Deciphering the molecular mechanisms controlling grain length and width in polyploid wheat

Jemima Florence Brinton

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

John Innes Centre

September 2017

This work is dedicated to my Grandma, Jean Brinton

Abstract

There is an urgent need to increase crop yields to address food insecurity. Grain weight, determined by grain length and width, is an important component of final grain yield. However, our understanding of the mechanisms that control grain weight in polyploid wheat is limited. The overall aim of this thesis was to understand the mechanisms that control grain length and width in hexaploid wheat through the characterisation of two previously identified grain weight quantitative trait loci (QTL) on chromosomes 5A and 6A.

Using near isogenic lines (NILs) we found that the 5A and 6A QTL act through different mechanisms to increase grain weight. The 5A QTL acts post-fertilisation, primarily to increase grain length (4.0%) through increased pericarp cell size. The 5A QTL also has a pleiotropic effect on grain width (1.5%) during late grain development. The 6A QTL acts during very early grain development, perhaps pre-fertilisation, and specifically increases final grain width (2.3%).

Fine-mapping reduced the QTL mapping intervals and revealed complex underlying genetic architectures. The 6A QTL mapped to a large linkage block in the centromeric region of chromosome 6A containing the known grain size gene, *TaGW2_A*, although we provide evidence to suggest that this is not the causal gene underlying the 6A QTL. Fine-mapping of the 5A QTL suggests that two tightly linked genes with an additive effect on grain length underlie the locus. A haplotype analysis suggests that the 5A QTL is not fixed in UK germplasm.

The corresponding physical intervals for both the 6A and 5A QTL remain large and contain several hundred genes, making speculation on candidates for the causal genes difficult. A transcriptomics study with the 5A NILs provided insight into the genes and pathways that are differentially regulated and hence may play a role in controlling the differences in grain weight.

Acknowledgements

Firstly, I would like to thank my supervisor, Cristobal Uauy, for his endless enthusiasm, support, and guidance. Thank you for introducing me to the world of wheat and for all the opportunities you have given me during my PhD, I have learnt so much and I could not have asked for a better supervisor. I would also like to thank my secondary supervisor, Alison Smith, for the insight and encouragement she has provided throughout this project in addition to the time I spent in her lab as an undergraduate, an experience which motivated me to embark on a PhD in the first place.

Thank you as well to all those involved in my rotation year. Tony Miller and Dale Sanders for my time in their lab, to my fellow rotation students: Nuno, Andrew, Jie and Marc and to Nick Brewin and Stephen Bornemann for guiding us through that first year.

I would also like to thank the John Innes Foundation and the Agriculture and Horticulture Development Board for funding my PhD. The AHDB student conferences and industry visits have been very enjoyable events and extremely valuable for gaining insights into the wider applications of research.

There are so many people who have contributed to this project and without whom it literally would not have been possible. In particular, James Simmonds, who not only did all the preliminary work that created the project in the first place, but also helped me find my feet in the lab when I started and has continued to put in a huge amount of help, time, work and support right to the end – thank you so much. The help of Peter Scott was also invaluable – my first field season without you felt very strange! Thank you to all of those who put so much time and hard work into the enormous harvesting and phenotyping efforts, special mentions to Tobin and Pam who have suffered it year on year! Thank you also to Francesca for her commitment to the cell size imaging/measuring, to Ricardo, Philippa and Nikolai for bioinformatics help and resources and to Abdul Kader for taking on the mutants. I am also very grateful to the horticultural staff, bioimaging group, genotyping service and field trials team for all of their help.

I have been extremely lucky during my PhD to be part of a brilliant lab group. You are a truly lovely group of people to work with; fun and helpful in equal measure ☺. Thank you to those I have had the pleasure of sharing an office with at various points over the last four years – Nikolai, Philippa, Sophie, Olu, Sarah and Emilie – I can always count on you for help/cake/laughter/gossip! A special thanks to Nikolai for keeping me motivated with baked goods in the final few weeks (and generally for all the time spent answering my many questions) and to Sophie for being there to listen and provide a positive outlook when the last few months have felt particularly overwhelming.

Outside of the lab, I have found an excellent group of friends in my fellow PhD students, with whom I have many fond memories and hopefully many more to come! Thank you especially to Rachel for her friendship over the last four years, from moral support to giving up your bank holiday weekend to help me in the field (even if it was just to test your new wellies) – you have been there through it all! A huge thank you also to Marc for all of the laughter – and helpful scientific input – you have kept me sane. Thank you as well to the friends who pre-date the PhD, both from university and the ‘Gaggle’ from home, for sticking around despite my terrible communication and providing an escape from the PhD bubble.

I would lastly like to thank my family, of whom there are far too many to mention one by one, although I wish I could. Thank you all for your love and support always, but especially in the last few weeks. The thought of making you all proud is the thing that has pushed me through to the end. I would like to thank in particular my grandparents. Grandma, I am so sad that you weren’t quite able to see me finish this but the last time we spoke you told me that I really just needed to get it over and done with and so that is what I have done. I know that Grandpa will be proud for both of you. Thank you to my brothers, Timothy and Jeremy, for always keeping me grounded and to my sister, Isabella, for being interested in everything and never failing to put a smile on my face. Finally, my biggest thank you of all is to my wonderful Mum for your constant encouragement and belief in me, for listening to my endless rants and worries, and for never being more than a phone call away. Thank you.

Table of contents

Table of figures

Table of tables

List of Abbreviations

BA	Bristol Axiom
BC	Back-cross
BCF	Binary call format
bp	Base pair
BS	Bristol SNP
cM	centimorgan
CSS	Chromosome Survey Sequence
CxB	Charger x Badger
DE	Differentially expressed
DH	Doubled haploid
dpa	days post anthesis
DW	Dry Weight
EMS	Ethyl methanesulfonate
FDR	False-discovery rate
FW	Fresh weight
Gbp	Giga base pair
GCD	Green canopy duration
GL1	Grain Length 1
GL2	Grain Length 2
GO	Gene ontology
HetRec	Heterozygous recombinant
HomRec	Homozygous recombinant
IWGSC	International Wheat Genome Sequencing Consortium
JB_RNASeq	Jemima Brinton RNASeq
JBHap	Jemima Brinton Haplotype
KASP	Kompetitive Allele Specific PCR
kb	Kilo base pair
LD	Linkage disequilibrium
LOD	Log-of-odds
Mbp	Mega base pair
MTME	Multi-trait multi-environment
NGS	Next generation sequencing
NIL	Near isogenic lines
PCD	Programmed cell death
POPSEQ	Population sequencing
PWM	Position weight matrix
QTL	Quantitative trait loci
RIL	Recombinant inbred line
RNAi	RNA interference
RNA-Seq	RNA Sequencing
SEM	Scanning election microscope
SES	Single ear samples
SNP	Single nucleotide polymorphism
SRA	Sequence read archive
SxR	Spark x Rialto
TF	Transcription factor

continued on next page

List of abbreviations continued

TFBS	Transcription factor binding site
TGW	Thousand Grain Weight
TILLING	Targeting Induced Local Lesions in Genomes
tpm	transcripts per million
Ub	Ubiquitin
UTR	Untranslated region
VCF	Variant call format
WT	Wild-type

1 General introduction

1.1 Crop yields must increase to meet global food demand

It is predicted that by 2050, the global human population will have exceeded nine billion and this is driving an increased demand for food production (United Nations, 2015). This increased demand is exacerbated by competition from crops used for biofuels, increased pressures on agricultural systems from climate change and changing dietary habits. Space for agricultural expansion is limited and therefore a sustainable route to meet this demand is to increase crop production on existing farmlands. Projections have shown that increasing crop yields on land already used for agriculture could significantly reduce the number of people at risk of hunger globally by increasing the available food supply and reducing prices (Rosegrant *et al.*, 2013). However, whilst huge improvements in yield were achieved during the Green Revolution, rates of increase in crop yields have slowed in recent years (Figure 1.1) and are currently insufficient to achieve the estimated doubling in crop production required by 2050 (Tilman *et al.*, 2011; Ray *et al.*, 2013). With one in nine people in the world currently living under food insecurity and the proportion of the global population suffering from chronic hunger increasing in 2016 for the first time in a decade (FAO *et al.*, 2017), it is urgent that we identify ways to increase crop yields.

1.2 Wheat is a crop of global importance

Wheat is one of the world's most important crops and is grown on all five non-polar continents (Figure 1.2), on more land area across the globe than any other crop (FAO, 2017). Wheat plays an important role in human nutrition, in fact most people consume 50 wheat plants every day. It provides one-fifth of the human calorific intake and more protein globally than all types of meat combined (FAO, 2017).

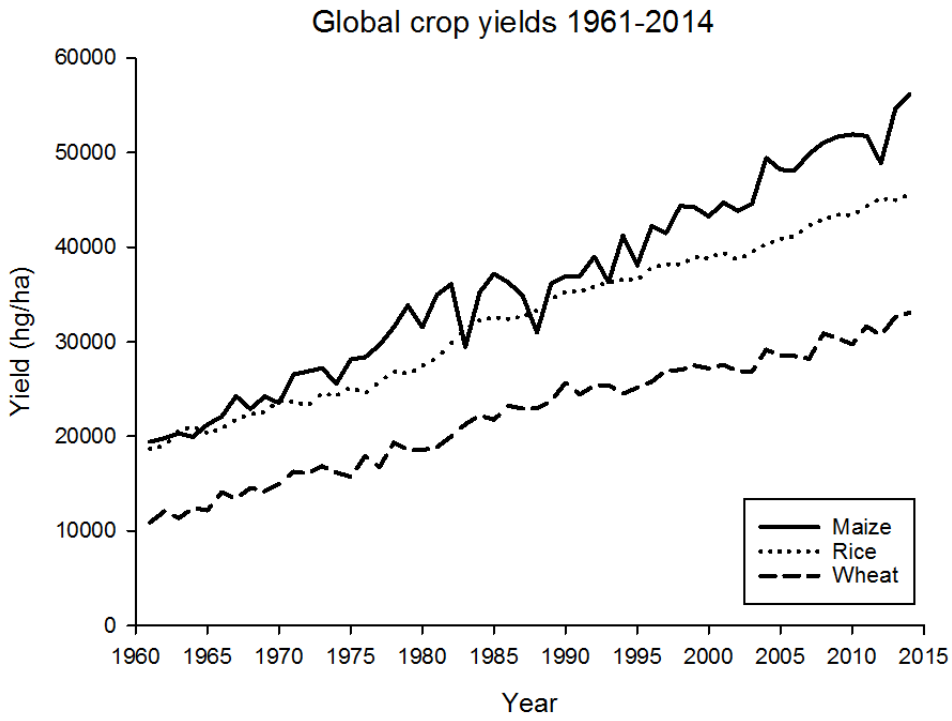


Figure 1.1: Global crop yields 1961-2014

Source http://www.fao.org/faostat/en/#data/QC, accessed 21-09-2017

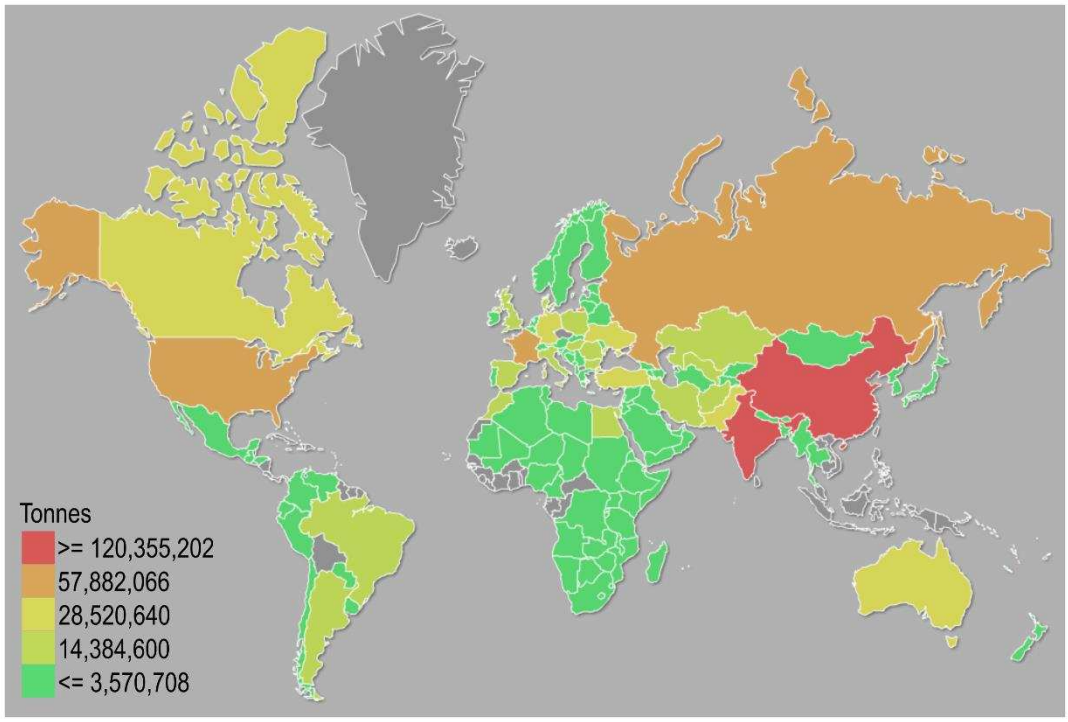


Figure 1.2: Global wheat production in 2010-2014

Source http://www.fao.org/faostat/en/#data/QC, accessed 21-09-2017

1.2.1 Modern cultivated wheat is an allopolyploid

As a result of its evolutionary history, modern wheat is an allopolyploid, i.e. a polyploid that has arisen through the hybridisation of chromosomes from different species (Comai, 2005). Around 400,000 years ago, a hybridisation event between two diploid grass species (*Triticum urartu* (AA)) and an unknown member of the *Sitopsis* family (BB) gave rise to the tetraploid wild emmer *Triticum turgidum* ssp. *dicoccoides* (AABB) (Haider, 2013). The selection for non-brittle rachis types that do not disperse seeds upon maturity led to domesticated emmer wheat (*Triticum turgidum* spp. *dicoccon* (AABB)) in the Fertile Crescent roughly 10,000 years ago (Luo *et al.*, 2007). Subsequent selection for free-threshing characteristics gave rise to modern day durum wheat *Triticum durum* (AABB); (Feldman, 2001). Semolina based products, such as pasta, are usually made from durum wheat, hence it is commonly known as pasta wheat. A second hybridisation event between emmer wheat (AABB) and *Aegilops tauschii* (DD) gave rise to the hexaploid wheat, *Triticum aestivum* (AABBDD) (Petersen *et al.*, 2006). Given that flour based products, including bread, are made from hexaploid wheat, *T. aestivum* is commonly referred to as bread wheat. However, most other forms of wheat consumption such as breakfast cereals, biscuits, pastries etc are also made from bread wheat with different industrial processing qualities. Bread wheat accounts for >95 % of wheat grown globally and was the focus of this work, hence we will refer to bread wheat as simply wheat throughout this thesis unless otherwise stated.

1.2.2 Wheat genomics resources

The three constituent genomes of hexaploid wheat (A, B and D) are referred to as homoeologous genomes each containing seven chromosomes, and share 96-98 % sequence similarity across coding regions (Krasileva *et al.*, 2013). This, along with the large (~17 Gbp) and highly repetitive nature of the wheat genome has meant that, until recently, wheat genomic resources were limited. However, this has changed drastically during the course of this PhD with many resources now becoming available (reviewed in Borrill *et al.*, 2015a; Uauy, 2017). The resources most relevant to this thesis are described below.

During my PhD, the available wheat genome assemblies have moved from highly fragmented assemblies containing millions of unordered contigs to 21 near-complete chromosome pseudomolecule sequences. In 2014 (the first year of my PhD), the International Wheat Genome Sequencing Consortium (IWGSC) released the Chromosome Survey Sequence (CSS) of the reference hexaploid wheat cultivar, Chinese Spring (IWGSC, 2014). The CSS assembly was generated through the flow-sorting and subsequent Illumina next generation sequencing (NGS) of individual chromosome arms. One major advantage of this approach over previously assemblies was that it allowed separation of the three homoeologous genomes, which previously had not been possible (Brenchley *et al.*, 2012). However, the major limitation of the CSS assembly is that it is non-contiguous, with the exception of chromosome 3B (Choulet *et al.*, 2014), containing millions of scaffolds with no physical order. Many of the scaffold sequences were anchored to a high density genetic map using population sequencing (POPSEQ; Mascher *et al.*, 2013). However, this

resulted in scaffolds being allocated to large unordered genetic bins and over half of the scaffolds having no positional information (Borrill *et al.*, 2015a). Gene models based on the CSS assembly were generated using RNA-Sequencing (RNA-Seq) data and information from related species. However, the accuracy of these gene models was limited by the highly fragmented nature of the CSS assembly and these gene models are incomplete with respect to more recent annotations, examined in more detail in Chapter 4 of this thesis.

In the last year, a number of more complete genome assemblies of Chinese Spring have been released. The first of these was the TGACv1 assembly (Clavijo *et al.*, 2017b), which used a whole genome shotgun sequencing (WGS) approach with the W2RAP assembly pipeline (Clavijo *et al.*, 2017a) to generate an assembly with scaffolds over 20 times longer than the CSS assembly (CSS N50 = 3.3 kb, TGACv1 N50 = 88 kb; Clavijo *et al.*, 2017b). Again, the TGACv1 scaffolds are not physically ordered but were genetically anchored in the same way as the CSS assembly. In addition to the increased contiguity, one of the biggest improvements of the TGACv1 assembly was the gene models that accompanied it. These are generally more complete than the CSS gene models and include > 20,000 genes that were not included in previous gene model sets (Clavijo *et al.*, 2017b). An even more complete assembly was released in July 2017, which combined long single-molecule sequencing reads with high coverage short reads to generate an assembly with at least ten times improved contiguity over the previous sequences (Zimin *et al.*, 2017). However, no gene models associated with this assembly have yet been released. Finally, the IWGSC have generated a whole genome assembly using Illumina sequencing and a proprietary assembly algorithm called DeNovoMAGIC. These sequences have been ordered using both POPSEQ and Hi-C (chromosome conformation capture) to generate 21 chromosome pseudomolecules. The most recent release of this assembly (IWGSC RefSeqv1.0; https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) integrates this sequence data with additional resources (including chromosome physical maps, BioNano optical maps, BAC sequences and genotyping-by-sequencing maps of a well characterised mapping population) to give a chromosomal scaffold N50 of 22.8 Mb. An annotation of the IWGSC RefSeqv1.0 assembly has also been generated and recently made publicly available (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations), although not in time for analysis as part of this thesis. Instead, we have used an *in silico* mapping of the TGACv1 gene models to the IWGSC RefSeqv1.0 (David Swarbreck and Gemy Kaithakottil, Earlham Institute; available at http://www.wheat-training.com/useful-wheat-links/).

All the genome assemblies described above are based on the same wheat cultivar, Chinese Spring, therefore not allowing insight into the variation that exists between different cultivars. The identification of variation between wheat cultivars is essential for genetic studies, and the availability of genetic markers has historically imposed a bottleneck on genetic studies in wheat. Genome assemblies have been generated using the same method as the TGACv1 assembly for four UK bread wheat varieties (Robigus, Paragon, Claire and Cadenza) and the tetraploid durum wheat variety, Kronos (available at http://opendata.earlham.ac.uk/Triticum_aestivum/). Genome

assemblies of other wheat varieties from across the world and progenitor species have also been released and many more are in the pipeline (Avni *et al.*, 2017; Montenegro *et al.*, 2017). To overcome the cost and time constraints imposed by whole genome sequencing, other strategies have been employed to identify variation between large numbers of different wheat cultivars. Particularly relevant to this PhD are the 90k iSelect (Wang *et al.*, 2014) and 820k Axiom (Winfield *et al.*, 2016) single nucleotide polymorphism (SNP) arrays, which have made used of reduced-representation sequencing approaches. The 90k iSelect used RNA-Seq reads from 19 hexaploid and 18 tetraploid wheat cultivars to call SNPs, whereas the 820k Axiom SNPs were identified using exome capture sequencing data from 43 wheat cultivars. Initially, a proportion of these SNPs were genetically positioned and have now been assigned physical positions with respect to the IWGSC RefSeqv1.0 (Ricardo Ramirez-Gonzalez; available at http://www.wheat-training.com/useful-wheat-links/).

Additional resources relevant to this PhD are the wheat expression databases (Pearce *et al.*, 2015; Borrill *et al.*, 2016) and exome-sequenced mutant populations (Krasileva *et al.*, 2017). In recent years, the reduced cost of NGS resulted in a huge amount of wheat RNA-Seq data being generated. However, despite raw sequencing reads being made publicly available in the NCBI sequence read archive (SRA), this data was not available to researchers in an easily accessible form. The wheat expVIP database (www.wheat-expression.com; Borrill *et al.*, 2016) includes 418 publically available wheat RNA-Seq samples from 16 different studies. All samples have been aligned in the same way to both the CSS and TGACv1 reference transcriptomes and will soon be updated with many more studies aligned to the IWGSC RefSeqv1.0 transcriptome reference. This allows the expression profiles of genes to be examined and compared across a wide range of tissues and developmental stages. This can be useful, for example, when prioritising candidate genes for further study, however, until recently, reverse genetics resources for such studies were not available in wheat. To address this, two exome-sequenced mutant populations were generated (www.wheat-tilling.com; Krasileva *et al.*, 2017). This functional genomics resource consists of 1,535 tetraploid (cv. Kronos) and 1,200 hexaploid (cv. Cadenza) EMS mutagenised lines. Exome capture followed by Illumina NGS was performed on these lines and SNPs have been called with respect to the CSS gene models to identify mutations and predict their effects. This resource allows the rapid identification of novel mutations in specific genes for functional characterisation.

These advances have opened up many new opportunities for wheat research, many of which were exploited in this work.

1.3 Wheat development and yield components

As discussed above, crop yields must increase to meet the food demands of a growing population. Final grain yield is a highly complex trait given its polygenic inheritance and strong environmental influence which translates into low heritability. Final yield represents the cumulative phenotype expression of the complete life cycle of the plant (Slafer, 2003) meaning that most traits will have

pleiotropic effects on yield. This has limited our understanding of the underlying genetic mechanisms governing yield in wheat.

1.3.1 Overview of wheat growth and development

From seed sowing and germination to the harvesting of the mature grain, the growth of a wheat plant progresses through three main phases: vegetative, reproductive and grain filling (Figure 1.3).

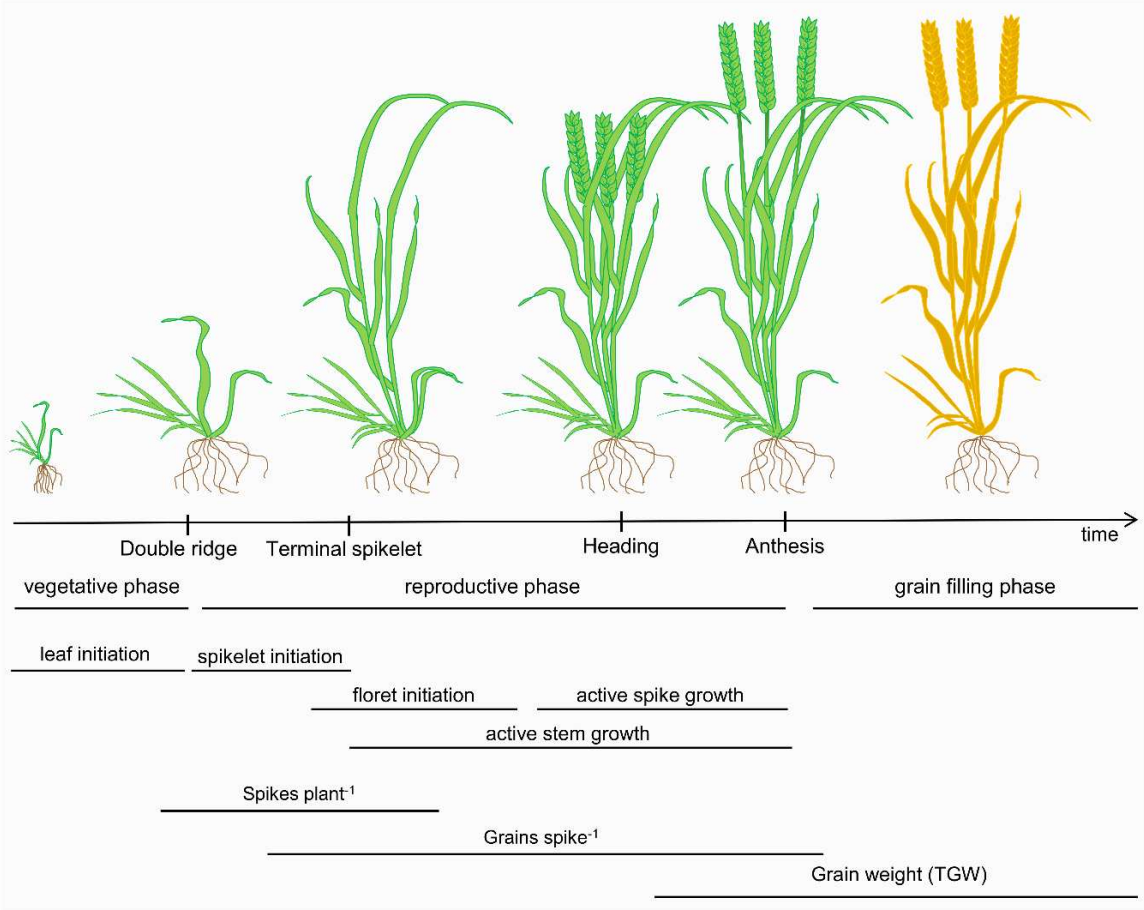


Figure 1.3: Wheat development and yield components

Adapted from Slafer (2003) and Guo *et al.* (2015)

1.3.1.1 Vegetative phase

During the vegetative phase, leaf emergence takes place and tiller initiation begins (Figure 1.3). Tillers are lateral shoots that emerge at the base of the stem with the potential to produce a wheat inflorescence (spike; Figure 1.4a) hence contributing to the number of spikes per plant (spikes plant^{-1}; Figure 1.3). Not all tillers will produce spikes, however, with usually at least the first three tillers to emerge producing fertile spikes (this will depend on planting density, soil fertility, among other factors). The duration of the vegetative phase can vary depending on whether the wheat variety is a 'winter' or 'spring' type. Winter wheats require a period of cold, known as vernalisation, to induce flowering and hence have a long vegetative phase. Spring wheats on the other hand do not require a period of vernalisation to flower and therefore develop more quickly

through the vegetative phase (reviewed in Distelfeld *et al.*, 2009). This cold requirement means that winter wheats are usually autumn-sown crops in the UK, whereas spring wheats are sown in late winter or early spring, hence their common names. The longer vegetative phase of winter wheats means that they generally produce more tillers, extra leaf area, and intercept more light across a growing season which can support higher grain yields in the mature plant. 95% of wheat grown in the UK is winter wheat.

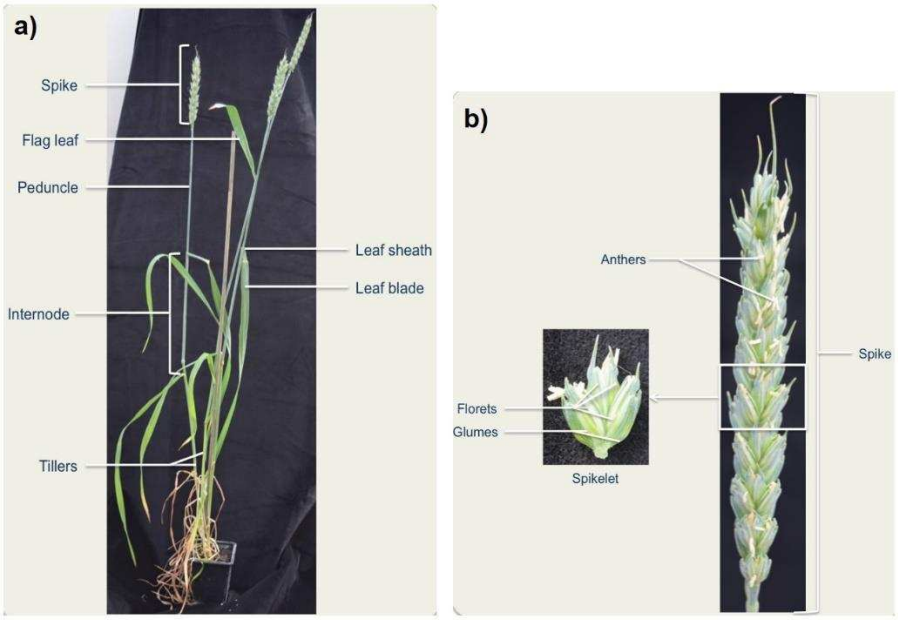


Figure 1.4: Structure of a wheat plant and spike

Figures from www.wheat-training.com; "Introduction to wheat"

1.3.1.2 **Reproductive phase**

The start of the reproductive phase is marked by floral/spikelet initiation. This stage is referred to as 'double ridge' and is the point at which the shoot apex transitions to produce floral structures, known as spikelets (Figure 1.4b, Figure 1.5). A single wheat spikelet includes two outer structures called glumes (Figure 1.5). Within the glumes there are several structures known as florets, each including structures called the palea and lemma. Each spikelet will initiate eight to-twelve floret primordia, with the two most basal florets (referred to as 1 and 2) arising on opposite sides of the spikelet meristem at roughly the same time. The third floret (3) initiates above floret 1 and the subsequent florets will initiate alternately on either side of the spikelet meristem (Figure 1.5). Only the first four to six florets are potentially fertile and will initiate a carpel (ovary) and three stamen (which include the anthers), which will develop between the palea and lemma during floret maturation. Spikelet initiation continues until the terminal spikelet stage, after which point no more spikelets will be initiated. Usually a wheat spike will produce around 20 spikelets, each with multiple florets that have the potential to hold grain. During the period from terminal spikelet stage to anthesis (flowering) the spike will experience a period of rapid growth, concurrent with a period of stem growth and elongation. In this period, some of the developing spikes and spikelets may abort (Kirby *et al.*, 1987).

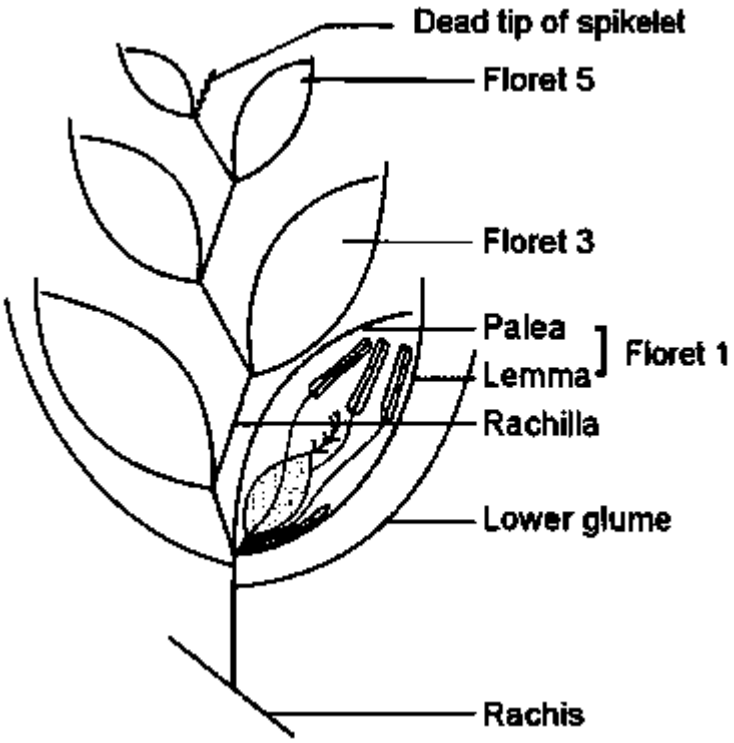


Figure 1.5: Wheat spikelet structure

Source Cereal Development Guide, Kirby *et al.* (1987)

The developing spike will remain enclosed in the developing leaf sheath (Figure 1.4) until the late stages of spike development. During stem elongation, the spike will be moved up the leaf sheath until heading, when the spike has fully emerged from the leaf sheath. Anthesis occurs a few days after heading, when pollen is released from the anthers and pollinates the carpel (Kirby *et al.*, 1987). In most cases, carpels will be pollinated with pollen from anthers in the same floret (self-pollination) and out-crossing is relatively rare (< 1 %) in wheat.

It is important to note that not all spikes and spikelets initiate and develop at the same time. The time between the initiation of the first and last spikelet can last several days or weeks, however the primordia grow and develop at different rates meaning that anthesis will occur within the space of a few days across a single spike. Within a single spike, spikelet differentiation and development begins in the middle of the spike and continues towards the top and bottom of the spike. Within a single spikelet, the floret development begins from the bottom (floret 1) and proceeds upwards (Bonnett, 1936). Similarly, anthesis first occurs in the spikelets in the middle of the spike and spreads towards the top and bottom tips. This has important consequences for sampling strategies when working with individual grains, as grains from different parts of the spike will be offset in their developmental stage. This is why we sampled grains from specific spikelet and floret positions in Chapter 2.

1.3.1.3 Grain development

Grain development begins with a "double fertilisation" event. This gives rise to the triploid endosperm nucleus (a single pollen nucleus fused with two polar nuclei in the embryo sac) and the diploid zygote embryo (the second pollen nucleus fused with the egg nucleus), which are surrounded by several tissues of maternal origin that originate from the ovary wall (Figure 1.6) (Shewry *et al.*, 2012).

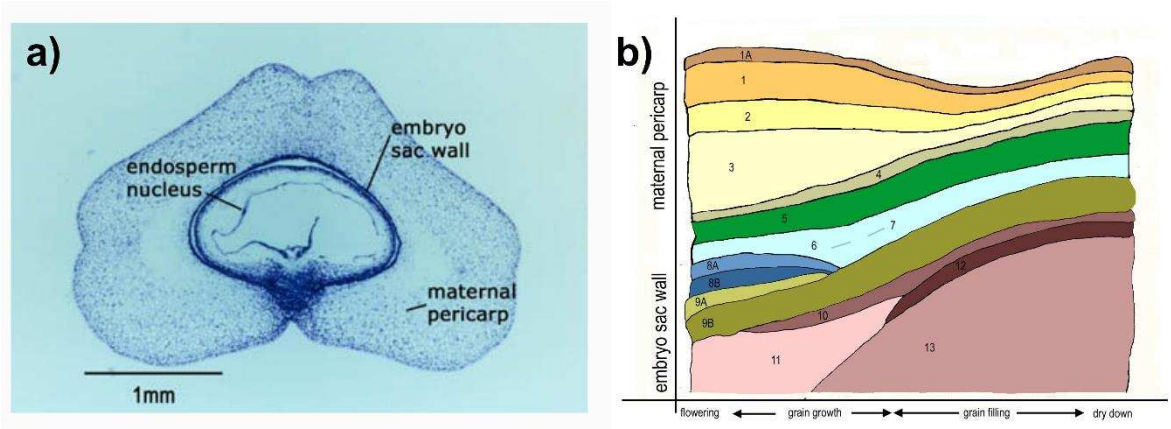


Figure 1.6: Early grain development

a) grain at three days post anthesis. b) cell layers inside the grain from anthesis to maturity. 1A-cuticle of outer epidermis, 1-outer epidermis, 2-hypodermis, 3-parenchyma, 4-intermediate cells, 5-cross cells, 6/7-inner epidermis/tube cells, 8A & 8B-outer integument, 9A & 9B-inner integument, 10-nucellar epidermis, 11-nucelllus, 12-aleurone, 13-starchy endosperm. Source: http://bio-gromit.bio.bris.ac.uk/cerealgenomics/cgi-bin/grain3.pl

In the first three to five days following fertilisation (days post anthesis; dpa), the endosperm nuclei undergo several rounds of mitosis in the absence of cell wall formation or cytokinesis to form the endosperm coenocyte (Olsen, 2001; Drea *et al.*, 2005), a multinucleate cell with a large vacuole (Figure 1.6). Cellularisation of the peripheral endosperm begins by six dpa and the central region previously occupied by the vacuole will contain nuclei at this stage (Drea *et al.*, 2005). From this point onwards the endosperm undergoes a period of rapid expansion, attributable to both cell division and expansion. Concurrent with the period of cellularisation, the endosperm also undergoes differentiation into four main cell types: starchy endosperm, aleurone, transfer cells and embryo surrounding region. The presence of four different cell types is one major difference between the endosperm of cereal grains and some dicots, including Arabidopsis, which only retain one major cell type in the endosperm of mature seeds (Olsen, 2001) .

In the first few days after fertilisation, the grain increases in size relative to the ovary but the shape remains similar ("a blunt inverted cone"; Drea *et al.*, 2005). The developing grain then lengthens significantly and reaches its maximum length at around 15 dpa (Rogers & Quatrano, 1983), by which time the basic structure of the grain has been established (Figure 1.7). This marks the

beginning of the grain filling period, which is most active in the UK between 14 and 28 days. During this time the dry weight of the grain roughly doubles through the accumulation of storage components (Shewry *et al.*, 2012), including starch and storage proteins, and the grain volume continues to increase but not in the longitudinal direction. The grain reaches physiological maturity at the end of the grain filling period, a stage which is characterised by maximum dry weight and approximately 40 % moisture content. The grain then undergoes desiccation for a period of 7-14 days until it reaches harvest ripeness (approximately 20 % moisture content).

Figure 1.7: Grain development between anthesis and 26 days post anthesis (dpa)

Example grains sampled in the 2014 grain development time course described in Chapter 2.

The mature seed therefore consists of the endosperm and the embryo, which are surrounded by the maternal outer layers (Figure 1.6b, Figure 1.8). These outer layers can be referred to broadly as the seed coat and the pericarp, although they are complex structures consisting of several different layers of cells (Figure 1.8). It has been shown in Arabidopsis that the seed coat plays several important roles in seed development and in cereal grains the pericarp takes on many of the key functions of the seed coat (reviewed in Radchuk & Borisjuk, 2014). However, despite their importance, the development of the outer layers of the grain have been much less intensively studied in wheat than the endosperm. The growth and development of the outer layers must happen in close coordination with the endosperm in order to accommodate the period of rapid growth and expansion. Studies in barley have shown that cell division in the pericarp reduces shortly after fertilisation, by around two dpa (Radchuk *et al.*, 2011). Pericarp growth subsequently continues through cell expansion, predominantly in the longitudinal direction. Programmed cell death (PCD) also occurs in the maternal seed tissue in coordination with endosperm development, thought to contribute both additional space and nutrients to the growing endosperm (Radchuk *et al.*, 2011; Radchuk *et al.*, 2017).

Whilst most phases of grain development have been extensively characterised phenotypically, the genetic and molecular basis of how these processes are controlled is not well understood in wheat.

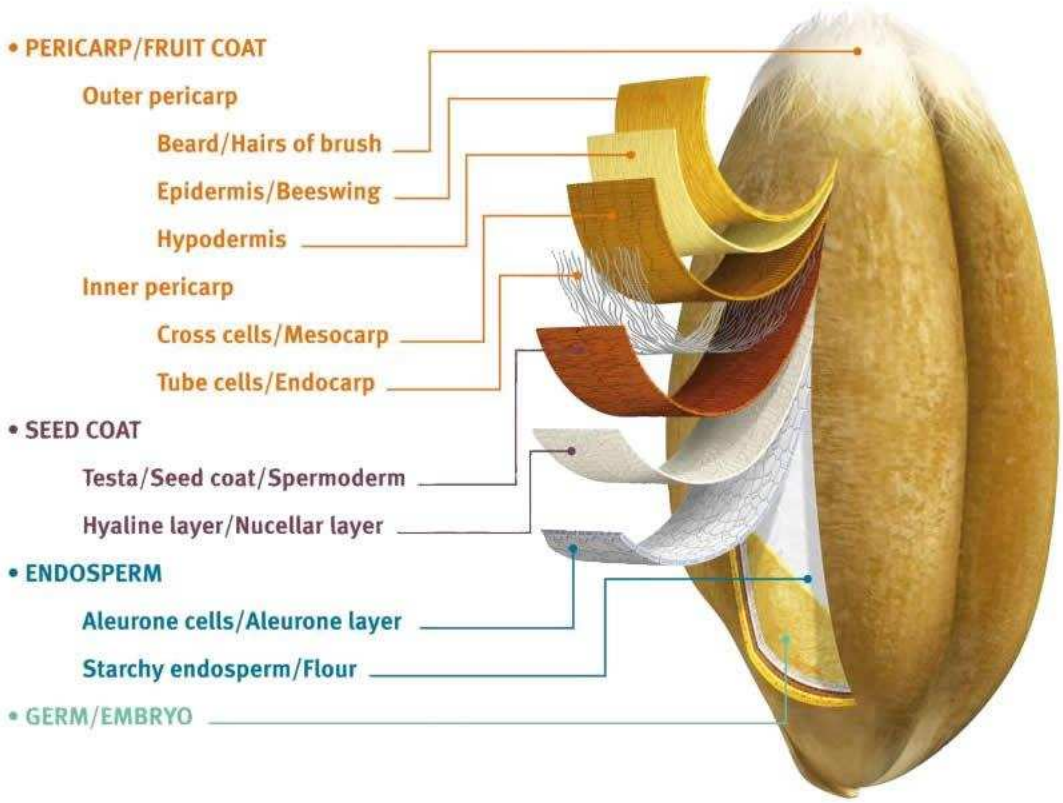


Figure 1.8: Diagram of a mature grain

Source: www.nabim.org.uk/wheat-structure

1.3.2 Factors affecting final grain yield

Grain yield is the ultimate result of plant growth and therefore essentially all genes will contribute towards yield either directly or indirectly. Genes affecting final yield can be separated into two main categories: yield adaptation and yield potential. Adaptation includes genes that determine the ability of the plant to perform well in a given environment, such as resistance to abiotic and biotic stresses or phenological adaptations e.g. flowering time (Slafer, 2003). Many of the large advances in yield gains in the past have come from phenological adaptations such as flowering time and height adaptation to maximise agronomic inputs. For example, the introduction of the *Rht* dwarfing genes allowed increased inorganic fertiliser application by reducing the susceptibility of a larger canopy to lodging (Hawkesford, 2014). Yield potential is concerned with genes that more directly control the productivity of the plant by affecting individual yield components (Slafer, 2003).

To facilitate its study, final grain yield can be broken down into three main yield components: spikes per plant (tiller number), grains per spike and individual grain weight (Figure 1.3). Spikes per plant and grains per spike together determine overall grain number. All three yield components interact and the periods in which they are determined during plant development partially overlap (Figure 1.3). This presents a challenge when trying to manipulate final yield as modification of a single component can result in negative pleiotropic effects on another component. For example,

increasing the number of spikes per plant could result in a decrease in the number of grains produced by each spike due to competition for resources, hence achieving no overall yield benefit. Indeed, negative correlations between grain number and grain weight are often observed (Kuchel *et al.*, 2007), although the two components are genetically separable (Griffiths *et al.*, 2015). Of course, this breakdown considers final grain yield based on a single isolated plant, whereas plants in the field constitute a canopy in which individual plants are in competition. This further complicates the determination of final grain yield in individual plants and spikes.

The number of spikes per plant is determined relatively early during plant development, followed by the number of grains per spike. The maximum number of potential grains per spike is ultimately determined by the number of spikelet primordia initiated (between double ridge and terminal spikelet) and the number of floret primordia initiated within each of the spikelets. Extending the time between double ridge and terminal spikelet can increase the number of spikelet primordia and hence the potential grain number (Serrago *et al.*, 2008; Gonzalez-Navarro *et al.*, 2016). The final grain number is determined by how many of the spikelets and florets retain fertility and undergo successful pollination. Maximum grain number is therefore fixed shortly after anthesis. Final grain weight is the last of the three components to be fixed during plant development and can be influenced until the grains reach physiological maturity. It has therefore been proposed that manipulating final grain weight may provide a cleaner route to increasing final grain yield in wheat. Indeed, final grain weight (measured as thousand grain weight; TGW) is more stably inherited than yield itself (Kuchel *et al.*, 2007).

1.4 Genetic control of grain weight

TGW is largely defined by the size of individual grains and can be broken down further into the morphometric components grain length, width, height and area, which are under independent genetic control (Gegas *et al.*, 2010). These grain size parameters are mainly controlled by the coordination of cell proliferation and expansion processes.

In rice, over 400 grain weight quantitative trait loci (QTL) have been identified, and several of the underlying genes have been cloned (reviewed in Xing & Zhang, 2010; Huang *et al.*, 2013). Studies in the model species, Arabidopsis, have also provided a deep molecular insight into the control of seed size (reviewed in Li & Li, 2015; Li & Li, 2016). These studies and others have revealed that seed/grain size is controlled by genes with a diverse range of molecular functions, some examples of which are described below.

Transcription factors (TFs) belonging to many different families have been shown to be involved in the control of seed/grain size, for example, the rice *SQUAMOSA PROMOTER-BINDING LIKE* (*SPL*) TF, *OsSPL16*. *OsSPL16* was cloned as the gene underlying the rice *GRAIN WIDTH 8* (*GW8*) QTL and positively regulates grain size through the promotion of cell proliferation (Wang *et al.*, 2012). Similarly, the Arabidopsis TF, *AINTEGUMENTA* (*ANT*) also promotes cell proliferation, acting as a positive regulator of seed size (Mizukami & Fischer, 2000). TFs that act to regulate

seed/grain size through the regulation of cell expansion have also been identified. *APETALA2* (*AP2*) and the WRKY TF, *TRANSPARENT TESTA GLABRA2 (TTG2)*, both act as negative regulators of seed size by limiting cell expansion in the integument in Arabidopsis (Johnson *et al.*, 2002; Garcia *et al.*, 2005; Ohto *et al.*, 2005).

Genes involved in the ubiquitin pathway are also important regulators of seed/grain size in many plant species (reviewed in Li & Li, 2014). This pathway acts to modify target proteins by the addition of a small protein called ubiquitin (Ub) through the sequential action of three enzymes: E1 (Ub activase), E2 (Ub conjugase) and E3 (Ub ligase). This modification has important regulatory functions in many cellular processes in plants and often involves the modified protein being targeted for degradation by the 26S proteasome (Hershko & Ciechanover, 1998). For example, *GW2*, a RING-type E3 Ub ligase, was cloned as the gene underlying a major rice grain weight QTL and negatively regulates grain width by limiting cell division (Song *et al.*, 2007). Orthologues of *GW2* in other species including Arabidopsis, wheat and maize also negatively regulate seed/grain weight (Li *et al.*, 2010; Xia *et al.*, 2013; Simmonds *et al.*, 2016) suggesting that this mechanism may be conserved across species. Downstream targets of the Arabidopsis *GW2* orthologue, *DA2*, have been identified that also regulate seed size, such as *DA1* and *UBIQUITIN SPECIFIC PROTEASE 15* (*UBP15*). *DA1* and *UBP15* interact genetically and physically and both regulate cell proliferation in the integument, however, *DA1* acts as a negative regulator whilst *UBP15* is a positive regulator (Liu *et al.*, 2008; Du *et al.*, 2014). *UBP15* is actually a deubiquitinating enzyme and other genes with deubiquitination activity have also been identified as regulators of grain size, such as *WIDE AND THICK GRAIN 1* (*WTG1*), which regulates grain size and shape in rice mainly through cell expansion (Huang *et al.*, 2017).

Components of several different signalling pathways have also been shown to play roles in the control of seed/grain size. Several studies have demonstrated roles for components of the G-protein signalling pathway, in which heterotrimeric G-protein complexes act with membrane bound G-protein coupled receptors to transduce extracellular signals to intracellular components (Trusov & Botella, 2016). Heterotrimeric G-protein complexes consist of three subunits: G_{α}, G_{β} and G_{γ} and roles in seed/grain size regulation have been identified for examples of all subunits in rice and Arabidopsis (reviewed in Botella, 2012). However, it is not clear if function is completely conserved across species. For example, an Arabidopsis G_{γ} subunit, *AGG3*, positively regulates seed size (Fang *et al.*, 2012), whilst the most similar rice G_{γ} subunits, *DEP1* and *GS3* appear to be negative regulators of seed size (Fan *et al.*, 2006; Huang *et al.*, 2009). Phytohormone signalling is also important in the control of seed/grain size with roles being demonstrated for auxin, brassinosteroid and cytokinin biosynthesis and signalling components (Riefler *et al.*, 2006; Schruff *et al.*, 2006; Jiang *et al.*, 2013). Other important signalling components have also been identified. For example *KLUH*, an Arabidopsis cytochrome P450, positively regulates seed size through promoting cell proliferation in the integuments (Adamski *et al.*, 2009) and this function appears to

be conserved in wheat (Ma *et al.*, 2016). Genes affecting epigenetic status have also been shown to have important roles in the control of seed/grain size (Xiao *et al.*, 2006).

Many of the components described above have been shown to act maternally to affect the final seed/grain size (reviewed in Li & Li, 2015) and it has been proposed in several species that the maternal outer tissues (i.e. seed coat or pericarp) set an upper limit to the final size of the seed/grain by physically restricting endosperm growth (Adamski *et al.*, 2009; Hasan *et al.*, 2011; Xia *et al.*, 2013).

1.4.1 Understanding of the genetic control of grain size in wheat

Despite the advances in Arabidopsis and rice, our understanding of the genetic mechanisms controlling grain size remains limited in wheat. Comparative genomics approaches and association studies have provided some insight (Breseghello & Sorrells, 2006; Sukumaran *et al.*, 2015; Ma *et al.*, 2016; Simmonds *et al.*, 2016; Arora *et al.*, 2017) and QTL associated with grain size and shape components (grain area, length and width) have been identified on almost every wheat chromosome (Breseghello & Sorrells, 2007; Gegas *et al.*, 2010; Simmonds *et al.*, 2014; Farré *et al.*, 2016; Kumar *et al.*, 2016; Brinton *et al.*, 2017). However, few of these QTL have been validated, none have been cloned and little is understood about the underlying mechanisms.

One of the major challenges to cloning grain size QTL in wheat and understanding the underlying mechanisms is the subtle nature of the effects compared to QTL in diploid species such as rice. Grain weight QTL in rice often have effects of > 20%, whilst grain size QTL in wheat usually have effects of ~ 5 % (Uauy, 2017) It has been proposed that the subtlety of these effects in wheat is due to functional redundancy between homoeologues resulting in the effects of variation in a single gene being masked by the effects of the remaining functional copies. Indeed, variation in the *GW2* gene in rice leads to grain weight differences of over 50% whereas a similar mutant in a single wheat homoeologue affects TGW by only 7 % in wheat (Song *et al.*, 2007; Simmonds *et al.*, 2016).

1.5 The 5A and 6A QTL for grain weight

Previously in the lab, two distinct major wheat grain weight QTL were identified on chromosomes 5A and 6A (henceforth referred to as the 5A QTL and 6A QTL, respectively; Simmonds *et al.*, 2014; Brinton *et al.*, 2017). Both QTL were identified in doubled haploid (DH) populations between UK hexaploid winter wheat cultivars and validated using near isogenic lines (NILs).

1.5.1 Identification of the 5A QTL

The 5A QTL was identified in a DH population developed between the UK cultivars ‘Charger’ and ‘Badger’ (CxB). The CxB DH population was evaluated for final yield and TGW across twelve environments: at least two years (yr) at five different locations (2 x England (3 yr), 1 x Scotland (2yr), 1 x France (2 yr) and 1 x Germany (2 yr)). A QTL analysis identified a region on chromosome 5A that was consistently associated with TGW, significant in seven out of twelve environments (based on the log-of-odds (LOD) score) and explaining 15.5 % of the phenotypic

variation. The QTL interval was confirmed in a multi-trait multi-environment analysis (MTME), with at least one marker in the QTL region being significantly associated with TGW in all twelve environments. In the CxB DH population, the 5A QTL increased TGW by 5.5 % with Badger providing the increasing allele.

Overall there was a significant correlation between TGW and final grain yield across all environments, but a yield QTL only co-located with the 5A QTL in two of the twelve environments in the QTL analysis. However, MTME analysis for yield showed a significant association between yield and at least one marker in the 5A QTL interval in seven out of twelve environments. It was concluded that the 5A QTL interval is associated with a consistent effect on TGW that often, but not always, translates to an increase in final grain yield. (Brinton *et al.*, 2017).

1.5.2 Identification of the 6A QTL

The 6A QTL was identified in a DH population between the UK cultivars 'Spark' and 'Rialto' (SxR) and was evaluated in the same twelve environments detailed above for the CxB DH population. A QTL analysis identified several TGW QTL in the SxR DH population present in at least five environments, but the TGW QTL on chromosome 6A co-located with a QTL for final grain yield. Across environments, there was a significant correlation between TGW and final grain yield in the SxR DH lines. MTME anlaysis found that markers within the 6A QTL interval were significantly associated with TGW and yield in ten and nine out of the twelve environments, respectively. In the SxR DH population, the 6A QTL increased TGW by 4.5 % and final yield by 3.8% with Rialto providing the increasing allele in both cases (Simmonds *et al.*, 2014).

Interestingly, *TaGW2_A*, the A genome wheat orthologue of *GW2* (rice E3 Ub ligase, described above), was located within the 6A QTL mapping interval (Simmonds *et al.*, 2014).

1.5.2.1 *TaGW2_A* as a potential candidate gene underlying the 6A QTL

At the beginning of my PhD, several studies had investigated the role of *TaGW2_A* in the control of grain size in wheat but contradictory results had been reported.

Multiple association studies had identified an A/G promoter SNP at the -593 bp position of *TaGW2_A* as associated with grain weight but had reached conflicting conclusions. One study found an association between the A allele and increased grain weight (Su *et al.*, 2011), whilst another identified the G allele as increasing grain weight (Zhang, X *et al.*, 2013). Contradictory results had also been produced as to whether the function of rice *GW2* as a negative regulator of grain weight is conserved in wheat. A natural missense mutation in exon 8 of *TaGW2_A* (Yang *et al.*, 2012) and downregulation of *TaGW2* expression by RNAi (Hong *et al.*, 2014) were both associated with an increase in grain weight, suggesting that *TaGW2_A* functions as a negative regulator of grain size in wheat. However, a separate RNAi study found that suppression of *TaGW2* expression resulted in smaller grains, suggesting positive regulation of grain size (Bednarek *et al.*, 2012). Therefore, although the evidence strongly suggested that *TaGW2_A* plays a role in the control of grain size, the precise function of the gene was not clear. Numerous studies had

identified grain weight QTL on wheat chromosome 6A and alluded to *TaGW2_A* as the possible causal gene (Mir *et al.*, 2012; Zhang, K *et al.*, 2013; Williams & Sorrells, 2014) but none had conclusively shown whether or not this was the case.

Although the parents of the SxR DH population, Spark and Rialto, do not have any coding region polymorphisms in *TaGW2_A*, they do carry the A/G -593 bp promoter SNP (Spark-A, Rialto-G; Simmonds *et al.*, 2014). Given the association of *TaGW2_A* with final grain size, its location within the 6A QTL mapping interval and the presence of the promoter SNP, we hypothesised that *TaGW2_A* could be a candidate for the causal gene underlying the 6A QTL.

1.6 Thesis aims

The overall aim of this thesis is to understand the mechanisms that control grain length and width in hexaploid wheat through the characterisation of the 5A and 6A QTL. Specifically, this thesis will combine phenotypic characterisation (Chapter 2), genetic mapping (Chapter 3) and transcriptomics (Chapter 4) to answer the following questions:

- Do the 5A and 6A QTL increase grain weight via the same or different mechanisms?
- What are the genes/pathways underlying the 5A and 6A QTL?
- Is *TaGW2_A* the gene underlying the 6A QTL?

2 Characterisation of 6A and 5A Near Isogenic Lines

All results described here regarding the 5A QTL and NILs have been published in the following manuscript (Appendix 1):

Brinton, J., Simmonds, J., Minter, F., Leverington-Waite, M., Snape, J., Uauy, C. 2017. Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat. New Phytologist, 215: 1026–1038. doi:10.1111/nph.14624

Additionally, the 2015 results regarding the *TaGW2_A* NILs have been published as part of the following manuscript:

Simmonds, J., Scott, P., Brinton, J., Mestre, T., Bush, M., del Blanco, A., Dubcovsky, J., Uauy, C. 2016. A splice acceptor site mutation in *TaGW2_A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. Theoretical and Applied Genetics, 129: 1099. https://doi.org/10.1007/s00122-016-2686-2

2.1 Chapter summary

In this chapter, a detailed characterisation of 6A, 5A and *TaGW2_A* NILs was conducted across multiple years of field trials to provide phenotypic insight into the mechanisms underlying the 6A and 5A QTL. We found that 6A NILs had a consistent difference in TGW (4.4 %) and that this was driven by an increase in final grain width (2.3 %). No differences in final grain length were observed between 6A NILs. The first differences in carpel/grain size parameters between 6A NILs were observed at the very early stages of carpel/grain development, possibly before fertilisation. *TaGW2_A* NILs showed a similar but distinct phenotype compared with 6A NILs suggesting that *TaGW2_A* may not be the causal gene underlying the 6A QTL. *TaGW2_A* NILs also had a consistent increase in TGW (6.7 %), but this was associated with increases in both final grain length (1.7 %) and width (1.9 %), which were established before fertilisation. The 5A QTL also had a robust effect on TGW (6.9 %) but this was driven by a primary effect on grain length (4.0 %) (established c. 12 dpa) and a pleiotropic late stage effect on grain width (1.5 %). This shows that the 6A and 5A QTL act to increase grain weight through distinct mechanisms. We also showed that the difference in grain length in 5A NILs is associated with longer cells in the pericarp suggesting that the 5A QTL acts to influence cell expansion.

2.2 Introduction

Many grain weight QTL have been identified in wheat (Breseghello & Sorrells, 2007; Simmonds *et al.*, 2014; Williams & Sorrells, 2014; Farré *et al.*, 2016), but very few have been validated and little mechanistic insight has been provided. Grain weight, like final yield, is a complex, polygenic trait and is largely defined by the size of the grain, which can be broken down into individual grain morphometric parameters (grain area, length and width). Studies in other species such as rice and Arabidopsis have shown that these parameters are defined during carpel/grain development through the coordination of cell expansion and proliferation processes by a diverse range of genes and

mechanisms (described in more detail in Chapter 1 and reviewed in Huang *et al.*, 2013; Li & Li, 2016). This chapter aims to provide insight into the mechanisms underlying the differences in grain weight associated with the 5A and 6A QTL.

In order to understand the contributions of specific genetic loci to complex polygenic traits, like grain weight, each locus must be separated and tested independently. This is particularly important when studying loci that have very subtle effects, such as those of the 5A and 6A QTL, which increase TGW by 5.5 % and 4.5 %, respectively. Both the 5A and 6A QTL were identified in DH populations (CxB and SxR, respectively). However, DH lines segregating for these QTL also segregated for other QTL and therefore are not suitable for specifically investigating the effects of the 5A and 6A QTL. In the same way, although the parental cultivars have differences in grain weight they also have sequence variation across the entire genome and therefore cannot be used to understand the effects of specific loci.

Near isogenic lines (NILs) are lines that differ only for a small segment of the genome and therefore allow the effects of that region to be studied without the confounding effects of background genetic diversity. NILs for both the 5A and 6A QTL had been developed and were available at the start of my PhD (Simmonds *et al.*, 2014; Brinton *et al.*, 2017), providing a valuable resource for dissecting the mechanisms underlying the grain weight effects by reducing the complexity imposed by background variation. The value of this reduction in complexity is illustrated in Figure 2.1, which shows the development of the 5A NILs.

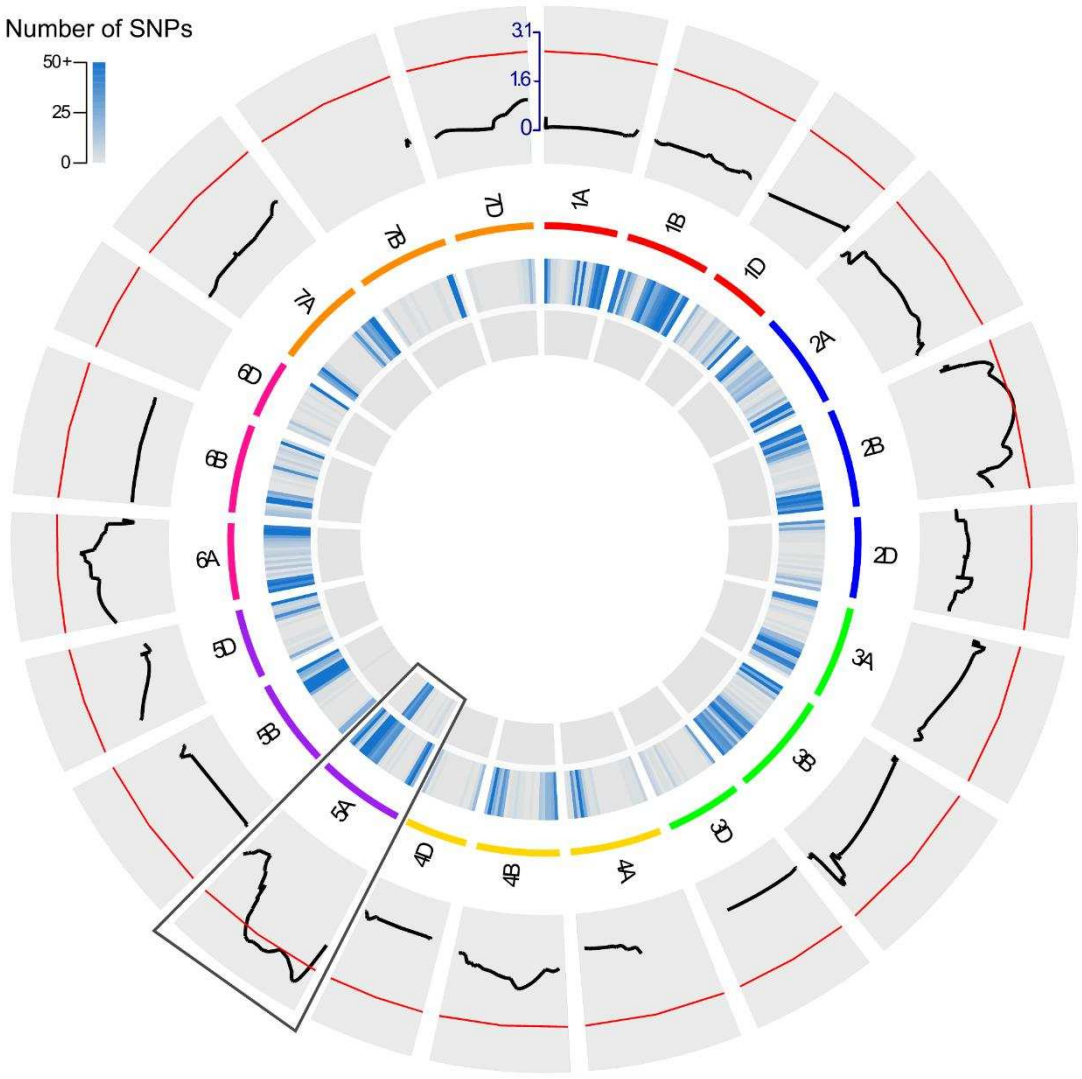


Figure 2.1: 5A QTL analysis and NIL development

Circos diagram showing the whole genome QTL scan and single nucleotide polymorphism (SNP) variation. Outer track is the mean log-of-odds (LOD) score for thousand grain weight (TGW) across all environments measured. The red line shows a LOD threshold of 2.5. This threshold was used to remain consistent with the default threshold of QTL cartographer (used for the MTME analysis discussed in 1.5.1) and was considered as appropriate for eliminating the background noise. Wheat chromosome groups are represented in different colours beneath the QTL scans. Two QTL exceeded the 2.5 LOD threshold (chromosomes 2B and 5A). The QTL on chromosome 5A (boxed segment) was selected for further analyses as it was the most significant (LOD score) and most stable across environments in addition to having a larger mean additive effect and accounting for more of the phenotypic variation than the QTL on chromosome 2B, Inner tracks correspond to heatmaps representing the number of iSelect SNPs in 30 Mb windows showing variation between Charger and Badger, parents of the doubled haploid population (outer) or a representative pair of 5A−/5A+ NILs (innermost). Physical positions of all markers (including those used in the QTL scan and iSelect markers) were determined using the IWGSC RefSeq v.1.0 sequence. Figure and legend text from Brinton *et al.* (2017).

Several grain weight QTL were identified in the CxB DH population in addition to the 5A QTL, in particular a major QTL on chromosome 2B (Figure 2.1, outer track). The parental cultivars (Charger and Badger) had sequence variation across the entire genome (7973 SNPs), whereas the NILs vary only at the location of 5A QTL (221 SNPs; 97.2% similar; Figure 2.1, inner tracks) (Brinton *et al.*, 2017). These NILs therefore eliminate the confounding effects of other genomic loci and allow specific examination of effects associated with the 5A QTL interval.

The 5A and 6A NILs were generated using the same overall strategy. Briefly, DH lines were selected that had the positive allele across the entire QTL interval (Badger (5A) or Rialto (6A)). These lines were then crossed to the recurrent (negative) parent (Charger (5A) or Spark (6A)), and heterozygous plants were backcrossed to the recurrent population for four generations (BC_4). After self-pollination, homozygous NILs with the alternative alleles across the QTL intervals were selected. From this point onwards, NILs with the positive allele (Badger or Rialto) across the interval will be referred to as 5A+ or 6A+ NILs, and those with the negative alleles (Charger or Spark) will be referred to as 5A- or 6A- NILs.

In addition to the 5A and 6A NILs, *TaGW2_A* mutant NILs were also available during this PhD. Previously in the lab, a screen of the tetraploid Kronos TILLING population identified a mutant line with a G to A transition in a splice acceptor site of *TaGW2_A* that led to mis-splicing and subsequent truncation of the *TaGW2_A* transcript (Simmonds *et al.*, 2016). Studying the effects of the mutant allele of *TaGW2_A* by comparing the mutant line to WT Kronos is confounded by the presence of background mutations, similar to the sequence variation between the parental varieties of the 5A and 6A QTL. The mutant allele *(gw2-A)* was therefore backcrossed into both tetraploid (cv Kronos) and hexaploid (cv Paragon) backgrounds to generate NILs for further characterisation (Simmonds *et al.*, 2016). The hexaploid *TaGW2_A* NILs were used during this PhD specifically to characterise the effects of *TaGW2_A* in the context of the hypothesis that it could be the gene underlying the 6A QTL. NILs with the non-functional mutant allele of *TaGW2_A* are referred to as *gw2-A* NILs and those with the wildtype (WT) allele are referred to as *GW2-A* NILs.

The aim of this chapter was to use a detailed characterisation of the 5A, 6A and *TaGW2_A* NILs to phenotypically answer the three main questions of this thesis. To this end, the effects of the QTL/mutations on final grain weight were dissected into individual grain size components. In the case of the 5A NILs the effects were broken down even further to determine whether the QTL acts to affect cell number or size. Further mechanistic insight was provided by examining the grain size and weight components of NILs during a time course of carpel/grain development. The 5A and 6A NILs were also assessed for a series of additional yield components and developmental traits to identify any pleiotropic effects of the QTL.

2.3 Materials and methods

2.3.1 Plant material and growth

The 5A, 6A and *TaGW2_A* NILs used in this chapter were generated by James Simmonds and are described in Brinton *et al.* (2017), Simmonds *et al.* (2014) and Simmonds *et al.* (2016), respectively. All NILs were evaluated at Church farm in Norwich (52.628 N, 1.171 E) with exact numbers of NILs grown in each year outlined in Table 2.1. All NILs were grown in large-scale yield plots (1.1 × 6 m) and a randomised complete block design was used with five replications.

Table 2.1: Summary of NILs grown in each year at Church farm

Year	5A NILs	6A NILs	*TaGW2_A* NILs
2012	10 BC_2	-	-
2013	10 BC_2	-	-
2014	12 BC_4	7 BC_4	-
2015	4 BC_4	4 BC_4	4 BC_2
2016	4 BC_4	4 BC_4	4 BC_4

2.3.2 Phenotyping

Grain morphometric measurements (grain width, length, area) and thousand grain weight (TGW) were recorded on the MARVIN grain analyser (GTA Sensorik GmbH, Germany) using approximately 400 grains obtained from the harvested grain samples of each plot. The plot average was used in the statistical analyses. Individual grain data from each plot sample was also extracted to examine distributions of grain size in the 5A and 6A NILs. Final grain yield was adjusted by plot size and moisture content. Other spike yield components and developmental traits measured include:

- Spikelet number* (all spikelets on the spike)
- Viable spikelets* (all spikelets containing grains)
- Grain number per spike* (Total grains from a single spike)
- Seeds per spikelet* (Total grains per spike/number of viable spikelets)
- Spike yield* (Total weight of all seeds per spike)
- Days to heading (days from sowing until 75% ear emergence of 75% of plot)
- Days to maturity (days from sowing until 75% plot senesced)
- Tiller number (measured as tillers (i.e. stems) m^{-2} after plots had been combine harvested)
- Crop height (measured at maturity)

*indicates measurements that were obtained from ten representative single ear samples (SES) taken from the field plots just before plots were combine harvested i.e. mature, dry spikes.

2.3.3 Carpel/grain developmental time courses

For the 5A and 6A carpel/grain developmental time courses, BC_4 NILs grown in 2014-2016 were used. For both QTL, two independent NILs carrying the negative allele (2 x 5A- or 2 x 6A-) and two independent NILs carrying the positive QTL allele (2 x 5A+ or 2 x 6A+) were used. The same NILs were used in all three years. For the *TaGW2_A* time courses, NILs grown in 2015 (BC_2) and 2016 (BC_4) were used. Again, two independent NILs were used for each genotype: two NILs carrying the WT allele of *TaGW2_A* (2x *GW2-A*) and two NILs carrying the non-functional A genome allele (2x *gw2-A*). In all experiments, 65 wheat inflorescences (referred to as ear or spike) per NIL were tagged across up to five blocks in the field at full ear emergence (peduncle just visible; Figure 2.2a) to ensure sampling at the same developmental stage. Ten spikes per NIL, per block, were sampled at each time point (i.e. 50 total spikes from the 65 tagged spikes). Exact time points taken are detailed in figure legends of the time courses (Figure 2.5, Figure 2.6, Figure 2.9). Spikes were kept on ice and taken to JIC for dissection. Ten carpels/grains were sampled from each spike from the outer florets (positions F1 and F2; Figure 2.2b) of spikelets located in the middle of the spike (Figure 2.2a) and placed in 2 mL eppendorf tubes. No clear developmental differences between F1 and F2 were observed during the sampling. Carpels/grains were weighed to obtain fresh weight (FW) and assessed for morphometric parameters (carpel/grain area, length and width) on the MARVIN grain analyser. Measurements were taken within 3 hours of dissection from the spike and kept at 4℃ in the intervening period to avoid moisture loss. Immediate measurement of a subset of carpels/grains and then re-smeasurement after several hours at 4℃ was performed to confirm that no grain size or weight parameters were affected by storage for this amount of time. Carpels/grains were then dried at 37 ℃ to constant weight (dry weight; DW). For each block at each time point, a total of ~100 carpels/grains were sampled (10 spikes per block x 10 carpels/grains per spike) per NIL. However, for the statistical analysis the average of the ~100 carpels/grain from each NIL within each block was used as the phenotypic value as the individual grains and spikes were considered as subsamples.

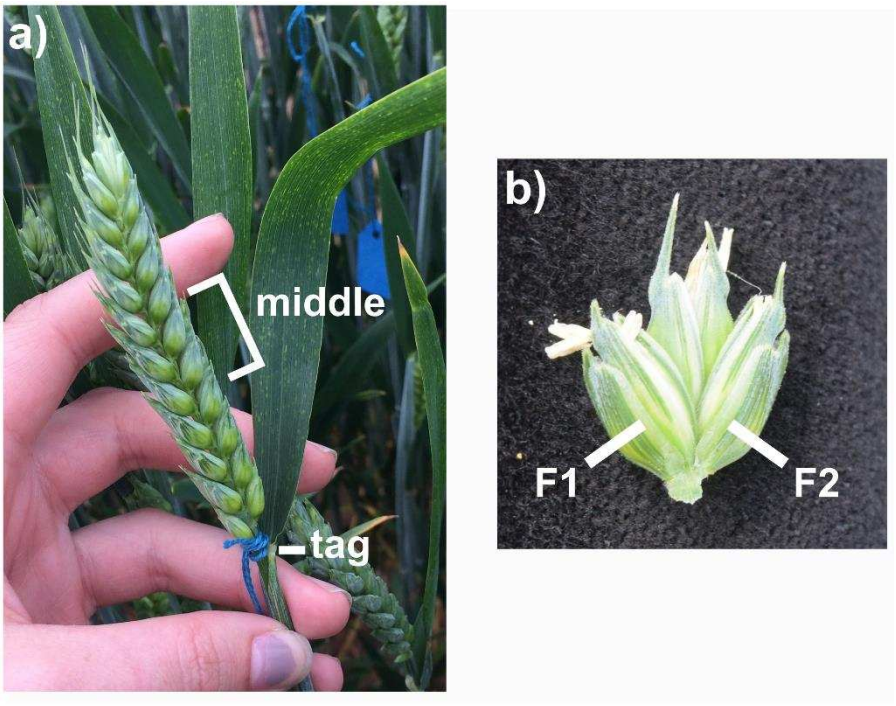


Figure 2.2: Sampling strategy for the carpel/grain development time courses

a) Spikes were tagged at full ear emergence (peduncle just visible). Grains were sampled from the middle of the spike. b) grains were sampled from the outer florets of spikelets (floret 1 (F1) and floret 2 (F2)).

2.3.4 Cell size measurements

One representative 5A- and 5A+ BC_4 NIL was used for cell size measurements. This pair of NILs were selected based on the consistency of the grain length effect across previous years. For each NIL, nine grains of average grain length were selected from the whole 2015 harvest sample from each block (groups 5A-/5A+ average). For the 5A- NIL, an additional nine grains were selected that had grain lengths equivalent to the average of the 5A+ NIL sample (5A- large). For the 5A+ NIL an additional nine grains were selected that had grain lengths equivalent to the average of the 5A- NIL sample (5A+ small). Grains of average length from three blocks of the 2016 harvest samples were also selected (nine grains from each block per genotype).

Grains were stuck crease down on to 12.5 mm diameter aluminium specimen stubs using 12 mm adhesive carbon tabs (both Agar Scientific), sputter-coated with gold using an Agar high resolution sputter coater (Figure 2.3b) and imaged using a Zeiss Supra 55 scanning election microscope (SEM). The surface (pericarp) of each grain was imaged in the top and bottom (embryo) half of the grain (Figure 2.3a, T and B, respectively), with images taken in at least three positions in each half. All images were taken at a magnification of 500x. Cell length was measured manually using the Fiji distribution of ImageJ (Schindelin *et al.*, 2012) (Figure 2.3c). Cell number was estimated for each grain using average cell length/grain length. For the statistical analyses, the average cell length of each individual grain was used.

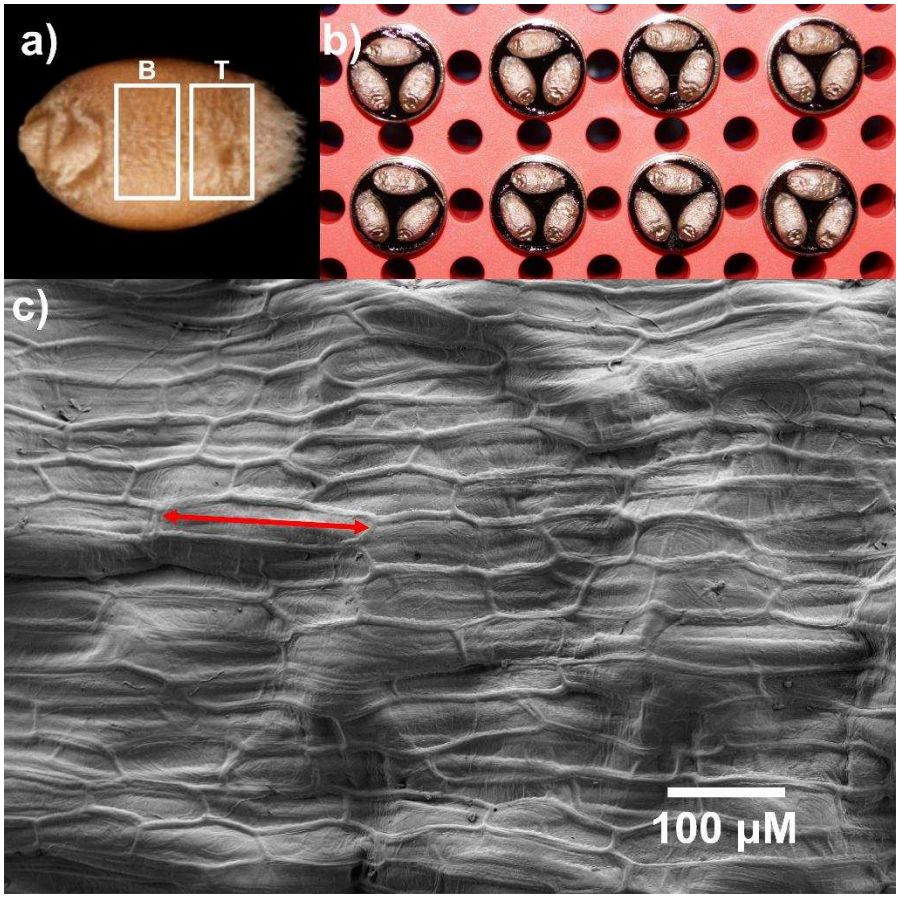


Figure 2.3: Scanning Electron Microscopy imaging for pericarp cell size measurements

a) Example grain showing the bottom (B) and top (T) half as used for imaging. b) Grains stuck crease down and sputter-coated with gold to be imaged. c) example scanning electron microscopy image taken for cell size measuring. Red arrow indicates how cell length was measured. Image taken at 500x magnification.

2.3.5 Statistical analysis

The NILs were evaluated using two-way ANOVAs across all years with the model including the interaction between environment and the genotype. For the evaluation of individual years the block and genotype were included in the model. Similarly, two-way ANOVAs, including genotype and block, were conducted for the developmental time courses and cell size measurements. Analyses were performed using R v3.2.5.

2.4 Results

2.4.1 Characterisation of the 6A QTL

2.4.1.1 6A NILs have a 4.4% difference in TGW

Across three years of replicated field trials, 6A+ NILs had significantly increased TGW compared with 6A- NILs (4.39%; P < 0.001; Table 2.2), ranging from 1.38% to 7.42% in individual years. However, when years were analysed individually, the increase in TGW was non-significant in 2016 (1.38%, P = 0.33). Across all three years the increase in TGW was associated with a 2.25%

increase in plot yield, although this was non-significant across years (P = 0.42) and in each year individually (Table 2.2).

Table 2.2: Mean Thousand Grain Weight (TGW), yield and grain morphometric parameters of 6A NILs

Year	Genotype	TGW (g)	Yield (kg/plot)	Grain area (mm^2)	Grain length (mm)	Grain width (mm)
2014	**6A-**	43.20	5.98	19.49	6.34	3.82
	6A+	45.29	6.08	20.01	6.37	3.91
		4.84%***	0.46%^{NS}	2.65%***	0.48%^{NS}	2.27%***
2015	**6A-**	38.22	6.66	15.66	5.94	3.26
	6A+	41.06	6.97	16.34	5.99	3.37
		7.42%***	4.71%^{NS}	4.37%***	0.77%^{NS}	3.35%***
2016	**6A-**	45.14	5.75	20.06	6.37	3.93
	6A+	45.77	5.76	20.29	6.33	3.99
		1.38%^{NS}	0.08%^{NS}	1.13%^{NS}	-0.62%^{NS}	1.49%*
Overall	**6A-**	42.19	6.13	18.40	6.22	3.67
	6A+	44.04	6.27	18.88	6.23	3.76
		4.39%***	2.25%^{NS}	2.58%***	0.20%^{NS}	2.31%***

%s indicate amount gained in 6A+ NILs compared with 6A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row) i.e. NS = Non-significant, * = P < 0.05, *** = P < 0.001

To identify potential pleiotropic effects of the QTL that could account for the absence of a significant yield effect, ten representative spikes from each plot of 6A NILs were assessed for a series of spike yield components (Table 2.3). Components included spikelet number, viable spikelet number, seeds per spikelet, grain number per spike and spike yield, although not all measurements were taken in all three years. In 2014, grain samples were compromised due to high levels of bunt infection at Church farm and in 2016 the spikelet number counting was performed incorrectly. Across two years, there was significant decrease in the number of viable spikelets in 6A+ NILs (-2.17%, equivalent to 0.44 spikelets per spike; P = 0.001) although this was driven by a strong effect in 2014. The decrease in viable spikelet number (and spikelet number overall) appears to have been compensated for in 2015 by an increase in the number of seeds per spikelet (3.98%) which resulted in one extra grain per spike (2.86%) in 6A+ NILs, although neither were significant (Table 2.3). Unfortunately, no grain number data is available for 2014 (when spikelet number was significantly reduced) so it is unclear whether this would have resulted in a significant reduction in grain number. Despite the fact that the grain number differences were not significant in 2015, the tendency towards more grains per spike combined with 8.7% higher TGW (P < 0.001) in the ten spike sample, resulted in significantly higher spike yield in 6A+ NILs (11.90%, P = 0.001). This translated into a higher overall plot yield (4.71%; Table 2.2), although this was not statistically significant.

Table 2.3: Spike yield components of ten representative single ear samples (SES) of 6A- and 6A+ BC_4 NILs

Year	Genotype	Spikelet number	Viable spikelets	Grain number per spike	Spike yield (g/spike)	Seeds per spikelet	SES-TGW (g)	SES-Grain Area (mm^2)	SES-Grain length (mm)	SES-Grain width (mm)
2014	**6A-**	22.12	21.00	-	-	-	-	-	-	-
	6A+	21.24	20.35	-	-	-	-	-	-	-
		-3.98%***	-3.10%**							
2015	**6A-**	21.53	19.15	33.52	1.265	1.75	37.80	19.11	6.39	3.71
	6A+	21.13	18.93	34.48	1.416	1.82	41.09	19.89	6.41	3.85
		-1.86%^{NS}	-1.15%^{NS}	2.86%^{NS}	11.90%**	3.98%^{NS}	8.71%***	4.06%**	0.31%^{MS}	3.65%**
Overall	**6A-**	21.83	20.08	-	-	-	-	-	-	-
	6A+	21.19	19.64	-	-	-	-	-	-	-
		-2.94%***	-2.17%**							

%s indicate amount gained in 6A+ NILs compared with 6A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row). i.e. NS = Non-significant, ** < 0.01, *** < 0.001. SES = Single Ear Samples. - = data not available

6A NILs were also measured in the field for a series of developmental traits. 6A+ NILs flowered c. one day earlier than 6A- NILs across years (measured as days to heading; P = 0.01; Table 2.4). 6A+ NILs also senesced c. one day later than 6A- NILs (measured as days to maturity; P = 0.006) although this was only measured in a single year (Table 2.4). No consistent significant effects were observed across years for crop height and tiller number, although there was a significant reduction in tiller number in 6A+ NILs in 2014 (Table 2.4). These results suggest that the 6A QTL acts to increase TGW in a stable manner across years, but the effects on final yield may be modulated by environmental interactions and negative effects on components such as spikelet number and tiller number.

