants compared to a reference *A. cruentus* genome. Pangenome analysis identified core, dispensable and private gene families. 51% of gene families were core, highlighting the value, in terms of gene discovery, of sequencing additional accessions and the inclusion of two *Amaranthus* species. A graphical pangenome was constructed using structural variants and demonstrated variation in copy number of genes with a likely role in disease resistance. This inter-specific pangenome will be highly valuable for future research on amaranth and facilitate usage of structural variants in trait mapping and causal gene discovery.

## Introduction

The majority of *Amaranthus* species originate in the Americas, but there are over 70 species now distributed around the world (Riggins and Mumm 2021). Three species – *Amaranthus cruentus*, *Amaranthus hypochondriacus* and *Amaranthus caudatus* – are commonly grown for their grain and considered pseudo-cereals. Species such as *Amaranthus tricolor* are grown as a leafy vegetable, however, leaf and grain amaranths cannot be distinguished purely by species, since there are accessions of grain amaranth species used for vegetable production (Lin *et al.* 2022).

In Sub-Saharan African, amaranths are viewed as a traditional crop and are one of the most commonly consumed African leafy vegetables. Their consumption is on the rise in East Africa but has been declining in South Africa, at least in part due to urban migration, shifts away from traditional diets and a lack of these crops in formal markets (Nyonje *et al.* 2022). However, leafy amaranth is recognized, along with other underutilised or orphan crops, for its potential in contributing to sustainable and healthy food systems (Mabhaudhi *et al.*

2016) and several amaranth leafy vegetable species were identified as priority crops for enhancing nutrition and climate-resilient agriculture in Sub-Saharan Africa (Zonneveld *et al.* 2023). Amaranths exhibit $C_4$ photosynthesis (Kadereit *et al.* 2003), meaning photosynthesis is more efficient under high temperatures, high light intensities and reduced water availability (Edwards, Franceschi and Voznesenskaya 2004). Amaranths grow well in low input agricultural systems typical of smallholder farmers and leaves contain high levels of vitamins, and antioxidants, as well as essential amino acids (Venskutonis and Kraujalis 2013; Achigan-Dako, Sogbohossou and Maundu 2014). Despite these beneficial traits, there are still challenges to overcome to realise the potential of this traditional crop. Amaranth has not been fully domesticated and the traits that have been targeted are those relevant to grain crops (e.g. seed colour (Anuradha *et al.* 2023)). Limited research has been carried out on leafy amaranth (Emmanuel and Babalola 2021), with limited improvement of lines and molecular genetic characterisation of traits (Riggins and Mumm 2021). Capitalising the potential of amaranths to enhance nutrition, dietary diversity, smallholder livelihoods and climate-resilient agriculture, will require the development of improved lines suitable for different agro-ecologies.

To date, three chromosomal-level *Amaranthus* reference genomes have been published: *A. hypochondriacus* (Lightfoot *et al.* 2017), *A. cruentus* (Ma *et al.* 2021) and *A. tricolor* (Wang *et al.* 2022). All are diploid with *A. hypochondriacus* and *A. cruentus* having current genome assembly sizes of 370 and 400 Mb and n=16 and n=17 chromosomes, respectively. *A. tricolor* has a larger genome size (current assembly of 520 Mb) distributed over 17 chromosomes. Near chromosomal-level assemblies have been published for *A. tuberculatus*, *A. hybridus* and *A. palmeri* – species renowned as invasive weeds (Montgomery *et al.* 2020). In addition to these genome assemblies, short read sequencing of 121 accessions of *A. caudatus*, *A. cruentus*, *A. hypochondriacus* and *A. quitensis* was carried out to investigate (incomplete) domestication (Stetter, Vidal-Villarejo and Schmid 2020) and a genotype-by-sequencing approach was used to clarify taxonomy of more than 700 accessions including tetraploid *A. dubius* (Lin *et al.* 2022).

Reference genomes are valuable genetic resources, but they do not capture the full genetic diversity of these species. Comparison and integration of high-quality assemblies of multiple accessions of a species (pangenomic studies) enables better understanding of genome evolution and phenotypic diversity and linking of genetic variation to agronomic traits. Pangenomes are particularly valuable for identification of structural variants (insertions, deletions, copy number variants and translocations), that are known to play a key role in phenotypic variation (Alonge *et al.* 2020). Multiple model plant and crop pangenomes have been assembled (Jayakodi *et al.* 2021) significantly expanding the scope of genetic diversity and improving trait mapping and prediction (Coletta *et al.* 2021). Pangenomic studies have also captured inter-specific variation between domesticated cultivars and wild relatives (Zhang *et al.* 2021).

We generated an inter-specific pangenome from high-quality genome assemblies of eight amaranth accessions, four *A. cruentus* and four *A. hypochondriacus*, the parents of a Multi-parent Advanced Generation Inter-Cross (MAGIC) population. We identified sequence and structural variants between the lines compared to the reference *A. cruentus* genome (Ma *et al.* 2021) and assembled both a graphical and gene-based pangenome. The accessions vary for leafy vegetable traits, including yield and metabolite profiles, and the pangenome

highlights copy number variation in disease resistance genes between the lines. Creation of this inter-specific amaranth pangenome will underpin use of the MAGIC population in developing improved lines of this traditional crop.

## Materials and Methods

### Field trial

The eight MAGIC population parent accessions, VI044371, VI044437A (*A. cruentus*), VI061487, VI061494A (*A. cruentus* admixture), VI062433, VI063769, VI050446 (*A. hypochondriacus*) and VI060472 (*A. hypochondriacus* admixture) were grown at WorldVeg for Eastern and Southern Africa in Arusha, Tanzania (coordinates: 3.37° south, 36.8° east). The trial took place in the warm season over 83 days, with the average time between the five harvests 13 days (Supp Dataset 2a). Three replicate plots of each accession were initially planted with 20 plants. At each harvest, the number of surviving plants was verified. Only one replicate of one line had less than 17 surviving plants at harvest 4, and four individual plots had less than 15 plants at the last harvest (Supp Dataset 2c). Leaf yield was measured in grams per plot and subsequently divided by the number of plants to estimate leaf yield per plant. The number of leaves per plant was counted. For leaf length and leaf width, 10 random leaves were picked from 10 random plants per plot at each harvest and measured. After checking for normality, trait data were analysed using linear mixed-effects models producing restricted maximum likelihood (REML) estimates determined by the lmer function in the lme4 package (Bates *et al.* 2015). The most complete model used to test significance of different variables on phenotypic traits was

$$y_{ijk} = \mu + Accession_i + Harvest_j + Acc.*Harvest_{ij} + Rep_k + e_{ijk}$$

$y_{ijk}$ is the trait prediction, $\mu$ is the mean, $Accession_i$ is the fixed effect of line (n=8), $Harvest_j$ is the fixed effect of Harvest (n=5), $Acc.*Harvest_{ij}$ is the fixed effect of accession by harvest interaction, $Rep_k$ is the random effect of replicate plot (n=3) and $e_{ijk}$ is the random residual assumed to be $\sim N(0, I\sigma^2{}_e)$, with I being an identity matrix and $\sigma^2{}_e$ the residual variance. Significance of the fixed variables was tested using Analysis of Variance with the model adjusted in a backward stepwise manner and the model with the lowest Akaike Information Criterion considered best fitting. $Rep_k$ was considered a random model effect. Tukey's HSD test was used for post-hoc comparisons.

### Metabolite profiling

The eight MAGIC parent accessions were grown in soil (Levington Advance, F2+S) in a glasshouse with supplemental lighting (240 mm $m^{-2}s^{-1}$ PPFD) for 12 hour days, at ambient humidity and air temperatures of 27°C (+/- 3°C) day; 25°C (+/- 3°C) night. The youngest fully expanded leaves were harvested from at least 3 independent 6-week old plants and methanolic extracts analysed in triplicate using liquid chromatography with data-dependent mass spectrometry (LC-MS/MS). Detailed methods are provided in Supplementary Information.

### RNA sequencing and analysis

RNA sequencing was carried out on the same samples as metabolite profiling, with an additional two samples per accession from two independent extractions of a second pool of leaves.

Three samples from *A. cruentus* "Arusha": leaf and roots from 6-week old plants and 2-week old seedlings were also included. Illumina short-read transcript sequencing on the NovaSeq

6000 platform generated a minimum of 28 million 150-bp paired-end reads for each sample. Read quality was assessed using FastQC (Wingett and Andrews 2018) and adapters removed using Trimmomatic (Bolger, Lohse and Usadel 2014). Subsequently, reads were aligned to the relevant genome (i.e. *A. cruentus* reference or a MAGIC population parent line) and transcript abundance measured using Salmon (Patro *et al.* 2017). Tximport R package was used to generate read counts and length-scaled transcript per million reads (TPMs)(Soneson, Love and Robinson 2016). Mean expression values were used in Supp Figure 6.

### Genome sequencing and assembly

Long-read PacBio HiFi sequencing was carried out for the eight founders, generating between 637,951 and 972,273 raw reads per accession. Genome assembly was performed using the *de novo* assembler Hifiasm v0.8-r279 (Cheng *et al.* 2022). Primary genome assemblies were purged to a full haploid genome by haplotype separation to obtain contig level assemblies using purge_dups VI.0.1 (Guan *et al.* 2020). GenomeScope 2.0 (Ranallo-Benavidez, Jaron and Schatz 2020) was used for assessing heterozygosity levels in the genomes and genome size of the eight founders with the raw PacBio HiFi sequencing reads. The quality of the genome assemblies was assessed using QUAST (Gurevich *et al.* 2013).

### Pseudochromosome construction and genome annotation

The assembled contigs of the four *A. cruentus* and four *A. hypochondriacus* accessions were anchored and oriented to chromosomes using the published reference genomes for each species and the reference-guided software RagTag V2.1.0 (Alonge *et al.* 2022) with default parameters. The repetitive sequences in the eight newly assembled genomes were identified by RepeatModeler v.2.0.1 (Flynn *et al.* 2020) based on the TE Dfam database and classified by RepeatMasker with the parameter: -e rmblast -a -s -norna -xsmall -gff.

An integrated gene annotation pipeline was used to annotate the protein-coding genes in the eight amaranth genomes. RNA-seq reads were aligned to their assembled genome using hisat2 v.2.1.0 (Kim *et al.* 2019) with parameters: --max-intronlen 500000 –dta and transcripts were assembled using StringTie v2.1.5 (Kovaka *et al.* 2019): -m 150 -t -f 0.3. Transdecoder v.5.0.2 (Haas *et al.* 2008) was used to predict the ORFs: --retain_blastp_hits. Homology-based prediction was performed using GeMoMa v.1.9 (Keilwagen *et al.* 2018) based on an improved annotation of the *A. cruentus* reference by manual gene curation. BRAKER2 (Brůna *et al.* 2021) was used for *de novo* prediction using model training based on the RNA-seq data after the annotated repeats were soft masked in the assembly. EVidenceModeler v.1.1.1 (Haas *et al.* 2008) was used to combine all gene model evidence obtained. Subsequently, manual curation of gene annotations in all eight MAGIC founder lines was performed using GenomeView (Abeel *et al.* 2012). Where there was clear RNAseq evidence, missing genes were included, fragmented genes repaired and adjacent joined genes separated. The quality of the annotation results was evaluated by BUSCO v4.0.4 using the embryophyta Odb10 gene set.

Putative gene function was identified using InterProScan v5.47-82 (Paysan-Lafosse *et al.* 2023) with different databases: PFAM, Gene3D, PANTHER, CDD, SUPERFAMILY, ProSite and GO. Functional annotation of predicted genes was also obtained by aligning amaranth protein sequences against sequences in public protein databases and the UniProt database using BLASTP (E-value <1 × 10−5).

### Gene-based pangenome

A gene-based pangenome was constructed with the ten amaranth assemblies by estimating core, dispensable and private gene sets based on gene family clustering using OrthoFinder v.2.3.3 (Emms and Kelly 2019). All protein sequences were subjected to homology searches using BLASTP with an expected value of 1 × 10−5. Protein sequences were clustered into paralogous and orthologous sequences using OrthoFinder with an inflation parameter of 1.5. GO-enrichment was conducted with the topGO package (Alexa and Rahnenfuhrer 2023).

