Exploring wild relative introgressions in wheat and their impact on genomic methodology

Benedict Coombes

This thesis is submitted for the degree of Doctor of Philosophy

University of East Anglia (UEA)

Earlham Institute

February 2024

Abstract

Deploying novel genetic variation, such as wild relative introgressions, into wheat breeding programmes could help to satisfy the global demands of wheat despite a rising population and a changing climate. Sequencing data, which has become much cheaper and more accessible over time, can play an important role by being used to identify and characterise introgressions in wheat that confer beneficial traits and can be deployed into breeding programmes. This thesis offers insights into wheat introgressions in the context of sequencing data, exploring how sequencing data can be used to detect and characterise introgressions and also how introgressions can interfere with the accurate processing of sequencing data in common genomic analyses.

First, I used whole-genome sequencing data to characterise a set of hexaploid wheat/*Ambylopyrum muticum* introgression lines to a high resolution, identifying introgressions and other structural changes in the lines. I then combined the sequencing data with rust resistance phenotype data to demonstrate how the region of introgressed genes underlying the phenotype can be identified and candidate genes proposed.

I then present findings on an important heat tolerance phenotype identified in wheat that is driven by three marker trait associations that together increase yield by over 50% under heat stress conditions. Using sequencing data, I discovered that one of these is driven by an *Aegilops tauschii* introgression. I then searched for candidate genes in multiple *Ae. tauschii* genomes, exposing the limits of relying on a single reference genome.

Finally, I found that the abundant introgressions across wheat accessions cause inaccurate RNA-seq read alignment that compromises research findings by leading to the underestimation of gene expression and the expression balance categories of triads being incorrectly assigned. To address this, I proposed a solution in which transcripts from multiple wheat cultivars are integrated into a pantranscriptome reference to use for RNA-seq read alignment.

Contents

List of Figures

List of Tables

Acknowledgements

The work presented in this thesis was supported by a Doctoral Training Partnership (DTP) PhD studentship from the Biotechnology and Biological Sciences Research Council (BBSRC). This research was also supported in part by the Research Computing at NBI through extensive use of the high-performance computing (HPC) cluster.

I would like to thank my supervisor Anthony Hall for his support and guidance over the last four years. He trusted in my abilities and ideas and gave me the freedom to explore research I found interesting or felt was likely to be fruitful. He has always made me feel like more than a student and an integral member of the research group. I would also like to express my gratitude to all of my collaborators and co-authors, without whom much of the work presented in this thesis would not have been possible.

I would also like to thank all the members of the Hall group, past and present, for playing a role in my PhD journey, and providing not only support and feedback on my research but offering much needed humour and fun to balance the challenges and relentless nature of academic work. Specifically, I want to thank Ryan Joynson for getting me up to speed at the beginning of my PhD and teaching me the foundations of wheat genomics research, and Rachel Rusholme-Pilcher, who's friendship has been invaluable to me throughout my PhD, both professionally and personally.

Finally, I would like to thank my parents for their unremitting love and support, without which I can confidently say I would not be where I am today.

1 Introduction

1.1 Why crop yields need to be increased

A rising human population and rising per capita income is increasing agricultural consumption over time. UN population projections according to the medium fertility model suggest that world population will increase to 9.7 billion by 2050 and to 10.9 billion by 2100, at which point the population will stop growing (United Nations, Department of Economic and Social Affairs, Population Division, 2019). Global average per capita consumption is also estimated to increase from 2831 to 3129 kcal per day between 2009 and 2050 (Pardey *et al.*, 2014) as more countries move out of poverty towards industrialisation (Brisson *et al.*, 2010). The structure of consumption will also change (Shiferaw et al., 2013); for example, rising incomes in Asia are associated with a convergence towards a Western diet, with increased consumption of wheat and animal products (Pingali, 2007). The rapidly changing climate will place additional pressure on farmers and breeders and will necessitate the development of crop varieties that are more resilient to abiotic stressors such as drought and heat (Kahiluoto *et al.*, 2019), and to pests and pathogens whose range and lifestyle may change due to climate change (Garrett *et al.*, 2006; Classen *et al.*, 2015; Surówka, Rapacz and Janowiak, 2020). Furthermore, climate change, along with the overuse of agricultural land, is expected to reduce the amount of arable land over time (Zhang and Cai, 2011), making improvements to yield and environmental stability of crops even more important.

1.2 Wheat as an important component of global food production

Triticum aestivum (bread wheat) is among the three most grown and consumed crops in the world, alongside *Zea mays* (maize) and *Oryza sativa* (rice). It is the most widely grown, cultivated on 217 million hectares and is the third most highly produced crop at 752 million tons per year (Erenstein *et al.*, 2022). Around 20% of the calories and protein consumed globally each year are derived from wheat either through direct consumption or via animal feed (Reynolds *et al.*, 2012). Wheat is a staple crop for around 35% of the global population (Grote *et al.*, 2021). The global demand for wheat increased by fourfold between the 1960s and 2009 and doubled between 1980 and 2009 (Shiferaw *et al.*, 2013). This equates to an average increase in demand of 2.24% per year since the 1960s. Wheat is the largest agricultural commodity on the global market; 194.4 million tons of

wheat are projected to be traded in 2023/24 (FAO, 2023). Therefore, wheat production has a vital role to play in supporting a rising and developing population.

1.3 Strategies to safeguard future global wheat production

To meet the global demand for wheat production, a multifaceted strategy is essential. This will involve genetic improvements to traits such as yield, biotic and abiotic stress resistance/resilience, and reduced need for inputs such as water and fertiliser. It will also involve optimising agronomic management practices (Shiferaw et al., 2013) and minimising food wastage (Kummu *et al.*, 2012). Additionally, wider availability of new agricultural technology (Ruzzante, Labarta and Bilton, 2021) and increased varietal turnover, particularly in the developing world where farmers often face slow replacement cycles (Atlin, Cairns and Das, 2017), will be of great importance.

As genomic scientists, our contribution to this will naturally come from genetic improvement, and more specifically, through leveraging large genomic datasets to facilitate increased understanding and better utilisation of wheat germplasm resources. Genetic gains to wheat yield are currently in the region of 0.5-1.0% per year (Reynolds *et al.*, 2017). Maintaining, or ideally improving this, despite growing environmental and demographic pressures, will be an important contribution to future food security.

1.4 Genomic origin of bread wheat

Bread wheat, *Triticum aestivum*, is an allohexaploid species (BBAADD genomes); this means its genome consists of three subgenomes derived from independent diploid species. Around 0.7-0.8mya, a hybridisation event between the male donor of the A subgenome, *Triticum urartu* (AA) and the female donor of the B subgenome, a species likely extinct but thought to be closely related to, but distinct from, *Aegilops speltoides* (SS), formed tetraploid wild emmer wheat, *Triticum turgidum, ssp. diccocoides* (BBAA) (Fig. 1-1) (Levy and Feldman, 2022). Between 8500 and 9000 years ago, hybridisation between the female donor of the A and B subgenomes, *T. turgidum* ssp. *durum*, (BBAA), a domesticated form of emmer wheat, and the male donor of the D subgenome, *Aegilops tauschii* (DD), a diploid wild goatgrass, gave rise to hexaploid *T. aestivum* (Fig. 1-1) (Levy and Feldman, 2022). *T. aestivum* was soon domesticated, beginning the thousands of years of intense cultivation and artificial selection for agronomic and end-use traits (Venske *et al.*, 2019) resulting in the bread wheat we grow and eat today.

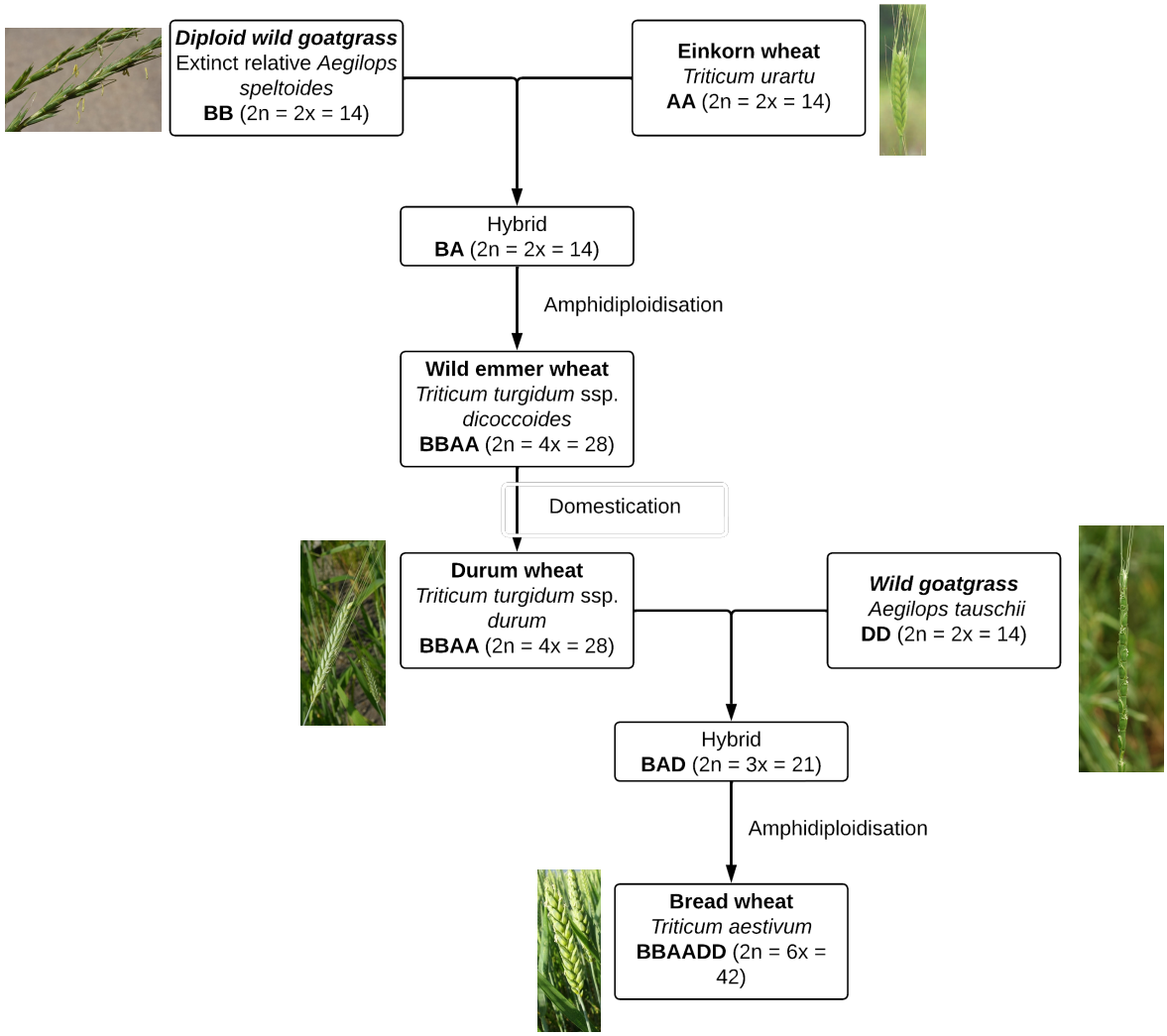


Figure 1-1. Simplified diagram of the polyploidisation events that led to the creation of *Triticum aestivum* (bread wheat).
Based on figure by Jauhar (2007) with updated information from Levy and Feldman (2022).

1.5 Limited genetic variation in modern wheat breeding material

Breeding new wheat varieties with higher yield, better end-use traits, resistance to pests and pathogens, and tolerance to abiotic stressors relies upon the presence of sufficient and appropriate genetic variation in the genepool that is accessible to breeders to incorporate into breeding programmes. All bread wheat grown today is thought to derive from just one or two rare hybridisation events between emmer wheat and *Ae. tauschii* (Charmet, 2011). This genetic bottleneck combined with intensive artificial selection has resulted in modern wheat material possessing less than a third of the nucleotide diversity seen in its wild progenitor species (Haudry *et al.*, 2007). The initial genetic diversity established in wheat upon its formation has been supplemented by mutation and sporadic hybridization events, primarily with wild populations of tetraploid wheat. In such

cases of hybridisation, almost all recombination takes place in the A and B subgenomes as the D genome has no homologous counterpart in tetraploid wheat. This has left the D subgenome with particularly low levels of genetic variation, around 16% of that of the A and B subgenomes (Yao Zhou *et al.*, 2020; Gaurav *et al.*, 2021).

The problem created by bottlenecking is compounded by pressure on breeders to prioritise advanced breeding material (Valkoun, 2001) for more rapid development of elite varieties that perform competitively and adhere to regulations about uniformity and quality (Cooper, Spillane and Hodgkin, 2001), limiting the introduction of genetic variation from external sources. Semi-dwarf, lodging-resistant varieties produced in the green revolution, although very high yielding, further limited genetic variation by creating a bottleneck in the Elite material that has been used by breeders since the 1960s (Sehgal *et al.*, 2015).

1.6 Novel sources of genetic variation for wheat breeding

The average yearly increase to wheat yield typically occurs through conventional breeding approaches. Minor effect genes are recombined through crossing Elite lines and selecting the best progeny based on important phenotypic traits, most important of which is usually grain yield (Sukumaran *et al.*, 2018; Reynolds *et al.*, 2020). Genetic markers can also be used to identify progressively beneficial allele combinations through genomic selection (Reynolds *et al.*, 2017). Additionally, major effect genes, encoding traits such as disease resistance, are identified and subsequently incorporated into varieties through controlled crosses between lines carrying the gene(s) of interest and breeding lines showing favourable agronomic trait profiles. Repeated backcrossing, using traditional phenotypic selection or modern marker-assisted selection then allows the gene of interest to be retained in a genetic background predominantly deriving from the line with more favourable agronomic characteristics (Tyagi *et al.*, 2014).

However, relying solely on Elite material as the source for these genes/alleles limits potential improvements to those that can be derived from the existing genetic variation present in the Elite breeding pool. This is particularly relevant in wheat due to its limited genetic diversity. To overcome this constraint, it is important to incorporate novel genetic variation into breeding programmes.

Such novel sources of genetic variation include landraces and wild and domesticated relatives of wheat. Landraces are locally adapted varieties of wheat, existing in genetically heterogeneous populations (Villa *et al.*, 2005) that have evolved under selection by farmers in local farming systems (Vikram *et al.*, 2016) and have not been through intensive selection by breeders for particular agronomic characteristics (Lopes et al., 2015). Landraces typically possess high tolerance to abiotic and biotic stresses and intermediate yield in a low input agricultural system (Zeven, 1998) and, due to being locally adapted and maintained rather than developed and distributed globally from a narrow collection of Elite lines, as a collection they contain far more genetic variation than modern Elite varieties (Reif *et al.*, 2005; Wingen *et al.*, 2014; Winfield *et al.*, 2016), much of which was left behind following the Green Revolution (Cseh *et al.*, 2019).

Wheat's wild relatives have not undergone the same genetic constraints as domesticated wheat. They haven't faced the intense selection in breeding programmes, and most have not faced the genetic bottlenecks of polyploidisation. Furthermore, they have undergone selection in a variety of environments in the presence of different abiotic and biotic selection pressures. Introducing genetic variation from wild relatives thus has the potential to bolster slowing gains and introduce novel resistance/tolerance phenotypes to pests and pathogens and abiotic stress (Valkoun, 2001; Nevo and Chen, 2010; Placido *et al.*, 2013; Zhang *et al.*, 2017; Cruppe *et al.*, 2019; He *et al.*, 2019; Fellers *et al.*, 2020; Narang *et al.*, 2020; Li *et al.*, 2022). Domesticated relatives of wheat, such as Rye, also offer valuable sources of genetic variation that is novel to wheat (Nkongolo *et al.*, 1992; Ren *et al.*, 2009; Yang *et al.*, 2009; Bertholdsson, Andersson and Merker, 2012; Crespo-Herrera *et al.*, 2013; Moskal *et al.*, 2021). Despite also having undergone selection pressure, they have mostly had experienced less severe genetic bottlenecks than bread wheat.

The relatives of wheat can be categorised by whether they belong to the primary, secondary, or tertiary genepool of wheat. Definitions of these genepools have changed over time and still vary between researchers (Ortiz *et al.*, 2008) but in this thesis they will be defined as follows. Species belonging to the primary genepool have a homologous subgenome in wheat for each of their own subgenomes. Those from the secondary genepool have a homologous subgenome in wheat for at least one, but not all, of its own subgenomes. Those from the tertiary genepool share no homologous genomes with wheat (Fig. 1-2). Despite the genetic distance, even members of the tertiary genepool can

be introgressed into wheat, although introgressing more distant relatives is more challenging and requires more sophisticated techniques.

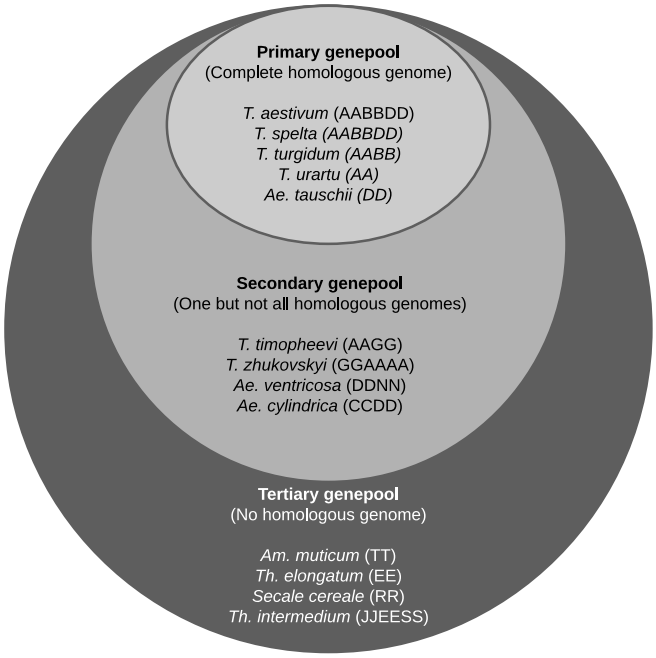


Figure 1-2. Definition of wheat's primary, secondary and tertiary genepools. Listed species are examples of members of each genepool. Subgenomes possessed by each species are in brackets.

1.7 History of incorporating wild relative variation in wheat

In the 1920s through to the 1940s Nikolai Vavilov, a Russian botanist and phytogeographer, collected seeds from wild wheat species from around the world to be preserved in the Leningrad Seedbank, whose name has since been changed to the NI Vavilov Institute of Plant Industry (Vavilov, 1940; Tanksley and McCouch, 1997; Mitrofanova, 2012). Vavilov developed the concept of the centre of origin of crop plants (Vavilov, 1926; Hummer and Hancock, 2015) and suggested that the diversity of crop wild relatives would be greatest near to these centres of origin. He was among the first to emphasise the potential future value of collections of plant genetic resources (Dzyubenko, 2018) and his work led to the creation of international genebanks. Genebanks are repositories of plant genetic resources maintained ex situ through seed storage (Mascher *et al.*, 2019). They contain seeds from plants from around the world with a focus on diverse crop varieties such as landraces and crop wild relatives which are likely to possess genes and alleles that will be useful for crop improvement (Hoisington *et al.*, 1999). CIMMYT (International Maize and Wheat Improvement Center) and ICARDA

(International Center for Agricultural Research in the Dry Areas) are examples of such genebanks that aim to characterise underutilised genetic variation in crops and mobilise this variation into global breeding programmes (Sehgal *et al.*, 2015). CIMMYT focuses on maize and wheat, whereas ICARDA has a broader focus, including dryland cereals, legumes and forage and rangeland species.

Wheat wild relative introgression lines, which are wheat varieties containing a chromosomal segment from the genome of a wild relative, date back to 1939 at the University of Saskatchewan, Canada, where the *Sr26* resistance gene from *Thinopyrum ponticum* (then *Agropyron elongatum*) was introduced into wheat (Shebeski and Wu, 1952; Knott, 1961). The resultant introgression lines were worked on by (Shebeski and Wu, 1952) and by (Knott, 1961), which led to the release of the cultivar Eagle in Australia, the first commercial cultivar containing *Sr26* (Dundas *et al.*, 2015). Many commercial lines containing *Sr26* have since been developed and grown in Australia, although its use has declined over time, possibly due to the associated yield reduction conferred by the *Th. ponticum* chromosome segment (Dundas *et al.*, 2015).

In the 1950s, Ernest Sears pioneered techniques for more easily incorporating distant wild relatives into the wheat genome. In 1956, he introgressed *Aegilops umbellulata* into wheat, conferring leaf-rust resistance (Sears, 1956). This was achieved by crossing emmer wheat with *Ae. umbellulata* and crossing the amphidiploid progeny with *T. aestivum*. Since Sears, there have been numerous examples of introducing beneficial traits by transferring chromosome segments from wheat's relatives into wheat. These include introducing an Eyespot disease resistance gene from *Aegilops ventricosa* into wheat (Doussinault *et al.*, 1983); wheat streak mosaic virus resistance from *Agropyron intermedium* into wheat (Friebe *et al.*, 1996); the leaf rust resistance gene *Lr19* from *Agropyron elongautum* into wheat (Reynolds *et al.*, 2001); powdery mildew and stripe rust resistance from the wheat-rye 1RS-1BL translocation (Han *et al.*, 2020); and the *Ug99* stem rust resistance gene from *Ae. speltoides* into tetraploid wheat (Klindworth *et al.*, 2012).

Due to the increasing appreciation for the potential value of wild relative genetic variation, organisations around the world have set up programmes to incorporate such variation more systematically rather than relying on previously incorporated material or small-scale efforts to introgress specific genes. For example, over the last few decades,

CIMMYT has incorporated exotic material into their vast germplasm through strategic crosses with landraces, wild relative introgression lines and, most notably, synthetic-derived lines (Dreccer *et al.*, 2007; Ortiz *et al.*, 2008). Synthetic hexaploid wheat is produced by crossing tetraploid durum wheat with diploid *Ae. tauschii* (Fig. 1-3), replicating the natural polyploidisation event that led to the creation of *T. aestivum* (Dreisigacker *et al.*, 2008). This process acts as a bridge to incorporate diversity present in durum wheat and populations of wild *Ae. tauschii* into Elite lines. Primary synthetic lines are typically crossed into an Elite background to produce advanced synthetic derivatives (ASDs) that are subjected to phenotypic screening. Over 1000 synthetic-derived lines were generated by CIMMYT as of 2016 (Das *et al.*, 2016). In 2018, over 62 synthetic-derived lines had been registered as cultivars worldwide (Li *et al.*, 2018). CIMMYT synthetic-derived lines possess significantly greater genetic diversity than original green revolution wheat lines (Warburton *et al.*, 2006). ASDs are commonly found in the pedigree history of varieties that, in international CIMMYT nurseries, outperform local varieties under diverse conditions (Manès *et al.*, 2012), including under and extreme heat stress (Cossani and Reynolds, 2015). This approach has also been successful in introducing disease resistance traits (Zhu *et al.*, 2014, 2016; Shamanin *et al.*, 2019). Landrace and synthetic-derived lines have been developed in recent years for drought, heat and yield potential conditions (Reynolds *et al.*, 2017; Molero *et al.*, 2019; Rosyara *et al.*, 2019) and many have been identified to have superior biomass compared to Elite lines under drought and heat conditions (Lopes and Reynolds, 2011; Cossani and Reynolds, 2015).

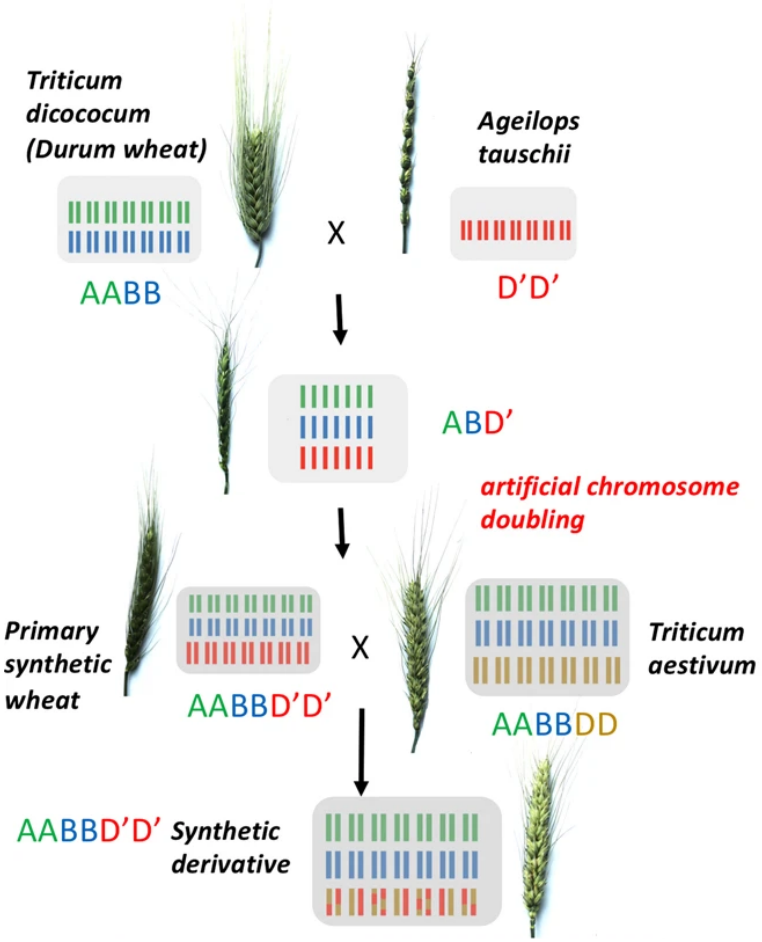


Figure 1-3. The process by which CIMMYT generate synthetic hexaploid wheat lines. Figure taken from Rosyara et al. (2019) with permissions granted by the Creative Commons Attribution 4.0 International License.

Another strategy is to systematically introgress entire wild relative genomes into wheat, including distant relatives belonging to the secondary and tertiary genepools. Researchers at the Wheat Research Centre at the University of Nottingham are pioneers of this approach. Through utilising recombination mutants and high-throughput genotyping methods, they create sets of introgression lines that possess most of a wild relative genome in variable, overlapping chromosomal and sub-chromosomal segments (King *et al.*, 2017, 2019). There are introgression lines currently available for *Ambylopyrum muticum*, *T. urartu*, *Ae. speltoides*, *Aegilops caudata*, *Aegilops comosa*, *Aegilops umbellulata*, *Thinopyrum bessarabicum*, *Secale anatolicum*, *Secale iranicum*, *Thinopyrum turcicum*, *T. turgidum*, and *Triticum timopheevii*. These introgression lines can be sent to researchers who phenotype them for different traits of interest. Segments conferring beneficial phenotypes can be further characterised and crossed into Elite varieties for deployment in breeding programmes.

1.8 Role of introgressions in wheat evolution and breeding

In addition to synthetically introduced introgressions, natural introgressions throughout wheat's history have played an important role in shaping the genetic diversity and adaptive potential of wheat, through the provision of novel genes and alleles. The extent of natural introgressions among wheat varieties has been the subject of several pieces of research.

For instance, He *et al.* (2019) conducted exome sequencing of 890 diverse wheat accessions, including landraces and cultivars, and identified abundant historic introgressions from wild emmer. They estimated that approximately 11.4% and 11.8% of the genome of each accession was composed of introgressions from wild emmer for cultivars and landraces, respectively. Introgressed regions exhibited elevated levels of genetic diversity and increased differentiation between accessions. Furthermore, the authors found that many of the introgressed regions displayed signatures of selection and have likely contributed to phenotypic variation and adaptation.

Cheng *et al.* (2019) analysed whole-genome sequencing (WGS) data from 93 accessions, including wheat landraces, wheat cultivars, and wheat relatives including wild emmer, *Ae. tauschii*, and durum wheat. They identified shared haplotypes between hexaploid wheat and various populations of wild emmer, which have heavily contributed to the genetic diversity in the A and B subgenomes of wheat and have likely introduced beneficial traits. In addition to introgressed haplotypes from wild emmer, they also detected introgressions from other wild relatives in all 63 bread wheat accessions studied. They highlighted specific introgressions that overlap with quantitative trait loci (QTLs) associated with important traits such as disease resistance and grain yield, indicating that these introgressions may have been under positive selection pressure as they confer important agronomic characteristics.

Zhou *et al.* (2020) analysed WGS data from all 25 subspecies of AA, BBAA and BBAADD genomes in the *Triticum* genus and two subspecies of *Ae. tauschii* (DD) to evaluate the proportion of bread wheat accession genomes that are composed of introgressions from diploid and tetraploid relatives from the primary genepool. They estimated that 4-32% of the bread wheat genome is composed of introgressions from populations of wild relatives from the primary genepool. Free threshing tetraploids and wild and domesticated emmer have significant representation in bread wheat accessions, compensating for the severe

genetic bottlenecks of hexaploidization and domestication. The majority of the gene flow from tetraploid relatives was found to be from free-threshing tetraploids in Europe and West Asia. 57% and 66% of the nucleotide diversity of wild emmer has been captured in wheat cultivars and landraces, respectively. This is in stark contrast with the 14% of nucleotide diversity of the D subgenome donor, *Ae. tauschii*, that is represented in bread wheat accessions.

Przewieslik-Allen *et al.* (2021) used genotyping data to construct and compare haplotype blocks of 358 wheat accessions and 113 wheat relatives from 44 species, covering the primary, secondary, and tertiary genepools. They classified near identical haplotype blocks between a wheat accession and a wheat relative as being an introgression. Using this methodology, they identified that 14.5-55.1% of the wheat accessions studied have evidence of introgression from tetraploid species. While not quite as extensive as tetraploid introgressions, introgressions from *Ae. tauschii* and species from the secondary and tertiary gene pools have made a sizeable contribution to wheat genomes and were associated with elevated levels of genetic diversity.

By comparing the total introgression size of each relative donor species between different collections of wheat accessions, grouped by release date, Przewieslik-Allen *et al.* (2021) found some interesting patterns. Gene flow from tetraploid species was more prominent in wheat accessions bred before 1960, whereas those bred after 1960 contained more exotic introgressions, including those from *Ae. tauschii* and species from the secondary and tertiary genepool. This reflects the change in breeding strategies following the green revolution, with a greater emphasis in introducing more diverse genetic variation into wheat. Notably, initiatives such as CIMMYT's synthetic wheat programme have accelerated the incorporation of *Ae. tauschii* genetic material into wheat accessions.

Several introgressions found in the chromosome-level genome assemblies generated as part of the 10+ wheat genomes project (Walkowiak *et al.*, 2020) have been well characterised. For example, there is a 33 Mbp telomeric *Ae. ventricosa* introgression on the distal end of the short arm of chr2A (Walkowiak *et al.*, 2020; Gao *et al.*, 2021; Keilwagen *et al.*, 2022). This particular introgression has been detected across several cultivars, including Jagger and Stanley. It confers wheat blast resistance (Cruz *et al.*, 2016) and contains other important resistance genes, including the *Lr37-Yr17-Sr38* gene cluster (Helguera *et al.*, 2003) that confers resistance against certain races of stripe, leaf, and

stem rust (Gao *et al.*, 2021); *Rkn3* that confers resistant against root-knot nematodes (Williamson *et al.*, 2013); and *Cre5* that confers resistance against pathotype Ha12 of the cereal cyst nematode (Jahier *et al.*, 2001). The frequency of this introgression in the CIMMYT spring wheat breeding programme, central USA regional winter wheat programs, and Kansas winter wheat germplasm has increased sizeably from the early 1990s to present day (Gao *et al.*, 2021), suggesting that the introgression confers traits under selection by breeders. The exception to this was in 2008-2010, where the introgression dropped in frequency, probably due to changes in virulence of the yellow rust pathogen leading to loss of *Yr17* resistance previously conferred by the introgression. Gao *et al.* (2021) also found that the presence of the introgression was associated with a small yield advantage. A very large introgression from *T. timopheevii* is found in the cultivar Lancer from the 10+ wheat genomes project. It spans most of chr2B (Watson-Haigh *et al.*, 2018; Walkowiak *et al.*, 2020; Keilwagen *et al.*, 2022) and contains the stem rust resistance gene *Sr36* (Bariana *et al.*, 2001; Chemayek *et al.*, 2017). There is also a *Th. ponticum* introgression at the distal end of the long arm of chr3D in Lancer (Walkowiak *et al.*, 2020; Keilwagen *et al.*, 2022). It contains the leaf rust resistance gene *Lr24* and the stem rust resistance gene *Sr24* (Walkowiak *et al.*, 2020).

1.9 Challenges associated with using wild relative introgressions for wheat improvement

Novel genetic variation from wheat's relatives that may be critical for future food security is often overlooked by commercial breeding companies due to the significant investment of time and resources required to incorporate unimproved genebank accessions into breeding programmes as parents (Atlin, Cairns and Das, 2017). Despite possessing genes conferring traits of interest for breeders, wild relative introgressions typically also possess genes that are deleterious in an agricultural setting and thus confer unfavourable phenotypes, a phenomenon known as linkage drag (Hao *et al.*, 2020).

This concern of linkage drag is compounded by the low recombination rates common within introgressions from distant relatives (McCouch *et al.*, 2020) caused by distant introgressed segments lacking a homologous chromosome in the gene pool with which it can recombine. To reduce the impact of linkage drag, introgressed segments can be broken up to retain the gene(s) of interest while removing deleterious introgressed genes, the feasibility of which has been demonstrated in several studies. Yasumuro *et al.*

(1981) induced homoeologous recombination to break up the *Th. Ponticum* chromosome segment in the Australian cultivar Eagle to separate *Sr26* from linked, deleterious genes. Similarly, (Khazan *et al.*, 2020), reduced the size of *Ae. sharonensis* introgressions while retaining leaf and stripe rust resistance genes.

1.10 The wheat genome and wheat genomic resources

As an allohexaploid (7n*3), the wheat genome is comprised of three independent subgenomes, A, B and D, which are genetically distinct but have a combinatorial effect on phenotype. Around 51.1% of the genes in wheat exist in triads (Ramírez-González *et al.*, 2018) which consist of three homoeologous genes, one belonging to each subgenome. Many other genes exist in more complex combinations of homoeologues, including dyads, where one homoeologue has been deleted, and tetrads, where one homoeologue has been duplicated (Juery *et al.*, 2021). The complexity of having three independent subgenomes is complicated further by the large genome size of around 16 Gbp and the high proportion of repetitive content with around 85% of the genome comprised of transposable elements, a mobile form of genetic sequence that can multiply and migrate within the genome over generations in a largely selfish manner (Wicker *et al.*, 2018).

After several genome assemblies of the Chinese wheat landrace Chinese Spring of increasing contiguity, completeness, and accuracy (Brenchley *et al.*, 2012; IWGSC *et al.*, 2014; Clavijo *et al.*, 2017; Zimin *et al.*, 2017), the International Wheat Genome Sequencing Consortium (IWGSC) released RefSeq v1.0 in 2018 (Appels *et al.*, 2018). This assembly was produced using the NRgene DeNovoMAGIC assembly algorithm, using Illumina sequencing reads from a variety of library prep methods to generate the contigs, and POPSEQ and Hi-C to order the contigs into 21 pseudomolecules, each representing a chromosome of wheat from chr1A to chr7D. Contigs that couldn't be placed on a pseudomolecule were placed in chrUn. The RefSeq v1.0 reference genome was accompanied by a high-quality gene annotation v1.0 which was followed by an improved annotation v1.1 which fixed several errors in the first annotation. This annotation contains 107891 high-confidence genes and 161537 low-confidence genes, classified as such based on their completeness, repeat content and similarity to genes found in DNA and protein databases. A refined version of RefSeq v1.0, RefSeq v2.1 was released in 2021 with increased contiguity due to the integration of additional optical mapping data and contigs generated from PACBIO long reads (Zhu *et al.*, 2021) and was accompanied by a

refined gene annotation. Due to this new reference not being available until part way through my PhD, the work in this thesis uses the genome assembly RefSeq v1.0 and the gene annotation RefSeq v1.1.

During my PhD, nine chromosome and four scaffold-level assemblies of wheat cultivars were generated as part of the 10+ wheat genomes project (Walkowiak *et al.*, 2020). These assemblies extend the genetic variation captured in the Chinese Spring reference genome, encompassing genetic variation from Elite wheat cultivars from around the world that differ due to past breeding selection (Walkowiak *et al.*, 2020). Using these assemblies, researchers can explore genes that are absent in the Chinese Spring reference genome, as well as genes varying in copy number between cultivars, and genes whose sequence varies between cultivars. They can also use the assemblies to replace Chinese Spring as the reference for mapping sequencing reads if the samples are more genetically similar to a cultivar with a chromosome-level genome assembly that is not Chinese Spring.

1.11 Genome sequencing for variant calling and gene expression analysis

A high-quality reference genome serves as a reference for mapping sequencing reads. This facilitates a variety of genomic analyses, including variant calling and gene expression analyses. Illumina paired-end short read sequencing is likely the most widely utilised sequencing technology. It is a second-generation sequencing, or next-generation sequencing (NGS), technology. NGS revolutionised genomics by offering cost-effective and extensive interrogation of genetic variation in target populations and high-throughput RNA sequencing (RNA-seq) for quantifying gene expression (Giani *et al.*, 2020). On the other hand, third-generation sequencing technologies such as PACBIO and Oxford Nanopore are both technologies that generate long reads and have seen extensive use in genome assembly and assessing structural variation (Gordon *et al.*, 2016; Jain *et al.*, 2018).

During Illumina paired-end sequencing, DNA is fragmented into short segments. These fragments are sequenced from both ends, providing a pair of reads that facilitates accurate mapping to the reference genome. Nowadays, these reads are typically 150 bp each, with an insert size of around 300-500bp. While alternative library designs are possible for specific experiments, this is a standard approach for variant calling. To determine where in the genome each read pair derived from, the reads are mapped to

the reference genome using alignment algorithms, implemented in bioinformatic tools such as BWA (Li, 2013) and Bowtie2 (Langmead and Salzberg, 2012). Read mapping involves finding the optimal position in the genome to align each read/read pair. Once reads are mapped, downstream analyses such as variant calling can be performed.

Variant calling is a fundamental application of genome sequencing, enabling the identification of genetic variants, such as single nucleotide polymorphisms (SNPs) and INDELs. This genotyping data is crucial for many genomic analyses, including associating genetic variants with phenotypes in genome-wide association studies and QTL mapping, marker-assisted selection, evolutionary and genetic diversity analyses, and genomic prediction (N. Wang *et al.*, 2020). While WGS provides comprehensive coverage of the entire genome, the large size of the wheat genome can make it cost-prohibitive, particularly when sequencing many lines at a sufficient depth of coverage. To overcome this limitation, reduced-representation methods are often used, reducing sequencing costs and/or enabling higher depth of coverage by only sequencing a subset of the genome, typically focusing on areas of higher interest (Borrill, Adamski and Uauy, 2015). Examples of such methods include genotyping-by-sequencing (GBS), exon capture or DNA capture designs that extend beyond the exome, such as the gene and putative promoter capture developed by Gardiner *et al.* (2019a).

Prior to the widespread adoption of cost-effective high-throughput sequencing, SNP genotyping arrays were the dominant method for genotyping a population. These arrays allowed simultaneous genotyping of a set of pre-selected SNPs obtained during a SNP discovery process. However, next-generation sequencing approaches offers greater resolution, as they discover more SNPs than are included in a genotyping array and can discover rare or novel variants rather than being limited to the variants in the array design.

1.12 The impact of unmapped and mismapped reads on genomic analyses in wheat

Sequencing reads are mapped to a reference genome based on their similarity to the reference sequence. Within-species mapping usually performs well for accurate genotyping or gene expression quantification, especially when samples are closely related. However, if the sample genome contains regions of high divergence compared to the reference genome, reads from these regions will map poorly, resulting in unmapped reads or reads mapping to the wrong locus. Mapping to the wrong locus may be

exacerbated in polyploids like wheat due to the presence of homoeologous sequences to which reads can falsely map.

The presence of introgressions is a common scenario that disrupts mapping accuracy, effectively making parts of the genome an inter-species sequencing mapping problem. These regions exhibit elevated SNP and INDEL densities and reduced synteny, causing reads to fall below the mapping threshold or leaving no proper reference locus for mapping. Adjusting mapping algorithm parameters allows mapping to be stricter or more lenient, but overly lenient mapping, in an attempt to force divergent reads to map, may compromise overall mapping accuracy. When working with diploid genomes, there is likely more tolerance for more increasing mapping leniency; however, when working with polyploids such as wheat, increasing mapping leniency will make it more difficult for reads to be assigned to the correct homoeologous region.

This issue of poor mapping is a problem for researchers when mapping samples that contain multiple introgressions as it could lead to inaccurate downstream results. This concern is particularly pertinent for a species like wheat, which contains an abundance of introgressions and homoeologous sequences which may provide an alternative incorrect mapping locus if one homoeologue is introgressed from a distant relative.

However, the phenomenon of reduced or elevated mapping coverage can also be exploited as a tool to detect divergent genome regions and identifying copy number variation within sequenced samples. Blocks of reduced mapping coverage are indicative of a deletion or an introgression in the sequenced sample while blocks of elevated mapping coverage are indicative of a duplication in the sequenced sample. This concept has been utilised in several publications.

For instance, Lemay *et al.* (2019) utilised mapping coverage information from mapped GBS reads to cost-effectively screen populations of soybean mutants for copy number variation. Following this, Keilwagen *et al.* (2019) detected large chromosomal modifications in sets of barley and wheat lines using GBS data sequenced at a very low sequencing depth. By identifying genomic windows with outlying mapping coverage – genomic windows with coverage significantly deviating from the median across the panel – they were able to identify large chromosomal modifications such as introgressions and deletions without requiring parental sequencing information. Reduced-representation sequencing methods like GBS are evidently effective at detecting large chromosomal

changes, even at low sequencing depths. However, some small structural changes may be difficult to identify as the whole genome is not sequenced.

1.13 Thesis aims and objectives

This thesis aims to leverage next-generation sequencing data to explore introgressions in wheat. This includes how sequencing data can be used to identify and characterise introgressions underlying important agronomic traits, and how the presence of introgressions presents challenges to the accurate processing of sequencing data.

These two broad aims are connected through the observation that sequencing reads derived from introgressions map poorly to a reference genome in which that introgression is not represented, which leads to signatures in the mapped sequencing data that can be used positively for introgression identification but causes problematic reference bias in ordinary genomic analyses.

The more specific aims of the thesis can be broken down by chapter:

- Chapter two:
    - Explore how WGS data generated from a set of synthetically-derived introgression lines can be used to characterise introgressed segments to a high resolution and test whether introgression junctions are enriched in specific genomic regions.
    - Assess structural changes that took place in the generation of the introgression lines.
    - Identify candidate introgressed regions and genes underlying rust resistance phenotypes.
- Chapter three:
    - Use sequencing data from a diverse mapping association panel to identify introgressions underlying heat tolerance MTAs, using a method based on the one developed in chapter two.
    - Explore the limitations of relying on a single reference genome to look for candidate genes following a GWAS.
- Chapter four:

- Following observations in the previous two chapters, determine the extent to which introgressions lead to reference bias in RNA-seq analyses in wheat.
- Develop a method to reduce reference bias caused by introgressions.

2 Pinpointing wild relative introgressions and chromosomal aberrations in hexaploid wheat/*Ambylopyrum muticum* introgression lines using whole-genome sequencing data

This chapter is an adaptation of work that has been published in *Plant Biotechnology* (Coombes *et al.*, 2022) (Appendix D1) and appears with permission granted by the Creative Commons Attribution 4.0 International License.

This work was a collaboration between the Anthony Hall group at the Earlham Institute and Julie and Ian King's group at the BBSRC Wheat Research Centre at the University of Nottingham. The introgression lines used in this chapter were generated by the Wheat Research Centre (King *et al.*, 2017, 2019). DNA and RNA for Illumina sequencing were extracted from the introgression lines by Cai-yun Yang and Stella Hubbart-Edwards, respectively, from the Wheat Research Centre. This DNA and RNA was sequenced by Genomics Pipelines at the Earlham Institute. John Fellers conducted high-molecular weight DNA extraction and Oxford Nanopore sequencing of *Am. muticum* and introgression line DH65 and. KASP™ genotyping was carried out by Surbhi Grewal. I carried out all the data analysis using the data generated by my collaborators. As I begun this project at the start of my PhD, I received guidance from Ryan Joynson regarding DNA read mapping and variant calling.

2.1 Abstract

To provide a source of novel genetic variation for the breeding community, the King group at the University of Nottingham BBSRC Wheat Research Centre generated a set of hexaploid wheat/*Ambylopyrum muticum* introgression lines. In this chapter, I have outlined an approach to identify these introgressions to a high resolution using WGS data from the introgression lines and the parent lines. Using this method, I characterised the macro-level structural landscape of seventeen introgression lines. This revealed previously characterised introgressions to a much higher resolution and revealed small, previously missed introgressions that were then validated using KASP™ markers. I discovered that introgression junctions are more likely to occur in and around gene bodies and that the development of the introgression lines resulted in many chromosomal aberrations, such as deletions, duplications, and homoeologous translocations. I then produced a draft genome assembly of *Am. muticum* using Oxford

Nanopore long reads and Illumina short paired-end reads. Using a combination of de novo, transcriptomic, and proteomic data, I produced a gene annotation of the assembly, followed by functional annotation and assignment of orthologue pairs between *Am. muticum* and wheat. I used this genome assembly and annotation, along with previously published rust resistance phenotype data, to identify candidate rust resistance genes introgressed exclusively into resistant lines.

2.2 Introduction

2.2.1 Ambylopyrum muticum

Ambylopyrum muticum [(Boiss.) Eig.; *Aegilops mutica* Boiss; 2n=2X=14; genome TT] is a diploid wild relative of wheat, belonging to wheat's tertiary genepool. It is one of the many wild relatives being used in the introgression breeding programme at the University of Nottingham's Wheat Research Centre. *Am. muticum* is of interest primarily for the transfer of abiotic and biotic stress tolerance traits to wheat (King *et al.*, 2017; Fellers *et al.*, 2020), which will likely be conferred by single large effect loci.

2.2.2 Introgression line production

The introgression lines studied here (Table 2-1) were developed at the Wheat Research Centre by the process described by King *et al.* (2017, 2019) (Fig. 2-1). *Am. muticum* accessions 2130004/2130012 were crossed with wheat varieties Pavon76 or Chinese Spring to produce F1 interspecific hybrids. To recover the introgressed *Am. muticum* segments in a predominantly wheat background, the F1 interspecific hybrids were backcrossed three times (resulting in BC3 lines) with combinations of the wheat varieties Paragon, Pavon76 and Chinese Spring. Genomic in-situ hybridisation (GISH) and KASP™ genotyping were used to ensure the presence of at least one introgressed segment in the final line. Ensuring the lines are homozygous is important to guarantee the stable inheritance of the introgressed segments. To do this, the BC3 lines were either made into double haploids, by pollinating them with maize followed by colchicine treatment, or were selfed. In this chapter, I used sequencing data generated from thirteen DH lines, three selfed lines, and one BC3 line (Table 2-1). Eight of the lines belong to a pair of lines (referred to here as DH pairs) that derive from seed from the same BC3 line.

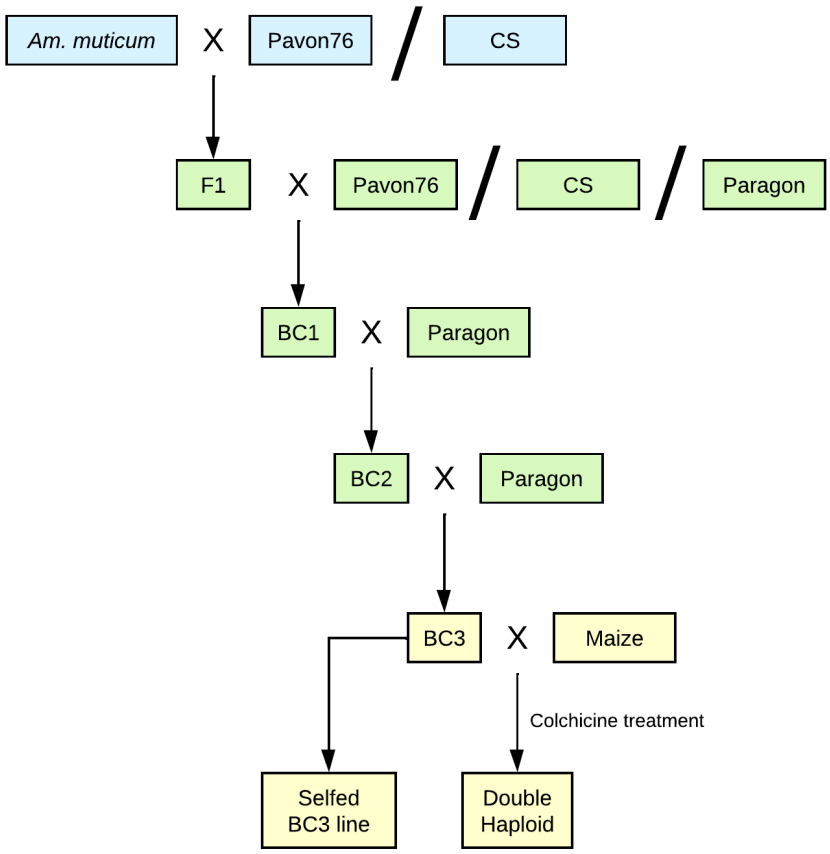


Figure 2-1. Process by which the *Am. muticum* introgression lines were generated. Adapted from King *et al.* (2017, 2019).

Table 2-1. Introgression lines analysed in this chapter. Lines with the same colour are in a DH pair, having been derived from the same BC3 line.

Line name	**Line sequencing name (used hereafter)**	***Am. muticum* accession**	**Cross history**
DHF1-8	DH8	2130012	(Pavon x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
DHF1-15	DH15	2130012	(Chinese Spring x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
DHF1-65	DH65	2130012	(Pavon x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
DHF1-86	DH86	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-92	DH92	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-96	DH96	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-121	DH121	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-123	DH123	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-124	DH124	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
DHF1-161	DH161	2130012	(Pavon x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
DHF1-355	DH355	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Maize (+ colchicine)
BC3-702-6	BC2F420	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon
BC3F2-137-2	BC3F326	2130012	(Pavon x *Am. muticum*) x Pavon x Paragon x Paragon x Self x Self

Table 2-1

DHF1-195	DH195	2130004	(Chinese Spring x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
DHF1-202	DH202	2130004	(Chinese Spring x *Am. muticum*) x Paragon x Paragon x Paragon x Maize (+ colchicine)
BC3F3-9-1	BC3F45	2130004	(Chinese Spring x *Am. muticum*) x Paragon x Paragon x Paragon x Self x Self x Self
BC3F3-10-1	BC3F46	2130004	(Chinese Spring x *Am. muticum*) x Paragon x Chinese Spring x Paragon x Self x Self x Self

2.2.3 Methods for identifying wild relative introgressions

Cytogenetic techniques, such as GISH, facilitate the visualisation of stained chromosomes under a microscope. In wheat, GISH enables differentiation between the three subgenomes of wheat and between wheat and wild relative chromosomes. This technique aids in verifying the presence of introgressed segments in a line, as well as detecting large translocations, duplications, and deletions. However, GISH suffers from limited resolution, struggling to discern chromosome changes below approximately 20 Mbp. Additionally, it is difficult to use GISH to distinguish between chromosome groups 1-7.

When genotyping data from the wild relative is available, introgression lines can be genotyped to identify wild relative segments that possess SNPs unique to the wild relative species. While simple sequence repeats were once used, they were expensive and time-consuming and have been largely replaced by SNP markers (Akhunov, Nicolet and Dvorak, 2009; Bevan and Uauy, 2013).

Genotyping to identify which SNPs are present in a set of samples has previously been achieved using SNP genotyping arrays. For example, Winfield *et al.* (2016) developed the Axiom® 820K HD array, which contains 819,571 SNPs derived from exome-captured sequencing data from hexaploid wheat Elite lines and landraces, as well as tetraploid and diploid progenitors and relatives of wheat. A subset of the 820K HD array, the Axiom® Wheat-Relative Genotyping Array, was formulated from the 36,711 most informative SNPs for detecting wheat relative introgressions (King *et al.*, 2017; Przewieslik-Allen *et al.*, 2019). These SNPs were chosen to be co-dominant. This means that both alleles at a locus can be detected and distinguished from one another, making heterozygous calls possible and allowing homoeologous genomes to be distinguished, a notoriously challenging task when working with polyploids like wheat that possess homoeologous gene copies (Kaur, Francki and Forster, 2012).

Numerous studies have utilised the Wheat-Relative Genotyping Array to identify introgressions from a variety of wild relative species in wheat (Grewal *et al.*, 2018a; Grewal *et al.*, 2018b; Cseh *et al.*, 2019; Devi *et al.*, 2019; Baker *et al.*, 2020). For example, Przewieslik-Allen *et al.* (2019) used the Wheat-Relative Genotyping array to screen hexaploid wheat lines for introgressions from *Aegilops* species. To achieve this, they compared genotype calls of *Aegilops* accessions to hexaploid wheat lines and calculated a

percentage match across each window of 10 SNPs. They used control lines containing known introgressions to determine that a match rate of 40% or higher across a 10 SNP window is indicative of an introgression.

Genotyping arrays are fairly inflexible due to their predetermined marker set and difficulties adapting them to varying sample sizes or specific marker subsets (Rasheed *et al.*, 2017). Consequently, some researchers have opted to use Kompetitive allele-specific PCR (KASP™) genotyping instead (Grewal *et al.*, 2020a). KASP™ genotyping is cheaper per sample and offers increased flexibility in terms of the number of samples sequenced per assay and the markers selected. Hundreds to thousands of samples can be genotyped with relatively few markers if needed. Instead of a fixed set of markers, newly discovered markers can be integrated easily, and a subset of markers chosen to target specific genomic regions.

SNPs from fixed chip platforms such as the Axiom arrays can be converted into KASP™ genotyping markers. Several publications have done this using SNPs from the Wild-Relative Genotyping Array to detect introgressed segments from a variety of wheat relatives (Grewal *et al.*, 2018a, 2020; Grewal *et al.*, 2018b; Grewal *et al.*, 2020a; Grewal *et al.*, 2021). Grewal *et al.* (2022) and King *et al.* (2022) improved on this by using WGS data of wild relatives, alongside a bespoke bioinformatics pipeline, to discover new SNPs and select those that are within sequences unique to a single wheat chromosome. This helps create co-dominant SNPs that can accurately detect interspecific introgressions without interference from homoeologous and paralogous sequences that are abundant in wheat. This approach led to sets of KASP™ markers that can confidently detect introgressions and evenly covers the wheat genome, with less than 60 Mbp between each marker.

KASP™ genotyping to identify segments has been conducted on many of the *Am. muticum* introgression lines studied here (King *et al.*, 2019; Grewal *et al.*, 2022). The work in Grewal *et al.* (2022) was conducted in parallel with the work presented in this chapter and segment identification through my work led to increased marker deployment using KASP™ genotyping in Grewal *et al.* (2022). The instances where this occurred will be outlined in the results section.

The resolution and reliability with which introgressions can be identified using genotyping is dependent on the density of markers available. While more expensive, WGS of introgression lines dramatically elevates the potential resolution for detecting

introgressions and could reveal small segments previously missed and allow the precise locations of segment junctions to be pinpointed. In the case of lines possessing overlapping segments but different phenotypes, this would be valuable in identifying the source of the introgressed gene(s) underlying the phenotype of interest. It would also allow the locations of junctions to be characterised to determine if they are enriched near certain genomic features such as genes.

As described in section 1.12, mapping coverage information from low coverage sequencing data can be used to identify chromosomal changes including introgressions, deletions and duplications (Keilwagen *et al.*, 2019). However, it can be challenging to differentiate between deletions and introgressions using coverage information alone. It should be possible to identify the introgressions by also using SNPs derived from WGS data that are specific to the introgressed species; introgressed regions should have both low mapping coverage and SNPs specific to the introgressed species. Additionally, SNP information will enable the origin of the donor species to be validated. This is the technique I employ in this chapter.

2.2.4 Chromosomal aberrations in introgression lines

The three subgenomes of wheat behave as diploids during meiosis due to the action of the *Pairing Homoeologous 1* (*Ph1*) locus, which ensures that only homologous chromosomes participate in synapsis and crossovers (Griffiths *et al.*, 2006; Rey *et al.*, 2017). Suppressing or deleting the *Ph1* locus has been used by scientists for decades as a tool to enable wheat chromosomes to recombine with non-homologous chromosomes from distant relatives and transfer genes from wild relatives into wheat (Martín *et al.*, 2017). This can be achieved by using a wheat parent with a mutated or deleted *Ph1* locus or a wheat relative that naturally confers *Ph1* suppression, such as *Am. muticum* and *Ae. speltoides* (Dover and Riley, 1972a; Dover and Riley, 1972b; Dvorak, Deal and Luo, 2006; Li *et al.*, 2017).

However, the freedom of chromosomal pairing that enables wild relative recombination also enables pairing and recombination between homoeologous chromosomes, leading to the exchange of chromatin between wheat subgenomes (Koo *et al.*, 2017; Koo, Friebe and Gill, 2020). In addition to reciprocal translocations, this process can also lead to deletions and duplications where the synteny between homoeologous chromosomes is

poor. Finally, forced chromosome pairings between wheat and relative chromosomes in the F1 crosses are also likely to induce chromosomal aberrations.

2.2.5 Rust pathogens in wheat

Ten of the introgression lines studied in this chapter have been screened for resistance to Kansas isolates of stem, stripe, and leaf rust (Fellers *et al.*, 2020), revealing resistances to all three. These resistance phenotypes were not observed in the wheat parent lines Paragon or Pavon76; therefore, the resistance present in these introgression lines is most likely derived from introgressed *Am. muticum* resistance genes. Identifying the regions within which the resistance genes lie and eventually identifying the gene underlying the resistance will be of high value to breeders aiming to utilise these introgression lines as sources of rust resistance.

Rust fungal pathogens from the *Puccinia* genus are devastating to global wheat production, with losses estimated between US$ 4.3 to 5.0 billion each year (Figueroa, Hammond-Kosack and Solomon, 2018). Members of this genus cause three wheat rust diseases: stripe (yellow) rust, stem (black) rust and leaf rust (brown rust), caused by *Puccinia striiformis* f. sp. *tritici*, *Puccinia graminis* f. sp. *tritici*, and *Puccinia triticina*, respectively. Stripe rust is the most economically significant of the three diseases. 88% of wheat produced globally is susceptible, 38.2% of which is produced in areas where the fungus can persist (Beddow *et al.*, 2015), and yield losses of infected fields can reach 100% (Chen, 2005). Stripe rust is of greatest concern in temperate regions (Figueroa, Hammond-Kosack and Solomon, 2018) and is among the pathogens most detrimental to winter wheat production (Chen *et al.*, 2014). Leaf rust has the widest distribution of the three pathogens and displays a high level of diversity with new virulence profiles and adaptability to climatic change posing a problem to establishing lasting, durable resistance. Stem rust is less common than the other two and tends to be well controlled throughout much of the world, but epidemics can be the most devastating (Dean *et al.*, 2012) and new virulence to many commercialised resistance genes, as seen for example in the Ug99 race, has revealed the vulnerability of popular wheat cultivars and the imminent threat posed by this and other newly evolving strains (Singh *et al.*, 2015). Identifying novel sources of rust resistance is a crucial component of continual wheat breeding and development.

Resistance to rust is conferred by plant resistance genes. There are many classes of these genes that act through different mechanisms. The most common genes implicated in plant pathogen resistance are nucleotide-binding site leucine-rich repeat (NLR) genes. However, a variety of genes can act in resistance. For example, LRR protein kinases and ABC transporters, in addition to NLRs, have been implicated in resistance against leaf, stripe, and stem rust (Krattinger *et al.*, 2009; Chen *et al.*, 2020; H. Wang *et al.*, 2020; Zhang *et al.*, 2021).

2.2.6 Chapter aims

- Develop method to identify introgressions using whole-genome sequencing data from the introgression lines and the parent lines.
- Use this method to characterise the introgression lines, identifying introgressions and other large chromosomal aberrations such as deletions and duplications.
- Pinpoint introgression junctions and validate the junction of one line using Oxford Nanopore long reads. Test whether introgression junctions are enriched near genes.
- Determine the minimum sequencing depth required to detect introgressions to a reasonable resolution using mapping coverage alone.
- Generate a draft genome assembly of *Am. muticum* using Oxford Nanopore long reads and Illumina paired-end short reads.
- Generate a gene annotation of the assembly, followed by functional annotation and assignment of orthologue pairs between *Am. muticum* and wheat.
- Identify introgressed regions underlying rust resistance and candidate rust resistance genes introgressed exclusively into rust resistant lines.

2.3 Results

2.3.1 Development of a pipeline to identify introgressions and large structural variants using whole genome sequencing data

2.3.1.1 Using mapping coverage deviation to identify introgressions, deletions, and duplications in introgression lines

First, I quantified and normalised read counts in each genomic window for the possible wheat parents (Chinese Spring, Paragon, and Pavon76) and for the introgression lines. I then compared normalised read counts for each introgression line and the two wheat parents in its crossing history (Paragon + Pavon76 or Paragon + Chinese Spring), resulting in mapping coverage deviation values. The value closest to 1 was chosen for each genomic window, assuming that the parent with mapping coverage closest to the introgression line is the donor parent of that window. The coverage deviation value reflects the relative copy number of wheat DNA within a given window compared to that of the wheat parent in that window. A value of 1 indicates similarity in DNA content, while values approaching 0 suggest possible deletions or introgressions, and values of around 2 suggest duplications. Intermediate values suggest heterozygous copy number changes.

In Fig. 2-2, coverage deviation values for introgression line DH65 are plotted in 1 Mbp windows across the genome. Low mapping coverage deviation values are seen at the start of chr4D, where an introgression has been previously identified. In addition, there are other windows with coverage deviation values outside of the normal range, such as a block of reduced coverage at the start of chr5D. However, reduced mapping coverage alone cannot guarantee the nature of the structural event that has taken place, be it an introgression or a deletion. For this, SNPs in the introgression line uniquely shared with the introgressed species are useful.

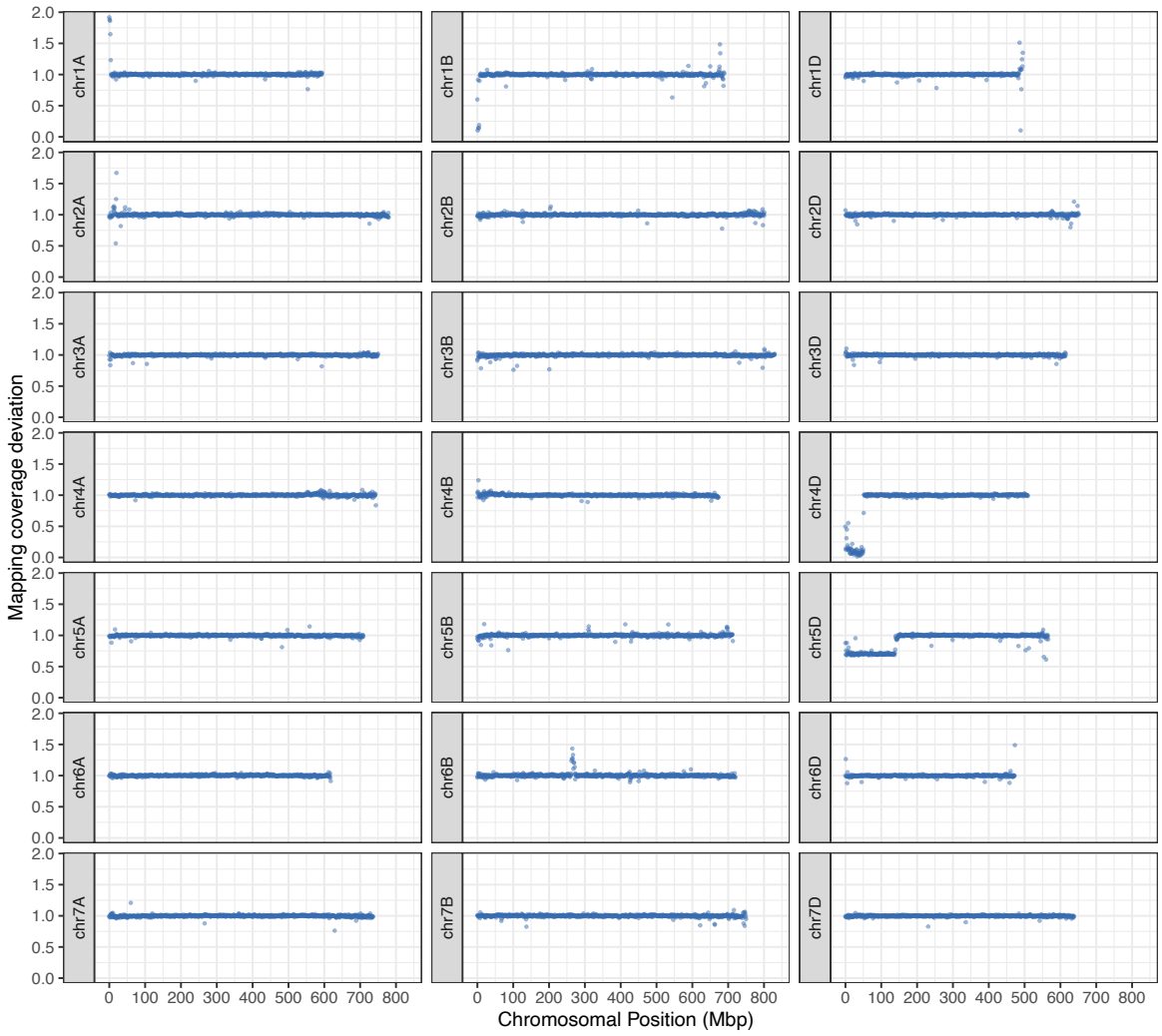


Figure 2-2. Mapping coverage deviation for introgression line DH65.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window in Chinese Spring RefSeq v1.0 compared to the wheat parent lines.

Fig. 2-3 shows IGV images of mapped reads within the introgressed region, highlighting how disruptions to synteny lead to the observed reduction in mapping coverage relative to the wheat parents that is characteristic of introgressions. Across most of an introgression, read mapping exhibits a distinctive pattern: islands of mapped reads surrounded by regions with fewer or no reads, where synteny is lower between the introgressed *Am. muticum* chromosome and the wheat chromosome that was replaced in the introgression process.

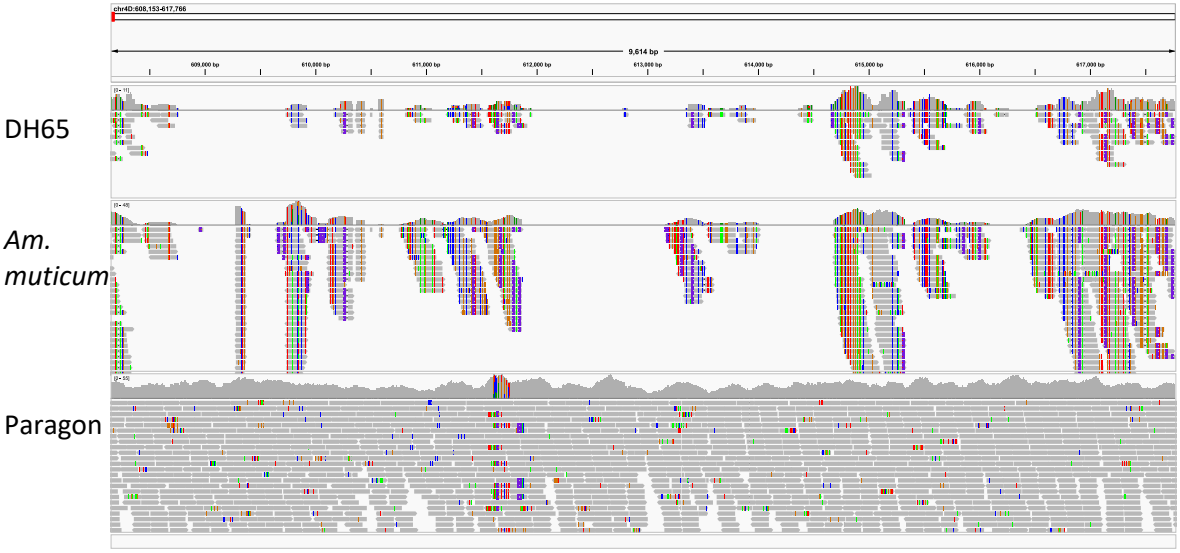


Figure 2-3. Figure 2 3. IGV image within the chr4D introgression in DH65, showing the Illumina paired-end short reads mapped to RefSeq v1.0 for DH65, *Am. muticum* and the wheat parent Paragon.

2.3.1.2 Generating *Am. muticum*-specific SNPs

To complement mapping coverage information, I identified SNPs unique to *Am. muticum*. These allow us to determine whether regions of reduced mapping coverage are introgressions or deletions. To generate SNPs that are unique to *Am. muticum*, I first conducted mapping and variant calling to identify SNPs using Illumina paired-end sequencing reads from Pavon76, Paragon and *Am. muticum*. *Am. muticum* SNPs not shared with Paragon or Pavon76 were classified as *Am. muticum*-specific SNPs. If at the same position and having the same allele as an *Am. muticum*-specific SNP, SNPs in each introgression line were classified as being *Am. muticum* specific. These were then divided into those that are homozygous and those that are heterozygous. As an example to show how homozygous and heterozygous *Am. muticum*-specific SNPs are found in different locations in the genome, in the introgression line DH65, the homozygous *Am. muticum*-specific SNPs were almost all located at the start of chr4D, within the region of the previously characterised introgression (Fig. 2-4). However, the heterozygous *Am. muticum*-specific SNPs were found in several locations in the genome, with the most densely packed region being on chr4B, which is homoeologous to the introgressed region on chr4D (Fig. 2-5).

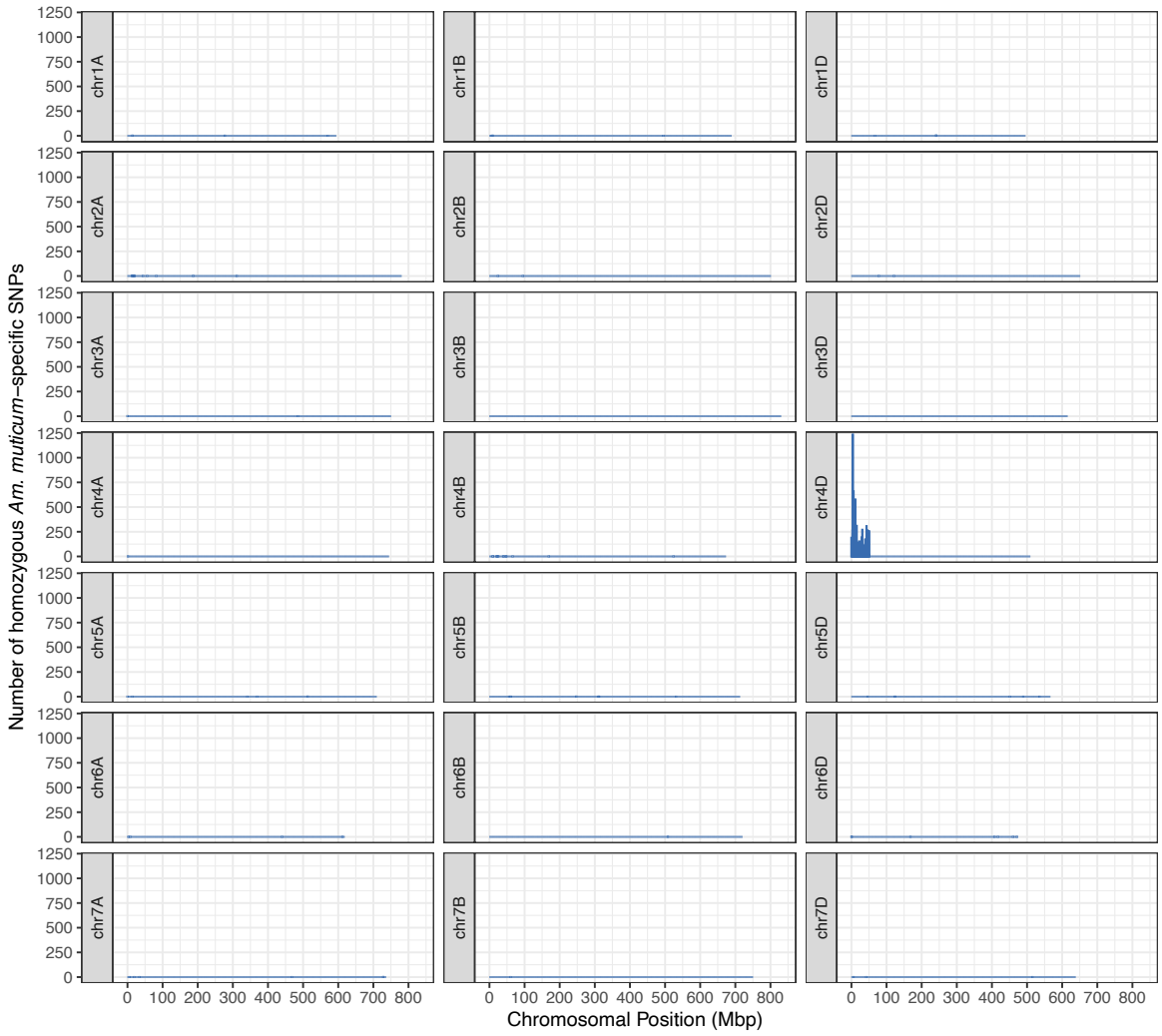


Figure 2-4. Homozygous *Am. muticum*-specific SNPs in introgression line DH65 which has a known introgression at the start of chr4D.

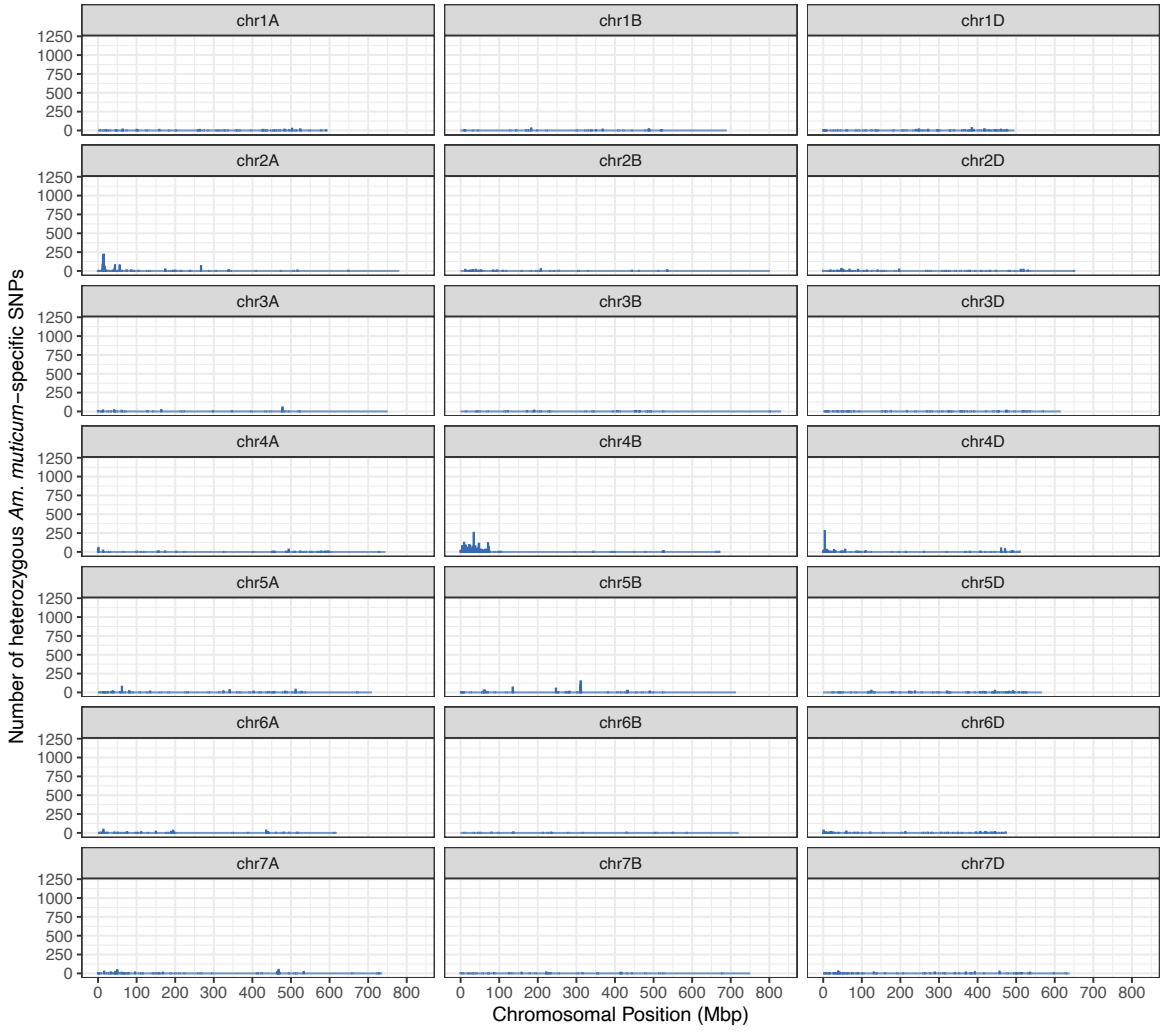


Figure 2-5. Heterozygous *Am. muticum*-specific SNPs in introgression line DH65 which has a known introgression at the start of chr4D.

I investigated the source of the heterozygous SNPs to ensure that they are artefacts of the mapping process and not genuine heterozygous SNPs. Due to the genetic distance between the wheat reference genome and *Am. muticum*, not all reads deriving from an introgression share the highest similarity to the introgressed site. These reads instead map to different regions of the genome, most notably to homoeologous regions on the other two subgenomes. However, since the homoeologous regions are typically not deleted in the introgression line, those *Am. muticum*-derived reads usually map at the same location as wheat-derived reads, leading to heterozygous SNPs being called, even if the SNP is homozygous in the wheat cultivar and in *Am. muticum*. This can be seen in IGV (Fig. 2-6) where heterozygous SNPs are called in the introgression line at sites that are homozygous in *Am. muticum* and in Paragon. Furthermore, where more than one SNP is present within a single read, the read either has the allele profile of *Am. muticum* or of Paragon across the variable sites and doesn't contain both an *Am. muticum* and a

Paragon SNP within the same read. This indicates that the reads mapped at this location are from different origins in the genome of the introgression line. This is in contrast to true introgressions, where the wheat DNA at the introgression site has effectively been deleted, so only *Am. muticum* reads, and not wheat reads, map to the introgression site. This results in true introgression sites containing homozygous *Am. muticum*-specific SNPs (Fig. 2-7).

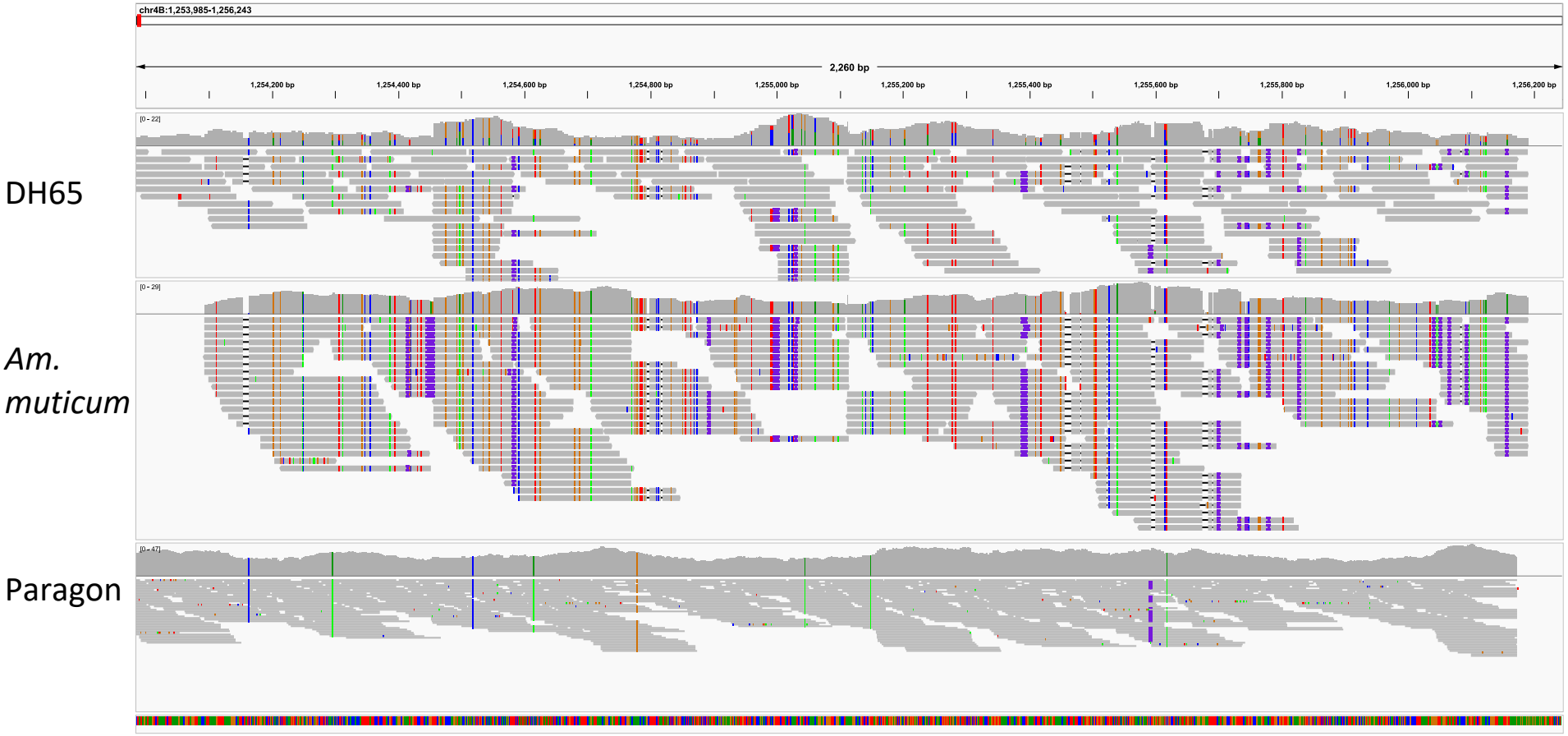


Figure 2-6. DH65 heterozygous *Am. muticum*-specific SNPs caused by reads from *Am. muticum* reads erroneously mapping to the same location as wheat reads have been mapped.

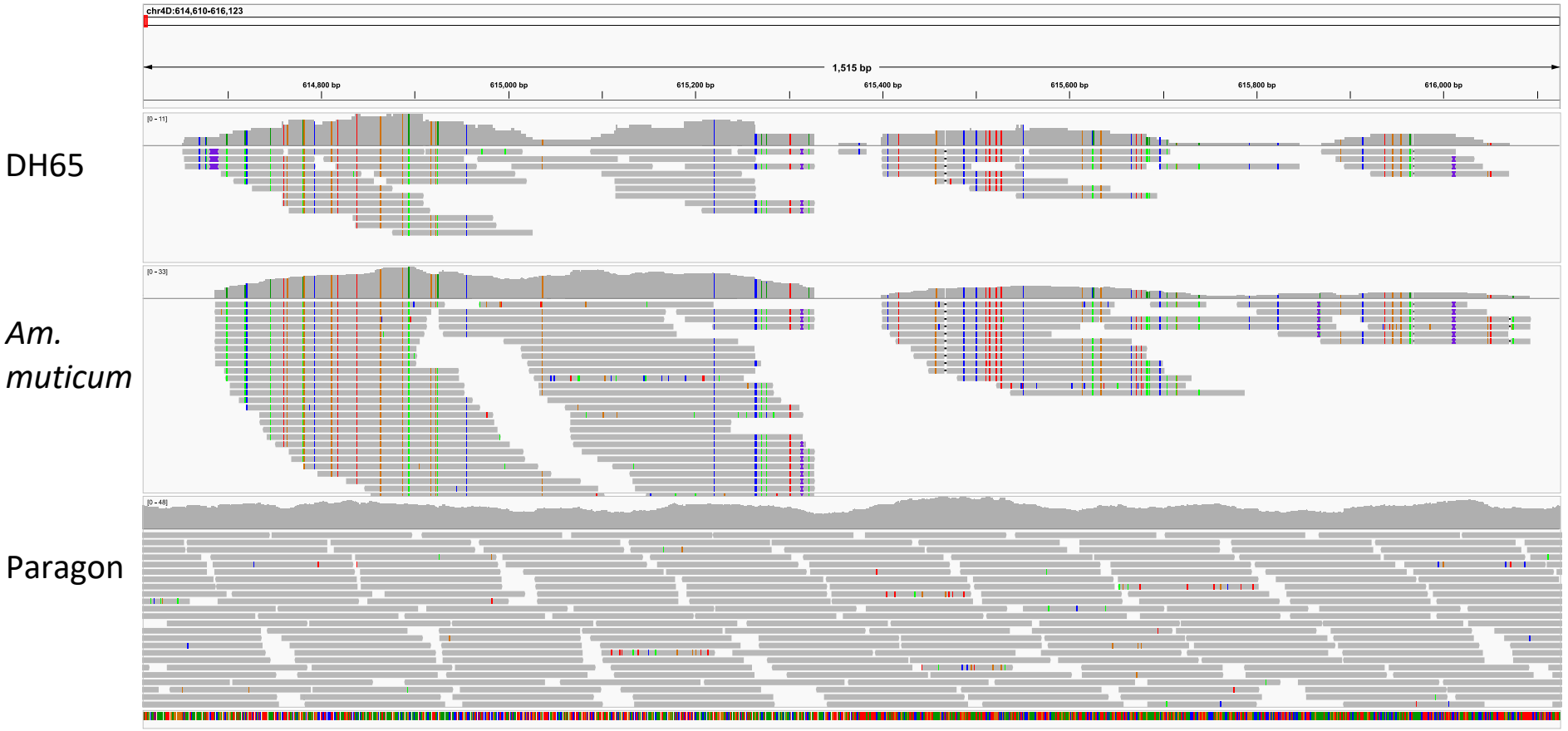


Figure 2-7. DH65 homozygous *Am. muticum*-specific SNPs at true introgression site.

2.3.1.3 Integrating *Am. muticum*-specific SNPs and mapping coverage information into pipeline

For the final identification of introgressions, I integrated the mapping coverage and SNP information by looking for blocks of genomic windows with low mapping coverage deviation values, a sufficient number of homozygous *Am. muticum*-specific SNPs, and few heterozygous *Am. muticum*-specific SNPs. Different parameters were tested until all previously detected introgressions were identified while no deletions previously verified were incorrectly classified as an introgression.

2.3.2 Whole genome sequencing allows introgressions to be detected with higher resolution

Using the pipeline, I identified introgressions in the 17 sequenced introgression lines to 1 Mbp resolution. I then defined the borders to a higher resolution, using 100 Kbp genomic windows, and pinpointed the introgression borders by hand, if possible, using IGV.

Using this approach, I confirmed the existence of 100% of the segments previously identified with KASP™ genotyping (Grewal *et al.*, 2022). However, I was able to resolve the locations of segment junctions to a much higher resolution than previous methods, due to the limited marker density available for KASP™ genotyping and the low resolution of GISH. In addition, I uncovered two previously unreported segments that were subsequently validated by KASP™ genotyping and included in Grewal *et al.* (2022); a 17.39 Mbp on the telomere of chr7D of DH195 and a 22.68 Mbp segment on the telomere of chr5D in DH121. I also identified a new 3.99 Mbp segment on chr6D of DH15. Surbhi Grewal validated this segment as real for this study using 2 KASP™ markers, WRC1873 and WRC1890. All precise segment positions are listed in Appendix A1.

Macro-level genome plots showing the introgressions detected by the pipeline in all of the introgression lines can be found in Appendix A2. Here I will show several examples for illustrative purposes. Fig. 2-8 shows DH65, which has a 51.29 Mbp segment on the telomere of the short arm of chr4D, and a 139.6 Mbp monosomic deletion on the short arm of chr5B. This is an example of a very simple line with a single clearly defined introgression.

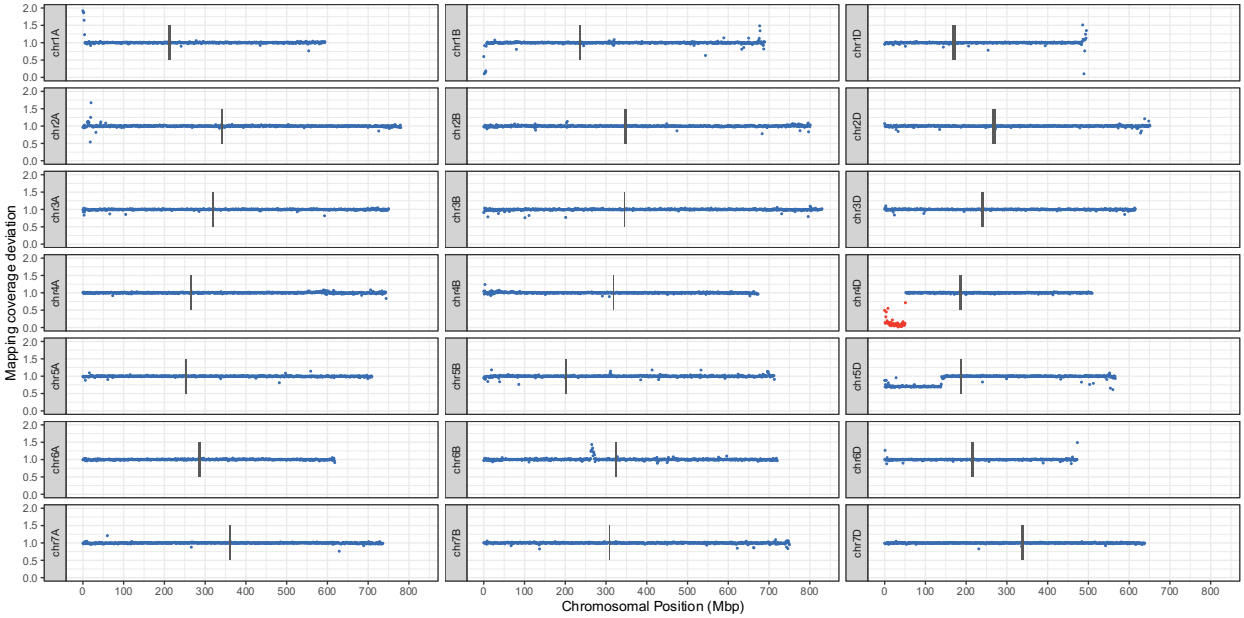


Figure 2-8. Macro-level chromosome plot for DH65.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted by Appels et al. (2018).

Fig. 2-9 shows line DH15. This line has two large introgressions: one on chr2A between 11.93 Mbp and 780.80 Mbp, and one on chr4B between around 3.00 Mbp and 635.87 Mbp. They both nearly span the whole chromosome, with a small section of wheat remaining at the start of the chr2A introgression and at the end of the chr4B introgression, showing that these have occurred through recombination rather than whole chromosome substitutions. The end of the chr2A introgression and the start of the chr4B introgression, while appearing to extend to the end of the chromosomes when using 1 Mbp windows, may have instead recombined very close to the end, based on a mapping coverage profile similar to Paragon at the very ends of the chromosome in IGV. However, there is no clear signal of coverage change from wheat to *Am. muticum* introgression as with most of the other junctions, so it is difficult to say definitively whether or not the introgression extends to the end of the telomeres, or to precisely pinpoint the position of these ends of the segments. DH15 also contains a very small introgression on chr6D between 470.63 Mbp to the end of the chromosome at 473.59 Mbp. This was the introgression that was previously missed due to lack of markers in this region but was confirmed by Surbhi Grewal using KASP™ assays.

