Exploring genetic loci for the improvement of inflorescence traits in the polyploid cereals wheat and tef

Maximillian R. W. Jones

A thesis submitted to the University of East Anglia

for the degree of Doctor of Philosophy

John Innes Centre

May 2025

Abstract

Genetic gains in both major and minor crops will be essential to meet the human population's future agricultural demands. In this thesis we explore how different types of variation for inflorescence traits – natural and engineered, recessive and dominant, coding and *cis*-regulatory – can contribute to yield for two polyploid cereal crops; wheat and tef (*Eragrostis tef*).

We initially describe resequencing and multi-location phenotyping of an important tef germplasm collection. Through *k*-mer-based genome-wide association we uncovered multiple marker-trait associations for inflorescence and grain traits. This included identification of a strong link between grain size and grain colour. We proposed the tef orthologues of *TRANSPARENT TESTA 2* as key regulators of these traits and detected natural variation in both homoeologs.

Next, we moved to wheat for an opportunity to study induced variation. The spikelets (grain-bearing inflorescence structures) of wheat are not created equally. The basal-most spikelets initiate first but quickly fall behind their central counterparts in development, ultimately producing smaller and fewer (or even zero) grains. Little natural variation has been identified which alters the resulting grain mass distribution across the spike. We conducted semi-spatial transcriptomics across a developmental time course of early wheat spikes and identified potential regulators of this process. We then manipulated two candidate genes via transgenic misexpression, and preliminary results suggest that one construct reduced the incidence of basal spikelets bearing no viable grain.

These results led us to conclude that reliable tissue and time-specific *cis*-regulomes are a powerful resource for testing developmental hypotheses and inducing novel, breeding-relevant variation. We pursued this goal by attempting to generate a genome-wide catalogue of *cis*-regulatory elements (CREs) relevant to early wheat spike development. Our methodology was to correlate chromatin accessibility with gene expression, integrating DNA hypomethylation as another CRE marker. This approach exhibited some capacity to detect enhancers, but not silencers.

Acknowledgements

Firstly, I am immensely grateful to my supervisor, Cristóbal Uauy. You have the most drive and the sharpest mind of anyone I have met, yet manage to combine this with a deep sense of care for your students and staff. Thank you for teaching me how to manage scientific projects, for supporting me when my confidence was in pieces, and for finally getting through to me that 'perfect is the enemy of good'. Having heroes seems to be out of fashion in our individualistic world, but I certainly place you among mine.

Many thanks also to my supervisory team, Ricardo and Steve, for offering valuable advice throughout the PhD process, and to Oluwaseyi, who became my close collaborator and an unofficial extra supervisor during our work on *Eragrostis tef* – it was a pleasure to work with you on this beautiful underutilised crop. Thanks also to Ash, Abel, Worku, Adanech, and the whole team involved in the tef GWAS work. I'm also highly grateful to all of the JIC platforms and staff that made my work possible, especially Horticultural Services, Wheat Transformation, Bioimaging, Laboratory Support, and Genotyping.

To the whole of the Uauy Lab past and present, thank you for being warm and welcoming from day one and for becoming great friends. I am incredibly lucky to have met so many wonderful people by being part of this group. The lab retreats, away days, and barbeques will be remembered fondly – and not just for the diverse, amazing food! To Nikolai and James, thank you for your support with all things lab and horticulture. So much of this thesis would not have been possible without your advice and direct input, and more still without your pep talks and reassurances. Huge thanks also to Pam and Sophie for help with crossing, harvesting, and threshing. Anna, Aura, Jesús, Andy, Marina, and Sophie, thank you for being fun, kind people to work with and for being inspiring senior PhD students to look up to. To the Biffen tearoom squad, thanks for creating a great space to decompress over lunch and have some whacky and inspiring conversations. To my PhD siblings, Isa and Katie, thanks for being there through all my highs and lows, being sounding boards for ideas and hypotheses, and just generally making work a fun place to be! I can't wait to hear about (and be proud of) what you both accomplish in the years to come.

Thanks also to Isa for being a wonderful housemate at both 452 and 146, alongside Liam, Sean, Daniella, and many others. Coming home to shared meals, movies, and board games has been a large part of making Norwich feel like home over the last four and a half years. Thank you to my friends outside of Norwich – Bradley, Smithy, Mark, Boo, and the Molesworth / Old Weston gang – for bearing with me and for lending me different perspectives on work and life. To Katie Lightfoot, thank you for everything you have taught me.

To my wonderful Steph, thank you for being my best friend and partner. It's a joy to have so much in common with you, from academia to climbing to a love of nature and the outdoors. I'm so grateful for all your help and encouragement during my PhD journey and I can't wait for the adventures we will have together in the future.

Lastly, none of this would have been possible without the love and patient support of my family throughout my life. Thank you so much Bill, Bunny, Eric, and Freda, and especially Mum, Dad, and Ben. I love you to the moon and back.

Table of contents

Table of figures

Table of tables

Supplementary data statement

Supplementary data, tables, and notes referenced in the text are made available at https://zenodo.org/records/15296805. Additional raw data not specifically quoted are also made available here.

Custom R and BASH scripts utilised in Chapter 4 are made available at https://github.com/maxrwjones/wheat_spike_ATACseq_RNAseq_correlation.

Abbreviations

3C	Chromosome conformation capture
ACR	Accessible chromatin region
ATAC-seq	Assay for transposase-accessible chromatin-seq
BLUP	Best linear unbiased predictor
BSS1	Basal Spike Specific 1 regulatory environment
BSS2	Basal Spike Specific 2 regulatory environment
CACR	Consensus ACR
CE	Carpel extension
CER	Controlled environment room
CN	Copy number
CNV	Copy number variation
CRE	*cis*-regulatory element
cv.	Cultivar
CV	Coefficient of variation
CYP93G1	*Cytochrome P450 93G1*
DE	Differentially expressed
DEG	Differentially expressed gene
EDR	Early double ridge
EGS	Effective genome size
EIAR	Ethiopian Institute of Agricultural Research
EMSA	Electrophoretic mobility shift assay
EPOD	9,10-epoxyoctadecanoic acid
FC	Fold-change
FIE-HRMS	Flow infusion electrospray high-resolution mass spectrometry
FM	Floral meristem
GE	Genome editing
GP	Glume primordia
GUS	β-glucuronidase
GW2	*GRAIN WIDTH and WEIGHT 2*
H^2	Broad-sense heritability
h^2	Narrow-sense heritability
HC	High confidence (gene model)
IGV	Integrative genomics viewer
IM	Inflorescence meristem
JIC	The John Innes Centre
KASP	Kompetitive allele specific PCR
KOROG	Kaempferol 3-O-rhamnoside-7-O-glucoside
LC	Low confidence (gene model)
LCM	Laser-capture microdissection

LD	Linkage disequilibrium
LDR	Late double ridge
LP	Lemma primordia
LTR	Long terminal repeat
MOF1	*MORE FLORET 1*
MTA	Marker-trait association
NBI	Norwich Bioscience Institutes
NF-YA3	*NUCLEAR FACTOR YA3*
NIL	Near-isogenic line
NLS	Nuclear localisation signal
OsBZR1	*Oryza sativa BRASSINAZOLE RESISTANT 1*
PCA	Principal component analysis
PE	Prime editing
PFCC	Primary fluorescent chlorophyll catabolite
PLS-DA	Partial least squares discriminant analysis
QTL	Quantitative trait loci
RBS	Rudimentary basal spikelets
RT	Retrotransposon
SBE	*Starch branching enzyme*
SD	Standard deviation
SEP	*SEPALLATA*
SM	Spikelet meristem
SS	Sum of squares
SVP	*SHORT VEGETATIVE PHASE*
TE	Transposable element
TF	Transcription factor
TGW	Thousand grain weight
TILLING	Targeting induced local legions in genomes
TPM	Transcripts per kilobase million
T-PMT	Transmitted-light photomultiplier tube (microscopy channel)
TSS	Transcription start site
TS	Terminal spikelet
TT2	*TRANSPARENT TESTA 2*
UMR	Unmethylated region
UTR	Untranslated region
W1, W2, W2.5, etc	Waddington wheat developmental stage
WGBS	Whole-genome bisulphite sequencing
WT	Wild-type
X-Gluc	5-bromo-4-chloro-3-indolyl-β-D-glucuronic acid cyclohexylammonium

1 – Introduction

1.1 – Improvements across major and minor crops are needed for sustainable food security

The human population will reach 9.7 billion by 2050 and will continue expanding until a peak of around 10.3 billion by 2085 (UN medium variant forecast; UN medium variant forecast; *World Population Prospects*, 2024). In order to meet the nutritional needs and changing dietary preferences of this population – especially rising meat consumption (Willett et al., 2019) – crop production will need to have increased by 35-56% between 2010 and 2050 (van Dijk et al., 2021), with further growth onward to 2085. Mounting demands on arable land for biofuel and biomaterial feedstocks are only intensifying these pressures (Hammond & Li, 2016; Das & Gundimeda, 2022; Ali et al., 2023; Jakrawatana et al., 2023). To achieve the required productivity growth in a climate and pollutant-responsible manner, despite widespread degradation of agricultural land, and without destroying the world's remaining wildernesses, is the challenge of a generation (D'Odorico et al., 2013; Willett et al., 2019).

Wheat will play a major role in this endeavour; it is one of the top three crops by acreage and tonnage worldwide (2013-2023 average; FAO, 2023), supplies over 20% of global calories, and provides more dietary protein than all meat and poultry combined (Brinton & Uauy, 2019). Today, over 800 million people are undernourished, meaning they have insufficient intake of calories and macronutrients (Willett et al., 2019). Continued genetic yield gains in the major cereals will be vital to prevent this figure swelling in the decades to come, providing a clear mandate for academic research.

Minor crops will also play an important role, particularly in tackling another form of malnutrition; micronutrient deficiency, sometimes termed 'hidden hunger'. Currently, over 2 billion people do not obtain adequate dietary vitamins and minerals, producing a plethora of preventable health conditions (Willett et al., 2019). An important factor in preventing micronutrient deficiency is ensuring people have access to a broad diversity of foods with different nutritional profiles (Remans et al., 2011; Kumar et al., 2015). The diets of individuals across the world have in fact become much more diverse over the last century, but, simultaneously, the global diet has become more homogenous (Vermeulen et al., 2020). Essentially, while the average person now consumes a greater number of crops and foods, different nations have trended towards consuming the same array of products in more similar ratios. For example, one report suggests that between 1961 and 2009, national food supplies have become 36% more similar (Figure 1.1; Khoury et al., 2014). This has been driven by globalisation, urbanisation, and rising income levels, but also partly by successes

in crop breeding; the creation of extremely high-yielding varieties in a relatively small number of species has incentivised farmers and supply chains to adopt these crops over less improved minor crops (Khoury et al., 2014; Cheng et al., 2017). Some foods have almost completely disappeared from national diets or even from global consumption, including many millets (e.g. raishan), *stenophylla* coffee, and many apple and banana varieties (Vetriventhan et al., 2020; Saladino, 2021). Re-diversifying global food systems is an important strategy for further mitigating micronutrient deficiencies (especially as many minor crops are highly nutritious), improving the resilience of global supply chains, and adapting agriculture to climate change (Willett et al., 2019; Vermeulen et al., 2020; McMullin et al., 2021).

To this end, academic research should also continue to support the improvement of minor (also termed 'orphan' or 'underutilised') crops. I became particularly interested in one such crop, the cereal tef (*Eragrostis tef*), because of its excellent nutritional properties, its esteemed place in Ethiopian culture, and its meiotically stable polyploid genome (discussed further below) – a feature shared with wheat, my primary study organism. Chapter 2 of my thesis covers work on this important orphan crop.

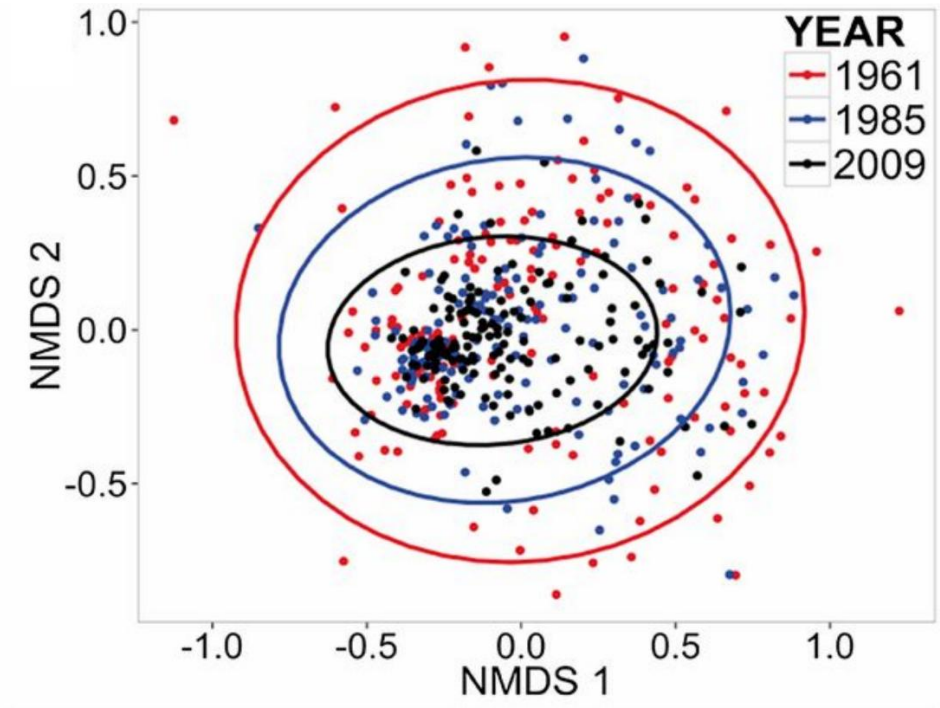


Figure 1.1 – National food supplies are increasingly homogeneous
Figure from (Khoury et al., 2014). Ordination by non-metric multidimensional scaling analysis of crop commodity composition in contribution to calories in national food supplies in 1961, 1985, and 2009. Points represent the multivariate commodity composition of each country in each year. Circles represent 95% confidence intervals around the centroid in each year. Between 1961 and 2009, the area contained within these 95% confidence intervals decreased by 68.8%, representing the decline in country-to-country variation of commodity composition (i.e., homogenization) over time. By a separate metric, mean among-country similarity increased by 35.7% during the same time period.

1.2 – Yield potential is governed by both source and sink traits

To achieve robust global food security without further land conversion, humanity will need to raise both the yield potential and yield stability of its crops while maintaining or improving quality traits. Yield stability refers to a crop cultivar's ability to maintain yield under unfavourable abiotic and biotic conditions, including pathogen exposure, herbivory, flooding, drought, high salinity, low nutrient availability, and extreme temperatures. Yield potential, meanwhile, refers to a cultivar's maximum yield under ideal conditions. Thankfully, while these two attributes can be antagonistic (Du et al., 2020; Z. H. Zhang et al., 2022), they are not necessarily so, with high yield potential varieties frequently shown to outperform lower yield potential varieties even in unfavourable environments (Weiner et al., 2021; Stella et al., 2023). In this thesis, we focus primarily on traits contributing to yield potential.

A recurring paradigm in plant research is the study of 'sources' and 'sinks' (Mason & Maskell, 1928; Foyer & Paul, 2001). Source tissues are those which acquire or produce resources for plant growth, such as roots for uptake of water and minerals or leaves and stems for photosynthetic fixing of energy and carbon backbones. Sink tissues, on the other hand, use resources derived from source tissues. Tissues may act as both sinks and sources, either for different resources (roots are sinks for photosynthates; Sonnewald & Fernie, 2018) and/or at different stages of development (Murchie et al., 2023). For example, the glumes, lemmas, and awns of cereal spikelets are initially net consumers of carbon and energy but later export photosynthates to the developing grains (Kohl et al., 2015).

For cereal crops, while vegetative material can variously be used for forage, biomass, or construction, the grain-bearing inflorescence is the primary structure of interest for human nutrition – and thus grains are most often what 'yield' refers to. Perhaps as a result of this, a simple dichotomy of describing the inflorescence as the cereal 'sink' and the remaining tissues as 'sources' has arisen (Murchie et al., 2023). In turn, researchers frequently categorise yield-related traits as contributing to either source or sink 'strength', meaning, respectively, the plant's ability to acquire resources or its capacity to convert those resources into grain yield.

There is much debate about the relative merits of trying to increase cereal yield potential by improving source and sink strength. These are reviewed extensively elsewhere but, in short, there appear to be valuable breeding targets in both spheres (Foulkes et al., 2011; Parry et al., 2011; Calderini et al., 2021; Paul, 2021; Reynolds et al., 2021; G. Liu et al., 2022; Murchie et al., 2023; Rosado-Souza et al., 2023; Slafer et al., 2023). We delve into this discussion

further in Chapter 3, reviewing whether the productivity of wheat's basal spikelets is limited by sink or source strength under common agricultural conditions.

Yield potential traits don't always neatly fit the source/sink strength dichotomy, though. Tef has notoriously tiny grains – the smallest of any cultivated cereal (Cheng et al., 2017) – and this is linked with considerable post-harvest losses during winnowing (Barretto et al., 2021; Tiguh et al., 2024). Breeding to increase tef grain size would very likely not raise sink strength or in-field yield due to trade-offs with grain number (Sadras, 2007; Gambín & Borrás, 2010; Griffiths et al., 2015; T. Guo et al., 2018), but it could boost net or 'take-home' yields for farmers by reducing post-harvest losses. In Chapter 2 we uncover numerous loci associated with grain size, specifically grain width, which we hope may contribute to this breeding aim in the future.

1.3 – Cereal architecture and yield components

The improvement of cereals is supported by a wealth of detailed knowledge about their development. Cereals are defined as grasses (Poaceae) cultivated for grain (Merriam-Webster), and grasses, like all plants, are composed of repeating units known as phytomers (Forster et al., 2007). Each phytomer is comprised of a node and an internode. Above ground, each node produces a leaf meristem and an axillary meristem, while the internodes are the stem segments between nodes (Evers & Vos, 2013). All aerial structures arise from the outgrowth, suppression, and interplay of node meristems and internodes.

Whilst it is important to understand the sequence of plant ontogeny, its temporal pace, and its environmental plasticity (e.g. (Kirby & Appleyard, 1984; O'Connor et al., 2020), we will here introduce relevant points of above-ground cereal architecture by examining a mature grass plant grown under favourable conditions. The basal-most nodes support full leaves from their leaf meristems, comprised of a blade and sheath, and some produce vegetative branches called tillers from their axillary meristems. Non-aborted tillers may go on to behave much like the main stem, including flowering and producing their own tillers (Evers & Vos, 2013). However, tillers are typically more resource-limited and less productive than the main stem. The total number of leaves and grain-bearing tillers produced varies greatly according to species, cultivar, and environmental conditions. Moving apically, both leaf and axillary meristems become progressively more suppressed in the ‘aerial’ nodes, whose internodes become elongated prior to heading. Lastly, there is a switch to denser patterning of phytomers for the inflorescence, with much less internode elongation. The final internode before the inflorescence is known as the peduncle and the main stem of the inflorescence is called the rachis (Figure 1.2).

Grass inflorescences all produce grain-bearing structures called spikelets but display highly variable architectures which strongly influence yield under agriculture or fitness in nature (Figure 1.2; Bartlett & Thompson, 2014). For example, many species, including rice and tef, produce panicles, which have primary, secondary, and potentially higher order branching from the main rachis, with spikelets borne spirally on the branches and sometimes on the upper rachis. Branch size and degree tend to decline acropetally. Meanwhile, the Triticeae tribe, including wheat, make inflorescences known as spikes, a defining phylogenetic characteristic of the group. Spikes are unbranched and produce sessile spikelets distichously directly from the rachis (Koppolu & Schnurbusch, 2019).

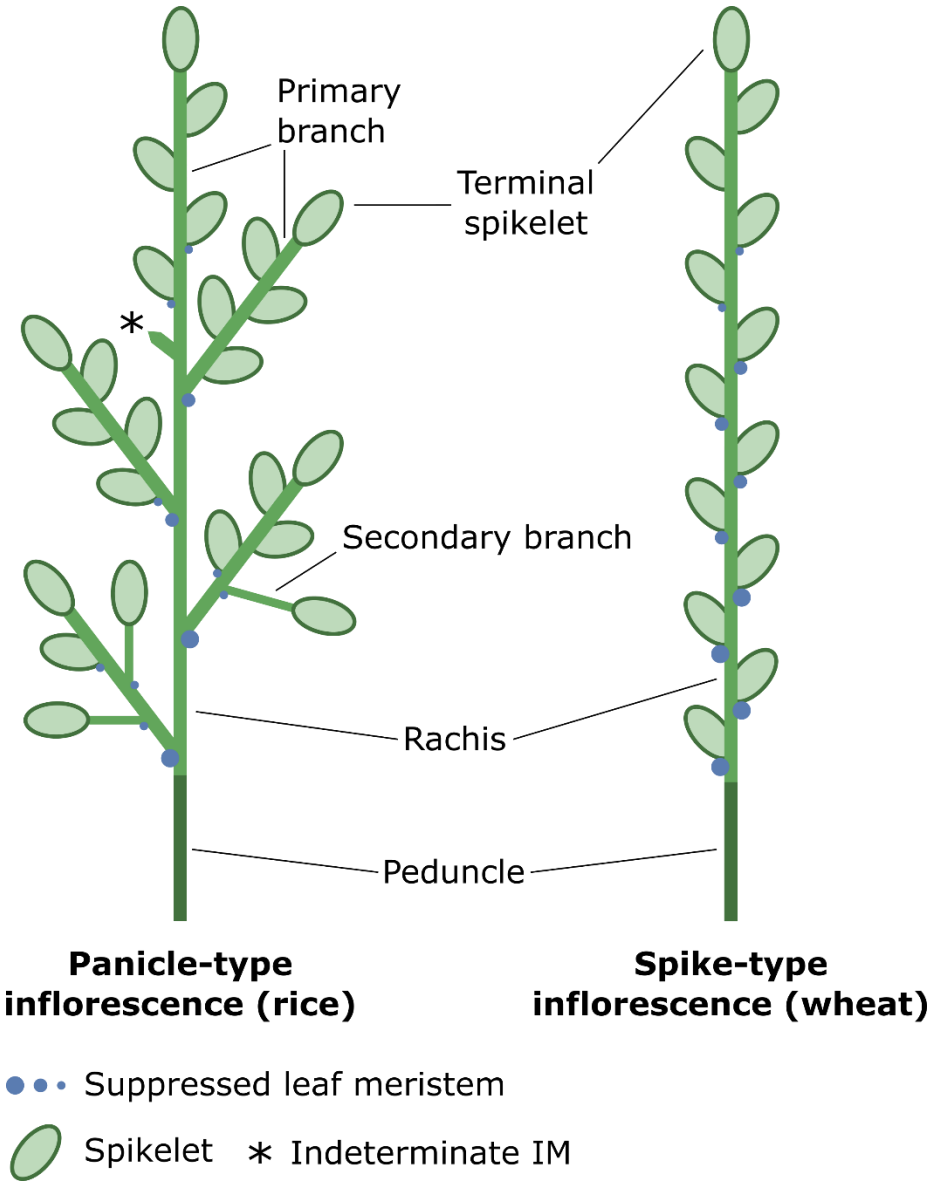


Figure 1.2 – Architectures of panicle and spike-type cereal inflorescences
Schematic illustrations of panicle and spike-type inflorescences, using rice and wheat as examples, respectively. Size and density of branches and spikelets not to scale. Blue dots denote the acropetally declining outgrowth of leaf meristems. IM = inflorescence meristem.

Additionally, the rachis inflorescence meristem (IM) and branch meristems may be determinate, ending in a terminal spikelet, or indeterminate, continually initiating spikelets until the meristem withers (Kellogg et al., 2013; Chun et al., 2022). Rachis and branch determinacy are independent. For example, rice branches are determinate, but the rachis is indeterminate (Kellogg, 2022). Branch angle and rachis node density are also highly variable, including within species. Spikelets too, may be determinate or indeterminate and produce variable numbers of florets. Barley and rice spikelets are determinate and produce a single floret per spikelet, those of maize are determinate and initiate two florets in both ear and silk, while wheat spikelets are indeterminate and may initiate 8-12 florets, though it is rare for more than five of these to set grain in even the most productive spikelets (Sakuma & Schnurbusch, 2020). While most species produce a single spikelet per node, this is also variable between species. For example, maize and sorghum produce two spikelets per node and six-rowed barley produces three (Sakuma & Schnurbusch, 2020). These variable architectural features are combined in myriad fashions across the grass family, producing a stunning array of inflorescence types and offering taxonomically informative characteristics.

In terms of the phytomer model, branches and spikelets both arise from axillary meristems, and are themselves recurring strings of phytomer units. The leaf meristems on the rachis and branches are usually highly suppressed, though as we will review in Chapter 3, this effect is weaker at the base of the rachis due to opposing gradients of vegetative and floral regulators. The internodes of spikelets comprise a short branch known as a rachilla, and the nodes again contain both axillary and leaf meristems. Typically, the axillary meristems of the first two spikelet nodes will be suppressed, and the leaf meristems will develop into bracts known as glumes (Kellogg, 2022). These are variable in size and complexity, and can be highly reduced, as seen in rice (Ciaffi et al., 2011). Later nodes develop both a floret from their axillary meristem and a lemma from their leaf meristem (from which an awn may extend). Florets are analogous to the flowers of other angiosperms and are typically bisexual, though monoecism (notably in maize), dioecism, and other reproductive systems exist at lower frequencies (Connor, 1979). Bisexual florets are comprised of four whorls of organs whose identity is controlled by a similar set of homoeotic genes (Ciaffi et al., 2011). Moving inwards, the whorls consist of a palea (equivalent to sepals or outer tepals), two or three lodicules (equivalent to petals or inner tepals), three or six stamens, and a gynoecium (set of fused carpels) containing a single ovule which may become a grain (Kellogg, 2015; Schrager-Lavelle et al., 2017; Kellogg, 2022).

Our understanding of cereal architecture allows us to frame the grain yield of a crop in terms of yield components of various levels. High-level yield components include plants per unit area, fertile tillers per plant, grain per tiller, and dry thousand grain weight. Grain per tiller can be further decomposed into number of inflorescence branches (for panicle-type inflorescences), spikelets per branch or per rachis, and grains per spikelet. Grain weight is, simplistically, a function of volume and density, but these too can be further dissected. Volume is affected by grain shape, height, and width, while density is affected by the relative sizes of different tissues and their compositions in terms of protein, starch, and other macromolecules.

As mentioned above for grain size, increases in one yield potential, yield stability, or quality component are often traded-off against decreases in others, potentially because of resource limitations (Sadras, 2007; Gambín & Borrás, 2010; Griffiths et al., 2015; T. Guo et al., 2018; Ke et al., 2020; Dwivedi et al., 2021; Bektas et al., 2023; Ren et al., 2023; Cao et al., 2024; Takai, 2024). Nonetheless, studying, breeding, and engineering specific yield components is a valid reductionist approach to crop improvement (Brinton & Uauy, 2019) and occasionally uncovers strategies which may decouple recurrent trade-offs (Kuzay et al., 2019; Liu et al., 2019; Sakuma et al., 2019; Calderini et al., 2021; Song et al., 2022; Takai et al., 2023). Reductionist approaches are particularly valuable for certain crops, including wheat and tef, which are made hard to study by their complex, polyploid genomes.

1.4 – Recessive variation is often masked in recent polyploids

Polyploidisation, also described as whole-genome duplication, is a phenomenon whereby an organism acquires an additional set of chromosomes compared with its progenitor(s). Autopolyploidy occurs through karyotype doubling within a single species, whereas allopolyploidy occurs through doubling after the hybridisation of distinct species (Van de Peer et al., 2017). Polyploids of both kinds may also be described with reference to the number of chromosome sets they have compared with their diploid progenitor(s); tetraploids have four sets, hexaploids have six, octaploids have eight, and so on.

Polyploidy is considered an important mode of speciation, particularly for plants, as polyploids are not typically able to cross successfully with their diploid progenitor species, creating an instantaneous and relatively strong form of reproductive isolation (Van de Peer et al., 2017). This has important implications for the breeding of polyploid crops, as variation is bottlenecked by a founder effect and reduced future gene flow in addition to the purifying effects of domestication and artificial selection (Borrill et al., 2019; Gaurav et al., 2022). Indeed, from the 1980s onward, much effort has been made to introduce novel diversity into the bread wheat gene pool by creating synthetic polyploids from diverse accessions of its progenitors (Dreisigacker et al., 2008; A. Li et al., 2018; Gaurav et al., 2022).

Bread wheat (*Triticum aestivum*) is a recent allohexaploid (2n=6x=42, AABBDD) that originated during the dawn of agriculture some 10,000 years ago (Figure 1.3; Marcussen et al., 2014; Levy & Feldman, 2022). Hexaploidization occurred between *Aegilops tauschii* (2n=2x=14, DD) and tetraploid emmer wheat (*Triticum turgidum*) (Marcussen et al., 2014; Levy & Feldman, 2022). Emmer is itself a recent polyploid (2n=4x=28, AABB), emerging less than half a million years ago from hybridisation of *T. urartu* (2n=2x=14, AA) and an unknown relative of *Ae. speltoides* (2n=2x=14, BB) (Huang et al., 2002; Dvorak & Akhunov, 2005; Avni et al., 2017). In contrast, durum wheat is a subspecies of *Triticum turgidum* which arose from emmer without further polyploidisation (Dubcovsky & Dvorak, 2007).

During allopolyploidisation, orthologous genes are combined into a single genome. These gene copies are termed homoeologs and they are frequently functionally redundant – at least initially – due to performing equivalent functions in the two progenitor species. Over evolutionary time, homoeologs typically diverge in function via the processes of pseudogenisation, neofunctionalisation, and subfunctionalisation (Comai, 2005; Flagel & Wendel, 2009; Roulin et al., 2013; Panchy et al., 2016).

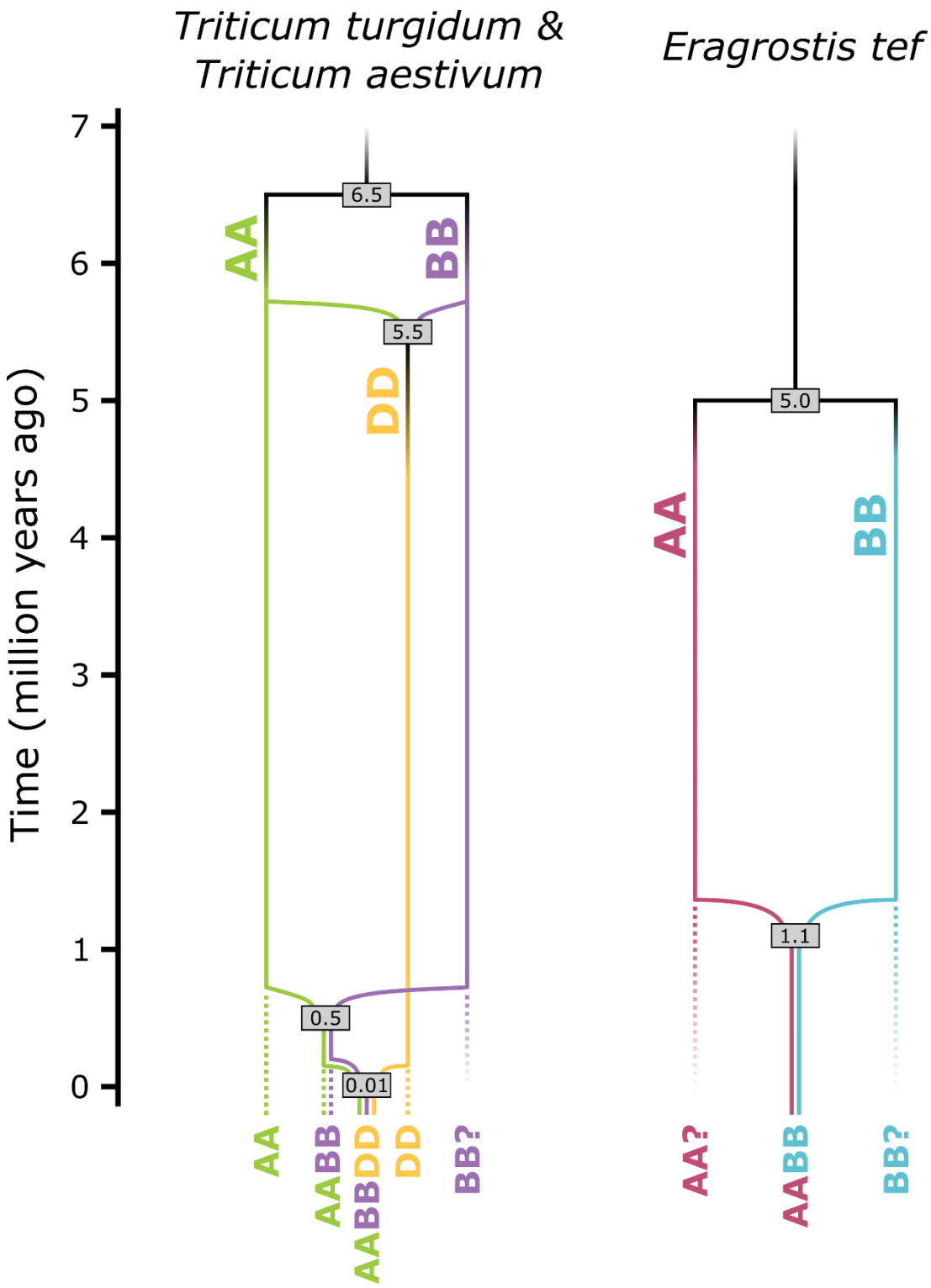


Figure 1.3 – Polyploidisation histories of *Triticum turgidum* (AABB), *Triticum aestivum* (AABBDD), and *Eragrostis tef* (AABB)

Estimated dates for speciation, hybridisation, and polyploidisation events are given in grey rectangles. Diploid lineages are indicated by coloured lines and labels. A and B are used for diploid lineages from Triticeae and *Eragrostis* by convention and are not related. Lines fading towards the present indicate that the diploid progenitor of a subgenome is unknown and may have become extinct. Data for Triticeae from (Marcussen et al., 2014; Avni et al., 2017; Levy & Feldman, 2022). Data for *Eragrostis* from (VanBuren et al., 2020). Figure design adapted from (Marcussen et al., 2014).

However, modern bread wheat still exhibits a high level of genetic redundancy for three reasons: firstly, all three contributing diploid genomes were highly related; the A and B genomes diverged just 6.5 million years ago, while *Ae. tauschii* (D genome) likely arose soon after as a homoploid hybrid of the A and B progenitors (Marcussen et al., 2014)*. Therefore, a high proportion of orthologs would still have played similar roles at the time of the polyploidisation events. Secondly, relatively little time – on evolutionary scales – has elapsed since the polyploidisations, allowing for little divergence. Bread wheat homoeologs retain 95-98% sequence identity across their coding regions (Appels et al., 2018). Lastly, the effective diploidisation of the genome by the homoeolog pairing locus *Ph1* (meaning dramatically reduced frequency of crossing over between homoeologous chromosomes) reduced chromosome elimination via aberrant recombination, leaving a large proportion (~63%) of homoeolog triads intact (Griffiths et al., 2006; Appels et al., 2018).

These characteristics also appear, broadly, to apply to tef (Figure 1.3). We note, though, that its evolution is much less highly studied, with most of the findings below only described in the tef reference genome paper (VanBuren et al., 2020). Tef is considered a domesticated species of *Eragrostis pilosa*, an allotetraploid that arose around 1.1 million years ago from unknown progenitors (Ingram & Doyle, 2003; VanBuren et al., 2020), making it a slightly older polyploid than emmer. However, the two contributing diploids diverged only ~5 million years ago – more recently than the wheat A and B progenitors. Overall, though, the dyads of tef exhibit weaker sequence similarity than the triads of bread wheat, averaging 93.9%. Lastly, much like wheat, tef is meiotically stable, displaying no detectable recombination of homoeologous chromosomes and strictly disomic inheritance. In fact, a much higher fraction (93.5%) of tef genes remain as homoeologous dyads than wheat genes remain homoeologous triads. Polyploidy is highly prevalent in the Chloridoideae subfamily to which tef belongs, occurring in an estimated 90% of species. Among the Chloridoideae allopolyploids, multivalent chromosome pairing is rare, suggesting this group has evolved some mechanism equivalent to the *Ph1* locus of wheat to stabilise its karyotype and maintain fertility (VanBuren et al., 2020), allowing it to repeatedly reap the emergent benefits of polyploidisation (Estep et al., 2014; Van de Peer et al., 2017).

Redundancy amongst homoeologs impacts breeding because recessive variation at one gene copy, such as a knockout, is masked by the other homoeologs, often resulting in minor

*This was evidenced by the finding that hundreds of gene trees across the genome supported A(B,D) and B(A,D) topologies approximately equally, while D(A,B) trees were roughly half as frequent, a pattern incompatible with incomplete lineage sorting alone (Marcussen et al., 2014). It is worth noting, however, that this claim has been disputed (L. F. Li et al., 2015) and then counter-disputed (Sandve et al., 2015) based on chloroplast lineage data since its publication. On balance, we favour the homoploid hybridisation interpretation.

or negligible phenotypic effects. This buffering may take the form of a dosage effect or of complete functional redundancy. A dosage effect occurs where homoeologs have similar cellular roles, but the remaining copies cannot compensate sufficiently to fully buffer a knockout. Accordingly, loss-of-function variation in one homoeolog has a small phenotypic effect, with the effect growing as additional homoeologs are mutated. Such effects may be additive, where phenotypic change is proportional to the number of gene copies knocked out, or non-additive, where change is not proportional, the typical case in bread wheat being small effects in single and double mutants, with a more extreme phenotype in the triple mutant. Alternatively, in cases of complete functional redundancy, knockout of one or two homoeologs produces no detectable phenotype and only the triple mutant is distinguishable from the wild-type (WT) parent (Borrill et al., 2019).

An example of an additive dosage-dependent gene is *GRAIN WIDTH and WEIGHT 2* (*GW2*), a major grain weight locus. In rice (*Oryza sativa*), which is diploid, near-isogenic lines (NILs) homozygous for a loss-of-function allele of *OsGW2* exhibit a 50% higher thousand grain weight (TGW) versus the WT parent (Song et al., 2007). In contrast, in the bread wheat 'Paragon', NILs for TILLING-derived (Uauy et al., 2009; Krasileva et al., 2017) knock-out mutations of single *TaGW2* homoeologs increased TGW by an average of just 5.3%. Only in the triple *TaGW2* mutant (i.e. aabbdd) was TGW affected by a similar magnitude as in rice, with a 20.7% increase (Simmonds et al., 2016; Wang et al., 2018).

Complete functional redundancy is illustrated by the gene *SBE-IIa*, which encodes a starch branching enzyme (SBE). Knocking out an SBE typically modifies the composition of starch drastically, increasing the ratio of weakly branched amylose to highly branched amylopectin. For *SBE-IIa* in bread wheat, this effect is only seen in the triple mutant, in which the ratio of amylose to amylopectin is four-fold that of the WT. No effect is seen in any double mutant combination, suggesting that the remaining homoeolog in each case is able to completely buffer the loss of two gene copies (Slade et al., 2012). Additional examples of dosage effects and functional redundancy are catalogued in (Borrill et al., 2019).

Regardless of the type of genetic buffering, introgressing loss-of-function variation for a single homoeolog into elite cultivars of polyploid species is associated with little, if any, reward. Achieving genetic gains commensurate to those seen in diploids requires stacking alleles across homoeologs, increasing breeding time and effort. More importantly, the 'hidden' nature of loss-of-function variation in polyploids makes detection and mapping of recessive quantitative trait loci (QTL) difficult in the first place (Borrill et al., 2019; Brinton & Uauy, 2019). The same difficulties apply to functional genetics studies and have precluded the development of gene knockout libraries, which have proved invaluable in other plants

(Ostergaard & Yanofsky, 2004; Wang et al., 2013; Matus et al., 2014; Li et al., 2017; X. Lu et al., 2018). Production and validation of double (tef or durum wheat) or triple mutants (bread wheat) knockouts, respectively, is time intensive, and is usually reserved for candidate genes whose knockout phenotypes have already been documented in a diploid cereal such as barley, rice, or maize.

1.5 – Domestication alleles in polyploid crops are frequently dominant

In contrast to recessive alleles, dominant alleles produce their full phenotypic effects when a single copy is present in the genome. Similarly, semi-dominant alleles produce large effects via a single homoeolog, but their effects are smaller when heterozygous with another allele. For example, the semi-dominant *P1*^{POL} allele of *VRT-A2* (TraesCS7A02G175200) increases glume length by 37% in homozygous Paragon NILs (amongst other pleiotropic effects; Adamski et al., 2021). Though by no means a like-for-like comparison, this is a considerably greater effect on a quantitative trait than that produced by homozygous single homoeolog knockouts of *GW2*.

Dominant alleles have therefore presumably been easier to detect and select for throughout wheat's agronomic history. Supporting this, the known major wheat domestication alleles are primarily dominant and semi-dominant (Peng et al., 1999; Yan et al., 2003; Fu et al., 2005; Doebley et al., 2006; Griffiths et al., 2006; Simons et al., 2006; Al-Kaff et al., 2008; Wilhelm et al., 2009), with notable exceptions at the *Tg* glume toughness loci (Dvorak et al., 2012; Faris et al., 2014), *Br* rachis fragility loci (Nalam et al., 2006; Avni et al., 2017), and senescence / grain protein content locus *Gpc-B1* (mapped to the gene *NAM-B1*; Uauy et al., 2006; Asplund et al., 2010; Lundström et al., 2017).

As an aside, no firm domestication loci or genes have yet been established for tef. *YAB2/SH1* was proposed as a candidate gene for reduced grain shatter (the primary distinguishing feature between *E. pilosa* and domesticated tef) based on sequence conservation and its role in sorghum and foxtail millet. However, knockout of this gene had no effect on abscission (Yu et al., 2023). In Chapter 2, we propose that grain colour – another important domestication trait in tef (Woldeyohannes, Desta, et al., 2022) – is strongly influenced by recessive variation at the tef orthologues of *TRANSPARENT TESTA 2*.

Dominant alleles are usually characterised, at the molecular level, by a gain-of-function. This can occur due to variation in both the coding and non-coding regions of a gene. For example, if a mutation in a gene's coding region disables a degradatory domain on the cognate protein, but does not compromise its catalytic or binding capacity, its functional activity will likely increase. Increased negative regulation at the transcriptional level or of the other homoeologs often fails to fully buffer such a change. If the protein acts as a transcription factor (TF), this could alter the expression of tens or hundreds of downstream genes. This mechanism underpins the mutations at the heart of the 1960s Green Revolution

in wheat; the *Rht1* alleles from Norin 10 were later each linked to dominant mutations in the binding domain of DELLA TFs that make them insensitive to gibberellic acid-mediated degradation. The mutant DELLA proteins constitutively suppress stem growth (among many pleiotropic effects), producing a semi-dwarf phenotype that prevents lodging under high fertiliser applications (Peng et al., 1999).

Dominant variation can also occur due to de-regulation at the translational (rather than post-translational, as above) level. The dominant *Q* allele, which confers free-threshing glumes and a compact, square spike, offers a clear example of this: it is an allele of the TF *APETALA2-5A* in which an exonic mutation reduces the mRNA's affinity to a degradatory microRNA (miR172), leading to elevated protein levels (Debernardi et al., 2017).

Lastly, changes in non-coding regulatory sequences (see next section; 1.6) can also produce dominant effects by altering transcription. Such mutations can cause genes to be expressed at altered levels in their original cell or tissue types, in a novel spatiotemporal pattern, or both. Again, when such a gene encodes a TF, the mutation can ultimately modify the expression of a multitude of downstream genes, potentially greatly altering the developmental trajectory of plant tissues. Several major wheat domestication traits arose from such mutations (Table 1.1). Though not a domestication allele, the semi-dominant $P1^{POL}$ allele discussed above was also found to be a product of *cis*-regulatory variation (Adamski et al., 2021). Copy number variation (CNV) can similarly produce dominant effects by altering gene expression. Duplicated genes can elevate transcript levels, or if duplicated into a novel genomic environment with new CREs, can lead to transcription in novel spatiotemporal patterns. CNV has produced dominant alleles for important traits such as flowering time (*VRN-A1* and *Ppd-B1*; Diaz et al., 2012), cold tolerance (*CBF* genes at *Fr-A2* locus; Wurschum et al., 2017), and awn development (*B1* locus; Li et al., 2023).

Of these four mechanisms for dominant variation – altering post-translational, translational, and transcriptional control, plus CNV – the latter two appear to have produced more pivotal alleles over the history of wheat cultivation. Taking a theoretical purview, this could be because eliminating negative regulation of a gene can only directly influence the tissues it is already expressed in or that its protein translocates to. In contrast, modifying a gene's regulatory elements can alter its transcriptional status in tissues it is not normally expressed in, ultimately offering greater flexibility for significant developmental effects.

Table 1.1 – Dominant wheat domestication alleles arising from non-coding regulatory variation
Where multiple homoeologs are given, the described mutations in either/any homoeolog produce a dominant effect.

Locus or gene names	Gene ID (IWGSC RefSeq v1.1)	Domestication trait	Domestication allele(s)	Reference
VRN-A1 ***VRN-B1*** ***VRN-D1***	TraesCS5A02G391700 TraesCS5B02G396600 TraesCS5D02G401500	Vernalisation insensitivity in spring wheat	Common deletions in intron 1 spanning an approx. 4kb region	Fu et al., 2005
PpD-A1 ***Ppd-D1***	TraesCS2A02G081900 TraesCS2D02G079600	Photoperiod insensitivity, earlier heading	Various dominant alleles share an equivalent ~900 bp deletion upstream of TSS. Alters circadian expression.	Beales et al., 2007; Wilhelm et al., 2009; Seki et al., 2011
Ph1	TraesCS5B02G255100	Subgenome stability, high fertility	*Ph1* maps to a fourth copy of *ZIP4* (a triad on the group 3 chromosomes). This copy inserted into a novel genomic context on Chr5B, giving a distinct expression profile.	Rey et al., 2017
WAPO-A1	TraesCS7A02G481600	Increased spikelet number per spike	*Wapo-A1b* allele is highly prevalent in bread wheat. Lacks 115 bp promoter deletion found in *Wapo-A1a* and is more highly transcribed. C47F amino acid change also contributes to phenotype.	Kuzay et al., 2019; Kuzay et al., 2022

1.6 – Crop engineering will benefit from improved understanding of *cis*-regulatory sequences

Regulatory non-coding sequences include both core promoter elements and specific regulatory elements. The former are necessary for initiating a basal level of transcription and are located within ~100 bp of transcription start sites (TSSs; Vedel & Scotti, 2011), while the latter modulate the expression of cognate gene(s) through the binding of specific TFs. The term *cis*-regulatory element (CRE) is sometimes applied to both of these classes of non-coding DNA and simply indicates that they lie on the same chromosome as the target gene, in contrast to *trans*-acting factors. In this thesis, we use CRE to describe only specific regulatory sequences and exclude core promoter elements.

CREs may be near or far from a target TSS – up to several Mbp away – and may be upstream, downstream, or within the gene, including in the 5' untranslated region (UTR), introns, and 3' UTR (Figure 1.4; Z. Lu et al., 2018). CREs are typically 50 to 1,500 bp in length and usually contain multiple TF-binding motifs (Vedel & Scotti, 2011; Weber et al., 2016). Distal CREs require DNA folding to be brought into physical proximity with the core promoter and, consequently, their activity is orientation independent (Weber et al., 2016). CREs are sometimes referred to as enhancers and silencers where the direction of their influence on a gene is known. However, a given CRE may act as an enhancer and a silencer for different genes or even for the same gene in different contexts.

A potential strategy for engineering variation, particularly dominant variation, is to alter the regulatory control of key genes, especially TFs, involved in development. This can be accomplished either through mutations in their associated CREs (achievable via genome editing), or by introducing a transgenic copy under the control of alternate promoter elements and CREs. However, apart from a handful of examples associated with intensively studied genes – such as the domestication loci described above – CREs coordinating gene expression in wheat and tef have remained poorly characterised. Due to (now largely historical) difficulties in studying chromatin properties in small tissues, this is especially true for CREs involved in inflorescence development, drastically limiting the scope for rational engineering. Nonetheless, in recent years, and especially throughout the duration of my PhD, the study of CREs in crop plants has greatly accelerated. This includes improved methods for CRE detection and for identification of their target genes. These methods are reviewed in Chapter 4 and we employ strategies for CRE and CRE-gene pair detection in Chapters 3 and 4. This knowledge informed our strategy for engineering variation in basal spikelet productivity, as presented in Chapter 3.

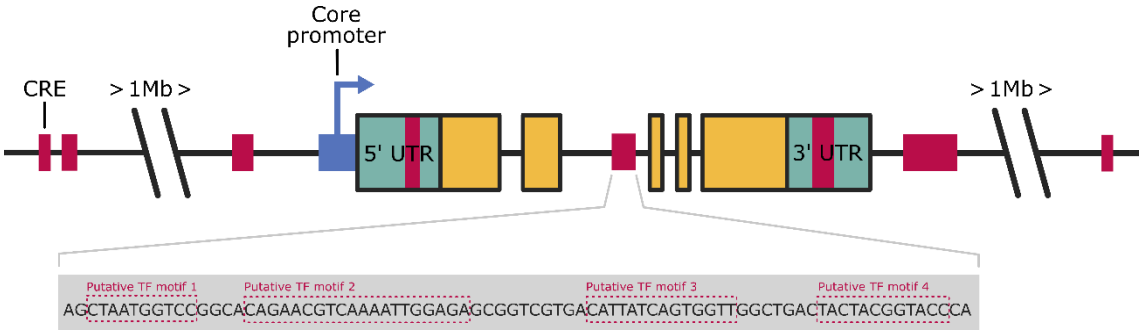


Figure 1.4 – *cis*-regulatory elements (CREs) are distinct from the core promoter and may occur upstream, downstream, or inside target genes
Arbitrary gene structure used to illustrate potential locations of cognate CREs. Exons denoted by tall boxes, with coding segments in yellow and untranslated regions (UTRs) in teal. Regulatory elements denoted by short boxes, with CREs in magenta and core promoter in blue. CREs may influence genes from up to and over 1 Mbp away. CREs are comprised of one or more transcription factor (TF) motifs.

1.7 – Thesis aims and objectives

The overall aim of this thesis is to explore different types of variation – natural and engineered, recessive and dominant, coding and *cis*-regulatory – for inflorescence traits in polyploid cereals. From the outset of my PhD, I had a strong interest in working on a major and a minor crop given the importance of both to robust global food security. To that end, this thesis includes work on both wheat and tef.

In Chapter 2 I aimed to uncover novel marker-trait associations using *k*-mer-based GWAS on an important tef germplasm collection. This led us to identify natural variation at a major locus for grain colour and grain size.

Objectives:

- Assess EIAR tef core collection for redundancy with Illumina sequencing data
- Calculate BLUPs for phenotypic data via linear mixed modelling
- Conduct k-mer-based GWAS, then evaluate associations and candidate genes

In Chapter 3, I used semi-spatial transcriptomics and engineered variation to investigate the phenomenon of low basal spikelet productivity in wheat. I transgenically overexpressed candidate positive regulators of spikelet development in the base of the wheat spike, aiming to produce a dominant gain in basal spikelet productivity.

Objectives:

- Collect, quality control, and analyse semi-spatial RNA-seq data across five key stages of wheat spike development
- Identify candidate regulators of spikelet development with weaker expression in the basal section of developing spikes versus the central section
- Identify additional genes with strong basal but weak central expression from which to develop novel basal spike-specific promoters
- Assemble misexpression constructs via Golden Gate cloning
- Phenotype transgenic wheat lines for basal spikelet productivity and off-target effects

Designing promoters for my transgenic constructs underscored the importance of understanding the wheat spike *cis*-regulatory landscape. In Chapter 4, I therefore sought to answer whether a combination of transcriptome, chromatin accessibility, and DNA methylation data is sufficient to generate a genome-wide catalogue of spike *cis*-regulatory elements and to link these to their target genes.

Objectives:

- Collect, quality control, and analyse RNA-seq and ATAC-seq data across 16 stages of wheat spike and carpel development
- Remap and analyse equivalent RNA-seq and ATAC-seq from (Lin et al., 2024)
- Analyse differentially expressed genes for enrichment of differentially accessible chromatin
- Develop a method to identify gene-CRE pairs by correlating their expression and chromatin accessibility trajectories across developmental stages

2 – Population genomics uncovers loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*)

2.1 – Chapter summary

Tef (*Eragrostis tef*) is an indigenous African cereal that is gaining global attention as a gluten-free "superfood" with high protein, mineral, and fibre contents. However, tef yields are limited by lodging and by losses during harvest owing to its small grain size (150× lighter than wheat). Breeders must also consider a strong cultural preference for white-grained over brown-grained varieties. Tef is relatively understudied with limited "omics" resources. Here, we resequence 220 tef accessions from an Ethiopian diversity collection and also perform multi-locational phenotyping for 25 agronomic and grain traits. Grain metabolome profiling reveals differential accumulation of fatty acids and flavonoids between white and brown grains. *k*-mer and SNP-based genome-wide association uncover important marker-trait associations, including a significant 70 kb peak for panicle morphology containing the tef orthologue of rice *qSH1*—a transcription factor (TF) regulating inflorescence morphology in cereals. We also observe a previously unknown relationship between grain size, colour, and fatty acids. These traits are highly associated with retrotransposon insertions in homoeologues of *TRANSPARENT TESTA 2*, a known regulator of grain colour. Our study provides valuable resources for tef research and breeding, facilitating the development of improved cultivars with desirable agronomic and nutritional properties.

Phenotyping, DNA extractions, and metabolite extractions were conducted at the Ethiopian Institute for Agricultural Research and were led by Worku Kebede. Abel Teshome (ILRI) assisted with SNP-based GWAS, performed SNP-based PCA, and prepared figure 2.4. Aiswarya Girija (IBERS) conducted metabolome analyses and prepared figures 2.9, 2.10, and 2.11. James Brown (JIC) advised on statistical analyses and performed heritability calculations. Oluwaseyi Shorinola (University Birmingham) assisted with formal analyses including LD, ADMIXTURE, and synteny analysis and prepared figures 2.2, 2.7, 2.8a, and 2.15.

All results described in this chapter have been published in the following manuscript:

Population genomics uncovers loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*)
Maximillian R. W. Jones, Worku Kebede, Abel Teshome, Aiswarya Girija, Adanech Teshome, Dejene Girma, James K. M. Brown, Jesus Quiroz-Chavez, Chris S. Jones, Brande B. H. Wulff, Kebebew Assefa, Zerihun Tadele, Luis A. J. Mur, Solomon Chanyalew, Cristobal Uauy and Oluwaseyi Shorinola, 2025. *Communications Biology*
https://doi.org/10.1038/s42003-025-08206-5

See Appendix A for publisher's version

2.2 – Introduction

Tef (*Eragrostis tef* (*Zucc*.) Trotter) is a cereal crop that has been grown in the Horn of Africa for millennia. It is a self-pollinating allotetraploid grass that is valued by farmers as a 'fail-safe' cash crop, resilient to marginal soils, waterlogging, high temperatures, and drought. Tef is a staple crop in Ethiopia, Africa's second most populous country, where it is grown on 3 million hectares (27% of cereal acreage) by around 6.7 million households, with annual production exceeding 5.5 million tonnes (Ethiopian Statistics Service, 2022a, 2022b). The crop acts as both feed and food, with the straw a prized forage for cattle and the whole-grain flour used to produce a fermented flatbread known as injera, which serves as a staple food for the majority of the country (Cheng et al., 2017).

Tef has also gained global attention as a 'superfood' thanks to its high protein, calcium, iron, and fibre contents, its low glycaemic index, and its lack of allergenic gluten (Cheng et al., 2017). Additionally, tef is rich in dietary antioxidants, including polyphenols and flavonoids, and essential polyunsaturated fatty acids like linoleic acid, which are not synthesised by the human body (Cotter et al., 2023). These nutritional features, combined with its climatic resilience, make tef an attractive crop for wider adoption. To date, government policies in Ethiopia have restricted export of tef germplasm and bulk grain to protect both its natural heritage and domestic consumption (Lee, 2018; Sankaranarayanan et al., 2020). However, tef cultivation is expanding beyond Ethiopia, notably in the USA, Australia, South Africa, and the Mediterranean regions (Barretto et al., 2021; Ruggeri et al., 2024). In these areas, tef is also used as a multi-harvest forage crop for producing premium-quality hay and silage (Wagali et al., 2023).

Tef is considered an underutilised crop because it has, so far, not benefited greatly from modern genomics-based approaches to breeding and research. However, as in other underutilised crops such as grass pea, yam, and lablab, this *status quo* is beginning to shift (Gonzales et al., 2024; Shorinola et al., 2024), with the generation of a high-quality reference genome (VanBuren et al., 2020) following on from a draft sequence (Cannarozzi et al., 2014). Notable progress has been made in breeding improved varieties of tef (Girma et al., 2014; Cheng et al., 2017), although advances have not been on the same scale as for major cereals like wheat or rice. Lodging under high nitrogenous fertiliser regimes is a major limiting factor but has so far been difficult to address through classical semi-dwarfing approaches, at least partially because of the value of tef straw as animal feed. Addressing lodging through improved root traits is also being explored (Bayable et al., 2020; Ben-Zeev et al., 2023). Panicle (inflorescence) morphology has been reported as a determinant of

lodging tolerance in tef. The species exhibits dramatic panicle diversity (Assefa et al., 2015), from open, highly lax panicles similar to wild *Eragrostis* species, to short-branched, compact panicles more akin to the spikes of Triticeae species (Figure 2.1a). Using a combination of controlled-environment phenotyping, mechanical testing, and crop modelling, Blosch et al. showed that tef varieties with compact panicles tended to be more resistant to lodging, suggesting an ideotype approach could be used to address this issue (Blösch et al., 2020).

Another consideration for breeders is grain size. Tef produces the smallest grains of any cultivated cereal, ranging from 1.0 to 1.7 mm in length, with a typical thousand grain weight (TGW) of 0.2–0.4 g, roughly 150-fold lower than that of wheat (Zanke et al., 2015; Cheng et al., 2017; Woldeyohannes, Desta, et al., 2022; Figure 2.1b). Indeed, the name tef is thought to derive from the Amharic word “teffa” meaning “lost” (Stallknecht et al., 1993); likely an allusion to the high levels of harvest and post-harvest losses (16–30%) experienced by tef farmers (Barretto et al., 2021; Tiguh et al., 2024). Breeding for larger grains could alleviate these losses and boost realised yields by improving the separation of grain and chaff during winnowing. Tef’s small grain size also makes it difficult to evenly broadcast the recommended >10 kg/ha during sowing. Farmers instead use high sowing rates (up to 30 kg/ha) that ultimately produce overcrowded fields prone to lodging (Ben-Zeev et al., 2020). Lastly, breeders must also address a strong cultural preference for white-grained over brown-grained varieties, which translates into a higher market price for the former (Jifar et al., 2015).

Here, we aim to use a population genomics approach to study the diversity and genetic architecture of agronomic, grain morphology, and grain metabolite traits in a representative Ethiopian tef collection. We therefore conducted short-read resequencing of 220 tef accessions from the Ethiopian Institute of Agricultural Research (EIAR) tef diversity collection. We characterised redundancy in this collection and produced a compact SNP panel that uniquely identifies the studied accessions. We combined this sequencing data with extensive in-field phenotyping across three trial locations, including precise grain morphology measurements and grain metabolome profiling (Figure 2.2). This led to the identification of important marker-trait associations for panicle morphology, grain size, grain colour, and multiple grain metabolites. Our analyses establish a previously unknown link between grain size and grain colour, including the co-association of these traits with multiple genomic loci. However, we also identify regions that decouple these traits, offering potential breeding opportunities. Our work delivers a set of genomic and phenotypic resources for a diverse panel of tef accessions and lays the groundwork for future studies to define causal genes and variants underlying loci of agronomic relevance.

Figure 2.1 – Diversity of panicle morphology and grain colour in tef
a, Comparison of a bread wheat spike (cv. Paragon, far left) with tef accessions exemplifying four categories of panicle morphology (from left to right; very lax, lax, semi-compact, and compact). **b**, Comparison of bread wheat grains (cv. Paragon, bottom) with grains from brown and white-grained tef varieties.

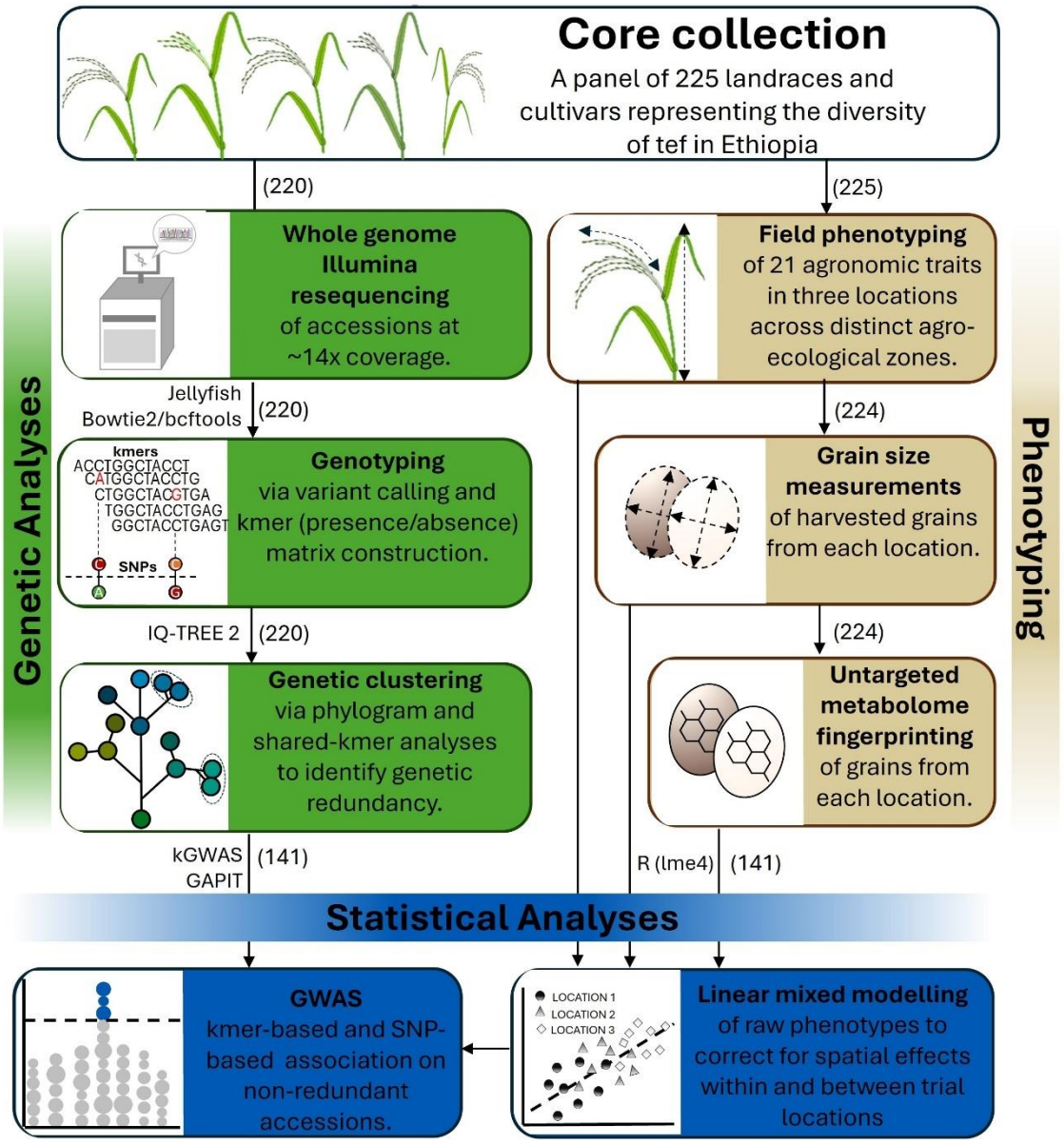


Figure 2.2 – Resequencing, phenotyping, and GWAS of the EIAR core tef collection
A representative panel of Ethiopian tef accessions was resequenced and phenotyped for agronomic, grain size, and metabolomic traits. Statistical modelling was used to correct for location and within-site spatial effects. The software and numbers of accessions used for each step are indicated above each box.

2.3 – Results

2.3.1 – SNP and *k*-mer-based methods identify redundancy in the EIAR core collection

Of the 225 accessions in the EIAR core collection, we sequenced (Illumina paired-end 150 bp) the genome of 220 accessions to an average depth of 8.85 Gbp (standard deviation (SD) = 0.77 Gbp), equivalent to 14.2-fold the estimated genome size (622 Mb) of the reference genome cultivar 'Dabbi' (VanBuren et al., 2020). The reads were mapped against the reference genome and variants were called. A quality-filtered and linkage-pruned set of 41,289 SNPs was prepared and used to investigate linkage disequilibrium (LD) decay and population structure. The genome-wide LD decay distance in the EIAR core collection was 46.3 kb (Figure 2.3). ADMIXTURE analysis revealed no clear population structure for two to 20 subpopulations (Alexander et al., 2009). Principal component analysis also did not suggest any distinct lineages, although there was a partial separation of brown and white-grained accessions by PC1 (Figure 2.4). This result was not unexpected, as a lack of strong population structure has previously been reported for other tef panels (M. D. Alemu et al., 2024).

An SNP-based phylogram was computed using IQ-TREE 2 (Minh et al., 2020). This revealed many groups of highly related accessions, with internal branch lengths close to zero nucleotide substitutions per site (Figure 2.5). A list of redundancy groups was defined such that the total branch length (phylogenetic distance) between any pair of accessions in a group was <0.005 substitutions per base pair. This resulted in 31 redundancy groups containing 2–19 accessions per group (Supp. Table 2.1). Two accessions were excluded from placement in redundancy groups because they had high apparent heterozygosities (22.7% and 24.9% heterozygous sites, versus an average of 1.5% (SD = 0.7%) for the other accessions) and likely represented seed mixtures rather than pure accessions.

To validate the redundancy groups defined above, a comprehensive *k*-mer (k = 51) presence/absence matrix was generated from the sequencing reads of the 220 sequenced accessions (Gaurav et al., 2022). We tested whether pairs of accessions from the previously defined groups tended to have more *k*-mer states in common than non-group pairs (Figure 2.5b,c, Supp. Data 2.1). We observed that all 264 intra-group pairs had a *k*-mer state identity rate above 96.0%, whereas 23,825 out of the 23,826 other pairs had a *k*-mer state identity rate below this threshold, with a mean and median of 85.4% (SD = 0.02%).

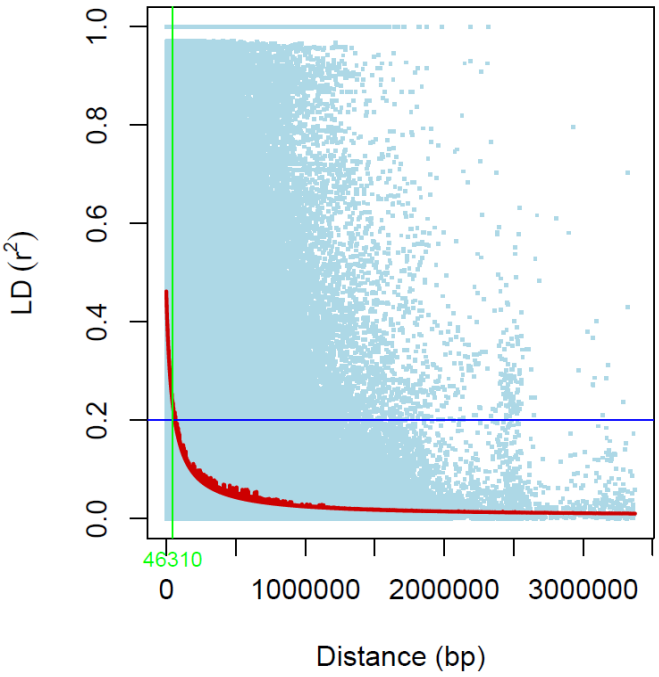


Figure 2.3 – Average genome-wide linkage disequilibrium (LD) decay
The x-axis represents the physical distance between pairs of SNPs in base pairs (bp). The y-axis represents the LD between SNPs, computed as r^2.

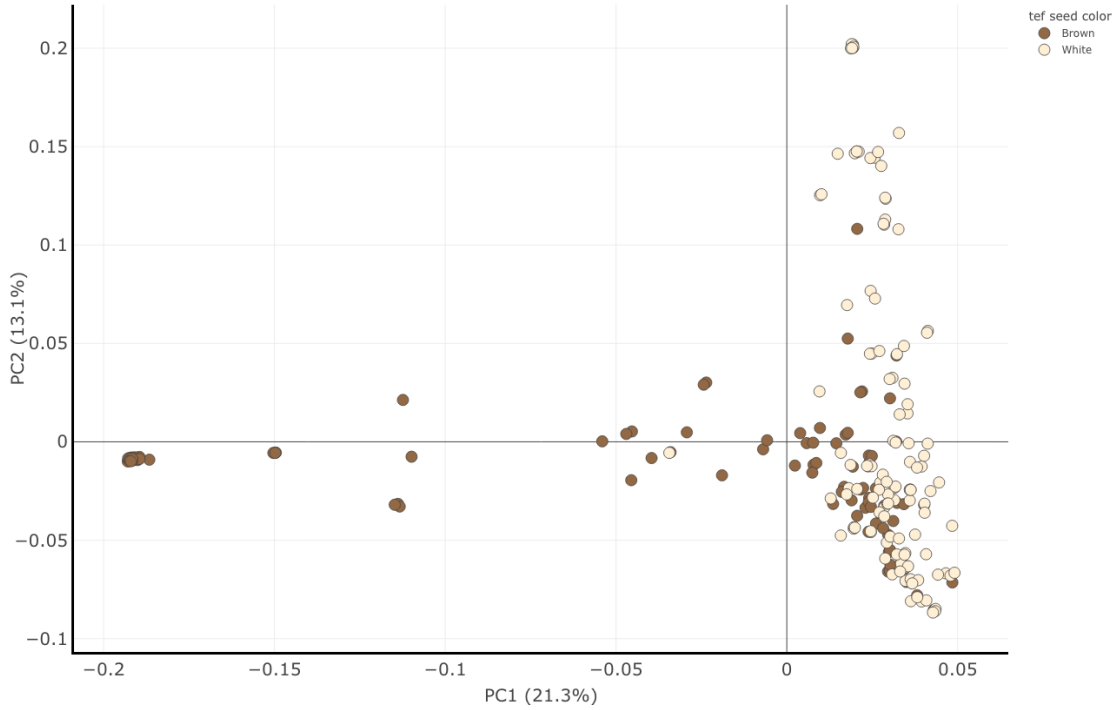


Figure 2.4 – Principal component analysis (PCA) does not strongly distinguish brown and white-grained accessions
PCA based on ~40K SNPs for 230 tef genotypes. Binary colour codes represent seed coat colour of the individual genotypes

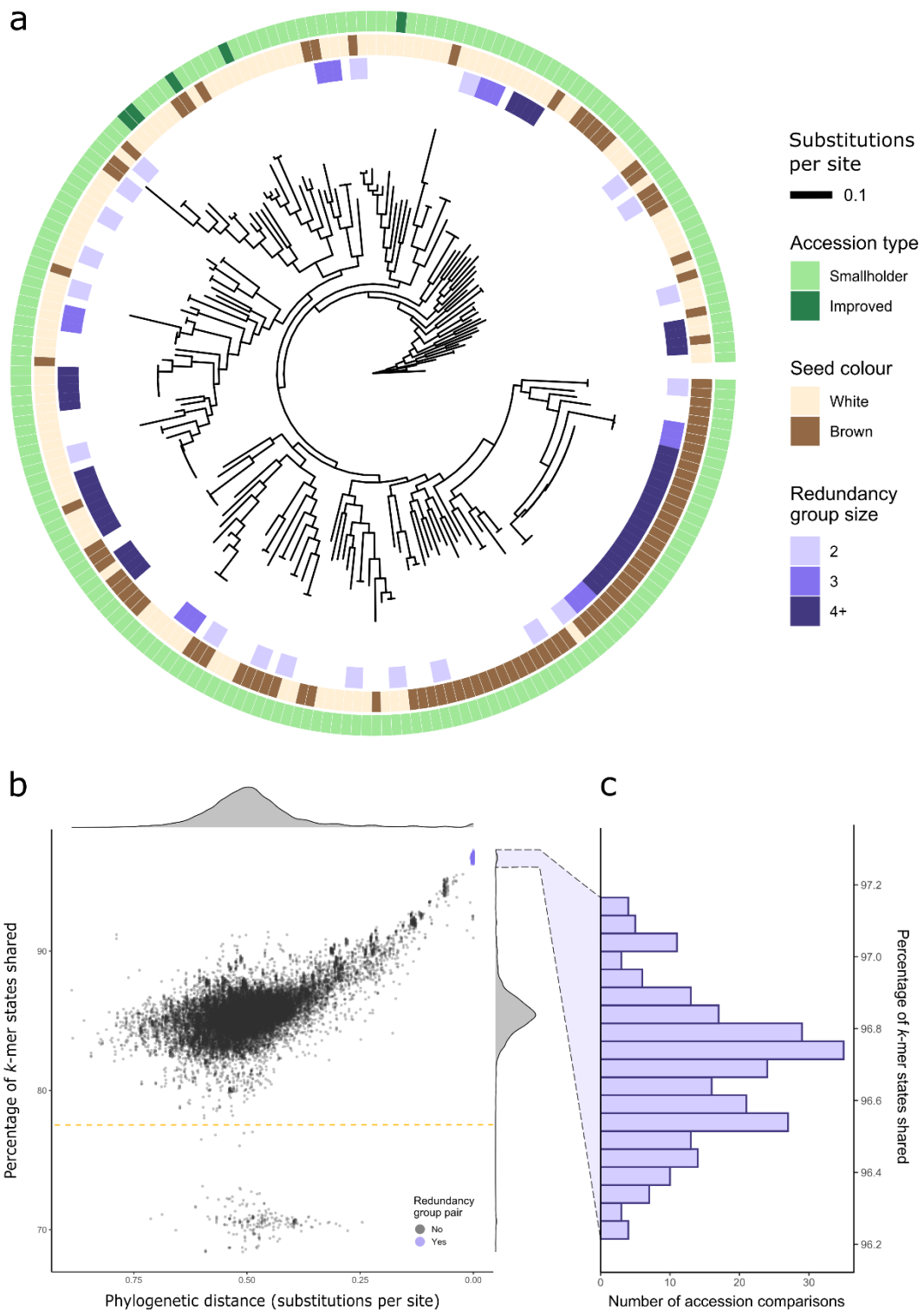


Figure 2.5 – Phylogenetic analyses identify redundancy in the EIAR core collection
a, Phylogram of 220 tef accessions, arbitrarily rooted against the accession 'Ada-T58' (insufficient SNP data were recovered for candidate outgroups, see Methods 2.5.3). White and brown-grained varieties are well-distributed across the phylogeny. A total of 32 redundancy groups, ranging in size from 2 to 19 accessions, were identified on the basis of small phylogenetic distances between pairs of accessions. **b,** Phylogenetic distance plotted against the percentage of *k*-mer states shared for all 24,090 pairwise comparisons between accessions. There is a strong correlation between these two relatedness metrics. Accession pairs from within the previously defined redundancy groups (purple) cluster together at uniquely high shared *k*-mer state rates, depicted in detail in (**c**). The points with particularly low percentages of shared *k*-mer states (<78%, dashed line) represent the full set of comparisons of 'DZ-01-1167' with other accessions.

The single pair exceeding this threshold ('DZ-01-91' and 'DZ-01-101'), still had a very low phylogenetic distance (0.007) compared to the overall distribution, so were added as an additional redundancy group. The broad agreement between these two metrics suggested that the redundancy groups would be better treated as single accession pools rather than distinct entities. This reduced our effective number of accessions from 220 to 150.

One accession, 'DZ-01-1167', produced notably low shared *k*-mer state rates when paired against all other accessions (Figure 2.5b, below the dashed line). DZ-01-1167 contains 294 million distinct *k*-mers versus an average of 160 million (SD = 3.5 million) for all other accessions (Figure 2.6), suggesting it contains many unique *k*-mers. This new diversity could derive from genetically distant tef accessions not otherwise captured in the panel or from interspecific introgression(s). Its source and utility could be further explored in future studies. This finding highlights the benefits of *k*-mer-based approaches, as this introgression is not apparent when comparing SNP-based phylogenetic distances.

Given the high levels of redundancy observed in the EIAR core collection, we selected a minimal panel of SNPs capable of distinguishing all 150 accession groups and singlets. This would allow other accessions belonging to the redundancy groups to be identified amongst the wider EIAR collection, as well as potentially facilitating some reconciliation with other tef collections. We identified a panel of 14 biallelic SNPs that could distinguish all 150 accession groups or singlets. To account for potential marker failures, we selected an additional 14 SNPs, making a total of 28. For each of these SNPs, all accession groups and singlets were homozygous for one allele or had missing data (Supp. Data 2.2, Supp. Note 2.1). All chromosomes were represented by at least one SNP except 5A, 7B, 8B, and 10A.

2.3.2 – Grain colour strongly correlates with plant height and grain morphometric traits

To capture phenotypic variation in the EIAR core collection, we phenotyped 17 phenological and morphological traits (Figure 2.7 and Supp. Table 2.7) at three field sites representing distinct agro-ecological zones (Supp. Table 2.6). Using high-resolution grain imaging, we also captured variation in eight grain size parameters relating to grain width, length and area. There was a significant effect (ANOVA p-value < 0.04) of location on all traits except for grain width and length. Trait coefficients of variation at each location ranged from 0.02 to 0.45, with phenology traits showing the least variation.

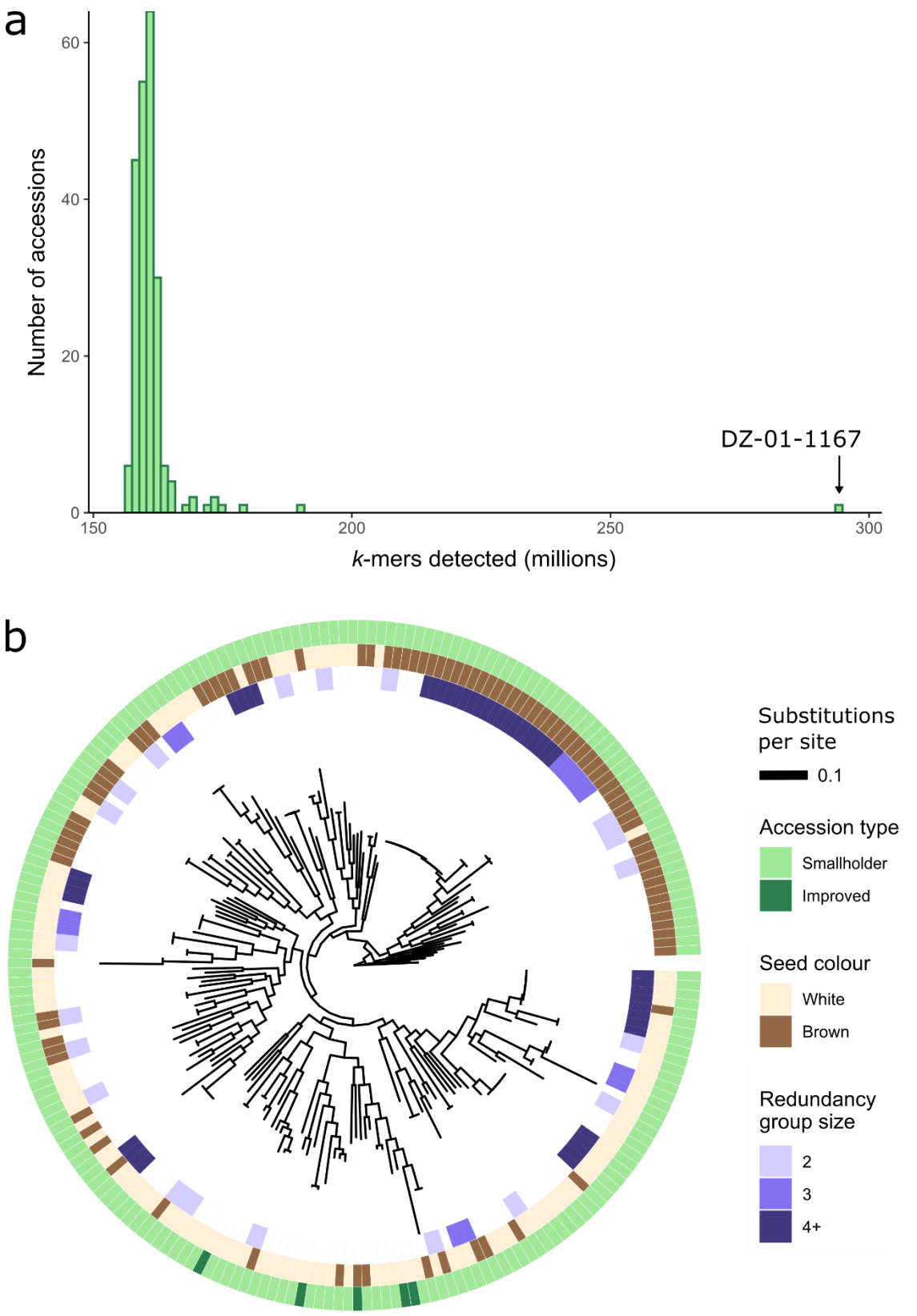


Figure 2.6 – Accession 'DZ-01-1167' contained a uniquely high number of distinct *k*-mers
a, Histogram showing number of distinct *k*-mers per sequenced tef accession (before collapsing redundancy groups). Accession DZ-01-1167 displayed a significantly higher number of distinct *k*-mers than any other accession (84% higher than the mean of other accessions). **b,** Phylogram of 220 tef accessions rerooted against DZ-01-1167. Brown and white accessions remained well-distributed and the same 32 redundancy groups were identified based on short phylogenetic distances.

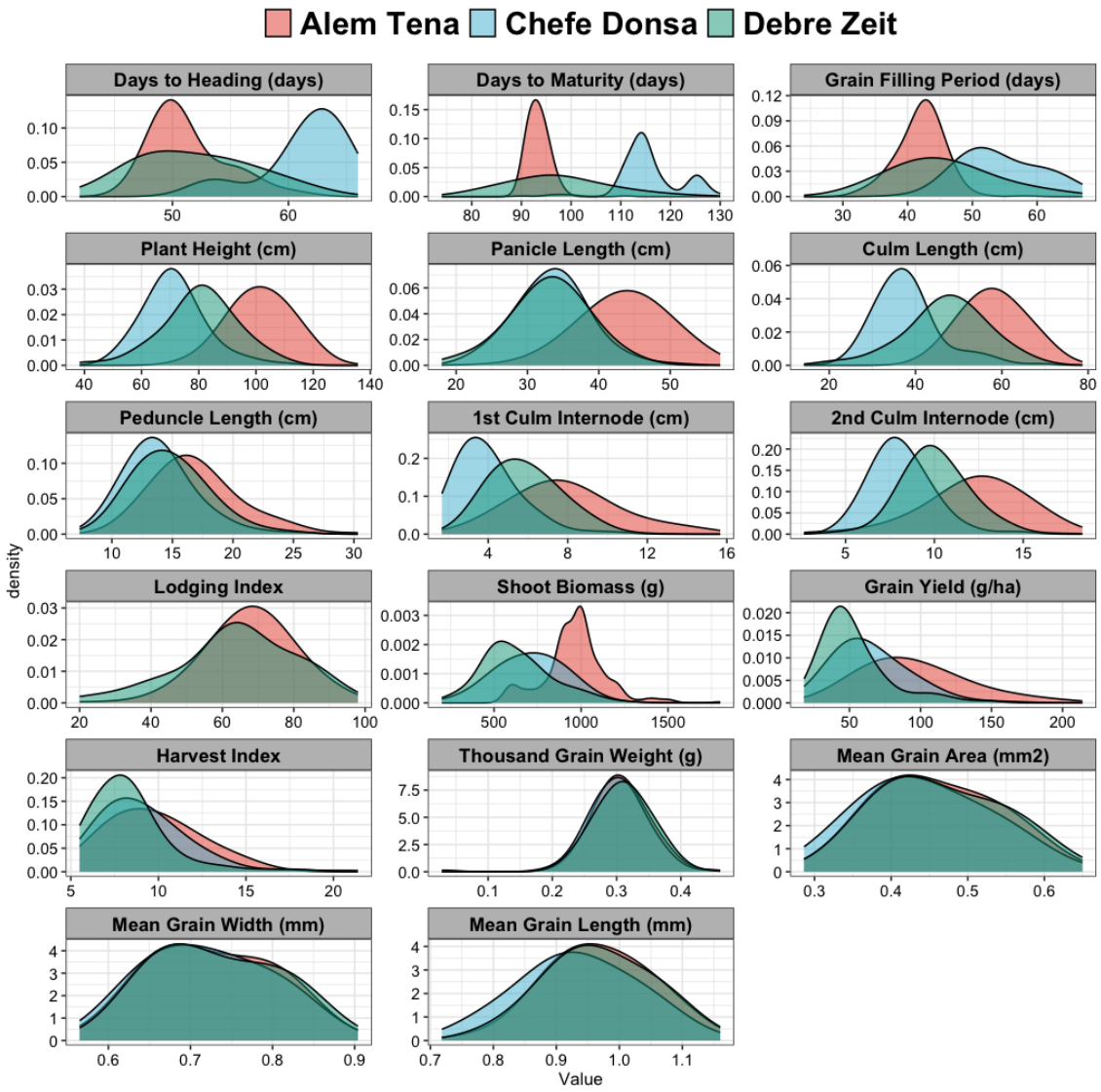


Figure 2.7 – Distribution of quantitative agronomic traits
Density plots showing the raw distributions of quantitative agronomic traits for each location.

To account for spatial variability within and between experimental locations, we used linear mixed modelling to generate genotypic best linear unbiased predictors (BLUPs) and broad-sense heritabilities (H^2) for each trait. The heritability for agronomic and grain morphometric traits ranged from 0.14 to 0.97 (Supp. Table 2.2). As expected, qualitative traits like panicle form and grain colour showed the highest heritability, 0.92 and 0.97, respectively. The heritability for grain morphometric traits, including grain length, width and area, was also high: 0.87, 0.90, and 0.89, respectively. Correlation analysis of trait BLUPs revealed expected associations between components of grain size (area, width, length) and weight, as well as between components of plant height (panicle length, plant height; Figure 2.8a).

Intriguingly, we also identified strong correlations between grain colour and both grain size and plant height. Plants of white-grained varieties were significantly taller than those of brown-grained varieties (Student's *t*-test, $p < 1 \times 10^{-5}$, Figure 2.8b). This has positive implications for straw yields but may increase lodging susceptibility. Brown-grained accessions are traditionally cultivated on more marginal soils, such as poorly drained vertisols, and, perhaps as a result, have come to be associated with smaller grains. However, our results indicated that brown-grained varieties tended to produce larger seeds than white-grained varieties when grown in common environments (Student's *t*-test, $p < 1 \times 10^{-15}$, Figure 2.8c). Despite this, there was no difference in TGW between white and brown-grained varieties (Student's *t*-test, $p = 0.22$), suggesting that, on average, white-grained varieties produce grains with higher densities.

Lodging index was, as expected, positively correlated with above-ground biomass, highlighting the trade-off faced by tef breeders between lodging rates and straw yields. However, we did not observe the previously reported correlation between lodging index and panicle morphology. This could be due to the relatively few lines (five) with the highest level of panicle compactness or the lower robustness of our lodging index BLUPs, given that this trait was only phenotyped at two of the three field sites.

2.3.3 – High-resolution metabolite fingerprinting shows differential metabolite accumulation in brown and white tef grains

The cultural preference for white-grained varieties in Ethiopia and Eritrea, as well as the growing international interest in tef's nutritional properties, motivated us to also explore variation in tef's grain metabolomes. We performed untargeted metabolite profiling using Flow Infusion Electrospray High-Resolution Mass Spectrometry (FIE-HRMS) analysis on

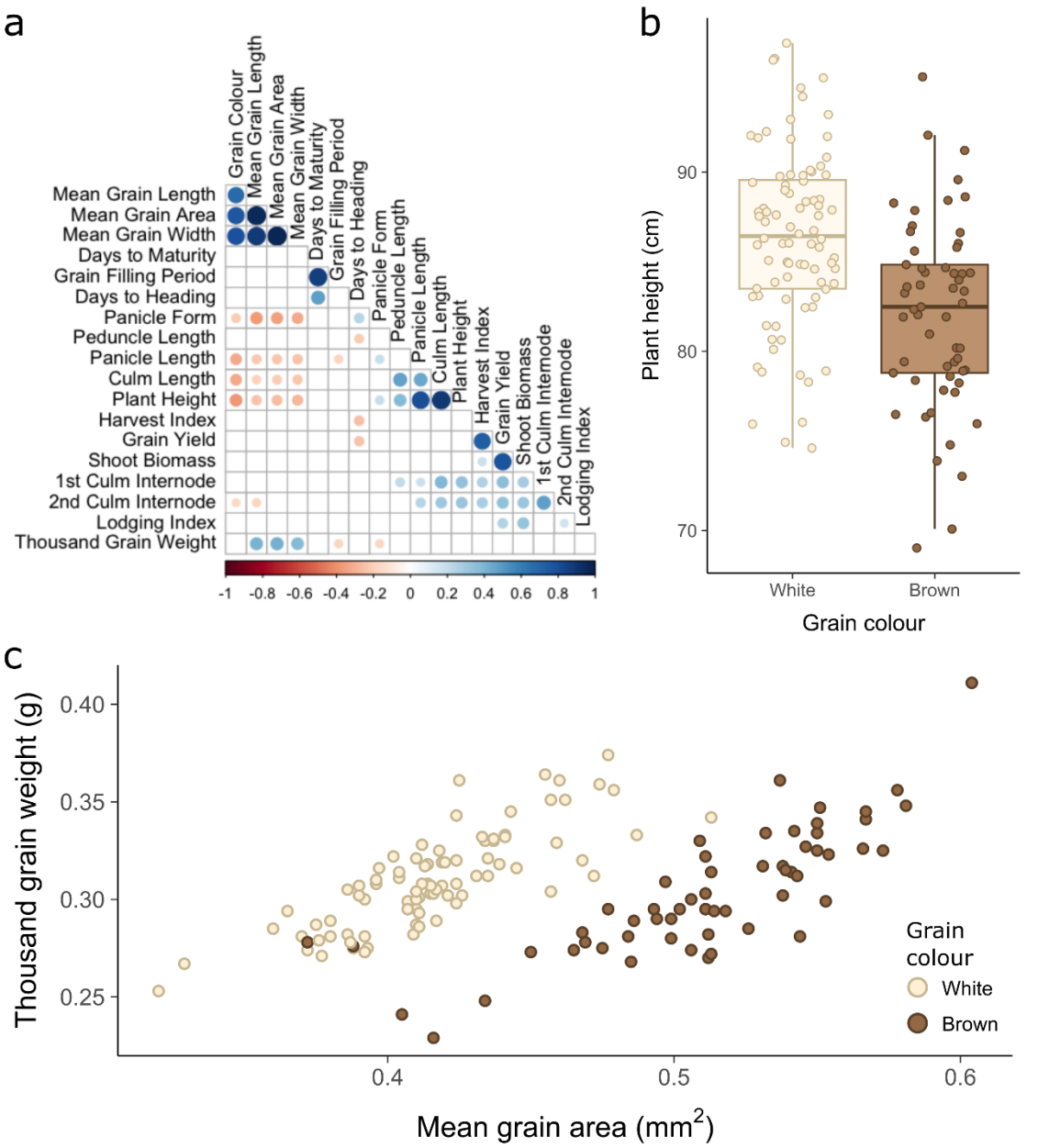


Figure 2.8 – Best linear unbiased predictors (BLUPs) reveal correlations between key agronomic traits

a-c, Analysis of BLUPs for $n = 141$ accessions and redundancy groups. **a,** Correlation tests were conducted between the BLUPs for 20 traits of interest. Significant correlations ($p < 0.05$) are indicated by circles whose size and colour represent the magnitude and direction of correlation. **b,** Boxplot of plant height BLUPs for white-grained ($n = 84$) and brown-grained ($n = 57$) varieties, with individual data points overlaid. White-grained accessions tended to produce taller plants. The centre line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles, and the whisker extends to 1.5 × interquartile range. **c,** Scatterplot of grain area BLUPs against thousand grain weight (TGW) BLUPs. A distinctly bimodal distribution is strongly explained by grain colour, with white-grained varieties tending to produce smaller grains. Despite this, their grains are of approximately the same mass as brown-grained varieties, suggesting higher grain densities.

grain samples from each plot of the three trial locations. A total of 1643 positively ionised mass-to-charge ratio (m/z) features and 1470 negatively ionised features were captured, of which 209 and 723, respectively, were differentially accumulated in brown and white-grained varieties (Student's *t*-test, FDR < 0.05).

From these differential *m/z* features, 183 could be tentatively identified using a rice (*Oryza sativa* ssp. *japonica*) reference metabolome library available in the KEGG public database (Kanehisa et al., 2023; Supp. Data 2.3). These differentially accumulated metabolites produced a clear separation of white and brown-grain samples when assessed by partial least squares discriminant analysis (PLS-DA; Figure 2.9a), but did not show differential accumulation between locations, suggesting little effect of locations on these metabolites (Figure 2.10). The differentially accumulated metabolites were enriched for various processes, including (unsaturated) fatty acid biosynthesis, linoleic acid metabolism, glutathione metabolism, porphyrin metabolism, flavone, and flavonol biosynthesis, and riboflavin metabolism (Figure 2.9b).