### Identification of variants and generation of graphical pangenome

The HiFi reads of the eight amaranth accessions were aligned to the *A. cruentus* reference using MUMmer v.4.0.0 (Marçais *et al.* 2018). SNPs were identified using Longshot (Edge and Bansal 2019) retaining only the variants labeled as "PASS" and filtered using the following criteria: QUAL (the sixth column) greater than 50 and DP greater than 10 (QUAL > 50 and DP > 10). PBSV ('PacificBiosciences/pbsv', 2024) and cuteSV (Jiang *et al.* 2023) were used to identify SVs. For PBSV the following filters were used: --min_insertion_len 50, --max_insertion_len 100000, --min_deletion_len 50, and --max_deletion_len 100000. Insertions and deletions (tags INS and DEL) were categorised as PAVs and duplications (tag DUP) as copy number variations (CNVs). PBSV and cuteSV VCF files were merged to get the non-redundant file using SURVIVOR (Jeffares *et al.* 2017), only retaining SVs < 100kb. Assemblytics version 1.2.1 (Nattestad and Schatz 2016) was used to identify PAV > 100kb. Genome alignment was performed using NUCmer (Marçais *et al.* 2018) with parameters: --maxmatch -l 100 -c 500. Raw alignments were further filtered with delta-filter using: -1 -i 90 -l 100 and then Assemblytics to parse the filtered delta file to obtain SVs. A graphical pangenome was constructed using the 8 MAGIC parent line genome assemblies, as well as the reference *A. cruentus* genome with updated annotation, with the Mini-graph Cactus pipeline (Hickey *et al.* 2023).

A PCA was made with an annotated biallelic SNP VCF with BCFtools v.1.10.2-foss-2018b (Danecek *et al.* 2021), filtered for quality > 1$^{st}$ quantile with VCFtools v.0.1.16-foss-2018b-Perl-5.28.0 (Danecek *et al.* 2011) and for LD with PLINK v.1.9_beta-6.16 (Chang *et al.*, 2015), with pairwise comparisons calculated in 10 kb windows, a step size of 1, and an R Squared ($R^2$) threshold of 0.5. The SNP VCF file was split per pangenome component (core and dispensable), and π was calculated in 100 kb windows for each VCF set with VCFtools v.0.1.16-foss-2018b-Perl-5.28.0.

Pairwise synteny analyses were performed using the JCVI v1.2.7 (Tang *et al.* 2024) implementation of MCScan (Wang *et al.* 2012). Protein sequences were used for ortholog detection under default settings followed by a maximum of four iterations for MCScan. Microsyntenic blocks were visualised through JCVI and manually supplemented with syntenic links using the output of OrthoFinder outlined above.

## Results

### The eight parent accessions of a *Amaranthus* MAGIC population show variation in agronomic traits

The World Vegetable Center (WorldVeg) conserves 1,455 *Amaranthus* accessions, with more than half of them classified as vegetable or leafy amaranth (Lin *et al.* 2022). From this large diversity panel, eight founder parent accessions were selected to generate a MAGIC population. These parents were classified in the WorldVeg genebank as four *Amaranthus*

*hypochondriacus* and four *Amaranthus cruentus* species, although two of the lines were reclassified following *de novo* SNP calling (Lin *et al.* 2022)(Supplementary Dataset 1). These two accessions (VI060472 and VI061494A), along with VI061487, did not appear to have a single ancestral genomic component in that study, with maximum Q values (proportion of an individual's genome originating from one population) of 0.8669, 0.5507 and 0.8147. The lines were characterised using morphological descriptors (as part of >700 lines assessed from the WorldVeg collection) and the 8 parents showed variation in a number of traits relevant to improvement for grain or leaf production (Supplementary Dataset 1). Furthermore, the combination of traits across the 8 lines differed, illustrating the diversity captured by these parent lines. *A. hypochondriacus* is typically a grain amaranth whereas *A. cruentus* is grown for both grain and leaf, and considered a traditional leafy vegetable in Sub-Saharan Africa (Zonneveld *et al.* 2021). Although, in general, *A. cruentus* flower later than grain amaranth, and are taller with larger leaves (Lin *et al.* 2022), within the 8 lines no traits clearly differ between the two species (Supplementary Dataset 1). However, individual accessions show traits that are of importance to development of improved lines for cultivation. VI063769 has multiple branches emanating from the base of the stem (rather than branches up the stem) which eases manual harvesting of leaves and VI044371 shows reduced seed shattering compared to other accessions. Several accessions show longer times to flowering (lengthening the vegetative phase) and pigmentation in the stem, leaf and petiole, which influences acceptance of fresh produce. The founder accessions also vary for seed colour (a key grain trait) (Supp Dataset 1).

Agronomic characteristics of the MAGIC population parent accessions were assessed in an irrigated randomised complete block design field trial at WorldVeg in Arusha, Tanzania during the warm season (November 2019 – February 2020, average temperature 21.6°C) (Supp Dataset 2). Leaves were harvested five times during the trial with regrowth of the plants in between; time between harvests varied from 11 to 30 days. The significance of the variation in traits was assessed using linear mixed-effects models including harvest and accession as fixed variables along with their interaction (Supp Table 1). Leaf yield, length, width and number all varied significantly by harvest, with leaf yield peaking at the third harvest and the first and fifth harvests having the lowest yield (Figure 1, Supp Figure 1a). Leaf number followed a similar pattern (Supp Figure 1b) whereas leaf length and width decreased from the first harvest (Supp Figure 1c, d). Crucially there was significant variation between accessions for leaf yield (Supp Table 1), with a significant difference between accessions VI044371 and VI050446 for third harvest yield (Figure 1a) and total yield (i.e. the sum of all five harvests) (Figure 1b). VI044371 and VI050446 also showed a significant difference in leaf length (but not width) across the whole trial (Supp Figure 1e, f), likely underlying the difference in yield as leaf number did not differ significantly between accessions (Supp Table 1).

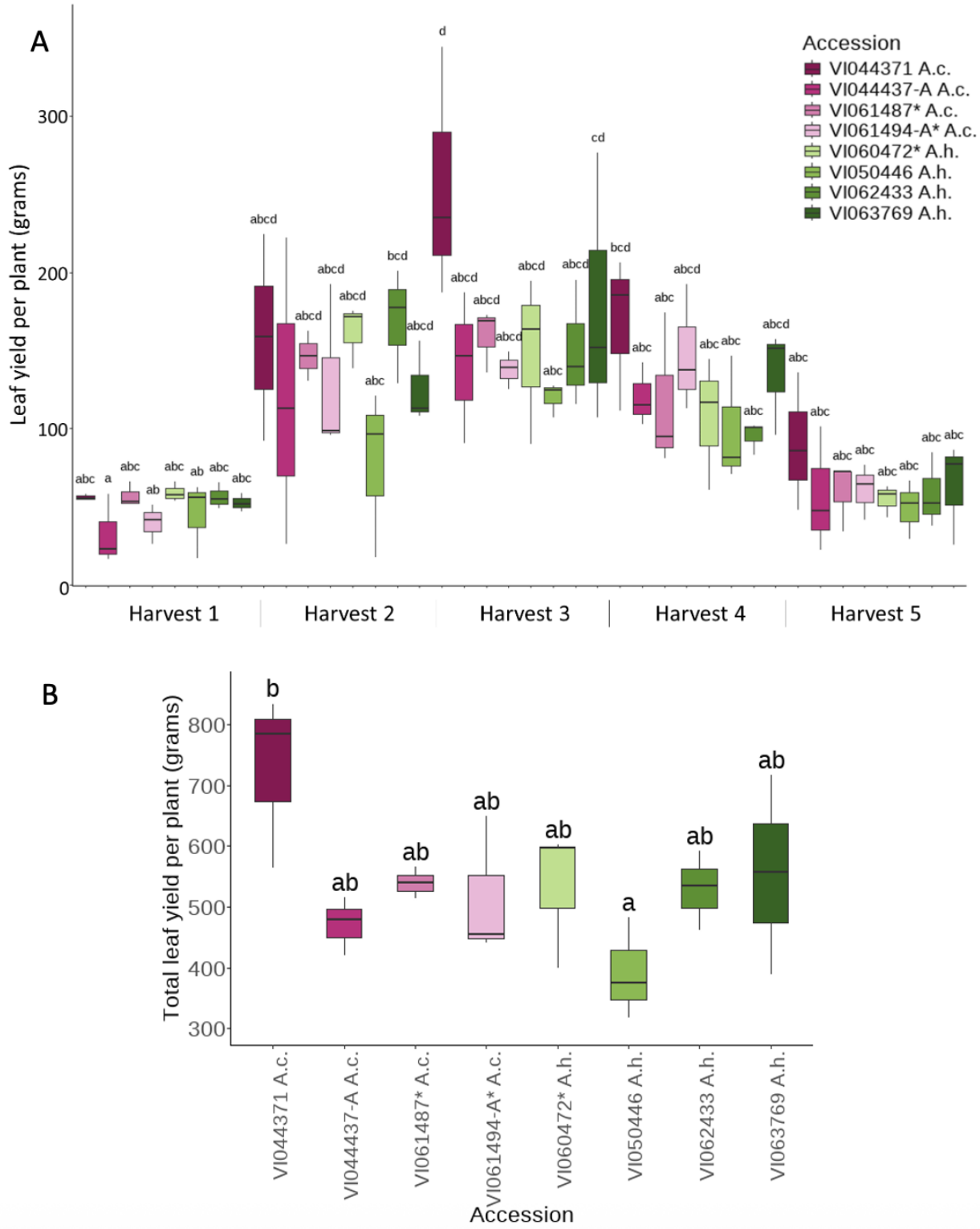



**Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A)** Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters indicate significant differences within each harvest as assessed by Tukey's honest significance test. Initials A.c. and A.h. indicate the species of each line with '*' indicating lines with admixture.

### The eight amaranth accessions vary in metabolite profiles relevant to human nutrition

The nutritional content of traditional crops, such as amaranth, is an important factor in the value of these crops to local food systems. We assessed the eight founder lines for variation in metabolite content of greenhouse-grown leaves using untargeted metabolite profiling of methanolic extracts with three replicates.

Liquid chromatography–mass spectrometry (LC-MS) analysis resulted in 3256 features from the positive and negative ionization modes after correcting for redundancy (Supp Dataset 3). Hierarchical clustering using normalised data for these 3256 non-redundant features demonstrates the similarity of replicates as well as clear separation of the two species of amaranth (Supp Figure 2a). Principal component analysis using the abundance of the 3256 non-redundant features again demonstrated clear variation between the two amaranth species (Supp Fig 2b).

After removing very low abundance features, 1421 non-redundant peaks remained of which 1227 varied significantly in abundance between the 8 parent accessions and 847 showed a significant difference in abundance between the two species (*A. cruentus* and *A. hypochondriacus*) (Supp Dataset 3a). Of the 1227, 122 peaks could be annotated and represented 18 major metabolite classes. After clustering of these 122 features on the basis of their abundance across the samples, features with the same annotation cluster together (Supp Figure 3) providing support for the accuracy of annotations. The abundance of the 122 annotated metabolites was also able to clearly distinguish between the two species groups, consistent with their revised species attribution (Lin et al. 2022) (Figure 2a). Figure 2b highlights the variation in abundance of these 122 metabolites both across and within species. Of interest to our aim of developing improved amaranth lines, many of these annotated metabolites are of relevance to human nutrition.

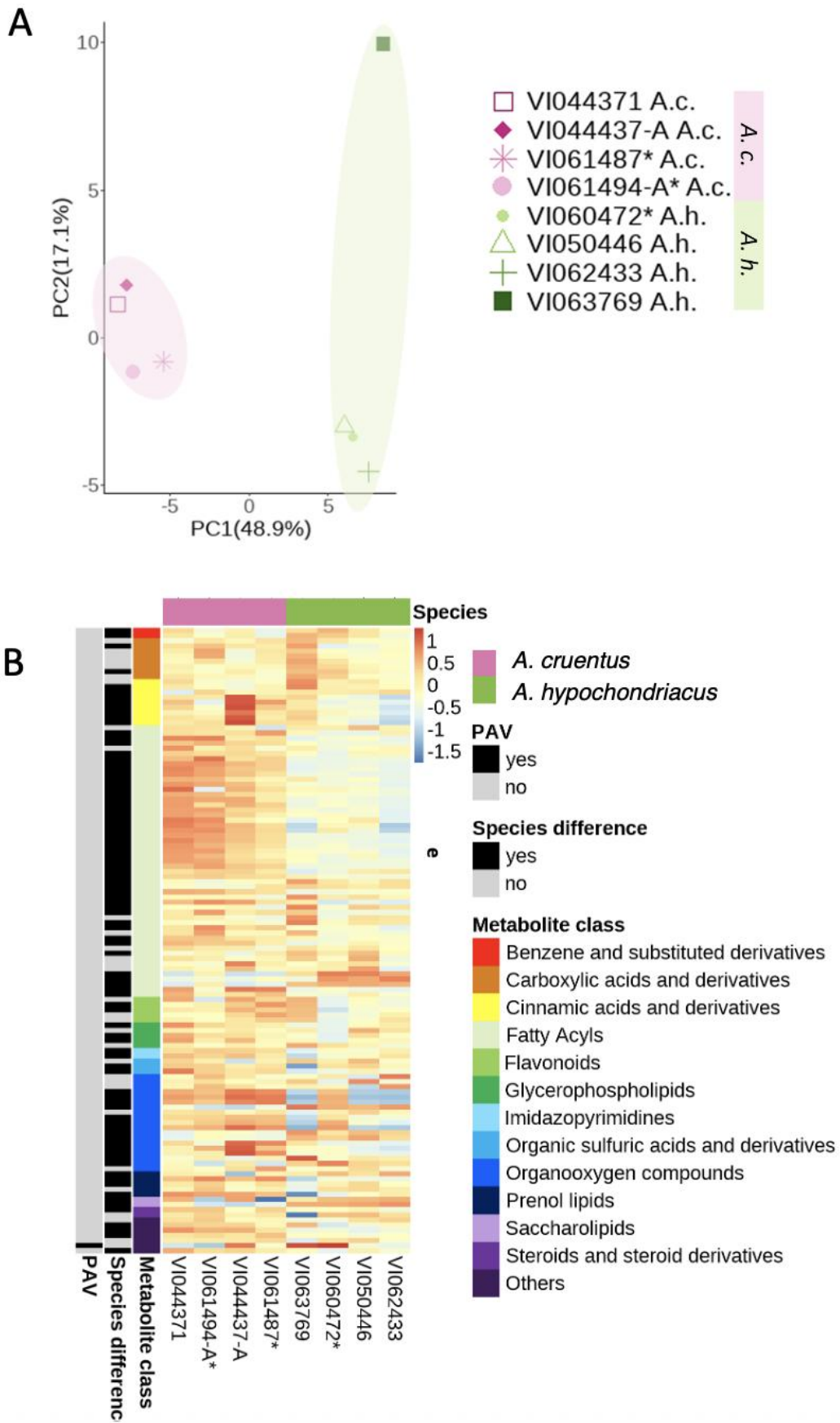



**Figure 2. Variation in metabolite content between the eight *Amaranthus* MAGIC population parent accessions. A)** Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

