Variant Detection in *Brassica Napus*

Katarina Gmeiner

MSc by Research

University of York

Biology

April 2022

Abstract

Background

The advent of Next Generation Sequencing (NGS) made molecular markers, such as Single Nucleotide Polymorphisms (SNPs) more commonly used. However, detection of SNPs in polyploid genomes may be challenging i.e. due to higher volumes of repetitive DNA. This project aims to create a computational pipeline for variant detection in the polyploid genome of *Brassica Napus*.

Results

A total of 304 Single Nucleotide Polymorphisms (SNPs) and 16 Insertions or Deletions (INDELs) were detected across 6 genes. For sequencing, 4 SNPs and 2 INDELs were selected. Despite the latter variation not passing the penultimate filter, sequencing data did reveal a 4bp deletion at position 153 of the FAE1 gene.

Conclusion

The presence of variation, which was computationally filtered out, indicates that too stringent filters were used. However, we demonstrated that our pipeline was able to accurately discard low-quality variants. To mitigate the effects of lenient filtering methods, we suggest further separating the pipeline into independent SNPs and INDELs pipelines to apply variation-specific filters. Moreover, we showed that FASTA clusters could be used as an effective tool to gain insights into complex genomes.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest appreciation to the entire Bancroft group – thank you for guiding me through this project with your patience and expertise.

I would also like to extend my thanks to my friends and family, who tirelessly supported me throughout my time at university.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

1. Introduction

1.1. The Need for Accelerated Crop Improvement

Since 2014 global undernourishment has been on the rise. It is estimated that between 720 and 811 million people globally faced hunger in 2020 (FAO 2021). Approximately one in three persons are affected by malnutrition, such as undernutrition, micronutrient deficiencies and/or obesity (FAO 2018). The 2030 Agenda for Sustainable Development set the goal (Sustainable Development Goal (SDG) #2) to end hunger and malnutrition by 2030 globally. However, given the current trends, this goal seems unattainable: All forms of malnutrition are expected to increase to one in two persons by 2030 worldwide (FAO 2018). The Global Hunger Index suggests that even a *low* global level of hunger will be impossible to achieve (von Grebmer et al. 2021).

Different factors are contributing towards food security. The FAO highlighted 4 key issues impacting global food security: Conflict, climate variability and extremes, economic slowdowns and downturns and the unaffordability of healthy diets. Three of these factors are exacerbated by poverty and inequality (FAO 2021). Lenaerts et al. (2019) addressed population growth and climate change as key issues: Due torising population levels, the demand for food and land is rising (Nations 2015).

Both reports mentioned climate change as a contributing factor towards food insecurity. Adverse effects of climate change may manifest in increasing levels of CO2 (Peng et al. 2004) and pests and diseases (Newton et al. 2011). Extreme weather events, such as floods and droughts, are also exaggerated by climate change (Hay et al. 2016). Altogether, climate change is likely to negatively impact crop production globally (Lobell and Gourdji 2012).

With various factors pressuring food security, there is a need for food production to be both resilient and sustainable (Smith 2013). While it is necessary to elevate levels of food production, future demand for cereal plants may not be met, even with an increase in productivity (Ray et al. 2013). In this light, a transformation of the current food system is required. As such, food production needs to become more efficient and productive. Food productivity relies on the development of new technologies, such as new crop varieties. However, plant breeding is a time consuming process. Generally, developing and releasing a new rice variety takes at least 10 years (Acquaah 2012). Conventional breeding methods will not be sufficient to meet future demands of crop production (Lenaerts et al. 2019).

Since the domestication of plant crops, plant breeding contributed towards the development of new cultivars. Pre-genomic breeding led to the development of new cultivars, which successfully improved the yield of most major crops in the 20th century. This success is due to the usage of natural and mutant induced genetic variation and the efficient selection of favourable traits. While plant breeding in the 20th century relied on the evaluation and identification of genetic variation based mostly on the phenotype, genetics has been revolutionising plant breeding of the 21st century (Pérez-de-Castro et al. 2012). New tools and techniques allow the study of the genotype and its relationship with the phenotype (Tester and Langridge 2010).

New breeding techniques are under the control of biotechnology, the study and use of DNA markers. Utilising molecular markers accelerates the development of new cultivars, as linked agronomic traits can be identified and used during the selection process in breeding programmes (Xia et al. 2019; Senthilvel et al. 2019). DNA markers have been used for a variety of crops, such as rice (*Oryza sativa;* Mackill et al. 1999), corn (*Zea mays;* Ortiz 2010), wheat (*Triticum aestivum;* Suwarno et al. 2015), and tomatoes (*Lycopersicon esculentum;* Landjeva et al. 2007).

1.2. Development of Marker-Assisted Technologies

With several factors pushing for more efficient food production, the need for new technologies is urgent. Genetic markers are one such technology: They can be utilised to select favourable genotypes and traits, which are difficult to measure using phenotype assays, and to eliminate linkage drag in backcrossing (Appleby et al. 2009). Further, they are also used to link diseases associated with certain mutations, paternity assessments and forensics (Collins et al. 2004). Lastly, they are crucial in the development of genome maps and haplotypes for regions of interest (Rafalski 2002).

Hedrich describes genetic markers as “specific DNA sequences with a known location on a chromosome” (Hedrich 2012). Genetic polymorphisms are directly impacting the availability of genetic markers. Polymorphisms are changes in the DNA sequence with a frequency greater than 1% (Hedrich 2012). However, such variation is often not visible on a phenotypic level. Therefore, several marker systems have been developed to exploit and analyse genetic markers (Lateef 2015; Appleby et al. 2009).

1.2.1. First Generation Markers

One of the first markers developed were termed Restriction Fragment Length Polymorphisms (RFLPs). Botstein et al. (1980) showed how cloned pieces of DNA could be used as genetic markers and so, RFLPs became widely used in the 1980s and 1990s (Lateef 2015). RFLPs utilise restriction enzymes, which cut DNA at specific sites. Changes in the DNA interferes with restriction enzymes, resulting in differently sized DNA fragments. Both single nucleotide changes and insertions/deletions can be detected with RFLPs. Some advantages of RFLPs are their co-dominance, easy reproducibility and high locus-specificity (Kochert 1991; Lateef 2015). Further, the variety of restriction enzymes allows adjustment to the experiment’s conditions and needs (length, symmetry, AT or GC bias, methylationsensitivity). For instance, choosing methylation-sensitive restriction enzymes avoids cutting repetitive sequences in plants (Davey et al. 2011).

However, RFLPs usage has been declining over the past decade. Newer techniques have been established, which are less time consuming, require less amounts of pure DNA and have simpler procedures (Kochert 1991; Lateef 2015).

1.2.2. Second Generation Markers

The development of new markers is tightly linked to the progression of sequencing technology. The advent of Next Generation Sequencing (NGS) led to an improvement of sequencing technologies. DNA Sequencing over the years has become less expensive, less laborious and more efficient. The abundance of genomic data and sequencing technologies revolutionised agricultural genomics, including complex plant species with limited public resources. While more sequencing data is being collected, new and improved techniques are required to analyse such data (Davey et al. 2011; Deschamps et al. 2012; Chung et al. 2017).

The second generation of markers has been developed with the increasing popularity of second generational sequencing methodology. The previous generation, Sanger sequencing (Sanger and Coulson 1975), utilised radioactive agents and slab gels and was slow by NGS standards. The second generation of sequencing allows to run multiple reactions in parallel, dramatically reducing the cost of sequencing. Although different second generation sequencing methods differ from each other, Sequencing by Synthesis (SBS), for instance, has a higher error rate compared to Sanger sequencing and produces much shorter reads (300-500 bases; Slatko et al. 2018).

Some of such second generational markers include Simple Sequence Repeats (SSRs) and Expressed Sequence Tags (ESTs). The former, also known as microsatellites, became a widely used marker in the 1990s. Long before the advent of SSRs, it was known that eukaryotic DNA contains a large number of repeating sequences (Britten and Kohne 1968). SSRs exploit such tandem repeats: They are short sequences (2-6 bp) of di-, tri- and tetra-nucleotide repeats (for instance (GT)n, (GTA)n and (GATA)n). These repeats often surround highly conserved DNA sequences, which is why primer specificity is a crucial requirement of SSRs. However, their resulting loci show high levels of allelic variation, making them valuable markers. SSRs markers are easily analysed by PCR and detected in high-resolution electrophoresis systems (eg. PAGE or AGE; Jiang 2013). Polymorphisms in SSRs are detected by varying number of repeats in different genotypes, hence making penta- and tetra-nucleotide more robust systems, as variation is easier to detect in longer repeats (Ellegren 2000; Koelling et al. 2012; Lateef 2015).

Advantages of SSRs markers include their hyper-variability, co-dominance, locus-specificity and reproducibility. Further, they only require small amounts of DNA samples (~100ng) and are low-priced for manual assays. However, SSRs marker development itself is laborious and expensive for large-scale, automated methods (Mir et al. 2013; Jiang 2013).

Due to the availability of Expressed Sequence Tags (ESTs), SSRs markers can be developed in silicio for many plant species. Sequencing data from EST sequencing projects are available online and can be used to scan for SSRs (Varshney et al. 2005; Rudd 2003). The resulting markers (EST-SSR or genic microsatellites) are inexpensive, as they are virtually by-products of publicly available ESTs sequencing data (Varshney et al. 2005).

ESTs are “fragments of mRNA sequences derived through single sequencing reactions performed on randomly selected clones from cDNA libraries” (Parkinson and Blaxter 2009). The first use of ESTs was recorded in the 80s, when Putney and colleagues sequenced inserts from rabbit muscle (Putney et al. 1983). With advancements in sequencing technology, ESTs became a viable addition to sequencing projects (Adams et al. 1991): As they represent the *expressed* region of genomes, ESTs have been used for gene identification and validation of gene predictions purposes. Further, they can be used as a cost-effective alternative to full genome sequencing (Parkinson and Blaxter 2009).

However, as ESTs only contain the expressed regions of a genome, information about regulatory sequences within introns gets lost. Further, the presence or absence of introns can greatly affect the quality of an EST: Introns interrupt the coding sequence of a gene and therefore the resulting EST may differ from the original gene (Jones et al. 2009).

1.2.3. Third Generation Markers

Further advancements in sequencing technology lead to a drop in the cost of genome sequencing. This enabled more genomes to be sequenced and resequenced for analysis of genomic diversity (Mardis 2008; Schatz et al. 2010). With a wealth of genomic information, Single Nucleotide Polymorphisms (SNPs) emerged as new markers.

According to the similarity with SNPs, markers can be classified into SNPs (due to sequence variation, eg. RFLP) and non-SNPs (due to length variation, eg. SSRs; Gupta et al. 2001; Jiang 2013).

1.3. Single Nucleotide Polymorphisms and INDELs

SNPs are the simplest form of polymorphisms: A single nucleotide change between two DNA sequences in a specific location in the genome. Nucleotide bases can be classified into one-ring pyrimidines (C and T) and two-ring purines (A and G). Depending on the affected base and through mutation resulting change, SNPs can be either transitions or transversions: Transitions are mutations that do not affect the number of rings of the nucleotide base (C/T or G/A). Transversions occur when the nucleotide type is changed from purine to pyrimidine, or vice versa (C/G, A/T, C/A, or T/G; Edwards et al. 2007, Guo et al. 2017). If a variation occurs at any given position, the two possible nucleotides are said to be alleles for this position. Although 4 different variants could, in theory, be involved in a single SNP, in practice they are usually biallelic. Despite this disadvantage over SSRs, SNPs are abundant in many genomes: In humans, SNPs are one of the most common types of mutations, influencing protein coding, transcriptional regulation, alternative splicing and non-coding RNA regulation (Xu et al. 2012). In plants SNPs are estimated to appear every 100-300bp (Oraguzie et al.2007; Xu 2010).

Changes in the nucleotide sequence are inheritable, hence why SNPs are used as genetic markers (Jiang 2013), in association genetics approaches, for creating linkage maps and identification of linkage disequilibrium. Because SNPs are evolutionary stable and therefore do not change much between generations, they are ideal for understanding complex genetic traits (Syvänen 2001; Trick et al. 2009; Appleby et al. 2009).

Further, SNPs do not require the use of restriction enzymes, which aids in identification of variation in polymorphisms-dense sequences: Usage of restriction enzymes in other techniques, such as Reduced-Representation Libraries (RRL), may cause loss of markers due to changes in fragment distribution and exclusion of size selection (Davey et al. 2011; Berthelot et al. 2014). RRL-approaches utilise restriction enzymes to re-sample specific subsets of the genome across many individuals with subsequent alignment to a reference sequence (Altshuler et al., 2000). Such approaches allow simultaneous screening and sequencing of thousands of SNPs. Despite its efficiency for non-model species, restriction enzyme-based variant detection may not be suitable for highly repetitive and high ploidy genomes: Only 48% of SNPs in rainbow trout were validated using RRL. This is due to the Whole-Genome Duplication (WGD) event, which resulted in doubling of the entire genome (Davey et al. 2011; Berthelot et al. 2014; Graham et al., 2020).

A different type of mutation commonly used in genetics are INDELs: The term INDEL refers to **IN**sertions and **DEL**etions in genomic DNA (Mills 2006). Naturally occurring INDELs are considered polymorphisms and are less than 1kb in length. Any insertions or deletions longer than 1kb are considered results of duplication or DNA fusion events (Reams and Roth 2015; Sehn 2015). If an INDEL occurs in the coding region of a gene and is divisible by 3, it is considered an “in-frame” mutation. Such mutations may have little or no effect, depending on the structural properties of the inserted/deleted amino acid residue. “Frameshift” polymorphisms are INDELs which cause an alteration in the DNA reading code: Any insertion or deletion (indivisible by 3) of DNA bases may shift the reading code and therefore produce nonsense or missense mutations or translation of a premature stop codon. While missense mutations result in an amino acid change, nonsense mutations generate a premature stop codon, preventing synthesis of the full-length protein (Minde et al., 2011; Sharma, Keeling and Rowe, 2020). The latter may trigger the mRNA degradation pathway and/or truncation of the protein (Brogna et al. 2016).

SNPs have become increasingly more popular and more widely used than previously established genetic markers. One reason for this is the amount of genomic information available: SNPs require sequence information, which became more reliable and available with the advent of newer sequencing technologies (Jones et al. 2009).

Table 1: Overview of presented markers. Adapted from (Jiang 2013).

Feature / Marker	**RFLPs**	**SSRs**	**ESTs**	**SNPs**
Genomic Abundance	High	Moderate - High	Moderate	Very High
Genomic Coverage	Low copy coding region	Whole Genome	Expressed Regions	Whole Genome
Type of Variation	Single base changes, INDELs	Changes in length of repeats	Single base changes, INDELs	Single base changes, INDELs
Technically demanding	Moderate	Low	Moderate - High	High
Time demanding	High	Low	Low	Low

1.4. Brassica Napus

The family of *Brassicaceae* (or Cruciferae) includes 14 families and 4440 species (Kiefer *et al.*, 2014). One of its genus, *Brassica*, is economically important for its wide use for nutrition, oil and bio fuel (Al-Shehbaz, 2012). *Brassica napus* is especially popular worldwide: Being the second most cultivated oilseed species in the world, *B. napus* is mainly used for human consumption and animal feed purposes (Wittkop, Snowdon and Friedt, 2009; Hossain *et al.*, 2018). *B. napus* (AACC, 2n=38) is thought to be derived ~7500 years ago after an hybridization event of two diploid genomes *Brassica rapa* (AA, 2n = 20; Wang et al. 2011) and *Brassica oleracea* (CC, 2n = 18; Liu et al. 2014), followed by genome doubling (Chalhoub et al. 2014). Although no wild *B.napus* species are known (Gomez-Campo 1999), the "original" rapeseed is thought to be of winter type (Lu et al. 2019).

The Triangle of U (Nagaharu 1935, see Fig. 1) describes the Brassica family and its relationship with each other: Other allopolyploid *Brassica* species have different combinations of the diploid genomes. Artificial fusion of both *B. rapa* and *B. oleracea* genomes generates *B. napus*, although those forms are not very viable and show reduced fertility (Olsson, 2010).

Comparison of the subgenomes of *B. napus* with its orthologues in *B. oleracea* and *B. nigra* suggest formation of *B. napus* around 7500 years ago (Chalhoub et al. 2014). *B. napus* was first documented in Europe around 400 years ago as a winter crop with a biennial life cycle and strong vernalization requirement. Around 100 years later, rapeseed was grown without vernalization (Gomez-Campo 1999). Later, rapeseed was introduced to China in the 1930s as a semi-winter species with moderate vernalization requirements and to Australia and Canada in 1960s and 70s as a spring crop (Liu 1985; Chen et al. 2008; Wei et al. 2017). *B. napus* can be further divided into three ecotypes, namely winter (requiring a prolonged cold period), semi-winter (requiring a short cold period) and spring (no cold required; Lu et al. 2019). Additionally, rapeseed can be categorised according to geographical location, for instance, European winter and spring, Asian semi-winter, Australian and Canadian (Zou et al. 2019).

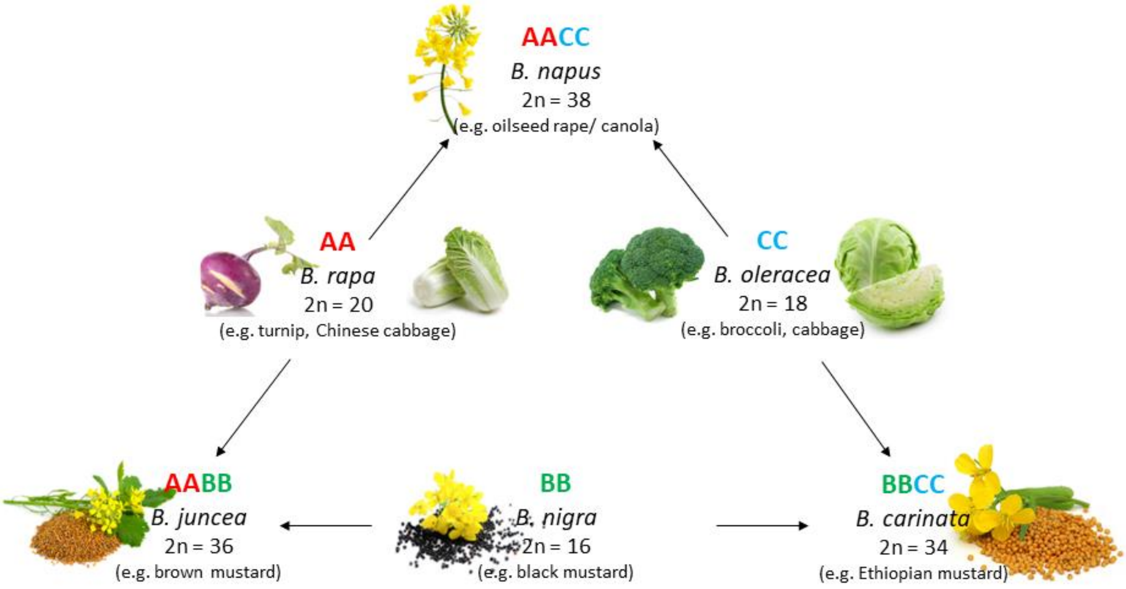


Figure 1: Triangle of U

Overview of the *Brassica* species. Hybridisation of two of three diploid ancestral species *Brassica rapa* (AA, 2n = 20), *Brassica nigra* (BB, 2n = 16) and *Brassica oleracea* (CC, 2n = 18) resulted in formation of the allopolyploid species *Brassica napus* (AACC, 2n = 38), *Brassica juncea* (AABB, 2n = 36) and *Brassica carinata* (BBCC, 2n = 34; Lu et al. 2019).

1.5. Genes of Interest

As the third largest oil crop, approximately 15% of human vegetable oil consumption stems from *B. napus* crops (Kaur et al. 2020). With a percentage of 95, Triacylglycerols (TAGs) are the main component of *B. napu*s oil. TAGs contain a glycerol backbone and three fatty acid chains: These fatty acids (FAs) differ in carbon length and saturation (Lu et al. 2019). Examples of common FAs in plants include palmitic (C16:0), stearic (C18:0), oleic (C18:1), linoleic (C18:2), linolenic (C18:3), eicosenoic (C20:1) and erucic acid (C22:1; Knutzon et al. 1992b)

The composition of fatty acids in seeds has been a focus in research for several years (Micha and Mozaffarian 2009; Gillingham, Harris-Janz, and Jones 2011). Oleic, erucic and linolenic acid have been of our particular interest (see Fig. 2):

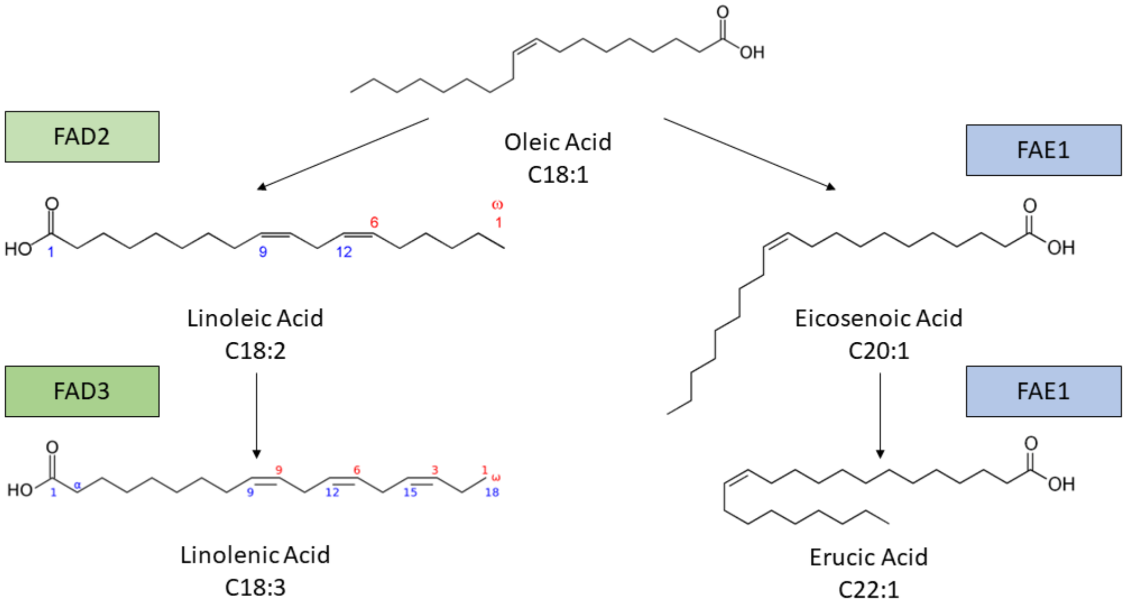


Figure 2: Oleic Acid Pathway.