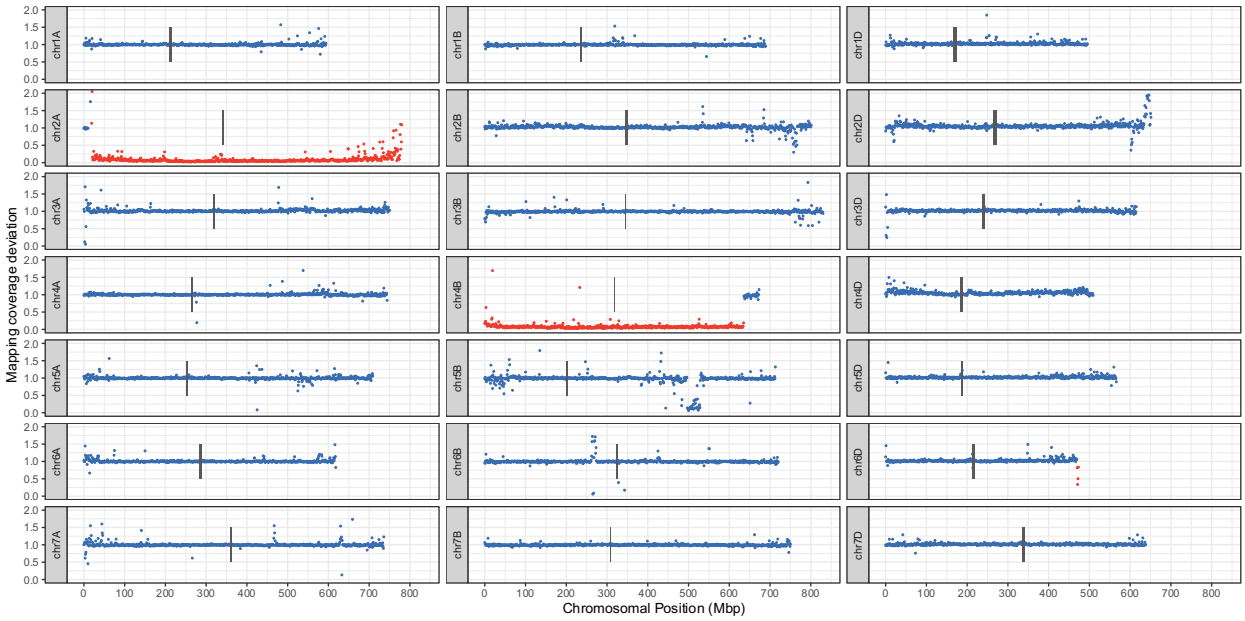


Figure 2-9. Macro-level chromosome plot for DH15.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted by Appels *et al.* (2018).

Fig. 2-10 shows line DH161. This line has a whole chromosome introgression on chr1A. Chr7D in this line is deleted, evidenced by the nearly absent mapping without homozygous *Am. muticum*-specific SNPs.

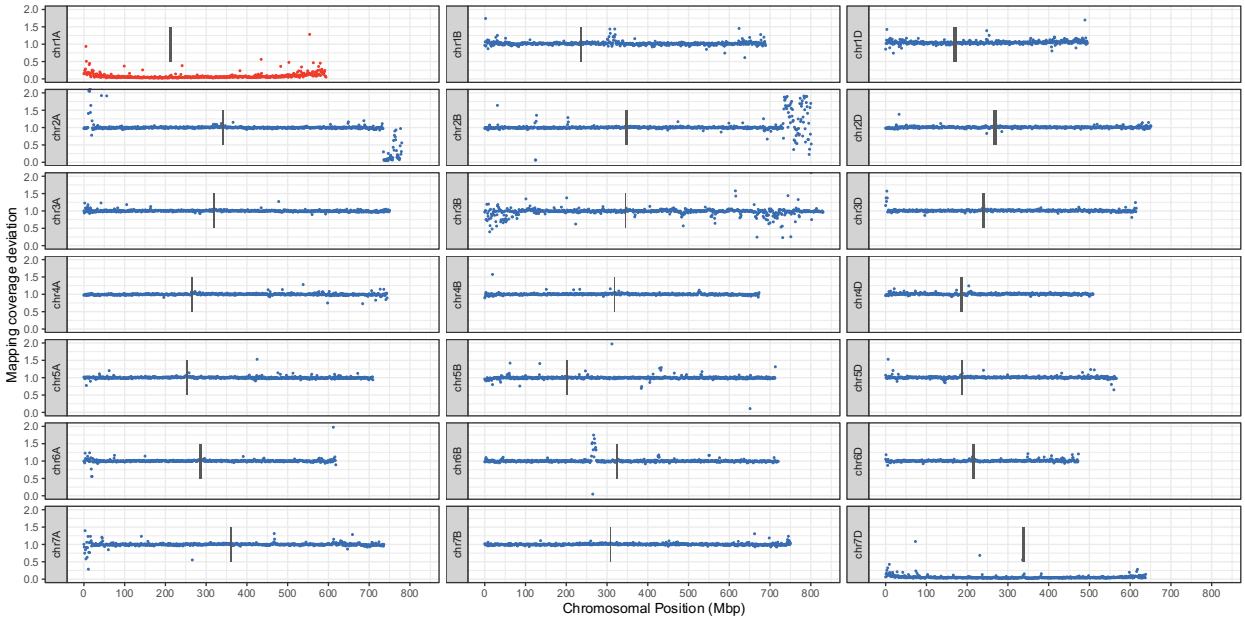


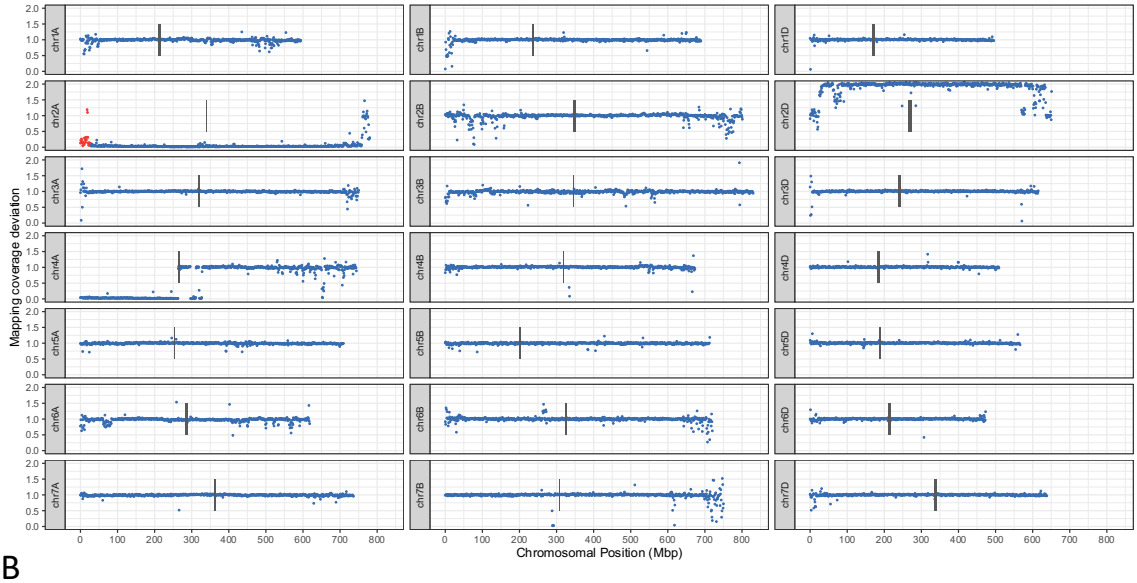
Figure 2-10. Macro-level chromosome plot for DH161.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted by Appels *et al.* (2018).

DH86 (Fig. 2-11A) is a good example of a line whose complex structure presents a challenge to this method of identifying introgressions. Based on mapping coverage deviation, most of chr2A appears to be deleted, while most of chr2D appears to be duplicated. However, there appears to be an introgression at the start of chr2A and a corresponding homoeologous region at the start of chr2D that, instead of being duplicated, has a coverage deviation of around 1. As the introgression can't exist in isolation without the rest of a chromosome, this introgression is actually at the start of chr2D. However, the duplication of chr2D and the deletion of chr2A led to the appearance of the introgression on chr2A.

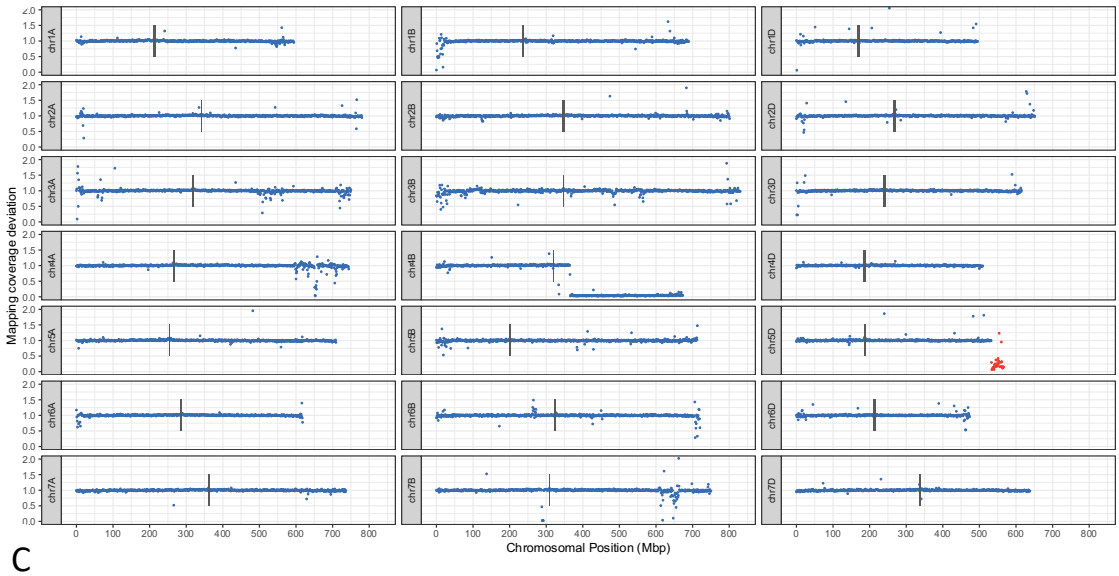
Looking at DH92 (Fig. 2-11B) alongside DH86 (Fig. 2-11A), which are a DH pair, I can infer what the BC3 line likely looked like and the two possible DH line states (Fig. 2-11C). Unlike DH86, DH92 has a normal set of group 2 chromosomes. This suggests that the BC3 line

had one copy of chr2D which possessed the introgression and was paired with chr2A, in addition to the having a normal pair of 2D chromosomes. When the chromosomes were segregated and doubled during the DH process, the resulting DH line could either have a normal pair of 2A, 2B and 2D chromosomes (as in DH92), or a pair of 2D chromosomes possessing the introgression at the start, alongside a pair of normal 2B and 2D chromosomes (as in DH86).

A


Mapping coverage deviation
Chromosomal Position (Mbp)

B


Mapping coverage deviation
Chromosomal Position (Mbp)

C

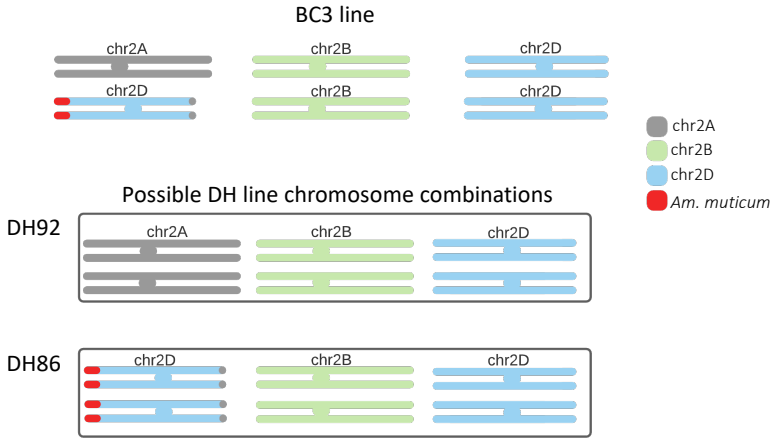

BC3 line
chr2A
chr2B
chr2D
Am. muticum
Possible DH line chromosome combinations
DH92
DH86

Figure 2-11. Macro-level chromosome plot for the DH pair A) DH86 and B) DH92. Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted by Appels et al. (2018). **C)** Inferred state of the BC3 line used to generate DH86 and DH92, and the possible DH line chromosome combinations.

Four lines; DH121, DH123, DH195 and DH202, contain overlapping introgressions on chr7D. (Fig. 2-12). DH121 and DH123 have identical segments on chr7D between 59.71 Mbp and the end of the chromosome at 638.69 Mbp. These lines are a DH pair, so this introgression is derived from the same initial cross and recombination event.

DH195 has a large introgression between the start of the chromosome and 500.82 Mbp and a small introgression between 621.30 Mbp and the end of the chromosome at 638.69 Mbp. Although DH202 is in a DH pair with DH195, the segments appear differently. DH202 is lacking the small introgression at the end of chr7D seen in DH195. This is of interest as DH195 exhibits complete adult resistance to leaf rust not seen in DH202 (Fellers *et al.*, 2020), suggesting the causal resistance gene(s) is in this segment. Furthermore, the beginning of the large chr7D segment was classified as being at the start of chr7A instead of chr7D. While this is possible, it seems more likely that the large chr7D segment remains intact and is also 500.82 Mbp in length as in DH195. The following scenario would make this true. In the BC3 line that was used to make DH195 and DH202, there is a translocation from the start of chr7D to the start of one copy of chr7A. Half of the resulting DH lines would then have two copies of chr7A with the chr7D translocation (as in DH202) and half would have two normal copies of chr7A (as in DH195). As there is *Am. muticum* at the start of chr7D, in the DH lines which have the chr7D-chr7A translocation, this would lead to homozygous *Am. muticum* SNPs called at the start of chr7A, which is homoeologous to the insertion site on chr7D. These homozygous *Am. muticum* SNPs, combined with the reduced mapping coverage caused by the absence of the start of chr7A, would lead to the start of chr7A being incorrectly classified as an introgression.

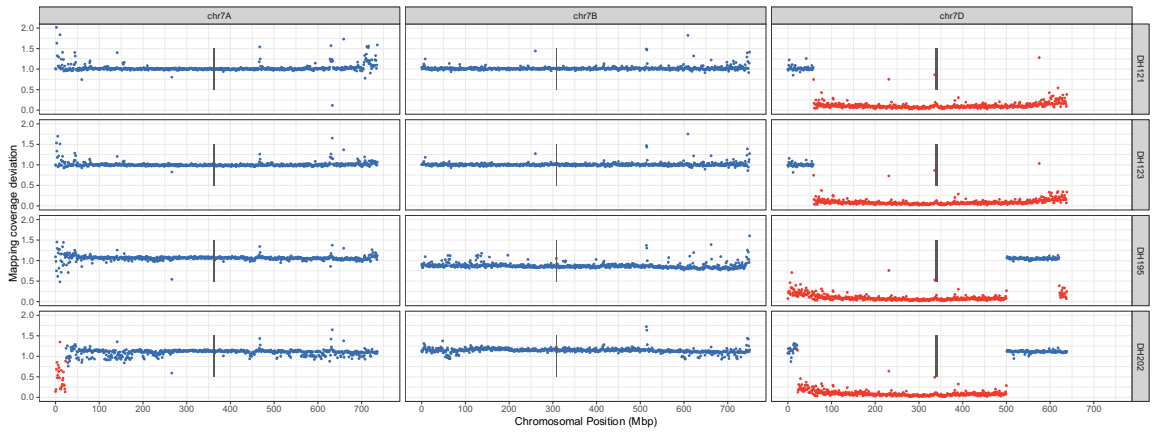


Figure 2-12. Macro-level chromosome plot of chromosome group 7 for the DH pair DH121 and DH123, and the DH pair DH195 and DH202.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted by Appels *et al.* (2018).

2.3.3 Pinpointing introgression borders

By manually searching the boundary regions using IGV, I was able to identify the positions of 33/42 segment ends (78.6%). Next, I tested whether the introgression junctions are more likely to be found near genes. For this, I focused on segment borders created through crossovers, and not those created by telomere substitutions. Also, for introgression borders derived from the same initial cross and thus at the same position in the genome, I only counted these once. Looking solely at crossovers and excluding junctions from non-independently derived segments, the precise crossover point of 12/17 (70.6%) junctions could be determined. Of the remaining five junctions, two were located to within 100 Kbp, while three, due to duplication events overlapping the introgressions, had structures too complex to precisely pinpoint. 11/12 of the junctions precisely located are within 670bp of a gene with 8 falling within the gene itself. The final junction was 6.75 Kbp from the nearest gene.

For line DH65, I validated the pinpointed junction using Oxford Nanopore long reads mapped to the RefSeq v1.0 along with the Illumina paired-end short reads (Fig. 2-13). Oxford Nanopore reads spanned the breakpoint between *Am. muticum* and wheat at the right-hand side of the 51.6 Mbp chr4D segment, adding confidence to the identification from Illumina reads alone. I assembled these mapped Oxford nanopore reads using wtdbg2 (Ruan and Li, 2020) with relaxed parameters to include reads that were clipped

due to high divergence between wheat and *Am. muticum*. The resulting contig spans the entire junction, including regions to which neither the Illumina reads from *Am. muticum* nor DH65 map. These regions appear to have elevated SNP density, explaining the gaps in mapping. This is also a good example of a junction falling within a gene.

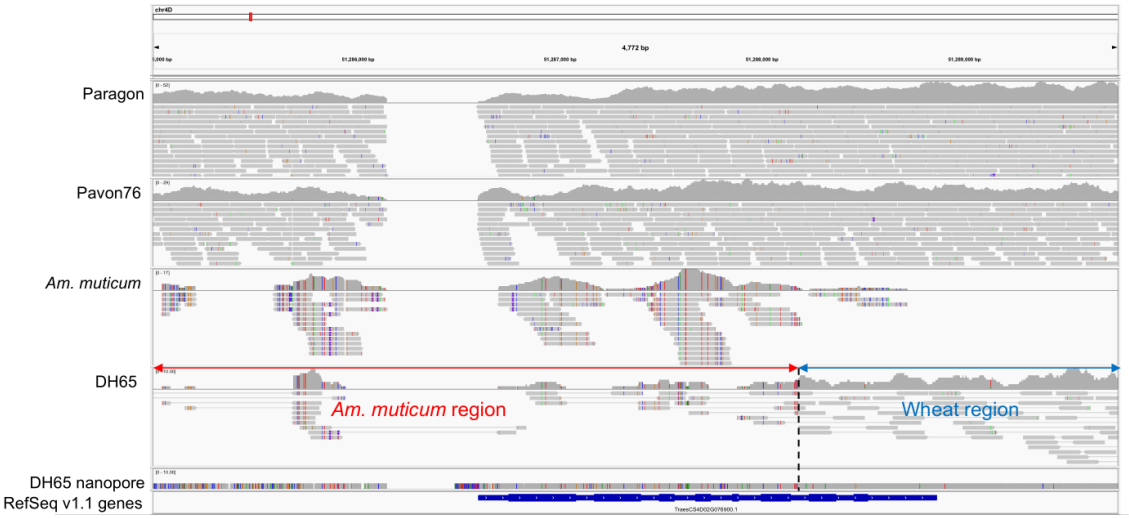


Figure 2-13. IGV image of the introgression junction on chr4D of line DH65. Illumina paired-end reads are mapped to RefSeq v1.0. The DH65 nanopore track shows assembled contigs aligned back to RefSeq v1.0.

2.3.4 Minimum sequencing depth required

WGS data offers an affordable means for breeders to locate the precise location and size of introgressed segments in wheat. This is particularly true when other genotyping data, such as KASP™, is also available as the minimum required sequencing depth is determined by that needed for the coverage deviation analysis rather than for SNP calling. I performed an analysis to determine the minimum sequencing depth required to locate the position and size of known introgressed segments using coverage deviation alone. To achieve this, I downsampled the Illumina paired-end sequencing data from DH65 and DH92, two lines for which the introgression borders are resolved to a high resolution, to 1x, 0.1x, 0.01x and 0.001x. I found that 0.01x was the lowest sequencing depth that still enabled the introgressions in DH65 and DH92 to be clearly identified (Fig. 2-14). At 0.001x, although the segments and the deletion can be seen, the noise becomes quite high. Therefore, if the segments were any more complex than those in DH65 and DH92, identification may become unreliable.

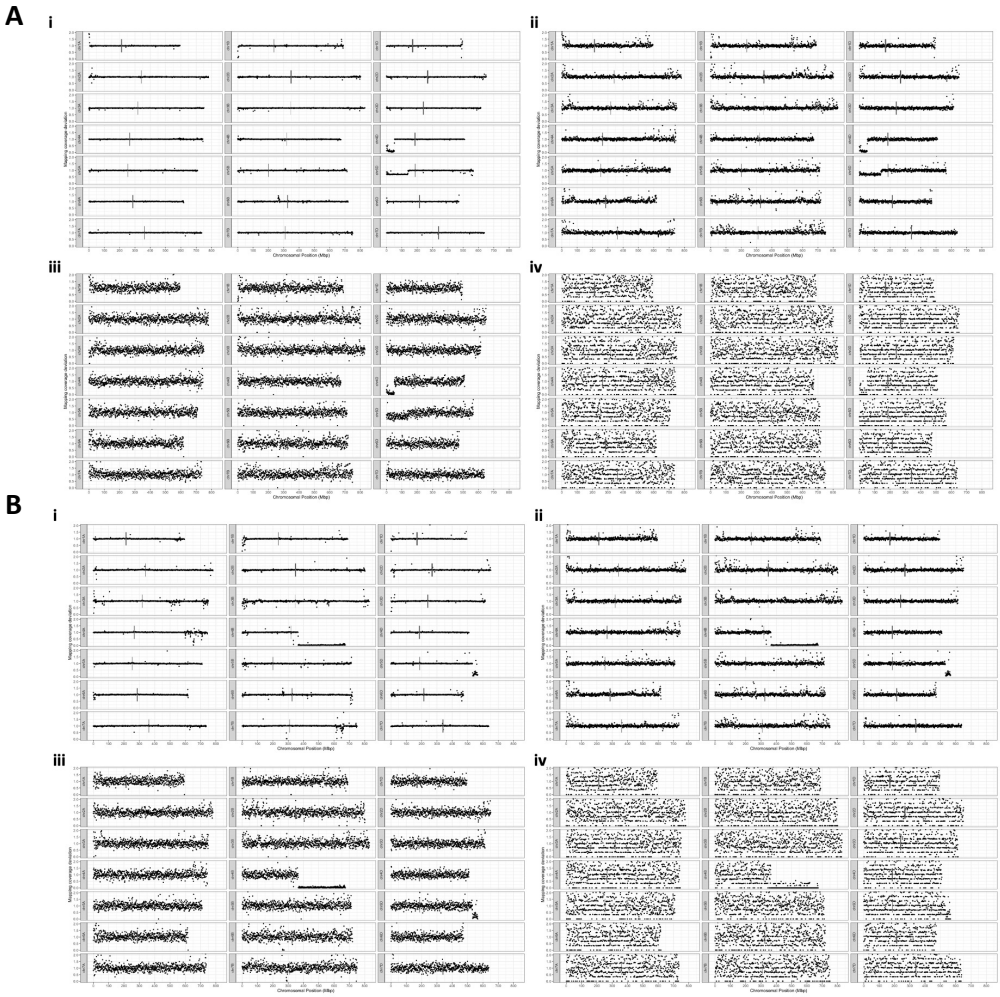


Figure 2-14. Minimum required sequencing depth to discover introgressed segments in *Am. muticum* introgression lines.
Each point represents the deviation in mapping coverage compared to the wheat parent lines in 1 Mbp windows across Chinese Spring RefSeq v1.0 for: **A)** DH65 and **B)** DH92. The Illumina paired-end reads were downsampled to **i.** 1x, **ii.** 0.1x, **iii.** 0.01x and **iv.** 0.001x.

2.3.5 Introgression crossing induces large chromosomal aberrations and homoeologous pairing/recombination

In addition to introgression sites, I used the mapping coverage information to identify large chromosomal aberrations such as deletions, duplications, and translocations, based on coverage deviation values that are outside of the normal range yet not attributable to introgressions. 12 of the 17 (70.6%) sequenced introgression lines were found to have one or more chromosomal aberrations exceeding 140 Mbp. For instance, the DH pair DH124 and DH355 possess a duplication of the majority of chr1A with a corresponding deletion of the homoeologous region of chr1B (Fig. 2-15A). In DH86, the short arm of chr4A is deleted and in DH92, which belongs to the same DH pair, the long arm of chr4B is deleted (Fig. 2-15B). DH65 and DH121, which are not in a DH pair, share an identical

monosomic deletion of the majority of the short arm of chr5D (Fig. 2-15C). DH195 has a monosomic deletion of chr1A (Fig. 2-15E).

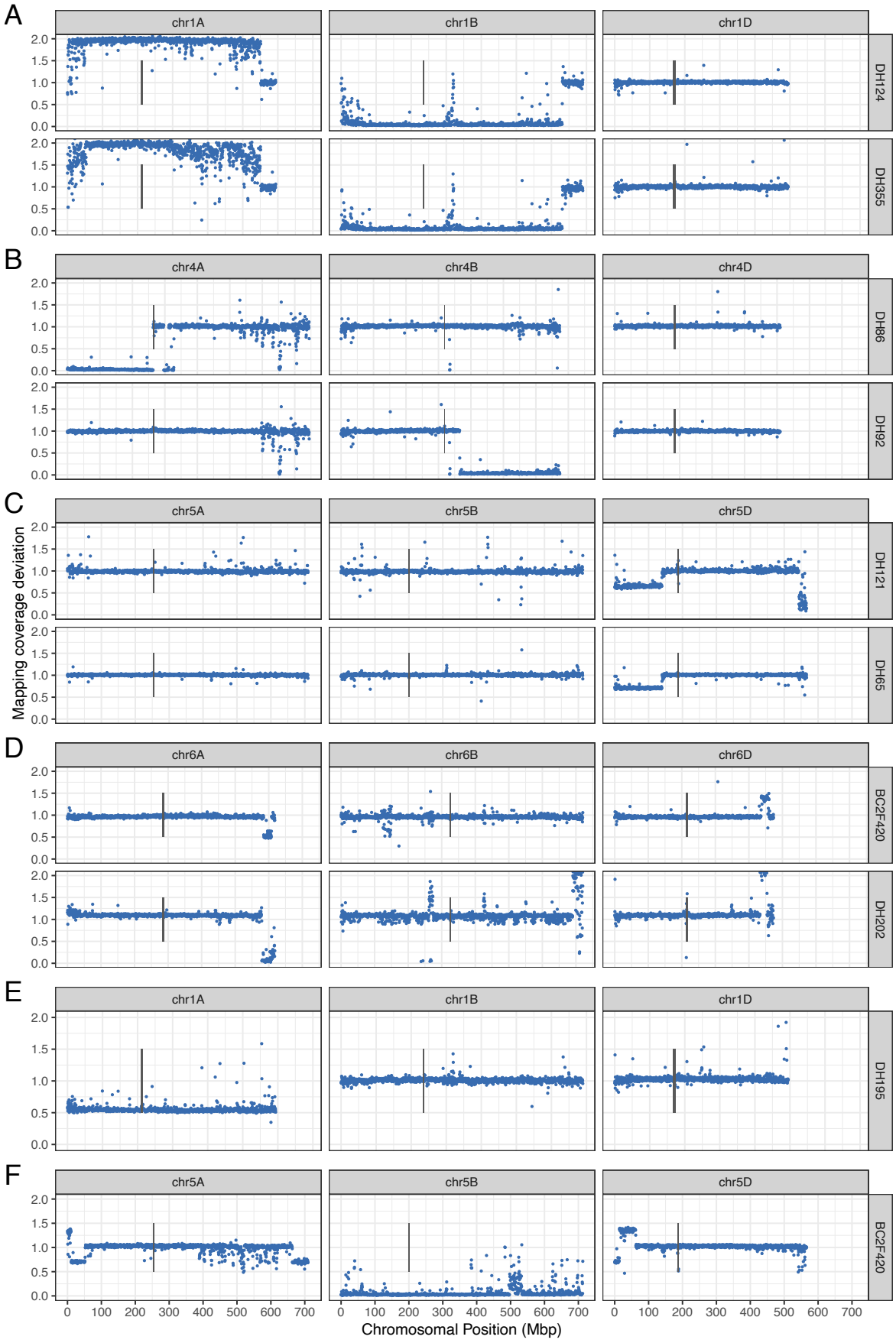


Figure 2-15. Large chromosomal aberrations in *Am. muticum* introgression lines. Each point shows mapping coverage deviation compared with the wheat parents in 500 Kbp windows across the genome. Estimated centromere positions from Appels., *et al*. (2018) are markers by vertical black bars. **A)** Corresponding duplication and deletion seen in both lines of the DH pair, caused by pairing of a duplicated chr1A and chr1B. Mapping coverage deviation of 1 at the end of chr1A and chr1B indicates a large translocation between chr1A and chr1B has taken place in duplicated chr1A + chr1B pair and discontiguous mapping coverage deviation change towards beginning of chr1A and chr1B suggests lots of smaller translocation events. **B)** Chromosome arm deletions on homoeologous chromosomes of DH pair. **C)** Monosomic deletions at the same position in two independently derived lines. **D)** Homoeologous exchange within homoeologous group 6, at similar positions in two independently derived lines. **E)** Monosomic deletion of chr1A in DH195. **F)** Homoeologous recombination event between chr5A and chr5D and a deleted chr5B.

Mapping coverage deviation can also be used to detect homoeologous translocations resulting in the non-reciprocal transfer of DNA, indicated by corresponding deletion and duplication at homoeologous regions. However, reciprocal translocations can't be detected in this way as while the DNA will be in a different position in the sample genome, the sequencing reads will map to the same place as normal, resulting in normal levels of mapping coverage.

For example, in DH124 and DH355, most of chr1A has been duplicated and most of chr1B has been deleted with a region at the end of each chromosome having normal coverage (Fig. 2-15A). This piece of chr1B left can't exist on its own. Therefore, the likely explanation for this is a translocation of chr1B onto a duplicated copy of chr1A, replacing the homoeologous region at normal coverage on chr1A. In BC2F420, recombination appears to have taken place between chr5A and chr5B, seen by the corresponding rise and fall of coverage deviation values in homoeologous regions (Fig. 2-15F).

In DH202, copy number variation affects homoeologous regions on chr6A, chr6B and chr6D (Fig. 2-15D). Chr6B and chr6D both have duplications, which together match the homoeologous region of chr6A that has been deleted. While this could have a number of biological causes, this could be caused by the two duplicated regions replacing the deleted region on chr6A side by side. If true, this suggests that homoeologous pairing between all three chromosomes took place. Similarly, BC2F420 also appears to have undergone homoeologous exchange at similar locations as DH202 (Fig. 2-15F); however, only chr6A and chr6D were involved.

2.3.6 Genome assembly and annotation of *Am. muticum*

To facilitate the identification of introgressed genes, I generated a draft genome assembly for *Am. muticum* accession 2130012. The initial set of contigs was generated using Flye (Kolmogorov *et al.*, 2019), which uses a graph-based approach to assemble Oxford Nanopore long reads. Due to the high error rate of Oxford Nanopore reads, the resulting contigs were polished using the same Oxford Nanopore reads with the built-in polisher in Flye, and then with two rounds of polishing with Pilon using the 102 Gbp of *Am. muticum* Illumina paired-end short reads that were also used for generating *Am. muticum*-specific SNPs in section 2.3.1.2. The final genome assembly was 2.53 Gbp in length, had an N50 of 75.5 Kbp, and was comprised of 96,256 scaffolds (Table 2-2). Although the genome assembly spans just 41.0% of the estimated genome size (based on a flow cytometry estimate of 6.174 Gbp (Pellicer and Leitch, 2020)), BUSCO analysis (Waterhouse *et al.*, 2018) revealed that 93.9% of the expected gene space was assembled unfragmented (Fig. 2-16).

Following the generation of the genome assembly, I conducted repeat annotation and masking followed by gene annotation. The annotation was produced by integrating evidence from root and shoot transcriptomic data, proteomic data, and ab initio predictions, and resulted in 86,841 predicted gene models, 32,385 of which were designated as high confidence (Table 2-3). 28,995 (89.5%) of the high-confidence genes were assigned functional annotation using eggnog (Huerta-Cepas *et al.*, 2017).

Table 2-2. Metrics of *Am. muticum* genome assembly.

Assembly length (Gbp)	**No. scaffolds**	**Contig N50 (bp)**	**Scaffold N50 (bp)**	**% BUSCO complete genes (% single copy)**	**No. high-confidence genes**	**No. low-confidence genes**	**No. high-confidence genes with functional annotation assigned**
2.53	96256	64199	75508	93.9% (88.7%)	32385	54456	28995

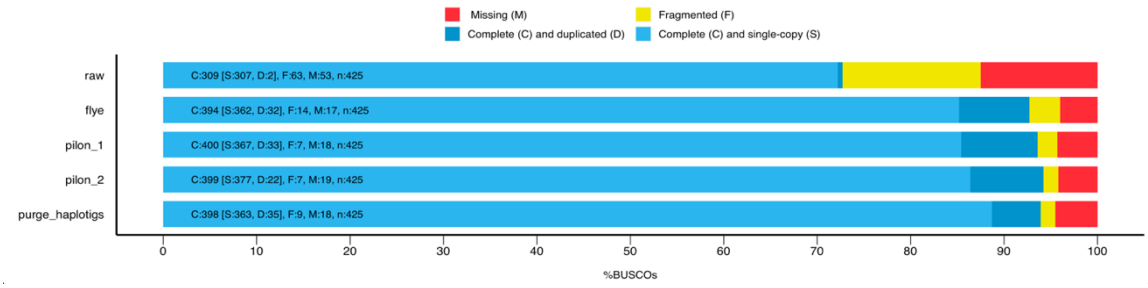


Figure 2-16. BUSCO results after each stage of the assembly.
Raw is the output of flye without any polishing. Flye is the output after long read polishing using flye's built-in polishing tool. Pilon_1 and pilon_2 are the output after the first and second round, respectively, of short-read polishing with Pilon. Purge_haplotigs is the output following collapsing of haplotigs in the polished assembly.

Table 2-3. Gene annotation metrics for high-confidence and low-confidence genes.

	High-confidence genes	**Low-confidence genes**
Total genes (no.)	32385	54456
Single exon (no.)	6695	27364
Multi exon genes (no.)	25690	27092
Mean gene length (bp)	3355	1642
Median gene length (bp)	2178	713
Mean CDS length (bp)	1198	716
Median CDS length (bp)	1000	502
Mean exons per transcript (no.)	4.81	2.39
Median exons per transcript (no.)	3	1
Mean exon length (bp)	249	307
Median exon length (bp)	131	196

2.3.7 Identifying candidate introgressed genes underlying novel resistances to stripe, leaf, and stem rust

Combining high resolution identification of segment positions with phenotypic data allows us to compare lines and predict the location of genes underlying the phenotype of interest. Combining this with a genome assembly of the introgressed species allows us to identify introgressed candidate genes for the phenotype of interest. I explored this approach using previously reported rust resistance phenotypes as a case study.

Two of the introgression lines that were sequenced, DH92 and DH121 (Fig. 2-17), show complete resistance to Kansas isolates of stripe/yellow rust at the seedling stage (Fellers *et al.*, 2020). In addition, DH92 exhibits partial resistance to stem rust and chlorotic adult resistance to leaf rust, neither of which were observed in DH121 (Fellers *et al.*, 2020). DH92 and DH121 share overlapping chr5D segments, the positions of which I refined to 533.2-566.1 Mbp (32.9 Mbp) in DH92, and 543.4-566.1 Mbp (22.7 Mbp) in DH121. It is

probable that the overlapping 22.7 Mbp region is the source of the stripe rust resistance, while the 10.2 Mbp region unique to DH92 is the source of the leaf/stem rust resistance.

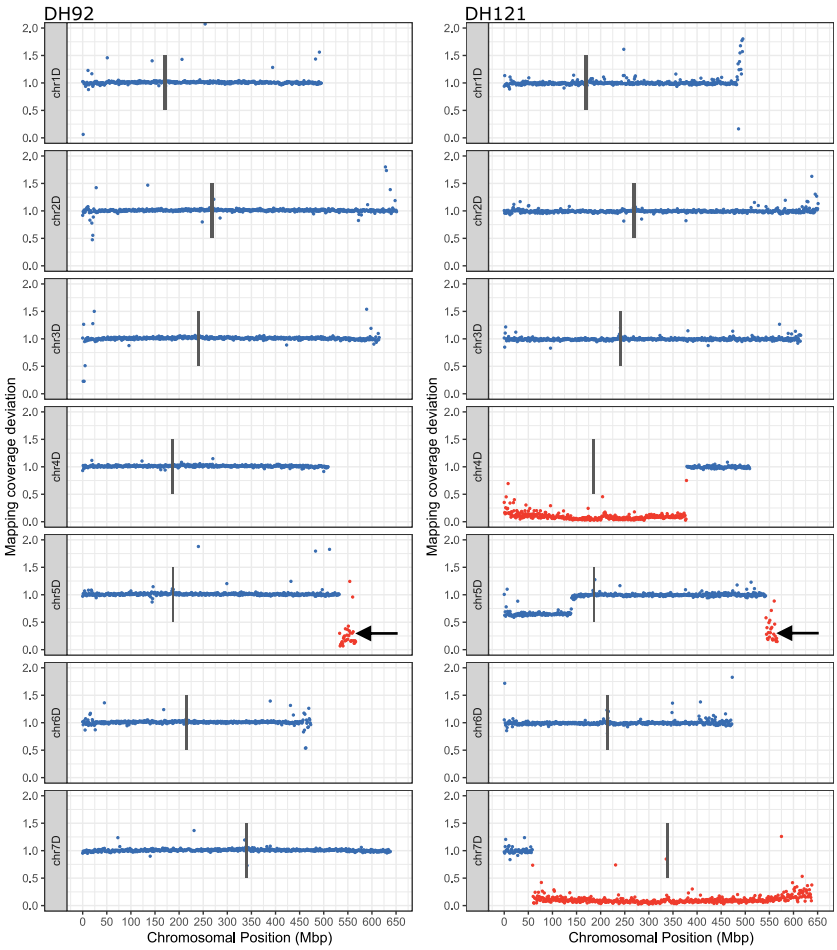


Figure 2-17. *Am muticum* introgressions detected in the D subgenome of DH92 and DH121.
Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within *Am. muticum* introgressions. The vertical black bars represent the position of the centromeres, predicted in Appels *et al*. (2018). The introgressions common to DH92 and DH121 and thus likely containing the causal resistance genes are labelled with a black arrow. Neither line has introgressions in the A or B subgenomes.

To identify genes from the *Am. muticum* genome assembly that are introgressed in both DH92 and DH121, I first concatenated the *Am. muticum* genome with the RefSeq v1.0 genome to generate a pseudo ABDT genome. To this genome I then mapped the DH92 and DH121 Illumina paired-end reads (Fig. 2-18). The principle of this method is that reads derived from genes present in the introgressions will map to the *Am. muticum* copy of that gene in the pseudo ABDT genome. If the wheat subgenomes contain orthologues of the introgressed *Am. muticum* gene, the reads will map preferentially to the *Am.*

muticum copy due to the divergence between the *Am. muticum* genome and the three wheat subgenomes. This is the same way in which read mapping correctly differentiates between homoeologues from the different wheat subgenomes in typical mapping of wheat reads to the wheat reference genome. Genes with sufficient mapping coverage in both DH92 and DH121 and not in any other lines phenotyped for stripe rust resistance were manually scrutinised using IGV and those confirmed to have even coverage across the gene in both lines were deemed to be introgressed in both.

Genes introgressed in both lines that also belonged to a family of genes previously implicated in rust resistance were classified as candidate genes for the stripe rust resistance common to both lines (Table 2-4). I identified thirteen complete NLRs that are exclusively introgressed in DH92 and DH121. Among these NLRs, twelve possess a syntenic wheat orthologue within the part of the chr5D segments that is common in both lines. Two of the NLRs have unique NB-ARC domains (<80.0% amino acid sequence identity) compared to wheat progenitor genomes *Ae. tauschii* (Luo *et al.*, 2017), *Triticum urartu* (Ling *et al.*, 2018) and *T. turgidum* ssp. *dicoccoides* (Avni *et al.*, 2017), and to Chinese Spring and fifteen other *T. aestivum* cultivars (Walkowiak *et al.*, 2020). Additionally, I found that ten of the NLRs are located within a 597.3 Kbp cluster, including the two NLRs with novel NB-ARC domains which are 19.06 Kbp from one another in the *Am. muticum* genome assembly.

Two ABC transporter genes uniquely introgressed in DH92 and DH121 were also identified, both of which have wheat orthologues on chr5D with greater than 97.5% protein sequence identity. Seven protein kinase genes were found uniquely introgressed in DH92 and DH121, six of which are LRR protein kinases, one with a malectin-like domain and a protein kinase domain. Three of the six LRR protein kinases have low protein sequence identity (52.2%, 74.2%, 77.0%) to the most similarly protein in wheat, indicating that these could be novel *Am. muticum* genes.

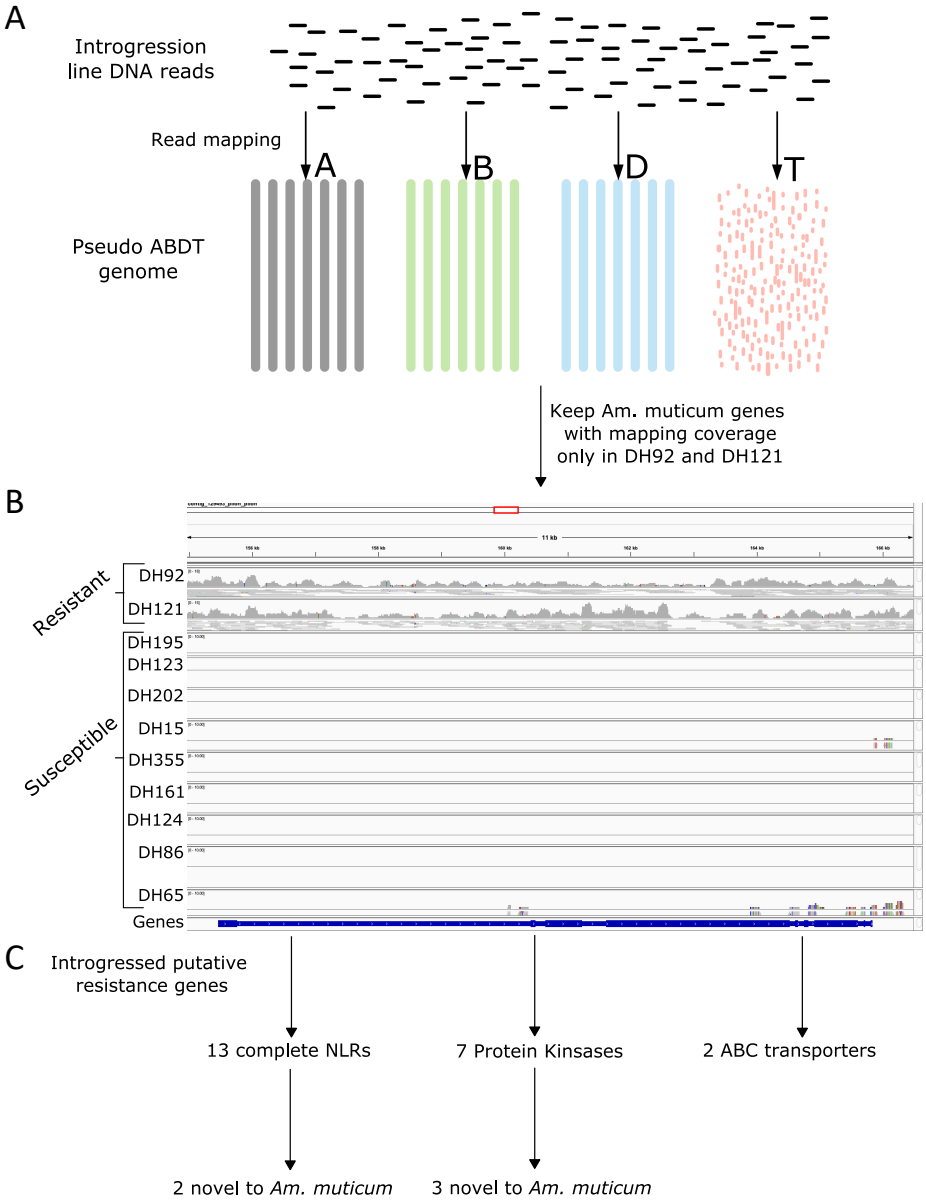


Figure 2-18. Method for identifying putative stripe rust resistance genes introgressed uniquely in DH92 and DH121.

A) Illumina paired-end short reads from the introgression lines are mapped to the pseudo ABDT genome, which was constructed by concatenating the RefSeq v1.0 wheat reference genome with the *Am. muticum* genome assembly. **B)** IGV image showing mapping coverage of Illumina paired-end short reads from eleven introgression lines mapped to the ABDT pseudo genome, here visualising mapping to a contig from the *Am. muticum* genome assembly. The bottom track shows the annotated *Am. muticum* genes. These images show coverage across the gene in DH92 and DH121 but little coverage in the other introgression lines. This suggests that the *Am. muticum* gene here is introgressed uniquely in DH92 and DH121. **C)** Number of putative *Am. muticum* resistance genes of each type identified as being uniquely introgressed in DH92 and DH121, along with the number that produce proteins novel to *Am. muticum* when compared to the proteins in wheat genomes.

Table 2-4. Candidate resistance genes uniquely introgressed in DH92 and DH121, which both exhibit stripe rust resistance due to a shared introgression on chr5D.

Gene name	**Gene type**	**Highest NB-ARC domain protein identity or overall protein identity for non-NLRs when aligned to fifteen wheat genomes and to *T. urartu*, *Ae. tauschii*, and *T. turgidum ssp. dicoccoides*.**	**Novel NB-ARC domain or novel protein for non-NLRs**
contig_104068_pilon_pilon.6	NLR	93.886	no
contig_104068_pilon_pilon.7	NLR	93.361	no
contig_105710_pilon_pilon.1	NLR	95.833	no
contig_122086_pilon_pilon.2	NLR	92.213	no
contig_122086_pilon_pilon.3	NLR	90.204	no
contig_127133_pilon_pilon.1	NLR	81.893	yes
contig_129493_pilon_pilon.7	NLR	98.75	no
contig_129493_pilon_pilon.9	NLR	94.068	no
contig_137938_pilon_pilon.3	NLR	88.571	no
contig_16723_pilon_pilon.4	NLR	87.446	no
contig_4003_pilon_pilon.2	NLR	73.214	yes
contig_40592_pilon_pilon.2	NLR	88.571	no
contig_51888_pilon_pilon:51710-63637	NLR	86.301	no
contig_120634_pilon_pilon.1	ABC transporter	97.5	no
contig_55354_pilon_pilon.1	ABC transporter	98.347	no
contig_132047_pilon_pilon.3	LRR + protein tyrosine and serine/threonine kinase	91.146	no
contig_132047_pilon_pilon.4	LRR + protein tyrosine and serine/threonine kinase	53.36	yes

Table 2-4

contig_132047_pilon_pilon.8	LRR + protein tyrosine and serine/threonine kinase	69.559	yes
contig_137761_pilon_pilon.4	LRR + protein tyrosine and serine/threonine kinase	97.272	no
contig_4003_pilon_pilon.11	LRR + protein tyrosine and serine/threonine kinase	95.66	no
contig_85897_pilon_pilon.5	LRR protein kinase	95.487	no
contig_85897_pilon_pilon.12	malectin-like and protein tyrosine kinase domains	77.861	yes

To identify candidate genes for the leaf and stem rust resistance observed in DH92 but not in DH121 (Table 2-5), I used the same approach but looked for genes with sufficient mapping coverage only in DH92. I detected three protein kinases and three wall-associated protein kinases (WAKs) that are uniquely introgressed in DH92 and likely within the 10.2 Mbp region of the chr5D segment not shared with DH121. Five of these six genes are located on the same 104.7 Kbp contig in the *Am. muticum* assembly. Two of the WAKs are orthologous to the wheat genes TaWAK388 and TaWAK390 on chr5D, while one is orthologous to TaWAK255 on chr4A.

In contrast to the other uniquely introgressed putative resistance genes, the WAKs have reads mapping to them in most of the introgression lines; however, only in DH92 is the coverage uniform across their lengths. This implies that the mapping in the other lines is caused by false mapping from similar genes. Therefore, the WAKs are still likely uniquely introgressed in DH92 and can remain as candidates for resistance. This is confirmed by the absence of mapping across the rest of the contig, on which the WAKs reside, in the other introgression lines.

Table 2-5. Candidate resistance genes uniquely introgressed in DH92, which exhibits leaf and stem rust resistance.

Gene name	**Gene type**	**Highest percentage identity when aligned to the proteins from fifteen wheat genomes and to *T. urartu*, *Ae. tauschii*, and *T. turgidum ssp. dicoccoides*.**	**Novel NB-ARC domain or novel protein for non-NLRs**
contig_147444_pilon_pilon.1	Protein kinase domain only	92.07	no
contig_147444_pilon_pilon.8	WAK	98.274	no
contig_147444_pilon_pilon.9	WAK	97.948	no
contig_147444_pilon_pilon.10	WAK	98.111	no
contig_147444_pilon_pilon.13	Protein kinase domain only	92.07	no
contig_51565_pilon_pilon.1	Protein kinase domain only	100	no

2.4 Discussion

2.4.1 Using whole genome sequencing to pinpoint wild relative introgressions in wheat – an affordable new tool to aid introgression breeding programmes

The current approach for studying synthetic introgression lines prior to deployment in breeding programmes relies on cytogenetic and genotyping techniques, such as GISH and KASP™ genotyping, respectively (King *et al.*, 2019; Grewal *et al.*, 2022). *De novo* discovery of SNPs to produce higher density KASP™ markers has improved the resolution and as demonstrated here, recent KASP™ genotyping (Grewal *et al.*, 2022) was able to detect most of the introgressed segments identified with my approach. However, these approaches lack the resolution needed to unpick the precise size and location of segments and will likely miss small segments without the guidance of WGS data to identify areas in which additional markers should be deployed. I found this with the new chr6D segment, the small chr7D segment in DH195 and the chr5D segment in DH121, the latter two of which are novel sources of disease resistance.

I demonstrated how whole genome sequencing data can be used to define introgressions to a very high resolution as well as resolve large-scale structural changes in these lines. Downsampling showed that if SNP information is not required, only 0.01x sequencing coverage is required to pinpoint the junctions of known introgressed segments to a comparable resolution. This agrees with Adhikari *et al.* (2022), who also found that 0.01x sequencing coverage was sufficient to identify the introgressions when only relying on coverage information. Overlaying this information with KASP™ genotyping will undoubtedly provide an affordable method to characterise sets of synthetic introgression lines more accurately and comprehensively.

Introgressed segments nested within complex genomic structures, such as in DH202 and DH86, can only be inferred in conjunction with cytogenetic data and/or segregation patterns of DH pairs. Some introgression segment boundaries, such as the left-hand border of chr2A in DH15, can be identified but structural complexities around the junction make them difficult to pinpoint precisely. Therefore, caution is advised when relying on the introgression assignments provided by WGS data alone, particularly for complex lines with several large introgressions/deletions/duplications. However, for most lines, where genomic structure is simpler, this approach appears to be robust. Besides, it is these simple lines that are likely to be of greater interest to breeders due to being

easier to introduce into breeding programmes, making the challenges of characterising the other lines less imperative.

Other approaches for identifying introgressions were published while I was undertaking this research. Adhikari *et al*. (2022) adopted a similar approach to mine. They used low-coverage sequencing data, with a mean of 0.025x, from 384 wheat-barley introgression lines to identify the position and copy number of barley introgressions. As in my pipeline, mapped read counts were normalised so that regions of the introgression line genome with no introgression or deletion/duplication have a value close to one. As a chromosome-level genome assembly is available for barley, they also mapped the introgression line reads to the barley genome and looked for a rise in mapping coverage to barley corresponding to the low mapping coverage against the wheat genome. This enabled the source of the introgression from the barley to be identified and is a useful addition to the pipeline for cases where a high-quality genome assembly is available for the donor species. The availability of the barley genome assembly means that SNP information derived from the introgression line sequencing reads are not required for confirming the donor species or for discriminating between introgressions and deletions. This allows for very low-coverage sequencing without an accompanying genotyping method such as KASP™.

Previous work has shown that crossovers between wheat and wild relatives are enriched in gene rich regions (Nyine *et al.*, 2020), which mirrors recombination rates along the genome (Gardiner *et al.*, 2019b). My analysis confirms this, with sufficient resolution to locate crossover sites within specific gene bodies. Outside of genic regions, synteny between wheat and *Am. muticum* is very low so genes may be the only place where synteny is sufficiently high for recombination to take place.

2.4.2 Genomic instability following alien introgression crossing

In this chapter, I showed that structural disruption is common in introgression pre-breeding material, including homoeologous pairing and recombination, and duplications and deletions up to chromosome size. This is likely caused by the *Am. muticum*-induced suppression of the *Ph1* locus which was utilised to induce recombination between wheat and *Am. muticum* chromosomes. This *Ph1* suppression enables pairing of chromosomes from different subgenomes, which can cause translocations between subgenomes and unstable chromosome pairings that can cause large deletions and duplications. Forced

pairings between wheat and wild relative chromosomes in the F1 cross are likely also responsible for structural changes. The DH process could be inducing structural change, but the fact we see these structural events in the selfed lines that weren't subjected to the DH process excludes this from being the primary explanation. We can be confident that the lack of a functional *Ph1* locus is contributing to the structural change due to the evidence of pairing and recombination between subgenomes. The backcrossing implemented in the development of these lines will have removed a lot of structural changes, so what we see here is a subset of the initially generated structural change. An awareness of chromosomal aberrations is important for breeders using these lines in their breeding programmes. It will be important to identify the location of the *Ph1* suppressor in *Am. muticum* and other wild relatives that have an innate *Ph1* suppression system, such as *Ae. speltoides* (Dvorak, Deal and Luo, 2006; Li *et al.*, 2017) to prevent segments being carried forward into breeding programmes that contain a *Ph1* suppressor that could generate further genomic disruption.

Smaller scale variation in mapping coverage suggests that structural disruption is not restricted to the large chromosomal aberrations. However, it is challenging to accurately assess smaller-scale structural variation, such as transposable element mobilisation and smaller INDELs with the data available. It may prove useful to assess the nature and extent of such variation in the future. To understand this type of variation, we will need a genome assembly of an introgression line and the wheat parents used in the cross, or a genome assembly of the wheat parent and long read sequencing data from the introgression line like produced here for DH65. As of the time of writing, these genome assemblies are not available and unfortunately, structural variation at this small scale between available chromosome-level genome assemblies and Paragon and Pavon76 is too great for structural variants arising from the creation of the introgression lines to be distinguished from existing structural variation between the cultivars.

2.4.3 A case study for uncovering candidate introgressed genes underlying phenotypes of interest

Combining high resolution detection of introgressed segment borders with phenotype information and a genic assembly of *Am. muticum* enabled me to identify likely introgressed regions for novel resistance phenotypes and produce a list of candidate resistance genes. This will help breeders develop markers to incorporate the phenotype

of interest into Elite varieties and will also facilitate further analysis to identify the causal resistance genes. It also acts as a template that can be built upon to unpick traits of interest in sets of introgression lines.

I identified the probable region of stripe rust resistance in DH92 and DH121 as being within the 22.68 Mbp overlapping region of the chr5D segment. The small size and telomeric position of this segment makes it conducive for use in breeding. Within this region, I identified candidate resistance genes, including 3 novel NLRs and 3 novel LRR Pkinase proteins. I searched for other classes of genes that have been cloned for stripe rust resistance (Zheng *et al.*, 2020), such as hexose transporter genes, wheat Kinase-START (WKS) genes, and tandem kinase-pseudokinase (TKP) genes, but found no examples of these genes uniquely introgressed in these lines.

The DH92 resistance to leaf rust that is not shared with DH121 is likely conferred by the portion of the 10.2 Mbp chr5D introgressed segment in DH92 that is not shared with DH121. The resistance is only seen in adult plants and to a composite of isolates (Fellers *et al.*, 2020); this race non-specific adult-plant resistance (APR) tends to be more durable and, in combination with the small segment size, makes this resistance another good target for further characterisation. I identified 3 WAKs and 3 protein kinases uniquely introgressed in DH92. Wall-associated kinases have previously been shown to confer resistance to leaf rust that looks similar to APR (Dmochowska-Boguta *et al.*, 2020) and protein kinase proteins, such as *Yr36*, have been implicated in APR (Ellis *et al.*, 2014).

It should be noted that the resistance could be conferred by genes that are absent from the assembly or the annotation or were missed in the detection pipeline. The main purpose of this analysis was to use the high-resolution introgression detection to identify where the resistance genes will be found and then to demonstrate the possibility of using sequencing data to probe phenotypes of interest in introgression lines. The analysis described here will work better with improved assemblies in which contiguous introgressed segments can be reconstructed and introgressed content fully assessed.

2.4.4 Future work

Several methods, including the one I have presented here, have arisen for characterising introgressions in sets of introgression lines. These work well; however, better or cheaper methods may be developed. The currently published methods have different advantages

and disadvantages, and the best approach to implement will depend on a number of factors, such as required resolution, prior existence of KASP™ genotyping, and cost limitations. For example, if KASP™ assays with sufficient marker density to identify most introgressed segments are available, but more precise estimates of size and position are required, low-cost, low coverage whole-genome skim sequencing may be sufficient and allow a larger number of lines to be assessed. Conversely, the skim sequencing could be used first, to locate regions of reduced mapping coverage. To confirm whether these regions are introgressions, these they could then be targeted with KASP™ markers, developing additional markers if current markers don't adequately cover the putative introgressed region. This combined approach is similar to what happened in practice between my work presented here and the work by Grewal *et al.* (2022). Alternatively, if a chromosome-level genome assembly is available for the donor species, skim sequencing can in most cases be used to identify introgressions without needing other genotyping.

To identify which introgressed *Am.* muticum genes underlie the rust resistance phenotypes, forward genetic screening should be conducted. For example, EMS mutagenesis can be used to generate a set of mutants for either DH92 or DH121, which can be phenotyped for rust resistance, followed by sequencing of the susceptible mutants. Loss-of-function mutations should be consistently found in the causal resistance gene in these susceptible mutants. Using the sequences of the candidate genes I identified to develop primers for amplicon sequencing could assist this process to reduce sequencing costs.

A chromosome-level genome assembly for *Am. muticum* is currently being generated at the Wheat Research Centre. This will allow the analyses presented here to be repeated with a higher quality. In particular, the identification of resistance genes introgressed in resistant introgression lines will be much more accurate using a higher quality assembly. This new assembly will be a good reference for forward genetic screening in the future to identify introgressed genes underlying phenotypes of interest.

2.5 Methods

2.5.1 DNA extraction and whole-genome sequencing

Am. muticum introgression lines and wheat parents were grown in a growth room (16h, 21°C day/8h, 18°C night). Genomic DNA from young leaves was isolated using extraction

buffer (0.1 m Tris–HCl pH 7.5, 0.05 m EDTA pH 8.0, 1.25% SDS). Samples were incubated at 65 °C for 1 h before being placed on ice and mixed with ice-cold 6 m NH4C2H3O for 15 minutes. The samples were then spun down, the supernatant was mixed with isopropanol to pellet the DNA and the isolated DNA was treated with RNase A and then purified with phenol/chloroform. DNAs were dissolved in TE (10mM Tris-HCl pH8.0, 0.1mM EDTA). PCR-free libraries were produced from this DNA with >600bp insert sizes (gel size-selection). These were sequenced by Genomics Pipelines on Illumina NovaSeq 6000 S4 flowcells to produce 150 bp paired-end reads for the introgression lines and Pavon76 and 250bp paired-end reads for *Am. muticum*. 150 bp paired-end reads for Chinese Spring (study PRJNA393343; runs SRR5893651 and SRR5893652) and Paragon (study PRJEB35709; runs ERR3728451, ERR3760033, ERR3760405 and ERR3728448) were produced in previous studies (Appels *et al.*, 2018; Walkowiak *et al.*, 2020) and were downloaded for use in this chapter.

2.5.2 Read processing, mapping and SNP calling

Adaptors from Illumina paired-end reads were removed and reads trimmed for quality using Trimmomatic v0.30 (Bolger, Lohse and Usadel, 2014) with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:40 CROP:150. Contaminants were screened and removed using Kontaminant v2.1.5 (Daly *et al.*, 2015) with *E. coli*, phage, and vector libraries. The cleaned reads were mapped to the Chinese Spring reference genome RefSeq v1.0 (Appels *et al.*, 2018) using BWA-MEM v0.7.13 (Li, 2013) with the -M parameter to enable duplicates to be marked. The alignment was filtered using samtools v1.4 (Li *et al.*, 2009): supplementary alignments, improperly paired reads, and non-uniquely mapped reads (mapping quality <= 10) were removed. PCR duplicates were detected and removed using Picard's MarkDuplicates v2.1.1 (Depristo *et al.*, 2011). Variants were called using mpileup from samtools v1.4 (Li *et al.*, 2009) and call from bcftools v1.3.1 (Li and Barrett, 2011) using the multiallelic calling model -m.

Initial variant filtering was performed using GATK v3.5.0 (Depristo *et al.*, 2011) to remove INDELs and to retain SNPs with a quality score >= 30 and a read depth >= 10 for the parental lines and >= 5 for the introgression lines. Read depth filtering differed between parental and introgression lines because introgression lines were sequenced to a lower depth than the parent lines. Introgression line SNPs were further filtered as follows:

homozygous SNPs were retained if 5 or more reads supported the alternative allele and the allele frequency was 1; heterozygous SNPs were retained if 3 or more reads were supporting each allele. Sites with 3 or more alleles were removed.

2.5.3 *In silico* karyotyping - calculating mapping coverage deviation compared to wheat parents

The number of mapped reads after filtering and duplicate removal was counted across genomic windows (1 Mbp and 100 Kbp) in RefSeq v1.0 using bedtools makewindows v 2.28.0 (Quinlan and Hall, 2010) and hts-nim-tools v0.0.1 (Pedersen and Quinlan, 2018) for the wheat parents (Chinese Spring, Paragon and Pavon76) and each introgression line. Mapped read counts were normalised by dividing by the total number of reads following duplicate removal. In each genomic window, normalised read counts for each introgression line were divided by the normalised count of each wheat parent in its crossing history (Paragon & Pavon76 or Paragon & Chinese Spring. The number closest to 1 was kept as the coverage deviation for that window, under the assumption that the parent with mapping coverage closest to the introgression line is the parental donor in that window. The resulting coverage deviation value reflects the copy number of wheat DNA in that window relative to the wheat parent. A value of 1 indicates that the DNA in that window is present in the same amount as in the parent line. A value approaching 0 suggests either a deletion or an introgression has occurred at that region, and a value of 2 suggests a duplication event has taken place. Intermediate values indicate heterozygous copy number changes.

2.5.4 Identifying *Am. muticum*-specific SNPs and matching them to introgression line SNPs

First, I generated 18,496,474 SNPs between *Am. muticum* and Chinese Spring that weren't shared with either Paragon or Pavon76. At heterozygous sites where one allele was unique to *Am. muticum*, the unique allele was retained. If two alternative alleles were present, both were kept provided both were specific to *Am. muticum*. Introgression line SNPs were then assigned as *Am. muticum* if matching an *Am. muticum*-specific SNP in position and allele. Sites exceeding 3x mean coverage level were removed as this is a signature of collapsed repeat expansion. These SNPs were then split into homozygous and heterozygous and binned into 1 Mbp windows using bedtools coverage v2.28.0 (Quinlan and Hall, 2010).

2.5.5 Assigning introgressed regions using coverage and SNP information

Coverage deviation blocks were defined based on contiguous blocks of 1 Mbp windows with coverage deviation < 0.7, with windows within 5 Mbp from the previous coverage deviation block being merged. The block was discarded if < 80.0% of constituent windows had a coverage deviation < 0.7. 1 Mbp windows within a coverage deviation block and containing >= 55 homozygous *Am. muticum* specific SNPs and a ratio of homozygous to heterozygous *Am. muticum* specific SNPs >= 4 were classified as candidate *Am. muticum* windows. Coverage deviation blocks with >= 14% windows assigned as *Am. muticum* using the parameters above were classed as an introgressed segment. These parameters are adjustable and were chosen because they revealed all previously known introgression segments and didn't falsely class any known deletions as introgressions. To locate the borders of the introgressions more precisely, the coverage deviation values in the 100 Kbp windows up and downstream of either end of the introgression blocks were examined to determine where the coverage deviation starts to decrease. To do this, and locate the precise position of the introgression junction if possible, the BAM alignment files for *Am. muticum*, Paragon, Pavon76, and the introgression line were loaded into IGV (Robinson *et al.*, 2011). I then searched to find the position where the coverage and SNP profile switch from those resembling the wheat parents to those resembling *Am. muticum*. In cases where the junction could not be precisely located, a range was given within which the junction position most likely falls.

2.5.6 KASP™ validation

Genomic DNA was isolated from leaf tissue of 10-day old seedlings in a 96-well plate as described by Thomson and Henry (Thomson and Henry, 1995). Introgression line DH15 was genotyped alongside the three parental wheat genotypes (Chinese Spring, Paragon and Pavon76) and the two *Am. muticum* accessions as controls. For each KASP™ assay, two allele-specific primers and one common primer were used (Appendix A3). The genotyping procedure was as described in (Grewal *et al.*, 2020b). In summary, the genotyping reactions were set up using the automated PIPETMAX® 268 (Gilson, UK) and performed in a ProFlex PCR system (Applied Biosystems by Life Technology) in a final volume of 5 µl with 1 ng genomic DNA, 2.5 µl KASP™ reaction mix (ROX), 0.068 µl primer mix and 2.43 µl nuclease free water. PCR conditions were set as 15 min at 94°C; 10 touchdown cycles of 10 s at 94°C, 1 min at 65–57°C (dropping 0.8°C per cycle); and 35

cycles of 15 s at 94°C, 1 min at 57°C. Fluorescence detection of the reactions was performed using a QuantStudio 5 (Applied Biosystems) and the data was analysed using the QuantStudio™ Design and Analysis Software v1.5.0 (Applied Biosystems).

2.5.7 Validation of introgression junction with Oxford nanopore long reads

DNA from introgression line DH65 was prepared using ligation sequencing kit SQK-LSK109 and sequenced to a depth of 7x on a MinION using a R9.4.1_RevD flow cell. Reads were filtered using NanoFilt (De Coster *et al.*, 2018) to remove reads shorter than 1 Kbp or with a quality score less than 7. Filtered reads were mapped to RefSeq v1.0 using minimap2 v2.7 (Li, 2018) with parameters -ax map-ont and --secondary=no. Mapped reads around the breakpoint (chr4D:51283000-51595000) of the chr4D introgression were extracted using samtools v1.4 (Li *et al.*, 2009), including clipped portions of mapped reads, and assembled using wtdbg2 (Ruan and Li, 2020). The resulting contigs were mapped to RefSeq v1.0 using minimap2 v2.7 (Li, 2018) with parameters -ax map-ont and –secondary=no and examined in IGV (Robinson *et al.*, 2011) along with the mapped Illumina paired-end reads from DH65 and the parent lines.

2.5.8 Genome assembly of *Am. muticum*

DNA from *Aegilops mutica* (now *Am. muticum*) line 2130012 (JIC) was prepared using ligation sequencing kit SQK-LSK109 and sequenced on a MinION using a R9.4.1_RevD flow cell. 178 Gbp of raw nanopore reads were filtered using NanoFilt v2.7.1 (De Coster *et al.*, 2018), removing reads shorter than 1 Kbp or with a quality score less than 7. Filtered reads were assembled using Flye v2.8.1-b1676 (Kolmogorov *et al.*, 2019). Following Oxford nanopore read polishing integrated into Flye, I conducted 2 rounds of Pilon v1.23 (Walker *et al.*, 2014) polishing using the 102 Gbp of *Am. muticum* Illumina paired-end short reads described earlier to correct systematic errors in the Oxford nanopore reads. Finally, haplotigs that were not collapsed in the assembly were detected and resolved using purge_haplotigs (Roach, Schmidt and Borneman, 2018). Gene completeness was assessed using BUSCO v3.0.2 (Waterhouse *et al.*, 2018) with parameters -l viridiplantae_odb10 –species wheat and -m geno.

2.5.9 Repeat annotation

A de novo library of transposable elements for *Am. muticum* was produced using EDTA v1.9.5 (Ou *et al.*, 2019), using CDS sequences from *T. aestivum* RefSeq v1.1 to exclude

protein-coding regions from the library. The resulting library was aligned to the assembly using BLASTn from blast+ v2.7.1 (Camacho *et al*., 2009) and sequences with fewer than 3 full-length hits (defined as over 80.0% similarity across more than 80.0% of the length of the element) were removed. The remaining sequences were clustered using cd-hit-est v4.6.7 (Li and Godzik, 2006) with parameters aL 0.8 aS 0.8 -c 0.8 to remove redundant sequences. The resulting library was used to mask the genome assembly using RepeatMasker v4.07 (Chen, 2004) with parameters -s -no_is -norna -nolow -div 40 -cutoff 225.

2.5.10 Gene annotation

Ab initio, protein homology and transcriptome evidence were combined to predict protein-coding genes in the *Am. muticum* assembly. First, the *Am. muticum* RNA reads (4 root replicates and 4 shoot replicates) were trimmed using Trimmomatic v0.30 (Bolger, Lohse and Usadel, 2014) with the parameters ILLUMINACLIP:BBDUK_adaptor.fa:2:30:12 SLIDINGWINDOW:4:20 MINLEN:20 AVGQUAL:20. The trimmed RNA reads were mapped to the masked genome using STAR v2.7.6a (Dobin *et al.*, 2013). Transcripts were assembled using four independent reference-guided approaches; Trinity v2.1.1 (Grabherr *et al.*, 2011); StringTie v2.1.4 (Pertea *et al.*, 2015), Cufflinks v2.2.1 (Trapnell *et al.*, 2012) and CLASS2 v2.1.7 (Song, Sabunciyan and Florea, 2016). Transdecoder v5.5.0 (Haas *et al.*, 2013) was used on each set of transcripts to produce coding ORFs. PORTCULLIS v1.2.0 (Mapleson *et al.*, 2018) was used to produce splice information. Uniprot (The UniProt Consortium, 2019) reference proteomes for *T. aestivum*, *Ae. tauschii*, *T. turgidum*, *Oryza sativa*, *Brachypodium distachyon*, and *Arabidopsis thaliana* were aligned to the genome using BLASTx from blast+ v2.7.1 (Camacho *et al*., 2009). Mikado (Venturini *et al.*, 2018) was used to merge and refine the transcripts produced by each tool, aided by the splice site information and the protein homology evidence. The same uniprot reference proteomes as above were aligned to the masked genome using TBLASTN (Camacho *et al*., 2009). Hits within 20 Kbp that had an e-value below 1e-5 and sequence identity greater than 75% were merged using bedtools merge v2.28.0 (Quinlan and Hall, 2010). These regions were searched using exonerate v2.4.0 (Slater and Birney, 2005) to refine protein alignments. To produce a set of transcript annotations to train the *ab initio* gene predictor Augustus v3.1.0 (Hoff and Stanke, 2019), Mikado protein coding genes were filtered to retain multi-exon genes that have < 80.0% amino acid identity with any other protein, have no stop codon in the ORF, and are at least 500bp away from another gene.

Of the remaining transcripts, the 2000 with the highest mikado score were retained and randomly split into 1800 training genes and 200 testing genes. The AUGUSTUS hidden Markov model (HMM) was trained using these transcripts using etraining followed by 10-fold cross validation using optimize_augustus.pl. To produce initial *ab initio* gene predictions, Augustus was run using this HMM, along with intron hints produced from the STAR bam file using bam2hints from Augustus. EvidenceModeler (Haas *et al.*, 2008) was used to integrate the *ab initio*, transcriptome, and protein evidence and produce a set of high-confidence gene models for model training. This set of genes was used to retrain Augustus as before, as well as GlimmerHMM v3.0.4 (Majoros, Pertea and Salzberg, 2004). Augustus, SNAP v2013_11_29 (Korf, 2004) (using the rice HMM) and GlimmerHMM were run to produce final *ab initio* gene predictions. These predictions were incorporated with the transcript and protein homology evidence using EvidenceModeler to produce final gene models with the following weightings: mikado transcriptome - weight 14; Augustus gene models - weight 4; exonerate protein evidence - weight 7; SNAP gene models - weight 1; GlimmerHMM gene models - weight 2. Predicted proteins were aligned to the TREP database of transposable elements and to TREMBL proteins using BLASTp from blast+ v2.7.1. Genes were classed as high-confidence if they had a complete protein-coding gene model, transcript evidence from one or more of the transcriptome assembly methods, and their encoded protein had one or more significant hits to TREMBL (Bairoch and Apweiler, 1997) and no significant hit to TREP (Wicker et al., 2002). The remaining predicted genes were classed as low-confidence. Gene functions were assigned using eggnog v5.0 (Huerta-Cepas *et al.*, 2019).

2.5.11 Assigning orthologue pairs

Am. muticum/wheat orthologue pairs were assigned using best reciprocal BLAST hits combined with OrthoFinder orthogroup assignments. OrthoFinder v2.5.2 (Emms and Kelly, 2019) was used with default settings to cluster the longest protein encoded by each high-confidence gene from *Am. muticum*, *Ae. tauschii*, *T. urartu*, *T. aestivum*, *O. sativa* and *B. distachyon* into orthogroups. *Am. muticum* proteins (extracted and translated from the GFF annotation file) and wheat proteins (taken from IWGSC 1.1 pep.fa file for wheat proteins) were aligned reciprocally using BLASTp from blast+ v2.7.1 (Camacho *et al.*, 2009) with parameters -outfmt 6 -max_hsps 3 -max_target_seqs 3 -evalue 1e-6. This was done for each wheat subgenome independently. Hits were retained if percentage identity >= 90.0% and alignment length was >= 80.0% query length. An *Am. muticum* gene was

placed in an orthologue pair with a wheat gene if it was in an orthogroup with that gene and the pair were each other's best reciprocal BLAST hit.

2.5.12 Identifying which *Am. muticum* genes are introgressed in each introgression line

The wheat reference genome RefSeq v1.0 and the draft *Am. muticum* assembly were concatenated to form a pseudo ABDT genome. Illumina paired-end short reads from the introgression lines were mapped to this genome and filtered using the same process as mapping to RefSeq v1.0 alone. *Am. muticum* genes in each introgression line were defined as introgressed if the introgression line reads, when mapped to the ABDT pseudo genome, had a mean coverage >= ~0.6 * mean sequencing depth (≥ 13.2x for DH202 and ≥3x for the remaining lines) across the gene. The gene also had to be found on a contig/scaffold in the *Am. muticum* genome assembly which has a gene (which could be the same gene) that passes the above coverage threshold and is in an orthologue pair with a wheat gene whose start position is within a region labelled as a *Am. muticum* introgression. This is a conservative classification to prevent inclusion of non-introgressed genes.

2.5.13 Identifying introgressed resistance genes

Am. muticum NLR genes were identified using two parallel methods. In the first, the gene models were scanned for typical NLR domains using hmmscan from the HMMER package (Finn, Clements and Eddy, 2011) in the filtered protein-coding gene models produced by EVM. In the second, the entire genome was scanned *de novo* for complete, functional NLR protein-coding regions using NLRAnnotator (Steuernagel *et al.*, 2020). This allowed several additional NLR genes whose gene model had been filtered out or was never constructed to be recovered. Other types of genes that have been previously implicated in rust resistance (Zheng *et al.*, 2020), such as ABC transporters, Pkinases, hexose transporters, wheat Kinase-START genes and tandem kinase-pseudokinase proteins, were identified from the eggnog functional annotation and validated by searching for pfam domains characteristic of each class of gene using hmmscan from hmmer v3.3.

Genes identified as potential resistance genes and identified as being introgressed above were manually checked using IGV to identify candidates with even sequencing coverage across the genes in DH92 and DH121 only, in the case of the shared stripe rust resistance, and across the genes in DH92 only, in the case of the DH92-specific leaf and stem rust

resistance. To reduce the number of genes to manually check, I removed any genes with less than 2x mean mapping coverage across their length in either DH92 or DH121. The gene models were manually curated using the available evidence. For NLR predictions with no gene model but transcriptomic and *ab initio* evidence, gene models were manually constructed. The novelty of the uniquely introgressed NLRs was tested by extracting the NB-ARC domains using hmmscan from hmmer v3.3 and aligning them using BLASTp from blast+ v 2.7.1 to the proteins of high-confidence genes from Chinese Spring and 15 other wheat cultivars (Walkowiak *et al.*, 2020; White *et al.*, 2024) and to the proteins of the wheat progenitor species: *Ae. tauschii* (Luo *et al.*, 2017); *T. urartu* (Ling *et al.*, 2018); and *T. turgidum* ssp. *dicoccoides* (Avni *et al.*, 2017). Hits below 85% identity were considered novel. The novelty of the other protein types was tested by aligning the whole amino acid sequence to the same protein set; proteins below <80.0% were considered novel.

3 Heat stress tolerance in field conditions derived from exotic alleles and *Ae. tauschii* introgression

This chapter is an adaptation of work that has been published in *Communications Biology* (Molero *et al.*, 2023) (Appendix D2) and appears with permission granted by the Creative Commons Attribution 4.0 International License.

This work was a collaboration between the Anthony Hall group at the Earlham Institute and the Wheat Physiology Group at CIMMYT, led by Matthew Reynolds. Gemma Molero and the other authors from CIMMYT (Francisco Pinto, Francisco J Piñera-Chávez and Carolina Rivera-Amado) grew and phenotyped the plants. Ryan Joynson conducted the GWAS analysis and discovered the effect of the three favourable alleles on yield and canopy temperature. I identified introgressed material, narrowed down the interval and conducted candidate gene work. I also conducted the statistical analyses presented here, under the advice of Gemma Molero. I made all of the figures, except for Fig. 3-6, which was made by Ryan Joynson. Figs. 3-1, 3-2, 3-3 and 3-5 were based on mock-ups made by Gemma Molero, and Fig. 3-7 was based on a mock-up made by Ryan Joynson.

3.1 Abstract

Increased heat stress driven by global warming is a major threat to wheat productivity. This necessitates the identification of heat tolerant alleles and the subsequent development of new wheat varieties that are more resilient to future climatic change. CIMMYT have developed an association mapping panel, HIBAP I, to represent the diversity found in their spring wheat collection, which includes exotic material such as synthetic-derived lines and landraces. HIBAP I was evaluated in the field under heat stress and yield potential conditions, demonstrating that exotic-derived lines performed better under heat stress than Elite lines. This phenotype data from the field was then combined with SNPs identified using enrichment capture sequencing data to conduct a genome-wide association study to reveal three exotic-derived pleiotropic loci underlying this heat tolerance which together boost yield by over 50% and reduce canopy temperature by around 2 °C. I then identified an *Ae. tauschii* introgression underlying the locus of largest effect size, which confers around 32.2% increased yield under heat stress. Finally, I extracted candidate genes and demonstrate the limitations of relying on the wheat

reference genome, particularly when divergent introgressed material is underlying the trait under investigation.

3.2 Introduction

3.2.1 Heat stress as a major challenge for future global wheat production

Heat is one of the major abiotic stresses that threatens wheat production. Even in the absence of future global warming, unusually hot seasons can be disastrous for wheat yield. For example, in 2010, during Russia's hottest summer in 130 years, wheat yield was reduced by 30% (Wegren, 2011). As global temperatures rise in coming years, heat stress will become an increasingly significant problem, with a predicted yield reduction of 6% for every 1°C rise in temperature and 9.1% ± 5.4% per 1°C rise in the most affected regions (Asseng *et al.*, 2014). Rising temperatures will also likely increase the occurrence of other abiotic stressors such as drought, salinity, and waterlogging (Wang, Vinocur and Altman, 2003; Chapman *et al.*, 2012; Lamaoui *et al.*, 2018; Lawas *et al.*, 2018; Nasser *et al.*, 2020; Surówka, Rapacz and Janowiak, 2020) as well as pathogens and pests whose range, lifestyle, and interactions with crops may change with the climate (Garrett *et al.*, 2006; Classen *et al.*, 2015; Surówka, Rapacz and Janowiak, 2020).

Both continual heat stress throughout the growing season and heat stress restricted to the reproductive phases (known as terminal heat stress) (Nesar *et al.*, 2022) negatively impact yield; both of which are predicted to increase in frequency and intensity in the future (Reynolds *et al.*, 2016). Heat stress negatively impacts yield by targeting a variety of physiological processes in plants (Akter and Rafiqul Islam, 2017). Heat stress impacts the vegetative, reproductive and grain filling phases of the crop cycle, although the reproductive and grain filling phases are especially susceptible to heat stress which is why terminal heat stress can be so devastating. In particular, pre-flowering and anthesis are the stages most affected by heat stress (Cossani and Reynolds, 2012; Bheemanahalli *et al.*, 2019).

Heat stress reduces the efficiency of photosynthesis and respiration, reduces leaf area and crop duration and accelerates leaf senescence (Reynolds *et al.*, 2016; Balla *et al.*, 2019; Degen, Orr and Carmo-Silva, 2021). Temperatures greater than 30°C during pollen development causes pollen abortion (Begcy *et al.*, 2018; Ullah *et al.*, 2022) and heat stress during grain development shortens the duration of grain filling (Dias, Bagulho and Lidon,

2008; Dias, Lidon and Ramalho, 2009; Ullah *et al.*, 2022) and lowers the accumulation of starch and protein in the grain (Ullah *et al.*, 2022) by lowering the rate of starch biosynthesis (Begcy *et al.*, 2018). This results in reduced floret fertility and grain weight (Reynolds *et al.*, 2016), critical components of yield, which have been found to be reduced by 40.17-41.19% under heat stress induced by delayed sowing (Shenoda *et al.*, 2021). High temperatures can also reduce the duration of the crop cycle. Heading dates are predicted to advance, on average, by 1 week between 2020-2049 and by 2-3 weeks by 2100 (Gouache *et al.*, 2012). A shorter grain-filling stage, less time between heading and maturity (Mohammadi *et al.*, 2012), and more generally less time between successive phenological events (Zahedi and Jenner, 2003) will also lead to reduced yield.

3.2.2 HiBAP I – CIMMYT's high biomass association panel

Exploring available germplasm for heat tolerance traits is an important step towards identifying genetic variation underlying such traits. CIMMYT have developed a spring wheat association mapping panel, HiBAP I (High Biomass Association Panel) (Molero *et al.*, 2019) that is representative of the diversity contained within CIMMYT's 75000 wheat cultivar collection (Lyra *et al.*, 2021). HiBAP I contains 149 lines which are either high-yielding Elite varieties or lines with a pedigree history including crosses with exotic parents (Fig. 3-1). These exotic parents include synthetic-derived lines, which possess up to 43% donor DNA from *Ae. tauschii* and durum wheat (Joynson *et al.*, 2021); introgression lines with introgressions from *S. cereale* (Rye) and/or *Th. Ponticum* (Joynson *et al.*, 2021); and landraces from Mexico and India. The Elite lines in the panel include 11 varieties released by CIMMYT between 1966 and 2007 and 42 lines identified during screening of CIMMYT breeding and pre-breeding material for high biomass/radiation use efficiency across different growth stages. Lines were chosen from a pre-panel of over 250 lines if, under yield potential field conditions (optimal conditions for maximising yield), the plants presented a favourable agronomic background and had similar height and phenology (time of transition between phenological stages). This was done to reduce the confounding effects that extreme height or phenology may have on traits of interest.

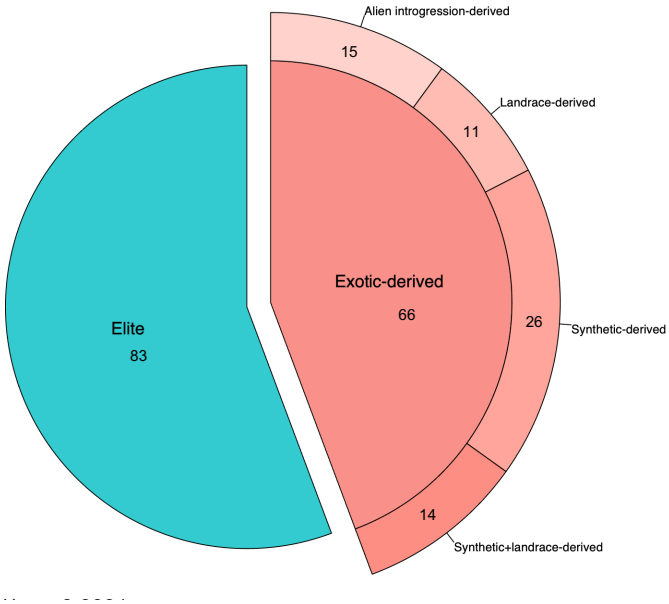


Figure 3-1. Number of lines in HiBAP I from each group.

This panel was developed as part of CIMMYT's long-standing objective to incorporate exotic variation to develop varieties of wheat more tolerant to abiotic stress and to support developing nations and future proof crop production against significant environmental changes. The panel is a good potential source of variation underlying abiotic stress tolerance as synthetic lines and landraces in CIMMYT germplasm have previously been identified that produce higher yield under heat stress and drought when compared to Elite varieties (Lopes and Reynolds, 2011; Cossani and Reynolds, 2015). However, only a small proportion of these promising genetic resources have been properly evaluated for climatic tolerance and as described in section 1.9, despite the promise of this material, breeders tend to avoid exotic material due to reduced recombination and possible linkage drag of introgressed segments (McCouch *et al.*, 2020).

3.2.3 Genome-wide association studies and typical downstream approaches for refining intervals

Genome-wide association studies aim to understand the genetic bases of complex traits by looking for associations between genetic markers and phenotypic variation across a population of individuals (Brachi, Morris and Borevitz, 2011). The central output of a GWAS is a set of significant MTAs. Each MTA consists of a single SNP within an interval of linked SNPs that are significantly associated with phenotypic variation in a given trait.

When the resolution of genotyping is high enough, provided population structure and size permit it, it can be possible to narrow down the MTA interval to a single gene (Brachi, Morris and Borevitz, 2011). However, in most cases, particularly when using a genotyping array or enrichment capture sequencing to generate genotyping data, population size is small, and/or linkage disequilibrium is high around the favourable allele in the population studied, a larger interval will be identified. It is also often the case that the causal SNP falls within intergenic regulatory regions, preventing the gene under regulation being identified.

In the context of pre-breeding, GWAS are useful for traits governed by a few genes of major effect, compared with traits like grain yield which are usually governed by many loci of minor effect. MTAs identified in a GWAS can be deployed in breeding programmes through marker-assisted selection (K. Singh *et al.*, 2021). They can also act as a starting point for fine mapping to identify the most important variants and genes underlying the variation in the trait (Broekema, Bakker and Jonkers, 2020), and can help lead to the elucidation of the molecular mechanism underlying the trait.

Following a GWAS, the identified intervals containing MTAs are typically searched to identify candidate genes, variation in which may be underlying the variation in phenotype for the trait under investigation. These then become priorities for further study. To guide the selection of candidate genes, previous research on the gene or orthologues of the gene in other species is examined to determine if they have previously been implicated in the trait under investigation. Additionally, genes harbouring functionally-relevant mutations, such as those that alter the amino sequence, are more likely to be chosen as candidates. Biparental mapping populations can be generated to validate MTAs and, in conjunction with forward genetic screening, reduce the size of the interval and identify the causal gene(s).

3.2.4 Chapter aims

- Analyse physiological data gathered in the field, comparing exotic-derived and Elite lines under heat stress and yield potential conditions.
- Conduct genome-wide association study to uncover loci underlying heat stress tolerance.
- Use mapping coverage and SNP information to identify introgressed material underlying any of the MTAs.

- Identify candidate genes underlying the MTAs and explore limitations of relying on a single reference genome.

3.3 Results

3.3.1 Physiological analysis of HiBAP I in heat stress and yield potential environments.

HiBAP I was grown and evaluated for two consecutive years under both yield potential and heat stress conditions, which were conferred by delaying sowing for 3 months. Across the two years, the heat stress severely impacted almost all measured physiological traits (Fig. 3-2). For example, yield was 48.1% lower than in the yield potential conditions, the duration of the crop cycle was 30.6% shorter, and there were 26.8% more infertile spikelets per spike.

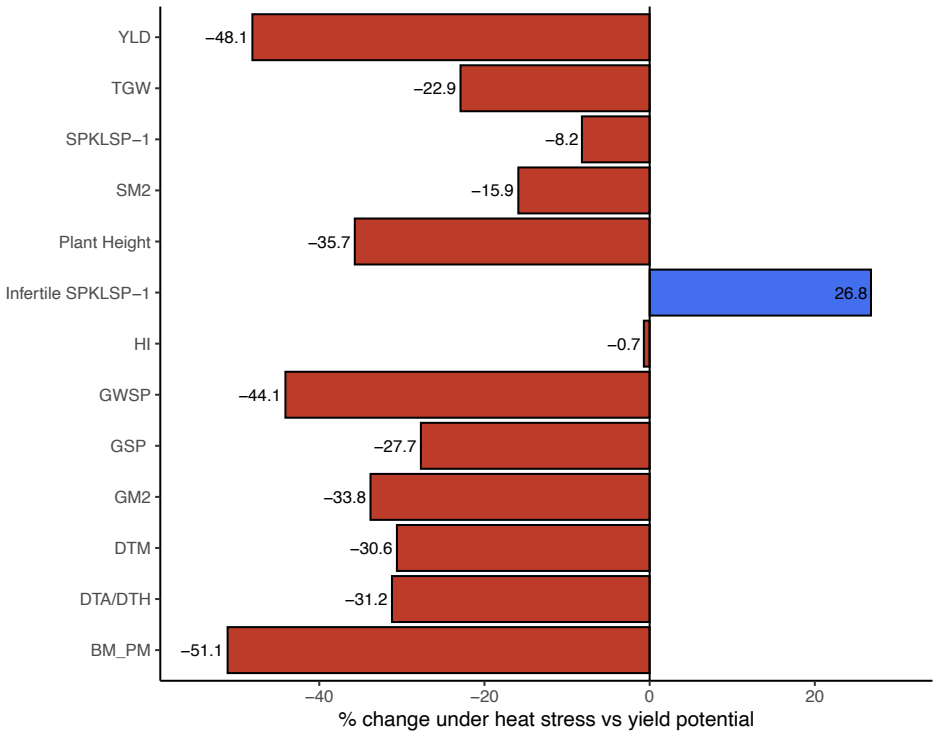


Figure 3-2. Effect of heat stress on various physiological traits.
Looks at yield (YLD), thousand grain weight (TGW), number of spikelets per spike (SPKLSP^{-1}), number of spikes per m^2 (SM2), plant height (Height), number of infertile spikelets per spike (infertile SPKLSP^{-1}), harvest index (HI), grain weight per spike (GWSP), number of grains per spike (GSP), grain number (GM2), days to physiological maturity (DTM), days to anthesis or days to heading (DTA/DTH for yield potential and heat stress experiments respectively), and biomass at physiological maturity (BM_PM), showing the percentage difference yield potential conditions compared to heat stress conditions.

To determine whether Elite and exotic-derived lines respond differently to heat stress, I compared trait values of Elite and exotic-derived lines under heat stress and yield potential conditions (Fig. 3-3, Table 3-1). This revealed that exotic-derived lines have significantly higher yield than Elite lines in the heat stress conditions but have the same yield under yield potential conditions. Exotic-derived lines had significantly higher thousand grain weight under both conditions, higher grain number under heat stress conditions but lower grain number under yield potential conditions, higher biomass under both conditions, and were taller under both conditions. Elite lines had higher harvest index under yield potential conditions but there was no difference under heat stress conditions.

