

Identifying and Understanding the Importance and Role of Bidirectional Promoters in Wheat

Katie Hawkins



UNIVERSITY OF LEEDS

Submitted in accordance with the requirements for the degree of
Master of Science by Research in Biological Sciences
The University of Leeds
School of Biological Sciences
September 2023

Intellectual Property and Publication Statement

The candidate confirms that the work submitted is his/her/their own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

Special thanks to Dr. Laura Dixon for her continued support through the year, and for making this project possible.

Abbreviations

BiP	= bidirectional promoter
BP	= base pair
CSD	= superoxide dismutase
DH	= drought heat study
GA	= gibberellic acid
GO	= gene ontology
IEA	= inferred from electronic annotation
Inr	= the initiator element
MS	= microspore study
TILLING	= Targeting Induced Local Lesions In Genomes
TPM	= transcript per million
TSS	= transcription start site

Abstract

In the context of a changing climate, understanding gene regulation in the globally important *Triticum aestivum* (bread wheat) is becoming increasingly important to inform breeding programs and further techniques and resources in genetic manipulation. Bidirectional promoters are one such gene regulatory element. Bidirectional promoters are promoters that initiate the transcription of two diverging genes simultaneously. In bioinformatic analysis genes with transcription start sites within 1000 base pairs of one another and are located on opposite strands are considered to be bidirectionally arranged and under the control of the same promoter. I conducted a whole genome analysis in *T. aestivum* identifying 1,050 bidirectional promoters representing 1.95% of the protein coding genes in wheat. I also performed this analysis for *Arabidopsis thaliana*, *Oryza sativa* Japonica (rice), *Avina sativa* (oat), *Zea mays* (maize), *Brassica Rapa* (field mustard), *Populus trichocarpa* (black cottonwood), *Solanum lycopersicum* (cherry tomato) in which bidirectional genes represented 16.7, 5.6, 0.8, 2.5, 7.9, 2.6, and 3.2% of all protein coding genes, respectively. The bidirectional arrangement was conserved in wheat, rice, and *Arabidopsis* in twenty-two pairs. Like bidirectional genes in previous studies, wheat bidirectional genes were found to be located predominantly in CpG islands, highly coexpressed, and frequently involved in functions relating to cellular maintenance. Bidirectional genes in wheat can also respond to temperature, with one pair confirmed through qPCR to be expressed more highly at 25°C than 15°C. Bidirectional promoters are enriched for particular motifs, four of which are consensus sequences for four transcription factor families: MADSbox; MIKC, AP2; ERF, TCP, and NAC; NAM. The former is known to regulate genes in response to temperature, while the latter two are known to regulate genes in response to stress in plants. Extreme temperature is a source of stress in plants. Therefore, bidirectional genes, in addition to other regulatory elements such as non-coding and alternative splicing, may be important in responding to stress in plants. Furthermore, a new definition for bidirectional genes in plants is suggested, with a cut off of 560 rather than 1000 base pairs. In conclusion, bidirectional genes in wheat have the broad characteristics defined in animals for bidirectional genes, but may play a more important role in response to stress and temperature response than animal species.

Contents

Chapter 1 - Introduction.....	7
1.0 Understanding gene regulation in bread wheat in the context of climate change....	7
1.1 Our current understanding of bidirectional promoters.....	12
1.1.0 Promoters.....	12
1.1.0 Bidirectional Promoters.....	13
1.2 An agronomically relevant polygenic trait: Coleoptile length.....	17
1.2.1 Genetic Influences: Rht.....	18
Chapter 2 - Detection of Bidirectional Promoters in Wheat.....	20
2.0 Bioinformatic detection of bidirectional promoters in wheat, rice, and thale cress... 20	
2.0.0 Materials and Methods.....	20
2.0.0.0 Identifying genes arranged in a bidirectional orientation.....	20
2.0.0.1 Genome density calculation.....	20
2.0.1 Results.....	20
2.0.2 Discussion.....	25
2.1 Orthology of bidirectional genes and the bidirectional gene arrangement.....	27
2.1.0 Methods.....	28
2.1.0.0 Identifying orthology of bidirectional promoters with <i>Arabidopsis thaliana</i> and <i>Oryza sativa Japonica</i>	28
2.1.0.1 Conserved bidirectional arrangement.....	28
2.1.1. Results.....	29
2.1.2 Discussion.....	31
2.2 Co-expression analysis of bidirectional genes from RNAseq data.....	33
2.2.0 Methods.....	33
2.2.1 Results.....	37
2.2.2 Discussion.....	40
2.3 Temperature dependent response of bidirectional gene pairs.....	42
2.3.0 Methods.....	42
2.3.1 Results.....	45
2.3.2 Discussion.....	53
Chapter 3 - Function and Mechanism of Bidirectionally Arranged Genes.....	54
3.0 GO term analysis, enriched promoter motifs, CpG islands, and pair alignment....	54
3.0.0 Methods.....	54
3.0.0.0 GO Term Analysis.....	55
3.0.0.2 Isolating the Intermediate Sequence Between Proximal Genes.....	57
3.0.0.3 Enriched Motifs.....	59
3.0.0.4 CG Content Calculation.....	59
3.0.0.5 CpG Island Calculation.....	59
3.0.1 Results.....	59
3.0.1.0 GO Term Enrichment and Functional Analysis.....	60
3.0.1.1 Alignment of Bidirectional and Proximal Pairs.....	62

3.0.1.2 Promoter Motif Enrichment.....	64
3.0.1.3 CpG Islands.....	66
3.0.2 Discussion.....	67
3.1 Expression of bidirectionally arranged genes in leaf and coleoptile samples.....	72
3.3.0 Methods.....	72
3.3.0.0 Growth Conditions and Sampling.....	72
3.3.0.1 RNA Extraction and cDNA Synthesis.....	72
3.3.0.2 qPCR.....	72
3.3.1 Results.....	74
Chapter 4 - Discussion.....	80
4.1 Whole Genome Identification of Bidirectionally Arranged Genes in Wheat.....	80
4.2 Calculation of Key Characteristics.....	81
4.3 RNAseq data and GO terms can be used to identify genes that respond to temperature.....	81
Supplementary.....	82
Glossary.....	82
References.....	83

Tables and Figures

In order of appearance

Chapter 1.0	
Figure 1. The genome arrangement of <i>Triticum aestivum</i> and The origin of the subgenomes	8
Chapter 1.1	
Figure 2. A schematic representation of two bidirectionally arranged genes	14
Table 1. Categories of Species Studied by Yang and Yu (2009)	15
Chapter 2.0	
Table 2. Number of bidirectionally arranged protein coding genes expressed as both an absolute number and a percentage of total protein coding genes	21
Figure 3. The genome density plotted against the percentage of bidirectional genes present in eight plant species	22
Figure 4. The gene density plotted against the percentage bidirectional genes for each chromosome	23
Figure 5. Representation of the location of bidirectional genes on the genome as a percentage of total genes	24
Figure 6. The location of the transcription start sites of bidirectionally arranged genes	25
Chapter 2.1	
Figure 7. Venn diagram where each circle represents the number of wheat bidirectional genes that have an orthologue (not necessarily bidirectionally arranged) in rice/ <i>Arabidopsis</i>	29

Table 3. Percentage of all, bidirectional (BiP), and proximal wheat protein coding genes that have orthologues in rice and Arabidopsis	30
Table 4. Conserved bidirectional arrangement between wheat, rice, and Arabidopsis	30
Figure 8. Adapted from Bolot et al., 2009. Synteny between wheat and rice	32
Chapter 2.2	
Table 5. References for the RNAseq data used in order to calculate co-expression coefficients	34-36
Figure 9. The proportional distribution of co-expression coefficients of bidirectional and proximal pairs	38
Figure 10. The proportional distribution of absolute co-expression coefficients of bidirectional and proximal pairs	39
Chapter 2.3	
Table 6. Summary of Supplementary Table 7D. Integration of temperature information for bidirectional genes from various sources	46
Table 7. A subset of Supplementary Table 7D, including only rows with a count > 5, corrected p-value (BH) < 0.05, and a co-expression coefficient ≥ 0.65	47
Figure 12. Expression in <i>Triticum aestivum</i> cultivar Azhurnaya in the first leaf and the coleoptile, and in the cultivar Chinese Spring in the leaf at the seedling stage for ten bidirectional genes	48
Table 8. The homoeologous arrangement of candidate genes	47
Figure 13. Expression of ten bidirectional genes measured in three different studies under different temperature and drought stress conditions	49
Table 9. BLASTn results of candidate bidirectional pair sequences	52
Chapter 3.0	
Figure 14. Demonstrates the two possible arrangements of proximal gene pairs	58
Figure 15. GO term enrichment analysis of the biological processes domain for bidirectional wheat genes	60
Figure 16. Percentage of bidirectional and proximal pairs that have related functions, unrelated functions, and not enough data	62
Table 10. Bidirectional gene pairs that align with one another using BLASTn	63
Figure 17. Number of common GO terms between the pairs plotted against the co-expression coefficient	64
Table 11. A list of enriched sequences in bidirectional promoters as detected by memsuite	65
Figure 18. Violin plot comparing the CG content of bidirectional and proximal promoters	66
Figure 19. The observed to expected ratio of CpG islands in bidirectional (BiP) and proximal promoters	67
Chapter 3.1	
Table 12. qPCR primers used to calculate expression of candidate genes	73-74
Figure 20. The coleoptile length of Paragon wheat on days 1 to 9 of growth in temperatures from 10°C to 30°C at 5 degree intervals	75
Figure 21. The expression in arbitrary units as determined by qPCR of candidate bidirectional pairs and their homoeologues	76
Figure 22. The expression in arbitrary units of bidirectional pair 407	77

Figure 23. Expression in arbitrary units of bidirectional pair 356, and pair 458, homoeologues of pair 307 78

Chapter 1 - Introduction

1.0 Understanding gene regulation in bread wheat in the context of climate change

Globally, there are two dominant species of domestic wheat: bread wheat *Triticum aestivum* and durum wheat *Triticum durum*. 90-95% of all wheat grown is *T. aestivum*, which is used to produce bread, biscuits, and other baked products (Giraldo *et al.*, 2019). Grains unsuitable for human consumption and straw are also valuable as feedstock. Around 5% of wheat grown is *T. durum* which is predominantly used to produce pasta (Almarri *et al.*, 2023). The remaining percentage is made up of other wheats including spelt (*T. spelta*), domestic einkorn (*T. monococcum*), and domestic emmer (*T. turgidum*), which besides spelt have negligible market value in the UK (Costanzo *et al.* 2019). These are marketed as health foods and grown as animal feed (Giraldo *et al.*, 2019; Ruibal-Mendieta *et al.*, 2004). In some places landrace varieties of *T. aestivum* are grown for local consumption (Baboev *et al.*, 2021). Wild and landrace varieties are also sources of genetic diversity in modern breeding programs (Baboev *et al.*, 2021).

Triticum aestivum is the most widely cultivated crop in the world with over 700 million tonnes being produced each year from 173 different countries (Shahbandeh 2023 *Statistica*). Around 20% of all calories and protein consumed globally are from bread wheat, and it's also rich in many B-vitamins (thiamine, riboflavin, pyridoxine, niacin, pantothenic acid, biotin, folates) and fibre, thus constituting a major part of people's diets (Batifoulier *et al.* 2006). For these reasons it is integral to food security and is economically valuable; globally, an average of 138 million tonnes of bread wheat was exported per year between 2005 and 2014, and nearly £50 billion worth of wheat was traded in 2021 (van der Velde *et al.*, 2018; The Observatory of Economic Complexity, 2023).

Modern hexaploid wheat *Triticum aestivum*, has the genome organisation AABBDD, which arose around 8,000 years ago as a result of a hybridisation event between domesticated emmer and *Aegilops tauschii* - a grass species from the same family as wheat, and the progenitor of the D subgenome (IWGSC 2013). Other subgenomes, including C, exist in wild relatives of wheat such as *Aegilops caudata* which is a diploid with the genome organisation CC (Grewal *et al.* 2020). Emmer is a tetraploid with the genome organisation AABB which itself arose from a hybridisation event between the diploid *T. urartu* (AA) and another, as of yet unknown, diploid species (BB) that occurred half a million years ago (Charmet 2011; Hong-Qing *et al.*, 2018). It is unknown whether the B subgenome progenitor is extant, as many of the *Sitopsis* grasses are morphologically similar to one another, so it is possible the progenitor is as of yet undiscovered (Levy and Feldman 2022). As a result of these hybridisation events *T. aestivum* has three subgenomes donated from different species

present in two copies each, making it an allohexaploid. Forty-seven percent of genes in *T. aestivum* are arranged as a triad, meaning there is one copy of a gene on each subgenome, the rest are present in other arrangements (IWGSC 2018). When a gene has a homologous copy on another subgenome these are referred to as homoeologues, and would have been orthologues in the donor species’.

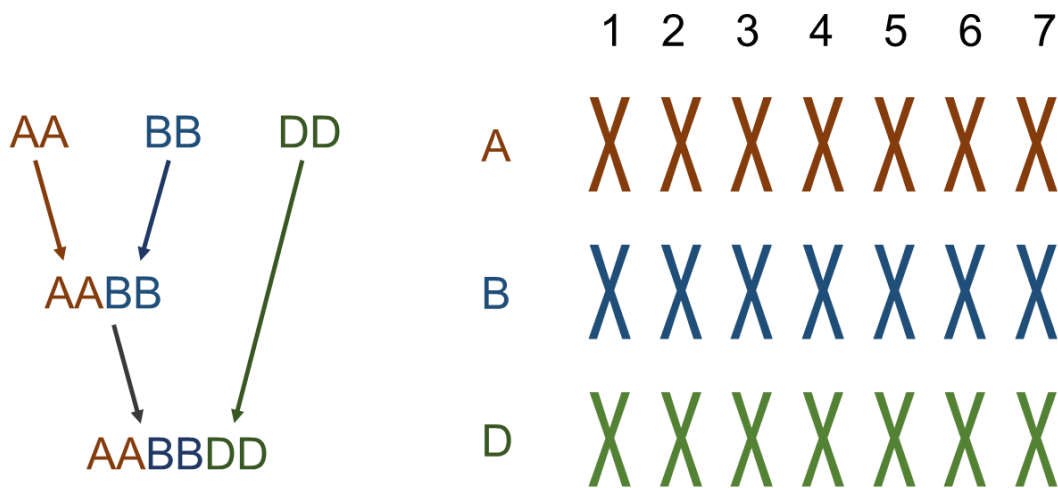


Figure 1. Right: The genome arrangement of *Triticum aestivum*. The 21 chromosomes are represented by crosses, seven in each subgenome A, B, D. Left: The origin of the subgenomes. The letter combinations represent the genomes present in *T. aestivum* (AABBDD) and its ancestors.

A polyploid nature confers both advantages and disadvantages on an organism, and will exist in the wild in cases where the advantages are sufficient to overcome the initial disadvantages as the organism adapts to the new genome organisation (Comai 2005). The onset of polyploidy necessarily causes massive disruption in the regulation and organisation of the genome (Feliner *et al.*, 2022). The genome is physically larger and thus takes up more space in the cells, leading to an increased nuclear and cell size. This is disadvantageous at the cell level, as the surface area to volume ratio of the components is affected (Comai 2005). In theory the doubling of the chromatin volume will lead to only a 1.6 times increase in the surface area of the nuclear envelope which can, for example, impact the interactions between lamins (proteins present on the inside of the nuclear envelope) and the chromatin (Comai 2005). This is hypothesised to have regulatory repercussions through the effect of gene dosage (Comai 2005). The increased chromatin volume also decreases its accessibility which in *T. aestivum* explains the reduced expression of some genes on the D subgenome compared to its *A. tauschii* donor (Lu *et al.*, 2020). The production of aneuploid cells in meiosis and mitosis is also more likely in polyploids (Comai 2005).

Some advantages of polyploidy include heterosis and gene redundancy (Comai 2005). Heterosis is a long-observed, widely exploited, yet poorly understood phenomenon in which hybrids have increased vigour compared to their parents; this is also observed in newly formed polyploid species (Birchler *et al.*, 2010). Gene redundancy refers to the fact that two

identical genes in the same genome cannot be stably retained as there is no fitness benefit to their retention, instead one will either be lost or undergo sub or neofunctionalisation (Brown 2002). Thus the polyploidisation events that occurred in wheat opened up opportunities for the evolution of new genes.

We can speculate that certain gene families have homoeologues that are more likely to be functional as they have greater retention than the average 35.8%, such as MIKC-type MADS-box genes, GRAS transcription factors, and argonaute genes, which have 62.7%, 79.8%, and 60.9% retention respectively (Schilling 2020; Liu and Wang 2021; Liu *et al.*, 2021). For example, the MADS-box genes *TaAGL17-B* and *TaAGL17-D* are upregulated in response to heat, whereas the A homoeologue is not, indicating that the B and D copies have gained a function that confers heat tolerance (Schilling 2020). Another example is *TaGRAS166*, which was induced in the leaves by infection with stripe rust and powdery mildew, as well as phosphorus starvation. Its homoeologues meanwhile, had stable expression, so *TaGRAS166* may have gained a function relating to tolerance to stress (Liu and Wang 2021). The argonaute gene *TaAGO4a* is not affected by heat stress while its homoeologues have reduced expression, suggesting a function related to heat tolerance for this gene (Liu *et al.*, 2021). The transcription factor family NAC have relatively high differing expression between homoeologues, suggesting neofunctionalisation in this family which is known to be involved in development and response to biotic and abiotic stress (Ma *et al.*, 2021).

Traits like tolerance to high temperature and drought are becoming increasingly important as the climate warms. For every 1°C increase *per se* in temperature wheat production decreases by 6% worldwide (Asseng *et al.*, 2015). For context, since 1880 the global temperature has increased by at least 1.1°C, with two-thirds of that warming occurring since 1975 (GISTEMP Team 2023; Lenssen *et al.*, 2019). The ideal growth temperature for wheat is between 16°C and 26°C depending on the growth stage, though it differs between cultivars and has also been reported to be between 12°C to 24°C (Khan *et al.*, 2021; Farooq *et al.*, 2011). At the temperature sensitive grain filling stage the optimum temperature is around 26°C, but a range of temperatures from 13°C to 30°C can be tolerated (Khan *et al.*, 2021). When the terminal spikelet is developing the optimal temperature is around 16°C, but again a range from 2.5°C to 20°C can be tolerated (Khan *et al.*, 2021).

Temperatures outside of the ideal range have adverse effects on grain number, filling, and quality (Akter and Islam, 2017). High temperatures reduce the growth period which reduces grain filling by an estimated 1.5 mg per day for every degree above 20°C (Farooq *et al.*, 2011). Above 30°C wheat experiences reduced seed set leading to reduced yield and can cause complete sterility in some cultivars (Farooq *et al.*, 2011). In the UK in 2022 summer temperatures reached above 40°C. The effect of heat stress on wheat will be even more pronounced in climates such as that of the Indo-Gangetic Plains which is currently favourable to wheat growth and accounts for 15% of production (Ortiz *et al.*, 2008). By 2050 half of this area could be reclassified as a heat stressed environment (Ortiz *et al.*, 2008). An additional consequence of climate change is that summer precipitation in the UK will decrease, leading to greater variability in wheat yield (Putelat *et al.*, 2021). Based on computer modelling wheat yield in the south-east of Great Britain will become “substantially unstable” by 2050 due to this region of the UK receiving relatively less rainfall (Putelat *et al.*,

2021). Between 1993 and 2013 four fifths of the lowest yielding years in Europe occurred because of heat or drought stress (van der Velde *et al.*, 2018).

Overall, climate change is resulting in more extreme and unpredictable weather conditions, such as drought, flooding, and increased temperature. Under these pressures the yield of wheat not only needs to be maintained but increased, to account for the increasing global population, which is presently at just under 8 billion, but is expected to reach 9 billion by 2050 (Charmet 2011). To feed this number of people, the yield of wheat needs to sustainably increase by 2% each year without decrease in nutrient content, which will require the combined efforts of genetic manipulation and changes to land management (Charmet 2011). Current yields of wheat in the UK are around 11 tonnes per hectare of an estimated 20 tonnes per hectare potential (Wheat Growth Guide, 2023). Yields of wheat are increasing, but only by 0.9% each year and global yield increases are stagnating (Charmet 2011; Ray *et al.*, 2012).

One of the main approaches to increasing wheat yield is through breeding programs, which aim to produce new commercial cultivars (Tessema *et al.*, 2020). Wheat breeding programs generate genetically diverse germplasm through chemically inducing random mutants or through the crossing of commercial cultivars with cultivars containing natural variation (Tessema *et al.*, 2020). Sources of this genetic variation can be landraces or wild relatives (Baboev *et al.*, 2021). The population containing these variations is then screened for desirable characteristics, which are then backcrossed into a desirable genetic background, thus producing a new commercial wheat cultivar (Tessema *et al.*, 2020).

This process can be complemented by the integration of genomic research. One way in which this is done is through QTL analyses, which are used to map particular phenotypes to genomic regions (Maccaferri *et al.*, 2022). Marker assisted selection can then be used in lieu of phenotyping as the trait can be accurately predicted using those markers (Maccaferri *et al.*, 2022). However, that genomic region may not be expressed, due to interaction with other genes or the environment (Maccaferri *et al.*, 2022). Thus the greater the understanding of the mechanisms that lead to a particular phenotype the more reliably and quickly it can be bred into commercial cultivars, a process that takes upwards of eight years (Tessema *et al.*, 2020).

Most efforts in this direction have focused on identifying genetic variation within genes (Rothamsted Research 2021; Hammond-Kosack *et al.*, 2021). However, there has been recent interest in the variation in regulatory regions between genes (Rothamsted Research 2021; Hammond-Kosack *et al.*, 2021). Even small changes in regulatory regions can produce profound changes in an organism by altering downstream expression of genes, so if beneficial mutations can be characterised and crossed into wheat, this will provide wheat breeding programs a plethora of as of yet untapped genetic variation (Rothamsted Research 2021; Hammond-Kosack *et al.*, 2021).

As discussed, gene expression can be regulated by transcription factors, but there are myriad other ways genes can be regulated including: through the accessibility of chromatin, alternative splicing, and non-coding RNA (ncRNA) (Tsompana and Buck 2014; Yang *et al.*, 2014; Zhang *et al.*, 2019). These can interact with plant stressors to confer advantages

under stressed environments, such as those that will become more common as the climate changes.

The first of these examples, accessibility of chromatin, is possible because the DNA in eukaryotic cells is compacted into nucleosomes, which consists of four pairs of histones, H2A, H2B, H3, and H4, with DNA wound around them (Tsompana and Buck 2014). Chromatin is composed of these nucleosomes (Tsompana and Buck 2014). In order for genes to be expressed the DNA needs to become less compacted so that transcription factors and RNA polymerase can bind, and thus genes are regulated by chromatin accessibility (Tsompana and Buck 2014). H2A.Z is a variant of the H2A histone that binds DNA more tightly but becomes dissociated from the DNA at higher temperatures, hence making downstream genes more accessible, and is thus important in regulating response to temperature (Kumar and Wigge 2010). While this was discovered in thale cress validation in budding yeast, *Saccharomyces cerevisiae*, suggests conservation across eukaryotes (Kumar and Wigge 2010). In *Brachypodium distachyon*, a commonly used model for crop species that is closely related to wheat, disruption of the H2A.Z histone variant was found to have reduced yield phenotypes comparable to growth at elevated temperatures, suggesting its importance in response to temperature in *B. distachyon* (Boden *et al.*, 2013). Thermosensitive genes important in grain filling in wheat, such as beta-amylase (AMY1), UDP-glucose pyrophosphorylase (UDP-GPP), and serpin 2A, are upregulated in both *B. distachyon* and *T. aestivum* when grown in conditions 5°C warmer than the control, demonstrating its suitability as a proxy for wheat in this experiment (Boden *et al.*, 2013). Therefore, H2A.Z is likely to be involved in integrating temperature information into gene regulation in wheat. Wheat histone variant TaH2A.7 when overexpressed in thale cress conferred drought tolerance, in part by promoting stomatal closure (Xu *et al.*, 2016). It also increased sensitivity to ABA, but didn't impact response to salt (Xu *et al.*, 2016). Therefore, genes can be regulated in response to heat stress in wheat via chromatin accessibility.

Alternative splicing is a process that occurs in eukaryotes where the exons are joined together in different combinations, producing different mRNA transcripts - known as isoforms - from the same gene (Yang *et al.*, 2014). RNAseq experiments under normal, heat stress, drought stress, and heat-drought stress conditions were carried out in wheat in order to identify alternative splicing that resulted from growing in stressed conditions (Liu *et al.* 2017). Isoforms that had greater than or equal to 30% variation between the normal and stressed conditions were considered to be induced by stress (Liu *et al.*, 2017). There were 251, 6618, and 7451 genes that showed alternative splicing under drought stress, heat stress, and heat-drought stress respectively, including heat shock transcription factors (Liu *et al.*, 2017).

TaHsfA2-7 (heat shock transcription factor A2-7) produces a truncated splice variant at higher temperatures through intron retention (Ma *et al.*, 2023). Expression of this protein in yeast and thale cress confers improved thermotolerance when exposed to temperatures of 50°C for 30 minutes and 45°C for 50 min respectively (Ma *et al.*, 2023). The alternative splicing for *TaHsfA2* is conserved between wheat, rice, and thale cress (Liu *et al.*, 2017). Alternative splicing of genes from other families can also confer heat tolerance, such as the transporter gene *TabZIP60*, as demonstrated by transgenic expression in thale cress (Geng *et al.*, 2018).

Non-coding RNAs are transcripts that are not translated into protein but have their own function (Zhang *et al.*, 2019). For example, tRNA is a ncRNA that folds into a clover shape and is a crucial component of the ribosome, where it links the amino acid and messenger RNA, allowing mRNA to be translated. ncRNAs can be classified as housekeeping, such as tRNA, or regulatory, such as microRNA (miRNA), small interfering RNA (siRNA), and long non-coding RNA (lncRNA) (Zhang *et al.*, 2019).

Wheat has approximately 40,970 lncRNAs, the majority of which are between 300-500 nucleotides in length (Jiang *et al.*, 2023). Of those, 273 lncRNAs were found to respond to high temperature in the seeds of wheat landrace WTB (Jiang *et al.*, 2023). How seeds respond to temperature is important because elevated temperature is the main factor that affects pre-harvest sprouting, which is when the grains begin to germinate before harvesting (Jiang *et al.*, 2023). Pre-harvest sprouting reduces the quality of the grain and is estimated to cost £0.8 million in global losses every year (Jiang *et al.*, 2023). lncRNAs also have a role to play in responding to low temperature through the crucial cold tolerance gene CSD1 (Superoxide dismutase 1) (Lu *et al.*, 2019). lncR9A regulates response to low temperature by competitively binding with Tae-miR398, which in turn downregulates CSD1 (Lu *et al.*, 2019).

miRNAs and siRNAs are known to be involved in high temperature response in crops (Singh *et al.*, 2021). In wheat, four siRNAs identified were tested for differential expression under heat, cold, dehydration, and NaCl stresses, and all demonstrated differential expression under at least two conditions (Yao *et al.*, 2010).

Therefore, gene regulatory elements regulate agronomically important traits in wheat, including those that may confer tolerance to abiotic stressors that are becoming more commonplace as the climate warms. A sufficient understanding of these elements will allow for the incorporation of advantageous variants of crucial regulatory regions into commercial cultivars, making them better suited for producing high yields in our new climate.

1.1 Our current understanding of bidirectional promoters

1.1.0 Promoters

A promoter is a region of DNA that is located upstream of a gene and to which transcription factors and RNA polymerase bind in order to initiate transcription of that gene. These transcription factors usually bind to conserved motifs that are found in different numbers, combinations and orientations (Inukai *et al.*, 2017). The specific combinations of motifs provide specificity for transcription factor binding and therefore the regulation of gene expression. Whether a gene is ultimately transcribed is controlled by the complex of proteins that is formed at the promoter, and its ability to recruit RNA polymerase which transcribes the DNA (Alberts *et al.*, 2002). This complex contains two elements: first, general transcription factors that bind core promoter sequences, such as the TATA box and Inr, which are common between many genes (Roeder 1996). A core promoter sequence is sufficient to induce a basal level of transcription *in vitro* (Roeder 1996). This is then modulated by the second element, which is gene-specific and regulates the basal expression whether by activation or repression of transcription (Roeder 1996).

TATA boxes are defined by the consensus sequence TATAWAW and bind with TATA box-binding proteins (TBF) (Rutherford and Van Duyne 2013). In addition, certain sequences on either side of the consensus sequence can improve the stability of TBF binding (Bareket-Samish *et al.*, 2000). This mechanism is well conserved in plants with 29% of thale cress promoters containing a TATA box, the number in wheat remains unknown (Molina and Grotewold 2005).

TATA boxes are found in CG depleted regions and thus can be juxtaposed with promoters found in CpG islands. CpG describes a C followed by a G in the sugar phosphate backbone in the 5' to 3' direction, and CpG islands are regions on DNA strands characterised by an unusually high number of these islands (Gardiner-Garden and Frommer 1987). The ratio of observed to expected is calculated using the following formula:

$$Obs/Exp CpG = \frac{Number\ of\ CpG}{Number\ of\ C \times Number\ of\ G} \times N$$

Where N is the total number of nucleotides in the sequence (Gardiner-Garden and Frommer 1987).

Generally in the literature a cutoff of 0.6 is used to define CpG islands (Gardiner-Garden and Frommer 1987). CpG islands are associated with a lack of methylation and the 5' ends of genes (Cross and Bird 1995). Methylation of DNA is a heritable epigenetic mark, has a role in regulating gene expression, and is essential for normal development (Jin *et al.*, 2011). Because CpG islands are often found at the 5' ends of genes, they are thought of as gene markers, although this was found in vertebrate genomes, and when applied to plants it only holds true for plants with small genomes like thale cress and not for larger plant genomes such as wheat (Ashikawa 2001). Genes with TATA box promoters tend to be tissue specific, whereas genes in CpG islands are more likely to be constitutively expressed (Schug *et al.*, 2005; Ashikawa 2001).

1.1.0 Bidirectional Promoters

A bidirectional promoter is defined as the region between two transcription start sites (TSS) of genes on opposite strands of the DNA within 1000 base pairs of one another (Trinklein *et al.*, 2004; *Fig. 2*). It is assumed that the region represents a single promoter that confers transcription in both directions, and therefore the sense/antisense distinction isn't applicable to bidirectionally arranged genes.

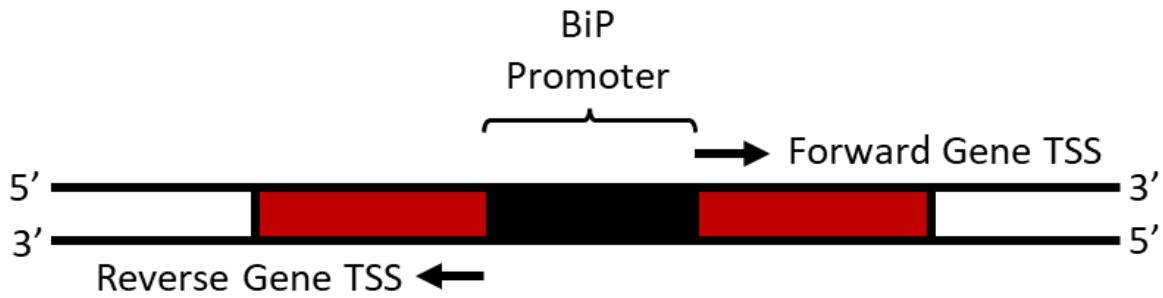


Figure 2. A schematic representation of two bidirectionally arranged genes. The red section represents two different genes, the black section represents the bidirectional promoter and the arrows represent the direction of transcription of the two genes. Note that the transcription must occur on opposite strands due to the 5' to 3' direction of transcription. The genes are referred to as “forwards” and “reverse” as one is being transcribed in the forwards direction, or what appears here as left to right, while the other is transcribed in the reverse direction, or what appears here as right to left.

This gene arrangement has long been noted incidentally, the earliest example being in mouse in 1985, but more recently the question of whether this gene arrangement is biologically important or occurs by chance in the genome has been asked (Crouse *et al.*, 1985). This has led to whole genome studies searching for bidirectionally arranged genes, as well as in some cases looking for enriched motifs and common features to begin to ascertain an underlying mechanism for bidirectional transcription (Dhadi *et al.*, 2009; Behura and Severson 2014). This has occurred in model organisms such as mouse, rat, fruit flies, yeast, and thale cress, but also in species of particular interest such as humans and crops including rice and maize (Dhadi *et al.*, 2009; Behura and Severson 2014; Liu *et al.*, 2014; Trinklein *et al.*, 2004; Li *et al.*, 2006).

Bidirectional promoters are present in varying quantities depending on the species, ranging from around 10% of all genes in human to around 50% of all genes in *Saccharomyces cerevisiae* (Trinklein *et al.*, 2004; Xu *et al.*, 2009). If genes were randomly distributed in a genome, some would be bidirectionally arranged through random chance, and this chance would be higher in more gene dense genomes. As expected, genome density is correlated with the fraction of bidirectional genes with a Spearman's correlation coefficient of 0.64 across eukaryotes, including representative species from vertebrates, insects, fungi, nematodes and plants (Yang and Yu 2009). The species included in the study are listed in the table below. The two plant species considered here, thale cress and rice, both had fewer bidirectional genes than predicted by genome density, suggesting that plants may have fewer bidirectionally arranged genes than other species (Yang and Yu 2009).

Table 1. Categories of Species Studied by Yang and Yu (2009).