Oleic acid can undergo two different pathways: Firstly, the conversion to Linolenic acid, catalysed by FAD2 and FAD3. Secondly, the elongation to Erucic acid, catalysed by FAE1 in multiple cycles.

1.5.1. Fatty Acid Desaturase 2 (FAD2)

Oils with high oleic acid levels (>75%) have several benefits, in comparison to low oleic acid levels, such as decreasing the risk of cardiovascular diseases in humans (Chang and Huang 1998) and prolonging the shelf life of the oil product, due to its antioxidant properties (Lauridsen et al. 1999). Currently, most rapeseed cultivars worldwide contain ~55–65% oleic acid (Long et al. 2018). Increasing oleic acid content by detecting novel markers is therefore a big objective in rapeseed research (Fu et al. 2021).

On the other hand, due to its oxidative properties, linolenic acid is an undesirable fatty acid: Linolenic acid is highly unsaturated and can therefore be easily oxidised, reducing the shelf life and causing off-flavour to the oil (Hu et al. 2006; Yang et al.

2012). The average canola oil contains about 20% linoleic acid and 10% linolenic acid (Hu et al. 2006). Reducing the linolenic acid content (<3%) of oils is therefore beneficial for prolonging shelf life (Wittkop et al. 2009).

In plants, fatty acids are synthesised from acetyl-CoA in plastids and later exported into the cytosol. Oil is synthesised in the endoplasmic reticulum (ER; Browse and Somerville 1991). In the stroma of plastids, 30 enzymatic reactions produce C16- and C18-carbon fatty acids, of which 75% are unsaturated (Ohlrogge and Browse 1995; Somerville 2000). Membrane-bound ER desaturases catalise the desaturation of membrane-bound phospholipids. The integral ER-membrane proteins FAD2 and FAD3 primarily desaturate additional chloroplast lipids (Los and Murata 1998; Shanklin and Cahoon 1998).
The initial desaturation step is catalysed by stearoyl-acyl carrier protein desaturase (SAD). SAD converts stearic acid (C18:0) to oleic acid (C18:1). FAD2 further desaturates oleic acid into linoleic acid (C18:2) in the ER. Finally, the conversion from linoleic acid to gamma-linolenic acid (C18:3) is catalysed by FAD3 in the ER (Zhang et al. 2012; Bhunia et al. 2016; Dar et al. 2017).

1.5.2. Fatty Acyl Coa Elongase 1 (FAE1)

Two major seed-oil types are present within *B. napus*: Low-erucic (<2%) and high-erucic acid types. Reducing erucic acid content has been the goal of several rapeseed breeding programmes (Yan et al. 2015; Zhao et al. 2019). Although erucic acid is an nutritionally unfavourable component (Badawy, Atta, and Ahmed 1994), high erucic acid cultivars are important material for industrial applications (Hristov et al. 2011). Erucic acid from high-erucic acid rapeseed can be processed into biodiesel, lubricants, surfactants, pharmaceuticals, cosmetics, soaps, rubber and nylon and has been in high demand to meet the needs for biodegradable and environmentally safe oil products (Hristov et al. 2011; Konkol et al. 2019; Lu et al. 2019). *B. napus* contains around 45-55% of erucic acid. To minimise the cost of purification, elevation of the 45% erucic acid content in *B. napus* would be an desirable outcome (Mietkiewska et al. 2007).

The membrane-bound Fatty Acyl Coa Elongase (FAE) complex catalyses the formation of long-chain monounsaturated fatty acids in the ER. In *B. napus*, oleic acid undergoes two cycles of elongation to form erucic acid (C22:1). The FAE complex catalyses four reactions per cycle:
In the first step 3-ketoacyl-CoA is generated through a condensation reaction of C18:1-CoA with malonyl-CoA. The resulting product is then reduced to a 3-hydroxyacyl-CoA derivative, followed by sequential dehydration and reduction to create the final acyl-CoA product (Katavic et al. 2002; Lu et al. 2019). FAE1 catalyses the first condensation reaction and is the rate-limiting step for erucic acid synthesis in *B.napus* (Millar and Kunst 1997).

1.6. Common Mutagens in Variant Detection

The present study utilised gamma radiation to induce mutation in *B. napus*. While gamma radiation panels are a commonly used methodology in the field of plant sciences, a range of different mutagenesis-inducing techniques exist. Examples are chemical mutagens and heavy-ion beams, which will be further discussed below.

1.6.1. Chemical Mutagens and Targeting Induced Local Lesions in Genomes (TILLING)

One method that uses chemical mutagens is Targeting Induced Local Lesions In Genomes (TILLING). TILLING is a reverse-genetics approach to identify mutations by utilising chemical agents (Till et al. 2004). PCR amplification using fluorescently labelled primers is followed by digestion by mismatch-specific enzymes. The approximate position of the mutation within the amplicon is revealed by the size of the fragments on polyacrylamide gels (Salgotra and Neal Stewart 2020; Gilchrist et al. 2006). The chemical agent, ethylmethanesulfonate (EMS), induces a high volume of mutations: Treatment with EMS results in ethylation of G residues, leading to G/C → A/T transitions. Notably, the use of EMS or MNU mainly causes point mutations (Harloff et al. 2012) and small INDELs (Till et al. 2004). While TILLING is more cost-efficient than other SNPs detection methods and can be carried out without expensive machinery or complicated procedures, reproducibility remains challenging. Data generated from TILLING is often not comparable, as different methods have been used to calculate mutation frequencies. Harloff et al. (2012) suggested using the abundance of G nucleotides to calculate mutation frequencies.

1.6.2. Gamma-rays and Heavy-ion beams

After the discovery that X-rays induce mutations by Muller (1927), ionising radiation such as gamma-rays became established in plant genetics. Since the publication of the first results of X-rays in maize (Stadler 1928), the field has been developing rapidly, making gamma-radiation the most used method in plant mutation breeding since the 1960s. Almost 50% of officially registered mutant cultivars on the FAO/IAEA mutant variety database (http://mvgs.iaea.org) account for cultivars obtained by gamma-ray radiation (Ahloowalia and Maluszynski 2001; Li et al. 2019).

In the past two decades a different approach, radiation by heavy-ion beams, has been established as another efficient mutagenesis methodology. Both gamma-rays and heavy-ion beams are ionising radiation, which cause lesions in the chemical bonds of molecules. Those lesions are caused by the release or capture of electrons. The resulting lesions include nucleotide base lesions and DNA single- and double-bond breaks. Especially the latter lesion is of interest, as the pathways involved in the repair of DNA double-bond breaks are error-prone, resulting in substitutions, INDELs and chromosome rearrangements (Rodgers and McVey 2016; Li et al. 2019).

Depending on the ion being used for irradiation, heavy ion beams have unique properties, such as the mass and electrical charge of the ion. Therefore, the Linear Energy Transfer (LET) can vary from 22.5 to 4000 keV·µm^{-1}, whereas the LET for gamma-rays is 0.2 keV·µm^{-1}. Studies (Yatagai 2004; Hagiwara et al. 2019) have compared the effects of low and high LET radiation in mammalian cells: Low LET irradiation cause randomly distributed DNA damage in the nucleus, whereas high LET irradiation leads to chromosome breaks, translocations and large INDELs. The mutagenic effectiveness, an index to compare mutagenic agents, of heavy ion beams is, due to the high-LET properties, higher than of gamma-radiation (Kazama et al. 2017; Li et al. 2019; Yamaguchi et al. 2009).

A study by Li et al. (2019) compared the effects of C-ion beams and gamma-rays irradiation in rice. Their findings showed a higher amount of SNPs and small INDELs (-4bp - +1bp) in gamma-radiated cultivars. These results agree with previous studies: Gamma irradiation tends to cause smaller INDELs (1-2bp) than C-ion radiation (>3bp) in *Arabidopsis* (Yoshihara et al. 2010).

Ultimately, the use of a specific mutagen depends on the experimental design and aims. Gamma-radiation was found to be the most suitable mutagen for our experiment: Although TILLING does cause mutations similar to those caused by gamma-radiation, the majority of registered mutant cultivars were generated using gamma-radiation. Therefore our data is comparable to a bigger number of cultivars. Intermediate and large INDELs and SVs usually cannot be detected using NGS data (Shigemizu et al. 2018), hence why heavy ion radiation would be unsuitable for the present project.

However, heavy-ion beams can be utilised for future experiments: As INDELs are difficult to computationally analyse, an approach to improve the INDEL-calling pipeline could be using data from heavy-ion irradiated cultivars, as these are more likely to show large INDELs and Structural Variants (SV; Li et al. 2019). Although large INDELs and SVs can also be found in gamma-ray irradiated rice cultivars (Morita et al. 2009), the number remains comparatively low (1 SV, 16.7% large deletions (9.4 - 129.7kbp). With more INDEL-data available, the focus of such experiments should be the accurate detection and validation of INDELs to create an efficient variant-calling pipeline.

1.7. The field of Bioinformatics

While NGS technologies are facilitating sequencing projects, a new challenge arises in analysing and interpreting the resulting information (Metzker 2010). Further, genetic maps, genotypes and genomic expression need to be processed to acquire the relevant biological information. The field of bioinformatics deals with such challenges: The first sequence analysis dates back to the beginning of Sanger sequencing and numerous genomes have been sequenced since. Now, more reads of lower quality are being generated by NGS technologies. To efficiently analyse such data, new approaches are needed to understand complex biological traits (Horner et al. 2010; Pérez-de-Castro et al. 2012). This is what, among other things, the field of bioinformatics deals with.

1.7.1. Challenges in Variant Detection

Despite all progress, SNP detection in plants remains challenging. Firstly, the number of possible genotypes is increased due to a higher number of alleles and resulting combinations. Secondly, the absence of physical linkage with heterozygous alleles complicates determination of the exact number of present alleles. The resulting low coverage per allele impedes detection of sequencing errors: NGS data show higher error rates than traditional Sanger sequencing, making sequence error detection a crucial step in the pipeline. This can be overcome by deeper sequencing and therefore increasing the confidence in a called SNP.

Further, NGS sequencing results in shorter reads than Sanger sequencing and therefore an increased risk of aligning sequences to the wrong region. Previous studies have suggested focusing on more accurate read mapping (Garrison et al. 2018) and usage of long-read sequences (Wenger et al. 2019) to increase the accuracy in variant calling. Li et al. (2011) developed a model that suggested sequencing more individuals at low depth (2-4x) was a viable alternative to sequencing fewer individuals at high depth (>30x) in association studies. However, this model cannot be applied to plants, as it assumes a diploid genome and many plant species are of polyploid nature (Wenger et al. 2019; Cooke et al.). Thus, finding a trade-off between deeper sequencing and accounting for wrong alignments remains a challenging task.

Similar challenges arise during INDEL detection. Although software specifically for INDEL detection in NGS has been developed (eg. SOAP; Li et al. 2008), NGS platform specific errors impede correct detection of INDELs. Most tools are able to account for substitution errors only and those who do account for insertion or deletion errors were found to be fairly inaccurate, showing in disagreement of detected INDELs. A proposed reason for inaccurate INDEL error handling is the fact that error programs utilise high data coverage to account for errors. Although various subcategories of error handling programs exist, which utilise different parameters and methods, many programs struggle with similar challenges: reads mapped to the wrong region in the genome, reads associated in low-covered regions and reads with high error rates. Further, NGS platforms that produce long reads, and therefore high quality assemblies, are especially prone to such errors (Allam et al. 2015).

An example of limited concordance of detected INDELs has been pointed out by (Ramakrishna et al. 2018). Their study researched INDELs in cultivated soybeans by whole genome resequencing. Despite showing Transition/Transversion (TS/TV) ratios comparable to previous studies, the average densities for SNPs and INDELs were found to be 79.04/Mb and 461.48 /Mb, respectively. Earlier experiments (Yadav et al. 2015) showed lower SNPs/Mb and higher INDELs/Mb densities than those found by (Ramakrishna et al. 2018).

Moreover, although INDELs were found to be the second most common type of mutations in the human genome (1 INDEL per 7.2Kb; (Mills 2006), it is estimated that a third of small INDELs in humans are undetected as per 2010 (Mullaney et al. 2010). Supporting this statement, a study (Jiang et al. 2015) analysing European and Yoruban High-Throughput Sequencing (HTS) data estimated that only 55% of INDELs in both genomes have been detected. The study assessed accuracy of INDEL detection and suggested that a third of all INDELs occur in long homopolymers, regions which impede INDEL detection. Further challenges in

INDEL detection have been listed, namely presence of repeats, short interspersed elements and homopolymers and dimers. Regions with a high number of repeats are known to be prone to sequencing errors, however the results of the study suggest that errors in INDEL detection in such regions are due to a fundamental nature, which requires new technologies and techniques (Jiang et al. 2015).

These findings highlight the need for a more accurate and precise detection of polymorphisms. Despite being common in all genomes and owning useful properties for further genetic analysis, the detection of genetic markers remains challenging in plants. (Lateef 2015) argues to detect as many SNPs as possible as “It is very likely that improvement of complex traits will depend on the ability to manipulate genes, which have minor effects, and show interaction with each other” (Lateef 2015).

1.8. Thesis Aims

Variant detection in polyploid organisms remains a challenging task. Variants often remain undetected due to the wealth of genomic data. Further, many programmes and algorithms are programmed for diploid genomes and therefore cannot be applied to polyploid organisms. Therefore, the aim of this project is to create an effective variant-detecting pipeline in *B.napus*. As such, we developed a pipeline to analyse 584 mutant lines arising from gamma radiation.

The aims of this project can be categorised into two parts:

1. **Development of a computational pipeline which accurately detects variants.**
   Here we evaluated different methods to detect variants: Variant detection by clustering of FASTA files and by calling and filtering from VCF files. While the latter method is common practise in Bioinformatics, the aforementioned challenges impede accurate detection. By evaluating different software and commands we aim to establish a computational pipeline that filters out sequencing noise but remains sensitive enough to detect true variants.
2. **Validation of computationally detected variants by sequencing.**
   To confirm the presence of variants, we will analyse sequences from mutant lines which have been detected by our pipeline. Here we utilise alignments of DNA sequences to validate computationally detected variants. Finally, we aim to confirm variants with sequencing data from selected mutational lines.

2. Materials and Methods

2.1. Plant Material

For this project the DNA sequences of the rapeseed variety “Maplus”, a european winter-habit type (low glucosinolate, high erucic acid type), was analysed. Seeds were exposed to varying levels of gamma rays (^{60}Co), 750 Gy, 1500 Gy, 1750 Gy and 2000 Gy, at the International Atomic Energy Agency (IAEA) in Austria. For comparison, seeds of the same variety were treated with four different doses of fast neurons (FN), 40 Gy, 60 Gy, 80 Gy and 100 Gy in the Budapest Research Reactor (BRR) at HAS Centre for Energy Research (AEKI), Hungary. To compare the effects of gamma irradiation and the corresponding FN dosages, 8 M_1 selfed plants were grown (M_2). Seeds from M_2 were harvested and prepared for DNA and RNA extraction, according to (He *et al.*, 2017). Following mRNA-Seq, performed by the Illumina sequencing platform, 150 base PE reads have been obtained from the HiSeq 4000 platform at Beijing Genomics Institute (BGI), China.

2.2. Computational Analysis

2.2.1. File Types

The following chapter introduces different file types used in this project. It will elaborate on the reasoning for the creation and its function of each file format. Further, the basic structure of each file format will be presented and its advantages and disadvantages will be reviewed. Lastly, it will discuss why these particular file formats have been chosen over different ones.

2.2.1.1. Sequence Alignment/Map (SAM) format and Binary Alignment/Map (BAM) format

The progress in sequencing and aligning technologies led to the development of several programmes, which impeded downstream analysis due to changes in file formats. With the start of the 1000 Genomes Project (1000 Genomes Project Consortium et al,.2010), Li *et al.*, (2009) developed the Sequence Alignment/Map (SAM) format. SAM supports different types of sequences, single- and paired-end reads and uniforms information about the quality of the sequence within one format. The Binary Alignment/MAP (BAM) format is a binary and compressed version of SAM which stores the same information.

Other file formats have emerged with the goal of standardising file formats. Examples are the Sequencing Reads Format (SRF, http://srf.sourceforge.net/), genotype likelihoods/posterior SNP probabilities (GLF; http://maq.sourceforge.net/glfProgs.shtml) and the Genome Variation Format (GVF; Reese *et al.*, 2010). Both SRF and GLF store variant information in non-standardised tabular formats, hence why they are unsuitable for comparison between genome analysis projects. GVF is an adaptation of the Generic Feature Format version 3 (GFF3), which uses Sequence Ontology to detect variants (Reese *et al.*, 2010). Although the format is not suitable for storing variants from different samples (Lorenc, 2015), GVF can be used for exchange of variant annotations between different genomic databases (Cunningham *et al.*, 2015). Lubin *et al.*, (2017) suggested that another file type, the Variant Call Format (VCF) has more options for file manipulation and a better application to clinical projects.

2.2.1.2. Variant Call Format (VCF)

After the standardisation of next-generation reads alignment by the SAM/BAM file format, the Variant Call Format (VCF) was proposed by Danecek *et al.*, (2011). The development of VCF aimed to create a standardised format for storing variant information, with rich annotations about the respective variant. As the GVF is not applicable for storing information about multiple samples, the VCF was designed for usage in polyploid organisms and various contexts. Further, indexing allows fast access to variations and easy data manipulation (Danecek *et al.*, 2011)

2.2.1.2.1. Structure of the VCF

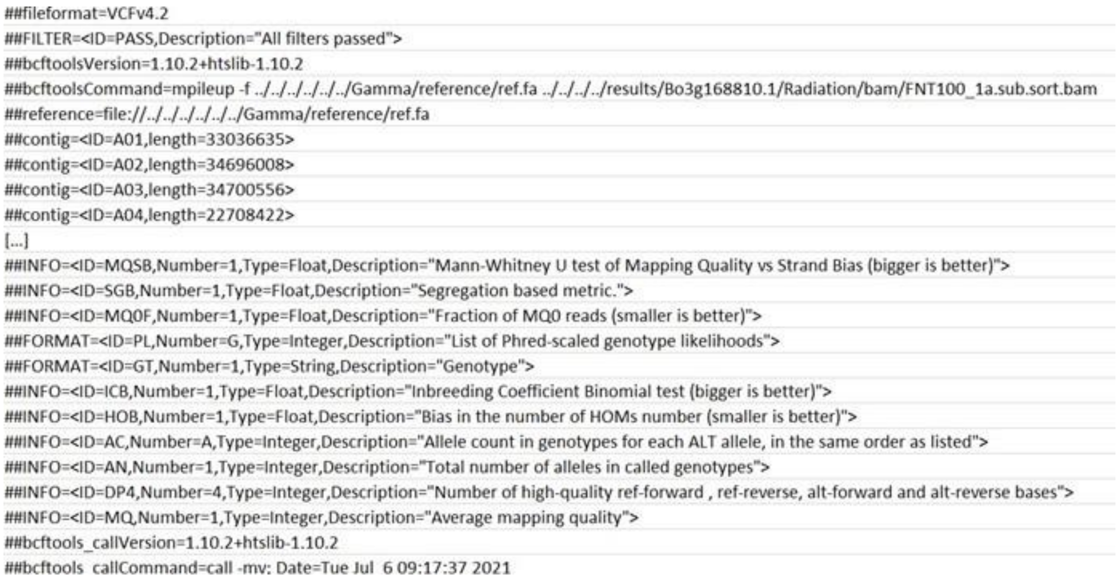
The VCF is a tab delimited text format, which stores all the information about a given sequence in 10 columns. The file can be divided into 3 parts (EMBL-EBI, no date; Ian_Maurer, 2020):

- Meta Information: Multiple lines prefixed by ##
- Column Header: Single line prefixed by #
- Data Lines: Information about variants

The information included ranges from the position of the called variant to detailed statistical information, which puts the identified site into context of the entire sequence and genome.

2.2.1.2.1.1. Meta Information

The meta information is the first part of the VCF. Identifiable by the ##-prefix, the meta information stores details about the content of the VCF. In the upper part of the section, specifications about the VCF and the commands used to generate the VCF file can be found. The middle section stores information about the sequences of the specific genome. The length of chromosomes and genes can be found there. The last part of the section describes abbreviations used in the INFO column of the Column Header section (see Fig. 3 and Table 2 for details).



```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.10.2+htslib-1.10.2
##bcftoolsCommand=mpileup -f ../../../../../../Gamma/reference/ref.fa ../../../../results/Bo3g168810.1/Radiation/bam/FNT100_1a.sub.sort.bam
##reference=file://../../../../../../Gamma/reference/ref.fa
##contig=<ID=A01,length=33036635>
##contig=<ID=A02,length=34696008>
##contig=<ID=A03,length=34700556>
##contig=<ID=A04,length=22708422>
[...]
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.10.2+htslib-1.10.2
##bcftools_callCommand=call -mv; Date=Tue Jul  6 09:17:37 2021
```

Figure 3: Example of the meta data in a VCF file.

The meta data section contains detailed information about file type specifications and used commands.

Table 2: Explanation of the Abbreviations used in VCF files.

The following abbreviations are used in the INFO column of the data lines section. The descriptions can be found in the meta data section and are presented down below (Team, 2015).