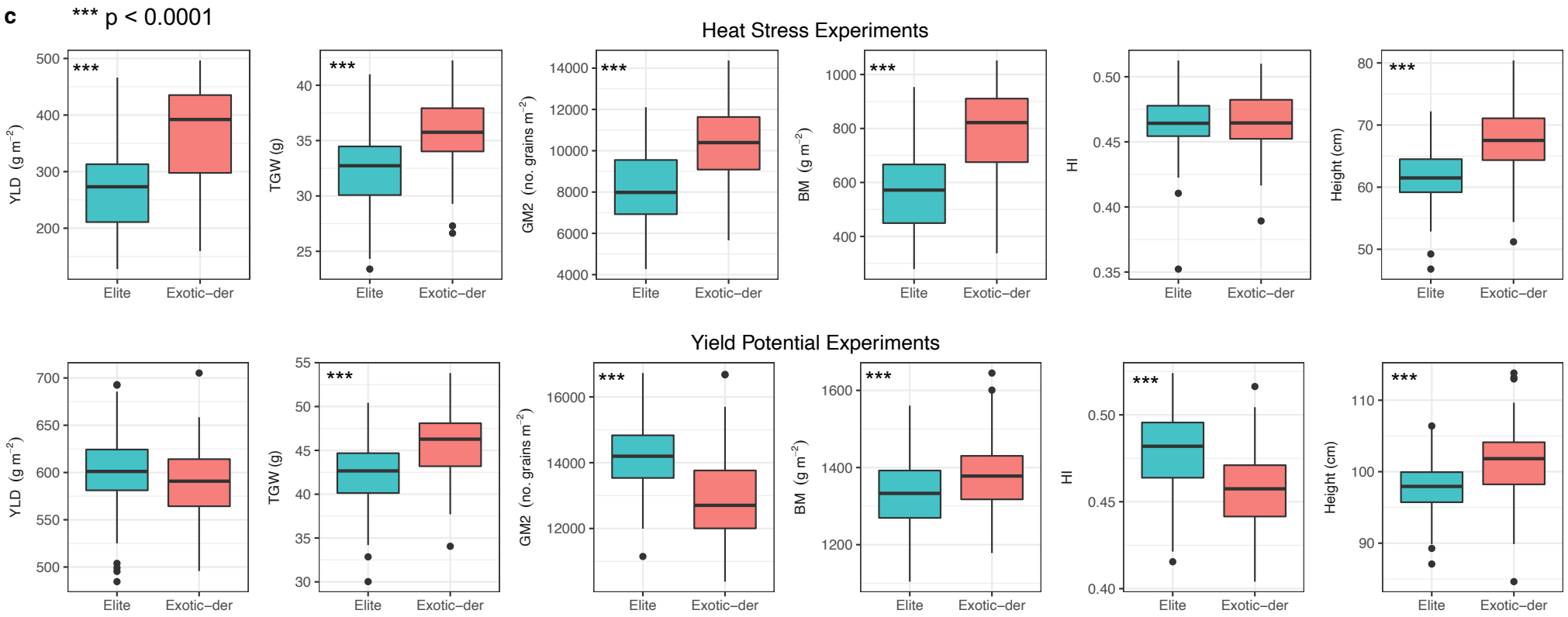

c
*** p < 0.0001
Heat Stress Experiments
Yield Potential Experiments
YLD (g m^{-2})
TGW (g)
GM2 (no. grains m^{-2})
BM (g m^{-2})
HI
Height (cm)
***
Elite
Exotic−der

Figure 3-3. Comparison of physiological traits between Elite and exotic-derived lines in HiBAP I measured under both heat stress and yield potential conditions.
Measured traits are yield (YLD), thousand grain weight (TGW), grain number (GM2), biomass at physiological maturity (BM_PM), harvest index (HI), and Height. The boxplots are defined as follows: centre line = median; box limits = upper and lower quartiles, whiskers = 1.5x interquartile range; points = outliers. The significance of the difference between Elite (*n* = 83 biologically independent lines) and exotic-derived (*n* = 66 biologically independent lines) lines for each trait was assessed using two-tailed *t* tests with no assumption of equal variance. *p*-values below 0.01 were considered significant (*), below 0.001 very significant (**) and below 0.0001 highly significant (***). Means, standard deviations, confidence intervals and *p*-values can be found in Table 3-1.

Table 3-1. Summary statistics for physiological traits measured under heat stress or yield potential conditions over 2 years. Confidence intervals and p-values were calculated using two-tailed t tests with no assumption of equal variance to compare Elite lines (N=83) with exotic-derived lines (N=66).

Trait	Condition	Elite mean ± S.D	Exotic mean ± S.D	95% confidence interval	p-value
YLD (g m^{-2})	Heat stress	265±69.6	365±88.1	-126, -73.6	9.61e-12
TGW(g)	Heat stress	32.3±3.36	35.7±3.27	-4.43, -2.28	7.66e-09
GM2(no. grains m^{-2})	Heat stress	8120±1913	10246±2054	-2776, -1476	1.685e-09
BM_PM (g m^{-2})	Heat stress	565±138	786±169	-272, -170	3.38e-14
HI	Heat stress	0.465±0.0232	0.466±0.0229	-0.0009151, 0.000588	0.668
Height (cm)	Heat stress	61.6±4.55	67.2±6.03	-7.34, -3.80	7.69e-09
YLD (g m^{-2})	Yield potential	560±42.0	590±35.6	-2.51, 22.6	0.116
TGW(g)	Yield potential	42.4±3.95	45.7±4.06	-4.69, -2.07	1.03e-06
GM2(no. grains m^{-2})	Yield potential	14228±1066	12986±1374	835, 1649	1.78e-08
BM_PM (g m^{-2})	Yield potential	1331±84.4	1385±97.2	-83.6, -23.7	5.54e-04
HI	Yield potential	0.478±0.0242	0.456±0.0225	0.0140 ,0.0291	9.21e-08
Height (cm)	Yield potential	97.9±3.99	102±4.98	-5.22, -2.24	2.41e-06

3.3.2 The relationship between yield and canopy temperature/NDVI under heat stress

During both vegetative and grain filling stages, exotic-derived lines had significantly higher normalised difference vegetation index (NDVI) and significantly lower canopy temperature than Elite lines under heat stress but not under yield potential conditions (Fig. 3-4, Table 3-2). NDVI was significantly positively correlated with yield and canopy temperature was significantly negatively correlated with yield during both vegetative and grain filling stages but only under heat stress conditions (Fig. 3-5, Table 3-3). The traits were not correlated under yield potential conditions. The exotic-derived lines had higher correlation coefficients than Elite lines under heat stress, suggesting that they had a stronger relationship between NDVI and yield and canopy temperature and yield.

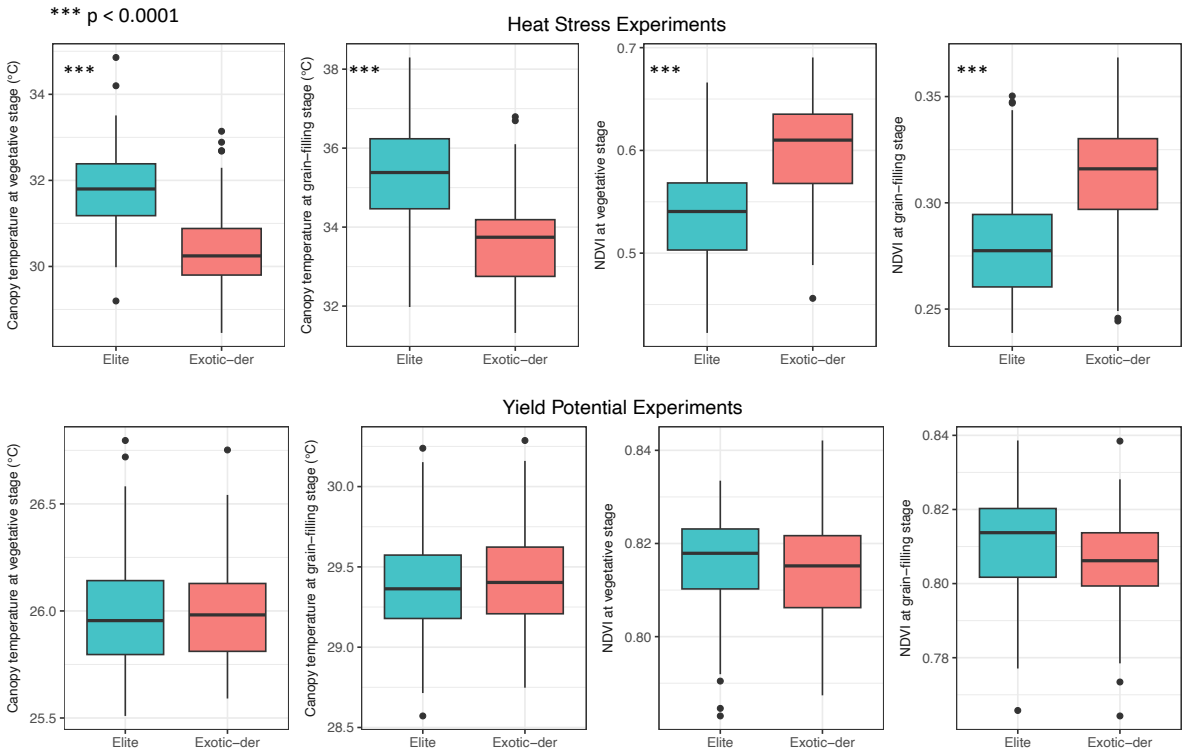


Figure 3-4. Comparison of canopy temperature and NDVI at the vegetative and grain-filling stages between Elite and exotic-derived lines in HiBAP I measured under both heat stress and yield potential conditions.
The boxplots are defined as follows: centre line = median; box limits = upper and lower quartiles, whiskers = 1.5x interquartile range; points = outliers. The significance of the difference between Elite (*n* = 83 biologically independent lines) and exotic-derived (*n* = 66 biologically independent lines) lines for each trait was assessed using two-tailed *t* tests with no assumption of equal variance. *p*-values below 0.01 were considered significant (*), below 0.001 very significant (**) and below 0.0001 highly significant (***). Means, standard deviations, confidence intervals and *p*-values can be found in Table 3-2.

Table 3-2. Summary statistics for canopy temperature and NDVI under heat stress or yield potential conditions over 2 years, during the vegetative and the grain-filling phenological stages. Confidence intervals and p-values were calculated using two-tailed t tests with no assumption of equal variance to compare Elite lines (N=83) with exotic-derived lines (N=66).

Trait	**Phenological Stage**	**Condition**	**Elite mean ± S.D**	**Exotic mean ± S.D**	**95% confidence interval**	**p-value**
CT (°C)	Vegetative	Heat stress	31.8±0.987	30.5±1.05	1.05, 1.71	1.681e-13
CT (°C)	Grain-filling	Heat stress	35.3±1.09	33.7±1.27	1.23, 2.01	2.122e-13
NDVI	Vegetative	Heat stress	0.535±0.0467	0.603±0.0530	-0.0838, -0.0510	3.079e-13
NDVI	Grain-filling	Heat stress	0.280±0.0248	0.311±0.0290	-0.0404, -0.0227	1.133e-10
CT (°C)	Vegetative	Yield potential	26.0±0.271	26.0±0.240	-0.0826, 0.0831	0.9949
CT (°C)	Grain-filling	Yield potential	29.4±0.331	29.4±0.305	-0.164, 0.0428	0.2488
NDVI	Vegetative	Yield potential	0.816±0.0105	0.814±0.0110	-0.00154, 0.00545	0.2712
NDVI	Grain-filling	Yield potential	0.810±0.0140	0.806±0.0132	1.93e-06, 8.85e-03	0.0499

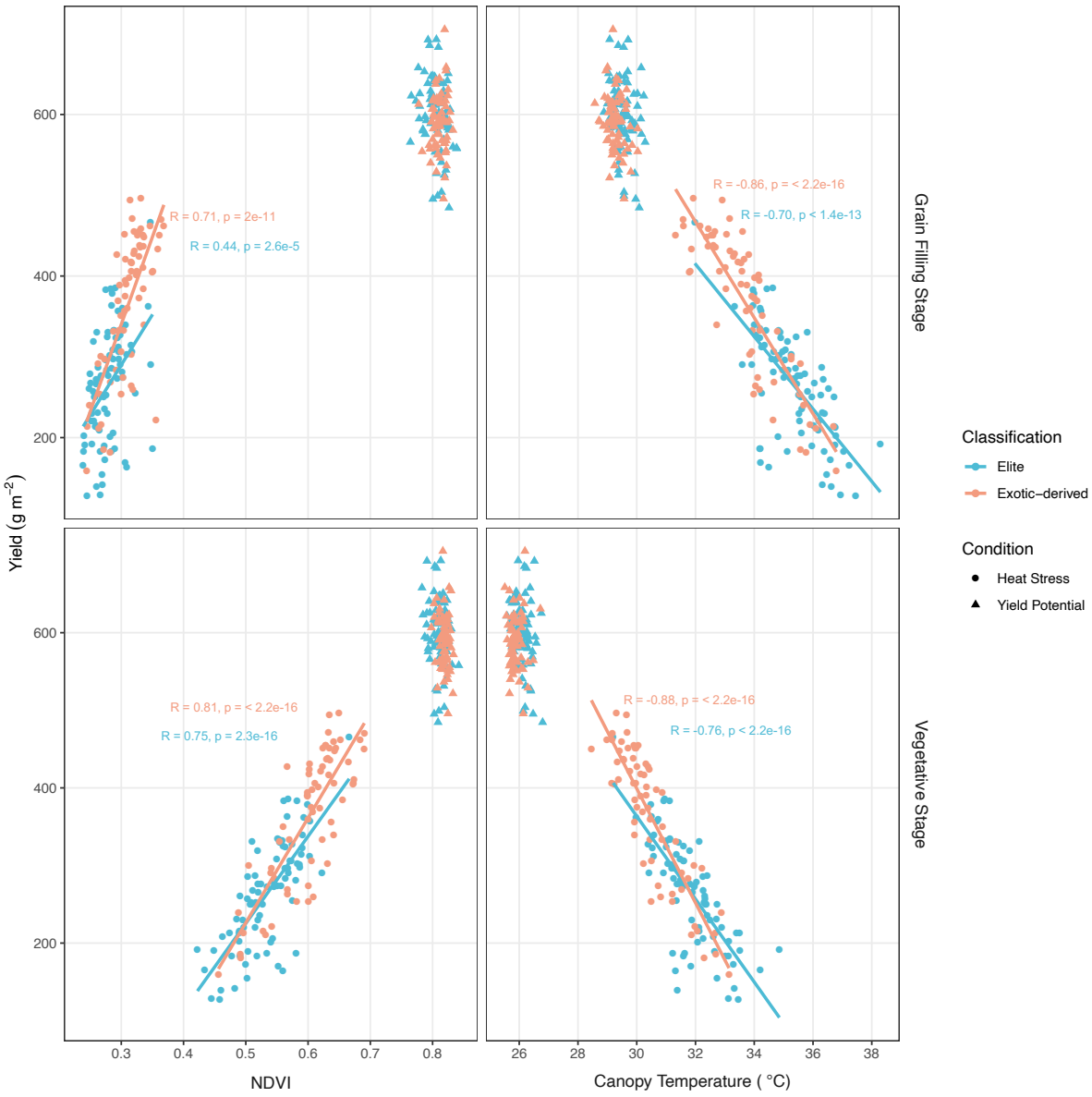


Figure 3-5. NDVI and canopy temperature were measured with UAVs at pre-heading (vegetative stage) and during grain filling.

Regression lines were calculated using Pearson's correlation coefficient between each pair of traits (n = 83 and 66 biologically independent lines for the Elite and exotic-derived groups, respectively) and added for classification/condition combinations with a significant correlation (p-value < = 0.01). The correlation coefficient, r, and the steepness of the line, ranges from −1 to 1, signifying very negatively correlated and very positively correlated, respectively. Correlation coefficients, confidence intervals and p-values can be found in Table 3-3.

Table 3-3. Pearson's correlation tests between NDVI and Yield and between Canopy temperature (CT) and yield under either heat stress or yield potential conditions and at either the vegetative or grain filling phenological stage.

Trait	Condition	Phenological stage	Line classification	Pearson's correlation coefficient (r)	95% confidence interval	p-value
NDVI	Heat stress	Vegetative	Elite	0.753	0.641, 0.833	2.29e-16
NDVI	Heat stress	Vegetative	Exotic-derived	0.814	0.712, 0.882	<2.2e-16
NDVI	Yield potential	Vegetative	Elite	-0.197	-0.396, 0.0197	0.0745
NDVI	Yield potential	Vegetative	Exotic-derived	-0.187	-0.410, 0.0579	0.133
NDVI	Heat stress	Grain Filling	Elite	0.444	0.252, 0.602	2.64e-05
NDVI	Heat stress	Grain Filling	Exotic-derived	0.712	0.568, 0.814	2.00e-11
NDVI	Yield potential	Grain Filling	Elite	-0.256	-0.447, -0.0425	0.0196
NDVI	Yield potential	Grain Filling	Exotic-derived	0.0943	-0.151 ,0.329	0.451
CT (°C)	Heat stress	Vegetative	Elite	-0.758	-0.837, -0.648	<2.2e-16
CT (°C)	Heat stress	Vegetative	Exotic-derived	-0.875	-0.922, -0.804	<2.2e-16

Table 3-3

CT (°C)	Yield potential	Vegetative	Elite	-0.260	-0.450, -0.0467	0.0177
CT (°C)	Yield potential	Vegetative	Exotic-derived	0.00350	-0.239, 0.245	0.978
CT (°C)	Heat stress	Grain Filling	Elite	-0.702	-0.797, -0.573	1.44e-13
CT (°C)	Heat stress	Grain Filling	Exotic-derived	-0.859	-0.912, -0.779	<2.2e-16
CT (°C)	Yield potential	Grain Filling	Elite	-0.206	-0.403, 0.0105	0.0621
CT (°C)	Yield potential	Grain Filling	Exotic-derived	-0.265	-0.476, -0.0243	0.0317

3.3.3 Genome-wide association study to identify genetic loci associated with heat tolerance

To identify genomic loci associated with the physiological traits under heat stress conditions, my colleague Ryan Joynson conducted a GWAS using genotyping data generated in (Joynson *et al.*, 2021) using enrichment capture Illumina paired-end reads, and the physiological data collected in the field. This GWAS revealed three marker trait associations (MTAs) at chr1B-63398861 (favourable allele = C; interval = 0.6-10 Mbp), chr2B-820002 (favourable allele = C; interval = chr2B:0-1 Mbp) and chr6D-6276646 (favourable allele = T; interval = chr6D:3-7.5 Mbp) (Molero et al., 2021) (Fig. 3-6). These MTAs were associated with many traits under heat stress (Table 3-4), including 5 yield traits, 3 stress tolerance indices, and NDVI and canopy temperature at both vegetative and grain-filling stages. The MTAs were not associated with phenological traits, suggesting that the observed heat tolerance was not driven by alterations to the timing of phenological stages.

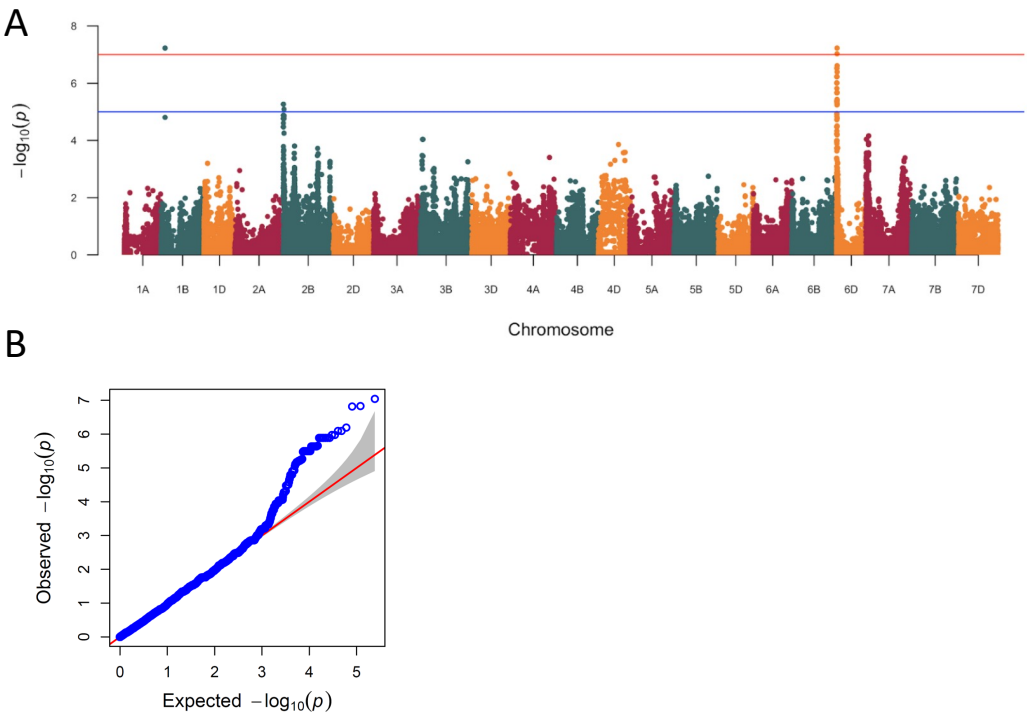


Figure 3-6. GWAS results for stress susceptibility index (SSI) under heat stress, one of the traits studied as an example.

A) Manhattan plot showing the -$\log_{10}$ of p-values for each SNP, sorted by genomic location. The horizontal blue line indicates an arbitrary cutoff of $-\log_{10}(p)$ of 5. The horizontal red line indicates the conservative Benjamini–Hochberg cutoff implemented by GAPIT. **B)** Quantile-quantile (Q-Q) plot for the GWAS. This Q-Q plot illustrates the distribution of p-values obtained from the GWAS for this trait. The x-axis represents the -$\log_{10}$ of expected p-values under the null hypothesis, assuming no true genetic associations. The y-axis represents the -$\log_{10}$ of observed p-values resulting from the GWAS analysis. The black dots represent the observed p-values of each SNP, sorted in ascending order by -$\log_{10}$ p-value. Dots deviating from the red line indicate significant associations between genetic markers and the trait values.

Table 3-4. Summary of Marker-Trait Associations (MTAs) for different physiological traits.

Trait	Chromosome	MTA ID	Position	p-value	Interval
Yield Traits					
YLD	chr1B	chr1B-63398861	63398861	8.32e-08	0.6-10 Mbp
	chr6D	chr6D-6276646	6276646	4.16e-07	3-7.5 Mbp
BM	chr1B	chr1B-63398861	63398861	3.4e-08	0.6-10 Mbp
	chr2B	chr2B-820002	820002	0.00000434	1 Mbp
	chr6D	chr6D-6276646	6276646	1.94e-07	3-7.5 Mbp
GFR	chr1B	chr1B-63398861	63398861	6.55e-07	0.6-10 Mbp
	chr6D	chr6D-6276646	6276646	0.00000439	3-7.5 Mbp
GM2	chr1B	chr1B-63398861	63398861	6.4806e-06	0.6-10 Mbp
	chr6D	chr6D-6276646	6276646	2.8069e-07	3-7.5 Mbp
SM2	chr1B	chr1B-63398861	63398861	4.1647e-07	0.6-10 Mbp
Stress Tolerance Index					
Stress_intYLD	chr1B	chr1B-63398861	63398861	5.9082e-08	0.6-10 Mbp
	chr2B	chr2B-820002	820002	5.4266e-06	1 Mbp
	chr6D	chr6D-6276646	6276646	5.9158e-08	3-7.5 Mbp
Stress_intBM	chr1B	chr1B-63398861	63398861	3.1404e-08	0.6-10 Mbp
	chr2B	chr2B-820002	820002	7.1692e-06	1 Mbp
	chr6D	chr6D-6276646	6276646	3.1404e-08	3-7.5 Mbp
Stress_intGM2	chr1B	chr1B-63398861	63398861	2.6569e-08	0.6-10 Mbp
	chr2B	chr2B-820002	820002	7.1074e-07	1 Mbp
	chr6D	chr6D-6276646	6276646	6.6123e-08	3-7.5 Mbp
UAV measurements					
UAV_CTvg_AV	chr1B	chr1B-63398861	63398861	1.8924e-07	0.6-10 Mbp

Table 3-4

	chr2B	**chr2B-820002**	**820002**	**9.0008e-06**	**1 Mbp**
	chr6D	chr6D-6276646	6276646	6.1095e-06	3-7.5 Mbp
UAV_CTgf_AV	chr1B	chr1B-63398861	63398861	1.7673e-07	0.6-10 Mbp
	chr2B	chr2B-820002	820002	2.2766e-06	1 Mbp
	chr6D	chr6D-6276646	6276646	1.8452e-07	3-7.5 Mbp
UAV_NDVIvg_AV	chr1B	chr1B-63398861	63398861	4.0814e-07	0.6-10 Mbp
	chr2B	chr2B-820002	820002	6.4281e-06	1 Mbp
UAV_NDVIgf_AV	chr6D	chr6D-6276646	6276646	5.7295e-06	3-7.5 Mbp

3.3.4 Effect of allele combinations on yield and canopy temperature

The favourable C allele at the chr1B and chr2B MTAs always co-occur. Lines with both of these alleles and the unfavourable A allele at the chr6D MTA have 24.3% higher yield under heat stress compared to lines with unfavourable allele at each of the three MTAs. Lines which also have the favourable T allele at the chr6D MTA have 56.5% higher yield under heat stress compared to lines with the unfavourable allele at each of the three MTAs (Fig. 3-7, Table 3-5). Assuming the three alleles do not interact epistatically and the T allele on chr6D can function independently of the other alleles, the T allele on chr6D can be assumed to increase yield under heat stress by 32.4%. Lines with the favourable allele at all three MTAs show a reduction in canopy temperature of 1.97 °C and 2.37 °C, at vegetative and grain filling stages, respectively, when compared to lines with the unfavourable allele at all three positions (Fig. 3-7, Table 3-5). Under yield potential conditions, no difference was observed between favourable and unfavourable allele combinations for yield or for canopy temperature (Fig. 3-7, Table 3-5).

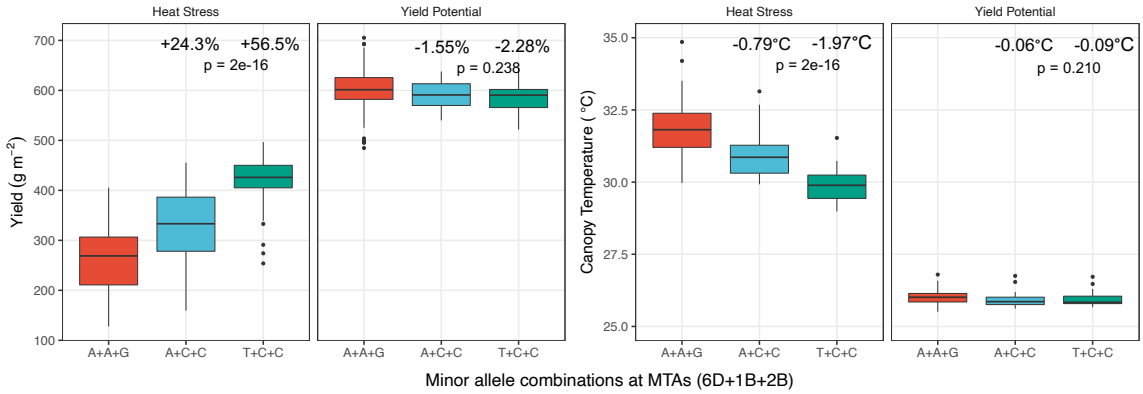


Figure 3-7. Effect of allele combinations from the three MTAs (chr6D-6276646, chr1B-63398861, and chr2B-820002) on yield and on canopy temperature under heat stress and yield potential conditions.

The combination of favourable alleles is T+C+C and the combination of unfavourable alleles is A+A+G. The boxplots are defined as follows: centre line = median; box limits = upper and lower quartiles, whiskers = 1.5x interquartile range; points=outliers. The percentage change and °C change is calculated compared to lines with the unfavourable alleles at all each of the three MTAs. The significance of the different between allele combinations was computed using a one-way ANOVA test (*n* = 87, 14, and 31 biologically independently lines for A+A+G, A+C+C and T+C+C, respectively). Means, standard deviations and *p*-values from Tukey's honest significance test can be found in Table 3-5.

Table 3-5. Yield and canopy temperature of the different allele combinations at the MTAs at chr6D-6276646, 1B chr1B-63398861, and 2B chr2B-820002. The combination of favourable alleles is T+C+C and the combination of unfavourable alleles is A+A+G. The significance of allele combinations was computed using a one-way ANOVA test and post-hoc test using Tukey's honest significance test.

			Mean ± S.D			**Tukey's HSD p-values**		
Trait	**Condition**	**Anova p-value**	**A+A+G**	**A+C+C**	**T+C+C**	A+C+C-A+A+G	T+C+C-A+A+G	T+C+C-A+C+C
Yield	Heat stress	<2e-16	263±67.3 g m^{-2}	327±79.5 g m^{-2}	411±59.8 g m^{-2}	2.58e-3	0.00	3.18e-4
Yield	Yield Potential	0.238	600±44.8 g m^{-2}	591±28.5 g m^{-2}	587±27.7 g m^{-2}	0.681	0.242	0.937
CT	Heat stress	<2e-16	31.8±0.987 °C	31.0±0.963 °C	29.9±0.556 °C	5.97e-3	0.00	2.21e-4
CT	Yield Potential	0.210	26.0±0.255 °C	26.0±0.318 °C	25.9±0.243 °C	0.676	0.211	0.921

The favourable allele at each of these MTA sites is almost exclusively found in exotic-derived lines. 50/55 (chr1B), 44/45 (chr2B) and 33/33 (chr6D) lines that possess the homozygous favourable allele at the chr1B, chr2B and chr6D MTAs, respectively, are classified as exotic-derived. 7 lines have a heterozygous SNP call (A/T) at 6D-6276646, 3 of which are classified as exotic-derived. The HiBAP lines are inbred to the F9 or F10 generation, so most sites are expected to be homozygous. As the sequencing data was derived from pooled samples of plants from each line, the heterozygous calls could be caused by homozygous and heterozygous alleles segregating at this locus. If this were true, we would expect to see a phenotype that is intermediate between that possessed by lines with the favourable allele and that possessed by lines with the unfavourable allele. However, there is no significant difference in yield or canopy temperature between lines with the A/T genotype and lines with the T/T genotype (Fig. 3-8). Without a larger number of heterozygous lines, it is difficult to draw definitive conclusions, but it seems likely that the 7 lines with the A/T genotype called at 6D-6276646 truly have that genotype.

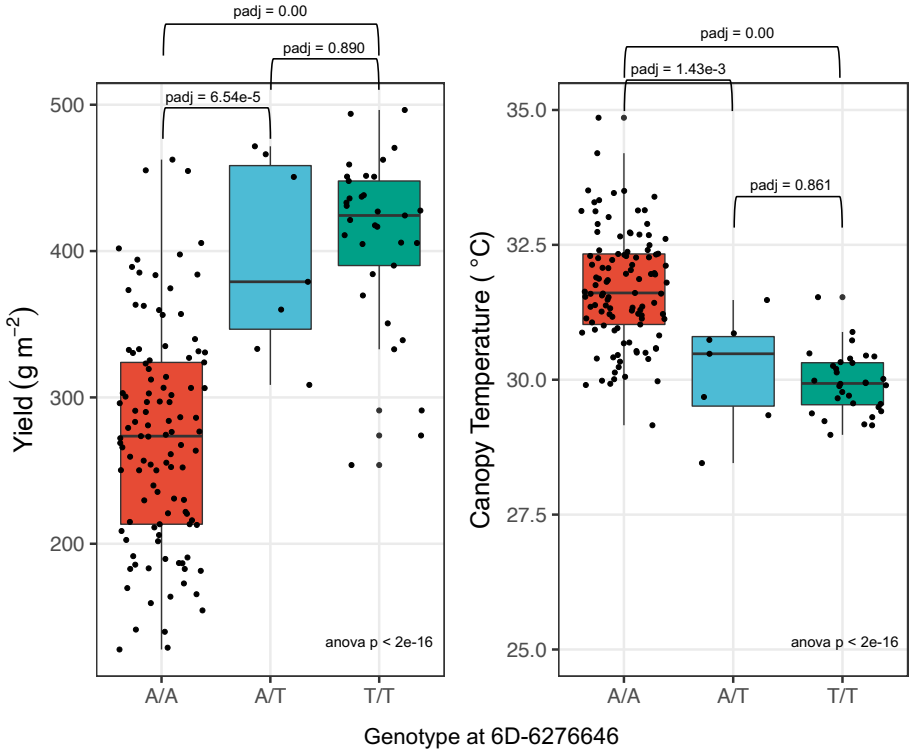


Figure 3-8. Yield and vegetative canopy temperature under heat stress conditions for lines with homozygous unfavourable allele (A/A), heterozygous for the favourable allele (A/T) and homozygous for the favourable allele (T/T).

Black points indicate individual data points; these were included due to the small sample size of lines with the A/T genotype. Significance was computed using a one-way ANOVA test (n=109, 7, and 32 biologically independently lines for A/A, A/T and T/T, respectively). Tukey's honest significance test was used to calculate adjusted p-values for each pairwise comparison.

3.3.5 Uncovering an *Ae. tauschii* introgression underlying the chr6D marker-trait association

The exotic material in the pedigree history of most of the lines possessing the favourable alleles suggests that the source of the alleles, and thus of the heat tolerance phenotype, may have been a wild or domesticated relative, either through historic introgressions or through CIMMYT's synthetic wheat programme.

As demonstrated in chapter two, deviation in mapping coverage of sequencing reads can be used to identify divergent regions of the genome in a sequenced line. However, in this case, due to the increased number of lines and lack of parental sequencing information, it is more appropriate to compare mapping coverage of each line to the median mapping coverage value for each window across the panel of 149 lines, instead of comparing mapping coverage to the parent lines as in chapter two. This number of lines also allows us to statistically determine outliers in mapping coverage using the outliers package in R.

To illustrate this, Fig. 3-9 shows the coverage deviation values across each chromosome for three HiBAP lines with large previously characterised introgressions (Ren *et al.*, 2009; Niu *et al.*, 2014; Joynson *et al.*, 2021). HiBAP_58 has a Rye introgression on chr1B between 0 and around 239 Mbp. HiBAP_39 has a *Th. ponticum* introgression on chr7D from around 340 Mbp to the end of the chromosome at 638.7 Mbp. HiBAP_2 contains both the Rye and *Th. ponticum* introgression. However, each of these shows intermediate levels of coverage deviation which suggests these introgressions are in a heterozygous state.

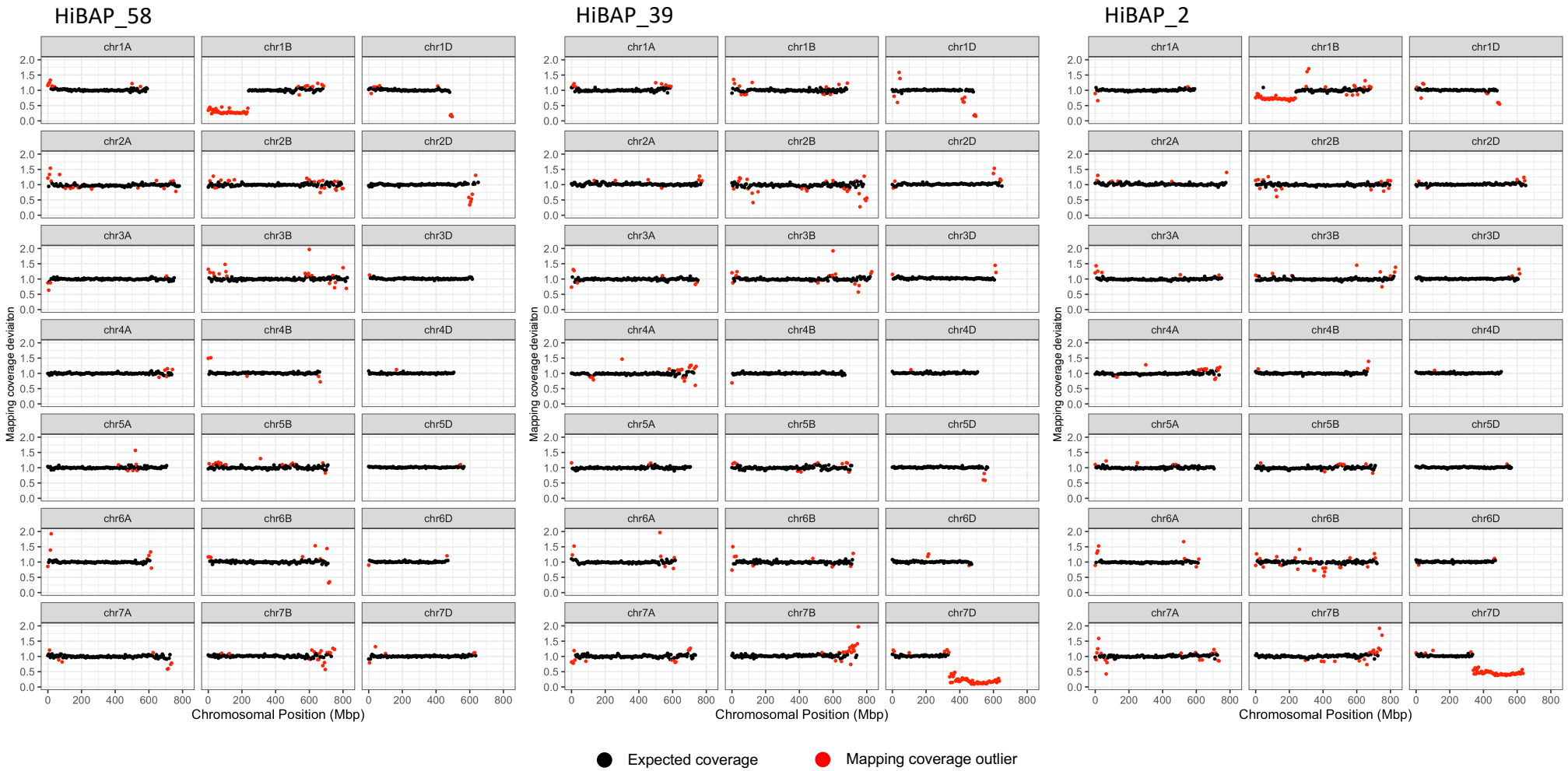

HiBAP_58
HiBAP_39
HiBAP_2
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D
Mapping coverage deviaiton
Chromosomal Position (Mbp)
Expected coverage
Mapping coverage outlier

Figure 3-9. Using mapping coverage deviation to identify divergent regions of the genome in sequenced lines from HiBAP I, using 5 Mbp genomic windows.
For each line, in each genomic window, mapping coverage deviation is calculated against the median of the panel for that genomic window. Red points are statistically significant outliers (n = 149 biologically independent lines). The lines displayed here, possess previously characterised introgressions: a Rye introgression on chr1B in HiBAP_58 and HiBAP_2 and a *Th. Ponticum* introgression on chr7D in HiBAP_39 and HiBAP_2. In HiBAP_2, the two introgressions appear to be heterozygous due to the intermediate level of mapping coverage deviation compared to the same introgressions in the other two lines.

Using this method, I searched for regions of reduced mapping coverage overlapping the three MTAs. This revealed a probable introgression at the start of chr6D, overlapping the chr6D MTA interval in multiple lines from HIBAP I, including all 33 lines with the T/T genotype and all 7 lines with the A/T genotype at MTA 6D-6276646. I found no evidence of introgressions overlapping the chr1B or the chr2B MTA intervals. To support the mapping coverage deviation and to identify the origin of the chr6D introgression, I used SNPs specific to wild/domesticated wheat relatives. I found that homozygous *Ae. tauschii* SNPs overlapped the region of reduced mapping coverage.

To illustrate this, I've used Sokoll (HiBAP_57), a synthetic-derived cultivar that has been released in Pakistan (Reynolds *et al.*, 2017). Blocks of mapping coverage deviation below 1 and homozygous *Ae. tauschii*-specific SNPs reveal several *Ae. tauschii* introgressions in the D subgenome which is expected due to its synthetic origin (Fig. 3-10). This includes *Ae. tauschii* at the start of chr6D, overlapping with the MTA. Due to *Ae. tauschii* belonging to wheat's primary genepool and being very genetically similar to the D subgenome, the reduction in mapping coverage seen within *Ae. tauschii* introgressions is much lower than seen in introgression from more distant species, such as the *Am. muticum* introgressions studied in chapter two.

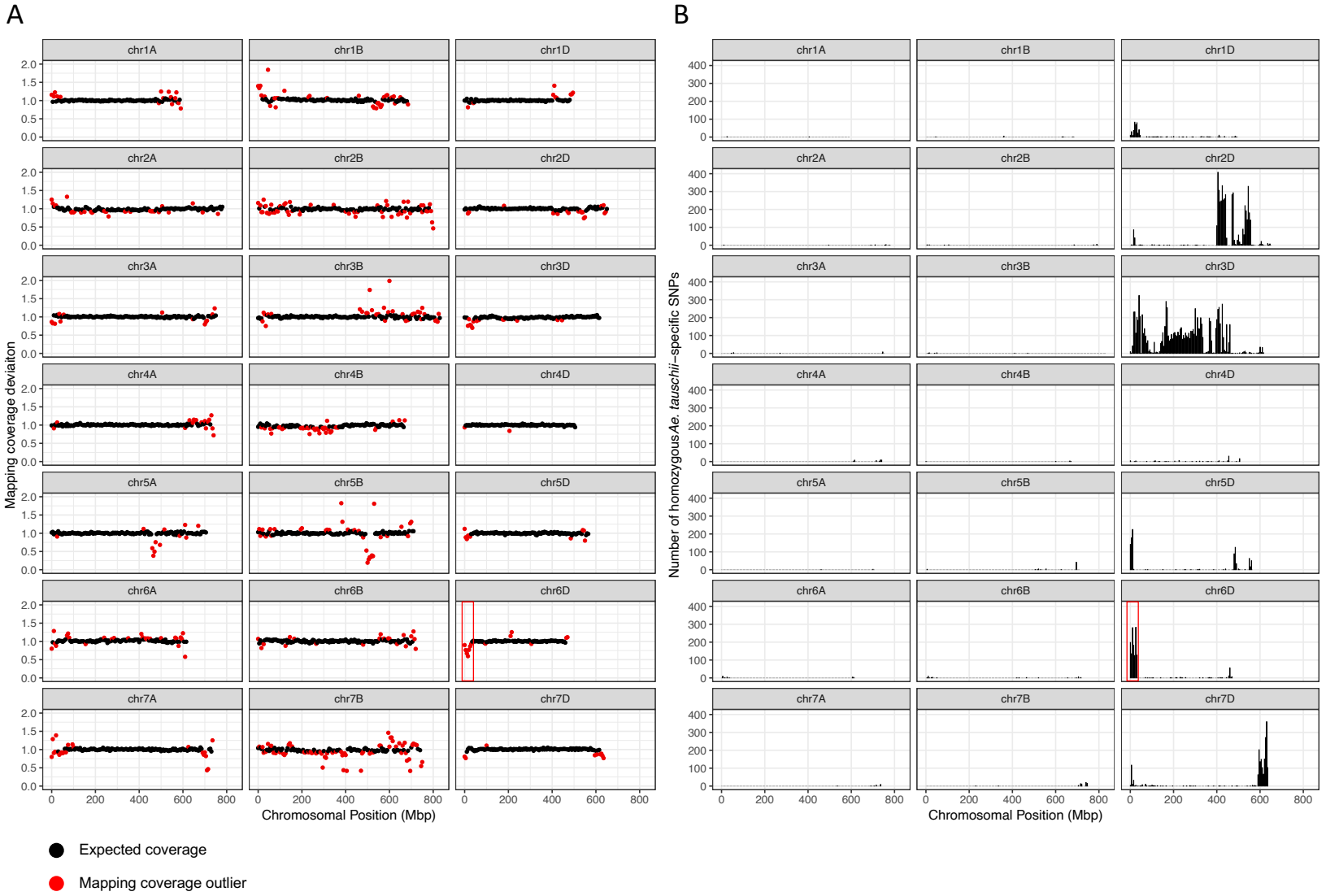

A
B
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D
Mapping coverage deviaiton
Number of homozygous Ae. tauschii–specific SNPs
Chromosomal Position (Mbp)
Expected coverage
Mapping coverage outlier

Figure 3-10. Ae. tauschii introgressions in Sokoll (HiBAP_57).
A) Mapping coverage between Sokoll and the median of the panel in 5 Mbp genomic windows. Red points are statistically significant outliers (n = 149 biologically independent lines). The red box indicates the *Ae. tauschii* introgression overlapping the chr6D MTA **B)** The number of homozygous *Ae. tauschii*-specific SNPs in each 5 Mbp genomic window. The red box indicates the *Ae. tauschii* introgression overlapping the chr6D MTA.

To examine the *Ae. tauschii* introgression in more detail, I used 1 Mbp genomic windows and looked specifically at the first 50 Mbp of chr6D (Fig. 3-11). In Sokoll (HiBAP_57), the segment is 31.6 Mbp in length. As *Ae. tauschii* is from wheat's primary genome and is thus more similar to the D subgenome than more distant wheat relatives are, not every 1 Mbp window is sufficiently lacking in synteny for reads to map poorly and produce significant coverage deviation below 1. This explains why some windows within the introgression have a coverage deviation of around 1. However, these windows still have *Ae. tauschii*-specific SNPs and are within a block of 1 Mbp windows in which most have significant coverage deviation below 1. Therefore, we can be confident that the introgression includes these windows. The fact that this is one contiguous introgression can also be seen in Fig. 3-10.

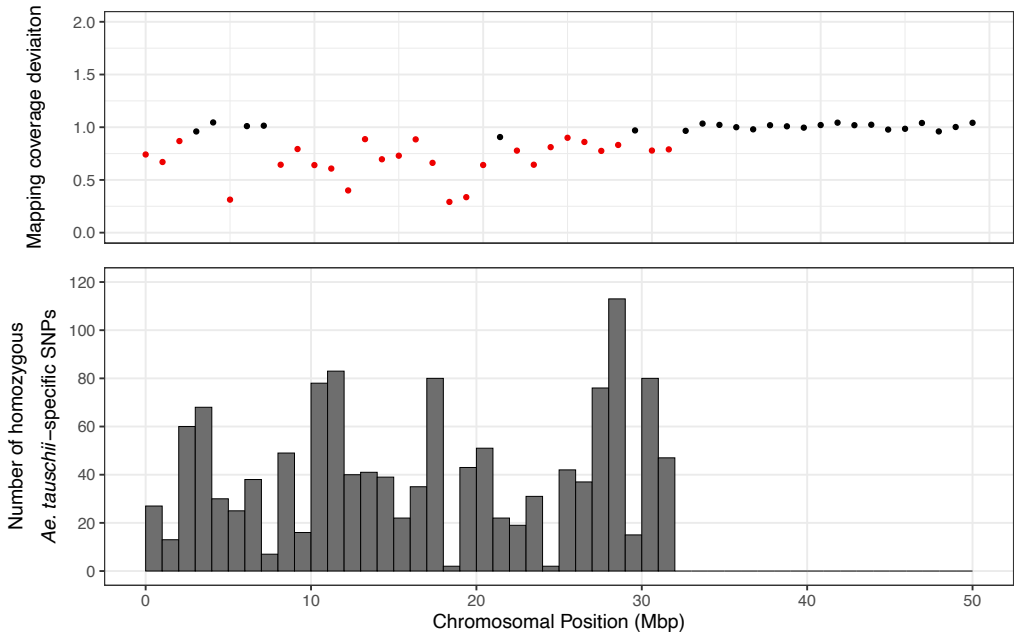


Figure 3-11. *Ae. tauschii* introgressions at the start of chr6D in Sokoll (HiBAP_57). The top panel shows deviation in mapping coverage between Sokoll and the median of the panel in 1 Mbp genomic windows. Red points are statistically significant outliers (*n* = 149 biologically independent lines). The bottom panel shows the number of homozygous *Ae. tauschii*-specific SNPs in each 1 Mbp genomic window.

Looking at the *Ae.* tauschii introgression in other HiBAP lines revealed that the segment is variable in length, indicating it has recombined since its introduction. The longest segment is 31.6 Mbp in length, as seen in Sokoll (HiBAP_57). Various reduced sizes of the segment are also found in some of the lines without the favourable allele. To determine which specific region of the segment was responsible for the phenotype, I compared the segment size and position across the lines. I found that all 40 lines with the favourable T allele in a homozygous or heterozygous state have *Ae. tauschii* between 5.05 Mbp and 6.85 Mbp on chr6D (based on Chinese Spring coordinates). Furthermore, this region is not introgressed in the lines homozygous for the unfavourable A allele. This can be seen in Fig. 3-12, in which all six HiBAP lines plotted possess *Ae. tauschii* at the start of chr6D but only the four with the T allele (HiBAP 57, 29, 48, and 65) have *Ae. tauschii* between 5.05 Mbp and 6.85 Mbp and have high yield under heat stress (369.67g m^{-2}, 438.18g m^{-2}, 451.43g m^{-2}, 459.15g m^{-2}), while the 2 lines with the A allele (HiBAP 92 and 103) lack this region and have low yield under heat stress (185.72g m^{-2}, 213.34g m^{-2}).

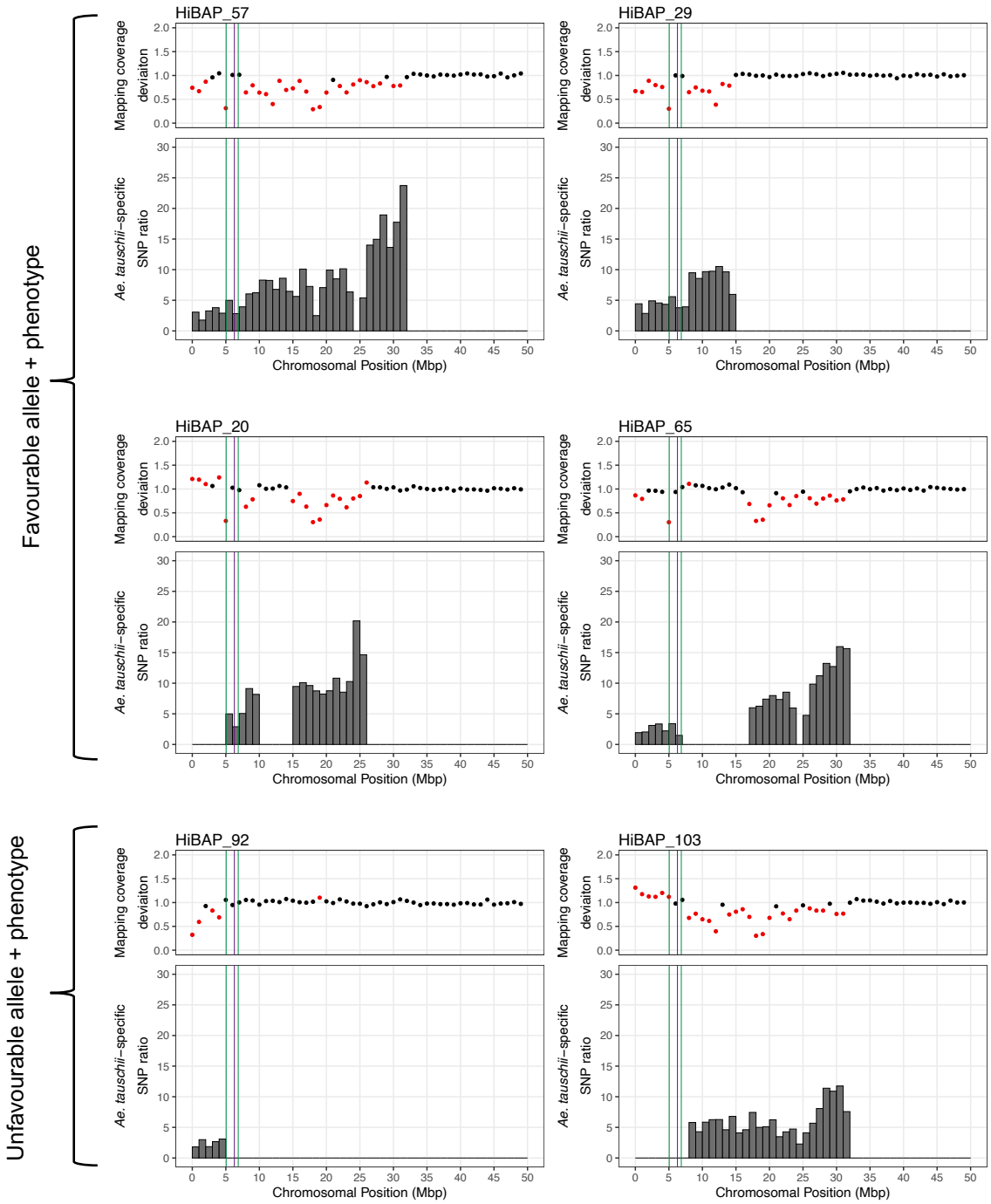

Favourable allele + phenotype
Unfavourable allele + phenotype
HiBAP_57
HiBAP_29
HiBAP_20
HiBAP_65
HiBAP_92
HiBAP_103
Mapping coverage deviaiton
Ae. tauschii–specific SNP ratio
Chromosomal Position (Mbp)
Expected coverage
Mapping coverage outlier
Core introgressed region
6D-6276646 MTA position

Figure 3-12. Visualising *Ae. tauschii* introgressions across the first 50 Mbp of chr6D in six HiBAP I lines.
Four of the HiBAP lines contain the favourable T allele at chr6D-6276646 (HiBAP 57, 29, 48, and 65) and two contain the unfavourable A allele at chr6D-6276646 (HiBAP 92 and 103). Mapping coverage deviation was computed between the HiBAP line and the median of the panel in 1 Mbp windows. Red points are statistically significant outliers (*n* = 149 biologically independent lines). *Ae. tauschii*-specific SNP ratio in each 1 Mbp window was calculated by dividing the number of homozygous *Ae. tauschii*-specific SNPs in that window by mean number of homozygous *Ae. tauschii*-specific SNPs in that window across the panel. Green lines mark the borders of the region common to all lines with the favourable T allele. The purple line indicates the MTA position.

To further explore the propensity for the introgressed segment to recombine, I looked at the segment in five lines whose parents were Sokoll and Weebil1 (Fig. 3-13). As Sokoll has the full-length, 31.6 Mbp segment and Weebil1 lacks the segment, variation in the segment size in the offspring indicates recombination has taken place within the segment. Recombination within the segment took place in all four lines whose parents were Sokoll and Weebil1, suggesting that the *Ae. tauschii* introgression readily recombines in a wheat background. The segment in all four lines looks different so there doesn't appear to be clear common recombination sites, although the number of samples is too low to properly assess this.

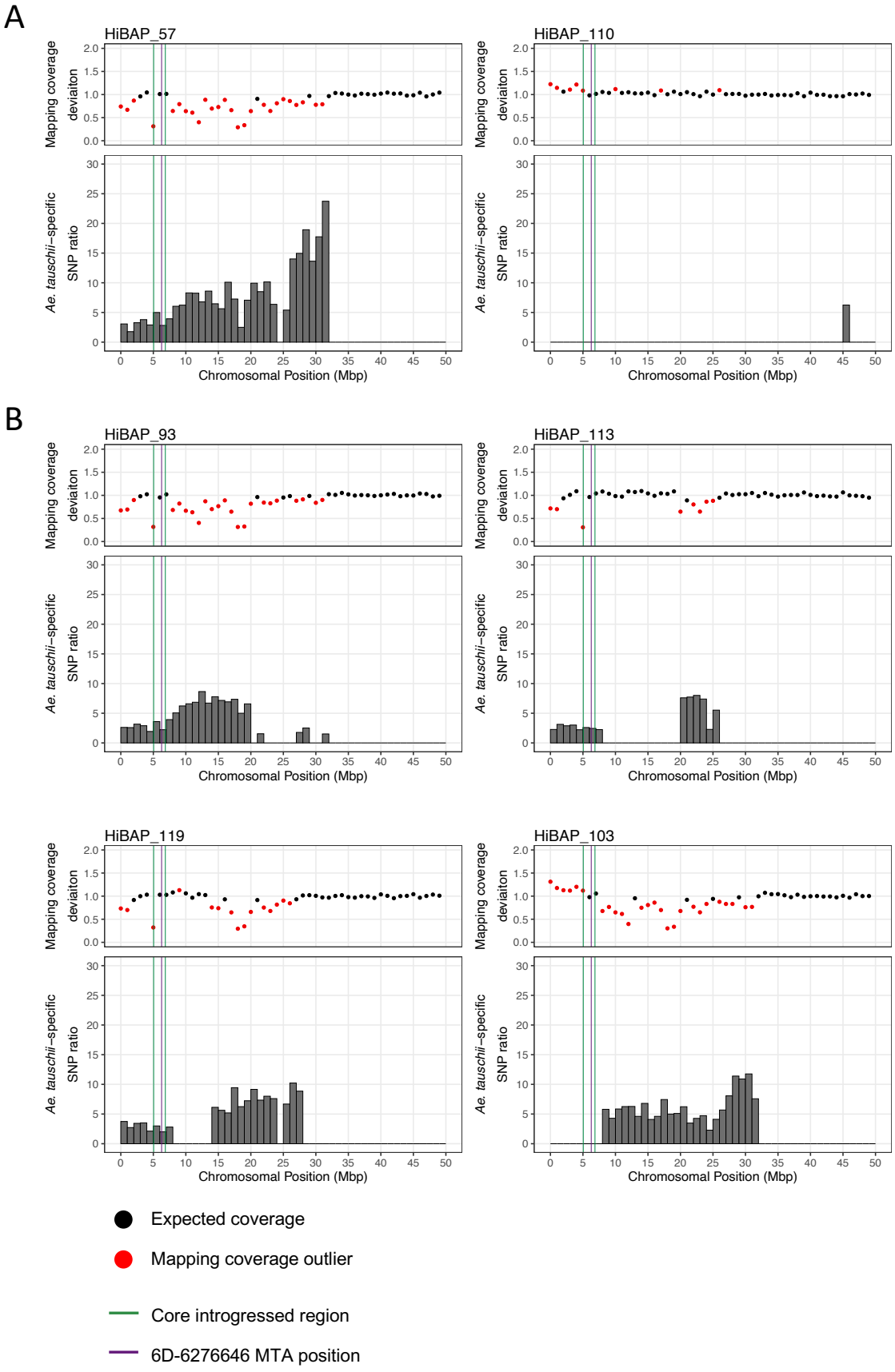

A
HiBAP_57
HiBAP_110
B
HiBAP_93
HiBAP_113
HiBAP_119
HiBAP_103
Mapping coverage deviaiton
Ae. tauschii–specific SNP ratio
Chromosomal Position (Mbp)
Expected coverage
Mapping coverage outlier
Core introgressed region
6D-6276646 MTA position

Figure 3-13. *Ae. tauschii* introgressions within 6D:0-50 Mbp in A) Sokoll (HiBAP_57) and Weebil1 (HiBAP_110) and B) four lines whose parents were Sokoll and Weebil1. Mapping coverage deviation was computed between the HiBAP line and the median of the panel in 1 Mbp windows. Red points are statistically significant outliers. *Ae. tauschii*-specific SNP ratio in each 1 Mbp window was calculated by dividing the number of homozygous *Am. muticum*-specific SNPs in that window by mean number of homozygous SNPs in that window across the panel. Green lines mark the borders of the region common to all lines with the favourable T allele in a homozygous or heterozygous state. The purple line indicates the 6D MTA position.

3.3.6 Anchoring the core introgressed region to the *Ae.* tauschii reference genome

So far, we have identified the core introgressed region relative to the Chinese Spring wheat reference genome. However, due to differences in synteny and variation in gene content between Chinese Spring and *Ae. tauschii*, identifying the corresponding region in *Ae. tauschii* will be more appropriate for candidate gene searches. To extract the corresponding region in *Ae. tauschii*, I aligned the proteins and chromosome sequence of chr6D in Chinese Spring with the *Ae. tauschii* reference genome. I then located where the borders and flanking sequence and proteins from the 1.80 Mbp region corresponded on the *Ae. tauschii* reference genome. The corresponding region is a 1.49 Mbp region on chr6 of the *Ae. tauschii* reference genome between 4.63 Mbp and 6.12 Mbp (Fig. 3-14B). This represents the probable introgressed chromosome segment, within which the gene(s) underlying the MTA is likely found.

As it is unknown which *Ae. tauschii* accession is introgressed into these lines, when looking for candidate genes, I used multiple *Ae. tauschii* genomes in case there were genes unique to some of the accessions. I extracted the corresponding region from chromosome-level genome assemblies of four other accessions (Zhou *et al.*, 2021) using alignments between the introgressed region from the *Ae. tauschii* reference genome and the other *Ae. tauschii* genomes.

Using the chromosome alignments, I also explored the synteny of the core introgressed region between Chinese Spring and *Ae. tauschii* (Fig. 3-14A). The first half of the core introgressed region lacks synteny between Chinese Spring and *Ae. tauschii*, with most of the DNA absent in *Ae. tauschii*. This corresponds with the poor mapping coverage (low coverage deviation value) at the start of the core introgressed region in Fig. 3-12. Synteny resumes in the second half of the core introgressed region, in line with better mapping coverage (higher coverage deviation value – close to 1) in Fig. 3-12. There is an inversion

of around 1 Mbp immediately outside of the core introgressed region. As recombination is suppressed within inversions (Li *et al.*, 2023), this may act as a natural boundary to the core region.

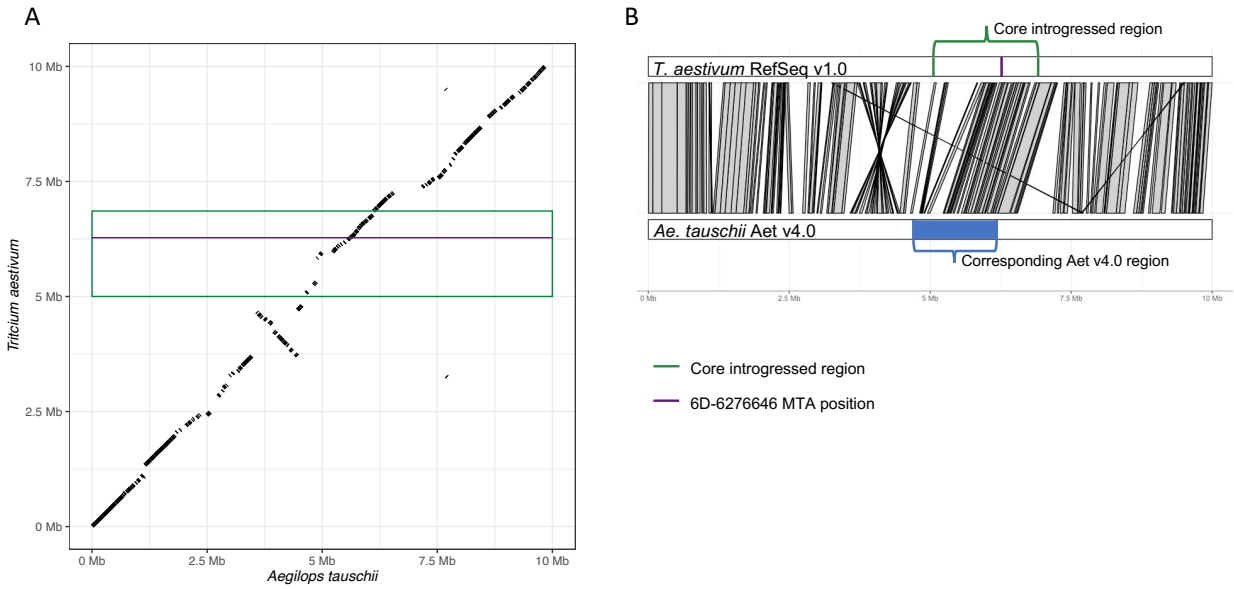


Figure 3-14. Alignments between *Ae. tauschii* and the wheat reference genome at the start of chr6/chr6D.
A) Synteny between 6D:1-10,000,000 in Chinese Spring RefSeq v1.0 (Appels *et al.*, 2018) and *Ae. tauschii* Aet v4.0 (Luo *et al.*, 2017). The green box indicates the 1.80 Mbp region (1.49 Mbp relative to *Ae. tauschii*) common to all lines with the favourable T allele, corresponding to the green region in Fig. 3-12. The purple line indicates the MTA position. **B)** Alignment of 6D:1-10,000,000 in Chinese Spring RefSeq v1.0 (Appels *et al.*, 2018) and 6D:1-10,000,000 in *Ae. tauschii* Aet v4.0 (Luo *et al.*, 2017), illustrating how the syntenic region in *Ae. tauschii* was identified and extracted.

3.3.7 Identifying candidate genes

The core introgressed region varied slightly between the five *Ae. tauschii* accessions; its length ranged from 1.49 Mbp to 1.82 Mbp and it contained between 26 and 33 genes (Appendix B1). Three genes were identified as potential candidates based on comprehensive literature searches of the functionality of related genes. One of these is a MIKC-type MADS-box transcription factor gene, AET6Gv20025600, related to *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*), which, when overexpressed in *Arabidopsis*, leads to chloroplast biogenesis, elevated photosynthesis, and heat tolerance (Ning *et al.*, 2021). Although the *Ae. tauschii* and wheat orthologues have identical protein sequences, regulatory differences in this gene could cause a phenotype similar to the overexpression phenotype seen in *Arabidopsis*.

I also found a novel mitogen-activated protein kinase (MAPK), AetT093_6Dv1G029500/AetAY17_6Dv1G019800, found in two of the *Ae. tauschii* accessions, AY17 and T093, and not found in wheat. MAPKs have been connected to the heat stress response in *Arabidopsis*, rice and maize (Mo *et al.*, 2021) and more specifically to oxidative stress tolerance in wheat exposed to prolonged heat stress treatment (Kumar *et al.*, 2021). This gene isn't found in the wheat cultivars from Walkowiak et al. (2020) and is only present in some of the *Ae. tauschii* accessions. This position in the accessory genome of *Ae.* tauschii suggests that this gene evolved fairly recently and is possibly involved in environmental adaptation.

Finally, I identified two type-B two-component response regulator receivers, AET6Gv20025700 and AET6Gv20025800, which are related to *Arabidopsis* response regulators (ARRs). Type-B ARRs are transcription factors that regulate cytokinin signalling pathways and influence response to abiotic stress (Argyros *et al.*, 2008; Nguyen *et al.*, 2016; Frank *et al.*, 2020). Loss of function mutants exhibit altered root structure, increased light sensitivity, and altered concentrations of chlorophyll and anthocyanins (Argyros *et al.*, 2008). Heat stress negatively impacts photosynthetic capacity through the reduction of chlorophyll content and photochemical efficiency (Liu *et al.*, 2020). It is thought that increased cytokinin, particularly in the leaves, plays a protective role against heat stress by clearing reactive oxygen species (ROS) and upregulating heat shock response proteins, allowing the plant to maintain normal growth under heat stress.

The interval surrounding the MTA on chr2B, contains the gene *DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN 1A* (*DREB1A*). *DREB1A* is in a family of transcription factors that bind DRE/CRT elements during the abiotic stress response. In wheat, overexpression of *DREB1* has been linked to increased photosynthetic efficiency and drought tolerance (Y. Zhou *et al.*, 2020) and lines overexpressing *DREB2* were more tolerant to heat and cold stress (Lee et al., 2010). Finally, *STEROL GLUCOSYLTRANSFERASE* (*SGT*) is present within the interval surrounding the MTA on chr1B. Knockout and overexpression studies in *Arabidopsis* suggest that SGT is involved in stability in response to heat (Mishra *et al.*, 2013; Misra *et al.*, 2016).

After identifying candidate genes, I looked whether there were any obvious differences between the *Ae. tauschii* and wheat orthologues that might underlie a functional difference. Both ARRs identified have syntenic wheat orthologues, but one of them,

AET6Gv20025700, has a myb-binding domain in the annotated gene model which is absent from its wheat orthologue TraesCS6D02G014900. If real, this difference could underlie a large functional difference in lines with AET6Gv20025700 introgressed. However, through manual reannotation to validate this, I found that the difference in domain was due to misannotation as in both *Ae. tauschii* and Chinese Spring. The myb-binding domain is present and expressed at the mRNA level. To check whether this leads to an intact translated protein or if it contains a premature stop codon, I assembled transcripts from the mapped mRNA reads and translated the coding regions. The Chinese Spring gene has an uninterrupted open reading frame, suggesting that this gene should be functional. A difference in this myb domain between lines is therefore unlikely to be involved in the heat tolerance trait.

3.4 Discussion

3.4.1 Physiological analysis reveals insights into heat tolerance

As expected, the heat stress imposed on the plants had a major negative effect on most of the measured traits, a finding supported by previous studies (Thapa *et al.*, 2020; Shenoda *et al.*, 2021). However, exotic-derived lines had significantly higher grain yield than Elite lines under heat stress with no penalty under yield potential conditions. Grain yield is considered a reliable criterion for assessing heat stress in wheat (Shenoda *et al.*, 2021) so its use here for that purpose is valid. Exotic-derived lines also had a significantly lower stress susceptibility index than Elite lines, adding additional support to the resilience of exotic-derived lines under heat stress. The better performance of exotic-derived lines is unsurprising as exotic parents have been routinely used to enhance genetic diversity in wheat pre-breeding programmes and previous research has demonstrated that their inclusion leads to enhanced performance under heat stress (Cossani and Reynolds, 2015; Pinto, Molero and Reynolds, 2017), drought (Reynolds, Dreccer and Trethowan, 2007; Lopes and Reynolds, 2011), and salinity (Colmer, Flowers and Munns, 2006).

The GWAS revealed three markers, predominantly found in exotic-derived lines that are associated with a range of traits under heat stress, including higher grain number, higher biomass throughout the crop cycle, reduced stress susceptibility index and cooler canopy temperature during both the vegetative and the grain-filling phenological stages. These alleles are likely responsible for the better performance of exotic-derived lines reported

in section 3.3.1. This is further supported by the more distinct difference in yield between the different allele combinations than between the exotic-derived lines and Elite lines, seen by the tighter distribution in the boxplot and the reduced overlap between groups. This is probably caused by not all the exotic-derived lines containing the favourable alleles at the MTAs. For example, only 40/66 (60.6%) of exotic-derived lines possess the favourable T allele in a homozygous or heterozygous state at the chr6D MTA; however, the remaining 26 exotic-derived lines contributed to the comparison between exotic-derived and Elite lines.

In this study, there was a negative correlation between canopy temperature and yield under heat stress conditions, which was stronger in exotic-derived lines than in Elite lines. Furthermore, the three MTAs were associated with canopy temperature, with the canopy of lines possessing the favourable alleles at the three MTAs being around 2°C cooler than that of lines lacking these alleles. Reduced canopy temperature has been previously connected to a higher tolerance to heat stress and drought (Pinto *et al.*, 2010), and variation in canopy temperature under heat stress has been previously identified at CIMMYT (Pinto *et al.*, 2010). Canopy temperature has also been found to be correlated with yield in warm environments under different levels of water availability (Mohammadi et al., 2012). Cooler canopies have also been associated with better optimised root distribution in wheat (Pinto and Reynolds, 2015). This may be because roots capable of supplying sufficient water to cool aerial structures enable increased transpiration and a maintenance of a cooler canopy (Amani, Fischer and Reynolds, 1996; Trethowan and Mahmood, 2011). Higher transpiration rates have also been associated with increased stomatal conductance which could drive increased photosynthesis and the higher biomass seen in the exotic-derived lines.

NDVI is another trait, like canopy temperature, that is easily phenotyped and can act as a proxy for heat/drought tolerance (Pinto *et al.*, 2010). It encompasses ground cover and canopy nitrogen content and can thus be used as an indirect estimate of the health of leaves involved in photosynthesis as healthy green leaves are indicative of stress tolerance for drought and heat conditions (Tao *et al.*, 2000). Plants with a staygreen phenotype have delayed senescence so can maintain green leaves and undergo photosynthesis for longer post anthesis, which improves grain filling (Kamal *et al.*, 2019) and yield stability (Kumar *et al.*, 2022), even in heat stress or drought conditions.

In this study, NDVI was positively correlated with yield under heat stress conditions with this correlation being stronger in exotic-derived lines than in Elite lines. Furthermore, the three MTAs were associated with NDVI, providing support for the idea that, in addition to keeping a cool canopy, the maintenance of a healthy canopy and green leaves is involved in the heat tolerance observed here.

3.4.2 Validity of using delayed sowing to induce heat stress

In this study, heat stress was imposed on the plants by delaying sowing by three months, with emergence registered in March rather than December. Delaying sowing influences not only temperature but also photoperiod and could therefore introduce confounding effects into the study. However, for several reasons, this is unlikely to have had a major effect on the results. Firstly, the lines used in this study are relatively insensitive to photoperiod, having been selected as such using CIMMYT's shuttle breeding technique. Using marker analysis, this insensitivity to photoperiod was confirmed, as the spring allele at *Vrn-B1* (*Vrn-B1A*) and *Vrn-D1* (*Vrn-D1a*) was present in around 90% of the HIBAP I panel (Dreisigacker *et al.*, 2021). Secondly, several other studies have used delayed sowing to effectively evaluate heat tolerance, both at CIMMYT's Obregon station (Reynolds *et al.*, 1994, 2016; Lillemo *et al.*, 2005; Mondal *et al.*, 2013; Cossani and Reynolds, 2015) and outside of CIMMYT (Shenoda *et al.*, 2021; Sinha and Kumar, 2022).

3.4.3 Marker-trait associations uncovered through genome-wide association study

Combining genotyping data generated through *de novo* SNP discovery and phenotyping data from the field in a genome-wide association study revealed three pleiotropic MTAs. When the favourable alleles at each MTA exist in the same line, they confer a 56.5% increase to yield and a 1.97 °C/2.37 °C reduction in canopy temperature under heat stress conditions compared to lines without the three favourable alleles. These markers are associated with multiple traits of agronomic importance, all indicating tolerance to heat stress, including yield, biomass, grain per square metre and grain filling rate.

Interestingly, although these markers are seemingly found in different genomic locations, the favourable alleles at the chr1B and chr2B MTAs always co-occur in the same line, and these lines also typically possess the favourable allele at the chr6D MTA, although not in every case. To explain the co-occurrence of the chr1B and chr2B favourable alleles, it is possible that the chr1B and chr2B favourable alleles are in close physical proximity on the

same chromosome if there is structural variation between Chinese Spring and the HiBAP lines that has caused reads to mismap and these alleles to be incorrectly assigned to different chromosomes. Alternatively, these alleles could be interacting epistatically and have been selected for a combined function. The latter could also explain the chr6D favourable allele typically co-occurring with the chr1B and chr2B favourable alleles.

Lines containing the favourable alleles are almost all exotic-derived lines. However, they are not found in any specific subcategory; they are found in synthetic-derived lines, introgressions lines, and landrace-derived lines. This brings the origin of the alleles into question and challenges the assumption that the chr6D allele originates solely from CIMMYT's synthetic breeding programme. However, it is possible that a synthetic-derived line is in the pedigree history of lines not explicitly labelled as synthetic-derived. Due to the challenges of accurately recording pedigree histories in complex pre-breeding programmes such as those conducted at CIMMYT, this seems very possible. For the chr6D allele, one contender for donor is Sokoll, which is commonly used as a synthetic-derived parent line and contains the chr6D *Ae. tauschii* segment in its longest observed length.

The MTAs identified here do not overlap with MTAs identified in previous association studies that used HIBAP I to identify associations with biomass (Molero *et al.*, 2019) or photosynthetic efficiency traits (Joynson *et al.*, 2021). When writing up this research, we became aware that the chr6D MTA is supported by a heat tolerance MTA nearby on chr6D in (Singh *et al.*, 2018; S. Singh *et al.*, 2021). Our work adds value to this by offering independent evidence in a different experiment, in addition to phenotyping more traits related to heat tolerance, showing the lack of downside of the allele under yield potential conditions, and uncovering the connection between canopy temperature and yield and between NDVI and yield. Furthermore, this work also revealed the other two MTAs, on chr1B and chr2B, which to our knowledge had not been previously reported.

3.4.4 *Ae. tauschii* introgression underlying the marker-trait association on chr6D

Using coverage deviation and SNP information from the HiBAP I lines, I identified that the chr6D MTA overlaps with an *Ae. tauschii* introgression that readily recombines within CIMMYT germplasm. This is promising for the successful deployment of introgressions from wheat's primary genepool into breeding programmes and may reduce concerns regarding linkage drag and poor recombination of such introgressions. The recombination allowed me to identify a small core region of *Ae. tauschii* only present in lines with the

favourable allele and the heat tolerant phenotype. The smallest segment in the panel that contains the favourable allele at the ch6D MTA is around 5 Mbp in length and due to the ease with which this region can recombine its size can likely be broken down further. Its small size, recombination potential, and telomeric location make it amenable for use in breeding programmes.

Singh et al. (2018) also independently suspected that *Ae. tauschii* was underlying this MTA. However, this relied upon pedigree-based inference and markers that look speculative. Here, I confirmed this speculation, providing additional experimental evidence to support the hypothesis. Together, the two independent studies provide strong evidence that *Ae. tauschii* is underlying this MTA.

3.4.5 The limitations of relying on the reference genome for candidate gene discovery

When searching for candidate genes within the chr6D MTA interval, instead of searching within the Chinese Spring wheat reference genome, I instead searched for genes within five *Ae. tauschii* genomes in the regions corresponding to the core introgressed region identified relative to Chinese Spring. While the search for candidate genes within the reference genome of the species being studied is generally deemed appropriate, in this case, as an *Ae. tauschii* introgression is underlying the MTA, it is not unlikely that novel genes or changes to gene order could result in the causal gene(s) not being present in the interval being searched in the reference genome.

The search resulted in finding a novel *Ae. tauschii* gene and the absence of the isoflavone reductase gene proposed by Singh et al. (2018) as the primary candidate gene which, due to gene order differences between Chinese Spring and *Ae. tauschii*, is not present within the core introgressed region, instead being around 3 Mbp upstream. Without discovering that an introgression was underlying this MTA and using the corresponding region within *Ae. tauschii* reference genomes as the source for candidate genes, the novel gene would have been missed and the isoflavone reductase gene may have been chosen as a candidate. This highlights the benefit of considering non-reference genomes, especially when introgressions from other species are involved.

Candidate genes are inherently speculative and the genes underlying the heat tolerance phenotype might not be any of those suggested here as candidates for several possible reasons. Firstly, the causal genes may not be present in the genomes surveyed. Secondly,

the causal genes could be missed during the literature search due to the phenotype being driven by an unknown or unexpected pathway, or there being a lack of functional information available about the genes. Finally, the MTA could be within a regulatory element that acts on a gene that is driving the phenotype but is outside the MTA interval, making it much more challenging to pinpoint the causal gene. Therefore, while candidate genes provide a starting point for follow-up work, they should be treated with caution and alternative genes should not be overlooked.

3.4.6 Future work

The identification of the three alleles presents an immediate opportunity for breeding by incorporating the heat resilience phenotype into Elite wheat cultivars. Currently, CIMMYT is designing KASP™ assays for these markers, and European breeding companies have expressed interest in using them. Initially, these KASP™ assays can be used to survey existing breeding germplasm and nursery collections to assess the presence of the favourable alleles. If these alleles are already present in Elite breeding material, their identification will accelerate the breeding process. Alternatively, if these alleles are not yet part of Elite breeding material, they can be incorporated through marker-assisted breeding programmes. If using marker-assisted breeding to incorporate these alleles, the selection of donor lines is important. They could be selected using the introgression mapping approach shown here to select lines that possess the least amount of divergent content while still containing the favourable alleles. This approach aims to minimise the introduction of linkage drag and streamline the crossing process.

Once the favourable alleles have been incorporated into Elite cultivars, their performance across diverse environments should be evaluated. This is a crucial step to ensure the heat tolerance phenotype is not exclusive to the Mexican climate or the specific field conditions of the study. For example, as this experiment was conducted in fully irrigated conditions, the effect of drought on the heat resilient phenotype is unknown.

Beyond immediate practical applications, further characterising the three MTAs will be of academic interest. An important part of this will be generating a chromosome-level genome assembly from a HiBAP line containing the favourable allele at all 3 MTAs, a task that is no longer prohibitively costly, using, for example, PACBIO HIFI sequencing. This new reference genome will provide us with the exact gene content of the heat tolerance lines and serve as a foundational resource for forward genetic screens to determine the

genes underlying the phenotype and for differential expression analyses comparing plants under heat stress and yield potential conditions to understand the transcriptional response to heat stress. It may also allow us to determine the cause of the common co-occurrence of favourable alleles. Due to the existence of introgressed material underlying the MTA of largest effect size, producing a genome assembly is particularly important, as gene content variation can obscure the causal genes and inaccurate read mapping, caused by reads mapping poorly to divergent introgressed gene models, can compromise the gene expression analysis.

3.5 Methods

3.5.1 Plant material and growth conditions

The High Biomass Association Mapping Panel HiBAP I consists of 149 spring wheat lines and is composed of elite high yielding lines and lines with exotic material in their pedigree history derived from CIMMYT breeding and pre-breeding programs (Molero *et al.*, 2019). These exotic lines include primary synthetic derivative lines, containing between 0.5% and 43% donor material (Joynson *et al.*, 2021); Mexican and other origin landrace derivative lines; and Elite lines containing an introgressed segment of *Th. ponticum* on chr7D and/or *S. cereale* on chr1B (Joynson *et al.*, 2021). The set of Elite lines contains 11 CIMMYT varieties released from 1966 until 2007 and additional lines selected during systematic screening of CIMMYT breeding and pre-breeding material under yield potential and heat stress field conditions. This allowed the identification of Elite genotypes with favourable expression of traits of interest such as high biomass/radiation use efficiency at different growth stages including final above ground biomass under both yield potential and heat stress conditions.

To construct the final panel, a pre-panel consisting of more than 250 lines from different sources were evaluated in the field under favourable conditions; lines with a favourable agronomic background and without extreme height or phenology under yield potential conditions were selected to reduce the confounding effect of extreme phenology or height on the expression of biomass and other traits. HiBAP I was evaluated during 2015/16 and 2016/17 under yield potential (YP16 and YP17) and heat stress conditions (Ht16 and Ht17). Heat stressed conditions were created with delayed sowing where emergence was registered in March instead of November or December as in a normal growing cycle.

The field experiments were carried out at IWYP-Hub (International Wheat Yield Partnership Phenotyping Platform) situated at CIMMYT's Experimental Station in Campo experimental Norman E. Borlaug (CENEB) in the Yaqui Valley, near Ciudad Obregon, Sonora, Mexico (27°24' N, 109°56' W, 38 masl) under fully irrigated conditions for both yield potential and heat stress experiments. The soil type at the experimental station is a coarse sandy clay, mixed montmorillonitic typic caliciorthid. It is low in organic matter and is slightly alkaline (pH 7.7) (Sayre, Rajaram and Fischer, 1997). Experimental design for all environments was an alpha-lattice. Yield potential experiments consisted of four replicates in raised beds (2 beds per plot, each 0.8 m wide and 4m long) with four (YP16) and two (YP17) rows per bed (0.1 m and 0.24 m between rows respectively) and 4 m long. For heat stress experiments, two replicates were evaluated for HiBAP I in 2 m × 0.8 m plots with three rows per bed. Seeding rates were 102 Kg ha^{-1} and 94 Kg ha^{-1} for YP and Ht experiments, respectively. Appropriate weed disease and pest control were implemented to avoid yield limitations. Plots were fertilised with 50 kg N ha^{-1} (urea) and 50 kg P ha^{-1} at soil preparation, 50 kg N ha^{-1} with the first irrigation and another 150 kg N ha^{-1} with the second irrigation.

3.5.2 Agronomic measurements

Phenology of the plots was recorded during the cycle using the Zadoks growth scale (GS) (Zadoks, Chang and Konzak, 1974), following the average phenology of the plot (when 50% of the shoots reached a certain developmental stage). The phenological stages recorded were heading for heat experiments (GS55, DTH), anthesis for yield potential experiments (GS65, DTA) and physiological maturity (GS87, DTM) for both experiments. Percentage of grain filling was calculated as the number of days between anthesis and physiological maturity divided by DTM.

Plant height was measured as the length of five individual shoots per plot from the soil surface to the tip of the spike excluding the awns. Spike, awn, and peduncle length were measured in five shoots per plot before physiological maturity (PM). Fertile ($SPKLSP^{-1}$) and infertile spikelets per spike ($InfSPKLSP^{-1}$) were also counted in five spikes per plot at PM.

At physiological maturity, grain yield and yield components were determined using standard protocols (Pask *et al.*, 2012). The samples of fertile shoots were oven-dried, weighed and threshed to allow calculation of harvest index, biomass at physiological maturity, spikes per square meter, grains per square meter, number of grains per spike

and grain weight per spike. Grain yield was determined on a minimum of 3.2 m^2 to a maximum of 4.8 m^2 under yield potential experiments and 1.6 m^2 under heat experiments. In yield potential experiments only, to avoid edge effects arising from border plants receiving more solar radiation, 50 cm of the plot edges were discarded before harvesting. From the harvest of each plot, a subsample of grains was weighed before and after drying (oven-dried to constant weight at 70 °C for 48 h) and the ratio of dry to fresh weight was used to determine dry grain yield and thousand grain weight. Grain number was calculated as (Yield/TGW) × 1000. Biomass at physiological maturity was calculated as yield/HI. Number of spikes per m^2 was calculated as biomass at physiological maturity / (shoot dry weight/shoot number).

3.5.3 Unmanned Aerial Vehicle (UAV) for canopy temperature and NDVI estimation

Aerial measurements for canopy temperature and NDVI were collected using different aerial platforms. Each year, logistics and availability determined which platform could be used. The multispectral and thermal cameras were calibrated onsite by measuring calibration panels placed on the ground before and after each mission. An exception to this was the aircraft missions, where a calibration performed at the airfield would not have been representative of the trial conditions. The flights were designed as a regular grid of north-south flightpaths covering the whole trial with images that overlapped 75% in all directions to ensure a good reconstruction of the orthomosaic. The flights were performed under clear sky conditions at solar noon ±2 h.

NDVI and canopy temperature orthomosaics were obtained from the aerial images using the software Pix4D. The orthomosaics were then exported to ArcGIS where a grid of polygons was adjusted on top of the image. To avoid the border effect, the polygons were buffered 0.5 m from the north and south border of the plot. Finally, the pixel values were extracted using the 'raster' package in R. The value of all the pixels enclosed within each polygon was extracted, possible outliers were removed and the average per plot was calculated.

3.5.4 Calculating stress tolerance indices

To determine the effect of heat stress evaluated across years and panels, the stress susceptibility index (SSI) was calculated for each HiBAP I line using data from yield potential (Yyp) and heat stress (Yht) experiments as follows:

$$SSI = \frac{1 - \frac{Yht}{Yyp}}{1 - \frac{\bar{Y}ht}{\bar{Y}yp}}$$

where Yht and Yyp are the yield of the HiBAP I line evaluated under heat stress and yield potential conditions, respectively, while $\bar{Y}$ht and $\bar{Y}$yp are the mean yield of lines from HiBAP I evaluated under heat stress and yield potential conditions, respectively (Fischer and Maurer, 1978).

3.5.5 DNA extraction, capture enrichment, and genotyping

All genotyping data was taken from Joynson *et al*. (2021). Flag leaf material from 10 plants per line was collected from field grown plots post anthesis and pooled prior to extraction with a CTAB-based protocol. DNA was extracted using a standard Qiagen DNEasy extraction preparation and quality and quantity assessed using a NanoDrop 2000 (Thermofisher Scientific) and the Quant-iTTM assay kit (Life Technologies). From this DNA, dual indexed Trueseq libraries with an average insert size of 450 bp were produced for each line and enriched using a custom MyBaits 12 Mbp (100,000 120 bp RNA probes) enrichment capture synthesised by Arbour Bioscience and using 8x pre-capture multiplexing. 90,000 of these probes were designed using an island strategy to target regions across the whole genome. A subgenome-collapsed reference was used to design these probe sequences to enable homoeologous regions to be targeted with a single probe. 10,000 of the probes were designed for selected genes, targeting both the gene body and 2 Kbp upstream. Post enrichment libraries were sequenced by Genomics Pipelines at the Earlham Institute using an S4 flowcell on an Illumina NovaSeq6000 producing 150 bp paired end reads.

To process the sequencing reads, first they were trimmed and low-quality reads were removed. These reads were mapped to the Chinese Spring RefSeq v1.0 wheat reference genome (Appels *et al.*, 2018) using BWA mem v0.7.13 (Li, 2013). Samtools v1.4 (Li and Durbin, 2009) was used to remove unmapped reads, supplementary alignments, improperly paired reads, and reads that didn't map uniquely (mapping quality < 10). PCR duplicates were removed using Picard's MarkDuplicates (Depristo *et al.*, 2011). SNPs were called using samtools mpileup and bcftools call (Li, 2011) with parameter -m. SNPs were filtered using GATK (Depristo *et al.*, 2011) to remove SNPs that were heterozygous, had a quality score <30 or a depth <5. A locus was designated as homozygous reference if no

alternative allele was found but 5 or more reads were mapped at that position. To create a set of shared SNPs for use in GWAS, SNPs for all lines were combined and loci with more than 10% missing data and a minor allele frequency (MAF) below 5% were removed. The remaining SNP loci were subjected to imputation using Beagle 5.0 (Browning, Zhou and Browning, 2018) to impute missing SNPs.

3.5.6 Genome-wide association study (GWAS)

STRUCTURE v2.3.4 (Pritchard, Stephens and Donnelly, 2000) was used to genetically infer the population structure of the panel and produce a population structure matrix. An admixture model was selected and run using 30,000 burn-in iterations and 50,000 Markov Chain Monte Carlo (MCMC) model repetitions for assumed subpopulations of 2–10 with 10 randomly selected, seeded iterations for each assumed subpopulation. The delta *k* method from (Evanno, Regnaut and Goudet, 2005) was applied to all 10 replicates to identify the most likely number of definable subpopulations. This was implemented using the STRUCTURE HARVESTER Python script (Earl and vonHoldt, 2012). Finally, CLUMPP v1.1.2 (Jakobsson and Rosenberg, 2007) was used with 10 independent STRUCTURE replicates to produce a consensus Q matrix for each assumed subpopulation number. GWAS analysis was conducted using the MLM model implemented in GAPITv3.0 (Lipka *et al.*, 2012). Principal component analysis eigenvectors 1–10 or membership coefficient matrices for 3-8 assumed subpopulations deduced above by STRUCTURE were used as covariates in the model to mitigate the effects of hidden familial relatedness. The EMMA method (Hyun *et al.*, 2008) was implemented in GAPIT to create a positive semidefinite kinship matrix required by the MLM model. Each MTA flanking interval was deduced by identifying the SNP position furthest upstream and downstream from the highest associated SNP that was above the -log P threshold of 5.

3.5.7 Identifying regions of divergence

The RefSeq v1.0 genome was split into n genomic windows based on window size (1 Mbp and 100 Kbp) using bedtools makewindows (Quinlan and Hall, 2010). Using the filtered alignments from above, the number of reads mapping to each window was computed using hts-nim-tools v0.0.1 (Pedersen and Quinlan, 2018). To normalise by the sequencing depth of each line, read counts were divided by the number of mapped reads that passed the filters, producing normalised read counts c. Different windows of the genome have variable mapping coverage rates, so to compute coverage deviation we must compare

each window to the same window in the other lines in the collection.
Median normalised read counts, m, were produced, containing the median for each genomic window. Mapping coverage deviation was then defined for each line as:

$$d_i = \frac{c_i}{m_i \cdot \varepsilon}$$

for window i ∈ {1, 2, …, n}, where ε is the median d value across the genome for the line. Statistically significant d values were calculated using the scores function from the R package 'outliers' with median absolute deviation (MAD) and a probability of 0.99. This method was based on Keilwagen *et al.* (2022).

3.5.8 Producing species-specific SNPs

WGS data for 6 *Ae. tauschii* accessions (Luo *et al.*, 2017; Zhou *et al.*, 2021) , 4 *S. cereale* accessions (Bauer *et al.*, 2017), *Secale. vavilovii* (Bauer *et al.*, 2017) and *Thinopyrum ponticum* (Walkowiak *et al.*, 2020), and *T. aestivum* cultivars Weebil (Walkowiak *et al.*, 2020), Norin61 (Walkowiak et al., 2020) and Pavon76 (Coombes *et al.*, 2022) were mapped to RefSeq v1.0 and were filtered and SNP called as described before for the genotyping. Homozygous SNPs were retained if they had between 10 and 60 reads supporting the alternative allele and had an allele frequency >= 0.8. Heterozygous SNPs were retained if they were biallelic with each allele having >= 5 reads in support and an allele frequency >= 0.3. SNPs belonging to a wheat relative and not shared with any of the other wheat relatives or Elite cultivars were retained as species-specific SNPs. SNP deviation scores were calculated by dividing the number of SNPs in each window matched to a species-specific SNP by the mean number of SNPs matched to that species in that window across all the HiBAP I lines.

3.5.9 Synteny between *Ae. tauschii* and *T. aestivum*

The first 10 Mb of chr6D from Chinese Spring and chr6 from *Ae. tauschii* Aet v4.0 (Luo *et al.*, 2017) were aligned using minimap2 (Li, 2018) with parameters -x asm10. Alignments <2.5 Kbp in length or with mapping quality <40 were discarded. The synteny plots were produced using pafr R package (Winter, 2021).

3.5.10 Extracting corresponding region and genes from *Ae. tauschii* genomes

Proteins encoded by genes in the first 10 Mbp of chr6D in Chinese Spring and chr6 in *Ae. tauschii* were aligned using BLASTp from blast+ v2.7.1 (Camacho *et al*., 2009). The protein alignments and the minimap2 alignments were used to anchor the borders of the region commonly introgressed in all lines with the 6D T allele from Chinese Spring to the *Ae. tauschii* genome. The commonly introgressed sequence was extracted from the *Ae. tauschii* reference genome and aligned to the other 4 chromosome-level *Ae. tauschii* genome assemblies using minimap2 (Li, 2018) with parameters -x asm5. Alignments shorter than 5 Kbp or with mapping quality <40 were removed. The coordinates of each orthologous regions were determined manually and the genes within these coordinates extracted by hand. The *Ae. tauschii* genes and their proteins within these segments are considered as candidate genes. BLASTp from blast+ v2.7.1 (Camacho *et al*., 2009) was used to compare these proteins to wheat proteins. Protein domains were identified analysed using HMMER hmmscan (Potter *et al.*, 2018) via ebi using Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, and TreeFam databases. The novelty of genes was determined by aligning the extracted protein sequence to the five *Ae. tauschii* genomes (Luo *et al.*, 2017; Zhou *et al.*, 2021) and the genomes from the 10+ wheat genomes project (Walkowiak *et al.*, 2020) using tBLASTn from blast+ v2.7.1 (Camacho *et al*., 2009).