Table 2.4: Developmental traits of 6A BC_4 NILs

Year	Genotype	Days to heading	Days to maturity	Tiller number	Crop Height (cm)
2014	**6A-**	243.80	296.00	82.10	75.70
	6A+	242.95	296.70	76.53	75.13
		-0.85**	0.70**	-5.57**	-0.58NS
2015	**6A-**	250.90	-	133.75	84.75
	6A+	250.00	-	136.10	83.75
		-0.90*		2.35NS	-1.00NS
2016	**6A-**	250.38	-	-	-
	6A+	249.89	-	-	-
		-0.49NS			
Overall	**6A-**	248.36	-	107.93	80.23
	6A+	247.61	-	106.31	79.44
		-0.75*		-0.75NS	-0.75NS

Differences indicate amount gained in 6A+ NILs compared with 6A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row) i.e. NS = Non-significant, * = P < 0.05, ** = P < 0.01. - = data not available

2.4.1.2 Grain width underlies the increase in TGW in 6A+ NILs

Grain morphometric parameters (grain area, length and width) of 6A NILs were measured to understand the contribution of the individual components to the overall increase in TGW (Table 2.2). 6A+ NILs had significantly increased grain area (P < 0.001) and grain width (P < 0.001) compared to 6A- NILs. No significant grain length differences were observed in any year. 6A+ NILs had 2.31% wider grains across all years ranging from 1.49-3.35% in individual years. Grain area differences ranged from 1.13 – 4.37 %, although the difference was non-significant in 2016, reminiscent of the non-significant TGW increase in 2016. These results were based on whole plot samples and were confirmed in ten representative ear samples taken before harvest (Table 2.3). The absence of any significant grain length effect suggests that grain width is the main factor underlying the increase in grain area and TGW in 6A+ NILs. However, the difference in grain area (2.58%; Table 2.2) did not fully account for the difference in TGW (4.39%; Table 2.2). This

discrepancy could be accounted for by an increase in grain height/thickness but this parameter was not measured, discussed further in section 2.5.3.

2.4.1.3 The 6A QTL affects grains uniformly within the spike

Distributions of grain width were compared between 6A NILs using measurements from individual seeds to determine whether the 6A QTL has a uniform effect on all grains within the spike. Violin plots of grain width showed some variation in distribution shape between years (Figure 2.4). However, distribution shapes within years were very similar between 6A- and 6A+ NILs suggesting that the QTL has a uniform and stable effect across the whole spike and within spikelets. In all years, the 6A+ distributions were shifted higher reflecting the higher average grain width and illustrating the fact that 6A+ NILs had both larger numbers of wider grains and fewer thinner grains than 6A- NILs. Note that individual distributions are not completely normally distributed since the plots are based on the multiple independent NILs used for each genotype.

Figure 2.4: Distribution of grain width of 6A NILs from whole plot samples

Violin plots showing the distribution of individual seed measurements of grain width across three field experiments of BC_4 6A near isogenic lines (NILs). Orange plots = 6A+ NILs, grey = 6A- NILs. Black lines within the boxes indicate the median value. All within year comparisons were significant (2014, 2015: $P < 0.001$; 2016: $P = 0.03$).

2.4.1.4 The 6A QTL acts during very early grain development to increase grain width

Grain developmental time courses were conducted to determine when the first differences in grain size and weight were established between 6A NILs. In 2014, at the first time point (4 dpa) the 6A+ NILs had significant increases in all components measured (grain length: 4.66%, P = 0.02; width: 4.71%, P = 0.008; area: 10.10%, P = 0.006; FW: 19.90%, P = 0.003; and DW: 21.1%, P = 0.008; Figure 2.5 and Table 2.5). Significant differences in grain width, area and FW were maintained throughout the rest of the time course until the final time point (30dpa) although the differences reduced in magnitude (grain width: 2.2%; P = 0.001, grain area: 2.1%, P = 0.01; FW: 6.1%, P = 0.02). These differences at the final time point were reflective of the measurements from mature grains (Table 2.2). The significant difference in DW was maintained until the penultimate time point (20 dpa), however by 30 dpa 6A+ NILs had 4.6% higher DW but the difference was non-significant (P = 0.16). No significant differences in grain length were observed after 4 dpa.

To determine whether differences in grain size and weight components were established before fertilisation, the first time points in 2015 and 2016 were taken at heading (full ear emergence, peduncle just visible), but before plants had reached anthesis (i.e. before fertilisation/flowering). In both years, there were increases in carpel width and FW in 6A+ NILs at heading although the differences were borderline non-significant (Width (2015: 2.66%, P = 0.06; 2016: 3.55%, P = 0.08), FW (2015: 13.52%, P = 0.09; 2016: 11.3%, P = 0.06)). Additionally, in 2016 6A+ NILs had 11.7% higher DW although this was again borderline non-significant (11.7 %, P = 0.09). In 2015 the first significant increase in any component was observed at 5 dpa when 6A+ NILs had significantly wider grains (2.93%, P = 0.038) and significantly heavier grains (DW: 11.9%, P = 0.049). At 12 dpa, despite overall increases in all components in the 6A+ NILs, none of the comparisons were significant. At the final time point (19 dpa) grain width (3.42%, P < 0.001), grain area (3.93%, P < 0.001), FW (8.05%, P <0.001) and DW (6.13%, P = 0.047) were all significantly increased in 6A+ NILs. No significant differences in grain length were observed at any time point in 2015. In 2016 however all measured components were significantly increased in 6A+ NILs at 2 dpa. Grain width, area, FW and DW remained significantly higher in 6A+ NILs throughout the time course, with the exception of DW which was only borderline significant at 19 dpa. From 5 dpa onwards there were no significant differences between 6A NILs in grain length.

Taking together the results from all three years suggests that the 6A QTL acts during very early grain development to increase grain width, area and weight. However, it remains unclear as to whether this mechanism acts pre or post-fertilisation.

Table 2.5: Differences between 6A NILs in grain size and weight parameters during carpel/grain development time courses

Year	Days Post Anthesis	Length (%)	Width (%)	Area (%)	FW (%)	DW (%)
2014	4	$4.66^{0.021}$	$4.71^{0.008}$	$10.10^{0.006}$	$19.90^{0.003}$	$21.10^{0.008}$
	10	$2.40^{0.070}$	$3.33^{0.013}$	$5.78^{0.020}$	$13.25^{0.012}$	$15.02^{0.031}$
	15	$0.89^{0.251}$	$3.22^{<0.001}$	$4.12^{0.006}$	$11.81^{0.004}$	$16.06^{0.019}$
	20	$0.57^{0.120}$	$2.87^{<0.001}$	$3.52^{<0.001}$	$8.72^{<0.001}$	$11.50^{0.003}$
	30	$-0.40^{0.240}$	$2.18^{0.001}$	$2.09^{0.015}$	$6.55^{0.020}$	$4.95^{0.157}$
2015	-5	$1.97^{0.164}$	$2.66^{0.065}$	$4.30^{0.108}$	$13.52^{0.094}$	$3.07^{0.641}$
	0	$1.08^{0.357}$	$2.28^{0.226}$	$3.29^{0.237}$	$4.82^{0.400}$	$7.81^{0.347}$
	2	$1.07^{0.738}$	$1.98^{0.461}$	$3.05^{0.555}$	$12.48^{0.150}$	$9.64^{0.177}$
	5	$2.24^{0.244}$	$2.93^{0.038}$	$5.91^{0.077}$	$9.65^{0.055}$	$11.89^{0.049}$
	12	$0.48^{0.651}$	$2.04^{0.079}$	$3.05^{0.193}$	$6.93^{0.078}$	$9.29^{0.080}$
	19	$0.27^{0.498}$	$3.42^{<0.001}$	$3.93^{<0.001}$	$8.05^{<0.001}$	$6.13^{0.047}$
2016	-3	$0.96^{0.602}$	$3.55^{0.084}$	$5.35^{0.157}$	$11.30^{0.060}$	$11.73^{0.090}$
	2	$9.04^{0.039}$	$10.15^{0.011}$	$19.09^{0.018}$	$32.96^{0.011}$	$37.56^{0.015}$
	5	$4.10^{0.123}$	$5.38^{0.021}$	$11.20^{0.040}$	$19.03^{0.017}$	$18.63^{0.038}$
	12	$1.31^{0.113}$	$4.99^{<0.001}$	$6.17^{0.002}$	$11.19^{0.002}$	$12.79^{0.008}$
	19	$0.26^{0.689}$	$3.93^{<0.001}$	$4.16^{0.006}$	$7.41^{0.009}$	$7.70^{0.073}$
	26	$0.44^{0.325}$	$4.09^{<0.001}$	$4.56^{<0.001}$	$8.68^{<0.001}$	$9.53^{0.004}$

%s indicate amount gained in 6A+ NILs compared with 6A- NILs. Superscripts are the ANOVA P-values of the comparison between 6A+ and 6A- NILs.

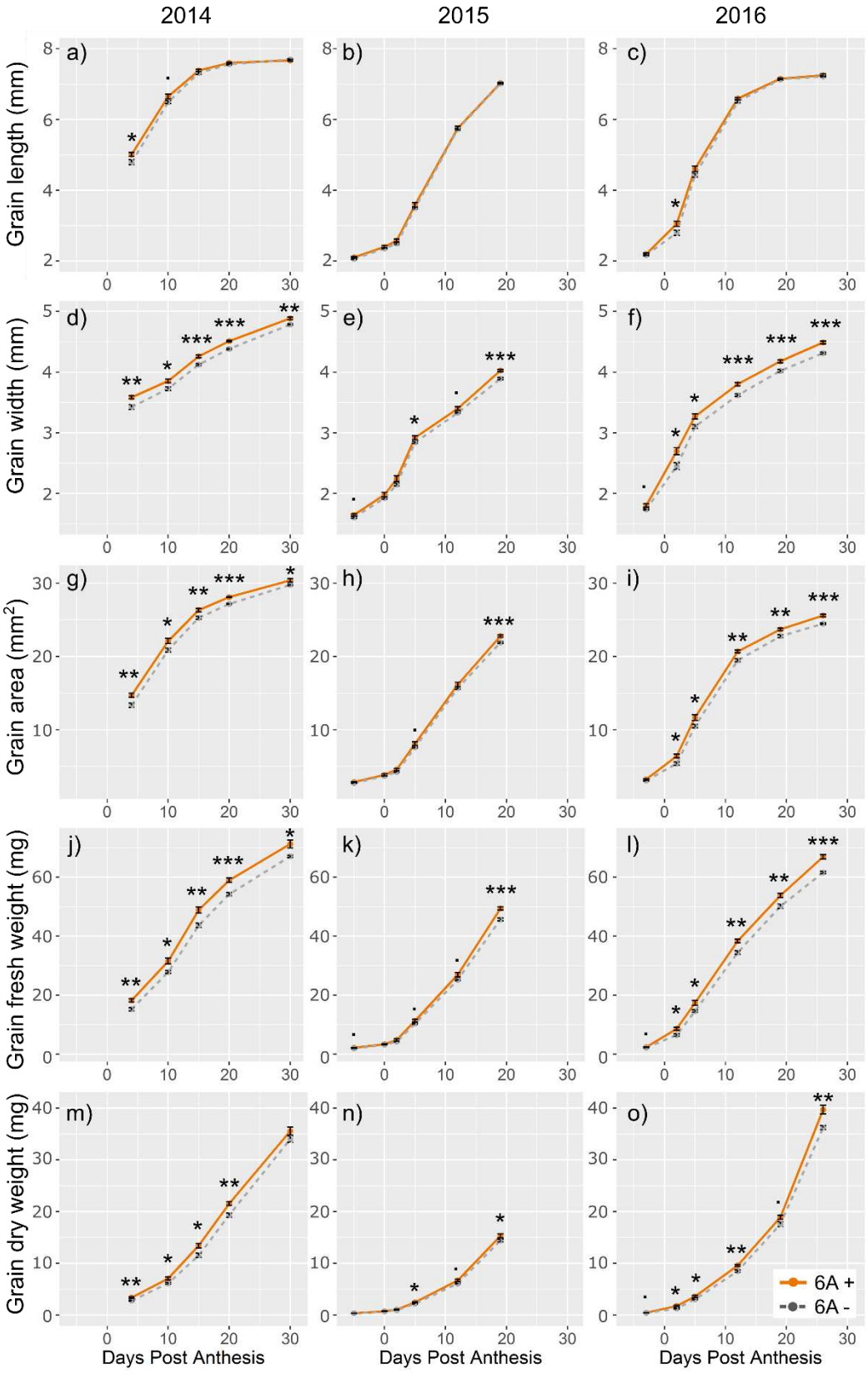


Figure 2.5: Carpel/grain development time courses of 6A NILs

Carpel/grain length (a, b, c), width (d, e, f), area (g, h, i), fresh weight (j, k, l) and dry weight (m, n o) of 6A- (grey, dashed line) and 6A+ (orange, solid line) BC_4 near isogenic lines (NILs) during carpel/grain development in 2014-2016 field trials. 2014 samples: 4, 10, 15, 20 and 30 days post anthesis (dpa)); 2015 samples: -5, 0 (anthesis), 2, 5, 12 and 19 dpa; 2016 samples: -3, 2, 5, 12, 19 and 26 dpa. . = $P < 0.10$, * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$. Error bars show one standard error above and below the mean.

2.4.1.5 *GW2-A* NILs show phenotypic differences compared to 6A NILs

The A genome copy of the RING-type E3 ubiquitin ligase *TaGW2* (*TaGW2_A)* genetically mapped to the original 6A QTL region (Simmonds *et al.*, 2014) and has previously been associated with the control of grain size (Su *et al.*, 2011; Bednarek *et al.*, 2012; Yang *et al.*, 2012; Zhang, X *et al.*, 2013; Hong *et al.*, 2014). Therefore, we hypothesised that *TaGW2_A* could be a candidate gene for the 6A QTL (as discussed in Chapter 1). To test this hypothesis phenotypically, *TaGW2_A* NILs were assessed for grain weight and morphometric parameters and carpel/grain development time courses were conducted in 2015 and 2016. The results from 2015 have been published in Simmonds *et al.* (2016).

The *gw2-A* (mutant) NILs carry a G-A mutation in the AG canonical splice acceptor site of exon 5 of *TaGW2_A*, resulting in mis-splicing of the transcript. The predominant splice variant arising from this mutation uses an alternative splice acceptor site located 4bp downstream of the wild type splice site, causing a frame-shift that generates a premature termination codon. The arising truncated *gw2-A* protein is 134 amino acids in length, compared to the 426 amino acid length wild type protein (Simmonds *et al.*, 2016). In most RING-type E3 ligases, the RING domain is located at the N-terminus and is the domain that binds the E2 conjugase whilst the rest of the protein (C-terminus) contains protein-protein interaction domains that are important for substrate recognition and binding (Stone *et al.*, 2005). The RING domain in *TaGW2_A* is predicted to be located at 61-104 amino acids therefore may not be disrupted in the *gw2-A* truncated protein, and consequently the mutant allele may retain ubiquitination activity. However, a mutation in a similar location in the rice *GW2* protein resulted in a truncated protein that retained intrinsic ubiquitination activity but still acted as a null allele due to the lack of the substrate binding domain (Song *et al.*, 2007). It is possible that the *gw2-A* mutation present in the *GW2-A* NILs used in this thesis has the same effect, but this remains to be experimentally validated.

2.4.1.5.1 gw2-A NILs have 6.7% higher TGW, driven by both grain length and width

Across two years of field trials, *gw2-A* (mutant) NILs had 6.65% higher TGW than *GW2-A* (WT) NILs ($P < 0.001$), ranging from 6.17-7.11% in each year. This was larger than the TGW differences observed between 6A NILs across years (4.4% higher TGW in 6A+ NILs across three years, 4.2 % in 2015-2016; Table 2.2). Similarly to the 6A NILS, no significant differences in yield were observed in either year.

Across years, *gw2-A* (mutant) NILs had significantly increased grain length (1.74%, $P <0.001$; Table 2.6) and grain width (1.94%, $P = 0.007$; Table 2.6), which combined to give a 3.57% ($P <0.001$) increase in grain area compared to *GW2-A* (WT) NILs. This was in contrast to 6A NILs, which showed significant differences in grain width and area, but no significant differences in grain length in each of the three years tested (Table 2.2). This would support the hypothesis that the 6A effect is distinct from *TaGW2_A*.

Given that the differences in grain length and grain width between *TaGW2_A* NILs were of a similar magnitude, these results suggest that the increase in TGW in *gw2-a* (mutant) NILs is driven by a combination of increases in both grain width and grain length.

Table 2.6: Mean Thousand Grain Weight (TGW), yield and grain morphometric parameters of TaGW2_A NILs

Year	Genotype	TGW (g)	Yield (kg/plot)	Grain area (mm^2)	Grain length (mm)	Grain width (mm)
2015	***GW2-A* (WT)**	43.869	5.195	20.275	6.698	3.699
	***gw2-A* (mut)**	46.573	5.328	20.909	6.785	3.767
		6.17%***	2.56%^{NS}	3.13%**	1.30%*	1.84%**
2016	***GW2-A* (WT)**	45.859	5.612	21.058	6.676	3.896
	***gw2-A* (mut)**	49.118	5.642	21.901	6.822	3.975
		7.11%***	0.53%^{NS}	4.00%***	2.18%***	2.03%**
Overall	***GW2-A* (WT)**	44.864	5.404	20.666	6.687	3.797
	***gw2-A* (mut)**	47.846	5.485	21.405	6.804	3.871
		6.65%***	1.51%^{NS}	3.57%***	1.74%***	1.94%**

%s indicate amount gained in *gw2-a (mutant)* NILs compared with *GW2-A* (WT) NILs. Superscripts indicate significance determined by ANOVA for either each year, or across both years (final row). ie. NS = Non-significant, * = P < 0.05, ** = P < 0.01, *** = P < 0.001. 2015 = BC_2, 2016 = BC_4.

2.4.1.5.2 TaGW2_A acts before fertilisation

Time courses of carpel/grain development were conducted in 2015 and 2016. In 2015, samples were taken at -4, 0 (anthesis), 3, 9, 16 and 23 dpa. In 2016, just three critical time points were sampled: -3, 0 and 20 dpa.

In 2015, *gw2-A* (mutant) NILs had significantly increased carpel length (5.55%), width (6.22%) and area (12.13%) at the first time point (-4 dpa, P < 0.001; Table 2.7 and Figure 2.6). These differences were maintained for the duration of the time course with the exception of carpel/grain length, which became non-significant by the final time point (P = 0.267). The differences in carpel/grain size components translated to increases in both carpel FW and DW, which were significantly increased in *gw2-A* (Mutant) NILs across the whole time course.

In 2016, significant differences were again observed in carpel length (3.2%), width (3.3%) and area (6.4%) at the first time point (-3 dpa), translating to an increase in carpel FW (8.36%). In contrast to 2015, carpel/grain length remained significantly higher in *gw2-A* (Mutant) NILs for the duration of the time course whilst differences in grain width and area were non-significant at the final time point (20 dpa). No significant differences in FW were observed after -3 dpa and no significant differences in DW were observed across the entire time course in 2016.

Despite the conflicting results of the post anthesis time points in 2015 and 2016, in both years it appears that *TaGW2_A* acts before anthesis to influence carpel length and width, which combine to modulate carpel area and weight, and ultimately final grain weight.

Table 2.7: Differences between TaGW2_A NILs during carpel/grain development time courses

Year	Days Post Anthesis	Length (%)	Width (%)	Area (%)	FW (%)	DW (%)
2015	**-4**	5.55%$^{<0.001}$	6.22%$^{<0.001}$	12.13%$^{<0.001}$	14.22%$^{0.02}$	27.99%$^{<0.001}$
	0	6.66%$^{<0.001}$	9.58%$^{<0.001}$	15.13%$^{<0.001}$	27.34%$^{<0.001}$	24.49%$^{0.002}$
	3	9.81%$^{0.004}$	9.43%$^{<0.001}$	18.95%$^{0.002}$	25.41%$^{0.005}$	20.90%$^{0.003}$
	9	4.23%$^{0.005}$	5.03%$^{<0.001}$	9.52%$^{<0.001}$	15.27%$^{<0.001}$	18.86%$^{<0.001}$
	16	1.78%$^{0.005}$	4.11%$^{0.001}$	5.61%$^{0.001}$	10.00%$^{0.002}$	9.48%$^{0.013}$
	23	1.02%$^{0.267}$	3.64%$^{<0.001}$	5.09%$^{<0.001}$	9.78%$^{<0.001}$	7.66%$^{0.003}$
2016	**-3**	3.20%$^{0.002}$	3.29%$^{0.002}$	6.43%$^{<0.001}$	8.36%$^{0.005}$	4.77%$^{0.175}$
	0	3.31%$^{0.024}$	2.56%$^{0.049}$	5.94%$^{0.026}$	7.12%$^{0.140}$	7.42%$^{0.165}$
	20	1.99%$^{<0.001}$	0.86%$^{0.458}$	2.64%$^{0.053}$	3.21%$^{0.187}$	-2.39%$^{.0412}$

%s indicate amount gained in *gw2-A* (Mutant) NILs compared with *GW2-A* (WT) NILs. Superscripts are the ANOVA P-values of the comparison between *GW2-A* (WT) and *gw2-A* (Mutant) NILs. 2015 = BC_2, 2016 = BC_4.

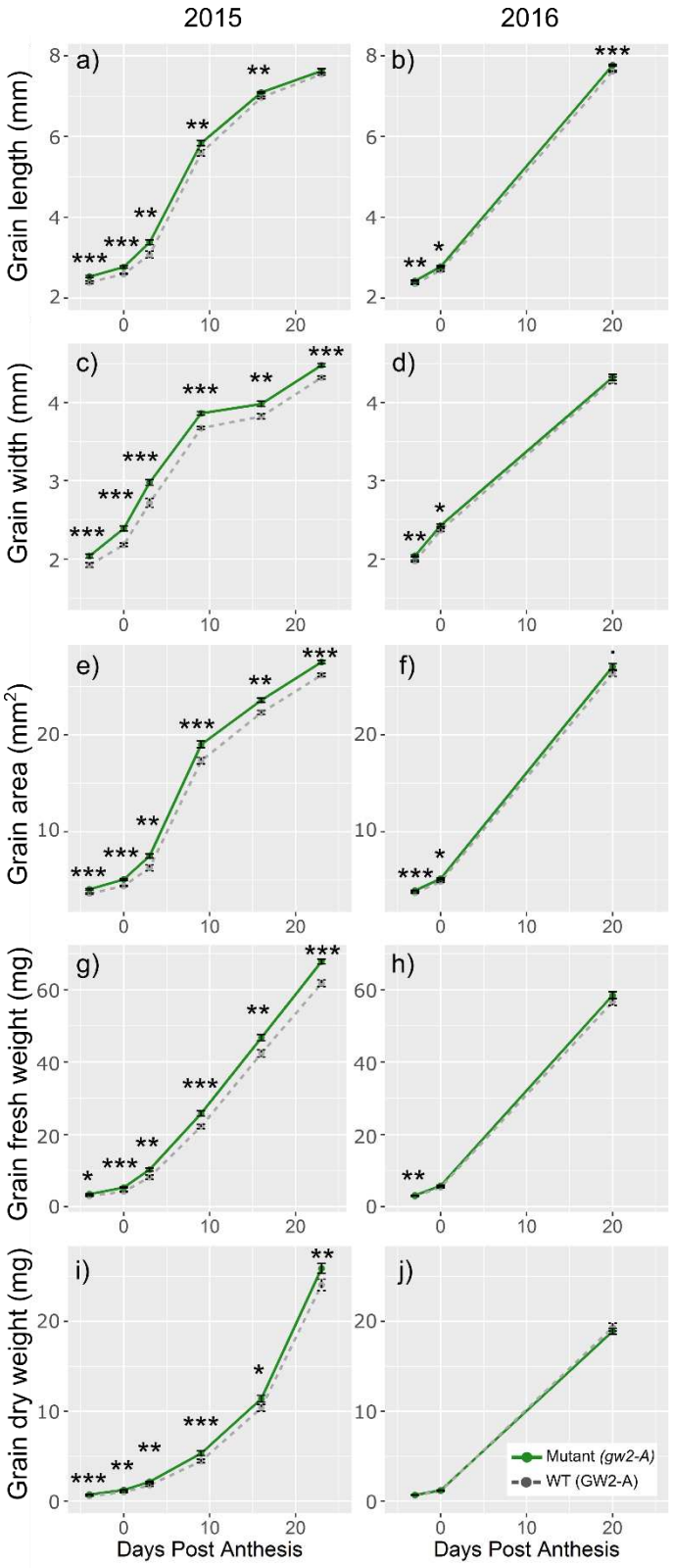


Figure 2.6: Carpel/grain development time courses of TaGW2_A NILs

Carpel/grain length (a, b), width (c, d), area (e, f), fresh weight (g, h) and dry weight (i, j) of GW2-A (WT; grey, dashed line) and gw2-a (mutant; green, solid line) BC_2 (2015) and BC_4 (2016) near isogenic lines (NILs) during carpel/grain development in 2015-2016 field trials. 2015 samples: -4, 0 (anthesis), 3, 9, 16 and 23 dpa; 2016 samples: -3, 0 and 20 dpa. * = P < 0.05, ** = P < 0.01, *** = P < 0.001. Error bars show one standard error above and below the mean.

2.4.2 Characterisation of the 5A QTL

All results described here relating to the 5A QTL have been published in Brinton *et al.* (2017). Results from the BC_2 NILs (2012-2013) were obtained by James Simmonds prior to the start of the PhD, but had not been previously published. These results were therefore analysed alongside results obtained during the PhD (BC_4 NILs, 2014-2016).

2.4.2.1 5A NILs have a 6.9% difference in TGW

Across five years of replicated field trials 5A+ NILs showed an average increase in TGW of 6.92% ($P < 0.001$) ranging from 4.00 to 9.28% (Table 2.8), and significant in all years. The difference in TGW was associated with a yield increase of 1.28% in 5A+ NILs across all years, although this effect was not significant ($P = 0.093$). The effect varied across years with a significant yield increase of 2.17% ($P = 0.046$) in 2014 and non-significant effects of between 0.02 to 1.72% in the other four years.

Table 2.8: Mean Thousand Grain Weight (TGW), yield and grain morphometric parameters of 5A NILs

Year	Genotype	TGW (g)	Yield (kg/plot)	Grain area (mm^2)	Grain length (mm)	Grain width (mm)
	5A-	38.027	4.408	18.755	6.625	3.475
2012	**5A+**	41.554	4.437	19.930	6.900	3.557
		9.28%***	0.66%^{NS}	6.26%***	4.15%***	2.35%**
	5A-	40.772	6.157	19.969	6.705	3.674
2013	**5A+**	43.544	6.159	20.979	6.963	3.727
		6.80%***	0.02%^{NS}	5.06%***	3.86%***	1.44%***
	5A-	47.368	6.495	21.493	6.798	3.930
2014	**5A+**	50.729	6.636	22.579	7.063	3.979
		7.09%***	2.17%*	5.05%***	3.90%***	1.25%**
	5A-	42.734	7.582	18.044	6.426	3.479
2015	**5A+**	46.201	7.712	19.293	6.730	3.554
		8.11%***	1.72%^{NS}	6.93%***	4.72%***	2.16%***
	5A-	49.292	5.974	19.829	6.580	3.735
2016	**5A+**	51.266	6.064	20.610	6.816	3.745
		4.00%*	1.50%^{NS}	3.94%**	3.58%***	0.27%^{NS}
	5A-	43.639	6.123	19.618	6.627	3.659
Overall	**5A+**	46.659	6.201	20.678	6.894	3.712
		6.92%***	1.28%^{NS}	5.41%***	4.04%***	1.45%***

%s indicate amount gained in 5A+ NILs compared with 5A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row). ie. NS = Non-significant, * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$. 2012-13 = BC_2 NILs, 2014-16 = BC_4 NILs.

5A NILs were measured for a series of spike yield components to determine possible pleiotropic effects associated with the 5A+ TGW effect. Within most years, there was no significant effect of the 5A+ allele on spike yield components such as spikelet number, seeds per spikelet or grain number per spike (Table 2.9). However, when all years were analysed together, there was a significant reduction in grain number (-3.55%, P = 0.04) and seeds per spikelet (-3.37%, P = 0.015) associated with the 5A+ QTL. This statistical significance was driven by a particularly strong negative effect in 2016 as grain number and seeds per spikelet were non-significant in the preceding four seasons (2012-15). Overall, however, the 5A+ QTL is associated with a consistent small decrease in these spike yield components. Taking into account the 6.92% effect of the 5A+ QTL on TGW and the tendency for decreases in some spike yield components, the overall spike yield increased by 2.33% (P = 0.032) across the five years. However, similar to grain number and seeds per spikelet, the statistical significance is driven by a single year (2014) despite overall positive effects in another three years (2012, 2013, and 2015).

Table 2.9: Spike yield components of ten representative single ear samples (SES) of 5A- and 5A+ NILs

Year	Genotype	Spikelet number	Viable Spikelets	Spike Length	Grain number per spike	Spike yield (g/spike)	Seeds per spikelet	SES-TGW (g)	SES-Grain Area (mm^2)	SES-Grain length (mm)	SES-Grain width (mm)
	5A-	24.69	22.33	9.41	64.15	2.49	2.60	38.70	19.71	6.66	3.62
2012	**5A+**	24.87	22.40	9.64	63.29	2.58	2.55	40.61	20.59	6.88	3.66
		0.73%^{NS}	0.36%^{NS}	2.43%^{NS}	-1.34%^{NS}	3.83%^{NS}	-1.82%^{NS}	4.92%*	4.46%***	3.32%***	1.26%^{NS}
	5A-	21.93	20.79	9.50	65.07	2.84	2.97	43.65	20.35	6.66	3.78
2013	**5A+**	22.00	20.64	9.63	62.72	2.90	2.85	46.33	21.33	6.92	3.83
		0.30%^{NS}	-0.72%^{NS}	1.32%^{NS}	-3.60%^{NS}	2.16%^{NS}	-3.92%^{NS}	6.14%***	4.84%***	3.88%***	1.09%**
	5A-	21.54	20.42	-	84.06	4.11	3.90	48.90	21.55	6.74	3.94
2014	**5A+**	21.59	20.31	-	82.43	4.36	3.81	52.90	22.76	7.02	4.01
		0.21%^{NS}	-0.53%^{NS}		-1.94%^{NS}	6.02%**	-2.24%^{NS}	8.17%***	5.60%***	4.08%***	1.73%***
	5A-	20.65	18.27	-	54.83	2.56	3.00	46.74	19.06	6.63	3.59
2015	**5A+**	20.52	18.09	-	53.89	2.64	2.98	48.97	20.12	6.91	3.63
		-0.61%^{NS}	-0.96%^{NS}		-1.71%^{NS}	3.03%^{NS}	-0.84%^{NS}	4.77%**	5.54%***	4.33%***	1.01%^{NS}
	5A-	23.25	22.55	-	83.77	3.86	3.72	46.04	19.59	6.63	3.65
2016	**5A+**	23.27	22.35	-	77.04	3.75	3.45	48.68	20.35	6.78	3.71
		0.10%^{NS}	-0.90%^{NS}		-8.04%**	-2.90%^{NS}	-7.24%**	5.72%**	3.86%**	2.21%***	1.56%*
	5A-	22.41	20.87	9.46	70.37	3.17	3.24	44.81	20.05	6.66	3.72
Overall	**5A+**	22.45	20.76	9.63	67.88	3.25	3.13	47.49	21.03	6.90	3.77
		0.17%^{NS}	-0.53%^{NS}	1.87%*	-3.55%*	2.33%*	-3.37%*	6.00%***	4.87%***	3.57%***	1.34%***

%s indicate amount gained in 5A+ NILs compared with 5A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row) i.e. NS = Non-significant, * = P < 0.05, ** = P < 0.01, *** = P < 0.001. 2012-13 = BC_2 NILs, 2014-16 = BC_4 NILs. - = data not available. SES = single ear sample

5A NILs were also measured for several developmental traits (Table 2.10). There was a significant reduction of 4 tillers per m across two years (P = 0.008) and a 1 cm increase in crop height in a single year (P = 0.038) in the 5A+ NILs (Table 2.10). No effect was seen for days to heading or days to maturity (Table 2.10). Taken together, these results suggest that the 5A+ QTL has a consistent positive effect on TGW and that the effects on yield are modulated by a series of smaller compensating negative effects on yield components such as grain number, seeds per spikelet and tiller number.

Table 2.10: Developmental traits of 5A NILs

Year	Genotype	Days to heading	Days to maturity	Tiller number	Crop Height (cm)
2012	**5A-**	250.2	318.5	126.7	76.9
	5A+	250.5	317.7	121.2	77.9
		0.3^{*}	-0.8^{*}	-5.5^{*}	1.0^{*}
2013	**5A-**	250.2	298.9	-	-
	5A+	250.5	298.7	-	-
		0.3^{NS}	-0.2^{NS}		
2014	**5A-**	236.4	293.3	69.3	-
	5A+	236.5	293.3	66.9	-
		0.1^{NS}	0.0^{NS}	-2.4^{NS}	
2015	**5A-**	246.5	-	-	-
	5A+	246.0	-	-	-
		-0.5^{NS}			
2016	**5A-**	242.5	-	-	-
	5A+	242.8	-	-	-
		0.3^{NS}			
Overall	**5A-**	245.2	303.6	98.0	76.9
	5A+	245.2	303.2	94.0	77.9
		0.1^{NS}	-0.3^{NS}	-4.0^{**}	-

Differences indicate amount gained in 5A+ NILs compared with 5A- NILs. Superscripts indicate significance determined by ANOVA for either each year, or across all years (final row) i.e. NS = Non-significant, * = P < 0.05, ** = P < 0.01. 2012-13 = BC_2 NILs, 2014-16 = BC_4 NILs. - = data not available

2.4.2.2 The TGW increase in 5A+ NILs is primarily due to increased grain length

NILs were assessed for grain morphometric parameters (length, width and area) (Table 2.8). 5A+ NILs had significantly increased grain length ($P < 0.001$), width ($P < 0.001$) and area ($P < 0.001$) compared to 5A- NILs across all years with the exception of width in 2016. On average, the 5A+ QTL increased grain length by 4.04% ($P < 0.001$), ranging from 3.58 to 4.72% ($P < 0.001$ in all years). Unlike the *TaGW2_A* NILs which had equivalent differences in length and width (Table 2.6), the 5A effect on width was smaller than that on length, averaging 1.45% ($P < 0.001$; range 0.27 to 2.35%) and significant in four out of five years (Table 2.8). The effects on length and width combined to increase grain area by an average of 5.41% ($P < 0.001$), significant in all five years. These results were based on combine harvested grain samples and were also confirmed in ten representative SES taken before harvest. TGW of the ten spikes correlated strongly with the whole plot samples ($r = 0.84$, $P < 0.001$) and showed a similar difference between NILs (6.00%, $P < 0.001$; Table 2.9). Across datasets, the effect of the 5A+ QTL on grain length was more than twice the size of the effect on grain width. This fact, together with the more consistent effect on grain length across years (Coefficient of variation length = 10.6%; width = 55.3%; TGW = 27.8%) suggests that the increase in grain length is the main factor driving the increase in grain area and TGW.

2.4.2.3 The 5A QTL has a uniform effect on grains within the spike

Violin plots for grain length showed variation in the shape of the distribution of individual seeds among years (Figure 2.7). However, within years the 5A- and 5A+ grain length distributions were very similar in shape, suggesting the 5A QTL affects all grains uniformly and in a stable manner across the whole spike and within spikelets, similar to the 6A QTL. In all years, the 5A+ grain length distributions were shifted higher than the 5A- NILs with an increase in longer grains and fewer shorter grains, in addition to the higher average grain length (Figure 2.7). Grain width distributions were also very similar in shape within years, but had a less pronounced shift between NILs (Figure 2.8) consistent with the overall smaller effect of the 5A QTL on grain width.

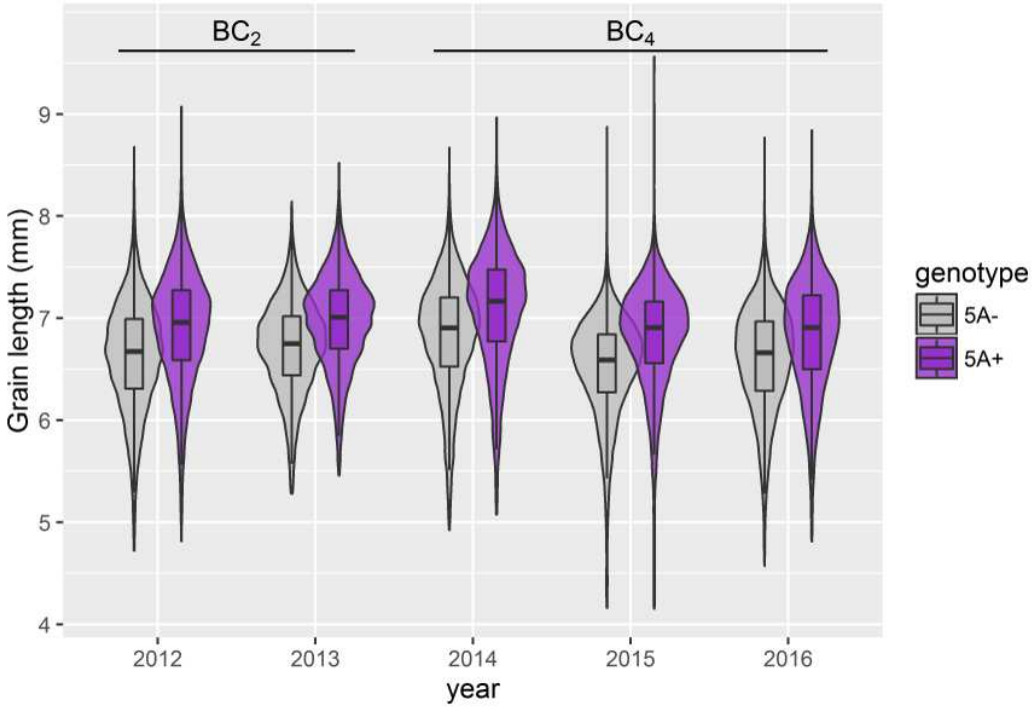


Figure 2.7: Distribution of grain length of 5A NILs from whole plot samples

Violin plots showing the distribution of individual seed measurements of grain length across the five field experiments of 5A BC_2 (2012-2013) and BC_4 (2014-2016) near isogenic lines (NILs). Purple = 5A+ NILs, grey plots = 5A- NILs. Black lines within the boxes indicate the median value. All within year comparisons between NILs were significant ($P < 0.001$).

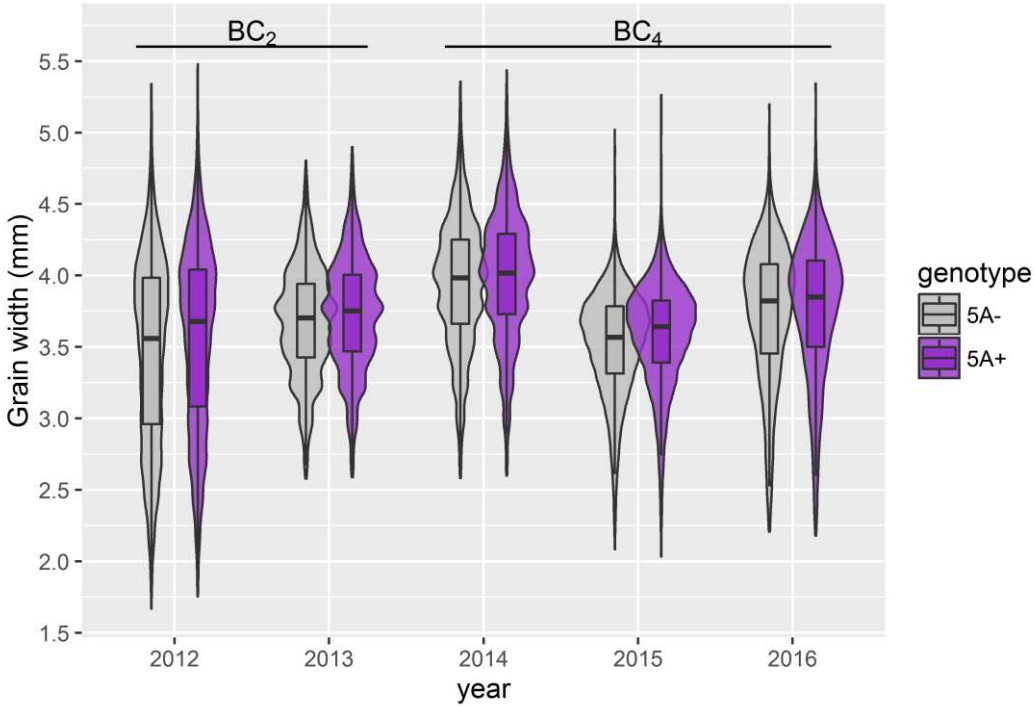


Figure 2.8: Distribution of grain width of 5A NILs from whole plot samples

Violin plots showing the distribution of individual seed measurements of grain width across the five field experiments of 5A BC_2 (2012-2013) and BC_4 (2014-2016) near isogenic lines (NILs). Purple = 5A+ NILs, grey plots = 5A- NILs. Black lines within the boxes indicate the median value. 2012-2015 within year comparisons between NILs were significant ($P < 0.01$). The 2016 comparison between NILs was non-significant.

2.4.2.4 The 5A QTL region acts during grain development to increase grain length

Grain development time courses of two 5A- and two 5A+ BC_4 NILs were conducted to determine when differences in grain morphometric parameters (grain length, width and area) between NILs are first established. Grain FW and DW were also measured. Grains were sampled in 2014, 2015 and 2016 from field plots at anthesis (with the exception of 2014) and at five further time points across grain development until the difference in grain size had been fully established. Exact time points are detailed in Figure 2.9 and Table 2.11. Similar profiles were observed in all years, with the first morphometric parameter to show a significant difference being grain length (Figure 2.9a-c) and any differences in grain width were not observed until the final time point (Figure 2.9d-f).

In 2014, the first significant difference in grain length was observed at 8 dpa with 5A+ NILs having 3.33% longer grains than 5A- NILs (P = 0.001; Table 2.11, Figure 2.9a). This was later in grain development than when the first significant differences in grain size were observed in both the 6A and *TaGW2_A* NILs. 5A+ grains remained significantly longer until the final time point (27 dpa; 4.46 % increase, P < 0.001; Figure 2.9a). Similarly, differences in grain area were observed at 8 dpa (5.57%, P = 0.001) and maintained until the final time point (6.52%, P < 0.001; Figure 2.9g). A significant difference in grain width was observed at the final time point only (2.47%, P = 0.001; Figure 2.9d). Grains from 5A+ NILs were also significantly heavier at 8 dpa (FW: 8.31%, P = 0.004; DW: 6.82%, P = 0.020). 5A+ grains remained heavier until the final time point, although the differences at 12 dpa were non-significant (Figure 2.9j,m).

In 2015, the first significant difference in grain length was observed at 12 dpa with 5A+ NILs having 1.49% longer grains than 5A- NILs (P =0.035). Although this was four days later than the first grain length difference in 2014, the mean grain lengths were similar at these time points in the two years (2014: 5A+ = 6.62 mm, 5A- = 6.41 mm; 2015: 5A+ = 6.50 mm, 5A- = 6.40 mm). The 2015 grain length effect increased to 4.35% at 19 dpa (P < 0.001) and was maintained at the final time point (26 dpa; 4.48 % increase, P < 0.001; Figure 2.9b). Significant differences in grain area were detected at 19 dpa (5.74 % increase; P < 0.001; Figure 2.9h) and this difference was maintained at the final time point (6.06 %, P < 0.001). No significant effects on grain width were observed until 26 dpa when 5A+ NILs increased grain width by 1.66 % (P = 0.015; Figure 2.9e). By the final time point 5A+ NILs also had significantly heavier grains (FW: 7.13 %, P < 0.001; DW: 3.71 %, P = 0.01; Figure 2.9k, n).

In 2016, the first differences in both grain length (2.88 %, P < 0.001) and grain area (4.15 %, P < 0.001) were observed at 15 dpa (Figure 2.9c, i). These differences increased to 4.02 % (grain length, P < 0.001) and 6.30 % (grain area, P < 0.001) at the final time point (21 dpa). There was also a 6.37 % increase in the FW of 5A+ grains at the final time point (P = 0.019). No significant differences were observed at any time point in grain width or dry weight in 2016, reminiscent of the non-significant difference in the grain width of mature grains in 2016 (Table 2.8).

In all years, the grain size and dry weight effects observed were consistent with the differences observed in mature grains (Table 2.8). The fact that the effects on width, area and weight were all

observed after the first significant difference in grain length in all three years further supports grain length as the main factor driving the increase in grain weight in 5A+ NILs.

Table 2.11: Differences between 5A NILs of grain size and weight parameters during grain development time courses

Year	Days Post Anthesis	Length (%)	Width (%)	Area (%)	FW (%)	DW (%)
2014	4	$2.28^{0.130}$	$0.77^{0.466}$	$3.24^{0.0695}$	$4.57^{0.0764}$	$2.32^{0.430}$
	8	$3.33^{0.00146}$	$1.69^{0.0512}$	$5.57^{0.001}$	$8.31^{0.004}$	$6.82^{0.020}$
	12	$3.57^{<0.001}$	$0.43^{0.670}$	$4.12^{0.0132}$	$5.07^{0.062}$	$2.84^{0.258}$
	18	$4.48^{<0.001}$	$1.16^{0.162}$	$5.72^{<0.001}$	$8.33^{0.003}$	$4.73^{0.031}$
	27	$4.46^{<0.001}$	$2.47^{0.001}$	$6.52^{<0.001}$	$8.47^{<0.001}$	$6.30^{<0.001}$
2015	0	$0.64^{0.386}$	$-0.23^{0.734}$	$0.80^{0.560}$	$1.58^{0.576}$	$1.65^{0.613}$
	4	$0.06^{0.993}$	$0.41^{0.751}$	$0.03^{0.983}$	$-1.79^{0.771}$	$-2.88^{0.876}$
	7	$1.22^{0.404}$	$0.24^{0.777}$	$1.57^{0.493}$	$1.95^{0.505}$	$-2.20^{0.472}$
	12	$1.49^{0.035}$	$-1.05^{0.237}$	$0.63^{0.651}$	$1.35^{0.684}$	$-2.29^{0.317}$
	19	$4.35^{<0.001}$	$1.26^{0.090}$	$5.74^{<0.001}$	$6.27^{0.006}$	$4.72^{0.077}$
	26	$4.48^{<0.001}$	$1.66^{0.015}$	$6.06^{<0.001}$	$7.13^{<0.001}$	$3.71^{0.01}$
2016	0	$2.65^{0.191}$	$1.16^{0.527}$	$3.20^{0.304}$	$2.22^{0.626}$	$2.80^{0.462}$
	3	$-1.21^{0.352}$	$-0.03^{0.926}$	$-1.16^{0.555}$	$-3.14^{0.345}$	$-3.51^{0.088}$
	8	$0.40^{0.743}$	$-0.35^{0.644}$	$-0.63^{0.750}$	$-2.10^{0.463}$	$-4.61^{0.168}$
	10	$1.35^{0.144}$	$-0.14^{0.919}$	$1.51^{0.455}$	$1.56^{0.592}$	$-1.44^{0.602}$
	15	$2.88^{<0.001}$	$0.88^{0.118}$	$4.15^{<0.001}$	$2.10^{0.379}$	$1.06^{0.705}$
	21	$4.02^{<0.001}$	$1.97^{0.063}$	$6.30^{<0.001}$	$6.37^{0.019}$	$1.48^{0.551}$

%s indicate amount gained in 5A+ NILs compared with 5A- NILs. Superscripts are the ANOVA P-values of the comparison between 5A+ and 5A- NILs.

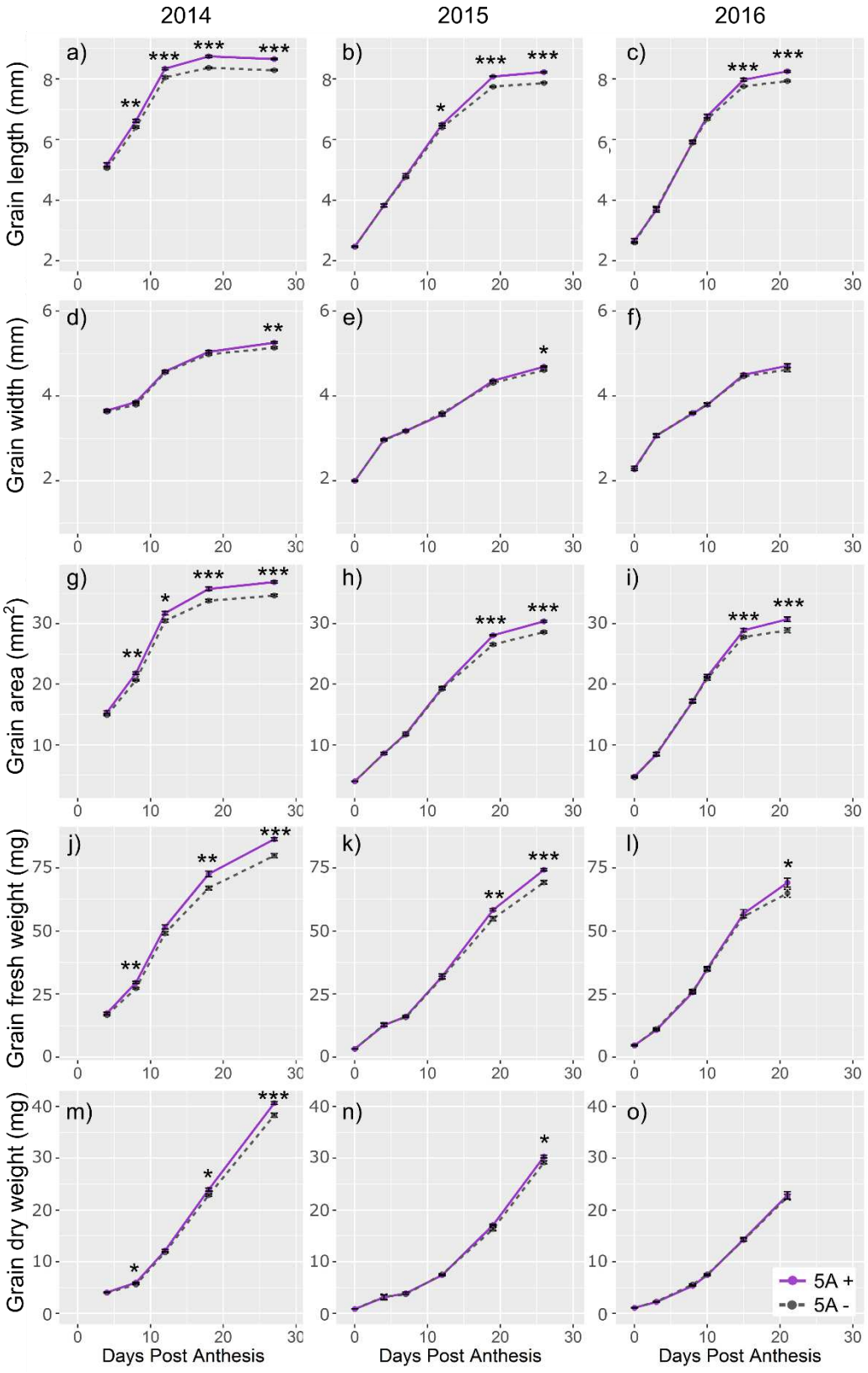


Figure 2.9: Grain developmental time courses of 5A NILs

Grain length (a, b, c), width (d, e, f), area (g, h, i), fresh weight (j, k, l) and dry weight (m, n o) of 5A- (grey, dashed line) and 5A+ (purple, solid line) BC_4 near isogenic lines (NILs) during grain development in 2014-2016 field trials. 2014 samples: 4, 8, 12, 18 and 27 days post anthesis (dpa)); 2015 samples: 0 (anthesis), 4, 7, 12, 19 and 26 dpa; 2016 samples: 0, 3, 8, 10, 15 and 21 dpa. * = $P < 0.05$, ** = $P < 0.01$, *** = $P < 0.001$. Error bars show one standard error above and below the mean.

2.4.2.5 5A+ NILs have increased pericarp cell length independent of absolute grain length

Grain size can be influenced by both cell proliferation and cell expansion. To understand which of these processes the 5A QTL affects, SEM was used to image pericarp cells and determine cell size of BC_4 5A- and 5A+ grains. Mature grains from the 2015 field experiment were selected from a 5A- and 5A+ NIL pair based on their grain length and using a variety of criteria to allow for distinct comparisons (Figure 2.10). The first comparison was between grains of average length from the 5A- and 5A+ NIL distributions (Figure 2.10a). Average 5A+ grains had an 8.33 % significant increase in mean cell length ($P = 0.049$) compared to average 5A- grains and this was reflected in a shift in the whole distribution of 5A+ cell lengths (Figure 2.10a). Next, cell lengths in grains of the same size from 5A- and 5A+ NILs were compared. Relatively long grains from the 5A- NIL distribution (Figure 2.10b; orange) that had the same grain length as the average 5A+ grains were selected. This comparison showed that 5A+ grains still had longer cells (9.53%, $P = 0.015$) regardless of the fact that the grain length of the two groups were the same (6.8 mm; Figure 2.10b). The opposite comparison was also made by selecting relatively short grains from the 5A+ NIL distribution (Figure 2.10c; green) and comparing them with average 5A- grains. Similar to before, the 5A+ grains had longer cells (8.61%), although this effect was borderline non-significant ($P = 0.053$; Figure 2.10c). Finally, a comparison of long 5A- grains and short 5A+ grains again showed that cells were longer in 5A+ grains (9.81%, $P = 0.011$; Figure 2.10d), even though the 5A+ grains used in this comparison were 7.65% shorter than the 5A- grains. Within genotype comparisons of cell length between grains of different lengths showed no significant differences in mean cell length (Figure 2.10e, f). The results were confirmed in 2016 where average 5A+ grains had a 24.6 % significant increase in mean cell length compared to average 5A- grains ($P < 0.001$; Figure 2.11). These results indicate that the 5A+ region from Badger increases the length of pericarp cells independent of absolute grain length. In 2015, average length grains of both 5A- and 5A+ NILs had the same number of cells (calculated as grain length / mean cell length; Figure 2.12a). However, in 2016, 5A- NILs had significantly more cells than 5A+ NILs (19.8 %, $P < 0.001$; Figure 2.12b).

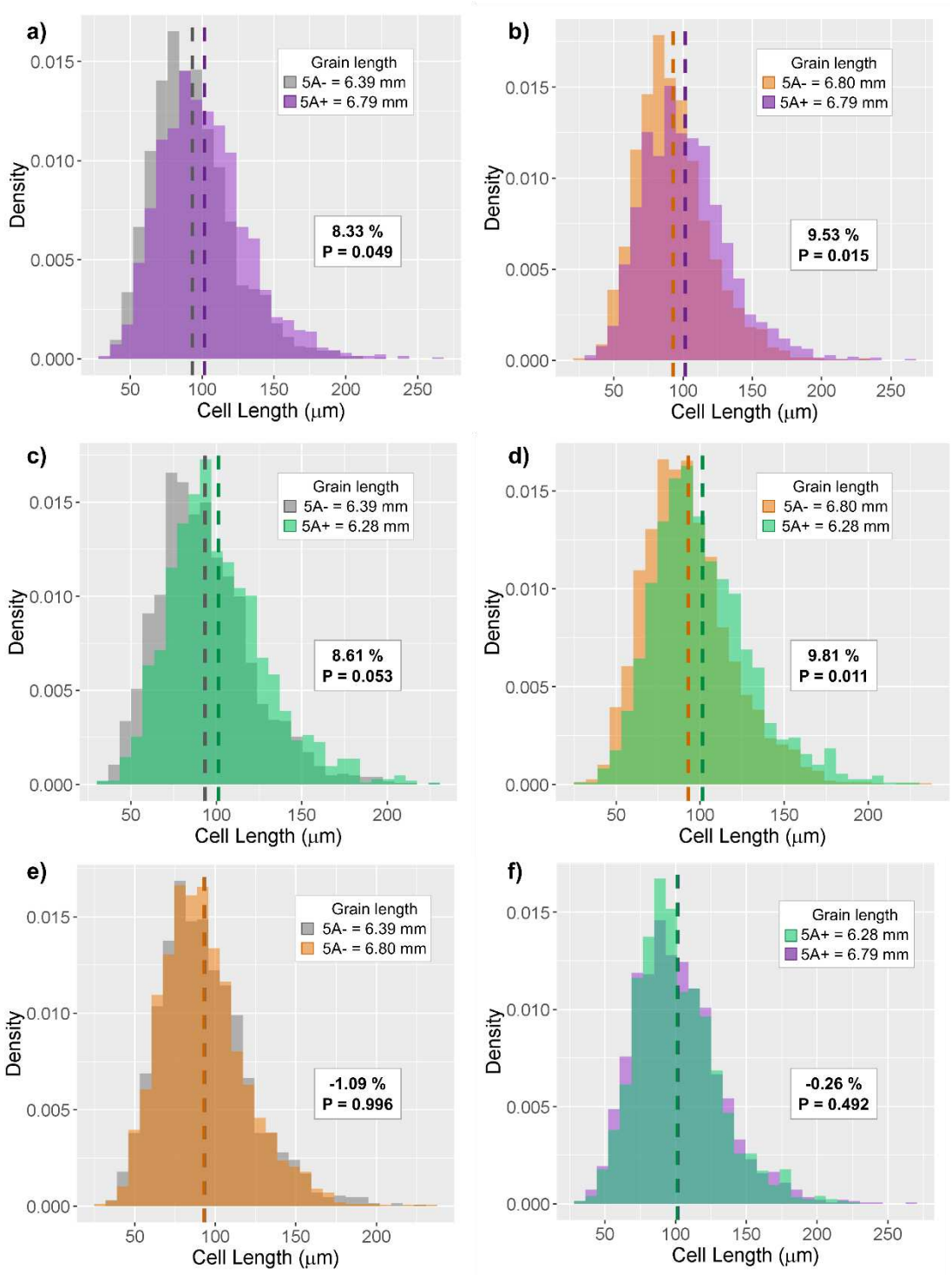


Figure 2.10: Comparisons of pericarp cell length in 5A NILs (2015)

Density plots of cell length measurements from 27 grains per genotype group in 2015; dashed line represents the mean. "Grain length" insets show the average grain length of each group of grains used for measurements. In panels a-d the increase in cell length of 5A+ near isogenic lines (NILs) relative to cell length of 5A- grains is shown as a percentage along with the P-values calculated using ANOVA to compare means of the two groups displayed. In panels e-f the percentage indicates the increase in cell length of the group with longer grains relative to the group with shorter grains. a) Grains of average length from 5A- and 5A+ NILs, b) average 5A+ grains and equivalent 5A- grains, c) average 5A- grains and equivalent 5A+ grains, d) long 5A- grains (length equivalent to average 5A+ grains) and short 5A+ grains (grain length equivalent to average 5A- grains), e) average 5A- grains and long 5A- grains, f) short 5A+ grains and average 5A+ grains.

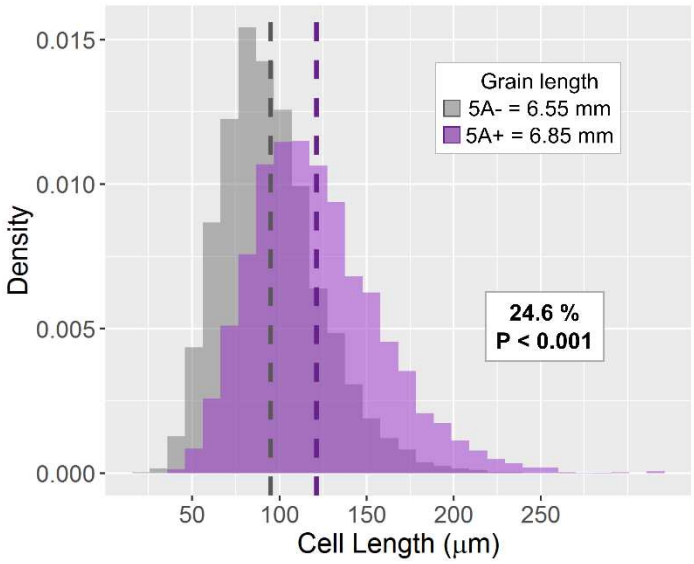


Figure 2.11: Comparison of pericarp cell length in 5A NILs (2016)

Density plots of cell length measurements from 2016 grains. Dashed line represents the mean. "Grain length" inset shows the average grain length of each group of grains used for measurements. Grains used were of average length from 5A- and 5A+. The increase in cell length of 5A+ NILs relative to cell length of 5A- grains is shown as a percentage along with the P-value calculated using ANOVA to compare means of the two groups displayed.

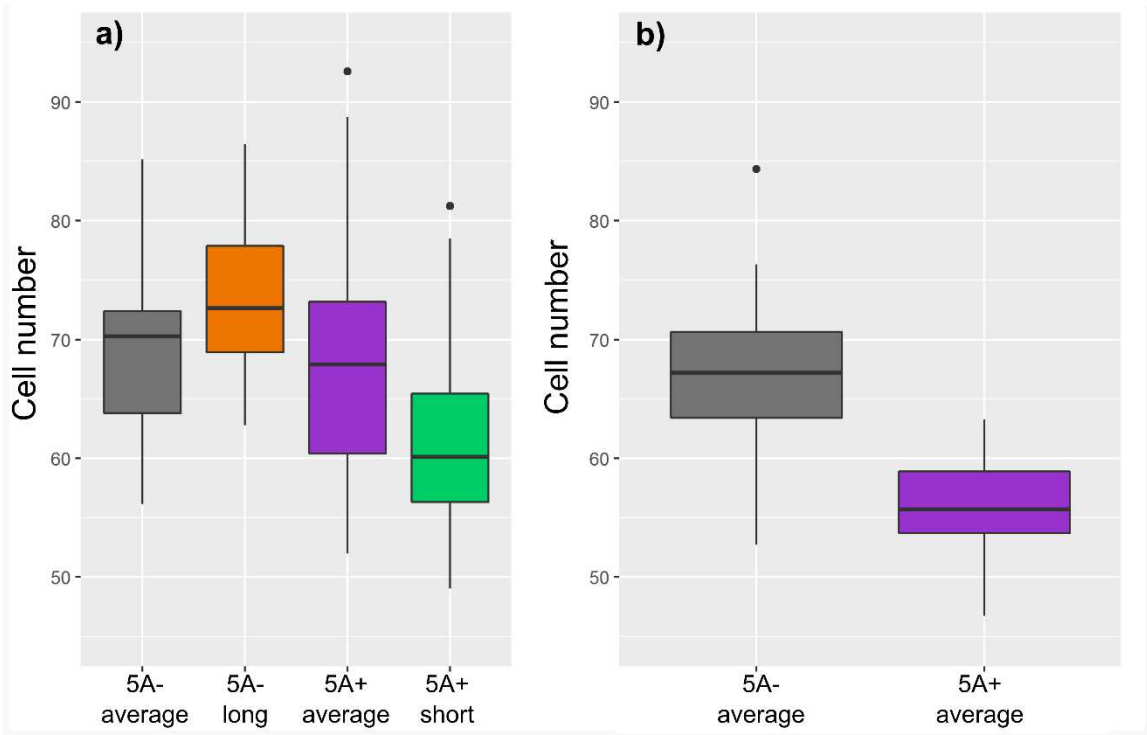


Figure 2.12: Comparison of pericarp cell number in 5A NILs (2015 and 2016)

Boxplots show the distribution of cell number (calculated as grain length/mean cell length) in the different groups of grains from which pericarp cell size was measured in 2015 (a) and 2016 (b). In 2015, there was no significant difference between 5A- average and 5A+ average cell numbers, whereas in 2016 the difference was significant ($P < 0.001$)

2.5 Discussion

2.5.1 The 6A and 5A QTL act through distinct mechanisms

An aim of this chapter was to gain mechanistic insight into two different grain weight QTL (6A and 5A) through detailed phenotypic characterisation of NILs. Both the 6A and 5A QTL are associated with a significant increase in TGW across years (4.4% and 6.7%, respectively), however, the increases in TGW are achieved through different mechanisms. The 6A QTL acts to increase grain width whilst the 5A QTL acts to increase grain length, supporting the fact that these grain size parameters are under independent genetic control in wheat (Gegas *et al.*, 2010). This is consistent from studies in other cereal crops such as maize and rice (Wang *et al.*, 2012; Chen *et al.*, 2016).

Significant differences in grain width and area were observed between 6A NILs but no significant differences in grain length were observed in any year. This, together with the fact that the magnitude of the differences in grain width and grain area were very similar (2.31 % and 2.58 %, respectively) strongly support grain width as the sole effect underlying the increase in grain area in 6A+ NILs and a major contributor to the final grain weight.

Conversely, differences in both grain length and grain width were observed in the mature grains of 5A NILs. However, taking all results together suggests that grain length is the primary driver of increased grain weight in 5A+ NILs. Across three years, the difference in grain length between 5A NILs was the first grain size component difference to be established. Only after this, were any differences in grain width or weight observed. These differences in final grain length were extremely consistent across years (despite average TGW values ranging from 39.8 to 50.3 g) compared to the more variable differences in grain width and weight. Additionally, the effect on grain length was double the size of the grain width effect. These results suggest that the 5A QTL increases TGW by a primary effect on grain length, which confers the potential for further enhancements by pleiotropic effects on grain width. The grain length effect is genetically controlled and stable across environments, whereas the pleiotropic effect on grain width occurs later in grain development and is more environmentally dependent and variable. Increases in grain length and grain width then combine to give a roughly additive effect on grain area (length: 4.04%, width: 1.45%, area: 5.41%). The final magnitude of the 5A grain weight increase (ranging from 4.0 to 9.3 %) is thus determined by the extent to which the late stage pleiotropic effect on grain width is manifested and the potential exploited. This could explain why the grain width increase was significantly correlated with the increase in TGW ($r = 0.98$, $p = 0.004$) whilst grain length was not ($r = 0.71$, $p = 0.18$). This hypothesis could be tested by evaluating the 5A NILs in different environments where factors such as location, sowing date, water levels and fertilisation regimes are directly manipulated. If the hypothesis is correct then the grain length difference would remain stable and present under all conditions, whereas the magnitude of the grain width and weight differences would vary with conditions, showing larger increases in more favourable conditions.

The first differences in grain width between 6A NILs were observed at earlier stages of grain development than the first grain size differences observed between 5A NILs. This supports the hypothesis that the two QTL act by distinct mechanisms as they not only influence different grain size components, but also act at different times during carpel/grain development. The early stage difference in width in 6A NILs suggests again that this difference in grain width is achieved by a different mechanism than the late increases in grain width in the 5A NILs. The grain width difference in 5A NILs only occurs at the later stages of grain development, likely to be because of an enhanced capacity for grain filling afforded by the increased grain length. In the 6A NILs, the difference in grain width is established before grain filling begins and might therefore be achieved by a more direct effect on organ size. Measuring the size of other organs in 6A NILs, such as leaves and total plant biomass, could provide insight into whether this is the case.

The increase in grain weight and length in 5A+ NILs was associated with increased pericarp cell length. In wheat and barley, pericarp cell division decreases shortly after fertilisation (2 to 6 days; Drea *et al.*, 2005; Radchuk *et al.*, 2011) and cell expansion plays the predominant role in increasing pericarp size during grain development. These results are consistent with a role of the 5A QTL on pericarp cell expansion given that significant differences in grain size are only established c. twelve days after fertilisation, once cell expansion has begun. However, a possible overlapping late effect on cell division cannot be discarded given the conflicting results in final pericarp cell number between years (Figure 2.12).

Due to time constraints, no cell size/number data was obtained for the 6A NILs however these studies are currently underway using field samples grown in 2017. As the differences in grain size are established during very early grain development in the 6A NILs, we hypothesise that 6A+ NILs have increased numbers of cells. At the very early stages of carpel/grain development, growth is mainly driven by cell division (Drea *et al.*, 2005) and only later does cell expansion take over, as seen in the grain development dynamics of the 5A NILs. An increase in either the rate or duration of cell division in 6A+ NILs could also account for the initial difference observed in carpel/grain length that is not present in mature grains. The subsequent rapid expansion of pericarp cells occurs mostly in the longitudinal direction (Pielot *et al.*, 2015), perhaps to a genetically determined grain length, hence the final grain length of 6A NILs is the same. The final grain width is achieved only after the final grain length has been established (Rogers & Quatrano, 1983), largely through grain filling processes (Shewry *et al.*, 2012). In the 6A NILs, it is possible that this filling process continues in the same way in both genotypes, but that the initial increase in cell number in 6A+ NILs allows a larger final grain size to be achieved.

Future work will also examine cell size and number in the developing carpels/grain. Assessment on the cellular level could determine whether the 6A QTL acts pre- or post-fertilisation. Further studies looking at the dynamics of cell proliferation and expansion across carpel/grain development time

courses for both QTL will provide insight into how exactly these processes interact to achieve the final differences in grain size.

2.5.2 6A NILs and *TaGW2_A* NILs show similar but distinct phenotypes

The A genome wheat orthologue of the rice E3 ubiquitin ligase, *GW2*, maps within the original 6A grain weight QTL interval (Simmonds et al, 2014). *GW2* negatively regulates grain width and weight in rice (Song *et al.*, 2007) and was therefore considered as a potential candidate for the causal gene underlying the 6A QTL. To test this hypothesis phenotypically, grain size parameters and carpel/grain development of NILs for a knock-out mutation of *TaGW2_A* and the 6A NILs were compared.

TaGW2_A NILs showed differences in both final grain length and grain width, whilst the 6A NILs had differences in grain width only. Unlike in the 5A NILs, which also had differences in both final grain length and grain width, the differences in these parameters in *TaGW2_A* NILs were of similar magnitude (1.74 % and 1.94 %, respectively) and were established at the same stage of carpel/grain development. This suggests that *TaGW2_A* acts to genetically increase both grain length and width, rather than the genetic control of one component leading to pleiotropic downstream effects on the other component as in the 5A NILs. Given that grain length and grain width are under independent genetic control (Gegas *et al.*, 2010), this could suggest that *TaGW2_A* and the 6A QTL act via different mechanisms. Likewise, the carpel/grain development profiles of the 6A NILs and *TaGW2_A* NILs showed unique patterns. Differences in carpel/grain length were seen throughout grain development in the *TaGW2_A* NILs (excluding the final time point in 2015), whilst carpel/grain length differences were rarely observed in the 6A NILs and only present at the very early stages of carpel/grain development. Additionally, *TaGW2_A* NILs displayed clear differences in carpel width and length at heading. These results suggest that *TaGW2_A* acts maternally to control grain size in wheat, consistent with the role of the Arabidopsis homologue, *DA2*, which acts to increase cell proliferation in the integument, a maternal tissue (Xia *et al.*, 2013). No significant differences between any grain size/weight components were observed at heading in 6A NILs. However, all components were higher in 6A+ NILs and many of the differences were borderline non-significant. It is therefore not possible to determine from these data whether the 6A QTL also acts on maternal tissue before fertilisation.

Currently, experiments are being performed to look at differences in cell size and cell number in *TaGW2_A* NILs, which could provide further information as to whether the mechanism is similar to the 6A QTL. As *TaGW2_A* acts during carpel development, this is again likely to be an effect on cell number. Additionally, both the rice and Arabidopsis orthologues of *TaGW2_A* influence organ size through modulation of cell division rather than expansion (Song *et al.*, 2007; Xia *et al.*, 2013).

Phenotypic differences alone cannot rule out *TaGW2_A* as a candidate gene for the 6A QTL. The *TaGW2_A* NILs only allow examination of the phenotype of a specific allele of *TaGW2_A* and it is

possible that the causal 6A QTL gene could be a different allelic variant of *TaGW2_A* that causes a more subtle and slightly altered phenotype. Whilst there are no SNPs present in the coding regions of *TaGW2_A* in the parental varieties of the 6A QTL, Spark and Rialto, there is a SNP located 593 bp upstream of the *TaGW2_A* start codon (Simmonds *et al.*, 2014). Rialto, the positive 6A parent, carries a G at this position and Spark, the negative 6A parent, has an A. This SNP has previously been associated with grain width and TGW although studies have generated contradictory results, finding different alleles associated with increased grain weight (Su *et al.*, 2011; Zhang, X *et al.*, 2013; discussed in more detail in the general discussion).

2.5.3 Differences in grain area do not fully account for differences in TGW

For both the 5A and 6A QTL and *TaGW2_A*, the differences in grain area between NILs did not fully account for the differences in grain weight (6A NILs: TGW = 4.39 %, grain area = 2.58 %; 5A NILs: TGW = 6.92 %, grain area = 5.41 %, *TaGW2_A* NILs: TGW = 6.65 %, grain area = 3.57 %). One possible explanation could be that increases in grain size are not directly proportional to increases in grain weight. A more likely explanation is that grain area only considers grain size in two dimensions, whereas the grain is actually a three-dimensional structure. Measurements of grain volume, taking into account differences in grain height/thickness as well as length and width, are required and could account for the 'missing' difference. Such 3D measurements are possible, for example, X-ray Computed Tomography (CT) scanning has been used to image grains *in situ* in the spike and can provide additional information about grain morphology such as volume and crease depth (Strange *et al.*, 2015). Studies in barley have also used magnetic resonance imaging (MRI) to obtain more detailed analyses of grain dimensions, in addition to processes taking place inside the grain (Pielot *et al.*, 2015). However, these measurements remain challenging in large scale studies that require high throughput methods due to the high numbers of replications required to detect the more subtle differences observed in wheat NILs.

2.5.4 Increases in TGW do not consistently translate into increases in final yield

For all NILs assessed, consistent effects on grain length or width and TGW did not always translate into increased yield. It has previously been shown that grain weight is more stably inherited than yield itself (Kuchel *et al.*, 2007). However, there can be trade-offs between different yield components, in particular grain weight and grain number, which could account for the lack of increase in final yield.

In the original DH analysis for the 5A QTL, the 5A TGW effect co-located with final yield in seven of the twelve environments in which the population was assessed (Brinton *et al.*, 2017). This overall positive trend was also reflected in the 5A NILs, although yield increases were only significant in 2014. Across years there were small negative effects in the 5A+ NILs on yield components such as tiller and grain number. Although these differences were not consistent across all years, it is possible that negative effects on these yield components modulate the overall effect

of the 5A QTL on yield. Similarly, in the original DH analysis for the 6A QTL, the 6A TGW effect co-located with an effect on final yield (Simmonds *et al.*, 2014). Additionally, Simmonds *et al.* (2014) found that across four years of field trials and a single glasshouse experiment, 6A+ NILs had significantly increased yield in three out of the five experiments. However, in the results presented here, there were no significant differences in final yield observed between 6A NILs. As with the 5A QTL, this could be due to modulation of final yield by a series of smaller negative effects on other yield components, for example 6A+ NILs had significantly fewer tillers in 2014 and in previous years (Simmonds *et al.*, 2014).

These negative pleiotropic effects could be due to additional genes within the wider QTL regions. For example, a minor QTL for tiller number has been detected in the 6A region (Simmonds *et al.*, 2014). If this is the case then identification of the causal genes will allow specific selection of the beneficial alleles, mitigating the negative effects of other genes on final yield. Alternatively, it could be that the effect on final yield is a result of pleiotropic effects caused by the genes underlying the 5A and 6A QTL themselves. If this is the case then it raises the question as to how these compensatory effects on components such as tiller and grain number could arise as they are determined before the phenotype is expressed in the grains (after anthesis (5A) or just before anthesis (6A)). It is possible that these genes do not function in a grain specific manner and may affect other developmental process in different tissues that we are currently not aware of. These effects could then be further modulated by environmental interactions.

Understanding these effects will be challenging until the underlying genes are identified as the subtle differences between NILs mean that it is difficult to separate small, but direct, effects of the genes from random biological variation. Identification of the underlying genes will allow a wider range of variation to be explored in order to determine the exact function of the genes. This has been seen previously in the cloning of the wheat grain protein content (*GPC*) QTL (Uauy *et al.*, 2006). The *GPC* QTL was associated with a 4-5 day difference in senescence timing and a 10 % difference in protein content, but it was not clear exactly how/if the QTL affected nutrient remobilisation dynamics and final yield due to the subtle effects observed. Identification of the underlying gene allowed an RNAi line to be generated, which had a much clearer phenotype (up to 30 day difference in senescence timing and 30 % difference in protein content) and clearly showed that the *GPC* gene itself had no effect on carbohydrate remobilisation or final grain yield (Uauy *et al.*, 2006; Borrill *et al.*, 2015b). In a similar way, the opportunity to directly manipulate the genes underlying the 5A and 6A QTL will allow us to determine whether the pleiotropic effects are due to the gene or environmental variation.

Understanding the precise functions of the underlying genes will be critical to identify how they can affect final yield and it is possible that the full potential of these QTL will be realised only under certain environments or in combination with other genes.

3 Fine-mapping of the 5A and 6A QTL

3.1 Chapter summary

In this chapter, fine-mapping was used to further define the 5A and 6A QTL to narrower genetic intervals and we used the latest genome references to define the physical sequence and genes within the regions. For both QTL, recombinant populations and genetic maps were available at the start of the PhD. Larger recombinant populations for both QTL were generated during the PhD and the marker density across the intervals were increased. We found that the 6A QTL mapped to a large linkage block located in the centromeric region of chromosome 6A. We tentatively mapped the 6A grain width phenotype to a 0.28 cM interval, corresponding to 61.2 Mbp and containing 396 genes. Importantly, this interval did not contain *TaGW2_A* suggesting that *TaGW2_A* is not the causal gene underlying the 6A QTL. We mapped the 5A grain length effect to an overall 6.6 cM interval on the long arm of chromosome 5A. However, recombinants within this interval suggested conflicting mapping positions leading to the hypothesis that there are two distinct genes, *GL1* and *GL2*, underlying the 5A QTL that have an additive effect on grain length. A haplotype analysis of the 5A QTL interval across 20 UK wheat cultivars suggests that the QTL is not fixed in UK germplasm and that *GL1* and *GL2* are not always inherited together.

3.2 Introduction

Map-based or positional cloning is a method for identifying the gene/genomic lesion responsible for a trait of interest through genetically mapping the lesion to a progressively narrower chromosomal interval by successively excluding other parts of the genome. This will continue either until the causal lesion is identified or until the interval is narrow enough that the candidates within it can be evaluated by other methods (Lukowitz *et al.*, 2000). Positional cloning can be used to identify the causal lesions originating from essentially any source, including chemical mutagenesis, radiation, transposon insertion or natural variation (Gallavotti & Whipple, 2015). In the context of this thesis and the 5A and 6A QTL, we are using positional cloning to identify a lesion resulting from natural variation between the parents of the two QTL DH mapping populations.

Positional cloning consists of two main phases: the first is a preliminary mapping to define a broad interval containing the locus of interest e.g. QTL analysis. The second phase is mapping on a finer scale (fine-mapping) focussing on a specific interval to identify the causal lesion. In the case of the 5A and 6A QTL discussed here, the initial broad mapping corresponds to the original identification of the QTL in the QTL analysis performed using the DH populations (Simmonds *et al.*, 2014; Brinton *et al.*, 2017). This chapter will discuss the progress made with the second phase: fine-mapping of the 5A and 6A QTL. For this analysis, we used populations of recombinant inbred line (RILs) that were generated alongside the development of the NILs and so recombination is specifically focussed on the 5A and 6A QTL intervals.

The success of positional cloning relies on three main components: the nature of the phenotype/trait of interest, the recombination frequencies around the causal locus and the availability of genetic markers/physical sequence in the region of interest.

In terms of the phenotype, it is critical that recombinant lines can be unambiguously scored, which can be challenging for quantitative traits. This is not as much of an issue for qualitative or Mendelian traits where plants can be easily scored in a binary manner. However, the differences in grain weight caused by the 5A and 6A QTL are subtle, and the distributions of grain size of the positive and negative NILs overlap (Figure 2.4, Figure 2.7). In this chapter, we have tried to accommodate the quantitative nature of the traits by allocating recombinant lines to a parental type (i.e. large or small grains), essentially "Mendelising" the grain size phenotype.

The resolution to which a trait can be mapped using positional cloning is largely dependent on the recombination rate around the causal locus. Genetic recombination occurs during meiosis and involves the exchange of genetic material between homologous chromosome pairs as a result of successful crossover events. In general, the size of the mapping population determines the mapping resolution as screening more individuals increases the likelihood of identifying a recombination event at a particular position. However, recombination rate is not constant across a chromosome. It has been observed in many species, including wheat, that recombination rates tend to be highest towards the telomeres of the chromosome whilst recombination is suppressed in centromeric regions. The recombination rate can also be affected by the genomic features located in a particular interval. For example, areas of higher recombination rates have been identified in gene-rich regions. A negative relationship has also been identified between recombination rates and repetitive element content. This is particularly relevant for wheat, which has a highly repetitive genome. Therefore, the position of the causal locus on the chromosome can present a major bottleneck to identifying the underlying lesion by positional cloning. Indeed, research is ongoing to better understand how recombination rates are controlled with a view to increasing them to overcome these problems (reviewed in Lambing *et al.*, 2017).

Until recently, the availability of genetic markers and genomic resources was a major limitation to positional cloning efforts in wheat. However, this has drastically changed in recent years with the availability of several high density SNP arrays (Wang *et al.*, 2014; Winfield *et al.*, 2016) and the release of a number of reference genome sequences (IWGSC RefSeq v1.0; IWGSC, 2014; Zimin *et al.*, 2017). The genome sequences are not only valuable for marker identification, but also as they allow the physical sequence across large mappings interval to be accessed without the prerequisite of generating a bespoke physical map. Of course, the reference genome sequence is a single cultivar (Chinese Spring) and this may not be the same as the varieties used for the positional cloning. Therefore there may be differences in sequence or larger scale rearrangements in cultivars of interest with respect to the reference sequence. This is being addressed by the generation of

genome sequences of other cultivars, and can also be complemented with variety specific exome capture data, for example.

The aim of this chapter was to use fine-mapping to refine the 5A and 6A QTL mapping intervals. The latest genomic resources were also used to reveal the genetic architecture underlying the QTL and define genes present in the physical intervals. Additionally, a SNP marker for *TaGW2_A* (Hap-P2; Su *et al.*, 2011) was used in the fine-mapping of the 6A QTL to genetically address the question of whether *TaGW2_A* could be the causal 6A gene. For the 5A QTL we also performed a haplotype analysis to determine how the QTL behaves in UK germplasm.