The largest group of annotated metabolites varying across the 8 accessions are lipids, including fatty acyls, glycerolipids, glycerophospholipids, prenols, saccharolipids and steroids. Lipids are obviously an important nutritional component, and many of these (particularly fatty acyls) appear to differ between the *A. cruentus* and *A. hypochondriacus* accessions (Figure 2b) with higher levels in the *A. cruentus* accessions. Polyphenolic compounds including flavonoids, benzenoids, sesquiterpenes, tannin and hydroxycinnamic acid esters vary across the 8 MAGIC accessions (Figure 2b). These compounds have been associated with multiple health benefits, most likely through their anti-oxidant properties (Monjotin *et al.* 2022).

### Generation of high-quality annotated genome assemblies for the eight Amaranth accessions

To aid association of phenotypic traits with genome sequence, we wanted to be able to identify both sequence and structural variants in the amaranth MAGIC population founder lines. We generated high-quality *de novo* assemblies of the eight amaranth accessions using PacBio HiFi sequencing data (Supplementary Dataset 4a,b, Supplementary Figure 4a). The resulting genomes range in size from 404 Mb to 420 Mb, with contig N50 values from 1.1 Mb to 1.8 Mb (Table 1), consistent with estimated genome sizes based on k-mer analysis (Table 1). The genomes were annotated using a combination of *ab initio* prediction, homology searches, and RNAseq evidence (Supplementary Dataset 4c), supplemented by rigorous manual curation. The resulting annotations identified between 23,756 and 24,973 gene models in each genome (Table 1). We also carried out manual curation of the *A. cruentus* reference genome (Ma et al. 2021) reducing the number of annotated genes from 25,477 to 23,814, but increasing the average gene length from 4,337 to 4,606 bp (Table 1). This manual curation resulted in the BUSCO score increasing from 90.5% complete genes to 94.8% (an additional 69 genes) and the functional annotation rate improving from 94.3% to 97%. The average length of genes and coding regions (CDS) were very consistent across the eight MAGIC population parent and *A. cruentus* reference genome, ranging from 4502-4606 bp and 1192-1240 bp respectively (Table 1). We compared the quality of our annotations to those of six other representative species in the *Amaranthaceae* family, as well as *Arabidopsis thaliana* and *Vitis vinifera* (Supplementary Figure 4b). With the exception of the *A. hypochondriacus* reference (Lightfoot et al., 2017), the distributions of CDS length and exon number of our amaranth genomes were highly consistent with those of other *Amaranthaceae, A. thaliana* and *V. vinifera*. Given the extensive validation and hence high quality of the *A. thaliana* genome (Zhang *et al.* 2022a), this consistency provides evidence for the quality of our annotations. Over 95% of annotated genes in the 8 parent accessions can be assigned functions using InterProScan (v5.47-82) and BUSCO (v4.0.4) scores of >95% demonstrate the high level of completeness of the annotated gene models in the eight genomes (Table 1).

**Table 1.** Summary of genome assemblies and annotation for the eight *Amaranthus* MAGIC population parent accessions and *A. cruentus* reference genome (Ma *et al.*, 2021). Ac = *A. cruentus*, Ah = *A. hypochondriacus* with * indicating accessions with genome admixture.

| Sample | *A. cruentus* reference | VI044371 | VI044437 | VI061487* | VI061494-A* | VI050446 | VI060472* | VI062433 | VI063769 |
|---|---|---|---|---|---|---|---|---|---|
| Species | Ac | Ac | Ac | Ac | Ac | Ah | Ah | Ah | Ah |
| **Assembly** | | | | | | | | | |
| Estimated genome size (Mb) | 400 | 410.3 | 412.4 | 404.5 | 425.5 | 416.4 | 410.6 | 403.2 | 406.3 |
| Total length (Mb) | 370.9 | 410.0 | 404.8 | 413.4 | 416.4 | 420.3 | 404.4 | 410.7 | 416.8 |
| No. of contigs | 625 | 695 | 766 | 875 | 1039 | 843 | 735 | 773 | 903 |
| Contig N50 (Mb) | 21.7 | 1.8 | 1.6 | 1.1 | 1.1 | 1.3 | 1.4 | 1.5 | 1.5 |
| **Annotation** | | | | | | | | | |
| Number of annotated genes | 23,814 | 24,973 | 23,756 | 24,475 | 24,595 | 24,494 | 23,988 | 24,103 | 24,698 |
| Average gene length (bp) | 4606 | 4502 | 4574 | 4523 | 4525 | 4590 | 4545 | 4577 | 4543 |
| Average mRNA length (bp) | 1240 | 1192 | 1240 | 1215 | 1217 | 1227 | 1216 | 1224 | 1212 |
| Average exon length (bp) | 238 | 231 | 236 | 233 | 233 | 233 | 231 | 233 | 231 |
| Average exons per gene | 5.19 | 5.15 | 5.26 | 5.21 | 5.21 | 5.26 | 5.26 | 5.25 | 5.23 |
| Average intron length (bp) | 799 | 798 | 784 | 786 | 787 | 790 | 782 | 790 | 788 |
| **BUSCO** | | | | | | | | | |
| Complete | 94.8% | 96.1% | 96.8% | 97.0% | 98.0% | 96.8% | 95.3% | 96.1% | 97.0% |
| Complete and single-copy | 91.9% | 93.4% | 93.4% | 93.8% | 94.9% | 93.6% | 92.3% | 93.1% | 93.8% |
| Complete and duplicated | 2.9% | 2.7% | 3.4% | 3.2% | 3.1% | 3.2% | 3.0% | 3.0% | 3.2% |
| Fragmented | 0.6 | 0.8% | 0.7% | 0.8% | 0.8% | 1.0% | 0.9% | 0.7% | 0.8% |
| Missing | 4.6% | 3.1% | 2.5% | 2.2% | 1.2% | 2.2% | 3.8% | 3.2% | 2.2% |
| **Functional annotation (%)** | 97 | 95.7 | 97.2 | 95.1 | 95 | 96.8 | 96.5 | 96.6 | 96.4 |

A similar proportion (~57%) of the assembled sequence of each accession was composed of repetitive sequences, with the highest proportion of known elements being long-terminal repeat retro-elements (LTRs)(Supplementary Dataset 5). The Copia content was consistently slightly higher than Gypsy content in all amaranth accessions, which is in contrast to other *Amaranthaceae* species like quinoa and beet (Ma *et al.* 2021). The eight newly sequenced amaranth accessions and the two reference genomes (Lightfoot *et al.* 2017; Ma *et al.* 2021) showed a high degree of pairwise co-linearity at the chromosomal scale, with approximately 80% of genes in co-linear blocks (Supp Figure 5a) and good concordance in gene space representation (Supp Figure 5b). Within each genome approx. 38% of genes were in co-linear blocks (representing duplicated genes), apart from the *A. hypochondriacus* reference genome with only 23.38% of genes (Supp Fig 5a). The co-linearity between *A. hypochondriacus* and the other accessions was also slightly lower than the overall average at 75% (Supp Figure 5a). A phylogenetic tree based on 5534 single-copy genes common to all 10 accessions separated the two species as expected, and in line with the metabolite data above and GBS data (Lin et al. 2022) (Supplementary Figure 5c).

### The genetic repertoire of an inter-specific Amaranth pangenome

Using Orthofinder (Emms and Kelly 2019), genes from the ten amaranth genomes were classified into 23,796 assigned families and 5,042 unassigned single-copy genes unique to one accession (Supplementary Dataset 6). 14,805 orthogroups were detected in all ten genomes and defined as the core gene families (51%), 8,979 were identified in 2 - 9 accessions, termed dispensable gene families (31%), and 12 gene families (41 genes) were detected in only one accession, which along with the single-copy genes unique to a single accession, resulted in 5,054 private orthogroups (18%)(Figure 3a, b). 86% of the core gene families (12,755) are single copy genes in all accessions. The percentage of core gene families declined as more amaranth accessions were added to the pangenome as expected, but did not plateau (Figure 3c) indicating further accessions are likely to bring additional variation in gene content. Approximately 70% of the genes in each genome were part of core gene families in the final pangenome of the 10 accessions (Figure 3d), with ~30% variable gene content. Evaluating conservation of genes within each amaranth species demonstrated a higher proportion of core genes in each species sub-pangenome compared to the full pangenome (Table 2).

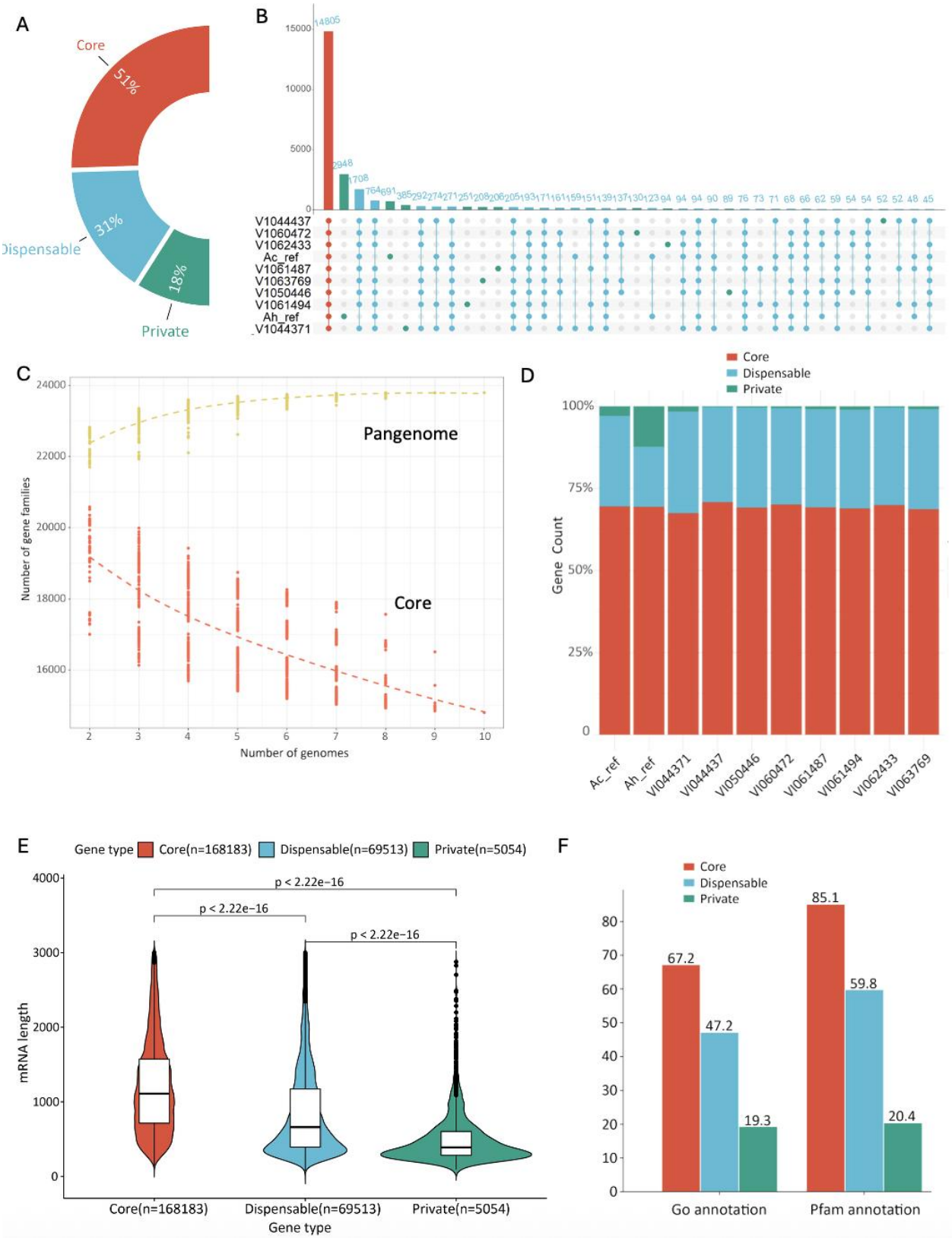



**Figure 3:** Analysis of pangenome components across ten *Amaranthus* genomes (eight MAGIC population parents and two published reference genomes, *A. cruentus* and *A. hypochondriacus).* **A)** The proportion of core, dispensable and private gene families (including single copy genes unique to one accession) across the ten amaranth genomes. **B**) The number of gene families shared between the ten amaranth genomes. **C**) The proportion of core and total gene families in the pangenome as the number of genomes compared increases. **D**) The proportion of genes in each genome that are classified as core, dispensable or private in the pangenome of 10 accessions. **E**) mRNA length of genes in the three pangenome components. **F)** The proportion of core, dispensable and private genes that are functionally annotated via Gene Ontology or the Pfam database.