Group	Species
Vertebrate	<i>Homo sapiens</i>
Vertebrate	<i>Pan troglodytes</i>
Vertebrate	<i>Mus musculus</i>
Vertebrate	<i>Rattus norvegicus</i>
Vertebrate	<i>Canis familiaris</i>
Vertebrate	<i>Gallus gallus</i>
Insect	<i>Apis mellifera</i>
Nematode	<i>Caenorhabditis elegans</i>
Fungi	<i>Saccharomyces cerevisiae</i>
Fungi	<i>Schizosaccharomyces pombe</i>
Fungi	<i>Eremothecium gossypii</i>
Fungi	<i>Kluyveromyces lactis</i>
Fungi	<i>Magnaporthe grisea</i>
Fungi	<i>Neurospora crassa</i>
Plant	<i>Arabidopsis thaliana</i>
Plant	<i>Oryza sativa</i>

Bidirectionally arranged genes can be found across all three domains of life and share some common features: they are found in CpG rich regions of the genome, are depleted for TATA boxes, and tend to encode a disproportionate number of housekeeping genes, particularly those relating to DNA repair and cell cycle regulation (Bagchi and Vishwanath 2016; Adachi *et al.*, 2002). An important aspect of the bidirectional promoter is considering how this shared promoter will impact the expression of the two genes. As they are thought to be under the control of the same promoter, it is assumed that they will be co-expressed at the mRNA level, and that this confers an advantage by allowing for the tightly controlled coordination of two genes. Bidirectionally arranged genes are indeed more likely to be co-expressed, but also are more likely to be anti-regulated, in other words, for two paired bidirectionally arranged genes it is more likely that as the expression of one gene increases, the other decreases, compared to two random genes that are located near to one another on the genome (Dash *et al.*, 2020). How this occurs is not known, but it has been speculated that the two genes may compete for RNA polymerase. Further complicating the situation is

the fact that although it is more likely, not every bidirectionally arranged gene pair is co- or anti-regulated, some have no correlation in expression (Dhadi *et al.*, 2009). This could be because the two genes are expressed in different tissues or at different developmental stages, thus despite being under the control of the same promoter are spatially or temporally separated. It could also be due to noise in the data, arising from the somewhat arbitrary cut-off of 1000 base pairs to define bidirectional promoters. Instead of a single promoter regulating both genes as predicted, two promoters could be present and acting independently, or regulated in another way.

To investigate the mechanism of bidirectional transcription promoter enrichment analyses have been carried out. Perhaps the most enlightening was a promoter enrichment analysis that found that the consensus sequence for the ets-related (erythroblast transformation specific) human transcription factor GA binding protein (GABP) was enriched (Collins *et al.*, 2007). Three cell types were assayed for binding of GABP to bidirectional promoters and 62-82% of bidirectional promoters were found to bind GABP depending on the cell type (Collins *et al.*, 2007). To validate these results, a GABP consensus sequence was introduced to six promoters, four of which had significantly increased transcription in the reverse direction, demonstrating that this consensus sequence is sufficient to cause bidirectional transcription from a single promoter (Collins *et al.*, 2007). This transcription factor however is not present in plants; the mechanism by which bidirectional transcription occurs must therefore be different between eukaryotes.

Promoter enrichment analyses have been carried out in thale cress, rice, and the tree species black cottonwood (Dhadi *et al.*, 2009). SORLIP2AT, consensus sequence GGGCC, and SITEIIATCYTC, consensus sequence TGGGCY, were found to be enriched between the three species for bidirectional promoters (Dhadi *et al.*, 2009). In addition there were several motifs that were enriched only for one or two of the three species, but without experimental verification it remains unknown if any of these play a role in bidirectional transcription in plants (Dhadi *et al.*, 2009).

The function of bidirectional genes has been determined in humans by calculating what fraction of characterised genes with different functions were bidirectional (Adachi *et al.*, 2002). While 10% of all genes in human are arranged bidirectionally, of 120 DNA repair genes 42% were bidirectionally arranged, of 14 DNA replication genes 50% were bidirectionally arranged, and in total 290 genes with housekeeping functions were found to be bidirectionally arranged 30% of the time (Adachi *et al.*, 2002). Therefore in the human genome bidirectionally arranged genes have housekeeping functions three times more frequently than expected. Housekeeping functions are those related to cellular maintenance, and this is one of the four criteria to be a housekeeping gene set out by Joshi *et al.* in 2022. The other three are: stability of expression, conservation across species, and that their function is essential (Joshi *et al.*, 2022). The fact that bidirectional genes are more likely to have housekeeping functions was validated in a subsequent study by GO term enrichment analysis (Lui *et al.*, 2010).

Evidence from GO enrichment analysis supports that the disproportionate number of housekeeping bidirectional genes is common across species. In insects housekeeping genes were found to be enriched, for example the most enriched GO term in *Drosophila melanogaster* was “reproduction” (Behura and Severson, 2014). In maize, the most relevant

GO terms in the biological processes domain as identified through visualisation on a direct acyclic graph, which were “nucleobase-containing compound metabolic processes”, “cellular protein metabolic processes” and “stress response”, the first two of which are housekeeping in function (Liu *et al.*, 2014). Direct acyclic graphs are similar to flow charts in that they represent a flow of data, but they do not allow cycles, which makes it suitable for representing the strictly hierarchical GO terms. The same study also compared the GO annotation in maize to rice, sorghum, thale cress, and soybean, finding the bidirectional genes in all five played similar roles in metabolic processes, therefore bidirectional genes are enriched for housekeeping functions in plants as well as animals (Liu *et al.*, 2014). Finally, this finding agrees with the fact that housekeeping genes tend to be controlled by CpG island promoters (Cross and Bird 1995).

The sharing of promoters is thought to be one advantage of gene clustering (Gluck-Thaler *et al.*, 2018). Gene clustering is when multiple genes involved in the same pathway or with a related function are present in close proximity in the genome (Lopez *et al.*, 2010). Genes with a similar expression pattern also tend to cluster together (Lopez *et al.*, 2010). These are sometimes referred to as “operon-like” but unlike prokaryotic operons the genes are not transcribed as one mRNA strand, but separately. This explains how it’s possible that although under specific conditions all genes within a cluster are co-expressed, under other conditions only select genes within this cluster are expressed (Nutzmann *et al.*, 2014). Therefore, bidirectional genes may act in concert with other genes within the same cluster.

In conclusion, bidirectional promoters are emerging as an important aspect of gene regulation, and while headway has been made in determining their function and mechanism in the human genome, they’ve been little studied in plants. There are two key differences between human and plant bidirectional promoters: firstly, the transcription factors which control the bidirectional transcription differ, and secondly, the function of bidirectional genes in plants may also be related to response to stress as well as having housekeeping functions. Therefore, investigation into this aspect of gene regulation in the crucially important crop *Triticum aestivum* will certainly improve our understanding of the way in which its complex genome is regulated, and perhaps uncover novel ways in which it responds to stress. In the long term this could help inform wheat breeding efforts to produce cultivars that are tolerant to stresses that are becoming more common due to climate change.

1.2 An agronomically relevant polygenic trait: Coleoptile length

The coleoptile is a sheath-like structure in monocotyledonous plants from which the first leaf, called the cotyledon or scutellum, emerges (Wei *et al.*, 2022). Its function is to protect the emerging cotyledon; coleoptiles that are too short result in poor emergence (Wei *et al.*, 2022). The length of the coleoptile therefore limits the depth at which the seed can be planted in the soil. This becomes a problem in arid regions because deeper planting is beneficial in drought-stressed environments as it allows for access to water at deeper levels. Coleoptile length is also a predictor of drought-appropriate phenotypes that happen much later in development, such as yield and fertile tiller number, so it’s attractive to wheat breeding programs (Abdolshahi *et al.*, 2021). Using coleoptile length to predict phenotypes such as fertile tiller number is less accurate than phenotyping for these traits specifically but it means the screen is substantially quicker as the wheat only needs to be grown for 10 days rather than until maturity (Abdolshahi *et al.*, 2021).

It's feasible to breed wheat with longer coleoptiles as the trait is highly heritable (Rebetzke *et al.*, 1999). It is also a polygenic trait, under the control of many genes. As of 2023, 114 QTLs for coleoptile length in wheat had been identified, and some molecular markers for breeding have been developed (Xu, D. *et al.*, 2023; Singh *et al.*, 2015).

1.2.1 Genetic Influences: *Rht*

There are difficulties in breeding the commonly grown semi dwarf varieties of wheat with longer coleoptiles. Semi dwarf varieties developed in the green revolution are less prone to lodging and have a greatly increased harvest index (yield). This is genetically underpinned by mutations in the reduced height genes, *Rht1* (*Rht-B1b*) and *Rht2* (*Rht-D1b*), which cause gibberellic acid (GA) insensitivity, and decrease stem extension (Rebetzke *et al.*, 1999). However, *Rht-B1b* also reduces root length and coleoptile length via multiple pathways including those related to circadian rhythm and sucrose metabolism, so the plants have poor emergence and the seeds can't be sewn as deeply (Xu *et al.*, 2023; Sukhikh *et al.*, 2021). Under deep sowing less than 40% of GA-insensitive, *Rht-B1b* wheat plants emerged (Rebetzke *et al.*, 2007). These phenotypes are unsuitable for drought-stressed conditions to the extent that wheat varieties with *Rht* genes don't produce a higher yield than non-modified wheat under drought conditions (Kurishbayev *et al.*, 2020). Therefore, farmers choose not to grow the semi-dwarf varieties that were a product of the green revolution in arid regions (Rebetzke *et al.*, 1999). This presents challenges as most elite wheat carries one of the semi-dwarfing alleles and is the genetic background to which most subsequent breeding improvements have been made, for example. disease resistance.

There are alternatives to *Rht-B1b* and *Rht-D1b* as there are more than eighteen *Rht* gene variants now known in wheat, which are divided into two categories: GA sensitive and GA insensitive (Sukhikh *et al.*, 2021). Some of these alleles, such as *Rht-B1c* and *Rht-D1c*, are unsuitable for wheat breeding as they cause a dwarf phenotype, resulting in stunted growth and low yield, but other *Rht* genes could be used, whether alone or in combination, to produce desired phenotypes (Sukhikh *et al.*, 2021). For example the GA sensitive *Rht8* and *Rht24* result in semi-dwarf varieties but no effect on coleoptile length has been found (Xu *et al.*, 2023).

It's therefore possible for farmers to be able to benefit from both the increased yield of dwarf varieties and drought resistance of long coleoptile varieties by using GA sensitive *Rht* genes to create semi dwarf wheat plants while still allowing for the selection of long coleoptile genes.

1.2.2 Environmental Influences: Drought and temperature

Genetic factors are most important in deciding coleoptile length, but some environmental variables also play a role (Botwright *et al.*, 2001). Coleoptile growth is inhibited by light, and thus coleoptiles reach a longer length when wheat is grown in complete darkness (Wei *et al.*, 2022; Pinthus and Abraham 1996). As the main function of the coleoptile is to protect the first leaf as it emerges from the soil, further extension is not needed once the coleoptile has broken the surface of the soil and is exposed to light.

When grown in darkness cooler temperatures promote longer coleoptiles (Pinthus and Abraham 1996). This was especially true in GA insensitive semi-dwarf varieties; semi-dwarf varieties that were short due to GA insensitivity had a reduction in coleoptile length of 30 mm when grown at 19°C compared to 11°C, meanwhile GA sensitive semi-dwarf varieties that were short due to mutations in other genes had a reduction in coleoptile length of only 15 mm at 19°C compared to 11°C (Botwright *et al.*, 2001). The increased length at cooler temperatures was due to a longer elongation period at 11°C that was not observed at 25°C (Pinthus and Abraham 1996). High temperature increases the elongation rate but decreases the duration of elongation by 55%, so even though the coleoptile grew faster at higher temperatures, the coleoptile was shorter overall (Pinthus and Abraham 1996).

Varieties of wheat were categorised as drought tolerant and intolerant based on DTC (drought tolerance coefficient) values, the tolerant varieties had no variation in coleoptile length in drought stress whereas the drought-sensitive cultivars had shorter coleoptiles under drought conditions (Wei *et al.*, 2022). In 662 varieties of wheat, including spring and winter cultivars, coleoptile length on day 10 of growth ranged between 34 mm to 114 mm in length, when grown in darkness at 22°C in germination paper (Mohan *et al.*, 2013). Coleoptiles generally reached their maximum height on day 5 or 6 (Wei *et al.*, 2022). Wheat was usually planted at a depth of 20-40mm. When planted at a depth of 150mm, after 15 days varieties with a coleoptile length between 91-110mm had the greatest emergence rate (around 40%) of any group, including those with longer coleoptiles (111-120mm) (Mohan *et al.*, 2013). After 21 days, the emergence rate was around 45% (Mohan *et al.*, 2013).

Therefore, a coleoptile length of 91-110mm was desirable for wheat varieties that were planted in environments that suffer from a lack of rainfall in the sowing period. Introducing this trait into commercial cultivars such as Paragon or Cadenza would give farmers a greater choice of drought-suitable varieties. Paragon is a GA sensitive spring wheat commonly grown in the UK and produces high quality grains used for bread making (Amalova *et al.*, 2022). Cadenza is a facultative spring wheat with the wild type *Rht-D1a* allele, so is also GA sensitive (Martinez *et al.*, 2021).

No whole-genome approach to identifying and characterising bidirectional promoters and genes in wheat has previously been carried out, and therefore it is as of yet unknown what role bidirectional promoters and genes play in wheat. In this study I was interested in pursuing the following aims: carrying out a whole genome identification of bidirectional promoters in wheat, measuring and calculating key characteristics of the promoters and genes (e.g. whether the promoter are CpG islands, whether the genes are enriched for housekeeping functions), and using RNAseq data and GO terms to identify genes that respond to temperature. The first two aims are a whole genome approach in which I attempted to broadly characterise bidirectional promoters and genes, whereas the final aim focused particularly on temperature and drought related functions, as they may prove useful in breeding climate resistant wheat. More specifically, I decided to focus on the coleoptile length phenotype, as it is important when growing wheat in arid regions, is linked to drought appropriate phenotypes later in development, and is known to vary at different temperatures. I focused on cultivars Paragon and Cadenza for growth cabinet experiments as they are widely used commercial cultivars, and Cadenza has an available TILLING (targeting induced local lesions in genomes) population, which is a population with known random mutations in genes.

Chapter 2 - Detection of Bidirectional Promoters in Wheat

2.0 Bioinformatic detection of bidirectional promoters in wheat, rice, and thale cress

I aimed to identify all bidirectional promoters in wheat, rice, and thale cress. Any genes with TSS within 100-1000 base pairs of one another and located on opposite strands were defined as bidirectionally arranged, and the sequence between the TSS considered a bidirectional promoter (Trinklein *et al.*, 2004). This definition is commonly used, except I made the amendment that the minimum distance must be at least 100 base pairs, which is considered the shortest length a promoter can be (Le *et al.*, 2019).

2.0.0 Materials and Methods

2.0.0.0 Identifying genes arranged in a bidirectional orientation

EnsemblPlants biomaRt function was used to obtain all protein-coding genes and their respective transcription start sites, location, and orientation for the following organisms: *Triticum aestivum* (IWGSC), *Arabidopsis thaliana* (TAIR10), *Oryza sativa* Japonica (IRGSP-1.0), *Avena sativa* (Oat_OT3098_v2), *Brassica rapa* (SCU_BraROA_2.3), *Populus trichocarpa* (Pop_tri_v3), *Solanum lycopersium* (SL3.0), and *Zea mays* (Zm-B73-REFERENCE-NAM-5.0). The following attributes were selected, in order: Gene stable ID, Strand, Chromosome/scaffold name, Gene description, Gene start (bp), Gene end (bp), and TSS. Using R version 4.2.1 non-overlapping genes with TSSs within 100-1000bp of one another and on opposite strands were extracted (RStudio 2020, Supplementary 1A: script, Supplementary 1B: input files, Supplementary 1C: folder of output files). The bidirectional pairs were assigned an arbitrary pair ID.

2.0.0.1 Genome density calculation

Gene density per chromosome was calculated using the TSS of the most distal protein coding gene on the chromosome as an estimate of its total length. The total number of protein coding genes and total genome length was taken from EnsemblPlants.

2.0.1 Results

In the bread wheat reference genome, cv. Chinese Spring, 2,100 protein coding genes were found to be bidirectionally arranged, which represents 1.95% of all protein coding genes (Table 2). Of the species tested thale cress had the most bidirectionally arranged genes by absolute value and percentage, while oat had the fewest by percentage and cherry tomato by absolute number (Table 2). Wheat had the third largest by absolute but second smallest by percentage due to its high predicted gene number, a probable consequence of hexaploidy.

Table 2. Number of bidirectionally arranged protein coding genes expressed as both an absolute number and a percentage of total protein coding genes.

Genome	BiP Genes	Total Protein Coding Genes	Percentage BiP
Wheat (<i>Triticum aestivum</i>)	2100	107546	1.95
Thale Cress <i>Arabidopsis thaliana</i>	4632	27628	16.77
Rice (<i>Oryza sativa Japonica</i>)	2008	35775	5.61
Oat (<i>Avina sativa</i>)	1592	190245	0.84
Maize (<i>Zea mays</i>)	1362	54667	2.49
Field Mustard (<i>Brassica Rapa</i>)	3412	43129	7.91
Black Cottonwood (<i>Populus trichocarpa</i>)	1468	57146	2.57
Cherry Tomato (<i>Solanum lycopersicum</i>)	1106	34429	3.21

As observed here, plant species have extremely different total gene numbers and overall gene size. It's expected that some genes would have a bidirectional arrangement purely by random chance, in other words, if genes were positioned entirely randomly on a genome some would end up in a head-to-head arrangement. If more genes are present on a smaller genome, a greater number of genes would be arranged head-to-head, and vice versa for fewer genes on a larger genome. In order to determine how far genome density explains the fraction of bidirectional genes in the genome, a scatter plot comparing the genome density in genes per a million bases to the percentage of bidirectional genes was created for the eight plant species. A perfect correlation would suggest bidirectional genes are a product of genome density rather than *de novo* regulation.

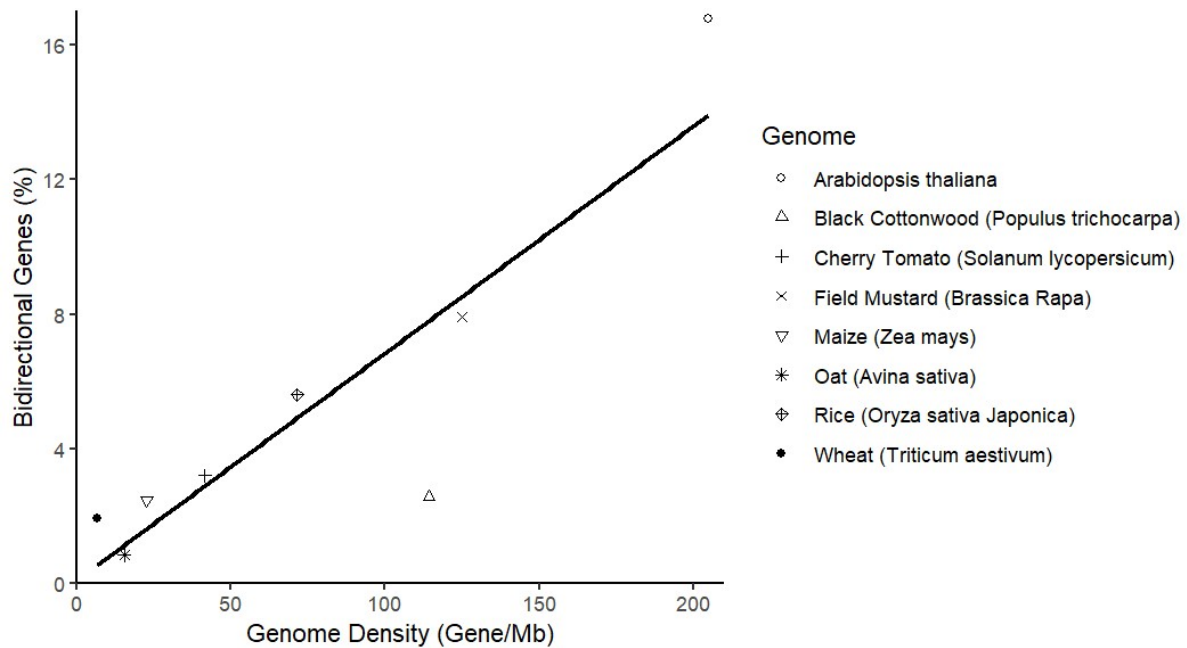


Figure 3. The genome density (number of genes divided by total gene length in Mb) plotted against the percentage of bidirectional genes present in eight plant species, each identified by a different shape. Pearson's correlation coefficient; $R^2 = 0.76$; p -value < 0.001 .

The genome density has a strong positive correlation with the percentage of bidirectional genes in this sample of plant species, with a more gene dense genome thale cress fits broadly with the pattern but has more bidirectional promoters than the trend would predict, while Black Cottonwood breaks pattern with a genome density of 114.3 genes/mb but only 2.57% bidirectional genes, barely more than the much less gene dense wheat, which has a genome density of 6.81 genes/mb (Figure 3).

To further investigate if genome density can explain the abundance of bidirectional genes within the genome, the wheat genome was separated into its chromosomes and the percentage of bidirectional genes was calculated for each (Figure 4). I hypothesised that if genome density can explain the distribution of bidirectional genes on each chromosome, the most gene dense chromosomes will have the most bidirectional genes.

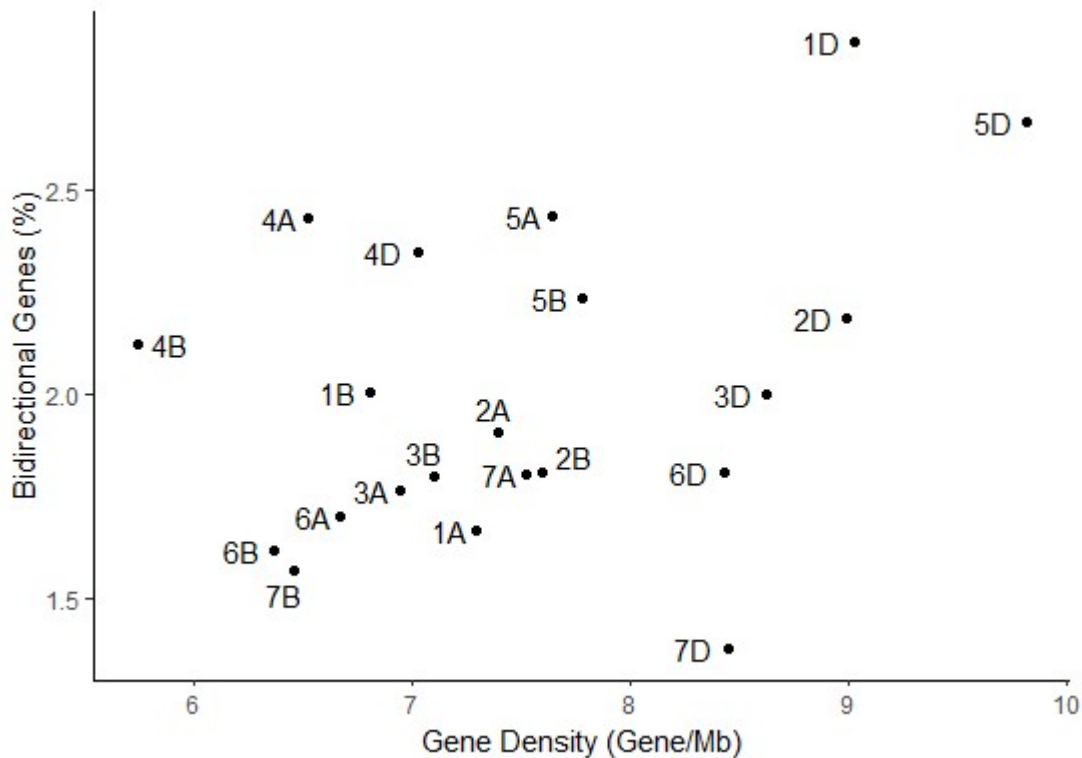


Figure 4. The gene density plotted against the percentage bidirectional genes for each chromosome. No significant correlation (Pearson's correlation coefficient; $R^2 = 0.11$; p -value > 0.05).

When split by chromosome the genome density did not explain the distribution of bidirectional genes. Chromosome 1D contains the highest number of bidirectional genes and 7D the least. The range in densities spans from 5.7 to 9.8 genes/mb. Chromosomes 4, 5, 2D and 1D have the highest fraction of bidirectional genes (Figure 4).

To visualise where on the genome these regions of high density bidirectional genes were, two different heatmaps were created, one in which the percentage of bidirectional genes per protein coding gene in an arbitrary region was indicated by colour (Figure 5), and one in which the absolute locations of the TSS were marked on the chromosomes (Figure 6). The former puts the number of bidirectional genes in the context of total genes, so accounts for the non-random distribution of genes on the genome, but doesn't provide any information about the actual number of bidirectional genes in that region, hence the creation of the latter plot.

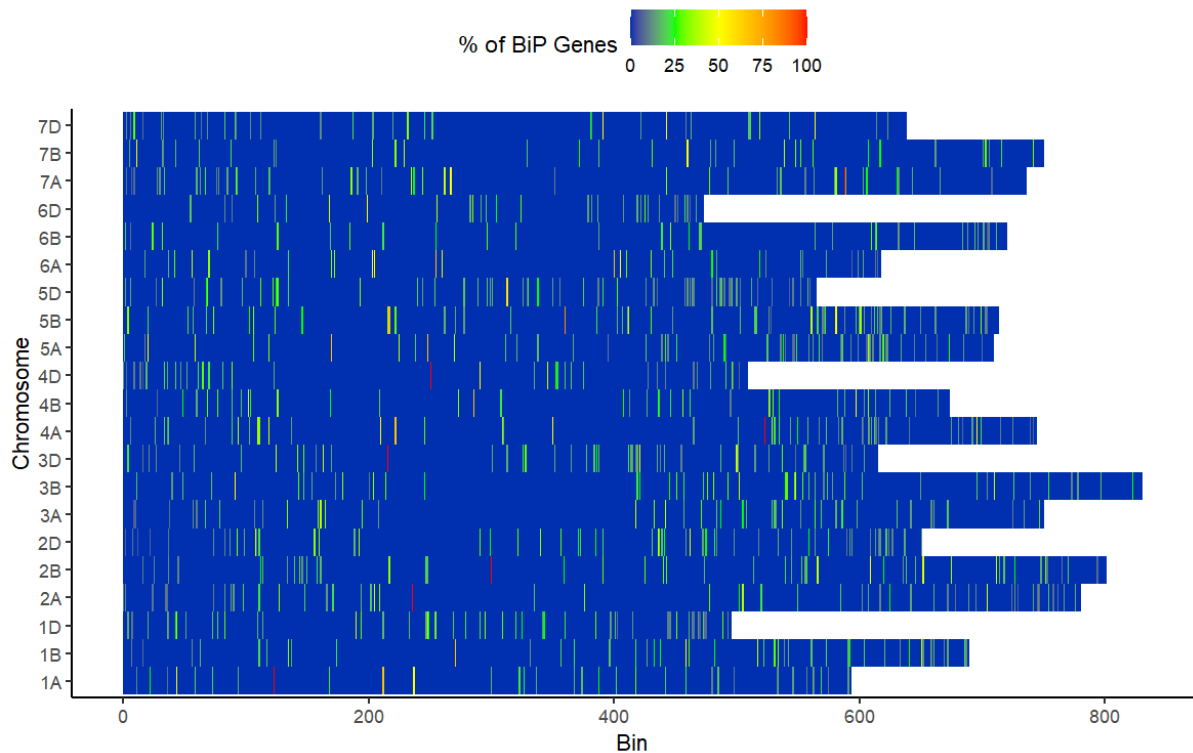


Figure 5. The height of each bar represents the length of each chromosome, which is split into bins representing 1 mb each. The chromosome length is therefore rounded up to the nearest million base pairs. The colour of the bins represents the percentage of genes in that region that are arranged bidirectionally (red=most dense; blue=least dense).

In seven regions the only two genes present are arranged bidirectionally, and thus 100% of the genes in those areas are bidirectionally arranged. These bins are 1A_123, 2A_236, 2B_300, 3D_216, 4A_523, 4D_251, and 7A_588, representing pairs 10, 169, 219, 437, 501, 596, and 967 respectively. The reverse gene of pair 219 was annotated as a malic enzyme.

Four are located within the centromeric region. The most gene rich region that had a fraction of bidirectional genes above 50% is on chromosome 5D (bin 5D_313) where eight of the fourteen genes are bidirectional, meaning 57% of genes in this region are bidirectional. The top three most gene rich regions were bin 1D_254, 5D_6, 3B_201 with 58, 56, and 48 protein coding genes respectively. Of these 12, 8, and 4 were bidirectional, respectively.

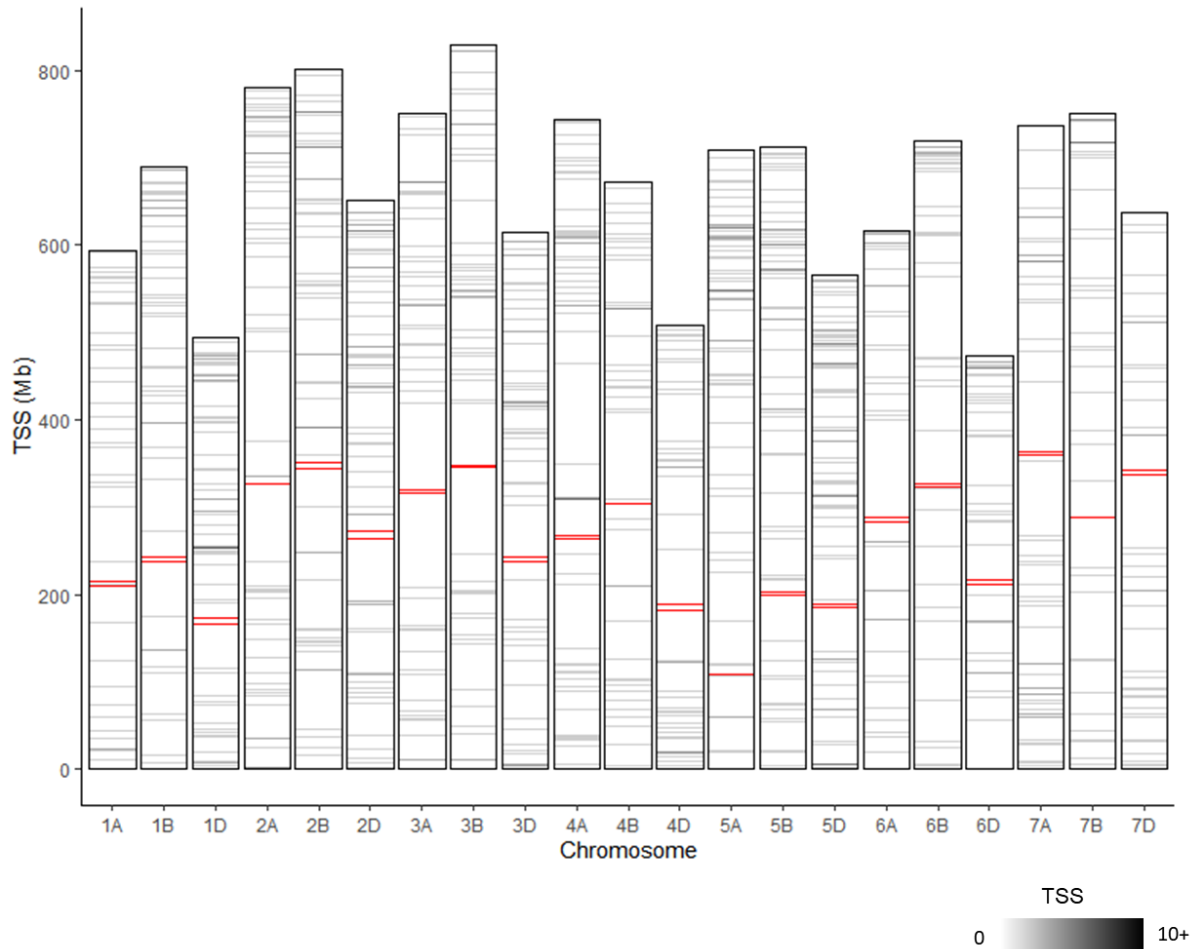


Figure 6. The location of the transcription start sites of bidirectionally arranged genes. Darker regions represent a greater number of overlapping TSS. The total length of the chromosomes is estimated by using the highest protein coding TSS. Red lines delineate the centromeric region.

Bidirectional genes are concentrated at the distal ends of the chromosomes, with the exception of the lower end of 6D, which doesn't have any bidirectional genes for the first 55mb, despite having unidirectional genes in that region. Interestingly, the highest concentrations of bidirectional genes are in the centromeric region of 1D and 4A, despite the centromeric regions being gene sparse compared to the rest of the genome.

2.0.2 Discussion

Wheat has fewer bidirectionally arranged genes in terms of percentage of its genome compared to thale cress and rice, which have smaller, more compact genomes than wheat. This is especially true of thale cress. Therefore I hypothesised that plants with greater genome density would have a greater fraction of bidirectional genes in their genome.

I compared the genome density of eight angiosperms representing three different orders to the percentage of bidirectional genes in its genome (Figure 3). The percentage of bidirectional genes was strongly correlated with genome density in the angiosperms tested

(Figure 3), but black cottonwood and thale cress deviated from this pattern, with the former having fewer bidirectional promoters than expected, and the latter having more. Perhaps the deviation from the pattern for blackwood cotton can be explained by the fact that it was the only species of tree included, but further data would need to be analysed to draw a conclusion. Wheat broadly fits with the pattern, but the predicted percentage of bidirectional promoters for wheat based on genome density is only 0.55% as compared to the overall average of 1.95%, around three and half times as much. Taken together, this evidences the fact that although genome density is correlated with bidirectionality, it's far from a perfect predictor, suggesting other factors influence how common bidirectional genes are in the genome.

One of these factors could be transposon activity. Discovered by Barbara McClintock in the 1940s they are genes that can move around in the genome either by a "copy and paste" mechanism in the case of retrotransposons or a "cut and paste" mechanism as in the case of P-element DNA transposons (Parhad and Theurkauf 2019). Their action is both useful in driving genome evolution and problematic, interrupting exons and promoters, resulting in genes being unexpressed or the production of truncated proteins (Dubin *et al.*, 2018). As transposons are one of the main drivers of increased genome size, transposons could give rise to bidirectionally arranged genes by random chance or through preferential insertion sites upstream of an existing gene. Bidirectional promoters could be advantageous by reducing the promoter length available for insertion, thus decreasing the likelihood of disruption of bidirectional genes, which often have housekeeping functions.

For those that have many more bidirectional genes than expected, like thale cress, other mechanisms than the action of transposons may be more important in the creation of new genes, such as gene duplication, which would make it more likely for the genes to be near to one another on the genome without substantially increasing genome size. Although, of course, duplicated genes would be on the same strand, so another process, such as the "evolutionary scrambling of genomes" by transposons, would have to occur to render them bidirectional (Fedorof 2012).

A similar correlation between genome density and ratio of bidirectional genes can be seen across eukaryotes, although each clade shows its own particular pattern (Yang and Yu 2009). For example insects tend to cluster together, all having similar gene densities and ratio of bidirectional genes (Yang and Yu 2009). Plants don't show this same clustering because they have a broad range of genome densities (Figure 3). Plants have a greater range of genome sizes than insects, animals, fungi, or any other group. While the DNA content per haploid genome in plants varies 2000 fold, the number of protein coding genes doesn't vary overmuch between taxa, thus resulting in a broader range of genome densities than other clades (Fedorof 2012). One of the main drivers of a large genome size, other than polyploidy, is the action of class 1 retrotransposons (Sung-II and Nam-Soo 2014).