3.3 Methods

3.3.1 Plant material and growth

The 6A recombinant populations used in this chapter were generated by James Simmonds alongside the development of 6A NILs, described in Simmonds *et al.* (2014). For the original 6A population, 212 BC_4F_2 plants were screened for recombination between markers *gwm334* and *gwm570*, encompassing the 6A genetic map developed during the initial identification of the 6A QTL (Simmonds *et al.*, 2014). 67 recombinants were identified and self-pollinated to generate homozygous BC_4F_3 RILs. The larger 6A RIL population was generated within this PhD in the same way, but screening a larger number of BC_4F_2 plants (2,674). These plants were screened for recombination between a narrower marker interval (*BS00010933-BS00066623*) identifying 892 recombinants. Further development of this population was carried out during the PhD and is therefore described in the results section.

The 5A RIL populations used in this chapter were also generated by James Simmonds alongside 5A NIL development, described in Brinton *et al.* (2017). Screening of 170 BC_4F_2 plants identified 60 recombinants between *gwm293* and *gwm186*, the markers used for the selection of NILs. Recombinant plants were self-pollinated to develop homozygous BC_4F_3 RILs. The larger 5A RIL population was developed in the same way, but screening a larger number of BC_4F_2 plants (1,140) and using a slightly narrower marker interval (*BS00075504* and *BS00183958).* 310 recombinant plants were identified. Again, further development of this population was carried out during the PhD and is described in the results section.

All RIL populations were evaluated at Church farm in Norwich (52.628 N, 1.171 E). Subsets of the original 6A RIL population were evaluated in five trials across four years: large-scale yield plots (1.1 x 6m) in 2013-2016 and an additional trial of 1.1 x 1m plots in 2015. In all five trials a randomised complete block design was used with at least five replications. The exact 6A RILs used in each trial are detailed in Table 3.1 (see Results section). The larger 6A RIL population was evaluated in 2016. RILs were grown in single 1m rows with up to three replications depending on seed availability. Subsets of the original 5A population were evaluated in four trials across three years: 1.1 x 1m plots in 2014 and 2015 and 1.1 x 6m plots in 2015 and 2016. In all four trials, a

randomised complete block design was used with at least five replications. The exact details of 5A RILs assessed in each trial are outlined in Table 3.2 (see Results section). The larger 5A RIL population was evaluated in 2016. RILs were grown in single 1m rows, replicated up to five times depending on seed availability.

3.3.2 Grain phenotyping

Grain morphometric measurements (grain width, length, area) and TGW were recorded on the MARVIN grain analyser (GTA Sensorik GmbH, Germany). For all full plots (1.1 x 6m and 1.1 x 1m) approximately 400 grains obtained from the combine harvested grain samples were used. For single rows, ten representative spikes were harvested from each row. The ten spikes were threshed together and the grains obtained from these samples were assessed.

3.3.3 Marker development

Genetic maps were available for both the 6A and 5A original RIL populations at the start of the PhD (details in Simmonds *et al.*, 2014; Brinton *et al.*, 2017). However these did not provide sufficient marker density across the intervals of interest and additional markers were developed. With the exception of a single marker, SNP markers used to genotype the RIL populations fall into four categories (BS, BA, JB_RNASeq and JBHap markers) which are described below.

3.3.3.1 BS and BA markers

BS (Bristol SNP) markers were developed based on data from 90K iSelect array genotyping of BC_4 6A and 5A NILs (Simmonds *et al.*, 2014; Brinton *et al.*, 2017). BA (Bristol Axiom) markers were developed based on data from 820k Axiom array genotyping of parental varieties of the QTL: Spark (6A-), Rialto (6A+), Charger (5A-) and Badger (5A+) (Winfield *et al.*, 2016). KASP primers for all SNPs in the iSelect and Axiom arrays have been designed previously by Ricardo Ramirez-Gonzalez using Polymarker (Ramirez-Gonzalez *et al.*, 2015) and are publicly available at http://polymarker.tgac.ac.uk/. Initially BS and BA markers across the 6A and 5A QTL intervals were selected based on the predicted genetic positions of markers (POPSEQ). However, with the release of more contiguous genome assemblies, markers were selected based on their physical positions across the intervals with respect to the reference sequence (details of how markers were positioned are below (3.3.4)).

3.3.3.2 JB_RNASeq markers

JB_RNASeq (Jemima Brinton RNASeq) markers used to genotype the 5A RILs were designed using RNA-Seq data from a pair of 5A NILs. Twelve RNA samples from grains were sequenced: one 5A- and one 5A+ NIL, each at two time points and with three biological replicates. The RNA-Seq experiment and detailed methods including RNA extraction and sequencing are described in detail in Chapter 4. Specifically for the SNP identification, RNA-Seq reads were aligned to the Chinese Spring Chromosome Survey Sequence cDNA reference (CSS; IWGSC, 2014) downloaded

from *Ensembl* plants release 29. Read alignment was performed using kallisto-0.42.3 (Bray *et al.*, 2016) with default parameters, 30 bootstraps (-b 30) and the –pseudobam option. Pseudobam files for each genotype (5A- and 5A+) were merged to generate a single BAM file for each genotype. SNP calling with respect to Chinese Spring was performed using the samtools-0.1.19 mpileup command followed by the bcftools-1.2 call command (Li *et al.*, 2009). Samtools mpileup was used with the -Agf options: -A includes improperly paired reads, -g computes the genotype likelihoods and outputs them in binary call format (BCF) and -f specifies a reference fasta file. The bcftools call command was used with -O u (to give an uncompressed output, essential for downstream processing) and -c (to call SNPs using Bayesian inference) options. BCF files were converted to variant call format (VCF) using bcftools view and VCF files were filtered with samtools vcfutils.pl using -d 10 -a 9 options to output SNPs with a minimum read depth of 10 and a minimum alternate read number of 9. A grep command was used to extract only SNPs with an allele frequency of 1 ('AF1=1') to filter for homozygous SNPs only. SNPs located in the 5A mapping interval were extracted and compared between genotypes to identify SNPs that were unique to either the 5A- or 5A+ NIL. This identified 145 SNPs between NILs in 34 gene models. However, after manual inspection of BAM files only SNPs in four of the genes looked to be real. Common reasons for discarding SNPs included small regions of mis-mapping or the SNP being present in both NILs but filtered out of the output for one NIL due to low read depth. All four SNPs were validated experimentally using KASP assays (designed using Polymarker (Ramirez-Gonzalez *et al.*, 2015) which were subsequently used as markers JBRNA_Seq1-4 (Appendix 2). JBRNA_Seq1, 2 and 4 were predicted to be non-synonymous SNPs resulting in missense mutations in the 5A- NIL. The three genes (1: Traes_5AL_6401EFD6F, 2: Traes_5AL_AEB344EBB, 4: Traes_5AL_632F49251) were predicted to encode a TATA binding protein, an Fe-S cluster protein and P-loop NTPase, respectively.

3.3.3.3 **JBHap markers**

The JBHap (Jemima Brinton Haplotype) markers were developed based on the haplotype analysis conducted across the 5A interval (described below). KASP assays were designed using Polymarker for 22 SNPs defining haplotypes across the 5A grain length mapping interval (Appendix 2).

3.3.3.4 **Hap-P2 marker**

Hap-P2 is an A/G SNP at the -593 bp position in the promoter of *TaGW2_A* and the original marker was designed as a cleaved amplified polymorphism sequence (CAPS) marker by Su *et al.* (2011). For ease of genotyping a KASP assay for the Hap-P2 SNP was designed using Polymarker and used to genotype the 6A RIL populations (Appendix 2).

3.3.4 Physical positions

To obtain physical locations, SNPs were positioned with respect to the recently released Chinese Spring sequence (IWGSC RefSeq v.1.0; https://wheat-urgi.versailles.inra.fr/Seq-

Repository/Assemblies). Physical positions of all iSelect and Axiom SNPs were obtained using BLASTN (Altschul *et al.*, 1990) to align the surrounding sequence (201 bp) to the RefSeq v.1.0 assembly, provided by Ricardo Ramirez-Gonzalez and available at http://www.wheat-training.com/useful-wheat-links/. The positions of all additional SNPs were determined in a similar way by using BLASTN to align 100-300 bp of surrounding sequence to RefSeq v1.0. Positions of TGACv1 gene models in RefSeq v.1.0 were obtained using GMAP (Wu & Watanabe, 2005) retaining the best hit position and using a 95% minimum similarity cut-off (David Swarbreck and Gemy Kaithakottil, Earlham Institute).

3.3.5 DNA extraction and KASP genotyping

DNA extraction and KASP genotyping were performed as previously described (Pallotta *et al.*, 2003; Trick *et al.*, 2012).

3.3.6 Exome capture for haplotype analysis

Exome capture data for 20 UK wheat cultivars were provided by Philippa Borrill. Alignment of data and SNP calling with respect to the CSS reference (IWGSC, 2014) were also performed by Philippa Borrill. Briefly, reads were aligned to the CSS reference using bowtie2 with the very-sensitive-local option (Langmead & Salzberg, 2012) followed by SNP calling using freebayes (Garrison & Marth, 2012) with the following options: --use-best-n-alleles 2 (only allow sites with up to two alleles), --min-mapping-quality 7 (only use reads with MAPQ>7) and --min-base-quality 20 (only use bases with quality > 20). Details of how SNPs defining haplotypes across the 5A grain length interval were identified are detailed in the results section. The position of SNPs with respect to the IWGSC RefSeq v1.0 were determined as described above (3.3.4).

3.3.7 Statistical analysis

RILs were evaluated using two-way ANOVAs. For the original 6A and 5A RIL populations, the model included the trial as a factor in the model. When individual trials were evaluated, the field block (replicate) was included as a factor in the model. Similarly, for the larger RIL populations (assessed in a single trial) the field block was included as a factor in the model. When RIL groups were assessed, independent RILs within each group were considered as replicates within the model. For the larger RIL populations, individual RILs belonging to a single RIL family were considered as replicates of a single independent RIL. RIL groups were assigned to parental genotypes using a *post hoc* Dunnett's test to compare with control groups. The specific control groups used for each comparison are described in the results section. All statistical analyses were performed using Minitab® Statistical Software.

3.4 Results

3.4.1 Genetic mapping of the 6A QTL for grain width

3.4.1.1 Grain width maps to a 4.6 cM interval on chromosome 6A

A set of 67 RILs with recombination between the microsatellite markers *gwm334* and *gwm570* were used initially to fine map the grain width QTL on chromosome 6A. These markers define the bounds of the genetic map of chromosome 6A developed in Simmonds *et al.* (2014) with the identification of the 6A TGW QTL. These 67 RILs were identified in a screen of 212 plants performed by James Simmonds, defining an interval of 15.8 cM. The recombination events in individual RILs were defined by the addition of 41 SNP markers across the 15.8 cM interval (details in methods; Figure 3.1a). This identified two linkage blocks within the interval, comprised of 13 (Linkage block 1; Figure 3.1a; circles) and twelve (Linkage block 2; Figure 3.1a; squares) markers each. Linkage block 2 contains the Hap-P2 marker, a marker previously described which maps the position of *TaGW2_A* (Su *et al.*, 2011) a proposed candidate gene for the 6A TGW QTL.

A subset of 41 RILs showing recombination between *BS00003635* and *BS00003835* were grown in five replicated field trials (6m plots 2013-2016, 1m plots 2015; details in Table 3.1) and phenotyped for grain weight and grain morphometric parameters. These markers encompass the introgressed region of the 6A NILs, and hence the interval to which the 6A TGW effect was initially mapped (Simmonds *et al.*, 2014). 38 of these RILs were unambiguously assigned to 13 distinct RIL groups based on their genotype at the 39 markers within the interval between *BS00003635* and *BS00003835* (SR Gr1.1-13; Figure 3.1b). RILs with either the Spark (6A-; S-Control) or Rialto (6A+; R-Control) genotype across the entire interval were selected as controls. Grain width was used as the grain morphometric parameter for mapping as it had previously been defined as the factor underlying the TGW difference in 6A NILs (Chapter 2; Simmonds *et al.*, 2014). Across all five trials and within each trial individually, there were significant differences in grain width observed between RIL groups ($P < 0.001$). Across all trials, the R-Control had 4.18% wider grains than the S-Control, ranging from 2.50% to 5.65% in individual trials, consistent with the grain width differences observed between 6A NILs (Chapter 2). Each RIL group was classified to a parental type (Spark, 6A-; Rialto, 6A+) using Dunnett's tests to both the S- and R-Controls. For example, RIL groups were classified as Spark-like if they were both significantly different to the R-Control and non-significantly different to the S-Control and vice versa. Of the thirteen RIL groups, eleven were unambiguously assigned as either Spark or Rialto-like (Figure 3.1c; grey and orange, respectively). Two of the groups (SR Gr1.2 and SR Gr1.9) were significantly different from both the S- and R- controls and therefore were classified as intermediate types (Figure 3.1c; hatched). Using this method, the grain width was mapped to the 4.6 cM interval between *BS00066522* and *BS00066623* (Figure 3.1a; green markers). The critical RIL groups defining this interval (SR Gr1.3,8,10) are indicated with green arrows in Figure 3.1c. The interval between *BS00066522* and *BS00066623* encompassed 26 additional markers, however 25 of these belong to

Linkage blocks 1 and 2 and notably the interval contained Hap-P2 (the *TaGW2_A* marker). Only two RIL groups had recombination within this interval: SR Gr1.2 (two independent RILs) which has recombination between *BA00363556* and Linkage block 2 (squares) and SR Gr1.9 (one RIL) which has recombination between Linkage block 1 (circles) and *BA00363556*. However, both these groups were classified as intermediate types and therefore could not be used to define the interval further.

Looking at the classification of each of the lines and trials individually (Table 3.1) shows that the classification of lines in both SR Gr1.2 and 1.9 was variable across trials. For example, in SR Gr1.2, SR21 was classified as S in 2013 and 2015 6m plots, SR in 2014 6m and 2015 1m and R in 2016 6m plots. SR Gr1.9 showed a slight tendency towards an S-like classification but was still variable (S in 2013-2015 6m plots, SR in 2015 1m plots and R in 2016 6m plots). Interestingly, no RIL groups were classified as S in the 2016 trial (only SR and R classifications could be assigned), however, reanalysing the data across trials without the 2016 data still resulted in the same overall classification of RIL groups (data not shown). The grain width interval on chromosome 6A could therefore not be defined further than the 4.6 cM interval between *BS00066522* and *BS00066623* due to limited recombination within this RIL population.

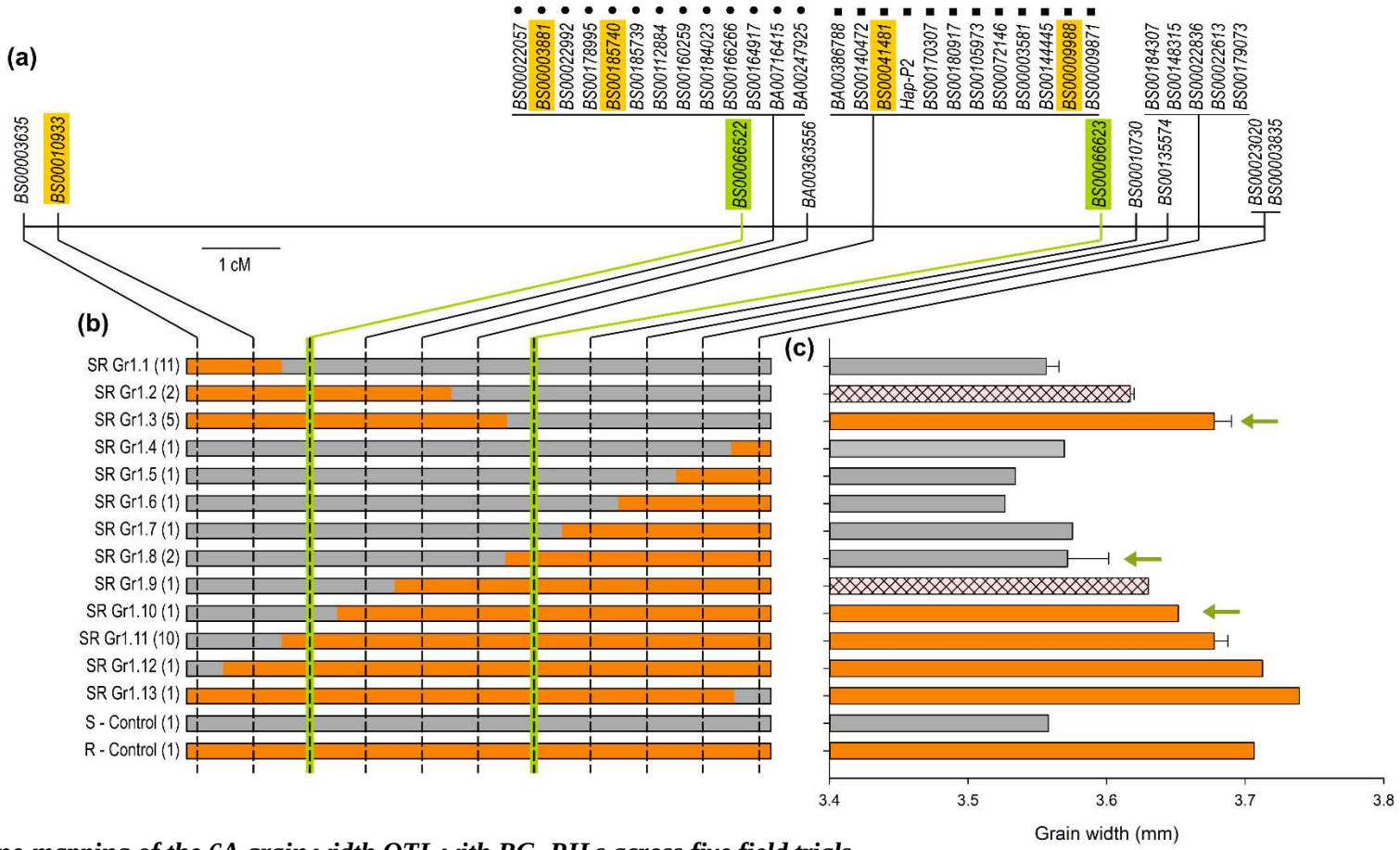


Figure 3.1: Initial fine-mapping of the 6A grain width QTL with BC_4 RILs across five field trials

(a) genetic map of the 6A QTL mapping interval based on the original set of BC_4 recombinant inbred lines (RILs) displayed in (b). Markers highlighted in yellow were used to screen the larger RIL population. Markers highlighted in green are the flanks for the fine-mapped grain width interval defined by this population. Markers with circles or squares adjacent belong to large linkage blocks. (b) Graphical genotypes of RIL groups with the number in brackets indicating the number of independent RILs in each RIL group. RILs were grouped based on genotype defined by having either the Spark-like (grey; 6A-) or Riato-like (orange; 6A+) allele at each marker across the interval. (c) ANOVA adjusted mean grain widths for each RIL group across five field trials (6m plots 2013-2016 and 1m plots 2015). Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as R-control (6A+; orange) or S-control (6A-; grey) like according to Dunnett's test. Hatched bars were classified as intermediate (SR). Green arrows indicate critical RIL groups that define the green highlighted markers as the flanks.

Table 3.1: ANOVA adjusted mean grain width and Dunnett classification of BC_4 RILs used for initial fine-mapping of the 6A grain width QTL

			2013 (6m plots)		2014 (6m plots)		2015 (6m plots)		2015 (1m plots)		2016 (6m plots)		Overall	
RIL group	**Class**	**RIL**	**Width (mm)**	**Class**	**Width (mm)**	**Class**	**Width (mm)**	**Class**	**Width (mm)**	**Class**	**Width (mm)**	**Class**	**Width (mm)**	**Class**
SR Gr1.1	S	SR6	3.555	S	3.878	S	3.274	S	3.394	SR	3.951	SR	3.607	SR
		SR19	-	-	-	-	-	-	3.340	SR	-	-	3.573	S
		SR23	-	-	-	-	-	-	3.348	SR	-	-	3.581	S
		SR25	-	-	-	-	-	-	3.271	S	-	-	3.505	S
		SR35	-	-	-	-	-	-	3.281	S	-	-	3.515	S
		SR36	-	-	-	-	-	-	3.309	S	-	-	3.542	S
		SR38	-	-	-	-	-	-	3.347	SR	-	-	3.580	S
		SR44	-	-	-	-	-	-	3.306	S	-	-	3.539	S
		SR46	-	-	-	-	-	-	3.336	SR	-	-	3.569	S
		SR66	-	-	-	-	-	-	3.334	SR	-	-	3.567	S
		SR67	-	-	-	-	-	-	3.307	S	-	-	3.541	S
SR Gr1.2	SR	SR1	3.595	S	3.848	S	3.333	SR	3.388	SR	3.928	SR	3.620	SR
		SR21	3.505	S	3.890	SR	3.303	S	3.372	SR	3.960	R	3.614	SR
SR Gr1.3	R	SR2	3.683	R	3.914	SR	3.356	R	3.478	R	4.036	R	3.690	R
		SR3	3.625	SR	3.925	SR	3.357	R	3.432	R	4.064	R	3.677	R
		SR4	3.601	S	3.889	SR	3.320	S	3.437	R	4.015	R	3.649	SR
		SR12	3.735	R	3.951	R	3.389	R	3.507	R	4.027	R	3.718	R
		SR13	-	-	-	-	-	-	3.422	R	-	-	3.656	R
SR Gr1.4	S	SR10	3.502	S	3.852	S	3.264	S	-	-	-	-	3.570	S
SR Gr1.5	S	SR63	-	-	-	-	-	-	3.301	S	-	-	3.534	S
SR Gr1.6	S	SR45	-	-	-	-	-	-	3.293	S	-	-	3.527	S
SR Gr1.7	S	SR57	-	-	-	-	-	-	3.342	SR	-	-	3.575	S

Table 3.1 continued on next page

Table 3.1 cont'd from previous page

RIL group	Class	RIL	2013 (6m plots)		2014 (6m plots)		2015 (6m plots)		2015 (1m plots)		2016 (6m plots)		Overall	
			Width (mm)	Class	Width (mm)	Class	Width (mm)	Class	Width (mm)	Class	Width (mm)	Class	Width (mm)	Class
SR Gr1.8	S	SR14	3.572	S	3.844	S	3.298	S	3.370	SR	3.942	SR	3.602	S
		SR54	-	-	-	-	-	-	3.284	S	3.910	SR	3.542	S
SR Gr1.9	SR	SR15	3.599	S	3.879	S	3.304	S	3.422	R	3.955	SR	3.630	SR
SR Gr1.10	R	SR39	-	-	-	-	3.381	R	3.383	SR	3.991	R	3.652	R
SR Gr1.11	R	SR17	3.707	R	3.938	R	3.427	R	3.504	R	4.054	R	3.722	R
		SR22	-	-	-	-	-	-	3.483	R	-	-	3.717	R
		SR24	-	-	-	-	-	-	3.458	R	-	-	3.691	R
		SR27	-	-	-	-	-	-	3.411	SR	-	-	3.644	R
		SR28	-	-	-	-	-	-	3.443	R	-	-	3.677	R
		SR32	-	-	-	-	-	-	3.389	SR	-	-	3.623	SR
		SR51	-	-	-	-	-	-	3.447	R	-	-	3.680	R
		SR52	-	-	-	-	-	-	3.464	R	-	-	3.697	R
		SR55	-	-	-	-	-	-	3.420	R	-	-	3.653	R
		SR58	-	-	-	-	-	-	3.439	R	-	-	3.672	R
SR Gr1.12	R	SR30	3.690	R	3.943	R	3.409	R	-	-	-	-	3.712	R
SR Gr1.13	R	SR9	3.750	R	3.984	R	3.393	R	-	-	-	-	3.739	R
S-Control (6A-)		SR10C	3.521	S	3.860	S	3.252	S	3.302	S	3.873	S	3.558	S
R-Control (6A+)		SR9C	3.720	R	3.956	R	3.427	R	3.447	R	3.997	R	3.707	R

Width (mm) are the ANOVA adjusted means of grain width in each trial (or overall in the final column) each incorporating at least five replicates. Classifications were assigned using Dunnett's test to compare each line to a control (S-Control (6A-; narrow grains) and R-Control (6A+: wide grains)): S = significantly different from the R-Control and not significantly different from the S-Control; R = significantly different from the S-Control and no significantly different from the R-Control; SR = intermediate i.e. not significantly different from both the S-and R-Controls, or significantly different from both the S- and R-Controls. - = data not available (i.e. RIL not grown in trial).

Using the same approach to fine map the 6A TGW effect (as opposed to the grain width effect) resulted in only four of thirteen RIL groups being unambiguously assigned to a parental type (Figure 3.2). Using these four lines, the TGW effect can be positioned between *BS00010933* and *BS00066623*. However, the Dunnett's tests did not identify the S- and R-Controls as significantly different from each other and therefore these results are not reliable. This highlights the importance of mapping using the grain width phenotype due to the increased phenotypic stability compared with TGW. All subsequent genetic mapping was therefore performed using grain width only.

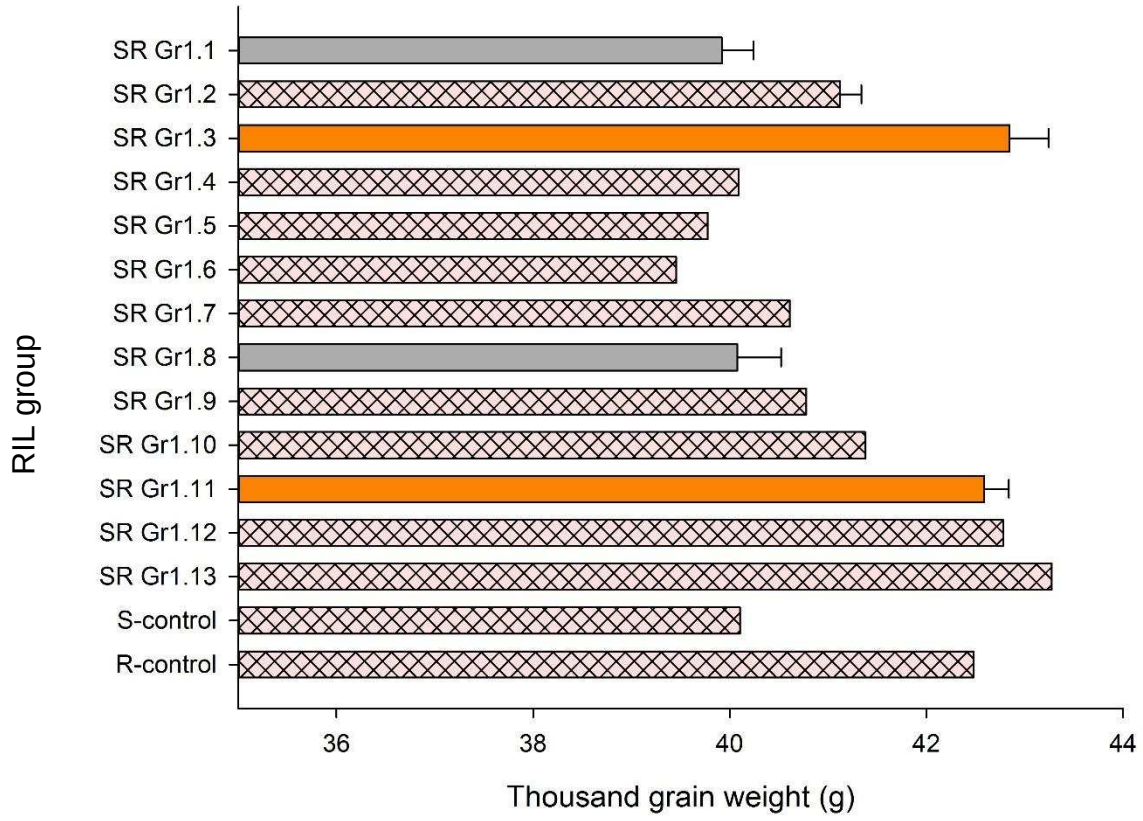


Figure 3.2: ANOVA adjusted mean thousand grain weight of the original 6A BC_4 RIL groups

ANOVA adjusted mean thousand grain weight (TGW) for each 6A RIL group across five field trials (6m plots 2013-2016 and 1m plots 2015). Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as R-Control (6A+; orange) or S-Control (6A-; grey) like according to Dunnett's test. Hatched bars were classified as intermediate (SR).

3.4.1.1 Generation of a larger RIL population to further define the 6A interval

To address the problem of limited recombination between *BS00066522* and *BS00066623* a larger RIL population was generated. In a screen of 2,674 BC_4F_2 plants (performed by James Simmonds), 892 heterozygous recombinants (HetRecs) were identified between *BS00010933* and *BS00066623*, corresponding to an interval of 16.7 cM. The interval between *BS00010933* and *BS00066623* encompasses the fine-mapped grain width interval defined in the previous RIL population.

To prioritise HetRecs for advancement to homozygous recombinants (HomRecs), HetRecs were screened with five additional markers between *BS00010933* and *BS00066623* to define recombination events (Figure 3.3). Priority was given to HetRecs with recombination between the flanking markers of the previously defined grain width mapping interval (*BS00066522* and *BS00066623*) and in particular to HetRecs with recombination between markers in Linkage block 1 or 2 (Figure 3.3a; circles or square, respectively). The additional genotyping showed that in the new RIL population, the grain width mapping interval corresponded to 8.3 cM compared to 4.6 cM in the original RIL population. In total, 224 HetRecs were selected to take forward to homozygous recombinants (HomRecs), with the exact distribution of genotypes shown in parentheses in Figure 3.3a. Each of the selected HetRecs were self-pollinated and twelve progeny were screened to identify HomRecs. For each family (defined as progeny from a single HetRec), at least two HomRecs and a control line (with a single parental allele across all screening markers) were selected where possible. In total, 556 HomRecs (RILs) belonging to 203 independent RIL families were selected, in addition to 26 and 36 independent S-like and R-like controls, respectively (Figure 3.3b).

Whilst grain was collected from RILs in the first generation, these single plants were grown in 96-well trays under glasshouse conditions. This resulted in grain number being compromised and so no reliable grain size phenotype could be obtained. To obtain a more reliable phenotype, all RILs plus six controls (3 x S-Control + 3 x R-Control) were grown in the field in 2016 in single 1m rows replicated in up to three blocks depending on seed availability. From each row, ten individual spikes were harvested at maturity for grain phenotyping.

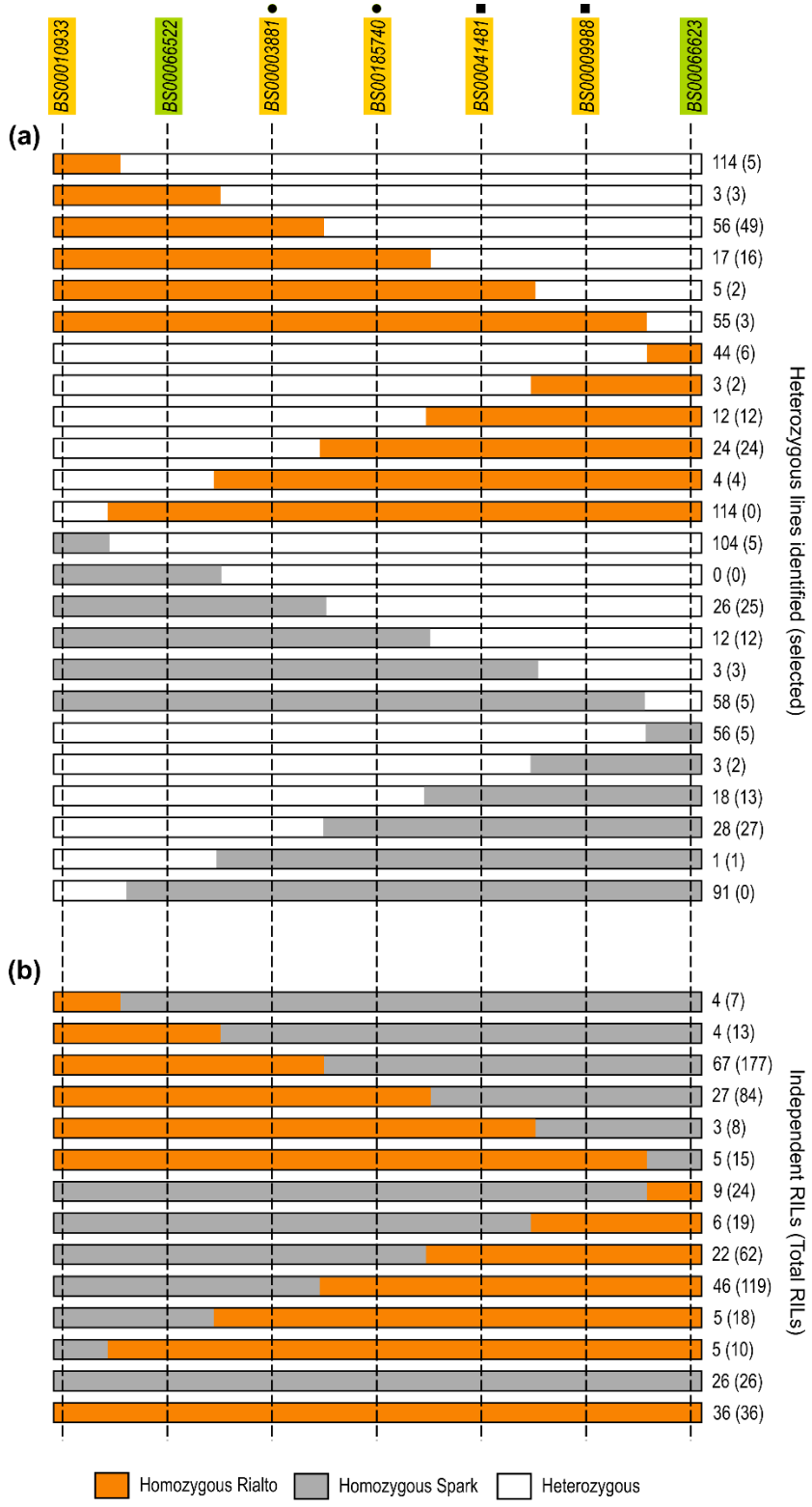


Figure 3.3: Generation of additional BC_4 6A RILs

(a) Graphical genotypes of heterozygous recombinants (HetRecs) identified between *BS00010933* and *BS0006623* from a screen of 2,674 plants. Numbers on the right hand side are the total number of HetRecs identified with each genotype with the number of each group selected to generate homozygous recombinants (recombinant inbred lines; RILs) shown in parentheses. (b) Graphical genotypes of RILs selected after self-pollination of the selected HetRecs. Numbers are the independent RIL families (i.e. the number of HetRec parents) with the total number of RILs with each genotype in parentheses. Markers highlighted in green indicate the flanking markers of the fine-mapped 6A grain width interval defined in Figure 3.1.

3.4.1.1.1 Increasing marker density to prioritise RILs for phenotyping

Due to the large number of RILs, further genotyping was performed to prioritise RILs for grain phenotyping of the 2016 field samples. A representative RIL from each independent RIL family showing recombination between *BS00066522* and *BS00066623* (Figure 3.3b; green markers) were selected and genotyped with an additional 19 SNP markers across the interval to further define the recombination events. Addition of the extra markers revealed that Linkage block 1 (defined in the original RIL population; Figure 3.1a. circles) could be separated into five genetic positions across a 5.7 cM interval in the new RIL population, although some linkage remained at three of these positions (Figure 3.4a, circles). Similarly, Linkage block 2 (Figure 3.1a, squares) could also be separated in the new RIL population although to a lesser extent, with a group of seven markers remaining linked (Figure 3.4a, squares). Based on the more detailed genotypes, a total of 150 RILs from 87 independent RIL families with a distribution of recombination events across the interval between *BS00066522* and *BS00066623* were selected as priority lines for grain phenotyping of the field samples (Figure 3.4b).

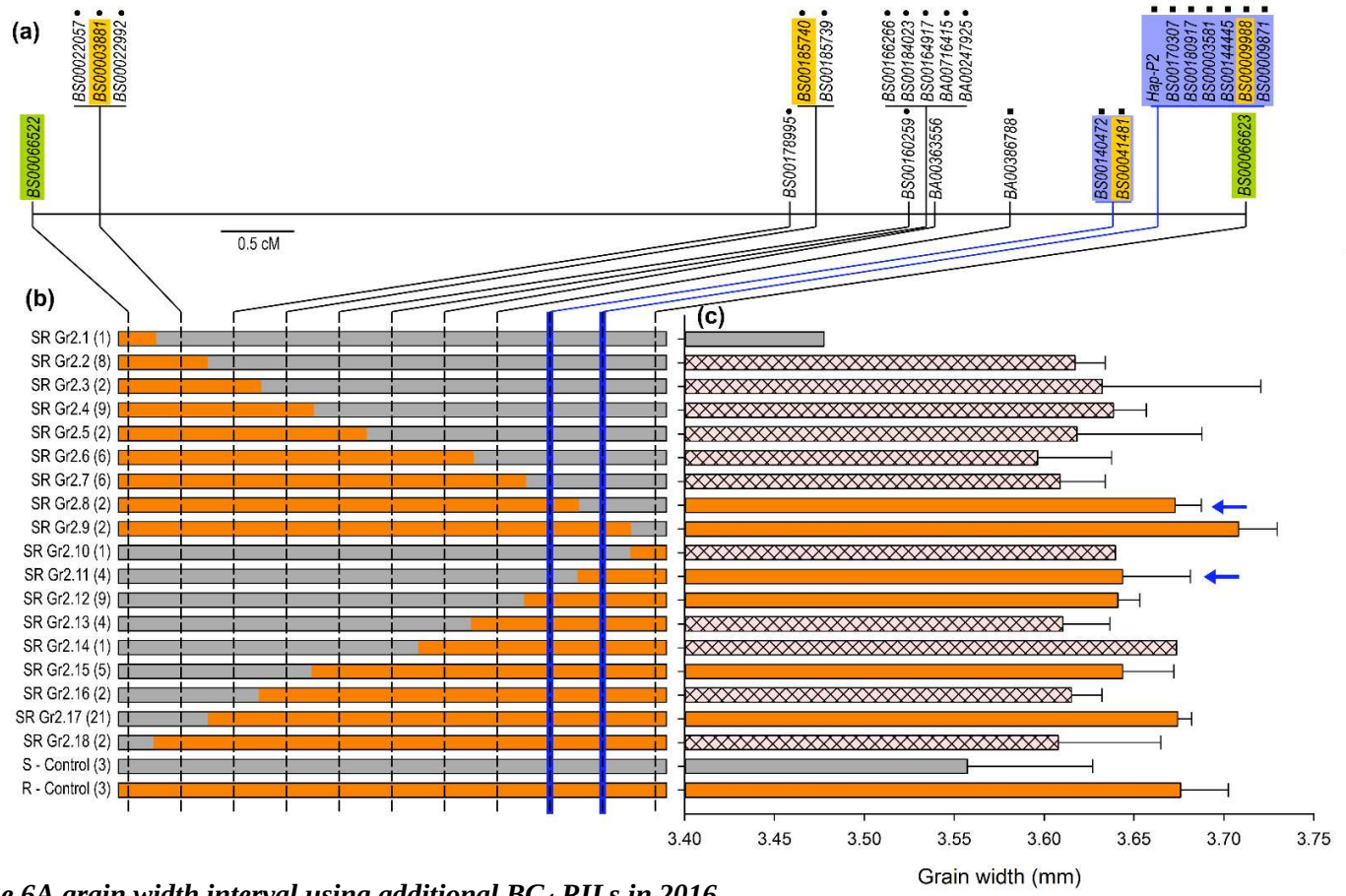


Figure 3.4: Fine mapping of the 6A grain width interval using additional BC_4 RILs in 2016

(a) genetic map of the 6A QTL mapping interval based on the additional BC_4 recombinant inbred lines (RILs) displayed in (b). Markers highlighted in yellow were used to screen lines during the generation of the RILs (Figure 3.3). Markers highlighted in green are the flanking markers for the fine-mapped grain width interval defined by the original RILs and circles/squares indicate markers that were genetically linked in the original RILs (Figure 3.1). Markers highlighted in blue are the flanks for the grain width interval defined by this RIL population. (b) Graphical genotypes of RIL groups with the number in brackets indicating the number of independent RILs in each RIL group. (c) ANOVA adjusted mean grain widths for each RIL group across replicated 1m rows in 2016 field trials. Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as R-control (6A+; orange) or S-control (6A-; grey) like according to Dunnett's test. Hatched bars were classified as intermediate. Blue arrows indicate groups that define the blue highlighted markers as the flanking markers.

3.4.1.1.2 Additional RILs tentatively fine map grain width to a 0.28 cM interval

For each of the 150 RILs, grains from the ten harvested spikes were phenotyped for grain morphometric parameters. Measurements from RILs belonging to the same RIL family were considered as replicates of a single independent RIL i.e. effectively 87 RILs grown in 3 field blocks, with individual RILs providing within block replication. Independent RILs were assigned to 18 RIL groups based on their genotypes across the interval between *BS00066522* and *BS00066623* (Figure 3.4b; number of independent RILs in each group shown in parentheses). Significant differences in grain width were identified between RIL groups (Figure 3.4c). The R-Control group had 3.38% wider grains than the S-Control group, similar to the differences observed in the original RIL population and the 6A NILs (3.4.1.1, Table 2.2). A *post hoc* Dunnett's test was used to classify each RIL group to a parental type, as described previously. Of the 18 RIL groups only seven could be unambiguously assigned to a parental type. The remaining eleven RIL groups were classified as non-significantly different from both the S- and R- control groups and therefore were considered intermediate (SR). However, using just the seven groups that could be assigned to a parental type allowed the grain width phenotype to be mapped to a 0.28 cM interval between two blocks of linked markers (Figure 3.4; highlighted in blue). The left flank of the interval corresponded to two linked markers and the right flank to seven linked markers. In the original RIL population, all nine markers were contained within Linkage block 2 (Figure 3.1a; squares). Notably, the markers in the right flank contain *Hap-P2*, suggesting that the 6A grain width phenotype can be separated from *TaGW2_A*. This interval was considered tentative as it was based on a single year of data and the majority of RIL groups could not be assigned to a parental type. However, if this tentative interval is correct then it would suggest that *TaGW2_A* is not the gene underlying the 6A QTL for grain width.

3.4.1.2 **Determining physical positions of markers across the 6A grain width interval**

Physical positions of the markers across the 6A interval were determined by using BLASTN to align the marker sequences to the latest wheat genome reference sequence: Chinese Spring IWGSC RefSeq v1.0 (Figure 3.5). The physical order of markers according to RefSeq v1.0 agreed with the genetic order of markers according to both RIL populations.

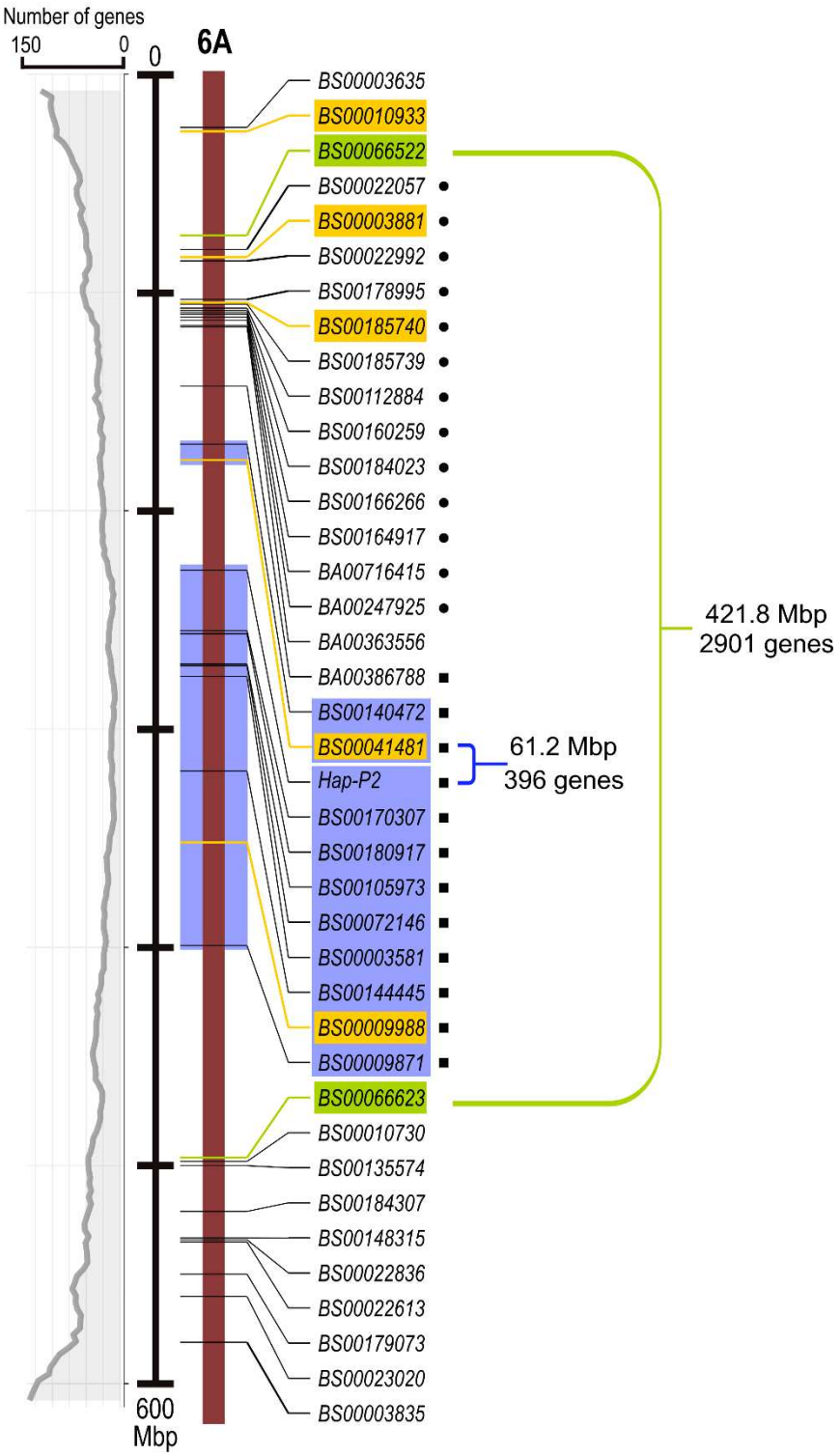


Figure 3.5: Physical positions of markers defining the grain width interval on chromosome 6A

Physical positions of markers on chromosome 6A according to IWGSC RefSeq v1.0. Markers highlighted in green are flanks of the fine-mapped grain width interval defined by the original BC_4 RILs (SR Gr1.1-1.13; Figure 3.1), markers highlighted in yellow were used in generation of additional RILs (Figure 3.3) and markers highlighted in blue are flanks of the grain width interval defined by the additional RILs in 2016 (Figure 3.4). Circles and squares indicate groups of markers that were genetically linked in the original BC_4 RILs (Figure 3.1). Line graph (grey) shows rolling mean of the number of genes located in 3 Mbp bins across chromosome 6A.

The high confidence 4.6 cM grain width interval between *BS00066522* and *BS00066623* corresponded to a 421.8 Mbp region according to RefSeq v1.0 (Figure 3.5; green markers). This interval contained 2,901 TGACv1 gene models based on an *in silico* mapping to RefSeq v1.0 provided by David Swarbreck and Gemy Kaithakottil (Earlham Institute). The expression profile of these genes was analysed using the wheat expVIP expression platform (Borrill *et al.*, 2016). This analysis showed that only 1,972 of the genes were expressed in any RNA-Seq sample in the database (>0.5 transcripts per million (tpm), n = 418) and 1,742 of the genes were expressed in at least one grain RNA-Seq sample (>0.5 tpm, n = 147).

The tentative interval between *BS00041481* and *Hap-P2* considerably reduced the size of the interval to 61.2 Mbp (Figure 3.5; blue markers), containing 396 TGACv1 gene models. Of these 396 genes, 266 were expressed above 0.5 tpm in any RNA-Seq sample in the expVIP expression platform and 233 were expressed in at least one grain RNA-Seq sample. Whilst functional annotations are available for these gene models, the intervals remain too large to begin speculating on any candidate genes based on function.

The physical positions of the markers also provided insights into the genetic architecture underlying the 6A QTL. The two linkage blocks identified in the original RIL population behaved quite differently when assigned physical positions. Linkage block 1 (Figure 3.5, circles), containing 13 markers in the original RIL population, spanned a physical interval of 34.5 Mbp. Those markers that remained linked in the larger RIL population spanned relatively small intervals, ranging from 145 bp – 5.1 Mbp. The relatively close physical proximity of these markers could explain why recombination was limited across this group. Conversely, Linkage block 2 (Figure 3.5, squares) spanned a much larger physical interval of 227.6 Mbp, over a third of the total size of the chromosome 6A pseudomolecule (618 Mbp). The seven markers that remained linked in the large RIL population, including *Hap-P2*, also covered a large distance (159.8 Mbp). This interval is located at the centre of chromosome 6A and appears to cover a relatively gene poor region, suggesting that this interval is centromeric. The limited recombination in this region could therefore be explained by lower rates of recombination in centromeric regions often observed in Triticeae genomes (Akhunov *et al.*, 2003; Mascher *et al.*, 2017).

Overall, the original RIL population enabled the fine-mapping of the 6A grain width effect to a 4.6 cM interval. This corresponded to a 421.8 Mbp interval encompassing a large centromeric linkage block containing the *Hap-P2* marker for *TaGW2_A*. To overcome the issue of limited recombination in this RIL population, a larger RIL population was generated. A single year of field data for the larger population tentatively reduced the interval to 61.2 Mbp and separated the grain width phenotype from the *Hap-P2* marker. The larger RIL population is being grown in field trials in 2017 to obtain a more robust phenotype and the identification of additional markers across the interval will allow further refinement of the mapping position.

3.4.2 Genetic mapping of the 5A QTL for grain length

3.4.2.1 Grain length maps to a 6.6 cM interval on chromosome 5A

A set of 60 BC_4 RILs showing recombination between *gwm293* and *gwm186* were used to fine map the grain length interval on chromosome 5A. These markers were selected as they were used for generation of the 5A NILs and therefore encompass the interval to which the 5A TGW and grain length effect were initially mapped. These 60 RILs were identified from a screen of 170 plants (performed by James Simmonds) defining a genetic distance of 17.65 cM between *gwm293* and *gwm186*.

The genotypes of the 60 RILs were further defined by the addition of 33 SNP markers across the interval between *gwm293* and *gwm186*. Genotyping with these 33 SNP markers defined the recombination events in the 60 RILs and, similar to the 6A interval discussed previously, revealed a linkage block of 14 markers along with several smaller groups of genetically linked markers (Figure 3.6a).

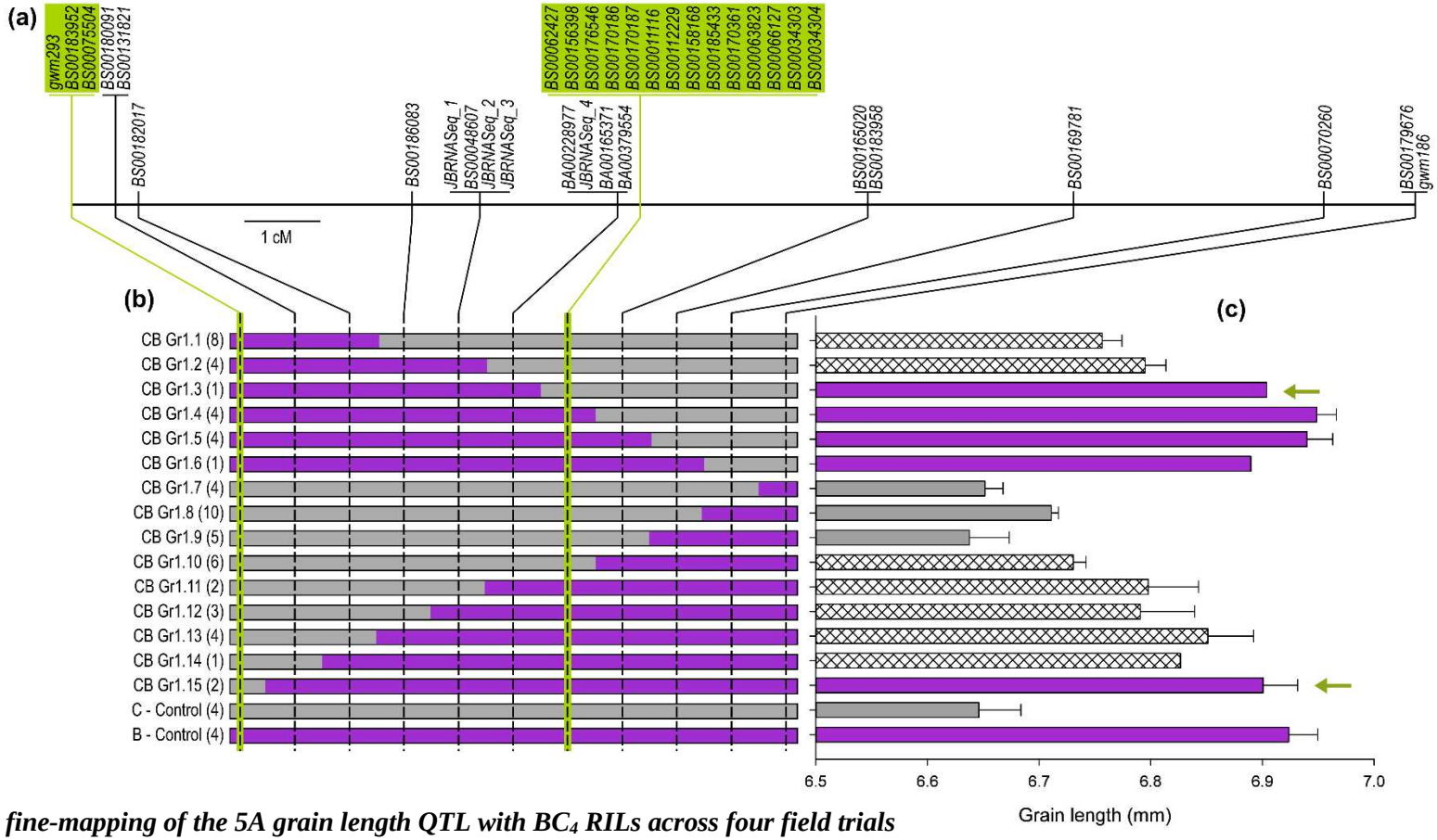


Figure 3.6: Initial fine-mapping of the 5A grain length QTL with BC_4 RILs across four field trials

(a) Genetic map of the 5A QTL mapping interval based on the original set of BC_4 recombinant inbred lines (RILs) displayed in (b). Markers highlighted in green are the flanks for the fine-mapped grain length interval defined by this population. (b) Graphical genotypes of RIL groups with the number in brackets indicating the number of independent RILs in each RIL group. RILs were grouped based on their genotypes defined by having either the Charger-like (grey; 5A-) or Badger-like (purple; 5A+) allele at each marker shown across the interval. (c) ANOVA adjusted mean grain lengths for each RIL group across four field trials (1m plots 2014-2015 and 6m plots 2015-2016). Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as C-control (5A-; grey) or B-control (5A+; purple) like according to a Dunnett's test. Hatched bars were classified as intermediate (CB). Green arrows indicate critical RIL groups that define the green highlighted markers as the flanks.

59 of the 60 RILs were unambiguously assigned to 15 RIL groups based on the genotype of each of the 33 markers across the interval between *gwm293* and *gwm186*, shown graphically in Figure 3.6b. Lines with either the Charger (5A-; C-Control) or Badger allele (5A+; B-Control) across all markers within the whole interval were used as controls. Additionally, in some trials 5A NILs were also used as controls (exact lines used in each trial are detailed in (Table 3.2)). The 59 RILs were grown and phenotyped for grain morphometric parameters across four field trials: 2014-2015 1m plots and 2015-2016 6m plots. As grain length is the main driver of the TGW difference between 5A NILs (Chapter 2; Brinton *et al.*, 2017), the grain length phenotype was used for fine-mapping.

Across trials and within each trial, significant differences in grain length between RILs were identified ($P < 0.001$; Table 3.2). Overall, the B-Control group had 4.18% longer grains than C-Control group (ranging 3.58 – 4.95%), reflective of the grain length differences observed between 5A NILs (4.04%, Table 2.8). In the same way as described previously for the 6A RILs, *post hoc* Dunnett's tests were used to classify RIL groups to a parental type. RIL groups significantly different from the B-Control group and non-significantly different from the C-Control group were classed as Charger-like (5A-, short grains; Figure 3.6c, grey bars). RIL groups significantly different from the C-Control group and non-significantly different from the B-Control group were classed as Badger-like (5A+, long grains; Figure 3.6c, purple bars). RIL groups that did not satisfy both conditions were classed as intermediate (CB; Figure 3.6c, hatched bars). In this way, eight of the RIL groups could be assigned to a parental type, whilst the remaining seven groups were classed as intermediate. The eight groups that could be classed as either Charger or Badger-like defined the grain length effect to a 7.49 cM interval between two groups of linked markers (Figure 3.6; green markers). The left flank included *gwm293* (the original left hand flank of the introgressed interval) and the right hand flank consisted of 14 linked markers.

Six RIL groups had recombination between the ten markers located within the 7.49 cM interval but all six were classed as intermediate and therefore could not be used to further fine map the grain length phenotype. However, looking at the individual RILs that comprised the six intermediate RIL groups showed that individual RILs had a range of classifications within a group, but within each RIL itself the classifications were relatively stable across trials (Table 3.2). In other words, unlike in the initial 6A fine-mapping where the intermediate groups (SR Gr1.2 and 1.9; Table 3.1) consisted of RIL lines that were themselves classed as intermediate, in this case with the 5A RILs intermediate groups were often classed as such because they contained RIL lines that had different classifications. For example, CB Gr1.12 was classed as intermediate (CB) and contained three independent RILs: HR-CB5, HR-CB30 and HR-CB29. HR-CB5 was classed as a B-type across all trials and in each of the four trials individually. HR-CB30 was classed as CB overall and in three of the four trials (B in 2016). Finally, HR-CB29 was classed as C-type overall and in two of the three trials in which it was grown (CB in 2014).

The phenotypic differences within a RIL group could be explained by the individual RILs within a group having different recombination events but the marker density across the interval was not high enough to identify this. For example, all lines in CB Gr1.12 have recombination between *BS00186083* and *JBRNASeq_1*, but the exact location of the recombination may be different in each of the three RILs. Further fine-mapping was therefore performed using the individual RIL lines from RIL groups with recombination across the fine-mapped 7.46 cM interval. Figure 3.7 shows the 14 individual RILs with recombination across the 7.46 cM interval that could be assigned unambiguously as C or B types. Using these 14 RILs, the grain length effect was fine-mapped to a slightly narrower 6.59 cM interval between *BS00182017* and a group of four linked markers (*BA00228977*, *JBRNASeq_4*, *BA00165371*, *BA00379554;* Figure 3.7a, blue markers). Several RILs classified as C or B had recombination within this interval but three of these RILs suggested conflicting mapping positions. HR-CB9 placed the grain length phenotype to the left of *BS00186083*, whilst HR-CB5 and HR-CB58 mapped grain length to the right of *BS00186083*. The grain length phenotype could therefore not be mapped to a narrower interval using this population. It is also worth noting that eleven of the individual RILs were classed themselves as intermediate.

Table 3.2: ANOVA adjusted mean grain length and Dunnett's test classification of individual BC_4 RILs used for initial 5A fine-mapping

			2014 (1m plots)		2015 (1m plots)		2015 (6m plots)		2016 (6m plots)		Overall	
RIL group	**Class**	**RIL**	**Length (mm)**	**Class**	**Length (mm)**	**Class**	**Length (mm)**	**Class**	**Length (mm)**	**Class**	**Length (mm)**	**Class**
CB Gr1.1	CB	HR-CB9	7.055	CB	6.805	CB	-	-	6.764	B	6.817	B
		HR-CB10	7.043	CB	6.832	B	-	-	6.702	CB	6.813	CB
		HR-CB26	7.030	C	6.739	CB	-	-	6.741	B	6.769	CB
		HR-CB14	6.993	C	6.720	CB	-	-	6.729	B	6.757	CB
		HR-CB46	6.992	C	6.762	CB	-	-	6.685	C	6.755	CB
		HR-CB56	6.936	C	6.771	CB	-	-	-	-	6.736	C
		HR-CB11	7.067	B	6.644	C	-	-	6.725	B	6.712	C
		HR-CB23	6.899	C	6.720	CB	-	-	6.552	C	6.666	C
CB Gr1.2	CB	HR-CB13	7.094	B	6.831	CB	6.633	CB	6.790	B	6.832	B
		HR-CB54	7.065	CB	6.857	CB	6.638	CB	6.744	B	6.828	B
		HR-CB2	6.988	C	6.845	CB	6.575	CB	6.670	C	6.771	CB
		HR-CB16	7.019	C	6.710	C	6.641	CB	6.725	B	6.762	CB
CB Gr1.3	B	HR-CB24	7.195	B	6.912	B	6.706	B	6.839	B	6.904	B
CB Gr1.4	B	HR-CB43	7.201	B	7.002	B	6.824	B	-	-	6.985	B
		HR-CB28	7.180	B	6.866	B	6.754	B	-	-	6.901	B
		HR-CB35	7.191	B	6.995	B	6.760	B	-	-	6.956	B
		HR-CB20	7.189	B	-	-	-	-	-	-	6.937	B
CB Gr1.5	B	HR-CB53	7.259	B	-	-	-	-	-	-	6.995	B
		HR-CB12	7.221	B	-	-	-	-	-	-	6.969	B
		HR-CB17	7.176	B	-	-	-	-	-	-	6.924	B
		HR-CB57	7.144	B	-	-	-	-	-	-	6.892	B
CB Gr1.6	B	HR-CB50	7.142	B	-	-	-	-	-	-	6.890	B

Table 3.2 continued on next page

Table 3.2 cont'd from previous page

RIL group	Class	RIL	2014 (1m plots)		2015 (1m plots)		2015 (6m plots)		2016 (6m plots)		Overall	
			Length (mm)	Class	Length (mm)	Class	Length (mm)	Class	Length (mm)	Class	Length (mm)	Class
CB Gr1.7	C	HR-CB21	6.934	C	-	-	-	-	-		6.682	C
		HR-CB34	6.925	C	-	-	-	-	-	-	6.673	C
		HR-CB31	6.898	C	-	-	-	-	-	-	6.634	C
		HR-CB1	6.865	C	-	-	-	-	-	-	6.613	C
CB Gr1.8	C	HR-CB33	7.004	C	-	-	-	-	-	-	6.752	CB
		HR-CB45	6.985	C	-	-	-	-	-	-	6.733	CB
		HR-CB39	6.986	C	-	-	-	-	-	-	6.721	CB
		HR-CB40	6.970	C	-	-	-	-	-	-	6.718	CB
		HR-CB37	6.960	C	-	-	-	-	-	-	6.708	CB
		HR-CB47	6.959	C	-	-	-	-	-	-	6.707	CB
		HR-CB41	6.956	C	-	-	-	-	-	-	6.704	C
		HR-CB36	6.955	C	-	-	-	-	-	-	6.703	C
		HR-CB42	6.938	C	-	-	-	-	-	-	6.686	C
		HR-CB32	6.938	C	-	-	-	-	-	-	6.686	C
CB Gr1.9	C	HR-CB19	7.026	CB	-	-	-	-	-	-	6.774	CB
		HR-CB3	6.872	C	-	-	-	-	-	-	6.620	C
		HR-CB60	6.883	C	-	-	-	-	-	-	6.618	C
		HR-CB59	6.857	C	-	-	-	-	-	-	6.605	C
		HR-CB52	6.819	C	-	-	-	-	-	-	6.567	C
CB Gr1.10	CB	HR-CB27	6.977	C	6.806	CB	6.582	CB	-	-	6.765	CB
		HR-CB55	7.017	C	6.722	CB	6.590	CB	-	-	6.743	CB
		HR-CB8	6.964	C	-	-	-	-	-	-	6.711	CB
		HR-CB15	7.010	C	6.704	C	6.518	C	-	-	6.711	C
		HR-CB22	6.981	C	6.642	C	6.593	CB	-	-	6.703	C
		HR-CB6	6.955	C	-	-	-	-	-	-	6.690	CB

Table 3.2 continued on next page

Table 3.2 cont'd from previous page

RIL group	Class	RIL	2014 (1m plots)		2015 (1m plots)		2015 (6m plots)		2016 (6m plots)		Overall	
			Length (mm)	Class	Length (mm)	Class	Length (mm)	Class	Length (mm)	Class	Length (mm)	Class
CB Gr1.11	CB	HR-CB58	7.031	CB	6.876	B	-	-	6.744	B	6.847	B
		HR-CB25	7.043	CB	6.766	CB	6.575	CB	6.647	C	6.758	CB
CB Gr1.12	CB	HR-CB5	7.138	B	6.953	B	6.681	B	6.746	B	6.876	B
		HR-CB30	7.028	CB	6.788	CB	6.588	CB	6.711	B	6.782	CB
		HR-CB29	7.050	CB	6.676	C	6.518	C	-	-	6.709	C
CB Gr1.13	CB	HR-CB7	7.150	B	7.016	B	-	-	6.744	B	6.912	B
		HR-CB44	7.077	B	6.943	B	-	-	6.823	B	6.898	B
		HR-CB38	7.118	B	6.932	B	-	-	6.807	B	6.897	B
		HR-CB18	7.012	C	6.721	CB	-	-	6.683	C	6.739	CB
CB Gr1.14	CB	HR-CB4	7.079	B	-	-	-	-	-	-	6.827	CB
CB Gr1.15	B	HR-CB48	7.128	B	-	-	-	-	-	-	6.864	CB
		HR-CB51	7.178	B	-	-	-	-	-	-	6.926	B
C-control (5A-)	C	HR-CB37-C	6.996	C	6.745	C	6.612	C	-	-	6.759	C
		HR-CB7-C	6.746	C	6.684	C	6.430	C	-	-	6.616	C
		BC_4-5 (NIL)	-		6.551	C	6.419	C	6.597	C	6.603	C
		BC_4-17 (NIL)	6.895	C	6.559	C	6.434	C	6.564	C	6.607	C
B-control (5A+)	B	HR-CB9-C	7.242	B	6.981	B	6.848	B	-	-	6.996	B
		HR-CB38-C	7.202	B	6.892	B	6.780	B	-	-	6.924	B
		BC_4-6 (NIL)	-	-	6.873	B	6.746	B	6.819	B	6.893	B
		BC_4-19 (NIL)	7.213	B	6.824	B	6.713	B	6.814	B	6.880	B

Length (mm) are the ANOVA adjusted means of grain length in each trial (or overall in the final column) each incorporating at least five replicates. Classifications were assigned using Dunnett's test to compare each line to a control (C-Control (5A-; short grains) and B-Control (5A+: long grains)): C = significantly different from the B-Control and not significantly different from the C-Control; B = significantly different from the C-Control and no significantly different from the B-Control; CB = not significantly different from the C-Control or the B-Control.

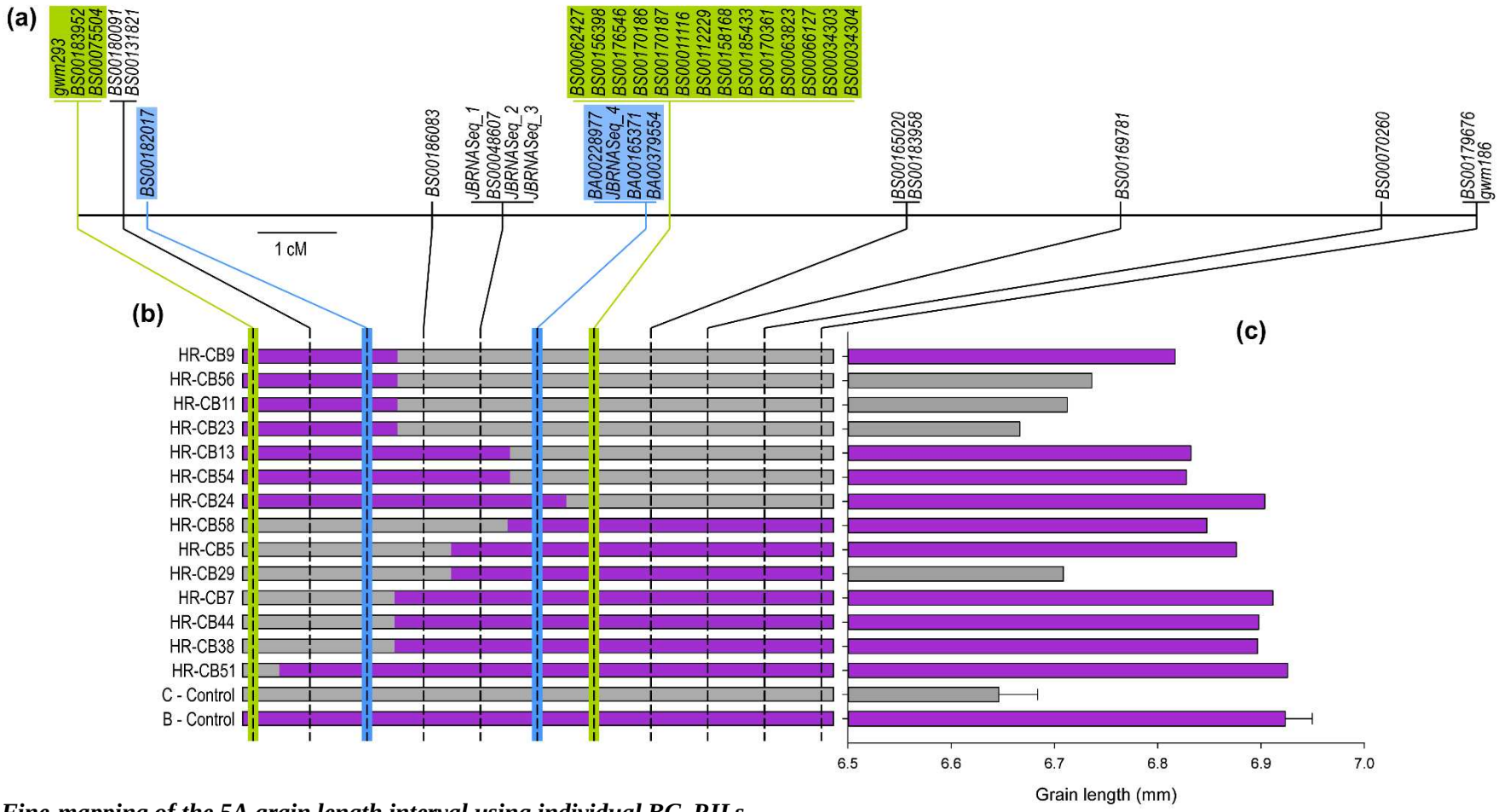


Figure 3.7: Fine-mapping of the 5A grain length interval using individual BC_4 RILs

(a) Genetic map of the 5A QTL mapping interval based on the original set of BC_4 recombinant inbred lines (RILs). Markers highlighted in green are the flanks for the fine-mapped grain length interval defined by mapping with RIL groups. Markers highlighted in blue are the flanks for the grain length interval defined using the individual RILs in (b). (b) Graphical genotypes of RILs showing the allele at each marker (Charger-like (grey; 5A-) or Badger-like (purple; 5A+)). (c) ANOVA adjusted mean grain lengths for each RIL across four field trials (1m plots 2014-2015 and 6m plots 2015-2016). Bars are coloured according to classification as C-Control (5A-; grey) or B-Control (5A+; purple) like according to a Dunnett's test. Error bars are standard error of lines within the control groups (n = 4)/

3.4.2.2 **Generation of a larger 5A RIL population**

As with the 6A QTL, a larger RIL population was generated to further map the 5A grain length QTL. In an initial screen of 1,140 BC_4F_3 plants performed by James Simmonds, 310 HetRecs were identified between *BS00075504* and *BS00183958*, defining a genetic distance of 13.60 cM (10.48 cM in the original RIL population described above). These markers encompassed the initial 7.49 cM fine-mapped grain length interval defined using the original RIL population. The genotypes of the HetRecs were defined further by the addition of seven markers across the interval (Figure 3.8a). All 310 HetRec plants were self-pollinated and twelve progeny of each were screened to identify HomRecs. Of the progeny originating from a single HetRec (i.e. independent RIL family) at least two HomRecs (RILs) and a control line (a single parental type across the whole interval) were selected where possible (Figure 3.8b). In total, 558 individual RILs from 272 independent RIL families with recombination across the interval between *BS00075504* and *BS00183958* were selected. In addition, 59 C and 64 B-Control lines were selected. All 558 RILs and six control lines (3 x C-Control, 3 x B-Control) were grown in field trials in 2016 in single 1m rows replicated up to five times depending on seed availability.

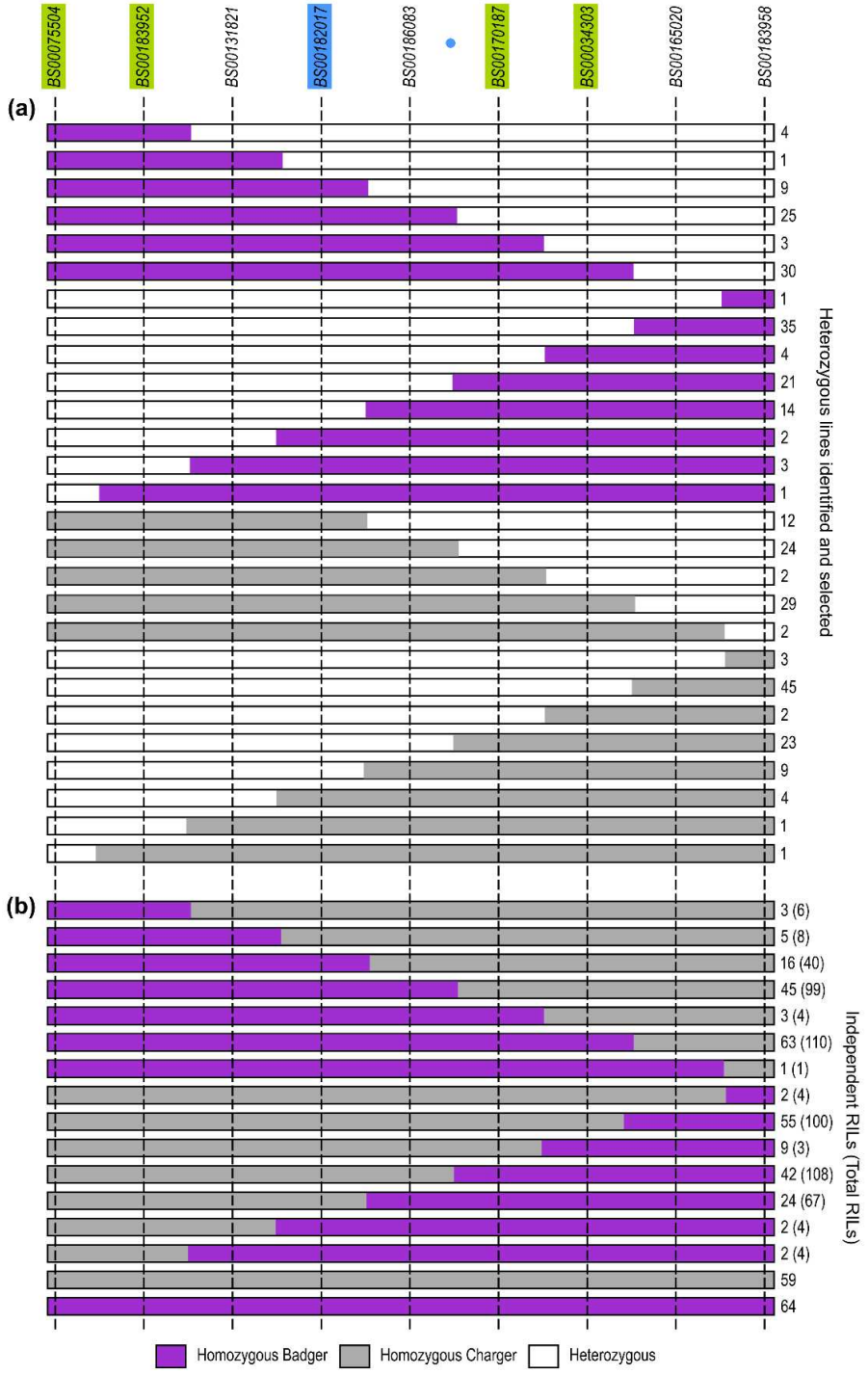


Figure 3.8: Generation of the larger 5A RIL population

(a) Graphical genotypes of heterozygous recombinants (HetRecs) identified between *BS00075504* and *BS00183958* from a screen of 1140 plants. Numbers on the right hand side are the total number of HetRecs identified with each genotype. All HetRecs were selected to generate homozygous recombinants (recombinant inbred lines; RILs). (b) Graphical genotypes of RILs selected after self-pollination of the selected HetRecs. Numbers are the independent RIL families (i.e. the number of HetRec parents) with the total number of RILs with each genotype in parentheses. Markers highlighted in green indicate the flanking markers of the initial 5A grain length interval defined with original RIL groups (Figure 3.6). The blue marker indicates the left flank of the narrower grain length interval. The group of right flanking markers (*BA00228977*, *JBRNASeq_4*, *BA00165371*, *BA00379554*) was not run but is shown as a blue circle based on its genetic position in the original RIL population (Figure 3.7).

3.4.2.2.1 Increasing marker density in a subset of the larger 5A RIL population

A subset of 290 RILs from 114 RIL families with recombination between *BS00182017* and *BS00170187* were selected for further genotyping. The subset was genotyped using eight additional markers within the interval not used for the screening of the larger RIL population. This genotyping showed that in this RIL population the linkage had been broken between the four markers that defined the right flank of the 6.6 cM grain length interval in the original RIL population (*BA00228977*, *JBRNASeq_4*, *BA00165371*, *BA00379554*; Figure 3.9a). The linkage between *JBRNASeq_1*, *BS00048607*, *JBRNASeq_4*, *JBRNASeq_3* was also partially broken, with two independent RIL families having recombination between *JBRNASeq_1* and the other three markers (0.11 cM). However *BS00048607*, *JBRNASeq_4*, *JBRNASeq_3* remained linked.

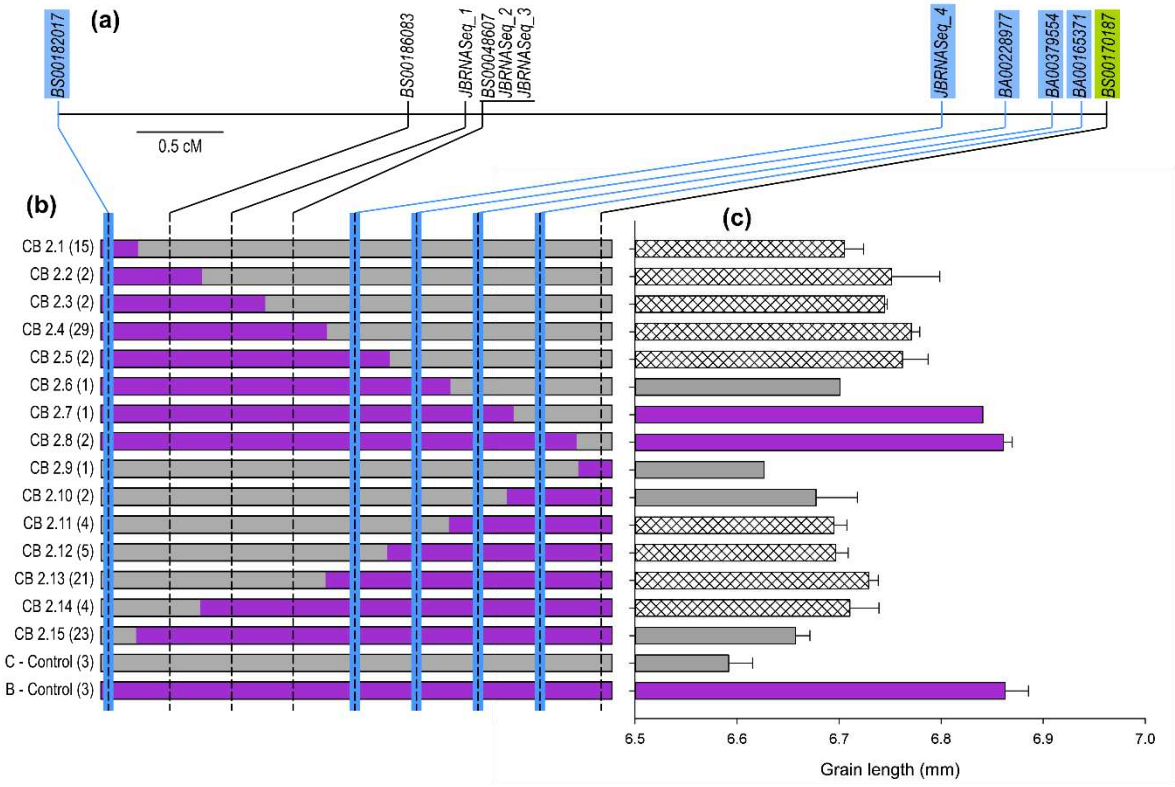


Figure 3.9: Initial fine-mapping of grain length using the larger 5A RIL population

a) Genetic map of the 5A QTL mapping interval based on the additional BC_4 recombinant inbred lines (RILs) displayed in (b). Green marker: flank of the 7.49 cM grain length interval defined in Figure 3.6, blue markers: flanks of the 6.6 cM grain length interval defined in Figure 3.7. (b) Graphical genotypes of RIL groups with the number in brackets indicating the number of independent RILs in each RIL group. (c) ANOVA adjusted mean grain length of each RIL group across replicated 1m rows in 2016 field trials. Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as C-Control (5A-; grey) or B-Control (5A+; purple) like according to Dunnett's test. Hatched bars were classified as intermediate (CB).

3.4.2.2.2 Fine-mapping with the larger 5A RIL population suggests conflicting mapping positions

The same subset of 290 RILs and six control lines were selected for assessment of grain morphometric parameters from the 1m row field samples collected in 2016. As with the original 5A RIL population, grain length was used as the parameter for fine-mapping. Independent RIL families were grouped into 15 distinct RIL groups based on their genotype across the interval between *BS00182017* and *BS00170187* (Figure 3.9b). Individual RILs within each independent RIL family were considered as replicates of the same RIL. Significant differences in grain length were observed between RIL groups ($P < 0.001$). The B-Control group had 4.11% longer grains than the C-Control group, similar to the differences observed between controls in the original 5A RIL population and between 5A NILs. RIL groups were again classified as either C, B or intermediate type using a *post hoc* Dunnett's test as described previously. In this way, six of the 15 groups were assigned unambiguously as a Charger or Badger type whilst the remaining nine groups were classed as intermediate (Figure 3.9c, hatched bars). However, the grain length interval, previously mapped to between *BS00182017* and the linkage block containing *JBRNASeq_4*, *BA00228977*, *BA00165371* and *BA00379554* (Figure 3.7a, blue markers) could not be defined further using the six C or B RIL groups. As with the original 5A RIL population, different RIL groups suggested that the grain length phenotype mapped to either side of *BS00186083*. Five of the RIL groups (CB2.6-2.10) positioned the grain length phenotype between *BA00228977* and *BA00165371*, to the right of *BS00186083* (Figure 3.9). However, CB 2.15 suggested that the grain length phenotype mapped to the left *BS00186083* and although this was only a single RIL group, the group contained 23 independent RILs compared to a single RIL in CB2.6, the RIL group that did not support this position.