**Table 2**: Composition of species-specific subpangenomes. The proportion of core, dispensable and private gene families (and single copy private genes) in the species-specific sub pangenomes, and subpangenomes of the *A. cruentus* and *A. hypochondriacus* MAGIC founder accessions only (i.e. not including the *A. cruentus* or *A. hypochondriacus* reference genomes of Ma et al. 2021 and Lightfoot et al. 2017, respectively).

| | **Full pangenome** | ***A. cruentus* accessions only** | ***A. hypo-chondriacus* accessions only** | ***A. cruentus* MAGIC founder lines only** | ***A. hypo-chondriacus* MAGIC founder lines only** |
|---|---|---|---|---|---|
| **Core gene families** | 51% | 72% | 64% | 80% | 86% |
| **Dispensable gene families** | 31% | 18% | 20% | 11% | 10% |
| **Private gene families and single copy private genes** | 18% | 10% | 16% | 9% | 4% |

We observed that the annotated mRNA length of core genes is significantly longer than that of dispensable and private genes (Figure 3e). Furthermore, using RNA-seq data from leaf tissue of the eight MAGIC founder accessions, we observed that overall core genes showed higher expression compared to dispensable genes but there was no obvious difference between expression of dispensable and private genes (Supp Figure 6a). Core genes also showed reduced nucleotide diversity compared to dispensable genes indicating a greater level of conservation (Supp Figure 6b). A number of gene families (1708) were found in all the amaranth accessions except the *A. hypochondriacus* published genome (Lightfoot *et al.* 2017)(Figure 3b) and over half of the private genes were found only in this genome (Supp Figure 6c). The species-specific subpangenome of just the four *A. hypochondriacus* founder lines had a much higher proportion of core gene families than an *A. hypochondriacus* subpangenome also including the reference genome (Table 2). The annotation of the *A. hypochondriacus* reference genome appears to be of lower quality than the other accessions, with shorter gene lengths compared to the *A. cruentus* reference (Ma *et al.* 2021), which is likely the reason for the high proportion of private genes in this genome. In contrast, the subpangenome of just the *A. cruentus* MAGIC founders (not including the reference genome) had a similar proportion of private gene families to the full *A. cruentus* subpangenome Table 2).

1,798 Benchmarking Universal Single-Copy Ortholog (BUSCO) genes were identified in the amaranth inter-specific pangenome and, as expected, most of these genes (68%, 1,230 genes) were part of core gene families (Supp Figure 6d). 27% (480 genes) were dispensable and 5% (88 genes) were classed as private genes. It is also clear from this analysis that the *A. hypochondriacus* reference genome in particular, lacks a significant proportion of BUSCO genes present in all other 9 genomes (Supp Figure 6d).

Gene Ontology (GO) and protein domain annotation (Pfam) analysis revealed that the proportion of genes with a functional annotation decreased from core to dispensable to private genes (Figure 3f). This signal, combined with the lower nucleotide diversity and

higher gene expression of core genes, is consistent with the core genes being essential. We analysed GO term enrichment among the genes assigned to different pangenome components. Single-copy genes and gene families were newly assigned to pangenome components based on their presence in the 9 amaranth accessions ignoring the *A. hypochondriacus* reference (Supp Dataset 6a) with the key change being a significant reduction in the number of private genes (Table 2). GO term enrichment analysis did not reveal striking differences in categories of gene function between the genes in the pangenome components, although the most significantly enriched term in the core gene families was DNA-binding transcription factors with over 80% of the pangenome transcription factor genes being part of core gene families (Supp Dataset 7). Other central biological processes were also enriched such as RNA binding, chromatin binding, protein deubiquitination, microtubule binding and nitrogen metabolism. Dispensable gene families included genes with a role in response to auxin, wounding and oxidative stress, as well as cell wall biogenesis, potentially underlying plant adaptation to changing environmental conditions (Supp Dataset 7).

### Sequence and structural variation in the amaranth pangenome

The HiFi sequencing reads of the eight MAGIC population parent accessions were aligned to the *A. cruentus* “Arusha” reference genome (Ma *et al.* 2021) to identify both sequence and structural variants. A total of 8,378,957 single nucleotide polymorphisms (SNPs) were identified: 1,767,497 were found in the four *A. cruentus* parent accessions, while 6,611,460 were detected in the four *A. hypochondriacus* parent accessions, indicating these genomes were less similar to the *A. cruentus* reference (Table 3). Furthermore, the two *A. cruentus* accessions showing genome admixture (VI061487 and VI061494, Li et al. 2022) contained a higher number of SNPs compared to the two other *A. cruentus* accessions, and conversely, the *A. hypochondriacus* accession with genome admixture (VI060472) had the lowest number of variants amongst the *A. hypochondriacus* parent genomes, although the difference was less pronounced. Despite the genome admixture, PCA analysis using the SNPs clearly split the MAGIC parent lines into two species-defined groups (Figure 4a) as seen with the metabolite profiling data (Figure 2a).

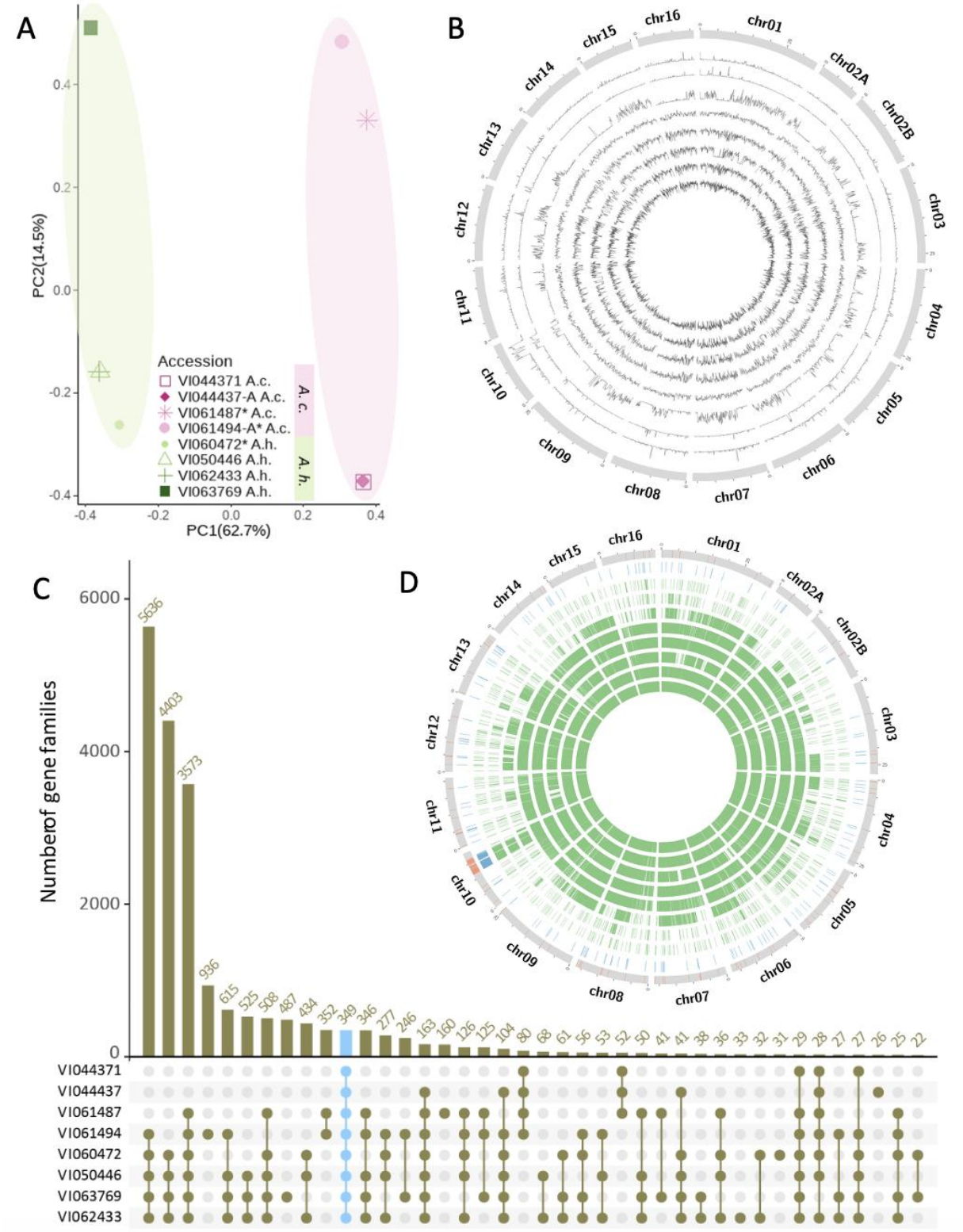



**Figure 4**. Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome. **A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups. **B**) the distribution of SNPs along the genome. From the outer to inner layer is V1044371, V1044437, V1061487, V1061494, V1050446, V1060472, V1062433 and V1063769. **C**) the number of gene families that contain SNPs in at least one of the eight MAGIC founder lines. Blue indicates (core) gene families where all MAGIC parents contain SNPs in the genes compared to the *A. cruentus* reference. **D**) the distribution of genes containing SNPs along the genome. Green is all genes with SNPs for each accession (accessions in the same order as in B). Blue indicates core genes with SNPs in all accessions and orange is single copy core genes with SNPs in all accessions.

**Table 3: The number of Single Nucleotide Polymorphisms (SNPs) and structural variants (SVs) in each MAGIC parent accession genome.** The variants are in comparison to the *A. cruentus* "Arusha" reference genome (Ma et al. 2021). The total number of SNPs, SNP density, SNPs in protein coding regions and the number of genes containing at least one SNP are given along with the number of presence absence variation (PAV), inversions (INV), copy number variations (CNV) and translocations (TRANS). * indicates an accession showing admixture (Li et al. 2022). Grey shading indicates the four *A. hypochondriacus* accessions.

| | SNPs | | | | SVs | | | |
|---|---|---|---|---|---|---|---|---|
| Accessions | Total # | SNP density (SNPs /kb) | # in protein coding region | # of gene loci with SNP(s) | PAV | INV | CNV | TRANS |
| V1044371 | 88,253 | 0.22 | 2511 | 824 | 22621 | 97 | 3093 | 188 |
| V1044437 | 105,345 | 0.26 | 3224 | 1,142 | 22860 | 101 | 3060 | 188 |
| V1061487* | 455,149 | 1.10 | 19314 | 6,336 | 31849 | 110 | 3842 | 468 |
| V1061494* | 1,118,750 | 2.69 | 40584 | 13,542 | 45210 | 138 | 5362 | 932 |
| V1050446 | 1,677,395 | 3.99 | 72845 | 17,734 | 63665 | 210 | 4833 | 1,160 |
| V1060472* | 1,541,975 | 3.81 | 66835 | 16,262 | 59253 | 200 | 4729 | 1,054 |
| V1062433 | 1,718,562 | 4.18 | 74016 | 17,914 | 64177 | 224 | 4916 | 1,230 |
| V1063769 | 1,673,528 | 4.02 | 76143 | 17,852 | 68351 | 208 | 5573 | 1,280 |
| Total | 8,378,957 | | 355,472 | 91,606 | 377,986 | 1,288 | 35,408 | 6,500 |

The four *A. hypochondriacus* genomes showed a relatively even distribution of SNPs along the genome (Figure 4b), although VI060472 (genetic admixture) contains some clear regions of reduced SNP density (e.g. regions on chromosome 1, 6, 11 and 14) which may represent regions inherited from an *A. cruentus* ancestor. The four *A. cruentus* accessions showed different patterns of variation density. VI061494 (admixture) showed an even distribution of SNPs along the genome (similar to the *A. hypochondriacus* accessions) whereas the VI061487 genome (admixture) showed specific regions of high SNP density. One region of variation (on chromosome 10) was clearly shared across three of the *A. cruentus* accessions (Figure 4b), with a large proportion of SNPs in the two accessions most closely related to the reference line (Supp Figure 4c) concentrated in this region.