It is known that genes can arise *de novo* from non-coding RNA, and although here the focus has been on protein coding genes, often a bidirectional promoter will transcribe a protein coding gene in one direction and a non-coding RNA in the other (Long *et al.*, 2013). Thus, it's plausible that the non-coding RNA becomes coding, resulting in two protein coding genes being bidirectionally transcribed by one promoter.

Bidirectional genes may be selected for in order to co-regulate gene pairs with related functions, enabling rapid and resource efficient regulation in response to stressors. Two genes under the same promoter would reduce the required cellular machinery and the myriad proteins that modulate gene expression. This could be particularly prevalent in some genomic contexts, such as that of wheat which must regulate, sometimes in a dosage dependent manner, thousands of genes across three genomes. The different selection pressures on bidirectional genes in different genomic contexts could explain some of the deviations from the correlation between percentage of bidirectional genes and genome density (Figure 3).

In conclusion, the relationship between gene size, number of genes, and number of bidirectional genes is influenced by how genes arise and are lost from the genome. Because of this, genome density is a good though not perfect predictor of the fraction of bidirectional genes in the genome in plants.

Genome density does not explain the distribution of bidirectional genes on chromosomes. Like all genes, more bidirectional genes are found at the distal ends of chromosomes, thus when split by chromosome the genome density does not explain the distribution of bidirectional genes (Sidhu and Gill 2005). Contributing to this is the fact that wheat genes tend to be found in gene rich regions, separated by long stretches of repetitive non-coding DNA (Sidhu and Gill 2005). Therefore, at the chromosome level the distribution of bidirectional genes is not random.

The D chromosomes are more gene dense than their counterparts, and the D genome has slightly more bidirectional genes than the other two. The older subgenomes A and B tend to be more similar to one another than either are to D in terms of gene density and percentage of bidirectional genes.

Bidirectionally arranged genes are present both independently and within gene rich regions. Bidirectional genes appear in seven instances as pairs of lone genes and are therefore not necessarily parts of larger gene clusters. Due to the genomic context of the genes however, they may be switched off and so not biologically important. On the other hand, recombination tends to occur more frequently in gene rich regions, and initiates at promoters, so having two housekeeping genes in a gene sparse region under the control of a single promoter may allow them to avoid recombination events that would alter their function (Shah and Hassan 2005). To investigate this hypothesis, I used expression data to determine if these genes are expressed, and make predictions as to their function. This will allow the hypothesis to be discounted if the genes are not expressed or do not have an integral function.

Other regions are gene dense and contain bidirectional promoters, potentially representing gene clusters with related functions. Software has been developed to identify gene clusters, which in combination with the data made available here, makes it possible for future studies to determine if bidirectional genes play a role in gene clustering (Chavali and Rhee 2017).

2.1 Orthology of bidirectional genes and the bidirectional gene arrangement

I aimed to identify whether bidirectionally arranged genes were more conserved than other genes, as higher than expected conservation is supporting evidence that these genes are

more likely to be housekeeping genes. I also wanted to determine if when one bidirectional gene was conserved, was its pair also conserved, and I refer to this as “the bidirectional gene arrangement” throughout. Finally, in order to use as a control, I identified proximal genes, which I defined as genes with TSS within a 1000 bp of one another, and on the same strand of the DNA.

2.1.0 Methods

2.1.0.0 Identifying orthology of bidirectional promoters with *Arabidopsis thaliana* and *Oryza sativa* Japonica

Proximal genes were identified for *T. aestivum*, which were defined as two protein coding genes with transcription start sites within 1000 base pairs of one another, located on the same DNA strand; they therefore could not be under the control of the same promoter. This was done by using a modified version of the code for identifying bidirectionally arranged genes located in Supplementary 1A, and simply changing the conditional statement from strandA != strandB (the genes must be on opposite strands) to strandA == strandB (the genes must be on the same strand). Four-hundred-and-sixty pairs or 938 genes were found to be proximal in wheat and can be found in Supplementary 1D > proximal_triticum_aestivum.csv.

The biomart function of EnsemblPlants release 57 (<https://plants.ensembl.org>) was used to obtain all *O. sativa* Japonica (IRGSP-1.0) and *A. thaliana* (TAIR10) orthologues for bidirectional and proximal wheat genes, which are listed in Supplementary 1C putative_BIP_triticum_aestivum.csv and 1D > proximal_triticum_aestivum.csv, respectively. On Ensembl the following attributes were selected: Gene stable ID, *Triticum aestivum* gene stable ID, *Triticum aestivum* orthology confidence [0 low, 1 high].

2.1.0.1 Conserved bidirectional arrangement

EnsemblPlants biomart was again used to obtain *T. aestivum* orthologues for bidirectionally arranged *A. thaliana* and *O. sativa* genes. The putative bidirectional genes identified for rice and *A. thaliana* were used as filters and the following attributes were of interest: Gene stable ID, *Triticum aestivum* gene stable ID, and *Triticum aestivum* orthology confidence [0 low, 1 high]. The resulting table was downloaded as two text files (Supplementary 2B).

The *T. aestivum* orthologues identified as outlined above are not necessarily themselves bidirectionally arranged, because gene arrangements differ between species. In order to filter these lists to retain only *A. thaliana* and *O. sativa* bidirectional genes with a bidirectional orthologue in *T. aestivum* a custom R script (RStudio Team 2020, Supplementary 2A) was used. The only rows retained were those in which the gene in the “Triticum aestivum gene stable ID” column was present in the list of bidirectionally arranged wheat genes, as previously determined and listed in Supplementary 2C.

In order to rank the quality of the orthologues, the total number of orthologues was recorded and the confidence ratings were summed, so that one to many mappings from *T. aestivum* to *A. thaliana* and rice were accounted for in a single row. Subsequently this data frame was subsetted to calculate the conservation between only *O. sativa* and only *A. thaliana*, in some

cases only one gene in a pair was conserved, so the final filtering step was to remove unpaired *T. aestivum* genes. If a *T. aestivum* bidirectionally arranged gene was present in the “*Triticum aestivum* gene stable ID” column, but its pair was not, the row was removed. It was assumed that if both genes had bidirectionally arranged orthologues that those two orthologues were bidirectionally arranged with each other, although that was not necessarily the case.

2.1.1. Results

The majority of bidirectional genes identified in *Triticum aestivum* have an orthologue in rice and thale cress.

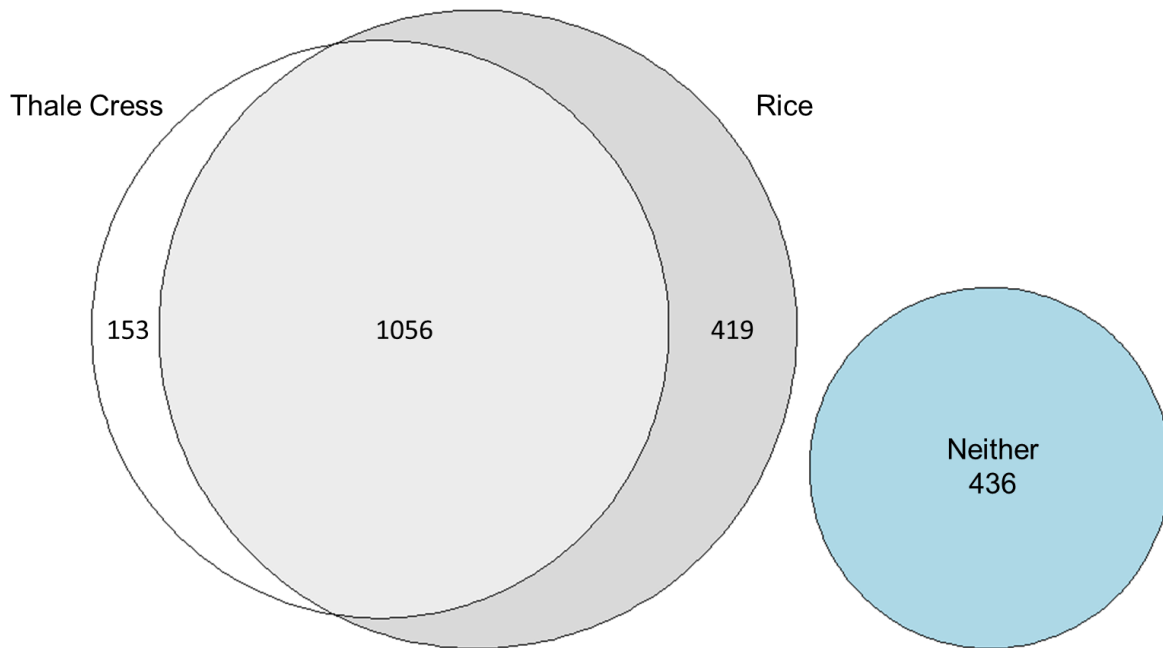


Figure 7. Venn diagram where each circle represents the number of wheat bidirectional genes that have an orthologue (not necessarily bidirectionally arranged) in rice/thale cress.

Thale cress and wheat had 153 orthologues that were not also orthologous with rice, rice and wheat had 419 orthologues that were not also orthologous with thale cress. Around 78% of bidirectional genes have at least one orthologue in rice or thale cress (Figure 7). Meanwhile, 57% have at least one orthologue in thale cress and 70% have at least one orthologue in rice (Table 3).

Table 3. Percentage of all, bidirectional (BiP), and proximal wheat protein coding genes that have orthologues in rice and thale cress.

	Rice (%)	Thale Cress (%)
All	65	47
BiP	70	57
Proximal	40	34

The analysis identified that bidirectional genes are more conserved than proximal genes (Table 3), and more conserved than average for all protein coding genes in the genome.

Table 4. Conserved bidirectional arrangement between wheat, rice, and thale cress. The first column represents the species with which the arrangement is conserved, the second is the number of bidirectional pairs in wheat that have a conserved bidirectional arrangement with that species and the final column represents the number of wheat paired genes as a percentage of all wheat bidirectional genes.

Conservation with:	Number of Conserved Wheat Pairs	As a Percentage of All Wheat Bidirectional Genes
<i>A. thaliana</i>	53	5.0
<i>O. sativa</i>	259	24.7
<i>A. thaliana</i> and <i>O. sativa</i>	22	2.1
<i>A. thaliana</i> or <i>O. sativa</i>	290	27.6

My analysis identified that 27.6% of bidirectional genes remain in a conserved arrangement. The upper limit is 78% which would mean that every conserved gene is bidirectionally arranged, the lower limit is 0% which would mean none of the conserved bidirectional genes have a conserved arrangement. The remaining fraction are those that have no orthologue in rice or thale cress and were therefore not considered conserved.

Twenty-two bidirectional pairs are conserved between wheat, rice, and thale cress. Of the 22 most conserved pairs, 16 of 44 genes are annotated by Ensembl, with 6 unique annotations which are summarised as follows: tousled-like protein kinase (TSL-like), protein-lysine N-methyltransferase, F-box/kelch-repeat protein SKIP4, dual-specificity phosphatase CDC25, DEAD-box ATP-dependent RNA helicase 57, and heat stress transcription factor.

2.1.2 Discussion

Bidirectional genes are well conserved (table 3). This is likely because they have housekeeping functions that are integral to the functioning of the cell. Housekeeping genes by definition are highly conserved and bidirectional genes are more likely to have housekeeping functions (Liu *et al.*, 2011). To confirm this hypothesis I carried out a GO term enrichment analysis.

The greater conservation of bidirectional genes between wheat and rice compared to wheat and thale cress is consistent with the fact that wheat and rice are more closely related; both are monocots belonging to the order *Poales* that diverged from one another around 60 million years ago (Charles *et al.*, 2009). Thale cress meanwhile is a dicot belonging to the order *Rosid* and diverged from rice and wheat around 150 million years ago, so is regulated very differently to rice and wheat (Chaw *et al.*, 2004).

The bidirectional arrangement is conserved between rice and wheat. Synteny among grasses has traditionally been considered to be high, based on marker evidence before the sequencing of the wheat genome (Kellogg 2001). But past studies using only markers could not account for whole genome duplications or distinguish easily between paralogues and orthologues, so overestimated the similarity between genomes (Minx *et al.*, 2005). Of 42,654 rice genes, which covered 83.1% of the rice genome, 1180 were conserved in wheat, representing 90.4% of the wheat genome (Bolot *et al.*, 2009). 27.2% of these differed from their expected position based on synteny, indicating genome rearrangement has occurred in these regions (Bolot *et al.*, 2009). Even so, the genomes of grasses are sufficiently similar to be aligned against one another, which has been done to make inferences about the last common ancestor of the grass family (Wang *et al.*, 2015).

One aspect that affects synteny between species is the gain and loss of chromosomes. The ancestor of all grasses, like thale cress, would have had five chromosomes (Salse *et al.*, 2008). A whole genome duplication event a little under 90 million years ago, before the divergence of rice and wheat, resulted in a chromosome number of 10 (Salse *et al.*, 2008). This accounts for the similarity of the wheat chromosomes 1 and 3, and their respective similarity with chromosomes 5 and 1 in rice (Salse *et al.*, 2008). Two fusion and breakage events resulted in a chromosome number of 12, which is still the case in rice today, but since the divergence of wheat and rice further breakage and fusion events have occurred in wheat ($n=7$) followed by two hybridisation events ($n=21$) that gave rise to modern hexaploid wheat (Salse *et al.*, 2008).

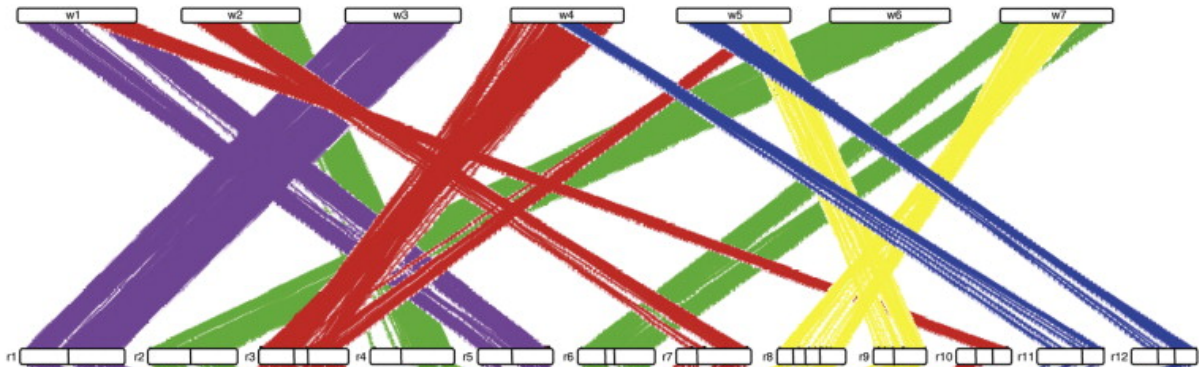


Figure 8. Adapted from Bolot *et al.*, 2009. Synteny between wheat and rice. W1-7 represents the seven wheat chromosomes, and r1-12 the twelve rice chromosomes. The strands linking the chromosomes represent orthologues. The colours differentiate between the five ancestral chromosomes of *A. thaliana* (A5=purple, A7=red, A11=blue, A8=yellow, A4=green).

The differences between grass genomes and thale cress are even greater due to their earlier divergence. Whereas rice and wheat genes are found in clusters separated by repetitive, gene sparse regions, genes in thale cress are more evenly distributed (Barakat *et al.*, 1998). There are no large syntenic blocks in common between thale cress and rice, and only 10 strictly similar collinear regions exist, with sizes ranging from 2 to 63 kb and a mean length of 25 kb (Liu *et al.*, 2001). Many of the similar regions coincided with gene clusters, for example, the cytochrome P450 cluster and the glycosyltransferases were more than 82% and 72% respectively were identical between rice and thale cress (Liu *et al.*, 2001).

Since rice has been evolving more slowly than wheat, and wheat has undergone additional chromosome breakage and fusion events as well as two hybridisation events, synteny between wheat and thale cress is likely to be absent, except in conserved gene clusters. Despite this 5.0% of wheat bidirectional pairs had a conserved arrangement with thale cress, or 1.7% if you roughly adjust for polyploidy by dividing by three. There are 19 pairs on the A subgenome, 18 on the B subgenome, and 16 on the D subgenome, and the conserved pairs are spread across all of the chromosomes, excepting 6D, so are not clustered together. Where the bidirectional arrangement is conserved between thale cress and wheat then appears not to be due to them appearing in conserved syntenic blocks, although more investigation is needed to be conclusive. The pairs could have instead escaped being separated in genomic rearrangements due to their proximity.

A much higher percentage of conservation exists between rice and wheat bidirectional genes, which agrees with their higher synteny and later divergence. Chromosome 2 has the greatest number of conserved pairs between wheat and rice with 63 pairs, and chromosome 4 the least, with only 23 pairs. The conservation, again, doesn't seem to be explained by syntenic blocks as then you would expect chromosome 3 to have the largest number of conserved pairs. To test the hypothesis that the proximity of the pairs explains their conservation the percentage of conservation in proximal pairs could be determined. A significantly higher number of conserved bidirectional pairs would indicate another factor is at play in maintaining their arrangement.

There were several BLAST searches that matched to characterised proteins. CDC25 and TSL-like proteins are involved in the normal functioning of the cell cycle and are highly conserved between plants and animals (Ehsan *et al.*, 2004; Landrieu *et al.*, 2004). Protein-lysine N-methyltransferases are involved in modifying chromatin structure and are therefore involved in numerous important pathways in eukaryotes (Zhou *et al.*, 2020). F-box proteins are one of the largest protein families in plants, thale cress containing about 700 of them, many in poorly characterised pathways, RNA helicases unwind double stranded RNA and are found in nearly every organism, and heat stress transcription factors regulate response to high temperature in animals and plants (Zhang *et al.*, 2019; Tanner and Linder 2001; Ye *et al.*, 2020; Ponomarenko and Kolchanov). These six annotations point to proteins that are well-conserved not only between plants but across all domains of life and are involved in essential processes.

Perhaps the most interesting to us are the heat stress transcription factors, which represent three homoeologous genes (TraesCS5A02G437900, TraesCS5B02G440700, TraesCS5D02G445100) that have unannotated pairs. The annotation is projected from Os03g0795900 in *O. sativa* where it encodes heat shock factor 12, also known as HSFA2E. Transforming thale cress with HSFA2E conferred increased thermotolerance at the cotyledon stage (Yokotani *et al.*, 2007).

2.2 Co-expression analysis of bidirectional genes from RNAseq data

In this section I aimed to determine whether bidirectional genes pairs are more likely to be coexpressed with one another than proximal pairs, as well as to identify correlations between coexpression and other known values such as promoter length.

2.2.0 Methods

Raw expression data for all wheat bidirectional and proximal genes was downloaded in batches from the wheat expression browser (<http://www.wheat-expression.com/>; references in table below). This was used to calculate the co-expression coefficients (Pearson's correlation coefficient) for all bidirectional and proximal gene pairs as listed in Supplementary 1A and 1D respectively, using a custom R script (Supplementary 5A; RStudio Team 2020). Pairs where either of the genes were not expressed at a level of 0.5 tpm or higher in any condition were considered not expressed and removed from the dataset before calculation. P-values were adjusted using Benjamini-Hochberg correction.

Table 5. References for the RNAseq data used in order to calculate co-expression coefficients.

RNA-seq study title	Reference
The transcriptional landscape of hexaploid wheat across tissues and cultivars	Ramirez-Gonzalez <i>et al.</i> , 2018
Genome interplay in the grain transcriptome of hexaploid bread wheat	Pfeifer <i>et al.</i> , 2014
Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family	Pearce <i>et al.</i> , 2015
Effect of the down-regulation of the high Grain Protein Content (GPC) genes on the wheat transcriptome during monocarpic senescence	Cantu <i>et al.</i> , 2011
Gene expression in the developing aleurone and starchy endosperm of wheat	Gillies <i>et al.</i> , 2012
Genome analyses of the wheat yellow (stripe) rust pathogen <i>Puccinia striiformis</i> f. sp. <i>tritici</i> reveal polymorphic and haustorial expressed secreted proteins as candidate effectors	Cantu <i>et al.</i> , 2013
Transcriptional Reprogramming of Wheat and the Hemibiotrophic Pathogen <i>Septoria tritici</i> during Two Phases of the Compatible Interaction	Yang <i>et al.</i> , 2013
Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (<i>Triticum aestivum</i> L.)	Kugler <i>et al.</i> , 2013
Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat	Leach <i>et al.</i> , 2014
Pistillody mutant reveals key insights into stamen and pistil development in wheat (<i>Triticum aestivum</i> L.)	Yang <i>et al.</i> , 2015
Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew	Zhang <i>et al.</i> , 2014
Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat	Liu <i>et al.</i> , 2015

(Continued overleaf)

RNA-seq study title	Reference
Transcriptomic analysis of wheat near-isogenic lines identifies PM19-A1 and A2 as candidates for a major dormancy QTL	Barrero <i>et al.</i> , 2015
Emergence of wheat blast in Bangladesh was caused by a South American lineage of <i>Magnaporthe oryzae</i>	Islam <i>et al.</i> , 2016
Analysis of wheat microspore embryogenesis induction by transcriptome and small RNA sequencing using the highly responsive cultivar "Svilena"	Seifert <i>et al.</i> , 2016
The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression	Dobon <i>et al.</i> , 2016
Characterisation of the wheat (<i>triticum aestivum L.</i>) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat	Oono <i>et al.</i> , 2013
A synthetic tuber-specific and cold-induced promoter is applicable in controlling potato cold-induced sweetening	Li <i>et al.</i> , 2013
Structural And Functional Partitioning Of Bread Wheat Chromosome 3b	Choulet <i>et al.</i> , 2014
An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations	Bernardo <i>et al.</i> , 2017
Suppressed recombination and unique candidate genes in the divergent haplotype encoding Fhb1, a major Fusarium head blight resistance locus in wheat	Schweiger <i>et al.</i> , 2016
The Fusarium crown rot pathogen Fusarium pseudograminearum triggers a suite of transcriptional and metabolic changes in bread wheat (<i>Triticum aestivum L.</i>)	Powell <i>et al.</i> , 2016
Exogenous Abscisic Acid and Gibberellic Acid Elicit Opposing Effects on <i>Fusarium graminearum</i> Infection in Wheat	Buhrow <i>et al.</i> , 2016
Transcriptome and Metabolite Profiling of the Infection Cycle of <i>Zymoseptoria tritici</i> on Wheat Reveals a Biphasic Interaction with Plant Immunity Involving Differential Pathogen Chromosomal Contributions and a Variation on the Hemibiotrophic Lifestyle Definition	Rudd <i>et al.</i> , 2015

(continued overleaf)

RNA-seq study title	Reference
Development and Validation of <i>Thinopyrum elongatum</i> –Expressed Molecular Markers Specific for the Long Arm of Chromosome 7E	Gou <i>et al.</i> , 2016
Transcriptome and Allele Specificity Associated with a 3BL Locus for Fusarium Crown Rot Resistance in Bread Wheat	Ma <i>et al.</i> , 2014
Understanding the Biochemical Basis of Temperature-Induced Lipid Pathway Adjustments in Plants	Li <i>et al.</i> , 2015
mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat	Li <i>et al.</i> , 2014
Identification of Transcription Factors Regulating Senescence in Wheat through Gene Regulatory Network Modelling	Borrill <i>et al.</i> , 2019
Genome-Wide Transcription During Early Wheat Meiosis Is Independent of Synapsis, Ploidy Level, and the Ph1 Locus	Martín <i>et al.</i> , 2018

R studio was used to calculate the co-expression coefficients. Data was imported from Supplementary file 2C which contains a list of wheat bidirectional genes, as well as corresponding information, including: pair IDs, description, strand, and distance (distance between the TSS, equal to the length of the promoter in the case of bidirectional pairs). Additionally, tpm data was imported, as downloaded from the wheat expression browser for bidirectional genes.

One issue with calculating expression coefficients is that two genes with no expression at all will have a perfect correlation of 1. This is misleading, as in conditions where those genes are expressed, they may not have a similar expression profile at all. To circumvent this problem, pairs where neither were expressed were excluded. Pairs where neither were expressed was defined as when neither pair had expression above 0.5 tpm in any condition. 0.5 was used as the cutoff because anything below this is likely to be an artefact of the experiment and is unlikely to reflect true expression.

To exclude these pairs of columns from the analysis an empty data frame was created and columns were added to it by pairs if they met the condition. This was done by looping through every unique pair ID in the bidirectional data frame, where each row represented a pair, pulling out the gene names and using these to index the tpm data frame and selecting the relevant columns. The maximum value for each of these two columns was calculated, and then the maximum value of those two values. If this number was greater than 0.5 those two columns were appended to the new data frame. This resulted in a matrix with tpm data for expressed bidirectional genes. This was converted to a data frame.

The function `rcorr()` from the `Hmisc` library was used to calculate the correlation coefficients (Harrell 2022). The function calculates the correlation coefficients for all provided genes against each other, so the co-expression coefficient for every possible combination of genes was calculated, not just pairs. The output co-expression coefficients and p-values were stored in two separate data frames.

The dataset was then reformatted so that each row represented one pair rather than one gene, so that the co-expression coefficients and p-values could be appended to the dataframe without repeating or missing cells. This was done by splitting the data frame into odd and even row numbers and then merging these data frames together by their pair ID. Any repeated columns were removed before merging, except “Gene_2” which was used to confirm the pairs had been matched correctly. A coefficient was calculated for 1040 of 1050 bidirectional pairs, and 448 of 469 proximal pairs.

In order to add the co-expression coefficients and p-values to the data frame a custom function was designed that takes a row index as an argument and uses it to select a single row of the data frame containing the gene pairs. The gene names are pulled out, used to select the correct co-expression coefficient and p-value from their respective data frames and then returned as a two element vector. This function was used in combination with the base R `sapply()` function to carry out the procedure for every row in that data frame. The resulting data frame was then transposed with the `t()` function from the `rlist` library, and these columns were bound to the gene pair data frame (Ren 2021).

The preceding steps were repeated for proximal genes, using data from Supplementary 1D. A “Type” column was added to both resulting data frames to be able to distinguish between proximal and bidirectional genes, then the data frames were bound together. With all the data in one place multiple test corrections were carried out (Benjamini and Hochberg False Discovery Rate) on the p-values. The co-expression coefficient was set to 0 in cases where the corrected p-value was less than 0.05. The distribution of co-expression coefficients were then compared for each category.

2.2.1 Results

As the main adaptive function for bidirectional promoters is thought to be the ability to quickly and resource efficiently co-regulate pairs of genes, I decided to calculate the co-expression coefficients (Pearson’s correlation coefficient) of bidirectional and proximal pairs. The co-expression coefficient uses expression data, such as that generated from qPCR or RNAseq, and is an estimate of how co-expressed two genes are, where -1 is a perfectly negative correlation, 0 is no correlation, and 1 is a perfectly positive correlation.

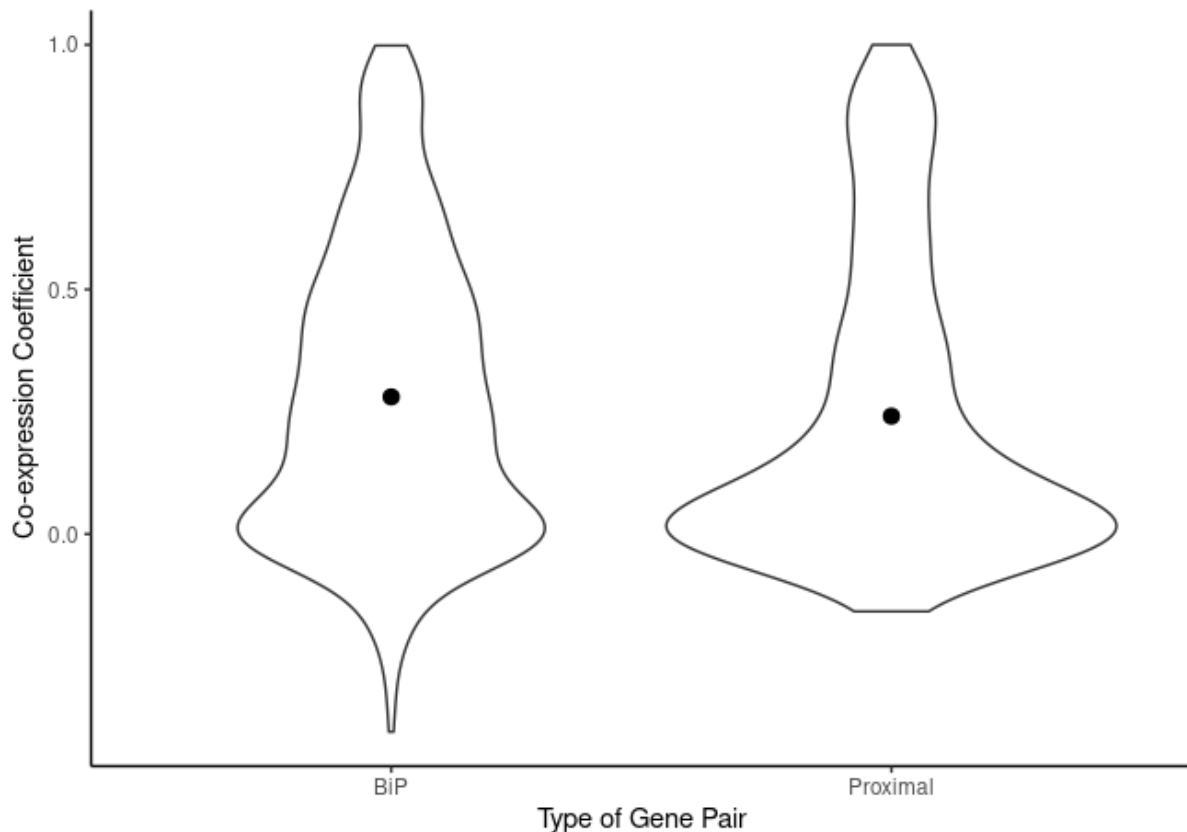


Figure 9. The proportional distribution of co-expression coefficients of bidirectional and proximal pairs, non-significant (Pearson's correlation coefficient; adjusted p -value > 0.05) co-expression coefficients were set to 0. The black dots represent the mean. Bidirectional pairs are more likely to be co-expressed (two sample t -test; p -value < 0.05).

I discovered that there were more anti-regulated bidirectional than proximal pairs, and the most anti-regulated bidirectional pair had a much lower co-expression coefficient than the proximal category: -0.4 compared to only -0.16 (Figure 9). Only 1.07% of proximal pairs had a significant negative coexpression coefficient after correction, so anti-regulated neighbouring genes are actually extremely uncommon in wheat, and even when they are present the correlation is relatively close to zero. In bidirectional genes however it's 5.5 times more common with 5.9% of bidirectional pairs being anti-regulated and the correlation can be much stronger.

Bidirectional pairs are also more likely to be co-regulated than proximal pairs ($p < 0.05$), although there are slightly fewer near perfect correlations, there are more pairs with a correlation between 0.2 and 0.7 and fewer with no significant correlation at all (Figure 9).

While the true co-expression coefficients are useful for distinguishing between co- and anti-regulated pairs, the mean tends towards zero because the co- and anti-regulated pairs cancel one another out. Therefore the absolute value of the co-expression coefficients was taken to negate this effect. In this case the strength but not the direction (positive or negative) of the correlation is taken into account.

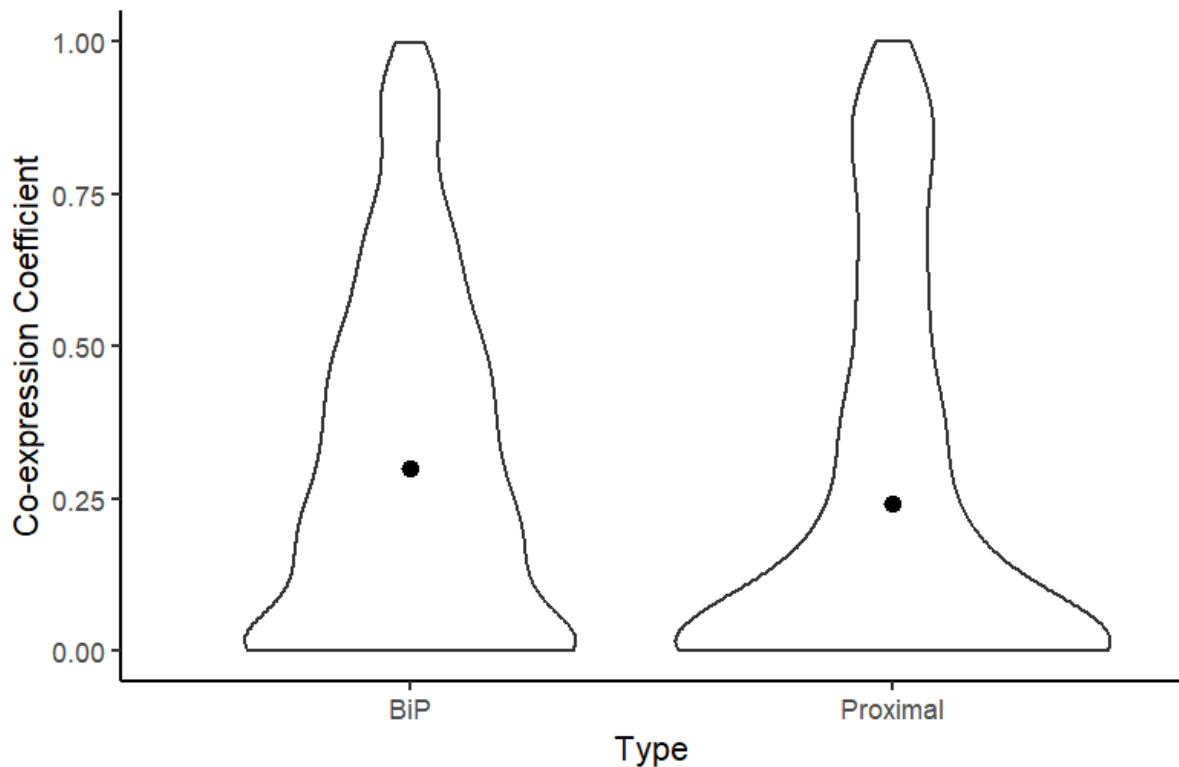


Figure 10. The proportional distribution of absolute co-expression coefficients of bidirectional and proximal pairs, non-significant (Pearson's correlation coefficient; adjusted p -value > 0.05) co-expression coefficients were set to 0. The black dots represent the mean, bidirectional (BiP) = 0.3; proximal = 0.24. Bidirectional pairs are more likely to be either anti- or co-regulated (two sample t -test; p -value < 0.001).