Similar to the original 5A RIL population, individual RILs within the intermediate RIL groups had a range of classifications (Table 3.3). All RIL groups were therefore divided into sub-groups based on the phenotype call of each RIL within the group. For example, RIL group CB 2.13 contained 21 independent RILs and was classified as intermediate (CB) overall. However, when looking at each of the 21 RILs individually, six were classed as C, seven as B and eight as CB. RIL group CB 2.13 was therefore split into three sub-groups, CB 2.13-C, CB 2.13-B and CB 2.13-CB, containing six, seven and eight RILs, respectively. In this way RILs were categorised into a total of 30 RIL sub-groups, twelve of which were intermediate (CB).

Table 3.3: ANOVA adjusted mean grain length and class of individual RILS in the larger 5A RIL population used for fine-mapping

Overall group	Call	RIL	N (sibling RILs)	Length (mm)	Class
CB Gr2.1	CB	CB107	2	6.668	C
		CB132	1	6.673	C
		CB143	3	6.636	C
		CB208	3	6.589	C
		CB221	2	6.715	C
		CB23	2	6.585	C
		CB266	1	6.631	C
		CB35	2	6.596	C
		CB80	2	6.665	C
		CB280	2	6.732	CB
		CB331	3	6.720	CB
		CB378	1	6.706	CB
		CB156	5	6.767	B
		CB239	6	6.832	B
		CB70	3	6.777	B
CB Gr2.2	CB	CB295	3	6.725	CB
		CB212	1	6.820	B
CB Gr2.3	CB	CB121	3	6.742	CB
		CB30	3	6.747	CB
CB Gr2.4 (cont'd on next page)	CB	CB152	1	6.654	C
		CB235	2	6.688	C
		CB351	2	6.704	C
		CB14	1	6.731	CB
		CB273	3	6.732	CB
		CB28	1	6.715	CB
		CB31	4	6.748	CB
		CB374	3	6.753	CB
		CB47	3	6.752	CB
		CB117	1	6.766	B
		CB157	2	6.797	B
		CB161	4	6.769	B
		CB202	3	6.769	B
		CB210	2	6.760	B
		CB215	2	6.792	B
		CB216	1	6.754	B
		CB317	1	6.841	B
		CB328	4	6.810	B
		CB345	2	6.770	B
		CB347	2	6.748	B
		CB43	2	6.764	B
		CB59	2	6.795	B
		CB61	1	6.812	B
		CB73	1	6.844	B

Table 3.3 cont'd on next page

Table 3.3 cont'd from previous page

Overall group	Call	RIL	N (sibling RILs)	Length (mm)	Class
CB Gr2.4 (cont'd from previous page)	CB	CB75	3	6.760	B
		CB78	2	6.826	B
		CB82	5	6.851	B
		CB92	4	6.768	B
		CB96	3	6.774	B
CB Gr2.5	CB	CB277	1	6.725	CB
		CB381	3	6.775	B
CB Gr2.6	CC	CB324	1	6.701	CB
CB Gr2.7	BB	CB118	4	6.841	B
CB Gr2.8	BB	CB243	2	6.871	B
		CB326	3	6.854	B
CB Gr2.9	CC	CB3	6	6.626	C
CB Gr2.10	CC	CB181	4	6.644	C
		CB370	3	6.725	CB
CB Gr2.11	CB	CB12	2	6.692	C
		CB338	1	6.657	C
		CB159	1	6.719	CB
		CB271	4	6.701	CB
CB Gr2.12	CB	CB214	2	6.666	C
		CB293	4	6.675	C
		CB318	3	6.678	C
		CB45	3	6.725	CB
		CB46	4	6.720	CB
CB Gr2.13	CB	CB130	3	6.697	C
		CB142	1	6.693	C
		CB147	3	6.673	C
		CB21	2	6.712	C
		CB267	1	6.680	C
		CB274	1	6.650	C
		CB135	1	6.729	CB
		CB144	6	6.730	CB
		CB160	5	6.718	CB
		CB183	1	6.736	CB
		CB291	4	6.701	CB
		CB298	4	6.702	CB
		CB40	2	6.733	CB
		CB91	5	6.745	CB
		CB166	1	6.749	B
		CB191	4	6.818	B
		CB300	1	6.779	B
		CB302	1	6.787	B
		CB49	2	6.751	B
		CB66	1	6.800	B
		CB71	1	6.747	B

Table 3.3 cont'd on next page

Table 3.3 cont'd from previous page

Overall group	Call	RIL	N (sibling RILs)	Length (mm)	Class
CB Gr2.14	CB	CB279	3	6.664	C
		CB366	1	6.694	C
		CB136	3	6.727	CB
		CB1	1	6.796	B
CB Gr2.15	CC	CB112	2	6.534	C
		CB139	4	6.650	C
		CB173	2	6.662	C
		CB174	4	6.525	C
		CB192	4	6.660	C
		CB275	2	6.708	C
		CB286	4	6.531	C
		CB325	1	6.561	C
		CB327	2	6.636	C
		CB329	5	6.688	C
		CB333	3	6.658	C
		CB353	2	6.597	C
		CB380	2	6.613	C
		CB41	3	6.593	C
		CB51	3	6.654	C
		CB54	1	6.658	C
		CB57	3	6.682	C
		CB11	4	6.699	CB
		CB169	1	6.707	CB
		CB190	2	6.715	CB
		CB304	6	6.730	CB
		CB330	2	6.727	CB
		CB27	4	6.769	B
C-Control	CC	CB243C	1	6.636	C
		CB351C	1	6.584	C
		CB202C	1	6.556	C
B-Control	BB	CB140C	1	6.908	B
		CB43C	1	6.845	B
		CB80C	1	6.835	B

Length (mm) are the ANOVA adjusted means of grain length. Classifications were assigned using Dunnett's test to compare each line to a control (C-Control (5A-; short grains) and B-Control (5A+: long grains)): C = significantly different from the B-Control and not significantly different from the C-Control; B = significantly different from the C-Control and no significantly different from the B-Control; CB = intermediate i.e. not significantly different from the C-Control or the B-Control, or significantly different from both.

The 18 -C and -B sub-groups were used to further fine map the grain length effect (Figure 3.10). Eleven of the sub-groups mapped grain length to the same interval as in the original RIL population, between *BS00182017* and *JBRNASeq_4* (5.10 cM, Figure 3.10, orange highlighted groups). However, sub-groups with additional recombination in this interval again produced the problem of two conflicting mapping positions. Using four sub-groups (CB2.1-B, CB2.2-B, CB2.14-C and CB2.15-C; Figure 3.10, yellow highlighted groups) along with the eleven orange sub-groups mapped grain length to a 2.02 cM interval between *BS00182017* and *BS00186083*. However, using three other sub-groups (CB2.4-C, CB2.13-B and CB2.14-B; Figure 3.10, pink highlighted groups) along with the eleven orange sub-groups mapped grain length to a 2.65 cM interval between *BS00048607, JBRNASeq_2, JBRNASeq_3* (all linked) and *JBRNASeq_4*.

One explanation for the conflicting mapping results could be that they are based on a single year of data, and so additional year datasets could provide further support to one of the mapping positions. However, this phenomenon was also observed in fine-mapping with the original 5A RIL population which was assessed in four trials across three different years. An alternative explanation (the 'two-gene' hypothesis) could be that both mapping positions are correct and that there are two genes within the interval between *BS00182017* and *JBRNASeq_4* that contribute additively to final grain length.

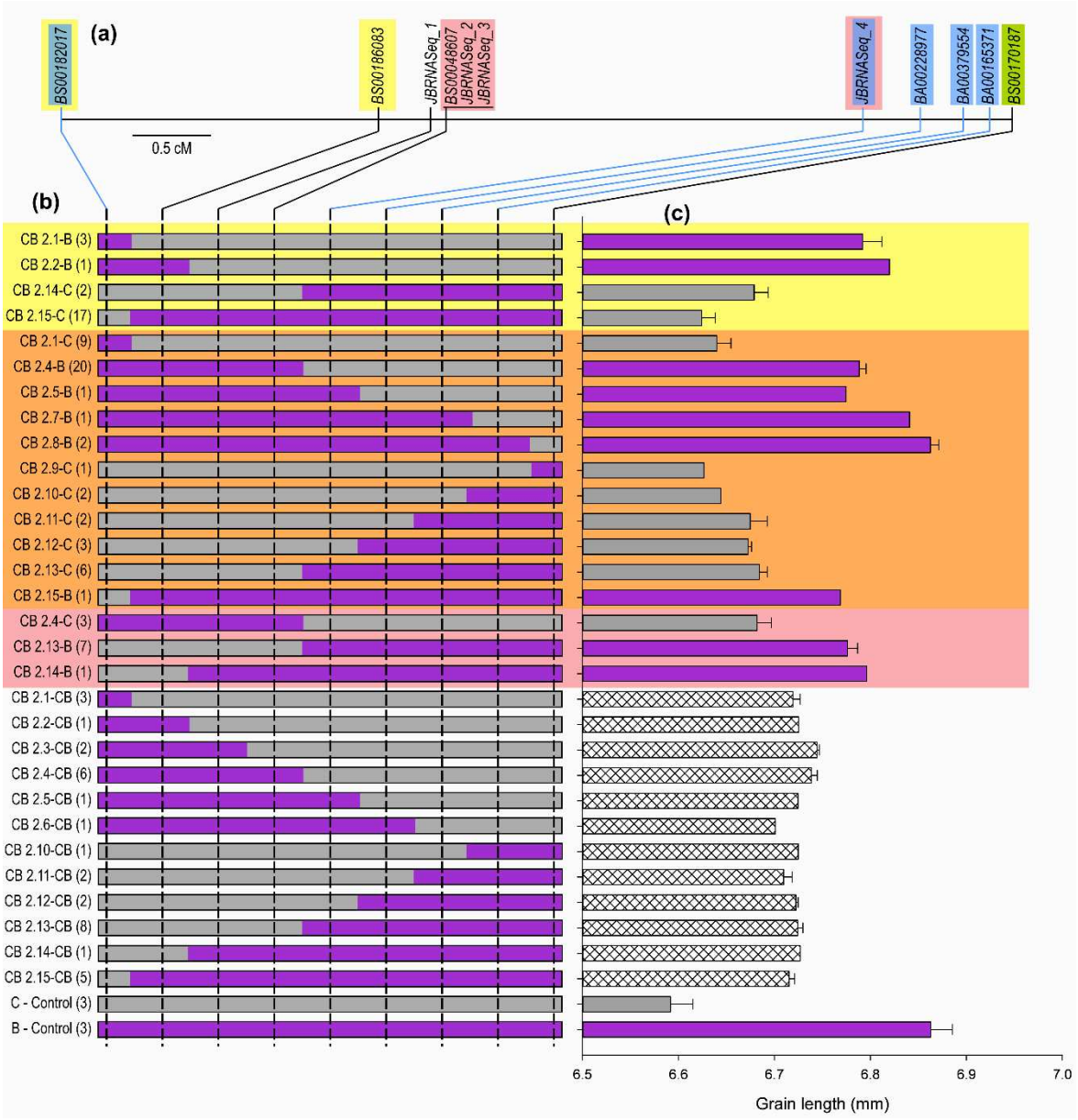


Figure 3.10: Fine-mapping of the 5A grain length effect using RIL sub-groups

a) Genetic map of the 5A QTL mapping interval based on the larger BC_4 recombinant inbred line (RIL) population. Markers highlighted in blue are the flanking markers of the 5A grain length interval. Pairs of markers highlight in yellow and pink show the two conflicting grain length mapping positions within this interval. (b) Graphical genotypes of RIL sub-groups with the number in brackets indicating the number of independent RILs in each RIL group. (c) ANOVA adjusted mean grain length of each RIL group across replicated 1m rows in 2016 field trials. Error bars are the standard error of all lines within the RIL group. Bars are coloured according to classification as C-Control (5A-; grey) or B-Control (5A+; purple) like according to Dunnett's test. Hatched bars were classified as intermediate (CB). In (b) and (c) groups that are highlighted in yellow support the mapping position between the two yellow markers, groups highlighted in pink support the mapping position between the pink markers. Groups highlighted in orange support either position.

3.4.2.3 **The ‘two-gene’ hypothesis**

If the ‘two-gene hypothesis’ is correct then one would expect that RILs showing recombination between the two genes/loci would have an intermediate grain length phenotype compared to lines with both the Charger or Badger alleles of each gene. The fact that many RILs with recombination between *BS00182017* and *JBRNASeq_4* in both the original and larger RIL populations could not be allocated unambiguously to a parental type supports this hypothesis.

To examine this more explicitly, RILs in the larger population were allocated to groups according to the genotype of each of the four markers defining the two internal mapping regions as defined by fine-mapping with this population (Region 1: *BS00182017* and *BS00186083*, Figure 3.11a-b, yellow markers; Region 2: *BS00048607*/*JBRNASeq_2/JBRNASeq_3* and *JBRNASeq_4*, Figure 3.11a-b, pink markers). For example, RILs with the Charger allele at all four markers were classed CCCC whereas RILs with the Charger allele at the left flank of Region 1 *(BS00182017*) and Badger at the other three markers were classed as CBBB. Significant differences in grain length were observed between groups ($P < 0.001$). A Dunnett’s test using the CCCC and BBBB groups as controls was used to categorise each group phenotypically. Four of the six groups with internal recombination were classed as intermediate, as predicted by the ‘two-gene’ hypothesis. However, two of the groups were classed as similar to the CCCC group.

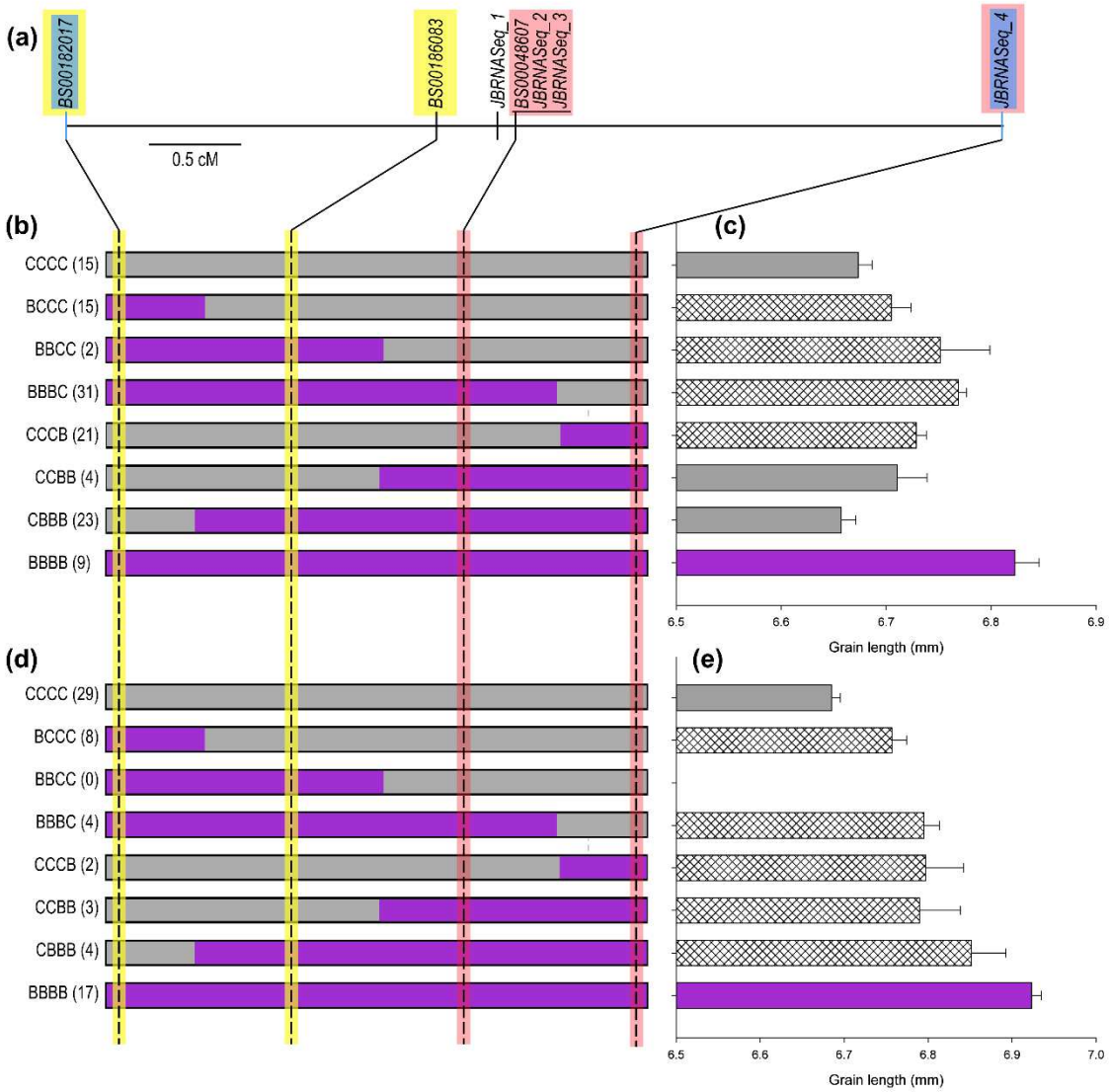


Figure 3.11: Genotype groups of RILs according to the 'two-gene' hypothesis

(a) Genetic map across the 5A grain length interval according to the larger 5A RIL population. Blue markers show the overall grain length interval defined in Figure 3.7, yellow markers define Region 1 and pink markers define Region 2. (b) Graphical genotypes of the larger 5A RIL population grouped according to the four markers that define the two internal mapping regions. Numbers in brackets are the number of independent RILs in each group. (c) ANOVA adjusted mean grain length of each genotype group across replicated 1m rows in 2016 field trials. Error bars are standard error of RILs within the group. Bars are coloured according to classification as CCCC (5A-; grey) or BBBB (5A+; purple) like according to Dunnett's test. Hatched bars are intermediate. (d) Graphical genotypes of the original 5A RIL population grouped according to the four markers that define the two internal mapping regions. Numbers in brackets are the number of independent RILs in each group. (e) ANOVA adjusted mean grain length of each genotype group across four field trials (1m plots 2014-2015 and 6m plots 2015-2016). Error bars are standard error of RILs within the group. Bars are coloured according to classification as CCCC (5A-; grey) or BBBB (5A+; purple) like according to Dunnett's test. Hatched bars are intermediate.

Testing this hypothesis using the same population and data that was used to generate it could be seen as a somewhat circular argument. To test the hypothesis in a more independent manner, the same analysis was performed using the original RIL population. RILs could be allocated to seven of eight possible groups; no RILs in the original population belonged to the BBCC group. Again, significant differences in grain length were observed between groups ($P < 0.001$). In this population, Dunnett's tests using CCCC and BBBB groups as controls classed all groups with internal recombination as having intermediate grain length phenotypes.

We designated the gene located in Region 1 as *Grain Length 1* (*GL1*) and the gene located in Region 2 as *Grain Length 2* (*GL2*). The Badger alleles (5A+; long grains) are indicated using uppercase letters (*GL1/GL2*) and the Charger alleles (5A-; short grains) are indicated using lowercase letters (*gl1/gl2*). The classifications described above based on the four markers defining Region 1 and Region 2 resulted in eight genotype groups. However, there are only four possible combinations for the two genes themselves: *GL1*/*GL2*, *GL1*/*gl2*, *gl1*/*GL2* and *gl1*/*gl2*. The CCCC, CCBB, BBCC and BBBB groups can be allocated immediately to these gene groups as they have no recombination within either region and so are fixed for either the Charger or Badger allele of each gene (*gl1/gl2, gl1/GL2, GL1/gl2, and GL1/GL2*, respectively). However, the CBBB, BCCC, CCCB, and BBBC groups are fixed for one of the genes but segregating for the other gene. For example, RILs in the CBBB group are fixed for the Badger allele of the gene located in Region 2 (*GL2*) but have recombination within Region 1 and therefore could have either the Charger or Badger allele of *GL1* (i.e. $\frac{GL1}{gl1}$/*GL2*). This is reflected in Figure 3.11e where the CBBB group has a higher mean grain length than the other intermediate groups, because it contains both *GL1*/*GL2* and *gl1*/*GL2* lines.

To try and determine which of the gene groups each RIL belonged to, RILs within each group were classified again using a Dunnett's test, but with different controls. The same analysis was performed on the original RIL population and the larger RIL population, although populations were analysed separately. RILs were first split into those with recombination in Region 1 ($\frac{GL1}{gl1}$/*gl2* and $\frac{GL1}{gl1}$/*GL2*) and those with recombination in Region 2 (*gl1/* $\frac{GL2}{gl2}$ and *GL1*/$\frac{GL2}{gl2}$). Each of the genotype groups were classified using the appropriate pair of controls defined by the possible gene groups (outlined in Table 3.4). For example, RILs in the CBBB group (i.e. $\frac{GL1}{gl1}$/*GL2*) could be either *gl1/GL2* or *GL1/GL2* and so lines in the CCBB (*gl1/GL2*) and BBBB (*GL1/GL2*) groups were used as controls. Due to the large number of RILs in the CCCC (*gl1/gl2*) and BBBB (*GL1/GL2*) groups, only the lines which has previously been used as C and B controls were used as controls for these groups. As previously, RILs were only classified if they were both significantly different from one control and non-significantly different from the other. For example, in the CBBB group described above RILs were only classed as *gl1/GL2* if they were significantly different from the *GL1/GL2* control and non-significantly different from the *gl1/GL2* control. As previously, whilst most RILs

could be unambiguously classified in this way, in some cases RILs did not satisfy both conditions and were considered intermediate. In the CBBB example, an intermediate group is expressed as $\frac{GL1}{gl1}$/*GL2* because the region 1 allele remains uncertain. In the original RIL population no RILs were in the BBCC group and therefore no *GL1/gl2* controls were available. Comparisons requiring a *GL1/gl2* control were therefore assigned using a single control and thus are lower confidence (Table 3.4).

In total, 93 independent RILs from the larger RIL population were assigned to one of four gene groups, whilst 27 lines could not be assigned. The *GL1/GL2* group had 2.53% longer grains than the *gl1/gl2* group (Figure 3.12a), similar to but slightly lower than the differences seen between NILs and control lines. As predicted, *GL1/gl2* and *gl1/GL2* groups had smaller increases in grain length with respect to lines with the complete Charger (*gl1/gl2)* than lines with Badger (*GL1/GL2)* interval (1.37 % and 0.58 %, respectively). The *GL1/gl2* group had a greater increase in grain length than the *gl1/GL2* group.

Table 3.4: Classification of RILs within two-gene genotype groups

						Classifications	
Population	**Genotype group**	**Region**	**n RILs**	**Possible gene groups**	**Control groups used**	**Gene group**	**n RILs**
Large RIL population	BCCC	1	15	*GL1/gl2*	BBCC (*GL1/gl2*)	*GL1/gl2*	6
				gl1/gl2	CCCC (*gl1/gl2*)	*gl1/gl2*	5
						$\frac{GL1}{gl1}$/*gl2*	4
	CBBB	1	23	*gl1/GL2*	CCBB (*gl1/GL2*)	*gl1/GL2*	21
				GL1/GL2	BBBB (*GL1/GL2*)	*GL1/GL2*	0
						$\frac{GL1}{gl1}$/*GL2*	2
	CCCB	2	21	*gl1/GL2*	CCBB (*gl1/GL2*)	*gl1/GL2*	17
				gl1/gl2	CCCC (*gl1/gl2*)	*gl1/gl2*	0
						gl1/$\frac{GL2}{gl2}$	4
	BBBC	2	31	*GL1/gl2*	BBCC (*GL1/gl2*)	*GL1/gl2*	13
				GL1/GL2	BBBB (*GL1/GL2*)	*GL1/GL2*	1
						GL1/$\frac{GL2}{gl2}$	17
Original RIL population	BCCC	1	8	*GL1/gl2*	-	*GL1/gl2*	5
				gl1/gl2 C'C'	CCCC (*gl1/gl2*)	*gl1/gl2*	3
						$\frac{GL1}{gl1}$/*gl2*	-
	CBBB	1	4	*gl1/GL2*	CCBB (*gl1/GL2*)	*gl1/GL2*	1
				GL1/GL2	BBBB (*GL1/GL2*)	*GL1/GL2*	4
						$\frac{GL1}{gl1}$/*GL2*	0
	CCCB	2	2	*gl1/GL2*	CCBB (*gl1/GL2*)	*gl1/GL2*	2
				gl1/gl2	CCCC (*gl1/gl2*)	*gl1/gl2*	0
						gl1/$\frac{GL2}{gl2}$	0
	BBBC	2	4	*GL1/gl2*	-	*GL1/gl2*	4
				GL1/GL2	BBBB (*GL1/GL2*)	*GL1/GL2*	0
						GL1/$\frac{GL2}{gl2}$	-

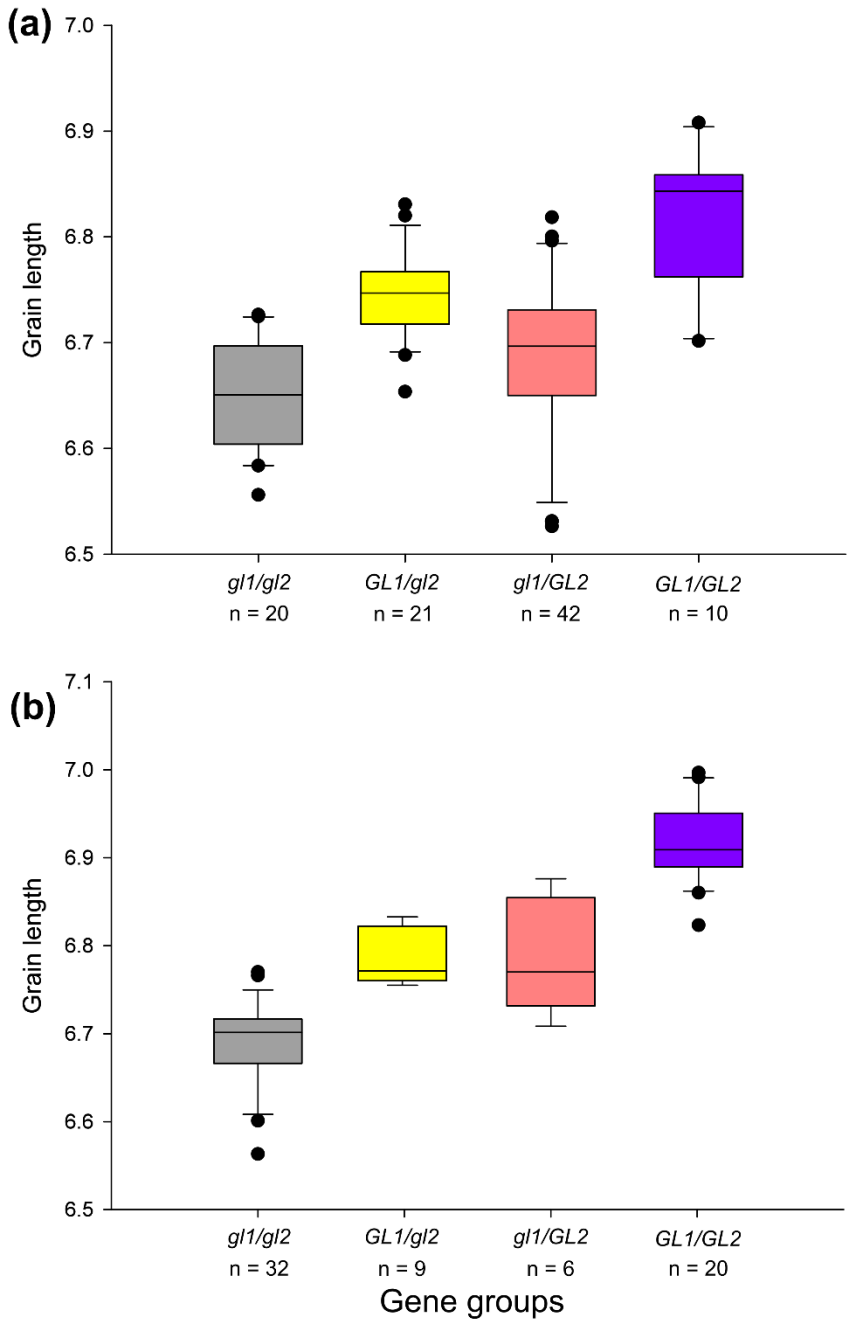


Figure 3.12: BC$_4$ 5A RIL lines phenotypically classified into 'two-gene' groups

Boxplots showing the grain length of the larger 5A RIL population (a) and the original 5A RIL population (b) classified using a Dunnett's test into four gene groups*: gl1/gl2* (grey): Charger alleles of *gl1* and *gl2*; *GL1/gl2 (*yellow): Badger allele of *GL1*, Charger allele of *gl2*; *gl1/GL2* (pink): Charger allele of *gl1*, Badger allele of *GL2*; *GL1/GL2* (purple): Badger allele of *GL1* and *GL2*. n is the number of independent RILs in each group.

The same trend was seen in the original RIL population where all lines could be assigned to one of the four gene groups (Figure 3.12b). The *GL1/GL2* group had 3.50 % longer grains than the *gl1/gl2* group, more reflective of the differences in grain length observed between the 5A NILs. Again, the *GL1/gl2* and *gl1/GL2* groups had smaller increases in grain length compared with the *gl1/gl2* group than the *GL1/GL2* group (1.60 % and 1.53 %, respectively). Similar to the larger RIL population, the increase in the *GL1/gl2* group was slightly higher than for *gl1/GL2* although this was less pronounced in the original RIL population. Less within group variation was observed in the original RIL population than the larger RIL population possibly due to the fact that this consisted of data from multiple trials.

Overall, despite some overlap in the range of grain lengths seen between each of the gene groups, the data appears to fit the expectations of the two-gene hypothesis. The gene groups which carry a single positive Badger allele (*GL1/gl2* and *gl1/GL2*) have an intermediate grain length phenotype compared to lines with both the Charger or Badger alleles at each gene (*gl1/gl2* and *GL1/GL2*). The fact that the percentage increases in grain length contributed by each of the individual genes do not completely account for the increase seen with the Badger allele of both genes could suggest that these genes act synergistically, although this is currently very speculative.

3.4.2.3.1 Further fine-mapping of GL1

Assuming that the 'two-gene' hypothesis is correct and assigning each RIL to a gene group based on phenotype allowed the fine-mapping of grain length to proceed separately for *GL1* and *GL2*. To do this, RILs were separated into *GL1* segregating RILs (i.e. the BCCC and CBBB groups) and *GL2* segregating RILs (i.e. the CCCB and BBBC groups). For *GL2*, the grain length effect could not be mapped any further at this stage as no additional markers could be identified between the flanking markers (*BS00048607*/*JBRNASeq_2*/*JBRNASeq_3* and *JBRNASeq_4*). However, eight additional markers were identified between the flanking markers of *GL1* (*BS00182017* and *BS00186083*; Figure 3.13a). *GL1* segregating RILs from both populations were genotyped with the additional markers (Figure 3.13b,d). Three of the markers were linked to the *BS00182017*, the proximal (left hand) flanking marker of *GL1*. Assessing the genotype of each RIL together with the gene group classification described allowed grain length to be fine-mapped to a slightly narrower interval between *BS00182017* and *BA00603545* in both populations (1.86 cM compared to 2.01 cM previously in the larger population; Figure 3.13; yellow interval). However, despite there being five additional markers across the interval, grain length could not be mapped further with confidence as again different RIL lines suggested conflicting mapping positions. Using just the high confidence RILs from the original population i.e. RILs phenotyped across multiple trials that were classified using two controls (HR-CB18, HR-CB44, HR-CB38 and HR-CB-7), *GL1* mapped between *JBHap011* and *BA00603545* (Figure 3.13d-e). However only two of the six lower confidence RILs in the original population (RILs classified using a single control) supported this mapping position. Similarly, only half of the RILs from the larger RIL population supported this mapping position,

although only based on a single year of data. Overall, further refinement of *GL1* is currently limited by the number of lines with informative recombination in the original RIL population and availability of phenotypic data for the larger RIL population.

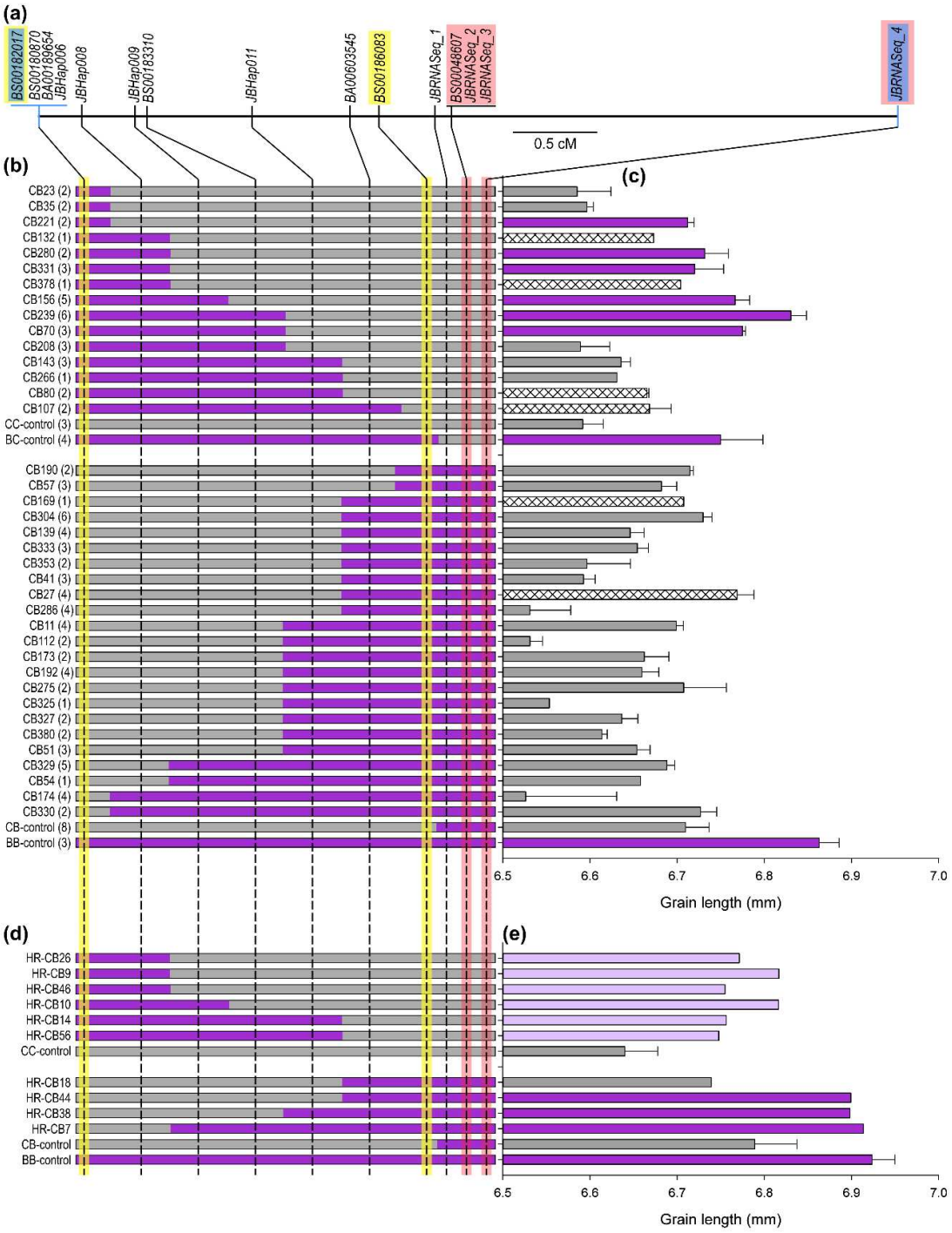


Figure 3.13: Fine-mapping of GL1

(a) Genetic map across the 5A grain length interval according to the larger 5A RIL population. Blue markers define the overall grain length interval defined in Figure 3.7, yellow markers define *GL1* and pink markers define *GL2*. (b) and (d) show graphical genotypes of independent RILs with recombination across Region 1 in the larger and original 5A BC_4 RIL populations, respectively. Number in brackets are the number of individual sibling RILs belonging to each family. (c) and (e) show the ANOVA adjusted mean grain length of each RIL. Bars are coloured according to a Dunnett's test to the control groups shown. Pale purple bars were classified as B-like but are lower confidence as only one control group (CC) was available. Error bars in (c) are standard error of individual RILs within the independent RIL family. Error bars in (e) are standard error of RILs in the control groups.

3.4.2.4 Determining physical positions of markers across the 5A grain length interval

The physical positions of the markers across the 5A grain length interval were determined by a BLASTN of the marker sequences against RefSeq v1.0 (Figure 3.14). For the majority of markers, the physical and genetic positions were in agreement (39 of 43 markers). However, the physical order of a group of four markers (*BA00228977*, *JBRNASeq_4*, *BA00165371*, *BA00379554*) did not agree with the genetic order based on the larger RIL population. In the original RIL population, these markers were linked and defined the right flank of the fine-mapped grain length interval. In the context of the 'two-gene' hypothesis, *JBRNASeq_4* defined the distal (right hand) flank of *GL2*. According to RefSeq v1.0 these four markers covered an interval of 14.6 Mbp in the order *BA00228977*, *JBRNASeq_4*, *BA00165371*, *BA00379554*. However, in the larger RIL population the genetic order of these markers was *JBRNASeq_4*, *BA00228977*, *BA00379554*, *BA00165371* (Figure 3.9a), supported by 15 independent RILs. This suggests some small scale rearrangements on chromosome 5A in the parental cultivars of the RIL populations (Charger and Badger) with respect to the reference cultivar, Chinese Spring. For the purposes of determining the physical size of the intervals and gene numbers, the genetic order was respected and *JBRNASeq_4* was used as the flanking marker of both mapping intervals.

The initial fine-mapping in the original 5A RIL population reduced the mapping interval from 367.5 Mbp (*gwm293-gwm186*) to an interval of 295.2 Mbp between *BS00075504* and *BS00062427* containing 1,929 TGACv1 genes. The further fine-mapping with individual RILs and the larger RIL population considerably reduced the size of the overall grain length mapping interval to 75.3 Mbp (*BS00182017-JBRNASeq_4*) containing 673 TGACv1 genes. Only 531 of these genes were expressed above 0.5 tpm in any of the RNA-Seq samples in the wheat expVIP database (n = 418) and 474 were expressed in at least one grain RNA-Seq sample (> 0.5 tpm; n = 147).

Within the overall grain length region, the initial intervals defining *GL1* (*BS00182017 - BS00186083*) and *GL2* (*JBRNASeq_3 – JBRNASeq_4*) correspond to 45.5 Mbp and 10.5 Mbp, respectively. The additional (tentative) fine-mapping of *GL1* reduced this interval to 33.6 Mbp (*BS00182017 – BA00603545*) containing 311 TGACv1 genes. Just 241 of the genes were expressed in an expVIP RNA-Seq sample and 220 were expressed in at least one grain RNA-Seq sample. The 10.5 Mbp *GL2* interval contained 106 TGACv1 genes, only 89 of which were expressed in any expVIP RNA-Seq sample and 74 in at least one grain sample (Borrill *et al.*, 2016).

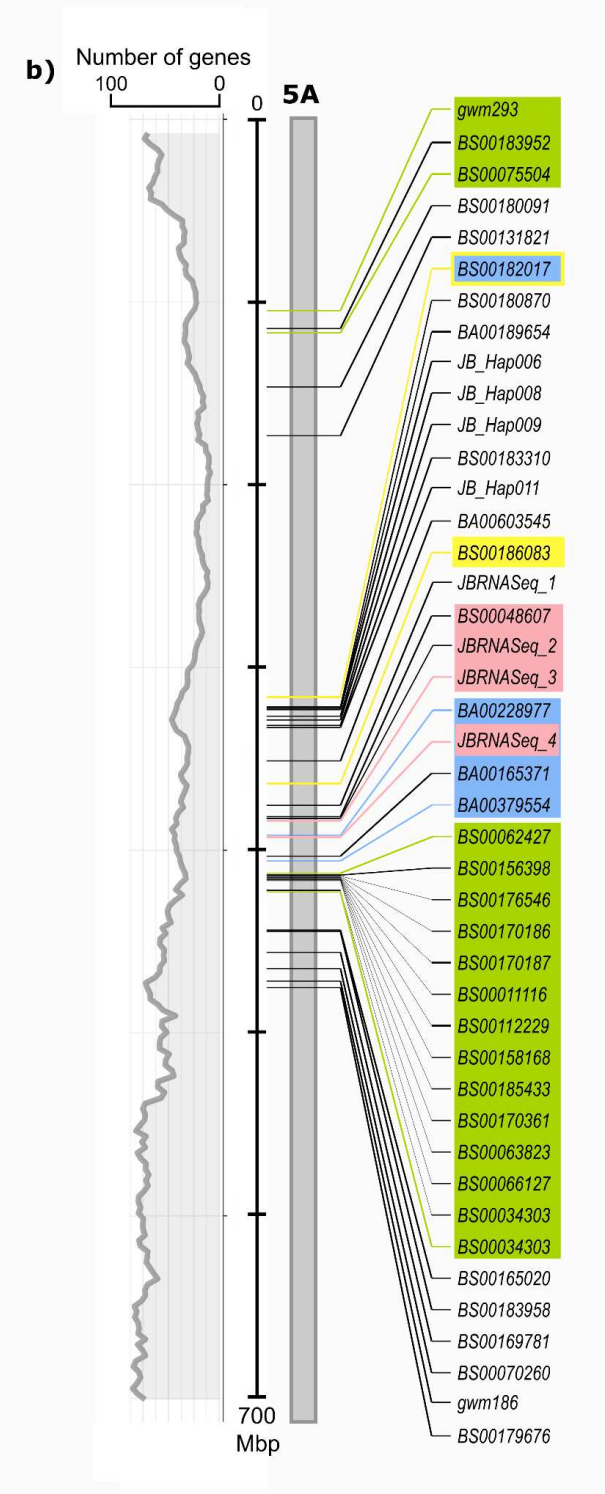


Figure 3.14: Physical positions of markers defining the grain length intervals on chromosome 5A

Physical positions of markers on chromosome 5A according to IWGSC RefSeq v1.0. Markers highlighted in green are flanks of the fine-mapped grain length interval defined by the original BC_4 RIL groups (Figure 3.6), and markers highlighted in blue are flanks of the grain length interval defined by individual RILs (Figure 3.7). Yellow highlighted markers are the flanks of *GL1* and pink highlighter markers are the flanks of *GL2*. Line graph (grey) shows rolling mean of the number of genes located in 3 Mbp bins across chromosome 5A.

3.4.2.5 Haplotype analysis of the 5A grain length interval

3.4.2.5.1 The 5A grain length interval is not fixed in UK germplasm

Defining the physical positions of markers across the interval allowed a haplotype analysis to be conducted to understand how the 5A grain length interval(s) behave in other UK wheat cultivars. Exome capture data from 20 UK wheat cultivars was used to identify SNPs with respect to IWGSC Chinese Spring Chromosome Survey Sequence (IWGSC, 2014) (Figure 3.15; cultivars with circles). The position of each SNP in RefSeq v1.0 was identified using BLAST. SNPs located between 300-400 Mbp on the chromosome 5A pseudomolecule were selected as this region encompassed the fine-mapped grain length interval. A total of 205 SNPs with respect to Chinese Spring were identified in this 100 Mbp region, however 122 of these SNPs were monomorphic in all 20 cultivars and therefore not informative for this analysis. The 83 remaining SNPs were summarised into 22 groups of SNPs that showed the same pattern across the 20 cultivars and a representative SNP was selected for each group (JB_Hap001-022; Figure 3.15). The 20 cultivars were assigned to 12 distinct haplotype groups based on their genotypes across the 22 SNPs, with over half of the cultivars contained within two groups (Group 2: four cultivars, Group 4: seven cultivars). To determine which haplotype groups the parental cultivars of the 5A QTL (Charger (5A-) and Badger (5A+)) belonged to, KASP markers were designed for each of the 22 SNPs (Appendix 2). Both the parental cultivars and a pair of 5A NILs were genotyped with the 22 haplotype markers. Using this genotyping Charger/5A- was assigned to Group 4, whilst Badger/5A+ was assigned to Group 12 (Figure 3.15; grey and purple highlighted cultivars). The fact that Charger and Badger fall into different haplotype groups suggests that the positive 5A grain length allele(s) are not yet fixed in UK germplasm. Additionally, Charger belonged to the largest haplotype group (4) whilst group 12 (containing Badger) was small and quite different from the other haplotype groups. This could suggest that the Charger allele (i.e. the negative 5A allele) is more prevalent within UK breeding programmes and so the selection of the Badger (positive) allele could offer improvements in grain size. An alternative explanation could be that selection for the grain length effect has eroded the long range haplotype of group 12, hence the positive allele is present in many cultivars but not visible in this analysis. However, there are no clear recombination breakpoints to suggest this alternative explanation in this data. Further analysis with additional cultivars and better defined mapping intervals will be required to establish exactly how this QTL has been selected during the breeding process.

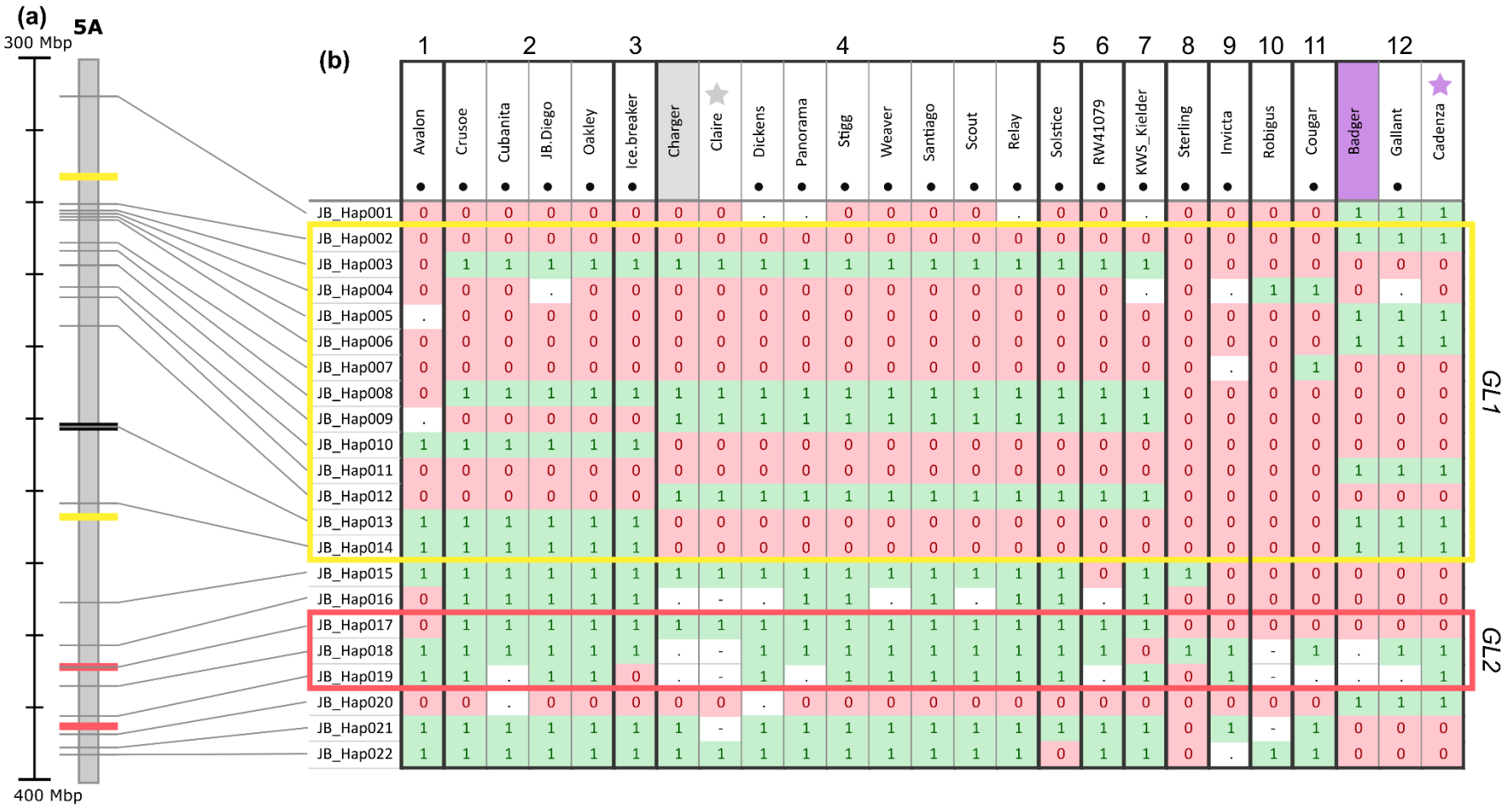


Figure 3.15: Haplotype analysis across the 5A grain length interval

(a) Grey lines are position of SNPs JB_Hap001-022 on chromosome 5A according to RefSeq v1.0. Yellow lines are the flanking markers of Region 1 (*BS00182017 - BS00186083*) and pink lines are the flanks of Region 2 (*JBRNASeq_3 – JBRNASeq_4*). (b) Genotype of each of the SNPs JB_Hap001-022 in different wheat cultivars. '1' indicates a SNP with respect to the Chinese Spring reference, '0' indicates the Chinese Spring allele and '.' indicates a missing data point. Numbers are groups of cultivars with the same genotype across all 22 SNPs. Cultivars with circles indicate the 20 UK cultivars for which exome capture was available. Cultivars with stars are sequenced cultivars in the same haplotype group as Charger or Badger.

3.4.2.5.2 Charger and Badger have the same haplotypes as sequenced varieties

Three additional cultivars were also characterised with the 22 haplotype markers: Claire, Cadenza and Robigus. These three cultivars were selected as they have been sequenced by the Earlham Institute. Claire shared a haplotype with Charger (group 4; Figure 3.15, grey star) whilst Cadenza had the same haplotype as Badger (group 12; Figure 3.15, purple star). Robigus was not identical to any of the haplotype groups and was allocated to its own group (10) although it was highly similar to groups 9 and 11. The fact that Charger and Badger have the same haplotypes as sequenced cultivars means that the genome sequences of Claire and Cadenza can be used as proxies for the parental cultivars of the 5A grain length QTL.

3.4.2.5.3 Regions 1 and 2 are not always inherited together

Haplotypes across the interval were assessed with respect to the two-gene hypothesis. Thirteen haplotype SNPs were located across the *GL1* interval (Figure 3.15, yellow box) and ten were polymorphic between Charger and Badger. Four of these SNPs (JBHap006, 008, 009, 011) were used to genotype RILs in the further fine-mapping of Region 1 discussed previously (3.4.2.3.1). Three haplotype SNPs were located in the *GL2* interval (Figure 3.15, pink box). However, only JB_Hap017 was polymorphic between Charger and Badger and this was located at the same position as JBRNASeq_3, the distal flank of *GL2*. The haplotype analysis suggested that *GL1* and *GL2* are not always inherited together. For example, cultivars such as Avalon (group 1) and Invicta (group 9) shared the Badger haplotype across *GL2* haplotype but did not have the Badger haplotype across *GL1*. This suggests that Avalon and Invicta could have the Badger (5A+) allele of *GL2* but not of *GL1*. No other haplotype groups had the Badger haplotype across *GL1*.

Overall, the 5A grain length effect was fine-mapped to an overall interval of 75.3 Mbp (*BS00182017-JBRNASeq_4*) containing 673 genes. Initial analysis of the larger RIL population suggests that this interval contains two distinct but closely linked genes that have an additive effect on grain length. This is supported by the original RIL population but further phenotypic data is required to confirm the hypothesis. A haplotype analysis across 20 UK wheat cultivars showed that the 5A grain length interval is not fixed in UK germplasm and suggested that the two genes underlying the QTL are not always inherited together. Charger and Badger both share haplotypes with wheat cultivars that have been fully sequenced, Claire and Cadenza, respectively. These genomes can now be used as proxies for the parental genomes and will help to advance the fine-mapping by revealing sequence variation that was not previously accessible, such as promoters and other regulatory regions.

3.5 Discussion

3.5.1 Fine-mapping reveals complex genetic architectures underlying both the 6A and 5A QTL

The main aim of this chapter was to further define the genetic intervals underlying the 6A and 5A QTL. In both cases this revealed complex underlying genetic architectures although in different ways.

3.5.1.1 The 6A QTL maps to a large linkage block on chromosome 6A

Fine-mapping of the 6A grain width QTL defined a high confidence mapping interval of 4.6 cM in the original RIL population, which contained large blocks of linked markers. The generation of a much larger RIL population broke the linkage to a certain extent, but many markers remained linked. Positioning the markers across this physical interval on the chromosome 6A pseudomolecule from the newly released wheat genome (IWGSC RefSeq v1.0) revealed that this 4.6 cM interval corresponded to a very large physical interval of 421.8 Mbp, over two-thirds of the chromosome 6A pseudomolecule (618 Mbp). A tentative sub-centimorgan (0.28 cM) interval was defined using the larger RIL population, but this still corresponded to a large physical interval of 61.2 Mbp predicted to contain 488 genes. The 6A interval is located close to the centromere, and it is well documented that centromeric regions are associated with lower rates of recombination in Triticeae hence leading to extended linkage disequilibrium (LD; Akhunov *et al.*, 2003; Mascher *et al.*, 2017). However, the linkage appears to extend across a large proportion of chromosome 6A and this could be due to additional factors as well as the centromeric position of the region. It would be interesting to examine this region more closely the exome capture data used in the 5A haplotype analysis to determine if this extended linkage also exists in other wheat cultivars. It would also be interesting to assess more generally the genetic diversity that exists across this region in UK germplasm and other germplasm pools to see if there has been any particularly strong selection placed on this region during the breeding process. The exome capture data available for the 20 UK cultivars represents a valuable starting point for these studies.

The extended linkage across the 6A interval has important implications for association and mapping studies aiming to identify genes located within this region. The high degree of linkage in this interval could result in spurious associations of a trait with a polymorphism in a specific candidate gene. The extended haplotype across the region would encompass hundreds of other genes in addition to the candidate gene that could potentially be underlying the trait leading to incorrect conclusions. The results of the 6A QTL fine-mapping illustrate how the limited recombination rate impedes positional cloning of genes within this region as the resolution to which traits can be mapped is not sufficiently high. Alternative mapping approaches that are not so dependent on recombination rate could be employed to overcome this, for example, the use of

deletion lines. The quantitative and subtle nature of the phenotype, however, makes this analysis difficult.

However, in the larger RIL population there are at least six independent RILs with recombination across the tentative 0.28 cM grain width interval. Additional phenotypic data for the RILs will be obtained from the 2017 field trials which can be used to confirm this position. Additional markers across the interval will also be identified to further define the 6A grain width effect to a narrower physical interval. Within a breeding context, the fact that strong linkage exists across this interval suggests that the flanking markers identified within this PhD would be sufficient to efficiently select the 6A positive haplotype. Breeders should be aware however that in doing so they are also selecting an additional 480 genes. It will be valuable to compare the sequenced UK cultivars to examine the consequences of this strategy and to define differences in sequence across these long range haplotypes.

3.5.1.2 There are potentially two genes underlying the 5A QTL that influence grain length

Similar to the 6A QTL, the 5A grain length effect could only be mapped to a relatively large physical interval of 75 Mbp *BS00182017* and *JBRNASeq_4* containing 673 genes. However, unlike the 6A QTL, this was not due to limited recombination across the interval. The mapping was instead limited by the fact that the majority of RILs across the interval had an intermediate grain length phenotype and therefore could not classified as a parental type. This effect was observed in both RIL populations and across multiple independent trials. These results led to the hypothesis that there are two tightly linked genes underlying the 5A QTL that have an additive or synergistic effect on grain length. This could explain why the increase in grain weight conferred by this QTL (6.9 %) is relatively large compared to other grain weight QTL in wheat that have more subtle effects (Simmonds *et al.*, 2014; Farré *et al.*, 2016).

The identification and separation of the two regions will allow fine-mapping of the grain length effect of the two genes to proceed separately. The selection of more appropriate controls for each region has already allowed more RILs to be classified phenotypically which is essential for fine-mapping (discussed below). Indeed, the separation of *GL1* and *GL2* has reduced complexity within the interval by dividing the large 75 Mbp region containing > 600 genes into two smaller regions containing 241 and 80 genes, respectively.

The phenomenon of two closely linked loci affecting a trait has been observed in relation to other traits in wheat. For example, two closely linked haplotype blocks on chromosome 5B were identified that interact to influence root biomass (Voss-Fels *et al.*, 2017). Interestingly, it seems that one particular combination of alleles across these haplotype blocks dominates in European wheat cultivars due to the strong selection of a QTL for heading date located between the two blocks. Another example is the close physical proximity of *TaMKK3-A* and the *PM19-A* genes on chromosome 4A that both influence seed dormancy. These genes were initially both proposed as

candidates for the *Phs-A1* pre-harvesting sprouting QTL on chromosome 4A. Although it was subsequently shown that *TaMKK3-A* was the causal gene underlying the *Phs-A1* locus (Shorinola *et al.*, 2017), transgenic *PM19-A* lines demonstrate a role for these genes in the control of seed dormancy as well (Barrero *et al.*, 2015).

It will be interesting to understand how *GL1* and *GL2* interact on a mechanistic level to influence grain length. For example, do they affect the same biological processes or distinct pathways? Characterisation of recombinants that separate the two genes in a similar way to the NILs described in Chapter 2 would provide insights into this e.g. by profiling grain development and cell size. A study examining pericarp cell size in a large number of the 5A RILs has been conducted using material grown in 2017 field trials. Currently it is not known whether these genes also have an additive effect on the pericarp cell size, or whether this phenotype is associated with just one of the genes. The interactions between genes could occur on a range of levels including physical interaction at the protein level, genetic interaction by influencing steps of the same pathway or by influencing independent pathways that both control grain length. Ultimately, identification and functional annotation of the genes themselves will allow hypotheses about how they interact to be generated and tested experimentally.

3.5.2 *TaGW2_A* does not map within the tentative 6A grain width interval

Another aim of this thesis is to determine whether *TaGW2_A* is the causal gene underlying the 6A QTL. In Chapter 2 we identified phenotypic differences between 6A NILs and *TaGW2_A* mutant NILs suggesting that they act through different mechanisms. In the current chapter, this hypothesis was tested genetically by using the *Hap-P2* promoter SNP discussed in Chapter 2 (Su *et al.*, 2011; Zhang, X *et al.*, 2013) to map *TaGW2_A* relative to the 6A QTL. The high confidence 4.6 cM grain width interval defined by the original RIL population included the *Hap-P2* marker therefore not eliminating *TaGW2_A* as a candidate gene. However, as discussed above this interval corresponds to a huge physical distance (421.8 Mbp) containing > 2,000 genes and extensive linkage. So although this interval does not eliminate *TaGW2_A* as a candidate, it equally does not provide evidence that *TaGW2_A* is the underlying gene as there are so many other genes contained within the interval. However, further fine-mapping using the larger RIL population defined a tentative interval of 0.28 cM (61.2 Mb, 488 genes) that does not include *TaGW2_A*. Two independent RIL families had recombination events that separated the grain width phenotype from the *Hap-P2* marker (Figure 3.4, SR Gr2.8). Both RILs had the Spark (6A-) allele of the Hap-P2 marker, but were classified phenotypically as Rialto-like (6A+). Data from additional trials (e.g. samples from the 2017 field trials) will be required to confirm and increase confidence in these results. However, taking together the genetic evidence presented here and the phenotypic differences in grain size parameters and development described in Chapter 2 seems to suggest that the *TaGW2_A* is not the gene underlying the 6A QTL and that they act through different mechanisms.

3.5.3 Dissecting grain weight to more stable components allows near-qualitative classification of RILs

For both the 6A and 5A QTL, fine-mapping was conducted using a more stable grain size component than TGW (grain width and grain length, respectively). Dissecting the TGW effect into these more stable components allowed RILs to be classified to a parental type in a qualitative/binary manner i.e. narrow/short grains or wide/long grains. This was essential to enable the fine-mapping of the loci as a similar classification was not possible using the TGW effect. This was demonstrated in the original 6A RIL population as using the grain width phenotype allowed eleven of thirteen RIL groups to be classified in a binary manner whilst this was only possible for two of the same thirteen groups using the TGW effect (Figure 3.2).

However, even using the individual grain size components presented challenges as unambiguous classification of lines to a parental type was not always possible. Although the effects on grain width and length of the two QTL are very stable, they are still very subtle (~4 % difference between control lines). Additionally, the differences in mean grain width/length between genotypes actually represent overlapping distributions of grain size which can be difficult to separate (Figure 2.4, Figure 2.7). The subtle effects of grain size/weight QTL present a major challenge to defining the mechanisms and genes underlying grain weight QTL in wheat. Grain size QTL that have been successfully cloned in diploid species such as rice have much larger effects, often with >20% differences between genotypes when examined in NILs (Song *et al.*, 2007). It has been proposed that this is due to the full effects of a single gene being masked or buffered by the functional redundancy conferred by homoeologous gene copies in polyploid wheat (Borrill *et al.*, 2015a).

The ‘two-gene’ hypothesis for the 5A QTL proposes that the difficulty in classifying RILs with recombination across the 5A interval to a parental type was due to the two genes having an additive effect on grain size. Separating the two regions in this interval and selecting more appropriate controls allowed more of the RILs to be unambiguously classified to a specific ‘gene group’. However, this exacerbates the issue described above as the differences between controls are even more subtle when considering the genes separately (~1%). In the Chapter 2, the grain length effect of the 5A QTL was shown to be driven by increased pericarp cell length. Importantly, the effect on pericarp cell size was shown to be independent of absolute grain length. This is therefore an even more stable phenotype of the 5A QTL and could greatly assist with the allocation of RIL lines to a parental/control type by reducing the phenotypic variation between grains that can mask the subtle effect of the grain length phenotype. Pericarp cell size phenotyping of 5A RILs was conducted during the 2017 field season and will be used to further define the 5A grain length intervals. It is also possible that the two genes act through different mechanisms and that the cell length phenotype will map to a single locus.

3.5.4 The value of advances in wheat genomics resources for genetic mapping

The recent advances in wheat genomics resources were invaluable for the fine-mapping performed in this chapter. The majority of SNP markers used were designed using data from genotyping of NILs or parental cultivars with the 90k iSelect and 820k high density SNP arrays (Simmonds *et al.*, 2014; Winfield *et al.*, 2016; Brinton *et al.*, 2017). Initially markers were selected based on their predicted genetic positions (POPSEQ) but many SNPs were not positioned in these datasets. The specific positioning of these SNPs with respect to the recently released wheat chromosome pseudomolecules (RefSeq v1.0) allowed many more SNPs across the two mapping intervals to be identified.

The positioning of markers onto the chromosome pseudomolecules allowed for the first time insight into the physical sequence underlying the QTL mapping intervals. Although the annotation of the IWGSC RefSeqv1.0 has only recently been made publically available, an *in silico* mapping of the TGACv1 gene models (Clavijo *et al.*, 2017b) allowed identification of the genes within the intervals. The TGACv1 gene models have also been functionally annotated, and expression data across a wide range of conditions and tissues was examined using the wheat expVIP database (Borrill *et al.*, 2016). However, the mapping intervals of both QTL currently remain too large to begin speculating on candidate genes based on predicted function and expression patterns.

One limitation of using the Chinese Spring reference genome to understand the 6A and 5A intervals is that there will be differences between this cultivar and the parental cultivars of the QTL. This was already observed across the 5A region where there were some discrepancies between the order of markers in the reference sequence and the genetic order determined using the RIL populations. This could suggest some small scale rearrangements in this region in Charger and/or Badger with respect to Chinese Spring. Alternatively, this could represent errors in the genome sequence. In addition to rearrangements, there may also be other differences including the presence/absence of certain genes. With this in mind, the wheat community is now moving towards the development of a wheat pan-genome, with the complete genome sequences of several wheat cultivars from across the world already being available and many more in production (http://opendata.earlham.ac.uk/Triticum_aestivum/; Montenegro *et al.*, 2017).

To overcome the cost and time constraints imposed by whole genome sequencing of cultivars reduced-representation sequencing, such as exome capture and RNA-Seq, can provide valuable insights into the variation that exists between cultivars. In the current chapter, RNA-Seq data of 5A NILs was used to identify additional polymorphisms between NILs as additional markers for fine-mapping (this experiment is discussed in more depth in Chapter 4). Additionally, SNP calling using exome capture data of 20 UK cultivars also allowed haplotype groups across the 5A interval to be defined. This haplotype analysis suggested that the 5A grain length QTL is not fixed in UK germplasm. This analysis also identified additional SNPs within the region and found that both Charger and Badger share haplotypes across the 5A interval with wheat cultivars that have been

fully genome sequenced. This opens up exciting new opportunities as it will essentially allow the comparison of the full genome sequences of the 5A parents across the interval whereas previous data had been limited just to the coding regions (RNA-Seq and exome capture). As no additional coding region variations across the 5A regions could be identified between NILs/parents using RNA-Seq and exome capture data, this could suggest that the causal SNPs are located in non-coding regions, such as promoters or introns.

4 Comparative transcriptomics of 5A NILs

All results described in this chapter have been submitted for publication and have been uploaded to the preprint server, bioRxiv as the following manuscript (Appendix 3):

Brinton J, Simmonds J, Uauy C. 2017. Ubiquitin-related genes are differentially expressed in isogenic lines contrasting for pericarp cell size and grain weight in hexaploid wheat. bioRxiv. doi.org/10.1101/175471

4.1 Chapter summary

In this chapter, RNA-Seq was performed on 5A NILs to identify differentially expressed genes and pathways that potentially influence pericarp cell size and grain weight. Grains were sampled at four and eight days post anthesis according to the 2014 time course described in Chapter 2. A specific set of 112 transcripts were differentially expressed between 5A NILs at either time point, including seven genes located in the fine-mapped interval(s) defined in Chapter 3. Many of the wheat genes identified belong to families that have been previously associated with seed/grain development in other species. However, few of these wheat genes are the direct orthologues and none have been previously characterised in wheat. Notably, differentially expressed transcripts were identified at almost all steps of the pathway associated with ubiquitin-mediated protein degradation.

4.2 Introduction

Transcriptomics is a powerful tool to gain insights into the complex gene regulatory networks that underlie specific traits and biological processes. Several studies have used transcriptomics approaches to look at the genes expressed during grain development in wheat (Laudencia-Chingcuanco *et al.*, 2007; Wan *et al.*, 2008; Pellny *et al.*, 2012; Shewry *et al.*, 2012; Liu *et al.*, 2014; Pfeifer *et al.*, 2014a; Yu *et al.*, 2016). However, these studies have mostly focussed on the later stages of grain development, often focussing on starch accumulation in the endosperm. Additionally, many of these studies were performed using microarrays (Laudencia-Chingcuanco *et al.*, 2007; Wan *et al.*, 2008; Yu *et al.*, 2016), which represent a fraction of the transcriptome and are unable to distinguish between homoeologous gene copies. More recent studies have used RNA-Seq (Pellny *et al.*, 2012; Pfeifer *et al.*, 2014a), which is an open-ended platform that provides homoeologue specific resolution. However, the accuracy of RNA-Seq is dependent on the availability of a high-quality reference sequence and accurate gene models. To date, the RNA-Seq grain development studies have used either expressed sequence tags (ESTs) (Pellny *et al.*, 2012; Liu *et al.*, 2014) or the Chromosome Survey Sequence (CSS) (Pfeifer *et al.*, 2014a) as references. However in hindsight, these annotations are incomplete with respect to the latest gene models (IWGSC RefSeq v1.0; Clavijo *et al.*, 2017b). These novel resources (introduced in more detail in Chapter 1) provide new opportunities for more detailed and accurate transcriptomic studies in wheat.

A potential drawback of transcriptomic studies is that comparisons across varieties, tissues or time points can result in a large number of transcripts being differentially expressed (DE). Whilst this informs our understanding of the biological mechanisms, it is difficult to prioritise specific genes for downstream analysis. Comparative transcriptomic approaches using more precisely defined genetic material, tissues and developmental time points can aid in this by defining a smaller set of differentially regulated transcripts. For example, a comparison of the flag leaf transcriptomes of wildtype and RNAi knockdown lines of the *Grain Protein Content 1* (*GPC*) genes was used to identify downstream targets of the *GPC* TFs (Cantu *et al.*, 2011). Similarly, the transcriptomes of NILs segregating for a major grain dormancy QTL on chromosome arm 4AL were compared and specific candidate genes underlying the QTL were identified (Barrero *et al.*, 2015). To our knowledge, no such experiments have been performed on isogenic lines with a known difference for grain size in wheat.

In this chapter, RNA-Seq was performed on the 5A grain length NILs characterised in Chapter 2. In Chapter 2, we established that the QTL acts during early grain development and that 5A+ NILs have significantly increased thousand grain weight (TGW; 7%), grain length (4%) and pericarp cell length (10%) compared to 5A- NILs (Brinton *et al.*, 2017). The 5A NILs carry an introgressed segment of ~490 Mb and in Chapter 3 we fine-mapped the grain length effect to a 75 Mb region on the long arm of chromosome 5A according to the IWGSC RefSeq v1.0. We also hypothesised that this interval contains two genes (*GL1* and *GL2*) that have an additive effect on grain length. The RNA-Seq experiment was only conducted on 5A NILs as we had clearly established the stage during grain development when the first phenotypic differences between NILs appear and hence hypothesised that this is when the 5A QTL acts. For the 6A NILs however, it was not possible to define exactly the developmental stage at which the 6A QTL acts and consequently to select the most appropriate sampling time.

The aim of this chapter was to identify biological pathways that potentially influence grain length and pericarp cell size by using RNA-Seq to identify genes that are differentially regulated between the 5A- and 5A+ NILs.

4.3 Methods

4.3.1 Plant material

The 5A BC_4 NILs used in this chapter were characterised in Chapter 2 and have been described previously (Brinton *et al.*, 2017). One genotype each for the 5A- (Charger allele, short grains) and 5A+ NIL (Badger allele, long grains) were used (the same NIL pair as used for the cell size measurements in Chapter 2). Plants were sampled at 4 (time point 1: T1) and 8 (time point 2: T2) days post anthesis (dpa) during the 2014 developmental time course outlined in Chapter 2 (Brinton *et al.*, 2017). Briefly, plants were grown in 1.1 x 6 m plots (experimental units) in a complete randomised block design with five replications, and spikes were tagged at full ear emergence. The

three blocks with the most similar flowering time were used for sampling. For each genotype, three grains from three separate spikes from different plants within the experimental unit were sampled. Each biological replicate, therefore, consisted of the pooling of nine grains per genotype. Grains were sampled from the outer florets (positions F1 and F2) from the middle section of each of the three spikes. Grains were removed from the spikes in the field, immediately frozen in liquid nitrogen and stored at -80ºC. In total, three biological replicates (from the three blocks in the field) were sampled for each NIL at each time point.

4.3.2 RNA extraction and sequencing

For each biological replicate, the nine grains were pooled and ground together under liquid nitrogen. RNA was extracted in RE buffer (0.1 M Tris pH 8.0, 5 mM EDTA pH8.0, 0.1 M NaCl, 0.5% SDS, 1% β-mercaptoethanol) with Ambion Plant RNA Isolation Aid (Thermo Fisher Scientific). The supernatant was extracted with 1:1 acidic Phenol (pH 4.3):Chloroform. RNA was precipitated at -80ºC by addition of Isopropanol and 3M NA Acetate (pH 5.2). The RNA pellet was washed twice in 70% Ethanol and resuspended in RNAse free water. RNA was DNAse treated and purified using RNeasy Plant Mini kit (Qiagen) according to the manufacturer's instructions. RNA QC, library construction and sequencing were performed by the Earlham Institute, Norwich. Library construction was performed on a PerkinElmer Sciclone using the TruSeq RNA protocol v2 (Illumina 15026495 Rev.F). Libraries were pooled (2 pools of 6) and sequenced on 2 lanes of a HiSeq 2500 (Illumina) in High Output mode using 100bp paired end reads and V3 chemistry. Initial quality assessment of the reads was performed using fastQC (Andrews, 2010).

4.3.3 Read alignment and differential expression analysis

Reads were aligned to two reference sequences from the same wheat variety, Chinese Spring: the Chromosome Survey Sequence (CSS; IWGSC, 2014) downloaded from *Ensembl* plants release 29) and the TGACv1 reference sequence (Clavijo *et al.*, 2017b). Read alignment and expression quantification were performed using kallisto-0.42.3 (Bray *et al.*, 2016) with default parameters, 30 bootstraps (-b 30) and the –pseudobam option. Kallisto has previously been shown to be suitable for the alignment of wheat transcriptome data in a homoeologue specific manner (Borrill *et al.*, 2016).

Differential expression analysis was performed using sleuth-0.28.0 (Pimentel *et al.*, 2017) with default parameters. Transcripts with a false-discovery rate (FDR) adjusted P-value (q value) < 0.05 were considered as differentially expressed. Transcripts with a mean abundance of < 0.5 tpm in all four conditions were considered not expressed and were therefore excluded from further analyses.

For each condition, the mean tpm of all three biological replicates was calculated. All heatmaps display mean expression values as normalised tpm, on a scale of 0 to 1 with 1 being the highest expression value of the transcript. Read coverage for gene models was obtained using bedtools-2.24.0 genome cov (Quinlan & Hall, 2010) for each pseudobam file and then combined to get a

total coverage value of each position. Coverage across a gene model was plotted as relative coverage on a scale of 0 to 1, with 1 being equivalent to the highest level of coverage for the gene model in question.

4.3.4 GO term enrichment

The R package GOseq v1.26 was used (Young *et al.*, 2010) to test for enrichment of gene ontology (GO) terms in specific groups of DE transcripts. Over-represented GO terms with a Benjamini Hochberg FDR adjusted P-value of < 0.05 were considered to be significantly enriched.

4.3.5 Functional annotation

Functional annotations of transcripts were obtained from the TGACv1 annotation (Clavijo *et al.*, 2017b). Additionally, for coding transcripts BLASTP against the non-redundant NCBI protein database and conserved domain database were performed, in each case the top hit based on e-value was retained. In cases where all three annotations were in agreement, the TGAC annotation is reported. In cases where the three annotations produced differing results, all annotations are reported. Orthologues in other species such as Arabidopsis and rice were obtained from *Ensembl* plants release 36. Eight of the 112 DE transcripts had no annotation or protein sequence similarity with other species. The remaining 104 DE transcripts were manually categorised based on their predicted function. Transcripts that fell into a category of size 1 were classed as 'other'. For the non-coding transcripts, BLASTN was used to identify potential miRNA precursors using a set of conserved and wheat specific miRNA sequences obtained from (Sun *et al.*, 2014). The -task blastn-short option of BLAST for short sequences was used and only hits of the full length of the miRNA sequence with no mismatches as were considered as potential precursors. The psRNAtarget tool (http://plantgrn.noble.org/psRNATarget/) was used to determine the miRNA targets.

4.3.6 Identification of transcription factor binding sites

1,000 bp of sequence upstream of the cDNA start site was extracted to search for transcription factor binding sites (TFBS). Transcripts with < 1,000 bp upstream in the reference sequence were not used in the analysis. The FIMO tool from the MEME suite (v 4.11.4; Grant *et al.*, 2011) was used with a position weight matrix (PWM) obtained from plantPAN 2.0 (http://plantpan2.itps.ncku.edu.tw/; Chow *et al.*, 2016). FIMO was run with a P-value threshold of <1e-4 (default), an increased max-stored-scores of 1,000,000 to account for the size of the dataset, and a –motif-pseudo of 1e-8 as recommended for use with PWMs (Peng *et al.*, 2016). The background model was generated using the fasta-get-markov command of MEME on all extracted promoter sequences.

4.3.7 Enrichment testing

Fisher's exact test was performed to test for enrichment of different categories of transcripts relative to all expressed transcripts using R-3.2.5. For functional annotation categories, enrichment

testing was only performed on categories that could be extracted using GO terms and key words based on their annotation in the TGAC reference. Only DE transcripts that could be extracted using this method were used in the enrichment tests. For example, 12 DE transcripts identified were associated with ubiquitin. The annotation of these transcripts was obtained through a combination of the TGAC annotation and manual annotation. However, only seven of these transcripts could be extracted using GO terms and key words from the whole reference annotation. Therefore, only seven transcripts were used for the enrichment test.

4.4 Results

4.4.1 RNA-sequencing of 5A near isogenic lines

RNA-seq was performed on whole grains from two of the 5A grain length NILs (Chapter 2; Brinton *et al.*, 2017). The time point when NILs showed the first significant differences in grain length (8 dpa; T2) and the preceding time point (4 dpa; T1) were selected to capture differences in gene expression occurring during this period (Figure 4.1). We hypothesised that although there was no significant difference in the grain length phenotype at T1, phenotypic differences were beginning to emerge and gene expression changes influencing this may already be occurring. Over 362 M reads across all 12 samples were obtained (two time points, two NILs, three biological replicates), with individual samples ranging from 15.0 M to 53.6 M reads and an average of 30.2 M reads (standard error ± 3.5 M reads) per sample (Table 4.1). Reads were aligned to two different transcriptome sequences from the reference wheat cultivar, Chinese Spring: the IWGSC Chromosome Survey Sequence (CSS; IWGSC, 2014) and TGACv1 (TGAC; Clavijo *et al.*, 2017b) reference. On average across samples, 69.8 ± 0.3 % of reads aligned to the CSS reference, whilst 84.4 ± 0.2 % of reads aligned to the TGAC reference.

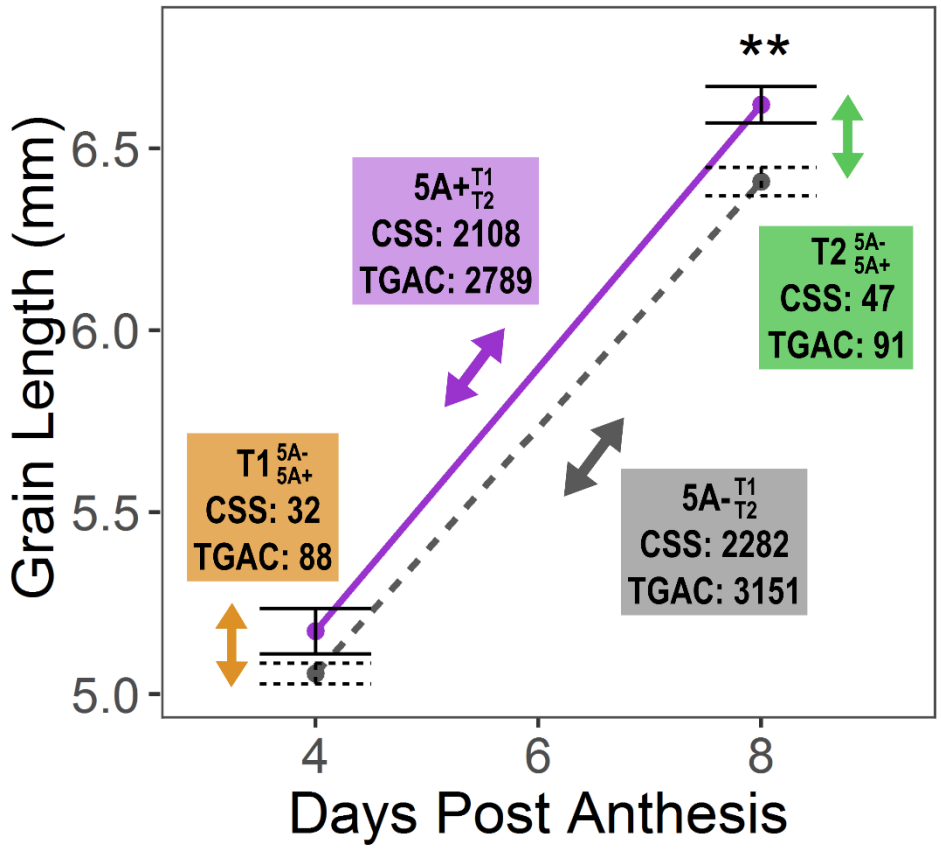


Figure 4.1: Differentially expressed genes between 5A NILs across time

RNA-seq was carried out on whole grain RNA samples taken in 4 different conditions: 5A- (short grains) and 5A+ (long grains) NILs at 4 days post anthesis (dpa; T1) and 8 dpa (T2). These were selected as the time point when the first significant difference ($P < 0.01$, asterisks) in grain length was observed between 5A- (grey, dashed line, short grains) and 5A+ (purple, solid line, long grains) and the preceding time point. Differentially expressed (DE) transcripts were identified for four comparisons (q value < 0.05). Coloured boxes indicate the numbers of DE transcripts identified for each comparison using alignments to either the IWGSC Chinese Spring Chromosome Survey Sequence (CSS) or the TGACv1 (TGAC) Chinese Spring reference transcriptomes. Two 'across time' comparisons: 5A- $^{T1}_{T2}$ (grey box; comparing T1 and T2 samples of the 5A- NIL) and 5A+ $^{T1}_{T2}$ (purple box; comparing T1 and T2 samples of the 5A+ NIL), and two 'between NIL' comparisons: T1 $^{5A-}_{5A+}$ (orange box; comparing 5A- and 5A+ NILs at T1) and T2 $^{5A-}_{5A+}$ (green box; comparing 5A- and 5A+ NILs at T2).