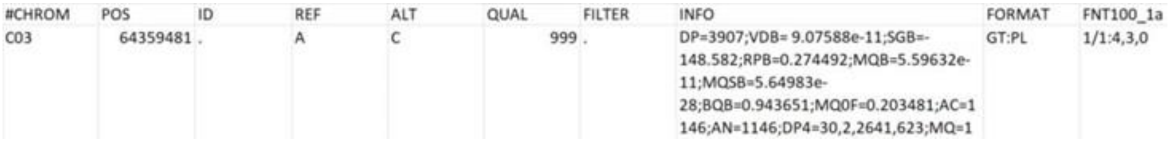
Abbreviation	**VCF Description**
IDV	Maximum number of raw reads supporting an indel
IMF	Maximum fraction of raw reads supporting an indel
DP	Raw read depth

VDB	Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)
RPB	Mann-Whitney U test of Read Position Bias (bigger is better)
MQB	Mann-Whitney U test of Mapping Quality Bias (bigger is better)
BQB	Mann-Whitney U test of Base Quality Bias (bigger is better)
MQSB	Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)
SGB	Segregation based metric.
MQ0F	Fraction of MQ0 reads (smaller is better)
PL	List of Phred-scaled genotype likelihoods
GT	Genotype
ICB	Inbreeding Coefficient Binomial test (bigger is better)
HOB	Bias in the number of HOMs number (smaller is better)
AC	Allele count in genotypes for each ALT allele, in the same order as listed
AN	Total number of alleles in called genotypes
DP4	Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases

MQ	Average mapping quality

2.2.1.2.1.2. Column Header and Data Lines

The remaining part of a VCF file consists of the column header and data lines. The column header contains a single, #-prefixed line. Each line in the data lines section represents a position in the genome and corresponds to the column header section. Missing values in data lines are normally represented by a dot (see Fig. 4 and Table 3).



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	FNT100_1a
C03	64359481	.	A	C	999	.	DP=3907;VDB= 9.07588e-11;SGB=-148.582;RPB=0.274492;MQB=5.59632e-11;MQSB=5.64983e-28;BQB=0.943651;MQ0F=0.203481;AC=1146;AN=1146;DP4=30,2,2641,623;MQ=1	GT:PL	1/1:4,3,0

Figure 4: Example of a SNP in a VCF file.

The variant shown was taken from Bo3g168810.1 (FAD2) of the *B. napus* genome. Given the data presented here, the following information can be read about the variant:

The detected variant is a A → C transition found at 64359481 on Chromosome 3 of the C genome. The expected genotype is homozygous (alternative) and unphased. The present SNP contains additional information about the phred-scaled genotype likelihood (PL).

Table 3: Explanation of the VCF Column Header.

The following abbreviations are found in the column header of a VCF file.

Abbreviation	**Definition**
CHROM	Chromosome
POS	Position of the variant
ID	Identifier
REF	Reference Allele - Base(s) found on the reference sequence
ALT	Alternative Allele - Base(s) found on the present sequence
QUAL	Quality Score (out of 100)
FILTER	Quality Filter - Indication which filter have passed or failed (semicolon)
INFO	Further Information - (see Table 1)
FORMAT	Genotype (GT) - Some VCF versions contain the additional information and its corresponding values in this and the following column, respectively.
Sample Data (optional)	GT values - Indicates which alleles are phrased (|) or unphased (/). 0 represents the reference, 1 the alternative allele.

Depending on the type of NGS mechanisms used, the resulting sequences may be prone to different errors. For instance, the Illumina platform shows fairly high misscall error rates at 1%. The SOLiD platform utilises fluorescent dye colour, which bias appears in later machine cycles. Further uncertainty can be introduced by high depth sequencing and alignment errors. Platform specific algorithms have been developed to improve error rates by 5-20% and reduce false-positive SNP calls. Examples for such are BayesCall for Illumina (Kao et al. 2009) and Rsolid for the SoLiD platform (Wu et al. 2010; Nielsen et al. 2011).

However, SNP calling requires quantification that accounts for any errors that may affect the accuracy and possibility of any called SNP. While some platforms designed algorithms specific to their platforms, the Phred quality score remains the standard measure of quality of sequences generated by DNA sequencing. The following formula allows conversion from platform-specific to the Phred quality score:

$$Q_{Phred} = 10 \log_{10} P$$

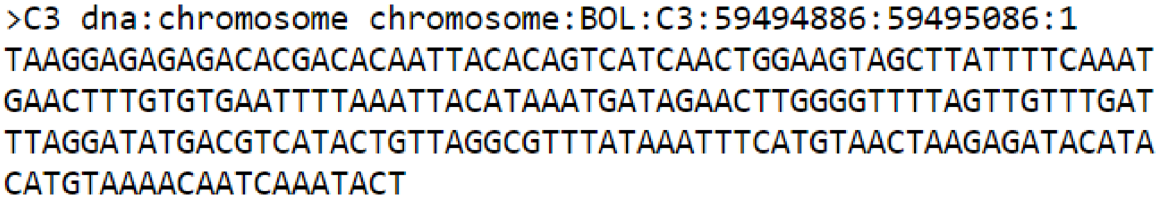
Where P represents base-calling error probability (Ewing et al. 1998). A 1% error rate in base calling conforms to a Phred score of 20.

The VCF was adapted by the 1000 Genomes Project as the standard file format for variant scoring (Lorenc, 2015). However, Moore *et al.*, (2012) argues that the VCF lacks the option to store further information about a called variant, such as its biological consequences. While the GVF possesses the infrastructure to store such annotations, notably, this may be of relevance in the field of medicine and healthcare. The annotations in the VCF are sufficient for this present project, as it does not aim to investigate the biological impact from mutations. This project focuses on improving a variant calling pipeline and identifying new variants. While conversion to GVF is possible, this additional step lies beyond the scope of the project. Future experiments may compare VCF and GVF pipeline to identify further advantages and disadvantages of each format in *B. napus*.

2.2.1.3. FASTA Format

The FASTA format was initially developed as an improvement to the FASTP format by Pearson and Lipman (1988). Changes in the algorithm lead to increased sensitivity and a more sophisticated alignment approach. However, the most striking feature of the FASTA format is the ability to quickly search and compare

DNA sequences in data bases. While amino acid sequences were still stored in FASTP files, FASTA files are nowadays universally used to store DNA and amino acid sequences. The simplicity of the format (see Fig. 5) allows easy processing and manipulation with common programming and scripting languages. Later, the FASTQ format emerged as an extension to FASTA: Additionally to DNA sequences, the FASTQ file includes numeric quality scores corresponding to each nucleotide (Cock et al., 2010).



```
>C3 dna:chromosome chromosome:BOL:C3:59494886:59495086:1
TAAGGAGAGAGACACGACACAATTACACAGTCATCAACTGGAAGTAGCTTATTTTCAAAT
GAACTTTGTGTGAATTTTAAATTACATAAATGATAGAACTTGGGGTTTTAGTTGTTTGAT
TTAGGATATGACGTCATACTGTTAGGCGTTTATAAATTTCATGTAACTAAGAGATACATA
CATGTAAAACAATCAAATACT
```

Figure 5: Example a FASTA file.

The presented example is a 200bp excerpt from the C3 chromosome of the *B. oleracea* (BOL) genome. A sequence in the FASTA file starts with a single-line identifier, followed by sequence data. The identifier is distinguished from the actual sequence by a greater-than (>) symbol at the beginning. Sequences are represented in the standard IUPAC code for amino acids and nucleotide bases. The identifier at the start describes the starting position of the sequence, as well as, chromosome, gene ID and (optionally) species (BLAST TOPICS, no date).

2.2.2. Development of a Variant Detecting Pipeline

The resulting Illumina sequencing reads were prepared for computational analysis. An overview of the pipeline can be found at Fig. 9. The reads in FASTQ were mapped to a reference genome using Borrows Wheeler Aligner (BWA-mem, version 0.7.17; Li and Durbin 2009). The reference sequence used was the *Brassica* pan-transcriptome developed by He et al. (2015). The resulting bam files have been aligned to the same reference sequence, sorted and shortened for variant calling. The shortened versions included regions of the genes of interest. For bam files manipulation SAMTools (version 1.10-foss-2018b; Li et al. 2009) and BCFTools (version 1.10.2-GCC-9.3.0; Li 2011), for subsequent variant calling VCFTools (version 0.1.15-foss-2018b-Perl-5.26.1; Danecek et al. 2011) have been used.

In the following steps VCF files have been processed to validate the reported variants. Using BEDTools (version 2.30.0-GCC-11.2.0; Quinlan and Hall 2010), variants which occurred in both control and “mutant” lines were removed. The resulting VCF files were separated by variant types, namely SNPs and INDELs. Next, SNPs which do not affect coding regions were removed. Filters were applied: For both SNPs and INDELs a QUAL < 30 filter was applied. For SNPs, additionally, a missing data < 50% was applied. The filters removed all variants with a quality score below 30 and SNPs which have more than 50% of their data missing. Lastly, utilising R, variants which occur more than 6 times across all samples were removed. All computational work was undertaken on the Viking Cluster, which is a high-performance compute facility provided by the University of York.

To prepare for Wet Lab validation, variants with mutant alleles which occurred at least 60% of the time were chosen. This was done comparing the AD (number of mutant alleles) and DP (total number of any alleles). Further, the mutant sequences were aligned to a reference sequence to visualise the mutation. As such, the following mutations and their lines have been selected to be validated (see Table 4):

Table 4: Variants selected for sequencing.

A05	25129372	733	G	A	FNT80-6
A05	25129085	446	G	A	FNT80_1a, G2000-136a, G2000-419-1

Despite G2000-419-1 not showing any variation in the alignment, we wanted to test the precision of variant detection utilising FASTA alignments. As the same mutation was detected at a high rate (60% < occurrences) in 2 different lines, we hypothesised that the same mutation will be visible in said mutational line.

2.2.3. Wet lab Validation

After detecting and analysing variants computationally, selected samples were prepared to be validated by sequencing. In the following protocol, DNA samples containing variants were amplified and prepared for sequencing.

2.2.3.1. Polymerase Chain Reaction (PCR)

The reactions were carried out in 0.5ml tubes containing:
7µl 1X Master Mix (Thermo Scientific), 1µl of forward and reverse primers each, 1µl cDNA and 5µl water each. The sequences of primers used can be found in Table 5.

The PCR was set at the following protocol:

SNPs:

- 94°C for 5 minutes
- 29 cycles of:

    - 94°C for 30 seconds
    - 57°C for 30 seconds
    - 75°C for 1 minute
- 72°C for 10 minutes
- Hold at 7°C

INDELs:

- 95°C for 5 minutes
- 14 cycles of:
    - 94°C for 30 seconds
    - 63°C for 30 seconds (-1°C per cycle)
    - 72°C for 1 minute
- 29 cycles of:
    - 94°C for 30 seconds
    - 53°C for 30 seconds
    - 72°C for 1 minute
- 72°C for 15 minutes
- Hold at 7°C

Table 5: Primers used.

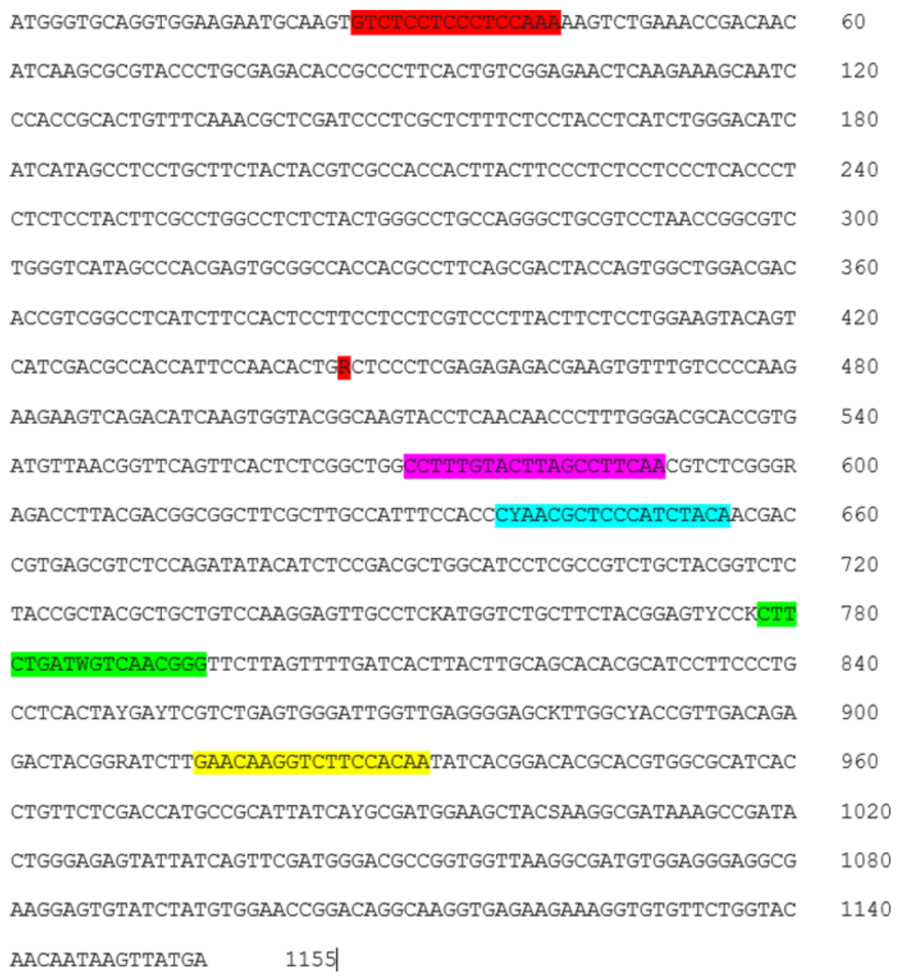
FAD2.A5-F	GTGTCTCCTCCCTCCAAA
FAD2.A5-R	CCTCATAACTTATTGTTGTACCAG
FAE1.C3-F	GCCGCTATTTTGCTCTCCAA
FAE1.C3-R	CCAATCAATTCGGGAGCCAC

2.2.3.2. Primer Design

Various primers were designed to test their amplification properties in the FAD2 and FAE1 gene. For the FAD2 gene, one forward (“forward”) and four reverse primers (R1-R4) were tested. Our aim was to test if the newly designed primers would be able to detect the SNP at location 446 (see Table 6). The previously used FAD2.A5 forward and reverse primers were utilised as control samples. For the FAE1 gene one forward (“Kati-F) and reverse primer (“Kati-R”) was designed. Here we aimed to amplify the INDEL at location 153 (data not shown).

Table 6: Primers tested.

Forward	GTCTCCTCCCTCCAAA
R1	TTGTGGAAGACCTTGTTC
R2	CCCGTTGACTATCAGAAG
R3	TGTAGATGGGAGCGTTAG
R4	TTGAAGGCTAAGTACAAAGG
Kati-F	ACACGAGTCTCTGACTTAC
Kati-R	GATTGATGTGCTAGAGAAGA



```
ATGGGTGCAGGTGGAAGAATGCAAGTGTCTCCTCCCTCCAAAAAGTCTGAAACCGACAAC    60
ATCAAGCGCGTACCCTGCGAGACACCGCCCTTCACTGTCGGAGAACTCAAGAAAGCAATC    120
CCACCGCACTGTTTCAAACGCTCGATCCCTCGCTCTTTCTCCTACCTCATCTGGGACATC    180
ATCATAGCCTCCTGCTTCTACTACGTCGCCACCACTTACTTCCCTCTCCTCCCTCACCCT    240
CTCTCCTACTTCGCCTGGCCTCTCTACTGGGCCTGCCAGGGCTGCGTCCTAACCGGCGTC    300
TGGGTCATAGCCCACGAGTGCGGCCACCACGCCTTCAGCGACTACCAGTGGCTGGACGAC    360
ACCGTCGGCCTCATCTTCCACTCCTTCCTCCTCGTCCCTTACTTCTCCTGGAAGTACAGT    420
CATCGACGCCACCATTCCAACACTGRCTCCCTCGAGAGAGACGAAGTGTTTGTCCCCAAG    480
AAGAAGTCAGACATCAAGTGGTACGGCAAGTACCTCAACAACCCTTTGGGACGCACCGTG    540
ATGTTAACGGTTCAGTTCACTCTCGGCTGGCCTTTGTACTTAGCCTTCAACGTCTCGGGR    600
AGACCTTACGACGGCGGCTTCGCTTGCCATTTCCACCCYAACGCTCCCATCTACAACGAC    660
CGTGAGCGTCTCCAGATATACATCTCCGACGCTGGCATCCTCGCCGTCTGCTACGGTCTC    720
TACCGCTACGCTGCTGTCCAAGGAGTTGCCTCKATGGTCTGCTTCTACGGAGTYCCKCTT    780
CTGATWGTCAACGGGTTCTTAGTTTTGATCACTTACTTGCAGCACACGCATCCTTCCCTG    840
CCTCACTAYGAYTCGTCTGAGTGGGATTGGTTGAGGGGAGCKTTGGCYACCGTTGACAGA    900
GACTACGGRATCTTGAACAAGGTCTTCCACAATATCACGGACACGCACGTGGCGCATCAC    960
CTGTTCTCGACCATGCCGCATTATCAYGCGATGGAAGCTACSAAGGCGATAAAGCCGATA    1020
CTGGGAGAGTATTATCAGTTCGATGGGACGCCGGTGGTTAAGGCGATGTGGAGGGAGGCG    1080
AAGGAGTGTATCTATGTGGAACCGGACAGGCAAGGTGAGAAGAAAGGTGTGTTCTGGTAC    1140
AACAATAAGTTATGA        1155
```

Figure 6: Visualisation of primer design.

Primers were designed to include the mutation (highlighted in red). The forward primer is highlighted in red, the reverse primers R4, R3, R2 and R1 in magenta, cyan, green and yellow, respectively.

2.2.3.3. Gel Electrophoresis

The length of PCR products was confirmed using Gel Electrophoresis in a 1% Agarose gel (see Fig. 7). The gel was prepared with a 0.5X TBE buffer and 2µl of ethidium bromide. Aliquots of 5µl of the PCR products were loaded onto the gel and run for 30 minutes at 130V.

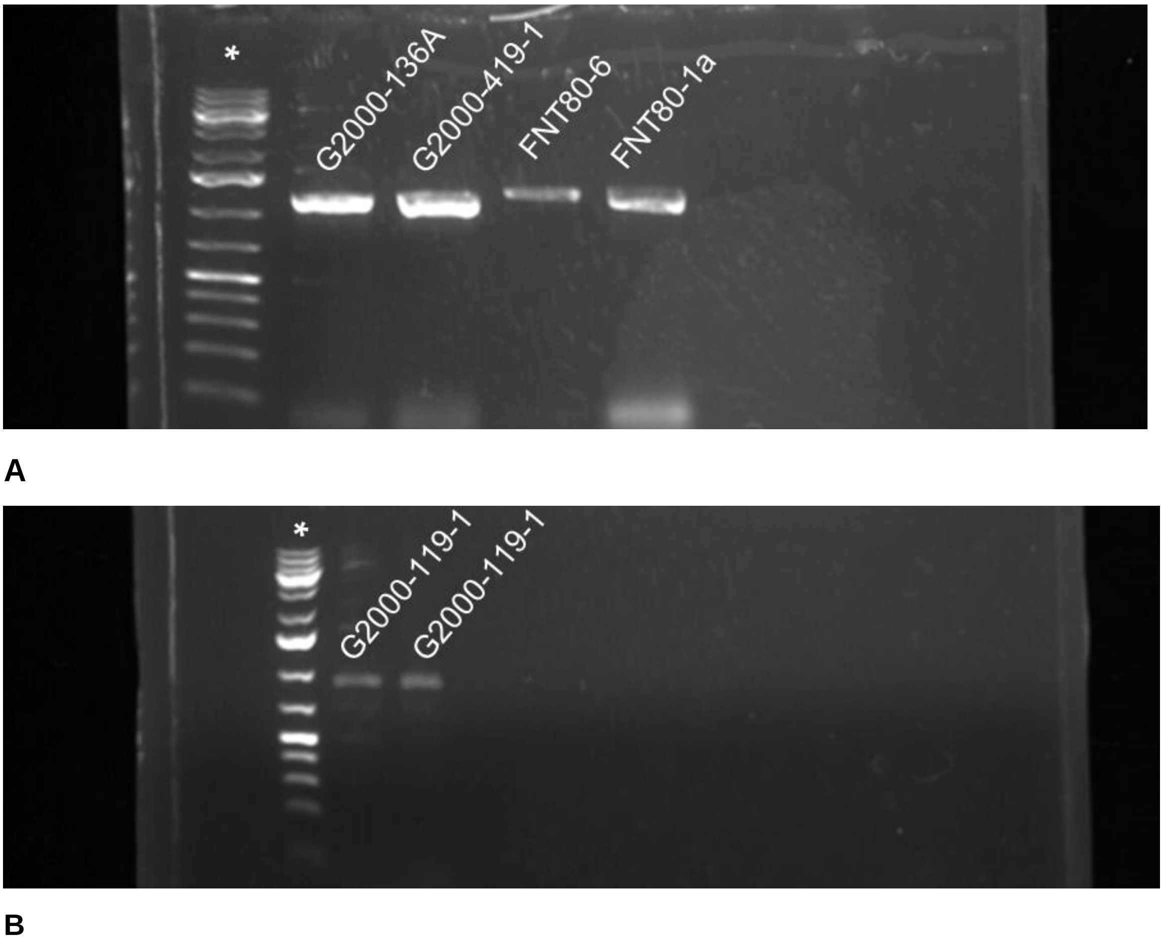


Figure 7: Gel Electrophoresis confirms successful amplification of genes.

*****: 10bp DNA ladder

A: The FAD2 gene of 4 mutational lines were amplified. (G2000-136A, G2000-419-1, FNT80-6, FNT80-1a)

B: Amplification of FAE1 genes of 2 mutational lines. (G2000-119-1, FNT40-5)
All PCR products match with the expected length of their respective genes.

2.2.3.4. Preparation of samples for sequencing

For each sample, 10µl of PCR product was mixed with 1µl SAP and 1µl Exonuclease I. After mixing, the samples were amplified using the following protocol:

- 37°C for 15 minutes
- 80°C for 15 minutes
- Hold at 7°C

After, 2µl of 10µM primers were added to the samples. For FAD2 samples the forward, for FAE1 samples the reverse primer were used (see Table 5). Lastly, water was added to reach a total volume of 15µl and sent to sequencing.