In agreement with other studies (Cotter et al., 2023), our results show that brown-grained varieties tend to have higher proportions of essential polyunsaturated omega-6 and omega-3 fatty acids (e.g., linoleic acid and alpha-linolenic acid, respectively) while white-grained varieties have higher levels of saturated fatty acids (e.g., caprylic acid and 9,10-epoxyoctadecanoic acid (EPOD)). Omega fatty acids have been associated with lowering cardiovascular disease, cancer, and autoimmune diseases (Day, 2004). However, the high levels of unsaturated fatty acids in brown-grained tef may also contribute to its increased proneness to rancidity and therefore its lower consumer appeal (Wallis et al., 2022; Figure 2.9c, Figure 2.11).

We also found differential accumulation of flavonoids between white and brown-grained varieties. These compounds can affect flavour and colour and act as antioxidants (Gebru et al., 2020). We found increased levels of the flavonoids rutin and 3-dehydroshikimate in brown-grained varieties. Meanwhile, in white-grained varieties, we observed elevated levels of apigenin and kaempferol 3-O-rhamnoside-7-O-glucoside (KOROG). Flavonols such as the latter have been known to contribute to white pigmentation (Dong & Lin, 2021; Figure 2.9c, Figure 2.11).

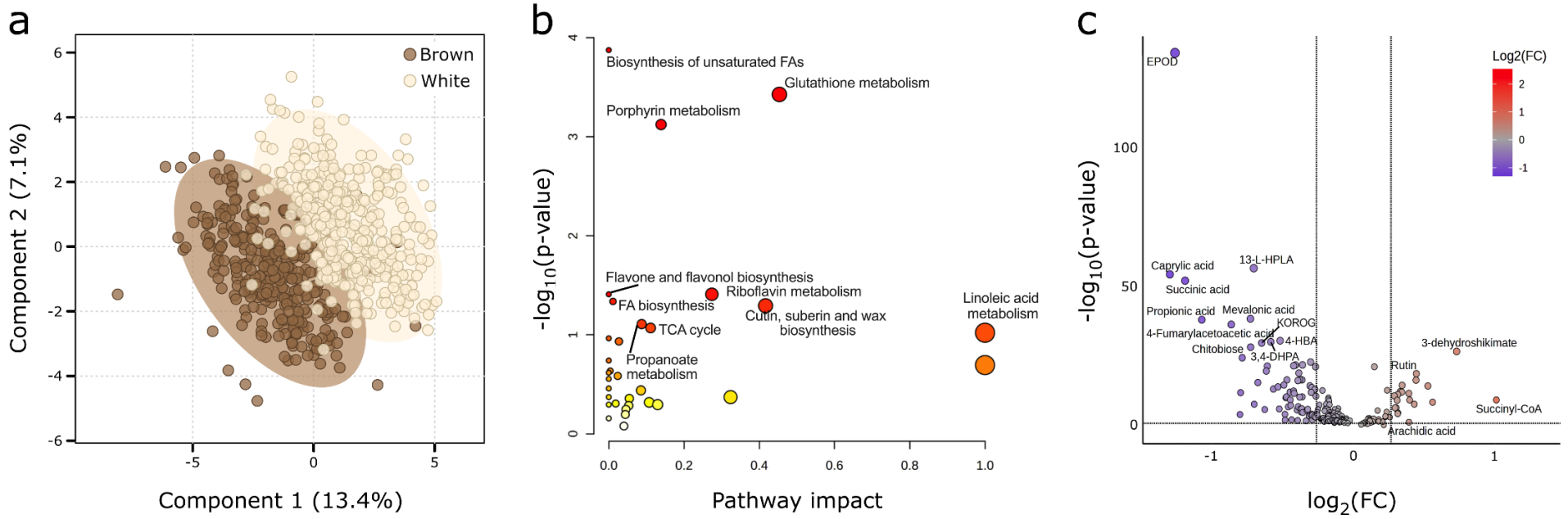


Figure 2.9 – Brown and white-grained tef accessions display differential metabolite accumulation
a, Partial least squares discriminant analysis (PLS-DA) of metabolites in grain samples of brown and white-grained accessions. Ellipses represent 95% confidence intervals around each group. **b,** Differentially accumulated metabolites show enrichment for several metabolic pathways, notably fatty acid and flavone metabolism. Point size scales with pathway impact and colour intensity scales with significance of pathway enrichment. **c,** Volcano plot for the 183 identifiable differentially accumulated metabolites. Fold-change (FC) was calculated as mean value in brown-grained varieties divided by that in white-grained varieties. Plotted FC thresholds are $\log_2(0.83)$ and $\log_2(1.2)$ and plotted FDR threshold is $\log_{10}(0.05)$.

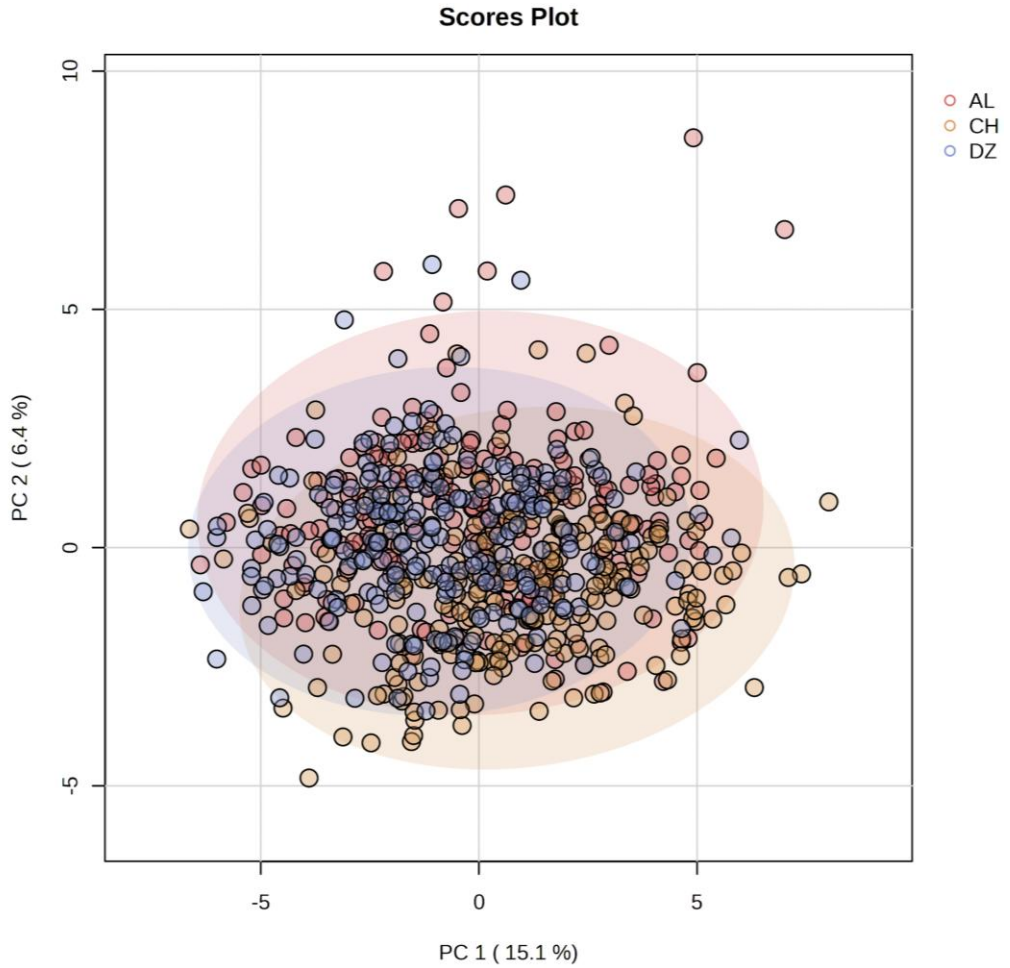


Figure 2.10 – Differentially accumulated metabolites between brown and white-grained accessions did not differ between locations
Principal component analysis of trial plots from all three locations based on the 183 annotated metabolites that were differentially accumulated between brown and white-grained accessions. AL, CH and DZ represent Alem Tena, Chefe Donsa and Debre Zeit, respectively. Points from the three locations were not well-separated, indicating little effect of location on the grain accumulation of these metabolites. Ellipses represent 95% confidence intervals around each group.

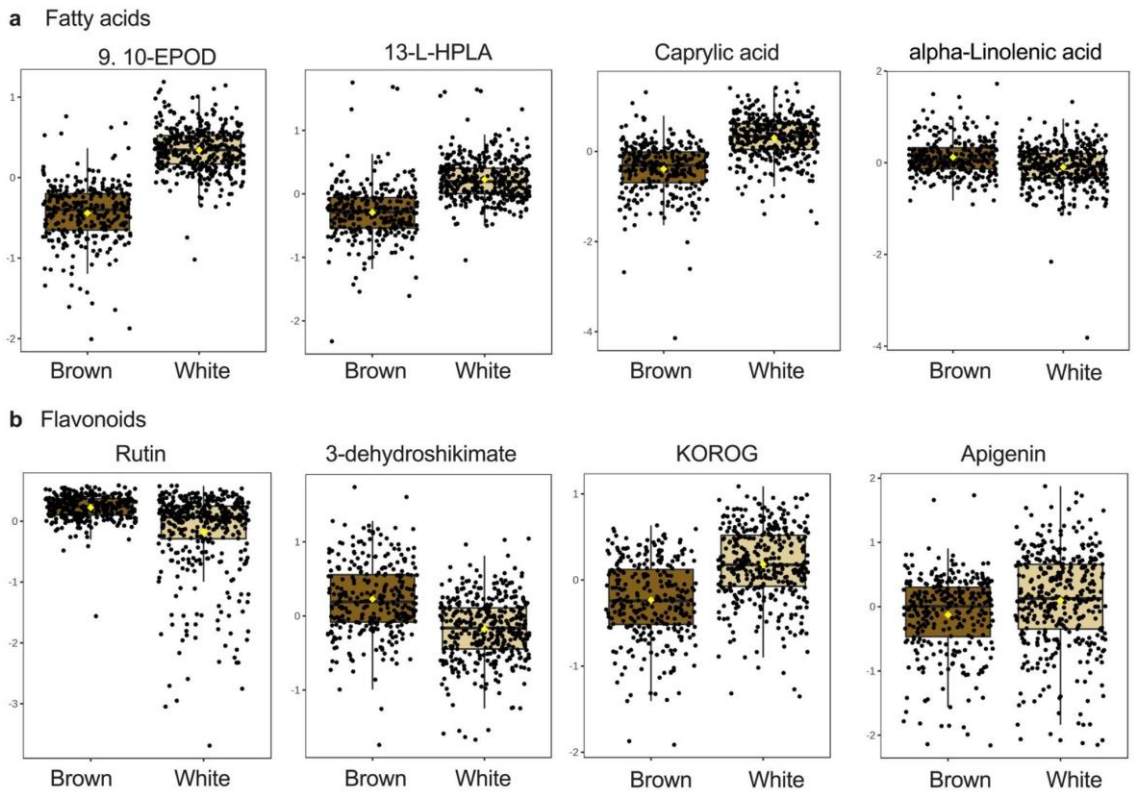


Figure 2.11 – Grain samples from brown and white-grained accessions differed in their accumulation of several fatty acids and flavonoids
a, fatty acid metabolites. **b,** flavonoid metabolites. Relative metabolite concentrations are shown as $\log_2$ transformed values and each dot represents biologically independent samples (brown n=303 and white n=389). The box ranges from 25% and the 75% percentiles; the 5% and 95% percentiles are indicated as error bars. Medians are indicated by horizontal lines within each box.

2.3.4 – *k*-mer-based GWAS identifies regions associated with panicle and grain morphologies

To identify genomic regions associated with the agronomic and grain morphology traits, we conducted a *k*-mer-based genome-wide association study (kGWAS) using the previously calculated BLUPs and a new *k*-mer matrix to account for the reduced number of non-redundant accessions. (See Gupta, 2021 and Karikari et al., 2023 for introductions to kGWAS and comparisons to traditional SNP-based GWAS.) Of the agronomic traits tested, we detected significant marker-trait associations (MTAs) for panicle morphology, grain morphology, and grain colour (Figure 2.12a, Supp. Table 2.3). Manhattan plots for the remaining traits are provided as Appendix B. We also carried out a SNP-based GWAS and identified four significant MTAs for panicle morphology, grain morphology, and lodging index (Supp. Table 2.4).

The control of panicle morphology appeared to be relatively simple, with a single highly associated 70-kb region on chromosome 3B (Figure 2.12b). The underlying reference *k*-mers negatively correlated with panicle morphology (scored as 1 to 4 for very lax to compact). This matched our expectations as the reference cultivar Dabbi produces very lax panicles. The significant region contains 13 gene models, including the tef orthologue (Et_3B_031395) of the rice gene *qSH1*/*RIL1* (*QTL for Seed Shattering on chromosome 1*/*RI-LIKE1*; Os01g0848400). *qSH1* is a *BEL1*-like homeobox TF linked with seed shattering and inflorescence architecture (Konishi et al., 2006; Ikeda et al., 2019). *qSH1* orthologues are also expressed in the inflorescence meristems of maize and wheat, suggesting a conserved role in inflorescence development across the grass family (Walley et al., 2016; Woodhouse et al., 2021).

There was a set of complex co-localised associations for grain morphology (Figure 2.12a). Most strikingly, a highly significant region (peak 7) supported by tens of thousands of *k*-mers was detected on chromosome 4B for grain colour and grain width (Figure 2.13a,b). This region was, as expected, also associated with grain area, but was not significant for grain length (Figure 2.14). There was also a region on chromosome 4A (peak 3) significantly associated with grain colour and width (Figure 2.13a,b). In addition, we found smaller regions on chromosome 3B (peak 2) and 4A (peak 6) co-associated with grain colour and grain width (Figure 2.12a, Figure 2.13a,b). The *k*-mers in each of the above regions were correlated with brown grain colour and increased grain size parameters. This supports the correlation between grain colour and grain size observed in the plotted BLUPs (Figure 2.8).

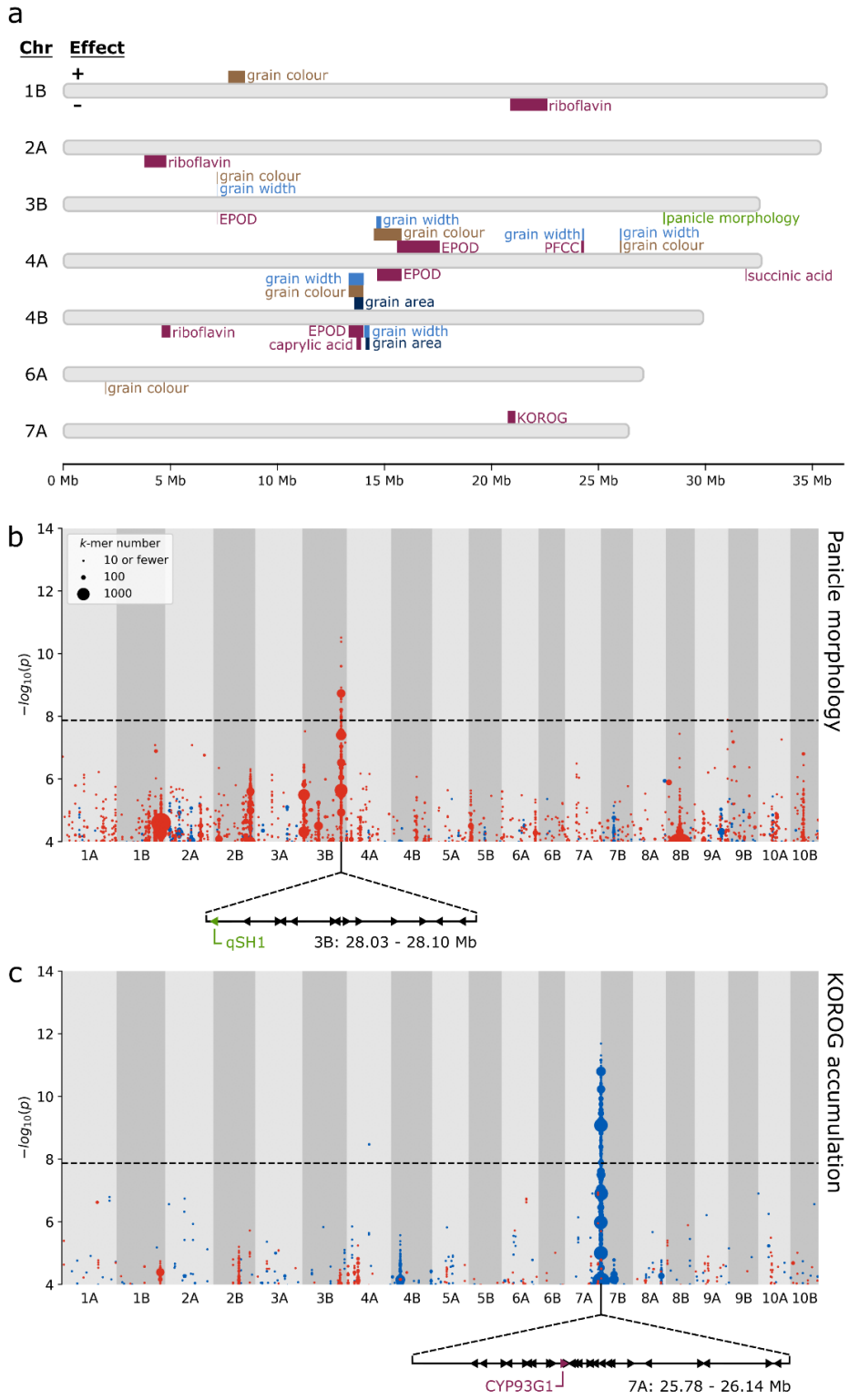
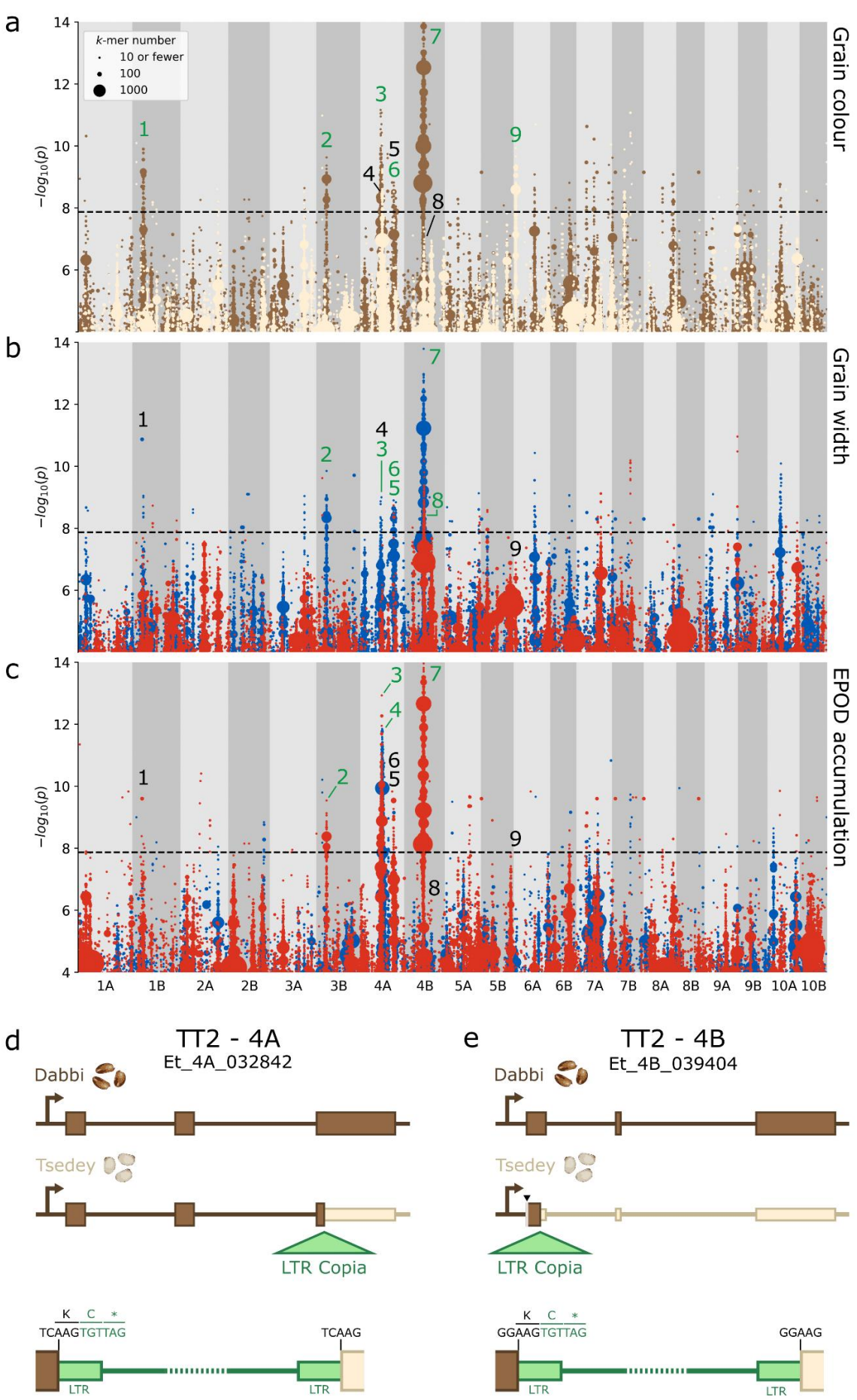


Figure 2.12 - *k*-mer-based GWAS identifies multiple marker-trait associations, including regions associated with panicle morphology and grain KOROG

a, Plot summarising all trait-associated regions identified by *k*-mer-based GWAS. Regions positively associated with traits are plotted above their respective chromosomes, while negatively associated regions are plotted below. For grain colour, positive and negative associations indicate brown and white, respectively. **b,** A region significantly associated with panicle morphology was detected on chromosome 3B. The arrangement of the 13 genes within this region is displayed below the plot. The candidate gene *qSH1* is highlighted. **c,** A region significantly associated with Kaempferol 3-O-rhamnoside-7-O-glucoside (KOROG) was detected on chromosome 7A. The arrangement of the 26 genes within this region is displayed below the plot. The candidate gene *CYP93G1* is highlighted. In **b** and **c**, *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association; red for panicle laxness or low KOROG, blue for panicle compactness or high KOROG. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10.



[Caption on next page]

Figure 2.13 – Co-association of grain colour, width, and EPOD concentration with multiple regions

Plots of *k*-mers associated with **a,** grain colour, **b,** grain width, and **c,**grain EPOD concentration. *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association. In (**a**), brown denotes association with brown grain colour and white with white grain colour. In (**b** and **c**), red denotes association with lower trait values, and blue with higher trait values. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10. Nine regions are labelled with black and green numbers, denoting whether the region is significant or not significant for the plotted trait, respectively. Diagrams of LTR Copia insertions into *TT2* homoeologs on **d,** chromosome 4A, **e,** chromosome 4B. Top: structure of *TT2* in 'Dabbi' (brown-grained). Centre: structure of *TT2* in 'Tsedey' (white-grained). Bottom: detail of LTR Copia insertions. Narrower exons indicate presumed protein truncations. Black DNA bases denote 5 bp target-site duplications. Green DNA bases denote the start of the retrotransposon insertions. Single-letter amino acid codes show the introduction of premature stop codons (*). The first 22 bp of the *TT2* open reading frame on 4B is not assembled in the Tsedey genome (greyed out, black arrowhead). Gene annotations derive from (VanBuren et al., 2020). and do not include 5' and 3' untranslated regions.

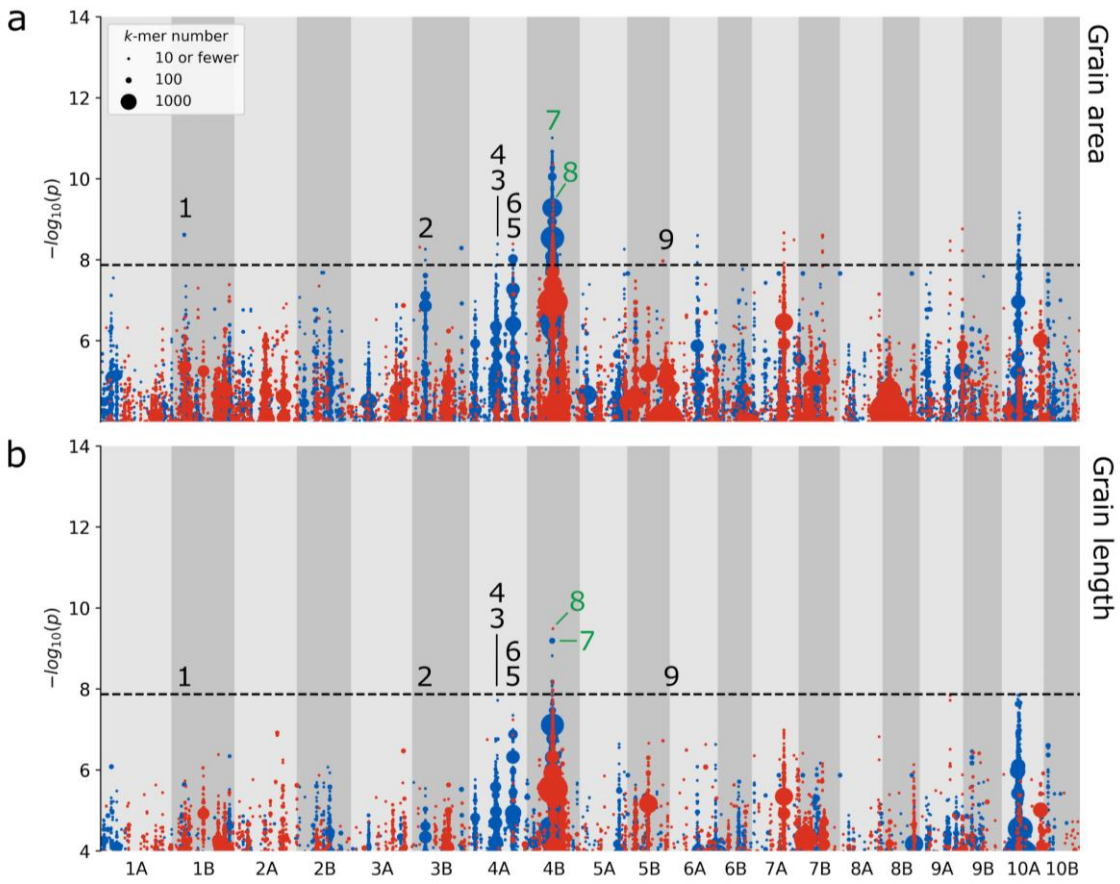


Figure 2.14 – *k*-mer-based GWAS identifies marker-trait associations for grain area, but not grain length

Plots of *k*-mers associated with **a**, grain area and **b**, grain length. *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association; red for association with lower trait values, and blue for association with higher trait values. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10. The nine highlighted regions from Figure 7 are labelled with black and green numbers, denoting whether the region is significant or insignificant for the plotted trait, respectively.

In contrast to the regions discussed above, there were also cases where grain colour and grain size were decoupled. A third positive grain width peak on chromosome 4A (Figure 2.13 peak 5) is well-separated from the upstream and downstream grain colour peaks, by 8460 kb and 1580 kb, respectively. On chromosome 4B there is a region negatively associated with grain width and area (Figure 2.13 peak 8) that is separated by just 10 kb from peak 7, a major peak for brown grain colour and higher grain width and area. A marginally insignificant peak for grain width and area exists on chromosome 10A (Figure 2.13b, Figure 2.14). Lastly, there are two grain colour peaks on chromosomes with no significant grain size peaks (Figure 2.13). This includes a 790 kb peak associated with brown grains on chromosome 1B (peak 1, 114 genes) and a 40 kb peak associated with white grains on chromosome 6A (peak 9, 6 genes). These two regions offer the strong possibility of breeding for grain colour independently of grain size.

2.3.5 – Associated regions for grain metabolites and grain morphology co-localise

We hypothesised that genomic loci associated with grain colour might also be associated with the differentially accumulated metabolites. To facilitate GWAS analysis, we calculated BLUPs and broad-sense heritabilities for 21 grain metabolites with high fold-change differences and/or which are known to be involved in pigmentation or important for human nutrition (Supp. Data 2.3). Heritability values were generally high, with 18 of the 21 metabolites having $H^2 > 0.50$ (Supp. Table 2.5). kGWAS revealed significant regions for riboflavin, EPOD, primary fluorescent chlorophyll catabolite (PFCC), succinic acid, caprylic acid, and KOROG (Figure 2.12a, Supp. Table 2.3). Manhattan plots for the remaining traits are provided as Appendix B. SNP-based GWAS also identified nine MTA for four metabolite traits, five of which overlap with the significant region for kGWAS (Supp. Table 2.4).

As we hypothesised, kGWAS associations for some metabolites, including EPOD, caprylic acid and PFCC, co-localise with regions associated with grain colour and/or size. Of these, EPOD showed the most consistent and significant overlap with grain colour and grain width, on chromosomes 3B (peak 2), 4A (peak 3) and 4B (peak 7, also overlapping a grain area peak; Figure 2.13c). In these co-localised regions, EPOD was negatively associated with grain colour (i.e brown-grained varieties had lower EPOD content). There was also a significant peak for EPOD on chromosome 4A (peak 4) that partially overlapped, and was positively correlated with, a grain colour peak (peak 3). The 4A peaks for EPOD (peaks 3 and 4) were also identified in our SNP-based GWAS. Caprylic acid was associated with a single locus that overlapped with peak 7 for EPOD, grain colour, and grain width. Similarly, PFCC

was associated with a single region in both *k*-mer and SNP-based GWAS that also overlapped with grain width.

We also found associations for other metabolites, including KOROG, succinic acid, and riboflavin, that did not overlap with grain colour and/or size. KOROG was associated in both *k*-mer and SNP-based GWAS with a single prominent 360 kb region on chromosome 7A (Figure 2.12c) containing 26 gene models. This included Et_7A_050580, which encodes a cytochrome P450 (*CYP93G1*) with flavanone 2-hydroxylase activity and has been previously shown to be involved in the biosynthesis of flavonol glycosides like KOROG (Lam et al., 2014). Succinic acid was also associated with a single region, in this case spanning 40 kb on chromosome 4A and containing eleven genes. Control of riboflavin accumulation appears more complex, with three regions on chromosomes 1B (1730 kb, 174 genes), 2A (1030 kb, 126 genes), and 4B (420 kb, 75 genes) associated with riboflavin accumulation but that are not associated with other traits.

2.3.6 – *TRANSPARENT TESTA 2* is a candidate for grain colour variation

Given that the most prominent associations for grain colour, size and metabolites content cluster at peak 3 (chr 4A: 14.48–15.79 Mb) and peak 7 (chr 4B: 13.34–14.05 Mb) (Figure 2.12a, Figure 2.13a,b, Supp. Table 2.3), we examined the gene content in these peaks to identify potential candidate genes. Interestingly, these peaks displayed partial homology; genes in the proximal end of peak 3 are homoeologous to genes in the distal end of peak 7 (Figure 2.15). Homoeologous gene pairs in these regions include Et_4A_032844/Et_4B_037039 and Et_4A_032842/Et_4B_039404, whose orthologues have been previously shown to regulate grain colour, size and fatty acid content. The former are orthologues of *NUCLEAR FACTOR YA3* (*NF-YA3*), which regulates seed oil content and seed size across diverse angiosperms, including *Arabidopsis* and oil palm (Yeap et al., 2017). The latter are orthologues of *TRANSPARENT TESTA 2* (*TT2*), a MYB TF that is associated with proanthocyanidin accumulation in the seeds of various Brassicaceae species (Nesi et al., 2001; Ren et al., 2017).

We compared the gene sequences of the *TT2* orthologues from the Dabbi reference genome (a brown-grained variety) to those from the published draft assembly of the white-grained variety ‘Tsedey’ (Cannarozzi et al., 2014). In the Tsedey assembly, we identified striking insertions of long terminal repeat (LTR) Copia superfamily retrotransposons (RTs) in the third and first exons of the A and B *TT2* homoeologues, respectively (Figure 2.13d,e, Supp. Data 2.4 and 2.5). The A subgenome RT introduces an in-frame cysteine and then a premature stop codon, truncating most of the final exon (162 codons). The B-subgenome

RT is inserted in the first exon, truncating most of the protein. The positions of the two elements in different exons suggest independent insertions occurring after subgenome divergence. In both RTs, the 5 bp target-site duplications and 106 bp LTRs remain undegraded, suggesting a relatively recent insertion (Wicker & Keller, 2007). This is consistent with the large number of recently active LTR RT families previously identified in tef. We did not find any protein-truncating mutation between Dabbi and Tsedey in the A and B homoeologues of *NF-YA3*; the B homoeologue contains one non-deleterious missense mutation, while the A homoeologue contains no missense mutation.

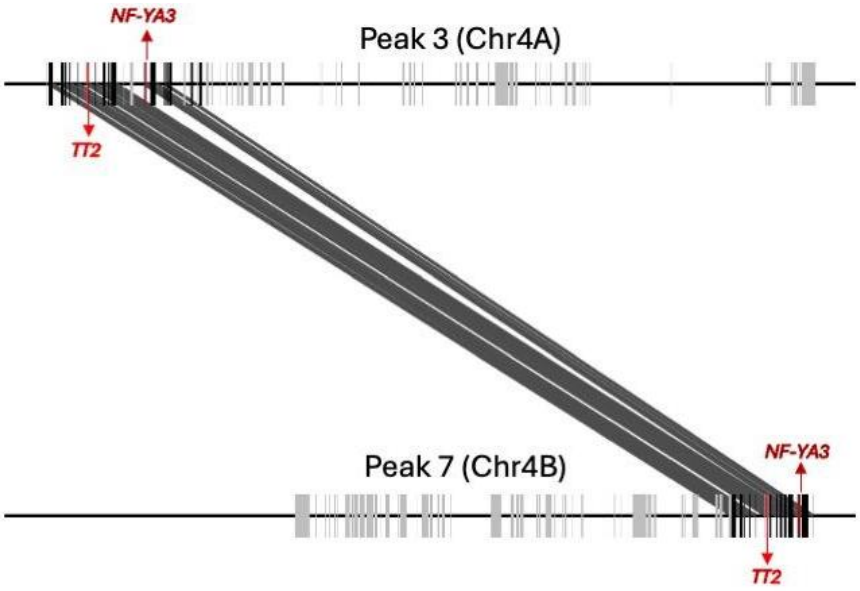


Figure 2.15 – Homoeology in associated regions for grain size, colour and metabolite on Chr4A and Chr4B
Homoeologous relationship between genes in peak 3 and peak 7. Vertical bars represent individual genes and diagonal lines connect genes with homoeologous relationships. Candidate genes within these intervals are denoted with red arrows pointing upwards (forward strand) or downwards (reverse strand).

2.4 – Discussion

Developing genomic resources for underutilised crops is crucial for accelerating their improvement, adoption, and utilisation, and will in turn boost the resilience of the interconnected global food system (Chapman et al., 2022; Shorinola et al., 2024). Our extensive phenotyping and whole genome resequencing of a diverse tef collection represents a valuable resource for germplasm characterisation and trait mapping in this locally vital and globally emerging crop.

While grown in Ethiopia and Eritrea for thousands of years, systematic collection and breeding of tef did not begin until the 1950s, with sampling of varieties directly from farmers' fields. Since then, numerous germplasm collections have been established, containing over 7000 accessions. These are primarily maintained in Ethiopia, but smaller collections exist elsewhere (Chanyalew et al., 2021; Woldeyohannes, Desta, et al., 2022). Correspondence of varieties between these collections is undocumented, preventing cross-utilisation of phenotyping and sequencing data. Genomics approaches have been invaluable for resolving such issues and for identifying redundancy within collections (Mascher et al., 2019; Milner et al., 2019). Our work reveals redundancies in the EIAR tef collection, which are likely due to repeated sampling of farmer-traded germplasm across modest geographical ranges. The compact SNP panel we have developed can be used to identify further redundancy within the EIAR collection. The resequencing data presented here can also be combined with existing mid-density genotyping data from other tef collections to assess redundancies, differences, and complementarity between the different tef collections globally (Woldeyohannes, Iohannes, et al., 2022; M. D. Alemu et al., 2024). This will shed further light on tef's breeding history, facilitate germplasm exchange, and inform the selection of accessions for a tef pan-genome to optimally capture tef diversity.

We identified a strong candidate gene for panicle morphology, Et_3B_031395, which is orthologous to the rice *qSH1* gene. qSH1 is a BEL1-like homeodomain protein that underlies variation in seed shattering in rice (Konishi et al., 2006) and is regulated by *SUPERNUMERARY BRACT* (Jiang et al., 2019), whose direct orthologue in wheat is the major inflorescence morphology gene *Q*, which controls both seed shattering and inflorescence compactness (Simons et al., 2006). The *Arabidopsis* orthologue of *qSH1*, *REPLUMLESS/PENNY* (Roeder et al., 2003), also known as *PENNYWISE* (Smith & Hake, 2003) or *BELLRINGER* (Byrne et al., 2003), is important for fruit development and dehiscence. *qSH1* has also been directly connected with inflorescence architecture, with

qsh1 ri (*verticillate rachis*) double mutants displaying abnormal timing and arrangement of primary branch meristems (Konishi et al., 2006; Ikeda et al., 2019). *qSH1* also strongly influences bract suppression in the inflorescence (Konishi et al., 2006; Ikeda et al., 2019). In addition, paralogs and orthologues of *qSH1* have been shown to control inflorescence patterning in maize and *Arabidopsis* (Byrne et al., 2003; Smith & Hake, 2003; Tsuda et al., 2017). Given the established roles of its orthologues in diverse plant species and its localisation within a narrow candidate region, Et_3B_031395 emerges as a promising candidate for the regulation of panicle morphology in tef. Nonetheless, functional validation will be necessary to confirm its role.

Most farmers in Ethiopia rely on manual broadcasting for sowing on small plots (0.25 to 1 hectare), as access to mechanisation is limited and such equipment is typically not optimised for the tiny seeds of tef (Fikadu et al., 2020; Tadele & Hibistu, 2021). This practice leads to inefficient seed use as farmers typically sow at higher rates than recommended to ensure good field coverage (15–25 kg/ha, instead of 5 kg/ha; Assefa et al., 2022). These high seeding rates also produce overcrowded fields of weak-stemmed plants more prone to lodging (Ben-Zeev et al., 2020). Larger grains would facilitate mechanised handling and ensure seedlings have sufficient nutrients for establishment from greater soil depths. Together, this would promote row-based drilling of tef, alleviating the above issues and supporting the ongoing transformation of tef cultivation practices. Indeed, agronomists have already experimented with pelleting tef seeds with inert material to enable mechanised sowing, highlighting the promise of this approach (Cannarozzi et al., 2018). Lastly, increasing grain size could help reduce the high grain loss rates experienced by smallholders during traditional threshing and winnowing processes (Tiguh et al., 2024).

Our kGWAS results offer potential breeding targets for increasing grain size. We identified six genomic regions significantly associated with grain width, two of which were also associated with grain area. However, we also identified a previously unknown link between grain size and grain colour, which could complicate this process. Brown-grained varieties tended to produce larger grains (0.51 mm^2) than white-grained varieties (0.42 mm^2), and this was reflected in the kGWAS; four of the regions positively associated with grain width were also associated with brown grain colour. This co-localisation presents an issue because there is a strong cultural preference in Ethiopia for white tef flour, and this translates into a market incentive for farmers to grow white-grained varieties. Introgression of grain size alleles into elite white-grained varieties at the cost of increased pigmentation would therefore not be favourable.

However, not all grain size and grain colour loci were co-localised. We identified one grain width locus on chromosome 4A that is very distant from the two grain colour regions on this chromosome, plus two grain colour loci on chromosomes which do not harbour grain size loci (1B and 6A). We also observed a region on chromosome 7A strongly associated with the metabolite KOROG. While our analysis did not find this region to be co-associated with grain colour, flavonols such as KOROG have previously been associated with white pigmentation (Dong & Lin, 2021). These regions offer opportunities to combine favourable grain size and colour alleles through introgression, though it is yet to be seen if the introduction of 'white grain' alleles into a brown-grained background would yield a dominant effect. It is more likely that positive breeding outcomes could be achieved by stacking multiple additive grain size loci. Uncovering further such loci should be a priority for future GWAS studies in tef. It is also important to note that the use of a high-throughput and high-accuracy phenotyping platform (MARViN grain analyser) was key to uncovering these grain size variations. While routine for major crops such as wheat, this is the first application of high-resolution, image-based grain measurements to a tef panel to our knowledge. This exemplifies the benefits that adopting robust and well-tested phenotyping methodologies from mainstream crops can bring to the research of underutilised crops (Shorinola et al., 2024).

Another route to achieving large-grained white tef varieties could be to knock out key TFs or enzymes linked with pigmentation in a brown-grained background (through mutation breeding, transgenesis, or genome editing). To implement such an approach, the tef research community will need to increase its understanding of relevant genes and their pleiotropic effects. Contributing to this, we identified a candidate pair of homoeologues present within each of the two major loci for grain size, colour, and the fatty acid EPOD. Et_4A_032842 and Et_4B_039404 are orthologous to *TRANSPARENT TESTA 2* (*TT2*), an R2R3 MYB TF known to regulate seed coat colour in the Brassicaceae family through proanthocyanidin formation and accumulation (Nesi et al., 2001; Ren et al., 2017).

We propose that variations in the two tef orthologues of *TT2* contribute to variation in grain colour. This aligns with a previous report of two "duplicate" genetic factors *B/b and B2/b2* as the major determinants of brown/white pigmentation (Berhe, 1981). Our hypothesis is further supported by our discovery of independent insertions of related LTR-Copia retrotransposons in both homoeologues of the white-grained tef variety Tsedey. These insertions, which are absent in the genome of the brown-grained variety Dabbi, would both lead to truncated and likely non-functional proteins. These *TT2* polymorphisms could also underlie the variation in fatty acid content and grain size, as has been shown in *Arabidopsis* and *Brassica napus* (Chen et al., 2012; Zhou et al., 2016).

This hypothesis could be tested by mutagenising *TT2* in a brown-grained variety. However, currently, we cannot preclude the association of these traits with alternative or additional candidate genes. The homoeologues Et_4A_032844 and Et_4B_037039 lie within the same two loci and are orthologous to *NF-YA3*, a gene which activates oil accumulation in oil palm mesocarp and increases oil content and seed size when overexpressed in *Arabidopsis* (Yeap et al., 2017). It is, therefore, possible that the correlated traits (grain colour, size and fatty acid) are modulated by two separate gene families in linkage blocks.

While the above manipulation is an intriguing possibility, we acknowledge that the original cultural preference for white-grained tef varieties is likely linked to flavour and baking properties in addition to aesthetics, although this has not been well-studied. It is therefore also important to study the tef metabolome beyond its direct contribution to colour. Future studies could utilise our extensive metabolome profiling data to conduct a more comprehensive metabolite GWAS (mGWAS) and deepen our understanding of the genetics underpinning differential metabolite accumulation in brown and white-grained tef.

Our work demonstrates the value that can be brought to underutilised crops such as tef by applying resequencing and population genomics in combination with large-scale phenotyping and metabolome profiling. We identify multiple genetic loci for morphological and nutritional traits and suggest how these could inform future research and breeding efforts, including contribution to new tef varieties with desirable plant architecture and consumer-preferred grain traits. Other underutilised crops could also greatly benefit from such methods to accelerate their domestication or improvement.

2.5 – Methods

2.5.1 – Germplasm and phenotyping

A core panel of 225 tef accessions was selected from the broader EIAR collection to capture a broad range of phenotypic diversity. This panel consisted of 220 smallholder varieties and 5 improved, registered varieties (Supp. Data 2.6). The EIAR collection is originally derived from 2175 tef germplasm accessions (Ketema, 1993), 35 cultivars (Ebba, 1975), and 10 released improved varieties (Assefa et al., 2011).

The accessions were grown in Ethiopia at three EIAR research sites: Alem Tena, Chefe Donsa, and Debre Zeit. Chefe Donsa and Debre Zeit sites represent non-stressed environments, while Alem Tena represents a moisture-stressed environment (Supp. Table 2.6). Each field trial was set up using an augmented block design (Supp. Data 2.7) and consisted of 1 m rows sown with 0.3 g grain and spaced 50 cm apart. Most accessions were sown once per field site, but five improved varieties (Ebba, Boset, Bora, Dagim and Felagot) were sown four times per field site as controls (incorporated via linear mixed modelling, see section 2.5.6).

Data were collected on a range of qualitative and quantitative traits (Supp. Data 2.8). Phenotyping methodology was derived from the Tef Breeding Manual (Chanyalew et al., 2021) and is described in detail in Supp. Table 2.7. Qualitative traits included basal stalk colour, grain colour, panicle colour, and panicle form. Quantitative traits included phenology (days to heading, days to maturity, and grain filling period) and agro-morphology (plant height, panicle length, peduncle length, culm length, first culm internode length, second culm internode length, above-ground shoot biomass, grain yield, harvest index, and lodging index). Lodging index was only assessed at Alem Tena and Debre Zeit.

Grain samples from each of the 720 rows across the three field trials were analysed using a MARViN Grain Analyzer. For each sample, 0.075–0.085 g of grain (mean of 262 grains) was evenly distributed on the imaging tray. Mean grain area and TGW were recorded, as well as mean, minimum, and maximum values for grain width and length. Insufficient grain was harvested from accession Trotteriana-T-138 for MARViN analysis.

2.5.2 – DNA extraction and resequencing

For each accession, ~0.7 g fresh leaf tissue was collected from 3-week-old plants from the Alem Tena field trial. DNA was extracted using DNeasy Plant or DNeasy Plant Pro kits

(QIAGEN) and eluted in 50 µL AE buffer. Sufficient high-quality DNA could not be extracted for the accession Trotteriana-T-138. For three further accessions ('DZ-01-170', 'DZ-01-1015', 'Gealamie-T-111'), the observed grain colour at Alem Tena did not match that observed at the other two field locations, suggesting heterogeneity in the seed stock. Given that the DNA sample would, therefore, not represent the majority of the phenotyping data, these accessions were not sequenced and therefore not used for GWAS. The remaining 221 DNA samples were sequenced by Novogene UK (Illumina paired-end 150 bp), and data were returned for 220 accessions (no data were produced for 'DZ-01-12').

2.5.3 – SNP calling, LD calculation, and phylogenetic analyses

Raw sequencing reads were trimmed using fastp (Chen et al., 2018) and mapped to the *Eragrostis tef* reference genome (cv. Dabbi) using Bowtie2 (v2.4.1; Langmead & Salzberg, 2012; VanBuren et al., 2020). The mapped reads were filtered for MAPQ scores >30 using SAMtools (v1.18), and a VCF file was generated using BCFtools (v1.18; Li et al., 2009; Bonfield et al., 2021; Danecek et al., 2021). VCF statistics were generated using VCFtools and examined using base R (v4.1.3; Danecek et al., 2011; R Core Team, 2023). Empirically derived filters were applied using VCFtools (--max-missing 0.90; --minQ 30; --minGQ 15; --min-meanDP 10; --max-meanDP 18; --minDP 5; --maxDP 23; --maf 0.025; Purcell et al., 2007) and linkage pruning was conducted using Plink (v1.90b4.6; --allow-extra-chr; --indep-pairwise 20 5 0.5). The final VCF file of 41,289 SNPs was converted to PHYLIP format using the tool vcf2phylip (v2.9; Ortiz, 2019).

TASSEL (v5.2.54; Bradbury et al., 2007) was used to compute pairwise intra-chromosome LD correlation coefficient (r^2) between SNP markers across the entire tef genome. LD decay scatterplot was then produced by plotting the r^2 values against physical distance (bp) using R software. The intersection point between the genome-wide LD curve and the r^2 threshold (0.2) determined the genome-wide LD decay value.

We generated a phylogram from the filtered SNPs using IQ-Tree 2 (v2.3.2; -B 10000; --msub nuclear; -m MFP + ASC; --seed 42; Minh et al., 2020). We attempted to root the phylogram against an outgroup, *Eragrostis curvula*, using short-read sequencing data from (Carballo et al., 2023; NCBI SRA SRR22846089). Unfortunately, after aligning this data to the *E. tef* 'Dabbi' reference genome, insufficient SNPs could be called to enable inclusion in the phylogram. We could not locate suitable sequencing data to test the use of additional *Eragrostis* species as outgroups. We therefore arbitrarily rooted the phylogram against 'Ada-T-58' based on alphabetical order, following the default settings for IQ-Tree 2. We also tested an unrooted phylogram for comparison but found that it was less compact and less readily

annotatable than a circular, rooted topology. As a result, it failed to convey the main message of the figure: that there was considerable redundancy within the studied germplasm collection. Given this aim for the figure – as opposed to a detailed exploration of the evolutionary and domestication history of this collection – use of an arbitrary root was deemed the best option.

The phylogram was visualised in R using ggtree (v3.10.1) and ggtreeExtra (v1.12.0; Yu et al., 2017; Xu et al., 2021). Population structure was investigated by applying ADMIXTURE analysis (v1.3.0, K = 1:20; --cv = 10; Alexander et al., 2009) and principal component analysis (SNPRelate v1.29.0; Zheng et al., 2012) to the same VCF file. The results were visualised using Pophelper (v2.3.1) and Plotly (v4.10.1), respectively (Francis, 2017; Sievert, 2020). After defining the redundancy groups, the name of a single accession from each group was arbitrarily assigned to represent the group in subsequent analyses (Supp Table 2.1). The phylogram in Figure 2.6 was generated and visualised using the same settings.

A *k*-mer presence/absence matrix was computed for all 220 sequenced accessions from trimmed reads using scripts from a previously published *k*-mer-GWAS pipeline (https://github.com/wheatgenetics/owwc/tree/master/kGWAS, section 1; Gaurav et al., 2022). Additional guidance on the use of this pipeline is provided at https://github.com/quirozcj/kmerGWAS_descriptions. Default parameters were used for all steps. Notably, the default *k*-mer length and minimum *k*-mer frequency per accession were not modified (-m 51 and -L 4, respectively). Shared *k*-mer state rates were computed using custom bash and R scripts (https://github.com/Uauy-Lab/tef_kGWAS_2024).

2.5.4 – Minimal SNP panel selection

The trimmed reads from accessions belonging to redundancy groups were pooled and subsampled down to the average read number per single library (29,350,529 paired-end reads). A VCF file was generated as above for the 150 redundancy groups and singlets. The same filters and linkage pruning were applied. This VCF file was input to the Minimal Marker pipeline (Winfield et al., 2020). This involves conversion to a genotype matrix, then selection of SNPs. However, prior to SNP selection, the pipeline was modified to convert all heterozygous calls to missing data. Heterozygous loci are unstable between generations and would therefore be unreliable markers for consistently identifying accessions across generations. In contrast, the original target species for the pipeline, apple (*Malus domestica*), is largely propagated vegetatively, so heterozygous loci are stably inherited between generations. Missing calls are not used by the pipeline to distinguish accessions,

so this change forced the pipeline to select a panel of SNPs that uniquely identifies the core collection using only loci homozygous across the 150 tef redundancy groups and singlets. The first run selected 14 SNPs fulfilling this remit. To provide redundancy, these SNPs were removed from the genotype matrix and a second run was conducted, leading to the selection of a further 14 SNPs.

An additional consideration was whether the genotypes of the individual members of the redundancy groups were consistent with the overall group genotype. This was investigated using custom Bash and R scripts (https://github.com/Uauy-Lab/tef_kGWAS_2024), and the results are summarised in Supp. Data 2.2 and Supp. Note 2.1. There were no cases where group members' genotypes compromised the utility of the SNP set for unique identification of the 150 non-redundant accessions.

2.5.5 – Metabolite extraction, profiling, and statistics

Methanolic metabolite extractions were conducted at the International Livestock Research Institute (ILRI, Ethiopia) on 40 mg grain from each trial plot following a previously published protocol (López-Álvarez et al., 2017). Briefly, tissue was ground (Tissue Lyser (QIAGEN), 25 Hz, 2 min), added to 1 mL pre-cooled 100% methanol (−20 °C), and placed on ice for 30 min with vortexing every 5 min. The extracts were then centrifuged, vacuum concentrated, and shipped to Aberystwyth University, Wales, UK, for high-resolution metabolite profiling. The samples were resuspended in 300 µL of pre-cooled 100% methanol, vortexed for 5 min and centrifuged at 1000×*g* at 4 °C for 5 min. An aliquot of 200 µL of each sample was used for untargeted metabolite fingerprinting using flow infusion electrospray high-resolution mass spectrometry (FIE-HRMS) mode using Q Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo-Scientific, UK) where data was captured in negative and positive ionisation mode. Quality controls were derived from a master mix sample where 10 mL of each extract was pooled and also "blanks" of 100% methanol. Three 20 µL injections were performed for each sample as technical replicates. FIE-HRMS metabolite fingerprints in both positive and negative ionisation modes in a single run. 20 µL of samples were injected into a flow of 100 mL min^{-1}. The acquisition of mass-to-charge ratio (*m/z*) data and their binning to discrete bins and peaks was conducted as previously described (Finch et al., 2022; Ferreira et al., 2023).

Good-quality metabolite data could not be produced for 15 samples (Supp. Data 2.8). A further set of ten samples (including the three Alem Tena discrepancies mentioned previously) was removed because their grain colour at one location did not match the grain colour at the other two locations (Supp. Data 2.8). *m/z* feature intensities from biologically

independent samples (brown *n* = 303 and white *n* = 389) were log10 transformed and Pareto scaled, and those differing significantly between brown and white-grained varieties were selected (Student's *t*-test, FDR < 0.05, mass tolerance of 5 ppm). Metabolite identities were assigned to the differential *m/z* features using the Mummichog algorithm (Li et al., 2013), with reference to the latest KEGG version of the *Oryza sativa japonica* (RefSeq) metabolite library (International Rice Genome Sequencing, 2005; Kanehisa et al., 2023). A mass tolerance of 5 ppm was used, and all possible adducts and isotopes were considered. Where *m/z* values could be matched to multiple metabolites, the metabolite with the smallest mass difference to the *m/z* value was selected. Statistical analysis, principal component analysis (PCA), PLS-DA, variable metabolite prediction, and volcano plot were carried out using the online R-based platform MetaboAnalyst 6.0 (https://www.metaboanalyst.ca; Pang et al., 2024).

2.5.6 – Statistical modelling for BLUP and heritability calculation

The original field trial design was updated to reflect the treatment of redundant accessions as combined redundancy groups (Supp. Data 2.7). Individual plots belonging to the same redundancy group were treated as biological replicates. Data points were removed for the metabolite analyses above. Genotypic BLUPs were calculated using the R package lme4 (v1.1.32; Bates et al., 2015) by fitting the following linear mixed model using restricted maximum likelihood (REML):

$$f(Y) = \alpha + \beta X + \gamma Z + \delta W + e$$

Where Y is the observed trait value, $f()$ is a transformation conducted for normalisation (either square root, natural log, or none), α is the global mean, β is location, X is a matrix of location effects, γ is block by location identity, Z is a matrix of block by location effects, δ is genotype identity, W is a matrix of genotype effects, and e is the residual. Location was modelled as a fixed effect due to the low number of factor levels, while block by location and genotype were modelled as random effects. The transformation $f()$ applied to each trait was selected to make residuals approximately normally distributed and independent of fitted values. Supp. Tables 2.2 and 2.5 describe the transformation applied to each trait and list any additional data points removed for specific traits prior to BLUP calculation. For example, Alem Tena data was removed prior to the calculation of BLUPs for DTH, DTM, and GFP, as this data made the traits unsuitable for linear mixed modelling even after transformation.

Modelling of BLUPs was not deemed appropriate for panicle morphology and grain colour, given their ordinal and binary encoding (respectively) and minimal variation between field sites. Instead, simple means were used as the genotypic values. To rationalise trait values for presentation, the global intercepts were added to the BLUPs for plant height, grain area, and thousand grain weight in Figure 2.8a,b. Raw BLUPs were used for kGWAS and SNP-based GWAS computations.

Broad-sense heritability (H^2) was calculated via the 'Cullis' method (Cullis et al., 2006) (quoted in the Results) and the BLUP-BLUE regression method (i.e., 'Walsh and Lynch' method; Walsh & Lynch, 2018). Calculations were performed in GenStat (*Genstat for Windows*, 2022) using the same transformations as for BLUP derivation. The selected methods are considered robust to unbalanced trial designs and produced similar results (mean difference 0.021, largest difference 0.076). Tef is highly selfing and we have demonstrated very low heterozygosity for this population (mean = 1.5%). Because of this, additive genetic variance (V_A) will predominate over dominance variance (V_D), so H^2 is expected to be approximately equal to (though slightly larger than) narrow-sense heritability (h^2).

2.5.7 – *k*-mer-based GWAS

A new *k*-mer presence/absence matrix was generated as above using the previously pooled and subsampled reads for the 150 accessions or redundancy groups. Association mapping, calculation of significance threshold, and plotting were conducted using the same kGWAS pipeline (Gaurav et al., 2022) on 141 accessions (nine redundancy groups were excluded because they contained both brown and white-grained accessions; Supp. Table 2.1). For the metabolite traits, a further three accessions were excluded because only one datapoint remained for BLUP calculation ('DZ-01-517', 'DZ-01-1376', 'Hotolla-T-135'). *k*-mers were projected onto the Dabbi reference genome and reported as the number of *k*-mers at a given association level per 10 kb genomic bin. The significance threshold for associations was calculated via Bonferroni correction as follows:

$$padj = \frac{0.05}{n} = 1.36 \times 10^{-8}$$

Where *n* is the number of *k*-mers utilised for association calculations (187,226,135) and *k* is the *k*-mer length (51). This threshold is plotted as $-\log_{10}(1.36 \times 10^{-8}) = 7.87$ on all Manhattan plots presented. For each trait, putative trait-associated regions were extended from the first bin on a chromosome containing significant *k*-mers and terminated at the point where the subsequent 500 kb contained zero significant *k*-mers. Additional putative regions were

then iteratively initiated from the next bin containing significant *k*-mers. Putative regions were then defined as significantly trait-associated if they contained ≥750 significant *k*-mers. Supp. Table 2.3 lists all significantly trait-associated regions. Dotplot sequence alignments of the LTR Copia insertions in the candidate genes in these regions were made with the dotplot function in R package SeqinR (v4.2-36; Charif & Lobry, 2007).

2.5.8 -SNP-based GWAS

SNP-based GWAS was carried out via GAPIT (v3; Wang & Zhang, 2021) with six different models: FarmCPU, BLINK, MLMM, SUPER, CMLM, and ECMLM. The significance threshold for associations was calculated via Bonferroni correction as follows:

$$padj = \frac{0.05}{n} = 1.01 \times 10^{-6}$$

Where n is the number of SNPs utilised for association calculations (49,660). The VCF file used was the same as that used to generate the minimal SNP panel, except that during linkage pruning, R^2 was set to 0.7 instead of 0.5. SNP-trait associations were considered significant when supported by at least two of the six models tested. We also considered nearby SNPs significant if they were supported by different models and separated by less than the LD decay distance (46 kb). Details of all significantly trait-associated SNPs are provided in Supp. Table 2.4.

2.5.9 – *TT2* Sequence analysis in white and brown-grained tef accessions

The homoeologous candidate genes, Et_4A_032842 and Et_4B_039404, located in the associated peaks on chr 4A and chr 4B, respectively, were identified as orthologue of *Arabidopsis TT2* gene based on sequence homology. We compared the sequences of tef *TT2* genes between the red-grained accession Dabbi and the white-grained accession, Tsedey. For this, *TT2* genomic sequences (Et_4A_032842 and Et_4B_039404) from the Dabbi reference genome (VanBuren et al., 2020) were obtained from Ensembl Plant and were used as query for a BLAST search against the draft genome assembly of Tsedey (Cannarozzi et al., 2014) available at CoGe (https://genomevolution.org/coge/). Scaffolds showing more than 90% percentage identity for each query were extracted from the Tsedey genome assembly using SAMtools faidx tool (Li et al., 2009). *TT2* sequences in the scaffolds were annotated using the gene model for Et_4A_032842 and Et_4B_039404 from the Dabbi assembly (Supp. Data 2.4 and 2.5). To ascertain if the insertions identified within the

annotated gene models were RTs, the insertion sequences were used as queries for BLAST search against a database of repeat elements from tef, available at PlantRep (Luo et al., 2022). Insertions with more than 90% identity to repeat elements in the database were considered as RTs. The identified RT insertions were manually annotated to highlight the LTR and tandem site duplications at either end of the insertions (Supp. Data 2.4 and 2.5).

3 – Semi-spatial transcriptomics reveals putative regulators of low basal spikelet productivity in wheat

3.1 – Chapter Summary

The low productivity of the basal spikelets in wheat is a developmental puzzle given that they initiate well before the more productive central spikelets. This phenomenon has also long been recognised as a potential breeding target. However, wheat displays little genetic variation for this trait when grown under high-yielding agronomic conditions. Previous work has suggested that gradients of floral and vegetative signals may compete in the base of the wheat spike, slowing the reproductive development of the basal spikelets. To identify further regulators of low basal spikelet productivity, we produced semi-spatial transcriptomic data from a developmental series of microdissected wheat spikes. Using this data, we identified transcription factors involved in spikelet development that were differentially expressed between central and basal sections of the early spike. By first developing two novel sets of regulatory elements, we then specifically overexpressed two candidate regulators of spikelet development – *SEP1-B6* and *MOF-B1* – in the base of incipient spikes. Preliminary evidence suggests that this targeted overexpression of *SEP1-B6* reduced the number of rudimentary basal spikelets, a key component of low basal spikelet productivity.

This work was partially conducted in collaboration with Dr Nikolai Adamski (JIC) and Dr Anna Backhaus (JIC). Microdissections and RNA extractions were conducted between us. Anna also contributed to trimming and pseudomapping of RNA-seq data. I was assisted in phenotyping MJ_GG18 and MJ_GG19 transgenic lines and crossing mof1 *TILLING lines by Pamela Crane (JIC). Phenotyping of* mof1 *TILLING lines was largely conducted by Alex Yu (visiting undergraduate). Many thanks to Dr Andrew Breakspear (JIC) for advice on fluorophore selection, assistance with agroinfiltration of* Nicotiana benthamiana*, and confocal imaging of positive controls. Dr Lowel O'Mallard (NBI) and Tom Betteridge (NBI) assisted my bioinformatic analyses by installing software onto the NBI computing cluster. I am grateful to Dr David Fischer (Medical University of Vienna, Austria) for advice on differential expression design matrices for DESeq2 and his own software ImpulseDE2. Plant growth was facilitated by JIC Horticultural Services and confocal microscopy by the JIC Bioimaging platform. I was particularly assisted in the latter by Dr Sergio Lopez. Novogene UK provided prompt, high-quality sequencing services.*

3.2 – Introduction

The lanceolate shape of the mature inflorescence (spike) is a defining characteristic of the *Triticum* genus in both wild and domesticated species. This shape arises both because the main growth axis is determinate – being capped with a terminal spikelet – and because the central spikelets are larger than those at the apex and base of the spike, which produce fewer and smaller grains (Bonnett, 1936). In fact, in typical field conditions, the basal-most two or more spikelets are completely infertile and are termed rudimentary basal spikelets (RBS) (Tamagno et al., 2024). This in turn is accounted for by the slower development of the spikelet and floret meristems at these positions prior to anthesis, one consequence of which is increased rates of floret abortion.

Floret abortion occurs during a 'critical period' between 20 and 10 days pre-anthesis and recent work suggests the decision to abort is based on whether florets successfully pass a minimum stage of development (~ Waddington (W) stage 5.5;(Waddington et al., 1983) prior to this checkpoint (Backhaus et al., 2023). This process increases uniformity in the developmental ages of the surviving florets and may be one of the mechanisms governing the relatively synchronous anthesis across the wheat spike. RBS are hypothesised to be a consequence of zero florets in a spikelet surviving this checkpoint (Backhaus et al., 2023).

The delayed development of the apical spikelets is a simple consequence of their later initiation compared to the central and basal spikelets. Similarly, more distal florets within a given spikelet initiate later than those proximal to the rachis, leading to delayed development and lower productivity. In contrast, the evolutionary and molecular reasons for the lag in development of the basal-most spikelets – which are initiated first – remain open questions. Nonetheless, recent experiments and reanalyses have increased our understanding of certain aspects, which we explore below.

An early concept in the study of low basal spikelet productivity was that these spikelets had the poorest or "lowest priority" (Stockman et al., 1983) access to photoassimilates. This was partly based on the observation that the basal-most 3-4 spikelets are only supplied by branches of the central vascular bundles in the rachis, while entire large bundles are diverted into the central spikelets (apical spikelets are supplied by whole, but small, bundles ;Whingwiri et al., 1981).

This notion has been challenged by recent experiments showing that, in field conditions at ~10 days pre-anthesis, sucrose, glucose, and fructose concentrations generally do not differ significantly between the basal and central spikelets (Backhaus et al., 2023). These

results agree with reanalysis of earlier work (Stockman et al., 1983) and remained consistent when total available photoassimilate levels were decreased by applying 12-13 day shading treatments prior to sample collection. In the minority of cases where significant spatial differences were observed for particular combinations of site, year, shading, tissue, and sugar, the basal sections had higher sugar concentrations.

These experiments suggest that the supply of photoassimilates to basal spikelets is not limited by their different vascular morphology, though there are caveats to this conclusion. Firstly, this data was acquired towards the end of the 'critical period' for abortion, so does not exclude the possibility that the basal spikelets have lower sugar concentrations in the developmental stages leading up to this checkpoint. Secondly, static snapshots of sugar concentrations are not indicative of photoassimilate flux; the basal spikelets may equilibrate at similar sugar concentrations as the central spikelets, but actually be receiving and incorporating less carbon per unit time. These reservations mean we cannot yet rule out vascular limitations on basal spikelet development, but such limitations are likely to be proximal rather than ultimate given that the lanceolate shape of the wheat spike is established by the glume primordia stage (W3), yet the vascular system has only just begun to differentiate by this point (Pizzolato, 1997, 1998). Future experiments could attempt to investigate spatial carbohydrate concentrations at the very early stages of spike development to elucidate if photoassimilate availability plays a role in establishing initial spikelet asymmetry.

Experiments manipulating the resources available to the developing wheat spike, such as the above shading treatments, have also helped to reinforce the concept that while the relatively low productivity of the basal spikelets are constitutive, the number of RBS is a facultative trait governed by the environment. (Backhaus et al., 2023) found that shading treatments applied 20 to 13 days pre-anthesis increased RBS number across multiple cultivars in three field trials. This appears to be a consequence of a spike-wide reduction in the number of florets surviving abortion due to shading, with a higher number of basal spikelets now maintaining zero fertile spikelets.

(Tamagno et al., 2024) also found that RBS number could be manipulated, in this case by growing wheat at dramatically lower densities (15 or 33 plants m^{-2}) than is used for high-yielding cropping (controls were grown at 250 or 480 plants m^{-2}). While not measured, this was presumed to increase assimilate supply per spike, even given much increased tillering rates. This treatment drastically decreased RBS number versus the controls in both the main and tiller spikes across 14 accessions. Five accessions even set an average of two grains in the basal-most spikelet of their main spike. In contrast, the authors note that the

typical productivity gradients of the wheat spike were maintained in the low-density treatment, i.e. grain set also increased for central and apical spikelets, supporting the constitutive nature of this trait.

The above study utilised accessions released between 1940 and 2021 and found that under control, high-density planting, there was no significant relationship between year of release and the grain number of the basal-most four spikelets. This is consistent with a study comparing a germplasm collection (n = 180) with modern European wheat varieties (n = 210), which showed that basal spikelet grain set is only marginally higher in modern varieties. Interestingly, the two populations have almost identical spatial profiles of relative yield contribution per spikelet, again highlighting the largely fixed nature of the lanceolate spike. Thus, historically, improvements in basal spikelet productivity only seem to have occurred as a by-product of spike-wide productivity gains.

Insights into the molecular mechanism underlying this constitutive trait have recently started to emerge. The *P1* locus underlying the high RBS, long glume, and long grain phenotypes of *Triticum turgidum* ssp. *polonicum* has been mapped to *VEGETATIVE TO REPRODUCTIVE TRANSITION 2* (*VRT-A2*; TraesCS7A02G175200), a member of the *SHORT VEGETATIVE PHASE* (*SVP*) subfamily of MADS-box transcription factors (TFs) (Adamski et al., 2021; J. Liu et al., 2021). Development of *Triticum aestivum* (cv. Paragon) near-isogenic lines (NILs) showed that the $P1^{POL}$ (*VRT-A2b*) allele indeed confers higher RBS number and longer floral organs versus $P1^{WT}$ (*VRT-A2a*) (Adamski et al., 2021; Backhaus et al., 2022). The same is true of transgenic lines (cv. 'Fielder') expressing exogenous copies of *VRT-A2b* (Adamski et al., 2021; Backhaus et al., 2022).

VRT-A2 therefore appears to be an important regulator of basal spikelet productivity. At the early double ridge (EDR; W2), late double ridge (LDR; W2.5), and glume primordia (GP; W3) stages, these genes display an acropetally weakening expression pattern, with strong expression at leaf primordia, but weakening through the incipient peduncle, basal spike, and onwards to the spike apex (Li et al., 2021; Backhaus et al., 2022). They have also been associated with repression of flowering and shown to be downregulated during the floral transition in Arabidopsis (Gregis et al., 2013), wheat (Li et al., 2021), other cereals (Sentoku et al., 2005; Trevaskis et al., 2007). This led (Backhaus et al., 2022) to propose a model whereby opposing gradients of flowering repressive *SVP* genes and flowering promoting *SEPALLATA* (*SEP*) genes lead to a weaker floral transition signal in the basal spikelets. This in turn delays their development in comparison with more central spikelets which are exposed to lower levels of *SVP* transcription. This is consistent with yeast two-hybrid data

showing indirect interaction between *SVP* and *SEP* proteins via competitive binding of a common partner TF (Li et al., 2021).

This effect may be a result of *VRT2* and its paralog *SVP1*'s role in repressing aerial branching. Single mutants in *vrt2* and *svp1,* and *vrt2 svp1* double mutants, show dramatically increased rates of axillary spikelets or whole spikes at sub-peduncle nodes (Li et al., 2021). The expression pattern of *SVP* genes described above correlates well with the repression of axillary meristems (AxM) in wild-type wheat; AxM at sub-peduncle aerial nodes are fully repressed, basal spikelet meristems are moderately repressed – leading to lower productivity – while central and apical spikelet meristems are not repressed and produce highly productive spikelets. This suggests that RBS are a byproduct of AxM repression caused by imprecise gene expression boundaries.

If we accept the hypothesis that source strength is adequate post-anthesis, there is face validity to the idea that yield could be boosted by optimising the allocation of limited *pre*-anthesis sink strength to raise the number of florets surviving the abortion checkpoint during the critical phase and thereby increase post-anthesis sink strength. The current literature suggests that the low productivity of basal spikelets may be interpreted as an architectural inefficiency created by overlapping floral and vegetative signalling gradients. Given this, upregulating their productivity has attracted increasing attention in recent years as a potential mechanism to achieve greater sink strength without significant pre-anthesis resource investment. This is crucial to avoid trade-offs such as increased abortion in other spikelets or a reduction in fertile tiller number.

How might this be achieved? As discussed, relatively little natural variation for basal spikelet productivity has been identified within modern wheat, meaning researchers must rely on induced variation. In a diploid species, this could be achieved by screening a mutagenised population for basal spikelet phenotypes. However, this approach is generally not effective in polyploid wheat due to the genetic redundancy provided by the (typically) four or six copies of each gene in tetraploid or hexaploid wheat, respectively. To our knowledge, this has not been conducted in a diploid wheat wild relative either.

Manipulating *SVP* expression also presents difficulties as these genes are highly pleiotropic, with crucial roles throughout the wheat lifecycle. *svp1 vrt2* double mutants produce weakly fertile – and therefore wasteful – aerial branches as previously discussed, but also exhibit delayed heading time and increased spikelets per spike. These arise, respectively, from *SVP* genes' roles in accelerating the vegetative to reproductive transition in the apical meristem and in converting the inflorescence meristem into a terminal spikelet (Li et al., 2021). Lastly,

downregulation of *VRT2* expression can reduce grain length and weight, as shown by recent work which both overexpressed a *VRT2* repressor and knocked-out an activator, leading to suppressed expression of all three *VRT2* homoeologs.

Another approach could be to identify regulators of spikelet development through comparative transcriptomics and then to tweak their expression to favour basal spikelet productivity. Differential gene expression analyses have been a mainstay for identifying plant stress response genes for many years and have uncovered genes up- or down-regulated following exposure to heat (Kino et al., 2020), cold (Y. Liu et al., 2022), drought (Fracasso et al., 2016), waterlogging (Arora et al., 2017), nutrient deficiency (Wang et al., 2019), pests (Divya et al., 2021), and disease (Matic et al., 2016). Such case-control pairs are not always apparent for developmental studies, but can be produced by employing developmental mutants where available (Marks et al., 2009; Z. Y. Li et al., 2015; Bräuning et al., 2018). In the present scenario, comparing RNA-seq datasets derived from central and basal tissues of incipient wheat spikes could be used to identify regulators of spikelet development that are differentially expressed (DE) in these two regions.

A previous study analysed the transcriptomes of single micro-dissected wheat spikes (cv. Paragon; hexaploid spring wheat) using low-input RNA-seq (Backhaus et al., 2022). They analysed basal, central, and apical spike sections at both LDR (W2.5) and GP (W3) stages. These stages were selected because they are, respectively, the last developmental stage without, and the first with, a lanceolate spike. Amongst other findings, this data contributed to the establishment of the *SVP-SEP* model described above.

Here, we hypothesised that we could identify additional genes mediating low basal spikelet productivity through the use of semi-spatial transcriptomics on a time course of central and basal wheat spike sections. We further hypothesised that we could specifically raise basal spikelet productivity by transgenically misexpressing such genes to reduce spatial differences in their expression. To achieve this, we first extended on the work of (Backhaus et al., 2022) with the aim of identifying novel regulators of wheat spikelet productivity. We produced RNA-seq time courses for basal and central sections of the spike across a broader range of wheat developmental stages, spanning from EDR (W2) to carpel extension (CE; W5). We utilised pools of spike tissue rather than single spike sections in order to improve upon the high variability between replicates observed in the previous study. By using near-isogenic lines (NILs) with either the $P1^{WT}$ or $P1^{POL}$ allele of *VRT-A2*, we were also able to identify potential downstream targets of this pleiotropic TF. Using our RNA-seq data, we selected two potential regulators of spikelet productivity – *MORE FLORET 1* and *SEP1-6* – and developed two candidate regulatory environments to achieve specific overexpression

of these genes in the basal portion of the spike. We tested these regulatory environments using multiple reporter systems and then phenotyped transgenic lines overexpressing our candidate developmental genes.

3.3 – Results

3.3.1 – Pooling multiple plants decreases the variability between replicates of wheat spike semi-spatial transcriptomics

We generated expression data for pools of central and basal sections of the wheat spike at five early stages of development (W2, W2.5, W3.25, W4, and W5; Figure 3.1A, adapted from (Kirby & Appleyard, 1984) with two or more biological replicates per stage-section combination (Table 3.1). This was conducted for two NILs (cv. Paragon) containing different *VRT-A2* (TraesCS7A02G175200) alleles. The $P1^{WT}$ NIL contains the wild-type *VRT-A2a* allele, while the $P1^{POL}$ NIL contains the *VRT-A2b* allele introgressed from *Triticum turgidum* ssp. *polonicum* (Adamski et al., 2021) discussed earlier.

We calculated read counts and transcripts per million (TPM) values for all genes in the IWGSC RefSeq v1.1 annotation (Appels et al., 2018) using the Kallisto pseudoaligner (Bray et al., 2016). All subsequent analyses were conducted for the 107,892 high confidence (HC) gene models only. Principal component analysis (PCA) showed that samples clustered by both stage (PC1; 18.0% variance) and section (PC2; 10.5% variance; Figure 3.1B). However, there was no strong effect of genotype in any major PCs (those explaining >5% of variance; Figure 3.1C,D). Genotypes were only well separated by PC10 which accounted for just 1.6% of variance (visualised up to PC15). The separation of spatial sections by PC2 decreased across development, suggesting that the transcriptional profiles operating in these sections are more distinct earlier in development. One replicate for the basal section of $P1^{POL}$ at W5 was removed as a suspected outlier (red circle in Figure 3.1B-D).

For the $P1^{WT}$ samples, individual samples showed expression (> 0.5 TPM) of 49,387 genes on average (SD = 1,049; Supp. Data 3.1), while 55,346 unique genes were expressed (>0.5 TPM average across replicates) across all samples. $P1^{POL}$ samples produced similar results (Supp. Data 3.1).

In their earlier study (Backhaus et al., 2022) identified a high level of variability between biological replicates (n = 2-4). They calculated a median coefficient of variation (CV) across gene-stage-section combinations of 0.39. They also calculated CV values for comparable whole-spike transcriptomic time courses which used pools of spikes. (Y. P. Li et al., 2018) pooled 100-200 spikes of winter wheat (cv. 'Kenong 9204') per biological replicate, producing a median CV of 0.14 (n = 2 replicates per sample category). (Feng et al., 2017) pooled 10-50 spikes (cv. Chinese Spring) per sample, yielding a median CV of 0.21 (n = 2).

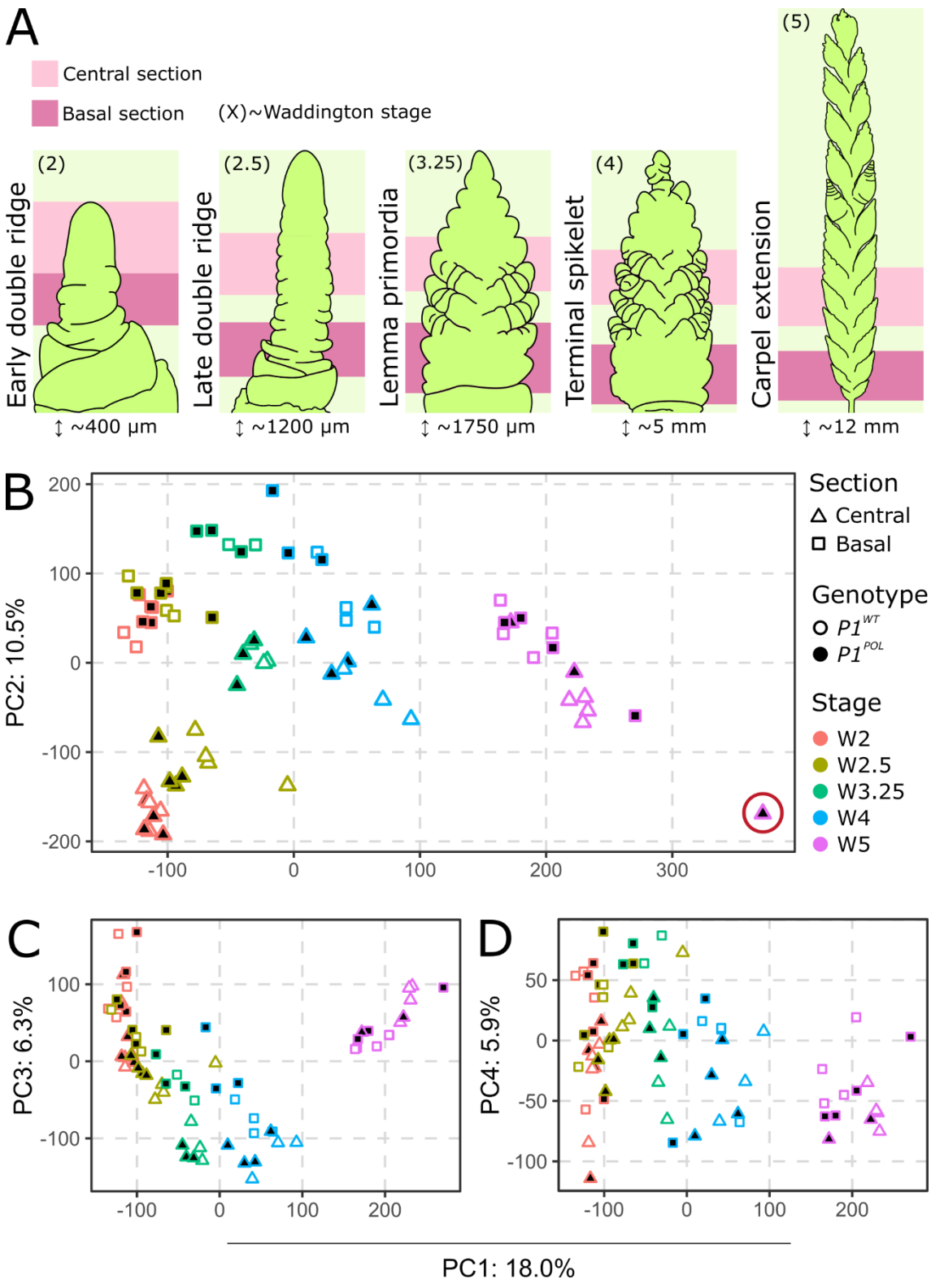


Figure 3.1 – RNA-seq samples cluster by spatial section and developmental stage
A, Diagram of spike stages and spatial sections dissected for RNA-seq. Illustrations for W2-W4 adapted from (Kirby & Appleyard, 1984). Illustrations are not to scale – approximate spike lengths are provided below illustrations. **B-D,** Scatterplots of samples by principal components 1 and 2 (**B**), 3 (**C**), or 4 (**D**). PCA was conducted on HC genes with non-zero variance. Subsequent PCs each accounted for less than 5% of total variation. Shape denotes section, fill colour denotes genotype, and outline colour denotes developmental stage. One outlier was removed prior to downstream analyses (red ring in **B**; removed from **C** and **D** to aid visualisation).

Table 3.1 – Samples obtained for wheat spike semi-spatial transcriptomics time course
Developmental stages are described by their names and Waddington (W) stage number. Dissection methodology detailed in Methods.

Genotype	Section	Developmental stage	Biological replicates
P1^{WT} (*VRT-A2a*)	Central	Early Double Ridge (W2)	4
		Late Double Ridge (W2.5)	4
		Lemma Primordia (W3.25)	3
		Terminal Spikelet (W4)	3
		Carpel Extension (W5)	4
	Basal	Early Double Ridge (W2)	4
		Late Double Ridge (W2.5)	4
		Lemma Primordia (W3.25)	2
		Terminal Spikelet (W4)	4
		Carpel Extension (W5)	4
P1^{POL} (*VRT-A2b*)	Central	Early Double Ridge (W2)	4
		Late Double Ridge (W2.5)	4
		Lemma Primordia (W3.25)	3
		Terminal Spikelet (W4)	4
		Carpel Extension (W5)	2*
	Basal	Early Double Ridge (W2)	4
		Late Double Ridge (W2.5)	4
		Lemma Primordia (W3.25)	3
		Terminal Spikelet (W4)	3
		Carpel Extension (W5)	4

**Remaining after outlier discussed in Figure 1.1 removed*

Backhaus et al. attributed the lower CV values in these studies to their use of pools of spikes, which would dampen variation between replicates compared with their own strategy of sequencing sections of single spikes.

Supporting this, the CV values calculated for each gene-stage-section-genotype combination in our data were also lower. After filtering out LC genes and weakly expressed genes (<0.5 TPM average across replicates) the median CV value across our data was 0.12. (N.B. it is unclear if similar filtering steps were applied prior to the calculations in Backhaus et al., 2022). The lower variability between biological replicates in our data is compatible with detection of smaller differential gene expression effect sizes.

3.3.2 – Central and basal spike transcriptomes are highly different at early spike development stages

Of the genes expressed (> 0.5 TPM) in at least one section-stage combination in $P1^{WT}$, 14,196 exhibited different expression patterns across development between the central and basal sections ($p < 0.001$, Benjamini-Hochberg corrected). When the minimum expression threshold was raised to > 5 TPM the number of DE genes remaining fell moderately to 10,534, suggesting most DE genes are relatively highly expressed in at least one section-stage combination.

Given that the development of the basal spikelets visibly lags versus the central spikelets by the GP (W3) stage, we wished to investigate which genes were DE between central and basal spikelets prior to this in order to identify genes driving the divergence in morphology. At EDR (W2), 4,106 genes were more highly expressed in the central section versus 3,199 which were more weakly expressed (padj < 0.001, IHW corrected; Figure 3.2). At LDR (W2.5), 4,160 genes were more highly expressed in the central section versus 2,250 which were more weakly expressed (padj < 0.001, IHW corrected; Figure 3.2). 3,869 of these DE genes were shared by both stages (53% of EDR DE genes, 60% of LDR DE genes).

In contrast, far fewer genes were DE between stages. For the central section, 1,160 genes were upregulated from EDR to LDR, while 470 were downregulated. For the basal section, 483 genes were upregulated from EDR to LDR, while 449 were downregulated. This suggests that there are greater differences in the spike development programmes between central and basal sections at EDR and LDR than there are between those stages. This result agrees with the work of (Backhaus et al., 2022), who also found greater numbers of DE genes between spike sections (basal, central, and apical) than between stages (LDR and GP).

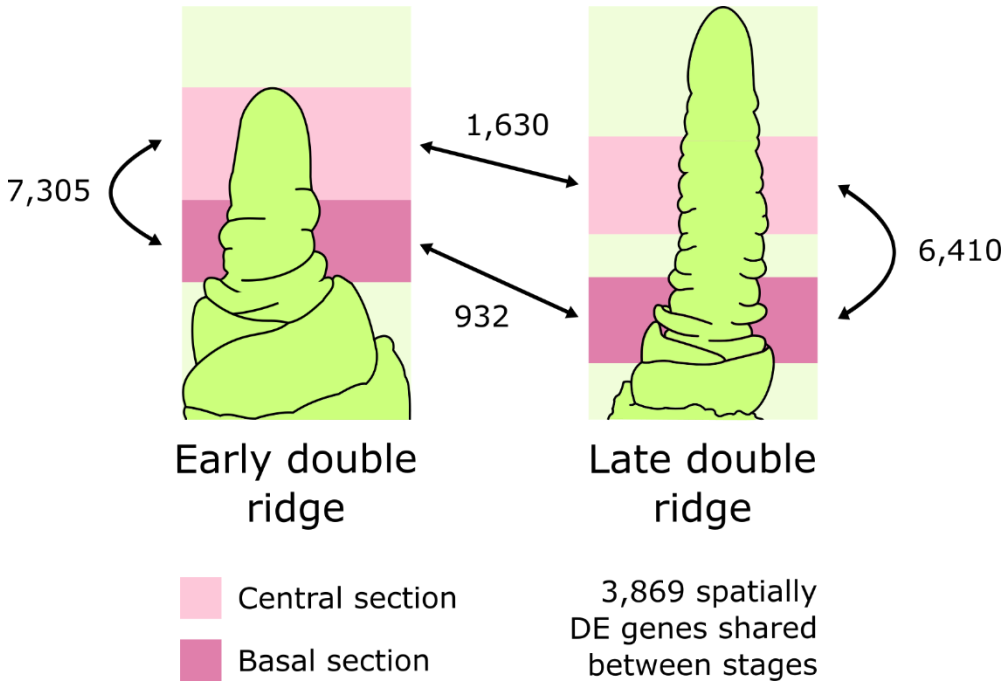


Figure 3.2 - High numbers of spatially differentially expressed genes were identified at early double ridge (W2) and late double ridge stages (W2.5) in P1^{WT} Paragon NILs
Arrows indicate number of differentially expressed genes (padj < 0.001, IHW corrected) between indicated tissues sections or genotypes. Illustrations adapted from (Kirby & Appleyard, 1984).

3.3.3 – Regulators of spikelet meristem determinacy *MOF1* and *SEP1-6* are more weakly expressed in basal spike sections

To further dissect which genes could be involved in regulating basal spikelet productivity, we focused on the 7,305 genes showing spatially different expression at EDR. As discussed previously, *VRT2* plays a crucial role in determining basal spikelet productivity, but its manipulation can induce adverse pleiotropic effects. To discover independent factors regulating basal spikelet productivity – including those that might decouple basal spikelet productivity and aerial branching – we wished to exclude genes downstream of *VRT2* from consideration. To identify potential downstream targets, we compared the developmental expression profiles (W2 to W5) of all HC genes between the $P1^{WT}$ and $P1^{POL}$ NILs using the R package ImpulseDE2 (Fischer et al., 2018). 969 genes were DE in one or both of the central and basal developmental series (padj < 0.001, Benjamini-Hochberg corrected). The relatively small number of DE genes identified agrees with the co-clustering of $P1^{WT}$ and $P1^{POL}$ samples by PCA. These genes are not all anticipated to be direct targets of *VRT-A2*; their transcript abundance may be perturbed by transcriptional, anatomical, or physiological changes more proximal to *VRT-A2*. We decided to remove these genes from further consideration to ensure we did not misdirect our efforts into 'rediscovering' the transcriptional cascade controlled by *VRT-A2*. While this was a conservative approach and may have discarded some important genes, we feel it was justified given the wealth of candidates available to explore in the remaining gene list.

Excluding these genes from the EDR set of spatially DE genes resulted in a total of 7,068 (not all of the 969 genes DE between $P1^{WT}$ and $P1^{POL}$ were present in the set of 7,305 spatially DE genes). This set was filtered to retain only TFs based on a previously published dataset (Evans et al., 2022), leaving 547 genes for further investigation. Rice orthologues were identified for 525 genes and, of these, 81 (orthologous to 45 rice genes) were linked to inflorescence-related annotations on the database FunRiceGenes (Huang et al., 2022).

These genes were then filtered for expression patterns of interest. A key feature of interest was lower basal versus central expression at EDR and LDR. Such genes could be positive regulators of spikelet development, with their weaker expression in basal spikelets causing a slower spikelet meristem to floret meristem transition. This would be more amenable to transgenic manipulation because introducing elevated or ectopic expression would be a dominant, gain-of-function trait. In contrast, manipulating genes which negatively regulate spikelet development (i.e. are more highly expressed in basal spikelets), such as *VRT2*, could require knock-outs or knock-downs across multiple redundant gene copies for large

effect sizes, delaying germplasm development. To this effect, we filtered for gene triads in which all three homoeologs showed non-zero mean expression in the central section at EDR and LDR, then for triads in which all three homoeologs showed lower mean basal versus central expression at those time points.

We found that all three homoeologs of *MORE FLORET 1 / MULTI-FLORET SPIKELET 2* (*MOF1/MFS2;* henceforth *MOF1*) showed weaker or temporally lagging patterns of gene expression in the basal section of the spike (Figure 3.3). *MOF1* is a MYB TF that was characterised simultaneously by two groups as a regulator of the spikelet meristem (SM) to floral meristem (FM) transition, via ensuring spikelet meristem determinacy (Li et al., 2020; Ren et al., 2020).

It was also shown that INDETERMINATE SPIKELET1 (OsIDS1) and *SUPERNUMERARY BRACT* (*SNB*), two key *APETALA2-like* TFs involved in the SM-to-FM transition, were downregulated in the rice *mof1* mutant (Li et al., 2020). The *mof1* mutant produced variable phenotypes, including some plants with multiple sets of floral organs per spikelet. This contrasts with WT rice which produces only one floret per spikelet. While wheat spikelets are already indeterminate, we hypothesised that the role for *MOF1* in accelerating the SM-to-FM transition may be conserved.

Another group of interest was the *LOFSEP* clade of *SEP* MADS-box TFs. Five *LOFSEP* genes showed weaker basal expression at EDR and LDR; *SEP1-D4*, *SEP1-D5*, and all three homoeologs of *SEP1-6*. While the *SEP* genes are involved in specifying floral organ identity in cereals via an E-class homeotic function, *LOFSEP* genes additionally act to specify the SM and production of its non-floral organs; the glumes and lemmas (Cui et al., 2010; Gao et al., 2010). For example, *SEP1-4/*5 (*OsMADS5*) and *SEP1-6* (*OsMADS34/PAP2*) cooperate to limit inflorescence branching and thus accelerate the maturation of spikelet meristems (Zhu et al., 2022). Interestingly, the *SEP1-6* genes display a gene expression trajectory that decreases from the late double ridge stage onwards for the central spike section, but for the basal section starts lower and increases until around the terminal spikelet (TS; W4) stage (Figure 3.3). *SEP1-6* is also the earliest *SEP* to be transcribed in rice (Kobayashi et al., 2010). This may suggest a more focussed role for *SEP1-6* in SM specification and early development over floral progression.

SEP1-6 has previously been constitutively overexpressed in hexaploid wheat (cv. ‘Kenong 199’; Y. G. Wang et al., 2017). Across 45 independent lines producing various levels of *SEP1-6* expression, expression was negatively correlated with the number of spikelets per spike.