3.5.11 Reannotation of type-B two component response regulator gene

To test whether the missing myb-binding domain in the TraesCS6D02G014900 annotation was real or an artefact, I manually reannotated the gene. I identified the exon containing the myb-binding domain in the wheat orthologue by aligning the coding sequence from the *Ae. tauschii* orthologue to Chinese Spring RefSeq v1.0 (Appels *et al.*, 2018) using tBLASTn from blast+ v2.7.1 (Camacho *et al*., 2009). I mapped Chinese Spring RNA-seq data from leaf, root and shoot to RefSeq v1.0 (Appels *et al.*, 2018) using HISAT2 (Kim *et al.*, 2019) and assembled transcripts using cufflinks (Trapnell *et al.*, 2012). I visually inspected the coding sequence and RNA-seq alignments using IGV (Robinson *et al.*, 2011), which showed that the myb-binding domain exon is present and expressed in wheat. To check whether the protein has a premature stop codon, I extracted the coding sequence from the assembled transcript and checked for the presence of a complete open reading frame with no stop codons using EMBOSS getorf (Rice, Longden and Bleasby, 2000). Finally, I

checked the presence of intact domains with HMMER hmmscan (Potter *et al.*, 2018) via ebi using the Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, and TreeFam databases.

4 Reference bias caused by introgressions is a major confounding effect in RNA-seq analyses in wheat

This chapter is an adaptation of work that has been published in *BMC Biology* (Coombes *et al.*, 2024) (Appendix D3) and appears with permission granted by the Creative Commons Attribution 4.0 International License.

This work was conceived and carried out by me, with the exception of using OrthoFinder to identify orthologues between the wheat cultivars, which was carried out by my collaborator Thomas Lux from the PGSB plant genomics group at Helmholtz Zentrum München, and Fig. 4-14 was made by my colleague Hannah Rees, using data I generated, while working on a publication that is published in *PLoS Biology* (Rees et al., 2022). Eduard Akhunov provided critical feedback on the accompanying manuscript.

4.1 Abstract

RNA-seq is a fundamental technique in genomics yet reference bias, where transcripts derived from non-reference alleles are quantified less accurately, can undermine the accuracy of RNA-seq quantification and thus the conclusions made downstream. Reference bias in RNA-seq analysis has yet to be explored in complex polyploid genomes such as wheat despite the ubiquity of introgressions in many of these genomes, which introduce blocks of highly divergent genes. Using both simulated and experimental data, I found that RNA-seq alignment in wheat suffers from widespread reference bias which is largely driven by divergent introgressed genes. This leads to the expression of many genes being underestimated, incorrect assessment of homoeologue expression balance, and false associations of expression levels with traits and genetic variants. By incorporating divergent gene models from ten wheat genome assemblies into a pantranscriptome reference, I present a novel method to reduce reference bias, which can be readily scaled to capture more variation as new genome and transcriptome data becomes available.

4.2 Introduction

4.2.1 Quantifying gene expression using RNA-seq

Quantifying gene expression using RNA-seq has become a fundamental technique in genomics research, facilitating insights into the functional landscape of genomes. It has

been employed in numerous publications across a range of biological systems to identify candidate genes underlying traits of interest, uncover transcriptional pathways and networks, and investigate hypotheses relating to gene and transcriptional evolution and adaptation.

The core of RNA-seq data processing is quantifying the level of expression for each transcript and/or gene in each sample. RNA is extracted from the biological sample, converted to cDNA, and sequenced. This is usually done using Illumina sequencing, where each read is 100-250bp in length and often paired. This pool of sequenced reads represents a snapshot of the expression of each gene at the time of sampling, with the number of reads derived from a gene proportional to the level of expression of that gene.

Various algorithms have been devised to process RNA-seq reads and produce expression counts for each transcript/gene. Two main approaches are currently in popular usage: alignment and pseudoalignment. Alignment algorithms such as HISAT2 (Kim *et al.*, 2019) and STAR (Dobin *et al.*, 2013) rely on traditional sequence alignment algorithms to map the RNA-seq reads to a reference genome. For eukaroytes, these tools perform spliced mapping, which takes splicing across exon and intron boundaries into account. This is necessary because the reads are derived from post-splicing mature mRNA but the reference to which they are mapped includes the intron sequences. Alignment-based algorithms are typically very precise but are computationally expensive, consuming a lot of memory and taking a long time. They are also highly sensitive to reference quality.

A faster and less computationally intensive approach is pseudoalignment, used by tools such as kallisto (Bray *et al.*, 2016) and Salmon (Patro *et al.*, 2017). Instead of mapping to a reference genome using alignment algorithms, they use k-mer based pseudoalignment to identify the best matches between each read and a transcript from a defined set of transcript sequences, modelling the probability of each read having arisen from each transcript. When speed and efficiency is important and neither the precise location of each read in the genome nor alternative splicing events are needed, pseudoalignment offers a useful alternative.

4.2.2 Relative homoeologue expression in wheat triads

As detailed in section 1.10, many of the genes in wheat, around 51.1% high-confidence genes in the RefSeq v1.1 annotation (Ramírez-González *et al*., 2018), exist in triads which

consist of three homoeologous genes, one belonging to each subgenome. The relative mRNA expression of homoeologues within a triad is biologically important as variation in the relative expression of homoeologues between tissues or environmental conditions may confer phenotypic plasticity (Ramírez-González *et al.*, 2018). Additionally, understanding the interaction between homoeologue expression is important for crop development, where genes present in triads may be targeted to alter agronomic traits.

The relative homoeologue expression of a triad can be described as belonging to one of seven categories: balanced, where all 3 homoeologues are expressed similarly; A, B or D supressed, where the supressed homoeologue is expressed much less than the other two homoeologues copies; and A, B or D dominant, where the dominant homoeologue is expressed much higher than the other two dominant homoeologues. While varying in different tissues and conditions, around 70% of triads show balanced expression. Of the remaining triads, more than twice the number of triads possess a suppressed homoeologue than a dominant homoeologue (Ramírez-González *et al*., 2018).

4.2.3 Reference bias

Making meaningful inferences from RNA-seq data relies upon the accuracy of alignment and quantification. Downstream analyses and their subsequent interpretation assume that the estimated gene expression levels reflect actual gene expression in the biological samples. However, nucleotide variation between the sequenced sample and the reference genome/transcriptome in the coding region of genes can cause errors in read assignment during the alignment/pseudoalignment step. Some reads may be unassigned, while others may be assigned to the wrong locus. This source of error is widely known as reference bias as transcripts derived from alleles present in the reference sequence will be quantified more accurately (Günther and Nettelblad, 2019).

The reduction in accuracy caused by reference bias has the potential to impact downstream analyses and lead to incorrect or incomplete findings. For example, Thorburn *et al.* (2023) found that mapping sequencing data to a single reference genome causes reference bias that results in inaccurate findings in population genomic studies. While this study focused on mapping DNA reads, it can be assumed to apply to RNA-seq reads. Zhan, Griswold and Lukens (2021) found that in maize, reference bias strongly affects the accuracy of transcript abundance estimates from RNA-seq reads. They reanalysed RNA-seq data generated from 105 lines from a B73xMo17 recombinant inbred

line population (Li *et al.*, 2013) and found that for about 50% of eQTL alleles detected by Li *et al*. (2013) their detection depended on the reference genome used to align the RNA-seq reads. This suggests that, as Li et al. (2013) only used the B73 reference genome, around 50% of the eQTLs they discovered are likely to be false positives. Munger *et al.* (2014) found that using a single reference genome to align RNA-seq reads from a multi-parent mouse population resulted in only 88.2% accuracy in eQTL assignment, compared to 98.3% when using individualised genomes.

4.2.4 Impact of reference bias in complex polyploid genomes like wheat

The impact of reference bias in RNA-seq analysis hasn't been assessed in complex polyploid genomes such as wheat despite these genomes having characteristics that may increase the extent and degree of reference bias relative to species with simpler genomes. As many genes in wheat have homoeologues in the other subgenomes, yet RNA-seq reads are derived from all subgenomes at once, read assignment must be able to distinguish reads deriving from the different subgenomes. Accurate discrimination of wheat homoeologue RNA-seq reads has been demonstrated with both pseudoalignment (Ramírez-González *et al.*, 2018; He *et al.*, 2022) (99.9% accuracy) and alignment-based (98% accuracy) (He *et al.*, 2022) methods when mapping reads back to the genome from which they derived. However, when mapping reads from a different genotype, unequal divergence between homoeologues relative to the reference genome may compromise the accuracy of the expression balance estimation between homoeologues.

As discussed in section 1.8, introgressions are very common in wheat. As introgressions introduce divergent gene models from different species, it is likely that they are a source of reference bias in RNA-seq analysis due to RNA-seq reads derived from these introgressed genes being unable to be assigned to the reference genome. Relaxing read mapping/alignment parameters to allow more reads from introgressions to map would reduce the number of unassigned reads; however this would increase the number of reads assigned to the wrong location in the genome. This would be much more pronounced in polyploid genomes because read mapping must be stringent enough for reads deriving from the different subgenomes to be correctly distinguished. Therefore, working with a species such as wheat which is both polyploid and has many introgressions creates a unique problem.

4.2.5 Chapter aims

- Assess the extent of reference bias in RNA-seq quantification in wheat using simulated datasets from ten chromosome-level genome assemblies.
- Construct pantranscriptome reference to reduce reference bias.
- Assess how well the pantranscriptome reference reduces reference bias in the simulated datasets.
- Quantify expression using experimentally-generated RNA-seq data using the Chinese Spring and pantranscriptome references.
- Explore the impact of reference bias in the experimentally-generated dataset, focusing on the findings relating to homoeologue expression bias.

4.3 Results

4.3.1 Using simulated RNA-seq data to estimate the extent of reference bias in wheat

To explore the impact of reference bias on the quantification of gene expression in wheat, I first simulated 1000 read pairs from the longest CDS sequence of each high-confidence gene in Chinese Spring RefSeq v1.1 (Appels *et al.*, 2018) and the nine chromosome-level genome assemblies produced from the wheat pangenome project (Walkowiak *et al.*, 2020; White *et al.*, 2024) if the transcript was at least 500bp. By simulating the same number of RNA-seq read pairs from each gene, all genes should be quantified to the same level. Genes whose estimated expression deviates from the expected value are assumed to have experienced errors during the alignment or quantification process. These reads were pseudoaligned or aligned to the Chinese Spring reference transcriptome or genome using kallisto or STAR, respectively. These algorithms were chosen as a representative of pseudoalignment and alignment-based methods as they are among the most common RNA-seq quantification tools used by the wheat research community.

When mapping simulated Chinese Spring reads to the Chinese Spring genome (hereafter referred to as self-mapping) no reference bias will be present, so the accuracy of quantification is determined by the ability of the alignment algorithm to correctly assign reads to a matched reference. Mapping RNA-seq reads from the other cultivars to the Chinese Spring reference (hereafter called cross-mapping) allows the extent of reference bias caused by variation between cultivar genomes to be determined. The difference

between the accuracy of self-mapping and the accuracy of cross-mapping can be assumed to be caused by reference bias.

For self-mapping, only the genes from which reads were simulated (due to having a transcript over 500bp in length) were examined to avoid lots of genes having an artificial gene count of zero and making quantification accuracy look lower than it actually is. Similarly, for cross-mapping, only the genes from which reads were simulated in that cultivar and are in a 1-to-1 orthologous relationship with a Chinese Spring gene from which reads were simulated were examined. The way these genes were chosen for analysis explains why fewer genes were examined for cross-mapping than for self-mapping.

4.3.1.1 Impact of reference bias on gene-level expression counts

As 1000 read pairs were simulated per gene, genes correctly quantified should have 1000 read pairs assigned during the alignment and have an estimated read count of 1000. In order to count the number of genes correctly and incorrectly quantified, a threshold had to be used. Genes with fewer than 500 read pairs assigned were classified as underestimated, genes with more than 1500 read pairs assigned were classified as overestimated, and genes with between 500 and 1500 read pairs assigned were classified as correctly quantified.

Self-mapping predictably yields very accurate estimates of gene expression, with kallisto slightly outperforming STAR (Figs. 4-1a, 4-1b, Table 4-1). Using kallisto, 88401/88443 (99.95%) genes were correctly quantified. 32 genes were underestimated, and 10 genes were overestimated. Using STAR, 87689/88443 (99.15%) were correctly quantified. 504 genes were underestimated, and 250 genes were overestimated.

Cross-mapping yielded much less accurate estimation of gene expression with a skew towards underestimation (Figs. 4-1a, 4-1b, Table 4-1). The percentage of genes correctly quantified ranged from 55773/63001 (88.53%) for Lancer, with 5700 (9.05%) and 1528 (2.43%) genes under and overestimated, respectively, to 58468/64077 (91.2%) for Norin61, with 2527 (3.94%) and 3082 (4.81%) genes under and overestimated, respectively. For cross-mapping, unlike self-mapping, STAR appears to perform better than kallisto; the proportion of correctly quantified genes ranged from 58390/63001 (92.68%) for Lancer, with 3916 and 695 genes under and overestimated, respectively, to

59648 (93.1%) for Norin61, with 2450 and 1979 genes under and overestimated, respectively.

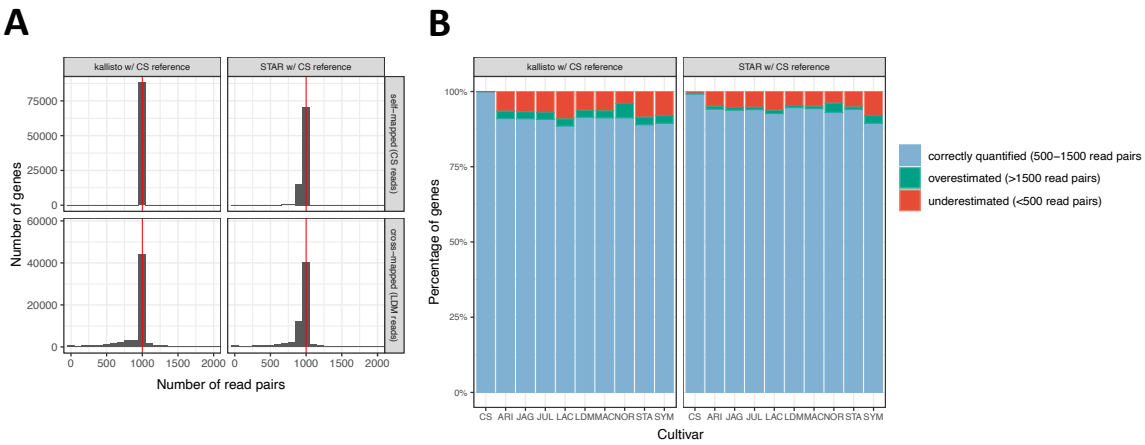


Figure 4-1. Estimating the extent of reference bias on gene-level read counts in wheat using simulated RNA-seq reads.

A) Distribution of read counts when self-mapping Chinese Spring simulated reads or cross-mapping Landmark simulated reads, aligned to Chinese Spring using either kallisto or STAR. If quantification is perfectly accurate, we expect to see a single bar at 1000 read pairs on the x axis. **B)** Percentage of genes with expression estimated correctly, expression underestimated (< 500 read pairs) and expression overestimated (> 1500 read pairs) for simulated reads from 10 cultivars aligned to Chinese Spring using either kallisto or STAR.

Table 4-1. Number of genes correctly quantified (500 - 1500 read pairs), underestimated (< 500 read pairs), and overestimated (> 1500 read pairs) from simulated RNA-seq data, using kallisto or STAR with the Chinese Spring reference.

Cultivar		Kaliisto with Chinese Spring reference			STAR with Chinese Spring reference		
	No. genes	Correctly estimated	Underestimated	Overestimated	Correctly estimated	Underestimated	Overestimated
ARI	59515	54154	3877	1484	56007	2895	613
CS	88443	88399	32	12	87681	506	256
JAG	62646	56955	4215	1476	58744	3328	574
JUL	63384	57505	4400	1479	59556	3324	504
LAC	63001	55756	5716	1529	58383	3923	695
LDM	63517	58073	3967	1477	60114	2969	434
MAC	63203	57655	4033	1515	59598	3075	530
NOR	64077	58455	2536	3086	59643	2455	1979
STA	63001	56107	5254	1640	59362	3237	502
SYM	59370	53095	4745	1530	53095	4745	1530

4.3.1.2 Impact of reference bias on assignment of triad expression balance

To explore the effect of reference bias on the quantification of homoeologue expression balance, I calculated the proportion of triads belonging to each category that defines a different state of relative homoeologue expression. As reads were simulated evenly across genes, all triads should be classified as balanced; therefore, triads classified as imbalanced (one or two homoeologues with expression greater than the other(s)) are considered incorrectly classified. The definition of each homoeologue balance state is the same as in (Ramírez-González *et al.*, 2018) and is described in section 4.5.2.

When self-mapping, 99.9% of triads were correctly classified using kallisto and 99.8% were correctly classified using STAR (Figs. 4-2a, 4-2b, Table 4-2). When cross-mapping, the percentage of correctly classified triads were much lower, ranging from 80.97% (Lancer) to 93.84% (Norin61) using kallisto and from 90.23% (Lancer) to 96.12% (Norin61) using STAR (Figs. 4-2a, 4-2b, Table 4-2). Across the cultivars, triads incorrectly classified as suppressed, where one homoeologue is estimated to be expressed less than the others, were far more common than triads incorrectly classified as dominant, where one homoeologue is estimated to be expressed more highly than the others (Fig. 4-2B). This reflects how the reference bias leads to more underestimated than overestimated genes.

The B subgenome has the most, and the D subgenome the fewest, number of triads incorrectly classified as suppressed. This still holds if we disregard cultivars with very large introgressions from wild relatives in the B subgenome, such as Lancer. This pattern is in line with observations of greater diversity in the A and B subgenomes, with the B subgenome having the highest level of diversity (Cheng *et al.*, 2019). This is largely caused by gene flow from wild tetraploid *T. dicoccoides* throughout the cultivation history of bread wheat without comparable gene flow to the D subgenome (Dvorak *et al.*, 2006; He *et al.*, 2019).

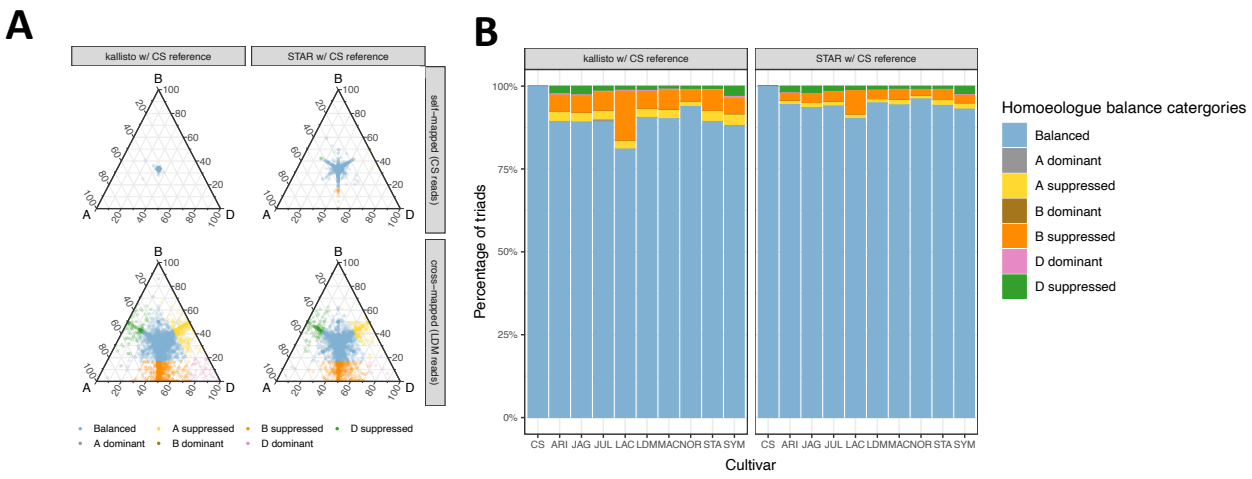


Figure 4-2. Estimating the extent of reference bias on the classification of triad expression balance in wheat using simulated RNA-seq reads.

A) Balance of homoeologue expression across triads when self-mapping Chinese Spring simulated reads or cross-mapping Landmark simulated reads, aligned to Chinese Spring using either kallisto or STAR. Each point on the ternary plot represents one triad. Points towards a corner indicate dominant expression of that homoeologue, while points opposite a corner indicate suppression of that homoeologue. If all triads were perfectly classified, we expect to see all dots at the centre of the plot. **B)** Percentage of triads in each expression category, using simulated reads from 10 cultivars aligned to Chinese Spring using either kallisto or STAR.

Table 4-2. Percentage of triads classified in each expression category from simulated RNA-seq data, using kallisto or STAR with the Chinese Spring reference. Values are rounded to three significant figures.

Cultivar	Kaliisto with Chinese Spring reference							STAR with Chinese Spring reference						
	Balanced	A dominant	B dominant	D dominant	A suppressed	B suppressed	D suppressed	Balanced	A dominant	B dominant	D dominant	A suppressed	B suppressed	D suppressed
ARI	89.3	0.38	0.0441	0.433	2.75	5.15	1.99	94.4	0.124	0.0177	0.141	1.27	2.37	1.68
CS	100	0.00	0.00	0.00	0.0062	0.00	0.00	99.8	0.00	0.00	0.00	0.0432	0.0987	0.0308
JAG	89.2	0.227	0.0805	0.344	2.68	5.21	2.30	93.5	0.0659	0.0146	0.132	1.51	2.78	1.98
JUL	89.2	0.932	0.0783	0.356	2.60	0.0783	1.21	94.0	0.192	0.0071	0.171	1.28	3.06	1.29
LAC	81.0	0.300	0.0572	0.321	2.39	15.0	0.922	90.2	0.15	0.0214	0.0857	1.25	7.27	1.00
LDM	90.5	0.0637	0.0283	0.361	2.75	5.32	0.956	94.9	0.0283	0.0071	0.092	1.27	2.80	0.864
MAC	90.2	0.0783	0.0712	0.477	2.84	5.79	0.577	94.3	0.0498	0.0427	0.178	1.71	3.07	0.634
NOR	93.8	0.0632	0.0562	0.176	1.62	3.58	0.667	96.1	0.0632	0.0211	0.0492	1.09	1.99	0.667
STA	89.3	0.0932	0.0932	0.308	3.23	6.26	0.695	94.2	0.043	0.0645	0.172	1.71	3.18	0.652
SYM	88.0	0.449	0.0793	0.599	3.22	4.92	2.78	93.1	0.141	0.0264	0.212	1.77	2.56	2.23

4.3.1.3 Impact of reference bias on cultivar comparison and introgressions as source of reference bias

To explore where in the genome the incorrectly quantified genes are and if they overlap with introgressions, I compared the estimated expression of Lancer and Jagger 1-to-1 orthologues, whose simulated reads were aligned to Chinese Spring using STAR as it performed the best for cross-mapping in section 4.3.1.1. A cultivar comparison was used here instead of one cultivar compared to expected read counts for two reasons. Firstly, this provides insights into how many genes are incorrectly quantified when comparing the expression of genes between two cultivars. Secondly, as these cultivars possess different introgressions, it allows the impact of introgressions on expression quantification to be explored in more detail. Genes with read counts > 1.5x or < 1/1.5x compared to the other cultivar were classified as incorrectly quantified. Using STAR, 4791/60338 (7.94%) genes were incorrectly quantified (Fig. 4-3A, 4-3B).

I identified introgressed regions by looking for blocks of reduced CDS nucleotide identity between pairs of 1-to-1 Lancer-Jagger orthologues (Fig. 4-3C). Very high gene-level divergence is strongly indicative of introgressed material; indeed several of these blocks correspond to previously characterised introgressions and likely additional introgressions that are yet to be characterised. These introgressions include (coordinates based on Chinese Spring RefSeq v1.0): *Ae. ventricosa* introgression in Jagger (chr2A:1-24643290) (Walkowiak *et al.*, 2020; Gao *et al.*, 2021; Keilwagen *et al.*, 2022); *T. timopheevii* introgression in Lancer (chr2B:89506326-756157100) (Walkowiak *et al.*, 2020; Keilwagen *et al.*, 2022); *Aegilops markgrafii* introgression in Jagger (chr2D:570141481-613325841) (Keilwagen *et al.*, 2022); and a *Th. ponticum* introgression in Lancer (chr3D:591971000-615552423) (Walkowiak *et al.*, 2020; Keilwagen *et al.*, 2022).

There is a clear overlap between blocks of incorrectly quantified genes and regions of high gene-level nucleotide divergence between the cultivars (Figs. 4-3a, 4-3c), suggesting the introgressions are a major source of the reference bias observed here. Genes with an introgressed copy in Lancer tend to be underestimated in Lancer and genes with an introgressed copy in Jagger tend to be underestimated in Jagger. To support the involvement of introgressions in reference bias, 1881/3054 (61.59%) genes known to be introgressed (belonging to one of the four previously characterised introgressions listed above) were incorrectly quantified between the two cultivars, compared to 2910/57284

(5.08%) genes outside of the known introgressions that were incorrectly quantified (chi-squared p-value < 2.2e-16) (Fig. 4-3D). In further support of CDS divergence being a predominant contributing factor to miscalled genes, I found genes that were miscalled to have a mean CDS identity between orthologue pairs of 97.3% compared to a mean of 99.9% for genes not miscalled (p-value < 2.2e-16, 95 CI = [2.45, 2.63]) (Fig. 4-3E). The percentage of genes miscalled ranges from 83.2% for genes with <96% CDS identity to just 2.9% for genes with >=99% identity (Fig. 4-3F).

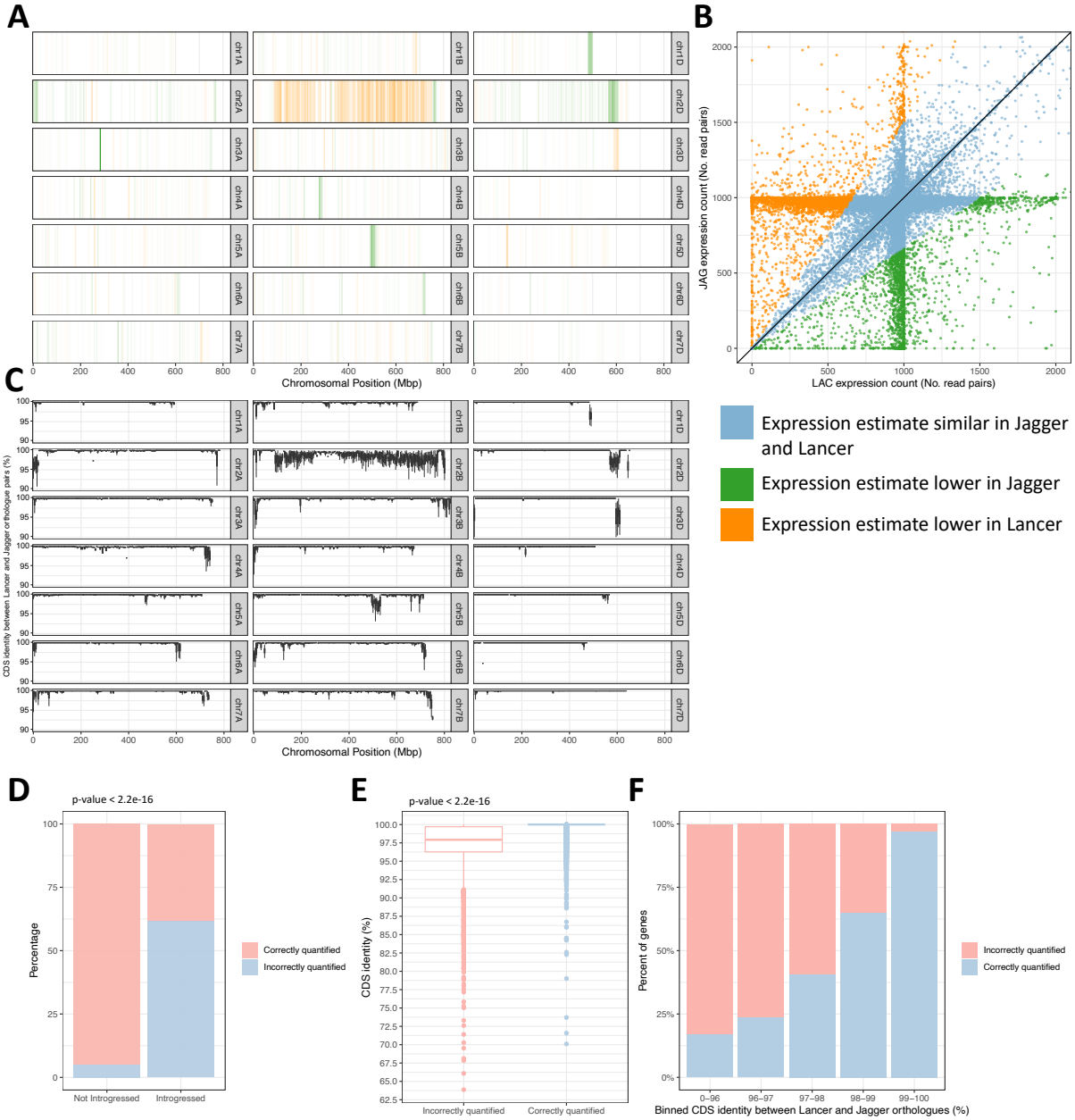


Figure 4-3. Exploring the impact of reference bias on expression differences between cultivars and enrichment of incorrectly quantified genes within introgressions.
A) The chromosome plot shows the distribution of incorrectly quantified genes in 5 Mbp windows, coloured by the cultivar in which the estimated expression is lower; orange blocks are underestimated in Lancer compared to Jagger, while green blocks are underestimated in Jagger compared to Lancer. The reads are aligned using STAR as this outperformed kallisto for cross-mapping. **B)** Scatter plot shows expression counts for Lancer-Jagger orthologue pairs. Genes are considered incorrectly quantified if their estimated read count is 1.5x or 1/1.5x the other cultivar. **C)** CDS nucleotide identity between Lancer and Jagger 1-to-1 orthologue pairs, binned into 5 Mbp genomic windows based on Chinese Spring RefSeq v1.0 coordinates. **D)** Percentage of genes correctly and incorrectly quantified in characterised introgressed regions and regions not characterised as introgressed. **E)** CDS nucleotide identity between Lancer and Jagger 1-to-1 orthologue pairs for those that are incorrectly quantified and for those that are correctly quantified. **F)** Percentage of genes correctly and incorrectly quantified, split into bins of different levels of CDS nucleotide identity.

4.3.2 Constructing a pantranscriptome reference to reduce reference bias

The 10+ wheat genomes project generated chromosome-level de novo assembled genomes for nine wheat cultivars to supplement the Chinese Spring reference genome (Walkowiak *et al.*, 2020). These genomes include numerous introgressions that contribute significantly to the observed reference bias. High-quality gene annotations for these genome assemblies have since been generated, using a comparable method to that used to annotate Chinese Spring (White *et al.*, 2024).

To reduce reference bias caused by divergent gene models, I constructed a new kallisto transcriptome reference that I've called the pantranscriptome reference. Transcripts from all 107891 Chinese Spring RefSeq v1.1 high-confidence genes were included as the base. To this, I added transcripts from the nine chromosome-level genome assemblies from Walkowiak *et al*. (2020) if the transcript is derived from a gene that is present in a 1-to-1 relationship with one of the 107891 Chinese Spring genes (Table 4-3). This resulted in a set of transcripts from 762877 genes from the 10 cultivars; 107891 genes from Chinese Spring and a mean of 72887 genes from each of the nine other cultivars (Fig. 4-4). 80211 Chinese Spring genes had at least one 1-to-1 orthologue in another cultivar, while 59639 Chinese Spring genes had a 1-to-1 orthologue in all nine other cultivars (Fig. 4-5, Table 4-3) based on OrthoFinder (Emms and Kelly, 2019) orthogroup assignments.

Table 4-3. Number of genes from each cultivar in the pantranscriptome reference.

Cultivar	**Number of genes in pantranscriptome reference**
CS	107891
ARI	69519
JAG	73193
JUL	73892
LAC	73522
LDM	74225
MAC	73732
NOR	74920
STA	73638
SYM	69345

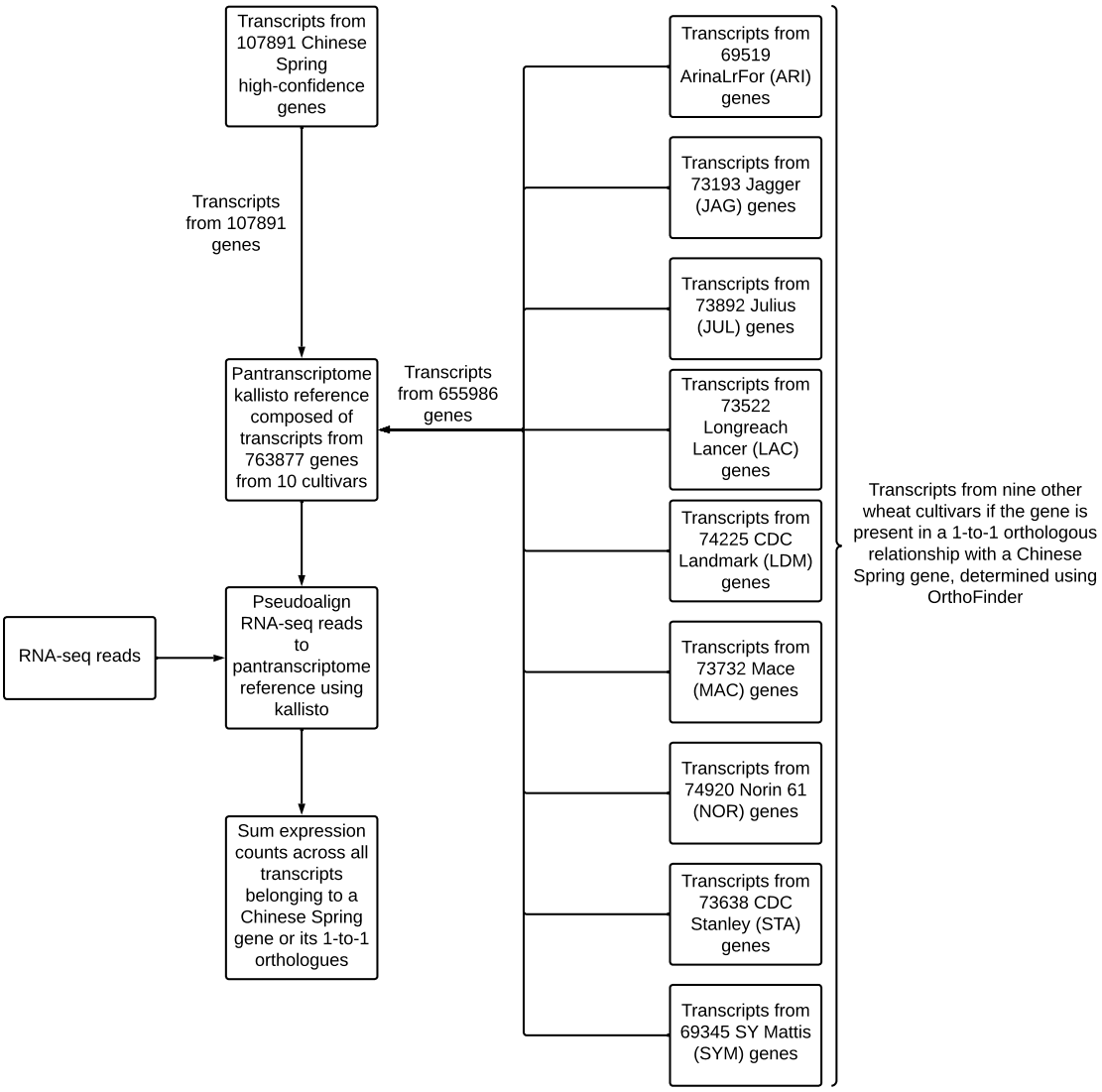


Figure 4-4. Creation of the pantranscriptome reference and how RNA-seq reads are aligned to it.

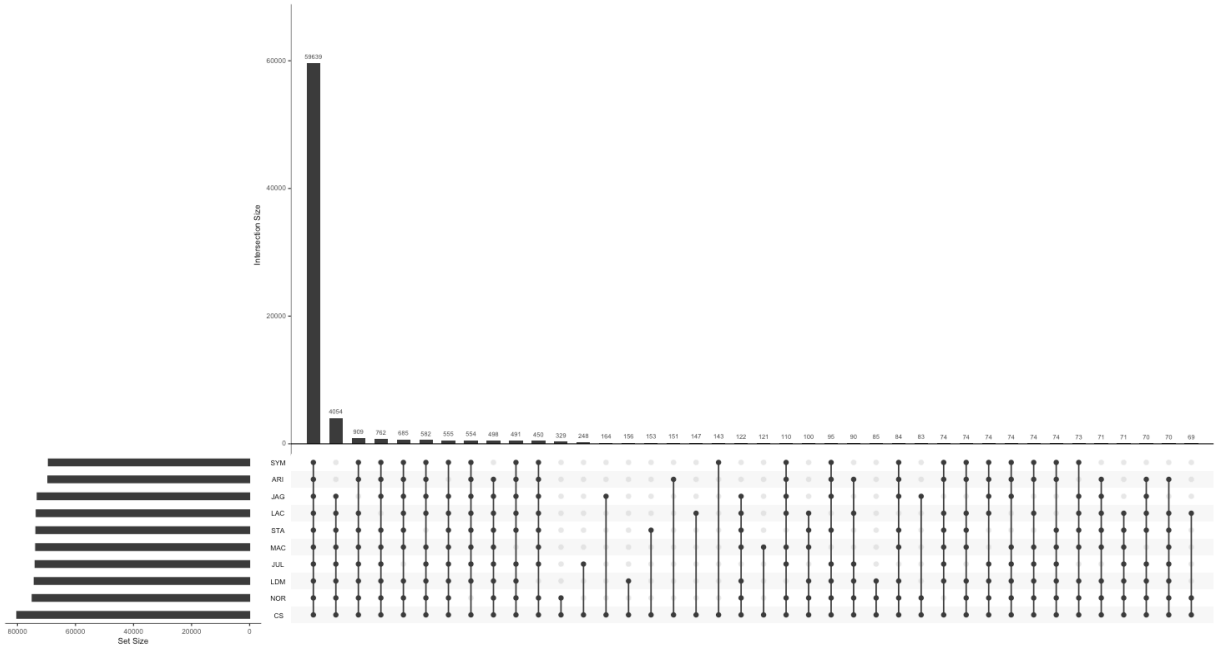


Figure 4-5. Upset plot of 1-to-1 orthologues used for the construction of the pantranscriptome reference.

Only genes in a 1-to-1 relationship with a Chinese Spring gene are included. Plot was generated using the UpSetR package in R.

Kallisto pseudoalignment was then conducted as usual but using this pantranscriptome reference instead of the Chinese Spring reference. After pseudoalignment, read counts and TPMs were summed across all transcripts corresponding to a given Chinese Spring gene, resulting in an expression matrix with the same number of genes and gene IDs as when using the Chinese Spring reference.

Kallisto splits read counts evenly across transcripts with an identical match so including identical redundant transcripts in the reference does not cause errors in read count quantification provided the reads counts are summed across the transcripts after kallisto pseudoalignment. All transcripts corresponding to a gene can thus be added without issue. To ensure this is the case and that the pantranscriptome reference doesn't introduce any additional errors from adding redundant transcripts, I compared quantified expression counts between four difference references: Chinese Spring; the pantranscriptome reference; Chinese Spring plus the Landmark transcripts from genes in a 1-to-1 relationship with a Chinese Spring gene; and the pantranscriptome reference without the Landmark transcripts. Simulated RNA-seq reads from Landmark were used for pseudoalignment.

Of these four references, the pantranscriptome reference resulted in the highest accuracy, correctly quantifying 97.5% of genes. Chinese Spring plus Landmark transcripts performed similarly, with 97.5% of genes correctly quantified. This result demonstrates that the inclusion of redundant transcripts introduces minimal errors in the kallisto pseudoalignment. Using the pantranscriptome reference without the Landmark transcripts resulted in slightly less accurate quantification, with 96.8% of genes correctly quantified. The difference is likely attributable to uniquely introgressed genes in Landmark that are absent from the other cultivars. Nevertheless, due to many introgressed genes being common between cultivars, it still significantly outperformed the use of Chinese Spring alone, which resulted in just 91.4% genes being correctly quantified.

4.3.3 Impact of the pantranscriptome reference on reference bias using simulated data

Using the pantranscriptome reference instead of Chinese Spring to quantify expression from the simulated RNA-seq reads resulted in much more accurate quantification for genes that were previously underestimated when cross-mapping, removing nearly all gene counts below 1000. There was little change in the number of genes overestimated

and minimal change in the distribution of read counts when self-mapping (Fig. 4-6, Table 4-4). For Lancer, the cultivar with the largest reference bias, using the pantranscriptome reference increased the number of genes correctly quantified from 58390/63001 (92.68%) using STAR to 61352/63001 (97.38%). Using the pantranscriptome reference, only 2 genes remained quantified below 500 read pairs compared to the 3916 genes quantified below 500 read pairs when using the Chinese Spring reference.

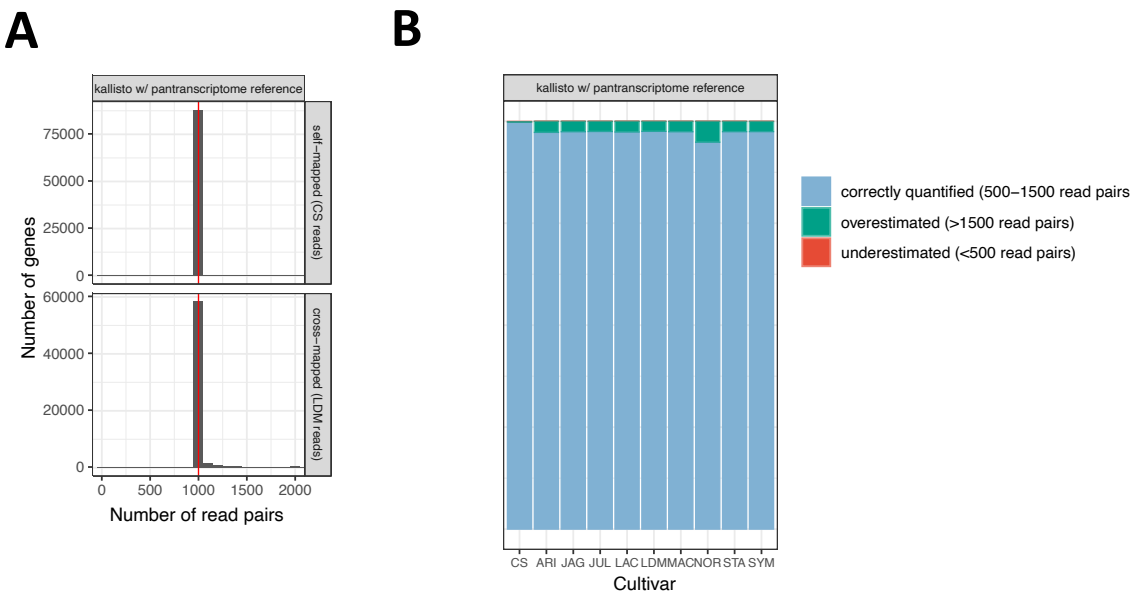


Figure 4-6. Estimating the remaining impact of reference bias on gene-level read counts in wheat using simulated RNA-seq reads when using the pantranscriptome reference. **A)** Distribution of read counts when using kallisto to align Chinese Spring simulated reads or Landmark simulated reads to the pantranscriptome reference. If quantification is perfectly accurate, we expect to see a single bar at 1000 read pairs on the x axis. **B)** Percentage of genes with expression estimated correctly, expression underestimated (< 500 read pairs) and expression overestimated (> 1500 read pairs) for simulated reads from 10 cultivars aligned to the pantranscriptome reference using kallisto.

Table 4-4. Number of genes correctly quantified (500-1500 read pairs), underestimated (< 500 read pairs), and overestimated (> 1500 read pairs) from simulated RNA-seq data, using kallisto with the pantranscriptome reference.

Cultivar	No. of genes	Correctly quantified	Underestimated	Overestimated
ARI	59515	57894	10	1611
CS	88443	88288	92	63
JAG	62646	61040	9	1597
JUL	63384	61796	9	1579
LAC	63001	61344	5	1652
LDM	63517	61948	6	1563
MAC	63203	61581	6	1616
NOR	64077	60789	7	3281
STA	63001	61490	15	1596
SYM	59370	57863	15	1492

The number of triads with correctly assigned homoeologue expression balance also greatly increased when using the pantranscriptome reference (Fig. 4-7, Table 4-5). All cross-mapped cultivars had at least 99.89% triads correctly assigned as balanced; this compares to between 80.97% and 93.84% of triads correctly assigned using kallisto with the Chinese Spring reference, and between 90.23% to 96.12% of triads correctly assigned using STAR to align to Chinese Spring.

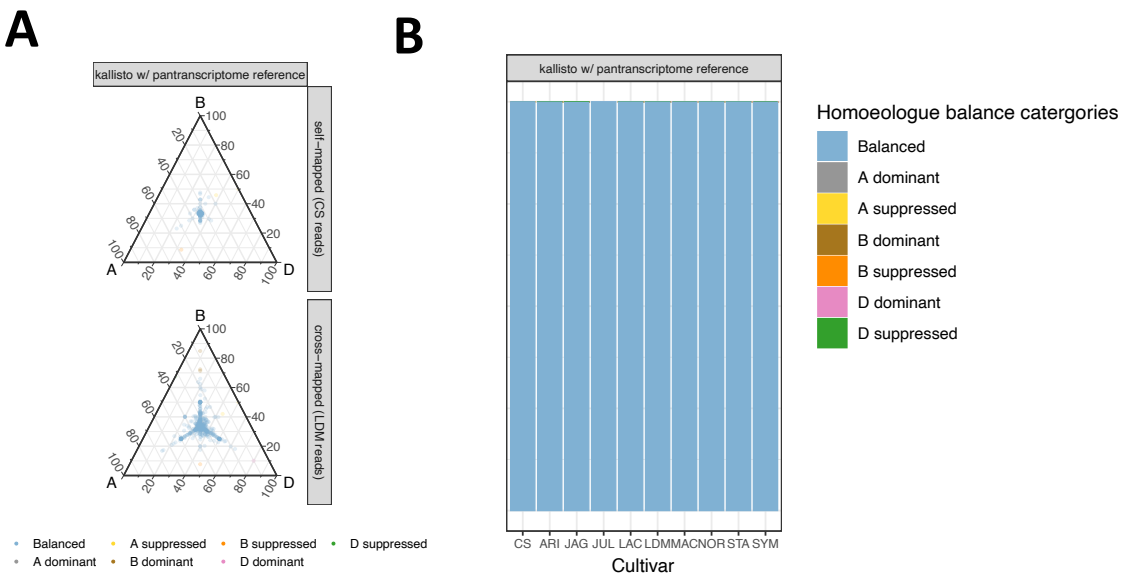


Figure 4-7. Estimating the extent of reference bias on the classification of triad expression balance in wheat using simulated RNA-seq reads.
A) Balance of homoeologue expression across triads when using kallisto to align Chinese Spring simulated reads or Landmark simulated reads to the pantranscriptome reference. Each point on the ternary plot represents one triad. Points towards a corner indicate dominant expression of that homoeologue, while points opposite a corner indicate suppression of that homoeologue. If all triads were perfectly classified, we would expect to see all dots at the centre of the plot. **B)** Percentage of triads in each expression category, using simulated reads from 10 cultivars aligned to the pantranscriptome reference using kallisto.

Table 4-5. Percentage of triads classified in each expression category from simulated RNA-seq data, using kallisto with the pantranscriptome reference. Values rounded to three significant figures.

Cultivar	Balanced	A dominant	B dominant	D dominant	A suppressed	B suppressed	D suppressed
ARI	99.9	0.0265	0.0088	0.00	0.0265	0.0265	0.0088
CS	100	0.00	0.00	0.00	0.0123	0.0062	0.00
JAG	99.9	0.0146	0.0219	0.0146	0.0146	0.0073	0.0073
JUL	99.9	0.0213	0.0142	0.0071	0.0071	0.0071	0.0142
LAC	99.9	0.0143	0.0214	0.0071	0.0071	0.0071	0.0071
LDM	99.9	0.0142	0.0212	0.0071	0.0212	0.0071	0.0071
MAC	99.9	0.0285	0.0214	0.0142	0.0214	0.00	0.00
NOR	99.9	0.00	0.007	0.007	0.014	0.007	0.0211
STA	99.9	0.0215	0.0072	0.0072	0.0215	0.00	0.0072
SYM	99.9	0.0264	0.0176	0.0088	0.0176	0.0441	0.00

Finally, I tested how the number of genes incorrectly quantified between Jagger and Lancer changed when using the pantranscriptome reference (Fig. 4-8). Using the pantranscriptome reference reduced the number of genes incorrectly quantified in one cultivar from 4971/60338 (7.94%) to 617 (1.02%) (Fig. 4-8). Only 23 genes (0.0381%) remain incorrectly quantified due to underestimation in one cultivar.

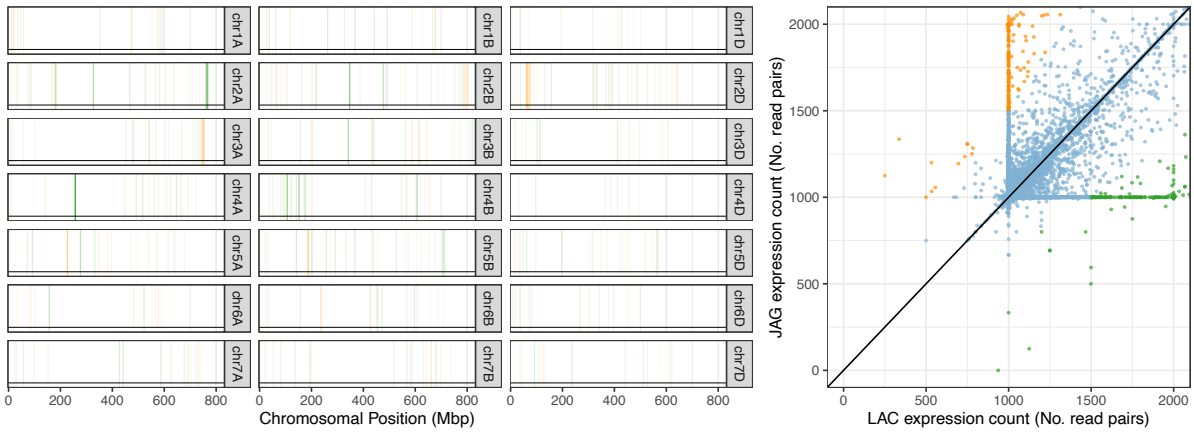


Figure 4-8. Remaining incorrectly quantified genes after correction using the pantranscriptome reference.
Scatter plot shows expression counts for simulated reads of Lancer-Jagger orthologue pairs when using kallisto with the pantranscriptome reference. Genes are considered incorrectly quantified if their estimated read count is 1.5x or 1/1.5x the other cultivar. The chromosome plot shows the distribution of incorrectly quantified genes in 5 Mbp windows, coloured by the cultivar in which the estimated expression is lower; orange blocks are underestimated in Lancer compared to Jagger, while green blocks are underestimated in Jagger compared to Lancer.

The error in quantification that remains when using the pantranscriptome reference is almost all due to the overestimation of gene expression. This is likely caused by copy number variation or presence/absence variation between cultivars, as opposed to divergence between orthologous gene models.

4.3.4 Exploring reference bias caused by introgressions in experimentally-generated RNA-seq data

Simulated RNA-seq data is unlikely to capture the complete picture of a real experiment (Srivastava *et al.*, 2020). While the simulations highlight theoretical errors, it is important to assess how reference bias impacts published findings and how using the pantranscriptome reference corrects errors in experimentally-generated data.

4.3.4.1 Reanalysing data from He *et al*. (2022)

The experimentally-generated dataset I used to further explore the impact of reference bias in wheat was generated by He *et al.* (2022). In this study, RNA-seq data from 198 diverse wheat accessions, alongside enrichment capture paired-end DNA reads was used to uncover expression quantitative trait loci (eQTLs) and link them to productivity traits and the relative expression of homoeologues. eQTLs are like QTLs/MTAs identified through genome-wide association studies, except the phenotype data is gene expression values instead of physiological traits. eQTLs can uncover genetic variation underlying the regulatory control of gene expression variation between accessions.

Of particular relevance for the work in this chapter, He *et al*. (2022) identified a set of genes whose expression was negatively correlated with one of their homoeologues due to one homoeologue having a low level of expression in a subset of accessions. The presence of the lowly expressed alleles in accessions was associated with various important productivity traits. The authors hypothesised that this expression dosage is likely driving the observed phenotypic variation and has itself been driven and maintained by selection.

This set contains 59 genes to which I have added *ELF3-D1* for two reasons. Firstly, although *ELF3-D1* didn't fall into the set of very negatively correlated 59 genes as defined by He *et al*. (2022), it was used as a case example due to its agronomic significance in determining heading date. Secondly, it still exhibited negative correlation with its B homoeologue *ELF3-B1*, with this expression bias associated with agronomic traits. This set of 60 genes is hereafter referred to as genes showing a lack of expression correlation.

Firstly, I set out to identify potential introgressed regions within these accessions to ascertain whether the genes showing a lack of expression correlation tend to be found within these regions. If so, this could indicate an increased likelihood for their classification as lacking expression correlation to have been affected by reference bias. To do this, I mapped the enrichment capture Illumina paired-end DNA reads to Chinese Spring RefSeq v1.0 and for each 1 Mbp genomic window, calculated the mapping coverage deviation between each line and the median for that window across the accessions, labelling windows with a coverage deviation value significantly below 1 in an accession as possessing an introgression or deletion (Fig. 4-9A). There is more divergent material in the A and B subgenomes than the D subgenome, which is expected based on

the higher levels of gene flow to the A and B subgenomes over the cultivation history of wheat (Dvorak *et al.*, 2006; He *et al.*, 2019; Wang *et al.*, 2022). I found that the 60 genes showing a lack of expression correlation are enriched within these windows (Fig. 4-9B), with 78.2% of these genes belonging to a genomic window identified as introgressed or deleted in 30 or more accessions. In contrast, just 12.3% of the rest of the genes in the genome are found within a genomic window identified as introgressed or deleted in 30 or more accessions.

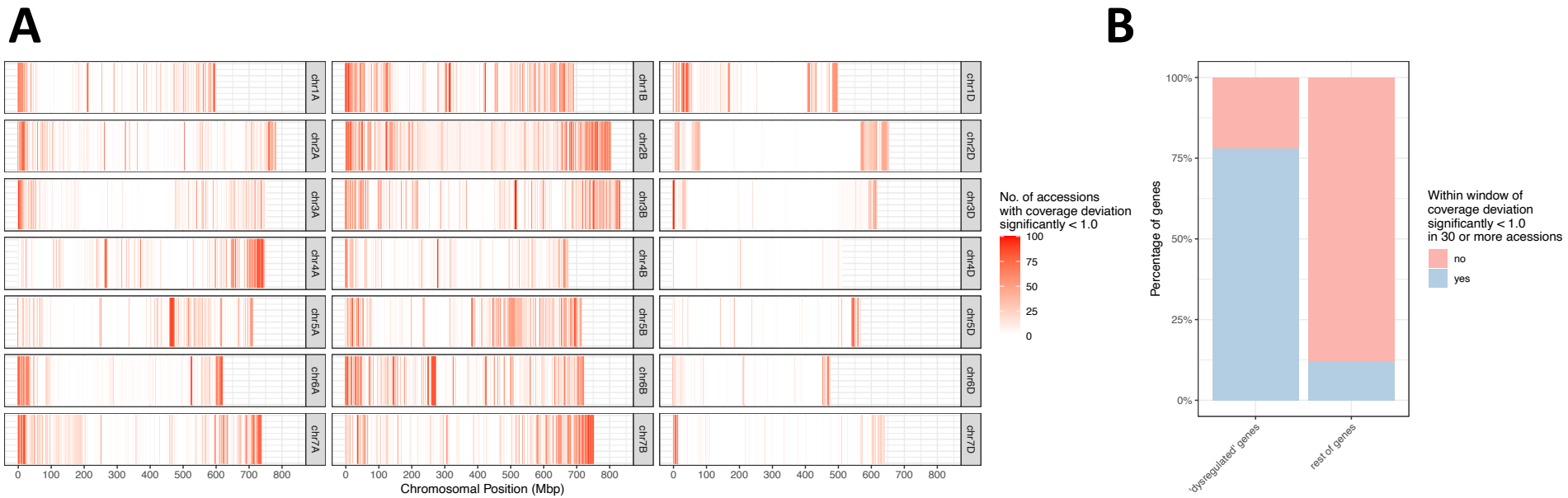

A
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D
0 100 200 300 400 500 600 700 800
Chromosomal Position (Mbp)
No. of accessions
with coverage deviation
significantly < 1.0
100
75
50
25
0
B
Percentage of genes
100%
75%
50%
25%
0%
'dysregulated' genes
rest of genes
Within window of
coverage deviation
significantly < 1.0
in 30 or more acessions
no
yes

Figure 4-9. Enrichment of genes showing a lack of expression correlation in He et al. (2022) in regions of divergence.
A) Chromosomal distribution of the number of accessions in each 1 Mbp genomic window which had mapping coverage deviation significantly less than 1 and are thus likely to contain divergent introgressed material or be deleted. **B)** The proportion of genes from the set of 60 genes showing a lack of expression correlation ('dysregulated' genes) identified by He et al. (2022) and the proportion of genes in the rest of the genome that are present in genomic windows identified as introgressed or deleted in 30 or more accessions.

To investigate the impact using the pantranscriptome reference has on the estimated expression of genes within this dataset, I pseudoaligned the leaf RNA-seq data from the wheat accessions to both the Chinese Spring and the pantranscriptome reference. Despite STAR outperforming kallisto for cross-mapping RNA-seq reads, I used Kallisto for aligning to Chinese Spring here instead of STAR to ensure consistency with the analysis performed by He *et al.* (2022).

Among the 60 genes showing a lack of expression correlation, 43/60 (71.7%) have, in 25 or more accessions, an estimated expression less than half when using the Chinese Spring reference compared to when using the pantranscriptome reference (Fig. 4-10). These are likely introgressed genes whose expression is underestimated when using Chinese Spring as the reference. Additionally, 6/60 (10.0%) of the genes have, in 25 or more accessions, an estimated expression more than double when using the Chinese Spring reference compared to when using the pantranscriptome reference (Fig. 4-10). This could be due to incorrect assignment of RNA-seq reads to a gene when using the Chinese Spring reference, which gets corrected when using the pantranscriptome reference as those reads now have a more appropriate target for assignment, resulting in fewer reads assigned to the first gene.

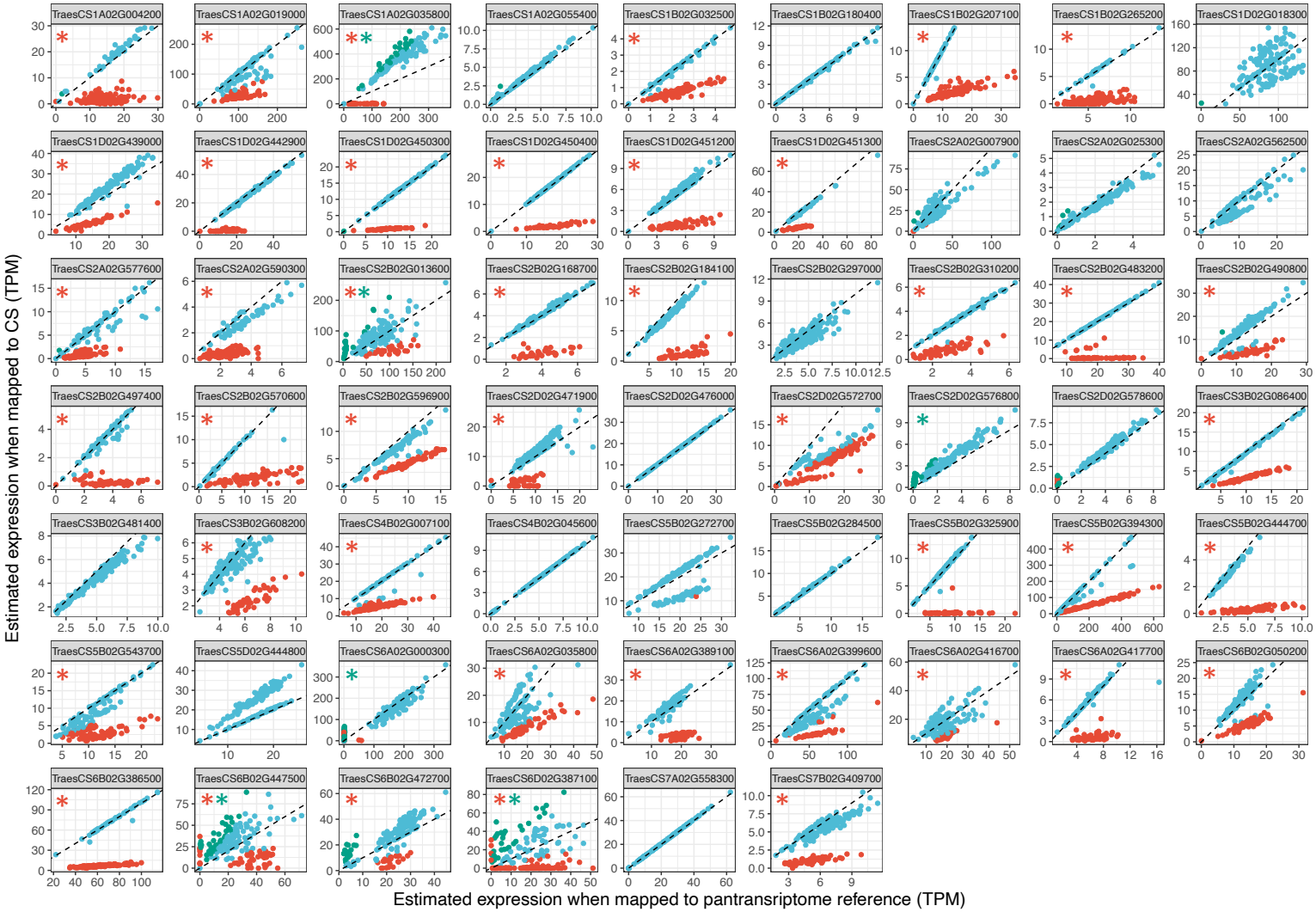

TraesCS1A02G004200
TraesCS1A02G019000
TraesCS1A02G035800
TraesCS1A02G055400
TraesCS1B02G032500
TraesCS1B02G180400
TraesCS1B02G207100
TraesCS1B02G265200
TraesCS1D02G018300
TraesCS1D02G439000
TraesCS1D02G442900
TraesCS1D02G450300
TraesCS1D02G450400
TraesCS1D02G451200
TraesCS1D02G451300
TraesCS2A02G007900
TraesCS2A02G025300
TraesCS2A02G562500
TraesCS2A02G577600
TraesCS2A02G590300
TraesCS2B02G013600
TraesCS2B02G168700
TraesCS2B02G184100
TraesCS2B02G297000
TraesCS2B02G310200
TraesCS2B02G483200
TraesCS2B02G490800
TraesCS2B02G497400
TraesCS2B02G570600
TraesCS2B02G596900
TraesCS2D02G471900
TraesCS2D02G476000
TraesCS2D02G572700
TraesCS2D02G576800
TraesCS2D02G578600
TraesCS3B02G086400
TraesCS3B02G481400
TraesCS3B02G608200
TraesCS4B02G007100
TraesCS4B02G045600
TraesCS5B02G272700
TraesCS5B02G284500
TraesCS5B02G325900
TraesCS5B02G394300
TraesCS5B02G444700
TraesCS5B02G543700
TraesCS5D02G444800
TraesCS6A02G000300
TraesCS6A02G035800
TraesCS6A02G389100
TraesCS6A02G399600
TraesCS6A02G416700
TraesCS6A02G417700
TraesCS6B02G050200
TraesCS6B02G386500
TraesCS6B02G447500
TraesCS6B02G472700
TraesCS6D02G387100
TraesCS7A02G558300
TraesCS7B02G409700
Estimated expression when mapped to CS (TPM)
Estimated expression when mapped to pantransriptome reference (TPM)
Estimated expression in accession higher using pantranscriptome reference
Estimated expression in accession the same using pantranscriptome reference
Estimated expression in accession lower using pantranscriptome reference

Figure 4-10. Estimated expression of the 60 genes showing a lack of expression correlation in He et al. (2022), using either the Chinese Spring RefSeq v1.1 transcriptome (y axis) or the pantranscriptome reference (x axis) as targets for kallisto pseudoalignment.
The dashed black line represents x=y, which is the expected value if the reference is not affecting the estimation of gene expression. An accession lying on this dashed line has this gene's expression estimated the same when using each reference. Red dots and green dots represent accessions in which a given gene has a TPM value <50% or >150%, respectively, when using the Chinese Spring reference than when using the pantranscriptome reference. A red star indicates that in 25 or more accessions, the gene has an estimated expression less than half when using the Chinese Spring reference compared to when using the pantransctipome reference. A green star indicates that in 25 or more accessions, the gene has an estimated expression more than double when using the Chinese Spring reference compared to when using the pantranscriptome reference.

While this analysis shows that using the Chinese Spring reference leads to the underestimation of expression of many of these genes, it is also important to explore the impact this underestimation has on correlation scores between homoeologues, as this is the metric used by He et al. (2022) to classify the genes as lacking expression correlation. I found that the Spearman's correlation coefficient (SCC) score between homoeologues from this set was -0.0990 when using the Chinese Spring reference and 0.407 using the pantranscriptome reference (Fig. 4-11). This difference was significantly different (p-value < 2.2e-16; 95% confidence interval ranges from -0.603 to -0.410). This suggests that the reference bias affecting many of these genes led to artificially low estimates of correlation between homoeologues in He et al. (2022).

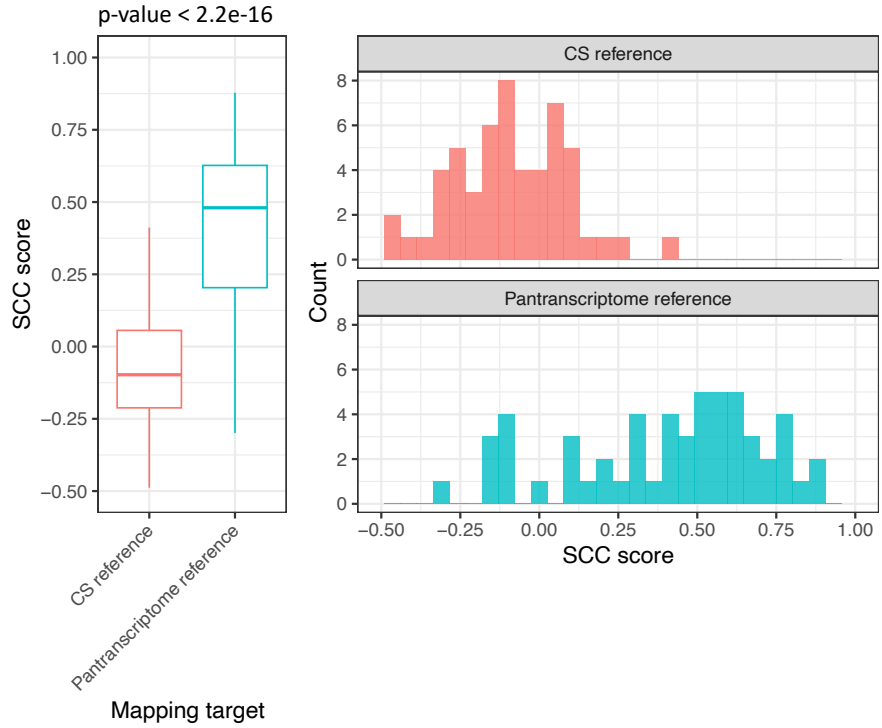


Figure 4-11. Spearman's correlation coefficient (SCC) between homoeologue pairs where one was identified as lacking expression correlation by He et al. (2022). SCC scores were computed between homoeologue pairs where one homoeologue is in the set of genes showing a lack of expression correlation identified by He et al. (2022). SCC scores were computed between AB, AD and BD homoeologue pairs and the lowest score was used. Triads in which any of the homoeologues were not present in the RefSeq v1.1 high-confidence gene annotation were excluded. The significance of the difference between SCC scores when using the Chinese Spring reference compared to when using the pantranscriptome reference was calculated using a two-tailed t-test with no assumption of equal variance.

4.3.4.2 Exploring estimated gene expression in the chr1D introgression

Several regions with poor mapping coverage (mapping coverage deviation significantly below 1) in multiple accessions analysed by He et al. (2022) overlap precisely with previously identified introgressions from cultivars for which chromosome-level genome assemblies were generated during the 10+ wheat genomes project (Walkowiak *et al.*, 2020). One such introgression is found at the end of chr1D (484,302,410bp-495,453,186 bp, based on RefSeq v1.0 coordinates), which I found to be present unbroken in 53/198 (26.8%) accessions (Appendix C1) and also in cultivars Jagger and Cadenza from the wheat pangenome (Fig. 4-12A).

I chose this introgression to study in more detail for several reasons. Firstly, its size is invariable between accessions; therefore, all accessions possessing the introgression have

the same genes introgressed. Additionally, this region featured prominently in He et al. (2022) as it contains 6 of the 60 genes showing a lack of expression correlation, including *ELF3-D1*, which was used as a case example due to its role in heading date (Wang *et al.*, 2016), an important agricultural trait.

In their paper, He *et al.* (2022) suggest this introgression is a terminal deletion. However, Wittern *et al.* (2023), identified that this region, including *ELF3-D1*, is in fact an introgression present in Cadenza and Jagger. They deduced that the donor of this introgression was either *T. timopheevii* or *Ae. speltoides*, based on the *ELF3-D1* gene model in Jagger sharing an intronic deletion with both of these species.

I explored this further to narrow down the potential donor species. I compared the proteins in the Jagger introgression with the *Ae. speltoides* proteins and found that the median protein identity between orthologues was only 91.6%. This strongly suggests that *Ae. speltoides* is not the donor species for this introgression. At the time of writing, there wasn't a genome assembly of *T. timopheevii* available so I could not perform the same analysis I performed for *Ae. speltoides*. However, I mapped *T. timopheevii* Illumina paired-end reads to the Jagger genome assembly and found that the read mapping was dense with an even coverage across the chr1D introgression (Fig. 4-13), suggesting that *T. timopheevii* is still a likely donor. As we can't be certain about the donor species, I will refer to this introgression as the chr1D introgression.

To evaluate how changing the reference changes the expression quantification of the introgressed genes, I compared the mean expression of genes from the chr1D introgression across accessions possessing the introgression to their 1-to-1 wheat orthologue across the accessions lacking the introgression. When using the Chinese Spring reference, the introgressed genes appear to be less expressed than their wheat orthologues; however, when using the pantranscriptome reference, no significant difference in expression was found between the genes (Fig. 4-12B, Appendix C2).

This reveals two important findings. Firstly, it supports the earlier work in this chapter showing that the expression of introgressed genes is underestimated when using the Chinese Spring reference. Secondly, it shows that introgressed genes, at least in this instance, are not expressed differently from the wheat orthologues they have replaced.

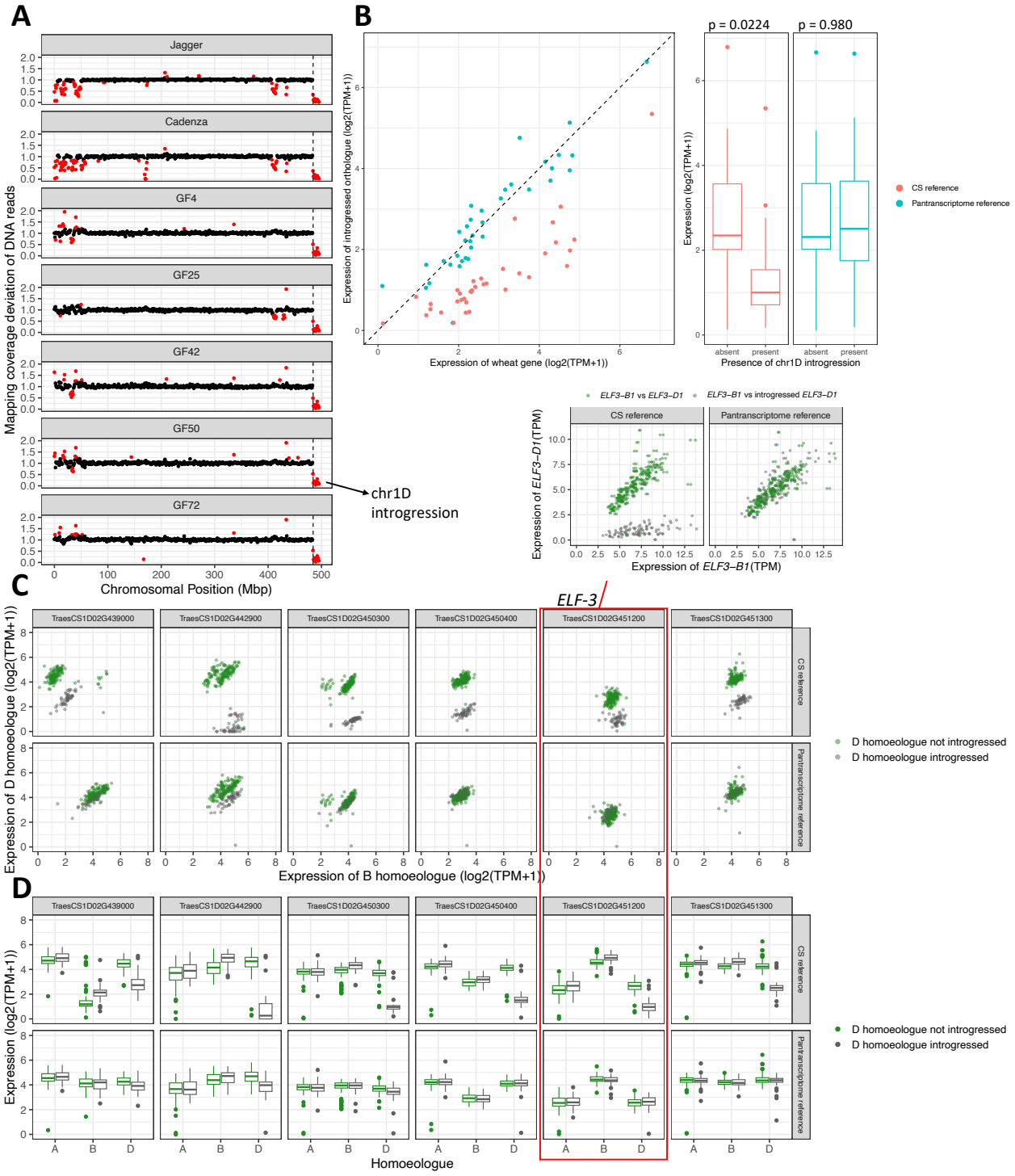

A
Jagger
Cadenza
GF4
GF25
GF42
GF50
GF72
Mapping coverage deviation of DNA reads
Chromosomal Position (Mbp)
chr1D introgression
B
Expression of introgressed orthologue (log2(TPM+1))
Expression of wheat gene (log2(TPM+1))
p = 0.0224
p = 0.980
Expression (log2(TPM+1))
absent
present
Presence of chr1D introgression
CS reference
Pantranscriptome reference
ELF3–B1 vs ELF3–D1
ELF3–B1 vs introgressed ELF3–D1
Expression of ELF3–D1 (TPM)
Expression of ELF3–B1 (TPM)
C
ELF-3
TraesCS1D02G439000
TraesCS1D02G442900
TraesCS1D02G450300
TraesCS1D02G450400
TraesCS1D02G451200
TraesCS1D02G451300
Expression of D homoeologue (log2(TPM+1))
Expression of B homoeologue (log2(TPM+1))
D homoeologue not introgressed
D homoeologue introgressed
D
Expression (log2(TPM+1))
Homoeologue
A
B
D

Figure 4-12. Introgressed genes falsely identified as less expressed due to reference bias.

A) Mapping coverage deviation of DNA reads across chr1D of Jagger, Cadenza, and 5 accessions from He et al. (2022). Each point is the coverage deviation value for a given 1 Mbp genomic window. Windows with a normalised coverage score significantly different to the median normalised coverage score for that window across the set of lines being compared are coloured red. Coverage deviation values significantly below one indicates an introgression is present or a deletion has taken place relative to the median of the rest of the set of lines. Coverage deviation values and significance values were calculated separately for the accessions and for the cultivars Jagger and Cadenza, the latter two being compared to mapping coverage values from the other cultivars whose genome was assembled in Walkowiak *et al.* 2020). The reduced coverage at the end of chr1D, the left-hand border of which is indicated by the vertical dashed black line, is an introgression common to 53 of the 198 accessions and cultivars Jagger and Cadenza. **B)** Expression of the wheat gene compared to its introgressed orthologue from the chr1D introgression, using either the Chinese Spring v1.1 transcriptome or the pantranscriptome reference as the target for kallisto pseudoalignment. Orthologue pairs with TPM < 1 in both the introgressed and the wheat gene when mapping to the pantranscriptome reference were excluded. The significance of the difference between introgressed and non-introgressed orthologues when using the Chinese Spring or the pantranscriptome reference was calculated using two-tailed t tests with no assumption of equal variance **C)** Estimated expression level of introgressed D homoeologues compared to the wheat B homoeologues and wheat D homoeologues compared to wheat B homoeologues, using either the Chinese Spring v1.1 transcriptome or the pantranscriptome reference as the target for kallisto pseudoalignment. Each point represents one accession. **D)** Expression level of triads in which the D homoeologue is an introgressed gene in a subset of lines, using either Chinese Spring or the pantranscriptome reference as the target for kallisto pseudoalignment. The centre line of the boxplots = the median; the box limits = the upper and lower quartiles; the whiskers = 1.5x interquartile range; and the points = the outliers.

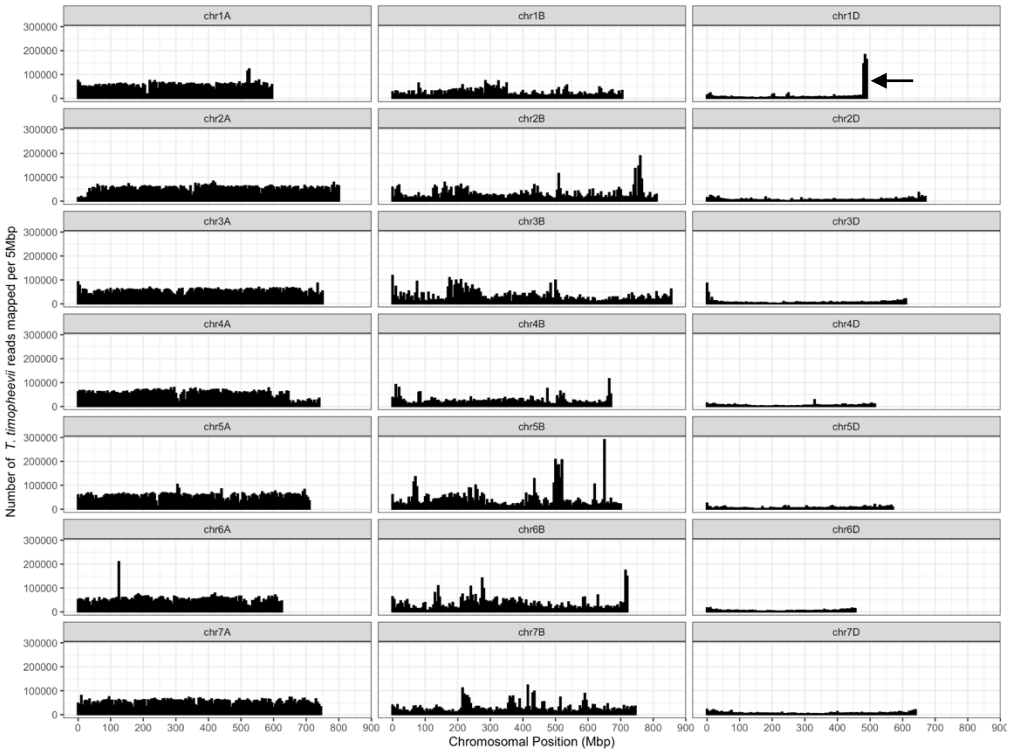


Figure 4-13. Reads from *T. timopheevii* accession P95 mapped to *T. aestivum* cv. Jagger (which contains the chr1D introgression) and binned into 5 Mbp genomic windows. The number of mapped reads was divided by the length of window to accurately reflect read density at the final window of each chromosome. The chr1D introgression with a putative origin of *T. timopheevii* is at 481585620-493450010 and is indicated by the black arrow. *T. timopheevii* is a tetraploid with genomes related to the A and B subgenomes of wheat. This is reflected in the greater mappability of *T. timopheevii* reads to the A and B subgenomes than the D subgenome. However, the higher read count across the 1D introgression suggests this region is more similar between *T. timopheevii* and the introgression than between *T. timopheevii* and the A and B subgenomes, lending support to the donor of this introgression being *T. timopheevii*.

Earlier in this chapter, I used simulated data to show how reference bias can lead to the incorrect assignment of expression balance across triads. To explore this phenomenon using experimental data and on a few specific triads, I explored the estimated expression across triads within the chr1D introgression that are also in the set of genes showing a lack of expression correlation from He et al. (2022). When the RNA-seq reads are pseudoaligned to Chinese Spring, in accessions possessing the chr1D introgression, *ELF3-D1* appears to have low expression and *ELF3-B1* appears to have slightly elevated levels of expression compared with accessions without the chr1D introgression (Figs. 4-12c, 4-12d). However, when using the pantranscriptome reference, in accessions possessing the chr1D introgression, the expression of *ELF3-D1* and *ELF3-B1* in accessions with the chr1D introgression appears very similar to the expression of *ELF3-D1* and *ELF3-B1* in accessions

without the chr1D introgression (Figs. 4-12c, 4-12d). The CDS sequence for the introgressed copy of *ELF3-D1* shares 97% sequence identity with *ELF3-D1* in Chinese Spring, 97.56% identity with *ELF3-A1* and 97.8% identity with *ELF3-B1*. The high divergence of *ELF3-D1* from the introgression and *ELF3-D1* from Chinese Spring and the greater similarity between *ELF3-D1* from the introgression with *ELF3-B1* from Chinese Spring explains how most reads were unable to be assigned, yet some were incorrectly assigned to *ELF3-B1*. This resulted in the slight increase in estimated expression of *ELF3-B1* when using the Chinese Spring reference. The five other genes showing a lack of expression correlation within the chr1D introgression show reduced homoeologue imbalance using the pantranscriptome reference and expression levels in line with triads in which the D homoeologue has not been introgressed. Four of these genes also show a slight decrease in estimated expression in their B homoeologue when mapping to the pantranscriptome reference, supporting the idea that incorrect assignment of reads from the introgressed gene to its homoeologue, in addition to reads not being assigned to the introgressed homoeologue, will have contributed to the artificially low correlation scores observed by He et al. (2022).

My colleague, Hannah Rees, conducted a study analysing an RNA-seq timecourse dataset from the wheat cultivar Cadenza in order to understand the transcriptional regulation of circadian clock genes in wheat. As Cadenza contains the chr1D introgression which included notable clock genes, *TaELF3-1D* and *TaSIG3-1D*, I used the timecourse dataset to examine whether the chr1D introgression was introducing reference bias that was altering the results of her study.

This analysis was conducted before I developed the pantranscriptome method. Therefore, to correct the estimated expression in this case, I concatenated the chr1D introgression from Jagger to the Chinese Spring RefSeq v1.0 reference genome, as it is the same introgression found in Cadenza, but Cadenza only had a scaffold-level, rather than a chromosome-level, genome assembly available. After aligning the RNA-seq reads to this concatenated reference using HISAT2, expression counts and TPM values for the 1-to-1 Chinese Spring and Jagger orthologues were summed.

When mapping to Chinese Spring, *TaELF3-1D* and *TaSIG3-1D* appear very lowly expressed, and they were not classified as rhythmically expressed. However, after aligning the reads to the reference including the chr1D introgression, the expression of both genes was

much higher and the rhythmicity could be properly assessed, revealing that *TaSIG3-1D* was rhythmically expressed like its homoeologues and *TaELF3-1D* was not rhythmically expressed, also like its homoeologues (Fig. 4-14).

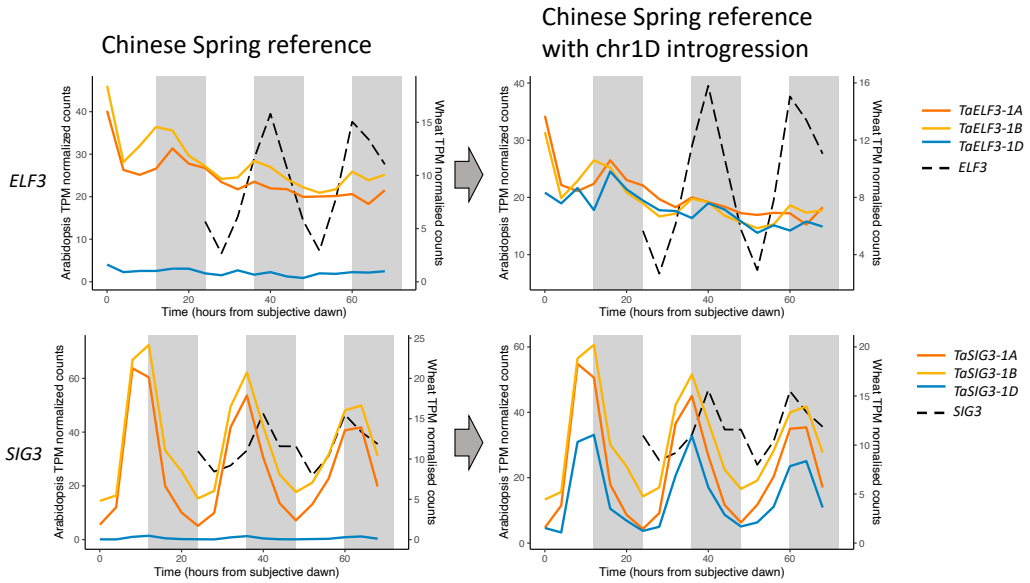


Figure 4-14. The impact of reference bias on the estimation of rhythmicity within the *ELF3* and *SIG3* triads.
Estimated expression level over time of the *Elf3* and *SIG3* triads in Cadenza when using the Chinese Spring reference or a combined reference which includes the chr1D introgression sequence. The chr1D introgression is found in Cadenza and includes *TaElf3-1D* and *TaSIG3-1D*. Wheat homoeologues are coloured according to their identity to either the A genome (orange), B genome (yellow), or D genome (blue) and grey and white blocks indicate subjective dark and light time periods under constant conditions. Data represent the mean of 3 biological replicates and transcript expression is collapsed to the gene level. The dashed black line represents the expression of the *Arabidopsis* orthologue of *Elf3* and *SIG3* from a circadian timecourse dataset generated by Romanowski *et al.* (2020).

The introgressed *TaELF3-1D* allele has been linked with a QTL for heading date (Wittern *et al.*, 2023). No significant difference was found between the mean expression level of the three *TaELF3* homoeologues when reads were aligned to the combined reference of Chinese Spring and Jagger ($F(2, 51) = 2.005$, $p = 0.145$, 1-way ANOVA). Therefore, any heading date phenotype conferred by this allele is likely to be due to altered protein function of the introgressed gene rather than expression-level differences.

4.4 Discussion

In the emerging era of plant pangenomics, chromosome-level assemblies are being generated for an increasing number of cultivars/accessions, which will facilitate a shift away from reference genome-centric methods. Here, I have demonstrated the importance of utilising these resources effectively for RNA-seq analyses in wheat to reduce reference bias.

4.4.1 RNA-seq reference bias in wheat

The quantification of gene expression from RNA-seq data in wheat is very accurate when the reference is of the same accession/cultivar as the sample. However, when the sample accessions differ to that used to generate the reference genome, a noticeable level of reference bias occurs. This bias affected both the quantification of individual genes and the correct assignment of triads to categories of homoeologue expression balance.

A primary culprit behind this reference bias is the presence of introgressions from wheat's wild and domesticated relatives, which possess blocks of genes highly divergent from wheat, resulting in challenges in the correct assignment of RNA-seq reads. Due to the severity of the reference bias observed here, these introgressed regions are effectively rendered inaccessible to any meaningful form of analysis and conclusions. However, they have hitherto been included in analyses and downstream conclusions.

In this work, kallisto outperformed STAR for self-mapping but when cross-mapping, STAR was better able to deal with divergence between genes, leading to more accurate quantification of gene expression. Similar limitations of alignment-free methods have been previously discussed; for example, Wu *et al.*, (2018) demonstrated that kallisto performs poorly for lowly expressed genes and for RNA reads with biological variation compared to the reference. Despite dealing with reference bias better than kallisto, STAR was unable to resolve the issue of reference bias.

It could be argued that relaxing alignment parameters would enable reads from divergent genes to be able to be assigned to the reference. However, this will likely undermine the accuracy of read assignment in other parts of the genome and will probably cause more incorrectly assigned reads despite decreasing unassigned reads. In particular, this would take place among homoeologues, as the introgressed gene is not always more genetically similar to the gene it replaced than to the replaced gene's homoeologues.

It could be argued that reference bias may not significantly affect differential expression analyses between conditions or across tissues within a single genotype as the ratio of estimated expression between conditions/tissues should remain similar regardless of reference bias. This may be true; however, this idea should be formally tested. At the very least, genes affected by reference bias so severely that their expression counts are close to zero may be ineligible for inclusion in differential expression calls, which could lead to biologically important genes being excluded from the analysis.

If interested in homoeologue expression balance, unequal divergence of homoeologues relative to the reference will lead to incorrect findings. Reference bias can also obscure complex expression patterns. For instance, in the Cadenza circadian clock timecourse dataset, I showed that *TaELF3-1D* and *TaSIG3-1D*, which are within an introgression in Cadenza, have very low estimates of expression when using the Chinese Spring reference, and the rhythmicity is difficult to ascertain. However, when using the combined reference, more reads aligned, the expression levels were raised to levels not significantly different to their A and B homoeologues, and the rhythmicity of the genes could be accurately assessed.

As genome assemblies for more wheat accessions become available, matching a sample to a more suitable reference genome will become increasingly feasible. However, in situations involving multiple accessions where a common reference genome is needed, or when the appropriate genome assembly is unavailable for within-accession analyses, it is crucial to exercise caution and investigate whether introgressed genes could be affecting the results. Going forwards, in addition to exercising caution, it is important to develop novel methodologies to address the issue of introgression-induced reference bias.

4.4.2 Using a pantranscriptome reference to reduce reference bias

Prior studies have highlighted the benefits brought by using enhanced references or individualised references for RNA-seq alignment. For instance, Vijaya Satya, Zavaljevski and Reifman (2012) constructed an enhanced reference genome for humans by incorporating alternative allele segments at known polymorphic loci. Other researchers have reported using individualised genomes or transcriptome references by updating them with SNPs, INDELs and splice site information for each individual (Munger *et al.*, 2014; Liu, MacLeod and Liu, 2018). Munger et al. (2014), when working with a multi-parent mouse population, found that mapping to individualised genomes substantially

increased the accuracy of eQTL assignment from 88.2% to 98.3% and corrected false-positive linkage signals. Kaminow *et al.* 2022) constructed a pan-human consensus genome by calculating the consensus allele for each variant, which considerably improved the accuracy of RNA-seq mapping compared to when the reference genome was used.

My approach follows in the vein of these pieces of research. However, individualised genomes or consensus genomes are not suitable for wheat as the extensive divergence introduced by introgressions makes the accurate genotyping that is necessary for constructing individualised or consensus genomes challenging. Instead, I constructed a pantranscriptome reference that incorporates transcripts from other wheat cultivars into the Chinese Spring reference transcriptome, provided they come from genes in a 1-to-1 orthologous relationship with a Chinese Spring gene. The low computational requirements of kallisto regardless of reference size allows for scalability as more genome and transcriptome data become available, while still running in a fraction of the time taken by alignment-based tools to align to a single reference genome.

The pantranscriptome reference corrects nearly all incorrectly underestimated genes that belong to an introgression present in the assembled pangenome cultivars and in a 1-to-1 relationship with a Chinese Spring gene. However, this approach currently has limitations. The pantranscriptome reference will not have captured all the genetic variation present in wheat germplasm around the world. In particular, it mainly includes transcripts from cultivars from Western countries and, except for Chinese Spring, doesn't include transcripts from Asian or African wheat accessions. There are several other publicly available, high-quality wheat genome assemblies whose transcripts could have been incorporated into the pantranscriptome (Guo *et al.*, 2020; Athiyannan *et al.*, 2022; Shi *et al.*, 2022; Jia *et al.*, 2023). However, I opted to only use genomes that were annotated using the same methodology to ensure accurate orthologue assignment.