Of the 8,378,957 SNPs identified, 355,473 were located within protein-coding genes, and could cause gain or loss of start or stop codons and splice donor or acceptor sites, as well as frameshift or premature stop codons. Figure 4c shows the number of gene families that contain SNPs in one or more of the amaranth accessions with 349 core gene families having SNPs in all accessions compared to the *A. cruentus* reference genome. The distribution of SNPs within protein-coding regions (Figure 4d) follows a similar distribution to that of all SNPs (Figure 4b). Variation within single copy genes could have a more significant impact on plant phenotype. Of the 12,755 single copy core gene families, 11,776 contain SNPs in one or more of the MAGIC parent lines. The apparent "hotspot" for variation on chromosome 10 is in the gene-dense (low TE) region of this chromosome (Ma *et al.* 2021) and many of the core gene families (and single copy core gene families) containing SNPs in each accession are located here (Figure 4d).

A total of 421,182 structural variants (SVs) of up to 100 kb were identified including 377,986 presence and absence variations consisting of 231,633 insertions (presence) and 146,353 deletions (absence) with respect to the reference. In addition, 1,288 inversions, 35,408 copy number variations and 6,500 translocations were identified amongst the MAGIC founder lines (Table 3). The number of SVs in each accession is much lower than the number of SNPs as expected, but follows a similar pattern of abundance (Supp Figure 7a). Both deletions and insertions were relatively evenly distributed across the genome with a small number of "hotspots". The hotspots for deletions and insertions were generally not co-located with the exception of regions on chromosome 10, 1 and 2b (Supp Figure 8a, b). Inversions appeared to be more frequently co-located across the accessions (Supp Figure 8c), perhaps indicating a key evolutionary difference between the MAGIC population parents and the *A. cruentus* line chosen as reference.

The SVs were abundant (62-67%) in repeat regions of the genome, while 33–38% of SVs in each accession overlapped genic and flanking regions (5 kb each side of a gene), suggesting a potential impact of SVs on gene content and regulation (Supp Figure 7b). To probe this further, we identified all genes that were within deleted regions (or contained a deleted region) of the parent lines compared to the *A. cruentus* reference. The number of genes ranged from 242 to 4374 with larger numbers in the *A. hypochondriacus* accessions as expected (Supp Figure 9a). Nearly half of the genes identified were shared by just the *A. hypochondriacus* accessions, with the rest showing a variable pattern of presence/absence (Supp Figure 9b). Genes absent or containing a deletion in *A. hypochondriacus* accessions were enriched for functions in lipid metabolism, DNA repair and response to stimulus, light and hormones, potentially impacting on the plant's response to environmental conditions. In contrast, deletions in *A. cruentus* accessions showed enrichment for functions in cellulose biosynthesis and ADP metabolism, both of which could impact yield. Genes deleted in the two *A. cruentus* accessions showing admixture were also enriched for the *A. hypochondriacus*-associated terms "response to stimulus" and "lipid metabolism".

Using Assemblytics (Nattestad and Schatz 2016) to identify large structural variants (SVs) greater than 100 kb, we identified a total of 16 deletions and 136 insertions compared to the reference *A. cruentus* genome (Table 4). Curiously, the number of insertions was higher in the *A. cruentus* accessions compared to the *A. hypochondriacus* accessions, which does not reflect their phylogenetic relationships. However, this is likely to simply reflect the relatively low number of large SVs in the genomes compared to SNPs and smaller SVs.

**Table 4: Large insertions and deletions detected in each MAGIC parent accession genome using Assemblytics.** The deletions are in comparison to the *A. cruentus* "Arusha" reference genome (Ma et al. 2021). * indicates an accession showing admixture (Li et al. 2022). Grey shading indicates the four *A. hypochondriacus* accessions.

| | Deletions | | | | Insertions | | | |
|---|---|---|---|---|---|---|---|---|
| **Accessions** | **100 - 500 Kb** | **500 - 1,000 Kb** | **> 1 Mb** | **Total (>100 Kb)** | **100 - 500 Kb** | **500 - 1,000 Kb** | **> 1 Mb** | **Total (>100 Kb)** |
| V1044371 | 0 | 0 | 0 | 0 | 19 | 8 | 2 | 29 |
| V1044437 | 1 | 0 | 0 | 1 | 18 | 6 | 1 | 25 |
| V1061487* | 0 | 0 | 0 | 0 | 15 | 7 | 0 | 22 |
| V1061494* | 1 | 0 | 1 | 2 | 14 | 4 | 3 | 21 |
| V1050446 | 2 | 1 | 2 | 5 | 8 | 2 | 1 | 11 |
| V1060472* | 0 | 1 | 1 | 2 | 5 | 3 | 2 | 10 |
| V1062433 | 1 | 1 | 2 | 4 | 6 | 2 | 2 | 10 |
| V1063769 | 1 | 0 | 1 | 2 | 6 | 1 | 1 | 8 |
| **Total** | **6** | **3** | **6** | **16** | **91** | **33** | **12** | **136** |

An amaranth graphical pangenome was constructed from the SVs identified using Minigraph Cactus (Hickey *et al.* 2023). To illustrate the value of such a pangenome, we identified two syntenic regions containing copy number variation, likely by tandem duplication and/or gene loss, of genes encoding nucleotide-binding leucine-rich repeat (NLR) proteins (Figure 5). NLR proteins act as immune receptors and are often arranged in clusters and vary considerably between and within plant species (Barragan and Weigel 2021). Information on key amaranth diseases is limited to a survey of fungal diseases (Appiah-Kubi *et al.* 2022) and damping-off in seedlings (Sealy, Kenerley and McWilliams 1988). However, the eight MAGIC parent lines do differ in susceptibility to *Pythium aphanidermatum* (a causal agent of damping-off)(Thanyasiriwat *et al.* 2025) and future analysis of the MAGIC population generated from these parent lines could enable identification of genetic variation underlying this trait. Capturing the diversity of genes and genome structure using pangenomes will aid in linking sequence and structural variation with multiple phenotypic traits and the breeding or engineering of improved crop lines.

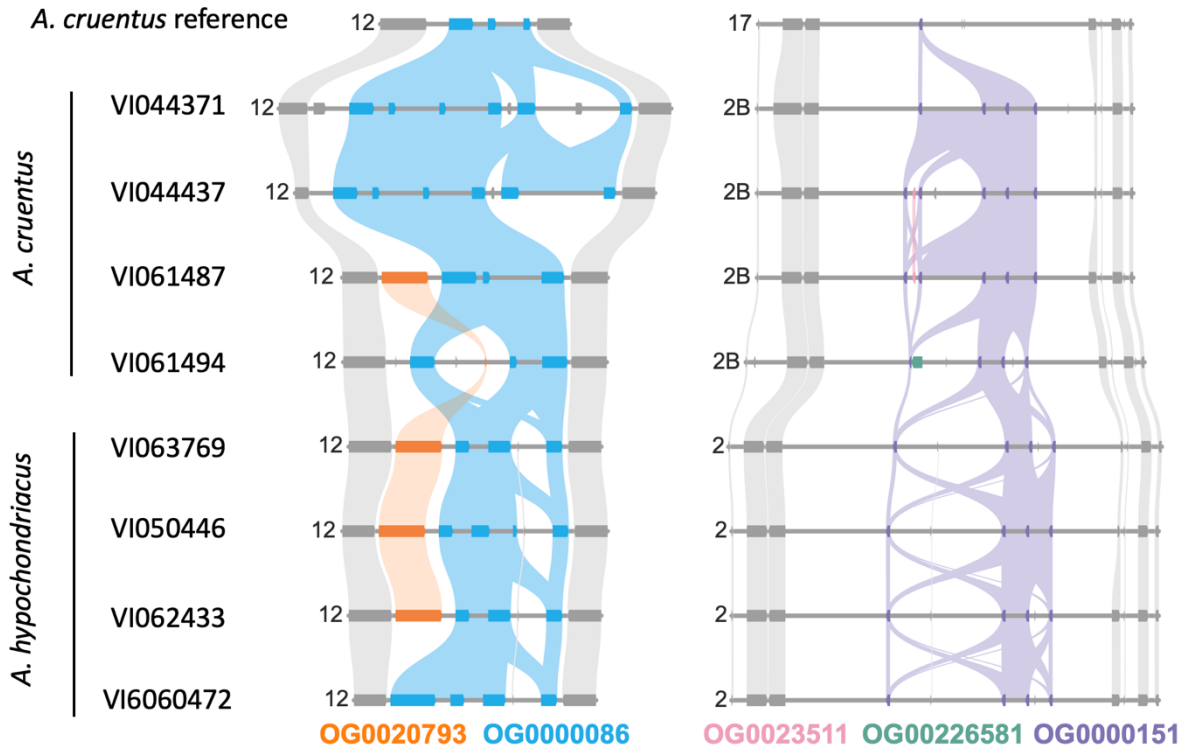



**Figure 5: Variation in copy number of putative disease resistance gene orthologues in the *Amaranthus* pangenome**. Genes encoding orthologous nucleotide binding and leucine-rich repeat (NLR) proteins in two genomic locations are indicated for each of the eight founder accessions and the *A. cruentus* reference. The corresponding orthogroup is indicated by colour, with non-NLR genes flanking the region of interest indicated in grey. The chromosomal location is indicated at the left end of each genomic region.

## Discussion

The 8 new genomes presented here represent a significant advance in the genomic resources available for amaranth, and constitute the first pangenome within this genus. Given the ability of *A. cruentus* and *A. hypochondriacus* to hybridise (Lin *et al.* 2022) we have generated an inter-specific pangenome enabling direct comparison of variants between multiple accessions of each species. These two species are thought to have been independently domesticated from *A. hybridus* but not completely as key domestication traits are still lacking (Stetter, Vidal-Villarejo and Schmid 2020). Indeed, several of the accessions in our study have black seed versus the typical cream-coloured seed of domesticated lines. As has been often stated, *Amaranthus* species are difficult to distinguish reliably based on morphology and *A. cruentus* and *A. hypochondriacus* are part of a species complex along with *A. caudatus* (another species independently domesticated from *A. hybridus*). The ability for these species to hybridise is clearly of value in capturing variation for breeding purposes and has also led to admixture within accessions. A genotype-by-sequencing approach (Lin *et al.* 2022) indicated that three of the lines used in this study (VI060472, VI061487 and VI061494A) contained genome components not typically associated with their given species classification. Consequences of this genomic admixture could be seen in the numbers of SNPs identified in each accession compared to the reference *A. cruentus* genome (Table 3) however, a PCA using the SNPs clearly distinguished the two species groups (Figure 4a).

Remarkably, a PCA using metabolite profiles from leaf tissue also clearly distinguished the two species groups (Figure 2a). This raises the possibility of using metabolite profiling (potentially cheaper, quicker and easier) for amaranth species classification where genome sequencing is not possible or too costly. Metabolite profiles from a larger diversity set with accompanying genome data would be useful to evaluate the broader value of this approach beyond the accessions in this study.

Our amaranth pangenome considerably expands the scope of known amaranth gene families with 18% of gene families (the vast majority of these just single copy genes) only detected in a single amaranth accession (Figure 3). This is a high proportion of private genes compared to many other plant pangenomes (*Raphanus*, jujube, soybean and spinach all have 1% or less (Liu *et al.* 2020; Zhang *et al.* 2021; Guo *et al.* 2024; She *et al.* 2024). This is likely due to the inclusion of two species in our pangenome, and a significant proportion of private genes coming from the reference *A. hypochondriacus* genome (Table 2). The number of dispensable/private gene families in our pangenome appears to plateau after the inclusion of 7 genomes, whereas the proportion of core gene families keeps decreasing. This is a similar pattern to that seen in a *Brassica napus* pangenome (Hurgobin *et al.* 2018) and may reflect additional genomes adding new gene variants rather than completely novel genes/gene families. At 51%, the proportion of core gene families (those present in all 9 accessions) was within the range seen in other plant pangenomes (Tao *et al.* 2019). However, this proportion does vary widely and is not necessarily correlated with the number of accessions included, for example, 24% in a Setaria pangenome from 110 accessions (He *et al.* 2023), 60% in an Arabidopsis pangenome from 69 accessions (Lian *et al.* 2024), and 35% in a pangenome assembled using 8 jujube accessions (Guo *et al.* 2024). The proportion of core gene families is still decreasing as more genomes are integrated (Figure 3c) indicating that this collection of 9 lines has not captured the full range of amaranth diversity and it is likely that incorporation of additional genomes will reduce the proportion of core genes below 51%.