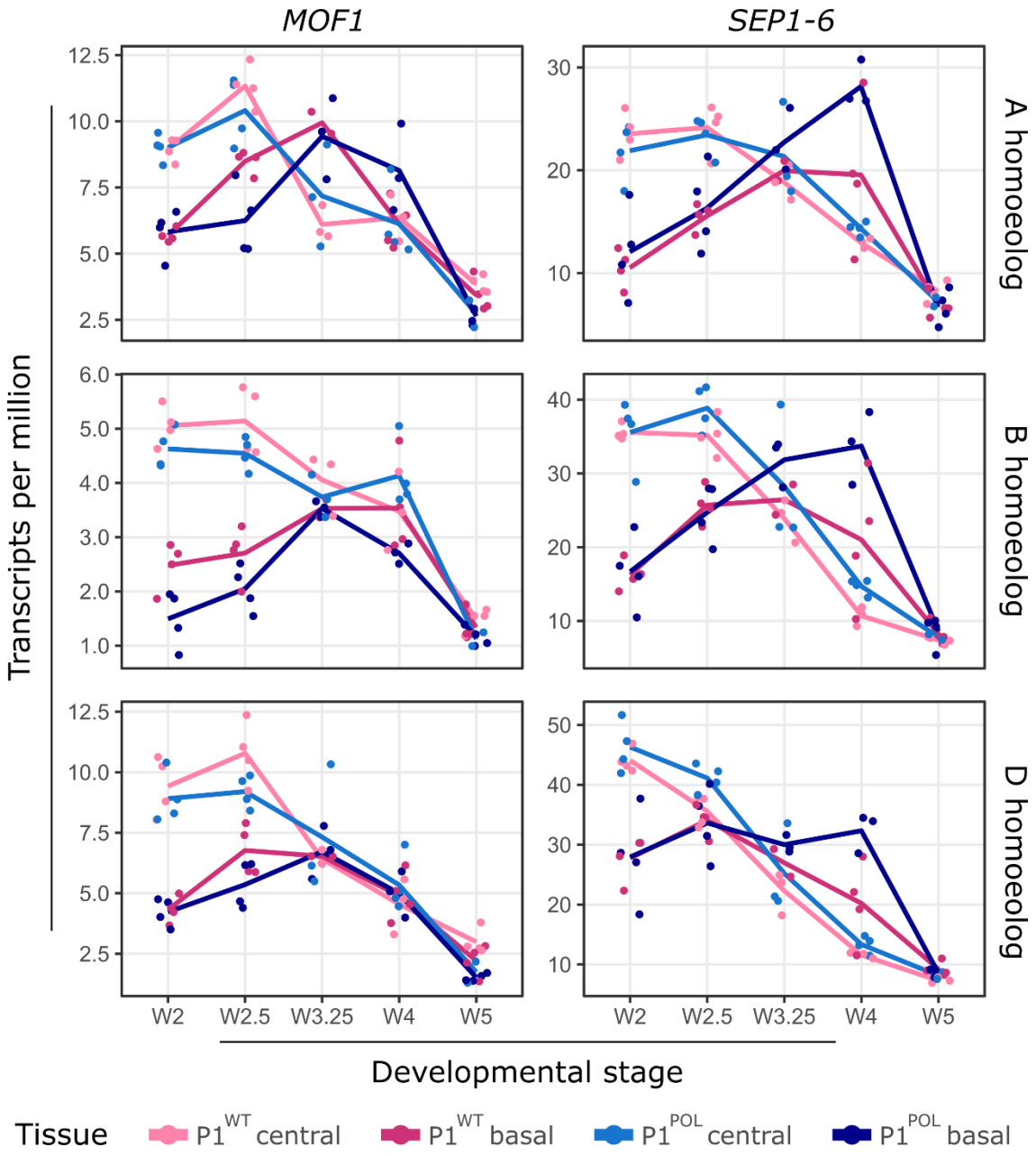


Figure 3.3 – All three homoeologs of *MOF-1* and *SEP1-6* are more strongly expressed in basal spike sections at early spike development stages
TPM values for individual samples (points) and stage-section-genotype means (lines) for *MOF-1* (left) and *SEP1-6* (right). First, second, and third rows depict A, B, and D subgenome homoeologs, respectively.

This appeared to be due to reductions in the durations of the double ridge (W2-2.5), floret primordia (W3.5), and terminal spikelet (W4) stages. This acceleration of inflorescence development would allow fewer spikelet meristems to form before the terminal spikelet is initiated and is consistent with the branch suppression role proposed in rice. Surprisingly, florets per spikelet were also reduced. This may be due to the ectopic expression of the gene by the constitutive construct, as in our RNA-seq time courses we have observed a decline in *SEP1-6* as spikelet development progresses. We therefore presume that strategies to manipulate *SEP1-6* should be careful to avoid elevated expression beyond roughly the floret primordia stage to avoid this effect.

Three other triads also matched our filtering criteria, but these either had less consistent trends between homoeologs (the triad orthologous to OsCUC1 (Os06g0344900)) or showed high TPM variability between experimental replicates (triads orthologous to OsbHLH98 (Os03g0797600) and OsGRAS-32/DLT (Os06g0127800)) so were less convincing candidates.

We hypothesised that raising the expression of *MOF1* or *SEP1-6* specifically in the basal section of the spike at the EDR and LDR stages could accelerate the SM-to-FM transition in the basal spikelets – and thereby boost their productivity – without additional deleterious effects on spike architecture.

3.3.4 – RNA-seq and ATAC-seq data can be used to inform the design of synthetic regulatory environments

Transgenic expression experiments in cereals generally use constitutive promoters that result in transcription of the target gene in all tissues and at all developmental stages. To achieve expression of transgenes that was spatially and temporally restricted to the basal spike at the EDR and LDR stages, we needed to test novel combinations of regulatory elements. We decided to examine our semi-spatial transcriptomic data to identify genes that naturally exhibit the intended expression pattern, with the aim of utilising their regulatory elements for targeted misexpression transgenics.

We filtered the set of genes with spatially differential expression at the EDR stage to retain only those with at least moderate expression (>1 TPM) at EDR and LDR and 3-fold higher mean expression in the basal section at EDR and LDR, leaving 586 genes. We then either increased the minimum basal versus central differential to 5-fold (leaving 196 genes) or introduced a requirement for low maximum expression (<5 TPM) in non-spike tissue categories (cv. 'Azhurnaya' mapped to IWGSC RefSeq v1.1, processed data from (Borrill et

al., 2016), leaving 108 genes. We then examined the gene expression profiles of these genes to identify those with the aforementioned characteristics.

Two candidate homoeolog groups were identified in which all three homoeologs showed the intended expression patterns. The first was orthologous to the barley *MANY-NODED DWARF 1* gene (*MND1* (HORVU.MOREX.r3.7HG0742750); also rice *OsgIHAT1/GW6a* (Os06g0650300)) The homoeologs in wheat are TraesCS7A02G506400, TraesCS7B02G413900, and TraesCS7D02G494500. The second had not been functionally characterised or named (TraesCS2A02G382500, TraesCS2B02G399800, and TraesCS2D02G378900). In both homoeolog groups, gene expression was close to zero in the central spike section throughout development, while in the basal section expression was highest at the EDR stage and then declined steadily (Figure 3.4).

Using previously published bread wheat ATAC-seq data (cv. Kenong 9204; discussed in Chapter 4; Lin et al., 2024), we then examined how chromatin accessibility around these genes varied across development with the aim of identifying regulatory regions that could be correlated with the desired expression pattern.

We hypothesised that enhancer regions contributing to the desired expression pattern would have high ATAC-seq signal at the EDR and LDR stages, with declining accessibility at later stages. In contrast, silencer regions would have low accessibility at EDR and LDR – permitting the high expression of the target genes observed at these stages – and then increasing accessibility later in spike development. We also examined these genes for unmethylated DNA (see Chapter 4 for dataset generation), as this has been shown to stably capture the superset of accessible chromatin regions (ACRs) which become accessible across different tissues and developmental stages (Crisp et al., 2019; Crisp et al., 2020). Unmethylated regions (UMRs) were therefore used as corroborating evidence for putative regulatory regions. Typically, the first 2 kb upstream of the transcription start site is used by default for initial promoter development. We hoped that by selecting regions more strategically we could more accurately recapitulate our target genes' expression patterns.

Ultimately, we selected the untranslated regions (5' and 3' UTRs) and additional non-coding regions adjacent to the two B genome homoeologs to drive expression of our target genes as these had the clearest sets of ATAC-seq peaks. The regions selected each partially overlapped with at least one UMR. For *MND-B1*, we selected an upstream region of 1,446 bp (7B:681857270-681858715) and a downstream region of 1,163 bp (7B:681854338-681855500; Figure 3.5).

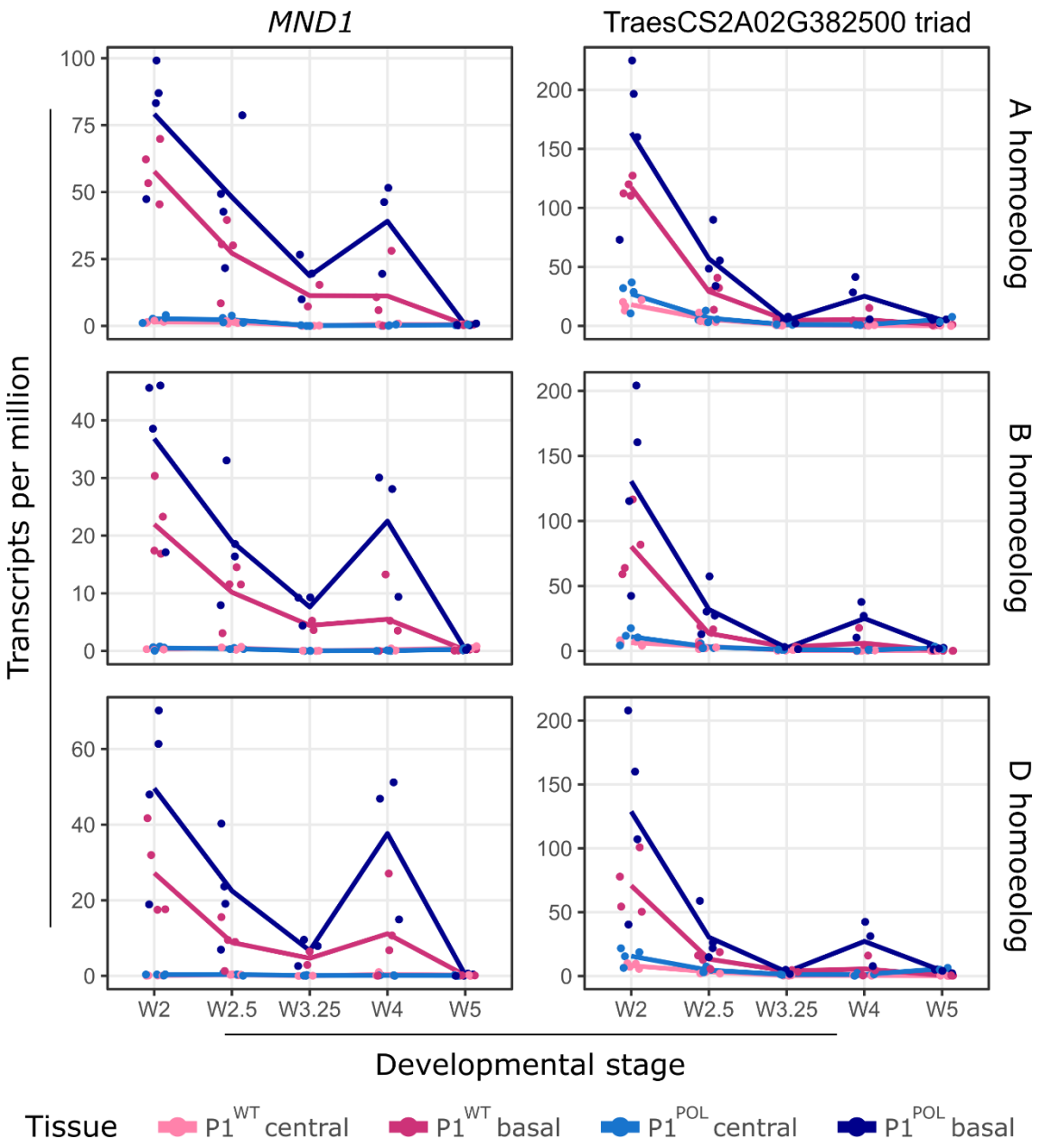


Figure 3.4 – All three homoeologs of *MND1* and an unnamed triad are more strongly expressed in basal spike sections at early spike development stages
TPM values for individual samples (points) and stage-section-genotype means (lines) for *MND-1* (left) and the uncharacterised triad of TraesCS2A02G382500, TraesCS2B02G399800, and TraesCS2D02G378900 (right). First, second, and third rows depict A, B, and D subgenome homoeologs, respectively.

For TraesCS2B02G399800 we selected two upstream regions of 2,734 bp (2B:568268278-568271011) and 2,002 bp (2B:568273035-568275036), and a downstream region of 1,782 bp (2B: 568275643-568277424), (Figure 3.6). The sequences extracted from the 5' UTR, basal promoter, and upstream region of *MND-B1* were synthesised as a Golden Gate L0 'Pro + 5U' part and denoted Basal Spike Specific 1 P5U (BSS1-5PU; Figure 3.5). Similarly, the sequences from the 3' UTR and downstream region of *MND-B1* were synthesised into a L0 '3U*' part (BSS1-3D). The equivalent parts for TraesCS2B02G399800 were denoted BSS2-5PU and BSS2-3D (Figure 3.6). In each case, the sequences were domesticated to remove BbsI and BsaI restriction sites.

In order to test the selected regulatory elements, four L2 Golden Gate constructs were assembled, with each set of regulatory elements driving either a fluorescent tdTomato or enzymatic β-glucuronidase (GUS) reporter gene (Table 3.2). tdTomato was selected because, in initial screens, wheat spike tissue did not autofluoresce under the laser used for tdTomato activation (561 nm). Additionally, tdTomato is a tandem dimer, meaning a single translation event results in two fluorescently responsive active sites, which we hoped would improve our ability to detect fluorescent signals.

The tdTomato sequence we used did not contain introns and was fused to a nuclear localisation signal (NLS), with the aim of concentrating recombinant protein to boost any fluorescent signal. GUS was selected as a back-up reporter because, as an enzyme, a single translation event can lead to many catalysis events, producing many blue dye molecules from the colourless 5-bromo-4-chloro-3-indolyl glucuronide (X-Gluc) substrate. Again, we hoped this would boost weaker signal to detectable levels. The GUS sequence we used contained introns to boost mRNA accumulation (Gallegos & Rose, 2015) and was not fused to a localisation sequence. An additional NOS terminator was included in each construct in case the native terminators in BSS1-3D or BSS2-3D were defective. These constructs were transformed into bread wheat (cv. Fielder) and assessed for copy number (CN) at the T0 stage.

To test the effect of misexpressing *MOF-1* and *SEP1-6*, a further four constructs were assembled, with each set of regulatory elements driving expression of either *MOF-B1* (TraesCS2B02G420900) or *SEP1-B6* (TraesCS5B02G396700) pre-mRNA (introns retained, domesticated to remove BbsI and BsaI restriction sites), each with a C-terminal 3x FLAG tag (Table 3.2). The numbers of independent transformation events and resultant plants are detailed in Table 3.2.

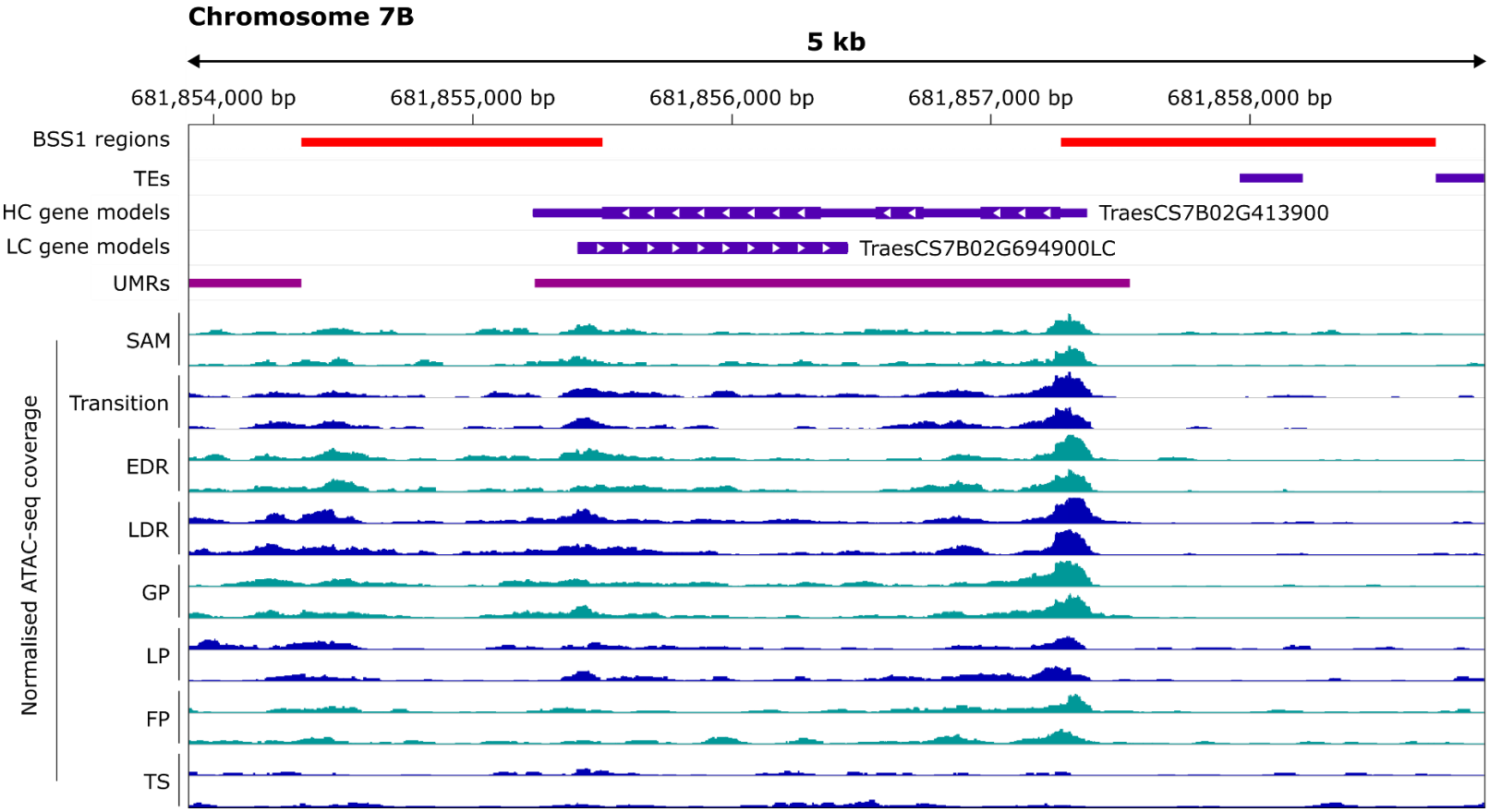


Figure 3.5 – Non-coding regions about *MND-B1* (TraesCS7B02G413900) selected for basal spike-specific regulatory environment 1 (BSS1)
Annotated 5 kb region of chromosome 7B (IWGSC RefSeq v1.0 assembly). Purple tracks indicate transposable elements (TEs) plus high confidence (HC) and low confidence (LC) gene models (RefSeq v1.1 annotation). Magenta track indicates stably unmethylated regions (UMRs). Teal and blue tracks (alternated for clarity) indicate ATAC-seq coverage for spike tissue chromatin at various developmental stages (two replicates per stage). ATAC-seq coverage was normalised against sample read number and tracks were standardised to a maximum normalised coverage of 50 to aid visual comparison of lanes. The red track indicates the final regions selected to become BSS1 regulatory parts. Data visualised using the Integrative Genomics Viewer (IGV; Robinson et al., 2011).

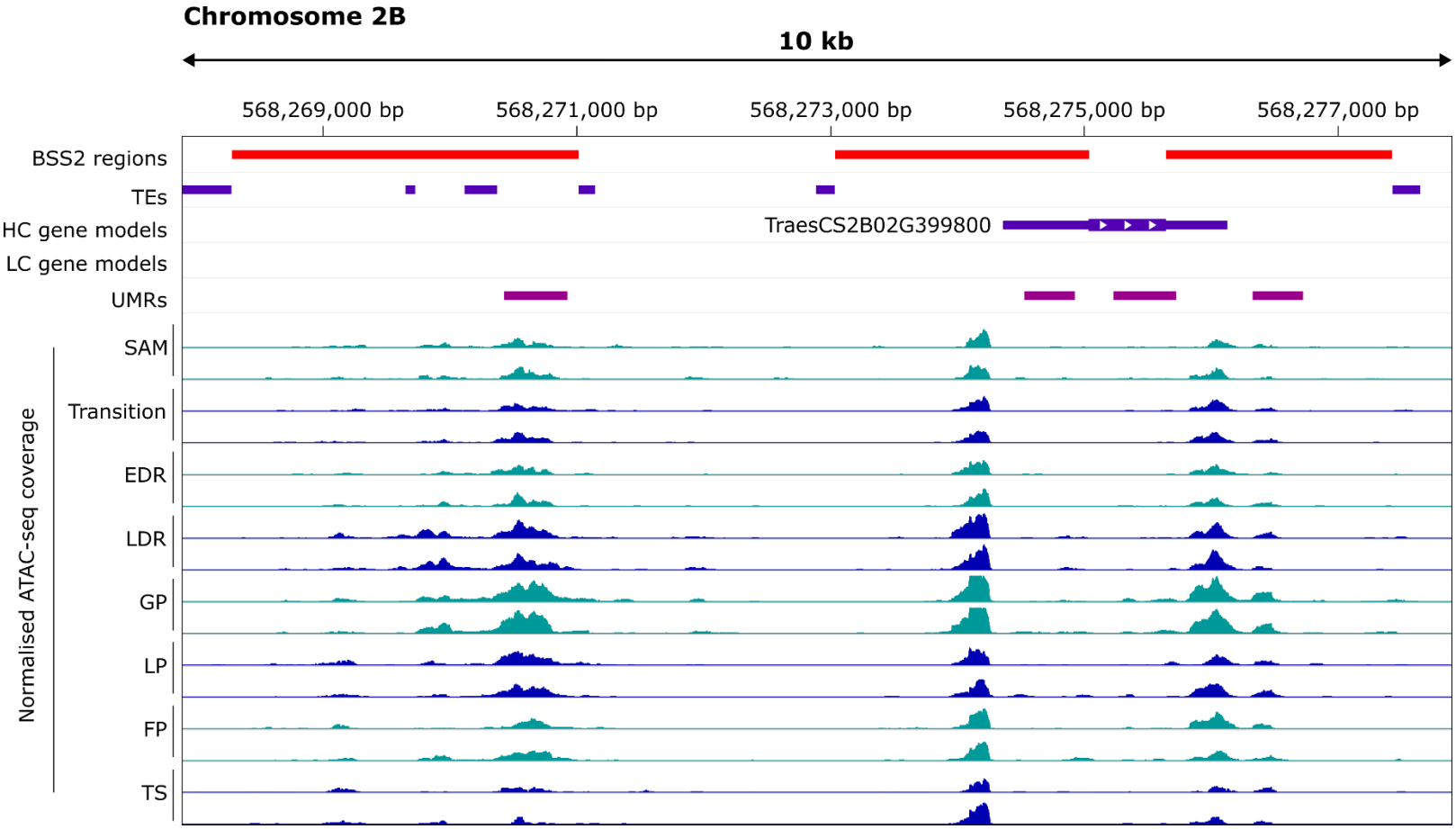


Figure 3.6 – Non-coding regions TraesCS2B02G399800 selected for basal spike-specific regulatory environment 2 (BSS2)
Annotated 10 kb region of chromosome 2B (IWGSC RefSeq v1.0 assembly). Purple tracks indicate transposable elements (TEs) plus high confidence (HC) and low confidence (LC) gene models (RefSeq v1.1 annotation). Magenta tracks indicates stably unmethylated regions (UMRs). Teal and blue tracks (alternated for clarity) indicate ATAC-seq coverage for spike tissue chromatin at various developmental stages (two replicates per stage). ATAC-seq coverage was normalised against sample read number and tracks were standardised to a maximum normalised coverage of 100 to aid visual comparison of lanes. The red track indicates the final regions selected to become BSS2 regulatory parts. Data visualised using the IGV (Robinson et al., 2011).

We assembled and validated these constructs as described in Methods 3.5.3. We handed over MJ_GG18, MJ_GG20, MJ_GG22, and MJ_GG24 to the JIC Wheat Transformation Team on 11/10/2023 as purified plasmids, and MJ_GG17, MJ_GG19, MJ_GG21, and MJ_GG23 on 20/12/2023 as Agrobacterium standard inoculums. For each construct, a minimum of two independent transgenic events were produced by the JIC Wheat Transformation Team, corresponding to a minimum of four T_0 plants (some calli produced multiple surviving plantlets; Table 3.2). Unfortunately, due to platform restructuring and reduced staffing levels, these plants were mostly delivered later than the expected six-month turn-around time (Table 3.2). In particular, MJ_GG17, MJ_GG20, and MJ_GG21 plants were only delivered in late November of 2024. Given the time needed for grain set and drying down, this precluded any analysis of T_1 plants from these lines within the timeframe of my PhD.

3.3.5 – BSS2, but not BSS1, drives spatiotemporally specific transcription and translation of reporter genes

To begin testing if our synthetic regulatory environments could drive spatiotemporally specific transcription of a target gene, we grew WT plants plus T_1 progeny from the independent T_0 plants MJ_GG24_2 (CN = 1), MJ_GG24_4 (CN = 7), and MJ_GG24_6 (CN = 8). These contained the tdTomato CDS under the control of the upstream and downstream BSS1 parts (BSS1::tdTomato). Similarly, we grew the T_1 offspring of MJ_GG23_1 (CN = 5), which drives tdTomato with the BSS2 parts (BSS2:tdTomato).

We extracted RNA from pools of whole spike tissue at the EDR, LDR, and GP stages. Four biological replicates were collected for each transgenic line, plus two biological replicates for WT Fielder. We also extracted RNA from leaf punches of three *Nicotiana benthamiana* plants transiently transformed with a construct for constitutive expression of nuclear localised NLS-tdTomato (Caldas et al., 2022). We obtained good quality RNA for all but one MJ_GG23_1 sample (Figure 3.7A).

Two DNA digestion steps were conducted prior to complementary DNA (cDNA) synthesis to ensure no residual genomic DNA (gDNA) remained. This precaution was taken because the tdTomato transgene did not contain introns, so gDNA and cDNA copies would not be distinguishable. We tested for residual gDNA using GAPDH primers which yield amplicons of different sizes from gDNA and cDNA (Figure 3.7B, Supp. Data 3.2). gDNA from WT Fielder was used as a control. All wheat spike samples except one MJ_GG24_4 replicate yielded the expected cDNA band and did not produce the larger gDNA bands.

Table 3.2 – Constructs for wheat misexpression transgenics

Parts are categorised by Golden Gate standard part nomenclature (Engler et al., 2014). 3U* is a non-standard part with 5' GCTT and 3' TAGA overhangs. TerP2 is a non-standard part with 5' TAGA and 3' CGCT overhangs. Independent event and plant counts only include cases with transgenic construct copy number > 0. Delivery date indicates date plants were transplanted to 11 cm (1 L) pots upon delivery by the JIC Wheat Transformation Team.

Construct ID	Pro + 5U	CDS1 ns	CT	3U*	TerP2	Independent events	Number of T_0 plants	Delivery date for T_0 plants	T_1 plants analysed?
MJ_GG17	BSS2-5PU	*MOF-B1*	3x FLAG	BSS2-3D	Nos	2	5	19/10/2024	
MJ_GG18	BSS1-5PU	*MOF-B1*	3x FLAG	BSS1-3D	Nos	3	6	04/04/2024	✓
MJ_GG19	BSS2-5PU	*SEP1-B6*	3x FLAG	BSS2-3D	Nos	3	4	07/08/2024	✓
MJ_GG20	BSS1-5PU	*SEP1-B6*	3x FLAG	BSS1-3D	Nos	5	15	24/10/2024	
		CDS1							
MJ_GG21	BSS2-5PU	GUS (with introns)		BSS2-3D	Nos	9	22	19/10/2024	
MJ_GG22	BSS1-5PU	GUS (with introns)		BSS1-3D	Nos	3 4	6 6	26/02/2024 04/04/2024	✓
		NT2	**CDS2**						
MJ_GG23	BSS2-5PU	NLS	tdTomato	BSS2-3D	Nos	1 1	1 3	07/08/2024 10/03/2025	✓
MJ_GG24	BSS1-5PU	NLS	tdTomato	BSS1-3D	Nos	2 1	4 1	26/02/2024 04/04/2024	✓

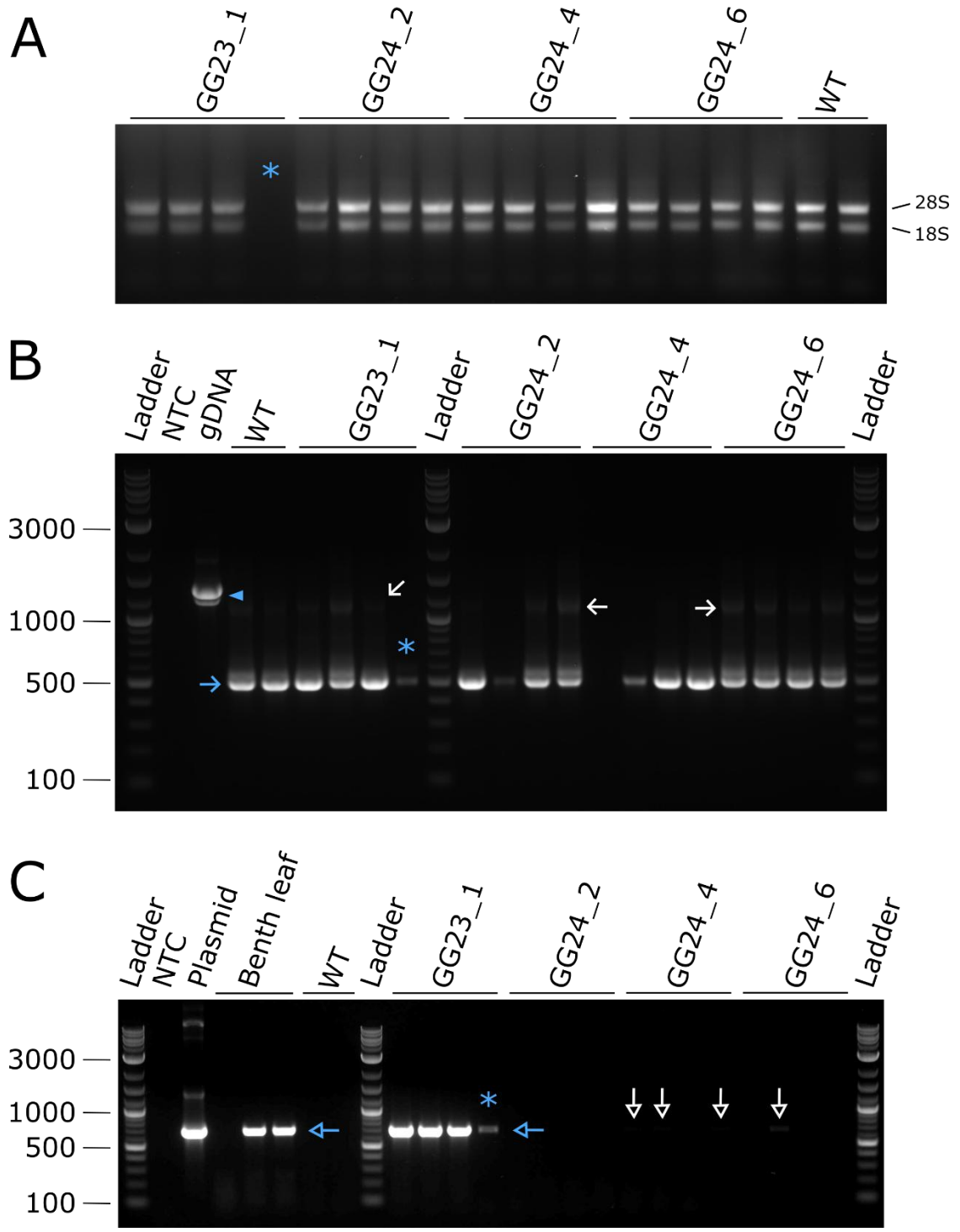


Figure 3.7 – BSS2, but not BSS1, drives strong expression of tdTomato transcripts in early wheat spikes

A, Total RNA extracted from pools of EDR, LDR, and GP spikes of WT Fielder and T_1 progeny of MJ_GG23_1 (BSS2::tdTomato) and MJ_GG24_2, _4, and _6 (BSS1::tdTomato). **B,** Amplification of genomic GAPDH and GAPDH transcripts to check for gDNA contamination of cDNA samples. 'gDNA' lane contains WT Fielder gDNA. Blue arrowhead indicates double band for genomic GAPDH (1384 bp for B homoeolog, 1259 bp for A and D). Blue arrow indicates 520 bp band for GAPDH transcripts. White arrows denote unexpected weak bands of <1,200 bp size. **C,** Amplification of tdTomato sequence. 'Plasmid' lane contains purified MJ_GG24 plasmid DNA. Unfilled blue arrows denote expected 696 bp tdTomato bands. Unfilled white arrows denote very weak tdTomato bands detected in some MJ_GG24_4 and MJ_GG24_6 replicates. Blue asterisks throughout denote an MJ_GG23_1 replicate for which little total RNA was recovered. Ladder units are bp. Ladder = NEB 1 kb Plus DNA Ladder. NTC = non-template control.

Some samples additionally produced very weak bands at higher fragment lengths that were nonetheless shorter than the predicted gDNA bands, potentially indicating residual, partially digested gDNA. Lastly, we tested samples for the presence or absence of tdTomato transcripts by PCR. Amplification was observed for three of four positive controls (three *N. benthamiana* cDNA samples and purified MJ_GG24 plasmid) and all four replicates of MJ_GG23_1 (BSS2; Figure 3.7C). Very weak amplification was detected for three replicates of MJ_GG24_4 and one replicate of MJ_GG24_6 (BSS1). This suggested that the BSS2 regulatory environment was driving strong spike expression of tdTomato at early spike stages, while BSS1 was driving only weak or no transgene expression.

We repeated this experiment with the T1 progeny of our other two BSS1::tdTomato lines, MJ_GG24_1 (CN = 10) and MJ_GG24_3 (CN = 5), which were derived from the same callus as MJ_GG24_2. For these, we extracted RNA from pools of whole spike tissue at the lemma and floret primordia stages (LP; W3.25 and FP; W3.5). This was a compromise as *MND-B1* – from which BSS1 was derived – is most highly expressed in EDR and LDR spikes, but collecting sufficient tissue from LP and FP spikes is considerably faster. We observed no amplification of tdTomato transcripts from the resulting cDNA samples (n = 4 per genotype; Supp. Data 3.3).

To further explore the efficacy of the BSS1 and BSS2 parts, we also imaged the young spikes and other tissues of our tdTomato lines by confocal microscopy. Leaf sections of one of the transiently transformed *N. benthamiana* plants were used as positive controls. We observed clear foci of tdTomato expression in the *N. benthamiana* samples (Figure 3.8), corresponding to nuclear localisation of the transgene. Similar foci could be observed in MJ_GG23_1 samples in the vegetative SAM and in the basal spikes at the transition, EDR, LDR, GP, and LP stages. In contrast, foci were much weaker and/or absent higher up the spike (Figure 3.8). There was also expression of tdTomato just below the spike in the upper stem, incipient peduncle, and leaf primordia. Foci were detected in these tissues throughout the above stages and up until CE (W5) stage. However, by the later W6, W7, and W7.5 stages, tdTomato foci were scarcely detectable in these tissues using the same microscope settings (Figure 3.9).

In contrast, no foci were detectable in MJ_GG24_2, MJ_GG24_4, MJ_GG24_6, or WT wheat spikes (Figure 3.8) at any spike stage using the same microscope settings. At higher laser powers, we did observe diffuse fluorescent signals in the tdTomato channel for all genotypes, particularly on cut surfaces and edges. However, given its presence in WT plants, this signal was presumed to be autofluorescence, potentially originating from a molecule produced during wounding responses.

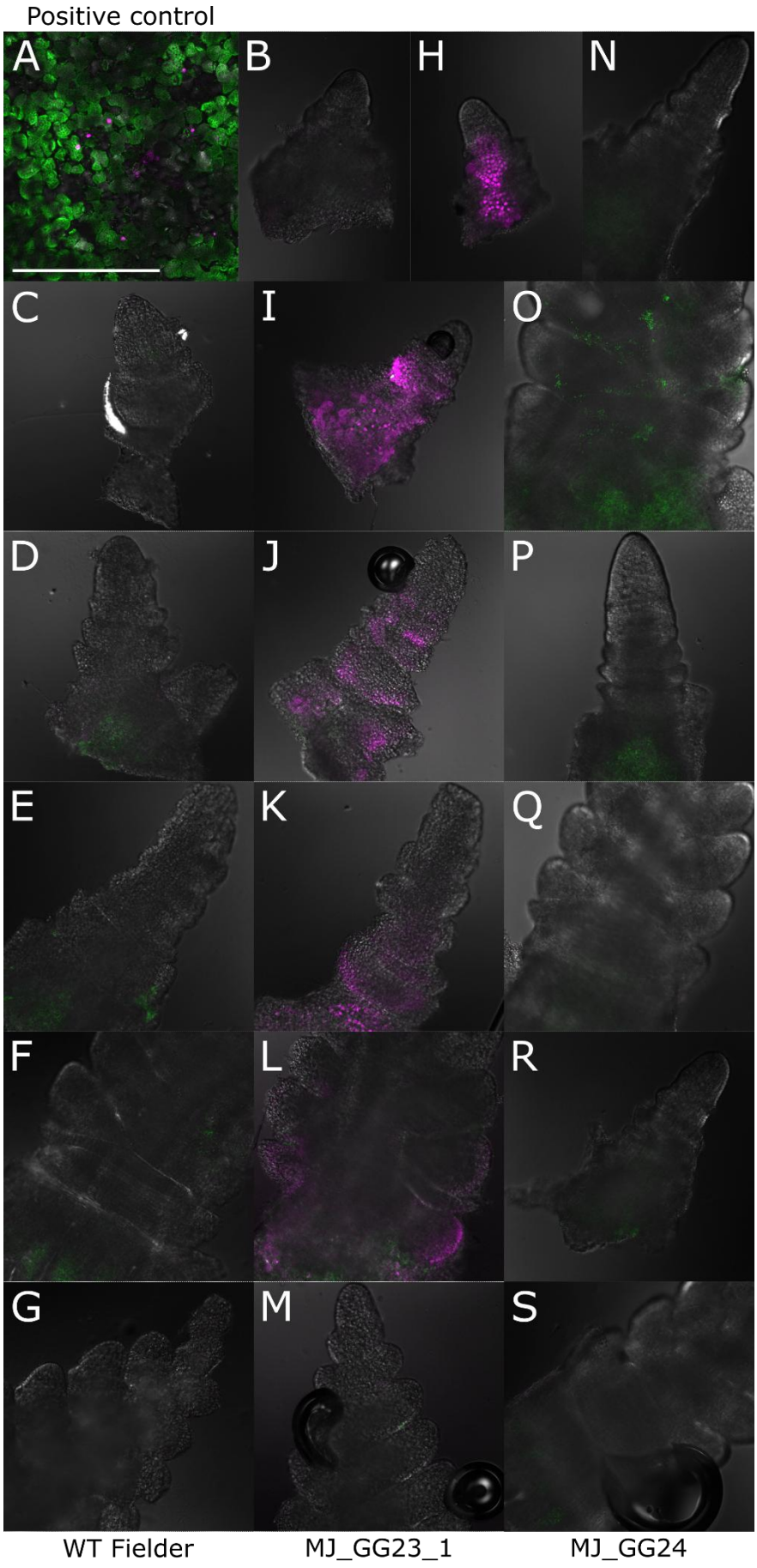


Figure 3.8 – BSS2 drives an acropetally weakening gradient of tdTomato protein production in wheat SAMs and early spikes Single plane multi-channel confocal images of NLS::tdTomato *Nicotiana benthamiana* mesophyll cells (**A**), WT Fielder (**B-G**), T_1 generation MJ_GG23_1 (**H-M**), MJ_GG24_2 (**N,O**), MJ_GG24_4, and (**P,Q**) MJ_GG24_6 (**R-S**). **B, H,** Vegetative SAM. **C, I,** Transition stage spikes (Waddington 1) **D, J, N, P, R,** Early double ridge stage spikes (W2). **E, K, Q,** Spikes between the late double ridge (W2) and glume primordia (W3) stages. **F, L, O, Q, S,** Basal sections of lemma primordia (W3.25) stage spikes. **G, M, S,** Apical sections of the same lemma primordia stage spikes. Greyscale channel = T-PMT, green channel = 647-721 nm (includes chlorophyll A emission maxima), magenta channel = 561-614 nm (includes tdTomato emission maximum). Scale bar represents 500 µm and is applicable to all images.

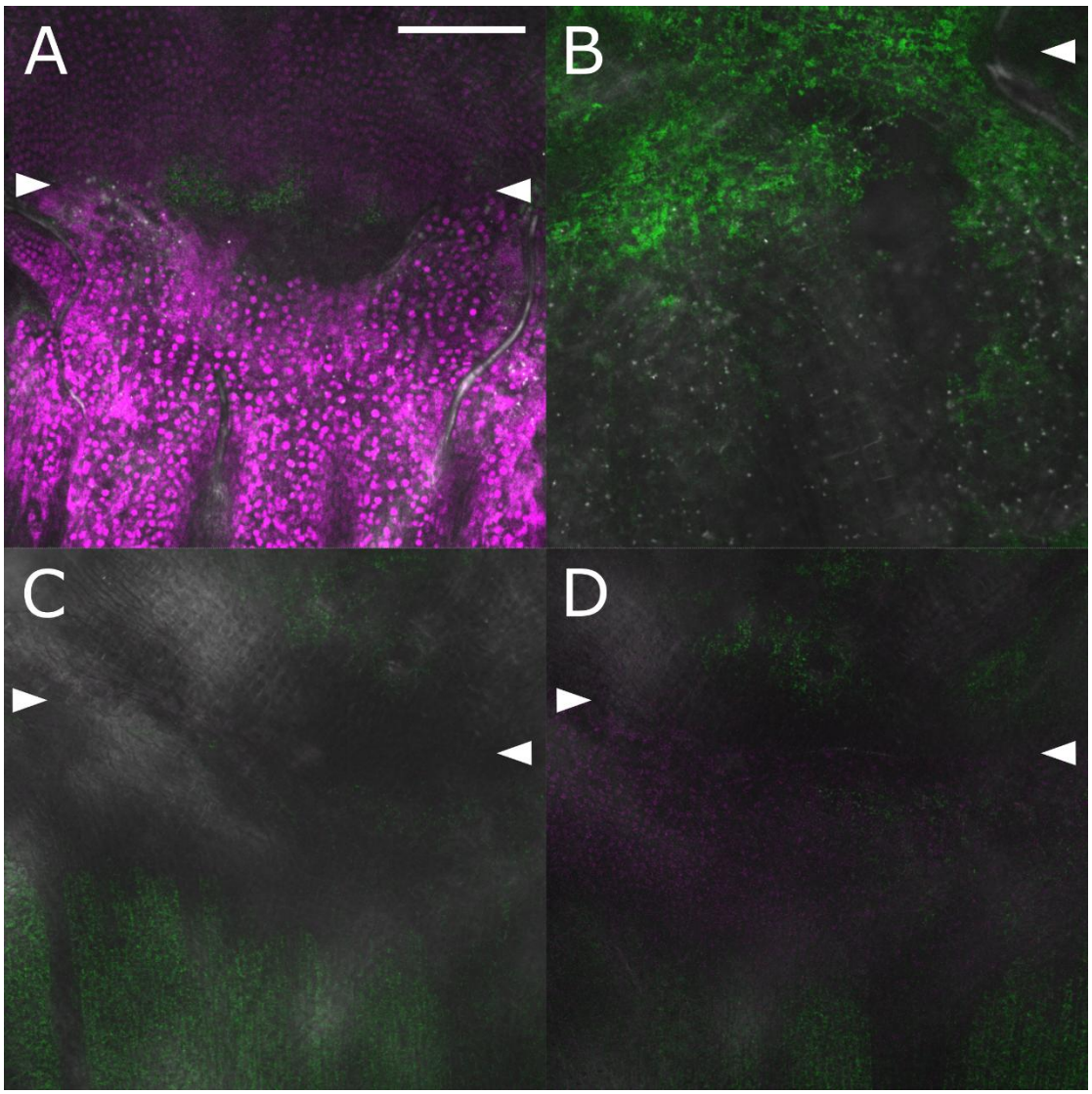


Figure 3.9 – BSS2 drives strong accumulation of tdTomato protein in the upper stem up until the Waddington (W) 5 stage, after which tdTomato foci are only visible using higher sensitivity settings

Single plane multichannel confocal images of rachis-stem boundaries of MJ_GG23_1 T_1 plants. **A,** W5 stage plant. **B,** W6-W7 stage plant. **C,D,** Two images of the same W7.5 stage plant. Microscope settings were identical for **A-C** while **D** was taken with higher laser power and master gain (Table 3.3). White arrowheads indicate approximate position of boundary between rachis (above) and stem (below). Greyscale channel = T-PMT, green channel = 647-721 nm (includes chlorophyll A emission maxima), magenta channel = 561-614 nm (includes tdTomato emission maximum). Scale bar represents 200 µm and is applicable to all images.

Damage-induced autofluorescence has been extensively reported in eudicots (Thomson et al., 1995; Bennett et al., 1996) and occasionally in cereals (Miller et al., 2023). Diffuse fluorescent signals, but no foci, were also observed in the seed coat and/or pericarp of WT and MJ_GG23_1 grain (Figure 3.10). Again, presence of fluorescence in the WT suggests autofluorescence rather than unexpected expression of the tdTomato transgene in MJ_GG23_1. No tdTomato foci were observed in WT or MJ_GG23_1 in various other tissues assayed, using either equivalent or more sensitive settings, including the root, 5th leaf, endosperm, or spikelet/floral organs at the W11 stage including glume, lemma body, lemma awn, stamen, and carpel (Figure 3.10).

Integration of transgene constructs into a genomic environment favourable for expression is a stochastic process. Therefore, while no signal was detected in the five MJ_GG24 lines described above (representing three independent events), we wanted to screen additional events to confirm whether or not the BSS1 regulatory environment could drive gene expression in wheat spikes. We therefore conducted GUS histochemical assays on T_1 spikes from 12 MJ_GG22 lines (BSS1::GUS$_{in}$; 7 independent events; CN = 1 to CN > 10). A transgenic wheat line with constitutive GUS expression was used as a positive control, while WT Fielder plants were used as negative controls. The positive control spikes turned a strong, deep blue during the GUS assay, while no blue colouration could be detected in either WT or MJ_GG22 spikes at the EDR, LDR, and GP stages (n = 4-8; Figure 3.11). T_1 seed from MJ_GG21 lines (BSS2::GUS$_{in}$) was not available in time to complete GUS assays.

Overall, these molecular and microscopy results suggest that BSS1 parts are not sufficient to drive basal spike specific gene expression. However, the detection of a very low level of tdTomato transcripts in two genotypes (MJ_GG24_4 and MJ_GG24_6) suggests that further transgenic events would need to be generated and screened to confirm this. On the other hand, BSS2 appears to drive strong, expression in the basal sections of developing wheat spikes which was largely spatiotemporally specific, though did extend to the upper stem / incipient peduncle and leaf primordia.

3.3.6 – Misexpression of *SEP1-B6* by BSS2 reduces RBS number

Given the detection of a weak PCR signal in two BSS1::tdTomato lines, we phenotyped mature T_1 plants grown in a controlled environment room (CER) from three independent BSS1::*MOF-B1* lines: MJ_GG18_1 (CN = 2), MJ_GG18_5 (CN = 8), and MJ_GG18_6 (CN = 2). The T_1 plants were assessed for copy number and those with CN > 0 (n = 10-12) were compared against WT plants (n = 12).

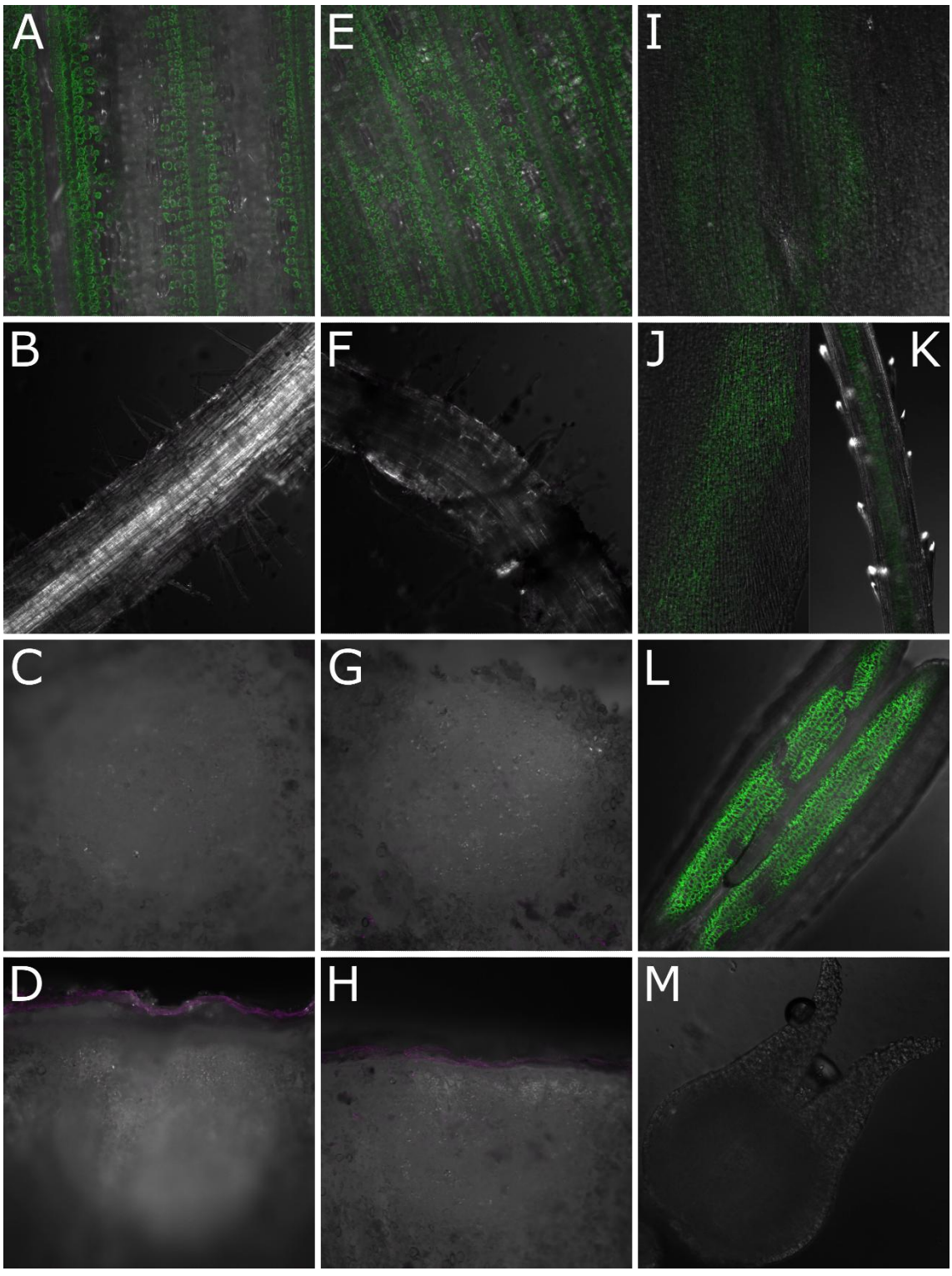


Figure 3.10 – BSS2 does not drive tdTomato protein accumulation in other non-target tissues
Single plane multichannel confocal images of WT Fielder (**A-D**) and T_1 generation MJ_GG23_1 (**E-M**). **A, E,** 5th leaf. **B, F,** Root. **C, G,** Mature grain endosperm. **D, H,** Mature grain endosperm, seed coat, and pericarp. **I-M,** Spikelet and floral organs at W11 stage, including glume (**I**), lemma body (**J**), lemma awn (**K**), anther (**L**), and carpel (**M**). Greyscale channel = T-PMT, green channel = 647-721 nm (includes chlorophyll A emission maxima), magenta channel = 561-614 nm (includes tdTomato emission maximum).

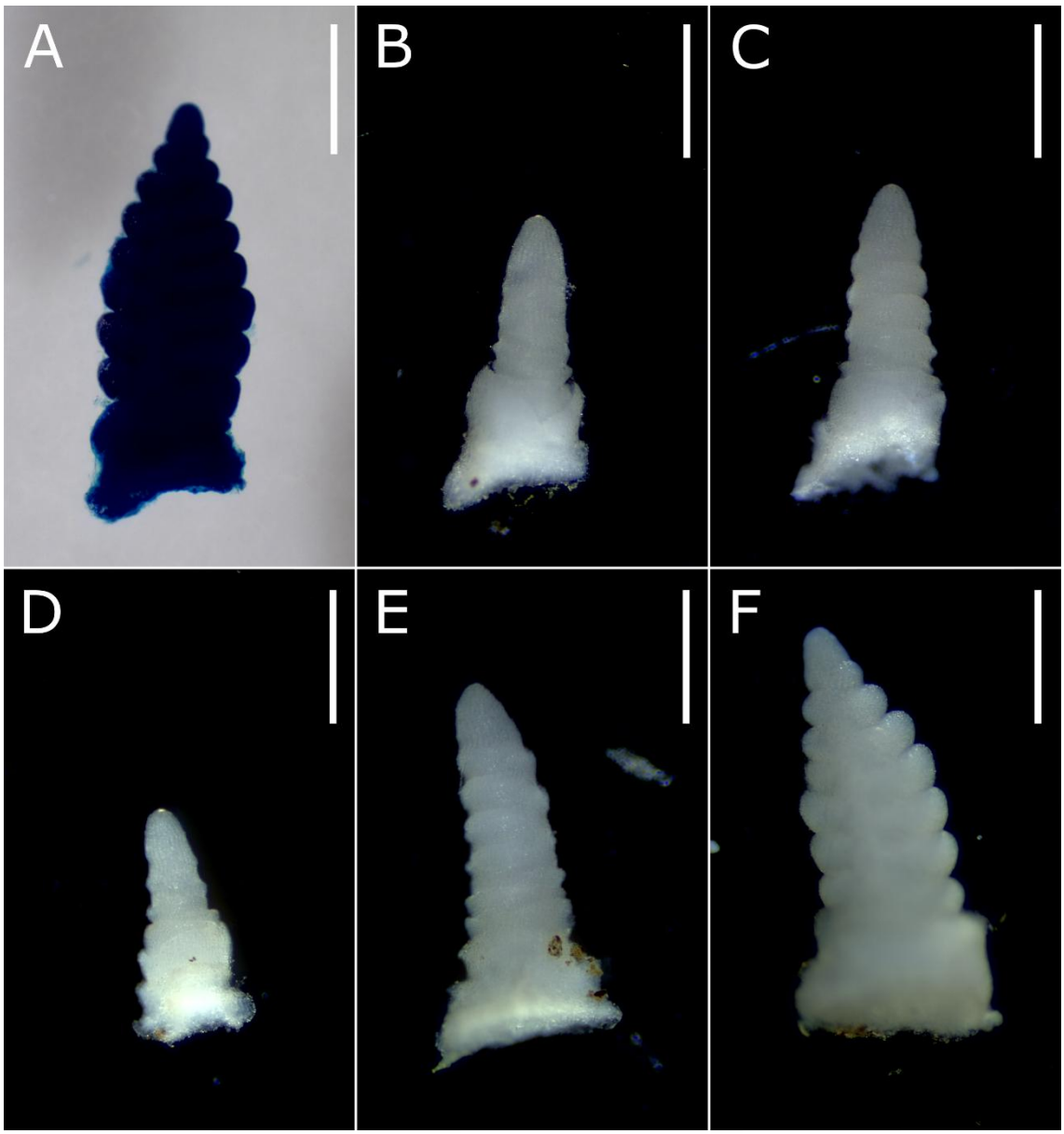


Figure 3.11 – BSS1 does not drive detectable expression of a GUS_{in} transgene in early wheat spikes

A-F, Detached early wheat spikes imaged after GUS staining and clearing with ethanol. **A,** Fielder with constitutive GUS_{in} construct. **B, C,** WT Fielder at EDR and LDR stages, respectively. **D, E, F,** T_1 progeny of MJ_GG22_2 (BSS1::GUS_{in}) at EDR, LDR, and GP stages, respectively. Scale bars represent 500 µm.

Presence of the BSS1::tdTomato construct had no significant effect on number of fertile tillers (includes main spike; p = 0.73; Figure 3.12A) or main spike length (p = 0.14) and an upward but marginally insignificant effect on main spike RBS number (p = 0.065; Figure 3.12C,D). There was, however, a significant negative effect of construct presence on main spike spikelet number (mean difference = -3.0; $p = 9.2 \times 10^{-4}$; Figure 3.12B), although there was no difference amongst the transgenic genotypes (p = 0.25). The equivalent tests were not conducted for BSS1::*SEP1-B6* lines due to time constraints.

We also phenotyped mature CER T_1 plants deriving from four BSS2::*SEP1-B6* lines (three independent transformation events). These were MJ_GG19_1 (CN = 2) and MJ_GG19_2 (CN = 3; same event), plus MJ_GG19_3 (CN = 2) and MJ_GG19_4 (CN = 22). The T_1 plants were assessed for copy number and those with CN > 0 (n = 6-13) were compared against WT plants (n = 12). Presence of the BSS2::*SEP1-B6* construct had no significant effect on spikelet number (p = 0.10; Figure 3.13B) or number of sterile apical spikelets (p = 0.25). However, construct presence was associated with a moderate, but significant, decrease in main spike peduncle length (mean difference = -52 mm, p = 0.035) and an increase in main spike length (mean difference = 11 mm, $p = 2.4 \times 10^{-4}$), although plant height did not change overall (p = 0.20; Figure 3.14). More strikingly, construct presence was linked to large decreases in fertile tiller number (mean difference = -6.7, $p = 2.0 \times 10^{-7}$; Figure 3.13A) and main spike RBS number (mean difference = -1.1, $p = 4.7 \times 10^{-4}$; Figure 3.13C-E). Amongst these, only fertile tiller number exhibited differences amongst the transgenic genotypes (p = 0.017). There was a strong correlation between fertile tiller number and RBS when calculated across both WT and BSS2::*SEP1-B6* – Pearson's correlation coefficient (r) was 0.52 ($p = 2.2 \times 10^{-4}$) and Spearman's correlation coefficient (r_s) was 0.57 ($p = 5.2 \times 10^{-5}$; Figure 3.15). However, upon exclusion of the WT plants, this correlation became weaker and non-significant (r = 0.26, p = 0.14; $r_s = 0.31$, p = 0.08). Repeat experiments in CER, glasshouse, and/or field conditions will be required to confirm the replicability of each of these results.

3.3.7 – A *mof1* mutant produces significantly larger floral organs

Given the different spikelet architecture in wheat (multiple florets initiated, indeterminate) compared with rice (single floret initiated, determinate), we wanted to test if the phenotypes observed in rice *mof1* mutants were conserved in wheat. We therefore aimed to generate a *mof1* knock-out mutant in wheat. Stop-gain or frame-shift mutations were not available for any *MOF1* homoeologs in either the tetraploid wheat (cv. 'Kronos') or hexaploid wheat (cv. 'Cadenza') TILLING populations (Uauy et al., 2009; Krasileva et al., 2017).

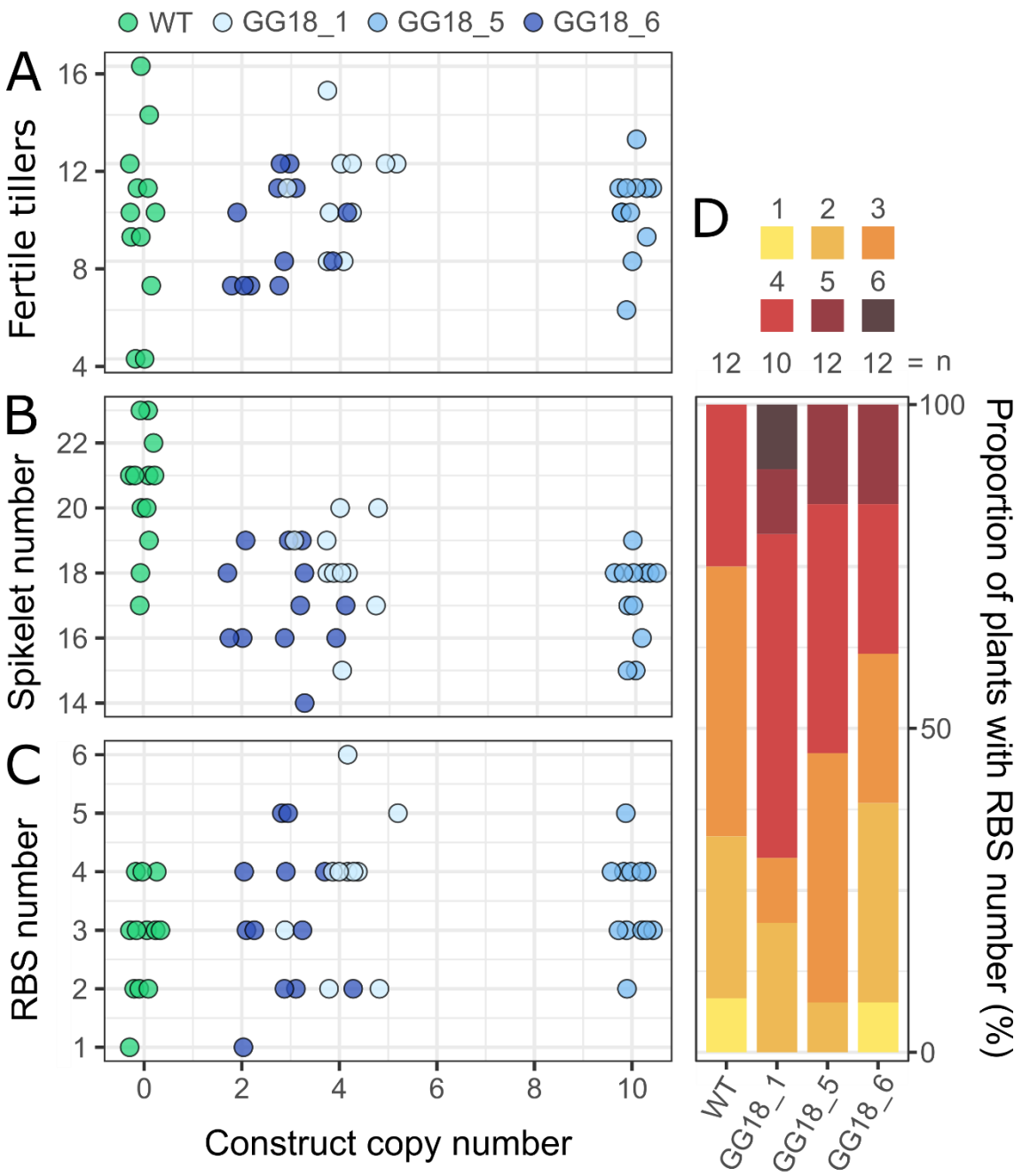


Figure 3.12 – Plants containing BSS1::*MOF-B1* constructs produce significantly fewer spikelets
A-C, Points represent measurements from independent plants taken after senescence. Copy numbers measured to integer values; points jittered to improve clarity. **A,** Number of fertile tillers (includes main spike). **B,** Spikelet number on main spike (includes non-grain-bearing spikelets). **C,** Number of rudimentary basal spikelets (RBS) on main spike. **D,** Proportion of plants from each genotype with different main spike RBS numbers (same data as **C**).

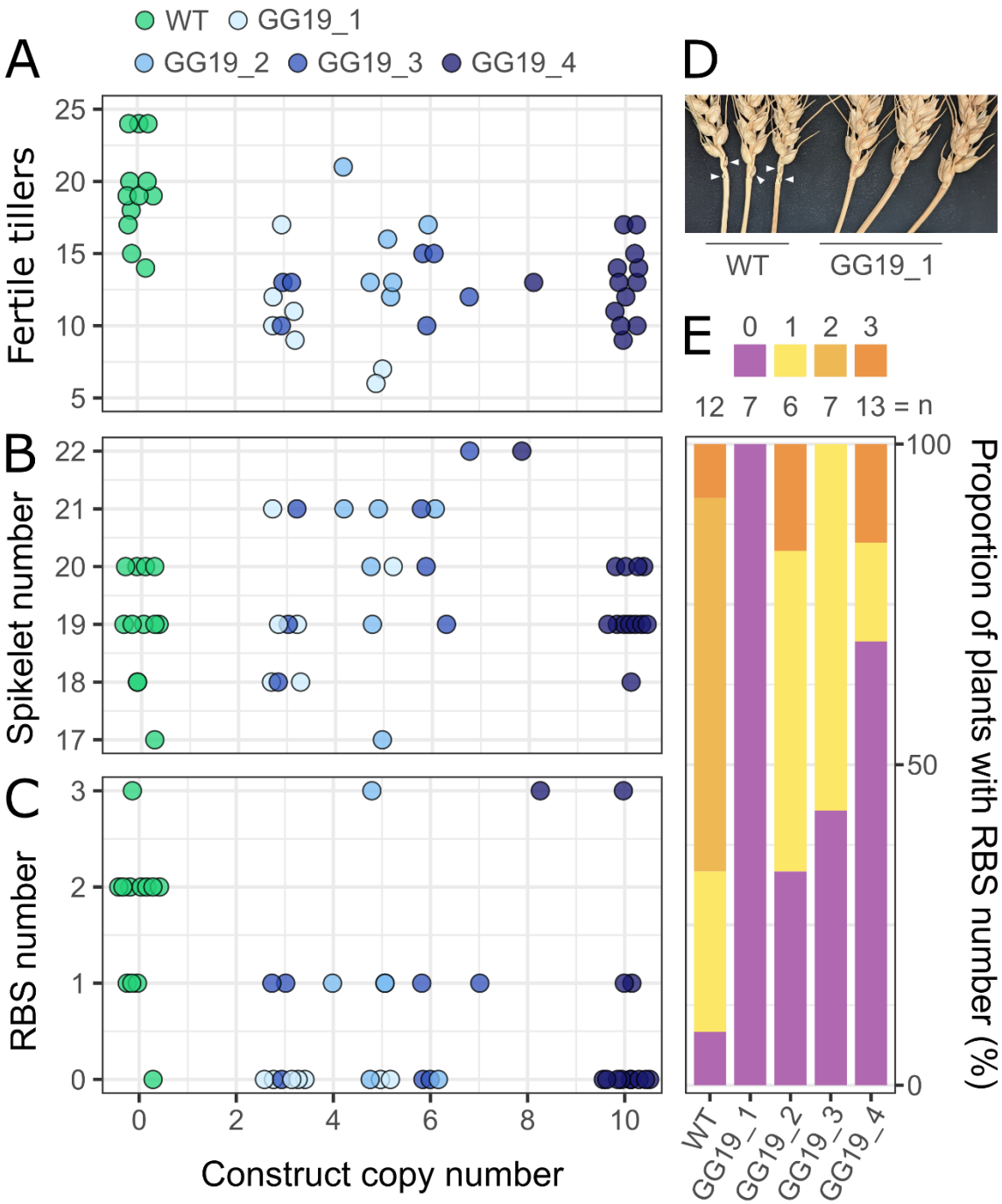


Figure 3.13 – Plants containing BSS2::*SEP1-B6* constructs produce significantly fewer main spike rudimentary basal spikelets (RBS)

A-C, Points represent measurements from independent plants taken after senescence. Copy numbers measured to integer values; points jittered on x-axis to improve clarity. **A,** Number of fertile tillers (includes main spike). **B,** Spikelet number on main spike (includes non-grain-bearing spikelets). **C,** Number of RBS on main spike. **D,** Basal spikelets of example WT and MJ_GG19_1 plants. White arrowheads denote RBS. **E,** Proportion of plants from each genotype with different main spike RBS numbers (same data as **C**).

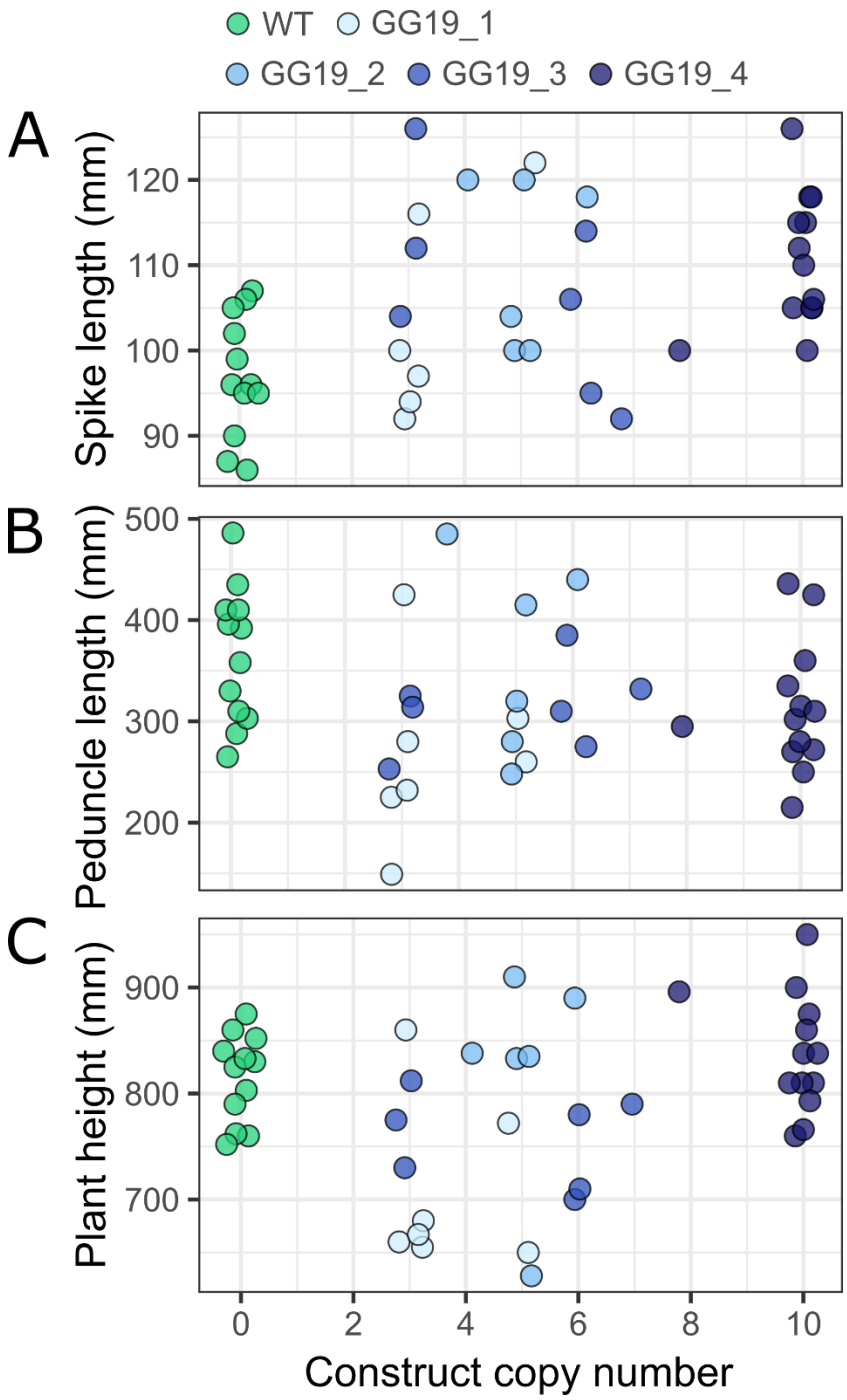


Figure 3.14 – Main spikes on BSS2::*SEP1-B6* plants are significantly longer but have shorter peduncles

A-C, Points represent measurements from independent plants taken after senescence. Copy numbers measured to integer values; points jittered on x-axis to improve clarity. **A,** Main spike length, excluding awns. **B,** Main spike peduncle length. **C,** Plant height from soil surface, excluding awns.

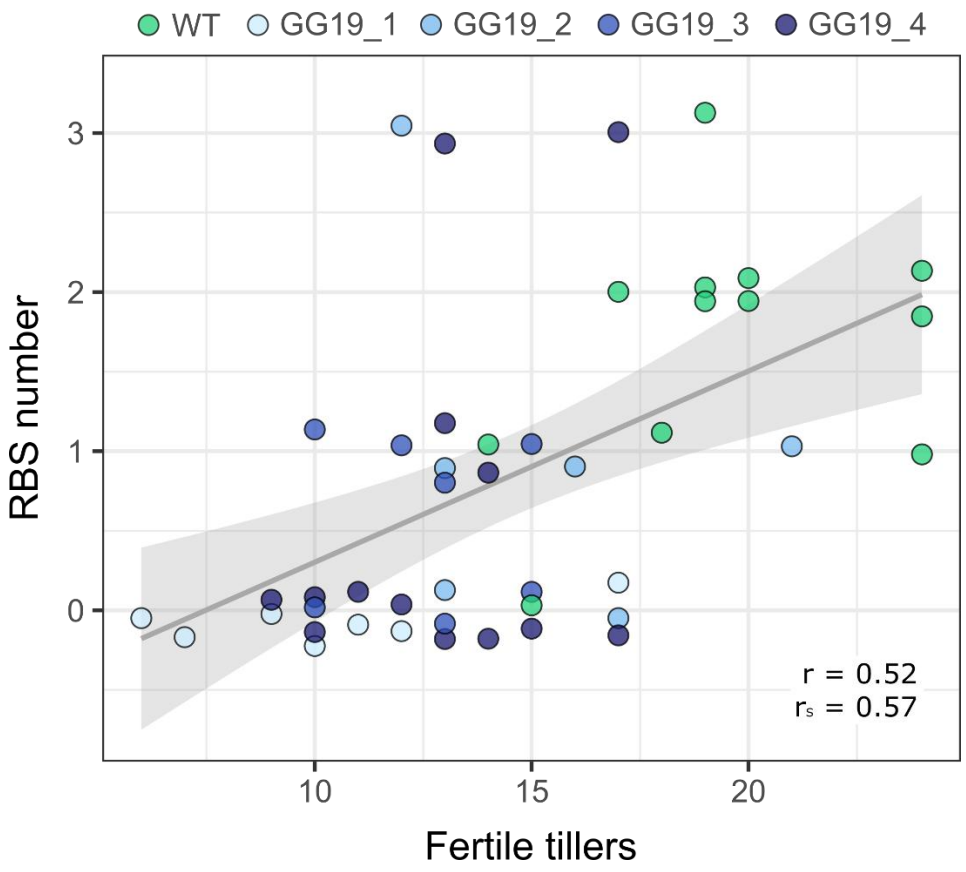


Figure 3.15 – Fertile tiller number is strongly positively correlated with main spike RBS number when measured across WT and BSS2::*SEP1-B6* lines
Points represent measurements from independent plants taken after senescence and are jittered on the y-axis to improve clarity. r is Pearson's correlation coefficient and r_s is Spearman's correlation coefficient. Grey line denotes the line of best fit and grey ribbon denotes the 95% confidence interval.

However, an exon 2 splice-donor variant for the A homoeolog (Kronos4375.chr2A.656967754) and an exon 3 splice-acceptor variant for the B homoeolog (Kronos2630.chr2B.603921626) were identified (Figure 3.16A,B). The next two possible splice-donor sites (GT) for the A homoeolog and the next three possible splice-acceptor sites (AG) for the B homoeolog are out of frame, suggesting that exons three to six could be translated out of frame for both genes.

Custom KASP markers were developed for the target mutations (Supp. Data 3.2) and used to produce double mutant (*mof1*) and double WT (*MOF1*) F_2 plants (Figure 3.16C). KASP genotyping was validated by Sanger sequencing amplicons of both homoeologs from F_2 *mof1* (n=5) and *MOF1* (n=3) plants.

Additionally, we wanted to confirm that the splice site mutations produced truncated or out of frame transcripts and would therefore represent effective knock-outs. RNA was extracted from EDR meristems of the selfed progeny of *mof1* and *MOF1* F_2 plants and cDNA was synthesised. We designed primer sets to amplify across the affected intron-exon boundaries, one specific to *MOF-A1*, one to *MOF-B1*, and one common to both homoeologs. Initial tests suggested the A subgenome and common primer sets amplified the target cDNA transcripts, while the B subgenome primers produced only very weak bands (Figure 3.17). For both the A subgenome and common primer sets, the amplicons from *mof1* plants were notably shorter than those from *MOF1* plants based on gel migration. This suggests that alternate splice sites were indeed utilised for at least *MOF-A1*. We aim to develop a working homoeolog-specific assay for *MOF-B1* and then to Sanger sequence amplicons from both homoeologs to confirm that alternate splice sites are used in our *mof1* plants and whether the remainder of the transcript would be translated out of frame or truncated in each case.

Glasshouse F_2 plants were then inspected at anthesis for wheat equivalents of the 'more floret' mutation observed in rice. In rice, this manifests as multiple sets of floral organs per spikelet, including two carpels, where normally rice determinately produces a single floret per spikelet (Li et al., 2020; Ren et al., 2020). However, it is unclear from these publications if rice *mof1* actually produces multiple florets per spikelet or if there is over-proliferation of floral organs within a single floret. Given that wild-type wheat already produces multiple florets per spikelet, we were interested to see if abnormalities may arise in the form of supernumerary floral organs within florets. However, examination of *mof1* spikelets and florets by stereo microscope did not reveal any developmental abnormalities.

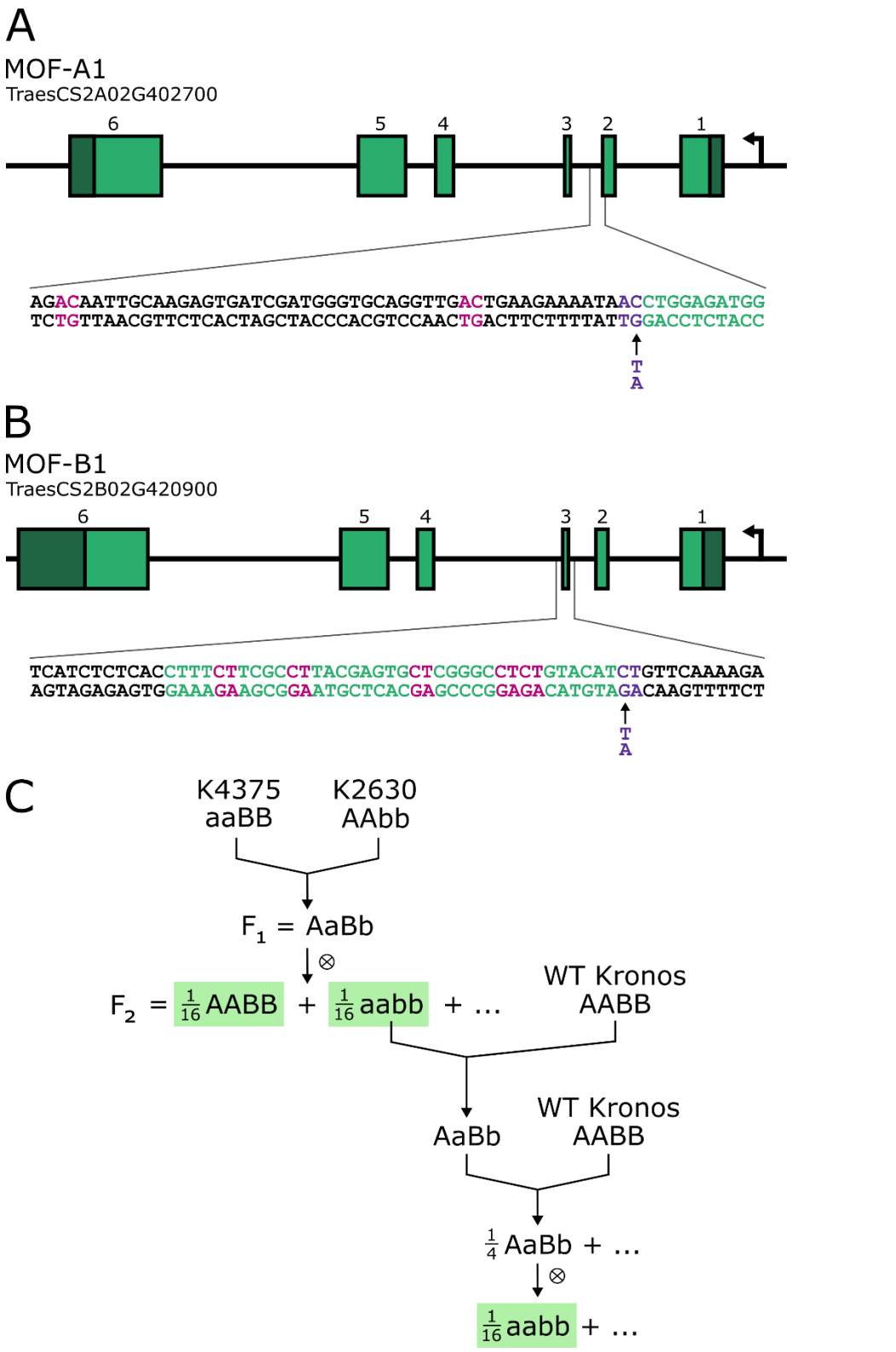


Figure 3.16 – Splice site SNPs are available for both *MOF1* homoeologs in the Kronos TILLING population

A,B, Location of selected SNPs at the exon 2 splice donor site of *MOF-A1* (**A**) and at the exon 3 splice acceptor site of *MOF-B1* (**B**). Green boxes and sequences represent exons, while dark green regions represent UTRs. Introns and exons are to scale. Arrows denote target SNPs within wild-type splice acceptor/donor sites (purple). Putative alternate downstream acceptor/donor sites within 50 bp are depicted in magenta. **C,** Crossing scheme for generation of *mof1* mutants from TILLING lines K4375 and K2630, plus subsequent backcrosses to WT Kronos. Target genotypes are highlighted in green.

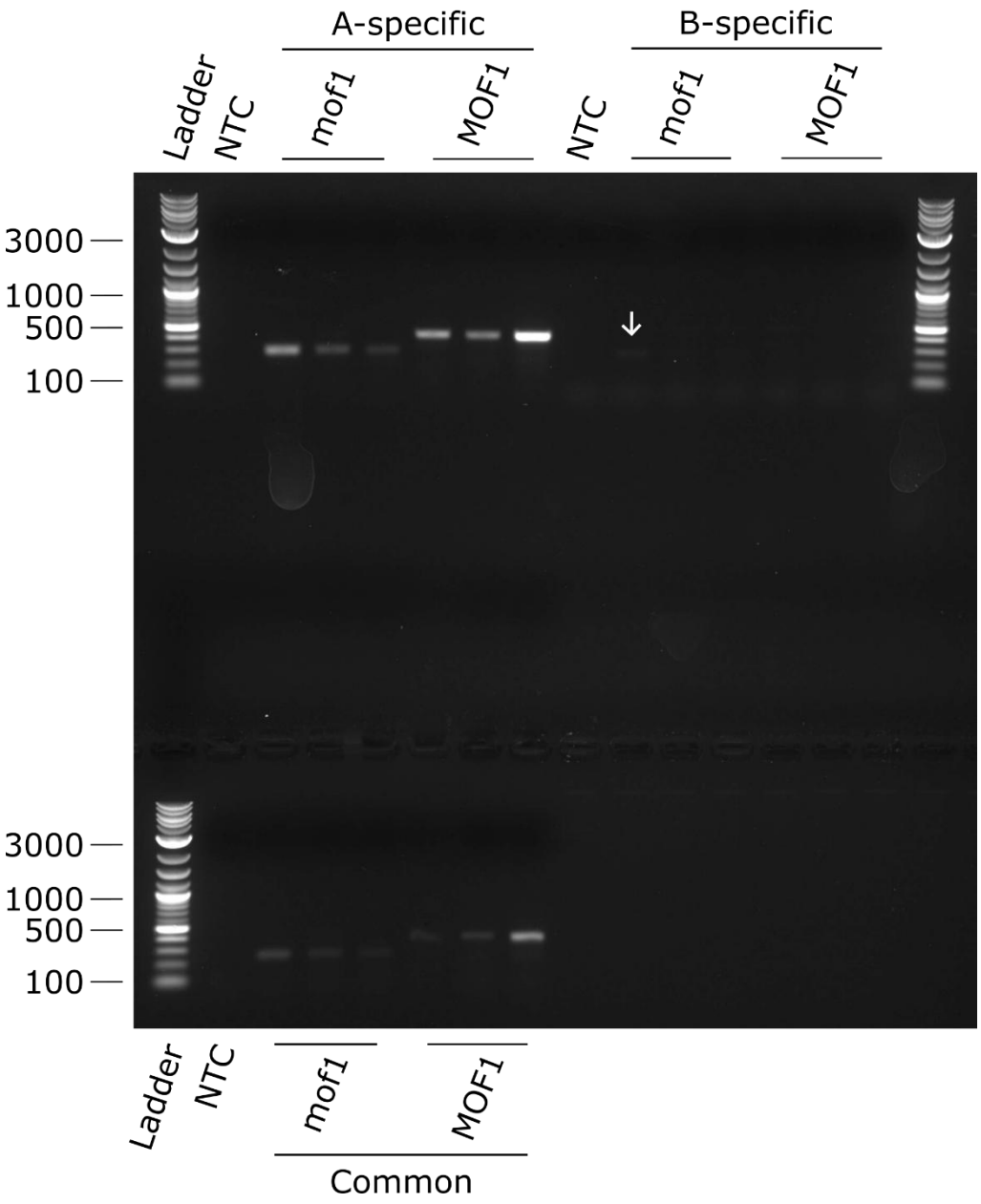


Figure 3.17 – Putative *mof1* mutants produce shorter MOF-A1 transcripts versus WT
Reverse transcription PCR using A subgenome-specific (*MOF-A1*) and common (*MOF-A1/B1*) primer sets produce smaller bands for putative *mof1* mutants than WT *MOF1* sister lines. B subgenome-specific (*MOF-B1*) primers amplified weakly (white arrow) or not at all. Ladder unit is bp. Ladder = NEB 1 kb Plus DNA Ladder. NTC = non-template control.

Given the lack of the rice *mof1*'s most striking phenotype, we also phenotyped F_2 plants at maturity for a variety of additional traits. No difference was found between *mof1* and *MOF1* for plant height (n = 22/15, p = 0.10), number of fertile tillers (n = 22/15, p = 0.45), main spike length (n = 13/8, p = 0.81), or main peduncle length (n = 13/8, p = 0.09). Unfortunately, most main spikes were not phenotyped for spikelet number, RBS, or number of sterile apical spikelets before they were destructively analysed (see below). However, there was no difference between *mof1* (22 plants) and *MOF1* (14 plants) for these traits across all fertile tiller spikes (p > 0.2). The growth conditions resulted in zero RBS for the majority of these spikes for both *mof1* (115 spikes; 58%) and *MOF1* (78 spikes total; 67%), suggesting they were inappropriate to assay RBS response. The large pots used (1 L) would have provided greater access to light and nutrients than typical field conditions, leading to high source strength and a reduced incidence of RBS, as observed in (Tamagno et al., 2024).

Additionally, we dissected mature main tiller spikes of *mof1* (n = 17) and *MOF1* (n = 10) and measured the lengths and widths of their glumes, lemmas, paleas, and grains (Figure 3.18). We compared organ size parameters between the two genotypes using three layers of ANOVA model (section 3.5.6). Model [1] contained the interaction terms genotype*spikelet and genotype*spikelet:floret, model [2] only contained the genotype*spikelet interaction, and model [3] contained no interactions.

We found that for model [1] there were very few significant interactions (p < 0.05) between genotype and spikelet-floret combinations (Supp. Data 3.4). For glume width, only spikelet1-floret2 and spikelet13-floret2 produced significant interactions with genotype, out of 14 possible combinations. For palea length, only spikelet13-organ3 produced a significant interaction with genotype, out of 42 possible combinations. Similarly, for both grain width and grain length, only 2/42 spikelet-floret combinations interacted significantly with genotype. There were no significant interactions for glume length, lemma width, lemma length, and palea width. We concluded that [1] was not appropriate for any organ size parameter, as very few combinations interacted with genotype and there was no pattern as to which combinations were significant.

We then checked model [2] for each organ size parameter to see if there were any significant interactions (p < 0.05) between genotype and spikelet position. Again, most organ size parameters showed zero significant interactions. Significant interactions were observed for glume width between genotype and spikelet 4, palea length between genotype and spikelet 13, and grain width between genotype and spikelets 2, 6, 7, and 10. We concluded that model [2] was not appropriate for most organ size parameters, though appeared to be useful for describing grain width.

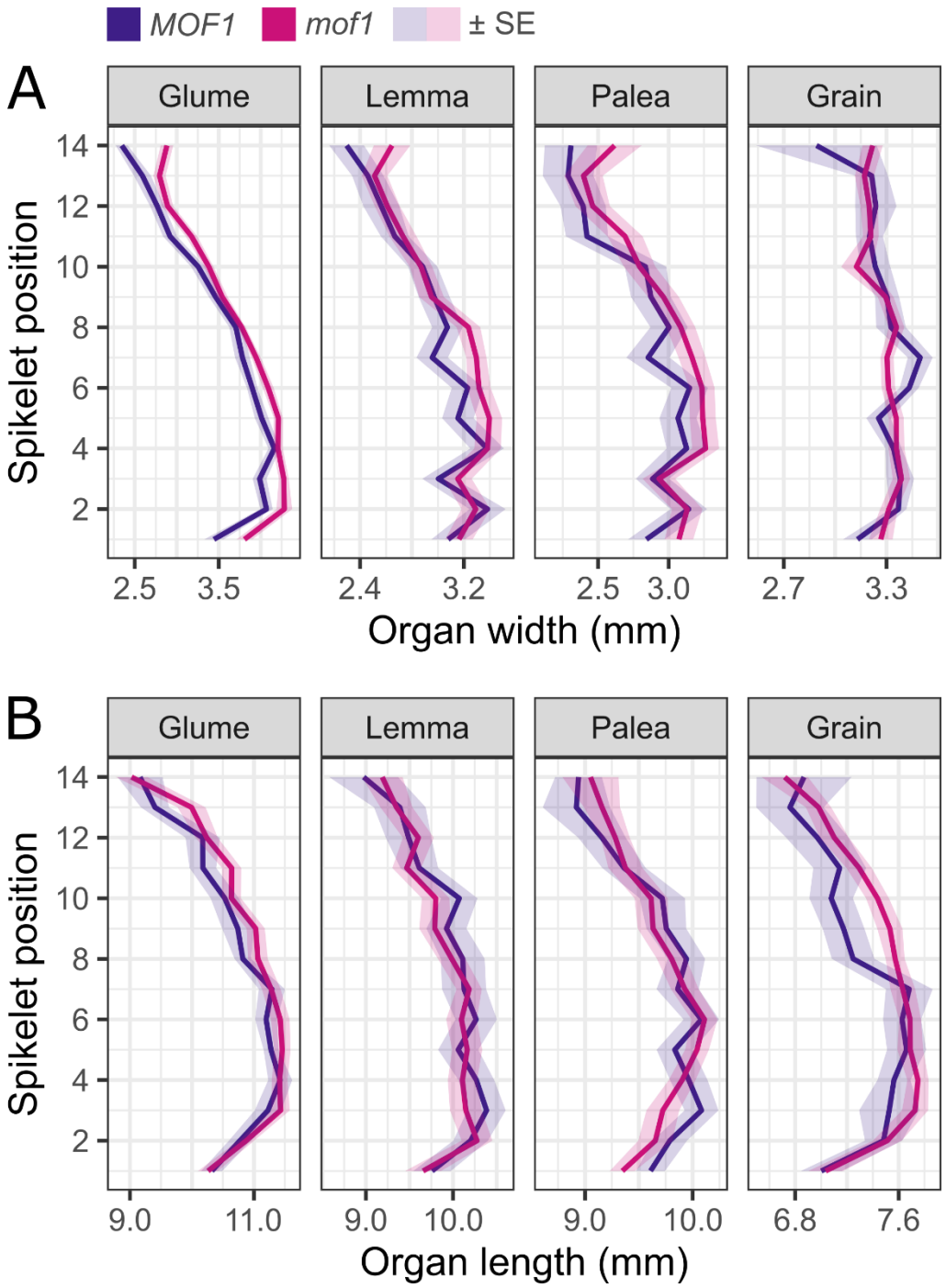


Figure 3.18 – Mature floral organs were wider and/or longer in *mof1* plants versus *MOF1*
Comparison of **A,** widths and **B,** lengths of *mof1* and *MOF1* floral organs at maturity. Darker lines represent the mean, while lighter ribbons represent ± 1 standard error. Data from florets 1 to 4 for spikelets 1 (basal) to 14 (apical) is presented. A small number of datapoints (1.18% of total) from 5th florets and spikelets 15 and 16 were excluded before plotting.

For model [3] (*y ~ genotype + spikelet + spikelet:floret*), genotype was a significant term for many organ size parameters, with *mof1* tending to produce slightly larger organ measurements. *mof1* exhibited significantly altered glume widths (+0.19 mm, $p = 1.4x10^{-10}$), glume lengths (+0.17 mm, p = 0.015), lemma widths (+0.068 mm, p = 0.036), lemma lengths (-0.13 mm, p = 0.030), palea widths (+0.13 mm, $p = 9.8x10^{-5}$), and grain lengths (+0.13 mm, p = 0.0012). Genotype was not significant in this model for palea length or grain width.