3. Development of Variant Detecting Methodologies

The aim of the present project is to develop variant calling pipelines for *B. napus*. As such we investigated two different approaches, variant detection by clustering and by computational analysis.

3.1. Variant detection by clustering

Firstly, we researched the possibilities to call variants with FASTA files. Utilising whole genome sequencing data from gamma-radiation panels, we assembled identical genomic sequences into clusters and compared them to each other using *logDet* pairwise distances (phangorn R package; Schliep 2021). This allowed us to plot said distances onto a phylogenetic tree (ggtree R package; Yu *et al.*, 2017, ggplot2 R package; Wickham, 2016) and to identify similar and dissimilar clusters. Subsequently we analysed erucic acid content in relation to cluster similarity: We identified sequences with high and low erucic acid content (either top or low 10% of all samples) and compared them against similarity to other sequences.

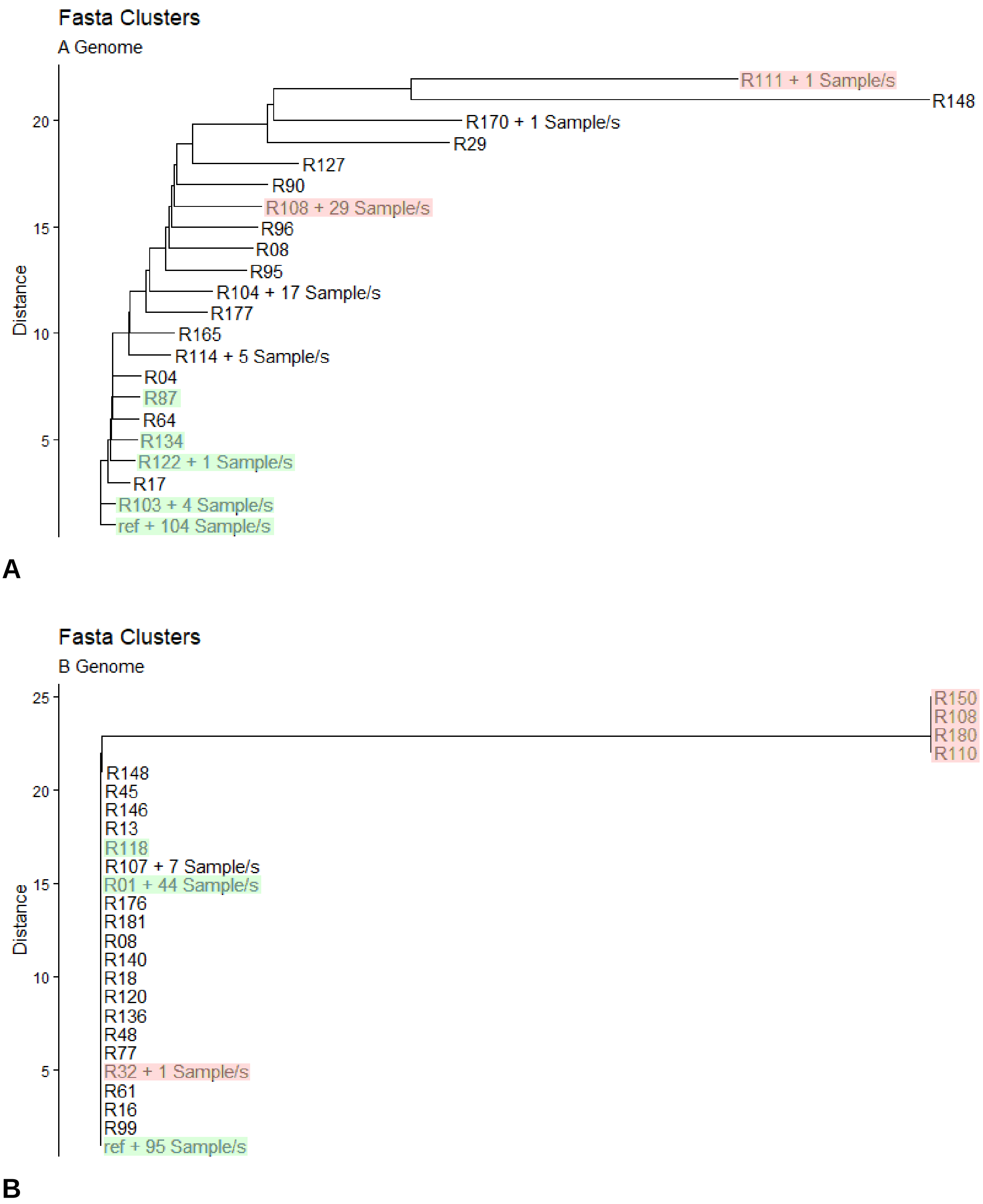


Figure 8: Comparison of sequence similarity in relation to erucic acid content. A+B: Phylogenetic trees were plotted to compare similarity between FASTA clusters. A total of 184 sequences were grouped into clusters with identical sequences. Single Sequences were numbered from (R) 1-184. Clusters were then plotted in relation to genomic distance from the reference cluster. Subsequently, 18 sequences (~10%) with the highest and lowest erucic acid content were identified and highlighted in green and red, respectively. Clusters containing sequences with

both high and low erucic acid content were highlighted only if sufficient sequences for one trait (twice as many) were present. Remaining clusters were highlighted if at least one top/bottom 10% sequence was found within the cluster. Genetic distances (as *dis.logDet* values) were scaled as branch lengths. Nodes represent branching points from sequence similarities (Baum, 2008; Yu et al., 2017).

While we were not able to identify any variants here, the FASTA clustering method provided us with an overview of the variation of sequences (see Fig. 8). Using clustering, we were able to see how many identical sequences were present: In both A and B genome, the biggest cluster group included sequences identical to the reference sequence. While more and bigger clusters are present in the A genome, almost all sequences of the B genome fall under the same interior node. On the contrary, the phylogenetic tree for genome A shows more nodes and branches of varying length, indicating greater genomic variation. Given the number of clusters with identical sequences, some common variants may be present in the sequences of the A genome. Interestingly, the A genome implies that higher dissimilarity to the reference sequence may result in decreased erucic acid content: Samples with high erucic acid content were found in clusters most similar to the reference cluster. Such a trend is not visible in the B genome, where samples with high erucic acid content are present regardless of genetic distance. Notably, in both genomes, sequences with the highest dissimilarity to the reference sequence all showed low erucic acid content.

3.2. Variant detection by computational analysis

Secondly, we followed a more established approach by developing a variant calling pipeline utilising bioinformatic software. The data used stems from the same gamma-radiation panel as used as in the clustering method. A schematic overview of the full pipeline can be found in Fig 9. Initially we mapped raw reads to the reference sequence. The resulting BAM files were aligned to a reference sequence to organise and locate sequences. Then, BAM files were shortened to include the region of interest only to save processing power. After preparing BAM files for variant calling we followed several steps to obtain variants of high quality: Firstly, variants which appeared in both control and mutant files were excluded to ensure that called variants were actually caused by radiation. Samples from control files were not treated with any mutagens. Next, we separated variants according to their type, namely SNPs and INDELs. By doing so different filters can be applied according to variant type. As variants in coding regions are more likely to impact the phenotype, we excluded SNPs in non-coding regions. Subsequently filters were applied to exclude low-quality variation, variation which is likely to be sequencing noise (see Chapter 2.2.1.2.1.2.): A QUAL < 30 filter was applied to both SNPs and INDELs, an additional missing-data < 50% was applied to SNP variants. The application of filters is a standard step in variant calling pipelines to exclude sequencing errors and low-quality variants. We decided to use the missing-data < 50% filter for SNPs to further exclude poorly sequenced samples and to ensure high quality of the dataset (Cercaet al. 2021).

Finally, we removed variants which occurred more than 6 times across all samples: PCR duplicates can arise if the same DNA fragment is sequenced multiple times. Because PCR duplicates can occur through PCR amplification bias, the presence of many duplicates may lead to misidentification of these as true variants (Ebbert et al. 2016). As our total sample size amounts to almost 600, we set a threshold of 1% to avoid any misidentification. Here we ran a R function that counts occurrences (using information from VCF files) and outputs those with less than 6 occurrences in a separate file. Subsequently, the BCFTools *-query -l* command matched variants from said file to the previously filtered VCF files. Variation with no matches were hereby removed.

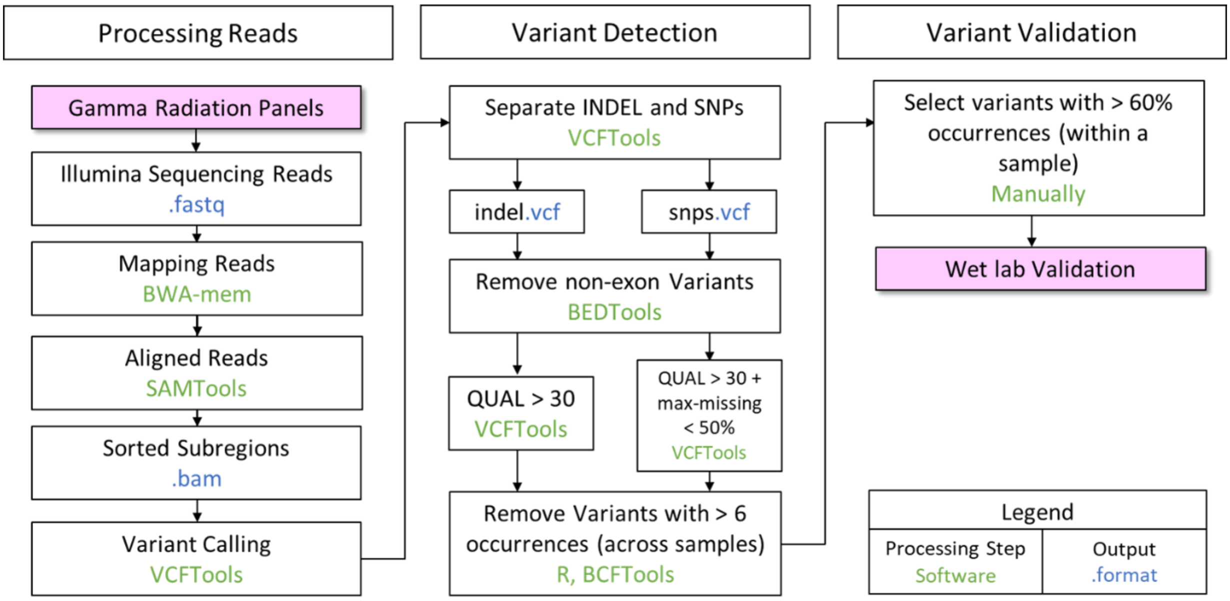
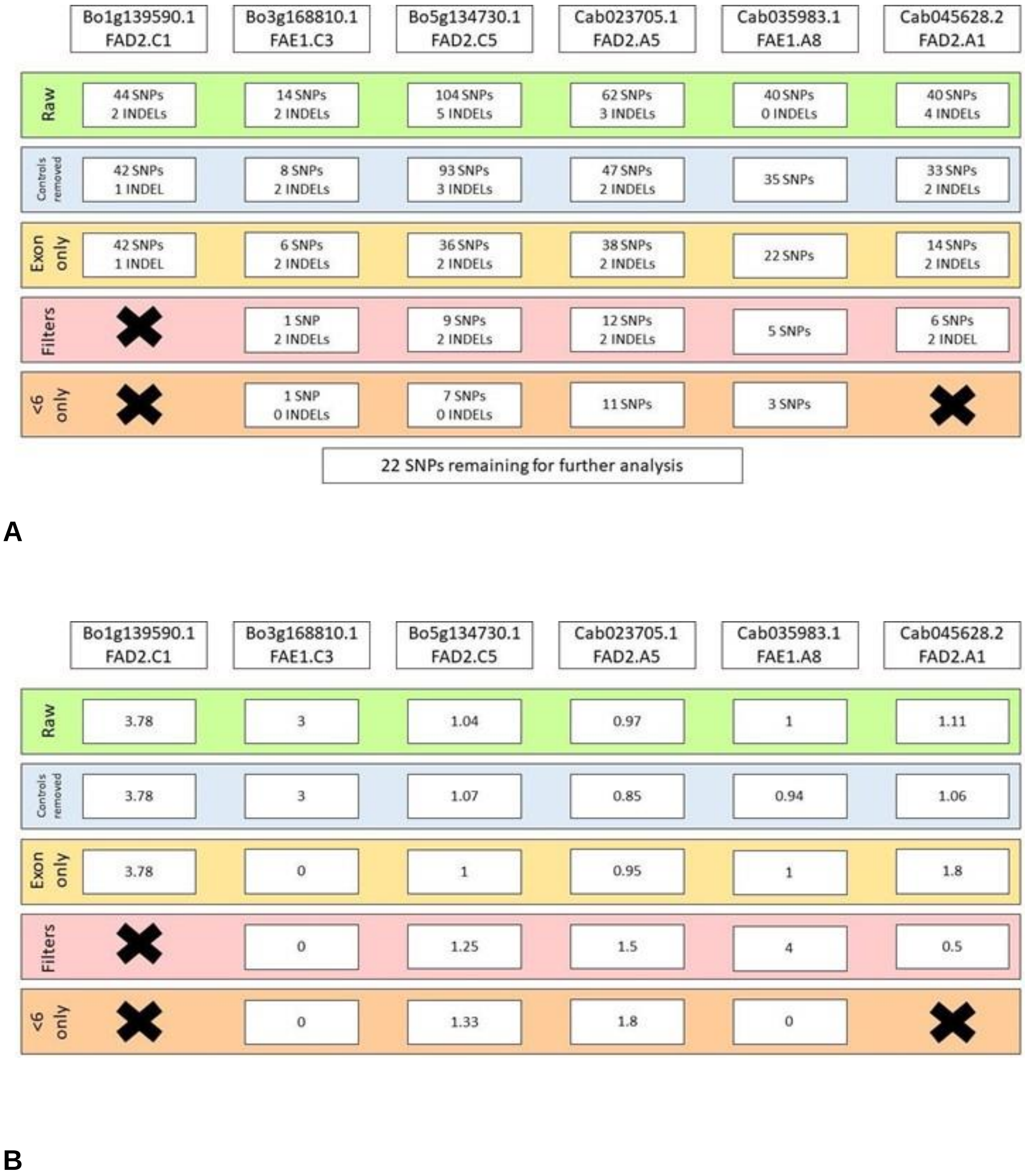


Figure 9: Schematic Overview of the Computational Pipeline.

The pipeline can be divided into initial data preparation and subsequent variant filtering and calling. In the first half of the pipeline datasets were converted from FASTQ to BAM files. Then, BAM files were prepared for subsequent variant calling. In the second half filters were applied to ensure high quality of called variation.



A

	Bo1g139590.1 FAD2.C1	Bo3g168810.1 FAE1.C3	Bo5g134730.1 FAD2.C5	Cab023705.1 FAD2.A5	Cab035983.1 FAE1.A8	Cab045628.2 FAD2.A1
Raw	44 SNPs 2 INDELs	14 SNPs 2 INDELs	104 SNPs 5 INDELs	62 SNPs 3 INDELs	40 SNPs 0 INDELs	40 SNPs 4 INDELs
Controls removed	42 SNPs 1 INDEL	8 SNPs 2 INDELs	93 SNPs 3 INDELs	47 SNPs 2 INDELs	35 SNPs	33 SNPs 2 INDELs
Exon only	42 SNPs 1 INDEL	6 SNPs 2 INDELs	36 SNPs 2 INDELs	38 SNPs 2 INDELs	22 SNPs	14 SNPs 2 INDELs
Filters	✖	1 SNP 2 INDELs	9 SNPs 2 INDELs	12 SNPs 2 INDELs	5 SNPs	6 SNPs 2 INDEL
<6 only	✖	1 SNP 0 INDELs	7 SNPs 0 INDELs	11 SNPs	3 SNPs	✖

22 SNPs remaining for further analysis

B

	Bo1g139590.1 FAD2.C1	Bo3g168810.1 FAE1.C3	Bo5g134730.1 FAD2.C5	Cab023705.1 FAD2.A5	Cab035983.1 FAE1.A8	Cab045628.2 FAD2.A1
Raw	3.78	3	1.04	0.97	1	1.11
Controls removed	3.78	3	1.07	0.85	0.94	1.06
Exon only	3.78	0	1	0.95	1	1.8
Filters	✖	0	1.25	1.5	4	0.5
<6 only	✖	0	1.33	1.8	0	✖

Figure 10: Overview of detected variation.

A: Number of SNPs and INDELs present after each filtering step, for each gene.

B: TS/TV ratios after each filtering step.

For each filtering step, the transition/transversion (TS/TV) ratio was calculated (see Fig. 10). The ratio was calculated taking the number of transitions divided by the number of transversions. If the TS/TV > 1, the number of transitions was greater than the number of transversions. This would give us an overview of the present variations and let us compare our pipeline to others.

A total number of 304 SNPs and 16 INDELs were detected. In the initial filtering step, variation aligning with those in control lines were removed. This means, all variation, which position and mutation (eg. A → G) matches to variation detected in control lines, were removed. In the initial step 44 SNPs and 6 INDELs were removed, leaving a total of 258 SNPs and 9 INDELs left for further filtering. The number of transitions and transversions remained fairly consistent across all genes, with FAD2.A1 showing the biggest difference in the TS/TV with -0.15. In the next step, (113) SNPs, which occurred outside of coding regions, were filtered out. Notably, this step filtered out all transitions in FAE1.C3 and reduced the number of transitions in FAD2.C5, while more transversions were filtered out for the remaining genes, except for FAD2.C1. Next, filters (QUAL < 30, missing-data < 50%) were applied. This step filtered out the most variation, with 123 SNPs and 1 INDEL being removed. Interestingly, more transversions were removed, with TS/TV ratios increasing for every gene, except for FAD2.A1, which TS/TV ratio lowered from 1.8 to 0.5. A total of 33 SNPs and 8 INDELs were left for the last filtering step. Lastly, variation which occurred in more than 6 mutant lines were removed. This step filtered out all remaining INDELs and 11 SNPs, leaving 22 SNPs left for further analysis.

3.3. Validation by Practical Work

After identification of variation by computational analysis, candidate mutations were selected to be validated in labs. Variants were picked according to their occurrences. While during computational analysis variants were filtered according to their occurrences across all experimental lines, here mutations were selected if they occurred more than 60% within a single experimental line. This is to ensure that the candidate variants arise enough times to be detected by wet lab experiments. Hence, the aim of the following set of experiments is to validate the mutations detected by the computational pipeline.

Firstly, FASTA sequences of the experimental lines were aligned to determine the position of the variant. Secondly, sequencing data was analysed to confirm the presence of the candidate mutations.

3.3.1. Primer Design

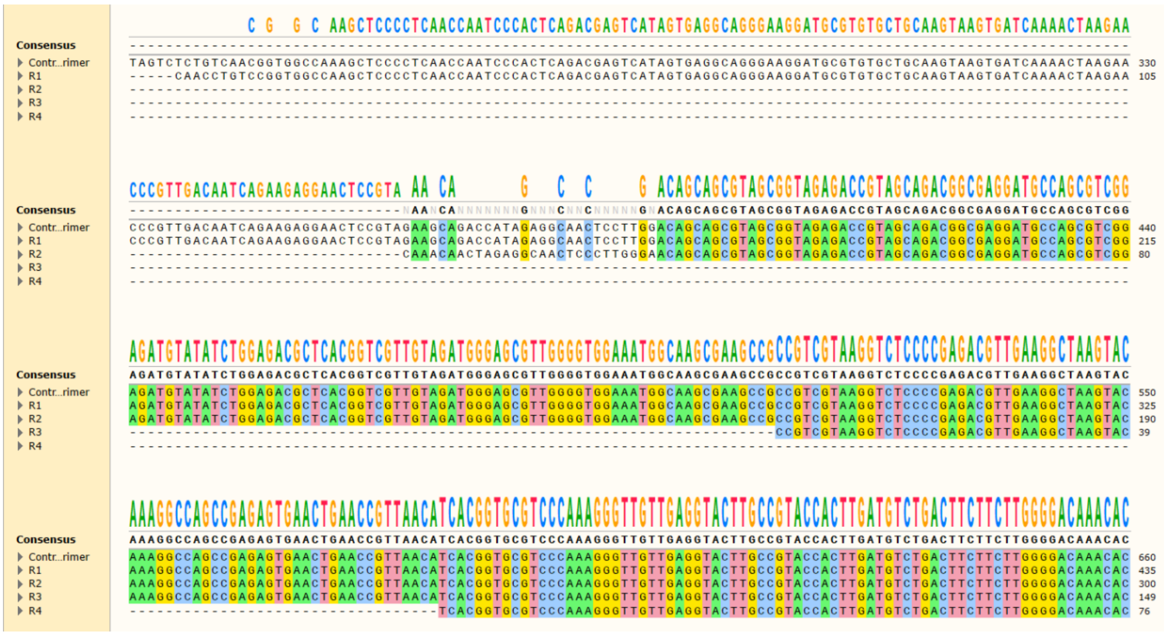


Figure 11: The designed primers amplified the expected genomic region. Primers were designed descending in length. Sequencing data confirmed that the designed primers amplified the region of interest.

For amplification by PCR, one forward and four different reverse primers were designed. The observed length corresponds with the expected length of PCR products: The control primers resulted in a fragment length of 1113 bases, R1, R2, R3 and R4 showed a length of 892, 757, 606 and 536, respectively (see Fig. 11 and 12).

Further, we were able to confirm that the tested primers were allele-specific: Except for a small region (see Fig. 12), all primers showed clean, single-peak amplification.