The difference between the means in this case is both more evident and more significant ($p < 0.001$) (Figure 10). The most coexpressed pair in any category was proximal pair 145 with a perfect co-expression coefficient of one, a closer look at the pair revealed the expression data for both genes was identical. This is likely because the raw data was recorded when a previous genome reference was in use, and a single gene from that reference has been mapped onto two genes from the newest reference genome. This data point is therefore not an accurate reflection of co-expression between the genes in pair 145. Excluding this data point, the top 5 most highly coexpressed pairs (coexpression coefficients: 0.977-0.998) were all bidirectional.

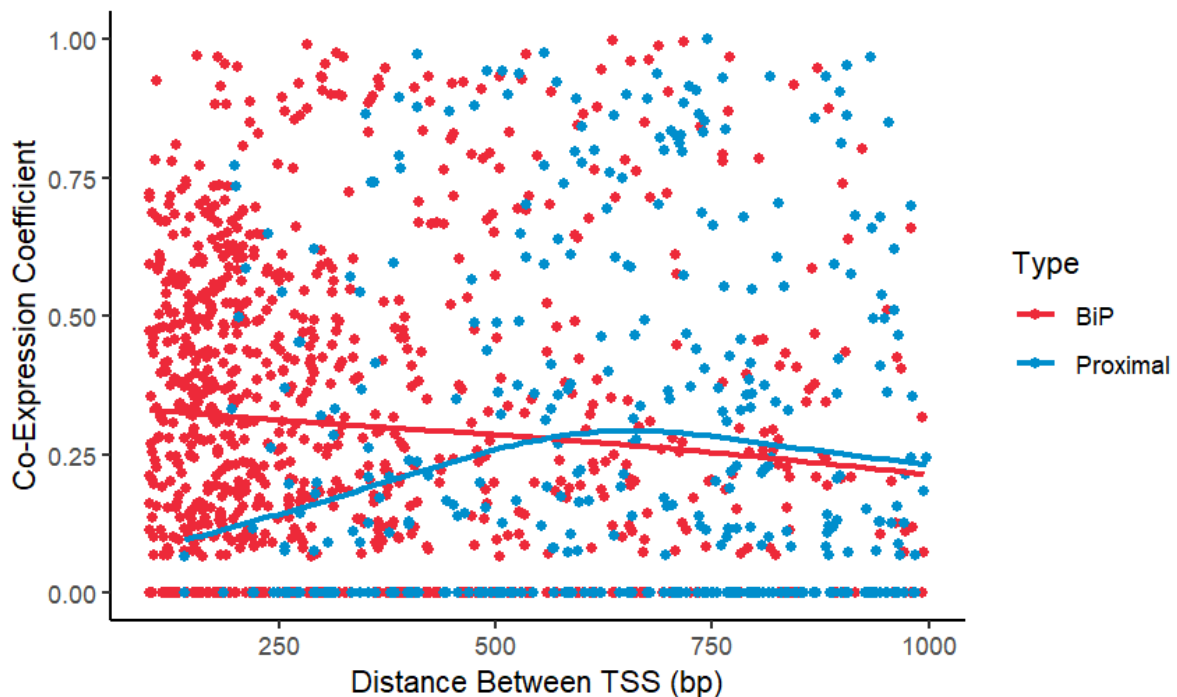


Figure 11. Absolute co-expression coefficient of bidirectional and proximal pairs correlated with the distance between transcription start sites. Red = bidirectional pairs (BiP); blue = proximal pairs. Lines of best fitted plotted using Lowess smoothing. BiP line was weakly but significantly negatively correlated using Spearman's rank ($\rho = -0.15$, p -value < 0.001).

Because between the two proximal TSS lies a gene, there are no proximal pairs with a distance of less than 140 bp. Besides this caveat both pairs can have any distance and any co-expression coefficient, however, broad trends are visible in the data. Bidirectional pairs have a weakly negative correlation ($\rho = -0.15$, p -value < 0.001) between co-expression coefficient and distance, which agrees with the hypothesis that the longer the distance the fewer pairs are truly bidirectional. The fitted curve for proximal pairs was not linear, increasing until about 600 bp before gradually declining. The lines of best fit for proximal and bidirectional pairs cross at approximately 560 bp, and subsequently follow a similar trajectory.

2.2.2 Discussion

Bidirectionally arranged genes are more likely to be either co-expressed or anti-regulated compared to proximal gene pairs (Figure 10). Therefore, the co-expression of bidirectional genes is not a consequence of the fact that genes proximal to one another on the genome are more likely to be coexpressed. Neighbouring genes, represented by proximal genes in this study, are coexpressed with their proximal gene pair due to the complexity of genome regulation in eukaryotes which is regulated by a combination of transcription factors, chromatin accessibility, and by the temporal and spatial organisation of chromosomal loci (Janga *et al.*, 2008). For example, transcription factors preferentially bind genes on specific chromosomes and neighbouring genes, especially those on the same strand, and neighbouring genes were found to have similar histone modifications in *S. cerevisiae* (Janga *et al.*, 2008; Deng *et al.*, 2010). Bidirectionally arranged genes are coexpressed more frequently than can be explained by these factors. The emerging picture is that the spatial

proximity of genes is beneficial in coordinating their regulation, whether this is by proximity on the genome or by the arrangement of the genetic material in 3D space.

The difference in co-expression coefficients between bidirectional and proximal pairs is not due to many perfectly correlated bidirectional pairs but an increase of correlation coefficients between 0.2 - 0.7, and fewer with no correlation at all (Figure 10). This indicates that being under the control of the same promoter increases the relative coregulation but the ultimate strength of that coregulation is influenced by other factors that are modulating gene expression in the cell.

As well as being more likely to be co-expressed, bidirectional genes were more likely to be anti-regulated. These anti-regulated pairs were particularly striking as they were almost exclusively bidirectional, with only five proximal genes being significantly anti-regulated, and all had absolute co-expression coefficients of less than 0.16 (Figure 9). None were taken forwards as candidates to look at in more detail in this study because a stringent cutoff of 0.65 was used, whereas the most anti-regulated pair in wheat had an absolute co-expression coefficient of 0.4. In light of these results I propose a more relaxed cut off of 0.2 should be used in future studies that are interested in anti-regulated pairs. Anti-regulation may be beneficial for the regulation of genes that are required in opposite conditions, or when interaction between two genes is detrimental, even though each expressed alone is advantageous.

As the co-expression coefficients in this study are calculated from 36 studies with different varieties of wheat grown under various conditions, a broad picture of the expression pattern is represented by this coefficient. However, the data is not exhaustive to every condition wheat could possibly be grown in, so the resulting co-expression coefficient is a product of the experimental conditions used in the different studies. For example, bidirectional pair 453 is the fifth most highly co-expressed pair ($\rho = 0.97$; adjusted p-value < 0.05), but a single point highly expressed in both conditions is responsible for the strength of this correlation. Without it, the co-expression coefficient is reduced to only 0.56. This demonstrates that a single additional point, if highly expressed in both paired genes, can have a large impact on the correlation coefficient. This may occur when a gene is responsive to a specific condition, for example a gene with a role in resistance to powdery mildew may only be expressed when the plant is infected with powdery mildew. Although this represents a minority of points in the data set, it should be noted as a source of noise and considered when selecting candidate pairs for future study.

A further consideration is that the cut off of 1000 bp between TSS to be defined as bidirectional has unclear origins (Polson *et al.*, 2011). One-thousand base pairs is often reported to be the upper limit for promoter length, even though many promoters longer than 1000 bp have been identified (Le *et al.*, 2019). In *S. cerevisiae* approximately 5% of all promoters are longer than 2000 bp, in *S. pombe* the average length of promoters of core environmental stress response genes is 1243 bp (Kristiansson *et al.*, 2009). However, such figures are not available for wheat. On the other hand, it's not clear if bidirectional promoters have the same distribution of lengths as unidirectional promoters. Bidirectional genes are reported in the literature as having housekeeping functions, and constitutively expressed housekeeping genes tend to have shorter promoters than those that regulate stress-response genes that require fine-tuned regulation in response to environmental

factors. (Kristiansson *et al.*, 2009). Meanwhile, bidirectional promoters that encode a gene in one direction and regulatory RNA in the other tend to be much shorter than 1000 bp in length, so evidence suggests that bidirectional promoters may be shorter than average (Julia Qüesta, personal communication). Evidence in humans demonstrated that the average distance between TSS of two genes that are next to one another on the genome has a bimodal distribution with the largest peak centering at just below 100,000 bp, but another smaller peak between 100-1000 bp, possibly representing bidirectional pairs (Trinklein *et al.*, 2004). In addition to this diversity, plant genome architecture is different to that of animals, and the core promoter architecture is known to differ between plants and mammals, so bidirectional promoters may differ in length between different kingdoms of life (Yamamoto *et al.*, 2007). Therefore, the 1000 bp cut off is an educated guess on how long bidirectional promoters are, and one that derives primarily from evidence in animals and yeast.

Because of this I was interested in confirming the somewhat arbitrary cut off of 1000 bp. I compared the distance between TSS for proximal and bidirectional pairs to their correlation coefficient. The lines of best fit mapping the correlation between the co-expression coefficients and distance between TSS for proximal and bidirectional pairs cross at approximately 560 bp, and subsequently follow a similar trajectory (Figure 11). This indicates that after 560 base pairs, the co-expression coefficient can be explained by the proximity of the genes on the genome, rather than any impact of a bidirectional promoter. This differs from the human genome where no correlation was found between the distance between TSS and the degree of co-expression (Trinklein *et al.*, 2004). Therefore, I suggest that the maximum length of bidirectional genes in plants is 560 rather than 1000 bp.

2.3 Temperature dependent response of bidirectional gene pairs

In this section I aimed to identify candidate bidirectional gene pairs that were likely to be coregulated and have differential expression at different temperatures.

2.3.0 Methods

Based on five sources of information, which consisted of four RNAseq studies and GO terms, candidate bidirectional genes with functions relating to temperature and drought conditions were selected (Li *et al.*, 2015; Liu *et al.*, 2015; Seifert *et al.*, 2016). In each of the RNAseq studies, a list of genes that were differentially expressed at different temperatures was produced, and could be taken forwards in this study. For the GO terms, a list of genes was produced as follows.

First, a list of temperature related GO terms was compiled in a spreadsheet by searching for the key terms “temperature”, “cold”, and “heat” on QuickGO (<https://www.ebi.ac.uk/QuickGO/annotations>) and copying in the resulting GO terms and descriptions (Supplementary 6A). Fifteen GO terms relating to the key term “temperature” were identified, 11 for “cold”, and 38 for “heat”.

For all bidirectional *T. aestivum* (IWGSC) genes as listed in Supplementary 1C, a list of their associated GO terms were obtained from EnsemblPlants release 57 biomart function by selecting the following attributes: Gene ID, GO term name, definition, accession, evidence code and domain. The results were saved as a csv file, which was then imported into R

along with the 64 temperature related GO terms (Supplementary 3A). A column of logical values was added to the bidirectional GO term data frame indicating whether the GO term in that row was one of the 64 temperature related GO terms, only those that were true were retained. The resulting data frame contained 9 genes, all unique (Supplementary 6C). Four had the term “response to heat” (GO:0009408), three had the term “response to cold” (GO:0009409), and two had the term “protein refolding” (GO:0042026). The protein refolding term is described as “the process carried out by a cell that restores the biological activity of an unfolded or misfolded protein, using helper proteins such as chaperones”. It’s related to heat stress as proteins begin to denature at high temperatures, and thus require intervention to remain correctly folded and functional. The evidence code for these GO terms was IEA, meaning they were inferred from electronic annotation, so the GO term is based on homology and sequence information.

The conditions of the RNAseq studies are described below, as well as the process of updating the gene ID names to the most recent genome annotation. Existing RNAseq data from an unpublished study within the lab group that grew *T. aestivum* cv. Buster for 4 weeks at 10°C and 14°C was used. The data included 3938 genes whose transcripts, taken from leaf tissue, were differentially expressed when grown at 10°C and 14°C. A list of differentially expressed genes between each condition was provided.

Three studies from the wheat expression browser, shortened to the reference names: cold, drought-heat, and microspore, were used as sources of differential expression information (Li *et al.*, 2015; Liu *et al.*, 2015; Seifert *et al.*, 2016). The differentially expressed genes were taken from the supplementary information from the respective papers.

The cold experiment grew *T. aestivum* cultivar Manitou, a Canadian spring wheat recommended for growth in north-west Canada (Li *et al.*, 2015). It was grown in growth cabinets at a temperature of 23°C for two weeks (Li *et al.*, 2015). The control plants were allowed to continue growing at 23°C, the cold stressed plants were moved to a growth cabinet set to 4°C (Li *et al.*, 2015). After two weeks leaf samples were taken and expression determined through qPCR (Li *et al.*, 2015). Genes and their log fold change in expression between the two temperature conditions were listed in Supplemental Data Set 7. Those with an adjusted p-value (q-value) of less than 0.05 were taken forwards. A final complication was that this study had been carried out in 2015, before the wheat genome was fully sequenced and before the current standard wheat genome annotation, refseq 1.1, was released. Although changes have been made to the annotation since 2015, the genes can be mapped to the modern annotation using EnsemblPlants’ ID history converter tool. I used this tool on the significant differentially expressed genes from the cold study (Supplementary 7A).

The drought heat experiment grew *T. aestivum* cultivar TAM_107, a wheat winter grown in central and southern USA (Liu *et al.*, 2015). It is known to be drought and temperature tolerant and was a leading variety in the 1980s (Liu *et al.*, 2015). It was grown in a 16h/8h light/dark photoperiod, with a day/night temperature of 22°C/18°C (Liu *et al.*, 2015). After 7 days they were subjected to one of four conditions: no change (control), heat stress, drought stress, or heat and drought stress (Liu *et al.*, 2015). Heat stress was achieved by moving the plants to a growth cabinet set to 40°C, drought stress was applied using 20% (m/V) PEG-6000 (Liu *et al.*, 2015). One hour after applying heat stress and 6 hours after applying

drought stress leaf samples were taken RNAseq was carried out (Liu *et al.*, 2015). A list of differentially expressed triplets was listed in additional file 3: Table S2, although no distinction was made about which of the three condition/s the gene was differentially expressed in, nor was whether it was up or down regulated (Liu *et al.*, 2015). In total, 4566 triplets were listed as being differentially expressed. The history converter tool was used for differentially expressed genes in the drought heat study, one subgenome at a time, the results of which were combined into one spreadsheet (Supplementary 7B).

The microspore experiment used *T. aestivum* cultivar Svilena, a bulgarian winter wheat (Seifert *et al.*, 2016). The aim of this study was to use cold stress to induce reprogramming of immature pollen grains, to generate an embryo-like structure (Seifert *et al.*, 2016). Samples were taken at three developmental stages: when the pollen grain was untreated, directly after a 10 day cold treatment, and at first visible karyokinesis (Seifert *et al.*, 2016). Differential expression between the first and second stage (T1) and the second and third stage (T2) was determined *via* RNAseq (Seifert *et al.*, 2016). The up and down regulated transcripts at T1 and T2 were listed in additional file six in with arbitrary transcript names. Conversions from the arbitrary identifier to the IWGSC26 wheat genome gene IDs were provided in Additional File 1 of the Seifert *et al.*, 2016 paper. A custom python script was used to translate the arbitrary identifiers in the document containing the differentially expressed genes.

The generated script read in the conversion document and created a dictionary, a method of storing data in key value pairs where a value can be accessed by using a particular key. Each key must be unique but the values may repeat. In this case, the key was the arbitrary code and the value was the gene ID. Instances of the dictionary where the value was blank, i.e. there was no equivalent gene ID listed in the document were removed. The document containing the differentially expressed genes was then read in. These were listed in four columns: T1 upregulated, T1 downregulated, T2 upregulated, T2 downregulated. I looped over each row and added the first, second, third, and final cell to four different lists. I defined a function that would remove any blank values from the list using a while loop and applied this function to each of the four lists. Thus each column of the spreadsheet was converted to a list. Next, I converted each arbitrary reference into the gene ID using the previously created dictionary. I defined a function that would iterate over a given list and use each value as a key to access the dictionary. This would return the value associated with that key, i.e. return the gene ID, which I then appended to a new list. The function would return this new list. I applied this function to each of the arbitrary gene references, and now had four lists of gene IDs. I wrote these lists to a csv file where each list was a column.

Once the arbitrary references were converted to IWGSC26 gene IDs EnsemblPlants' history ID converter was used to convert each column of gene IDs to the refseq1.1 gene IDs, then compiled the four resulting files into one spreadsheet (Supplementary 7C).

Now I had a total of five temperature related gene lists: four of wheat genes differentially expressed in different temperature conditions, and one of bidirectional genes associated with temperature related GO terms. Any wheat gene could appear in any one of these five lists, and the more lists it appears in the stronger the evidence its function is temperature related. To consider bidirectional pairs of genes I multiplied the number of times one gene appeared in these lists (0-5) by the number of times its pair appeared (0-5) to generate a "count"

statistic that could take any value from 0-25. To calculate the count statistic and to integrate it with information about the bidirectional pairs I used R (RStudio 2020).

The co-expression data was imported from Supplementary 5B, and proximal pairs removed, resulting in a data frame with a bidirectional pair on each row. The temperature information was added in four columns, one with a list of the studies in which the forward gene was differentially expressed, one that counted how many studies in which the forward gene was differentially expressed, and then another two with the same information but for the reverse gene. Instead of listing the full title of the study each was given a label, where possible indicating the conditions in which differential expression occurred: unpublished = 10v14, cold = 4v23, drought heat = DH, and microspore = MS. I first created four empty columns: DE_forward, DE_reverse, F_Count, and R_Count. To populate them a function was defined that would update these columns for a given dataset. The function took the following arguments: the path where the dataset is located, the name of the column that contained the differentially expressed genes, and the label that should be used to ID that study. The dataset was imported as a data frame using the path argument, then the column name argument was used to get the index of the column with the list of differentially expressed genes, and was used in subsequent steps to select that column. I created a logical column that checked these gene IDs against a list of bidirectional gene IDs, and only kept rows that were true. For all forward genes in the bidirectional data frame that were also in the given differentially expressed gene column I updated the DE_Foward column by adding the existing contents of the column to the given label, separated by a space. I added one to the F_Count column. I repeated this for the reverse genes, and the function returns an updated version of the bidirectional data frame. I applied this function for each of the four differentially expressed data sets and also for the GO term data set. I then multiplied the F_Count and R_Count columns to create simply a Count column that would allow the bidirectional pairs to be ordered from most likely to be temperature regulated pairs to least likely. The F and R_Count columns were removed and this data frame was written to a file that can be found in Supplementary 7D.

2.3.1 Results

Candidate genes with high co-expression and differential expression at different temperatures were identified.

Table 6. Summary of Supplementary Table 7D. The count statistic is a measure of how likely the gene pairs are to be related to temperature. The frequency is how many bidirectional gene pairs had that count statistic. The scale goes from 0-25 but no gene pair had a score greater than 6.

Count	Frequency
0	712
1	120
2	119
3	32
4	51
5	0
6	16

All counts were in a range from 0-6, with the majority of the gene pairs not being related to temperature (count 0, 712 pairs). No pair had a count of greater than 6, and therefore the most promising pairs for temperature were the 16 with a count of six. Of these 16, some had a low co-expression coefficient, so were unlikely to be coregulated. In order to select candidates I narrowed this 16 down to five by applying a cut off of ≥ 0.65 for the co-expression coefficient.

Table 7. Summarises the likelihood of each bidirectional pair having a function related to temperature based on RNAseq and GO term data. Only those with a count of greater than 5 and a co-expression coefficient of greater than or equal to 0.65 have been shown. The full table is available in Supplementary 7D.

Pair ID	Forward	Reverse	Distance between TSS	Coexpression Coefficient	B DE H Forward ^a	DE Reverse ^a	Count ^b
Pair 189	TraesCS 2A02G45 5800	TraesCS 2A02G4 55700	562	0.70	0 MS 4v23 DH	4v23 DH	6
Pair 356	TraesCS 3A02G34 9900	TraesCS 3A02G3 49800	412	0.70	0 4v23 DH	MS 4v23 DH	6
Pair 407	TraesCS 3B02G38 2500	TraesCS 3B02G3 82400	426	0.67	0 4v23 DH	MS 4v23 DH	6
Pair 458	TraesCS 3D02G34 3900	TraesCS 3D02G3 43800	491	0.68	0 4v23 DH	MS 4v23 DH	6
Pair 584	TraesCS 4D02G06 7700	TraesCS 4D02G0 67600	231	0.65	0 MS 4v23 DH	4v23 DH	6

^aIndication of studies in which evidence for response to temperature was found, 4v23 = cold, MS = microspore study, DH = drought heat study (unpublished data; Li *et al.*, 2015; Liu *et al.*, 2015; Seifert *et al.*, 2016); ^bNumber of studies in DE Forwards multiplied by the number of studies in DE Reverse.

All five pairs were differentially regulated between 4°C and 23°C and in heat/drought stress compared to control (Table 7). In addition one gene from each pair was also differentially expressed in the microspore study (Table 7).

As my aim in a future experiment was to measure the expression of candidate pairs in leaf and coleoptile tissue at different temperatures, I confirmed the suitability of the candidates by examining if these pairs of genes were expressed in the leaf and coleoptile at the seedling stage using existing RNAseq data from the wheat expression browser (Ramirez-Gonzalez *et al.*, 2018). Two studies measured expression throughout development in non-stressed conditions, these were carried out in the varieties Azhurnaya and Chinese Spring.

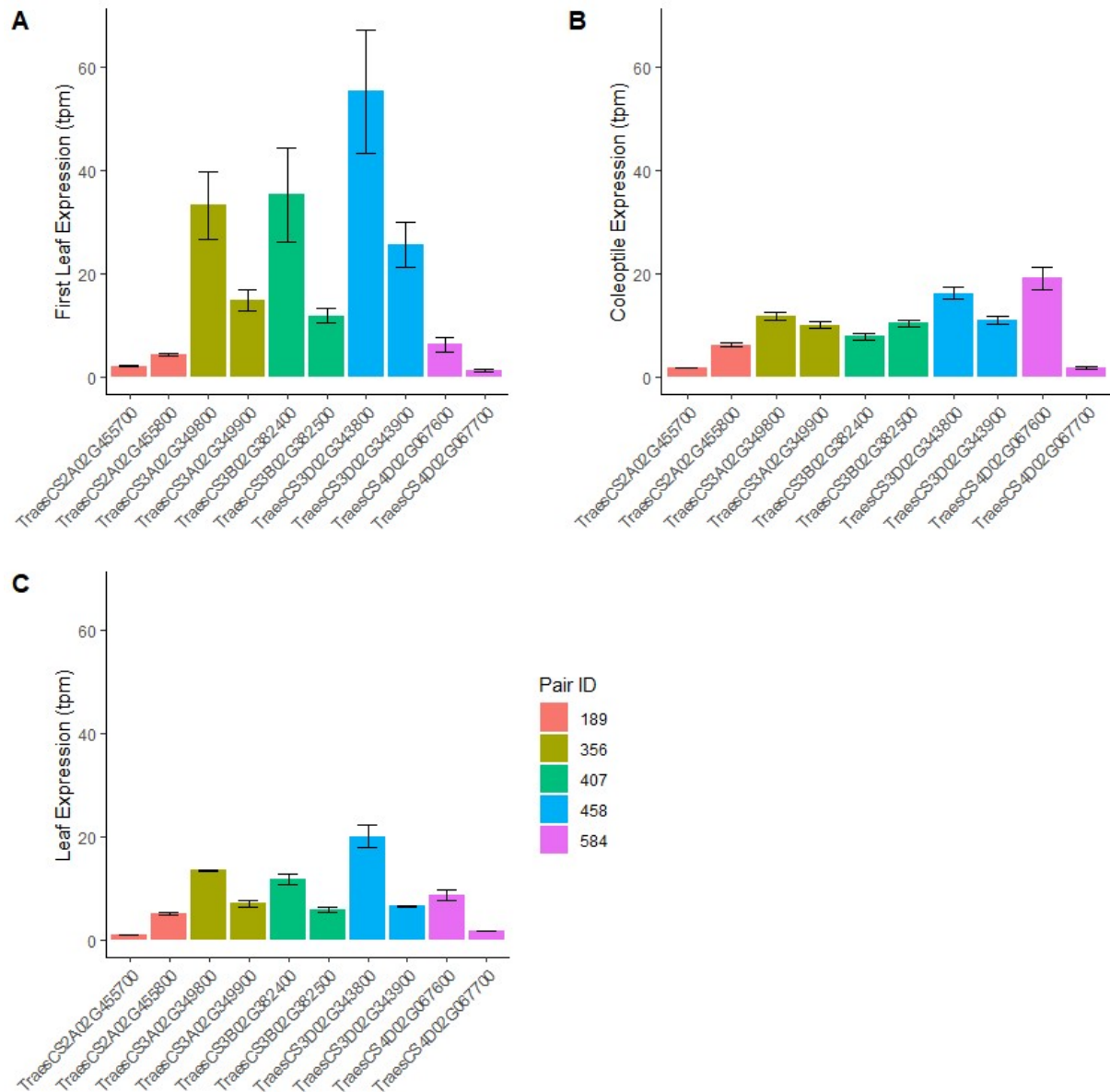


Figure 12. Expression in *Triticum aestivum* cultivar *Azhurnaya* in tpm in the first leaf (panel A) and the coleoptile (panel B), and in the cultivar *Chinese Spring* in the leaf at the seedling stage (panel C) for ten bidirectional genes. Bars represent the standard error.

The ten bidirectional genes (5 pairs) were expressed in both the coleoptile and the first leaf in both cultivars. In *Azhurnaya*, pair 458 was the most highly expressed and pair 189 the least in both tissues. The gene pairs 356, 407, and 458 were much more highly expressed in the *Azhurnaya* leaf than in the coleoptile or the *Chinese Spring* leaf. Pair 189 was consistent in all three cases, and pair 584 was most highly expressed in the coleoptile. The genes generally maintain their relative expression to one another across the two varieties and tissues, except for pair 407 in the coleoptile (Figure 12).

Having established that the genes were expressed in the coleoptile and leaf, and thus decided these were the final candidate genes, I determined if these genes had homoeologues and whether these were themselves bidirectionally arranged.

Table 8. The homoeologous arrangement of candidate genes. Candidate genes are highlighted in bold, each block of three in each column represents an homologous triplet, or doublet, when there is no homoeologous gene on one of the genomes. Where the homoeologous are also bidirectionally arranged, the pair ID is given in the first column, or NA if the homoeologous are unpaired.

Pair ID	Reverse	Forward
189	TraesCS2A02G455700	TraesCS2A02G455800
243	TraesCS2B02G477700	TraesCS2B02G477800
NA	none	TraesCS2D02G456100
NA	TraesCS4A02G246200	TraesCS4A02G246300
NA	TraesCS4B02G068500	TraesCS4B02G068600
584	TraesCS4D02G067600	TraesCS4D02G067700
356	TraesCS3A02G349800	TraesCS3A02G349900
407	TraesCS3B02G382400	TraesCS3B02G382500
458	TraesCS3D02G343800	TraesCS3D02G343900

The five pairs are composed of three homoeologous triplet bidirectional pairs, one pair which has no bidirectionally paired homoeologues, and one pair which has one bidirectionally arranged homoeologous pair (Table 8). This homoeologous pair had a correlation coefficient of 0.49 and a count of 6.

Next, I took a closer look at the differential expression in the drought heat, cold, and microspore studies to form an idea of what temperature conditions trigger their up or down regulation (Figure 13).

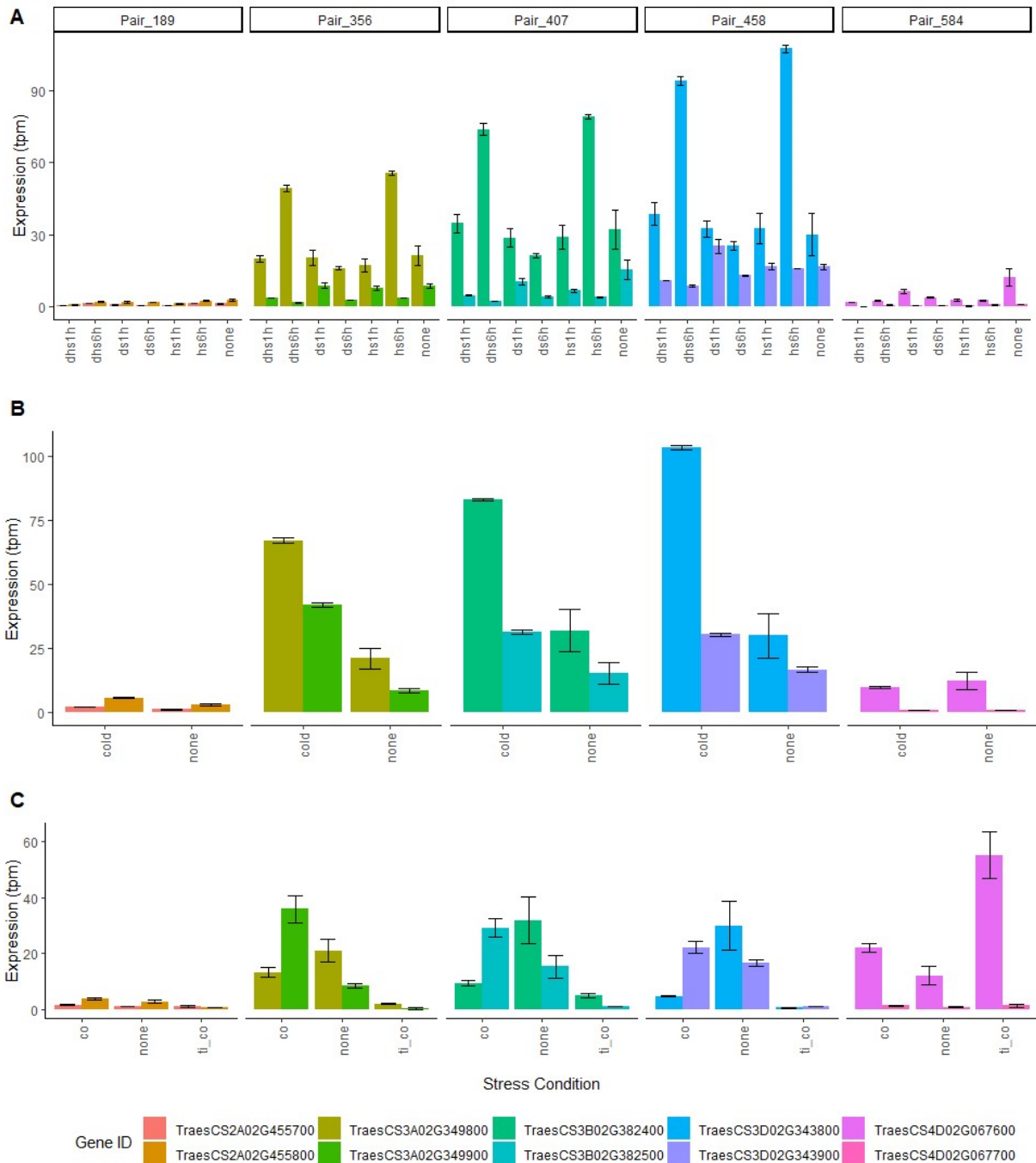


Figure 13. Expression of 10 bidirectional genes measured in tpm in three different studies under different temperature and drought stress conditions. The genes are grouped into bidirectional pairs (left to right: pair 189, 356, 407, 458, 584). Panel A. expression from the drought heat study. After 7 days wheat plants cv. TAM_107 were exposed to drought and/or heat and samples were taken after 1 and 6 hours. Dhs1h = drought and heat stress for 1 hour; dhs6h = drought and heat stress for 6 hours; ds1h = drought stress for 1 hour; ds6h = drought stress for 6 hours; hs1h = heat stress for 1 hour; hs6h = heat stress for 6 hours (Liu et al., 2015). Panel B. After 2 weeks at 23°C wheat plants cv. Manitou were subjected to 4°C temperatures for 2 weeks and samples taken. Cold = plants subjected to 4°C; none = plants that remained at 23°C (Li et al., 2015). Panel C. Immature pollen of wheat cv. Svilena was exposed to a 10 day cold treatment and samples taken directly after and at first visible

karyokinesis. None = before cold treatment; co = after cold pretreatment; co_ti = at first visible karyokinesis (Seifert et al., 2016). Bars represent the standard error.

Pair 189 had relatively little expression in any of these conditions, but were slightly more expressed in cold stressed conditions than the control (Figure 13).

TraesCS3A02G349800 in pair 356 was strongly upregulated in response to heat and drought heat stress after six hours, but demonstrated no change in response to drought stress alone. Its pair either remained the same or decreased for all heat/drought conditions, so the correlation in expression under these conditions is poor (Figure 13). Being homoeologous, pairs 407 and 458 had similar expression profiles. The relative expression of each gene in each condition remained constant, but pair 456 was more highly expressed than 407, which in turn was more highly expressed than pair 356. All three homoeologous pairs in the cold study were upregulated in response to cold. In the microspore study, all three homoeologous pairs had a reverse in which gene in the pair was most highly expressed after the cold treatment (Figure 13).

Gene TraesCS3D02G343800 in pair 458 had the most dramatic response, more than doubling in expression under heat stress after six hours; this was also seen in combined drought heat stress and cold stress, but expression was reduced when exposed only to drought (Figure 13). The expression in drought heat stress after 6 hours was slightly lower than in heat stress alone. Its pair had much more modest changes in expression, not changing in response to heat stress alone, reducing in response to combined drought heat stress and initially increasing before returning to its usual expression after six hours in drought stress alone.

In pair 584 one gene was expressed extremely minimally and so didn't have any notable differences between the conditions in any study, but TraesCS4D02G067600 was downregulated in response to both cold and heat stress conditions (Figure 13).

To gain a final insight into the function of these genes I used BLASTn to search for the best hit by sequence homology (Table 9).

Table 9. BLASTn results of candidate bidirectional pair sequences. The best hit with an informative description is listed with the pair ID, gene ID, query cover, e-value, and percentage identity.

