Wheat haplotype diversity by a *k*-mer based approach

Jesús Quiroz-Chávez

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

John Innes Centre

December 2022

This work is dedicated to my father: Rigoberto Quiroz Dominguez and mother: Maria Petra Chavez Gomez. To my two amazing brothers, Everardo Quiroz Chávez, and Ernesto Rigoberto Quiroz Chavez, and to my sister Petra Ramona Quiroz Chávez. To Maria Jose Guillen Espinoza (My friend of live), and a very special dedication to Manuel Norberto Quiroz-Chávez, my fourth brother!

Abstract

Wheat is the second most widely cultivated crop, and it is a staple food across the globe. The hexaploid form has the largest, polyploid, complex, and highly repetitive genome. Due to this complexity and size, wheat lagged in genomic studies. With advances in NGS genomics progress substantially in daily basis for many crops, including bread wheat. We now face the challenge on how to better exploit these resources for breeding to benefit food security. The main objective of this work was to develop a method to define haplotypes and a database in wheat to explore the genetic diversity in landraces and modern cultivars and link genome information with phenotypes. We embraced the challenge of using whole genome sequencing at ~12-fold coverage of more than >1,000 WGS genomes.

We developed IBSpy, a method to detect genetic variations using raw reads by *k*-mers. We benchmarked this method with previous genome alignments to detect regions which are identical by state (>99.99% sequence identity). We characterized parameters that impact in the results and provide further guidance to implement at specific situations. Our method detects variations at the resolution as with fully genome assemblies and condenses multiple types of sequences and types of variations into a single form.

Using these variations, we defined haplotypes at 1 Mbp resolution by a multi-genome approach and built a haplotype database using the >1,000 genotypes. We tracked haplotypes from landraces into modern cultivars and found that large haplotype blocks were brought into modern cultivars from landraces and are maintained through >80 years of breeding. Using these haplotypes, we conducted a haplotype GWAS, and detected genome regions associated to disease (wheat blast and yellow rust) and spike related traits. Novel unexploited haplotypes were identified in landraces absent in modern cultivars. This method integrates pangenome informed haplotypes to capture genome regions private to each assembly and can handle large WGS data.

We proved IBSpy to efficiently detect known and novel hybridisations/introgressions in the wheat pangenome and landraces at 50 Kbp resolution. We characterized a collection of *Triticum monococcum*, *Aegilops tauschii,* and large introgressions from multiple wild relatives and propose candidate genotypes to be the closest donors of those hybridisations/introgressions. Using these haplotypes, we identified novel hybridisations of *Ae. tauschii* in the D subgenome of wheat absents in the pangenome references. These results demonstrated the utility of our haplotype calls using an alternative approach to the conventional aliments methods. We created a flexible and wide haplotype database based on *k*-mers to which novel 12-fold WGS genotypes can be added and easily integrated in the context to this haplotype database.

Acknowledgements

First, I would like to thank my supervisor, Cristobal Uauy for giving me this great opportunity of learning Bioinformatics from the scratch! Thank you for your support, guidance, and patience. I could not have asked for a better lab group and supervisor.

Many thanks to Ricardo Ramirez-Gonzalez who introduced me to the world of programming with lots of patience. Thanks for all your suggestions, recommendations. This project would not have never reached this point and impact without your great contribution.

I would also like to thank my second supervisor Richard Morris for his guidance and great recommendations during my reviews. I would like to thank Brande Wulff and Kumar Gaurav for their contributions in the initial part of the project and for introducing me to the world of *k*-mers.

I would also like to thank the "Consejo Nacional de Ciencia y Tecnologia (CONACYT)" and John Innes International Scholarship for funding my PhD. Thanks to the DFW project for funding this project.

Thanks to all members of the Uauy Lab and special thanks to, James Simmonds, Tobin and Pam and the field team for their support with field and glasshouse work. Thanks to Nikolai A. for his guidance in the lab-related things.

Thanks to all the initial unknown in lab students Aura, Marina, Anna, Bijan, and Andy, who became amazing friends along these four years!

Outside the lab, I have found an excellent group of friends in my fellow PhD students, with whom I have many fun memories and hopefully many more to come.

Last, but not least, to my great family who always support me in any decisions and in the new challenges I embrace. They are truly my main motivation to progress and do good things in life! Thanks to Maria Jose for all these years together and for your patience with all the difficulties in these four years away! We made it! Huge thanks to my childhood friends (the MVS-PBS neighbourhood; sifo, bufon, drama, sello, fakso, chon, jace, formek, pelon, irak) who stayed in my life forever and are the ones to talk about anything when things are difficult. Really a second family and lucky us!!

Table of contents

List of figures

List of tables

Supplemental figures

Supplemental Tables

List of Abbreviations

2NS/2As	*Aegilops ventricosa* Chromosome Arm 2NS Segment Translocated to Wheat Chromosome Arm 2AS
AgRenSeq	Association Genetics RenSeq
AP	Affinity Propagation
API	Application Programming Interface
BBSRC	Biotechnology and Biological Sciences Research Council
cDNA	Complementary DNA
CS	Chinese Spring
DFW	Designing Future Wheat
dmp	Damping
dNTPs	Deoxyribonucleotide Triphosphate
EMS	Ethyl Methanesulfonate
ESTs	Expressed Sequenced Tags
FN	False Negative
FP	False Positive
Gb	Giga Basepair
GBS	Genotyping By Sequencing
GediFlux	Genetic Diversity in Agriculture: Temporal Flux, Sustainable Productivity And Food Security
GISH	Genomic In Situ Hybridization
GLM	Generalized Linear Model
GMM	Gaussian Mixture Models
GWAS	Genome Wide Association Study
hapGWAS	Haplotype GWAS
HAWK	Hitting Associations With K-Mers
HiFi	High Fidelity
IBSpy	Identity By State in Python
InDels	Insertion And Deletions
IWGSC	International Wheat Genome Sequencing Consortium
KASP	Kompetitive Allele Specific
Kbp	Kilo Basepair
kGWAS	K-Mer GWAS
L1	Lineage 1
L2	Lineage 2
L2-SA	Lineage 2 Subgroup A
L2-SB	Lineage 2 Subgroup B
L3	Lineage 3
LD	Linkage Disequilibrium
LNISKS	Longer Needle in A Scanter *K*-Stack

Mbp	Mega Basepair
NAM	Nested Association Mapping
NB-ARC	Nucleotide-Binding Domain Shared With APAF-1, Various R-Proteins And CED-4
NBS-LRR	Nucleotide-Binding Leucine Rich Repeat
NIAB	National Insitute of Agicultural Botany
NIKS	Needle In The *K*-Stack
NL	National List
NLR	Nucleotide-Binding Leucine-Rich Repeat
NVPT	National Variety Performance Trials
PCA	Principal Component Analysis
Ph1	Pairing Homoeologous 1
QTL	Quantitative Trait Loci
RHT1	Reduced Height 1
RL	Recommended List
RNAseq	RNA Sequencing
SC	Silhouette Coefficient
SD	Segregation Distortion
SMRT	Single Molecule Real-Time
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SOM	Self-Organizing Maps
SRA	Super Race Avirulence
SSD	Single Seed Descent
SSR	Simple Sequence Repeat
SV	Structural Variants
TN	True Negative
TP	True Positive
UPGMA	Unweighted Pair Group Method With Arithmetic Mean
VCF	Variant Call Format
WatSeq	Watkins Sequecing
WGS	Whole Genome Sequencing
YR7	Yellow Rust Disease Resistance Gene 7
UK	United Kingdom
ssp	Subspecies
CIMMYT	International Centre For Wheat And Maize Improvement
BGI	Beijing Genomics Institute
DNBSeq	DNA Nanoball Sequencing
MAS	Marker Assisted Selection
GS	Genomic Selection
NIKS	Needle In The *K*-Stack

RenSeq	Resistance Genes Enrichment Sequencing
IL	Isogenic Lines
MUMmer	Maximal Unique Matches
IBS	Identity By State
CENH3	Centromeric Histone H3 Variant
DArTseq	Diversity Array Technologies Sequencing
JIC	John Innes Centre
YR5	Yellow Rust Disease Resistance Gene 5
WAPO1	*Wheat Ortholog Of ABERRANT PANICLE ORGANIZATION1 (APO1) Gene*
DPI	Days Post Inoculation
INRA	Institut National De La Recherche Agronomique
RFLP	Restriction Fragment Length Polymorphism
NGS	Next-Generation Sequencing
CPU	Central Processing Unit
RAM	Random-Access Memory
PHG	Practical Haplotype Graph

1. General introduction

1.1. Wheat germplasm resources

1.1.1. Wheat as a major crop

Wheat is the second most widely cultivated crop with over 220M hectares harvested worldwide in 2021. The largest wheat producers are China, India, Russian Federation, USA, and France (FAOSTAT, 2022, https://www.fao.org/faostat/en/#data). The global wheat production in 2021 was 770M tonnes, with productions above 700M tonnes over the past decade. The main use of wheat is for human consumption and is divided mainly in pasta (*Triticum durum*; about 5% of global consumption) and bread (*Triticum aestivum*; roughly 95% of global consumption) wheat and it is used for feeding in some regions, such as the United Kingdom (UK).

Yields of wheat production has been affected by biotic and abiotic stresses since its domestication and these factors have been intensified in recent years. Important pathogens affecting wheat production include yellow rust, stem rust, leaf rust (*Puccinia* ssp.), wheat blast (*Magnaporthe oryzae Triticum*), and fusarium head blight (*Fusarium* ssp). It is estimated that these diseases generate between 10-28% loses on yield production (Savary et al., 2019). The immediate strategy used by farmers is to employ chemical fungicides and pesticides to fight back those threats. However, breeding efforts to develop resistant varieties and managing fields rotating crops are increasingly becoming more important with recent bans on the use of chemical inputs. In addition to the biotic threats, wheat production its threatened by abiotic stresses such as droughts, heat, and extreme colds in spring wheat growing regions. It is predicted that those threats will continue with the unpredictable changes in climate. Alongside, wheat production faces a new challenge with the recent conflict between countries, such as the Russian invasion of Ukraine (third and sixth worldwide wheat producers, respectively). As a result, several developing countries are suffering the lack of this basic staple food for their daily basis (Bentley et al., 2022).

1.1.2. The origin of hexaploid wheat

Bread hexaploid wheat consist of three sub genomes originated from two main interspecific hybridizations (*Triticum aestivum L. AABBDD*, 2n = 6x = 42). The first hybridization originated between *Triticum urartu* (AA), the main donor of the A subgenome, and a closely wild relative of *Ae. speltoides* (BB), the B subgenome donor (Daud & Gustafson, 1996; Miki et al., 2019). This event gave rise to the wild tetraploid *Triticum turgidum* (AABB) (Huang et al., 2002). A second hybridization event from a cultivated tetraploid wheat *T. turgidum* ssp. *dicoccum* with the *Ae.*

tauschii (DD) genome gave rise to the hexaploid bread wheat *T. aestivum* (Miki et al., 2019). Thus, the D subgenome was originated from the diploid *Ae. tauschii*.

Subgenomes A and B diverged from a common ancestor ~7 million years ago while the D genome was originated from a hybridisation and subsequent speciation event between the A and B genome donors ~5 million years ago (Fig. 1.1)(Li & Gill, 2006; Marcussen et al., 2014). Historical records and the absence of wild hexaploid wheats suggest that modern *T. aestivum* originated after the first wheat domestication ~8-10 thousand years ago (Salamini et al., 2002). This hypothesis is supported by historical and archaeological records of hexaploid wheat cultivation in the region of the Fertile Crescent in the North of Iran. It is hypothesized that the hexaploidy wheat expanded from this place to several other regions worldwide including Europe and Asia as a result of its wide adaptation from its polyploid genome nature (Dubcovsky & Dvorak, 2007).

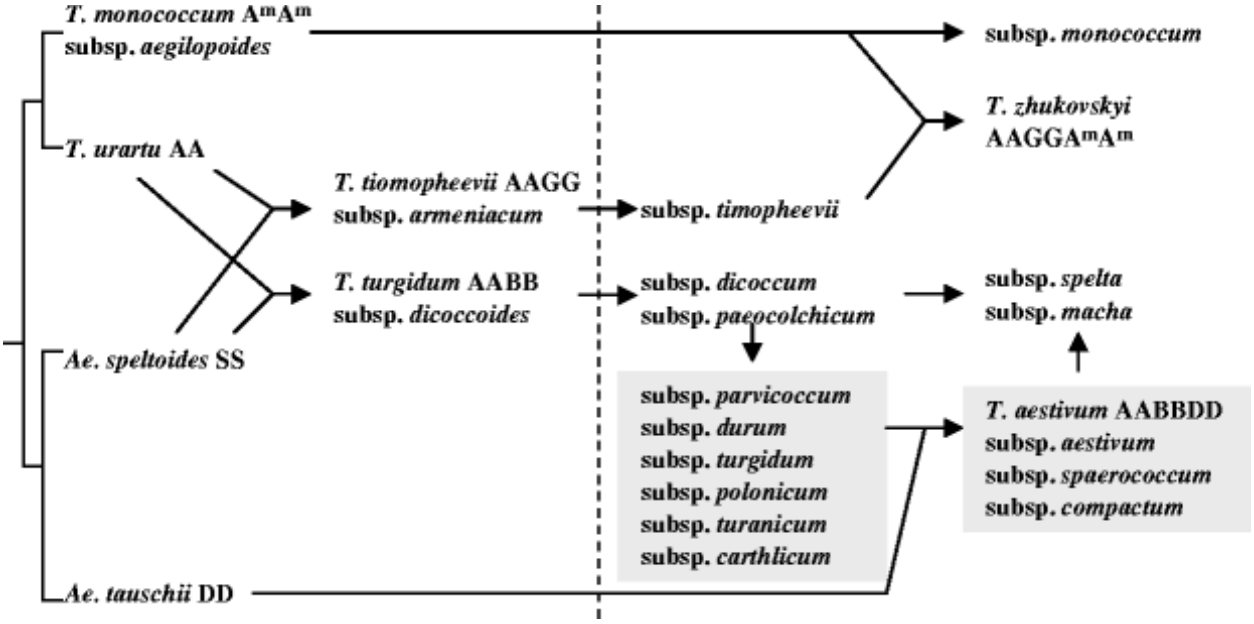


Fig. 1. 1 From (Li & Gill, 2006). Evolutionary relationships among different wheats and their domestication.
"The vertical dashed line separates the wild species (left) from the domesticated forms (right). The species and subspecies marked with grey background are free threshing. The genome formula follows the species".

1.1.3. The Watkins Landraces Collection

A landrace is a locally grown cultivar of a crop commonly cultivated by farmers who keep and propagate their "best seeds" after each generation of cultivation. These landraces are usually well adapted to small regions and due to the nature of being open pollinated, maintain a relatively high level of heterogeneity depending on the species and type of cultivation compared to developed

elite modern varieties. Large collections of landraces for different important crops or "orphan crops" are maintained in seedbanks as a germplasm reservoir. In several cases those accessions remain unexploited by breeders or geneticist due to the difficulty and time consuming to "clean-off" undesirable traits. It has been demonstrated, however, that landrace collections maintain unvaluable agronomically important alleles particularly against pathogens and for nutritional value (Sansaloni et al., 2020; Würschum et al., 2022).

Historically, landrace collections were developed by individual researchers or geneticists for different crops. Later those collections were created systematically by institutions or universities in attempts to capture the widest genetic diversity of a species (Langridge & Waugh, 2019; Mascher et al., 2019; Schulthess et al., 2022). Wheat is not the exception, and large collections of landraces and cultivars have been maintained and are available worldwide. For example, important landrace collections of wheat are maintained at CIMMYT (Sansaloni et al., 2020; Vikram et al., 2016), INRA (Balfourier et al., 2019), IPK (Schulthess et al., 2022), Central Europe (Cseh et al., 2021), GediFlux (Aradottir et al., 2017), and in JIC to mention some. The former maintains and curate the Watkins (Wingen et al., 2017) collection.

In this work, we employed the Watkins collection created in the UK. The Watkins collection initially consisted of more that 7,000 bread and durum wheat accessions collected by A.E. Watkins in the 1930s at the School of Agriculture in Cambridge. It includes cultivars collected from local markets from 32 countries covering Asia, Europe, Africa, and the American continent (Wingen et al., 2014). Thousands of accessions from the collection were lost and 1,291 accession remain accessible at the John Innes Centre. From this set of lines, a process of selection and stabilization by Single Seed Descent (SSD) to remove some of the mixture and heterogenicity was applied by the Griffith's group (Dr. Simon Griffiths) and 827 remain as the "Watkins Stabilised Collection of Hexaploid Landrace Wheats". Although this process of stabilization removed most of the heterogenicity, after a century of regeneration and seed propagation, some pollen cross contamination may have occurred among them. In our study having homozygous genotypes is of importance since our method to detect genetic variations is affected by heterogenicity since it cannot differentiate between a heterozygous or homozygous loci in a genotype. Instead, our method will detect a locus as present regardless of its level of heterogenicity which could impact on the results and conclusions. This topic will be discussed in Chapter 2.

Analysis between the SSD sister lines revealed on average 35% alleles differing among them when using 41 simple sequence repeat (SSR) markers (Wingen et al., 2014). Out of 827 accessions, 86% are spring and 14% are winter habit. Initially were grouped into nine ancestral geographical groups by using reduced genotypic data but were condensed later to seven when using whole genome

re-sequencing data (unpublished data). Analysis of the 41 SSR markers revealed on average 22.4 alleles per loci with a minimum of 3 and a maximum of 61 in the entire 827 Watkins accessions. The collection was also characterised with single nucleotide polymorphism (SNP) arrays by (Wingen et al., 2017) and (Przewieslik-Allen et al., 2021) who identified a series of wild wheat introgressions, chromosome rearrangements, and deletions which we will further explore in Chapter 4. From this collection, a core set sub-collection of 119 accessions that captures ~97% of the total genetic diversity was subcategorized. This genetic diversity has been consistently reflected on the phenotypes observed among the accessions for different traits (Wingen et al., 2014). This core set it is included in the WatSeq initiative and in this thesis.

Although, it is difficult to predict the year of origin from landraces and its cultivation, several accessions from the WatSeq collection may have acted as founders for important breeding programs as documented in history books of breeding. Therefore, it is likely that many intact genome regions made their way through into modern cultivars and it won't be unexpected to find large intact genome regions being selected by breeders into modern cultivars. There may be, however, an invaluable set of novel alleles still unexploited in the collection that could help to improve novel modern varieties for yield and other agronomically important traits.

1.1.4. Wheat breeding and current modern cultivars

Breeding is defined as the act of selecting genotypes with desirable traits to directly or indirectly satisfy human needs (Adams, 1962). Plant breeding has had great impact on food production by the release of new cultivars worldwide for different crops with better yields, disease resistance, nutritional value, and several other important traits (Borlaug, 1983). Early plant improvements were made directly by farmers and these "early breeders" selected traits empirically observing and keeping the best plants for subsequent generations. As described in section 1.1.3., these early and locally adapted groups of plants selected by farmers are known as landraces. For example, historical evidence suggests that the free-threshing spikes on cereals was one of the first trait selected during domestication (Peng et al., 2011). Selecting and keeping this trait facilitated manual work to separate grains from the lemma, palea, and glumes during the post-harvest process.

Later, with the progress in science and new discoveries in genetics with the work of Mendel and other geneticists (e.g., Biffen, Saunders, Bateson), selection followed by hybridization in wheat improvement was initiated in the late 19th century simultaneously in France, Germany, Canada, US, and Australia. Early breeding methods evolved from bulking and pure line selection to directed

hybridizations controlling the process of pollination. By region, these early wheat breeding schemes were dominated by the private sectors in Europe and mainly by the government in Canada, US, and Australia. The progress was constant and for the middle of the 20th century the release of pure lines was increasingly common in both, public and private sector (Baenziger & Principal, 2009).

The Watkins landrace collection described in section 1.1.3 is a representation of the genetic diversity in wheat before modern cultivars and elite breeding lines. After landraces, early cultivars dominated the acreage of agronomically important regions worldwide. One example from the Northern Europe is the GediFlux collection which stands for "Genetic Diversity in Agriculture: Temporal Flux, Sustainable Productivity and Food Security". These early modern cultivars are composed of two main datasets: 1) The Euro-Recommended List cultivated from 1945 to the 2000s including 282 accessions, and 2), the UK National List (NL) which comprises 197 accessions from the 1990s. The complete collection contains winter wheat accessions from Austria, Belgium, Germany, Denmark, France, Great Britain England, Netherland, and Sweden. Consistently, genetic analysis and population structure of this collection differentiates two main groups, the EU recommended list and the UK national list groups (Wingen et al., 2014).

In parallel with these early breeding cultivars in Europe, wheat breeding experienced a substantial change between the 1960-1970s just after a series of new methods for breeding took place with the shuttle breeding implemented at CIMMYT (Spanish acronym for International Centre for Wheat and Maize Improvement) by Norman Borlaug. This was the "Green Revolution" period when a consecutive set of semi-dwarf wheat varieties were released in Mexico and in the US. First, a variety named Gaines was released in Washington State US by O.A. Vogel and colleagues in 1961. Pitic 62 and Penjamo 62 were released in CIMMYT Mexico in 1962. Sonora 64, Lerma Rojo 64, Super X, and Siete Cerros in 1964. All of these contained one or two genes conferring the dwarfism trait from the Japanese winter variety Norin 10.

The advantage of these new varieties was a 100% increase on yield over previous varieties positively impacting on food production (Borlaug, 1983). The physiological benefit of these genotypes was an increase in tiller number, high grain-filled spikes, and short stems, providing lodging resistance. This "new" plant architecture had a direct impact on the "harvest index", a transformation and allocation of dry matter into the grain, which was facilitated by the assimilation and transformation of the high nitrogen and other fertilizations schemes recently introduced in those periods. Immediately after the release of those early short stem varieties, several wheat breeding programs worldwide introgressed these alleles into their germplasm. At that time Norin 10 was the only source of the dwarf genes which would bring a novel problem of genetic diversity

bottleneck since several other genome regions would be fixed reducing the number of alleles in the region if some genes were linked to the dwarfism genes.

In this context of genetic diversity, breeding for high yield cultivars and fixing favourable alleles has impacted on the genetic diversity of wheat as demonstrated with the use of dwarf genes across different breeding programs. Although this trend of plant improvement is similar to other crops, quantifying the breeding impact on genetic diversity may differ depending on the breeding program, geographic region, and the method employed to evaluate the diversity. For example, cultivars released from 1800 to 2000 in Europe, were documented to have less genetic diversity compared to landraces when using 609 microsatellites markers (Roussel et al., 2004). Considering both types of germplasm the average number of alleles per locus was 14.5. Comparing landraces vs cultivars, the effect was 25% fewer alleles in cultivars compared to landraces. This decrease on genetic diversity was remarkable in varieties released in the 1960s (Roussel et al., 2004).

Contemplating a relatively wide geographically collection of cultivars from Europe released from 1840 to 2000, Roussel et al., (Roussel et al., 2005) found an average of 16.4 alleles per locus. This number of alleles was stable until 1960 followed by a decrease in allele number when including varieties released in the 2000s. Controversial results in other programs indicate that breeding has impacted negatively on genetic diversity at specific periods of time followed by a subsequent period of restoration by an increase on diversity (Reif et al., 2005). A possible explanation may be due to the reintroduction and use of wild relatives or landraces in recent breeding programs to select for disease resistance in elite cultivars.

In summary, most of the early and modern breeding cultivars can be tracked back from landrace accessions. These early cultivars were subsequently improved further during the Green Revolution with the extensive use of semi-dwarfing alleles which further impacted on the genetic diversity of modern cultivars. Most modern cultivars are the result of inter-crossing mainly among elite cultivars. However, with advances in genomics resources which facilitates genome characterization and selection, modern global breeding programs are in constant use of landraces and alien introgressions from wild relatives to select resistant genes and increase genetic diversity.

1.1.5. The WatSeq initiative

The WatSeq stands for the Watkins Sequencing, a project initiative originated between the John Innes Centre, UK led by Dr. Simon Griffiths and the Agricultural Genomics Institute at Shenzhen (AGIS) in China led by Shifeng Sheng in an effort to sequence the entire Watkins landrace collection. The objective was to explore the genetic diversity of the collection and link genotypes

with field phenotypes to exploit and facilitate their use by the wheat community and breeders. The sequenced accessions include 827 landraces from the collection described in section 1.1.3 including the Watkins stabilized group, which is a highly homozygous group. It also includes 218 modern cultivars from the GediFlux collection as described in 1.1.4.

The entire collection has been phenotyped for several agronomically important traits including yield, nutritional value, and disease resistance related traits. Phenotypes of the collection has been collected over >10 years by different research groups across collaborative institutions globally. Seeds of the collection have been distributed across several other institutions worldwide to be characterized for specialized traits including nutritional value of the grain, root morphology, biotic and abiotic stresses, and disease related. Importantly, the phenotypic information collected by JIC is publicly available and it is maintained in the website https://wisplandracepillar.jic.ac.uk/results_resources.htm as part of the BBSRC Designing Future Wheat (DFW) initiative. Users can access to this information for the trait of interest and request the germplasm freely available. Importantly, after publication of the main manuscript of the WatSeq, the WGS information will also be publicly available for the wheat community expanding the genome information repertoire for wheat genomics. This provides an invaluable phenotypic and germplasm information with additional sequencing material for further characterization and exploration of the collection.

In previous studies a series of genotyping analysis of the entire or partial collection were done using SNPs arrays, exome capture, or microsatellites. However, the WatSeq initiative is the first project that involves the complete collection sequencing the whole genome at 12-fold coverage. The objective of the project was to create a whole genome haplotype map including a highly diverse collection of landraces and important modern cultivars. The target sequence coverage was stablished to be of 12-fold short reads (150 bp) using the DNBSEQ technology. The main approach of the project was to follow the routine variant calling pipeline using a genome reference (Chinese Spring reference) followed by variant calling to define haplotypes.

Alongside this haplotype map by conventional methods, the research objective of this thesis is to develop an alternative method identify longe-range haplotypes which will be described step by step in the following chapters. Both methods are importantly complementary identifying unique and overlapping results and advantages and disadvantages that will be discussed across this thesis. Using phenotypes collected by different collaborative groups, genotype-phenotype associations have been identified with both methods, the routine variant calling and our method using GWAS associations studies. Result of this analysis will be discussed in Chapter 3.

1.2. Sequencing and genotyping technologies

1.2.1. Sequencing technologies

Sequencing technologies have evolved considerably in the last two decades (van Dijk et al., 2018). As a result, many tools have been developed to exploit genomic information for crop improvement. SNP arrays, Genotyping by Sequencing (GBS), targeted capture probes, and Whole Genome Sequencing (WGS) are common examples (e.g., Sansaloni *et al.,* 2020; Mascher *et al.,* 2021). Novel chemistry methods for sequencing have been developed impacting positively on sequencing costs thanks to the novelty in chemistry reactions, equipment scale, computer power, and commercial competitions among emerging sequencing companies. For example, after the SANGER sequencing technology, Illumina, a second-generation platform for sequencing short reads (~150 to 250 bp), predominated the market with its high throughput, low cost, and high accuracy sequencing approach compared to other platforms. Analogously, recently, the Beijing Genomics Institute (BGI) developed a chemistry termed DNA nanoball sequencing (DNBSeq) which has a very high throughput for short reads. The drawback or commercial strategy of this technology is that it is only accessible through BGI labs limiting its extended use by users worldwide.

In the market of long reads sequencing, PacBio and Oxford Nanopore have led the market. While Nanopore technology offers much longer sequencing reads (>1 Mbp) than its competitors, it has the disadvantage of having high error rate base calls and lower throughput. Similarly, PacBio offers long reads sequencing and offers higher accuracy than its competitor Nanopore. PacBio recently developed the high-fidelity consensus sequencing technology (HiFi), a breakthrough that has gained popularity in genomics by offering long reads (~10 kb) at very high accuracy (99.9%). As a result, HiFi sequencing is becoming the gold standard particularly for genome assemblies. Furthermore, to increase the long reads sequencing throughput, in 2023 PacBio deployed the Revio system which can generate large amounts of data per unit of time (Baker, 2010; Mardis, 2017; Shendure et al., 2017; van Heyningen, 2019). Across this thesis we employed sequencing data mainly from short read sequencing from BGI (DNBSeq) and Illumina, but we benchmarked results with PacBio long reads HiFi data.

At the time of writing this thesis the progress on sequencing technologies advances considerably and it is possible that our method developed in this research would need to be updated or become obsolete since sequencing and genome assembly would be the routine standard in genomics. Alternatively, it could be that the sequencing of short reads, in which this project is mainly based, will be more affordable in costs and sequencing at 12-fold for other wheat collections will be added to our database increasing the power of our approach to differentiate genome regions

among thousands of individuals from worldwide collections including other crops species or wheat wild relatives.

1.2.2. SNPs, KASP, and MAS

SNPs arrays, initially designed from expressed sequenced tags (ESTs) and complementary DNA (cDNA) are routinely used in some breeding programs for marker assisted selection (MAS) or genome selection (GS). Initial SNPs arrays called variations within the gene coding/UTR regions (Allen et al., 2017; Sun et al., 2020; Wang et al., 2014) and were restricted to variants present in the discovery panel, which used to be a reduced representation of individuals of a crop or species. More recent SNPs arrays have incorporated genome information outside gene regions from WGS data and integrate genomic information from larger collections including wild relatives and landraces for different crops (Sun et al., 2020; Winfield et al., 2016). Using these novel SNPs arrays, breeders and geneticists are able to precisely conduct genetic analysis including population structure analysis and phylogenetic studies for evolution or detect introgressions in modern cultivars. The disadvantage, however, is that still the variations queried are restricted to the SNPs included in the panel array and novel alleles are missed.

An alternative approach to identify novel variants not present in SNP panel arrays at relatively low costs is Genotyping By Sequencing (GBS), which is a particularly useful approach for complex genomes by sequencing a reduced representation of the complete genome (Elshire et al., 2011). The sequencing data, however, is bias towards certain regions in the genome, and does not always capture consistent regions among different samples which complicates their analysis as some real information is cleaned-off during pre-step filter analysis (Lachance & Tishkoff, 2013). Some of these limitations can partially be addressed using different methods of imputation or by combining with other genomic data like the SNPs arrays (Negro et al., 2019). This approach can significantly reduce the genotyping cost and may be affordable for some large-scale breeding programs that need to genotype thousands of samples reducing the cost per sample at large scale. Small to medium genotyping projects breeding programs, however, may not be able to afford these genotyping methods in hundreds of samples yet.

The Kompetitive Allele-Specific PCR (KASP) genotyping method developed by LGC genomics is an approach that relies on SNP variations flexible to genotype a single, a few SNPs, or thousands of variants at low cost. The versatility of this genotyping method is employed by small projects and breeding programs and can be fine-tuned by selecting customized SNPs by the user. This is beneficial when a breeder desires to incorporate a few SNPs associated to major QTLs detected in

their germplasm into the KASP panel for MAS or GS. Furthermore, the method is flexible enough that if a new QTL or gene is cloned, a new KASP assay can be quickly designed and incorporated into the panel. The flexibility of this approach has generated that several previous SSR markers be transformed into KASP markers and are routinely used across multiple breeding programs. Additionally, KASP markers can be easily set it up to run into low throughput labs with basic lab equipment similar to the facilities used for SSRs markers. The difference is that KASP markers are straightforward to use and can differentiate between a single SNP and easily genotype hundreds of samples depending on the lab capacity. A minor disadvantage, however, is that this technology is based mainly on SNPs, which are bi-allelic. A partial solution is that users can design two or more KASP markers in a region inherited together and call multi allelic haplotypes by the combination of two or more SNPs.

1.2.3. Targeting sequencing (capture probes)

Analogous to GBS, captured-based sequencing (exome-promoter capture followed by sequencing) is a genome reduction and sequencing approach that target regions of a genome allowing costs optimization. It is an alternative for genetic variations discovery and enables sequencing of large number of samples at high coverage. The advantage over GBS is its consistency to often capture similar genome regions across multiple genotypes (Gardiner et al., 2019; F. He et al., 2019; Krasileva et al., 2017). First-generation capture probes were designed based on ESTs/cDNA and did not capture non-coding sequences, which are known to be important for agronomic traits (Cao et al., 2021; Chen et al., 2020). More recent probes design integrates outside regions such as promoter sequences (Gardiner et al., 2019; Hammond-Kosack et al., 2021; Zhang et al.). The difficulty of this approach is that it requires previous steps for capture probe design which can be more expensive and laborious than GBS or SNPs arrays. Additionally, if a new genome reference annotation integrates further gene information absent in the capture panel, this would require to be updated each time. Finally, capture probes overlook large portion of the genome and the genetic variants called are bias towards genome references used in their design. Recent advances in pangenome projects of several crops have demonstrated that multiple individuals of a species are required to capture a comprehensive genome information from the species core genes (Bayer et al., 2020; Ebler et al., 2022). Therefore, further capture probes design will be required to be considered using a pangenome reference instead of individual references.

Although the genotypes methods here briefly described are powerful tools in the routine genotyping projects when screening hundreds to thousands of breeding samples, they do not capture the complete genome information to *de novo* investigate large structural variations or

large haplotypes and are bias toward genome regions and genome references. Therefore, to gain a compressive understanding of the genome information of a species, whole genome sequencing would be required (Ahmed et al., 2023; Gaurav et al., 2022). As sequencing costs decrease with the progress on novel technologies, full genome re-sequencing projects are becoming an additional alternative for genotyping large germplasm collections in several important crops including wheat.

1.2.4. Whole Genome Sequencing (WGS)

WGS is a method where the full genome of a sample is sequenced at a specific coverage depth and reads length. Initial WGS projects were restricted mostly for genome assemblies, and they were expensive for large collections of genotypes (Jiao & Schneeberger, 2017; Sohn & Nam, 2016). Whit the progress on NGS technology WGS for genotyping was initially possible for some model crops (Cao et al., 2011) with relatively small genome species (Bukowski et al., 2017; Yano et al., 2016). Applying WGS to large genome crops such as the hexaploid wheat or to thousands of samples, was still expensive ten years ago. In recent years, the number of WGS projects have increased tremendously as a result of technological developments on sequencing platforms, progress on computational power, and novel algorithms and software to efficiently analyse large datasets (Hu et al., 2021; Wenger et al., 2019).

WGS for genotyping projects commonly uses lower coverage than genome-capture sequencing methods (~5 - 15x) and have been on demand for many crops (Alonge et al., 2020; Bayer et al., 2020; Lozano et al., 2021; Wei et al., 2021). This coverage, however, is still expensive for large genome crops such as wheat. To leverage full genome information, some projects have employed WGS at shallow coverage at <1x for costs optimization in sequencing capturing the whole genome information when combined with variants imputation (Adhikari et al., 2022; Franco et al., 2020). This approach can alleviate to capture whole genome information, but it is still expensive to applying WGS to thousands of samples in large populations in routine breeding programs. If the progress on NGS and computational developments progress at similar rates as in recent years, WGS might be the routine method to employ in the near future since it does not require extra steps and it is straightforward to adapt to common variant calling pipelines. It is also possible that WGS in a routine basis will allow for further *de novo* genome assemblies in multiple samples using recent advances in long read sequencing platforms (Wenger et al., 2019).

1.2.5. Whole Genome Assemblies

In the last ten-years, genome assemblies have become common for many organisms including orphan crops, something that was unthinkable 20 year ago. Breakthroughs in sequencing technologies and chemistry in the last five years has brought genome assemblies even further to the level where is its common to generate genome assemblies in a relatively short period of time and for multiple samples of the same specie (pangenomes). These capabilities bring novel opportunities for easily and cheaply generate chromosome-scale assemblies allowing to answer novel questions in genome structure and to have a broad genome reference representation of an organism. In this context, examples in cereals are the pangenome assemblies of 15 hexaploid wheat genomes (Walkowiak et al., 2020), 20 barley cultivars (Jayakodi et al., 2020), and 26 Maize elite lines (Hufford et al., 2021). More recently, as discussed in section 1.2.1, long read sequencing HiFi with high accuracy has allowed to deploy even further genome assemblies a lower cost and time (Amarasinghe et al., 2020; Cheng et al., 2021) than short reads. It is predicted that in the coming years there will be a surge in the number of assemblies generated for genotypes of multiple species (Rhie et al., 2021) and including large genomes such as wheat.

These new capabilities, however, brings novel challenges on data manipulation or comparisons among different studies. For example, it was a routine approach to compare genetic variants using a common single genome reference across different projects. Now with multiple references, users can run genome alignments against multiple references for comparison (Armstrong et al., 2020), or to select a single reference that would suit for their analysis. Although this may not be a big problem in small-genome reference species, running alignments for hundreds or thousands of samples to multiple references, individually, using current aligners algorithms for large genome references (e.g., 16 GB genome in hexaploid wheat), represents a challenge for computing resources and time. Additionally, to take advantage of multiple genomes assemblies information, users will need to align to each of the references or identify variations using a genome reference that represents better their genotypes since most genome aligners use a single reference.

Alternative ways to analyse the growing number of public sequence databases such as the constant addition of genome assemblies, is the use of novel algorithms employing multiple references to directly identify variations in a genome of the size of wheat. Proposals to approach this challenge are to call variants from a unified genome graph, which condenses pangenome information from multiple genome individuals in a graph-based genome (Bradbury et al., 2022; Liu et al., 2023; Wang et al., 2022). Depending on the project objective, is also possible to directly compare genome assembly vs genome assembly in a pairwise manner (Brinton et al., 2020), but this is more expensive and represent bias towards the quality of the assemblies since some

information may be lost during assembly steps. An alternative method to detect genetic variations with or without genome references among individuals is the use of *k*-mers, a sub-sequence of a sequence of the *n* size in bp (Gaurav et al., 2022; Voichek & Weigel, 2020).

1.2.6. *k*-mer methods

Routinely, genetic variants are identified after sequence alignment of raw reads to a genome reference. However, variant calling by aligning to a single reference can be problematic for many reasons; (i) when a genome reference is large, polyploid, repetitive, or incomplete, important regions can be missed, (ii) highly divergent individuals to the reference can produce poor alignments, (iii) accurate variant calling requires sequencing depth ~>5x to distinguish between sequencing errors and real variants, and finally (iv), large structural variants (SV) are difficult to detect, particularly with short sequencing reads (Saxena et al., 2014).

Alternatively, variants between two genotypes can be detected by reference-free alignment methods. For example, a direct raw reads sequence comparison has been developed to identify genetic variations that can detect SV based on sequence breakpoint patterns between samples (Shimmura et al., 2020). A second strategy is the use of *k*-mers (Zielezinski et al., 2019). *k*-mers are sub-sequences of a sequence of *k* length. *k*-mers can be represented as a presence/absence or numerically to identify natural genetic variants or mutations for association mapping. For example, the NIKS (needle in the *k*-stack) algorithm compares WGS *k*-mers between mutated and non-mutated bulked samples to identify variants (Nordström et al., 2013). Kestrel, quantifies the *k*-mer distribution to characterize highly polymorphic regions and structural variations (Audano et al., 2017). HAWK; hitting associations with *k*-mers (Rahman et al., 2018) and AgRenSeq; association genetics (Arora et al., 2019) conduct GWAS analysis by presence-absence of *k*-mers followed by mapping or local assembly of the associated *k*-mers to identify functional sequence regions. LNISKS (longer needle in a scanter *k*-stack) developed based on NIKS (Suchecki et al., 2019) was implemented to identify EMS mutants in the wheat genome. And finally, GWAS using *k*-mers from WGS data for large genomes have been implemented (Gaurav et al., 2022; Voichek & Weigel, 2020). One of the limitations of working with *k*-mers, WGS, and large genomes, is computational burden, but improvements in this area has been one of the main focus of researchers (Denti et al., 2019; Mehrab et al., 2021; Pajuste et al., 2017; Standage et al., 2019).

Regardless of the genotyping approach, in many cases, the goal of variants calling and genotyping, is to associate genotype with phenotypes, understand genome dynamics of an organisms, or genome population studies. In recent years genotype-phenotype association studies are common

using haplotypes instead of individual SNPs. To our knowledge, at the time of writing this thesis there is not a method to define whole genome haplotypes using *k*-mers.

1.3. Haplotype-based selection for breeding

1.3.1. Defining Haplotypes

The use of SNP markers is the most common genotyping approach coupled with phenotypic information for QTL detection, GWAS, or GS (Sansaloni et al., 2020). Regardless of the genotyping technology, most of the current methods employ individual SNP markers for genotype-phenotype associations studies. One disadvantage is that individual SNP markers are bi-allelic but having thousands of SNPs allows to combine two or more into a haplotype. A haplotype is defined as two or more genetic polymorphisms clustered as a locus under linkage disequilibrium (LD) inherited together with limited chances of recombination (Patil et al., 2001). A haplotype can be set arbitrarily allocating all the SNPs that present within a defined fixed genome region, sliding window, (Guo et al., 2009), number of continuous SNPs, or based on LD (Gabriel et al., 2002; Kim et al., 2018; Pook et al., 2019).

Depending on the method employed to build haplotypes some advantages or disadvantages might be encountered. The most common approach is the use of LD after SNP calls. This method is reasonable straightforward and can handle large number of genotypes in small to medium size genomes or with low density SNP arrays. It represents a challenge however, for large genomes, dense SNPs markers, and large number of genotypes in a population (Bhat et al., 2021). A disadvantage if the genotyping information is not uniform across the genome can result in misclassified and extended haplotypes. For example, haplotypes in the D wheat sub genome based on SNPs usually detect longer haplotype blocks than in chromosome A and B (Brinton et al., 2020). Similarly, haplotypes in centromere regions tend to be larger than in telomere regions (Balfourier et al., 2019). Although, these haplotypes reflect the real low recombination rates and real haplotype blocks, the lack of SNPs representing centromere regions in SNPs arrays or in capture probes designed based on gene content, have an impact on the size and number of haplotypes identified (Brinton et al., 2020).

A second method commonly used is the arbitrarily fixed genome window size. In this case a specific range size in bp is defined and the number of markers within the window are counted as a haplotype (Brinton et al., 2020; Huang et al., 2007). Using this approach uniform haplotypes are defined across the genome including centromere regions. This approach is straightforward for computing load and feasible for large genome and high number of samples. The difficulty,

however, is that additional work for window size optimization needs to be validated in previous steps. Additionally, using this approach real haplotypes can be split resulting in the loss of resolution in haplotype-phenotype associations analysis, particularly in telomere regions where haplotypes are often shorter than in centromere regions caused by high recombination rates. A partial solution for the former at the expenses of time and computer resources, is the use of sliding window sizes (Bhat et al., 2021; Huang et al., 2007).

A similar approach to the arbitrarily defined window size, is the arbitrary use continuous number of SNPs. This approach can be more straight forward to implement but like the LD approach, can result in large haplotype blocks if not uniform SNP information is captured across the genome. Similar to the defined window size approach, counting a defined number of SNPs for haplotypes can result in the split of real haplotypes on telomere regions (Meuwissen et al., 2014).

Novel methods to define haplotypes have been developed with advances in technological improvements at genome-wide scale (Cheng et al., 2021; Garg et al., 2021; Sinha et al., 2020), and with RNA-seq data (Berger et al., 2020). Recent methods employing multiple references or multiple genotypes as "reference graphs" to define haplotypes from consensus genomes instead of individual assemblies have been developed (Rakocevic et al., 2019; Shang et al., 2022). This has been possible with the release of multiple genome assemblies in many important crops (Bayer et al., 2020).

Despite its size (~16 Gb), wheat has not been the exception, and in 2020 the wheat pangenome project made it possible to develop whole genome haplotypes from 15 important cultivars (Brinton et al., 2020; Walkowiak et al., 2020). Haplotypes were defined by using whole genome assemblies and pairwise alignments among cultivars. To define haplotypes a set of parameters were adjusted including the window size and multiple pre-filtering steps. The median sequence identity above 99.99% in a 5 Mbp window based on genome alignments was considered as identical-by-state (IBS) region or haplotype between two cultivars. This stringency was flexible enough to account for single nucleotides gaps (Ns') and sequence errors in the assemblies. Near-IBS regions were detected as having sequence similarity of ~99.95% (1 SNP in 10 Kbp) while sequence similarity < 99.5 % were predicted to be introgressions from wild or close wheat relatives.

An important discovery of the Brinton et al., 2020 study found that gene based genomic information was not sufficient to differentiate between IBS and near-IBS regions among cultivars. Adding flanking (2 Kbp flanking sequence) regions information to the coding sequences (CDS) improved the detection of real haplotypes. Using 5 Mbp window size the median haplotype size for those 15 wheat cultivars was 9.34 Mbp and 196 genes per block. Haplotypes were 15.43 Mbp

on average in centromere regions and these extended blocks contained hundreds of genes. This information is of important for breeding since centromeres usually may remain fixed in breeding populations due to low recombination in these regions.

Analysis of this multi-reference project revealed that worldwide wheat cultivars share >60% of their genomes with at least one of the pangenome genotypes (Brinton et al., 2020). In the same study, large "haploblocks" stretching thousands of bp between pairwise comparison were detected in genotypes that were both related and not directly related by pedigree (Fig. 1.2) reflecting the limited genetic diversity exploited in modern global germplasm for breeding. Furthermore, Brinton et al., 2020., in a case study, demonstrated that haplotype informed analysis identified novel haplotypes which were unable to be detected from previous SNPs arrays and capture-probes approaches. A set of novel haplotypes present exclusively in the Watkins landrace collection evidenced the importance of including outside gene information and haplotypes instead of individual SNPs for breeding information. This is of importance since usually genotype-phenotype associations are commonly carried out using individual SNPs. However, as mentioned before, in recent years there is a recent interest to develop haplotype-based associations in genomics studies instead of individual SNPs (Bevan et al., 2017; Bhat et al., 2021; Brinton et al., 2020; Jordan et al., 2021; Mayer et al., 2020).

Fig. 1. 2. From Brinton et al., 2020. Haplotypes across the "highly conserved" region of chromosome 6A.

"**a** Physical position of productivity-related QTL (rectangles) and GWAS hits (triangles) mapped to the highly conserved region on chromosome 6A (see "Methods"). *; grain-size mapping interval based on UK cultivars Spark and Rialto. **b** Diagrammatic representation of all haplotype blocks on chromosome 6A in the 15 sequenced cultivars (based on 5-Mbp bin haplotypes; scaled to the longest chromosome 6A). Regions with the same colour at the same position share common haplotypes (except for white regions which are not contained within haplotype blocks). Vertical grey line indicates the position of TaGW2-A (237 Mbp). Labels H1–H7 indicate haplotype groups based on the minimum haplotype block (beige bar; 187–445 Mbp)".

1.3.2. Genotype-phenotype associations by haplotypes

Linking genomes with phenotypes is one of the main objectives in genome studies. Followed by variant calling, a common step is to associate genotypes with phenotypes and in some cases determine alleles functions. This can be achieved by QTL mapping, GS, GWAS, or Isogenic Lines (IL) with natural or induced variations (Adamski et al., 2020; Aglawe et al., 2021; Arora et al., 2019). A constrain in genotype-phenotype associations analysis is that molecular markers rarely are the causative of a phenotypic change or explain low phenotypic variation of the trait under study. Instead, those markers are in LD in proximity with the functional variant. The low marker-phenotype effect can be attributed to the low LD between the causal polymorphism and the SNP

used in the genetic analysis, epistatic interactions, multiple genes with minor effect, or environmental interactions (Bevan et al., 2017; N'Diaye et al., 2018; Sallam et al., 2020). Therefore, a single SNP often do not capture the complete phenotypic effect (Kearsey & Farquhar, 1998; Kumar et al., 2017). On the contrary, haplotypes combine multiple variants in a genome region which can help to capture the causal polymorphism as multiple SNP combinations in a locus can be tested for phenotype associations. This can result in greater trait variation effect explained by haplotypes compared to the single marker approach (Bevan et al., 2017; N'Diaye et al., 2018; Sallam et al., 2020).

An additional advantage of haplotypes over individual SNPs is on MAS in breeding programs. For example, a single marker during MAS can be no longer informative if a breeder incorporates new germplasm having a SNP in the same locus position of a trait informative marker but this novel SNP source is not associated with the beneficial allele effect. On the contrary, a combination of two or more SNPs into a haplotype have a better probability to capture the true beneficial allele since multiple combinations of SNPs are less likely to occur by chance in new germplasm sources not associated with the beneficial allele (Hasan et al., 2021). Similarly, using haplotype would help to capture epistatic interactions between loci resulting in better genome trait association predictions in GS (Bevan et al., 2017; Meuwissen et al., 2014; Sehgal et al., 2020; Voss-Fels et al., 2019).

1.4. Wheat introgressions and haplotypes phenotypic value

Wild wheat relatives are invaluable reservoirs of genetic diversity for agronomically important traits, particularly for disease resistant genes (Leigh et al., 2022). Genetic material from more than fifty species have been successfully introgressed into wheat and several wild wheat relatives hybridize well with tetraploid and hexaploid genomes (Wulff & Moscou, 2014). Tetraploid wheat is considered as the primary gene pool for hexaploid wheat since crosses between these two species are easily carried out. On the other hand, secondary and tertiary gene pools infrequently hybridize naturally with modern hexaploidy wheat and are more commonly used by breeders and geneticist to transfer genetic diversity into cultivated elite varieties by specialized methods (Hao et al., 2020). Despite these barriers rare hybridizations between tertiary gene pools and hexaploid wheat naturally exist and contribute to its genetic diversity.

Historically, wild wheat relatives have been used in different breeding programs worldwide since the 1900's (Doussinault et al., 1983) to introduce genetic diversity. With the advent of novel genome technologies, the use of these wild relatives can be employed more efficiently by tracking

introgression into chromosome physical regions in the wheat genome precisely. As a result, there is an increase interest to exploit and integrate novel genetic diversity from introgressions and large collections of novel synthetic wheats between wild relatives and modern wheats across breeding programs (Devi et al., 2019).

Large number of hexaploid wheats have been shown to harbour introgressions from tetraploid wheats (Przewieslik-Allen et al., 2021) particularly *T. timopheevii* and *Ae. ventricosa*. It has been hypothesized that the tetraploid wheat *T. timopheevii* (AAGG) was originated from a second independent hybridization event from the same progenitors of *T. turgidum* and *T. aestivum*, *T. urartu* (AA) and *Ae. speltoides* (Feldman, 1966). Cytological and genomic studies have demonstrated that the A genome of *T. timopheevii* recombines more frequently with the wheat A subgenome than the G genome with the B subgenome (King et al., 2022). Therefore, the gene pool of *T. timopheevii* has served as a donor for several agronomically important traits into wheat such as disease resistance. As a result, hybrids between *T. timopheevii* and *T. aestivum* are frequently used to introgress novel genetic variation via homoeologous recombination but natural hybridizations also occur (Brown-Guedira et al., 2003; Chemayek et al., 2017; Järve et al., 2000).

A second important and frequently wild relative used in wide crosses against wheat is the tetraploid *Ae. ventricosa* (2n = 4x = 28, NNDD). Early breeding programs have employed *Ae. ventricosa* to introduce segments into wheat cultivars to exploit mainly disease resistant traits. An example is the famous cultivar named VPM1 (Doussinault et al., 1983). A widely exploited introgression from this cultivar, is the 2AS/2NvS translocation on chr2A of wheat involved in several disease resistant and yield related traits (Xue et al., 2018). A second example from *Ae. ventricosa* is the α-amylase gene introgressed into chr7D conferring resistance to eyespot (*Oculimacula acuformis* and *O. yallundae*) (Gale et al., 1984). These introgressions have been widely exploited in modern breeding and it is hypothesized to be present in several other important cultivars worldwide still not documented (Cruz et al., 2016; Przewieslik-Allen et al., 2021) and probably involved in other agronomically important traits (D. Singh et al., 2019).

In Chapter 4 of this thesis, we extended the analysis of the contribution into the wheat genome of these two important wild wheat relative species and demonstrated that there are still hidden introgressions present in several important modern cultivars and germplasm bank collections unexploited.

1.5. General Aim

The main objective of this project was to define haplotypes to build a haplotype database of wheat to elucidate the diversity between cultivars and landraces. We focused on **1)** developing a method to build a haplotype database, **2)** explore the diversity between landraces, early cultivars (1900's) and modern varieties (after 2000's), and **3)** detect genome-wide wild wheat introgressions/hybridizations and their impact on shaping the genomes of landraces and modern cultivars. We hope that the resources here generated will contribute to the wheat community for a more targeted and genome-based breeding approach.

2. Alignment and *k*-mer methods to identify variations.

In this chapter we describe our bioinformatic approach named Identity by State in python (IBSpy). We thank Dr. Kumar Gaurav for his contribution on how to employ *k*-mers to detect variations in genome assembly comparisons. Initial scrips of IBSpy were done based on this method following his algorithm to capture variations between genome assemblies. We also thank Dr. Brande Wulff for his contribution and feedback on initial results during early stages of IBSpy participating as thesis second supervisor.

After adjusting different parameters of the pilot scripts, the final version of IBSpy was written by Dr. Ricardo Ramirez-Gonzales and was uploaded in the public repository of the Uauy Lab with additional scores: "*observed_kmers*" and "*kmer_distance*". Contributors of IBSpy are Luca Venturini and Luis Yanes as described in https://github.com/Uauy-Lab/IBSpy. Cong Feng from Shifeng Cheng's group wrote the IBScpp version in the C++ language. https://github.com/Uauy-Lab/IBSpy/tree/main/IBScpp. We thank Dr. Simon Griffiths and Shifeng Cheng for providing us early access to the WatSeq dataset raw sequences. This data was pivotal for the pilot tests to validate IBSpy at a large-scale using genotypes with different levels of relatedness apart of the publicly available pangenome cultivars. The main manuscript entitled "***Harnessing Landrace Diversity Empowers Wheat Breeding for Climate Resilience***" for the WatSeq data has been submitted for publication (Cheng et al., under revision). A portal for this public dataset for further exploration is on https://wwwg2b.com/.

Analysis to translate the percentage of sequence identity from whole genome assembly alignments to IBSpy *variations* for the B and D sub genome was analysed by Dr. Xiaoming Wang during his scientific visit to the Uauy's Lab in 2022-2023.

2.1. Chapter summary

IBSpy is a method to directly detect whole genome variations between a genome reference and raw reads from a query sample using *k*-mers. We demonstrate that this approach benchmarks well against previously established methods using whole genome alignments of eleven genome assemblies of the wheat pangenome. Using IBSpy, we detected regions which are identical by state (>99.99% similarity based on sequence alignment) as having approximately <10 IBSpy *variations* in consecutive 50 Kbp windows. We validated our method to combine different sequencing platform, scaffolds, or genome level assemblies, using raw reads of >150 bp length. The optimal *k*-mer size was defined to range between *k*=25-mer to *k*=51-mer, and we decided to use *k*=31-mer to leverage the already available *k*-mer databases and to account for computational load. We

established an optimal sequencing coverage of raw data to be ~12-fold for 150-bp reads in wheat when using 31-mers and when removing unique *k*-mers. Above this coverage, minor improvements are obtained, however, if enough coverage is provided, our method can detect variations at the same resolution as with full genome assemblies and detect genome misassemblies. We demonstrated that with long read sequencing, less coverage is needed as seen in genome assembly methods. IBSpy condenses multiple types of sequences and structural variations into a single type allowing them to be integrated in downstream analyses. We acknowledge that our approach does not discriminates a few or a single SNP in 50 Kbp windows and we provide key points for further improvements in resolution. The overlooking of these few SNPs in a window can be used as a feature to detect long-range haplotypes and fine tune variations among samples sharing the long-range haplotypes having different phenotypes to narrow down causal variations within a haplotype (we provide a case study in Chapter 3).

2.2. Introduction

2.2.1. Alignment methods and variant calling

In the past five years, chromosome-scale genome assemblies of multiple representative accessions of important crops have been made available (Hufford et al., 2021; Jayakodi et al., 2020; Walkowiak et al., 2020). Studies on whole genome sequencing (WGS) from collections (i.e., not assembled) with multiple accessions in different crops and wild relatives are also becoming increasingly common (Gaurav et al., 2022; W. Wang et al., 2018; Zhou et al., 2015). These collections include species with complex genomes such as wheat (*T. aestivum*), its wild relatives, and other complex crop genomes (Peng et al., 2022; Zhao et al., 2022; Zhou et al., 2020). Generating this volume of data, it is having great impact in genomic analyses and genetic studies in different areas, however, the amount of data and how to analyse it represent new challenges for computing resources and software development.

Having high quality chromosome-level assemblies is of importance particularly for allopolyploid species such as wheat, which has the ABD subgenomes sharing ~98% of sequence similarity among them within the coding sequence (Ramírez-González et al., 2018). High quality assemblies can alleviate some of the pitfalls of alignments methods to call variations in these repetitive and complex genomes. However, the main limitations to detect variations using alignment-based calling are: 1) many genome regions can be missed if the genome reference is not representative of the species, 2) if the genotype being tested diverged from the reference genome it can lead to poor alignments, 3) structural variations (insertion and deletions (InDels), copy number variation,

and genome inversions) are commonly not identified with the alignment based methods, and 4) aligning reads against a genome reference to identify variations is computing demanding for large and complex genomes such as the hexaploid wheat (16 Gb).

In the multiple genome-reference era, researchers can now decide to select a specific genome assembly to use as a reference, or to align reads against multiple references. However, aligning samples to multiple references can be challenging in species with large genomes such as wheat. Furthermore, comparisons and reproducibility among multiple studies will become complicated as researchers will need to align to different references. To face these challenges, novel methods have been developed that involve aligning raw reads to a haplotype graph genome representation to call and identify variations (Bradbury et al., 2022; Shang et al., 2022). These haplotype graphs require the availability of multiple genome assemblies of a species.

Methods to call variation without genome references using *k*-mers are also becoming more common (Arora et al., 2019; Gaurav et al., 2022; Rahman et al., 2018; Voichek & Weigel, 2020). In addition to not relying on a single genome reference, *k*-mers can detect multiple types of variations and compare individuals from highly divergent samples. The disadvantage, however, is that some *k*-mer pipelines are still not well adapted into the routine variant calling software. Instead, analysis is performed by implementing in-house scripts requiring large computational infrastructure. More developed *k*-mer based software are less straightforward to use than alignment-based methods or output files are usually not compatible with the routine bioinformatic tools.

Additionally, variation discovery can be challenging if the species to study is polyploid, highly repetitive, or heterozygous since sequencing similarities can be a confounding factor since reads often align to more than one region. In genomic studies, a *k*-mer analysis can be used to detect the repetitiveness of the genome assembly, estimate the genome size, sequencing coverage, raw reads heterogenicity and quality (Pflug et al., 2020). In genome assemblies, most genome assemblers use a pre step to define the optimal *k*-mer size to use before generating the assembly. This because the repetitiveness of a genome affects the uniqueness *k*-mers at specific size. Longer *k*-mers are required if a genome is highly repetitive as they have less chances to occur across the genome than short *k*-mers. There is a trade-off, however, since longer *k*-mers require more computing load. Furthermore, longer *k*-mers are prone to capture sequencing errors at reads edges affecting the quality of the final assembly.

Most of these methods to call variations or to generate genome assemblies are influenced by the type of sequencing used to genotype, or in cases of reference-dependent variant calling, by the quality of the assemblies, read lengths, and depth coverage. Fine tuning of these variables to

define the optimal combinations will be dependent on the organisms being studied and the objective of the project. For example, read length for alignments and variant calling are less influenced by the read length and are more dependent on reads quality depth. On the other hand, long (>10 Kbp) reads with lower sequence depth than short reads (~150 bp) are suitable to extend the contiguity in genome assembly projects (De Coster et al., 2021).

A common procedure in alignments methods is to perform a pre-step to verify the reads quality, measure the coverage, and to clean reads-off with high sequencing error rate. In genome assemblies and *k*-mer analysis a critical pre-filtering step is to remove unique *k*-mers as they commonly originate from sequencing errors. However, sequencing data often yields non-uniform coverage of the genome with *k*-mers displaying a Poisson distribution. Therefore, if the sequencing coverage is relatively low, removing unique *k*-mers can be detrimental since some real genome information is still represented as unique *k*-mers (Lee et al., 2020).

In this chapter we implemented a novel approach to call variations and used it to build a *variations* database of >1,000 genotypes based on *k*-mers from raw reads based on multiple chromosome-scale references. Our approach allowed us to integrate genome information from wild wheat relatives, landraces, and modern wheat varieties into the databases and unify multiple types of genome variations.

The general workflow of this chapter is depicted in **Fig. 2.1.** In brief, we first evaluated the *k*-mer profiles of the 15 wheat pangenome assemblies (both chromosome and scaffold assemblies) and from raw reads of the Watkins Sequencing (WatSeq) dataset which includes >1,000 accessions. Across the analysis we integrated different types of sequencing data. We next developed IBSpy (Identity by state in python) which uses the *k*-mer databases described and a genome reference to generate three types of scores based on *k*-mer presence/absence. Using IBSpy we ran pairwise comparisons among chromosome-scale assemblies and among the wheat subgenomes (A, B, and D). Using the Brinton et al., 2020 chromosome-scale alignments, we translated sequence similarity to IBSpy scores. We next validated IBSpy to use raw reads to detect variations and studied the effect of different raw read types, depth, and *k*-mer sizes on the *variations* count.

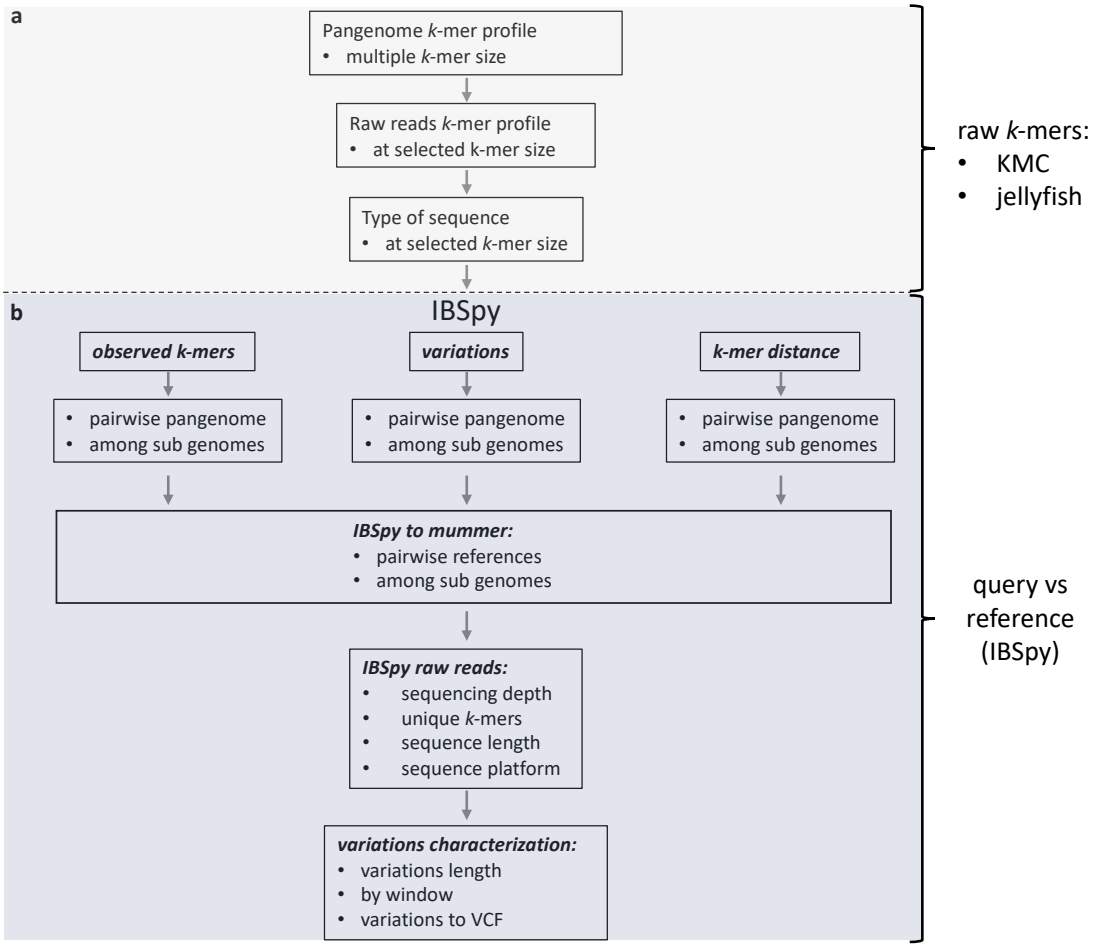


Fig. 2. 1 Chapter 2 workflow.