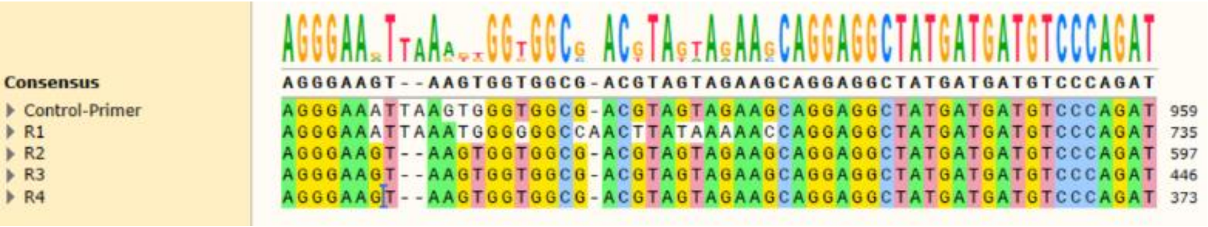
3.3.2. Selected Variants

The following variants and experimental lines were picked to be analysed through practical work (see Table 7). After selection of variants with an occurrence of more than 60% within a sample (see Fig. 9), 4 SNPs in the FAD2 gene remained. To test INDEL detection in silico, we included 2 FAE1 INDELs. Due to occurring more than 6 times, across all samples, the aforementioned INDELs were discarded (see Fig. 10). Including said INDELs allowed us to compare SNP to INDEL detection, as well as variants in the FAD2 and FAE1 gene. Further, the results would give an indication of the effectiveness and accuracy of the last filter used.

Table 7: Overview of variants selected for further analysis.

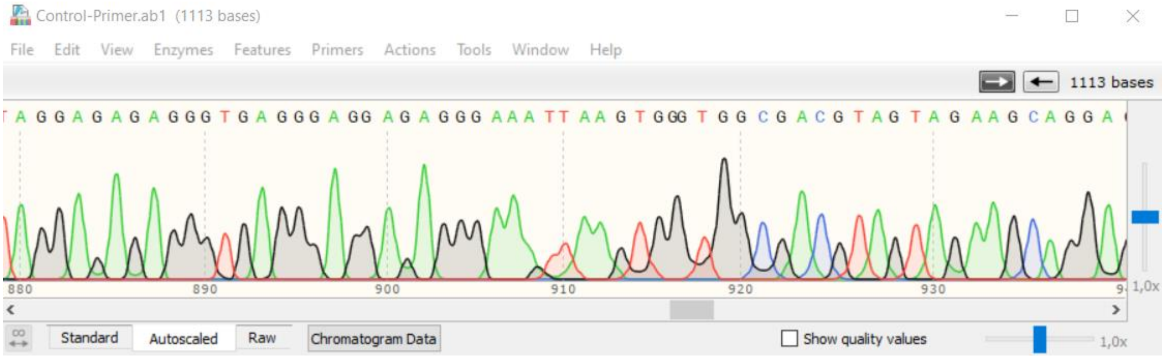
A total of 6 variants, 4 SNPs and 2 INDELs, were selected to be sequenced. All SNPs were identified in the FAD2 gene and both INDELs were detected in the FAE1 gene.

Experimental Line	**Mutation**	**Position**	**Gene**
FNT80-6	G → A	733	FAD2
FNT80-1a	G → A	446	FAD2
G2000-136a	G → A	446	FAD2
G2000-419-1	G → A	446	FAD2
G2000-119-1	4bp deletion	153	FAE1
FNT40-5	4bp deletion	153	FAE1



```
Consensus         AGGGAAGT--AAGTGGTGGCG-ACGTAGTAGAAGCAGGAGGCTATGATGATGTCCCAGAT
▸ Control-Primer  AGGGAAATTAAGTGGGTGGCG-ACGTAGTAGAAGCAGGAGGCTATGATGATGTCCCAGAT  959
▸ R1              AGGGAAATTAAATGGGGGGCCAACTTATAAAAACCAGGAGGCTATGATGATGTCCCAGAT  735
▸ R2              AGGGAAGT--AAGTGGTGGCG-ACGTAGTAGAAGCAGGAGGCTATGATGATGTCCCAGAT  597
▸ R3              AGGGAAGT--AAGTGGTGGCG-ACGTAGTAGAAGCAGGAGGCTATGATGATGTCCCAGAT  446
▸ R4              AGGGAAGT--AAGTGGTGGCG-ACGTAGTAGAAGCAGGAGGCTATGATGATGTCCCAGAT  373
```

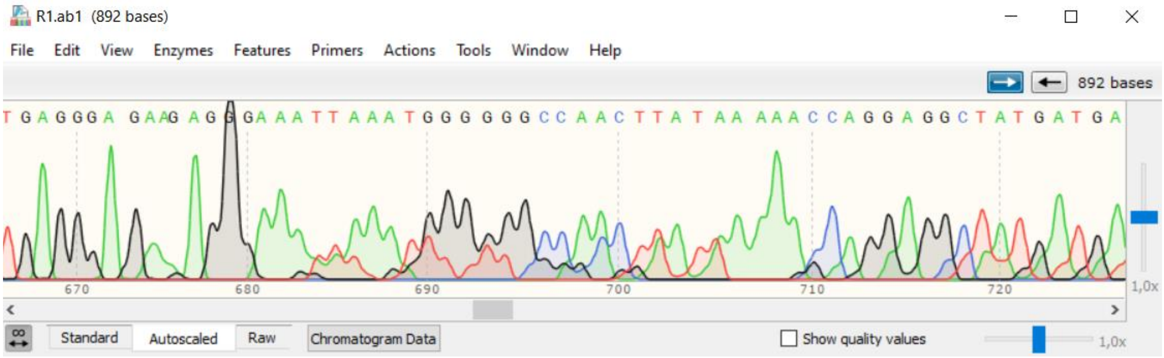
A



Control-Primer.ab1 (1113 bases)

File Edit View Enzymes Features Primers Actions Tools Window Help
1113 bases
T A G G A G A G A G G G T G A G G G A G G A G A G G G A A A T T A A G T G G G T G G C G A C G T A G T A G A A G C A G G A
880
890
900
910
920
930
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

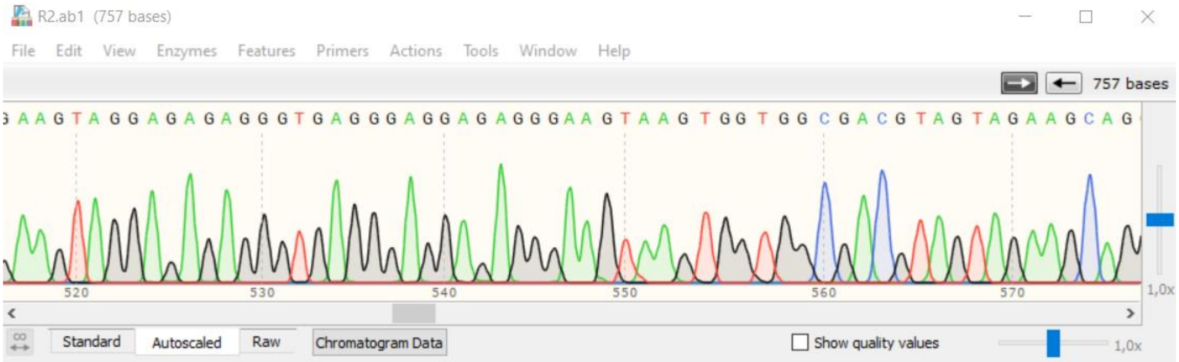
B



R1.ab1 (892 bases)

File Edit View Enzymes Features Primers Actions Tools Window Help
892 bases
T G A G G G A G A A G A G G G A A A T T A A A T G G G G G G C C A A C T T A T A A A A A C C A G G A G G C T A T G A T G A
670
680
690
700
710
720
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

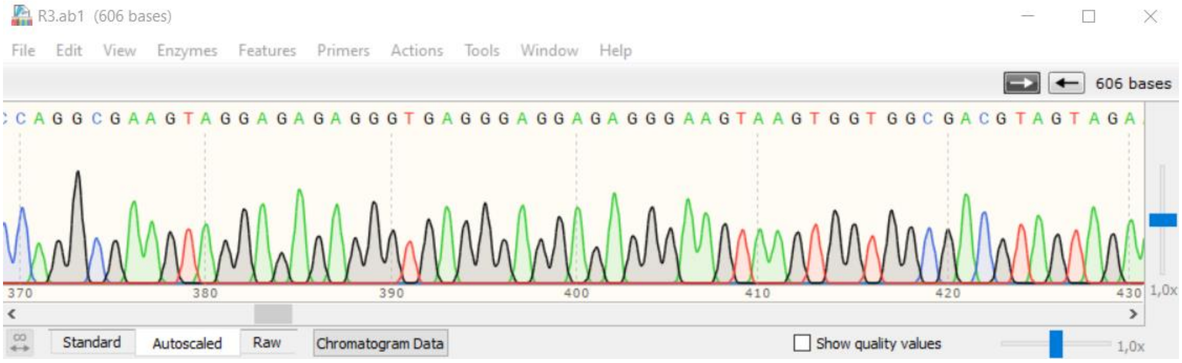
C



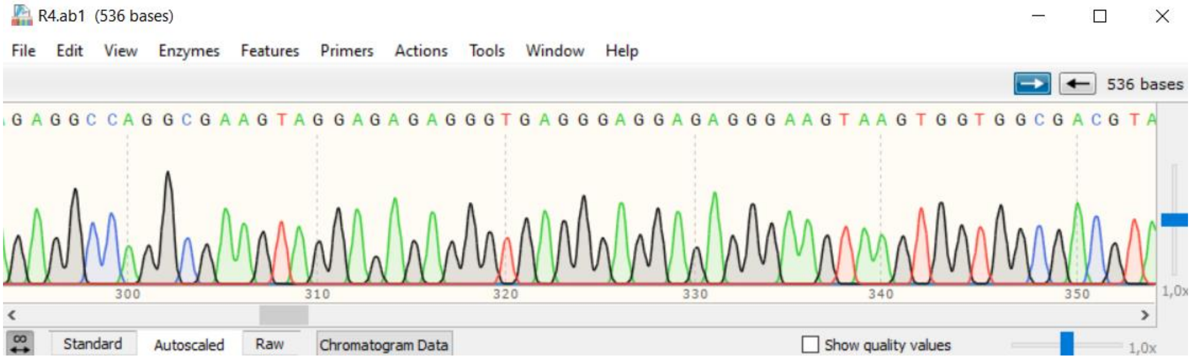
R2.ab1 (757 bases)

File Edit View Enzymes Features Primers Actions Tools Window Help
757 bases
G A A G T A G G A G A G A G G G T G A G G G A G G A G A G G G A A G T A A G T G G T G G C G A C G T A G T A G A A G C A G
520
530
540
550
560
570
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

D


R3.ab1 (606 bases)
File Edit View Enzymes Features Primers Actions Tools Window Help
606 bases
CAGGCGAAGTAGGAGAGAGGGTGAGGGAGGAGAGGGAAGTAAGTGGTGGCGACGTAGTAGA
370 380 390 400 410 420 430
1,0x
Standard Autoscaled Raw Chromatogram Data
□ Show quality values
1,0x

E


R4.ab1 (536 bases)
File Edit View Enzymes Features Primers Actions Tools Window Help
536 bases
GAGGCCAGGCGAAGTAGGAGAGAGGGTGAGGGAGGAGAGGGAAGTAAGTGGTGGCGACGTA
300 310 320 330 340 350
1,0x
Standard Autoscaled Raw Chromatogram Data
□ Show quality values
1,0x

F

Figure 12: Control and R1 primers showed slight discrepancies in sequencing.

A: Alignment of sequences showed some discrepancies for control and R1 primers.

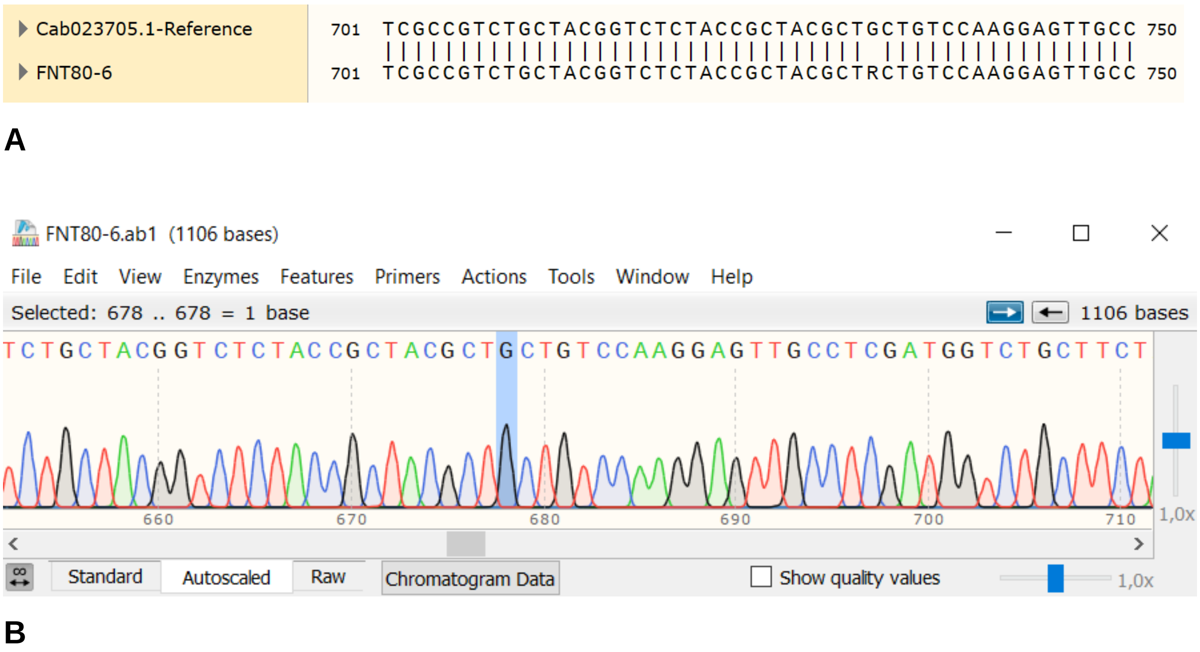
B: Non-specific amplification from position 907-922.

C: Non-specific amplification from 682-709.

D-F: The remaining primers showed no signs of non-specific amplification for the respective region.

3.3.2.1. Variants in FAD2

For the FAD2 gene 4 SNPs were identified and picked for further analysis. Although one SNP was not detected in the FASTA alignment, we decided to still sequence the mutation line to determine the accuracy of variant detection using FASTA alignments.



```
▸ Cab023705.1-Reference   701  TCGCCGTCTGCTACGGTCTCTACCGCTACGCTGCTGTCCAAGGAGTTGCC 750
                               |||||||||||||||||||||||||||||||| |||||||||||||||||
▸ FNT80-6                 701  TCGCCGTCTGCTACGGTCTCTACCGCTACGCTRCTGTCCAAGGAGTTGCC 750
```

A

FNT80-6.ab1 (1106 bases)

Selected: 678 .. 678 = 1 base

1106 bases

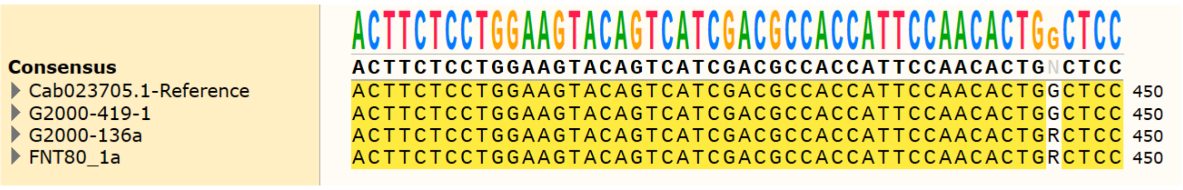
TCTGCTACGGTCTCTACCGCTACGCTGCTGTCCAAGGAGTTGCCTCGATGGTCTGCTTCT

B

Figure 13: No variation was detected at position 733 in the FAD2 gene.

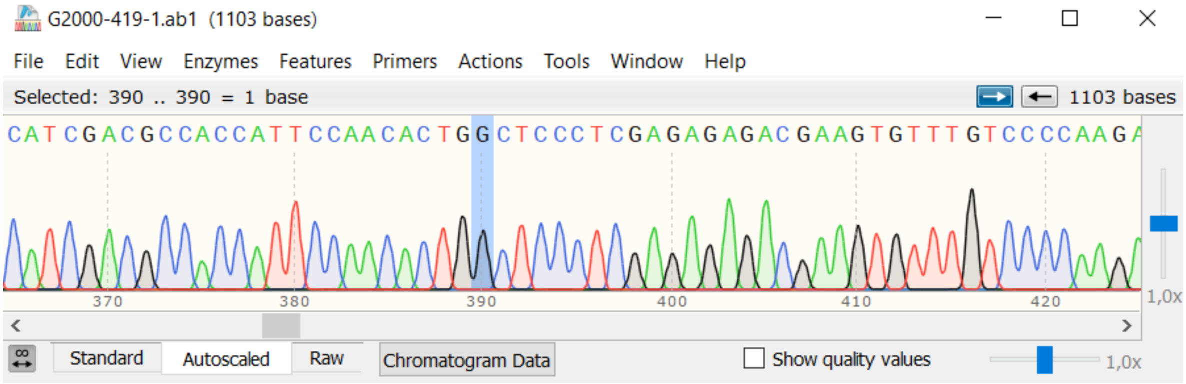
A: Alignment of the experimental line with the reference sequence showed a G → A transition at position 733 of the FAD2 gene.

B: Sequencing data did not show a variation at the corresponding position within the gene.



```
                             ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGGCTCC
Consensus                    ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGNCTCC
Cab023705.1-Reference        ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGGCTCC 450
G2000-419-1                  ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGGCTCC 450
G2000-136a                   ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGRCTCC 450
FNT80_1a                     ACTTCTCCTGGAAGTACAGTCATCGACGCCACCATTCCAACACTGRCTCC 450
```

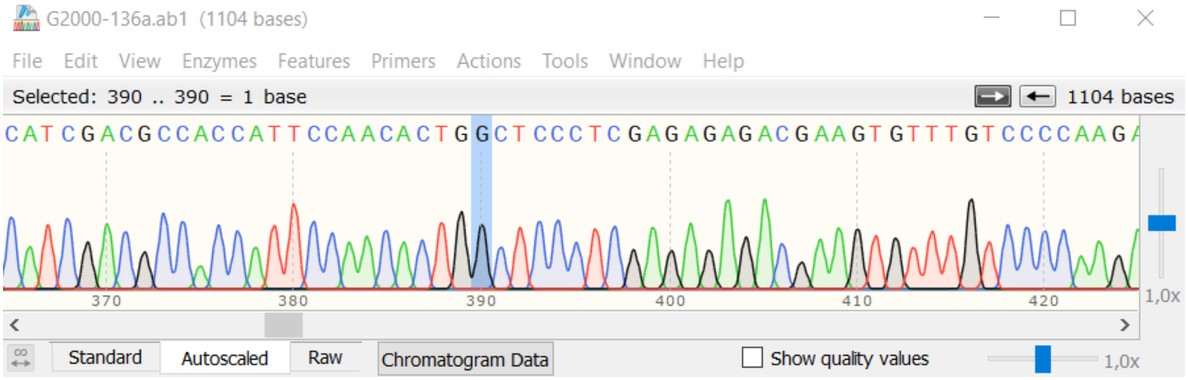
A


G2000-419-1.ab1 (1103 bases)
File Edit View Enzymes Features Primers Actions Tools Window Help
370
380
390
400
410
420
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

Selected: 390 .. 390 = 1 base — 1103 bases

CATCGACGCCACCATTCCAACACTGGCTCCCTCGAGAGAGACGAAGTGTTTGTCCCCAAGA

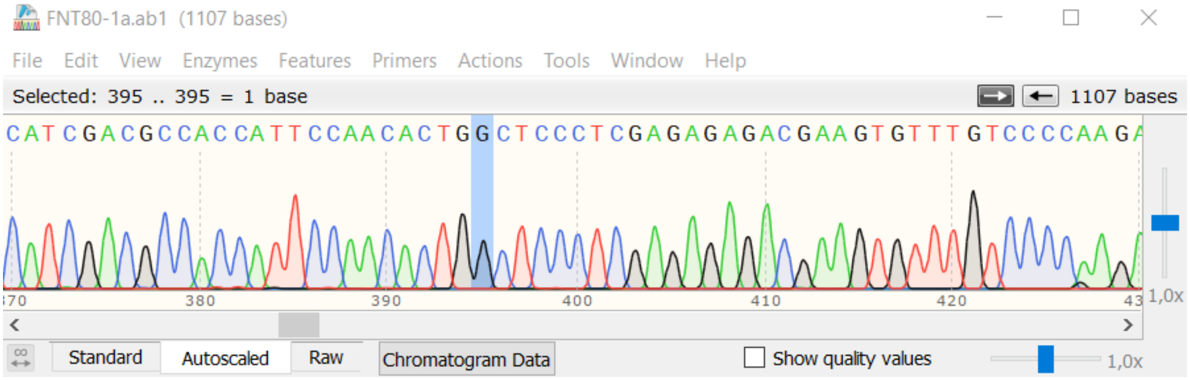
B


G2000-136a.ab1 (1104 bases)
File Edit View Enzymes Features Primers Actions Tools Window Help
370
380
390
400
410
420
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

Selected: 390 .. 390 = 1 base — 1104 bases

CATCGACGCCACCATTCCAACACTGGCTCCCTCGAGAGAGACGAAGTGTTTGTCCCCAAGA

C


FNT80-1a.ab1 (1107 bases)
File Edit View Enzymes Features Primers Actions Tools Window Help
370
380
390
400
410
420
43
1,0x
Standard
Autoscaled
Raw
Chromatogram Data
□ Show quality values
1,0x

Selected: 395 .. 395 = 1 base — 1107 bases

CATCGACGCCACCATTCCAACACTGGCTCCCTCGAGAGAGACGAAGTGTTTGTCCCCAAGA

D

Figure 14: No variation was detected at position 446 in the FAD2 gene.