As additional genomes and/or transcriptomes are sequenced and other existing genomes are re-annotated to provide consistent gene annotations, the pantranscriptome reference can be expanded to encompass a broader range of genetic variation. This may eventually reach a saturation point where most commonly segregating variation is covered within the reference. Another limitation of the pantranscriptome reference is its inability to correct reference bias caused by copy number variation such as tandem duplications or presence/absence variation. Instead, it is limited to correcting reference

bias caused by divergent genes. This is because, to ensure additional errors were not introduced, I elected to only add transcripts from other cultivars to the pantranscriptome reference if they came from genes in a 1-to-1 orthologous relationship with a Chinese Spring gene. This results in most of the genes whose expression is overestimated when using Chinese Spring as the reference remaining overestimated when using the pantranscriptome reference. While overcoming this limitation is important, doing so is challenging as it involves resolving intricate orthologue and paralogue relationships, and it remains unclear how novel genes and genes with varying copy numbers between cultivars should be represented in the pantranscriptome reference.

Entirely different and superior solutions to the problem of RNA-seq reference bias in wheat may emerge in the future. For instance, the field of graph genomes is rapidly developing (Garrison *et al.*, 2018), including methods to align RNA-seq reads to graph genomes (Sibbesen *et al.*, 2022). Martiniano *et al.* (2020) used a sequence variation graph containing human variants from the 1000 Genome Project, reducing reference bias by creating a balanced representation of alleles of polymorphic sites. However, successfully creating graphs for genomes as large and complex as wheat remains a major challenge. It is also a much heavier-weight solution compared to the pantranscriptome pseudoalignment approach. At the very least, my approach offers a temporary way to improve the accuracy of RNA-seq alignment and explore the impact of reference bias, particularly for genes comprising the core genome. Following further development and the incorporation of new data, it may evolve into a long-term alternative, more lightweight approach to emerging graph-based methods.

4.4.3 Examining reference bias in experimentally-generated RNA-seq data

By utilising the valuable dataset generated by He et al. (2022), I demonstrated the presence of reference bias in experimentally-generated datasets as well as in simulated datasets. The diverse nature of wheat accessions sequenced by He et al. (2022) may have rendered this study particularly susceptible to the effects of reference bias. Indeed, my analysis indicated that regions displaying divergence were very prevalent across these accessions. However, the prevalence of introgressions in this dataset may be typical for collections of wheat accessions as introgressions are a common feature in most wheat germplasm, including commercially distributed Elite cultivars. Wheat accessions containing diverse introgressions hold significant importance in wheat research, as they

can serve as valuable sources of genetic variation for breeders, not to mention the insights they can provide into the evolutionary dynamics of wheat genomes. Therefore, the ability to accurately study them is important.

Among the 60 genes showing a lack of expression correlation identified by He et al. (2022), 78.2% were found to be enriched in genomic regions identified as introgressed or deleted in 30 or more accessions. Furthermore, I demonstrated that most of these genes exhibited much higher levels of estimated expression when using the pantranscriptome reference as opposed to the Chinese Spring reference. Additionally, the use of the pantranscriptome reference led to increased correlation scores between homoeologue pairs. As the pantranscriptome reference likely doesn't contain all the introgressions present in the accessions reanalysed, it is possible that the impact of reference bias has been underestimated here.

These findings may necessitate the revision of the explanation as to why these genes were associated with variation in important productivity traits. While some of these triads may still demonstrate genuine dysregulation of homoeologues and dosage effects, it appears likely that for many of these genes, variation in the gene sequence itself underlies the observed trait variation as opposed to changes in expression dosage among homoeologues. It appears that He et al. (2022) have inadvertently observed the correlation of introgressed genes with agronomic productivity traits, adding evidence that introgressions present in wheat accessions are important drivers of trait variation. This finding also has implications for our understanding of the evolutionary and selection mechanisms implicated in the control of these traits.

4.4.4 chr1D introgression and Elf3

To gain a more precise understanding of how the quantification of introgressed genes is influenced by the choice of reference, I conducted an analysis focusing on genes located within the chr1D introgression. This was selected due to its presence in approximately a quarter of the accessions and a lack of variation in size across the accessions possessing it, due to an absence of recombination within the introgression. Additionally, this introgression contained 6 of the 60 genes showing a lack of expression correlation identified by He et al. (2022). I showed that when using the Chinese Spring reference, the introgressed genes appear to be less expressed than the wheat orthologue they replaced. However, when the pantranscriptome reference was used, which includes the transcripts

from the introgressed genes, there was no significant difference in estimated expression between these genes.

Furthermore, the correction of the quantification of these genes had a notable impact on the estimated expression balance across triads in which the D homoeologue was introgressed. This correction resulted in an increase in the estimated expression of the D homoeologue and in most cases, a slight decrease in the estimated expression of the B homoeologue, due to incorrectly assigned reads from the D homoeologue when using the Chinese Spring reference.

It would not have been surprising to see, even after removing reference bias, that introgressed genes were expressed differently than the wheat orthologue they replace, perhaps due to the divergence in regulatory sequences. However, this finding suggests that, at least for this introgression, this is not the case. These findings have implications for any RNA-seq studies using wheat accessions containing introgressions, and also more specifically for studies looking at the expression of introgressed genes and what mechanisms underlie the phenotype they confer.

4.4.5 Future work

The pantranscriptome reference as presented in this chapter offers an improvement over aligning RNA-seq reads from many diverse accessions to a single reference genome. However, as discussed in section 4.4.2, there is room for improving this method. First of all, it can be extended to incorporate transcripts from more accessions. To ensure accurate orthologue assignments are maintained, this may involve consolidating the annotations of other genomes, so they are comparable with the annotations of Chinese Spring and the cultivars from the 10+ wheat genomes project. Another important improvement will be to develop a way to reduce reference bias caused by copy number variation. It may be that this won't be able to be solved until graph-based methodology improves sufficiently to handle large genomes such as wheat.

While the work in this chapter is focused on wheat, similar issues may be encountered when working on other species with a polyploid genome and/or many introgressions. Therefore, similar analyses on other species could offer valuable insights for their respective research communities.

4.5 Methods

4.5.1 Read simulation, alignment, and quantification

Reads were simulated from the longest transcript from each high-confidence gene in the Chinese RefSeq v1.1 annotation and the nine pseudomolecule genome assemblies (White *et al.*, 2024) if the transcript length >= 500bp. 1000 pairs of 150 bp reads with an insert size of 400bp and no errors were simulated for each transcript using wgsim from samtools v1.9 (Li *et al.*, 2009).

The kallisto index was produced from the CDS sequences from the RefSeq v1.1 high-confidence gene annotations using kallisto v0.44.0 (Bray *et al.*, 2016). Reads were pseudoaligned to this index using the default settings and 100 bootstraps. Read counts and TPMs were summed across transcripts to generate gene level counts.

To construct the pantranscriptome reference, my collaborator Thomas Lux ran Orthofinder (Emms and Kelly, 2019) with standard parameters to define orthogroups based on the longest isoform protein sequences of the high-confidence genes from Chinese Spring and the 9 chromosome-level pangenome cultivars. If a gene was found in a 1-to-1 relationship with a Chinese Spring gene, I added its transcripts to the Chinese Spring RefSeq v1.1 high-confidence transcript FASTA file. A kallisto index was built and reads were pseudoaligned as above. Read counts and TPMs were each summed across all transcripts of a gene and its 1-to-1 orthologues to generate gene-level counts.

The STAR index was built for RefSeq v1.0 with the RefSeq v1.1 high-confidence gene annotation using STAR v2.7.6a (Dobin *et al.*, 2013) using default parameters except for --limitGenomeGenerateRAM 200000000000 and --genomeSAindexNbases 12. The simulated reads from the 10 cultivars were aligned to this index using STAR and the predicted splice junctions from all were merged and then filtered to remove non-canonical junctions, junctions supported by 2 or fewer uniquely mapping reads and reads already annotated in the original genome annotation. The index was rebuilt using these discovered splice sites in addition to the annotated splice sites. The simulated reads from the 10 cultivars were aligned to this new index with parameters --quantMode TranscriptomeSAM and --outSAMunmapped Within. Gene-level read counts were generated using RSEM v1.2.28 (Li and Dewey, 2011).

For comparisons between self-mapping and cross-mapping, the following criteria were used to determine whether a gene was present in the analysis. For self-mapping, Chinese Spring genes from which RNA-seq reads were simulated in Chinese Spring were included. For cross-mapping, for a Chinese Spring gene to be included it had to be in a 1-to-1 orthologous relationship with a gene from the cross-mapped cultivar, and RNA-seq reads must have been simulated from both the Chinese Spring gene and the orthologue from the cross-mapped cultivar.

4.5.2 Defining triad balance

Triads in Chinese Spring were taken from Ramirez-Gonzalez *et al*. (2018). For each cultivar, triads were retained if RNA-seq reads were simulated from all three homoeologues. Triad balance was computed in the same way as in (Ramírez-González *et al.*, 2018) except for the use of read counts rather than TPMs due to the way the reads were simulated. The relative read count of each homoeologue within a triad was calculated as follows:

$$A_{norm} = \frac{A}{A + B + D}$$

$$B_{norm} = \frac{B}{A + B + D}$$

$$D_{norm} = \frac{D}{A + B + D}$$

where A, B and D are the read counts of the A, B and D homoeologues, respectively. Euclidean distance was then used to calculate the distance between each set of normalised expression values across a triad and an ideal normalised read count bias for each of seven categories (Table 4-6). A triad was assigned to an expression bias category by selecting the category with the shortest Euclidean distance between the observed and the ideal bias.

Table 4-6. Ideal normalised read count bias for each triad expression category

Category	A	B	D
Balanced	**0.33**	**0.33**	**0.33**
A suppressed	**0**	**0**	**0**
B suppressed	**0.5**	**0.5**	**0.5**
D suppressed	**0.5**	**0.5**	**0.5**
A dominant	**1**	**0**	**0**
B dominant	**0**	**1**	**0**
D dominant	**0**	**0**	**1**

4.5.3 Binning incorrectly quantified genes

For each 5 Mbp genomic window in Chinese Spring RefSeq v1.0, a score was calculated based on the number of underestimated (read count < 500) and overestimated (read count > 1500) genes within that window:

$$(-1 * no.of\ underestimated\ genes) + (1 * no.overestimated\ genes)$$

4.5.4 Calculating CDS identity

BLASTn from blast+ v2.7.1 (Camacho *et al.*, 2009) was used to align the nucleotide sequences of the longest transcript of pairs of 1-to-1 orthologues between Lancer and Jagger. The identity of the best hit between pairs was binned into 5 Mbp genomic windows.

4.5.5 Processing sequencing data from He et al. (2022)

198 accessions had both leaf RNA-seq data and enrichment capture short paired-end DNA reads. The RNA-seq data from the 198 lines was pseudoaligned to both Chinese Spring RefSeq v1.1 and the pantranscriptome reference as above for the simulated reads. Accessions GF25, GF270, GF32, GF37, GF41 and GF73 were excluded for RNA-seq analyses as in He et al. (2022).

DNA reads were mapped to RefSeq v1.0 and to the pseudo introgression genome and filtered as above for the simulated DNA reads. Accessions GF294, GF342, GF366, GF380, GF381, GF383, GF38 were excluded for DNA analyses as in He et al. (2022).

4.5.6 Identifying chr1D introgression

For the pangenome cultivars, I simulated paired-end 150 bp reads with a 500bp insert size and no errors from all fourteen genome assemblies (ArinaLrFor, Cadenza, Claire, Jagger, Julius, Lancer, Landmark, Mace, Norin61, Paragon, Robigus, Stanley, SY Mattis, and Weebil) generated by Walkowiak *et al.* (2020) to a depth of 10x using WGSim within samtools v1.9 (Li *et al.*, 2009). These reads were mapped to Chinese Spring RefSeq v1.0 (Appels *et al*., 2018). The alignments were filtered using samtools v1.4 (Li et al., 2009): supplementary alignments, improperly paired reads, and non-uniquely mapped reads (mapping quality less than 10) were removed. Mapping coverage deviation and significance for these cultivars and for the accessions from He *et al*. (2022) were calculated as in section 3.5.7 in two separate analyses, one for the accessions from He *et al*. (2022) and one for the cultivars from Walkowiak et al. (2020).

4.5.7 Locating coordinates of introgression boundaries

To detect the precise locations of the chr1D, chr2A *Ae. ventricosa*, and chr2D *Ae. markgrafii* introgressions in Jagger; and the chr2B *T. timopheevii* and chr3D *Th. ponticum* introgressions in Lancer, I used the simulated read mappings for Jagger and Lancer from section 4.5.6. Read depths were binned into 5 Mbp and 1 Mbp windows using bedtools makewindows v2.28.0 (Quinlan and Hall, 2010) and hts-nim-tools v0.0.1 (Pedersen and Quinlan, 2018). The window in which read depth drops, signifying the start/end of the introgression, was identified for each introgression and IGV (Robinson *et al.*, 2011) was used to precisely identify the position where the coverage profile changed. To locate the location of the introgressions relative to the Jagger/Lancer genomes in order to identify which genes have been introgressed, I extracted the Chinese Spring sequences 1 Mbp either side of the precisely located border position (or until the end of the chromosome) for each introgression and aligned them to the Jagger or Lancer genome assembly using minimap2 (Li, 2018) with parameters -x asm5. These alignments were used to determine the borders of the introgressed region as they appear in Jagger or in Lancer.

4.5.8 Characterising the chr1D introgression donor species

Blastp from blast+ v2.7.1 (Camacho *et al*., 2009) was used to align the *Ae. speltoides* proteins with the longest isoforms of the Jagger proteins of genes found in the chr1D introgression. The best hit for each Jagger protein was kept. Paired-end Illumina DNA short reads from *T. timopheevii* (King *et al.*, 2022) were mapped to Chinese Spring RefSeq v1.0 using BWA mem v0.7.13 (Li and Durbin, 2009). Samtools v1.4 (Li *et al.*, 2009) was used to filter the alignments to retain mapped reads, primary alignments, properly paired reads and uniquely mapping reads (mapping quality > 10). PCR duplicates were found and removed using the Picard Tools v2.1.1 MarkDuplicates function (Depristo *et al.*, 2011). Read depths were binned into 5 Mbp windows using bedtools makewindows v2.28.0 (Quinlan and Hall, 2010) and hts-nim-tools v0.0.1 (Pedersen and Quinlan, 2018) and divided by window length to account for windows at ends of chromosomes which are less than 5 Mbp in length.

4.5.9 Calculating SCC scores between homoeologues

SCC scores for triads including the 60 genes identified as lacking expression correlation by He *et al*. (2022) were calculated between AB, AD, and BD homoeologue pairs using the cor.test function in R with the 'spearman' method; the lowest SCC value of the three comparisons was used. Triads were excluded if any of the homoeologues were not found in the RefSeq v1.1 high-confidence annotation.

4.5.10 Analysis of clock genes from Cadenza timecourse RNA-seq dataset

The details of plant growth and sampling can be found in Rees *et al.* (2022). Briefly, Cadenza was grown under 12hlight:12hdark conditions for circadian entrainment, followed by sampling under constant light conditions of all aerial tissue every 4 hours for 3 days, with 3 biological replicates per time point. mRNA was extracted and sequenced by Genomics Pipelines on a NovaSeq S2 flow cell by Genomics Pipelines at the Earlham Institute to generate 150 bp paired-end reads with an average sequencing depth of 85 million reads per replicate.

The reads were filtered and trimmed using Trimmomatic v0.30 (Bolger, Lohse and Usadel, 2014). Filtered and trimmed reads were aligned using HISAT2 v2.0.4 (Kim *et al.*, 2019) to Chinese Spring RefSeq v1.0, and to Chinese Spring RefSeq v1.0 with the chr1D introgression from Jagger concatenated as an additional chromosome. HISAT2 was used

for consistency with the rest of the analysis in Rees *et al.* (2022). Non-uniquely mapping reads were removed using samtools v1.3 and gene-level abundances were quantified using StringTie v2.1.4 (Pertea *et al.*, 2015). When reads were aligned to the combined reference, for Chinese Spring genes in the chr1D introgression region with a 1-to-1 orthologue in Jagger, TPM values were summed across the transcripts of the orthologue pairs. Gene-level TPM values were then averaged across the three biological replicates. Genes with 0 TPM at all time points were removed. To identify rhythmically expressed genes, the R package MetaCycle (Wu *et al.*, 2016) was run using the following parameters; minper = 12, maxper = 35, adjustPhase = "predictedPer." Transcripts were defined as rhythmic if they had *q*-values < 0.05 and high confidence rhythmic transcripts if they have *q*-values < 0.01.

5 Conclusions and outlook

Introgressions from wheat's wild and domesticated relatives are important sources of novel genetic variation and have played an important role in the evolution of wheat. They are now being more deliberately introduced and utilised to assist in the effort to continually drive improvements to wheat varieties. The rapid development of sequencing technologies has provided important tools for studying introgressions. It has also revealed how abundant introgressions are across wheat material, leading to the possibility that failing to account for them may be negatively impacting genomic analyses. This thesis has demonstrated the value of sequencing data for identifying and characterising introgressions in wheat and has shown how introgressions lead to reference bias that negatively impacts the processing of sequencing data in common genomic analyses. For each of these, I have provided methods and ideas that other researchers can build upon in the future and apply to their own research.

Chapter two presents, to the best of my knowledge, the highest resolution identification of introgression junctions to date. The high-resolution allowed crossover locations to be visualised within gene bodies, uncovered small segments previously missed by lower resolution methods, and showed small differences between overlapping segments between lines with different rust resistance phenotypes. The characterisation of these introgression lines will be useful to those using those specific lines and the methodology adds to several other pieces of work, giving researchers more ideas to draw from when looking to characterise sets of introgression lines in the future. It also provides ideas for how one can investigate introgression lines in conjunction with phenotyping data, which can be improved upon as chromosome-level genome assemblies become available for more wild relatives.

In chapter three, I reported the discovery of three MTAs, which together increase yield under heat stress by more than 50%. I uncovered an *Ae. tauschii* introgression underlying an MTA associated with heat tolerance and took advantage of recombination that has taken place within the introgression to narrow down which region is underlying the phenotype. Looking within the region in *Ae. tauschii* genomes assemblies showed differences in gene content and order in this region between wheat and *Ae. tauschii*. This finding highlights the limitations of relying on reference genomes in studies like this. This is particularly relevant when divergent material is present in the sampled lines, which

increases the likelihood of presence/absence variation or gene order differences between the samples and the reference genome.

The heat tolerance trait itself is potentially of very high value, due to the large effect size of the three MTAs and the importance of climate tolerance traits to future food security in the face of climate change. It also demonstrates how loci of large effect can underlie traits that might typically be considered to be complex and driven only by many small-effect loci, and thus difficult to unpick using genome-wide association study. Further work will be required to validate the phenotype in different environments, after which the alleles can be incorporated into breeding programmes.

The dramatic impact of these introgressions on read mapping led to the idea that when the goal is not identifying introgressions, the presence of introgressions in the samples being studied should negatively impact the analysis by reducing read mapping rates and introducing read mapping errors. This led to the development of the work presented in chapter four, where I demonstrated the large impact of introgressions on the accuracy of RNA-seq quantification in wheat and presented a method to partially reduce the reference bias caused by these introgressions. By reanalysing gene expression data from diverse wheat accessions generated by He et al. (2022), I demonstrated that some of their findings may have been falsely caused by reference bias. Instead of expression bias across homoeologues driving trait variation, it seems that the appearance of this homoeologue expression bias was at least partly caused by reference bias caused by introgressions. This suggests that the observed trait variation is probably driven by the introgressed alleles themselves.

The work presented in chapter four will likely impact future RNA-seq studies in wheat and may lead to the revision of previous findings. Tools to utilise new genomes will develop rapidly over the coming years and will reduce the issues that introgressions cause to the accuracy of read mapping. My work brings awareness to the issue now to prevent errors being proliferated in the literature, even if only by demonstrating the potential for errors and advising caution to be taken by researchers. While currently useful, my solution of using a pantranscriptome kallisto reference to reduce reference bias is partial in the type of errors remedied and incomplete in the variation currently captured in the available genome assemblies. I look forward in anticipation to how it may be improved, or surpassed by alternative approaches, such as the use of pangenome graphs. I expect this

to be a highly active area of research and discussion and I feel my work is an important contribution to its early stages.

Availability of data and materials

Chapter two:

Sequencing data generated for this work, along with the *Am. muticum* assembly is available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2021-10-08_wheatxmuticum. *Am. muticum* illumina sequencing data available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Grewal_et_al_2021-09-13_Amybylopyrum_muticum/. The Chinese Spring sequencing data used is available from ENA (study PRJNA393343; runs SRR5893651 and SRR5893652). The Paragon sequencing data used is available from ENA (study PRJEB35709; runs ERR3728451, ERR3760033, ERR3760405 and ERR3728448). Custom scripts used for introgression detection are available at: https://github.com/benedictcoombes/alien_detection

Chapter three:

Publicly available sequencing data used in this work is available at the European Nucleotide Archive (ENA): HiBAP I enrichment capture sequencing data - PRJEB38874; *Th. ponticum*—SRR13484812; *S. vavilovii:* ERR505040, ERR505041, ERR505042; *S. cereale* accession Lo90: ERR504990, ERR504991, ERR504992; *S. cereale* accession Lo176: ERR505005, ERR505006, ERR505007; *S. cereale* accession Lo282: ERR505015, ERR505016, ERR505017; *S. cereale* accession Lo351: ERR505035, ERR505036, ERR505037; *Ae. Tauschii* accession XJ65*:* SRR13961980; Y173: SRR13962062; SX60: SRR13962012; AY29: SRR13961834; KU2832: SRR13961928; Y215: SRR13962048; Weebil1: PRJEB35709; Norin61: PRJNA492239; Pavon76: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2021-10-08_wheatxmuticum/PIP-2495/200812_A00478_0126_AHN5W3DRXX/A10948_1_1/; *T. aestivum* cv. Chinese Spring RNAseq data: Root - SRP133837; SRR6799264; SRR6799265; Leaf - SRR6799258; SRR6799259; SRR6799260; Spike - SRR6802608; SRR6802609; SRR6802610; SRR6802611.

VCF and hapmap genotype files for HiBAP I are available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2022-04-08_HiBAP_genotyping/

The phenotypic data for the HIBAP I panel evaluated under yield potential and heat stress conditions can be found in the Dataverse CIMMYT Research Data Repository at https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/10548643.

Chapter four:

The pantranscriptome reference, along with a python script to sum expression counts across all transcripts of a given Chinese Spring gene and its 1-to-1 orthologues, can be found at https://doi.org/10.6084/m9.figshare.24242767.

The RNA-seq data and DNA sequencing data from He et al. (2022) reanalysed here are stored in the NCBI SRA under project codes PRJNA670223 and PRJNA787276.

References

Adhikari, L. *et al.* (2022). 'A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations', *Scientific Reports*, 12(17583). doi: 10.1038/s41598-022-19858-2.

Akhunov, E., Nicolet, C. and Dvorak, J. (2009). 'Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay', *Theoretical and Applied Genetics*, 119, pp. 507–517. doi: 10.1007/s00122-009-1059-5.

Akter, N. and Rafiqul Islam, M. (2017). 'Heat stress effects and management in wheat. A review', *Agronomy for Sustainable Development*, 37(5). doi: 10.1007/s13593-017-0443-9.

Amani, I., Fischer, R. A. and Reynolds, M. P. (1996). 'Evaluation of canopy temperature as a screening tool for heat tolerance in spring wheat', *Journal of Agronomy and Crop Science*, 176(2), pp. 119–129.

Appels, R. *et al.* (2018). 'Shifting the limits in wheat research and breeding using a fully annotated reference genome', *Science*, 361(6403). doi: 10.1126/SCIENCE.AAR7191.

Argyros, R. D. *et al.* (2008). 'Type B Response Regulators of Arabidopsis Play Key Roles in Cytokinin Signaling and Plant Development', *The Plant Cell*, 20(8), pp. 2102-2116. doi: 10.1105/tpc.108.059584.

Asseng, S. *et al.* (2014). 'Rising temperatures reduce global wheat production', *Nature Climate Change*, 5(2), pp. 143–147. doi: 10.1038/nclimate2470.

Athiyannan, N. *et al.* (2022). 'Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning', *Nature Genetics*, 54, pp. 227–231. doi: 10.1038/s41588-022-01022-1.

Atlin, G. N., Cairns, J. E. and Das, B. (2017). 'Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change', *Global Food Security*, 12, pp. 31–37. doi: 10.1016/j.gfs.2017.01.008.

Avni, R. *et al.* (2017). 'Wild emmer genome architecture and diversity elucidate wheat evolution and domestication', *Science*, 357(6346), pp. 93–97. doi: 10.1126/science.aan0032.

Baker, L. *et al.* (2020). 'Exploiting the genome of *Thinopyrum elongatum* to expand the gene pool of hexaploid wheat', *Theoretical and Applied Genetics*, 133, pp. 2213–2226. doi: 10.1007/s00122-020-03591-3.

Balla, K. *et al.* (2019). 'Heat stress responses in a large set of winter wheat cultivars (*Triticum*

aestivum L.) depend on the timing and duration of stress', *PLoS ONE*, 14(9). doi: 10.1371/journal.pone.0222639.

Bariana, H. S. *et al.* (2001). 'Mapping of durable adult plant and seedling resistances to stripe rust and stem rust diseases in wheat', *Australian Journal of Agricultural Research*, 52(12), pp. 1247–1255. doi: 10.1071/AR01040.

Bauer, E. *et al.* (2017). 'Towards a whole-genome sequence for rye (*Secale cereale* L.)', *The Plant Journal*, 89(5), pp. 853–869. doi: 10.1111/TPJ.13436.

Beddow, J. M. *et al* (2015). 'Research investment implications of shifts in the global geography of wheat stripe rust', *Nature Plants*, 1(15132). doi: 10.1038/nplants.2015.132.

Begcy, K. *et al.* (2018). 'Compared to Australian Cultivars, European Summer Wheat (*Triticum aestivum*) Overreacts When Moderate Heat Stress Is Applied at the Pollen Development Stage', *Agronomy*. *Agronomy*, 8(99). doi: 10.3390/agronomy8070099.

Bertholdsson, N., Andersson, S. C. and Merker, A. (2012). 'Allelopathic potential of Triticum spp., Secale spp. and Triticosecale spp. and use of chromosome substitutions and translocations to improve weed suppression ability in winter wheat', *Plant Breeding*, 131(1), pp. 75–80. doi: https://doi.org/10.1111/j.1439-0523.2011.01895.x.

Bevan, M. W. and Uauy, C. (2013). 'Genomics reveals new landscapes for crop improvement', *Genome Biology*, 14(206). doi: 10.1186/gb-2013-14-6-206.

Bheemanahalli, R. *et al.* (2019). 'Quantifying the Impact of Heat Stress on Pollen Germination, Seed Set, and Grain Filling in Spring Wheat', *Crop Science*, 59(2), pp. 684–696. doi: 10.2135/cropsci2018.05.0292.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Borrill, P., Adamski, N. and Uauy, C. (2015). 'Genomics as the key to unlocking the polyploid potential of wheat', *New Phytologist*, 208(4), pp. 1008–1022. doi: https://doi.org/10.1111/nph.13533.

Brachi, B., Morris, G. P. and Borevitz, J. O. (2011). 'Genome-wide association studies in plants: the missing heritability is in the field', *Genome Biology*, 12(232). doi: 10.1186/gb-2011-12-10-232.

Bray, N. L. *et al.* (2016). 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. doi: 10.1038/nbt.3519.

Brenchley, R. *et al.* (2012). 'Analysis of the bread wheat genome using whole-genome shotgun sequencing', *Nature*, 491, pp. 705–710. doi: 10.1038/nature11650.

Brisson, N. *et al.* (2010). 'Why are wheat yields stagnating in Europe? A comprehensive data analysis for France', *Field Crops Research*, 119(1), pp. 201–212. doi: https://doi.org/10.1016/j.fcr.2010.07.012.

Broekema, R. V., Bakker, O. B. and Jonkers, I. H. (2020). 'A practical view of fine-mapping and gene prioritization in the post-genome-wide association era', *Open Biology*, 10(190221). doi: 10.1098/rsob.190221.

Browning, B. L., Zhou, Y. and Browning, S. R. (2018). 'A One-Penny Imputed Genome from Next-Generation Reference Panels', *American Journal of Human Genetics*. Cell Press, 103(3), pp. 338–348. doi: 10.1016/j.ajhg.2018.07.015.

Camacho, C. *et al*. (2009). 'BLAST+: architecture and applications.', *BMC Bioinformatics*, 10(421). doi: 10.1186/1471-2105-10-421.

Chapman, S. C. *et al.* (2012). 'Plant adaptation to climate change—opportunities and priorities in breeding', *Crop and Pasture Science*, 63(3), pp. 251-268. doi: 10.1071/CP11303.

Charmet, G. (2011). 'Wheat domestication: Lessons for the future', *Comptes Rendus Biologies*, 334(3), pp. 212–220. doi: 10.1016/J.CRVI.2010.12.013.

Chemayek, B. *et al.* (2017). 'Tight repulsion linkage between Sr36 and Sr39 was revealed by genetic, cytogenetic and molecular analyses', *Theoretical and Applied Genetics*, 130, pp. 587–595. doi: 10.1007/s00122-016-2837-5.

Chen, N. (2004). 'Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences.' *Current protocols in bioinformatics*, 5(4.10.1-4.10.14). doi: 10.1002/0471250953.bi0410s05.

Chen, S. *et al.* (2020). 'Wheat gene Sr60 encodes a protein with two putative kinase domains that confers resistance to stem rust', *New Phytologist*, 225(2), pp. 948–959. doi: https://doi.org/10.1111/nph.16169.

Chen, W. *et al.* (2014). 'Wheat stripe (yellow) rust caused by *Puccinia striiformis* f. sp. *tritici*', *Molecular Plant Pathology*, 15(5), pp. 433–446. doi: 10.1111/mpp.12116.

Chen, X. M. (2005). 'Epidemiology and control of stripe rust [Puccinia striiformis f. sp. tritici] on wheat', *Canadian Journal of Plant Pathology*, 27(3), pp. 314–337. doi: 10.1080/07060660509507230.

Cheng, H. *et al.* (2019). 'Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat', *Genome Biology*, 20(136). doi: 10.1186/s13059-019-1744-x.

Classen, A. T. *et al.* (2015). 'Direct and indirect effects of climate change on soil microbial and soil microbial-plant interactions: What lies ahead?', *Ecosphere*, 6(130). doi: 10.1890/ES15-00217.1.

Clavijo, B. 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. doi: 10.1101/gr.217117.116.

Colmer, T. D., Flowers, T. J. and Munns, R. (2006). 'Use of wild relatives to improve salt tolerance in wheat', *Journal of Experimental Botany*, 57(5), pp. 1059–1078. doi: 10.1093/JXB/ERJ124.

Coombes, B. *et al.* (2022). 'Whole genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/ *Ambylopyrum muticum* introgression lines', *Plant Biotechnology Journal*, 21(3), pp. 482-496. doi: 10.1111/pbi.13859.

Coombes, B. *et al.* (2024). 'Introgressions lead to reference bias in wheat RNA-seq analysis', *BMC Biology*, 22(56). doi: 10.1186/s12915-024-01853-w.

Cooper, H. D., Spillane, C. and Hodgkin, T. (2001). 'Broadening The Genetic Base Of Crop Production', New York: CABI Publishing; Rome: Food and Agriculture Organization of the United Nations (FAO); Rome: International Plant Genetic Resources Institute (IPGRI).

Cossani, C. M. and Reynolds, M. P. (2012). 'Physiological traits for improving heat tolerance in wheat', *Plant physiology*, 160(4), pp. 1710–1718. doi: 10.1104/PP.112.207753.

Cossani, C. M. and Reynolds, M. P. (2015). 'Heat Stress Adaptation in Elite Lines Derived from Synthetic Hexaploid Wheat', *Crop Science*, 55(6), pp. 2719–2735. doi: 10.2135/CROPSCI2015.02.0092.

De Coster, W. *et al.* (2018). 'NanoPack: visualizing and processing long-read sequencing data', *Bioinformatics*, 34(15), pp. 2666–2669. doi: 10.1093/bioinformatics/bty149.

Crespo-Herrera, L. A. *et al.* (2013). 'Resistance to multiple cereal aphids in wheat–alien substitution and translocation lines', *Arthropod-Plant Interactions*, 7, pp. 535–545. doi: 10.1007/s11829-013-9267-y.

Cruppe, G. *et al.* (2019). 'Novel Sources of Wheat Head Blast Resistance in Modern Breeding Lines and Wheat Wild Relatives', *Plant Disease*, 104(1), pp. 35–43. doi: 10.1094/PDIS-05-19-0985-RE.

Cruz, C. D. *et al.* (2016). 'The 2NS Translocation from Aegilops ventricosa Confers Resistance to

the Triticum Pathotype of Magnaporthe oryzae', *Crop Science*, 56(3), pp. 990–1000. doi: https://doi.org/10.2135/cropsci2015.07.0410.

Cseh, A. *et al.* (2019). 'Development of a New A^m–Genome-Specific Single Nucleotide Polymorphism Marker Set for the Molecular Characterization of Wheat–*Triticum monococcum* Introgression Lines', *The Plant Genome*, 12(180098). doi: https://doi.org/10.3835/plantgenome2018.12.0098.

Daly, G. M. *et al.* (2015). 'Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data', *PLOS ONE*. Public Library of Science, 10(6), p. e0129059. Available at: https://doi.org/10.1371/journal.pone.0129059.

Das, M. K. *et al.* (2016). 'Genetic diversity among synthetic hexaploid wheat accessions (Triticum aestivum) with resistance to several fungal diseases', *Genetic Resources and Crop Evolution*. Springer Netherlands, 63(8), pp. 1285–1296. doi: 10.1007/S10722-015-0312-9/FIGURES/4.

Dean, R. *et al.* (2012). 'The Top 10 fungal pathogens in molecular plant pathology', *Molecular Plant Pathology*, 13(4), pp. 414–430. doi: 10.1111/j.1364-3703.2011.00783.x.

Degen, G. E., Orr, D. J. and Carmo-Silva, E. (2021). 'Heat-induced changes in the abundance of wheat Rubisco activase isoforms', *New Phytologist*, 229(3), pp. 1298–1311. doi: 10.1111/nph.16937.

Depristo, M. A. *et al.* (2011). 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, 43(5), pp. 491–498. doi: 10.1038/ng.806.

Devi, U. *et al.* (2019). 'Development and characterisation of interspecific hybrid lines with genome-wide introgressions from *Triticum timopheevii* in a hexaploid wheat background', *BMC Plant Biology*, 19(183). doi: 10.1186/s12870-019-1785-z.

Dias, A. S., Bagulho, A. S. and Lidon, F. C. (2008). 'Ultrastructure and biochemical traits of bread and durum wheat grains under heat stress', *Brazilian Journal of Plant Physiology*, 20(4), pp. 323–333. doi: 10.1590/s1677-04202008000400008.

Dias, A. S., Lidon, F. C. and Ramalho, J. C. (2009). 'I. Heat stress in Triticum: kinetics of Ca and Mg accumulation', *Brazilian Journal of Plant Physiology*. 21(2), pp. 123–134. doi: 10.1590/S1677-04202009000200005.

Dmochowska-Boguta, M. *et al.* (2020). 'TaWAK6 encoding wall-associated kinase is involved in wheat resistance to leaf rust similar to adult plant resistance', *PLoS ONE*, 15(1), p. e0227713. doi: https://doi.org/10.1371/journal.pone.0227713.

Dobin, A. *et al.* (2013). 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Doussinault, G. *et al.* (1983). 'Transfer of a dominant gene for resistance to eyespot disease from a wild grass to hexaploid wheat', *Nature*, 303(5919), pp. 698–700. doi: 10.1038/303698a0.

Dover, Gabriel A. and Riley, R. (1972). 'Prevention of pairing of homoeologous meiotic chromosomes of wheat by an activity of supernumerary chromosomes of *Aegilops*', *Nature*, 240, pp. 159–161. doi: 10.1038/240159a0.

Dover, G. A. and Riley, R. (1972). 'Variation at two loci affecting homoeologous meiotic chromosome pairing in triticum aestivum × aegilops mutica hybrids', *Nature New Biology*, 235, pp. 61–62. doi: 10.1038/newbio235061a0.

Dreccer, M. F. *et al.* (2007). 'CIMMYT-selected derived synthetic bread wheats for rainfed environments: Yield evaluation in Mexico and Australia', *Field Crops Research*, 100(2-3), pp. 218–228. doi: https://doi.org/10.1016/j.fcr.2006.07.005.

Dreisigacker, S. *et al.* (2008). 'Use of synthetic hexaploid wheat to increase diversity for CIMMYT bread wheat improvement', *Australian Journal of Agricultural Research*, 59(5), pp. 413-420. doi: 10.1071/AR07225.

Dreisigacker, S. *et al.* (2021). 'Effect of flowering time-related genes on biomass, harvest index, and grain yield in CIMMYT elite spring bread wheat', *Biology*, 10(855). doi: 10.3390/biology10090855.

Dundas, I. *et al.* (2015). 'Chromosome Engineering and Physical Mapping of the *Thinopyrum ponticum* Translocation in Wheat Carrying the Rust Resistance Gene *Sr26*', *Crop Science*, 55(2), pp. 648–657. doi: 10.2135/cropsci2014.08.0590.

Dvorak, J. *et al.* (2006). 'Molecular Characterization of a Diagnostic DNA Marker for Domesticated Tetraploid Wheat Provides Evidence for Gene Flow from Wild Tetraploid Wheat to Hexaploid Wheat', *Molecular Biology and Evolution*, 23(7), pp. 1386–1396. doi: 10.1093/molbev/msl004.

Dvorak, J., Deal, K. R. and Luo, M. C. (2006). 'Discovery and Mapping of Wheat Ph1 Suppressors', *Genetics*, 174(1), pp. 17–27. doi: 10.1534/genetics.106.058115.

Dzyubenko, N. I. (2018). 'Vavilov's Collection of Worldwide Crop Genetic Resources in the 21st century', *Biopreservation and Biobanking*, 16(5), pp. 377–383. doi: 10.1089/bio.2018.0045.

Earl, D. A. and vonHoldt, B. M. (2012). 'STRUCTURE HARVESTER: A website and program for

visualizing STRUCTURE output and implementing the Evanno method', *Conservation Genetics*, 4, pp. 359–361. doi: 10.1007/s12686-011-9548-7.

Ellis, J. G. *et al.* (2014). 'The past, present and future of breeding rust resistant wheat', *Frontiers in Plant Science*, 5(641). doi: 10.3389/fpls.2014.00641.

Emms, David M. and Kelly, S. (2019). 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biology*, 20(238). doi: 10.1186/s13059-019-1832-y.

Erenstein, O. *et al.* (2022). 'Global Trends in Wheat Production, Consumption and Trade', *Wheat Improvement*. Springer International Publishing, pp. 47–66. doi: 10.1007/978-3-030-90673-3_4.

Evanno, G., Regnaut, S. and Goudet, J. (2005). 'Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study', *Molecular Ecology*, 14(8), pp. 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x.

FAO (2023). 'Food Outlook - Biannual report on global food markets', *Food Outlook*. doi: https://doi.org/10.4060/cc8589en.

Fellers, J. P. *et al.* (2020). 'Resistance to wheat rusts identified in wheat/*Amblyopyrum muticum* chromosome introgressions', *Crop Science*, 60(4), pp. 1957–1964. doi: https://doi.org/10.1002/csc2.20120.

Figueroa, M., Hammond-Kosack, K. E. and Solomon, P. S. (2018). 'A review of wheat diseases—a field perspective', *Molecular Plant Pathology*, 19(6), pp. 1523–1536. doi: https://doi.org/10.1111/mpp.12618.

Finn, R. D., Clements, J. and Eddy, S. R. (2011). 'HMMER web server: interactive sequence similarity searching.', *Nucleic Acids Research*, 39, pp. W29-37. doi: 10.1093/nar/gkr367.

Fischer, R. A. and Maurer, R. (1978). 'Drought Resistance in Spring Wheat Cultivars, I. Grain Yield Responses', *Australian Journal of Agricultural Research*, 29, pp. 897–912. doi: 10.1071/AR9780897.

Frank, M. *et al.* (2020). 'Root-derived trans-zeatin cytokinin protects Arabidopsis plants against photoperiod stress', *Plant, Cell & Environment*, 43(11), pp. 2637–2649. doi: 10.1111/PCE.13860.

Friebe, B. *et al.* (1996). 'Transfer of Wheat Streak Mosaic Virus Resistance from *Agropyron intermedium* into Wheat', *Crop Science*, 36(4), pp. 857-861. doi: https://doi.org/10.2135/cropsci1996.0011183X003600040006x.

Gao, L. *et al.* (2021). 'The *Aegilops ventricosa* 2N^{v}S segment in bread wheat: cytology, genomics

and breeding', *Theoretical and Applied Genetics*, 134, pp. 529–542. doi: 10.1007/s00122-020-03712-y.

Gardiner, L-J., Brabbs, T., *et al.* (2019a). 'Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat', *GigaScience*, 8(4). doi: 10.1093/gigascience/giz018.

Gardiner, L.-J., Wingen, L. U., *et al.* (2019b). 'Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency', *Genome Biology*, 20(69). doi: 10.1186/s13059-019-1675-6.

Garrett, K. A. *et al.* (2006). 'Climate change effects on plant disease: Genomes to ecosystems', *Annual Review of Phytopathology*, pp. 489–509. doi: 10.1146/annurev.phyto.44.070505.143420.

Garrison, E. *et al.* (2018). 'Variation graph toolkit improves read mapping by representing genetic variation in the reference', *Nature Biotechnology*, 36(9), pp. 875–879. doi: 10.1038/nbt.4227.

Gaurav, K. *et al.* (2021). 'Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement', *Nature Biotechnology 2021*, 40, pp. 422–431. doi: 10.1038/s41587-021-01058-4.

Giani, A. M. *et al.* (2020). 'Long walk to genomics: History and current approaches to genome sequencing and assembly', *Computational and Structural Biotechnology Journal*, 18, pp. 9–19. doi: 10.1016/j.csbj.2019.11.002.

Gordon, D. *et al.* (2016). 'Long-read sequence assembly of the gorilla genome', *Science*, 352(6281). doi: 10.1126/science.aae0344.

Gouache, D. *et al.* (2012). 'Evaluating agronomic adaptation options to increasing heat stress under climate change during wheat grain filling in France', *European Journal of Agronomy*, 39, pp. 62–70. doi: 10.1016/j.eja.2012.01.009.

Grabherr, M. G. *et al.* (2011). 'Full-length transcriptome assembly from RNA-Seq data without a reference genome.', *Nature biotechnology*, 29(7), pp. 644–652. doi: 10.1038/nbt.1883.

Grewal, S., Hubbart-Edwards, S., *et al.* (2018). 'Detection of *T. urartu* Introgressions in Wheat and Development of a Panel of Interspecific Introgression Lines', *Frontiers in Plant Science*, 9(1565). doi: https://www.frontiersin.org/articles/10.3389/fpls.2018.01565.

Grewal, S., Yang, C., *et al.* (2018). 'Characterisation of *Thinopyrum bessarabicum* chromosomes through genome-wide introgressions into wheat', *Theoretical and Applied Genetics*, 131, pp. 389–

406. doi: 10.1007/s00122-017-3009-y.

Grewal, S., Othmeni, M., *et al.* (2020). 'Development of Wheat-*Aegilops caudata* Introgression Lines and Their Characterization Using Genome-Specific KASP Markers', *Frontiers in Plant Science*, 11(606). doi: https://doi.org/10.3389/fpls.2020.00606.

Grewal, S., Hubbart-Edwards, S., *et al.* (2020). 'Rapid identification of homozygosity and site of wild relative introgressions in wheat through chromosome-specific KASP genotyping assays', *Plant Biotechnology Journal*. 18(3), pp. 743–755. doi: https://doi.org/10.1111/pbi.13241.

Grewal, S. *et al.* (2021). 'Generation of Doubled Haploid Wheat-*Triticum urartu* Introgression Lines and Their Characterisation Using Chromosome-Specific KASP Markers', *Frontiers in Plant Science*, 12(643636). doi: https://doi.org/10.3389/fpls.2021.643636.

Grewal, S. *et al.* (2022). 'Chromosome-specific KASP markers for detecting *Amblyopyrum muticum* segments in wheat introgression lines', *The Plant Genome*, 15(e20193). doi: https://doi.org/10.1002/tpg2.20193.

Griffiths, S. *et al.* (2006). 'Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat', *Nature*, 439, pp. 749–752. doi: 10.1038/nature04434.

Grote, U. *et al.* (2021). 'Food Security and the Dynamics of Wheat and Maize Value Chains in Africa and Asia', *Frontiers in Sustainable Food Systems*, 4(617009). doi: 10.3389/fsufs.2020.617009.

Guo, W. *et al.* (2020). 'Origin and adaptation to high altitude of Tibetan semi-wild wheat', *Nature Communications*, 11(5085). doi: 10.1038/s41467-020-18738-5.

Haas, B. J. *et al.* (2008). 'Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments', *Genome Biology*, 9(R7). doi: 10.1186/gb-2008-9-1-r7.

Haas, B. J. *et al.* (2013). '*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.', *Nature protocols*, 8, pp. 1494–1512. doi: 10.1038/nprot.2013.084.

Han, G. *et al.* (2020). 'Identification of an Elite Wheat-Rye T1RS·1BL Translocation Line Conferring High Resistance to Powdery Mildew and Stripe Rust', *Plant Disease*, 104(11), pp. 2940–2948. doi: 10.1094/PDIS-02-20-0323-RE.

Hao, M. *et al.* (2020). 'The Resurgence of Introgression Breeding, as Exemplified in Wheat

Improvement', *Frontiers in Plant Science*, 11(252). doi: 10.3389/fpls.2020.00252.

Haudry, A. *et al.* (2007). 'Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication', *Molecular Biology and Evolution*, 24(7), pp. 1506–1517. doi: 10.1093/molbev/msm077.

He, F. *et al.* (2019). 'Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome', *Nature Genetics*, 51, pp. 896–904. doi: 10.1038/s41588-019-0382-2.

Helguera, M. *et al.* (2003). 'PCR Assays for the *Lr37-Yr17-Sr38* Cluster of Rust Resistance Genes and Their Use to Develop Isogenic Hard Red Spring Wheat Lines', *Crop Science*, 43(5), pp. 1839–1847. doi: https://doi.org/10.2135/cropsci2003.1839.

Hoff, K. J. and Stanke, M. (2019). 'Predicting Genes in Single Genomes with AUGUSTUS.', *Current Protocols in Bioinformatics*, 65(e57). doi: 10.1002/cpbi.57.

Hoisington, D. *et al.* (1999). 'Plant genetic resources: What can they contribute toward increased crop productivity?', *PNAS*, 96(11), pp. 5937–5943. doi: 10.1073/pnas.96.11.5937.

Huerta-Cepas, J. *et al.* (2017). 'Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper', *Molecular Biology and Evolution*, 34(8), pp. 2115–2122. doi: 10.1093/molbev/msx148.

Huerta-Cepas, J. *et al.* (2019). 'eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.', *Nucleic Acids Research*, 47(D1), pp. D309–D314. doi: 10.1093/nar/gky1085.

Hummer, K. E. and Hancock, J. F. (2015). 'Vavilovian Centers of Plant Diversity: Implications and Impacts', *HortScience*, 50(6), pp. 780–783. doi: 10.21273/hortsci.50.6.780.

Hyun, M. K. *et al.* (2008). 'Efficient Control of Population Structure in Model Organism Association Mapping', *Genetics*, 178(3), pp. 1709–1723. doi: 10.1534/GENETICS.107.080101.

IWGSC (The International Wheat Genome Sequencing Consortium) (2014). 'A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome', *Science*, 345(1251788). doi: 10.1126/science.1251788.

Jahier, J. *et al.* (2001). 'The *Aegilops ventricosa* segment on chromosome 2AS of the wheat cultivar "VPM1" carries the cereal cyst nematode resistance gene *Cre5*', *Plant Breeding*, 120(2), pp. 125–128. doi: https://doi.org/10.1046/j.1439-0523.2001.00585.x.

Jain, M. *et al.* (2018). 'Nanopore sequencing and assembly of a human genome with ultra-long reads', *Nature Biotechnology*, 36(4), pp. 338–345. doi: 10.1038/nbt.4060.

Jakobsson, M. and Rosenberg, N. A. (2007). 'CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure', *Bioinformatics*, 23(14), pp. 1801–1806. doi: 10.1093/bioinformatics/btm233.

Jauhar, P. P. (2007). 'Meiotic Restitution in Wheat Polyhaploids (Amphihaploids): A Potent Evolutionary Force', *Journal of Heredity*, 98(2), pp. 188–193. doi: 10.1093/jhered/esm011.

Jia, J. *et al.* (2023). 'Genome resources for the elite bread wheat cultivar Aikang 58 and mining of elite homeologous haplotypes for accelerating wheat improvement', *Molecular Plant*, 16(12), pp. 1893–1910. doi: 10.1016/j.molp.2023.10.015.

Joynson, R. *et al.* (2021). 'Uncovering candidate genes involved in photosynthetic capacity using unexplored genetic variation in Spring Wheat', *Plant Biotechnology Journal*, 19(8), pp. 1537–1552. doi: 10.1111/PBI.13568.

Juery, C. *et al.* (2021). 'New insights into homoeologous copy number variations in the hexaploid wheat genome', *The Plant Genome*, 14(e20069). doi: https://doi.org/10.1002/tpg2.20069.

Kahiluoto, H. *et al.* (2019). 'Decline in climate resilience of European wheat', *PNAS*, 116(1), pp. 123–128. doi: 10.1073/pnas.1804387115.

Kamal, N. M. *et al.* (2019). 'Stay-Green Trait: A Prospective Approach for Yield Potential, and Drought and Heat Stress Adaptation in Globally Important Cereals', *International Journal of Molecular Sciences*, 20(5837) doi: 10.3390/ijms20235837.

Kaminow, B. *et al.* (2022). 'Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses', *Genome Research*, 32(4), pp. 738–750. doi: 10.1101/gr.275613.121.

Kaur, S., Francki, M. G. and Forster, J. W. (2012). 'Identification, characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species', *Plant Biotechnology Journal*, 10(2), pp. 125–138. doi: https://doi.org/10.1111/j.1467-7652.2011.00644.x.

Keilwagen, J. *et al.* (2019). 'Detecting Large Chromosomal Modifications Using Short Read Data From Genotyping-by-Sequencing', *Frontiers in Plant Science*, 10(1133). doi: 10.3389/FPLS.2019.01133/BIBTEX.

Keilwagen, J. *et al.* (2022). 'Detecting major introgressions in wheat and their putative origins

using coverage analysis', *Scientific Reports*, 12(1908). doi: 10.1038/s41598-022-05865-w.

Khazan, S. *et al.* (2020). 'Reducing the size of an alien segment carrying leaf rust and stripe rust resistance in wheat', *BMC Plant Biology*, 20(153). doi: 10.1186/s12870-020-2306-9.

Kim, D. *et al.* (2019). 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology*, 37, pp. 907–915. doi: 10.1038/s41587-019-0201-4.

King, J. *et al.* (2017). 'A step change in the transfer of interspecific variation into wheat from *Amblyopyrum muticum*', *Plant Biotechnology Journal*, 15(2), pp. 217–226. doi: https://doi.org/10.1111/pbi.12606.

King, J. *et al.* (2019). 'Development of Stable Homozygous Wheat/*Amblyopyrum muticum* (*Aegilops mutica*) Introgression Lines and Their Cytogenetic and Molecular Characterization', *Frontiers in Plant Science*, 10(34). doi: 10.3389/fpls.2019.00034.

King, J. *et al.* (2022). 'Introgression of the *Triticum timopheevii* Genome Into Wheat Detected by Chromosome-Specific Kompetitive Allele Specific PCR Markers', *Frontiers in Plant Science*, 13(919519). doi: 10.3389/fpls.2022.919519.

Klindworth, D. L. *et al.* (2012). 'Introgression and characterization of a goatgrass gene for a high level of resistance to ug99 stem rust in tetraploid wheat', *G3*, 2(6), pp. 665–673. doi: 10.1534/g3.112.002386.

Knott, D. R. (1961). 'THE INHERITANCE OF RUST RESISTANCE. VI. THE TRANSFER OF STEM RUST RESISTANCE FROM *AGROPYRON ELONGATUM* TO COMMON WHEAT', *Canadian Journal of Plant Science*, 41(1), pp. 109–123. doi: 10.4141/cjps61-014.

Kolmogorov, M. *et al.* (2019). 'Assembly of long, error-prone reads using repeat graphs', *Nature Biotechnology*, 37, pp. 540–546. doi: 10.1038/s41587-019-0072-8.

Koo, D-H. *et al.* (2017). 'Homoeologous recombination in the presence of *Ph1* gene in wheat', *Chromosoma*, 126, pp. 531–540. doi: 10.1007/s00412-016-0622-5.

Koo, D-H., Friebe, B. and Gill, B. S. (2020). 'Homoeologous Recombination: A Novel and Efficient System for Broadening the Genetic Variability in Wheat', *Agronomy*, 10(1059). doi: 10.3390/agronomy10081059.

Korf, I. (2004). 'Gene finding in novel genomes', *BMC Bioinformatics*, 5(59). doi: 10.1186/1471-2105-5-59.

Krattinger, S. G. *et al.* (2009). 'A Putative ABC Transporter Confers Durable Resistance to Multiple

Fungal Pathogens in Wheat', *Science*, 323(5919), pp. 1360–1363. doi: 10.1126/science.1166453.

Kumar, R. *et al.* (2022). 'Stay-green trait serves as yield stability attribute under combined heat and drought stress in wheat (Triticum aestivum L.)', *Plant Growth Regulation*, 96(1), pp. 67–78. doi: 10.1007/s10725-021-00758-w.

Kumar, R. R. *et al.* (2021). 'Characterizing the putative mitogen-activated protein kinase (*MAPK*) and their protective role in oxidative stress tolerance and carbon assimilation in wheat under terminal heat stress', *Biotechnology Reports*, 29(e00597). doi: 10.1016/J.BTRE.2021.E00597.

Kummu, M. *et al.* (2012). 'Lost food, wasted resources: Global food supply chain losses and their impacts on freshwater, cropland, and fertiliser use', *Science of The Total Environment*, 438, pp. 477–489. doi: https://doi.org/10.1016/j.scitotenv.2012.08.092.

Lamaoui, M. *et al.* (2018). 'Heat and Drought Stresses in Crops and Approaches for Their Mitigation', *Frontiers in Chemistry*, 6(26). doi: 10.3389/fchem.2018.00026.

Langmead, B. and Salzberg, S. L. (2012). 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lawas, L. M. F. *et al.* (2018). 'Molecular mechanisms of combined heat and drought stress resilience in cereals', *Current Opinion in Plant Biology*, 45(Part B), pp. 212–217. doi: 10.1016/j.pbi.2018.04.002.

Lemay, M-A. *et al.* (2019). 'Screening populations for copy number variation using genotyping-by-sequencing: A proof of concept using soybean fast neutron mutants', *BMC Genomics*, 20(634). doi: 10.1186/s12864-019-5998-1.

Levy, A. A. and Feldman, M. (2022). 'Evolution and origin of bread wheat.', *The Plant cell*, 34(7), pp. 2549–2567. doi: 10.1093/plcell/koac130.

Li, A. *et al.* (2018). 'Synthetic Hexaploid Wheat: Yesterday, Today, and Tomorrow', *Engineering*, 4(4), pp. 552–558. doi: 10.1016/J.ENG.2018.07.001.

Li, B. and Dewey, C. N. (2011). 'RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics*, 12(323). doi: 10.1186/1471-2105-12-323.

Li, H. *et al.* (2009). 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078-2079. doi: 10.1093/BIOINFORMATICS/BTP352.

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

2987-2993. doi: 10.1093/BIOINFORMATICS/BTR509.

Li, H. (2013). 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv*. doi: 10.48550/arxiv.1303.3997.

Li, H. *et al.* (2017). 'Introgression of the *Aegilops speltoides Su1-Ph1* Suppressor into Wheat', *Frontiers in Plant Science*, 8(2163). doi: https://doi.org/10.3389/fpls.2017.02163.

Li, H. (2018). 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*, 34(18), pp. 3094-3100. doi: 10.1093/BIOINFORMATICS/BTY191.

Li, H. *et al.* (2023). 'Heterozygous inversion breakpoints suppress meiotic crossovers by altering recombination repair outcomes', *PLoS Genetics*, 19(4), p. e1010702. doi: https://doi.org/10.1371/journal.pgen.1010702.

Li, H. (2011). 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27(21), pp. 2987–2993. doi: 10.1093/BIOINFORMATICS/BTR509.

Li, H. and Durbin, R. (2009). 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*. Bioinformatics, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, J. *et al.* (2022). 'Molecular and Cytogenetic Identification of Stem Rust Resistant Wheat–*Thinopyrum intermedium* Introgression Lines', *Plant Disease*, 106(9), pp. 2447–2454. doi: 10.1094/PDIS-10-21-2274-RE.

Li, L. *et al.* (2013). 'Mendelian and Non-Mendelian Regulation of Gene Expression in Maize', *PLoS Genetics*, 9(1), p. e1003202. doi: https://doi.org/10.1371/journal.pgen.1003202.

Li, W. and Godzik, A. (2006). 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.', *Bioinformatics*, 22(13), pp. 1658–1659. doi: 10.1093/bioinformatics/btl158.

Lillemo, M. *et al.* (2005). 'Differential Adaptation of CIMMYT Bread Wheat to Global High Temperature Environments', *Crop Science*, 45(6), pp. 2443–2453. doi: 10.2135/cropsci2004.0663.

Ling, H.-Q. *et al.* (2018). 'Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*', *Nature*, 557, pp. 424–428. doi: 10.1038/s41586-018-0108-0.

Lipka, A. E. *et al.* (2012). 'GAPIT: genome association and prediction integrated tool', *Bioinformatics*, 28(18), pp. 2397–2399. doi: 10.1093/BIOINFORMATICS/BTS444.

Liu, X., MacLeod, J. N. and Liu, J. (2018). 'iMapSplice: Alleviating reference bias through personalized RNA-seq alignment', *PLoS ONE*, 13(8), p. e0201554. doi: 10.1371/journal.pone.0201554.

Liu, Y. *et al.* (2020). 'Research Progress on the Roles of Cytokinin in Plant Response to Stress', *International Journal of Molecular Sciences*, 21(6574). doi: 10.3390/ijms21186574.

Lopes, M. S. and Reynolds, M. P. (2011). 'Drought Adaptive Traits and Wide Adaptation in Elite Lines Derived from Resynthesized Hexaploid Wheat', *Crop Science*, 51(4), pp. 1617–1626. doi: 10.2135/CROPSCI2010.07.0445.

Luo, M-C. *et al.* (2017). 'Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*', *Nature*, 551, pp. 498–502. doi: 10.1038/nature24486.

Lyra, D. H. *et al.* (2021). 'Gene-based mapping of trehalose biosynthetic pathway genes reveals association with source- and sink-related yield traits in a spring wheat panel', *Food and Energy Security*, 10(3), p. e292. doi: https://doi.org/10.1002/fes3.292.

Majoros, W. H., Pertea, M. and Salzberg, S. L. (2004). 'TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.', *Bioinformatics*, 20(16), pp. 2878–2879. doi: 10.1093/bioinformatics/bth315.

Manès, Y. *et al.* (2012). 'Genetic Yield Gains of the CIMMYT International Semi-Arid Wheat Yield Trials from 1994 to 2010', *Crop Science*, 52(4), pp. 1543–1552. doi: https://doi.org/10.2135/cropsci2011.10.0574.

Mapleson, D. *et al.* (2018). 'Efficient and accurate detection of splice junctions from RNA-seq with Portcullis', *GigaScience*, 7(12), p. giy131. doi: 10.1093/gigascience/giy131.

Martín, A. C. *et al.* (2017). 'Dual effect of the wheat *Ph1* locus on chromosome synapsis and crossover', *Chromosoma*, 126, pp. 669–680. doi: 10.1007/s00412-017-0630-0.

Martiniano, R. *et al.* (2020). 'Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph', *Genome Biology*, 21(250). doi: 10.1186/s13059-020-02160-7.

Mascher, M. *et al.* (2019). 'Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding', *Nature Genetics*, 51, pp. 1076–1081. doi: 10.1038/s41588-019-0443-6.

McCouch, S. *et al.* (2020). 'Mobilizing Crop Biodiversity', *Molecular Plant*, 13(10), pp. 1341–1344.

doi: 10.1016/J.MOLP.2020.08.011.

Mishra, M. K. *et al.* (2013). 'Overexpression of WsSGTL1 Gene of *Withania somnifera* Enhances Salt Tolerance, Heat Tolerance and Cold Acclimation Ability in Transgenic *Arabidopsis* Plants', *PLoS ONE*, 8(4), p. e63064. doi: https://doi.org/10.1371/journal.pone.0063064.

Misra, P. *et al.* (2016). 'Functional Analysis and the Role of Members of SGT Gene Family of Withania somnifera', in Jha, S. (ed) *Transgenesis and Secondary Metabolism*, Cham: Springer, pp. 1–14. doi: 10.1007/978-3-319-27490-4_16-1.

Mitrofanova, O. P. (2012). 'Wheat genetic resources in Russia: Current status and prebreeding studies', *Russian Journal of Genetics: Applied Research*, 2, pp. 277–285. doi: 10.1134/S2079059712040077.

Mo, S. *et al.* (2021). 'Mitogen-activated protein kinase action in plant response to high-temperature stress: a mini review', *Protoplasma*, 258, pp. 477–482. doi: https://doi.org/10.1007/s00709-020-01603-z.

Mohammadi, M. *et al.* (2012). 'Effective application of canopy temperature for wheat genotypes screening under different water availability in warm environments', *Bulgarian Journal of Agricultural Science*, 18(6), pp. 934–941.

Molero, G., Joynson, R., Pinera-Chavez, Francisco J., *et al.* (2019). 'Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring wheat and its role in yield potential', *Plant Biotechnology Journal*, 17(7), pp. 1276–1288. doi: 10.1111/PBI.13052.

Molero, G. *et al.* (2023). 'Exotic alleles contribute to heat tolerance in wheat under field conditions', *Communications Biology*, 6(21). doi: 10.1038/s42003-022-04325-5.

Mondal, S. *et al.* (2013). 'Earliness in wheat: A key to adaptation under terminal and continual high temperature stress in South Asia', *Field Crops Research*, 151, pp. 19–26. doi: 10.1016/j.fcr.2013.06.015.

Moskal, K. *et al.* (2021). 'The Pros and Cons of Rye Chromatin Introgression into Wheat Genome', *Agronomy*, 11(456). doi: 10.3390/agronomy11030456.

Munger, S. C. *et al.* (2014). 'RNA-Seq Alignment To Individualized Genomes Improves Transcript Abundance Estimates In Multiparent Populations', *Genetics*, 198(1), pp. 59–73. doi: 10.1534/GENETICS.114.165886/-/DC1.

Narang, D. *et al.* (2020). 'Discovery and characterisation of a new leaf rust resistance gene

introgressed in wheat from wild wheat *Aegilops peregrina*', *Scientific Reports*, 10(7573). doi: 10.1038/s41598-020-64166-2.

Nasser, L. M. *et al.* (2020). 'Combining Ability of Early-Maturing Yellow Maize Inbreds under Combined Drought and Heat Stress and Well-Watered Environments', *Agronomy*, 10(1585). doi: 10.3390/agronomy10101585.

Nesar, N. A. *et al.* (2022). 'Terminal Heat Stress and Its Mitigation Options through Agronomic Interventions in Wheat Crop: A Review', *International Journal of Environment and Climate Change*, 12(11), pp. 131–139. doi: 10.9734/ijecc/2022/v12i1130955.

Nevo, E. and Chen, G. (2010). 'Drought and salt tolerances in wild relatives for wheat and barley improvement', *Plant, Cell & Environment*, 33(4), pp. 670–685. doi: https://doi.org/10.1111/j.1365-3040.2009.02107.x.

Nguyen, K. H. *et al.* (2016). 'Arabidopsis type B cytokinin response regulators ARR1, ARR10, and ARR12 negatively regulate plant responses to drought', *PNAS*, 113(11), pp. 3090–3095. doi: 10.1073/PNAS.1600399113/-/DCSUPPLEMENTAL.

Ning, G. *et al.* (2021). 'Genetic manipulation of *Soc1*-like genes promotes photosynthesis in flowers and leaves and enhances plant tolerance to high temperature', *Plant Biotechnology*, 19(1), pp. 8–10. doi: 10.1111/PBI.13432.

Niu, Z. *et al.* (2014). 'Development and characterization of wheat lines carrying stem rust resistance gene *Sr43* derived from *Thinopyrum ponticum*', *Theoretical and Applied Genetics*, 127, pp. 969–980. doi: 10.1007/s00122-014-2272-4.

Nkongolo, K. K. *et al.* (1992). 'Identification of Rye Chromosomes Involved in Tolerance to Barley Yellow Dwarf Virus Disease in Wheat × Triticale Hybrids', *Plant Breeding*, 109(2), pp. 123–129. doi: https://doi.org/10.1111/j.1439-0523.1992.tb00162.x.

Nyine, M. *et al.* (2020). 'Genomic Patterns of Introgression in Interspecific Populations Created by Crossing Wheat with Its Wild Relative', *G3*, 10(10), pp. 3651–3661. doi: 10.1534/g3.120.401479.

Ortiz, R. *et al.* (2008). 'Wheat genetic resources enhancement by the International Maize and Wheat Improvement Center (CIMMYT)', *Genetic Resources and Crop Evolution*, 55, pp. 1095–1140. doi: 10.1007/s10722-008-9372-4.

Ou, S. *et al.* (2019). 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome Biology*, 20(275). doi: 10.1186/s13059-019-1905-y.

Pardey, P. G. *et al.* (2014). 'A Bounds Analysis of World Food Futures: Global Agriculture Through to 2050', *Australian Journal of Agricultural and Resource Economics*, 58(4), pp. 571–589. doi: https://doi.org/10.1111/1467-8489.12072.

Pask, A. *et al.* (2012). 'Physiological breeding II: a field guide to wheat phenotyping', *CIMMYT*. Available at: https://repository.cimmyt.org/bitstream/handle/10883/1288/96144.pdf?sequence=3&isAllowed=y.

Pedersen, B. S. and Quinlan, A. R. (2018). 'hts-nim: scripting high-performance genomic analyses', *Bioinformatics*, 34(19), pp. 3387-3389. doi: 10.1093/BIOINFORMATICS/BTY358.

Pellicer, J. and Leitch, I. J. (2020). 'The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies', *New Phytologist*, 226(2), pp. 301–305. doi: https://doi.org/10.1111/nph.16261.

Pertea, M. *et al.* (2015). 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.

Pingali, P. (2007). 'Westernization of Asian diets and the transformation of food systems: Implications for research and policy', *Food Policy*, 32(3), pp. 281–298. doi: https://doi.org/10.1016/j.foodpol.2006.08.001.

Pinto, R. S. *et al.* (2010). 'Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects', *Theoretical and Applied Genetics*, 121, pp. 1001–1021. doi: 10.1007/s00122-010-1351-4.

Pinto, R. S., Molero, G. and Reynolds, M. P. (2017). 'Identification of heat tolerant wheat lines showing genetic variation in leaf respiration and other physiological traits', *Euphytica*, 213(76). doi: 10.1007/S10681-017-1858-8.

Pinto, R. S. and Reynolds, M. P. (2015). 'Common genetic basis for canopy temperature depression under heat and drought stress associated with optimized root distribution in bread wheat', *Theoretical and Applied Genetics*, 128(4), pp. 575–585. doi: 10.1007/S00122-015-2453-9.

Placido, D. F. *et al.* (2013). 'Introgression of Novel Traits from a Wild Wheat Relative Improves Drought Adaptation in Wheat', *Plant Physiology*, 161(4), pp. 1806–1819. doi: 10.1104/pp.113.214262.

Potter, S. C. *et al.* (2018). 'HMMER web server: 2018 update', *Nucleic Acids Research*, 46(W1), pp. W200-W204. doi: 10.1093/NAR/GKY448.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics*, 155(2), pp. 945–959. doi: 10.1093/genetics/155.2.945.

Przewieslik-Allen, A. M. *et al.* (2019). 'Developing a High-Throughput SNP-Based Marker System to Facilitate the Introgression of Traits From *Aegilops* Species Into Bread Wheat (*Triticum aestivum*)', *Frontiers in Plant Science*, 9(1993). doi: 10.3389/fpls.2018.01993.

Przewieslik-Allen, A. M. *et al.* (2021). 'The role of gene flow and chromosomal instability in shaping the bread wheat genome', *Nature Plants*, 7, pp. 172–183. doi: 10.1038/s41477-020-00845-2.

Quinlan, A. R. and Hall, I. M. (2010). 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/BIOINFORMATICS/BTQ033.

Ramírez-González, R. H. *et al.* (2018). 'The transcriptional landscape of polyploid wheat', *Science*, 361(6403). doi: 10.1126/science.aar6089.

Rasheed, A. *et al.* (2017). 'Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives', *Molecular Plant*, 10(8), pp. 1047–1064. doi: 10.1016/j.molp.2017.06.008.

Rees, H. *et al.* (2022). 'Circadian regulation of the transcriptome in a complex polyploid crop', *PLoS Biology*, 20(10), p. e3001802. doi: https://doi.org/10.1371/journal.pbio.3001802.

Reif, J. C. *et al.* (2005). 'Wheat genetic diversity trends during domestication and breeding', *Theoretical and Applied Genetics*, 110, pp. 859–864. doi: 10.1007/S00122-004-1881-8.

Ren, T. *et al.* (2009). 'Development and characterization of a new 1BL.1RS translocation line with resistance to stripe rust and powdery mildew of wheat', *Euphytica*, 169, pp. 207–213. doi: 10.1007/s10681-009-9924-5.

Rey, M-D. *et al.* (2017). 'Exploiting the *ZIP4* homologue within the wheat *Ph1* locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids', *Molecular Breeding*, 37(95). doi: 10.1007/s11032-017-0700-2.

Reynolds, M. *et al.* (2020). 'Breeder friendly phenotyping', *Plant Science*, 295(110396). doi: 10.1016/J.PLANTSCI.2019.110396.

Reynolds, M. *et al.* (2012). 'Achieving yield gains in wheat', *Plant, Cell & Environment*, 35(10), pp. 1799–1823. doi: https://doi.org/10.1111/j.1365-3040.2012.02588.x.

Reynolds, M., Dreccer, F. and Trethowan, R. (2007). 'Drought-adaptive traits derived from wheat wild relatives and landraces', *Journal of Experimental Botany*, 58(2), pp. 177–186. doi:

10.1093/JXB/ERL250.

Reynolds, M. P. *et al.* (1994). 'Physiological and morphological traits associated with spring wheat yield under hot, irrigated conditions', *Australian Journal of Plant Physiology*, 21, pp. 717–730. doi: 10.1071/PP9940717.

Reynolds, M. P. *et al.* (2001). 'Physiological Basis of Yield Gains in Wheat Associated with the *Lr19* Translocation from *Agropyron Elongatum*, in Bedö, Z. and Láng, L. (eds) *Wheat in a Global Environment: Proceedings of the 6th International Wheat Conference*, 5–9 June 2000, Budapest, Hungary', Dordrecht: Springer Netherlands, pp. 345–351. doi: 10.1007/978-94-017-3674-9_44.

Reynolds, M. P. *et al.* (2016). 'An integrated approach to maintaining cereal productivity under climate change', *Global Food Security*. Elsevier, 8, pp. 9–18. doi: 10.1016/J.GFS.2016.02.002.

Reynolds, Matthew P *et al.* (2017). 'Strategic crossing of biomass and harvest index—source and sink—achieves genetic gains in wheat', *Euphytica*, 213(257). doi: 10.1007/s10681-017-2040-z.

Rice, P., Longden, L. and Bleasby, A. (2000). 'EMBOSS: the European Molecular Biology Open Software Suite', *Trends in Genetics*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

Roach, M. J., Schmidt, S. A. and Borneman, A. R. (2018). 'Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies', *BMC Bioinformatics*, 19(460). doi: 10.1186/s12859-018-2485-7.

Robinson, J. T. *et al.* (2011). 'Integrative genomics viewer', *Nature Biotechnology*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.

Romanowski, A. *et al.* (2020). 'Global transcriptome analysis reveals circadian control of splicing events in *Arabidopsis thaliana*', *The Plant Journal*, 103(2), pp. 889–902. doi: https://doi.org/10.1111/tpj.14776.

Rosyara, U. *et al.* (2019). 'Genetic Contribution of Synthetic Hexaploid Wheat to CIMMYT's Spring Bread Wheat Breeding Germplasm', *Scientific Reports*, 9(12355). doi: 10.1038/s41598-019-47936-5.

Ruan, J. and Li, H. (2020). 'Fast and accurate long-read assembly with wtdbg2', *Nature Methods*, 17, pp. 155–158. doi: 10.1038/s41592-019-0669-3.

Ruzzante, S., Labarta, R. and Bilton, A. (2021). 'Adoption of agricultural technology in the developing world: A meta-analysis of the empirical literature', *World Development*, 146(105599). doi: https://doi.org/10.1016/j.worlddev.2021.105599.

Sayre, K. D., Rajaram, S. and Fischer, R. A. (1997). 'Yield Potential Progress in Short Bread Wheats in Northwest Mexico', *Crop Science*, 37, pp. 36–42. doi: 10.2135/CROPSCI1997.0011183X003700010006X.

Sears, E. R. (1956). 'The transfer of leaf-rust resistance from Aegilops umbellulata to wheat.', *Brookhaven Symposia in Biology*, 9, pp. 1-21.

Sehgal, D. *et al.* (2015). 'Exploring and Mobilizing the Gene Bank Biodiversity for Wheat Improvement', *PLoS ONE*, 10(7), p. e0132112. doi: 10.1371/JOURNAL.PONE.0132112.

Shamanin, V. *et al.* (2019). 'Primary hexaploid synthetics: Novel sources of wheat disease resistance', *Crop Protection*, 121, pp. 7–10. doi: https://doi.org/10.1016/j.cropro.2019.03.003.

Shebeski, L. H. and Wu, Y. S. (1952). 'Inheritance in Wheat of Stem Rust Resistance Derived from Agropyron Elongatum', *Scientific Agriculture*, 32(1), pp. 26–35. doi: 10.4141/sa-1952-0003.

Shenoda, J. E. *et al.* (2021). 'Effect of long-term heat stress on grain yield, pollen grain viability and germinability in bread wheat (*Triticum aestivum* L.) under field conditions', *Heliyon*, 7(6), p. e07096. doi: 10.1016/j.heliyon.2021.e07096.

Shi, X. *et al.* (2022). 'Comparative genomic and transcriptomic analyses uncover the molecular basis of high nitrogen-use efficiency in the wheat cultivar Kenong 9204', *Molecular Plant*, 15(9), pp. 1440–1456. doi: 10.1016/j.molp.2022.07.008.

Shiferaw, B. *et al.* (2013). 'Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security', *Food Security*, 5, pp. 291–317. doi: 10.1007/s12571-013-0263-y.

Sibbesen, J. A. *et al.* (2022). 'Haplotype-aware pantranscriptome analyses using spliced pangenome graphs', *bioRxiv*. doi: 10.1101/2021.03.26.437240.

Singh, K. *et al.* (2021). 'WheatQTLdb: a QTL database for wheat', *Molecular Genetics and Genomics*, 296, pp. 1051–1056. doi: 10.1007/s00438-021-01796-9.

Singh, R. P. *et al.* (2015). 'Emergence and Spread of New Races of Wheat Stem Rust Fungus: Continued Threat to Food Security and Prospects of Genetic Control', *Phytopathology*, 105(7), pp. 872–884. doi: 10.1094/PHYTO-01-15-0030-FI.

Singh, Sukhwinder *et al.* (2018). 'Harnessing genetic potential of wheat germplasm banks through impact-oriented-prebreeding for future food and nutritional security', *Scientific Reports*, 8(12527). doi: 10.1038/S41598-018-30667-4.

Singh, S. *et al.* (2021). 'Direct introgression of untapped diversity into elite wheat lines', *Nature Food*, 2, pp. 819–827. doi: 10.1038/s43016-021-00380-z.

Sinha, S. K. and Kumar, K. R. R. (2022). 'Heat Stress in Wheat: Impact and Management Strategies Towards Climate Resilience', in Roy, S. *et al*. (eds) *Plant Stress: Challenges and Management in the New Decade*. Springer International Publishing, pp. 199–214. doi: 10.1007/978-3-030-95365-2_13.

Slater, G. S. C. and Birney, E. (2005). 'Automated generation of heuristics for biological sequence comparison', *BMC Bioinformatics*, 6(31). doi: 10.1186/1471-2105-6-31.

Song, L., Sabunciyan, S. and Florea, L. (2016). 'CLASS2: accurate and efficient splice variant annotation from RNA-seq reads', *Nucleic Acids Research*, 44(10), p. e98. doi: 10.1093/nar/gkw158.

Srivastava, A. *et al.* (2020). 'Alignment and mapping methodology influence transcript abundance estimation', *Genome Biology*, 21(239). doi: 10.1186/s13059-020-02151-8.

Steuernagel, B. *et al.* (2020). 'The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire.', *Plant physiology*, 183(2), pp. 468–482. doi: 10.1104/pp.19.01273.

Sukumaran, S. *et al.* (2018). 'Genetic analysis of multi-environmental spring wheat trials identifies genomic regions for locus-specific trade-offs for grain weight and grain number', *Theoretical and Applied Genetics*, 131(4), pp. 985–998. doi: 10.1007/s00122-017-3037-7.

Surówka, E., Rapacz, M. and Janowiak, F. (2020). 'Climate change influences the interactive effects of simultaneous impact of abiotic and biotic stresses on plants', in Hasanuzzaman, M. (ed) *Plant Ecophysiology and Adaptation under Climate Change: Mechanisms and Perspectives I*, Singapore: Springer Singapore, pp. 1–50. doi: 10.1007/978-981-15-2156-0_1.

Tanksley, S. D. and McCouch, S. R. (1997). 'Seed banks and molecular maps: Unlocking genetic potential from the wild', *Science*, 277(5329), pp. 1063–1066. doi: 10.1126/science.277.5329.1063.

Tao, Y. Z. *et al.* (2000). 'Identification of genomic regions associated with stay green in sorghum by testing RILs in multiple environments', *Theoretical and Applied Genetics*, 100, pp. 1225–1232. doi: 10.1007/s001220051428.

Thapa, S. *et al.* (2020). 'Impacts of sowing and climatic conditions on wheat yield in Nepal', *Malaysian Journal of Halal Research*, 3(1), pp. 38–40. doi: 10.2478/mjhr-2020-0006.

The UniProt Consortium (2019). 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids*

Research, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Thorburn, D-M. J. *et al.* (2023). 'Origin matters: Using a local reference genome improves measures in population genomics', *Molecular Ecology Resources*, 23(7), pp. 1706–1723. doi: https://doi.org/10.1111/1755-0998.13838.

Trapnell, C. *et al.* (2012). 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7, pp. 562–578. doi: 10.1038/nprot.2012.016.

Trethowan, R. M. and Mahmood, T. (2011). 'Genetics Options for Improving the Productivity of Wheat in Water-Limited and Temperature-Stressed Environments', in Yadav, S. S. *et al*. (eds) *Crop Adaptation to Climate Change*, Oxford, UK: Wiley-Blackwell, pp. 218–237. doi: 10.1002/9780470960929.ch16.

Tyagi, S. *et al.* (2014). 'Marker-assisted pyramiding of eight QTLs/genes for seven different traits in common wheat (*Triticum aestivum* L.)', *Molecular Breeding*, 34, pp. 167–175. doi: 10.1007/s11032-014-0027-1.

Ullah, A. *et al.* (2022). 'Heat stress effects on the reproductive physiology and yield of wheat', *Journal of Agronomy and Crop Science*, 208(1), pp. 1–17. doi: 10.1111/jac.12572.

United Nations, Department of Economic and Social Affairs, Population Division (2019). 'World Population Prospects 2019: Highlights'. ISBN: 978-92-1-148316-1.

Valkoun, J. J. (2001). 'Wheat pre-breeding using wild progenitors', *Euphytica*, 119, pp. 17–23. doi: 10.1023/A:1017562909881.

Vavilov, N. I. (1940). 'The new systematics of cultivated plants', in Huxley, J. (ed) *The New Systematics*, Oxford: Oxford University Press.

Vavilov, N. I. (1926). 'Centers of Origin of Cultivated Plants', *Inst. Appl. Bot. Plant breed.*, 16(2).

Venske, E. *et al.* (2019). 'Bread wheat: a role model for plant domestication and breeding', *Hereditas*, 156(16). doi: 10.1186/s41065-019-0093-9.

Venturini, L. *et al.* (2018). 'Leveraging multiple transcriptome assembly methods for improved gene structure annotation', *GigaScience*, 7(8), p. giy093. doi: 10.1093/gigascience/giy093.

Vijaya Satya, R., Zavaljevski, N. and Reifman, J. (2012). 'A new strategy to reduce allelic bias in RNA-Seq readmapping', *Nucleic Acids Research*, 40(16), p. e127. doi: 10.1093/nar/gks425.

Vikram, P. *et al.* (2016). 'Unlocking the genetic diversity of Creole wheats', *Scientific Reports*, 6(23092). doi: 10.1038/srep23092.

Villa, T. C. C. *et al.* (2005). 'Defining and identifying crop landraces', *Plant Genetic Resources*, 3(3), pp. 373–384. doi: DOI: 10.1079/PGR200591.

Walker, B. J. *et al.* (2014). 'Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.', *PloS one*, 9(11), p. e112963. doi: 10.1371/journal.pone.0112963.

Walkowiak, S. *et al.* (2020). 'Multiple wheat genomes reveal global variation in modern breeding', *Nature*, 588, pp. 277–283. doi: 10.1038/s41586-020-2961-x.

Wang, H. *et al.* (2020). 'An ankyrin-repeat and WRKY-domain-containing immune receptor confers stripe rust resistance in wheat', *Nature Communications*, 11(1353). doi: 10.1038/s41467-020-15139-6.

Wang, J. *et al.* (2016). '*TaELF3-1DL*, a homolog of *ELF3*, is associated with heading date in bread wheat', *Molecular Breeding*, 36(161). doi: 10.1007/s11032-016-0585-5.

Wang, N. *et al.* (2020). 'Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding', *Scientific Reports*, 10(16308). doi: 10.1038/s41598-020-73321-8.

Wang, W., Vinocur, B. and Altman, A. (2003). 'Plant responses to drought, salinity and extreme temperatures: Towards genetic engineering for stress tolerance', *Planta*, 218, pp. 1–14. doi: 10.1007/s00425-003-1105-5.

Wang, Z. *et al.* (2022). 'Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference', *Nature Communications*, 13(3891). doi: 10.1038/s41467-022-31581-0.

Warburton, M. L. *et al.* (2006). 'Bringing wild relatives back into the family: Recovering genetic diversity in CIMMYT improved wheat germplasm', *Euphytica*, 149, pp. 289–301. doi: https://doi.org/10.1007/s10681-005-9077-0.

Waterhouse, R. M. *et al.* (2018). 'BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.', *Molecular Biology and Evolution*, 35(3), pp. 543–548. doi: 10.1093/molbev/msx319.

Watson-Haigh, N. S. *et al.* (2018). 'DAWN: a resource for yielding insights into the diversity among wheat genomes', *BMC Genomics*, 19(941). doi: 10.1186/s12864-018-5228-2.

Wegren, S. K. (2011). 'Food Security and Russia's 2010 Drought', *Eurasian Geography and Economics*, 52(1), pp. 140–156. doi: 10.2747/1539-7216.52.1.140.

White, B. *et al.* (2024). 'De novo annotation of the wheat pan-genome reveals complexity and diversity within the hexaploid wheat pantranscriptome', *bioRxiv*. doi: 10.1101/2024.01.09.574802.

Wicker, T. *et al.* (2018). 'Impact of transposable elements on genome structure and evolution in bread wheat', *Genome Biology*, 19(103). doi: 10.1186/s13059-018-1479-0.

Williamson, V. M. *et al.* (2013). 'An *Aegilops ventricosa* Translocation Confers Resistance Against Root-knot Nematodes to Common Wheat', *Crop Science*, 53(4), pp. 1412–1418. doi: https://doi.org/10.2135/cropsci2012.12.0681.

Winfield, M. O. *et al.* (2016). 'High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool', *Plant Biotechnology Journal*, 14(5), pp. 1195–1206. doi: 10.1111/pbi.12485.

Wingen, L. U. *et al.* (2014). 'Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat', *Theoretical and Applied Genetics*, 127, pp. 1831–1842. doi: 10.1007/S00122-014-2344-5/FIGURES/3.

Winter, D. (2021). 'pafr: Read, Manipulate and Visualize "Pairwise mApping Format" Data'. Available at: https://dwinter.github.io/pafr/.

Wittern, L. *et al.* (2023). 'Wheat *EARLY FLOWERING 3* affects heading date without disrupting circadian oscillations, *Plant Physiology*, 191(2), pp. 1383-1403. doi: https://doi.org/10.1093/plphys/kiac544.

Wu, D. C. *et al.* (2018). 'Limitations of alignment-free tools in total RNA-seq quantification', *BMC Genomics*, 19(510). doi: 10.1186/s12864-018-4869-5.

Wu, G. *et al.* (2016). 'MetaCycle: an integrated R package to evaluate periodicity in large scale data', *Bioinformatics*, 32(21), pp. 3351–3353. doi: 10.1093/bioinformatics/btw405.

Yang, Z-J. *et al.* (2009). 'Molecular cytogenetic characterization of wheat–*Secale africanum* amphiploids and derived introgression lines with stripe rust resistance', *Euphytica*, 167, pp. 197–202. doi: 10.1007/s10681-008-9861-8.

Yasumuro, Y. *et al.* (1981). 'INDUCED PAIRING BETWEEN A WHEAT (*TRITICUM AESTIVUM*) AND AN *AGROPYRON ELONGATUM* CHROMOSOME ', *Canadian Journal of Genetics and Cytology*, 23(1),

pp. 49–56. doi: 10.1139/g81-006.

Zadoks, J. C., Chang, T. T. and Konzak, C. F. (1974). 'A decimal code for the growth stages of cereals', *Weed Research*, 14(6), pp. 415–421. doi: 10.1111/J.1365-3180.1974.TB01084.X.

Zahedi, M. and Jenner, C. F. (2003). 'Analysis of effects in wheat of high temperature on grain filling attributes estimated from mathematical models of grain filling', The *Journal of Agricultural Science*, 141(2), pp. 203–212. doi: 10.1017/S0021859603003411.

Zeven, A. C. (1998). 'Landraces: A review of definitions and classifications', *Euphytica*, 104, pp. 127–139. doi: 10.1023/A:1018683119237.

Zhan, S., Griswold, C. and Lukens, L. (2021). '*Zea mays* RNA-seq estimated transcript abundances are strongly affected by read mapping bias', *BMC Genomics*, 22(285). doi: 10.1186/S12864-021-07577-3/TABLES/4.

Zhang, J. *et al.* (2021). 'A recombined *Sr26* and *Sr61* disease resistance gene stack in wheat encodes unrelated *NLR* genes', *Nature Communications*, 12(3378). doi: 10.1038/s41467-021-23738-0.

Zhang, P. *et al.* (2017). 'Chromosome Engineering Techniques for Targeted Introgression of Rust Resistance from Wild Wheat Relatives', in Periyannan, S. (ed) *Wheat Rust Diseases: Methods and Protocols*, New York: Springer New York, pp. 163–172. doi: 10.1007/978-1-4939-7249-4_14.

Zhang, X. and Cai, X. (2011). 'Climate change impacts on global agricultural land availability', *Environmental Research Letters*, 6(014014). doi: 10.1088/1748-9326/6/1/014014.

Zheng, S. *et al.* (2020). 'Characterization and diagnostic marker development for *Yr28-rga1* conferring stripe rust resistance in wheat', *European Journal of Plant Pathology*, 156(2), pp.623-634. doi: 10.1007/s10658-019-01912-x.

Zhou, Yongbin *et al.* (2020). 'Overexpression of soybean *DREB1* enhances drought stress tolerance of transgenic wheat in the field', *Journal of experimental botany*, 71(6), pp. 1842–1857. doi: 10.1093/JXB/ERZ569.

Zhou, Yao *et al.* (2020). '*Triticum* population sequencing provides insights into wheat adaptation', *Nature Genetics*, 52, pp. 1412–1422. doi: 10.1038/s41588-020-00722-w.

Zhou, Y. *et al.* (2021). 'Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement', *Nature Plants*, 7, pp. 774–786. doi: 10.1038/s41477-021-00934-w.

Zhu, T. *et al.* (2021). 'Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring

genome assembly', *The Plant Journal*, 107(1), pp. 303–314. doi: https://doi.org/10.1111/tpj.15289.

Zhu, Z. *et al.* (2014). 'Mapping resistance to spot blotch in a CIMMYT synthetic-derived bread wheat', *Molecular Breeding*, 34, pp. 1215–1228. doi: 10.1007/s11032-014-0111-6.

Zhu, Z. *et al.* (2016). 'Characterization of Fusarium head blight resistance in a CIMMYT synthetic-derived bread wheat line', *Euphytica*, 208, pp. 367–375. doi: 10.1007/s10681-015-1612-z.

Zimin, A. V *et al.* (2017). 'The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*', *GigaScience*, 6(11), p. gix097. doi: 10.1093/gigascience/gix097.

Appendix A

Appendix A1. Introgressed *Am. muticum* segments identified in 17 introgression lines

Line name	chromosome	start	end	length	precise identification left junction	precise identification right junction
DH8	chr2D	591023410	651852609	60.83 Mbp	yes - TraesCS2D02G493600	no
DH15	chr2A	11934615	780798557	768.86 Mbp	difficult to resolve due to structural complexity	no - probably not telomere
DH15	chr4B	by 300000 but not telomere	635868588	635.08 Mbp	no but not telomere	yes - TraesCS4B02G342400
DH15	chr6D	470634719	473592718	3.99 Mbp	yes - TraesCS6D02G401600	yes - telomere
DH65	chr4D	1	51288177	51.29 Mbp	yes - telomere	yes - TraesCS4D02G076900
DH86	chr2D	1	24800000	24.8 Mbp	no	no
DH92	chr5D	533204927	566080677	32.88 Mbp	yes - TraesCS5D02G508200	yes - telomere
DH96	chr4D	1	45649275	45.65 Mbp	yes - telomere	yes - TraesCS4D02G070200

Appendix A1.

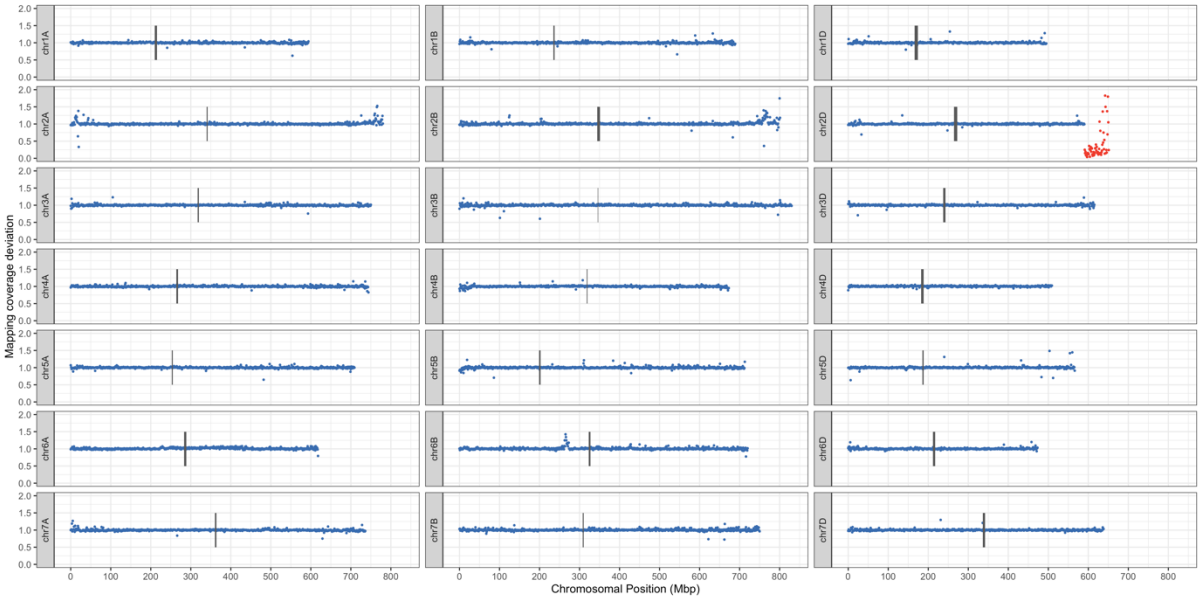
DH121	chr4D	1	378269537	378.27 Mbp	yes - telomere	yes - TraesCS4D02G221100
DH121	chr5D	543405627	566080677	22.68 Mbp	no	yes - telomere
DH121	chr7D	59705175	638686055	578.98 Mbp	yes	yes - telomere
DH123	chr7D	59705175	638686055	578.98 Mbp	yes	yes - telomere
DH161	chr1A	1	594102056	594.1 Mbp	yes - telomere	yes - telomere
BC2F420	chr2D	591023474	651852609	60.83 Mbp	yes - TraesCS2D02G493600	yes - telomere
BC3F326	chr3D	395642516	615552423	219.90 Mbp	yes - TraesCS3D02G286600	yes - telomere
DH195	chr7D	1	500819199	500.82 Mbp	yes - telomere	yes - TraesCS7D02G386500
DH195	chr7D	621299250	638686055	17.39 Mbp	yes - TraesCS7D02G525800	yes - telomere
DH202	chr7D	1	500819199	500.82 Mbp	yes - telomere	yes - TraesCS7D02G386500
BC3F45	chr3A	708695000	750843639	42.14 Mbp	no	yes - telomere

Appendix A1.