Table 4.1: Mapping summary of RNA-Seq samples

				CSS gene models		TGAC gene models	
Genotype	**Time point**	**Replicate**	**Reads**	**Reads pseudoaligned**	**% reads pseudoaligned**	**Reads pseudoaligned**	**% reads pseudoaligned**
5A -	1	1	24,443,658	17,072,939	69.85	20,549,681	84.07
5A -	1	2	34,441,799	23,349,288	67.79	28,483,090	82.70
5A -	1	3	23,462,705	16,220,597	69.13	19,664,859	83.81
5A -	2	1	21,333,672	14,839,724	69.56	18,052,324	84.62
5A -	2	2	14,967,302	10,632,519	71.04	12,803,552	85.54
5A -	2	3	35,522,754	25,491,523	71.76	30,297,336	85.29
5A +	1	1	19,267,564	13,520,181	70.17	16,317,352	84.69
5A +	1	2	22,299,102	15,479,234	69.42	18,780,525	84.22
5A +	1	3	30,531,539	20,789,582	68.09	25,436,453	83.31
5A +	2	1	51,637,607	36,192,489	70.09	43,739,451	84.70
5A +	2	2	53,575,232	37,956,887	70.85	45,497,914	84.92
5A +	2	3	30,553,421	21,604,895	70.71	25,984,674	85.05
		Total	362,036,355	253,149,858	-	305,607,211	-
		Mean	30,169,696	21,095,822	69.87	25,467,268	84.41

4.4.2 Comparison between Chinese Spring reference transcriptomes

A transcript was defined as expressed if it had an average abundance of > 0.5 tpm in at least one of the four conditions (2 NILs x 2 time points). This resulted in 62.5 % (64,020) and 37.1% (101,652) of the transcripts being expressed in the CSS and TGAC transcriptomes, respectively. DE transcripts (q value < 0.05) were defined using sleuth (Pimentel *et al.*, 2017) and four pairwise comparisons were performed: two 'across time' and two 'between NIL' comparisons. The 'across time' analyses consisted of a comparison between T1 and T2 samples of the 5A- NIL (hereafter symbolised as 5A- $^{T1}_{T2}$; Figure 4.1, grey) and the corresponding comparison for the 5A+ NIL samples (hereafter 5A+ $^{T1}_{T2}$; Figure 4.1, purple). In both cases, the T1 sample was used as the control condition, so transcripts were considered as upregulated or downregulated with respect to T1. The 'between NIL' analyses consisted of a comparison between the 5A- and 5A+ NILs at T1 (hereafter T1 $^{5A-}_{5A+}$; Figure 4.1, orange), and a comparison between the 5A- and 5A+ NILs at T2 (hereafter T2 $^{5A-}_{5A+}$; Figure 4.1, green). In both cases, the recurrent parent 5A- NIL was used as the control genotype. In all cases, more DE transcripts were identified in the TGAC compared with the CSS transcriptome, and similar trends were observed for both references across the four comparisons (Figure 4.1).

The comparison with the fewest DE transcripts (T1 $^{5A-}_{5A+}$; 32 and 88 DE transcripts for CSS and TGAC, respectively) was selected to conduct a more in depth analysis of the alignments and references. For all DE transcripts from each alignment the equivalent transcript/gene model was identified in the other reference sequence using *Ensembl* plants release 35 and the gene models were compared (Table 4.2).

Table 4.2: Comparison between TGAC and CSS gene models

TGAC transcript ID	CSS transcript ID	DE in TGAC?	TGAC q value	DE in CSS?	CSS q value	Comparison class
TRIAE_CS42_5AL_TGACv1_378188_AA1251790.1	Traes_5AL_CA424FE08.2	Y	2.92E-02	N	1.00E+00	CSS 3' truncation
TRIAE_CS42_2DL_TGACv1_157942_AA0502830.1	Traes_2DL_E1640BFDC.1	Y	4.21E-02	N	1.00E+00	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_374078_AA1189690.1	Traes_5AL_8BF894427.2	Y	6.35E-07	N	1.00E+00	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_374446_AA1200420.1	Traes_5AL_0573B44BE.1	Y	2.81E-02	N	1.00E+00	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_374560_AA1203240.1	Traes_5AL_32B5C730F.1	Y	3.22E-02	N	1.00E+00	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_377520_AA1247660.1	Traes_5AL_55BB0BEFC.1	Y	4.71E-02	N	1.00E+00	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_377986_AA1250630.1	Traes_5AL_999D96884.1	Y	6.14E-08	Y	1.62E-09	CSS 5' and 3' truncation
TRIAE_CS42_5AL_TGACv1_378334_AA1252720.1	Traes_5AL_1639C7AB0.1	Y	1.59E-10	N	NA	CSS 5' and 3' truncation
TRIAE_CS42_5AS_TGACv1_393493_AA1273190.4	Traes_5AS_9D5B8EA01.1	Y	7.60E-10	Y	1.69E-12	CSS 5' and 3' truncation
TRIAE_CS42_5DL_TGACv1_433728_AA1420750.1	Traes_5DL_531A38273.1	Y	1.63E-02	N	NA	CSS 5' and 3' truncation
TRIAE_CS42_7DS_TGACv1_622195_AA2034920.1	Traes_7DS_E14CFC6F2.2	Y	1.48E-04	Y	3.77E-05	CSS 5' and 3' truncation
TRIAE_CS42_5AS_TGACv1_392838_AA1265240.1	Traes_5BL_6C1EFA808.1	Y	1.60E-07	Y	4.03E-02	CSS 5' and 3' truncation + chromosome
TRIAE_CS42_5AS_TGACv1_392838_AA1265240.2	Traes_5BL_6C1EFA808.1	Y	2.92E-02	Y	4.03E-02	CSS 5' and 3' truncation + chromosome
TRIAE_CS42_5BS_TGACv1_427448_AA1393420.1	Traes_1AS_2B7CD7B59.1	Y	7.13E-07	N	1.00E+00	CSS 5' and 3' truncation + chromosome
TRIAE_CS42_5AL_TGACv1_373986_AA1186560.2	Traes_5AL_3AA4476D6.1	Y	2.24E-05	Y	6.41E-51	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_373986_AA1186560.3	Traes_5AL_3AA4476D6.1	Y	4.64E-07	Y	6.41E-51	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_373995_AA1186970.1	Traes_5AL_D57725ABD.1	N	1.00E+00	Y	2.03E-05	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_374080_AA1189800.1	Traes_5AL_F1F202C88.1	Y	1.65E-36	Y	4.13E-29	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_374080_AA1189810.1	Traes_5AL_5DE16F8EA.2	Y	6.35E-07	Y	1.20E-10	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_374097_AA1190260.1	Traes_5AL_1F7681FE3.1	Y	4.64E-07	NA	NA	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_374319_AA1196780.1	Traes_5AL_FCDD18A4D.1	Y	1.92E-03	Y	2.36E-03	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_374542_AA1202810.4	Traes_5AL_00CC4E7C6.1	Y	3.61E-07	Y	3.69E-11	CSS 5' truncation

Table 4.2 continued on next page

Table 4.2 continued from previous page

TGAC transcript ID	CSS transcript ID	DE in TGAC?	TGAC q value	DE in CSS?	CSS q value	Comparison class
TRIAE_CS42_5AL_TGACv1_375361_AA1220430.2	Traes_5AL_385883702.1	Y	4.21E-02	Y	3.33E-04	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_375845_AA1227980.1	Traes_5AL_2EDDF65BE.2	Y	1.00E-03	Y	7.99E-06	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_375845_AA1227990.1	Traes_5AL_AC299D3FF.1	Y	4.83E-15	Y	7.37E-11	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_376402_AA1236390.1	Traes_5AL_DD1665D87.2	Y	7.07E-04	Y	2.34E-02	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_376619_AA1239170.1	Traes_5AL_B8B668113.1	N	1.00E+00	Y	4.52E-02	CSS 5' truncation
TRIAE_CS42_5AL_TGACv1_376953_AA1242690.2	Traes_5AL_F09A6AECA.2	Y	1.53E-13	Y	9.28E-38	CSS 5' truncation
TRIAE_CS42_1AS_TGACv1_019083_AA0060340.1	NC	Y	1.42E-05	NC	NC	CSS missing
TRIAE_CS42_2AL_TGACv1_095650_AA0313320.1	NC	Y	4.87E-04	NC	NC	CSS missing
TRIAE_CS42_2AL_TGACv1_095650_AA0313330.1	NC	Y	4.94E-03	NC	NC	CSS missing
TRIAE_CS42_2AL_TGACv1_095956_AA0316040.1	NC	Y	5.16E-09	NC	NC	CSS missing
TRIAE_CS42_2AL_TGACv1_095956_AA0316050.1	NC	Y	1.11E-10	NC	NC	CSS missing
TRIAE_CS42_2DS_TGACv1_177196_AA0568430.1	NC	Y	4.57E-06	NC	NC	CSS missing
TRIAE_CS42_2DS_TGACv1_177488_AA0578600.1	NC	Y	4.65E-02	NC	NC	CSS missing
TRIAE_CS42_3AS_TGACv1_211411_AA0689940.1	NC	Y	1.29E-02	NC	NC	CSS missing
TRIAE_CS42_3DL_TGACv1_250330_AA0866270.1	NC	Y	2.65E-02	NC	NC	CSS missing
TRIAE_CS42_3DL_TGACv1_250633_AA0871340.1	NC	Y	9.45E-04	NC	NC	CSS missing
TRIAE_CS42_4BL_TGACv1_321219_AA1057420.1	NC	Y	2.23E-02	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374025_AA1188070.1	NC	Y	8.44E-04	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374097_AA1190230.1	NC	Y	2.32E-05	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374249_AA1195190.1	NC	Y	8.01E-26	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374374_AA1198360.1	NC	Y	8.15E-04	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374446_AA1200410.1	NC	Y	4.06E-10	NC	NC	CSS missing

Table 4.2 continued on next page

Table 4.2 continued from previous page

TGAC transcript ID	CSS transcript ID	DE in TGAC?	TGAC q value	DE in CSS?	CSS q value	Comparison class
TRIAE_CS42_5AL_TGACv1_374657_AA1205780.1	NC	Y	4.65E-02	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_374675_AA1206250.1	NC	Y	1.63E-02	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_375085_AA1215790.1	NC	Y	5.97E-04	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_375493_AA1222690.2	NC	Y	1.83E-06	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_375721_AA1226170.1	NC	Y	3.46E-04	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_375857_AA1228100.1	NC	Y	3.43E-03	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_376076_AA1231790.1	NC	Y	1.08E-03	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_376877_AA1241920.1	NC	Y	2.33E-02	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_376877_AA1241930.1	NC	Y	2.47E-30	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_376953_AA1242680.1	NC	Y	1.62E-17	NC	NC	CSS missing
TRIAE_CS42_5AL_TGACv1_376953_AA1242690.1	NC	Y	1.59E-13	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393235_AA1270150.1	NC	Y	3.09E-05	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393473_AA1272910.1	NC	Y	1.80E-05	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393577_AA1274040.1	NC	Y	1.13E-26	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393580_AA1274130.1	NC	Y	1.17E-18	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393696_AA1275280.2	NC	Y	1.67E-35	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393726_AA1275550.1	NC	Y	3.48E-10	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393783_AA1275990.1	NC	Y	4.55E-22	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_393897_AA1277010.1	NC	Y	7.13E-07	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_394352_AA1279770.1	NC	Y	1.79E-04	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_394531_AA1280840.1	NC	Y	6.00E-03	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_395074_AA1282530.1	NC	Y	6.69E-18	NC	NC	CSS missing

Table 4.2 continued on next page

Table 4.2 continued from previous page

TGAC transcript ID	CSS transcript ID	DE in TGAC?	TGAC q value	DE in CSS?	CSS q value	Comparison class
TRIAE_CS42_5AS_TGACv1_395084_AA1282570.1	NC	Y	7.37E-03	NC	NC	CSS missing
TRIAE_CS42_5AS_TGACv1_402215_AA1284070.1	NC	Y	2.32E-05	NC	NC	CSS missing
TRIAE_CS42_6AL_TGACv1_471516_AA1510240.1	NC	Y	3.11E-03	NC	NC	CSS missing
TRIAE_CS42_6BL_TGACv1_503194_AA1627460.1	NC	Y	1.12E-05	NC	NC	CSS missing
TRIAE_CS42_7AS_TGACv1_571015_AA1843630.1	NC	Y	1.55E-02	NC	NC	CSS missing
TRIAE_CS42_7BL_TGACv1_577371_AA1873630.1	NC	Y	5.06E-05	NC	NC	CSS missing
TRIAE_CS42_7BS_TGACv1_592547_AA1940160.1	NC	Y	2.39E-04	NC	NC	CSS missing
TRIAE_CS42_7DS_TGACv1_621701_AA2023630.1	NC	Y	6.64E-05	NC	NC	CSS missing
TRIAE_CS42_U_TGACv1_641674_AA2101090.1	NC	Y	6.67E-05	NC	NC	CSS missing
TRIAE_CS42_2BS_TGACv1_148693_AA0494610.1	Traes_2BS_009718F07.2	Y	2.65E-02	N	1.00E+00	CSS split
TRIAE_CS42_2BS_TGACv1_148693_AA0494610.1	Traes_2BS_2272AAEE2.2	Y		N	1.00E+00	CSS split
TRIAE_CS42_5AL_TGACv1_374231_AA1194360.2	Traes_5AL_C1E3FCB4F.1	Y	4.60E-03	N	1.00E+00	CSS split
TRIAE_CS42_5AL_TGACv1_374231_AA1194360.2	Traes_5AL_CD19FF15F.1	Y		N	1.00E+00	CSS split
TRIAE_CS42_5AL_TGACv1_374321_AA1196890.1	Traes_5AL_4B1DD2A62.1	Y	3.09E-12	N	NA	CSS split
TRIAE_CS42_5AL_TGACv1_374321_AA1196890.1	Traes_5AL_1D8F705CE.1	Y		Y	2.27E-05	CSS split
TRIAE_CS42_5AL_TGACv1_374321_AA1196890.1	Traes_5AL_2EC36FC33.1	Y		Y	1.89E-11	CSS split
TRIAE_CS42_5AL_TGACv1_374321_AA1196890.1	Traes_5AL_E24DCCFF0.1	Y		N	1.10E-01	CSS split
TRIAE_CS42_5AL_TGACv1_375394_AA1220930.4	Traes_5AL_67878B82B.1	Y	0.00E+00	N	1.00E+00	CSS split
TRIAE_CS42_5AL_TGACv1_375394_AA1220930.4	Traes_5AL_0C2D144B0.1	Y		N	1.00E+00	CSS split
TRIAE_CS42_5AS_TGACv1_393119_AA1268700.1	Traes_5AS_AF0876292.1	Y	3.11E-03	N	1.00E+00	CSS split
TRIAE_CS42_5AS_TGACv1_393119_AA1268700.1	Traes_5AS_25E2451D6.1	Y		N	1.00E+00	CSS split
TRIAE_CS42_7AL_TGACv1_556075_AA1754700.1	Traes_7AL_A29227860.2	Y	1.25E-02	N	1.00E+00	CSS split

Table 4.2 continued on next page

Table 4.2 continued from previous page

TGAC transcript ID	CSS transcript ID	DE in TGAC?	TGAC q value	DE in CSS?	CSS q value	Comparison class
TRIAE_CS42_7AL_TGACv1_556075_AA1754700.1	Traes_7AL_773C8EC8C.1	Y		N	NA	CSS split
TRIAE_CS42_5AL_TGACv1_374233_AA1194500.1	Traes_5AL_158704A70.1	N	NA	Y	5.32E-13	CSS structure change
TRIAE_CS42_5AL_TGACv1_374413_AA1199590.1	Traes_5AL_A9CF39101.1	Y	3.27E-02	N	1.00E+00	CSS structure change
TRIAE_CS42_1BS_TGACv1_049354_AA0149980.1	Traes_1BS_C59E4945B.2	Y	4.32E-02	Y	3.00E-02	No change
TRIAE_CS42_5AL_TGACv1_375845_AA1227940.1	Traes_5AL_1C4AB8F62.1	N	1.00E+00	Y	2.40E-03	No change
TRIAE_CS42_5AL_TGACv1_376107_AA1232210.1	Traes_5AL_70C442FE1.2	N	NA	Y	2.27E-03	No change
TRIAE_CS42_5AS_TGACv1_392558_AA1260860.1	Traes_5AS_34C5341E4.1	Y	3.90E-47	Y	3.57E-51	No change
TRIAE_CS42_5AS_TGACv1_394776_AA1281770.1	Traes_5AS_78CA97493.2	Y	1.43E-06	Y	6.58E-07	No change
TRIAE_CS42_5AS_TGACv1_394776_AA1281770.3	Traes_5AS_78CA97493.2	Y	4.11E-02	Y	6.58E-07	No change
TRIAE_CS42_5AS_TGACv1_393572_AA1273920.1	Traes_5AS_BD279FFF4.2	Y	2.41E-13	Y	6.05E-16	TGAC 5' truncation
NC	Traes_5AL_78644A5C4.1	NC	NC	Y	3.20E-08	TGAC missing
NC	Traes_5AL_BAB11D9B4.3	NC	NC	Y	2.10E-02	TGAC missing
NC	Traes_5BS_3B409615C.1	NC	NC	Y	1.95E-05	TGAC missing
NC	Traes_2AS_8F1446457.2	NC	NC	Y	1.19E-04	TGAC missing
NC	TRAES3BF002600020CFD_t1	NC	NC	Y	4.44E-02	TGAC missing
TRIAE_CS42_5AL_TGACv1_374727_AA1207530.1	Traes_5AL_F8182F2FB.1	Y	2.43E-03	Y	2.15E-02	CSS fused
TRIAE_CS42_5AL_TGACv1_374727_AA1207540.1	Traes_5AL_F8182F2FB.1	N	1.00E+00	Y		CSS fused
TRIAE_CS42_5AL_TGACv1_376411_AA1236550.1	Traes_5AL_58BA759B9.5	Y	1.06E-10	Y	1.82E-64	CSS fused
TRIAE_CS42_5AL_TGACv1_376411_AA1236560.1	Traes_5AL_58BA759B9.5	Y	1.20E-03	Y		CSS fused

For 64 of the TGAC DE transcripts no equivalent CSS DE transcript was identified, either because there was no corresponding CSS gene model (47 transcripts) or the expression change between NILs was non-significant for the CSS transcript. Analogously, eleven CSS DE transcripts did not have an equivalent TGAC gene model DE, five of which were due to there being no corresponding TGAC gene model annotated. Combining both sets identified 42 groups of equivalent gene models, 26 of which were differentially expressed in both alignments. Comparing these 42 groups and taking into account fused and split gene models within each dataset, there were 97 gene models in both datasets (50 CSS + 47 TGAC) (Figure 4.2a, Table 4.2). Of these, only six were identical between the CSS and TGAC references. All other discrepant gene models fell under categories included truncations in either reference, gene models that were split/fused in one reference sequence, and gene models that differed drastically in their overall structure.

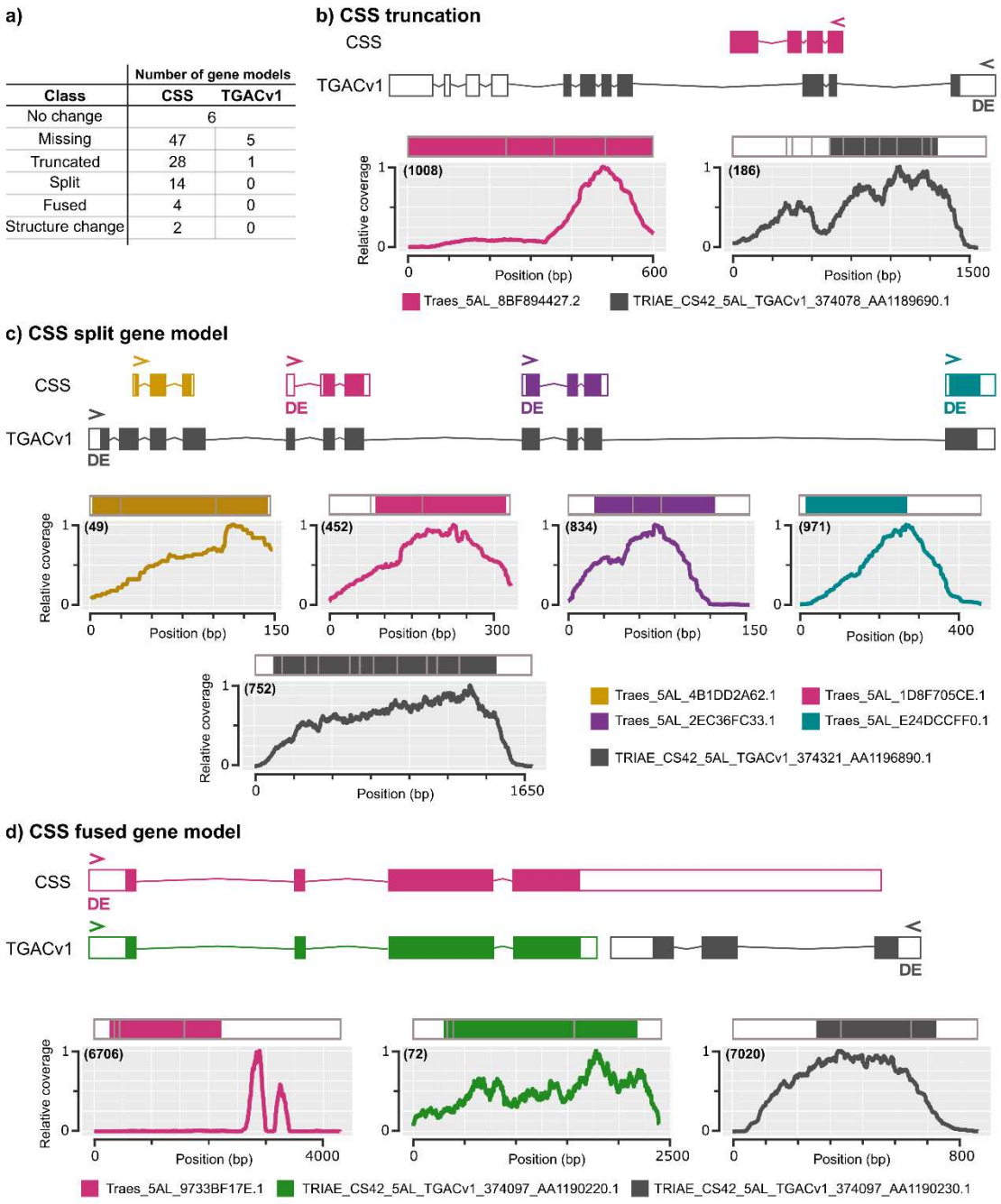


Figure 4.2: Comparison between CSS and TGACv1 gene models

a) Discrepancies identified between gene models in the CSS and TGAC reference sequences and the number of gene models falling into categories. Panels b), c) and d) show specific examples of discrepancies. In each panel, a representation of the unspliced gene model is shown with exons as coloured boxes, untranslated regions as white boxes, and introns as thin lines. Graphs show the relative read coverage across the spliced transcript with the structure represented diagrammatically directly above each graph. The number in brackets shows the maximum absolute read depth for each gene model. > and < in the gene structures indicate the direction of transcription and a 'DE' indicates that the gene model was differentially expressed in T1 $^{5A-}_{5A+}$ (q value < 0.05). For each panel transcript names are shown in the coloured legends.

For all discrepant gene models, transcriptome read mapping and an interspecies comparison was used to determine which gene model seemed most plausible. Figure 4.2b shows an example of the most commonly identified discrepancy where a gene model was truncated in the CSS reference (pink) relative to the TGAC reference (grey). The DE TGAC gene model was supported by this transcriptome data as read coverage was observed across the whole gene model whilst the coverage across the CSS gene model dropped at the position where an intron is predicted in the TGAC model. Another common discrepancy was a single gene model in one reference being split into multiple gene models in the other reference. Figure 4.2c shows an instance where a single DE TGAC gene model comprised four separate CSS gene models. In this case, all five gene models had coverage across the entire gene body, however the single TGAC gene model was more similar to proteins from other species, suggesting that this single gene model was most likely correct. The final example (Figure 4.2d) shows two TGAC gene models that were fused into a single CSS gene model. The coverage across the CSS gene model was inconsistent, with most reads concentrated in the 3' untranslated region (UTR). The two TGAC gene models had more consistent coverage across the entire gene models and were both supported by protein alignments with other species. Interestingly, only the shorter TGAC gene model was DE (Figure 4.2d, grey), suggesting that differential expression of the CSS gene model was driven by the reads mapping to the putative 3' UTR rather than the coding regions of the transcript (Figure 4.2d, pink). Taking together the fact that a higher percentage of reads mapped to the TGAC gene models and that many more of the examined TGAC gene models were supported by interspecies comparison and expression data than the CSS gene models, all further analysis used the alignments to the TGAC gene models only.

4.4.3 Many DE transcripts during early grain development are shared between NILs

3,151 and 2,789 DE transcripts were identified across early grain development in 5A-$^{T1}_{T2}$ and 5A+$^{T1}_{T2}$, respectively (Figure 4.1, Figure 4.3a). The DE transcripts were evenly distributed across the 21 chromosomes, showing no overall bias towards any chromosome group or subgenome (Figure 4.3b). Approximately 60% (1,832) of the DE transcripts were shared between 5A-$^{T1}_{T2}$ and 5A+$^{T1}_{T2}$ (Figure 4.3a) and 84% (1,532) of the shared transcripts were upregulated across time (Figure 4.3c). 41 significantly enriched GO terms were identified in the upregulated transcripts (Table 4.3). Sixteen of the GO terms were associated with biological process and could be grouped under three parent GO terms: metabolic process (GO:0008152), defence response (GO:0006952) and biological regulation (GO:0065007) (Table 4.3; Figure 4.3c). Within metabolic process we found terms associated with carbohydrate (GO:0005975) and pyruvate metabolism (GO:0006090), vitamin E (GO:0010189) and triglyceride biosynthesis (GO:0019432), mRNA catabolism (GO:0006402), proteolysis (GO:0006508) and phosphorylation (GO:0016310). Downregulated transcripts (300) were enriched for seven GO terms, four of which were associated with biological process: potassium ion transport (GO:0006813), signal transduction (GO:0007165), phosphorelay

signal transduction (GO:0000160) and carbohydrate metabolism (GO:0005975) (Figure 4.3c, Table 4.4). The overlap between enriched GO terms in the upregulated and downregulated transcripts (e.g. carbohydrate metabolism) suggests that different aspects of these processes are being differentially regulated during this early stage of grain development.

Many transcripts identified were only DE across early grain development in one of the two genotypes (i.e. unique to either the 5A- $^{T1}_{T2}$ or 5A+ $^{T1}_{T2}$ comparisons). However, many of these transcripts were borderline non-significant in the opposite genotype comparison illustrated by the fact that the distributions of q values were skewed towards significance (Figure 4.4). Additionally, the uniquely DE transcripts were enriched for GO terms similar to the shared transcripts (Table 4.5, Table 4.6). Some GO terms, however, were only enriched in the uniquely DE transcripts, for example, cell wall organisation or biosynthesis (GO:0071554) and response to abiotic stimulus (GO:0009628). Overall, these results suggest that although there were some differences between genotypes, broadly similar biological processes were taking place in the grains of both the 5A NILs at the early stages of grain development.

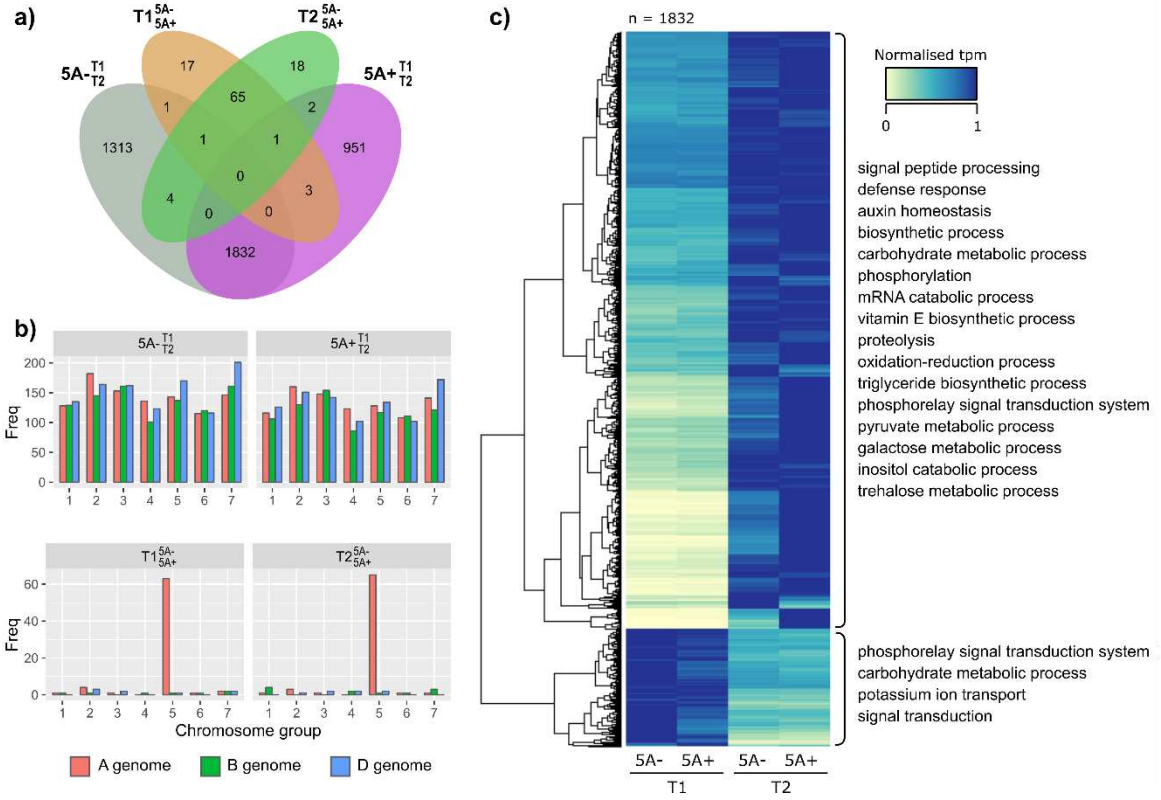


Figure 4.3: Overview of differentially expressed transcripts

a) Venn diagram of differentially expressed (DE) transcripts ($q < 0.05$) identified in 4 pairwise comparisons: T1 $^{5A-}_{5A+}$ (orange), T2 $^{5A-}_{5A+}$ (green), 5A- $^{T1}_{T2}$ (grey) and 5A+ $^{T1}_{T2}$ (purple). b) Number of DE transcripts located on each chromosome for all comparisons. The 5A- $^{T1}_{T2}$ and 5A+ $^{T1}_{T2}$ DE transcripts (top graphs) are evenly distributed across all 21 chromosomes whereas T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ DE transcripts (bottom graphs) are concentrated on chromosome 5A. c) Heatmap of normalised tpm (transcripts per million) of common DE transcripts in 5A- $^{T1}_{T2}$ and 5A+ $^{T1}_{T2}$ (n = 1,832). Hierarchical clustering separated these into transcripts that were upregulated (n = 1,532) and downregulated (n = 300) across time. Significantly enriched GO terms (biological function only) for each group are shown on the right of the heatmap.

Table 4.3: Enriched gene ontology (GO) terms in common upregulated transcripts differentially expressed (DE) in the 5A-$\frac{T1}{T2}$ and 5A+$\frac{T1}{T2}$ comparisons (n = 1,532)

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0006952	72	162	defence response	BP	3.07E-83
GO:0000160	32	292	phosphorelay signal transduction system Biological regulation	BP	1.21E-15
GO:0006508	64	1455	Proteolysis Metabolic process	BP	8.87E-12
GO:0006465	8	27	signal peptide processing Metabolic process	BP	1.02E-06
GO:0055114	90	3409	oxidation-reduction process Metabolic process	BP	1.24E-05
GO:0005975	46	1455	carbohydrate metabolic process Metabolic process	BP	2.52E-04
GO:0006402	6	27	mRNA catabolic process Metabolic process	BP	2.52E-04
GO:0010189	3	3	vitamin E biosynthetic process Metabolic process	BP	3.22E-04
GO:0010252	3	5	auxin homeostasis Biological regulation	BP	2.62E-03
GO:0009058	17	372	biosynthetic process Metabolic process	BP	4.22E-03
GO:0016310	7	66	phosphorylation Metabolic process	BP	4.22E-03
GO:0019432	3	6	triglyceride biosynthetic process Metabolic process	BP	4.63E-03
GO:0006090	4	20	pyruvate metabolic process Metabolic process	BP	1.27E-02
GO:0006012	5	43	galactose metabolic process Metabolic process	BP	2.61E-02
GO:0005991	2	3	trehalose metabolic process Metabolic process	BP	3.60E-02
GO:0019310	2	3	inositol catabolic process Metabolic process	BP	3.60E-02
GO:0030014	6	16	CCR4-NOT complex	CC	1.22E-05
GO:0005787	5	14	signal peptidase complex	CC	1.55E-04
GO:0030904	3	8	retromer complex	CC	1.15E-02
GO:0004857	56	164	enzyme inhibitor activity	MF	9.44E-57
GO:0004869	23	67	cysteine-type endopeptidase inhibitor activity	MF	1.77E-22
GO:0004190	36	308	aspartic-type endopeptidase activity	MF	1.39E-18

Table 4.3 continued on next page

Table 4.3 continued from previous page

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0004871	15	104	signal transducer activity	MF	1.42E-08
GO:0016813	6	10	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines	MF	5.91E-07
GO:0045735	9	40	nutrient reservoir activity	MF	1.34E-06
GO:0008233	12	96	peptidase activity	MF	4.38E-06
GO:0004867	7	24	serine-type endopeptidase inhibitor activity	MF	7.52E-06
GO:0020037	33	767	heme binding	MF	1.29E-05
GO:0030170	18	273	pyridoxal phosphate binding	MF	2.82E-05
GO:0016705	25	523	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	MF	6.06E-05
GO:0051741	3	3	2-methyl-6-phytyl-1,4-benzoquinone methyltransferase activity	MF	3.22E-04
GO:0050242	4	11	pyruvate, phosphate dikinase activity	MF	1.32E-03
GO:0005506	27	738	iron ion binding	MF	2.20E-03
GO:0008483	8	84	transaminase activity	MF	3.04E-03
GO:0008237	7	66	metallopeptidase activity	MF	4.22E-03
GO:0008234	11	222	cysteine-type peptidase activity	MF	3.45E-02
GO:0004555	2	3	alpha,alpha-trehalase activity	MF	3.60E-02
GO:0050113	2	3	inositol oxygenase activity	MF	3.60E-02
GO:0016772	7	99	transferase activity, transferring phosphorus-containing groups	MF	3.87E-02
GO:0003978	4	29	UDP-glucose 4-epimerase activity	MF	4.23E-02
GO:0004553	26	881	hydrolase activity, hydrolyzing O-glycosyl compounds	MF	4.84E-02

DE transcripts =DE transcripts associated with the GO term, Total transcripts = all transcripts associated with the GO term. BP = Biological Process, CC = Cellular Component, MF = Molecular Function. Superscripts are a common parent GO term. Adjusted P-values were calculated using the Benjamini Hochberg procedure.

Table 4.4: Enriched gene ontology (GO) terms in common downregulated transcripts differentially expressed (DE) in the 5A- $\frac{T1}{T2}$ and 5A+ $\frac{T1}{T2}$ comparisons (n = 300)

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0005975	17	1455	carbohydrate metabolic process [Metabolic process]	BP	1.61E-03
GO:0000160	8	292	phosphorelay signal transduction system [Biological regulation]	BP	1.61E-03
GO:0007165	8	404	signal transduction [Biological regulation]	BP	1.18E-02
GO:0006813	3	33	potassium ion transport [Cation transport]	BP	4.03E-02
GO:0004553	15	881	hydrolase activity, hydrolyzing O-glycosyl compounds	MF	1.84E-04
GO:0043531	18	1396	ADP binding	MF	2.97E-04
GO:0005249	3	21	voltage-gated potassium channel activity	MF	1.18E-02

DE transcripts =DE transcripts associated with the GO term, Total transcripts = all transcripts associated with the GO term. BP = Biological Process, MF = Molecular Function. Superscripts are a common parent GO term. Adjusted P-values were calculated using the Benjamini Hochberg procedure.

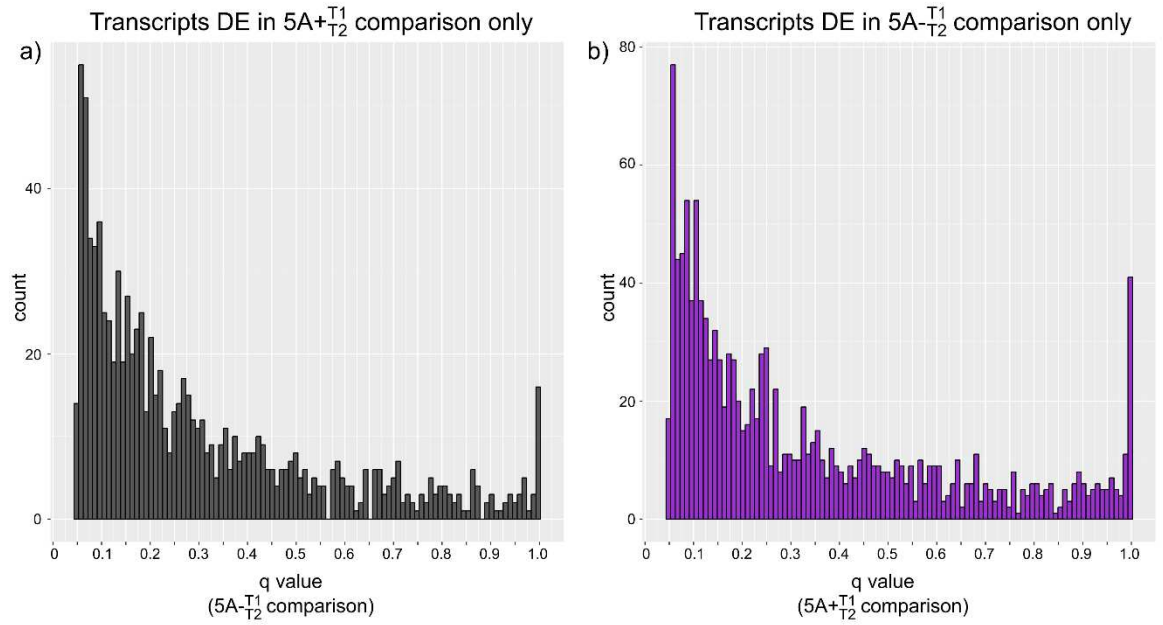


Figure 4.4: Distributions of q values of uniquely differentially expressed transcripts in the 5A-$\frac{T1}{T2}$ and 5A+$\frac{T1}{T2}$ comparisons

a) Distribution of 5A-$\frac{T1}{T2}$ q values for transcripts that were differentially expressed (DE) only in the 5A+$\frac{T1}{T2}$ comparison (and not the 5A-$\frac{T1}{T2}$ comparison). b) Distribution of 5A-$\frac{T1}{T2}$ q values for DE transcripts across time in the 5A-$\frac{T1}{T2}$ comparison only. The fact that both distributions are skewed towards lower q values shows suggests that many of the DE genes within a single comparison were borderline non-significant in the opposite comparison.

Table 4.5: Enriched gene ontology (GO) terms in transcripts uniquely differentially expressed (DE) in the 5A- $\frac{T1}{T2}$ ***comparison (n = 1,319)***

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0005975	53	1455	carbohydrate metabolic process Metabolic process	BP	8.71E-08
GO:0042546	10	77	cell wall biogenesis	BP	3.97E-05
GO:0010411	9	64	xyloglucan metabolic process Metabolic process	BP	5.56E-05
GO:0006073	9	71	cellular glucan metabolic process Metabolic process	BP	9.07E-05
GO:0000160	17	292	phosphorelay signal transduction system Biological regulation	BP	9.07E-05
GO:0019538	8	64	protein metabolic process Metabolic process	BP	3.50E-04
GO:0009664	5	22	plant-type cell wall organisation	BP	1.41E-03
GO:0009765	8	91	photosynthesis, light harvesting Metabolic process	BP	4.06E-03
GO:0009688	3	6	abscisic acid biosynthetic process Metabolic process	BP	5.06E-03
GO:0009415	3	6	response to water	BP	5.06E-03
GO:0006952	9	162	defence response	BP	3.22E-02
GO:0034551	2	3	mitochondrial respiratory chain complex III assembly	BP	4.37E-02
GO:0009638	2	3	phototropism	BP	4.37E-02
GO:0055114	68	3409	oxidation-reduction process Metabolic process	BP	4.88E-02
GO:0005576	13	132	extracellular region	CC	1.77E-05
GO:0005618	12	130	cell wall	CC	5.56E-05
GO:0048046	10	84	apoplast	CC	5.56E-05
GO:0004553	37	881	hydrolase activity, hydrolyzing O-glycosyl compounds	MF	9.15E-07
GO:0004857	13	164	enzyme inhibitor activity	MF	8.72E-05
GO:0016762	9	72	xyloglucan:xyloglucosyl transferase activity	MF	9.07E-05
GO:0004556	5	23	alpha-amylase activity	MF	1.59E-03
GO:0009540	3	6	zeaxanthin epoxidase [overall] activity	MF	5.06E-03

Table 4.5 continued on next page

Table 4.5 continued from previous page

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0020037	25	767	heme binding	MF	5.06E-03
GO:0004871	8	104	signal transducer activity	MF	7.69E-03
GO:0005506	22	738	iron ion binding	MF	8.71E-08

DE transcripts =DE transcripts associated with the GO term, Total transcripts = all transcripts associated with the GO term. BP = Biological Process, MF = Molecular Function. Superscripts are a common parent GO term. Adjusted P-values were calculated using the Benjamini Hochberg procedure.

Table 4.6: Enriched gene ontology (GO) terms in transcripts uniquely differentially expressed (DE) in the 5A+ $\frac{T1}{T2}$ comparison (n = 957)

GO term	DE transcripts	Total transcripts	Term ID	Ontology	Adjusted P-value
GO:0005978	7	45	glycogen biosynthetic process [Metabolic process]	BP	2.16E-04
GO:0005975	35	1455	carbohydrate metabolic process [Metabolic process]	BP	1.41E-03
GO:0055114	58	3409	oxidation-reduction process [Metabolic process]	BP	3.29E-02
GO:0010155	3	10	regulation of proton transport [Cation transport]	BP	3.37E-02
GO:0004553	30	881	hydrolase activity, hydrolyzing O-glycosyl compounds	MF	2.09E-05
GO:0008878	6	28	glucose-1-phosphate adenylyltransferase activity	MF	2.16E-04
GO:0016491	39	2012	oxidoreductase activity	MF	3.29E-02

DE transcripts =DE transcripts associated with the GO term, Total transcripts = all transcripts associated with the GO term. BP = Biological Process, MF = Molecular Function. Superscripts are a common parent GO term. Adjusted P-values were calculated using the Benjamini Hochberg procedure.

4.4.4 DE transcripts between NILs are concentrated on chromosome 5A

88 and 91 DE transcripts were identified between the NILs in T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$, respectively, many fewer than identified in 5A- $^{T1}_{T2}$ or 5A+ $^{T1}_{T2}$. This was expected as the NILs are genetically very similar and therefore the difference in developmental stage between the T1 and T2 time points results in greater changes in gene expression. Of these 179 DE transcripts, 67 were common between T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$, whereas 45 DE transcripts between genotypes were unique and identified only at a single time point (resulting in 112 DE transcripts between NILs at any time point; Figure 4.3a, Figure 4.5a). No GO terms were significantly enriched in these groups. Of the 67 common DE transcripts, 54 (80%) were located on chromosome 5A (Figure 4.3b, Figure 4.5a), whilst in both the T1 and T2 unique groups less than 50% were located on chromosome 5A (Figure 4.5a). Similar numbers of DE transcripts were more highly expressed in either genotype, with no distinct patterns observed between the unique or common groups.

Of the 74 DE transcripts located on chromosome 5A all were located within the 491 Mbp introgressed region of the NILs (Figure 4.5b). Higher numbers of DE transcripts were identified in regions of increased SNP density between the 5A NILs. In the previous chapter, the grain length effect was fine-mapped to a 75 Mbp interval on 5AL (between *BS00182017* (317 Mbp) and *JBRNASeq_4* (393 Mb)) and eight of the DE transcripts were located within this interval. Six of the transcripts were located in *GL1* interval and two in the *GL2* interval with respect to the 'two-gene' hypothesis proposed in Chapter 3 (Figure 4.5b). Of the eight transcripts, three were more highly expressed in the 5A+ NILs ($5A+_{high}$ transcripts), two of which were transcript variants of the same gene (a kinesin-like protein; only .2 variant shown in Figure 4.5b). The other $5A+_{high}$ transcript was annotated as a putative retrotransposon protein. One of the five transcripts more highly expressed in the 5A- NIL ($5A-_{high}$ transcripts) had no annotation and the remaining four were annotated as a non-coding RNA, a RING/U-box containing protein, a TauE-like protein and a DUF810 family protein.

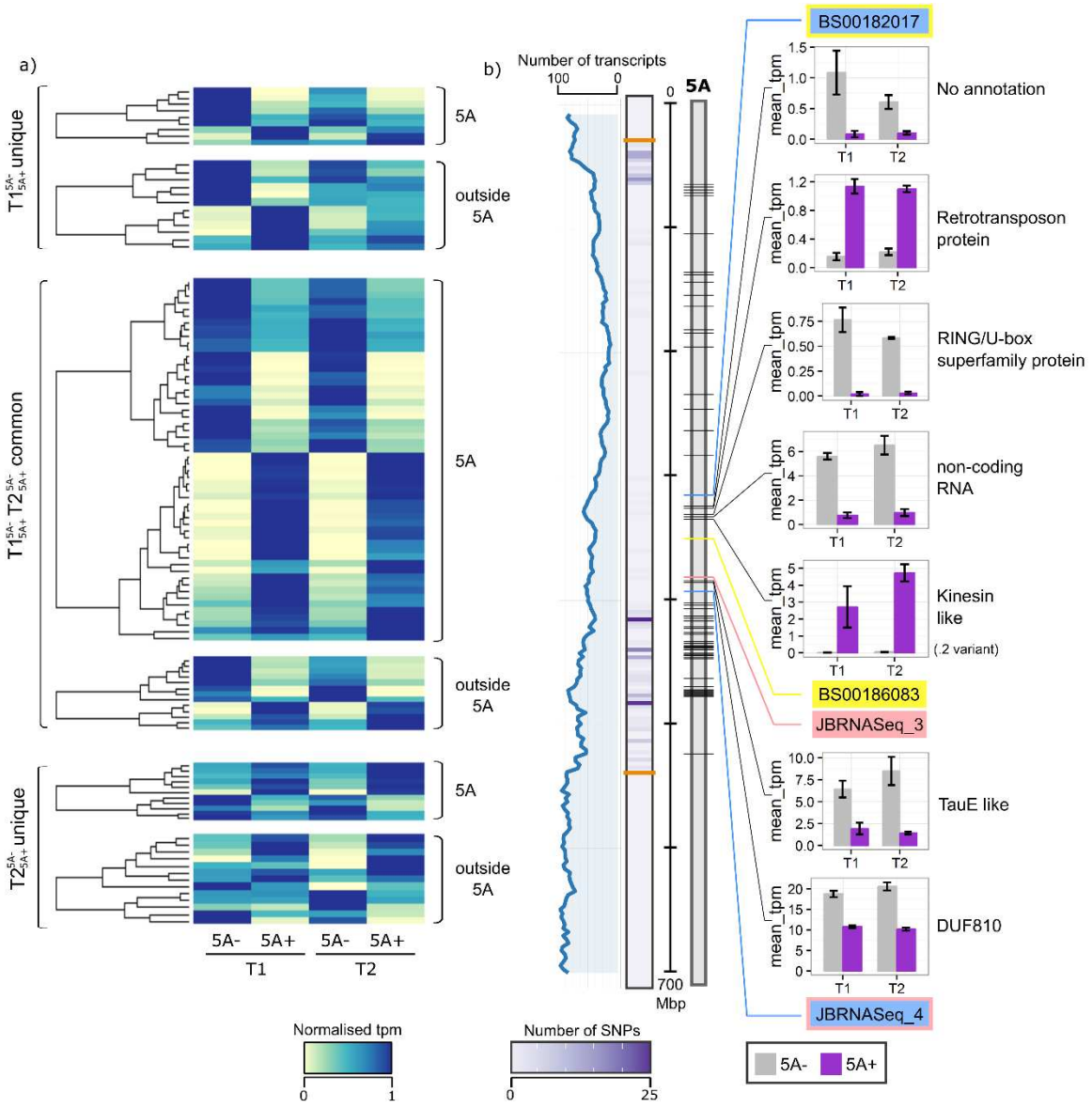


Figure 4.5: Differentially expressed transcripts between 5A NILs at T1 and T2

a) Heatmap of normalised tpm (transcripts per million) of DE (differentially expressed) transcripts between NILs (T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ comparisons). Transcripts are first grouped based on whether they were differentially expressed at both time points (T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ common) or at only T1 or T2 (T1 $^{5A-}_{5A+}$ unique and T2 $^{5A-}_{5A+}$ unique, respectively), and then whether they are located on chromosome 5A or not. b) Location of DE transcripts on chromosome 5A (black lines on grey rectangle). Line graph (blue) shows rolling mean of the number of transcripts located in 3 Mbp bins across chromosome 5A, alongside heatmap which shows the number of 90k iSelect SNPs between the 5A- and 5A+ NILs in similar sized bins. Orange lines on the SNP heatmap define the 491 Mbp introgression which differs between then NILs. Blue lines on the chromosome indicate the positions of the flanking markers of the overall fine-mapped region of the 5A grain length QTL (BS00182017 and JBRNASeq_4). The yellow and pink lines indicate the internal flanks of *GL1* and *GL2*, respectively (BS00186083 and JBRNASeq_3) of the 'two-gene' hypothesis. Bar charts show the mean tpm values at T1 and T2 of DE transcripts located in the fine-mapped region (5A- NILs in grey, 5A+ NILs in purple). Only one transcript variant (.2) of the kinesin-like gene is shown. Error bars are standard error of the three biological replicates.

4.4.5 DE transcripts outside of chromosome 5A are enriched in specific transcription factor binding sites

As all the DE transcripts on chromosome 5A were located within the 491 Mbp introgressed region, it is possible that the differential expression was a direct consequence of sequence variation between the NILs e.g. in the promoter regions. However, the 38 DE transcripts located outside of chromosome 5A have the same nucleotide sequence as they are identical by descent (BC_4 NILs confirmed with 90k iSelect SNP array data; Brinton *et al.*, 2017). It was hypothesised that these DE transcripts located outside of the 5A introgression are downstream targets of genes, such as transcription factors (TFs), located within the 5A introgression.

To assess this, transcription factor binding sites (TFBS) were identified in the promoter regions of these 38 DE transcripts. The TFBS identified in this group of transcripts were associated with 91 distinct TF families (Table 4.7), five of which were enriched relative to all expressed transcripts (Table 4.7; adjusted $P < 0.05$). The enriched TFBS families were C2H2, Myb/SANT, AT-Hook, YABBY and MADF/Trihelix.

Table 4.7: Enriched transcription factor binding sites in the promoters of differentially expressed located outside of 5A

TFBS family	Outside 5A DE transcripts (n=38)	All expressed transcripts (n=101,653)	Adjusted P-value
C2H2	36	77987	0.021
Myb/SANT	38	88575	0.021
AT-Hook	38	90203	0.028
YABBY	15	19447	0.034
MADF;Trihelix	13	16632	0.042

Values are the number of transcripts in which binding sites associated with the specified transcription factor (TF) family are present. Adjusted P-values were calculated using the Benjamini Hochberg procedure

To determine potential candidates for upstream regulators all annotated TFs located within the introgressed region on chromosome 5A were identified (Borrill *et al.*, 2017). A total of 200 annotated TFs were identified, belonging to 35 TF families (Table 4.8). Of these, four families (across 29 genes) overlapped with enriched TFBS families. Four of the 29 TFs were located within the fine-mapped grain length interval on chromosome 5A, including C2H2, MYB and MYB_related TFs (Table 4.8). Of these, the MYB and MYB_related TFs were located within the *GL1* interval. None of the TFs located in the *GL2* interval overlapped with TF families with enriched binding sites in the outside 5A DE transcripts.

Table 4.8: Transcription factors identified in the 5A NIL introgression

	Number of genes			
TF family	**Introgression**	**Overall grain length interval**	***GL1* interval**	***GL2* interval**
AP2	2	-	-	-
B3	14	1	-	-
bHLH	18	-	-	-
bZIP	13	3	1	-
C2H2*	10	2	-	-
C3H	4	-	-	-
CO-like	1	1	1	-
CPP	2	-	-	-
DBB	1	-	-	-
Dof	3	1	1	-
E2F/DP	1	-	-	-
EIL	1	-	-	-
ERF	30	3	-	-
FAR1	20	2	2	-
G2-like	6	3	1	-
GATA	1	-	-	-
GeBP	1	-	-	-
GRAS	3	-	-	-
HB-other	1	-	-	-
HD-ZIP	4	-	-	-
HSF	3	-	-	-
LBD	3	1	1	-
MIKC	2	-	-	-
MYB*	10	1	1	-
MYB_related*	8	1	1	-
NAC	13	2	2	-
NF-YC	1	-	-	-
SBP	1	-	-	-
SRS	1	-	-	-
TALE	1	-	-	-
TCP	3	-	-	-
Trihelix*	1	-	-	-
WOX	2	1	1	-
WRKY	11	6	4	2
ZF-HD	4	-	-	-

A * indicates that transcription factor (TF) binding sites associated with the TF family were significantly enriched in the promoters of transcripts that were differentially expressed between NILs and located outside of chromosome 5A

4.4.6 Functional annotation of DE transcripts

Having analysed DE transcripts between NILs based on chromosome location, the 112 DE transcripts were examined based on their functional annotations. Multiple categories of annotations were identified including transcripts associated with ubiquitin-mediated protein degradation, cell cycle, metabolism, transport, transposons and non-coding RNAs (Table 4.9; full annotations in Table 4.11). Few categories were exclusively located on/outside 5A or had exclusively higher expression in either the 5A- or 5A+ NIL.

Table 4.9: Categories of DE transcripts between NILs based on predicted function

Category	number of transcripts	Adjusted P-value	5A/not 5A	NIL with higher expression: 5A-/5A+
non-coding RNA	15	0.141	10/5	6/9
transposon-associated	14	0.008	4/10	5/9
ubiquitin	12**	0.008	10/2	8/4
cell cycle	5	-	5/0	2/3
histone-related	5	-	3/2	3/2
heat shock	5	-	3/2	2/3
protease	4	-	3/1	3/1
transport	4	-	3/1	2/2
metabolism	5	-	5/0	4/1
homeobox	4	0.001	3/1	1/3
cell wall	3	-	2/1	2/1
transcription	3	-	2/1	0/3
non-translating	2	-	0/2	1/1
peroxisome	2	-	0/2	0/2
other*	20	-	14/6	11/9
No annotation	8	-	4/4	5/3

Adjusted P-values displayed are based on an enrichment test of the functional categories relative to all expressed transcripts followed by P-value adjustment using the Benjamini Hochberg procedure. - indicates that an enrichment test was not performed as categories were based on bespoke annotations. * includes transcripts with annotations that could not be grouped by function with other transcripts. ** only the seven transcripts that were annotated as ubiquitin-related in the TGAC annotation were used in the enrichment test (see methods).

The category with the most DE transcripts was non-coding RNA (ncRNA, 15 transcripts), although this was not significantly enriched relative to all expressed transcripts. All ncRNA transcripts were classed as long non-coding RNAs (>200bp; Guttman & Rinn, 2012). Four of the ncRNAs overlapped with coding transcripts (two in the antisense direction) and one ncRNA was a putative miRNA precursor (Ta-miR132-3p, 5'-3' mature sequence: TATAAACTTGGTCAAAGTTTG; Sun *et al.*, 2014). Thirteen transcripts (belonging to nine genes) were identified as putative targets of Ta-miR132-3p in the TGAC reference but none of these target transcripts were differentially expressed in this dataset (Table 4.10). The second largest transcript category was transposon-associated (14 transcripts; adjusted P = 0.008), whereas the third largest category was DE transcripts related to ubiquitin and the proteasome (12 transcripts; P = 0.008). DE transcripts annotated as homeobox were also enriched (4 transcripts; adjusted P = 0.001). Interestingly, homeodomain TFBS were identified in the promoters of 27 of the 38 outside 5A DE transcripts although this was not significantly enriched (adjusted P = 0.166).

Table 4.10: Putative targets of Ta-miR132-3p

Transcript	TGAC Annotation
TRIAE_CS42_2AL_TGACv1_093400_AA0279320.1	Protein phosphatase 2C containing protein
TRIAE_CS42_2BL_TGACv1_134090_AA0443680.1 TRIAE_CS42_2BL_TGACv1_134090_AA0443680.2	Germin-like protein 4-1, Uncharacterized protein
TRIAE_CS42_2BS_TGACv1_146052_AA0454150.1	Glycosyltransferase
TRIAE_CS42_6BS_TGACv1_516121_AA1673900.3	Mitochondrial import inner membrane translocase subunit tim22, Uncharacterized protein
TRIAE_CS42_7AS_TGACv1_569800_AA1824270.1	ABC transporter C family member 10, Uncharacterized protein
TRIAE_CS42_7BL_TGACv1_577198_AA1868480.1	Isoleucyl-tRNA synthetase, cytoplasmic, Uncharacterized protein
TRIAE_CS42_7DL_TGACv1_603074_AA1975250.1	Isoleucyl-tRNA synthetase, cytoplasmic, Uncharacterized protein
TRIAE_CS42_7DS_TGACv1_622139_AA2033500.1 TRIAE_CS42_7DS_TGACv1_622139_AA2033500.2 TRIAE_CS42_7DS_TGACv1_622139_AA2033500.3 TRIAE_CS42_7DS_TGACv1_622139_AA2033500.4	Shikimate kinase, Uncharacterized protein
TRIAE_CS42_U_TGACv1_641065_AA2084010.1	Glycosyltransferase

Table 4.11: Functional annotation of differentially expressed transcripts in the T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ comparisons

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_5AS_TGACv1_393645_AA1274860.1	1A	428,214,494	T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_1BL_TGACv1_032057_AA0124830.1	1B	261,775,404	T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_1BL_TGACv1_030315_AA0086470.1	1B	290,813,728	T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_2AS_TGACv1_112274_AA0334670.1	2A	313,287,504	T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_3DL_TGACv1_250633_AA0871340.1	3D	279,882,343	T1	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_3DL_TGACv1_250330_AA0866270.1	3D	608,417,264	T1 + T2	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AS_TGACv1_393897_AA1277010.1	5A	85,149,732	T1 + T2	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AS_TGACv1_393783_AA1275990.1	5A	104,011,459	T1 + T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AS_TGACv1_394352_AA1279770.1	5A	139,518,885	T1 + T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AS_TGACv1_393726_AA1275550.1	5A	162,029,855	T1 + T2	5A+	non-coding RNA	ncRNA; repeat associated	TGAC
TRIAE_CS42_5AL_TGACv1_375857_AA1228100.1*	5A	334,343,515	T1 + T2	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AL_TGACv1_376953_AA1242680.1	5A	427,317,205	T1 + T2	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AL_TGACv1_374498_AA1201570.1	5A	434,793,971	T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AL_TGACv1_376877_AA1241920.1	5A	447,314,890	T1	5A-	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_5AL_TGACv1_374413_AA1199590.1	5A	475,323,308	T1 + T2	5A+	non-coding RNA	ncRNA	TGAC
TRIAE_CS42_1AS_TGACv1_019083_AA0060340.1	1A	27,390,992	T1 + T2	5A-	transposon-associated	Repeat associated	TGAC
TRIAE_CS42_1BS_TGACv1_049354_AA0149990.1	1B	185,585,375	T2	5A+	transposon-associated	Retrotransposon-like	Manual
TRIAE_CS42_2AL_TGACv1_095650_AA0313320.1	2A	398,328,221	T1 + T2	5A-	transposon-associated	AT-hook motif-containing protein, putative, Putative helicase; retrotransposon-like	TGAC; Manual
TRIAE_CS42_2DS_TGACv1_177196_AA0568430.1	2D	174,418,517	T1 + T2	5A+	transposon-associated	Repeat associated	TGAC
TRIAE_CS42_3AS_TGACv1_211411_AA0689940.1	3A	547,984,779	T1	5A+	transposon-associated	Transposon protein, putative, mutator sub-class	TGAC

Table 4.11 continued on next page

Table 4.11 continued from previous page

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_4BL_TGACv1_321219_AA1057420.1	4B	527,725,294	T1 + T2	5A-	transposon-associated	Replication factor-A carboxy-terminal domain protein	TGAC
TRIAE_CS42_5AS_TGACv1_393577_AA1274040.1	5A	71,474,627	T1 + T2	5A+	transposon-associated	Retrotransposon protein; Ty3-gypsy subclass	TGAC
TRIAE_CS42_5AS_TGACv1_394531_AA1280840.1	5A	199,145,177	T1	5A-	transposon-associated	zinc ion binding, nucleic acid binding; putative retrotransposon	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_374025_AA1188070.1*	5A	328,818,968	T1	5A+	transposon-associated	Retrotransposon, putative, Ty1-copia subclass	TGAC
TRIAE_CS42_5AL_TGACv1_374249_AA1195190.1	5A	436,603,964	T1 + T2	5A+	transposon-associated	Transposon protein, CACTA, En/Spm sub-class	TGAC
TRIAE_CS42_5DL_TGACv1_433728_AA1420750.1	5D	443,143,427	T1	5A+	transposon-associated	Transposon-related, RICESLEEPER 2-like	Manual
TRIAE_CS42_5DL_TGACv1_435337_AA1449220.1	5D	458,092,027	T2	5A-	transposon-associated	Transposon-related	Manual
TRIAE_CS42_6AL_TGACv1_471516_AA1510240.1	6A	555,225,210	T1 + T2	5A+	transposon-associated	Retrotransposon protein-like	Manual
TRIAE_CS42_7DS_TGACv1_621701_AA2023630.1	7D	28,807,835	T1	5A+	transposon-associated	Transposon protein, Mutator sub-class	TGAC
TRIAE_CS42_1BS_TGACv1_049354_AA0149980.1	1B	185,548,016	T1 + T2	5A-	ubiquitin	UBCc domain; E2 ubiquitin conjugating enzyme	Manual
TRIAE_CS42_5AL_TGACv1_375493_AA1222690.2	5A	265,539,274	T1 + T2	5A-	ubiquitin	BTB/POZ domain-containing protein	TGAC
TRIAE_CS42_5AL_TGACv1_374542_AA1202810.4*	5A	333,439,847	T1 + T2	5A-	ubiquitin	RING/U-box superfamily protein; putative E3 ligase	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_375361_AA1220430.2	5A	413,416,970	T1 + T2	5A-	ubiquitin	RAD23, ubiquitin receptor, proteasome associated	Manual
TRIAE_CS42_5AL_TGACv1_374374_AA1198360.1	5A	421,856,090	T1 + T2	5A-	ubiquitin	F-box protein-like	Manual
TRIAE_CS42_5AL_TGACv1_374097_AA1190230.1	5A	439,551,515	T1 + T2	5A+	ubiquitin	Ubiquitin, Polyubiquitin 14; NEDD8-like protein RUB1	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_374675_AA1206250.1	5A	439,853,084	T1	5A-	ubiquitin	RING/U-box superfamily protein, Zinc finger protein-like protein	TGAC
TRIAE_CS42_5AL_TGACv1_378415_AA1253190.1	5A	470,075,745	T2	5A+	ubiquitin	F-box protein	TGAC

Table 4.11 continued on next page

Table 4.11 continued from previous page

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_5AL_TGACv1_377520_AA1247660.1	5A	475,464,797	T1 + T2	5A+	ubiquitin	eIF3 n terminal; PCI/PINT associated module	Manual
TRIAE_CS42_5DL_TGACv1_434064_AA1428250.1	5D	207,168,354	T2	5A+	ubiquitin	E3 ubiquitin protein ligase SDIR1	TGAC
TRIAE_CS42_5AS_TGACv1_402215_AA1284070.1	NA	NA	T1 + T2	5A-	ubiquitin	Ubiquitin, Polyubiquitin 4	TGAC
TRIAE_CS42_5AS_TGACv1_393696_AA1275280.2	NA	NA	T1 + T2	5A-	ubiquitin	Polyubiquitin, Ubiquitin	TGAC
TRIAE_CS42_5AS_TGACv1_393572_AA1273920.1	5A	73,805,941	T1 + T2	5A-	cell cycle	Kinesin-like protein	TGAC
TRIAE_CS42_5AL_TGACv1_373986_AA1186560.2*	5A	336,456,148	T1 + T2	5A+	cell cycle	Kinesin-like protein	TGAC
TRIAE_CS42_5AL_TGACv1_373986_AA1186560.3*	5A	336,456,148	T1 + T2	5A+	cell cycle	Kinesin-like protein	TGAC
TRIAE_CS42_5AL_TGACv1_374322_AA1196910.1	5A	408,190,618	T2	5A-	cell cycle	IMP dehydrogenase/GMP reductase; HAUS augmin-like complex subunit 5	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_376411_AA1236560.1	5A	473,423,951	T1 + T2	5A+	cell cycle	SHAGGY-like kinase	Manual
TRIAE_CS42_3AL_TGACv1_195567_AA0651320.1	3A	746,292,914	T2	5A+	histone-related	Histone H2A	Manual
TRIAE_CS42_5AS_TGACv1_394776_AA1281770.1	5A	248,528,260	T1 + T2	5A-	histone-related	Histone deacetylase 14 isoform	Manual
TRIAE_CS42_5AS_TGACv1_394776_AA1281770.3	5A	248,528,260	T1 + T2	5A-	histone-related	Histone deacetylase 14 isoform	Manual
TRIAE_CS42_5AL_TGACv1_374195_AA1193180.2	5A	477,295,158	T2	5A+	histone-related	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein; Lysine-specific demethylase JMJ30	TGAC; Manual
TRIAE_CS42_U_TGACv1_643349_AA2131010.1	Un	103,526,863	T2	5A-	histone-related	Histone superfamily protein; Histone H4 superfamily	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_376851_AA1241550.4	5A	414,772,190	T2	5A+	heatshock	TPR repeat-containing thioredoxin TDX; heatshock related	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_375394_AA1220930.4	5A	474,541,233	T1 + T2	5A+	heatshock	HSP90 superfamily	Manual
TRIAE_CS42_5AL_TGACv1_375394_AA1220930.5	5A	474,541,233	T2	5A+	heatshock	HSP90 superfamily	Manual
TRIAE_CS42_4DL_TGACv1_343485_AA1135140.1	Un	100,323,297	T2	5A-	heatshock	HSP70, DnaK	Manual
TRIAE_CS42_7BL_TGACv1_578302_AA1892720.1	NA	NA	T2	5A-	heatshock	Retrotransposon putative; HEAT-STRESS-ASSOCIATED 32	TGAC; Manual
TRIAE_CS42_2DS_TGACv1_177488_AA0578600.1	2D	13,913,395	T1	5A-	protease	Protease domain; ankryin repeats	Manual

Table 4.11 continued on next page

Table 4.11 continued from previous page

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_5AL_TGACv1_374321_AA1196890.1	5A	422,062,336	T1 + T2	5A-	protease	Aspartyl aminopeptidase; Zinc peptidase-like superfamily	TGAC; Manual
TRIAE_CS42_5AL_TGACv1_375845_AA1227990.1	5A	444,535,471	T1 + T2	5A-	protease	Abi superfamily, CAAX protease	Manual
TRIAE_CS42_5AL_TGACv1_374078_AA1189690.1	5A	476,435,013	T1 + T2	5A+	protease	Serine carboxypeptidase-like protein 9	TGAC
TRIAE_CS42_5AS_TGACv1_393493_AA1273190.4	5A	77,084,934	T1 + T2	5A+	transport	Calcium transporting ATPase	TGAC
TRIAE_CS42_5AL_TGACv1_375949_AA1229270.2*	5A	387,399,431	T2	5A-	transport	TauE superfamily	Manual
TRIAE_CS42_5AL_TGACv1_374155_AA1191930.1	5A	448,605,465	T2	5A-	transport	Potassium transporter	TGAC
TRIAE_CS42_7AL_TGACv1_556075_AA1754700.1	7A	567,001,946	T1	5A+	transport	ABC transporter G family	Manual
TRIAE_CS42_5AS_TGACv1_395074_AA1282530.1	5A	143,730,436	T1 + T2	5A+	metabolism	Glyoxylate reductase	TGAC
TRIAE_CS42_5AL_TGACv1_376402_AA1236390.1	5A	404,531,058	T1 + T2	5A-	metabolism	Acetate--CoA ligase	Manual
TRIAE_CS42_5AL_TGACv1_374560_AA1203240.1	5A	417,045,347	T1	5A-	metabolism	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	TGAC
TRIAE_CS42_5AL_TGACv1_374231_AA1194360.2	5A	438,141,289	T1 + T2	5A-	metabolism	Monogalactosyldiacylglycerol synthase	Manual
TRIAE_CS42_5AS_TGACv1_393119_AA1268700.1	5D	180,466,588	T1 + T2	5A-	metabolism	quinolinate synthase	TGAC
TRIAE_CS42_5AS_TGACv1_392838_AA1265240.1	5A	181,399,775	T1 + T2	5A+	homeobox	Homeobox protein knotted-1-like 2	TGAC
TRIAE_CS42_5AS_TGACv1_392838_AA1265240.2	5A	181,399,775	T1 + T2	5A+	homeobox	Homeobox protein knotted-1-like 2	TGAC
TRIAE_CS42_5AL_TGACv1_374741_AA1207970.1	5A	463,451,614	T2	5A-	homeobox	homeobox-leucine zipper protein	TGAC
TRIAE_CS42_7BL_TGACv1_577371_AA1873630.1	7B	692,600,853	T1 + T2	5A+	homeobox	Homeobox protein knotted-1-like 2	TGAC
TRIAE_CS42_5AL_TGACv1_375085_AA1215790.1	5A	475,524,596	T1 + T2	5A+	cell wall	Fascilin-like arabinogalactan protein	Manual
TRIAE_CS42_5AL_TGACv1_374319_AA1196780.1	5A	476,667,345	T1 + T2	5A-	cell wall	Pectin lyase-like superfamily protein; Polygalacturonase-like	TGAC; Manual
TRIAE_CS42_7DS_TGACv1_622195_AA2034920.1	7D	106,414,593	T1	5A-	cell wall	Callose synthase	Manual
TRIAE_CS42_5AL_TGACv1_378188_AA1251790.1	2B	667,276,330	T1 + T2	5A+	transcription	Far1-related sequence 5-like protein	TGAC
TRIAE_CS42_4DS_TGACv1_361025_AA1159110.4	4D	79,211,678	T2	5A+	transcription	LIM-domain binding protein, SEUSS orthologue	Manual

Table 4.11 continued on next page

Table 4.11 continued from previous page

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_5AS_TGACv1_392558_AA1260860.1	5A	185,647,816	T1 + T2	5A+	transcription	SSXT protein; GRF1-interacting-factor	TGAC; Manual
TRIAE_CS42_2DL_TGACv1_157942_AA0502830.1	2D	614,383,115	T1	5A+	non-translating	Non-translating	TGAC
TRIAE_CS42_7AS_TGACv1_571015_AA1843630.1	7A	32,210,288	T1 + T2	5A-	non-translating	Non-translating	TGAC
TRIAE_CS42_5AL_TGACv1_374446_AA1200410.1	5A	473,092,584	T1 + T2	5A+	peroxisome	Putative peroxisomal targeting signal 1 receptor	TGAC
TRIAE_CS42_5AL_TGACv1_374446_AA1200420.1	5A	473,195,032	T1	5A+	peroxisome	Putative peroxisomal targeting signal 1 receptor	TGAC
TRIAE_CS42_2AL_TGACv1_095650_AA0313330.1	2A	398,347,440	T1	5A-	other	homologue of yeast autophagy 18 (ATG18) G	TGAC
TRIAE_CS42_2AL_TGACv1_095956_AA0316040.1	2A	398,971,137	T1	5A-	other	DEA(D/H)-box RNA helicase family protein	TGAC
TRIAE_CS42_2AL_TGACv1_095956_AA0316050.1	2A	398,979,949	T1 + T2	5A-	other	DNA binding protein-like	TGAC
TRIAE_CS42_2BS_TGACv1_148693_AA0494610.1	2B	58,987,101	T1	5A-	other	transducin family protein, WD-40 repeat family	Manual
TRIAE_CS42_3DL_TGACv1_254173_AA0896070.1	3D	276,574,630	T2	5A+	other	IQM1; Calmodulin binding	Manual
TRIAE_CS42_5AS_TGACv1_393235_AA1270150.1	5A	138,079,992	T1 + T2	5A+	other	Hydroxyproline-rich glycoprotein family protein	TGAC
TRIAE_CS42_5AS_TGACv1_393580_AA1274130.1	5A	155,209,029	T1 + T2	5A+	other	Endoplasmic reticulum, stress-associated Ramp4	TGAC
TRIAE_CS42_5AL_TGACv1_374163_AA1192120.1	5A	284,690,697	T2	5A-	other	Rho GTPase-activating protein gacA	TGAC
TRIAE_CS42_5AL_TGACv1_374727_AA1207530.1*	5A	388,637,659	T1 + T2	5A-	other	DUF810 family protein	TGAC
TRIAE_CS42_5AL_TGACv1_376076_AA1231790.1	5A	403,283,773	T1	5A+	other	tyrosine--tRNA ligase 1	Manual
TRIAE_CS42_5AL_TGACv1_376953_AA1242690.1	5A	427,319,739	T1 + T2	5A-	other	Hypersensitive induced response protein 3	TGAC
TRIAE_CS42_5AL_TGACv1_376953_AA1242690.2	5A	427,320,059	T1 + T2	5A-	other	Hypersensitive induced response protein 3	TGAC
TRIAE_CS42_5AL_TGACv1_374097_AA1190260.1	5A	439,569,122	T1 + T2	5A+	other	DNA-3-methyladenine glycosylase	TGAC
TRIAE_CS42_5AL_TGACv1_375721_AA1226170.1	5A	439,767,852	T1 + T2	5A+	other	Bet1-like SNARE 1-1	TGAC
TRIAE_CS42_5AL_TGACv1_378334_AA1252720.1	5A	444,849,607	T1 + T2	5A-	other	Vacuolar processing enzyme 4	TGAC
TRIAE_CS42_5AL_TGACv1_376411_AA1236550.1	5A	473,427,021	T1 + T2	5A+	other	ribonucleoside--diphosphate reductase large subunit partial match	Manual
TRIAE_CS42_5AL_TGACv1_374080_AA1189800.1	5A	473,625,152	T1 + T2	5A+	other	Vacuolar protein sorting-associated protein; Sec3 superfamily	TGAC; Manual

Table 4.11 continued on next page

Table 4.11 continued from previous page

Transcript ID	Chr	Position	Time point DE	NIL with higher expression	Category	Annotation	Source
TRIAE_CS42_5AL_TGACv1_374080_AA1189810.1	5A	473,637,539	T1 + T2	5A+	other	Glycosyltransferase protein 2-like	TGAC
TRIAE_CS42_5AL_TGACv1_377986_AA1250630.1	5A	524,318,116	T1 + T2	5A-	other	Generative cell specific-1; Hapless 2	TGAC; Manual
TRIAE_CS42_6BL_TGACv1_503194_AA1627460.1	6B	623,718,695	T1 + T2	5A-	other	Allene oxide cyclase 4	TGAC
TRIAE_CS42_4BS_TGACv1_327817_AA1075010.1	4B	95,073,178	T2	5A-	No annotation	NA	NA
TRIAE_CS42_5AS_TGACv1_395084_AA1282570.1	5A	70,454,003	T1 + T2	5A-	No annotation	NA	NA
TRIAE_CS42_5AS_TGACv1_393473_AA1272910.1	5A	237,953,207	T1 + T2	5A-	No annotation	NA	NA
TRIAE_CS42_5AL_TGACv1_374657_AA1205780.1*	5A	328,545,606	T1	5A-	No annotation	NA	NA
TRIAE_CS42_5AL_TGACv1_375845_AA1227980.1	5A	444,539,760	T1	5A-	No annotation	NA	NA
TRIAE_CS42_5AL_TGACv1_376877_AA1241930.1	5A	447,310,958	T1 + T2	5A-	No annotation	NA	NA
TRIAE_CS42_5BS_TGACv1_427448_AA1393420.1	5B	52,914,927	T1 + T2	5A+	No annotation	NA	NA
TRIAE_CS42_U_TGACv1_641674_AA2101090.1	6B	719,335,147	T1	5A+	No annotation	NA	NA
TRIAE_CS42_7BS_TGACv1_592547_AA1940160.1	7B	198,610,669	T1 + T2	5A+	No annotation	NA	NA

* indicates that the transcript is located in the fine-mapped interval for grain length. Chromosome (Chr) and position of the transcripts are based on an *in silico* mapping of TGACv1 gene models to IWGSC RefSeq v1.0 (NA indicates that no position could be assigned using this method). NA in the annotation and source columns indicates that no annotation could be obtained.

The DE transcripts related to ubiquitin were of particular interest as ubiquitin-mediated protein turnover has previously been associated with the control of seed/grain size in wheat (Simmonds *et al.*, 2016) and other species including rice and Arabidopsis (Disch *et al.*, 2006; Song *et al.*, 2007; Xia *et al.*, 2013). The pathway acts through the sequential action of a cascade of enzymes (see Figure 4.6a legend) to add multiple copies of the protein ubiquitin (Ub) to a substrate protein that is then targeted for degradation by the proteasome. DE transcripts were identified at almost all steps of this pathway (excluding E1): two ubiquitin proteins and one ubiquitin-like protein, one E2 conjugase, six potential E3 ligase components and two putative components of the proteasome (Figure 4.6). In addition to these, we also identified four DE transcripts annotated as proteases (Figure 4.6), which are known substrates regulated by this pathway (Du *et al.*, 2014; Dong *et al.*, 2017; Huang *et al.*, 2017) and that influence organ size through the regulation of cell proliferation. Most of the components of the ubiquitin pathway that were differentially expressed were more highly expressed in the 5A- NIL (11/16, including proteases) (Figure 4.6b).

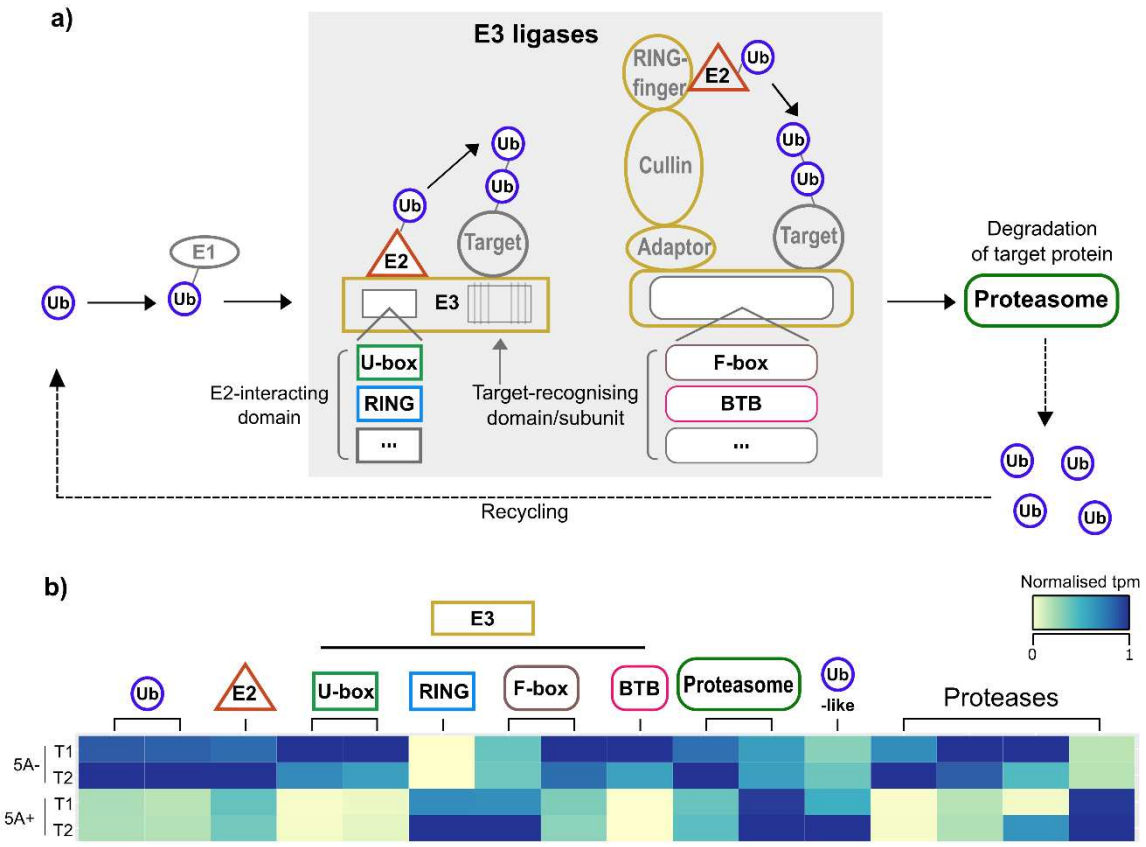


Figure 4.6: Differential regulation of the ubiquitin pathway in 5A NILs

a) Differentially expressed (DE) transcripts with functional annotations related to ubiquitin-mediated protein turnover were enriched relative to the whole genome (a). This pathway acts to add multiple copies of the protein Ubiquitin (Ub) to a substrate protein through the sequential action of a cascade of three enzymes: E1 (Ub activating enzymes), E2 (Ub-conjugating enzymes) and E3 (Ub ligases). The tagged substrate is then targeted for degradation by the 26S proteasome and the Ub proteins are recycled. The E3 ligases are the most diverse of the three enzymes and both single subunit proteins and multi-subunit complexes exist. A subset of these classes is shown in the grey box in (a), selected based on the annotations of DE transcripts. Single subunit E3 ligases have an E2-interacting domain (e.g. U-box, RING, etc. (…)) and a substrate-recognising domain. Multi-subunit complexes also have E2-interacting complexes and substrate-recognising subunits (e.g. F-box, BTB, etc. (…)). In the context of organ size control, some proteases have been identified as downstream targets of this pathway (e.g. DA1, UBP15 (Du *et al.*, 2014; Dong *et al.*, 2017)). b) Heatmap of normalised tpm of DE transcripts associated with ubiquitin, the proteasome and proteases.

4.5 Discussion

In this chapter, RNA-Seq was performed on the developing grains of 5A NILs. In Chapter 2 we established that these NILs have a difference in pericarp cell size, grain length and final grain weight and that the first phenotypic differences between NILs arose during early grain development (Brinton *et al.*, 2017). The aim of this chapter was to identify genes that are differentially expressed between 5A NILs at these early stages of grain development in order to identify specific genes and pathways that affect pericarp cell size and grain size at the transcriptional level.