We also backcrossed F_2 *mof1* plants to WT Fielder twice then self-pollinated these to produce *mof1* mutants in a cleaner background (i.e. fewer non-target TILLING mutations; Figure 3.16C). Unfortunately, it was not possible to conduct phenotyping on these before my thesis hand-in date.

3.4 – Discussion

3.4.1 – Semi-spatial RNA-seq is a useful tool for hypothesis generation and candidate gene selection

Here, we utilised RNA-seq on pooled sections of early wheat spikes to create an improved atlas of semi-spatial gene expression across development. This dataset covers a broader range of spike developmental stages than previous work (Backhaus et al., 2022), from EDR to CE (W2 to W5), utilises two genotypes contrasting for our trait of interest, and offers lower variability between replicates, enabling our dataset to detect smaller differences in gene expression between sample types. This was achieved by pooling multiple spikes rather than having each replicate derive from a single spike section. This approach has previously been considered technically challenging (Backhaus et al., 2022) due to the difficulty of accurately staging and sectioning many spikes (up to 30 for a single EDR replicate) in a repeatable fashion. Nonetheless, our PCA suggests we achieved good clustering of like samples and good separation of unlike samples. This may be partially due to setting a strict formula for harvesting semi-spatial sections (see Methods).

Our data supports previous findings of greater differences in gene expression between spatial sections of a single spike stage than there are between stages (Backhaus et al., 2022). Our PCA also indicated that this effect is stronger earlier in development (EDR and LDR stages) and weaker later on (TS, CE), suggesting a spike-wide convergence on common gene regulatory networks. This was reaffirmed when we later conducted further central-basal differential expression analyses on all stages to support (Long et al., 2024); the number of DE genes (padj < 0.001, IHW corrected) was highest at EDR (7,250) and LDR (6,410), dropped sharply by LP (3,057), and diminished further by TS (1,872) and CE (2,086) (Supp. Data 3.5). It is tempting to speculate that this network convergence may play a role in enabling the relatively synchronous imposition of the floret abortion checkpoint at 20-10 days pre-anthesis. An alternative interpretation is that because correct staging is more complex at later stages and fewer spikes were required per replicate, the pooled tissue collected was more variable, reducing our power to detect spatial differences.

This data allowed us to identify candidate genes whose misexpression might modulate basal spikelet productivity. *MOF-1* and *SEP1-6* were initially captured in a set of 547 TFs with differential expression between basal and central portions of the spike at EDR. These genes also passed a further round of filtering which scraped the FunRiceGenes database for rice orthologs with functionally validated inflorescence development annotations. Lastly, both

genes showed considerably weaker expression in the basal versus the central spike at EDR and LDR, a criterion imposed to identify putative positive, rather than negative, regulators of spikelet productivity. Given the *LOFSEP* clade's involvement in the basipetal floral signalling gradient opposing the *SVP* clade's acropetal vegetative gradient, the appearance of *SEP1-6* in our final candidate set offered some validation of our methodology. Previous work in rice has also ascribed a role in the SM-to-FM transition to *MOF1*, though no prior work been conducted on this gene in wheat. These factors made this an interesting, though more speculative, candidate to follow up on.

Alternative methods to mine our RNA-seq data for potential regulators of basal spikelet productivity can also be envisaged. For example, stage-section combinations could be explored for enrichment for genes involved in plant hormonal processes using an existing catalogue of orthology-based annotations for wheat (Jones et al., 2022). This would be highly complementary with existing projects to utilise hormone biosensors in wheat (Dao et al., 2023; Dr Stephen Pearce, personal communication). More speculatively, this latter approach could be combined with recently developed techniques for live imaging of developing wheat spikes *ex situ* (unpublished data) for spatiotemporal tracking of hormone concentrations.

We additionally used our RNA-seq data in concert with complementary sequencing datasets to facilitate our attempts at complex manipulations of *MOF1* and *SEP1-6* expression. Given a dearth of previously characterised wheat spike promoters, we explored our data for genes with expression patterns we wished to impart upon our candidate spikelet productivity regulators. We selected genes with strong basal expression at EDR and LDR – declining in subsequent stages – plus low or no central expression throughout the timecourse. To further minimise pleiotropic effects, we also aimed to select genes with low expression in other tissues such as the roots, leaves, stem, floral organs, and grains. This was achieved by analysing publicly available TPM data hosted on the ExpVIP expression browser (Borrill et al., 2016) and available for bulk download on the Grassroots repository (Bian et al., 2017). Lastly, we utilised published ATAC-seq data to identify gene-proximal regions which might be contributing to the desired expression pattern and incorporated these into size-efficient synthetic regulatory environments. These approaches highlight the wealth of public, wheat-specific sequencing datasets now available and the value of integrating these inform experimentation.

Beyond this work, our data and methodology has already proved useful for additional projects. Firstly, our data was recently used to inform the selection of a gene panel for MERFISH spatial transcriptomics (Long et al., 2024). This technique was used to achieve

cellular resolution transcript detection for 200 genes across four stages of wheat spike development. Given the authors' interest in examining spatial expression patterns, our data on central-basal DE genes supplied the majority of genes used for probe design. Our data therefore contributed to a resource which overcomes its own major limitation; that it aggregates gene expression over many cell and tissue types, obscuring subtle patterns of gene expression in incipient organs. The unprecedented level of precision afforded by MERFISH overcomes these issues and will likely allow fresh insights into spikelet development and the phenomenon of low basal spikelet productivity.

Our transcriptomic timecourse and use of existing ATAC-seq data have also informed an ongoing gene-editing project at the John Innes Centre. The aim of this project is to mutagenise the CREs of known spike development regulators to generate novel allelic series with subtle phenotypes more likely to be beneficial in agronomic contexts. By investigating the temporal gene expression and chromatin accessibility patterns of these genes, we identified multiple putative CREs in each for targeting by CRISPR-Cas9. Inspired by a multiplex gene editing approach named 'BREEDIT' (breeding and gene editing; Lorenzo et al., 2023), guides targeting multiple genes have been incorporated into a small number of constructs to allow investigation of many CREs with a reduced number of expensive transformation events.

3.4.2 – Identification and exclusion of downstream targets of *VRT-A2*

VRT-A2 has previously been shown to play a major role in low basal spikelet productivity, and, consequently, in RBS formation (Backhaus et al., 2022). As a result, basal spikelet productivity might be increased by downregulating *VRT-A2* expression. However, knocking out *VRT-A2* and/or other *SVP* genes has deleterious pleiotropic effects, including delayed heading time and production of additional spikelets or spikes from the AxM of sub-peduncle nodes (Li et al., 2021). In this study, we aimed to discover candidate regulators of basal spikelet productivity independent of *VRT-A2*, both to expand our knowledge of basal spikelet productivity regulators and in case manipulating downstream genes also produced negative pleiotropies.

We attempted to do this by comparing the semi-spatial transcriptomic time courses of two NILs with different *VRT-A2* alleles; $P1^{WT}$ found in most wheat varieties and $P1^{POL}$ found in *Triticum turgidum* spp. *polonicum*. The latter produces ectopic expression of *VRT-A2* in the wheat spike, so we hypothesised that genes active in the spike which are regulated directly or indirectly by *VRT-A2* would show differential expression patterns between the two NILs. 969 genes were DE between the two time courses. Of these, 237 were also DE between

central and basal spike sections at EDR in *P1*^{WT}, so were removed from consideration as candidates for transgenic manipulation.

Given the multitude of roles *VRT-A2* plays throughout floral transition and development in wheat, these candidate targets may themselves prove important regulators of spike architecture. Future studies should further investigate this gene set both bioinformatically – mining for known interactors, interesting domains, and mutant phenotypes in other crops – and experimentally, including through the production of novel wheat mutants.

Interestingly, recent work has revealed that knocking out an upstream negative regulator of *VRT-A2* called *MULTI-FLORET SPIKELET 1* (*MFS1*; unrelated to *MOF1/MFS2*), moderately boosts spike *VRT-A2* expression across a range of developmental stages (0.5 cm spikes to 4.5 cm) (J. Liu et al., 2025). This resulted in a separation-of-function genotype in which the lengths of glumes and grains were boosted, as in *P1*^{POL} plants, but RBS number was not elevated. While this does not directly provide a mechanism by which to raise basal spikelet productivity agronomically, it does suggest that the pleiotropic effects of *VRT-A2* could also be decoupled by other interventions. Mutagenesis or misexpression of the downstream targets of *VRT-A2* we have identified could be explored for such effects, as different targets may execute *VRT-A2*'s different functions.

3.4.3 – *MOF1* and *SEP1-6* can be manipulated to influence spike traits

To date, we have only grown and phenotyped a single CER trial of T_1 *BSS1:MOF-B1* and *BSS2:SEP1-B6* plants and we fully appreciate the need to conduct further trials in similar conditions, glasshouses, and the field to assess the reproducibility of our results. Additionally, rather than comparing our positive transgenic plants to untransformed WT plants, it would be preferable to use either null transformant lines or sibling T_1 plants in which the construct was lost by segregation. Nonetheless, these preliminary results suggest that our approach of misexpressing candidate regulators of the SM-to-FM transition can alter spike architecture.

BSS1::*MOF-B1* plants exhibited, on average, three fewer spikelets than WT plants in their main spikes. Additionally, RBS number trended slightly higher in these lines versus WT, though the difference was marginally insignificant. Further work is needed to confirm whether these results are replicable and to explore additional traits such as floral organ dimensions, grain number and mass, tiller productivity, and per-plant yields. In particular, other metrics for basal spikelet productivity in addition to RBS number should be explored, such as total productivity of the basal most three, four, or five spikelets. We note that the

BSS1 regulatory environment failed to drive expression of tdTomato (assayed by confocal microscopy, and PCR) and GUS (assayed by histochemical staining) reporter genes, so the observed phenotypes may not be caused by misexpression of *MOF-B1*. Instead, they may be a result of epigenetic changes arising from the transformation and regeneration process. Comparison against additional control lines could resolve this ambiguity.

In contrast to the generally deleterious traits observed in BSS1::*MOF-B1* lines, those observed in BSS2::*SEP1-B6* lines were more mixed. Main spike peduncle length decreased by over 50 mm, while spike length increased modestly by 11 mm. However, this was not associated with a concomitant decrease in overall plant height, despite a weak trend for reduced height in the transgenics (-25 mm mean difference, $p = 0.20$). Though not measured, the transgenic lines may exhibit compensatory increases in sub-peduncle internode length. Future work should explore whether the different partitioning of plant height in WT and BSS2::*SEP1-B6* may alter associated traits such as harvest index and lodging resistance. One caveat is that the effects on spike and peduncle lengths did not differ amongst the four lines tested despite their varying construct copy-numbers (ranging from 3 to 10+). This may indicate that these effects were an artefact of transformation rather than a direct result of transgene misexpression, given that the latter might be expected to scale with dosage. Still, CN is not always proportional to transcription, so qPCR analyses would be needed to confirm the relative dosage of basal *SEP1-B6* produced in each line.

BSS2::*SEP1-B6* lines also showed a strong reduction in main spike RBS number versus WT in agreement with our original hypothesis. This may indicate an acceleration of basal spikelet development in agreement with the proposed role for *SEP1-6* in promoting the SM-to-FM transition. However, preliminary visual inspections of spikes between the EDR and CE stages did not reveal any clear differences in spike architecture. Further work will be needed to understand the mechanism by which RBS is reduced, including microdissection at later stages up to anthesis. Subtler phenotypes could be identified using quantitative morphogenic analysis, for example via MorphoGraphX software.

In this trial, majority of transgenic plants exhibited zero RBS (64%), versus a single WT plant (8%). Given that RBS cannot be reduced lower than zero, the conditions tested might, therefore, not allow the full effect on this trait to be observed. This could be remedied in future experiments by utilising poorer growing conditions such as sub-optimal lighting/temperature or reduced nutrient availability (e.g. via reduced fertiliser application, smaller pots for CER or glasshouse trials, or high sowing densities in field trials). Additionally, further traits should be measured, including per-plant productivity and other metrics of basal spikelet productivity. Again, though, this effect also did not differ amongst

transgenic genotypes, raising the same caveat described above for spike and peduncle lengths.

Lastly, BSS2::*SEP1-B6* lines displayed strong reductions in fertile tiller number (12.7 versus 19.4 for WT). Unlike the previous traits, this effect did differ significantly amongst the transgenic genotypes, but there was still no correlation with construct copy number. We were concerned that the reduction in fertile tiller number might be driving these lines' lower RBS number by allowing greater soil resources to be allocated to each spike, much like in the experiments described by Tamago et al. (2024). We therefore calculated Pearson's and Spearman's correlation coefficients between fertile tiller number and RBS number both with and without inclusion of the WT plants. With the inclusion of WT plants there were strong correlations between these traits. However, these weakened and became non-significant when looking at the transgenics alone, suggesting that RBS number was more strongly influenced by construct presence/absence than by fertile tiller number itself. To test this more directly, we could conduct a tiller removal experiment to equalise the number of flowering tillers across genotypes.

If these effects are independent, it is worth noting that while potentially deleterious, the observed reduction in fertile tiller numbers may not be relevant under typical agronomic conditions where, generally, only one to two tillers set grain (Tamagno et al., 2024). This again suggests the need to test these lines under more stressful conditions than those in our CER trial, particularly through the use of agronomically relevant planting densities in field settings. We hypothesise that if RBS number is still reduced under conditions where WT and transgenic lines produce similar, low numbers of fertile tillers, this could produce a measurable yield increase.

Future work should also assay the reciprocal transgenics BSS1::*SEP1-B6* and BSS2::*MOF-B1*. The latter would be of particular interest given the strong, specific expression of tdTomato transcripts and protein that this regulatory environment conferred and we hypothesise would produce different or additional phenotypes compared with BSS1::*MOF-B1*.

We also explored the effects of knocking out *MOF1* in tetraploid wheat. *mof1* mutants showed no difference in spike architecture traits compared with *MOF1* plants of the same generation, but did display alterations in floral organ morphology. This was particularly clear using model [3], which showed small, though significant, differences in glume width, glume length, lemma width, lemma length, palea width, and grain length. There were generally few interactions between genotype and either floret or spikelet position, leading us to drop all

interaction terms for most organ size parameters. The few scattered interacting combinations showed little structure, suggesting that while they might be statistically significant, they may not be biologically relevant. If, on the other hand, a spatial group of spikelets or florets showed interactions with genotype, perhaps along with a gradated effect size, we would consider this of greater interest. Grain width came closest to this, with the central spikelets 6, 7, and 10 all interacting significantly with genotype. However, spikelet 2 also produced a significant interaction, disrupting this pattern. It is also worth noting that in model [3], there was no significant effect of genotype sans interactions on grain width. In the future, it would be interesting to cross *mof1* and $P1^{POL}$ lines to see if their observed effects on floral organ sizes are additive.

Overall, we have identified putative mediators of low basal spikelet productivity in addition to the well-characterised gene *VRT-A2*. We have manipulated two of these genes, both transgenically and via TILLING mutations, and have begun to phenotypically characterise the resulting lines. Our initial results suggest that we may be able to rescue RBS number by basal misexpression of *SEP1-B6*, though this requires replication under the same and alternative conditions.

3.4.4 – Expanding wheat's promoter toolkit will facilitate more effective transgenic manipulation

Despite the precautions taken to achieve spatiotemporal specificity, we still observed that BSS2 led to expression of tdTomato protein in non-target tissues and timepoints; the incipient peduncle, upper stem, and the bases of leaf primordia showed strong concentrations of nuclear tdTomato protein from the vegetative stage through to W9. This highlights the limited spatiotemporal resolution in the RNA-seq data we utilised to select regulatory environment sequences.

The previously mentioned MERFISH data later confirmed that TraesCS2B02G399800, from which BSS2 was derived, is indeed expressed in these additional tissues at all timepoints assayed (W2.5, W3.25, W4, and W5; Figure 3.19; Long et al., 2024). In fact, expression appeared to be stronger in the bases of leaf primordia than in the basal spike. While spatially distinct, it appears that the expression of *MND-B1* (used for BSS1 elements) is also strongest just below the spike at these timepoints rather than in the basal spikelets, becoming highly specific to the peduncle at W5 (Figure 3.19).

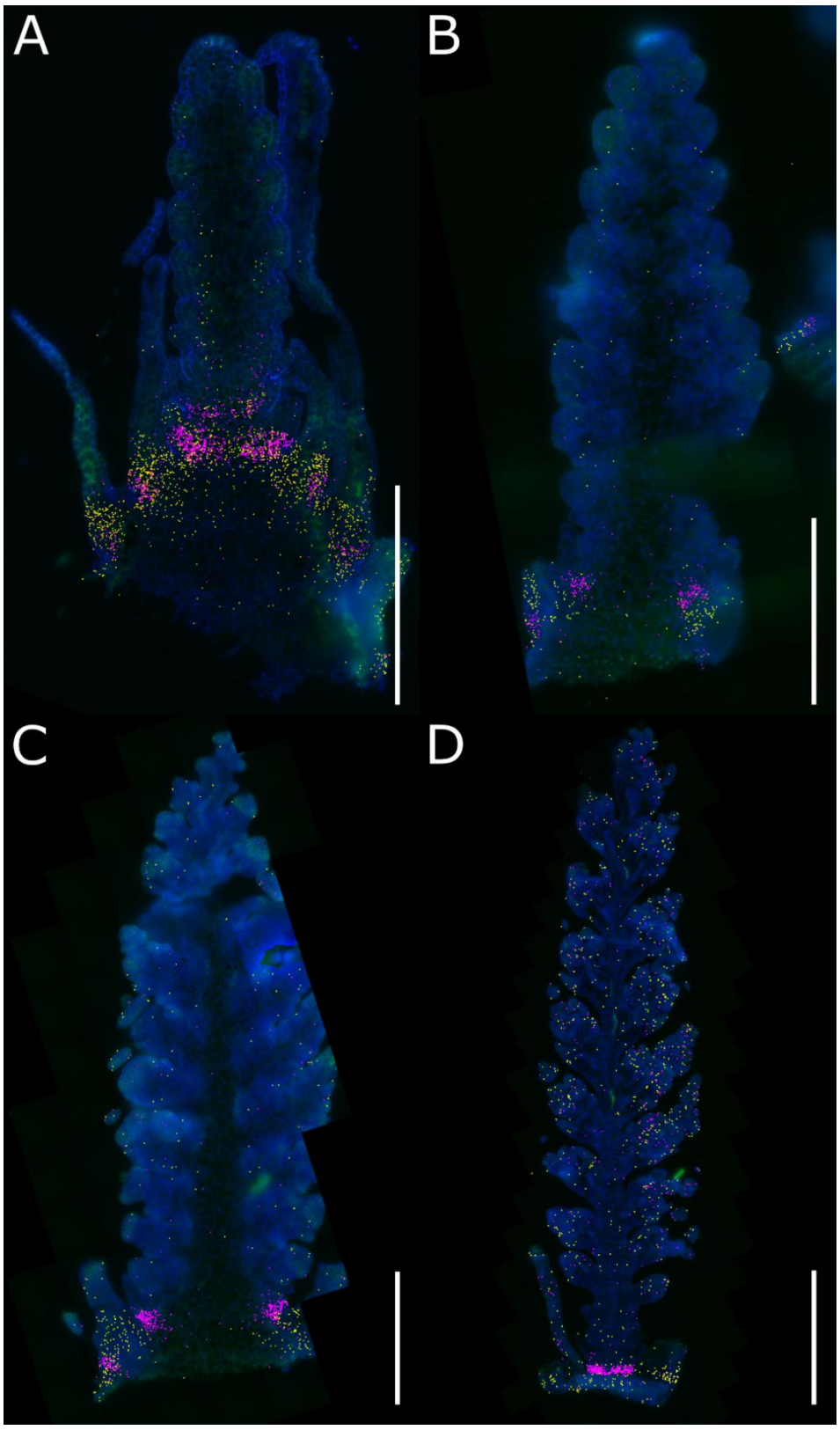


Figure 3.19 – MERFISH data on TraesCS2B02G399800 (used for BSS2 elements) transcripts matches observations of BSS2::tdTomato expression pattern
10 µm thick cryosections of WT bread wheat spikes (cv. Paragon) hybridised with MERFISH probes. **A,** Late double ridge stage. **B,** lemma primordia stage. **C,** Terminal spikelet stage **D,** Carpel extension stage. Blue channel = DAPI, green channel = PolyT. Pink dots (*MND-B1* + homoeologs, used for BSS1) and yellow dots (TraesCS2B02G399800 + homoeologs, used for BSS2) represent decoded transcript assignments. Data is from (Long et al., 2024) and visualised using VizGen MERSCOPE® Visualizer software. Scale bars represent 500 µm in **A-C** and 1000 µm in **D**.

This is also observed in *VRT-A2* (Li et al., 2021; Long et al., 2024) and perhaps indicates that the absence of precise expression boundaries between the basal spike and incipient peduncle is common to many genes, which in turn relates to the observed gradients of leaf ridge outgrowth and spikelet productivity.

These findings suggest that our attempts to misexpress MOF-B1 and SEP1-B6 were less precise than intended, potentially introducing additional pleiotropic effects and complicating interpretation of our results. Their real expression patterns could be ascertained at a later date by *in situ* hybridisation or single gene FISH. These assays could potentially be made specific to the introduced transgenic copies given our inclusion of C-terminal 3x FLAG tags. While our transgenic lines have produced phenotypes of interest and are still a valuable resource, it would be useful to develop promoters or regulatory environments that are more precise to the basal spikelets for future explorations of basal spikelet productivity.

This raises a general issue for the study of developmental processes in wheat. Unlike model organisms such as Arabidopsis (Marquès-Bueno et al., 2016; Schürholz et al., 2018; Yaschenko et al., 2024) and *Marchantia polymorpha* (Sauret-Güeto et al., 2020; Romani et al., 2024) and, to some extent, rice (Singha et al., 2022) and maize (Liu, 2009; Yassitepe et al., 2021), wheat still lacks a diverse toolkit of well-characterised regulatory environments for tissue and cell-type specific expression. (We stress 'regulatory environment' rather than just 'promoter' because the inclusion of introns, signal peptides, and distal sequences is crucial to achieving strong, specific expression patterns.)

There has been progress in developing regulatory environments for certain tissues (phloem, endosperm and other grain tissues) or in response to some stressors (cold, dehydration, wounding, fungal pathogens), largely thanks to promoters co-opted from other monocotyledonous species (Hensel et al., 2011b). Overall, though, our capacity to achieve spatiotemporally specific expression in wheat remains limited, and the majority of wheat transgenic experiments to date have utilised strong constitutive elements such as maize *UBIQUITIN-1* or rice *ACTIN1*, both of which are promoter-intron fusions (Hensel et al., 2011b).

The value of such resources has been amply demonstrated over multiple decades in both model and crop species. For example, the C_4 Rice Project, which aims to engineer C_4 photosynthesis in rice, has conducted a number of impressive manipulations of leaf cell ultrastructure using tissue and cell-type specific promoters (P. Wang et al., 2017; Lee et al., 2021). For instance, consortium researchers have developed a bundle sheath-specific

promoter (Hua et al., 2024) and have since used this to manipulate brassinosteroid signalling via the transcription factor *BRASSINAZOLE RESISTANT 1* (*OsBZR1*) (Cackett et al., 2024). Constitutive overexpression of this gene led to a desirable increase in bundle sheath chloroplast area, but also caused rapid leaf senescence and reduced grain set. In contrast, bundle sheath-specific expression of *OsBZR1* produced only the positive chloroplast area effect.

This example and many others highlight the role an enhanced wheat toolkit could play in accelerating the exploration of developmental processes and facilitating the generation of novel, agronomically useful traits. However, given the high cost of producing transgenic wheat lines, efficient strategies for screening novel regulatory environments are essential. The use of cell-type specific transcriptomes has proved beneficial in rice and elsewhere to identify gene candidates for regulatory element mining (Hua et al., 2021). These can be produced either by laser-capture microdissection (LCM) or through single-cell transcriptomics combined with judicious use of cell-type marker genes. Existing LCM-derived transcriptomes for the vegetative and spikelet ridges of barley could be a starting point for producing novel promoters specific to particular cell populations of the Triticeae spike (Thiel et al., 2021). Spatial transcriptomics may also facilitate such candidate selections in the future (Long et al., 2024). Additionally, the use of transient transformation of protoplasts derived from specific tissues could reduce some of the need to produce costly stable transgenics, as it has been demonstrated that protoplasts retain their tissue-specific gene expression patterns to some extent (Diaz et al., 1995; Faraco et al., 2011).

However, even once a candidate gene with a desirable expression pattern is identified, researchers must still identify the CREs necessary and sufficient to recapitulate that pattern. As demonstrated in this chapter, one potential avenue for achieving this is the examination of tissue-specific patterns of chromatin accessibility proximal to the gene of interest. My next chapter explores this approach in more detail, and evaluates its potential for generating a genome-wide atlas of spike-relevant CREs for wheat.

3.5 – Methods

3.5.1 – Semi-spatial RNA-seq

Paragon NILs containing either the *VRT-A2a* (*P1*^{WT}) or the *VRT-A2b* (*P1*^{POL}) allele were grown in a controlled environment chamber under a long day photoperiod (16 / 8 h; 20 /15 °C; light / dark) at 300 µmol m^{-2} s^{-1} incident radiation and 60% humidity. The first three stages (W2, W2.5, W3.25; Figure 3.1) were collected from plants in ~75 mL cells. Later stages were collected from plants potted on after ~21 days into 1 L pots. The media used was 'John Innes Cereal Mix' (65% peat, 25% loam, 10% grit, 3 kg m-3 dolomitic limestone, 1.3 kg m−3 Yara PG MixTM 14-16-18, 3 kg m-3 Osmocote® Exact).

Tissue for RNA-seq was collected as described previously (Faci, Backhaus, et al., 2024). Developing spikes were collected from the main tillers only using a dissecting microscope and ophthalmic microsurgery knives. Equipment and gloves were frequently sanitised using Blitz RNase removal spray (Severn Biotech Ltd.) and always between dissecting different genotypes or stages. For the early double ridge stage (W2), spikes were bisected to produce basal and central/apical sections (Figure 3.1). For the late double ridge (W2.5), lemma primordia (W3.25), terminal spikelet (W4) and carpel extension (W5) stages, the basal section consisted of the most basal four spikelets from each spike. Two spikelets were skipped, then the subsequent four spikelets were harvested to comprise the central section (Figure 3.1). To provide sufficient tissue for RNA extraction, different numbers of spikes were dissected for each biological replicate at each stage: W2 = 30, W2.5 = 20, W3.25 = 10, W4 = 5, W5 = 2. Four replicates were collected for each stage-section-genotype combination. Samples were collected into tubes on dry ice, then flash frozen in liquid nitrogen and stored at -70 °C.

Tissue was homogenised by bead beating using a Genogrinder then total RNA was extracted using RNeasy Plant Mini (QIAGEN) and Direct-zol RNA Microprep (Zymo Research) kits as described in the manufacturers' manuals. Total RNA (1 µg) was sent to Novogene UK for PCR-free library preparation and Illumina sequencing (PE150; 50M reads per sample). RNA-seq data from *P1*^{WT} was deposited in BioProject PRJNA1201104 on the NCBI SRA as part of another publication (Long et al., 2024).

3.5.2 – Bioinformatic analyses

RNA-seq reads were trimmed using cutadapt (v1.9.1; Martin, 2011) and quality assessment was conducted before and after trimming using FastQC (v0.11.8; Andrews, 2010). Reads

were pseudomapped using Kallisto (v0.44.0; Bray et al., 2016), supplying a k=31 index of the IWGSC RefSeq v1.1 gene annotations (IWGSC, 2018) and setting '--bootstrap-samples=30'. TPM and read count values were summed across annotated alternate transcripts to produce a single figure per gene. Each PCA was conducted using only HC genes with non-zero counts in at least one target sample. Read counts were transformed using the rlog transformation from DESeq2 (v1.34.0; Love et al., 2014) to prevent highly expressed genes from having a disproportionate effect and PCs calculated using the base R function prcomp (R Core Team, 2023).

Gene trajectories were compared using ImpulseDE2 in "case-control" mode (v 3.6.1; Fischer et al., 2018) and the default Benjamini-Hochberg correction was utilised. Pairwise comparisons were made using DESeq2 and adjusted p-values were calculated using independent hypothesis weighting (IHW R package v1.22.0; Ignatiadis et al., 2016), using the DESeq2 baseMean variable (mean normalised count values) as a covariate. The complete set of FunRiceGenes annotations was downloaded on 08/02/2023. Developmental time course expression data for the cultivar Azhurnaya was downloaded from the Grassroots Data Repository (Borrill et al., 2016; Bian et al., 2017). ATAC-seq data was processed as previously discussed and visualised using the Interactive Genome Browser (IGV) (Nassar et al., 2023).

3.5.3 – Construct assembly and transformation

Constructs were assembled following the plant Golden Gate assembly standard (Engler et al., 2008; Engler et al., 2014). See Figure 2 of (Engler et al., 2014) for correspondence between part names and sticky-end overhangs. Level 0 parts for *MOF-B1* (CDS1 ns), *SEP1-B6* (CDS1 ns), BSS1-5PU (Pro + 5U), BSS1-3D (non-standard 3U*), BSS2-5PU (Pro + 5U), and BSS2-3D (non-standard 3U*) were synthesised by Twist Bioscience. These were each domesticated for internal BsaI and BbsI sites. Additionally, one SNP was introduced to BSS1-P5U (7B:681,858,347 A→G) to disrupt a 23 bp repeat which was predicted to interfere with synthesis. Each of these was cloned into the in-house 'pTwist-Kan-HC' high copy number backbone, except for BSS2-5PU which was cloned into the 'pTwist-Kan-MC' medium copy number backbone due to poor *E. coli* growth with the previous backbone.

Eight Level 1 constructs were assembled as described in Table 3.2, using the above parts and the additional L0 parts pICSL62001 (Nos terminator; non-standard secondary terminator), pICH75111 (GUS with two introns; CDS1; Addgene #50327), pICSL50007 (3xFLAG C-terminal tag; CT; Addgene #50308), fp08009 (tdTomato; CDS2; Faulkner Lab JIC), and fp08064 (NLS; NT2; Faulkner Lab JIC). The L1 backbone was pICH47742

(Addgene #48001). Eight Level 2 binary vectors were assembled by cloning these Level 1 parts into the P2 position of pGoldenGreenGate-M (Addgene #165422; Smedley et al., 2021). A Level 1 hpt hygromycin resistance cassette (Addgene #165423; Smedley et al., 2021) was cloned into P1 and the constructs were sealed using a P2 end-linker (pICH41744; Addgene #48017). Annotated sequences for custom L0 parts and L2 constructs are provided in Supp. Data 3.6 and are available to order on Addgene (https://www.addgene.org/Cristobal_Uauy/). For conversions between the codes used in this thesis (e.g. MJ_GG19) and published Addgene identifiers, see Table 3.3.

Constructs were transformed into the hexaploid wheat cultivar Fielder without the use of the *GRF-GIF* system (Hayta et al., 2019) by the JIC Wheat Transformation Team.

3.5.4 – Phenotyping of transgenic reporter lines

For tdTomato PCR experiments, wheat plants were grown and whole spike meristems collected as described above, except plants were not potted on. For LP/FP samples six spikes were pooled, while for EDR/LDR/GP samples 24 spikes were pooled. *Agrobacterium tumefaciens* containing the construct fp08024 (Caldas et al., 2022) was infiltrated into two leaves of three *Nicotiana benthamiana* plants as previously described (Caldas et al., 2022) and leaf discs weighing approximately 15 mg were taken using a #3 cork borer. RNA extractions were conducted using Direct-zol RNA Microprep kits (Zymo Research) as described in the manufacturer's manual except that an additional centrifugation (13,000 x g, 1 min) was conducted after the final wash buffer step to remove excess ethanol. RNA quality was assessed by Nanodrop (Thermo Fisher Scientific) and by running 400 ng total RNA on an agarose gel. cDNA was synthesised using 1 mg total RNA per sample via QuantiTect Reverse Transcription kits (QIAGEN). Previously designed GAPDH primers (Harrington, 2019) were used for detection of residual gDNA (Supp. Data 3.2). Custom tdTomato primers were designed using Primer3 (Koressaar & Remm, 2007) (Supp. Data 3.2). All gels were 1% agarose in TAE buffer.

Confocal imaging of WT and tdTomato lines was conducted using a Zeiss LSM 880. Microscope slides (1.0 mm; VWR® catalogue number 631-1552) were prepared for use by creating small rectangular 'wells' of approximately 50 x 15 mm using strips of double-sided sticky tape then adding ~250 µL tap water. Spikes were uncovered from leaf sheaths using ophthalmic microsurgery knives, bisected using a razor blade fragment, deposited onto a slide, then covered with a thickness 1.5 borosilicate cover slip (VWR® catalogue number 631-0138). Leaf, root, grain, and spikelet/floral organ tissues from WT and tdTomato lines were similarly prepared and imaged.

Table 3.3 – Conversions between thesis construct codes and Addgene identifiers

Level	Construct code	Addgene name
L0	MOF-B1	pTwist-Kan-HC-L0-TaMOFB1(int)
L0	SEP1-B6	pUC-GW-Kan-L0-TaSEP1B6(int)
L0	BSS1-P5U	pTwist-Kan-HC-L0-TaBSS1-5PU
L0	BSS1-3U	pTwist-Kan-HC-L0-TaBSS1-3D
L0	BSS2-P5U	pTwist-Kan-MC-L0-TaBSS2-5PU
L0	BSS2-3U	pTwist-Kan-HC-L0-TaBSS2-3D
L2	MJ_GG18	pGGG-TaBSS1-TaMOFB1(int)
L2	MJ_GG20	pGGG-TaBSS1-TaSEP1B6(int)
L2	MJ_GG22	pGGG-TaBSS1-GUS(int)
L2	MJ_GG24	pGGG-TaBSS1-tdTom
L2	MJ_GG17	pGGG-TaBSS2-TaMOFB1(int)
L2	MJ_GG19	pGGG-TaBSS2-TaSEP1B6(int)
L2	MJ_GG21	pGGG-TaBSS2-GUS(int)
L2	MJ_GG23	pGGG-TaBSS2-tdTom

A leaf section from one of the previously described *N. benthamiana* plants was imaged as a positive control for familiarisation with the expected appearance of nuclear foci (Zeiss LSM 800). For all samples, a 633 nm laser was used to image chlorophyll autofluorescence (647-721 nm filters) and a 561 nm laser was used to image tdTomato fluorescence (561-614 nm filters), with the latter also generating the transmitted light (T-PMT) channel. Microscope settings including laser power, master gain, and pinhole size, were kept constant for comparisons between genotypes. For specific settings see Table 3.4.

To analyse BSS1::GUS_{in} lines, wheat spikes were collected as above and then stained and cleared as previously described (Hayta et al., 2021). WT Fielder was used as a negative control, while a Fielder line containing a construct for constitutive GUS expression (Addgene #165418; Hayta et al., 2021) was used as a positive control.

3.5.5 – Phenotyping of transgenic developmental misexpression lines

BSS1::*MOF-B1* and BSS2::*SEP1-B6* plants were grown and potted on as described above. Plants were assessed for construct copy number (CN) by the JIC Genotyping platform using qPCR against the hpt hygromycin resistance gene. This CN detection methodology has poor resolution for higher CN values, so any CN above 10 were set to 10.

Plant height was measured from the soil surface to the tip of the terminal spikelet of the tallest tiller. Main spike length was measured from the end of the peduncle to the tip of the terminal spikelet. Awns were excluded from these two measurements. We included the main spike in fertile tiller number counts and defined fertile tillers as those bearing one or more filled grains. Spikelet number counts included RBS and sterile apical spikelets. RBS and sterile apical spikelets were defined as basal and apical spikelets, respectively, containing zero filled grains.

3.5.6 – *mof1* mutant generation and phenotyping

Tetraploid *T. turgidum* (cv. Kronos) TILLING lines Kronos2630 and Kronos4375 were obtained from the John Innes Centre Germplasm Resources Unit (GRU). KASP markers for the mutations of interest were designed using Primer3 (Koressaar & Remm, 2007) and Benchling (Supp. Data 3.2). Plants homozygous for these mutations were crossed and then the F_1 generation was self-pollinated. As an aside, of the parental generation, only one of six Kronos2630 plants headed at the same time as the Kronos4375 plants, with the remainder heading many weeks later.

Table 3.4 – Acquisition parameters for confocal microscopy

Micrograph description	Figure panels	Microscope	Magnification	Pinhole (µm)	633 nm laser line attenuator transmission (%)	633 nm master gain	561 nm laser line attenuator transmission (%)	561 nm master gain	T-PMT master gain
***N. benthamiana* leaf**	Figure 3.8A	Zeiss LSM 800	10x	33	1.1	750	2.2	840	240
Wheat spike	Figure 3.8B-S Figure 3.9A-C	Zeiss LSM 880	10x	32.8	0.8	700	1.0	700	430
Wheat spike (higher sensitivity)	Figure 3.9D				0.8	700	1.8	800	380
Wheat leaf	Figure 3.10A,E				0.4	700	1.5	700	430
Wheat root	Figure 3.10B,F				1.5	700	1.5	700	380
Wheat endosperm, seed coat, glume, lemma body, lemma awn, anther, carpel	Figure 3.10C-D Figure 3.10G-M				0.8	700	1.0	700	430

This may be due to a stop gain mutation in VRN-A1 which is also present in this TILLING line (Kronos2630.chr5A.587412769). We speculate that this could have influenced the vernalisation requirements of this spring wheat cultivar.

The F_2 generation was grown in a greenhouse from 19/02/2024 using the same soil and potting strategy described above. Supplementary lighting and heat was provided to achieve day/night temperatures ~20/15 °C and a photoperiod of 16/8 h. Plants were genotyped using the same KASP markers, then allowed to self and set grain. Additional leaf tissue samples were collected from F_2 double WT plants (n = 3) and double *mof1* mutants (n = 5) for DNA extraction and Sanger sequencing (GENEWIZ®), which confirmed the validity of our KASP genotyping. To assess the effects of the *MOF1* splice site mutations, spike tissue (EDR stage, 20 spikes) was collected from the F_3 generation of self-pollinated F_2 double WT and double *mof1* plants. Three biological replicates were prepared per genotype. cDNA was prepared as described above and used to develop reverse transcription PCR assays for the A and B homoeologs of *MOF1*/*mof1* (Supp. Data 3.2).

Gross morphology traits were measured as described above for transgenic misexpression plants. We adapted a previously published methodology to measure the sizes of spikelet organs in *mof1* and *MOF1* by arranging dissected organs on PCR plate films (Adamski et al., 2021; Supp. Data 3.7). We included datapoints from spikelets 1 (basal) to 14 (apical) and excluded datapoints from spikelets 15 and 16 as these contributed only 39 organ measurements out of a total of 3588 (1.1%) and were not dramatically overrepresented in one genotype over the other. Similarly, we excluded one lemma datapoint and two palea datapoints from 5th florets as these comprised a small fraction of total measurements (0.08%).

We then compared the widths and lengths of glumes, lemmas, paleas, and grains from *mof1* and *MOF1* using three levels of ANOVA model:

[1] $$y \sim genotype + spikelet + spikelet{:}\,floret + genotype * spikelet + genotype(spikelet{:}\,floret)$$

[2] $$y \sim genotype + spikelet + spikelet{:}\,floret + genotype * spikelet$$

[3] $$y \sim genotype + spikelet + spikelet{:}\,floret$$

The experimental units were individual plants in pots distributed randomly on a single glasshouse bench, so no block term was included in our ANOVAs. For each organ-parameter combination (e.g. palea width), we tested whether model [1] identified significant interactions between genotype and floret ID (nested within spikelet). In all cases

there were very few interactions (Supp. Data 3.4), leading us to eliminate an interaction term and test model [2]. Again, there were mostly few interactions, except for grain length which had four significant genotype-spikelet combinations. Apart from grain length, we then tested each combination using model [3] to see if there was a significant effect of genotype aggregated across all spikelets and organs.

4 – ATAC-seq for genome-wide discovery of spike-relevant *cis*-regulatory elements in wheat

4.1 – Chapter Summary

Genome-wide characterisation of the *cis*-regulatory element (CRE)-gene pairs active during early wheat spike growth would facilitate experimental investigation of this developmental process by allowing more precise transgenic and GE-based manipulation of target genes. Candidate regions for active CREs can be detected through chromatin accessibility assays and then corroborated and/or linked to target genes by additional methods. Here, we trialled an approach for discovering CRE-gene pairs based on ATAC-seq data to detect chromatin accessibility, profiling of stably unmethylated regions to support candidate CREs, and RNA-seq data to examine how gene expression varies with CRE chromatin accessibility. We produced our own ATAC-seq data across 16 stages of wheat spike and carpel development, but, ultimately, the data displayed low signal-to-noise ratios and was unsuitable for our intended analyses. Instead, we reanalysed an independent dataset that was produced concurrently by a collaborating research group. We found that differentially expressed genes were enriched for differentially accessible CRE candidates in close sequence proximity, and that this enrichment was greater when using CRE candidates supported by unmethylated region data. Lastly, we showed a modest ability to detect putative enhancer-gene relationships, but no capacity to detect silencer-gene pairs.

I was supported in this work by Isabel Faci (JIC) and Neil McKenzie (JIC). Isabel developed our nuclei extraction protocol and we jointly carried out the microdissections, RNA extractions, and nuclei extractions described here. Neil developed our ATAC-seq tagmentation protocol and carried out tagmentation of all samples. I am also grateful to Dr Peter Crisp (University of Queensland, Australia) for producing UMR calls from wheat WGBS data and Dr Tahrang Mehta (Earlham Institute) for providing advice on ATAC-seq data analysis. Plant growth was facilitated by JIC Horticultural Services and nuclei imaging by the JIC Bioimaging platform. Novogene UK provided prompt, high-quality sequencing services.

4.2 – Introduction

In the previous chapter (3.4.4), we established one application of an improved capacity to identify CREs underlying spatiotemporally specific patterns of gene expression in wheat. We argued that this would facilitate more rational design of regulatory environments for transgenic mis-expression experiments, which in turn would support explorations of wheat developmental processes. Another application is the engineering of target gene expression by mutagenising CREs *in situ* through genome editing (GE) techniques. Such approaches can produce allelic series of variation with subtler phenotypic effects and, consequently, more agronomic utility than complete knock-outs or transgenic overexpressor lines.

This approach was elegantly demonstrated in a series of experiments on *Solanum* species by applying GE to re-create key domestication alleles in wild relative of tomato, *S. pimpinellifolium* (Rodriguez-Leal et al., 2017). For example, the weak gain-of-function *locule number* (*lc*) QTL was phenocopied by disrupting the same CArG repressor element as in the natural allele. This caused upregulation of the transcription factor (TF) *WUSCHEL* (of the classical *CLV-WUS* stem cell circuit), leading to enlarged apical meristems, and, ultimately, increased fruit size. Expanding on this, an allelic series of novel locule number variation was generated in modern tomato, *S. lycopersicum*, by mutagenising the promoter of *SlCLV3* using eight arbitrary gRNAs across 2 kbp. These did not target specific CREs, but still produced novel alleles with phenotypes spanning between and extending beyond the range of trait values achieved through breeding. This demonstrates the potential for GE of non-coding sequences to produce novel germplasm for research and, potentially, breeding. Refining such approaches by targeting specific, rationally selected CREs could further boost their value.

Variations of CRISPR-Cas9 could enable even more precise CRE edits. For example, base editing can precisely substitute one base for another. By combining various deaminases with catalytically impaired nucleases, all four transition mutations can be produced (i.e. C→T, T→C, A→G, G→A) without inducing a double-stranded break and therefore with a very low rate of indel generation (Molla & Yang, 2019). This technology has been applied in plants, including wheat (Zong et al., 2017), and could be used to modify TF binding sites towards or away from a consensus sequence to subtly tweak their transcriptional effects. Base editors with altered or relaxed PAM requirements have been created to expand the range of sites that can be targeted. Another technology, prime editing (PE), expands upon this through a different mechanism. Early PE systems enabled all twelve transition and transversion point mutations to be produced, and, moreover, could precisely introduce

small deletions and insertions of a specific sequence (< 50 bp) (Anzalone et al., 2020; Villiger et al., 2024). Newer PE variants are continually upgrading the range and size of possible manipulations, as well as their efficiency (Villiger et al., 2024; Zeng et al., 2024). While most development to date has occurred in mammalian systems (Y. Zhao et al., 2025), PE provides the prospect of engineering plant gene expression by inserting entirely new TF binding sites, effectively generating novel CREs. This was recently demonstrated in wheat in an allele replacement experiment. The $P1^{WT}$ allele (of *VRT-A2*) in Fielder was replaced with the $P1^{POL}$ allele from *Triticum turgidum* ssp. *Polonicum*. This involved the precise deletion of a 567 bp sequence and insertion of a 157 bp sequence containing five intron-mediated enhancement elements (Y. Zhao et al., 2025).

Given these exciting applications, how can CREs relevant to the gene expression patterns of a particular developmental process be identified in a high-throughput manner? CREs are associated with a range of chromatin properties, though some of these differ between eukaryotic kingdoms. Such chromatin features were initially characterised by studying CREs discovered with early, low-throughput methods, such as reporter constructs, enhancer trapping, electrophoretic mobility shift assays (EMSAs), promoter cut-down analyses, and QTL mapping (Weber et al., 2016). However, now that these features are known, methods have been developed to detect them genome-wide, enabling high-throughput discovery of novel CREs.

A prominent feature of active CREs across eukaryotes is that they are usually situated in nucleosome-depleted, accessible chromatin regions (ACRs). This is thought to be because most TFs require un-occluded DNA to bind their cognate motifs. In animal models, ACRs have been convincingly demonstrated to correlate with CREs at the genome-wide scale and high-throughput ACR detection has been extensively used for CRE discovery (Lee et al., 2004; Ozsolak et al., 2007; Li et al., 2011; Sheffield & Furey, 2012; Thurman et al., 2012; Yue et al., 2014). Such approaches were applied much later in plants, beginning in 2012 with the genome-wide mapping of open chromatin in rice (Zhang, Wu, et al., 2012), though they have since exploded in popularity (Bubb & Deal, 2020). There is also a wealth of functional validation demonstrating that the positions of active CREs strongly overlap with ACRs in plants. For example, one study detected over 10,000 gene-distal ACRs in Arabidopsis, then functionally validated a subset using a β-glucuronidase reporter. 71% (10/14) candidate CREs were found to possess enhancer activity *in planta*, vs 0/10 control regions (Zhu et al., 2015). It is now widely accepted that active CREs primarily reside within ACRs in plant genomes (Z. Lu et al., 2018). However, chromatin is highly dynamic and therefore a CRE that

is accessible and active in one tissue at a given developmental stage or in a given environment may be inaccessible and inactive in another (Sullivan et al., 2014).

Numerous methods have been developed for high-throughput detection of ACRs, most notably DNase-seq, MNase-seq, and ATAC-seq. The general principle of such methods is to expose intact chromatin to an enzyme that marks DNA – e.g. by digestion, ligation, or chemical labelling. However, much like TFs, these enzymes preferentially bind DNA where they are not sterically hindered by nucleosomes – i.e. within ACRs (Minnoye et al., 2021; Mansisidor & Risca, 2022). The digested or tagged DNA can then be sequenced using next-generation technologies, the resultant reads mapped to a reference genome, and 'peaks' (clusters of reads) corresponding to ACRs called. The same data can be used to map nucleosome positioning by examining the complement of the accessible DNA. More sensitive techniques can also detect TF footprints within ACRs.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) has dramatically risen in popularity since it was first described (Buenrostro et al., 2013; Yan et al., 2020). Approximately 400 ATAC-seq datasets were deposited in PubMed in 2019 – roughly four-fold the number from the three next-most popular methods combined (Minnoye et al., 2021). Accordingly, it offers several advantages over techniques like DNase-seq or MNase-seq; notably, bulk ATAC-seq can be performed using orders of magnitude fewer cells, requires less arduous experimental calibration, and employs a much faster protocol (Bajic et al., 2018; Bubb & Deal, 2020; Marinov & Shipony, 2021; Minnoye et al., 2021). Nonetheless, other methods do have their own advantages. For example, DNase-seq is still preferential for detecting TF-footprints within ACRs (Marinov & Shipony, 2021; Minnoye et al., 2021).

The reduced calibration requirements of ATAC-seq can be attributed to its unique mechanism amongst ACR detection techniques. The method employs an engineered hyperactive Tn5 transposase which inserts pre-loaded sequencing adapters directly into accessible chromatin, simultaneously tagging and fragmenting ('tagmenting') the DNA (Candela-Ferre et al., 2024; Figure 4.1). This combination of steps into a single reaction reduces the number of stages requiring adjustment, while the transposition process is generally more robust to non-optimal enzyme concentrations than nucleases, which are prone to over- or under-digesting the DNA (Bajic et al., 2018; Minnoye et al., 2021).

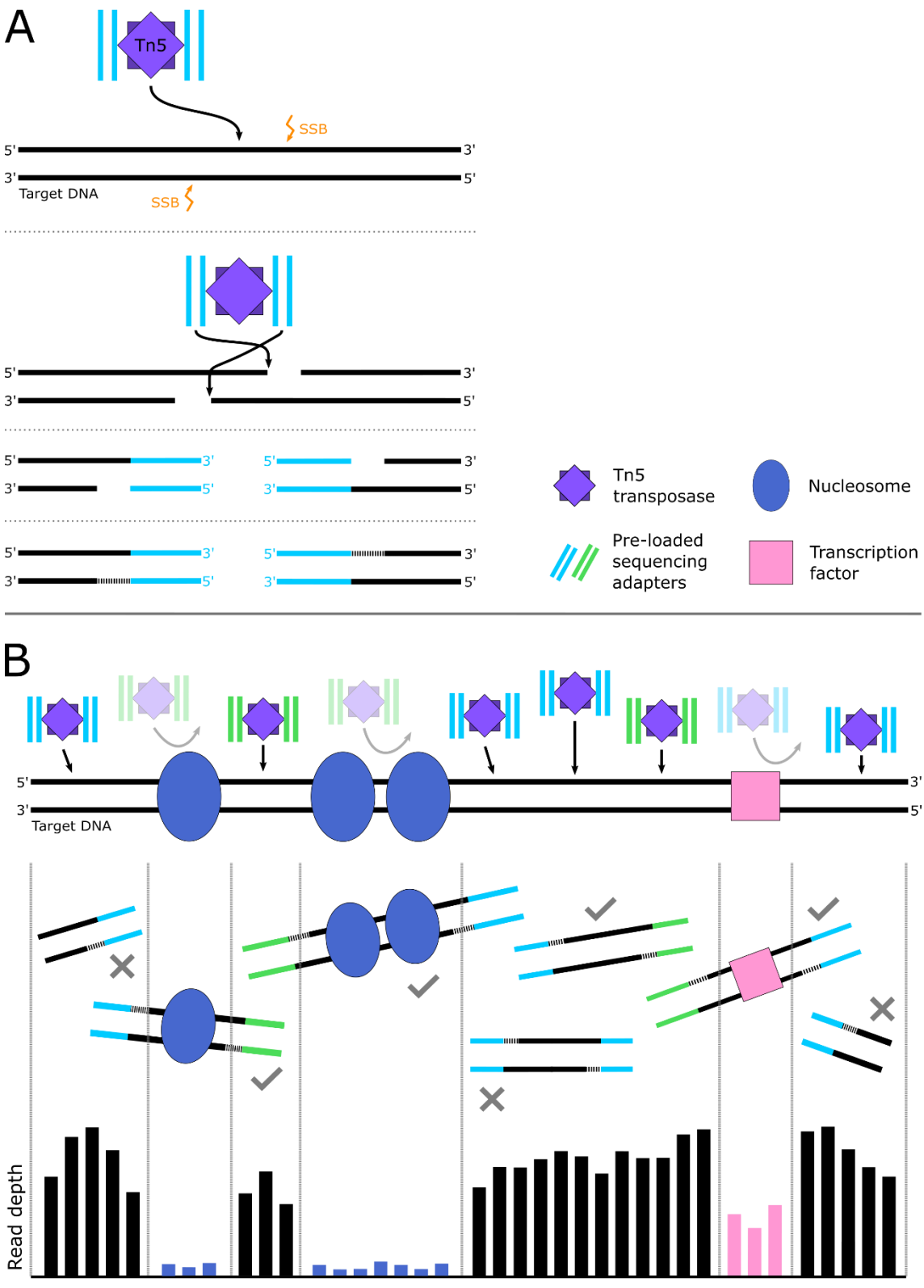


Figure 4.1 – Engineered Tn5 transposase can be used to 'tagment' accessible chromatin
A, Mechanism of a single tagmentation reaction. A Tn5 transposase dimer creates staggered single strand breaks and ligates dsDNA adapters onto the cut ends. A library preparation PCR then resolves the single strand 9 bp gaps while adding indices and flow cell-binding oligos. **B,** Only fragments containing both types of adapter can be amplified. Tn5 preferentially inserts into DNA not bound by nucleosomes or transcription factors. Peaks are called from *in silico* tags centred about the points of insertion – not the centre of read pairs – to preserve this preference.

Despite the strong evidence that most active CREs reside within ACRs, not all ACRs contain CREs. Of those that do, embedded CREs may only span a small fraction of the accessible region (Z. Lu et al., 2018). Fortunately, the accuracy of CRE prediction can be enhanced by coupling ACR profiling with additional methodologies (Weber et al., 2016; Z. Lu et al., 2018). Five such approaches are discussed here.

Firstly, one can identify regions where ACRs and other CRE-associated chromatin features overlap. Enrichment of the histone variant H2A.Z in flanking nucleosomes and depletion of DNA methylation appear to be recurring features of plant CREs (Ricci et al., 2019). This was demonstrated most comprehensively in (Lu et al., 2019)'s *tour de force* which analysed chromatin patterns across 13 plant species spanning eudicotyledons, grasses, and non-grass monocotyledons. Subsequently, it was proposed that stable profiles of DNA hypomethylation (<10% methylation across all cytosine contexts) capture the superset of all CREs accessible in various tissues, developmental stages, and environments (Crisp et al., 2019; Crisp et al., 2020). Thus, whole-genome bisulphite sequencing (WGBS) of, say, leaf tissue can be used to probe for CREs that only become accessible during spike development. This approach was initially validated in five monocotyledonous species, including maize, rice, and barley (Crisp et al., 2020). Unmethylated regions (UMRs) could be used to corroborate ACR calls where these regions intersect, or this method could be used as an initial insight for tissue-timepoint combinations which are difficult to conduct chromatin accessibility profiling on.

Histone modifications, such as acetylation or methylation of specific amino acids, have also been proposed as signatures for plant CREs. However, unlike in animals, a reliable histone modification code has yet to be established in plants. Amongst other inconsistencies, CRE histone modifications seem to vary with their distance from their target TSS(s) (Weber et al., 2016; Marand et al., 2017; Lu et al., 2019; Ricci et al., 2019). Despite this lack of uniformity, researchers have suggested that CREs can be detected by identifying coordinated histone modification patterns between CRE-gene pairs. In wheat, the densities of activating (H3K4me3, H3K9ac, H3K27ac) and repressive (H3K27me3) histone modifications were shown to be highly similar between CREs and their cognate genes (Wang et al., 2021). This method was validated using Hi-C data (discussed below) and was shown to perform better than the traditional approach of assigning CREs to genes based on sequence proximity. Alongside Hi-C, this is one of the only CRE detection methods that directly links CREs with the genes they regulate. At present, the relative merits of this method versus Hi-C are unclear, though there is some suggestion that this approach better tracks the current transcriptional activity of genes. An important consideration will be the

relative quantities and qualities of tissue required for the respective methods, as this may promote the use of one method over the other for answering certain biological questions.

Secondly, ACRs can be analysed for known TF-binding motifs. While TF-binding motifs do intuitively seem a good predictor of CREs, there are some drawbacks to this approach. The presence of a motif does not confirm that the cognate TF(s) binds at that location *in vivo* or has functional significance. Binding motifs are very short, typically less than 10 bp in length, meaning they can arise by chance. Prioritising clusters of putative motifs can reduce false discovery rates as these are less likely to occur stochastically. Additionally, not all TFs exhibit strongly sequence-specific binding, making binding sites hard to predict from sequence alone. Motif scanning is also limited by the relatively small number of motifs confirmed (via reporter assays, mobility shift assays, ChIP, or DAP-seq) in non-model plants (Weber et al., 2016). An alternative approach is to apply *de novo* motif prediction tools originally developed for use on ChIP-seq data, such as MEME (Bailey et al., 2006) or DREME (Bailey, 2011), to define recurring motifs within ACRs as these are also good indicators of functional CREs (Maher et al., 2018; Han et al., 2020; Parvathaneni et al., 2020; Marand et al., 2021). Other features of functional significance can also be used to pinpoint CREs within a set of ACRs, such as conservation of non-coding sequences between species, which can be determined using tools such as BLSSpeller (De Witte et al., 2015). Inter-species conservation of ACRs themselves can also be informative.

Thirdly, chromosome conformation capture (3C;(Dekker et al., 2002) methodologies can be utilised to assay for physical interactions between gene-distal ACRs and gene promoters. Such interactions are indicative of CREs, because TF motifs must be brought into close spatial proximity with their target genes for bound TFs and their co-regulators to influence the basal transcription machinery and thus affect transcriptional outcomes (Weber et al., 2016; Marand et al., 2017). An early study employing this approach, which paired an ACR detection technique called FAIRE (formaldehyde-assisted isolation of regulatory elements) with a large array of 3C assays, revealed novel CREs of the classical maize *booster 1* (*b1*) flavonoid biosynthesis locus, in addition to confirming a previously mapped hepta-repeat regulatory region (Louwers et al., 2009). High-throughput, sequence-agnostic implementations of 3C, such as Hi-C, now allow genome-wide mapping of chromosome interactions (Belton et al., 2012). These can be correlated with genome-wide ACR data to substantiate putative CREs (Dong et al., 2017; Ricci et al., 2019; Han et al., 2020; Marand et al., 2021). As discussed above, a major advantage of Hi-C is that it is one of very few methods that can directly link a putative ACR with its target gene(s).

Fourthly, functional assays can be used to confirm whether candidate ACRs have transcription-enhancing activities. Enhancer-trapping is one such technology, which utilises a transposase to produce random genomic insertions of a reporter gene with a minimal promoter. Insertion near an endogenous enhancer can increase expression of the reporter above basal levels, allowing its detection. The location of the reporter in these positive hits can be determined by inverse PCR, but the enhancer interacting with it can still be difficult to identify because of CREs' ability to act over long distances (Springer, 2000; Chudalayandi, 2011). This, combined with its relatively low-throughput, means enhancer trapping is not appropriate for genome-wide discovery of CREs. However, the technique still finds utility in generating lines with cell-lineage-specific labelling for developmental studies (Amalraj et al., 2020). An alternative approach, enhancer-reporter constructs, resolves the former issue by cloning a known putative enhancer into a construct containing a reporter and minimal promoter. This is then transiently transformed into plant tissues or protoplasts to assess expression of the reporter. Traditionally this was a low-throughput process used to validate a few candidates at a time, but, recently, massively parallel reporter assays have been developed to increase throughput, such as STARR-seq, which can test millions of putative CREs or even examine whole genomes (Inoue & Ahituv, 2015; Muerdter et al., 2015). Nonetheless, a confounding factor in these functional studies is that query sequences are not trialled in their native genomic contexts, meaning effects on enhancer activity due to DNA looping, chromatin features, other CREs, or different promoter structures will not be taken into account. Furthermore, the TF(s) that mediate an enhancer's effects may simply not be expressed in the tissue or protoplasts that the construct is transformed into – or may not be encoded by the genome at all if constructs are tested in a different species for ease of transformation (Jores et al., 2020). Additionally, most current implementations of the methods discussed above are unable to detect transcriptional silencers (Doni Jayavelu et al., 2020).

Lastly, RNA-seq data can suggest whether putative CREs influence transcription *in vivo*. As discussed above, ACRs' size, accessibility, or even presence/absence may differ across different tissues, environments, and developmental stages. If an ACR's accessibility correlates, positively or negatively, with changes in expression of its putative target genes (for example, the closest genes up- and downstream in the absence of additional data), this is strong evidence that the ACR is a functional CRE (Zhang, Zhang, et al., 2012). To home in on CREs relevant to biological processes of interest, one can compare ACR and RNA-seq data from appropriate tissues across a developmental time series or set of environmental conditions. For example, this strategy has been used to identify CREs involved in salt

tolerance in Arabidopsis (Uygun et al., 2019), inflorescence development in maize (Parvathaneni et al., 2020), and cold response across several grasses (Han et al., 2020).

Here, we attempted to use ATAC-seq data in combination with UMR and RNA-seq data to detect spike-relevant CRE-gene interactions. We first produced 10 trial ATAC-seq libraries, troubleshot these both in the wet lab and computationally, then produced a final set of ATAC-seq data for a series of 16 wheat spike developmental stages. We also generated a companion RNA-seq dataset for the same tissues. Unfortunately, our ATAC-seq data was highly noisy and, ultimately, inadequate for reliable ACR detection. We therefore also analysed data from a collaborating group's concurrent ATAC-seq and RNA-seq project. From these samples we were able to detect >100,000 ACRs for each of seven developmental stages, with consistent numbers of ACRs per stage. We additionally leveraged previously generated wheat WGBS data to call UMRs for the corroboration of detected ACRs. Finally, we explored the utility of correlating the differential accessibility patterns of ACRs with the expression patterns of nearby genes for the detection of functional CRE-gene pairs. This method showed limited promise for the detection of enhancer elements, but not silencers.

4.3 – Results

4.3.1 – Spike and carpel RNA-seq replicates cluster by developmental stage and reveal gene expression differences across development

We collected microdissected tissues from bread wheat (cv. 'Chinese Spring') at 16 stages of spike development (Figure 4.2) with the aim of producing ATAC-seq and RNA-seq datasets. This comprised seven stages of whole spikes tissue (W1-W4), one stage where spikelets were collected (W5), and eight stages comprised of isolated carpels (W6-W10; Waddington et al., 1983). We pooled tissue from multiple plants as per Table 4.1.

We extracted RNA and sequenced four biological replicates per stage, with between 49.2 and 71.1 million 150 bp paired-end raw reads obtained per replicate. We calculated read counts and TPMs for all genes in the IWGSC RefSeq v1.1 annotation (Appels et al., 2018) using Kallisto (Bray et al., 2016). Principal component analysis (PCA) revealed that the samples clustered strongly by timepoint in all major PCs (>5% variance, Figure 4.3), indicating that spikes were staged accurately during tissue collection. A clear developmental trajectory was visible across PC1 and PC2 space, though PC3 better separated whole spike stages. Carpel data (W6-W10) was utilised in a separate project and is not further referenced.

For the seven whole-spike timepoints (W1-W4), of the 269,428 gene models, 153,582 genes were expressed (i.e. had non-zero counts in at least one replicate of at least one stage). ImpulseDE2 (Fischer et al., 2018) was used to identify 42,838 genes that were differentially expressed (DE) across the timecourse, i.e. with expression profiles significantly different to a flat intercept (padj < 0.001). Retaining only high-confidence (HC) genes expressed at > 0.5 transcripts per million (TPM) on average in at least one developmental stage reduced this number to 32,667. Of these, 2,229 were TFs as classified by (Evans et al., 2022).

4.3.2 – Low-coverage ATAC-seq for wheat spikes and carpels was ineffective for detecting accessible chromatin regions

We prepared ten trial ATAC-seq libraries from seven of the above developmental stages and conducted 150 bp paired-end sequencing on each. We aimed for 50 million paired reads per sample, equivalent to 15 Gbp or just under 1x coverage of the experimentally determined Chinese Spring haploid genome size (15.8 Gb; Appels et al., 2018). We obtained between 38.7 and 64.3 million raw read pairs, averaging 47.4 million.

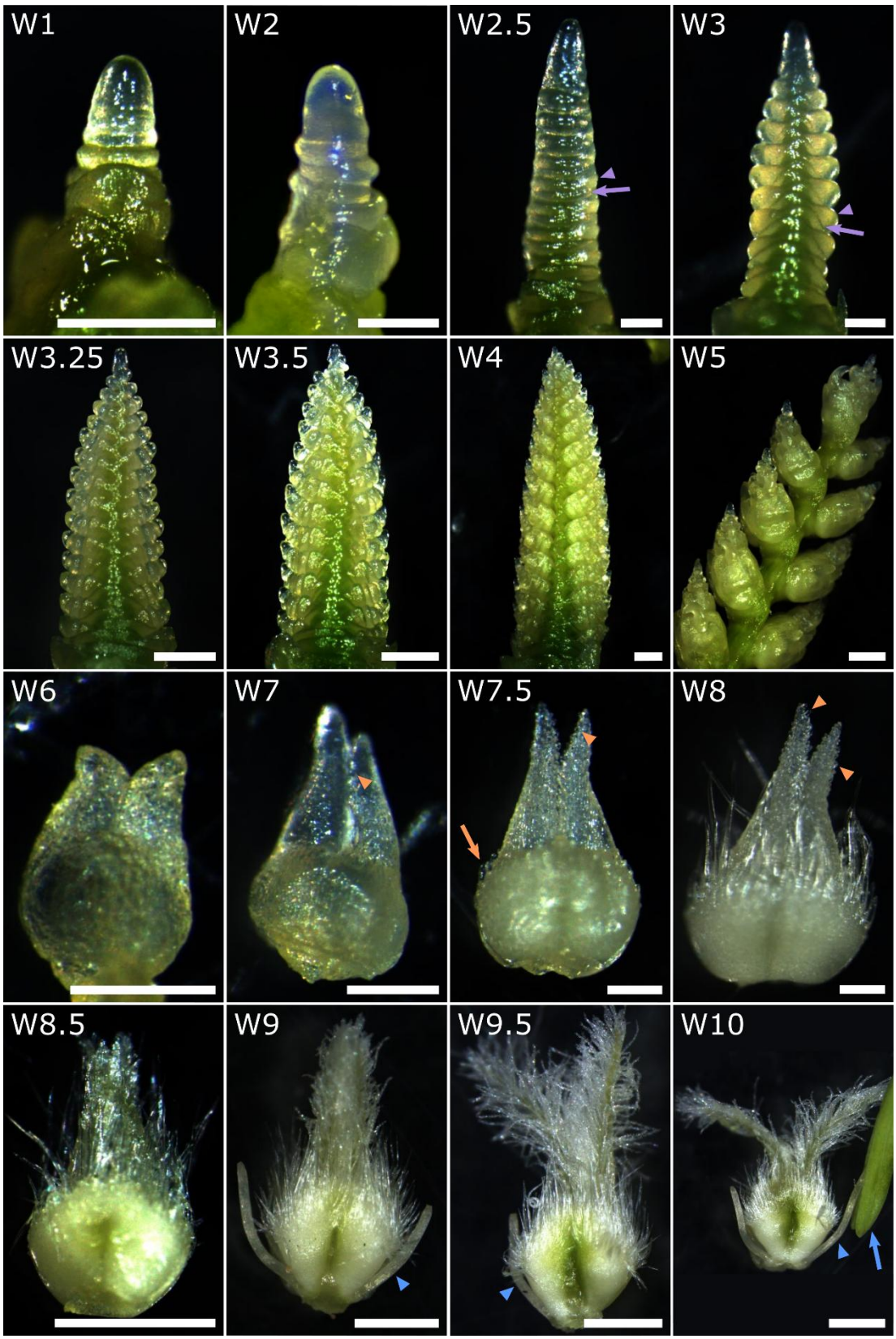


Figure 4.2 – Micrographs of microdissected wheat tissues (cv. 'Chinese Spring')
Scale bars represent 200 µm for row 1 (W1-W3), 500 µm for row 2 (W3.25-W5), 200 µm for row 3 (W6-W8), and 1,000 µm for row 4 (W8.5-W10). Purple arrowheads indicate spikelet primordia, while purple arrows indicate leaf primordia. Orange arrowheads indicate initiating stigmatic branches, and the orange arrow indicates initiating ovary wall hairs. Blue arrowheads indicate stamen filaments, while the blue arrow indicates an anther – these were removed prior to collection of carpel tissue. Image for W2.5 is rotated 90° compared with other spike images. 'W' indicates Waddington stage.

Table 4.1 – Descriptions of the wheat developmental stages microdissected for ATAC-seq and RNA-seq, including tissue requirements

Waddington stage	Stage name	Description	Tissue per replicate
W1	Transition	Apical meristem transitioning into a floral spike. Squat dome approx. 200 µm across, 250 µm in length. No or few ridges.	100 spikes (ATAC-seq) 50 spikes (RNA-seq)
W2	Early double ridge	Single, alternating ridges visible. Spike ~400 µm in length.	44 spikes (ATAC-seq) 30 (RNA-seq)
W2.5	Late double ridge	Clear 'double' ridges along >75% of length. Upper ridge = spikelet primordia, lower = bracteal primordia. Spike ~1250 µm in length.	35 spikes (ATAC-seq) 12 spikes (RNA-seq)
W3	Glume primordia	>80% of spikelet primordia swollen into bulbous structures. Growth of bracteal primordia terminates. Spike ~1400 µm in length.	25 spikes
W3.25	Lemma primordia	Spikelet primordia visibly segmented into ~3 sections. Spike ~1750 µm in length.	11 spikes
W3.5	Floret primordia	Spikelet primordia visibly segmented into ~4 sections. Spike 2.5-3.5 mm in length.	5 spikes
W4	Terminal spikelet	Individual florets visible without dissection of spikelets. Middle of spike wider than base. Spike ~5 mm in length.	3 spikes
W5	Carpel extension	Individual spikelets well separated from each other. Central spikelets ~1.5 mm in length.	12 spikelets
W6	Narrow canal	Stylar canal remains as a narrow opening. Tips of styles bend outwards. Carpels ~ 300 µm in length.	43 carpels (ATAC-seq) 20 carpels (RNA-seq)
W7	Stigmatic branches differentiating	Swollen cells just visible on styles, these are stigmatic branches just differentiating. Carpels 500-600 µm in length.	31 carpels (ATAC-seq) 17 carpels (RNA-seq)
W7.5	Stigmatic branches elongating	Unicellular hairs just differentiating on ovary wall. Stigmatic branches slightly elongated. Carpels 800-1000 µm in length, ovaries ~500 µm across.	41 carpels
W8	Hairs elongating	Hairs on ovary wall elongated, longest ≥2/3 length of styles. Carpels ~1000 µm in length, ovaries ~750 µm across.	25 carpels
W8.5	Branches form tangled mass	Styles move closer together, stigmatic branches form a tangled mass. Carpels ~1500-1750 µm in length, ovaries ~1000-1200 µm across	13 carpels
W9	Branches erect	Styles and stigmatic branches erect, styles held together. Additional hairs continue to differentiate and existing hairs elongate. Carpels ~3 mm in length, ovaries ~1.5 mm across.	11 carpels
W9.5	Styles spreading outwards	Styles and stigmatic branches spreading outwards. Stigmatic hairs well developed. Carpels ~4mm in length (if include full length of bending style), ovaries ~1.5mm across.	9 carpels
W10	Pollen	Styles curved outwards, stigmatic branches spread wide. Pollen grains visible on some stigmatic hairs. Carpels ~4mm in length (if include full length of bending style), ovaries ~1.5mm across.	9 carpels

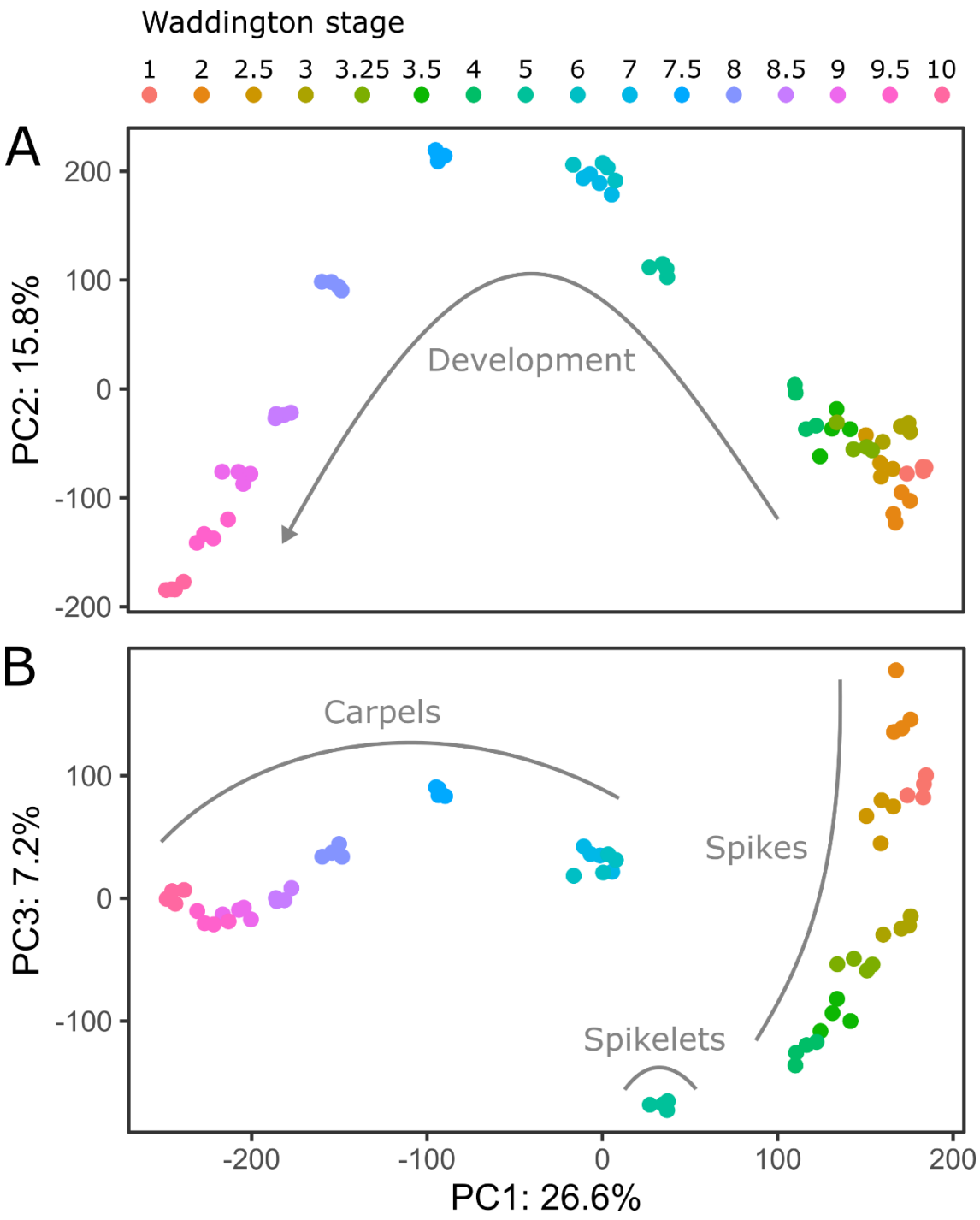


Figure 4.3 – Chinese Spring spike and carpel RNA-seq samples cluster by tissue and developmental stage

Scatterplots of samples by principal components 1 and 2 (**B**) or 1 and 3 (**C**). PCA was conducted on HC genes with non-zero variance. Subsequent PCs each accounted for less than 5% of total variation.

We also downloaded the raw reads from a previous study which conducted ATAC-seq on bread wheat leaf samples (cv. Paragon; Lu et al., 2020). This dataset included three samples from intact chromatin, plus a 'naked' DNA sample – genomic DNA with highly degraded chromatin – which can be used to control for sequence biases in the background Tn5 insertion rate. On average, these samples contained 45 million read pairs. We then conducted a series of quality assessment and filtering steps to determine the efficacy of our ATAC-seq protocol versus this published data (Table 4.2). We also examined a bread wheat seedling ATAC-seq dataset (cv. Chinese Spring, no replicates; Concia et al., 2020), but we identified that a very high proportion of reads were duplicates (99.7%) and therefore did not utilise it as an additional reference dataset (Table 4.2).

Firstly, we trimmed reads to remove adapter read-through and then filtered out reads shorter than 38 bp as this is the minimum spacing of transposition events (Adey et al., 2010). From our samples, 82-95% of reads were retained following adapter trimming, averaging 92%. For comparison, 95% of read pairs passed this post-trimming threshold for the three intact chromatin samples from (Lu et al., 2020). These results indicated slight over-tagmentation of our samples and we therefore revised our tagmentation protocol to allow the volume of Tn5 transposase used to scale with the number of nuclei in the target replicate (see Methods 4.3.3).

We then mapped reads to the IWGSC RefSeq v1.0 assembly. The majority of read pairs in our samples mapped concordantly with a mean of 93%. However, after filtering for mapping quality (see Methods 4.3.5), the retained fractions were more modest, between 35% and 50% of the raw read number (mean 44%). In comparison, the reference leaf samples retained just 19% on average. Next, we assessed the number of reads mapping to the mitochondrial or plastidial genomes (we appended organellar genome sequences to the RefSeq v1.0 before read mapping; Ogihara et al., 2002; Ogihara et al., 2005). The organellar genomes do not possess eukaryotic chromatin and are very accessible to the Tn5 transposase. As a result, ATAC-seq is not informative of organellar gene regulation and contamination of nuclei preparations with mitochondria or chloroplasts can absorb a high proportion of tagmentation reactions and sequencing reads, sometimes above 50% (Montefiori et al., 2017; Rickner et al., 2019; Smith et al., 2021). In our data, just 0.09% of concordant pairs mapped to the organellar genomes on average, indicating that our nuclei isolations were highly clean. This was much lower than the average for Lu et al.'s samples (3.5% of concordant pairs). Reads mapping to unassembled scaffolds ('ChrUn') were then also removed.

Table 4.2 – Bioinformatic statistics for trial ATAC-seq samples and reference leaf samples

Source	Sample	Raw read pairs	Post-trimming	MAPQ>30 on nuclear pseudo-molecules	Deduplicated	Percent of raw read pairs remaining	Peaks called
This thesis	W3_trial	49,411,342	46,852,749	22,816,281	19,059,242	38.6	104
	W3.25_trial_1	50,780,575	46,986,303	20,973,842	17,698,729	34.9	72
	W3.25_trial_2	44,646,024	41,162,291	19,179,774	16,168,057	36.2	41
	W3.5_trial_1	43,014,390	41,054,547	19,767,102	16,843,743	39.2	401
	W3.5_trial_2	38,741,979	31,864,025	10,787,352	8,436,747	21.8	30
	W4_trial	52,722,894	46,756,686	19,432,903	16,011,053	30.4	1,388
	W5_trial_1	41,265,640	37,200,399	17,274,149	14,633,807	35.5	39
	W5_trial_2	64,311,911	59,899,110	22,440,395	19,835,792	30.8	18
	W8_trial	48,885,874	45,598,659	20,825,985	17,365,144	35.5	223
	W9_trial	40,205,948	37,244,690	13,807,710	11,918,142	29.6	779
Lu et al., 2020	Leaf_1	46,512,936	43,381,842	11,397,004	9,437,318	20.3	180,737
	Leaf_2	41,051,085	39,359,388	5,451,071	4,454,887	10.9	158,564
	Leaf_3	44,854,458	43,099,959	5,791,248	4,721,575	10.5	156,038
	Leaf_naked	45,604,062	43,574,596	18,000,719	15,837,478	34.7	N/A
Concia et al., 2020	Seedling	271,129,872	264,854,516	356,824,112	1,236,848	99.7	N/A

Lastly, we filtered duplicate reads (which may arise as PCR duplicates or sequencer optical errors). On average, 16% of the remaining reads from our samples were duplicates. This is a relatively low rate, indicating that sufficient nuclei were present to achieve the target sequencing depth and that the numbers of PCR cycles we used for library preparation were appropriate. In comparison, 18% of reads were duplicates in the samples from (Lu et al., 2020).

We then used the peak calling software MACS2 (Zhang et al., 2008; Gaspar, 2018) to identify ACRs. This included comparing each of our samples against the naked DNA control sample from (Lu et al., 2020) to account for the background insertion rate for the Tn5 transposase. Our trial samples yielded, on average, just 310 peaks, though this ranged greatly, from 18 to 1,388. In contrast, Lu et al.'s samples produced 165,113 peaks on average. This was despite these samples retaining, on average, fewer than half the number of read pairs as our samples (8.6 million versus 15.8 million). To investigate why so few peaks could be called from our data, the mapped, filtered, and depth-normalised read coverage profiles were visualised using the Integrative Genome Viewer (IGV; Robinson et al., 2011). Across all chromosomes from all samples, our samples exhibited a highly rarefied distribution of reads, spread across both genic and intergenic regions. This bore a strong resemblance to Lu et al.'s control sample, though there were occasional regions of high read density. These often, but not always, coincided with ACRs in Lu et al.'s non-control samples, in which reads were largely confined to distinct ACRs close to genes and were mostly absent from intergenic regions. Example intergenic and genic (*VRT-A2;* TraesCS7A02G175200) regions are shown in Figure 4.4 and Figure 4.5.

This suggested that despite the nuclei for our samples appearing intact when assessed with DAPI staining (Faci, Jones, et al., 2024), their chromatin was partially degraded prior to tagmentation. As a result, it appears that most regions across the genome were accessible to the transposase in at least some nuclei – even where this would not have been the case *in vivo* – leading to a low background level of noise. The difference in ATAC-seq data quality between our samples and those from (Lu et al., 2020) could be due to the different methodologies employed (see Discussion).

However, the presence of some peaks and sites of clear preferential insertion indicated that the chromatin produced by our methodology could still be sufficiently intact to support detection of biologically relevant accessible chromatin. We predicted that obtaining higher sequencing coverage on similar samples would yield greater numbers of reads mapping to ACRs, while the read depth of the background would increase only marginally as additional random insertions would fill in gaps in the noise.

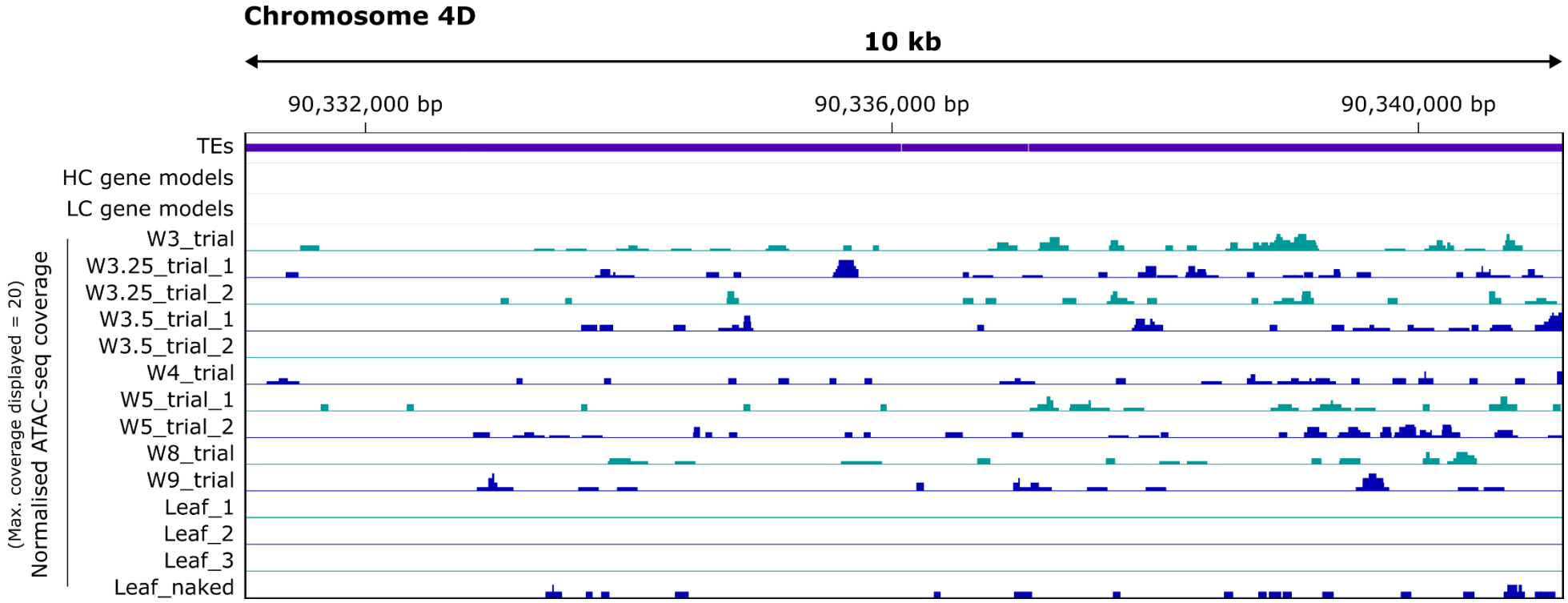


Figure 4.4 – Trial spike and carpel ATAC-seq samples were much noisier in intergenic regions than reference leaf samples
Annotated 10 kb region of chromosome 4D (IWGSC RefSeq v1.0 assembly). Teal and blue tracks (alternated for clarity) indicate coverage of our trial ATAC-seq samples and reference leaf samples from (Lu et al., 2020). Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 20 to aid visual comparison of lanes. Data visualised using the IGV (Robinson et al., 2011).

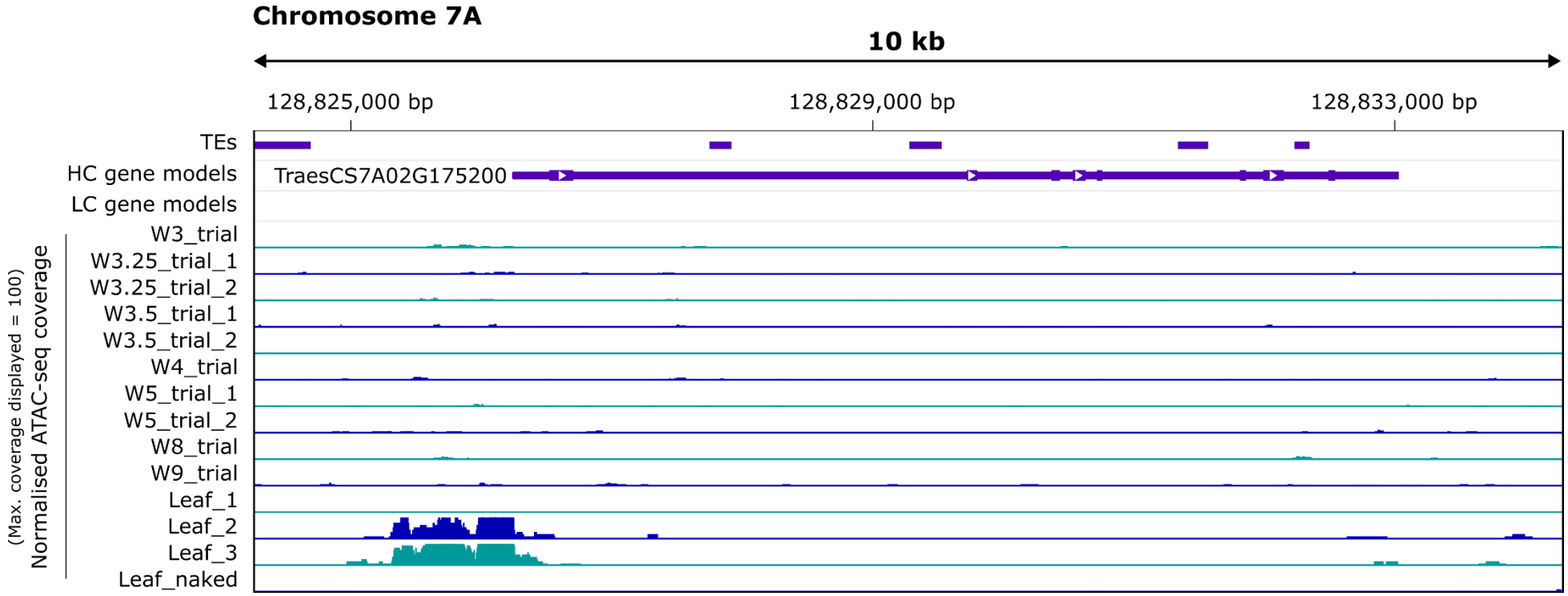


Figure 4.5 – Trial spike ATAC-seq samples did not show clear sites of enrichment, unlike reference leaf samples
Annotated 10 kb region of chromosome 7A (IWGSC RefSeq v1.0 assembly) containing *VRT-A2*. Teal and blue tracks (alternated for clarity) indicate coverage of our trial ATAC-seq samples and reference leaf samples from (Lu et al., 2020). Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 100 to aid visual comparison of lanes. Data visualised using the IGV (Robinson et al., 2011).

Ultimately, we thought this would yield greater contrast between ACRs and noise, allowing more statistically significant ACRs to be called.

4.3.3 – Higher-coverage ATAC-seq data only moderately improved detection of accessible chromatin regions

Given the above results, we proceeded to generate further libraries for the previously described 16 developmental stages (cv. Chinese Spring) and sequenced these to a higher depth. We produced libraries for five biological replicates for each stage, but not all samples passed pre-sequencing quality controls. Ultimately, 1-4 replicates were sequenced for each stage, totalling 50 samples (Table 4.3). We aimed for 200 million paired reads (60 Gbp) per sample, i.e. four-fold the target for our trial samples. Between 184 and 318 million paired reads were delivered for most samples, though one sample produced many more reads (W3_rep4; 410 million) and one far fewer (W2.5_rep4; 54 million) (Table 4.3; Figure 4.6). Excluding these, the average number of raw read pairs was 218 million (SD = 25 million). We also sequenced four naked DNA controls derived from different stages to a target depth of 50 million paired reads, achieving an average of 46 million.

We trimmed, mapped, filtered, and deduplicated these samples as described above. Omitting the mentioned outlier values, this left, on average, 69 million read pairs (SD = 8.0 million; Table 4.3, Figure 4.6). For the four control samples, an average of 20 million read pairs (SD = 5.0 million) remained. We then called peaks for each developmental stage by combining replicates, supplying the four controls to provide estimates of the background insertion rate. After normalising for sample read depth, this data appeared less noisy than our trial ATAC-seq samples when viewing the same intergenic regions (Figure 4.7). However, they did still exhibit more noise than Lu et al.'s leaf samples.

We expected that a greater sequencing depth would provide a greater signal-to-noise ratio, allowing us to call higher numbers of ACRs. This was true to some extent, as more ACRs, on average, were called from these samples than from our lower coverage samples; 9,289 vs 307 (Table 4.3). However, the numbers of ACRs called for different developmental stages were highly varied, from 314 to 49,174, and even adjacent developmental stages exhibited major differences in peak numbers. This precluded the explanation that the large range of peak numbers could be explained by global up- or downregulation of total ACR number across spike development. This suggested that we could not use the emergence or disappearance of ACRs through time as markers for differentially active CREs. We also still called considerably fewer peaks versus the benchmark Lu et al. (2019) samples.