The pangenome not only expands our knowledge of amaranth genes, but highlights sequence and structural variation that is likely to impact gene function and regulation. The density of SNPs identified in the eight MAGIC founder lines (compared to the *A. cruentus* reference genome) varied from 0.22 to 4 SNPs per kb (Table 3). Although only 4% of the SNPs were in protein-coding genes, SNPs were identified in one or more of the amaranth lines in 20,221 of the 23,796 gene families in the pangenome (Figure 4c). The density of SVs was significantly lower (0.06 to 0.18 SVs per kb) and the majority of these were found in repeat regions of the genome, however, at least 1/3 of the SVs in each genome are found overlapping with coding genes and/or their 5 kb flanking regions (Supp Figure 7) suggesting these SVs would impact on presence/absence and/or gene regulation. Further sequencing of the amaranth accessions using methods that capture physical association between DNA sequences (e.g. Hi-C) would likely identify additional SVs and particularly enable us to pinpoint larger genomic rearrangements between the lines.

The 8 genomes that have been newly sequenced, assembled, annotated and compared to the reference *A. cruentus* genome, are the founders of a MAGIC population. Our inter-specific pangenome will therefore be of immediate value for genotyping the population and facilitating the mapping of phenotypic traits. The value of a MAGIC population is dependent on the genetic and phenotypic diversity captured within the parent lines. Genetic diversity is highlighted above both in terms of variants and the presence/absence of gene families.

Although several accessions exhibit genomic admixture, we anticipate this will have minimal impact on trait mapping as all accessions show very low levels of heterozygosity (Supp Dataset 4). Our study also demonstrates that the founders vary in key phenotypic traits that are relevant to the cultivation and consumption of amaranth as a leafy vegetable. These include leaf yield, as well as leaf size (length/width) and flowering time (i.e. days to flowering). Smallholder farmers in South Africa prefer fewer larger leaves (rather than more smaller leaves) for ease of harvesting. Similarly, the branching habit of VI063769 (many branches from the base of the stem rather than branches up the stem) makes manual harvesting easier and quicker. A recent study using six of the parent lines demonstrated variation in drought tolerance (assessed by leaf yield under different soil water regimes)(Khoza *et al.* 2025) . The MAGIC founder lines differ in content of nutritionally-relevant metabolites (Figure 2). More than 3000 non-redundant features could be detected in our metabolomic analysis, but only 169 of these could be annotated (of which 122 differed between the amaranth accessions). The annotation rate (5%) reflects a current assessment of annotation ability in plant metabolomics (Alseekh and Fernie 2023) but is lower than studies in other underutilised crops such as jujube (Zhang *et al.* 2022b). This may be due to different annotation methods; we applied a stringent annotation workflow by manually reconciling MS1 matches against a restricted plant-derived database to chemical class matches assigned using MS2 fragment data from Sirius 4.0 (Dührkop *et al.* 2019). Some of the metabolite classes differing between the eight amaranth accessions (e.g. polyphenolics) are of importance to human nutrition but also play important roles in flavour and in defence against pathogens and herbivory (Stiller *et al.* 2021). Both roles are of relevance to the use of amaranth in food systems.

In conclusion, the inter-specific pangenome generated here from diverse amaranth accessions, expands the known genetic repertoire of amaranth, highlights variation in gene families between accessions, and provides a valuable resource for the investigation of the genetic basis underlying agronomic traits. The use of two amaranth species captures greater genetic (including gene family) diversity than one species with the potential to identify the genomic mechanisms underlying the striking species differences in metabolite profiles.

## Acknowledgments

The project was funded by a Biotechnology and Biological Sciences Research Council award to KD, SV, MB and SM (BB/R020345/1) and Grand Challenges Research Fund QR Funding distributed by the University of York. We thank the African Orphan Crop Consortium (AOCC, http://africanorphancrops.org/) for their collaboration and support of the genome sequencing. YVdP acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). MdK is funded by the BBSRC White Rose Doctoral Training Partnership grant (BB/T007222/1). WorldVeg was supported by the Ministry of Agriculture Taiwan and by the long-term strategic donors Taiwan, UK aid from the UK government, United States Agency for International Development (USAID), Australian Centre for International Agricultural Research (ACIAR), Germany, Thailand, Philippines, Korea, and Japan. We acknowledge Dr. Oanh Nguyen from the University of California, Davis Genome Center for her input and sequencing of the genomes.

## Author Contributions

The study was conceived and designed by KD, AvD, RS, SM, WJvR, AG, MB, SV and YvdP. Experimental work was carried out by FV, FD, LC-y, RS, TL and MdK, with genome assembly by YL. Genome annotation, pangenome construction and comparative analyses

were carried out by XM, JC, TS, SS and MdK, with input from DJ and metabolite analysis by MdK and TL. Funding for the research was obtained by KD, SV, MB, SM, YVdP and AvD. The manuscript was written by XM, MdK and KD with input from all authors.

## Data availability statement

Data presented in this paper can be found in the National Center for Biotechnology Information under Bioproject PRJNA1193012 (genome sequencing, assembly and RNA sequencing data) and in the MASSIVE database under datasets MSV000096703, MSV000096711, MSV000096712, MSV000096713, MSV000096714 and MSV000096870 (as outlined in the Supplementary Information).

All genome assemblies, annotations and protein-coding genes are deposited in the ORCAE database at https://bioinformatics.psb.ugent.be/gdb/Amranthus/

All sequence and structural variants, as well as the graphical pangenome, will be available at Zenodo (submission underway).

## Conflict of interests

The authors declare no conflicts of interest.

## Supplementary information

Supplementary Methods: Detailed Metabolite Profiling methodology

Supplementary Data 1; Data from Lin et al. (2022) for the 8 MAGIC parent accessions.

Supplementary Data 2; MAGIC population parent accession field trial information and data.

Supplementary Data 3; Features detected from metabolome profiling of the 8 amaranth MAGIC parent accessions.

Supplementary Data 4; Sequencing statistics for the 8 *Amaranthus* accessions.

Supplementary Data 5; Repetitive element content of the 8 Amaranth MAGIC population parent genomes and the *A. cruentus* reference.

Supplementary Data 6; Orthologous gene families and single copy genes (orthogroups) within the 8 amaranth MAGIC population parent lines and the reference *A. cruentus* and *A. hypochondriacus* genomes.

Supplementary Data 7; Gene Ontology terms showing significant enrichment in genes from the core, dispensable and private pangenome components.

Supplementary Dataset 8: Custom metabolite database for annotation.

Supplementary Figure 1; Leaf trait data from the irrigated field trial of the eight MAGIC parent accessions.

Supplementary Figure 2; Accession and metabolite sample clustering based on normalised data for 3256 non-redundant detected features.

Supplementary Figure 3; Clustering of 122 annotated metabolites based on their abundance across samples from the 8 Amaranth MAGIC population founder lines.

Supplementary Figure 4; *Amaranthus* MAGIC population parent accession genome assembly and annotation.

Supplementary Figure 5; Co-linearity and phylogenetic relationships between the eight Amaranth MAGIC population parent genomes.

Supplementary Figure 6; Analysis of pangenome components.

Supplementary Figure 7; Structural variants in the eight MAGIC population founder lines.

Supplementary Figure 8; Density of structural variants in the eight MAGIC population founder accession genomes.

Supplementary Figure 9. Genes contained within deletion regions (or containing deleted regions) of the MAGIC parent genomes compared to the *A. cruentus* reference genome.

## References


Abeel T, Parys TV, Saeys Y *et al.* GenomeView: A Next-Generation Genome Browser. *Nucleic Acids Research* 2012;**40**:e12.

Achigan-Dako EG, Sogbohossou OED, Maundu P. Current Knowledge on *Amaranthus* Spp.: Research Avenues for Improved Nutritional Value and Yield in Leafy Amaranths in Sub-Saharan Africa. *Euphytica* 2014;**197**:303–17.

Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. 2023, DOI: 10.18129/b9.bioc.topgo.

Alonge M, Lebeigle L, Kirsche M *et al.* Automated Assembly Scaffolding Using RagTag Elevates a New Tomato System for High-Throughput Genome Editing. *Genome Biology* 2022;**23**:258.

Alonge M, Wang X, Benoit M *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 2020;**182**:145-161.e23.

Alseekh S, Fernie AR. Expanding Our Coverage: Strategies to Detect a Greater Range of Metabolites. *Curr Opin Plant Biol* 2023;**73**:102335.

Anuradha, Kumari M, Zinta G *et al.* Genetic Resources and Breeding Approaches for Improvement of Amaranth (*Amaranthus* Spp.) and Quinoa (*Chenopodium quinoa*). *Frontiers in Nutrition* 2023;**10**:1129723.

Appiah-Kubi Z, Adomako J, Dwamena H *et al.* Survey of fungal diseases associated with amaranth (*Amaranthus* species) in peri-urban vegetable farms in Kumasi and Tamale metropolis of Ghana. *Afr J Agric Res* 2022;**18**:792–802.

Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell* 2021;**33**:814–31.

Bates D, Mächler M, Bolker B *et al.* Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 2015;**67**:1–48.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Brůna T, Hoff KJ, Lomsadze A *et al.* BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* 2021;**3**:lqaa108.

Cheng H, Jarvis ED, Fedrigo O *et al.* Haplotype-Resolved Assembly of Diploid Genomes without Parental Data. *Nature Biotechnology* 2022;**40**:1332–5.

Coletta RD, Qiu Y, Ou S *et al.* How the Pan-Genome Is Changing Crop Genomics and Improvement. *Genome Biology* 2021;**22**:3.

Danecek P, Auton A, Abecasis G *et al.* The Variant Call Format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.

Danecek P, Bonfield JK, Liddle J *et al.* Twelve Years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008.

Dührkop K, Fleischauer M, Ludwig M *et al.* SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nature Methods* 2019;**16**:299–302.

Edge P, Bansal V. Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing. *Nature Communications* 2019;**10**:4660.

Edwards GE, Franceschi VR, Voznesenskaya EV. Single-cell C4 photosynthesis versus the dual-cell (Kranz) paradigm. *Annual Review of Plant Biology* 2004;**55**:173–96.

Emmanuel OC, Babalola OO. Amaranth Production and Consumption in South Africa: The Challenges of Sustainability for Food and Nutrition Security. *International Journal of Agricultural Sustainability* 2021;**0**:1–12.

Emms DM, Kelly S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biology* 2019;**20**:238.

Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families. *Proc Natl Acad Sci U S A* 2020;**117**:9451–7.

Guan D, McCarthy SA, Wood J *et al.* Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies. *Bioinformatics* 2020;**36**:2896–8.

Guo M, Lian Q, Mei Y *et al.* Analyses of Pan-Genome and Resequencing Atlas Unveil the Genetic Basis of Jujube Domestication. *Nature Communications* 2024;**15**:9320.

Gurevich A, Saveliev V, Vyahhi N *et al.* QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* 2013;**29**:1072–5.

Haas BJ, Salzberg SL, Zhu W *et al.* Automated Eukaryotic Gene Structure Annotation Using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 2008;**9**:R7.

He Q, Tang S, Zhi H *et al.* A Graph-Based Genome and Pan-Genome Variation of the Model Plant Setaria. *Nature Genetics* 2023;**55**:1232–42.

Hickey G, Monlong J, Ebler J *et al.* Pangenome Graph Construction from Genome Alignments with Minigraph-Cactus. *Nat Biotechnol* 2023:1–11.

Hurgobin B, Golicz AA, Bayer PE *et al.* Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol J* 2018;**16**:1265–74.

Jayakodi M, Schreiber M, Stein N *et al.* Building Pan-Genome Infrastructures for Crop Plants and Their Use in Association Genetics. *DNA Research* 2021;**28**:dsaa030.

Jeffares DC, Jolly C, Hoti M *et al.* Transient Structural Variations Have Strong Effects on Quantitative Traits and Reproductive Isolation in Fission Yeast. *Nature Communications* 2017;**8**:14061.

Jiang T, Cao S, Liu Y *et al.* Regenotyping Structural Variants through an Accurate Force-Calling Method. 2023:2022.08.29.505534.

Kadereit G, Borsch T, Weising K *et al.* Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C4 Photosynthesis. *International Journal of Plant Sciences* 2003;**164**:959–86.

Keilwagen J, Hartung F, Paulini M *et al.* Combining RNA-seq Data and Homology-Based Gene Prediction for Plants, Animals and Fungi. *BMC Bioinformatics* 2018;**19**, DOI: 10.1186/s12859-018-2203-5.

Khoza LN, Mavengahama S, Rensburg WSJ van *et al.* Yield and growth of different amaranth genotypes under varying water regimes. *Acta Hortic* 2025:425–36.

Kim D, Paggi JM, Park C *et al.* Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 2019;**37**:907–15.

Kovaka S, Zimin AV, Pertea GM *et al.* Transcriptome Assembly from Long-Read RNA-seq Alignments with StringTie2. *Genome Biology* 2019;**20**:278.