In an initial step **a**) the query *k*-mer databases are created either from genome assemblies or from raw reads using KMC or Jellyfish software. These *k*-mers are used to characterize the pangenome genome profiles, raw reads, and the different types of raw reads used in this study. These *k*-mer database are also the input of IBSpy as a query. **b**) characterization of IBSpy scores using either genome assemblies or raw reads of different types.

2.3. Methods

In all cases, the scripts used are in the following link with the scrip name of the analysis within the folder "scripts": https://github.com/quirozcj/PhD_thesis_JQCH_2022.

2.3.1. Germplasm & Sequencing data

In this analysis we included 11 chromosome and 5 scaffold level assemblies from (Walkowiak et al., 2020). During the PhD project, additional chromosome assemblies with different qualities have been released. In total, across this project we employed 20 assemblies for different analysis (**Table 2.1**). This research relies mainly on the Watkins Sequencing (WatSeq) project composed of 218

modern cultivars and 827 landraces from the Watkins collection (Wingen et al., 2014). The sequencing coverage ranged from 12 to 15-fold DNBSeq 150 bp reads (Supplemental Table S2.1).

Table 2. 1. Genome assemblies used in this study.

ID*	Line	Assembly Type	Growth Habit	Origin	Use	Analysis	Publication
mace	Mace	chromosome-scale	Spring	Australia	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
lancer	LongReach Lancer	chromosome-scale	Spring	Australia	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
stanley	CDC Stanley	chromosome-scale	Spring	Canada	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
landmark	CDC Landmark	chromosome-scale	Spring	Canada	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
julius	Julius	chromosome-scale	Winter	Germany	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
norin61	Norin 61	chromosome-scale	Facultative Spring	Japan	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
arinaLrFor	Arina*LrFor*	chromosome-scale	Winter	Switzerland	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
spelta	PI190962 (spelt wheat)	chromosome-scale	Winter	Central Europe	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
jagger	Jagger	chromosome-scale	Winter	USA	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
sy_mattis	SY Mattis	chromosome-scale	Winter	France	reference and query	Synteny windows	https://doi.org/10.1038/s41586-020-2961-x
chinese	Chinese Spring	chromosome-scale	Spring	IWGSC	reference and query	Synteny windows	DOI: 10.1126/science.aar7191
cadenza	Cadenza	Scaffold	Facultative Spring	UK	query	query	https://doi.org/10.1038/s41586-020-2961-x
paragon	Paragon	Scaffold	Spring	UK	query	query	https://doi.org/10.1038/s41586-020-2961-x
robigus	Robigus	Scaffold	Winter	UK	query	query	https://doi.org/10.1038/s41586-020-2961-x
claire	Claire	Scaffold	Winter	UK	query	query	https://doi.org/10.1038/s41586-020-2961-x
weebil	Weebill 1	Scaffold	Spring	CIMMYT	query	query	https://doi.org/10.1038/s41586-020-2961-x
tibetan	Zang1817	chromosome-scale	Spring	Tibetan	query	introgressions	https://doi.org/10.1038/s41467-020-18738-5
renan	renan	chromosome-scale	Winter	France	reference and query	introgressions	https://doi.org/10.1093/gigascience/giac034
borlaug	borlaug	chromosome-scale	Spring	CIMMYT	reference and query	introgressions	
kariega	kariega	chromosome-scale	Spring		reference and query	reads length	https://doi.org/10.1038/s41588-022-01022-1

ID* indicates names used for each reference in this study. The "Use" indicates if the genotype was used as a reference, query, or both. The 'Analysis" column indicates a specific analysis for specific assemblies.

In addition to wheat genotypes, in in this thesis we leveraged the publicly available data of 265 *Ae. tauschii* (D genome progenitor of hexaploid wheat) accessions (Gaurav et al., 2022) (Supplemental Table S2.2) to explore the D genome diversity. During the development of this project, we integrated 218 additional accessions to detect introgressions from *T. monococcum* (A^mA^m genome) into wheat and two chromosome scale assemblies of one domesticated and one wild *T. monoccocum* accession (Ahmed et al., 2023) (Supplemental Table S2.3). In-depth analysis of wild wheat relatives will be addressed in Chapter 4.

We also included publicly available datasets of wild relatives from different publications (Walkowiak et al., 2020), including one rye (*Secale cereale*) accession Lo7 (Rabanus-Wallace et al., 2021) (Supplemental Table S2.4). The quality of the raw read sequences was determined using fastqc (v.0.11.8). Samples of the WatSeq project were processed by collaborators AGIS. (Scripts: https://github.com/quirozcj/PhD_thesis_JQCH_2022).

2.3.2. *k*-mer variant calling pipeline.

We used jellyfish v.2.2.6 (Marçais & Kingsford, 2011) or KMC v3.0.1 (Kokot et al., 2017) to create the *k*-mer databases from individual samples. For genome assemblies, we kept unique *k*-mers and screened for differences on *k*-mer sizes (Supplemental Table S2.5). For the WatSeq raw reads samples, we removed unique *k*-mers to reduce computational burden and leverage on the already available datasets of 31-mers. Only when the sequence depth was <10-fold with read length 150 bp, we kept unique *k*-mers to compensate for the lack of coverage. We used Python3 scripts to plot histograms for the *k*-mer counts to verify samples coverage and sequences quality profiles (Supplemental Table S2.6). Scripts: https://github.com/quirozcj/PhD_thesis_JQCH_2022.

2.3.3. Code for Identity by State in python (IBSpy)

The code for IBSpy is publicly available in GitHub (https://github.com/Uauy-Lab/IBSpy) and was co-developed with Dr. Ricardo Ramirez-Gonzalez.

2.3.4. Alignments to IBSpy *variations*

To translate the IBSpy *variations* equivalence to alignments sequence similarity, we compared the published (Brinton et al., 2020) pairwise MUMmer alignments among ten chromosome-scale pangenome cultivars (ArinaLrFor, Chinese Spring (CS), Jagger, Julius, Lancer, Landmark, Mace, Norin61, Stanley, Mattis) with the corresponding *variations* counts from IBSpy outputs. In total, there were 90 pairwise alignments analysed across the A, B, and D genomes. We analysed the data in 500 Kbp windows and kept those windows with at least 60% breadth of alignment in the MUMmer output (77.8%). For each 500 Kbp window, we had the average sequence identity between the pangenome reference and the other nine pangenome query samples (if over 60% breadth of alignment) alongside the IBSpy *variations* for the equivalent comparisons using the pangenome reference assembly and the *k*-mer database. The over 60% bread alignment was selected since eventually we had alignments that did not cover significant region of the 500 Kbp window. For example, we had 500 Kbp windows that had alignments covering only 10% of the window and those would not be informative since all the other 90% would be highly different in sequence identity and therefore no alignments were possible. In addition to the 500 Kbp window, we tested 100 Kbp which tended to include overlapping alignments longer than 100 Kbp. We also tested 1,000 kbp windows, however, in this case alignments were often too short and most of the window was often not covered. Therefore, we decide to use 500 Kbp which were overall well covered with mummer alignments lengths.

2.4. Results

2.4.1. The wheat *k*-mer landscape

2.4.1.1. Pangenome *k*-mer distribution by size

When working with *k*-mers, the genome size, complexity and ploidy of the species, the quality of the assembly (pseudomolecules or scaffold length), and the error rate of the reads can all impact and determine the optimal *k*-mer size. An optimal *k*-mer size also depends on the type of analysis or question to address. In this analysis we investigated the *k*-mer profiles of the wheat assemblies using different *k*-mer size with the aim to efficiently detect variations and differentiate among multiple genomes and raw reads data. For our purposes, the optimal size would be a high percentage of unique *k*-mers in the genome, across multiple genome references considering a trade-off with computer burden. We used the eleven chromosome-scale assemblies to find the *k*-mer distribution by size using a range of *k*-mers from *k*=15 to *k*=101-mer. Historically, the genome reference of Chinese Spring (CS) was the first high-quality assembly (and annotation) at chromosome-scale level, therefore we explored its *k*-mer profiles first.

We observed that for the CS reference at 21-mer size, 25% of the genome is represented as a unique *k*-mers. Five and ten cumulative *k*-mers capture 43% and 50%, respectively, of the genome content. With 31-mers ~37% of the genome is represented as a unique *k*-mers. 101-mers captured ~80% of the genome as a unique *k*-mers and ~90% at five or ten cumulative (**Fig. 2.2).** *k*-mer abundance of the remaining chromosome-scale pangenome assemblies demonstrated that overall, independently of the genotype, they have similar *k*-mer abundance profiles (**Fig. 2.3**). This similarity on *k*-mer profiles may be because a similar pipeline or similar sequencing method were used to assemble them. Comparisons with genome assemblies using different procedures and sequencing reads (e.g., PacBio HiFi or Nanopore long reads) would be required to confirm this. Regarding the *k*-mer size, as expected, 101-mers have the highest unique *k*-mers representation in all genomes with ~80%.

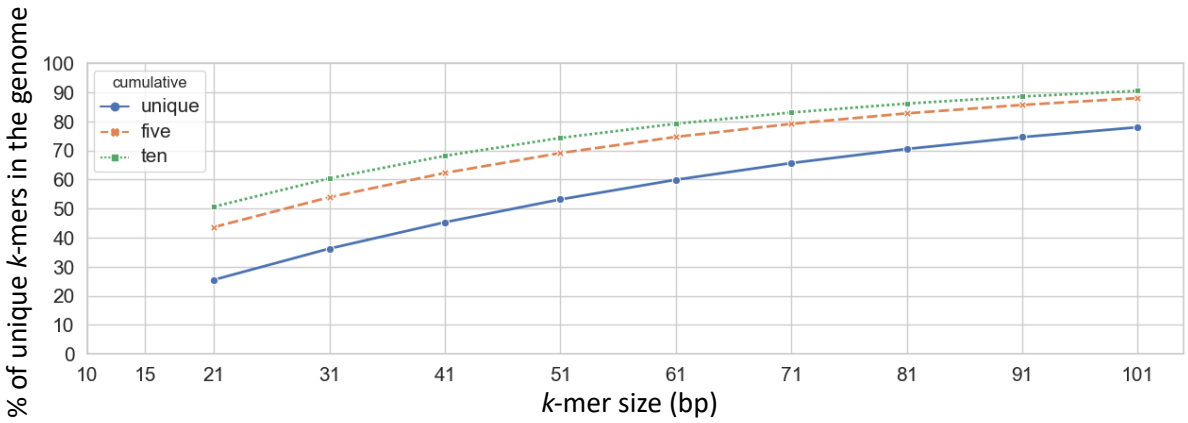


Fig. 2. 2. Genome representation at different *k*-mer size in Chinese Spring(RefSeq.v1.0) reference.

The graph represents the percentage (%) of the genome of CS that is represented as a unique (blue) *k*-mer. For example, ~35% of the CS genome is represented in 31-mers as a unique *k*-mer (no other sequence in the genome has those *k*-mers). The orange line indicates the cumulative five which means that a particular percentage in *y*-axis of the genome of CS that is represented in five or less *k*-mers at given *k*-mer size. For example, using 31-mers ~55% of the genome is represented five times or less. Similarly for the ten cumulative (green line).

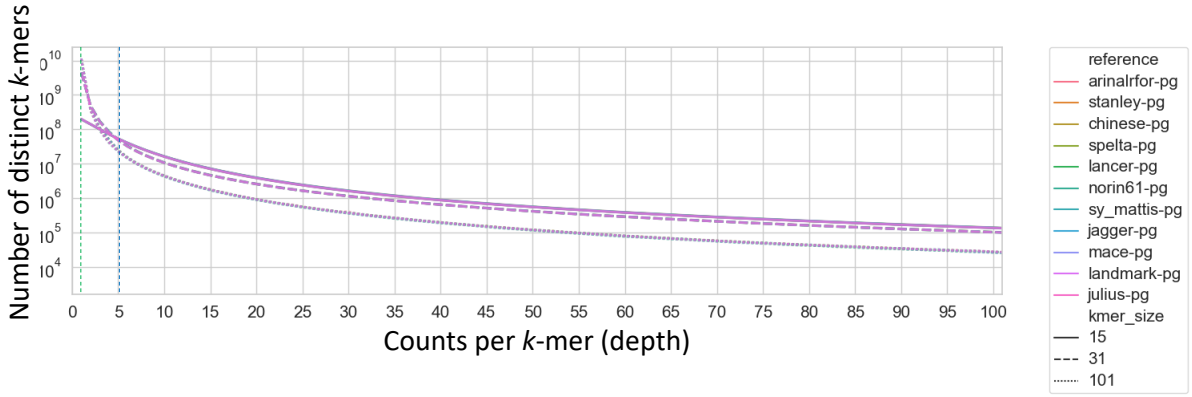


Fig. 2. 3. *k*-mer frequency distribution of the eleven chromosome-scale assemblies.

y-axis, *k*-mer count (Log) of the *k*-mer depth (frequency) in *x*-axis. High proportion of the *k*-mers is represented as unique as indicated with the highest values at the position 1 of *x*-axis (green line). After the first 5 (blue vertical line) occurrences (depth) the counts drops and stabilizes. Note that all genotypes are shown in the figure, but the curves overlap completely giving the impression of a single curve for each *k*-mer size.

The number of unique *k*-mers in all pangenomes increases considerably after ~19 *k*-mer size and continue a smoothly up to 101 *k*-mer size (**Fig. 2.4a)**. Although, very similar among pangenomes, *k*-mer profiles indicates slightly different levels of genome assemblies quality or real genome compositions among the wheat references. As expected, scaffold-level assemblies had overall more unique *k*-mers, which may reflect that scaffold level assemblies most likely do not assemble more complex repetitive regions and hence have a higher proportion of non-repetitive genome

regions than chromosome-scale assemblies. It may also indicate a high number of misassembled regions in the scaffold level assemblies giving rise to a high proportion if unique *k*-mers (**Fig. 2.4b)**.

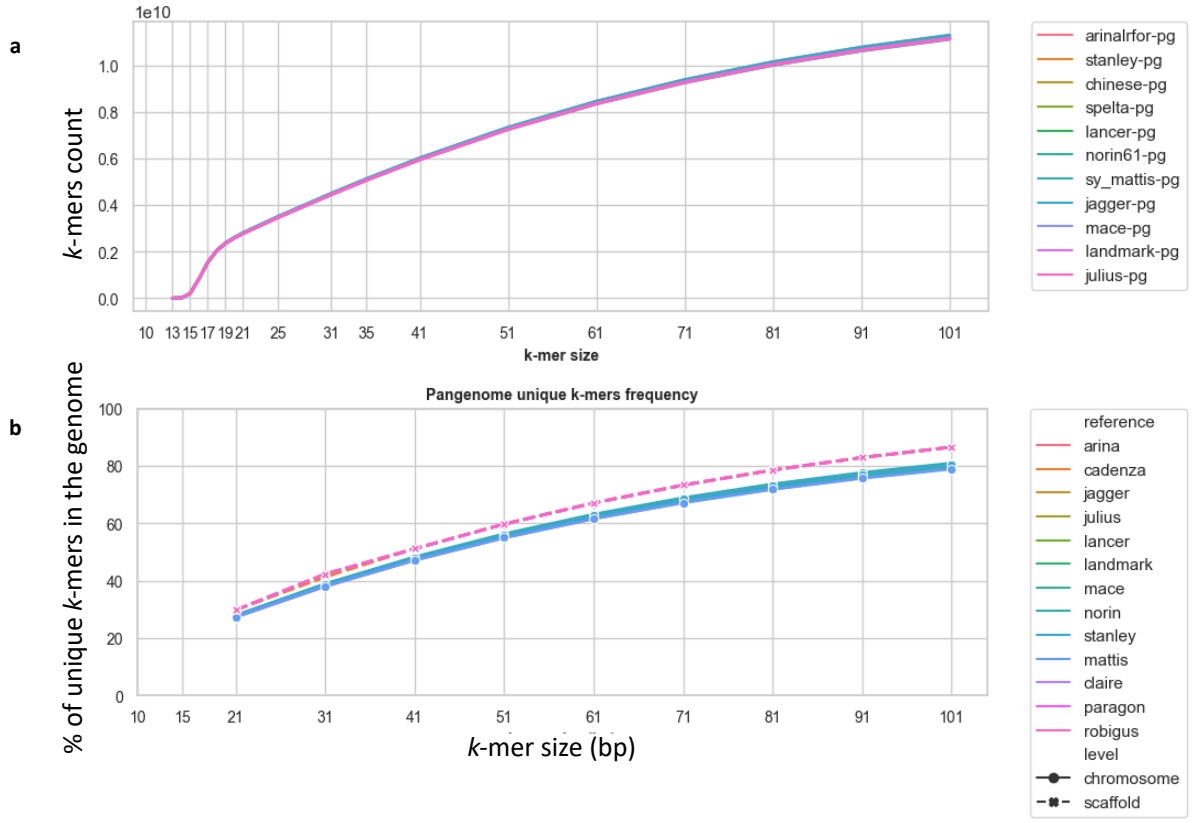


Fig. 2. 4. **Unique *k*-mer frequency distribution in the wheat pangenome.**
a) proportion of unique *k*-mers in the entire genome at different *k*-mer sizes. **b)** *x*-axis, percentage of unique *k*-mers in each of the references at different *k*-mer lengths. Solid lines are chromosome-level and dashed lines are scaffold-level assemblies.

In summary, our analysis revealed that the wheat chromosome-scale assemblies have a considerable representation of unique *k*-mers above >20-mers. Other studies have used this *k*-mer size as a default. The length of the *k*-mer determines the total fraction of unique *k*-mers found in the entire genome. Ideally, longer *k*-mer sizes are preferred as they capture the uniqueness of a DNA sequence in a genome which is of importance to differentiate within genome regions (e.g., for genome assemblies) and to differentiate among genotypes (e.g., to find variations by *k*-mers). However, longer *k*-mers is high computing demanding and there is a limitation due to sequencing read lengths. Increasing the *k*-mer size escalates the probability of capturing sequencing errors from reads edges inside *k*-mer sequences. Therefore, there is a limit and trade-off with computer

burden, genome size, and read length to select for an optimal *k*-mer size. As described in our methods, we will leverage the already available 31-mers from the WatSeq project. In this analysis we confirmed that 31-mers are in the range of optimal *k*-mer size found which is >20-mers and a reasonable computer burden for hexaploid wheat. In addition, considering that the WatSeq samples available are DNBSeq short reads of 150 bp, *k*-mers larger than 51 bp would be a limitation since longer *k*-mer size in short reads has the risk to extend regions prone to errors at read edges. This would result in less resolution to detect *variations* since real information will be lost. Therefore, in this analysis we will employ 31-mers as our default for downstream analyses.

In addition, our results also demonstrate a very high similarity on the *k*-mer profiles of the 11 chromosome scale assemblies. This might be due to the use of the same pipeline to create all the assemblies and therefore prone to similar errors (Walkowiak et al., 2020). A second explanation would be that the wheat pangenome assemblies have overall similar genome identity since most of them are important cultivars selected, a part of Norin61 and CS. This hypothesis would be possible to test as other genome assemblies are released particularly from landraces and from wild relatives. Although not explored in this thesis, a more in depth analysis comparing *k*-mer profiles per chromosome and/or per genome region would reveal if genome differences among cultivars are masked by the highly repetitive regions nature of the wheat genome (e.g., transposons content) (Appels et al., 2018).

2.4.1.2. *k*-mer distribution of raw reads

In addition to the genome complexity described above, other aspects impact on the *k*-mer profiles results when working with raw reads instead of genome assemblies. For example, analysing *k*-mer profiles using raw reads is influenced by the sequencing platforms used to generate the reads, sequencing quality, and read length and depth. In our previous analysis we screened the wheat pangenome for *k*-mer profiles and overall, all the chromosome level assemblies had a similar *k*-mer frequency. In this analysis we investigated the *k*-mer differences from raw reads. We aimed to detect sequencing differences in quality in our datasets to apply normalization/filtering criteria based on this information in downstream analysis.

From the WatSeq panel, we explored the *k*-mer (31-mer) profiles of 375 randomly selected genotypes which have on average ~12-fold depth coverage based on the number of reads (**Fig. 2.5**). Our results indicate that overall, most of the genotypes have similar coverage and *k*-mer profiles with few exceptions ranging from 9-fold to 14-fold. A few accessions had high multiple repetitive *k*-mers, which may be indicative of low-quality reads with a high error rate and low

coverage (Table S2.6). Across these 375 samples we found that the peak of the *k*-mer distribution was between 9 and 11-fold depth. This slightly lower than the 12-fold coverage based on read counts is expected since a pre-filtering step to the initial raw reads was applied to remove low quality reads to reduce the computer burden for *k*-mer and mapping analysis (Cheng et al., under revision).

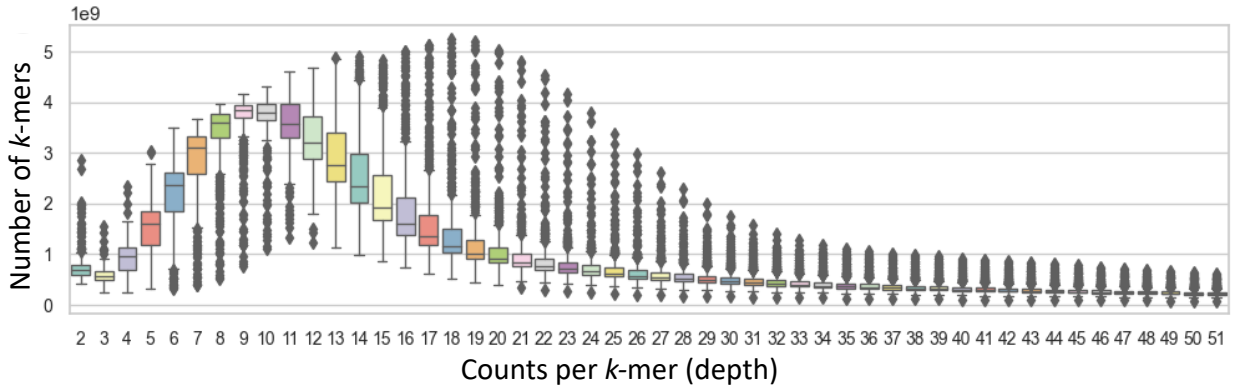


Fig. 2. 5. 31-mers distribution of raw reads.

The data is from 375 random WatSeq raw reads samples at ~10-15-fold coverage (average 12-fold) plus 23 pangenome subsampled raw reads at approximately equivalent coverage (12-fold). Unique *k*-mers were removed from this dataset. A peak is reflected at ~9 to 11x depth indicating the approximately coverage of the reads after pre-filtering step of cleaning and removing low quality reads from the initial raw reads.

2.4.2. Implementation of Identity By State in python (IBSpy) to detect variations.

2.4.2.1. Types of variations captured by *k*-mers.

There are different *k*-mer based methods to identify and quantify genetic variations between genome sequences. Here, we describe a new algorithm using presence/absence of *k*-mers in genomic intervals. This approach registers three types of scores using a genome reference: "*observed_kmers*", "*variations*", and "*kmer_distance*". In all cases, the scores are measured in 50 Kbp windows using a genome reference chromosome physical position.

The rationale of using 50 Kbp window as our base starting point was the work demonstrated in (Brinton et al., 2020) using pairwise chromosome alignments. In their analysis, they found that when examining windows <50 Kbp, it was not possible to differentiate between identical-by-state (IBS) and near-IBS regions in wheat due to lack of sufficient variation in small windows (50 Kbp). For example, to define an IBS region between two genotypes (Brinton et al., 2020) set a cut-off of 1 SNP in 10,000 bp which accounts for 99.99% sequence similarity. A one SNP in 10,000 bp allowed

flexibility for misassembles and sequencing errors. Using this criterion, they explored different window sizes (1 Mbp, 2.5 Mbp, and 5 Mbp) and selected 5 Mbp as optimal to detect IBS regions. Therefore, as starting we decided to initiate to explore 50 Kbp and expand larger windows as needed to define IBS regions by combining multiple 50 Kbp sub windows and will be described in Chapter 3.

IBSpy uses two types of input data, a genome assembly and a *k*-mer database. The database can be generated from raw reads of any type, quality, or from genome assemblies (scaffold or chromosome scale). The algorithm parses the genome assembly on the go and splits chromosomes into 50 Kbp windows to record all possible *k*-mers ignoring unassembled regions with "Ns". Based on these two sets of information, the three types of scores are calculated. We explain each in turn:

2.4.2.2. Observed *k*-mers

Based on the genome assembly and *k*-mer database, IBSpy counts the number of *k*-mers in the 50 Kbp window from the reference also present in the *k*-mer database of a query sample. If all the 31-mers are present, the score for the *observed_kmers* is equal to all possible *k*-mers in the 50 Kbp window. An example of how *observed_kmers* are calculated is shown in Fig. 2.6.

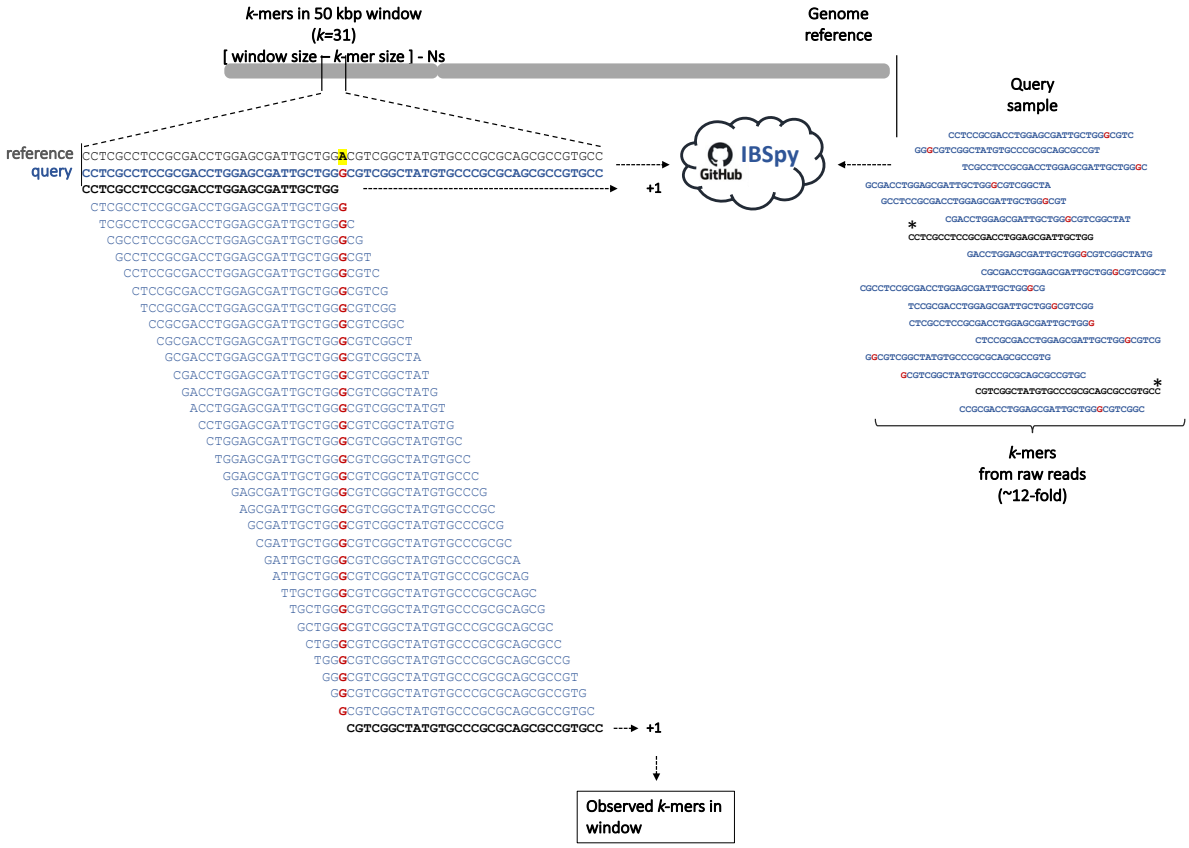


Fig. 2. 6. IBSpy "*observed_kmers*" in window.

Using 50 Kbp windows of the genome reference (grey), IBSpy compares all possible *k*-mers in a window reference sequence to the *k*-mers of a query sample (blue) (from raw reads, scaffold-level, or chromosome-scale assemblies) and count the number of observed *k*-mers within each window. The example shows a SNP (in red) between the reference (grey) and a query (blue) and how only the observed *k*-mers present in both, the reference, and the query, are counted. Next, the counts and the chromosome position are recorded, IBSpy moves to the next 50 Kbp window and the process is repeated until reaching the end of the chromosome and subsequently to the whole genome.

2.4.2.3. *Variations*

Similarly, as with the *observed_kmers*, to register the *variations* score, IBSpy parses 50 Kbp window of the reference assembly, but this time it reads each *k*-mer in the 50 Kbp from the start comparing one *k*-mer sequence at a time against the *k*-mer database of the query sample. If one or a set of "*contiguous overlapping k-mers*" from the reference are not present in the query sample, this is registered as a single *variation*. This would be equivalent to a '*short-phased sequence*' captured. However, for simplicity, we will refer to them as a "*variation*". This score includes the count and the window position from the reference to which they were mapped (**Fig. 2.7**). To differentiate between genetic variations as a more general term and IBSpy *variations*, we will refer as IBSpy *variations* in italics.

A *variation* can have different sequence lengths falling into three possible categories.

a) A single SNP yields a *variation* = 2n-*k*. For example, when using 31-mers, a single SNP will generate a *variation* of 61 bp (**Fig.2.7**).
b) A deletion in the query sample will generate a *variation* size = 2n-*k* + deletion size (**Fig. 2.8**).
c) An insertion in the query sample (deletion in the reference) will not be detected.
d) Two SNPs whose distance is less than the *k*-mer size used (in this example SNPs closer than 31 bp) will generate a variation size of 2n-*k* + the distance from the first to the last SNP found. This will be similar to the *"kmer_distance"* score illustrated in **Fig. 2.9**. In all cases the *variations* score will record a single *variation* for all these types of polymorphisms described.

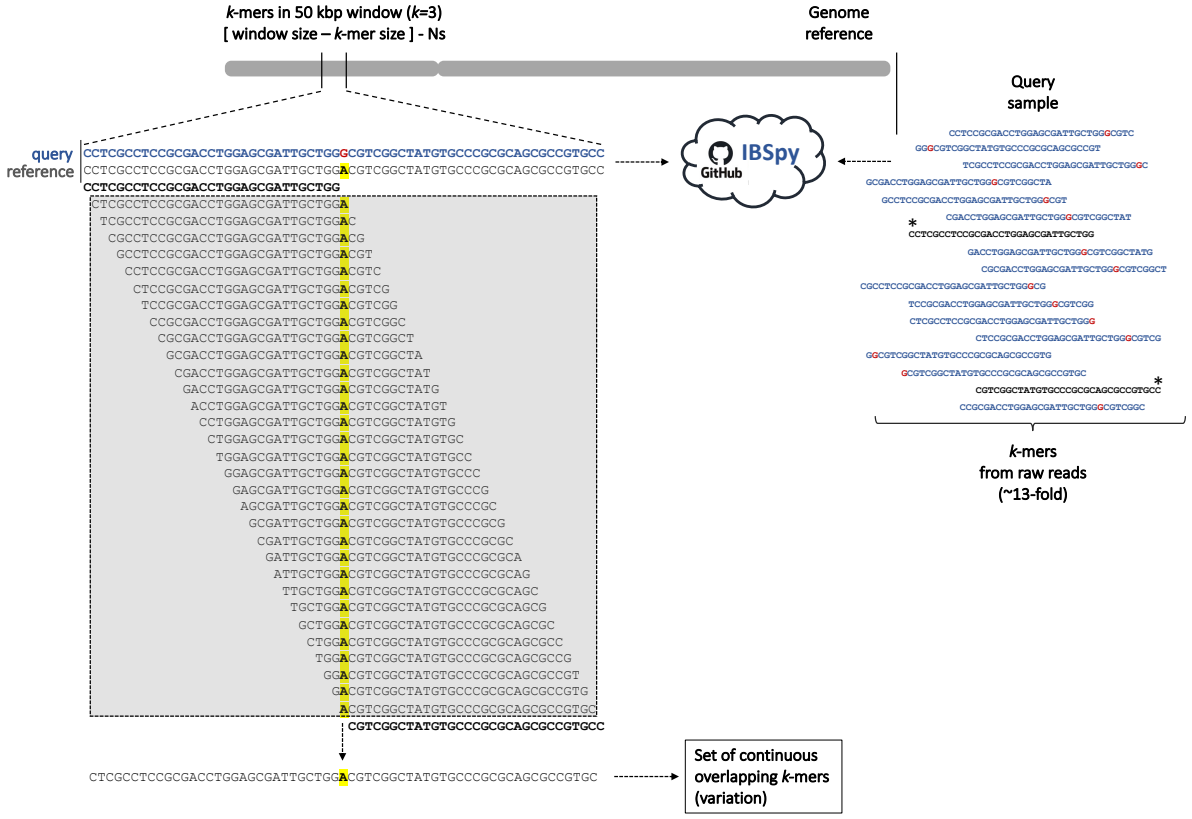


Fig. 2. 7. IBSpy *variations* score.

Using 50 Kbp windows and *31*-mers, IBSpy compares *k*-mers in a reference sequence to the *k*-mers of any query sample and count the number of *variations* within each 50 Kbp window. A *variation* is defined as a set of continuous overlapping *k*-mers from the reference completely absent in the query (square box). In this example, a single *variation* of 61 bp length is shown (at the bottom), which is the condensed score of all 31-mers having the A nucleotide (in yellow) not present in the *k*-mers of the query sample from raw reads (blue sequence). After recording this *variation*,

IBSpy continues to scan the sequence of the reference across the 50 Kbp window until a new set of continuous overlapping *k*-mers are absent in the query sample. At the end of the 50 Kbp window, the reference chromosome position and the total number of *variations* are recorded before moving to the next 50 Kbp window and the process is repeated. Low *variations* count indicates high similarity between the 50 Kbp reference window in the assembly and the query sample, whereas high *variation* count indicates low sequence similarities.

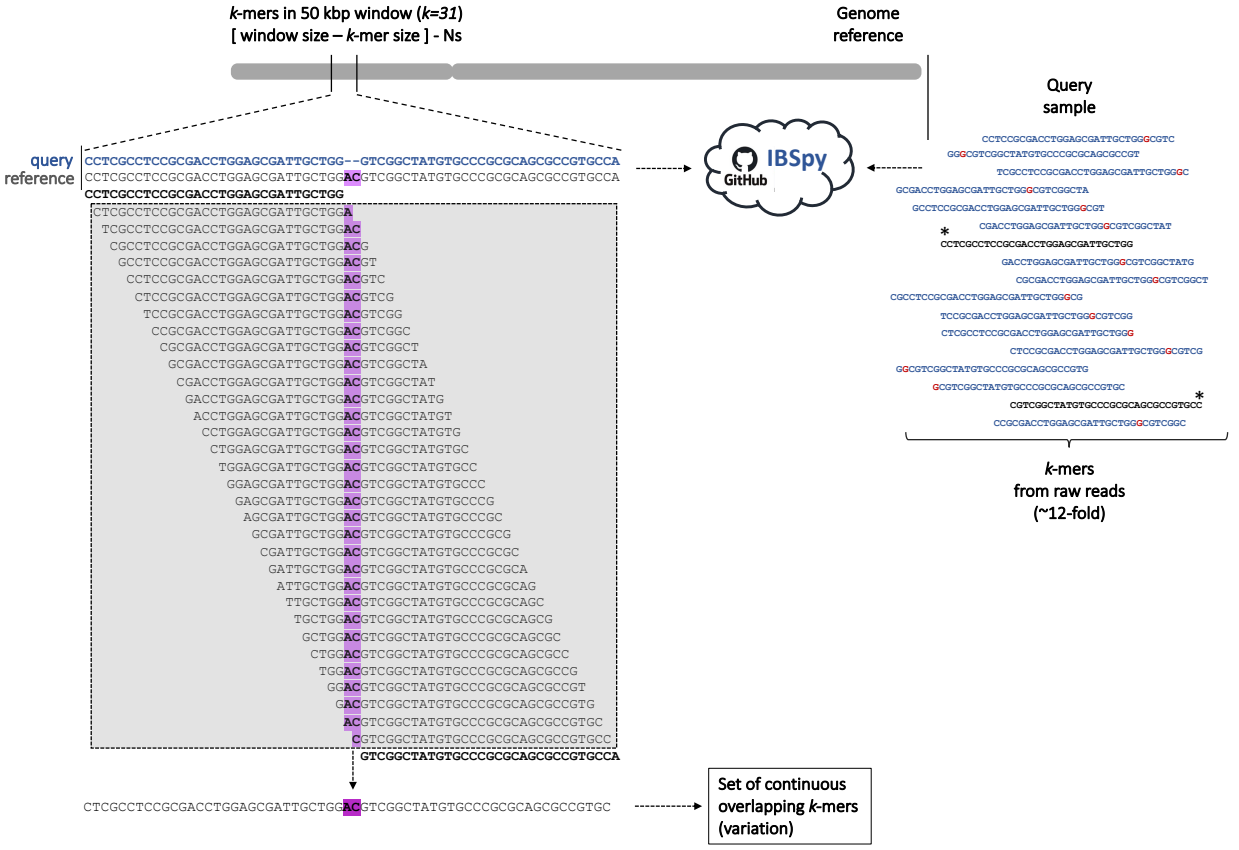


Fig. 2. 8. A 2 bp deletion counts as a single *variation* but has two *kmer_distance* counts.
Example of a 2 bp deletion in the query sample (blue). The *observed_kmers* will score minus two from the total window size (as in **Fig. 2.6**), *variations* will be counted as one, and the *k*-mer distance will capture two counts. In all cases, these scores will be added to the other counts in the 50 Kbp window.

2.4.2.4. *k*-mer distance

When a two or more SNPs are closer than the *k*-mer size they are registered as a single variation. However, *kmer_distance* will be longer as all the *k*-mers between the SNPs are missing in the query sample. Therefore, the *kmer_distance* is calculated by the distance between the first and the last SNP within a *variation*. If there are additional SNPs in between the flanking SNPs, the distance will be the same (**Fig. 2.9**).

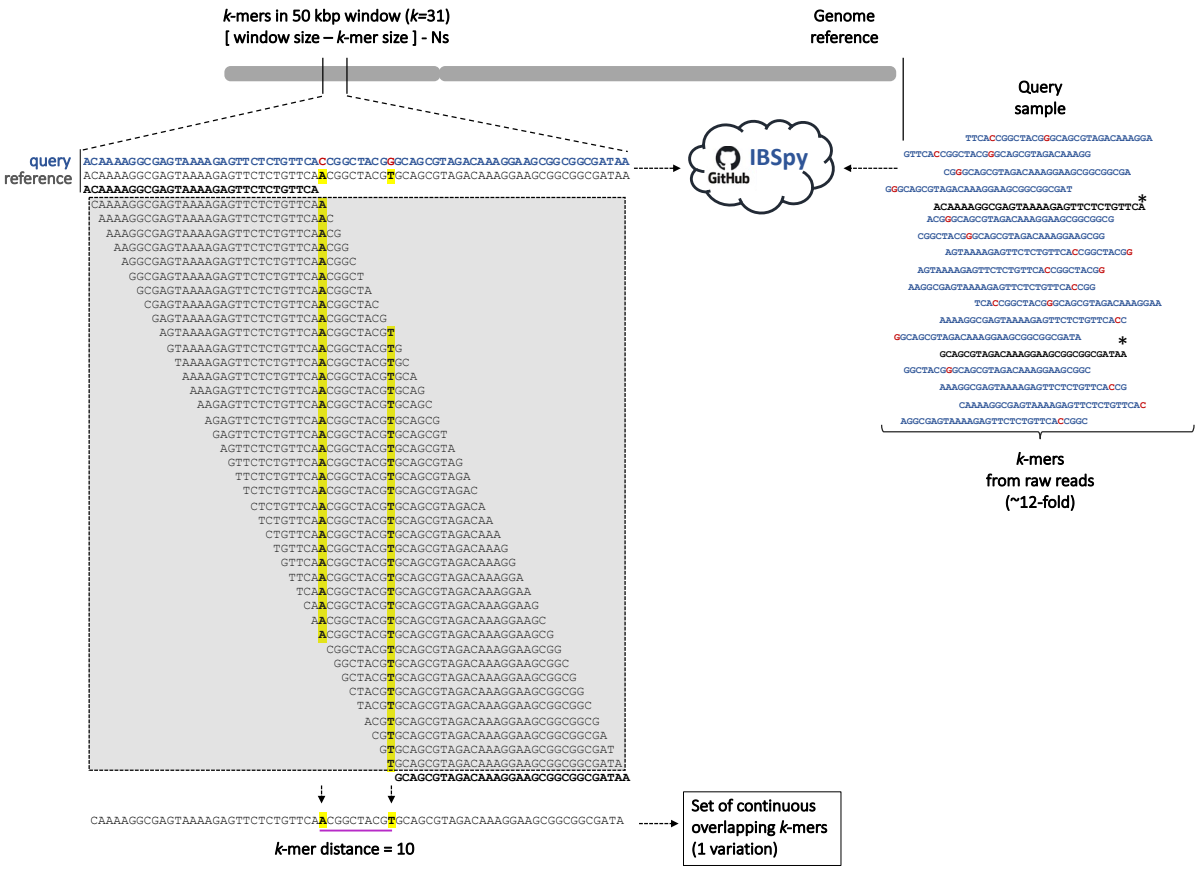


Fig. 2. 9. IBSpy *kmer_distance* score. Example of two SNPs closer than the *k*-mer size (31-mers).

The two SNPs are 10 bp away of each other (pink bar) and therefore, *kmer_distance* will count 10. *observed_kmers* will count two and *variations* will count one.

In summary, we stablished a method called IBSpy which captures three scores in a 50 Kbp window based on a genome assembly and a *k*-mer database: *observed_kmers*, *variations*, and *kmer_distance*. Each of the scores register similar, albeit slightly different features based on the type and number of variations between two samples. The output file from IBSpy is a plain tab separated file with seven columns. In the following analysis, we will explore differences captured by each of the scores and document each of them for their purposes. An example of an output file is shown in **Table S2.9** with further descriptions https://github.com/Uauy-Lab/IBSpy.

IBSpy *variations*: a fingerprint for the wheat genome and among subgenome homeologs.

As an entry point to this analysis, we first explored the *variations* score. Using their chromosome positions and 50 Kbp window, we analysed the *variations* count distributions between two samples across the whole genome. We hypothesized that if there were differences between two

samples at the sequence level (SNPs or InDels), these will be captured by the number of presence/absence *k*-mers in a genome region (e.g., genome window size) by the appropriate *k*-mers size and yield a genome wide fingerprint of *variations* from the different genome regions.

To test this hypothesis, we used a single reference genome (Mattis) and compared it to two chromosome-scale assemblies, one of them known to be highly related to the reference (Julius) and the second known to be a more distant genotype Chinese Spring (CS) based on (Brinton et al., 2020). We focused on the homeologs chromosome 6 (chr6) since chr6A is well characterized in our group. Julius and Mattis share two IBS blocks in telomeric regions of chr6A, two large IBS blocks on chr6B, and one main large block on chr6D. On the other hand, CS, does not share those blocks with Mattis.

First, using the *variations* distribution in 50 Kbp window, we observed that chr6A had several windows with high *variations* across the chromosome physical positions, often with values higher than 100 *variations*. These windows with high variations counts were adjacent among them until a count drop was observed from high to low *variations* as we moved through the chromosome. Interestingly, telomere regions had more shifts in high and low *variations* counts (and vice versa) than centromere regions (**Fig. 2.10a** top). This could be indicative of recombination breakpoints and consistent with literature which shows higher recombination in telomeric regions of wheat chromosomes (Choulet et al., 2014). When exploring chr6B, in addition to the patterns of high and low *variations* detected on chr6A, we observed several adjacent 50 Kbp windows with low *variations* count close to "0" matching the IBS regions defined by (Brinton et al., 2020) and (http://www.crop-haplotypes.com/Wheat/haplotype/6B) from 50 to 255 Mbp and from 480 to 665 Mbp (**Fig. 2.10a** middle). These patterns gave our first insight to detect IBS by counting IBSpy *variations* in windows. Exploring chr6D, we observed fewer windows with high *variations* (e.g., >30 count) than in chr6A or chr6B, and these were mostly located at the end of the chromosome. The *variations* count between Mattis, and Julius were mostly low and tended to be lower than 30 *variations* (**Fig. 2.10a** bottom).

Overall, considering the three homeologs chromosomes we spotted four main levels of *variations* on wheat which were consistent across multiple 50 Kbp windows (**Table 2.2**).

a) low and close to 0 variations in a continuous set of windows which often had values <10 (e.g., chr6A below purple line, **Fig. 2.10a** top),
b) below <30 with median at ~15 and which did not have windows with zero values (e.g., chr6A between yellow and purple lines, **Fig. 10a** top),
c) high above >30 and median at 120; this category was relatively rare (e.g., chr6A between green and yellow lines, **Fig. 10a** top), and

d) very high >120 with median at ~350 variations (e.g., chr6A between blue and green lines, Fig. 10a top). An additional, and more distant category was detected at >350 with median 500 when comparing *variations* among homeolog chromosomes (data not shown). The histogram of the distribution by chromosome clearly distinguished those levels of *variations* with peaks at "0", at ~15 and <30, and at median 120. The 350 median level of *variations* is hardly noticeable since there were a few regions having this level of variation (**Fig. 2.10b**, *x*-axis Log scale). Comparing the histogram distribution of CS as a query, we observed that there were no peaks distributions with "0" counts and this comparisons had the highest peak with median of ~15 and <30 *variations* count category (**Fig. 2.10c**).

Table 2. 2. Different levels of *variations* detected in 50 Kbp windows among genome assemblies and the hypothetical relatedness.

Level	Cut-off	Proposed relatedness
1	<10	Pairwise IBS
2	<30	Immediate gene pool
3	~120	Intermediate gene pool
4	~350	Distant wild relatives
5	>350	Homeologs A B D and deletions

To validate the levels of variations detected and obtain a clearer picture of the *variations*, we tested the *variations* profile across the whole genome. Overall, we detected similar patterns with roughly higher *variations* in the B sub genome than in A and D subgenomes. At the same time, we detected more *variations* in subgenome A than in D (**Fig. S2.1**).

In summary, our results show that the *variations* score by *k*-mers can detect genome differences between two genotypes and across the whole genome. Regardless of the relatedness of the genotypes employed, those differences could be detected. In general, five main levels of variations are distinguishable, one of them with *variation* values close to "0" in adjacent windows matching IBS regions defined in (Brinton et al., 2020). Consistent with previous (Akhunov et al., 2010; Balfourier et al., 2019; Cseh et al., 2021) results, IBSpy *variations* agrees that A and B subgenomes to have more diversity than D subgenome in terms of genetic variations detected by the common alignments methods.

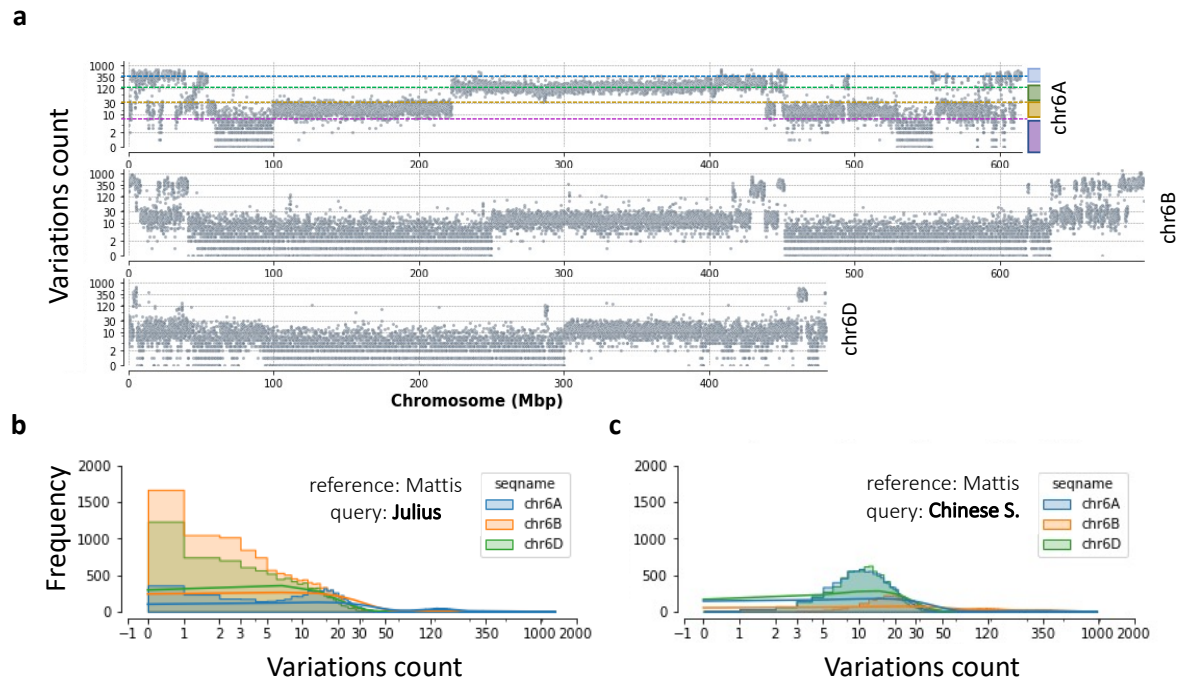


Fig. 2. 10. IBSpy *variations* detects four main levels of genetic diversity in wheat.

In all plots are the *variations* data from chr6A, chr6B, and chr6D using Mattis as a reference. **a)** *y*-axis variations count (Log) across chromosome physical positions (*x*-axis), **b)** histogram distribution of the *variations* in **Julius** query sample and in **c)** Chinese Spring. *x*-axis in **a** and **b** is Log scaled. Straight lines indicate different levels of *variations* as described in **Table 2.2**, pink = <10, yellow <30, green ~120, and blue ~350. Above 350 *variations* count are differences among A, B, and D subgenomes and large deletions. Note that Mattis (reference) shares higher sequence similarity to Julius **b**) than to Chinese Spring **c**) in agreement with (Brinton et al., 2020).

Relatedness among IBSpy scores

We next explored if *observed_kmers* and *kmer_distance* produced similar results to *variations* scores and how are they related among them. Plotting the distributions along chromosome physical positions, we detected that *observed_kmers* detected similar block patterns in the IBS regions on chr6B as with the *variations* score (**Fig. 2.11ab,** yellow boxes**)**. The *observed_kmer* in a 50 Kbp window was >99.95% with adjacent windows as seen with the *variations*. In (Brinton et al., 2020), these two IBS blocks had >99.99% sequence identity similarity.

Comparing *kmer_distance* score, we observed a similar block pattern, but the blocks were less defined and had several 50 Kbp window with high values mixing with the second level of *variations* (**Fig. 2.11c,** yellow boxes). One of the reasons that *kmer_distance* does not properly differentiate between categories may be the missing reads due to misassembles in the query or the reference and accumulation of counts. In comparison the *variations* score compensates for those errors and

"buffers" by combining continuous overlapping variations into one type. Therefore, from this analysis onwards, we will discard the use of *kmer_distance* in downstream analysis.

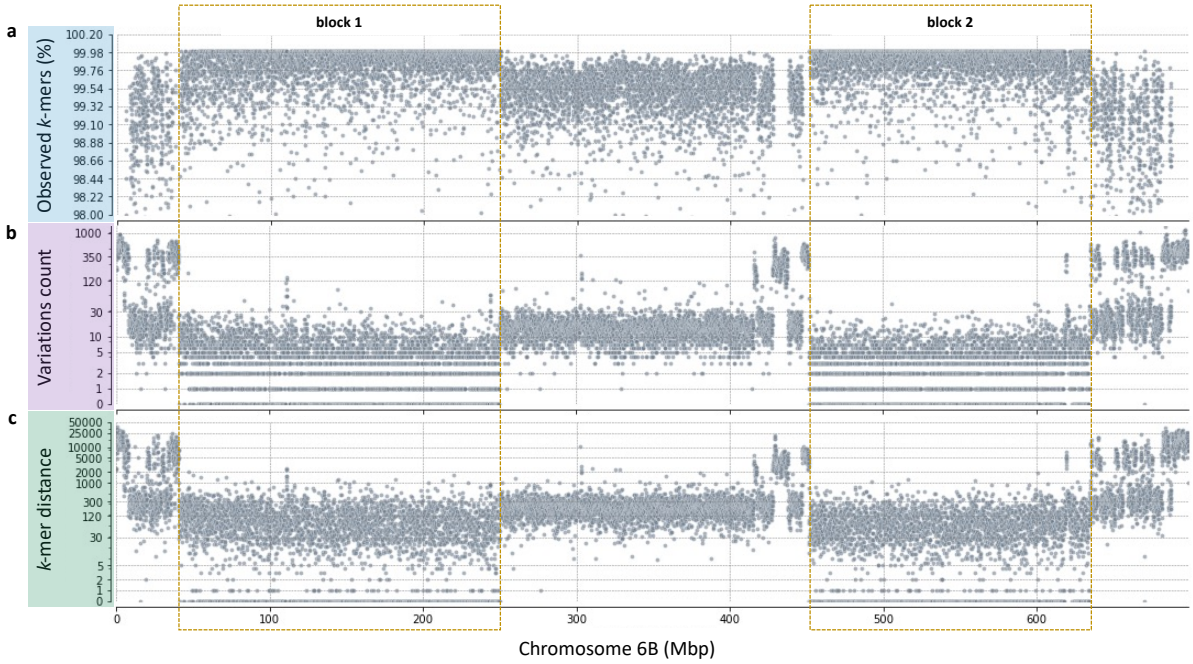


Fig. 2. 11. IBSpy scores comparison using Mattis reference, Julius as query sample, and chromosome 6B example. In the three plots, each dot represents a 50 Kbp window and their corresponding score between Mattis reference physical positions (*x*-axis) and Julius query sample. **a)**, *observed_kmers* score transformed to similarity percentage (*y*-axis) and zoomed in from 98-100% between the two samples. **b)**, *Variations* count score count (*y*-axis Log), and **c)** *k*-mer distance score. Yellow boxes represent the two regions on chr6B that were identified as IBS by (Brinton et al., 2020) and have low *variations* count and low *k*-mer distance.

2.4.3. IBSpy *variations* with raw reads

In the previous analysis we used the IBSpy output based on the chromosome-scale assemblies both as a reference and as a query samples. We next wanted to characterize IBSpy *variations* derived from raw reads. The aim was to validate IBSpy to detect *variations* using raw reads and characterize which factors impact on the results. Across these analyses we made use of IBS regions and sequence similarity among genome references defined in (Brinton et al., 2020) as positive controls. We investigated several parameters including: (1) sequence depth coverage, (2) removing or keeping unique *k*-mers, (3) *k*-mer lengths under different read lengths to detect *variations*, (4) different sequencing platforms and a combination of reads and, (5) using scaffold level assemblies as a query. In the end we selected a set of parameters for downstream analyses

based on the data available in this project and propose key points for further improvements in future developments.

2.4.3.1. Sequence coverage (depth)

Alignments methods and SNP calling are affected by sequencing depth. Genome assemblies and *k*-mer based algorithms are usually affected by read length where less coverage is needed with longer reads. In previous analysis, we explored *k*-mer profiles and characterized IBSpy *variations* equivalence to alignments sequence identity using chromosome-scale assemblies. In this analysis we tested and validated IBSpy to detect *variations* using raw reads with different sequence depths and defined an optimal sequence coverage by comparing known IBS regions. When working with *k*-mers and raw reads, it's common to remove unique *k*-mers to avoid computational burden since unique *k*-mers are mostly originated from sequencing errors. As an entry point into this analysis, we made initial tests on IBSpy by removing unique *k*-mers, and we discuss further down the effect of including them.

We first determined whether our approach was able to detect *variations* in a similar way as using chromosome-scale assemblies. We used a set of subsampling reads from 1 to 20-fold, a range of sequence depth commonly used in alignments and SNP calling pipelines. To account for similar dataset as in our genome assembly analysis, we subsampled raw reads that were previously used to generate the 11 pangenome assemblies (Walkowiak et al., 2020). For consistency, as an example, we will show the analysis between Mattis as a reference and Julius raw reads as a query genotype.

We first plotted the variations count in 50 Kbp window across the chromosome physical position. We observed that above >4-fold coverage, a separation between two levels of *variations* were detected. The separation of these levels of *variations* were similar to the observed in chromosome-scale assemblies at ~10 – 12-fold coverage (**Fig. 2.12**). Similarly, the regions defined as IBS by Brinton et al., 2020 and available in http://www.crop-haplotypes.com, started to be detected as having several adjacent 50 Kpb window close to <10 *variations* count (**Fig. 2.12,** yellow boxes).

Analysis of the same region but using the *observed_kmers* score detected a block with high percentage of *observed_kmers* in the same regions defined as IBS on chromosome 2B detecting two main blocks as seen in the *variations* count score (**Fig. 2.13a**). A detailed analysis demonstrated that the *observed_kmers* were above >99.98% in the IBS region when using 10 to 12-fold coverage (**Fig. 2.13b**).

We next counted the *variations* distributions of the three six homoelogues (A B D) chromosomes (with unique *k*-mers removed). We noticed that with low coverage, a single distribution of the data was formed with a peak at ~350 *variations* count in all. Above ~10-fold we started to see a separation of the data and two distributions of the data were formed in some comparisons. One of the main peaks was located at ~30 and the second at <10 variations. Those distributions were clearly separated at ~12x. Above the 12-fold coverage few improvements to separate these distributions was gained (**Fig. 2.17**). Although these distributions were similar to the observed in the chromosome-scale assemblies analysis, we noticed that at 20-fold raw read coverage we obtained better resolution than with chromosome-scale assemblies. This would indicate that the chromosome scale assemblies had some misassembles or had missing data.

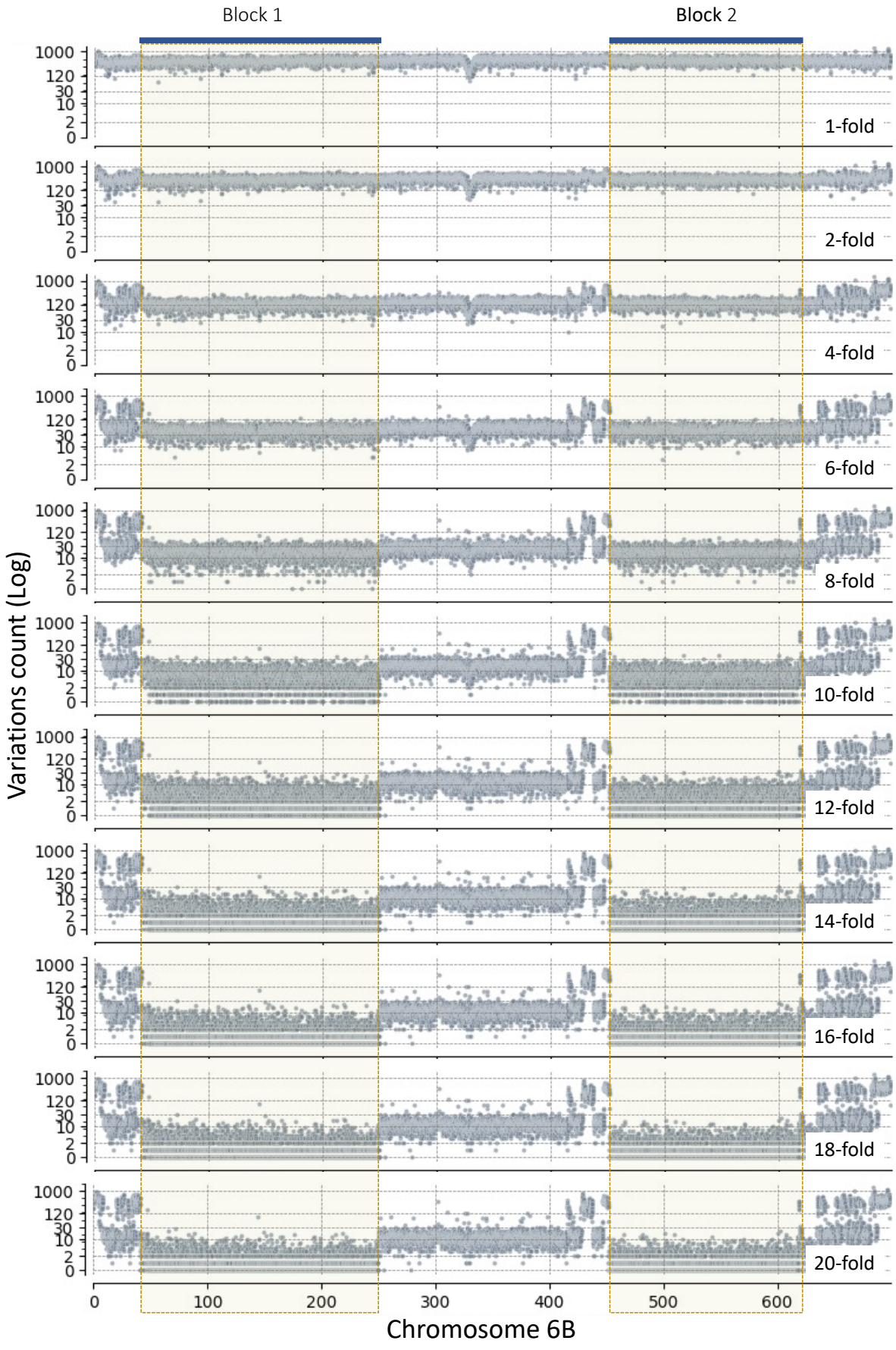

Block 1
Block 2
1-fold
2-fold
4-fold
6-fold
8-fold
10-fold
12-fold
14-fold
16-fold
18-fold
20-fold
1000
120
30
10
2
0
Variations count (Log)
0
100
200
300
400
500
600
Chromosome 6B

Fig. 2. 12. IBSpy *variations* score at different sequencing depths (removing unique *k*-mers).

Illumina short reads (250 bp) of Julius (query) against Mattis (reference). **a)** *x*-axis (Log), IBSpy *variations* count in 50 Kbp window at defined fold coverage from 1-fold to 20-fold. *Y*-axis chromosome physical position. The blue bars and the yellow boxes indicate the IBS regions depicted in (Brinton et al., 2020).

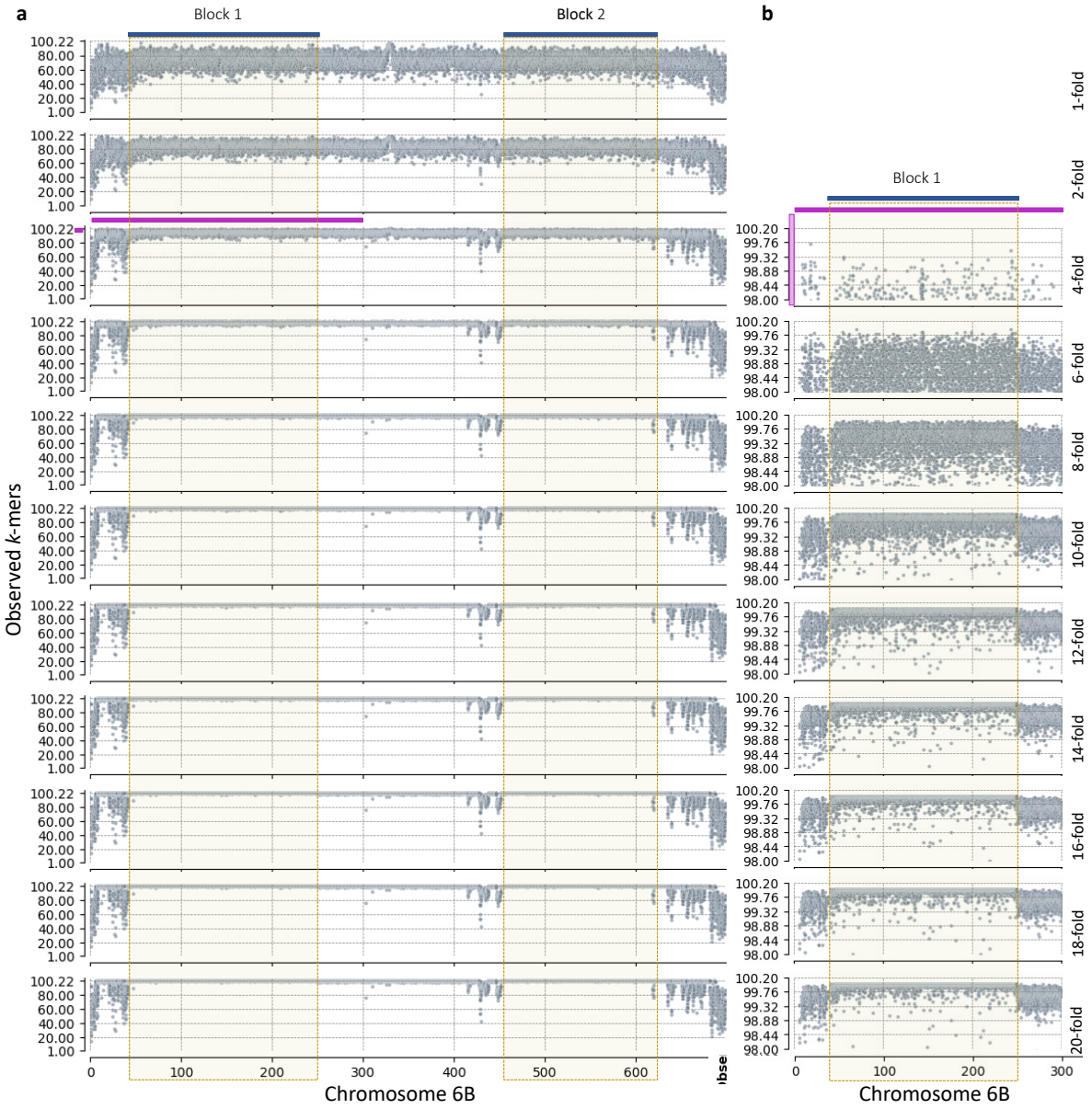


Fig. 2. 13. IBSpy *observed k-mers* score at different sequencing depths (removing unique *k*-mers).