Table 4.3 – Bioinformatic statistics for final ATAC-seq samples (continued on next page)

Stage	Replicate	Raw read pairs	Post-trimming	MAPQ>30 on nuclear pseudo-molecules	Deduplicated	Percent of raw read pairs remaining	Peaks called
W1	_2	222,453,938	207,629,690	90,572,548	70,988,825	31.9	1,603
	_4	267,820,886	253,368,313	110,414,004	88,478,702	33.0	
W2	_3	222,036,656	211,865,137	91,648,149	64,814,346	29.2	22,632
	_4	263,741,537	249,026,235	106,977,759	83,943,695	31.8	
	_5	221,964,830	207,621,599	92,215,715	73,517,672	33.1	
W2.5	_2	208,692,061	189,557,068	83,304,774	69,230,968	33.2	314
	_3	239,390,857	213,946,404	97,412,440	76,379,293	31.9	
	_4	53,531,517	51,202,392	17,346,675	14,451,308	27.0	
W3	_1	215,463,194	185,192,194	72,707,333	56,673,529	26.3	332
	_4	410,033,194	342,693,709	143,450,089	118,181,554	28.8	
W3.25	_3	228,070,904	218,199,201	79,132,632	64,504,617	28.3	6,278
	_4	208,754,116	192,521,273	84,861,820	70,825,648	33.9	
	_5	198,196,995	183,721,245	75,903,777	62,158,141	31.4	
W3.5	_1	203,176,005	188,011,359	86,813,305	71,883,274	35.4	4,146
	_2	211,318,714	196,882,568	86,226,194	69,821,998	33.0	
	_3	202,182,759	187,796,870	80,257,623	64,860,743	32.1	
	_4	203,211,165	190,962,044	83,638,183	66,752,464	32.8	
W4	_1	225,661,492	212,334,889	95,096,335	76,347,350	33.8	18,352
	_2	206,120,548	193,054,324	85,976,012	69,048,717	33.5	
	_3	204,940,041	192,055,211	82,304,704	67,576,145	33.0	
	_5	239,617,369	219,249,649	94,190,521	70,122,583	29.3	
W5	_1	217,982,567	206,460,426	93,012,550	68,280,302	31.3	5,731
	_3	209,792,825	196,055,619	91,711,001	69,678,887	33.2	
	_4	205,850,577	193,601,307	89,262,992	71,033,390	34.5	
	_5	214,996,567	190,632,652	91,131,086	74,602,467	34.7	
W6	_1	198,797,573	180,320,882	75,983,379	56,698,336	28.5	996
	_2	199,660,090	187,339,531	83,496,889	65,596,832	32.9	

Stage	Replicate	Raw read pairs	Post-trimming	MAPQ>30 on nuclear pseudo-molecules	Deduplicated	Percent of raw read pairs remaining	Peaks called
W7	_1	318,312,793	293,305,018	139,235,501	72,450,904	22.8	496
	_2	198,796,432	188,811,220	91,404,632	69,065,369	34.7	
	_3	205,261,863	182,740,039	85,052,609	59,026,075	28.8	
	_4	184,063,302	146,170,172	61,028,887	47,842,275	26.0	
W7.5	_1	214,292,510	182,934,151	76,051,696	58,462,110	27.3	1,126
	_3	187,264,323	151,044,104	66,059,898	55,153,143	29.5	
	_4	195,912,903	169,236,401	77,518,675	59,819,983	30.5	
W8	_2	218,160,521	193,228,487	91,750,852	69,655,257	31.9	18,662
	_3	202,425,605	175,190,370	82,759,974	63,267,573	31.3	
	_4	224,513,868	203,173,081	100,680,899	77,310,950	34.4	
	_5	212,697,832	189,193,023	91,400,418	71,362,628	33.6	
W8.5	_1	282,492,264	269,133,823	119,166,286	87,076,637	30.8	49,174
	_2	204,863,167	190,864,337	90,048,478	70,603,243	34.5	
	_3	206,691,717	191,349,401	92,753,868	73,540,218	35.6	
W9	_1	196,515,446	181,188,879	84,434,431	64,866,105	33.0	13,358
	_2	231,433,641	196,389,175	90,397,787	72,281,941	31.2	
	_3	221,507,974	206,078,165	99,344,947	76,152,347	34.4	
	_4	225,945,058	199,646,794	92,511,467	69,398,865	30.7	
W9.5	_1	209,153,589	195,113,553	93,524,674	75,364,703	36.0	1,126
	_3	210,169,414	193,436,213	91,010,048	69,991,868	33.3	
	_4	213,572,563	196,869,490	86,523,151	57,204,684	26.8	
	_5	251,730,650	216,361,081	100,614,563	79,547,196	31.6	
W10	_2	202,969,805	176,457,853	82,376,575	67,654,294	33.3	4,302
Naked DNA control	W2.5_naked	59,090,136	57,050,927	27,278,198	23,966,312	40.6	NA
	W5_naked	36,425,164	34,843,969	18,274,154	16,095,345	44.2	
	W7.5_naked	35,520,971	34,168,889	18,428,538	15,903,633	44.8	
	W9.5_naked	54,811,855	53,011,051	28,655,256	25,247,356	46.1	

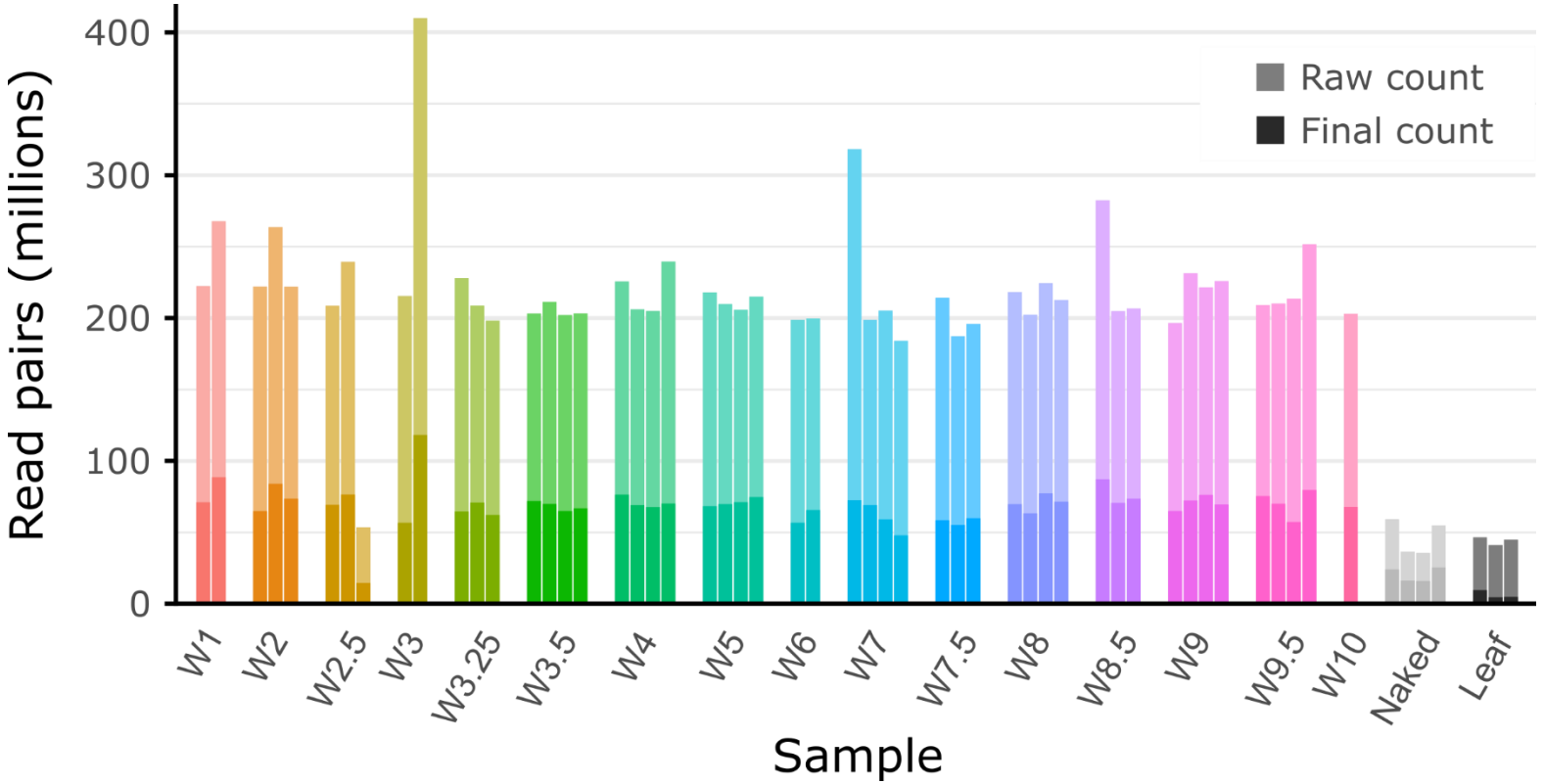


Figure 4.6 – Our final ATAC-seq samples contained 5-fold the raw reads and 8-fold the filtered reads of reference leaf samples
Bar plots displaying raw (lighter) and filtered (darker) read pair numbers for our final samples and three leaf samples from (Lu et al., 2020)

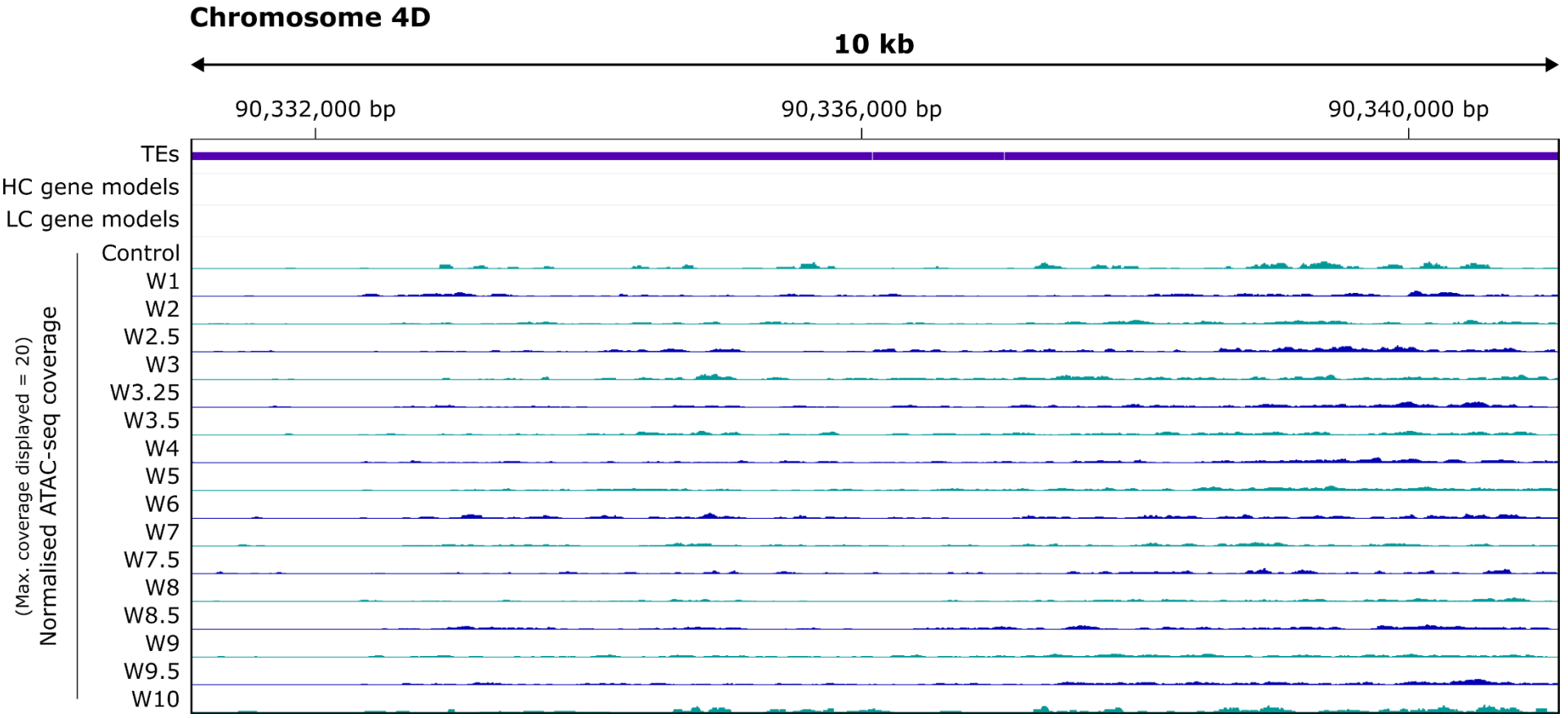


Figure 4.7 – Final spike and carpel ATAC-seq samples show relatively little noise in intergenic regions
Annotation of the same 10 kb region of chromosome 4D (IWGSC RefSeq v1.0 assembly) illustrated in Figure 4.4. Teal and blue tracks (alternated for clarity) indicate coverage of pooled replicates from our final ATAC-seq experiments. Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 20 to aid visual comparison of lanes. Data visualised using the IGV (Robinson et al., 2011).

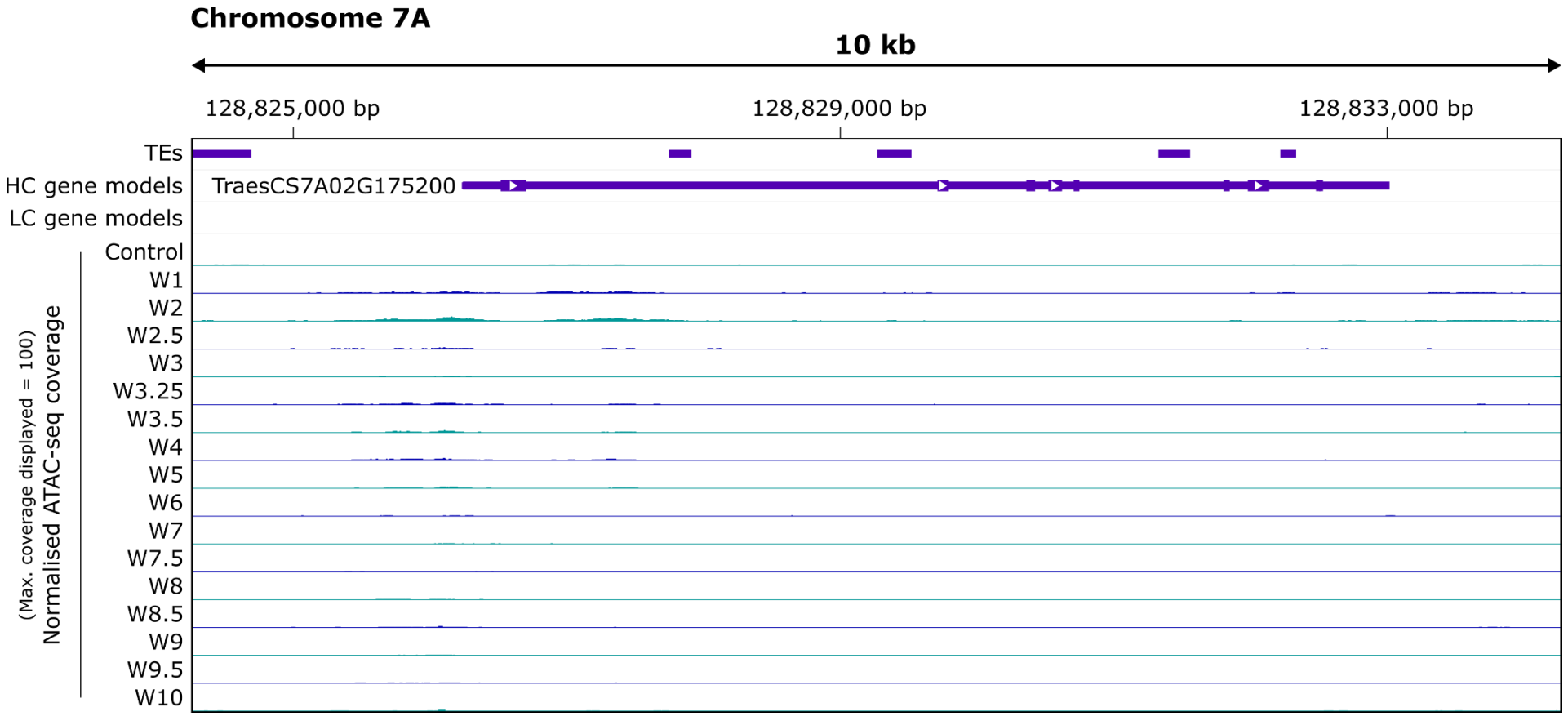


Figure 4.8 – Final spike and carpel ATAC-seq samples did not show clear sites of enrichment
Annotation of the same 10 kb region of chromosome 7A (IWGSC RefSeq v1.0 assembly) containing *VRT-A2* illustrated in Figure 4.5. Teal and blue tracks (alternated for clarity) indicate coverage of pooled replicates from our final ATAC-seq experiments. Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 100 to aid visual comparison of lanes. Data visualised using the IGV (Robinson et al., 2011)

These results were reflected in examinations of test genic regions. For example, while slight enrichments above background level could be observed upstream of the TSS of *VRT-A2* and inside its first intron (Figure 4.8) – which harbours a known, functional CRE (Adamski et al., 2021; J. Liu et al., 2025) – peaks were only called here for the W2 stage. Normalised coverage was significantly stronger here in the samples from Lu et al. (2019). Ultimately, the data appeared to be of limited use for our intended application of quantitative methods to catalogue genome-wide spike-relevant CREs based on correlation of ATAC-seq and RNA-seq signal.

4.3.4 – High-quality ATAC-seq data suggests that differentially expressed genes are enriched for differentially accessible chromatin regions

While we were producing and analysing the above ATAC-seq and RNA-seq datasets, another lab produced a similar resource for bread wheat (cv. Kenong 9204) for eight spike stages between SAM and W4 (Lin et al., 2024). Given the limitations of our own data, we decided to remap and analyse this data using our own pipeline to enable comparison with our own data. We processed the ATAC-seq data (2 biological replicates per stage) as above, except that no naked DNA control was used. The raw data contained an average of 205 million read pairs (SD = 41 million), while the final filtered data contained 56 million pairs on average (SD = 16 million).

The mapped reads displayed a low level of background noise, with reads clustered mostly at ACRs (Figure 4.9, Figure 4.10). The number of ACRs called per stage was relatively consistent with an average of 294,821 (SD = 40,479). There was also no developmental trend for increasing or decreasing numbers of peaks, with the minimum number of peaks at W3.5 (253,569) and the maximum at W2.5 (347,771) (Table 4.4).

Next, putative CREs were identified by demarcating consensus ACRs (CACRs), defined as regions containing an ACR between W1 and W4. This strategy assumed that downregulated CREs would still produce some signal versus background heterochromatin. 200,198 CACRs were identified. This appeared to be a somewhat appropriate order of magnitude given that there are 107,892 HC gene models in the RefSeq v1.1 annotation. To quantify the chromatin accessibility of the CACRs, we extracted the maximum read coverage values (normalised by each sample's final read depth) underlying each CACR for each replicate of each stage.

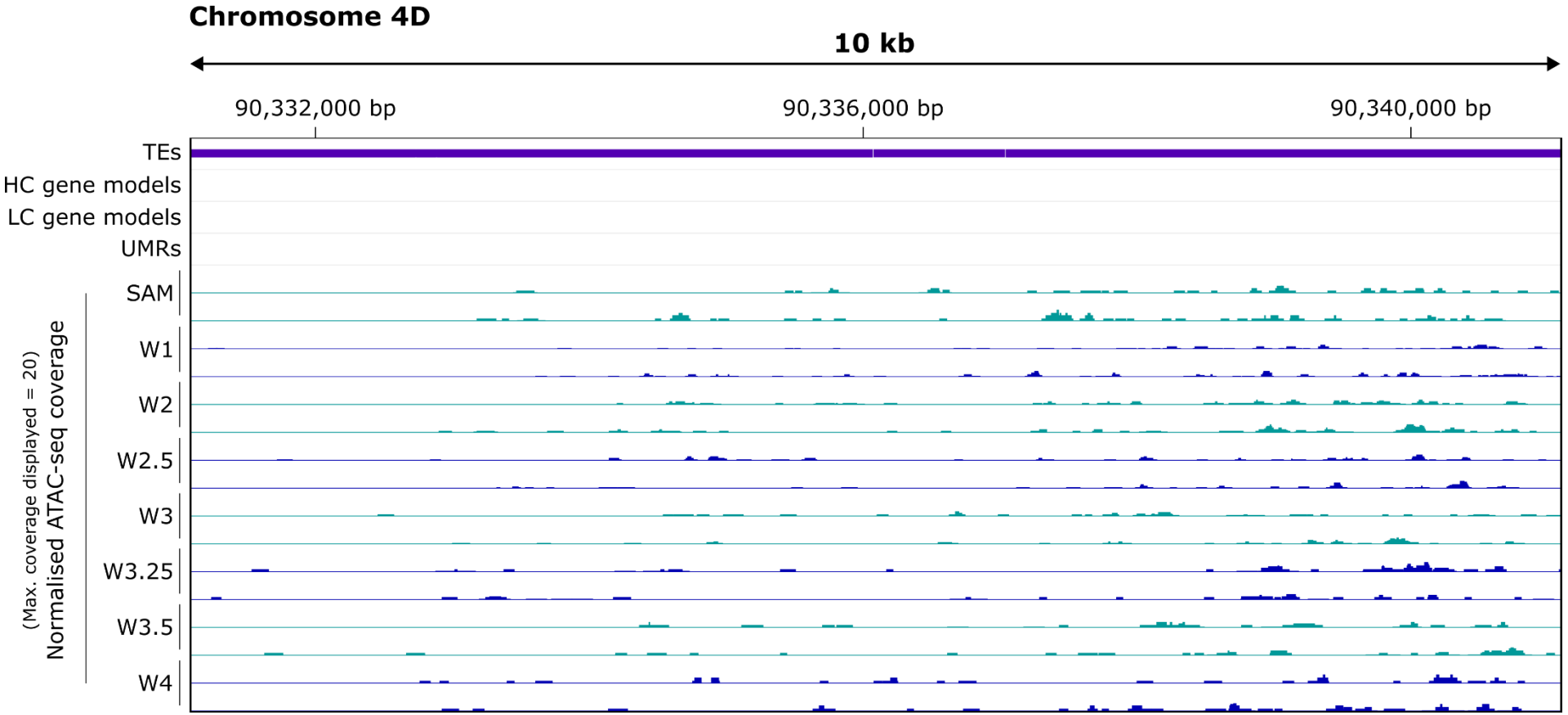


Figure 4.9 – ATAC-seq samples from Lin et al. (2024) show relatively little noise in intergenic regions
Annotation of the same 10 kb region of chromosome 4D (IWGSC RefSeq v1.0 assembly) illustrated in Figure 4.4 and Figure 4.7. Teal and blue tracks (alternated for clarity) indicate coverage of ATAC-seq samples from (Lin et al., 2024). Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 20 to aid visual comparison of lanes. A track for the inferred UMRs was also included (magenta). Data visualised using the IGV (Robinson et al., 2011).

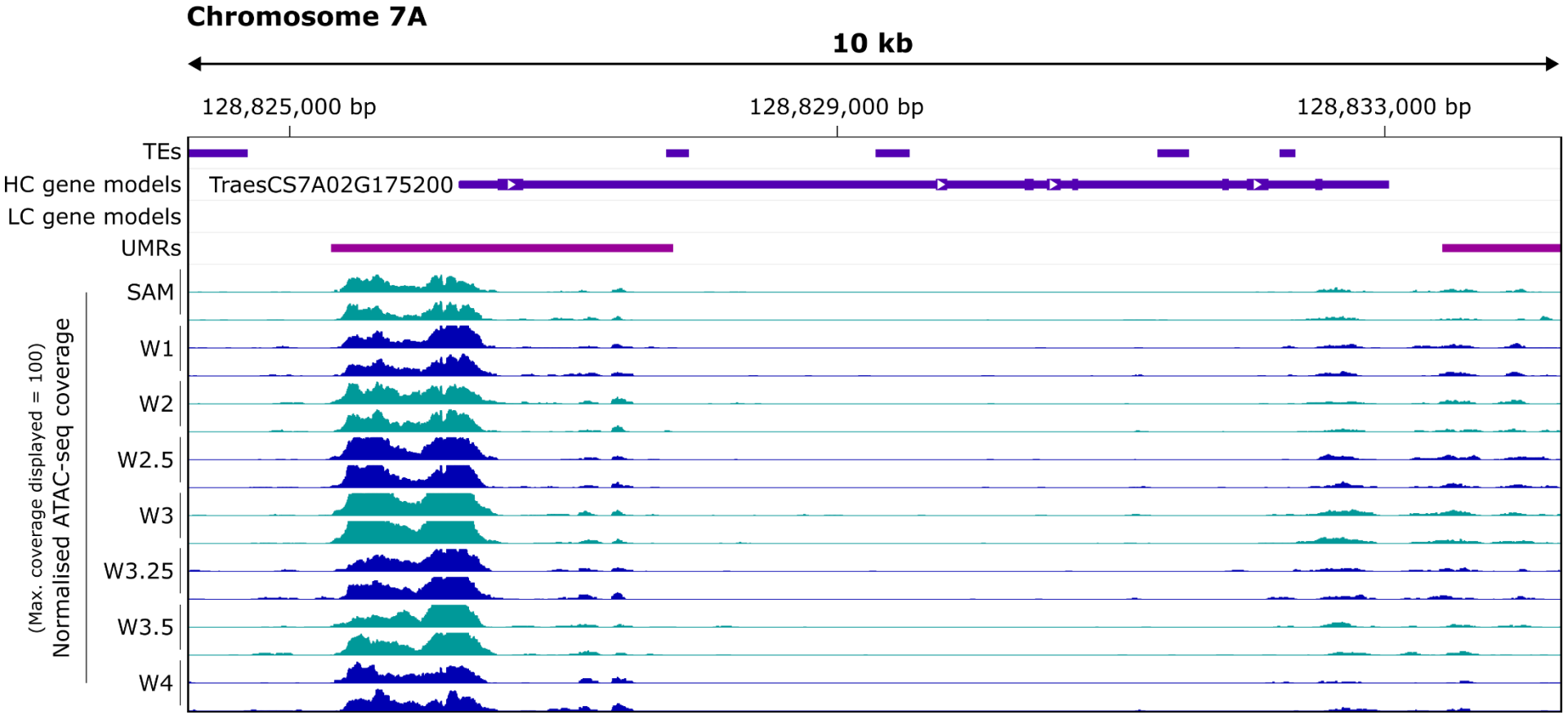


Figure 4.10 – ATAC-seq samples from Lin et al. (2024) show clear sites of enrichment
Annotation of the same 10 kb region of chromosome 7A (IWGSC RefSeq v1.0 assembly) containing *VRT-A2* illustrated in Figure 4.5 and Figure 4.8. Teal and blue tracks (alternated for clarity) indicate coverage of ATAC-seq samples from (Lin et al., 2024). Coverage was normalised against total read number and tracks were standardised to a maximum normalised coverage of 100 to aid visual comparison of lanes. A track for the inferred UMRs was also included (magenta). Data visualised using the IGV (Robinson et al., 2011)

Table 4.4 – Bioinformatic statistics for Kenong 9204 samples from Lin et al. (2024)

Stage	Replicate	Raw read pairs	Post-trimming	MAPQ>30 on nuclear pseudo-molecules	Deduplicated	Percent of raw read pairs remaining	Peaks called
SAM	_1	183,015,076	176,181,685	51,550,203	44,832,402	24.5	246,912
	_2	151,336,308	145,437,626	44,538,589	38,766,931	25.6	
W1	_1	275,376,868	264,340,130	91,183,674	78,191,670	28.4	328,960
	_2	281,879,188	271,608,892	85,980,884	74,614,919	26.5	
W2	_1	248,235,631	238,611,656	75,178,767	66,027,143	26.6	318,081
	_2	265,658,572	254,686,429	79,520,142	68,663,044	25.8	
W2.5	_1	187,379,764	180,706,785	84,777,119	72,311,778	38.6	347,771
	_2	188,943,894	181,225,616	86,975,757	74,077,091	39.2	
W3	_1	179,749,120	173,069,295	79,453,374	67,342,158	37.5	330,303
	_2	197,526,228	190,919,155	88,656,992	74,022,994	37.5	
W3.25	_1	217,944,875	212,979,894	45,787,959	39,655,754	18.2	273,478
	_2	178,266,352	174,356,401	51,395,230	43,982,183	24.7	
W3.5	_1	191,295,779	186,873,461	46,453,015	39,406,256	20.6	253,569
	_2	208,417,431	203,195,747	55,781,942	46,488,516	22.3	
W4	_1	154,551,291	149,858,597	41,353,042	36,403,600	23.6	259,492
	_2	172,976,380	168,469,187	44,826,974	38,704,569	22.4	

We also pseudomapped the Kenong 9204 RNA-seq data (3 biological replicates per stage) to the RefSeq v1.1 transcriptome to generate transcript counts. 23,908 HC genes on chromosome pseudomolecules were DE ($p < 0.001$) across the seven spike development stages of interest. Next, a set of extended coordinates was generated for each gene, extending from the 5' UTR to the next upstream gene and from the 3' UTR to the next downstream gene (while also accounting for certain edge cases, see Methods). These were intersected with the CACRs to assign them a set of neighbouring genes. For most CACRs lying in intergenic regions, this meant that two neighbouring genes were assigned. This could increase the false positive rate observed in subsequent tests, but permitting CREs to be assigned to both upstream and downstream genes is consistent with their known position- and orientation-independent mode of action (Schmitz et al., 2021). Nonetheless, restricting CACRs to upstream of gene transcription start sites (TSS), or even to within a set distance of the TSS, should also be investigated in the future.

The resulting 385,947 partially redundant CACRs were then tested for differential chromatin accessibility across spike development, with 95,395 exhibiting differential expression (differentially accessible CACRs; i.e. dCACRs). Crucially, significantly differentially expressed genes (DEGs) were enriched for both total number of CACRs (4.95 on average, 37% more) and for dCACRs (1.35 on average, 61% more) versus genes with constant but non-zero expression (3.61 and 0.84 on average, respectively; Figure 4.11A).

To confirm the validity of this finding, 500 additional datasets were simulated and tested. In each simulation, each CACR-gene pair was randomly reassigned a new 'neighbouring' HC gene, then the number of CACRs and dCACRs associated with the original list of 23,908 DEGs was calculated. On average, DEGs were associated with 3.67 CACRs (SD = 0.011) and 0.91 dCACRs (SD = 0.005) in these shuffled datasets. The results from the real dataset therefore lay well outside the distribution of simulated means for both metrics, at +112 SD for CACRs and +82 SD for dCACRs (Figure 4.11B).

Previous work has shown that most unmethylated DNA regions (UMRs) are stable across the plant lifecycle and represent the superset of regions that become accessible in different tissues and at different life cycle stages (Crisp et al., 2020). We hypothesised that CACRs overlapping UMRs would be more likely to represent real CREs and thus be more likely to influence nearby genes. We therefore generated a wheat UMR set using published leaf WGBS data (Appels et al., 2018) as previously described, with UMRs defined as regions ≥300 bp containing <10% cytosine methylation across CG, CHG, and CHH contexts (Crisp et al., 2020; Eglitis-Sexton et al., 2024). This yielded 335,231 UMRs, ranging from the minimum size of 300 bp up to 13,800 bp (mean = 822 bp, SD = 731 bp).

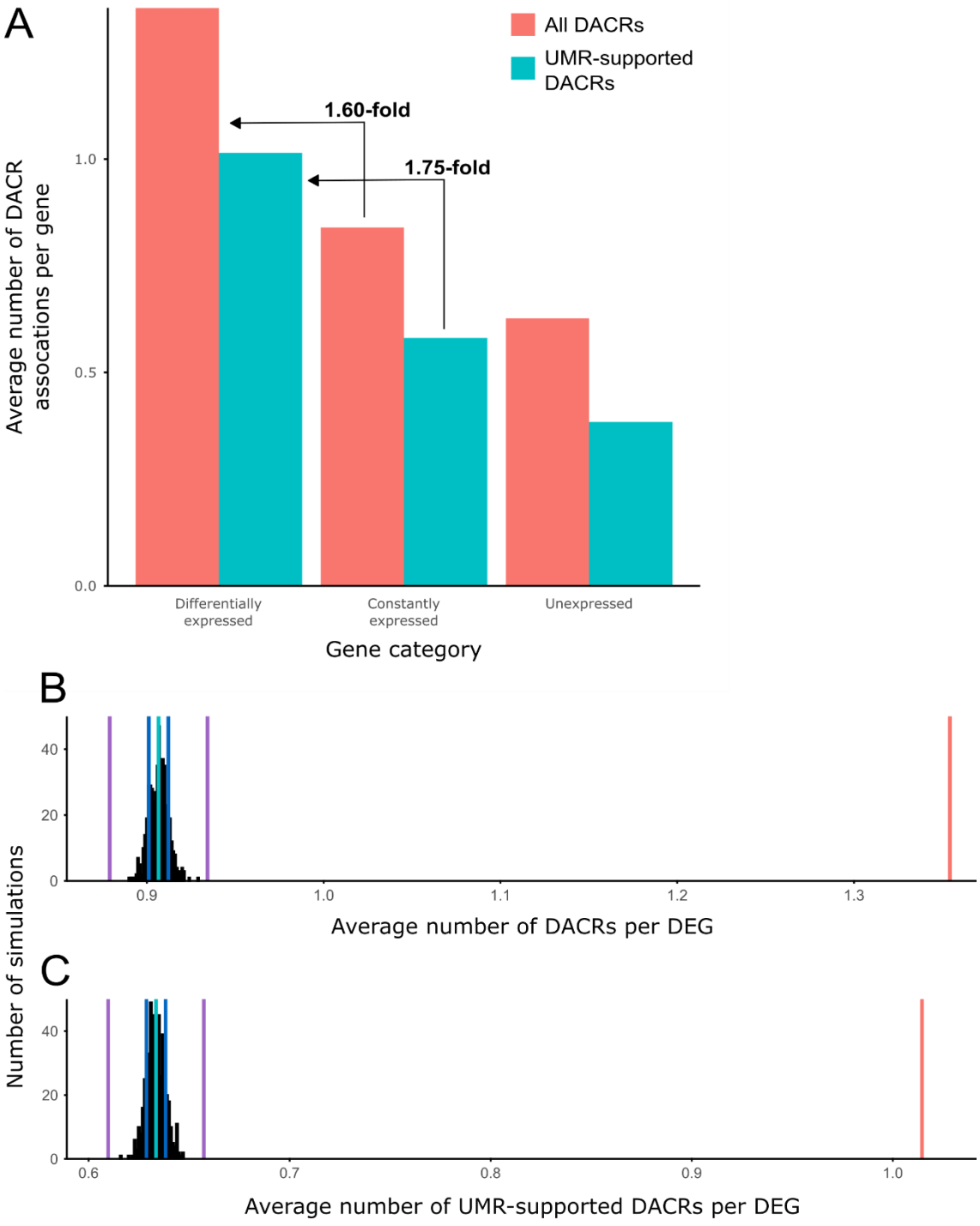


Figure 4.11 – Differentially expressed genes (DEGs) are enriched for dCACRs versus constantly expressed or non-expressed genes

A, DEGs are enriched for dCACRs and UMR-supported dCACRs versus constantly expressed and unexpressed genes. **B,C,** the real dataset (red) has a higher average number of dCACRs (**B**) or UMR-supported dCACRs per DEG (**C**) than was observed in 500 simulations. Light blue, dark blue, and purple lines denote the mean, mean ± 1 SD, and mean ± 5 SD, respectively.

The total UMR space observed in diploid plants assayed to date has typically been around 100 Mb regardless of total genome size, reflecting the general dearth of UMRs in TE-rich regions (Eglitis-Sexton et al., 2024). Given that bread wheat is hexaploid, the size of our UMR set, 276 Mb, was roughly in line with expectation. This suggests that the WGBS data we used was of sufficient quality and coverage for this application.

We found that by using the subset of CACRs overlapping UMRs, DEGs were further enriched for total CACRs (3.33 on average, 44% more) and dCACRs (1.01 on average, 75% more) versus constantly expressed genes (Figure 4.11A). However, this filtering process did diminish the absolute number of dCACRs associated with DEGs, from 32,376 to 24,255. We then simulated 500 random datasets in the same manner as above. In these shuffled datasets, DEGs were paired with 2.33 UMR-supported CACRs (SD = 0.009) and 0.63 (SD = 0.005) UMR-supported dCACRs on average. The real data therefore outperformed the shuffled datasets, with results lying outside the distribution of null means at +217 SD for CACRs and +171 SD for dCACRs (Figure 4.11C).

These tests suggested that there was some capacity to detect real CRE-gene pairs by analysing correlations between this ATAC-seq and RNA-seq data, but a method was still required to parse these CREs from the background noise.

4.3.5 – Correlation of ATAC-seq and RNA-seq trajectories can be used to rank confidence in enhancers, but not silencers

Average ATAC-seq maximum coverage and RNA-seq TPM trajectories were standardised to the same scale by mean-centring and dividing by the standard deviation (Figure 4.12). To investigate which dCACRs might act as enhancer elements, a mean trajectory was calculated for each dCACR-DEG pair by taking the average of their standardised maximum coverage and TPM trajectories. The sum of squares distance between the standardised datapoints and the average trajectory was calculated for each of the seven timepoints. The mean of these sums of squares was then calculated as a metric of trajectory similarity. dCACR-DEG pairs with lower mean sums of squares have better correlations between ATAC-seq and RNA-seq data and we hypothesised that these are more likely to represent real enhancers. Across the 32,376 dCACR-DEG pairs, the mean of the mean sums of squares was 0.66 (Figure 4.13A).

To validate this approach, we generated 100 simulated datasets where each dCACR was randomly paired with a DEG and their mean sums of squares calculated as above. Averaging across these datasets, the mean of the mean sums of squares was 0.82 (1 SD = 0.002).

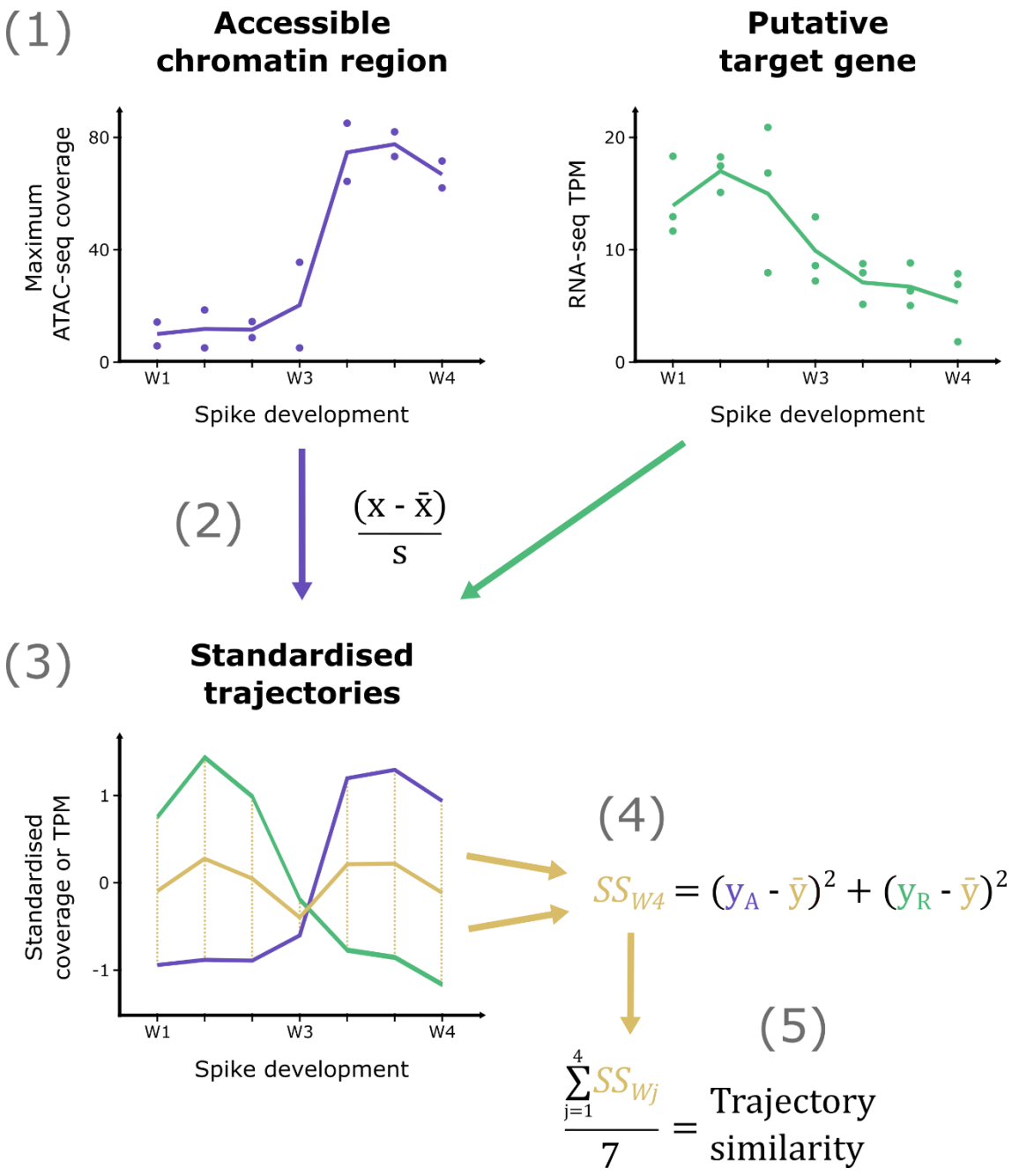


Figure 4.12 – A sums of squares approach for calculating trajectory shape similarity between ACR accessibility and target gene expression

(1) The maximum normalised ATAC-seq coverage found within a dCACR is calculated for each replicate (purple points). The average maximum per stage is then calculated (purple line). Similarly, for each gene an average TPM per stage (green line) is calculated from the three RNA-seq replicates (green points).

(2) The per-stage averages for each dCACR and gene are standardised by mean-centring and dividing by the sample standard deviation

(3) For each candidate dCACR-gene pair, an average of the standardised coverage and TPM trajectories is calculated (gold line).

(4) For each stage, the sum of squares (*SS*) distance between the average and coverage/TPM trajectories is calculated.

(5) The mean of these per-stage sums of squares is calculated as a metric of trajectory similarity

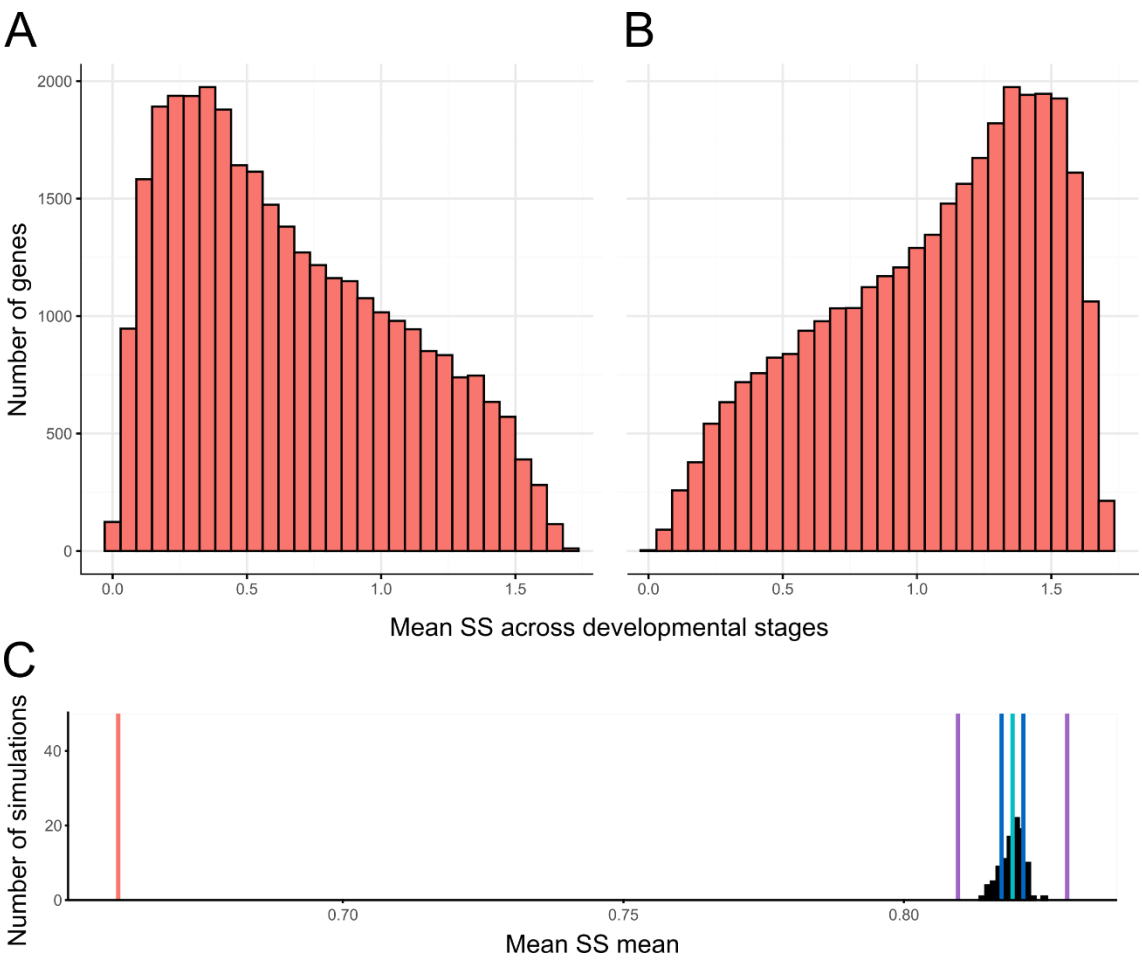


Figure 4.13 – Average sum of squares (SS) distances can be used as a metric of correlation between gene expression and chromatin accessibility
A, modelling dCACRs as enhancers produces a positively-skewed distribution of mean SS distances between profiles of DEG gene expression and neighbouring dCACR accessibility. **B**, modelling dCACRs as silencers produces a negatively-skewed distribution of SS distances between profiles of DEG gene expression and neighbouring dCACR accessibility. **C,** real DEG-dCACR pairs produce a lower average mean SS distance (red line) than randomly paired DEGs and dCACRs. The mean (light blue), mean ± 1 SD (dark blue), and mean ± 5 SD (purple) for the distribution of simulated average mean SS distances are marked.

The real result therefore lay -80 SD away from the mean result of the simulations, showing that real dCACR-DEG pairs produce tighter fits on average between chromatin accessibility and gene expression (Figure 4.13C). Together, these results suggest that the confidence of a particular dCACR acting as an enhancer for a neighbouring gene can be ranked against other pairs using this sums of squares approach.

In contrast to enhancers, this method did not appear suitable for ranking the confidence of dCACRs acting as silencers. For this test, the standardised ATAC-seq maximum coverage values were multiplied by negative one to invert the accessibility profile. If an inverted chromatin accessibility profile closely tracks the expression of a given gene, it can be inferred that the dCACR is acting to negatively regulate that gene – i.e. it is acting as a silencer. Using this approach, the mean of the mean sums of squares for the 32,376 dCACR-DEG pairs was 1.05. In comparison, for a set of 100 simulated datasets, the mean of the mean sums of squares was 0.89 (1 SD = 0.002), suggesting that the real dataset produced worse fits than randomly paired dCACRs and DEGs. Given the positive result for enhancers and the simple inversive transformation applied, this result was expected – the distribution of mean sums of squares for silencers must be a mirror image of that observed for enhancers (Figure 4.13B). In contrast, if we had observed a bimodal distribution, this would have suggested a capacity to detect both enhancers and silencers. This result is consistent with the fact that far fewer silencers have been documented to date in plants (Schmitz et al., 2021) and indeed are generally regarded as rarer than enhancers (Biłas et al., 2016). Real silencers are likely present in the data, but it appears that our methodology cannot parse these from background noise.

4.4 – Discussion

4.4.1 – Our Chinese Spring ATAC-seq data has limited utility

Overall, we found that our ATAC-seq data produced from pooled Chinese Spring spikes, spikelets, and carpels was of limited use. Our filtered reads did not strongly cluster into well-defined peaks and instead mapped at a low density across most of the genome, even in deep intergenic regions comprised almost exclusively of transposable elements (TEs). There is extensive evidence that, *in vivo*, such regions are largely bound into constitutive heterochromatin and therefore inaccessible to Tn5. This suggested that even though we obtained high numbers of visually intact nuclei via our protocol, the chromatin within was at least partially degraded at the point of tagmentation.

In contrast, Lu et al. (2019) produced high-quality data, with reads mapping largely within statistically significant ACRs. We believe that this could be due to the different methods used, as Lu et al. (2019) isolated nuclei from protoplasts obtained from fresh leaf tissue. As a result, there were no freeze-thaw cycles in their protocol. Conversely, given the time taken to obtain sufficient numbers of pooled spikes/carpels for each replicate, we found it necessary to collect spikes into super-cooled tubes to prevent severe tissue dehydration (which we presumed would be deleterious to chromatin structure). We cooled tubes by placing them in dry ice and periodically immersing them in liquid nitrogen. We hypothesise that including a freeze-thaw cycle, plus the temperature changes between dry ice and liquid nitrogen, may have damaged the chromatin within our tissue samples. Our protocol also differed in that we isolated nuclei directly from ground tissue, rather than via an intermediate protoplast stage.

We later observed that a concurrent attempt at wheat spike ATAC-seq (Lin et al., 2024) was able to produce high-quality data despite the challenges of collecting a pool of spike tissue for each replicate. This group collected ~20 spikes for each biological replicate across all stages and it appears that spikes were collected into tubes on water ice. The combination of a smaller maximum number of spikes (we collected up to 100 spikes into a single tube; Table 4.1) and placing tubes on ice appears to have been sufficient to prevent dehydration without freezing. It would be interesting to learn how quickly samples were able to be collected and by how many experimenters as further points of comparison. Another point of comparison is that the publication cited for their methodology used only ~5,000 nuclei per tagmentation reaction (Zhao et al., 2023), whereas we used up to ~60,000.

4.4.2 – Further adjusting ACR calling parameters could boost sensitivity

We were able to call high, consistent numbers of ACRs using the samples from Lu et al. (2019) and even greater numbers using the higher-coverage samples from Lin et al. (2024). Nonetheless, additional research suggested we could further optimise our scripts to boost the sensitivity of ACR detection. This is because, despite being highly popular for ATAC-seq peak calling, the software we used – MACS2 – was originally written for identifying peaks from ChIP-seq data and, accordingly, makes certain assumptions about expected read distributions (Yan et al., 2020).

Putative peaks are identified as regions with higher read density than the global background rate (calculated as total reads divided by the effective genome size). These are then tested against the local background rate in the control sample to see if the peak arose through sequence biases or is a true enrichment of ChIP-seq or ATAC-seq reads (Zhang et al., 2008). However, when no background control is supplied (as with our analysis of the data from Lin et al. 2024), MACS2 computes the local density from the target sample itself, using a 10 kb window centred on the putative peak. This is appropriate for ChIP-seq given the sharp peaks expected from this technique, but may be less so for ATAC-seq as ACRs can be much broader. Using the local 10 kb region as the background control might therefore prevent true ACRs from being detected, raising the false negative rate (i.e. lowering sensitivity). This could be alleviated by instructing MACS2 to use a broader area, say 100 kb, for its calculation of local read densities. Alternatively, more recent ATAC-seq specific tools, for example HMMRATAC (Tarbell & Liu, 2019), may also provide better sensitivity when no control data is available. This tool was previously computationally intensive and relatively difficult to install, but a HMMRATAC-based approach has now been introduced into the MACS suite (v3.0.0; October 2022).

A second limitation of our analysis is that we did not specify an effective genome size (EGS) specific to bread wheat. As described above, putative peaks are established by comparing a target region to the global background read density. This is itself calculated using the EGS; the total number of bases in the genome that are mappable and non-repetitive. MACS2 uses a default EGS of 2.7 Gbp (an estimate for the human EGS). Coincidentally, this is not a bad estimate for wheat given its high levels of redundancy due to polyploidy and TE proliferation: RefSeq v1.0 is 14.5 Gbp and estimated to be 85% repetitive DNA, giving a rough EGS estimate of 2.2 Gbp (Appels et al., 2018). The default EGS therefore likely represents a slight overestimate, which could potentially raise our false positive rate (i.e. lowering specificity)

by leading to a lower global background. Still, in order to be reported, putative peaks must also be significantly enriched versus the local background (either in the same sample or in a control file if available), so we anticipate the effects of this limitation to be minor. The other wheat ATAC-seq studies we investigated either did not report the EGS used (Lu et al., 2020; Lin et al., 2024) or used a much higher EGS similar in magnitude to the full genome size of wheat (17 Gbp; Concia et al., 2020), which we predict would have lowered ACR detection sensitivity. A more bespoke EGS parameter could be calculated by making a mappability mask via software such as GEM (Guo et al., 2012) or deepTools (Ramirez et al., 2016).

It is interesting to consider if more robust peak calling could have produced better results in our ATAC-seq – RNA-seq correlation analyses (which are discussed separately below). However, we believe the above alterations would be insufficient to increase the utility of our own ATAC-seq datasets given their highly noisy read distributions.

4.4.3 – Combining high-quality ATAC-seq and RNA-seq data may allow detection of enhancers specific to the target tissues and development stages

Our results suggest there is a limited capacity to identify CRE-gene pairs from equivalent time series of bulk RNA-seq and ATAC-seq data. Genes displaying differential gene expression over the time series (versus genes with constant expression levels) showed enrichment for nearby total accessible chromatin regions and, more strongly, differentially accessible chromatin regions. These results were supported by analyses of simulated datasets which showed that the real results lay well outside the distributions associated with the null hypothesis.

Our approach also provided some evidence for which dCACRs were most likely to be upregulating neighbouring gene expression. The accessibility and expression profiles of real dCACR-DEG pairs tracked each other more closely on average than those of randomly assigned dCACR-DEG pairs, with a positively skewed distribution of mean sum of squares values. This suggests that the best (lowest scoring) dCACR-DEG pairs are more likely to represent CREs that upregulate neighbouring gene expression. Setting an absolute level of confidence on these dCACR-DEG pairs by applying false-discovery rate statistics would be of future interest. In contrast, this method had no capacity to detect dCACRs which downregulate neighbouring gene expression. This could be due to silencer elements being generally rarer than enhancers in plants (Biłas et al., 2016).

The sums of squares analyses above assume a linear relationship between CRE chromatin accessibility and target gene expression. The biological reality may be more complex than this. For example, might a given two-fold increase in ATAC-seq coverage correspond to a step-change in TF binding that dramatically increases target gene expression? Could this kind of relationship be better described by alternative models, for example as a logistic relationship between chromatin accessibility and gene expression? Such hypotheses could be tested by applying various transformations to the input accessibility or expression data prior to calculating sums of squares.

However, we predict that any such modifications could only moderately improve the performance of this method given that the absolute differences in dCACR association between DEGs and constantly expressed genes were disappointing. For example, in the analysis that did not consider UMR support, DEGs were associated with 1.35 dCACRs on average, while constantly expressed genes were associated with an average of 0.84 dCACRs. This ~0.5 difference suggests that, on average, only one in every two genes has a single dCACR that is potentially causally associated with gene expression changes. Even if these true positives could be identified with total confidence, this drastically limits the utility of this approach for identifying CREs to use in synthetic promoters or to manipulate with gene editing approaches, as target genes may not be associated with any functional DACRs.

Another confounding factor is that many of the ACRs and dCACRs identified here may represent core promoter elements rather than tuneable CREs, as the core promoters of actively transcribed genes are usually highly accessible, with accessibility co-varying with gene expression (Schmitz et al., 2021). If the method had proved more promising, a follow-up experiment would be to exclude CACRs lying within around 100 bp of the parent genes' transcription start sites and recalculate the results accordingly.

Both of these limitations suggest that the input CACR-gene pairs need to be refined. Here, pairs were assigned based on genomic adjacency; each gene was assigned CACRs which lie within their transcribed regions or before their next upstream or downstream neighbour. However, lessons learned from existing resource-intensive epigenetic studies could improve upon this approach. For example, in maize it has been shown that enhancers preferentially occur upstream and adjacent (i.e. without intervening genes) of target genes because successive rounds of TE insertion have progressively separated them from their cognate genes (Lu et al., 2019; Ricci et al., 2019). Thus, while CREs *can* act independent of position and orientation according to accepted dogma (Schmitz et al., 2021), in practice they may exhibit stereotyped patterning in cereal genomes. A future strategy could therefore

be to focus solely on CACR-gene pairs in which the CACR is directly upstream of the potential target gene.

Interestingly, after we had completed our analyses, deployment of a very similar method in rice (*Oryza sativa*) to good effect was reported (Zhu et al., 2024). This study profiled chromatin accessibility and gene expression in three distinct cultivars for "23 distinct tissues" [sic; we count 25] from across the rice life cycle. Detected ACRs covered ~15% of the rice genome, notably higher than previous estimates in other plants such as maize (~4%) and Arabidopsis (~4%). Next, any ACRs within 20 kbp upstream or downstream of a gene's TSS were assigned as putative regulators. Pearson's correlation coefficients (r) were calculated for each ACR-gene pair across all expression/accessibility samples, and high-confidence links were defined as those with $|r| \geq 0.4$ and $p < 0.05$ based on permutation testing. This resulted in 59,075 ACR-gene links, which a variety of additional analyses suggest are indeed enriched for real CRE-gene interactions. For example, the ACR-gene set showed an overlap with published expression QTL-gene pairs that was significantly greater than expected by chance ($p < 2 \times 10^{-6}$).

Some of the differences between this study and ours would be difficult to adopt retroactively. For example, (Zhu et al., 2024) used an updated ATAC-seq methodology called UMI-ATAC-seq which improves read quantitation and TF footprinting (Zhu et al., 2020). Additionally, for defining candidate ACR-gene pairs, they set a specific genomic distance (± 20 kbp from TSS) based on the reported ranges of rice topologically associated domains (TADs, i.e. chromatin folding distances). The average size of TADs in wheat has been estimated, but is much larger (> 200 kb (Concia et al., 2020; Jia et al., 2021), potentially limiting its utility. This study also leveraged a much higher number of ATAC-seq/RNA-seq experiment pairs for their correlation analysis; 66 tissue/cultivar combinations versus our seven spike stages. Lastly, given rice's much smaller genome size, each replicate in (Zhu et al., 2024) was also sequenced to a much higher coverage relative to the target reference genomes (24.7x vs 4.2x). However, given that a large fraction of the wheat genome is comprised of highly inaccessible TE arrays, this may be a poor approximation of any functional difference in sequencing depth for the purposes of ATAC-seq. Calculating target sequencing depths on the basis of previously reported accessible genome fractions will help future studies achieve a good balance between sufficient coverage and unnecessary expenditure.

Given the data we have available, though, it would be of future interest to apply their simpler correlation methodology in favour of our sums-of-squares model. We are unclear, however, how their cut-off value for the Pearson's correlation coefficient was reached and whether

this would be appropriate for our data. Despite this method's utility, it does still have limitations. Only ~48% of annotated genes were assigned an ACR, which the authors attribute to a lack of variation in the accessibility of housekeeping genes' regulatory elements, plus the limitations of assuming that gene expression scales linearly with CRE accessibility. The reality is likely much more complex, especially because most genes are regulated by multiple CREs simultaneously. This limitation also applies to our method, as previously discussed. Despite this, the use of a similar method (albeit using DNase-seq) across many human cell-types associated 98% of genes with at least one ACR and yielded a median of 19 ACRs per gene (Sheffield et al., 2013), suggesting greater scope for this class of methods than has so far been demonstrated in plants. Lastly, this paper only describes enhancers, with no mention of silencers, despite allowing for negative correlations in their model. It is thus unclear if silencers were found but not described explicitly, or whether this approach failed to detect them, like our own.

Overall, then, while not all of the modifications presented by (Zhu et al., 2024) are applicable to the data we utilised or to wheat more generally – and noting that the approach still exhibits some limitations – this work suggests that our general method for cataloguing the wheat spike *cis*-regulome is a viable one which could be refined in the future.

Alternatively, the use of additional data types could be used to refine the input ACR-gene candidates. For example, Hi-C could be used to confirm which genes each ACR physically interacts with. However, gathering such additional data can be prohibitively expensive, time-consuming, and, in the case of small tissues, technically difficult. To partially overcome this, one could explore the overlap between CREs identified in leaf Hi-C data and those found in spike ACR data. Histone modification data (via ChIP-seq) is another option for linking CREs and genes, as regions that are brought into spatial proximity often share chromatin characteristics (Wang et al., 2021). The ATAC-seq and RNA-seq datasets we utilised was published alongside histone modification data which could be used for this purpose (Lin et al., 2024). This comprised data for five histone marks (H3K4me3, H3K27ac, H3K27me3, H3K36me3, H2A.Z.) across five developmental stages (SAM, W2.5, W3, W3.25, W4). This was achieved using 'CUT&Tag', a method for epigenomic profiling in small tissue samples (Kaya-Okur et al., 2019). Another advantage of Hi-C and histone modification data is that they would allow detection of (potentially rarer) cases where ACRs interact with genes separated by one or more intervening genes. These methods and others (plus their use in combination), are rapidly improving in viability and utility in plants. Amongst other topics spanning this thesis, the next chapter will discuss further options for cataloguing genome-wide spike CREs that may become available in the near future.

4.5 – Methods

4.5.1 – Plant materials and growth conditions

Hexaploid bread wheat (cv. Chinese Spring) seeds were stratified (4 °C) and germinated on Petri plates in the lab, then grown on in soil in a controlled environment room under a long day photoperiod (16 h / 8 h; 20 °C / 15 °C; light / dark) at 300 µmol m^{-2} s^{-1} incident radiation and 60% humidity. The first four stages (W1, W2, W2.5, W3.25; Figure 4.2, Table 4.1) were collected from plants in ~75 mL cells. Later stages were collected from plants potted on after 21 days into 1 L pots. The media used was 'John Innes Cereal Mix' (65% peat, 25% loam, 10% grit, 3 kg m-3 dolomitic limestone, 1.3 kg m−3 Yara PG MixTM 14-16-18, 3 kg m-3 Osmocote® Exact).

4.5.2 – Tissue collection

Tissue for ATAC-seq and RNA-seq was collected as described previously (Faci, Backhaus, et al., 2024). Briefly, tissue was collected from the main tillers only using a dissecting microscope and ophthalmic microsurgery knives. Equipment and gloves were frequently sanitised using Blitz RNase removal spray (Severn Biotech Ltd.) and always between dissecting different genotypes or stages. Samples were collected into tubes on dry ice and placed into liquid nitrogen at intervals of no more than 20 min to ensure samples remained completely frozen. Completed samples were stored at -70 °C.

When harvesting spikes and spikelets (stages W1 to W5), care was taken to ensure no vegetative, non-spike tissue was captured. When harvesting carpels (stages W6 to W10), care was taken to ensure no glume, lemma, palea, lodicule, or stamen tissue was captured. Carpels were only harvested from florets 1 and 2 and only from the middle ~75% of spikelets.

Initially, ten samples from various stages were collected for initial ATAC-seq trials (Table 4.2). Later, five replicates were collected for each developmental stage for both ATAC-seq and RNA-seq. For most stages, enough tissue was collected in a single tube to provide one RNA-seq and one ATAC-seq replicate, with tissue divided after homogenisation (~1:2 ratio). We hoped that this 'pairing' of samples could improve our ability to detect correlations between the two types of data. Tissue was collected separately, however, for the five stages with the smallest input tissues (W1, W2, W2.5, W6, and W7) as these were too difficult to divide accurately after homogenisation.

4.5.3 – ATAC-seq library preparation, RNA extraction, and sequencing

Samples for ATAC-seq were homogenised by bead beating using a TissueLyser II bead mill (QIAGEN; 1,650 rpm, 30 s). TissueLyser adaptor blocks were supercooled prior to use. Nuclei were extracted and checked for quality and quantity as described previously (Faci, Jones, et al., 2024). Tagmentation was then performed using reagents from the Illumina Nextera DNA Library Prep Kit (FC-121-1030/15028212) but with a custom protocol, see Supp. Note 4.1. Libraries were stored at -20 °C until sequenced.

Samples for RNA extraction were homogenised as above, except that two rounds of bead beating were conducted. Total RNA was extracted using Direct-zol RNA Microprep (Zymo Research) kits as described in the manufacturer's manual, except that RNA elution was conducted twice, with 14 µL nuclease-free water each time. Total RNA was stored at -70 °C until sequenced.

All sequencing was conducted by Novogene UK using an Illumina NovaSeq 6000 to produce 150 bp paired-end reads. We requested 50 million paired reads for both our ATAC-seq trial samples and our RNA-seq samples. We requested 200 million paired reads for our final ATAC-seq samples.

4.5.4 – RNA-seq data analysis

Raw RNA-seq reads were trimmed using fastp (v0.23.1; Chen et al., 2018) and quality assessment was conducted before and after trimming using FastQC (v0.11.8; Andrews, 2010). Default settings were used except that '--detect_adapter_for_pe' was specified. Kallisto (v0.44.0; Bray et al., 2016) was then utilised to quantify transcripts, supplying a k=31 Kallisto index of the IWGSC RefSeq v1.1 gene annotations (Appels et al., 2018). Additional options specified were '--bias' (enables sequence-based bias correction) and '--bootstrap-samples=30'. TPM and read count values were summed across annotated alternate transcripts to produce a single figure per gene.

PCAs were conducted using only HC genes with non-zero counts in at least one target sample. Read counts were transformed using the rlog transformation from DESeq2 (v1.34.0; Love et al., 2014) (to prevent highly expressed genes from having a disproportionate effect) and PCs calculated from these using the base R function prcomp (R Core Team, 2023). Gene trajectories were analysed for significant deviation from a flat intercept using ImpulseDE2 in "case-only" mode (v1.10.0 using R v3.6.1; Fischer et al., 2018), setting a significance threshold of padj < 0.001 (Benjamini–Hochberg corrected).

4.5.5 – ATAC-seq data analysis

Raw ATAC-seq reads were trimmed and assessed for quality as above, except that reads shorter than 38 bp after trimming were removed (--length-required 38). Surviving reads were mapped using Bowtie2 (v2.4.1; Langmead & Salzberg, 2012), with '--maxins' set to 750 and mapping effort set to '--very-sensitive'. Mapped reads were then filtered to retain concordant pairs (-F 12, -f 3) exceeding a minimum MAPQ score (-q 30) using Samtools (version 1.12, Li et al., 2009). Reads mapping to unmapped scaffolds ('ChrUn') or to the mitochondrial or chloroplast genomes were removed. Reads were then sorted, deduplicated, and indexed using Samtools. These BAM files were then converted to BED files using BEDTools (v2.29.2; Quinlan & Hall, 2010). This ensures both reads of each pair are utilised independently for peak calling (Delisle et al., 2019). Peaks were called using MACS2 with the following options: --format BED, --nomodel, --extsize 150, --shift -75, --keep-dup all (v2.2.7.1; Zhang et al., 2008; Gaspar, 2018; Delisle et al., 2019). We controlled for background Tn5 insertion rates in our lower coverage tests by supplying MACS2 with the naked DNA leaf sample from (Lu et al., 2020). For analysis of our higher coverage samples, we supplied our own four naked DNA samples (cv. Chinese Spring). For analysis of the previously published Kenong 9204 data (Lin et al., 2024), we did not supply a background control.

To make visualisation comparable between samples, we used BEDTools to calculate read coverage per base normalised against final sequencing depth using the following formula:

$$normalised\ coverage = coverage \times \frac{100{,}000{,}000}{final\ read\ pairs \times 2}$$

The resulting bedGraph files were sorted and converted to bigWig binary format using tools written by W. J. Kent (Kent et al., 2010). Normalised coverage was then visualised using the Integrative Genomics Viewer (IGV; Robinson et al., 2011). All coverage tracks were set to the same height and data scaling within any given visualisation.

4.5.6 – Correlation of ATAC-seq and RNA-seq data

A set of CACRs was generated using the ACRs already called from Lin et al. (2024)'s data. CACRs were defined as regions containing a peak in all stages between W1 and W4 and were computed using BEDTools multiinter and BEDTools intersect. Maximum coverage per CACR for each replicate of each stage was appended to these BED files using BEDTools intersect. HC gene coordinates were extended upstream and downstream to the next adjacent genes using a custom script in R (v4.1.3; R Core Team, 2023), then separate BED

files were generated for each gene. See associated script 04_extend_gene_coords.r for methods used to deal with edge cases (e.g. first gene on chromosome, e.g. preceding gene ends inside current gene). CACRs were assigned to genes based on intersection with these single gene BED files using BEDTools intersect. dCACRs were called using ImpulseDE2, setting a significance threshold of padj < 0.05 (Benjamini–Hochberg corrected). Enrichment analyses and dataset simulations were conducted in R.

To generate a set of wheat UMRs, we obtained a published set of leaf WGBS data (Appels et al., 2018) and called UMRs as described previously (Crisp et al., 2020; Eglitis-Sexton et al., 2024). We used the upper of the two suggested thresholds for a 100 bp tile to not be counted as missing data (minimum of 5x coverage per cytosine).

Correlation of ATAC-seq maximum coverage and RNA-seq TPM trajectories, including normalisation of data types to the same scale and implementation of the sums of squares-based method, was conducted using custom BASH and R scripts. See accompanying GitHub repository for custom scripts: https://github.com/maxrwjones/wheat_spike_ATACseq_RNAseq_correlation.

5 – General Discussion

5.1 – Thesis summary

In this thesis, we have explored how different types of variation can be utilised to improve inflorescence traits in cereal crops. Such variation may be any combination of natural or engineered, dominant or recessive, and coding or *cis*-regulatory. We first uncovered natural variation by conducting a *k*-mer-based GWAS (kGWAS) study on a tef germplasm collection (Chapter 2). We identified 26 marker-trait associations across 10 morphological traits and grain metabolites. Of these, we noted recurring co-localisation of grain colour and grain size associations, matching our earlier observation of a strong correlation between brown colour and larger grains in the BLUPs. This finding has strong relevance to tef breeding given that white grain colour and larger grain size are both considered desirable traits. Within two homoeologous marker-trait associated regions, we identified the tef orthologues of *TRANSPARENT TESTA 2* (*TT2*) and proposed these as candidate drivers of grain colour and size variation given their documented roles in other species. By comparing published assembled genomes of a brown and a white-grained cultivar, we identified coding sequence variation in these genes in the form of TE-induced nonsense mutations. We predict the mutation for white grain colour is recessive given that it putatively causes a loss of protein function. However, given tef's tetraploid nature, it will be interesting to examine whether the loss of one, two, or three gene copies will show dosage responses or complete functional redundancy. We also discovered considerable redundancy in the studied collection and developed a compact SNP panel that can be used to uniquely identify each distinct accession.

In contrast to this natural, presumably recessive, coding sequence variation, in Chapter 3 we attempted to introduce engineered, dominant, regulatory variation to manipulate the phenomenon of low basal spikelet productivity in wheat. We hypothesised that, given the narrow range of natural variation for this trait, we could raise basal spikelet productivity by alleviating the spatial differences in expression of genes promoting the spikelet meristem to floral meristem transition. To achieve this, we developed a semi-spatial transcriptomics time course of early wheat spike development. This allowed us not only to identify candidate regulators of low basal spikelet productivity, but to select additional genes with desirable expression patterns from which to develop specialised promoters (or, more precisely, regulatory environments) for basal spike-specific misexpression. Preliminary results suggest our misexpression of *SEPALLATA1-B6* (*SEP1-B6*) may reduce the number of rudimentary basal spikelets (RBS) versus wildtype, though this will need to be confirmed by repeat experiments under various conditions. Overall, our combination of semi-spatial transcriptomics and ATAC-seq data has proved a useful tool for hypothesis generation.

Using these data, we have already supported additional projects including cellular resolution spatial transcriptomics of the developing wheat spike (Long et al., 2024) and multiplexed genome editing of domestication gene *cis*-regulatory elements (CREs).

One of the regulatory environments we developed, BSS1, did not drive strong reporter expression in any tissue examined, while BSS2 produced expression in the peduncle in addition to our target tissue. This highlighted the importance of generating a repertoire of reliable promoters or regulatory environments for use in fundamental wheat research. In turn, this suggested the need for methods to characterise the wheat *cis*-regulome across various tissues and developmental stages in order to prioritise candidates for regulatory environment development. In Chapter 4 we experimented with one such approach – correlation of chromatin accessibility and gene expression data – aiming to catalogue the *cis*-regulome of the developing wheat spike. While our own ATAC-seq data was of limited utility, we leveraged equivalent data from collaborators (Lin et al., 2024) to show enrichment of accessible and differentially accessible non-coding regions in the genomic neighbourhood of differentially expressed genes (versus constantly or non-expressed genes). We extended this to scoring correlations between accessibility and expression profiles with the aim of detecting functional CRE-gene pairs. Our method showed a limited capacity to detect enhancers, but not silencers, though this is consistent with the lower prevalence of the latter in plant genomes. More complex correlation models and refinement of the input CRE-gene candidates could improve the utility of this approach.

Throughout this thesis, we have explored methods to identify, create, and utilise variation for inflorescence traits in polyploid cereals. Despite our work being relatively fundamental, we believe our findings and methods may indirectly be of use for future breeding efforts in wheat, tef, and other cereal crops. In the remainder of this chapter, we examine future prospects for the above research areas and discuss some broader themes.

5.2 – Studying natural variation without and beyond reference genomes

Studying natural variation in germplasm collections is a time-tested method for identifying loci and genes involved in developmental processes. In Chapter 2 of this thesis, we uncovered novel loci associated with tef inflorescence and grain variation using kGWAS. With the ever-declining costs of sequencing, particularly short read sequencing-by-synthesis (Fletcher, 2025), kGWAS is becoming appropriate for an increasing range of research scenarios. This will soon extent to analysing complete bread wheat genomes, as discussed below, and could therefore also be leveraged to investigate the traits discussed in Chapter 3.

Firstly, for the many orphan or underutilised crops now being studied through the lens of modern genomics (reviewed by (Shorinola et al., 2024) and others), sequencing can bypass the costly and lengthy development of genotyping platforms (e.g. SNP arrays), as exemplified by our own work. Reads are decomposed into *k*-mers and the presence or absence of each are used directly as markers. Where SNP data has other specific utilities – such as for phylogeny reconstruction – SNP sets can still be developed *in silico* from mapped sequencing reads according to various parameters such as linkage pruning and data missingness. Such a phylogeny proved highly valuable in our work, as we were able to identify considerable redundancy within the EIAR tef core collection (Chapter 2.3.1). For future studies investigating germplasm collections which have not previously been genetically characterised, we recommend that these sequencing and *in silico* genotyping steps are conducted well before phenotyping to avoid duplication of phenotyping effort. For example, this could be carried out on leaf tissue from the individual plants being grown for seed bulk after single seed descent. Collection of high-quality phenotyping data across multiple sites and/or years is often a limiting factor for the power of GWAS studies, so such savings could have a tangible effect on the quality of research outputs.

Next, kGWAS can be used to identify novel trait-associated loci not present in reference genomes, e.g. due to INDELs or larger structural variations (Rahman et al., 2018). While simpler implementations of kGWAS only ask “which accessions have each of the *k*-mers found in the reference genome?”, the approach can be expanded to explore “which accessions have each and every *k*-mer found across all accessions in the study?” More complex pipelines can also utilise *k*-mer counts to infer copy number variation (Rahman et al., 2018; He et al., 2024). Though we did not explore this capacity in the work presented here, we anticipate it would uncover additional loci of interest, particularly as we have

already identified a putative large introgression within our panel that is not present in the reference genome cultivar (Chapter 2.3.1, Figure 2.5, Figure 2.6).

This faculty also lends kGWAS to the study of highly diverse collections of well-studied species such as wheat. The A. E. Watkins collection comprises 827 bread wheat landraces collected from 32 countries in the 1920s and 30s. Resequencing and analysis of this collection has revealed that the haplotypes within modern bread wheat varieties are largely derived from just two of seven 'ancestral groups' (Cheng et al., 2024). Mining such a collection through SNP GWAS would therefore likely miss many key loci, even if read mapping was conducted against many assembled varieties from the growing wheat pangenome (Walkowiak et al., 2020; Jiao et al., 2025), as these too are almost exclusively derived from the same two ancestral groups. In contrast, kGWAS could compare presence and absence of *k*-mers not found in a given reference genome, ensuring more genetic variation is analysed against the phenotypic data. To date, kGWAS has not been conducted on any bread wheat population, largely because of the very high storage and processing requirements to build and handle the core *k*-mer matrices. However, the JIC Informatics team has recently completely recoded the kGWAS pipeline we utilised for our study (Gaurav et al., 2022), realising efficiency gains that make conducting kGWAS on the Watkins collection sequencing data feasible (Dr Burkhard Steuernagel, personal communication). We will be particularly interested to see if this approach reveals novel variation for basal spikelet productivity or other deviations from lanceolate spike shapes given their scarcity in modern germplasm (Philipp et al., 2018).

Going further, kGWAS can also be conducted entirely in the absence of a reference genome ('reference-free'). This again has clear applicability for orphan crops, many of which do not yet have high-quality reference genomes. Still, this advantage may be short-lived, as the generation of chromosome-scale reference genomes for orphan crops is accelerating (Chapman et al., 2022; Shorinola et al., 2024), with recent additions including lablab (Njaci et al., 2023), grasspea (Edwards et al., 2023; Rajarammohan et al., 2023), and guar (J.-H. Li et al., 2024). This is due to a plethora of factors, including better bioinformatic tools for genome assembly, wider familiarity with their use, rising global interest in crop diversity, and the falling cost, better accuracy, and higher read lengths of long-read sequencing technologies such as Oxford Nanopore and PacBio circular consensus sequencing (CCS), plus auxiliary scaffolding methods such as Hi-C.

Accordingly, pangenomes are beginning to emerge for many hitherto underutilised or under-researched crops, inviting us to consider whether these designations remain appropriate. Examples include sesame (Yu et al., 2019), pigeon pea (Zhao et al., 2020), white lupin

(Hufnagel et al., 2021), foxtail millet (He et al., 2023), pearl millet (Yan et al., 2023), broomcorn millet (Chen et al., 2023), cowpea (Liang et al., 2024), common bean (Cortinovis et al., 2024), peanut (K. Zhao et al., 2025), and lablab (Chapman, 2025). Partial assembly of non-reference sequences has also been conducted for chickpea (Varshney et al., 2021), mung bean (C. Liu et al., 2022), and *Pisum* (Yang et al., 2022). As part of a wider collaboration, we (the team from (Jones et al., 2025) have also begun working in collaboration with researchers from Sant'Anna School of Advanced Studies (Italy) on a tef pangenome to bring the benefits of pangenomics (Chapman et al., 2022; Tay Fernandez et al., 2022; Schreiber et al., 2024; Hu et al., 2025) to this valuable orphan crop (data not shown). Of particular relevance to this discussion are the insights this will bring into the prevalence and diversity of structural and copy-number variation across the diversity of genebank tef varieties. Capturing intraspecific diversity through pangenomics is also important for genome editing, as presence-absence and copy-number variation can inform the editing strategy and allows the design of cultivar-specific sgRNAs.

5.3 – Transgenesis is an essential research tool, especially in polyploids, but its use remains constrained in many crops

Once a promising locus correlated with natural variation is identified, there are multiple methods available to plant scientists to validate causality, including allele introgression using marker-assisted selection or examining mutant collections for suitable knock-out lines. However, for polyploid species, these approaches require multiple generations to combine alleles across homoeologs in addition to the backcross generations needed to remove additional, confounding variation. This greatly lengthens the time horizons needed for validation (Uauy et al., 2017). Transformation-based approaches offer more precise, potentially faster, solutions. Genome editing can be used to knock-out multiple homoeologs simultaneously, RNAi or VIGS can be used for milder knock-down across homoeologs, and transgenic complementation of existing mutants can be used for further validation (Adamski et al., 2020). For example, in Chapter 2.3.6, we proposed that *TT2*'s potential role in tef grain colour, size, and EPOD fatty acid accumulation, could be tested by mutagenising this pair of homoeologs in a brown-grained cultivar via CRISPR-Cas9. Genome editing even offers the possibility of direct allele replacement in crop species via techniques such as prime editing, as demonstrated by (Y. Zhao et al., 2025), who exchanged the $P1^{WT}$ allele in the bread wheat cultivar Fielder for the $P1^{POL}$ allele from *Triticum turgidum* ssp. *polonicum*.

An extension of the above is the application of crop transformation to recapitulate known natural variation – for example from landraces or crop wild relatives – directly in breeding-compatible material. This again bypasses the lengthy process of introgression and backcrossing, but furthermore prevents linkage drag issues where the target beneficial variation and adjacent deleterious variation are rarely separated by recombination. A clear application of this approach is the recreation and stacking of disease resistance into elite material where the underlying '*R*' gene(s) are known (Dracatos et al., 2023). This extends to 'transferring' resistance between species that do not readily hybridise. For example, the multipathogen resistance gene *Lr34* from bread wheat has been transgenically introduced into barley (Risk et al., 2013) and rice (Krattinger et al., 2016). Some types of variation can also be recapitulated by genome editing, which may be more appropriate in some jurisdictions. For example, the UK recently passed secondary legislation for the 'Genetic Technology (Precision Breeding) Act', which allows genome edited crops to be grown and sold in England, while transgenic plants are still prohibited outside of research contexts. As an aside, cisgenic plants are also now permitted, allowing gene transfer approaches to be

used on species which can hybridise. Based on this, work at The Sainsbury Laboratory (UK) is aiming to stack *R* genes from compatible wild relatives into elite potatoes (Prof. Jonathan Jones, personal communication). Researchers at the JIC are already using genome editing to recreate and fix key wheat domestication alleles in select lines of the Watkins collection to increase their utility in pre-breeding pipelines (Dr Simon Griffiths, personal communication). These include manipulating *Rht1* for semi-dwarfism and *Tamyb10* (historically known as *R*/*r*) for reduced pre-harvest sprouting. Notably, these approaches hinge on the ability to efficiently transform a diverse range of target germplasm, discussed further below.

Transformation also allows researchers to ask and answer questions not possible with known natural variation. In Chapter 3, we investigated the phenomenon of low basal spikelet productivity in wheat. As previously discussed, little genetic variation altering the distribution of grain productivity along the spike axis – which produces its characteristic lanceolate shape – has been identified in bread wheat (Philipp et al., 2018). We therefore utilised novel regulatory environments to drive specific overexpression of *MORE FLORET 1* (*MOF1*) and *SEP1-6* in the base of the wheat spike, aiming to amend their weaker expression in the basal spikelets versus the central (Chapter 3.3). We predicted that this gain-of-function manipulation would produce dominant variation, allowing us to observe phenotypes in hexaploid wheat immediately in T_0 and T_1 plants. An alternative transformation-based approach could have been to try and *reduce* these genes' expression in the central spikelets. This could potentially be achieved by RNAi techniques, but, as this would be a loss-of-function change, it may have been necessary to target all six native gene copies.

Given the utility of transformation to developmental studies and breeding, both are limited where crop transformation cannot be achieved regularly, cheaply, and by many independent research groups. By 'transformation', we have so far meant stable (i.e. inherited) transformation, rather than transient. Stable transformation of plants can be achieved in several ways, including *in planta* transformation of reproductive tissue by biolistic bombardment (e.g. (Hamada et al., 2018) or exposure to *Agrobacterium* via immersion or injection – an example of the latter being Arabidopsis floral-dip (Rafiei et al., 2024). However, the most popular and efficient methods for cereals are based on *Agrobacterium*-mediated transformation of excised tissue followed by regeneration in tissue culture.

Stable tef transformation was first reported in 2013 through the use of *Agrobacterium* and tissue culture (Gebre et al., 2013), but has not become routine. One obstacle is that grain-

derived immature embryos – which show high regeneration efficiency in other cereals – are too small for mechanical isolation in tef (Hayta, 2023). Recently, two groups at IBERS (UK) and NIAB (UK) have been developing higher efficiency protocols for tef transformation with some success (Dr Aiswarya Girija and Dr Stéphanie Swarbreck, respectively, personal communication). The latter has focused efforts on the white-grained cultivar Tsedey to integrate resources for molecular characterisation of tef, as a draft genome (Cannarozzi et al., 2014) and a TILLING mutant collection (Cannarozzi et al., 2018) are already available for this cultivar. We are also aware that an updated, chromosome-scale assembly for Tsedey has been produced by Corteva Agriscience (Prof. Zerihun Tadele, personal communication), though the company has not made their publication intentions known. We hope that these more recent efforts lead to efficient, reproducible, and open-access protocols for tef transformation.

Even for wheat, one of the three major global cereals, stable transformation remains expensive (£6,000-12,000 in the UK; Prof. Cristobal Uauy, personal communication) and the province of a small number of highly specialised labs. This is primarily because most donor tissues are highly recalcitrant to tissue culture, necessitating a year-round supply of immature embryos, which is difficult and costly to maintain (Hayta, 2023; Rafiei et al., 2024). Additionally, most transformation and regeneration protocols are highly genotype-dependent, allowing only certain cultivars to be transformed with reasonable efficiencies (Hayta, 2023; Rafiei et al., 2024). For example, (Hayta et al., 2021) reported a 33% transformation efficiency for Fielder, but only 10% for Kronos and 4% for the bread wheat cultivar 'Cadenza', even with additional protocol modifications.

Nonetheless, great progress has been made in the last five years to overcome these strong tissue and genotype dependencies of wheat transformation protocols. Including a constitutively expressed chimeric *GRF4-GIF1* gene in transformation vectors partially overcomes existing protocols' issues with genotype-dependency, raising transformation efficiencies to acceptable levels for direct genome editing in diverse cultivars (Debernardi et al., 2020; Hayta et al., 2021; Qiu et al., 2022; Biswal et al., 2023). Use of this fusion protein has also improved regeneration efficiencies for other recalcitrant species including triticale (Debernardi et al., 2020) and sorghum (J. Li et al., 2024). A limitation of this system is that it pleiotropically affects phenotypes beyond callus regeneration. The original *GRF4-GIF1* study noted reduced grain number per spike and increased grain weight (Debernardi et al., 2020), while (Chen, 2023) noted effects on heading time, spike morphology, and grain width, plus low seed set. This may be particularly detrimental to the study and manipulation of inflorescence development, so would necessitate either backcrossing rounds for transgene

segregation or the experimental inclusion of unedited sister lines or 'empty vector' transformants (with only *GRF4-GIF1* and no other transgenes) as additional controls. (Chen, 2023) suggests that assuring initial vector copy numbers are kept low will be an essential feature for widespread adoption of the system.

Progress has also been made in reducing the reliance on immature embryos. Recently, (Ye et al., 2023) reported transformation efficiencies >7% on mature embryos from field-grown seed by centrifugation-assisted *Agrobacterium* inoculation. While these findings will need to be replicated by other research groups, this method appears promising and may be further improved in combination with the *GRF4-GIF1* system. In other cereals, including tef, advances have also been made in the use of leaf fragments as explants through the use of two additional morphogenic regulators, *Wuschel2* (*Wus2*) and *Baby Boom* (*Bbm*) (Wang et al., 2023). While they failed to regenerate bread wheat using their method, it is possible that the use of a *GRF4-GIF1* construct in a similar protocol could yield better results. Advances in reducing tissue specificity may also come from combining *GRF4-GIF1* with *Wus2*, *Bbm*, and/or ternary helper vectors (Chen et al., 2022; Vandeputte et al., 2024).

Improved systems for transient transformation of cereals would also be of great value for in-depth characterisation of protein localisation and interactions. Current alternatives for protein characterisation include *in vitro* pull-downs, yeast two-hybrid assays, or experiments in transiently transformed eudicot systems such as *Nicotiana benthamiana*, including subcellular localisation of fluorescently tagged proteins, luciferase or bimolecular fluorescence complementation, and Co-IP. Combining these methods can produce high-quality evidence for protein behaviour, as demonstrated for regulators of *VRT-A2* by (J. Liu et al., 2025). Nonetheless, the results from such techniques may not be representative of protein behaviour in the original cereal, for example because of missing cofactors. Options for transient transformation of cereals include *Agrobacterium* infiltration, biolistic bombardment of attached leaf tissue, and protoplast transformation, but each is currently limited by either protocol complexity and/or low efficiency (Hensel et al., 2011a; Kirienko et al., 2012; Miller et al., 2023). Agroinfiltration has been successfully expanded to a range of eudicot species (Hoshikawa et al., 2019; Suzaki et al., 2019; Zhang, Chen, et al., 2020), and it would be highly beneficial to adapt this relatively simple transformation method to cereals. It may, however, be necessary to engineer model cereal lines for enhanced receptivity to agroinfiltration, for example by reducing deposition of epidermal cuticular wax, lowering silica content, increasing the volume of intercellular space, and introducing autonomous production of *vir* gene-inducing phenolics such as

acetosyringone (Smith & Hood, 1995; Andrieu et al., 2012; Hwang et al., 2017; Sharma et al., 2020).

5.4 – Deep characterisation of crop *cis*-regulomes is rapidly advancing

As we argued previously (Chapter 3.4.4, Chapter 4.2), genome-wide catalogues of CREs in tissues and developmental stages of interest can greatly support transgenic or genome editing interventions for developmental studies. In Chapter 4, we explored one approach for defining a wheat spike *cis*-regulome, namely correlation of differential gene expression with differential chromatin accessibility of adjacent regions. We concluded that our approach had some capacity to detect enhancer-gene pairs, though we would need to explore statistical approaches for defining confidence in these links. In contrast, we could not detect silencer elements with our method. We also reviewed the implementation of a similar method in rice (Zhu et al., 2024) in Chapter 4.4.3. Overall, we concluded that while this class of methods has strong potential for the detection CRE-gene pairs, our own implementation was limited by the number of sample types, the assumption of a linear response between CRE accessibility and gene expression, and the choice of input ACR-gene candidates.

The latter issue could be overcome through the use of additional data types. We covered how the use of Hi-C or profiling of correlated histone modifications could be used to refine ACR-gene candidates, though noted the difficulty and cost of applying these methods to a diverse range of often tiny tissues. Still, during the course of this PhD, global interest in producing wheat 'omics' data has risen dramatically (Yao et al., 2025; Zhang et al., 2025), while the cost of methods like Hi-C has declined (N. Liu et al., 2021) due to methodological refinements, proliferation of commercial kits, and ongoing decreases in sequencing costs. We therefore anticipate that reliable catalogues of wheat CRE-gene interactions – including those relevant to wheat spike development – will soon be generated using combinations of the above methods and their single-cell and spatial counterparts (X. Liu et al., 2025; Nobori, 2025).

The continued adoption of newer methods for CRE exploration in plant and crop research makes achieving this goal ever more likely. For example, there is increasing use of nascent transcription profiling methods in plants, including GRO-seq/PRO-seq, plant NET-seq (pNET-seq), and CB RNA-seq (Hetzel et al., 2016; Zhu et al., 2018; Qin et al., 2022; Liu et al., 2023). These methods seek to profile RNA molecules as they are transcribed, allowing the detection of unstable, short-lived RNA species such as enhancer-derived RNAs (eRNAs) that are not detectable in typical steady-state RNA-seq experiments (Oka et al., 2017). eRNA levels appear to more accurately predict enhancer activity than chromatin

accessibility or histone modifications in animals (Arner et al., 2015; Henriques et al., 2018) and possibly in plants (Xie et al., 2022; Zhu et al., 2025). Therefore, while they do not link CREs to genes, they could provide supporting information to facilitate triaging of tissue-specific CREs for mutagenesis by GE or for designing novel promoters. eRNA profiling has been reported for wheat leaves (Xie et al., 2022), so is likely feasible in additional tissues. Contrastingly, others have suggested that eRNAs are rare in plants and may not be useful for studying enhancers (Hetzel et al., 2016; Zhu et al., 2018; McDonald et al., 2024).

Another class of techniques which are increasingly used in plant research seem ideally placed for the characterisation of CRE-gene pairs. A profusion of protocols have been developed to simultaneously detect and contact-map accessible chromatin regions (ACRs), combining the virtues of chromatin accessibility profiling and chromatin conformation capture techniques. These include OCEAN-C (T. Li et al., 2018), trac-looping (Lai et al., 2018), Hi-CoP (Zhang, Li, et al., 2020), HiCAR (Wei et al., 2022), ChIATAC (Chai et al., 2023), and TAC-C (Kang et al., 2025). Two studies have conducted OCEAN-C on leaf tissue of bread wheat, durum wheat, and *Aegilops tauschii*, providing novel insights into chromatin interactions between subgenomes and how polyploidisation sets up homoeolog biases, including biased levels of expression, accessibility, and distal interactions (Yuan et al., 2022; Jiao et al., 2024). TAC-C was developed in cereals and has been tested on rice, sorghum, maize, and wheat, with data from the latter suggesting a key role for SBP family TFs in stabilising chromatin loops, which was functionally explored with a *spl7 spl15* double mutant (Kang et al., 2025). Another fascinating insight was that genes with crucial roles in C_4 photosynthesis showed higher loop frequencies in the two C_4 plants studied (maize and sorghum) versus the two C_3 plants (wheat and rice), pointing to the increased requirement for coordination of their transcription (Kang et al., 2025).

As multi-omics data has risen in complexity, researchers across biology have increasingly turned to machine learning approaches, including large language models, to integrate and draw inferences from different data sets and data types (Kang et al., 2022; Li et al., 2022; Babu & Snyder, 2023; Feldner-Busztin et al., 2023; Acharya & Mukhopadhyay, 2024; Panahi et al., 2024; Peleke et al., 2024; Wu et al., 2024). This has included methods for the detection of CREs and linking them to their cognate genes (T. Li et al., 2024; Zhu et al., 2024; Hale et al., 2025). These approaches will continue to evolve alongside the epigenomic assays we have discussed and will play an increasingly important role in crop *cis*-regulome characterisation in the years ahead.

Overall, improved characterisation of *cis*-regulatory landscapes will enable faster, higher confidence development of novel promoters and regulatory environments. This will in turn

promote hypothesis testing directly in crop species via transgenic lines, much as we did in Chapter 3 as we sought to manipulate wheat basal spikelet development.

5.5 – Ideotype design and engineering is a key contribution of academic research to breeding

Breeding companies and the CGIAR network have been, and continue to be, incredibly successful at recombining variation to achieve consistent yield gains and robust disease resistance, mostly through the stacking of minor, unmapped alleles (Nelson et al., 2018; Cowger & Brown, 2019; Brown, 2021). This is achieved by the steady adoption of new tools, for example, advances in genomic selection (R2D2 Consortium et al., 2021; Merrick et al., 2022; A. Alemu et al., 2024). However, the typical breeding process of iterative refinement by combining promising parental lines does not allow for the testing of certain traits or trait combinations, for example because intermediate steps are highly deleterious to breeding metrics. In evolutionary terminology, breeding may converge on local maxima but miss greater opportunities in global trait space. This was already recognised in the 1960s, with C. M. Donald writing "if we can sensibly postulate a model, albeit but a crude attempt at perfection, then we have the opportunity to devise and examine a combination of characters which otherwise may not occur in breeders' plots for centuries" (Donald, 1968). In this paper, Donald defined such a model as an 'ideotype'; a set of characteristics which if realised in a crop – and aligned with appropriate changes to management practices – could maximise productivity. Writing in the aftermath of the Green Revolution, Donald's ideotype concept was inspired by the development of semi-dwarf varieties of wheat and rice which could capitalise on high rates of nitrogenous fertiliser application without experiencing severe lodging.