Pair ID	Gene ID	Query Cover (%)	E-Value	Percentage ID	Description
189	TraesCS2A 02G455800	49	0.0	99.50	PREDICTED: Triticum urartu protein MULTIPLE CHLOROPLAST DIVISION SITE 1 (LOC125539686), mRNA
189	TraesCS2A 02G455700	NA	NA	NA	[only matches to uncharacterised genes]
356	TraesCS3A 02G349900	47	0.0	100.00	PREDICTED: Triticum aestivum internal alternative NAD(P)H-ubiquinone oxidoreductase A1, mitochondrial-like (LOC123062482), mRNA
356	TraesCS3A 02G349800	45	0.0	100.00	PREDICTED: Triticum aestivum UPF0051 protein slr0074-like (LOC123062484), mRNA
407	TraesCS3B 02G382500	47	0.0	100.00	PREDICTED: Triticum aestivum internal alternative NAD(P)H-ubiquinone oxidoreductase A1, mitochondrial-like (LOC123071289), mRNA
407	TraesCS3B 02G382400	48	0.0	100.00	PREDICTED: Triticum aestivum UPF0051 protein slr0074-like (LOC123071288), mRNA
458	TraesCS3D 02G343900	36	0.0	96.88	NAD(P)H-ubiquinone oxidoreductase A1, mitochondrial-like (LOC123071289), mRNA
458	TraesCS3D 02G343800	39	0.0	100.00	PREDICTED: Triticum aestivum UPF0051 protein slr0074-like (LOC123079630), mRNA
584	TraesCS4D 02G067700	50	0.0	95.57	PREDICTED: Triticum aestivum telomerase Cajal body protein 1-like (LOC123086930), transcript variant X1, mRNA
584	TraesCS4D 02G067600	NA	NA	NA	[only matches to uncharacterised genes]

The percentage IDs represent the similarity between the sequences which was perfect or near perfect for most hits. The query cover, the percentage of the query sequence that

matches the reference sequence, ranged from 36 to 50%, so the hits are not likely to be the same gene as its hit, but may have shared domains.

2.3.2 Discussion

Bidirectional pairs can be temperature dependent (Table 6; Figure 13). This is despite the fact that many are housekeeping in function so have a theoretically stable expression in all conditions. One gene in pair 189 matched to a chloroplast division site protein, which is essential in plants as chloroplasts are the site of photosynthesis without which the plant cannot survive. It therefore meets one of the four criteria to be a housekeeping gene (Joshi *et al.*, 2022) and yet is differentially regulated between 4°C and 23°C, as well as under drought/heat stress and in the microspore study. However, it's unsurprising that genes encoding proteins essential for chloroplast division are not expressed uniformly across time and tissues. Evidently, photosynthesis cannot occur in root tissues due to the lack of light, and in wheat once senescence has occurred chloroplasts are disassembled as resources are translocated from the leaves to the grain, so chloroplast division genes would not be expressed in the root tissues or during senescence (Gregersen *et al.*, 2013). This demonstrates essential genes are not necessarily constitutively expressed and that the classification of housekeeping is still not strictly defined, as acknowledged in a 2022 study attempting to more rigorously define housekeeping genes (Joshi *et al.*, 2022). Therefore, although genes may be housekeeping in the sense that they are essential, it doesn't necessitate that they do not respond to temperature. In fact, temperature, along with photoperiod, are two of the most important environmental signals that plants respond to in order to regulate developmental processes such as senescence (Gregersen *et al.*, 2013; Gustavo 1996).

Response to temperature is vitally important in plants for both normal development and surviving extreme temperature stress, which impacts plants more immediately than other stresses such as nutrient or water stress (Nievola *et al.*, 2017). Photosystem II is particularly sensitive to heat stress, because the increase in membrane fluidity disrupts the light harvesting proteins inserted in the membrane, and because the electron transport chain is disrupted such that electrons backflow to photosystem II (Mathur *et al.*, 2014; Yamauchi *et al.*, 2011). Wheat is no exception, experiencing reduced photosystem II efficiency on hot days (Ristic *et al.*, 2007). To examine the physiological effects of high temperature on the photosynthetic apparatus heat sensitive wheat cultivar Karl 92 and heat tolerant cultivar Ventor were first grown in 25/15°C light/dark conditions, then at the onset of flowering the temperature was increased to 35/24°C light/dark for 12 days (Narayanan *et al.*, 2016). This resulted in reduced seed weight and reduced grain yield per plant, although this was less pronounced in the heat tolerant cultivar (Narayanan *et al.*, 2016). This was accompanied by physiological changes; both cultivars had damaged thylakoid membranes but only the heat sensitive cultivar had no increase in reactive oxygen and nitrogen species, indicating that wheat can evolve adaptations to partially overcome the heat sensitivity of photosystem II (Narayanan *et al.*, 2016). Such defences are activated only when temperatures are high, as the trade off for greater heat tolerance is usually a reduced yield and/or growth (Nievola *et al.*, 2017). They are activated by modifications of the transcriptome, proteome, and metabolome (Nievola *et al.*, 2017). Therefore a fast response to heat is essential, at all levels including that of transcription, where the bidirectional mechanism may play a role in coordinating gene pairs in response to temperature.

The three homoeologous pairs identified in my analysis (Table 8) matched to the same proteins, NAD(P)H-ubiquinone oxidoreductase A1 and predicted *T. aestivum* UPF0051 protein slr0074-like. NADH ubiquinone oxidoreductase is a large multi-subunit enzyme involved in the citric acid cycle and is required for aerobic respiration. It is inserted in the inner mitochondrial membrane and pumps protons into the intermembrane space in order to create a proton concentration gradient and drive ATP synthase (Subrahmanian *et al.*, 2016).

It's unclear where the UPF0051 protein slr0074-like protein name was inferred from originally, but the only exact match on uniprot is to *Drosophila kikkawai*. According to panther (<https://pantherdb.org/tools/compareToRefList.jsp>) it contains two domains, iron-sulfer cluster assembly (PTHR30508) and protein ABCI8 (PTHR30508), chloroplastic-related. Searching the panther database for those two domains, and looking for *T. aestivum* genes assigned to those domains, leads back to the three homoeologous genes identified in Table 8. ABCI8 (Os05g0400600) is a chloroplast ABC transporter involved in iron-sulfur cluster assembly in rice, which when mutated results in accumulation of iron, so may be involved in iron transport, agreeing with the other panther domain iron-sulfur cluster assembly (Zeng *et al.*, 2017). Other ABC transporters, such as ATM3/ABCB23 (Os06g0128300) are known to be essential in iron-sulfur cluster assembly in rice, and losing the encoding gene is lethal (Zuo *et al.*, 2017). However, neither of these rice ABC transporters are orthologous with the three wheat genes in question. They are orthologous with the uncharacterised Os01g0830000 from rice and both AT5G44316 (ABCI9) and AT4G04770 (ABCI8) from thale cress that encode a “stabiliser of iron transporter SufD superfamily protein” and “ATP binding cassette protein 1” according to EnsemblPlants. Iron-sulfur clusters are cofactors of proteins that are important to photosynthesis and respiration, including to ubiquinone oxidoreductase, which has 8 iron-sulfur clusters (Subrahmanian *et al.*, 2016). I therefore hypothesised that wheat gene TraesCS3A02G349800 and its homoeologues encode an ABC transporter that localises to the mitochondria, rather than the chloroplast, where it is involved in iron transport for the formation of iron-sulfur clusters required for ubiquinone oxidoreductase.

TraesCS4D02G067700 matched to “telomerase Cajal body protein 1-like” while its pair only matched to uncharacterised genes. Cajal bodies are a non-membrane bound organelle involved in myriad processes, but are essential to only mice and zebrafish embryogenesis so are thought to increase the speed of processes that occur inside the cell (Love *et al.*, 2017).

Chapter 3 - Function and Mechanism of Bidirectionally Arranged Genes

3.0 GO term analysis, enriched promoter motifs, CpG islands, and pair alignment

I aimed to determine if bidirectional genes are more likely to have functions relating to cellular maintenance, if bidirectional gene pairs are more likely to have related functions than proximal genes pairs, to identify enriched motifs that may bind transcription factors that confer bidirectional transcription, and to calculate the similarity between the sequences of bidirectional pairs to determine if they arose of gene duplication events.

3.0.0 Methods

3.0.0.0 GO Term Analysis

As described in Chapter 2.3 GO terms associated with BiP and proximal genes were obtained from EnsemblPlants (Supplementary 3A). This was repeated for proximal pairs, which were used as a control (Supplementary 3C). RStudio was used to determine how many bidirectional pairs had GO terms associated with both genes in the pair, and what fraction of those had at least one overlapping GO term (RStudio 2020).

The GO terms were imported from Supplementary 3A and 3C, the evidence code column was removed. A list of bidirectional and proximal genes were imported with their corresponding pair IDs in order to merge it with the GO term data frames, so they now both have a column with the pair ID of each gene (Supplementary 1D). Duplicate rows were removed, as were genes with no GO annotation. A custom function was used in combination with the `sapply` function to filter out genes that were no longer in pairs, so if one gene had GO annotations and its pair didn't, it was removed. A pair of empty data frames, one for bidirectional and one for proximal pairs was created, with the column headings "overlap", to be marked true or false, and "pair ID". A pair of custom functions was written that subsetted each data frame by the pair ID given as an argument. The `table` function was then applied to the column containing the GO terms, which counts the instance of each element. This information was stored as a dataframe. A column with the pair ID was added, and an overlap column was added if the element appeared twice, once for each gene. This was appended to the relevant data frame with the pair ID. These functions were used in combination with the base R `sapply` function to apply the function to every bidirectional and proximal pair with GO terms.

The overlapping GO term information was then integrated with the co-expression data (Supplementary 5B). The co-expression data was imported and merged by pair ID with the overlapping GO term information. Only significant co-expression coefficients (adjusted p-value < 0.05).

GO enrichment processes for all wheat bidirectionally arranged genes was carried out using Gene Ontology Resource (<http://geneontology.org/>; Thomas *et al.*, 2022). Bonferroni correction for multiple testing was used.

3.0.0.1 Pair Alignment

The gene sequences for all proximal and bidirectional genes were downloaded from EnsemblPlants using the `biomart` function to search the IWGSC database. Bidirectional forward and reverse genes and proximal genes 1 and 2 as listed in Supplementary 5B were used as external reference IDs in the gene filter section, and the selected attributes were Gene ID and the unspliced gene sequence.

Command line blast was used to align the genomic sequences by writing linux shell scripts and submitting them to ARC4, part of the high performance computing facilities at the University of Leeds, UK (Supplementary 9B). Only the top hit is returned, or nothing if there is no close match. The output files can be found in Supplementary 9C.

Isolating Bidirectional Promoter Sequences

The sequences 1000 bp upstream of the transcription start site were downloaded from EnsemblPlants biomaRt for all forward orientated bidirectional genes. Forward genes referring to those with an orientation of 1 on EnsemblPlants. The wheat database IWGSC was selected, forwards bidirectional genes as listed in Supplementary 5B were used as external reference IDs in the gene filter section. The selected attributes were flank-coding region, upstream flank, and 1000 was entered into the input box. The resulting data was downloaded as a text file. The same was done for the bidirectional reverse gene IDs, i.e. those with an orientation of -1 on Ensembl.

A custom python 3.11 script was then used to cut these sequences so only the sequence between the transcription start site of the forwards and the reverse genes was retained, i.e. the bidirectional promoter (Supplementary 8B; Van Rossum and Drake 2009). The text files were looped through line by line, if a line in the text file started with a ">" character the line was saved as a variable, and the following lines until the next ">" character had their newline characters stripped and the lines were combined so the sequence was saved as one unbroken string. The variable starting with the ">" character had this stripped, leaving only the gene name which was used as a key and the sequence used as a value. In this way two dictionaries of key value pairs containing the gene name and the sequence respectively were created, one for the forwards and one for the reverse genes.

The csv module was used to import the co-expression data from Supplementary 5B, the data from which was used to create two dictionaries, one containing the pair IDs as keys and the distances (between the TSS) as values, and another containing the pair IDs as keys and the two gene IDs for that pair as a list as values. At this stage the following information has been imported: 1000 bp upstream sequence of all BIP genes, pair IDs, key values, and distances between TSS for all pairs.

A custom function was defined that returned the promoter sequence for a given bidirectional gene of either orientation. The function took the gene ID as an argument. From this the pair ID was obtained using one of the previously populated dictionaries, and then the pair ID used to obtain the distance from another dictionary in turn. For forwards bidirectional genes the sequence was obtained and characters from 1000 minus the distance to the 1000th character were cut out, and for the reverse genes the sequence was first reversed, then characters from the start to the value for distance were cut out. For the distinction between forwards and reverse bidirectional genes, see Figure 2. The promoter sequence was then returned from the function and saved as a value to a dictionary that can be called by the pair ID. The function was carried out for all bidirectional genes.

Because the forward and reverse genes in a pair share a promoter, and the reverse gene upstream sequence was reversed as part of the process, the resulting promoter sequences from each one will give the same promoter sequence in the same direction but a different strand. So the two sequences will be complementary to one another. A spot check was done to confirm the script was behaving as intended.

The forward and reverse promoter sequences were written to a file in fasta format. The title line contained the pair ID and two genes IDs separated by a space. The files can be found in Supplementary 8D.

3.0.0.2 Isolating the Intermediate Sequence Between Proximal Genes

Proximal genes were defined for the purpose of this study as two genes on the same strand with transcription start sites no more than 1000 bp away from one another. Each can therefore be assumed to be under the control of its own promoter. The sequence between the end of one gene and the transcription start site of the next was used as a control. The gene with the 3' end facing the intermediate sequence was referred to as the "tail" gene, and the gene with the TSS end facing the intermediate sequence was referred to as the "head" gene. The intermediate sequence therefore will contain the promoter of the head gene, the promoter of the tail gene was not considered.

First a python script was used to define each proximal gene as either head or tail, converting it from the way of referring to the two genes as gene 1 and 2 where gene 1 is the one with the smaller TSS and appears on the left hand side when viewed in Ensembl. As some genes are present in two proximal pairs, they will be both a head and a tail gene depending on which pair you're considering.

The csv module was used to import a list of proximal genes listed pairwise from Supplementary 5B. The information was stored in a series of dictionaries accessed using the pair or gene ID, and two other dictionaries were created to access the pair ID from the gene ID and vice versa. I multiplied the TSS of each gene by its orientation, so forward genes (strand = 1) would not change and the reverse genes would become negative. Then, if the calculated value was smaller for the first gene than the second, then gene 1 was the tail gene and gene 2 was the head gene, or if the calculated value was greater for the first gene than the second vice versa.

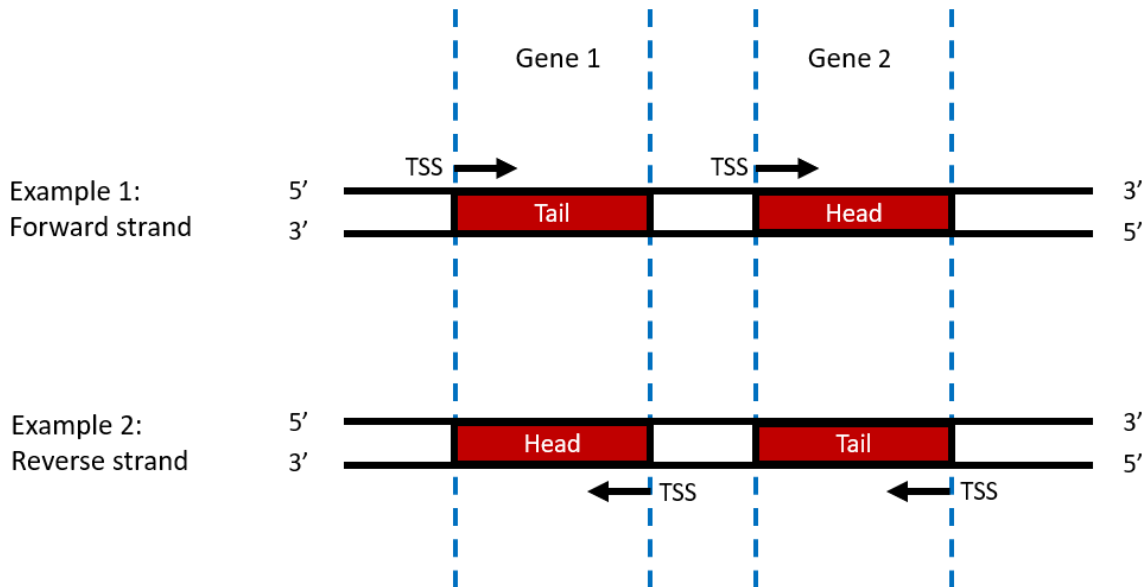


Figure 14. Demonstrates the two possible arrangements of proximal gene pairs, with genes being represented in red. Either the two genes are on the forward strand, as in example 1, or the two genes are on the reverse strand as in example 2. The black arrows represent the direction of transcription from the transcription start site. The blue dashed lines delineate gene 1 and gene 2. The new classification, with “head” being the gene with the TSS facing the intermediate sequence between the genes and “tail” being the gene with the 3’ end of the gene facing the intermediate gene sequence, are labelled directly on the gene in white. Note that the head and tail are reversed in example two when the gene is on the reverse strand.

A second python script used the head tail classification to cut the intermediate sequence from between the proximal pairs. The relevant modules fasta, pandas, and math were imported. The fasta file containing the proximal gene sequences was imported using pandas, a data handling module in python that allows data to be imported as a data frame. I subset the data frame so I had one containing the head genes and one containing the tail genes. The gene length was calculated for the tail genes by creating a custom function that subtracts the gene start column from the gene end column and saves the information to a dictionary. The function was applied to the dataset using the apply function with lambda function to apply it to each row. A pair distance dictionary was created in the same way. The intermediate sequence length was calculated for all tail genes using the distance between TSS and tail gene length just calculated. A custom function was written to take the tail gene length from the distance between TSS and save the information to a dictionary. It was applied to every row in the data frame using a lambda function with the apply function. A head to pair ID and a pair ID to sequence dictionary were created, again using a custom function with apply and lambda. Then a for loop was used for every pair ID and using the created dictionaries the head sequences were cut between character 1000 minus the length of the intermediate sequence to the 1000th character. The resulting promoter sequences that were longer than 100 bp in length were written to a fasta file (Supplementary 8E).

3.0.0.3 Enriched Motifs

Mememotif STREME was used to identify enriched motifs in proximal and bidirectional promoters (<https://meme-suite.org/meme/>). The forwards promoter sequence in Supplementary 8D was used as an input for the bidirectional set of promoters and the head promoter sequence in Supplementary 8E was used for the proximal set. Motifs between 5 and 15 bp were specified and the right aligned option was selected, all other options were as default. The output was downloaded in html format (Supplementary 8A). The TF/TFBS search tool of Plantpan 3.0 was used to identify if the motif belonged to a particular transcription factor family and the motif ID if one existed (<http://plantpan.itps.ncku.edu.tw/plantpan3/index.html>; Chow *et al.*, 2019).

3.0.0.4 CG Content Calculation

Bidirectional promoter sequences were imported into R from Supplementary 8D and proximal promoter sequences from Supplementary 8E. R's CG function from the seqinr package was used to calculate the CG content of proximal and bidirectional promoters (RStudio Team 2020; Charif and Lobry 2007). Unknown bases (N) were excluded from the analysis. The C to G ratio was calculated by dividing the number of C bases by the number of G bases.

3.0.0.5 CpG Island Calculation

R with the packages seqinr, stringr, and tidyr packages were used (RStudio Team 2020; Charif and Lobry 2007; Wickham 2022; Wickham and Girlich 2022). Bidirectional promoter sequences were imported into R from Supplementary 8D and proximal promoter sequences from Supplementary 8E as a list of SeqFastadna sequences. A data frame with the values required to calculate the observed to expected ratio of CpG islands for bidirectional and proximal promoters was created by using custom and existing functions to generate each column. A custom function countBase was defined to count the number of a particular base in a SeqFastadna sequence and return that number, it took the SeqFastadna list and the base to be counted as arguments. A second custom function countCpG was defined to calculate the number of CpG islands in a given SeqFastadna object. A CpG island was defined as a cytosine followed by a guanine base in the 5' to 3' direction. Then the data frame was created. The pair ID column was created by using the name function on the list of promoter sequences, the length by using the getLength function from the seqinr package, the C column was calculated by using the custom countBase function, as was the G and N column (N being undetermined bases), the CpG count column was created by using the custom function countCpG. Using these values the Expected/Observed ratio of CpG islands was calculated for proximal and bidirectional promoters using the following formula:

$$Obs/Exp \text{ CpG} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times N$$

If the fraction of unknown bases was greater than 50%, the obs/exp value was set to NA.

3.0.1 Results

3.0.1.0 GO Term Enrichment and Functional Analysis

The number of GO terms in the bidirectional subset was compared to the GO terms for the entire wheat genome, and if there were significantly more the GO term was considered to be enriched.

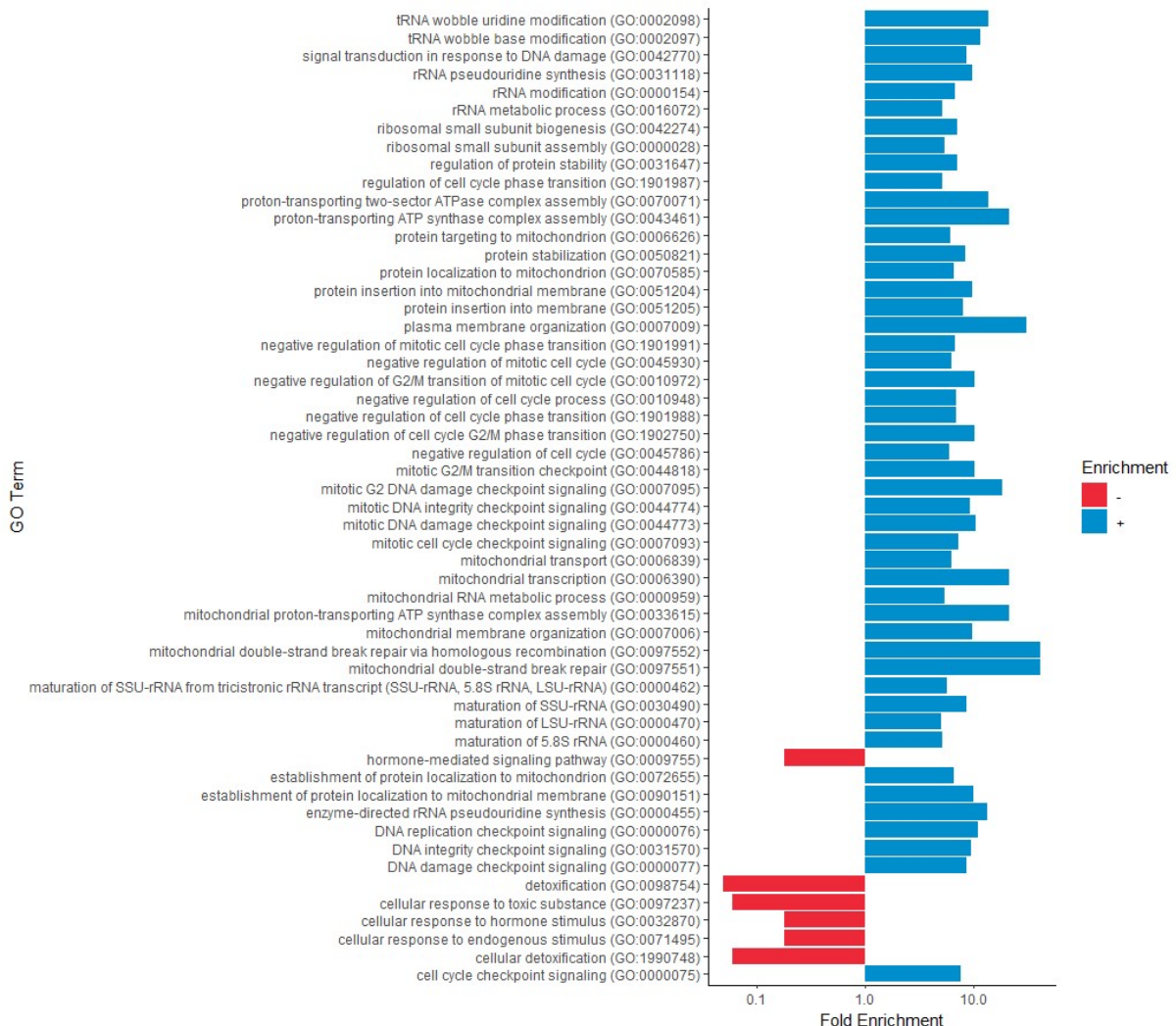


Figure 15. GO term enrichment analysis of the biological processes domain for bidirectional wheat genes, displayed on a log scale for clarity. Only those with more than 5 fold enrichment or depletion are shown. Red (-) = depletion; Blue (+) = enrichment. Fisher's exact test; adjusted p-value < 0.05.

The most highly enriched GO term with a 41.49 fold enrichment was “mitochondrial double-strand break repair via homologous recombination” and its parent term “mitochondrial double-strand break repair” (Figure 14). There are only five genes in wheat associated with this GO term, and four of them are bidirectionally arranged, compared to the expected 0.1 (Supplementary 3B). All four also have the associated GO term “mitotic G2 DNA damage checkpoint signaling”, which is also enriched. This implies these genes are crucial in delaying the proceeding of the cell cycle when there is excessive DNA damage.

None of these four genes are paired with each other - three comprise a homologous triad, and the other is located on chromosome seven and is homologous with the only unidirectional gene with this GO term. The pairs of these genes are uncharacterised, so it is unknown if they also have DNA repair functions.

The two most common GO terms that were more than five-fold enriched for bidirectional promoters were “rRNA metabolic process” and “ribosomal small subunit biogenesis” with 68 and 37 associated bidirectional genes as opposed to the expected 13 and 5, representing a 5.13 and 7.08 fold enrichment respectively (Supplementary 3B).

The three most depleted GO terms were “detoxification”, “cellular response to toxic substance”, and their parent term “cellular detoxification” (Figure 14). Only a single bidirectional gene, TraesCS4D02G103100, was associated with these GO terms, as opposed to an expected number of twenty-one, seventeen, and seventeen respectively (Supplementary 3B). There were three other GO terms depleted more than five-fold, having only four genes rather than the expected twenty-one (Supplementary 3B). All three of these GO terms were related to hormone signalling and response, and all three were associated with the same four genes. These four genes were not bidirectionally arranged with one another.

A GO term enrichment analysis gives an idea of what functions bidirectional genes are involved in more frequently than a random wheat gene, but it doesn't give any information about what pairs have overlapping functions. The functions of bidirectional gene pairs are thought to be related, hence the requirement for them to be coregulated. To investigate not only the function of bidirectional genes but whether bidirectionally paired genes have related functions, I calculated how many bidirectional pairs had at least one GO term each, then of those, I calculated how many shared at least one GO term from any of the three domains, cellular, molecular, or biological. If the gene pairs shared at least one GO term they were said to have a related function, unrelated if they didn't share at least one GO term, and undetermined if either pair did not have a GO term each to compare. This was also done for proximal pairs as a control.

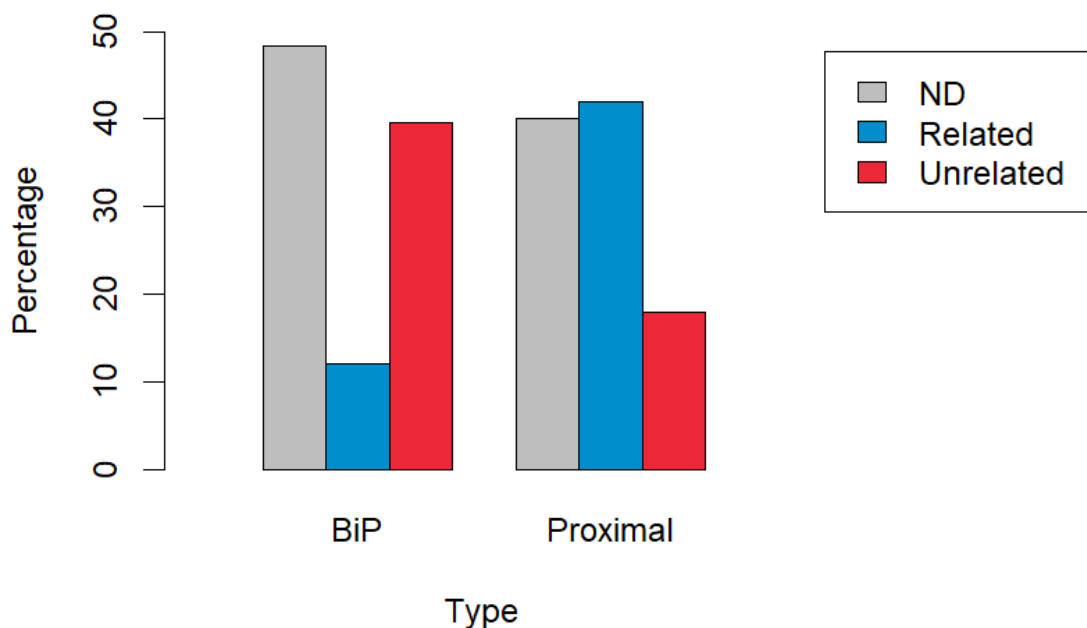


Figure 16. Percentage of bidirectional (BiP) and proximal pairs that have related functions, unrelated functions, and not enough data (ND).

Just over 50% (568) of the bidirectional pairs had at least one GO term associated with both genes in the pair. Of these 23% (132) had between one and nine GO terms in common (Figure 16). For comparison, in proximal pairs 60% (281) had at least one GO term associated with both genes in the pair (Figure 16). Of the pairs that had at least one GO term associated with them 70.1% (197) had between one and fifteen GO terms in common (Figure 16). Therefore, according to GO terms proximal gene pairs are more likely to have related functions than bidirectional gene pairs. This is unexpected as both proximal and bidirectional pairs are in close proximity and therefore more likely to have related functions, so the expectation is that they would be equal, or bidirectional pairs more likely to be related as they have an additional layer of coregulation.

3.0.1.1 Alignment of Bidirectional and Proximal Pairs

The proximal genes could more often have related functions because they arise from gene duplication more frequently than bidirectional genes. In addition, one of the ways in which both categories of genes may arise is through gene duplication. In order to investigate whether bidirectional and proximal gene pairs are likely to be formed as a result of gene duplication I used BLASTn to align the genomic sequences of each pair, assuming that highly similar sequences are likely to be a result of gene duplication.

Table 10. Bidirectional gene pairs that align with one another using BLASTn. A cut off of query cover >50% was used. The co-expression coefficient is a measure of co-expression and all included were significant with a p-value < 0.001.

Forward	Reverse	Percentage ID	Query Cover	E-Value	Pair ID	Coexpression Coefficient
TraesCS3B 02G358800	TraesCS3B 02G358700	100	100	0	403	0.909144
TraesCS3D 02G322300	TraesCS3D 02G322200	100	100	0	455	0.765104
TraesCS4B 02G334400	TraesCS4B 02G334300	95.457	70	0	570	0.767725
TraesCS3D 02G010900	TraesCS3D 02G010800	83.216	56	9.65E-14 8	421	0.995123
TraesCS7B 02G262400	TraesCS7B 02G262300	78.649	58	2.18E-73	994	0.842401
TraesCS2D 02G368000	TraesCS2D 02G367900	76.851	68	0	295	0.945972

Only six of 1050 bidirectional pairs had similar sequences at the genome level, in contrast to 22 of 469 proximal pairs, which is 0.6% for bidirectional pairs and 4.7% for proximal pairs. Two bidirectional pairs and ten proximal pairs had 100% ID and 100% query cover, indicating that the sequence of one gene is entirely contained within the other. The two bidirectional pairs with 100% ID and query cover are homoeologues of one another, with no third homoeologue present on the A genome for either gene. Of those with similar sequences only two of the proximal pairs had a co-expression coefficient of greater than or equal to 0.65, whereas all of the bidirectional genes did. There are so few paralogous genes that they won't make a large impact on the relative fraction of pairs with related and unrelated functions.

Following the assumptions made about bidirectional promoters, I have so far established that bidirectional pairs are more likely to be co-expressed (Figure 10) but less likely to have related functions compared to proximal pairs (Figure 16). Next, I investigated whether those with related functions were more likely to be co-expressed. To do this, I compared the number of GO terms in common for bidirectional and proximal pairs to the co-expression coefficient. 568 bidirectional pairs were analysed, or 54% and 281 proximal pairs were analysed, or 60%, as the remainder had fewer than one GO term per gene in the pair.

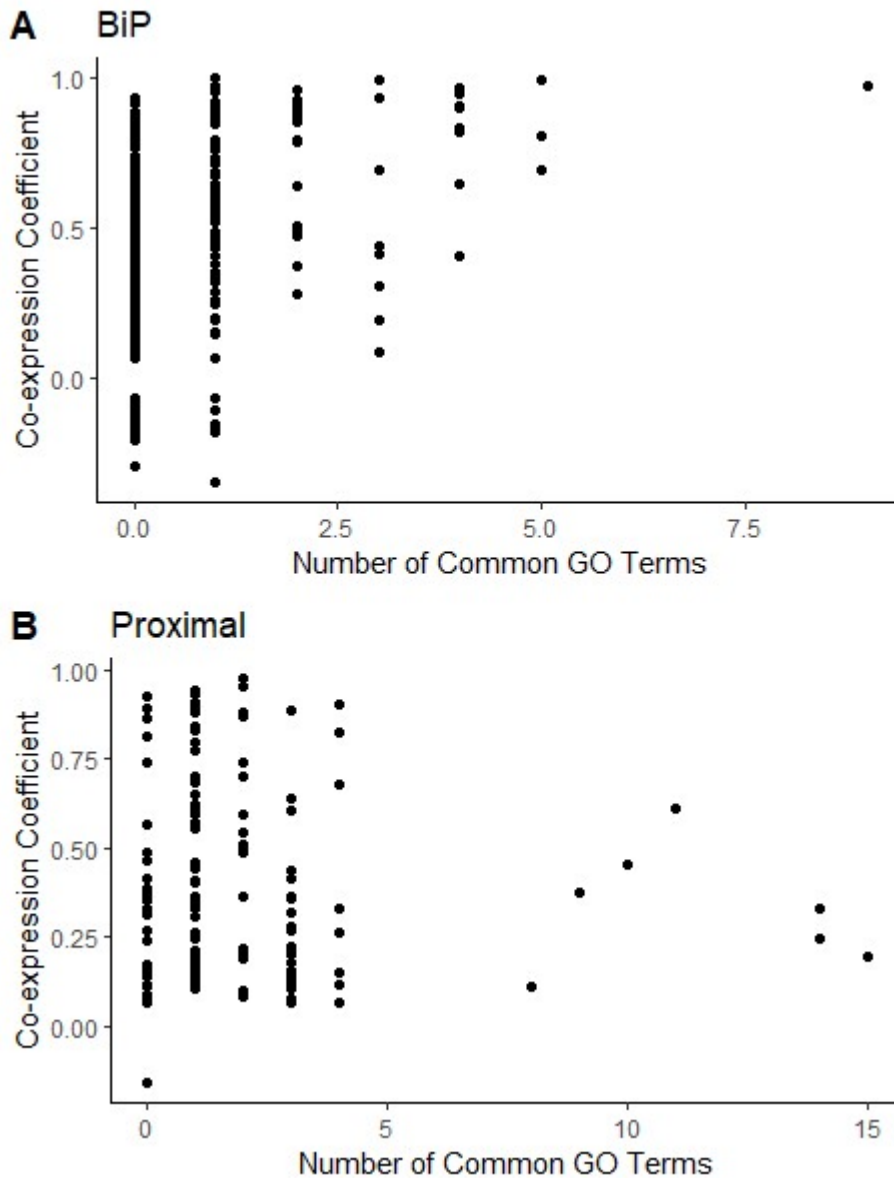


Figure 17. For all bidirectional (panel A) and proximal (panel B) gene pairs with GO terms for both genes, the number of common GO terms between the pairs is plotted against the co-expression coefficient (only co-expression coefficients with an adjusted p -value < 0.05 , insignificant co-expression coefficients were set to zero).