Illumina short reads (250 bp) of Julius (query) against Mattis (reference). **a**) *x*-axis, IBSpy *variations* count in 50 Kbp window at defined fold coverage from 1-fold to 20-fold. *Y*-axis chromosome physical position. In yellow two IBS regions depicted in (Brinton et al., 2020) on chr6B. **b**), a zoom from **a**) in the 98 to 100 % region and from 0 to 300 Mbp chromosome physical position as depicted by the pinks bars. 1-fold and 2-fold were omitted in **b**) since thy had

zero datapoints in the zoom in region at > 98% *observed k-mers*. The same analysis but keeping unique *k*-mers is shown in **Fig. S2.2**.

We next compared the *variations* count of raw reads of Julius against the chromosome-scale of Mattis (**Fig. 2.14**) and scaffold level assemblies (**Fig. 2.19**). We observed that the *variations* detected using raw reads at 12-fold matched the *variations* counts by chromosome-scale assemblies (**Fig. 2.14**). In our previous analysis we explored know IBS regions among pangenomes (based on alignments in Brinton et al., 2020), and did a comparison with chromosome-scale IBSpy *variations* to determine that values of <10 in adjacent 50 Kbp windows were indicative of IBS regions. Those regions were from 50 Mbp to 250 Mbp and from 450 to 635 Mbp in chr6B between Mattis and Julius. These two genotypes share 23.1% IBS regions on the entire genome and 53.6% on chr6B. When we compared the IBS *variations* from raw reads, we observed consistent results as with chromosome scale to have *variations* count <10.

When plotting the IBSpy *variations* from raw reads across chromosome positions, we observed similar profiles to the IBS *variations* regions from chromosome-scale assemblies. These variations matched the IBS regions defined in Brinton et al., 2020 to have *variations* <10 counts and continuous zero *variations* in 50 Kbp windows. These results were consistent in all the pangenome comparisons and across the whole genome. Our results suggest that IBSpy detects *variations* with raw reads to a similar extent as with chromosome level assemblies and that an optimal coverage for raw reads would be ~12-fold for wheat removing unique *k*-mers.

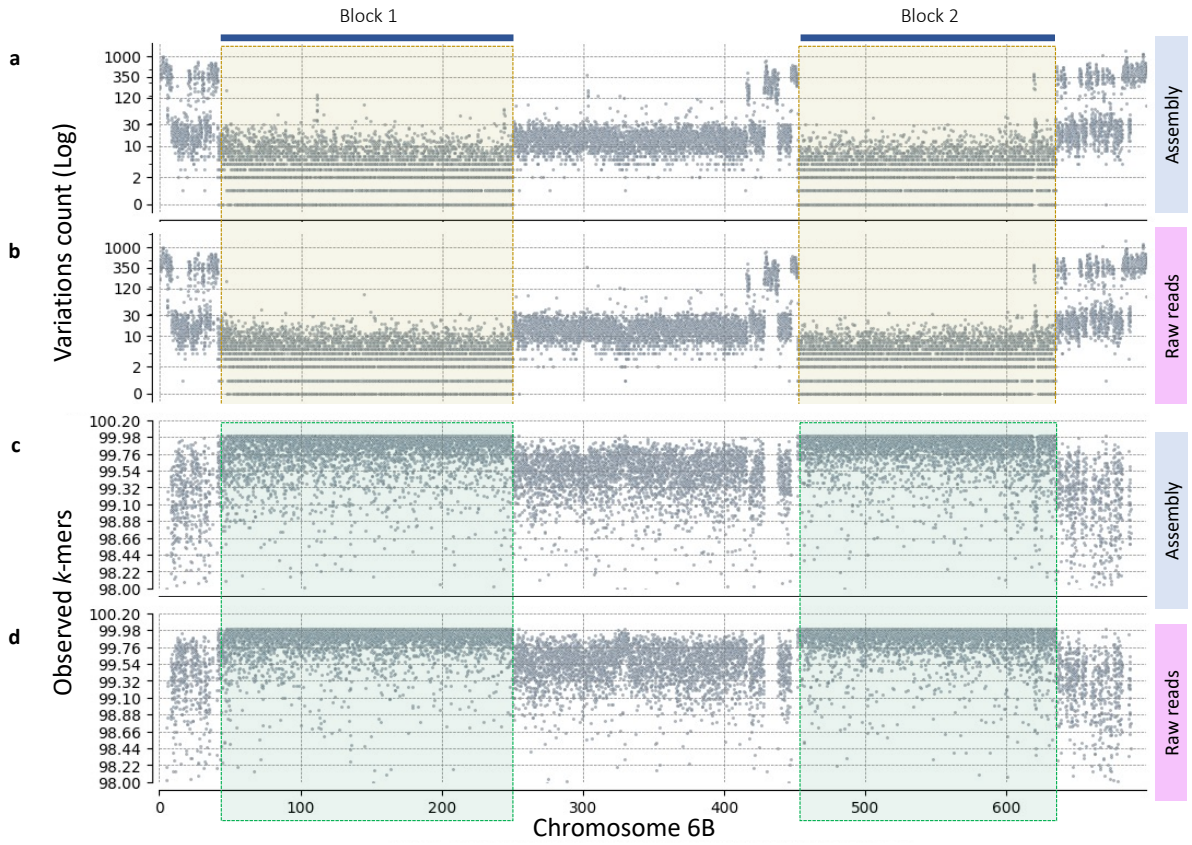


Fig. 2. 14. Raw reads at 12-fold vs chromosome-scale.

Comparison of raw reads vs chromosome-scale assembly Julius vs Mattis reference. **a**) *variations* count in 50 Kbp window across chromosome 6B of Mattis physical positions. *k*-mers from the genome reference of Julius against the Mattis genome assembly. **b**) *k*-mers from 12-fold raw reads of Julius against Mattis genome reference. **c**) and **d**) similar to **a** and **b** but plotting the *observed_k*mers score in 50 Kbp window from 99 to 100% interval.

We then tested if using the same genotype as a reference and querying raw reads from its own genotypes would detect *variations*. Our results demonstrated that there were mostly low *variations* across the whole genome. However, the *variations* were not zero and levels of *variations* were slightly higher at telomeric regions. These levels of *variations* were slightly higher with ~10-fold raw data coverage than with 20-fold. The histogram distribution indicates that most of the *variations* fall in the range of <10 variations, which is similar to the level of *variations* detected for IBS regions in our previous analysis. This provides an empirical value to determine our threshold for detecting IBS regions for downstream analysis (**Fig. 2.15**). These results suggest that our improvements in resolution detected by 20-fold compared to the chromosome-scale assemblies described above, may be due to misassembles in the reference or by high coverage reads having more probabilities to match a *k*-mer in the reference by chance.

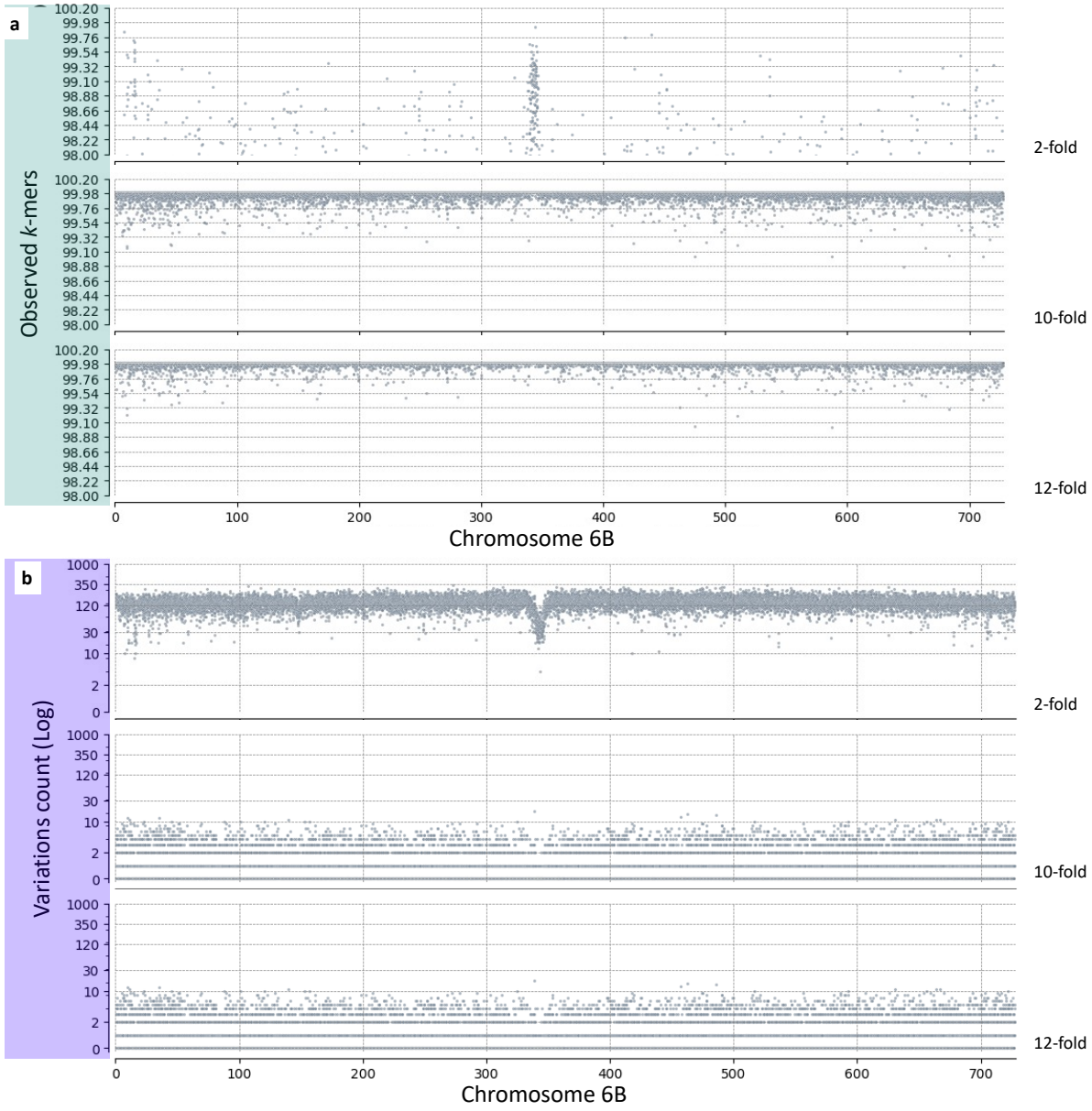


Fig. 2. 15. Julius raw reads against Julius reference at different sequencing depths keeping unique *k*-mers as a quality control for IBSpy.

a) Observed *k*-mers score of Julius at 2, 10, and 12-fold raw reads. As expected, at 10 and 12-fold most close to 100% of the observed *k*-mer score. **b**). *Variations* count of Julius at 2, 10, and 12-fold raw reads. Similarly, *variations* count <10 is present across the chromosome as indication of our background noise. This is most likely due to sequencing coverage, error, and reference misassembles. The *variations* <10 is consistent as the expected IBS regions found in other comparisons between two different genotypes.

2.4.3.2. Removing or keeping unique *k*-mers

Reports have demonstrated that removing unique *k*-mers in *k*-mer analysis when low coverage raw reads are employed can lead to real sequence information being lost (Lee et al., 2020). This occurs because sequencing is not uniform across the genome and unique *k*-mers resulting from non-overlapping reads can be lost. In this analysis we compared the effect of keeping unique *k*-mers using 12-fold coverage.

We focused specifically on the distribution of data in the <10 and the <30 categories as this is where we hypothesised the removal of unique *k*-mers could have the biggest impact. As expected, when comparing raw data where unique *k*-mers were maintained vs removed, we observed that the distributions of the <10 and <30 data were more clearly separated with unique *k*-mers at 12-fold. We observed an increase in the number of *variations* uniformly across the genome (Fig. 2.16) when removing unique *k*-mers, demonstrating that by keeping unique *k*-mers the resolution to separate the two distributions was improved. Keeping unique *k*-mers at 12-fold had similar resolution as with a chromosome-scale genome reference (Fig. 2.16, 2.17). Keeping unique *k*-mers and using 10-fold coverage had similar resolution as with 12-fold removing unique *k*-mers.

Depending on the read quality, in general, removing unique *k*-mers decreases the resolution by the equivalence of ~3x sequence coverage when using the IBSpy *variations* data. These results suggest that for datasets with <12-fold coverage, it is recommended to keep unique *k*-mers to compensate for the lack of sequencing coverage. Keeping unique *k*-mers when coverage is low (<12-fold) is particularly important to discriminate between *variations* counts ~<10 from the <30 which would impact int the ability to distinguish IBS and near-IBS regions. To deal with computer burden and keeping unique *k*-mers, smaller *k*-mers sizes may be required, but this topic was not explored in this thesis. Since there is a computer cost to storage data and there is also a cost to generate 3-fold more sequencing data, therefore, users may need to set a trade-off in each particular case.

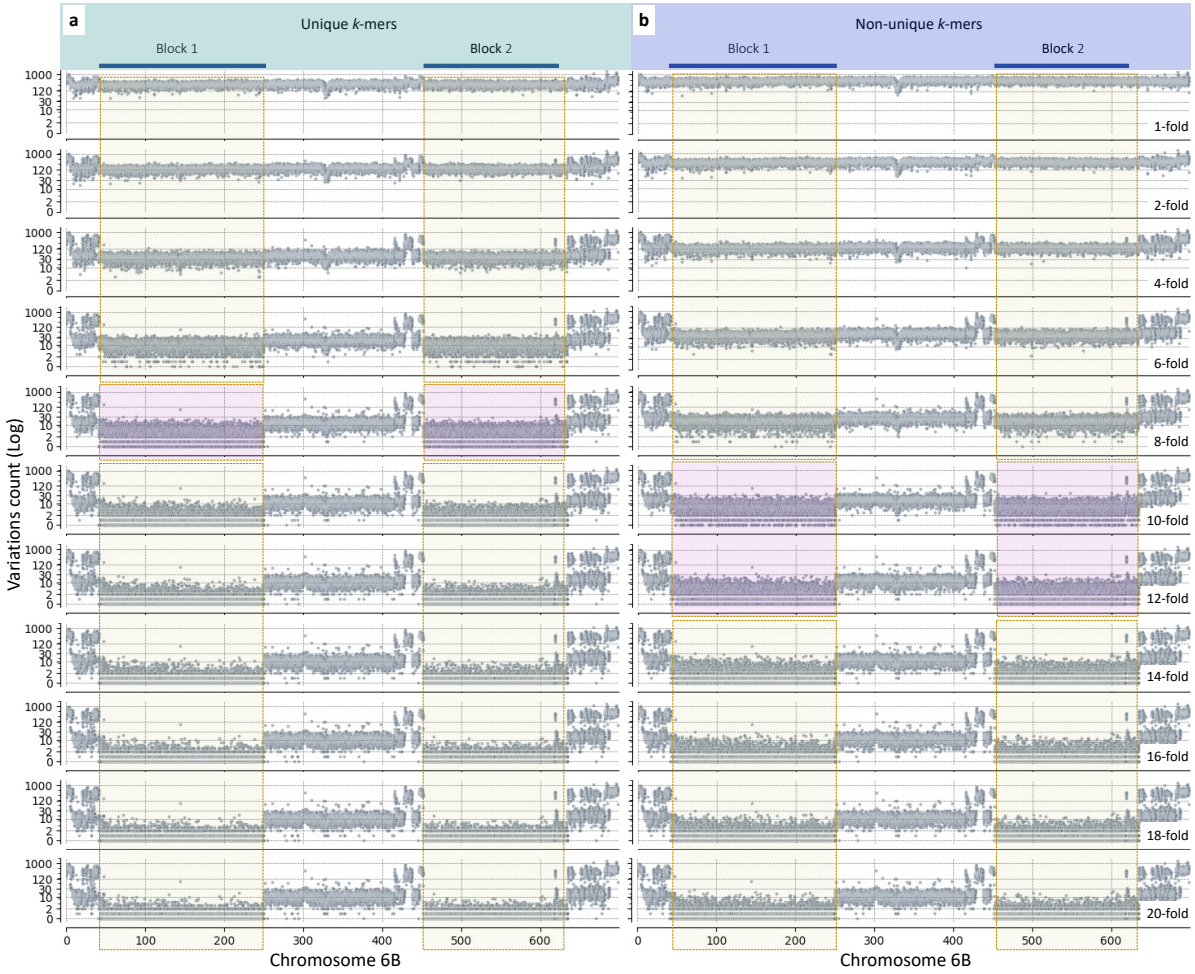


Fig. 2. 16. Removing unique *k*-mers impacts on the *variations* count captured by IBSpy.

Variations count at different sequencing coverage using Mattis as a reference vs Julius raw reads from 1 to 20-fold as a query. **a**) keeping unique *k*-mers and **b**) removing unique *k*-mers. The yellow boxes indicate the IBS regions between Mattis and Julius on chromosome 6B. Purple boxes in **a**) at 8-fold coverage including unique *k*-mers indicate the equivalence on level of *variations* detected in **b**) at 8 – 12-fold coverage removing unique *k*-mers as sign of loss of real sequence information in unique *k*-mers important to differentiate IBS regions.

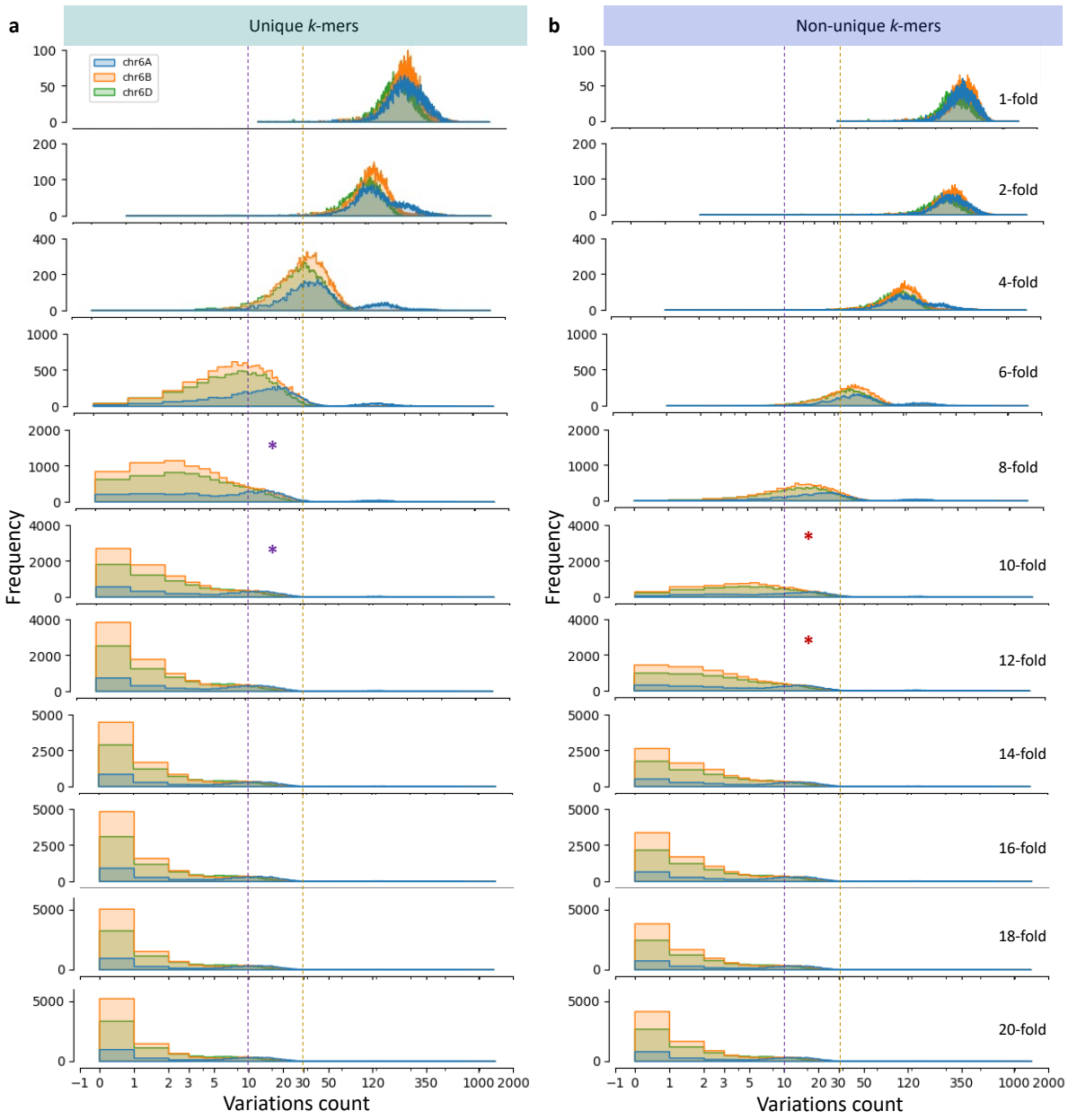


Fig. 2. 17. ***Variations*** **distributions of unique** ***k*****-mers vs non-unique** ***k*****-mers at different sequencing depth.**
a) histogram distribution of the *variations* counts in 50 Kbp window of the chromosome six triad. **b)** distribution removing unique *k*-mers. Purple lines indicate the <10 *variations* count threshold as IBS regions in pairwise comparison. Red line indicates the 30 variations count threshold. Purple asterisk in **a**) are equivalent to red asterisks at 8 and 10-fold keeping unique *k*-mers to 10 and 12-fold removing unique *k*-mers.

2.4.3.3. Sequencing platform

In our previous analysis we demonstrated that IBSpy detects *variations* using paired Illumina short reads of different sizes. We showed that the length of the reads impact on the *k*-mer size used,

and the coverage needed. In this analysis we tested different sequencing platforms and defined the coverage needed when long reads (~10 Kbp) are available. We employed, PacBio HiFi reads and Illumina 250 bp reads. As an example, we used Kariega, a hexaploid wheat assembly which has HiFi reads publicly available. We compared the results of using long reads versus using their genome assemblies. Our results demonstrated that IBSpy can efficiently differentiate *variations* levels employing different sequencing platforms. We demonstrated that with longer reads (e.g., 10 Kbp) less coverage is needed than with short reads (e.g., 150 bp).

In routine sequence assemblies, the sequence length impacts the quality of the assembly contigs generated (Athiyannan et al., 2022; Walkowiak et al., 2020). This is due to the overlapping *k*-mers among reads and the *k*-mer size used. In this analysis we tested and validated the effect of read length on the *variations* detected by IBSpy. In our previous analyses we tested IBSpy to detect *variations* at different *k*-mer lengths with chromosome assemblies. We also validated that our pipeline effectively detects variations with ~12x, ~250 bp raw reads. However, in a pilot test we observed an increase on the number of *variations* detected across the genome when using 150 bp reads. Since the WatSeq data and many publicly available re-sequencing projects use ~150 bp DNBSeq reads, in this analysis we aimed to investigate if the read length and *k*-mer size had an impact on the number of *variations* detected by IBSpy. In this analysis, a higher number of *variations* is a negative feature as it reduces our ability to distinguish between the IBS and near-IBS categories.

For this analysis we employed the publicly available HiFi reads of Kariega genome assembly (Athiyannan et al., 2022). To compare with our previous analysis, we employed sequencing depth from 2 to 20-fold coverage using either unique *k*-mers and removing unique *k*-mers. Our results demonstrates that with HiFi reads we would be able to define IBS regions using sequencing from 4 to 6-fold coverage when removing unique *k*-mers. This based on the level of *variations* count <10 in 50 Kbp continuous windows (**Fig. 2.18**). This would be equivalent to 2 to 4-fold coverage when keeping unique *k*-mers (**Fig S2.2 - S2.6**). A few improvements were detected after 8-folds with HiFi reads. At 20-fold HiFi reads showed almost better resolution than with chromosome-scale assemblies.

It's important to mention that keeping unique *k*-mers with HiFi reads is less computational demanding since there is less coverage needed and the sequencing errors at the end of each read are reduced. On the contrary, shotgun short reads generate much higher number of individual reads thus accumulating the sequencing errors from this high number of reads which are common at reads edges. Additionally, HiFi reads have much les error sequencing than the previous version of PacBio sequencing.

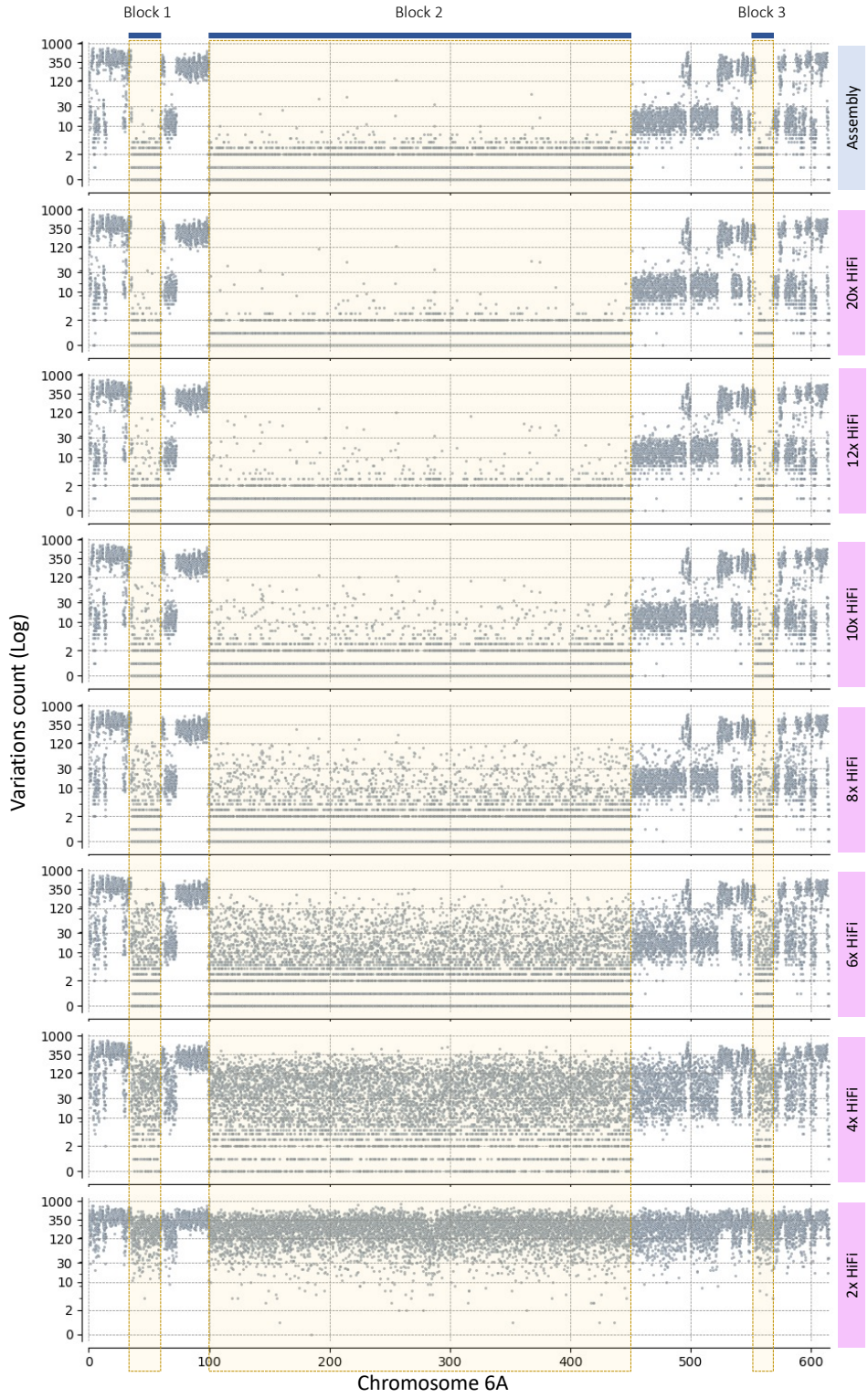


Fig. 2. 18. Long reads sequencing requires less coverage to efficiently detect IBSpy *variaitons*.

HiFi reads at different sequence coverage Mattis and Kariega (removing unique *k*-mer). For comparison with Illumina short reads, in this example we removed unique *k*-mers. Keeping the unique *k*-mers of this analysis is on **Fig S2.2 - S2.6**.

2.4.3.4. Scaffold level assembly as a query

In our previous analysis we explored using scaffold level assemblies to be used as a reference to detect variations and capture novel genome information not seen in the chromosome-scale assemblies. In this analysis we compared the level of resolution of scaffold assemblies as a query sample. We tested *k*-mers derived from the scaffold level assemblies of cultivars Robigus, Cadenza, Paragon, Claire, and Weebill and compared them using raw reads of these genotypes. As a common reference we use Mattis which shares 22.6%, 17.6%, 14.2%, 35% and 5.7% IBS haplotypes in 5 Mbp similarity with Robigus, Cadenza, Paragon, Claire and Weebill, respectively based on (Brinton et al., 2020). Our results showed that in all cases, using *k*-mers derived from raw reads outperformed the scaffold-level assemblies. In all cases the number of variations was higher in scaffold-level assemblies compared with raw reads. As a case study we will show the results of Claire vs Mattis (**Fig. 2.19**). The reason may be because scaffold assemblies include several misassembles or may lack many sequences lost during the assembly process from raw reads. These analyses suggest that it would be preferred to use raw reads than scaffold assemblies to detect IBSpy *variations*. Furthermore, as NGS and long-sequencing progress, scaffold level assemblies are becoming less common.

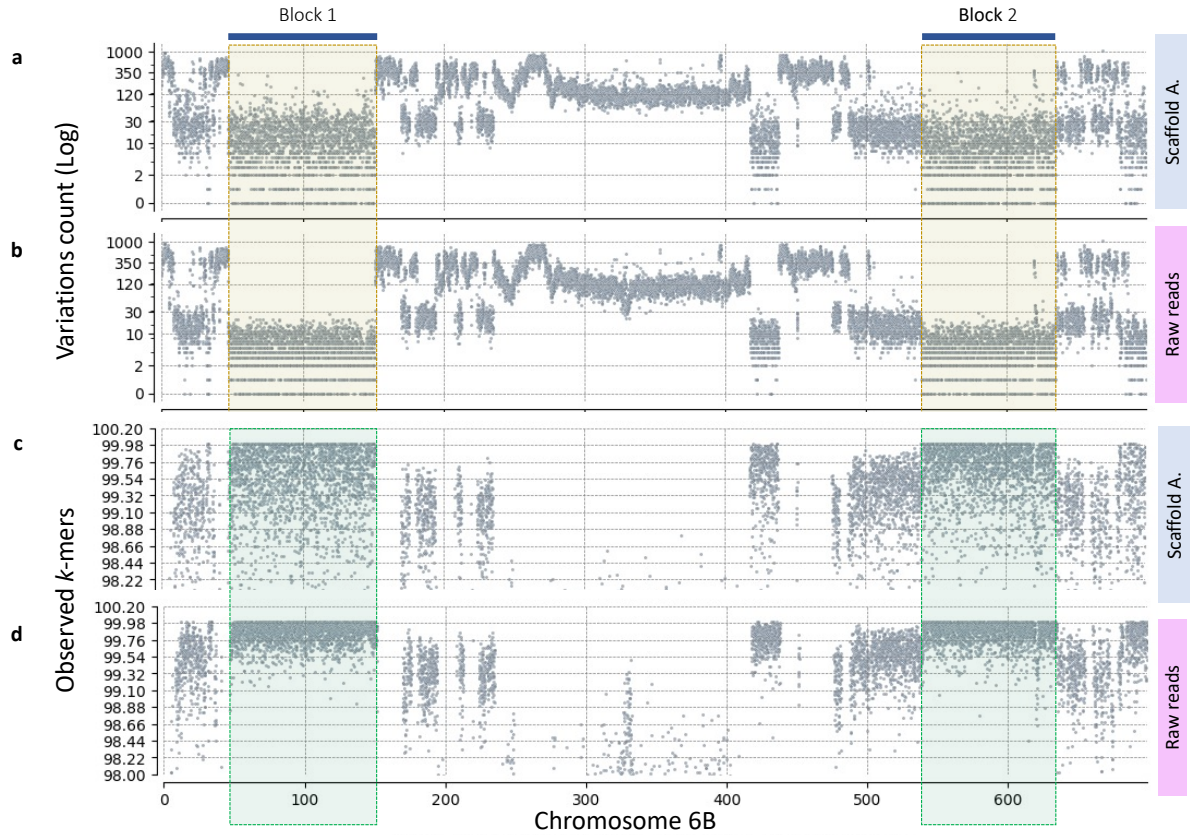


Fig. 2. 19. Raw reads overcome scaffold-scale assemblies to detect IBSpy *variations*.
Raw reads had 12-fold coverage. **a**) *variations* count comparison of scaffold assembly of Claire vs Mattis reference and **b**) raw reads of Claire vs Mattis reference. In **a)** and **b**) *y*-axis variations count. **c**) *observed_kmers* of scaffold level assembly and **d**) *observed_kmers* of raw reads against Mattis reference. In **c**) and **d**) percentage of observed *k*-mers. In all cases *x*-axis chromosome physical positions. In this analysis we removed unique *k*-mers from raw reads. These were *k*-mers derived from scaffold assemblies and the comparisons in from *k*-mers from raw reads.

2.4.4. Alignment to IBSpy variations comparison

IBSpy to sequence identity: After validating IBSpy to detect *variations* across the whole genome and selecting the appropriated parameters we wanted to define its equivalence with sequence identity. Therefore, to extrapolate IBSpy results into a common alignments and sequence identity between genotypes, in this analysis we compared the output of sequence alignments between fully assembled references to the IBSpy *variations* data. We analysed the published (Brinton et al., 2020) pairwise MUMmer alignments among the ten pangenome cultivars (ArinaLrFor, CS, Jagger, Julius, Reach Lancer, Landmark, Mace, Norin61, Stanley, Mattis) with the corresponding *variations* counts from IBSpy outputs to compare the sequence identity against *variations* counts. In total, there were 90 pairwise alignments analysed per subgenome. We analysed the data in 500 Kbp windows (a total of 890,793 windows for the A genome) and kept those windows with at least 60% breadth (coverage_prc) of alignment in the MUMmer output (77.8%; 693,102 500 Kbp windows for the A genome). For the B subgenome we analysed 1,197,901 windows in total and kept 814,519.0 (68%). For the D subgenome we analysed 854,385 in total and kept 751,509 87 (96%). The tables with the conversions for the A genome are in:

https://opendata.earlham.ac.uk/wheat/under_license/toronto/Uauy_2022-09-24_IBSpy_Triticum_monococcum_introgressions/data/nucmer_to_IBSpy_variations/.

Tables for the B and D subgenomes were generated later in the project and are available in our group upon request freely available.

For each 500 Kbp window, we had the average sequence identity between the pangenome reference and the other nine pangenome query samples (if over 60% breadth of alignment), alongside the IBSpy *variations* for the equivalent comparisons using the pangenome reference assembly and the *k*-mer database. We grouped the data based on the number of *variations* in increments of 10 variations per bin and determined the distribution of the sequence identity in each bin (**Fig. 2.20**).

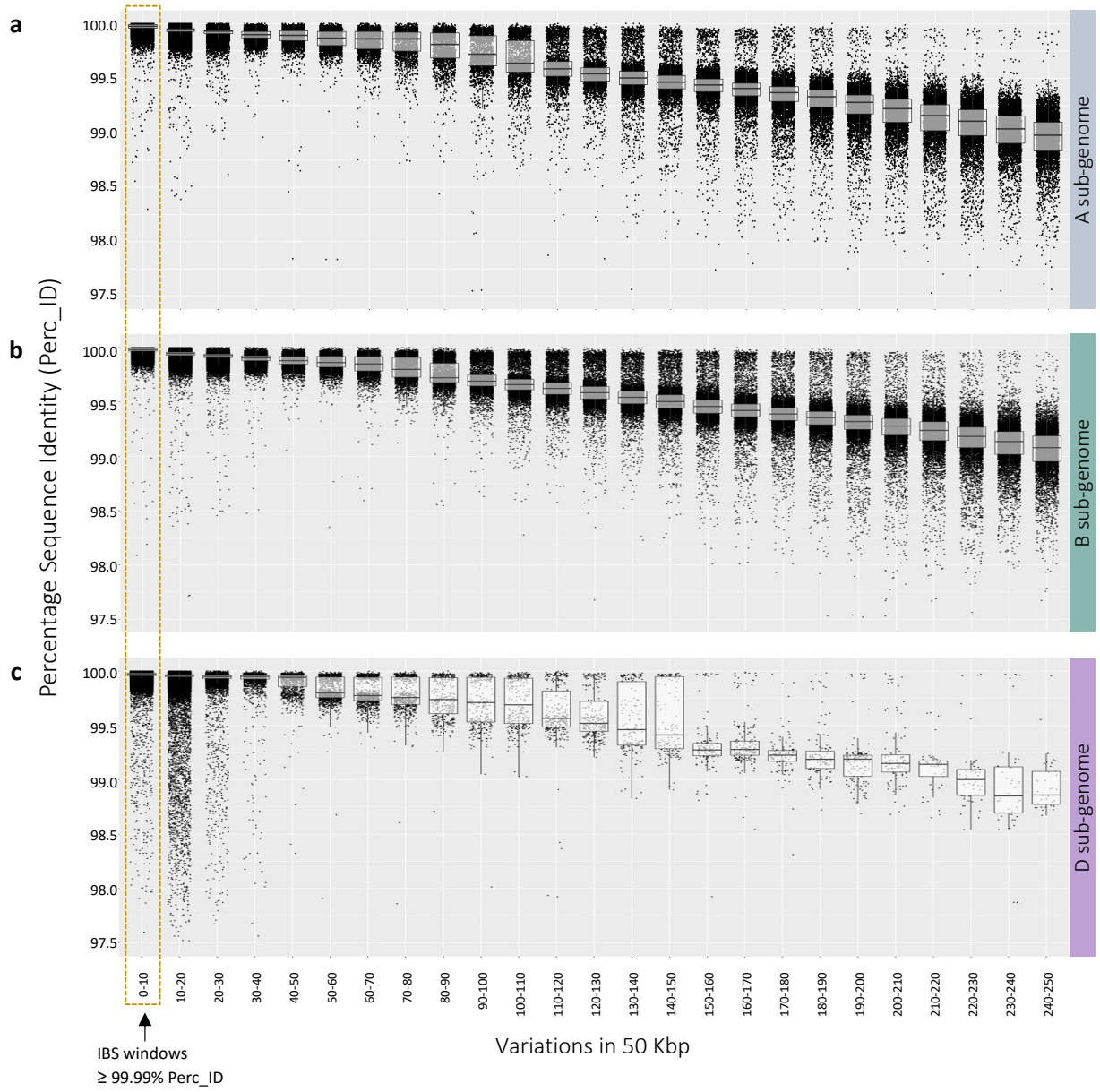


Fig. 2. 20. IBSpy *variations* to sequence similarity.

Relationship between IBSpy variations in bins of 10 and the percentage sequence identity of pairwise alignments in the A **a),** B **b),** and D **c)** genomes of hexaploid wheat. The data is filtered for alignments with at least 97.5% sequence identity across 60% of the 500 Kbp window and less than 250 *variations* per 50 Kbp. Percentage sequence identity ≥ 99.99% was considered as IBS in pairwise whole genome comparison in Brinton et al., 2020. This sequence similarity it is equivalent to have 0-10 variations in 50 Kbp windows in our approach (yellow box) and therefore our hypothetical IBS regions.

Whole genome vs whole genome: To compare multiple levels of variations among references we used the IBS regions defined in Brinton et al., 2020. These regions are IBS to have median > 99.99% sequence similarity in a pairwise comparisons. Sequences <99.99% but >99.95% would be considered as a near-IBS or sequences in the immediate wheat gene pool. Sequences <99.5% were considered as a more distant sequence similarity commonly found from wild relatives.

Using these criteria, we identified the following variations patterns:

a) IBSpy *variations* ≤ 10: Windows with ≤ 10 *variations* had median sequence similarity of ≥ 99.99% based on the pairwise alignment values (1 SNP every 5,000 bp). This would be equivalent to IBS in Brinton et al., 2020.
b) IBSpy *variations* between 10 and 30: This would be equivalent to two assemblies alignments of ≥ 99.95% (1 SNP per 1000 bp) and we would consider them related within the immediate gene pool (variations between cultivars and accessions). This would be equivalent to near-IBS regions in Brinton et al., 2020.
c) IBSpy *variations* between 30 and 120: Windows under this criterion would be equivalent of sequence alignments > 99.55%, or roughly 1 SNP every 225 bp.
d) IBSpy *variations* > 120: This would be equivalent to alignments with less than 99.55% sequence identity and are most likely reflective of comparisons between wheat and sequences derived from wild relative hybridizations in one of the samples (reference or the query sample).

For example, for the A subgenome using these classifications, we assigned 28.1%, 48.7%, 5.2% and 18% to the four categories outlined above in the ten pangenome references, respectively.

Subgenome IBSpy variations similarity: To quantify the sequence similarity among the wheat genome in terms of IBSpy *variations*, we split and ran IBSpy by subgenome using the corresponding reference. For example, subgenome A from Mattis reference vs subgenome B from Mattis reference. Multiple comparisons indicate that *variations* between the A vs B, A vs D, and B vs D genome comparisons have >350 *variations* in 50 Kbp window (**Fig. 2.21**). This is equivalent to sequence alignments of <98.4%, which is consistent with the comparison expectations between the homeologs (Ramírez-González et al., 2018). These results suggest that roughly the three subgenomes are equally distant to each other. Using these subgenome vs subgenome comparisons and using the A or B genomes as a reference, we detected a *variation* count drop roughly in chromosomes centres. We hypothesize that those drops in *variation* counts to be more highly conserved centromeric histone H3 variant (CENH3) sequences (**Fig. 2.21,** yellow bar). Surprisingly, we did not detect this low variation peak when using the D subgenome as a reference. A possible explanation for this could be that the D subgenome centromere regions were not

assembled as accurately as A and B sub-genomes and sequences of this regions went into the chromosome unanchored (chrUn) or lost during the assembly. Other explanation would be that the D centromeres are more distinct to the A and B subgenomes due to a more distant divergence.

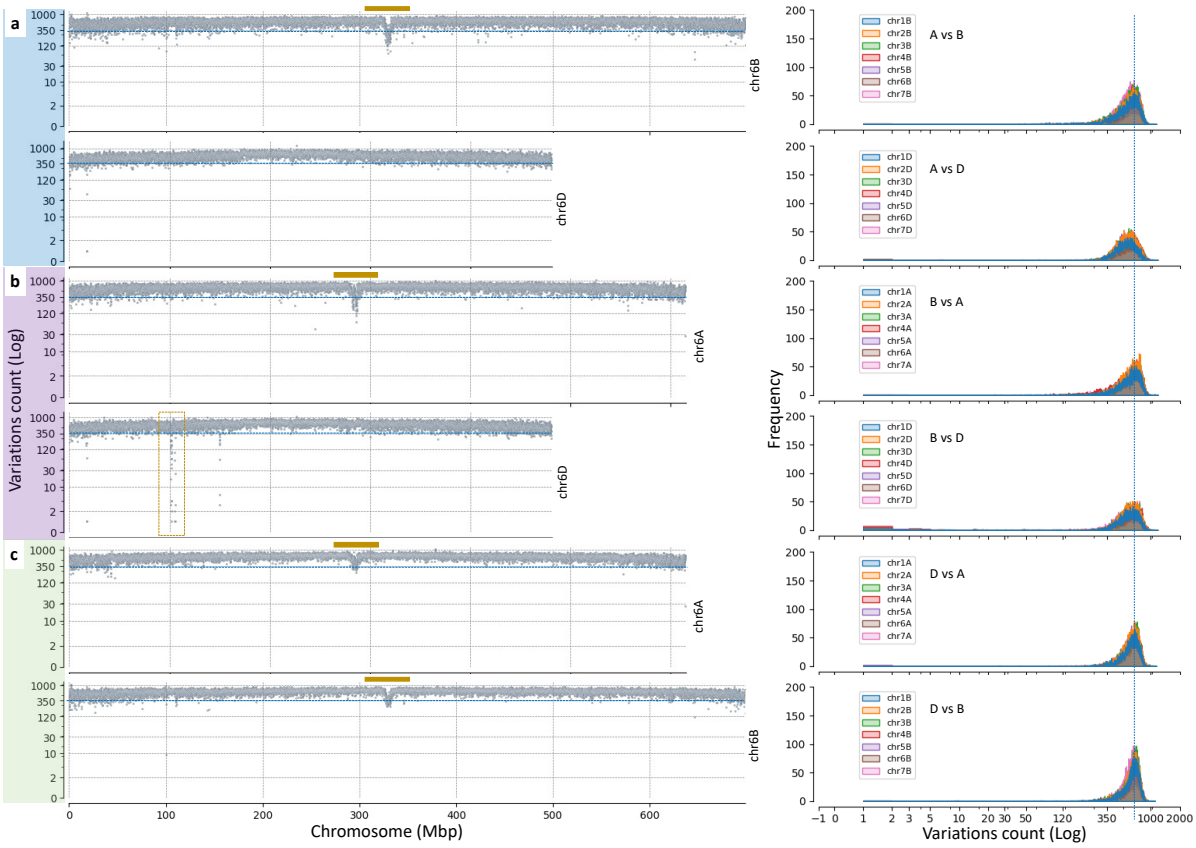


Fig. 2. 21. ***Variations*** **fingerprint among wheat homeologs subgenomes count (using genome of Mattis as an example).**

From top to bottom and from left to right in all cases; **a**) representative of the A subgenome (query) *variations* count against chr6B and chr6D (reference) across the chromosome physical positions. (right) *variations* histogram distribution of A subgenome vs B subgenome and D subgenome. **b)** B subgenome against chr6A and chr6D. (right) B subgenome (query) vs A subgenome (reference). **c**), D subgenome against chr6A and chr6B. (right) D subgenome (reference) vs A subgenome and B subgenome. Yellow bars indicate low variations count hypothesized as being the centromere regions in the corresponding subgenome used as a reference. Yellow box in **b**) indicates a region with low *variations* count which could be an indication of conserved region between D subgenome and B subgenome. Alternatively, it could indicate a misassembles from B subgenome misplaced in chr6D.

In summary, we detected variations at the same level as with chromosome-scale assemblies. This is crucial since genome assemblies requires much higher sequence coverage, cost, and are time consuming to generate. We defined 12-fold and 150 bp reads as an optimal coverage to differentiate among *variations* levels of sequence identity in 50 Kbp windows and differentiate IBS

regions among cultivars. No major improvements were seen after 12-fold to 20-fold coverage; however, 20-fold may be useful to differentiate more precisely between hypothetical IBS regions (<10 *variations*) and near-IBS (<30 *variations*) if the reads are available. We demonstrated that our method can be validated by comparing chromosome raw reads against the genome assembly of the same genotype. This quality control test demonstrated that *variations* <10 are most likely background noise across the genome using raw data without unique *k*-mers. Further validations of the ideal coverage can be tested in a telomere-to-telomere genome assemblies against their raw reads at high coverage. Coverage of 10-fold and 150 bp reads may also be used to detect IBS regions, although unique *k*-mers would need to be included in this case. Long reads (>10 Kbp) need less coverage (in the range of 4 to 8-fold) to differentiate among levels of variations and IBS regions. Keeping unique *k*-mers is less computing demanding in long reads HiFi data because the low base call error rate and low number of reads compared to shot gun short reads.

2.5. Discussion

2.5.1. The wheat *k*-mer landscape

Genome assembly research in wheat has advanced considerably in the last ten years. First, with the assembly and annotation of a chromosome-scale reference in 2018 (Appels et al., 2018) and two years later in 2020 with the pangenome project assembling 15 high quality genomes of important cultivars (Walkowiak et al., 2020). These projects have generated invaluable resources to the wheat community allowing to rapidly clone functional genes and use alternative genomes for QTL mapping and GWAS analysis based on modern cultivars instead of landraces (Walkowiak et al., 2020). Additionally, comparisons of the genome content among assemblies validated the repetitiveness of the wheat genome and the conservation of large haplotype blocks and near-IBS regions (Brinton et la 2020). This repetitiveness of a genome can be problematic for genome assemblies or alignments using short reads since they usually map to more than one position in the genome.

In this project we employed these 15 genomes to differentiate between genomic variations based on *k*-mers rather than the more common alignment-based SNP calling. As an entry point, we first explored the *k*-mer distributions of these 15 genomes when using different *k*-mer sizes. We demonstrated that 40% of the wheat genome in the 11 chromosome-level assemblies is represented as unique *k*-mers when using 31-mers. Similarly, (Chapman et al., 2015) in an early genome assembly of a hexaploid wheat found that 45% of the genome is represented in unique *k*-mers when using 51-mers. In our study, when using 51-mers of the 11 chromosome-scale

assemblies, we found that almost 60% of the genome is represented as unique *k*-mers as demonstrated in **Fig. 2.4**. Comparisons with the maize (*Zea mays*) and barley (*Hordeum vulgare*) genome assemblies in other studies indicates similar profiles (Hufford et al., 2021; Jayakodi et al., 2020; Liu et al., 2017). This may reflect the higher accuracy of modern algorithms and higher sequencing quality of the later assemblies compared to those ten years ago. Future assemblies may improve this accuracy as more genomes are becoming routinely assembles with long high-quality sequencing reads.

(Chapman et al., 2015) suggested that the three hexaploid wheat subgenomes are largely differentiated by 51-mers. In our analysis, we demonstrated that with 31-mers the three subgenomes are largely differentiated, and they show high differences on *variations* fingerprints using IBSpy as depicted in **Fig. 2.21.** We hypothesize that we can discriminate among subgenomes most likely because the conservation of the three genome copies is within genes and which out of those ~55-50 % are in triads. Outside these regions there is a high level of genome diversity accumulated during the wheat genome evolution and gene flow by natural and induced hybridizations from closely related wild relatives as previously suggested by (Dubcovsky & Dvorak, 2007).

2.5.2. Variations and methods to detect them.

Genomes maintain a large repertoire of genetic variations among individuals of a specie. After the NGS revolution, SNPs became the most prevalent and therefore the most common type of polymorphism employed in population genomics and phenotype-genotype associations studies in plants (Tibbs Cortes et al., 2021) and other organisms (Uffelmann et al., 2021). In wheat this is not the exception. For example, recent studies using a collection of 3,990 wheat accessions from 106 countries characterized genome wide SNPs using “exome capture” (F. He et al., 2019), “DArTseq” technology (Sansaloni et al., 2020), and SNPs arrays (Shorinola et al., 2022; Soleimani et al., 2022). In a different study using a collection of 298 bread Iranian wheat varieties and landraces and employing GBS, detected 46,862 SNPs (Rahimi et al., 2019). In their study they found that Iranian landraces harbours more genetic variation than elite varieties where the B and A genomes had more SNPs than the D subgenome in agreement with other studies. Using these set of variations, Rahimi et al., 2019 grouped all the accessions in three main groups. Group 1 having mainly modern varieties meanwhile group 2 and 3 containing mostly landraces. Similarly, using an array of 20K SNPs Cseh et al., 2021 defined six ancestral groups in a collection of winter wheat landraces from central Europe and identified that the European winter cultivars originated mainly from four

ancestral groups (Cseh et al., 2021). Although informative, SNPs often does not capture large structural variations and are bias towards gene content regions in the genome.

There are different approaches to detect genome variations (De Coster & Van Broeckhoven, 2019; Hwang et al., 2015; McKenna et al., 2010). At the time of writing this thesis, SNPs arrays, GBS, and capture probe sequencing are the most common and affordable platforms to detect variations between genotypes. However, in recent years WGS started to emerge in several important crops as the top choice for genotyping. As a result of the later, there is a vast amount of publicly available sequencing data which has generated a new challenge and opportunity on how to exploit these resources. When sequencing, either by genome reduction (GBS or capture-probes sequencing) or WGS, the most common pipeline is to align raw reads to a reference assembly with a subsequent step for variant calling. However, alignments-based methods introduce bias towards the reference used, intergenic regions with high polymorphisms often produce poor alignments (Armstrong et al., 2019), and represent a computer burden challenge in large genome sizes such as in wheat. In the SNP calling approach, a popular software to call *variations* after genome alignments is Freebayes (Garrison & Marth, 2012).

In this research we propose an alternative approach to the alignment methods based on *k*-mers called IBSpy. It is useful for large datasets to detect *variations,* condenses multiple types of variations from the whole genome (including intergenic regions) into one, and can be employed among genetically distant individuals such as landraces and wild relatives. Under the right parameters, and enough sequencing coverage, this method can differentiate between IBS and near-IBS regions with almost half of the computer burden compared to WGS alignments. Another advantage of IBSpy, is to easily combine and compare genome *variations* using multiple genome references. Something that was envisioned for alignment methods but again, constrained by computer burden in large genomes.

A method to align short reads simultaneously against multiple references was proposed by (Schneeberger et al., 2009). Later, variations calling based on multiple references as a graph representation started to emerge with the progress and low cost of NGS and recent advances on long-read sequencing technologies (Ebert et al., 2021; Garg et al., 2021). For example, PanGenie, is an algorithm that uses haplotype-resolved assemblies to detect variations by inference. This approach is based on *k*-mer and short reads sequences and it has the advantage to detect a wide type of genetic variations (Ebler et al., 2022).

Alternative methods to detect variations purely using *k*-mer started to emerge and those methods either involved a genome reference or directly comparing *k*-mer in within raw reads (Gaurav et al., 2022; Rahman et al., 2018; Voichek & Weigel, 2020). The disadvantage however is that the

context of the genome information is unknown and to define the region or the sequence there is an extra step to map *k*-mers back to a reference or by locally assembling raw reads, which is again, computing demanding and bias to genome assemblies. Similarly, a novel method called BayesTyper (Sibbesen et al., 2018) combines graph representation and *k*-mers. It uses a graph representation of a reference in conjunction to the pre-defined variations to genotype and capture structural variations by *k*-mers.

Apart of SNP variations, a few studies integrate structural variations (SV) into genomic analyses. This because most software are based on SNPs scores and the difficulties to detect these SV in the genome accurately (Jakubosky, Smith, et al., 2020). Consequently, several important functional variations other than SNPs, are overlooked during genomic analyses (Jakubosky, D'Antonio, et al., 2020; Voichek & Weigel, 2020). SVs are important drivers of crop domestications since they are associated in key agronomically important traits. However, SVs present a challenge to detect and integrate into genome analysis. To embrace this challenge alternative methods to the common SNP calling were developed to call SVs. For example, (Eggertsson et al., 2019) developed GrapTyper2, a method to detect structural variations in humans using a genome graph. This approach can detect either large or small SVs in large populations of individuals. With IBSpy we unify all types of variations into a single type including SVs. This can be beneficial for downstream analysis such as GWAS analysis since large SV are often linked to important phenotypes (Jakubosky, D'Antonio, et al., 2020).

In our study, we integrate different types of genome variations into a single score called *variations* or *observed k-mers* using presence/absence of *k*-mers. Using these *variations* calls, we translated the equivalence of our *variations* to sequence similarity of alignment methods. This was 99.99%, a SNP in 5 Kbp, in (Brinton et al., 2020) sequence identity to the equivalence of ~<10 *variations* in 50 Kbp windows in our study using IBSpy. These results are of importance because in our method we employed raw reads instead of genome assemblies. These *variations* encompass SNPs and InDels mainly. Copy number, duplications, and chromosome rearrangements are integrated as a single *variation* count and do not reflect the actual size or chromosome position of these large SV. Therefore, the large chromosome re-arrangements or duplications will not be detected intact as they are from raw reads. This because the presence/absence of the *k*-mer signal will be detected regardless of the chromosome physical position in the query sample.

(Brinton et al., 2020) suggested that near-IBS regions to have >99.95% and <99.99% sequence identity in a pairwise whole genome alignment comparison which is the expected diversity after 10,000 years of evolution and the mutation rate divergence which is 99.968%. They also suggested that <99.5% sequence identity may come from mor distant gene pools from wild wheat relatives.

In our analysis we found that *variations* <30 in 50 Kbp window, is the genetic variation expected after 10,000 years of genetic divergence. We detected two more levels of *variations* above >120 and 350 count that we hypothesize come from distant gene pools from wild relatives. A more in depth on this topic will be addressed in Chapter 4.

In the present thesis we mainly explored the IBSpy *variations* score. The *observed_kmers* and *kmer_distance* were added later to the IBSpy software and were compared against IBSpy *variatons*. In a pilot study, we noticed that *observed_kmers* gave similar result to the *variations* score allowing to differentiate between IBS and near-IBS regions. Similarly, these two scores allowed to compare introgressed regions and detect the hypothetical haplotype blocks similar to the observed in Brinton et al., 2020. The *observed_kmers* score measured similarity in percentage similar to the sequence identity score Brinton et al., 2020 as shown in figures 2.11, 2.13-14, and 2.19.

Due to time constrains, in the present study we decided to explore in dept the *variations* score. However, *observed_kmers* will be analysed further in a follow up project. On the other hand, the *kmer_distance* score resulted in a low resolution to differentiate among genome regions (IBS, near-IBS, and introgressions), therefore, we discarded its use early in the project. An additional pilot analysis was done combining two or two of the three scores to call haplotypes. However, we obtained similar haplotype calls with the constrain of an increase in computer burden. Haplotype calls will be addressed in Chapter 3.

2.5.3. Effect of raw reads on genome studies

Innovation in sequencing chemistry has revolutionised the analysis of genomes. Different sequencing platforms use specific chemistry reactions to read or predict nucleic acids and report specific genome sequence reads. Those differences are reflected in the sequence read length and base call quality prone to different error rates. Depending on the study objective and budget, users may decide to select a particular sequencing platform to use or to combine two or more methods. For example, long reads are preferred for genome assemblies (Athiyannan et al., 2022; Aury et al., 2022) meanwhile short reads are common in population genetics or GWAS studies (Zhao et al., 2022; Zhou et al., 2021; Zhou et al., 2020).

When detecting variations from raw reads, the type of reads, sequencing platform, read length, and depth can impact on genome variation types identified, number, and accuracy. Additionally, conditional to the organism, the sequencing coverage needed to call variations differs. For example, to call SNP variants in the wheat genome usually ~10-fold is required. Less than 10-fold

it is also employed with constrains in SNP calling accuracy. Alternatively, methods to detect SNPs variations with ~<1-fold and imputations are also becoming popular to reduce costs to genotype large populations (Adhikari et al., 2022; Bradbury et al., 2022).

In this thesis chapter, we evaluated the IBSpy software under different parameters and types of sequencing and described the main features to fine tune accordingly to the objective of the analysis which was to differentiate IBS vs near-IBS and introgression regions. These parameters were evaluated for hexaploid wheat, but a similar road map can be followed to determine the best parameters in other organisms. We measured different parameters of raw reads and described some of the most important factors influencing IBSpy results. As shown in other studies, there is a trade-off among read depth coverage, sequence length, computing resources, and costs. In our analysis we defined that with sequencing length of 150 bp reads the optimal coverage ranges from 12 to 15-fold depending on the quality of the reads and removing unique *k*-mers. This coverage is similar to studies calling SNP variations with the routine alignment methods with the advantage of a reduced computer burden in our method.

Similarly, depending on the objective of the project and species, a specific sequencing depth may be required. For example, to assembly a wheat genome a coverage of ~30-fold using PacBio HiFi reads is usually required which has on average read length of 15.7 Kbp (Athiyannan et al., 2022). On the other hand, using Illumina short reads at 250 bp, >150x is commonly used (Walkowiak et al., 2020). These requirements are similar in other cereals such as maize with ~22-fold and 15.6 Kbp average length using PacBio HiFi reads (Hon et al., 2020). Similar sequencing coverage and length were required for barley (Jayakodi et al., 2020) and in cucurbitaceous like watermelon (Deng et al., 2022).

Another parameter to consider in genome studies is the quality of the raw reads. This is of particular importance in genome assemblies. A pre-step to reduce the sequencing error in raw reads is to remove unique *k*-mers to avoid assembly errors. In our approach, IBSpy parses the presence/absence of the *k*-mers in the query sample from raw reads and compare it with the presence of those *k*-mer in a genome assembly. Therefore, unique *k*-mers from sequenceing errors in raw reads are mostly unnoticeable and have minor impact on *variations* detection. In this analysis we tested the impact on IBSpy by keeping or removing unique *k*-mers. We demonstrated that unique *k*-mers are informative in our analysis and therefore, the optimal coverage when keeping unique *k*-mers ranged from 10 to 12-fold for 150 bp read length.

Similar to genome assemblies, the length of the reads impact on IBSpy to detected *variations*. For example, we demonstrated that an optimal coverage for 150 bp reads is in the range of 10 to 12-fold as mentioned before. However, for 250 bp reads the optimal coverage ranges from 8 to 10-

fold keeping unique *k*-mers. Using HiFi reads, which have on average ~10 Kbp read length, the optimal coverage ranges from 4 to 6-fold keeping unique *k*-mer while removing unique *k*-mers range from 6 to 8-fold. Altogether these analysis revealed information to consider when using IBSpy to detect *variations* based on and raw reads. Depending on the aim of the project, computer infrastructure, and budget, users may select a trade-of on storage and sequencing costs.

In this analysis we provide a new approach to count variations across the genome comparing any genotype having 12-fold coverage and short 150 bp raw reads length. This method focuses on a new score to count *variations* in 50 Kbp windows. Comparisons with other methods are feasible providing a sensitivity and specificity benchmarking against the routine SNP calling approach for example. In our analysis we provided an example of the equivalence of our *variations* scores against the sequence identity comparing whole genome mummer alignments as described in section 2.4.4. Our results demonstrates that we can differentiate genome regions among cultivars and that IBS regions detected by whole genome sequence identity analysis are equivalent to have ≤ 10 variations in 50 Kbp consecutive windows as shown in Fig. 2.11. Further evaluations would compare the number of SNP detected in 50 Kbp by the routine mapping analysis. However, since our method integrates InDels into the count as a *variation* this would need to be considered when analysing. A sensitivity and specificity metric would consider those SNPs as a ground truth and IBSpy *variations* count as a test to detect an equivalent level of variation. By the time of writing this thesis a VCF file integrating those SNPs are in progress to be publicly available and will allow further benchmarking including the haplotype calls described in Chapter 3.

3. IBSpy: a multi-genome approach to call haplotypes in wheat.

In this chapter Dr. Ricardo Ramirez-Gonzalez generated the files of the syntenic windows among the 11 chromosome-scale pangenome assemblies using the mummer alignments and gene projections generated by Dr. Jemima Brinton described in Brinton et al., 2020 and in this chapter in Fig. 3. 14.

We thank members of the Uauy Lab by generating and providing the spike morphology data for haplotype GWAS analysis in section 3.4.2.1 and 3.4.2.2. In particular, we thank Anna Backhaus, Andy Chen, and James Simmons who led the project for sowing and collecting the Watkins phenotypic data within our group.