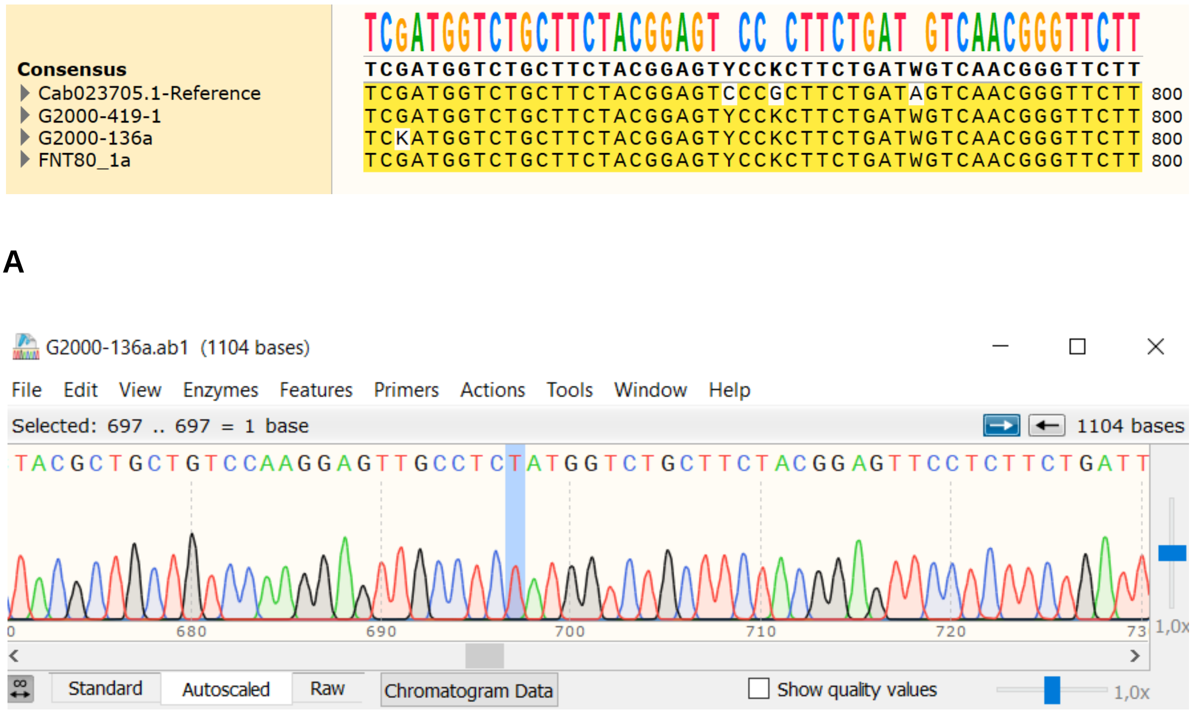
A: A G → A transition was detected in mutational lines FNT80-1a, G2000-136a and G2000-419-1. The latter did not show the respective mutation in the corresponding FASTA file.

B-D: Sequencing data shows a guanine, instead of an expected adenine nucleotide. Therefore no mutation was present in the corresponding mutational lines.

Despite passing all quality filters and variation being present in the FASTA alignments, we were not able to validate any SNPs in the FAD2 gene (see Fig. 13 and 14).

Although almost all FASTA alignments showed variation (except for G2000-419-1, see Fig. 14), no SNPs were visible in the sequencing data corresponding to the mutational lines.

During analysis of the alignments, we identified a possible SNP at position 697 in the FAD2 gene (see Fig. 15). We sequenced the corresponding mutant line (G2000-136a) to determine if an actual SNP was present. However, DNA sequencing showed no variation.



```
                          TCGATGGTCTGCTTCTACGGAGT CC CTTCTGAT GTCAACGGGTTCTT
Consensus                 TCGATGGTCTGCTTCTACGGAGTYCCKCTTCTGATWGTCAACGGGTTCTT
  Cab023705.1-Reference   TCGATGGTCTGCTTCTACGGAGTCCCGCTTCTGATAGTCAACGGGTTCTT 800
  G2000-419-1             TCGATGGTCTGCTTCTACGGAGTYCCKCTTCTGATWGTCAACGGGTTCTT 800
  G2000-136a              TCKATGGTCTGCTTCTACGGAGTYCCKCTTCTGATWGTCAACGGGTTCTT 800
  FNT80_1a                TCGATGGTCTGCTTCTACGGAGTYCCKCTTCTGATWGTCAACGGGTTCTT 800
```

A

G2000-136a.ab1 (1104 bases)

Selected: 697 .. 697 = 1 base

1104 bases

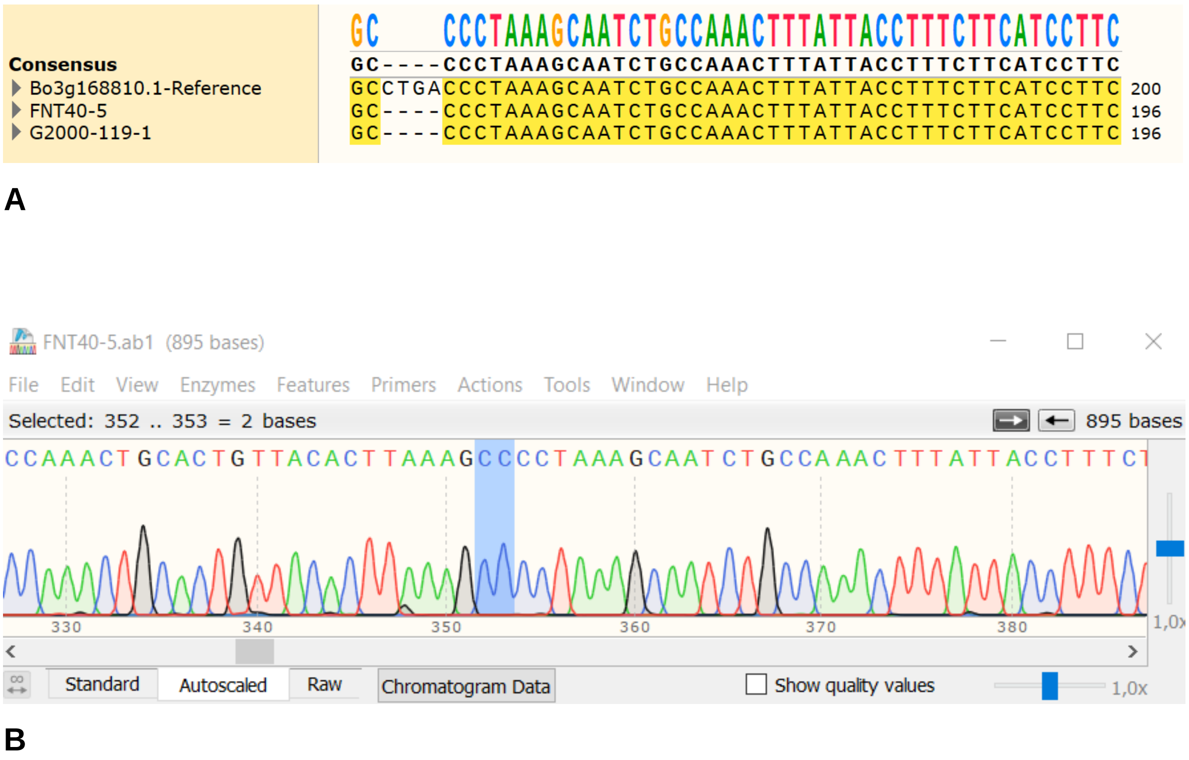
```
TACGCTGCTGTCCAAGGAGTTGCCTCTATGGTCTGCTTCTACGGAGTTCCTCTTCTGATT
```

B

Figure 15: The computational pipeline correctly discarded control variation.
A: A G → T transversion was identified in the G2000-136a in the alignment.
B: While DNA sequencing did show the mutation identified in the alignment, the variant was filtered out during the first step of the variant analysis process, as the same SNP occurred in the control line (data not shown).

3.3.2.2. Variants in FAE1

For the FAE1 gene, 2 INDELs was chosen to be sequenced. Both FASTA alignment and sequencing data showed the presence of a 4bp deletion (see Fig. 16). The respective variation was discarded in the last filtering step, due to occurring more than 6 times. However, our sequencing data indicates said INDEL to be a true variant.



```
Consensus               GC----CCCTAAAGCAATCTGCCAAACTTTATTACCTTTCTTCATCCTTC
▸ Bo3g168810.1-Reference GCCTGACCCTAAAGCAATCTGCCAAACTTTATTACCTTTCTTCATCCTTC 200
▸ FNT40-5               GC----CCCTAAAGCAATCTGCCAAACTTTATTACCTTTCTTCATCCTTC 196
▸ G2000-119-1           GC----CCCTAAAGCAATCTGCCAAACTTTATTACCTTTCTTCATCCTTC 196
```

A

FNT40-5.ab1 (895 bases)

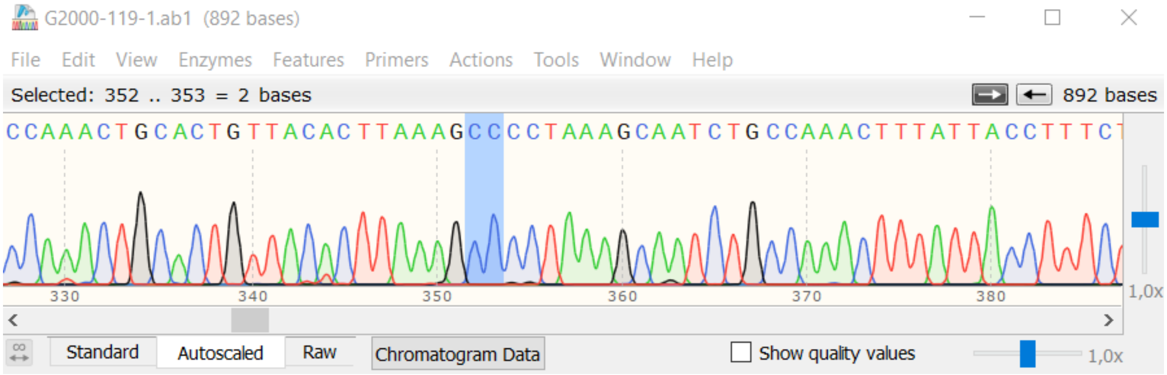
Selected: 352 .. 353 = 2 bases

895 bases

CCAAACTGCACTGTTACACTTAAAGCCCCTAAAGCAATCTGCCAAACTTTATTACCTTTCT

□ Show quality values

B



Selected: 352 .. 353 = 2 bases

CCAAACTGCACTGTTACACTTAAAGCCCCTAAAGCAATCTGCCAAACTTTATTACCTTTCT

C

Figure 16: A 4bp deletion was detected at position 153 of the FAE1 gene.

A: In the alignment a 4bp deletion was visible at position 153 in mutant lines FNT40-5 and G2000-119-1.

B+C: In the sequencing data for both mutant lines, a 4bp deletion was detected.

4. Discussion

4.1. Computational Pipeline

4.1.1. Variant Detection by Clustering

In the initial step for this project we developed clusters to test variant detection using FASTA sequences. This method was found to be not as precise as variant detection using VCF files, as no information about the quality of the variants is included in FASTA files. However, sequence clustering allowed us to gain an insight into mutant lines with high and low erucic acid content: We found that sequences most dissimilar to the reference showed low erucic acid content. Contrary, sequences identical to the reference sequence showed high erucic acid content. However, high dissimilarity to the reference sequence does not necessarily result in decreased erucic acid content. As seen in the B genome (see Fig. 8), lines with high erucic acid content were identified, despite showing great dissimilarity to the reference sequence. Further, samples with low erucic acid content were found regardless of high similarity to the reference sequence.
Interestingly, samples in the A genome imply a relationship between similarity and erucic acid content: Out of 7 clusters, with sequences highly similar or identical to the reference sequence, 5 clusters contained lines with high erucic acid content.

Unfortunately, our data is not conclusive enough to indicate a relationship between erucic acid content and FASTA clusters. One approach to gather more significant data could be to further divide clusters according to their erucic acid content (eg. in upper and lower 10%, 20%, etc.). By doing so, all of the available data would be utilised, which could make any detected trends more conclusive. We decided to highlight only those clusters with sequences with erucic acid content in the upper and lower 10%.

Notably, we found sequences with low and high erucic acid content in the same clusters. Such clusters were not coloured. These findings show a need for more precision. While further division by erucic acid content may increase discrepancies within clusters, it may indicate a tendency towards high or low erucic acid content within the cluster. An explanation for the discrepancies within single clusters may be the combined effect of all mutations: Our clusters were separated by subgenomes, however, some mutations in a different subgenome may have an impact big enough to influence the alignment analysis of the other genome.

Masood and Khan (2015) describe clustering as “an unsupervised data mining technique which is used to place individual artifacts into relevant groups without prior knowledge of distinct group properties to explore structure in the data.” (Masood and Khan 2015). Biologists have been utilising clustering for exploring genomes and orthologues (Zou et al. 2020): Orthologues for the FAE1 gene have been found across members of *Brassicaceae* using sequence clustering (Singh et al. 2017). Further, Li et al. (2014) used SNP clustering according to genotype, to identify SNPs relating to glucosinolate content in *Arabidopsis thaliana*. Then, paralogues of the affected gene (HAG1) were compared to “pseudomolecules” (representative of 19 *B.napus* chromosomes): As a result, four paralogues in *B.napus* were identified.

Clustering approaches can also be utilised to identify rare variants: By clustering individuals into subgroups according to phenotypic distance from the control group, the number of true positives was increased by 17% in comparison to a non-clustering approach (Sun et al. 2016).

The aforementioned studies demonstrated that clustering of various genetic information can be used to explore genomes and identify key-regulating loci. While we also based our clustering on genetic distance, we were not able to identify any SNPs. To identify SNPs, future experiments may take the average erucic acid content of each cluster and map it against genetic distance. Mutations and loci that impact erucic acid content have been discovered: Two (out of six) isoforms of FAE1 are known to regulate erucic acid content in seeds (Lu et al. 2019). Moreover, a 2bp deletion was found to stop erucic acid production (Wu et al. 2015). Utilising this information in combination with genetic distance could help to better understand genetic distances themselves and allow us to draw conclusions about variants. Notably, in our experiments we defined genetic distance as sequence similarity to the reference sequence: Future experiments may find other definitions and approaches more suitable, such as clustering by genotype.

Because of the nature of clustering approaches we used for our experiments, we were not able to identify the exact loci accounting for genetic distance (and therefore sequence similarity).

While we separated sequences based on their position in the genome, a possible next step in the development of FASTA clusters could be to cluster the sequences for FAD2 and FAE1 genes separately. By doing so, the precision would be increased and the results could provide a better understanding of sequence similarity and its relationship to erucic acid content. Further, as FAD2 and FAE1 have orthologues in both A and B genome, the results could help pinpoint relevant loci in each genome. Clusters with high or low content could then be easily analysed and their sequences further investigated to detect variation.

4.1.1.1. Reference Genomes and Alignment-free Methods

One main advantage of FASTA clusters is that there is no requirement for a reference sequence. Alignment of sequences requires a high-quality reference genome to effectively call SNPs and genotypes (Chen et al. 2013). While we did utilise the available reference sequence to identify identical sequences, this is purely optional. This step allowed us to filter out sequences with no variation present.

Although reference genomes for many species have been published, new and improved versions are being continuously released. In *B. napus*, the latest Darmor-bzh reference genome was assembled using long-read sequences and nanopore technology (Rousseau-Gueutin et al. 2020). However, Song et al. (2020) claims that most reference genomes still lack accuracy and completeness, impeding the detection of structural variants (SVs). Pan-genomes are proposed to aid identification of SVs: Per definition, pan-genomes represent core genes and dispensable genes of a species (Tettelin et al. 2005). In the human genome pan-genomes were found to improve variant calling in highly polymorphic regions significantly (Valenzuela et al. 2018). These results are promising for polyploid species: Polyploids often include large genomes and high similarity between duplicated chromosomes (Paape et al. 2018). Further, the use of short sequences may lead to the break of contigs in polymorphic regions (Claros et al. 2012). Pan-genomes overcome such challenges in variant detection. Additionally, pan-genomes aid in understanding the global complexity of polyploid genomes (Morgante et al. 2007). However, assembly of a pan-genome for large genomes is both computationally demanding and costly, hence why transcriptomes are a viable option for larger genomes. Transcriptomes model the expressed genes of a species using pan-genome data (Contreras-Moreira et al. 2017). The used transcriptome for this project was developed by He et al. (2015) and uses coding DNA sequences (CDS) of *Brassica* A and C genomes. Combining gene models from *B. oleracea* TO100 and *B. napus* Darmor-bzh resulted in increased collinearity, especially in comparison to Darmor-bzh resources. Therefore, the present transcriptome is a suitable reference data for our project.

Alignment of sequences of polyploid species is especially challenging: Polyploid species require twice the depth as diploid species to ensure full genome coverage, resulting in increased cost to reach the sufficient depth required. Moreover, the use of short read sequencing technologies further impedes the ability to appropriately call variants. A high Minor Allele Frequency (MAF) threshold distinguishes between heterozygotes from sequencing errors in diploid genomes, but it cannot be applied to polyploid species: Low MAFs often indicate variation on rare alleles. This challenge is worsened by the low availability of high-quality polyploid genomes (VanWallendael and Alvarez 2022).

One way to bypass these obstacles is to utilise alignment-free variant calling techniques. Here, k-mers, subsamples of sequences with length k, are being calculated. Analyses which utilise k-mers were found to be more flexible and computing power-efficient than their alignment-based counterparts (Leimeister et al. 2014; Vinga and Almeida 2003). Studies (Ranallo-Benavidez et al. 2020) in polyploid species also successfully integrated k-mers based methods.

So far, there is little research available on alignment-free methodologies in polyploid species. However, the current data suggests a great potential for such analyses: Until high-quality genomes from long read NGS are assembled, alignment-free methods can detect variation from short read sequences (Voichek and Weigel 2020). Notably, alignment-free methods analyse the entire sequencing set, whereas established variant-calling pipelines normally consider only loci with a high depth coverage. As a result, poorly sequenced samples show higher similarity to each other. VanWallendael and Alvarez (2022) suggests normalisation by library size to mitigate the influence of poorly sequenced reads.

For our studies on variant detection by alignment, the use of k-mers may add the flexibility required to analyse polyploid sequences: We used the *dist.logDet* command from the phangorn R package (Schliep 2021) to calculate pairwise distances. Sequences have been converted to a phyDat object prior. We decided to use the phangorn package because its algorithms were the most sensitive: Different packages, nor other commands within the package, were not able to detect any differences between the sequences. The used command, *dis.logDet*, is based on the calculations of (Lockhart et al. 1994). While this algorithm mitigates the effects of base composition bias among sequences (Kaltenpoth et al. 2012), Jermiin et al. (2009) argues the simplicity of the calculations may lead to false results due to homology of sites in alignments being inappropriately accounted for.

4.1.2. Software Used

The present study developed a computational pipeline to detect SNPs and INDELs in *B.napus*. The computational analysis was carried out using, among others, SAMTools and VCFTools. While SAMTools is a commonly used tool in the field of bioinformatics, the Genome Analysis Toolkit (GATK; McKenna et al. 2010) is yet another popular software. GATK’s HaplotypeCaller command is frequently used to call SNPs. Further, for INDELs GATK offers the IndelRealigner command, which corrects mapping errors and makes INDEL containing regions more consistent. However, we decided to use SAMTools for this projects due to multiple reasons: Firstly, as SAMTools is widely used, plenty of resources are available online. Secondly, the initial alignment has been carried out with BWA-mem, making SAMTools the most compatible variant calling tool, which could have been used: A study (Yao et al. 2020) researched the precision of different alignment tools and the ability of various variant calling tools to accurately detect variation downstream. This research concluded that for polyploid organisms, BWA-mem and SAMTools are the most accurate tools to call variants with. GATK showed, next to SAMTools and FreeBayes, the lowest number of missed calls among all mapping tools. As previously mentioned, SAMTools was found to be the most precise tool with BWA-mem, whereas GATK showed less missed calls in Bowtie2.

4.1.3. Transitions and Transversions in *Brassica Napus*

To classify the detected variants and to compare our results to those from similar studies, we calculated the TS/TV ratio for each gene after application of filters (see Fig. 10). The overall amount of transitions and transversions detected in this study matches with those present in other studies: In the raw dataset we observed 55,6% transitions (48% A/G, 52% T/V) and 44,4% transversions (33% A/C, 18% A/T, 26% G/C, 23% G/T), resulting in a TS/TV of 1.25. Using Restriction-site Associated DNA (RAD) sequencing, Bus et al. (2012) reported 58,2% transitions (49,7% A/G, 50,3% C/T) and 41,8% transversions (26,5% A/C, 29,7% A/T, 17% G/C and 26,8 G/T), totalling to a TS/TV ratio of 1.39. Another study (Barchi et al. 2011) utilising RAD sequencing in eggplants reported an even higher TS/TV ratio of 1.65. However, in a genome-wide study on polymorphisms in *B. rapa* (Park et al. 2010), over 21,000 SNPs were detected and a TS/TV ratio of 1.03 was reported. Therefore we can conclude that our observed ratio fits into the range of ratios reported by other studies. Notably, Park et al. (2010) reported a TS/TV ratio of 1.03 in both exon and introns. After filtering, a ratio of 1.63 in exons only was observed. This aligns with our observations: For SNPs in coding regions only, we report a TS/TV ratio of 1.5, 61% transitions (54% A/G, 46% T/C) and 39% transversions (47% A/C, 13% A/T, 21% G/C and 19% G/T). Bus et al. (2012) observed a higher TS/TV ratio (1.6) as well.

These findings give rise to the question why more transitions are found in coding regions than in introns. Park et al. (2010) reports a higher frequency of transitions in exon regions (61.9%) compared to intron regions (52.7%). The study further analysed the ratio of nonsynonymous to synonymous SNP rates per site (Ka/Ks), as selection pressure is variable across loci in *B.rapa.* Genes with a mean Ka/Ks lower than 0.1 include protein and nucleic acid binding proteins, meaning those sites are subject to high selective pressure. This study reported an average Ka/Ks ratio of 0.18, indicating strong natural selection in *B.rapa.*

Although TS/TV ratios are only an approximate measure of quality (Carson et al. 2014), high quality datasets are expected to have TS/TV ratios between 2.8 and 3 (DePristo et al. 2011). However, the authors were analysing human NGS data; Research in other polyploid plants observed lower TS/TV ratios. It is therefore unclear if the proposed ideal ratio is applicable to polyploid species.