4.5.1 The importance of a high-quality reference sequence

The RNA-Seq data was initially mapped to two different reference transcriptomes: CSS and TGAC. The TGAC outperformed the CSS transcriptome both in terms of the number of reads that aligned and in the gene models themselves. This was most likely due to the significant improvement in terms of sequence contiguity of the TGAC reference over the CSS (N50= 88.8 vs < 10 kb, respectively), allowing more accurate prediction of gene models. The results of this chapter highlight the practical importance of this improvement as 64 more DE transcripts were detected using the TGAC reference, in most cases, due to the absence of a corresponding gene model in the CSS reference (46 transcripts). There were also cases where incorrect gene models in the CSS reference led to misleading results. For example, in the CSS fused gene model case study (Figure 4.2d) a single DE transcript from the CSS reference had a large accumulation of reads mapping to the 3' UTR. This gene was the orthologue of Arabidopsis *NPY1*, which plays a role in auxin-regulated organogenesis (Cheng *et al.*, 2007) and could therefore be related to the control of grain size. However, in the TGAC reference, in addition to the *NPY1* orthologue, an alternative gene model was annotated in place of the 3' UTR. This alternative gene model was differentially expressed whilst the *NPY1* orthologue was expressed at a very low level and was not differentially expressed.

As shown in Chapter 3, the improvements in scaffold size, contiguity and gene annotation open up new opportunities in wheat research. Here the new physical sequence was used to assign locations to 107 of 112 DE transcripts identified between NILs, allowing us to determine which DE transcripts were located within the QTL fine-mapped interval(s) defined in Chapter 3. Likewise, the analysis of promoter sequences enabled new hypothesis generation for this specific biological process and will also aid in the understanding of how promoter differences across genomes affect the relative transcript abundance of the different homoeologues. It will also be interesting to explore the differences in the promoter sequences between the sequenced 'parental' varieties (Claire and Cadenza, as defined in Chapter 3) in light of the results of the current chapter. These results exemplify the importance of correctly annotated gene models and improved genome assemblies in gaining a more accurate view of the underlying biology.

4.5.2 Differential expression analysis provides an insight into the biological processes occurring during early grain development

Grains were sampled at 4 and 8 dpa to encompass the developmental stage at which the first significant difference in grain length between 5A NILs is observed. During this stage, increases in grain size are largely driven by cell expansion in the pericarp (Drea *et al.*, 2005; Radchuk *et al.*, 2011), consistent with the finding that increased pericarp cell size underlies the difference in final grain length (discussed previously in Chapter 2). These time points are also relatively early compared to other grain related RNA-Seq studies which have focussed on later grain filling processes (Pellny *et al.*, 2012; Pfeifer *et al.*, 2014b; Yu *et al.*, 2016). The 'across time' comparisons ($5A - {}^{T1}_{T2}$ and $5A + {}^{T1}_{T2}$) identified > 2,700 DE transcripts in each NIL, and there was a large overlap in the biological processes being differentially regulated. Most of the DE transcripts were upregulated over time and many of these were associated with metabolism and biosynthesis consistent with grains undergoing a period of rapid growth and the start of endosperm cellularisation at this stage of development (Shewry *et al.*, 2012). Transcripts associated with proteolysis and mRNA catabolism were also upregulated across time consistent with increases in specific proteases and other hydrolytic enzymes at this stage of grain development (Dominguez & Cejudo, 1996). These could be indicative of programmed cell death which occurs in both the nucellus and pericarp of the developing grain up to 12 dpa (Radchuk *et al.*, 2011; Radchuk *et al.*, 2017). An upregulation of transcripts associated with defence response and oxidation-reduction process was also identified, consistent with previous reports of accumulation of proteins associated with defence against both pathogens and oxidative stress during the early-mid stages of grain development (Kaspar-Schoenefeld *et al.*, 2016). Transcriptional studies always have the caveat that changes in gene expression may not translate to changes in protein level (Pires & Conant, 2016). However, proteomic analyses of similar stages of grain development have identified the differential regulation of similar ontologies (Kaspar-Schoenefeld *et al.*, 2016; Yang *et al.*, 2017) suggesting that these transcriptional changes are reflective of overall protein status in the grain.

4.5.3 Comparative transcriptomics as a method to identify candidate genes underlying the 5A grain length QTL

The use of highly isogenic material allowed the direct comparison of the effect of the 5A introgression on gene expression at each time point (T1 ${}^{5A-}_{5A+}$ and T2 ${}^{5A-}_{5A+}$). This resulted in a defined set of 112 DE transcripts between genotypes. The majority of T1 ${}^{5A-}_{5A+}$ and T2 ${}^{5A-}_{5A+}$ DE transcripts were located on chromosome 5A and all of these were located within the 5A introgression. This is expected given that the sequence variation in the NILs was restricted to the chromosome 5A region.

DE transcripts located within the fine-mapped interval(s) on chromosome 5A represent good candidates for further characterisation. The kinesin-like gene and RING/U-box superfamily protein

are particularly strong candidates based on their functional annotations. Previous studies have demonstrated that kinesin-like proteins can regulate grain length and cell expansion through involvement with microtubule dynamics (Kitagawa *et al.*, 2010; Li *et al.*, 2011; Fujikura *et al.*, 2014). The RING/U-box protein is a putative E3 ligase, a class of enzymes which have been associated with the control of grain size e.g. *TaGW2_A* discussed in the previous chapters and discussed in more detail later (Song *et al.*, 2007; Simmonds *et al.*, 2016).

It is premature, however, to speculate on the identity of a 5A causal gene(s) at this stage. It is difficult to predict whether DE transcripts in the fine-mapped interval are truly associated with the effect of the 5A QTL or are simply a consequence of sequence variations between the parental cultivars, i.e. 'guilt by association'. A relevant example was the recent use of transcriptomics to define a candidate gene underlying a grain dormancy QTL (*PM19*) (Barrero *et al.*, 2015). Subsequent studies showed that a different gene in close physical proximity (*TaMKK3*) (Torada *et al.*, 2016) was responsible for the natural variation observed (Shorinola *et al.*, 2017). The mis-interpretation of the transcriptomics data was due to complete linkage between the DE *PM19* gene and the causal *TaMKK3* gene in the germplasm used in the original study. Additionally, the causal gene(s) underlying the 5A QTL may not be differentially expressed between the 5A NILs and could be a result of allelic variation that alters the function of the gene independent of expression level. The SNP calling performed using the RNA-Seq data described in the previous chapter did not identify coding region polymorphisms between any genes predicted to be located within the interval(s). Analysis of the genomic sequences of the two 'parental' varieties will provide further insights but ultimately further fine-mapping of the 5A loci will be required to identify the underlying gene(s).

4.5.4 DE transcripts outside chromosome 5A are candidates for downstream targets of the 5A QTL

DE transcripts outside of chromosome 5A were considered as candidates for downstream targets of genes located in the 5A introgression because the differential expression are unlikely to have arisen through sequence variation. These included genes located in the A, B and D genomes implying that there is cross-talk at the transcriptional level between the three genomes. In the promoters of these genes, there was enrichment of TF binding sites associated with TF families that have all previously been shown to play diverse roles in the control of organ development (Barg *et al.*, 2005; Kaplan-Levy *et al.*, 2012). For example YABBY genes, a plant specific family of TFs, play a critical role in patterning and the establishment of organ polarity (Sarojam *et al.*, 2010) and fruit size (Cong *et al.*, 2008). Another example are the C2H2 TFs, *NUBBIN* and *JAGGED*, which are involved in determining carpel shape in Arabidopsis (Dinneny *et al.*, 2006). AT-Hook TFs play roles in floral organ development in both maize and rice (Gallavotti *et al.*, 2011; Jin *et al.*, 2011) and modulate cell elongation in the Arabidopsis hypocotyl (Street *et al.*, 2008). Few of these TF

families have been characterised in wheat, and although these interactions need to be experimentally validated, they could be potential targets for the manipulation of grain size.

4.5.5 DE transcripts have functions related to the control of seed/organ size

Studies in species such as rice and Arabidopsis have shown that seed size is regulated by a complex network of genes and diverse mechanisms, ultimately through the coordination of cell proliferation and expansion (reviewed in Huang *et al.*, 2013; Li & Li, 2015). 5A+ NILs have significantly longer pericarp cells, suggesting that the underlying gene influences cell expansion (Chapter 2; Brinton *et al.*, 2017). Genes that physically modify the cell wall have been shown to directly control cell expansion (reviewed in Cosgrove, 2005) and three of the DE transcripts between 5A NILs have potential roles in cell wall synthesis and remodelling. There were also a number of DE transcripts associated with the cell cycle and the control of cell proliferation. During seed development, a number of cell cycle types in addition to the typical mitotic cycle are observed. One such alternative cycle type is endoreduplication, characterised by the replication of chromosomes in the absence of cell division, which is associated with cell enlargement (reviewed in Dante *et al.*, 2014). Two of the DE transcripts were the closest wheat orthologues of Arabidopsis genes that have specific roles in organ development: a *GRF*-interacting factor (*GIF*) and *SEUSS* (*SEU*). In Arabidopsis, the GIF genes interact with the *GROWTH-REGULATING FACTOR* (*GRF*) TFs and act as transcriptional co-activators to regulate organ size through cell proliferation (Lee *et al.*, 2009). Conversely, *SEU* acts a transcriptional co-repressor and interacts with important regulators of development to control many processes, including floral organ development (Bao *et al.*, 2010).

Seed development requires the coordination of processes across multiple tissues, namely the seed coat, endosperm and embryo. The development and growth of these tissues is inherently interlinked, and it has been proposed that the mechanical constraint imposed by the maternal seed coat/pericarp places an upper limit on the size of the seed/grain (Adamski *et al.*, 2009; Hasan *et al.*, 2011; Brinton *et al.*, 2017). Epigenetic regulation appears to play an important role in the cross-talk and coordination of these tissues (Locascio *et al.*, 2014). The differential expression of 34 non-coding transcripts, transposons and histone-related transcripts between NILs could suggest a difference in epigenetic status associated with the control of pericarp cell size. Additional work to characterise these non-coding RNAs would be warranted to establish their role in grain development.

The ubiquitin-mediated control of seed/grain size has been documented in a number of species (reviewed in Li & Li, 2014), including wheat (Yang *et al.*, 2012; Simmonds *et al.*, 2016). DE transcripts associated with the ubiquitin pathway were significantly enriched in the 5A NILs. The pathway tags substrate proteins with multiple copies of the ubiquitin protein through the sequential action of a cascade of enzymes: E1 (Ub activating), E2 (Ub conjugases) and E3 (Ub ligases). The ubiquitinated substrate proteins are then targeted to the 26S proteasome for degradation (Hershko & Ciechanover, 1998). *TaGW2_A*, described in previous chapters as a potential candidate

underlying the 6A grain width QTL, is a RING-type E3 ligase. As previously discussed, the rice and Arabidopsis orthologues of this gene (*GW2* and *DA2*, respectively) act to influence grain/seed size through the modulation of cell proliferation. Another E3 ligase, *EOD1/BB* also negatively regulates seed size in Arabidopsis (Disch *et al.*, 2006). In general, the E3 ligase determines the specificity for the substrate proteins (Hershko & Ciechanover, 1998) and *DA2* and *EOD1* may have different substrate targets, however they converge and both target the ubiquitin-activated protease *DA1*. *DA1* also negatively regulates cell proliferation and acts synergistically with both *DA2* and *EOD1*, although it is not clear whether the two E3 ligases act via independent genetic pathways or as part of the same mechanism *(Xia et al., 2013; Vanhaeren et al., 2016; Dong et al., 2017).* *UBP15* (a ubiquitin specific protease) is a downstream target of this pathway and conversely acts as a positive regulator of seed size through the promotion of cell proliferation (Du *et al.*, 2014). Other ubiquitin-associated regulators of organ/grain size have been identified, including components of the 26S proteasome, enzymes with deubiquitinating activity and proteins that have been shown to bind ubiquitin *in vitro* (Weng *et al.*, 2008; Kurepa *et al.*, 2009; Huang *et al.*, 2017). The DE transcripts associated with this pathway are not direct orthologues of these previously characterised genes. As such the functional characterisation of these putative novel components could provide new insights into the ubiquitin-mediated control of grain size in cereals. A subset of these genes have been selected for further characterisation using TILLING mutants, discussed further in Chapter 5.

5 General discussion

The overall aim of this thesis was to understand the mechanisms that control grain length and width in hexaploid wheat through the characterisation of two distinct grain weight QTL located on chromosomes 5A and 6A. Specifically, this PhD combined phenotypic characterisation, genetic mapping and transcriptomics to answer the following questions:

- Do the 5A and 6A QTL increase grain weight via the same or different mechanisms?
- What are the genes/pathways underlying the 5A and 6A QTL?
- Is *TaGW2_A* the gene underlying the 6A QTL?

5.1 Mechanisms and genes underlying the 6A and 5A QTL

As determined in Chapter 2, the 6A and 5A QTL act to increase grain weight through different mechanisms, consistent with previous reports that these grain size parameters are under independent genetic control (Gegas *et al.*, 2010). The 5A QTL acts primarily to increase grain length during early grain development, but post-fertilisation, through increased pericarp cell size. The 5A QTL also has a pleiotropic effect on grain width during late grain development, which is smaller than the effect on length. This late-season width effect is potentially more sensitive to environmental variation and determines the magnitude of the final grain weight increase. On the other hand, the 6A QTL acts during very early grain development, perhaps before fertilisation (i.e. during carpel development), and specifically increases grain width, with no differences observed in final grain length. Although no cell size/number data was obtained for the 6A QTL, we hypothesised that this is likely to be an effect on cell number due to the timing of initial grain size differences, although this has not yet been tested experimentally.

5.1.1 Genes and pathways underlying the 6A QTL

Speculating on the identity of candidates for the causal genes underlying the 6A QTL remains challenging, as the high confidence fine-mapped interval contains > 2,000 genes and even the tentative narrower interval contains > 400 genes. The carpel/grain developmental time courses showed that the QTL acts during the very early stages of carpel/grain development, highlighting the importance of this early stage in determining final grain weight (discussed further in section 5.1.4). However, it was not possible to define the exact time during development when these differences are first established and this meant that we could not select time points for RNA-Seq studies in the same way as for the 5A QTL. This means that we have limited information about genes that might be regulated differently between 6A NILs and therefore related to the final grain weight phenotype.

Based on previous studies that show that predominantly cell proliferation is occurring at the very early stages of grain development (Drea *et al.*, 2005; Radchuk *et al.*, 2011), we hypothesise that the 6A QTL acts to influence cell number. Studies in species including rice and Arabidopsis have

identified genes that influence seed/grain weight through the control of cell proliferation. These genes have a diverse range of functions including transcription factors, G-protein signalling, phytohormone signalling, cell cycle components, cytochromes and proteases (Mizukami & Fischer, 2000; Schruff *et al.*, 2006; Adamski *et al.*, 2009; Huang *et al.*, 2009; Qi *et al.*, 2012; Xia *et al.*, 2013; Du *et al.*, 2014) For example, in rice a *SPL* TF, *OsSPL16*, was found to influence grain size through positive regulation of cell proliferation by modulating the expression of certain components of the cell cycle machinery (Wang *et al.*, 2012). Negative regulators of cell proliferation have also been identified as important for the control of grain size, such as the E3 ubiquitin ligase, *GW2* (Song *et al.*, 2007), the A genome wheat orthologue of which (*TaGW2_A*) was considered as a potential candidate gene for the 6A QTL.

5.1.1.1 Is *TaGW2_A* the causal gene underlying the 6A QTL?

One of the aims of this thesis was to determine whether *TaGW2_A* is the causal gene underlying the 6A QTL as it mapped within the original 6A grain weight QTL interval (Simmonds *et al.*, 2014). This hypothesis was assessed phenotypically and genetically in Chapters 2 and 3. Taking together all the evidence we concluded that *TaGW2_A* is unlikely to be the gene underlying the 6A QTL and that they act through different mechanisms (Chapter 3). The main lines of evidence leading to this conclusion were that:

- *TaGW2_A* maps outside of the 0.28 cM fine-mapped 6A grain width interval
- *TaGW2_A* has no coding region polymorphisms in the parental varieties, Spark and Rialto (Simmonds *et al.*, 2014; non-coding polymorphisms discussed further later)
- *TaGW2_A* NILs have significantly different final grain width and length (two experiments) whilst differences in final grain length were not observed in 6A NILs (three experiments)
- *TaGW2_A* NILs have differences in carpel/grain length throughout carpel/grain development but differences in carpel/grain length were rarely observed between 6A NILs
- *TaGW2_A* NILs have clear differences in carpel width and length at heading, whereas no significant differences in carpel size or weight were observed between 6A NILs at heading

However, this conclusion is subject to confirmation of the 0.28 cM fine-mapped interval using additional phenotypic data from the larger 6A RIL population from the 2017 field trials. This confirmation is particularly critical because the high confidence 4.6 cM fine-mapped interval does include *TaGW2_A* and it cannot be excluded as the causal gene based on phenotypic differences alone, as discussed in Chapter 2. One of the main arguments for *TaGW2_A* as the causal gene underlying the 6A QTL, aside from its effect on grain weight in general, is the presence of an A/G promoter SNP at the -593 bp position between the parental varieties, Spark and Rialto (Simmonds *et al.*, 2014). Previous work in wheat and other species has shown that SNPs in regulatory regions (i.e. non-coding sequence) can underlie major QTL and be responsible for dramatic phenotypic differences. For example, a single SNP in the 5' regulatory region of the *qSH1* gene in rice was found to underlie a major QTL for seed shattering (Konishi *et al.*, 2006). The -593 bp SNP in the

upstream region of *TaGW2_A* has previously been associated with grain width and TGW in Chinese germplasm, however studies have generated contradictory results (Su *et al.*, 2011; Zhang, X *et al.*, 2013). These studies used association analysis in similar panels of Chinese germplasm to identify a putative effect of the *TaGW2_A* promoter SNP on TGW. Su *et al.* found that the A allele at the -593 position was associated with increased grain weight, whilst Zhang *et al.* found that the G allele was associated with increased grain weight. This could be explained by the extended LD that exists in the *TaGW2_A* region given its proximal position on chromosome 6A as observed in the 6A RIL populations in Chapter 3. This would determine extended haplotypes that could encompass hundreds of genes in addition to *TaGW2_A*, any of which could underlie the observed variation in grain weight.

Direct manipulation of *TaGW2_A* through induced (Simmonds *et al.*, 2016; Chapter 2) and natural missense mutations (Yang *et al.*, 2012) has clearly established a role of *TaGW2_A* on grain size in wheat. Molecular studies have also shown that the ubiquitination activity of rice *GW2* is conserved in *TaGW2_A* (Bednarek *et al.*, 2012). However, it is still an open question as to whether the association effects in the two contradictory studies are due to allelic differences in *TaGW2_A* itself or in a linked gene across the haplotype block. Indeed, the same logic can be applied to the 4.6 cM fine-mapped grain width interval on chromosome 6A in this thesis: in a region that contains > 2,000 genes it is not possible to say whether *TaGW2_A* is the causal gene or not regardless of the presence of the promoter SNP. This is reminiscent of the cloning of the pre-harvest sprouting QTL (*Phs-A1;* discussed previously in Chapter 3.5.1.2) where *PM19-A1* was incorrectly identified as the causal gene due to the presence of a promoter deletion and a demonstrated effect of the gene on grain dormancy through direct manipulation (Barrero *et al.*, 2015). However, it was subsequently shown that *PM19-A1* was in fact linked to the true causal gene, *TaMKK3*, in the germplasm studied resulting in a spurious association with the QTL phenotype and *PM19* promoter deletion (Shorinola *et al.*, 2017).

If data from the 2017 field trials confirm that *TaGW2_A* maps separately from the 6A QTL this will open up some interesting new avenues for potential further studies. An important question to ask will be precisely which aspects of grain development the 6A gene and *TaGW2_A* affect. Given that the two pairs of NILs seem to have similar phenotypic differences, it is possible that the two genes may influence the same processes. Characterisation of the NILs on a cellular level during carpel/grain development will provide insights into this and these studies are currently underway. Additionally, it would be interesting to understand how the two genes interact and whether beneficial alleles of both genes can be combined to give additive or synergistic increases in TGW. If so, then this could have implications in breeding as well as providing mechanistic insight. Currently, breeders are selecting for a large physical region on chromosome 6A, encompassing both the 6A grain weight effect and *TaGW2_A*. The separation of these two loci could allow for

novel combinations of alleles to be deployed, although this would still be limited by the low rates of recombination observed across this region on chromosome 6A.

5.1.1.2 Future steps to identify genes and pathways underlying the 6A QTL

The fact that genes with so many diverse functions can influence cell proliferation makes it premature to speculate on the identity of a candidate gene from 488 genes solely based on predicted function. In addition to the additional phenotypic data for the larger RIL population, the marker density across the interval will also be increased. No gene based SNP calling has yet been conducted on the 6A NILs/parental cultivars and this will be performed using exome capture data. This will be useful for identifying additional markers and also will identify genes with potentially deleterious mutations, which could assist in prioritising candidate genes for further study. In order to identify non-coding polymorphisms, a promoter capture array will also be employed to access variation in the 2 kb upstream of all genes in this region and this will again help to identify additional markers and potential candidate genes. As more complete genome sequences of additional wheat varieties become available, sequence variation in other regulatory regions will also be explored.

5.1.2 Genes and pathways underlying the 5A QTL

More insight was gained into the potential genes and mechanisms underlying the 5A QTL during this PhD. We found that the 5A QTL acts primarily to increase grain length and this was associated with increased cell length in the pericarp. The first differences in grain length were observed at around 12 dpa (8 – 15 dpa across years, ~ 6.5 mm), which is consistent with a role of the QTL in cell expansion as cell proliferation in the pericarp decreases shortly after fertilisation (Drea *et al.*, 2005; Radchuk *et al.*, 2011). Similar to the 6A and *TaGW2_A* data, the results from the 5A NILs emphasise the importance of the early stages of grain development in determining the final grain size.

Overall, these results suggest that the gene(s) underlying the 5A QTL either directly or indirectly regulate cell expansion in the pericarp, a mechanism that is known to be a key determinant of grain/seed size in several species. Some genes, such as expansins and XTH (xyloglucan endotransglucosylase/hydrolases), affect cell expansion directly by physically modifying or "loosening" the cell wall (reviewed in Cosgrove, 2005), and the expression of these enzymes has been associated with pericarp cell expansion in wheat and barley (Lizana *et al.*, 2010; Radchuk *et al.*, 2011; Munoz & Calderini, 2015). The properties of the cell wall can also be modified, for example accumulation of certain tannins in the cell wall can change its competence for elongation. The Arabidopsis WRKY transcription factor, *TTG2*, regulates some steps of the tannin biosynthesis pathway. *ttg2* mutants have smaller seeds due to smaller cells in the seed coat, likely due to a reduced capacity of the cell wall for elongation due to altered tannin levels (Johnson *et al.*, 2002; Garcia *et al.*, 2005). In rice, *SRS3*, a kinesin 13 protein, was shown to regulate grain length through

cell size likely through the regulation of microtubule dynamics (Kitagawa *et al.*, 2010). Other genes regulate pericarp/seed coat cell size through more indirect mechanisms, for example through the regulation of sugar metabolism and subsequent accumulation in the vacuole (Ohto *et al.*, 2005; Ohto *et al.*, 2009) and endoreduplication (Chevalier *et al.*, 2014). Many of the genes identified within the fine-mapped region(s) for grain length have functional annotations similar to these genes, but as with the 6A QTL, the intervals remain too large to speculate on the identity of the causal gene(s) based on function alone.

As discussed briefly in Chapter 4, seed/grain development requires the coordination of processes across the pericarp/seed coat, endosperm and embryo. It has been proposed in multiple species, that the size of the maternal pericarp/seed coat exerts its influence on final grain size by physically restricting endosperm growth (Calderini *et al.*, 1999; Adamski *et al.*, 2009; Hasan *et al.*, 2011). Grain size in rice is limited by the size of the spikelet hull in an analogous way (Song *et al.*, 2005). In wheat, both pericarp width (Gegas *et al.*, 2010; Simmonds *et al.*, 2016) and length (Lizana *et al.*, 2010; Hasan *et al.*, 2011) have been proposed as key determinants of final grain size. The results from this thesis support this idea and we hypothesise that the 5A cell expansion effect increases the physical space available for endosperm growth during the middle and late stages of grain development. This increased physical capacity could then lead to the increase in grain width that is only established at the later stages of grain development, consistent with the time during grain development associated with grain filling and endosperm growth (Olsen, 2001; Shewry *et al.*, 2012). It is not clear whether the increased capacity for grain filling is utilised by increasing the rate or duration of grain filling in 5A+ NILs. A more detailed time course of grain development with more frequent time points and continuing until the final grain weight had been achieved would be required to determine this. Additionally, time courses would ideally be measured in degree days rather than absolute days to properly calculate grain filling rates, especially in order to compare across years whilst accounting for environmental variation in temperature. Unfortunately, uninterrupted weather data from the weather station at Church farm across the entire time course was not available in any year.

It has been shown that the cross-talk between the endosperm and pericarp/seed coat extends beyond purely mechanical constraints and increased cell size in the seed coat/pericarp can be achieved as an indirect effect of increased endosperm growth. For example, the *HAIKU (IKU)* genes act to promote endosperm growth in Arabidopsis. *iku* mutants have smaller seeds due to reduced endosperm growth and indirectly reduce cell elongation in the integument/seed coat (Garcia *et al.*, 2003). The indirect effect on cell size in the seed coat (a maternal tissue) was determined by demonstrating that *iku* double mutants pollinated with WT pollen had WT-like seeds, therefore showing that the *iku* mutations do not have a direct effect on the maternal integument. Already this could suggest a level of communication between the two tissues (Garcia *et al.*, 2003). The *IKU* genes interact on a genetic basis with *TTG2* (described above) and *iku ttg2* double mutants have

seeds even smaller than *iku* mutants, due to the *ttg2* mutation compromising the elongation capacity of integument cell walls hence restricting endosperm growth further. This is in accordance with the size of the pericarp imposing a physical constraint on endosperm. However, combining the *iku* mutations with lines that have reduced cell proliferation in the integuments (due to overexpression of *KIP RELATED PROTEIN2*) did not show an additive effect on seed size and instead the reduction in cell number in the integument was compensated for by increased cell elongation (Garcia *et al.*, 2005). This suggests that in some cases the size of the pericarp/seed coat can be adjusted to accommodate the growth of the endosperm, providing additional evidence there must be communication/signalling between the tissues. An example of communication from seed coat to endosperm/embryo in cereals can be seen in the control of seed dormancy. A group of three genes, known as the R genes, are responsible for determining grain colour specifically by controlling pigmentation in the seed coat. It is proposed that a pleiotropic effect of the seed coat pigmentation is to regulate grain dormancy (Flintham, 2000). The exact nature of the communication between tissues is not fully understood. Whilst much progress has been made in species such as Arabidopsis, with roles demonstrated for phytohormones, epigenetic factors and sugars (amongst others; reviewed in Nowack *et al.*, 2010; Locascio *et al.*, 2014; Radchuk & Borisjuk, 2014), still relatively little is understood about the molecular basis of this signalling in cereals. Caution should be exercised when translating insight gained from Arabidopsis into cereals as it is possible that not all these processes and mechanisms are conserved, particularly as there are fundamental differences in the final composition of the seed/grain. For example, the Arabidopsis endosperm consists of a single cell type whilst the endosperm of mature wheat grains contains four major cell types (Olsen, 2001).

From the results presented in this thesis, it is therefore not possible to say conclusively whether the increased pericarp cell size in 5A+ NILs is due to a direct effect on cell expansion in the pericarp or an indirect effect of increased endosperm growth. As discussed previously, the early stage at which the grain length phenotype appears would suggest a direct effect on pericarp cell size, but this will need to be confirmed genetically. This could be tested through the assessment of pericarp cell size in F_1 hybrids from reciprocal crosses between 5A- and 5A+ NILs. As the pericarp is an exclusively maternal tissue, if the 5A QTL directly affects cell size in the pericarp then F_1 grains resulting from a 5A+ NIL pollinated by a 5A- NIL would have the 5A+ large pericarp cell size phenotype, whilst the reciprocal cross would not. Conversely, if the 5A QTL affects pericarp cell size as an indirect effect of endosperm growth then only F_1 grains from the 5A- NIL pollinated by the 5A+ NIL would have the large pericarp cell size phenotype. These experiments are currently being conducted. Usually, studies of this nature are challenging in wheat as the subtle phenotypic differences associated with QTL in polyploids can be masked by the phenotypic variation observed between individual F_1 grains (e.g. ~ 5% difference in grain size components in the case of the 5A and 6A QTL). However, the robust effect on pericarp cell size in the 5A NILs that is independent

of absolute grain length could overcome this and opens up new opportunities for parent of origin studies in wheat.

It would also be interesting to assess how the development of the endosperm is affected by the 5A QTL. Whilst the work in this PhD has provided insights into the mechanisms underlying both QTL by breaking down overall grain yield into its constituent parts in the form of specific grain size components, the understanding remains mostly on a whole grain level, both from the phenotype and transcriptome points of view. The next steps to take would be to dissect this down even further to look at the individual tissues within the grain such as the endosperm, embryo and pericarp. Indeed, even breaking the grain down into the three main tissues remains quite a simplistic view as each tissue is composed of several different layers and cell types (Figure 1.8). It would be very interesting to examine these tissues microscopically during carpel/grain development to understand the effects of the QTL in more mechanistic detail and this could be complemented by tissue specific expression studies.

5.1.2.1 Genes selected for further characterisation using TILLING mutants

Combining information about the 5A QTL obtained from the phenotypic characterisation, genetic mapping and transcriptomic study we selected a subset of 14 genes to characterise further through the generation of TILLING mutants (Table 5.1). We identified lines in the exome-sequenced tetraploid TILLING population (Krasileva *et al.*, 2017) with deleterious mutations in A and B homoeologues of each of the genes and generated double mutants where possible. These candidate genes were largely selected from the set of DE genes identified in the transcriptomic study based on their location in the fine-mapped interval or having a functional annotation potentially related to the control of grain size. One gene was selected due to having a missense SNP between NILs and being located in the 5A grain length fine-mapped region (*GL2* interval). These selections were made using the 2014 CSS gene models, before the more complete TGAC gene models were available. Hence, some omissions might have been made due to the lack of information at the time of selection.

Table 5.1: Genes selected to generate double knock-out mutants in the tetraploid TILLING lines

TGAC gene name	Annotation	Reason
TRIAE_CS42_5AL_TGACv1_373986_AA1186560	Kinesin-like protein	DE; function & *GL1*
TRIAE_CS42_5AL_TGACv1_375266_AA1218860*	Trehalose-6-Phosphate	DE; function
TRIAE_CS42_5AL_TGACv1_375845_AA1227940*	RING domain	DE; function
TRIAE_CS42_5AL_TGACv1_375845_AA1227990	Abi superfamily, CAAX protease	DE; function
TRIAE_CS42_1BS_TGACv1_049354_AA0149980	UBCc domain; E2 ubiquitin conjugating enzyme	DE; function
TRIAE_CS42_5AL_TGACv1_374542_AA1202810	RING/U-box superfamily protein; putative E3 ligase	DE; function &*GL1*
TRIAE_CS42_5AL_TGACv1_374321_AA1196890	Aspartyl aminopeptidase; Zinc peptidase-like superfamily	DE; function
TRIAE_CS42_5AL_TGACv1_374097_AA1190230	Ubiquitin, Polyubiquitin 14; NEDD8-like protein RUB1	DE; function
TRIAE_CS42_5AL_TGACv1_375361_AA1220430	RAD23, ubiquitin receptor, proteasome associated	DE; function
TRIAE_CS42_5AS_TGACv1_392558_AA1260860	SSXT protein; GRF1-interacting-factor	DE; function
TRIAE_CS42_5AL_TGACv1_376107_AA1232210*	RNA-binding protein 25; splicing related	DE; function
TRIAE_CS42_5AL_TGACv1_374727_AA1207530	DUF810 family protein	DE; *GL2*
TRIAE_CS42_5AL_TGACv1_375949_AA1229270	TauE superfamily	DE; *GL2*
TRIAE_CS42_5AL_TGACv1_377065_AA1243680	TATA binding protein	SNP; *GL2*

DE = differentially expressed in the 5A RNA-Seq experiment, SNP = predicted missense SNP between 5A NILs, *GL1* = located in the *GL1* fine-mapped interval, *GL2* = located in the *GL2* fine-mapped interval, function = selected based on function related to grain/organ size, * indicates that the gene was DE in the original CSS analysis but not in the TGAC reference (mutants were selected prior to the release of the TGACv1 reference).

For each gene, the A and B single mutant lines have been crossed (by visiting fellow Abdul Kader Alabdullah) and F_1 seeds self-pollinated to generate F_2 populations segregating for the mutations in various combinations. The F_2 populations are currently being phenotyped for grain size components to see if there are any associations with the mutated genes. In this way we will be able to determine if any of these genes have a potential role in the control of grain size. It is possible that none of the selected genes are the causal gene(s) underlying the QTL, particularly as only five of them lie within the fine-mapped interval(s). However, any genes that are found to be associated with differences in grain size components will be interesting novel candidates to characterise further in the context of grain size control in wheat.

5.1.3 Maternal control of grain size

All three pairs of NILs assessed (5A, 6A and *TaGW2_A*) point towards the maternal control of final grain size. Differences in carpel size were observed between *TaGW2_A* NILs before heading suggesting that *TaGW2_A* acts on maternal tissue. Borderline non-significant differences in carpel size were observed between 6A NILs, suggesting that this QTL could also act on maternal tissue before fertilisation. Although differences in grain length were established after fertilisation in 5A

NILs, the QTL is associated with larger cells in the pericarp, a maternal tissue. Although as discussed above, this may or may not be as a result of a direct effect on pericarp cell expansion.

The maternal control of seed/grain size has been demonstrated both genetically and phenotypically in many species including Arabidopsis, rice, wheat and maize (Hasan *et al.*, 2011; Li & Li, 2015; Zhang *et al.*, 2016). Studies have shown that the maternal control of seed/grain size can be exerted through a range of mechanisms, affecting cells in maternal tissues both pre- or post-fertilisation. For example, the Arabidopsis gene, *KLUH* acts maternally to increase seed size through the positive regulation of cell proliferation in the integument (Adamski *et al.*, 2009) and studies suggest that this function is conserved in the wheat orthologue, *TaCYP78A* (Ma *et al.*, 2016). Conversely, the Arabidopsis *ARF2* gene acts as a negative regulator of seed size with a loss-of-function mutant producing 20-40% heavier seeds. The increase in seed weight was associated with increased numbers of cells in the seed coat as a result of increased cell proliferation in the integument/ovule before fertilisation (Schruff *et al.*, 2006). Similarly, *GW2* in rice and its orthologue in Arabidopsis (*DA2*) influence grain/seed size through restriction of cell proliferation in the maternal tissue (Song *et al.*, 2007; Xia *et al.*, 2013). This is consistent with the results from this thesis and Simmonds *et al.* (2016) that *TaGW2_A* acts on maternal tissue, although the effect on cell size and number has not yet been determined. *DA1*, a target of *DA2* in Arabidopsis, also acts synergistically with *DA2* to limit cell proliferation in the integument (Xia *et al.*, 2013; Dong *et al.*, 2017). Genes have also been identified that act maternally to influence cell expansion in maternal tissue, for example *TTG2* and *AP2* (discussed above; Garcia *et al.*, 2005; Ohto *et al.*, 2005).

Programmed cell death (PCD) in the pericarp tissue has also been shown to be important for the maternal control of grain size. It has been proposed that PCD is an important step for enlargement of the pericarp to accommodate endosperm growth. Downregulation of *VACUOLAR-PROCESSING ENZYME 4* (*VPE4*) by RNAi in barley resulted in delayed PCD in the pericarp and consequently smaller grains (Radchuk *et al.*, 2017). One of the DE genes between 5A NILs was annotated as *VPE4* (Table 4.11) and taking this together with the differential regulation observed of proteolytic components, this could suggest a role of PCD in regulation of grain size in the 5A NILs. However, the most extensive PCD occurs during the later stages of grain development (Radchuk *et al.*, 2011; Radchuk *et al.*, 2017) and so this could be a downstream effect. Differences in the progression of PCD in 5A NILs could be identified by a histological analysis of developing grains as in Radchuk *et al.* (2017). However, this might be challenging due to the number of samples that may be required to detect subtle differences between NILs.

The assignment of the 5A, 6A and *TaGW2_A* effects to the maternal parent will need to be confirmed genetically and this could be determined with the F_1 experiments described above. The maternal parent may contribute to final seed size through other mechanisms in addition to the presence of the pericarp/seed coat and the mechanical constraints it imposes. The mother plant

plays many important roles in the development of the grain/seed including provisioning of nutrients to the developing grain, responses to the environment during grain development and the imprinting of genes after fertilisation, all of which have been shown to influence final grain size (discussed in Zhang *et al.*, 2016). If the effects of either the 6A or 5A QTL can be assigned to the maternal parent then it will be interesting to understand exactly how the maternal parent contributes to the final phenotype. Identifying genes that act maternally to influence grain size could have advantages in a breeding context, particularly with respect to hybrid seed generation.

5.1.4 Importance of early grain development

Regardless of the putative direct maternal effects of the 5A and 6A QTL, all three pairs of NILs highlight the importance of early carpel/grain development in determining final grain size, consistent with previous studies in wheat and other cereals (Calderini *et al.*, 1999; Golan *et al.*, 2015; Simmonds *et al.*, 2016). However, despite the importance of these early stages, relatively little is known about the mechanisms underlying early grain development. Studies have characterised these stages phenotypically to a certain extent (Drea *et al.*, 2005; Radchuk *et al.*, 2011) but most characterisation has focussed on the later stages of grain development, mainly on endosperm development. The same is especially true in terms of characterisation on the transcriptional and molecular level, as discussed briefly in Chapter 4. Although numerous wheat grain RNA-Seq studies have been performed, very few have focussed on stages of grain development as early as those described in Chapter 4. This is evidenced by the fact that of 148 grain RNA-Seq samples in the wheat expVIP database only six were taken at stages earlier than 8 dpa, four of which formed part of the same study (Gillies *et al.*, 2012; Choulet *et al.*, 2014). Additionally, there are no RNA-Seq samples in the expVIP database from ovules i.e. pre-anthesis (Borrill *et al.*, 2016). The results from this thesis strongly suggest that understanding the mechanisms underlying these early stages will be critical to identify ways to manipulate final grain size. Based on the ability of grains to compensate for early events in grain development, it is tempting to speculate that manipulating genes and pathways that affect these early stages could provide grain size increases that are more robust to environmental variation.

5.2 Potential consequences of increasing grain size and pleiotropic effects of the 5A and 6A QTL

Increases in grain weight are often associated with pleiotropic effects either on the grain itself or on other plant organs. When considering the pleiotropic effects of QTL, the effects could be due to other genes within the QTL interval. Alternatively, they could be due to the gene(s) controlling grain size themselves, either as an indirect result of increasing grain size or as a direct effect of the gene in another part of the plant.

5.2.1 Pleiotropic effects on yield components

Despite grain weight being more stably inherited than overall yield itself, increases in grain weight have previously been associated with negative pleiotropic effects on other yield components such as grain number and spike number (Kuchel *et al.*, 2007). However, results from this PhD and other studies have shown that these components are not inherently linked and can be genetically separated (Griffiths *et al.*, 2015). Indeed, the fact that alterations in spike and grain number have downstream effects on grain size, likely due to competition for resources, does not necessarily mean that changes in grain size will affect grain and spike number. During this PhD, these components were assessed in the 5A and 6A NILs. In the 6A NILs there were no consistent negative effects across years on either grain number or spike (tiller) number, although there were some negative effects in individual years (Table 2.3). This was consistent with the previous studies of these NILs (Simmonds *et al.*, 2014). In terms of the 5A NILs, 5A+ NILs had significantly reduced grain and tiller number across years although these effects were both driven by particularly strong effects in a single year (Table 2.9). We hypothesised that the combination of these smaller negative effects could explain why neither the 6A or 5A NILs had consistent differences in final grain yield, despite consistent increases in TGW. Alternatively, our evaluation of yield components based on a ten spike sampling might not be robust enough to allow us to detect differences. This could be due to the fact that we usually select ten spikes, corresponding to the main tiller in most cases. For these spikes, we observe increase in spike yield (Table 2.3, Table 2.9). However, by using this sampling strategy we could be missing pleiotropic effects on spikes further behind in development (e.g. third or fourth spike), which could arise from compensation effects from the larger grains in the main spikes of the 6A+ and 5A+ NILs. The negative effects on other yield components could either be as a result of additional genes in the introgressed regions of the NILs or as an effect of the causal genes themselves. For example, a minor QTL for tiller number was identified in the 6A introgression, but this mapped distal to the QTL for TGW (Simmonds *et al.*, 2014). This suggests that in the 6A NILs the pleiotropic effect on tiller number is due to another gene in the interval rather than an effect of the 6A causal gene itself.

5.2.2 Pleiotropic developmental effects

We also observed developmental differences between 6A NILs, including differences in flowering time and senescence, with 6A+ NILs flowering earlier and senescing later. This could suggest that the 6A QTL is associated with an extended grain filling period. However, this was not assessed directly and it has previously been shown that an increase in the time between flowering and senescence (green canopy duration; GCD) does not always result in an increased duration of grain filling (Borrill *et al.*, 2015b). Similar to the tillering effect, a QTL for GCD was identified in the 6A NIL introgression but did not show any correlation with TGW in the original QTL analysis and so the two traits are likely to be under independent genetic control (Simmonds *et al.*, 2014). That said, some genes that influence grain/seed size in other species have also been shown to affect other

developmental traits such as senescence and flowering time. For example, in Arabidopsis, *DA1* acts a negative regulator of seed size but also promotes senescence with mutants having larger seeds and delayed senescence (Li *et al.*, 2008; Vanhaeren *et al.*, 2016). Additionally, *DA1* also affects the size of other organs in addition to the seed, for example it acts as a negative regulator of petal and leaf size suggesting that it is a general regulator of organ growth rather than specifically seed size.

It is not known whether the 5A and 6A genes have grain specific effects or whether they could be general regulators of organ size, for example, these genes could influence leaf and root size. This could have implications both in positive and negative ways. For example, a non-grain specific effect could be seen as wasteful with resources going into non-grain biomass production. Alternatively, plants with larger leaves could have increased photosynthetic capacity through increased area for light interception (reviewed in Long *et al.*, 2006) and a larger root system could also be beneficial. If these genes do have similar effects on the development of grains and leaves e.g. the 5A gene(s) could increase cell expansion in both tissues, then this could be a useful tool for determining the mechanism by which the genes act. The leaf could act as a more tractable system for performing experiments than the grain, and this has proven to be a useful tool for understanding the function of genes that control seed size in Arabidopsis and maize such as *DA1*, *DA2* and *BIG BROTHER* (Rachel prior pers comm; Vanhaeren *et al.*, 2016; Xie *et al.*, 2017). It would be useful to investigate this possibility in subsequent studies and determine if putative effects are sufficiently strong and robust to be properly quantified.

5.2.3 Pleiotropic effects on grain nutrient composition

An avenue that was not explored in this PhD was the effect of these QTL on the composition of the grain itself, aside from increasing the overall size and weight. For example, we did not examine the effect that manipulating the grain size has on the micronutrient, protein or starch content of the grain. Negative correlations between grain weight and grain protein content have been documented (Simmonds, 1995), proposed to be a dilution effect of increased starch in the grain. It is therefore possible that the increases in grain weight associated with the 6A and 5A QTL could be associated with a decrease in nutritional value and quality. Based on the fact that both QTL act during very early grain development, before grain filling and starch accumulation has begun (Drea *et al.*, 2005; Shewry *et al.*, 2012), I would hypothesise that these QTL act to enhance grain filling capacity rather than a particular aspect of grain filling itself. Therefore, I would expect the grain filling process to proceed in the same way, with the relative proportions of protein, starch and micronutrients etc. remaining roughly the same.

5.2.4 Understanding the causes of pleiotropic effects

These pleiotropic effects could be assessed in the 5A and 6A NILs, but this could be challenging for a number of reasons. It will not be possible using the NILs to separate effects that are due to the causal gene(s) themselves from those effects that area a result of other genes within the

introgressed intervals. Additionally, as the phenotypic differences between NILs are very subtle it may be difficult to separate truly biological effects from random biological variation. Identifying the causal gene(s) will allow larger variation to be explored and open new opportunities for studying the function of the gene in more detail. Techniques such as RNAi, CRISPR or TILLING (reviewed in Uauy, 2017) will provide routes to explore a wide range of variation in the underlying genes ranging from understanding the effects of knock-out mutations to exploring how more subtle allelic variations can affect gene function.

5.3 Combining beneficial alleles

Understanding the specific biological mechanisms and genes underlying the 5A and 6A QTL allows hypotheses about combining beneficial alleles of genes to be generated and tested in an informed and targeted way. This can occur on many different levels.

5.3.1 Combining homoeologues

Identifying the causal genes of the 5A and 6A QTL will allow the B and D homoeologues to be identified. This will be important due to the subtle effects of grain weight QTL in hexaploid wheat compared to grain weight QTL in diploid species (Borrill *et al.*, 2015a; Uauy, 2017). Simultaneously modulating the function of all three homoeologues has the potential to expand the range of phenotypic variation and achieve effects comparable to those in diploids, for example *NAM-B1*, the gene underlying the GPC QTL discussed previously (Uauy *et al.*, 2006; Avni *et al.*, 2014). The increased phenotypic range will be important both for understanding gene function and also for providing breeders with novel allelic combinations as simultaneous beneficial mutations in all three homoeologues are unlikely to occur naturally. Alternatively, the three homoeologues may not have completely redundant functions as certain copies may have diverged in function and/or regulation. Although the causal genes underlying the 5A and 6A QTL have not yet been identified, this concept is being explored with the 5A candidate genes discussed above (Table 5.1). Additionally, studies in the lab are currently investigating the effects of combining *TaGW2_A* with TILLING knock-out mutations in the B and D homoeologues (*TaGW2_B* and *_D*). Other groups have also generated lines with mutations in all three *TaGW2* homoeologues using CRISPR (Liang *et al.*, 2017), which provide an alternative and complementary method to understand the function and interaction between the homoeologues.

5.3.2 Combining components of pathways involved in grain size regulation

Different components of the same pathway could be combined to give additive effects on grain size. As discussed above, in Arabidopsis, *DA1* is a target of the E3 ubiquitin ligase *DA2.* Individually, both act to negatively regulate organ size through the suppression of cell proliferation and *da1 da2* double mutants have a synergistic effect on organ size (Xia *et al.*, 2013; Dong *et al.*, 2017). Combining components of the same pathway may not always provide additive/synergistic effects and mutations could be epistatic (i.e. the phenotypic effect of one gene is dependent on the

presence of the second modifier gene). Regardless, identifying specific pathways will allow the function of different components to be investigated and manipulated to fine-tune the final grain size. The ubiquitin-related differentially expressed genes from the 5A RNA-Seq study represent good candidates for initial investigation in this way.

Additionally, genes affecting different pathways and grain size components could be combined to give additive/synergistic effects. For example, combining genes that regulate cell expansion with genes regulating cell proliferation or genes that increase grain length with genes that increase grain width. In the lab, NILs have been generated that combine the 5A QTL (grain length; cell expansion) and the 6A QTL (grain width; possibly cell proliferation) in a common genetic background. Initial results suggest that combining the two QTL does have an additive effect on grain weight through increased grain width and grain length. In the 2017 field trials we also assessed the cell size phenotype of these NILs, the results of which are currently being analysed. We hypothesise that NILs with both the 5A and 6A QTL will have increased cell number and cell size. Interestingly, combining the two QTL seems to have a 'stabilising' effect on the final grain weight. In 2016, the 6A grain width effect did not perform well alone and 6A+ NILs did not have significantly increased grain weight (Table 2.2). However, combining the 5A QTL and 6A QTL still had significantly higher grain weight than the 5A QTL alone, suggesting that there could be some interaction between the two QTL. This was also seen when analysing historical data from the UK public Avalon x Cadenza population (Simmonds, unpublished results). Identifying the genes underlying these QTL will allow the exact nature of this interaction to be investigated further.

5.3.3 Combining grain size genes with other aspects of plant development

Combining genes that affect different yield components could provide a solution to overcome the negative pleiotropic effects associated with increasing individual yield components. For example, increases in grain number are often associated with decreases in grain size and consequently no increase in overall grain yield is achieved. Combining a gene that increases grain number and a gene that influences grain size could act to increase grain number whilst maintaining or enhancing the grain size. Similar approaches could be taken to maintaining or increasing the nutritional value of the grain.

Lastly, whilst this thesis has focussed on the genetic mechanisms underlying grain size and yield, the agronomic aspects should not be ignored. Breeders select to maximise yield under specific planting densities and agronomy conditions. When developing NILs, we modify a single region of the genome which is extremely useful to study the trait in question (grain size in this thesis), but could alter the overall balance of the canopy that was selected to maximise yield. Therefore, it is likely that changes in agronomy practices may be required to maximise the chance that the positive effect on grain size seen in NILs will translate into yield. This is currently being tested for the 2017-2018 field season by modifying seeding rates and fertilisation regimes to better understand the interactions between genetics, environment and agronomy management.

5.4 Concluding statement

Overall, this thesis has provided new insights into the mechanisms controlling grain size in wheat through the characterisation of two distinct grain size QTL in multiple different ways. The results presented here highlight the importance of early grain development in determining final grain size in wheat, and provide direct genetic evidence for the importance of the pericarp tissue. Fine-mapping of the two QTL revealed complex underlying genetic architectures. Although the causal genes were not identified, the intervals were reduced and the new flanking markers have been shared with breeders to facilitate more efficient selection of the beneficial regions. The 5A transcriptomic study identified differentially expressed genes and pathways that could be involved in the control of grain size, a subset of which are now being functionally characterised.

Ultimately, identifying the genes and pathways that control grain size and understanding how they interact will allow breeders to manipulate and fine-tune final grain yield in wheat in novel ways.

6 References

Adamski NM, Anastasiou E, Eriksson S, O'Neill CM, Lenhard M. 2009. Local maternal control of seed size by *KLUH/CYP78A5*-dependent growth signaling. *Proceedings of the National Academy of Sciences* **106**(47): 20115-20120.

Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, Gill BS, Miftahudin, Gustafson JP, Lazo G, Chao S, et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research* **13**(5): 753-763.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403-410.

Andrews S 2010. FastQC: a quality control tool for high throughput sequence data.

Arora S, Singh N, Kaur S, Bains NS, Uauy C, Poland J, Chhuneja P. 2017. Genome-Wide Association Study of Grain Architecture in Wild Wheat Aegilops tauschii. *Frontiers in Plant Science* **8**: 886.

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, et al. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**(6346): 93-97.

Avni R, Zhao RR, Pearce S, Jun Y, Uauy C, Tabbita F, Fahima T, Slade A, Dubcovsky J, Distelfeld A. 2014. Functional characterization of *GPC-1* genes in hexaploid wheat. *Planta* **239**(2): 313-324.

Bao F, Azhakanandam S, Franks RG. 2010. *SEUSS* and *SEUSS-LIKE* Transcriptional Adaptors Regulate Floral and Embryonic Development in *Arabidopsis*. *Plant physiology* **152**(2): 821-836.

Barg R, Sobolev I, Eilon T, Gur A, Chmelnitsky I, Shabtai S, Grotewold E, Salts Y. 2005. The tomato early fruit specific gene *Lefsm1* defines a novel class of plant-specific SANT/MYB domain proteins. *Planta* **221**(2): 197-211.

Barrero JM, Cavanagh C, Verbyla KL, Tibbits JFG, Verbyla AP, Huang BE, Rosewarne GM, Stephen S, Wang P, Whan A, et al. 2015. Transcriptomic analysis of wheat near-isogenic lines identifies *PM19-A1* and *A2* as candidates for a major dormancy QTL. *Genome Biology* **16**(1): 93.

Bednarek JP, Boulaflous A, Girousse C, Ravel C, Tassy C, Barret P, Mouzeyar MFB, Said. 2012. Down-regulation of the TaGW2 gene by RNA interference results in decreased grain size and weight in wheat. *Journal of Experimental Botany* **63**(16): 5945-5955.

Bonnett O. 1936. The development of the wheat spike. *J. agric. Res* **53**: 445-451.

Borrill P, Adamski N, Uauy C. 2015a. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist* **208**(4): 1008-1022.

Borrill P, Fahy B, Smith AM, Uauy C. 2015b. Wheat Grain Filling Is Limited by Grain Filling Capacity rather than the Duration of Flag Leaf Photosynthesis: A Case Study Using *NAM* RNAi Plants. *PloS one* **10**(8): e0134947.

Borrill P, Harrington SA, Uauy C. 2017. Genome-Wide Sequence and Expression Analysis of the NAC Transcription Factor Family in Polyploid Wheat. *G3: Genes|Genomes|Genetics* **7**(9): 3019-3029.

Borrill P, Ramirez-Gonzalez R, Uauy C. 2016. expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant physiology* **170**(4): 2172-2186.

Botella JR. 2012. Can heterotrimeric G proteins help to feed the world? *Trends in plant science* **17**(10): 563-568.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**(5): 525-527.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**.

Breseghello F, Sorrells ME. 2006. Association Mapping of Kernel Size and Milling Quality in Wheat (*Triticum aestivum L.*) Cultivars. *Genetics* **172**(2): 1165-1177.

Breseghello F, Sorrells ME. 2007. QTL analysis of kernel size and shape in two hexaploid wheat mapping populations. *Field Crops Research* **101**(2): 172-179.

Brinton J, Simmonds J, Minter F, Leverington-Waite M, Snape J, Uauy C. 2017. Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat. *New Phytologist* **215**(3): 1026-1038.

Calderini D, Abeledo L, Savin R, Slafer GA. 1999. Effect of temperature and carpel size during pre-anthesis on potential grain weight in wheat. *The Journal of Agricultural Science* **132**(04): 453-459.

Cantu D, Pearce SP, Distelfeld A, Christiansen MW, Uauy C, Akhunov E, Fahima T, Dubcovsky J. 2011. Effect of the down-regulation of the high *Grain Protein Content* (*GPC*) genes on the wheat transcriptome during monocarpic senescence. *BMC genomics* **12**(1): 492.

Chen J, Zhang L, Liu S, Li Z, Huang R, Li Y, Cheng H, Li X, Zhou B, Wu S, et al. 2016. The Genetic Basis of Natural Variation in Kernel Size and Related Traits Using a Four-Way Cross Population in Maize. *PloS one* **11**(4): e0153428.

Cheng Y, Qin G, Dai X, Zhao Y. 2007. *NPY1*, a BTB-NPH3-like protein, plays a critical role in auxin-regulated organogenesis in *Arabidopsis*. *Proceedings of the National Academy of Sciences* **104**(47): 18825-18829.

Chevalier C, Bourdon M, Pirrello J, Cheniclet C, Gévaudant F, Frangne N. 2014. Endoreduplication and fruit growth in tomato: evidence in favour of the karyoplasmic ratio theory. *Journal of Experimental Botany* **65**(10): 2731-2746.

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**.

Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, Chiang-Hsieh Y-F, Hou P-F, Yang T-Y, Chang W-C. 2016. PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Research* **44**(Database issue): D1154-D1160.

Clavijo BJ, Garcia Accinelli G, Wright J, Heavens D, Barr K, Yanes L, Di Palma F. 2017a. W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv*.

Clavijo BJ, Venturini L, Schudoma C, Garcia Accinelli G, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, et al. 2017b. 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**: 885-896.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature reviews. Genetics* **6**(11): 836-846.

Cong B, Barrero LS, Tanksley SD. 2008. Regulatory change in *YABBY*-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nature genetics* **40**(6): 800-804.

Cosgrove DJ. 2005. Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **6**(11): 850-861.

Dante RA, Larkins BA, Sabelli PA. 2014. Cell cycle control and seed development. *Frontiers in Plant Science* **5**: 493.

Dinneny JR, Weigel D, Yanofsky MF. 2006. *NUBBIN* and *JAGGED* define stamen and carpel shape in *Arabidopsis*. *Development* **133**(9): 1645-1655.

Disch S, Anastasiou E, Sharma VK, Laux T, Fletcher JC, Lenhard M. 2006. The E3 ubiquitin ligase *BIG BROTHER* controls *Arabidopsis* organ size in a dosage-dependent manner. *Current Biology* **16**(3): 272-279.

Distelfeld A, Li C, Dubcovsky J. 2009. Regulation of flowering in temperate cereals. *Current opinion in plant biology* **12**(2): 178-184.

Dominguez F, Cejudo FJ. 1996. Characterization of the Endoproteases Appearing during Wheat Grain Development. *Plant physiology* **112**(3): 1211-1217.

Dong H, Dumenil J, Lu FH, Na L, Vanhaeren H, Naumann C, Klecker M, Prior R, Smith C, McKenzie N, et al. 2017. Ubiquitylation activates a peptidase that promotes cleavage and destabilization of its activating E3 ligases and diverse growth regulatory proteins to limit cell proliferation in Arabidopsis. *Genes Dev* **31**(2): 197-208.

Drea S, Leader DJ, Arnold BC, Shaw P, Dolan L, Doonan JH. 2005. Systematic spatial analysis of gene expression during wheat caryopsis development. *The Plant cell* **17**(8): 2172-2185.

Du L, Li N, Chen L, Xu Y, Li Y, Zhang Y, Li C, Li Y. 2014. The Ubiquitin Receptor *DA1* Regulates Seed and Organ Size by Modulating the Stability of the Ubiquitin-Specific Protease *UBP15/SOD2* in *Arabidopsis*. *The Plant cell* **26**(2): 665-677.

Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, Li X, Zhang Q. 2006. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics* **112**(6): 1164-1171.

Fang W, Wang Z, Cui R, Li J, Li Y. 2012. Maternal control of seed size by *EOD3/CYP78A6* in *Arabidopsis thaliana*. *The Plant Journal* **70**(6): 929-939.

FAO 2017. Online statistical database: Food balance. FAOSTAT.

FAO, IFAD, UNICEF, WFP, WHO. 2017. The State of Food Security and Nutrition in the World 2017. Building resilience for peace and food security. Rome, FAO

Farré A, Sayers L, Leverington-Waite M, Goram R, Orford S, Wingen L, Mumford C, Griffiths S. 2016. Application of a library of near isogenic lines to understand context dependent expression of QTL for grain yield and adaptive traits in bread wheat. *BMC Plant Biology* **16**.

Feldman M. 2001. The world wheat book: a history of wheat breeding. *Edited by: Bonjean AP, Angus WJ*: 3-53.

Flintham JE. 2000. Different genetic components control coat-imposed and embryo-imposeddormancy in wheat. *Seed Science Research* **10**(1): 43-50.

Fujikura U, Elsaesser L, Breuninger H, Sánchez-Rodríguez C, Ivakov A, Laux T, Findlay K, Persson S, Lenhard M. 2014. Atkinesin-13A Modulates Cell-Wall Synthesis and Cell Expansion in *Arabidopsis thaliana* via the THESEUS1 Pathway. *PLOS Genetics* **10**(9): e1004627.

Gallavotti A, Malcomber S, Gaines C, Stanfield S, Whipple C, Kellogg E, Schmidt RJ. 2011. *BARREN STALK FASTIGIATE1* is an AT-Hook Protein Required for the Formation of Maize Ears. *The Plant cell* **23**(5): 1756-1771.

Gallavotti A, Whipple CJ. 2015. Positional cloning in maize (Zea mays subsp. mays, Poaceae). *Applications in Plant Sciences* **3**(1): apps.1400092.

Garcia D, Gerald JNF, Berger F. 2005. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in Arabidopsis. *The Plant cell* **17**(1): 52-60.

Garcia D, Saingery V, Chambrier P, Mayer U, Jürgens G, Berger F. 2003. *Arabidopsis haiku* Mutants Reveal New Controls of Seed Size by Endosperm. *Plant physiology* **131**(4): 1661-1670.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *ARXIV*.

Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S, Sayers L, Doonan JH, Snape JW. 2010. A genetic framework for grain size and shape variation in wheat. *The Plant cell* **22**(4): 1046-1056.

Gillies SA, Futardo A, Henry RJ. 2012. Gene expression in the developing aleurone and starchy endosperm of wheat. *Plant biotechnology journal* **10**(6): 668-679.

Golan G, Oksenberg A, Peleg Z. 2015. Genetic evidence for differential selection of grain and embryo weight during wheat evolution under domestication. *Journal of Experimental Botany* **66**(19): 5703-5711.

Gonzalez-Navarro OE, Griffiths S, Molero G, Reynolds MP, Slafer GA. 2016. Variation in developmental patterns among elite wheat lines and relationships with yield, yield components and spike fertility. *Field Crops Research* **196**(Supplement C): 294-304.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**(7): 1017-1018.

Griffiths S, Wingen L, Pietragalla J, Garcia G, Hasan A, Miralles D, Calderini DF, Ankleshwaria JB, Waite ML, Simmonds J. 2015. Genetic dissection of grain size and grain number trade-offs in CIMMYT wheat germplasm. *PloS one* **10**(3): e0118847.

Guo Z, Chen D, Schnurbusch T. 2015. Variance components, heritability and correlation analysis of anther and ovary size during the floral development of bread wheat. *Journal of Experimental Botany*.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482**(7385): 339-346.

Haider N. 2013. The origin of the B-genome of bread wheat (*Triticum aestivum L.*). *Russian Journal of Genetics* **49**(3): 263-274.

Hasan AK, Herrera J, Lizana C, Calderini DF. 2011. Carpel weight, grain length and stabilized grain water content are physiological drivers of grain weight determination of wheat. *Field Crops Research* **123**(3): 241-247.

Hawkesford MJ. 2014. Reducing the reliance on nitrogen fertilizer for wheat production. *Journal of Cereal Science* **59**(3): 276-283.

Hershko A, Ciechanover A. 1998. The ubiquitin system. *Annu Rev Biochem* **67**: 425-479.

Hong Y, Chen L, Du L-p, Su Z, Wang J, Ye X, Qi L, Zhang Z. 2014. Transcript suppression of *TaGW2* increased grain width and weight in bread wheat. *Functional & integrative genomics* **14**(2): 341-349.

Huang K, Wang D, Duan P, Zhang B, Xu R, Li N, Li Y. 2017. *WIDE AND THICK GRAIN 1*, which encodes an otubain-like protease with deubiquitination activity, influences grain size and shape in rice. *The Plant Journal* **91**(5): 849-860.

Huang R, Jiang L, Zheng J, Wang T, Wang H, Huang Y, Hong Z. 2013. Genetic bases of rice grain shape: so many genes, so little known. *Trends in plant science* **18**(4): 218-226.

Huang X, Qian Q, Liu Z, Sun H, He S, Luo D, Xia G, Chu C, Li J, Fu X. 2009. Natural variation at the *DEP1* locus enhances grain yield in rice. *Nature genetics* **41**(4): 494-497.

IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**(6194): 1251788.

Jiang W-B, Huang H-Y, Hu Y-W, Zhu S-W, Wang Z-Y, Lin W-H. 2013. Brassinosteroid regulates seed size and shape in Arabidopsis. *Plant physiology* **162**(4): 1965-1977.

Jin Y, Luo Q, Tong H, Wang A, Cheng Z, Tang J, Li D, Zhao X, Li X, Wan J, et al. 2011. An AT-hook gene is required for palea formation and floral organ number control in rice. *Developmental Biology* **359**(2): 277-288.

Johnson CS, Kolevski B, Smyth DR. 2002. *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *The Plant cell* **14**(6): 1359-1375.

Kaplan-Levy RN, Brewer PB, Quon T, Smyth DR. 2012. The trihelix family of transcription factors – light, stress and development. *Trends in plant science* **17**(3): 163-171.

Kaspar-Schoenefeld S, Merx K, Jozefowicz AM, Hartmann A, Seiffert U, Weschke W, Matros A, Mock H-P. 2016. Label-free proteome profiling reveals developmental-dependent patterns in young barley grains. *Journal of Proteomics* **143**: 106-121.

Kirby EJM, Appleyard M, Unit NACGBA. 1987. *Cereal development guide*: Arable Unit, National Agricultural Centre.

Kitagawa K, Kurinami S, Oki K, Abe Y, Ando T, Kono I, Yano M, Kitano H, Iwasaki Y. 2010. A Novel Kinesin 13 Protein Regulating Rice Seed Length. *Plant and Cell Physiology* **51**(8): 1315-1329.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP Caused Loss of Seed Shattering During Rice Domestication. *Science* **312**(5778): 1392-1396.

Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F. 2013. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology* **14**.

Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang X, Borrill P, et al. 2017. Uncovering hidden variation in polyploid wheat. *Proceedings of the National Academy of Sciences* **114**(6): E913-E921.

Kuchel H, Williams KJ, Langridge P, Eagles Ha, Jefferies SP. 2007. Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theoretical and Applied Genetics* **115**(8): 1029-1041.

Kumar A, Mantovani EE, Seetan R, Soltani A, Echeverry-Solarte M, Jain S, Simsek S, Doehlert D, Alamri MS, Elias EM, et al. 2016. Dissection of Genetic Factors underlying Wheat Kernel Shape and Size in an Elite x Nonadapted Cross using a High Density SNP Linkage Map. *Plant Genome* **9**(1).

Kurepa J, Wang S, Li Y, Zaitlin D, Pierce AJ, Smalle JA. 2009. Loss of 26S Proteasome Function Leads to Increased Cell Size and Decreased Cell Number in Arabidopsis Shoot Organs. *Plant physiology* **150**(1): 178-189.

Lambing C, Franklin FCH, Wang C-JR. 2017. Understanding and Manipulating Meiotic Recombination in Plants. *Plant physiology* **173**(3): 1530-1542.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4): 357-359.

Laudencia-Chingcuanco DL, Stamova BS, You FM, Lazo GR, Beckles DM, Anderson OD. 2007. Transcriptional profiling of wheat caryopsis development using cDNA microarrays. *Plant Molecular Biology* **63**(5): 651-668.

Lee BH, Ko J-H, Lee S, Lee Y, Pak J-H, Kim JH. 2009. The Arabidopsis *GRF-INTERACTING FACTOR* Gene Family Performs an Overlapping Function in Determining Organ Size as Well as Multiple Developmental Properties. *Plant physiology* **151**(2): 655-668.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**.

Li J, Jiang J, Qian Q, Xu Y, Zhang C, Xiao J, Du C, Luo W, Zou G, Chen M, et al. 2011. Mutation of Rice *BC12/GDD1*, Which Encodes a Kinesin-Like Protein That Binds to a GA Biosynthesis Gene Promoter, Leads to Dwarfism with Impaired Cell Elongation. *The Plant cell* **23**(2): 628-640.

Li N, Li Y. 2014. Ubiquitin-mediated control of seed size in plants. *Frontiers in Plant Science* **5**: 332.

Li N, Li Y. 2015. Maternal control of seed size in plants. *Journal of Experimental Botany* **66**(4): 1087-1097.

Li N, Li Y. 2016. Signaling pathways of seed size control in plants. *Current opinion in plant biology* **33**: 23-32.

Li Q, Li L, Yang X, Warburton ML, Bai G, Dai J, Li J, Yan J. 2010. Relationship, evolutionary fate and function of two maize co-orthologs of rice *GW2* associated with kernel size and weight. *BMC Plant Biology* **10**(1): 143.

Li YH, Zheng LY, Corke F, Smith C, Bevan MW. 2008. Control of final seed and organ size by the *DA1* gene family in Arabidopsis thaliana. *Genes & Development* **22**(10): 1331-1336.

Liang Z, Chen K, Li T, Zhang Y, Wang Y, Zhao Q, Liu J, Zhang H, Liu C, Ran Y, et al. 2017. Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nature Communications* **8**: 14261.

Liu W, Zhihui Wu, Yufeng Zhang, Dandan Guo, Yuzhou Xu, Weixia Chen, Haiying Zhou, Mingshan You, Li B. 2014. Transcriptome analysis of wheat grain using RNA-Seq. *Frontiers of Agricultural Science and Enginnering* **1**(3): 214-222.

Liu Y, Wang F, Zhang H, He H, Ma L, Deng XW. 2008. Functional characterization of the Arabidopsis ubiquitin-specific protease gene family reveals specific role and redundancy of individual members in development. *The Plant Journal* **55**(5): 844-856.

Lizana XC, Riegel R, Gomez LD, Herrera J, Isla A, McQueen-Mason SJ, Calderini DF. 2010. Expansins expression is associated with grain size dynamics in wheat (*Triticum aestivum L.*). *Journal of Experimental Botany* **61**(4): 1147-1157.

Locascio A, Roig-Villanova I, Bernardi J, Varotto S. 2014. Current perspectives on the hormonal control of seed development in Arabidopsis and maize: a focus on auxin. *Frontiers in Plant Science* **5**: 412.

Long SP, Zhu X-G, Naidu SL, Ort DR. 2006. Can improvement in photosynthesis increase crop yields? *Plant, Cell and Environment* **29**(3): 315-330.

Lukowitz W, Gillmor CS, Scheible W-R. 2000. Positional Cloning in Arabidopsis. Why It Feels Good to Have a Genome Initiative Working for You. *Plant physiology* **123**(3): 795-806.

Luo M-C, Yang Z-L, You FM, Kawahara T, Waines JG, Dvorak J. 2007. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theoretical and Applied Genetics* **114**.

Ma M, Zhao H, Li Z, Hu S, Song W, Liu X. 2016. *TaCYP78A5* regulates seed size in wheat (*Triticum aestivum*). *Journal of Experimental Botany* **67**(5): 1397-1410.

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, et al. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**(7651): 427-433.

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J* **76**.

Mir RR, Kumar N, Jaiswal V, Girdharwal N, Prasad M, Balyan HS, Gupta PK. 2012. Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. *Molecular Breeding* **29**(4): 963-972.

Mizukami Y, Fischer RL. 2000. Plant organ size control: *AINTEGUMENTA* regulates growth and cell numbers during organogenesis. *Proceedings of the National Academy of Sciences* **97**(2): 942-947.

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, et al. 2017. The pangenome of hexaploid bread wheat. *The Plant Journal* **90**(5): 1007-1013.

Munoz M, Calderini DF. 2015. Volume, water content, epidermal cell area, and *XTH5* expression in growing grains of wheat across ploidy levels. *Field Crops Research* **173**: 30-40.

Nowack Moritz K, Ungru A, Bjerkan Katrine N, Grini Paul E, Schnittger A. 2010. Reproductive cross-talk: seed development in flowering plants. *Biochemical Society Transactions* **38**(2): 604-612.

Ohto M-a, Fischer RL, Goldberg RB, Nakamura K, Harada JJ. 2005. Control of seed mass by *APETALA2*. *Proceedings of the National Academy of Sciences of the United States of America* **102**(8): 3123-3128.

Ohto M-a, Floyd SK, Fischer RL, Goldberg RB, Harada JJ. 2009. Effects of *APETALA2* on embryo, endosperm, and seed coat development determine seed size in Arabidopsis. *Sexual Plant Reproduction* **22**(4): 277-289.

Olsen OA. 2001. ENDOSPERM DEVELOPMENT: Cellularization and Cell Fate Specification. *Annu Rev Plant Physiol Plant Mol Biol* **52**: 233-267.

Pallotta M, Warner P, Fox R, Kuchel H, Jefferies S, Langridge P 2003. Marker assisted wheat breeding in the southern region of Australia. *Proceedings of the 10th international wheat genetics symposium, Paestum, Italy*: Istituto Sperimentale per la Cerealicultura Roma, Italy. 789-791.

Pearce S, Vazquez-Gross H, Herin SY, Hane D, Wang Y, Gu YQ, Dubcovsky J. 2015. WheatExp: an RNA-seq expression database for polyploid wheat. *BMC Plant Biology* **15**.

Pellny TK, Lovegrove A, Freeman J, Tosi P, Love CG, Knox JP, Shewry PR, Mitchell RAC. 2012. Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome. *Plant physiology* **158**(2): 612-627.

Peng FY, Hu Z, Yang R-C. 2016. Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. *BMC Genomics* **17**: 573.

Petersen G, Seberg O, Yde M, Berthelsen K. 2006. Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). *Molecular Phylogenetics and Evolution* **39**(1): 70-82.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KFX, Olsen O-A. 2014a. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KFX, Olsen Oa. 2014b. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**(6194): 1250091.

Pielot R, Kohl S, Manz B, Rutten T, Weier D, Tarkowská D, Rolčík J, Strnad M, Volke F, Weber H, et al. 2015. Hormone-mediated growth dynamics of the barley pericarp as revealed by magnetic resonance imaging and transcript profiling. *Journal of Experimental Botany* **66**(21): 6927-6943.

Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods* **14**(7): 687-690.

Pires JC, Conant GC. 2016. Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage in the Evolution of Genomes. *Annual review of genetics* **50**: 113-131.

Qi P, Lin Y-S, Song X-J, Shen J-B, Huang W, Shan J-X, Zhu M-Z, Jiang L, Gao J-P, Lin H-X. 2012. The novel quantitative trait locus *GL3.1* controls rice grain size and yield by regulating *Cyclin-T1;3*. *Cell Research* **22**(12): 1666-1680.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Radchuk V, Borisjuk L. 2014. Physical, metabolic and developmental functions of the seed coat. *Frontiers in Plant Science* **5**: 510.

Radchuk V, Tran V, Radchuk R, Diaz-Mendoza M, Weier D, Fuchs J, Riewe D, Hensel G, Kumlehn J, Munz E, et al. 2017. *Vacuolar processing enzyme 4* contributes to maternal control of grain size in barley by executing programmed cell death in the pericarp. *New Phytologist*.

Radchuk V, Weier D, Radchuk R, Weschke W, Weber H. 2011. Development of maternal seed tissue in barley is mediated by regulated cell expansion and cell disintegration and coordinated with endosperm growth. *Journal of Experimental Botany* **62**(3): 1217-1227.

Ramirez-Gonzalez RH, Uauy C, Caccamo M. 2015. PolyMarker: A fast polyploid primer design pipeline. *Bioinformatics* **31**(February): 2038-2039.

Ray DK, Mueller ND, West PC, Foley Ja. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PloS one* **8**(6): e66428-e66428.

Riefler M, Novak O, Strnad M, Schmülling T. 2006. Arabidopsis cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *The Plant cell* **18**(1): 40-54.

Rogers SO, Quatrano RS. 1983. Morphological Staging of Wheat Caryopsis Development. *American Journal of Botany* **70**(2): 308-311.

Rosegrant MW, Tokgoz S, Bhandary P, Msangi S 2013. Looking ahead: Scenarios for the future of food. *2012 Global Food Policy Report*. Washington, D.C.: International Food Policy Research Institute 88-101.

Sarojam R, Sappl PG, Goldshmidt A, Efroni I, Floyd SK, Eshed Y, Bowman JL. 2010. Differentiating *Arabidopsis* Shoots from Leaves by Combined *YABBY* Activities. *The Plant cell* **22**(7): 2113-2130.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, et al. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**(7): 676-682.

Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, Scott RJ. 2006. The *AUXIN RESPONSE FACTOR 2* gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. *Development* **133**(2): 251-261.

Serrago RA, Miralles DJ, Slafer GA. 2008. Floret fertility in wheat as affected by photoperiod during stem elongation and removal of spikelets at booting. *European Journal of Agronomy* **28**(3): 301-308.

Shewry PR, Mitchell RaC, Tosi P, Wan Y, Underwood C, Lovegrove A, Freeman J, Toole Ga, Mills ENC, Ward JL. 2012. An integrated study of grain development of wheat (cv. Hereward). *Journal of Cereal Science* **56**(1): 21-30.

Shorinola O, Balcárková B, Hyles J, Tibbits JFG, Hayden MJ, Holušova K, Valárik M, Distelfeld A, Torada A, Barrero JM, et al. 2017. Haplotype Analysis of the Pre-harvest Sprouting Resistance Locus *Phs-A1* Reveals a Causal Role of *TaMKK3-A* in Global Germplasm. *Frontiers in Plant Science* **8**(1555).

Simmonds J, Scott P, Brinton J, Mestre TC, Bush M, Blanco A, Dubcovsky J, Uauy C. 2016. A splice acceptor site mutation in *TaGW2-A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theoretical and Applied Genetics* **129**(6): 1099-1112.

Simmonds J, Scott P, Leverington-Waite M, Turner AS, Brinton J, Korzun V, Snape J, Uauy C. 2014. Identification and independent validation of a stable yield and thousand grain

weight QTL on chromosome 6A of hexaploid wheat (*Triticum aestivum L.*). *BMC Plant Biology* **14**(1): 191.

Simmonds NW. 1995. The relation between yield and protein in cereal grain. *Journal of the Science of Food and Agriculture* **67**(3): 309-315.

Slafer GA. 2003. Genetic Basis of Yield as Viewed From a Crop Physiologist 's Perspective. *Annals of Applied Biology* **142**(2): 117-128.

Song QJ, Shi JR, Singh S, Fickus EW, Costa JM, Lewis J, Gill BS, Ward R, Cregan PB. 2005. Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics* **110**(3): 550-560.

Song X-J, Huang W, Shi M, Zhu M-Z, Lin H-X. 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature genetics* **39**(5): 623-630.

Stone SL, Hauksdóttir H, Troy A, Herschleb J, Kraft E, Callis J. 2005. Functional Analysis of the RING-Type Ubiquitin Ligase Family of Arabidopsis. *Plant physiology* **137**(1): 13-30.

Strange H, Zwiggelaar R, Sturrock C, Mooney SJ, Doonan JH. 2015. Automatic estimation of wheat grain morphometry from computed tomography data. *Functional Plant Biology* **42**(5): 452-452.

Street IH, Shah PK, Smith AM, Avery N, Neff MM. 2008. The AT-hook-containing proteins *SOB3/AHL29* and *ESC/AHL27* are negative modulators of hypocotyl growth in Arabidopsis. *The Plant Journal* **54**(1): 1-14.

Su Z, Hao C, Wang L, Dong Y, Zhang X. 2011. Identification and development of a functional marker of *TaGW2* associated with grain weight in bread wheat (*Triticum aestivum L.*). *Theoretical and Applied Genetics* **122**(1): 211-223.

Sukumaran S, Dreisigacker S, Lopes M, Chavez P, Reynolds MP. 2015. Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoretical and Applied Genetics* **128**(2): 353-363.

Sun F, Guo G, Du J, Guo W, Peng H, Ni Z, Sun Q, Yao Y. 2014. Whole-genome discovery of miRNAs and their targets in wheat (*Triticum aestivum L.*). *BMC Plant Biology* **14**(1): 142.

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences* **108**(50): 20260-20264.

Torada A, Koike M, Ogawa T, Takenouchi Y, Tadamura K, Wu J, Matsumoto T, Kawaura K, Ogihara Y. 2016. A Causal Gene for Seed Dormancy on Wheat Chromosome 4A Encodes a MAP Kinase Kinase. *Current Biology* **26**(6): 782-787.

Trick M, Adamski NM, Mugford SG, Jiang C-C, Febrer M, Uauy C. 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biology* **12**(1): 14-14.

Trusov Y, Botella JR. 2016. Plant G-Proteins Come of Age: Breaking the Bond with Animal Models. *Frontiers in Chemistry* **4**(24).

Uauy C. 2017. Wheat genomics comes of age. *Current opinion in plant biology* **36**: 142-148.

Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J. 2006. A *NAC* gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**(5803): 1298-1301.

United Nations 2015. World Population Prospects: The 2015 Revision.

Vanhaeren H, Nam Y-J, De Milde L, Chae E, Storme V, Weigel D, Gonzalez N, Inzé D. 2016. Forever Young: The Role of Ubiquitin Receptor *DA1* and E3 Ligase *BIG BROTHER* in Controlling Leaf Growth and Development. *Plant physiology*: 01410.

Voss-Fels KP, Qian L, Parra-Londono S, Uptmoor R, Frisch M, Keeble-Gagnère G, Appels R, Snowdon RJ. 2017. Linkage drag constrains the roots of modern wheat. *Plant, Cell & Environment* **40**(5): 717-725.

Wan Y, Poole RL, Huttly AK, Toscano-Underwood C, Feeney K, Welham S, Gooding MJ, Mills C, Edwards KJ, Shewry PR. 2008. Transcriptome analysis of grain development in hexaploid wheat. *BMC genomics* **9**.

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, et al. 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnology journal*: 787-796.

Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, Zeng R, Zhu H, Dong G, Qian Q, et al. 2012. Control of grain size, shape and quality by *OsSPL16* in rice. *Nature genetics* **44**(8): 950-954.

Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X, et al. 2008. Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. *Cell Research* **18**(12): 1199-1209.

Williams K, Sorrells ME. 2014. Three-Dimensional Seed Size and Shape QTL in Hexaploid Wheat (Triticum aestivum L.) Populations. *Crop Science* **54**(1): 98-110.

Winfield MO, Allen AM, Burridge AJ, Barker GLA, Benbow HR, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G, et al. 2016. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant biotechnology journal* **14**(5): 1195-1206.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(9): 1859-1875.

Xia T, Li N, Dumenil J, Li J, Kamenski A, Bevan MW, Gao F, Li Y. 2013. The Ubiquitin Receptor *DA1* Interacts with the E3 Ubiquitin Ligase *DA2* to Regulate Seed and Organ Size in Arabidopsis. *The Plant cell* **25**(9): 3347-3359.

Xiao W, Brown RC, Lemmon BE, Harada JJ, Goldberg RB, Fischer RL. 2006. Regulation of Seed Size by Hypomethylation of Maternal and Paternal Genomes. *Plant physiology* **142**(3): 1160-1168.

Xie G, Li Z, Ran Q, Wang H, Zhang J. 2017. Over-expression of mutated *ZmDA1* or *ZmDAR1* gene improves maize kernel yield by enhancing starch synthesis. *Plant biotechnology journal*.

Xing Y, Zhang Q. 2010. Genetic and molecular bases of rice yield. *Annual review of plant biology* **61**: 421-442.

Yang M, Gao X, Dong J, Gandhi N, Cai H, von Wettstein DH, Rustgi S, Wen S. 2017. Pattern of Protein Expression in Developing Wheat Grains Identified through Proteomic Analysis. *Frontiers in Plant Science* **8**: 962.

Yang Z, Bai Z, Li X, Wang P, Wu Q, Yang L, Li L, Li X. 2012. SNP identification and allelic-specific PCR markers development for *TaGW2*, a gene linked to wheat kernel weight. *Theoretical and Applied Genetics* **125**(5): 1057-1068.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**(2): R14.

Yu Y, Zhu D, Ma C, Cao H, Wang Y, Xu Y, Zhang W, Yan Y. 2016. Transcriptome analysis reveals key differentially expressed genes involved in wheat grain development. *The Crop Journal* **4**(2): 92-106.

Zhang K, Wang J, Zhang L, Rong C, Zhao F, Peng T, Li H, Cheng D, Liu X, Qin H, et al. 2013. Association Analysis of Genomic Loci Important for Grain Weight Control in Elite

Common Wheat Varieties Cultivated with Variable Water and Fertiliser Supply. *PloS one* **8**(3): e57853.

Zhang X, Chen J, Shi C. 2013. Function of *TaGW2-6A* and its effect on grain weight in wheat (*Triticum aestivum L.*). 347-357.

Zhang X, Hirsch CN, Sekhon RS, de Leon N, Kaeppler SM. 2016. Evidence for maternal control of seed size in maize from phenotypic and transcriptional analysis. *Journal of Experimental Botany* **67**(6): 1907-1917.

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *bioRxiv*.