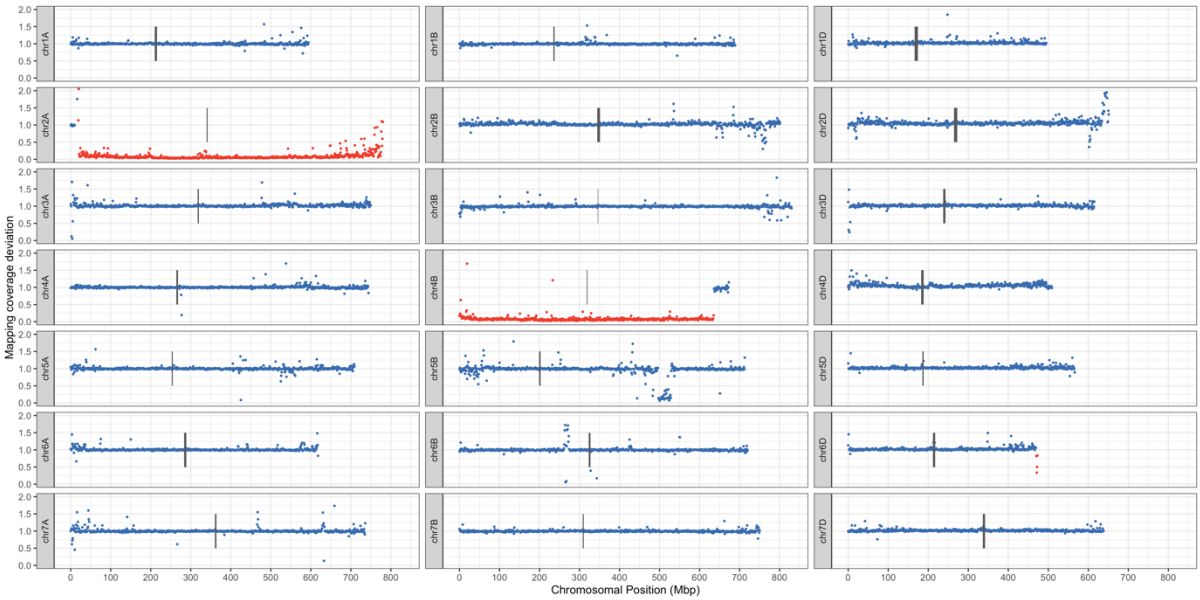
BC3F45	chr5D	1	399328510	399.33 Mbp	yes - telomere	yes - TraesCS5D02G304200
BC3F46	chr3D	1000000-2000000	549713134	549.71 Mbp	no - difficult to resolve due to possible repeat expansions or duplications	yes - TraesCS3D02G439400 (6.75 Kbp away)

Appendix A2. Macro-level plots for each introgression line studied in chapter two. Each dot shows the mapping coverage deviation value of a 1 Mbp genomic window compared to the wheat parent lines. Red dots are windows within a block identified as an *Am. muticum* introgression. The vertical black bars represent the position of the centromeres, predicted in Appels et al. (2018).

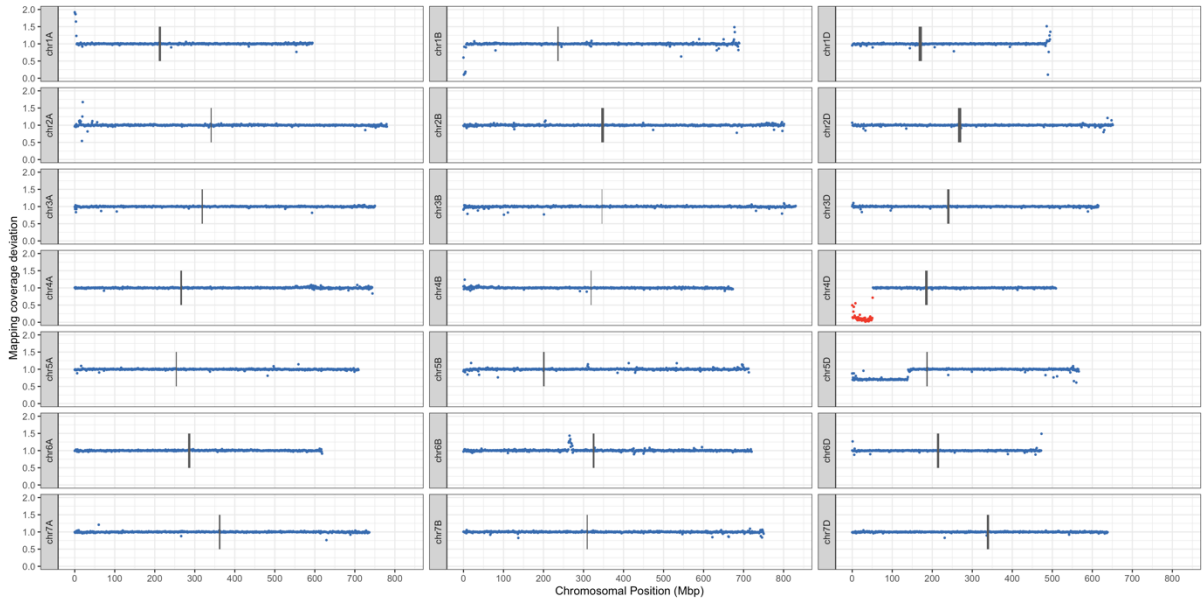
DH8


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

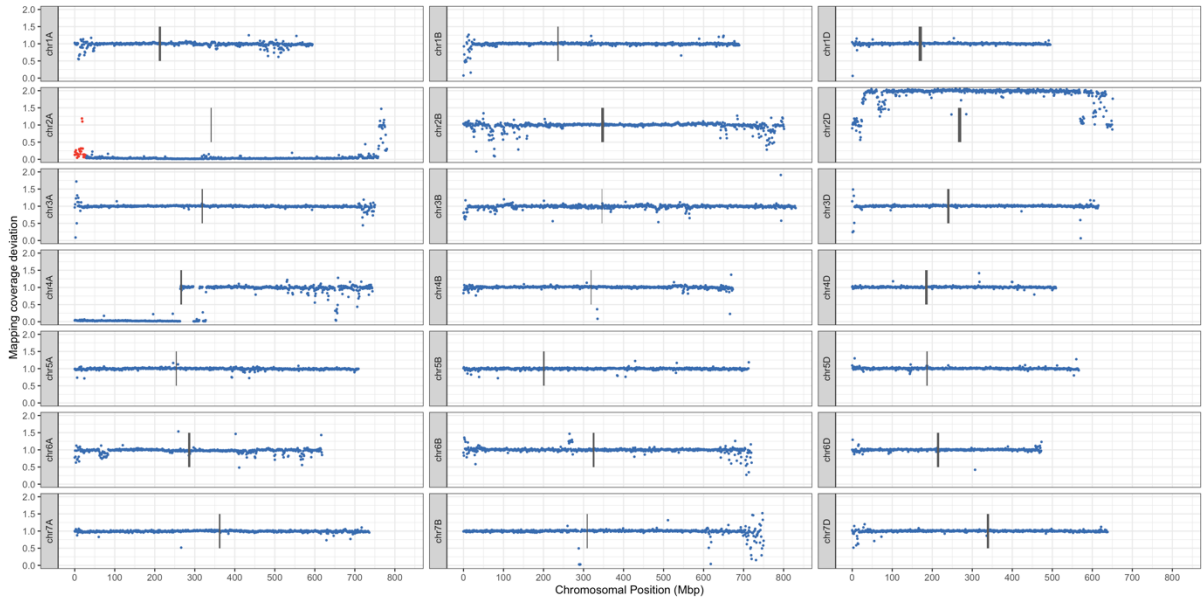
DH15


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

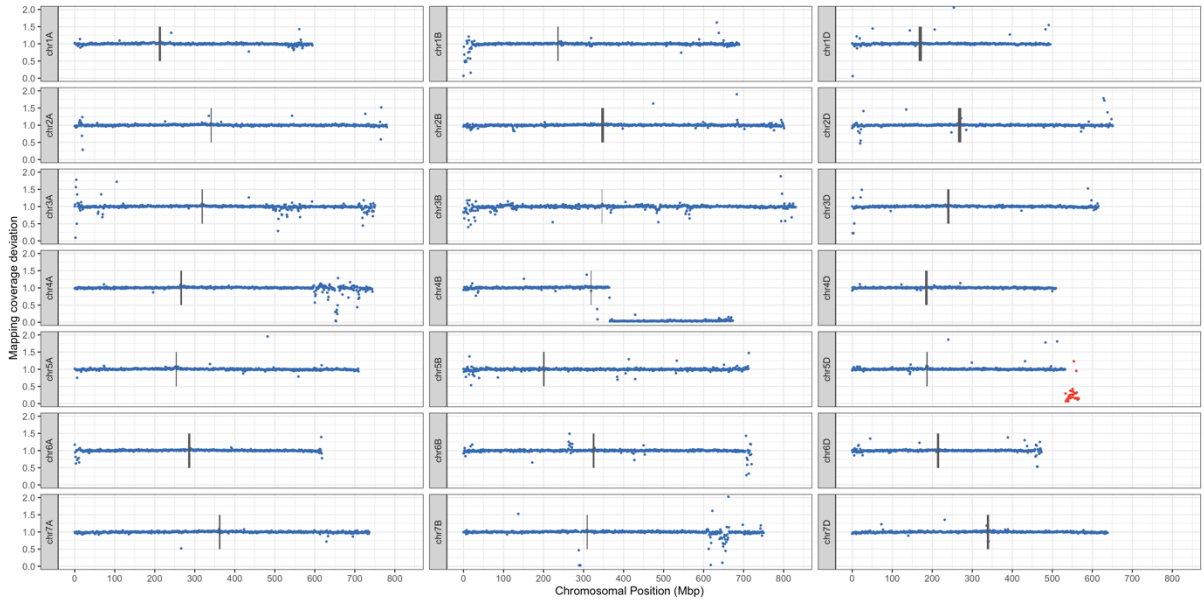
DH65


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

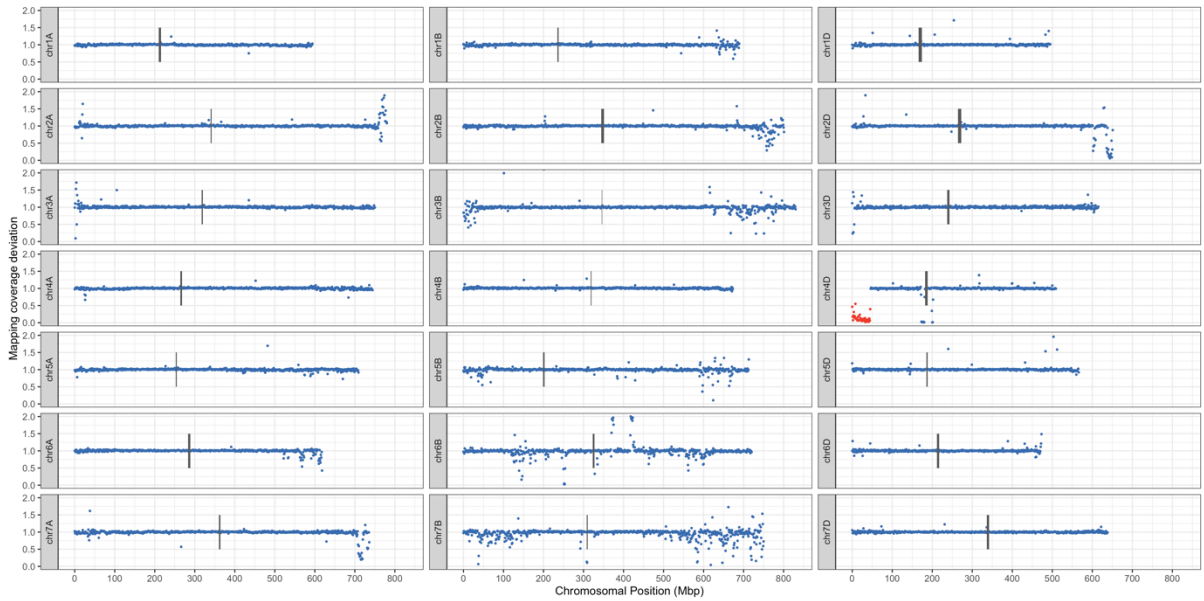
DH86


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

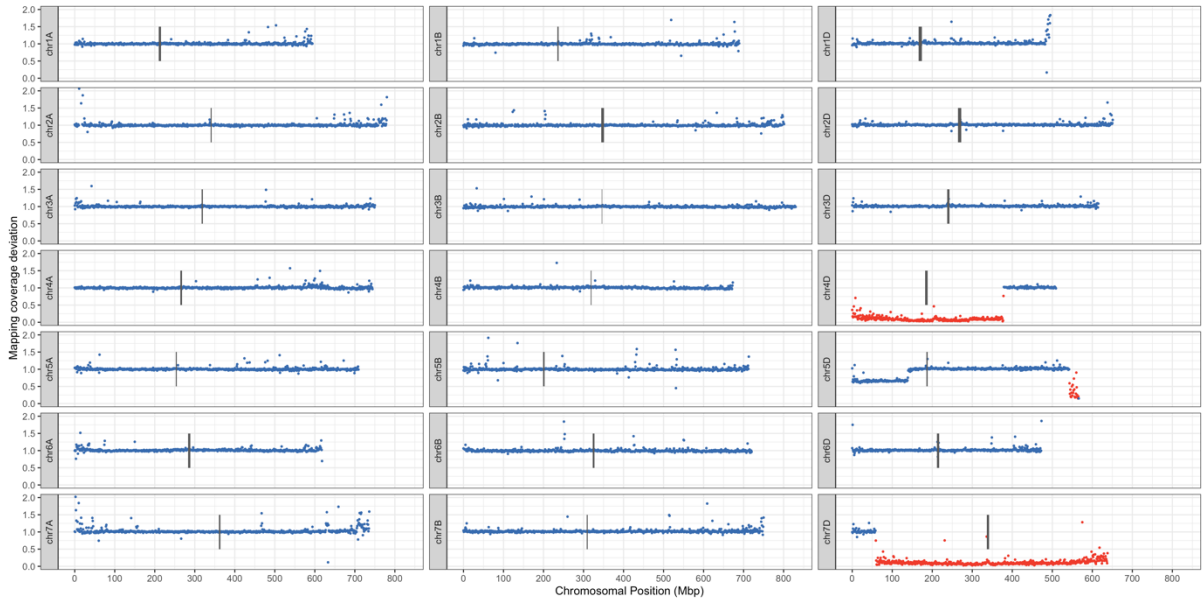
DH92


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

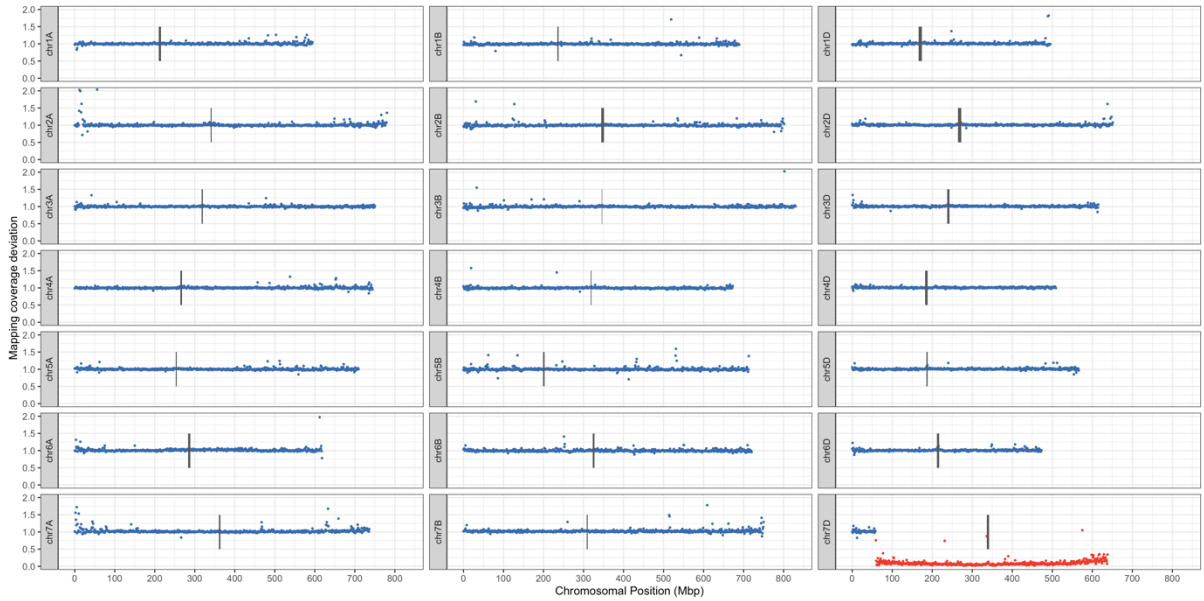
DH96


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

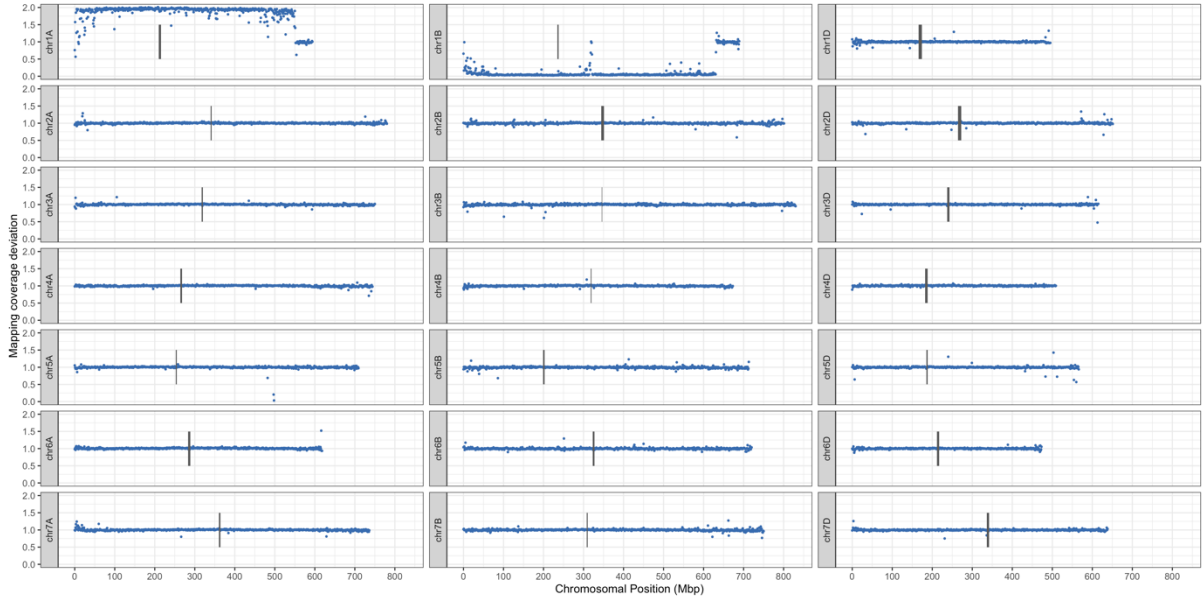
DH121


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

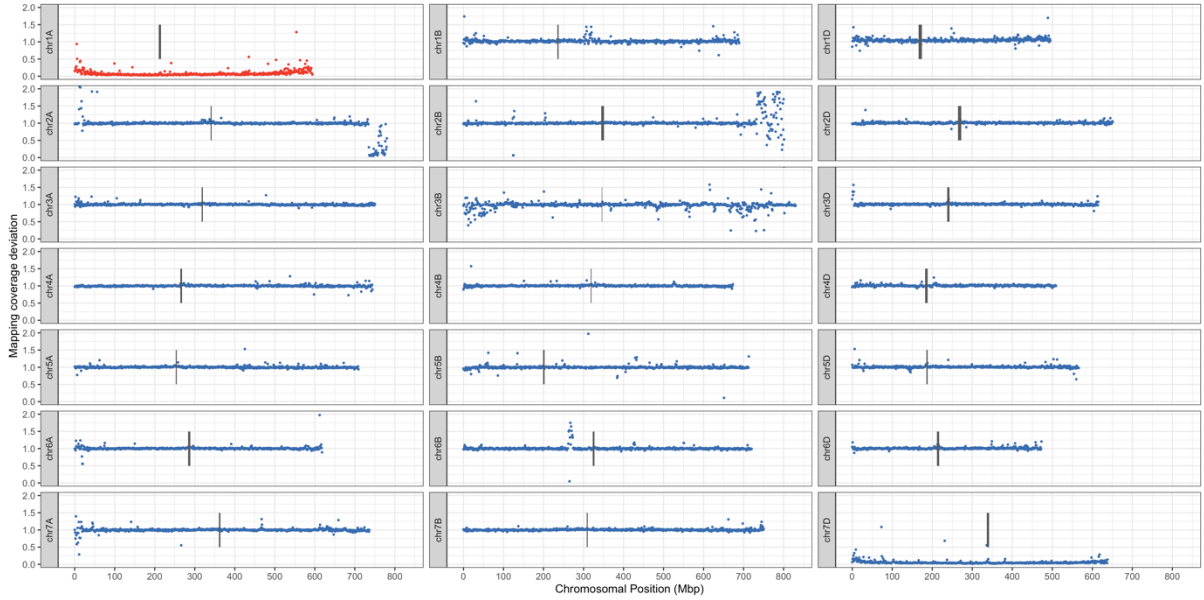
DH123


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

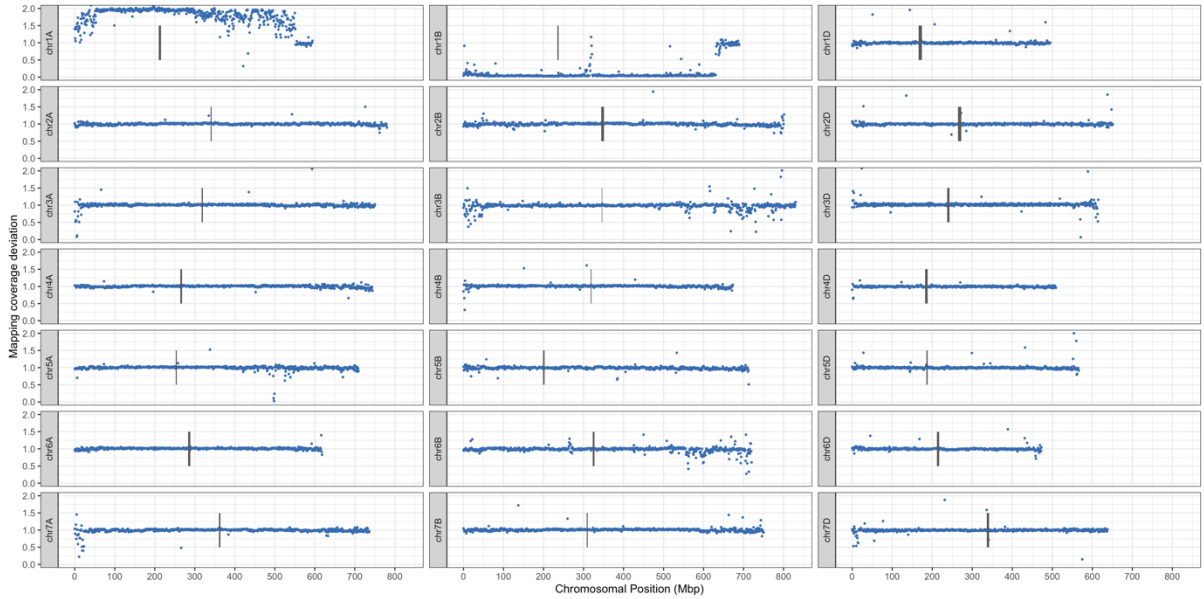
DH124


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

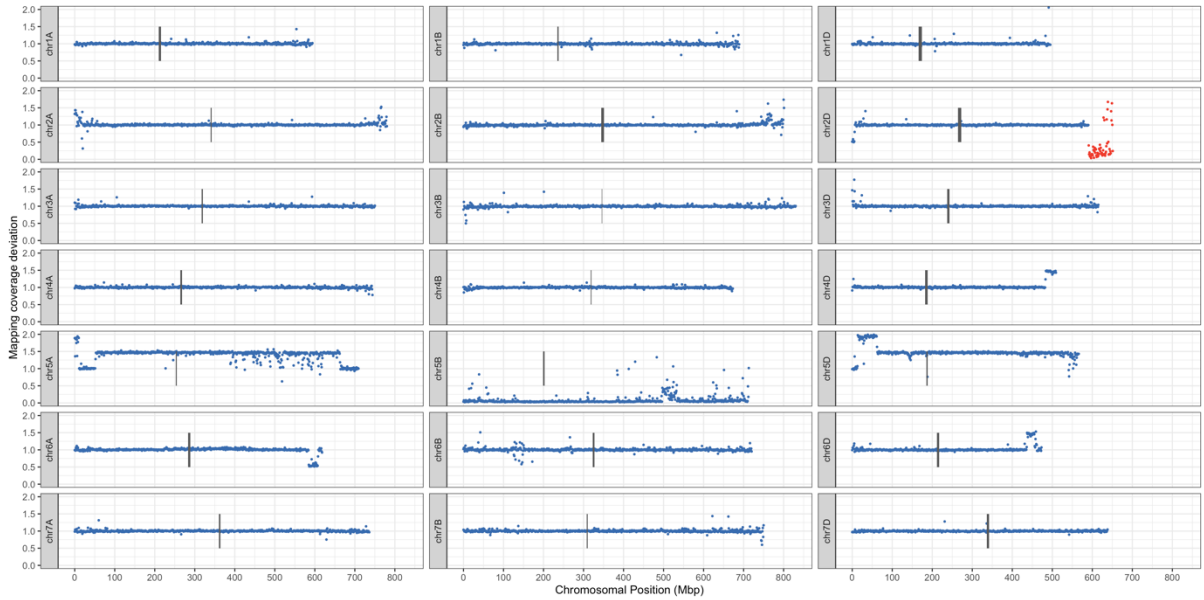
DH161


Mapping coverage deviation
Chromosomal Position (Mbp)
chr1A
chr1B
chr1D
chr2A
chr2B
chr2D
chr3A
chr3B
chr3D
chr4A
chr4B
chr4D
chr5A
chr5B
chr5D
chr6A
chr6B
chr6D
chr7A
chr7B
chr7D

DH355



BC2F420



Appendix A3. Primer details for the KASP™ assays used for genotyping of DH15

KASP_ID	Allele_specific_primer_1	Allele_specific_primer_2	Common_primer
WRC1890	gtatggtcatacgtgatagcC	cggtatggtcatacgtgatagcT	Gcagttcccgccgaaataaa
WRC1873	ccaccactctctcaagtaaggC	ccaccactctctcaagtaaggT	Gcatgcagcagctttgagtc

Appendix B

Appendix B1. *Ae. tauschii* genes present in five *Ae. tauschii* genome assemblies, within the core introgressed region underlying the chr6D MTA for heat tolerance traits.

Accession name	Gene	Gene type
AL8/78	AET6Gv20022000	NLR
AL8/78	AET6Gv20022200	NLR
AL8/78	AET6Gv20023100	Unknown
AL8/78	AET6Gv20023900	Potential DNA-binding domain
AL8/78	AET6Gv20024100	Pyridoxal-dependent decarboxylase
AL8/78	AET6Gv20024800	Unknown
AL8/78	AET6Gv20025000	Protein of unknown function
AL8/78	AET6Gv20025100	Unknown
AL8/78	AET6Gv20025200	Unknown
AL8/78	AET6Gv20025300	Unknown
AL8/78	AET6Gv20025500	Ribosome inactivating protein
AL8/78	AET6Gv20025600	MIKC-type MADS-box transcription factor
AL8/78	AET6Gv20025700	Two-component response regulator
AL8/78	AET6Gv20025800	Two-component response regulator
AL8/78	AET6Gv20026000	WS/DGAT C-terminal domain
AL8/78	AET6Gv20026400	Unknown
AL8/78	AET6Gv20026500	Pyridoxal-dependent decarboxylase
AL8/78	AET6Gv20026600	Alginate lyase
AL8/78	AET6Gv20027000	Cytochrome P450
AL8/78	AET6Gv20027200	Cytochrome P450
AL8/78	AET6Gv20027300	NmrA-like family
AL8/78	AET6Gv20027400	LRR
AL8/78	AET6Gv20027600	LRR
AL8/78	AET6Gv20027700	LRR
AL8/78	AET6Gv20027800	LRR
AL8/78	AET6Gv20027900	FAR1 DNA-binding domain
AL8/78	AET6Gv20028000	Unknown

Appendix B1.

AY17	AetAY17_6Dv1G017100	unknown
AY17	AetAY17_6Dv1G017200	NLR
AY17	AetAY17_6Dv1G017300	LRR
AY17	AetAY17_6Dv1G017400	None
AY17	AetAY17_6Dv1G017600	NLR
AY17	AetAY17_6Dv1G017900	Potential DNA-binding domain
AY17	AetAY17_6Dv1G018300	Pyridoxal-dependent decarboxylase
AY17	AetAY17_6Dv1G018500	RNA recognition motif
AY17	AetAY17_6Dv1G018600	Unknown
AY17	AetAY17_6Dv1G018700	None
AY17	AetAY17_6Dv1G018800	Unknown
AY17	AetAY17_6Dv1G018900	Robisome inactivating protein
AY17	AetAY17_6Dv1G019000	MIKC-type MADS-box transcription factor
AY17	AetAY17_6Dv1G019100	Two-component response regulator
AY17	AetAY17_6Dv1G019200	WS/DGAT C-terminal domain
AY17	AetAY17_6Dv1G019300	Pyridoxal-dependent decarboxylase
AY17	AetAY17_6Dv1G019600	Two-component response regulator
AY17	AetAY17_6Dv1G019700	Cytochrome P450
AY17	AetAY17_6Dv1G019800	Mitogen-activated Protein Kinase
AY17	AetAY17_6Dv1G019900	LigB
AY17	AetAY17_6Dv1G020000	Cytochrome P450
AY17	AetAY17_6Dv1G020100	NmrA-like family
AY17	AetAY17_6Dv1G020200	LRR
AY17	AetAY17_6Dv1G020300	LRR
AY17	AetAY17_6Dv1G020400	NLR
AY17	AetAY17_6Dv1G020600	Myb/SANT-like DNA-binding domain
AY61	AetAY61_6Dv1G0024900	Cytochrome P450
AY61	AetAY61_6Dv1G0024300	Cytochrome P450
AY61	AetAY61_6Dv1G0024700	Cytochrome P450
AY61	AetAY61_6Dv1G0024100	GRF zinc finger

Appendix B1.

AY61	AetAY61_6Dv1G0023600	MIKC-type MADS-box transcription factor
AY61	AetAY61_6Dv1G0024600	LigB
AY61	AetAY61_6Dv1G0021600	LRR
AY61	AetAY61_6Dv1G0025500	LRR
AY61	AetAY61_6Dv1G0025200	LRR
AY61	AetAY61_6Dv1G0026000	Myb/SANT-like DNA-binding domain
AY61	AetAY61_6Dv1G0021000	NLR
AY61	AetAY61_6Dv1G0021400	NLR
AY61	AetAY61_6Dv1G0022700	Potential DNA-binding domain
AY61	AetAY61_6Dv1G0022300	Potential DNA-binding domain
AY61	AetAY61_6Dv1G0023000	Protein of unknown function
AY61	AetAY61_6Dv1G0023900	Pyridoxal-dependent decarboxylase
AY61	AetAY61_6Dv1G0022500	Pyridoxal-dependent decarboxylase
AY61	AetAY61_6Dv1G0025700	Retrotransposon gag protein
AY61	AetAY61_6Dv1G0023500	Ribosome inactivating protein
AY61	AetAY61_6Dv1G0023200	RNA recognition motif
AY61	AetAY61_6Dv1G0024200	Two-component response regulator
AY61	AetAY61_6Dv1G0021700	Unknown
AY61	AetAY61_6Dv1G0022900	Unknown
AY61	AetAY61_6Dv1G0023400	Unknown
AY61	AetAY61_6Dv1G0021100	Unknown
AY61	AetAY61_6Dv1G0023300	Unknown
AY61	AetAY61_6Dv1G0022000	Unknown
AY61	AetAY61_6Dv1G0021200	Unknown
AY61	AetAY61_6Dv1G0022200	Unknown
AY61	AetAY61_6Dv1G0021500	Unknown
AY61	AetAY61_6Dv1G0020700	Unknown
AY61	AetAY61_6Dv1G0025600	Unknown
AY61	AetAY61_6Dv1G0023800	WS/DGAT C-terminal domain
XJ02	AetXJ02_6Dv1G022300	Cytochrome P450

Appendix B1.

XJ02	AetXJ02_6Dv1G022500	Cytochrome P450
XJ02	AetXJ02_6Dv1G021800	MIKC-type MADS-box transcription factor
XJ02	AetXJ02_6Dv1G023100	LRR
XJ02	AetXJ02_6Dv1G019100	LRR
XJ02	AetXJ02_6Dv1G022900	LRR
XJ02	AetXJ02_6Dv1G023500	Myb-SANT-like DNA-binding domain
XJ02	AetXJ02_6Dv1G019900	NLR
XJ02	AetXJ02_6Dv1G019300	NLR
XJ02	AetXJ02_6Dv1G020000	NLR
XJ02	AetXJ02_6Dv1G019000	NLR
XJ02	AetXJ02_6Dv1G019400	NLR
XJ02	AetXJ02_6Dv1G019600	NLR
XJ02	AetXJ02_6Dv1G022600	NmrA-like family
XJ02	AetXJ02_6Dv1G020600	Papain family cysteine protease
XJ02	AetXJ02_6Dv1G020400	Potential DNA-binding domain
XJ02	AetXJ02_6Dv1G021000	Protein of unknown function
XJ02	AetXJ02_6Dv1G023300	Proton-conducting membrane transporter
XJ02	AetXJ02_6Dv1G022100	Pyridoxal-dependent decarboxylase conserved domain
XJ02	AetXJ02_6Dv1G023200	Retrotransposon gag protein
XJ02	AetXJ02_6Dv1G021100	Ribosomal L28e
XJ02	AetXJ02_6Dv1G021700	Ribosome inactivating protein
XJ02	AetXJ02_6Dv1G021300	RNA recognition motif
XJ02	AetXJ02_6Dv1G019200	Unknown
XJ02	AetXJ02_6Dv1G021200	Unknown
XJ02	AetXJ02_6Dv1G021400	Unknown
XJ02	AetXJ02_6Dv1G021600	Unknown
XJ02	AetXJ02_6Dv1G021500	Unknown
XJ02	AetXJ02_6Dv1G020700	Unknown

Appendix B1.

XJ02	AetXJ02_6Dv1G020900	Unknown
XJ02	AetXJ02_6Dv1G022000	WS/DGAT C-terminal domain
T093	AetT093_6Dv1G029300	Cytochrome P450
T093	AetT093_6Dv1G029700	Cytochrome P450
T093	AetT093_6Dv1G028400	MIKC-type MADS-box transcription factor
T093	AetT093_6Dv1G029600	LigB
T093	AetT093_6Dv1G030100	LRR
T093	AetT093_6Dv1G030300	LRR
T093	AetT093_6Dv1G030700	Myb-SANT-like DNA-binding domain
T093	AetT093_6Dv1G029900	NmrA-like family
T093	AetT093_6Dv1G027200	Papain family cysteine protease
T093	AetT093_6Dv1G029500	Mitogen-activated Protein Kinase
T093	AetT093_6Dv1G027600	Protein of unknown function
T093	AetT093_6Dv1G028800	Pyridoxal-dependent decarboxylase
T093	AetT093_6Dv1G030400	Retrotransposon gag protein
T093	AetT093_6Dv1G027700	Ribosomal L28e protein family
T093	AetT093_6Dv1G028300	Ribosome inactivating protein
T093	AetT093_6Dv1G027900	RNA recognition motif
T093	AetT093_6Dv1G028600	Two-component response regulator
T093	AetT093_6Dv1G029200	Two-component response regulator
T093	AetT093_6Dv1G025300	Unknown
T093	AetT093_6Dv1G025800	Unknown
T093	AetT093_6Dv1G026800	Unknown
T093	AetT093_6Dv1G027000	Unknown
T093	AetT093_6Dv1G027300	Unknown
T093	AetT093_6Dv1G027500	Unknown
T093	AetT093_6Dv1G027800	Unknown
T093	AetT093_6Dv1G028000	Unknown
T093	AetT093_6Dv1G028100	Unknown
T093	AetT093_6Dv1G028200	Unknown
T093	AetT093_6Dv1G029100	Unknown

Appendix B1.

T093	AetT093_6Dv1G030500	Unknown
T093	AetT093_6Dv1G028700	WS-DGAT C-terminal domain

Appendix C

Appendix C1. Accessions from the He et al., (2022) dataset that do and do not contain the chr1D introgression

Lines with chr1D introgression	Lines without chr1D introgression
GF4	GF76
GF5	GF77
GF25	GF78
GF42	GF75
GF46	GF71
GF50	GF70
GF51	GF7
GF57	GF69
GF72	GF68
GF73	GF67
GF74	GF64
GF82	GF66
GF83	GF63
GF84	GF62
GF85	GF125
GF89	GF61
GF90	GF6

Appendix C1.

GF92	GF60
GF101	GF231
GF106	GF124
GF119	GF123
GF120	GF387
GF121	GF39
GF127	GF383
GF152	GF381
GF168	GF380
GF184	GF38
GF189	GF374
GF196	GF370
GF197	GF37
GF230	GF364
GF256	GF363
GF262	GF361
GF263	GF362
GF279	GF36
GF285	GF359
GF289	GF358
GF294	GF348

Appendix C1.

GF297	GF347
GF300	GF345
GF301	GF344
GF304	GF336
GF307	GF12
GF309	GF335
GF311	GF332
GF313	GF331
GF330	GF327
GF333	GF325
GF342	GF323
GF360	GF320
GF366	GF32
GF371	GF317
GF391	GF319
	GF315
	GF312
	GF310
	GF31
	GF117
	GF100

Appendix C1.

	GF8
	GF305
	GF30
	GF299
	GF298
	GF115
	GF295
	GF293
	GF29
	GF287
	GF283
	GF110
	GF282
	GF281
	GF278
	GF277
	GF276
	GF27
	GF270
	GF11
	GF267

Appendix C1.

	GF265
	GF264
	GF249
	GF248
	GF247
	GF24
	GF238
	GF107
	GF235
	GF234
	GF228
	GF23
	GF226
	GF225
	GF224
	GF221
	GF213
	GF215
	GF208
	GF21
	GF207

Appendix C1.

	GF204
	GF205
	GF104
	GF201
	GF202
	GF190
	GF19
	GF188
	GF186
	GF180
	GF175
	GF103
	GF170
	GF154
	GF157
	GF151
	GF15
	GF138
	GF14
	GF97
	GF93

Appendix C1.

	GF13
	GF102
	GF91
	GF9
	GF88
	GF87
	GF86
	GF128
	GF81
	GF308
	GF306

Appendix C2. Expression values for genes within the chr1D introgression for accessions from the He et al. (2022) dataset, using either the Chinese Spring or the pantranscriptome reference. Accessions are split based on whether or not they contain the chr1D introgression.

Gene	Chinese Spring reference		Pantranscriptome reference	
	No introgression	Introgression	No introgression	Introgression
TraesCS1D02G436900	1.30	0.298	1.29	2.08
TraesCS1D02G437000	1.96	0.360	2.09	2.28
TraesCS1D02G437100	0.135	0.0129	0.150	0.323
TraesCS1D02G437200	3.21	0.691	3.48	3.66
TraesCS1D02G437300	0.291	0.145	0.282	0.436
TraesCS1D02G437400	0.0867	0.127	0.0765	1.14
TraesCS1D02G437600	0.889	0.422	0.569	0.443
TraesCS1D02G437700	5.15	1.2	5.04	3.99
TraesCS1D02G437800	0.668	0.255	0.587	0.580
TraesCS1D02G437900	1.47	0.572	1.42	1.24
TraesCS1D02G438000	9.52	5.78	8.88	11.2
TraesCS1D02G438300	0.00	0.0221	0.00150	0.0115
TraesCS1D02G438500	0.000400	0.00	0.000000200	0.00
TraesCS1D02G438800	0.000500	0.00260	0.000400	0.00260
TraesCS1D02G438900	0.00260	0.00	0.00150	0.0237
TraesCS1D02G439000	22.1	7.32	18.9	15.1
TraesCS1D02G439600	3.41	0.719	3.24	2.28
TraesCS1D02G440200	3.56	0.618	3.52	2.45
TraesCS1D02G440400	0.0562	0.558	0.00860	0.866
TraesCS1D02G440500	10.3	1.66	10.4	26.0
TraesCS1D02G440600	1.46	0.439	1.28	1.08

Appendix C2.

TraesCS1D02G441000	0.00190	0.00	0.00260	0.0110
TraesCS1D02G441300	0.0252	0.0244	0.00710	0.0128
TraesCS1D02G442400	7.58	1.87	7.26	8.58
TraesCS1D02G442500	0.120	0.0247	0.123	0.110
TraesCS1D02G442900	24.8	2.02	26.0	14.5
TraesCS1D02G443100	4.32	1.11	3.96	7.45
TraesCS1D02G444100	7.95	1.01	7.91	10.1
TraesCS1D02G444300	0.000600	0.00470	0.000700	0.00440
TraesCS1D02G444400	0.65	0.0513	0.651	0.598
TraesCS1D02G444500	0.921	0.779	2.47	2.08
TraesCS1D02G445200	4.02	0.983	3.93	5.66
TraesCS1D02G445300	28.2	3.74	27.3	19.0
TraesCS1D02G445500	2.67	0.137	2.62	0.139
TraesCS1D02G445600	0.0198	0.00120	0.0201	0.0233
TraesCS1D02G445700	4.89	1.43	4.93	6.77
TraesCS1D02G448700	2.91	0.994	2.91	2.60
TraesCS1D02G449200	26.1	2.93	26.0	34.1
TraesCS1D02G449600	3.92	0.944	3.88	3.61
TraesCS1D02G449700	0.612	0.155	0.645	0.531
TraesCS1D02G449800	0.0246	0.0330	0.0252	0.0326
TraesCS1D02G450100	0.524	0.0309	0.554	0.181
TraesCS1D02G450200	4.15	1.33	4.08	4.08
TraesCS1D02G450300	12.4	1.49	12.4	10.2
TraesCS1D02G450400	16.7	2.74	16.7	17.0
TraesCS1D02G450800	0.00980	0.0148	0.000100	0.00520

Appendix C2.

TraesCS1D02G451000	0.567	0.0642	0.560	0.577
TraesCS1D02G451100	2.92	0.644	3.06	2.00
TraesCS1D02G451200	5.45	1.23	5.04	5.36
TraesCS1D02G451300	19.2	5.37	21.4	19.2
TraesCS1D02G451600	0.00220	0.00	0.00230	0.00630
TraesCS1D02G451700	3.99	0.939	3.96	3.13
TraesCS1D02G451800	20.3	3.52	18.4	12.0
TraesCS1D02G451900	110	39.7	101	98.6
TraesCS1D02G452000	0.000300	0.00140	0.00000500	0.000500
TraesCS1D02G452100	0.00	0.00	0.00	0.00280
TraesCS1D02G452400	3.52	0.356	3.61	4.95
TraesCS1D02G452700	0.00110	0.000500	0.00110	0.00310
TraesCS1D02G452800	0.00	0.000900	0.000300	0.00180
TraesCS1D02G453100	3.09	0.881	3.05	4.41
TraesCS1D02G453600	3.80	0.375	3.71	2.40
TraesCS1D02G454200	0.00320	0.00570	0.00210	0.00740
TraesCS1D02G454900	0.000500	0.000200	0.000300	0.000200

Appendix D: Publications

Appendix D1. Coombes, B. *et al.* (2022). 'Whole genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/ *Ambylopyrum muticum* introgression lines', *Plant Biotechnology Journal*, 21(3), pp. 482-496. doi: 10.1111/pbi.13859.

Appendix D2. Molero, G.*, Coombes, B.* *et al.* (2023). 'Exotic alleles contribute to heat tolerance in wheat under field conditions', *Communications Biology*, 6(21). doi: 10.1038/s42003-022-04325-5.

Appendix D3. Coombes, B. *et al.* (2024). 'Introgressions lead to reference bias in wheat RNA-Seq analysis', *BMC Biology*, 22(56). doi: 10.1186/s12915-024-01853-w.

* Indicates equal contribution

Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2023) **21**, pp. 482–496 doi: 10.1111/pbi.13859

Whole-genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/ *Ambylopyrum muticum* introgression lines

Benedict Coombes[1], John P. Fellers[2], Surbhi Grewal[3], Rachel Rusholme-Pilcher[1], Stella Hubbart-Edwards[3], Cai-yun Yang[3], Ryan Joynson[1], Ian P. King[3], Julie King[3] and Anthony Hall[1,*]

[1]*Earlham Institute, Norwich, Norfolk, NR4 7UZ, UK*
[2]*USDA–ARS Hard Winter Wheat Genetics Research Unit, Manhattan, Kansas, 66506, USA*
[3]*School of Biosciences, The University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire, LE12 5RD, UK*

Received 22 November 2021;
revised 28 April 2022;
accepted 15 May 2022.
**Correspondence* (Tel +44(0)1603450001;
email anthony.hall@earlham.ac.uk)

Keywords: Wheat, introgression, wild relative, resistance, breeding, genomics.

Summary

Wheat is a globally vital crop, but its limited genetic variation creates a challenge for breeders aiming to maintain or accelerate agricultural improvements over time. Introducing novel genes and alleles from wheat's wild relatives into the wheat breeding pool via introgression lines is an important component of overcoming this low variation but is constrained by poor genomic resolution and limited understanding of the genomic impact of introgression breeding programmes. By sequencing 17 hexaploid wheat/*Ambylopyrum muticum* introgression lines and the parent lines, we have precisely pinpointed the borders of introgressed segments, most of which occur within genes. We report a genome assembly and annotation of *Am. muticum* that has facilitated the identification of *Am. muticum* resistance genes commonly introgressed in lines resistant to stripe rust. Our analysis has identified an abundance of structural disruption and homoeologous pairing across the introgression lines, likely caused by the suppressed *Ph1* locus. mRNAseq analysis of six of these introgression lines revealed that novel introgressed genes are rarely expressed and those that directly replace a wheat orthologue have a tendency towards downregulation, with no discernible compensation in the expression of homoeologous copies. This study explores the genomic impact of introgression breeding and provides a schematic that can be followed to characterize introgression lines and identify segments and candidate genes underlying the phenotype. This will facilitate more effective utilization of introgression pre-breeding material in wheat breeding programmes.

Introduction

Triticum aestivum L. (bread wheat) is a vital crop, providing around 20% of calories and 25% of protein consumed globally (Reynolds *et al.,* 2012). Improvements to wheat since the late 19th century have largely come from conventional breeding strategies, but these improvements rely on ample genetic variation in the primary gene pool (Hao *et al.,* 2020). The hexaploid bread wheat grown today derives from just one or two polyploidization events ~10 000 years ago between the tetraploid *Triticum turgidum* and the diploid *Aegilops tauschii* (Charmet, 2011). The limited diversity stemming from this genetic bottleneck has been compounded over time by intensive breeding. Pressure on breeders to prioritize advanced breeding material (J. Valkoun, 2001) for more rapid development of uniform, high-quality varieties have limited the introduction of genetic variation from external sources. The genetic variation that does exist in modern wheat material is rapidly being exhausted, evident in plateauing yield improvements that left unchecked, will be insufficient to meet global demands (Ray *et al.,* 2013). Wild relative introgression breeding will be a major component of overcoming this genetic constraint in the years to come, enabling breeders to access the secondary and tertiary gene pools of wheat (Hao *et al.,* 2020; J. Valkoun, 2001) and incorporate novel alleles or genes into modern breeding material.

There are many examples of the successful transfer of wild relative genes into wheat since first pioneered by E.R. Sears (Doussinault *et al.,* 1983; Fatih, 1983; Friebe *et al.,* 1996; Klindworth *et al.,* 2012; Sears, 1956). However, challenges associated with the high-throughput production and verification of introgression lines, in addition to the linkage drag of introgressed segments, have limited the widespread adoption of introgression breeding. Utilizing recombination mutants and high-throughput marker methods, introgressing entire wild relative genomes into wheat as stably inherited, homozygous segments is now possible (King *et al.,* 2017, 2019). These sets of lines provide the raw material required for the incorporation of alien variation into breeding programmes. Segments in these lines that confer phenotypes of interest can be identified. Lines with overlapping segments can then be crossed to break down large segments (Khazan *et al.,* 2020), resulting in genes of interest captured in short introgressed segments with reduced linkage drag, ready to be deployed in breeding programmes.

Identifying the introgressed content of each introgression line is important for the effective utilization of these lines. Insufficient marker density for genotyping approaches such as Kompetitive allele-specific PCR (KASP) and low resolution of genomic *in situ*

hybridization (GISH) limits the resolution at which segments can be identified. Determining the precise size and positions of segments and refining positions of overlap between introgression lines is important when relating to phenotypic data to narrow down regions containing genes of interest. Identifying introgression boundaries at a higher resolution will allow lines with overlapping segments to be identified; these can be crossed to break down segments and capture genes of interest in smaller segments with reduced linkage drag.

Wild relative genes have undergone selection in a different environment to the agricultural setting in which elite wheat lines are selected and thus may be deleterious, or be, at the very least imperfect replacements of their wheat orthologue when deployed in field conditions. Therefore, many genes introgressed along with a gene of interest will contribute to reduced agronomic performance of a line. This reduced performance will be driven by differences both in the encoded protein and in the pattern of expression of the introgressed gene compared to the wheat orthologue it replaced. In addition to these direct changes to gene expression caused by introgression, disruptions to established regulatory networks and the resulting indirect effects on the expression of wheat genes in the genomic background will likely contribute to altered performance.

Ordinarily, hexaploid wheat behaves as a diploid during meiosis. The *Pairing Homoeologous 1* (*Ph1*) locus is largely responsible for this behaviour, restricting synapsis and crossovers to homologous chromosomes (Rey *et al.,* 2017). A suppressed or deleted *Ph1* locus enables recombination between wheat chromosomes and non-homologous wild relative chromosomes and is a major tool used to transfer wild relative genes into wheat (Martín *et al.,* 2017). However, this also enables homoeologous chromosomes to pair and recombine leading to transmission of chromatin between the subgenomes of wheat (Koo *et al.,* 2020) and deletions/duplications where synteny between homoeologous chromosomes breaks.

Here, we have conducted a high-resolution genomic analysis on 17 hexaploid wheat/*Am. muticum* introgression lines (King *et al.,* 2017, 2019), utilizing whole-genome sequencing (WGS) data from the introgression lines and the parent lines and a draft genome assembly of *Amblyopyrum muticum* [(Boiss.) Eig.; *Aegilops mutica* Boiss; 2n = 2X = 14; genome TT], a wild relative of wheat belonging to its tertiary gene pool. Phenotypic screening of *Am. muticum* introgression lines (Fellers *et al.,* 2020) has revealed resistances to leaf, stem and stripe rust not observed in the parental wheat lines and thus likely conferred by introgressed genes. KASP genotyping to identify segments has been conducted on many of these lines (Grewal *et al.,* 2021). Through this analysis, we have pinpointed introgression segment junctions to a higher resolution than previously possible, in many cases within a single pair of reads, demonstrating segments of variable size that overlap between introgression lines, which explains some differences in resistance phenotype seen between lines. These overlaps will enable these segments to be further broken down by crossing introgression lines together. Using *in silico* karyotyping, we have shown that large-scale structural disruption is ubiquitous across the lines, including deletions and duplications up to whole-chromosome size and homoeologous recombination likely facilitated by *Ph1* suppression. A genome assembly and gene annotation of *Am. muticum* has enabled us to identify introgressed resistance genes in stripe, stem and leaf rust-resistant lines that may represent novel resistance conferred by *Am. muticum* genes. Analysis of gene expression of six introgression lines compared with the wheat parent lines has revealed that novel introgressed genes are less likely to be expressed than introgressed genes replacing an orthologue. Introgressed genes directly replacing a wheat orthologue show a tendency to be downregulated, with no significant balancing of the homoeologous copies in the remaining subgenomes.

Results

Whole-genome sequencing facilitates high-resolution introgression detection

To reveal *Am. muticum* segments within introgression lines using WGS data, we developed a workflow that utilizes mapping coverage and single nucleotide polymorphism (SNP) information from the introgression line and the wheat parents. If a wheat segment is replaced by an *Am. muticum* segment the mapping coverage will drop in that region due to structural variation and breaks in synteny between wheat and *Am. muticum*. Due to the homozygous nature of the lines, homozygous muticum-specific SNPs are indicative of the site of introgression. Reads derived from an introgressed segment that aberrantly map to a non-introgressed region will map at the same position as the wheat reads coming from that region and result in heterozygous SNP calls with muticum-specific and wheat-specific alleles found at the same position. Therefore, to locate introgressions, we searched for genomic blocks with reduced mapping coverage, homozygous *Am. muticum*-specific SNPs and few heterozygous *Am. muticum*-specific SNPs. We identified introgressions using 1Mbp genomic windows and then defined the borders to a higher resolution using 100Kbp genomic windows. This was performed on 17 double haploids (DH) or backcrossed (BC) *Am. muticum* introgression lines from which Illumina paired-end short reads were produced to an average depth of around 5x. Figure 1a shows an example of this macro-level visualization of introgression line DH65, which has a 51.29Mbp segment on the telomere of the short arm of chr4D, and a 139.6Mbp monosomic deletion on the short arm of chr5B. Macro-level genome plots for all lines can be seen in Figure S1.

Using this approach, we confirm the existence of 100% of segments previously identified with KASP genotyping (Grewal *et al.,* 2021). However, we were able to resolve the locations of segment junctions to a much higher resolution than previous methods, due to the limited marker density available for KASP genotyping and the inability of GISH to resolve segments below ~20Mbp. In addition, we were able to uncover two previously unreported segments that have been subsequently validated by KASP genotyping (Grewal *et al.,* 2021); a 17.39Mbp on the telomere of chr7D of DH195 and a 22.68Mbp segment on the telomere of chr5D in DH121. We also identified a new 3.99Mbp segment on chr6D of DH15 that we validated using 2 KASP markers, WRC1873 and WRC1890 (Table S3). All precise segment positions are listed in Table S2.

To explore junction regions of segments in fine detail, we used the Integrative Genome Viewer (IGV) (Robinson *et al.,* 2011), an interactive browser that allows sequencing reads mapped to a genome within a specified interval to be manually interrogated. Using IGV to explore the junction regions, we were able to precisely identify 33/42 segment ends (78.6%). As some segment ends are telomere substitutions as opposed to crossovers and some segments are derived from the same initial cross, we just looked at uniquely-derived crossover junctions and found that we could identify the precise crossover point between wheat and

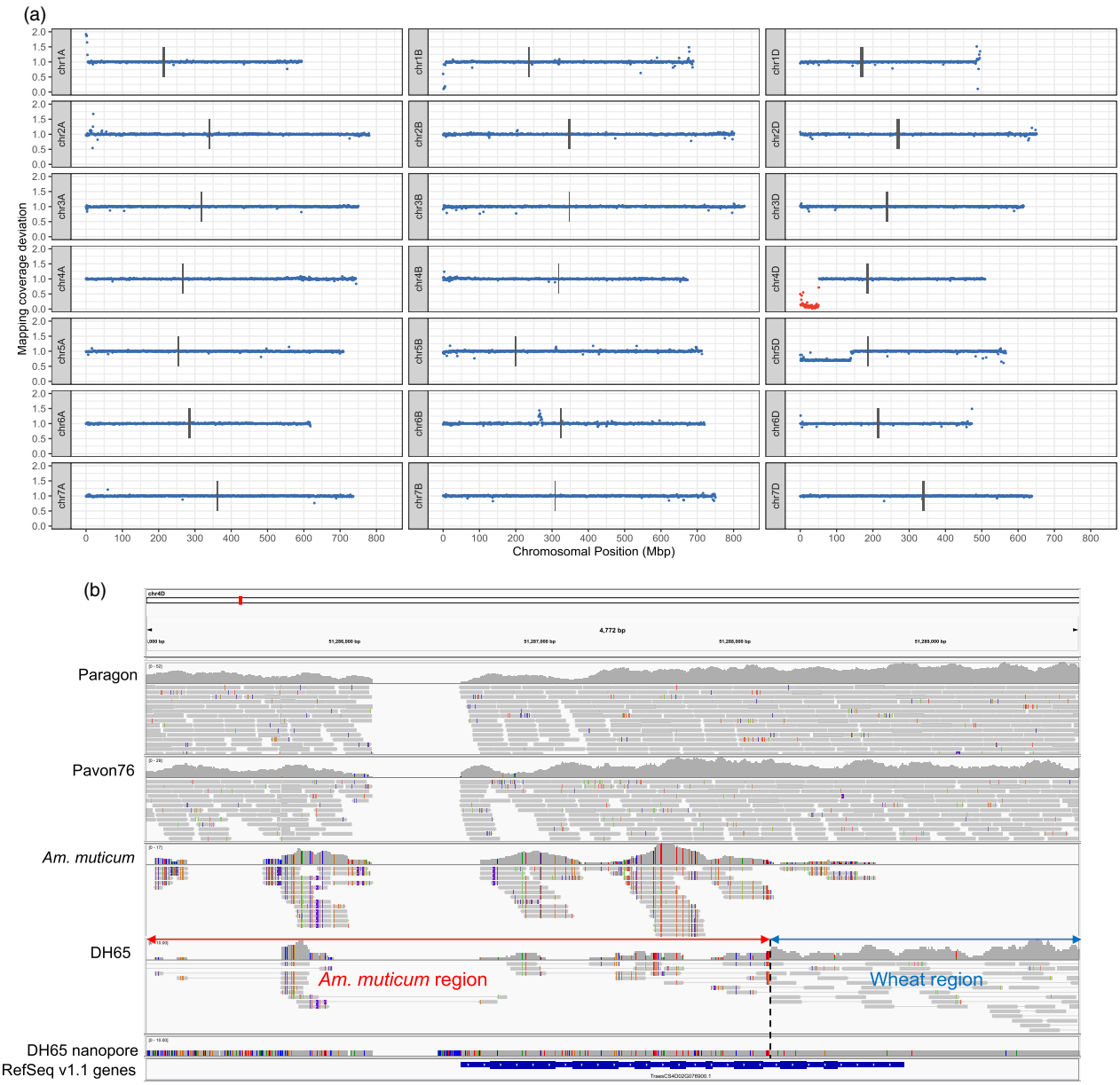


Figure 1 Identifying introgressed *Am. muticum* segments using whole-genome sequencing data. (a) Introgression line DH65, which has a 51.29Mbp introgressed segment on chr4D and a 139.6Mbp monosomic deletion on chr5D. Each point represents the deviation in mapping coverage with the wheat parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red. (b) IGV image showing junction at the right-hand side of chr4D segment in the introgression line DH65 (Figure 1a), spanned by both Illumina paired-end reads and Oxford Nanopore reads from DH65. The first four tracks show mapped illumina WGS data, the fifth track shows assembled contig from aligned Oxford Nanopore reads for DH65, and the bottom track shows high confidence genes from the RefSeq v1.1 annotation.

Am. muticum in 12/17 (70.6%) cases. Of the remaining junctions, two were narrowed down to within 100kbp and three had complex structures with duplication events that prohibited precise localization. Out of the 12 high-resolution junctions, 11 (91.7%) were within 670 bp upstream or downstream of a wheat gene, with 8 falling within the gene body itself, suggesting that crossovers may be localized to genes. The remaining junction was 6.75Kbp downstream from the nearest gene.

For line DH65, the pinpointed junction was validated with Oxford Nanopore long reads mapped to RefSeq v1.0 along with the Illumina paired-end short reads (Figure 1b). Oxford Nanopore reads spanned the breakpoint between *Am. muticum* and wheat at the right-hand side of the 51.6Mbp chr4D segment, adding confidence to the identification from Illumina reads alone. We assembled these mapped Oxford Nanopore reads using wtdbg2 (Ruan and Li, 2020) with relaxed parameters to include reads that were clipped due to high divergence between wheat and *Am. muticum*, producing a contig that spans the junction. This contig spans the entire junction, including regions to which neither the Illumina reads from *Am. muticum* nor DH65 map. These regions appear to have elevated SNP density, explaining the gaps in mapping.

For introgression lines with KASP genotyping verification, WGS data may offer an affordable tool to aid breeders to identify precise location and size of these segments. To assess the sequencing depth requirements to locate the position and size of introgressed segments using coverage deviation alone, we downsampled the Illumina paired-end short reads from 2 lines for which we have identified very precise positions of the junction borders; DH65 and DH92, to 1x, 0.1x, 0.01x and 0.001x to choose the lowest depth at which we could still resolve segment position. 0.01x was the lowest depth that still provided comparable resolution (Figure S3).

Introgression breeding process induces homoeologous pairing and large chromosomal aberrations

In addition to introgression sites, we have identified large deletions and duplications, many of which were whole chromosome arm or whole chromosome in scale, based on the deviations in mapping coverage not attributable to introgressions. Within the 17 lines examined, 12 lines (70.6%) have one or more very large chromosomal aberrations exceeding 140Mbp. These include duplication of most of chr1A with a deletion of the homoeologous region of chr1B in DH pair DH124 + DH355 (Figure 2a); deletion of the short arm of chr4A in DH86 and deletion of the long arm of chr4B in its DH pair, DH92 (Figure 2b); monosomic deletion of most of the short arm of chr5D in DH121 and DH65 (Figure 2c), which are not a DH pair, indicating that this event has occurred multiple times at the same position; and a monosomic deletion of chr1A in DH195 (Figure 2e).

Homoeologous translocations resulting in the non-reciprocal transfer of genetic material can be detected through mapping coverage deviation, indicated by a duplication and deletion in corresponding homoeologous regions. We can also use differences within a double haploid (DH) pair (Table S1) to infer what genetic events must have taken place to give rise to the segregation patterns we see from DH lines derived from the same BC3 line. We see evidence of homoeologous pairing both from duplicated/deleted pairs of chromosomes, such as in DH355 and DH124 (Figure 2a) and from corresponding deletions/duplications at homoeologous positions (Figure 2d, f). In BC2F420 (Figure 2f), recombination has taken place between chr5A and chr5D and chr5B has been deleted.

Genome assembly and annotation of *Am. muticum*:

To facilitate the identification of introgressed genes both for differential expression analysis and to find candidate introgressed resistance genes, we produced a draft genome assembly for *Am. muticum* 2130012 comprising most of the gene space. After polishing with long and short reads and resolving haplotigs, the assembly comprised 96 256 contigs and was 2.53Gbp in length, with an N50 of 75.5Kbp (Table S5). We estimated the size of the *Am. muticum* genome through two independent methods: mapping the Oxford Nanopore reads back to the assembly and computing coverage across single-copy genes; and based on k-mer counts within the Illumina paired-end reads (Figure S4). These resulted in estimates of 4.90Gbp and 4.57Gbp, respectively, compared with flow cytometry estimate of 6.174Gbp (Pellicer and Leitch, 2020). Although the genome spans just 53.4% of the estimated genome size (mean of our two estimates), BUSCO analysis (Waterhouse *et al.*, 2018) revealed that 94.2% of the expected gene space was assembled unfragmented (Figure S5). Gene annotation using evidence from root and shoot transcriptomic data, proteomic data, and ab initio predictions resulted in 86 841 gene models, 32 385 of which were designated as high confidence (HC) (Table 1). 28 995 (89.8%) of the HC genes were assigned functional annotation.

To identify *Am. muticum* genes not present in wheat and gene families that have undergone expansion in *Am. muticum*, both of which could be contributing novel variation in introgression lines, we used OrthoFinder (Emms and Kelly, 2019) to construct 31 616 orthogroups from the proteins encoded by the HC genes from *Am. muticum*, *Triticum aestivum*, *Triticum urartu*, *Aegilops tauschii*, *Oryza sativa* and *Brachypodium distachyon* (Figure S6). 93.8% of *Am. muticum* genes were placed in an orthogroup. 3873 *Am. muticum* genes are not present in wheat and 108 orthogroups, comprising 867 *Am. muticum* genes, have undergone expansion in *Am. muticum* compared to wheat. Enrichment analysis of GO Slim terms (Figure S7) revealed that the novel *Am. muticum* genes were enriched most significantly for terms associated with metabolic processes.

Expression of introgressed genes and impact on the background wheat transcriptome

To explore how introgressed genes are expressed and to understand the impact of the introgression breeding programme on the wheat transcriptome, we produced mRNAseq data for six of the introgression lines and the wheat parent lines. *Am. muticum* genes introgressed into each line were identified using orthologue assignments and DNA read mapping evidence. RNA reads were mapped to a pseudo genome (ABDT) constructed by concatenating the wheat reference genome, RefSeq v1.0, with the draft *Am. muticum* genome assembly; this allows us to distinguish between RNA deriving from wheat genes and from *Am. muticum* genes in the same way that we can distinguish between wheat homoeologues.

Across all six lines, 1750/4989 (35.1%) introgressed genes were expressed. Splitting the introgressed genes into those with an orthologue in wheat and those that are novel revealed that while 1627/3691 (44.1%) introgressed genes with a wheat orthologue were expressed, only 123/1298 (9.48%) novel introgressed genes were expressed (Figure 3a). For introgressed genes that do have a wheat orthologue, those that are more diverged from the orthologue are less likely to be expressed (Figure 3b), ranging from 21.5% of genes with no wheat orthologue >90% protein identity being expressed to 64.8% of genes with an orthologue in wheat with ≥99% protein identity being expressed.

To test whether *Am. muticum* genes that have directly replaced a wheat orthologue are expressed differently to that orthologue, we called differential expression between each introgression line and the wheat parent lines using DESeq2 (Love *et al.*, 2014) after summing the expression count of each replaced wheat gene with that of its introgressed *Am. muticum* orthologue. Between 13.3% and 23.1% (mean of 19.3% across all lines) of introgressed genes were called as differentially expressed ($abs(\log_2 FC) \geq 1$ and adj. *P*-value ≤ 0.05 in both parental comparisons) when compared to the expression in the parent lines of the wheat orthologue they replaced (Figure 3c). Between 54.5% and 87.8% (mean of 69.8% across the lines) of these differentially expressed introgressed genes were downregulated in the introgression line.

We hypothesized that the suppression of wheat genes in an introgressed or deleted region would lead to a change in the expression of homoeologous copies of that gene in the other subgenomes to compensate for the loss of expression. The results

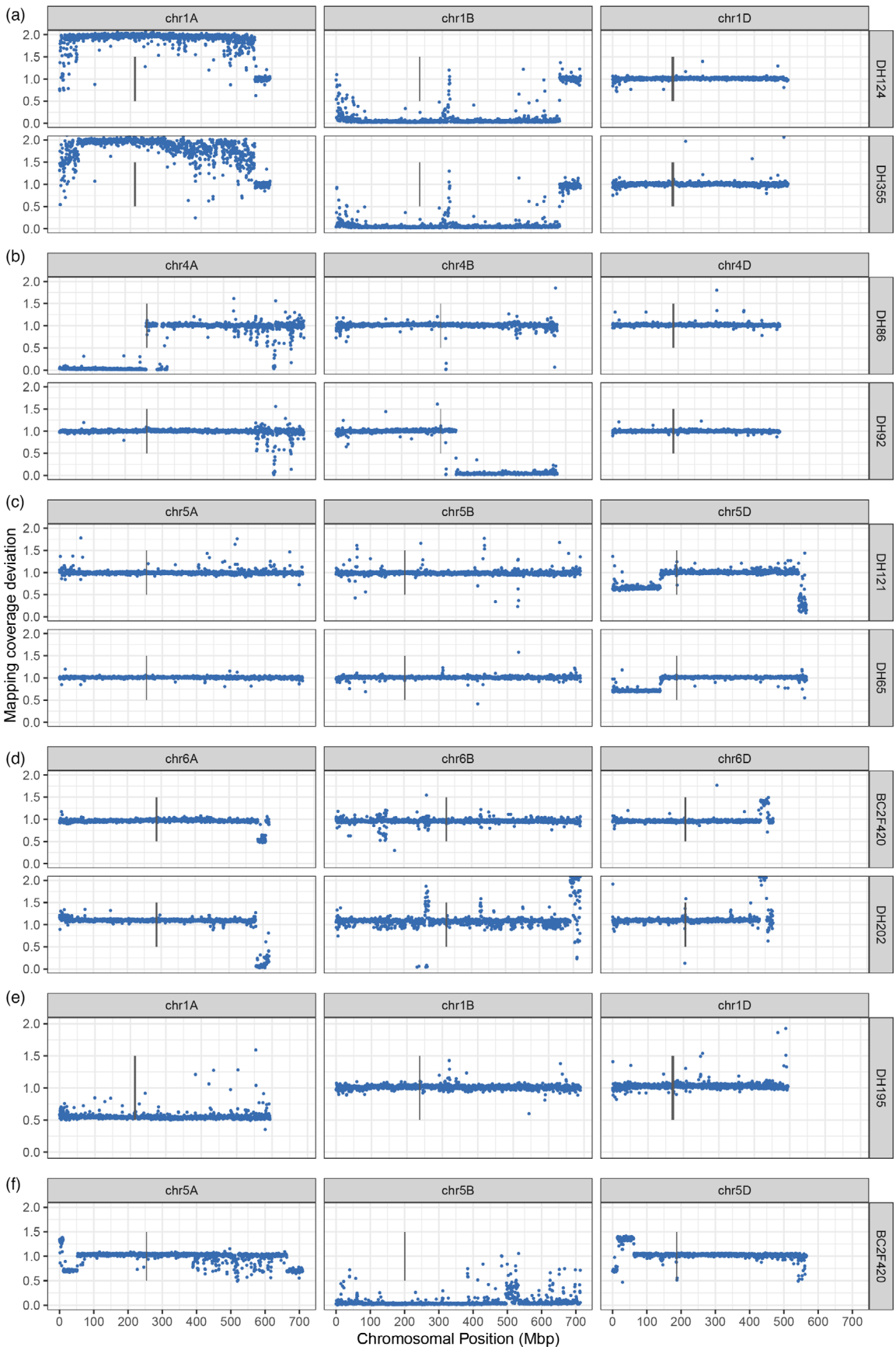


Figure 2 Large chromosomal aberrations in *Am. muticum* introgression lines. Each point shows mapping coverage deviation compared with the wheat parents in 500Kbp windows across the genome. (a) Corresponding duplication and deletion seen in both lines of the DH pair, caused by pairing of a duplicated chr1A and chr1B. Mapping coverage deviation of 1 at the end of chr1A and chr1B indicates a large translocation between chr1A and chr1B has taken place in duplicated chr1A + chr1B pair and discontiguous mapping coverage deviation change towards beginning of chr1A and chr1B suggests lots of smaller translocation events. (b) Chromosome arm deletions on homoeologous chromosomes of DH pair. (c) Monosomic deletions at the same position in two independently derived lines. (d) Homoeologous exchange within homoeologous group 6, at similar positions in two independently derived lines. (e) Monosomic deletion of chr1A in DH195. (f) Homoeologous recombination event between chr5A and chr5D and a deleted chr5B.

of multiple approaches support a lack of overall rebalancing of triad expression following suppression of one of the copies, both in the 400Mbp introgressed region on chr5D of BC3F45 and in the deletion of chr7D in DH161 (Figure 4). In triads where the D homoeologue has been replaced by a *Am*. muticum gene or been deleted, there is an overall reduction in expression on the D homoeologue (Figure 4a iii, b iii), though to a much lesser degree in the introgression where introgressed *Am. muticum* orthologues are being expressed. In the introgression, there were 74 triads with the D homoeologue introgressed and called as downregulated; none of these triads had any homoeologues called as upregulated. This was compared with 10 953 control triads, where no homoeologues are introgressed or deleted and the D homoeologue is not differentially expressed, of which 17 (0.155%) triads had the A or B homoeologue upregulated. For the deletion, out of 1294 triads with the D homoeologue deleted and therefore not expressed, just 6 triads had one or more homoeologues upregulated (0.464%); this compares to 37 (0.369%) out of 10 031 control triads having the A or B homoeologue upregulated. These differences are not significant (Fisher's exact test two-tailed *P*-values of 1.00 and 0.628, respectively).

To complement the above approach and consider homoeologues whose expression may have changed but not sufficiently to be called as significant by DESeq2, we looked at the $\log_2$ fold change ($\log_2$FC) in DESeq2 normalized expression counts. Plotting the $\log_2$FC of DESeq2 normalized expression counts in 10Mbp windows (Figure 4a i, b i) across the chromosomes illustrates the overall stability of expression in homoeologous regions of introgressions and deletions. For the introgression and the deletion, we compared the $\log_2$FC of the A and B homoeologues of triads where the D homoeologue had been introgressed or deleted with the $\log_2$FC of the A and B homoeologues of a control set of triads defined as above (Figure S9). We found no statistically significant difference between the test and control sets (two-tailed *t* test *P*-values: deletion = 0.209; introgression = 0.252). This indicates that, like the proportion of DEGs, the change in expression counts of homoeologues in which the D homoeologue has been downregulated/silenced does not change beyond that expected by chance.

We also looked at genes in genomic windows not deviating in coverage compared with the wheat parent lines (Figure 3) to explore whether the introgressions and structural changes induced by the introgression breeding programme had indirectly affected the expression of remaining wheat genes. Between 0.181% and 2.40% (106–1261 genes; mean of 0.860% across lines) of these wheat genes were differentially expressed compared with the wheat parents. To assess whether any specific gene functions were enriched in the differentially expressed genes we looked for enriched GO Terms (Figure S8). We found some terms to be enriched, suggesting a non-stochastic impact on background transcription; however, differences between lines suggest that the nature of the impact on background transcription depends on the genes introgressed/disrupted elsewhere in the genome. Some terms are enriched in more than 3 lines, suggesting these are commonly affected. These are oxidoreductase activity, oxidation–reduction process, tetrapyrrole binding, catalytic activity, carbohydrate metabolic process, cofactor binding, which are enriched in downregulated genes; and ion binding, hydrolase activity and catalytic activity, which are enriched in upregulated background genes.

Table 1 Metrics for *Am. muticum* high-confidence (HC) and low-confidence (LC) gene models

	HC	LC
Total genes (no.)	32 385	54 456
Single exon (no.)	6695	27 364
Multi exon genes (no.)	25 690	27 092
Mean gene length (bp)	3355	1642
Median gene length (bp)	2178	713
Mean CDS length (bp)	1198	716
Median CDS length (bp)	1000	502
Mean exons per transcript (no.)	4.81	2.39
Median exons per transcript (no.)	3	1
Mean exon length (bp)	249	307
Median exon length (bp)	131	196

Identifying candidate introgressed genes underlying *Am. muticum* derived rust resistance

Two of the lines that we sequenced, DH92 and DH121 (Figure 5a), have complete resistance at the seedling stage to Kansas isolates of *Puccinia striiformis tritici* (stripe/yellow rust) (Fellers *et al.*, 2020). DH92 also displays chlorotic adult resistance to leaf rust and partial resistance to stem rust, that is absent in DH121. These lines have overlapping 5D segments, the positions of which were refined to 533.2–566.1Mbp (32.9Mbp) in DH92 and 544.1–566.1Mbp (22Mbp) in DH121. Therefore, the source of the stripe rust resistance is likely within the overlapping 22.68Mbp region, and the source of leaf/stem rust resistance is likely within the 10.9Mbp region unique to DH92.

Using a mapping-based approach to the pseudo-ABDT genome (Figure 5b) and combining with functional annotation, we identified 13 complete nucleotide-binding, leucine-rich repeat (NLR) immune receptors uniquely introgressed in these two lines. 12 of these have a syntenic wheat orthologue within the overlapping region of the 5D segments and 2 displayed unique

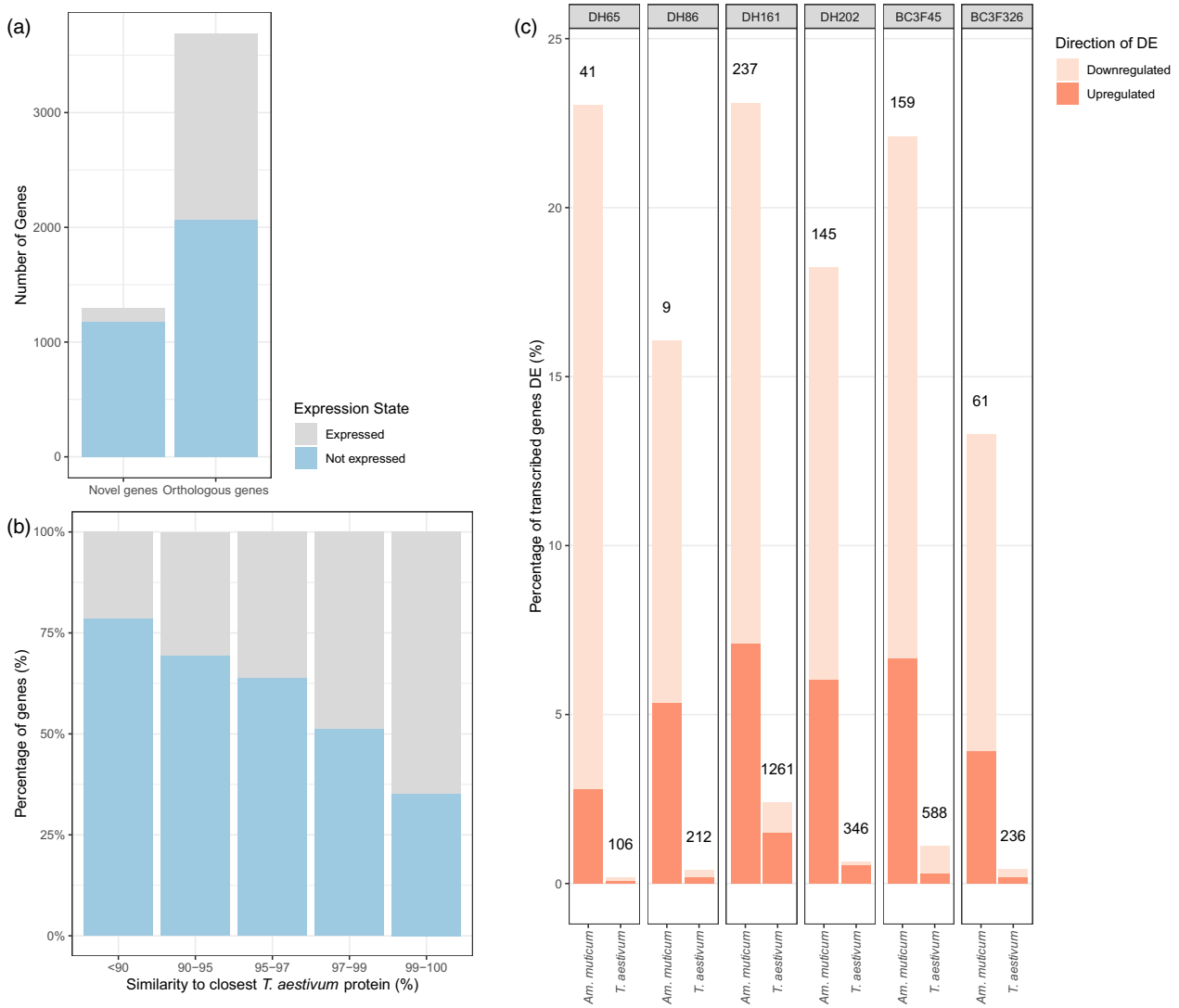


Figure 3 Expression of introgressed *Am. muticum* genes. (a) Expression state (Expressed or Not Expressed) of novel introgressed genes and introgressed genes in an orthogroup with a wheat gene (b) Expression state (Expressed or Not Expressed) of introgressed genes within an orthogroup with a wheat gene, binned by the protein identity between the *Am. muticum* protein and the most similar protein in the wheat reference genome annotation RefSeq v1.1. (c) Differential expression in 6 introgression lines, looking at introgressed genes compared to the orthologue they replaced in the parent lines, and background wheat genes compared with the expression in the wheat parent lines. The height of the bar represents the percentage of genes differentially expressed within introgressed and background regions for each line. The number above each bar is the number of genes called as differentially expressed.

NB-ARC domain signatures. 10 of the NLRs are within a 597.34Kbp cluster, including the 2 novel NLRs. We also identified 2 ABC transporters uniquely introgressed, both of which have 5D orthologues with over 97.5% protein identity, and 7 protein kinase genes uniquely introgressed, 3 of which are highly diverged at the protein level compared with the closest protein in wheat (52.2%, 74.2% and 77.0%). NLRs, ABC transporters and LRR protein kinases have all been previously implicated in resistance to stripe, leaf and stem rust (Chen *et al.*, 2020; Krattinger *et al.*, 2009; Wang *et al.*, 2020). Gene candidates are detailed in Table S6.

We identified 3 wall-associated protein kinases (WAKs), and 3 protein kinases uniquely introgressed in DH92 with orthologoues or proximal to orthologues of wheat genes in the 10.9Mbp non-overlapping region of the 5D segment. 2 of the WAKs are orthologues of TaWAK388 and TaWAK390 on 5D and 1 is orthologous to TaWAK255 on chr4A. Wall-associated kinases have previously been associated with leaf rust adult plant resistance (APR) (Dmochowska-Boguta *et al.*, 2020). Two of the protein kinases are identical at the protein level and are most similar to TaWAK387 just upstream of the TaWAK388 and TaWAK390. These may be truncated tandem duplications of this WAK. Unlike the other uniquely introgressed genes identified, the WAKs have some reads mapping to them in most of the introgression lines but only in DH92 is the coverage uniform across their lengths. This likely suggests that these are uniquely introgressed in DH92 and thus can remain as resistance candidates, but similar *Am. muticum* WAKs present in other lines are falsely mapping to these. This is supported by a lack of mapping across the rest of the contig in the other lines, unlike in DH92. Gene candidates are detailed in Table S7.

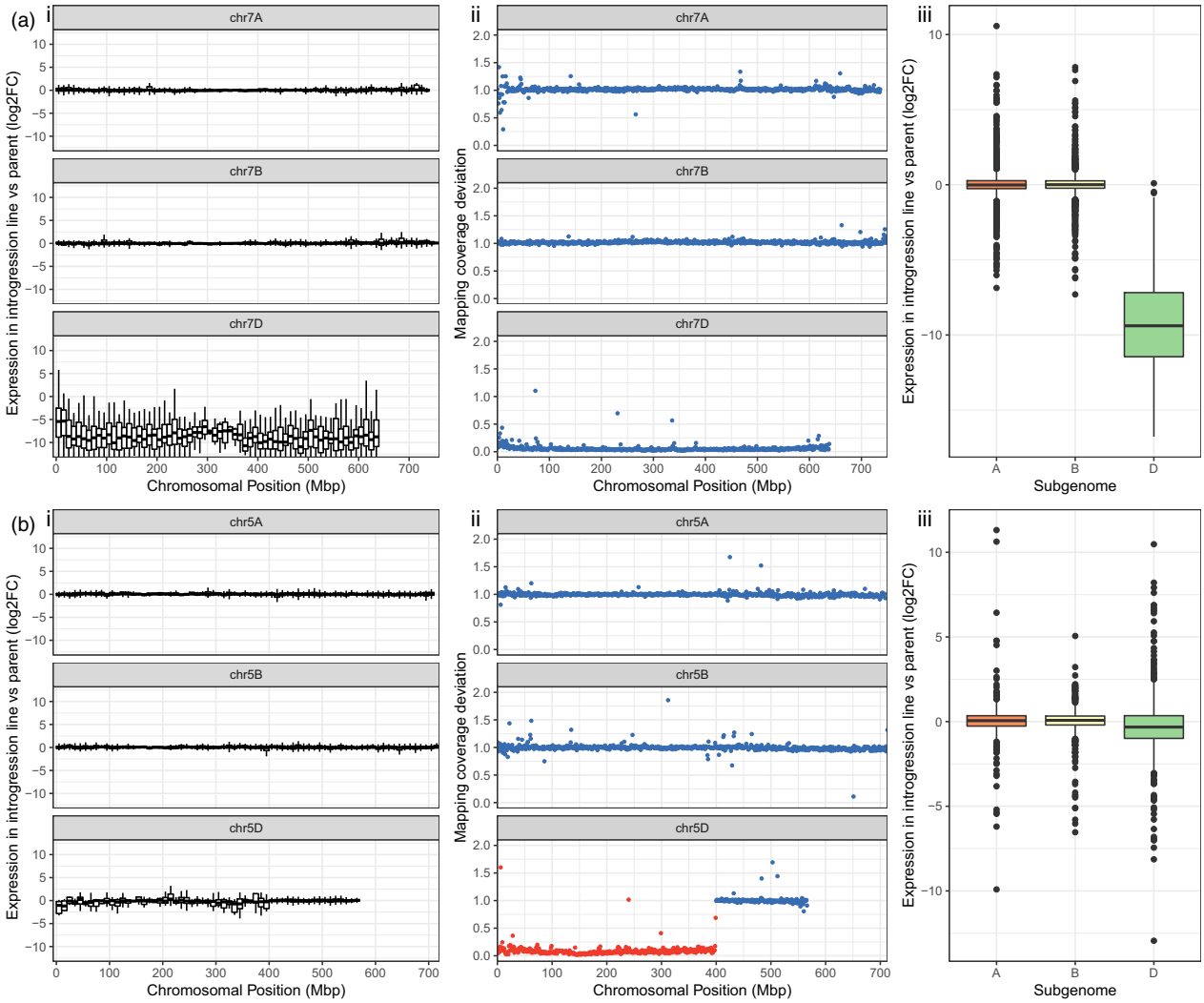


Figure 4 Expression profile across introgression and a deleted region and their homoeologous regions. (a) chr5A, chr5B and chr5D in BC3F45, with a chr5D:1-400Mbp introgression where chr5D genes have been replaced by *Am. muticum* orthologues (b) chr7A, chr7B and chr7D in DH161 where chr7D has been deleted. i DESeq2 processed log_2FC (introgression line/Paragon) of expression compared with Paragon binned into 10Mbp window ii Macro level structure in 1Mbp windows. Each point represents the deviation in mapping coverage compared to the parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red; iii. log_2FC (introgression line/Paragon) of A, B and D homoeologues belonging to triads in which the D copy has been deleted or replaced by an *Am. muticum* gene.

Discussion

Using whole-genome sequencing to pinpoint wild relative introgressions in wheat – An affordable approach to better characterize introgression lines

The current approach for studying synthetic introgression lines prior to deployment in breeding programmes relies on cytogenetic and genotyping techniques, namely GISH and KASP (Grewal *et al.*, 2021; King *et al.*, 2019). *De novo* discovery of SNPs to produce higher density KASP markers has improved the resolution but are insufficient for unpicking the precise size and location of segments and will likely miss small segments without the guidance of WGS data to identify areas in which additional markers should be deployed. We observe this with the new chr6D segment, the small chr7D segment in DH195 and chr5D segment in DH121, the latter two of which are sources of novel disease resistance.

We have demonstrated how whole-genome sequencing data can be used to define introgressions to a very high resolution as well as resolve large-scale structural changes in these lines. Downsampling has shown that if we do not require SNP information, only 0.01x sequencing coverage is required to pinpoint the junctions of known introgressed segments to a comparable resolution. Overlaying this information with KASP genotyping will undoubtedly provide an affordable method to characterize sets of synthetic introgression lines more accurately and comprehensively.

Introgressed segments nested within complex genomic structures, such as in DH202 (Figure S2), can only be inferred in conjunction with cytogenetic data and/or segregation patterns of DH pairs. Some introgression segment boundaries, such as the left-hand border of chr2A in DH15, can be identified but are difficult to pinpoint precisely due to structural complexities around the junction. Therefore, caution is advised when relying on

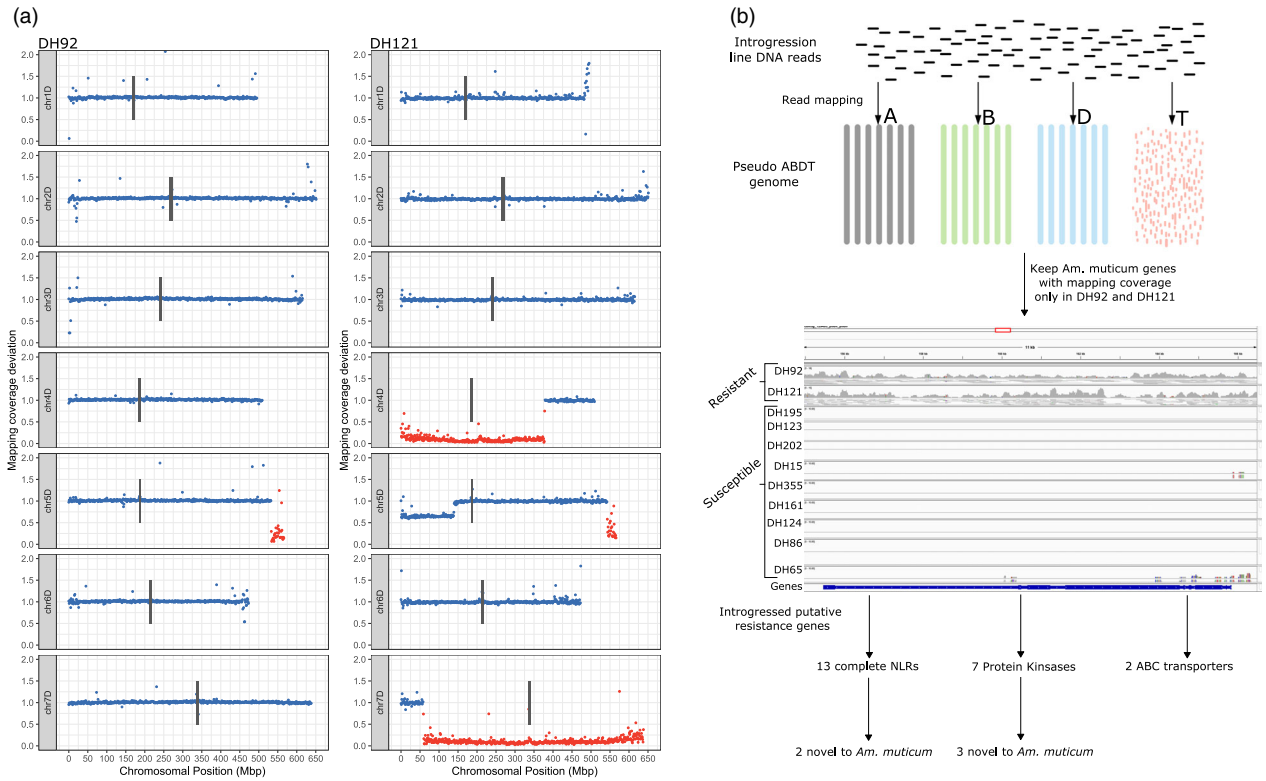


Figure 5 Identifying candidate introgressed resistance genes. Introgression lines DH92 and DH121 possess a partially overlapping introgressed segment on chr5D, a common resistance phenotype to stripe rust but a differential resistance phenotype to leaf and stripe rust. (a) Macro-level structure of the D subgenome of DH92 and DH121 (no segments on A or B subgenomes). Each point represents the deviation in mapping coverage compared with the parent lines in 1Mbp windows across Chinese Spring RefSeq v1.0. Windows within assigned *Am. muticum* introgression blocks are coloured red. (b) Identifying resistance genes uniquely introgressed in DH92 and DH121 and thus candidates for the stripe rust resistance shared between the two lines.

introgression assignments provided by WGS data alone, particularly for complex lines with several large introgressions/deletions/duplications. However, for most lines, where genomic structure is simpler, this approach is robust and is nevertheless an improvement on lower resolution methods, even if only to identify confounding structural complexities that would otherwise have been missed.