4.2. Validation by Practical Work

4.2.1. Variant Detection by Alignment

A total of 4 SNPs and 2 INDELs were identified and chosen to be sequenced. These variants passed all filters of the computational pipeline. Although one SNP was not detected in the FASTA alignment (G2000-419-1, see Fig. 14, we decided to still sequence the mutation line to determine the accuracy of using FASTA alignments for variant detection. The same SNP was detected in other mutant lines and in the VCF file for the corresponding mutant line, hence why we decided to still sequence the mutation. This would also show an indication of the accuracy of FASTA alignments for variant detection. Further, a SNP was identified during alignment of

FASTA sequences (see Fig. 14). The resulting sequencing data did show a SNP at the corresponding site. However, after manual inspection, the in the FASTA alignment detected SNP was found in the VCF for control lines. This indicates that our pipeline successfully filtered out control variation.

For our project we used sequencing data from cultivars, which were not treated with radiation. As such, we were able to identify variation which was not induced by gamma-radiation and therefore discarded. However, variant detection by alignment of FASTA sequences impedes this distinction: FASTA files include no information about variants, therefore variants can only be detected by alignment. While removing control variation with FASTA files is computationally possible, it would require a much higher effort to achieve this. As previously shown, the detected variation does not necessarily equal true variation. Further sequencing information is required to confirm the presence of a true variant – information which is stored in VCF files. We can therefore conclude that alignment of DNA sequences is too imprecise to be used for variant detection. However, this approach can be used to confirm previously detected mutations.

Notably, the used FASTA sequences did include ambiguity codes. Ambiguity codes are used if at any given position more than one allele was observed. However, the difference needs to be recorded sufficient times for it to be included in the sequence (through the ambiguity code). For a more stringent variant calling approach, the ambiguity code could be altered utilising the information present in FASTQ and BAM files. Various R packages and scripts are already available to disambiguate FASTA files. Using the sequencing information available, the threshold to include or exclude variant alleles could be changed. For instance, the ambiguity code could be completely left out and a variant allele could be called if the respective variant occurs in more than 70% of observations. As a result, variation detected by sequence alignment would more likely be actual variation detectable by sequencing. While this approach would reduce the number of detected variants, the ones present are more likely to be true variants.

A less stringent approach could be to reduce the number of filters used, which would shift the focus onto detection by sequencing. This approach increases the laboratory work, however, it may lead to detection of variation, which would have been discarded otherwise.

We aimed to develop a rather stringent pipeline: While similar filters were used in other experiments (QUAL > 20; Bus et al. 2012, QUAL > 30, depth > 5; Yu et al. 2021), we further filtered variation out according to their occurrences across all mutational lines and within a single mutational line. Firstly, we filtered out all variations occurring more than 6 times across all sequences: Because we were using roughly 600 mutational lines, we set a threshold of 1% to avoid identifying false positives. As seen as in Fig. 16, we identified one INDEL in the FAE1 gene, however, the said variation was discarded due to high occurrence. This raises the question of whether our pipelines used too stringent filters: Ebbert et al. (2016) demonstrated that PCR duplicate removal did not have a significant effect on the accuracy of variant datasets. Notably, their findings are based on human Whole Genome Sequencing (WGS) data, so far there is no data supporting their claims in polyploid species. Still, our results do indicate that PCR duplicate removal is not a necessary step in our variant calling pipeline: None of the SNPs that passed all filters were proven to be true variants, yet one INDEL that was filtered out was found in sequenced DNA. Therefore, we recommend following a less stringent approach, as usage of too many or stringent filters leads to removal of true variation. One way to bypass PCR amplification bias is the incorporation of unique molecular identifiers (UMIs): Because molecules in the starting pool are barcoded with a unique UMI, reads with the same UMI must be PCR duplicates (Fu et al. 2018). Hereby a less stringent approach can be accurately used, while reducing the number of false positive variation.

We also sequenced variation which was only present in the respective FASTA sequence (see Fig. 14) to test whether variant detection by alignment was a considerable alternative to established pipelines. While the confirmed INDEL was present in FASTA alignments, so were SNPs. This indicates that variant detection by alignment of sequences was too imprecise to be used for our experiment: Although sequence alignments and clusters can be useful in gaining an overview of the data, the FASTA file type lacks the information to ensure a certain standard of quality is met. Further, as the file format does not include any details about the sequences, its usage is restricted to sequence alignments. Contrarily, the parameters present in VCF files can be utilised for a range of different experiments. For instance, a Genome-Wide Association Study (GWAS; Wang et al. 2018) used a Minor Allele Frequency (MAF) > 0.05 filter to collect variation relevant to their type of study.

Interestingly, Jones et al. (2009) states that alignment of sequences was the simplest way to detect SNPs. Jones argues that alignments can also be used to detect SNPs in non-coding genomic regions. However, this holds true for variant detection by VCF as well. In fact, we had to include a filtering step to exclude variants in non-coding regions, due to the scope of the experiment. Certainly, detection of variants in introns is crucial in understanding the relationship between genes and their cis- and trans-regulatory elements. Investigating polymorphisms in non-coding regions is particularly important for the detection of INDELs and understanding their effects:

One motivation to research polymorphisms is the development of genetic markers. Although both SNPs and INDELs can be used as such, it is suggested that INDELs may be stronger genetic markers, due to their bigger impact on protein structure and function than SNPs (Rokas and Holland 2000). Moreover, the resulting conformational changes in the protein's structure may also lead to significant changes in a trait's expression, as seen in mitochondrial genes. Lastly, depending on the secondary structure of the protein, INDELs can have a varying effect on the overall protein structure. INDELs within α-helices and ß-sheets can have a bigger effect on protein structure than those occurring in loops and turns (Kim and Guo 2010).

4.3. Limitations and Future Work

Although the present computational pipeline does differ in filtering of SNPs and INDELs, further differentiation may improve INDEL calling. While applying the same quality filters for both SNPs and INDELs is a common approach (Li et al. 2019; Yao et al. 2020), more INDEL-specific filters could be applied to improve detection of INDELs. For instance, GATK offers a local realignment step which removes frameshifts and therefore reduces the number of false-positive calls (Polyanovsky et al. 2011). Moreover, GATK includes a variant quality score recalibration (VQSR), which further helps differentiating between true variants and sequencing noise (McKenna et al. 2010; Clevenger et al. 2015).

As mentioned in Chapter 4.1.2, we used BWA-mem and SAMTools for reference mapping and variant calling, respectively. This was because the combination of said software was found to be the most precise for variant calling (Yao et al. 2020). Although BWA-mem performed better than Bowtie2 (Clevenger et al. 2015; Yao et al. 2020), Bowtie2 showed better handling of INDELS. Further, GATK was found to perform best when used in combination with Bowtie2. Therefore, future INDEL-detecting pipelines should utilise GATKs INDEL-specific commands in combination with Bowtie2.

Future studies might follow a less stringent approach and filter variants by their predicted effect: Software, such as the Ensembl Variant Effect Predictor (VEP; McLaren et al. 2016), predict the impact variants may have on the protein sequence of a gene. VEP, for instance, is available as a downloadable Perl script and can be easily implemented into any variant detecting pipeline. Including this step may draw a connection between FASTA clusters and variant calling: By selecting variants with great impact on the protein and identifying those in clusters, more conclusive results may arise. Further, by separating variants according to genome, clusters of the respective subgenomes and their relationships with orthologues may be better understood. Finally, VEP data may confirm the presence of the identified INDEL in FAE1 and simulate the effect said variation has on the oleic acid pathway.

5. Conclusion

The present project researched variant detection in the allopolyploid species *B.napus*. As, due to the wealth of genomic data, variant detection in polyploid species remains challenging, this project aimed to develop a pipeline to call variants with. Our pipline was able to filter false-positive variants out, yet 2 INDELs, which were present in sequencing data, were discarded as well. With adjustments in filtering methods, we can conclude that the present variant calling pipeline offers a basis for future pipelines.

Depending on the expectations of following experiments, the present pipeline can be developed to be less or more stringent: A more stringent approach may reduce the amount of detected SNPs but increases the likelihood of detecting true SNPs. On the other hand, a less stringent procedure increases the number of variants and therefore laboratory efforts to validate variations. However, the latter approach may give rise to variations which would have been discarded otherwise. Given our results, we recommend using less stringent filters to minimize the risk of discarding true variants. UMIs and INDEL-specific commands may mitigate the effects usage of more lenient filters have on data quality and accuracy.

We showed that FASTA clustering offers a simple but valuable insight into sequence variation. While this approach may not be precise enough to detect exact loci of variants, we suggest combining clustering with variant effect information. Other clustering approaches, such as clustering by genotype or separation by subgenome, may help in understanding orthologues and loci that impact erucic acid content.

6. References

Acquaah, G. (2012) ‘Breeding self-pollinated species’, *Principles of Plant Genetics and Breeding. Wiley-Blackwell, Oxford, UK*, pp. 303–336.

Adams, M.D. *et al.* (1991) ‘Complementary DNA sequencing: expressed sequence tags and human genome project’, *Science*, 252(5013), pp. 1651–1656.

Ahloowalia, B.S. and Maluszynski, M. (2001) ‘Induced mutations--A new paradigm in plant breeding’, *Euphytica/ Netherlands journal of plant breeding*, 118(2), pp. 167–173.

Allam, A., Kalnis, P. and Solovyev, V. (2015) ‘Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data’, *Bioinformatics* , 31(21), pp. 3421–3428.

Al-Shehbaz, I.A. (2012) ‘A generic and tribal synopsis of the Brassicaceae (Cruciferae)’, *TAXON*, pp. 931–954. doi:10.1002/tax.615002.

Altshuler, D. et al. (2000) ‘An SNP map of the human genome generated by reduced representation shotgun sequencing’, Nature, pp. 513–516. doi:10.1038/35035083.

Appleby, N., Edwards, D. and Batley, J. (2009) ‘New Technologies for Ultra-High Throughput Genotyping in Plants’, *Plant Genomics*, pp. 19–39. doi:10.1007/978-1-59745-427-8_2.

Badawy, I.H., Atta, tsand Ahmed, W.M. (1994) ‘Biochemical and toxicological studies on the effect of high and low erucic acid rapeseed oil on rats’, *Die Nahrung*, 38(4), pp. 402–411.

Barchi, L. *et al.* (2011) ‘Identification of SNP and SSR markers in eggplant using RAD tag sequencing’, *BMC Genomics*. doi:10.1186/1471-2164-12-304.

Baum, D. (2008) “Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups”, *Nature Education* 1(1):190.

Berthelot, C. *et al.* (2014) ‘The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates’, *Nature communications*, 5, p. 3657.

Bhunia, R.K., Kaur, R. and Maiti, M.K. (2016) ‘Metabolic engineering of fatty acid biosynthetic pathway in sesame (Sesamum indicum L.): assembling tools to develop nutritionally desirable sesame seed oil’, *Phytochemistry Reviews*, 15(5), pp. 799–811.

BLAST TOPICS (no date). Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp (Accessed: 28 August 2022).

Botstein, D. *et al.* (1980) ‘Construction of a genetic linkage map in man using restriction fragment length polymorphisms’, *American journal of human genetics*, 32(3), pp. 314–331.

Britten, R.J. and Kohne, D.E. (1968) ‘Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms’, *Science*, 161(3841), pp. 529–540.

Brogna, S., McLeod, T. and Petric, M. (2016) ‘The Meaning of NMD: Translate or Perish’, *Trends in genetics: TIG*, 32(7), pp. 395–407.

Browse, J. and Somerville, C. (1991) ‘Glycerolipid Synthesis: Biochemistry and Regulation’, *Annual review of plant physiology and plant molecular biology*, 42(1), pp. 467–506.

Bus, A. *et al.* (2012) ‘High-throughput polymorphism detection and genotyping in Brassica napus using next-generation RAD sequencing’, *BMC genomics*, 13, p. 281.

Carson, A.R. *et al.* (2014) ‘Effective filtering strategies to improve data quality from population-based whole exome sequencing studies’, *BMC bioinformatics*, 15, p. 125.

Cerca, J. *et al.* (2021) ‘Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms’, *Methods in ecology and evolution / British Ecological Society* [Preprint], (2041-210X.13562). doi:10.1111/2041-210x.13562.

Chalhoub, B. *et al.* (2014) ‘Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome’, *Science*, pp. 950–953. doi:10.1126/science.1253435.

Chang, N.W. and Huang, P.C. (1998) ‘Effects of the ratio of polyunsaturated and monounsaturated fatty acid to saturated fatty acid on rat plasma and liver lipid concentrations’, *Lipids*, 33(5), pp. 481–487.

Chen, S. *et al.* (2008) ‘Divergent patterns of allelic diversity from similar origins: the case of oilseed rape (Brassica napus L.) in China and Australia’, *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, 51(1), pp. 1–10.

Chen, X. *et al.* (2013) ‘Detection and genotyping of restriction fragment associated polymorphisms in polyploid crops with a pseudo-reference sequence: a case study in allotetraploid Brassica napus’, *BMC genomics*, 14, p. 346.

Chung, Y.S. *et al.* (2017) ‘Genotyping-by-sequencing: a promising tool for plant genetics research and breeding’, *Horticulture, Environment, and Biotechnology*, 58(5), pp. 425–431.

Claros, M.G. *et al.* (2012) ‘Why assembling plant genome sequences is so challenging’, *Biology*, 1(2), pp. 439–459.

Clevenger, J. *et al.* (2015) ‘Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations’, *Molecular plant*, 8(6), pp. 831–846.

Cock, P.J.A. et al. (2010) ‘The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants’, Nucleic acids research, 38(6), pp. 1767–1771.

Collins, A., Lau, W. and De La Vega, F.M. (2004) ‘Mapping genes for common diseases: the case for genetic (LD) maps’, *Human heredity*, 58(1), pp. 2–9.

Contreras-Moreira, B. *et al.* (2017) ‘Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species’, *Frontiers in plant science*, 8, p. 184.

Cooke, D.P., Wedge, D.C. and Lunter, G. (no date) ‘Benchmarking small-variant genotyping in polyploids’. doi:10.1101/2021.03.29.436766.

Cunningham, F. et al. (2015) ‘Improving the Sequence Ontology terminology for genomic variant annotation’, Journal of biomedical semantics, 6, p. 32.

Danecek, P. *et al.* (2011) ‘The variant call format and VCFtools’, *Bioinformatics* , 27(15), pp. 2156–2158.

Dar, A.A. *et al.* (2017) ‘The FAD2 Gene in Plants: Occurrence, Regulation, and Role’, *Frontiers in plant science*, 8, p. 1789.

Davey, J.W. *et al.* (2011) ‘Genome-wide genetic marker discovery and genotyping using next-generation sequencing’, *Nature reviews. Genetics*, 12(7), pp. 499–510.

DePristo, M.A. *et al.* (2011) ‘A framework for variation discovery and genotyping using next-generation DNA sequencing data’, *Nature genetics*, 43(5), pp. 491–498.

Deschamps, S., Llaca, V. and May, G.D. (2012) ‘Genotyping-by-Sequencing in Plants’, *Biology*, 1(3), pp. 460–483.

Ebbert, M.T.W. *et al.* (2016) ‘Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches’, *BMC bioinformatics*, 17 Suppl 7, p. 239.

Edwards, D. *et al.* (2007) ‘What Are SNPs?’, *Association Mapping in Plants*, pp. 41–52. doi:10.1007/978-0-387-36011-9_3.

Ellegren, H. (2000) ‘Microsatellite mutations in the germline: implications for evolutionary inference’, *Trends in genetics: TIG*, 16(12), pp. 551–558.

EMBL-EBI (no date) *Understanding VCF format*. Available at: https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/understanding-vcf-format/ (Accessed: 10 July 2021).

Ewing, B. *et al.* (1998) ‘Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment’, *Genome Research*, pp. 175–185. doi:10.1101/gr.8.3.175.

Food and Agriculture Organization of the United Nations (2018) *2017 The State of Food Security and Nutrition in the World: Building resilience for peace and food security*. Food & Agriculture Org.

Food and Agriculture Organization of the United Nations *et al.* (2021) *The State of Food Security and Nutrition in the World 2021: Transforming food systems for food security, improved nutrition and affordable healthy diets for all*. Food & Agriculture Org.

Fu, Y. et al. (2018) ‘Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers’, BMC genomics, 19(1), p. 531.

Fu, Y. *et al.* (2021) ‘Identification and Development of KASP Markers for Novel Mutant Alleles Associated With Elevated Oleic Acid in’, *Frontiers in plant science*, 12, p. 715633.

Garrison, E. *et al.* (2018) ‘Variation graph toolkit improves read mapping by representing genetic variation in the reference’, *Nature Biotechnology*, pp. 875–879. doi:10.1038/nbt.4227.

Gilchrist, E.J. *et al.* (2006) ‘TILLING is an effective reverse genetics technique for Caenorhabditis elegans’, *BMC genomics*, 7, p. 262.

Gillingham, L.G., Harris-Janz, S. and Jones, P.J.H. (2011) ‘Dietary Monounsaturated Fatty Acids Are Protective Against Metabolic Syndrome and Cardiovascular Disease Risk Factors’, *Lipids*, pp. 209–228. doi:10.1007/s11745-010-3524-y.

Gomez-Campo, C. (1999) *Biology of Brassica Coenospecies*. Elsevier.

Graham, C.F. et al. (2020) ‘How “simple” methodological decisions affect interpretation of population structure based on reduced representation library DNA sequencing: A case study using the lake whitefish’, PLOS ONE, p. e0226608. doi:10.1371/journal.pone.0226608.

von Grebmer Jill Bernstein Miriam Wiemers Tabea Schiffer Asja Hanano Olive Towey Réiseal Ní Chéilleachair Connell Foley Seth Gitter Kierstin Ekstrom Heidi Fritschel, K. (2021) ‘2021 GLOBAL HUNGER INDEX HUNGER AND FOOD SYSTEMS IN CONFLICT SETTINGS’.

Guo, C. et al. (2017) ‘Transversions have larger regulatory effects than transitions’, BMC genomics, 18(1), p. 394.

Gupta, P.K., Roy, J.K. and Prasad, M. (2001) ‘Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants’, *Current science*, 80(4), pp. 524–535.

Hagiwara, Y. *et al.* (2019) ‘Clustered DNA double-strand break formation and the repair pathway following heavy-ion irradiation', *Journal of radiation research*, 60(1), pp. 69–79.

Harloff, H.-J. *et al.* (2012) 'A mutation screening platform for rapeseed (Brassica napus L.) and the detection of sinapine biosynthesis mutants', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 124(5), pp. 957–969.

Hay, J.E. *et al.* (2016) 'Introduction to the special issue: Observed and projected changes in weather and climate extremes', *Weather and Climate Extremes*, pp. 1–3. doi:10.1016/j.wace.2015.08.006.

Hedrich, H.J. (2012) *The Laboratory Mouse*. Academic Press.

He, Z. *et al.* (2015) 'Construction of Brassica A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research', *Data in brief*, 4, pp. 357–362.

He, Z. *et al.* (2017) 'Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNA seq-based visualization', *Plant biotechnology journal*, 15(5), pp. 594–604.

H-L, L. (1985) 'Rapeseed Genetics and Breeding', *Shanghai Science and Technology Publishing House*, pp. 9–43.

Horner, D.S. *et al.* (2010) 'Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing', *Briefings in Bioinformatics*, pp. 181–197. doi:10.1093/bib/bbp046.

Hossain, Z. *et al.* (2018) 'Transcriptome profiling of Brassica napus stem sections in relation to differences in lignin content', *BMC genomics*, 19(1), p. 255.

Hristov, A.N. *et al.* (2011) 'Effect of replacing solvent-extracted canola meal with high-oil traditional canola, high-oleic acid canola, or high-erucic acid rapeseed meals on rumen fermentation, digestibility, milk production, and milk fatty acid composition in lactating dairy cows', *Journal of dairy science*, 94(8), pp. 4057–4074.

Hu, X. *et al.* (2006) 'Mapping of the loci controlling oleic and linolenic acid contents and development of fad2 and fad3 allele-specific markers in canola (Brassica napus L.)', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 113(3), pp. 497–507.

Ian_Maurer (2020) 'What is a Variant Call Format (VCF) file?' GenomOncology, 9 April. Available at: https://www.genomoncology.com/post/what-is-a-variant-call-format-vcf-file (Accessed: 10 July 2021).

Jermiin, L.S. *et al.* (2009) 'SeqVis: A Tool for Detecting Compositional Heterogeneity Among Aligned Nucleotide Sequences', in Posada, D. (ed.) *Bioinformatics for DNA Sequence Analysis*. Totowa, NJ: Humana Press, pp. 65–91.

Jiang, G.-L. (2013) ‘Molecular Markers and Marker-Assisted Breeding in Plants’, in Andersen, S.B. (ed.) *Plant Breeding from Laboratories to Fields*. Rijeka: IntechOpen.

Jiang, Y., Turinsky, A.L. and Brudno, M. (2015) ‘The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection’, *Nucleic Acids Research*, pp. 7217–7228. doi:10.1093/nar/gkv677.

Jones, N. *et al.* (2009) ‘Markers and mapping revisited: finding your gene’, *The New phytologist*, 183(4), pp. 935–966.

Kaltenpoth, M. *et al.* (2012) ‘Accelerated evolution of mitochondrial but not nuclear genomes of Hymenoptera: new evidence from crabronid wasps’, *PloS one*, 7(3), p. e32826.

Kao, W.-C., Stevens, K. and Song, Y.S. (2009) ‘BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing’, *Genome research*, 19(10), pp. 1884–1895.

Katavic, V. *et al.* (2002) ‘Restoring enzyme activity in nonfunctional low erucic acid Brassica napus fatty acid elongase 1 by a single amino acid substitution’, *European journal of biochemistry / FEBS*, 269(22), pp. 5625–5631.

Kaur, H. *et al.* (2020) ‘The impact of reducing fatty acid desaturation on the composition and thermal stability of rapeseed oil’, *Plant biotechnology journal*, 18(4), pp. 983–991.

Kazama, Y. *et al.* (2017) ‘Different mutational function of low- and high-linear energy transfer heavy-ion irradiation demonstrated by whole-genome resequencing of Arabidopsis mutants’, *The Plant journal: for cell and molecular biology*, 92(6), pp. 1020–1030.