Appendices

Appendix 1

Brinton, J., Simmonds, J., Minter, F., Leverington-Waite, M., Snape, J., Uauy, C. 2017. Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat. New Phytologist, 215: 1026–1038. doi:10.1111/nph.14624

Research

New Phytologist

Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat

Jemima Brinton, James Simmonds, Francesca Minter, Michelle Leverington-Waite, John Snape and Cristobal Uauy

John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

Author for correspondence:
Cristobal Uauy
Tel: +44 0 1603 45019
Email: cristobal.uauy@jic.ac.uk

Received: *17 March 2017*
Accepted: *26 April 2017*

New Phytologist (2017) **215:** 1026–1038
doi: 10.1111/nph.14624

Key words: cell size, grain length, grain size, pericarp, quantitative trait loci (QTL), thousand grain weight (TGW), wheat, yield.

Summary

- Crop yields must increase to address food insecurity. Grain weight, determined by grain length and width, is an important yield component, but our understanding of the underlying genes and mechanisms is limited.
- We used genetic mapping and near isogenic lines (NILs) to identify, validate and fine-map a major quantitative trait locus (QTL) on wheat chromosome 5A associated with grain weight. Detailed phenotypic characterisation of developing and mature grains from the NILs was performed.
- We identified a stable and robust QTL associated with a 6.9% increase in grain weight. The positive interval leads to 4.0% longer grains, with differences first visible 12 d after fertilization. This grain length effect was fine-mapped to a 4.3 cM interval. The locus also has a pleiotropic effect on grain width (1.5%) during late grain development that determines the relative magnitude of the grain weight increase. Positive NILs have increased maternal pericarp cell length, an effect which is independent of absolute grain length.
- These results provide direct genetic evidence that pericarp cell length affects final grain size and weight in polyploid wheat. We propose that combining genes that control distinct biological mechanisms, such as cell expansion and proliferation, will enhance crop yields.

Introduction

By 2050, it is predicted that the human population will have exceeded 9 billion people (United Nations, 2015). This is driving an increased demand for food production that is exacerbated by the use of crops for fuel and animal feed, and the pressures on agricultural systems resulting from climate change. With land for agricultural expansion being limited, increasing crop yields provides a sustainable route towards meeting this demand. However, rates of yield increase have slowed in recent years and are currently insufficient to achieve the estimated doubling in crop production that will be required by 2050 (Tilman *et al.*, 2011; Ray *et al.*, 2013). Projections show that increasing productivity on existing farmlands would increase the available food supply and lower prices, significantly reducing the number of people at risk of hunger globally (Rosegrant *et al.*, 2013). With one in nine people currently living under food insecurity (FAO *et al.*, 2015), it is urgent that we identify ways to increase crop yields.

Final crop yield is a complex quantitative trait strongly influenced by interacting genetic and environmental factors. For cereal crops, seed/grain weight (measured as thousand grain weight, TGW) is a major yield component and is more stably inherited than final yield itself (Kuchel *et al.*, 2007). Grain weight is largely defined by the size of individual grains and the morphometric components of grain area, length and width. A number of genes controlling these traits have been cloned from major grain weight quantitative trait loci (QTL) in rice (Fan *et al.*, 2006; Song *et al.*, 2007; Weng *et al.*, 2008; Wang *et al.*, 2012). For example, GW2, a RING-type E3 ubiquitin ligase, acts as a negative regulator of cell division and was identified as the gene underlying a major QTL for rice grain width and weight (Song *et al.*, 2007). These studies, in addition to those in model species, have shown that seed size is controlled through diverse mechanisms and genetic pathways (reviewed by Xing & Zhang, 2010; Li & Li, 2015). In Arabidopsis, the *AINTEGUMENTA* (*ANT*) transcription factor increases seed size through increased cell proliferation (Mizukami & Fischer, 2000), whilst the *APETELA2* (*AP2*) transcription factor regulates seed size by limiting cell expansion (Ohto *et al.*, 2005). Other genes include those involved in phytohormone biosynthesis and signalling (Riefler *et al.*, 2006; Schruff *et al.*, 2006; Jiang *et al.*, 2013) and G-protein signalling pathways (Huang *et al.*, 2009). Interestingly, many of these genes have been shown to act maternally (reviewed by Li & Li, 2015) and it has been proposed that the seed coat/pericarp (a maternal tissue) sets an upper limit to the final size of the seed/grain (Adamski *et al.*, 2009; Hasan *et al.*, 2011; Xia *et al.*, 2013).

Despite these advances, our understanding of the control of grain size is more limited in important crop species such as wheat (*Triticum aestivum*). Wheat provides *c.* 20% of the calories consumed by humans and more protein globally than all types of meat combined (FAO, 2017). Many QTL for grain weight and, more recently, individual grain size/shape components have been

identified in wheat (Breseghello & Sorrells, 2007; Gegas *et al.*, 2010; Simmonds *et al.*, 2014; Farre *et al.*, 2016; Kumar *et al.*, 2016). However, no mechanistic insight has been provided for these QTL, few have been validated (Simmonds *et al.*, 2014) and, as yet, none have been cloned.

A major challenge to validate and define the mechanisms governing grain weight QTL in polyploid wheat has been that their effects are often subtle compared with QTL identified in diploid species such as rice (Uauy, 2017). One explanation is that wheat has a more limited capacity for increasing grain size than rice. An alternative, and more likely, scenario is that the effect of variation in an individual gene is masked by functional redundancy from homoeologous gene copies (Borrill *et al.*, 2015); bread wheat is a hexaploid species with three homoeologous genomes (A, B and D) that share 96–98% sequence similarity across genes (Krasileva *et al.*, 2013). In addition, the size (17 Gb) and highly repetitive nature of the wheat genome has meant that, until recently, the genomic resources available in wheat have been limited. However, in the last few years there has been a radical change in the wheat genomics landscape with resources now including complete genome sequences and high-quality gene models (IWGSC RefSeq v.1.0; IWGSC, 2014; Clavijo *et al.*, 2017), transcriptomic databases (Pearce *et al.*, 2015b; Borrill *et al.*, 2016), high-density single nucleotide polymorphism (SNP) arrays (Wang *et al.*, 2014; Winfield *et al.*, 2016) and exome-sequenced mutant populations (Krasileva *et al.*, 2017).

In this study, we identified a stable and robust QTL for grain weight in hexaploid wheat, which is driven by an increase in grain length. The QTL affects cell expansion in the grain and acts to increase the length of cells in the pericarp (maternal seed coat). We genetically mapped the effect to an interval on chromosome 5A, and used the latest wheat genome sequences and gene models to define the genes within the physical space. This detailed characterisation of the QTL provides direct genetic evidence that pericarp cell expansion affects final grain size, offering new insights into the mechanisms controlling grain weight in polyploid wheat.

Materials and Methods

Plant material

A doubled haploid (DH) mapping population was developed from the cross between two UK hexaploid winter wheat (*Triticum aestivum* L.) cultivars, 'Charger' and 'Badger'. The population was created using the wheat × maize technique from F_1 plants (Laurie & Bennett, 1988) and comprised 129 individuals, 92 of which were genotyped and used for evaluation. The 5A QTL was validated with the development of near isogenic lines (NILs). Two DH lines (CB53 and CB89) homozygous for the positive Badger loci across the complete linkage group were crossed to Charger and heterozygous F_1 plants were backcrossed to the Charger recurrent parent for four generations (BC_4). Heterozygous plants were selected at each generation using markers *Xgwm293* and *Xgwm186*. After BC_2 and BC_4, heterozygotes were self-pollinated and NILs homozygous for the alternative alleles across the interval were extracted (BC_2F_2 and BC_4F_2). In total, 10 BC_2 NILs were generated, six of which carried the *Xgwm293* to *Xgwm186* Badger-positive interval. An additional 12 BC_4 NILs were generated from the two DH lines (six Badger and six Charger interval). Two representative BC_4 NILs with alternative haplotypes were genotyped with the 90K iSelect array (Wang *et al.*, 2014) to confirm the introgression and identify additional segregating genomic regions. Recombinant BC_4F_2 plants between the flanking markers were also selected and self-pollinated for the development of homozygous BC_4F_3 recombinant inbred lines (RILs). Screening 170 plants with flanking markers *Xgwm293* and *Xgwm186* yielded 60 recombinants within the interval, defining a genetic interval of 17.65 cM.

Genetic map construction and QTL analysis

The Charger × Badger genetic map was developed using simple sequence repeat (SSR) markers. From 650 SSRs tested, 239 from the JIC/*psp* (Bryan *et al.*, 1997; Stephenson *et al.*, 1998), IPK Gatersleben/*gwm*/*gdm* (Roder *et al.*, 1998; Pestsova *et al.*, 2000), Wheat Microsatellite Consortium/*wmc* (http://wheat.pw.usda.gov/ggpages/SSR/WMC/), Beltsville Agricultural Research Station/*barc* (Song *et al.*, 2005) and INRA/*cfa*/*cfd* (Guyomarc'h *et al.*, 2002) collections were polymorphic between parental lines. Consensus maps (Somers *et al.*, 2004) were used to select 212 SSR markers which maximised genome coverage with an approximate marker density of one SSR every 20 cM. In addition, nine sequence-tagged microsatellite profiling (STMP) markers (Hayden & Sharp, 2001) were incorporated into the map. To increase marker density, 75 Kompetitive Allele Specific Primers (KASP) markers were utilised. Markers with assigned chromosome locations (Allen *et al.*, 2011) were targeted to fill gaps in the genetic map.

DNA extractions and genotyping procedures were performed as in Simmonds *et al.* (2014). Likewise, map construction, QTL detection and multi-trait multi-environment (MTME) analysis was conducted as in Simmonds *et al.* (2014). Significant QTL effects were detected above a 2.5 log-of-odds (LOD) threshold (QTL Cartographer default).

SSR and KASP markers used in the QTL analyses were positioned with respect to the newly released Chinese Spring sequence through a Blast search of 100–300 bp encompassing each SNP against the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v.1.0 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). In most cases, the order on the reference sequence agreed with the genetic order in the Charger × Badger population. For discrepancies, we used the genetic position to order markers. In cases of no hits to the RefSeq v.1.0 assembly, we inferred a physical position based on the two closest markers and the relative distance of all three markers based on their centiMorgan positions. Similarly, physical positions of all iSelect SNPs were obtained using Blast to align the surrounding sequence (201 bp) to the RefSeq v.1.0 assembly. TGACv1 gene models were positioned on RefSeq v.1.0 with GMAP (Wu & Watanabe, 2005) using best hit position and

95% minimum similarity cut-off (D. Swarbreck and G. Kaithakottil, Earlham Institute, Norwick, UK).

Field evaluation and phenotyping

The DH population was evaluated in the field in a randomised complete block design with three replications at five sites (Norwich and Sandringham, England; Balmonth, Scotland; Bohnshausen, Germany; and Froissy, France (Simmonds *et al.*, 2014)). The experiments were continued for 3 yr (2001–2003) at Norwich and Sandringham, and 2 yr (2002–2003) at the other three sites. The field trials were sown in large-scale yield plots (1.1 × 6 m) and treated with standard farm pesticide and fertiliser applications to reproduce commercial practice. All trials were sown by grain number for comparable plant densities per plot (275 seeds m^{-2}). Plots were measured for final plot yield after adjustment for plot size, and TGW was calculated by counting and weighing 100 seeds from each plot.

The NILs were evaluated at Norwich in 2012 and 2013 (10 BC_2 NILs), 2014 (12 BC_4 NILs) and 2015 and 2016 (four BC_4 NILs), while BC_4 RILs were analysed in 2014–2016. For both NIL and RIL experiments, a randomised complete block design was used with five replications. NILs were grown in large-scale yield plots (1.1 × 6 m), whereas RILs were grown in 1.1 × 1 m plots in 2014 and large-scale yield plots in 2015 and 2016. Final grain yield (adjusted by plot size and moisture content) was determined for NILs across the 5 yr. Developmental traits were also measured for NILs in 2012–2016, although not all traits were measured in each year (Supporting Information Table S1). For all NILs (2012–2016) and RILs (2014–2016), grain morphometric measurements (grain width, length, area) and TGW were recorded on the MARVIN grain analyser (GTA Sensorik GmbH, Neubrandenburg, Germany) using *c.* 400 grains obtained from the harvested grain samples. For all NILs (2012–2016), 10 representative spikes per field plot were also measured for spike yield components (spikelet number, number of viable spikelets, spike length, grain number per spike, spike yield and seeds per spikelet), TGW and grain morphometric parameters. The data from the 10 representative spikes were consistent with the whole plot values.

Grain developmental time courses

The BC_4 NILs grown in 2014–2016 were used for the grain developmental time courses. Two Charger (5A−) and two Badger (5A+) NILs were used, and the same NILs were used in all three years. We tagged 65 ears per NIL across each of four blocks in the field at full ear emergence (peduncle just visible) to ensure sampling at the same developmental stage. Ten spikes per NIL, per block, were sampled at each of five (2014) or six (2015–2016) time points. The 2014 time points included 4, 8, 12, 18 and 27 d post-anthesis (dpa). The 2015 time points included anthesis (0 dpa), and 4, 7, 12, 19 and 26 dpa. The 2016 time points included 0, 3, 8, 10, 15 and 21 dpa. Ten grains were sampled from each spike from the outer florets (positions F_1 and F_2) of spikelets located in the middle of the spike. Grains were weighed to obtain fresh weight, assessed for morphometric parameters (grain area, length and width) on the MARVIN grain analyser and then dried at 37°C to constant weight (dry weight). For each block at each time point, a total of *c.* 100 grains were sampled (10 spikes × 10 grains) per NIL. However, for the statistical analysis the average of each NIL within each block was used as the phenotypic value as the individual grains and spikes were considered as subsamples.

Cell size measurements

One representative 5A− and 5A+ BC_4 NIL was used for cell size measurements. We selected mature grains from three blocks of the 2015 harvest samples based on a variety of criteria. For each NIL, we selected nine grains of average grain length from the whole harvest sample from each block (groups 5A−/5A+ average). For the 5A− NIL, an additional nine grains were selected that had grain lengths equivalent to the average of the 5A+ NIL sample (5A− large). For the 5A+ NIL an additional nine grains were selected that had grain lengths equivalent to the average of the 5A− NIL sample (5A+ small). We also selected grains of average length from three blocks of the 2016 harvest (nine grains were selected from each block per genotype). Grains were stuck crease-down on to 12.5 mm diameter aluminium specimen stubs using 12 mm adhesive carbon tabs (both Agar Scientific, Stansted, UK), sputter coated with gold using an Agar high-resolution sputter coater and imaged using a Zeiss Supra 55 scanning electron microscope. The surface (pericarp) of each grain was imaged in the top and bottom half of the grain, with images taken in at least three positions in each half. All images were taken at a magnification of ×500. Cell length was measured using the Fiji distribution of ImageJ (Schindelin *et al.*, 2012) (Fig. S1). Cell number was estimated for each grain using grain length/average cell length. For the statistical analyses, we considered the average cell length of each individual grain as a subsample within the block.

Statistical analysis

DH lines homozygous across the genetic interval for the two major QTL, *Qtgw-cb.2B* (*Xgwm259-Xstm119tgag*) and *Qtgw-cb.5A* (*Xgwm443-XBS00000435*) were classified by genotype. Using this classification, general linear model ANOVAs were performed for TGW incorporating environment and year as factors for each individual QTL, and for lines with both increasing alleles compared with those with neither. Pearson's correlation coefficient was calculated to assess the correlation between yield and TGW. All analyses performed on DH lines was carried out using Minitab v.17.3.1 (Minitab Inc., Coventry, UK).

The NILs and RILs were evaluated using two-way ANOVAs, with the model including the interaction between environment and the 5A QTL. RIL groups were assigned as having a Charger- or Badger-like grain length phenotype using a *post hoc* Dunnett's test to compare with C- and B-control groups. Similarly, two-way ANOVAs, including genotype and block, were conducted for the developmental time courses and cell size measurements.

Analyses were performed using GenStat, 15th edition (VSN International, Hemel Hemstead, UK) and R v3.2.5.

Results

A QTL on chromosome 5A is associated with increased grain weight

A genetic map was developed for the Charger × Badger DH population comprising 296 polymorphic molecular markers. Linkage analysis resulted in 32 linkage groups that were assigned to 21 chromosomes, covering a genetic distance of 1296 cM. The only chromosome with no marker coverage was 6D.

QTL analysis identified two regions with consistent variation for TGW, chromosomes 2B (*Qtgw-cb.2B*) and 5A (*Qtgw-cb.5A*), based on the mean LOD score across environments (Fig. 1) and co-localisation of significant QTL (Table S2). *Qtgw-cb.2B* was identified in seven of the 12 sites yr^{-1} environments, providing a mean of 11% of the explained variation when significantly expressed and a mean additive effect of 1.26 g per 1000 grains, with Charger providing the increasing allele. The peak LOD for the QTL was located at markers *Xgwm148* and *Xgwm120* depending on the environment. *Qtgw-cb.5A* was also significant at seven of the 12 environments and accounted for 15.5% of the phenotypic variation with a mean additive effect of 1.6 g per 1000 grains. The peak for *Qtgw-cb.5A* was defined by markers *Xgwm293* (20.6 cM) and *Xbarc180* (25.3 cM; Fig. 2a), with Badger providing the increasing allele.

Analysis of DH lines homozygous across the wider QTL regions for both *Qtgw-cb.2B* (*Xgwm259-Xstm119tgag)* and *Qtgw-cb.5A* (*Xgwm443-XBS00000435*) demonstrated that the increasing alleles of each individual QTL provided a significant 4.1% and 5.5% increase in TGW ($P < 0.001$), respectively. DH lines containing both QTL ($n = 9$) produced a 10% increase ($P < 0.001$) over lines with neither ($n = 10$), suggesting *Qtgw-cb.2B* and *Qtgw-cb.5A* are additive when combined.

There was a significant correlation ($P < 0.001$) between grain yield and TGW across all datasets, but significant QTL were only co-located for both traits in France 2003 (2B) and England-Norwich 2002/Scotland 2002(5A) (Table S3). This suggests that although TGW was an important component regulating yield in this DH population, it was also influenced by other yield components. As *Qtgw-cb.5A* had a larger mean additive effect and accounted for more of the phenotypic variation than *Qtgw-cb.2B*, we selected *Qtgw-cb.5A* for further analyses.

MTME analysis defines *Xgwm293* as the peak marker of *Qtgw-cb.5A*

MTME analysis was conducted on chromosome 5A for both TGW and grain yield. For TGW, markers above the significance threshold (LOD > 2.5) ranged from *Xgwm293* (20.6 cM) to

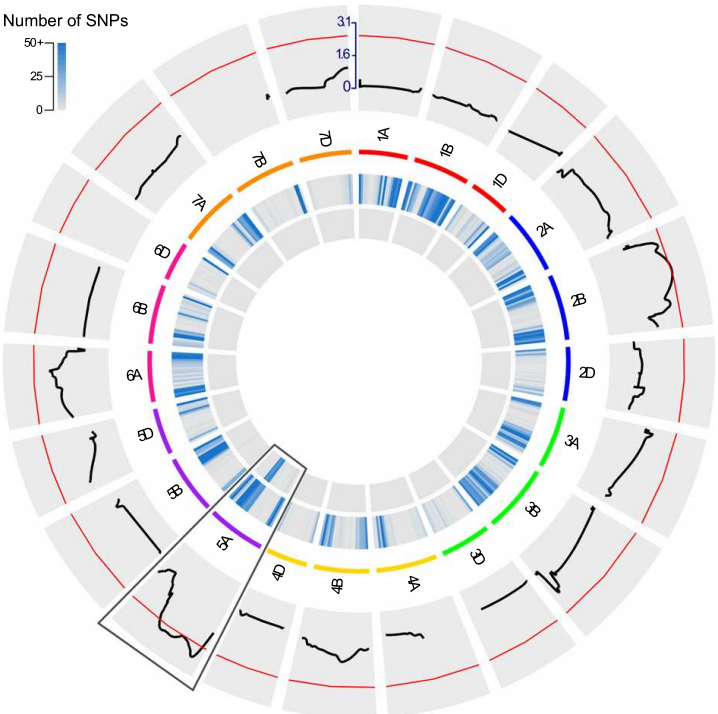


Fig. 1 Quantitative trait loci (QTL) analysis and near-isogenic line (NIL) development. Circos diagram showing the whole genome QTL scan and single nucleotide polymorphism (SNP) variation. Outer track is the mean log-of-odds (LOD) score for thousand grain weight (TGW) across all environments measured. The red line shows an LOD threshold of 2.5. Wheat chromosome groups are represented in different colours beneath the QTL scans. The most significant and stable QTL identified was on chromosome 5A (boxed segment). Inner tracks correspond to heatmaps representing the number of iSelect SNPs in 30 Mb windows showing variation between Charger and Badger, parents of the doubled haploid population (outer) or a representative pair of 5A−/5A+ NILs (innermost). Physical positions of all markers (including those used in the QTL scan and iSelect markers) were determined using the IWGSC RefSeq v.1.0 sequence.

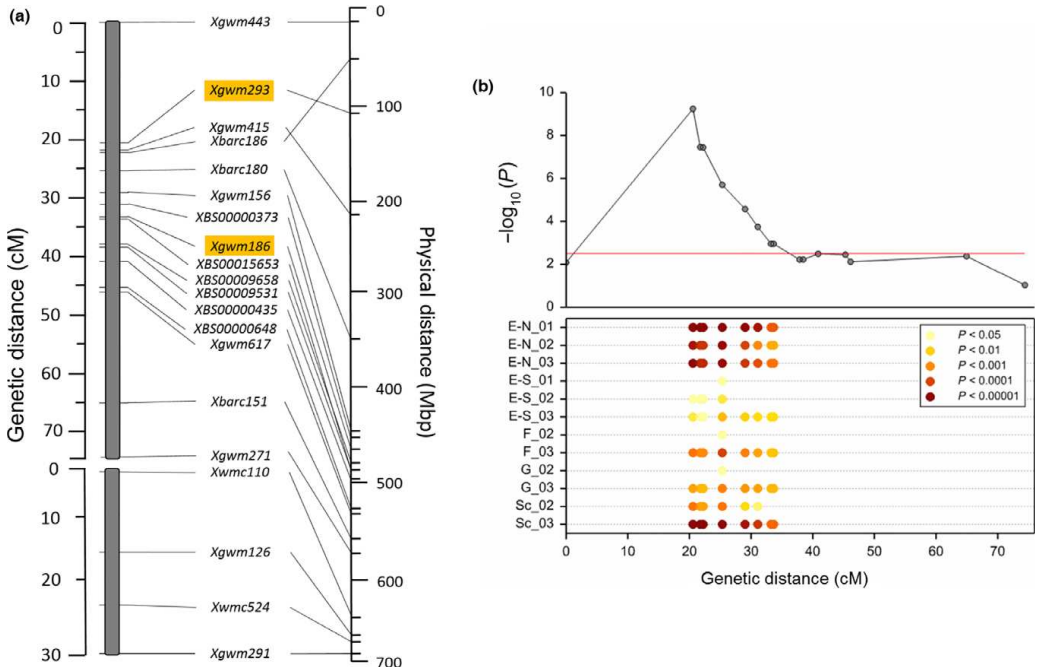


Fig. 2 Chromosome 5A genetic/physical map and thousand grain weight (TGW) multi-trait multi-environment (MTME) analysis. (a) Genetic and physical map of wheat chromosome 5A. The left-hand side represents the genetic map, comprising two linkage groups with calculated distances between markers in centiMorgans (linkage group 1: 0–74.4 cM; linkage group 2: 0–30.2 cM). The right-hand side represents the physical map according to the Chinese Spring IWGSC RefSeq v.1.0 sequence. Markers highlighted in orange indicate those used for near isogenic line (NIL) development. (b) MTME quantitative trait loci (QTL) analysis of the 5A QTL for TGW across Linkage group 1. The red line indicates a log-of-odds (LOD) threshold of 2.5. (c) Markers with significant additive effects are shown for each environment for those markers above the LOD threshold in (b). The intensity of the colour (yellow to brown) indicates the level of the significance as indicated by the legend. E-N, England-Norwich; E-S, England, Sandringham; F, France; G, Germany; Sc, Scotland.

XBS00015653 (33.7 cM) with the peak being at *Xgwm293* (Fig. 2b,c). At least one of the markers within the identified region was significant at each of the 12 environments, with Badger always providing the beneficial alleles. For yield, MTME analysis identified a significant QTL in the *Qtgw-cb.5A* region, with the peak marker (*Xgwm293)* being the same as for TGW. Significant increases in the additive effect of Badger were observed in seven environments (Fig. S2), contrasting with only two in the previous single-environment analysis. It is worth noting that in two environments (England-Sandringham 2001 and 2003), the alternative parent Charger had a borderline significant effect on yield in the MTME analysis. Taken together, these results suggest that the Badger *Qtgw-cb.5A* interval is associated with a consistent effect on TGW across environments which often, but not always, translates into a yield benefit.

NILs differing for *Qtgw-cb.5A* show a 6.9% difference in TGW

To independently validate and further investigate the effect of *Qtgw-cb.5A* (hereafter 5A QTL) on TGW, BC_2 and BC_4 NILs differing for the QTL region were developed using markers *Xgwm293* and *Xgwm186* and Charger as the recurrent parent. Pairs of BC_4 NILs carrying the Charger (5A−) or Badger (5A+) segment were genotyped using the iSelect 90K SNP array (Wang *et al.*, 2014) and found to be 97.2% similar, only showing variation in 221 markers across the 5A QTL, compared with 7973 SNPs between the parents (Fig. 1, inner tracks). These NILs therefore provide a valuable resource for specifically studying the effects of the 5A QTL in more depth.

Across 5 yr of replicated field trials, 5A+ NILs showed an average increase in TGW of 6.92% ($P < 0.001$) ranging from 4.00 to 9.28% (Table 1), and significant in all years. The difference in TGW was associated with a yield increase of 1.28% in 5A+ NILs across all years, although this effect was not significant ($P = 0.093$). The effect varied across years with a significant yield increase of 2.17% ($P = 0.046$) in 2014 and nonsignificant effects of 0.02–1.72% in the other four years. The positive effect of the QTL on yield was similarly subtle in the DH population as described previously.

We measured the NILs for a series of spike yield component traits to determine possible pleiotropic effects associated with the 5A+ TGW effect. Within most years, there was no significant effect of the 5A+ allele on spike yield components such as spikelet

Table 1 Mean thousand grain weight (TGW), yield and grain morphometric parameters of 5A near isogenic lines (NILs)

Year	Genotype	TGW (g)	Yield (kg per plot)	Grain area (mm^2)	Grain length (mm)	Grain width (mm)
2012	5A−	38.027	4.408	18.755	6.625	3.475
	5A+	41.554	4.437	19.930	6.900	3.557
		9.28%***	0.66%^{ns}	6.26%***	4.15%***	2.35%**
2013	5A−	40.772	6.157	19.969	6.705	3.674
	5A+	43.544	6.159	20.979	6.963	3.727
		6.80%***	0.02%^{ns}	5.06%***	3.86%***	1.44%***
2014	5A−	47.368	6.495	21.493	6.798	3.930
	5A+	50.729	6.636	22.579	7.063	3.979
		7.09%***	2.17%*	5.05%***	3.90%***	1.25%**
2015	5A−	42.734	7.582	18.044	6.426	3.479
	5A+	46.201	7.712	19.293	6.730	3.554
		8.11%***	1.72%^{ns}	6.93%***	4.72%***	2.16%***
2016	5A−	49.292	5.974	19.829	6.580	3.735
	5A+	51.266	6.064	20.610	6.816	3.745
		4.00%*	1.50%^{ns}	3.94%**	3.58%***	0.27%^{ns}
Overall	5A−	43.639	6.123	19.618	6.627	3.659
	5A+	46.659	6.201	20.678	6.894	3.712
		6.92%***	1.28%^{ns}	5.41%***	4.04%***	1.45%***,[1]

[1]Percentages (%) indicate amount gained in 5A+ NILs compared with 5A− NILs. Asterisks indicate significance determined by ANOVA for either each year, or across all years (final row). ns, nonsignificant; *, $P<0.05$; **, $P<0.01$;***, $P<0.001$; 2012–2013, BC_2-NILs; 2014–2016, BC_4-NILs.

number, seeds per spikelet or grain number per spike (Table S4). However, when all years were analysed together, we observed a significant reduction in grain number (−3.55%, $P=0.04$) and seeds per spikelet (−3.37%, $P=0.015$) associated with the 5A+ interval. This statistical significance was driven by a particularly strong negative effect in 2016 as grain number and seeds per spikelet were nonsignificant in the preceding four seasons (2012–2015). Overall, however, the 5A+ interval is associated with a consistent small decrease in these spike yield components.

Taking into account the 6.92% effect of the 5A+ QTL on TGW and the tendency for decreases in some spike yield components, the overall spike yield increased by 2.33% ($P=0.032$) across the five years. However, similar to grain number and seeds per spikelet, the statistical significance is driven by a single year (2014) despite overall positive effects in another three years (2012, 2013, and 2015). We also measured tiller numbers and found a significant reduction of 4 tillers m^{-1} in the 5A+ NILs across 2 yr ($P=0.008$) (Table S1). No effect was seen for spikelet number and additional phenology traits (Table S4). Taken together, these results suggest that the 5A+ interval has a consistent positive effect on TGW and that the effects on yield are modulated by a series of smaller compensating negative effects on yield components such as grain number, seeds per spike and tiller number.

The TGW increase in 5A+ NILs is primarily due to increased grain length

TGW is determined by individual components including physical parameters such as grain length and width. To understand the relative contribution of these components to the increase in TGW, NILs were assessed for these grain morphometric parameters (length, width and area) using a two-dimensional imaging system (Table 1). 5A+ NILs had significantly increased grain length ($P<0.001$), width ($P<0.001$) and area ($P<0.001$) compared with 5A− NILs across all years with the exception of width in 2016. On average, the 5A+ QTL increased grain length by 4.04% ($P<0.001$), ranging from 3.58 to 4.72% ($P<0.001$ in all years). The effect on width was smaller, averaging 1.45% ($P<0.001$; range 0.27–2.35%) and significant in four out of five years (Table 1). The effects on length and width combined to increase grain area by an average of 5.41% ($P<0.001$), significant in all five years. These results were based on combine harvested grain samples and were also confirmed in 10 representative single ear samples taken before harvest. TGW of the 10 spikes correlated strongly with the whole plot samples ($r=0.84$, $P<0.001$) and showed a similar difference between NILs (6.00%, $P<0.001$; Table S4). Across datasets, the effect of the 5A+ QTL on grain length was more than twice the size of the effect on grain width. This fact, together with the more consistent effect on grain length across years (coefficient of variation length = 10.6%; width = 55.3%; TGW = 27.8%; Table S5) suggests that the increase in grain length is the main factor driving the increase in grain area and TGW.

We compared the distribution of grain length and width using data from individual seeds to determine whether the QTL affects all grains uniformly. Violin plots for length showed variation in distribution shape among years (Fig. 3). However, within years the 5A− and 5A+ grain length distributions were very similar in shape, suggesting that the QTL affects all grains uniformly and in a stable manner across the ear and within spikelets. In all years, the 5A+ grain length distributions were shifted higher than the 5A− NILs with an increase in longer grains and fewer shorter grains, in addition to the higher average grain length (Fig. 3). Grain width distributions were also very similar in shape within years, but had a less pronounced shift between NILs (Fig. S3), consistent with the overall smaller effect of the 5A QTL on grain width.

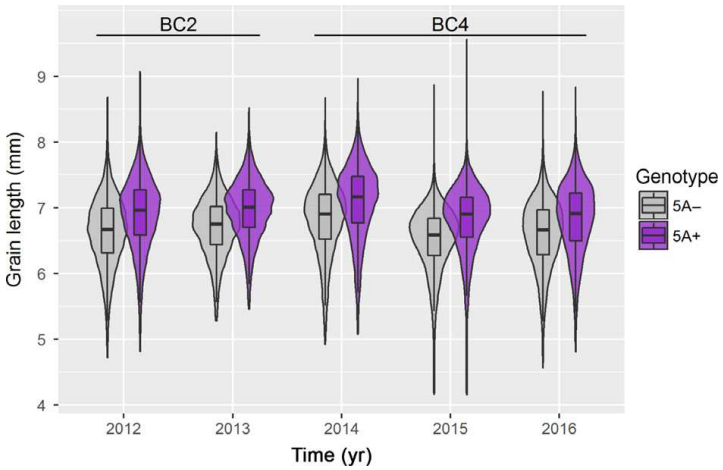


Fig. 3 Distribution of grain length of wheat near isogenic lines (NILs) from whole plot samples. Violin plots showing the distribution of individual seed measurements of grain length across the five field experiments of BC_2 (2012–2013) and BC_4 (2014–2016) NILs. Purple, 5A+ NILs; grey plots, 5A− NILs. All within-year comparisons between NILs were significant ($P < 0.001$) .

The 5A QTL region acts during grain development to increase grain length

To determine when differences in grain morphometric parameters between NILs are first established, we conducted grain development time courses of two 5A− and two 5A+ BC_4 NILs. Grains were sampled in 2014, 2015 and 2016 from field plots at anthesis and at five further time points across grain development until the difference in grain size had been fully established. Data from 2015 are shown in Fig. 4 as a representative year (samples taken at anthesis (0 dpa), and at 4, 7, 12, 19 and 26 dpa). The first significant difference in grain length was observed at 12 dpa with 5A+ NILs having 1.5% longer grains than 5A− NILs ($P = 0.034$). This effect increased to 4.4% at 19 dpa ($P < 0.001$) and was maintained at 26 dpa (4.5% increase, $P < 0.001$; Fig. 4a). No significant effects on grain width were observed until 26 dpa when 5A+ NILs increased grain width by 1.7% ($P = 0.015$; Fig. 4b). Significant differences in grain area were detected at 19 dpa (5.7% increase; $P < 0.001$; data not shown) and this difference was maintained at the final time point, 26 dpa (6.1%, $P < 0.001$). By the final time point, 5A+ NILs also had significantly heavier grains (3.7%, $P = 0.01$; Fig. 4c). These effects were all consistent with the grain size and weight differences observed in mature grains in 2015 (Table 1) and were also observed in 2014 and 2016 (Figs S4, S5). The fact that the effects on width, area and weight are all after the first significant difference on grain length in all three years further supports grain length as the main factor driving the increase in grain weight.

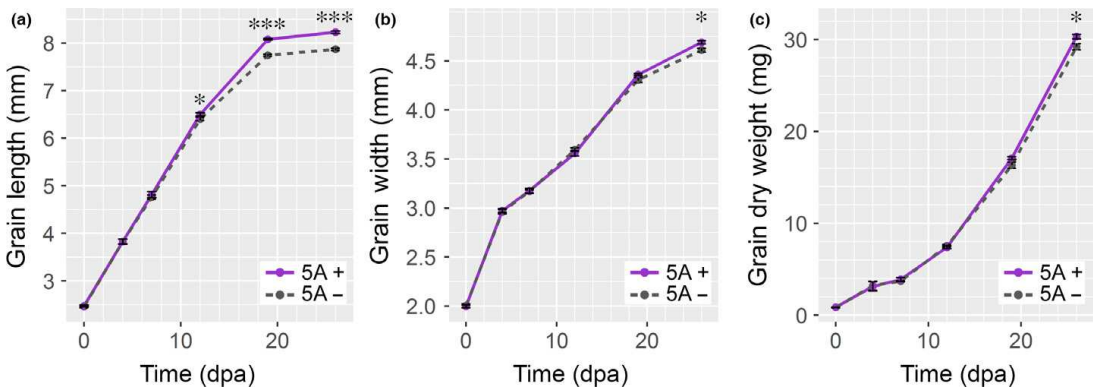


Fig. 4 Grain development time course of 5A− and 5A+ near isogenic lines (NILs). (a) Grain length, (b) grain width and (c) grain dry weight of 5A− (grey, dashed line) and 5A+ (purple, solid line) BC_4 wheat NILs during grain development with samples taken at anthesis (0 d post-anthesis, dpa), and 4, 7, 12, 19 and 26 dpa in 2015 field trials. *, $P < 0.05$; ***, $P < 0.001$. Error bars show ± SEM.

5A+ NILs have increased pericarp cell length independent of absolute grain length

We used scanning electron microscopy to image pericarp cells and determine cell size of BC_4 5A− and 5A+ grains. Mature grains from the 2015 field experiment were selected from a 5A− and 5A+ NIL pair based on their grain length and using a variety of criteria to allow for distinct comparisons (Fig. 5). First, we compared grains of average length from the 5A− and 5A+ NIL distributions (Fig. 5a). We found that average 5A+ grains had an 8.33% significant increase in mean cell length ($P = 0.049$) compared with average 5A− grains and that this was reflected in a shift in the whole distribution of 5A+ cell lengths (Fig. 5a). Next, we compared cell lengths in grains of the same size from 5A− and 5A+ NILs. We selected relatively long grains from the 5A− NIL distribution (Fig. 5b; orange) that had the same grain length as the average 5A+ grains. This comparison showed that 5A+ grains still had longer cells (9.53%, $P = 0.015$) regardless of the fact that the grain length of the two groups was the same (6.8 mm; Fig. 5b). We also made the opposite comparison by

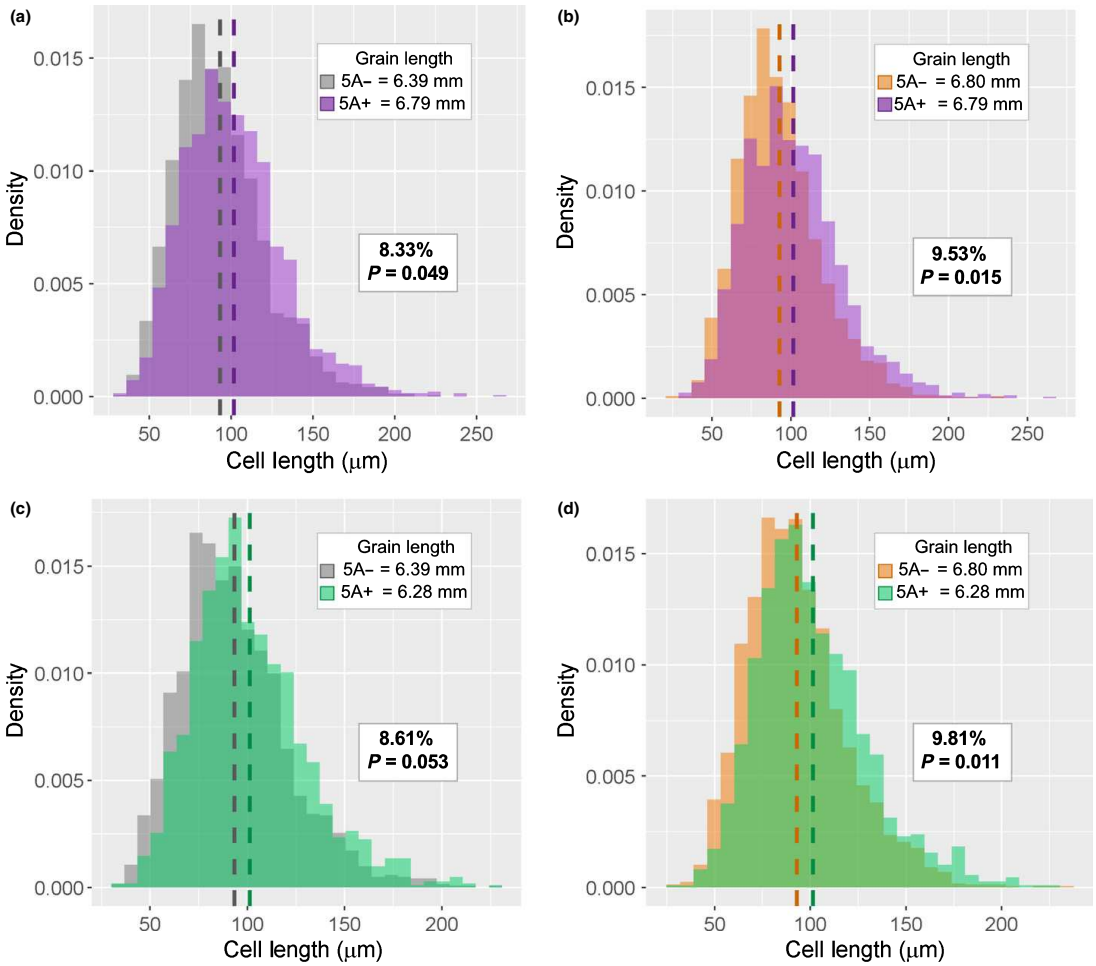


Fig. 5 Comparisons of cell length between 5A− and 5A+ near isogenic lines (NILs). Density plots of cell length measurements from 27 grains per genotype group; dashed line represents the mean. 'Grain length' insets show the average grain length of each group of grains used for measurements. The increase in cell length of 5A+ wheat NILs relative to cell length of 5A− grains is shown as a percentage along with the *P*-values calculated using ANOVA to compare means of the two groups displayed. (a) Wheat grains of average length from 5A− and 5A+ NILs, (b) average 5A+ grains and equivalent 5A− grains, (c) average 5A− grains and equivalent 5A+ grains, (d) long 5A− grains (length equivalent to average 5A+ grains) and short 5A+ grains (grain length equivalent to average 5A− grains).

selecting relatively short grains from the 5A+ NIL distribution (Fig. 5c; green) and comparing them with average 5A− grains. Similar to before, the 5A+ grains had longer cells (8.61%), although this effect was borderline nonsignificant ($P = 0.053$; Fig. 5c). Finally, a comparison of long 5A− grains and short 5A+ grains again showed that cells were longer in 5A+ grains (9.81%, $P = 0.011$), even though the 5A+ grains used in this comparison were 7.65% shorter than the 5A− grains. Within-genotype comparisons of cell length between grains of different lengths showed no significant differences in mean cell length (Fig. S6). The results were confirmed in 2016 where average 5A+ grains had a 24.6% significant increase in mean cell length compared with average 5A− grains ($P < 0.001$; Fig. S7). These results indicate that the 5A+ region from Badger increases the length of pericarp cells independent of absolute grain length. Using grain length and mean cell length to calculate cell number, we determined that the average length grains of both 5A− and 5A+ had the same number of cells in 2015. However, in 2016, 5A− NILs had significantly more cells than 5A+ NILs (Fig. S8).

The grain length QTL maps to a 75 Mb/4.3 cM genetic interval

We used a set of 60 homozygous RILs to map the grain length phenotype to a narrower genetic interval within the 5A QTL region (17.65 cM, 367 Mbp). KASP markers were developed for 25 additional SNPs between the two original QTL flanking markers (*Xgwm293* and *Xgwm186*; Fig. 6a) based on data from the iSelect genotyping of BC_4 NILs and 820K Axiom Array genotyping of Charger and Badger (Winfield *et al.*, 2016). Based on the genotype of these 25 markers, 49 of the RILs were assigned to 11 distinct recombination groups represented as graphical genotypes in Fig. 6(a). Control RILs were selected based on having either the Charger (5A−) or Badger (5A+) genotypes across the interval (C-control and B-control, respectively).

RILs were phenotyped for grain length in three field seasons and we found significant differences between RIL groups ($P < 0.001$). The overall average grain length of the B-control group was 4.06% higher than the C-control group ($P < 0.001$; Fig. 6b), consistent with the differences in grain length observed between the NILs (Table 1). Each RIL group was classified based on Dunnett's tests to both control groups: for example, an RIL group was classified as Charger-like only if it was both significantly different from the B-control *and* nonsignificantly different from the C-control. Using this classification, we assigned unambiguously the 11 RIL groups to a parental type and genetically mapped the grain length phenotype between markers *XBS00182017* and *XBA00228977* (Fig. 6). This represents a genetic distance of 4.32 cM corresponding to a physical interval of 74.6 Mb in the Chinese Spring RefSeq v.1.0 sequence.

This 74.6 Mb interval contains 811 TGACv1 gene models (Clavijo *et al.*, 2017) based on *in silico* mapping to the Chinese Spring reference (Notes S1). We analysed the expression profile of these genes on the wheat expVIP expression platform (Borrill *et al.*, 2016) and found that 439 of these genes are expressed (> 2 transcripts per million) in at least one grain RNA-seq sample

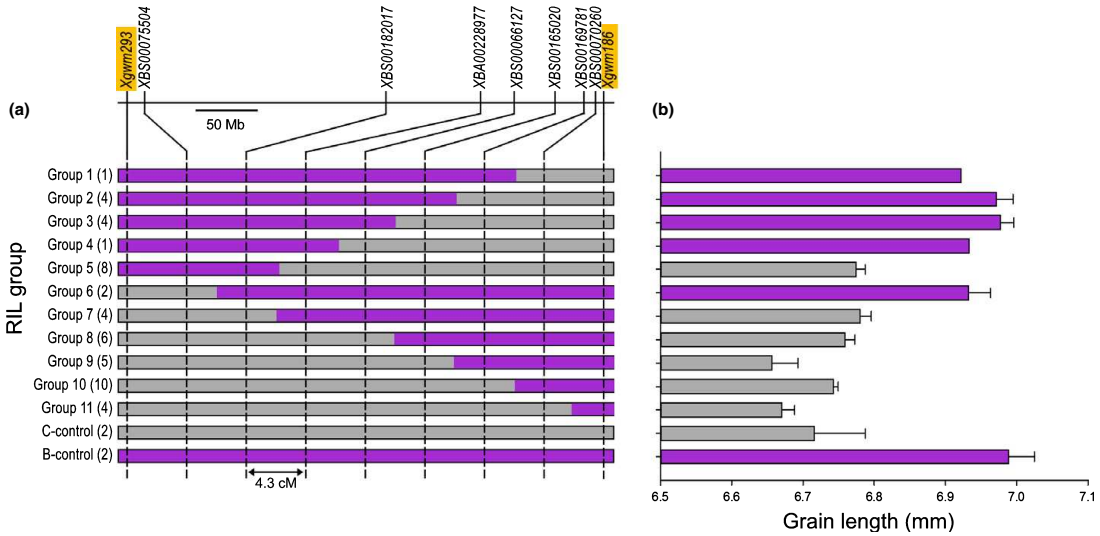


Fig. 6 Grain length maps to a 4.3 cM interval on wheat chromosome 5A. (a) Graphical genotypes of recombinant inbred line (RIL) groups with the number of lines in each group shown in parentheses. RILs were grouped based on their genotypes defined by having either the Charger-like (grey) or the Badger-like (purple) allele at each marker shown across the interval. Markers highlighted in orange indicate markers used for near isogenic line (NIL) development. (b) ANOVA adjusted mean grain length of RIL groups across all experiments. Bars are coloured based on a Charger- or Badger-like phenotype, determined by Dunnett's test. Purple, Badger-like; grey, Charger-like. Error bars represent SEM.

($n = 147$). The developmental time courses suggest that the 5A QTL acts at *c.* 12 dpa and we found 405 of these transcripts expressed in grain samples taken at around this time (4–15 dpa, $n = 59$), with 298 genes expressed in the pericarp tissue (Pearce *et al.*, 2015a) (Notes S1; Fig. S9).

Discussion

In this study we identified a stable and robust QTL associated with a 6.9% increase in grain weight. This increase is driven by longer grains associated with increased pericarp cell length. In wheat and barley pericarp cell division decreases shortly after fertilization (2–6 d; Drea *et al.*, 2005; Radchuk *et al.*, 2011) and cell expansion plays the predominant role in increasing pericarp size during grain development. Our results are consistent with a role of the 5A gene on pericarp cell expansion given that significant differences in grain size are only observed 12 d after fertilization, once cell expansion has begun. However, we cannot discard an overlapping late effect on cell division given the conflicting results in final pericarp cell number between years.

Overall, our results suggest that the gene underlying this locus regulates, either directly or indirectly, cell expansion in the pericarp (seed coat), a mechanism that is known to be a key determinant of grain/seed size in several species. Some genes, such as expansins and XTH (xyloglucan endotransglucosylase/hydrolases), affect cell expansion directly by physically modifying or 'loosening' the cell wall (reviewed by Cosgrove, 2005), and the expression of these enzymes has been associated with pericarp cell expansion in wheat and barley (Lizana *et al.*, 2010; Radchuk *et al.*, 2011; Munoz & Calderini, 2015). Other genes regulate pericarp/seed coat cell size indirectly, for example through the regulation of sugar metabolism and subsequent accumulation in the vacuole (Ohto *et al.*, 2005) and endoreduplication (Chevalier *et al.*, 2014). Our results provide direct genetic evidence that pericarp cell expansion affects final grain size and weight in polyploid wheat.

The maternal control of grain/seed size has been well documented in rice and Arabidopsis (Li & Li, 2015), as well as in wheat through physiological and genetic studies (Hasan *et al.*, 2011; Simmonds *et al.*, 2016). This can affect cell proliferation and/or cell expansion of maternal tissues, such as the wheat pericarp, both before and after fertilisation (Garcia *et al.*, 2005; Adamski *et al.*, 2009; Ma *et al.*, 2016). For example, *GW2* in rice and its orthologue in Arabidopsis (*DA2*) affect grain/seed size through suppression of cell proliferation (Song *et al.*, 2007; Xia *et al.*, 2013). Similarly in wheat, a knock-out mutant of the *GW2* orthologue has larger carpels than wild-type plants, suggesting that the gene acts on maternal tissue before fertilisation (Simmonds *et al.*, 2016). The effect of the wheat *GW2* gene on cell size and number has not been determined however.

The direct assignment of the 5A effect to the maternal parent will require additional studies, including analysis of F_1 hybrids from reciprocal crosses. These studies are not routinely performed in wheat given that the phenotypic variation between individual F_1 grains often surpasses the relatively subtle phenotypic effects of most grain size QTL (usually < 5% in wheat). The identification of a robust effect on pericarp cell length in this study, which is independent of the individual grain size, opens up a new approach to explore these parent-of-origin effects in polyploid wheat.

It has been proposed, in multiple species, that the size of the pericarp/seed coat determines final grain size by restricting endosperm growth (Calderini *et al.*, 1999; Adamski *et al.*, 2009; Hasan *et al.*, 2011). This is analogous to the way in which grain size in rice is limited by the size of the spikelet hull (Song *et al.*, 2005). Both the length (Lizana *et al.*, 2010; Hasan *et al.*, 2011) and the width (Gegas *et al.*, 2010; Simmonds *et al.*, 2016) of the pericarp have been proposed as key determinants of final grain weight in wheat. Our results provide genetic evidence for the importance of the maternal pericarp tissue and show that length is the underlying component for the 5A locus. Across three years, the difference in grain length between NILs was the first grain size component difference to be established. Only after this did we observe any differences in grain width, weight or grain filling rate. These differences in grain length were extremely consistent across years (despite average TGW values ranging from 39.8 to 50.3 g) compared with the more variable differences in grain width and weight. Based on these results we hypothesise that the 5A locus increases grain weight by a primary effect on grain length, which confers the potential for further enhancements by pleiotropic effects on grain width. The grain length effect is genetically controlled and stable across environments, whereas the pleiotropic effect on grain width occurs later in grain development and is more environmentally dependent and variable. The final magnitude of the 5A grain weight increase (ranging from 4.0 to 9.3%) is thus determined by the extent to which the late-stage pleiotropic effect on grain width is manifested and the potential exploited. This could explain why the grain width increase was significantly correlated with the increase in TGW ($r = 0.98$, $P = 0.004$) whilst grain length was not ($r = 0.71$, $P = 0.18$).

By dissecting TGW to a more stable yield component (grain length) we were able to classify RILs in a qualitative/binary manner (i.e. 'short' or 'long' grains) which enabled the fine mapping of the 5A locus to a genetic distance of 4.3 cM. We identified *c.* 400 genes in this interval that are expressed in the grain, several of which have annotations associated with genes implicated in the control of grain/seed size. Although it is premature to speculate on potential candidate genes, identification of the causal polymorphism will provide functional insight into the specific mechanism by which pericarp cell size and grain weight are controlled in polyploid wheat.

The consistent effect of the 5A locus on grain length and weight did not always translate into increased yield. In the original DH analysis, the 5A TGW effect co-located with final yield in seven of the 12 environments. This overall positive trend was also reflected in the NILs, although yield increases were only significant in 2014. We concluded that the effects on yield are modulated by a series of smaller negative effects on other yield components which could be due to additional genes within the broader 5A region. Alternatively, it could be that the full

potential of the grain length effect will be realised only under certain environments or in combination with other genes.

By understanding the biological mechanism by which the 5A locus achieves increased grain size, hypotheses can be generated to combine genes in an informed and targeted way. For example, we are combining the 5A grain length/pericarp cell expansion effect with the *TaGW2* mutants which affect grain width (presumably through pericarp cell proliferation) to determine if they act in an additive or synergistic manner. Identifying the 5A gene will also allow the function of the homoeologous copies on chromosomes 5B and 5D to be determined. This is important because the effects of grain weight QTL in polyploid wheat are often very subtle compared with those in diploid species (Borrill *et al.*, 2015; Uauy, 2017). Modulating the function of all three homoeologues simultaneously holds the potential to expand the range of phenotypic variation and achieve effects comparable to those in diploids, for example *NAM-B1* (Uauy *et al.*, 2006; Avni *et al.*, 2014; Liang *et al.*, 2014). Ultimately, identifying the genes and alleles that control specific yield components and understanding how they interact amongst them and with the environment will allow breeders to manipulate and fine-tune wheat yield in novel ways.

Acknowledgements

This work was supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) grants BB/J003557/1, BB/J004588/1, BB/P013511/1 and BB/P016855/1. J.B. was supported by the UK Agriculture and Horticulture Development Board (AHDB) and the John Innes Foundation. We thank Ricardo Ramirez-Gonzalez for the BLAST of iSelect SNPs to the IWGSC REFSEQ v.1.0, David Swarbreck and Gemy Kaithakottil (Earlham Institute) for *in silico* mapping of TGACv1 gene models to the IWGSC for pre-publication access to REFSEQ v.1.0, the JIC Field Trials Horticultural services for technical support in glasshouse and field experiments.

Author contributions

J.B. conducted the developmental time courses, fine-mapping, analysed the data and wrote the manuscript. J. Simmonds developed the germplasm used in this study, performed phenotypic assessments and QTL analyses, analysed the data and wrote the manuscript. F.M. conducted cell size measurements. M.L-W. led the mapping of the Charger × Badger DH population. J. Snape coordinated and conceived the DH population field trials. C.U. conceived the experiments, analysed the data and wrote the manuscript. All authors read and approved the final manuscript.

References

Adamski NM, Anastasiou E, Eriksson S, O'Neill CM, Lenhard M. 2009. Local maternal control of seed size by *KLUH/CYP78A5*-dependent growth signaling. *Proceedings of the National Academy of Sciences, USA* **106**: 20115–20120.

Allen AM, Barker GLA, Berry ST, Coghill JA, Gwilliam R, Kirby S, Robinson P, Brenchley RC, D'Amore R, McKenzie N *et al.* 2011. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal* **9**: 1086–1099.

Avni R, Zhao RR, Pearce S, Jun Y, Uauy C, Tabbita F, Fahima T, Slade A, Dubcovsky J, Distelfeld A. 2014. Functional characterization of *GPC-1* genes in hexaploid wheat. *Planta* **239**: 313–324.

Borrill P, Adamski N, Uauy C. 2015. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist* **208**: 1008–1022.

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

Breseghello F, Sorrells ME. 2007. QTL analysis of kernel size and shape in two hexaploid wheat mapping populations. *Field Crops Research* **101**: 172–179.

Bryan GJ, Collins AJ, Stephenson P, Orry A, Smith JB, Gale MD. 1997. Isolation and characterisation of microsatellites from hexaploid bread wheat. *Theoretical and Applied Genetics* **94**: 557–563.

Calderini D, Abeledo L, Savin R, Slafer GA. 1999. Effect of temperature and carpel size during pre-anthesis on potential grain weight in wheat. *Journal of Agricultural Science* **132**: 453–459.

Chevalier C, Bourdon M, Pirrello J, Cheniclet C, Gévaudant F, Frangne N. 2014. Endoreduplication and fruit growth in tomato: evidence in favour of the karyoplasmic ratio theory. *Journal of Experimental Botany* **65**: 2731–2746.

Clavijo BJ, Venturini L, Schudoma C, Garcia Accinelli G, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H *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**: 885–896.

Cosgrove DJ. 2005. Growth of the plant cell wall. *Nature Reviews: Molecular Cell Biology* **6**: 850–861.

Drea S, Leader DJ, Arnold BC, Shaw P, Dolan L, Doonan JH. 2005. Systematic spatial analysis of gene expression during wheat caryopsis development. *Plant Cell* **17**: 2172–2185.

Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, Li X, Zhang Q. 2006. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics* **112**: 1164–1171.

FAO, IFAD, WFP. 2015. *The state of food insecurity in the World 2015.* Meeting the 2015 international hunger targets: taking stock of uneven progress. Rome, Italy: FAO.

FAO. 2017. *Online statistical database: food balance.* FAOSTAT [WWW document] URL http://www.fao.org/faostat/en/ [accessed 27 February 2017].

Farre A, Sayers L, Leverington-Waite M, Goram R, Orford S, Wingen L, Mumford C, Griffiths S. 2016. Application of a library of near isogenic lines to understand context dependent expression of QTL for grain yield and adaptive traits in bread wheat. *BMC Plant Biology* **16**: 161.

Garcia D, Gerald JNF, Berger F. 2005. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in Arabidopsis. *Plant Cell* **17**: 52–60.

Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S, Sayers L, Doonan JH, Snape JW. 2010. A genetic framework for grain size and shape variation in wheat. *Plant Cell* **22**: 1046–1056.

Guyomarc'h H, Sourdille P, Charmet G, Edwards KJ, Bernard M. 2002. Characterisation of polymorphic microsatellite markers from *Aegilops tauschii* and transferability to the D-genome of bread wheat. *Theoretical and Applied Genetics* **104**: 1164–1172.

Hasan AK, Herrera J, Lizana C, Calderini DF. 2011. Carpel weight, grain length and stabilized grain water content are physiological drivers of grain weight determination of wheat. *Field Crops Research* **123**: 241–247.

Hayden MJ, Sharp PJ. 2001. Sequence-tagged microsatellite profiling (STMP): a rapid technique for developing SSR markers. *Nucleic Acids Research* **29**: e43.

Huang X, Qian Q, Liu Z, Sun H, He S, Luo D, Xia G, Chu C, Li J, Fu X. 2009. Natural variation at the *DEP1* locus enhances grain yield in rice. *Nature Genetics* **41**: 494–497.

IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**: 1251788.

Jiang W-B, Huang H-Y, Hu Y-W, Zhu S-W, Wang Z-Y, Lin W-H. 2013. Brassinosteroid regulates seed size and shape in Arabidopsis. *Plant Physiology* **162**: 1965–1977.

Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, Akhunov E, Uauy C *et al.* 2013. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biology* **14**: R66.

Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang X, Borrill P *et al.* 2017. Uncovering hidden variation in polyploid wheat. *Proceedings of the National Academy of Sciences, USA* **114**: E913–E921.

Kuchel H, Williams KJ, Langridge P, Eagles HA, Jefferies SP. 2007. Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theoretical and Applied Genetics* **115**: 1029–1041.

Kumar A, Mantovani EE, Seetan R, Soltani A, Echeverry-Solarte M, Jain S, Simsek S, Doehlert D, Alamri MS, Elias EM *et al.* 2016. Dissection of genetic factors underlying wheat kernel shape and size in an elite × nonadapted cross using a high density SNP linkage map. *Plant Genome* **9**: 1.

Laurie DA, Bennett MD. 1988. The production of haploid wheat plants from wheat × maize crosses. *Theoretical and Applied Genetics* **76**: 393–397.

Li N, Li Y. 2015. Maternal control of seed size in plants. *Journal of Experimental Botany* **66**: 1087–1097.

Liang C, Wang Y, Zhu Y, Tang J, Hu B, Liu L, Ou S, Wu H, Sun X, Chu J *et al.* 2014. *OsNAP* connects abscisic acid and leaf senescence by fine-tuning abscisic acid biosynthesis and directly targeting senescence-associated genes in rice. *Proceedings of the National Academy of Sciences, USA* **111**: 10013–10018.

Lizana XC, Riegel R, Gomez LD, Herrera J, Isla A, McQueen-Mason SJ, Calderini DF. 2010. Expansins expression is associated with grain size dynamics in wheat (*Triticum aestivum* L.). *Journal of Experimental Botany* **61**: 1147–1157.

Ma M, Zhao H, Li Z, Hu S, Song W, Liu X. 2016. *TaCYP78A5* regulates seed size in wheat (*Triticum aestivum*). *Journal of Experimental Botany* **67**: 1397–1410.

Mizukami Y, Fischer RL. 2000. Plant organ size control: *AINTEGUMENTA* regulates growth and cell numbers during organogenesis. *Proceedings of the National Academy of Sciences, USA* **97**: 942–947.

Munoz M, Calderini DF. 2015. Volume, water content, epidermal cell area, and *XTH5* expression in growing grains of wheat across ploidy levels. *Field Crops Research* **173**: 30–40.

Ohto M-a, Fischer RL, Goldberg RB, Nakamura K, Harada JJ. 2005. Control of seed mass by *APETALA2*. *Proceedings of the National Academy of Sciences, USA* **102**: 3123–3128.

Pearce S, Huttly AK, Prosser IM, Li Y-d, Vaughan SP, Gallova B, Patil A, Coghill JA, Dubcovsky J, Hedden P *et al.* 2015a. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the *GA3ox* family. *BMC Plant Biology* **15**: 130.

Pearce S, Vazquez-Gross H, Herin SY, Hane D, Wang Y, Gu YQ, Dubcovsky J. 2015b. WheatExp: an RNA-seq expression database for polyploid wheat. *BMC Plant Biology* **15**: 299.

Pestsova E, Ganal MW, Roder MS. 2000. Isolation and mapping of microsatellite markers specific for the D genome of bread wheat. *Genome* **43**: 689–697.

Radchuk V, Weier D, Radchuk R, Weschke W, Weber H. 2011. Development of maternal seed tissue in barley is mediated by regulated cell expansion and cell disintegration and coordinated with endosperm growth. *Journal of Experimental Botany* **62**: 1217–1227.

Ray DK, Mueller ND, West PC, Foley JA. 2013. Yield trends are insufficient to double global crop production by 2050. *PLoS ONE* **8**: e66428.

Riefler M, Novak O, Strnad M, Schmülling T. 2006. Arabidopsis cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *Plant Cell* **18**: 40–54.

Roder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW. 1998. A microsatellite map of wheat. *Genetics* **149**: 2007–2023.

Rosegrant MW, Tokgoz S, Bhandary P, Msangi S. 2013. Looking ahead: scenarios for the future of food. In: *2012 Global Food Policy Report*. Washington, DC, USA: International Food Policy Research Institute, 88–101.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B *et al.* 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**: 676–682.

Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, Scott RJ. 2006. The *AUXIN RESPONSE FACTOR 2* gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs. *Development* **133**: 251–261.

Simmonds J, Scott P, Brinton J, Mestre TC, Bush M, Blanco A, Dubcovsky J, Uauy C. 2016. A splice acceptor site mutation in *TaGW2-A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theoretical and Applied Genetics* **129**: 1099–1112.

Simmonds J, Scott P, Leverington-Waite M, Turner AS, Brinton J, Korzun V, Snape J, Uauy C. 2014. Identification and independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of hexaploid wheat (*Triticum aestivum* L.). *BMC Plant Biology* **14**: 191.

Somers DJ, Isaac P, Edwards K. 2004. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **109**: 1105–1114.

Song X-J, Huang W, Shi M, Zhu M-Z, Lin H-X. 2007. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genetics* **39**: 623–630.

Song QJ, Shi JR, Singh S, Fickus EW, Costa JM, Lewis J, Gill BS, Ward R, Cregan PB. 2005. Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics* **110**: 550–560.

Stephenson P, Bryan G, Kirby J, Collins A, Devos K, Busso C, Gale M. 1998. Fifty new microsatellite loci for the wheat genetic map. *Theoretical and Applied Genetics* **97**: 946–949.

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences, USA* **108**: 20260–20264.

Uauy C. 2017. Wheat genomics comes of age. *Current Opinion in Plant Biology* **36**: 142–148.

Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J. 2006. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**: 1298–1301.

United Nations Department of Economic and Social Affairs, Population Division. (2015. *World population prospects: the 2015 revision.* Working Paper no. ESA/P/WP.241. [WWW document] URL https://esa.un.org/unpd/wpp/ [accessed 27 February 2017].

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L *et al.* 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal* **12**: 787–796.

Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, Zeng R, Zhu H, Dong G, Qian Q *et al.* 2012. Control of grain size, shape and quality by *OsSPL16* in rice. *Nature Genetics* **44**: 950–954.

Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X *et al.* 2008. Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. *Cell Research* **18**: 1199–1209.

Winfield MO, Allen AM, Burridge AJ, Barker GLA, Benbow HR, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G *et al.* 2016. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal* **14**: 1195–1206.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**: 1859–1875.

Xia T, Li N, Dumenil J, Li J, Kamenski A, Bevan MW, Gao F, Li YH. 2013. The ubiquitin receptor DA1 interacts with the E3 ubiquitin ligase DA2 to regulate seed and organ size in Arabidopsis. *Plant Cell* **25**: 3347–3359.

Xing Y, Zhang Q. 2010. Genetic and molecular bases of rice yield. *Annual Review of Plant Biology* **61**: 421–442.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Example scanning electron microscope image for pericarp cell size measurements.

Fig. S2 Yield MTME analysis.

Fig. S3 Distribution of grain width of NILs from whole plot samples.

Fig. S4 Grain development time course of 5A− and 5A+ NILs (2014 field trials).

Fig. S5 Grain development time course of 5A− and 5A+ NILs (2016 field trials).

Fig. S6 Comparisons of cell length within genotypes between different sized groups of grains.

Fig. S7 Comparisons of cell length between 5A− and 5A+ grains in 2016.

Fig. S8 Comparison of cell number in the pericarp in 2015 and 2016 samples.

Fig. S9 Expression analysis of genes in the fine mapped interval on 5A between markers *XBS00182017* and *XBA00228977*.

Table S1 Developmental traits of 5A− and 5A+ NILs

Table S2 Significant QTL identified for TGW in the Charger x Badger doubled haploid population

Table S3 Significant QTL identified for yield in the Charger x Badger doubled haploid population

Table S4 Spike yield components 10 representative single ear samples of 5A− and 5A+ NILs

Table S5 Coefficient of variation for thousand grain weight (TGW), yield and grain morphometric parameters of 5A− and 5A+ NILs

Notes S1 TGACv1 genes in the fine mapped interval and associated expression data.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Appendix 2

Primer table of KASP markers developed during the PhD

Marker name	Primer	Sequence (5'-3')
Hap-P2	F	GAAGGTGACCAAGTTCATGCTAGGGTGAGACGAAAATAAATCGA
	H	GAAGGTCGGAGTCAACGGATTAGGGTGAGACGAAAATAAATCGG
	C	GGACTTGGTAGCTTTCACTTTATGA
JBRNA_Seq1	F	GAAGGTGACCAAGTTCATGCTCCTCTTGCATCATCACTACCAC
	H	GAAGGTCGGAGTCAACGGATTCCTCTTGCATCATCACTACCAT
	C	CCTGCTGGTGTTAGTGGATCT
JBRNA_Seq2	F	GAAGGTGACCAAGTTCATGCTGCCTCCCATCCTTTGACGAG
	H	GAAGGTCGGAGTCAACGGATTGCCTCCCATCCTTTGACGAA
	C	GTGCTATCTTGGACATCTTGTCT
JBRNA_Seq3	F	GAAGGTGACCAAGTTCATGCTTGTAATTTGTTTGCTGCAGAGAC
	H	GAAGGTCGGAGTCAACGGATTTGTAATTTGTTTGCTGCAGAGAA
	C	CTTAGCAGATGGTTCTTTAGTATGC
JBRNA_Seq4	F	GAAGGTGACCAAGTTCATGCTTACTGCCTCTCCTTTCAGCCC
	H	GAAGGTCGGAGTCAACGGATTTACTGCCTCTCCTTTCAGCCT
	C	CCATTTCAGGTCTTGGCTGGTAT
JBHap001	F	GAAGGTGACCAAGTTCATGCTCACTTTCTCATTGCAATGCCATA
	H	GAAGGTCGGAGTCAACGGATTCACTTTCTCATTGCAATGCCATG
	C	TTAGAAAATTCGATGATGCACACT
JBHap002	F	GAAGGTGACCAAGTTCATGCTCCTGAGAGCTATCACCCTCC
	H	GAAGGTCGGAGTCAACGGATTCCTGAGAGCTATCACCCTCT
	C	GCTCTCTTCTCTTCCTTTCGTAC
JBHap003	F	GAAGGTGACCAAGTTCATGCTAGAGTGGAGAGGAAGACCGG
	H	GAAGGTCGGAGTCAACGGATTAGAGTGGAGAGGAAGACCGA
	C	TCACCGCGGCAATGGCTA
JBHap004	F	GAAGGTGACCAAGTTCATGCTAAGGCGATTTCTCAACAGGAT
	H	GAAGGTCGGAGTCAACGGATTAAGGCGATTTCTCAACAGGAC
	C	ATGCCTTCCTACTTCCCTGG
JBHap005	F	GAAGGTGACCAAGTTCATGCTGCACAAAACCAGAGCTAAACCG
	H	GAAGGTCGGAGTCAACGGATTGCACAAAACCAGAGCTAAACCA
	C	GGCTGTTTATTGCAGTTGCC
JBHap006	F	GAAGGTGACCAAGTTCATGCTGGCAGTGCAGGATACGGT
	H	GAAGGTCGGAGTCAACGGATTGGCAGTGCAGGATACGGC
	C	AGCGTAGAAAAGCCACAAGAAG
JBHap007	F	GAAGGTGACCAAGTTCATGCTAACGGCAATTAATATCGATGGAAG
	H	GAAGGTCGGAGTCAACGGATTAACGGCAATTAATATCGATGGAAA
	C	GCACCACACAGTTAGCTTAAAGAT
JBHap008	F	GAAGGTGACCAAGTTCATGCTTGAACCCAACGAAGCAGAATC
	H	GAAGGTCGGAGTCAACGGATTTGAACCCAACGAAGCAGAATT
	C	GCGGCGAAATTTATGTGGTTG
JBHap009	F	GAAGGTGACCAAGTTCATGCTACCTTTCACATAAATTTGAGGTGT
	H	GAAGGTCGGAGTCAACGGATTACCTTTCACATAAATTTGAGGTGC
	C	ACTCTCGGGTTAAATACAGAACA
JBHap010	F	GAAGGTGACCAAGTTCATGCTTCCAGCATGTGATTAACTACGATAT
	H	GAAGGTCGGAGTCAACGGATTTCCAGCATGTGATTAACTACGATAA
	C	AGTTGCTATTCCAGTTTTCCCAT

Marker name	Primer	Sequence (5'-3')
JBHap011	F	GAAGGTGACCAAGTTCATGCTTCCTGTTGACACAGAAAGATCAGC
	H	GAAGGTCGGAGTCAACGGATTTCCTGTTGACACAGAAAGATCAGT
	C	AACTGTAACCGACTCGGAGC
JBHap012	F	GAAGGTGACCAAGTTCATGCTACTATGCTCGATTCTCAACACAA
	H	GAAGGTCGGAGTCAACGGATTACTATGCTCGATTCTCAACACAC
	C	TCGTTGAACAATGCAGTGCA
JBHap013	F	GAAGGTGACCAAGTTCATGCTGTTTCAACAAGACCTCCCGG
	H	GAAGGTCGGAGTCAACGGATTGTTTCAACAAGACCTCCCGA
	C	ACATGAAGCTCTCTGCCTG
JBHap014	F	GAAGGTGACCAAGTTCATGCTGTCTCCTGATTTTGGTCGTGC
	H	GAAGGTCGGAGTCAACGGATTGTCTCCTGATTTTGGTCGTGT
	C	AGATTTGCCAGATATCGATGACA
JBHap015	F	GAAGGTGACCAAGTTCATGCTCCTCTTGCATCATCACTACCAC
	H	GAAGGTCGGAGTCAACGGATTCCTCTTGCATCATCACTACCAT
	C	CCTGCTGGTGTTAGTGGATCT
JBHap016	F	GAAGGTGACCAAGTTCATGCTGCCTCCCATCCTTTGACGAG
	H	GAAGGTCGGAGTCAACGGATTGCCTCCCATCCTTTGACGAT
	C	GTGCTATCTTGGACATCTTGTCT
JBHap017	F	GAAGGTGACCAAGTTCATGCTTGTAATTTGTTTGCTGCAGAGAC
	H	GAAGGTCGGAGTCAACGGATTTGTAATTTGTTTGCTGCAGAGAA
	C	CTTAGCAGATGGTTCTTTAGTATGC
JBHap018	F	GAAGGTGACCAAGTTCATGCTGGACGGGTTAACAACAGTACAATAG
	H	GAAGGTCGGAGTCAACGGATTGGACGGGTTAACAACAGTACAATAT
	C	CCTCAACTATCAGGCTGGGA
JBHap019	F	GAAGGTGACCAAGTTCATGCTTGCCTCTGCTGTGATGGT
	H	GAAGGTCGGAGTCAACGGATTTGCCTCTGCTGTGATGGG
	C	CAACGTTAATACTTCTGCACTTACA
JBHap020	F	GAAGGTGACCAAGTTCATGCTTTCTTGCACATTCTTTAATGGAGAG
	H	GAAGGTCGGAGTCAACGGATTTTCTTGCACATTCTTTAATGGAGAT
	C	TTGAGACTCTGGATCATGCG
JBHap021	F	GAAGGTGACCAAGTTCATGCTGGCCATAATTCTTTTGAAAGCACG
	H	GAAGGTCGGAGTCAACGGATTGGCCATAATTCTTTTGAAAGCACA
	C	TCAGGAACGCCCTCTCCG
JBHap022	F	GAAGGTGACCAAGTTCATGCTTACTGCCTCTCCTTTCAGCCC
	H	GAAGGTCGGAGTCAACGGATTTACTGCCTCTCCTTTCAGCCT
	C	CCATTTCAGGTCTTGGCTGGTAT

The primer column indicates the primer type. F = FAM, H = HEX and C = common. F and H primers are labelled probes, with the first 21 bp of each F and C primer sequence being the dye probe sequences

Appendix 3

Brinton J, Simmonds J, Uauy C. 2017. Ubiquitin-related genes are differentially expressed in isogenic lines contrasting for pericarp cell size and grain weight in hexaploid wheat. bioRxiv. doi.org/10.1101/175471

Ubiquitin-related genes are differentially expressed in isogenic lines contrasting for pericarp cell size and grain weight in hexaploid wheat

Jemima Brinton, James Simmonds and Cristobal Uauy

John Innes Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom

jemima.brinton@jic.ac.uk

james.simmonds@jic.ac.uk

cristobal.uauy@jic.ac.uk

Corresponding author: Cristobal Uauy

Abstract

Background

There is an urgent need to increase global crop production. Identifying and combining genes controlling indidviual yield components, such as grain weight, holds the potential to enhance crop yields. Transcriptomics is a powerful tool to gain insights into the complex gene regulatory networks that underlie such traits, but relies on the availability of a high-quality reference sequence and accurate gene models. Previously, we identified a grain weight QTL on wheat chromosome 5A (5A QTL) which acts during early grain development to increase grain length through cell expansion in the pericarp. In this study, we performed RNA-sequencing on near isogenic lines (NILs) segregating for the 5A QTL and used the latest gene models to identify differentially expressed (DE) genes and pathways that potentially influence pericarp cell size and grain weight in wheat.

Results

We sampled grains at four and eight days post anthesis and found genes associated with metabolism, biosynthesis, proteolysis and defence response to be upregulated during this stage of grain development in both NILs. We identified a specific set of 112 transcripts DE between 5A NILs at either time point, including seven potential candidates for the causal gene underlying the 5A QTL. The 112 DE transcripts had functional annotations including non-coding RNA, transpon-associated, cell-cycle control, and ubiquitin-related processes. Many of the wheat genes identified belong to families that have been previously associated with seed/grain development in other species. However, few of these wheat genes are the direct orthologs and none have been previously characterised in wheat. Notably, we identified DE transcripts at almost all steps of the pathway associated with ubiquitin-mediated protein degradation. In the promoters of a subset of DE transcripts we identified enrichment of binding sites associated with C2H2, MYB/SANT, YABBY, AT-HOOK and Trihelix transcription factor families.

Conclusions

In this study, we identified DE transcripts with a diverse range of predicted biological functions, reflecting the complex nature of the pathways that control early grain development. Further functional characterisation of these candidates and how they interact could provide new insights into the control of grain size in cereals, ultimately improving crop yield.

Keywords

Wheat, RNA-seq, Ubiquitin, Grain weight, Pericarp, Transcriptomic, Grain size, near isogenic lines

Background

Crop production must increase to meet the demands of a global population estimated to exceed nine billion by 2050 [1]. Indeed, one in nine people currently live under food insecurity [2]. With limited opportunity for agricultural expansion, increasing yields on existing land could significantly reduce the number of people at risk of hunger [3]. It is estimated that at least a 50% increase in crop production is required by 2050 [4, 5], however current rates of yield increase are insufficient to achieve this goal [6]. It is therefore critical and urgent that we identify ways to increase crop yields.

Final crop yield is influenced by the interaction of many genetic and environmental factors. This complexity hinders its study and has meant that the mechanisms controlling this trait are not well understood. Grain weight, however, an important component of final yield, is more stably inherited and is better understood than yield itself [7]. Grain weight is mainly determined by grain size, which itself is controlled by the coordination of cell proliferation and expansion processes. Studies in both crop and model species have shown that these processes are regulated by a wide range of genes and molecular mechanisms (reviewed in [8, 9]). Control at the transcriptional level has been demonstrated, with the rice transcription factor (TF) *OsSPL16* influencing grain size through cell proliferation [10], whilst a WRKY domain TF, *TTG2,* influences cell expansion in the integument of

the Arabidopsis seed [11]. Important pathways relating to protein turnover have also been identified, for example the E3 ubiquitin-ligase, *GW2,* which negatively regulates grain weight and width in rice through the control of cell division [12]. *GW2* orthologues in other species, including Arabidopsis and wheat, also act as negative regulators of seed/grain size suggesting that these mechanisms may be conserved across species [13, 14]. Other pathways/mechanisms which affect grain size include microtubule dynamics [15, 16], G-protein signalling [17, 18] and phytohormone biosynthesis and signalling [19-21].

Wheat is a crop of global importance, accounting for approximately 20 % of the calories consumed by the human population [22]. However, our understanding of the mechanisms controlling grain size remains limited in wheat, compared to rice and Arabidopsis. Comparative genomics approaches have provided some insight [13, 23] and many quantitative trait loci (QTL) associated with grain size and shape components (grain area, length and width) have been identified [24-29]. However, none of these QTL have been cloned and little is understood about the underlying mechanisms. Previously, we identified a QTL associated with increased grain weight on wheat chromosome 5A. Using BC_4 near isogenic lines (NILs) we determined that the QTL acts during the early stages of grain development to increase grain length through increased cell expansion in the pericarp [29]. This and other studies [13, 30, 31] suggest that the early stages of grain/ovule development are important for determining final grain size/shape in wheat.

Transcriptomics is a powerful tool to gain insights into the complex gene regulatory networks that underlie specific traits and biological processes. Several studies have used transcriptomics approaches to look at the genes expressed during grain development in wheat [32-38]. However, these studies have mostly focused on the later stages of grain development, often focusing on starch accumulation in the endosperm. Additionally, many of these studies were performed using microarrays [33, 36, 37], which represent a fraction of the transcriptome and are unable to distinguish between homoeologous gene copies. More recent studies have used RNA-seq [35, 34],

which is an open-ended platform that provides homoeolog specific resolution. However, the accuracy of RNA-seq is dependent on the availability of a high-quality reference sequence and accurate gene models. Until recently, the large (~17 Gb) and highly repetitive nature of the hexaploid wheat genome meant that genomic resources were limited and incomplete. However, this has changed drastically in the last few years with the release of several whole genome sequences and annotations [39-42]. To date, the RNA-seq grain development studies have used either expressed sequence tags (ESTs) [35, 38] or the Chromosome Survey Sequence (CSS) [34] as references. However in hindsight, these annotations are incomplete with respect to the latest gene models [41, 39]. These novel resources provide new opportunities for more detailed and accurate transcriptomic studies in wheat.

A potential drawback of transcriptomic studies is that comparisons across varieties, tissues or time points can result in a large number of transcripts being differentially expressed. While this informs our understanding of the biological mechanisms, it is difficult to prioritise specific genes for downstream analysis. Comparative transcriptomic approaches using more precisely defined genetic material, tissues and developmental time points can aid in this by defining a smaller set of differentially regulated transcripts. For example, a comparison of the flag leaf transcriptomes of wild-type and RNAi knockdown lines of the *Grain Protein Content 1* (*GPC*) genes was used to identify downstream targets of the *GPC* TFs [43]. Similarly, the transcriptomes of NILs segregating for a major grain dormancy QTL on chromosome arm 4AL were compared and specific candidate genes underlying the QTL were identified [44]. To our knowledge, no such experiments have been performed on isogenic lines with a known difference for grain size in wheat.

In this study, we performed RNA-seq on NILs segregating for a major grain weight QTL on chromosome arm 5AL. Previously, we showed that the QTL acts during early grain development and that NILs carrying the positive 5A allele (5A+ NILs) have significantly increased thousand grain weight (TGW; 7%), grain length (4%) and pericarp cell length (10%) compared to NILs carrying the negative

5A allele (5A- NILs) [29]. The NILs carry an introgressed segment of ~490 Mb and using recombinant inbred lines we fine-mapped the grain length effect to a 75 Mb region on the long arm of chromosome 5A according to the IWGSC RefSeq v1.0. The aim of the present study was to identify biological pathways that potentially influence grain length and pericarp cell size by using RNA-seq to identify genes that are differentially regulated between the 5A- and 5A+ NILs.

Results

RNA-sequencing of 5A near isogenic lines

We performed RNA-seq on whole grains from two 5A NILs which contrast for grain length [29]. We chose the time point when NILs show the first significant differences in grain length (8 days post anthesis (dpa); T2) and the preceding time point (4 dpa; T1) to capture differences in gene expression occurring during this period (Figure 1). We hypothesised that although there was no significant difference in the grain length phenotype at T1, phenotypic differences were beginning to emerge and gene expression changes influencing this may already be occurring. We obtained over 362 M reads across all 12 samples (two time points, two NILs, three biological replicates), with individual samples ranging from 15.0 M to 53.6 M reads and an average of 30.2 M reads (standard error ± 3.5 M reads) per sample (Table 1). We aligned reads to two different transcriptome sequences from the Chinese Spring reference accession, the IWGSC Chromosome Survey Sequence (CSS) [40] and TGACv1 (TGAC) [41] reference. On average across samples, 69.8 ± 0.3 % of reads aligned to the CSS reference, whilst 84.4 ± 0.2 % of reads aligned to the TGAC reference.