We have found that crossover points between wheat and *Am. muticum* mostly take place within or adjacent to genes. Previous work has shown crossovers between wheat and wild relatives are enriched in gene rich regions (Nyine *et al.*, 2020), which mirrors recombination rates along the genome (Gardiner *et al.*, 2019). Here we have achieved sufficient resolution to reveal these wild relative crossovers are taking place not only in regions of open chromatin and increased recombination rate, but within the genes themselves. Interestingly, this follows the same pattern previously identified for crossovers between homoeologous chromosomes (Zhang *et al.*, 2020), in contrast to homologous crossovers which, while enriched in subtelomeric regions and at recombination hotspot motifs, are not specifically enriched in and adjacent to gene bodies.

Genomic instability generated through introgression breeding programme

We have illustrated that structural disruption is common in introgression pre-breeding material, including homoeologous pairing and recombination, and duplications and deletions up to chromosome size. This is likely caused by the *Am. muticum*-induced suppression of the *Ph1* locus (Dover and Riley, 1972a, 1972b), however forced chromosome pairings in the F1 cross and the DH process may also be involved, although we see similar disruption in non-DH lines. An awareness of chromosomal aberrations is important for breeders using these lines in their breeding programmes. It will be important to identify the location of the *Ph1* suppressor in *Am. muticum* and other wild relatives that have an innate *Ph1* suppression system, such as *Ae. speltoides* (Li *et al.*, 2017) to prevent segments being carried forward into breeding programmes that contain a *Ph1* suppressor that could generate further genomic disruption. For introgression lines conferring specific phenotypes of interest, it may be important to remove the chromosomal aberrations through further backcrossing or to characterize which wheat genes have been deleted or duplicated as these may have large effects on phenotype.

Smaller scale variation in mapping coverage suggests there is structural variation taking place that we cannot accurately assess with our available data, such as transposable element mobilization. It will be important to assess the nature and extent of such variation in the future. Unfortunately, structural variation between available chromosome-level genome assemblies and Paragon/Pavon76 is too great for structural variants arising from genome shock to be distinguished from existing structural variation between the cultivars. To study this type of variation,

we will need genome assemblies of an introgression line and the wheat parents used in the cross, or a genome assembly of the wheat parent and long read data from the introgression line.

Identification of novel introgressed genes and gene expression profile of introgression lines

We identified 3873 novel *Am. muticum* genes that could underlie novel traits to introduce into wheat. The gene expression analysis revealed that in the introgression lines, these novel genes are much less likely to be expressed than introgressed genes with an orthologue in wheat. For introgressed genes that do have an orthologue in wheat, there is a further relationship between level of divergence and likelihood of being expressed. This may reflect a lack of required regulatory elements or less efficient transcription factor binding due to divergence between the *Am. muticum* and wheat genes. However, some of this relationship could be driven by the confounding effect of more conserved genes having more core functions and therefore being more constitutively expressed (Luna and Chain, 2021). It will be important to explore this further to begin to determine whether traits identified in wild relatives may present differently when introgressed into wheat.

Many of the introgressed genes are differentially expressed compared with the wheat orthologue replaced, far exceeding the proportion of wheat genes in the background that are differentially expressed. This makes sense biologically due to the different genomic background the *Am. muticum* genes have been placed in. Two previous studies have explored the expression of genes in wheat introgression lines with barley (Rey *et al.,* 2018) and *Aegilops longissima* (Dong *et al.,* 2020) introgressed. Due to the differences in methods used for different studies, it is difficult to compare the total proportion of DEGs. However, both previous studies also show that many introgressed genes are differentially expressed with most of these being downregulated or silenced. Despite the elevated levels of differential expression among introgressed orthologues, it is important to note that the majority of introgressed genes replacing a direct orthologue were not differentially expressed, suggesting a remarkable similarity in expression compared to the replaced gene in the majority of cases.

We did not see a significant change in homoeologue expression in response to introgression or deletion events. This is in line with previous results showing a lack of compensation in homoeologue expression following aneuploidy (Zhang *et al.,* 2017). This lack of response suggests that if large-scale balancing of triad expression does take place, it must require selection pressure, which these synthetic lines lack. Now that genome assemblies are available for wheat cultivars possessing many wild relatives introgressions (Walkowiak *et al.,* 2020) that have undergone extensive artificial selection in a wheat background, it will be interesting to analyse how these introgressed regions are expressed and whether balancing of triad expression arises after a period of selection.

We see some commonly enriched GO Terms in the genomic background that may be linked with cellular stress or loss of cellular homeostasis; this conclusion is supported by conclusions drawn in the wheat/barley introgression line (Rey *et al.,* 2018). The lines without these enriched GO Terms have less disruption overall, with fewer differentially expressed genes in the background and thus may either not have sufficient genomic stress to trigger these responses or lack sufficient sample size of DEGs to call significance.

A case study for uncovering candidate introgressed genes underlying phenotypes of interest

Combining high-resolution detection of introgressed segment borders with phenotypic information and a genic assembly of *Am. muticum* has enabled us to identify likely regions for novel rust resistances and produce lists of candidate genes. We identified the probable region of stripe rust resistance in DH92 and DH121 as being within the 22.68Mbp overlapping region of the chr5D segment. The small size and telomeric position of this segment makes it conducive for use in breeding. Within this region, we have identified candidate resistance genes, including 3 novel NLRs and 3 novel LRR Pkinase proteins. We did not find evidence for other classes of resistance genes that have been cloned for stripe rust resistance (Zheng *et al.,* 2020) uniquely introgressed in these lines. The DH92 resistance to leaf rust, that is not shared with DH121, is only seen in adult plants and to a composite of isolates; this race non-specific APR tends to be more durable and, in combination with the small segment size, makes this resistance another good target for further characterization. We identified 3 WAKs and 3 protein kinases uniquely introgressed in DH92. Wall-associated kinases have previously been shown to confer resistance to leaf rust that looks similar to APR (Dmochowska-Boguta *et al.,* 2020) and protein kinase proteins, such as Yr36, have been implicated in APR (Ellis *et al.,* 2014). If only interested in either the stripe rust or leaf/stem rust resistance, DH92 and DH121 could be crossed to recover the desired resistance in a smaller segment with less linkage drag.

In addition to narrowing down the source of resistance genes and identifying introgressed resistance candidates, this method acts as a case study that can be built on to aid the dissection of traits in sets of introgression lines. These lines as well as many other sets of synthetic introgression lines are being phenotyped for a variety of agronomically important traits and genome assemblies for additional wild relatives are likely to be produced in the coming years. The analysis we have described here will work better with improved assemblies in which contiguous introgressed segments can be reconstructed and introgressed content fully assessed.

Experimental procedures

Introgression line selection

Am. muticum/hexaploid wheat introgression lines were produced as in (King *et al.,* 2017, 2019) and summarized in Method S1. 13 DH lines, 3 selfed lines and 1 heterozygous BC line, along with *Am. muticum*, Paragon, Pavon76 and Chinese Spring, were selected for DNA whole-genome sequencing (Table S1). 12 of the lines belong to a pair or a trio of lines (referred to in this manuscript as DH pairs) that derive from seed from the same BC1 cross, so common segments are not independently derived. 4 DH and 2 BC lines (Table S1), along with *Am. muticum*, Paragon, Pavon76 and Chinese Spring, were selected for RNA extraction and sequencing.

Whole-genome sequencing, mapping and SNP calling

DNA from young leaf tissue was extracted and sequenced on Illumina NovaSeq 6000 S4 flowcells to produce 150 bp paired-end reads for the introgression lines and Pavon76 and 250 bp paired-end reads for *Am. muticum* (Method S2). 150 bp paired-end reads from Chinese Spring and Paragon were previously produced.

Reads were mapped to the Chinese Spring reference genome RefSeq v1.0 (International Wheat Genome Sequencing Consortium, 2018), followed by SNP calling and filtering (Method S3).

In silico karyotyping - calculating mapping coverage deviation compared to wheat parents

The number of mapped reads post-filtering and duplicate removal was counted across genomic windows (1Mbp and 100Kbp) in RefSeq v1.0 using bedtools makewindows (Quinlan and Hall, 2010) and hts-nim-tools (Pedersen and Quinlan, 2018) for the wheat parents (Chinese Spring, Paragon and Pavon76) and each introgression line. Mapped read counts were normalized by dividing by the total read number post-duplicate removal. Normalized counts of each introgression line were divided by the normalized count of each wheat parent in its crossing history (Paragon + Pavon76 or Paragon + Chinese Spring) and the number closest to 1 was kept as the coverage deviation for that window, under the assumption that the parent with mapping coverage closest to the introgression line is the parental donor in that window. The resulting number reflects the copy number of wheat DNA in that window relative to the wheat parent. A number of 1 indicates that the DNA in that window is present in the same amount as in the parent line. A number approaching 0 suggests either a deletion or an introgression has occurred at that region, and a number of 2 suggests a duplication event has taken place. Intermediate values indicate heterozygous copy number change. We defined windows with a coverage deviation between 0.8 and 1.2 as being 'normal' and not in copy number variation compared with the wheat parents.

Identifying *Am. muticum*-specific SNPs and assigning introgressed regions

A set of custom python scripts were used to analyse the coverage deviation files and vcfs and identify the introgression segments in each line. These scripts, alongside more detailed methods, are available at: https://github.com/benedictcoombes/alien_detection. First, we produced 18 496 474 SNPs between *Am. muticum* and Chinese Spring that were not shared with either Paragon or Pavon76 (Method S4). Introgression line SNPs were then assigned as *Am. muticum* if matching an *Am. muticum* specific SNP in position and allele. Sites exceeding 3x mean coverage level were removed as this signifies collapsed repeat expansion. These SNPs were then split into homozygous and heterozygous and binned into 1Mbp windows using bedtools coverage (Quinlan and Hall, 2010).

Coverage deviation blocks were defined based on contiguous blocks of 1Mbp windows with coverage deviation <0.7, with windows within 5Mbp from the previous coverage deviation block being merged. The block was discarded if <80.0% of constituent windows had a coverage deviation <0.7. Coverage deviation blocks were assigned as *Am. muticum* based on the presence of homozygous *Am. muticum*-specific SNPs and a high ratio of homozygous to heterozygous *Am. muticum*-specific SNPs, within 1Mbp windows across the block (Method S5). Coverage deviation in 100Kbp windows either side of the larger block was used to define the borders of the segment. To locate the precise position of this junction, the BAM alignment files for *Am. muticum*, Paragon, Pavon76 and the introgression line were loaded into IGV (Robinson *et al.,* 2011). The region around the border identified above was searched manually to find the position where the coverage and SNP profile switches from that of the wheat parents to that of *Am. muticum*.

KASP validation

To validate the newly identified segment that had not been previously validated, a KASP™ genotyping assay was conducted as described in (Grewal *et al.,* 2020) (Method S6) (Table S4).

Junction validation using Oxford nanopore long reads

DNA from introgression line DH65 extraction was prepared using ligation sequencing kit SQK-LSK109 and sequenced to a depth of 7x on a MinION using the R9.4.1_RevD flow cell. Reads were filtered using NanoFilt (De Coster *et al.,* 2018) to remove reads below a quality score of 7 or a length of 1Kbp. Filtered reads were mapped to RefSeq v1.0 using minimap2 (Li, 2018) with parameters -axe map-ont and --secondary = no. Mapped reads around the breakpoint (chr4D:51283000–51 595 000) were extracted using samtools (Li *et al.,* 2009), including clipped portions of mapped reads, and assembled using wtdbg2 (Ruan and Li, 2020). The resulting contigs were mapped to RefSeq v1.0 using minimap2 (Li, 2018) with parameters -axe map-ont and visualized in IGV (Robinson *et al.,* 2011) along with the mapped Illumina paired-end short reads from the parent lines and DH65.

Genome assembly of *Am. muticum*

DNA from *Aegilops mutica* (now *Am. muticum*) line 2130012 (JIC) was prepared using ligation sequencing kit SQK-LSK109 and sequenced on a MinION using the R9.4.1_RevD flow cell. 178Gbp of raw Oxford Nanopore long reads were filtered using NanoFilt (De Coster *et al.,* 2018), removing reads below a quality score of 7 or a length of 1Kbp. Filtered reads were assembled using the Flye assembler (Kolmogorov *et al.,* 2019). Following polishing integrated into Flye using Oxford Nanopore reads, we conducted 2 rounds of pilon (Walker *et al.,* 2014) polishing using 102Gbp of Illumina paired-end short reads to correct systematic errors in the Oxford Nanopore reads. Finally, haplotigs that were not collapsed in the assembly were detected and resolved using purge_haplotigs (Roach *et al.,* 2018). Gene completeness was assessed using BUSCO 3.0.2 (Waterhouse *et al.,* 2018) with parameters -l viridiplantae_odb10 –species wheat and -m geno. Genome size of *Am. muticum* accession 2130012 was estimated by mapping back the Oxford Nanopore reads to putative single-copy genes and through a k-mer based approach (Method S7).

Gene annotation

Following annotation and masking of transposable elements (Method S8), gene annotation was performed using *ab initio*, protein homology and transcriptome evidence from *Am. muticum* root and shoot mRNAseq data (Method S9). These were sources of evidence were integrated using EvidenceModeler (Haas *et al.,* 2008) and partitioned into high- and low-confidence genes.

Protein family analysis

OrthoFinder (Emms and Kelly, 2019) was used with default settings to cluster the longest protein encoded by high-confidence genes from *Am. muticum*, *Ae. tauschii*, *T. urartu*, *T. aestivum*, *O. sativa* and *B. distachyon* into orthogroups. *Am. muticum* genes were classified as novel if in an orthogroup without a wheat protein or not assigned to an orthogroup. An orthogroup was determined to have expanded in *Am. muticum* compared to wheat if the orthogroup contained 4 or more *Am. muticum* proteins more than twice the number of proteins than wheat.

Assigning orthologue pairs

First, we computed best reciprocal blast hits between *Am. muticum* and each wheat subgenome independently. *Am. muticum* proteins (extracted and translated from gff) and wheat proteins (taken from IWGSC 1.1 pep.fa file) were aligned reciprocally using blastp (Camacho *et al.,* 2009) with parameters -outfmt 6 -max_hsps 3 -max_target_seqs 3 -evalue 1e-6. Hits were retained if percentage identity ≥90.0% and alignment length was ≥ 80.0% query length. An *Am. muticum* gene was placed in an orthologue pair with a wheat gene if it was in an orthogroup with that gene and the pair were each other's best reciprocal blast hit.

Classifying introgressed genes

The wheat reference genome RefSeq v1.0 and the draft *Am. muticum* assembly were concatenated to form a pseudo ABDT genome. Illumina paired-end short reads from the introgression lines were mapped to this genome and filtered using the same process as mapping to RefSeq v1.0 alone. Introgressed *Am. muticum* genes in each line were defined as those with mean depth across their length ≥ 13.2x in DH202 and ≥3x for the remaining lines (≥~0.6 * mean sequencing depth) from the ABDT pseudo genome mapping above and on a contig/scaffold with a gene assigned to an orthologue pair with a wheat gene whose start position is within a region labelled as a *Am. muticum* introgression and also passes the coverage threshold above. This is a conservative classification to prevent inclusion of non-introgressed genes.

mRNA extraction, sequencing, alignment and quantification

mRNA was extracted and sequenced in triplicate from leaf tissue of six introgression lines, Chinese Spring, Paragon and Pavon76 (Method S10). RNA reads were trimmed using Trimmomatic (Bolger *et al.,* 2014) with the parameters ILLUMINACLIP: BBDUK_adaptor.fa:2:30:12 SLIDINGWINDOW:4:20 MINLEN:20 AVGQUAL:20. The gff3 for the high confidence CS genes was concatenated with the gff3 for *Am. muticum* genes. Splice site hints for HISAT2 were produced using extract splice sites.py from HISAT-2.0.4 (Kim *et al.,* 2019). The trimmed reads were mapped to the pseudo ABDT genome using HISAT2 with the splice hint file provided and parameters -k 101 --dta --rna-strandness RF. Non-uniquely mapping reads were removed using samtools view -q 40. Stringtie (Pertea *et al.,* 2015) was used to compute gene-level abundances, outputting both raw counts and transcript-per-million (TPM) values.

Expression of introgressed *Am. muticum* genes

The protein sequences encoded by introgressed *Am. muticum* genes were aligned to the proteins encoded by RefSeq v1.1 HC genes using blastp (Camacho *et al.,* 2009). The identity of the best hit for each protein was retained, with an identity of 0 assigned to proteins with no hit. TPM values for each gene were taken as the mean of the three replicates. Genes with mean TPM greater than 1.0 were classified as expressed.

Differential expression analysis

For each wheat gene in a region identified as introgressed, they were either removed if not in an orthologue pair with an introgressed *Am. muticum* gene or their expression count was summed with that of its *Am. muticum* orthologue. Differential expression analysis between each introgression line and its two wheat parents was performed using DESeq2 (Love *et al.,* 2014). A gene was classified as differentially expressed if it had an adjusted *P*-value below 0.05 and an absolute $\log_2$FC ≥ 1 in both parental comparisons. Differentially expressed genes were partitioned into those in introgressed regions, and in the unaffected wheat background where coverage deviation is between 0.8 and 1.2.

Testing triad expression balancing

To examine whether genes belonging to triads that have homoeologues that have been replaced by a *Am. muticum* gene or have been deleted, we took test sets of triads (Ramírez-González *et al.,* 2018) that satisfied the following conditions: the D copy is introgressed or deleted and called as downregulated; the A and B homoeologues are in normal copy number regions (coverage deviation between 0.8 and 1.2); and all homoeologues have normalized expression count across samples ≥1. These were compared to control sets of triads that satisfied the same conditions except the D homoeologue was within a normal copy number region and was not called as differentially expressed. These sets were used for both the comparison of number of triads with A and/or B homoeologue upregulated and for the comparisons of the mean $\log_2$FC of the A and B homoeologues between the test and control sample of triads. The significance of these comparisons was tested using two-tailed Fisher's exact test and two-tailed *t* test, respectively.

GO term analysis

We transferred functional GO Term annotation from genes in the RefSeq v1.0 annotation to genes in the RefSeq v1.1 annotation if they shared greater that 99% similarity across greater than 90.0% of their length. Statistically enriched GO Terms within the differentially expressed background gene set were computed using the R package topGO (Alexa and Rahnenfuhrer, 2021) with the following parameters: nodeSize = 10; classicFisher test $P < 0.05$ and algorithm= 'parentchild'. Enrichment for GO Terms involved in biological processes was tested against all background genes that fall within windows with mapping coverage deviation between 0.8 and 1.2. For novel *Am. muticum* genes, GO terms were extracted from the eggnog functional annotation and converted to GO Slim terms using owltools Map2Slim (https://github.com/owlcollab/owltools). Enrichment was performed as above but against all *Am. muticum* HC genes.

Identifying introgressed resistance genes

Potential resistance genes in the *Am. muticum* assembly, including NLRs, Protein Kinases and ABC transporters were identified (Method S11). Resistance genes were manually checked using IGV to identify candidates with even sequencing coverage across the genes in DH92 and DH121 only, in the case of the shared stripe rust resistance, and across the genes in DH92 only, in the case of the DH92-specific leaf and stem rust resistance. To reduce the number of genes to manually check, we removed any genes with less than 2x mean mapping coverage across their length in either DH92 or DH121. The gene models were manually curated using the available evidence. For NLRs revealed by NLRAnnotator (Steuernagel *et al.,* 2020) with no gene model but transcriptomic and *ab initio* evidence, gene models were manually constructed. The novelty of the uniquely introgressed NLRs was tested by extracting the NB-ARC domains using hmmscan (Finn *et al.,* 2011) and aligning them using blastp (Camacho

et al., 2009) to the proteins of HC genes from 10 wheat cultivars (Walkowiak *et al.,* 2020). Hits below 85% identity were considered novel. The novelty of the other protein types was tested by aligning the whole amino acid sequence to the same protein set; here, hits below <80.0% were considered novel.

Acknowledgements

We would like to acknowledge BBS/E/T/000PR9816 (NC1 - Supporting EI's ISPs and the UK Community with Genomics and Single Cell Analysis) for data generation and BB/CCG1720/1 for the physical HPC infrastructure and data centre delivered via the NBI Computing infrastructure for Science (CiS) group.

Funding

BBSRC Core Capability Grant BB/CCG1720/1 (AH, RJ, RR-P). BBSRC funded Norwich Research Park Biosciences Doctoral Training Partnership grant BB/M011216/1 (BC). BBSRC Designing Future Wheat grant BB/P016855/1 and its constituent work packages DFW WP4 Data Access and Analysis (AH, RJ, RR-P, JK, IPK, SG, SE, CY). BBSRC grant BB/J004596/1 as part of the Wheat Improvement Strategic Programme (WISP) (JK, IPK, SG, SE, CY). USDA-ARS CRIS 3020–21000-011-000-D (JF).

Competing interests

The authors declare that they have no competing interests.

Author contributions

AH, JL, IPK and BC contributed to conceptualization. BC, RJ and RR-P contributed to methodology. BC contributed to formal analysis, visualization and writing—original draft. JF, SG, CY and SE contributed to investigation. BC, AH, RR-P, JK and JF contributed to writing—review and editing. AH contributed to supervision. AH, JK and IPK contributed to funding acquisition.

Data availability

Sequencing data produced as part of this study, along with the *Am. muticum* assembly is available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2021-10-08_wheatxmuticum. *Am. muticum* Illumina short-read sequencing reads available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Grewal_et_al_2021-09-13_Amybylopyrum_muticum/. The Chinese Spring sequencing data used is available from ENA (study PRJNA393343; runs SRR5893651 and SRR5893652). The Paragon sequencing data used are available from ENA (study PRJEB35709; runs ERR3728451, ERR3760033, ERR3760405 and ERR3728448). Custom scripts used for introgression detection are available at: https://github.com/benedictcoombes/alien_detection.

References

Alexa, A., Rahnenfuhrer, J., (2021) *topGO: enrichment analysis for gene ontology*. R package version 2.38.1.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Charmet, G. (2011) Wheat domestication: lessons for the future. *C. R. Biol.* **334**, 212–220.

Chen, S., Rouse, M.N., Zhang, W., Zhang, X., Guo, Y., Briggs, J. and Dubcovsky, J. (2020) Wheat gene Sr60 encodes a protein with two putative kinase domains that confers resistance to stem rust. *New Phytol.* **225**, 948–959.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669.

Dmochowska-Boguta, M., Kloc, Y., Zielezinski, A., Werecki, P., Nadolska-Orczyk, A., Karlowski, W.M. and Orczyk, W. (2020) TaWAK6 encoding wall-associated kinase is involved in wheat resistance to leaf rust similar to adult plant resistance. *PLoS One* **15**, e0227713. https://doi.org/10.1371/journal.pone.0227713

Dong, Z., Ma, C., Tian, X., Zhu, C., Wang, G., Lv, Y., Friebe, B. *et al.* (2020) Genome-wide impacts of alien chromatin introgression on wheat gene transcriptions. *Sci. Rep.* **10**, 4801.

Doussinault, G., Delibes, A., Sanchez-Monge, R. and Garcia-Olmedo, F. (1983) Transfer of a dominant gene for resistance to eyespot disease from a wild grass to hexaploid wheat. *Nature* **303**, 698–700.

Dover, G.A. and Riley, R. (1972a) Prevention of pairing of homoeologous meiotic chromosomes of wheat by an activity of supernumerary chromosomes of aegilops. *Nature* **240**, 159–161.

Dover, G.A. and Riley, R. (1972b) Variation at two loci affecting homoeologous meiotic chromosome pairing in Triticum aestivum x Aegilops mutica hydrids. *Nature New Biol.* **235**, 61–62.

Ellis, J.G., Lagudah, E.S., Spielmeyer, W. and Dodds, P.N. (2014) The past, present and future of breeding rust resistant wheat. *Front. Plant Sci.* **5**, 641.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238.

Fatih, A.M. (1983) Analysis of the breeding potential of wheat-Agropyron and wheat-Elymus derivatives. I. Agronomic and quality characteristics. *Hereditas* **98**, 287–295.

Fellers, J.P., Matthews, A., Fritz, A.K., Rouse, M.N., Grewal, S., Hubbart-Edwards, S., King, I.P. *et al.* (2020) Resistance to wheat rusts identified in wheat/ *Amblyopyrum muticum* chromosome introgressions. *Crop. Sci.* **60**, 1957–1964.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37.

Friebe, B., Gill, K.S., Tuleen, N.A. and Gill, B.S. (1996) Transfer of wheat streak mosaic virus resistance from *Agropyron intermedium* into wheat. *Crop. Sci.* **36**, 857–861.

Gardiner, L.-J., Wingen, L.U., Bailey, P., Joynson, R., Brabbs, T., Wright, J., Higgins, J.D. *et al.* (2019) Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency. *Genome Biol.* **20**, 69.

Grewal, S., Coombes, B., Joynson, R., Hall, A., Fellers, J., Yang, C., Scholefield, D., Ashling, S., Isaac, P., King, I., King, J., (2021) *A novel approach to develop wheat chromosome-specific KASP markers for detecting Amblyopyrum muticum segments in doubled haploid introgression lines. BioRxiv*. doi: https://doi.org/10.1101/2021.09.29.462370

Grewal, S., Othmeni, M., Walker, J., Hubbart-Edwards, S., Yang, C.-Y., Scholefield, D., Ashling, S. *et al.* (2020) Development of wheat-Aegilops caudata introgression lines and their characterization using genome-specific KASP markers. *Front. Plant Sci.* **11**, 606. https://doi.org/10.3389/fpls.2020.00606

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7.

Hao, M., Zhang, L., Ning, S., Huang, L., Yuan, Z., Wu, B., Yan, Z. *et al.* (2020) The resurgence of introgression breeding, as exemplified in wheat improvement. *Front. Plant Sci.* **11**, 252.

International Wheat Genome Sequencing Consortium (IWGSC). (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191. https://doi.org/10.1126/science.aar7191

Valkoun, J.J. (2001) Wheat pre-breeding using wild progenitors. *Euphytica* **119**, 17–23.

Khazan, S., Minz-Dub, A., Sela, H., Manisterski, J., Ben-Yehuda, P., Sharon, A. and Millet, E. (2020) Reducing the size of an alien segment carrying leaf rust and stripe rust resistance in wheat. *BMC Plant Biol.* **20**, 153.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915.

King, J., Grewal, S., Yang, C.-Y., Hubbart, S., Scholefield, D., Ashling, S., Edwards, K.J. *et al*. (2017) A step change in the transfer of interspecific variation into wheat from Amblyopyrum muticum. *Plant Biotechnol. J.* **15**, 217–226.

King, J., Newell, C., Grewal, S., Hubbart-Edwards, S., Yang, C.-Y., Scholefield, D., Ashling, S. *et al*. (2019) Development of stable homozygous wheat/Amblyopyrum muticum (Aegilops mutica) introgression lines and their cytogenetic and molecular characterization. *Front. Plant Sci.* **10**, 34.

Klindworth, D.L., Niu, Z., Chao, S., Friesen, T.L., Jin, Y., Faris, J.D., Cai, X., Xu, S.S., (2012) Introgression and characterization of a goatgrass gene for a high level of resistance to ug99 stem rust in tetraploid wheat *G3 (Bethesda)*, **2** 665–673. doi:Placeholder Text

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546.

Koo, D.-H., Friebe, B. and Gill, B.S. (2020) Homoeologous recombination: a novel and efficient system for broadening the genetic variability in wheat. *Agronomy* **10**, 1059. https://doi.org/10.3390/agronomy10081059

Krattinger, S.G., Lagudah, E.S., Spielmeyer, W., Singh, R.P., Huerta-Espino, J., McFadden, H., Bossolini, E. *et al*. (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science* **323**, 1360–1363.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.

Li, H., Deal, K.R., Luo, M.-C., Ji, W., Distelfeld, A. and Dvorak, J. (2017) Introgression of the Aegilops speltoides Su1-Ph1 suppressor into wheat. *Front. Plant Sci.* **8**, 2163.

H Li, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.

Luna, S.K. and Chain, F.J.J. (2021) Lineage-specific genes and family expansions in Dictyostelid genomes display expression bias and evolutionary diversification during development. *Genes (Basel)* **12**, 1628. https://doi.org/10.3390/genes12101628

Martín, A.C., Rey, M.-D., Shaw, P. and Moore, G. (2017) Dual effect of the wheat Ph1 locus on chromosome synapsis and crossover. *Chromosoma* **126**, 669–680.

Nyine, M., Adhikari, E., Clinesmith, M., Jordan, K.W., Fritz, A.K., Akhunov, E., (2020) Genomic Patterns of Introgression in Interspecific Populations Created by Crossing Wheat with Its Wild Relative *G3 (Bethesda)*, **10**, 3651–3661. doi: Placeholder Text

Pedersen, B.S. and Quinlan, A.R. (2018) Hts-nim: scripting high-performance genomic analyses. *Bioinformatics* **34**, 3387–3389.

Pellicer, J. and Leitch, I.J. (2020) The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* **226**, 301–305.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295.

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

RH Ramírez-González, Borrill, P., Lang, D., Harrington, S.A., Brinton, J., Venturini, L., Davey, M. *et al*. (2018) The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089. https://doi.org/10.1126/science.aar6089

Ray, D.K., Mueller, N.D., West, P.C. and Foley, J.A. (2013) Yield trends are insufficient to double global crop production by 2050. *PLoS One* **8**, e66428. https://doi.org/10.1371/journal.pone.0066428

E Rey, Abrouk, M., Keeble-Gagnère, G., Karafiátová, M., Vrána, J., Balzergue, S., Soubigou-Taconnat, L. *et al*. (2018) Transcriptome reprogramming due to the introduction of a barley telosome into bread wheat affects more barley genes than wheat. *Plant Biotechnol. J.* **16**, 1767–1777.

Rey, M.-D., Martín, A.C., Higgins, J., Swarbreck, D., Uauy, C., Shaw, P. and Moore, G. (2017) Exploiting the ZIP4 homologue within the wheat Ph1 locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids. *Mol. Breed.* **37**, 95.

Reynolds, M., Foulkes, J., Furbank, R., Griffiths, S., King, J., Murchie, E., Parry, M. *et al*. (2012) Achieving yield gains in wheat. *Plant Cell Environ.* **35**, 1799–1823.

Roach, M.J., Schmidt, S.A. and Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26.

Ruan, J. and Li, H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158.

Sears, E.R., (1956) *The Transfer of Leaf-Rust Resistance from Aegilops Umbellulata to Wheat*. pp. 1–22. Brookhaven National Laboratory (Upton: Biology Dept): Genetics in Plant Breeding.

Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.-J., Yu, G., Baggs, E. *et al*. (2020) The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A. *et al*. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963. https://doi.org/10.1371/journal.pone.0112963

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H. *et al*. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283.

Wang, H., Zou, S., Li, Y., Lin, F. and Tang, D. (2020) An ankyrin-repeat and WRKY-domain-containing immune receptor confers stripe rust resistance in wheat. *Nat. Commun.* **11**, 1353.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. *et al*. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548.

Zhang, A., Li, N., Gong, L., Gou, X., Wang, B., Deng, X., Li, C. *et al*. (2017) Global analysis of gene expression in response to whole-chromosome aneuploidy in Hexaploid wheat. *Plant Physiol.* **175**, 828–847.

Zhang, Z., Gou, X., Xun, H., Bian, Y., Ma, X., Li, J., Li, N. *et al*. (2020) Homoeologous exchanges occur through intragenic recombination generating novel transcripts and proteins in wheat and other polyploids. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 14561–14571.

Zheng, S., Wu, Y., Zhou, M., Zeng, L., Liu, R., Li, Y., Liu, Z. *et al*. (2020) Characterization and diagnostic marker development for Yr28-rga1 conferring stripe rust resistance in wheat. *Eur. J. Plant Pathol.* **156**, 623–634.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Whole genome macro-level plot for all 17 hexaploid wheat/*Am. muticum* introgression lines.

Figure S2 Macro structure of chr7A, chr7B and chr7D in four introgression lines (two DH pairs: DH195+DH202, and DH121+DH123).

Figure S3 Minimum required sequencing depth to uncover introgressed segments in introgression lines.

Figure S4 k-mer distribution of *Am. muticum* Illumina paired-end short reads used to estimate genome size.

Figure S5 BUSCO results using the viridiplantae_odb10 dataset after each sequential round of the assembly.

Figure S6 Interspecies intersection of orthogroups produced by OrthoFinder.
Figure S7 GO Slim terms enriched in novel *Am. muticum* genes.
Figure S8 GO terms enriched in differentially expressed wheat genes.
Figure S9 Testing compensation in homoeologue expression following deletion or introgression.
Table S1 Introgression lines sequenced in this study.
Table S2 Segments identified in each introgression line included in this study. If junction within or nearby a gene, the gene name is included.
Table S3 Genotyping results of DH15 with newly discovered small 6D segment. 'a' indicates presence of homozygous wheat-specific alleles; 'b' indicates presence of homozygous *Am. muticum*-specific alleles; '-' indicates absence of a genome-specific allele for a particular KASP assay.
Table S4 Primer details for the KASP assays used for genotyping of DH15.
Table S5 Metrics of *Am. muticum* genome assembly.
Table S6 Potential resistance genes uniquely introgressed in DH92 and DH121, which share resistance to stripe rust.
Table S7 Potential resistance genes introgressed in DH92 but absent from DH121 and the other lines. DH92 has stem and leaf rust resistance not seen in DH121.
Method S1 Introgression line production.
Method S2 DNA extraction and whole-genome sequencing.
Method S3 Read mapping and SNP calling.
Method S4 Producing *Am. muticum*-specific SNPs.
Method S5 Assigning coverage deviation blocks as *Am. muticum*.
Method S6 KASP genotyping.
Method S7 Estimating genome size.
Method S8 Repeat annotation and masking.
Method S9 Gene annotation.
Method S10 mRNA extraction and sequencing.
Method S11 Identifying resistance genes.

communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-04325-5 OPEN

Exotic alleles contribute to heat tolerance in wheat under field conditions

Gemma Molero[1,3,5], Benedict Coombes[2,5], Ryan Joynson[2,4], Francisco Pinto[1], Francisco J. Piñera-Chávez[1], Carolina Rivera-Amado[1], Anthony Hall[2✉] & Matthew P. Reynolds[1✉]

Global warming poses a major threat to food security and necessitates the development of crop varieties that are resilient to future climatic instability. By evaluating 149 spring wheat lines in the field under yield potential and heat stressed conditions, we demonstrate how strategic integration of exotic material significantly increases yield under heat stress compared to elite lines, with no significant yield penalty under favourable conditions. Genetic analyses reveal three exotic-derived genetic loci underlying this heat tolerance which together increase yield by over 50% and reduce canopy temperature by approximately 2 °C. We identified an *Ae. tauschii* introgression underlying the most significant of these associations and extracted the introgressed *Ae. tauschii* genes, revealing candidates for further dissection. Incorporating these exotic alleles into breeding programmes could serve as a pre-emptive strategy to produce high yielding wheat cultivars that are resilient to the effects of future climatic uncertainty.

[1] International Maize and Wheat Improvement Center (CIMMYT), Texcoco 56237, Mexico. [2] The Earlham Institute, Norwich NR4 7UZ, UK. [3] Present address: KWS Momont Recherche, 59246 Mons-en-Pévèle, Hauts-de-France, France. [4] Present address: Limagrain Europe, Clermont-Ferrand, France. [5] These authors contributed equally: Gemma Molero, Benedict Coombes. ✉email: anthony.hall@earlham.ac.uk; m.reynolds@cgiar.org

Wheat is among the most widely cultivated crops in the world with more than 216 million hectares grown annually[1], most of which is produced under temperate conditions[2]. Heat stress is one of the major abiotic stressors that impacts global wheat production, reducing leaf area, crop duration and the efficiency of photosynthesis and respiration[3] as well as reducing floret fertility and individual grain weight[4]. Together, these physiological consequences negatively impact productivity[3] with potential devastating effects. For example, in 2010, Russia saw a 30% reduction in wheat yield during their hottest summer in 130 years[5]. Cases like this could become commonplace as global warming causes temperatures to rise and extreme weather events to become more frequent. Simulations predict that global yields will fall by on average 6% for each 1 °C increase in temperature[6], with some regions reaching 9.1% ± 5.4% per 1 °C rise[7]. Adaptation to future climate scenarios is vital to ensure global food security[8]. Climatic instability, combined with environmental constraints, such as restricted supplies of irrigation water and arable land loss, emphasises the need for breeding strategies that deliver both increased yield potential during favourable cycles and resilience to abiotic stress and environmental constraints.

Such adaptation relies on genetic variation underlying the traits of interest; however, modern elite wheat material typically has limited genetic variation, particularly in the D genome[9], due to historic genetic bottlenecks[10,11] compounded by intensive artificial selection by breeders[12]. A strategy employed by the International Maize and Wheat Improvement Center (CIMMYT) to increase the genetic diversity of wheat pre-breeding material is to incorporate exotic parents in their germplasm via strategic crosses[11,13]. The most common exotic parents used are Mexican and other origin landraces[14] and primary synthetics, which are produced by hybridising tetraploid durum wheat with *Aegilops tauschii*, the ancestral donor of the D genome, to recreate hexaploid bread wheat[15]; these synthetic lines act as a bridge to introduce durum and *Ae. tauschii* variation into modern hexaploid wheat. This approach has been successful in introducing disease resistance as well as drought and heat adaptive traits[16,17]. Landrace and synthetic material have been identified with superior biomass in comparison to elite lines under drought and heat conditions[18,19] and elite lines that include landrace or synthetic material in their background have been developed in recent years for drought, heat, and yield potential conditions[20–22].

Challenges remain for the effective deployment of landrace and synthetic material. Only a small fraction of these vast collections of crop genetic resources have been evaluated for climate resilience traits and potential tradeoffs under favourable conditions have not been assessed. Currently, most of these genetic resources are unused[23] as breeders tend to avoid exotic materials because of large regions of poor recombination and a fear of linkage drag[24]. Furthermore, despite evidence of the contribution of exotic material in wheat improvement, the physiological and genetic bases of heat tolerance in this material remain unclear.

Here, we evaluate a spring wheat panel in the field containing contrasting material controlled for phenology and plant height under heat stress and yield potential conditions. We explore yield and related physiological traits and compare exotic-derived lines with elite lines. We conduct a genome-wide association study to reveal marker trait associations (MTA) with heat tolerance traits and evaluate their impact under favourable conditions. Finally, we identify introgressed *Ae. tauschii* underlying an MTA and employ in silico mapping downstream of the GWAS to narrow down the interval, explore recombination and identify candidate genes.

Results

Physiological evaluation of HiBAP I under heat stress. To estimate the contribution of exotic material to heat tolerance and identify its genetic bases, we evaluated the High Biomass Association Panel I (HiBAP I) for two consecutive years under yield potential and heat stressed irrigated conditions in NW-Mexico (Supplementary Table 1). The HiBAP I panel represents an unprecedented resource of genetic diversity[25]. It contains 149 lines, some of which are elite while others contain exotic material from landraces, synthetics, and wild relatives (Fig. 1a, Supplementary Data 1). All lines have agronomically acceptable backgrounds and a restricted range of phenology and plant height under yield potential conditions[21] which allows traits of interest to be evaluated without confounding effects.

Heat stress was imposed by delayed sowing compared to the check environment (Supplementary Fig. 1) and, across both years of evaluation, this reduced yield by 48.1% and shortened the crop cycle duration by more than 30% (Fig. 1b, Supplementary Table 3). When we analysed the response to heat stress of the lines based on their pedigree, exotic-derived lines exhibited an average of 37.7% higher yield compared to elite lines under heat stressed conditions (Fig. 1c, upper). Biomass, the trait most affected by heat stress, was 39% higher in exotic-derived lines, and other yield components, except for harvest index (HI), were significantly higher in exotic-derived lines than elite lines (Fig. 1c, upper). Under yield potential conditions, exotic-derived lines did not show a yield penalty compared to elite lines, as reported in[21] (Fig. 1c, lower). Exotic lines were on average 5.6 cm and 3.8 cm taller than elite lines under heat stressed and yield potential conditions, respectively. No differences in phenology were observed between the groups in either of the environments. Plant height was not correlated with yield under yield potential conditions ($r = -0.007$, $p > 0.05$), but positive correlations were observed between plant height and yield under heat stressed environments ($r = 0.699$, $p < 0.001$). The better performance of exotic-derived lines was validated using the stress susceptibility index (SSI). This measure is negatively correlated with yield under heat stressed conditions; thus, lower SSI values indicate higher tolerance to a stressful environment. Compared to elite lines, exotic-derived lines had significantly lower SSI values for yield, grains per m^2 and biomass at physiological maturity, but not for thousand grain weight (Table 1).

Additional physiological traits were measured in the experiments to help understand the physiological basis of the superiority of the exotic-derived lines under heat stressed conditions. Exotic-derived lines had significantly higher normalised difference vegetative index (NDVI), a proxy for biomass, and significantly lower canopy temperature during both vegetative and grain filling stages under heat stressed conditions but not under yield potential conditions (Fig. 2). NDVI measured during vegetative and grain filling stages was positively correlated with yield (Fig. 2) while canopy temperature was negatively correlated with yield at both stages (Fig. 2). These correlations were present under heat stressed conditions but not under yield potential conditions. The correlations were steeper for exotic-derived lines than elite lines for both traits across both phenological stages, suggesting that NDVI and canopy temperature are having a higher impact on yield in exotic-derived lines compared with elite lines. Under heat stressed environments, both NDVI and canopy temperature presented similar correlations with biomass at physiological maturity, grain number, and other yield components (Supplementary Table 5) but no correlation was observed under yield potential. The stess susceptibility index index calculated for yield was negatively correlated with agronomic and physiological traits except for canopy temperature, where positive correlations were observed

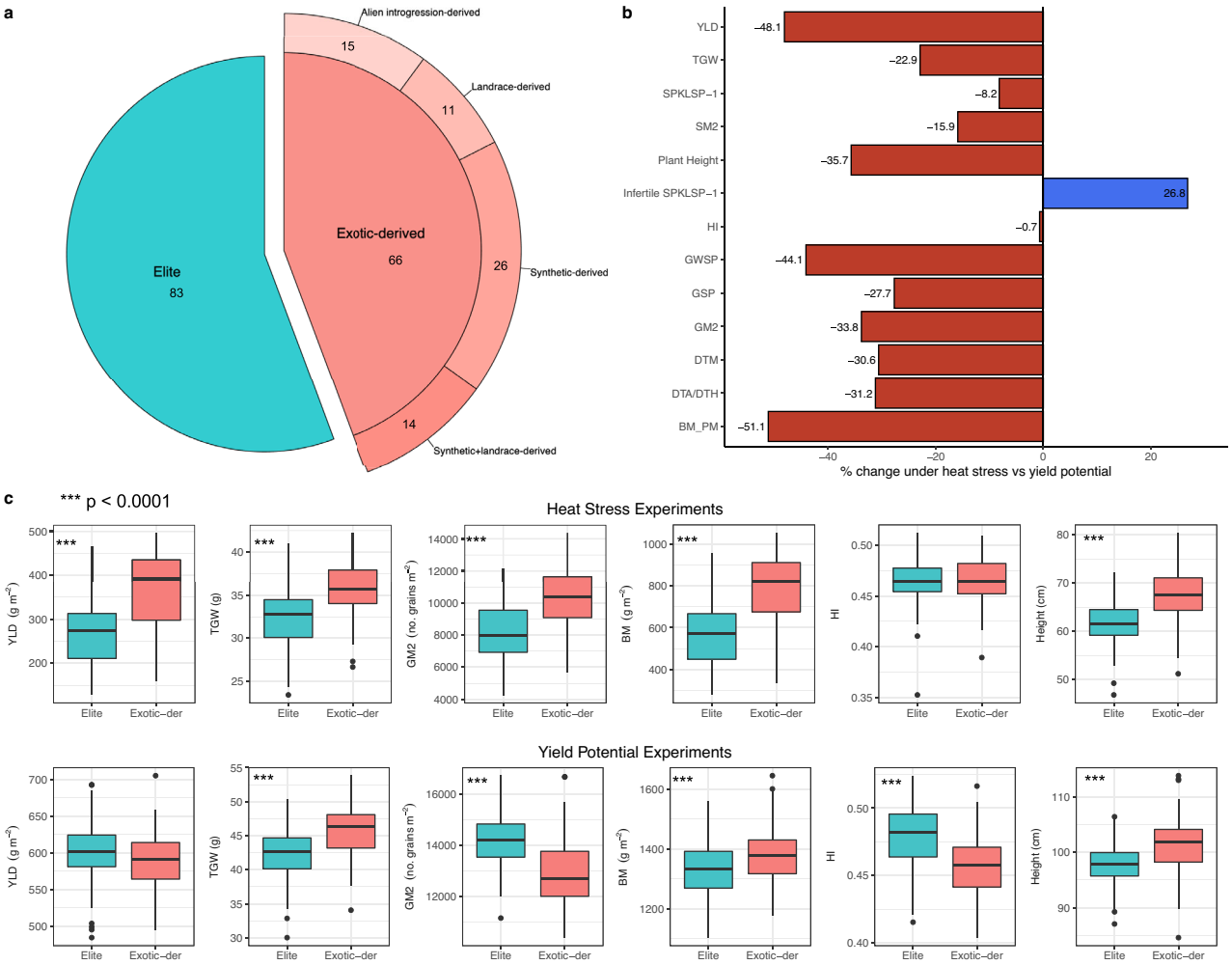


Fig. 1 Physiological assessment of HiBAP I panel, comparing elite and exotic-derived lines under heat stressed and yield potential conditions. **a** Number of lines from each group. **b** Effect of heat stress on yield (YLD), thousand grain weight (TGW), number of spikelets per spike (SPKLSP^{-1}), number of spikes per m^2 (SM2), plant height (Height), number of infertile spikelets per spike (infertile SPKLSP^{-1}), harvest index (HI), grain weight per spike (GWSP), number of grains per spike (GSP), grain number (GM2), days to physiological maturity (DTM), days to anthesis or days to heading (DTA/DTH for yield potential and heat stress experiments respectively), and biomass at physiological maturity (BM_PM), showing the percentage difference compared to yield potential conditions. **c** Comparison of yield (YLD), thousand grain weight (TGW), grain number (GM2), biomass at physiological maturity (BM_PM), harvest index (HI), and Height between elite and exotic-derived lines in HiBAP I measured under both heat stress and yield potential conditions. The boxplots are defined as follows: centre line = median; box limits = upper and lower quartiles, whiskers = 1.5x interquartile range; points = outliers. The significance of the difference between Elite ($n = 83$ biologically independent lines) and exotic-derived ($n = 66$ biologically independent lines) lines for each trait was assessed using two-tailed *t* tests with no assumption of equal variance. *p*-values below 0.01 were considered significant (*), below 0.001 very significant (**) and below 0.0001 highly significant (***). Means, standard deviations, confidence intervals and *p*-values can be found in Supplementary Table 2.

indicating that more tolerant lines had consistently cooler canopies (Supplementary Table 5).

Genome-wide association analysis reveals genetic associations under heat stress. To explore the genetic bases of these exotic-derived heat tolerance traits, marker-trait association analyses were performed using Best Linear Unbiased Estimator (BLUE) means from two or four replicates for each measured trait over two growing seasons. The most relevant MTAs are shown in Supplementary Table 6, and all Manhattan plots are shown in Supplementary Fig. 2. We found 3 pleiotropic markers (Fig. 3a) on chr1B (chr1B-63398861: C), chr2B (chr2B-820002: C) and chr6D (chr6D-6276646: T). These MTAs were associated with all three heat stress indices along with multiple yield traits, including yield and canopy temperature, at both vegetative and grain filling stages, and were not associated with harvest index or phenology (Fig. 3c, Fig. S2). The favourable allele at each position was the minor allele.

Lines with the favourable C allele on 1B and 2B and the unfavourable A allele on 6D have 24.3% higher yield under heat stress; lines which also have the favourable T allele on 6D have 56.5% higher yield under heat stress compared to lines with the three unfavourable alleles (Fig. 3b). Assuming the three alleles do not interact epistatically, the T allele on 6D can be estimated to increase yield under heat stress by 32.4%. Lines with the favourable allele at all three MTAs show a reduction in canopy temperature of 1.97 °C and 2.37 °C, at vegetative and grain filling stages, respectively, when compared to lines with the unfavourable allele at all three positions (Fig. 3b). Under yield potential conditions, no difference was observed between favourable and unfavourable allele combinations for yield or for canopy temperature (Fig. 3b). The

Table 1 Stress susceptibility index (SSI) calculated for yield (YLD), thousand grain weight (TGW), grains per m^2 (GM2) and biomass at physiological maturity (BM_PM) of elite and exotic-derived lines obtained from adjusted means for two years of data in each environment.

Trait	r_p (YLD_Heat)	Elite		Exotic	
		$n = 83$		$n = 66$	
SSI_YLD	−0.976	1.16 ± 0.24	a	0.79 ± 0.32	b
SSI_TGW	−0.439	1.03 ± 0.2	a	0.95 ± 0.27	a
SSI_GM2	−0.951	1.26 ± 0.42	a	0.60 ± 0.52	b
SSI_BM_PM	−0.946	1.13 ± 0.20	a	0.85 ± 0.23	b

Letters indicate the statistical significance between Elite and Exotic groups. Means followed by different letters are significantly different (p-value < 0.01) according to pairwise t tests. r_p corresponds to the phenotypic correlation with the yield obtained under heat environments. Data represents the mean ± S.D. Sample size, n, indicates the number of biologically independent lines in each group.

favourable allele at each of these MTAs is predominantly found in exotic-derived lines with 50/55 (1B), 44/45 (2B) and 33/33 (6D) lines with the favourable allele classified as exotic-derived. 7 lines appear to be heterozygous (A/T) at 6D-6276646. The HiBAP lines are inbred to at least the F9 or F10 generation so, as sequencing data was generated from pooled samples, this observation could be the result of alleles segregating at this locus. However, we observe no significant difference in yield or canopy temperature under heat stress between lines that are heterozygous and lines that are homozygous for this allele (Supplementary Fig. 3). This suggests that these lines are indeed heterozygous for the favourable allele and also suggests that the phenotype may be dominantly inherited.

Aegilops tauschii introgression underlies 6D MTA. Due to the better performance of exotic-derived lines under heat stress and exotic-derived lines possessing alleles for heat tolerance, we searched for introgressed material overlapping the MTAs. We detected introgressed material in HiBAP I lines by looking for genomic blocks containing windows with SNPs specific to *Ae. tauschii*, *Th. ponticum* or *S. cereale* and reduced mapping coverage, seen as coverage deviation (mapping coverage compared to the median mapping coverage across the panel) significantly below 1, which indicates breaks in synteny between wheat and the introgressed chromosome segment. Using this approach, we identified introgressed *Ae. tauschii* material at the beginning of 6D in all 33 lines with the T/T genotype and all 7 lines with the A/T genotype at MTA 6D-6276646, where T is the favourable allele. As *Ae. tauschii* is from wheat's primary genome and thus very similar to the D subgenome, not every 1Mbp window is sufficiently lacking in synteny for reads to map poorly and produce significant coverage deviation below 1. This explains why some windows within the introgression have coverage deviation of around 1. However, these windows still have *Ae. tauschii*-specific SNPs and are within a block of 1Mbp windows in which most have significant coverage deviation below 1. Therefore, we can be confident that the introgression includes these windows.

The full-length, unbroken segment is 31.6Mbp in length, as seen in Sokoll (HiBAP_57) (Fig. 4a). The segment size within independent Sokoll Weebil1 crosses show that recombination occurs readily within the segment, breaking it up into variable sizes (Supplementary Fig. 4). By comparing the overlapping segments between lines, we found a 1.80Mbp core introgressed region between 5.05Mbp and 6.85Mbp that is present in all lines with the T/T or A/T genotype at 6D-6276646 and absent in all the lines with the A/A genotype (Fig. 4a). In A/T lines, the introgression itself, in addition to the favourable allele, appears to be heterozygous, evidenced by intermediate mapping coverage deviation compared to the homozygous lines and by heterozygous SNPs whose alternative alleles are specific to *Ae. tauschii*. Using chromosome and protein alignments, we anchored this 1.80Mbp core region from the wheat RefSeq v1.0 genome to the *Ae. tauschii* reference genome, Aet v4.0[26], and extracted the syntenic 1.49Mbp region between 4.63Mbp and 6.12Mbp. This represents the probable introgressed content of the core introgressed region and likely contains the gene(s) responsible for the MTA (Fig. 4b, c). We found no evidence of introgressed material overlapping the 1B or 2B MTAs.

Candidate genes for MTAs in 1B, 2B and 6D. For the 6D MTA, we identified the syntenic region in the *Ae. tauschii* genome and a list of genes that had been introgressed (Fig. 4c). As we are unaware of the *Ae. tauschii* accession that has been introgressed, we also looked at the genes within the same region in four other available chromosome-level *Ae. tauschii* assemblies[27]. Between accessions, this region varies between 1.49Mbp and 1.82Mbp in length and contains between 26 and 33 genes (Supplementary Data 2). These include a MIKC-type MADS-box gene orthologous to *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*); a mitogen-activated protein kinase (*MAPK*) gene found in two of the *Ae. tauschii* accessions with no orthologue in wheat; and a pair of type-B two-component response regulator receiver proteins, orthologous to type-B *Arabidopsis* response regulators (ARRs) with closest similarity to *ARR-11*. One member of the pair, AET6Gv20025700, appeared to have a myb-binding domain that is missing from the wheat orthologue gene model. However, after manual reannotation, we found that this difference was a misannotation in wheat so likely not causing a functional difference. We also found that both ARR genes were expressed in spike and grain in both *Ae. tauschii* and wheat but not in leaf or root, whereas the other candidate genes were expressed across leaf, root, spike and grain tissues. This might exclude the ARR genes for involvement in the heat tolerant phenotype which is established during the vegetative stage and maintained through grain filling. For the 1B and 2B MTAs, as they were not within an introgression, we submitted the sequence 1Mbp up and downstream of the MTA to Knetminer, a gene discovery tool[28]. Within the 2B interval, we identified *DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN 1A* (*DREB1A*) and *STEROL GLUCOSYLTRANSFERASE* (*SGT*) as promising candidate genes. The functional evidence of candidacy for each candidate gene is outlined in Supplementary Note 1.

Discussion

Exotic parents are routinely used to increase genetic diversity in wheat pre-breeding pipelines and their enhanced performance has been demonstrated under salinity[29], drought[14,19] and heat stress[18,30]. In the present study, exotic-derived lines performed better under heat stress than elite lines with no yield penalty under yield potential conditions. This increased yield under heat stress was associated with a range of factors, including higher biomass throughout the crop cycle, higher grain number and cooler canopy temperature during both vegetative and grain filling stages. Contrary to other studies[31,32] higher pre or post anthesis biomass (NDVI) or lower canopy temperature was not associated with higher yields under favourable conditions. Cooler canopies have been previously associated with higher tolerance to drought and heat irrigated environments[33] and with optimised root distribution in bread wheat[34]. Plants with an optimised root system are able to satisfy the high evaporative demand through elevated transpiration rates under hot irrigated conditions and thus maintain cooler canopies[35]. Higher transpiration rates are associated with increased

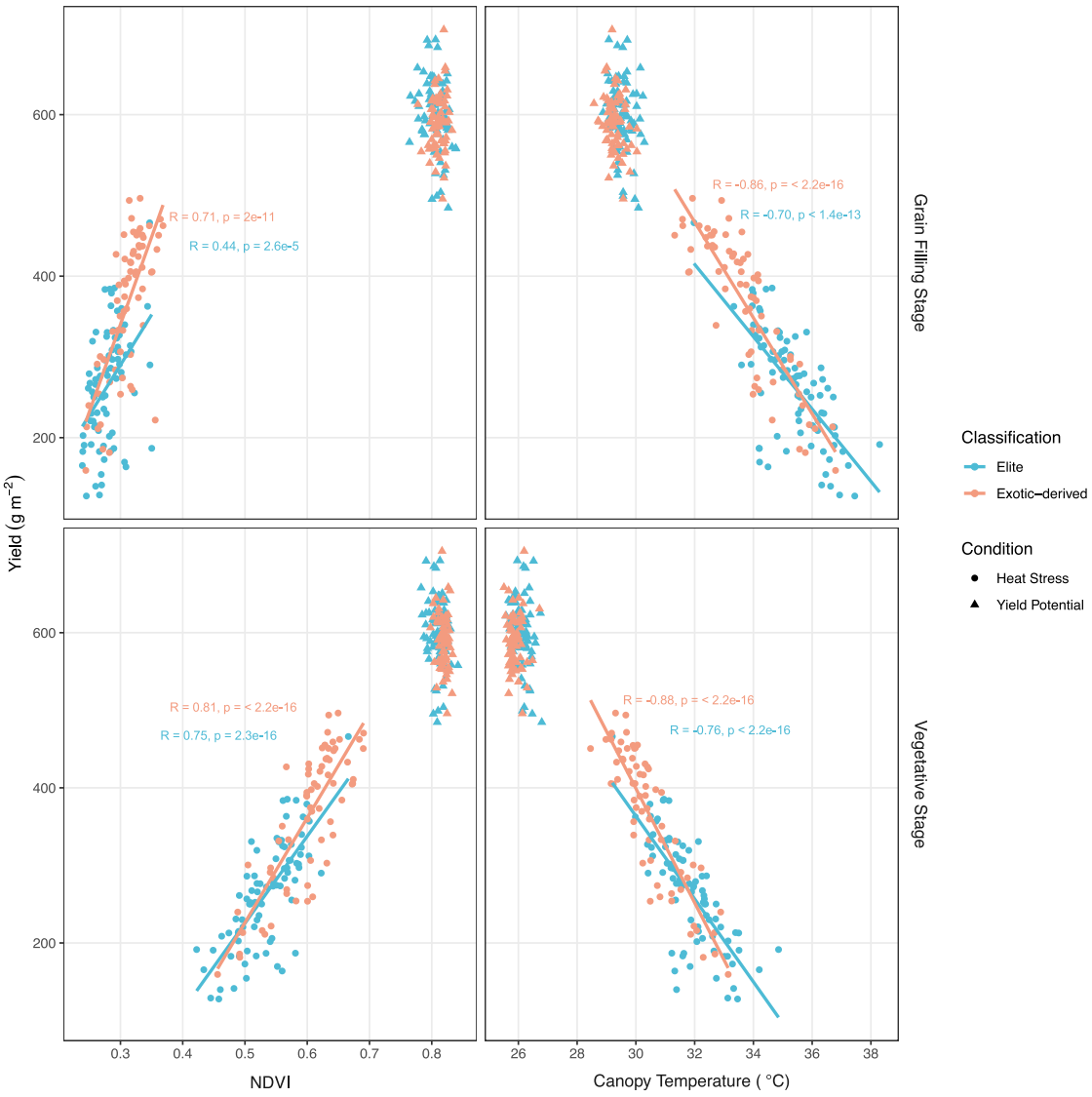


Fig. 2 Relationship between normalised difference vegetation index (NDVI) and yield and between average canopy temperature and yield at both grain-filling and vegetative stages. NDVI and canopy temperature were measured with UAVs at pre-heading (vegetative stage) and during grain filling. Regression lines were calculated using Pearson's correlation coefficient between each pair of traits (n = 83 and 66 biologically independent lines for the Elite and exotic-derived groups, respectively) and added for classification/condition combinations with a significant correlation (p-value < = 0.01). The correlation coefficient, r, and the steepness of the line, ranges from −1 to 1, signifying very negatively correlated and very positively correlated, respectively. Pearson's correlation coefficient, confidence intervals and p-values for all comparisons can be found in Supplementary Table 4.

stomatal conductance that, in turn, is associated with higher photosynthesis that can explain the higher biomass observed in exotic-derived lines in comparison with elite lines. However, according to temperature response models in wheat[6], the observed reduction in plant temperature of approximately 2 °C would be unlikely to account alone for the >50% increased yield of exotic lines[6].

Despite variation in plant height being restricted, exotic-derived lines were taller than elite lines in both environments. Plant height and phenology were restricted under yield potential conditions, but the variation under heat stress environments was not initially considered for the panel construction. Interestingly, lines that performed well under heat stress had the lowest difference in plant height between yield potential and heat stress conditions. Taller plants have better light interception and a better light distribution in comparison with shorter plants, and this has been associated with increased photosynthesis[36]. Therefore, plant height may be influencing the better performance of exotic-derivatives. Among all stress indices, the stress susceptibility index (SSI) is thought to be the most useful index for evaluating tolerant cultivars. Exotic-derived lines had significantly lower SSI than elite lines, adding additional support to the resilience of this exotic material under heat stress.

In the present study, heat stress was achieved by delaying sowing by more than three months. This could have introduced confounding effects as delayed sowing changes not only temperature but also photoperiod. However, the photoperiod effect in this study is considered minimal for several reasons. Firstly, the lines presented in this study were selected using the shuttle breeding technique that characterises CIMMYT's wheat breeding strategy and selects lines relatively insensitive to photoperiod and vernalisation response. This is because one selection site has a long photoperiod and negligible vernalizing cold[37]. Secondly, insensitivity to the photoperiod was confirmed by marker analysis where among the known major adaptation genes, the spring allele at *Vrn-B1* (*Vrn-B1a*) and *Vrn-D1* (*Vrn-D1a*) and the Ppd-insensitive allele at *Ppd-D1* (*Ppd-D1a*) were present in ~90% of

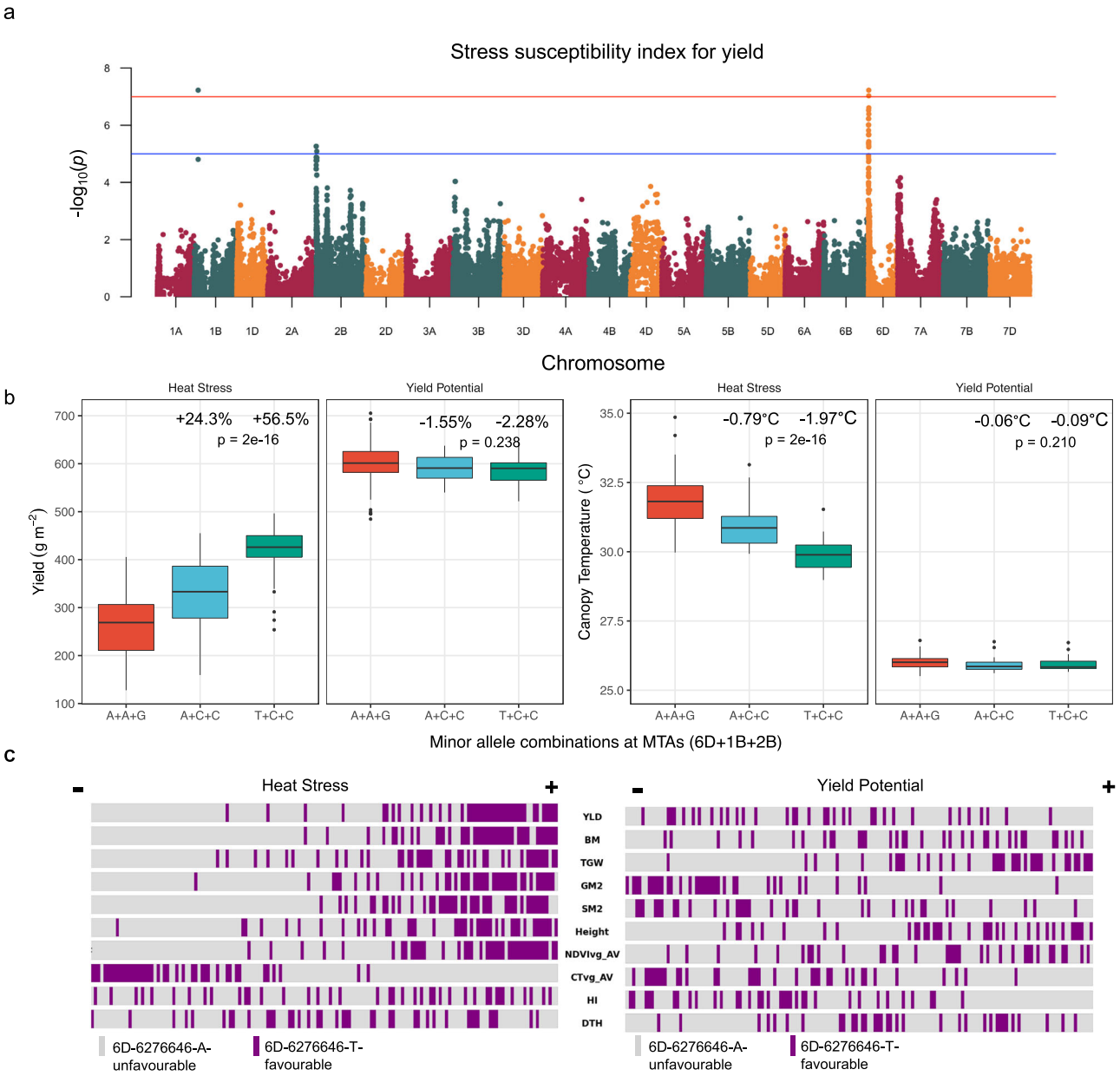


Fig. 3 Genome-wide association study reveals genetic markers underlying heat tolerance traits. a Manhattan plot showing marker trait associations for stress susceptibility index (SSI) for yield under heat stress. The horizontal blue line indicates an arbitrary cutoff of $-\log_{10}(p)$ of 5. The horizontal red line indicates the conservative Benjamini–Hochberg cutoff implemented by GAPIT. **b** Specific marker allelic variants effects on yield and on canopy temperature under heat stress and yield potential conditions in chromosomes 6D (chr6D-6276646), 1B (chr1B-63398861), and 2B (chr2B-820002), where the combination of favourable alleles is T+C+C and the combination of unfavourable alleles is A+A+G. The boxplots are defined as follows: center line = median; box limits = upper and lower quartiles, whiskers = 1.5x interquartile range; points=outliers. The percentage change and °C change is calculated compared to lines with the major alleles at all three MTAs. Significance of allele combinations was computed using a one-way ANOVA test ($n = 87$, 14, and 31 biologically independently lines for A+A+G, A+C+C and T+C+C, respectively). Means, standard deviations and p-values from Tukey's honest significance test can be found in Supplementary Table 7. **c** Phenotype distribution under heat stress and yield potential conditions highlighting the rank of 6D minor allele carriers for each phenotype where lines in the panel are ordered from lowest to highest value for each trait.

the HiBAP I panel[38]. Finally, delayed sowing at CIMMYT's Obregon station is routinely used for evaluating heat-tolerant breeding material, and several studies confirm the value of late sowing at this experimental site to develop germplasm adapted to different heat stressed environments worldwide[4,18,39–41].

De novo SNP discovery is the process of generating SNP markers from high-throughput next-generation sequencing as opposed to using lower density genotyping arrays. The value of this approach in breeding efforts is starting to be more widely recognised. In conjunction with high throughput phenotyping methods[42], high density, unbiased markers can be leveraged to discover MTAs or to narrow pre-existing QTL intervals to provide more robust markers for global breeding programs[43,44]. Using these methods, we have identified alleles at three pleiotropic MTAs on chromosomes 1B, 2B and 6D that when stacked increase yield by 56.5% and reduce canopy temperature by 1.97 °C/2.37 °C under heat stress conditions when compared to lines containing the three major alleles at these positions (Fig. 3b). These markers were associated with multiple agronomically important traits under heat stress including yield, grain per

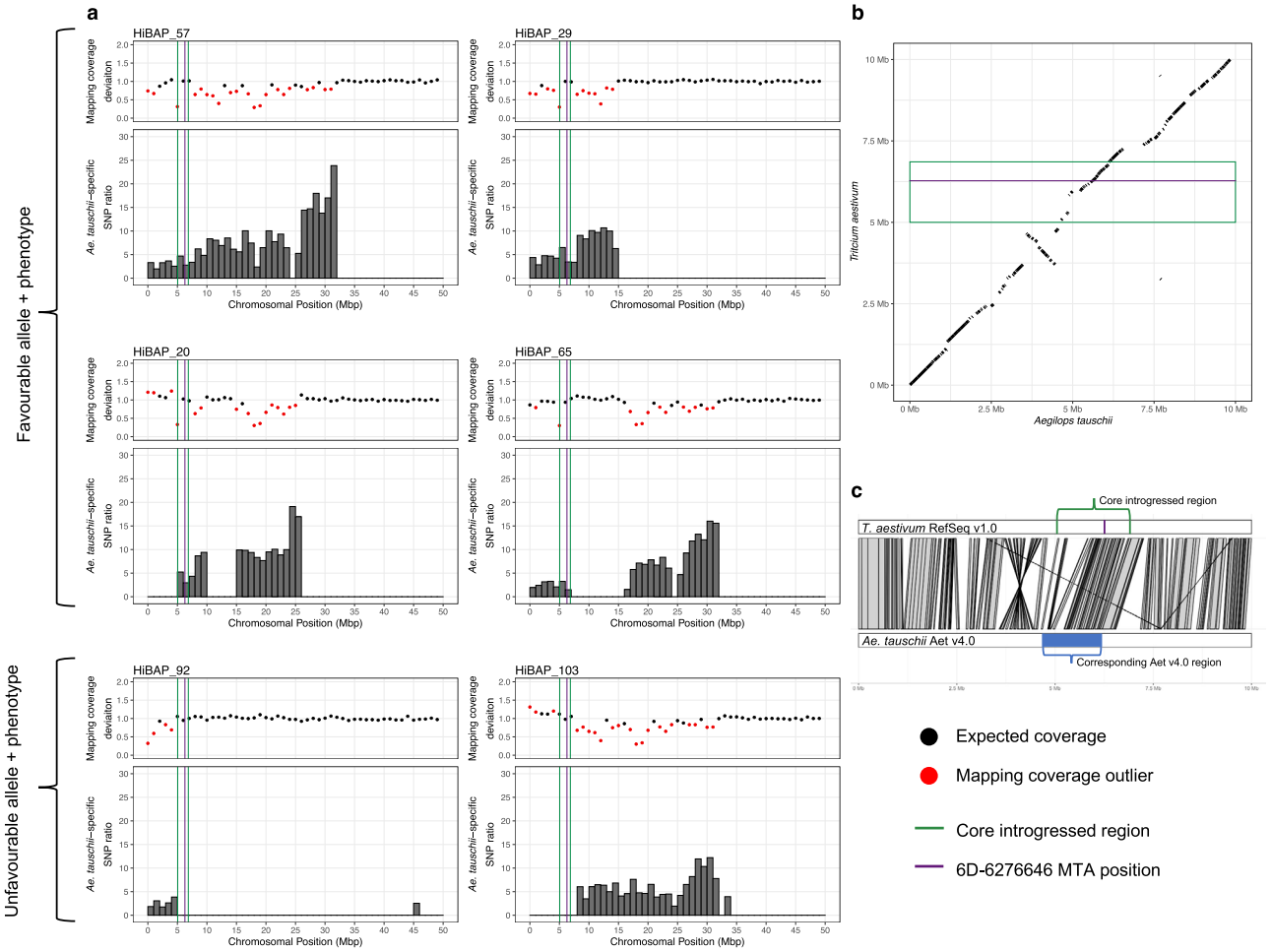


Fig. 4 *Aegilops tauschii* introgression underlies chr6D-6276646 MTA. a Visualising *Ae. tauschii* introgressions across the first 50Mbp of chr6D in six HiBAP I lines, four containing the favourable T allele at chr6D-6276646 (HiBAP 57, 29, 48, and 65) and two containing the unfavourable A allele at chr6D-6276646 (HiBAP 92 and 103). Mapping coverage deviation was computed between the HiBAP line and the median of the panel in 1Mbp windows. Red points are statistically significant outliers ($n = 149$ biologically independent lines). *Ae. tauschii*-specific SNP ratio in each 1Mbp window was calculated by dividing the number of homozygous *Ae. tauschii*-specific SNPs in that window by mean number of homozygous *Ae. tauschii*-specific SNPs in that window across the panel and then removing values below 1.45. Green lines mark the borders of the region common to all lines with the favourable T allele, corresponding to a 1.80Mbp region in wheat RefSeq v1.0 and a 1.49Mbp in *Ae. Tauschii* Aet v4.0. The purple line indicates the MTA position. **b** Synteny between 6D:1-10,000,000 in CS RefSeq v1.0[52] and *Ae. tauschii* Aet v4.0[26]. The green box indicates the 1.80Mbp region (1.49Mbp relative to *Ae. tauschii*) common to all lines with the favourable T allele, corresponding to the green region in (**a**). The purple line indicates the MTA position. **c** Alignment of 6D:1-10,000,000 in CS RefSeq v1.0[52] and 6D:1-10,000,000 in *Ae. tauschii* Aet v4.0[26], illustrating how the syntenic region in *Ae. tauschii* was identified and extracted.

square metre, grain filling rate and biomass (Fig. S2). Despite being in apparently disparate regions of the genome, the 1B and 2B favourable alleles always occur together and the 6D favourable allele usually occurs with the 1B and 2B favourable alleles. This suggests that there may be functional linkage between the markers. All three MTAs are predominantly found in exotic-derived lines but are not exclusive to any of the exotic categories as we see them in synthetic, introgression line and landrace derivatives. This brings their origin into question as their most recent pedigree suggests that the favourable alleles may have come from different sources. These MTAs do not overlap with MTAs previously identified for HiBAP I for photosynthetic efficiency[25] or biomass traits[21]. We identified several individuals that appear to be heterozygous for the introgression and for the favourable allele on 6D. As sequencing was conducted on pooled samples of 10 individuals per line, lines that appear heterozygous might instead be segregating for presence/absence of a homozygous introgression. As the phenotype under heat stress appears to be the same between lines that are homozygous and lines that are heterozygous at this locus, it seems likely that these lines are heterozygous for the introgression and allele. In addition, this also indicates that the heat tolerant phenotype contributed by 6D may be dominantly inherited; however, additional work would be needed to validate this. The zygosity of the allele in these lines can be verified in future work by developing markers and observing how they segregate in subsequent generations.

By utilising mapping coverage information and species-specific SNPs, we identified that the MTA on 6D was within an *Ae. tauschii* introgression. We show that this introgression readily recombines within CIMMYT germplasm by comparing the introgressed segment in different offspring of the same cross. Due to concerns regarding linkage drag and lack of recombination of wild relative introgressions[44], this is promising for the deployment of introgressed segments from the primary genepool into breeding programmes. The recombination enabled us to reduce the size of the interval responsible for the MTA by looking for the

region always present in lines with the favourable genotype. The smallest segment in the panel that contains the MTA is around 5Mbp and can likely be broken down further; the small size and it's telomeric location make it amenable for deployment in breeding programmes.

The longest unbroken segment is present in Sokoll, a commonly used advanced synthetic-derived line. Recombination within the segment takes place in all Sokoll × Weebil1 crosses yet appears unbroken in Sokoll. Therefore, Sokoll may be the donor line for this marker in many of the lines in HiBAP I. This would make sense given its presence in many of the pedigree histories of CIMMYT's synthetic-derived lines (Supplementary Data 1). Some of the HiBAP I lines contain an *Ae. tauschii* segment that contains both the 6D MTA and a resistance gene upstream that underlies an MTA from a recent GWAS in *Ae. tauschii*[45]. If the accession of *Ae. tauschii* in the HiBAP I lines confers the same resistance these lines could be used as donors for both traits.

The 6D MTA uncovered is supported by an MTA for heat tolerance reported in[13,22] nearby on 6D. Singh et al., 2018[13] state that the 6D MTA overlapped with an *Ae. tauschii* introgression, using speculative markers and pedigree-based inference. Here, we confirm this speculation and then demonstrate its ability to recombine and narrow down the introgressed region conferring the heat resilient phenotype through in silico introgression mapping.

Following the identification of the core introgressed region, we extracted the syntenic region from five *Ae. tauschii* chromosome-level assemblies and used these, as opposed to the wheat reference genome, as our source for putative candidate genes underlying the 6D MTA. Extensive literature searches on the introgressed *Ae. tauschii* genes or the wheat genes within the interval (for the 1B and 2B MTAs) uncovered several candidate genes for further dissection. Candidate genes are by their nature speculative but may provide a starting point for follow-up studies aiming to map the causal genes. As the 6D *Ae. tauschii* segment appears to be actively recombining, it should be possible to precisely dissect this region and map the causal gene. Our proposed candidate genes for the 6D MTA differ from the gene proposed by Singh et al.[13]. By using the *Ae. tauschii* genomes rather than relying solely on the wheat reference genome, we have demonstrated that the isoflavone reductase gene Singh et al.[13] proposed is not present in the core introgressed region. This difference and the introgressed candidate gene not found in wheat identified highlight the importance of considering non-reference genomes downstream of a GWAS, particularly when divergent material has been introduced, as the variation underlying the trait of interest might be absent from the reference genome.

These three markers can be deployed into marker-assisted breeding or introgression pipeline programmes to incorporate heat resilience traits into elite cultivars. The fact that no yield penalty was identified under more favourable conditions adds value to their deployment, especially given the negative impact that has been documented in terms of yield stability under increasing temperatures using extensive international data[46]. The donor lines for these markers will be selected using our introgression mapping approach to introduce minimal linkage drag alongside the traits of interest. Efforts to develop KASP markers for the favourable MTA alleles are currently ongoing at CIMMYT. The germplasm is available to the community through IWYP.org request.

Methods

Plant material and growth conditions. The High Biomass Association Mapping Panel HiBAP I consists of 149 spring wheat lines (Supplementary Data 1) and is composed of elite high yielding lines and lines with exotic material in their pedigree history derived from CIMMYT breeding and pre-breeding programs[21]. These exotic lines include primary synthetic derivative lines, containing between 0.5% and 43% donor material[25]; Mexican and other origin landraces derivatives; and Elite lines containing an introgressed segment of *Th. ponticum* on chr7D and/or *S. cereale* on chr1B[25]. The set of Elite lines contain 11 CIMMYT varieties released from 1966 until 2007 and additional lines derived from the systematic screening under yield potential and heat stressed field conditions of CIMMYT breeding and pre-breeding material. This allowed the identification of elite genotypes with favourable expression of traits of interest such as high biomass/RUE at different growth stages including final above ground biomass under both yield potential and heat stressed conditions. In general, pre-breeding material is derived from crosses where one of the parents was selected for expressing low canopy temperature and/or high yield or biomass under heat stressed environments.

To construct the final panel, a pre-panel consisting of more than 250 lines from different sources were evaluated in the field under favourable conditions; lines with a favourable agronomic background and without extreme height or phenology under yield potential conditions were selected to reduce the confounding effect of extreme phenology or height on the expression of biomass and other traits. HiBAP I was evaluated during 2015/16 and 2016/17 under yield potential (YP16 and YP17) and heat stressed conditions (Ht16 and Ht17). Heat stressed conditions were created with delayed sowing where emergence was registered in March instead of November or December as in a normal growing cycle (Supplementary Table 1, Supplementary Fig. 1).

The field experiments were carried out at IWYP-Hub (International Wheat Yield Partnership Phenotyping Platform) situated at CIMMYT's Experimental Station in Campo experimental Norman E. Borlaug (CENEB) in the Yaqui Valley, near Ciudad Obregon, Sonora, Mexico (27°24' N, 109°56' W, 38 masl) under fully irrigated conditions for both yield potential and heat stressed experiments. The soil type at the experimental station is a coarse sandy clay, mixed montmorillonitic typic calciorthid. It is low in organic matter and is slightly alkaline (pH 7.7)[47]. Experimental design for all environments was an alpha-lattice. Yield potential experiments consisted of four replicates in raised beds (2 beds per plot each 0.8 m wide) with four (YP16) and two (YP17) rows per bed (0.1 m and 0.24 m between rows respectively) and 4 m long. For heat stressed experiments, two replicates were evaluated for HiBAP I in 2 m × 0.8 m plots with three rows per bed (Supplementary Table 1). Seeding rates were 102 Kg ha^{-1} and 94 Kg ha^{-1} for YP and Ht experiments, respectively. Appropriate weed disease and pest control were implemented to avoid yield limitations. Plots were fertilised with 50 kg N ha^{-1} (urea) and 50 kg P ha^{-1} at soil preparation, 50 kg N ha^{-1} with the first irrigation and another 150 kg N ha^{-1} with the second irrigation. Rainfall, radiation, maximum, minimum and mean temperature by month for all the years of evaluation are presented in Supplementary Fig. 1.

Agronomic measurements. Phenology of the plots was recorded during the cycle using the Zadoks growth scale (GS)[48], following the average phenology of the plot (when 50% of the shoots reached a certain developmental stage). The phenological stages recorded were heading for heat experiments (GS55, DTH), anthesis for yield potential experiments (GS65, DTA) and physiological maturity (GS87, DTM) for both experiments. Percentage of grain filling (PGF) was calculated as the number of days between anthesis and physiological maturity divided by DTM.

Plant height was measured as the length of five individual shoots per plot from the soil surface to the tip of the spike excluding the awns. Spike, awn and peduncle length were measured in five shoots per plot before physiological maturity (PM). Fertile ($SPKLSP^{-1}$) and infertile spikelets per spike ($InfSPKLSP^{-1}$) were also counted in five spikes per plot at PM.

At physiological maturity, grain yield and yield components were determined using standard protocols[49]. Samples of 100 (YP16), 50 (YP17) or 30 (Ht16, Ht17) fertile shoots were taken from the harvested area at physiological maturity to estimate yield components. The sample was oven-dried, weighed and threshed to allow calculation of harvest index, biomass at physiological maturity, spikes per square meter, grains per square meter, number of grains per spike and grain weight per spike. Grain yield was determined on a minimum of 3.2 m^2 to a maximum of 4.8 m^2 under yield potential experiments and 1.6 m^2 under heat experiments. In yield potential experiments only, to avoid edge effects arising from border plants receiving more solar radiation, 50 cm of the plot edges were discarded before harvesting. From the harvest of each plot, a subsample of grains was weighed before and after drying (oven-dried to constant weight at 70 °C for 48 h) and the ratio of dry to fresh weight was used to determine dry grain yield and thousand grain weight. Grain number was calculated as (Yield/TGW) × 1000. Biomass at physiological maturity was calculated from yield/HI. Number of spikes per m^2 was calculated as biomass at physiological maturity /(shoot dry weight/shoot number).

Unmanned Aerial Vehicle (UAV) for canopy temperature and NDVI estimation. Aerial measurements data for canopy temperature and NDVI was collected using different aerial platforms. Each year, the logistics and availability determined which platform could be used for measuring the heat trials. A summary of the platforms used, together with the cameras and the achieved resolutions, is presented in Supplementary Table 8. The multispectral and thermal cameras were calibrated onsite by measuring over calibration panels placed on the ground before and after each mission. An exception were the aircraft missions, where a calibration performed at the airfield would not be representative of the trial conditions. The

flights were designed as a regular grid of north-south flightpaths covering the whole trial with images that overlapped 75% in all directions to ensure a good reconstruction of the orthomosaic. The flights were performed under clear sky conditions at solar noon ±2 h.

NDVI and canopy temperature orthomosaics were obtained from the aerial images using the software Pix4D. The orthomosaics were then exported to ArcGIS where a grid of polygons representing each polygon was adjusted on top of the image. To avoid the border effect, the polygons were buffered 0.5 m from the north and south border of the plot. Finally, the pixel values were extracted using the 'raster' package in R. We extracted the value of all the pixels enclosed within each polygon and removed possible outliers and calculated the average per plot.

Stress tolerance Indices. To determine the effect of heat stress in the genotypes evaluated across years and panels, Stress susceptibility index (SSI) was calculated using data from yield potential (Yyp) and heat stressed (Yht) experiments as follows (Eq. 1):

$$\mathrm{SSI} = \frac{1 - \frac{Yht}{Yyp}}{1 - \frac{\bar{Y}ht}{\bar{Y}yp}} \quad (1)$$

where $\bar{Y}$ht and $\bar{Y}$yp are the mean yields of wheat lines evaluated under heat stress and yield potential conditions, respectively[50].

Statistics and reproducibility. Data from both panels was analysed by using a mixed model for computing the least square means (LSMEANS) for each line across both years using the program Multi Environment Trial Analysis with R for Windows (METAR[51],). When its effect was significant, DTA/DTH was used as a covariate (fixed effect) except for phenology. Broad sense heritability (H^2) was estimated for each trait across both years as follows (Eq. 2):

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma^2}{re}} \quad (2)$$

where r is the number of repetitions, e is the number of environments (years), σ^2 is the error variance, σ^2g is the genotypic variance and σ^2ge = G×Y variance. Unpaired t tests for stress index (SSI) were conducted with the means across years to determine if the elite and exotic groups presented statistical differences with p-value < 0.001.

DNA extraction, capture enrichment and genotyping. All genotyping data was taken from ref. [25]. Flag leaf material from 10 plants per line was collected from field grown plots post anthesis and pooled prior to extraction with a CTAB-based protocol. DNA was extracted using a standard Qiagen DNEasy extraction preparation and quality and quantity assessed using a NanoDrop 2000 (Thermofisher Scientific) and the Quant-iTTM assay kit (Life Technologies). From this DNA, dual indexed Trueseq libraries with an average insert size of 450 bp were produced for each line and enriched using a custom MyBaits 12Mbp (100,000 120 bp RNA probes) enrichment capture synthesised by Arbour Bioscience and using 8x pre-capture multiplexing. 90,000 of these probes were designed using an island strategy to target regions across the whole genome. A subgenome-collapsed reference was used to design these probe sequences to enable homoeologous regions to be targeted with a single probe, thus expanding the design space. The final 10,000 probes were designed for selected genes, targeting both the gene body and 2Kbp upstream. Post enrichment libraries were sequenced using an S4 flowcell on an Illumina NovaSeq6000 producing 150 bp paired end reads.

Sequencing reads were trimmed and low-quality reads removed. These reads were mapped to the Chinese Spring RefSeq v1.0 wheat reference genome[52] using BWA mem v0.7.13[53]. Samtools v1.4[54] was used to remove unmapped reads, supplementary alignments, improperly paired reads, and reads that didn't map uniquely ($q < 10$). PCR duplicates were removed using Picard's MarkDuplicates[55]. SNPs were called using samtools mpileup and bcftools call[56] with parameter -m. SNPs were filtered using GATK[55] to remove SNPs that were heterozygous, had a quality score <30 or a depth <5. A locus was designated as homozygous reference if no alternative allele was found but 5 or more reads were mapped at that position. To create a set of shared SNPs for use in GWAS, SNPs for all lines were combined and loci with more than 10% missing data and a minor allele frequency (MAF) below 5% were removed. The remaining SNP loci were subjected to imputation using Beagle 5.0[57].

Genome-wide association study (GWAS). STRUCTURE v2.3.4[58] was used to genetically infer the population structure of the panel and produce a population structure matrix. An admixture model was selected and run using 30,000 burn-in iterations and 50,000 Markov Chain Monte Carlo (MCMC) model repetitions for assumed subpopulations of 2–10 for 10 randomly selected, seeded iterations for each assumed subpopulation. The delta k method from[59] was applied to all 10 replicates to identify the most likely number of definable subpopulations. This was implemented using the STRUCTURE HARVESTER Python script[60]. Finally, CLUMPP v1.1.2[61] was used with 10 independent STRUCTURE replicates to produce a consensus Q matrix for each assumed subpopulation number. GWAS analysis was conducted using the MLM model implemented in GAPITv3.0[62]. Principal component analysis eigenvectors 1–10 or membership coefficient matrices for 3-8 assumed subpopulations deduced above by STRUCTURE were used as covariates in the model to mitigate the effects of hidden familial relatedness. The EMMA method[63] was implemented in GAPIT to create a positive semidefinite kinship matrix required by the MLM model. Each MTA flanking interval was deduced by identifying the SNP position furthest upstream and downstream from the highest associated SNP that was above the -log P threshold of 5.

Identifying regions of divergence. RefSeq v1.0[52] was split into n genomic windows using bedtools makewindows[64]. Using the alignments produced in ref. [25] and detailed above, the number of reads mapping to each window was computed using hts-nim-tools[65]. To normalise by the sequencing depth of each line, read counts were divided by the number of mapped reads that passed the filters, producing normalised read counts c. Different windows of the genome have variable mapping coverage rates, so to compute coverage deviation we must compare each window to the same window in the other lines in the collection. Median normalised read counts, m, were produced, containing the median for each genomic window across the 149 lines. Mapping coverage deviation, d, was then defined for each line as follows (Eq. 3):

$$d_i = \frac{c_i}{m_i \cdot \varepsilon} \quad (3)$$

for window i ∈ {1, 2, …, n}, where ε is the median d value across the genome for the line. Statistically significant d values were calculated using the scores function from the R package 'outliers' with median absolute deviation (MAD) and probability of 0.99. This method was based on ref. [66].

Producing species-specific SNPs. Paired-end whole-genome sequencing data for the *Ae. tauschii* reference accession AL8/78[26] and 5 additional accessions that represent 5 different clades[27], 4 *Secale cereale* accessions[67], *Secale. vavilovii*[67], *Thinopyrum ponticum*[68], and *T. aestivum* cultivars Weebil[68], Norin61[68] and Pavon76[69] were mapped to RefSeq v1.0[52], filtered and SNP called as described for the genotyping above and in[25]. Homozygous SNPs were retained if they had between 10 and 60 reads supporting the alternative allele and an allele frequency >= 0.8. Heterozygous SNPs were retained if they had between 10 and 60 reads mapped and were biallelic with each allele having >= 5 reads in support and an allele frequency >= 0.3. SNPs from one relative species not shared with any of the other species or wheat cultivars were retained as species-specific SNPs. These species-specific SNPs were assigned to HiBAP SNPs if they matched in position and allele. Species-specific SNP ratios were calculated by dividing the number of SNPs in each window matched to a species-specific SNP by the mean number of SNPs matched to that species in that window across HiBAP I. SNP ratio scores below 1.45 were removed to keep enriched scores only.

Synteny between Ae. tauschii and T. aestivum. The first 10 Mb of 6D from CS and *Ae. tauschii* Aet v4.0[26] were aligned using Minimap2[70] with parameters -x asm10. Alignments <2.5 Kb in length or with mapping quality <40 were discarded. The dot plot was produced using pafr R package[71]. Proteins encoded by genes in the first 10Mbp of 6D in *Ae. tauschii* and CS were aligned using BLASTp[72]. Protein alignments and minimap2 alignments were used to anchor either side of the region commonly introgressed in all lines with the 6D T genotype to anchor the region from CS to *Ae. tauschii*. The *Ae. tauschii* genes and their proteins within this segment are considered as candidate genes. BLASTp[72] was used to compare these proteins to wheat proteins. Protein domains were identified using HMMER hmmscan[73] via ebi using Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, and TreeFam databases.

Extracting corresponding region and genes from Ae. tauschii genomes. Proteins encoded by genes in the first 10Mbp of 6D in *Ae. tauschii* and CS were aligned using BLASTp[72]. Protein alignments and minimap2 alignments were used to anchor either side of the region commonly introgressed in all lines with the 6D T allele to anchor the region from CS to *Ae. tauschii*. The sequence extracted from the *Ae. tauschii* reference genome was aligned to the other 4 chromosome-level assemblies using minimap2[70] with parameters -x asm5. Alignments below length 5000 or quality of 40 were removed. The coordinates of each orthologous region were determined manually and the genes within these coordinates extracted from the respective gff files. The *Ae. tauschii* genes and their proteins within these segments were considered as candidate genes for functional exploration. BLASTp[72] was used to compare these proteins to wheat proteins. Protein domains were identified using HMMER hmmscan[73] via ebi using Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, and TreeFam databases. Novelty of genes was determined by aligning the extracted protein sequence to each genome using tblastn[72].

Exploring functionality of candidate genes. The genes in each identified interval (except for those in the 6D interval) were submitted to Knetminer[28]. The knowledge networks created for each gene were then studied to identify links to the trait from which each MTA was deduced including their biochemical function and orthologous genes being linked in other organisms such as Rice and *Arabidopsis*

thaliana. For the *Ae. tauschii* genes introgressed into the 6D interval, we conducted extensive literature searches to identify genes with links to heat stress response based on functional studies of related genes.