3.0.1.2 Promoter Motif Enrichment

The mechanism of bidirectional promoters is poorly understood, but it is known that transcription factor GABP is sufficient to induce bidirectional transcription in four of six cases, so I decided to look for common motifs in bidirectional promoters that could bind an analogous transcription factor in wheat (Collins *et al.*, 2007).

Table 11. A list of enriched sequences in bidirectional promoters as detected by memesuite, the top motif sequence ID and family as identified by plantpan 3.0 and the occurrence, i.e. how many bidirectional promoters the motif occurs in according to memesuite.

BIP - compared to shuffled sequence	Motif Sequence ID	Family	Occurrence (%)
GTCCA	none	none	56.1
CTTCTCC	none	none	34.9
AGGARGAGGAGR	none	none	33.6
CCCGGCCCA	none	none	30.2
AATGGGCCGCA	TF_motif_seq_0211	none	29.7
RTCGAY	none	none	27.8
AGCCCA	none	none	24.4
GTCGAC	TF_motif_seq_0322	none	23.7
GKATATATA	none	none	17.1
AAAAAAGAAAAAA	TFmatrixID_0274	MADSbox; MIKC	16
CGCCGCCGCCGCC S	TFmatrixID_0601	AP2; ERF	14.3
CAGGTGGGCCCGG B	TFmatrixID_0435	TCP	13.6
MAGCCCAAC	none	none	13.3
TCAAAAAAAAAAR	none	none	12.4
YTACTACTAC	none	none	11.5
AAACCCTAGC	TF_motif_seq_0402	none	11.4
CCGTCCGATCG	none	none	9.4
CTCTCTCTCTCY	TF_motif_seq_0165	none	8.7
AGCAGCARCA	none	none	8.1
AGTTACGTRWW	TFmatrixID_0381	NAC; NAM	5.1
AGAATCCACTTC	none	none	4.7

Twenty-one motifs were identified, four of which were known binding sites for transcription factor families. These transcription factor families were MADSbox; MIKC, AP2; ERF, TCP, and NAC; NAM. The occurrence of these transcription factors ranged from 4.7% to 56.1%. Only one sequence, GTCCA, occurred in more than half of all bidirectional promoters, much higher than the second most common sequence of CTTCTCC, which occurred in 34.9% of promoters (Table 11).

3.0.1.3 CpG Islands

The fact that bidirectional promoters tend to be located in regions that lack TATA boxes and instead be located in CpG islands is thought to be related to their mechanism (Ahmad *et al.*, 2020). CpG islands are defined as regions that are unusually dense in cytosine followed by guanine bases in the DNA backbone in the 5' to 3' direction (Plass and Rush 2005). I calculated the CG content and compared the observed to expected ratio of CpG islands in bidirectional and proximal promoters in order to determine if bidirectional promoters in wheat are also found predominantly in CpG islands (Figure 19; Figure 20). Regions of DNA with an observed to expected CpG island ratio of greater than 0.6 is considered to be CpG rich in the literature (Gardiner-Garden and Frommer 1987).

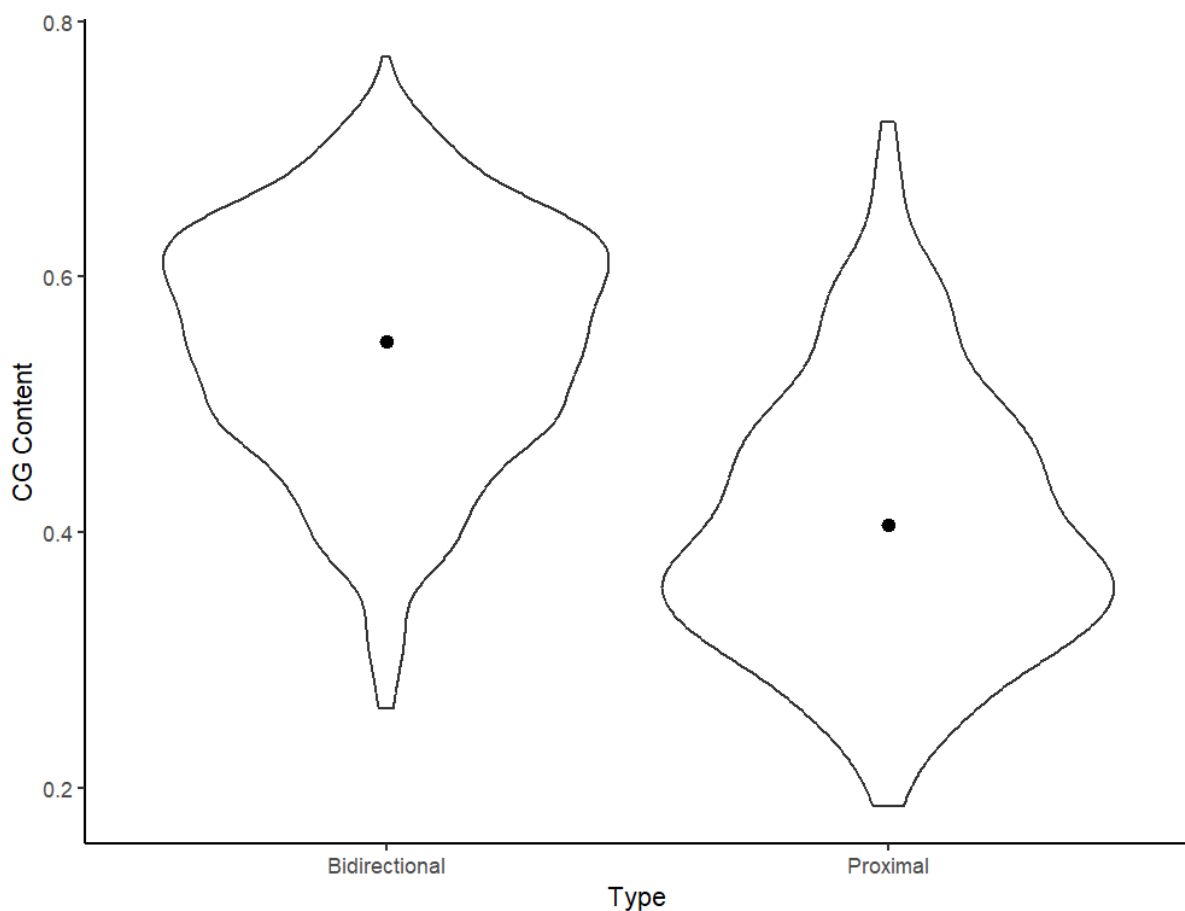


Figure 18. Violin plot comparing the CG content of bidirectional and proximal promoters. The black dot represents the mean value. The difference is significant (two sample t-test) with a p-value approaching zero (p -value < 0.001).

The mean CG content for bidirectional promoters is 0.55 and 0.41 for proximal (Figure 18). Therefore bidirectional promoters have a significantly greater CG content than proximal promoters.

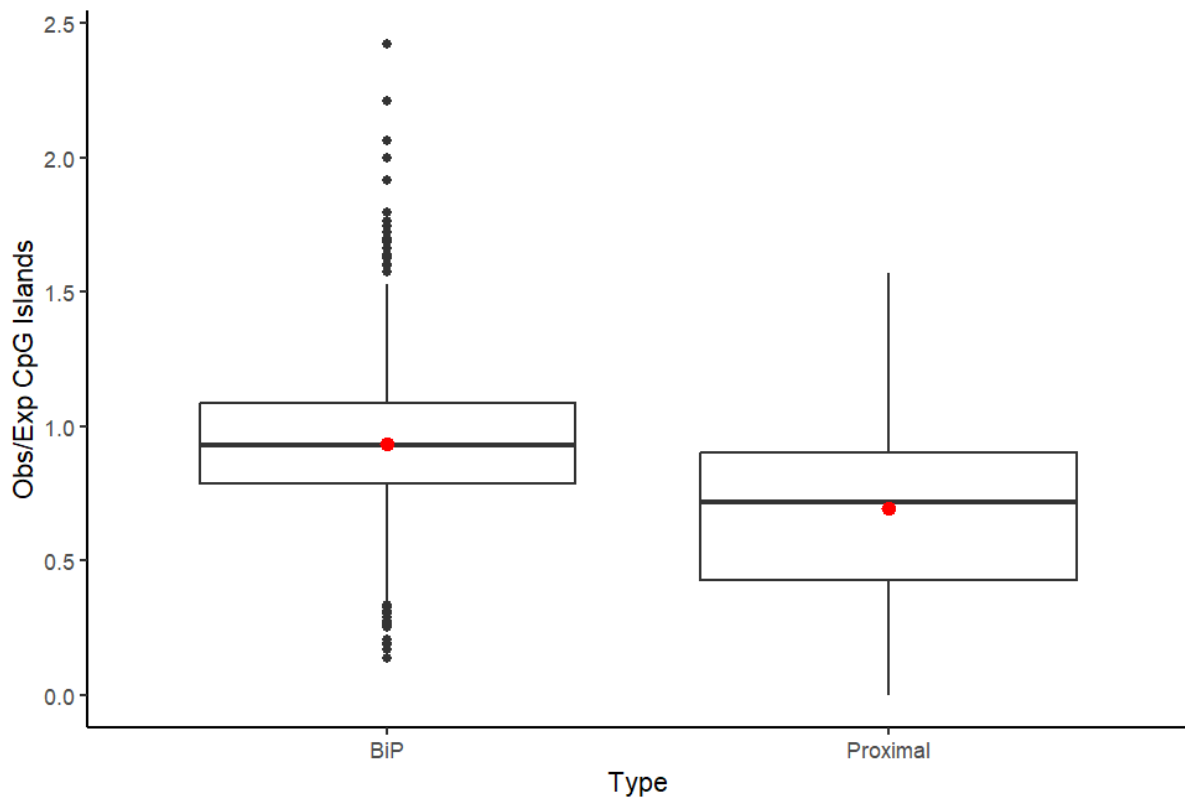


Figure 19. The observed to expected ratio of CpG islands in bidirectional (BiP) and proximal promoters. The red dot represents the mean and the black horizontal line the median. The difference between the observed to expected ratio for proximal and bidirectional promoters is significantly different (two sample t-test) with a p -value < 0.001 .

Bidirectional promoters have a mean of 0.93 observed to expected CpG islands, which is significantly more than the mean of 0.69 for proximal promoters. Therefore both pass the cut off of 0.6 to be considered CpG islands but proximal promoters don't have a high enough CG content whereas bidirectional promoters do. So, in wheat bidirectional promoters tend to be CpG islands, unlike proximal promoters.

3.0.2 Discussion

In wheat, GO terms relating to tRNA and rRNA metabolic processes were enriched in bidirectional genes, which were the first two processes to which the term “housekeeping gene” originally applied (Supplementary 3B). GO terms relating to DNA repair, cell-cycle regulation, ribosome assembly and biogenesis, were also enriched and so agreeing with previous studies that found bidirectional genes are more likely to be involved in cellular maintenance (Supplementary 3B; Liu *et al.*, 2014).

The analyses used in past studies to conclude bidirectional genes are more likely to be housekeeping genes have also largely been based on GO term enrichment, which of the

four criteria to be a housekeeping gene can only test one: being involved in cellular maintenance (Joshi *et al.*, 2022). Therefore, the findings here agree with the literature. To conclusively say if bidirectional genes in all species meet the other three criteria for housekeeping genes: their conservation, essentiality, and stability of expression need to be investigated. The method of taking housekeeping genes and determining what fraction of them are bidirectionally arranged, as done in human studies, could allow one to determine if the genes are essential, if it was reported whether the gene when mutated was lethal (Adachi *et al.*, 2002). Since the genes used in this kind of study are necessarily ones that are studied well enough for their function to be elucidated, this information may already be known so it would be a resource free approach to begin asking this question.

The stability of expression could also be investigated bioinformatically with existing data. A proposed way to measure the stability of expression is by using the Gini coefficient, where 0 represents stable expression and 1 represents differential expression, and then take the bottom 20th percentile of genes, referred to as Gini genes, as these genes would have the most stable expression, being closer to zero (Joshi *et al.*, 2022). Doing this for nine different animal species allowed for identification of housekeeping genes common across the nine species, as well as those common to only primates or non-primates, demonstrating this approach can produce biologically meaningful results (Joshi *et al.*, 2022). GO term enrichment of Gini genes have similar GO term enrichment outputs, suggesting GO term analysis is a valid way of detecting housekeeping genes.

It's striking that in bidirectional gene pairs, genes with related functions as predicted by GO terms were always co-expressed, with the minimum degree of co-expression increasing as the degree of relation of function, as approximated by the number of common GO terms, increased (Figure 17). In other words, bidirectional genes with related functions are more likely to be co-expressed than proximal genes with related functions (Figure 17). In proximal pairs, the degree of relation of function and co-expression was unrelated (Figure 17). Therefore while proximal pairs are more likely to have related functions, it's not an indication of coexpression, whereas a common function in bidirectional pairs is indicative of high co-expression.

The four pairs that shared the most common GO terms were pairs 86, 421, and 424 had five common GO terms each, and pair 85 that shared nine (data not shown). Pair 85 was particularly notable as the genes involved have been characterised (Feuillet *et al.*, 2001). Its overlapping GO terms were as follows: ATP binding, integral component of membrane, kinase activity, membrane, nucleotide binding, phosphorylation, protein kinase activity, protein phosphorylation, and transferase activity. This pair is highly co-expressed (co-expression coefficient = 0.97). The genes were found to be *Tak14* and *Lrk14*, which play roles in plant defence (Feuillet *et al.*, 2001). The two genes share 80% identity with 48% query cover, so while they are distinct genes they may have arisen as a result of gene duplication.

To understand if gene duplication is the predominant mechanism by which genes become bidirectionally arranged I aligned the sequences of bidirectional and proximal pairs and found that both had few hits. I suspect this is an underestimate because proximal genes are known to arise from duplication events and yet only 22 of 469 were detected. This is partially a result of the way in which the analysis was conducted. While nucleotide and peptide

sequences are both used for the detection of duplicate genes, using the peptide sequence is more common as it reduces noise from introns which have a high rate of mutation. It also discounts synonymous substitutions in the exons. Furthermore the cut off I used for the percentage identity was highly stringent, 50% when often 30% or 40% is used (Lallemand *et al.*, 2020). Therefore redoing the analysis in this way could help reduce the number of false negatives.

It would not be a perfect solution, however, as while homology is often used for the detection of duplicate genes, some homologues, especially genes that arose as a consequence of less recent duplication events, will have less than 30 percent identity (Lallemand *et al.*, 2020). Continuing to lower the cut off however begins to introduce an unacceptable number of false positives into the dataset. Therefore different methods have been proposed such as Rost's formula that based homology on similarity of secondary structure rather than sequence (Rost 1999). For low throughput analysis common domains are often a good indicator of homology, so for candidate pairs the domains present can be compared using databases such as NCBI (Lallemand *et al.*, 2020). Alternatively, or in addition, a whole genome gene duplication analysis could be carried out in wheat, as has already been done for particular gene families, and these compared to bidirectionally arranged genes to gain a picture of the origin of bidirectional genes (Ma *et al.*, 2023).

Gene duplication can arise as a result of various mechanisms: whole genome duplication, segmental duplication, tandem duplications, and transposable elements (Lallemand *et al.*, 2020). Whole genome duplication is common in plants, and was crucial to the evolution of modern hexaploid wheat (IWGSC 2018). However, it seems unlikely to give rise to bidirectional promoters because a subsequent genome rearrangement would have to occur where the gene is transferred to a different subgenome, within a thousand base pairs of one of its homoeologues. Furthermore, bidirectional pairs 407, 356, and 407, are homoeologous with one another, and there is no underrepresentation of homoeologous triads in bidirectional pairs (Table 8). Finally, as many of the bidirectional pairs are conserved with rice and thale cress it suggests that these pairs were formed before the divergence of these species, so while ancient whole genome duplications could have played a role in their formation, the two most recent events in wheat haven't, which demonstrates that whole genome duplication doesn't give rise to the bidirectional arrangement.

The most likely mechanisms of gene duplication to give rise to bidirectional genes are tandem duplication or through the action of transposable elements, because these mechanisms can generate duplicate genes immediately next to the original. Tandem duplications occur due to unequal crossing over, and are known to have produced gene clusters such as that at the Rhg1 locus on chromosome 18 of soybean, which confers resistance to cyst nematode (Cook *et al.*, 2012). Tandem duplication is involved preferentially in creating genes that respond to environmental stimuli and bidirectional promoters are enriched for motifs that bind to transcription factors that regulate stress responses (Hanada *et al.*, 2008). Non-tandem mechanisms such as whole genome duplication preferentially give rise to genes with intracellular regulatory functions (Hanada *et al.*, 2008). Tandem genes (along with whole genome duplicated genes) have been found in thale cress and rice to have less divergence in expression than other duplicated genes, this could be in part due to the action of bidirectional promoters (Wang *et al.*, 2011).

Transposable elements, or “jumping genes” are mobile elements within the genome by which genes can move within the genome through either a cut and paste (class II) or copy and paste (class I or retrotransposons) mechanism (Ladd and Bordoni 2023). Transposable elements are common in plant genomes, and more than 80% of the wheat genome is made up of transposable elements, 67% of the wheat genome being composed specifically of long-terminal repeat retrotransposons (Zhang *et al.*, 2021; Bariah *et al.*, 2020).

Retrotransposons can result in copy number variation, which is a source of variation in a population in which different numbers of copies of the same gene are present in different individuals (Bariah *et al.*, 2020). Multiple copies can become fixed in a population overtime and can lead to the formation of new genes, so bidirectional genes could arise in this manner.

In both the case of transposable elements and tandem repeats the new gene would be in the same orientation as the original gene so an inversion mutation would have to occur for the genes to be bidirectionally arranged.

If bidirectional pairs indeed have very little sequence similarity then they must arise from other mechanisms than gene duplication, i.e. from genome rearrangement mutations, these include: inversions, transpositions, chromosome fission and fusion, double-cut joins, and translocations (Hartmann *et al.*, 2017). Inversions are mutations in which the direction of the gene is reversed, so proximal genes could become bidirectionally arranged in this way. Translocations involve the moving of genetic information in the genome through chromosome breakage and re-attachment, possibly leading to genes becoming bidirectionally arranged (Villoutriex *et al.*, 2021; Hartmann *et al.*, 2017).

Overall there are a variety of mechanisms by which the bidirectional genome arrangement may arise, but fundamentally either the gene must be created in a bidirectional arrangement or be moved from elsewhere in the genome thus becoming bidirectionally arranged. The first step in resolving this is to calculate the homology of bidirectional pairs in order to determine if bidirectional pairs result as a consequence of gene duplication. Preliminary investigation suggests that genome rearrangement rather than gene duplication is responsible for the bidirectional gene arrangement but results were not conclusive.

Another avenue to explore the mechanism by which bidirectional genes function is the enriched motifs within bidirectional promoters. Of the motifs enriched in bidirectional promoters four represent the binding sequence of characterised transcription factor families: MADS, NAC, TCP, and AP2 (Table 11). MADS boxes are found in a broad range of eukaryotic organisms and in plants regulate floral development, and are involved in regulating response to temperature and other abiotic stresses (Zhao *et al.*, 2020; Castelán-Muñoz *et al.*, 2019). AP2 are found predominantly in plants and are involved in morphogenesis, stress response, and signal transduction (Riaz *et al.*, 2021). The remaining two families, NAC and TCP, are both plant-specific transcription factors involved in plant development and stress response, which are known to interact with one another, either by forming heterodimers or by NAC acting downstream of TCP (Nakashima *et al.*, 2012; Spears *et al.*, 2022).

In contrast to animals, plants have a flexible body architecture that can be adapted to suit their environment, so transcription factors that regulate genes related to stress response

often also regulate development due to crosstalk between the two processes, thus potentially explaining the apparent discrepancy between bidirectional genes being involved in both cellular maintenance and stress response (Munné-Bosch and Müller 2013).

Of the motifs identified 17 didn't belong to a characterised family, but four of those did have a corresponding motif sequence indicating these sequences had been identified as potential cis-regulatory elements in the literature. Three had imperfect matches where a substring of the enriched motif had been described in the literature, but that substring was not identified itself as enriched motif in bidirectional promoters. The final one, GTCGAC, has been described in barley (Xue 2003). GTCGAC, a CRT/DRE motif (C-repeat/dehydration-responsive element), was identified as interacting with the gene *HvCBF2*, which encodes a transcriptional activator in barley that is transiently upregulated in response to cold (Xue 2003). At 25°C minimal binding between *HvCBF2* and the CRT/DRE motif was detected and a decrease in temperature resulted in increased binding, with maximum expression at 0°C (Xue 2003).

Up1 (GGCCCAWW) is a substring of the enriched motif AATGGGCCGCA, and Up2 (AAACCCTA) is a substring of the enriched motif AAACCCTAGC. Both were found to be enriched in thale cress genes that were upregulated in response to main stem decapitation (Tatematsu *et al.*, 2005). Up1 has been found to be enriched in rice and thale cress bidirectional promoters and Up2 in thale cress (Dhadi *et al.*, 2009). However, neither Up1 nor Up2 themselves were identified as enriched in bidirectional promoters in wheat. S1 (TCTCTCTCT) is a substring of the motif CTCTCTCTCTCY and is found in the 35S promoter of cauliflower mosaic virus where it enhances gene expression and is able to bind to plant nuclear proteins (Pauli *et al.*, 2004). Therefore, perhaps this motif also acts as an enhancer in wheat, particularly in bidirectional promoters.

Of the eight motifs for which previous literature exists all but one is involved in stress response and/or development. One or many of the 13 uncharacterised motifs may be responsible for inducing bidirectional promotion. GTCCA is of particular interest because it's the only one found in the majority of bidirectional promoters.

Finally, I calculated CG content and the ratio of observed to expected CpG islands. A value higher than 0.6 indicates a CpG island, along with a CG content of greater than 50% (Gardiner-Garden and Frommer 1987). Bidirectional promoters, on average, are CpG islands. Most eukaryotes have significantly less than 50% CG content in their genome, and wheat is no exception with the CG content of the wheat genome being less than 48% (Brenchley *et al.*, 2012). This is because methylcytosine is mutagenic hotspot for C to T substitution mutations, in intergenic regions this leads to CpG deficiency overtime, but in genes and promoters there is a greater selection pressure against this mutation, which results in CpG islands being associated with genes more frequently (Deaton and Bird 2011).

CpG islands tend to be constitutively expressed genes more frequently than tissue specific genes which agrees with the idea that bidirectional genes are more likely to be housekeeping genes (Schug *et al.*, 2005).

3.1 Expression of bidirectionally arranged genes in leaf and coleoptile samples

In this section I aimed to take the candidates identified in section and confirm if they were differentially expressed between 15°C and 25°C, as well as co-expressed.

3.3.0 Methods

3.3.0.0 Growth Conditions and Sampling

Triticum aestivum cv. Paragon was grown in Sanyo PMC and Sanyo 4 cabinets at 15°C and 25°C respectively with a 12h light/12h hour dark photoperiod. The soil used was John Innes Cereal Mix which was composed of: 40% medium grade peat, 40% sterilised soil, 20% horticultural grit, 1.3kgm³ PG mix 14-16-18 +Te base fertiliser, 1kgm³ osmocote mini 16-8-11 2mg + Te 0.02% B, wetting agent, 2kgm³ maglime, 300gm³ exemptor. Seeds were planted one seed to a cell just below the surface of the soil in a tray with 5x4 cells, each with dimensions 50x40x30mm. 50mm was the depth at the deepest point. Light intensity was set to 4, which measured as 100±15 µmol.m⁻².s⁻¹, on day 1 approximating a “normal” light level. The Sanyo PMC cabinet was lit with MITSUBISHI/OSRAM FL4OSS.W/37 fluorescent tubes and Sanyo 4 with PHILIPS MASTER TL-D 26W/835 fluorescent tubes. Watering occurred on day 0 and subsequently whenever the soil was dry to the touch.

Leaf and coleoptile samples were taken from each condition on day 4 of growth, where day 0 was the time of planting in soil and immediately flash frozen in liquid nitrogen in 1.5ml Eppendorf tubes. Three biological replicates were taken for each condition, with five coleoptiles/leaves in each replicate in the 25°C condition and seven in the 15°C condition, due to slower growth resulting in less tissue at 15°C. Samples were stored at -80°C.

3.3.0.1 RNA Extraction and cDNA Synthesis

RNA extraction was performed using a Sigma-Aldrich Spectrum Plant Total RNA Kit (product number STRN250-IKT) as per protocol A outlined in the manufacturer’s instructions. cDNA was synthesised via reverse transcription as follows: 5µl RNA was incubated for 30 minutes at 37°C with 1µl DNase (Promega, Wisconsin, USA) and 1.5µl 5xFS buffer (Promega), then 1µl DNase stop solution (Promega, Wisconsin) was added to each sample, which were incubated for 10 minutes at 65°C, followed by the addition of 0.75µl oligodTs and 0.75µl dNTPs to each sample and an incubation of 5 minutes at 65°C. The samples were cooled immediately on ice for around 5 minutes. 3µl 5xFS buffer (Invitrogen), 0.75µl DTT (Invitrogen), 0.75µl Superscript III reverse transcriptase (Invitrogen), 0.75µl RNase out (Invitrogen), and 0.75µl dH₂O was added to each sample and incubated at 25°C for 5 minutes, 50°C for 50 minutes, and 70°C for 15 minutes. The resulting cDNA was diluted by a factor of 10 with nuclease free water and stored at -20°C.

3.3.0.2 qPCR

Expression as approximated by quantity of mRNA transcripts was measured by RT-qPCR using the Biorad C1000 Thermal Cycler and CFX96 Real-Time System. Each well contained 2.5µl cDNA, 2.5µl of primer mix (4µl F, 4µl R and 92µl nuclease free water) and 5µl GoTaq qPCR master-mix (Promega, USA). Gene expression was normalised using housekeeping

gene TraesCS5A02G015600 with primers as designed by Borrill *et al.*, 2016. All other primers were designed using primerBLAST and confirmed to have no off site targets with BLAST.

Table 12. qPCR primers used to calculate expression of candidate genes.

Gene	Direction	Sequence
TraesCS5A02G015600 (Control)	Forward	TCTAAATGTCCAGGAAGCTGTTA
TraesCS5A02G015600 (Control)	Reverse	CCTGTGGTGCCCAACTATT
TraesCS2B02G477700	Forward	ACATGCCGGTTAACAACGT
TraesCS2B02G477700	Reverse	TCCTACTGCATTTCTCTATGGC
TraesCS2A02G455700	Forward	CGCTCCGGGAATGAACTTG
TraesCS2A02G455700	Reverse	CACGTCTGGTTGCATAGCTG
TraesCS2D02G456100	Forward	CGTCTTCTCCATCCCCTCTC
TraesCS2D02G456100	Reverse	AGCTTTAACACGGGATGGGA
TraesCS2B02G477800	Forward	CATTGCAGGCTTCAGTTTCG
TraesCS2B02G477800	Reverse	GTTTTGCAATCCCACGTTGC
TraesCS2A02G455800	Forward	CGCTTAATTCCGCCAGATG
TraesCS2A02G455800	Reverse	GTTCCGGCTTTTACACGGGAC
TraesCS3D02G343900	Forward	TACCCGGCATATCAGAGGGT
TraesCS3D02G343900	Reverse	TCCCCGCTAAATTCAACCCC
TraesCS3B02G382500	Forward	CTGGCATATCAGAGGGCGAG
TraesCS3B02G382500	Reverse	CCCCGCTAAATTCAACCCCT
TraesCS3A02G349900	Forward	GGATGCTGCTGTGTTACCG
TraesCS3A02G349900	Reverse	GTGTGACTTGGGAAATGGCA
TraesCS4D02G067700	Forward	GATGCCACCACCAGCGAG
TraesCS4D02G067700	Reverse	ACCAAACCTTCTGTGCCCA
TraesCS4B02G068600	Forward	GATCCGGCCACCTGTGTA

Gene	Direction	Sequence
TraesCS4B02G068600	Reverse	GAGCAGCAGTTCTCATCGC
TraesCS4B02G068500	Forward	TGAAGGTGACCAGCAAACAGA
TraesCS4B02G068500	Reverse	GCATAGTGGCTTGACAGGGAT
TraesCS4D02G067600	Forward	AAGCCTTGCCGAAACCATTC
TraesCS4D02G067600	Reverse	GACTGCTCCTCTTCTGTTTGC
TraesCS3D02G343800	Forward	CTCCACGTCCCTCTCGGT
TraesCS3D02G343800	Reverse	AAGACGGCTCTACAGAACCC
TraesCS3A02G349800	Forward	GCCGGATTATATCCAAGGGCA
TraesCS3A02G349800	Reverse	TTCATCTGGACCAACCCACG
TraesCS3B02G382400	Forward	TTCAGGTGCAGAGAATGCGT
TraesCS3B02G382400	Reverse	ATAGGTGTTAGCAGCAGCGT

3.3.0.4 Wildtype Paragon Experiment: Coleoptile Length

Paragon was grown at 10°C, 15°C, 20°C, 25°C, and 30°C, all other conditions were as described above. The growth cabinets used in all cases were Sanyo. Coleoptile length measurements were taken from day 1 to 9 and measured to the nearest millimetre.

3.3.1 Results

An experiment in which no samples were taken but the coleoptile was allowed to develop to its maximum length at various temperatures was carried out. This was done to determine the day in which most growth was taking place so in a future experiment samples could be taken on this day, as well as to establish the growth habits of the coleoptile in Paragon under different temperatures.

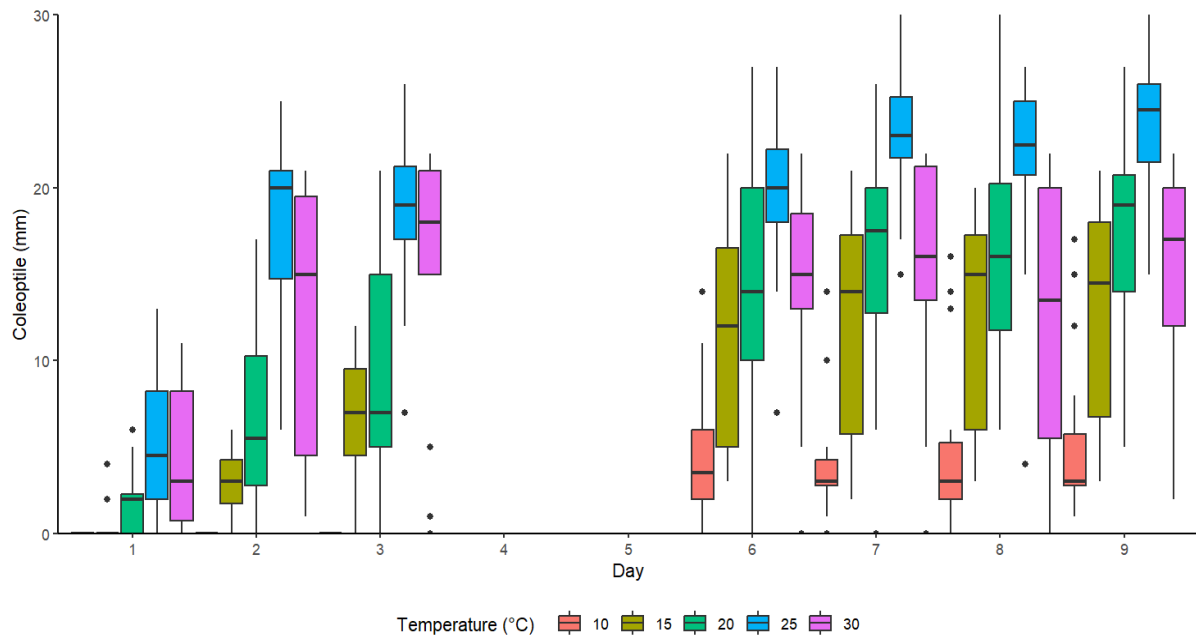


Figure 20. The coleoptile length of Paragon wheat on days 1 to 9 of growth in temperatures from 10°C to 30°C at 5 degree intervals. Red = 10°C, Yellow = 15°C, green = 20°C, blue = 25°C, purple = 30°C. Black horizontal line represents the median, the coloured boxes the upper and lower quartile and the vertical line represents the range of the data, with anomalies, represented by black dots, excluded.

The coleoptile length increased until day 6 where it started to plateau for all temperatures (Figure 20). As the temperature increased the maximum coleoptile length increased, reaching a maximum at 25°C after which it started to decrease (Figure 20). The shortest coleoptile length occurred at 10°C (Figure 20). The days with the greatest rate of growth appeared to be on day 4 or 5, so day 4 was chosen to take RNA samples (Figure 20).