We thank Simon Berry from Limagrain for providing the rust phenotypic data as part of the collaboration with the WatSeq project for the haplotype GWAS analysis described in 3.4.2.3.

In section "3.4.2.4, Wheat Blast" of this chapter, Dr. Paul Nicholson and Tom O'Hara generated the phenotype scores as part of a collaboration. At the time of writing this thesis, the results of this collaboration are in progress for publication using the IBSpy *variations* and haplotype based GWAS from a subset of the Watkins collection for wheat blast resistance: "The wheat powdery mildew resistance gene *Pm4* also confers resistance to wheat blast". (O'Hara et al.).

3.1. Chapter summary

In this chapter, we defined a method to call haplotypes using the *variations* detected by *k*-mers described in Chapter 2. We tested different algorithms to predict haplotypes and describe the advantages and disadvantages of each. We built a database using collection of more than >1,000 genotypes including wild wheat relatives, landraces, and modern wheat cultivars. We used this database to track haplotypes from landraces into modern cultivars. The method has been validated to call haplotypes at 1 Mbp resolution using multi-genome assemblies information. This parameter can be adjusted for different windows size and/or sliding windows, however, care must be had to avoid losing accuracy. Our results suggest that large haplotype blocks were brought into modern cultivars from landraces and those blocks have been maintained in modern elite cultivars through >80 years of breeding. We tracked haplotype blocks inherited from parents and relatives from three generations and these haplotypes are consistent with publicly available pedigree information. We successfully validated the haplotypes calls using the analysis of the ten pangenome assemblies in previous reports in wheat and a collection of *Ae. tauschii* accessions. Novel unexploited haplotypes were identified in landraces. As expected, we identified a higher number of haplotypes in telomeric regions than in centromeric regions where haplotypes blocks

were extended in physical size. Using phenotypic information, we conducted a haplotype GWAS analysis and detected genome regions with associations to disease (wheat blast, yellow rust) and quantitative traits (spikelet number, max floret number) at 1 Mbp resolution. Using wheat blast as a case study, we identified significant associations using reference genomes which did not carry the resistance haplotype. Together, these results demonstrated the utility of our haplotype calls using an alternative approach to the conventional methods using *k*-mers instead of alignments and SNPs methods. This method complements with the already stablished approaches and has the advantages to integrate a pangenome informed haplotype calls which are useful in genome-phenotype associations studies to capture genome regions private to each assembly and can handle large genome information from WGS data and large genomes.

3.2. Introduction

3.2.1. Methods for Haplotype building

With advances in sequencing technology, high density genotyping markers are now available for several important crops at low cost which allows to genotype thousands of individuals (Rasheed et al., 2017). These high-density markers are becoming common to build haplotype maps in crops and there are different methods to define them and vary depending on the data available and the purpose of the study. Haplotypes can be as short as two adjacent SNPs, large chromosome blocks regions (Brinton et al., 2020), or include the whole chromosome (Garg, 2021).

Common approaches to define haplotypes involve using SNPs and by LD in population studies (Pritchard & Przeworski, 2001). Historically, SNPs arrays were predominantly used to reconstruct haplotypes based on a population basis from multiple individuals and these methods relied mainly on LD (Balfourier et al., 2019; Cseh et al., 2021). With the relatively cheap sequencing, modern methods to reconstruct haplotypes based on raw reads are now possible. Short reads capture limited information of haplotypes compared to long read sequencing (Kronenberg et al., 2021). Therefore, using short reads, it is common to map to a genome reference followed by SNP calling and haplotype reconstruction (Garrison & Marth, 2012). However, with sufficient sequencing depth of short reads, haplotypes can also be reconstructed by local assembling (Gaurav et al., 2022; Voichek & Weigel, 2020). More recently, with advances in third generation sequencing such as Oxford Nanopore Technologies (Wenger et al., 2019) and PacBio long-read sequencing (Y. Wang et al., 2021), chromosome-scale haplotypes are now possible. For example, the HiFi technology generate sequences of 10-20 Kbp with accuracy >99% at great scale with the recent technological improvements (e.g., https://www.pacb.com/revio/). These advances are particular of importance and useful to define haplotypes in heterozygous organisms or polyploids and will become prevalent in the coming years.

While a maximum of two haplotypes would be present in a child from two homozygous parents, multiple haplotypes will be present at the population level at a given chromosome region. Thus, depending on the type of reads and objective of the study, different methods are preferred. Selecting linked SNP based on LD has been one of the most widely employed method to define haplotypes in plants (Al Bkhetan et al., 2019; Caldwell et al., 2006; Hyten et al., 2007; Kim et al., 2007).

3.2.2. Crop haplotype maps

Haplotype maps can be defined as the characterization of common genetic variations in an organism or a population in the context of genome recombination hotspots and linkage disequilibrium (Altshuler et al., 2005). The first haplotype map in plants was reported in 2007 (Clark et al., 2007) in the model plant *Ae. thaliana*. With novel methods to generate high-density markers, variation characterization of thousands of individuals rapidly expanded and the availability of haplotype maps of several crops have been released with pivotal information for genetic studies (Bukowski et al., 2017; Gore et al., 2009).

For example, the maize haplotype map revealed that high genetic diversity exists among important maize cultivars. These highly divergent haplotypes and heterozygosity has been mainly influenced by the high recombination rate of this crop (Gore et al., 2009). The rice haplotype study identified regions in the genome most likely driving its domestication and demonstrated high level of diversity in wild populations (Huang et al., 2012). The tomato haplotype map revealed LD differences between wild tomatoes and domesticated tomatoes suggesting that modern varieties specialization derived from market preferences lead to the genetic differences in this crop (Robbins et al., 2010). Similarly, the characterization of the genetic variations in the barley genome and the pangenome assemblies from multiple accessions identified polymorphisms that suggest the geographic expansion and breeding drivers of specific structural variations (Jayakodi et al., 2020).

3.2.3. Wheat haplotype map

Despite its importance for global food production, until recently, wheat has lagged on genomic studies due to the size and complexity of its genome (e.g., polyploidy, high percentage of repeats). The wheat haplotype map from a reduced representation of its genome using exome capture, was made available in 2015 (Jordan et al., 2015). In the same year a consensus map of wheat using RIL populations was made available by GBS (Li et al., 2015). Later, with the release of the first complete chromosome-level assembly and annotation in 2018, an extensive haplotype map using targeted re-sequencing of 890 worldwide hexaploid, tetraploid wheats, and wild relatives, was release (F. He et al., 2019). In addition, WGS projects with high density markers are becoming prevalent and high-density maps are now possible also for this crop. With the release of ten additional whole chromosome assemblies in 2020 (usually referred as the wheat pangenome), a further characterization of large structural variations in wheat have been unravelled (Walkowiak et al., 2020). Together, all these events will help to further explore the wheat genome impacting on

genetic population studies and breeding allowing to characterize large germplasm collections, which remained until recently, unexplored in germplasm banks (Vikram et al., 2016; Wingen et al., 2014).

3.2.4. The Watkins haplotype diversity

As briefly introduced in Chapter 1, the wheat Watkins collection is a reservoir of genetic diversity for wheat in the UK and worldwide. Until recently this collection has remained mostly unexplored at the genome level due to the complexity and size of the wheat genome. An initial genetic diversity characterization was carried out using a reduced number of microsatellite markers (Wingen et al., 2014). Using this genetic information, 119 accession representing most of the diversity of the collection was defined as a core set. Employing this core set, a nested association mapping (NAM) population of 60 biparental populations was developed using the spring cultivar "Paragon" as the common parent. These populations were genotypes with ~200 SNP markers and were used to create a consensus map for wheat. Characterization of these maps corroborated wheat as having high collinearity, but translocations events were also common in the multiple bi-parental populations (Wingen et al., 2017). After this analysis, no further characterization has been done at the whole genome level of the collection.

Recently, in a case study to validate the use of the already available genotyping SNP array data, Brinton et al., 2020 (Brinton et al., 2020) used this data to assign haplotypes to the Watkins collection in a specific QTL region across chromosome 6A. Although informative, the array was not able to differentiate several haplotypes in the panel (which were known to be different based on the pangenome assemblies). To increase the resolution to differentiate among haplotypes, Brinton et al., 2020 designed 17 haplotype-informed markers based on the pangenome assemblies. These markers identified common haplotypes to the Watkins and modern wheats and a wide array of Watkins specific haplotypes which were absent in modern cultivars. Importantly, private Watkins haplotypes were associated with positive effects for agronomically important traits. In conjunctions, these studies revealed the potential of landrace collections and the limitations of current genotyping platforms to capture novel variations. Therefore, a whole genome haplotype analysis of the collection would be valuable to identify novel variations and facilitate the integration into modern varieties.

In summary, in this chapter we addressed the challenge to define IBS regions and build a novel method to call haplotypes using IBSpy *variations* count based on *k*-mers from WGS raw reads at ~12-fold coverage. We explored different methods to define IBS across the whole genome,

multiple genome references, and more than >1,000 landraces and modern wheat accessions. We characterized the wheat genome based on these haplotypes and propose a global haplotype database along with the "*variations fingerprint*" defined in Chapter 2 for each of the accessions that can be used to put in context of any new genotype including hexaploid, tetraploid, or wild wheat relatives. We give case study examples of the haplotype-phenotype analysis for qualitative and quantitative traits and highlight the usefulness of the Watkins collections to deliver unexploited genomic regions into modern wheats and to track historically used haplotypes into modern wheats.

3.3. Methods

3.3.1. Germplasm

To evaluate the haplotypes, we created different groups of sub samples. Each group is summarized in Supplemental Table S3.1. Sub-grouping was made to answers different questions regarding the pipeline and how it was influenced by the number of genotypes, types of samples, and reads used in each case.

3.3.2. Gaussian Mixture Models (GMM)

We employed the GMM which implements the Expectation Maximization algorithm from the API sklearn.mixture.GaussianMixture package in Python. We tested the model using IBSpy raw *variations* counts and the Log transformation of the data. The code to automatically call IBS and non-IBS using the GMM model is under the IBSpy software, and it is described in https://github.com/Uauy-Lab/IBSpy section IBSplot.

3.3.3. Precision and Recall

We used the IBS and non-IBS regions defined in (Brinton et al., 2020) as our positive control and compared against the IBS regions defined by IBSpy. As output, we had four categories:

a) True Positive (TP) when the two methods agree in an IBS region,
b) False Positive (FP) when only IBSpy calls a region as IBS, but not Brinton et al., 2020.
c) True Negative (TN) when the two methods agree in non-IBS regions, and
d) False Negative (FN) when IBSpy calls a region non-IBS but Brinton et al., 2020 called this region as IBS (Fig. 3.1).

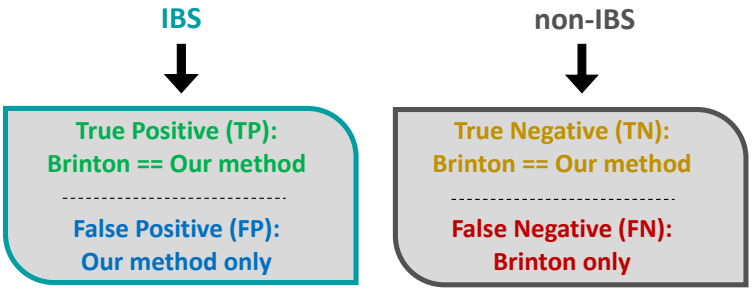


Fig. 3. 1. Precision-Recall expected outputs comparing IBSpy vs Brinton et al., 2020 IBS regions.

3.3.4. Clustering algorithms

For the hierarchically-clustering heatmap we used the API described in: https://seaborn.pydata.org/generated/seaborn.clustermap.html. We used the default parameters, which employs the '*Euclidean'* distance metric to calculate the spatial distance among query samples. The linkage method to calculate cluster uses the '*average'* method also called UPGMA (unweighted pair group method with arithmetic mean) (Sokal, 1958).

To calculate the Affinity Propagation (AP) haplotypes, we used the API described: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

We used the default parameter and under different "damping factors" (dmp) as follow:

a) high_dmp: 0.5, 0.6, 0.7, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97
b) inter_dmp: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9
c) low_dmp: 0.5, 0.52, 0.54, 0.56, 0.58, 0.6, 0.65, 0.7

To calculate the Silhouette Coefficient score (SC) we used the API in: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. The algorithm was described in (Rousseeuw, 1987). The implementation of the complete pipeline is in the final version of IBSpy (v.0.4.6) in https://github.com/Uauy-Lab/IBSpy.

3.3.5. Phenotypic data

We used phenotypic data collected from 2020 and 2021 in Norwich UK for spike morphology. Briefly, 1 m rows of the Watkin collection were grown at the Church Farm Experimental Station in Bawburgh, Norfolk. At harvest, 10 spikes from each accession were harvested and total spikelets and grains per spikelet (termed maximum number of florets) were counted. Rust phenotypic data

was provided by Limagrain based on pathology assays using *Puccinia striiformis fsp. triticeae* isolates defined as Pink and Red as described in (Hubbard et al., 2015). Wheat blast data was provided by Tom O'Hara (JIC) and was from leaf inoculation assays using a *Magnaporthe oryzae Triticum* Super Race Avirulence (SRA) isolates.

3.3.6. hapGWAS

We adjusted the AP numerical haplotype calls by window to presence/absence and assigned unique names. E.g., if we had 10 haplotypes in a chromosome physical position from 0 to 1 Mbp windows (1 Mbp window), we transformed those haplotype calls to presence "1" or absence "0" of all genotypes in the WatSeq collection to have it (1) or not have it (0). These ten haplotypes had a unique name and had the same chromosome physical position from 0 to 1 Mbp. The next window had the position 1 to 2 Mbp and so forth. Then, we adjusted the kGWAS (https://github.com/wheatgenetics/owwc/tree/master/kGWAS) described in (Gaurav et al., 2022) to run associations with our haplotypes using a presence/absence matrix which uses a Generalized Linear Model (GLM). PCA dimensions to account for population structure was constructed by using a VCF file generated by our collaborators using the same set of samples (Cheng et al., under revision) by mapping Illumina raw reads against CS reference (RefSeq v1.0). We ran hapGWAS with the default parameters as defined in (Gaurav et al., 2022).

3.4. Results

3.4.1. Calling haplotypes

In this section we determined a model to categorize IBS regions and comment on the pros and cons of the different methods tested. To evaluate each approach, we did benchmark across the models and chromosome positions under different parameters and evaluated their performance by comparing to the haplotypes called by (Brinton et al., 2020). In this section we aimed to identify regions across the genome where our pipeline confidentially calls haplotypes or where errors may occur and select the parameters to build a final haplotype database.

3.4.1.1. Binary category; IBS and non-IBS

In our previous analysis when counting *variations* in 50 Kbp we observed genome regions with low *variations* counts <10 that matched the IBS (haplotypes) regions defined in (Brinton et al., 2020)

in multiple genome assemblies. These *variations* count data had two or three differentiable mixture distribution shapes on histogram plots of the data. We hypothesised that the low *variation* counts distribution is an indication of IBS regions in the genome between a query sample and a reference for IBSpy data.

Exploring the data using a scatter plot across chromosome physical positions, four levels of *variations* categories were detected: <10, <30, <120, and >350 variation counts. We hypothesize that the category <30 is the diversity between elite hexaploid wheat vs elite or vs elite and landraces (1 SNP in 2,000 bp) *variations*. Windows with *variations* values >120 threshold are likely wheat hybridizations with distant relatives including tetraploid and diploid ancestors. Based on previous observations we also hypothesize that <10 *variations* are IBS regions. To validate this hypothesis, we classified IBS and non-IBS regions in pairwise comparisons using the pangenome assemblies and compared them to IBS regions defined by chromosome alignments in (Brinton et al., 2020).

To automatically define IBS regions, we used the Gaussian Mixture Models (GMM) for data categorization into IBS or non-IBS against the genome references. We explored window size of 50 Kbp and ranging from 100 to 1000 Kbp. We found that using this model (GMM) we could differentiate with the least error using 500 Kbp in pairwise comparisons. As an example, we show the histogram distribution of the *variations* count in 500 Kbp window in (**Fig. 3.2a**). In this example we could differentiate two main distributions of the data. The *variations* count in 500 Kbp window was similar to when using 50 Kbp window. With 500 Kbp windows, we observed a clearer separation of the data with the constrain of losing resolution. In these resultis, the low *variations* count matched precisely the IBS regions defined in (Brinton et al., 2020) between Mattis and Claire chr6A as a case study (**Fig. 3.2b**).

The histogram distribution of the data defined by the GMM model as an IBS grouped <50 *variations* count in 500 Kbp window as an IBS region. This is equivalent to have <10 *variations* count in 50 Kbp window (**Fig. 3.3a**). When plotting the IBS windows defined by the GMM model across the chromosome physical position we detected that they matched with the IBS block defined by Brinton et al., 2020 (**Fig. 3.3b**). Using these windows defined by the GMM model we reconstructed the entire haplotype block as depicted in (**Fig. 3.3c**).

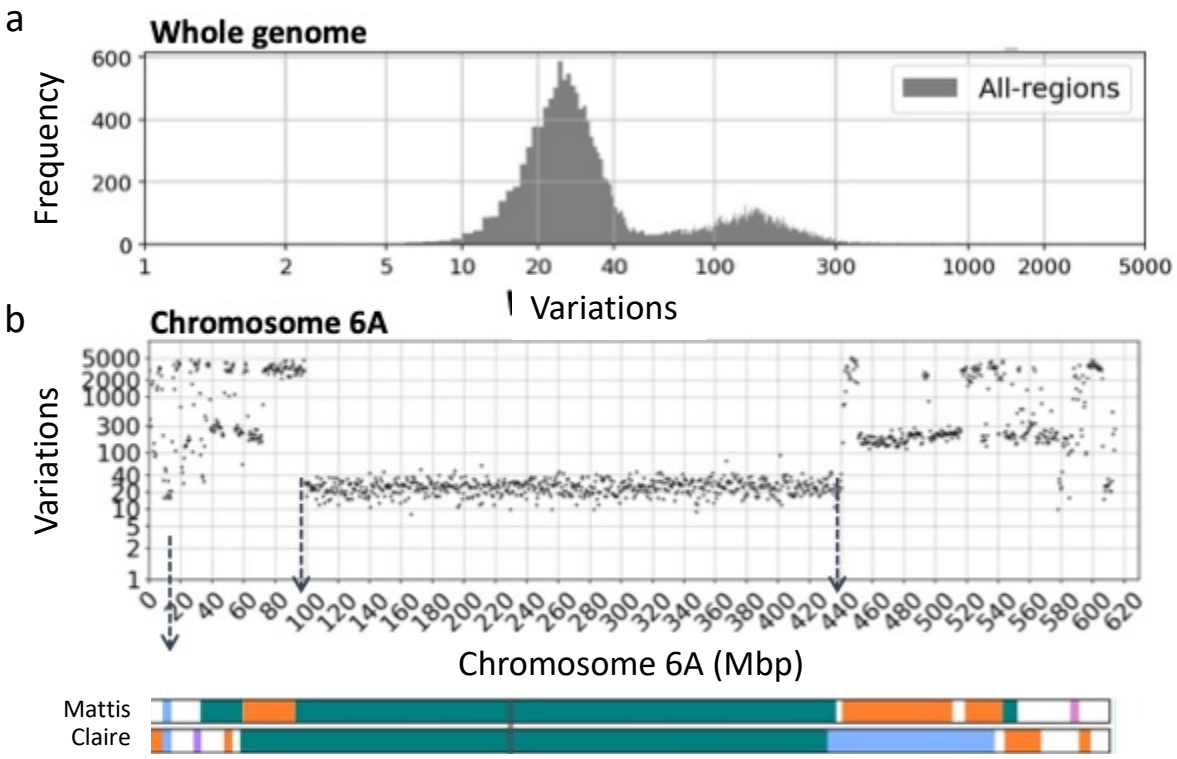


Fig. 3. 2. Low *variations* intervals match previous defined IBS regions (Brinton *et al.,* 2020).

Mattis (reference) vs Claire (query) *variations* count in 500 Kbp window. **a**) histogram distribution in 500 Kbp window. **b**). *Variations* count across chromosome 6A physical positions. The arrows indicate the IBS regions defined in Brinton et al., 2020 as depicted in green bar in **b** (bottom).

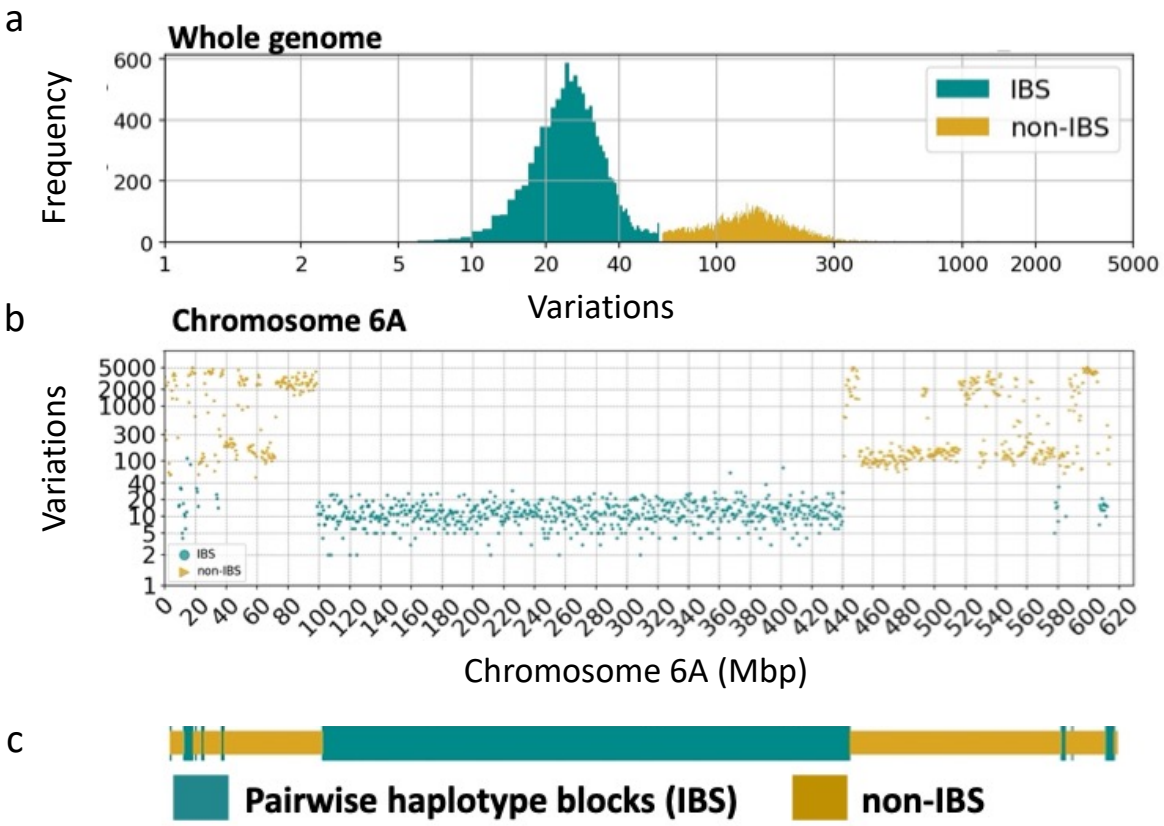


Fig. 3. 3. Haplotype blocks generated by the GMM model using the *variations* count score.

In this example we show IBS and non-IBS regions on chromosome 6A of Mattis as a reference and Claire as a query. **a**) histogram distribution in 500 Kbp window coloured as IBS (cyan) regions and non-IBS regions (yellow) defined by the GMM model. **b**). *Variations* count across chromosome 6A physical positions by colour in each category. **c**) final reconstruction of the haplotype blocks defined by the GMM model.

3.4.1.2. Benchmarking (Precision and Recall)

To validate the defined IBS regions based on GMM, we used the Precision-Recall metric (Saito & Rehmsmeier, 2015). In (Brinton et al., 2020), IBS regions were called based on 5 Mbp windows. In our method we selected 500 Kbp. Regions where the two methods called an IBS block were defined as True Positives (**TP**), regions where only IBSpy called IBS were defined as False Positives (**FP**), regions called as non-IBS by the two methods were True Negative (**TN**) and regions called IBS only by alignment-based method were defined as False Negative (**FN**). Using these categories, overall, we recall ~80% of haplotypes defined by alignments at ~80% accuracy. Using those categories, we plotted each of the categories across the chromosome physical positions (**Fig. 3.4**). Using the whole genome of Mattis vs Claire, we observed that most of the FP were located at haplotype block edges. Considering that Brinton et al., 2020 used 5 Mbp to define IBS regions and

the GMM model uses 500 Kbp, ten times more resolution, this is consistent. In addition to the haplotype edges differences, we detected FP regions in the middle of extended large blocks. These FP blocks are most likely because Brinton et al., 2020 used the gene content of the scaffold assemblies in Claire as a criterion to assign IBS blocks, therefore, it could be that the FP regions found within blocks does not contain genes and were not included in the Brinton et al., 2020 analysis (**Fig. 3.4**).

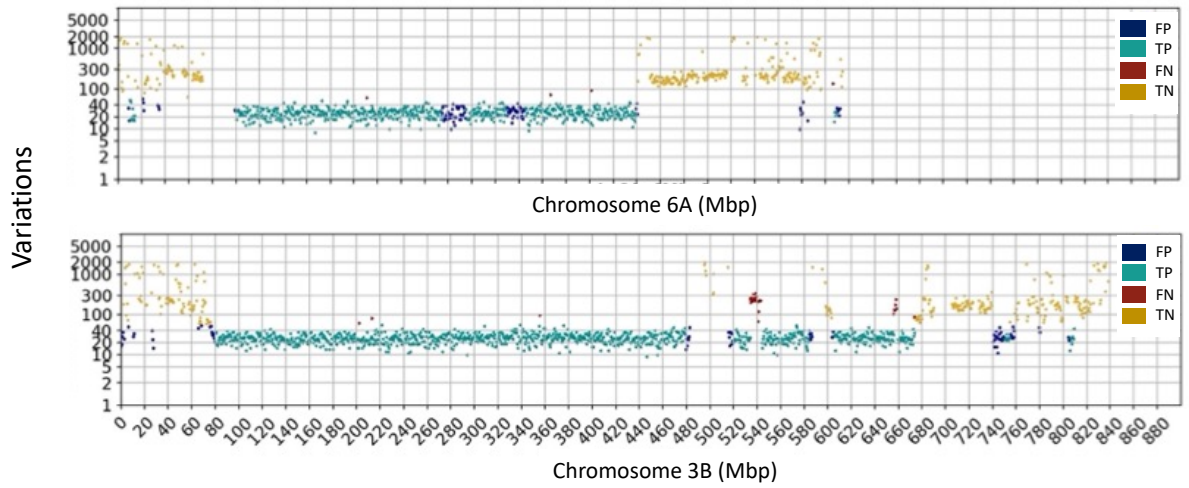


Fig. 3. 4. Precision-Recall outputs chromosome physical positions.

Here we are exemplifying two chromosomes: chr6A (top) and chr3B (bottom) of raw reads (12-fold) 250 bp of Claire (query) against Mattis reference assembly. Using haplotype regions from Brinton et al., 2020 as a positive control to test our GMM model we expected four outputs. For IBS: TP when the two methods call and IBS region. FP when only IBSpy call an IBS region. Similarly, non-IBS generates TN and FN when the two methods and only IBSpy call a non-IBS region, respectively. *Variations* counts (*y*-axis) within 500 kbp across chromosomes physical position (*x*-axis in Mbp). Blue dots indicate IBS regions captured by our approach only. Red dots indicate regions called as IBS in Brinton et al., 2020 not captured by IBSpy. The cyan dots where the two methods agree to call IBS region, and the yellow dots where the two methods agree to call non-IBS region.

3.4.1.3. Tracking haplotypes in modern wheat

Using the IBS regions defined by the GMM model, we next wanted to evaluate if we could track back genome regions in pedigree related cultivars. In our dataset we had three modern cultivars: Flame and Wasp which are the direct parents of Claire, an important UK variety (**Fig 3.5**). Using this pedigree as a case study, we called IBS regions using the GMM model and compared regions shared to a common reference (Mattis). Our hypothesis was that if Claire shared an IBS region with Mattis, at least one of the parents (Flame or Wasp) should also share this IBS region with Mattis.

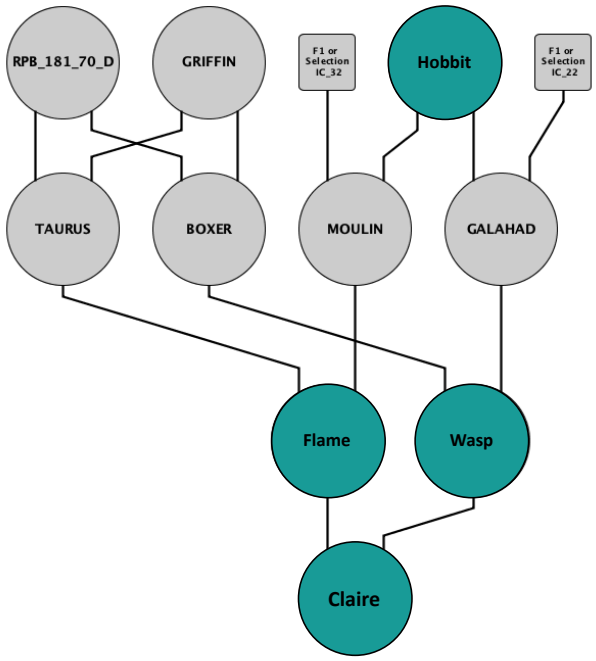


Fig. 3. 5. Claire (UK variety) pedigree.

A simplified pedigree of Claire based on information from http://wheatpedigree.net and Helium software (Shaw et al., 2014) to construct the hierarchical pedigree. Flame and Wasp are the direct parents of Clare. Flame and Wasp have a common parent, Hobbit, which is one generation above. In this project we have WGS (~12-fold) of Hobbit, Wasp, Flame and Claire. Furthermore, we have the genome scaffold assembly of Claire from the plus 10 pangenome project.

Using chr6A as an example, we observed that Claire shares a block from 100 to 440 Mbp with Mattis and this region was also shared by Flame and Mattis but not by Wasp and Mattis (Fig 3.6a, cyan colours and black dashed box). These results suggest that Flame was the donor of chr6A into Claire in the 100 to 440 Mbp region. In addition to the common IBS region to Mattis, we detected a "*variations fingerprint*" that extended outside the IBS region from 440 to 580 Mbp in Claire and Flame with high similar level of variation. Therefore, we hypothesize that the region inherited to Clare from Flame must be larger (**Fig. 3.6a,** grey box).

To further validate our hypothesis, we anchored the scaffold level assembly of Claire to CS (RefSeq.v.1.0) as a high-quality chromosome reference to obtain the coordinates positions of the scaffolds in Claire. Using these projections, we ran Claire as a reference and compared the IBS regions to Flame and Wasp. As predicted, the IBS region between Flame and Claire extended from 60 to 580 Mbp (**Fig. 3.6b**, blue dashed box). In addition, we detected that Wasp shares the region with Claire from 0 to 60 Mbp and from 560 Mbp to the end of the chromosome (**Fig. 3.6b**, red boxes). These results support our hypothesis and indicates that two main recombinations took place on chr6A between Flame and Wasp that gave rise to the chr6A of Claire: one at 60 Mbp and

the second between 560 to 580 Mbp. Our results also detected a region shared by both parents (560 to 580 Mbp) and hence we cannot assign to a unique parental source (**Fig. 3.6b**, between dashed lines).

To further investigate the fixed regions in Clare at 560 to 580, we analysed the IBS calls from Hobbit into Claire. Hobbit is a common parent of Flame and Wasp. As expected, we found that Hobbit shares the IBS block at 560 to 580 Mbp with Claire which further supports that this region was brought intact from Hobbit to Claire either through Flame or Wasp (**Fig 6b,** blue bar). We propose that those blocks are inherited to Claire either from Flame, Wasp, or both in **Fig. 6c**, where some regions are fixed as depicted by asterisks.

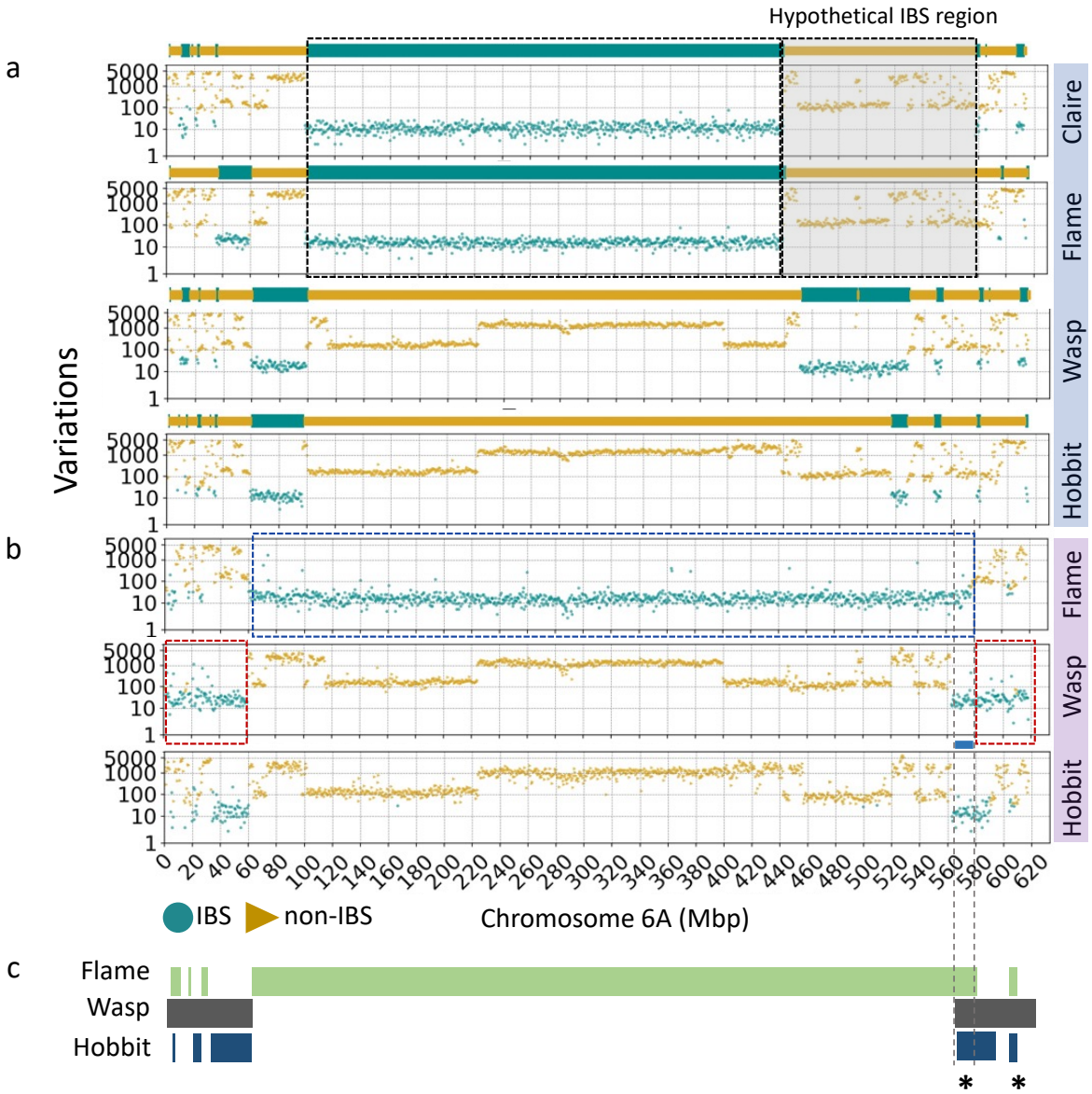


Fig. 3. 6. Large IBS blocks are maintained through generations.

GMM prediction of IBS regions (cyan) and non-IBS (yellow) using a 500 Kbp window *variations* count. **a**), *Variations* across chr6A of Flame, Wasp, and Hobbit against Mattis reference. Note, Mattis is not directly involved in the Claire pedigree, but it is used as a common reference to detect IBS regions. **b**), IBS regions using scaffold of Claire reference projected into CS reference chromosome positions. **c**) proposed haplotypes passed intact from Flame (green) and Wasp (grey) into Claire, and from Hobbit (blue) into Claire through Flame and Wasp. Dotted lines and asterisk indicate regions that are IBS in the parents and hence fixed in the progeny of Claire.

In summary, with this analysis we detected IBS regions in pairwise chromosome-scale references at high Precision and Recall rate using the GMM model similarly to the regions in Brinton et al., 2020. Importantly, we also detected IBS regions using raw reads at ~12-fold coverage. Using this IBS calls we tracked back a fixed IBS region which indicates that large blocks are kept intact intentionally or unintentionally by breeding selection. This is of significance because if a gene or a group of genes involved in agronomically important traits are located at the fixed region, the chances to improve a new cultivar will be restrained. On the other hand, if a breeder wants to maintain this region intact and only recombine and select outside the region, this information will be of pivotal interest for a breeding decision.

Equally important, we indirectly detected a *variations fingerprint* in Claire and Flame using Mattis as common referent parent and validated this "*fingerprint*" between the two cultivars (Flame and Clare) to be IBS. These results are of importance because, in theory, we could determine IBS regions or haplotypes in any pairwise comparisons using raw reads not only against the reference, but among raw reads samples without the need of a direct genome assembly of each accession.

3.4.1.4. One reference; multiple queries

In our previous tests, we classified IBS and non-IBS regions based on a reference assembly. Importantly, we realized that some genotypes had a similar *variations fingerprint* across defined chromosome regions when using a common genome reference. This profile was maintained between two query samples regardless of being IBS or non-IBS to the reference genome. Furthermore, such *variations fingerprints* were maintained between the two query samples irrespectively of the genome reference tested (**Fig 3.6a**, black and grey dashed squares). We realised that two genotypes that share the same *variations fingerprint* against the same reference would be most likely IBS between them. We hypothesized that these differences and similarities among samples could be measured by similarity distances metrics which are common in clustering algorithms and can handle data from multiple individuals at once instead of pairwise comparisons.

To address this hypothesis, we tested the hierarchical clustering algorithm which is common in data analysis clustering (Bar-Joseph et al., 2001; Müllner, 2011; Murtagh & Contreras, 2012). Hierarchical clustering builds clusters by merging or splitting samples successively using different metrics of measure defined by the user. To test this clustering as a case study to our data, we focused on genomic regions surrounding known QTLs of haplotypes reported previously in the literature, the *YR7* yellow rust resistance loci (Marchal et al., 2018) which represents a highly diverse region, and the *RHT-B1* locus which is highly conserved by breeding selection in several WatSeq genotypes.

The yellow rust disease resistance YR7 locus: The *YR7* locus is located on chr2B and has been characterised previously by (Marchal et al., 2020; Marchal et al., 2018). This region harbours several NLR genes, including *Yr7* and *Yr5a*/*Yr5b*, and it is considered as highly diverse region from a sequence conservation viewpoint. In CS (RefSeq v.1.0) the boundary of the region is between the *TraesCS2B02G486000* (683,034,442 bp) and *TraesCS2B02G490200* (687,635,975 bp) genes. As a case study and controls, we used the defined haplotypes (Marchal et al., 2020; Marchal et al., 2018) and pedigree information where, Cadenza, Paragon, Thatcher, Grafton, Skyfall, and Remy share the same *Yr7* haplotype (C G A). The original source of *Yr7* comes from a tetraploid wheat cultivar Iumillo introduced into Thatcher (hexaploid).

Our clustering results confirmed that these genotypes share the same haplotype as they were clustered together across the region (Cadenza, Paragon, Thatcher, Grafton, Skyfall, and Remy). Watkins WATDE0052, WATDE0050**,** and WATDE0067 corresponding to the ID names W397, W387, and W496, respectively, in Marchal et al., 2018, are also reported to carry the *Yr7* haplotype (C G A) and were identical to Cadenza. Consistently, in our analysis WATDE0052**,** WATDE0050 and WATDE0067 are identical among them in the exact *Yr7* region (**Fig. 3.7,** blue bar group), but they are slightly different to Cadenza when adding ± 3 Mbp flanking region (**Fig. 3.8,** blue and red boxes groups). Marchal et al., 2018 defined WATDE0033 (W246) to have the "G A G" non-*Yr7* haplotype. However, in our analysis using both the exact *Yr7* regions and the ± 3 Mbp flanking region we observed that WATDE0033 is identical to WATDE0052 which has the C G A *Yr7* haplotype (**Fig. 3.8,** red group).

The Mace and Wyalkatchen genotypes clustered in the same group as with Cadenza *Yr7* group of lines when using the exact *Yr7* region. However, using the ± 3 Mbp flanking region, they formed a separated cluster group. (Marchal et al., 2020) found a 99.98% sequence similarity in the *Yr7* region between Landmark, Stanley, and Mace against Cadenza, therefore a near-IBS region. It could be that our clustering analysis cannot differentiate a few SNPs within the exact *Yr7* region and genome flanking information is needed to differentiate these "long-range blocks". These

blocks could be informative to further investigate similar and more distant haplotype blocks to the *Yr7* locus for phenotypic characterization against the yellow rust disease.

Most of the other samples are also consistently clustered together in agreement with our previous results when using the GMM model and pedigree information described in our previous section. For example, the cluster of Claire includes Flame, Wasp, Hobbit, and Revelation, which are known to be related by pedigree among them are in the same group either using the exact *Yr7* region and with the flanking region. These results support our hypothesis that clustering algorithms can be used to identify genotypes sharing sequence identity similarities or identical haplotypes using a common reference as a "template". Importantly, these sequence similarities can be corroborated if the same set of samples are compared against to other genome references when available as with pangenome assemblies that can have different genome information in their syntenic regions.

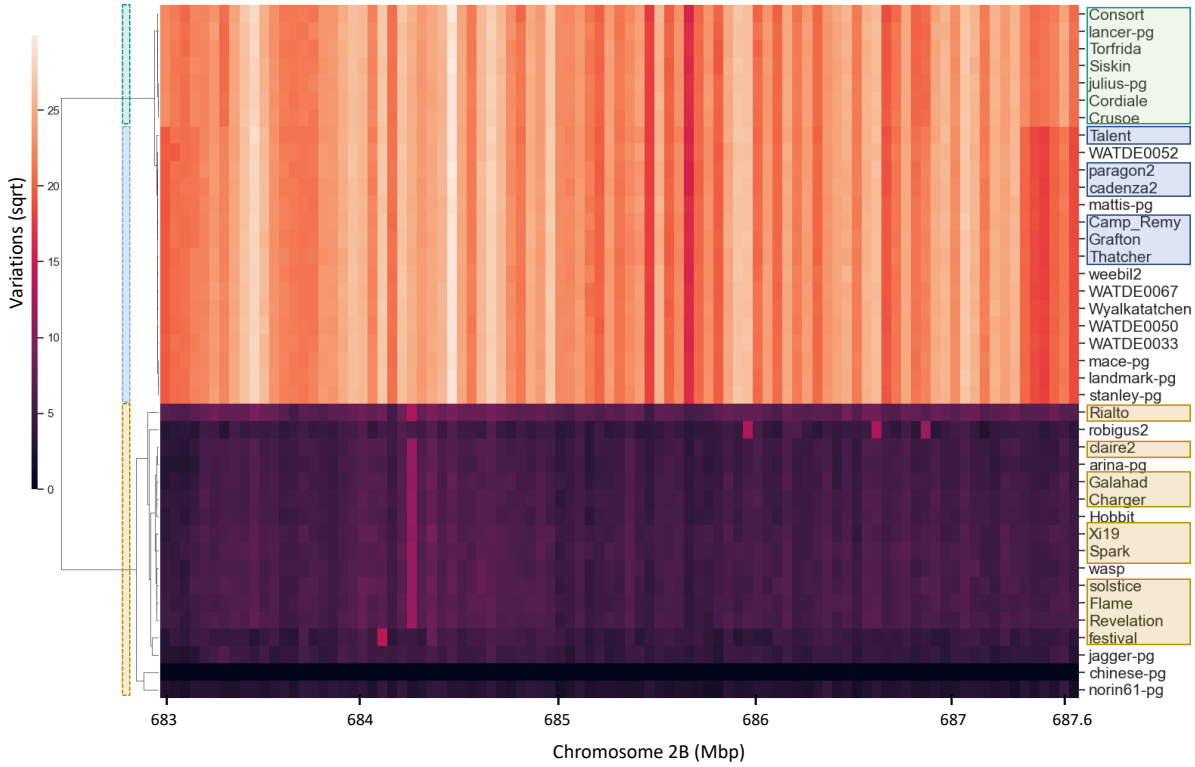


Fig. 3. 7. *YR7* locus exact region using CS reference.

In blue *Yr7* carriers detected in (Marchal et al., 2018) In green and yellow, non-Yr7 carriers. Three main clusters are formed using the exact gene boundaries of the *YR7* locus from *TraesCS2B02G486000* (683,034,442 bp) to *TraesCS2B02G490200* (687,635,975 bp) based on CS (RefSeq v1.0). IBSpy *variations* were transformed to sqrt for the clustering.

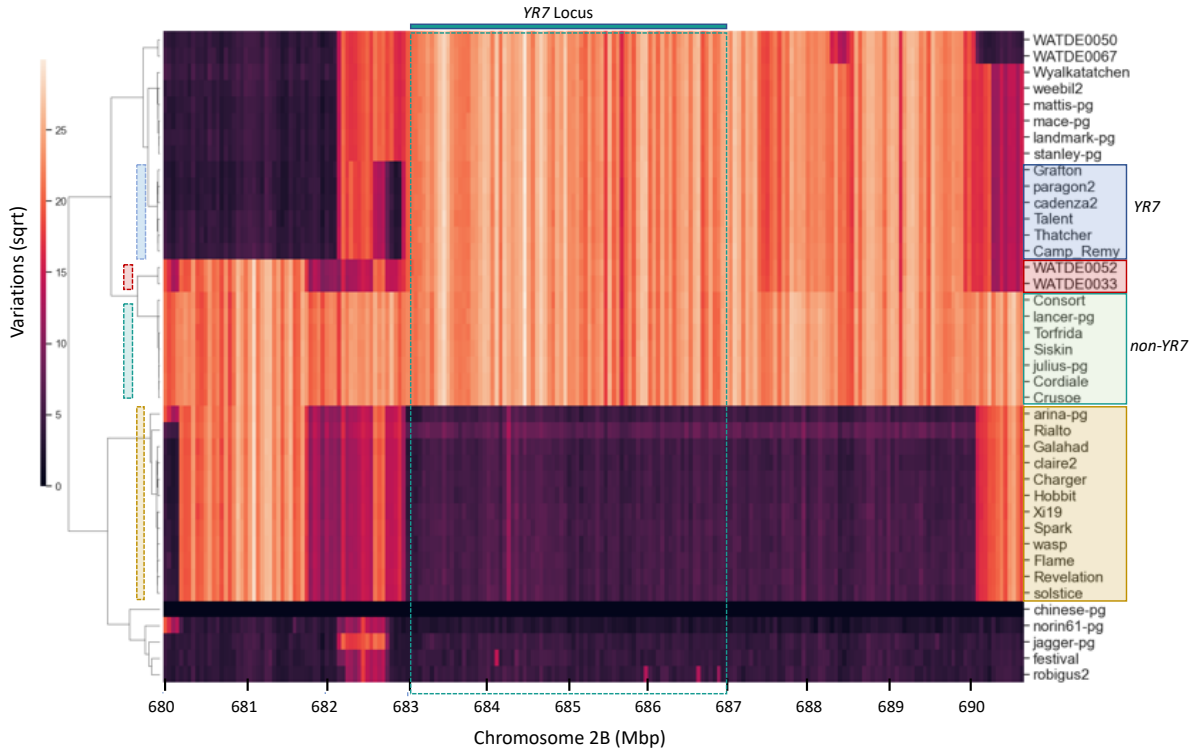


Fig. 3. 8. *YR7* locus ± 3 Mbp flanking region using CS reference.

In blue *Yr7* carriers detected in Marchal et al., 2018. In green and yellow, non-*Yr7* carriers. Seven main clusters are formed using 3 Mbp flanking region of the gene boundaries from the *YR7* locus from *TraesCS2B02G486000* (683,034,442 bp) to *TraesCS2B02G490200* (687,635,975 bp) based on CS (RefSeq v1.0). The blue bar indicates the exact *YR7* locus region. IBSpy variations were transformed to sqrt for the clustering.

In summary, we validated that clustering algorithms can group similar genotypes sharing a genome region by using IBSpy *variations* counts from a comparison against a common reference (CS in this example). Changing the flanking region size, some of the samples clustered in a different group, as it would be expected if recombination had occurred on the edges of the *Yr7* haplotype interval. Overall, we observed a consistency with previously predicted haplotypes (Marchal et al., 2020; Marchal et al., 2018) and found that clusters were often related by pedigree. Genotypes having known *Yr7* haplotypes clustered together.

3.4.1.5. Multiple reference clustering (hierarchical clustering)

To combine multiple references in our analysis we used the gene projections based on CS reference (RefSeq v1.0) to anchor syntenic regions from each of the other pangenome genotypes. To find the syntenic regions among references we used a common reference as a “template”, and we named it “assembly” to differentiate among the other references. For example, when using CS

as a template assembly, we use the gene annotations in CS to find the syntenic regions in other pangenome references by searching the projected gene location. We only integrated reference regions if the corresponding gene was present and located in the same chromosome as in the assembly. Using this additional information, we hypothesized that it would help us to discriminate among the tested genotypes samples more precisely. For example, in our previous analysis when using the individual genome reference of CS, we could not differentiate Mace, Landmark, Stanley, and Wyalkatchen from the Cadenza *Yr7* group of lines when using the exact *Yr7* region (**Fig. 3.7**). On the other hand, using the syntenic windows from multiple reference of the *Yr7*, although still very similar clusters, they are clearly different to the Cadenza group (**Fig. 3.9**). Marchal *et al.,* 2020 found a 99.98% sequence similarity (a near-IBS region) in the *Yr7* locus between Landmark, Stanley, and Mace vs Cadenza. Our results with the multiple reference analysis support these differences as they are located in a separated cluster from Cadenza. These differences are evidenced by the Landmark, Stanley, Mace, and Mattis references where H1 have a block with high level of *variations* are shown in orange-clear colour (**Fig. 3.9**).

Using this multi-genome reference approach clustering, in addition to the *Yr7* locus, we could differentiate five clusters which we hypothesize belong to different haplotypes. Therefore, we manually classified as being H1 - H5 haplotypes in the region where H1 is the Yr7 haplotype which includes Thatcher, Grafton, Cadenza, Paragon, Remy, and Talent (**Fig. 3.9,** coloured boxes). These haplotype carriers are consistent with the haplotypes defined in (Marchal et al., 2020; Marchal et al., 2018) and are pedigree related.

In summary, our results were consistent when testing multiple genome references individually or combined depending on the genome information included (exact locus region or adding flanking regions). However, using the multi-reference approach and the exact boundaries of the region we could differentiate the *Yr7* carriers precisely. This analysis supports that using multiple pangenome references in a single clustering analysis allows to differentiate highly similar samples, e.g., the Mace from the Cadenza *Yr7* haplotype. Furthermore, using multi-references we identified additional cluster groups suggesting that there are multiple haplotypes in the YR7 locus. It is important to mention that in this analysis we used a reduced number of samples to validate our method using known haplotypes. However, in our WatSeq dataset, we have > 1000 genotypes. It is likely that out of the haplotypes identified in this pilot analysis, there will be multiple additional and private haplotypes either from landraces or modern cultivars still unexploited just in this locus.

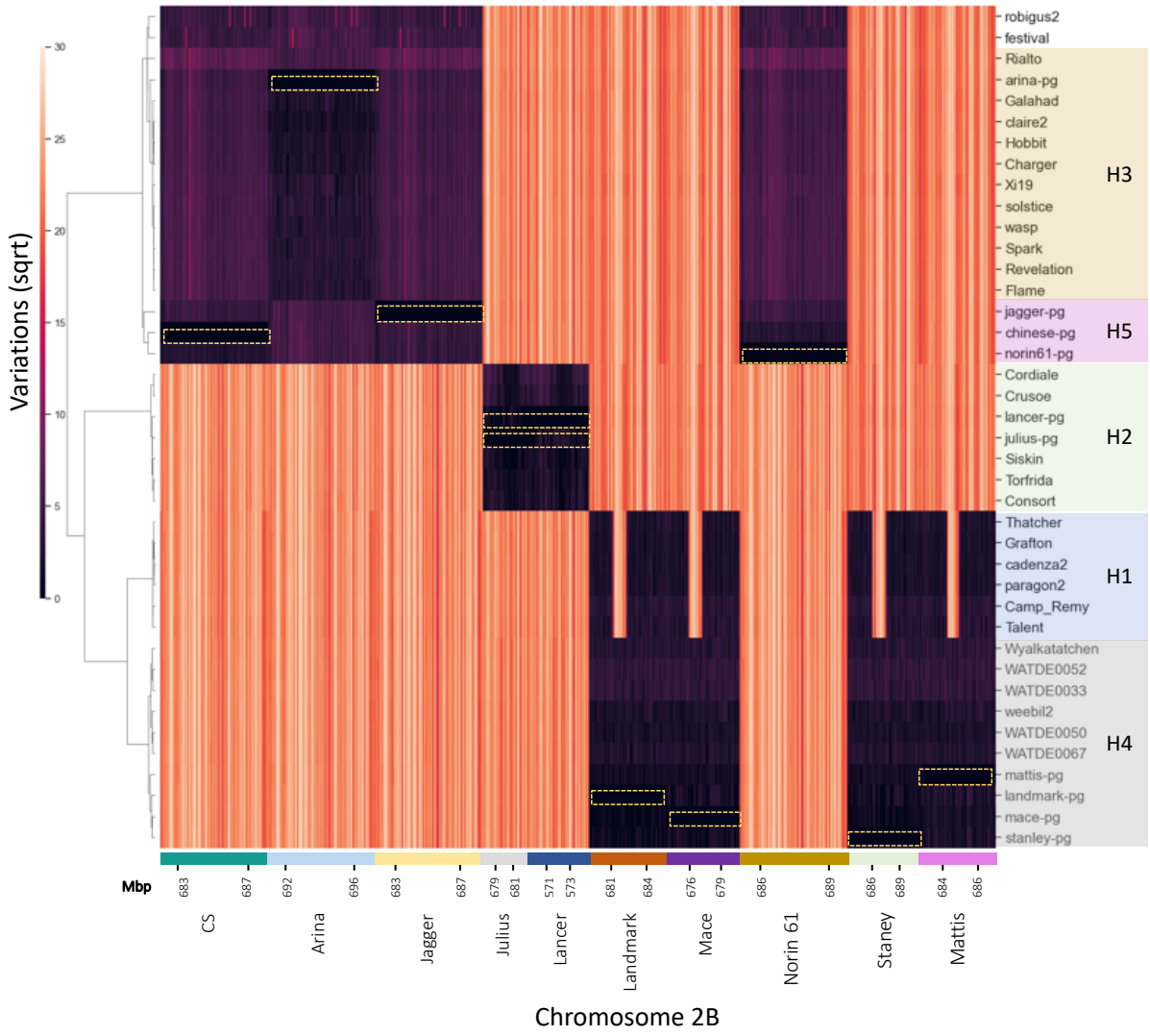


Fig. 3. 9. Clustermap using syntenic regions from multi-references of the *YR7* locus exact region.

IBSpy *variations* count were *sqrt* transformed before clustering. The exact region ranges from the *TraesCS2B02G486000* (683,034,442 bp) to *TraesCS2B02G490200* (687,635,975 bp) gene position in CS (equivalent region as in **Fig. 3.7**). *y-axis*, clustered groups of different cultivars and our hypothetical haplotype groups. In blue, haplotype H1, *Yr7* carriers detected in (Marchal et al., 2018). The other Haplotypes depicted in colours (H2 – H5) are non *Yr7* carriers. *x-axis,* each of the references having a syntenic region of the *YR7* locus. Different sizes of the colour bars indicate physical genome differences in length. Yellow boxes indicate our quality control of references used both as a query and as a reference. E.g., *k*-mers created from the genome assembly of CS was used as a query against the CS (green bar) genome reference and therefore no *variations* are expected (black colour in heatmap).

3.4.1.6. The *REDUCED HEIGHT-1* (*RHT1*) loci

In our previous example we explored the *YR7* locus, a rust resistance region which is highly diverse. Here we validated our approach to discriminate known *RHT* alleles from a region that has been

intensively selected by wheat breeders. There are two *RHT1* genes that have been strongly selected in wheat; *RHT-B1* which correspond to *TraesCS4B02G043100* in CS, and *RHT-D1* that correspond to the gene *TraesCS4D02G040400.* For this analysis we will focus in the *RHT-B1* only (Fig. 3.10). Brinton et al., 2020 defined different *RHT-B1* alleles for the 15 pangenome genotypes by alignment comparison of the 300 Kbp flanking region of each gene. In this analysis the wild type allele (e.g., tall phenotypes) is described as the "a" allele, whereas the reduced height phenotype allele is described by the "b" allele, as follows (Table 3.1.).

Table 3. 1 Analysis of *RHT-B1* sequences from (Brinton et al., 2020).

			Sequence used for alignment (bp)		Ns in sequence		Maximum alignable sequence (bp)		SNPs/		Total align	**Sequence**	Aligned/Max sequence (%)		NUCmer	NUCmer aligned/	
Allele	Cultivars	Start	Jagger/ Mace	*B1a/ D1a*	Jagger/ Mace	*B1a/ D1a*	Jagger/ Mace	*B1a/ D1a*	indels (bp)	Matches (bp)	sequence (bp)	**identity (%)**	Jagger/ Mace	*B1a/ D1a*	aligned (bp)	Total align (%)	Comments on BLASTn
Rht-B1a_1	**Chinese Spring**, Stanley, Norin61	30,711,017	300,001	299,779	2,060	1,491	297,941	298,288	135	297,573	297,708	**99.955%**	99.9%	99.8%	272,760	91.6%	
Rht-B1a_2	**ArinaLrFor**	31,343,647	300,001	299,319	2,060	233	297,941	299,086	98	297,870	297,968	**99.967%**	100.0%	99.6%	270,337	90.7%	Excluding 151 bp indel
Rht-B1a_4	**Julius**, Paragon	30,483,308	300,001	298,752	2,060	1,607	297,941	297,145	97	297,215	297,312	**99.967%**	99.8%	100.1%	242,440	81.5%	Excluding 151 bp and 602 bp indel
Rht-B1a_8	**Mace**, Cadenza	30,656,226	300,001	298,300	2,060	681	297,941	297,619	137	297,379	297,516	**99.954%**	99.9%	100.0%	267,754	90.0%	Excluding 602 bp indel

"Comparison between ~300 Kbp sequence surrounding *RHT-B1* in sequenced cultivars. Table shows the comparison between different *RHT-B1a* alleles (representative cultivar in bold) and *RHT-B1b* (Jagger). The table indicates the sequences used for BLASTn alignments, number of Ns in each sequence and the maximum sequence (total minus Ns). Total matches and SNPs/indels are indicated and the percentage sequence identity is calculated alongside the breadth of the BLASTn alignment with respect to the maximum. Total sequenced aligned from the tabulated NUCmer output (NUCmer aligned) alongside the breadth of the NUCmer alignment, is shown. Where relevant, indels that break the alignment are indicated. These are not included in the calculation of sequence identity."

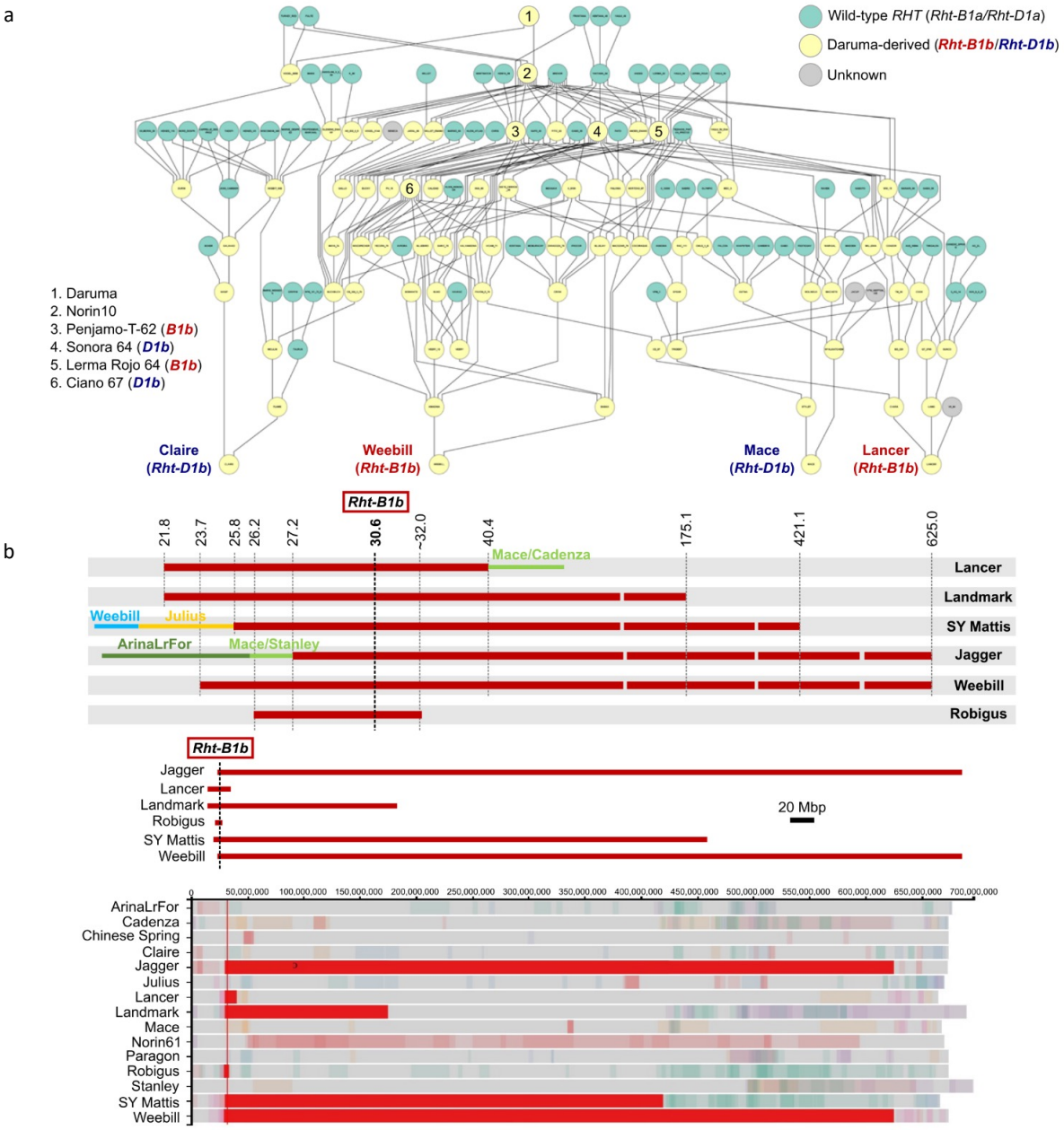


Fig. 3. 10. Analysis of *RHT-B1* pedigree and haplotypes from (Brinton et al., 2020) (Supplementary Fig. 2.).

a, "Pedigree of Lancer (*RHT-B1b*, Australia), Weebill (*RHT-B1b*, Mexico), Claire (*RHT-D1b*, UK) and Mace (*RHT-D1b*, Australia) tracing back to the common accession, Daruma (Japan), which is the donor of *RHT-B1b/RHT-D1b* (Wilhelm et al., 2013). Lines which are derived from Daruma are indicated in yellow, whereas lines with the wild-type *RHT-B1a* and *RHT-D1a* alleles are indicated in teal. Unknown genotypes are in grey. Important accessions in CIMMYT breeding which are shared in the pedigree of the four sequenced cultivars included in the tree are indicated with numbers. Pedigree generated with the Helium software. **b**, Shared haplotype block on chromosome 4B in six of the 15 sequenced cultivars which carry *RHT-B1b* (Jagger, Lancer, Landmark, Robigus, SY-Mattis and Weebill). Diagram shows the relative size of the shared haplotype among cultivars, with *RHT-B1b* indicated by the vertical black line. Due to the difference in scale, detailed breakpoints are indicated diagrammatically, whereas the middle panel shows at actual

scale. Haplotype blocks identified in this study are shown in the bottom panel as visualised in www.crop-haplotypes.com."