Reannotating candidate gene and assessing tissue-specific expression. To test whether the missing myb-binding domain in the TraesCS6D02G014900 annotation was real or an artefact, we manually reannotated the gene. We identified the exon containing the MYB-binding domain in the wheat orthologue by aligning the coding sequence from the tauschii orthologue to Chinese Spring RefSeq v1.0[52] using tblastn[72]. We mapped Chinese Spring RNAseq data from leaf, root and shoot to RefSeq v1.0[52] using HISAT2[74] and assembled transcripts using cufflinks[75]. We visually inspected the coding sequencing and RNA-Seq alignments using IGV[76], which showed that the MYB-binding domain exon is present and expressed in wheat. To check whether the protein has a premature stop codon, we extracted the coding sequence from the assembled transcript and checked for the presence of a complete open reading frame with no stop codons using EMBOSS getorf[77]. Finally, we checked the presence of intact domains with HMMER hmmscan[73] via ebi using Pfam, TIGRFAM, Gene3D, Superfamily, PIRSF, and TreeFam databases. To explore qualitative expression of candidate genes, we mapped *Ae. tauschii* RNAseq data from leaf, root, seedling and developing grain 10dd (PRJEB23317) to Aet v4.0[26] as above and abundances were counted using StringTie[78], taking the mean transcripts per million (TPM) across the replicates. Qualitative expression of the CS orthologues was explored using Wheat Expression Browser[79] and the previously leaf, root and shoot RNAseq data mapped above.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Publicly available sequencing data used in this study is available at the European Nucleotide Archive (ENA): HiBAP I enrichment capture sequencing data - PRJEB38874; *Th. ponticum*—SRR13484812; *S. vavilovii:* ERR505040, ERR505041, ERR505042; *S. cereale* accession Lo90: ERR504990, ERR504991, ERR504992; *S. cereale* accession Lo176: ERR505005, ERR505006, ERR505007; *S. cereale* accession Lo282: ERR505015, ERR505016, ERR505017; *S. cereale* accession Lo351: ERR505035, ERR505036, ERR505037; *Ae. Tauschii* accession XJ65: SRR13961980; Y173: SRR13962062; SX60: SRR13962012; AY29: SRR13961834; KU2832: SRR13961928; Y215: SRR13962048; Weebil1: PRJEB35709; Norin61: PRJNA492239; Pavon76: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2021-10-08_wheatxmuticum/PIP-2495/200812_A00478_0126_AHN5W3DRXX/A10948_1_1/; *Ae. tauschii* RNAseq data: PRJEB23317; *T. aestivum* cv. Chinese Spring RNAseq data: Root - SRP133837; SRR6799264; SRR6799265; Leaf - SRR6799258; SRR6799259; SRR6799260; Spike - SRR6802608; SRR6802609; SRR6802610; SRR6802611.

VCF and hapmap genotype files for HiBAP I are available at: https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2022-04-08_HiBAP_genotyping/

Phenotypic data presented in this paper for the HIBAP I panel evaluated under yield potential and heat stressed environments can be found in the Dataverse CIMMYT Research Data Repository at https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/10548643[80].

The source data used to generate the main figures can be found on zenodo at https://zenodo.org/record/7333888#.Y3dmbILP1O6[81], the GitHub repository: https://github.com/benedictcoombes/Exotic_alleles_contribute_to_heat_tolerance_in_wheat_under_field_conditions and in Supplementary Data 3.

Code availability

The code needed to reproduce the main figures can be found on Zenodo at https://zenodo.org/record/7333888#.Y3dmbILP1O6[81] and at the github repository: https://github.com/benedictcoombes/Exotic_alleles_contribute_to_heat_tolerance_in_wheat_under_field_conditions.

Received: 4 March 2022; Accepted: 30 November 2022;
Published online: 09 January 2023

References

1. FAOSTAT (2021) (January 18, 2022).
2. Jägermeyr, J. et al. Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nat. Food 2021 2:11* **2**, 873–885 (2021).
3. Cossani, C. M. & Reynolds, M. P. Physiological traits for improving heat tolerance in wheat. *Plant Physiol.* **160**, 1710–1718 (2012).
4. Reynolds, M. P. et al. An integrated approach to maintaining cereal productivity under climate change. *Glob. Food Sec* **8**, 9–18 (2016).
5. Wegren, S. K. Food Security and Russia's 2010 Drought. *Eurasia. Geogr. Econ.* **52**, 140–156 (2011).
6. Asseng, S. et al. Rising temperatures reduce global wheat production. *Nat. Clim. Change* **5**, 143–147 (2014).
7. Zhao, C. et al. Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl Acad. Sci. USA* **114**, 9326–9331 (2017).
8. Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327**, 818–822 (2010).
9. Hart, G. E., Dvorak, J., Luo, M.-C., Yang, Z.-L. & Zhang, H.-B. Communicated by the structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. *Theor. Appl. Genet.* **97**, 657–670 (1998).
10. Charmet, G. Wheat domestication: Lessons for the future. *C. R. Biol.* **334**, 212–220 (2011).
11. Sehgal, D. et al. Exploring and mobilizing the Gene Bank Biodiversity for wheat improvement. *PLoS One* **10**, e0132112 (2015).
12. Valkoun, J. J. Wheat pre-breeding using wild progenitors. *Euphytica* **119**, 17–23 (2001).
13. Singh, S. et al., Harnessing genetic potential of wheat germplasm banks through impact-oriented-prebreeding for future food and nutritional security. *Sci. Rep.* **8**, 12527 (2018).
14. Reynolds, M., Dreccer, F. & Trethowan, R. Drought-adaptive traits derived from wheat wild relatives and landraces. *J. Exp. Bot.* **58**, 177–186 (2007).
15. Trethowan, R. M. & Mujeeb-Kazi, A. Novel germplasm resources for improving environmental stress tolerance of hexaploid wheat. *Crop Sci.* **48**, 1255–1265 (2008).
16. Ortiz, R. et al. Climate change: Can wheat beat the heat? *Agric Ecosyst. Environ.* **126**, 46–58 (2008).
17. Aberkane, H. et al. Evaluation of durum wheat lines derived from interspecific crosses under drought and heat stress. *Crop Sci.* **61**, 119–136 (2021).
18. Cossani, C. M. & Reynolds, M. P. Heat stress adaptation in elite lines derived from synthetic hexaploid wheat. *Crop Sci.* **55**, 2719–2735 (2015).
19. Lopes, M. S. & Reynolds, M. P. Drought adaptive traits and wide adaptation in elite lines derived from resynthesized hexaploid wheat. *Crop Sci.* **51**, 1617–1626 (2011).
20. Reynolds, M. P. et al. Strategic crossing of biomass and harvest index—source and sink—achieves genetic gains in wheat. *Euphytica* **213**, 1–23 (2017).
21. Molero, G. et al. Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring wheat and its role in yield potential. *Plant Biotechnol. J.* **17**, 1276–1288 (2019).
22. Singh, S. et al. Direct introgression of untapped diversity into elite wheat lines. *Nat. Food* **2**, 819–827 (2021).
23. Reynolds, M. et al. Raising yield potential in wheat. *J. Exp. Bot.* **60**, 1899–1918 (2009).
24. McCouch, S. et al. Mobilizing crop biodiversity. *Mol. Plant* **13**, 1341–1344 (2020).
25. Joynson, R. et al. Uncovering candidate genes involved in photosynthetic capacity using unexplored genetic variation in Spring Wheat. *Plant Biotechnol. J.* **19**, 1537–1552 (2021).
26. Luo, M. C. et al. Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* **551**, 498–502 (2017).
27. Zhou, Y. et al. Introgressing the Aegilops tauschii genome into wheat as a basis for cereal improvement. *Nat. Plants* **7**, 774–786 (2021).
28. Hassani-Pak, K. et al. KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol. J.* **19**, 1670–1678 (2021).
29. Colmer, T. D., Flowers, T. J. & Munns, R. Use of wild relatives to improve salt tolerance in wheat. *J. Exp. Bot.* **57**, 1059–1078 (2006).
30. Pinto, R. S., Molero, G. & Reynolds, M. P. Identification of heat tolerant wheat lines showing genetic variation in leaf respiration and other physiological traits. *Euphytica* **213** (2017).
31. Tattaris, M., Reynolds, M. P. & Chapman, S. C. A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. *Front Plant Sci.* **7**, 1131 (2016).
32. Rutkoski, J. et al. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes Genomes Genet.* **6**, 2799–2808 (2016).
33. Pinto, R. S. et al. Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects. *Theor. Appl Genet* **121**, 1001–1021 (2010).
34. Pinto, R. S. & Reynolds, M. P. Common genetic basis for canopy temperature depression under heat and drought stress associated with optimized root distribution in bread wheat. *Theor. Appl Genet.* **128**, 575–585 (2015).
35. Amani, I., Fischer, R. A. & Reynolds, M. P. Evaluation of canopy temperature as a screening tool for heat tolerance in spring wheat. *J. Agron. Crop Sci.* **176**, 119–129 (1996).
36. Song, Q. et al. Optimal crop canopy architecture to maximise canopy photosynthetic CO2 uptake under elevated CO2 – a theoretical study using a

mechanistic model of canopy photosynthesis. *Funct. Plant Biol.* **40**, 108–124 (2013).
37. Khush, G. S. Green revolution: the way forward. *Nat. Rev. Genet.* **2**, 815–822 (2001).
38. Dreisigacker, S. et al. Effect of flowering time-related genes on biomass, harvest index, and grain yield in CIMMYT elite spring bread wheat. *Biology (Basel)* **10**, 855 (2021).
39. Lillemo, M., Ginkel, M., Trethowan, R. M., Hernandez, E. & Crossa, J. Differential adaptation of CIMMYT bread wheat to global high temperature environments. *Crop Sci.* **45**, 2443–2453 (2005).
40. Mondal, S. et al. Earliness in wheat: A key to adaptation under terminal and continual high temperature stress in South. *Asia. Field Crops Res.* **151**, 19–26 (2013).
41. Reynolds, M. P., Balota, M., Delgado, M. I. B., Amani, I. & Fischer, R. A. Physiological and morphological traits associated with spring wheat yield under hot, irrigated conditions. *Aust. J. Plant Physiol.* **21**, 717–730 (1994).
42. Reynolds, M. et al. Breeder friendly phenotyping. *Plant Sci.* **295**, 110396 (2020).
43. Gardiner, R. & Hall, J. A. Next-generation sequencing enabled genetics in hexaploid wheat. *Applications of Genetic and Genomic Research in Cereals*, 49–63 (2019).
44. Hao, M. et al. The resurgence of introgression breeding, as exemplified in wheat improvement. *Front Plant Sci.* **11**, 252 (2020).
45. Gaurav, K. et al., Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat. Biotechnol.* **40**, 422–431 (2021).
46. Xiong, W. et al. Increased ranking change in wheat breeding under climate change. *Nat. Plants* **7**, 1207–1212 (2021).
47. Sayre, K. D., Rajaram, S. & Fischer, R. A. Yield potential progress in short bread wheats in Northwest Mexico. *Crop Sci.* **37**, 36–42 (1997).
48. Zadoks, J. C., Chang, T. T. & Konzak, C. F. A decimal code for the growth stages of cereals. *Weed Res.* **14**, 415–421 (1974).
49. Pask, A., Pietragalla, J., Mullan, D. & Reynolds, M. P. *Physiological Breeding II: A Field Guide to Wheat Phenotyping* (CIMMYT, 2012).
50. Fischer, R. A. & Maurer, R. Drought resistance in spring wheat cultivars. I. Grain yield responses. *Aust. J. Agric. Res.* **29**, 897–912 (1978).
51. Alvarado, G. et al. META-R: A software to analyze data from multi-environment plant breeding trials. *Crop J.* **8**, 745–756 (2020).
52. Appels, R. et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
54. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).
55. Depristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987 (2011).
57. Browning, B. L., Zhou, Y., Browning, S. R. & One-Penny, A. Imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
58. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
59. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
60. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
61. Jakobssonn, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
62. Lipka, A. E. et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
63. Hyun, M. K. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
64. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
65. Pedersen, B. S. & Quinlan, A. R. hts-nim: scripting high-performance genomic analyses. *Bioinformatics* **34**, 3387 (2018).
66. Keilwagen, J. et al. Detecting large chromosomal modifications using short read data from genotyping-by-sequencing. *Front. Plant Sci.* **10**, 1133 (2019).
67. Bauer, E. et al. Towards a whole-genome sequence for rye (Secale cereale L.). *Plant J.* **89**, 853–869 (2017).
68. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
69. Coombes, B. et al. Whole genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/ Ambylopyrum muticum introgression lines. *Plant Biotechnol. J.* https://doi.org/10.1111/pbi.13859 (2022).
70. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094 (2018).
71. Winter, D. pafr: Read, Manipulate and Visualize 'Pairwise mApping Format' Data. R package version 0.0.2. https://dwinter.github.io/pafr/ (2021).
72. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 1–9 (2009).
73. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200 (2018).
74. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
75. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
76. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
77. Rice, P., Longden, L. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
78. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
79. Borrill, P., Ramirez-Gonzalez, R. & Uauy, C. expVIP: a customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiol.* **170**, 2172–2186 (2016).
80. Molero, G. et al., Phenotypic data of HIBAP I panel under yield potential and heat stress conditions. https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/10548643.
81. Molero, G. et al. Exotic alleles contribute to heat tolerance in wheat under field conditions. https://doi.org/10.5281/ZENODO.7333888 (2022).

Acknowledgements

This research was supported by the International Wheat Yield Partnership (IWYP) and by the Sustainable Modernization of Traditional Agriculture (MasAgro) an initiative from the Secretariat of Agriculture and Rural Development (SADER) and CIMMYT. Foundation for Food and Agriculture Research under the Grant ID: DFs-19-0000000013. A.H. was supported by BBSRC Core Strategic Programme Grant (Genomes to Food Security) BB/CSP1720/1; A.H. and R.J. was supported by the BBSRC Designing Future Wheat grant BB/P016855/1, BBS/E/T/000PR9783 (DFW WP4 Data Access and Analysis). A.H. and R.J. were also supported by BBSRC/IWYP BB/N020871/1. B.C. was supported by the BBSRC funded Norwich Research Park Biosciences Doctoral Training Partnership grant BB/M011216/1.

Author contributions

A.H., G.M. and MPR conceived of the idea and designed the experiment. G.M., F.P., F.J.P.C. and C.R.A. collected the field data. G.M. analysed the physiological data. R.J. conducted the genome-wide association study and Knetminer searches. B.C. conducted introgression analysis and introgressed candidate gene searches. B.C., G.M. and R.J. wrote the manuscript. A.H. and M.P.R. were responsible for funding and supervision. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-04325-5.

Correspondence and requests for materials should be addressed to Anthony Hall or Matthew P. Reynolds.

Peer review information *Communications Biology* thanks Robbie Waugh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: George Inglis. Peer reviewer reports are available.

BMC Biology

RESEARCH ARTICLE

Open Access

Introgressions lead to reference bias in wheat RNA-seq analysis

Benedict Coombes[1*], Thomas Lux[2], Eduard Akhunov[3] and Anthony Hall[1*]

Abstract

Background RNA-seq is a fundamental technique in genomics, yet reference bias, where transcripts derived from non-reference alleles are quantified less accurately, can undermine the accuracy of RNA-seq quantification and thus the conclusions made downstream. Reference bias in RNA-seq analysis has yet to be explored in complex polyploid genomes despite evidence that they are often a complex mosaic of wild relative introgressions, which introduce blocks of highly divergent genes.

Results Here we use hexaploid wheat as a model complex polyploid, using both simulated and experimental data to show that RNA-seq alignment in wheat suffers from widespread reference bias which is largely driven by divergent introgressed genes. This leads to underestimation of gene expression and incorrect assessment of homoeologue expression balance. By incorporating gene models from ten wheat genome assemblies into a pantranscriptome reference, we present a novel method to reduce reference bias, which can be readily scaled to capture more variation as new genome and transcriptome data becomes available.

Conclusions This study shows that the presence of introgressions can lead to reference bias in wheat RNA-seq analysis. Caution should be exercised by researchers using non-sample reference genomes for RNA-seq alignment and novel methods, such as the one presented here, should be considered.

Keywords Wheat, RNA-seq, Reference bias, Genomics, Introgressions, Polyploidy

Background

Quantification of gene expression using RNA-seq is a fundamental technique in genomics research. It has been employed in numerous publications across a range of biological systems to identify candidate genes underlying traits of interest, uncover transcriptional pathways and networks, and investigate hypotheses relating to gene and transcriptional evolution and adaptation. In RNA-seq experiments, mRNA, which represents a snapshot of the expression of each gene at the time of sampling, is extracted from the biological sample, converted to cDNA and sequenced. The number of resulting RNA-seq reads deriving from each gene/transcript are quantified, with the number of reads proportional to the level of expression of that gene/transcript. Quantifying the expression level of each transcript and/or gene typically involves alignment of sequencing reads to the reference genome or transcriptome of the sequenced species using spliced alignment tools such as HISAT2 [1] and STAR [2] or pseudoalignment tools such as kallisto [3] and Salmon [4]. Despite these tools typically being developed and benchmarked with human data, they are widely used across numerous biological systems, often without

*Correspondence:
Benedict Coombes
benedict.coombes@earlham.ac.uk
Anthony Hall
anthony.hall@earlham.ac.uk
[1] Earlham Institute, Norwich, Norfolk NR4 7UZ, UK
[2] Plant Genome and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany
[3] Department of Plant Pathology, Kansas State University, Manhattan, KS, USA

consideration for how they will behave with specific challenges the genomes of different species present.

Making meaningful inferences from RNA-seq data relies upon the accuracy of alignment and quantification; downstream analyses and subsequent interpretation assumes that the estimated gene expression reflects actual gene expression in the biological samples. However, nucleotide variation in the coding region of genes between the sequenced sample and the reference genome/transcriptome leads to errors in read assignment during the alignment/pseudoalignment step. Some reads may be unassigned, while others may be assigned to the wrong locus. This source of error is widely known as reference bias as transcripts derived from alleles present in the reference sequence will be quantified more accurately [5].

The reduction in accuracy caused by reference bias has the potential to negatively impact downstream analyses and lead to incorrect findings. For example, Thorburn et al. [6] demonstrated how using a single reference genome to map sequencing data from genetically diverse individuals causes reference bias that negatively impacts downstream analyses in population genomic studies. While this study looked at mapping DNA reads, the same can be assumed to be true about RNA-seq data. Zhan, Griswold and Lukens [7] found that accurate estimates of transcript abundances from RNA-seq reads in maize are strongly affected by reference bias. By reanalysing RNA-seq data from a B73xMo17 recombinant inbred line population, they found that the detection of around 50% of expression quantitative trait loci (eQTLs) alleles depended on which reference genomes was used: B73 or Mo17. As the previous study [8] used B73 as the reference, Zhan et al. [7] estimated that 50% of the detected eQTLs may be false positives. Munger et al. [9] found that mapping RNA-seq reads to individualised genomes instead of a single reference genome substantially increased the accuracy of eQTL assignment in mouse from 88.2 to 98.3%, removing false positive results that appeared when using a single reference genome.

The impact of reference bias in RNA-seq analysis has not been assessed in complex polyploid genomes such as wheat despite these genomes having characteristics that may increase the extent and degree of reference bias relative to species with simpler genomes. Polyploidisation increases the number of alleles per gene, typically resulting in a pair of alleles, known as homoeologues, in each subgenome; however, subsequent gene duplications or deletions can change the relative copy number of homoeologues between the subgenomes. As RNA-seq reads are derived from all subgenomes at once, read assignment must be able to distinguish reads deriving from homoeologues. Accurate discrimination of wheat homoeologue RNA-seq reads has been demonstrated with both pseudoalignment [10, 11] (99.9% accuracy) and alignment-based (98% accuracy) [11] methods when mapping reads back to the genome from which they derived. However, when mapping reads from a different genotype, unequal divergence between homoeologues relative to the reference genome may compromise the accuracy of the expression balance estimation between homoeologues. Being able to accurately estimate homoeologue expression balance is important for wheat research as variation in the relative mRNA expression of homoeologues within a triad may confer phenotypic plasticity [10] and variation in agronomic traits, the understanding of which has important applications for crop improvement.

Introgression events, the introduction of genetic material from one species to another [12], are common among plants; in fact, its frequency is thought to be higher in plants than in animals, due to higher rates of interspecific hybridisation success [13]. Additionally, novel genetic variation is commonly introgressed into plants by breeders and researchers for crop improvement [14]. Several studies have demonstrated how common introgressions are in wheat accessions with some accessions being comprised of up to 34% introgressed material [15–19]. The production of chromosome-level genome assemblies of modern elite wheat cultivars confirmed this, revealing introgressions from wild and domesticated relatives, including species outside of the *Triticum* and *Aegilops* genera, present in one or multiple cultivars [20, 21]. These introgressions introduce greater sequence divergence between varieties than observed between varieties at non-introgressed regions; this increased divergence likely leads to an increased proportion of reads that are unable to be assigned correctly.

Using simulated and experimentally generated RNA-seq data, we identify non-trivial levels of reference bias in RNA-seq mapping in wheat which can largely be attributed to introgressions. This leads to incorrect estimates of relative expression between homoeologues and incorrectly called differences in expression between cultivars. By constructing a pantranscriptome reference composed of Chinese Spring transcripts and transcripts from the assemblies generated as part of the 10+ wheat genomes project [20], we demonstrate how reference bias caused by divergent alleles can be reduced.

Results

Reference bias in wheat is driven by divergent genes introduced via *introgressions and results in underestimation of gene expression*

To explore the impact of reference bias on the quantification of gene expression in wheat, we simulated 1000 read pairs from each high-confidence (HC) gene

in Chinese Spring RefSeq v1.1 and the nine chromosome-level genome assemblies generated as part of the 10+ wheat genomes project [20, 22] if the longest transcript of the gene is at least 500 bp. These reads were pseudoaligned or aligned to the Chinese Spring reference transcriptome or genome using kallisto or STAR, respectively. These algorithms represent pseudoalignment and alignment-based methods and are among the most commonly used tools for RNA-seq quantification in the wheat community.

Mapping Chinese Spring reads to Chinese Spring, hereafter referred to as self-mapping, yields very accurate estimates of gene expression, with kallisto slightly outperforming STAR (Fig. 1a, b, Additional file 1: Table S1). Using kallisto, 88,401/88,443 (99.95%) of genes were correctly quantified (between 500 and 1500 read pairs). Thirty-two genes were underestimated (< 500 read pairs) and 10 genes were overestimated (> 1500 read pairs). Using STAR, 87,689/88,443 (99.15%) were correctly quantified with 504 and 250 genes underestimated and overestimated, respectively.

Mapping reads generated from the other cultivars to Chinese Spring, hereafter called cross-mapping, yielded much less accurate estimation of gene expression with a skew towards underestimation (Fig. 1a, b, Additional file 1: Table S1). The percentage of genes correctly quantified ranged from 55,773/63,001 (88.53%) for Lancer, with 5700 (9.05%) and 1528 (2.43%) under- and overestimated, respectively, to 58,468/64,077 (91.2%) for Norin61, with 2527 (3.94%) and 3082 (4.81%) genes under and overestimated, respectively. For cross-mapping, unlike self-mapping, STAR appears to perform better than kallisto; the proportion of correctly quantified genes ranged from 58,390/63,001 (92.68%) for Lancer, with 3916 and 695 under and overestimated, respectively, to 59,648/64,077 (93.1%) for Norin61, with 2450 (3.82%) and 1979 (3.09%) genes under and overestimated, respectively.

To explore the effect of reference bias on the quantification of homoeologue expression balance, we calculated the proportion of triads belonging to each category that defines a different state of relative homoeologue expression. As reads were simulated evenly across genes, all

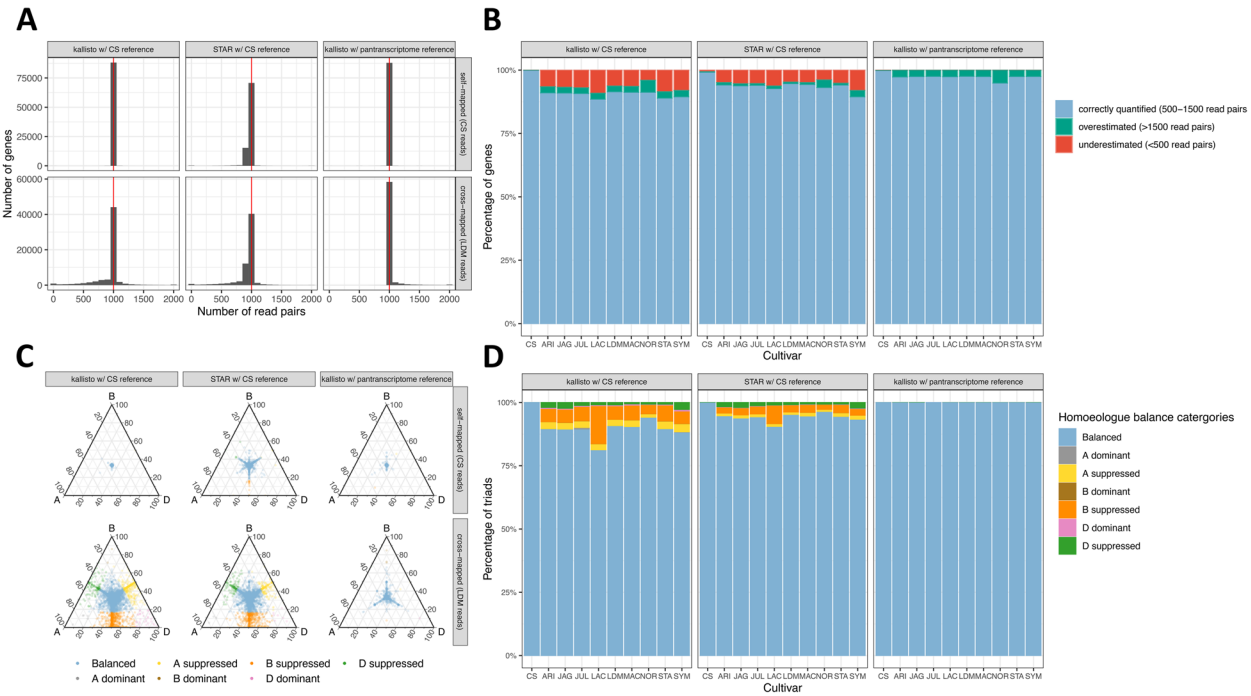


Fig. 1 Assessing the extent of reference bias in wheat. **A** Distribution of read counts when self-mapping Chinese Spring simulated reads or cross-mapping Landmark simulated reads. Comparing STAR and kallisto using the Chinese Spring RefSeq v1.0 reference and RefSeq v1.1 transcriptome and kallisto using the pantranscriptome reference. **B** Percentage of genes with expression estimated correctly, expression underestimated (< 500 read pairs) and expression overestimated (> 1500 read pairs) for simulated reads from 10 cultivars aligned to Chinese Spring with kallisto and STAR or to the pantranscriptome reference with kallisto. **C** Balance of homoeologue expression across triads when self-mapping Chinese Spring or cross-mapping Landmark simulated reads, comparing STAR and kallisto using the Chinese Spring RefSeq v1.0 reference and RefSeq v1.1 transcriptome and kallisto using the pantranscriptome reference. Each point on the ternary plot represents one triad. Points towards a corner indicate dominant expression of that homoeologue, and points opposite a corner indicate suppression of that homoeologue. **D** Percentage of triads in each expression category, using simulated reads from 10 cultivars aligned to Chinese Spring with kallisto and STAR or to the pantranscriptome reference with kallisto

triads should be classified as balanced; therefore, triads classified as imbalanced (one or two homoeologues with expression greater than the other(s)) are considered incorrectly classified. The percentage of correctly classified triads varies between 80.97% (Lancer) and 93.84% (Norin61) using kallisto and between 90.23% (Lancer) and 96.12% (Norin61) using STAR (Fig. 1c, d, Additional file 1: Table S2). Across the cultivars, triads incorrectly classified as suppressed, where one homoeologue is estimated to be expressed less than the others, were far more common than triads incorrectly classified as dominant, where one homoeologue is estimated to be expressed more highly than the others (Fig. 1d, Additional file 1: Table S2). This reflects how the reference bias leads to more underestimated than overestimated genes.

The B subgenome has the most, and the D subgenome the fewest, number of triads incorrectly classified as suppressed. This is in line with observations of greater diversity in the A and B subgenomes, with the B subgenome having the highest [16]. This difference is largely caused by gene flow from wild tetraploid *T. dicoccoides* to *T. aestivum* during the history of its cultivation, without comparable gene flow to the D subgenome [17, 19, 23]. This finding suggests the historic gene flow from tetraploid wheat likely contributes to reference bias in RNA-seq analyses.

To explore the extent of errors when comparing two cultivars mapped to a common reference, we compared the estimated expression of Lancer and Jagger genes, whose simulated reads were both aligned to Chinese Spring using STAR (Fig. 2a, b). Genes with read counts > 1.5 × or < 1/1.5 × compared to the other cultivar were classified as incorrectly quantified. Using STAR, 4791/60,338 (7.94%) genes were incorrectly quantified between the two cultivars; of these genes, 2747 and 2044 genes had a lower read count in Lancer and Jagger, respectively.

We observed a clear overlap between clusters of incorrectly quantified genes and regions of divergence between the cultivars (Fig. 2a, c), identified by blocks of reduced CDS nucleotide identity between pairs of orthologues between Lancer and Jagger. Such gene-level divergence is indicative of introgressed material; indeed, several of these blocks correspond to previously characterised introgressions. These introgressions include (coordinates based on Chinese Spring RefSeq v1.0) the following: *Aegilops ventricosa* introgression in Jagger (chr2A:1–24,643,290) [20, 21, 24]; *Triticum timopheevii* introgression in Lancer (chr2B:89,506,326–756157100) [20, 21]; *Aegilops comosa* introgression in Jagger (chr2D:570,141,481–613325841) [21]; and a *Thinopyrum ponticum* introgression in Lancer (chr3D:591,971,000–615552423) [20, 21]. 1881/3054 (61.59%) of introgressed genes (those belonging to one of the four previously characterised introgressions listed above) were incorrectly quantified between the two cultivars, compared to 2910/57,284 (5.08%) non-introgressed genes incorrectly quantified (Fig. 2d; chi-squared p-value < 2.2e − 16). Genes with an introgressed copy in Lancer tend to be underestimated in Lancer and genes with an introgressed copy in Jagger tend to be underestimated in Jagger.

In further support of CDS divergence being a predominant contributing factor to incorrect quantification, we found that incorrectly quantified genes have a mean CDS identity between orthologue pairs of 97.3% compared to a mean of 99.9% for genes correctly quantified (Fig. 2e; p-value < 2.2e − 16; 95% confidence interval ranges from 2.45 to 2.63). The percentage of genes incorrectly quantified ranges from 83.2% for genes with < 96% CDS identity between orthologues to just 2.9% for genes with ≥ 99% identity between orthologues (Fig. 2f).

Reducing reference bias by constructing a pantranscriptome reference

The 10+ wheat genomes project generated chromosome-level de novo assembled genomes for nine wheat cultivars in addition to the reference cultivar Chinese Spring [20]. These include numerous introgressions that are the predominant source of reference bias we observe. High-quality gene annotations for these genome assemblies have been produced [22]. We constructed a pantranscriptome reference by taking the transcripts from the 107,891 Chinese Spring HC genes and adding transcripts from the nine cultivars with a chromosome-level genome assembly generated as part of the 10+ wheat genomes project [20] if that transcript's gene exists in a 1-to-1 relationship with a gene from Chinese Spring, based on OrthoFinder [25] orthogroup assignments. This resulted in a set of transcripts from 763,877 genes from 10 cultivars, 107,891 from Chinese Spring and a mean of 72,887 from each of the nine other cultivars (Fig. 3). A total of 80,211 Chinese Spring genes had at least one 1-to-1 orthologue in another cultivar, while 59,639 Chinese Spring genes had a 1-to-1 orthologue in all nine other cultivars (Additional file 2: Fig. S1). The pantranscriptome reference was used as the transcriptome reference for kallisto pseudoalignment. After pseudoalignment, read counts and TPMs were summed across all transcripts corresponding to a given Chinese Spring gene. Kallisto splits read counts evenly across transcripts with an identical match so redundancy of transcripts does not cause problematic multi-mapping; all transcripts corresponding to a gene can thus be added.

To ensure using this pantranscriptome reference does not introduce any additional mapping errors from adding redundant transcripts, we compared quantified

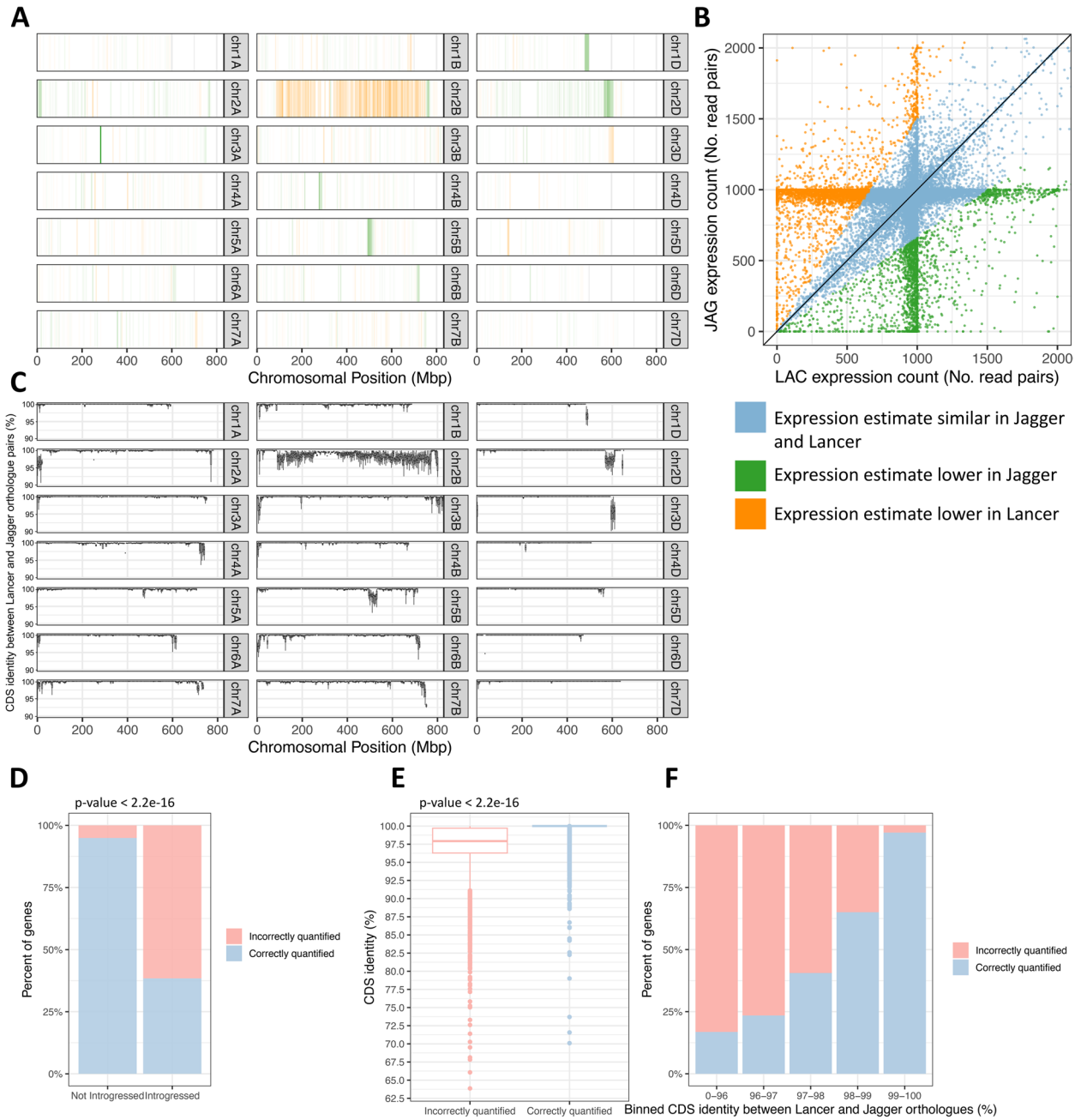


Fig. 2 The impact of reference bias on expression differences between cultivars and enrichment of incorrectly quantified genes within introgressions. **A** The distribution of incorrectly quantified genes in 5-Mbp windows, coloured by the cultivar in which the estimated expression is lower; orange blocks are underestimated in Lancer compared to Jagger, while green blocks are underestimated in Jagger compared to Lancer. The reads are aligned using STAR as this outperformed kallisto for cross-mapping. **B** Expression counts for Lancer-Jagger orthologue pairs. Genes are considered incorrectly quantified if their estimated read count is 1.5 × or 1/1.5 × the other cultivar. **C** CDS nucleotide identity between Lancer and Jagger 1-to-1 orthologue pairs, binned into 5-Mbp genomic windows based on Chinese Spring RefSeq v1.0. **D** Percentage of genes incorrectly quantified and correctly quantified in characterised introgressed regions and regions not characterised as introgressed. **E** CDS nucleotide identity between Lancer and Jagger 1-to-1 orthologue pairs for those that are incorrectly quantified and those that are correctly quantified. **F** Percentage of genes incorrectly quantified and correctly quantified, split into bins of different levels of CDS nucleotide identity

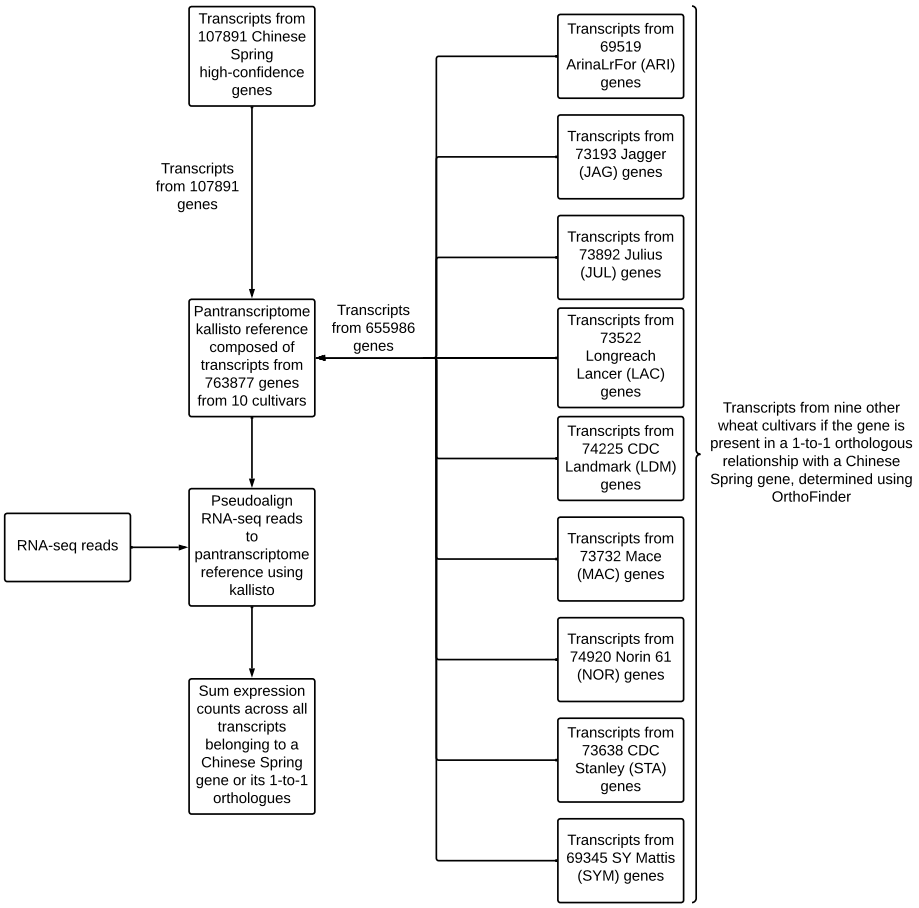


Fig. 3 Creation of the pantranscriptome reference and how RNA-seq reads are aligned to it

expression counts between four difference references: Chinese Spring, the pantranscriptome reference, Chinese Spring plus the Landmark transcripts from genes in a 1-to-1 relationship with a Chinese Spring gene, and the pantranscriptome reference without the Landmark transcripts. The simulated reads from Landmark were used for pseudoalignment. Of these four references, the pantranscriptome reference performed the best, with 97.53% of genes correctly quantified. Chinese Spring plus Landmark transcripts were very similar, with 97.50% of genes correctly quantified. This demonstrates that adding redundant transcripts and summing the read counts does not introduce errors in the kallisto mapping. Using the pantranscriptome reference without Landmark transcripts resulted in a slightly lower level of correct quantification, with 96.84% correctly quantified. The difference is likely due to uniquely introgressed genes in Landmark that are not present in the other cultivars. Nevertheless, due to many introgressed genes being common between cultivars, it still performed much better than just using Chinese Spring, which had 91.43% genes correctly quantified.

Using the pantranscriptome reference instead of Chinese Spring to quantify expression from the simulated RNA-seq reads resulted in much more accurate quantification for genes that were previously underestimated when cross-mapping, removing nearly all gene counts below 1000 (Fig. 1a, b). There was little change in the number of genes overquantified when cross-mapping and little difference in the distribution of read counts when self-mapping (Fig. 1a, b). The distribution of read counts shows that for Lancer, the most error-prone cultivar, the number of genes correctly quantified increased from 58,390/63,001 (92.68%) using STAR to 61,352/63,001 (97.38%) using the pantranscriptome reference. Using the pantranscriptome reference, only 2 genes remained quantified below 500 read pairs compared to 3916 genes when using the Chinese Spring reference. The number of triads correctly assigned to the balanced expression category also greatly increased when using the pantranscriptome reference (Fig. 1d). All cross-mapped cultivars had at least 99.89% triads correctly assigned as balanced; this compares to between 80.97 and 93.84% using kallisto, and between

90.23 and 96.12% using STAR to align to Chinese Spring.

Comparing Jagger and Lancer as before, this approach reduced the number of genes incorrectly quantified in one cultivar from 4971/60,338 (7.94%) to 617 (1.02%) (Additional file 2: Fig. S2). Only 23 genes (0.0381%) remain incorrectly quantified due to underestimation in one cultivar. Almost all the remaining error in both cross-mapped read counts and incorrectly quantified genes between cultivars is due to overestimation of gene expression, likely caused by copy number variation or presence/absence variation between cultivars, as opposed to divergence between orthologous gene models.

Exploring reference bias caused by introgressions in experimentally generated RNA-seq data

Simulated RNA-seq data is unlikely to capture the complete picture of a real experiment [26]. While our simulations highlight theoretical errors, it is important to assess how reference bias impacts published findings and how using the pantranscriptome reference corrects errors in real data. We reanalysed the sequencing data generated by He et al. [11]. He et al. [11] analysed RNA-seq data from 198 diverse wheat accessions, alongside enrichment capture paired-end DNA reads, to uncover eQTLs linked with homoeologue expression bias and variation in important productivity traits. Crucially for our work, they identified a set of genes whose expression exhibited negative correlation with its homoeologue across the panel. A subset of accessions possessed lowly expressed alleles in one of the homoeologues and the presence of the lowly expressed alleles was linked to various important productivity traits. This set contains 59 genes to which we have added *ELF3-D1*. While *ELF3-D1* did not fall into the set of very negatively correlated 59 genes, it was used as case example due to its agronomic significance. Also, it still did show a negative correlation with its B homoeologue, with this expression bias associated with agronomic traits. This set of 60 genes is hereafter referred to as genes showing lack of expression correlation.

Firstly, to identify potential introgressed regions within these accessions, we mapped the enrichment capture paired-end DNA reads to Chinese Spring RefSeq v1.0 and for each 1-Mbp genomic window, calculated the mapping coverage deviation between each line and the median for that window across the accessions (Fig. 4a). Blocks of windows with coverage deviation values significantly below 1 have few reads that have mapped in this region relative to the other accessions. This is indicative of an introgression (which introduces divergent DNA that maps less well) or a deletion. We observed more divergent material in the A and B subgenomes, which is expected based on the higher levels of gene flow to the A and B subgenomes (Fig. 4a) [17, 19, 23]. The genes showing lack of expression correlation identified by He et al. [11] are enriched in genomic windows identified as introgressed or deleted (Fig. 4b), with 78.2% of these genes in a genomic window identified as introgressed or deleted in 30 or more accessions. In the rest of the genome, only 12.3% of genes are found in a genomic window identified as introgressed or deleted in 30 or more accessions.

To explore the impact of the pantranscriptome reference on estimated expression, we pseudoaligned the leaf RNA-seq data from the 198 wheat accessions to both Chinese Spring and to the pantranscriptome reference. Kallisto was used for aligning to Chinese Spring instead of STAR for consistency with the analysis by He et al. [11]. 43/60 (71.7%) of genes showing lack of expression correlation (Fig. 5a) have, in 25 or more accessions, an estimated expression less than half when mapping to

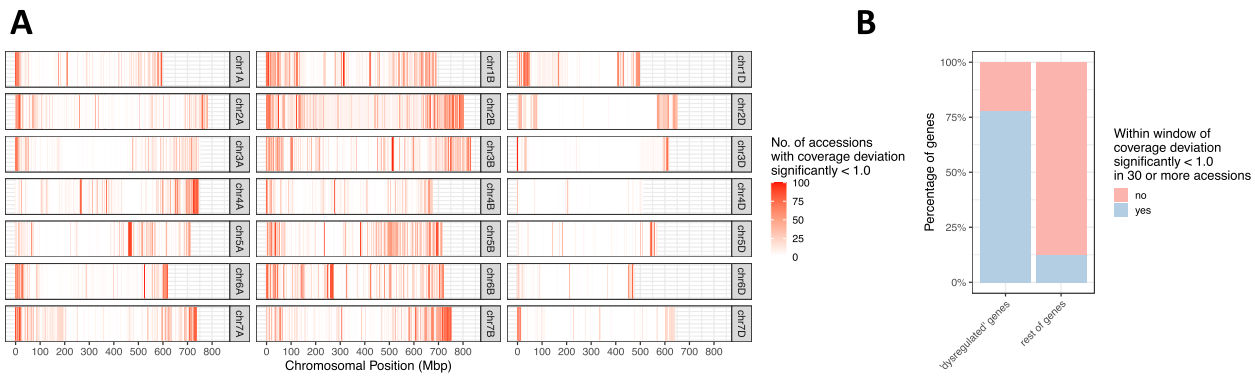


Fig. 4 Enrichment of genes showing a lack of expression correlation in He et al. [11] within regions of divergence. **A** Chromosomal distribution of the number of accessions in each 1-Mbp genomic window which had mapping coverage deviation significantly less than 1 and are thus likely to contain divergent introgressed material or be deleted. **B** The number of genes from the set of 60 genes showing lack of expression correlation identified by He et al. [11] that are present in genomic windows identified as introgressed or deleted in 30 or more accessions

Chinese Spring compared to when mapping to the pantranscriptome reference. These are likely introgressed genes whose expression is underestimated when using Chinese Spring as the reference. 6/60 (10.0%) of the genes have, in 25 or more accessions, an estimated expression more than double when mapping to Chinese Spring compared to when mapping to the pantranscriptome reference (Fig. 5a). This may arise if, when using the Chinese Spring reference, RNA-seq reads were incorrectly assigned to a gene because the correct gene is too divergent and then, when using the pantranscriptome reference, those incorrectly assigned reads now have another more appropriate gene to be assigned to, resulting in fewer reads assigned to the first gene.

While this shows that using Chinese Spring as the reference leads to underestimation of many of these genes, it is important to look at the impact of this on the calculated correlation between homoeologues that led to them being classified as genes of interest by He et al. [11]. We found that the SCC score between homoeologues from this set was −0.0990 when using the Chinese Spring reference and 0.407 using the pantranscriptome reference (Fig. 5b; p-value < 2.2e−16; 95% confidence interval ranges from −0.603 to −0.410). Even though this SCC value remains lower than the mean SCC (~0.8) reported for the entire set of homoeologues [11], it indicates that the usage of pantranscriptome as reference increases expression correlation estimates between homoeologues compared to single reference estimates.

Several regions with poor mapping coverage (mapping coverage deviation significantly below 1) in multiple accessions overlap precisely with previously identified introgressions from cultivars assembled in the 10+ wheat genomes project [20]. One such introgression is at the end of chr1D (484,302,410–495,453,186 bp, based on RefSeq v1.0 coordinates), present unbroken in 53/198 (26.8%) accessions (Additional file 1: Table S3) and shared with cultivars Jagger and Cadenza (Fig. 6a). The precise overlap of the blocks of the reduced mapping coverage in the accessions and in Jagger and Cadenza suggests that this introgression has the same origin in all these lines, and that no recombination has taken place within the introgression since its introduction. This lack of variation in its size makes it a good candidate for the following analysis. Additionally, this region was highlighted by He et al. [11] as it contains 6 of the genes showing lack of expression correlation, including *ELF3-D1*, which was used as a case example due to its role in heading date [27]. He et al. [11] suggest this is a terminal deletion; however, Wittern et al. [28] identified that the terminal region, including *ELF3-D1*, is an introgression in Cadenza and Jagger, deriving from either *Triticum timopheevii* or *Aegilops speltoides*, based on the *ELF3-D1* gene model possessing an intronic deletion shared with both of these species. We can exclude *Ae. speltoides* as the donor species as protein alignments between the Jagger introgression and *Ae. speltoides* proteins showed a median protein identity of just 91.6%. As *T. timopheevii* does not have a genome assembly available, we cannot confirm it is the donor; however, the mapping profile of *T. timopheevii* reads to the Jagger genome assembly suggest it is a likely match (Additional file 2: Fig. S3). As we cannot be certain about the donor species, we will hereafter refer to this introgression as the chr1D introgression.

We compared the mean expression of genes from the chr1D introgression across accessions that possess the introgression to their 1-to-1 wheat orthologue across the accessions lacking the introgression. When using the Chinese Spring reference, the introgressed genes appear to be less expressed than their wheat orthologues (p-value = 0.0224, 95% confidence interval ranges from −8.65 to −0.679); however, when using the pantranscriptome reference, no significant difference in expression was found between the genes (Fig. 6b, Additional file 1: Table S4; p-value = 0.980, 95% confidence interval ranges from −4.94 to 4.82).

(See figure on next page.)

Fig. 5 The impact of reference bias on the quantification of gene expression in the accessions sequenced by He et al. [11]. **A** Estimated expression of the 60 genes identified as showing a lack of expression correlation by He et al. [11], using either the Chinese Spring RefSeq v1.1 transcriptome or the pantranscriptome reference as targets for kallisto pseudoalignment. The dashed black line represents $x = y$, which is the expected value if the reference is not affecting the estimation of gene expression. An accession lying on this dashed line has this gene's expression estimated the same when using each reference. Red dots and green dots represent accessions in which a given gene has a TPM value < 50 or > 150%, respectively, when mapping to Chinese Spring than when mapping to the pantranscriptome reference. A red star indicates that in 25 or more accessions, the gene has an estimated expression less than half when mapping to Chinese Spring compared to when mapping to the pantranscriptome reference. A green star indicates that in 25 or more accessions, the gene has an estimated expression more than double when mapping to Chinese Spring compared to when mapping to the pantranscriptome reference. **B** Spearman's correlation coefficient (SCC) between homoeologue pairs where one homoeologue is in the set of genes showing a lack of expression correlation identified by He et al. [11]. SCC scores were computed between AB, AD and BD homoeologue pairs and the lowest score was used. Triads in which any of the homoeologues were not present in the RefSeq v1.0 HC gene annotation were excluded. The significance of the difference between SCC scores when using the Chinese Spring reference compared to when using the pantranscriptome reference was calculated using a two-tailed t-test with no assumption of equal variance

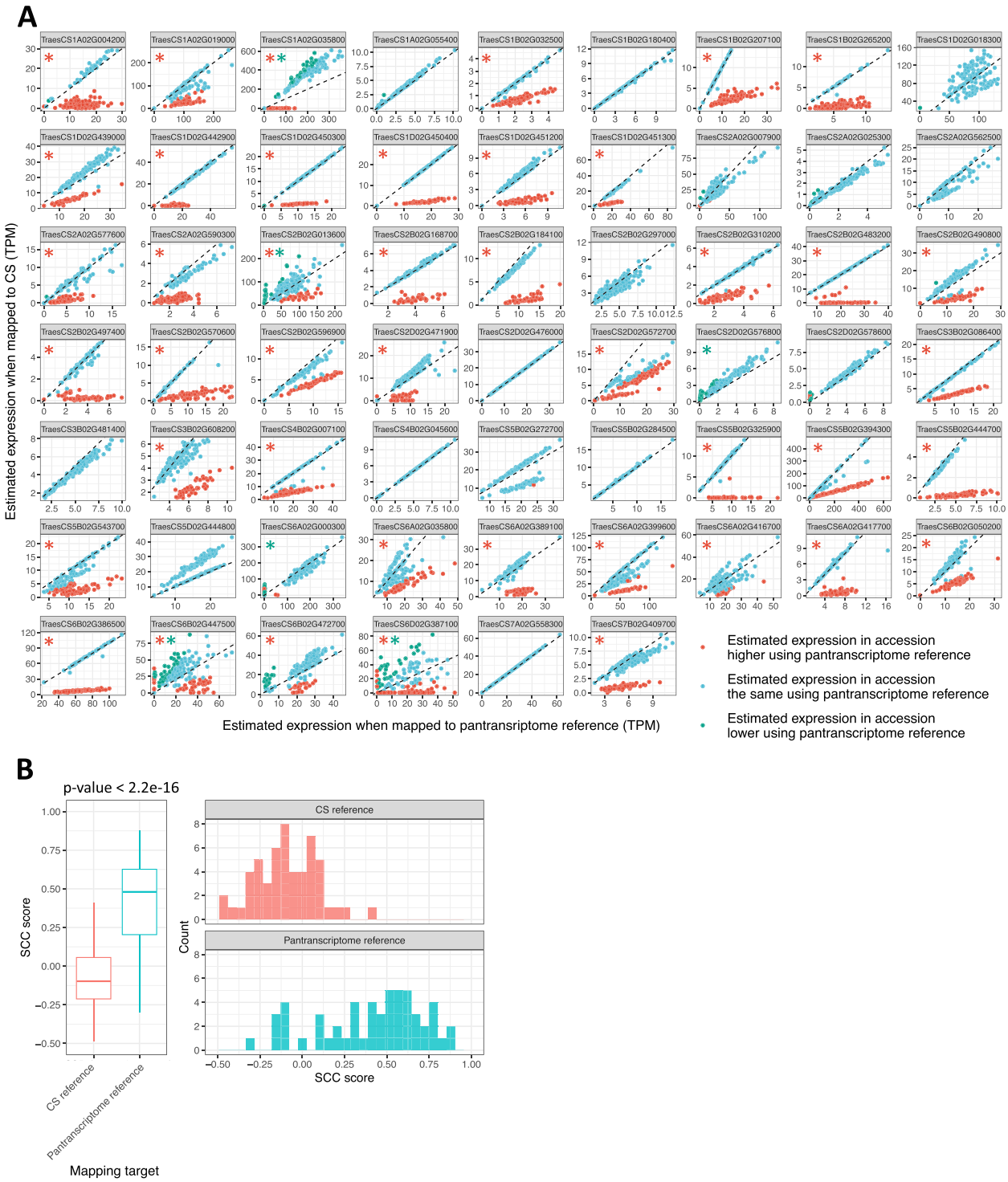


Fig. 5 (See legend on previous page.)

Earlier, using simulated data, we demonstrated that reference bias can lead to incorrect assignment of expression balance across triads. To examine this phenomenon in real data, we examined the estimated expression across triads within the chr1D introgression that are also in the set of genes showing lack of expression correlation identified by He et al. [11]. When the RNA-seq reads are pseudoaligned to Chinese Spring,

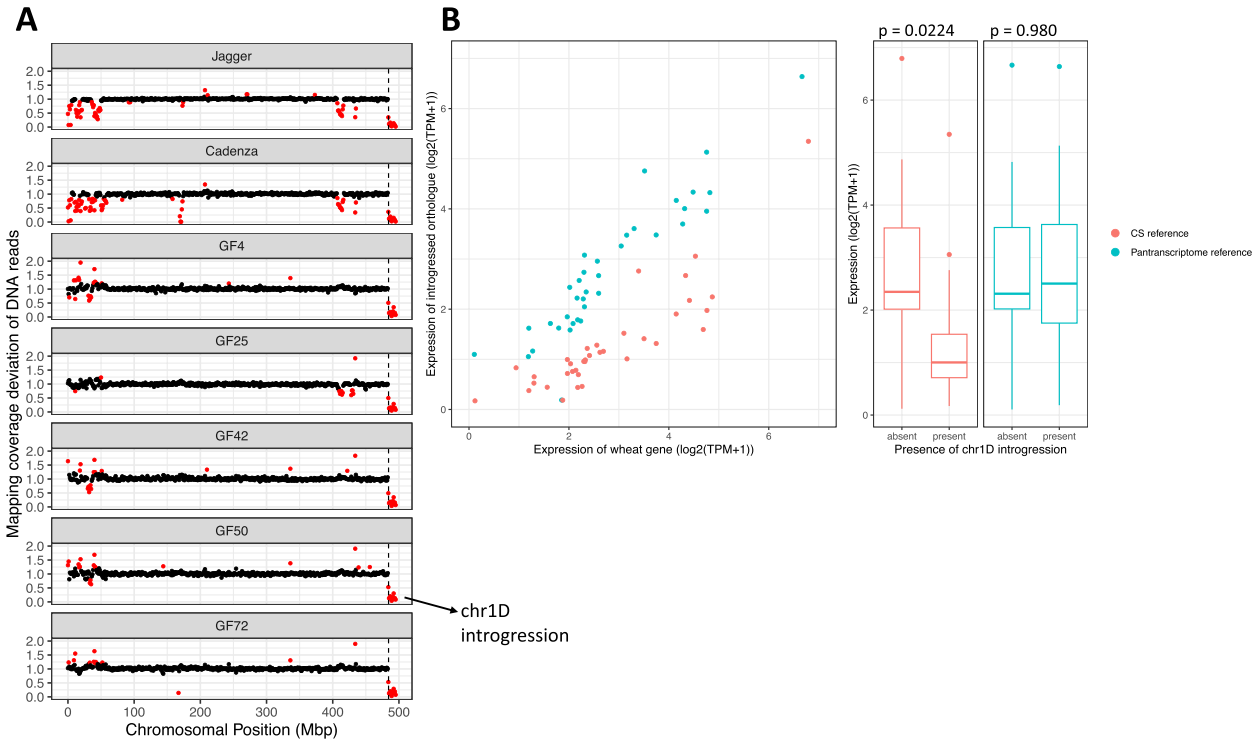


Fig. 6 Introgressed genes falsely identified as being less expressed due to reference bias. **A** Mapping coverage deviation of DNA reads across chr1D of Jagger, Cadenza, and 5 of the accessions analysed by He et al. [11]. Each point is the coverage deviation value for a given 1-Mbp genomic window. Windows with a normalised coverage score significantly different to the median normalised coverage score for that window across the set of lines being compared are coloured red. Coverage deviation values significantly below one indicates divergent material is present or a deletion has taken place, relative to the median of the rest of the set of lines. Coverage deviation values and significance values were calculated separately for the accessions and for the cultivars Jagger and Cadenza, the latter two being compared to mapping coverage values from the other cultivars whose genomes were assembled as part of the 10+ wheat genomes project [20]. The reduced coverage at the end of chr1D, the left-hand border of which is indicated by the vertical dashed black line, is the chr1D introgression, common to 53 of the 198 accessions and Jagger and Cadenza which were assembled as part of the 10+ wheat genomes project. **B** Expression of the wheat gene compared to its introgressed orthologue from the chr1D introgression, using either Chinese Spring or the pantranscriptome reference as targets for kallisto pseudoalignment. Orthologue pairs with TPM < 1 in both the introgressed and the wheat copy when mapping to the pantranscriptome reference were excluded. The significance of the difference between introgressed and non-introgressed orthologues when using the Chinese Spring or the pantranscriptome reference was calculated using two-tailed *t* tests with no assumption of equal variance

in lines with the chr1D introgression, *ELF3-D1* appears to be lowly expressed and the expression of *ELF3-B1* appears slightly elevated compared to accessions without the chr1D introgression. However, when mapped to the pantranscriptome reference, the expression of *ELF3-D1* and *ELF3-B1* in accessions with the chr1D introgression appears very similar to that in accessions without the chr1D introgression (Fig. 7a, b). The CDS sequence for *ELF3-D1* from the introgression shares 97.0% sequence identity with *ELF3-D1* in Chinese Spring, 97.6% identity with *ELF3-A1* and 97.8% identity with *ELF3-B1*. The high divergence of *ELF3-D1* from the introgression and *ELF3-D1* from Chinese Spring and the greater similarity between *ELF3-D1* from the introgression with *ELF3-B1* from Chinese Spring explains how most reads were unable to be assigned, yet some were incorrectly assigned to the *ELF3-B1*, hence the slight increase in estimated expression of *ELF3-B1* when using the Chinese Spring reference. The five other genes showing lack of expression correlation within the chr1D introgression also showed reduced homoeologue imbalance using the pantranscriptome reference and expression level in line with accessions without the chr1D introgression, in which the triad does not contain an introgressed D homoeologue. Four of these genes also showed a slight decrease in estimated expression in the B homoeologue when mapping to the pantranscriptome reference, supporting the idea that false mapping from the introgressed gene to its homoeologue will be driving false negative correlation scores in addition to artificially low expression of the introgressed homoeologue.

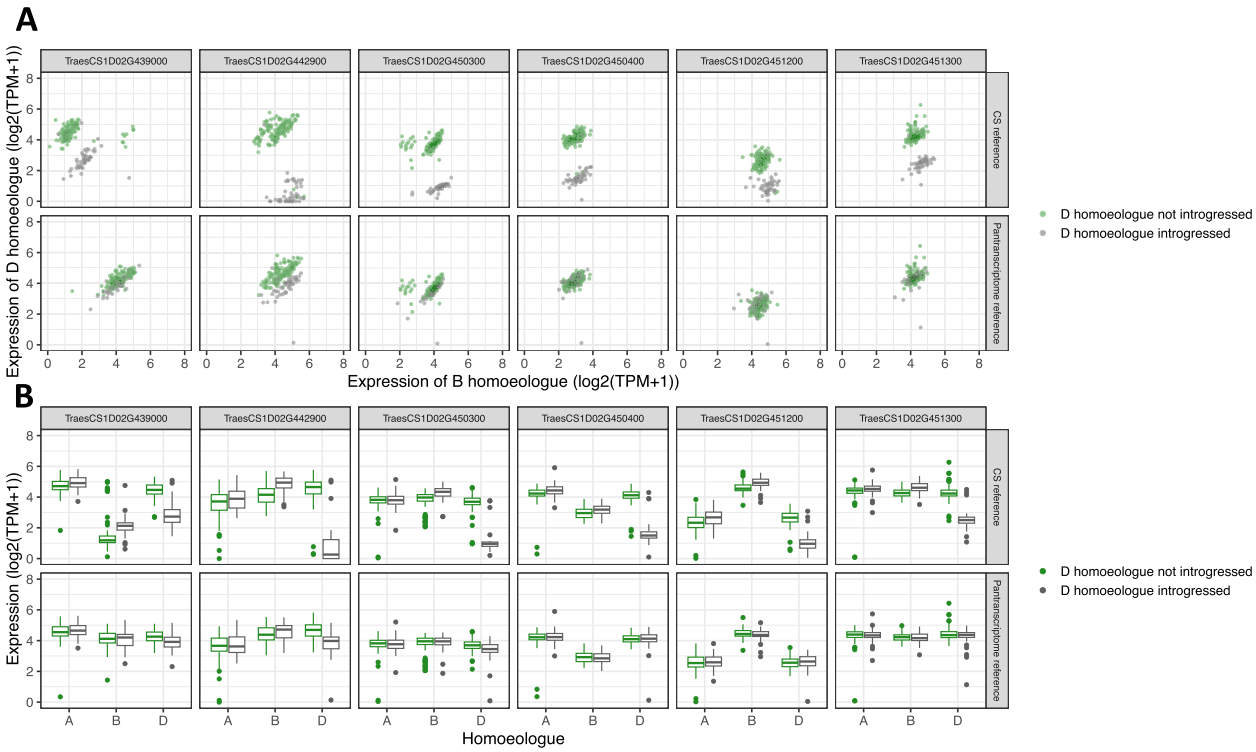


Fig. 7 The impact of reference bias on the quantification of triads in which one homoeologue has been introgressed. **A** Estimated expression level of introgressed D homoeologues compared to the wheat B homoeologues and wheat D homoeologues compared to wheat B homoeologues, using either Chinese Spring or pantranscriptome reference as targets for kallisto pseudoalignment. Each point represents one accession. **B** Expression level of triads from where the D homoeologue is an introgressed gene in a subset of lines, using either Chinese Spring or the pantranscriptome reference as targets for kallisto pseudoalignment. The centre line of the boxplots = the median; the box limits = the upper and lower quartiles, the whiskers = 1.5 × interquartile range; and the points = outliers

Discussion

In the emerging era of plant pangenomics, chromosome-level assemblies are being generated for an increasing number of cultivars/accessions, which will facilitate a shift away from reference genome-centric methods. Here we have demonstrated the importance of utilising these resources effectively for RNA-seq analyses in wheat to reduce reference bias.

RNA-seq reference bias in wheat

Quantification of gene expression from RNA-seq reads in wheat is very accurate when the matching reference genome for the sample is available. However, cross-mapping RNA-seq reads leads to detectable levels of reference bias, seen both at the individual gene level and also when assigning triads to categories of homoeologue expression balance. A major cause of this bias appears to be introgressions of diverged gene orthologues from wheat's wild and domesticated relatives. In some cases, references bias within introgressions could be severe enough to have a strong impact on downstream analyses and conclusion drawn based on these analyses. This analysis was conducted on wheat but other species with substantial introgressed content and/or polyploid genomes may suffer from the same problem. Similar analyses on other species may thus provide value for their respective communities.

Kallisto performed better for self-mapping but when cross-mapping, STAR was better able to deal with divergence between genes, although was far from resolving the issue of reference bias. Similar limitations of alignment-free methods have been previously discussed; for example, Wu et al. [29] demonstrated that kallisto performs poorly for lowly expressed genes and for RNA reads with biological variation compared to the reference.

A future exploration of the impact of reference bias on differential expression calls in wheat will be useful. Reference bias may have little impact on differential expression between conditions or across tissues within a single genotype, as, even if incorrectly quantified, the ratio of estimated expression between conditions/tissues should remain very similar regardless of reference. However, this needs to be assessed formally. If interested in homoeologue expression balance, however, unequal divergence of

homoeologues relative to the reference will lead to incorrect findings. Reference bias also makes complex patterns more difficult to discern. For example, in a previous study [30], we demonstrated how the rhythmicity of *ELF3-1D* and *SIG3-1D* in a Cadenza timecourse RNA-seq dataset was difficult to ascertain as the reads mapped so poorly to Chinese Spring. However, when using adding in the introgression to the reference, the reads mapped more correctly, and the rhythmicity could be accurately assessed.

Matching a sample to a more appropriate reference genome will become increasingly possible as genome assemblies for more wheat accessions become available. However, analyses involving two or more accessions require a common reference genome to which the RNA-seq reads can be aligned. In this situation, or when the appropriate genome assembly is not available for within-accession analyses, it is important to exercise caution and check whether introgressed genes might be impacting conclusions drawn. In the long term, it is important to work towards overcoming this issue of introgression-induced reference bias by implementing novel methods.

Using a pantranscriptome reference to reduce reference bias

Previous work has shown the benefit of using enhanced references or individualised references as targets for RNA-seq mapping. Vijaya Satya, Savaljevski and Reifman [31] constructed an enhanced reference genome for human by including alternative allele segments at known polymorphic loci. Other publications have reported mapping to individualised genomes/transcriptomes by updating the reference with SNPs, INDELs and/or splice sites for each individual [9, 32]. By using individualised genomes instead of a single reference genome, Munger et al. [9] increased the accuracy of eQTL detection in a multi-parent mouse population from 88.2 to 98.3%. Kaminow et al. [33] constructed a pan-human consensus genome by calculating the consensus allele for each variant; this significantly improved the accuracy of RNA-seq mapping when compared to the reference genome. Similar approaches have been used for reducing reference bias when mapping DNA reads [34, 35].

Our approach follows in this vein. However, individualised genomes or consensus genomes are not suitable for wheat as the degree of divergence introduced by introgressions prohibits the accurate genotyping necessary for creating said genomes. Instead, we built a pantranscriptome reference that includes transcripts from other wheat cultivars in the Chinese Spring reference transcriptome. The low resource requirements of kallisto regardless of reference size enables a highly scalable approach as more genome and transcriptome data are generated, while still running in a fraction of the time that alignment-based tools take to align to one reference genome.

The pantranscriptome reference corrects almost all expression values underestimated for genes belonging to an introgression present in the assembled pangenome cultivars and in a 1-to-1 relationship with a Chinese Spring gene. However, this approach does currently have limitations. The pantranscriptome reference will not currently contain all introgressions present across wheat accessions. The pantranscriptome reference is not representative of wheat germplasm around the world; for example, it lacks, with the exception of Chinese Spring, transcripts from Asian and African wheat cultivars. There are several such genomes whose transcripts could be incorporated into the pantranscriptome [36–39]. However, we opted to include only those genomes annotated using the same methodology to ensure accurate orthologue assignment.

As more genomes and/or transcriptomes are sequenced and other existing genomes are re-annotated to provide consistent gene annotations, transcripts can be added to the pantranscriptome reference to broaden the scope of genetic variation covered. This may lead to a saturation point at which most of the commonly segregating variation is captured within the reference and it can be considered complete for most use cases. This approach also only addresses errors caused by divergent genes and not those caused by copy number variation such as tandem duplications, and presence/absence variation caused by a cultivar having a gene deletion or a novel gene. This is because, to ensure additional errors were not introduced, we elected to only add transcripts from other cultivars to the pantranscriptome reference if they came from genes in a 1-to-1 orthologous relationship with a Chinese Spring gene. Developing a way to overcome this limitation is important but also challenging because it requires resolving complex orthologue and paralogue relationships, and it is unclear how novel genes and genes with varying copy number between cultivars should be represented in the pantranscriptome reference.

Different solutions entirely to the problem of RNA-seq reference bias in wheat may emerge as being superior. For example, the field of graph genomes is developing rapidly [40, 41], including methods to align RNA-seq reads to a graph genome [42]. However, graphs for genomes as large and as complex as wheat are yet to be created successfully. It is also a much heavier-weight solution compared to the pantranscriptome pseudoalignment approach. At the very least, our approach provides a temporary way to improve the accuracy of RNA-seq alignment, particularly for those genes comprising the core genome. With further development and the incorporation of new

data, it may evolve into an alternative, more lightweight approach to emerging graph-based methods.

Examining reference bias in experimentally generated RNA-seq data

Using the valuable dataset generated by He et al. [11], we were able to show that reference bias is present in experimentally generated datasets as well as simulated datasets. The diverse nature of the wheat accessions sequenced may have made this work particularly prone to the effects of reference bias; after all, we demonstrated that divergent regions are abundant across the accessions. However, the ubiquity of introgressions is not exclusive to this set of accessions as introgressions are common across most wheat germplasm, including Elite cultivars. Indeed, wheat accessions containing diverse introgressions are very important in wheat research as it may be the source of beneficial variation for breeders, not to mention sources of insight into the evolution of wheat genomes.

The homoeologous sets of genes showing lack of expression correlation identified by He et al. [11] were enriched in genomic regions identified as introgressed or deleted in many of the accessions with 78.2% falling in such regions. We also showed that most of these genes had much higher expression when using the pantranscriptome reference instead of the Chinese Spring reference. Using the pantranscriptome reference also increased the SCC scores calculated between homoeologue pairs. These findings may alter the interpretation of why these genes are associated with productivity traits. While some of these triads may still exhibit genuine dysregulation of homoeologues and homoeologue dosage effects, it is likely that, for at least some of these genes, variation in the gene sequence itself is underlying this trait variation, rather than alteration of expression dosage between homoeologues. This also has implications for the evolutionary and selection mechanisms implicated in the control of these traits.

To more precisely examine how the quantification of introgressed genes changes with the reference used, we focused on genes in the chr1D introgression due to its presence in around a quarter of the accessions and constant size across accessions possessing it. We showed that when using Chinese Spring as the reference, it appears as though introgressed genes are less expressed than the wheat orthologues they replaced. However, when using the pantranscriptome reference, which contains the introgressed gene models as the cultivar Jagger also contains this introgression, there is no significant difference between the expression of these genes. Correcting the quantification of these genes also altered the estimated expression balance across triads in which the D homoeologue is introgressed by raising the estimated expression of the D homoeologue. It would not have been surprising to see, even after removing reference bias, that introgressed genes were expressed differently than the wheat orthologue they replace, perhaps due to the divergence in regulatory sequences. However, this finding suggests that, at least for this introgression, that is not the case. This has implications for any RNA-seq studies using wheat accessions containing introgressions, and also more specifically for studies looking at the expression of introgressed genes and what mechanisms underlie the phenotype they confer.

Conclusions

Our results highlight the problem of reference bias in wheat RNA-seq alignment which, when relying on a single reference genome, lead to inaccurate gene expression quantification and incorrect assignment of homoeologue expression balance. This effect was shown using both simulated and experimentally generated data. As divergent introgressed genes play a major role in this reference bias, incorporating divergent gene models from different wheat cultivars into the transcriptome reference reduced the extent of reference bias and provides a novel method which can be further developed as high-quality genome assemblies become available for more cultivars.

Methods

Read simulation, alignment and quantification

Reads were simulated from the longest transcript from each HC gene in Chinese Spring RefSeq v1.0 [43] (with RefSeq v1.1 annotation) and the nine pseudomolecule genome assemblies [22] if the transcript ≥ 500 bp. Wgsim from samtools v1.9 [44] was used to simulate 1000 pairs of 150 bp reads per gene with an insert size of 400 bp and no errors.

The kallisto index was produced from the CDS sequences from the RefSeq v1.1 high-confidence gene annotations using kallisto v0.44.0 [3]. Reads were pseudoaligned to this index using 100 bootstraps and default settings. Read counts and TPM values were summed across transcripts to generate gene level counts and TPM values.

To construct the pantranscriptome reference, we first ran Orthofinder [25] with standard parameters to define orthogroups based on the longest isoform protein sequences of the HC genes from Chinese Spring and the nine cultivars for which chromosome-level genome assemblies were generated as part of the 10+ genome project [20]. If a gene was found in a 1-to-1 relationship with a Chinese Spring gene, its transcripts were added to the Chinese Spring RefSeq v1.1 HC transcript fasta file. A kallisto index was built and reads pseudoaligned as above. Read counts and TPMs were each summed across

all transcripts of a gene and its 1-to-1 orthologues using the custom python script *sum_orthologue_transcript_counts.py* [53] to generate gene-level counts.

The STAR index was built for RefSeq v1.0 with the RefSeq v1.1 HC gene annotation using STAR v2.7.6a [2] using default parameters except for –limitGenomeGenerateRAM 200000000000 and –genomeSAindexNbases 12. The simulated reads from the 10 cultivars were aligned to this index using STAR and the predicted splice junctions from all were merged and then filtered to remove non-canonical junctions, junctions supported by 2 or fewer uniquely mapping reads and reads already annotated in the original genome annotation. The index was rebuilt using these discovered splice sites in addition to the annotated splice sites. The simulated reads from the 10 cultivars were aligned to this new index with parameters –quantMode TranscriptomeSAM and –outSAMunmapped Within. Gene-level read counts were generated using RSEM v1.2.28 [45].

For read count comparisons between self-mapping and cross-mapping, the following criteria were used to determine whether a gene was present in the analysis. For self-mapping, all genes from which reads were simulated were used. For cross-mapping, genes from which reads were simulated in that cultivar and that are in a 1-to-1 relationship with a gene in Chinese Spring from which reads were also simulated were used.

Defining triad balance

Triads in Chinese Spring were taken from Ramírez-González et al. [10]. For each cultivar, triads were retained if all three homoeologues were used to simulate RNA-seq reads. Triad balance was computed in the same way as [10] except for the use of read counts rather than TPMs due to the way we simulated the reads. The relative read count of each homoeologue within a triad was calculated as follows:

$$A_{norm} = \frac{A}{A+B+D}$$

$$B_{norm} = \frac{B}{A+B+D}$$

$$D_{norm} = \frac{D}{A+B+D}$$

where A, B and D are the read counts of the A, B and D homoeologues, respectively. Euclidean distance was then used to calculate the distance between each set of normalised expression values across a triad to an ideal normalised read count bias for each of seven categories (Table 1). A triad is assigned to an expression bias category by selecting the category with the shortest Euclidean distance between the observed and the ideal bias.

Table 1 Ideal normalised read count bias for each triad expression category

Category	A	B	D
Balanced	0.33	0.33	0.33
A suppressed	0	0.5	0.5
B suppressed	0.5	0	0.5
D suppressed	0.5	0.5	0
A dominant	1	0	0
B dominant	0	1	0
D dominant	0	0	1

Calculating CDS identity

Blastn from blast+v2.7.1 [46] was used to align the nucleotide sequence of the longest transcripts of pairs of orthologues between Chinese Spring RefSeq v1.1 and Lancer. The identity of the best hit between pairs was taken and binned into 5-Mbp genomic windows.

Binning incorrectly quantified genes

The RefSeq v1.0 genome [43] was split into 5-Mbp genomic windows using bedtools makewindows [47] and for each window, a score was calculated based on the number of under (read count < 500) and overestimated (read count > 1500) genes within that window:

$$(-1 * no.of\ \ underestimated\ genes) + (1 * no.overestimated\ genes)$$

Processing sequencing data generated by He et al. [11]

One hundred ninety-eight accessions had both leaf RNA-seq data and enrichment capture short paired-end DNA reads. The RNA-seq data from the 198 lines was pseudoaligned to both Chinese Spring RefSeq v1.1 and the pantranscriptome reference as above for the simulated reads. TPMs were summed across transcripts to generate gene level counts. Accessions GF25, GF270, GF32, GF37, GF41 and GF73 were excluded for RNA-seq analyses as in [11].

DNA reads were mapped to Chinese Spring RefSeq v1.0 [43]. The alignment was filtered using samtools [44]: supplementary alignments, improperly paired reads, and non-uniquely mapped reads (mapping quality less than 10) were removed. PCR duplicates were detected and removed using the Picard Tools v2.1.1 MarkDuplicates

function [48]. Accessions GF294, GF342, GF366, GF380, GF381, GF383 and GF38 were excluded for DNA analyses as in [11].

Using mapping coverage deviation to identify divergent regions of the genome

To generate DNA sequencing reads for the cultivars assembled as part of the 10+ wheat genomes project, we simulated paired-end 150-bp reads with 500-bp insert and no errors from all fourteen *Triticum aestivum* genome assemblies (ArinaLrFor, Cadenza, Claire, Jagger, Julius, Lancer, Landmark, Mace, Norin61, Paragon, Robigus, Stanley, SY Mattis and Weebil) [20] to a depth of 10x using WGSim within samtools v1.9 [44]. Reads were mapped to RefSeq v1.0 as above.

The RefSeq v1.0 genome [43] was split into 1-Mbp genomic windows using bedtools makewindows [47]. Using the filtered read mappings for the cultivars from the 10+ wheat genomes [20] project and for the accessions analysed by He et al. [11], the number of reads mapping to each window was computed using hts-nim-tools [49]. To normalise by the sequencing depth of each line, read counts were divided by the number of mapped reads that passed the filters, producing normalised read counts. Different windows of the genome have variable mapping coverage rates, so to compute coverage deviation we must compare each window to the same window in the other lines in the collection. Median normalised read counts, m, were produced, containing the median for each genomic window. Mapping coverage deviation was then defined for each line as:

$$d_i = \frac{C_i}{m_i \cdot \varepsilon}$$

for window $i \in \{1, 2, \ldots, n\}$, where ε is the median d value across the genome for the line. Statistically significant d values were calculated using the scores function from the R package 'outliers' using median absolute deviation and probability of 0.99. Mapping coverage deviation and significance values were computed separately for the cultivars from the 10+ wheat genomes project [20] and for the accessions analysed by He et al. [11].

Locating coordinates of introgression boundaries

To detect the precise locations of the chr1D, chr2A *Ae. ventricosa*, and the chr2D *Ae. markgrafii* introgressions in Jagger, and the chr2B *T. timopheevii* and the chr3D *Th. ponticum* introgression in Lancer, I used the alignments for the simulated Jagger and Lancer reads generated above. Read depths were binned into 5- and 1-Mbp windows using bedtools makewindows [47] and hts-nim-tools [49]. The window in which read depth drops, signifying the start/end of the introgression, was identified for each introgression and IGV was used to precisely identify the position where the coverage profile changes. To locate the location of the introgressions relative to the Jagger/Lancer genomes in order to identify which genes have been introgressed, I extracted Chinese Spring sequence 1Mbp either side of the precisely located border position (or until the end of the chromosome) for each introgression and aligned them to the Jagger or Lancer genome assembly using minimap2 [50] with parameters -x asm5. These alignments were used to determine the borders of the introgressed region as they appear in their donor genomes.

Characterising the chr1D introgression donor species

Blastp from blast+ v2.7.1 [46] was used to align the *Ae. speltoides* proteins with the longest isoforms of the Jagger HC proteins. The best hit for each Jagger protein was kept. Paired-end Illumina DNA reads from *T. timopheevii* [51] were mapped to Chinese Spring RefSeq v1.0 [43] using BWA mem v0.7.13 [52]. Samtools v1.4 [44] was used to filter the alignments to retain mapped reads, primary alignments, properly paired reads and uniquely mapping reads (mapping quality greater than 10). PCR duplicates were found and removed using the Picard Tools v2.1.1 MarkDuplicates function [48]. Read depths were binned into 5-Mbp windows using bedtools makewindows [47] and hts-nim-tools [49] and divided by window length to account for windows at ends of chromosomes which are less than 5Mbp in length.

Calculating SCC between homoeologues

SCC scores were calculated between AB, AD and BD homoeologue pairs for triads where one homoeologue was in the set of genes showing lack of expression correlation identified by He et al. [11]. This was done using the cor.test function in R with the 'Spearman' method and the lowest SCC value of the three comparisons was taken. Triads were excluded if any of the homoeologues were not found in the HC RefSeq v1.1 annotation.

Statistical tests

The significance of the difference in the proportion of genes that were correctly quantified between introgressed and non-introgressed regions was calculated using a chi-squared test with a sample size of 60,338. The significance of the difference between mean CDS nucleotide identity between orthologue pairs when correctly quantified compared to incorrectly quantified was calculated using two-tailed t tests with no

assumption of equal variance and a sample size of 60,338. The significance of the difference in Spearman correlation scores between homoeologue pairs when using the Chinese Spring reference compared to the pantranscriptome reference was calculated using a two-tailed *t* test with no assumption of equal variance and a sample of 55. The significance of the difference between introgressed and non-introgressed orthologues when using the Chinese Spring or the pantranscriptome reference was calculated using two-tailed *t* tests with no assumption of equal variance with a sample size of 63.

Abbreviations

eQTL	Expression quantitative trait locus
HC	High confidence
SCC	Spearman's correlation coefficient

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-01853-w.

Additional file 1: Table S1. Number of genes correctly quantified, underestimated, and overestimated from simulated RNA-Seq data, using Kallisto with the Chinese Spring reference, STAR with the Chinese Spring reference, or kallisto with the pantranscriptome reference. **Table S2.** Percentage of triads classified in each expression category from simulated RNA-Seq data, using Kallisto with the Chinese Spring reference, STAR with the Chinese Spring reference, or kallisto with the pantranscriptome reference. **Table S3.** Accessions from the He *et al*. [11] dataset that do and do not contain the chr1D introgression. **Table S4.** Expression values for genes within the chr1D introgression for accessions from the He *et al*. [11] dataset, using either the Chinese Spring or the pantranscriptome reference. Accessions are split based on whether or not they contain the chr1D introgression.

Additional file 2: Fig. S1. Upset plot of 1-to-1 orthologue assignments used for the construction of the pantranscriptome reference. **Fig. S2.** Remaining incorrectly quantified genes after correction using the pantranscriptome reference. **Fig. S3.** Reads from *T. timopheevii* accession P95 mapped to *T. aestivum* cv. Jagger and binned into 5Mbp genomic windows.

Acknowledgements

We would like to thank Jose De Vega and Rachel Rusholme-Pilcher for providing feedback on an earlier version of the manuscript.

Authors' contributions

BC conceived of the study and conducted analysis, prepared figures and wrote the manuscript. TL identified 1-to-1 orthologues between Chinese Spring and the cultivars assembled in the 10+ wheat genomes project. AH provided supervision and edited the manuscript. EA was involved in discussions and edited the manuscript. All authors read and approved the final version of the manuscript.

Funding

BC was supported by the BBSRC funded Norwich Research Park Biosciences Doctoral Training Partnership grant BB/M011216/1. AH was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; Earlham Institute Strategic Programme Grant BBX011089/1 and BBS/E/ER/230002B (*Decode WP2 Genome Enabled Analysis of Diversity to Identify Gene Function, Biosynthetic Pathways And Variation In Agri/Aquacultural Traits)*. EA is supported by the Agriculture and Food Research Initiative Competitive Grants 2022–68013-36439 (WheatCAP) and grant INV-004430 from Bill and Melinda Gates Foundation.

Availability of data and materials

The pantranscriptome reference, along with a python script to sum expression counts across all transcripts of a given Chinese Spring gene and its 1-to-1 orthologues, can be accessed via figshare at https://doi.org/10.6084/m9.figshare.24242767 [53].
The RNA-seq data and DNA sequencing data generated by He et al. [11] are stored in the European Nucleotide Archive under project codes PRJNA670223 [54] and PRJNA787276 [55].
The wheat cultivar genomes and annotations generated as part of the 10+ wheat genomes project [20] can be accessed on Ensembl Plants release 58 via https://plants.ensembl.org/Triticum_aestivum/Info/Cultivars [56].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 20 October 2023 Accepted: 21 February 2024
Published online: 07 March 2024

References

1. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
3. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
4. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
5. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. 2019;15(7): e1008302.
6. Thorburn DMJ, Sagonas K, Binzer-Panchal M, Chain FJJ, Feulner PGD, Bornberg-Bauer E, et al. Origin matters: Using a local reference genome improves measures in population genomics. Mol Ecol Resour. 2023;23:1706–23.
7. Zhan S, Griswold C, Lukens L. Zea mays RNA-seq estimated transcript abundances are strongly affected by read mapping bias. BMC Genomics. 2021;22:285.
8. Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, et al. Mendelian and non-mendelian regulation of gene expression in Maize. PLoS Genet. 2013;9(1): e1007234.
9. Munger SC, Raghupathy N, Choi K, Simons AK, Gatti DM, Hinerfeld DA, et al. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. Genetics. 2014;198(1):59–73.
10. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, et al. The transcriptional landscape of polyploid wheat. Science. 2018; 361(6403):eaar6089.
11. He F, Wang W, Rutter WB, Jordan KW, Ren J, Taagen E, DeWitt N, Sehgal D, Sukumaran S, Dreisigacker S, Reynolds M, Halder J, Sehgal SK, Liu S, Chen J, Fritz A, Cook J, Brown-Guedira G, Pumphrey M, Carter A, Sorrells M, Dubcovsky J, Hayden MJ, Akhunova A, Morrell PL, Szabo L, Rouse M, Akhunov E. Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. Nat Commun. 2022;13(826). https://doi.org/10.1038/s41467-022-28453-y.
12. Edelman NB, Mallet J. Prevalence and adaptive impact of introgression. Ann Rev Genet. 2021;55:265–83.
13. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005;20(5):229–37.

14. Hao M, Zhang L, Ning S, Huang L, Yuan Z, Wu B, et al. The resurgence of introgression breeding, as exemplified in wheat improvement. Front Plant Sci. 2020;11:252.
15. Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, et al. Triticum population sequencing provides insights into wheat adaptation. Nat Genet. 2020;52(12):1412–22.
16. Cheng J, Liu J, Wen J, Nie X, Xu L, Chen N, Li Z, Wang Q, Zheng Z, Li M, Cui L, Liu Z, Bian J, Wang Z, Xu S, Yang Q, Appels R, Han D, Song W, Sun Q, Jiang Y. Frequency intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol. 2019;20(136). https://doi.org/10.1186/s13059-019-1744-x.
17. He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet. 2019;51:896–904.
18. Przewieslik-Allen AM, Burridge AJ, Wilkinson PA, Winfield MO, Shaw DS, McAusland L, et al. Developing a High-Throughput SNP-based marker system to facilitate the introgression of traits from aegilops species into bread wheat (Triticum aestivum). Front Plant Sci. 2019;9:1993.
19. Wang Z, Wang W, Xie X, Wang Y, Yang Z, Peng H, et al. Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference. Nat Commun. 2022;13:3891.
20. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature. 2020;588(7837):277–83.
21. Keilwagen J, Lehnert H, Berner T, Badaeva E, Himmelbach A, Börner A, et al. Detecting major introgressions in wheat and their putative origins using coverage analysis. Sci Rep. 2022;12:1908.
22. White B, Lux T, Rusholme-Pilcher R, Kaithakottil G, Duncan S, Simmonds J, et al. De novo annotation of the wheat pan-genome reveals complexity and diversity within the hexaploid wheat pan-transcriptome. BioRxiv. 2024. https://doi.org/10.1101/2024.01.09.574802.
23. Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo M-C. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Mol Biol Evol. 2006;23(7):1386–96.
24. Gao L, Koo D-H, Juliana P, Rife T, Singh D, Lemes da Silva C, et al. The Aegilops ventricosa 2NvS segment in bread wheat: cytology, genomics and breeding. Theor Appl Genet. 2021;134(2):529–42.
25. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
26. Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, et al. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 2020;21:239.
27. Wang J, Wen W, Hanif M, Xia X, Wang H, Liu S, et al. TaELF3-1DL, a homolog of ELF3, is associated with heading date in bread wheat. Mol Breed. 2016;36:161.
28. Wittern L, Steed G, Taylor LJ, Ramirez DC, Pingarron-Cardenas G, Gardner K, et al. Wheat EARLY FLOWERING 3 affects heading date without disrupting circadian oscillations. Plant Physiol. 2023;191(2):1383–403.
29. Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO. Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics. 2018;19:510.
30. Rees H, Rusholme-Pilcher R, Bailey P, Colmer J, White B, Reynolds C, et al. Circadian regulation of the transcriptome in a complex polyploid crop. PLoS Biol. 2022;20(10): e3001802.
31. Vijaya Satya R, Zavaljevski N, Reifman J. A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Res. 2012;40(16): e127.
32. Liu X, MacLeod JN, Liu J. iMapSplice: Alleviating reference bias through personalized RNA-seq alignment. PLoS ONE. 2018;13:8.
33. Kaminow B, Ballouz S, Gillis J, Dobin A. Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. Genome Res. 2022;32:738–50.
34. Chen NC, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. Genome Biol. 2021;22:8.
35. Vaddadi NSK, Mun T, Langmead B. Minimizing Reference Bias with an Impute-First Approach. bioRxiv. 2023. https://doi.org/10.1101/2023.
36. Athiyannan N, Abrouk M, Boshoff WHP, Cauet S, Rodde N, Kudrna D, et al. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat Genet. 2022;54:227–31.
37. Guo W, Xin M, Wang Z, Yao Y, Hu Z, Song W, et al. Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat Commun. 2020;11:5085.
38. Shi X, Cui F, Han X, He Y, Zhao L, Zhang N, et al. Comparative genomic and transcriptomic analyses uncover the molecular basis of high nitrogen-use efficiency in the wheat cultivar Kenong 9204. Mol Plant. 2022;15(9):1440–56.
39. Jia J, Zhao G, Li D, Wang K, Kong C, Deng P, et al. Genome resources for the elite bread wheat cultivar Aikang 58 and mining of elite homeologous haplotypes for accelerating wheat improvement. Mol Plant. 2023;16(12):1893–910.
40. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36(9):875–81.
41. Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. 2020;21:250.
42. Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, et al. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat Methods. 2023;20:239–47.
43. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361(6403):eaar7191.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
45. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics. 2009;10:421.
47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
48. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
49. Pedersen BS, Quinlan AR. hts-nim: scripting high-performance genomic analyses. Bioinformatics. 2018;34(18):3387–9.
50. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
51. King J, Grewal S, Othmeni M, Coombes B, Yang CY, Walter N, Ashling S, Scholefield D, Walker J, Hubbart-Edwards S, Hall A, King IP. Introgression of the Triticum timopheevii Genome Into Wheat Detected by Chromosome-Specific Kompetitive Allele Specific PCR Markers. Front Plant Sci. 2022;13(919519). https://doi.org/10.3389/fpls.2022.919519.
52. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
53. Coombes B, Lux T, Akhunov E, Hall A. Supplementary Data for paper titled 'Introgressions lead to reference bias in wheat RNA-Seq analysis'. 2023. figshare https://doi.org/10.6084/m9.figshare.24242767.v1.
54. RNA-seq data for a wheat diversity panel. ENA https://www.ebi.ac.uk/ena/browser/view/PRJNA670223 (2022).
55. Regulatory sequence diversity in the wheat genome. ENA https://www.ebi.ac.uk/ena/browser/view/PRJNA787276 (2020).
56. Yates DY, Allen J, Amode RM, Azov AG, Barba M, Becerra A, et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res. 2022;50:D996–1003.

Publisher's Note