While breeding companies do sometimes embark on projects to specifically engineer complex traits – see KWS' development of the highly efficient 'blue aleurone' method for hybrid seed production (international patent WO 2019/043082 A1) – the range of possible traits to explore is vast. We therefore argue that a key role played by applied academic research on crops is the design, engineering, and testing of ideotype components. Current large-scale projects in academic crop ideotype engineering include the C_4 Rice Project, ENSA (Enabling Nutrient Symbioses in Agriculture), and RIPE (Realising Increased Photosynthetic Efficiency). In turn, fundamental research is vital for informing our concepts of what might be valuable ideotype components to test. For instance, modelling of canopy photosynthesis has improved our understanding of optimal chlorophyll concentrations across soybean plants, inviting subsequent testing of this updated ideotype component (Walker et al., 2018). Predicting beneficial ideotype components for future climates and growing conditions is another valuable contribution of academic research (Semenov &

Stratonovitch, 2013; Tao et al., 2017; Senapati et al., 2019). For example, we may be due a re-evaluation of the benefits (improved water use efficiency and spike photosynthesis) and drawbacks (potentially increased susceptibility to some fungal diseases) of awns for UK wheat given future climactic predictions (Maydup et al., 2010; Rebetzke et al., 2016; Sanchez-Bragado et al., 2016; Carbajal-Friedrich & Burgess, 2024). Modelling of future ideotypes is another area in which machine learning may soon play an important role (Streich et al., 2020; Zhang et al., 2021; L. Zhang et al., 2022).

In this thesis, we have explored two potential ideotype components; larger grains for white tef varieties (Chapter 2) and increased basal spikelet productivity for wheat (Chapter 3). The first may not be achieved in current tef breeding programmes because, as we have demonstrated (Chapters 2.3.2 and 2.3.4), there is a general trend for increased pigmentation with increased grain size, and many loci for grain size are also associated with grain colour. Thus, breeding programmes may unintentionally avoid incorporating many alleles which increase grain size. There is also no guarantee that increased grain size will align with current tef breeding aims. We anticipate that increasing grain size will not result in increased in-field (gross) yield due to compensatory reductions in grain number per unit area, as observed in other crops (Sadras, 2007; Gambín & Borrás, 2010; Griffiths et al., 2015; T. Guo et al., 2018). Instead, the benefits of increased grain size lie in characteristics that may not currently be evaluated. For example, increased grain size could improve the separation of grain and chaff in traditional winnowing processes, reducing post-harvest losses and thereby increasing 'in-bag' (net) yields. Additionally, as we have argued previously, increased grain size could help reduce manual over-sowing or make tef more amenable to mechanised sowing using existing machinery. As a result of these factors, it may initially be the province of academic research to develop compelling evidence and appropriate germplasm for the adoption of this ideotype component.

We can also make a case for basal spikelet productivity being a trait more suitable for academic research. As previously noted, there is relatively little natural variation in modern and even landrace germplasm for alterations to the classic lanceolate shape of the wheat spike (Philipp et al., 2018). Other studies also suggest low heritabilities for the grain set (Guo et al., 2015) and maximum floret primordia (Z. Guo et al., 2018) of basal spikelets versus other positions along the spike. Therefore, to test whether this ideotype concept was worth pursuing, we had to generate novel variation through specific transgenic manipulations. Our pilot experiment suggests that targeted overexpression of *SEP1-B6* in the base of the spike reduces the number of RBS (Chapter 3.3.6), though we will need to validate this with additional trials in glasshouse and field environments, ideally with sibling zero copy number

lines as controls. Only if these tests show that we can successfully and consistently improve basal spikelet fertility in field conditions will we then be able to answer our original question – does this trait actually contribute to overall yield or other breeding goals? If so, a more distant future goal could be to recapitulate the phenotype through genome editing to allow its use in additional markets. The ongoing advances in wheat transformation and CRE characterisation discussed in Chapters 5.3 and 5.4, respectively, will prove essential to such an undertaking. Similar advances will enable higher throughput testing of ideotype components in other cereals too, facilitating this important function of academic crop research.

5.6 – Concluding statement

“The first essential component of social justice is adequate food for all mankind”

- Norman Borlaug, 1970

In this thesis I have explored how different types of variation can be used to study cereal inflorescence development and improve inflorescence traits. I have identified natural variation for important grain traits in tef, engineered variation for basal spikelet productivity in wheat, and tested a method for profiling the *cis*-regulome of wheat spike development. The study of cereal inflorescence development has advanced greatly in the genomics era, and the rate of progress has only risen during my PhD. Increased application of epigenomic techniques, single-cell and spatial transcriptomics, and machine learning approaches are all contributing to this acceleration. I believe that academic research will continue to play an important role in improving cereal yields and quality during the remainder of this challenging century. This will include fundamental developmental studies, work on minor cereals, and the development of novel ideotype components for step-changes in crop performance.

6 – Bibliography

Acharya, D., & Mukhopadhyay, A. (2024). A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. *Brief Funct Genomics*, *23*(5), 549-560. https://doi.org/10.1093/bfgp/elae013

Adamski, N. M., Borrill, P., Brinton, J., Harrington, S. A., Marchal, C., Bentley, A. R., Bovill, W. D., Cattivelli, L., Cockram, J., Contreras-Moreira, B., Ford, B., Ghosh, S., Harwood, W., Hassani-Pak, K., Hayta, S., Hickey, L. T., Kanyuka, K., King, J., Maccaferrri, M.,...Uauy, C. (2020). A roadmap for gene functional characterisation in crops with large genomes: Lessons from polyploid wheat. *Elife*, *9*. https://doi.org/10.7554/eLife.55646

Adamski, N. M., Simmonds, J., Brinton, J. F., Backhaus, A. E., Chen, Y., Smedley, M., Hayta, S., Florio, T., Crane, P., Scott, P., Pieri, A., Hall, O., Barclay, J. E., Clayton, M., Doonan, J. H., Nibau, C., & Uauy, C. (2021). Ectopic expression of *Triticum polonicum VRT-A2* underlies elongated glumes and grains in hexaploid wheat in a dosage-dependent manner. *The Plant Cell*, *33*(7), 2296-2319. https://doi.org/10.1093/plcell/koab119

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. Q., & Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density transposition. *Genome Biology*, *11*(12). https://doi.org/10.1186/gb-2010-11-12-r119

Al-Kaff, N., Knight, E., Bertin, I., Foote, T., Hart, N., Griffiths, S., & Moore, G. (2008). Detailed dissection of the chromosomal region containing the Ph1 locus in wheat Triticum aestivum: with deletion mutants and expression profiling. *Annals of Botany*, *101*(6), 863-872. https://doi.org/10.1093/aob/mcm252

Alemu, A., Astrand, J., Montesinos-Lopez, O. A., Isidro, Y. S. J., Fernandez-Gonzalez, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J., Ortiz, R., & Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. *Mol Plant*, *17*(4), 552-578. https://doi.org/10.1016/j.molp.2024.03.007

Alemu, M. D., Ben-Zeev, S., Hellwig, T., Barak, V., Shoshani, G., Chen, A., Razzon, S., Herrmann, I., Vorobyova, A., Hübner, S., & Saranga, Y. (2024). Genomic dissection of productivity, lodging, and morpho-physiological traits in *Eragrostis tef* under contrasting water availabilities. *Plants People Planet*. https://doi.org/10.1002/ppp3.10505

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

Ali, S. S., Abdelkarim, E. A., Elsamahy, T., Al-Tohamy, R., Li, F., Kornaros, M., Zuorro, A., Zhu, D., & Sun, J. (2023). Bioplastic production in terms of life cycle assessment: A state-of-the-art review. *Environ Sci Ecotechnol*, *15*, 100254. https://doi.org/10.1016/j.ese.2023.100254

Amalraj, B., Govindaraju, P., Krishna, A., Lavania, D., Linh, N. M., Ravichandran, S. J., & Scarpella, E. (2020). GAL4/GFP enhancer-trap lines for identification and manipulation of cells and tissues in developing Arabidopsis leaves. *Developmental Dynamics*, *249*(9), 1127-1146. https://doi.org/10.1002/dvdy.181

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. In http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Andrieu, A., Breitler, J. C., Sire, C., Meynard, D., Gantet, P., & Guiderdoni, E. (2012). An in planta, Agrobacterium-mediated transient gene expression method for inducing gene silencing in rice (Oryza sativa L.) leaves. *Rice (N Y)*, *5*(1), 23. https://doi.org/10.1186/1939-8433-5-23

Anzalone, A. V., Koblan, L. W., & Liu, D. R. (2020). Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, *38*(7), 824-844. https://doi.org/10.1038/s41587-020-0561-9

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Pozniak, C., Barad, O., Baruch, K., Keeble-Gagnere, G., Mascher, M., Ben-Zvi, G., Josselin, A. A.,...Team, M. W. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), 661-+. https://doi.org/10.1126/science.aar7191

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T. C., Alhendi, A. M., Arner, P., Axton, R., Baillie, J. K., Beckhouse, A., Bodega, B., Briggs, J., Brombacher, F.,...Hayashizaki, Y. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, *347*(6225), 1010-1014. https://doi.org/10.1126/science.1259418

Arora, K., Panda, K. K., Mittal, S., Mallikarjuna, M. G., Rao, A. R., Dash, P. K., & Thirunavukkarasu, N. (2017). RNAseq revealed the important gene pathways controlling adaptive mechanisms under waterlogged stress in maize. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-10561-1

Asplund, L., Hagenblad, J., & Leino, M. W. (2010). Re-evaluating the history of the wheat domestication gene *NAM-B1* using historical plant material. *Journal of Archaeological Science*, *37*(9), 2303-2307. https://doi.org/10.1016/j.jas.2010.04.003

Assefa, K., Cannarozzi, G., Girma, D., Kamies, R., Chanyalew, S., Plaza-Wüthrich, S., Blösch, R., Rindisbacher, A., Rafudeen, S., & Tadele, Z. (2015). Genetic diversity in tef [*Eragrostis tef* (Zucc.) Trotter]. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00177

Assefa, K., Yu, J.-K., Zeid, M., Belay, G., Tefera, H., & Sorrells, M. E. (2011). Breeding tef [*Eragrostis tef* (Zucc.) trotter]: conventional and molecular approaches. *Plant Breeding*, *130*(1), 1-9. https://doi.org/https://doi.org/10.1111/j.1439-0523.2010.01782.x

Assefa, M., Mehiret, T., Purba, J. H., Bahta, M., & Haille, A. (2022). Economic Analysis of Tef Yield Response to Different Sowing Methods and Seed Rates in Eastern Amhara, Ethiopia. *Agro Bali*, *5*(3). https://doi.org/https://doi.org/10.37637/ab.v5i3.868

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., MacLachlan, R. P., Sharpe, A. G., Fritz, A.,...Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, *357*(6346), 93-97. https://doi.org/10.1126/science.aan0032

Babu, M., & Snyder, M. (2023). Multi-Omics Profiling for Health. *Molecular and Cellular Proteomics*, *22*(6), 100561. https://doi.org/10.1016/j.mcpro.2023.100561

Backhaus, A. E., Griffiths, C., Vergara-Cruces, A., Simmonds, J., Lee, R., Morris, R. J., & Uauy, C. (2023). Delayed development of basal spikelets in wheat explains their increased floret abortion and rudimentary nature. *bioRxiv*, 2023.2002.2017.528935. https://doi.org/10.1101/2023.02.17.528935

Backhaus, A. E., Lister, A., Tomkins, M., Adamski, N. M., Simmonds, J., Macaulay, I., Morris, R. J., Haerty, W., & Uauy, C. (2022). High expression of the MADS-box gene VRT2 increases the number of rudimentary basal spikelets in wheat. *Plant Physiology*, *189*(3), 1536-1552. https://doi.org/10.1093/plphys/kiac156

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, *27*(12), 1653-1659. https://doi.org/10.1093/bioinformatics/btr261

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, *34*(Web Server issue), W369-373. https://doi.org/10.1093/nar/gkl198

Bajic, M., Maher, K. A., & Deal, R. B. (2018). Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. *Methods Mol Biol*, *1675*, 183-201. https://doi.org/10.1007/978-1-4939-7318-7_12

Barretto, R., Buenavista, R. M., Rivera, J. L., Wang, S., Prasad, P. V. V., & Siliveru, K. (2021). Teff (*Eragrostis tef*) processing, utilization and future opportunities: a review. *International Journal of Food Science & Technology*, *56*(7), 3125-3137. https://doi.org/https://doi.org/10.1111/ijfs.14872

Bartlett, M. E., & Thompson, B. (2014). Meristem identity and phyllotaxis in inflorescence development. *Frontiers in Plant Science*, *5*, 508. https://doi.org/10.3389/fpls.2014.00508

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/DOI 10.18637/jss.v067.i01

Bayable, M., Tsunekawa, A., Haregeweyn, N., Ishii, T., Alemayehu, G., Tsubo, M., Adgo, E., Tassew, A., Tsuji, W., Asaregew, F., & Masunaga, T. (2020). Biomechanical Properties and Agro-Morphological Traits for Improved Lodging Resistance in Ethiopian Teff (*Eragrostis tef* (Zucc.) Trottor) Accessions. *Agronomy-Basel*, *10*(7). https://doi.org/10.3390/agronomy10071012

Beales, J., Turner, A., Griffiths, S., Snape, J. W., & Laurie, D. A. (2007). A *Pseudo-Response Regulator* is misexpressed in the photoperiod insensitive *Ppd-D1a* mutant of wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*, *115*(5), 721-733. https://doi.org/10.1007/s00122-007-0603-4

Bektas, H., Hohn, C. E., Lukaszewski, A. J., & Waines, J. G. (2023). On the Possible Trade-Off between Shoot and Root Biomass in Wheat. *Plants (Basel)*, *12*(13). https://doi.org/10.3390/plants12132513

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268-276. https://doi.org/10.1016/j.ymeth.2012.05.001

Ben-Zeev, S., Kirby, N., Diehn, S., Shtein, I., Elbaum, R., & Saranga, Y. (2023). Unraveling the central role of root morphology and anatomy in lodging of tef (*Eragrostis tef*). *Plants People Planet*, *7*(3), 654-665. https://doi.org/10.1002/ppp3.10389

Ben-Zeev, S., Rabinovitz, O., Orlov-Levin, V., Chen, A., Graff, N., Goldwasser, Y., & Saranga, Y. (2020). Less Is More: Lower Sowing Rate of Irrigated Tef (Eragrostis tef) Alters Plant Morphology and Reduces Lodging. *Agronomy*, *10*(4).

Bennett, M., Gallagher, M., Fagg, J., Bestwick, C., Paul, T., Beale, M., & Mansfield, J. (1996). The hypersensitive reaction, membrane damage and accumulation of autofluorescent phenolics in lettuce cells challenged by Bremia lactucae. *The Plant Journal*, *9*(6), 851-865. https://doi.org/https://doi.org/10.1046/j.1365-313X.1996.9060851.x

Berhe, T. (1981). *Inheritance of Lemma Color, Seed Color and Panicle Form Among Four Cultivars of Eragrostis Tef (Zucc.) Trotter* [Doctoral Thesis, University of Nebraska-Lincoln]. https://digitalcommons.unl.edu/dissertations/AAI8118060/

Bian, X., Tyrrell, S., Olvera, D., & Davey, R. P. (2017). *The Grassroots life science data infrastructure* https://grassroots.tools

Biłas, R., Szafran, K., Hnatuszko-Konka, K., & Kononowicz, A. K. (2016). Cis-regulatory elements used to control gene expression in plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, *127*(2), 269-287. https://doi.org/10.1007/s11240-016-1057-7

Biswal, A. K., Hernandez, L. R. B., Castillo, A. I. R., Debernardi, J. M., & Dhugga, K. S. (2023). An efficient transformation method for genome editing of elite bread wheat cultivars. *Frontiers in Plant Science*, *14*, 1135047. https://doi.org/10.3389/fpls.2023.1135047

Blösch, R., Plaza-Wüthrich, S., de Reuille, P. B., Weichert, A., Routier-Kierzkowska, A. L., Cannarozzi, G., Robinson, S., & Tadele, Z. (2020). Panicle Angle is an Important Factor in Tef Lodging Tolerance. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00061

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., & Davies, R. M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab007

Bonnett, O. (1936). The development of the wheat spike. *Journal of Agricultural Research*, *53*, 445–451.

Borlaug, N. (1970). Norman Borlaug - Noble Lecture. In *Les Prix Nobel en 1970*. The Nobel Foundation. https://www.nobelprize.org/prizes/peace/1970/borlaug/lecture

Borrill, P., Harrington, S. A., & Uauy, C. (2019). Applying the latest advances in genomics and phenomics for trait discovery in polyploid wheat. *Plant Journal*, *97*(1), 56-72. https://doi.org/10.1111/tpj.14150

Borrill, P., Ramirez-Gonzalez, R., & Uauy, C. (2016). expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiology*, *170*(4), 2172-2186. https://doi.org/10.1104/pp.15.01667

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633-2635. https://doi.org/10.1093/bioinformatics/btm308

Bräuning, S., Catanach, A., Lord, J. M., Bicknell, R., & Macknight, R. C. (2018). Comparative transcriptome analysis of the wild-type model apomict (*Hieracium praealtum* and its *loss of parthenogenesis* (*lop*) mutant. *BMC Plant Biology*, *18*. https://doi.org/10.1186/s12870-018-1423-1

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525-527. https://doi.org/10.1038/nbt.3519

Brinton, J., & Uauy, C. (2019). A reductionist approach to dissecting grain weight and yield in wheat. *J Integr Plant Biol*, *61*(3), 337-358. https://doi.org/10.1111/jipb.12741

Brown, J. K. M. (2021). Achievements in breeding cereals with durable disease resistance in Northwest Europe. In R. Oliver (Ed.), *Achieving durable disease resistance in cereals* (1st ed.). Burleigh Dodds Science Publishing. https://doi.org/https://doi.org/10.1201/9781003180715

Bubb, K. L., & Deal, R. B. (2020). Considerations in the analysis of plant chromatin accessibility data. *Current Opinion in Plant Biology*, *54*, 69-78. https://doi.org/10.1016/j.pbi.2020.01.003

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding

proteins and nucleosome position. *Nature Methods*, *10*(12), 1213-1218. https://doi.org/10.1038/nmeth.2688

Byrne, M. E., Groover, A. T., Fontana, J. R., & Martienssen, R. A. (2003). Phyllotactic pattern and stem cell fate are determined by the Arabidopsis homeobox gene BELLRINGER. *Development*, *130*(17), 3941-3950. https://doi.org/10.1242/dev.00620

Cackett, L., Luginbuehl, L. H., Hendron, R.-W., Plackett, A. R. G., Stanley, S., Kelly, S., & Hibberd, J. M. (2024). Increased chloroplast area in the rice bundle sheath through cell specific perturbation of brassinosteroid signalling. *bioRxiv*, 2024.2008.2014.607565. https://doi.org/10.1101/2024.08.14.607565

Caldas, P. A. R., Zhu, J., Breakspear, A., Thapa, S. P., Toruño, T. Y., Perilla-Henao, L. M., Casteel, C., Faulkner, C. R., & Coaker, G. (2022). Effectors from a Bacterial Vector-Borne Pathogen Exhibit Diverse Subcellular Localization, Expression Profiles, and Manipulation of Plant Defense. *Molecular Plant-Microbe Interactions*, *35*(12), 1067-1080. https://doi.org/10.1094/Mpmi-05-22-0114-R

Calderini, D. F., Castillo, F. M., Arenas, M. A., Molero, G., Reynolds, M. P., Craze, M., Bowden, S., Milner, M. J., Wallington, E. J., Dowle, A., Gomez, L. D., & McQueen-Mason, S. J. (2021). Overcoming the trade-off between grain weight and number in wheat by the ectopic expression of expansin in developing seeds leads to increased yield potential. *New Phytologist*, *230*(2), 629-640. https://doi.org/10.1111/nph.17048

Candela-Ferre, J., Diego-Martin, B., Perez-Alemany, J., & Gallego-Bartolome, J. (2024). Mind the gap: Epigenetic regulation of chromatin accessibility in plants. *Plant Physiology*, *194*(4), 1998-2016. https://doi.org/10.1093/plphys/kiae024

Cannarozzi, G., Chanyalew, S., Assefa, K., Bekele, A., Blösch, R., Weichert, A., Klauser, D., Plaza-Wüthrich, S., Esfeld, K., Jöst, M., Rindisbacher, A., Jifar, H., Johnson-Chadwick, V., Abate, E., Wang, W., Kamies, R., Husein, N., Kebede, W., Tolosa, K.,...Tadele, Z. (2018). Technology generation to dissemination: lessons learned from the tef improvement project. *Euphytica*, *214*(2), 31. https://doi.org/10.1007/s10681-018-2115-5

Cannarozzi, G., Plaza-Wuthrich, S., Esfeld, K., Larti, S., Wilson, Y. S., Girma, D., de Castro, E., Chanyalew, S., Blosch, R., Farinelli, L., Lyons, E., Schneider, M., Falquet, L., Kuhlemeier, C., Assefa, K., & Tadele, Z. (2014). Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics*, *15*(1), 581. https://doi.org/10.1186/1471-2164-15-581

Cao, S., Liu, B., Wang, D., Rasheed, A., Xie, L., Xia, X., & He, Z. (2024). Orchestrating seed storage protein and starch accumulation toward overcoming yield-quality trade-off in cereal crops. *J Integr Plant Biol*, *66*(3), 468-483. https://doi.org/10.1111/jipb.13633

Carbajal-Friedrich, A. A. J., & Burgess, A. J. (2024). The role of the ideotype in future agricultural production [Hypothesis and Theory]. *Frontiers in Plant Physiology*, *Volume 2 - 2024*. https://doi.org/10.3389/fphgy.2024.1341617

Carballo, J., Bellido, A. M., Selva, J. P., Zappacosta, D., Gallo, C. A., Albertini, E., Caccamo, M., & Echenique, V. (2023). From tetraploid to diploid, a pangenomic approach to identify genes lost during synthetic diploidization of. *Frontiers in Plant Science*, *14*. https://doi.org/ARTN 1133986

10.3389/fpls.2023.1133986

Cereal. In. (n.d.). *Merriam-Webster*. Retrieved 19/05/2025, from https://www.merriam-webster.com/dictionary/cereal

Chai, H., Tjong, H., Li, P., Liao, W., Wang, P., Wong, C. H., Ngan, C. Y., Leonard, W. J., Wei, C. L., & Ruan, Y. (2023). ChIATAC is an efficient strategy for multi-omics mapping of 3D epigenomes from low-cell inputs. *Nat Commun*, *14*(1), 213. https://doi.org/10.1038/s41467-023-35879-5

Chanyalew, S., Kebede, W., Fikre, T., Genet, Y., Jifar, H., Demissie, M., Tolossa, K., Tadesse, M., Tadele, Z., & Assefa, K. (2021). *Tef Breeding Manual*. Ethiopian Institute of Agricultural Research.

Chapman, M. A. (2025). Novel breeding resources for the underutilised legume, lablab, based on a pangenome approach. *Breeding Science*, *75*(1), 61-66. https://doi.org/10.1270/jsbbs.24055

Chapman, M. A., He, Y., & Zhou, M. (2022). Beyond a reference genome: pangenomes and population genomics of underutilized and orphan crops for future food and nutrition security. *New Phytologist*, *234*(5), 1583-1597. https://doi.org/10.1111/nph.18021

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In U. Bastolla, M. Porto, H. E. Roman, & M. Vendruscolo (Eds.), *Structural Approaches to Sequence Evolution:*

Molecules, Networks, Populations (pp. 207-232). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35306-5_10

Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., Chen, W., Liao, Y., Zhang, W., Gao, Y., Dong, K., Ren, R., Yang, T., Zhang, L., Qi, M., Li, Z., Zhao, M., Wang, H., Wang, J.,...Diao, X. (2023). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nature Genetics*, *55*(12), 2243-2254. https://doi.org/10.1038/s41588-023-01571-z

Chen, M., Wang, Z., Zhu, Y., Li, Z., Hussain, N., Xuan, L., Guo, W., Zhang, G., & Jiang, L. (2012). The effect of *TRANSPARENT TESTA2* on seed fatty acid biosynthesis and tolerance to environmental stresses during young seedling establishment in *Arabidopsis*. *Plant Physiology*, *160*(2), 1023-1036. https://doi.org/10.1104/pp.112.202945

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Y. (2023). *Exploit natural variation within the Triticum genus to increase the size of maternal floral organs and wheat grain* [Doctoral Thesis, University of East Anglia]. https://ueaeprints.uea.ac.uk/id/eprint/97644/

Chen, Z., Debernardi, J. M., Dubcovsky, J., & Gallavotti, A. (2022). The combination of morphogenic regulators BABY BOOM and GRF-GIF improves maize transformation efficiency. *bioRxiv*, 2022.2009.2002.506370. https://doi.org/10.1101/2022.09.02.506370

Cheng, A., Mayes, S., Dalle, G., Demissew, S., & Massawe, F. (2017). Diversifying crops for food and nutrition security - a case of teff. *Biological Reviews*, *92*(1), 188-198. https://doi.org/10.1111/brv.12225

Cheng, S., Feng, C., Wingen, L. U., Cheng, H., Riche, A. B., Jiang, M., Leverington-Waite, M., Huang, Z., Collier, S., Orford, S., Wang, X., Awal, R., Barker, G., O'Hara, T., Lister, C., Siluveru, A., Quiroz-Chavez, J., Ramirez-Gonzalez, R. H., Bryant, R.,...Griffiths, S. (2024). Harnessing landrace diversity empowers wheat breeding. *Nature*, *632*(8026), 823-831. https://doi.org/10.1038/s41586-024-07682-9

Chudalayandi, S. (2011). Enhancer trapping in plants. *Methods Mol Biol*, *701*, 285-300. https://doi.org/10.1007/978-1-61737-957-4_16

Chun, Y., Kumar, A., & Li, X. (2022). Genetic and molecular pathways controlling rice inflorescence architecture. *Frontiers in Plant Science*, *13*, 1010138. https://doi.org/10.3389/fpls.2022.1010138

Ciaffi, M., Paolacci, A. R., Tanzarella, O. A., & Porceddu, E. (2011). Molecular aspects of flower development in grasses. *Sex Plant Reprod*, *24*(4), 247-282. https://doi.org/10.1007/s00497-011-0175-y

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews: Genetics*, *6*(11), 836-846. https://doi.org/10.1038/nrg1711

Concia, L., Veluchamy, A., Ramirez-Prado, J. S., Martin-Ramirez, A., Huang, Y., Perez, M., Domenichini, S., Rodriguez Granados, N. Y., Kim, S., Blein, T., Duncan, S., Pichot, C., Manza-Mianza, D., Juery, C., Paux, E., Moore, G., Hirt, H., Bergounioux, C., Crespi, M.,...Benhamed, M. (2020). Wheat chromatin architecture is organized in genome territories and transcription factories. *Genome Biol*, *21*(1), 104. https://doi.org/10.1186/s13059-020-01998-1

Connor, H. E. (1979). Breeding systems in the grasses: a survey. *New Zealand Journal of Botany*, *17*(4), 547-574. https://doi.org/10.1080/0028825X.1979.10432571

Cortinovis, G., Vincenzi, L., Anderson, R., Marturano, G., Marsh, J. I., Bayer, P. E., Rocchetti, L., Frascarelli, G., Lanzavecchia, G., Pieri, A., Benazzo, A., Bellucci, E., Di Vittori, V., Nanni, L., Ferreira Fernandez, J. J., Rossato, M., Aguilar, O. M., Morrell, P. L., Rodriguez, M.,...Papa, R. (2024). Adaptive gene loss in the common bean pan-genome during range expansion and domestication. *Nat Commun*, *15*(1), 6698. https://doi.org/10.1038/s41467-024-51032-2

Cotter, C. J., Wright, A. J., Romanov, A. V., Graf, T. N., Whisnant, E. D., Flores-Bocanegra, L., Doldron, M. S., Oberlies, N. H., Jia, Z., & Ligaba-Osena, A. (2023). Evaluating the Antioxidant Properties of the Ancient-Crop Tef (Eragrostis tef) Grain Extracts in THP-1 Monocytes. *Antioxidants (Basel)*, *12*(8). https://doi.org/10.3390/antiox12081561

Cowger, C., & Brown, J. K. M. (2019). Durability of Quantitative Resistance in Crops: Greater Than We Know? *Annual Review of Phytopathology*, *57*, 253-277. https://doi.org/10.1146/annurev-phyto-082718-100016

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., & Springer, N. M. (2020). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant

genomes. *Proceedings of the National Academy of Sciences*, *117*(38), 23991-24000. https://doi.org/doi:10.1073/pnas.2010250117

Crisp, P. A., Noshay, J. M., Anderson, S. N., & Springer, N. M. (2019). Opportunities to Use DNA Methylation to Distil Functional Elements in Large Crop Genomes. *Mol Plant*, *12*(3), 282-284. https://doi.org/10.1016/j.molp.2019.02.006

Cui, R., Han, J., Zhao, S., Su, K., Wu, F., Du, X., Xu, Q., Chong, K., Theissen, G., & Meng, Z. (2010). Functional conservation and diversification of class E floral homeotic genes in rice (Oryza sativa). *Plant Journal*, *61*(5), 767-781. https://doi.org/10.1111/j.1365-313X.2009.04101.x

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural Biological and Environmental Statistics*, *11*(4), 381-393. https://doi.org/10.1198/108571106x154443

D'Odorico, P., Bhattachan, A., Davis, K. F., Ravi, S., & Runyan, C. W. (2013). Global desertification: Drivers and feedbacks. *Advances in Water Resources*, *51*, 326-344. https://doi.org/10.1016/j.advwatres.2012.01.013

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Dao, T. Q., Drapek, C., Jones, A., & Leiboff, S. (2023). Comparing hormone dynamics in cereal crops via transient expression of hormone sensors. *bioRxiv*, 2023.2011.2014.567063. https://doi.org/10.1101/2023.11.14.567063

Darvey, N., Zhang, P., Trethowan, R., Dong, C., Lage, J., Bird, N., Tapsell, C., & Hummel, A. (2019). *Improved blue aleurone and other segregation systems*.

Das, P., & Gundimeda, H. (2022). Is biofuel expansion in developing countries reasonable? A review of empirical evidence of food and land use impacts. *Journal of Cleaner Production*, *372*, 133501. https://doi.org/https://doi.org/10.1016/j.jclepro.2022.133501

Day, L. (2004). Lipid chemistry. In C. Wrigley (Ed.), *Encyclopedia of Grain Science* (pp. 157-165). Elsevier.

De Witte, D., Van de Velde, J., Decap, D., Van Bel, M., Audenaert, P., Demeester, P., Dhoedt, B., Vandepoele, K., & Fostier, J. (2015). BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements. *Bioinformatics*, *31*(23), 3758-3766. https://doi.org/10.1093/bioinformatics/btv466

Debernardi, J. M., Lin, H., Chuck, G., Faris, J. D., & Dubcovsky, J. (2017). microRNA172 plays a crucial role in wheat spike morphogenesis and grain threshability. *Development*, *144*(11), 1966-1975. https://doi.org/10.1242/dev.146399

Debernardi, J. M., Tricoli, D. M., Ercoli, M. F., Hayta, S., Ronald, P., Palatnik, J. F., & Dubcovsky, J. (2020). A GRF-GIF chimeric protein improves the regeneration efficiency of transgenic plants. *Nature Biotechnology*, *38*(11), 1274-1279. https://doi.org/10.1038/s41587-020-0703-0

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, *295*(5558), 1306-1311. https://doi.org/10.1126/science.1067799

Delisle, L., Doyle, M., & Heyl, F. (2019, 14/06/2024). *ATAC-Seq data analysis (Galaxy Training Materials)*. https://training.galaxyproject.org/training-material/topics/epigenetics/tutorials/atac-seq/tutorial.html

Diaz, A., Zikhali, M., Turner, A. S., Isaac, P., & Laurie, D. A. (2012). Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). *PLoS One*, *7*(3), e33234. https://doi.org/10.1371/journal.pone.0033234

Diaz, I., Royo, J., Delahoz, P. S., & Carbonero, P. (1995). Gene Specificity Is Maintained in Transient Expression Assays with Protoplasts Derived from Different Tissues of Barley. *Euphytica*, *85*(1-3), 203-207. https://doi.org/Doi 10.1007/Bf00023949

Divya, D., Sahu, N., Reddy, P. S., Nair, S., & Bentur, J. S. (2021). RNA-Sequencing Reveals Differentially Expressed Rice Genes Functionally Associated with Defense against BPH and WBPH in RILs Derived from a Cross between RP2068 and TN1. *Rice*, *14*(1). https://doi.org/10.1186/s12284-021-00470-3

Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell*, *127*(7), 1309-1321. https://doi.org/10.1016/j.cell.2006.12.006

Donald, C. M. (1968). The breeding of crop ideotypes. *Euphytica*, *17*(3), 385-403. https://doi.org/10.1007/BF00056241

Dong, N. Q., & Lin, H. X. (2021). Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. *J Integr Plant Biol*, *63*(1), 180-209. https://doi.org/10.1111/jipb.13054

Dong, P., Tu, X., Chu, P. Y., Lu, P., Zhu, N., Grierson, D., Du, B., Li, P., & Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol Plant*, *10*(12), 1497-1509. https://doi.org/10.1016/j.molp.2017.11.005

Doni Jayavelu, N., Jajodia, A., Mishra, A., & Hawkins, R. D. (2020). Candidate silencer elements for the human and mouse genomes. *Nat Commun*, *11*(1), 1061. https://doi.org/10.1038/s41467-020-14853-5

Dracatos, P. M., Lu, J., Sanchez-Martin, J., & Wulff, B. B. H. (2023). Resistance that stacks up: engineering rust and mildew disease control in the cereal crops wheat and barley. *Plant Biotechnology Journal*, *21*(10), 1938-1951. https://doi.org/10.1111/pbi.14106

Dreisigacker, S., Kishii, M., Lage, J., & Warburton, M. (2008). Use of synthetic hexaploid wheat to increase diversity for CIMMYT bread wheat improvement. *Australian Journal of Agricultural Research*, *59*(5), 413-420. https://doi.org/10.1071/AR07225

Du, Y.-L., Xi, Y., Cui, T., Anten, N. P. R., Weiner, J., Li, X., Turner, N. C., Zhao, Y.-M., & Li, F.-M. (2020). Yield components, reproductive allometry and the tradeoff between grain yield and yield stability in dryland spring wheat. *Field Crops Research*, *257*, 107930. https://doi.org/https://doi.org/10.1016/j.fcr.2020.107930

Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, *316*(5833), 1862-1866. https://doi.org/10.1126/science.1143986

Dvorak, J., & Akhunov, E. D. (2005). Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance. *Genetics*, *171*(1), 323-332. https://doi.org/10.1534/genetics.105.041632

Dvorak, J., Deal, K. R., Luo, M. C., You, F. M., von Borstel, K., & Dehghani, H. (2012). The origin of spelt and free-threshing hexaploid wheat. *Journal of Heredity*, *103*(3), 426-441. https://doi.org/10.1093/jhered/esr152

Dwivedi, S. L., Reynolds, M. P., & Ortiz, R. (2021). Mitigating tradeoffs in plant breeding. *iScience*, *24*(9), 102965. https://doi.org/10.1016/j.isci.2021.102965

Ebba, T. (1975). T'ef (Eragrostis Tef) Cultivars: Morphology and Classification Part 2. In *Experiment Station Bulletin* (Vol. 66). Addis Ababa University, College of Agriculture.

Edwards, A., Njaci, I., Sarkar, A., Jiang, Z., Kaithakottil, G. G., Moore, C., Cheema, J., Stevenson, C. E. M., Rejzek, M., Novak, P., Vigouroux, M., Vickers, M., Wouters, R. H. M., Paajanen, P., Steuernagel, B., Moore, J. D., Higgins, J., Swarbreck, D., Martens, S.,...Emmrich, P. M. F. (2023). Genomics and biochemical analyses reveal a metabolon key to beta-L-ODAP biosynthesis in Lathyrus sativus. *Nat Commun*, *14*(1), 876. https://doi.org/10.1038/s41467-023-36503-2

Eglitis-Sexton, J., Mangila, L., Andrews, H., Hickey, L., & Crisp, P. A. (2024). Utilisation of Methylome Data to Identify Stably Unmethylated Regions in Plant Genomes. *Bio-Protocol*, *14*(4). https://doi.org/10.21769/bioprotoc.4944

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, *3*(11), e3647. https://doi.org/10.1371/journal.pone.0003647

Engler, C., Youles, M., Gruetzner, R., Ehnert, T. M., Werner, S., Jones, J. D., Patron, N. J., & Marillonnet, S. (2014). A golden gate modular cloning toolbox for plants. *ACS Synth Biol*, *3*(11), 839-843. https://doi.org/10.1021/sb4001504

Estep, M. C., McKain, M. R., Vela Diaz, D., Zhong, J., Hodge, J. G., Hodkinson, T. R., Layton, D. J., Malcomber, S. T., Pasquet, R., & Kellogg, E. A. (2014). Allopolyploidy, diversification, and the Miocene grassland expansion. *Proc Natl Acad Sci U S A*, *111*(42), 15149-15154. https://doi.org/10.1073/pnas.1404177111

Ethiopian Statistics Service. (2022a). *Agricultural Sample Survey - Area and Production of Major Crops - 2021/22 (2014 E.C.)*.

Ethiopian Statistics Service. (2022b). *Agricultural Sample Survey - Land Utilisation - 2021/22 (2014 E.C.)*.

Evans, C. E. B., Arunkumar, R., & Borrill, P. (2022). Transcription factor retention through multiple polyploidization steps in wheat. *G3-Genes Genomes Genetics*, *12*(8). https://doi.org/10.1093/g3journal/jkac147

Evers, J. B., & Vos, J. (2013). Modeling branching in cereals. *Frontiers in Plant Science*, *4*, 399. https://doi.org/10.3389/fpls.2013.00399

Faci, I., Backhaus, A. E., & Uauy, C. (2024). Wheat spike meristem microdissection. *protocols.io*. https://doi.org/dx.doi.org/10.17504/protocols.io.3byl49r2zgo5/v2

Faci, I., Jones, M. R. W., & Uauy, C. (2024). Intact and clean nuclei isolation from wheat meristems. *protocols.io*. https://doi.org/dx.doi.org/10.17504/protocols.io.yxmvm3renl3p/v1

FAO. (2023). *Crops and livestock products* https://www.fao.org/faostat/en/#data/QCL

Faraco, M., Di Sansebastiano, G. P., Spelt, K., Koes, R. E., & Quattrocchio, F. M. (2011). One protoplast is not the other! *Plant Physiology*, *156*(2), 474-478. https://doi.org/10.1104/pp.111.173708

Faris, J. D., Zhang, Z., & Chao, S. (2014). Map-based analysis of the tenacious glume gene Tg-B1 of wild emmer and its role in wheat domestication. *Gene*, *542*(2), 198-208. https://doi.org/10.1016/j.gene.2014.03.034

Feldner-Busztin, D., Firbas Nisantzis, P., Edmunds, S. J., Boza, G., Racimo, F., Gopalakrishnan, S., Limborg, M. T., Lahti, L., & de Polavieja, G. G. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics*, *39*(2). https://doi.org/10.1093/bioinformatics/btad021

Feng, N., Song, G. Y., Guan, J. T., Chen, K., Jia, M. L., Huang, D. H., Wu, J. J., Zhang, L. C., Kong, X. Y., Geng, S. F., Liu, J., Li, A. L., & Mao, L. (2017). Transcriptome Profiling of Wheat Inflorescence Development from Spikelet Initiation to Floral Patterning Identified Stage-Specific Regulatory Genes. *Plant Physiology*, *174*(3), 1779-1794. https://doi.org/10.1104/pp.17.00310

Ferreira, L. C., Santana, F. M., Scagliusi, S. M. M., Beckmann, M., & Mur, L. A. J. (2023). Metabolomics links induced responses to the wheat pathogen – tan spot – (Pyrenophora tritici-repentis) to the biosynthesis of flavonoids and bioenergetic metabolism [Preprint]. *Research Square*. https://doi.org/https://doi.org/10.21203/rs.3.rs-3105957/v1

Fikadu, A. A., Heckelei, T., & Woldeyohanes, T. B. (2020). Technical Efficiency of Teff Farms Controlling for Neighborhood effects in Ethiopia [Preprint]. *Research Square*. https://doi.org/https://doi.org/10.21203/rs.3.rs-30863/v1

Finch, J. P., Wilson, T., Lyons, L., Phillips, H., Beckmann, M., & Draper, J. (2022). Spectral binning as an approach to post-acquisition processing of high resolution FIE-MS metabolome fingerprinting data. *Metabolomics*, *18*(8), 64. https://doi.org/10.1007/s11306-022-01923-6

Fischer, D. S., Theis, F. J., & Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Research*, *46*(20), e119-e119. https://doi.org/10.1093/nar/gky675

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, *183*(3), 557-564. https://doi.org/10.1111/j.1469-8137.2009.02923.x

Fletcher, L. (2025, 07/05). https://frontlinegenomics.com/the-100-genome-wheres-the-limit/.

Forster, B. P., Franckowiak, J. D., Lundqvist, U., Lyon, J., Pitkethly, I., & Thomas, W. T. (2007). The barley phytomer. *Annals of Botany*, *100*(4), 725-733. https://doi.org/10.1093/aob/mcm183

Foulkes, M. J., Slafer, G. A., Davies, W. J., Berry, P. M., Sylvester-Bradley, R., Martre, P., Calderini, D. F., Griffiths, S., & Reynolds, M. P. (2011). Raising yield potential of wheat. III. Optimizing partitioning to grain while maintaining lodging resistance. *Journal of Experimental Botany*, *62*(2), 469-486. https://doi.org/10.1093/jxb/erq300

Foyer, C. H., & Paul, M. J. (2001). Source–sink relationships. *Plant Psychol*, *78*, 519-524.

Fracasso, A., Trindade, L. M., & Amaducci, S. (2016). Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biology*, *16*. https://doi.org/10.1186/s12870-016-0800-x

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour*, *17*(1), 27-32. https://doi.org/10.1111/1755-0998.12509

Fu, D., Szucs, P., Yan, L., Helguera, M., Skinner, J. S., von Zitzewitz, J., Hayes, P. M., & Dubcovsky, J. (2005). Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. *Molecular Genetics and Genomics*, *273*(1), 54-65. https://doi.org/10.1007/s00438-004-1095-4

Gallegos, J. E., & Rose, A. B. (2015). The enduring mystery of intron-mediated enhancement. *Plant Science*, *237*, 8-15. https://doi.org/10.1016/j.plantsci.2015.04.017

Gambín, B. L., & Borrás, L. (2010). Resource distribution and the trade-off between seed number and seed weight: a comparison across crop species. *Annals of Applied Biology*, *156*(1), 91-102. https://doi.org/https://doi.org/10.1111/j.1744-7348.2009.00367.x

Gao, X., Liang, W., Yin, C., Ji, S., Wang, H., Su, X., Guo, C., Kong, H., Xue, H., & Zhang, D. (2010). The SEPALLATA-like gene OsMADS34 is required for rice inflorescence and spikelet development. *Plant Physiology*, *153*(2), 728-740. https://doi.org/10.1104/pp.110.156711

Gaspar, J. M. (2018). Improved peak-calling with MACS2. *bioRxiv*, 496521. https://doi.org/10.1101/496521

Gaurav, K., Arora, S., Silva, P., Sanchez-Martin, J., Horsnell, R., Gao, L., Brar, G. S., Widrig, V., John Raupp, W., Singh, N., Wu, S., Kale, S. M., Chinoy, C., Nicholson, P., Quiroz-Chavez, J., Simmonds, J., Hayta, S., Smedley, M. A., Harwood, W.,...Wulff, B. B. H. (2022). Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. *Nature Biotechnology*, *40*(3), 422-431. https://doi.org/10.1038/s41587-021-01058-4

Gebre, E., Gugsa, L., Schlüter, U., & Kunert, K. (2013). Transformation of tef (*Eragrostis tef*) by *Agrobacterium* through immature embryo regeneration system for inducing semi-dwarfism. *South African Journal of Botany*, *87*, 9-17. https://doi.org/https://doi.org/10.1016/j.sajb.2013.03.004

Gebru, Y. A., Sbhatu, D. B., & Kim, K. P. (2020). Nutritional Composition and Health Benefits of Teff (*Eragrostis tef* (Zucc.) Trotter). *Journal of Food Quality*, *2020*. https://doi.org/10.1155/2020/9595086

Genstat for Windows. In. (2022). VSN International.

Girma, D., Assefa, K., Chanyalew, S., Cannarozzi, G., Kuhlemeier, C., & Tadele, Z. (2014). The origins and progress of genomics research on Tef (Eragrostis tef). *Plant Biotechnology Journal*, *12*(5), 534-540. https://doi.org/10.1111/pbi.12199

Gonzales, L. Y. R., Cannarozzi, G., Jäggi, L., Assefa, K., Chanyalew, S., Dell'Acqua, M., & Tadele, Z. (2024). The role of omics in improving the orphan crop tef. *Trends in Genetics*, *40*(5), 449-461. https://doi.org/10.1016/j.tig.2024.03.003

Gregis, V., Andrés, F., Sessa, A., Guerra, R. F., Simonini, S., Mateos, J. L., Torti, S., Zambelli, F., Prazzoli, G. M., Bjerkan, K. N., Grini, P. E., Pavesi, G., Colombo, L., Coupland, G., & Kater, M. M. (2013). Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. *Genome Biol*, *14*(6). https://doi.org/10.1186/gb-2013-14-6-r56

Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., Colas, I., & Moore, G. (2006). Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature*, *439*(7077), 749-752. https://doi.org/10.1038/nature04434

Griffiths, S., Wingen, L., Pietragalla, J., Garcia, G., Hasan, A., Miralles, D., Calderini, D. F., Ankleshwaria, J. B., Waite, M. L., Simmonds, J., Snape, J., & Reynolds, M. (2015). Genetic dissection of grain size and grain number trade-offs in CIMMYT wheat germplasm. *PLoS One*, *10*(3), e0118847. https://doi.org/10.1371/journal.pone.0118847

Guo, J., Zhang, Y., Shi, W., Zhang, B., Zhang, J., Xu, Y., Cheng, X., Cheng, K., Zhang, X., Hao, C., & Cheng, S. (2015). Association Analysis of Grain-setting Rates in Apical and Basal Spikelets in Bread Wheat (Triticum aestivum L.). *Frontiers in Plant Science*, *6*, 1029. https://doi.org/10.3389/fpls.2015.01029

Guo, T., Chen, K., Dong, N. Q., Shi, C. L., Ye, W. W., Gao, J. P., Shan, J. X., & Lin, H. X. (2018). *GRAIN SIZE AND NUMBER1* Negatively Regulates the OsMKKK10-OsMKK4-OsMPK6 Cascade to Coordinate the Trade-off between Grain Number per Panicle and Grain Size in Rice. *Plant Cell*, *30*(4), 871-888. https://doi.org/10.1105/tpc.17.00959

Guo, Y., Mahony, S., & Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, *8*(8), e1002638. https://doi.org/10.1371/journal.pcbi.1002638

Guo, Z., Zhao, Y., Roder, M. S., Reif, J. C., Ganal, M. W., Chen, D., & Schnurbusch, T. (2018). Manipulation and prediction of spike morphology traits for the improvement of grain yield in wheat. *Sci Rep*, *8*(1), 14435. https://doi.org/10.1038/s41598-018-31977-3

Gupta, P. K. (2021). GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and

-mers. *Bioessays*, *43*(11). https://doi.org/ARTN e2100109

10.1002/bies.202100109

Hale, C. O., Hsu, S.-K., Zhai, J., Schulz, A. J., Aubuchon-Elder, T., Costa-Neto, G., Gelfond, A., El-Walid, M., Hufford, M., Kellogg, E. A., La, T., Marand, A. P., Seetharam, A. S., Scheben, A., Stitzer, M., Wrightsman, T., Romay, M. C., & Buckler, E. S. (2025). Extensive modulation of a conserved *cis*-regulatory code across 589 grass species. *bioRxiv*, 2025.2004.2023.650228. https://doi.org/10.1101/2025.04.23.650228

Hamada, H., Liu, Y., Nagira, Y., Miki, R., Taoka, N., & Imai, R. (2018). Biolistic-delivery-based transient CRISPR/Cas9 expression enables in planta genome editing in wheat. *Sci Rep*, *8*(1), 14422. https://doi.org/10.1038/s41598-018-32714-6

Hammond, G. P., & Li, B. (2016). Environmental and resource burdens associated with world biofuel production out to 2050: footprint components from carbon emissions and land use to waste arisings and water consumption. *Glob Change Biol Bioenergy*, *8*(5), 894-908. https://doi.org/10.1111/gcbb.12300

Han, J., Wang, P., Wang, Q., Lin, Q., Chen, Z., Yu, G., Miao, C., Dao, Y., Wu, R., Schnable, J. C., Tang, H., & Wang, K. (2020). Genome-Wide Characterization of DNase I-Hypersensitive Sites and Cold Response Regulatory Landscapes in Grasses. *Plant Cell*, *32*(8), 2457-2473. https://doi.org/10.1105/tpc.19.00716

Harrington, S. A. (2019). *Understanding the molecular and genetic mechanisms regulating senescence in wheat* University of East Anglia]. John Innes Centre.

Hayta, S. (2023). Leaf transformation in grasses. *Nat Plants*, *9*(2), 197-198. https://doi.org/10.1038/s41477-023-01349-5

Hayta, S., Smedley, M. A., Clarke, M., Forner, M., & Harwood, W. A. (2021). An Efficient *Agrobacterium*-Mediated Transformation Protocol for Hexaploid and Tetraploid Wheat. *Current Protocols*, *1*(3). https://doi.org/10.1002/cpz1.58

Hayta, S., Smedley, M. A., Demir, S. U., Blundell, R., Hinchliffe, A., Atkinson, N., & Harwood, W. A. (2019). An efficient and reproducible *Agrobacterium*-mediated transformation method for hexaploid wheat (*Triticum aestivum* L.). *Plant Methods*, *15*(1). https://doi.org/10.1186/s13007-019-0503-z

He, C., Washburn, J. D., Schleif, N., Hao, Y., Kaeppler, H., Kaeppler, S. M., Zhang, Z., Yang, J., & Liu, S. (2024). Trait association and prediction through integrative k-mer analysis. *Plant Journal*, *120*(2), 833-850. https://doi.org/10.1111/tpj.17012

He, Q., Tang, S., Zhi, H., Chen, J., Zhang, J., Liang, H., Alam, O., Li, H., Zhang, H., Xing, L., Li, X., Zhang, W., Wang, H., Shi, J., Du, H., Wu, H., Wang, L., Yang, P., Xing, L.,...Diao, X. (2023). A graph-based genome and pan-genome variation of the model plant Setaria. *Nature Genetics*, *55*(7), 1232-1242. https://doi.org/10.1038/s41588-023-01423-w

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development*, *32*(1), 26-41. https://doi.org/10.1101/gad.309351.117

Hensel, G., Himmelbach, A., Chen, W., Douchkov, D. K., & Kumlehn, J. (2011a). Transgene expression systems in the Triticeae cereals. *Journal of Plant Physiology*, *168*(1), 30-44. https://doi.org/10.1016/j.jplph.2010.07.007

Hensel, G., Himmelbach, A., Chen, W. X., Douchkov, D. K., & Kumlehn, J. (2011b). Transgene expression systems in the *Triticeae* cereals. *Journal of Plant Physiology*, *168*(1), 30-44. https://doi.org/10.1016/j.jplph.2010.07.007

Hetzel, J., Duttke, S. H., Benner, C., & Chory, J. (2016). Nascent RNA sequencing reveals distinct features in plant transcription. *Proc Natl Acad Sci U S A*, *113*(43), 12316-12321. https://doi.org/10.1073/pnas.1603217113

Hoshikawa, K., Fujita, S., Renhu, N., Ezura, K., Yamamoto, T., Nonaka, S., Ezura, H., & Miura, K. (2019). Efficient transient protein expression in tomato cultivars and wild species using agroinfiltration-mediated high expression system. *Plant Cell Reports*, *38*(1), 75-84. https://doi.org/10.1007/s00299-018-2350-1

Hu, H., Zhao, J., Thomas, W. J. W., Batley, J., & Edwards, D. (2025). The role of pangenomics in orphan crop improvement. *Nat Commun*, *16*(1), 118. https://doi.org/10.1038/s41467-024-55260-4

Hua, L., Stevenson, S. R., Reyna-Llorens, I., Xiong, H., Kopriva, S., & Hibberd, J. M. (2021). The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis. *Plant Journal*, *107*(1), 268-286. https://doi.org/10.1111/tpj.15292

Hua, L., Wang, N., Stanley, S., Donald, R. M., Kumar Eeda, S., Billakurthi, K., Borba, A. R., & Hibberd, J. M. (2024). A transcription factor quintet orchestrating bundle sheath expression in rice. *bioRxiv*, 2024.2006.2017.599020. https://doi.org/10.1101/2024.06.17.599020

Huang, F., Jiang, Y., Chen, T., Li, H., Fu, M., Wang, Y., Xu, Y., Li, Y., Zhou, Z., Jia, L., Ouyang, Y., & Yao, W. (2022). New Data and New Features of the FunRiceGenes (Functionally Characterized Rice Genes) Database: 2021 Update. *Rice (N Y)*, *15*(1), 23. https://doi.org/10.1186/s12284-022-00569-1

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., & Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci U S A*, *99*(12), 8133-8138. https://doi.org/10.1073/pnas.072223799

Hufnagel, B., Soriano, A., Taylor, J., Divol, F., Kroc, M., Sanders, H., Yeheyis, L., Nelson, M., & Peret, B. (2021). Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnology Journal*, *19*(12), 2532-2543. https://doi.org/10.1111/pbi.13678

Hwang, H. H., Yu, M., & Lai, E. M. (2017). Agrobacterium-mediated plant transformation: biology and applications. *Arabidopsis Book*, *15*, e0186. https://doi.org/10.1199/tab.0186

Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, *13*(7), 577-580. https://doi.org/10.1038/nmeth.3885

Ikeda, T., Tanaka, W., Toriba, T., Suzuki, C., Maeno, A., Tsuda, K., Shiroishi, T., Kurata, T., Sakamoto, T., Murai, M., Matsusaka, H., Kumamaru, T., & Hirano, H. Y. (2019). BELL1-like homeobox genes regulate inflorescence architecture and meristem maintenance in rice. *Plant Journal*, *98*(3), 465-478. https://doi.org/10.1111/tpj.14230

Ingram, A. L., & Doyle, J. J. (2003). The origin and evolution of Eragrostis tef (Poaceae) and related polyploids: evidence from nuclear waxy and plastid rps16. *American Journal of Botany*, *90*(1), 116-122. https://doi.org/10.3732/ajb.90.1.116

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, *106*(3), 159-164. https://doi.org/10.1016/j.ygeno.2015.06.005

International Rice Genome Sequencing, P. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793-800. https://doi.org/10.1038/nature03895

Jakrawatana, N., Ngammuangtueng, P., Vorayos, N., & Gheewala, S. H. (2023). Replacing single-use plastics with biomaterial packaging in Thailand and impacts on the water-energy-climate Nexus. *Sustainable Production and Consumption*, *39*, 506-520. https://doi.org/https://doi.org/10.1016/j.spc.2023.05.036

Jia, J., Xie, Y., Cheng, J., Kong, C., Wang, M., Gao, L., Zhao, F., Guo, J., Wang, K., Li, G., Cui, D., Hu, T., Zhao, G., Wang, D., Ru, Z., & Zhang, Y. (2021). Homology-mediated inter-chromosomal interactions in hexaploid wheat lead to specific subgenome territories following polyploidization and introgression. *Genome Biol*, *22*(1), 26. https://doi.org/10.1186/s13059-020-02225-7

Jiang, L., Ma, X., Zhao, S., Tang, Y., Liu, F., Gu, P., Fu, Y., Zhu, Z., Cai, H., Sun, C., & Tan, L. (2019). The APETALA2-Like Transcription Factor SUPERNUMERARY BRACT Controls Rice Seed Shattering and Seed Size. *Plant Cell*, *31*(1), 17-36. https://doi.org/10.1105/tpc.18.00304

Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., Zhao, L., Wang, Z., Zhang, Y., Li, T., Fu, J., Chitikineni, A., Hou, J., Liu, H., Dwivedi, G., Liu, X., Jia, J., Mao, L., Wang, X.,...Zhang, X. (2025). Pan-genome bridges wheat structural variations with habitat and breeding. *Nature*, *637*(8045), 384-393. https://doi.org/10.1038/s41586-024-08277-0

Jiao, W., Lu, K., Wen, M., Mao, J., Ni, Z., Chen, Z. J., Wang, X., Song, Q., & Yuan, J. (2024). Ploidy variation induces butterfly effect on chromatin topology in wheat. *Plant Journal*, *119*(5), 2450-2463. https://doi.org/10.1111/tpj.16932

Jifar, H., Assefa, K., & Tadele, Z. (2015). Grain yield variation and association of major traits in brown-seeded genotypes of tef [Eragrostis tef (Zucc.)Trotter]. *Agriculture & Food Security*, *4*(1), 7. https://doi.org/10.1186/s40066-015-0027-3

Jones, M. R. W., Kebede, W., Teshome, A., Girija, A., Teshome, A., Girma, D., Brown, J. K. M., Quiroz-Chavez, J., Jones, C. S., Wulff, B. B. H., Assefa, K., Tadele, Z., Mur, L. A. J., Chanyalew, S., Uauy, C., & Shorinola, O. (2025). Population genomics uncovers loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*). *Communications Biology*, *8*(1), 807. https://doi.org/10.1038/s42003-025-08206-5

Jones, M. R. W., Long, K., Phillips, A., & Uauy, C. (2022). *Collation and orthology-based identification of hormone-related genes in bread wheat* (Version 1.0). https://doi.org/https://doi.org/10.5281/zenodo.7082849

Jores, T., Tonnies, J., Dorrity, M. W., Cuperus, J. T., Fields, S., & Queitsch, C. (2020). Identification of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves. *Plant Cell*, *32*(7), 2120-2131. https://doi.org/10.1105/tpc.20.00155

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, *51*(D1), D587-D592. https://doi.org/10.1093/nar/gkac963

Kang, J., Zhang, Z., Lin, X., Liu, F., Song, Y., Zhao, P., Lin, Y., Luo, X., Li, X., Li, Y., Wang, W., Liu, C., Xu, S., Liu, X., & Xiao, J. (2025). TAC-C uncovers open chromatin interaction in crops and SPL-mediated photosynthesis regulation. *bioRxiv*, 2025.2002.2010.637364. https://doi.org/10.1101/2025.02.10.637364

Kang, M., Ko, E., & Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Brief Bioinform*, *23*(1). https://doi.org/10.1093/bib/bbab454

Karikari, B., Lemay, M. A., & Belzile, F. (2023). *k*-mer-Based Genome-Wide Association Studies in Plants: Advances, Challenges, and Perspectives. *Genes*, *14*(7). https://doi.org/10.3390/genes14071439

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*, *10*(1), 1930. https://doi.org/10.1038/s41467-019-09982-5

Ke, Y., Yuan, M., Liu, H., Hui, S., Qin, X., Chen, J., Zhang, Q., Li, X., Xiao, J., Zhang, Q., & Wang, S. (2020). The versatile functions of OsALDH2B1 provide a genic basis for growth–defense trade-offs in rice. *Proceedings of the National Academy of Sciences*, *117*(7), 3867-3873. https://doi.org/10.1073/pnas.1918994117

Kellogg, E. A. (2015). Flowering Plants - Monocots - Poaceae. In K. Kubitzki (Ed.), *The Families and Genera of Vascular Plants* (Vol. XIII). Springer.

Kellogg, E. A. (2022). Genetic control of branching patterns in grass inflorescences. *Plant Cell*, *34*(7), 2518-2533. https://doi.org/10.1093/plcell/koac080

Kellogg, E. A., Camara, P. E., Rudall, P. J., Ladd, P., Malcomber, S. T., Whipple, C. J., & Doust, A. N. (2013). Early inflorescence development in the grasses (Poaceae). *Frontiers in Plant Science*, *4*, 250. https://doi.org/10.3389/fpls.2013.00250

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204-2207. https://doi.org/10.1093/bioinformatics/btq351

Ketema, S. (1993). *Tef (Eragrostis Tef): Breeding, Agronomy, Genetic Resources, Utilization, and Role in Ethiopian Agriculture*. Institute of Agricultural Research.

Khoury, C. K., Bjorkman, A. D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., Rieseberg, L. H., & Struik, P. C. (2014). Increasing homogeneity in global food supplies and the implications for food security. *Proc Natl Acad Sci U S A*, *111*(11), 4001-4006. https://doi.org/10.1073/pnas.1313490111

Kino, R. I., Pellny, T. K., Mitchell, R. A. C., Gonzalez-Uriarte, A., & Tosi, P. (2020). High post-anthesis temperature effects on bread wheat (*Triticum aestivum* L.) grain transcriptome during early grain-filling. *BMC Plant Biology*, *20*(1). https://doi.org/10.1186/s12870-020-02375-7

Kirby, E. J. M., & Appleyard, M. (1984). *Cereal Development Guide* (2nd ed.). National Agricultural Centre Arable Unit.

Kirienko, D. R., Luo, A., & Sylvester, A. W. (2012). Reliable transient transformation of intact maize leaf cells for functional genomics and experimental study. *Plant Physiology*, *159*(4), 1309-1318. https://doi.org/10.1104/pp.112.199737

Kobayashi, K., Maekawa, M., Miyao, A., Hirochika, H., & Kyozuka, J. (2010). PANICLE PHYTOMER2 (PAP2), encoding a SEPALLATA subfamily MADS-box protein, positively controls spikelet meristem identity in rice. *Plant & Cell Physiology*, *51*(1), 47-57. https://doi.org/10.1093/pcp/pcp166

Kohl, S., Hollmann, J., Erban, A., Kopka, J., Riewe, D., Weschke, W., & Weber, H. (2015). Metabolic and transcriptional transitions in barley glumes reveal a role as transitory resource buffers during endosperm filling. *Journal of Experimental Botany*, *66*(5), 1397-1411. https://doi.org/10.1093/jxb/eru492

Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., & Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. *Science*, *312*(5778), 1392-1396. https://doi.org/10.1126/science.1126410

Koppolu, R., & Schnurbusch, T. (2019). Developmental pathways for shaping spike inflorescence architecture in barley and wheat. *J Integr Plant Biol*, *61*(3), 278-295. https://doi.org/10.1111/jipb.12771

Koressaar, T., & Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics*, *23*(10), 1289-1291. https://doi.org/10.1093/bioinformatics/btm091

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., Simmonds, J., Ramirez-Gonzalez, R. H., Wang, X. D., Borrill, P., Fosker, C., Ayling, S., Phillips, A. L., Uauy, C., & Dubcovsky, J. (2017). Uncovering hidden variation in polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(6), E913-E921. https://doi.org/10.1073/pnas.1619268114

Krattinger, S. G., Sucher, J., Selter, L. L., Chauhan, H., Zhou, B., Tang, M., Upadhyaya, N. M., Mieulet, D., Guiderdoni, E., Weidenbach, D., Schaffrath, U., Lagudah, E. S., & Keller, B. (2016). The wheat durable, multipathogen resistance gene Lr34 confers partial blast resistance in rice. *Plant Biotechnology Journal*, *14*(5), 1261-1268. https://doi.org/10.1111/pbi.12491

Kumar, N., Jody, H., & and Rawat, R. (2015). If They Grow It, Will They Eat and Grow? Evidence from Zambia on Agricultural Diversity and Child Undernutrition. *The Journal of Development Studies*, *51*(8), 1060-1077. https://doi.org/10.1080/00220388.2015.1018901

Kuzay, S., Lin, H., Li, C., Chen, S., Woods, D. P., Zhang, J., Lan, T., von Korff, M., & Dubcovsky, J. (2022). WAPO-A1 is the causal gene of the 7AL QTL for spikelet number per spike in wheat. *PLoS Genetics*, *18*(1), e1009747. https://doi.org/10.1371/journal.pgen.1009747

Kuzay, S., Xu, Y., Zhang, J., Katz, A., Pearce, S., Su, Z., Fraser, M., Anderson, J. A., Brown-Guedira, G., DeWitt, N., Peters Haugrud, A., Faris, J. D., Akhunov, E., Bai, G., & Dubcovsky, J. (2019). Identification of a candidate gene for a QTL for spikelet number per spike on wheat chromosome arm 7AL by high-resolution genetic mapping. *Theoretical and Applied Genetics*, *132*(9), 2689-2705. https://doi.org/10.1007/s00122-019-03382-5

Lai, B., Tang, Q., Jin, W., Hu, G., Wangsa, D., Cui, K., Stanton, B. Z., Ren, G., Ding, Y., Zhao, M., Liu, S., Song, J., Ried, T., & Zhao, K. (2018). Trac-looping measures genome structure and chromatin accessibility. *Nature Methods*, *15*(9), 741-747. https://doi.org/10.1038/s41592-018-0107-y

Lam, P. Y., Zhu, F. Y., Chan, W. L., Liu, H., & Lo, C. (2014). Cytochrome P450 93G1 Is a Flavone Synthase II That Channels Flavanones to the Biosynthesis of Tricin O-Linked Conjugates in Rice. *Plant Physiology*, *165*(3), 1315-1327. https://doi.org/10.1104/pp.114.239723

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

Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D., & Lieb, J. D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, *36*(8), 900-905. https://doi.org/10.1038/ng1400

Lee, D. Y., Hua, L., Khoshravesh, R., Giuliani, R., Kumar, I., Cousins, A., Sage, T. L., Hibberd, J. M., & Brutnell, T. P. (2021). Engineering chloroplast development in rice through cell-specific control of endogenous genetic circuits. *Plant Biotechnology Journal*, *19*(11), 2291-2303. https://doi.org/10.1111/pbi.13660

Lee, H. (2018). Teff, A Rising Global Crop: Current Status of Teff Production and Value Chain. *The Open Agriculture Journal*, *12*. https://doi.org/doi:10.2174/1874331501812010185

Levy, A. A., & Feldman, M. (2022). Evolution and origin of bread wheat. *Plant Cell*, *34*(7), 2549-2567. https://doi.org/10.1093/plcell/koac130

Li, A., Liu, D., Yang, W., Kishii, M., & Mao, L. (2018). Synthetic Hexaploid Wheat: Yesterday, Today, and Tomorrow. *Engineering*, *4*(4), 552-558. https://doi.org/https://doi.org/10.1016/j.eng.2018.07.001

Li, G., Jain, R., Chern, M., Pham, N. T., Martin, J. A., Wei, T., Schackwitz, W. S., Lipzen, A. M., Duong, P. Q., Jones, K. C., Jiang, L., Ruan, D., Bauer, D., Peng, Y., Barry, K. W., Schmutz, J., & Ronald, P. C. (2017). The Sequences of 1504 Mutants in the Model Rice Variety Kitaake Facilitate Rapid Functional Genomic Studies. *Plant Cell*, *29*(6), 1218-1231. https://doi.org/10.1105/tpc.17.00154

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J.-H., Li, M.-J., Li, W.-L., Li, X.-Y., Ma, Y.-B., Tan, X., Wang, Y., Li, C.-X., & Ma, X.-R. (2024). Leguminous industrial crop guar (Cyamopsis tetragonoloba): The chromosome-level reference genome de novo assembly. *Industrial Crops and Products*, *216*, 118748. https://doi.org/https://doi.org/10.1016/j.indcrop.2024.118748

Li, J., Pan, W., Zhang, S., Ma, G., Li, A., Zhang, H., & Liu, L. (2024). A rapid and highly efficient sorghum transformation strategy using GRF4-GIF1/ternary vector system. *Plant Journal*, *117*(5), 1604-1613. https://doi.org/10.1111/tpj.16575

Li, J., Xin, X., Sun, F., Zhu, Z., Xu, X., Yang, J., Xie, X., Yu, J., Wang, X., Li, S., Tian, S., Li, B., Xie, C., & Ma, J. (2023). Copy number variation of B1 controls awn length in wheat. *The Crop Journal*, *11*(3), 817-824. https://doi.org/https://doi.org/10.1016/j.cj.2022.10.007

Li, K., Debernardi, J. M., Li, C., Lin, H., Zhang, C., Jernstedt, J., Korff, M. v., Zhong, J., & Dubcovsky, J. (2021). Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development. *The Plant Cell*, *33*(12), 3621-3644. https://doi.org/10.1093/plcell/koab243

Li, L. F., Liu, B., Olsen, K. M., & Wendel, J. F. (2015). A re-evaluation of the homoploid hybrid origin of *Aegilops tauschii*, the donor of the wheat D-subgenome. *New Phytologist*, *208*(1), 4-8. https://doi.org/10.1111/nph.13294

Li, R., Li, L., Xu, Y., & Yang, J. (2022). Machine learning meets omics: applications and perspectives. *Brief Bioinform*, *23*(1). https://doi.org/10.1093/bib/bbab460

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P., & Pulendran, B. (2013). Predicting network activity from high throughput metabolomics. *PLoS Comput Biol*, *9*(7), e1003123. https://doi.org/10.1371/journal.pcbi.1003123

Li, T., Jia, L., Cao, Y., Chen, Q., & Li, C. (2018). OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. *Genome Biol*, *19*(1), 54. https://doi.org/10.1186/s13059-018-1430-4

Li, T., Xu, H., Teng, S., Suo, M., Bahitwa, R., Xu, M., Qian, Y., Ramstein, G. P., Song, B., Buckler, E. S., & Wang, H. (2024). Modeling 0.6 million genes for the rational design of functional cis-regulatory variants and de novo design of cis-regulatory sequences. *Proc Natl Acad Sci U S A*, *121*(26), e2319811121. https://doi.org/10.1073/pnas.2319811121

Li, X. Y., Thomas, S., Sabo, P. J., Eisen, M. B., Stamatoyannopoulos, J. A., & Biggin, M. D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol*, *12*(4), R34. https://doi.org/10.1186/gb-2011-12-4-r34

Li, Y. F., Zeng, X. Q., Li, Y., Wang, L., Zhuang, H., Wang, Y., Tang, J., Wang, H. L., Xiong, M., Yang, F. Y., Yuan, X. Z., & He, G. H. (2020). MULTI-FLORET SPIKELET 2, a MYB Transcription Factor, Determines Spikelet Meristem Fate and Floral Organ Identity in Rice. *Plant Physiology*, *184*(2), 988-1003. https://doi.org/10.1104/pp.20.00743

Li, Y. P., Fu, X., Zhao, M. C., Zhang, W., Li, B., An, D. G., Li, J. M., Zhang, A. M., Liu, R. Y., & Liu, X. G. (2018). A Genome-wide View of Transcriptome Dynamics During Early Spike Development in Bread Wheat. *Scientific Reports*, *8*. https://doi.org/10.1038/s41598-018-33718-y

Li, Z. Y., Wang, J. B., Zhang, X. Q., & Xu, L. (2015). Comparative Transcriptome Analysis of Anthurium "Albama" and Its Anthocyanin-Loss Mutant. *PLoS One*, *10*(3). https://doi.org/10.1371/journal.pone.0119027

Liang, Q., Munoz-Amatriain, M., Shu, S., Lo, S., Wu, X., Carlson, J. W., Davidson, P., Goodstein, D. M., Phillips, J., Janis, N. M., Lee, E. J., Liang, C., Morrell, P. L., Farmer, A. D., Xu, P., Close, T. J., & Lonardi, S. (2024). A view of the pan-genome of domesticated Cowpea (Vigna unguiculata [L.] Walp.). *Plant Genome*, *17*(1), e20319. https://doi.org/10.1002/tpg2.20319

Lin, X., Xu, Y., Wang, D., Yang, Y., Zhang, X., Bie, X., Gui, L., Chen, Z., Ding, Y., Mao, L., Zhang, X., Lu, F., Zhang, X., Uauy, C., Fu, X., & Xiao, J. (2024). Systematic identification of wheat spike developmental regulators by integrated multi-omics, transcriptional network, GWAS, and genetic analyses. *Molecular Plant*, *17*(3), 438-459. https://doi.org/10.1016/j.molp.2024.01.010

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J., Cao, Z., Gao, Y., Wang, X., Li, S., Su, Q., Zhang, Z., Wang, S., Wu, X., Shang, Q., Shi, H., Shen, Y., Wang, B., & Tian, J. (2022). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Commun*, *3*(6), 100352. https://doi.org/10.1016/j.xplc.2022.100352

Liu, D. (2009). Design of gene constructs for transgenic maize. *Methods Mol Biol*, *526*, 3-20. https://doi.org/10.1007/978-1-59745-494-0_1

Liu, G., Yang, Y., Guo, X., Liu, W., Xie, R., Ming, B., Xue, J., Wang, K., Li, S., & Hou, P. (2022). Coordinating maize source and sink relationship to achieve yield potential of 22.5 Mg ha-1. *Field Crops Research*, *283*, 108544. https://doi.org/https://doi.org/10.1016/j.fcr.2022.108544

Liu, J., Chen, Z. Y., Wang, Z. H., Zhang, Z. H., Xie, X. M., Wang, Z. H., Chai, L. L., Song, L., Cheng, X. J., Feng, M., Wang, X. B., Liu, Y. H., Hu, Z. R., Xing, J. W., Su, Z. Q., Peng, H. R., Xin, M. M., Yao, Y. Y., Guo, W. L.,...Ni, Z. F. (2021). Ectopic expression of *VRT-A2* underlies the origin of *Triticum polonicum* and *Triticum petropavlovskyi* with long outer glumes and grains. *Molecular Plant*, *14*(9), 1472-1488. https://doi.org/10.1016/j.molp.2021.05.021

Liu, J., Dong, C., Liu, X., Guo, J., Chai, L., Guo, W., Ni, Z., Sun, Q., & Liu, J. (2025). Decoupling the pleiotropic effects of VRT-A2 during reproductive development enhances wheat grain length and weight. *Plant Cell*, *37*(2). https://doi.org/10.1093/plcell/koaf024

Liu, M., Shi, Z., Zhang, X., Wang, M., Zhang, L., Zheng, K., Liu, J., Hu, X., Di, C., Qian, Q., He, Z., & Yang, D. L. (2019). Inducible overexpression of *Ideal Plant Architecture1* improves both yield and disease resistance in rice. *Nat Plants*, *5*(4), 389-400. https://doi.org/10.1038/s41477-019-0383-2

Liu, M., Zhu, J., Huang, H., Chen, Y., & Dong, Z. (2023). Comparative analysis of nascent RNA sequencing methods and their applications in studies of cotranscriptional splicing dynamics. *Plant Cell*, *35*(12), 4304-4324. https://doi.org/10.1093/plcell/koad237

Liu, N., Low, W. Y., Alinejad-Rokny, H., Pederson, S., Sadlon, T., Barry, S., & Breen, J. (2021). Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics Chromatin*, *14*(1), 41. https://doi.org/10.1186/s13072-021-00417-4

Liu, X., Lin, X., Kang, J., Long, K. A., Yue, J., Chen, C., Wang, D., Lister, A., Macaulay, I. C., Liu, X., Uauy, C., & Xiao, J. (2025). Decoding cellular transcriptional regulatory networks governing wheat inflorescence development. *bioRxiv*, 2025.2001.2018.633750. https://doi.org/10.1101/2025.01.18.633750

Liu, Y., Cai, Y. J., Li, Y. Z., Zhang, X. L., Shi, N., Zhao, J. Z., & Yang, H. C. (2022). Dynamic changes in the transcriptome landscape of *Arabidopsis thaliana* in response to cold stress. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.983460

Long, K. A., Lister, A., Jones, M. R. W., Adamski, N. M., Ellis, R. E., Chedid, C., Carpenter, S. J., Liu, X., Backhaus, A. E., Goldson, A., Knitlhoffer, V., Pei, Y., Vickers, M., Steuernagel, B., Kaithakottil, G. G., Xiao, J., Haerty, W., Macaulay, I. C., & Uauy, C. (2024). Spatial Transcriptomics Reveals Expression Gradients in Developing Wheat Inflorescences at Cellular Resolution. *bioRxiv*, 2024.2012.2019.629411. https://doi.org/10.1101/2024.12.19.629411

López-Álvarez, D., Zubair, H., Beckmann, M., Draper, J., & Catalán, P. (2017). Diversity and association of phenotypic and metabolomic traits in the close model grasses *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Annals of Botany*, *119*(4), 545-561. https://doi.org/10.1093/aob/mcw239

Lorenzo, C. D., Debray, K., Herwegh, D., Develtere, W., Impens, L., Schaumont, D., Vandeputte, W., Aesaert, S., Coussens, G., De Boe, Y., Demuynck, K., Van Hautegem, T., Pauwels, L., Jacobs, T. B., Ruttink, T., Nelissen, H., & Inze, D. (2023). BREEDIT: a multiplex genome editing strategy to improve complex quantitative traits in maize. *Plant Cell*, *35*(1), 218-238. https://doi.org/10.1093/plcell/koac243

Louwers, M., Bader, R., Haring, M., van Driel, R., de Laat, W., & Stam, M. (2009). Tissue- and expression level-specific chromatin looping at maize b1 epialleles. *Plant Cell*, *21*(3), 832-842. https://doi.org/10.1105/tpc.108.064329

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, F. H., McKenzie, N., Gardiner, L. J., Luo, M. C., Hall, A., & Bevan, M. W. (2020). Reduced chromatin accessibility underlies gene expression differences in homologous chromosome arms of diploid and hexaploid wheat. *Gigascience*, *9*(6). https://doi.org/10.1093/gigascience/giaa070

Lu, X., Liu, J., Ren, W., Yang, Q., Chai, Z., Chen, R., Wang, L., Zhao, J., Lang, Z., Wang, H., Fan, Y., Zhao, J., & Zhang, C. (2018). Gene-Indexed Mutations in Maize. *Mol Plant*, *11*(3), 496-504. https://doi.org/10.1016/j.molp.2017.11.013

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*(12), 1250-1259. https://doi.org/10.1038/s41477-019-0548-z

Lu, Z., Ricci, W. A., Schmitz, R. J., & Zhang, X. (2018). Identification of *cis*-regulatory elements by chromatin structure. *Current Opinion in Plant Biology*, *42*, 90-94. https://doi.org/https://doi.org/10.1016/j.pbi.2018.04.004

Lundström, M., Leino, M. W., & Hagenblad, J. (2017). Evolutionary history of the *NAM-B1* gene in wild and domesticated tetraploid wheat. *BMC Genetics*, *18*. https://doi.org/10.1186/s12863-017-0566-7

Luo, X., Chen, S., & Zhang, Y. (2022). PlantRep: a database of plant repetitive elements. *Plant Cell Reports*, *41*(4), 1163-1166. https://doi.org/10.1007/s00299-021-02817-y

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018). Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. *Plant Cell*, *30*(1), 15-36. https://doi.org/10.1105/tpc.17.00581

Mansisidor, A. R., & Risca, V. I. (2022). Chromatin accessibility: methods, mechanisms, and biological insights. *Nucleus*, *13*(1), 236-276. https://doi.org/10.1080/19491034.2022.2143106

Marand, A. P., Chen, Z., Gallavotti, A., & Schmitz, R. J. (2021). A cis-regulatory atlas in maize at single-cell resolution. *Cell*, *184*(11), 3041-3055 e3021. https://doi.org/10.1016/j.cell.2021.04.014

Marand, A. P., Zhang, T., Zhu, B., & Jiang, J. (2017). Towards genome-wide prediction and characterization of enhancers in plants. *Biochim Biophys Acta Gene Regul Mech*, *1860*(1), 131-139. https://doi.org/10.1016/j.bbagrm.2016.06.006

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing, C., Jakobsen, K. S., Wulff, B. B., Steuernagel, B., Mayer, K. F., & Olsen, O. A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*(6194), 1250092. https://doi.org/10.1126/science.1250092

Marinov, G. K., & Shipony, Z. (2021). Interrogating the Accessible Chromatin Landscape of Eukaryote Genomes Using ATAC-seq. *Methods Mol Biol*, *2243*, 183-226. https://doi.org/10.1007/978-1-0716-1103-6_10

Marks, M. D., Wenger, J. P., Gilding, E., Jilk, R., & Dixon, R. A. (2009). Transcriptome analysis of *Arabidopsis* wild-type and *gl3–sst sim* trichomes identifies four additional genes required for trichome development. *Molecular Plant*, *2*(4), 803-822. https://doi.org/10.1093/mp/ssp037

Marquès-Bueno, M. M., Morao, A. K., Cayrel, A., Platre, M. P., Barberon, M., Caillieux, E., Colot, V., Jaillais, Y., Roudier, F., & Vert, G. (2016). A versatile Multisite Gateway-compatible promoter and transgenic line collection for cell type-specific functional genomics in Arabidopsis. *Plant Journal*, *85*(2), 320-333. https://doi.org/10.1111/tpj.13099

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads [next generation sequencing; small RNA; microRNA; adapter removal]. *2011*, *17*(1), 3. https://doi.org/10.14806/ej.17.1.200

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C., & Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nature Genetics*, *51*(7), 1076-1081. https://doi.org/10.1038/s41588-019-0443-6

Mason, T. G., & Maskell, E. J. (1928). Studies on the Transport of Carbohydrates in the Cotton Plant: II. The Factors determining the Rate and the Direction of Movement of Sugars. *Annals of Botany*, *42*(167), 571-636. http://www.jstor.org/stable/43237122

Matic, S., Bagnaresi, P., Biselli, C., Orru, L., Carneiro, G. A., Siciliano, I., Valé, G., Gullino, M. L., & Spadaro, D. (2016). Comparative transcriptome profiling of resistant and susceptible rice genotypes in response to the seedborne pathogen *Fusarium fujikuroi*. *BMC Genomics*, *17*. https://doi.org/10.1186/s12864-016-2925-6

Matus, J. T., Ferrier, T., & Riechmann, J. L. (2014). Identification of Arabidopsis knockout lines for genes of interest. *Methods Mol Biol*, *1110*, 347-362. https://doi.org/10.1007/978-1-4614-9408-9_20

Maydup, M. L., Antonietta, M., Guiamet, J. J., Graciano, C., López, J. R., & Tambussi, E. A. (2010). The contribution of ear photosynthesis to grain filling in bread wheat (Triticum aestivum L.). *Field Crops Research*, *119*(1), 48-58. https://doi.org/https://doi.org/10.1016/j.fcr.2010.06.014

McDonald, B. R., Picard, C. L., Brabb, I. M., Savenkova, M. I., Schmitz, R. J., Jacobsen, S. E., & Duttke, S. H. (2024). Enhancers associated with unstable RNAs are rare in plants. *Nat Plants*, *10*(8), 1246-1257. https://doi.org/10.1038/s41477-024-01741-9

McMullin, S., Stadlmayr, B., Mausch, K., Revoredo-Giha, C., Burnett, F., Guarino, L., Brouwer, I. D., Jamnadass, R., Graudal, L., Powell, W., & Dawson, I. K. (2021). Determining appropriate

interventions to mainstream nutritious orphan crops into African food systems. *Global Food Security*, *28*, 100465. https://doi.org/https://doi.org/10.1016/j.gfs.2020.100465

Merrick, L. F., Herr, A. W., Sandhu, K. S., Lozada, D. N., & Carter, A. H. (2022). Utilizing Genomic Selection for Wheat Population Development and Improvement. *Agronomy*, *12*(2), 522. https://www.mdpi.com/2073-4395/12/2/522

Miller, S., Ronager, A., Holm, R., Fontanet-Manzaneque, J. B., Cano-Delgado, A. I., & Bjarnholt, N. (2023). New methods for sorghum transformation in temperate climates. *AoB Plants*, *15*(3), plad030. https://doi.org/10.1093/aobpla/plad030

Milner, S. G., Jost, M., Taketa, S., Mazon, E. R., Himmelbach, A., Oppermann, M., Weise, S., Knupffer, H., Basterrechea, M., Konig, P., Schuler, D., Sharma, R., Pasam, R. K., Rutten, T., Guo, G., Xu, D., Zhang, J., Herren, G., Muller, T.,...Stein, N. (2019). Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics*, *51*(2), 319-326. https://doi.org/10.1038/s41588-018-0266-x

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Minnoye, L., Marinov, G. K., Krausgruber, T., Pan, L., Marand, A. P., Secchia, S., Greenleaf, W. J., Furlong, E. E. M., Zhao, K., Schmitz, R. J., Bock, C., & Aerts, S. (2021). Chromatin accessibility profiling methods. *Nat Rev Methods Primers*, *1*. https://doi.org/10.1038/s43586-020-00008-9

Molla, K. A., & Yang, Y. (2019). CRISPR/Cas-Mediated Base Editing: Technical Considerations and Practical Applications. *Trends in Biotechnology*, *37*(10), 1121-1142. https://doi.org/10.1016/j.tibtech.2019.03.008

Montefiori, L., Hernandez, L., Zhang, Z. J., Gilad, Y., Ober, C., Crawford, G., Nobrega, M., & Sakabe, N. J. (2017). Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-02547-w

Muerdter, F., Boryn, L. M., & Arnold, C. D. (2015). STARR-seq - principles and applications. *Genomics*, *106*(3), 145-150. https://doi.org/10.1016/j.ygeno.2015.06.001

Murchie, E. H., Reynolds, M., Slafer, G. A., Foulkes, M. J., Acevedo-Siaca, L., McAusland, L., Sharwood, R., Griffiths, S., Flavell, R. B., Gwyn, J., Sawkins, M., & Carmo-Silva, E. (2023). A 'wiring diagram' for source strength traits impacting wheat yield potential. *Journal of Experimental Botany*, *74*(1), 72-90. https://doi.org/10.1093/jxb/erac415

Nalam, V. J., Vales, M. I., Watson, C. J., Kianian, S. F., & Riera-Lizarazu, O. (2006). Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat (Triticum turgidum L.). *Theoretical and Applied Genetics*, *112*(2), 373-381. https://doi.org/10.1007/s00122-005-0140-y

Nassar, L. R., Barber, G. P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Lee, C. M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B. J., Schmelter, D., Speir, M. L.,...Kent, W. J. (2023). The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research*, *51*(D1), D1188-D1195. https://doi.org/10.1093/nar/gkac1072

Nelson, R., Wiesner-Hanks, T., Wisser, R., & Balint-Kurti, P. (2018). Navigating complexity to breed disease-resistant crops. *Nature Reviews: Genetics*, *19*(1), 21-33. https://doi.org/10.1038/nrg.2017.82

Nesi, N., Jond, C., Debeaujon, I., Caboche, M., & Lepiniec, L. (2001). The *Arabidopsis* TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell*, *13*(9), 2099-2114. https://doi.org/10.1105/tpc.010098

Njaci, I., Waweru, B., Kamal, N., Muktar, M. S., Fisher, D., Gundlach, H., Muli, C., Muthui, L., Maranga, M., Kiambi, D., Maass, B. L., Emmrich, P. M. F., Domelevo Entfellner, J. B., Spannagl, M., Chapman, M. A., Shorinola, O., & Jones, C. S. (2023). Chromosome-level genome assembly and population genomic resource to accelerate orphan crop lablab breeding. *Nat Commun*, *14*(1), 1915. https://doi.org/10.1038/s41467-023-37489-7

Nobori, T. (2025). Exploring the untapped potential of single-cell and spatial omics in plant biology. *New Phytologist*. https://doi.org/10.1111/nph.70220

O'Connor, K., González-Suárez, P., & Dixon, L. E. (2020). Temperature Control of Plant Development. *Annual Plant Reviews online*, *3*(3), 563-606. https://doi.org/https://doi.org/10.1002/9781119312994.apr0745

Ogihara, Y., Isono, K., Kojima, T., Endo, A., Hanaoka, M., Shiina, T., Terachi, T., Utsugi, S., Murata, M., Mori, N., Takumi, S., Ikeo, K., Gojobori, T., Murai, R., Murai, K., Matsuoka, Y., Ohnishi, Y., Tajiri, H., & Tsunewaki, K. (2002). Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molecular Genetics and Genomics*, *266*(5), 740-746. https://doi.org/10.1007/s00438-001-0606-9

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., Terachi, T., Shiina, T., Miyashita, N., Nasuda, S., Nakamura, C., Mori, N., Takumi, S., Murata, M., Futo, S., & Tsunewaki, K. (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. *Nucleic Acids Research*, *33*(19), 6235-6250. https://doi.org/10.1093/nar/gki925

Oka, R., Zicola, J., Weber, B., Anderson, S. N., Hodgman, C., Gent, J. I., Wesselink, J. J., Springer, N. M., Hoefsloot, H. C. J., Turck, F., & Stam, M. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol*, *18*(1), 137. https://doi.org/10.1186/s13059-017-1273-4

Ortiz, E. M. (2019). *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. In (Version 2.0) Zenodo.