Using these known *RHT-B1* haplotypes as a case study, we further explored the region by adding 0.5 Mbp flanking genome region to run a multi-genome reference clustering. When combining the multiple genome references, we observed that the *Rht-B1b* haplotype group of lines (Jagger, Lancer, Mattis, Landmark, and Weebill), based on (Brinton et al., 2020), clustered together (**Fig. 3.11,** blue box, H1). An additional set of lines including Santiago (*Rht-B1b* carrier) clustered in the same group as Robigus. Santiago shares the Robigus *Rht-B1b* allele (Würschum et al., 2017) based on molecular marker analysis, and it has a common pedigree three generations below Robigus.

Other groups that clustered together, e.g., Claire, Fame, Revelation, and Boxer (**Fig. 3.11,** H3), could be expected based on pedigree information. Wasp, Hobbit, and Galahad formed a different group supporting that they come from the same pedigree (H2) meanwhile Wyalkatchen and Stanley form a unique group (**Fig. 3.11,** H5).

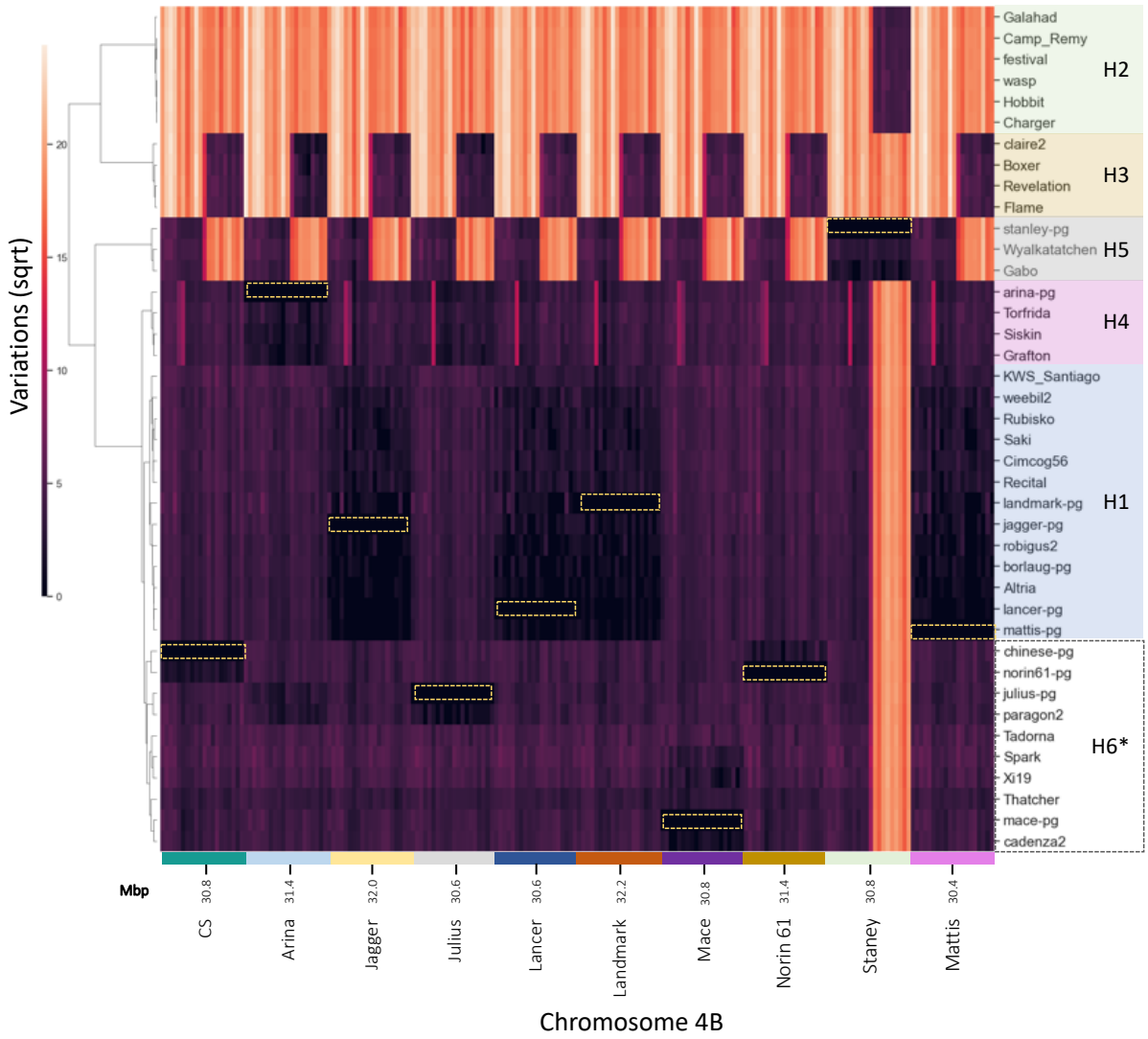


Fig. 3. 11. *RHT-1B* locus clustermap using multiple references.

We use the *TraesCS4B02G043100* gene in CS to locate the projected corresponding genes in the other genomes. In the *x-axis* are the genome regions in each reference. The values in Mbp are the corresponding *Rht-B1* position in each reference. Colours indicate the similarity of query samples in the *y-axis* to the corresponding reference. Purple-dark colours are more similar and orange-light colours are more different (have more variations) to each of the references synteny regions. Each of the colours in the *y-axis* indicates the hypothetical haplotypes corresponding to each cluster. H1 has the *Rht-1Bb* haplotype allele. H6* form a closely related cluster with the H1 *Rht-1Bb* carriers, but they do not have the *Rht-1Bb* allele. The yellow squares correspond to the references included in the analysis as quality control, and as expected they have no variations against themselves.

3.4.1.7. Clustering metrics

Based on the analysis above, we determined that multiple genome references allowed to cluster the known *RHT1* and *Yr7* carriers. Using hierarchical clustering we identified clusters of genotypes

in agreement to the hypothesized relationship based on pedigree and literature. The challenge, however, is to define a specific threshold to estimate a cut-off where a cluster should be separated and considered as a different group.

To define if we could use alternative method to separate haplotype groups in an unbiased manner, we analysed different metrics for similarity distances including: 'braycurtis', 'canberra', 'chebyshev', 'correltion', 'cosine', 'euclidean', 'minkowski', 'seuclidean' 'cityblock','sqeuclidean'. For correlation metrics we tested the 'pearson', 'kendall', and 'spearman' using the entire locus and multiple samples.

Looking into the analysis and using Jagger as a known *Rht-B1b* allele, we detected four distance metrics (euclidean, cosine, seuclidean, sqeuclidean) giving similar results and grouped the known genotypes carrying the *Rht-B1b* allele. Surprisingly, in this group a Watkins genotype also seemed to carry the *Rht-B1b* allele (data not shown). We knew that the source of *Rht-B1b* came from an old Japanese variety Norin 10 (from landrace Daruma) (**Fig. 3.10)**, therefore, it's unlikely that a Watkins line would have the same haplotype. However, this may be the case where a similar or near-identical haplotype has been maintained in this variety which was not identified before. Also, we cannot rule it out that a cross contamination could have happened during seed handling since its collection in the 1900s.

In the correlation analysis we observed better results than with the distance metrics. With *spearman* and *kendall*, the clustering of the known *Rht-B1b* genotypes is consistent and the separation between samples groups is better compared with the other metrics. We hypothesize that a threshold using spearman would range between the ~0.90, or Kendall = ~0.70 (**Fig. 3.12**). However, we noticed that when comparing each sample against each other, the threshold or correlation was slightly different (data not shown here). These results suggested that we would need to adjust a global threshold for each region.

In summary, the *spearman* correlation metric is the best to group the targeted alleles with a threshold close to~0.9. In the distance metric analysis, overall *euclidean*, *cosine*, *seuclidean*, *sqeuclidean* clustered most of the know samples alleles and perform similarly. In addition, these case studies suggest that a bespoke threshold would need to be defined for each locus.

Fig. 3. 12. Spearman correlation of the "*variations fingerprint*" among multiple genotypes of known *RHT-B1b* allele carriers.

In blue *Rht-B1b* carriers showing high correlation on the *variations* count from the multiple-reference syntenic regions against a particular genotype in *y*-axis as a "target comparison" vs *variations* count profiles genotypes in *x*-axis. For example, in the first comparison the "jagger-pg" genotype *variations* counts against all other genotypes in the *x*-axis were measured for their correlation using the "*variation fingerprint*" within the *RHT-B1* syntenic region from multiple references. Gray = 0.96, Red = 0.90, Green = 0.80, and Blue = 0.70 Spearman correlations thresholds.

3.4.1.8. Syntenic genome regions

As shown before, using the IBSpy "*variations fingerprint*" from multiple references can help to predict clusters and correlations among genotypes. This is because when two samples are IBS between them, they will always have the same level of *variations* to any reference used irrespectively of how similar or dissimilar they are to a reference. For example, if two genotypes are non-IBS in a defined genome region, both can be equally different to a ***x*** genome reference **(Fig. 3.13a**, left**)**, but they will not be equally distant to a second reference ***w*** **(Fig. 3.13a**, right**)**. In an example where genotype ***y*** and ***z*** are IBS in a specific genomic interval; if genotype ***y*** has a low *variation* count to a reference ***d***, and high *variations* count to reference ***w*** for the syntenic genomic interval, then genotype ***y*** and ***z*** should have similar level of low *variations* to reference ***d*** and equivalent high *variations* count to genotype ***w***. If the two genotypes are further compared to a ***x*** reference, they should be equally different (or similar) to ***x*** (**Fig. 3.13b**).

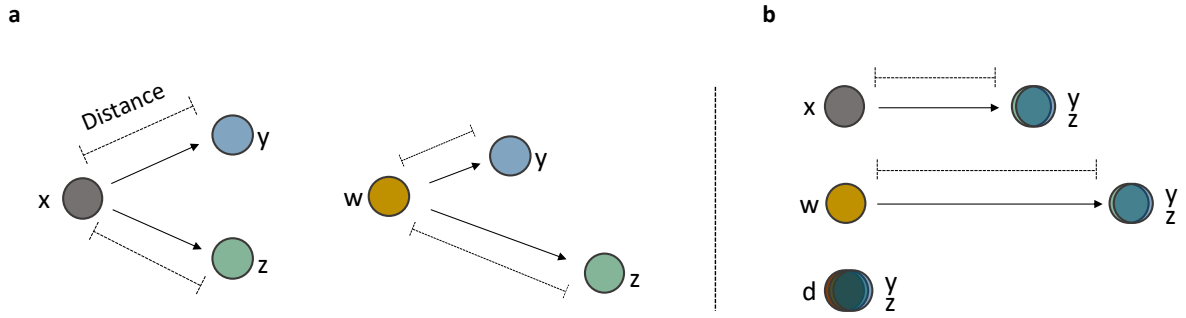


Fig. 3. 13. Distance similarities among samples.

a) non-IBS example of a pairwise comparison between *y* and *z* using *x* and *w* references as a common reference to indirectly measure the distance between them (*y* and *x*). Left, *y* (query) and *z* (query) genotypes comparison similarity distance based on variations count against *x* (reference). Right, the same query comparisons against a *w* reference depicting different distance similarities between *y* and *z*. **b**), IBS example of *y* and *z* genotypes comparisons against three references; *x*, *w*, *d* where against the three references *y* and *z* have similar distance indicating that they (*y* and *z*) are truly IBS in a defined genome region.

To perform multiple-genome IBSpy "*variations fingerprint*" analysis, we used the syntenic genome regions described in **Supplementary Fig. 9** of (Brinton et al., 2020) (**Fig. 3.14a**) based on RefSeq v1.0 and the RefSeq v1.1 annotations as a common factor among all references. In brief, first the genes in a region from the RefSeq v1.0 and the RefSeq v1.1 annotations are projected into a block of a "target reference", which can be any other pangenome reference . If one gene is missing in the target block, it doesn't break the block. Gaps of maximum of 20 genes are allowed among adjacent blocks, but the adjacent blocks must have at least 10 projected genes to be integrated in

the adjacent blocks. If there are no adjacent blocks in the region, blocks with more than 10 projected genes are kept (**Fig. 3.14a**).

Using the projected gene regions, we then built blocks of syntenic windows. For example, IBSpy *variations* are run on 50 Kbp windows. For clustering and calling haplotypes in 1 Mbp windows, the algorithm uses 20 windows of a reference (20 x 50 Kbp = 1 Mbp) and finds all the possible syntenic windows in all other pangenome references. Depending on the assembly used as the template reference, and the region of the genome, some regions may not be found in the corresponding references or the number of sub-windows of 50 Kbp will be low (**Fig. 3.14b**). Hence the number of windows used for the syntenic blocks can range from 20 (i.e., using only the 20 windows of the reference) to 220 (where all the ten references also have the maximum of 20 syntenic windows).

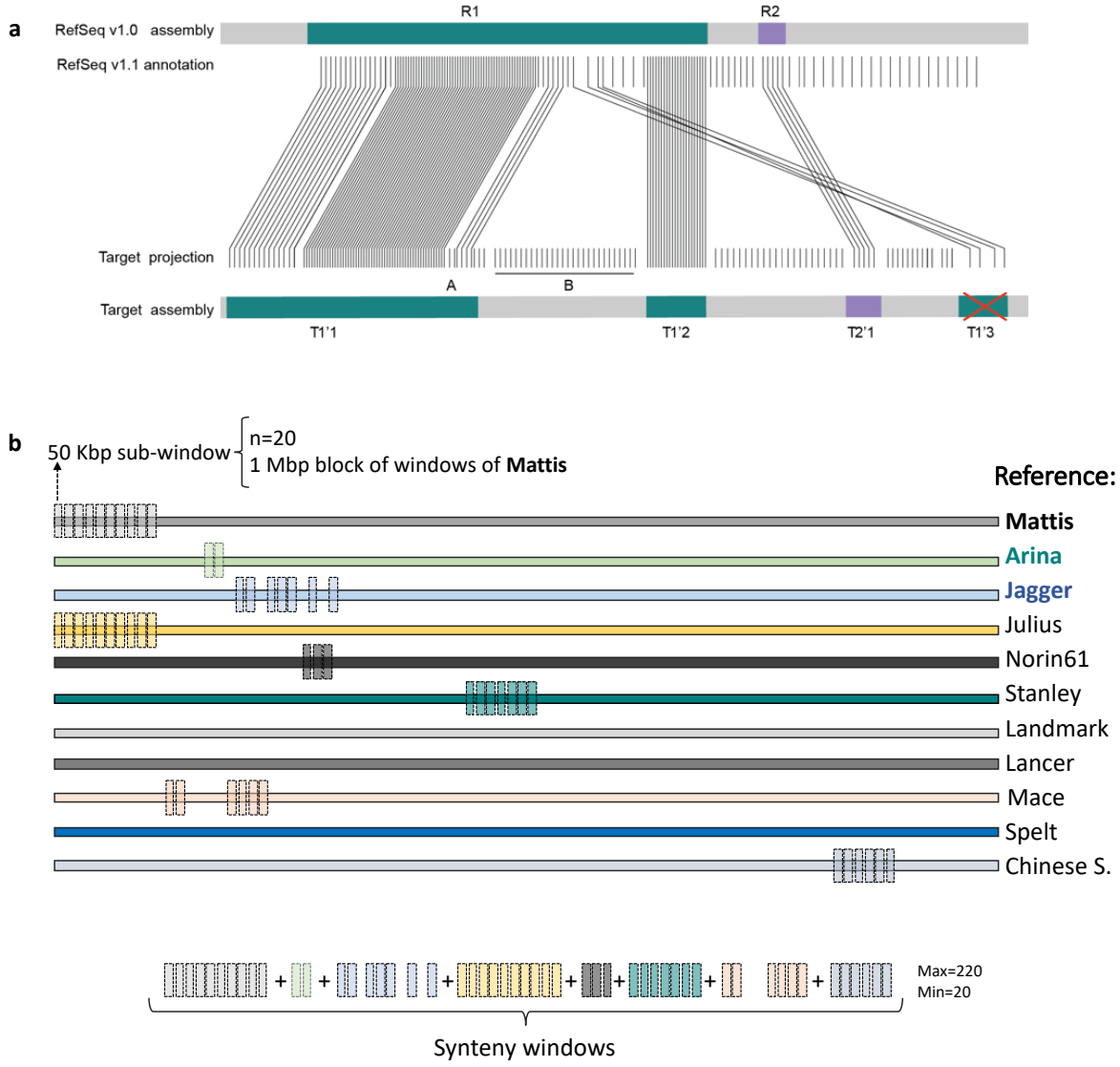


Fig. 3. 14. Pangenome syntenic regions.

a) from (Brinton et al., 2020). "The genes in region R1 are projected in Regions T1'1, T1'2 and T1'3. T1'1 has a gene (A) not from the haplotype block, but a single extra gene doesn't break the projection. There is a stretch of over 20 genes (B) between T1'1 and T1'2 which breaks the block R1. Likewise, region T1'3 is shorter than the minimum 10 genes to consider a block and there are other projected blocks from R1, hence T1'3 is removed. R2 is projected to T2'1, which contains less than the minimum 20 genes required to keep a projection. However, T2'1 is kept as it is the only possible projection for R2". **b)** Blocks of syntenic windows are assembled based on the combination of the twenty 50 Kbp windows from the reference being used (here Mattis), and all syntenic 50 Kbp windows from the ten additional chromosome scale assemblies. At the bottom, the compiled hypothetical windows found in synteny based on the 1 Mbp of Mattis. *Variations* of the compiled syntenic windows are used to cluster the multiple query samples.

To analyse the distribution of the syntenic windows we used Chinese Spring (RefSeq v.1.0) genome reference as a "template" to capture all possible corresponding windows in the ten genome assemblies of wheat pangenome using 1 Mbp block. We identified a tendency of fewer windows not being in synteny at telomere regions compared to centromere regions in the three sub genomes (**Fig. 3.15**, left**)**. This is in consistency as centromere regions are more conserved than telomeres due to high recombination rates at chromosome edges. The B genome had overall less syntenic windows than the A and D genomes. This is consistently as the B genome being the most diverse. As expected, the D genome had slightly more syntenic windows captured than the A and the B genomes (**Fig. 3.15**, right). This agrees to be the D genome of wheat the most recent hybridization that took place from a few wild progenitors of *Ae. tauschii* donors and no major introgressions have taken place.

Exploring the Lancer genome assembly as a template, we found low syntenic windows in regions where introgressions have been reported. For example, we found low syntenic windows almost exclusively to Lancer in Chr2B from ~95 to 600 Mbp where the previous introgression from *T. timophevii* was reported in (Walkowiak et al., 2020). Although, in low proportion, the syntenic windows found from this region in other cultivars could be from conserved genomic regions or misassemblies (**Fig. 15b**, green bar). Similarly, when using Mattis as a reference, we found low number of syntenic windows (~60 windows) on chr2A where the *Ae. ventricosa* introgression is located (**Fig. 15c**, green bar). This is in consistency with three genome references (Stanley, Mattis, and Jagger) having the introgressed block (Keilwagen et al., 2022; Walkowiak et al., 2020). These results suggest that other uniquely syntenic windows in the genome reference might be unknown introgressions or translocations blocks not reported before. This is of importance since often small genome regions were uniquely detected or present only in a few references suggesting that those small blocks could be additional introgressions (1 Mbp) and that IBSpy syntenic blocks could help to systematically identify them which are difficult to track by other methods. These syntenic

windows with introgression are also variable depending on the subgenome and reference used as a “template”.

The average number of syntenic windows was on average ~125 for the A and the B, and ~160 for the D genomes when using Chinese Spring as a reference. However, when using Spelta reference as a “template” we found the fewest syntenic windows across the genome (~60 on average) compared to all the other genome references (**Fig 15d**). This is expected as Spelta is a more distant genotype compared to modern cultivars. Adding more genome assemblies from landraces and wild relatives would reveal which of those genome regions are the most shared and reveal the possible introgressions donors.

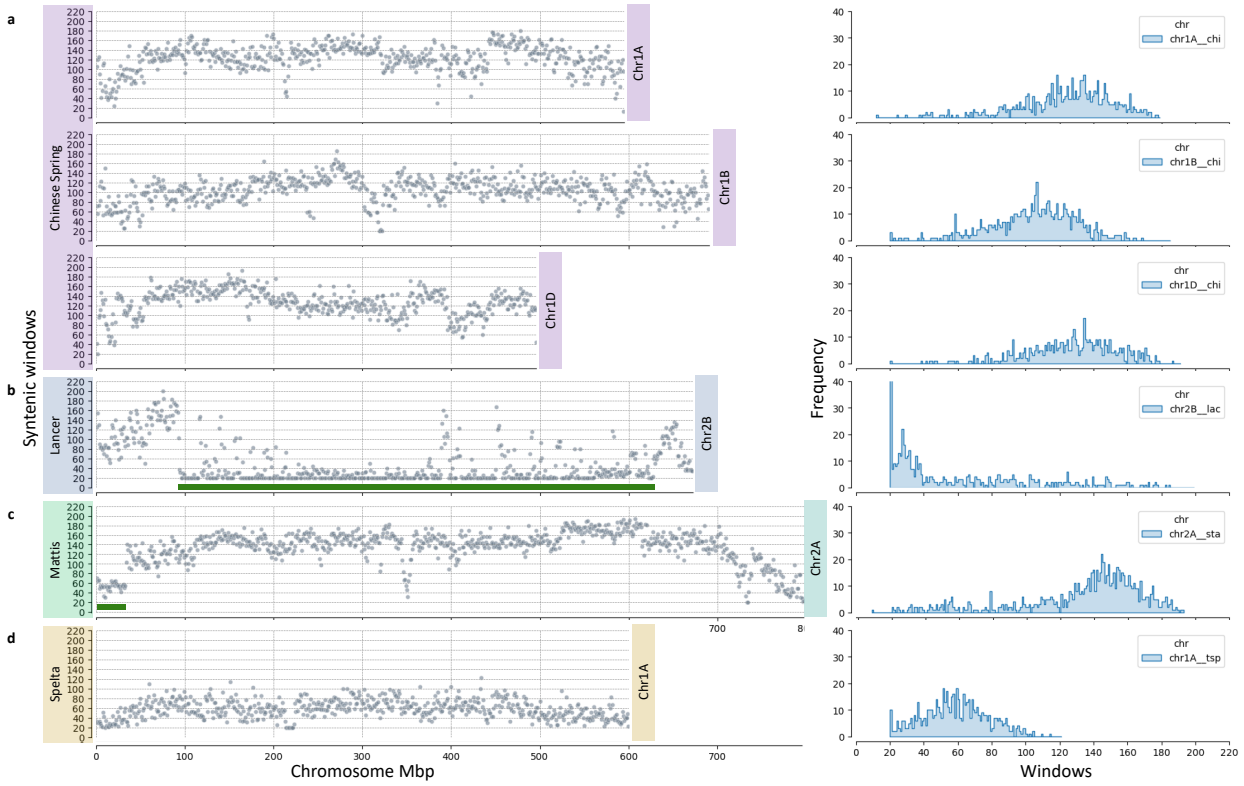


Fig. 3. 15. Syntenic windows genomic distribution.

In left are the number of 50 Kbp syntenic windows (*y*-axis) plotted across chromosome physical position in 1 Mbp blocks (*x*-axis). We expect to have a minimum of 20 windows of 50 Kbp per genome reference in 1 Mbp when there is not a syntenic window in any of the other references. We expect a maximum of 220 windows when a genome region is present in all references. **a**), an example of syntenic windows on chromosome one triad using Chinese Spring as a reference template. **b**), chr2B of Lancer showing the low syntenic (almost null) in the *T. timopheevii* introgression (green bar). **c**), chr2A of Stanley depicting the 60 syntenic window captured at the beginning of the chromosome where the *Ae. ventricosa* 2AS/2N^{v}S is located reflecting the other two references (Jagger and Mattis) in the pangenome having the introgressed block (green bar). **d**), the most distant genotype in the pangenome, Spelta, having

low number of syntenic windows across the whole genome (~60 on average) but chr1A is shown as an example. In the right side, the syntenic windows histograms of the corresponding chromosomes in the left are depicted.

3.4.1.9. Affinity Propagation (AP)

In a pilot analysis, we first evaluated different clustering algorithms such as self-organizing maps (SOM) and K-means, however the disadvantage was that they needed a predefined arbitrary number of clusters. We did not include the analyses here, but overall, we were unsatisfied by their performance and the requirement of inputting a predefined number of clusters. Therefore, we moved on to explore alternative algorithms that could automatically predict number of haplotypes per genome region.

A promising algorithm with this feature is the Affinity Propagation (AP). AP was implemented to detect patterns in different datasets (Frey & Dueck, 2007) and it is an algorithm that does not requires an arbitrary number of clusters as an input. Instead, it predicts the number of clusters based on observations from a dataset and uses a similarity distance metric between pairs of data samples to create clusters. Since our aim is to predict haplotypes by windows using *variations* counts among multiple samples, we implemented AP to automatically detect haplotypes across the genome, using either, a single reference or multiple references and using IBSpy *variations* count as an input information to build clusters.

An additional feature of AP it is that it requires a few parameters to select before its use. This can be an advantage or disadvantage. For example, users do not need to extensively explore multiple parameter combinations which can reduce the time during algorithm optimization. However, having a few parameters can lead to a few options to adjust with reduced chances to find the optimal combination for a defined dataset. One of the features of AP is the use of different metric distance, however as described in section 1.4.1.6, we selected the Euclidean distance as our default which uses the negative squared Euclidean distance between vectors. We did not test other metrics in combination with the AP algorithm in our analysis due to time, but this is something that may be a point for optimization in further developments of IBSpy. Another parameter that can be tuned in AP, is the "*preference*"; however, we used the median of the input similarity as a default since we have a range of different values for each window in our dataset.

A third parameter to adjust on AP is the "*damping*" factor (*dmp*). As a default AP uses *dmp* = 0.5 and the range can be selected from 0.5 to 1.0. As described in (Frey & Dueck, 2007), "the *dmp* parameter is the extent to which current values are maintained relative to incoming values to

avoid numerical oscillations and overfitting the data with new values". Therefore, the new values are "damped" down to avoid this. The optimal *dmp* value depends on the type of data and needs to be fine tunned accordingly to a case study. In our analysis we aimed to identify cluster groups among the WatSeq samples by using IBSpy *variations* and a defined window size. Therefore to select the optimal *dmp*, after calling IBSpy haplotypes in 1 Mbp windows intervals, we evaluated the quality of the clusters predicted by a range of *dmps* values from 0.5 to 0.97 and select the best *dmp* based on the Silhouette Coefficient (SC; see below) score metric (Rousseeuw, 1987). This process is repeated each 1 Mbp window across the whole genome always using the same > 1,000 genotypes and same parameters.

The SC score is a measure for the quality of the clusters based on the density and separation using the distance among other clusters and the distance among members within the cluster itself. High SC values indicates a well separated and condensed cluster. This evaluation is done for the clusters (haplotypes) called by each of the *dmp* values described above and on a 1 Mbp window. Then, based on the SC score, IBSpy selects the "best" *dmp* (with the highest SC score) and discard the others. Large *dmps* integrates members (genotypes) in the same cluster (haplotype) in a more relaxed manner incorporating members that are highly similar but probably may be not identical (e.g., may be near-IBS). On the contrary, low *dmp* values allocate similar (or near-IBS) genotypes into different clusters (different haplotypes) calling more haplotypes per window but impacting only those windows where the genotypes are very similar or windows with low diversity. Windows that are well separated due to high level of *variations* will remain as the original. We integrate AP into IBSpy-0.4.0 to automatically call haplotypes by a group of samples and flexible number of references.

AP Precision and Recall

For the precision/recall analysis we focused on the pangenome lines with whole genome assemblies to allow for comparisons between the IBS blocks identified in (Brinton et al., 2020) and the AP haplotypes. In our previous analysis on genome regions across multiple windows, we detected that the number of clusters changed based on the *dmp* parameter of AP. In a pilot study, running precision and recall, we detected that the "optimal" *dmp* value was different across individual windows. This is expected since each window may be independent in the absence of LD among genotypes and multiple samples *variation* counts changed in their distribution from window to window due to recombination, introgressions, deletions, and lack of genetic variation. To adjust for the "optimal" *dmp* in this analysis we tested multiple *dmp* per window and selected the *dmp* with the highest SC score clustering metric per 1 Mbp as described above. Testing several

windows across the genome, our results suggest that selecting the *dmp* with the highest SC score per window yields high and stable *fscore* across multiple Precision and Recall comparisons (**Fig. 3.16**).

In our example in **Fig. 3.16** we selected three chromosomes as an example for comparisons that shared high (52.2%), intermedium (10.8%), and low IBS regions based on 5 Mbp haplotypes described in (Brinton et al., 2020). Overall, these results were similar for the A and the B genomes which have higher genome diversity than the D genome. As shown in (**Fig. 3.16**) the scores precision, recall, and F1 scores changed among chromosomes selected and among different *dmps.* This is expected since the two genotypes Mattis vs Julius selected for the test varies in their sequence identity and IBS regions shared among chromosomes. However, selecting the "SC" as our default haplotype calls, we obtain relatively similar and high scores among chromosomes. Regardless of the genotypes tested our method usually fails to detect high score metrics for the D genome. This is because the low diversity of this genome as a result from the recent hybridization into hexaploid wheat from a few *Ae. tauschii* donors. Thus, most of the wheat genotypes will have a similar level of variations in 50 Kbp windows which impacts to the AP clustering algorithm to discriminate among haplotypes. To correct for the lack of genome diversity, we integrated an *Ae. tauschii* panel into the analysis as a query as shown in **Fig. 3.19.**

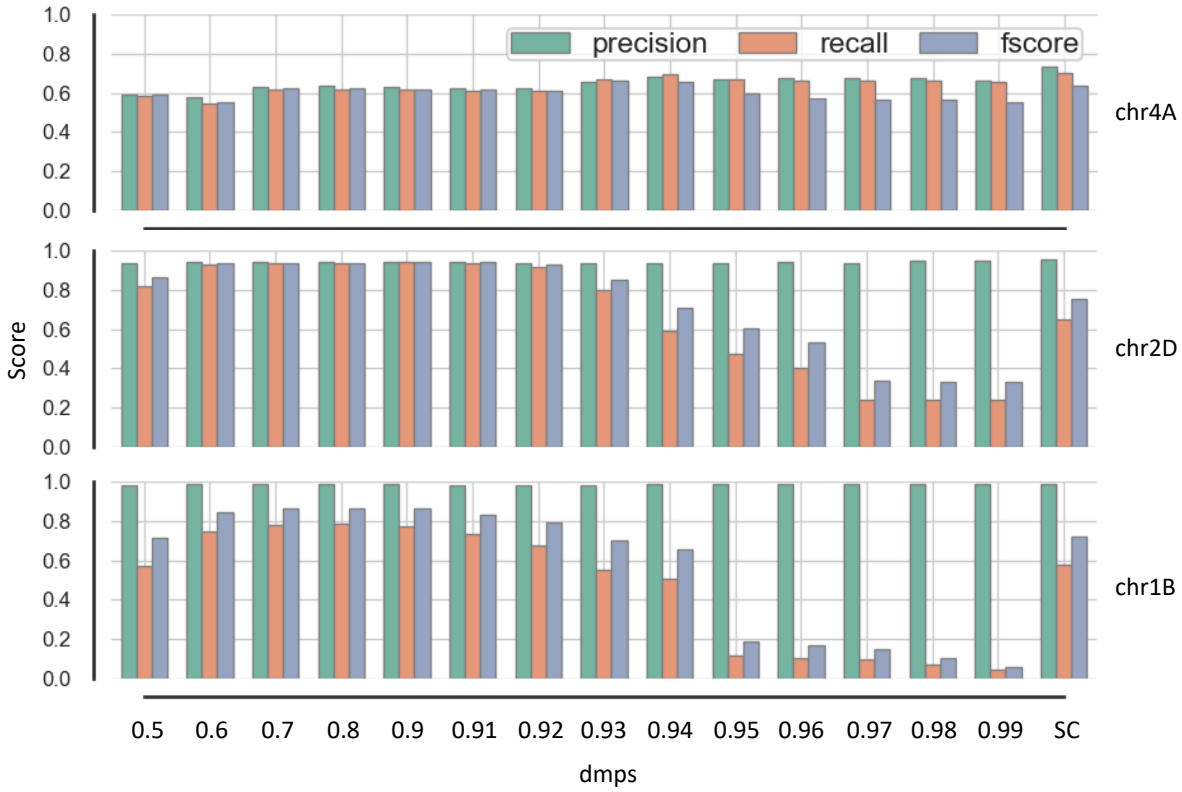


Fig. 3. 16. Precision and Recall based on alignments from (Brinton et al., 2020) vs IBSpy haplotypes.
For illustration we used the comparison from Mattis vs Julius. SC is the score selecting the *dmp* by window with the highest Silhouette Coefficient (SC) score. Mattis vs Julius share 52.2% IBS from chr4A, 10.8% in chr2D, and 3.9% in chr1B based on 5 Mbp haplotypes described in Brinton et al., 2020.

3.4.1.10. Redundancy test (*Ae. tauschii*)

To further validate the AP haplotype calls, we tested a highly diverse wild *Ae. tauschii* collection to determine if we could detect redundant genotypes which were previously identified by (Gaurav et al., 2022). In their study they reported several genotypes to be redundant based on KASP markers and 100,000 randomly selected SNPs. We used 265 accessions including 150 genotypes from lineage 2 and 115 genotypes from lineage 1 from their study and called IBSpy haplotypes to test for their redundancy. Lineage 2 genotypes were reported in (Gaurav et al., 2022) to be the closest donors of the D wheat genome. For this analysis, we included the unique *k*-mers when creating the *k*-mer databases from raw reads since several accessions had ~10-fold depth (Chapter 2, Supplemental Table S2.3). To call haplotypes we used the D genome syntenic regions of the 11 chromosomes-scale wheat assemblies (Chapter 2, Supplemental Table S2.1).

We tested the haplotype calls using different groups of lines to identify the effect on the number of lines used. We ran 45 comparisons across the whole genome using 1 Mbp window against the 11 genome references. These 45 comparisons included 42 comparisons between accessions which were identified as redundant in (Gaurav et al., 2022) (35 were from lineage 2 and six from lineage 1), and we randomly selected three comparisons not reported as redundant as controls (Supplemental Table S3.2).

Using the haplotype calls in 1 Mbp across the genome, in the comparison of lineage 2 samples we detected only two genotypes that had ≤ 99.4% haplotype calls similarity. All the other genotypes had ≥ 99.4% haplotype calls similarity. Two comparisons, comparison C7 (BW_01084 vs BW_01141) and C9 (BW_01141 vs BW_01189) had 61.8% and 61.9% haplotype calls similarity, respectively. These two comparisons had 98.93% and 98.02% similarity respectively when using KASP markers in (Gaurav et al., 2022), but it had no 100,000 random SNP data. In our analysis this similarity score was low across each of the chromosomes with chr6D having the least similarity (36.7%) in both comparisons (174 1 Mbp blocks out of 474 based on CS reference). Investigating the regions where these comparisons are different, we detected that different haplotype calls extended several windows in a block. As an example, we showed the region from 1 to 7 Mbp of chr6A (Fig. 3.17a). We also observed that these two comparisons called different haplotypes in the two comparisons exactly in the same windows. Since these two comparisons have a common

genotype (BW_01141), we could predict that BW_01084 and BW_01189 are redundant between them. This was confirmed in the C8 (BW_01084 vs BW_01189) having 99.7% similarity.

Comparisons in lineage 1 resulted in lower similarity scores compared to lineage 2 analysis with an average of 99.1% haplotype similarity among the six comparisons. By chromosome, chr3D had the average lowest similarity score with 98.6% where comparison C4 (BW_23932 vs BW_23934) had the lowest score with 98.0%. Investigating the genome regions, we detected a region from 291 to 352 Mbp common in all the similarity comparisons where AP called them differently. As an example, we showed the region from 289 to 358 Mbp of chr6A (Fig 3.17b). In (Gaurav et al., 2022) these set of comparisons had no KASP similarity analysis and the 100,000 random SNPs test had on average 99.95 %. This percentage was lower than the observed in Lineage 2 average comparison, which is 99.99% using the same 100,000 random SNPs. The three randomly comparisons (C43-C45) had on average 1.2% similarity (Supplemental Table S3.2).

Overall, our results suggest that AP efficiently detects redundant genotypes and the error call rate between redundant genotypes comparisons is relatively low. An explanation for this is that the genetic differences among *Ae. tauschii* accessions are high due to natural genetic variation accumulation and recombination over time. In the future, it would be valuable to incorporate the *Ae. tauschii* genome assemblies to capture more variations and structural variations absent in the current D wheat genome. This will impact breeding in the future since several research programs worldwide are increasing the use of synthetic hexaploid wheat lines derived from *Ae. tauschii* and other wheat wild relatives.

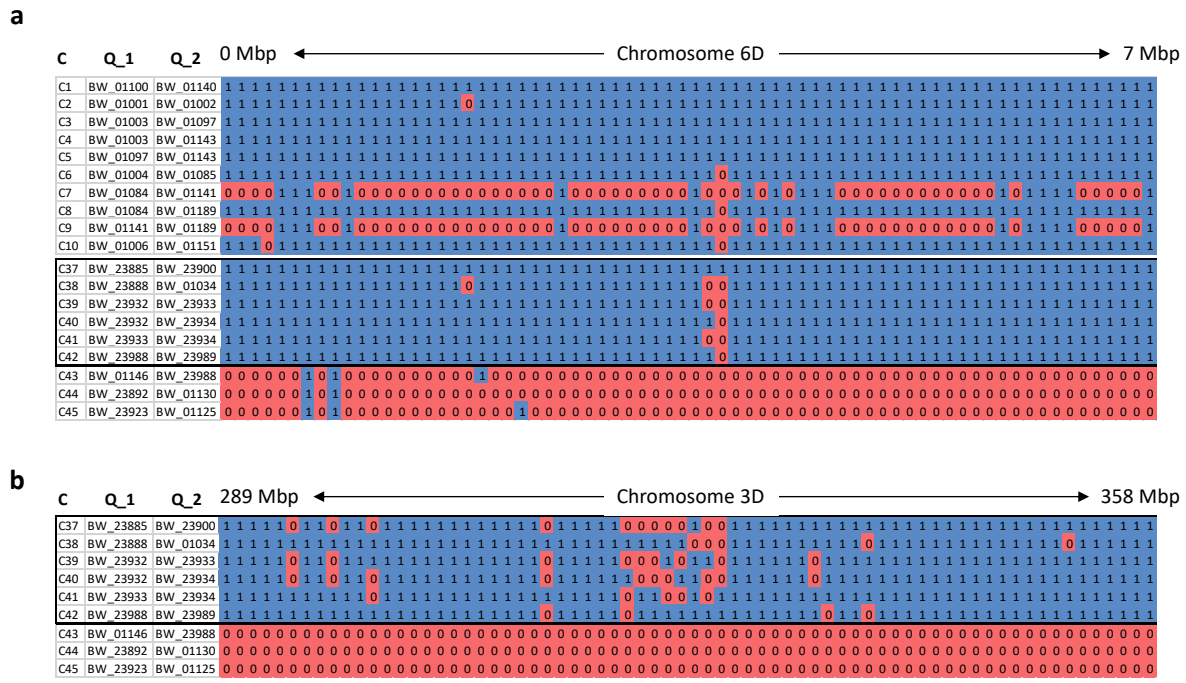


Fig. 3. 17. The AP haplotypes efficiently identify redundant genotypes.

a) Chromosome 6D haplotype calls from 0 to 7 Mbp. Comparison number ID (c), Q_1 (query_1) vs Q_2 (query_2) genotype comparisons. In blue with number "1" indicates that the two genotypes have the same haplotype in a specific window. "0s" in red indicate they are different and do not share the haplotype. C1 - C10 are L2 genotypes as described in (Gaurav et al., 2022). C37-C42 are L1 genotypes and C43 - C45 are randomly selected genotypes. **b**) Chromosome 3D haplotype calls from 289 to 358 Mbp.

3.4.1.11. Parent-child test (quality control).

The objective of this analysis was to complement and validate our quality control test of the haplotypes called by AP across the genome. This analysis involves the use of closely related modern wheat genotypes based on pedigree information. For this analysis, we identified genotypes for which we had raw data, and we could establish a clear parent-child relationship among them. For example, we could evaluate if the "child" accession was always similar or IBS to at least one of the two parents based on the AP haplotype calls. The aim was to detect the percentage and position of genome regions where at least one of the parents could be assigned as the donor or regions where the two parents were the donors (fixed regions among the parents). We also expected a percentage of regions defined as *"orphan regions"*, where none of the parents were the donors. The later was our error index since all regions from the child must come from at least one parent. We tested the same datasets and parameters as with precision and recall window and we used *dmps* with the highest SC score by window. We used three comparisons (**Table 3.2**).

Table 3. 2 Parent-child genotypes group comparisons.

Group	Parents	Child
G1	Rialto	Xi19
	Cadenza	Xi19
G2	Holdfast	Maris_Widgeon
	Cappelle_desprez	Maris_Widgeon
G3	Flame	Claire
	Wasp	Claire

For multiple comparisons we employed different group datasets for haplotype calling. In this initial test we summarize the results with the *"WatSeq_Pangenome_RAGT_ABD"* group. This group of samples includes 1,123 genotypes (Table S3.1). Our analysis was done by chromosome and by subgenome. In our G1 comparison (using Chinese Spring coordinates as our assembly template) we observed that Cadenza uniquely shares many more regions to its child Xi19 in most of the chromosomes with 6,241 windows (1 Mbp window size) in total. Only chr2A, chr2B, chr3B, and chr7A, Rialto uniquely shares more than >140 1 Mbp windows to Xi19. In total Rialto shares 892 windows (1 Mbp) across the whole genome. This was expected since the pedigree of Xi19 is Cadenza//Rialto/Cadenza (i.e., the Rialto * Cadenza F1 is backcrossed to Cadenza). The total number and average of 1 Mbp windows across the genome shared by the two parents were 2,598, and 123 respectively. The total number of windows detected as "*orphans*" was 4,344 which was 30% of the whole genome of Chinese Spring. Most of this percentage was from the D subgenome with more than half (58.6 %) being in this category compared to the regions shared by Rialto (2.6 %), Cadenza (33.7 %), and both parents (fixed regions) with 5.0 % (Supplemental Table S3.3).

In G2, Cappelle_Dezprez and Holdfast uniquely shared 20.5 % and 15.2 % of the whole genome with Maris Widgeon (child) respectively. Both parents shared 46.8 % of a common region while 17.3 % of the genome was in the "*orphans*" category. Overall, these two parents shared equal genome regions to their child. Again, the D subgenome had many more "*orphans*" windows than the A and B sub genomes. An example of the chromosome 2B of G2 comparison is depicted in (Fig. 3.18). In this comparison 56% of the chromosome is shared by both parents to the child which is equivalent to 376 Mbp in genome regions. 4.93% is shared uniquely by Holdfast (33 Mbp), and 34.54% is shared uniquely by Cappelle D. (231 Mbp). 4.44% was categorized in the "*orphan category*". These shared percentages are reflected in in the "*variations fingerprint*" heatmap comparisons and the hypothecia shared blocks based on similarity (Fig. 3.18b).

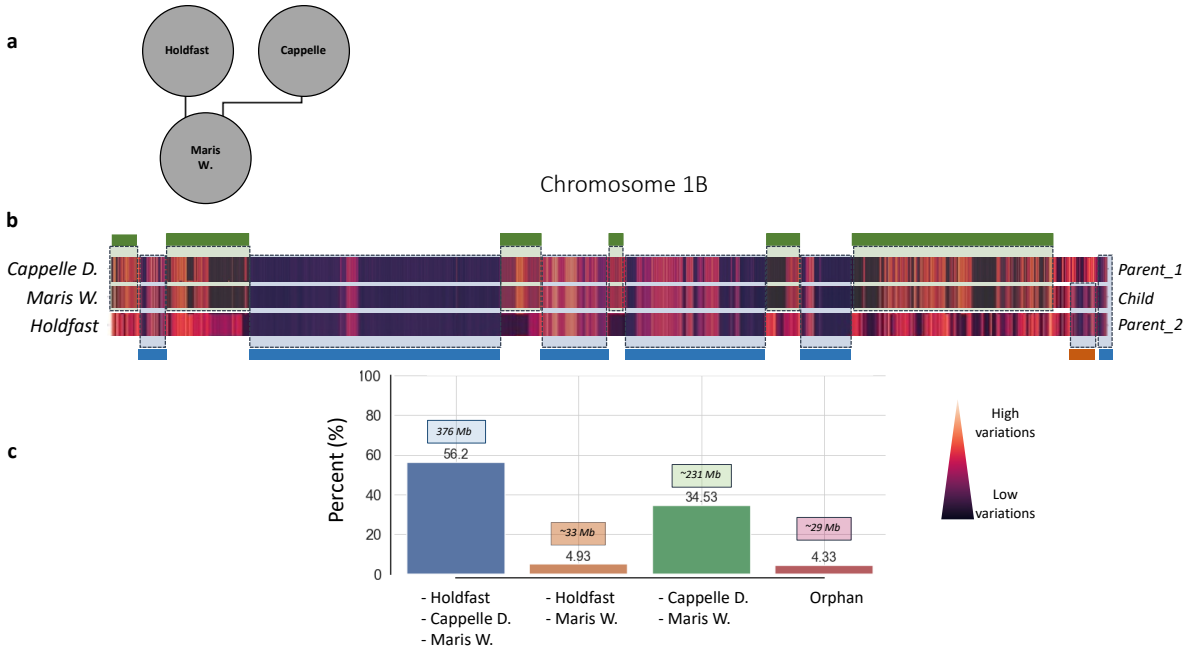


Fig. 3. 18. The parent-child test of Maris Widgeon pedigree sharing haplotypes across Chr1B.

Example of shared haplotype blocks on chromosome 1B from Cappelle Desprez (Parent_1) and Holdfast (Parent_2) to Maris Widgeon (Child). **a**) Maris W. pedigree. **b**) heatmaps of the three genotypes against CS reference depicting genome similarity based on *variations* count and the hypothetical blocks uniquely inherited by Cappelle D. to Maris W. (green bars) and from Holdfast to Maris W. (orange bars). Blue bars indicate hypothetical regions shared by the two parents (fixed regions). **c**) The percentage of the haplotypes called from each of the parents as being identical to the child based on the AP haplotypes. The "*orphan*" category indicates the regions not assigned by any of the parents (error calls).

In the G3, the category of two parents sharing the region to Claire (the child) was the highest with 63.5 % of the whole genome. Flame and Wasp uniquely share 14.8 % and 8.8 % respectively. The "*orphan category*" had 12.6 %, this being the group with the lowest error rate (orphan) of the three comparisons. The A, B, and D sub genomes had 9.2, 9.4, and 21% in the orphan category (**Supplemental Table S3.3**). The high rate seen in the two parents sharing regions agrees on both Flame and Wasp sharing a common parent one generation above (Hobbit) as described in section **3.4.1.3.**

To further investigate the genome regions with error rates, we then investigated the *variations* count basic statistics in 1 Mbp window used to call haplotypes. We noticed that the number of haplotypes per window (**dmp_num_avr_total**) in the *orphan* category had higher number of haplotypes compared to the other three categories (parent_1, parent_2, or shared). This was true for the three comparisons (G1, G2, and G3). Within this category, the D subgenome had the highest average number of haplotypes per 1 Mbp window with 46, 41, 42.8 in the G1, G2, and G3

respectively. The average standard deviation (std_avr_total), mean (mean_avr_total), and median (median_avr_total), were much smaller in the *orphan* than in the individual parents or both parent category in all subgenomes. However, it was much smaller in the D subgenome than in the A, and B subgenomes in all three groups comparisons (G1, G2, and G3). Finally, we investigated the skewedness average (skew_avr_total) of the *variations* in the 1 Mbp window blocks and detected that the *orphan* category and the D subgenome had overall higher skewness of the *variations* count data compared to the uniquely parent category, both parents category, and the A and B subgenomes (Supplemental Table S3.3).

In summary, our results suggest that the parent-child test can efficiently detect genome regions with haplotype calls defined erroneously by the AP analysis in the *orphan* category. Our results show that the D subgenome has the most problems to correctly call haplotypes as this is where most of the errors (i.e., *orphan* windows) are located. Investigating into the *variations* statistic by blocks of windows, we detected data features specific for the erroneous windows. These features are mostly related to the low diversity of the D subgenome and skewedness of the data and therefore low *variations* count in the 1 Mbp window.

Adding diversity to the Parent-child test with *Ae. tauschii*

Based on our previous results on haplotypes error call rates, we hypothesised that error calls were due to low diversity and skewedness in the distribution of the *variations* data within the 1 Mbp window. To address this hypothesis, we incorporated the 265 *Ae. tauschii* genotypes (which have on average ~10-fold coverage) from (Gaurav et al., 2022) into the haplotype calls analysis. Wild relatives harbour high genetic diversity and therefore we predicted that this would alleviate the low diversity of our wheat D genome dataset. For this analysis we used the group "WatSeq_Pangenome_RAGT_AeTau_D" which includes the same set of genotypes as used in our previous example plus the *Ae. tauschii* collection. This analysis was done only for the D subgenome since it was the genome with the highest error rate calls.

After adding the *Ae. tauschii* genotypes and calling haplotypes, we evaluate different features of the 1 Mbp window. We found that overall, there is no change in the number of windows assigned to each of the parent category before and after adding the *Ae. tauschii*. For example, out of the total number of 1 Mbp windows assigned to Cappelle D. as being the donor of Maris W., after adding the *Ae. tauschii* data, this value did not change and the number of windows in the 1 Mbp was similar to the *orphan* category (errors). Therefore, we discarded the number of windows used in the haplotype call to have a high impact on the erroneous haplotypes called (Fig. 3.19a). A second variable explored was the number of haplotypes per 1 Mbp window. Our results indicate

that the number of haplotypes per 1 Mbp was similar in all categories (Cappelle. D. Holdfast, all_ibs, and *orphans*) before adding the *Ae. tauschii* data, and the number increased slightly after across all four categories (Fig. 3.19b). This was expected since there will more haplotypes coming from the *Ae. tauschii* accessions absent in hexaploid wheat. The standard deviation (STD) and the mean for the *variations* count in the 1 Mbp windows increased after adding the *Ae. tauschii* data (Fig. 3.19c, d). Conversely, the skewness was reduced (Fig. 3.19e, Supplemental Table S3.3).

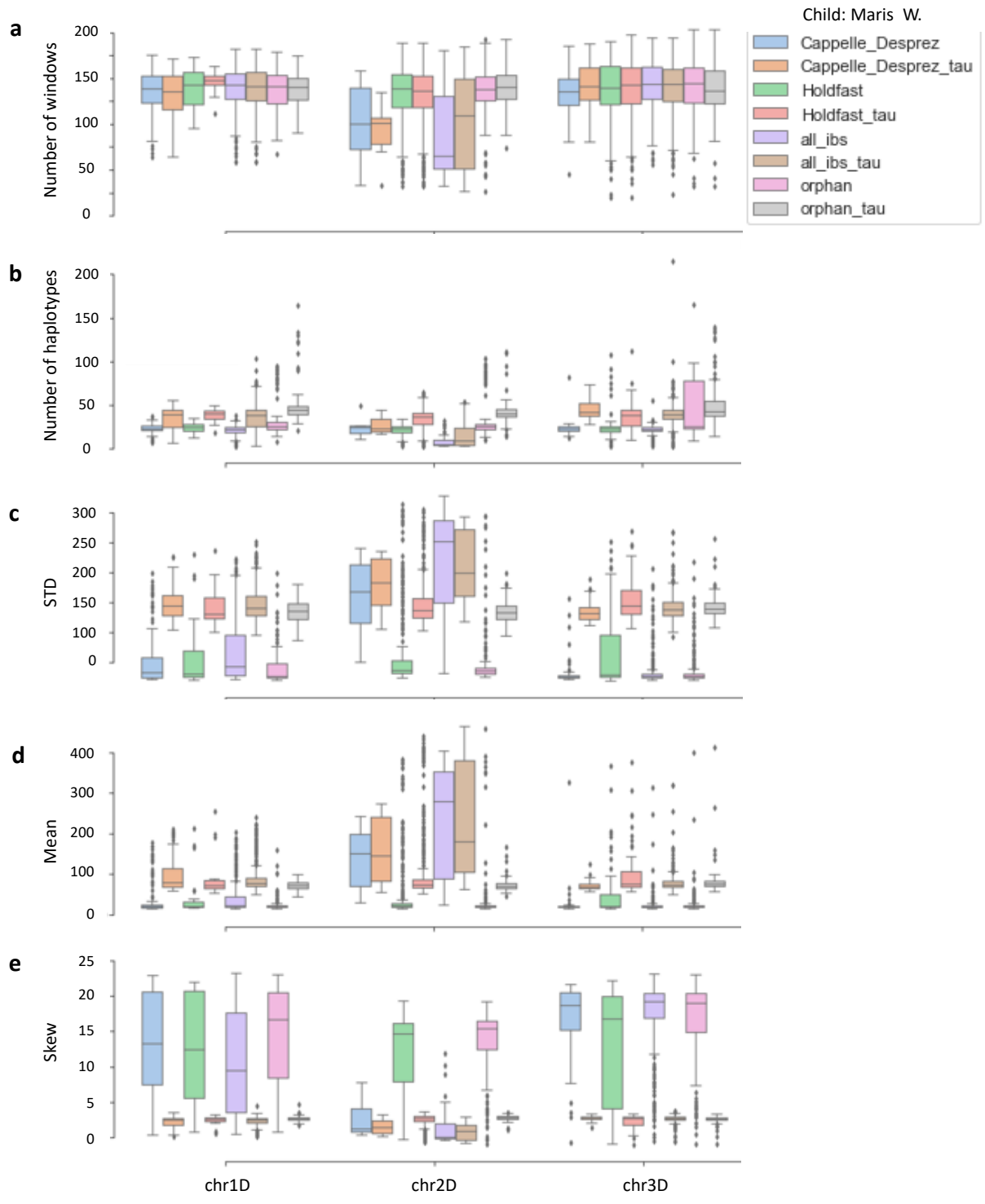


Fig. 3. 19. D subgenome parent-child analysis before and after adding 265 *Ae. tauschii* accessions.

As an example, we show data from three chromosomes (chr1D, chr2D, and chr3D). In the categories box, names having the suffix "_tau" is the data after adding the *Ae. tauschii* samples to the analysis. The category "all_ibs" includes windows where both parents had the same haplotype as the child (Maris W.). **a**) number of syntenic windows window per category. **b**) number of haplotypes in 1 Mbp window across chromosome. **c**) and **d**) standard deviation (STD) and

Mean of the *variations* count in the 1 Mbp window respectively. **e**) skewness of the *variations* count data in the 1 Mbp window.

3.4.1.12. Tracking pedigree haplotypes

Before, in section **3.4.1.3** we tracked back IBS regions into Claire from Flame, Wasp, and Hobbit. However, the analysis was limited to a small number of samples against a single reference in a pairwise comparisons. Despite this limitation, we still were able to detect large IBS regions in agreement with (Brinton et al., 2020) haplotypes based on mummer alignments. Later in our analysis we demonstrated that we could detect haplotypes in 1 Mbp windows from multiple genotypes by clustering AP. Using these haplotypes calls, in this analysis we tracked back haplotypes from landraces into modern wheats. As an example, we investigated the conformation of chromosome 6A of the reference Mattis. We aimed to identify regions from landraces that are maintained in modern cultivars. As a common reference we use CS to call haplotypes and we focused on the landraces queries only.

We found that out of 827 Watkins landraces, 15 make up the chr6A of Mattis and share large intact haplotypes blocks. A relatively small block at the start of the chromosome was shared entirely with WATDE0306 with a total of 20 consecutive 1 Mbp haplotype blocks (**Fig. 3.20a**). WATDE0950 and WATDE0786 shared the largest blocks where WATDE0950 shares a 357 Mbp intact block (**Fig. 3.20b, c,** teal bar) and WATDE0786 shares two main blocks of 34 Mbp and 74 Mbp (light blue bars) on either side of the WATDE0950 segment respectively. The remaining landraces share relatively short and fractionated blocks across the chromosome.

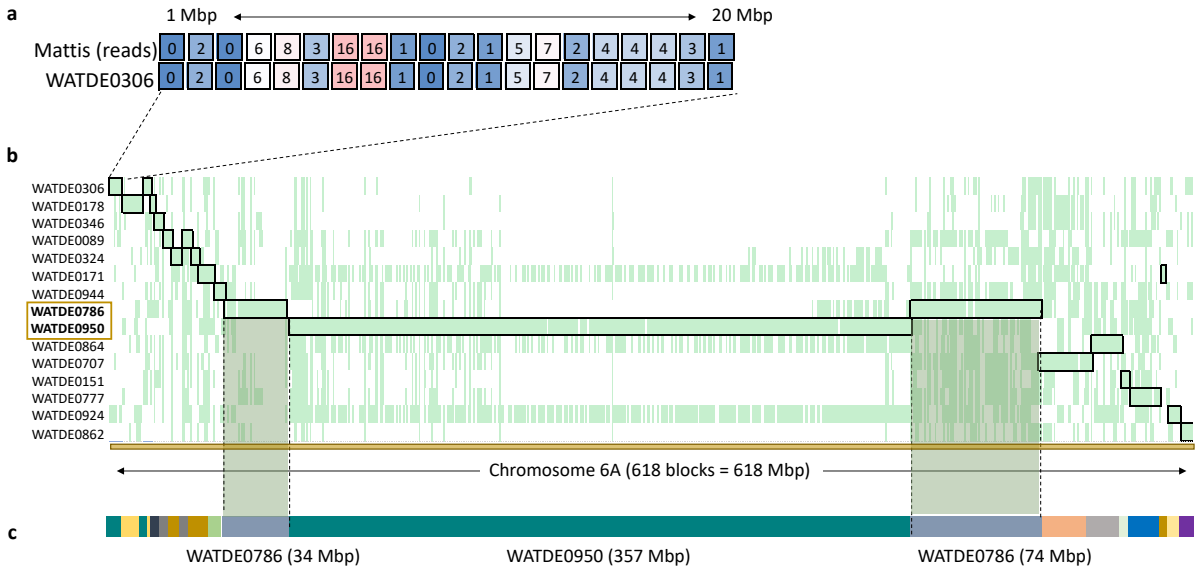


Fig. 3. 20. Large haplotype blocks are maintained into modern wheats from landraces.

a) AP haplotype calls comparison of a landrace (WATDE0306) vs the cultivar Mattis (raw reads) in 1 Mbp window on chr6A using CS as a reference. Twenty consecutive IBS windows shared between the two genotypes are shown as indicated by having the same "haplotype number" in each of the window. **b**) haplotype calls of the 15 landraces having the same haplotypes as Mattis across the chromosome were transformed and coloured in green. In total, 15 landraces make up the chr6A of Mattis (black boxes). WATDE0306 shares the start of the chromosome as depicted in **a**). WATDE0950 and WATDE0786 (boxes in yellow) shares the largest blocks. WATDE0950 shares 357 Mbp with Mattis **c**) teal bar, and WATDE0786 shares two main blocks of 34 Mbp and 71 Mbp (light blue bars). The remaining landraces share smaller blocks across the chromosome.

3.4.2. Haplotype based GWAS

In our previous analysis we *de novo* called haplotypes using IBSpy and the WatSeq dataset. We detected major haplotype blocks that have been maintained intact from landraces into modern elite cultivars. In this analysis, we aimed to determine if IBSpy haplotypes could be employed to run haplotype-based genome wide associations (hapGWAS).

For phenotypic data, we used field data collected in our group in 2020 and 2021 for several agronomically important traits (**Supplemental Table S3.4)** (Backhaus A. & Chen A., unpublished data) and used the 1 Mbp window IBSpy haplotypes. This phenotypic dataset was previously used in our lab to identify several GWAS associations genome regions (unpublish data) using SNP variants generated from the alignments of the WatSeq raw reads to the CS reference (RefSeq v.1.0) genome and publicly available software. To run hapGWAS, we adjusted the numeric haplotypes by window to presence/absence and unique names (See methods and scripts). We also adjusted the kGWAS (https://github.com/wheatgenetics/owwc/tree/master/kGWAS) described in (Gaurav et al., 2022) to run associations with IBSpy haplotypes using a presence/absence matrix. Since our pipeline uses the pangenome assemblies, we ran hapGWAS using each of the eleven references in turn. This feature of hapGWAS haplotypes is of importance since it can run multiple tests capturing genome information that might be private to a specific genome reference. On the other hand, alignment methods would require running and call SNPs to each of the references before a common GWAS which is computing demanding for a genome of the size of hexaploid wheat.

3.4.2.1. Spikelet number

Multiple publications have reported strong genome associations between the *WAPO* gene and spikelet number in wheat (Kuzay et al., 2022; Zhang et al., 2018). In our lab, this trait has been detected consistently in two years of phenotypic data using SNP based GWAS and different

software (Backhaus A. & Chen A., unpublished data). Therefore, we used this trait as our positive control.

Our results with hapGWAS detected a major hit at the end of chr7A, consistent with the previous analyses (**Fig. 3.21a** phenotype scores **Supplemental Table S3.4**). This hit was present in the 11 references used. In previous studies, the gene responsible for the trait (*TraesCS7A02G481600, WAPO-A1*) was detected at 674,081,462 – 674,082,918 bp (RefSeq v1.0). In the hapGWAS analysis, the haplotype with the strongest association (chr7A__chi_673000005) was haplotype 5 located at 672 - 673 Mbp. However, this haplotype was negatively associated with the trait. Haplotypes chr7A__chi_675000010 and chr7A__chi_675000006 were negatively and positively associated with the trait, respectively, and were in the 674 – 675 Mbp window (**Fig. 3.21a**). These results suggest that the region harbours several variations surrounding the responsible gene and therefore, multiple alleles that might have different effects on spikelet number.

3.4.2.2. Max floret number

A second strong and stable association hit was detected for maximum number of florets (**Fig. 3.21b,** phenotype scores **Supplemental Table S3.5**). This hit was equally strong in Jagger, Julius, and Landmark located at 645 - 646 Mbp in Jagger and Julius and at 652 – 653 Mbp in Landmark. The hit in CS (RefSeq v1.0) was located at 648 - 649 Mbp. We investigated if the *WAPO-B1* gene was located in each of the corresponding window in the other pangenomes. We found that *WAPO-B1* gene in Jagger is located at 651 Mbp and at 651 Mbp in Julius which is different to 645-646 Mbp hapGWAS hits. The gene in Landmark is at 658 Mbp and in Chinese at 649 Mbp. Therefore, the hapGWAS hit only overlaps with *WAPO-1B* in the CS reference. The corresponding region appears to be translocated in reference ArinaLrFor at position 104 – 105 Mbp and the *WAPO-B1* in ArinaLrFor coincides to be close to the region at 103 Mbp. The gene content of ArinaLrFor is similar to CS in the candidate region. The set of accession having the favourable haplotype are similar and consistent in each of the references. The favourable allele in Jagger is chr7B__jag_646000011 and in CS is chr7B__chi_649000011. Overall, although the exact position with the *WAPO-B1* did not coincide in all the references, in all cases the proximity with the strongest hit was close.

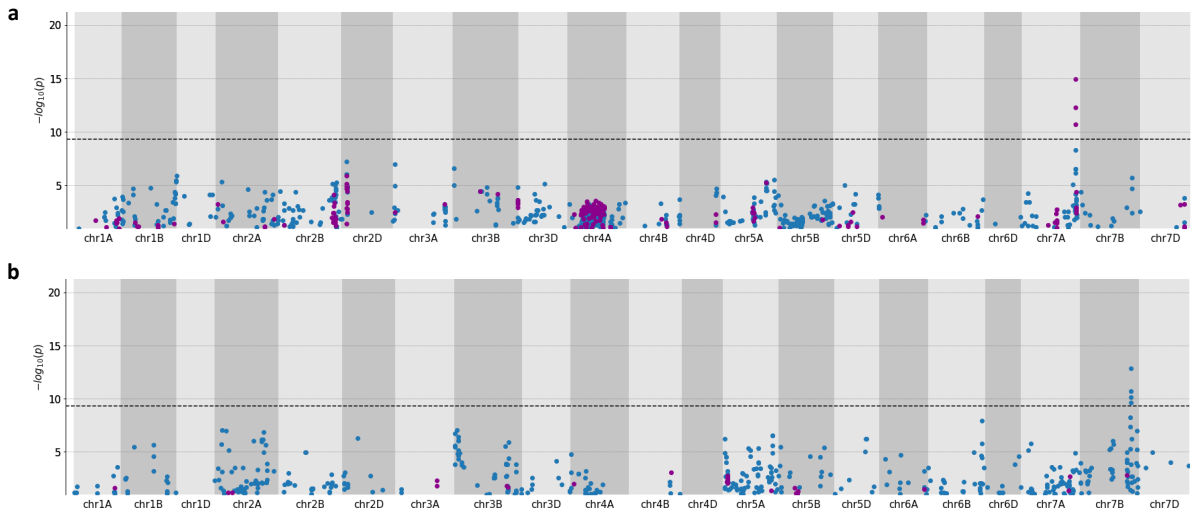


Fig. 3. 21. Validation of hapGWAS using known spike related traits hits.

a) Spikelet number hapGWAS located at 673 Mbp on chr7A based in CS (RefSeq v1.0) genome reference. **b**) Max floret number. Spikelet number hapGWAS based in Jagger reference chr7B. The hit in CS for Max floret number was located at 649 Mbp matching the *WAPO-B1* location. *x*-axis indicates the (*-log10(p)*) association. The horizontal line indicates the Bonferroni-adjusted *–log P* value threshold between 9.1 and 9.3 as described in (Gaurav et al., 2022) for *k*-mer comparisons. At the time of writing this thesis we haven't adjusted the threshold for the multiple testing for the number of haplotypes. This will be a follow up analysis for a continuation of the project.