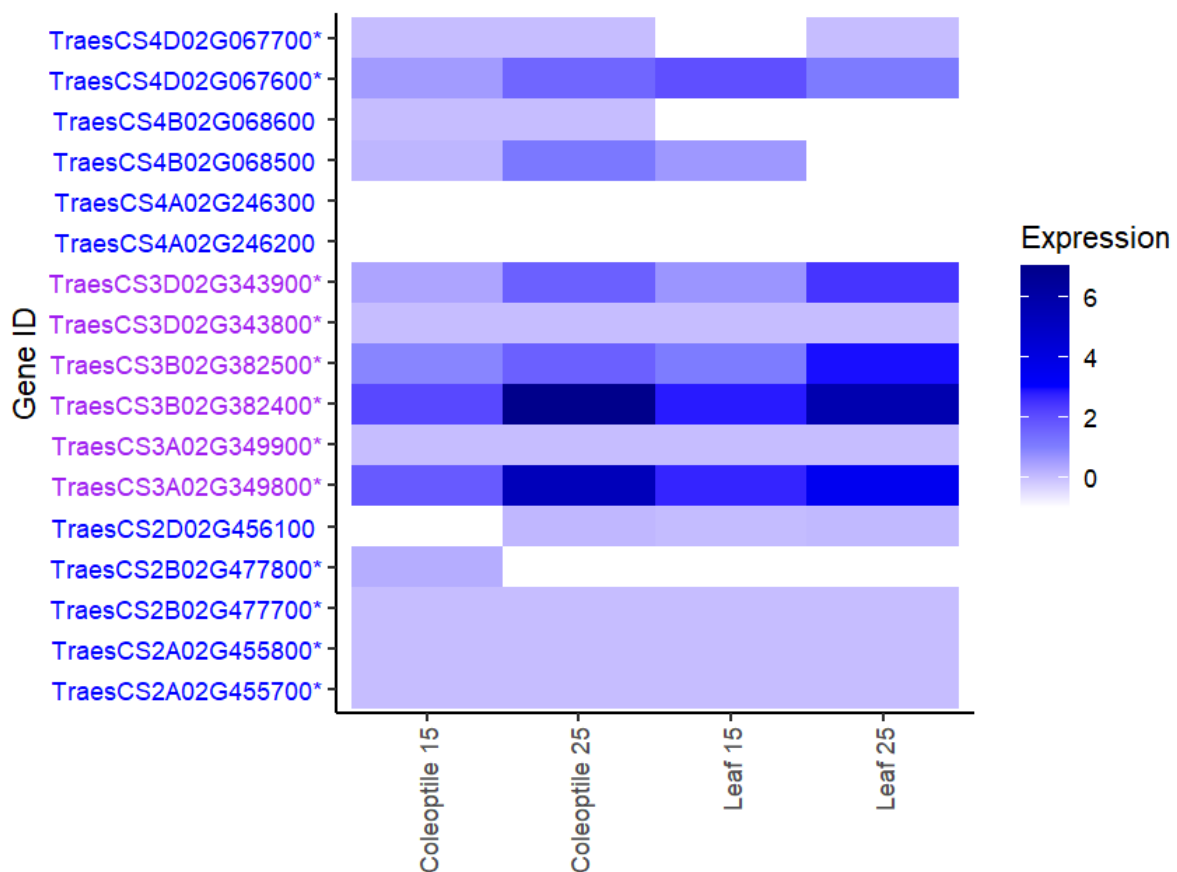


Figure 21. The expression in arbitrary units as determined by qPCR of candidate bidirectional pairs and their homoeologues. Two different tissue types, coleoptile and leaf, were sampled at two different temperatures, 15°C and 25°C. Low to high expression is represented by lilac to dark blue shading. White indicates no data was collected for that point. The colours on the y axis delineate different sets of homoeologues, i.e. one bidirectional pair and the homoeologues for each form one group. Bidirectionally arranged genes are marked with an asterisk (*).

The set of five genes on chromosome 2 are essentially not expressed with the highest expression being 0.3 for gene TraesCS2B02G477800 in the coleoptile at 15°C (Figure 21). Expression data for the set on chromosome 4 is only available for four of six genes due to lack of suitable primers (Figure 21). Only one gene in the bidirectional pair and its homoeologue was expressed, and its expression was higher and 25°C in the coleoptile but lower at 25°C in the leaf (Figure 21). The chromosome 3 set was the only one where a pair (407) was expressed to some degree in all conditions (Figure 21). Only one of the two genes in the other two pairs was expressed (Figure 21). As the most promising pair, the expression of pair 407 and its homoeologues was looked at in more detail.

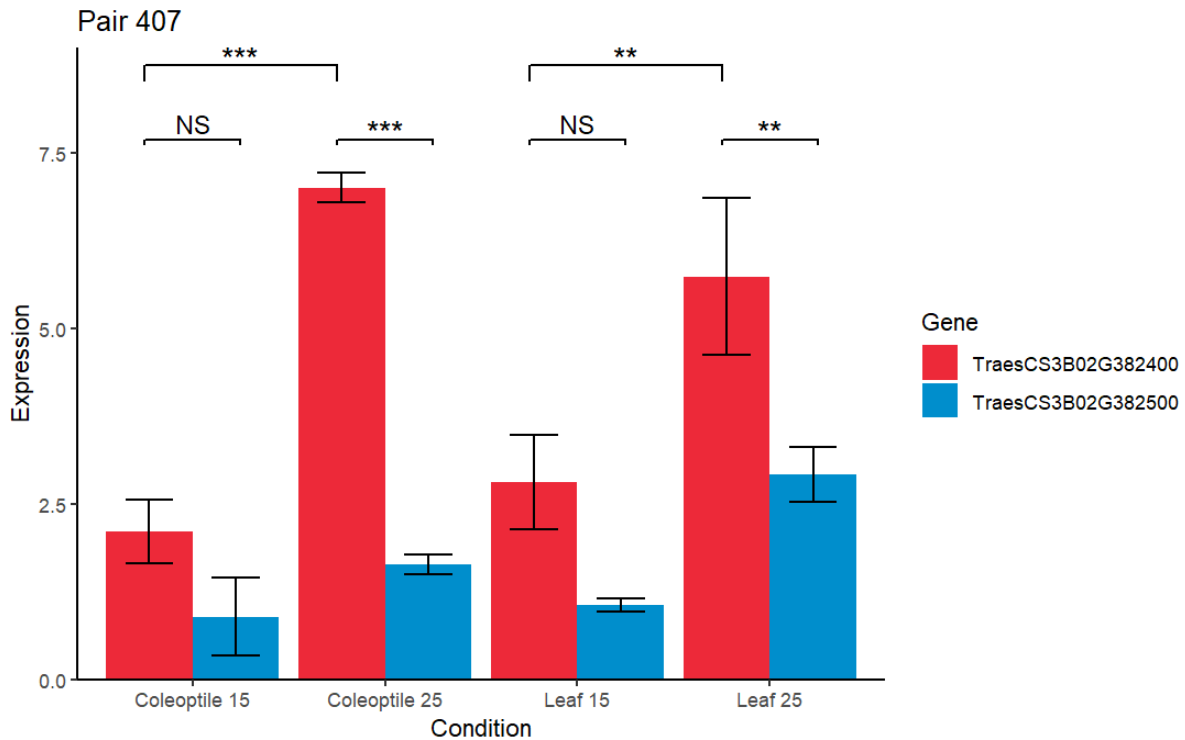


Figure 22. The expression in arbitrary units of bidirectional pair 407, genes *TraesCS3B02G382400* (red) and *TraesCS3B02G382500* (blue), as determined by qPCR. The condition is labelled with the tissue and the temperature. Error bars represent standard error. Levels of significance are indicated as follows: NS = not significant; ** $p < 0.05$; *** $p < 0.001$.

The ratio of gene expression between the pairs in each condition to 1s.f. is: 0.4, 0.2, 0.4, and 0.5, respectively (Figure 22).

At 15°C there is no significance between the genes in pair 407 in either tissue (Figure 22). At 25°C the reverse gene, *TraesCS3B02G382400*, is more highly expressed than the forwards gene in both tissues (Figure 22). This also means that the reverse gene is more highly expressed in both tissues at 25°C compared to 15°C (Figure 22). The co-expression seems to break down at 25°C as the difference between the genes becomes significant, and the expression of *TraesCS3B02G382400* is significantly higher at 25°C than 15°C, though the broad pattern of expression between the two pairs is maintained as *TraesCS3B02G382500* expression does also increase but not significantly (Figure 22).

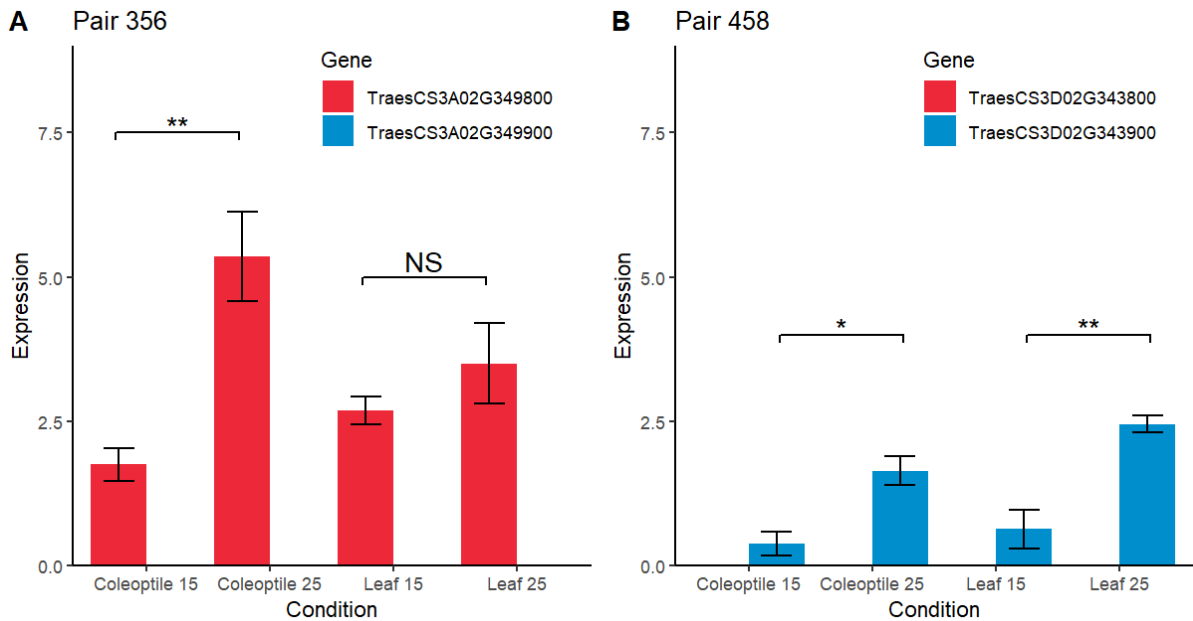


Figure 23. A) Expression in arbitrary units of bidirectional pair 356, genes *TraesCS3A02G349800* (red) and *TraesCS3A02G349900* (blue), as determined by qPCR. These genes are the subgenome A homoeologues of candidate pair 407. B) Expression in arbitrary units of bidirectional pair 458, genes *TraesCS3D02G343800* (red) and *TraesCS3D02G343900* (blue). All) The conditions are labelled with the tissue and the temperature. Error bars represent standard error. Levels of significance are indicated as follows: NS = not significant; * $p < 0.1$; ** $p < 0.05$.

Both homoeologous pairs are unidirectional, being transcribed in only one direction, and in opposite directions for each gene (Figure 23). For the remaining gene the pattern of expression mimics that of pair 407 genes, where the expression increases in all cases at 25°C compared to 15°C, although in one of the four cases this was not found to be significant (Figure 23).

3.3.2 Discussion

3.3.2.0 Coleoptile Length

I found that the coleoptile length increased at higher temperatures until 25°C, which is the opposite trend to what is observed in the literature where a faster growth rate at higher temperatures is counteracted by a longer elongation period, resulting in a longer coleoptile overall at cold temperatures (Figure 20; Pinthus and Abraham 1996). This is usually especially pronounced in GA sensitive cultivars such as Paragon (Botwright *et al.* 2001).

There is a distinct plateau in the data between day 6 and 9 for all temperatures measured, which demonstrates that the elongation period was over at all temperatures (Figure 20). However, to confirm this, the experiment could be repeated with measurements taken until day 20.

While the trend of longer coleoptiles at cooler temperatures has been observed in many different varieties of wheat, none to my knowledge have been conducted in Paragon. Therefore the explanation may either be genetic or environmental. To confirm which, the experiment could be repeated with Paragon and a cultivar that has been previously found to have longer coleoptiles at shorter temperatures. If the trend is inverted in both, the differences are due to environmental conditions. If Paragon has shorter coleoptiles at cooler temperatures and the other cultivar longer, the differences are genetic.

There are some differences between the standard experimental set up and the conditions I used in this study. As I was aiming to mimic real world conditions where possible the wheat seeds were grown in soil in a 12h/12h light/dark photoperiod. Generally in studies measuring coleoptile length the seeds are grown in germination paper entirely in the dark (Amram *et al.*, 2015). Germination paper is used for ease of measurement, and the seeds are germinated in the dark as light has an inhibitory effect on coleoptile growth (Wei *et al.*, 2022). Studies that do plant the coleoptiles in soil tend to do so at a depth of around 30mm, which is the depth seeds are planted at in the field (KWS 2023). The exceptions are those studies related to deeper sowing, where the sowing depth can be up to 200mm (Mohan *et al.*, 2013). In this study the seeds were planted just under the surface of the soil, in order to measure coleoptile length non-destructively. Therefore the coleoptiles were almost immediately exposed to the light, which perhaps had an inhibitory effect across temperatures that is not observed when seeds are grown entirely in the dark, explaining the fact a longer elongation period was not observed at cooler temperatures.

One study found a reversal in the trend of cooler temperatures resulting in longer coleoptiles, wherein most genotypes had longer coleoptiles at higher temperatures when the light intensity was $220 \mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$ as opposed to $140 \mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$ or in the dark (Pinthus and Abraham 1996). The light intensity for this experiment was $100\pm 15 \mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$, which wasn't sufficient to reverse the trend in this study, but genotype also had an effect (Pinthus and Abraham 1996). In the cultivar Maris Huntsman with an *rht* allele (tall phenotype) the coleoptiles were longer with increased temperature, as opposed to *rht* 1, 2, 3, *rht* 1 and 2 in combination and *rht* 2 and 3 in combination (Pinthus and Abraham 1996). Therefore light intensity, temperature, and genotype interact. At low light intensity, coleoptiles decrease in length with temperature, at high light intensity coleoptiles increase in length with temperature, and the light intensity at which this reversal occurs is genotype dependent. In natural conditions the light intensity will be much greater than in growth cabinet conditions, varying from 1,000-100,000 lux, depending on time of day and degree of shading, so in nature the coleoptile length will always increase in response to increased temperature (bios 2023). Therefore, the most likely explanation for the contradictory result is that the relationship between temperature and coleoptile length is often reported on in an absence of light, whereas here, due to shallow planting, light intensity becomes an important factor.

Additionally the coleoptiles were short compared to previously reported values. In GA sensitive cultivars planted at a depth of 100mm coleoptiles for 10 different varieties were between 100mm and 140mm in length on average whereas in this study they ranged between 1-39mm by day 9 (Amram *et al.*, 2015; Figure 20). This can be explained by the shallow planting, resulting in immediate exposure to light once the coleoptiles begin to grow, and light is known to have an inhibitory effect of coleoptile elongation.

3.3.2.1 Expression in Leaf and Coleoptile Tissue

The lack of expression in the chromosome 2 set is likely due to these genes being unexpressed in the chosen conditions and tissue. This was despite RNAseq data from the wheat expression browser that demonstrated that these genes are expressed in the first leaf and coleoptile in Chinese Spring and Azhurnaya, although no data is available for the growth conditions used in this experiment, or in the cultivar Paragon. Perhaps these genes have a different expression pattern in Paragon compared to other cultivars. Due to the lack of expression no conclusion can be drawn about the bidirectionality of this set.

The chromosome 4 pair is unidirectionally transcribed in these conditions, the paired gene not being expressed at all. The homoeologues follow this pattern in expression. However, the pair has a relatively high correlation coefficient of 0.68 so this was counter to expectation, and the correlation persists when narrowed down to only coleoptile tissue or only considering expression for days 2, 3, and 4, ruling out co-expression in only a specific developmental condition or tissue.

In the chromosome 3 set the two homoeologous pairs have the reverse expression pattern, with only one gene being expressed in a temperature dependent manner but the opposite gene in each pair is the one expressed. The two promoter sequences have an 84% identity and 71% query cover, matched on the plus/plus strands, indicating the direction of the promoter is the same in both pairs. This suggests that effect of the orientation of the promoter on directionality is less important than other factors, which agrees with the fact that 92% of transcription factors are non-directional, i.e. even when the direction of the binding motif is reversed, the direction of transcription is maintained (Sharon *et al.*, 2012).

The final pair in this set, 407, fit the bidirectional model where when the expression of one gene increased so did its pair, although not always significantly. To be able to use the promoter of this pair for transgenics where two genes are activated at high temperatures further characterisation of how genes under its control are affected by temperature in different tissues needs to be carried out in order to be able to reliably predict to what extent genes under its control will be expressed.

Chapter 4 - Discussion

This study set out to identify and characterise bidirectional promoters and associated genes in wheat, then focus more directly on those with temperature related functions. This was accomplished in three stages: whole genome identification of bidirectionally arranged genes in wheat (Section 2.0), measurement and calculation of key characteristics (Chapter 3; Chapter 2.1-2.2) and use of RNAseq data and GO terms to identify genes that respond to temperature (Chapter 2.3).

4.1 Whole Genome Identification of Bidirectionally Arranged Genes in Wheat

The reference genome sequence of *Triticum aestivum* landrace cultivar Chinese Spring was used to bioinformatically identify genes with TSS within 1000 bp of one another on opposite strands. This is the standard definition for a bidirectionally arranged gene pair, which is a genomic feature that has attracted attention for its role in genome regulation and role in

regulating oncogenes in humans (Greene *et al.*, 2007). Bidirectionally arranged genes are assumed to be under the control of the same promoter; certainly the bioinformatic definition will include false positives where this is not the case, but the noise created from these is supposed to be minimal in such a large dataset as whole genomes. Bidirectional promoters have been identified in a number of plant species including thale cress, rice, maize, and poplar (Dhadi *et al.*, 2009). Here, I aimed to identify them in the agriculturally important crop wheat and I found it has 1,050 bidirectional promoters, or 2,100 gene pairs, representing around 2% of genes in wheat. This is relatively few compared to thale cress in which 16.8% of genes are bidirectionally arranged, and rice in which 5.6% of genes are bidirectionally arranged (Table 2). This is owing to the fact wheat has a large, gene poor, genome (Sandhu and Gill 2002). Accounting for this factor of gene density wheat has more than the expected number of bidirectional promoters (Figure 3).

4.2 Calculation of Key Characteristics

Three characteristics of bidirectional promoters and genes are common across all species: they are enriched for functions relating to cellular maintenance, the promoters are CpG islands rather than TATA box controlled, and they are co-regulated more frequently than other genes (Bagchi and Vishwanath 2016; Adachi *et al.*, 2002). Wheat bidirectional promoters and genes followed these criteria (Figure 15; Figure 18; Figure 19; Figure 10). Additional characteristics have been suggested for bidirectional genes: that pairs have similar functions, that they arise as a consequence of gene duplication, and that they are under the control of a transcription factor that induces bidirectional transcription (Koyanagi *et al.*, 2005; Collins *et al.*, 2007). Bidirectional pairs in this study were not found to have more similar functions than proximal pairs, which were defined for the purpose of this study as genes that had transcription start sites within a 1000 bp of one another but were on the same strand, and thus due to the 5-3' direction of transcription could not be under the control of a single promoter situated between the two genes (Figure 16; Figure 2). It was however found that as the functional relatedness of bidirectional genes as determined by GO terms increased, so did the minimum level of co-expression (Figure 17). Gene duplication as a mechanism by which bidirectional promoters arise was investigated by alignment of the bidirectional pairs, but the results were not conclusive (Table 10). A motif enrichment analysis was carried out on the promoter regions of bidirectional pairs and enriched motifs were identified (Table 11). Consensus motifs of transcription factors that regulate stress response were enriched (Figure 8). Bioinformatic detection alone is not enough to conclude if any of these are involved in binding transcription factors that induce divergent transcription and future experimental work is required to determine what binds these motifs.

4.3 RNAseq data and GO terms can be used to identify genes that respond to temperature

Triticum aestivum is a globally important crop that is losing productivity as temperature conditions become more extreme due to the impact of climate change (Asseng *et al.*, 2015). Wheat breeding programs have always aimed to increase yield and as we begin to experience the impact of climate change more emphasis is being placed on developing cultivars that are tolerant to temperature and the often related drought stress. Therefore, I focused on identifying those bidirectional genes that responded to temperature using GO terms and RNAseq data. I used an additional criteria of high co-expression, assuming they would represent true bidirectional genes rather than false positives. Of three sets (candidate

pairs plus homoeologues) one pair demonstrated a high degree of bidirectional transcription and differential expression (Figure 22).

Identification of genes and regulatory elements that can be targeted in breeding programs to increase tolerance to heat stress can aid such programs in developing new cultivars. As bidirectional promoters are an important regulatory element that tie the expression of two genes together, understanding the advantages and mechanisms of regulating genes in this manner will allow them to be manipulated. For example, introducing additional motifs may increase the degree of co-expression between two genes. Another use of bidirectional promoters is in creating transgenics, where controlling two genes with one promoter can contribute to the problem of gene silencing when stacking genes (Rajeevkumar *et al.*, 2015). Promoters that induce transcription only under particular conditions are sometimes used, for example, at high or low temperature (Li *et al.*, 2013).

My research has enabled the identification of putative bidirectional promoters in the agriculturally important wheat plant. Through future characterisation of these promoters they could be used in the development of transgenic lines which can control the expression of two genes, or development of lines through selective breeding to alter the existing regulatory pathways in temperature response.

Supplementary

Supplementary information can be found at
<https://github.com/bskh1212/Bidirectional-Promoters.git>

Glossary

Bidirectional Gene Pair

Two genes on opposite DNA strands with transcription start sites within 1000 bp of one another.

CpG Island

Regions that are unusually dense in cytosine followed by guanine bases in the DNA backbone in the 5' to 3' direction and have a high CG content.

Dictionary (python)

A data type that stores values that can be accessed by unique keys. Cannot be accessed via indexing.

For Loop (programming)

A conditional statement that will perform a certain task when conditions are met.

Function (programming)

Code segments that perform a specific task.

Head Gene

A proximal gene with the transcription start site orientated towards its pair and the 3' end orientated away from its pair.

Index (programming)

A number that allows one to reference stored data, for example from a list.

Lambda Function (python)

A small, anonymous function.

Library (python)

A collection of additional code for a specific purpose that is not included in the python standard library, e.g. the library “pandas” can be imported for dataframe handling. Equivalent to a package in R.

List (python)

A way of storing heterogeneous variables sequentially, can be accessed via indexing.

Package (R)

A collection of additional code for a specific purpose that is not included in base R. e.g. the seqinr package for handling sequences of nucleic acids. Equivalent to a library in python.

Promoter

A region of DNA upstream to a gene that bind regulatory elements required to initiate transcription.

Proximal Gene Pair

Two genes on the same DNA strand with transcription start sites within 1000 bp of one another.

String (programming)

A data type used to represent text rather than numbers.

Tail Gene

A proximal gene with its transcription start site orientated away from its pair, and the 3' end of the gene orientated towards its pair.

References

Abdolshahi, R., Foroodi-Safat, S., Mokhtarifar, K, Ataollahi, R., Maghsoudi Moud, A., Kazempour, A., Pourseyedi, S. and Rahmani, A. 2021. Challenges of breeding for longer coleoptile in bread wheat (*Triticum aestivum* L.), *Genetic Resources and Crop Evolution*. 68(), pp. 1517–1527.

Adachi, N. and Lieber, M.R. 2002. Bidirectional gene organization: A common architectural feature of the human genome, *Cell*, 109(7), pp. 807-809.

AHDB. 2023. Wheat Growth Guide. Accessed 25/10/2023.

[https://projectblue.blob.core.windows.net/media/Default/Imported%20Publication%20Docs/AHDB%20Cereals%20&%20Oilseeds/General/Wheat%20growth%20guide%20\(2023\).pdf](https://projectblue.blob.core.windows.net/media/Default/Imported%20Publication%20Docs/AHDB%20Cereals%20&%20Oilseeds/General/Wheat%20growth%20guide%20(2023).pdf)

Ahmad, S.S., Samia, N.S.N, Khan, A.S, Turjya, R.R. and Khan, M.A. 2020. Bidirectional promoters: an enigmatic genome architecture and their roles in cancers, *Genes (Basel)*. 11(9), p. 1046.

Akter, N. and Islam, M.R. 2017. Heat stress effects and management in wheat. A review, *Agronomy for Sustainable Development*. 37(), pp.

- Alberts, B., Johnson, A., Lewis, J., Raff M., Roberts, K. and Walter, P. 2002. *Molecular Biology of the Cell* (4th edition): From DNA to RNA, New York, Garland Science.
- Almarri, N.B., Alghamdi, S.S., ElShal, M.H. and Afzal, M. 2023. Estimating genetic diversity among durum wheat (*Triticum durum* desf.) landraces using morphological and SRAP markers, *Journal of the Saudi Society of Agricultural Sciences*. 22(5), pp. 273-282.
- Amalova, A., Yermekbayev, K., Griffiths, S., Abugalieva, S., Babkenov, A., Fedorenko, E., Abugalieva, A. and Turuspekov, Y. 2022. Identification of quantitative trait loci of agronomic traits in bread wheat using a Pamyati Azieva × Paragon mapping population harvested in three regions of Kazakhstan, *PeerJ*. 10().
- Amram, A., Fadida-Myers, A. Golan, G., Nashef, K., Ben-David, R. and Peleg, Z. 2015. Effect of GA-sensitivity on wheat early vigor and yield components under deep sowing, *Frontiers in Plant Science*. 6().
- Ashikawa, I. 2001. *The Plant Journal*. Gene-associated CpG islands in plants as revealed by analyses of genomic sequences, 26(6), pp. 617-625.
- Asseng, S. *et al.* 2014. Rising temperatures reduce global wheat production, *Nature Climate Change*. 5(), pp. 143–147.
- Baboev, S., Muminjanov, H., Turakulov, K. Buronov, A., Mamatkulov, I., Koc, E., Ozturk, I, Dreisigacker, S., Shepelev, S. and Morgounov, A. 2021. Diversity and sustainability of wheat landraces grown in Uzbekistan, *Agronomy for Sustainable Development*. 41(). Pp.
- Bagchi, D.A. and Vishwanath I.R, 2016. The Determinants of Directionality in Transcriptional Initiation, *Trends in Genetics*, 32(6), pp. 322–333.
- Bareket-Samish, A., Cohen, I. and Haran, T.E. 2000. Signals for TBP/TATA box recognition, *Journal of Molecular Biology*, 299(4), pp. 965-977.
- Bariah, I., Keidar-Friedman, D. and Kashkush, K. 2020. Where the Wild Things Are: Transposable Elements as Drivers of Structural and Functional Variations in the Wheat Genome, *Frontiers in Plant Science*. 11().
- Barrero, J.M., Cavanagh, C., Verbyla, K.L., Tibbits, J.F.G., Verbyla, A.P., Huang, B.E., Rosewarne, G.M., Stephen, S., Wang, P., Whan, A., Rigault, P., Hayden M.J. and Gubler, F. 2015. Transcriptomic analysis of wheat near-isogenic lines identifies PM19-A1 and A2 as candidates for a major dormancy QTL, *Genome Biology*. 16(93).
- Barrero, Y.M., Cavanagh, C., Verbyla, K.L., Tibbits, J.F.G., Verbyla, A.P., Huang, B.E., Rosewarne, G.M., Stephen, S., Wang, P., Whan, A., Rigault, P., Hayden, M.J. and Gubler, F. 2015. Transcriptomic analysis of wheat near-isogenic lines identifies PM19-A1 and A2 as candidates for a major dormancy QTL, *Genome Biology*. 16(93).
- Batifoulier, F., Verny, M.A., Chanliaud, E., Rémésy, C. and Demigné, C. 2006. Variability of B vitamin concentrations in wheat grain, milling fractions and bread products, *European Journal of Agronomy*. 25(2), pp. 163-169.

Behura, S. K. and Severson, D. W. 2014. Bidirectional Promoters of Insects: Genome-Wide Comparison, Evolutionary Implication and Influence on Gene Expression. *Journal of Molecular Biology*, 427, 521-536.

Bernardo, J. *et al.* 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations, *Genome Research*. 27(). pp. 885-896.

Bios. How to Measure Light Intensity, *Architectural Lighting*.
<https://bioslighting.com/how-to-measure-light-intensity/architectural-lighting/>. Accessed 01/09/2023.

Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D. and Veitia, R.A. *et al.* 2010. Heterosis, *The Plant Cell*. 22(7) pp. 2105–2112.

Boden, S. A., Kavanová, M., Finnegan, E.J., and Wigge, P.A. 2013. Thermal stress effects on grain yield in *Brachypodium distachyon* occur via H2A.Z-nucleosomes, *Genome Biology*. 14(6).

Bolot, S., Abrouk, M, Masood-Quraishi, U., Stein, N., Messing, J., Feuillet, C., and Salse, J. 2009. The 'inner circle' of the cereal genomes, *Current Opinion in Plant Biology*. 12(2), pp. 119-125.

Borrill, P., Ramirez-Gonzalez, R. and Uauy, C. 2016. expVIP: a customizable RNA-seq data analysis and visualization platform, *Plant Physiology*. 170(4), pp. 2172-2186.

Borrill, P., Harrington, S.A., Simmonds, J. and Uauy, C. 2019. Identification of Transcription Factors Regulating Senescence in Wheat through Gene Regulatory Network Modelling, *Plant Physiology*. 180(3), pp. 1740–1755.

Botwright, T. L., Rebetzke, G.J., Condon, A.G., and Richards, R.A. 2001. Influence of variety, seed position and seed source on screening for coleoptile length in bread wheat (*Triticum aestivum* L.), *Euphytica*. 119(), pp. 349–356.

Brenchley, R. *et al.* 2012. Analysis of the bread wheat genome using whole genome shotgun sequencing, *Nature*. 491(7426), pp. 705–710.

Brown, T.A. 2002. Genomes 2nd edition: Chapter 15: How Genomes Evolve, Garland Science.

Buhrow, L.M. Cram, D., Tulpan, D., Foroud, N.A., Loewen, M.C. 2016. Exogenous Abscisic Acid and Gibberellic Acid Elicit Opposing Effects on *Fusarium graminearum* Infection in Wheat, *APS Publications*. 106(9).

Cantu, D., Pearce, S.P, Distelfeld, A., Christiansen, M.W., Uauy, C., Akhunov, E., Fahima, T. and Dubcovsky, J. 2011. Effect of the down-regulation of the high Grain Protein Content

(GPC) genes on the wheat transcriptome during monocarpic senescence, *BMC Genomics*. 12(492).

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., Dubcovsky, J., Saunders, D. and Uauy, C. 2013. Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors, *BMC Genomics*. 14(270).

Castelán-Muñoz, M., Herrera, J., Cajero-Sánchez, W., Arrizubieta, M., García-Ponce, C.T.P., de la Paz Sánchez, M. Álvarez-Buylla, E.R. and Garay-Arroyo, A. 2019. MADS-Box Genes Are Key Components of Genetic Regulatory Networks Involved in Abiotic Stress and Plastic Developmental Responses in Plants, *Frontiers in Plant Science*. 10().

Charif, D. and Lobry, J.R. 2007. A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, *Structural approaches to sequence evolution: Molecules, networks, populations*. Pp. 207-232.

Charmet, G. 2011. Wheat domestication: Lessons for the future, *Comptes Rendus Biologies*. 334(3), pp. 212-220.

Chavali, A. K. and Rhee, S.Y. 2017. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites, *Briefings in Bioinformatics*. 19(5), pp. 1022–1034.

Chaw, S.M., Chang, C., Chen, H. and Li, W. 2004. Dating the Monocot–Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes, *Journal of Molecular Evolution*. 58(), pp. 424–441.

Choulet, F. *et al.* 2014. Structural And Functional Partitioning Of Bread Wheat Chromosome 3b, *Science*. 345(6194).

Chow, C. N., Lee, Z., Hung, Y., Li, G., Tseng, K., Liu, Y., Kuo, P., Zheng, H. and Chang, W. 2019. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants, *Nucleic Acids Research*.

Collins, P. J., Kobayashi, Y., Nguyen, L., Trinklein, N.D. and Myers, R.M. 2007. The ets-related transcription factor GABP directs bidirectional transcription, *Plos Genetics*, 3(11) pp. 2247-2255.

Costanzo, A., Amos, D.C., Dinelli, G., Sferrazza, R.E., Accorsi, G., Negri, L. and Bosi, S. 2019. Performance and Nutritional Properties of Einkorn, Emmer and Rivet Wheat in Response to Different Rotational Position and Soil Tillage, *Sustainability*. 11(22).

Comai, L. 2005. The advantages and disadvantages of being polyploid, *Nature Reviews Genetics*. 6(), pp. 836–846.

Cook, D. E., Lee, T.J., Guo, X, Melito, S, Wang, K, M. Bayless, A.M, Wang, J, Hughes, T.J, Willis, D.K, Clemente, T.E, Diers, B.W, Jiang, J, Hudson, M.E., and Bent, A.F. 2012. Copy

number variation of multiple genes at Rhg1 mediates nematode resistance in soybean, *Science*. 338(6111).

Cross, S.H. and Bird, A.P. 1995. CpG islands and genes, *Current Opinion in Genetics and Development*. 5(3), pp. 309-314.

Crouse, G.F., Leys, E.J., McEwan, R.N, Frayne, E.G. and Kellems, R.E. 1985. Analysis of the mouse dhfr promoter region: existence of a divergently transcribed gene, *Molecular and cellular biology*, 5(8), pp. 1847–1858.

Dash, A., Gurdaswani, V., D'Souza J,S. and Ghag, S.B. 2020. Functional characterization of an inducible bidirectional promoter from *Fusarium oxysporum* f. sp. *Cubense*, *Scientific Reports*, 10(1).

Deaton, A. M. and Bird, A. 2011. CpG islands and the regulation of transcription. *Genes and Development*. 25(10), pp. 1010–1022.

Deng, Y., Dai, X., Xiang, Q., Dai, Z., He, C., Wang, J. and Feng, J. 2010. Genome-wide analysis of the effect of histone modifications on the coexpression of neighboring genes in *Saccharomyces cerevisiae*, *BMC Genomics*. 11(550).

Dhadi S.R. Krom, N. and Ramakrishna, W. 2009. Genome-wide comparative analysis of putative bidirectional promoters from rice, *Arabidopsis* and *Populus*, *Gene*, 429(15), pp. 65-73.

Dobon, A. Bunting, D.C.E., Cabrera-Quio, L.E., Uauy, C. and Saunders, D.G.O. 2016. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression, *BMC Genomics*. 17(380).

Dubin, M.J. Scheid, O.M., Becker, C. 2018. Transposons: a blessing curse, *Current Opinion in Plant Biology*. 42(), pp. 23-29.

Ehsan, H., Reichheld, J.P., Durfee, T., Roe, J.L. 2004. TOUSLED Kinase Activity Oscillates during the Cell Cycle and Interacts with Chromatin Regulators, *Plant Physiology*.134(4), pp. 1488–1499.

Farooq, M., Bramley, H., Palta, J.A. and Siddique, K.H.M. 2011. Heat Stress in Wheat during Reproductive and Grain-Filling Phases, *Critical Reviews in Plant Sciences*. 30(6), pp. 491-507.