Lian Q, Huettel B, Walkemeier B *et al.* A Pan-Genome of 69 *Arabidopsis thaliana* Accessions Reveals a Conserved Genome Structure throughout the Global Species Range. *Nature Genetics* 2024;**56**:982–91.

Lightfoot DJ, Jarvis DE, Ramaraj T *et al.* Single-Molecule Sequencing and Hi-C-based Proximity-Guided Assembly of Amaranth (*Amaranthus hypochondriacus*) Chromosomes Provide Insights into Genome Evolution. *BMC Biology* 2017;**15**:74.

Lin Y-P, Wu T-H, Chan Y-K *et al.* De Novo SNP Calling Reveals the Genetic Differentiation and Morphological Divergence in Genus *Amaranthus*. *Plant Genome* 2022;**15**:e20206.

Liu Y, Du H, Li P *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* 2020;**182**:162-176.e13.

Ma X, Vaistij FE, Li Y *et al.* A Chromosome-Level *Amaranthus cruentus* Genome Assembly Highlights Gene Family Evolution and Biosynthetic Gene Clusters That May Underpin the Nutritional Value of This Traditional Crop. *The Plant Journal* 2021;**107**:613–28.

Mabhaudhi T, O'Reilly P, Walker S *et al.* Opportunities for Underutilised Crops in Southern Africa's Post–2015 Development Agenda. *Sustainability* 2016;**8**:302.

Marçais G, Delcher AL, Phillippy AM *et al.* MUMmer4: A Fast and Versatile Genome Alignment System. *PLoS computational biology* 2018;**14**:e1005944.

Monjotin N, Amiot MJ, Fleurentin J *et al.* Clinical Evidence of the Benefits of Phytonutrients in Human Healthcare. *Nutrients* 2022;**14**:1712.

Montgomery JS, Giacomini D, Waithaka B *et al.* Draft Genomes of *Amaranthus tuberculatus*, *Amaranthus hybridus*, and *Amaranthus palmeri*. *Genome Biology and Evolution* 2020;**12**:1988–93.

Nattestad M, Schatz MC. Assemblytics: A Web Analytics Tool for the Detection of Variants from an Assembly. *Bioinformatics* 2016;**32**:3021–3.

Nyonje WA, Yang R-Y, Kejo D *et al.* Exploring the Status of Preference, Utilization Practices, and Challenges to Consumption of Amaranth in Kenya and Tanzania. *Journal of Nutrition and Metabolism* 2022;**2022**:2240724.

Patro R, Duggal G, Love MI *et al.* Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat Methods* 2017;**14**:417–9.

Paysan-Lafosse T, Blum M, Chuguransky S *et al.* InterPro in 2022. *Nucleic Acids Research* 2023;**51**:D418–27.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes. *Nat Commun* 2020;**11**:1432.

Riggins CW, Mumm RH. Amaranths. *Curr Biol* 2021;**31**:R834–5.

Sealy RL, Kenerley CM, McWilliams M EL. Evaluation of *Amaranthus* accessions for resistance to damping-off by *Pythium myriotylum*. *Plant Disease* 1988;**72**:985–9.

She H, Liu Z, Xu Z *et al.* Pan-Genome Analysis of 13 *Spinacia* Accessions Reveals Structural Variations Associated with Sex Chromosome Evolution and Domestication Traits in Spinach. *Plant Biotechnology Journal* 2024;**22**:3102–17.

Soneson C, Love MI, Robinson MD. Differential Analyses for RNA-seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Research* 2016;**4**:1521.

Stetter MG, Vidal-Villarejo M, Schmid KJ. Parallel Seed Color Adaptation during Multiple Domestication Attempts of an Ancient New World Grain. *Molecular Biology and Evolution* 2020;**37**:1407–19.

Stiller A, Garrison K, Gurdyumov K *et al.* From Fighting Critters to Saving Lives: Polyphenols in Plant Defense and Human Health. *International Journal of Molecular Sciences* 2021;**22**:8995.

Tang H, Krishnakumar V, Zeng X *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* 2024;**3**:e211.

Tao Y, Zhao X, Mace E *et al.* Exploring and Exploiting Pan-genomics for Crop Improvement. *Molecular Plant* 2019;**12**:156–69.

Thanyasiriwat T, Kawicha P, MacDonald S *et al.* Pathogenicity and genome assembly of a *Pythium aphanidermatum* isolate causing damping-off in amaranth in controlled environment agriculture. 2025, DOI: 10.1101/2025.04.10.648180.

Venskutonis PR, Kraujalis P. Nutritional Components of Amaranth Seeds and Vegetables: A Review on Composition, Properties, and Uses. *Compr Rev Food Sci F* 2013;**12**:381–412.

Wang H, Xu D, Wang S *et al.* Chromosome-Scale *Amaranthus tricolor* Genome Provides Insights into the Evolution of the Genus *Amaranthus* and the Mechanism of Betalain Biosynthesis. *DNA Research* 2022:dsac050.

Wang Y, Tang H, DeBarry JD *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**:e49.

Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000research* 2018;**7**:1338.

Zhang R, Kuo R, Coulter M *et al.* A High-Resolution Single-Molecule Sequencing-Based Arabidopsis Transcriptome Using Novel Methods of Iso-seq Analysis. *Genome Biol* 2022a;**23**:149.

Zhang X, Liu T, Wang J *et al.* Pan-Genome of *Raphanus* Highlights Genetic Variation and Introgression among Domesticated, Wild, and Weedy Radishes. *Mol Plant* 2021;**14**:2032–55.

Zhang Z, Shi Q, Wang B *et al.* Jujube Metabolome Selection Determined the Edible Properties Acquired during Domestication. *The Plant Journal* 2022b;**109**:1116–33.

Zonneveld M van, Kindt R, McMullin S *et al.* Forgotten Food Crops in Sub-Saharan Africa for Healthy Diets in a Changing Climate. *Proc National Acad Sci* 2023;**120**:e2205794120.

Zonneveld M van, Kindt R, Solberg SØ *et al.* Diversity and Conservation of Traditional African Vegetables: Priorities for Action. *Divers Distrib* 2021;**27**:216–32.

## Figure legends

**Figure 1. Leaf yield of the eight MAGIC population founder accessions in an irrigated field trial. A)** Boxplots show yield measured as fresh weight of harvested leaves per plant. Yield was harvested from three replicate plots of 20 plants with yield per plant calculated to adjust for any loss of plants in a plot. Harvest number indicates each subsequent harvest and regrowth of the plants. **B)** Total leaf yield (i.e. the sum of all five harvests) per plant per accession. Letters indicate significant differences within each harvest as assessed by Tukey's honest significance test. Initials *A.c.* and *A.h.* indicate the species of each line with '*' indicating lines with admixture.

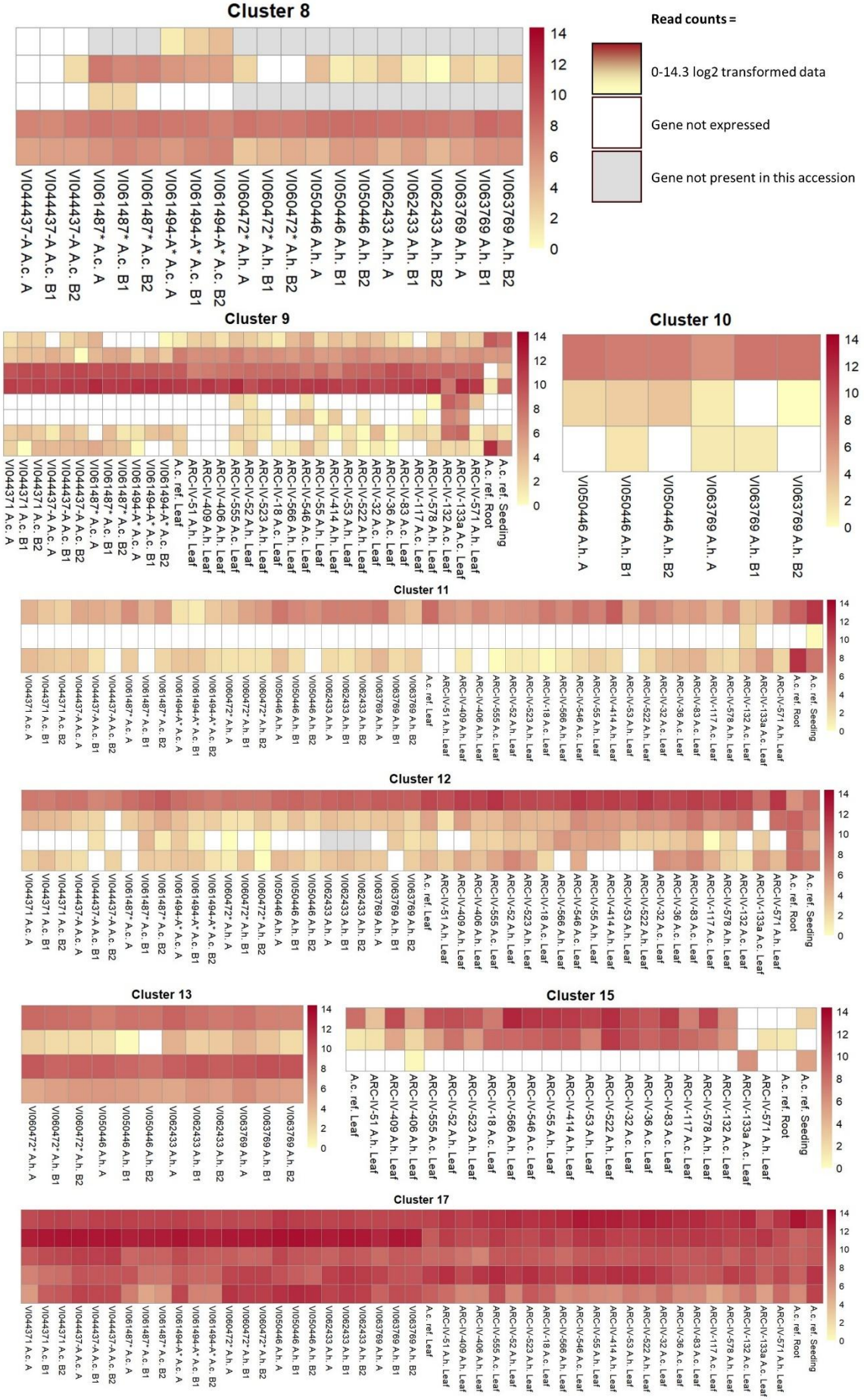
**Figure 2. Variation in metabolite content between the eight *Amaranthus* MAGIC population parent accessions. A)** Principal component analysis using the abundance of the 122 annotated metabolites significantly differing between MAGIC parent accessions. **B)** The relative abundance of the 122 metabolites across the 8 amaranth accessions organized by metabolite class. Presence absence values (PAV) based on raw data and metabolites varying significantly between the two species are indicated. '*' indicates accessions with genomic admixture.