3.4.2.3. Yellow rust GWAS QTLs

Several publications have reported genome associations for rust resistance genes in wheat (Marchal et al., 2018). Those analysis have mainly employed single SNPs. In this analysis we aimed to evaluate the hapGWAS pipeline using phenotypic data for rust resistance (**Supplemental Table S3.6**). This dataset was previously used by our collaborators and detected GWAS hits across the genome using SNPs (Cheng et al., under revision) using the CS reference (RefSeq v1.0). The phenotypes were scored against two yellow rust isolates here referred from the ancestral lineages groups Pink and Red defined using field phenotypic data from the UK described in (Hubbard et al., 2015).

Rust Pink PST lineage

Using IBSpy haplotypes and hapGWAS for the rust Pink lineage, we detected consistent and strong -log10*(p)* hits on chr7A of Jagger, Landmark, Norin61, Spelta, and Stanley. The haplotype with the strongest association was chr7A__jag_101000022 located at 100 – 101 Mbp window. Within this window there are two NLR genes; *TraesJAG7A03G03834120* (NLR) and *TraesJAG7A03G03834160*

(NLR). NLR genes are well known to be involved in disease resistance traits. These two genes are also present in the other references with the associations mentioned above and therefore, conserved. Previous reports have detected two genes on chr7A involved in rust resistance (Lr20 and Lr47) (Kumar et al., 2022). However, those genes are located at the end of the chromosome and provide resistance to leaf rust; hence they are very unlikely to be the same gene or locus.

In our hapGWAS analysis, a second association was detected on chr4A from 739 – 740 Mbp of Jagger. 12 genotypes had the favourable haplotype for the resistance in the 739 Mbp window. Interestingly, WATDE0088 has the haplotype but it is a susceptible phenotype. This would be a candidate to explore further since it could be that a few mutations in the resistant haplotype would have led to the loss of the resistance. Svevo has a similar haplotype to the resistant (H16) which could indicate that the origin of the haplotype conferring the resistance phenotype is a tetraploid wheat (Fig. 3.22a, b, c)

An additional GWAS hit was detected in Julius at 751 - 752 Mbp. Haplotype chr4A__jul_752000016 had the strongest association and with the favourable affect (Fig. 3.22d). Within this genome interval there are several disease-related genes: *TraesJUL4A03G02241820* (NLR), *TraesJUL4A03G02241900* (NLR), *TraesJUL4A03G02241950* (Kinase), *TraesJUL4A03G02241990* (NLR), *TraesJUL4A03G02242000* (NB-ARC), *TraesJUL4A03G02242030* (Kinase), *TraesJUL4A03G02242110* (NLR), *TraesJUL4A03G02242130* (NRL). This hit was also detected in ArinaLrFor, CS, and Lancer references. Further investigation will be required to validate if some of these genes are the causal of the rust resistance in wheat against the Pink PST.

A third haplotype association was detected on chr4D of CS, Julius, Lancer, Mace, Norin61, Spelta, Stanley, and Mattis references. However, no canonical disease resistant genes were detected or characterized in the windows with the strongest association. It could be that the gene(s) or polymorphism responsible is not present or assembled in the pangenome references but hapGWAS still detects the signal based on the surrounding genome information and the *variations* profiles of the samples with the positive alleles are captured as a novel haplotype. We will show a case study on this topic in section 3.4.2.4 related to this hypothesis.

Rust Red PST lineage

A second lineage called “Red” was used to phenotype the same set of Watkins accessions and score the disease effect. Our results running using these phenotypes and hapGWAS, detected three main hits in all comparisons. A very consistent hit was detected on chr7B long arm in ArinaLrFor, Chinese Spring, Jagger, Julius, Lancer, Landmark, Mace, Norin61, Stanley, and Mattis.

The hit on Julius was located at 539 – 540 Mbp window. The haplotype with the strongest association was chr7B__jul_540000005 (**Fig. 3.22e**). In this window there were several disease-resistant related genes including: *TraesJUL7B03G04227170* (NLR), *TraesJUL7B03G04227520* (Kinase), *TraesJUL7B03G04227680* (kinase), *TraesJUL7B03G04227720* (Kinase). A strong hit was also detected in ArinaLrFor at 217 – 218 Mbp where only the *TraesARI7B03G04104640* (Kinase) gene related to disease is located. For this association hit, haplotype chr7B__ari_218000019 was identified as having the favourable effect.

A second and consistent hit was on chr3B of ArinaLrFor, CS, Jagger, Julius, Lancer, Landmark, Mace, Norin61, Spelta, Stanley, and Mattis. Window 106 - 107 Mbp and 107 – 108 Mbp had the strongest associations based on Julius reference where the haplotypes chr3B__jul_107000018 and chr3B__jul_108000023 had the strongest association.

The last hit was detected on chr4D in CS, Jagger, Julius, Lancer, Mace, Norin61, Spelta, and Mattis. The hit was located in the same window described before with the rust Pink lineage. This suggest that a QTL in this chromosome may confer resistance to both lineages. The hit in Julius reference was located at 465 – 466 Mbp with haplotype chr4D__jul_466000015 having the favourable effect.

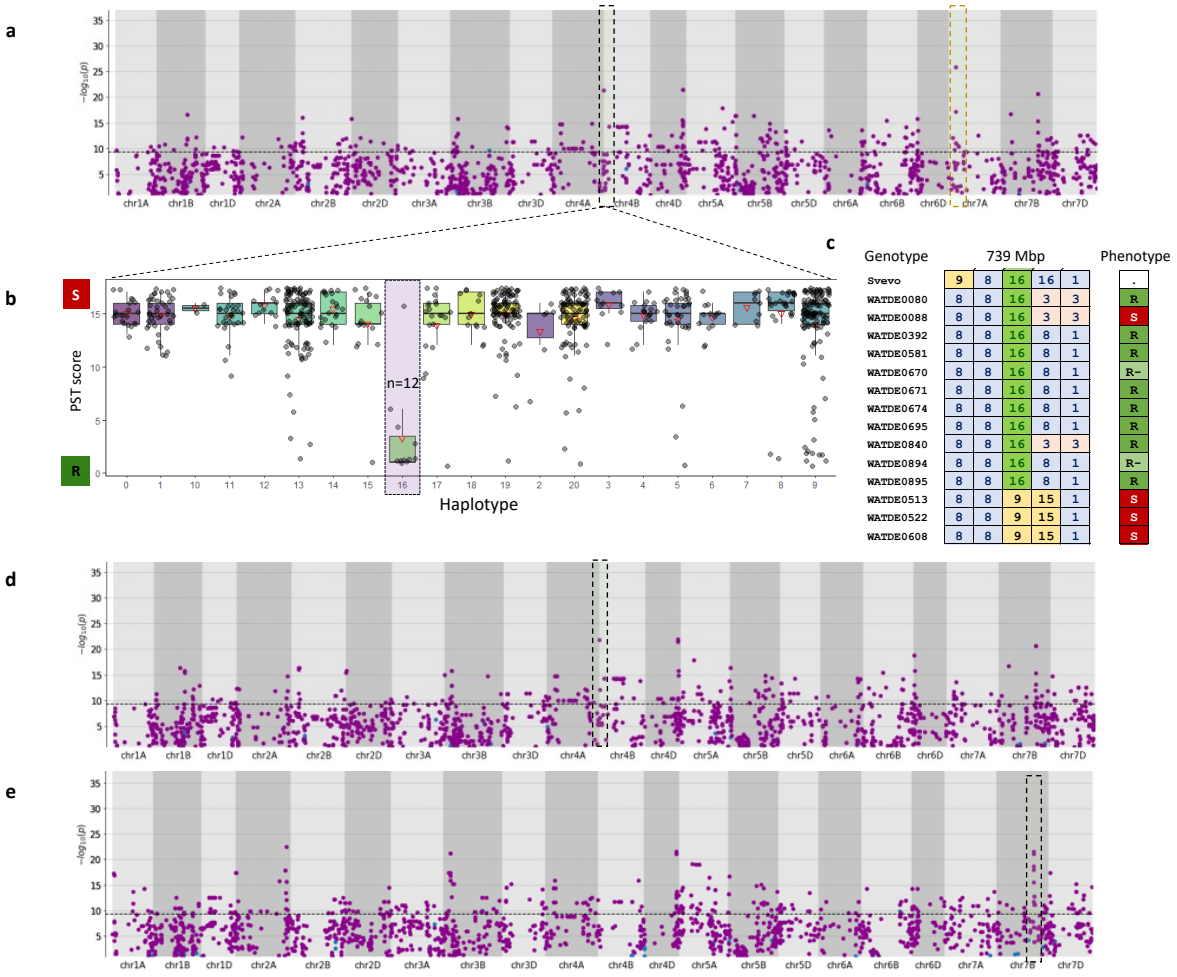


Fig. 3. 22. A novel rust resistant associations detected by hapGWAS at 1 Mbp resolution.

a) Jagger reference hapGWAS hit at 100 Mbp of chr7A (red box) and 739 – 740 of chr4A (black box). **b**) Pink rust lineage PST score and haplotypes. Resistant score (R, green), susceptible score (S, red). Haplotype H16 have n=12 resistant genotypes (purple box) and one susceptible. **c**) Genotypes having the favourable haplotype for the resistance in the 739 Mbp window. Svevo has the same haplotype for the resistance (H16). WATDE0088 has the haplotype for the resistance but it has a susceptible phenotype. **d**) hapGWAS hit in the Julius reference at the similar positions as in Jagger. **e**), hapGWAS association hit at 539 Mbp on chr7B of Julius reference using the Red rust PST lineage.

3.4.2.4. Wheat Blast

Wheat blast is caused by *Magnaporthe oryzae Triticum (MoT*) and affects wheat spikes resulting in bleached spikes and grains. It was first detected in Brazil from where it has been propagated to other wheat producing countries. There are a few wheat blast resistant QTLs at seedling stage on chromosomes 2B, 4B, 5A and 6A (Goddard et al., 2020; Juliana et al., 2020). (Juliana et al., 2020) identified QTLs on chromosome arms 2AS, 3BL, 4AL, and 7BL using two isolates (Bolivia and Bangladesh). Several of the markers associated are located at ch2A which is known to be the site

of the 2NS translocation from *Ae. ventricosa* (Cruz et al., 2016). Historically five resistance genes have been identified: *Rmg2*, *Rmg3*, *Rmg7* (Tagle et al., 2015), *Rmg8* (Anh et al., 2015), and *RmgGR119* (S. Wang et al., 2018), but the underlying identity of the genes remains unknown.

In a pilot analysis, our collaborators Dr. Paul Nicholson and Tom O'Hara previously identified a strong association using *k*-mers and GWAS on chr2A and blast resistance to the *Super Race Avirulence* (SRA) *MoT* isolate (Supplemental Table S3.7). Based on phenotypic evaluations, they hypothesised that Mattis (one of the pangenome cultivars) carried the resistance haplotype against the SRA isolate. In their pilot analysis using RenSeq data from a subset of Watkins landraces (O'Hara et al., under revision), they detected WATDE0056, WATDE0527, WATDE0568, WATDE0592, WATDE0720, and WATDE0786 to be susceptible SRA isolated and having a *k*-mer signal when running a *k*-mer GWAS equivalent to the resistant haplotype on chr2A QTL using Mattis as a reference. Therefore, the hypothesis was that if those susceptible lines had the positive *k*-mer signal, these genotypes may have a single or a few SNPs in the regions that cannot be detect with the raw *k*-mer analysis.

We therefore evaluated the region using the Mattis reference genome and the IBSpy *variations* count data, which has the resistant phenotype to SRA and should have the favourable allele in the region. In our initial haplotype analysis, we observed (in a cluster map and by visualization analysis on haplotype calls) that a set of lines had two main blocks (we called them: *region 1* and *region 2*) at the end of chr2A consistently present in the resistant lines using Mattis as a reference. A recombination between these two main blocks was present is several Watkins lines but it was kept intact in several modern cultivars based on Mattis (Fig.3.23a).

A third group of modern cultivars and three Watkins had the block *region 1* but lacked the *region 2* block. We observed that Stanley, an additional chromosome-level assembly also had the positive block. In our set of samples, we also included publicly available data of wild wheat relatives, and we noticed that *T. timopeevii* accession 33255 had a similar block in this region (Fig. 3.23a, top genotype R1 and R2 regions). Therefore, we hypothesised that the resistance haplotype may have originated from a tetraploid genotype.

Alignment analysis and results: In a follow up analysis to further explore the QTL region, we aimed to identify SNPs by aligning raw reads of a set of accessions of both resistant and susceptible genotypes against the Mattis genome. We aligned and detected that three of the susceptible lines having the haplotype block at region 788 - 789 position had a SNP in one of the candidate genes *TraesSYM2A03G00828360*, a kinase ATP binding protein. This suggested and supported the hypothesis that this gene had a mutation that could lead to the loss of resistance in the susceptible genotypes which otherwise have the favourable *k*-mer signal and IBSpy haplotype block.

Additional lines resistant to SRA: In addition to the Watkins accessions, we noticed that this block was also present in several other modern cultivars not included in the initial GWAS analysis with raw *k*-mers, these lines were mainly from the UK. To assess if those modern cultivars were also resistant, we tested 26 additional genotypes against the SRA isolated in a detached leaf blast assay (O'Hara et al., under revision). The results from this analysis demonstrated that all genotypes tested having the haplotype block in the region are resistant to SRA at 6 days post inoculation (6 DPI) which further supports the hypothesis of the candidate region (Supplemental Table S3.8). Interestingly, one modern cultivar, Epson, also displayed a very high resistance against SRA. Epson is a recent (2014) variety of Syngenta breeding company which could be an indication of a similar pedigree to Mattis (also from Syngenta).

Mattis reference: Using the haplotypes by 1 Mbp window obtained from IBSpy, we ran hapGWAS using the phenotypic data from the blast SRA isolated including only Watkins lines to further validate the genome region association with the phenotype. In the initial test, we used Mattis as a reference since it has the positive allele for the resistance. As expected, we obtained four major hits at the end of chr2A. The highest hit was located at 794 – 794.2 Mbp and the haplotype with the strongest association was chr2A__sym_794250005 (Fig.3.23b). The presence/absence correlation of having this haplotype and the resistant phenotype was -0.7. This indicates that having this haplotype the disease score is low (resistant). In this window, there was a gene named *TraesSYM2A03G00830550*, an NLR, at position 2A:794,036,223 - 794,041,272. This window corresponds to the initial candidate "*region 2*" that we detected by looking at the variations cluster map at the end of the chromosome.

A second candidate window with strong association that matched our initial candidate *"region 1"* was at 788 - 789 Mbp (Fig.3.23b) and the haplotypes with the favourable alleles were chr2A__sym_789000015 and chr2A__sym_789000019. In total, 26 genotypes had one of those haplotypes and four of them were susceptible. This 1 Mbp window has the *TraesSYM2A03G00828360* gene, a kinase ATP binding protein. This gene is of interest because we previously detected SNPs in three susceptible genotypes which have the same haplotype block (WATDE0592, WATDE0568, and WATDE0527). This region also falls within the QTL interval from the initial *k*-mer based analysis.

In total, there were 27 haplotypes in the associated window and three of them had the positive effect: chr2A__sym_794250003 (H3), chr2A__sym_789000015 (H15), and chr2A__sym_789000019 (H19). 28 modern elites had H3 (Fig.3.23c, purple) and are SRA resistant. H15 and H19 contained only Watkins genotypes which are predominately resistant (Fig.3.23c). We included 348 Watkins genotypes with scores in the hapGWAS analysis. The 24 elites were not

included in the hapGWAS, but they had the positive haplotype (H3). Finally, we observed in our previous analysis (cluster map in Mattis) that the entire block region where the GWAS hits were identified, is maintained in multiple accessions but not in all, suggesting that a recombination took place between these two GWAS hits and there may be two independent QTLs. This would be supported by the significant haplotype associations in region 1 and region 2.

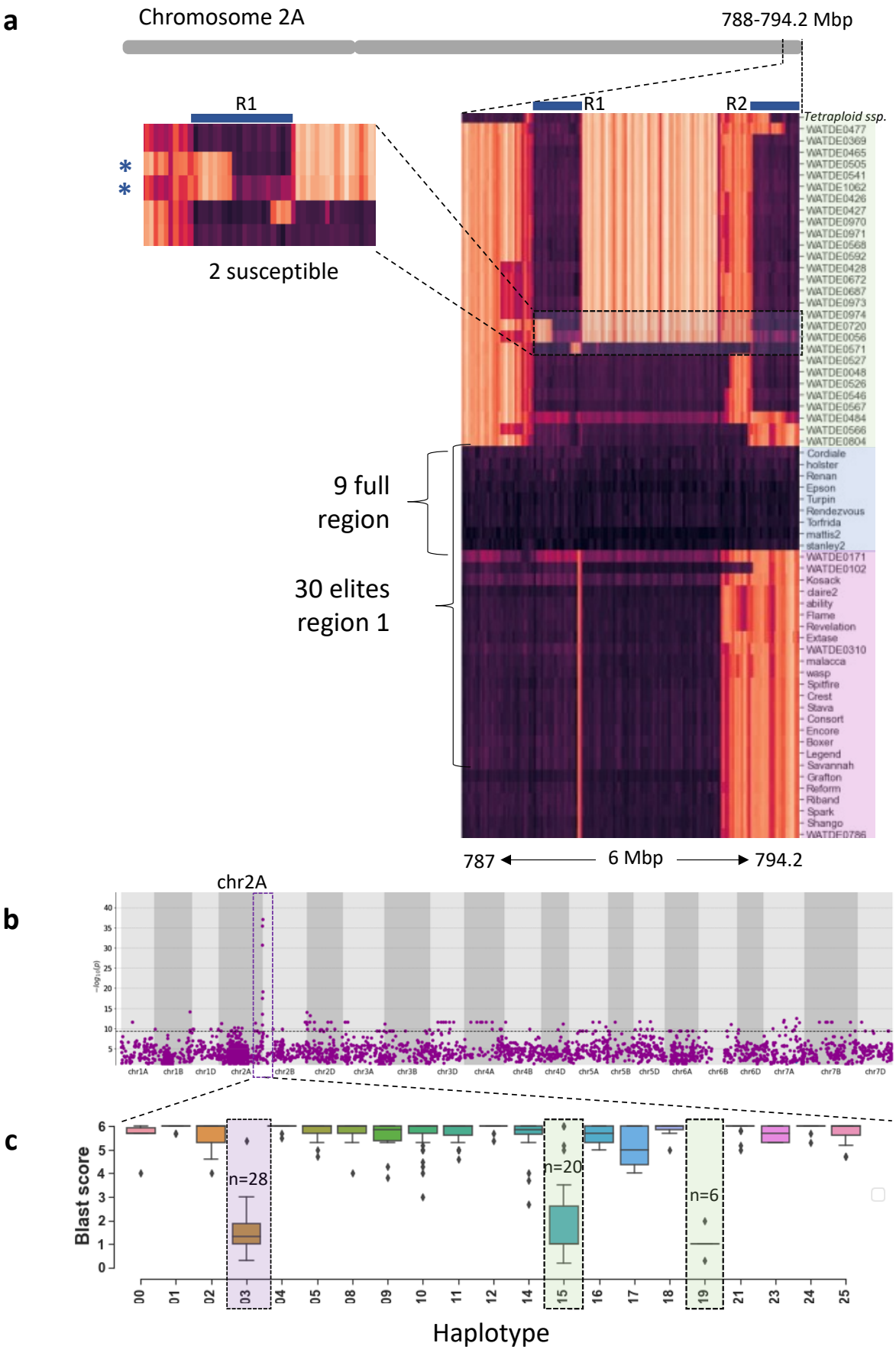

a
Chromosome 2A
788-794.2 Mbp
R1
R1
R2
Tetraploid ssp.
2 susceptible
9 full region
30 elites region 1
787
6 Mbp
794.2
b
chr2A
c
Blast score
n=28
n=20
n=6
Haplotype

Fig. 3. 23. Detection of wheat blast resistant (SRA isolated) associations by hapGWAS.

a) chromosome physical position of a QTL candidate to wheat blast resistance (788 - 794.2 Mbp) in Mattis reference. A cluster map within the region groups three main clusters (green, blue, and purple). The cluster in blue have the entire region similar to Mattis (low *variations*) including Stanley reference. Mattis and Stanley are known to be blast resistant to the SRA isolates. Most of the initial evaluated resistant genotypes are clustered in the green group, which are Watkins. These groups have two separated blocks similar to Mattis here defied as Region 1 (R1) and Region 2 (R2) (blue bars). The third group shares a similar region to Mattis only in R1. In this group there are 30 modern elite cultivars which were shown to be resistant to SRA after we discovered them to have the R1 region. In the zoom in block of R1 in **a**), there are three genotypes that have portion of R1 only. Two of them marked with asterisks in blue are susceptible to SRA. At the top of the cluster map a tetraploid wheat previously reported as *T. timophevii* in (Walkowiak et al., 2020) but we hypothesize to be *T. carthlicum* and hence is labelled as Tetraploid ssp. **b**) hapGWAS hit on chr2A that overlaps with the QTL region in **a**) in both R1 and R2. **c**) R1 hapGWAS haplotypes by genotypes. There were 27 haplotypes in the window and three of them have the positive effect: chr2A__sym_794250003 (H3), chr2A__sym_789000015 (H15), and chr2A__sym_789000019 (H19). 28 modern elites have H3 (purple) and are SRA resistant. H15 and H19 contain only Watkins genotypes which are predominately resistant. We include 348 genotypes with scores in the analysis, 324 Watkins and 24 elites. In (**b**), blue dots are positive correlated to have a high score (susceptible). Purple dots indicate that the phenotype has low score (resistant), in other words, it has the resistant haplotype. NOTE: the horizontal line was our initial threshold to consider the association to be true based on raw *k*-mers. Work is in progress to adjust the threshold to the number of haplotypes used for hapGWAS.

Beyond the chr2A hits in the Mattis reference, additional strong associations were identified in chr2B at 798 – 799 Mbp which has the *TraesSYM2B03G01095480*, a protein kinase ATP binding. We ran a pairwise alignment between the nucleotide sequence of the two genes (*TraesSYM2A03G00828360* and *TraesSYM2B03G01095480*) and found 82.48% nucleotide sequence similarity. The protein sequence similarity alignment of the two genes was 71.88 % (Supplemental Fig. S3.2), suggesting that these are not homeologs genes but probably a conserved gene with similar function.

Pangenome hapGWAS: We next wanted to test if it was possible to detect the chr2A associations using the additional pangenome references to get an insight of gene conservation. Our result demonstrated that out of eleven comparisons, the chr2A hit was detected above the *9 -log10(P)* threshold using ArinaLrFor, Mattis, CS, and Stanley as references only. These results suggest that the gene (s) might present a low conservation although a more comprehensive analysis with additional genome references or RNAseq data would validate this hypothesis. Or analysis also demonstrates that despite ArinaLrFor and Chinese Spring not having the favourable haplotype to SRA in chr2A, the hapGWAS pipeline was able to detect the signal by indirectly using the *variations* counts of the surrounding regions and the syntenic regions of the other genome references.

Additionally, we noticed that the reference ArinaLrFor, Julius, and Mace captured the hit on chr2B (Fig. 3.24). The haplotypes with the positive alleles in Julius for this hit were in two contiguous 1 Mbp windows. The first is chr2B__jul_783000030 located at 782 – 783 Mbp and the second is chr2B__jul_782000004 at 781 -782 Mbp. Looking into the gene content of the strongest candidate window, we noticed that again, this is a hotspot of NLRs genes; *TraesJUL2B03G01084260*, *TraesJUL2B03G01084320*, *TraesJUL2B03G01084350*, *TraesJUL2B03G01084380, TraesJUL2B03G01084390*. Interestingly, no hit was detected on chr2A of Julius. These results suggest that additional resistant gene is conserved in Julius, ArinaLrFor, and Mace, and that they may have the susceptible haplotype while some of the Watkins might have the resistant haplotype. This haplotype may not be present in the Mattis reference and some of the resistant Watkins may have one or the two functional alleles in the chr2A and chr2B homeologs genes. Further analysis on the SNPs within the candidate genes of Julius, ArinaLrFor, and Mace, by alignments would support this hypothesis (not addressed here).

Results with the Mace Reference also detected a very similar hit to the one in Julius. This appears to be the syntenic corresponding window of Julius because it has the same hotspot of several NLRs. The genes in the window with the strongest association are *TraesMAC2B03G01076480* TraesMAC2B03G01076690, *TraesMAC2B03G01076750*, *TraesMAC2B03G01076780*, *TraesMAC2B03G01076810, TraesMAC2B03G01076820, TraesMAC2B03G01076980.* The remaining references had a hit on chr2B but less strong than in chr2A and the gene content was different.

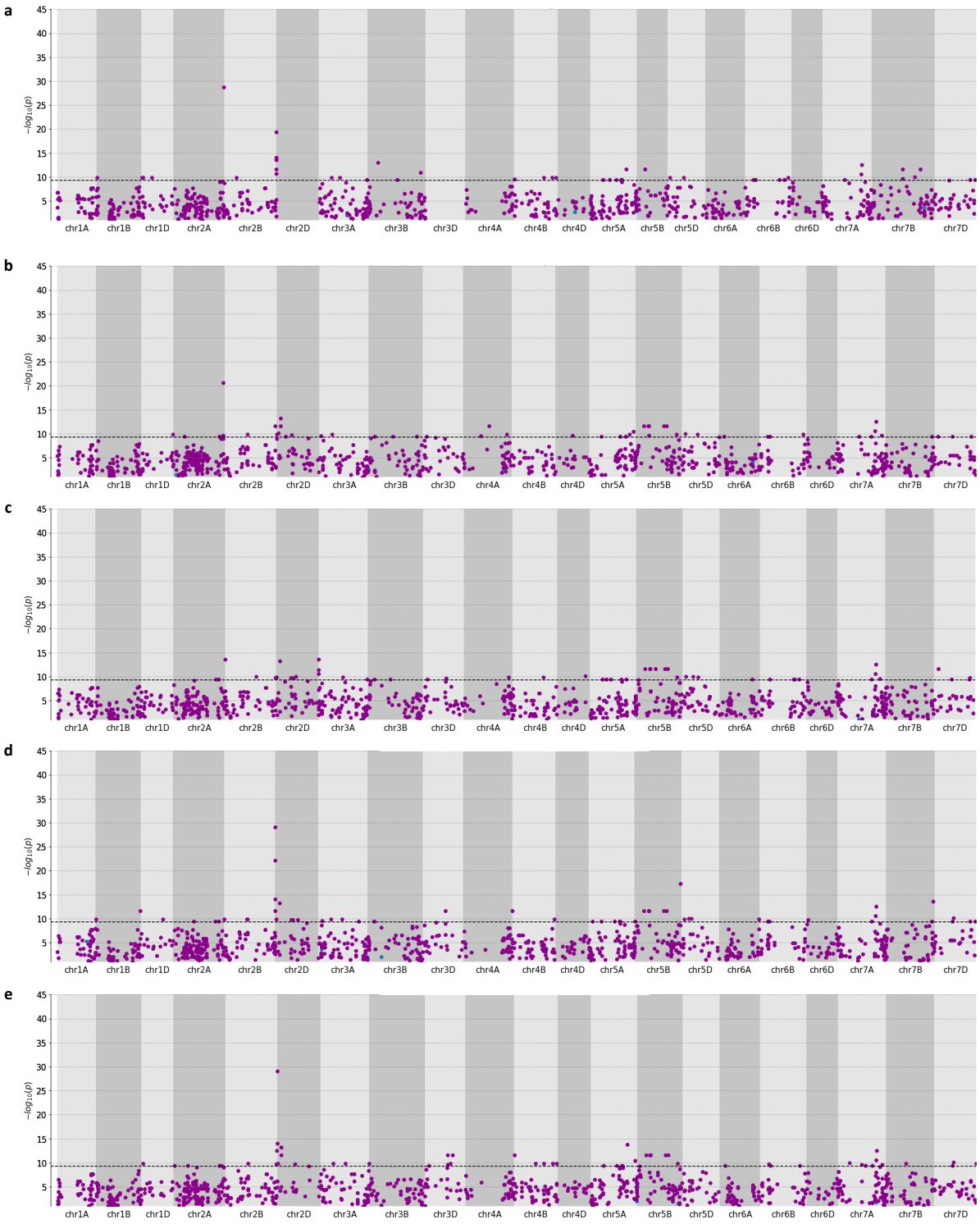


Fig. 3. 24. Pangenome GWAS detects unique hapGWAS associations for SRA blast resistance.

From top to bottom GWAS results from the SRA scores using the references **a**) ArinaLrFor, **b**) CS, **c**) Stanley, **d**) Julius, and **e**) Mace.

3.5. Discussion

3.5.1. Methods to define Haplotypes.

NGS technologies provide the tools to generate large genomic data for high density genotyping at relatively low cost. This allows users to combine multiple types of genetic variations to help to define haplotypes which can be more powerful to detect genotype-phenotype associations than traditional individual SNPs (Bhat et al., 2021). There are different methods to define haplotypes but until now, there is not a global protocol. Most recent methods employ NGS-based or array-based approaches (Rasheed et al., 2017). For example, using SNP array data, Balfourier et al., (Balfourier et al., 2019) built haplotypes employing the HaploView algorithm, a method based on LD and population statistics (Barrett et al., 2004). Similarly, using the same software, (Cseh et al., 2021) called haplotypes individually by chromosome using a confidence interval algorithm (Gabriel et al., 2002) which measures the LD among markers in a arbitrarily defined blocks.

Most of these methods still rely on SNP calling approaches and therefore identify variations in the same way. In the present study, we developed an alternative approach to build haplotypes based on presence/absence *k*-mers instead of alignments and SNP calling methods. To our knowledge, there are no reports on directly using *k*-mers to define haplotype blocks, but instead they employ raw *k*-mers to directly run GWAS or population analysis (Gaurav et al., 2022; Voichek & Weigel, 2020). Our approach uses a clustering algorithm based on "*variations fingerprint*" counts that automatically call haplotypes using split or sliding window size defined by the user.

A recent study in our group defined haplotypes of the wheat pangenome (15 assemblies). They have shown that large haplotypes in wheat are maintained intact regardless of the origin of the country and the growth habit of the genotype (Brinton et al., 2020). Most of these genotypes are extensively being used in several breeding programs globally which indicates that breeding programs are exchanging similar germplasm and haplotypes. Although the Brinton et al., 2020 results were informative, their approach still relied on whole genome assemblies and alignments, which at the time of writing this thesis are still expensive and time consuming for large germplasm collections and large genomes.

One of the aims of this thesis was to identify those haplotypes using raw reads at relatively low coverage (~12-fold). Our haplotypes defined in this chapter coincided with the results of (Brinton et al., 2020) on the large haplotype blocks being maintained in the pangenome cultivars. Furthermore, our analysis revealed that these blocks are prevalent in several other important cultivars from Northern Europe, CIMMYT, and in landraces collected from multiple countries. Similar observations were detected in (Brinton et al., 2020) using haplotype guided KASP markers

in a specific chromosome 6A QTL region comparing modern cultivar and the Watkin landraces collection.

Additionally, until recently, most haplotypes were built based on a single reference (Bhat et al., 2021). However, genome regions are missed if the genome reference does not represent the full genome diversity of a species. Based on this limitation, on the single reference problem, recently multiple genome references of different important crops have been developed. The use of multiple references to build haplotypes graphs are undergoing in humans and some important crops which have relatively small genomes (Bradbury et al., 2022; Rakocevic et al., 2019). However, the genome of wheat is polyploid and has three highly similar subgenomes, with a total size of ~16 Gbp. Therefore, using genome graphs is still challenging to incorporate WGS of a high number of genotypes for hexaploid wheat. An initial analysis to embrace this challenge was made using exome-capture of 65 genotypes in (Jordan et al., 2021) but it has the limitation to only call haplotypes based on the gene content regions.

In our analysis we leveraged the wheat pangenome to build haplotypes using multiple references and WGS data of >1,000 wheat accessions. Our approach is flexible to incorporate novel genotypes and evaluated them in the context of our current collection. As the sequencing cost continues to decrease, wheat accessions can be added which will allow to detect a comprehensive wheat genome as a species in a single haplotype database. This will facilitate comparisons among different studies and enable better reuse of publicly available data.

Although not presented in this thesis dues to time, during our fine tuning of parameters to determine haplotypes calls thresholds, we conducted a resampling pilot test to determine the effect of sample size using sub-groups genotypes. We also tested if including sequences of wild relatives had an impact on the number of haplotype calls. We divided this analysis into 22 sub-groups (Supplemental Table S3.1). As our default in most of the analysis we employed the *"WatSeq_Pangenome_RAGT_ABD"* group since we observed the most consistent results as it contains the greatest number of genotypes including only wheat accessions (1,123 accessions). To analyse the D sub genome, however, we included 265 additional *Ae. tauschii* accessions to improve the haplotype calls precision and recall as described in section 3.4.1.11 when adding diversity to the D genome. This was required since the hexaploid wheat D subgenome lacks genome diversity as it was hybridized between a reduced gene pool of *Ae. tauschii* genotypes with ancestral tetraploids (Marcussen et al., 2014). This low diversity in the D genome is persistent in landraces and modern cultivars as naturally hybridizations between hexaploids and *Ae. tauschii* rarely occur (Akhunov et al., 2010; Wang et al., 2013).

As described in section 3.4.1.9, Affinity Propagation automatically calls haplotypes based on the "*dmp*" parameter. The number of haplotypes per genome region relies on the number of genotypes included, as expected. In our IBSpy pipeline we automatically adjusted this parameter by testing the Silhouette Coefficient (SC) score for each haplotype per window selecting the best *dmp*. In this manner, our approach set an optimal "*dmp"* per genome region to call haplotypes. For example, in genome regions where there is a high level of diversity, high values of *dmp* would not impact drastically the results. However, in a low diversity genome region, low values of *dmp* would be preferred, as it will separate near-IBS regions better.

Depending on the purpose of the analysis, the practicality of adjusting these parameters can beneficial. For example, if the objective is to detect similar (but not identical) fragments of DNA such as old introgressions from wild ancestors, the *dmp* parameter can be set to include high values only. In this manner, similar genotypes will be grouped as having similar long-range haplotypes (near-IBS) and introgressions can be detected. We will show a case study in Chapter 4, section 4.4.2 on this topic.

In general, we observed that the number of haplotypes agreed with the number of samples included in each analysis. Haplotypes calls with the AP method is possible with low number of samples as long as it can generate well defined clusters allowing the algorithm to converge. In our case, the minimum of samples tested was 114 including only modern cultivars. However, using this low number of samples we noticed an increase in error calls, based on the parent-child tests (orphan category) and redundant test. Furthermore, in this study we included a maximum of 11 genome assemblies information using syntenic regions. We noticed that using fewer genome assemblies we were not able to discriminate different genotypes in specific genome regions as described in section 3.4.1.4. It might be possible that haplotype calls with <100 query samples will be possible as more genome references are added into the databases since more genome information will help to discriminate and create condensed and well separated clusters during the AP calls.

The precision-recall and F1 scores varied by genome region and chromosome depending on the *dmp* used per each 1 Mbp block during the AP clustering. As described in 3.4.1.9, IBSpy selects the best *dmp* based on the SC score. As the wheat genetic diversity varies among and within chromosomes due to natural variations and introgressions, each genome region had their "optimal" *dmp* value. In Fig. 3.16 we showed an example of three chromosomes, but we observed a similar trend in other chromosomes and other genome references tested. Overall, similar to the parent-child test discussed above, we had lower precision and recall in the genome D compared to A and B genomes due to the low diversity of the D genome. This problem was partially solved

by adding diversity with the 265 additional *Ae. tauschii* accessions as a query during the AP haplotype calls. Adding further *Ae. tauschii* genome assemblies as a reference by capturing syntenic windows as described in section 3.4.1.8 was not tested in the present study. However, we hypothesize that this may help to increase the scores in the precision and recall metric as more genome information with high variation is present in wild *Ae. tauschii* than in the current D hexaploid wheat allowing to discriminate better among wheat cultivars D genomes.

Going forward, if researchers are aiming to define haplotypes across QTL or GWAS regions and put them in the context of a wider haplotype database, it will allow others elsewhere to determine if they also have these haplotypes within their locally adapted germplasm. This should help better define which haplotypes would be worthwhile introducing into their germplasm pool, and which haplotypes are already present (Wimalanathan & Lawrence-Dill, 2021). Likewise, it would facilitate better understanding of G*E interactions since often studies in one location define a "positive effect" QTL, but this does not necessarily translate into a beneficial effect in other locations (Sukumaran et al., 2018). These haplotype-based G*E analysis would be of even greater value for large datasets within breeding programs which operate genomic selection models across multiple environments.

In addition to the genome information, huge advances on gene annotation pipelines and predictions based on gene networks have been developed (Hummel et al., 2023; Kotera et al., 2012; Theodoris et al., 2023). Our *variations* and haplotype calls here described can be used to investigate gene regulations based on haplotype blocks. For example, using IBSpy haploblocks users can investigate what is the gene content on the different haplotype blocks present in different cultivars. Those genes could also be interrogated for their gene expression changes to shed light if having similar haplotypes with different genome context affect the expression and therefore the gene effect. The impact on gene expression based on different haplotype lengths could also be interrogated. It could also be possible to test previous identified gene functions extending different haplotype blocks. This information could help to update and improve predictions on gene networks and genotype-phenotype associations.

3.5.2. Haplotype diversity in wheat

During the wheat evolution from diploid species to hexaploid, bread wheat has experienced at least two genome polyploidization events, the first from two diploid species; *T. urartu* and the extinct species related to *T. speltoides* and the second with the D-donor *Ae. tauschii*. Following these events of polyploidization during domestication, the genetic diversity of cultivated

accessions was reduced, and a percentage of this diversity was maintained in landraces. Later with the development of cultivars and the release of varieties by public and private institutions a further reduction of this diversity impacted in modern varieties through intense selection in breeding programs (Cseh et al., 2021; Haudry et al., 2007; Vikram et al., 2016).

Despite this loss of diversity, rich natural genetic diversity is still available in germplasm banks from collections and are valuable for wheat researchers (Schulthess et al., 2022). This genetic diversity can be indirectly studied measuring phenotypes in a population but phenotypes are influenced by environmental factors, therefore, prone to human errors and bias (Plekhanova et al., 2017). To partially alleviate this problem, researchers commonly use molecular markers in place to better study genetic diversity and to re-classify organisms and species (Kesawat & Das Kumar, 2009). Relatively cheap array-based genotyping and WGS are the predominant platforms used in these studies and depending on the objective, these molecular markers can be used individually or in haplotypes by combining two or more polymorphisms.

Haplotypes are DNA recombination blocks inherited together in subsequent generations instead of individual nucleotides (Bhat et al., 2021). In 2015 Jordan *et al.,* (Jordan et al., 2015), studied the haplotype diversity in a collection of 62 accessions using exome capture genotyping technology. They demonstrated that the B genome has on average a higher number of haplotypes than the A and D genomes, consistently with the hypothesis that the A and B genomes being more genetically diverse than the D genome (Akhunov et al., 2010). These number of haplotypes were concentrated mostly at telomere regions with a reduced number in centromeres.

In 2020 Brinton et al., demonstrated that current genotyping platforms (35K SNP array) do not properly allow to discriminate between haplotypes in modern UK accessions or landraces, and can be bias towards certain chromosome regions (Brinton et al., 2020). A study comparing landraces (n=199) vs modern cultivars (n=67), Cseh *et al.,* (Cseh et al., 2021) found that overall, landraces harbours more haplotypes and genetic diversity than modern cultivars. In this study, a collection of European and landrace accessions genotyped with a 20K SNP array containing 17,267 SNPs was employed. In their study they found 94.48 haplotype blocks per chromosome, ranging from 5 (on chr4D) to 178 (on chr5B) and the haplotype diversity (H_d) was 0.46 based on the method defined in (Nei & Tajima, 1981). The number of SNP per haplotype block was 297.52 ranging from 24 to 585 and the total number of haplotypes in the genome was 1,984.

In a broader analysis, (Balfourier et al., 2019) studied the haplotype diversity of a bread wheat collection of 4,506 accessions from the Institut National de la Recherche Agronomique (INRA) including worldwide (105 different countries) landraces (n=632), traditional cultivars (n= 965), and modern cultivars. Using a SNP array containing 280,226 SNPs, they found that 85% of the

haplotype blocks were shorter than 1 Mb having on average 4 haplotypes (alleles) per block and ranging from 2 to 20.

In 2014 Wingen *et al.,* (Wingen et al., 2014) measured the genetic diversity of 826 Watkins landraces using 41 microsatellite markers. They found a high level of genotypic diversity compared to modern wheat varieties released from 1945 to 2000 in Europe. This study revealed nine ancestral groups in the entire collection, information that was used to build a core collection of 116 accessions representing most of the genetic diversity from the initial dataset.

Later with the advent of NGS a high-density molecular marker in SNP-arrays (Winfield et al., 2016) was designed to genotype and characterize elite genotypes, landraces, and multiple species from the second and tertiary gene pool of wheat. Using a pairwise similarity matrix, different groups were detected by grouping among *Ae. tauschii*, *T. aestivum*, *T. turgidum* and wild relatives. Using this array, it was possible to differentiate among winter and spring wheats and landraces. Later, a more practical version of this SNP-array was developed focusing only in the most informative markers to reduce costs and genotype large number samples and it was named the "Wheat Breeders' Array" (Allen et al., 2017).

In our study using the WatSeq dataset which includes 827 Watkins landraces and 218 modern cultivars from the North of Europe, we defined haplotypes in 1 Mbp window. This window size is not static and can be easily modified by users adjusting optional parameters in IBSpy. The genomic information in our study was WGS and therefore incorporate the full genomic information compared to Exome or SNP based arrays genotyping. On average we found ~20 haplotypes per 1 Mbp window. In agreement with (Balfourier et al., 2019) and other studies, we found a higher number of haplotypes in telomere regions than in centromeres. Our method can also be more stringent or relaxed when using similarity scores among genotypes based on "*variations count profiles*" by *k*-mers to explore near-IBS regions. Fine tuning this parameter and modifying the number or type of genotypes while calling haplotypes, as expected, the number of haplotypes per 1 Mbp window varies accordingly.

3.5.3. Haplotype-phenotype associations

Regardless of the method to define haplotypes, they can be used to study genetic diversity, investigate the impact of breeding, or phenotype-genotype associations using GWAS or genomic selection (GS). In recent years, haplotype-based studies instead of single markers associations studies, became prevalent for different crops and traits (Bhat et al., 2021; Contreras-Soto et al., 2017; S. He et al., 2019; Jensen et al., 2020; N'Diaye et al., 2018). However, there is still a debate

if using haplotypes versus single markers is more efficient to detect marker-trait associations or if they can explain the phenotypic variation better.

For example, the use of haplotypes was reported to increase the prediction accuracy on GS (Matias et al., 2017; Won et al., 2020) and improved the detection of genomic loci in GWAS studies (Hamazaki & Iwata, 2020). An explanation for this is attributed to SNPs in array-based haplotypes to be filtered for minor allele frequency (MAF) in early steps of the analysis. Also, haplotypes are multi-allelic and can capture the "true" combination of variations which gives better associations to QTLs than individual SNPs. This is in agreement with single SNPs in GWAS analysis as they often do not represent the causal molecular variant of the phenotype (Korte & Farlow, 2013) and because haplotypes could benefit from local epistatic among QTLs within the haplotype.

In our analysis we used the AP haplotypes defined using IBSpy *variations* to run hapGWAS and found strong associations for spike related traits and disease resistant. We showed that our haplGWAS for spikelet number phenotype coincided with the *WAPO-A1 (TraesCS7A02G481600)* gene position on chr7AL (Kuzay et al., 2022; Zhang et al., 2018), the orthologous gene *ABERRANT PANICLE ORGANIZATION1 (APO1)* detected in *Oryza sativa* (Ikeda et al., 2007). In our analysis using 1 Mbp windows haplotypes, we found two haplotypes strongly associated with the trait at 674 – 675 Mbp in CS (RefSeq v1.0) which coincides with the chromosome physical position of *WAPO-A1.* The association was detected consistently in all the 11 chromosome-scale assemblies which suggest a strong conservation of the locus.

Interestingly, (Kuzay et al., 2022) found three main haplotypes in wheat (H1, H2, and H3) to be associated with spikelet number. H1 has a 115-bp deletion in the promoter affecting the expression of the gene in developing spikelets and this polymorphism is associated with reduced spikelet number. Conversely, H2, predominantly in modern hexaploid wheat, has a positive effect on spikelet number and has a stronger effect than H3. In our analysis when using hapGWAS, we detected two haplotypes, chr7A__chi_675000010 and chr7A__chi_675000006, negatively and positively associated with the trait respectively. We hypothesize that chr7A__chi_675000010 is linked to H1 having the 115-bp deletion detected by (Kuzay et al., 2022) and chr7A__chi_675000006 is related to H2 in the same study. We did not identify any hit for the *WAPO1* loci in the homeologs genes of wheat on subgenomes A or B. However, we found a hit for Max floret number in CS (RefSeq v1.0) located at 648 - 649 Mbp on chr7B. This position overlaps with the *WAPO-B1* locus. Analysis of *WAPO1* mutants in (Kuzay et al., 2022) found a wide range of floral abnormalities which suggest that *WAPO1* may be involved in other floral related traits. Therefore, we hypothesize that the Max floret number detected in *WAPO-B1* locus in our analysis having a positive allele affect to the trait, could be related.

For the disease resistant GWAS hits, we found that most of the hits contained multiple disease-related genes such as NLRs and Kinase related genes. It is well documented that NLRs are more dynamic and evolve more rapidly that other functional genes (Marchal et al., 2020; Marchal et al., 2018). In this analysis we did not expand to validate any of the regions with hits, but it will be a starting point and continuation for a follow up project for validation.

The current version of hapGWAS in this study validated the usefulness of the haplotypes called. As a starting point we made use of the already available software and adapted to our haplotype calls formats when needed. However, to correct for population structure in our hapGWAS study we employed a PCA matrix generated from a SNP information which were called based on CS (RefSeq.v1.0). It was highlighted in (Bhat et al., 2021) and (Meuwissen et al., 2014) that genomic predictions may benefit from a relationship matrix by haplotypes instead of single SNPs. This may also improve the results in our hapGWAS analysis.

Our method here described to define haplotype relies on multiple genome references. Therefore, hapGWAS is a reference-based approach. Our method differs with common SNPs methods because it uses genome information from multiple genome assemblies at once and is based on *k*-mers information. This can be a disadvantage for crops or orphan crops where there is no genome reference available. However, as discussed in section 1.2.1, sequencing technologies has had a huge progress, and it is predicted that genome assemblies for orphan crops and pangenomes for other important crops will be released in the following years since genome assemblies are becoming the new routine approach in genomics. The challenge will be if users using the routine SNP based calling genotyping approach wants to perform GWAS analysis using individually each of the genome references, particularly for large genomes such as hexaploid wheat (~16 Gb).

On the contrary, our method here developed benefit of having multiple genome references to call haplotypes by multiple comparisons from genome information extracted from each of the references. Therefore, the advantage of hapGWAS analysis is to integrate genome regions private to one or some references. For instance, the large deletions or chromosome introgressions from wild relatives assembled uniquely in some genome references as described in the wheat pangenome project (Walkowiak et al., 2020). Although, our method includes information from multiple references, a single reference is used as a "genome *template*" to give chromosome physical position to the haplotype calls and to find the syntenic regions from other genomes. Depending on the "*template genome*" used, our method can detect differences on associations using hapGWAS as exemplified in section 3.3.6. Furthermore, running hapGWAS using each of the references as a "*template genome*" is straightforward and multiple comparisons can be easily run as exemplified in section 3.4.2.4, **Fig. 3.24** using hexaploid wheat.

4. Wheat alien introgressions

In this chapter we collaborated with Simon Krattinger and Hanin Ahmed from the King Abdullah University of Science and Technology (KASUT). Simon Krattinger and Hanin Ahmed generated the sequencing data for the 218 accessions of *T. monococcum.* Hanin Ahmed assembled the chromosome scale references of the two *T. monococcum* accessions; TA299 and TA10622.

Part of this analysis was published in (Ahmed et al., 2023) ***"Einkorn genomics sheds light on history of the oldest domesticated wheat"***.

4.1. Chapter summary

In this chapter we describe a pipeline to detect and characterize hybridisations/introgressions into the ten pangenome assemblies using IBSpy. Our approach includes an initial step to set a threshold using alignments of chromosome level assemblies from hexaploid wheats and wild relatives. Sequence similarity cut-offs of >99.99% (one SNP per 10 Kbp) between two wheat genotypes corresponds to identical by state (IBS) regions, whereas similarity of 99.95% (five SNPs per 10 Kbp) corresponds to a typical level of variation observed among wheat cultivars or landraces. In our IBSpy-based analysis, we set a threshold to call hybridisations/introgressions regions as having >120 *variations* in a 50 Kbp window. This threshold is equivalent to < 99.90% sequence similarity between two accessions (ten SNPs per 10 Kbp). Using this criterion, we characterized a collection of *T. monococcum* accessions, a publicly available set of *Ae. tauschii* accessions and known large alien introgressions from wild wheat relatives. Our results detected known and novel hybridisation/introgression regions in elite wheat cultivars, with evidence that breeders have been selecting fragments from the initial introgression. Furthermore, we detected the presence of wild wheat relatives genome regions in landraces not reported before. We speculate that some of these hybridisations have been guided with purpose (i.e., introgressions in modern wheats), but many may have occurred from natural hybridisations in landraces. We propose candidate genotypes to be the closest donors of these introgressions and natural hybridisations. For the known *Ae. ventricosa* 2AS/2NvS translocation, we detected the actual donor in four of the pangenome cultivars. Importantly, using the IBSpy haplotype calls, we identified novel hybridisations from *Ae. tauschii* in the D sub genome of wheat which are not present in the ten pangenome references.

4.2. Introduction

Hexaploid wheat (*Triticum aestivum*, 2n = 6x = 42 chromosomes) consists of three subgenomes (A, B, and D) and it is highly repetitive having >80% TEs and the gene coding regions representing only <2% of its genome (Appels et al., 2018). The A and B genome diverged from a common ancestor ~7 million years ago while the D genome originated from a hybridisation, and subsequent speciation event, between the A and B genome donors ~5 million years ago (Marcussen et al., 2014). Modern hexaploid wheat originated by different hybridisations events, first between *T. urartu* (AA) and a close relative of *Ae. speltoides* (BB) giving rise to wild emmer wheat (*T. turgidum* ssp. *dicoccoides*; AABB) approx. 400 thousand years ago (Huang et al., 2002). A later hybridisation arose between this species (or its domesticated form *T. turgidum* ssp. *dicoccum*) and the D genome donor *Ae. tauschii* approximately 10,000 years ago that coincided with the rise of modern agriculture (Marcussen et al., 2014). This is supported by multiple archaeological records and the absence of hexaploid wheats in wild populations (Salamini et al., 2002). Although the allopolyploid genome of hexaploid wheat contains three homoeologous subgenomes, they do not recombine among them due to the presence of the *PAIRING HOMOEOLOGOUS 1* (*Ph1*) gene which prevents chromosome homoeologs pairing (Martinez-Perez et al., 2001). It therefore behaves as a diploid organism during meiosis which allows for genome stability in the polyploid state.

4.2.1. The contribution of *T. monococcum* to the modern A wheat genome

Archaeological evidence indicates that *T. monoccocum* (2n = 2x = 14x), or einkorn wheat, was also domesticated ~10,000 years ago in the Fertile Crescent (Heun et al., 1997; Lev-Yadun et al., 2000) and it is closely related to *T. urartu*, the A genome donor of *T. durum* (Ling et al., 2018) and hexaploid wheat. Individuals of *T. monococcum* exist both in wild and domesticated forms and the wild forms are mainly classified in three morphologically and genetically distinct races termed alpha (α), beta (β), and gamma (γ) races. From these races, β was the origin of domesticated einkorn (Kilian et al., 2007) and is proposed to have contributed to the genome of modern wheat via gene flow by natural hybridisations.

T. monococcum provides an important reservoir of genetic diversity for hexaploid wheat, especially since genes discovered in this species often function in the polyploid wheat context. Natural hybridisations from *T. monococcum* into polyploid wheat have played an important role for agronomically important traits. Despite the evidence on the usefulness of *T. monococcum* to accelerate wheat research, until recently there was no high-quality genome assembly of this specie. As part of this thesis, we collaborated with international partners to use IBSpy to detect *T.*

monococcum genome regions present in modern wheat cultivars. This analysis was based on the sequencing of two chromosome-scale assemblies from *T. monococcum*, one from a wild accession and one from domesticated einkorn, and the whole genome sequencing (WGS) of 218 diverse accessions.

4.2.2. *Aegilops tauschii*: the wheat D genome donor

It its well-recognised that the diploid *Ae. tauschii* is the main donor of hexaploid wheat D subgenome. Natural populations of *Ae. tauschii* have been characterized into two major lineages: members of *Ae. tauschii spp. tauschii* into Lineage 1 (L1) and members of *Ae. tauschii ssp. strangulata* into the Lineage 2 (L2). More recently Gaurav *et al.,* in 2022 identified a third Lineage L3 (Gaurav et al., 2022) from the same *Ae. tauschii ssp. strangulata* classification. A study by (Zhou et al., 2021) categorized a different *Ae. tauschii* collection into five subgroups and created chromosome-scale genome assemblies of their genome references named AY17, XJ02, T093, and AY61 to represent each of their lineages described: L1W, L1EX, L1EY, and L2E, respectively. A fifth subgroup categorized as L2W was represented by the AL878 genotype which already has a full genome assembly (Luo et al., 2017). Using these references and the re-sequencing of a panel of 278 accessions, they assigned each of the accessions of the panel to the groups mentioned above.

Population studies suggest that more than one event of natural hybridisation between a tetraploid durum wheat and individuals of *Ae. tauschii* lineages gave rise to modern hexaploid wheat. The main donors are thought to be members with origin in the Southern Caspian region and distributed from Transcaucasia (Armenia and Azerbaijan) to eastern Caspian Iran. On the other hand, L1 members have been reported to contribute only a small portion of its genome from 0.8% to 2.7% (N. Singh et al., 2019; Wang et al., 2013).

After the initial hybridisations of 4x and 2x that gave rise to the hexaploid wheat (6x), there was a reproductive barrier between 6x and 2x *Ae. tauschii* which limited the extent to which the wild D genome could contribute into the 6x D genome pool. Hence there is a reduced diversity in the D genome of modern-day wheat when compared to the A and B genome diversity which has been recovered through natural hybridisations between tetraploid and hexaploid wheat. To compensate for the lack of diversity, wheat geneticists frequently employ *Ae. tauschii* to bring genetic diversity into the D genome of hexaploid wheat. This is achieved by crossing a tetraploid wheat with *Ae. tauschii* followed by a chromosome doubling step (Li et al., 2018) or by a direct cross between a hexaploid wheat and *Ae. tauschii* (Gill & Raupp, 1987) resulting in a synthetic hexaploid wheat. Currently and historically breeding programs releasing new synthetics wheats

and bringing these genetic materials into modern cultivars have been successfully accomplished (Dreisigacker et al., 2008). For these materials, however, there is still uncertainty regarding the genome location and the size of the *Ae. tauschii* fragments integrated into wheat (when crossing hexaploid wheat directly with *Ae. tauschii*) and how to best select them in a precise manner to avoid genetic drag of undesirable traits.

NGS and genome sequencing resources provide a route to address this challenge of defining introgression blocks boundaries into wheat more precisely. For example, in 2013 a 10K Illumina Infinium SNP array was created based on *Ae. tauschii* accession AS75 from L1 collected in central China (Wang et al., 2013). In 2017 the first chromosome-scale assembly of *Ae. tauschii* accession AL8/78 from Armenia (L2) was published by (Luo et al., 2017) and in 2021 an improved assembly and annotation was achieved for the same genotype by (L. Wang et al., 2021). Using these genomic resources, breeders, and geneticists can design molecular markers to tag genome regions and follow in introgressions into wheat.

Despite these efforts, there is still a debate if these accessions with genome information are truly representative of the donor gene pool of the D subgenome of wheat. In this analysis we employed IBSpy to explore the genetic diversity of a panel of 265 accessions of *Ae. tauschii* and the contribution of members into modern wheat D genome pool. We suggest the closest gene pool that may have given origin to the D genome of wheat and validated that the previous genome assembly from AL8/78 is not the closest donor. Therefore, generating novel genome information of a wide representatives of *Ae. tauschii* would be of value. Furthermore, using IBSpy haplotype calls in combination with the WatSeq genotypes, we uncover two novel additional hybridisations events that may have contributed to the D genome gene flow, one from L3 and one from L2 members.

4.2.3. Tracking introgressions in heaxploid wheat

Wheat lost genetic diversity, first during domestication from wild relatives into landraces and then from landraces into modern elite varieties (Reif et al., 2005). On the contrary, wheat wild relatives maintain a high level of genetic diversity unexplored for agronomically important traits in wild populations (Leigh et al., 2022). During early breeding, wheat geneticists realized this potential of wild populations and started to make use of them incorporating traits into their breeding programs (Doussinault et al., 1983). From the early 1960s, crosses between wheat and wild relatives were successfully achieved and cultivars having those alien introgressions were dispersed worldwide in different breeding programs, both public and commercially (Gao et al., 2021). Initially, these large

induced introgressions were tracked using cytological techniques such as FISH (Badaeva et al., 2008) and more recently by molecular markers (Allen et al., 2017; Grewal et al., 2020). Different studies have detected and reported these large introgressions to be present in modern wheat cultivars populations either intact or fragmented selected by breeders to reduce their size and genetic drag (Keilwagen et al., 2022).

Large introgressions usually encompass several functional genes with different benefits and are particularly common for disease resistant traits or grain quality traits (Helguera et al., 2003). Evidence of this is that some large introgressions are actively selected and maintained in breeding programs (Gao et al., 2021). In addition to selection, large introgressions are prevalent within the wheat genome due to the lack or limited recombination between wheat and wild relatives. *Ph1* mutants, however, can be used to circumvent this issue (Rey et al., 2017). Although these alien introgressions have proven to be of pivotal value in modern varieties, still some breeding programs are reluctant to incorporate them due to the lack of recombination with the wheat genome and the time-consuming task for eliminating undesirable traits. There is an expectation that these challenges will be partially overcome with recent sequencing and novel molecular biology technologies (Hao et al., 2020). As a result, efforts to develop a wide repertoire of induced introgressions on wheat by different breeding programs are in progress (Devi et al., 2019; Gaurav et al., 2022; Zhou et al., 2021). Furthermore, methods to detect them in a precise manner using novel sequencing technologies are also becoming more prevalent (Grewal et al., 2020).

In this study, we demonstrated the potential of WGS combined with IBSpy to detect large introgressions at 50 Kbp resolution and provide some case studies where they are associated with beneficial haplotypes and traits. We hypothesize that, despite the large number of samples explored in this analysis (>1,000 wheat genotypes), as WGS becomes more affordable for other wheat germplasms banks, novel unexploited large introgressions already present within the wheat genome will be revealed. This will be easily accomplished by putting those novel sequences into the context of the current IBSpy *variations* database being built in this project.

The aim of this chapter is to validate IBSpy to detect large introgressions and elucidate historical hybridisations from wheat wild relatives into landraces and modern wheat. We aimed to translate the level of *variations* in IBSpy to the sequence similarity from alignment methods to set a cut-off between the immediate gene pool among wheat cultivars and landraces and compared with the level of *variations* in wheat wild relatives. Using this criterion, we investigated the contribution of *T. monococcum*, one of the closest A subgenomes donors, into hexaploid wheat. We used the same criteria to investigate the controversial historical number of hybridisations events that gave rise to the D subgenome from *Ae. tauschii.* Finally, we validated IBSpy to detect large

introgressions and large deletions and compare the level of variations of deletions vs a wild relative introgression region.

4.3. Methods

4.3.1. Sequence data

We used the collection of 265 publicly available *Ae. tauschii* accessions described in (Gaurav et al., 2022) to compare against the D subgenome of wheat (Chapter 2, Table 3). For the analysis of the A subgenome, we included 218 accessions of *T. monococcum* generated by Ahmed et al., 2023 (Chapter 2, Table S2.3). For large introgressions, we used the public available reads of *T. timopheevii* and *Ae. ventricosa* described in (Walkowiak et al., 2020). For the *Ae. ventricosa* genotype ventricosaCGB116981 we use raw reads from (Aury et al., 2022) (Table 5). The wheat samples were described in Chapter 2 WatSeq dataset.

4.3.2. IBSpy and methods to detect introgressions

We used two methods; the first is a *k*-mer mapping based approach, and the second method is based in IBSpy *variations*. The detailed steps are described in https://github.com/Uauy-Lab/monococcum_introgressions.

The *k*-mer mapping approach was developed by Hanin Ahmed. From the Methods of Ahmed et al., 2023. "For generating *k*-mer datasets, we used the whole-genome sequencing data from all domesticated einkorn accessions in the panel and *T. urartu* accessions (Zhou et al., 2020). *k*-mers (k=51) were counted from the Illumina raw data per accession using jellyfish (v2.2.10) (Marçais & Kingsford, 2011). We extracted the *k*-mer nucleotide sequences from each accession of *T. monococcum* and *T. urartu*. We concatenated all *k*-mers sequences from all *T. monococcum* accessions and *T. urartu* accessions into two separate files per species and kept one representative per *k*-mer. We removed common *k*-mers between *T. monococcum* and *T. urartu* and obtained a list of specific einkorn and *T. urartu k*-mers sequences, respectively. The lists of specific *k*-mers were later converted into fasta files. Each fasta file (*T. monococcum* and T. *urartu*) was mapped against the bread wheat genomes (Walkowiak et al., 2020) using BWA mem (v0.7.17) (Li & Durbin, 2010) requiring mapping of only full length of *k*-mers with no mismatches. Mapped *k*-mers in each bam file (*T. monococcum* and *T. urartu*) were analysed for the coverage in genomic windows of 1 Mbp using mosdepth (Pedersen & Quinlan, 2018) and visualized in R (v4.0.4) with ggplot2. Putative introgressions were identified as an increased coverage of mapped *k*-mers from *T. monococcum*

(with average coverage ≥ 5), but depleted mapping of *T. urartu* specific *k*-mers. Two or more regions were grouped into one if they were no more than 1 Mbp apart."

For IBspy we first created the *k*-mer databases (*k* = 31), we used jellyfish v.2.2.6 (Marçais & Kingsford, 2011) or KMC v3.0.1 (Kokot et al., 2017). We used IBSpy as described previously in the thesis and in https://github.com/Uauy-Lab/IBSpy. To estimate the IBSpy *variations* cut-off to define *T. monococcum* introgressions into the hexaploid pan-genome cultivars, we compared the output of sequence alignments between fully assembled references (Brinton et al., 2020) to the IBSpy *variations* data. We used *variations* ≤ 30 as a cut-off to detect putative introgressions. For each introgression block, we determined the number of *T. monococcum* accessions belonging to each of the six STRUCTURE groups described in **Fig. S4.1** from Hamed et al., 2023 (publication accepted), (**Fig. 4.1b**, **Table 4.1**). An accession was assigned as having an introgression block if it had at least 20 % of the 50 Kbp windows within the block with *variations* values ≤ 30. For example, if an introgression block had 60 windows, an accession would be classified as having the introgression if 12 or more 50 Kbp windows (60 × 20 % = 12 windows) had *variations* values of 30 or less.