Fedorof, N. 2012. Transposable Elements, Epigenetics, and Genome Evolution, *Science*. 338(6108), pp. 758-767.

Feliner, G.N., Casacuberta, J. and Wendel, J.F. 2020. Genomics of Evolutionary Novelty in Hybrids and Polyploids, *Frontiers in Genetics*. 11(), pp.

Feuillet, C., Penger, A., Gellner, K., Mast, A. and Keller, B. 2001. Molecular Evolution of Receptor-Like Kinase Genes in Hexaploid Wheat. Independent Evolution of Orthologs after

Polyploidization and Mechanisms of Local Rearrangements at Paralogous Loci1, *Plant Physiology*. 125(3), pp. 1304–1313.

Gardiner-Garden, M. and Frommer, M. 1987. CpG Islands in vertebrate genomes, *Journal of Molecular Biology*. 196(2), pp. 261-282.

Gaut, B.S. 2002. Evolutionary dynamics of grass genomes, *New Phytologist*. 154(1), pp. 15-28.

Geng, X., Zang, X., Li, H., Liu, Z., Zhao, A., Liu, J., Peng, H., Yao, Y., Hu, Z., Ni, Z., Sun, Q. and Xin, M. 2018. Unconventional splicing of wheat TabZIP60 confers heat tolerance in transgenic Arabidopsis, *Plant Science*. 274(), pp. 252-260.

Gillies, S.A., Futardo, A. and Henry, R.J. 2012. Gene expression in the developing aleurone and starchy endosperm of wheat, *Plant Biotechnology Journal*, 10(6).

Giraldo, P., Benavente, E., Manzano-Agugliaro, F. and Gimenez, E. 2019. Worldwide Research Trends on Wheat and Barley: A Bibliometric Comparative Analysis, *Agronomy*. 9(7), pp. 352-.

GISTEMP Team. 2023. GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies [Accessed 21 June 2023]. Available at: <https://earthobservatory.nasa.gov/world-of-change/global-temperatures>.

Gluck-Thaler, E. and Slot, J.C. 2018. Specialized plant biochemistry drives gene clustering in fungi, *Nature: The Isme Journal*, 12(7), pp.1694-1705.

Gou, L., Hattori, J., Fedak, G., Balcerzak, M., Sharpe, A., Visendi, P., Edwards, D., Tinker, N., Wei, Y.M., Chen, G.Y. and Ouellet, T. 2016. Development and Validation of Thinopyrum elongatum–Expressed Molecular Markers Specific for the Long Arm of Chromosome 7E, *Crop Science*. 56(1).

Greene, W.K., Sontani, Y., Sharp, M.A., Dunn, D.S., Kees, U.R., Bellgard, M.I. 2007. A promoter with bidirectional activity is located between TLX1/HOX11 and a divergently transcribed novel human gene, *Gene*. 391(1–2), pp. 223-232.

Gregersen, P.L., Culetic, A., Boschian, L. and Krupinska, K. 2013. Plant senescence and crop productivity, *Plant Molecular Biology*. 82(), pp. 603–622.

Grewal, S., Othmeni M., Walker J., Hubbard-Edwards, S., Yang, C., Scholefield, D., Ashling, S., Isaac, P., King, I.P. and King, J. 2020. Development of Wheat-Aegilops caudata Introgression Lines and Their Characterization Using Genome-Specific KASP Markers, *Frontiers in Plant Science*. 11().

Gustavo A. S. and Rawson H.M. 1996. Responses to photoperiod change with phenophase and temperature during wheat development, *Field Crops Research*. 46(1-3), pp. 1-13.

- Hammond-Kosack, M.C.U., King, R., Kanyuka, K. and Hammond-Kosack, K.E. 2021. Exploring the diversity of promoter and 5'UTR sequences in ancestral, historic and modern wheat, *Plant Biotechnology Journal*. 19(12), pp. 2469-2487.
- Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K. and Shiu, S.H. 2008. Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli, *Plant Physiology*. 148(2), pp. 993-1003.
- Harrell, Jr. F. 2022. `_Hmisc: Harrell Miscellaneous_`. R package, version 4.7-2, <<https://CRAN.R-project.org/package=Hmisc>>.
- Hartmann, T., Middendorf, M. and Bernt, M. 2017. Genome Rearrangement Analysis: Cut and Join Genome Rearrangements and Gene Cluster Preserving Approaches, *Methods in Molecular Biology*. 1704().
- Hong-Qing, L. *et al.*, 2018. Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*, *Nature*. 557(), pp. 424–428.
- Inukai, S., Kock, K.H. and Bulyk, M.L. 2017. Transcription factor–DNA binding: beyond binding site motifs, *HHS Author Manuscripts*. 43(), pp. 110–119.
- Islam, M.T. *et al.*, 2016. Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae*, *BMC Biology*. 14(84).
- IWGSC. 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation, *Nature*. 496(), pp. 91–95.
- IWGSC. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome, *Science*. 361(6403).
- IWGSC. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome, *Science*. 361(6403).
- Janga, S.C., Collado-Vides, J. and Babu, M.M. 2008. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes, *Biophysics and Computational Biology*. 105 (41) pp. 15761-15766.
- Jiang, H., Gao, W., Jiang, B.L, Liu, X., Jiang, Y.T, Zhang, L.T, Zhang, Y., Yan, S.N., Cao, J.J, Lu, J. Ma, C.X., Chang, C. and Zhang, H.P. 2023. Identification and validation of coding and non-coding RNAs involved in high-temperature-mediated seed dormancy in common wheat, *Frontiers in Plant Science*. 14().
- Jin, B., Li, Y. and Robertson, K.D. 2011. DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?, *Genes and Cancer*. 2(6), pp. 607–617.
- Joshi, C.J., Ke, W., Drangowska-Way, A., O'Rourke, E.J., Lewis, N.E. 2022. What are housekeeping genes?, *PLoS Computational Biology*. 18(7).

- Kellogg, E. A. 2001. Evolutionary History of the Grasses, *Plant Physiology*. 125(3), pp. 1198–1205.
- Khan, A., Ahmed, M., Ahmed, M. and Hussain, M.I. 2021. Rising Atmospheric Temperature Impact on Wheat and Thermotolerance Strategies, *Plants*. 10(1), p. 43.
- Koyanagi, K.O., Hagiwara, M., Itoh, T., Gojobori, T. and Imanishi, T. 2005. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system, *Gene*. 353(2), pp. 169-176.
- Kristiansson, E., Thorsen, M., Tamás, M.J. and Nerman, O. 2009. Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements, *Molecular Biology and Evolution*. 26(6), pp. 1299–1307.
- Kugler, K.G., Siegwart, G., Nussbaumer, T., Ametz, C., Spannagl, M., Steiner, B., Lemmens, M., Mayer, K.F.X., Buerstmayr, H. and Schweiger, W. 2013. Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (*Triticum aestivum* L.), *BMC Genomics*. 14(728).
- Kumar, S. V. and Wigge, P. A. 2010. H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis, *Cell*. 140(1), pp. 136-147.
- Kurishbayev, J. S. *et al.*, 2020. Green Revolution ‘Stumbles’ In A Dry Environment: Dwarf Wheat With Rht Genes Fails To Produce Higher Grain Yield Than Taller Plants Under Drought, *Plant, Cell and Environment*. 43(10), pp. 2355-2364.
- KWS 2023. Sowing wheat - Information about seed rates, seed dates and soil preparation. <https://www.kws.com/gb/en/consulting/sowing/sowing-wheat> accessed 01/09/2023. Ladd, M. and Bordoni, B. 2023. Genetics, Transposons, *StatPearls Publishing*.
- Lallemand, T., Leduc, M., Landès, C., Rizzon, C. and Lerat, E. and 2020. An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice, *Genes*. 11(9), p. 1046.
- Landrieu, I. Hassan, S., Sauty, M., Dewitte, F., Wieruszeski, J.M., Inzé, D., De Veylder, L. and Lippens, G. 2004. Characterization of the Arabidopsis thaliana Arath; CDC25 dual-specificity tyrosine phosphatase, *Biochemical and Biophysical Research Communications*. 322(3), pp. 734-9.
- Le, N.Q.K., Yapp, E.K.Y., Nagasundaram, N. and Yeh, H.Y. 2019. Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams, *Frontiers in Bioengineering and Biotechnology*. 7().
- Leach, L.J., Belfield, E.J., Jiang, C., Brown, C., Mithani, A. and Harberd N.P. 2014. Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat, *BMC Genomics*. 15(276).

- Lenssen, N., Schmidt, G.A., Hansen, J.E., Menne, M.J., Persin, A., Ruedy, R. and Zyss, D. 2019: Improvements in the GISTEMP uncertainty model, *Journal of Geophysical Research*. 124(12), pp. 6307-6326.
- Levy, A.A and Feldman, M. 2022. Evolution and origin of bread wheat, *The Plant Cell*. 34(7), pp. 2549-2567.
- Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., Yin, L., Zhang, R., Wu, L., Zheng, Y. and Mao, L. 2014. mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat, *The Plant Cell*. 26(5), pp. 1878–1900.
- Li, H.Z. Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., Yin, L., Zhang, R., Wu, L., Zheng, Y. and Mao, L. 2013. Evaluation of Assembly Strategies Using RNA-Seq Data Associated with Grain Development of Wheat (*Triticum aestivum* L.), *PLOS One*.
- Li, M., Song, B., Zhang, Q., Liu, X., Lin, Y., Ou, Y., Zhang, H. and Liu, J. 2013. A synthetic tuber-specific and cold-induced promoter is applicable in controlling potato cold-induced sweetening. *Plant Physiology and Biochemistry*. 67(), pp. 41-47.
- Li, Q., Zheng, Q., Shen, W., Cram, D., Fowler, D.B., Wei, Y. and Zou, J. 2015. Understanding the Biochemical Basis of Temperature-Induced Lipid Pathway Adjustments in Plants, *The Plant Cell*. 27(1), pp. 86–103.
- Li, Y.Y., Yu, H., Guo, Z.M., Guo, T.Q., Tu, K. and Li, Y.X. 2006. Systematic Analysis of Head-to-Head Gene Organization: Evolutionary Conservation and Potential Biological Relevance, *PLOS Computational Biology*.
- Liu, B.C., Chen, J. and Shen, B. 2010. Genome-wide Analysis of the Transcription Factor Binding Preference of Human Bidirectional Promoters and Functional Annotation of the Related Gene Pairs, *Computational Systems Biology*. 13(), pp. 81-92.
- Liu, B., Chen, J. and Shen, B. 2011. Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs, *BMC Systems Biology*. 5(1).
- Liu, H., Sachidanandam, R. and Stein, L. 2001. Comparative Genomics Between Rice and Arabidopsis Shows Scant Collinearity in Gene Order, *Genome Research*. 11(12), pp. 2020–2026.
- Liu, X., Zhou, X., Li, Y., Tian, J., Zhang, Q., Li, S., Wang, L., Zhao, J., Chen, R. and Fan, Y. 2014. Identification and functional characterization of bidirectional gene pairs and their intergenic regions in maize, *BMC Genomics*. 15(338).
- Liu, Y., Wang, L.M., Zhao, L.Z., Wang, W. and Zhangcorresponding, H.X. 2021. Genome-Wide Identification and Evolutionary Analysis of Argonaute Genes in Hexaploid Bread Wheat, *BioMed Research International*. Published online.

- Liu, Y and Wang, W. 2021. Characterization of the GRAS gene family reveals their contribution to the high adaptability of wheat, *Bioinformatics and Genomics*. 11(5), p. 553.
- Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y. and Sun, Q. 2015. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.), *BMC Plant Biology*. 15(152).
- Liu, Z., Qin, J., Tian, X., Xu, S., Wang, Y., Li, H., Wang, X., Peng, H., Yao, Y., Hu, Z., Ni, Z., Xin, M. and Sun, Q. 2017. Global profiling of alternative splicing landscape responsive to drought, heat and their combination in wheat (*Triticum aestivum* L.), *Plant Biotechnology Journal*. 25(10), pp. 3640-3656.
- Long, M., VanKuren, N.W., Chen, S. and Vibranovski, M.D. 2013. New Gene Evolution: Little Did We Know, *HHS Author Manuscripts*. 47(), pp. 307–333.
- Lopez, M.D., Guerra, J.J.M. and Samuelsson, T. 2010. Analysis of Gene Order Conservation in Eukaryotes Identifies Transcriptionally and Functionally Linked Genes, *Plos One*. 5(5).
- Love, A.J., Yu, C., Petukhova, N.V., Kalinina, N.O., Chen, J. and Talianskya, M.E. 2017. Cajal bodies and their role in plant stress and disease responses, *RNA Biology*. 14(6), pp. 779–790.
- Lu, F.H., McKenzie, N., Gardiner, L.J., Luo, M.C., Hall, A. and Bevan M.W. 2020. Reduced chromatin accessibility underlies gene expression differences in homologous chromosome arms of diploid *Aegilops tauschii* and hexaploid wheat, *Gigascience*. 9(6), Published online.
- Lu, Q., Guo, F., Xu, Q. and Cang, J. 2019. LncRNA improves cold resistance of winter wheat by interacting with miR398, *Functional Plant Biology*. 47(6), pp. 544-557.
- Ma, J., Guo, F., Xu, Q. and Cang, J. 2023. Genome-wide characterization of the VQ genes in Triticeae and their functionalization driven by polyploidization and gene duplication events in wheat, *International Journal of Biological Macromolecules*. 243().
- Ma, J.H., Yuan, M., Sun, B., Zhang, D., Zhang, J., Li, C., Shao, Y., Liu, W. and Jiang, L. 2021. Evolutionary Divergence and Biased Expression of NAC Transcription Factors in Hexaploid Bread Wheat (*Triticum aestivum* L.), *Plants*. 10(2), p. 382.
- Ma, Z., Li, M., Zhang, H., Zhao, B., Liu, Z., Duan, S., Meng, X., Li, G. and Guo, X. 2023. Alternative Splicing of TaHsfA2-7 Is Involved in the Improvement of Thermotolerance in Wheat, *International Journal of Molecular Sciences*. 24(2), pp. 1014-1028.
- Ma J., Stiller, J., Zhao, Q., Feng, Q., Cavanagh, C., Wang, P., Gardiner, D., Choulet, F., Feuillet, C., Zheng, Y.L., Wei, Y., Yan, G., Han, B., Manners, J.M. and Liu, C. 2014. Transcriptome and Allele Specificity Associated with a 3BL Locus for Fusarium Crown Rot Resistance in Bread Wheat, *PLOS One*. 9(11).

- Maccaferri, M., Bruschi, M. and Tuberosa, R. 2022. Sequence-Based Marker Assisted Selection in Wheat, *Wheat Improvement*. Springer, Cham. pp. 513–538.
- Martín, A.C., Borrill, P., Higgins, J., Alabdullah, A., Ramírez-González, R.H., Swarbreck, D., Uauy, C., Shaw, P. and Moore, G. 2018. Genome-Wide Transcription During Early Wheat Meiosis Is Independent of Synapsis, Ploidy Level, and the Ph1 Locus, *Frontiers in Plant Science*. 9().
- Martinez, A.F., Lister, C., Freeman, S., Ma, J., Berry, S., Wingen, L. and Griffiths, S. 2021. Resolving a QTL complex for height, heading, and grain yield on chromosome 3A in bread wheat.
- Mathur, S., Agrawal, D. and Jajoo, A. 2014. Photosynthesis: Response to high temperature stress, *Journal of Photochemistry and Photobiology*. 137(), pp. 116-126.
- Minx, P. *et al.*, 2005. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species, *Genome Research*. 15(), pp. 1284-1291.
- Mohan, A., Schillinger, W.F. and Gill, K.S. 2013. Wheat Seedling Emergence from Deep Planting Depths and Its Relationship with Coleoptile Length, *PLOS One*. 8(9).
- Molina, C. and Grotewold, E. 2005. Genome wide analysis of Arabidopsis core promoters, *BMC Genetics*. 6(25), published online.
- Nakashima, K., Takasaki, H., Mizoi, J., Shinozaki, K. and Yamaguchi-Shinozaki, K. 2012. NAC transcription factors in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 1819(2), pp. 97-103.
- Narayanan, S., Tamura, P. J., Roth, M.R., Prasad, P.V.V. and Welti, R. 2016. Wheat leaf lipids during heat stress: I. High day and night temperatures result in major lipid alterations, *Plant, Cell, and Environment*. 39(4), pp. 787–803.
- Nievola, C.C., Carvalho, C.P., Carvalho, V. and Rodrigues, E. 2017. Rapid responses of plants to temperature changes, *Temperature: Medical Physiology and Beyond*. 4(4), pp. 371–405.
- Nutzmann, H.W. and Osbourn, A. 2014. Gene clustering in plant specialized metabolism, *Current Opinion in Biotechnology*, 26(), pp. 91-99.
- Oono, Y., Kobayashi, F., Kawahara, Y., Yazawa, T., Handa, H., Itoh, T. and Matsumoto, T. 2013. Characterisation of the wheat (*triticum aestivum* L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat, *BMC Genomics*. 14(77).
- Ortiz, R., Sayre, K.D., Govaerts, B., Gupta, R., Subbarao, G.V., Ban, T., Hodson, D., Dixon, J.M., Ortiz-Monasterio, J.I., Reynolds, M. 2008. Climate change: Can wheat beat the heat?, *Agriculture, Ecosystems & Environment*. 126(1–2), pp. 46-58.

- Parhad, S.S. and Theurkauf, W.E. 2019. Rapid evolution and conserved function of the piRNA pathway, *Open Biology*. 9().
- Pearce, S., Huttly, A.K., Prosser, I.M., Li, Y., Vaughan, S.P., Gallova, B., Patil, A., Coghill, J.A., Dubcovsky, J., Hedden, P. and Phillips, A.L. 2015. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family, *BMC Plant Biology*. 15(130) .
- Pfeifer, M., Kugler, K.G., Sandve, S.R., Zhan, B., Rudi, H., Hvidsten, T.R., IWGSC, Mayer, K.F.X. and Olsen, O.A. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat, *Science*. 345(6194).
- Pinthus, M.J. 1967. Evaluation of winter wheat as a source of high yield potential for the breeding of spring wheat, *Euphytica*. 16(), pp. 231–251.
- Pinthus, M.J. and Abraham, M. 1996. Effects of light, temperature, gibberellin (GA3) and their interaction on coleoptile and leaf elongation of tall, semi-dwarf and dwarf wheat, *Plant Growth Regulation*. 18(), pp. 239–247.
- Plass, C. and Rush, L.J. 2005. CpG Islands, *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*. pp. 344–346.
- Polson, A., Durrett, E. and Reisman, D. 2011. A bidirectional promoter reporter vector for the analysis of the p53/WDR79 dual regulatory element, *Plasmid*. 66(3), pp. 169-179.
- Ponomarenko, M. and Kolchanov, N. 2013. Heat Shock Proteins, *Brenner's Encyclopedia of Genetics* (Second Edition).
- Powell, J.J., Carere, J., Fitzgerald, T.L., Stiller, J., Covarelli, L., Xu, Q., Gubler, F., Colgrave, M.L., Gardiner, D.M., Manners, J.M., Henry, R.J. and Kazan, K. 2016. The Fusarium crown rot pathogen *Fusarium pseudograminearum* triggers a suite of transcriptional and metabolic changes in bread wheat (*Triticum aestivum* L.), *Annals of Botany*. 119(5), pp. 853–867.
- Putelat, T., Whitmore, A.P., Senapati, N. and Semenov, M.A. 2021. Local impacts of climate change on winter wheat in Great Britain, *The Royal Society*. 8().
- Quan, C., Chen, G., Li, S., Jia, Z., Yu, P., Tu, J., Shen, J., Yi, B., Fu, T., Dai, C. and Ma, C. 2022. Transcriptome shock in interspecific F1 allotriploid hybrids between Brassica species, *Journal of Experimental Botany*. 73(8), pp. 2336-2353.
- Rajeevkumar, S., Anunanthini, P. and Sathishkumar, R. 2015. Epigenetic silencing in transgenic plants, *Frontiers in Plant Science*. 6(693).
- Ramirez-Gonzalez, R. *et al.*, 2018. The transcriptional landscape of hexaploid wheat across tissues and cultivars, *Science*. 361(6403).
- Ray, D.K., Ramankutty, N., Mueller, N.D., West, P.C. and Foley, J.A. 2012. Recent patterns of crop yield growth and stagnation, *Nature Communications*. 3(1293).

Rebetzke, G.J., Richards, R.A., Fischer, V.M. and Mickelson, B.J. 1999. Breeding long coleoptile, reduced height wheats, *Euphytica*. 106(), pp. 159–168.

Rebetzke, G.J., Richards, R.A., Fettell, M., Long, A.G., Condon, R.I., Forrester, T.L. and Botwright. 2007. Genotypic increases in coleoptile length improves stand establishment, vigour and grain yield of deep-sown wheat, *Field Crops Research*. 100(1), pp. 10-23.

Ren, K. 2021. `_rlist: A Toolbox for Non-Tabular Data Manipulation_`. R package version 0.4.6.2.

Riaz, M.W., Lu, J., Shah, L., Yang, L., Chen, C., Mei, X.D., Xue, L., Manzoor, M. A., Abdullah, M., Rehman, S., Si, H., and Ma, C. 2021. Expansion and Molecular Characterization of AP2/ERF Gene Family in Wheat (*Triticum aestivum* L.), *Frontiers in Genetics*. 12().

Ristic, Z., Bukovnik, U., and Vara Prasad, P.V. 2007. Correlation between Heat Stability of Thylakoid Membranes and Loss of Chlorophyll in Winter Wheat under Heat Stress, *Crop Science*. 47(5), pp. 2067-2073.

Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II, *Trends in Biochemical Science*. 21(9), pp. 327-335.

Rost, B. 1999. Twilight zone of protein sequence alignments, *Protein Engineering Design and Selection*. 12(2), pp. 85–94.

Rothamsted Research. 2021. New Targets For Crop Genetic Improvement Found. 23 October. [Online]. [Accessed 4 July 2023]. Available from: <https://www.rothamsted.ac.uk/news/new-targets-crop-genetic-improvement-found#PUBLICATION->

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

Rudd, J.J., Kanyuka, K., Hassani-Pak, K., Derbyshire, M., Andongabo, A., Devonshire, J., Lysenko, A., Saqi, M., Desai, N.M., Powers, S.J., Hooper, J., Ambroso, L., Bharti, A., Farmer, A., Hammond-Kosack, K.E., Dietrich, R.A. and Courbot, M. 2015. Transcriptome and Metabolite Profiling of the Infection Cycle of *Zymoseptoria tritici* on Wheat Reveals a Biphasic Interaction with Plant Immunity Involving Differential Pathogen Chromosomal Contributions and a Variation on the Hemibiotrophic Lifestyle Definition, *Plant Physiology*. 167(3), pp. 1158–1185.

Ruibal-Mendieta, N.L., Rozenberg, R., Delacroix, D.L., Petitjean, G., Dekeyser, A., Baccelli, C., Marques, C., Delzenne, N.M., Meurens, M., Habib-Jiwan, J.L. and Quetin-Leclercq, J. 2004. Spelt (*Triticum spelta* L.) and winter wheat (*Triticum aestivum* L.) wholemeals have similar sterol profiles, as determined by quantitative liquid chromatography and mass spectrometry analysis, *Journal of Agricultural and Food Chemistry*.

- Rutherford, K. and Van Duyne, G.D. 2013. DNA Sequence Recognition by Proteins, *Encyclopedia of Biological Chemistry (Second Edition)*. Pp. 149-153.
- Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegue, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M. and Feuillet, C. 2008. Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution, *The Plant Cell*. 20(1), pp. 11–24.
- Sandhu, D and Gill, K.S. 2002. Gene-containing regions of wheat and the other grass genomes, *Plant Physiology*. 128(3), pp. 803-811.
- Schilling, S., Kennedy, A., Pan, S., Jermiin, L.S. and Melzer, R. 2020. Genome-wide analysis of MIKC-type MADS-box genes in wheat: pervasive duplications, functional conservation and putative neofunctionalization, *New Phytologist*. 225(1), pp. 511-529.
- Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M. and Stoeckert Jr, C.J. 2005. Promoter features related to tissue specificity as measured by Shannon entropy, *Genome Biology*. 6().
- Schweiger, W., Steiner, B., Vautrin, S., Nussbaumer, T., Siegwart, G., Zamini, M., Jungreithmeier, F., Gratl, V., Lemmens, M., Mayer, K.F.X., Bérgeès, H., Adam, G., Buerstmayr, H. 2016. Suppressed recombination and unique candidate genes in the divergent haplotype encoding Fhb1, a major Fusarium head blight resistance locus in wheat, *Theoretical and Applied Genetics*. 129(), pp. 1607-1623.
- Seifert, F., Bössow, S., Kumlehn, J., Gnad, H. and Scholten, S. 2016. Analysis of wheat microspore embryogenesis induction by transcriptome and small RNA sequencing using the highly responsive cultivar “Svilena”, *BMC Plant Biology*. 16(97).
- Shah, M.M. and Hassan, A. 2005. Distribution of genes and recombination on wheat homoeologous group 6 chromosomes: a synthesis of available information, *Molecular Breeding*. 15(), pp. 45–53.
- Shahbandeh, M. 2023. Global wheat production from 1990/1991 to 2022/2023, *Statistica*. <https://www.statista.com/statistics/267268/production-of-wheat-worldwide-since-1990/>. Accessed 05-05-2023.
- Sharon, E. Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A. and Segal, E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters, *Nature*, 30(), pp. 521–530.
- Sidhu, D. and Gill, K. S. 2005. Distribution of genes and recombination in wheat and other eukaryotes, *Plant Cell, Tissue and Organ Culture*. 79(), pp. 257–270.
- Singh, K., Shukla, S., Kadam, S., Semwal, V.K., Singh, N.K. and Khanna-Chopra, R. 2015. Genomic regions and underlying candidate genes associated with coleoptile length under

deep sowing conditions in a wheat RIL population, *Journal of Plant Biochemistry and Biotechnology*. 24(), pp. 324–330.

Singh, R.K., Prasad, A., Maurya, J. and Prasad, M. 2021. Regulation of small RNA-mediated high temperature stress responses in crop plants, *Plant Cell Reports*. 41(), pp. 765–773.

Spears, B.J., McInturf, S.A., Collins, C., Chlebowski, M., Cseke, L.J., Su, J., Mendoza-Cózatl, D.G. and Gassmann, W. 2022. Class I TCP transcription factor AtTCP8 modulates key brassinosteroid-responsive genes, *Plant Physiology*. 190(2), pp.1457–1473.

Subrahmanian, N., Remacle, C. and Hamel, P.P. 2016. Plant mitochondrial Complex I composition and assembly: A review, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. 1857(7), pp. 1001-1014.

Sukhikh, I.S., Vavilova, V.J., Blinov, A. and Goncharov, N.P. 2021. Diversity and Phenotypical Effect of Allelic Variants of Rht Dwarfing Genes in Wheat, *Russian Journal of Genetics*. 57(), pp. 127-138.

Sung-II, L. and Nam-Soo, K. 2014. Transposable Elements and Genome Size Variations in Plants, *Genomics and Informatics*. 12(3), pp. 87-97.

Tanner N.K. and Linder, P. 2001. DExD/H Box RNA Helicases. 8(2), pp. 251-262.

Tessema, B.B., Liu, H., Sørensen, A.C., Andersen, J.P. and Jensen, J. 2020. Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat, *Frontiers in Genetics*. 11(). Pp.

The Observatory of Economic Complexity, 2023. Wheat, <https://oec.world/en/profile/hs/wheat>. Accessed 05-05-2023.

Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.P. and Mi, H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all, *Protein Science*. 31(1), pp. 8-22.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otilar, R.P. and Myers, R.M. 2004. An Abundance of Bidirectional Promoters in the Human Genome, *Genome Research*. 4(1), pp. 62–66.

Tsompana, M. and Buck, M. J. 2014. Chromatin accessibility: a window into the genome, *Epigenetics and Chromatin*. 7(33), pp.

Van der Velde, M., Baruth, B., Bussay, A., Ceglar, A., Garcia Condado, S., Karetos, S., Lecerf, R., Lopez, R., Maiorano, A., Nisini, L., Seguni, L. and van den Berg, M. 2018. In-season performance of European Union wheat forecasts during extreme impacts, *Nature Portfolio*. Published online.

Van Rossum, G. and Drake, F. L. 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

- Villoutreix, R. Ayala, D., Joron, M., Gompert, Z., Feder, J.L. and Nosil, P. 2021. Inversion breakpoints and the evolution of supergenes, *Molecular Ecology*.
- Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.H., Liu, T. and Paterson, A.H. 2015. Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events, *Molecular Plant*. 8(6), pp. 885-898.
- Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S.P., Feltus, F.A. and Paterson, A.H. 2011. Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms, *Plos One*. 6(12).
- Wei, N., Zhang, S., Liu, Y., Wang, J., Wu, B., Zhao, J., Qiao, L., Zheng, X., Wang, J. and Zheng, J. 2022. Genome-wide association study of coleoptile length with Shanxi wheat, *Crop and Product Physiology*. 13().
- Wickham, H. 2022. `_stringr`: Simple, Consistent Wrappers for Common String Operations_. R package version 1.4.1. <<https://CRAN.R-project.org/package=stringr>>.
- Wickham H, Girlich M (2022). `_tidyr`: Tidy Messy Data_. R package. version<<https://CRAN.R-project.org/package=rlist>>. 1.2.1, <<https://CRAN.R-project.org/package=tidyr>>.
- Xu, D., Hao, Q., Yang, T., Lv, X., Qin, H., Wang, Y., Jia, C., Liu, W., Dai, X., Zeng, J., Zhang, H., He, Z., Xia, X., Cao, S., and Ma, W. 2023. Impact of “Green Revolution” gene Rht-B1b on coleoptile length of wheat, *Frontiers in Plant Science*. 14().
- Xu, W., Li, Y., Cheng, Z., Xia, G. and Wang, M. 2016. A wheat histone variant gene TaH2A.7 enhances drought tolerance and promotes stomatal closure in Arabidopsis, *Plant Cell Reports*. 35(9), pp. 1799-1826.
- Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., Steinmetz, L.M. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature*. 457(7232), pp. 1033–1037.
- Yamamoto, Y.Y. Ichida, H., Abe, T. and Suzuki, Y. 2007. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis, *Nucleic Acids Research*. 35(18), pp. 6219–6226.
- Yamauchi, Y., Kimura, Y., Akimoto, S., Marutani, Y., Mizutani, M. and Sugimoto, Y. 2011. Plants switch photosystem at high temperature to protect photosystem II, *Nature Proceedings*.
- Yang, F., Li, W. and Jørgensen, H.J.L. 2013. Transcriptional Reprogramming of Wheat and the Hemibiotrophic Pathogen *Septoria tritici* during Two Phases of the Compatible Interaction, *PLOS One*.

- Yang, L. and Yu, J. 2009. A comparative analysis of divergently-paired genes (DPGs) among *Drosophila* and vertebrate genomes. *BMC Evolutionary Biology*, (9)55.
- Yang, Z., Peng, Z., Wei, S., Liao, M., Yu, Y. and Jang, Z. 2015. Pistillody mutant reveals key insights into stamen and pistil development in wheat (*Triticum aestivum* L.), *BMC Genomics*. 16(211).
- Yang W., Liu, J., Huang, B., Xu, Y.M., Li, J., Huang, L.F., Lin, J., Zhang, J., Min, Q.H., Yang, W.M. and Wang, X.Z. 2014. Mechanism of alternative splicing and its regulation, *Biomedical Reports*. 3(2), pp. 152–158.
- Yao Y., Ni, Z., Peng, H., Sun, F., Xin, M., Sunkar, R., Zhu, J.K. and Sun, Q. 2010. Non-coding small RNAs responsive to abiotic stress in wheat (*Triticum aestivum* L.), *Functional and Integrative Genomics*. 10(), pp. 187–190.
- Ye, J., Yang, X., Hu, G., Liu, Q., Li, W., Zhang, L., and Song, X. 2020. Genome-Wide Investigation of Heat Shock Transcription Factor Family in Wheat (*Triticum aestivum* L.) and Possible Roles in Anther Development, *International Journal of Molecular Studies*. 21(2), p. 608.
- Yokotani, N. Ichikawa, T., Kondou, Y., Matsui, M., Hirochika, H., Iwabuchi, M. and Oda, K. 2007. Expression of rice heat stress transcription factor OsHsfA2e enhances tolerance to environmental stresses in transgenic *Arabidopsis*, *Planta*. 227(), pp. 957–967.
- Zeng, X., Tang, R., Guo, H., Ke, S., Teng, B., Hung, Y.H., Xu, Z., Xie, X.M., Hsieh, T.F. and Zhang, X.Q. 2017. A naturally occurring conditional albino mutant in rice caused by defects in the plastid-localized OsABC18 transporter, *Plant Molecular Biology*. 94(1-2).
- Zhang, H., Yang, Y., Wang, C., Liu, M., Li, H., Fu, Y., Wang, Y., Nie, Y., Liu, X. and Ji, W. 2014. Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew, *BMC Genomics*. 15(898).
- Zhang, P., Wu, W., Chen, Q. and Ming Chen. 2019. Non-Coding RNAs and their Integrated Networks, *Journal of Integrative Bioinformatics*. 16(3), published online.
- Zhang, X., Gonzalez-Carranza, Z.H., Zhang, S., Miao, Y., Liu, C.J. and Roberts, J.A. 2019. F-box proteins in plants, *Annual Plant Reviews*. 2(), pp. 1-21.
- Zhang, Y., Li, Z., Zhang, Y., Lin, K., Peng, Y., Ye, L., Zhuang, Y., Wang, M., Xie, Y., Guo, J., Teng, W., Tong, Y., Zhang, W., Xue, Y., Lang, Z. and Zhang, Y. 2021. Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements, *Genome Research*. 31(12), pp. 2276–2289.
- Zhao, Y., Broholm, S.K., Wang, F., Rijpkema, A.S., Lan, T., Albert, V.A., Teeri, T.H., Elomaa, P. 2020. TCP and MADS-Box Transcription Factor Networks Regulate Heteromorphic Flower Type Identity in *Gerbera hybrida*. *Plant Physiology*. 184(3), pp. 1455-1468.

Zhou, H., Liu, Y., Liang, Y., Zhou, D., Li, S., Lin, S., Dong, H. and Huang, L. 2020. The function of histone lysine methylation related SET domain group proteins in plants, *Protein Science*. 29(5), pp. 1120–1137.

Zuo, J., Wu, Z., Li, Y., Shen, Z., Feng, X., Zhang, M. and Ye, H. 2017. Mitochondrial ABC Transporter ATM3 Is Essential for Cytosolic Iron-Sulfur Cluster Assembly¹, *Plant Physiology*. 173(4), pp. 2096–2109.