Ostergaard, L., & Yanofsky, M. F. (2004). Establishing gene function by mutagenesis in *Arabidopsis thaliana*. *Plant Journal*, *39*(5), 682-696. https://doi.org/10.1111/j.1365-313X.2004.02149.x

Ozsolak, F., Song, J. S., Liu, X. S., & Fisher, D. E. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, *25*(2), 244-248. https://doi.org/10.1038/nbt1279

Panahi, B., Hosseinzadeh Gharajeh, N., Mohammadzadeh Jalaly, H., & Golkari, S. (2024). Leveraging multi-omics and machine learning approaches in malting barley research: From farm cultivation to the final products. *Current Plant Biology*, *39*, 100362. https://doi.org/https://doi.org/10.1016/j.cpb.2024.100362

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of Gene Duplication in Plants. *Plant Physiology*, *171*(4), 2294-2316. https://doi.org/10.1104/pp.16.00523

Pang, Z., Xu, L., Viau, C., Lu, Y., Salavati, R., Basu, N., & Xia, J. (2024). MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. *Nat Commun*, *15*(1), 3675. https://doi.org/10.1038/s41467-024-48009-6

Parry, M. A., Reynolds, M., Salvucci, M. E., Raines, C., Andralojc, P. J., Zhu, X. G., Price, G. D., Condon, A. G., & Furbank, R. T. (2011). Raising yield potential of wheat. II. Increasing photosynthetic capacity and efficiency. *Journal of Experimental Botany*, *62*(2), 453-467. https://doi.org/10.1093/jxb/erq304

Parvathaneni, R. K., Bertolini, E., Shamimuzzaman, M., Vera, D. L., Lung, P. Y., Rice, B. R., Zhang, J., Brown, P. J., Lipka, A. E., Bass, H. W., & Eveland, A. L. (2020). The regulatory landscape of early maize inflorescence development. *Genome Biol*, *21*(1), 165. https://doi.org/10.1186/s13059-020-02070-8

Paul, M. J. (2021). Improving Photosynthetic Metabolism for Crop Yields: What Is Going to Work? *Frontiers in Plant Science*, *12*, 743862. https://doi.org/10.3389/fpls.2021.743862

Peleke, F. F., Zumkeller, S. M., Gultas, M., Schmitt, A., & Szymanski, J. (2024). Deep learning the cis-regulatory code for gene expression in selected model plants. *Nat Commun*, *15*(1), 3488. https://doi.org/10.1038/s41467-024-47744-0

Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., Beales, J., Fish, L. J., Worland, A. J., Pelica, F., Sudhakar, D., Christou, P., Snape, J. W., Gale, M. D., & Harberd, N. P. (1999). 'Green revolution' genes encode mutant gibberellin response modulators. *Nature*, *400*(6741), 256-261. https://doi.org/10.1038/22307

Philipp, N., Weichert, H., Bohra, U., Weschke, W., Schulthess, A. W., & Weber, H. (2018). Grain number and grain yield distribution along the spike remain stable despite breeding for high yield in winter wheat. *PLoS One*, *13*(10). https://doi.org/10.1371/journal.pone.0205452

Pizzolato, T. D. (1997). Procambial initiation for the vascular system in the spike of wheat. *International Journal of Plant Sciences*, *158*(2), 121-131. https://doi.org/Doi 10.1086/297421

Pizzolato, T. D. (1998). Procambial initiation for the vascular system in the spikelet of wheat. *International Journal of Plant Sciences*, *159*(1), 46-56. https://doi.org/Doi 10.1086/297520

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559-575. https://doi.org/10.1086/519795

Qin, Y., Long, Y., & Zhai, J. (2022). Genome-wide characterization of nascent RNA processing in plants. *Current Opinion in Plant Biology*, *69*, 102294. https://doi.org/10.1016/j.pbi.2022.102294

Qiu, F., Xing, S., Xue, C., Liu, J., Chen, K., Chai, T., & Gao, C. (2022). Transient expression of a TaGRF4-TaGIF1 complex stimulates wheat regeneration and improves genome editing. *Sci China Life Sci*, *65*(4), 731-738. https://doi.org/10.1007/s11427-021-1949-9

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

R2D2 Consortium, Fugeray-Scarbel, A., Bastien, C., Dupont-Nivet, M., & Lemarie, S. (2021). Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience. *Front Genet*, *12*, 629737. https://doi.org/10.3389/fgene.2021.629737

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. https://www.R-project.org

Rafiei, F., Wiersma, J., Scofield, S., Zhang, C., Alizadeh, H., & Mohammadi, M. (2024). Facts, uncertainties, and opportunities in wheat molecular improvement. *Heredity (Edinb)*, *133*(6), 371-380. https://doi.org/10.1038/s41437-024-00721-1

Rahman, A., Hallgrimsdottir, I., Eisen, M., & Pachter, L. (2018). Association mapping from sequencing reads using k-mers. *Elife*, *7*. https://doi.org/10.7554/eLife.32920

Rajarammohan, S., Kaur, L., Verma, A., Singh, D., Mantri, S., Roy, J. K., Sharma, T. R., Pareek, A., & Kandoth, P. K. (2023). Genome sequencing and assembly of Lathyrus sativus - a nutrient-rich hardy legume crop. *Sci Data*, *10*(1), 32. https://doi.org/10.1038/s41597-022-01903-4

Ramirez, F., Ryan, D. P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dundar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160-165. https://doi.org/10.1093/nar/gkw257

Rebetzke, G. J., Bonnett, D. G., & Reynolds, M. P. (2016). Awns reduce grain number to increase grain size and harvestable yield in irrigated and rainfed spring wheat. *Journal of Experimental Botany*, *67*(9), 2573-2586. https://doi.org/10.1093/jxb/erw081

Remans, R., Flynn, D. F., DeClerck, F., Diru, W., Fanzo, J., Gaynor, K., Lambrecht, I., Mudiope, J., Mutuo, P. K., Nkhoma, P., Siriri, D., Sullivan, C., & Palm, C. A. (2011). Assessing nutritional diversity of cropping systems in African villages. *PLoS One*, *6*(6), e21235. https://doi.org/10.1371/journal.pone.0021235

Ren, D., Ding, C., & Qian, Q. (2023). Molecular bases of rice grain size and quality for optimized productivity. *Sci Bull (Beijing)*, *68*(3), 314-350. https://doi.org/10.1016/j.scib.2023.01.026

Ren, D., Rao, Y., Yu, H., Xu, Q., Cui, Y., Xia, S., Yu, X., Liu, H., Hu, H., Xue, D., Zeng, D., Hu, J., Zhang, G., Gao, Z., Zhu, L., Zhang, Q., Shen, L., Guo, L., & Qian, Q. (2020). MORE FLORET1 Encodes a MYB Transcription Factor That Regulates Spikelet Development in Rice. *Plant Physiology*, *184*(1), 251-265. https://doi.org/10.1104/pp.20.00658

Ren, Y., He, Q., Ma, X., & Zhang, L. (2017). Characteristics of Color Development in Seeds of Brown- and Yellow-Seeded Heading Chinese Cabbage and Molecular Analysis of *Brsc*, the Candidate Gene Controlling Seed Coat Color [Original Research]. *Frontiers in Plant Science*, *Volume 8 - 2017*. https://doi.org/10.3389/fpls.2017.01410

Rey, M. D., Martin, A. C., Higgins, J., Swarbreck, D., Uauy, C., Shaw, P., & Moore, G. (2017). Exploiting the *ZIP4* homologue within the wheat *Ph1* locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids. *Molecular Breeding*, *37*(8), 95. https://doi.org/10.1007/s11032-017-0700-2

Reynolds, M., Atkin, O. K., Bennett, M., Cooper, M., Dodd, I. C., Foulkes, M. J., Frohberg, C., Hammer, G., Henderson, I. R., Huang, B., Korzun, V., McCouch, S. R., Messina, C. D., Pogson, B. J., Slafer, G. A., Taylor, N. L., & Wittich, P. E. (2021). Addressing Research Bottlenecks to Crop Productivity. *Trends in Plant Science*, *26*(6), 607-630. https://doi.org/10.1016/j.tplants.2021.03.011

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., Noshay, J. M., Galli, M., Mejía-Guerra, M. K., Colomé-Tatché, M., Johannes, F., Rowley, M. J., Corces, V. G., Zhai, J., Scanlon, M. J., Buckler, E. S., Gallavotti, A., Springer, N. M., Schmitz, R. J., & Zhang, X. (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants*, *5*(12), 1237-1249. https://doi.org/10.1038/s41477-019-0547-0

Rickner, H. D., Niu, S. Y., & Cheng, C. S. (2019). ATAC-seq Assay with Low Mitochondrial DNA Contamination from Primary Human CD4+T Lymphocytes. *Jove-Journal of Visualized Experiments*(145). https://doi.org/10.3791/59120

Risk, J. M., Selter, L. L., Chauhan, H., Krattinger, S. G., Kumlehn, J., Hensel, G., Viccars, L. A., Richardson, T. M., Buesing, G., Troller, A., Lagudah, E. S., & Keller, B. (2013). The wheat Lr34 gene provides resistance against multiple fungal pathogens in barley. *Plant Biotechnology Journal*, *11*(7), 847-854. https://doi.org/10.1111/pbi.12077

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Rodriguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., & Lippman, Z. B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell*, *171*(2), 470-480 e478. https://doi.org/10.1016/j.cell.2017.08.030

Roeder, A. H., Ferrandiz, C., & Yanofsky, M. F. (2003). The role of the REPLUMLESS homeodomain protein in patterning the *Arabidopsis* fruit. *Current Biology*, *13*(18), 1630-1635. https://doi.org/10.1016/j.cub.2003.08.027

Romani, F., Sauret-Güeto, S., Rebmann, M., Annese, D., Bonter, I., Tomaselli, M., Dierschke, T., Delmans, M., Frangedakis, E., Silvestri, L., Rever, J., Bowman, J. L., Romani, I., & Haseloff, J. (2024). The landscape of transcription factor promoter activity during vegetative development in Marchantia. *Plant Cell*, *36*(6), 2140-2159. https://doi.org/10.1093/plcell/koae053

Rosado-Souza, L., Yokoyama, R., Sonnewald, U., & Fernie, A. R. (2023). Understanding source-sink interactions: Progress in model plants and translational research to crops. *Mol Plant*, *16*(1), 96-121. https://doi.org/10.1016/j.molp.2022.11.015

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., Stacey, G., Doerge, R. W., & Jackson, S. A. (2013). The fate of duplicated genes in a polyploid plant genome. *Plant Journal*, *73*(1), 143-153. https://doi.org/10.1111/tpj.12026

Ruggeri, R., Rossini, F., Ronchi, B., Primi, R., Stamigna, C., & Danieli, P. P. (2024). Potential of teff as alternative crop for Mediterranean farming systems: Effect of genotype and mowing time on forage yield and quality. *Journal of Agriculture and Food Research*, *17*. https://doi.org/10.1016/j.jafr.2024.101257

Sadras, V. O. (2007). Evolutionary aspects of the trade-off between seed size and number in crops. *Field Crops Research*, *100*(2), 125-138. https://doi.org/https://doi.org/10.1016/j.fcr.2006.07.004

Sakuma, S., Golan, G., Guo, Z., Ogawa, T., Tagiri, A., Sugimoto, K., Bernhardt, N., Brassac, J., Mascher, M., Hensel, G., Ohnishi, S., Jinno, H., Yamashita, Y., Ayalon, I., Peleg, Z., Schnurbusch, T., & Komatsuda, T. (2019). Unleashing floret fertility in wheat through the mutation of a homeobox gene. *Proc Natl Acad Sci U S A*, *116*(11), 5182-5187. https://doi.org/10.1073/pnas.1815465116

Sakuma, S., & Schnurbusch, T. (2020). Of floral fortune: tinkering with the grain yield potential of cereal crops. *New Phytologist*, *225*(5), 1873-1882. https://doi.org/10.1111/nph.16189

Saladino, D. (2021). *Eating to Extinction: The World's Rarest Foods and Why We Need to Save Them*. Farrar, Straus and Giroux.

Sanchez-Bragado, R., Molero, G., Reynolds, M. P., & Araus, J. L. (2016). Photosynthetic contribution of the ear to grain filling in wheat: a comparison of different methodologies for evaluation. *Journal of Experimental Botany*, *67*(9), 2787-2798. https://doi.org/10.1093/jxb/erw116

Sandve, S. R., Marcussen, T., Mayer, K., Jakobsen, K. S., Heier, L., Steuernagel, B., Wulff, B. B. H., & Olsen, O. A. (2015). Chloroplast phylogeny of *Triticum/Aegilops* species is not incongruent with an ancient homoploid hybrid origin of the ancestor of the bread wheat D-genome. *New Phytologist*, *208*(1), 9-10. https://doi.org/10.1111/nph.13487

Sankaranarayanan, S., Zhang, Y., Carney, J., Nigussie, Y., Esayas, B., Simane, B., Zaitchik, B., & Siddiqui, S. (2020). What Are the Domestic and Regional Impacts From Ethiopia's Policy on the Export Ban of Teff? *Frontiers in Sustainable Food Systems*, *4*. https://doi.org/10.3389/fsufs.2020.00004

Sauret-Güeto, S., Frangedakis, E., Silvestri, L., Rebmann, M., Tomaselli, M., Markel, K., Delmans, M., West, A., Patron, N. J., & Haseloff, J. (2020). Systematic Tools for Reprogramming Plant Gene Expression in a Simple Model, *Marchantia polymorpha*. *ACS Synthetic Biology*, *9*(4), 864-882. https://doi.org/10.1021/acssynbio.9b00511

Schmitz, R. J., Grotewold, E., & Stam, M. (2021). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *The Plant Cell*, *34*(2), 718-741. https://doi.org/10.1093/plcell/koab281

Schrager-Lavelle, A., Klein, H., Fisher, A., & Bartlett, M. (2017). Grass flowers: An untapped resource for floral evo-devo. *Journal of Systematics and Evolution*, *55*(6), 525-541. https://doi.org/https://doi.org/10.1111/jse.12251

Schreiber, M., Jayakodi, M., Stein, N., & Mascher, M. (2024). Plant pangenomes for crop improvement, biodiversity and evolution. *Nature Reviews: Genetics*, *25*(8), 563-577. https://doi.org/10.1038/s41576-024-00691-4

Schürholz, A. K., López-Salmerón, V., Li, Z. N., Forner, J., Wenzl, C., Gaillochet, C., Augustin, S., Barro, A. V., Fuchs, M., Gebert, M., Lohmann, J. U., Greb, T., & Wolf, S. (2018). A Comprehensive Toolkit for Inducible, Cell Type-Specific Gene Expression in Arabidopsis. *Plant Physiology*, *178*(1), 40-53. https://doi.org/10.1104/pp.18.00463

Seki, M., Chono, M., Matsunaka, H., Fujita, M., Oda, S., Kubo, K., Kiribuchi-Otobe, C., Kojima, H., Nishida, H., & Kato, K. (2011). Distribution of photoperiod-insensitive alleles Ppd-B1a and Ppd-D1a and their effect on heading time in Japanese wheat cultivars. *Breed Sci*, *61*(4), 405-412. https://doi.org/10.1270/jsbbs.61.405

Semenov, M. A., & Stratonovitch, P. (2013). Designing high-yielding wheat ideotypes for a changing climate. *Food and Energy Security*, *2*(3), 185-196. https://doi.org/https://doi.org/10.1002/fes3.34

Senapati, N., Brown, H. E., & Semenov, M. A. (2019). Raising genetic yield potential in high productive countries: Designing wheat ideotypes under climate change. *Agric For Meteorol*, *271*, 33-45. https://doi.org/10.1016/j.agrformet.2019.02.025

Sentoku, N., Kato, H., Kitano, H., & Imai, R. (2005). *OsMADS22*, an *STMADS11*-like MADS-box gene of rice, is expressed in non-vegetative tissues and its ectopic expression induces spikelet meristem indeterminacy. *Molecular Genetics and Genomics*, *273*(1), 1-9. https://doi.org/10.1007/s00438-004-1093-6

Sharma, R., Liang, Y., Lee, M. Y., Pidatala, V. R., Mortimer, J. C., & Scheller, H. V. (2020). Agrobacterium-mediated transient transformation of sorghum leaves for accelerating functional genomics and genome editing studies. *BMC Res Notes*, *13*(1), 116. https://doi.org/10.1186/s13104-020-04968-9

Sheffield, N. C., & Furey, T. S. (2012). Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. *Genes (Basel)*, *3*(4), 651-670. https://doi.org/10.3390/genes3040651

Sheffield, N. C., Thurman, R. E., Song, L., Safi, A., Stamatoyannopoulos, J. A., Lenhard, B., Crawford, G. E., & Furey, T. S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*, *23*(5), 777-788. https://doi.org/10.1101/gr.152140.112

Shorinola, O., Marks, R., Emmrich, P., Jones, C., Odeny, D., & Chapman, M. A. (2024). Integrative and inclusive genomics to promote the use of underutilised crops. *Nature Communications*, *15*(1), 320. https://doi.org/10.1038/s41467-023-44535-x

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall / CRC. https://plotly-r.com

Simmonds, J., Scott, P., Brinton, J., Mestre, T. C., 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*(6), 1099-1112. https://doi.org/10.1007/s00122-016-2686-2

Simons, K. J., Fellers, J. P., Trick, H. N., Zhang, Z., Tai, Y. S., Gill, B. S., & Faris, J. D. (2006). Molecular characterization of the major wheat domestication gene Q. *Genetics*, *172*(1), 547-555. https://doi.org/10.1534/genetics.105.044727

Singha, D. L., Das, D., Sarki, Y. N., Chowdhury, N., Sharma, M., Maharana, J., & Chikkaputtaiah, C. (2022). Harnessing tissue-specific genome editing in plants through CRISPR/Cas system: current state and future prospects. *Planta*, *255*(1). https://doi.org/10.1007/s00425-021-03811-0

Slade, A. J., McGuire, C., Loeffler, D., Mullenberg, J., Skinner, W., Fazio, G., Holm, A., Brandt, K. M., Steine, M. N., Goodstal, J. F., & Knauf, V. C. (2012). Development of high amylose wheat through TILLING. *BMC Plant Biology*, *12*, 69. https://doi.org/10.1186/1471-2229-12-69

Slafer, G. A., Foulkes, M. J., Reynolds, M. P., Murchie, E. H., Carmo-Silva, E., Flavell, R., Gwyn, J., Sawkins, M., & Griffiths, S. (2023). A 'wiring diagram' for sink strength traits impacting wheat yield potential. *Journal of Experimental Botany*, *74*(1), 40-71. https://doi.org/10.1093/jxb/erac410

Smedley, M. A., Hayta, S., Clarke, M., & Harwood, W. A. (2021). CRISPR-Cas9 Based Genome Editing in Wheat. *Current Protocols*, *1*(3). https://doi.org/10.1002/cpz1.65

Smith, H. M., & Hake, S. (2003). The interaction of two homeobox genes, BREVIPEDICELLUS and PENNYWISE, regulates internode patterning in the Arabidopsis inflorescence. *Plant Cell*, *15*(8), 1717-1727. https://doi.org/10.1105/tpc.012856

Smith, J. P., Corces, M. R., Xu, J., Reuter, V. P., Chang, H. Y., & Sheffield, N. C. (2021). PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments. *Nar Genomics and Bioinformatics*, *3*(4). https://doi.org/10.1093/nargab/lqab101

Smith, R. H., & Hood, E. E. (1995). Agrobacterium tumefaciens Transformation of Monocotyledons. *Crop Science*, *35*(2), cropsci1995.0011183X003500020001x. https://doi.org/https://doi.org/10.2135/cropsci1995.0011183X003500020001x

Song, X., Meng, X., Guo, H., Cheng, Q., Jing, Y., Chen, M., Liu, G., Wang, B., Wang, Y., Li, J., & Yu, H. (2022). Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. *Nature Biotechnology*, *40*(9), 1403-1411. https://doi.org/10.1038/s41587-022-01281-7

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. https://doi.org/10.1038/ng2014

Sonnewald, U., & Fernie, A. R. (2018). Next-generation strategies for understanding and influencing source-sink relations in crop plants. *Current Opinion in Plant Biology*, *43*, 63-70. https://doi.org/10.1016/j.pbi.2018.01.004

Springer, P. S. (2000). Gene traps: tools for plant development and genomics. *Plant Cell*, *12*(7), 1007-1020. https://doi.org/10.1105/tpc.12.7.1007

Stallknecht, G. F., Gilbertson, K. M., & Eckhoff, J. L. (1993). Tef: Food Crop for Humans and Animals. In J. Jamick & J. E. Simon (Eds.), *New Crops*. Wiley.

Stella, T., Webber, H., Eyshi Rezaei, E., Asseng, S., Martre, P., Dueri, S., Rafael Guarin, J., Pequeno, D. N. L., Calderini, D. F., Reynolds, M., Molero, G., Miralles, D., Garcia, G., Slafer, G., Giunta, F., Kim, Y.-U., Wang, C., Ruane, A. C., & Ewert, F. (2023). Wheat crop traits conferring high yield potential may also improve yield stability under climate change. *in silico Plants*, *5*(2), diad013. https://doi.org/10.1093/insilicoplants/diad013

Stockman, Y. M., Fischer, R. A., & Brittain, E. G. (1983). Assimilate Supply and Floret Development within the Spike of Wheat (Triticum-Aestivum L). *Australian Journal of Plant Physiology*, *10*(6), 585-594. https://doi.org/Doi 10.1071/Pp9830585

Streich, J., Romero, J., Gazolla, J. G. F. M., Kainer, D., Cliff, A., Prates, E. T., Brown, J. B., Khoury, S., Tuskan, G. A., Garvin, M., Jacobson, D., & Harfouche, A. L. (2020). Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Current Opinion in Biotechnology*, *61*, 217-225. https://doi.org/https://doi.org/10.1016/j.copbio.2020.01.010

Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., Sandstrom, R., Thurman, R. E., Neph, S., Reynolds, A. P., Stergachis, A. B., Vernot, B., Johnson, A. K., Haugen, E., Sullivan, S. T., Thompson, A., Neri, F. V., 3rd, Weaver, M., Diegel, M.,...Stamatoyannopoulos, J. A. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep*, *8*(6), 2015-2030. https://doi.org/10.1016/j.celrep.2014.08.019

Suzaki, T., Tsuda, M., Ezura, H., Day, B., & Miura, K. (2019). Agroinfiltration-based efficient transient protein expression in leguminous plants. *Plant Biotechnol (Tokyo)*, *36*(2), 119-123. https://doi.org/10.5511/plantbiotechnology.19.0220b

Tadele, E., & Hibistu, T. (2021). Empirical review on the use dynamics and economics of teff in Ethiopia. *Agriculture & Food Security*, *10*(1), 40. https://doi.org/10.1186/s40066-021-00329-2

Takai, T. (2024). Potential of rice tillering for sustainable food production. *Journal of Experimental Botany*, *75*(3), 708-720. https://doi.org/10.1093/jxb/erad422

Takai, T., Taniguchi, Y., Takahashi, M., Nagasaki, H., Yamamoto, E., Hirose, S., Hara, N., Akashi, H., Ito, J., Arai-Sanoh, Y., Hori, K., Fukuoka, S., Sakai, H., Tokida, T., Usui, Y., Nakamura, H., Kawamura, K., Asai, H., Ishizaki, T.,...Uga, Y. (2023). *MORE PANICLES 3*, a natural allele of *OsTB1/FC1*, impacts rice yield in paddy fields at elevated CO_2 levels. *Plant Journal*, *114*(4), 729-742. https://doi.org/10.1111/tpj.16143

Tamagno, S., Carrera, C. S., Marchese, S. I., Savin, R., & Slafer, G. A. (2024). Sterility of basal spikelets in wheat: predetermined fate or a matter of resources? *Journal of Experimental Botany*, *75*(22), 7160-7173. https://doi.org/10.1093/jxb/erae373

Tao, F., Rötter, R. P., Palosuo, T., Díaz-Ambrona, C. G. H., Mínguez, M. I., Semenov, M. A., Kersebaum, K. C., Nendel, C., Cammarano, D., Hoffmann, H., Ewert, F., Dambreville, A., Martre, P., Rodríguez, L., Ruiz-Ramos, M., Gaiser, T., Höhn, J. G., Salo, T., Ferrise, R.,...Schulman, A. H. (2017). Designing future barley ideotypes using a crop model ensemble. *European Journal of Agronomy*, *82*, 144-162. https://doi.org/https://doi.org/10.1016/j.eja.2016.10.012

Tarbell, E. D., & Liu, T. (2019). HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research*, *47*(16). https://doi.org/10.1093/nar/gkz533

Tay Fernandez, C. G., Nestor, B. J., Danilevicz, M. F., Gill, M., Petereit, J., Bayer, P. E., Finnegan, P. M., Batley, J., & Edwards, D. (2022). Pangenomes as a Resource to Accelerate Breeding of Under-Utilised Crop Species. *International Journal of Molecular Sciences*, *23*(5). https://doi.org/10.3390/ijms23052671

Thiel, J., Koppolu, R., Trautewig, C., Hertig, C., Kale, S. M., Erbe, S., Mascher, M., Himmelbach, A., Rutten, T., Esteban, E., Pasha, A., Kumlehn, J., Provart, N. J., Vanderauwera, S., Frohberg, C., & Schnurbusch, T. (2021). Transcriptional landscapes of floral meristems in barley. *Science Advances*, *7*(18). https://doi.org/10.1126/sciadv.abf0832

Thomson, N., Evert, R. F., & Kelman, A. (1995). Wound healing in whole potato tubers: a cytochemical, fluorescence, and ultrastructural analysis of cut and bruise wounds. *Canadian Journal of Botany*, *73*(9), 1436-1450. https://doi.org/10.1139/b95-156

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K.,...Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75-82. https://doi.org/10.1038/nature11232

Tiguh, E. E., Delele, M. A., Ali, A. N., Kidanemariam, G., & Fanta, S. W. (2024). Assessment of harvest and postharvest losses of teff (*Eragrostis tef (Zucc.)*) and methods of loss reduction: A review. *Heliyon*, *10*(9), e30398. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e30398

Trevaskis, B., Tadege, M., Hemming, M. N., Peacock, W. J., Dennis, E. S., & Sheldon, C. (2007). *Short Vegetative Phase*-like MADS-box genes inhibit floral meristem identity in barley. *Plant Physiology*, *143*(1), 225-235. https://doi.org/10.1104/pp.106.090860

Tsuda, K., Abraham-Juarez, M. J., Maeno, A., Dong, Z., Aromdee, D., Meeley, R., Shiroishi, T., Nonomura, K. I., & Hake, S. (2017). KNOTTED1 Cofactors, BLH12 and BLH14, Regulate Internode Patterning and Vein Anastomosis in Maize. *Plant Cell*, *29*(5), 1105-1118. https://doi.org/10.1105/tpc.16.00967

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. https://doi.org/10.1126/science.1133649

Uauy, C., Paraiso, F., Colasuonno, P., Tran, R. K., Tsai, H., Berardi, S., Comai, L., & Dubcovsky, J. (2009). A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biology*, *9*, 115. https://doi.org/10.1186/1471-2229-9-115

Uauy, C., Wulff, B. B. H., & Dubcovsky, J. (2017). Combining Traditional Mutagenesis with New High-Throughput Sequencing and Genome Editing to Reveal Hidden Variation in Polyploid Wheat. *Annual Review of Genetics*, *51*, 435-454. https://doi.org/10.1146/annurev-genet-120116-024533

Uygun, S., Azodi, C. B., & Shiu, S. H. (2019). Cis-Regulatory Code for Predicting Plant Cell-Type Transcriptional Response to High Salinity. *Plant Physiology*, *181*(4), 1739-1751. https://doi.org/10.1104/pp.19.00653

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews: Genetics*, *18*(7), 411-424. https://doi.org/10.1038/nrg.2017.26

van Dijk, M., Morley, T., Rau, M. L., & Saghai, Y. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010-2050. *Nat Food*, *2*(7), 494-501. https://doi.org/10.1038/s43016-021-00322-9

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A. E., Wang, H., Chaluvadi, S. R., Han, G., Bryant, D., Edger, P. P., Messing, J., Sorrells, M. E., Mockler, T. C., Bennetzen, J. L., & Michael, T. P. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid

Ethiopian cereal teff. *Nature Communications*, *11*(1), 884. https://doi.org/10.1038/s41467-020-14724-z

Vandeputte, W., Coussens, G., Aesaert, S., Haeghebaert, J., Impens, L., Karimi, M., Debernardi, J. M., & Pauwels, L. (2024). Use of GRF-GIF chimeras and a ternary vector system to improve maize (Zea mays L.) transformation frequency. *Plant Journal*, *119*(4), 2116-2132. https://doi.org/10.1111/tpj.16880

Varshney, R. K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., Singh, N. P., Du, X., Upadhyaya, H. D., Khan, A. W., Wang, Y., Garg, V., Fan, G., Cowling, W. A., Crossa, J., Gentzbittel, L., Voss-Fels, K. P., Valluri, V. K., Sinha, P.,...Liu, X. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature*, *599*(7886), 622-627. https://doi.org/10.1038/s41586-021-04066-1

Vedel, V., & Scotti, I. (2011). Promoting the promoter. *Plant Science*, *180*(2), 182-189. https://doi.org/10.1016/j.plantsci.2010.09.009

Vermeulen, S. J., Park, T., Khoury, C. K., & Bene, C. (2020). Changing diets and the transformation of the global food system. *Annals of the New York Academy of Sciences*, *1478*(1), 3-17. https://doi.org/10.1111/nyas.14446

Vetriventhan, M., Azevedo, V. C. R., Upadhyaya, H. D., Nirmalakumari, A., Kane-Potaka, J., Anitha, S., Ceasar, S. A., Muthamilarasan, M., Bhat, B. V., Hariprasanna, K., Bellundagi, A., Cheruku, D., Backiyalakshmi, C., Santra, D., Vanniarajan, C., & Tonapi, V. A. (2020). Genetic and genomic resources, and breeding for accelerating improvement of small millets: current status and future interventions. *The Nucleus*, *63*(3), 217-239. https://doi.org/10.1007/s13237-020-00322-3

Villiger, L., Joung, J., Koblan, L., Weissman, J., Abudayyeh, O. O., & Gootenberg, J. S. (2024). CRISPR technologies for genome, epigenome and transcriptome editing. *Nature Reviews: Molecular Cell Biology*, *25*(6), 464-487. https://doi.org/10.1038/s41580-023-00697-6

Waddington, S. R., Cartwright, P. M., & Wall, P. C. (1983). A Quantitative Scale of Spike Initial and Pistil Development in Barley and Wheat. *Annals of Botany*, *51*(1), 119-130. https://doi.org/10.1093/oxfordjournals.aob.a086434

Wagali, P., Ngomuo, G., Kilama, J., Sabastian, C., Ben-Zeev, S., Ben-Meir, Y. A., Argov-Argaman, N., Saranga, Y., & Mabjeesh, S. J. (2023). The effect of teff (*Eragrostis tef*) hay inclusion on feed intake, digestibility, and milk production in dairy cows. *Frontiers in Animal Science*, *4*. https://doi.org/10.3389/fanim.2023.1260787

Walker, B. J., Drewry, D. T., Slattery, R. A., VanLoocke, A., Cho, Y. B., & Ort, D. R. (2018). Chlorophyll Can Be Reduced in Crop Canopies with Little Penalty to Photosynthesis. *Plant Physiology*, *176*(2), 1215-1232. https://doi.org/10.1104/pp.17.01401

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D.,...Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, *588*(7837), 277-283. https://doi.org/10.1038/s41586-020-2961-x

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., Nery, J. R., Smith, L. G., Schnable, J. C., Ecker, J. R., & Briggs, S. P. (2016). Integration of omic networks in a developmental atlas of maize. *Science*, *353*(6301), 814-818. https://doi.org/10.1126/science.aag1125

Wallis, J. G., Bengtsson, J. D., & Browse, J. (2022). Molecular Approaches Reduce Saturates and Eliminate trans Fats in Food Oils. *Frontiers in Plant Science*, *13*, 908608. https://doi.org/10.3389/fpls.2022.908608

Walsh, B., & Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford University Press. https://doi.org/10.1093/oso/9780198830870.001.0001

Wang, J., Qin, Q., Pan, J. J., Sun, L. J., Sun, Y. F., Xue, Y., & Song, K. (2019). Transcriptome analysis in roots and leaves of wheat seedlings in response to low-phosphorus stress. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-019-56451-6

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics Proteomics Bioinformatics*, *19*(4), 629-640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, M., Li, Z., Zhang, Y., Zhang, Y., Xie, Y., Ye, L., Zhuang, Y., Lin, K., Zhao, F., Guo, J., Teng, W., Zhang, W., Tong, Y., Xue, Y., & Zhang, Y. (2021). An atlas of wheat epigenetic regulatory elements reveals subgenome divergence in the regulation of development and stress responses. *Plant Cell*, *33*(4), 865-881. https://doi.org/10.1093/plcell/koab028

Wang, N., Long, T., Yao, W., Xiong, L., Zhang, Q., & Wu, C. (2013). Mutant resources for the functional analysis of the rice genome. *Mol Plant*, *6*(3), 596-604. https://doi.org/10.1093/mp/sss142

Wang, N., Ryan, L., Sardesai, N., Wu, E., Lenderts, B., Lowe, K., Che, P., Anand, A., Worden, A., van Dyk, D., Barone, P., Svitashev, S., Jones, T., & Gordon-Kamm, W. (2023). Leaf transformation for efficient random integration and targeted genome modification in maize and sorghum. *Nat Plants*, *9*(2), 255-270. https://doi.org/10.1038/s41477-022-01338-0

Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C. P., Bandyopadhyay, A., Quick, W. P., Furbank, R., Sage, T. L., & Langdale, J. A. (2017). Re-creation of a Key Step in the Evolutionary Switch from C_3 to C_4 Leaf Anatomy. *Current Biology*, *27*(21), 3278-+. https://doi.org/10.1016/j.cub.2017.09.040

Wang, W., Simmonds, J., Pan, Q., Davidson, D., He, F., Battal, A., Akhunova, A., Trick, H. N., Uauy, C., & Akhunov, E. (2018). Gene editing and mutagenesis reveal inter-cultivar differences and additivity in the contribution of *TaGW2* homoeologues to grain size and weight in wheat. *Theoretical and Applied Genetics*, *131*(11), 2463-2475. https://doi.org/10.1007/s00122-018-3166-7

Wang, Y. G., Yu, H. P., Tian, C. H., Sajjad, M., Gao, C. C., Tong, Y. P., Wang, X. F., & Jiao, Y. L. (2017). Transcriptome Association Identifies Regulators of Wheat Spike Architecture. *Plant Physiology*, *175*(2), 746-757. https://doi.org/10.1104/pp.17.00694

Weber, B., Zicola, J., Oka, R., & Stam, M. (2016). Plant Enhancers: A Call for Discovery. *Trends in Plant Science*, *21*(11), 974-987. https://doi.org/10.1016/j.tplants.2016.07.013

Wei, X., Xiang, Y., Peters, D. T., Marius, C., Sun, T., Shan, R., Ou, J., Lin, X., Yue, F., Li, W., Southerland, K. W., & Diao, Y. (2022). HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. *Molecular Cell*, *82*(6), 1225-1238 e1226. https://doi.org/10.1016/j.molcel.2022.01.023

Weiner, J., Du, Y. L., Zhao, Y. M., & Li, F. M. (2021). Allometry and Yield Stability of Cereals. *Frontiers in Plant Science*, *12*, 681490. https://doi.org/10.3389/fpls.2021.681490

Whingwiri, E. E., Kuo, J., & Stern, W. R. (1981). The Vascular System in the Rachis of a Wheat Ear. *Annals of Botany*, *48*(2), 189-201. https://doi.org/DOI 10.1093/oxfordjournals.aob.a086113

Wicker, T., & Keller, B. (2007). Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Research*, *17*(7), 1072-1081. https://doi.org/10.1101/gr.6214107

Wilhelm, E. P., Turner, A. S., & Laurie, D. A. (2009). Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (Triticum durum Desf.). *Theoretical and Applied Genetics*, *118*(2), 285-294. https://doi.org/10.1007/s00122-008-0898-9

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L. J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J. A., De Vries, W., Majele Sibanda, L.,...Murray, C. J. L. (2019). Food in the Anthropocene: the EAT-*Lancet* Commission on healthy diets from sustainable food systems. *The Lancet*, *393*(10170), 447-492. https://doi.org/10.1016/S0140-6736(18)31788-4

Winfield, M., Burridge, A., Ordidge, M., Harper, H., Wilkinson, P., Thorogood, D., Copas, L., Edwards, K., & Barker, G. (2020). Development of a minimal KASP marker panel for distinguishing genotypes in apple collections. *PLoS One*, *15*(11), e0242940. https://doi.org/10.1371/journal.pone.0242940

Woldeyohannes, A. B., Desta, E. A., Fadda, C., Pè, M. E., & Dell'Acqua, M. (2022). Value of teff (*Eragrostis tef*) genetic resources to support breeding for conventional and smallholder farming: a review. *Cabi Agriculture & Bioscience*, *3*(1). https://doi.org/10.1186/s43170-022-00076-9

Woldeyohannes, A. B., Iohannes, S. D., Miculan, M., Caproni, L., Ahmed, J. S., de Sousa, K., Desta, E. A., Fadda, C., Pè, M. E., & Dell'Acqua, M. (2022). Data-driven, participatory characterization of farmer varieties discloses teff breeding potential under current and future climates. *Elife*, *11*. https://doi.org/10.7554/eLife.80009

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., 2nd, Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, *21*(1), 385. https://doi.org/10.1186/s12870-021-03173-5

World Population Prospects (Version Online Edition). (2024). https://population.un.org/wpp/downloads?folder=Standard%20Projections&group=Population

Wu, C., Luo, J., & Xiao, Y. (2024). Multi-omics assists genomic prediction of maize yield with machine learning approaches. *Molecular Breeding*, *44*(2), 14. https://doi.org/10.1007/s11032-024-01454-z

Wurschum, T., Longin, C. F., Hahn, V., Tucker, M. R., & Leiser, W. L. (2017). Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat. *Plant Journal*, *89*(4), 764-773. https://doi.org/10.1111/tpj.13424

Xie, Y., Chen, Y., Li, Z., Zhu, J., Liu, M., Zhang, Y., & Dong, Z. (2022). Enhancer transcription detected in the nascent transcriptomic landscape of bread wheat. *Genome Biol*, *23*(1), 109. https://doi.org/10.1186/s13059-022-02675-1

Xu, S. B., Dai, Z. H., Guo, P. F., Fu, X. C., Liu, S. S., Zhou, L., Tang, W. L., Feng, T. Z., Chen, M. J., Zhan, L., Wu, T. Z., Hu, E. Q., Jiang, Y., Bo, X. C., & Yu, G. C. (2021). ggtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Molecular Biology and Evolution*, *38*(9), 4039-4042. https://doi.org/10.1093/molbev/msab166

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol*, *21*(1), 22. https://doi.org/10.1186/s13059-020-1929-3

Yan, H., Sun, M., Zhang, Z., Jin, Y., Zhang, A., Lin, C., Wu, B., He, M., Xu, B., Wang, J., Qin, P., Mendieta, J. P., Nie, G., Wang, J., Jones, C. S., Feng, G., Srivastava, R. K., Zhang, X., Bombarely, A.,...Huang, L. (2023). Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nature Genetics*, *55*(3), 507-518. https://doi.org/10.1038/s41588-023-01302-4

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., & Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene *VRN1*. *Proc Natl Acad Sci U S A*, *100*(10), 6263-6268. https://doi.org/10.1073/pnas.0937399100

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., Pandey, M. K., Ge, S., Xu, Q., Li, N., Li, G., Huang, Y., Saxena, R. K., Ji, Y., Li, M., Yan, X., He, Y., Liu, Y., Wang, X.,...Zong, X. (2022). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nature Genetics*, *54*(10), 1553-1563. https://doi.org/10.1038/s41588-022-01172-2

Yao, Y., Guo, W., Gou, J., Hu, Z., Liu, J., Ma, J., Zong, Y., Xin, M., Chen, W., Li, Q., Wang, Z., Zhang, R., Uauy, C., Baloch, F. S., Ni, Z., & Sun, Q. (2025). Wheat2035: Integrating pan-omics and advanced biotechnology for future wheat design. *Mol Plant*, *18*(2), 272-297. https://doi.org/10.1016/j.molp.2025.01.005

Yaschenko, A. E., Alonso, J. M., & Stepanova, A. N. (2024). Arabidopsis as a model for translational research. *Plant Cell*. https://doi.org/10.1093/plcell/koae065

Yassitepe, J. E. D. T., da Silva, V. C. H., Hernandes-Lopes, J., Dante, R. A., Gerhardt, I. R., Fernandes, F. R., da Silva, P. A., Vieira, L. R., Bonatti, V., & Arruda, P. (2021). Maize Transformation: From Plant Material to the Release of Genetically Modified and Edited Varieties. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.766702

Ye, X., Shrawat, A., Moeller, L., Rode, R., Rivlin, A., Kelm, D., Martinell, B. J., Williams, E. J., Paisley, A., Duncan, D. R., & Armstrong, C. L. (2023). *Agrobacterium*-mediated direct transformation of wheat mature embryos through organogenesis. *Frontiers in Plant Science*, *14*, 1202235. https://doi.org/10.3389/fpls.2023.1202235

Yeap, W. C., Lee, F. C., Shabari Shan, D. K., Musa, H., Appleton, D. R., & Kulaveerasingam, H. (2017). WRI1-1, ABI5, NF-YA3 and NF-YC2 increase oil biosynthesis in coordination with hormonal signaling during fruit development in oil palm. *Plant Journal*, *91*(1), 97-113. https://doi.org/10.1111/tpj.13549

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28-36. https://doi.org/https://doi.org/10.1111/2041-210X.12628

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881-892. https://doi.org/10.1111/pbi.13022

Yu, Y., Beyene, G., Villmer, J., Duncan, K. E., Hu, H., Johnson, T., Doust, A. N., Taylor, N. J., & Kellogg, E. A. (2023). Grain shattering by cell death and fracture in Eragrostis tef. *Plant Physiology*, *192*(1), 222-239. https://doi.org/10.1093/plphys/kiad079

Yuan, J., Sun, H., Wang, Y., Li, L., Chen, S., Jiao, W., Jia, G., Wang, L., Mao, J., Ni, Z., Wang, X., & Song, Q. (2022). Open chromatin interaction maps reveal functional regulatory elements and chromatin architecture variations during wheat evolution. *Genome Biol*, *23*(1), 34. https://doi.org/10.1186/s13059-022-02611-3

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A.,...Mouse, E. C. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355-364. https://doi.org/10.1038/nature13992

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., Argillier, O., Stiewe, G., Hinze, M., Neumann, F., Eichhorn, A., Polley, A., Jaenecke, C., Ganal, M. W., & Roder, M. S. (2015). Analysis of main effect QTL for thousand grain weight in European winter wheat (*Triticum aestivum* L.) by genome-wide association mapping. *Frontiers in Plant Science*, *6*, 644. https://doi.org/10.3389/fpls.2015.00644

Zeng, H., Daniel, T. C., Lingineni, A., Chee, K., Talloo, K., & Gao, X. (2024). Recent advances in prime editing technologies and their promises for therapeutic applications. *Current Opinion in Biotechnology*, *86*, 103071. https://doi.org/10.1016/j.copbio.2024.103071

Zhang, L., Zhang, Z., Tao, F., Luo, Y., Cao, J., Li, Z., Xie, R., & Li, S. (2021). Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning. *Environmental Research Letters*, *16*(12), 124043. https://doi.org/10.1088/1748-9326/ac32fd

Zhang, L., Zhang, Z., Tao, F., Luo, Y., Zhang, J., & Cao, J. (2022). Adapting to climate change precisely through cultivars renewal for rice production across China: When, where, and what cultivars will be required? *Agricultural and Forest Meteorology*, *316*, 108856. https://doi.org/https://doi.org/10.1016/j.agrformet.2022.108856

Zhang, N., Tang, L., Li, S., Liu, L., Gao, M., Wang, S., Chen, D., Zhao, Y., Zheng, R., Soleymaniniya, A., Zhang, L., Wang, W., Yang, X., Ren, Y., Sun, C., Wilhelm, M., Wang, D., Li, M., & Chen, F. (2025). Integration of multi-omics data accelerates molecular analysis of common wheat traits. *Nat Commun*, *16*(1), 2200. https://doi.org/10.1038/s41467-025-57550-x

Zhang, W., Wu, Y., Schnable, J. C., Zeng, Z., Freeling, M., Crawford, G. E., & Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. *Genome Research*, *22*(1), 151-162. https://doi.org/10.1101/gr.131342.111

Zhang, W., Zhang, T., Wu, Y., & Jiang, J. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell*, *24*(7), 2719-2731. https://doi.org/10.1105/tpc.112.098061

Zhang, Y., Chen, M., Siemiatkowska, B., Toleco, M. R., Jing, Y., Strotmann, V., Zhang, J., Stahl, Y., & Fernie, A. R. (2020). A Highly Efficient Agrobacterium-Mediated Method for Transient Gene Expression and Functional Studies in Multiple Plant Species. *Plant Commun*, *1*(5), 100028. https://doi.org/10.1016/j.xplc.2020.100028

Zhang, Y., Li, Z., Bian, S., Zhao, H., Feng, D., Chen, Y., Hou, Y., Liu, Q., & Hao, B. (2020). HiCoP, a simple and robust method for detecting interactions of regulatory regions. *Epigenetics Chromatin*, *13*(1), 27. https://doi.org/10.1186/s13072-020-00348-6

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Z. H., Palta, J. A., Lu, P., Ren, M. J., Zhu, X. T., & He, J. (2022). Traditional soybean (Glycine max) breeding increases seed yield but reduces yield stability under non-phosphorus supply. *Functional Plant Biology*, *49*(2), 132-144. https://doi.org/10.1071/FP21116

Zhao, J., Bayer, P. E., Ruperao, P., Saxena, R. K., Khan, A. W., Golicz, A. A., Nguyen, H. T., Batley, J., Edwards, D., & Varshney, R. K. (2020). Trait associations in the pangenome of pigeon pea (Cajanus cajan). *Plant Biotechnology Journal*, *18*(9), 1946-1954. https://doi.org/10.1111/pbi.13354

Zhao, K., Xue, H., Li, G., Chitikineni, A., Fan, Y., Cao, Z., Dong, X., Lu, H., Zhao, K., Zhang, L., Qiu, D., Ren, R., Gong, F., Li, Z., Ma, X., Wan, S., Varshney, R. K., Wei, C., & Yin, D. (2025). Pangenome analysis reveals structural variation associated with seed size and weight traits in peanut. *Nature Genetics*. https://doi.org/10.1038/s41588-025-02170-w

Zhao, L., Yang, Y. M., Chen, J. C., Lin, X. L., Zhang, H., Wang, H., Wang, H. Z., Bie, X. M., Jiang, J. F., Feng, X. Q., Fu, X. D., Zhang, X. S., Du, Z., & Xiao, J. (2023). Dynamic chromatin regulatory

programs during embryogenesis of hexaploid wheat. *Genome Biology*, *24*(1). https://doi.org/10.1186/s13059-022-02844-2

Zhao, Y., Huang, Z., Zhou, X., Teng, W., Liu, Z., Wang, W., Tang, S., Liu, Y., Liu, J., Wang, W., Chai, L., Zhang, N., Guo, W., Liu, J., Ni, Z., Sun, Q., Wang, Y., & Zong, Y. (2025). Precise deletion, replacement and inversion of large DNA fragments in plants using dual prime editing. *Nat Plants*, *11*(2), 191-205. https://doi.org/10.1038/s41477-024-01898-3

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, L., Li, Y., Hussain, N., Li, Z., Wu, D., & Jiang, L. (2016). Allelic Variation of BnaC.TT2.a and Its Association with Seed Coat Color and Fatty Acids in Rapeseed (Brassica napus L.). *PLoS One*, *11*(1), e0146661. https://doi.org/10.1371/journal.pone.0146661

Zhu, B., Zhang, W., Zhang, T., Liu, B., & Jiang, J. (2015). Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures. *Plant Cell*, *27*(9), 2415-2426. https://doi.org/10.1105/tpc.15.00537

Zhu, J., Liu, M., Liu, X., & Dong, Z. (2018). RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis. *Nat Plants*, *4*(12), 1112-1123. https://doi.org/10.1038/s41477-018-0280-0

Zhu, M., Liu, M., & Dong, Z. (2025). Monitoring transcription by nascent RNA sequencing in crop plants. *New Crops*, *2*, 100031. https://doi.org/https://doi.org/10.1016/j.ncrops.2024.100031

Zhu, T., Liao, K., Zhou, R., Xia, C., & Xie, W. (2020). ATAC-seq with unique molecular identifiers improves quantification and footprinting. *Commun Biol*, *3*(1), 675. https://doi.org/10.1038/s42003-020-01403-4

Zhu, T., Xia, C., Yu, R., Zhou, X., Xu, X., Wang, L., Zong, Z., Yang, J., Liu, Y., Ming, L., You, Y., Chen, D., & Xie, W. (2024). Comprehensive mapping and modelling of the rice regulome landscape unveils the regulatory architecture underlying complex traits. *Nat Commun*, *15*(1), 6562. https://doi.org/10.1038/s41467-024-50787-y

Zhu, W., Yang, L., Wu, D., Meng, Q., Deng, X., Huang, G., Zhang, J., Chen, X., Ferrandiz, C., Liang, W., Dreni, L., & Zhang, D. (2022). Rice SEPALLATA genes OsMADS5 and OsMADS34 cooperate to limit inflorescence branching by repressing the TERMINAL FLOWER1-like gene RCN4. *New Phytologist*, *233*(4), 1682-1700. https://doi.org/10.1111/nph.17855

Zong, Y., Wang, Y., Li, C., Zhang, R., Chen, K., Ran, Y., Qiu, J. L., Wang, D., & Gao, C. (2017). Precise base editing in rice, wheat and maize with a Cas9-cytidine deaminase fusion. *Nature Biotechnology*, *35*(5), 438-440. https://doi.org/10.1038/nbt.3811

7 – Appendices

Appendix A

Population genomics uncovers loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*)

Maximillian R. W. Jones, Worku Kebede, Abel Teshome, Aiswarya Girija, Adanech Teshome, Dejene Girma, James K. M. Brown, Jesus Quiroz-Chavez, Chris S. Jones, Brande B. H. Wulff, Kebebew Assefa, Zerihun Tadele, Luis A. J. Mur, Solomon Chanyalew, Cristobal Uauy and Oluwaseyi Shorinola, 2025. *Communications Biology*.
https://doi.org/10.1038/s42003-025-08206-5

communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08206-5

Population genomics uncovers loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*)

Maximillian R. W. Jones[1,11], Worku Kebede[2,3,11], Abel Teshome[1,4,11], Aiswarya Girija[5,6], Adanech Teshome[2], Dejene Girma[2], James K. M. Brown[1], Jesus Quiroz-Chavez[1], Chris S. Jones[7], Brande B. H. Wulff[8], Kebebew Assefa[2], Zerihun Tadele[9], Luis A. J. Mur[6], Solomon Chanyalew[2] ✉, Cristobal Uauy[1] ✉ & Oluwaseyi Shorinola[4,10] ✉

Tef (*Eragrostis tef*) is an indigenous African cereal that is gaining global attention as a gluten-free "superfood" with high protein, mineral, and fibre contents. However, tef yields are limited by lodging and by losses during harvest owing to its small grain size (150× lighter than wheat). Breeders must also consider a strong cultural preference for white-grained over brown-grained varieties. Tef is relatively understudied with limited "omics" resources. Here, we resequence 220 tef accessions from an Ethiopian diversity collection and also perform multi-locational phenotyping for 25 agronomic and grain traits. Grain metabolome profiling reveals differential accumulation of fatty acids and flavonoids between white and brown grains. *k*-mer and SNP-based genome-wide association uncover important marker-trait associations, including a significant 70 kb peak for panicle morphology containing the tef orthologue of rice *qSH1* —a transcription factor regulating inflorescence morphology in cereals. We also observe a previously unknown relationship between grain size, colour, and fatty acids. These traits are highly associated with retrotransposon insertions in homoeologues of *TRANSPARENT TESTA 2*, a known regulator of grain colour. Our study provides valuable resources for tef research and breeding, facilitating the development of improved cultivars with desirable agronomic and nutritional properties.

Tef (*Eragrostis tef* (*Zucc.*) Trotter) is a cereal crop that has been grown in the Horn of Africa for millennia. It is a self-pollinating allotetraploid grass that is valued by farmers as a 'fail-safe' cash crop, resilient to marginal soils, waterlogging, high temperatures, and drought. Tef is a staple crop in Ethiopia, Africa's second most populous country, where it is grown on 3 million hectares (27% of cereal acreage) by around 6.7 million households, with annual production exceeding 5.5 million tonnes[1]. The crop acts as both feed and food, with the straw a prized forage for cattle and the whole-grain flour used to produce a fermented flatbread known as injera, which serves as a staple food for the majority of the country[2].

Tef has also gained global attention as a 'superfood' thanks to its high protein, calcium, iron, and fibre contents, its low glycaemic index, and its lack of allergenic gluten[2]. Additionally, tef is rich in dietary antioxidants, including polyphenols and flavonoids, and essential polyunsaturated fatty acids like linoleic acid, which are not synthesised by the human body[3]. These nutritional features, combined with its climatic resilience, make tef an attractive crop for wider adoption. To date, government policies in Ethiopia have restricted export of tef germplasm and bulk grain to protect both its natural heritage and domestic consumption[4,5]. However, tef cultivation is expanding beyond Ethiopia, notably in the USA, Australia, South Africa,

[1]John Innes Centre, Norwich Research Park, Norwich, UK. [2]Ethiopian Institute of Agricultural Research (EIAR), Addis Ababa, Ethiopia. [3]Institute of Plant Sciences, Scuola Superiore Sant'Anna, Pisa, Italy. [4]International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia. [5]Institute of Biological, Environmental & Rural Sciences (IBERS), Plas Gogerddan, Aberystwyth University, Ceredigion, UK. [6]Department of Life Sciences, Penglais Campus, Aberystwyth University, Aberystwyth, UK. [7]International Livestock Research Institute, Nairobi, Kenya. [8]Plant Science Program, Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [9]University of Bern, Institute of Plant Sciences, Bern, Switzerland. [10]School of Biosciences, University of Birmingham, Birmingham, UK. [11]These authors contributed equally: Maximillian R. W. Jones, Worku Kebede, Abel Teshome. ✉e-mail: solchk2@gmail.com; cristobal.uauy@jic.ac.uk; o.shorinola@bham.ac.uk

Fig. 1 | Diversity of panicle morphology and grain colour in tef. a Comparison of a bread wheat spike (cv. 'Paragon', far left) with tef accessions exemplifying four categories of panicle morphology (from left to right; very lax, lax, semi-compact, and compact). b Comparison of bread wheat grains (cv. Paragon, bottom) with grains from brown and white-grained tef varieties.

and the Mediterranean regions[6,7]. In these areas, tef is also used as a multi-harvest forage crop for producing premium-quality hay and silage[8].

Tef is considered an underutilised crop because it has, so far, not benefited greatly from modern genomics-based approaches to breeding and research. However, as in other underutilised crops such as grass pea, yam, and lablab, this *status quo* is beginning to shift[9,10], with the generation of a high-quality reference genome[11] following on from a draft sequence[12]. Notable progress has been made in breeding improved varieties of tef[2,13], although advances have not been on the same scale as for major cereals like wheat or rice. Lodging under high nitrogenous fertiliser regimes is a major limiting factor but has so far been difficult to address through classical semi-dwarfing approaches, at least partially because of the value of tef straw as animal feed. Addressing lodging through improved root traits is also being explored[4,15]. Panicle (inflorescence) morphology has been reported as a determinant of lodging tolerance in tef. The species exhibits dramatic panicle diversity[16], from open, highly lax panicles similar to wild *Eragrostis* species, to short-branched, compact panicles more akin to the spikes of Triticeae species (Fig. 1a). Using a combination of controlled-environment phenotyping, mechanical testing, and crop modelling, Blosch et al. showed that tef varieties with compact panicles tended to be more resistant to lodging, suggesting an ideotype approach could be used to address this issue[17].

Another consideration for breeders is grain size. Tef produces the smallest grains of any cultivated cereal, ranging from 1.0 to 1.7 mm in length, with a typical thousand grain weight (TGW) of 0.2–0.4 g, roughly 150-fold lower than that of wheat[2,18,19] (Fig. 1b). Indeed, the name tef is thought to derive from the Amharic word "teffa" meaning "lost"[20]; likely an allusion to the high levels of harvest and post-harvest losses (16–30%) experienced by tef farmers[6,21]. Breeding for larger grains could alleviate these losses and boost realised yields by improving the separation of grain and chaff during winnowing. Tef's small grain size also makes it difficult to evenly broadcast the recommended >10 kg/ha during sowing. Farmers instead use high sowing rates (up to 30 kg/ha) that ultimately produce overcrowded fields prone to lodging[22]. Lastly, breeders must also address a strong cultural preference for white-grained over brown-grained varieties, which translates into a higher market price for the former[23].

Here, we aim to use a population genomics approach to study the diversity and genetic architecture of agronomic, grain morphology, and grain metabolite traits in a representative Ethiopian tef collection. We therefore conducted short-read resequencing of 220 tef accessions from the Ethiopian Institute of Agricultural Research (EIAR) tef diversity collection. We characterised redundancy in this collection and produced a compact SNP panel that uniquely identifies the studied accessions. We combined this sequencing data with extensive in-field phenotyping across three trial locations, including precise grain morphology measurements and grain metabolome profiling (Fig. 2). This led to the identification of important marker-trait associations for panicle morphology, grain size, grain colour, and multiple grain metabolites. Our analyses establish a previously unknown link between grain size and grain colour, including the co-association of these traits with multiple genomic loci. However, we also identify regions that decouple these traits, offering potential breeding opportunities. Our work delivers a set of genomic and phenotypic resources for a diverse panel of tef accessions and lays the groundwork for future studies to define causal genes and variants underlying loci of agronomic relevance.

Results

SNP and *k*-mer-based methods identify redundancy in the EIAR core collection

Of the 225 accessions in the EIAR core collection, we sequenced (Illumina paired-end 150 bp) the genome of 220 accessions to an average depth of 8.85 Gbp (SD = 0.77 Gbp), equivalent to 14.2-fold the estimated genome size (622 Mb) of the reference genome cultivar 'Dabbi'[11]. The reads were mapped against the reference genome and variants were called. A quality-filtered and linkage-pruned set of 41,289 SNPs was prepared and used to investigate linkage disequilibrium (LD) decay and population structure. The genome-wide LD decay distance in the EIAR core collection was 46.3 kb (Supplementary Fig. 1). ADMIXTURE analysis revealed no clear population structure for two to 20 subpopulations[24]. Principal component analysis also did not suggest any distinct lineages, although there was a partial separation of brown and white-grained accessions by PC1 (Supplementary Fig. 1). This result was not unexpected, as a lack of strong population structure has previously been reported for other tef panels[25].

An SNP-based phylogram was computed using IQ-TREE 2. This revealed many groups of highly related accessions, with internal branch lengths close to zero nucleotide substitutions per site (Fig. 3a)[26]. A list of redundancy groups was defined such that the total branch length (phylogenetic distance) between any pair of accessions in a group was <0.005 substitutions per base pair. This resulted in 31 redundancy groups containing 2–19 accessions per group (Supplementary Table 1). Two accessions were excluded from placement in redundancy groups because they had high apparent heterozygosities (22.7% and 24.9% heterozygous sites, versus an average of 1.5% (standard deviation (SD) = 0.7%) for the other accessions) and likely represent seed mixtures rather than pure accessions.

To validate the redundancy groups defined above, a comprehensive *k*-mer (k = 51) presence/absence matrix was generated from the sequencing reads of the 220 sequenced accessions[27]. We tested whether pairs of accessions from the previously defined groups tended to have more *k*-mer states in common than non-group pairs (Fig. 3b, c, Supplementary Data 1). We observed that all 264 intra-group pairs had a *k*-mer state identity rate above 96.0%, whereas 23,825 out of the 23,826 other pairs had a *k*-mer state identity rate below this threshold, with a mean and median of 85.4% (SD = 0.02%). The single pair exceeding this threshold (DZ-01-91 and DZ-01-101), still had a very low phylogenetic distance (0.007) compared to the overall distribution, so were added as an additional redundancy group. The broad agreement between these two relatedness metrics suggested that the redundancy groups would be better treated as single accession pools rather than distinct entities. This reduced our effective number of accessions from 220 to 150.

One accession, DZ-01-1167, produced notably low shared *k*-mer state rates when paired against all other accessions (Fig. 3b, below the dashed line). DZ-01-1167 contains 294 million distinct *k*-mers versus an average of

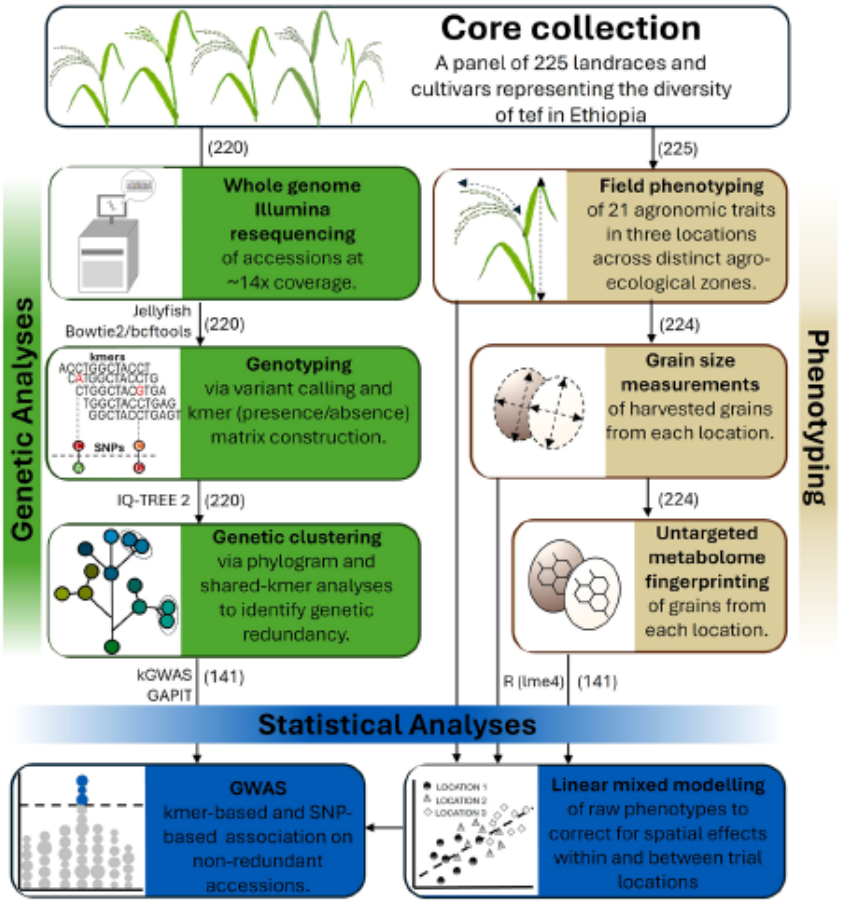


Fig. 2 | Resequencing, phenotyping, and GWAS of the EIAR core tef collection. A representative panel of Ethiopian tef accessions was resequenced and phenotyped for agronomic, grain size, and metabolomic traits. Statistical modelling was used to correct for location and within-site spatial effects. The software and numbers of accessions used for each step are indicated above each box. Vector images of the tef plant and sequencer were created by Wanda Pelin Canila/Shutterstock.com and Jaitham/Shutterstock.com, respectively.

160 million (SD = 3.5 million) for all other accessions (Supplementary Fig. 3), suggesting it contains many unique *k*-mers. This new diversity could derive from genetically distant tef accessions not otherwise captured in the panel or from interspecific introgression(s). Its source and utility could be further explored in future studies. This finding highlights the benefits of *k*-mer-based approaches, as this introgression is not apparent when comparing SNP-based phylogenetic distances.

Given the high levels of redundancy observed in the EIAR core collection, we selected a minimal panel of SNPs capable of distinguishing all 150 accession groups and singlets. This would allow other accessions belonging to the redundancy groups to be identified amongst the wider EIAR collection, as well as potentially facilitating some reconciliation with other tef collections. We identified a panel of 14 biallelic SNPs that could distinguish all 150 accession groups or singlets. To account for potential marker failures, we selected an additional 14 SNPs, making a total of 28. For each of these SNPs, all accession groups and singlets were homozygous for one allele or had missing data (Supplementary Data 2, Supplementary Note 1). All chromosomes were represented by at least one SNP except 5A, 7B, 8B, and 10A.

Grain colour strongly correlates with plant height and grain morphometric traits

To capture phenotypic variation in the EIAR core collection, we phenotyped 17 phenological and morphological traits at three field sites representing distinct agro-ecological zones (Supplementary Figs. 4, 5). Using high-resolution grain imaging, we also captured variation in eight grain size parameters relating to grain width, length and area. There was a significant effect (ANOVA p-value < 0.04) of location on all traits except for grain width and length. Trait coefficients of variation at each location ranged from 0.02 to 0.45, with phenology traits showing the least variation.

To account for spatial variability within and between experimental locations, we used linear mixed modelling to generate genotypic best linear unbiased predictors (BLUPs) and broad-sense heritabilities (H^2) for each trait. The heritability for agronomic and grain morphometric traits ranged from 0.14 to 0.97 (Supplementary Table 2). As expected, qualitative traits like panicle form and grain colour showed the highest heritability, 0.92 and 0.97, respectively. The heritability for grain morphometric traits, including grain length, width and area, was also high: 0.87, 0.90, and 0.89, respectively. Correlation analysis of trait BLUPs revealed expected associations between components of grain size (area, width, length) and weight, as well as between components of plant height (panicle length, plant height) (Fig. 4a).

Intriguingly, we also identified strong correlations between grain colour and both grain size and plant height. Plants of white-grained varieties were significantly taller than those of brown-grained varieties (Student's *t*-test, $p < 1 \times 10^{-5}$, Fig. 4b). This has positive implications for straw yields, but may increase lodging susceptibility. Brown-grained accessions are traditionally cultivated on more marginal soils, such as poorly drained vertisols, and, perhaps as a result, have come to be associated with smaller grains. However, our results indicated that brown-grained varieties tended to produce larger seeds than white-grained varieties when grown in common environments (Student's *t*-test, $p < 1 \times 10^{-15}$, Fig. 4c). Despite this, there was no difference in TGW between white and brown-grained varieties (Student's *t*-test, $p = 0.22$), suggesting that, on average, white-grained varieties produce grains with higher densities.

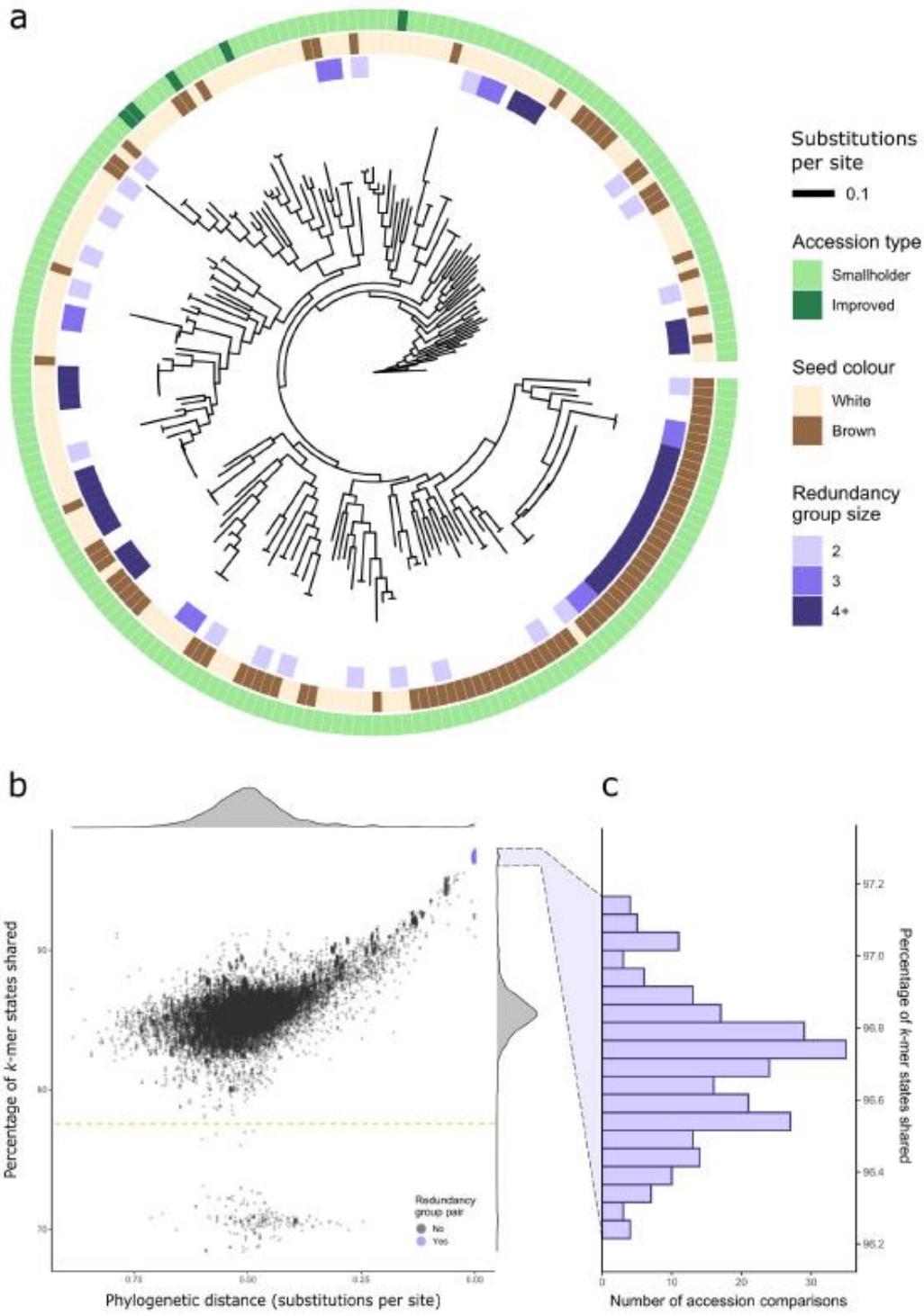


Fig. 3 | Phylogenetic analyses identify redundancy in the EIAR core collection. **a** Phylogram of 220 tef accessions, arbitrarily rooted against the accession "Ada-T58". White and brown-grained varieties are well-distributed across the phylogeny. A total of 32 redundancy groups, ranging in size from 2 to 19 accessions, were identified on the basis of small phylogenetic distances between pairs of accessions. **b** Phylogenetic distance plotted against the percentage of *k*-mer states shared for all 24,090 pairwise comparisons between accessions. There is a strong correlation between these two relatedness metrics. Accession pairs from within the previously defined redundancy groups (purple) cluster together at uniquely high shared *k*-mer state rates, depicted in detail in (c). The points with particularly low percentages of shared *k*-mer states (<78%, dashed line) represent the full set of comparisons of "DZ-01-1167" with other accessions.

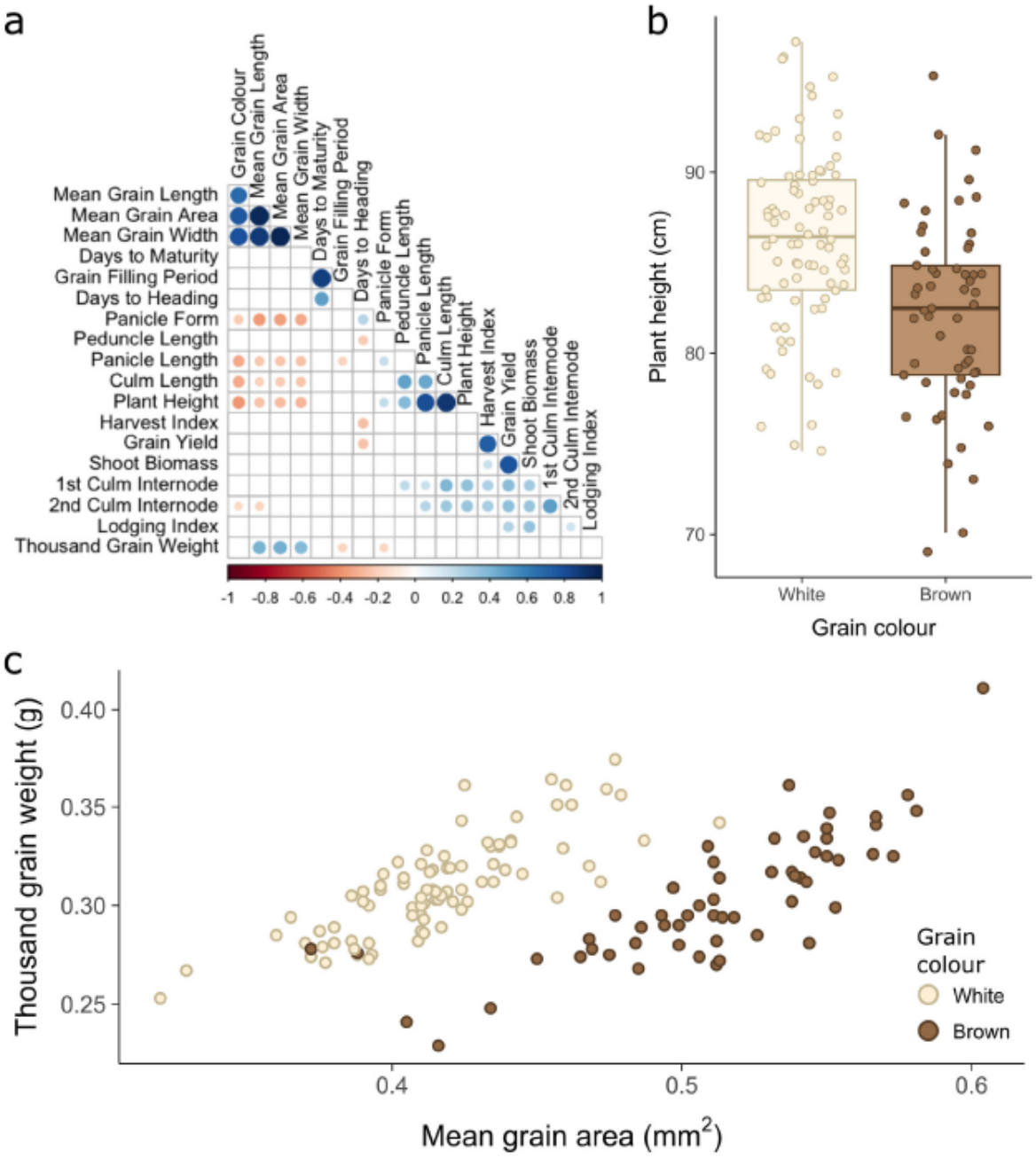


Fig. 4 | Best linear unbiased predictors (BLUPs) reveal correlations between key agronomic traits. a–c Analysis of BLUPs for $n = 141$ accessions and redundancy groups. **a** Correlation tests were conducted between the BLUPs for 20 traits of interest. Significant correlations ($p < 0.05$) are indicated by circles whose size and colour represent the magnitude and direction of correlation. **b** Boxplot of plant height BLUPs for white-grained ($n = 84$) and brown-grained ($n = 57$) varieties, with individual data points overlaid. White-grained accessions tended to produce taller plants. The centre line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles, and the whisker extends to 1.5* Interquantile range (IQR). **c** Scatterplot of grain area BLUPs against thousand grain weight (TGW) BLUPs. A distinctly bimodal distribution is strongly explained by grain colour, with white-grained varieties tending to produce smaller grains. Despite this, their grains are of approximately the same mass as brown-grained varieties, suggesting higher grain densities.

Lodging index was, as expected, positively correlated with above-ground biomass, highlighting the trade-off faced by tef breeders between lodging rates and straw yields. However, we did not observe the previously reported correlation between lodging index and panicle morphology. This could be due to the relatively few lines (five) with the highest level of panicle compactness or the lower robustness of our lodging index BLUPs, given that this trait was only phenotyped at two of the three field sites.

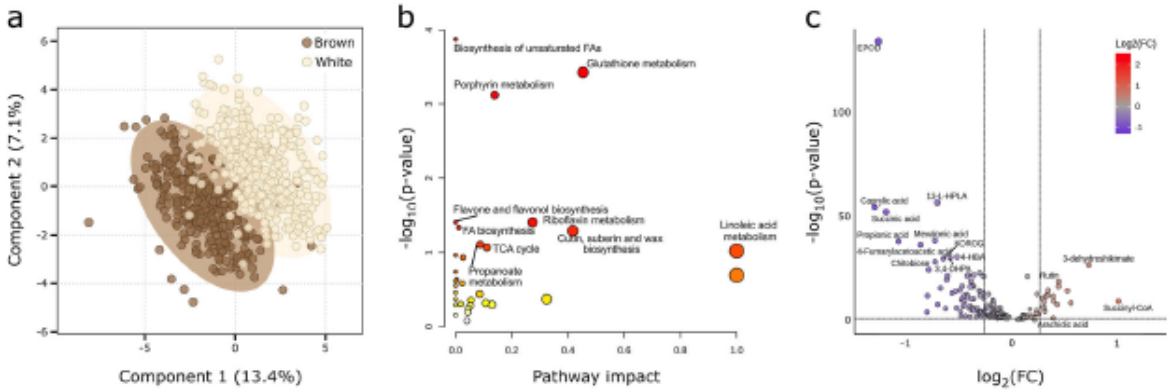


Fig. 5 | Brown and white-grained tef accessions display differential metabolite accumulation. **a** Partial least squares discriminant analysis (PLS-DA) of metabolites in grain samples of brown and white-grained accessions. Ellipses represent 95% confidence intervals around each group. **b** Differentially accumulated metabolites show enrichment for several metabolic pathways, notably fatty acid and flavone metabolism. Point size scales with pathway impact and colour intensity scales with significance of pathway enrichment. **c** Volcano plot for the 183 identifiable differentially accumulated metabolites. Fold-change (FC) was calculated as mean value in brown-grained varieties divided by that in white-grained varieties. Plotted FC thresholds are $\log_2(0.83)$ and $\log_2(1.2)$ and plotted FDR threshold is $\log_{10}(0.05)$.

High-resolution metabolite fingerprinting shows differential metabolite accumulation in brown and white tef grains

The cultural preference for white-grained varieties in Ethiopia and Eritrea, as well as the growing international interest in tef's nutritional properties, motivated us to also explore variation in tef's grain metabolomes. We performed untargeted metabolite profiling using Flow Infusion Electrospray High-Resolution Mass Spectrometry (FIE-HRMS) analysis on grain samples from each plot of the three trial locations. A total of 1643 positively ionised mass-to-charge ratio (*m/z*) features and 1470 negatively ionised features were captured, of which 209 and 723, respectively, were differentially accumulated in brown and white-grained varieties (Student's *t*-test, FDR < 0.05).

From these differential *m/z* features, 183 could be tentatively identified using a rice (*Oryza sativa* ssp. *japonica*) reference metabolome library available in the KEGG public database[28] (Supplementary Data 3). These differentially accumulated metabolites (DAMs) produced a clear separation of white and brown-grain samples when assessed by partial least squares discriminant analysis (PLS-DA) (Fig. 5a), but did not show differential accumulation between locations, suggesting little effect of locations on these metabolites (Supplementary Fig. 6). The DAMs were enriched for various processes, including (unsaturated) fatty acid biosynthesis, linoleic acid metabolism, glutathione metabolism, porphyrin metabolism, flavone, and flavonol biosynthesis, and riboflavin metabolism (Fig. 5b).

In agreement with other studies[3], our results show that brown-grained varieties tend to have higher proportions of essential polyunsaturated omega-6 and omega-3 fatty acids (e.g., linoleic acid and alpha-linolenic acid, respectively) while white-grained varieties have higher levels of saturated fatty acids (e.g., caprylic acid and 9,10-epoxyoctadecanoic acid (EPOD)). Omega fatty acids have been associated with lowering cardiovascular disease, cancer, and autoimmune diseases[29]. However, the high levels of unsaturated fatty acids in brown-grained tef may also contribute to its increased proneness to rancidity and therefore its lower consumer appeal[30] (Fig. 5c, Supplementary Fig. 7a).

We also found differential accumulation of flavonoids between white and brown-grained varieties. These compounds can affect flavour and colour and act as antioxidants[31]. We found increased levels of the flavonoids rutin and 3-dehydroshikimate in brown-grained varieties. Meanwhile, in white-grained varieties, we observed elevated levels of apigenin and kaempferol 3-O-rhamnoside-7-O-glucoside (KOROG). Flavonols such as the latter have been known to contribute to white pigmentation[32] (Fig. 5c, Supplementary Fig. 7b).

k-mer-based GWAS identifies regions associated with panicle and grain morphologies

To identify genomic regions associated with the agronomic and grain morphology traits, we conducted a *k*-mer-based genome-wide association study (kGWAS) using the previously calculated BLUPs and a new *k*-mer matrix to account for the reduced number of non-redundant accessions. Of the agronomic traits tested, we detected significant marker-trait associations (MTAs) for panicle morphology, grain morphology, and grain colour (Supplementary Table 3). We also carried out an SNP-based GWAS and identified four significant MTAs for panicle morphology, grain morphology, and lodging index (Supplementary Table 4).

The control of panicle morphology appeared to be relatively simple, with a single highly associated 70-kb region on chromosome 3B (Fig. 6b). The underlying reference *k*-mers negatively correlated with panicle morphology (scored as 1 to 4 for very lax to compact). This matched our expectations as the reference cultivar Dabbi produces very lax panicles. The significant region contains 13 gene models, including the tef orthologue (Et_3B_031395) of the rice gene *qSH1/RIL1* (*QTL for Seed Shattering on chromosome 1/RI-LIKE1*; Os01g0848400). *qSH1* is a *BEL1*-like homeobox transcription factor linked with seed shattering and inflorescence architecture[33,34]. *qSH1* orthologues are also expressed in the inflorescence meristems of maize and wheat, suggesting a conserved role in inflorescence development across the grass family [35,36].

There was a set of complex co-localised associations for grain morphology (Fig. 6a). Most strikingly, a highly significant region (peak 7) supported by tens of thousands of *k*-mers was detected on chromosome 4B for grain colour and grain width (Fig. 7a, b). This region was, as expected, also associated with grain area, but was not significant for grain length (Fig. 7b, Supplementary Fig. 8a, b). There was also a region on chromosome 4A (peak 3) significantly associated with grain colour and width (Fig. 7a, b). In addition, we found smaller regions on chromosome 3B (peak 2) and 4A (peak 6) co-associated with grain colour and grain width (Figs. 6a, 7a, b). The *k*-mers in each of the above regions were correlated with brown grain colour and increased grain size parameters. This supports the correlation between grain colour and grain size observed in the plotted BLUPs (Fig. 4c).

In contrast to the regions discussed above, there were also cases where grain colour and grain size were decoupled. A third positive grain width peak on chromosome 4A (Fig. 7b peak 5) is well-separated from the upstream and downstream grain colour peaks, by 8460 kb and 1580 kb, respectively. On chromosome 4B there is a region negatively associated with grain width and area (Fig. 7b peak

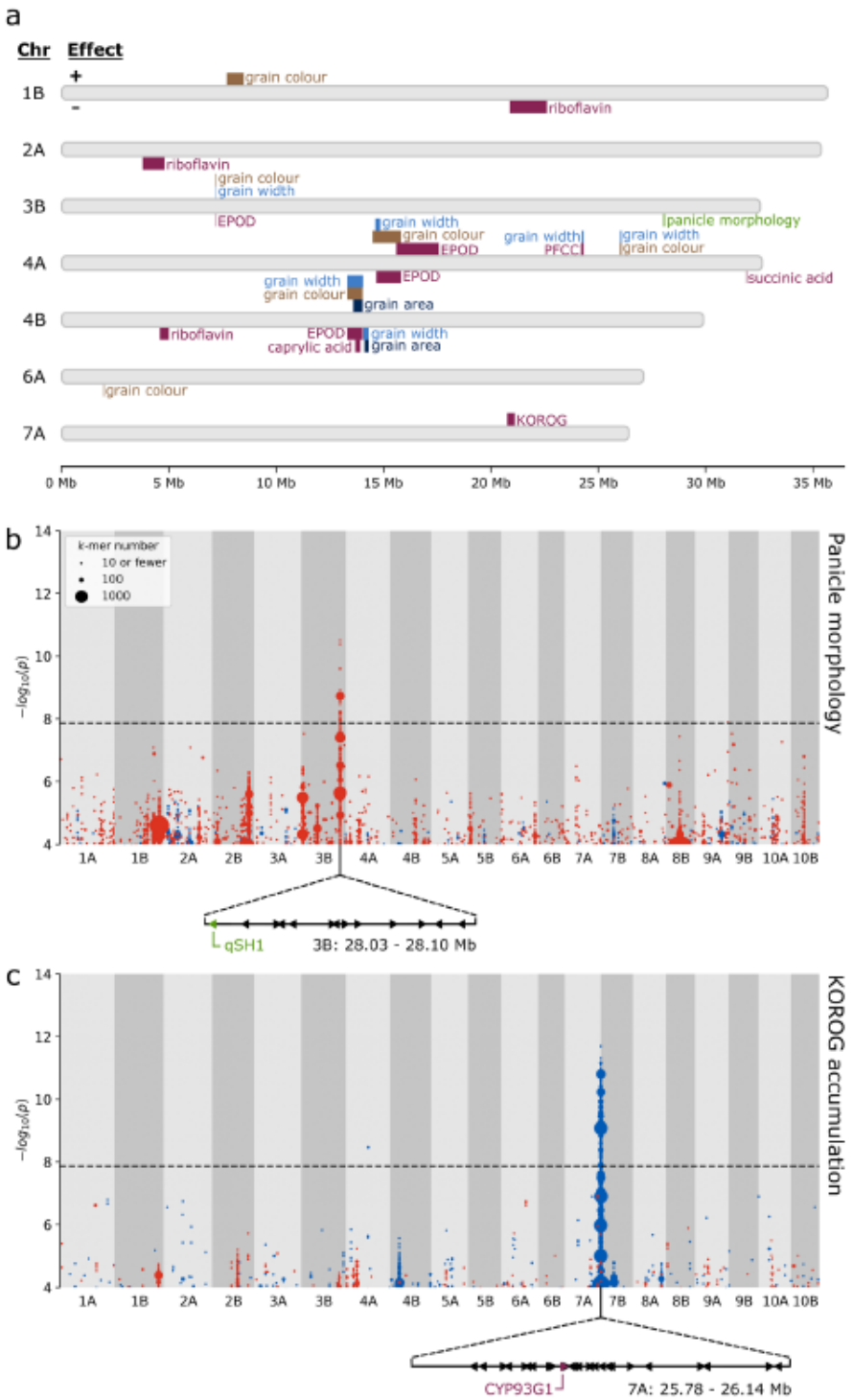
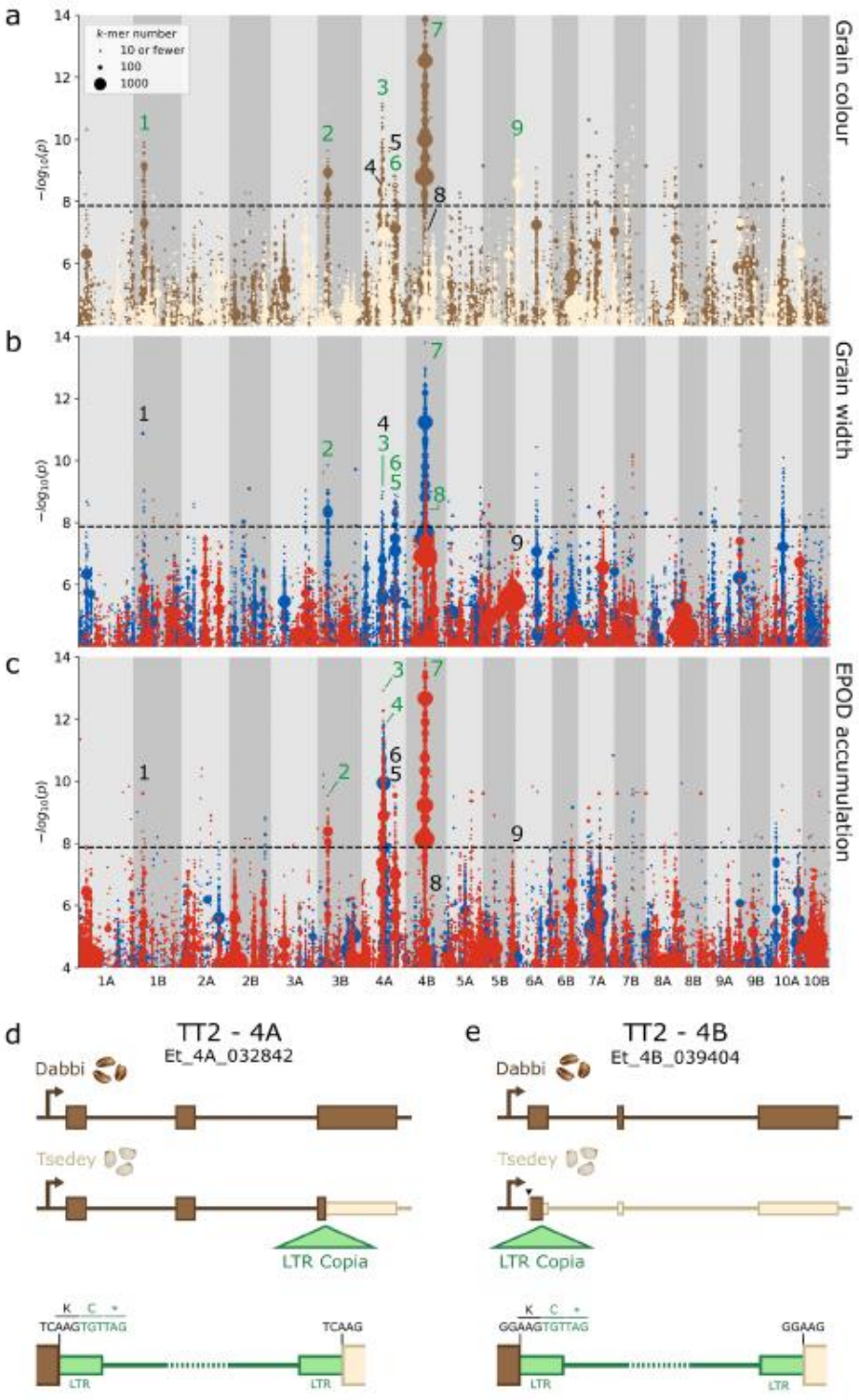


Fig. 6 | *k*-mer-based GWAS identifies multiple marker-trait associations, including regions associated with panicle morphology and grain KOROG. a Plot summarising all trait-associated regions identified by *k*-mer-based GWAS. Regions positively associated with traits are plotted above their respective chromosomes, while negatively associated regions are plotted below. For grain colour, positive and negative associations indicate brown and white, respectively. **b** A region significantly associated with panicle morphology was detected on chromosome 3B. The arrangement of the 13 genes within this region is displayed below the plot. The candidate gene *qSH1* is highlighted. **c** A region significantly associated with Kaempferol 3-O-rhamnoside-7-O-glucoside (KOROG) was detected on chromosome 7A. The arrangement of the 26 genes within this region is displayed below the plot. The candidate gene *CYP93G1* is highlighted. In (**b** and **c**), *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association; red for panicle laxness or low KOROG, blue for panicle compactness or high KOROG. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10.



8) that is separated by just 10 kb from peak 7, a major peak for brown grain colour and higher grain width and area. A marginally insignificant peak for grain width and area exists on chromosome 10A (Fig. 7b, Supplementary Fig. 8). Lastly, there are two grain colour peaks on chromosomes with no significant grain size peaks (Fig. 7a). This includes a 790 kb peak associated with brown grains on chromosome 1B (114 genes) and a 40 kb peak associated with white grains on chromosome 6A (6 genes). These two regions offer the strong possibility of breeding for grain colour independently of grain size.

Fig. 7 | Co-association of grain colour, width, and EPOD concentration with multiple regions. Plots of *k*-mers associated with a grain colour, b grain width, and c grain EPOD concentration. *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association. In (a), brown denotes association with brown grain colour and white with white grain colour. In (b and c), red denotes association with lower trait values, and blue with higher trait values. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10. Nine regions are labelled with black and green numbers, denoting whether the region is significant or not significant for the plotted trait, respectively. Diagrams of LTR Copia insertions into *TT2* homoeologs on d chromosome 4A, e chromosome 4B. Top: structure of *TT2* in Dabbi (brown-grained). Centre: structure of *TT2* in Tsedey (white-grained). Bottom: detail of LTR Copia insertions. Narrower exons indicate presumed protein truncations. Black DNA bases denote 5 bp target-site duplications. Green DNA bases denote the start of the retrotransposon insertions. Single-letter amino acid codes show the introduction of premature stop codons (*). The first 22 bp of the *TT2* open reading frame on 4B is not assembled in the Tsedey genome (greyed out, black arrowhead). Gene annotations derive from ref. 11. and do not include 5' and 3' untranslated regions.

Associated regions for grain metabolites and grain morphology co-localise

We hypothesised that genomic loci associated with grain colour might also be associated with the differentially accumulated metabolites. To facilitate GWAS analysis, we calculated BLUPs and broad-sense heritabilities for 21 grain metabolites with high fold-change differences and/or which are known to be involved in pigmentation or important for human nutrition (Supplementary Data 3). Heritability values were generally high, with 18 of the 21 metabolites having $H^2 > 0.50$ (Supplementary Table 5). kGWAS revealed significant regions for riboflavin, EPOD, primary fluorescent chlorophyll catabolite (PFCC), succinic acid, caprylic acid, and KOROG (Fig. 6a, Supplementary Table 3). SNP-based GWAS also identified nine MTA for four metabolite traits, five of which overlap with the significant region for kGWAS (Supplementary Table 4).

As we hypothesised, kGWAS associations for some metabolites, including EPOD, caprylic acid and PFCC, co-localise with regions associated with grain colour and/or size. Of these, EPOD showed the most consistent and significant overlap with grain colour and grain width, on chromosomes 3B (peak 2), 4A (peak 3) and 4B (peak 7, also overlapping a grain area peak) (Fig. 7c). In these co-localised regions, EPOD was negatively associated with grain colour (i.e brown-grained varieties had lower EPOD content). There was also a significant peak for EPOD on chromosome 4A (peak 4) that partially overlapped, and was positively correlated with, a grain colour peak (peak 3). The 4A peaks for EPOD (peaks 3 and 4) were also identified in our SNP-based GWAS. Caprylic acid was associated with a single locus that overlapped with peak 7 for EPOD, grain colour, and grain width. Similarly, PFCC was associated with a single region in both *k*-mer and SNP-based GWAS that also overlapped with grain width.

We also found associations for other metabolites, including KOROG, succinic acid, and riboflavin, that did not overlap with grain colour and/or size. KOROG was associated in both *k*-mer and SNP-based GWAS with a single prominent 360 kb region on chromosome 7A (Fig. 6c) containing 26 gene models. This included Et_7A_050580, which encodes a cytochrome P450 (*CYP93G1*) with flavanone 2-hydroxylase (F2H) activity and has been previously shown to be involved in the biosynthesis of flavonol glycosides like KOROG[37]. Succinic acid was also associated with a single region, in this case spanning 40 kb on chromosome 4A and containing eleven genes. Control of riboflavin accumulation appears more complex, with three regions on chromosomes 1B (1730 kb, 174 genes), 2A (1030 kb, 126 genes), and 4B (420 kb, 75 genes) associated with riboflavin accumulation but that are not associated with other traits.

TRANSPARENT TESTA 2 is a candidate for grain colour variation

Given that the most prominent associations for grain colour, size and metabolites content cluster at peak 3 (chr 4A: 14.48–15.79 Mb) and peak 7 (chr 4B: 13.34–14.05 Mb) (Figs. 6a, 7a, b, Supplementary Table 3), we examined the gene content in these peaks to identify potential candidate genes. Interestingly, these peaks displayed partial homology; genes in the proximal end of peak 3 are homoeologous to genes in the distal end of peak 7 (Supplementary Fig. 9). Homoeologous gene pairs in these regions include Et_4A_032844/Et_4B_037039 and Et_4A_032842/Et_4B_039404, whose orthologues have been previously shown to regulate grain colour, size and fatty acid content. The former are orthologues of *NUCLEAR FACTOR YA3* (*NF-YA3*), which regulates seed oil content and seed size across diverse angiosperms, including *Arabidopsis* and oil palm[38]. The latter are orthologues of *TRANSPARENT TESTA 2* (*TT2*), a MYB transcription factor that is associated with proanthocyanidin accumulation in the seeds of various Brassicaceae species[39,40].

We compared the gene sequences of the *TT2* orthologues from the Dabbi reference genome (a brown-grained variety) to those from the published draft assembly of the white-grained variety 'Tsedey'[12]. In the Tsedey assembly, we identified striking insertions of long-terminal repeat (LTR) Copia superfamily retrotransposons (RTs) in the third and first exons of the A and B *TT2* homoeologues, respectively (Fig. 7d, e, Supplementary Data 4 and 5). The A-subgenome RT introduces an in-frame cysteine and then a premature stop codon, truncating most of the final exon (162 codons). The B-subgenome RT is inserted in the first exon, truncating most of the protein. The positions of the two elements in different exons suggest independent insertions occurring after subgenome divergence. In both RTs, the 5 bp target-site duplications and 106 bp LTRs remain undegraded, suggesting a relatively recent insertion[41]. This is consistent with the large number of recently active LTR RT families previously identified in tef. We did not find any protein-truncating mutation between Dabbi and Tsedey in the A and B homoeologues of *NF-YA3;* the B homoeologue contains one non-deleterious missense mutation, while the A homoeologue contains no missense mutation.

Discussion

Developing genomic resources for underutilised crops is crucial for accelerating their improvement, adoption, and utilisation, and will in turn boost the resilience of the interconnected global food system[10,42]. Our extensive phenotyping and whole genome resequencing of a diverse tef collection represents a valuable resource for germplasm characterisation and trait mapping in this locally vital and globally emerging crop.

While grown in Ethiopia and Eritrea for thousands of years, systematic collection and breeding of tef did not begin until the 1950s, with sampling of varieties directly from farmers' fields. Since then, numerous germplasm collections have been established, containing over 7000 accessions. These are primarily maintained in Ethiopia, but smaller collections exist elsewhere[18,43]. Correspondence of varieties between these collections is undocumented, preventing cross-utilisation of phenotyping and sequencing data. Genomics approaches have been invaluable for resolving such issues and for identifying redundancy within collections[44,45]. Our work reveals redundancies in the EIAR tef collection, which are likely due to repeated sampling of farmer-traded germplasm across modest geographical ranges. The compact SNP panel we have developed can be used to identify further redundancy within the EIAR collection. The resequencing data presented here can also be combined with existing mid-density genotyping data from other tef collections to assess redundancies, differences, and complementarity between the different tef collections globally[25,46]. This will shed further light on tef's breeding history, facilitate germplasm exchange, and inform the selection of accessions for a tef pan-genome to optimally capture tef diversity.