We also performed a similar analysis as in (Brinton et al., 2020) by generating pairwise MUMmer (v4.0.0.2) (parameters: --mum --delta and delta-filter -l 20000 for filtering, i.e., retain only alignments ≥ 20 Kbp in length) alignments per chromosome between the assemblies from the ten pan-genome cultivars and the two *T. monococcum* assemblies generated by our collaborators (accessions TA299 and TA10622). For the large introgression detection of *T. timopheevii* and *Ae. ventricosa*, we used IBSpy *variations* count only since they do not have genome assemblies available at the time of writing this thesis.

4.4. Results

4.4.1. The contribution of *T. monococcum* to the wheat gene pool

We used IBSpy to identify *T. monococcum* introgressions in the ten hexaploid wheat pangenome cultivars (note that we describe these as introgressions although they can be considered as hybridisations more likely). We built *k*-mer databases from multiple genotypes, including the Illumina raw data of 218 *T. monococcum* accessions, two *T. monococcum* chromosome-scale assemblies, and ten genome assemblies of wheat (Walkowiak et al., 2020). Using 50 Kbp windows, we compared the *k*-mers in the reference sequence to the *k*-mers of each query genotype database and counted the number of *variations* within each window. As described previously in **Chapter 2**, a *variation* is defined as a set of continuous *k*-mers (k=31) from the reference

completely absent in the query. Low *variations* count indicates high similarity between the 50 Kbp reference assembly and the query sequence (hence can be used to identify introgressions), whereas high *variation* counts indicate lower sequence similarity.

To estimate the *variations* cut-off to detect *T. monococcum* introgressions into the hexaploid pangenome cultivars, we compared the output of sequence alignments between fully assembled references to the IBSpy *variations* data. We compared the published (Brinton et al., 2020) pairwise MUMmer alignments among the ten pan-genome cultivars (ArinaLrFor, CS, Jagger, Julius, Lancer, Landmark, Mace, Norin61, Stanley, Mattis) with the corresponding *variations* counts from IBSpy outputs to compare the sequence identity with the *variations* count. In total, there were 90 pairwise alignments analysed, and we focused on the seven A subgenome chromosomes. We analysed the data in 500 Kbp windows (a total of 890,793 windows) and kept those windows with at least 60% breadth of alignment in the MUMmer output (77.8%; 693,102 500 Kbp windows). We grouped IBSpy data into a window size of 500 Kbp to have more information to compare with the alignments, but we will describe the cut-off values as *variations* per 50 Kbp to be consistent across the thesis.

For each 500 Kbp window, we had the average sequence identity between the pangenome reference and the other nine pangenome query samples (if over 60% breadth of alignment), alongside the IBSpy *variations* for the equivalent comparisons using the pangenome reference assembly and the *k*-mer database. We grouped the data base on the number of *variations* (in increments of 10 *variations* per bin) and determined the distribution of the sequence identity in each bin (Chapter 2 Fig. 2.20). We identified that most of the 500 Kbp windows (532,248; 76.8%) had 30 or less *variations* per 50 Kbp. On average, these windows with 30 or less *variations* determined by the IBSpy method had sequence identity of >99.95% when their full genome assemblies were compared. The data distribution was such that a 500 Kbp window with ≤ 30 IBSpy *variations* per 50 Kbp has a 0.926 probability that the alignment of this window will have at least 99.9% sequence identity; and a 0.997 probability that the sequence identity will be at least 99.8%.

We performed a similar analysis by generating pairwise MUMmer (v4.0.0.2) alignments between the assemblies from the ten pangenome cultivars and the two *T. monococcum* assemblies generated as part of this collaboration (TA299 and TA10622). As expected, across the two pairwise alignments, a few 500 Kbp windows had alignments which covered at least 60% of the window (n= 238 windows; 0.3% of a total of 197,954 windows). In parallel, we used the raw reads of these two *T. monococcum* accessions to run IBSpy and compare the results with the MUMmer alignments. Using the ≤ 30 cut-off, we identified 201 windows that had on average 99.91% sequence identity; and a 0.970 probability that the sequence identity between pairwise alignments will be at least

99.8%. As such, we consider pairwise comparisons with IBSpy values of ≤ 30 variations per 50 Kbp as being identical or near identical by state, both in hexaploid wheat and between hexaploid and *T. monococcum* comparisons.

Next, we used the *variations* count ≤ 30 criteria to identify continuous windows that belong to an introgression block across the A genome of the ten wheat genome assemblies. For each 50 Kbp window, we determined the minimum number of *variations* in the raw data of the 218 accessions and the two assemblies of *T. monococcum*. Based on this "Einkorn_min" value, we identified 50 Kbp windows in which this value was equal to or lower than the 30 *variations* cut-off. These windows were considered as *T. monococcum* introgressions into the corresponding reference sequence. We next called introgressions blocks (Table. 4.1, Supplemental Table. S4.1) by stitching together 50 Kbp windows with *variations* ≤ 30 that were separated by less than ten non-introgression windows (i.e., with *variations* > 30). This was done as these "non-introgression" windows often had values just above the 30 *variations* cut-off. Two or more regions were grouped into one if they were <500 Kbp apart.

Among the ten wheat cultivars, ArinaLrFor showed the highest quantity of einkorn introgression with cumulative size of ~77 and 95 Mb based on the IBSpy *k*-mer variations and *k*-mer mapping, respectively (Fig. 4.1a; Table 2). The number of *T. monococcum* introgressions (identified by IBSpy) into the ten wheat cultivars ranged from 13 (Stanley) to 25 in ArinaLrFor. Their size also varied between 32.8 Mbp (Landmark) to 76.6 Mbp (ArinaLrFor). In total we found 1,714 genes across the introgressions blocks of ArinaLrFor genome. The longest block was block_20 located in chr6A at the end of the chromosome at 608.8 Mbp with 11.34 Mbp length and this block contains 714 genes with different gene functions. The introgression block with the most genes per sequence length was block_14 in chr6A at 1.5 Mbp in the short arm and had length of 0.65 Mbp. This block contained 33 genes with different gene functions including NLRs genes which are known to be involved in defence against pathogens. Per chromosome, chr6A had the most gene content per introgression blocks with 714 in chr 6A and had the most gene-dense block. In total across the introgression blocks we found 165 transposable elements genes which correspond to 9.63% of the gene content in the block (Supplemental S4.1.1, S4.1.2).

The fact that we detected *T. monococcum* introgressions in all ten pangenome wheat cultivars suggests that this is a widespread phenomenon. Considering a genome size of 15.4 Gbp (Appels et al., 2018) this means that between 0.21 and 0.50% of the wheat genome comes from *T. monococcum*.

Table 4. 1. *T. monococcum* introgressions identified in the ten wheat genomes using IBSpy.

	T. monococcum	
Cultivar	N	Cumulative size (bp)
Arina*LrFor*	25	76,615,321
Chinese Spring	14	38,698,556
Jagger	14	37,300,000
Julius	21	53,233,765
Lancer	18	49,751,962
Landmark	14	32,800,000
Mace	19	52,087,782
Norin61	15	36,350,000
Stanley	13	56,309,189
SY_Mattis	20	42,756,208

Overall, both IBSpy *k*-mer variations and *k*-mer mapping methods detected the same introgression blocks, with few exceptions being identified by only one method. The different results between the two methods are likely due to the genomic window size. With IBSpy, we looked at 50 Kbp resolution windows, whereas with the *k*-mer mapping approach we used 1 Mbp windows. Furthermore, for IBSpy we used all the accessions in the panel (n=218) whereas only the domesticated accessions (n=61) were used for *k*-mer mapping due to differences in the approaches. With IBSpy we identify specific regions from different groups and individual accessions and assigned the genetic gene pool into the ten genome references with the "dom_g1" group as the main donors (**Fig. 4.1bc**). IBSpy analysis across the ten wheat cultivars revealed regions on chromosome 5A with evidence of hybridisations with wild einkorn race γ being the putative donor (**Supplemental Table 4.1**). Because some domesticated einkorn accessions contain portions of γ race on chromosome 5A (**Fig. S4.1**), the *γ* introgressions in bread wheat could thus have been introduced through domesticated einkorn.

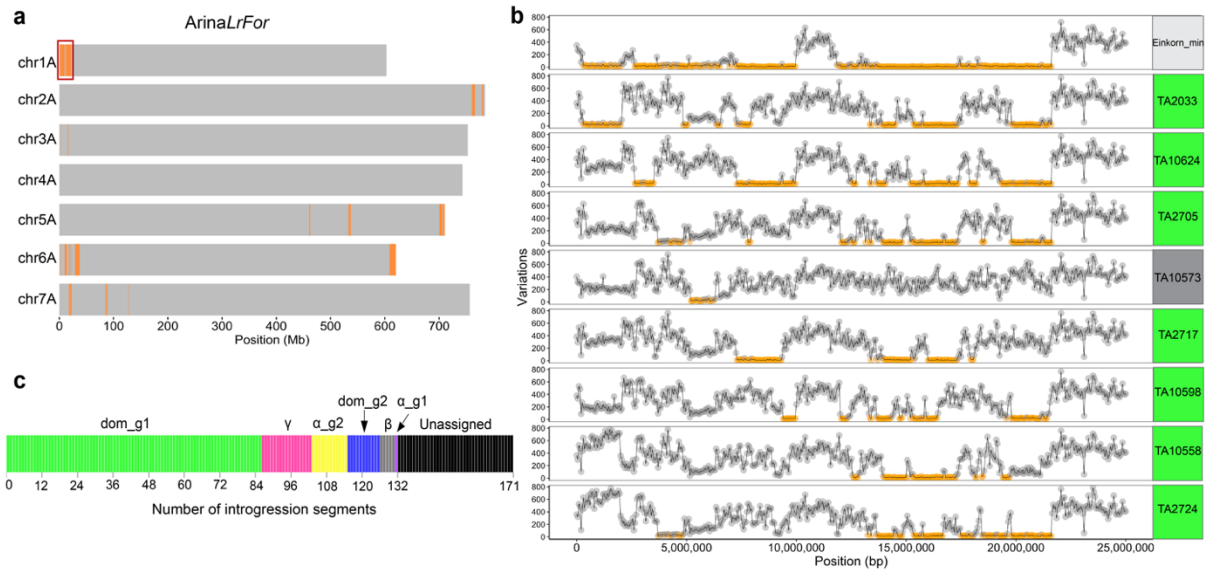


Fig. 4. 1. Einkorn introgressions into bread wheat.

a, Einkorn introgression (highlighted in orange) into ArinaLrFor identified by the *k*-mer *variations* approach (IBSpy). The red square at 1AS corresponds to the region shown in detail in (**b**). **b**, IBSpy variations between ArinaLrFor (chromosome 1A, position 0-25 Mb) and einkorn along chromosome arm 1AS. Regions with ≤30 variations are indicated in orange, corresponding to einkorn introgressions. Einkorn_min is the minimum number of variants across all re-sequenced einkorn accessions. The remaining plots illustrate the variations between ArinaLrFor and eight einkorn accessions. Accession names highlighted in green, and grey belong to domesticated groups 1 (dom_g1) and β, respectively. **c**, Number of introgression segments that could be assigned to a particular einkorn group (total =132 out of 171 segments) into the ten wheat genome references.

4.4.2. The origin of the D wheat genome

In this analysis we investigated the origin of the wheat D genome from the diploid genome of *Ae. tauschii* using a panel of 265 (242 non-redundant) accessions published in (Gaurav et al., 2022) and the *variations* detected by IBSpy. We aimed to determine the closest *Ae. tauschii* accessions or gene pool to the original donors of the wheat D genome using the 11 chromosome-scale assemblies (Walkowiak et al., 2020). We investigated the possibility of different pangenomes to share a specific set of IBS or near-IBS blocks with accessions in the panel and determine the similarity of the haplotype blocks to each wheat reference.

We used the lineage (L2) classes define by (Gaurav et al., 2022; Zhou et al., 2021) described in section 4.2.2 of this Introduction (**Table 4.2**). We carried out a detailed analysis to compared with our IBSpy *variations* focusing on the D sub genome of wheat. First, we defined the regions on the eleven wheat references that share similarity with any of the accessions of *Ae. tauschii* in the panel. Combining all the accessions we counted and filtered genome regions in the eleven

references that have ≤30 or ≤120 *variations* and name it *"Aet_min_30"* and *"Aet_min_120"* respectively. This criterion was stablished in our previous analysis with *T. monococcum* (section **4.4.1**). *Variations* counts ≥120 would be considered as non-*Ae. tauschii* sequence in the hexaploid pangenome cultivars.

Table 4. 2. Lineage representative accessions and corresponding class group in each study.

Accession	This study	Gaurav et al., 2022	Zhou et al., 2021
BW_01158	L2-SA	L2	AY61 (L2E)
BW_01096	L2-SB	L2	AL878 (L2W)
BW_01028	L3	L3	None
BW_23898	L1	L1	AY17 (L1W)
None	None	None	XJ02 (L1EX)
None	None	None	T093 (L1EY)

According to (Gaurav et al., 2022) regions in the D genome of wheat can have an origin from L1, L2, or L3, with L2 being the main donor. After defining the regions to be ≤30 of the entire panel, we next defined regions to be similar only to L3 and L1 in the eleven references by counting the regions that have ≤30 or 120 variations counts. For example, if a defined region in the D genome has *variations* counts ≤30 when comparing to a L3 genotype but has ≥120 variations count when comparing with all L2 genotypes, then we assigned the region as having an L3 origin. As a positive control we used the introgressed blocks into the D sub genome of the eleven chromosome-scale wheat assemblies from (Gaurav et al., 2022) which were assigned by using 100 Kbp non-overlapping window and lineage-specific *k*-mers.

In agreement with (Gaurav et al., 2022), we identified most of the regions to be from L2 accessions. Our results validated that a few regions are shared from L3 to the D wheat subgenome (**Supplemental Table. S4.2**). In addition, we also detected regions shared specifically with L3 on six genome assemblies on chr1D (**Fig. 4.2).** Unfortunately, in our dataset we only had sequence data from a single L3 genotype. Therefore, adding L3 genotypes to the collection might reveal novel L3 regions into wheat.

We next investigated if we could detect a genotype or group of accessions from the panel sharing significant larger blocks and high similarity to the eleven pangenome assemblies. Using hierarchical cluster maps on the *variations* count, we observed that L2 accessions formed two main subgroups of genotypes, suggesting that there is a subdivision within L2 accessions (**Fig. 4.2** in blue and green). The first subgroup of L2 accessions shared large blocks with a large proportion

of *variations* ≤30 and high similarity across multiple chromosomes of all D genome on the eleven chromosome references. We propose those genotypes to be the closest to the D wheat donor (s) and we called this group Lineage 2 subgroup A (L2-SA, **Fig. 4.2** in blue). The lineage specific genotypes clearly clustered separated among them. We propose four main groups as L2, L2-SA, L2-SB, and L3. The group L2-SA sharing the most similarity with the wheat D sub genome of the pangenome assemblies (PG). In **Fig. 4.2** we show ten representatives *Ae. tauschii* accession from each group and only one representative of L3 with publicly available data.

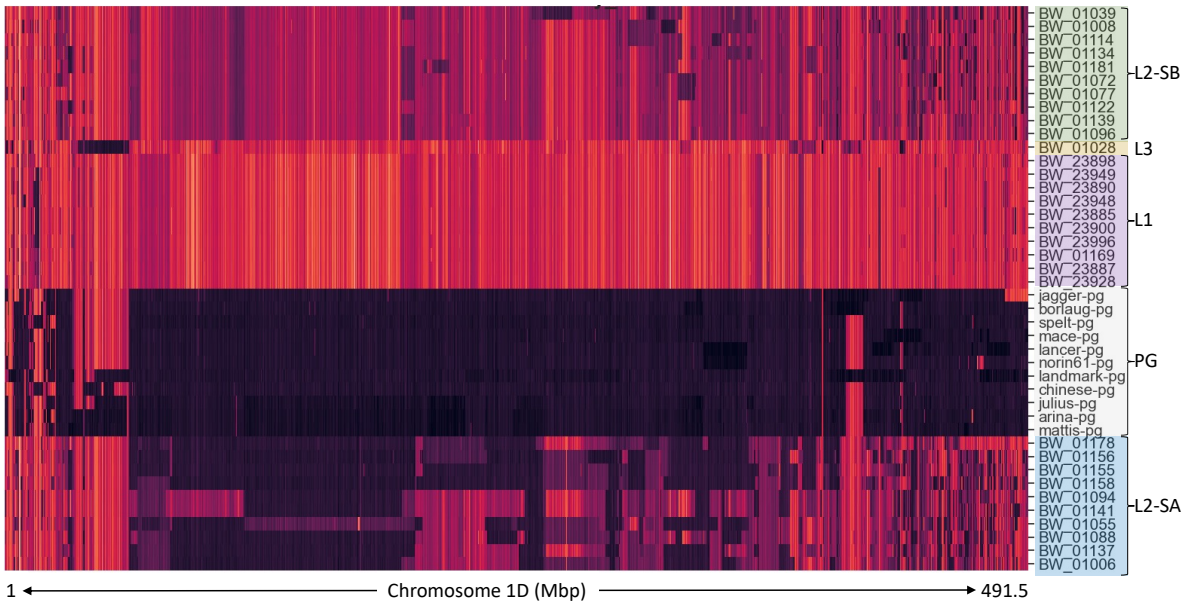


Fig. 4. 2. Lineage specific cluster map.

Ten representative genotypes of the proposed lineages and sub lineages clustered using IBSpy *variations* on chromosome 1D of reference Stanley. L1 in purple, L2-SA (blue), L2-SB (green), L3 (yellow) and the pangenome wheat assemblies (PG). In the dataset we have one L3 accession available. Darker colours indicate low *variations* against the chr1D of Lancer while orange-clear colours indicate high *variations* count across the chromosome physical position.

We next plot the *variations* count distribution of the whole genome in 50 Kbp windows of one representative of the L2-SA group (BW_01158). The distribution showed three main peaks, one at ~ ≤30 *variations* count (**Fig. 4.3a.,** blue panel) and a second at ≥120 (**Fig. 4.3a.,** purple panel**)** with a few datapoints between 30 and 120 *variations* count category **(Fig. 4.3a.,** green panel). The third peak was at > ~400 *variations* count corresponding to the data against the A and B wheat subgenomes (yellow panel). Removing the A and B comparisons left the two main peaks (**Fig. 4.3b).** Plotting the *variations* distribution by chromosomes detect slightly differences where some of them have more data in the ~≤30 *variations* count category (**Fig. 4.3c).** The *variations* count across the chromosome physical position of Staley showed similar blocks-like as when comparing

wheat vs wheat where *variations* count ≤30 extended several 50 Kbp windows across the chromosome (**Fig. 4.3d).**

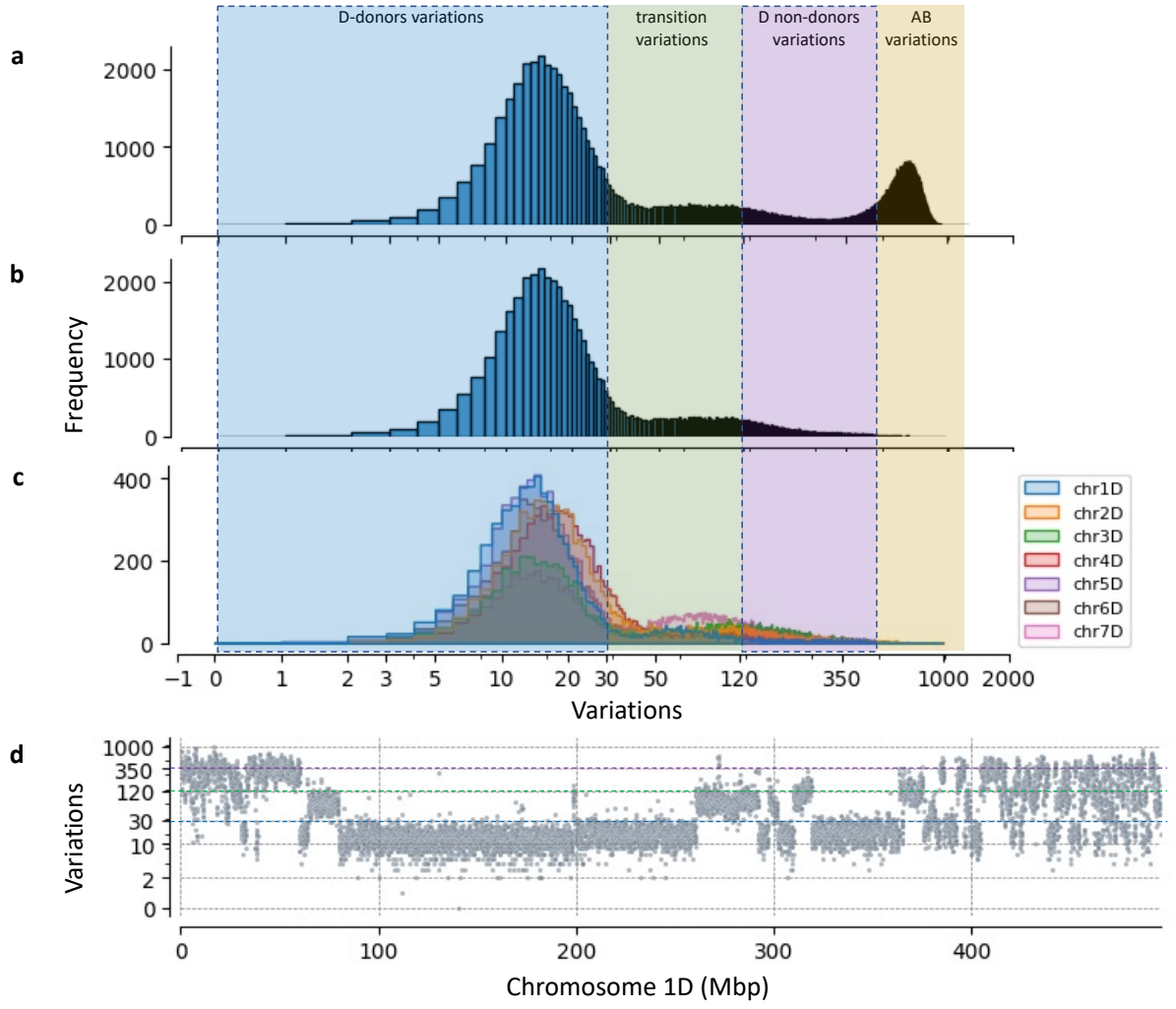


Fig. 4. 3. The IBSpy *variations* landscape of *Ae. tauschii* vs Stanley reference from a L2-SA group representative. **a**) *Variations* distribution of BW_01158 genotype, one of the closest D-donors belonging to the L2-SA sub lineage vs one of the pangenome assemblies (Stanley). Different colours show the hypothetical subdivisions based on *variations* levels. *Variations* ≤30 correspond to the predicted D-donors (blue), in green >30 and <120 *variations* (transitions variations). In purple >120, <500, and median ~350 D non-donors and distant genotypes. In yellow *variations* between D and A and D vs B genome *variations* >500. **b**) Showing the *variations* count from D genome only and **c**) distributions by chromosome. **d**) Shows the *variations* count across chromosome 1D of Stanley. Each dot corresponds to the *variations* in 50 Kbp window. The blue line indicates the ≤30 cut-off of the hypothetical *variations* counts of the wheat D subgenome donors while the red (>120) and purple (~350) lines are a more distant region. In the (Gaurav et al., 2022) dataset we identified 62 genotypes to belong to this group (L2-SA).

The second subgroup shared only small portion of *variations* ≤30 and have most of its windows in the range of >30 and <300, with a median of ~120 *variations*. We called this group Lineage 2 subgroup B (L2-SB) (**Fig. 4.2**, in green). This group shared a few large blocks with ≤30 *variations*.

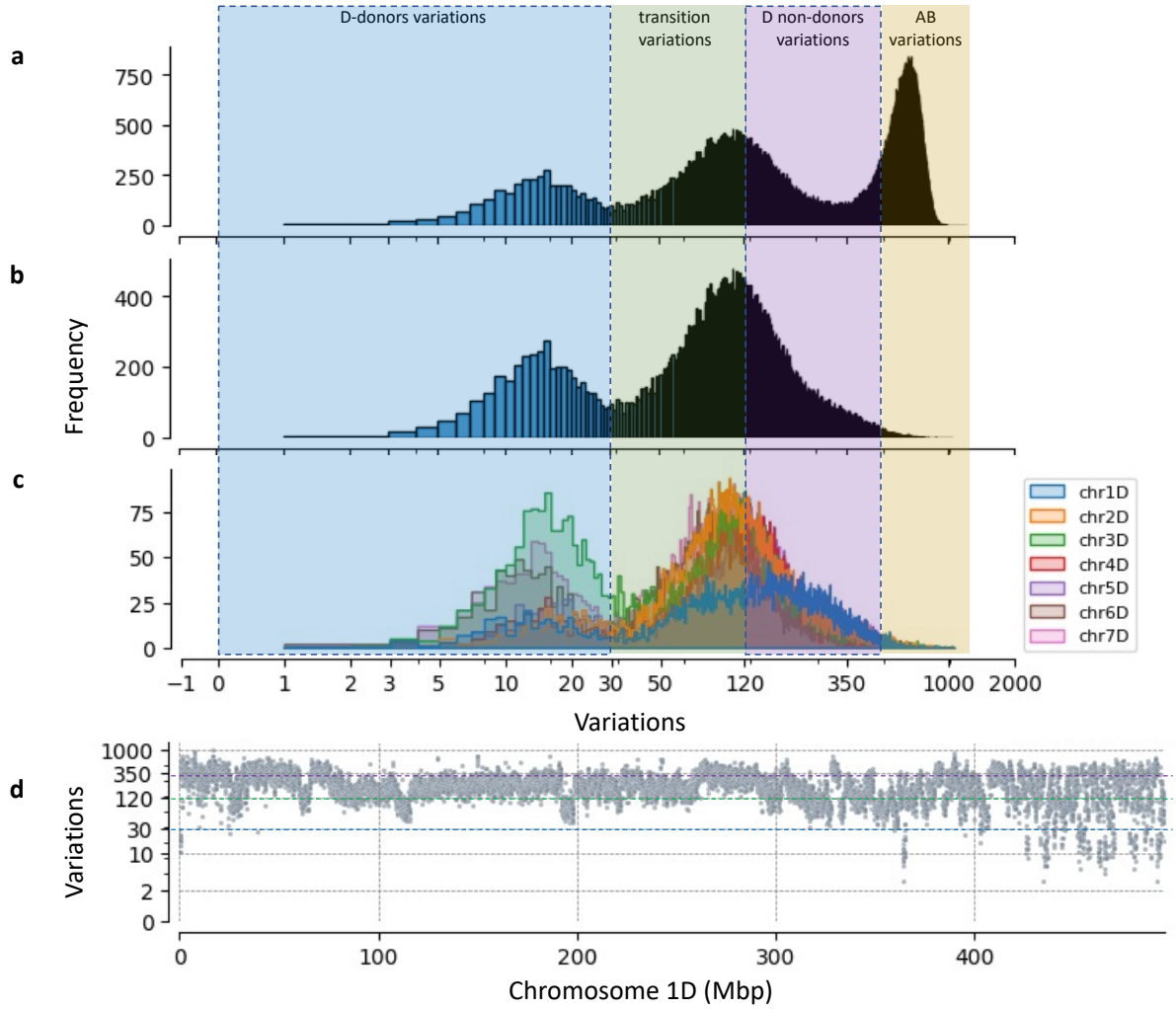


Fig. 4. 4. Example of L2-SB (BW_01182) representative.

This is an example where a genotype from the L2 (L2-SB in our study) described in Gaurav et al., 2022 has a *variations* distribution predominantly > 30 *variations* count, and we hypothesize is not the D subgenome wheat donor. In this case the peak distribution is at ~120 *variations* which is not the ≤30 variations cut-off of the D subgenome wheat donor. A few regions across the chromosome have *variations* ≤30 as shown in the scatter plot in **d**. In the Gaurav et al., 2022 dataset we identified 87 genotypes belonging to this group (Supplementary table: https://github.com/quirozcj/PhD_thesis_JQCH_2022/tree/main/chapter_4/Ae_tauschii).

On the other hand, genotypes from lineages L3 and L1 clustered in an independent group in the cluster map having high number of *variations* compared against the D genome of Stanley as depicted in **Fig. 4.2** (in yellow and purple respectively). The L3 genotypes shared few ≤30 *variations* block, and the *variations* distributions ranged from >120 to ~500 with a median of ~350 (**Fig. 4.5**). L1 rarely shared blocks ≤30 with any chromosome and the median *variations* was also similar to L3 at ~350 variations having the similar *variations* distribution (**Fig. 4.6**). The A and B sub genomes had a third distribution shape with variations >500 to 1,000 and median at ~700. A comparison of the *variations* distribution of representative genotypes for each subgroup is shown in **Fig. 4.7.**

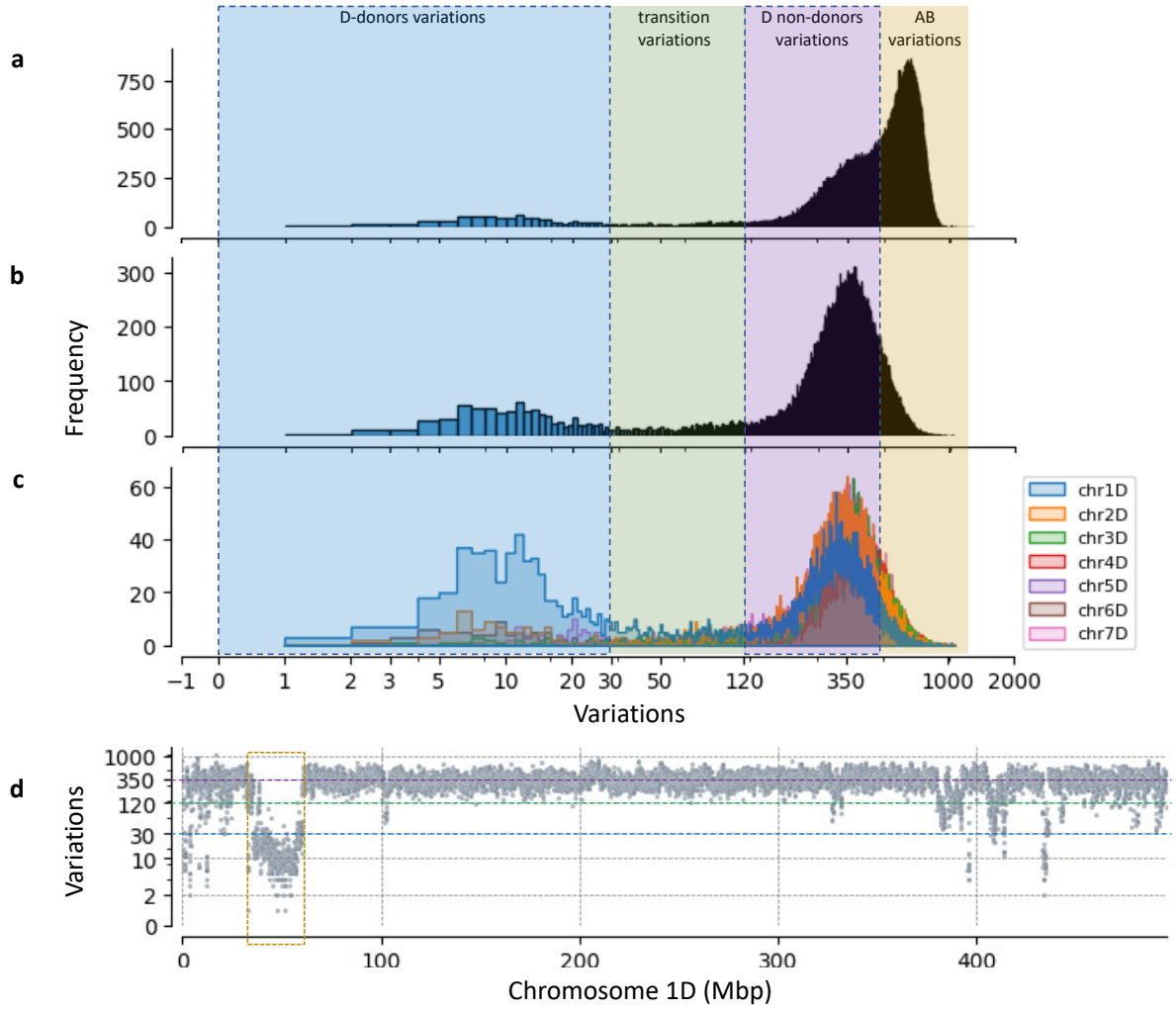


Fig. 4. 5. A representative of L3 (BW_01028) vs Stanley D genome of wheat.

In **a**) and **b**) the high proportion of *variations* count located in the purple panel which is the category of D subgenome non-donors. **c**), In blue, chromosome 1D, which has a region with low *variations* frequency ≤30 introgressed from L3 and is indicated by the yellow rectangle in **d**) chromosome physical position.

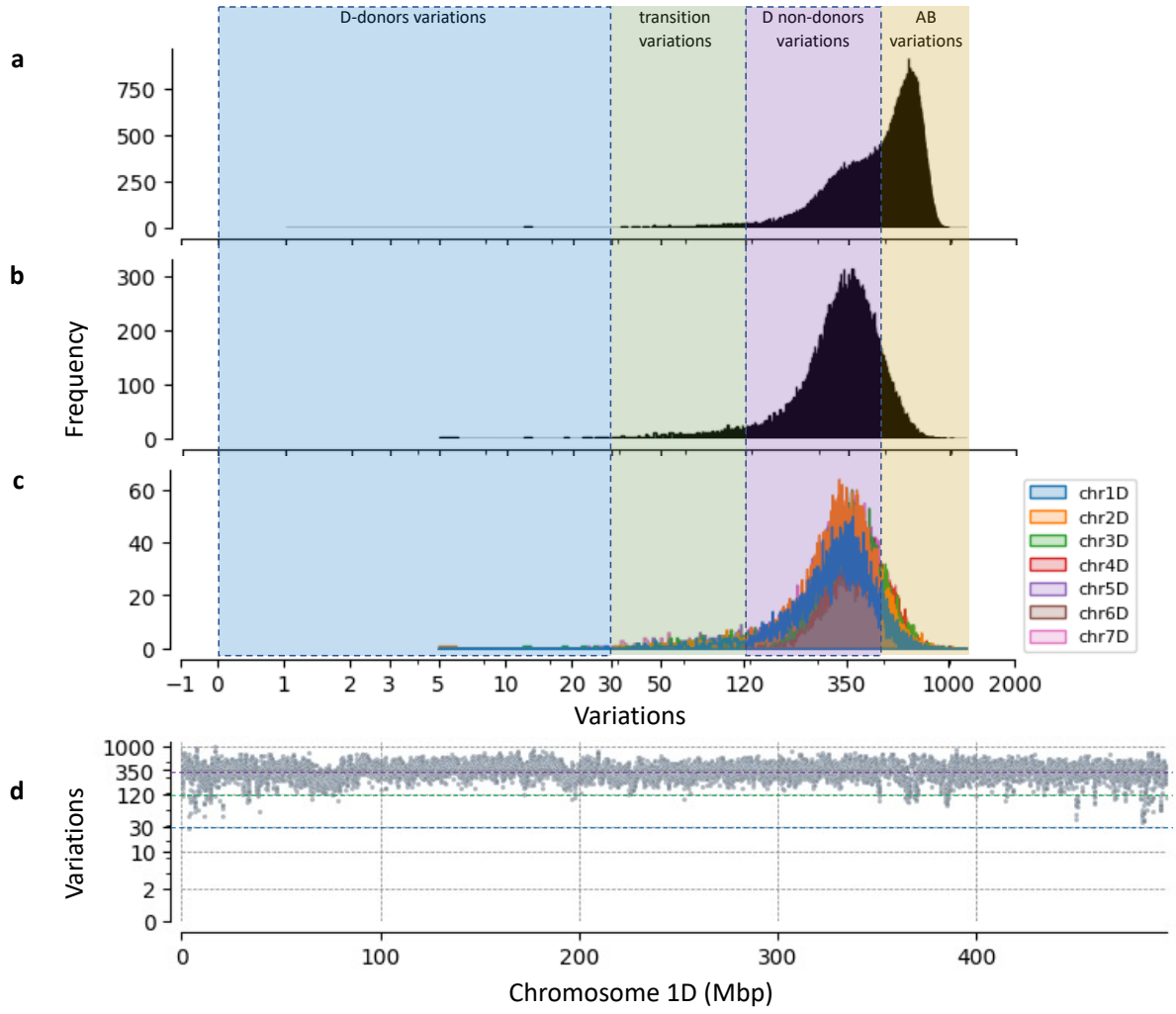


Fig. 4. 6. A representative of L1 (BW_23898) vs Stanley D genome of wheat.

a) *Variations* profile including the A and B sub genomes of wheat for comparison**.** Similar to L3, in this case a L1 genotype shows a peak with mean at ~350 *variations* in **b** and **c**, indicating that this is also a D non-donor. The position of the peak suggests that L3 and L1 are equally distant to the D wheat genome. This L1 accession has almost no windows with ≤30 *variations*. **d**), The *variations* count in the scatterplot suggests there are no segments of L1 on Stanley chr1D.

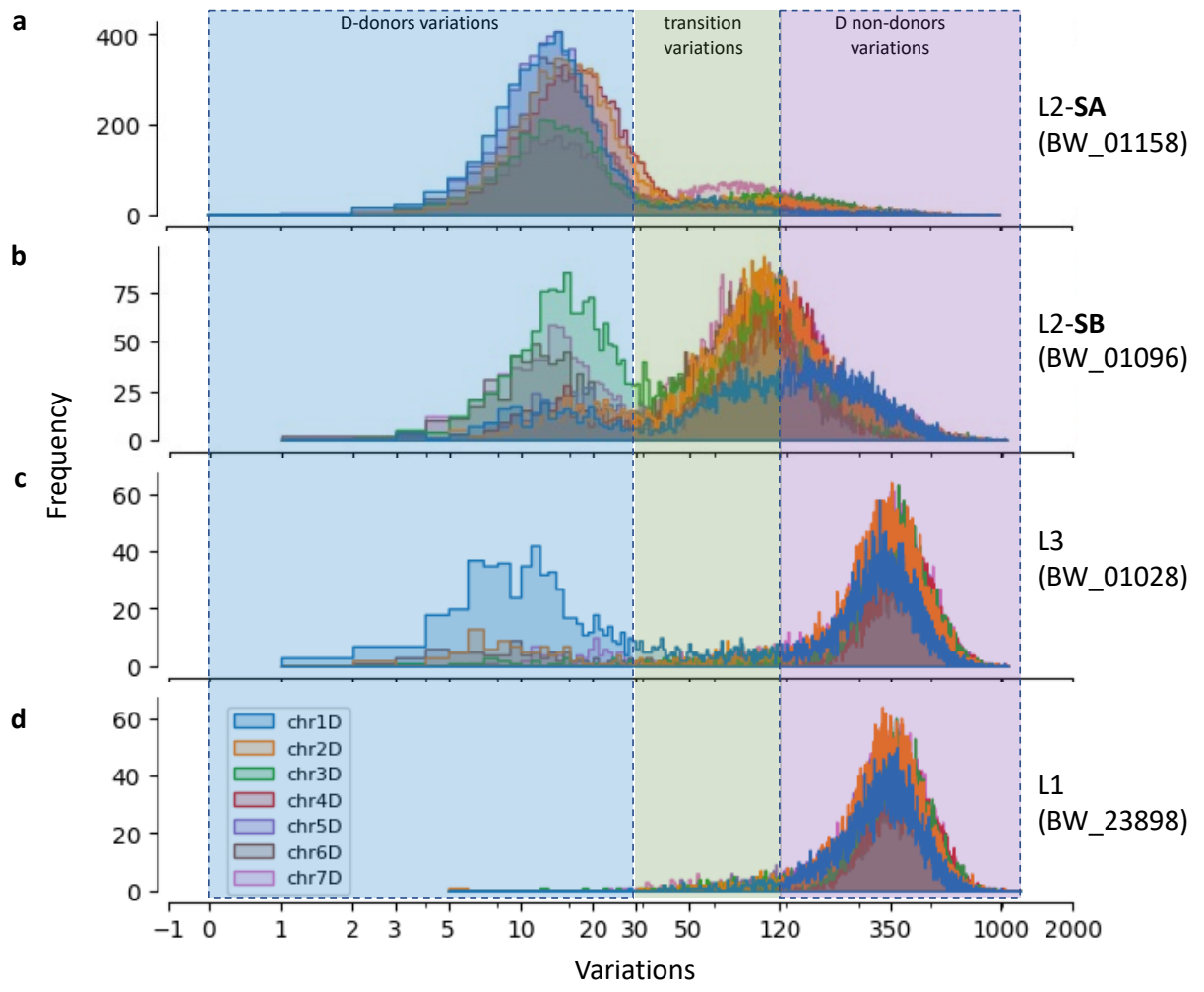


Fig. 4. 7. *Variations* similarity among *Ae. tauschii* sub lineages vs the D wheat genome (Stanley).

Variations distributions of representative genotypes belonging to each of the lineage using Stanley as a reference. **a**) L2-SA (BW_01158), the closest donor to the D subgenome of wheat, **b**) L2-SB (BW_01096) similar to the D donor, but not the closest wheat. **c**) L3 (BW_01028) and **d**) L1 (BW_23898) have equally level of variations against the D subgenome but L3 (**c**) having some regions with *variations* in the ≤ 30 category (blue) indicating introgressions into the D wheat genome by natural hybridisations.

To validate our results of differences in lineages, we used the *Ae. tauschii* references reported in Zhou et al., 2021 (described in the Introduction). In their study Zhou et al., assembled and assigned genome references to represent each of the *Ae. tauschii* lineages: AY61 (L2E), AY17 (L1W), XJ02 (L1EX), and T093 (L1EY). In our study, we first used the AY61 reference (which is the closest related to wheat D genome) to visualize the *variations* distributions. Although the *variations* distribution indicated that these two genotypes, L2-SA (BW_01158) and L2E, belong to the same lineage (large blocks of ≤30 *variations*), the variation distributions across the chromosome position indicates a

poor-quality of the assembly **(Fig. 4.8a).** Using a genotype from the group L2-SB (BW_01182) we obtained a distribution with *variations* peak at ~120 (**Fig. 4.8b**). On the contrary, when comparing a genotype from group L3 (BW_01028) against the AY61 reference, we observed that the distribution with the highest peak had a median at ~350 variations (**Fig. 4.8c**). Similar profile was detected when using a genotype from the L1 (BW_23898) with *variations* at ~350 (**Fig. 4.8d**). These data are consistent with (Zhou et al., 2021) showing that L2E group corresponds to the L2-SA group presented in our analysis.

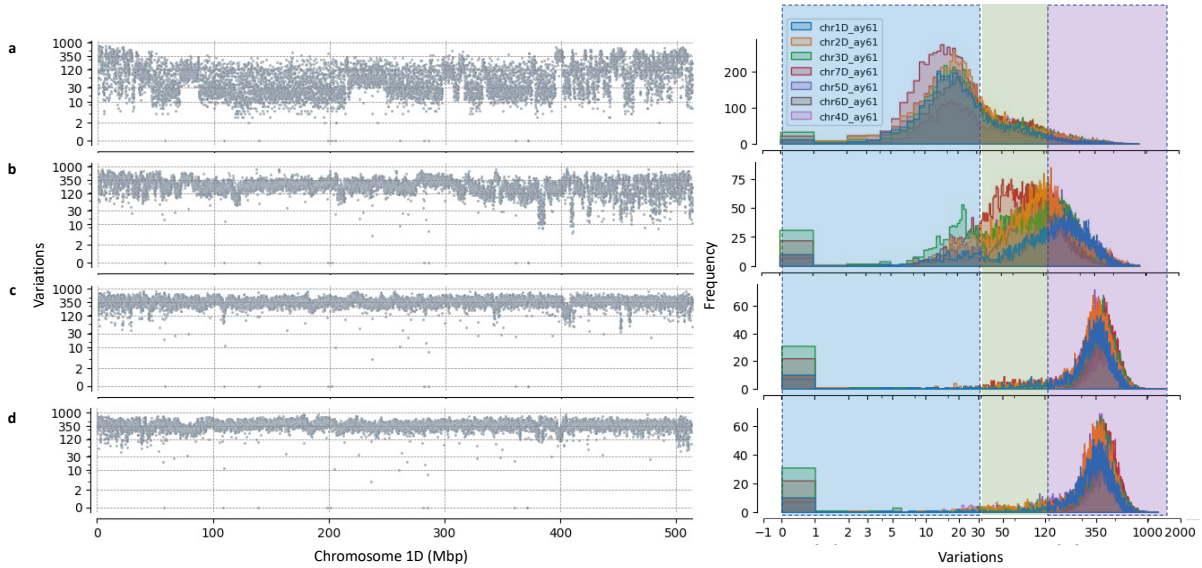


Fig. 4. 8. Comparison of *variations* profile using reference AY61 (L2E) and representatives of each lineage class. **a**) L2-SA (BW_01158) vs genome reference AY61 (L2E class in Zhou *et al.,* 2021) corresponds to the closest D wheat donor class. **b**) L2-SB (BW_01096) which correspond to the AL878 (L2W) group in Zhou et al., 2021. **c**) L3 (BW_01028) for which there is no representative accessions in Zhou et al., 2021. **d**) L1 (BW_23898) corresponds to the AY17 (L1W) in Zhou et al., 2021 group. **c** and **d** are equally distant to AY61 (L2E) group and there are no regions of shared hybridisations between those two groups against AY61 (L2E).

Different studies have proposed AL878 as one of the genetically closest *Ae. tauschii* accession to the hexaploid wheat D genome. On this basis was chosen as the accession to develop the *Ae. tauschii* reference genome (Luo et al., 2017). However, using previous observations in a pilot study with IBSpy, we hypothesize that AL878 assembly is part of the L2-SB sub lineage, which is not the closest donor of D wheat. To confirm this, we queried a L2-SB genotypes (BW_01182) against the AL878 reference. As predicted, we observed that the mean variations were ~<30 (**Fig. 4.9b**). On the contrary, using a L2-SA genotype (BW_01158), the median variations were at ~120 (**Fig. 4.9a**).

Comparing AL878 against L3 (BW_01028) or L1 (BW_23898) confirmed that they are highly different (distant) with median variation at ~350 (**Fig. 4.9cd**).

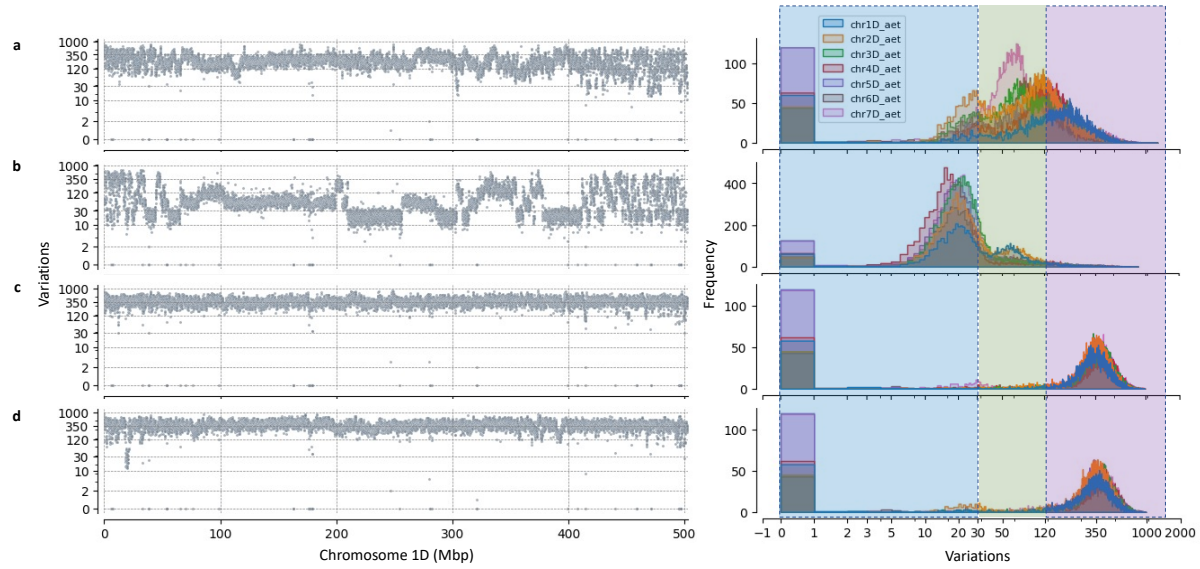


Fig. 4. 9. AL878 reference belongs to the L2-SB lineage class.

a) L2-SA (BW_01158) vs genome reference AL878 (L2W group in Zhou et al., 2021). **b**) L2-SB (BW_01096) which corresponds to the AL878 (L2W) group in Zhou et al., 2021. **c**) L3 (BW_01028) and for which there is no representative accessions in Zhou et al., 2021. **d**) L1 (BW_23898) correspond to the AY17 (L1W) in Zhou et al., 2021 group. **c** and **d** are equally distant to AY61 (L2E) group and there is no evidence of shared hybridisations between those two groups against AY61 (L2E).

Analysis using the AY17 (L1W), XJ02 (L1EX), and T093 (L1EY) references indicates that the L1 accessions in our dataset correspond mainly to the L1W group (**Fig. 4.10d**). However, it also suggests that the accessions in the L1 group is a mixture of L1s with different levels of variations and that a subgroup within this group may also exist sharing near-IBS like regions as shown in chr7D of AY17 in (**Fig. 4.11,** yellow rectangle).

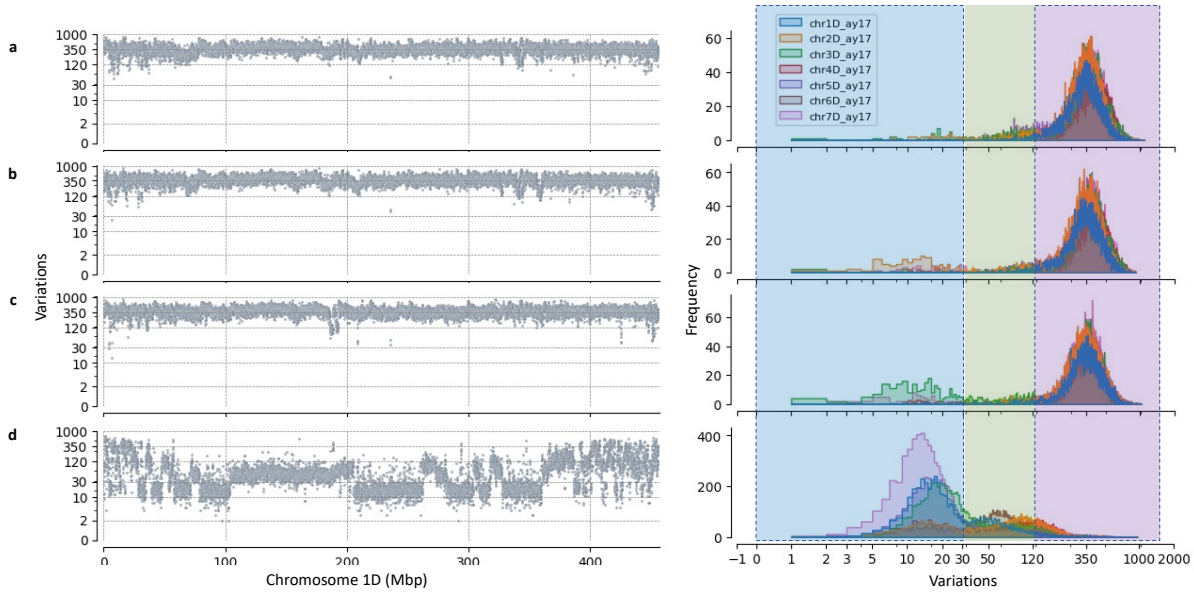


Fig. 4. 10. AY17 (L1W, in Zhou *et al.,* 2021) reference corresponds to L1 in Gaurav *et al.,* 2022 and in this study.
a) L2-SA (BW_01158) vs genome reference AY17 (L1W, in Zhou *et al.,* 2021). **b**) L2-SB (BW_01096) which correspond to the AL878 (L2W) group in Zhou *et al.,* 2021. **c**) L3 (BW_01028) and for which there is no representative accessions in Zhou *et al.,* 2021. **d**) L1 (BW_23898) corresponds to the AY17 (L1W) in Zhou *et al.,* 2021 group. **a**, **b** and **c** are equally distant to the AY17 (L1W) group and there is no evidence of shared hybridisations among those two groups against AY17.

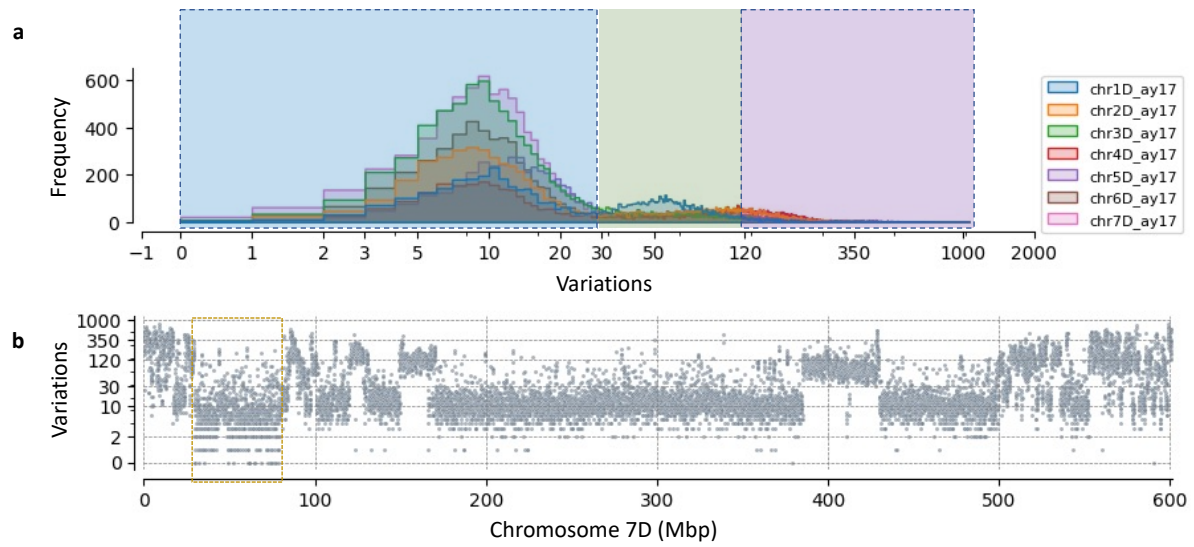


Fig. 4. 11. AY17 (L1W) vs BW_23933, a L1 genotype.
a) a L1 genotype showing high similarity to the reference AY17 (L1W) on different chromosomes having the <30 *variations* count cut-off (blue box). **b**) Chromosome 7D example of a region highly similar to the reference AY17 (yellow rectangle) which is indicative of a near-IBS region between two wild *Ae. tauschii* accessions.

Comparing L3 genotypes against the five *Ae. tauschii* genome assemblies available in this study and against the Stanley reference, we observed that the median *variation* was ~ 350 count indicating that the L3 is a completely different lineage on sequencing similarity to L1 or L2 (**Fig 4.12**). Given that L3 is equally distant to L1 and L2, this might suggest that L3 first diverged from a common L1 and L2 ancestor.

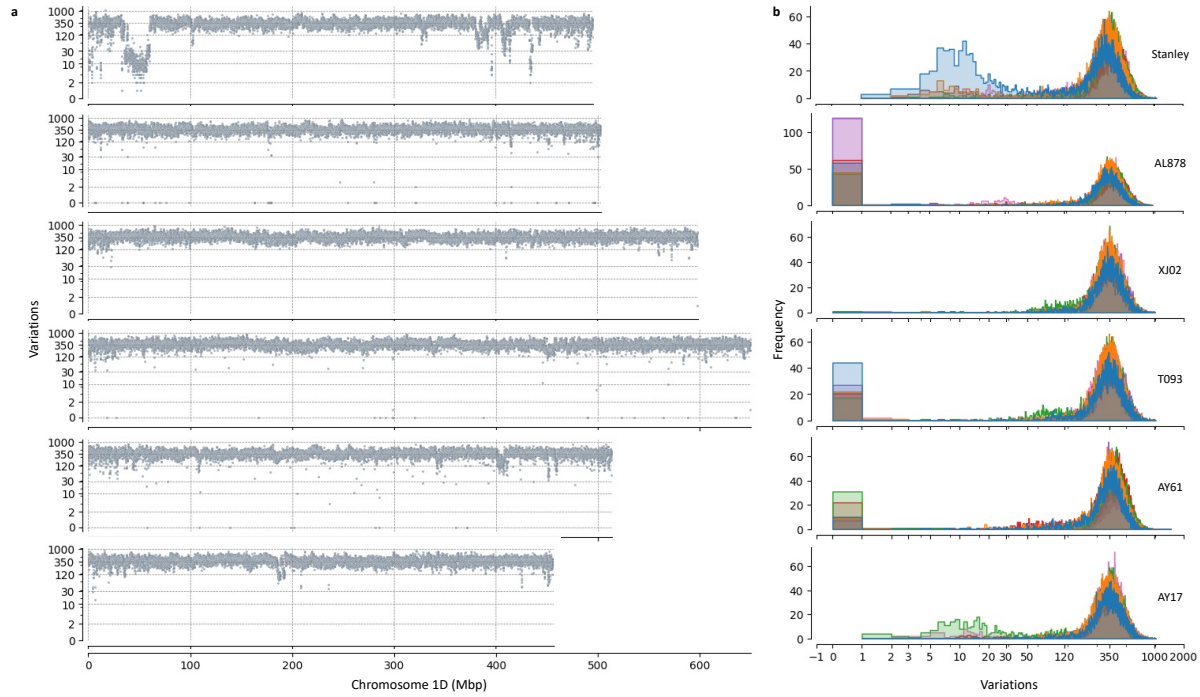


Fig. 4. 12. **L3 (BW_01028) is equally distant to the five genome assemblies of *Ae. tauschii* and the wheat D subgenome.**

In all cases the common query L3 genotype BW_01028 was used against the other five genome assemblies**. a)** from top to bottom as indicated in **b)** corresponding reference, BW_01028 accession vs Stanley, AL878 (L2-SB in this study, L2 in (Gaurav et al., 2022), and L2W in (Zhou et al., 2021), XJ02 (L1EX), T093 (L1EY), AY61 (L2E), and AY17 (L1W).

Zhou *et al.,* 2021 reported that the AY61 reference to be part of the closest donor gene pool to D genome due to a high proportion of shared orthologous gene pairs (density of orthologous gene pairs). Analysis of the synonymous mutation rate (Ks) between ortholog genes (subspecies divergence) showed the highest peak with Ks=~0.0009. i.e., a few Ks changes. In the same analysis AL878 had Ks=~0.0073, and T093 and XJ02 vs CS had Ks=~0.0077. In our IBSpy *variations* count analysis, we propose that Ks=~0.0009 is equivalent to having <30 IBSpy *variations* in 50 Kbp window, Ks=~0.0073 to have >30 and <300 but having median of ~120, and Ks=~0.0077 would be equivalent to have median of ~350 variations (**Fig. 4.13**). As a point of reference for comparisons with other important crops, B73, a maize elite line compared to the elite Mo17 has Ks=~0.025

(Zhou et al., 2021). These two maize genotypes belong to two distinct heterotic groups that were bred independently and combine well to form hybrids. In the future, it would be interesting to test if two highly distant genotypes of wheat or wild relatives results in some form of heterosis by testing highly diverse wheat genotypes developed in synthetics from crosses with wild distant relatives.

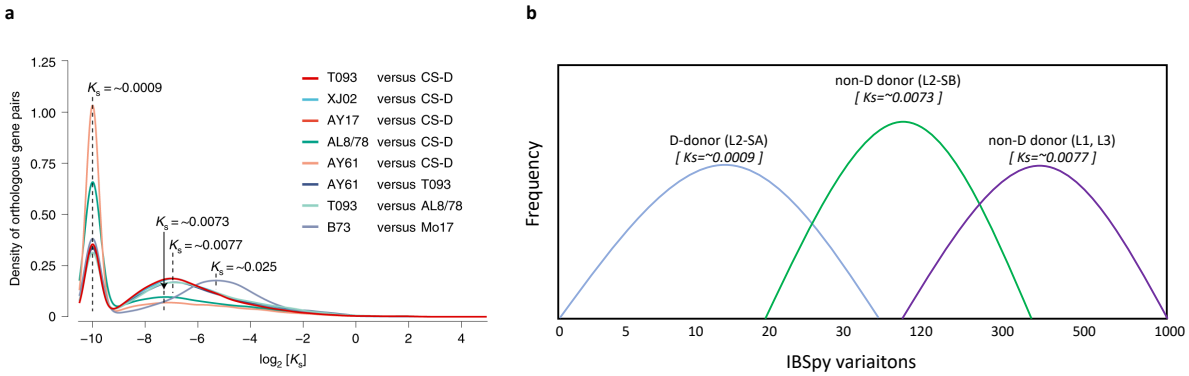


Fig. 4. 13. Ks to IBSpy. Zhou *et al.,* 2021 subspecies divergence mutation rate (Ks) analysis to IBSpy variations.
Data in **a**) was taken from Zhou *et al.,* 2021. **b**) is our proposed equivalence of **a**) with IBSpy *variations* of *Ae. tauschii* lineage relationship to the D-wheat genome.

To further investigate if the set of lines on L2-SA and L2-SB have different geographic distributions, we compared their collected regions. We found that accessions belonging to the L2-SA have their origin mainly in the North of Iran at ~ 36.695300 (latitude) - 53.536500 (longitude). On the other hand, accessions of the L2-SB were mainly collected from a region in Azerbaijan at 40.631900 (latitude) - 48.636400 (longitude) with a few exceptions in each subgroup. This information suggests that these two subgroups have been hybridizing separately in nature and a few genotypes might have had some crosspollinations over the years by seed dispersal naturally or by humans. Another explanation of the presence of mixtures in the two groups would be the misclassification during seed propagation and labelling. These results open new questions regarding the *Ae. tauschii* evolution and the D wheat closest donor (s) whiting subgroups.

In summary, our results validate IBSpy to differentiate among the D wheat subgenome progenitors and lineages. We found lineage specific donors into the wheat pangenome assemblies in agreement with (Gaurav et al., 2022). Our results suggest accessions from the L2-SA group from the North of Iran to be the closest donors of the D wheat genome. Variation analysis supports our previous findings that <30 IBSpy *variations* count in 50 Kbp window between two genotypes to belong to the same immediate gene pool. For example, genetic variation or sequence identity normally found between two wheat cultivars or a landrace with variations accumulated <10,000

years ago. *Variations* >30 would indicate mutations accumulated >10,000 year ago, hence most likely from hybridizations from wild relatives accumulated before the hexaploid polyploidization in tetraploids or diploids ancestors. These results also provide information to propose members of L3 as candidates for further genome assemblies projects to maximise the discovery and exploitation of D progenitors genome diversity.

4.4.3. Genetic diversity of the D genome in the WatSeq dataset

In our example above (**4.4.3.**), we detected unique blocks from L2 and L3 into six pangenome references similarly to (Gaurav et al., 2022). Due to the reduced number of genome assemblies compared to the vast number of modern and landraces accession in germplasm collections, we hypothesize that these lineage specific blocks, in addition to novel undetected genome regions, may be present in other wheat genotypes with similar or extended block sizes and recombination positions.

Gaurav et al., 2022 (Gaurav et al., 2022) identified regions in the wheat genome originated exclusively from L3, suggesting that at least two hybridisation events gave rise to the D wheat genome. To extend the analysis on the diversity of the D subgenome of wheat and determine if there are additional regions not detected in the pangenome assemblies, we explored the D genome regions unique to modern or landraces absent in the wheat pangenomes. We compared the *variations* fingerprint of the WatSeq panel along the *Ae. tauschii* accessions and explored the level of similarity among the *Ae. tauschii,* landraces, and modern wheat cultivars.