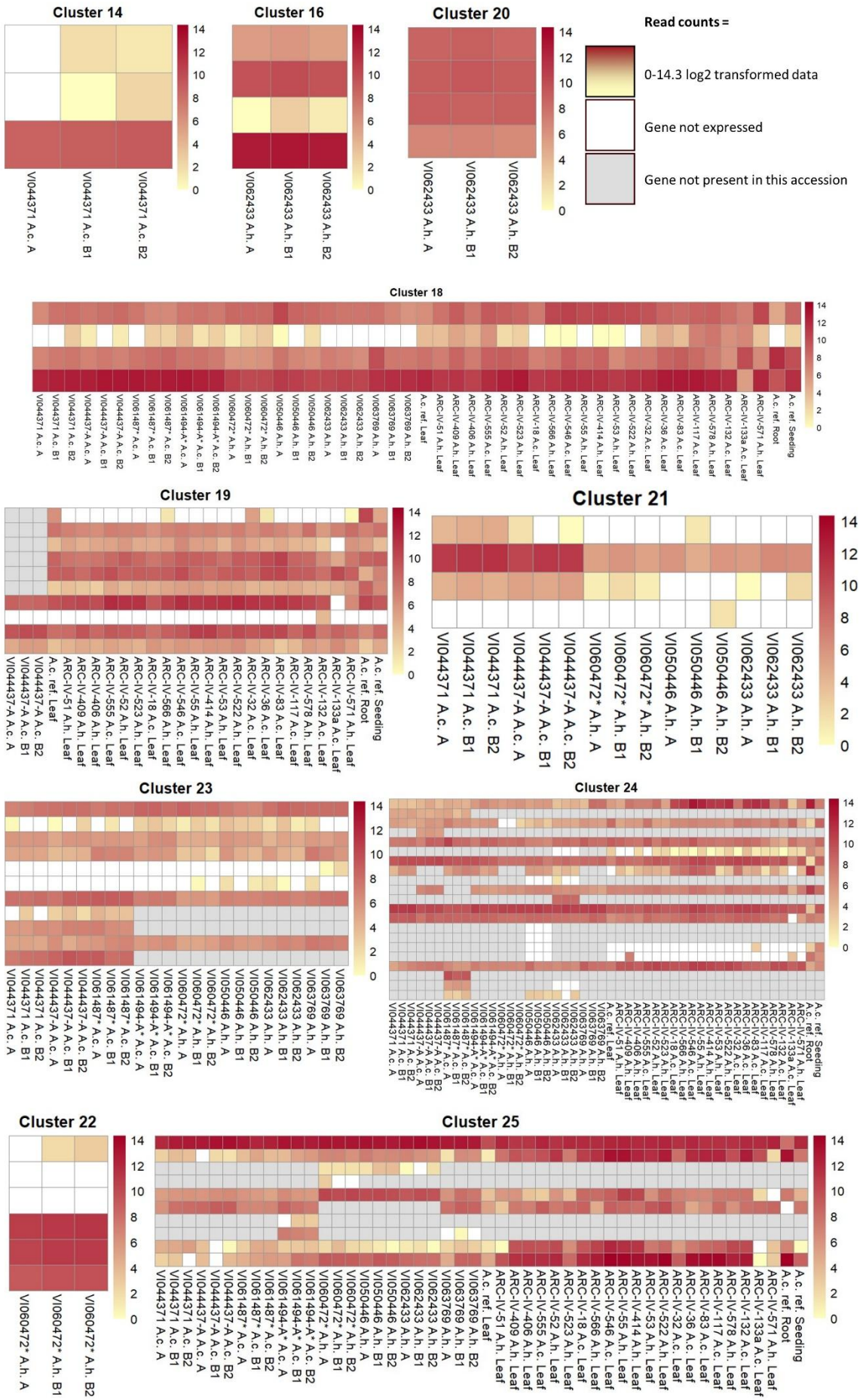
**Figure 3: Analysis of pangenome components across ten *Amaranthus* genomes (eight MAGIC population parents and two published reference genomes, *A. cruentus* and *A. hypochondriacus).* A)** The proportion of core, dispensable and private gene families (including single copy genes unique to one accession) across the ten amaranth genomes. **B**) The number of gene families shared between the ten amaranth genomes. **C**) The proportion of core and total gene families in the pangenome as the number of genomes compared increases. **D**) The proportion of genes in each genome that are classified as core, dispensable or private in the pangenome of 10 accessions. **E**) mRNA length of genes in the three pangenome components. **F)** The proportion of core, dispensable and private genes that are functionally annotated via Gene Ontology or the Pfam database.

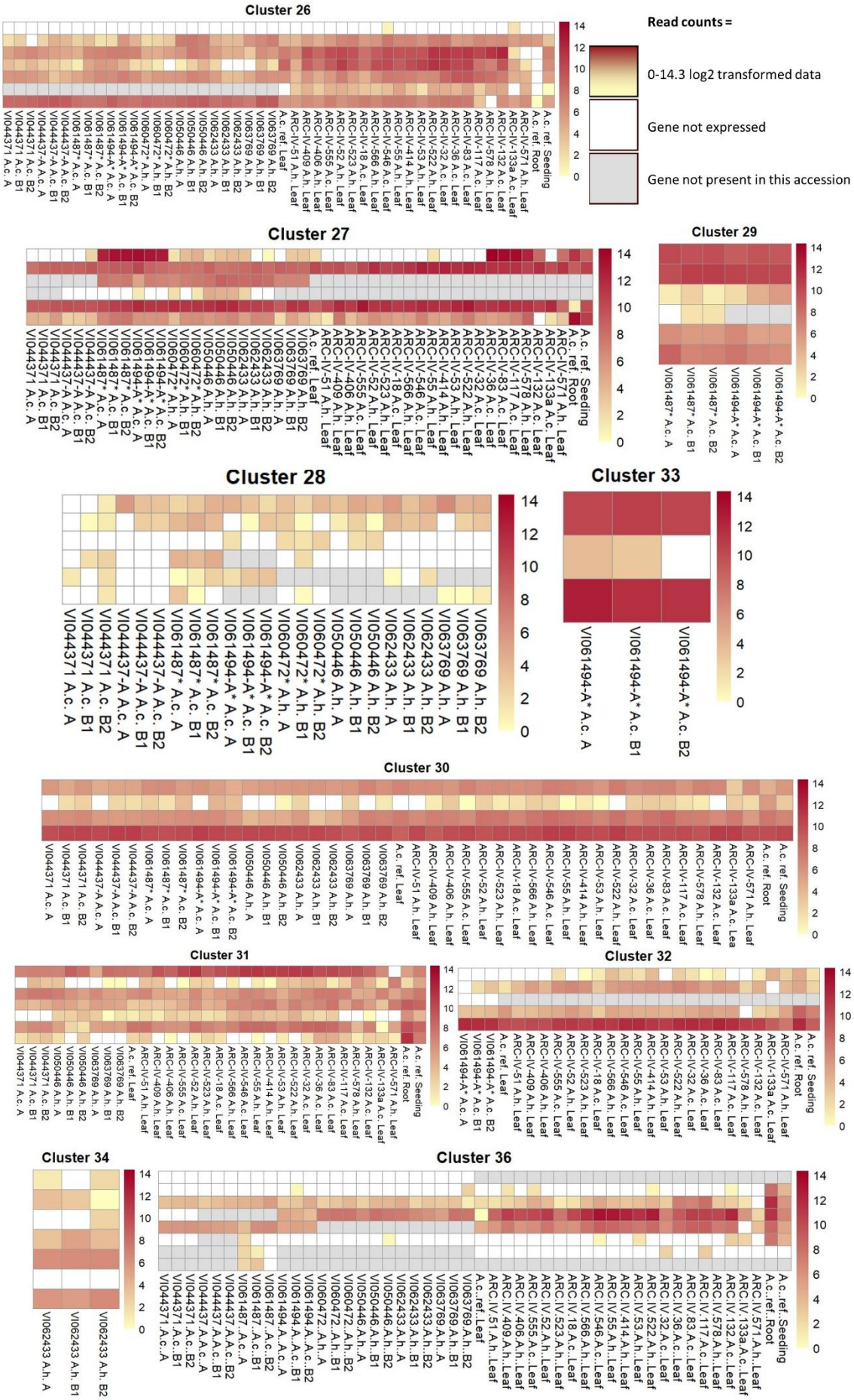
**Figure 4**. **Single Nucleotide Polymorphisms (SNPs) between the MAGIC founder lines and reference *A. cruentus* genome*.* A)** Principal component analysis using SNPs indicating clear separation of the MAGIC founder lines into two species groups. **B**) the distribution of SNPs along the genome. From the outer to inner layer is V1044371, V1044437, V1061487, V1061494, V1050446, V1060472, V1062433 and V1063769. **C**) the number of gene families that contain SNPs in at least one of the eight MAGIC founder lines. Blue indicates (core) gene families where all MAGIC parents contain SNPs in the genes compared to the *A. cruentus* reference. **D**) the distribution of genes containing SNPs along the genome. Green is all genes with SNPs for each accession (accessions in the same order as in B). Blue indicates core genes with SNPs in all accessions and orange is single copy core genes with SNPs in all accessions.

**Figure 5: Variation in copy number of putative disease resistance gene orthologues in the *Amaranthus* pangenome**. Genes encoding orthologous nucleotide binding and leucine-rich repeat (NLR) proteins in two genomic locations are indicated for each of the eight founder accessions and the *A. cruentus* reference. The corresponding orthogroup is indicated by colour, with non-NLR genes flanking the region of interest indicated in grey. The chromosomal location is indicated at the left end of each genomic region.

# Figures


Cluster 8
Read counts =
0-14.3 log2 transformed data
Gene not expressed
Gene not present in this accession
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 15
Cluster 17


Cluster 14
Cluster 16
Cluster 20
Read counts =
0-14.3 log2 transformed data
Gene not expressed
Gene not present in this accession
Cluster 18
Cluster 19
Cluster 21
Cluster 23
Cluster 24
Cluster 22
Cluster 25


Cluster 26
Read counts =
0-14.3 log2 transformed data
Gene not expressed
Gene not present in this accession
Cluster 27
Cluster 29
Cluster 28
Cluster 33
Cluster 30
Cluster 31
Cluster 32
Cluster 34
Cluster 36

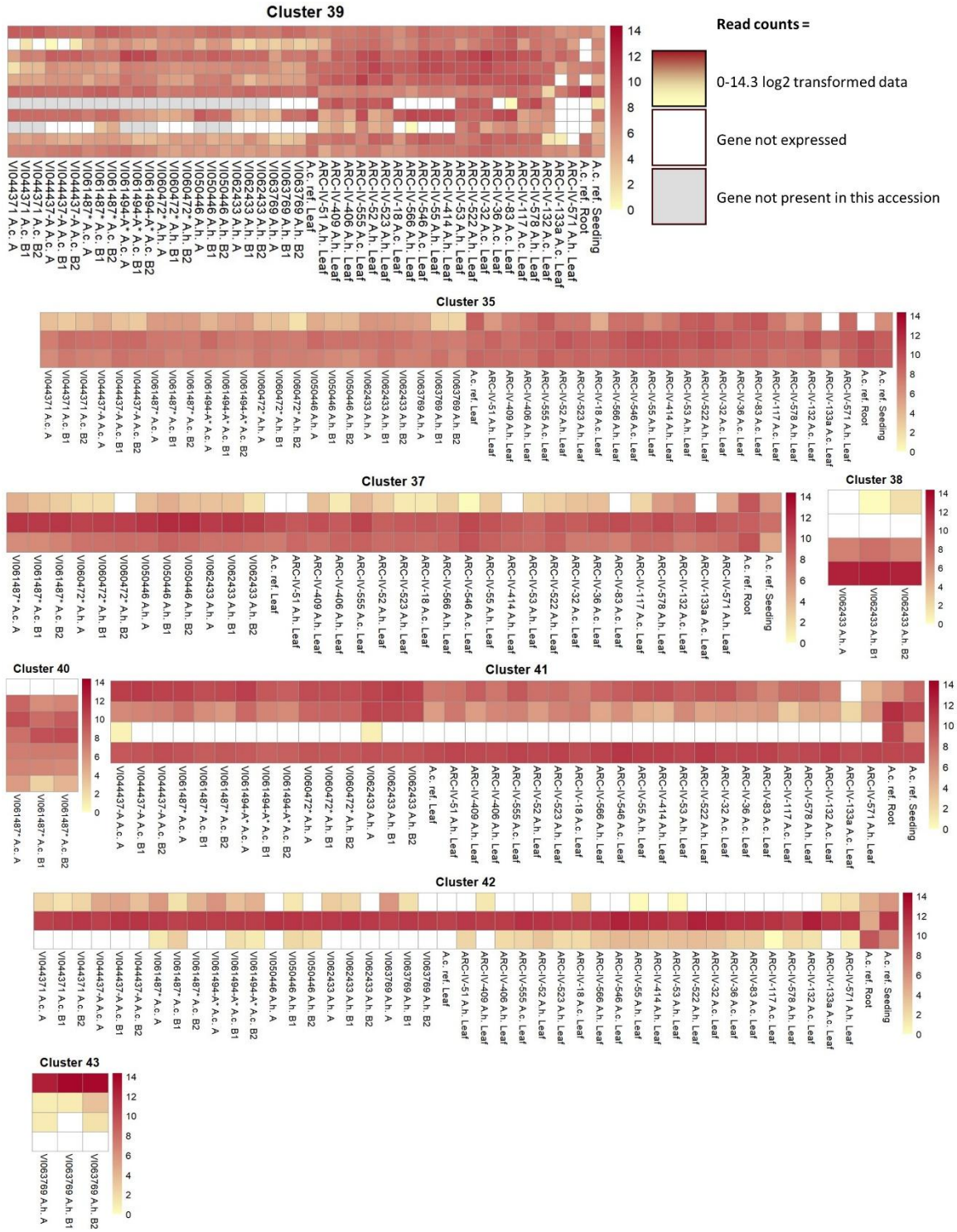



**Figure Appendix.** RNA expression heatmaps of genes in 43 putative biosynthetic gene clusters (BGC) (**Supp table BGC**) identified in one to 30 different accessions. Samples were taken from leaf tissue only with the exception of Arusha, i.e., *A. cruentus* reference (Ma *et al.*, 2021), for which it was also taken at root and seedling tissue. The three leaf samples of the MAGIC parents are physically different samples (A and B) and a repeated sample for B (B1 and B2). The 21 diversity set accessions are labelled with the assumed *A. h.* or *A. c.* species but ordered by their new species classification as indicated in (**Figure 5.4**). Expression patterns are therefore less consistent towards the right side of the heatmaps with four non *A. h.* or *A. c.* diversity set accessions and two non-leaf tissue samples of Arusha. Clusters of interest with similar expression patterns are highlighted in **Figure 5.5**.