We identified a strong candidate gene for panicle morphology, Et_3B_031395, which is orthologous to the rice *qSH1* gene. qSH1 is a BEL1-like homeodomain protein that underlies variation in seed shattering in rice[33] and is regulated by *SUPERNUMERARY BRACT*[47], whose direct orthologue in wheat is the major inflorescence morphology gene Q, which controls both seed shattering and inflorescence compactness[48]. The *Arabidopsis* orthologue of *qSH1*, *REPLUMLESS/PENNY*[49], also known as

PENNYWISE[50] or *BELLRINGER*[51], is important for fruit development and dehiscence. *qSH1* has also been directly connected with inflorescence architecture, with *qsh1 ri* (*verticillate rachis*) double mutants displaying abnormal timing and arrangement of primary branch meristems[33,34]. *qSH1* also strongly influences bract suppression in the inflorescence[33,34]. In addition, paralogs and orthologues of *qSH1* have been shown to control inflorescence patterning in maize and *Arabidopsis*[50–52]. Given the established roles of its orthologues in diverse plant species and its localisation within a narrow candidate region, Et_3B_031395 emerges as a promising candidate for the regulation of panicle morphology in tef. Nonetheless, functional validation will be necessary to confirm its role.

Most farmers in Ethiopia rely on manual broadcasting for sowing on small plots (0.25 to 1 hectare), as access to mechanisation is limited and such equipment is typically not optimised for the tiny seeds of tef[53,54]. This practice leads to inefficient seed use as farmers typically sow at higher rates than recommended to ensure good field coverage (15–25 kg/ha, instead of 5 kg/ha)[55]. These high seeding rates also produce overcrowded fields of weak-stemmed plants more prone to lodging[22]. Larger grains would facilitate mechanised handling and ensure seedlings have sufficient nutrients for establishment from greater soil depths. Together, this would promote row-based drilling of tef, alleviating the above issues and supporting the ongoing transformation of tef cultivation practices. Indeed, agronomists have already experimented with pelleting tef seeds with inert material to enable mechanised sowing, highlighting the promise of this approach[56]. Lastly, increasing grain size could help reduce the high grain loss rates experienced by smallholders during traditional threshing and winnowing processes[21].

Our kGWAS results offer potential breeding targets for increasing grain size. We identified six genomic regions significantly associated with grain width, two of which were also associated with grain area. However, we also identified a previously unknown link between grain size and grain colour, which could complicate this process. Brown-grained varieties tended to produce larger grains (0.51 mm^2) than white-grained varieties (0.42 mm^2), and this was reflected in the kGWAS; four of the regions positively associated with grain width were also associated with brown grain colour. This co-localisation presents an issue because there is a strong cultural preference in Ethiopia for white tef flour, and this translates into a market incentive for farmers to grow white-grained varieties. Introgression of grain size alleles into elite white-grained varieties at the cost of increased pigmentation would therefore not be favourable.

However, not all grain size and grain colour loci were co-localised. We identified one grain width locus on chromosome 4A that is very distant from the two grain colour regions on this chromosome, plus two grain colour loci on chromosomes which do not harbour grain size loci (1B and 6A). We also observed a region on chromosome 7A strongly associated with the metabolite KOROG. While our analysis did not find this region to be co-associated with grain colour, flavonols such as KOROG have previously been associated with white pigmentation[32]. These regions offer opportunities to combine favourable grain size and colour alleles through introgression, though it is yet to be seen if the introduction of 'white grain' alleles into a brown-grained background would yield a dominant effect. It is more likely that positive breeding outcomes could be achieved by stacking multiple additive grain size loci. Uncovering further such loci should be a priority for future GWAS studies in tef. It is also important to note that the use of a high-throughput and high-accuracy phenotyping platform (MARViN grain analyser) was key to uncovering these grain size variations. While routine for major crops such as wheat, this is the first application of high-resolution, image-based grain measurements to a tef panel to our knowledge. This exemplifies the benefits that adopting robust and well-tested phenotyping methodologies from mainstream crops can bring to the research of underutilised crops[10].

Another route to achieving large-grained white tef varieties could be to knock out key transcription factors or enzymes linked with pigmentation in a brown-grained background (through mutation breeding, transgenesis, or genome editing). To implement such an approach, the tef research community will need to increase its understanding of relevant genes and their pleiotropic effects. Contributing to this, we identified a candidate pair of homoeologues present within each of the two major loci for grain size, colour, and the fatty acid EPOD. Et_4A_032842 and Et_4B_039404 are orthologous to *TRANSPARENT TESTA 2* (*TT2*), an R2R3 MYB transcription factor known to regulate seed coat colour in the Brassicaceae family through proanthocyanidin formation and accumulation[39,40].

We propose that variations in the two tef orthologues of *TT2* contribute to variation in grain colour. This aligns with a previous report of two "duplicate" genetic factors *B/b and B2/b2* as the major determinants of brown/white pigmentation[57]. Our hypothesis is further supported by our discovery of independent insertions of related LTR-Copia retrotransposons in both homoeologues of the white-grained tef variety Tsedey. These insertions, which are absent in the genome of the brown-grained variety Dabbi, would both lead to truncated and likely non-functional proteins. These *TT2* polymorphisms could also underlie the variation in fatty acid content and grain size, as has been shown in *Arabidopsis* and *Brassica napus*[58,59].

This hypothesis could be tested by mutagenising *TT2* in a brown-grained variety. However, currently, we cannot preclude the association of these traits with alternative or additional candidate genes. The homoeologues Et_4A_032844 and Et_4B_037039 lie within the same two loci and are orthologous to *NF-YA3*, a gene which activates oil accumulation in oil palm mesocarp and increases oil content and seed size when overexpressed in *Arabidopsis*[38]. It is, therefore, possible that the correlated traits (grain colour, size and fatty acid) are modulated by two separate gene families in linkage blocks.

While the above manipulation is an intriguing possibility, we acknowledge that the original cultural preference for white-grained tef varieties is likely linked to flavour and baking properties in addition to aesthetics, although this has not been well-studied. It is therefore also important to study the tef metabolome beyond its direct contribution to colour. Future studies could utilise our extensive metabolome profiling data to conduct a more comprehensive metabolite GWAS (mGWAS) and deepen our understanding of the genetics underpinning differential metabolite accumulation in brown and white-grained tef.

Our work demonstrates the value that can be brought to underutilised crops such as tef by applying resequencing and population genomics in combination with large-scale phenotyping and metabolome profiling. We identify multiple genetic loci for morphological and nutritional traits and suggest how these could inform future research and breeding efforts, including contribution to new tef varieties with desirable plant architecture and consumer-preferred grain traits. Other underutilised crops could also greatly benefit from such methods to accelerate their domestication or improvement.

Methods

Germplasm

A core panel of 225 tef accessions was selected from the broader EIAR collection to capture a broad range of phenotypic diversity. This panel consisted of 220 smallholder varieties and 5 improved, registered varieties (Supplementary Data 6). The EIAR collection is originally derived from 2175 tef germplasm accessions[60], 35 cultivars[61], and 10 released improved varieties[62].

Field phenotyping

The accessions were grown in Ethiopia at three EIAR research sites: Alem Tena, Chefe Donsa, and Debre Zeit. Chefe Donsa and Debre Zeit sites represent non-stressed environments, while Alem Tena represents a moisture-stressed environment (Supplementary Table 6). Each field trial was set up using an augmented block design (Supplementary Data 7) and consisted of 1 m rows sown with 0.3 g grain and spaced 50 cm apart. Most accessions were sown once per field site, but five improved varieties (Ebba, Boset, Bora, Dagim and Felagot) were sown four times per field site as spatial controls.

Data were collected on a range of qualitative and quantitative traits (Supplementary Data 8). Phenotyping methodology was derived from the Tef Breeding Manual[43] and is described in detail in Supplementary Table 7. Qualitative traits included basal stalk colour, grain colour, panicle colour, and panicle form. Quantitative traits included phenology (days to heading, days to maturity, and grain filling period) and agro-morphology (plant height, panicle length, peduncle length, culm length, first culm internode length, second culm internode length, above-ground shoot biomass, grain yield, harvest index, and lodging index). Lodging index was only assessed at Alem Tena and Debre Zeit.

Grain size measurement

Grain samples from each of the 720 rows across the three field trials were analysed using a MARViN Grain Analyzer. For each sample, 0.075–0.085 g of grain (mean of 262 grains) was evenly distributed on the imaging tray. Mean grain area and TGW were recorded, as well as mean, minimum, and maximum values for grain width and length. Insufficient grain was harvested from accession Trotteriana-T-138 for MARViN analysis.

DNA extraction and resequencing

For each accession, ~0.7 g fresh leaf tissue was collected from 3-week-old plants from the Alem Tena field trial. DNA was extracted using DNeasy Plant or DNeasy Plant Pro kits (QIAGEN) and eluted in 50 µL AE buffer. Sufficient high-quality DNA could not be extracted for the accession Trotteriana-T-138. For three further accessions (DZ-01-170, DZ-01-1015, Gealamie-T-111), the observed grain colour at Alem Tena did not match that observed at the other two field locations, suggesting heterogeneity in the seed stock. Given that the DNA sample would, therefore, not represent the majority of the phenotyping data, these accessions were not sequenced and therefore not used for GWAS. The remaining 221 DNA samples were sequenced by Novogene UK (Illumina paired-end 150 bp), and data were returned for 220 accessions (no data were produced for DZ-01-12).

SNP calling, LD calculation, and phylogenetic analyses

Raw sequencing reads were trimmed using fastp[63] and mapped to the *Eragrostis tef* reference genome (cv. Dabbi) using Bowtie2 (v2.4.1)[11,64]. The mapped reads were filtered for MAPQ scores >30 using SAMtools (v1.18), and a VCF file was generated using BCFtools (v1.18)[65–67]. VCF statistics were generated using VCFtools and examined using base R (v4.1.3)[68,69]. Empirically derived filters were applied using VCFtools (--max-missing 0.90; --minQ 30; --minGQ 15; --min-meanDP 10; --max-meanDP 18; --minDP 5; --maxDP 23; --maf 0.025)[70] and linkage pruning was conducted using Plink (v1.90b4.6; --allow-extra-chr; --indep-pairwise 20 5 0.5). The final VCF file of 41,289 SNPs was converted to PHYLIP format using the tool vcf2phylip (v2.9)[71].

TASSEL (v5.2.54)[72] was used to compute pairwise intra-chromosome LD correlation coefficient (r^2) between SNP markers across the entire tef genome. LD decay scatterplot was then produced by plotting the r^2 values against physical distance (bp) using R software. The intersection point between the genome-wide LD curve and the r^2 threshold (0.2) determined the genome-wide LD decay value.

A phylogram was generated using IQ-Tree 2 and arbitrarily rooted against Ada-T-58 based on alphabetical order (v2.3.2; -B 10000; --msub nuclear; -m MFP + ASC; --seed 42)[26]. The phylogram was visualised in R using ggtree (v3.10.1) and ggtreeExtra (v1.12.0)[73,74]. Population structure was investigated by applying ADMIXTURE analysis (v1.3.0, K = 1:20; --cv = 10)[24] and principal component analysis (SNPRelate v1.29.0)[75] to the same VCF file. The results were visualised using Pophelper (v2.3.1) and Plotly (v4.10.1), respectively[76,77]. After defining the redundancy groups, the name of a single accession from each group was arbitrarily assigned to represent the group in subsequent analyses (Supplementary Table 1).

A *k*-mer presence/absence matrix was computed for all 220 sequenced accessions from trimmed reads using scripts from a previously published *k*-mer-GWAS pipeline (https://github.com/wheatgenetics/owwc/tree/master/kGWAS, section 1)[27]. Additional guidance on the use of this pipeline is provided at https://github.com/quirozcj/kmerGWAS_descriptions. Default parameters were used for all steps. Notably, the default *k*-mer length and minimum *k*-mer frequency per accession were not modified (-m 51 and -L 4, respectively). Shared *k*-mer state rates were computed using custom bash and R scripts (https://github.com/Uauy-Lab/tef_kGWAS_2024).

Minimal SNP panel selection

The trimmed reads from accessions belonging to redundancy groups were pooled and subsampled down to the average read number per single library (29,350,529 paired-end reads). A VCF file was generated as above for the 150 redundancy groups and singlets. The same filters and linkage pruning were applied. This VCF file was input to the Minimal Marker pipeline[78]. This involves conversion to a genotype matrix, then selection of SNPs. However, prior to SNP selection, the pipeline was modified to convert all heterozygous calls to missing data. Heterozygous loci are unstable between generations and would therefore be unreliable markers for consistently identifying accessions across generations. In contrast, the original target species for the pipeline, apple (*Malus domestica*), is largely propagated vegetatively, so heterozygous loci are stably inherited between generations. Missing calls are not used by the pipeline to distinguish accessions, so this change forced the pipeline to select a panel of SNPs that uniquely identifies the core collection using only loci homozygous across the 150 tef redundancy groups and singlets. The first run selected 14 SNPs fulfilling this remit. To provide redundancy, these SNPs were removed from the genotype matrix and a second run was conducted, leading to the selection of a further 14 SNPs.

An additional consideration was whether the genotypes of the individual members of the redundancy groups were consistent with the overall group genotype. This was investigated using custom Bash and R scripts (https://github.com/Uauy-Lab/tef_kGWAS_2024), and the results are summarised in Supplementary Data 2 and Supplementary Note 1. There were no cases where group members' genotypes compromised the utility of the SNP set for unique identification of the 150 non-redundant accessions.

Metabolite extraction and profiling

Methanolic metabolite extractions were conducted at the International Livestock Research Institute (ILRI, Ethiopia) on 40 mg grain from each trial plot following a previously published protocol[79]. Briefly, tissue was ground (Tissue Lyser (QIAGEN), 25 Hz, 2 min), added to 1 mL pre-cooled 100% methanol (−20 °C), and placed on ice for 30 min with vortexing every 5 min. The extracts were then centrifuged, vacuum concentrated, and shipped to Aberystwyth University, Wales, UK, for high-resolution metabolite profiling. The samples were resuspended in 300 µL of pre-cooled 100% methanol, vortexed for 5 min and centrifuged at 1000×*g* at 4 °C for 5 min. An aliquot of 200 µL of each sample was used for untargeted metabolite fingerprinting using Flow Infusion Electrospray High-resolution Mass Spectrometry (FIE-HRMS) mode using Q Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo-Scientific, UK) where data was captured in negative and positive ionisation mode using Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo-Scientific, UK). Quality controls (QC) were derived from a master mix sample where 10 mL of each extract was pooled and also "blanks" of 100% methanol. Three 20 µL injections were performed for each sample as technical replicates. FIE-HRMS metabolite fingerprints in both positive and negative ionisation modes in a single run. 20 µL of samples were injected into a flow of 100 mL min^{-1}. The acquisition of mass-to-charge ratio (*m/z*) data and their binning to discrete bins and peaks was conducted as previously described[80,81].

Statistics and reproducibility

Statistical analysis on annotated metabolites. Good-quality metabolite data could not be produced for 15 samples (Supplementary Data 8). A further set of ten samples (including the three Alem Tena discrepancies mentioned previously) was removed because their grain colour at one location did not match the grain colour at the other two locations

(Supplementary Data 8). *m/z* feature intensities from biologically independent samples (brown $n = 303$ and white $n = 389$) were log10 transformed and Pareto scaled, and those differing significantly between brown and white-grained varieties were selected (Student's *t*-test, FDR < 0.05, mass tolerance of 5 ppm). Metabolite identities were assigned to the differential *m/z* features using the Mummichog algorithm[82], with reference to the latest KEGG version of the *Oryza sativa japonica* (RefSeq) metabolite library[28,83]. A mass tolerance of 5 ppm was used, and all possible adducts and isotopes were considered. Where *m/z* values could be matched to multiple metabolites, the metabolite with the smallest mass difference to the *m/z* value was selected. Statistical analysis, principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), variable metabolite prediction, and volcano plot were carried out using the online R-based platform MetaboAnalyst 6.0 (https://www.metaboanalyst.ca)[84].

Statistical modelling for BLUP and heritability calculation. The original field trial design was updated to reflect the treatment of redundant accessions as combined redundancy groups (Supplementary Data 7). Individual plots belonging to the same redundancy group were treated as biological replicates. Data points were removed for the metabolite analyses above. Genotypic BLUPs were calculated using the R package lme4 (v1.1.32)[85] by fitting the following linear mixed model using restricted maximum likelihood (REML):

$$f(Y) = \alpha + \beta X + \gamma Z + \delta W + e$$

Where Y is the observed trait value, $f()$ is a transformation conducted for normalisation (either square root, natural log, or none), α is the global mean, β is location, X is a matrix of location effects, γ is block by location identity, Z is a matrix of block by location effects, δ is genotype identity, W is a matrix of genotype effects, and e is the residual. Location was modelled as a fixed effect due to the low number of factor levels, while block by location and genotype were modelled as random effects. The transformation $f()$ applied to each trait was selected to make residuals approximately normally distributed and independent of fitted values. Supplementary Tables 2 and 5 describe the transformation applied to each trait and list any additional data points removed for specific traits prior to BLUP calculation. For example, Alem Tena data was removed prior to the calculation of BLUPs for DTH, DTM, and GFP, as this data made the traits unsuitable for linear mixed modelling even after transformation.

Modelling of BLUPs was not deemed appropriate for panicle morphology and grain colour, given their ordinal and binary encoding (respectively) and minimal variation between field sites. Instead, simple means were used as the genotypic values. To rationalise trait values for presentation, the global intercepts were added to the BLUPs for plant height, grain area, and thousand grain weight in Fig. 4a, b. Raw BLUPs were used for kGWAS and SNP-based GWAS computations.

Broad-sense heritability (H^2) was calculated via the "Cullis" method[86] (quoted in the Results) and the BLUP-BLUE regression method (i.e., "Walsh and Lynch" method)[87]. Calculations were performed in GenStat[88] using the same transformations as for BLUP derivation. The selected methods are considered robust to unbalanced trial designs and produced similar results (mean difference 0.021, largest difference 0.076). Tef is highly selfing and we have demonstrated very low heterozygosity for this population (mean = 1.5%). Because of this, additive genetic variance (V_A) will predominate over dominance variance (V_D), so H^2 is expected to be approximately equal to (though slightly larger than) narrow-sense heritability (h^2).

k-mer-based GWAS. A new *k*-mer presence/absence matrix was generated as above using the previously pooled and subsampled reads for the 150 accessions or redundancy groups. Association mapping, calculation of significance threshold, and plotting were conducted using the same kGWAS pipeline on 141 accessions (nine redundancy groups were excluded because they contained both brown and white-grained accessions; Supplementary Table 1)[27]. For the metabolite traits, a further three accessions were excluded because only one datapoint remained for BLUP calculation (DZ-01-517, DZ-01-1376, Hotolla-T-135). *k*-mers were projected onto the Dabbi reference genome and reported as the number of *k*-mers at a given association level per 10 kb genomic bin. The significance threshold for associations was calculated via Bonferroni correction as follows:

$$padj = \frac{0.05}{n / k} = 1.36 \times 10^{-8}$$

Where n is the number of *k*-mers utilised for association calculations (187,226,135) and k is the *k*-mer length (51). This threshold is plotted as $-\log_{10}(1.36 \times 10^{-8}) = 7.87$ on all Manhattan plots presented. For each trait, putative trait-associated regions were extended from the first bin on a chromosome containing significant *k*-mers and terminated at the point where the subsequent 500 kb contained zero significant *k*-mers. Additional putative regions were then iteratively initiated from the next bin containing significant *k*-mers. Putative regions were then defined as significantly trait-associated if they contained ≥750 significant *k*-mers. Supplementary Table 3 lists all significantly trait-associated regions. Dotplot sequence alignments of the LTR Copia insertions in the candidate genes in these regions were made with the dotplot function in R package SeqinR (v4.2-36)[89].

SNP-based GWAS. SNP-based GWAS was carried out via GAPIT (v3)[90] with six different models: FarmCPU, BLINK, MLMM, SUPER, CMLM, and ECMLM. The significance threshold for associations was calculated via Bonferroni correction as follows:

$$padj = \frac{0.05}{n} = 1.01 \times 10^{-6}$$

Where n is the number of SNPs utilised for association calculations (49,660). The VCF file used was the same as that used to generate the minimal SNP panel, except that during linkage pruning, R^2 was set to 0.7 instead of 0.5. SNP-trait associations were considered significant when supported by at least two of the six models tested. We also considered nearby SNPs significant if they were supported by different models and separated by less than the LD decay distance (46 kb). Details of all significantly trait-associated SNPs are provided in Supplementary Table 4.

TT2 Sequence analysis in white and brown-grained tef accessions

The homoeologous candidate genes, Et_4A_032842 and Et_4B_039404, located in the associated peaks on chr 4A and chr 4B, respectively, were identified as orthologue of *Arabidopsis TT2* gene based on sequence homology. We compared the sequences of tef *TT2* genes between the red-grained accession Dabbi and the white-grained accession, Tsedey. For this, *TT2* genomic sequences (Et_4A_032842 and Et_4B_039404) from the Dabbi reference genome[11] were obtained from Ensembl Plant and were used as query for a BLAST search against the draft genome assembly of Tsedey[12] available at CoGe (https://genomevolution.org/coge/). Scaffolds showing more than 90% percentage identity for each query were extracted from the Tsedey genome assembly using SAMtools faidx tool[67]. *TT2* sequences in the scaffolds were annotated using the gene model for Et_4A_032842 and Et_4B_039404 from the Dabbi assembly (Supplementary Data 4 and 5). To ascertain if the insertions identified within the annotated gene models were RTs, the insertion sequences were used as queries for BLAST search against a database of repeat elements from tef, available at PlantRep[91]. Insertions with more than 90% identity to repeat elements in the database were considered as RTs. The identified RT insertions were manually annotated to highlight the long-terminal repeat and tandem site duplications at either end of the insertions (Supplementary Data 4 and 5).

Ethics and inclusion statement

This research is highly relevant to local partners in Ethiopia at the Ethiopian Institute for Agricultural Research (EIAR) and the International Livestock Research Institute (ILRI). Both institutes conduct research on and/or breeding of tef, which is a major cereal crop in Ethiopia. Local partners were involved throughout the study design, implementation, data analysis and writing. Roles and responsibilities were agreed upon amongst collaborators during initial grant acquisition processes, though additional researchers became involved throughout the research. Raw data has been made available to all collaborators throughout the research. The field trials for this study were conducted at an existing breeding station in Ethiopia and presented no known additional risks to research staff or the environment. These trials did not require special permissions from local authorities or review by a local ethics review committee. No human or animal research was conducted. Tef DNA and metabolite samples were transferred from Ethiopia to the UK with the Ethiopian government's permission. Raw data obtained from these samples has been shared with local researchers. Regional research has been utilised extensively in this study and cited accordingly.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Sequencing data is available via NCBI SRA under BioProject ID PRJNA1150514. Raw phenotypic data can be found in the Supplementary Data. Raw metabolomic data, VCF files and source data for graphs and charts in the main figure are available on Zenodo: https://doi.org/10.5281/zenodo.13837318[92].The tef germplasm used in this study is available for research purposes only upon permission by the Ethiopian government (particularly the Ethiopian Biodiversity Institute) and signing of a Material Transfer Agreement.

Code availability

Custom Bash and R scripts are available at https://github.com/Uauy-Lab/tef_kGWAS_2024.

Received: 5 October 2024; Accepted: 12 May 2025;
Published online: 26 May 2025

References

1. Ethiopian Statistics Service. *Report on Area and Production of Major Crops*. https://www.statsethiopia.gov.et/our-survey-reports/ (2022)
2. Cheng, A., Mayes, S., Dalle, G., Demissew, S. & Massawe, F. Diversifying crops for food and nutrition security—a case of teff. *Biol. Rev. Camb. Philos. Soc.* **92**, 188–198 (2017).
3. Cotter, C. J. et al. Evaluating the antioxidant properties of the ancient-crop tef (*Eragrostis tef*) grain extracts in THP-1 monocytes. *Antioxidants* **12**, 1561 (2023).
4. Sankaranarayanan, S. et al. What are the domestic and regional impacts from Ethiopia's policy on the export ban of teff? *Front. Sustain. Food Syst.* **4**, 4 (2020).
5. Lee, H. Teff, A rising global crop: current status of teff production and value chain. *Open Agric. J.* **12**, 185–193 (2018).
6. Barretto, R. et al. Teff (*Eragrostis tef*) processing, utilization and future opportunities: a review. *Int. J. Food Sci. Technol.* **56**, 3125–3137 (2021).
7. Ruggeri, R. et al. Potential of teff as alternative crop for Mediterranean farming systems: effect of genotype and mowing time on forage yield and quality. *J. Agric. Food Res.* **17**, 101257 (2024).
8. Wagali, P. et al. The effect of teff (*Eragrostis tef*) hay inclusion on feed intake, digestibility, and milk production in dairy cows. *Front. Anim. Sci.* **4**, 1260787 (2023).
9. Ramírez Gonzales, L. Y. et al. The role of omics in improving the orphan crop tef. *Trends Genet.* **40**, 449–461 (2024).
10. Shorinola, O. et al. Integrative and inclusive genomics to promote the use of underutilised crops. *Nat. Commun.* **15**, 320 (2024).
11. VanBuren, R. et al. Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.* **11**, 884 (2020).
12. Cannarozzi, G. et al. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* **15**, 581 (2014).
13. Girma, D. et al. The origins and progress of genomics research on Tef (*Eragrostis tef*). *Plant Biotechnol. J.* **12**, 534–540 (2014).
14. Bayable, M. et al. Biomechanical properties and Agro-morphological traits for improved lodging resistance in Ethiopian teff (*Eragrostis tef* (Zucc.) Trottor) accessions. *Agronomy***10**, 1012 (2020).
15. Ben-Zeev, S. et al. Unraveling the central role of root morphology and anatomy in lodging of tef (*Eragrostis tef*). *Plants People Planet* **7**, 654–665 (2023).
16. Assefa, K. et al. Genetic diversity in tef [*Eragrostis tef* (Zucc.) Trotter]. *Front. Plant Sci.* **6**, 177 (2015).
17. Blösch, R. et al. Panicle angle is an important factor in tef lodging tolerance. *Front. Plant Sci.* **11**, 61 (2020).
18. Woldeyohannes, A. B., Desta, E. A., Fadda, C., Pè, M. E. & Dell'Acqua, M. Value of teff (*Eragrostis tef*) genetic resources to support breeding for conventional and smallholder farming: a review. *CABI Agric. Biosci.* **3**, 27 (2022).
19. Zanke, C. D. et al. Analysis of main effect QTL for thousand grain weight in European winter wheat (*Triticum aestivum* L.) by genome-wide association mapping. *Front. Plant Sci.* **6**, 644 (2015).
20. Stallknecht, G. F., Gilbertson, K. M. & Eckhoff, J. L. Teff: food crop for humans and animals in *New Crops* (eds Jamick, J. & Simon, J. E.) 231–234 (Wiley, 1993).
21. Tiguh, E. E., Delele, M. A., Ali, A. N., Kidanemariam, G. & Fenta, S. W. Assessment of harvest and postharvest losses of teff (*Eragrostistef* (Zucc.)) and methods of loss reduction: a review. *Heliyon* **10**, e30398 (2024).
22. Ben-Zeev, S. et al. Less is more: lower sowing rate of irrigated tef (*Eragrostis tef*) alters plant morphology and reduces lodging. *Agronomy***10**, 570 (2020).
23. Jifar, H., Assefa, K. & Tadele, Z. Grain yield variation and association of major traits in brown-seeded genotypes of tef [*Eragrostis tef* (Zucc.) Trotter]. *Agric. Food Secur.***4**, 7 (2015).
24. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
25. Alemu, M. D. et al. Genomic dissection of productivity, lodging, and morpho-physiological traits in *Eragrostis tef* under contrasting water availabilities. *Plants People Planet* https://doi.org/10.1002/ppp3.10505 (2024).
26. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
27. Gaurav, K. et al. Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. *Nat. Biotechnol.* **40**, 422–431 (2022).
28. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
29. Day, L. Lipid chemistry in *Encyclopedia of Grain Science* (ed. Wrigley, C.) 157–165 (Elsevier, 2004).
30. Wallis, J. G., Bengtsson, J. D. & Browse, J. Molecular approaches reduce saturates and eliminate trans fats in food oils. *Front. Plant Sci.* **13**, 908608 (2022).
31. Gebru, Y. A., Sbhatu, D. B. & Kim, K.-P. Nutritional composition and health benefits of teff (*Eragrostis tef* (Zucc.) Trotter). *J. Food Qual.* **2020**, 1–6 (2020).

32. Dong, N.-Q. & Lin, H.-X. Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. *J. Integr. Plant Biol.* **63**, 180–209 (2021).
33. Konishi, S. et al. An SNP caused loss of seed shattering during rice domestication. *Science* **312**, 1392–1396 (2006).
34. Ikeda, T. et al. BELL1-like homeobox genes regulate inflorescence architecture and meristem maintenance in rice. *Plant J.* **98**, 465–478 (2019).
35. Walley, J. W. et al. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).
36. Woodhouse, M. R. et al. A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol.* **21**, 385 (2021).
37. Lam, P. Y., Zhu, F.-Y., Chan, W. L., Liu, H. & Lo, C. Cytochrome P450 93G1 is a flavone synthase II that channels flavanones to the biosynthesis of tricin O-linked conjugates in rice. *Plant Physiol.* **165**, 1315–1327 (2014).
38. Yeap, W.-C. et al. WRI1-1, ABI5, NF-YA3 and NF-YC2 increase oil biosynthesis in coordination with hormonal signaling during fruit development in oil palm. *Plant J.* **91**, 97–113 (2017).
39. Nesi, N., Jond, C., Debeaujon, I., Caboche, M. & Lepiniec, L. The *Arabidopsis* TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* **13**, 2099–2114 (2001).
40. Ren, Y., He, Q., Ma, X. & Zhang, L. Characteristics of color development in seeds of brown- and yellow-seeded heading Chinese cabbage and molecular analysis of Brsc, the candidate gene controlling seed coat color. *Front. Plant Sci.* **8**, 1410 (2017).
41. Wicker, T. & Keller, B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.* **17**, 1072–1081 (2007).
42. Chapman, M. A., He, Y. & Zhou, M. Beyond a reference genome: pangenomes and population genomics of underutilized and orphan crops for future food and nutrition security. *N. Phytol.* **234**, 1583–1597 (2022).
43. Chanyalew, S. et al. *Tef Breeding Manual* (Ethiopian Institute of Agricultural Research, 2021).
44. Milner, S. G. et al. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* **51**, 319–326 (2018).
45. Mascher, M. et al. Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081 (2019).
46. Woldeyohannes, A. B. et al. Data-driven, participatory characterization of farmer varieties discloses teff breeding potential under current and future climates. *Elife* **11**, e80009 (2022).
47. Jiang, L. et al. The APETALA2-like transcription factor SUPERNUMERARY BRACT controls rice seed shattering and seed size. *Plant Cell* **31**, 17–36 (2019).
48. Simons, K. J. et al. Molecular characterization of the major wheat domestication gene Q. *Genetics* **172**, 547–555 (2006).
49. Roeder, A. H. K., Ferrándiz, C. & Yanofsky, M. F. The role of the REPLUMLESS homeodomain protein in patterning the *Arabidopsis* fruit. *Curr. Biol.* **13**, 1630–1635 (2003).
50. Smith, H. M. S. & Hake, S. The interaction of two homeobox genes, BREVIPEDICELLUS and PENNYWISE, regulates internode patterning in the *Arabidopsis* inflorescence. *Plant Cell* **15**, 1717–1727 (2003).
51. Byrne, M. E., Groover, A. T., Fontana, J. R. & Martienssen, R. A. Phyllotactic pattern and stem cell fate are determined by the *Arabidopsis* homeobox gene BELLRINGER. *Development* **130**, 3941–3950 (2003).
52. Tsuda, K. et al. KNOTTED1 cofactors, BLH12 and BLH14, regulate internode patterning and vein anastomosis in maize. *Plant Cell* **29**, 1105–1118 (2017).
53. Fikadu, A. A., Heckelei, T. & Woldeyohanes, T. B. Technical efficiency of Teff farms controlling for neighborhood effects in Ethiopia. Research Square https://doi.org/10.21203/rs.3.rs-30863/v1 (2020).
54. Tadele, E. & Hibistu, T. Empirical review on the use dynamics and economics of teff in Ethiopia. *Agric. Food Secur.* **10**, 40 (2021).
55. Assefa, M., Mehiret, T., Purba, J. H., Bahta, M. & Haille, A. Economic analysis of tef yield response to different sowing methods and seed rates in Eastern Amhara, Ethiopia. *Agro Bali* **5**, 434–442 (2022).
56. Cannarozzi, G. et al. Technology generation to dissemination: lessons learned from the tef improvement project. *Euphytica* **214**, 31 (2018).
57. Berhe, T. *Inheritance of lemma color, seed color and panicle form among four cultivars of Eragrostis tef (Zucc.) Trotter* (University of Nebraska, 1981).
58. Chen, M. et al. The effect of *TRANSPARENT TESTA2* on seed fatty acid biosynthesis and tolerance to environmental stresses during young seedling establishment in *Arabidopsis*. *Plant Physiol.* **160**, 1023–1036 (2012).
59. Zhou, L. et al. Allelic variation of BnaC.TT2.A and its association with seed coat color and fatty acids in rapeseed (*Brassica napus* L.). *PLoS ONE* **11**, e0146661 (2016).
60. Ketema, S. *Tef (Eragrostis tef): Breeding, Agronomy, Genetic Resources, Utilization, and Role in Ethiopian Agriculture* (Institute of Agricultural Research, 1993).
61. Ebba, T. T.'ef (Eragrostis tef): Cultivars: Morphology and Classification. Part 2 Volume 66 of Experiment station bulletin, Addis Ababa university, College of agriculture (1975).
62. Assefa, K. et al. Breeding tef [*Eragrostis tef* (Zucc.) trotter]: conventional and molecular approaches. *Plant Breed.* **130**, 1–9 (2011).
63. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
65. Bonfield, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, giab007 (2021).
66. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
67. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
68. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
69. R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (2020).
70. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
71. Ortiz, E. M. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. Zenodo https://doi.org/10.5281/zenodo.2540861 (2019).
72. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
73. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
74. Xu, S. et al. GgtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* **38**, 4039–4042 (2021).
75. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
76. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).
77. Li, R. & Bilal, U. Interactive web-based data visualization with R, plotly, and shiny (Carson Sievert). *Biometrics* **77**, 776–777 (2021).

78. Winfield, M. et al. Development of a minimal KASP marker panel for distinguishing genotypes in apple collections. *PLoS ONE* **15**, e0242940 (2020).
79. López-Álvarez, D., Zubair, H., Beckmann, M., Draper, J. & Catalán, P. Diversity and association of phenotypic and metabolomic traits in the close model grasses *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Ann. Bot.* **119**, 545–561 (2017).
80. Ferreira, L. C., Santana, F. M., Scagliusi, S. M. M., Beckmann, M. & Mur, L. A. J. Metabolomics links induced responses to the wheat pathogen—tan spot—(*Pyrenophora tritici-repentis*) to the biosynthesis of flavonoids and bioenergetic metabolism. Research Square https://doi.org/10.21203/rs.3.rs-3105957/v1 (2023).
81. Finch, J. P. et al. Spectral binning as an approach to post-acquisition processing of high resolution FIE-MS metabolome fingerprinting data. *Metabolomics* **18**, 64 (2022).
82. Li, S. et al. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **9**, e1003123 (2013).
83. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
84. Pang, Z. et al. MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. *Nat. Commun.* **15**, 3675 (2024).
85. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
86. Cullis, B. R., Smith, A. B. & Coombes, N. E. On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* **11**, 381–393 (2006).
87. Walsh, B. & Lynch, M. *Evolution and Selection of Quantitative Traits* (Oxford University Press, 2018).
88. Genstat for Windows 22nd edition (VSN International, 2022).
89. Charif, D. & Lobry, J. R. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds Bastolla, U. et al.) 207–232 (Springer, 2007).
90. Wang, J. & Zhang, Z. GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinformatics* **19**, 629–640 (2021).
91. Luo, X., Chen, S. & Zhang, Y. PlantRep: a database of plant repetitive elements. *Plant Cell Rep.* **41**, 1163–1166 (2022).
92. Jones, M. et al. Variant, metabolite and source data for: Population genomics uncover loci for trait improvement in the indigenous African cereal tef (*Eragrostis tef*). Zenodo https://doi.org/10.5281/ZENODO.13837318 (2025).

Acknowledgements

The authors would like to thank Phil Robinson for help with photography, Martin Vickers for help with depositing sequence data with SRA, and Stephanie Williams for figure design advice. The authors would also like to thank Doni Hinsene and Tadelech Bizuneh for help with greenhouse tef planting and DNA extraction. Metabolite profiling was supported by Manfred Beckmann and Helen Phillips (Aberystwyth University). This work was supported by the Royal Society FLAIR Collaboration Grants 2020 (FCG \R1\201032); Biotechnology and Biological Sciences Research Council (BBSRC) through the Delivering Sustainable Wheat (BB/X011003/1) and Building Robustness in Crops (BB/X01102X/1) Institute Strategic Programmes; European Research Council (ERC-2019-COG-866328); Strategic Program for Resilient Crops: Grains for Health BBSRC grant, BBS/E/IB/230001B; Advancing Plant Health (BB/X010996/1); CGIAR Initiative on Sustainable Animal Productivity for Livelihoods, Nutrition and Gender inclusion (SAPLING). CGIAR research is also supported by contributions to the CGIAR Trust Fund. O.S. was also supported by the Royal Society FLAIR Fellowship (FLR_R1_191850). A.G. was also supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie H2020-MSCA-IF-2018 grant agreement No 842118, SUPERTEFF. J.Q.C. was supported by the Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT; 2018-000009-01EXTF-00306) and the JIC International Scholarship (2018–2022).

Author contributions

M.R.W.J.: Formal Analysis, Software, Investigation, Data Curation, Writing—Original Draft, Visualisation, Project administration; Ab.T.: Formal Analysis, Investigation, Supervision, Writing—Original Draft, Visualisation; A.G.:Formal Analysis, Investigation, Data Curation, Writing—Original Draft, Visualisation; W.K.: Investigation, Data Curation, Resources, Writing—Original Draft; Ad.T.: Investigation, Data Curation, Writing—Original Draft; D.G.: Conceptualisation, Supervision, Resources, Writing—Review & Editing; J.K.M.B.: Formal Analysis, Writing—Review & Editing; J.Q.C.: Methodology, Software; C.S.J.: Resources, Supervision, Writing—Review & Editing; B.B.H.W.: Conceptualisation, Writing—Review & Editing, Funding Acquisition; S.C.: Conceptualisation, Resources, Investigation, Project Administration, Writing—Review & Editing; K.A.: Conceptualisation, Resources, Investigation, Writing—Review & Editing; Z.T.: Conceptualisation, Resources, Writing—Review & Editing; L.A.J.M.: Conceptualisation, Resources, Formal Analysis, Supervision, Writing—Original Draft, Funding Acquisition. C.U.: Conceptualisation, Formal Analysis, Resources, Supervision, Writing—Original Draft, Project Administration, Funding Acquisition; O.S.: Conceptualisation, Formal Analysis, Investigation, Supervision, Writing—Original Draft, Visualisation, Project Administration, Funding Acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08206-5.

Correspondence and requests for materials should be addressed to Solomon Chanyalew, Cristobal Uauy or Oluwaseyi Shorinola.

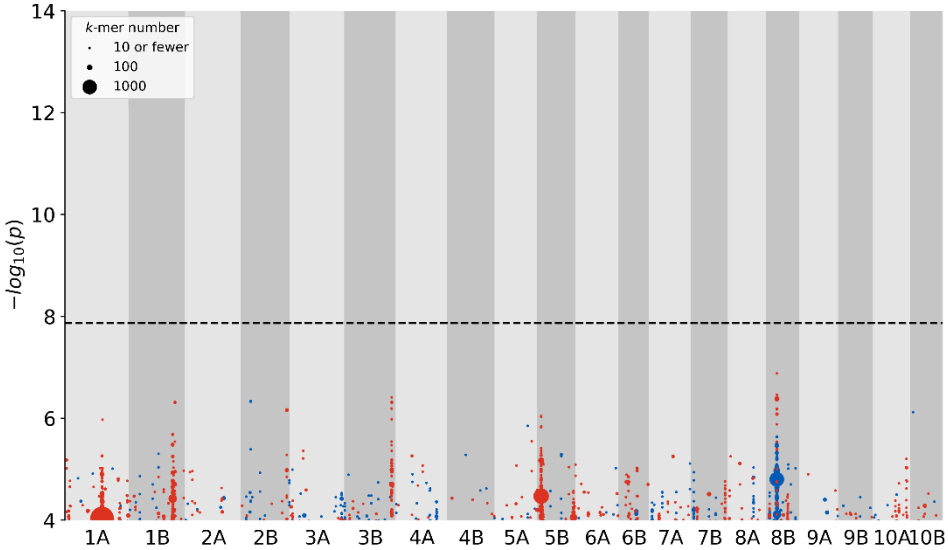
Peer review information *Communications Biology* thanks Shiran Ben-Zeev and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary handling editors: Jorge Duitama and Laura Rodríguez Pérez. A peer review file is available.

Appendix B

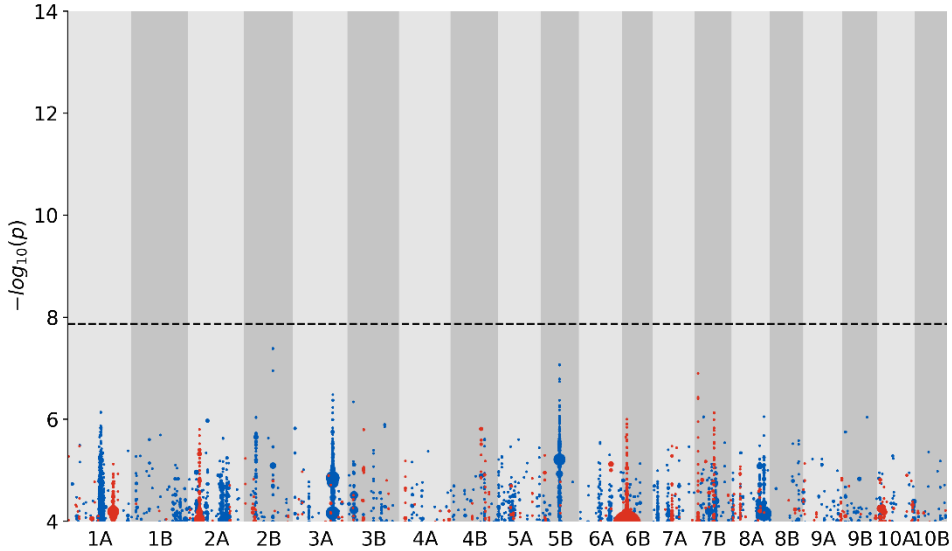
k-mer GWAS Manhattan plots for additional tef traits

In all plots below, *k*-mers are grouped according to their association level and genomic coordinates (10 kb bins) and coloured according to the direction of association; red for negative associations, blue for positive associations. Point size is proportional to the number of *k*-mers rounded upwards to the nearest 10.

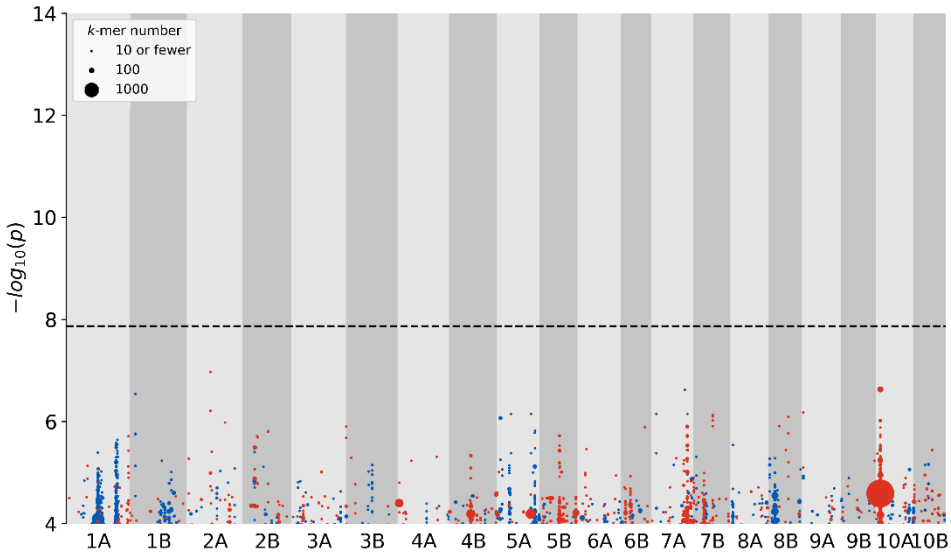
Culm length



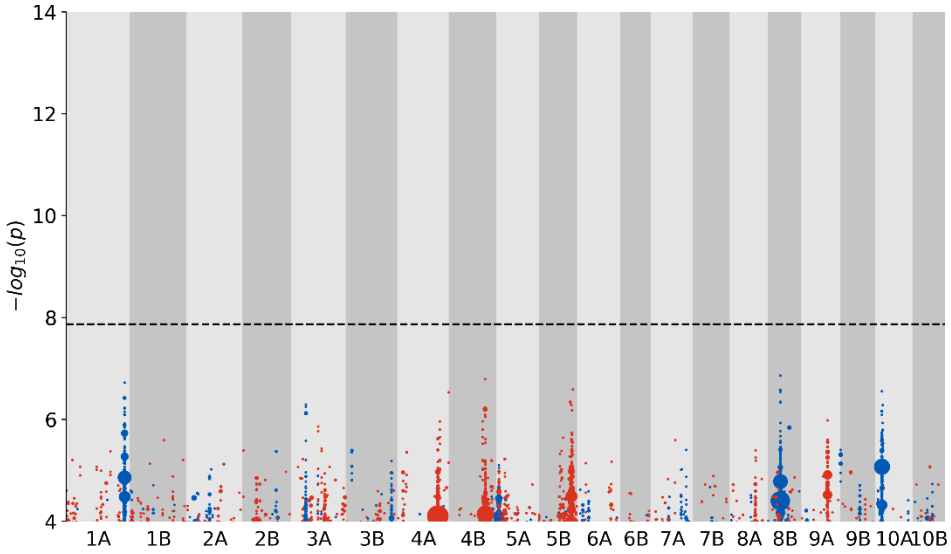
DTH



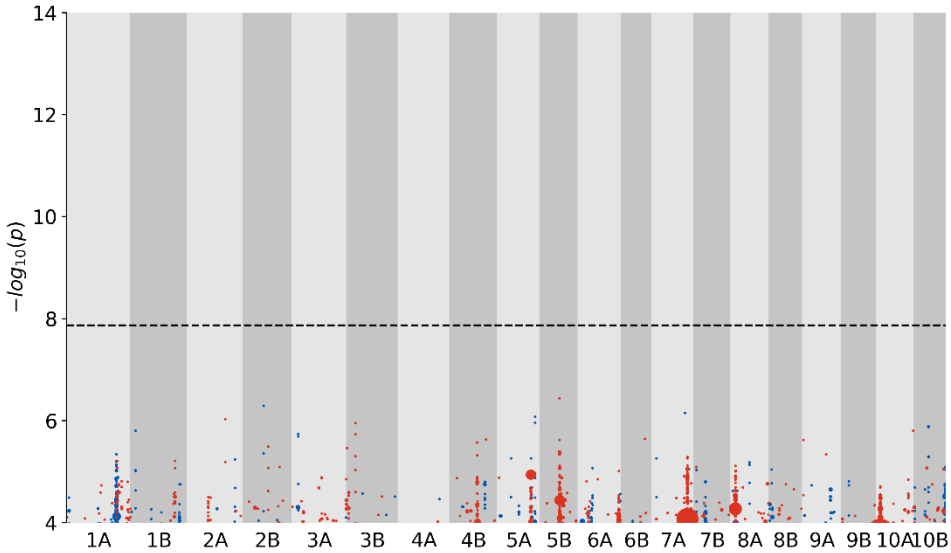
DTM



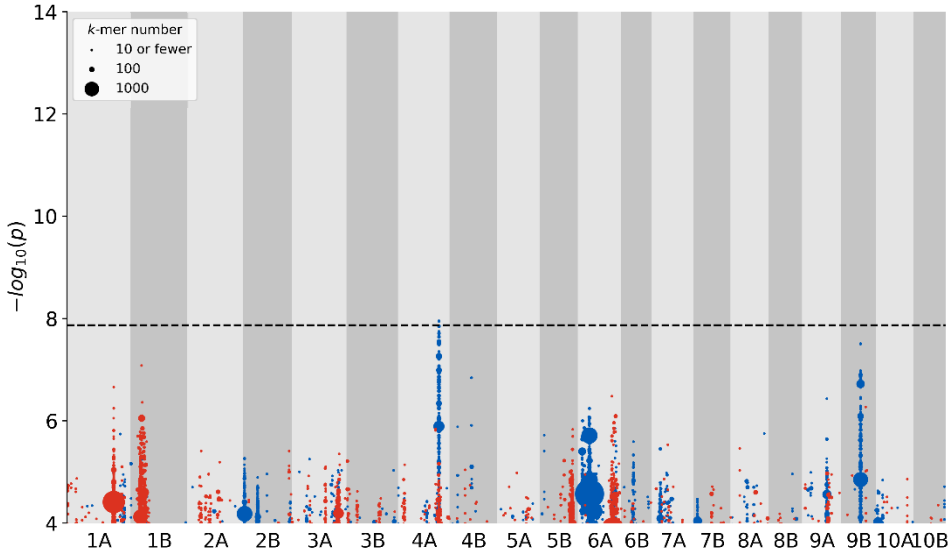
FCI length



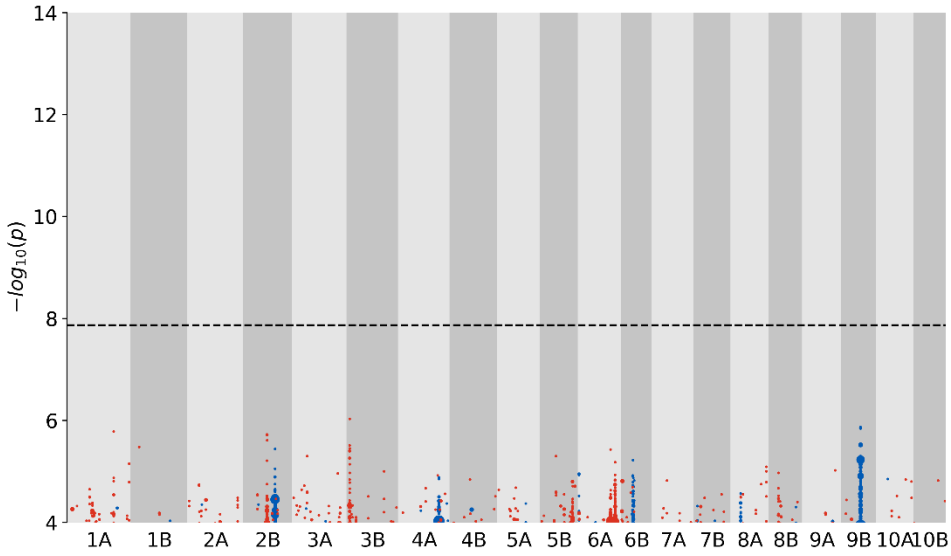
GFP



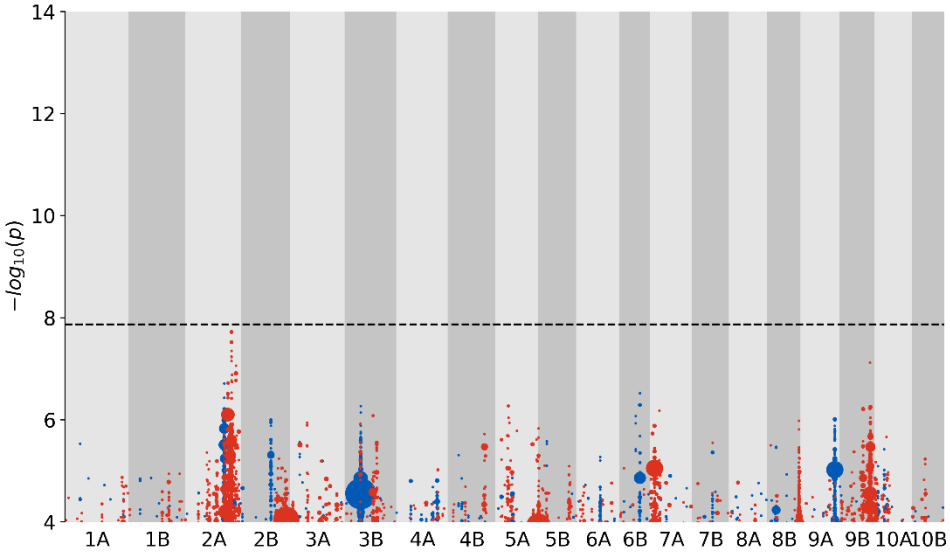
Grain yield



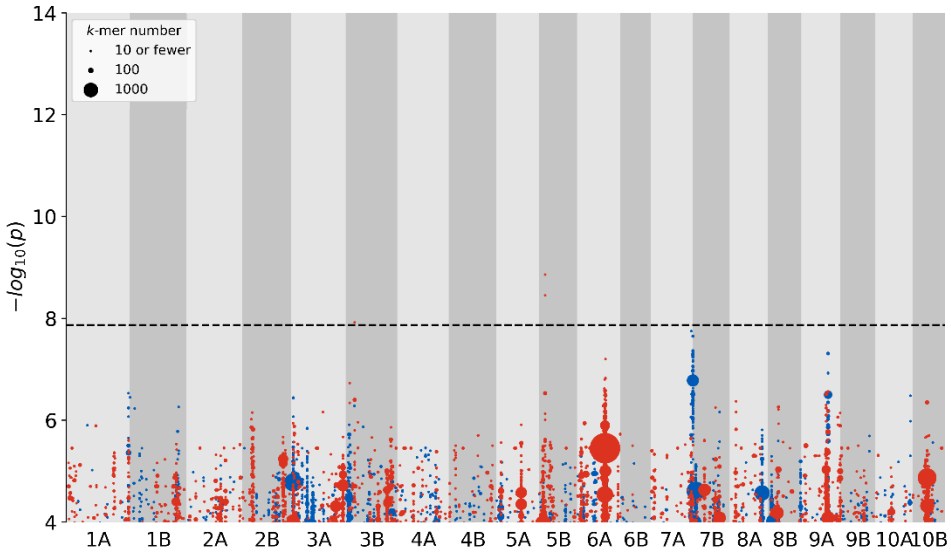
Harvest index



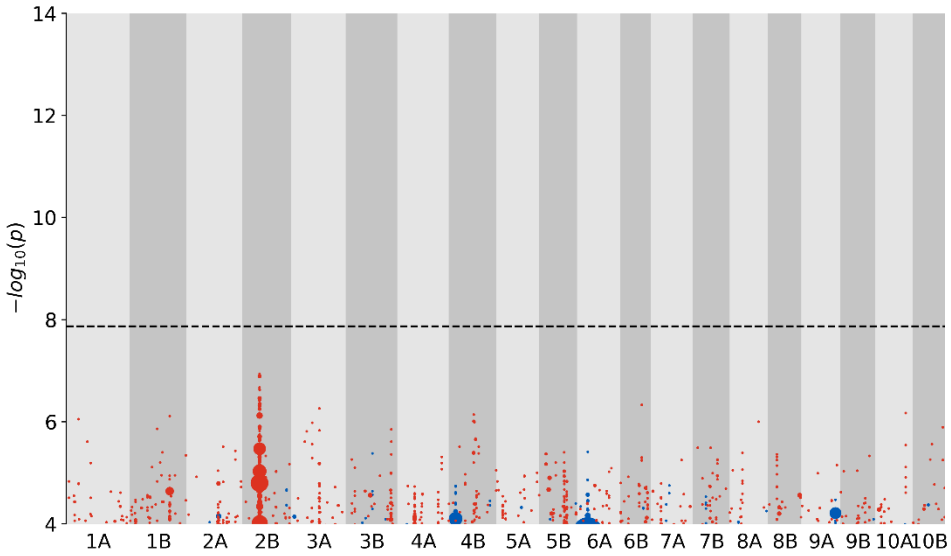
Lodging index



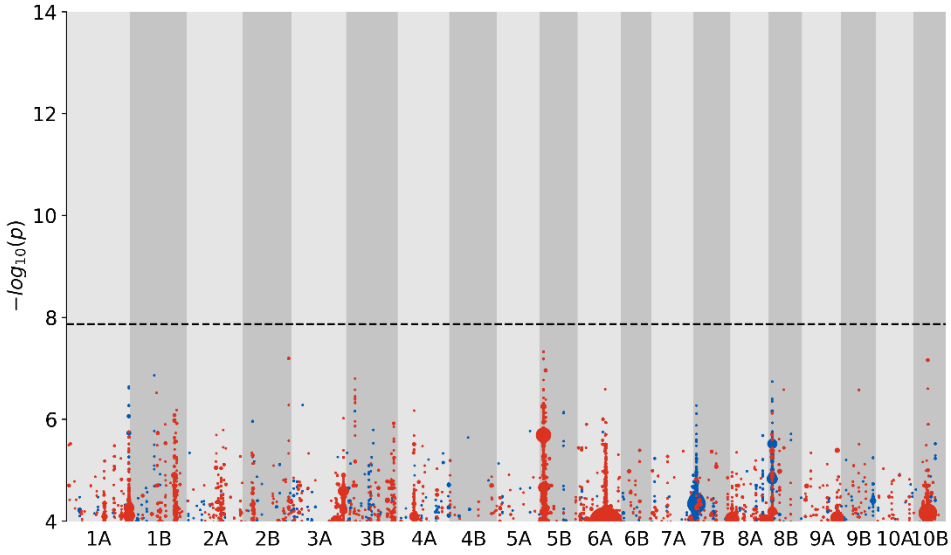
Panicle length



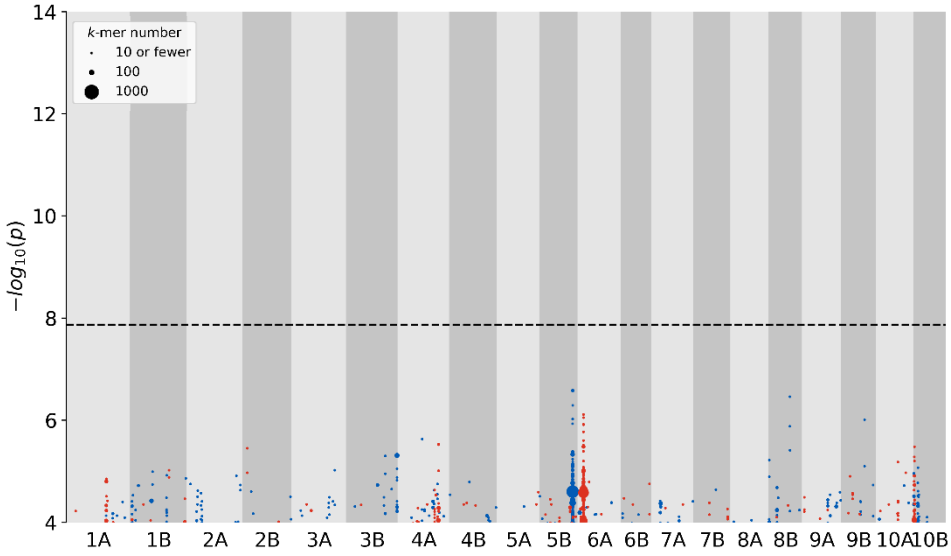
Peduncle length



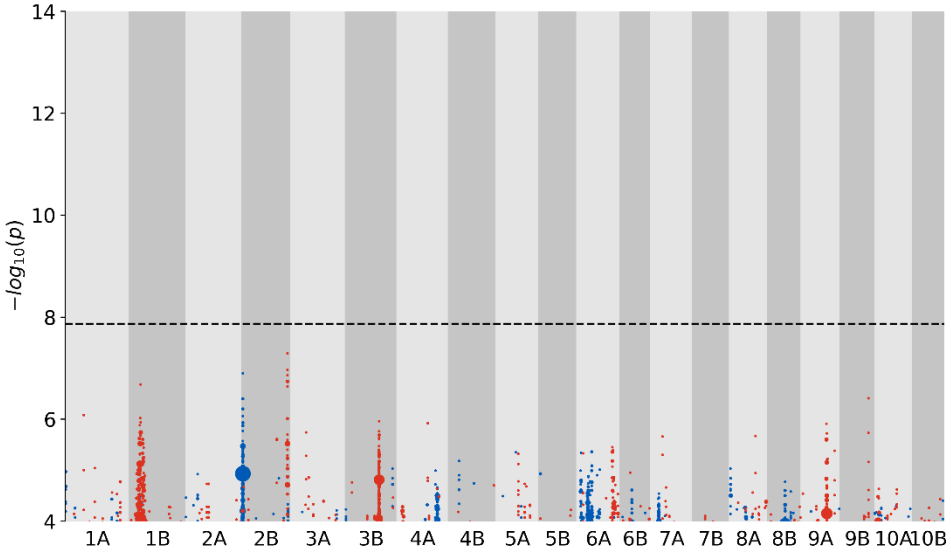
Plant height



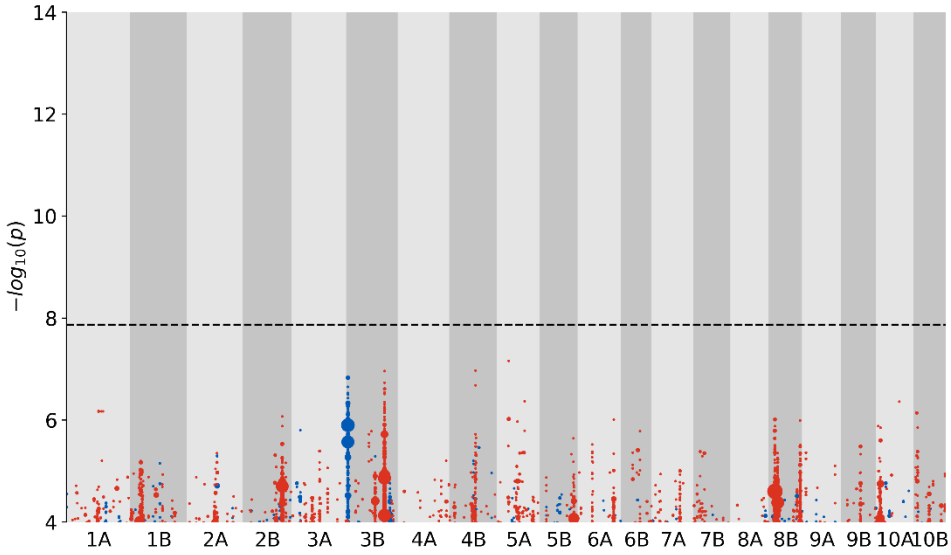
SCI length



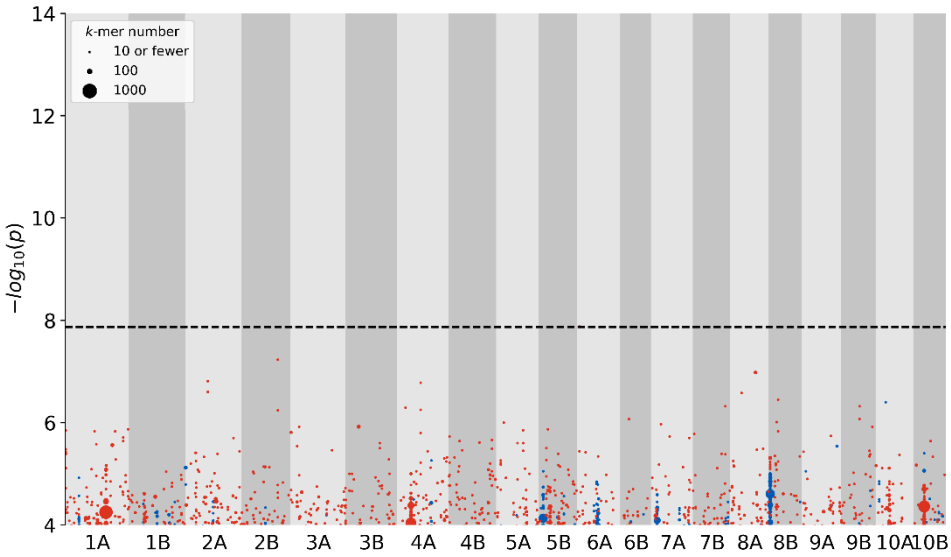
Shoot biomass



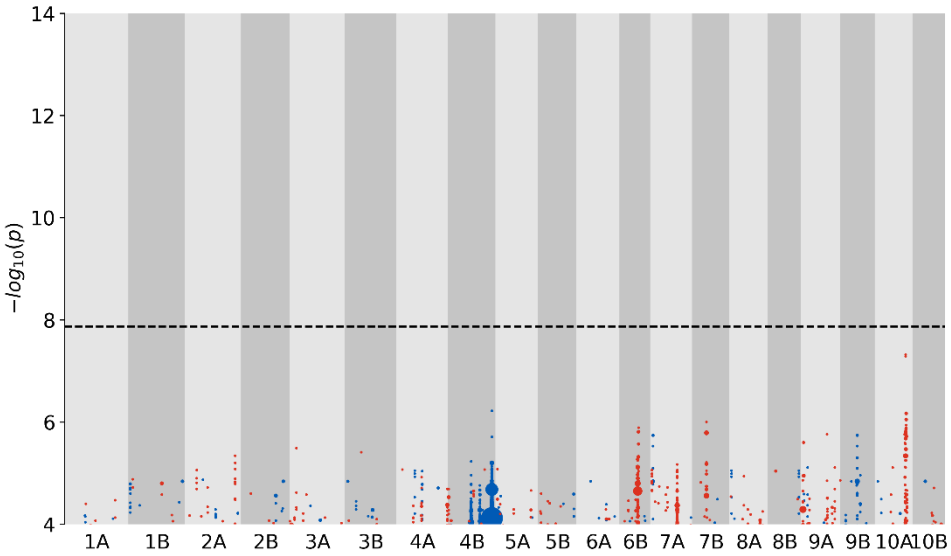
Thousand grain weight



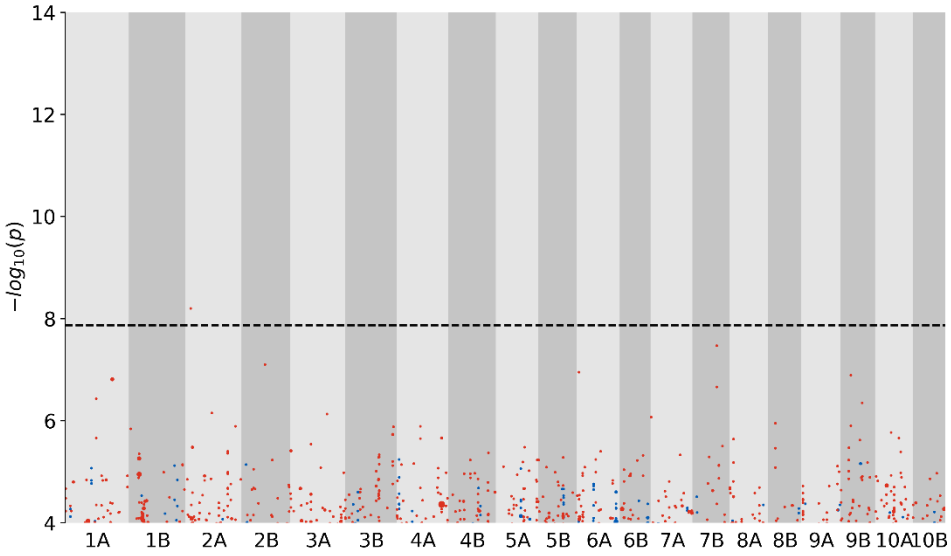
Alpha-linolenic acid (ALA)



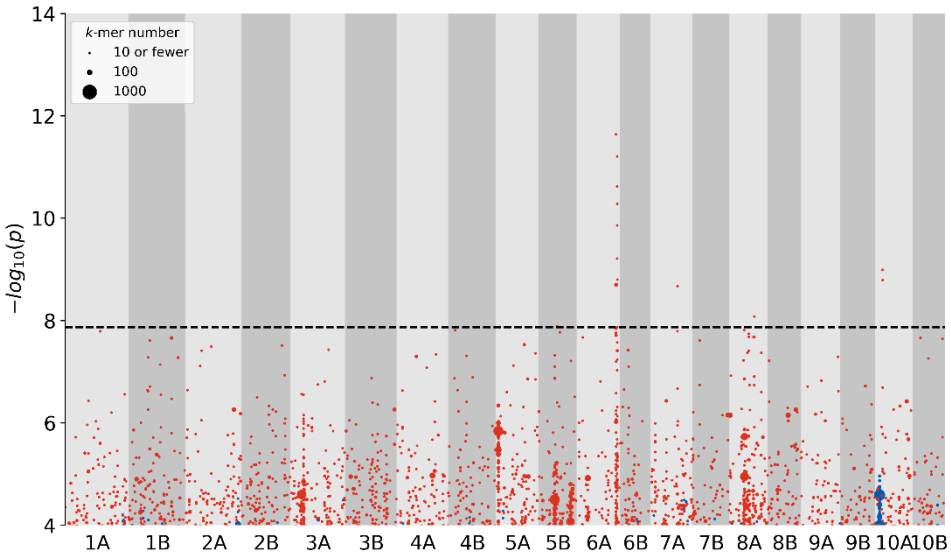
Apigenin



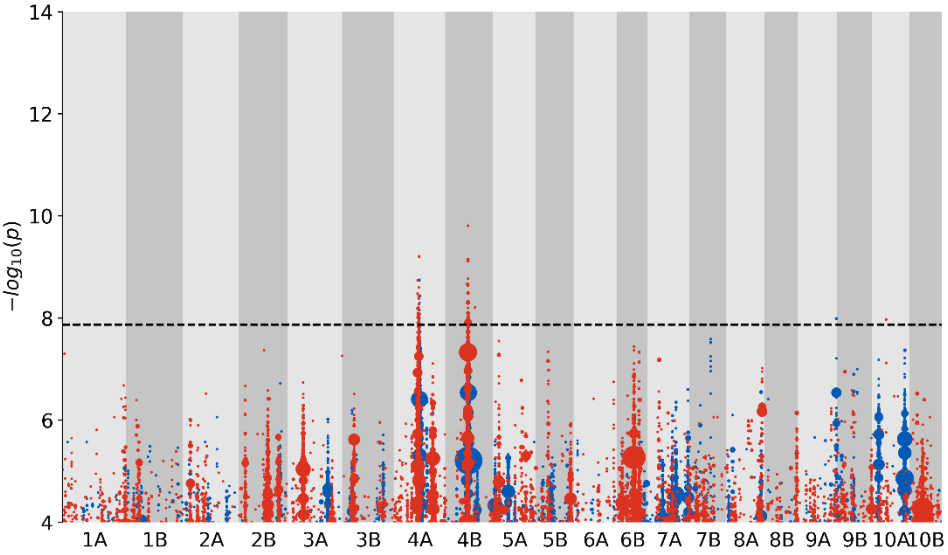
Arachidic acid



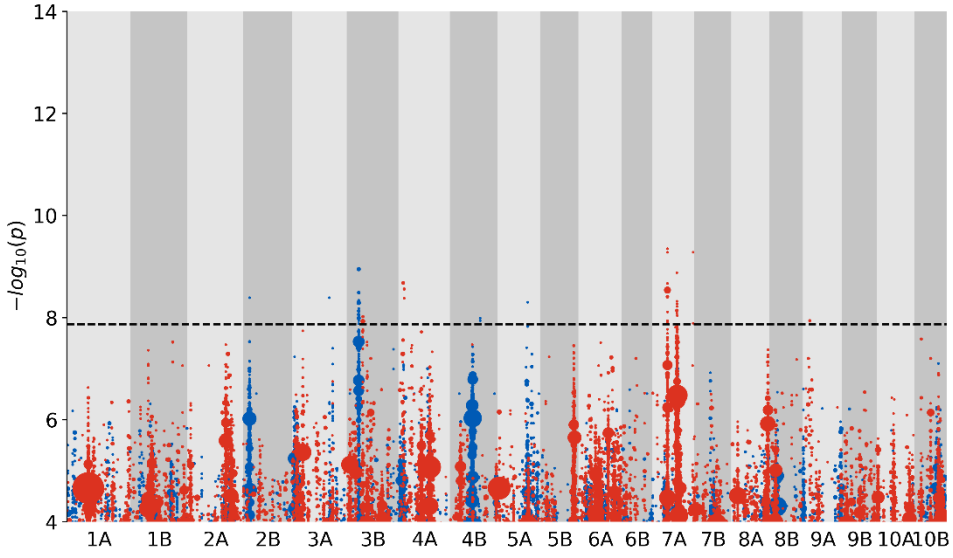
Ascorbic acid



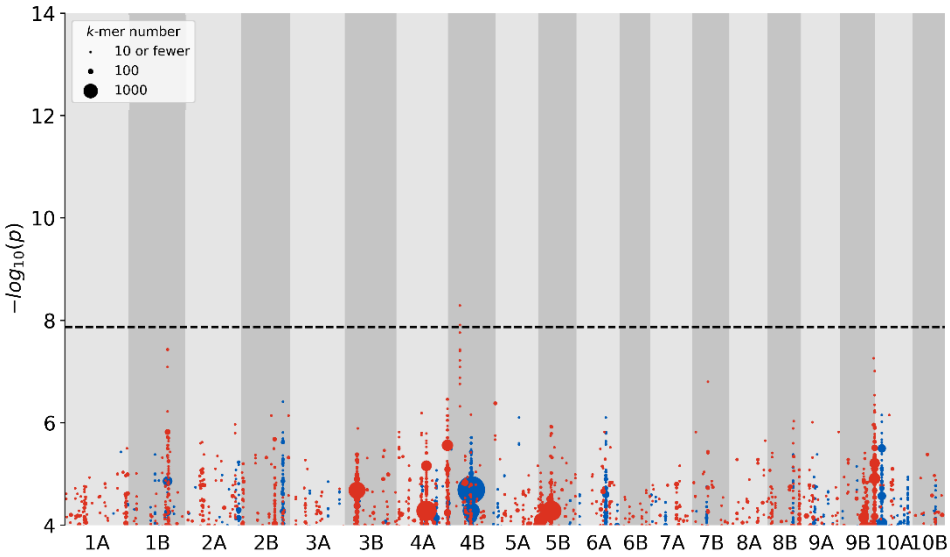
Caprylic acid



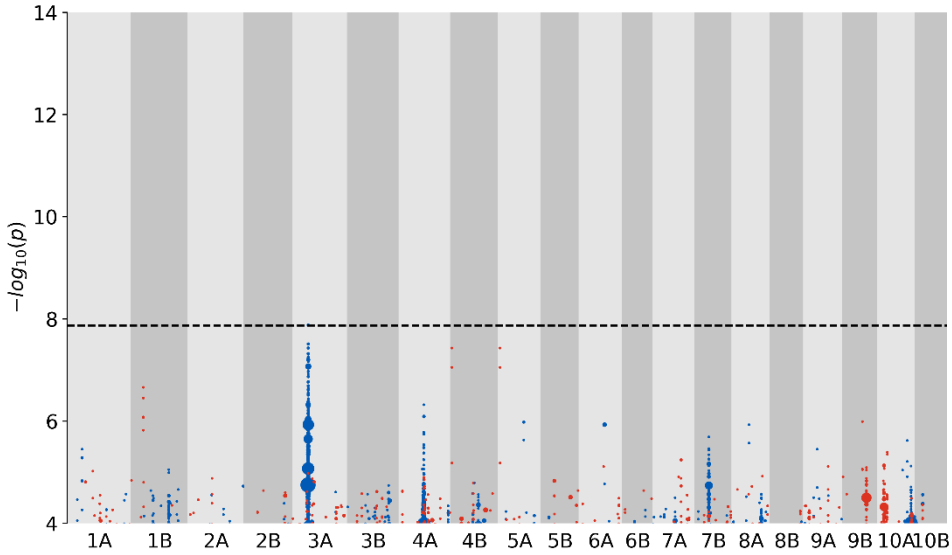
3-dehydroshikimate (DHS)



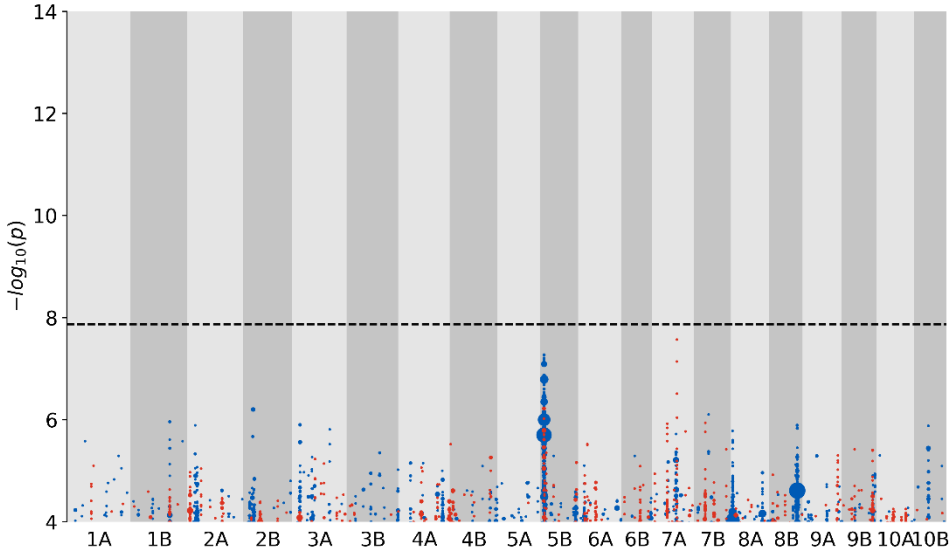
9,10-Epoxystearic acid (ESA)



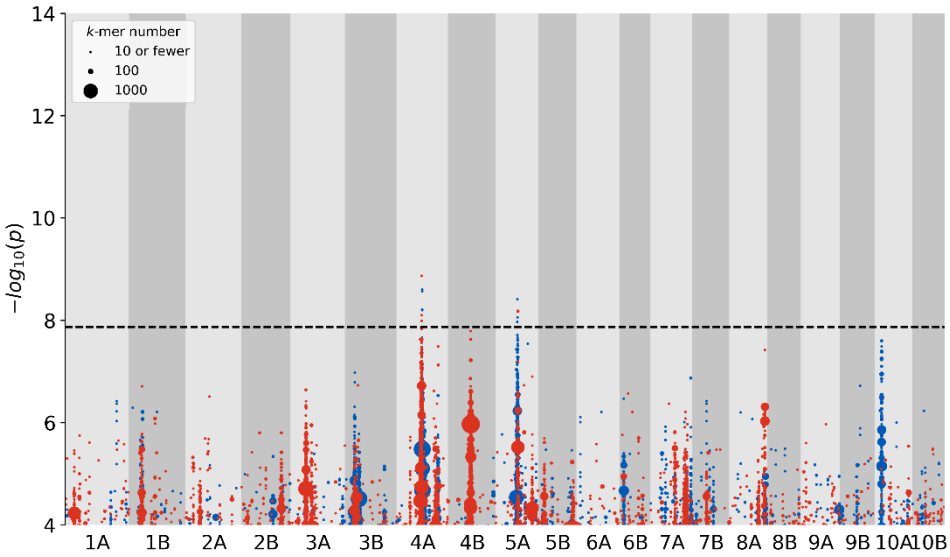
Glutathione



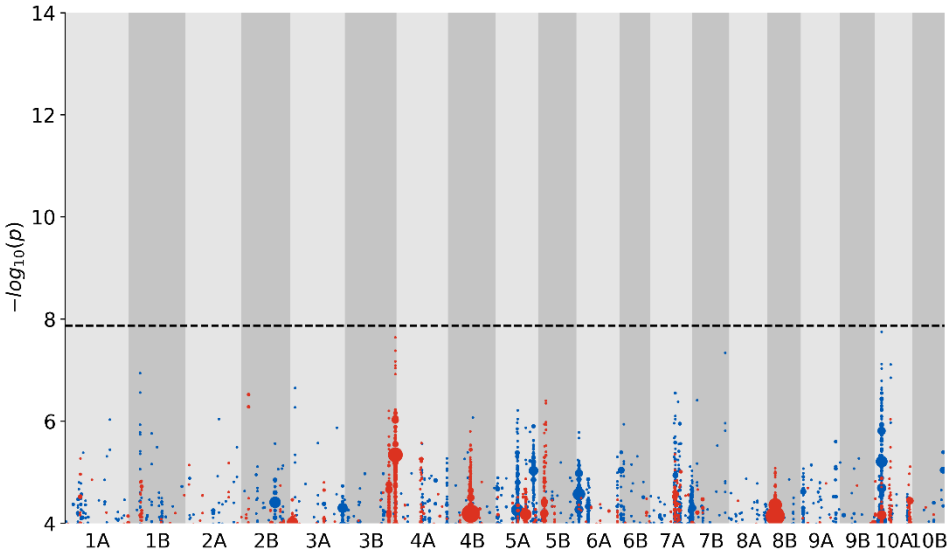
Heme



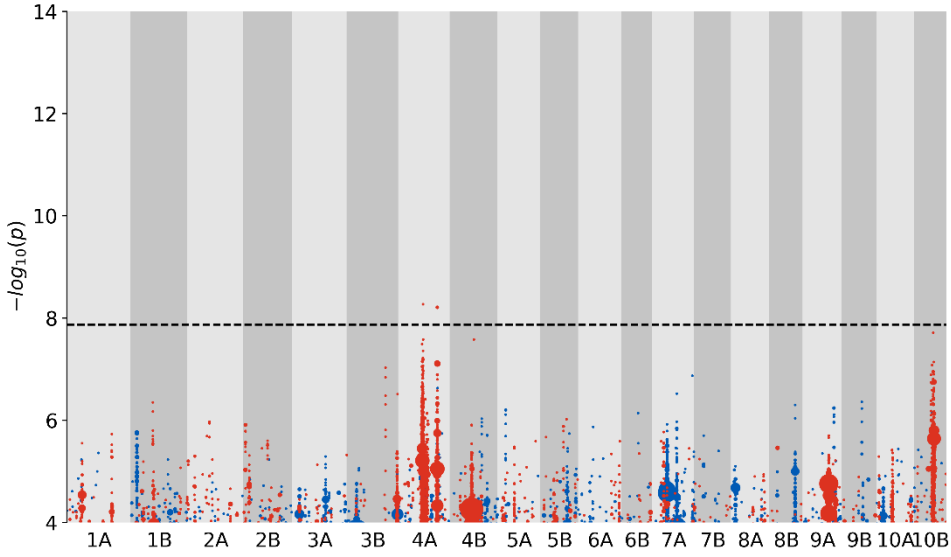
13-L-Hydroperoxylinoleic acid (HPLA)



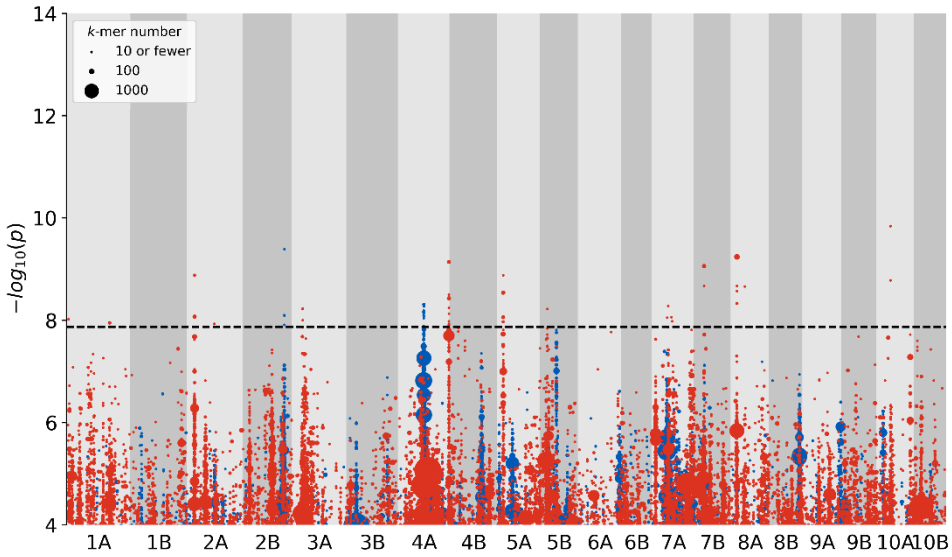
Jasmonic acid



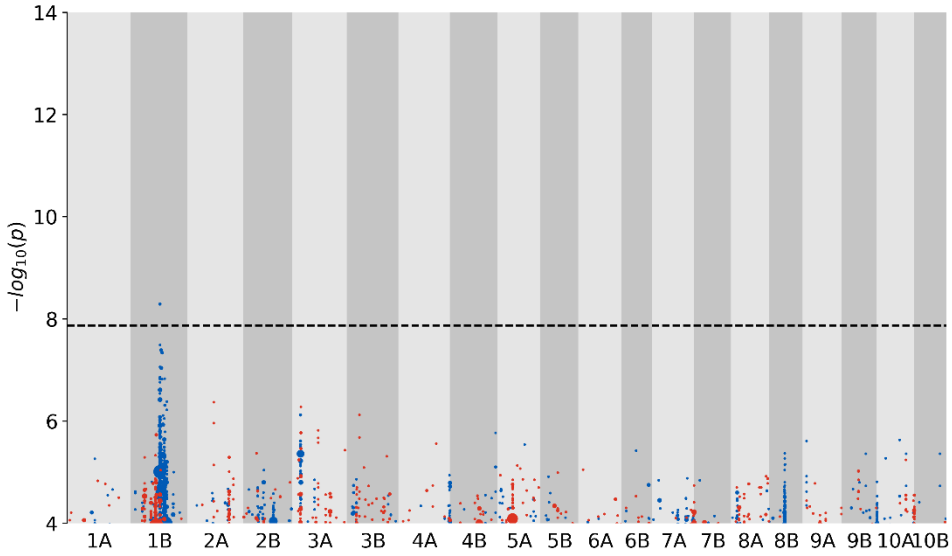
Mevalonic acid



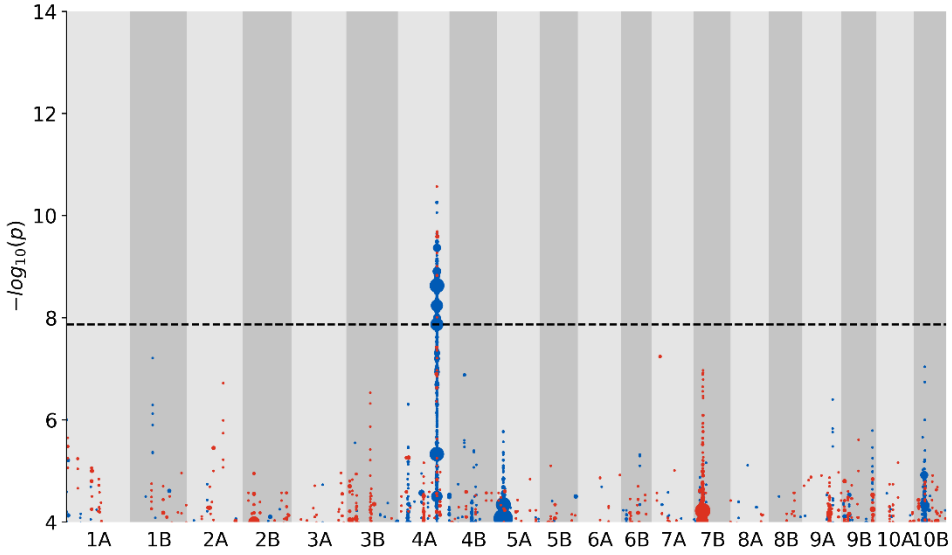
Propionic acid



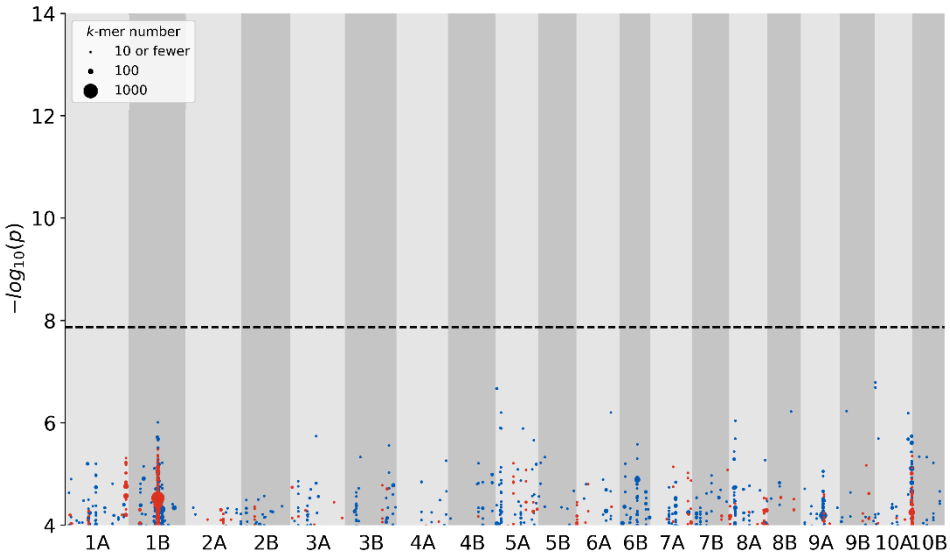
Oxidized glutathione (OG)



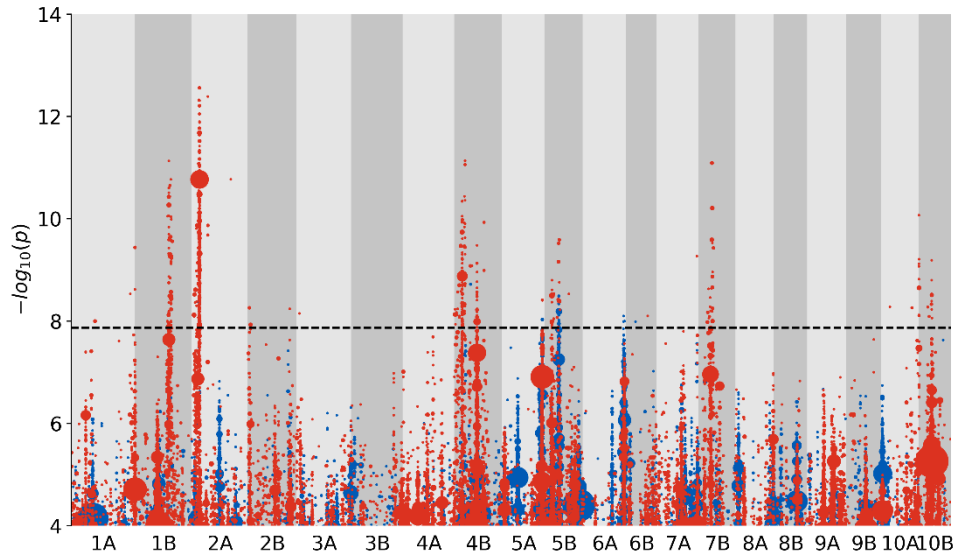
Primary fluorescent chlorophyll catabolite (PFCC)



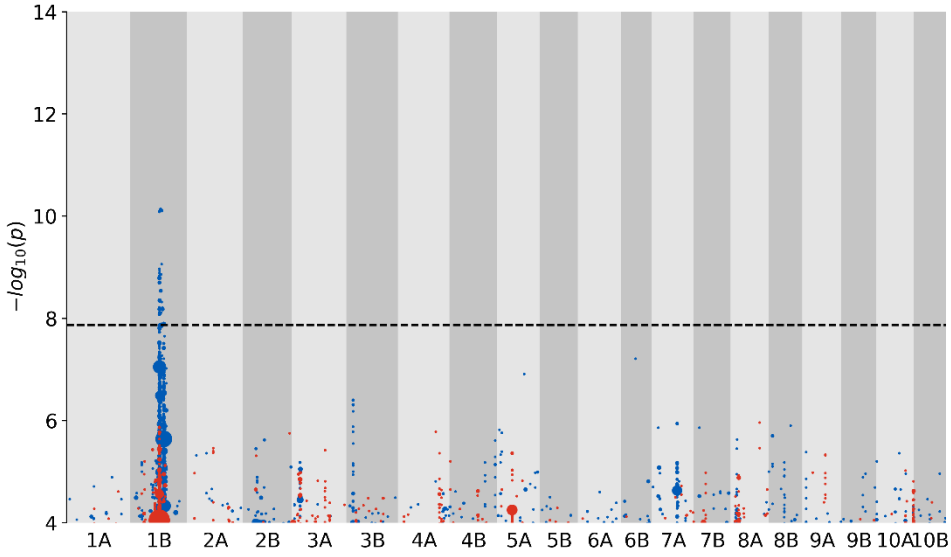
Quercitrin



Riboflavin



Rutin



Succinic acid