Kiefer, M. *et al.* (2014) ‘BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution’, *Plant & cell physiology*, 55(1), p. e3.

Kim, R. and Guo, J.-T. (2010) ‘Systematic analysis of short internal indels and their impact on protein folding’, *BMC structural biology*, 10, p. 24.

Knutzon, D.S. *et al.* (1992) ‘Modification of Brassica seed oil by antisense expression of a stearoyl-acyl carrier protein desaturase gene’, *Proceedings of the National Academy of Sciences of the United States of America*, 89(7), pp. 2624–2628.

Kochert, G. (1991) ‘Restriction Fragment Length Polymorphism in Plants and Its Implications’, in Biswas, B.B. and Harris, J.R. (eds) *Plant Genetic Engineering*. Boston, MA: Springer US, pp. 167–190.

Koelling, J. *et al.* (2012) ‘Development of new microsatellite markers (SSRs) for Humulus lupulus’, *Molecular breeding: new strategies in plant improvement*, 30(1), pp. 479–484.

Konkol, D. *et al.* (2019) ‘Biotransformation of rapeseed meal leading to production of polymers, biosurfactants, and fodder', *Bioorganic chemistry*, 93, p. 102865.

Landjeva, S., Korzun, V. and Börner, A. (2007) 'Molecular markers: actual and potential contributions to wheat genome characterization and breeding', *Euphytica*, pp. 271–296. doi:10.1007/s10681-007-9371-0.

Lateef, D.D. (2015) 'DNA Marker Technologies in Plants and Applications for Crop Improvements', *Journal of Biosciences and Medicines*, pp. 7–18. doi:10.4236/jbm.2015.35002.

Lauridsen, C. *et al.* (1999) 'Antioxidative and oxidative status in muscles of pigs fed rapeseed oil, vitamin E, and copper', *Journal of animal science*, 77(1), pp. 105–115.

Leimeister, C.-A. *et al.* (2014) 'Fast alignment-free sequence comparison using spaced-word frequencies', *Bioinformatics* , 30(14), pp. 1991–1999.

Lenaerts, B., Collard, B.C.Y. and Demont, M. (2019) 'Review: Improving global food security through accelerated plant breeding', *Plant science: an international journal of experimental plant biology*, 287, p. 110207.

Li, F. *et al.* (2014) 'Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (Brassica napus L.)', *DNA research: an international journal for rapid publication of reports on genesand genomes*, 21(4), pp. 355–367.

Li, F. *et al.* (2019) 'Comparison and Characterization of Mutations Induced by Gamma-Ray and Carbon-Ion Irradiation in Rice (L.) Using Whole-Genome Resequencing', *G3* , 9(11), pp. 3743–3751.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, pp. 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics* , 27(21), pp. 2987–2993.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics* , 25(14), pp. 1754–1760.

Li, R. *et al.* (2008) 'SOAP: short oligonucleotide alignment program', *Bioinformatics* , 24(5), pp. 713–714.

Liu, S. *et al.* (2014) 'The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes', *Nature Communications*. doi:10.1038/ncomms4930.

Li, Y. *et al.* (2011) 'Low-coverage sequencing: implications for design of complex trait association studies', *Genome research*, 21(6), pp. 940–951.

Lobell, D.B. and Gourdji, S.M. (2012) 'The influence of climate change on global crop productivity', *Plant physiology*, 160(4), pp. 1686–1697.

Long, W. *et al.* (2018) ‘Identification and Functional Analysis of Two New Mutant BnFAD2 Alleles That Confer Elevated Oleic Acid Content in Rapeseed’, *Frontiers in Genetics*. doi:10.3389/fgene.2018.00399.

Lorenc, M. (2015) ‘The development and application of bioinformatics methods and software tools for computational single nucleotide polymorphism discovery’. Available at: https://core.ac.uk/download/pdf/43382761.pdf.

Los, D.A. and Murata, N. (1998) ‘Structure and expression of fatty acid desaturases’, *Biochimica et biophysica acta*, 1394(1), pp. 3–15.

Lubin, I.M. *et al.* (2017) ‘Principles and Recommendations for Standardizing the Use of the Next-Generation Sequencing Variant File in Clinical Settings’, *The Journal of molecular diagnostics: JMD*, 19(3), pp. 417–426.

Lu, K. *et al.* (2019) ‘Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement’, *Nature communications*, 10(1), p. 1154.

Lu, S. *et al.* (2019) ‘Heterogeneous Distribution of Erucic Acid in Brassica napus Seeds’, *Frontiers in plant science*, 10, p. 1744.

Mackill, D.J., Nguyen, H.T. and Zhang, J. (1999) ‘Use of molecular markers in plant improvement programs for rainfed lowland rice’, *Field Crops Research*, pp. 177–185. doi:10.1016/s0378-4290(99)00058-1.

Mardis, E.R. (2008) ‘The impact of next-generation sequencing technology on genetics’, *Trends in Genetics*, pp. 133–141. doi:10.1016/j.tig.2007.12.007.

Masood, M.A. and Khan, M.N.A. (2015) ‘Clustering techniques in bioinformatics’, *IJ Modern Education and Computer Science*, 1, pp. 38–46.

McKenna, A. *et al.* (2010) ‘The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data’, *Genome research*, 20(9), pp. 1297–1303.

McLaren, W. *et al.* (2016) ‘The Ensembl Variant Effect Predictor’, *Genome biology*, 17(1), p. 122.

Metzker, M.L. (2010) ‘Sequencing technologies - the next generation’, *Nature reviews. Genetics*, 11(1), pp. 31–46.

Micha, R. and Mozaffarian, D. (2009) ‘Trans fatty acids: effects on metabolic syndrome, heart disease and diabetes’, *Nature reviews. Endocrinology*, 5(6), pp. 335–344.

Mietkiewska, E. *et al.* (2007) ‘Cloning and functional characterization of the fatty acid elongase 1 (FAE1) gene from high erucic Crambe abyssinica cv. Prophet’, *Plant biotechnology journal*, 5(5), pp. 636–645.

Millar, A.A. and Kunst, L. (1997) ‘Very-long-chain fatty acid biosynthesis is controlled through the expression and specificity of the condensing enzyme’, *The*

Plant journal: for cell and molecular biology, 12(1), pp. 121–131.

Mills, R.E. (2006) ‘An initial map of insertion and deletion (INDEL) variation in the human genome’, *Genome Research*, pp. 1182–1190. doi:10.1101/gr.4565806.

Minde, D.P. et al. (2011) ‘Messing up disorder: how do missense mutations in the tumor suppressor protein APC lead to cancer?’, Molecular Cancer. doi:10.1186/1476-4598-10-101.

Mir, R.R. *et al.* (2013) ‘Evolving Molecular Marker Technologies in Plants: From RFLPs to GBS’, *Diagnostics in Plant Breeding*, pp. 229–247. doi:10.1007/978-94-007-5687-8_11.

Moore, B. *et al.* (2012) ‘Using GVF for Clinical Annotation of Personal Genomes’, *AIMM* [Preprint]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.7246&rep=rep1&type=pdf.

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.

Morita, R. *et al.* (2009) ‘Molecular characterization of mutations induced by gamma irradiation in rice’, *Genes & genetic systems*, 84(5), pp. 361–370.

Mullaney, J.M. *et al.* (2010) ‘Small insertions and deletions (INDELs) in human genomes’, *Human molecular genetics*, 19(R2), pp. R131–6.

Muller, H.J. (1927) ‘ARTIFICIAL TRANSMUTATION OF THE GENE’, *Science*, 66(1699), pp. 84–87.

Nagaharu, U. (1935) ‘Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization’, *Journal of Japanese Botany*, 7(7), pp. 389–452.

Nations, U. (2015) ‘World population prospects: The 2015 revision’, *Population newsletter / issued by the Population Division of the Department of Economic and Social Affairs, United Nations*, 33(2), pp. 1–66.

Newton, A.C., Johnson, S.N. and Gregory, P.J. (2011) ‘Implications of climate change for diseases, crop yields and food security’, *Euphytica*, pp. 3–18. doi:10.1007/s10681-011-0359-4.

Nielsen, R. *et al.* (2011) ‘Genotype and SNP calling from next-generation sequencing data’, *Nature reviews. Genetics*, 12(6), pp. 443–451.

Ohlrogge, J. and Browse, J. (1995) ‘Lipid biosynthesis’, *The Plant cell*, 7(7), pp. 957–970.

Olsson, G. (2010) ‘SPECIES CROSSES WITHIN THE GENUS BRASSICA’, *Hereditas*, pp. 171–223. doi:10.1111/j.1601-5223.1960.tb03082.x.

Oraguzie, N.C. *et al.* (2007) *Association Mapping in Plants*. Springer Science & Business Media.

Ortiz, R. (2010) ‘Molecular Plant Breeding’, *Crop Science; Madison volume*. search.proquest.com, pp. 2196–2197. Available at: https://search.proquest.com/openview/27f7b6d4a6653cd2077871427be9aab7/1?pq-origsite=gscholar&cbl=30013.

Paape, T. *et al.* (2018) ‘Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica’, *Nature communications*, 9(1), p. 3909.

Parkinson, J. and Blaxter, M. (2009) ‘Expressed Sequence Tags: An Overview’, in Parkinson, J. (ed.) *Expressed Sequence Tags (ESTs): Generation and Analysis*. Totowa, NJ: Humana Press, pp. 1–12.

Park, S. *et al.* (2010) ‘Genome-wide discovery of DNA polymorphism in Brassica rapa’, *Molecular genetics and genomics: MGG*, 283(2), pp. 135–145.

Pearson, W.R. and Lipman, D.J. (1988) ‘Improved tools for biological sequence comparison’, Proceedings of the National Academy of Sciences, pp. 2444–2448. doi:10.1073/pnas.85.8.2444.

Peng, S. *et al.* (2004) ‘Rice yields decline with higher night temperature from global warming’, *Proceedings of the National Academy of Sciences of the United States of America*, 101(27), pp. 9971–9975.

Pérez-de-Castro, A.M. *et al.* (2012) ‘Application of genomic tools in plant breeding’, *Current genomics*, 13(3), pp. 179–195.

Polyanovsky, V.O., Roytberg, M.A. and Tumanyan, V.G. (2011) ‘Comparative analysis of the quality of a global algorithm and a local algorithm for alignment oftwo sequences’, *Algorithms for molecular biology: AMB*, 6(1), p. 25.

Putney, S.D., Herlihy, W.C. and Schimmel, P. (1983) ‘A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing’, *Nature*, 302(5910), pp. 718–721.

Quinlan, A.R. and Hall, I.M. (2010) ‘BEDTools: a flexible suite of utilities for comparing genomic features’, *Bioinformatics* , 26(6), pp. 841–842.

Rafalski, A. (2002) ‘Applications of single nucleotide polymorphisms in crop genetics’, *Current Opinion in Plant Biology*, pp. 94–100. doi:10.1016/s1369-5266(02)00240-6.

Ramakrishna, G. *et al.* (2018) ‘Genome-wide identification and characterization of InDels and SNPs in Glycine max and Glycine soja for contrasting seed permeability traits’, *BMC plant biology*, 18(1), p. 141.

Ranallo-Benavidez, T.R. *et al.* (2020) ‘GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes’, *Nature Communications*. doi:10.1038/s41467-020-14998-3.

Ray, D.K. *et al.* (2013) ‘Yield Trends Are Insufficient to Double Global Crop

Production by 2050', *PloS one*, 8(6), p. e66428.

Reams, A.B. and Roth, J.R. (2015) 'Mechanisms of Gene Duplication and Amplification', *Cold Spring Harbor Perspectives in Biology*, p. a016592. doi:10.1101/cshperspect.a016592.

'Recovering evolutionary trees under a more realistic model of sequence evolution' (1994) *Molecular Biology and Evolution* [Preprint]. doi:10.1093/oxfordjournals.molbev.a040136.

Reese, M.G. *et al.* (2010) 'A standard variation file format for human genome sequences', *Genome biology*, 11(8), p. R88.

Rodgers, K. and McVey, M. (2016) 'Error-Prone Repair of DNA Double-Strand Breaks', *Journal of Cellular Physiology*, pp. 15–24. doi:10.1002/jcp.25053.

Rokas, A. and Holland, P.W.H. (2000) 'Rare genomic changes as a tool for phylogenetics', *Trends in Ecology & Evolution*, pp. 454–459. doi:10.1016/s0169-5347(00)01967-4.

Rousseau-Gueutin, M. *et al.* (2020) 'Long-read assembly of the Brassica napus reference genome Darmor-bzh', *GigaScience*, 9(12). doi:10.1093/gigascience/giaa137.

Rudd, S. (2003) 'Expressed sequence tags: alternative or complement to whole genome sequences?', *Trends in Plant Science*, pp. 321–329. doi:10.1016/s1360-1385(03)00131-6.

Salgotra, R.K. and Neal Stewart, C. (2020) ‘Functional Markers for Precision Plant Breeding’, *International Journal of Molecular Sciences*, p. 4792. doi:10.3390/ijms21134792.

Sanger, F. and Coulson, A.R. (1975) ‘A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase’, *Journal of molecular biology*, 94(3), pp. 441–448.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) ‘Assembly of large genomes using second-generation sequencing’, *Genome Research*, pp. 1165–1173. doi:10.1101/gr.101360.109.

Schliep, K. (no date) *phangorn: Phylogenetic analysis in R*. Github. doi:10.1111/2041-210X.12760.

Sehn, J.K. (2015) ‘Insertions and Deletions (Indels)’, *Clinical Genomics*, pp. 129–150. doi:10.1016/b978-0-12-404748-8.00009-5.

Senthilvel, S. *et al.* (2019) ‘Development and validation of an SNP genotyping array and construction of a high-density linkage map in castor’, *Scientific Reports*. doi:10.1038/s41598-019-39967-9.

Shanklin, J. and Cahoon, E.B. (1998) ‘DESATURATION AND RELATED MODIFICATIONS OF FATTY ACIDS1’, *Annual review of plant physiology and plant molecular biology*, 49, pp. 611–641.

Sharma, J., Keeling, K.M. and Rowe, S.M. (2020) ‘Pharmacological approaches for targeting cystic fibrosis nonsense mutations’, European journal of medicinal chemistry, 200, p. 112436.

Shigemizu, D. *et al.* (2018) ‘IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis’, *Scientific reports*, 8(1), p. 5608.

Singh, N.K. *et al.* (2017) ‘Comparative Genomics and Synteny Analysis of KCS17-KCS18 Cluster Across Different Genomes and Sub-genomes of Brassicaceae for Analysis of Its Evolutionary History’, *Plant molecular biology reporter / ISPMB*, 35(2), pp. 237–251.

Slatko, B.E., Gardner, A.F. and Ausubel, F.M. (2018) ‘Overview of Next-Generation Sequencing Technologies’, *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 122(1), p. e59.

Smith, P. (2013) ‘Delivering food security without increasing pressure on land’, *Global Food Security*, pp. 18–23. doi:10.1016/j.gfs.2012.11.008.

Somerville, C. (2000) ‘Browse J, Jaworski JG, Ohlrogge JB. Lipids’, *Biochemistry and molecular biology of plants*, 1.

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.

Stadler, L.J. (1928) ‘Genetic Effects of X-Rays in Maize’, *Proceedings of the National Academy of Sciences of the United States of America*, 14(1), pp. 69–75.

Sun, R. *et al.* (2016) ‘A clustering approach to identify rare variants associatedwith hypertension’, *BMC proceedings*, 10(Suppl 7), pp. 153–157.

Suwarno, W.B. *et al.* (2015) ‘Genome-wide association analysis reveals new targets for carotenoid biofortification in maize’, *Theoretical and Applied Genetics*, pp. 851–864. doi:10.1007/s00122-015-2475-3.

Syvänen, A.-C. (2001) ‘Accessing genetic variation: genotyping single nucleotide polymorphisms’, *Nature Reviews Genetics*, pp. 930–942. doi:10.1038/35103535.

Team, F.F.T. (2015) ‘The Variant Call Format (VCF) Version 4.2 Specification’.

Tester, M. and Langridge, P. (2010) ‘Breeding technologies to increase crop production in a changing world’, *Science*, 327(5967), pp. 818–822.

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 of the United States of America*, 102(39), pp. 13950–13955.

Till, B.J. *et al.* (2004) ‘Discovery of induced point mutations in maize genes by TILLING’, *BMC plant biology*, 4, p. 12.

Trick, M. *et al.* (2009) ‘Single nucleotide polymorphism (SNP) discovery in the polyploidBrassica napususing Solexa transcriptome sequencing’, *Plant Biotechnology Journal*, pp. 334–346. doi:10.1111/j.1467-7652.2008.00396.x.

Valenzuela, D. *et al.* (2018) ‘Towards pan-genome read alignment to improve variation calling’, *BMC genomics*, 19(Suppl 2), p. 87.

VanWallendael, A. and Alvarez, M. (2022) ‘Alignment-free methods for polyploid genomes: Quick and reliable genetic distance estimation’, *Molecular ecology resources*, 22(2), pp. 612–622.

Varshney, R.K., Graner, A. and Sorrells, M.E. (2005) ‘Genic microsatellite markers in plants: features and applications’, *Trends in Biotechnology*, pp. 48–55. doi:10.1016/j.tibtech.2004.11.005.

Vinga, S. and Almeida, J. (2003) ‘Alignment-free sequence comparison-a review’, *Bioinformatics* , 19(4), pp. 513–523.

Voichek, Y. and Weigel, D. (2020) ‘Identifying genetic variants underlying phenotypic variation in plants without complete genomes’, *Nature genetics*, 52(5), pp. 534–540.

Wang, B. *et al.* (2018) ‘Dissection of the genetic architecture of three seed-quality traits and consequences for breeding in Brassica napus’, *Plant biotechnology journal*, 16(7), pp. 1336–1348.

Wang, X. *et al.* (2011) ‘The genome of the mesopolyploid crop species Brassica rapa', *Nature Genetics*, pp. 1035–1039. doi:10.1038/ng.919.

Wei, D. *et al.* (2017) 'A genome-wide survey with different rapeseed ecotypes uncovers footprints of domestication and breeding', *Journal of experimental botany*, 68(17), pp. 4791–4801.

Wenger, A.M. *et al.* (2019) 'Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome', *Nature Biotechnology*, pp. 1155–1162. doi:10.1038/s41587-019-0217-9.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wittkop, B., Snowdon, R.J. and Friedt, W. (2009) 'Status and perspectives of breeding for enhanced yield and quality of oilseed crops for Europe', *Euphytica/ Netherlands journal of plant breeding*, 170(1-2), p. 131.

Wu, H., Irizarry, R.A. and Bravo, H.C. (2010) 'Intensity normalization improves color calling in SOLiD sequencing', *Nature methods*, 7(5), pp. 336–337.

Wu, L. *et al.* (2015) 'Molecular evidence for blocking erucic acid synthesis in rapeseed (Brassica napus L.) by a two-base-pair deletion in FAE1 (fatty acid elongase 1)', *Journal of integrative agriculture*, 14(7), pp. 1251–1260.

Xia, W. *et al.* (2019) 'Development of High-Density SNP Markers and Their Application in Evaluating Genetic Diversity and Population Structure in', *Frontiers in plant science*, 10, p. 130.

Xu, F. *et al.* (2012) 'A fast and accurate SNP detection algorithm for next-generation sequencing data', *Nature communications*, 3, p. 1258.

Xu, Y. (2010) *Molecular Plant Breeding*. CABI.

Yadav, C.B. *et al.* (2015) 'Genome-wide SNP identification and characterization in two soybean cultivars with contrasting Mungbean Yellow Mosaic India Virus disease resistance traits', *PloS one*, 10(4), p. e0123897.

Yamaguchi, H. *et al.* (2009) 'Mutagenic effects of ion beam irradiation on rice', *Breeding Science*, pp. 169–177. doi:10.1270/jsbbs.59.169.

Yan, G. *et al.* (2015) 'Characterization of FAE1 in the zero erucic acid germplasm of Brassica rapa L', *Breeding science*, 65(3), pp. 257–264.

Yang, Q. *et al.* (2012) 'Identification of FAD2 and FAD3 genes in Brassica napus genome and development of allele-specific markers for high oleic and low linolenic acid contents', *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 125(4), pp. 715–729.

Yao, Z. *et al.* (2020) 'Evaluation of variant calling tools for large plant genome re-sequencing', *BMC Bioinformatics*. doi:10.1186/s12859-020-03704-1.

Yatagai, F. (2004) 'Mutations induced by heavy charged particles', *Uchu Seibutsu Kagaku*, 18(4), pp. 224–234.

Yoshihara, R. *et al.* (2010) ‘Mutational effects of different LET radiations in rpsL transgenic Arabidopsis’, *International journal of radiation biology*, 86(2), pp. 125–131.

Yu, F. *et al.* (2021) ‘Identification of Two Major QTLs in Brassica napus Lines With Introgressed Clubroot Resistance From Turnip Cultivar ECD01’, *Frontiers in plant science*, 12, p. 785989.

Yu, G. *et al.* (2017) ‘ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data’, *Methods in Ecology and Evolution*, pp. 28–36. doi:10.1111/2041-210x.12628.

Zhang, J. *et al.* (2012) ‘Arabidopsis fatty acid desaturase FAD2 is required for salt tolerance during seed germination and early seedling growth’, *PloS one*, 7(1), p. e30355.

Zhao, Q. *et al.* (2019) ‘A novel quantitative trait locus on chromosome A9 controlling oleic acid content in Brassica napus’, *Plant biotechnology journal*, 17(12), pp. 2313–2324.

Zou, J. *et al.* (2019) ‘Genome-wide selection footprints and deleterious variations in young Asian allotetraploid rapeseed’, *Plant biotechnology journal*, 17(10), pp. 1998–2010.

Zou, Q. *et al.* (2020) ‘Sequence clustering in bioinformatics: an empirical study’, *Briefings in bioinformatics*, 21(1), pp. 1–10.

1000 Genomes Project Consortium *et al.* (2010) ‘A map of human genome variation from population-scale sequencing’, *Nature*, 467(7319), pp. 1061–1073.