Comparison between Chinese Spring reference transcriptomes

We defined a transcript as expressed if it had an average abundance of > 0.5 transcripts per million (tpm) in at least one of the four conditions (2 NILs x 2 time points). This resulted in 62.5 % (64,020) and 37.1% (101,652) of the transcripts being expressed in the CSS and TGAC transcriptomes, respectively. We defined differentially expressed (DE) transcripts (q value < 0.05) using sleuth [45]

and performed four pairwise comparisons: two 'across time' and two 'between NIL' comparisons. The 'across time' analyses consisted of a comparison between T1 and T2 samples of the 5A- NIL (hereafter symbolised as 5A- $^{T1}_{T2}$; Figure 1, grey) and the corresponding comparison for the 5A+ NIL samples (hereafter 5A+ $^{T1}_{T2}$; Figure 1, purple). In both cases, the T1 sample was used as the control condition, so transcripts were considered as upregulated or downregulated with respect to T1. The 'between NIL' analyses consisted of a comparison between the 5A- and 5A+ NILs at T1 (hereafter T1 $^{5A-}_{5A+}$; Figure 1, orange), and a comparison between the 5A- and 5A+ NILs at T2 (hereafter T2 $^{5A-}_{5A+}$; Figure 1, green). In both cases, the recurrent parent 5A- NIL was used as the control genotype. In all cases, more DE transcripts were identified in the TGAC compared with the CSS transcriptome, and similar trends were observed for both references across the four comparisons (Figure 1).

We selected the comparison with the fewest DE transcripts (T1 $^{5A-}_{5A+}$; 32 and 88 DE transcripts for CSS and TGAC, respectively) to conduct a more in depth analysis of the alignments and references. For all DE transcripts from each alignment we identified the equivalent transcript/gene model in the other reference sequence using *Ensembl* plants release 35 and compared the gene models (Additional file 1). For 64 of the TGAC DE transcripts we did not identify an equivalent CSS DE transcript, either because there was no corresponding CSS gene model (47 transcripts) or the expression change between NILs was non-significant for the CSS transcript. Analogously, eleven CSS DE transcripts did not have an equivalent TGAC gene model DE, five of which were due to there being no corresponding TGAC gene model annotated. Combining both sets, we identified 42 groups of equivalent gene models, 26 of which were differentially expressed in both alignments. Comparing these 42 groups and taking into account fused and split gene models within each dataset, there were 97 gene models in both datasets (50 CSS + 47 TGAC) (Figure 2a, Additional file 1). Of these, only six were identical between the CSS and TGAC references. All other discrepant gene models fell under categories included truncations in either reference, gene models that were split/fused in one reference sequence, and gene models that differed drastically in their overall structure.

For all discrepant gene models we used transcriptome read mapping and an interspecies comparison to determine which gene model seemed most plausible. Figure 2b shows an example of the most commonly identified discrepancy where a gene model was truncated in the CSS reference (pink) relative to the TGAC reference (grey). The DE TGAC gene model was supported by our transcriptome data as we observed read coverage across the whole gene model whilst the coverage across the CSS gene model dropped at the position where an intron is predicted in the TGAC model. Another common discrepancy was a single gene model in one reference being split into multiple gene models in the other reference. Figure 2c shows an instance where a single DE TGAC gene model comprised four separate CSS gene models. In this case, all five gene models had coverage across the entire gene body, however the single TGAC gene model was more similar to proteins from other species, suggesting that this single gene model was most likely correct. The final example (Figure 2d) shows two TGAC gene models that were fused into a single CSS gene model. The coverage across the CSS gene model was inconsistent, with most reads concentrated in the 3' untranslated region (UTR). The two TGAC gene models had more consistent coverage across the entire gene models and were both supported by protein alignments with other species. Interestingly, only the shorter TGAC gene model was DE (Figure 2d, grey), suggesting that differential expression of the CSS gene model was driven by the reads mapping to the putative 3' UTR rather than the coding regions of the transcript (Figure 2d, pink). Taking together the fact that a higher percentage of reads mapped to the TGAC gene models and that many more of the examined TGAC gene models were supported by interspecies comparison and expression data than the CSS gene models, we decided to continue our analysis using the alignments to the TGAC gene models only.

Many DE transcripts during early grain development are shared between NILs

We identified 3,151 and 2,789 DE transcripts across early grain development in 5A- $\frac{T1}{T2}$ and 5A+ $\frac{T1}{T2}$, respectively (Figure 1, Figure 3a). The DE transcripts were evenly distributed across the 21 chromosomes, showing no overall bias towards any chromosome group or subgenome (Figure 3b).

Approximately 60% (1,832) of the DE transcripts were shared between 5A- $\frac{T1}{T2}$ and 5A+ $\frac{T1}{T2}$ (Figure 3a) and 84% (1,532) of the shared transcripts were upregulated across time (Figure 3c). We identified 41 significantly enriched gene ontology (GO) terms in the upregulated transcripts (Additional file 2). Sixteen of the GO terms were associated with biological process and could be grouped under three parent GO terms: metabolic process (GO:0008152), defence response (GO:0006952) and biological regulation (GO:0065007) (Figure 3c). Within metabolic process we found terms associated with carbohydrate (GO:0005975) and pyruvate metabolism (GO:0006090), vitamin E (GO:0010189) and triglyceride biosynthesis (GO:0019432), mRNA catabolism (GO:0006402), proteolysis (GO:0006508) and phosphorylation (GO:0016310). Downregulated transcripts (300) were enriched for seven GO terms, four of which were associated with biological process: potassium ion transport (GO:0006813), signal transduction (GO:0007165), phosphorelay signal transduction (GO:0000160) and carbohydrate metabolism (GO:0005975). The overlap between enriched GO terms in the upregulated and downregulated transcripts (e.g. carbohydrate metabolism) suggests that different aspects of these processes are being differentially regulated during this early grain development stage.

We also identified many transcripts that were only DE across early grain development in one of the two genotypes (i.e. unique to either the 5A- $\frac{T1}{T2}$ or 5A+ $\frac{T1}{T2}$ comparisons). However, many of these transcripts were borderline non-significant in the opposite genotype comparison illustrated by the fact that the distributions of q-values were skewed towards significance (Additional file 3). Additionally, the uniquely DE transcripts were enriched for GO terms similar to the shared transcripts (Additional file 2). Some GO terms, however, were only enriched in the uniquely DE transcripts, for example, cell wall organisation or biosynthesis (GO:0071554) and response to abiotic stimulus (GO:0009628). Overall, these results suggests that although there were some differences between genotypes, broadly similar biological processes were taking place in the grains of both the 5A NILs at the early stages of grain development.

DE transcripts between NILs are concentrated on chromosome 5A

We identified 88 and 91 DE transcripts between the NILs in T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$, respectively, many fewer than identified in 5A- $^{T1}_{T2}$ or 5A+ $^{T1}_{T2}$. This was expected as the NILs are genetically very similar and therefore the difference in developmental stage between the T1 and T2 time points results in greater changes in gene expression. Of these 179 DE transcripts, 67 were common between T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$, whereas 45 DE transcripts between genotypes were unique and identified only at a single time point (resulting in 112 DE transcripts between NILs at any time point). No GO terms were significantly enriched in these groups. Of the 67 common DE transcripts, 54 (80%) were located on chromosome 5A, whilst in both the T1 and T2 unique groups less than 50% were located on chromosome 5A (Figure 4a). Similar numbers of DE transcripts were more highly expressed in either genotype, with no distinct patterns observed between the unique or common groups.

We looked specifically at the positions of the 74 DE transcripts located on chromosome 5A and found that all were located within the 491 Mbp introgressed region of the NILs (Figure 4b). Higher numbers of DE transcripts were identified in regions of increased SNP density between the 5A NILs. Previously, we fine-mapped the grain length effect to a 75 Mbp interval on 5AL (between BS00182017 (317 Mbp) and BA00228977 (392 Mbp; [29]) and eight of the DE transcripts were located within this interval. Three of these transcripts were more highly expressed in the 5A+ NILs (5A+$_{high}$ transcripts), two of which were transcript variants of the same gene (a kinesin-like protein; only .2 variant shown in Figure 4b). The other 5A+$_{high}$ transcript was annotated as a putative retrotransposon protein. One of the five transcripts more highly expressed in the 5A- NIL (5A-$_{high}$ transcript) had no annotation and the remaining four were annotated as a non-coding RNA, a RING/U-box containing protein, a TauE-like protein and a DUF810 family protein.

DE transcripts outside of chromosome 5A are enriched in specific transcription factor binding sites

As all the DE transcripts on chromosome 5A were located within the 491 Mbp introgressed region, it is possible that the differential expression was a direct consequence of sequence variation between the NILs e.g. in the promoter regions. However, the 38 DE transcripts located outside of chromosome 5A have the same nucleotide sequence as they are identical by descent (BC_4 NILs confirmed with 90k iSelect SNP marker data [29]). We hypothesised that these transcripts are downstream targets of DE genes, such as transcription factors (TFs), located within the 5A introgression.

To assess this, we identified transcription factor binding sites (TFBS) present in the promoter regions of these 38 DE transcripts. We identified TFBS associated with 91 distinct TF families present in this group of transcripts (Additional file 4), five of which were enriched relative to all expressed transcripts (Table 2; FDR adjusted $P < 0.05$). The enriched TFBS families were C2H2, Myb/SANT, AT-Hook, YABBY and MADF/Trihelix.

To determine potential candidates for upstream regulators we identified all TFs located within the introgressed region on chromosome 5A [46]. We identified a total of 200 annotated TFs, belonging to 35 TF families. Of these, four families corresponding to 29 TF overlapped with enriched TFBS families. Four of the 29 TFs were located within the fine-mapped grain length region on chromosome 5A, including C2H2, MYB and MYB_related TFs (Additional file 5). However, none of them were DE between NILs at the two time points.

Functional annotation of DE transcripts

Having analysed DE transcripts between NILs based on chromosome location, we looked at the 112 DE transcripts based on their functional annotations (Additional file 6). We identified multiple categories including transcripts associated with ubiquitin-mediated protein degradation, cell cycle,

metabolism, transport, transposons and non-coding RNAs (Table 3). Few categories were exclusively located on/outside 5A or had exclusively higher expression in the either the 5A- or 5A+ NIL.

The category with the most DE transcripts was non-coding RNA (ncRNA, 15 transcripts), although this was not enriched relative to all expressed transcripts. All ncRNA transcripts were classed as long non-coding RNAs (>200bp, [47]) and we found that four of the ncRNAs overlapped with coding transcripts (two in the antisense direction) and one ncRNA was a putative miRNA precursor (Ta-miR132-3p; [48]). We identified 13 transcripts as putative targets of Ta-miR132-3p in the TGAC reference but none of these target transcripts were differentially expressed in our dataset. The second largest transcript category was transposon-associated (14 transcripts; FDR-adjusted p = 0.008), whereas the third largest category was DE transcripts related to ubiquitin and the proteasome (12 transcripts; p = 0.008). DE transcripts annotated as homeobox were also enriched (4 transcripts; FDR-adjusted p = 0.001). Interestingly, we identified homeodomain TFBS in 27 of the 38 outside 5A DE transcripts although this was not significantly enriched (FDR-adjusted p = 0.166, Additional file 4).

The DE transcripts related to ubiquitin were of particular interest as ubiquitin-mediated protein turnover has previously been associated with the control of seed/grain size in wheat [13] and other species including rice and Arabidopsis [14, 12, 49]. The pathway acts through the sequential action of a cascade of enzymes (see Figure 5a legend) to add multiple copies of the protein ubiquitin (ub) to a substrate protein that is then targeted for degradation by the proteasome. We identified differential expression of transcripts at almost all steps of this pathway (excluding E1): two ubiquitin proteins and one ubiquitin-like protein, one E2 conjugase, six potential E3 ligase components and two putative components of the proteasome (Figure 5). In addition to these, we also identified four DE transcripts annotated as proteases (Figure 5), which are known substrates regulated by this pathway [50-52] and that influence organ size through the regulation of cell proliferation. Most of

279 the components of the ubiquitin pathway that were differentially expressed were more highly
280 expressed in the 5A- NIL (11/16, including proteases) (Figure 5b).

281 Discussion

282 In this study, we performed RNA-seq on the grains of 5A NILs with a known difference in pericarp
283 cell size, grain length and final grain weight. We previously determined that the first phenotypic
284 differences between NILs arose during early grain development [29]. We hypothesised that
285 differences in gene expression between NILs during these early stages would allow us to identify
286 specific genes and pathways that that affect pericarp cell size and grain size at the transcriptional
287 level.

288 The importance of a high-quality reference sequence

289 We initially mapped the RNA-seq data to two different reference transcriptomes: CSS and TGAC. We
290 found that TGAC outperformed the CSS transcriptome both in term of the number of reads that
291 aligned and in the gene models themselves. This was most likely due to the significant improvement
292 in terms of sequence contiguity of the TGAC reference over the CSS (N50= 88.8 vs < 10 kb,
293 respectively), allowing more accurate prediction of gene models. Our study highlights the practical
294 importance of this improvement as we detected 64 more DE transcripts using the TGAC reference, in
295 most cases, due to the absence of a corresponding gene model in the CSS reference (46 transcripts).
296 We also identified cases where incorrect gene models in the CSS reference led to misleading results.
297 For example, in the CSS fused gene model case study (Figure 2d) a single DE transcript from the CSS
298 reference had a large accumulation of reads mapping to the 3' UTR. This gene was the orthologue of
299 Arabidopsis *NPY1*, which plays a role in auxin-regulated organogenesis [53] and could therefore be
300 related to the control of grain size. However, in the TGAC reference, in addition to the *NPY1*
301 orthologue, an alternative gene model was annotated in place of the 3' UTR. This alternative gene

model was differentially expressed whilst the *NPY1* orthologue was expressed at a very low level and was not differentially expressed.

The improvements in scaffold size, contiguity and gene annotation open up new opportunities in wheat research. Here we used the new physical sequence to assign locations to 107 of 112 DE transcripts identified between NILs, allowing us to determine which DE transcripts were located within the QTL fine-mapped interval. Likewise, the analysis of promoter sequences enabled new hypothesis generation for this specific biological process and will also aid in the understanding of how promoter differences across genomes affects the relative transcript abundance of the different homoeologs. This exemplifies the importance of correctly annotated gene models and improved genome assemblies in gaining a more accurate view of the underlying biology.

Differential expression analysis provides an insight into the biological processes occurring in early grain development

We sampled grains at 4 and 8 dpa to encompass the developmental stage at which the first significant difference in grain length between 5A NILs is observed. During this stage, increases in grain size are largely driven by cell expansion in the pericarp [54, 55], consistent with our previous finding that increased pericarp cell size underlies the difference in final grain length. These time points are also relatively early compared to other grain related RNA-seq studies which have focused on later grain filling processes [36, 56, 35]. The 'across time' comparisons (5A-$\frac{T1}{T2}$ and 5A+ $\frac{T1}{T2}$) identified > 2,700 DE transcripts in each NIL, and there was a large overlap in the biological processes being differentially regulated. We found that most DE transcripts were upregulated over time and many of these were associated with metabolism and biosynthesis consistent with grains undergoing a period of rapid growth and the start of endosperm cellularisation at this stage of development [32]. Transcripts associated with proteolysis and mRNA catabolism were also upregulated across time consistent with increases in specific proteases and other hydrolytic enzymes at this stage of grain development [57]. These could be indicative of programmed cell death which

occurs in both the nucellus and pericarp of the developing grain up to 12 dpa [54]. We also identified an upregulation of transcripts associated with defence response and oxidation-reduction process, consistent with previous reports of accumulation of proteins associated with defence against both pathogens and oxidative stress during the early-mid stages of grain development [58]. Transcriptional studies always have the caveat that changes in gene expression may not translate to changes in protein level [59]. However, proteomic analyses of similar stages of grain development have identified the differential regulation of similar ontologies [58, 60] suggesting that these transcriptional changes are reflective of overall protein status in the grain.

Comparative transcriptomics as a method to identify candidate genes underlying the 5A grain length QTL

The use of highly isogenic material allowed the direct comparison of the effect of the 5A introgression on gene expression at each time point (T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$). This resulted in a defined set of 112 DE transcripts between genotypes. The majority of T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ DE transcripts were located on chromosome 5A and all of these were located within the 5A introgression. This is expected given that the sequence variation in the NILs was restricted to the chromosome 5A region.

DE transcripts located within the fine-mapped interval on chromosome 5A represent good candidates for further characterisation. The kinesin-like gene and RING/U-box superfamily protein are particularly strong candidates based on their functional annotations. Previous studies have demonstrated that Kinesin-like proteins can regulate grain length and cell expansion through involvement with microtubule dynamics [15, 16, 61]. The RING/U-box protein is a putative E3 ligase, a class of enzymes which have been associated with the control of grain size (discussed in more detail later; [12, 13]).

It is premature, however, to speculate on the identity of a 5A causal gene(s) at this stage. It is difficult to predict whether DE transcripts in the fine-mapped interval are truly associated with the

effect of the 5A QTL or are simply a consequence of sequence variations between the parental cultivars, i.e. 'guilt by association'. A relevant example was the recent use of transcriptomics to define a candidate gene underlying a grain dormancy QTL (*PM19*) [44]. Subsequent studies showed that a different gene in close physical proximity (*TaMKK3*) [62] was responsible for the natural variation observed [63]. The mis-interpretation of the transcriptomics data was due to complete linkage disequilibrium between the DE *PM19* gene and the causal *TaMKK3* gene in the germplasm used in the original study. Additionally, the causal gene may not be differentially expressed between the 5A NILs and could be a result of allelic variation that alters the function of the gene independent of expression level. Ultimately, further fine-mapping of the 5A locus will be required to identify the underlying gene.

DE transcripts outside chromosome 5A are candidates for downstream targets of the 5A QTL

We considered DE transcripts outside of chromosome 5A as candidates for downstream targets of genes located in the 5A introgression because the differential expression could not have arisen through sequence variation. These included genes located on the A, B and D genomes implying that there is cross-talk at the transcriptional level between the three genomes. We identified, in the promoters of these genes, enrichment of TF binding sites associated with TF families which have all previously been shown to play diverse roles in the control of organ development [64, 65]. For example YABBY genes, a plant specific family of TFs, play a critical role in patterning and the establishment of organ polarity [66] and fruit size [67]. Another example are the C2H2 TFs, *NUBBIN* and *JAGGED,* which are involved in determining carpel shape in Arabidopsis [68]. AT-Hook TFs play roles in floral organ development in both maize and rice [69, 70] and modulate cell elongation in the Arabidopsis hypocotyl [71]. Few of these transcription factor families have been characterised in wheat, and although these interactions need to be experimentally validated, they could be potential targets for the manipulation of grain size.

DE transcripts have functions related to the control of seed/organ size

Studies in species such as rice and Arabidopsis have shown that seed size is regulated by a complex network of genes and diverse mechanisms, ultimately through the coordination of cell proliferation and expansion (reviewed in [8, 9]). 5A+ NILs have significantly longer pericarp cells, suggesting that the underlying gene influences cell expansion [29]. Genes that physically modify the cell wall have been shown to directly control cell expansion (reviewed in [72]) and we identified three DE transcripts that have potential roles in cell wall synthesis and remodelling. We also identified a number of DE transcripts associated with the cell cycle and the control of cell proliferation. During seed development, a number of cell cycle types in addition to the typical mitotic cycle are observed. One such alternative cycle type is endoreduplication, characterised by the replication of chromosomes in the absence of cell division, which is associated with cell enlargement (reviewed in [73]). Two of the DE transcripts were the closest wheat orthologues of Arabidopsis genes that have specific roles in organ development: a *GRF*-interacting factor (*GIF*) and *SEUSS* (*SEU*). In Arabidopsis, the GIF genes interact with the *GROWTH-REGULATING FACTOR* (*GRF*) TFs and act as transcriptional co-activators to regulate organ size through cell proliferation [74]. Conversely, *SEU* acts a transcriptional co-repressor and interacts with important regulators of development to control many processes, including floral organ development [75].

Seed development requires the coordination of processes across multiple tissues, namely the seed coat, endosperm and embryo. The development and growth of these tissues is inherently interlinked, and it has been proposed that the mechanical constraint imposed by the maternal seed coat/pericarp places an upper limit on the size of the seed/grain [30, 76, 29]. Epigenetic regulation appears to play an important role in the cross talk and coordination of these tissues [77]. The differential expression of 34 non-coding transcripts, transposons and histone-related transcripts between NILs could suggest a difference in epigenetic status associated with the control of pericarp

cell size. Additional work to characterise these non-coding RNAs would be warranted to establish their role in grain development.

The ubiquitin-mediated control of seed/grain size has been documented in a number of species (reviewed in [78]), including wheat [13, 79]. DE transcripts associated with the ubiquitin pathway were significantly enriched in the 5A NILs. The pathway tags substrate proteins with multiple copies of the ubiquitin protein through the sequential action of a cascade of enzymes: E1 (Ub activating), E2 (Ub conjugases) and E3 (Ub ligases). The ubiquitinated substrate proteins are then targeted to the 26S proteasome for degradation [80]. *GW2*, a RING-type E3 ligase, negatively regulates cell division and was identified as the causal gene underlying a QTL for grain width and weight in rice [12]. The Arabidopsis orthologue, *DA2,* acts via the same mechanism to regulate seed size in Arabidopsis [14]. Another E3 ligase, *EOD1/BB* also negatively regulates seed size in Arabidopsis [49]. In general, the E3 ligase determines the specificity for the substrate proteins [80] and *DA2* and *EOD1* may have different substrate targets, however they converge and both target the ubiquitin-activated protease DA1. DA1 also negatively regulates cell proliferation and acts syngergistically with both DA2 and EOD1, although it is not clear whether the two E3 ligases act via independent genetic pathways or as part of the same mechanism [14, 81, 50]. *UBP15* (a ubiquitin specific protease) is a downstream target of this pathway and conversely acts as a positive regulator of seed size through the promotion of cell proliferation [51]. Other ubiquitin-associated regulators of organ/grain size have been identified, including components of the 26S proteasome, enzymes with deubiquitinating activity and proteins that have been shown to bind ubiquitin *in vitro* [82, 52, 83]. The DE transcripts associated with this pathway are not direct homologs of these previously characterised genes. As such the functional characterisation of these putative novel components could provide new insights into the ubiquitin-mediated control of grain size in cereals.

Conclusions

In this study we have both generated candidates for the causal gene underlying the 5A QTL, and have also identified potential downstream pathways controlling grain size. A subset of these candidates is being tested functionally using TILLING mutants [84] and other approaches to provide novel insights into the control of grain size in cereals. Ultimately identifying the individual components and pathways that regulate grain size and understanding how they interact will allow us to more accurately manipulate final grain yields in wheat.

Methods

Plant material

The 5A BC_4 NILs used in this study have been described previously [29]. Briefly, the NILs were generated from a doubled haploid population between the UK cultivars 'Charger' and 'Badger' using Charger as the recurrent parent. The NILs differ for an approximately 491 Mbp interval on chromosome 5A. We used one genotype each for the 5A- (Charger allele, short grains) and 5A+ NIL (Badger allele, long grains). Plants were grown in 1.1 x 6 m plots (experimental units) in a complete randomised block design with five replications, and spikes were tagged at full ear emergence [29]. The three blocks with the most similar flowering time were used for sampling. Plants were sampled at 4 (time point 1: T1) and 8 (time point 2: T2) days post anthesis (dpa) based on the 2014 developmental time course outlined in [29]. For each genotype, we sampled three grains from three separate spikes from different plants within the experimental unit. Each biological replicate therefore, consisted of the pooling of nine grains per genotype. Grains were sampled from the outer florets (positions F1 and F2) from the middle section of each of the three spikes. Grains were removed from the spikes in the field, immediately frozen in liquid nitrogen and stored at -80^{o}C. In total, three biological replicates (from the three blocks in the field) were sampled for each NIL at each time point.

RNA extraction and sequencing

For each of the three biological replicate, the nine grains were pooled and ground together under liquid nitrogen. RNA was extracted in RE buffer (0.1 M Tris pH 8.0, 5 mM EDTA pH8.0, 0.1 M NaCl, 0.5% SDS, 1% β-mercaptoethanol) with Ambion Plant RNA Isolation Aid (Thermo Fisher Scientific). The supernatant was extracted with 1:1 acidic Phenol (pH 4.3):Chloroform. RNA was precipitated at -80⍰C by addition of Isopropanol and 3M NA Acetate (pH 5.2). The RNA pellet was washed twice in 70% Ethanol and resuspended in RNAse free water. RNA was DNAse treated and purified using RNeasy Plant Mini kit (Qiagen) according to the manufacturer's instructions. RNA QC, library construction and sequencing were performed by the Earlham Institute, Norwich (formerly The Genome Analysis Centre). Library construction was performed on a PerkinElmer Sciclone using the TruSeq RNA protocol v2 (Illumina 15026495 Rev.F). Libraries were pooled (2 pools of 6) and sequenced on 2 lanes of a HiSeq 2500 (Illumina) in High Output mode using 100bp paired end reads and V3 chemistry. Initial quality assessment of the reads was performed using fastQC [85].

Read alignment and differential expression analysis

Reads were aligned to two reference sequences from the same wheat accession, Chinese Spring: the Chromosome Survey Sequence (CSS; [40] downloaded from *Ensembl* plants release 29) and the TGACv1 reference sequence [41]. We performed read alignment and expression quantification using kallisto-0.42.3 [86] with default parameters, 30 bootstraps (-b 30) and the –pseudobam option. Kallisto has previously been shown to be suitable for the alignment of wheat transcriptome data in a homoeolog specific manner [87].

Differential expression analysis was performed using sleuth-0.28.0 [45] with default parameters. Transcripts with a false-discovery rate (FDR) adjusted p-value (q value) < 0.05 were considered as differentially expressed. Transcripts with a mean abundance of < 0.5 tpm in all four conditions were considered not expressed and were therefore excluded from further analyses.

For each condition the mean tpm of all three biological replicates was calculated. All heatmaps display mean expression values as normalised tpm, on a scale of 0 to 1 with 1 being the highest expression value of the transcript. Read coverage for gene models was obtained using bedtools-2.24.0 genome cov [88] for each pseudobam file and then combined to get a total coverage value of each position. Coverage across a gene model was plotted as relative coverage on a scale of 0 to 1, with 1 being equivalent to the highest level of coverage for the gene model in question.

GO term enrichment

We used the R package GOseq v1.26 [89] to test for enrichment of GO terms in specific groups of DE transcripts. We considered over-represented GO terms with a Benjamini Hochberg FDR adjusted p-value of < 0.05 to be significantly enriched.

Functional annotation

Functional annotations of transcripts were obtained from the TGACv1 annotation [41]. Additionally, for coding transcripts we performed BLASTP against the non-redundant NCBI protein database and conserved domain database, in each case the top hit based on e-value was retained. In cases where all three annotations were in agreement, the TGAC annotation is reported. In cases where the three annotations produced differing results, all annotations are reported. Orthologues in other species such as Arabidopsis and rice were obtained from *Ensembl* plants release 36. Eight of the 112 DE transcripts had no annotation or protein sequence similarity with other species. We manually categorised the remaining 104 DE transcripts based on their predicted function. Transcripts that fell into a category of size 1 were classed as 'other'. For the non-coding transcripts, we used BLASTN to identify potential miRNA precursors using a set of conserved and wheat specific miRNA sequences obtained from Sun et al, 2014 [48]. We used the -task blastn-short option of BLAST for short sequences and only considered hits of the full length of the miRNA sequence with no mismatches as potential precursors. We used the psRNAtarget tool (http://plantgrn.noble.org/psRNATarget/) to determine the miRNA targets.

Identification of transcription factor binding sites

We extracted 1,000 bp of sequence upstream of the cDNA start site to search for transcription factor binding sites (TFBS). Transcripts with < 1,000 bp upstream in the reference sequence were not used in the analysis. We used the FIMO tool from the MEME suite (v 4.11.4; [90]) with a position weight matrix (PWM) obtained from plantPAN 2.0 (http://plantpan2.itps.ncku.edu.tw/; [91]). We ran FIMO with a p value threshold of <1e-4 (default), increased the max-stored-scores to 1,000,000 to account for the size of the dataset, and used a –motif-pseudo of 1e-8 as recommended by Peng et al [92] for use with PWMs. We generated a background model using the fasta-get-markov command of MEME on all extracted promoter sequences.

Enrichment testing

To test for enrichment of different categories of transcripts relative to all expressed transcripts we performed Fisher's exact test using R-3.2.5. For functional annotation categories, enrichment testing was only performed on categories that could be extracted using GO terms and key words based on their annotation in the TGAC reference. Only DE transcripts that could be extracted using this method were used in the enrichment tests. For example, we identified 12 DE transcripts associated with ubiquitin. The annotation of these transcripts was obtained through a combination of the TGAC annotation and manual annotation. However, only seven of these transcripts could be extracted using GO terms and key words from the whole reference annotation. Therefore, only seven transcripts were used for the enrichment test.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data sets supporting the results of this article are included withing the article and its additional files. Raw sequence reads have been deposited in the NCBI Sequence Read Archive under the Bioproject PRJNA396738.

Competing interests

The authors declare that they have no competing interests

Authors contributions

JB designed the research, performed RNA extractions, analysed the data, performed statistical analyses and wrote the manuscript; JS coordinated the field trials and developed the germplasm used in this study; CU designed the research and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) Grants BB/P013511/1 and BB/P016855/1 and the International Wheat Yield Partnership (grant IWYP76). JB was supported by the UK Agriculture and Horticulture Development Board (AHDB) and the John Innes Foundation. We thank David Swarbreck and Gemy Kaithakottil (Earlham Institute) for *in silico* mapping of TGACv1 gene models and to the IWGSC for pre-publication access to RefSeq v.1.0

References

1. United Nations, Departments of Economic and Social Affairs, Population Division. World Population Prospects: The 2015 Revision 27 February 2017.

2. FAO, IFAD, WFP. The State of Food Insecurity in the World 2015. Meeting the 2015 international hunger targets: taking stock of uneven progress. Rome, FAO 2015.

3. Rosegrant MW, Tokgoz S, Bhandary P, Msangi S. Looking ahead: Scenarios for the future of food. 2012 Global Food Policy Report. Washington, D.C.: International Food Policy Research Institute 2013. p. 88-101.

4. Tilman D, Balzer C, Hill J, Befort BL. Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences. 2011;108(50):20260-4. doi:10.1073/pnas.1116437108.

5. FAO. Global agriculture towards 2050. Rome, FAO. 2009.

6. Ray DK, Mueller ND, West PC, Foley Ja. Yield Trends Are Insufficient to Double Global Crop Production by 2050. PloS one. 2013;8(6):e66428-e. doi:10.1371/journal.pone.0066428.

7. Kuchel H, Williams KJ, Langridge P, Eagles Ha, Jefferies SP. Genetic dissection of grain yield in bread wheat. I. QTL analysis. Theoretical and Applied Genetics. 2007;115(8):1029-41. doi:10.1007/s00122-007-0629-7.

8. Li N, Li Y. Maternal control of seed size in plants. Journal of Experimental Botany. 2015;66(4):1087-97. doi:10.1093/jxb/eru549.

9. Huang R, Jiang L, Zheng J, Wang T, Wang H, Huang Y et al. Genetic bases of rice grain shape: so many genes, so little known. Trends in Plant Science. 2013;18(4):218-26. doi:http://dx.doi.org/10.1016/j.tplants.2012.11.001.

10. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X et al. Control of grain size, shape and quality by *OsSPL16* in rice. Nature Genetics. 2012;44(8):950-4. doi:http://dx.doi.org/10.1038/ng.2327.

562 11. Garcia D, Gerald JNF, Berger F. Maternal control of integument cell elongation and zygotic
563 control of endosperm growth are coordinated to determine seed size in Arabidopsis. The Plant Cell.
564 2005;17(1):52-60. doi:10.1105/tpc.104.027136.
565 12. Song X-J, Huang W, Shi M, Zhu M-Z, Lin H-X. A QTL for rice grain width and weight encodes a
566 previously unknown RING-type E3 ubiquitin ligase. Nature Genetics. 2007;39(5):623-30.
567 doi:10.1038/ng2014.
568 13. Simmonds J, Scott P, Brinton J, Mestre TC, Bush M, Blanco A et al. A splice acceptor site mutation
569 in *TaGW2-A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and
570 longer grains. Theoretical and Applied Genetics. 2016;129(6):1099-112. doi:10.1007/s00122-016-
571 2686-2.
572 14. Xia T, Li N, Dumenil J, Li J, Kamenski A, Bevan MW et al. The Ubiquitin Receptor *DA1* Interacts
573 with the E3 Ubiquitin Ligase *DA2* to Regulate Seed and Organ Size in Arabidopsis. The Plant Cell.
574 2013;25(9):3347-59. doi:10.1105/tpc.113.115063.
575 15. Kitagawa K, Kurinami S, Oki K, Abe Y, Ando T, Kono I et al. A Novel Kinesin 13 Protein Regulating
576 Rice Seed Length. Plant and Cell Physiology. 2010;51(8):1315-29. doi:10.1093/pcp/pcq092.
577 16. Fujikura U, Elsaesser L, Breuninger H, Sánchez-Rodríguez C, Ivakov A, Laux T et al. Atkinesin-13A
578 Modulates Cell-Wall Synthesis and Cell Expansion in *Arabidopsis thaliana* via the THESEUS1 Pathway.
579 PLOS Genetics. 2014;10(9):e1004627. doi:10.1371/journal.pgen.1004627.
580 17. Fan C, Xing Y, Mao H, Lu T, Han B, Xu C et al. *GS3*, a major QTL for grain length and weight and
581 minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein.
582 Theoretical and Applied Genetics. 2006;112(6):1164-71. doi:10.1007/s00122-006-0218-1.
583 18. Huang X, Qian Q, Liu Z, Sun H, He S, Luo D et al. Natural variation at the *DEP1* locus enhances
584 grain yield in rice. Nature genetics. 2009;41(4):494-7. doi:10.1038/ng.352.
585 19. Riefler M, Novak O, Strnad M, Schmülling T. Arabidopsis cytokinin receptor mutants reveal
586 functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin
587 metabolism. The Plant Cell. 2006;18(1):40-54.

588 20. Schruff MC, Spielman M, Tiwari S, Adams S, Fenby N, Scott RJ. The *AUXIN RESPONSE FACTOR 2*
589 gene of Arabidopsis links auxin signalling, cell division, and the size of seeds and other organs.
590 Development. 2006;133(2):251-61. doi:10.1242/dev.02194.
591 21. Jiang W-B, Huang H-Y, Hu Y-W, Zhu S-W, Wang Z-Y, Lin W-H. Brassinosteroid regulates seed size
592 and shape in Arabidopsis. Plant physiology. 2013;162(4):1965-77.
593 22. FAO. Online statistical database: Food balance. FAOSTAT. 2017. http://www.fao.org/faostat/en/.
594 Accessed 27 February 2017.
595 23. Ma M, Zhao H, Li Z, Hu S, Song W, Liu X. *TaCYP78A5* regulates seed size in wheat (*Triticum*
596 *aestivum*). Journal of Experimental Botany. 2016;67(5):1397-410. doi:10.1093/jxb/erv542.
597 24. Simmonds J, Scott P, Leverington-Waite M, Turner AS, Brinton J, Korzun V et al. Identification and
598 independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of
599 hexaploid wheat (*Triticum aestivum L.*). BMC Plant Biology. 2014;14(1):191. doi:10.1186/s12870-
600 014-0191-9.
601 25. Breseghello F, Sorrells ME. QTL analysis of kernel size and shape in two hexaploid wheat mapping
602 populations. Field Crop Res. 2007;101(2):172-9. doi:10.1016/j.fcr.2006.11.008.
603 26. Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S et al. A genetic framework for grain
604 size and shape variation in wheat. The Plant cell. 2010;22(4):1046-56. doi:10.1105/tpc.110.074153.
605 27. Farré A, Sayers L, Leverington-Waite M, Goram R, Orford S, Wingen L et al. Application of a
606 library of near isogenic lines to understand context dependent expression of QTL for grain yield and
607 adaptive traits in bread wheat. BMC Plant Biology. 2016;16. doi:https://doi.org/10.1186/s12870-
608 016-0849-6.
609 28. Kumar A, Mantovani EE, Seetan R, Soltani A, Echeverry-Solarte M, Jain S et al. Dissection of
610 Genetic Factors underlying Wheat Kernel Shape and Size in an Elite x Nonadapted Cross using a High
611 Density SNP Linkage Map. Plant Genome-Us. 2016;9(1). doi:10.3835/plantgenome2015.09.0081.

612 29. Brinton J, Simmonds J, Minter F, Leverington-Waite M, Snape J, Uauy C. Increased pericarp cell
613 length underlies a major quantitative trait locus for grain weight in hexaploid wheat. New
614 Phytologist. 2017;215(3):1026-38. doi:10.1111/nph.14624.
615 30. Hasan AK, Herrera J, Lizana C, Calderini DF. Carpel weight, grain length and stabilized grain water
616 content are physiological drivers of grain weight determination of wheat. Field Crop Res.
617 2011;123(3):241-7. doi:https://doi.org/10.1016/j.fcr.2011.05.019.
618 31. Guo Z, Chen D, Schnurbusch T. Variance components, heritability and correlation analysis of
619 anther and ovary size during the floral development of bread wheat. Journal of Experimental Botany.
620 2015. doi:10.1093/jxb/erv117.
621 32. Shewry PR, Mitchell RaC, Tosi P, Wan Y, Underwood C, Lovegrove A et al. An integrated study of
622 grain development of wheat (cv. Hereward). Journal of Cereal Science. 2012;56(1):21-30.
623 doi:10.1016/j.jcs.2011.11.007.
624 33. Wan Y, Poole RL, Huttly AK, Toscano-Underwood C, Feeney K, Welham S et al. Transcriptome
625 analysis of grain development in hexaploid wheat. BMC Genomics. 2008;9. doi:10.1186/1471-2164-
626 9-121.
627 34. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR et al. Genome interplay in the grain
628 transcriptome of hexaploid bread wheat. Science. 2014;345. doi:10.1126/science.1250091.
629 35. Pellny TK, Lovegrove A, Freeman J, Tosi P, Love CG, Knox JP et al. Cell walls of developing wheat
630 starchy endosperm: comparison of composition and RNA-Seq transcriptome. Plant Physiology.
631 2012;158(2):612-27. doi:10.1104/pp.111.189191.
632 36. Yu Y, Zhu D, Ma C, Cao H, Wang Y, Xu Y et al. Transcriptome analysis reveals key differentially
633 expressed genes involved in wheat grain development. The Crop Journal. 2016;4(2):92-106.
634 doi:http://dx.doi.org/10.1016/j.cj.2016.01.006.
635 37. Laudencia-Chingcuanco DL, Stamova BS, You FM, Lazo GR, Beckles DM, Anderson OD.
636 Transcriptional profiling of wheat caryopsis development using cDNA microarrays. Plant Mol Biol.
637 2007;63(5):651-68. doi:10.1007/s11103-006-9114-y.

638 38. Liu W, Zhihui Wu, Yufeng Zhang, Dandan Guo, Yuzhou Xu, Weixia Chen et al. Transcriptome
639 analysis of wheat grain using RNA-Seq. Frontiers of Agricultural Science and Enginnering.
640 2014;1(3):214-22. doi:10.15302/j-fase-2014024.
641 39. IWGSC RefSeq v1.0. https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies.
642 40. IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*)
643 genome. Science. 2014;345(6194):1251788. doi:10.1126/science.1251788.
644 41. Clavijo BJ, Venturini L, Schudoma C, Garcia Accinelli G, Kaithakottil G, Wright J et al. An improved
645 assembly and annotation of the allohexaploid wheat genome identifies complete families of
646 agronomic genes and provides genomic evidence for chromosomal translocations. Genome
647 Research. 2017;27:885-96. doi:10.1101/gr.217117.116.
648 42. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the
649 hexaploid bread wheat genome, *Triticum aestivum*. bioRxiv. 2017. doi:10.1101/159111.
650 43. Cantu D, Pearce SP, Distelfeld A, Christiansen MW, Uauy C, Akhunov E et al. Effect of the down-
651 regulation of the high *Grain Protein Content* (*GPC*) genes on the wheat transcriptome during
652 monocarpic senescence. BMC Genomics. 2011;12. doi:10.1186/1471-2164-12-492.
653 44. Barrero JM, Cavanagh C, Verbyla KL, Tibbits JFG, Verbyla AP, Huang BE et al. Transcriptomic
654 analysis of wheat near-isogenic lines identifies *PM19-A1* and *A2* as candidates for a major dormancy
655 QTL. Genome Biology. 2015;16(1):93. doi:10.1186/s13059-015-0665-6.
656 45. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq
657 incorporating quantification uncertainty. Nat Meth. 2017;14(7):687-90. doi:10.1038/nmeth.4324.
658 46. Borrill P, Harrington SA, Uauy C. Genome-Wide Sequence and Expression Analysis of the NAC
659 Transcription Factor Family in Polyploid Wheat. G3: Genes|Genomes|Genetics. 2017.
660 doi:10.1534/g3.117.043679.
661 47. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature.
662 2012;482(7385):339-46.

663 48. Sun F, Guo G, Du J, Guo W, Peng H, Ni Z et al. Whole-genome discovery of miRNAs and their
664 targets in wheat (*Triticum aestivum L.*). BMC Plant Biology. 2014;14(1):142. doi:10.1186/1471-2229-
665 14-142.
666 49. Disch S, Anastasiou E, Sharma VK, Laux T, Fletcher JC, Lenhard M. The E3 ubiquitin ligase BIG
667 BROTHER controls Arabidopsis organ size in a dosage-dependent manner. Curr Biol. 2006;16(3):272-
668 9. doi:10.1016/j.cub.2005.12.026.
669 50. Dong H, Dumenil J, Lu FH, Na L, Vanhaeren H, Naumann C et al. Ubiquitylation activates a
670 peptidase that promotes cleavage and destabilization of its activating E3 ligases and diverse growth
671 regulatory proteins to limit cell proliferation in Arabidopsis. Genes Dev. 2017;31(2):197-208.
672 doi:10.1101/gad.292235.116.
673 51. Du L, Li N, Chen L, Xu Y, Li Y, Zhang Y et al. The Ubiquitin Receptor *DA1* Regulates Seed and Organ
674 Size by Modulating the Stability of the Ubiquitin-Specific Protease *UBP15/SOD2* in *Arabidopsis*. The
675 Plant Cell. 2014;26(2):665-77. doi:10.1105/tpc.114.122663.
676 52. Huang K, Wang D, Duan P, Zhang B, Xu R, Li N et al. *WIDE AND THICK GRAIN 1*, which encodes an
677 otubain-like protease with deubiquitination activity, influences grain size and shape in rice. The Plant
678 Journal. doi:10.1111/tpj.13613.
679 53. Cheng Y, Qin G, Dai X, Zhao Y. *NPY1*, a BTB-NPH3-like protein, plays a critical role in auxin-
680 regulated organogenesis in *Arabidopsis*. Proceedings of the National Academy of Sciences.
681 2007;104(47):18825-9. doi:10.1073/pnas.0708506104.
682 54. Radchuk V, Weier D, Radchuk R, Weschke W, Weber H. Development of maternal seed tissue in
683 barley is mediated by regulated cell expansion and cell disintegration and coordinated with
684 endosperm growth. Journal of Experimental Botany. 2011;62(3):1217-27. doi:10.1093/jxb/erq348.
685 55. Drea S, Leader DJ, Arnold BC, Shaw P, Dolan L, Doonan JH. Systematic spatial analysis of gene
686 expression during wheat caryopsis development. The Plant cell. 2005;17(8):2172-85.
687 doi:10.1105/tpc.105.034058.

688 56. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR et al. Genome interplay in the grain
689 transcriptome of hexaploid bread wheat. Science. 2014;345(6194):1250091.
690 doi:10.1126/science.1250091.
691 57. Dominguez F, Cejudo FJ. Characterization of the Endoproteases Appearing during Wheat Grain
692 Development. Plant Physiology. 1996;112(3):1211-7. doi:10.1104/pp.112.3.1211.
693 58. Kaspar-Schoenefeld S, Merx K, Jozefowicz AM, Hartmann A, Seiffert U, Weschke W et al. Label-
694 free proteome profiling reveals developmental-dependent patterns in young barley grains. Journal
695 of Proteomics. 2016;143:106-21. doi:http://dx.doi.org/10.1016/j.jprot.2016.04.007.
696 59. Pires JC, Conant GC. Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage
697 in the Evolution of Genomes. Annu Rev Genet. 2016;50:113-31. doi:10.1146/annurev-genet-120215-
698 035400.
699 60. Yang M, Gao X, Dong J, Gandhi N, Cai H, von Wettstein DH et al. Pattern of Protein Expression in
700 Developing Wheat Grains Identified through Proteomic Analysis. Front Plant Sci. 2017;8:962.
701 doi:10.3389/fpls.2017.00962.
702 61. Li J, Jiang J, Qian Q, Xu Y, Zhang C, Xiao J et al. Mutation of Rice *BC12/GDD1*, Which Encodes a
703 Kinesin-Like Protein That Binds to a GA Biosynthesis Gene Promoter, Leads to Dwarfism with
704 Impaired Cell Elongation. The Plant Cell. 2011;23(2):628-40. doi:10.1105/tpc.110.081901.
705 62. Torada A, Koike M, Ogawa T, Takenouchi Y, Tadamura K, Wu J et al. A Causal Gene for Seed
706 Dormancy on Wheat Chromosome 4A Encodes a MAP Kinase Kinase. Curr Biol. 2016;26(6):782-7.
707 doi:http://dx.doi.org/10.1016/j.cub.2016.01.063.
708 63. Shorinola O, Balcarkova B, Hyles J, Tibbits JFG, Hayden MJ, Holusova K et al. Association mapping
709 and haplotype analysis of the pre-harvest sprouting resistance locus *Phs-A1* reveals a causal role of
710 *TaMKK3-A* in global germplasm. bioRxiv. 2017. doi:10.1101/131201.
711 64. Kaplan-Levy RN, Brewer PB, Quon T, Smyth DR. The trihelix family of transcription factors – light,
712 stress and development. Trends in Plant Science. 2012;17(3):163-71.
713 doi:http://dx.doi.org/10.1016/j.tplants.2011.12.002.

714 65. Barg R, Sobolev I, Eilon T, Gur A, Chmelnitsky I, Shabtai S et al. The tomato early fruit specific
715 gene *Lefsm1* defines a novel class of plant-specific SANT/MYB domain proteins. Planta.
716 2005;221(2):197-211. doi:10.1007/s00425-004-1433-0.
717 66. Sarojam R, Sappl PG, Goldshmidt A, Efroni I, Floyd SK, Eshed Y et al. Differentiating Arabidopsis
718 Shoots from Leaves by Combined YABBY Activities. The Plant Cell. 2010;22(7):2113-30.
719 doi:10.1105/tpc.110.075853.
720 67. Cong B, Barrero LS, Tanksley SD. Regulatory change in YABBY-like transcription factor led to
721 evolution of extreme fruit size during tomato domestication. Nature Genetics. 2008;40(6):800-4.
722 doi:http://www.nature.com/ng/journal/v40/n6/suppinfo/ng.144_S1.html.
723 68. Dinneny JR, Weigel D, Yanofsky MF. *NUBBIN* and *JAGGED* define stamen and carpel shape in
724 *Arabidopsis*. Development. 2006;133(9):1645-55. doi:10.1242/dev.02335.
725 69. Jin Y, Luo Q, Tong H, Wang A, Cheng Z, Tang J et al. An AT-hook gene is required for palea
726 formation and floral organ number control in rice. Developmental Biology. 2011;359(2):277-88.
727 doi:http://dx.doi.org/10.1016/j.ydbio.2011.08.023.
728 70. Gallavotti A, Malcomber S, Gaines C, Stanfield S, Whipple C, Kellogg E et al. *BARREN STALK*
729 *FASTIGIATE1* is an AT-Hook Protein Required for the Formation of Maize Ears. The Plant Cell.
730 2011;23(5):1756-71. doi:10.1105/tpc.111.084590.
731 71. Street IH, Shah PK, Smith AM, Avery N, Neff MM. The AT-hook-containing proteins *SOB3/AHL29*
732 and *ESC/AHL27* are negative modulators of hypocotyl growth in Arabidopsis. The Plant Journal.
733 2008;54(1):1-14. doi:10.1111/j.1365-313X.2007.03393.x.
734 72. Cosgrove DJ. Growth of the plant cell wall. Nat Rev Mol Cell Bio. 2005;6(11):850-61. doi:DOI
735 10.1038/nrm1746.
736 73. Dante RA, Larkins BA, Sabelli PA. Cell cycle control and seed development. Front Plant Sci.
737 2014;5:493. doi:10.3389/fpls.2014.00493.

738 74. Lee BH, Ko J-H, Lee S, Lee Y, Pak J-H, Kim JH. The Arabidopsis *GRF-INTERACTING FACTOR* Gene
739 Family Performs an Overlapping Function in Determining Organ Size as Well as Multiple
740 Developmental Properties. Plant Physiology. 2009;151(2):655-68. doi:10.1104/pp.109.141838.
741 75. Bao F, Azhakanandam S, Franks RG. *SEUSS* and *SEUSS-LIKE* Transcriptional Adaptors Regulate
742 Floral and Embryonic Development in *Arabidopsis*. Plant Physiology. 2010;152(2):821-36.
743 doi:10.1104/pp.109.146183.
744 76. Adamski NM, Anastasiou E, Eriksson S, O'Neill CM, Lenhard M. Local maternal control of seed
745 size by *KLUH/CYP78A5*-dependent growth signaling. Proceedings of the National Academy of
746 Sciences. 2009;106(47):20115-20. doi:10.1073/pnas.0907024106.
747 77. Locascio A, Roig-Villanova I, Bernardi J, Varotto S. Current perspectives on the hormonal control
748 of seed development in Arabidopsis and maize: a focus on auxin. Front Plant Sci. 2014;5:412.
749 doi:10.3389/fpls.2014.00412.
750 78. Li N, Li Y. Ubiquitin-mediated control of seed size in plants. Front Plant Sci. 2014;5:332.
751 doi:10.3389/fpls.2014.00332.
752 79. Yang Z, Bai Z, Li X, Wang P, Wu Q, Yang L et al. SNP identification and allelic-specific PCR markers
753 development for *TaGW2*, a gene linked to wheat kernel weight. Theoretical and Applied Genetics.
754 2012;125(5):1057-68. doi:10.1007/s00122-012-1895-6.
755 80. Hershko A, Ciechanover A. The ubiquitin system. Annu Rev Biochem. 1998;67:425-79.
756 doi:10.1146/annurev.biochem.67.1.425.
757 81. Vanhaeren H, Nam Y-J, De Milde L, Chae E, Storme V, Weigel D et al. Forever Young: The Role of
758 Ubiquitin Receptor *DA1* and E3 Ligase *BIG BROTHER* in Controlling Leaf Growth and Development.
759 Plant Physiology. 2016. doi:10.1104/pp.16.01410.
760 82. Kurepa J, Wang S, Li Y, Zaitlin D, Pierce AJ, Smalle JA. Loss of 26S Proteasome Function Leads to
761 Increased Cell Size and Decreased Cell Number in Arabidopsis Shoot Organs. Plant Physiology.
762 2009;150(1):178-89. doi:10.1104/pp.109.135970.

763 83. Weng J, Gu S, Wan X, Gao H, Guo T, Su N et al. Isolation and initial characterization of *GW5*, a
764 major QTL associated with rice grain width and weight. Cell Research. 2008;18(12):1199-209.
765 84. Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L et al. Uncovering hidden
766 variation in polyploid wheat. Proceedings of the National Academy of Sciences. 2017;114(6):E913-
767 E21. doi:10.1073/pnas.1619268114.
768 85. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.
769 http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed September 9 2015.
770 86. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat
771 Biotech. 2016;34(5):525-7. doi:10.1038/nbt.3519.
772 87. Borrill P, Ramirez-Gonzalez R, Uauy C. expVIP: a Customizable RNA-seq Data Analysis and
773 Visualization Platform. Plant Physiology. 2016;170(4):2172-86.
774 88. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
775 Bioinformatics. 2010;26(6):841-2. doi:10.1093/bioinformatics/btq033.
776 89. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting
777 for selection bias. Genome Biology. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14.
778 90. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics.
779 2011;27(7):1017-8. doi:10.1093/bioinformatics/btr064.
780 91. Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y et al. PlantPAN 2.0: an update of
781 plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants.
782 Nucleic Acids Research. 2016;44(Database issue):D1154-D60. doi:10.1093/nar/gkv1035.
783 92. Peng FY, Hu Z, Yang R-C. Bioinformatic prediction of transcription factor binding sites at
784 promoter regions of genes for photoperiod and vernalization responses in model and temperate
785 cereal plants. BMC Genomics. 2016;17:573. doi:10.1186/s12864-016-2916-7.

Figure legends

Figure 1: Differentially expressed genes between 5A NILs across time.

RNA-seq was carried out on whole grain RNA samples taken in 4 different conditions: 5A- (short grains) and 5A+ (long grains) NILs at 4 days post anthesis (dpa; T1) and 8 dpa (T2). These were selected as the time point when the first significant difference ($P < 0.01$, asterisks) in grain length was observed between 5A- (grey, dashed line, short grains) and 5A+ (purple, solid line, long grains) and the preceding time point. Differentially expressed (DE) transcripts were identified for four comparisons (q-value < 0.05). Coloured boxes indicate the numbers of DE transcripts identified for each comparison using alignments to either the IWGSC Chinese Spring Survey Sequence (CSS) or the TGACv1 (TGAC) Chinese Spring reference transcriptomes. Two 'across time' comparisons: 5A- $^{T1}_{T2}$ (grey box; comparing T1 and T2 samples of the 5A- NIL) and 5A+ $^{T1}_{T2}$ (purple box; comparing T1 and T2 samples of the 5A+ NIL), and two 'between NIL' comparisons: T1 $^{5A-}_{5A+}$ (orange box; comparing 5A- and 5A+ NILs at T1) and T2 $^{5A-}_{5A+}$ (green box; comparing 5A- and 5A+ NILs at T2).

Figure 2 Comparison between CSS and TGACv1 gene models

a) Discrepancies identified between gene models in the CSS and TGAC reference sequences and the number of gene models falling into categories. Panels b), c) and d) show specific examples of discrepancies. In each panel, a representation of the unspliced gene model is shown with exons as coloured boxes, untranslated regions as white boxes, and introns as thin lines. Graphs show the relative read coverage across the spliced transcript with the structure represented diagrammatically directly above each graph. The number in brackets shows the maximum absolute read depth for each gene model. > and < in the gene structures indicate the direction of transcription and a 'DE' indicates that the gene model was differentially expressed in T1 $^{5A-}_{5A+}$ (q value < 0.05). For each panel transcript names are shown in the coloured legends.

Figure 3: Overview of differentially expressed transcripts

a) Venn diagram of differentially expressed (DE) transcripts ($q < 0.05$) identified in 4 pairwise comparisons: T1 $^{5A-}_{5A+}$ (orange), T2 $^{5A-}_{5A+}$ (green), 5A- $^{T1}_{T2}$ (grey) and 5A+ $^{T1}_{T2}$ (purple). b) Number of DE transcripts located on each chromosome for all comparisons. The 5A- $^{T1}_{T2}$ and 5A+ $^{T1}_{T2}$ DE transcripts (top graphs) are evenly distributed across all 21 chromosomes whereas T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ DE transcripts (bottom graphs) are concentrated on chromosome 5A. c) Heatmap of normalised tpm (transcripts per million) of common DE transcripts in 5A- $^{T1}_{T2}$ and 5A+ $^{T1}_{T2}$ (n = 1,832). Hierarchical clustering separated these into transcripts that were upregulated (n = 1,532) and downregulated (n = 300) across time. Significantly enriched GO terms (biological function only) for each group are shown on the right of the heatmap.

Figure 4: Differentially expressed transcripts between 5A NILs at T1 and T2

a) Heatmap of normalised tpm (transcripts per million) of DE (differentially expressed) transcripts between NILs (T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ comparisons). Transcripts are first grouped based on whether they were differentially expressed at both time points (T1 $^{5A-}_{5A+}$ and T2 $^{5A-}_{5A+}$ common) or at only T1 or T2 (T1 $^{5A-}_{5A+}$ unique and T2 $^{5A-}_{5A+}$ unique, respectively), and then whether they are located on chromosome 5A or not. b) Location of DE transcripts on chromosome 5A (black lines on grey rectangle). Line graph (blue) shows rolling mean of the number of transcripts located in 3 Mbp bins across chromosome 5A, alongside heatmap which shows the number of 90k iSelect SNPs between the 5A- and 5A+ NILs in 3 similar sized bins. Orange lines on the SNP heatmap define the 491 Mbp introgression which differs between then NILs. Red lines on the chromosome indicate the positions of the flanking markers of the fine-mapped region of the 5A grain length QTL (BS00182017 and BA00228977). Bar charts show the mean tpm values at T1 and T2 of DE transcripts located in the fine mapped region

(5A- NILs in grey, 5A+ NILs in purple). Only one transcript variant (.2) of the kinesin-like gene is shown. Error bars are standard error of the three biological replicates.

Figure 5: Differential regulation of the ubiquitin pathway in 5A NILs

a) Differentially expressed (DE) transcripts with functional annotations related to ubiquitin-mediated protein turnover were enriched relative to the whole genome (a). This pathway acts to add multiple copies of the protein Ubiquitin (Ub) to a substrate protein through the sequential action of a cascade of three enzymes: E1 (Ub-activating enzymes), E2 (Ub-conjugating enzymes) and E3 (Ub ligases). The tagged substrate is then targeted for degradation by the 26S proteasome and the Ub proteins are recycled. The E3 ligases are the most diverse of the three enzymes and both single subunit proteins and multi-subunit complexes exist. A subset of these classes is shown in the grey box in (a), selected based on the annotations of DE transcripts. Single subunit E3 ligases have an E2-interacting domain (e.g. U-box, RING, etc. (...)) and a substrate-recognising domain. Multi-subunit complexes also have E2-interacting complexes and substrate-recognising subunits (e.g. F-box, BTB, etc. (...)). In the context of organ size control, some proteases have been identified as downstream targets of this pathway (e.g. DA1, UBP15 [50, 51]). b) Heatmap of normalised tpm of DE transcripts associated with ubiquitin, the proteasome and proteases.

Tables

Table 1 Mapping summary of RNA-seq samples

Genotype	Time point	Replicate	Reads	CSS gene models		TGAC gene models	
				Reads pseudoaligned	% reads pseudoaligned	Reads pseudoaligned	% reads pseudoaligned
5A -	1	1	24443658	17072939	69.85	20549681	84.07
5A -	1	2	34441799	23349288	67.79	28483090	82.70
5A -	1	3	23462705	16220597	69.13	19664859	83.81
5A -	2	1	21333672	14839724	69.56	18052324	84.62
5A -	2	2	14967302	10632519	71.04	12803552	85.54
5A -	2	3	35522754	25491523	71.76	30297336	85.29
5A +	1	1	19267564	13520181	70.17	16317352	84.69
5A +	1	2	22299102	15479234	69.42	18780525	84.22
5A +	1	3	30531539	20789582	68.09	25436453	83.31
5A +	2	1	51637607	36192489	70.09	43739451	84.70
5A +	2	2	53575232	37956887	70.85	45497914	84.92
5A +	2	3	30553421	21604895	70.71	25984674	85.05
		Total	362036355	253149858	-	305607211	-
		Mean	30169696	21095822	69.87	25467268	84.41

Table 2: Enriched transcription factor binding sites in promoters of DE transcripts located outside of 5A

Values are the number of transcripts in which binding sites associated with the specified transcription factor (TF) family are present.

TF family	Observed in all expressed transcripts (n=101,653)	Expected in outside 5A DE transcript (n=38)	Observed in outside 5A DE transcripts (n=38)	FDR adjusted p-value
C2H2	77987	29	36	0.021
Myb/SANT	88575	33	38	0.021
AT-Hook	90203	34	38	0.028
YABBY	19447	7	15	0.034
MADF;Trihelix	16632	6	13	0.042

858 **Table 3: Categories of DE transcripts between NILs based on predicted function**

859 Adjusted p-values displayed are based on an enrichment test of the functional categories relative to
860 all expressed transcripts. - indicates that an enrichment test was not performed as categories were
861 based on bespoke annotations. * includes transcripts with annotations that could not be grouped by
862 function with other transcripts. ** only the 7 transcripts that were annotated as ubiquitin-related in
863 the TGAC annotation were used in the enrichment test (see methods).

Category	number of transcripts	Adjusted p-value	5A/not 5A	NIL with higher expression: 5A-/5A+
non-coding RNA	15	0.141	10/5	6/9
transposon-associated	14	0.008	4/10	5/9
ubiquitin	12**	0.008	10/2	8/4
cell cycle	5	-	5/0	2/3
histone-related	5	-	3/2	3/2
heat shock	5	-	3/2	2/3
protease	4	-	3/1	3/1
transport	4	-	3/1	2/2
metabolism	5	-	5/0	4/1
homeobox	4	0.001	3/1	1/3
cell wall	3	-	2/1	2/1
transcription	3	-	2/1	0/3
non-translating	2	-	0/2	1/1
peroxisome	2	-	0/2	0/2
other*	20	-	14/6	11/9
No annotation	8	-	4/4	5/3

Additional files

Additional file 1.xlsx - **Comparison of CSS and TGAC gene models.**

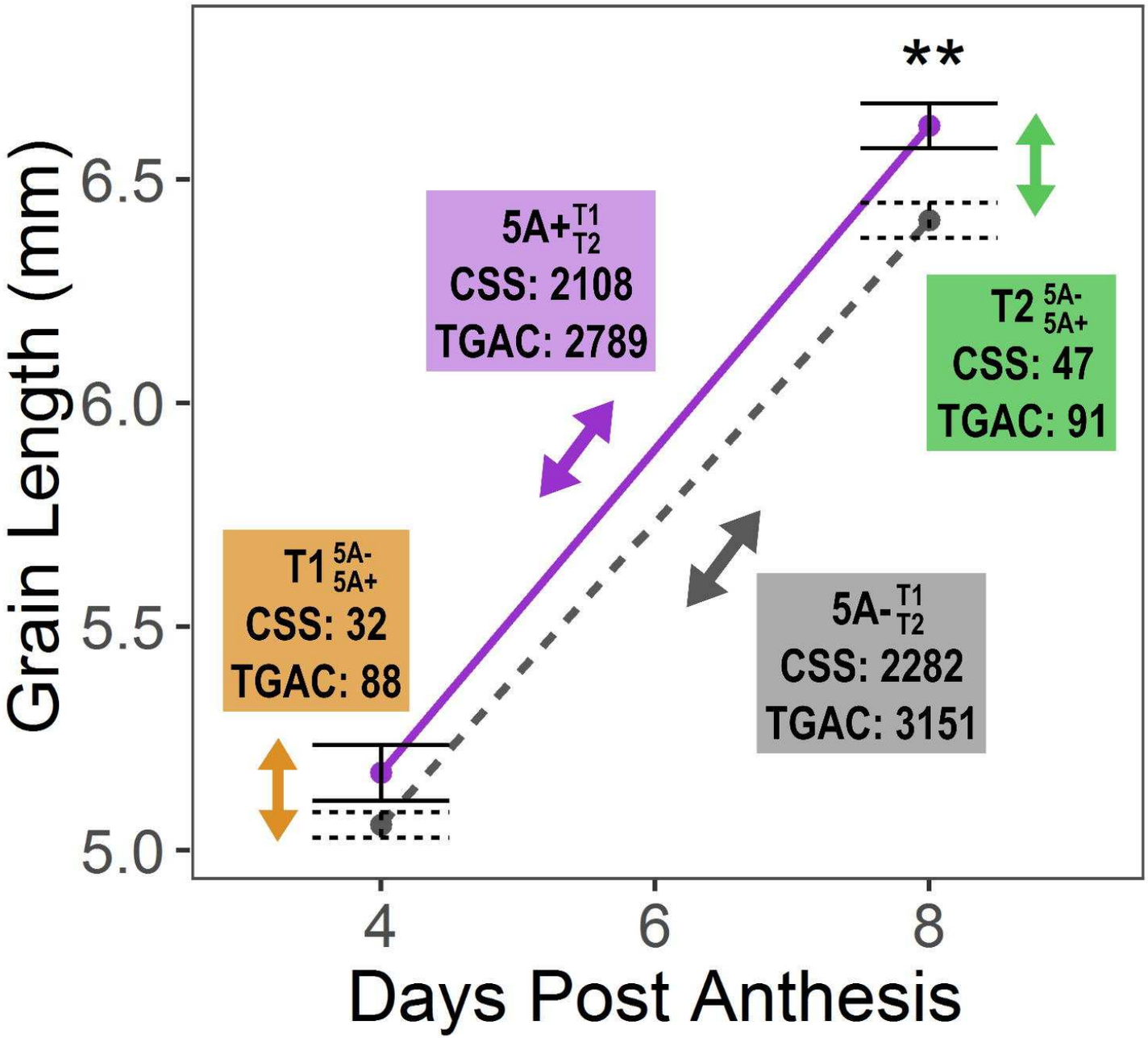
Additional file 2.xlsx – **Enriched GO terms in the across time comparisons**

Additional file 3.docx – **q-value distributions of uniquely differentially expressed transcripts across time**

Additional file 4.xlsx – **Transcription factor binding sites identified in outside 5A DE transcripts**

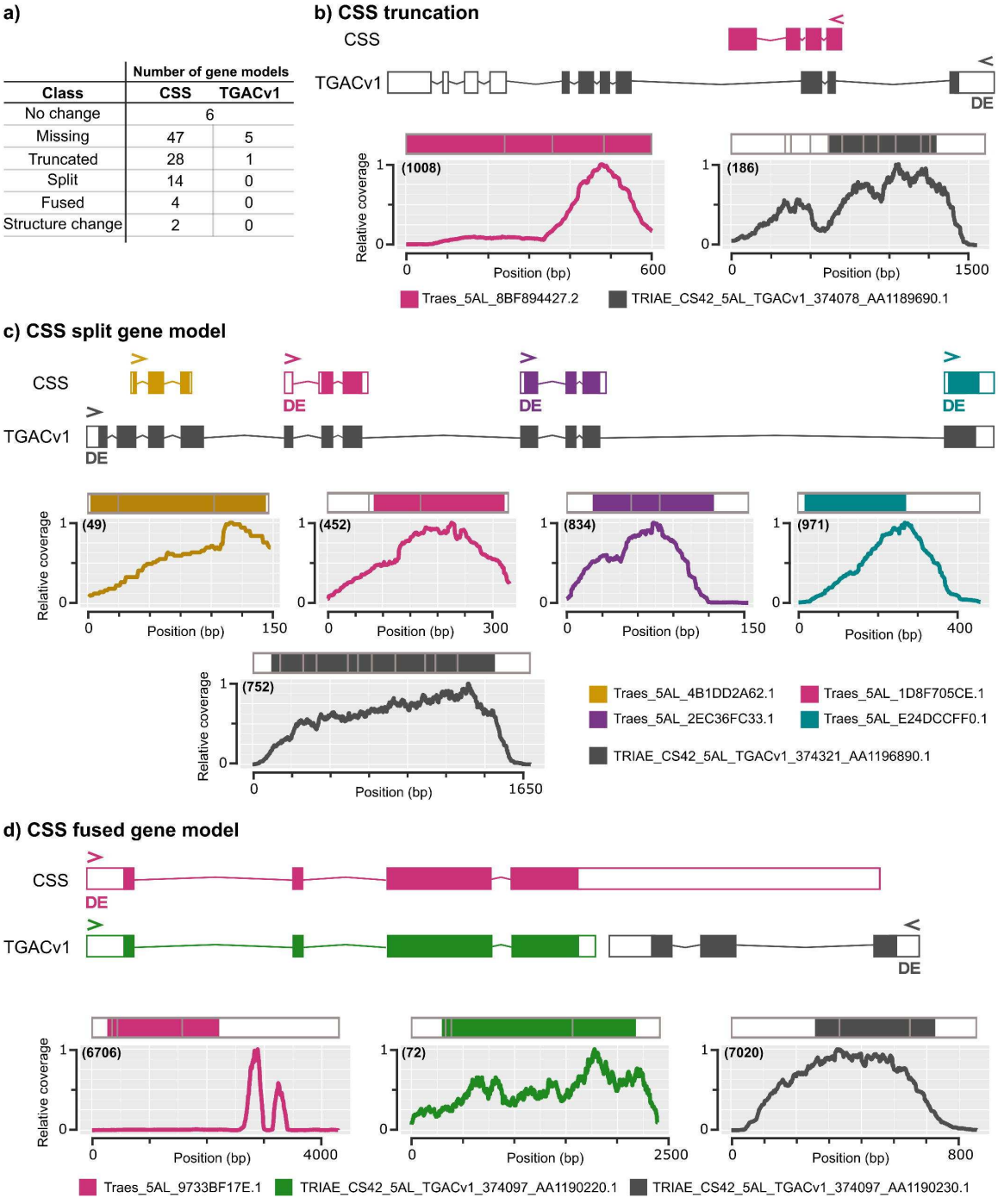
Additional file 5.xlsx – **Transcription factors present in the 5A introgression**

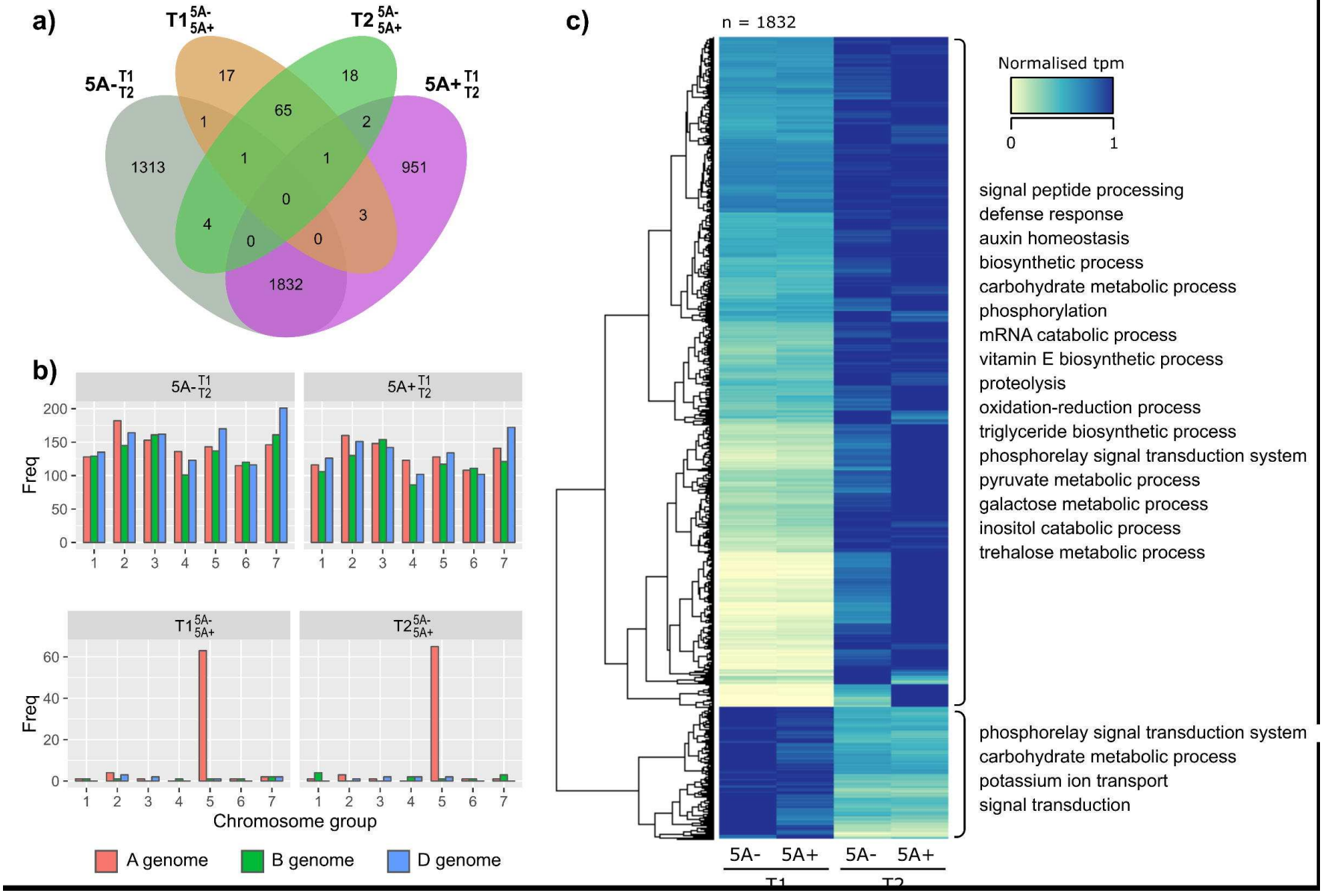
Additional file 6.xlsx – **Functional annotation of DE transcripts between NILs**

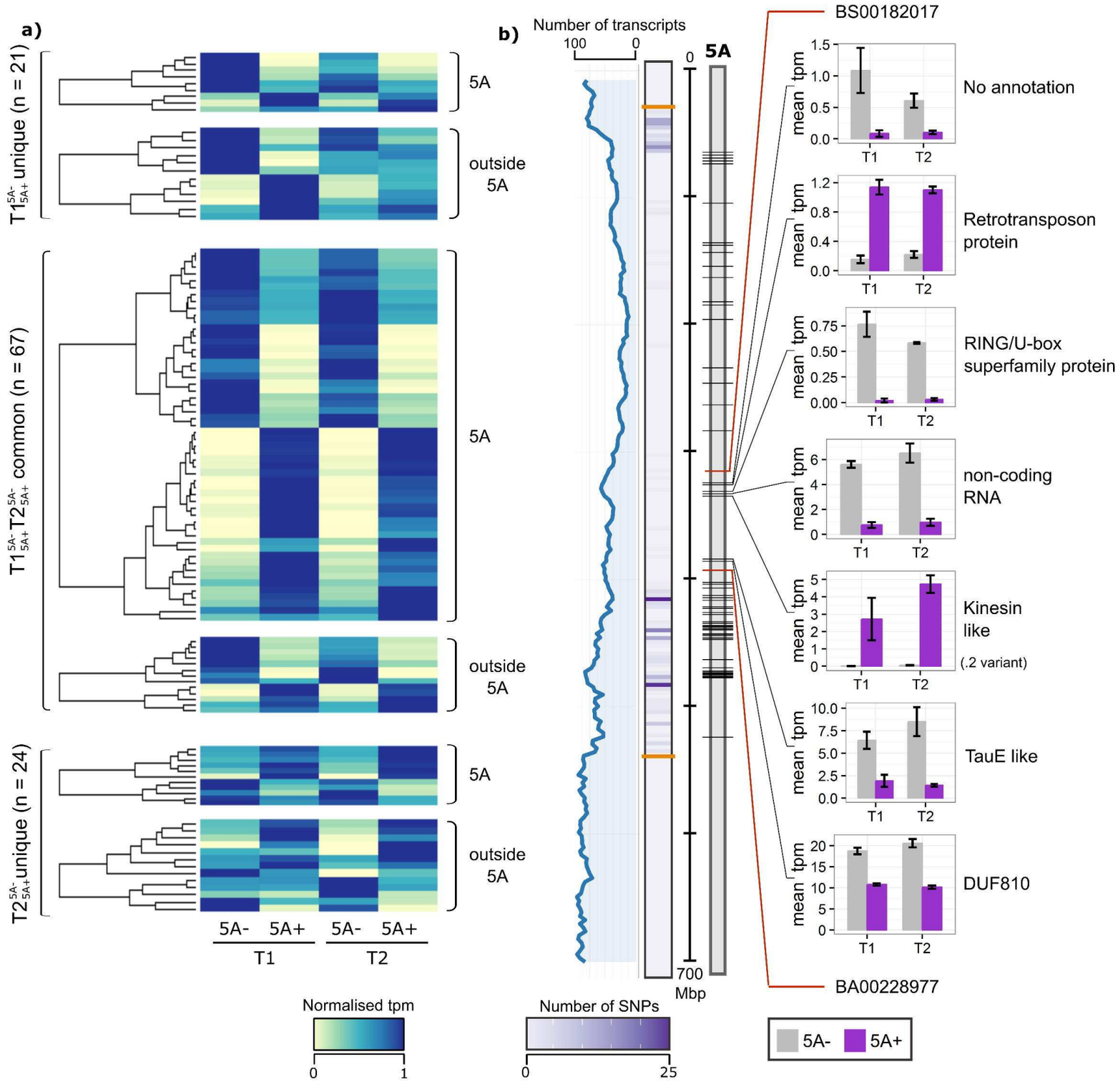

**
Grain Length (mm)
6.5
6.0
5.5
5.0
4
6
8
Days Post Anthesis
5A+ T1 T2
CSS: 2108
TGAC: 2789
T2 5A- 5A+
CSS: 47
TGAC: 91
T1 5A- 5A+
CSS: 32
TGAC: 88
5A- T1 T2
CSS: 2282
TGAC: 3151

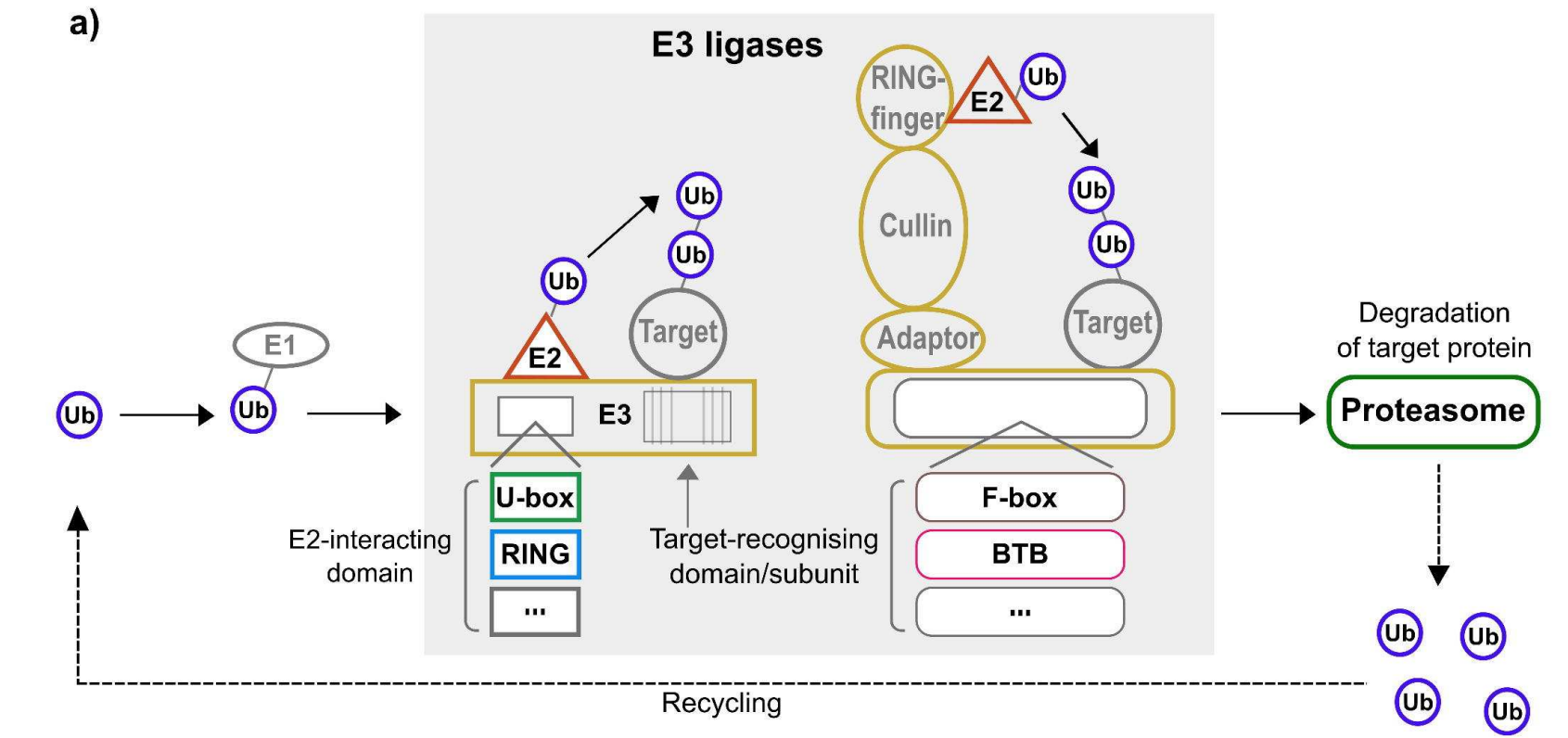
a)

Class	Number of gene models	
	CSS	TGACv1
No change	6	
Missing	47	5
Truncated	28	1
Split	14	0
Fused	4	0
Structure change	2	0




a)
T1 5A- 5A+
T2 5A- 5A+
5A- T1 T2
5A+ T1 T2
17
18
65
1
2
1313
1
1
951
0
3
4
0
0
1832
b)
5A- T1 T2
5A+ T1 T2
Freq
200
150
100
50
0
1 2 3 4 5 6 7
T1 5A- 5A+
T2 5A- 5A+
60
40
20
0
Chromosome group
A genome
B genome
D genome
c)
n = 1832
Normalised tpm
0
1
signal peptide processing
defense response
auxin homeostasis
biosynthetic process
carbohydrate metabolic process
phosphorylation
mRNA catabolic process
vitamin E biosynthetic process
proteolysis
oxidation-reduction process
triglyceride biosynthetic process
phosphorelay signal transduction system
pyruvate metabolic process
galactose metabolic process
inositol catabolic process
trehalose metabolic process
phosphorelay signal transduction system
carbohydrate metabolic process
potassium ion transport
signal transduction
5A- 5A+ 5A- 5A+
T1
T2


a)
b)
T1 5A- 5A+ unique (n = 21)
T1 5A- 5A+ T2 5A- 5A+ common (n = 67)
T2 5A- 5A+ unique (n = 24)
5A
outside 5A
5A- 5A+ 5A- 5A+
T1
T2
Normalised tpm
0
1
Number of transcripts
100
0
5A
0
700
Mbp
Number of SNPs
0
25
BS00182017
BA00228977
mean tpm
No annotation
Retrotransposon protein
RING/U-box superfamily protein
non-coding RNA
Kinesin like
(.2 variant)
TauE like
DUF810
5A-
5A+


a)
E3 ligases
Ub
E1
E2
Target
E3
U-box
RING
...
E2-interacting domain
Target-recognising domain/subunit
RING-finger
Cullin
Adaptor
F-box
BTB
Degradation of target protein
Proteasome
Recycling
b)
Normalised tpm
0
1
E3
Ub
E2
U-box
RING
F-box
BTB
Proteasome
Ub -like
Proteases
5A-
T1
T2
5A+
T1
T2