In this example, we used the chromosome 1D of Stanley reference as our case study since it has lineage specific regions from L2 and L3 reported in (Gaurav et al., 2022). More specifically in this analysis, we use the *High molecular weight glutenin (Glu-D1) locus Glu-D1* haplotype block gene region inherited exclusively from L3 (BW_01028) reported in (Delorean et al., 2021). The *Glu-D1* locus is involved in the quality of dough for bread and there are mainly two *Glu-D1* haplotypes in wheat, the *2+12* with origin from L2, and the *5+10* allele from L3. In total, there are 43 haplotypes described in a collection of 273 *Ae. tauschii* accessions reported by (Delorean et al., 2021). The *5+10* allele is preferred for superior dough quality in the process of bread making, and it is present in Stanley, Landmark and Jagger. The other eight pangenome references carry the *2+12* allele.

In this analysis, we first explored the prevalence of each of the haplotypes in the WatSeq dataset using IBSpy *variations* count. The *Glu-D1* is located at ~411 Mbp in Stanley reference, but we considered the entire block from L3 (BW_01028) in Stanley to be from 407 to 415 Mbp based on the IBSpy *variations* fingerprint (**Fig. 4.14a**, blue bar). This is shown in the L3 accession, BW_01028,

having low *variation* count (i.e., purple-dark colour) across the 407 to 415 Mbp interval when compared to Stanley as a reference. To simplify the analysis, we considered genotypes having the *5+10* allele if they were similar to Stanley reference otherwise having the *2+12* or unknown category (which could have another origin than L2 or being a deletion in the region). Our results of the > 1,000 hexaploid genotypes demonstrates that there is a high proportion of accessions having the *5+10* allele both in landraces and modern cultivars.

In agreement with (Delorean et al., 2021), we detected the entire *5+10* allele haplotype block on chr1D of Stanley, Landmark, Jagger, and additional lines, Paragon, Cadenza, Weebill and Borlaug. In addition to the modern genotypes, we identified several landraces having the entire or fragments of the block (**Fig. 4.14b,** blue bar); for example, genotypes WATDE0498, WATDE0669 and WATDE0047 (**Fig. 4.14b**, bottom). This indicates that the *5+10* haplotype has been selected in landraces probably by local farmers in specific regions for its bread-making qualities. Furthermore, the distinct block sizes detected in the genome region, suggest that multiple recombinations have taken place after the initial hybridisations before commercial breeding started. It is important to remark that in this pilot analysis, we considered only two haplotype alleles (5+10 and 2+12) but these results on recombined blocks indicates that additional haplotypes in the region may have risen in the region with novel functional *Glu-D1* haplotypes untapped in landraces. In future analysis it would be worth to explore individual SNPs in the region and associate for distinct phenotypes in bread-making qualities in those landraces and other modern wheats cultivars.

Exploring outside the genome region, we detected additional blocks in the Stanley *GLU-D1* downstream and upstream regions which are similar to L3 (BW_01028) (**Fig 4.14b**, pink bars on top; dark colour in the BW_01028 row). In addition to the similar blocks to the L3 blocks and Stanley, we noticed that several landrace accessions and other modern cultivars had high *variations* counts (> 120) outside of the blocks defined in (Delorean et al., 2021) or (Gaurav et al., 2022). We hypothesize that those high *variations* count are the result of recombined blocks missed in the genome references with origin, either from non L2 genotypes or from deletions, but remain in landraces and other cultivars. This because often those genotypes had an L3 block in the region described in (Gaurav et al., 2022) (**Fig. 4.14ab** green and pink bars).

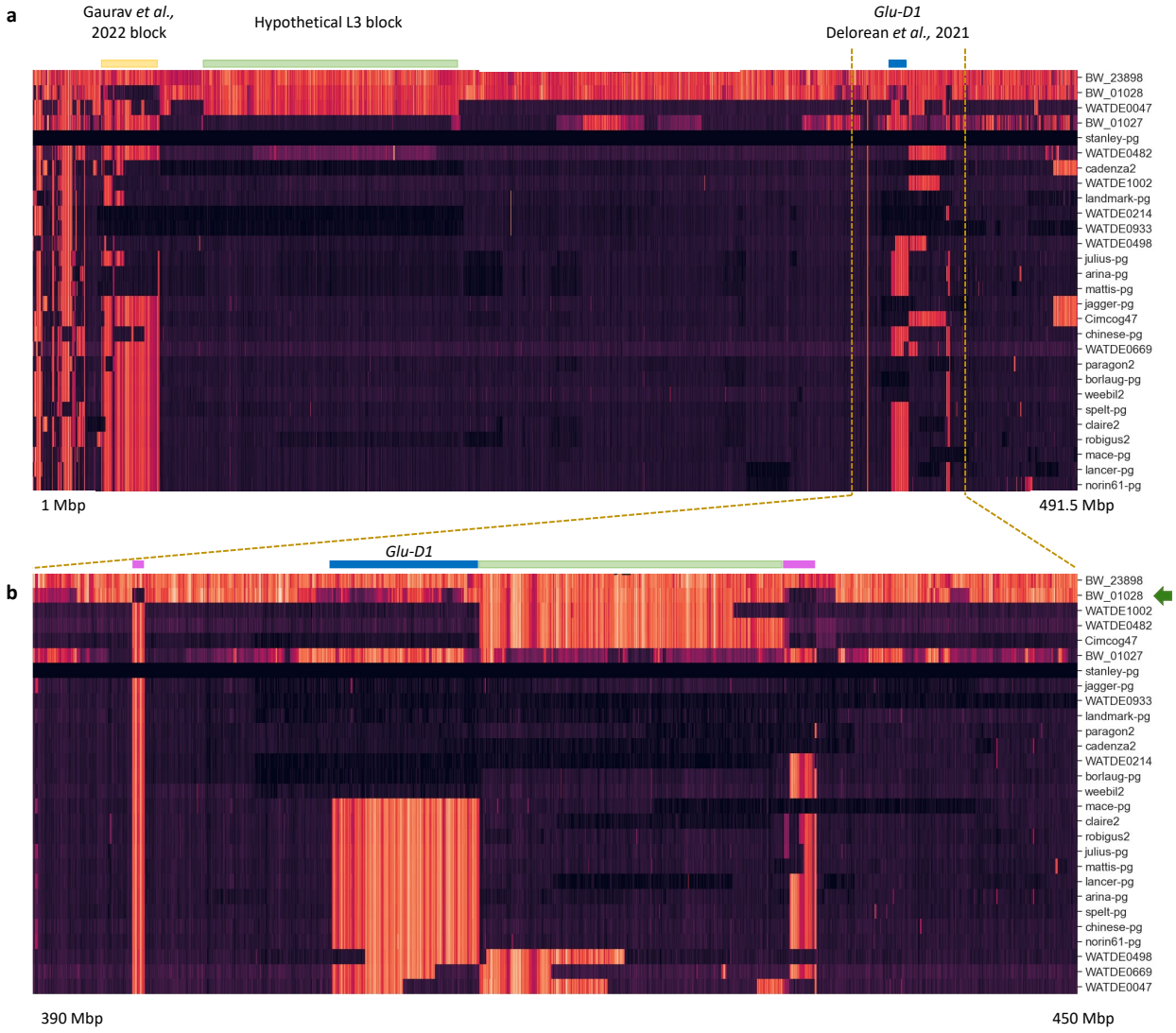


Fig. 4. 14. Lineage specific haplotype blocks in chromosome 1D in the WatSeq collection (Stanley as a reference). The heatmaps colours indicate similar to Stanley reference (dark purple) and very different to Stanley in orange clear. **a**) *Variations* count profile of the complete chromosome 1D in Stanley vs a subset of Watkins and modern from the WatSeq dataset. The blue bar indicates the Glu-D1 gene identified in (Delorean et al., 2021), a L3 region. The yellow bar indicates the L3 region identified in Gaurav et al., 2022 in some of the pangenome references, including Stanley. The green bar indicates hypothetical L3 blocks into the Watkin WATDE0047. **b**), a zoom in into the *Glu-D1* locus. Similar to **a**), the green bar indicates hypothetical non-L2 region in some of the Watkins as they are different to Stanley. L2 regions are the most common in wheat, and anything different to L2 regions would be candidates to come from other lineages, introgressions, or deletions. Pink bars indicate additional L3 blocks in Stanley as they are very similar to BW_01028 (L3 genotype, green arrow).

To validate our previous hypothesis and explore lineage specific regions absent in the pangenome assemblies, we used the affinity propagation (AP) haplotype calls described in **Chapter 3** including the *Ae. tauschii* accessions along the WatSeq samples. To account for the differences due to accumulated mutations over the last ~10,000 years, we focused on the haplotypes with "high_dmp" only. This parameter will group near-IBS samples in the same cluster and therefore the same haplotype. If a WatSeq genotype has similar *variations* pattern (but not identical IBS) to one of the *Ae. tauschii* accessions, we would be able to detect them with the "high_dmp" AP haplotype calls. Therefore, in this analysis we focused on the group of accessions highly similar (albeit not identical) to L3 accession BW_01028.

For this part we analysed the flanking region of the L3 block reported in (Gaurav et al., 2022) (**Fig. 4.15a**, yellow bar) since we observed that some Watkins accessions had high variations across a much wider region from 0 to ~200 Mbp on chr1D. Our results on this analysis, as predicted, the AP haplotype calls grouped 12 Watkins genotypes to have the same haplotypes as BW_01028 genotype (L3) from 104 to 202 Mbp based on the Stanley genome assembly (**Fig 4.15b,** accessions in green). In ten of these Watkins accessions, out of 99 1-Mbp windows in the predicted region, only a few windows were not assigned to the same haplotypes as BW_01028. An explanation for those missing windows may be due to the accumulation of *variations* after ~10,000 years of the initial hybridisation or because the actual L3 donor is not BW_01028, but rather a very closely related individual from the L3 population not present in this analysis. Two Watkins (WATDE1031 and WATDE1032) had different haplotypes to BW_01028 from 104 to 113 Mbp but after that region, they were grouped with BW_01028 until the end of the block (**Fig 4.15b**, black rectangle). This was consistent with the *variations* fingerprint observed in **Fig. 4.15a**, grey and green bar.

For comparison in this analysis, we included the BW_23898, which is a Lineage 1 (L1) genotype on the haplotype calls and as expected, this genotype is different to all other Watkins included in the analysis (**Fig 4.15b**, pink genotype). Therefore, these results support our hypothesis that BW_01028 is a close donor of this novel uncovered 99 Mbp block (**Fig 4.15a**, green rectangle) in landrace accessions. The *variations* fingerprint in the cluster map highlighted with the green bar, indicates that the introgressed haplotype may extend further, but the AP haplotype calls did not include it in the same group. The reason for this could be that the donor is a L3 genotype different to BW_01028 and a recombination between these two genotypes may have taken place before the hybridization with the D what genome. This would not be unexpected as we only have a single L3 genotype (BW_01028) and the original L3 donor of the wider interval could be more similar to BW_01028 in the more peri-centromeric interval and then recombined with another L3 individual distant to BW_01028. It could also be that the Watkins are more different in that region, and we

cannot detect it because none of the genome references have the entire block. Further genome assemblies from one of the Watkins carrying this L3 block or an *Ae. tauschii* L3 assemblies will resolve our hypothesis.

Finally, in in the same region, we detected a third block in a few Watkin (**Fig 4.15a,** green arrows) accessions, WATDE0900 and WATDE0549, which have slightly intermediate *variations* values over >30 count, not from the L3 but from a Lineage (L2-SA) genotype similar to BW_01055. This haplotype is very rare in the whole WatSeq dataset but due to the *variations* profile level, we hypothesize that this may come from a third hybridisation event and not from the most common L2-SA into D hexaploid wheat. The most common *variations* fingerprint into wheat comes from *Ae. tauschii* genotypes similar to the BW_01027 (L2-SA) and is the prevalent haplotype in all the pangenomes assemblies. The AP calls, however, did not cluster the entire block of the Watkins with the BW_01055 L2-SA genotype, only six consecutive 1 Mbp windows had the same group at 131 to 136 Mbp (**Fig 4.15b,** black rectangle). This may be because these Watkins are not close enough to BW_01055. The rationale of this third hybridization hypothesis is based on our previous observations that suggest that two genotypes from the same immediate gene pool in wheat after 10,000 years of hybridization would have <30 IBSpy *variations* count.

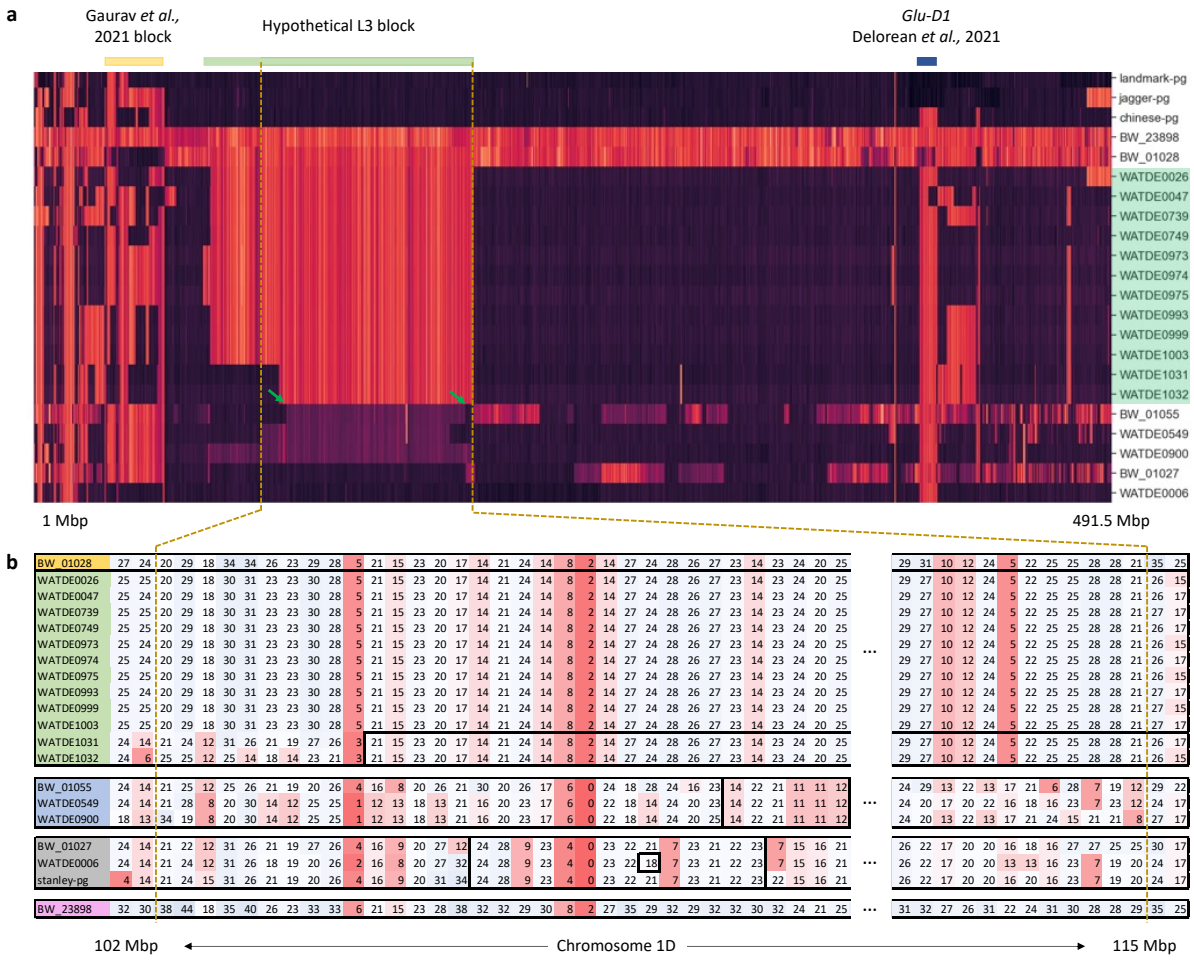


Fig. 4. 15. Landraces maintain extended L3 hybridisations blocks (Stanley as a reference).

The heatmaps colour indicates similar to Stanley reference (dark purple) and very different to Stanley in orange clear colours. **a**) the complete chromosome 1D in Stanley. The blue bar indicates the *Glu-D1* locus identified in (Delorean et al., 2021), a L3 region. The yellow bar indicates the L3 region identified in Gaurav et al., 2022 on Stanley. The green bar indicates hypothetical L3 blocks into some Watkins (green box). The green arrows indicate the boundaries of a third hypothetical hybridization having >30 *variations* count and similar to the BW_01055 genotype from the L2-SA lineage group. **b**), a zoom into the hypothetical L3 region indicated by the green bar**.** The table are the haplotype calls by AP. In yellow the L3 genotype (BW_01028) and its haplotypes across the chromosome physical positions in 1 Mbp window. The table is a portion of 115 Mbp representation of a larger region. In green the Watkins genotypes having several haplotypes calls identical to the L3 genotype in the hypothetical region. In blue, three Watkins genotypes having a reduced block size of the region similar to L3. In pink, a L1 genotype (BW_23898).

In summary, we validated that the wheat D genome donors and close related accessions of the hexaploid wheat have <30 IBSpy *variations* count. This *variation* threshold is similar to the observed in the *T. monococcum* panel comparison. These results support our hypothesis that close relatives of the wheat D and A genomes which hybridized <10,000 years ago share this level of

variations (<30). IBSpy *variations* blocks over >30 are older than 10k years ago in the wheat and wild relatives genomes and have a distant gene pool origin. We identified a set of genotypes, here named L2-SB, having >30 *variations* but having a mean of ~120 *variations*. We hypothesize that this level of *variations* diverged from the *Ae. tauschii* donors at > 10,000 years ago but are still genetically close related to lineage LS-SA, the closest D wheat donor. Future work would explore to categorize the level of *variations* of a wide set of the wild wheat relatives and their subpopulations to gain insight into the evolution of wheat from a broader perspective to improve our understanding for future breeding and wheat evolution.

Importantly, we demonstrated that our method detects introgressions using AP haplotype calls without the introgressions being present in the reference assembly. Using this method, we detected novel large L3 introgressions blocks into the D genome which are only present in the Watkins landraces. Although our findings of extended blocks from L3 are promising, intriguingly, it is still not understood why it is uncommon to find L3 blocks in hexaploid wheat outside of chromosome 1D. An open question remains to explore if L3 introgressed blocks into hexaploid wheat originated through an initial hybridisation with a L2-SA accessions (yet to be sequenced) or if a separate hybridisation between a tetraploid and L3 occurred. Hence, why does L3 blocks are only found in chr1D? Regardless of the explanation, there are expectations that advances in sequencing projects will help to uncover those unexploited natural hybridisations for breeding and to elucidate the hexaploid wheat origin.

4.4.4. Large wild wheat introgressions and deletions

Several introgressions from wild relatives have been reported in wheat and many of them are well known. Most of those studies to detect the introgressions have been carried out using individual SNPs or by cytological analysis (Badaeva et al., 2008; Przewieslik-Allen et al., 2021). Here, we validated some of the historically important introgressions for wheat breeding using IBSpy. Across this analysis we compared regions detected by IBSpy against other methods and describe novel introgressions detected in the WatSeq dataset. Additionally, we investigated a diverse set of wild ancestors and their contribution into the wheat genome (pangenome, modern wheats, and landraces). We interrogated the a) *T. timopheevii,* and b) *Ae. ventricosa* introgressions, and c) large deletions reported in (Przewieslik-Allen et al., 2021; Winfield et al., 2018).

4.4.4.1. *T. timopheevii* introgressions

A large 427 Mbp size introgression was reported in cultivar LongReach Lancer in (Walkowiak et al., 2020) (cultivar Lancer hereafter). This introgression carries the stem rust resistance gene *Sr36* from *T. timopheevii.* Lancer also carries the 60 Mbp *Lr24* (leaf rust) and *Sr24* (stem rust) introgression derived from *Thinopyrum ponticum* on chr3D. Since these two introgressions are historically important for wheat breeding to provide a wide resistance against the Ug99, in this analysis we extended the analysis to identify additional carriers of those fragments from the WatSeq dataset. In our results, we identified genotypes having different blocks sizes that would suggest that breeders are actively selecting for it. Our results with IBSpy detected the boundaries at 50 Kbp resolution in different modern genotypes. This will allow a more efficient use of the current available germplasm in future line development. We used the publicly available data of *T. timophevii* accession from (Walkowiak et al., 2020)

In our analysis, we first examined the chr2B introgression of *T. timopheevii* present in the Lancer pangenome reference using IBSpy *variations* count. We first compared the *variations* counts of *T. timopheevii* against Lancer and found values >120 for both distal ends of chr2B (**Fig 4.16**, >120 in grey). However, between ~93 Mbp to 626 Mbp we found that *T. timopheevii* had variation counts below <30, which is consistent with the presence of a *T. timopheevii* introgression in Lancer (**Fig 4.16**, orange) reported in Walkowiak et al., 2020. To further support the introgression site, we compared additional modern cultivars which do not carry the *T. timopheevii* segment introgression. As expected, these accessions had *variation* count overs >120 within the ~93 Mbp to 626 Mbp interval. With these results we validated the boundaries of the *T. timopheevii* introgression in Lancer. Then, we searched for additional genotypes in the WatSeq dataset to have the fragment and found lines having *variations* <30 count in the introgression interval. One of the accessions named Diablo had the largest introgressed block having *variations* count <30 from 403 to 608 Mbp. A second cultivar was Stava, which had the introgression from 509 to 608 Mbp. These two genotypes had the largest *T. timopheevii* introgression after the reference Lancer not reported before. Other cultivars had smaller introgressed blocks. One of them was Moisson which had the shortest introgression (**Fig. 4.16a,** purple box). Interestingly, we identified several UK cultivars which had identical fragments of *T. timopheevii* (e.g., Riband, Malacca, Cordiale and Crusoe) (**Fig. 4.16a,** blue box). When verifying their relatedness, consistently, we found that they share a common pedigree (**Fig. 4.16b**). Grafton, which is also within this pedigree, had a recombination event which removed part of the *T. timopheevii* introgression, as indicated by the shorter orange segment with respect to its parent Cordiale and sibling Crusoe (**Fig. 4.16a,** green box). This block was also present in Julius, a result which was previously unreported. We demonstrated with these cultivars that IBSpy can detect introgressions at 50 Kbp resolution (**Fig. 4.16c**). Our results also suggest that breeders are actively selecting for this fragment.

Table 4. 3. *T. timophevii* accessions used in this analysis from (Walkowiak et al., 2020).

Accession Number	Reads	Publication
Triticum timopheevii 33255	SRR13484808	Walkowiak *et al.*, 2020
Triticum timopheevii 15832	SRR13484807	Walkowiak *et al.*, 2020
Triticum timopheevii 10728	SRR13484803	Walkowiak *et al.*, 2020
Triticum timopheevii 10558	SRR13484817	Walkowiak *et al.*, 2020
Triticum timopheevii 22438	SRR13484809	Walkowiak *et al.*, 2020
Triticum timopheevii 14352	SRR13484806	Walkowiak *et al.*, 2020
Triticum timopheevii 3708	SRR13484804	Walkowiak *et al.*, 2020
Triticum timopheevii 10827	SRR13484805	Walkowiak *et al.*, 2020
Triticum timopheevii 17024	SRR13484818	Walkowiak *et al.*, 2020

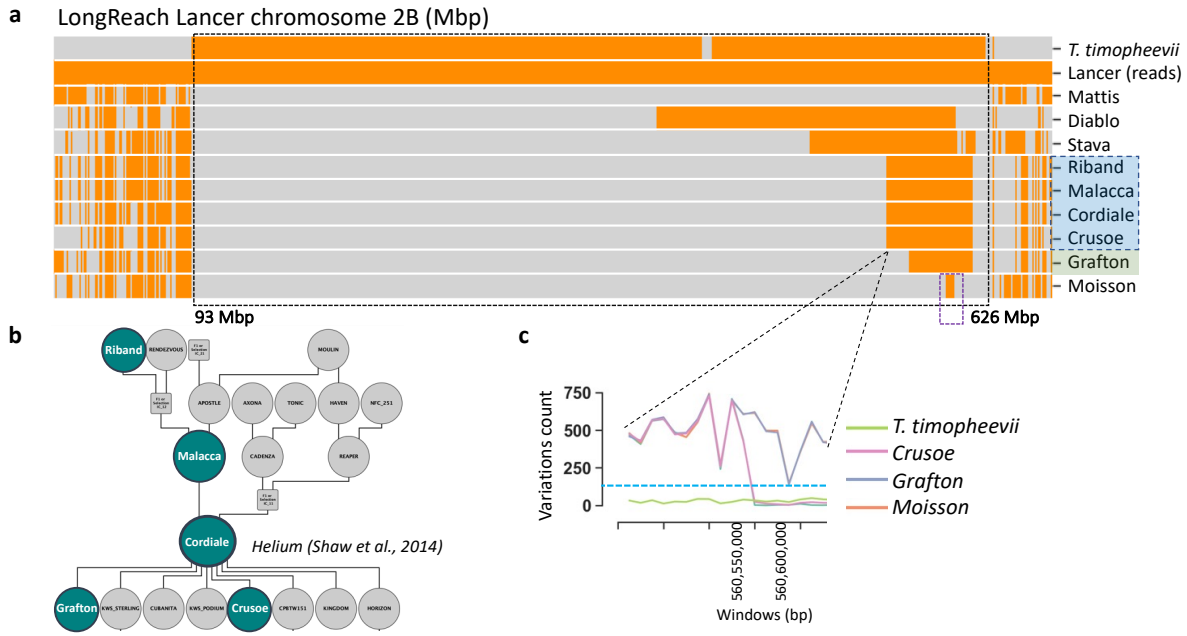


Fig. 4. 16. *T. timopheevii* introgression into wheat and WatSeq genotypes.

a) In orange IBSpy *variations* count <120 using Lancer as a reference indicating the introgression a region of chr2B. The square box indicates the introgression boundaries chr2B (~93 Mbp to 626 Mbp). Eight modern wheat cultivars are shown as having different introgression block sizes. **b**) Pedigree relationship of five genotypes with the introgression and showing Grafton with a reduced version of the block. **c**) Zoom in of the *variations* count depicting the exact 50 Kbp windows where the introgression starts in Crusoe as it becomes <120 *variations* count to the reference Lancer. The horizontal blue line indicates 120 *variations* count. In this fragment of the reference Lancer in chr2B (~93 Mbp to 626 Mbp), the *Sr36* gene is located as introgressed from the *T. timopheevii*. Orange colours outside

the introgression boundaries suggest old hybridizations from other wild wheat relatives either, in the reference Lancer or in the modern cultivars depicted.

4.4.4.2. The *Ae. ventricosa* 2AS/2N^{v}S translocation into the WatSeq

The 2AS/2N^{v}S translocation from *Ae. ventricosa* was introduced originally by (Doussinault et al., 1983) into the French cultivar VPM1. It has been documented that the 2AS/2N^{v}S introgression is beneficial for yield in CIMMYT and Kansas State (US) breeding programs (Gao et al., 2021) and it is present in several European modern cultivars. In this analysis we report further characterisation on the presence and frequency of the *Ae. ventricosa* introgression in the WatSeq modern GediFlux germplasm collection. We investigate the different translocations sizes reported in previous studies in the eleven pangenome assemblies and provide detailed introgression boundaries and discuss the absence of different blocks sizes present in the germplasm here used. Furthermore, we reported additional chromosome blocks with putative *Ae. ventricosa* introgressions hypothesized to be originally from VPM1 and that have been maintained in multiple modern cultivars. Some of these additional translocations have been reported before but the actual donor or 50 Kbp resolution region was unknown.

In our dataset we incorporated six *Ae. ventricosa* accessions which includes the *ventricosaCGB116981* genotype, which we hypothesize is the actual donor of the VPM1 cross (Table 4.4). For this analysis we used Mattis as our reference for the IBSpy analysis as it is well known to carry the *Ae. ventricosa* 2AS/2N^{v}S introgression on chr2A. Using a similar criterion, we confirmed that, *Ae. ventricosa* has *variations* count <120 against the Mattis reference from 0 to 32.6 Mp on chr2A. This same introgression was additionally confirmed in Mace, Stanley, and Jagger references. Using these four references with the detected *Ae. ventricosa* introgression and IBSpy, we identified in total 34 genotypes in the WatSeq dataset to have the 2AS/2N^{v}S introgression. Interestingly, all samples with the presence of the 2AS/2N^{v}S block had the complete region without apparently recombination from 0 - 32.6, 0 -33, 0 - 33.6, 0 - 31.8 Mbp based on Jagger, Mace, Stanley and Mattis, respectively. Four genotypes with CIMMYT origin (Becard_Kachu, Borlaug, Cimcog26, Cimcog49) had the introgression.

Table 4. 4. *Ae. ventricosa* accessions used in this analysis.

Accession Number	Reads	Publication
Aegilops ventricosa 2067	SRR13484802	Walkowiak et al., 2020
Aegilops ventricosa 2181	SRR13484816	Walkowiak et al., 2020
Aegilops ventricosa 2210	SRR13484813	Walkowiak et al., 2020

Aegilops ventricosa 2211	SRR13484815	Walkowiak et al., 2020
Aegilops ventricosa 2234	SRR13484814	Walkowiak et al., 2020
Aegilops ventricosa CGB116981	ERR7747980	Aury *et al.,* 2022

Consistently, most of the genotypes with the *Ae. ventricosa* 2AS/2N^{v}S introgression were related by pedigree and can be classified as descendants of VPM1 (http://wheatpedigree.net). Analysis in the recent modern cultivars from private breeding companies from the UK and Europe, we detected the 2AS/2N^{v}S introgression on the cultivars Extase, Wolverine, Piko, Rubisko, Sacramento, Saki, Skyscraper, Siskin, Santiago, and Revelation which reveals its extensive current use on modern European breeding. Consistent with our previous observations, the 2AS/2N^{v}S in these modern cultivars had the same introgressions size regardless of the reference used (**Table. 4.5**). These results suggest that there is a single introgression that has been maintained in all the cultivars and that the slightly different sizes seen in the genome references may be due to misassembles. This is supported by the additional fragments detected on the chrUn in Borlaug and Jagger similar to *ventricosaCGB116981* (not shown). Therefore, the precise introgression size may be determined once an *Ae. ventricosa* assembly is released. All this analysis, suggest that there is a low (or null) recombination in this region or that breeders are actively selecting for the complete segment for its multiple agronomic benefits. Unexpectedly, two Watkins accessions (WATDE0786 and WATDE0791) also carried this introgression. These may be due to an unintended cross pollination during seed propagation, DNA contamination, or mislabelling of samples. This because seeds of the Watkins collection have been propagated during almost ~100 years in close contact with other modern wheats.

Additional to the 2AS/2N^{v}S introgression, we identified seven genotypes to have an *Ae. ventricosa* fragment on chr7D from 616.0 - 618.6 Mbp (based on Jagger genome coordinates) and from 631.1 – 635 Mbp (based on Mattis coordinates). This region corresponds to the well-known eyespot *Pch1* gene region (Pasquariello et al., 2017). Although recombination in the region can rarely occur (Pasquariello et al., 2020), breeders have maintained the region intact in multiple cultivars including Jagger, Holster, Piko, Rendezvous, Revelation, and Renan (**Table. 4.5, Fig 4.17b**).

Other additional block was detected on chr2B from 7.1 - 8.8 Mbp based on the Jagger reference in the same set of six genotypes. These two blocks detected in two different chromosomes in the same set of samples may be an indication of linkage disequilibrium between the two blocks. Alternatively, it could be that the region on chr2B belongs to chr7D in Jagger or vice versa and is therefore a misassemble.

Previous studies detected a putative introgression on the telomeric region of the short arm on chr3D in the Mattis reference, but its origin was unclear. In this analysis, we detected this introgression and could assign it to *Ae. ventricosa* based on the raw data of *Ae. ventricosa* having IBSpy *variations* values below <10 in the 0 Mbp to 47.8 Mbp interval which suggest that *ventricosaCGB116981* is the actual donor (or a closely related) accession of the introgression. (Table 4.5, Fig 4.17f). Using Mattis as the reference, we also detected a small block of this introgression from 42.5 to 47.8 Mbp in modern cultivar Extase (Table 4.5, Fig. 4.17c).

Table 4.5 Introgressions from *Ae. ventricosa* (*ventricosaCGB116981*) into the wheat pangenome and the WatSeq modern lines.

Reference	Jagger			Mace	Stanley	Mattis		
Chromosome	chr2A	chr2B	chr7D	chr2A	chr2A	chr2A	chr3D	chr7D
Genotype								
Alert	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Becard_Kachu	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
borlaug-pg	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Cimcog26	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Cimcog49	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Cordiale	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Crusoe	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Drake	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Equinox	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Extase	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp	42.5 - 47.8 Mbp	
Fiorello	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
holster	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		631.1 - 635 Mbp
jagger-pg	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
KWS_Santiago	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
mace-pg	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
mattis-pg	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp	0 - 47.8 Mbp	631.1 - 635 Mbp
Piko	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		631.1 - 635 Mbp
Prophet	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Reedling	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Reform	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Renan	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		631.1 - 635 Mbp
Rendezvous	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		631.1 - 635 Mbp
Revelation	0 - 32.6 Mbp	7.1 - 8.8 Mbp	616.0 - 618.6 Mbp	0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		631.1 - 635 Mbp
Rubisko	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Sacramento	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Saki	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Savannah	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Siskin	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Skyscraper	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Spitfire	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
stanley-pg	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Torfrida	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Turpin	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
ventricosa-10x_	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
WATDE0786	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
WATDE0791	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		
Wolverine	0 - 32.6 Mbp			0 - 33.0 Mbp	0 - 33.6 Mbp	0 - 31.8 Mbp		

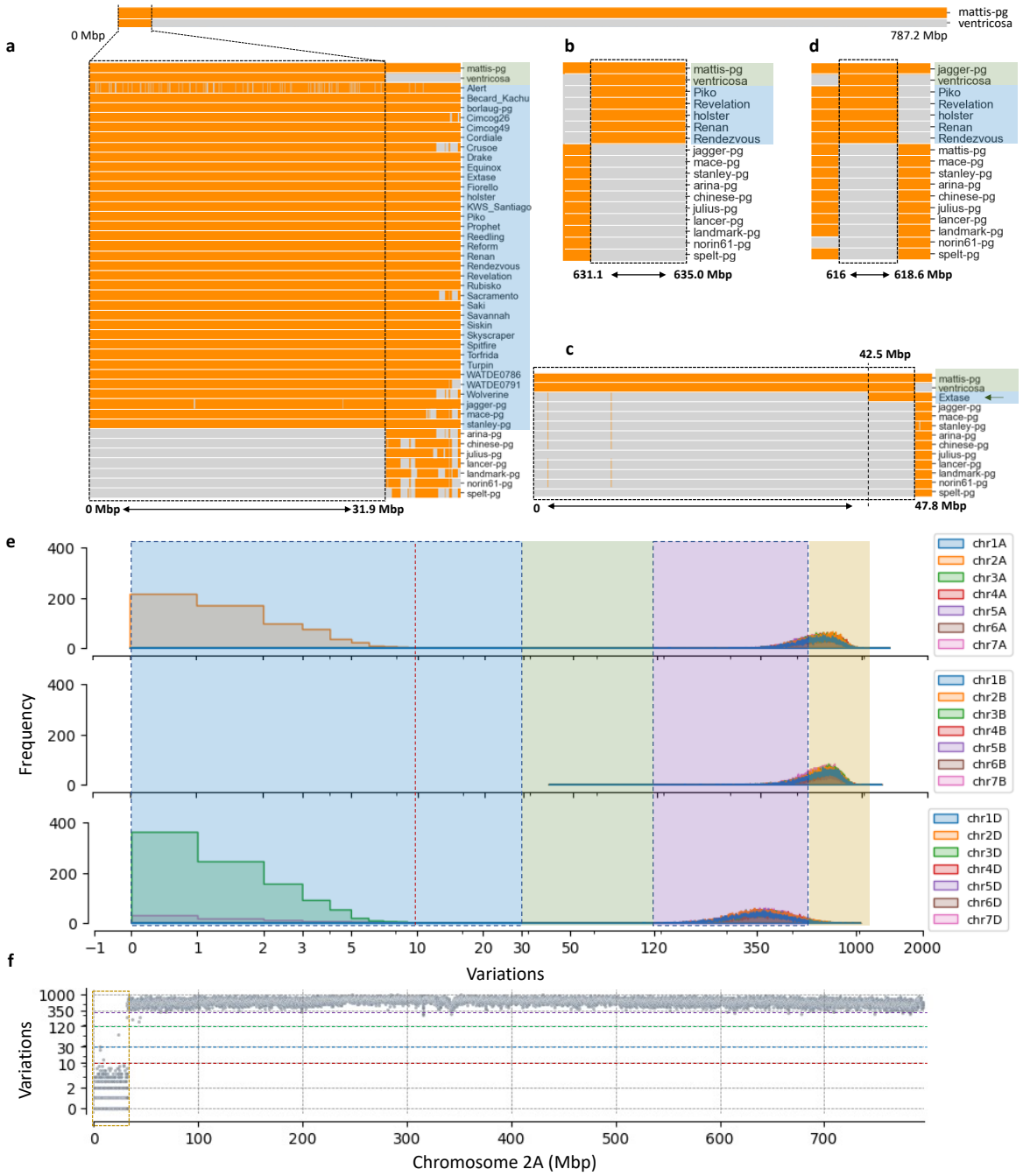


Fig. 4. 17. *Ae. ventricosa*, the donor of the 2AS/2N^{v}S introgression into wheat.

The reference Mattis has the introgression 2AS/2N^{v}S from 0 to 31.9 Mbp. The orange colours indicate <120 *variations* count of any other query sample against Mattis reference, which is an indication of having the introgression (DNA sequence similar to Mattis). **a**) 34 elite genotypes from the WatSeq dataset have the complete introgression block, including the references Jagger, Mace, and Stanley. **b**) A set of six genotypes have an additional introgression block on chr7D of Mattis (from 631.1 to 635 Mbp). **c**) Additional large introgression from *Ae. ventricosa* was detected in Mattis on ch3D from 0 to 47.8 Mbp and Extase (modern elite cultivar) having small portion of this introgression from 42.5 to 47.8 Mbp. **d**) The *Pch1* interval from Jagger on chr7D was also detected in six additional cultivars. These six

cultivars also have the fragment on chr7D of Mattis from 631.1 to 635 which indicates that these they may have the complete block and that was split in Mattis and Jagger. **e**) *variations* histogram distribution of *Ae. ventricosa* as a query against Mattis reference per sub genome indicating that only A and D sub genomes have introgression depicted by the *variations* count at ~<10 consistently with **a, b, c**. And **f**) *variations* count in 50 Kbp windows of reference Mattis vs *Ae. ventricosa* depicting very low variations (<10) indicative to be IBS region and therefore the donor accession of the introgression.

4.4.4.3. IBSpy on WatSeq large deletions

We next wanted to investigate how the known deletions previously identified in modern lines would emerge and differentiate from introgression with IBSpy. For this analysis we focused on the deletions reported by (Allen et al., 2017; Przewieslik-Allen et al., 2021; Winfield et al., 2018) and using the cultivars reported in literature to carry such deletions. In this analysis we queried the candidates for large deletions against the eleven pangenomes. However, as a case study we describe the results using the Julius reference as it shows remarkable IBS region against WATDE0117 discussed.

In our result using IBSpy we identified that Watkins_816 (in previous reports), here WATDE0117, has very high *variations* count against Julius across chr2D (**Fig. 4.18c, b, d)**. Based on the work of (Allen et al., 2017) and (Winfield et al., 2018), we interpret this as a whole chromosome deletion of chr2D in WATDE0117. However, we detected a small block at ~515 Mbp similar to Julius, which suggests that a remanent of the chromosome is present (**Fig. 4.18c,** green arrow). This could be due to a misassembly in Julius, a translocation in WATDE0117 (i.e., this chr2D region translocated elsewhere in WATDE0117 before the chromosome was lost) or it could be indicative of a highly repetitive region. The latter is unlikely since the region spans several 50 Kbp windows. Interestingly, we noticed that WATDE0117 is very different (high *variations* counts) across the whole genome compared to all pangenome cultivars except for chr2A (**Fig. 4.18a**), which is hypothesized to be a whole chromosome duplication according to Winfield *et al.,* 2018.

Additional results on the eleven pangenome assemblies suggest that chr2A of WATDE0117 is the only chromosome that shares IBS regions. Julius reference shares the largest with more than >50% IBS regions, a large block from ~160 to 535 Mbp, and from ~590 to 627 Mbp of chr2A (**Fig. 4.18a**, yellow bars). Norin61 and CS share mainly telomere regions, which was not seen in other pangenomes. This analysis reveals that only chr2A has been used in modern breeding based on the pangenome references. In future work, it would be worth it to explore additional Watkins to have the same haplotype on chr2A and therefore, a candidate duplication to extend the analysis in the phenotypic effect. In Winfield et al., 2018 there was an additional indication that

WATDE0117 may have other translocation in chr2B, however, our results with IBSpy do not show any evidence of a translocation in this chromosome (**Fig. 4.18b**).

An additional cultivar named CIMCOG29 was reported in Allen *et al.,* 2021 to have chromosome number variations detected by GISH. In our dataset we did not have this genotype, however, we identified CIMCOG32 to have a large deletion on chromosome arm chr7DL (**Fig. 4.18e**). This was the interpretation we gave to the region after 340 Mbp where *variation* counts are higher than >500 in most of the 50 Kbp windows (**Fig. 4.18f**) matching the deletion *variations* distributions detected in **Fig. 4.18d**. This genotype shares several large chromosome blocks with Mace, a cultivar from Australia, in other chromosomes. This is not unexpected since Australian germplasm has been influenced by CIMMYT lines. Those large blocks and similarity to Mace were seen in CIMMYT lines CIMCOG47 and CIMCOG56 (data not shown).

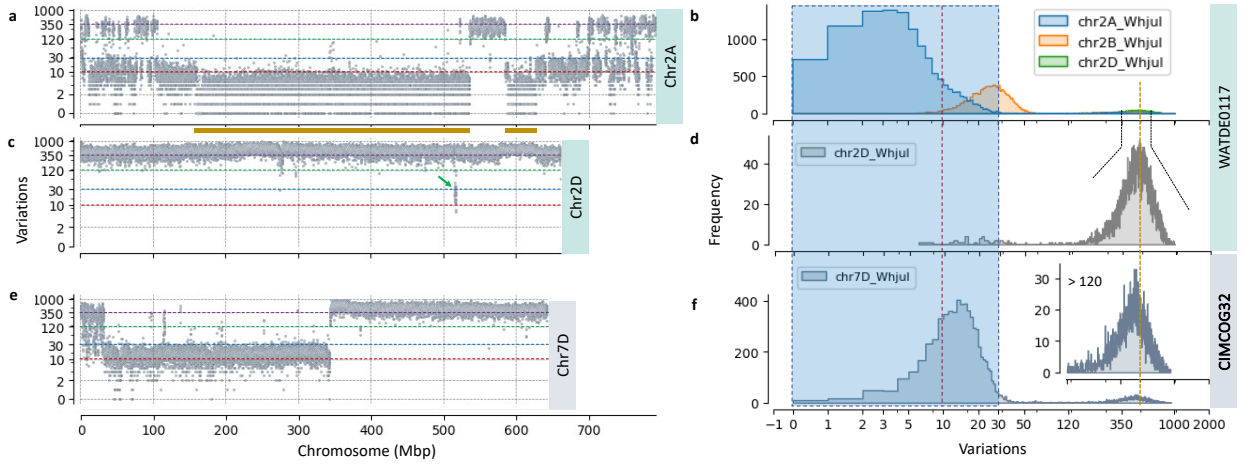


Fig. 4. 18. IBSpy detects large deletions.

a) chromosome 2A *variations* fingerprint of WATDE0117 vs Julius reference and **b**) the histogram distributions of the chr2 triad. **c**), chr2D of WATDE0117 vs Julius which has the entire block deleted with a remanent (green arrow). **d**), histogram distributions of chr2D deleted which has *variations* peak at ~500 count. **e**), chr7D of CIMCOG32, a cultivar from CIMMYT vs Julius. CIMCOG32 has a deletion of half chromosome from ~350 Mbp to the end of the chromosome. **f**), histogram distribution of chr7D of CIMCOG32 and a zoom in of the distribution filtering > 120 *variations* count. **d** and **f** *variations* peak are indicated by the yellow dashed line which falls at similar *variations* count. The red line indicates the threshold for the IBS hypothetical regions shared with WATDE0117 and Julius on chr2A (which was predicted to have a duplication in Allen et al., 2020). The blue square is the threshold defined for wheat cultivars not having an introgression.

4.5. Discussion

4.5.1. Methods to detect introgressions

Crop wild relatives are a valuable source of alleles still unexploited in many crops and breeding programs (Kilian et al., 2021). Interspecific or wide crosses have been used to transfer these alleles for desirable traits into crops mainly for disease or abiotic resistant phenotypes (Hao et al., 2020). Traditionally, these induced introgressions or natural hybridisations have been detected by cytological methods (Friebe et al., 1996) or C-banding patterns (Badaeva et al., 2008). Although valuable, cytological methods to detect introgressions are low throughput to screen large number of samples. Furthermore, the resolution and chromosome physical position of the introgressed blocks are low. They were employed mainly to detect presence absence at the scale of chromosome breakpoints blocks (Friebe et al., 1996). With progress on NGS, recent methods have been developed based on retrotransposon genome content analysis (Walkowiak et al., 2020) or by SNP-based methods (Przewieslik-Allen et al., 2021; Scholten et al., 2016).

Additionally, with the availability of high-quality assemblies of wheat and wild relatives, methods to detect introgressions without the need of cytological analysis have been developed. For example, Keilwagen et al., 2022 (Keilwagen et al., 2022) used mapping by alignments of sequencing reads to detect introgressions and predict the putative donors into descendant individuals. Similarly, (Keilwagen et al., 2019) used GBS and read coverage mapping to detect introgressions in barley and wheat. These methods rely on reference assemblies and the computer burden of alignments methods which represent a challenge for large genomes such as the hexaploid wheat (~16GB). An additional constrain with alignment-based approaches is the level of resolution that can be accomplished since the exact boundaries of the introgressions is mainly in magnitudes of Mbp in wheat. In our study, we used a *k*-mer based *variations* approach to detect introgressions at 50 Kbp resolution that does not rely on the alignments of sequencing reads. Instead, we used presence absence of *k*-mers which have been proven to be advantageous to detect variations on highly diverse genotypes because do not rely on read mapping to a genome reference and integrate genome structural variations information to the analysis (Gaurav et al., 2022; Rahman et al., 2018; Voichek & Weigel, 2020).

Furthermore, with our method, we could detect the actual donor, as we demonstrated with an example of the *Ae. ventricosa* in four of the pangenome references by detecting an IBS region with >99.99% sequence similarity (**Fig. 4.17f**). Using the chromosome-scale assemblies as an indirect scaffold, we detected 34 WatSeq genotypes having this exact introgression. Surprisingly, none of the genotypes having the introgression had a difference on the size of the introgressed fragment. Traditionally, this would be explained by the lack of recombination between wheat and *Ae.*

ventricosa. However, studies of the *Ae. ventricosa Pch1* region have shown categorically that recombination is possible between these genomes, albeit it is rare (Pasquariello et al., 2020). The fact that recombination is possible, and that the *Ae. ventricosa* introgression is located in the telomere region, a highly recombinogenic distal region of the chromosome, suggests that breeders are selecting the entire block, possibly, for multiple beneficial QTLs in the region. This is consistently with the analysis of the 2AS/2N^{v}S translocation present in multiple germplasms reported by (Gao et al., 2021) and shown to be beneficial for yield.

A different approach using low sequencing coverage was implemented by (Zhou et al., 2021) detecting introgressions by *k*-mers presence/absence of the parents inherited to the progeny with a subsequent step of mapping back to a genome reference to define the introgressed region. Interestingly, this method could detect introgressions at 0.1-fold coverage, but it relies on having the information of the actual donor parents of the progeny. Similarly, using *k*-mers presence/absence, (Gaurav et al., 2022) identified specific lineage regions into the D sub genome. With the current version of our approach, we can efficiently detect introgression at 50 Kbp resolution. However, if the actual donor(s) of the introgression is unknown, we still rely on ~10-fold coverage to detect it. With coverage of ~5-fold with Illumina 150 bp short reads we can detect the signal of an introgression with the constrain of not being able to detect the actual donor.

An importantly feature in our approach is that we can identify introgressions without the need of the assembly either from the donor or the accession having the introgression. Instead, we accomplish this by using a clustering approach and haplotype calls as demonstrated with an example from the *Ae. tauschii* into wheat in section **4.2.3.** This is an important step forward as all previous methods required to have a reference assembly with the putative introgression to be able to identify it. The haplotype-based approach outlined here will allow us to systematically identify regions where wheat accessions with only raw-reads cluster together with wild wheat relatives, and therefore candidate introgressions. The current resolution is 1 Mbp, but as outlined in **Chapter 3**, the AP pipeline can be adapted for smaller intervals.

The identification of novel *Ae. tauschii* segments in Watkins accessions also provides a roadmap for future crosses to incorporate this genetic variation into modern cultivars. We identified ten Watkins accession with a novel 99 Mbp from Lineage L3 region which is absent in modern cultivars. This region was maintained in the landraces but perhaps did not participate in the initial crosses performed by breeders in the early 1900s which were the founders of important modern cultivars. Likewise, we identified two Watkins accessions which carry putative novel lineage L2-SA regions which are absent from modern wheat. The ability to detect this novel variation in the absence of

a genome reference is an important step towards the targeted use of variation in breeding programmes.

4.5.2. The evolution of the D genome of hexaploid wheat

Historical and archaeological evidence suggest that wheat was domesticated in the Fertile Crescent (Brown et al., 2009; Tanno & Willcox, 2006). Most studies point to *Ae. tauschii* as the donor of the D sub genome of hexaploid wheat by a few events of independent of allopolyploidization 8,000 to ~10,000 years ago coinciding with the period of wheat domestication (Gaurav et al., 2022; Huang et al., 2002; MCFADDEN & SEARS, 1946; Pont et al., 2019) and the absence of historical records of wild hexaploid wheat accessions.

Studies suggest that only a few accessions contributed to the D genome of wheat from a specific *Ae. tauschii* lineage from the North of Iran (Lineage 2). In 2021, (Zhou et al., 2021), suggested that a reduced number of accessions from a L2 sub-group contributed the most to the D genome of wheat based on genetic similarity analysis. In our analysis we found similar results where accessions from a subgroup of L2, here named L2-SA, had the most similarity with the D genome of wheat. Other studies suggested that lineage L3 also contributed to the modern D wheat genome (Gaurav et al., 2021). Although to date there is no report of the actual donor and the number of hybridisations that took place, based on our results on novel introgressions blocks detected, we concluded that a single event may not explain the whole D sub genome variation observed. This is supported by the evidence of large L3 blocks on chr1D and other L2 introgressed blocks with different levels of *variations* in some Watkins landraces.

A particular observation from the analysis and level of *variations* among L1, L2, and L3 *Ae. tauschii* in this analysis detected with IBSpy suggest that the tree lineages are genetically equally distant among them. Interestingly, this similarity distance of the *variations* was similar to the observed in two heterotic groups in Maize. In future analysis it would be worth to test if controlled crosses among those L1, L2, and L3 lineages provides some level of hybrid vigour for agronomically important traits. Most importantly, the synthetics hexaploid already available (Gaurav et al., 2022) for the three lineages lines are good candidates to test the proposed hypothesis and detect the benefit (if some) in a very close to a breeding line in wheat. For example, we could test the above hypothesis using the following groups focusing on a hybrid vigour for yield or other traits:

a) cross a synthetic (L1) x synthetic (L2) and test for hybrid vigour.
b) cross a synthetic (L1) x synthetic (L3) and test for hybrid vigour.
c) cross a synthetic (L2) x synthetic (L3) and test for hybrid vigour.

4.5.3. The contribution of *T. monococcum* to the wheat A genome

T. urartu is the A sub genome donor of tetraploid and hexaploid wheat (Huang et al., 2002). The second closest relative of *T. urartu* and the A sub genome of wheat is *T. monococcum,* also known as einkorn. There is evidence that einkorn has contributed to the genome of wheat genome by natural or induced hybridisations (Kolmer et al., 2010; Zhang et al., 2010). However, until today, there is a lack of an extensive study at the whole genome or population level that demonstrates the extent to which the genome regions and gene flow were incorporated into modern wheat. Furthermore, there is a limitation on defining the exact boundaries of the hybridization regions in the physical chromosome positions that could help to design molecular markers for beneficial alleles for breeding.

In 2021, with the availability of the wheat pangenome, (Chen et al., 2021) found introgressions from different accessions of *T. monoccocum* to be present in the ArinaLrFor and Mattis genome assemblies on chr5AL with both genomes having a 9.5 Mbp fragment. The regions detected were from 700.7 Mbp to the end of the chromosome in ArinaLrFor and from 693.1 Mbp to the end of the chromosome in Mattis. In our analysis we demonstrated that a more precise genome region size would be of 10,124,532 bp in ArinaLrFor, and 9,606,209 bp in Mattis.

In agreement with this study of different fragments lengths of *T. monoccocum* being present in wheat, we suggest that an initial hybridisation took place with a *T. monococcum* donor that is now absent from the natural population or was not included in the collection here studied. Hence, cross-pollination followed by recombination among different *T. monococcum* accessions with the original donor took place and we detect those fragmented blocks in our study, but without the actual donor sharing the entire region. This is in consistency with the fact that (i) single donors could be detected in a few cases where the *T. monococcum* introgressions were relatively small, and (ii) the *T. monococcum* introgressions where concentrated in the telomeric regions which have higher recombination rates in Triticeae genomes (Choulet et al., 2014). Furthermore, the lack of fragments in centromere regions in most of the pangenome here explored suggest that a few hybridisations took place, and those genome regions may have been selected against due to undesirable traits in modern breeding or by natural selection affecting the fitness of those genotypes.

The domestication of einkorn was from wild members from a specific race named beta in Kilian et al., 2007 (Kilian et al., 2007). In our study, we propose that the gene pool from *T. monococcum* into wheat was brought from wild accessions into domesticated einkorn and from cocultivation of domesticated einkorn and wheat into the A wheat sub genome. This is supported by the number

of accessions sharing large proportions of blocks coinciding with the domestication region and cultivation of wheat (Balfourier et al., 2019; Brown et al., 2009; Marcussen et al., 2014) where early farmers may have selected for superior agronomically traits.

In our analysis, we identified ~0.2 to 0.5% of the wheat genome as *T. monococcum*. Then say that using the <120 cut-off about 30-35% of the genome are predicted to be introgressions. Some of these can now be attributed to *T. monococcum* (small 0.2 – 0.5%), and a few additional regions to *T. timopheevii* (chr2B) and *Ae. ventricosa* (Chr2A, chr3D, chr7D). As we showed in our section 4.4.2, we now can identify introgression that are not present in the genome references with the AP method and assign additional regions that come from wild relatives into the wheat genome. Therefore, we can predict introgressions into the complete WaSeq panel or any wheat accession with ~12-fold coverage not included here. This will be of importance since we will not only rely on pangenome assemblies to have the introgression.

In this analysis we explored the *Yr34* gene (synonym *Yr48*), which confers resistance to the yellow rust pathogen and was identified in the "Mediterranean" landrace from *T. monoccocum* into chromosome 5AL. This *T. monococcum* segment is present in several European cultivars, including the pangenome cultivars Arina*LrFor* and Mattis, and is hypothesised be originated from a single hybridisation event (Chen et al., 2021). *T. monoccocum* historically has been used for introgressions into wheat for agronomic traits. For example, an isogenic line of Thatcher (hexaploid wheat) named RL6137, received a translocation from *T. monococcum* into chromosome arm 3AS which confers leaf rust resistance (*Lr63*) (Kolmer et al., 2010). The relevance of *T. monococcum* is further exemplified by the identification of the stem rust resistance gene *Sr35* introgressed into wheat on chr3A which confers resistance against the Ug99 race of the stem rust pathogen (Zhang et al., 2010) (Saintenac et al., 2013). These examples document the relevance of *T. monococcum* natural genetic variation for modern wheat breeding.

4.5.4. The contribution of large introgressions into wheat

Modern wheat has a reduced genetic diversity compared to its wild relative ancestors. The hybridisation and exchange of genetic material of species that share highly similar genomes its possible (Badaeva et al., 2008). Wild relatives have been used in wheat since early breeding mainly to introduce disease resistant traits and several breeding programs have been benefited from these early induced and natural occurred introgressions. An example is the well-known *Ae. ventricosa* 2AS/2N^{v}S translocation on chr2A and the large fragment from *T. timopheevii* on chr2B Lancer pangenome.

The D genome of the tetraploid species *Ae. ventricosa Tausch* (DvDvNvNv) and the D sub genome of wheat pairs successfully and recombination and exchange of genetic material is possible in hybrids of these species (Badaeva et al., 2008; Gao et al., 2021). On the contrary, hybridisation among the Nv genome of *Ae. ventricosa* and the A wheat genome is less frequent. Different genes from *Aegilops* species have been introduced in wheat background (Doussinault et al., 1983; Friebe et al., 1996; Gale et al., 1984). For example, the 2AS/2N^{v}S was introduced in 1967 by Nicole Maìa and René Ecochard (Gao et al., 2021) from the *Ae. ventricosa*, accession Vent10 into a line named VPM-1. The genes *Yr17*, *Lr37*, and *Sr38* were mapped in the 2N and were demonstrated to be derived from *Ae. ventricosa* into wheat (Bariana & McIntosh, 1994).

In our analysis presented in this thesis we successfully detected and validated those large introgressions using IBSpy and discovered novel fragments not reported before. We detected the almost complete chromosome introgression from *T. timopheevii* into the pangenome Lancer chr2B reported by (Walkowiak et al., 2020). In addition, we detected the boundaries at 50 Kbp resolution of the introgression and detected large blocks present in Diablo an Stava cultivars. Several other modern cultivars had different fragment sizes of the introgression that suggest breeders are selecting for this block. Similarly, we detected the 2AS/2N^{v}S in more than >30 cultivars from the WatSeq panel and evidenced that the complete fragment has been passed intact in modern cultivars maybe by targeted selection.

In summary, our analysis here presented revealed the prevalence of natural and induced introgressions into the wheat genome. We evidenced that the wheat genome harbours large portions of introgressions across its entire genome still unexplored and that novel sequenced cultivars outside the pangenome references will reveal those alien fragments extensively. With these series of analysis, we demonstrate the usefulness of IBSpy to detect difficult introgressions at 50 Kbp resolution in wheat from distant wild relatives. We hope that our approach would provide the tools to further explore the wheat genome and other important crops and impact on breeding decisions for trait selection.

5. General discussion

The main objective of this PhD was to develop a method to build a haplotype database for wheat and elucidate the genetic diversity between cultivars, modern varieties, landraces, and associate genotypes with phenotypes. We developed a novel method based on *k*-mers presence/absence to detect genetic variations represented in three scores: "*observed_kmers*", "*variations*", and "*kmer_distance*". Since we use those scores to predict identical by state (IBS) genome regions among cultivars using mainly the python programming language, we named it IBSpy. Using the

variations score we extended IBSpy to define haplotypes using multi-genome references syntenic regions. Exploiting these haplotypes, we tracked back haplotypes in modern elite varieties from early cultivars and landraces. Using IBSpy haplotypes, we implemented hapGWAS to detect genome associations to qualitative and quantitative phenotypes. Finally, in Chapter 4, we exploited IBSpy *variations* fingerprint counts, and haplotype calls to detected whole genome introgressions at 50 Kbp resolution and define novel introgressions absent in the current genome reference assemblies available.

Specifically, this research focused on the challenge of identifying IBS regions using whole genome sequencing raw reads data at relatively low coverage (~12-fold) and automate haplotype calling of genome regions among >1,000 wheat genotypes. Using this haplotype database, we addressed our initial question of how genetic diversity has been changed in modern elite cultivars to target and incorporate those novel alleles into breeding. We benchmarked our pipeline with commonly used single-reference alignment and variant calling methods. Our approach complements alignment methods already established to explore genome diversity in population genomics and for genotype-phenotype associations studies. Particularly, our approach provides an alternative to exploit large genome datasets which are computationally challenging and time demanding. Our method benefits of integrating multi-genome information and large collections datasets for large genome (~16 Gb).

5.1. Challenges on variations discovery

Before the 1980s molecular markers were private to protein polymorphisms detected mainly by gel electrophoresis (Gottlieb, 1981). In this same decade Jeffreys made it possible to detect polymorphisms at the genome level through the discovery of DNA fingerprints (Jeffreys et al., 1985). Polymorphisms were distinguished mainly by the variation in DNA fragment lengths cut by restriction enzymes (restriction fragment length polymorphism; RFLP) (Botstein et al., 1980; Cooper & Schmidtke, 1984). Different versions of RFLPs were rapidly developed after its first discovery and although these methods are now considered low throughput genotyping, at that time, human fingerprint using those techniques had a huge impact for different applications on DNA-based discoveries.

These methods to detect polymorphism rapidly reached their use in plants to fingerprint varietal identity of crop varieties (Smith & Smith, 1992). As the advent on DNA sequenced progressed, methods such as DNA-arrays and sequencing of short DNA fragments started to emerge (Kehoe et al., 1999). In the last two decades however, NGS has considerably impacted on the discovery of

DNA markers and genotyping technologies at high scale. This has led to accumulation of large amounts of sequencing data in publicly available databases and to the capacity of generate large volumes of data in a short period of time (Fan et al., 2014).

As the amount of data continue to increase with NGS advances, novel algorithms are released to detect SNPs and structural variations. The software to use in each case depends on the data available, computer resources, and the objective of the study. The most common methods are based on alignments of short reads to a reference genome (Chiang et al., 2015; DePristo et al., 2011) and more recently to a pangenome graphs (Jordan et al., 2021; Kim et al., 2019; Rakocevic et al., 2019).

In this project we took advantage of NGS and publicly available data to explore genetic variations in wheat. The challenge of this study was on how to integrate and develop a method to call variations in a unified manner bearing in mind the wheat genome size (~16 Gbp) to avoid computer burden. Considering that at the moment of this study there were eleven chromosome-scale, and five scaffold scale assemblies generated by the pangenome project (Walkowiak et al., 2020) and additional five hexaploid wheat chromosome-scale at different quality assemblies generated by other groups (Athiyannan et al., 2022; Aury et al., 2022; Guo et al., 2020; Kale et al., 2022; Sato et al., 2021), the computational load to exploit and integrate all the information available was becoming prohibited.

In several species including wheat, read alignments studies for SNP calling usually uses ~ >10-fold sequencing coverage. In this study, the WatSeq project data has on average ~12-fold coverage of 150 bp DNBSeq reads, and more than > 1,000 genotypes were sequenced in the collection. To align a single sample to a single reference of hexaploid wheat (16 Gb) with this coverage roughly requires 96 CPU hours, ~120 Gb RAM, and 1 CPU using the most common aligners (Langmead et al., 2009; Li & Durbin, 2009) depending on the computing infrastructure. In our study, we implemented and used IBSpy to detect variations. This method calls variations by using *k*-mer presence/absence. It requires ~48 CPU hours, and 60 RAM, and 1 CPU to detect variations for a single sample against one genome reference for the hexaploid wheat genome (16 Gbp size).

Our method is simplified into two steps, first creating a *k*-mer databases directly from raw reads using either KMC or Jellyfish with a following step for querying this *k*-mers databases to a reference through IBSpy. These *variations* are the input to call haplotypes in downstream analysis optional to the user as described in Chapter 3. Compared with alignment methods, our approach simplifies the pipeline avoiding pre filtering steps such as cleaning and removing low quality reads. In downstream analysis, it avoids processing large BAM files and calling variations followed by a filtering step commonly required in alignment methods (McKenna et al., 2010).

Although IBSpy can detect *variations* in 50 Kbp windows, we acknowledge that the current version of IBSpy does not allow to detect variations at single or a few base pairs between two samples. While this may not be a constrain to detect IBS regions and long-range haplotype blocks at >1 Mbp, it will be a limitation if a user is interested in detecting a single SNPs or point mutations between highly similar genotypes. This limitation, however, can be used as a useful feature in haplotypes GWAS analysis as we described in Chapter 3. For example, building haplotypes can integrate a group of genotypes having the same long-range haplotype but contrasting phenotypes. This could be used to detect the putative polymorphism underlying the phenotype within the group of genotypes having the same haplotype. For example, if a set of ten lines having the same haplotype are associated with the resistant phenotypes, but out of ten, nine are resistant and one is susceptible it could provide us with clues of the mutation underling the trait as we would expect that the susceptible line should have a few SNPs different to the other nine resistant genotypes.

It is well known that large scale structural variations are important for agronomically important traits (Yang et al., 2019) and played essential roles during crop domestication (Gaut et al., 2018; Zhou et al., 2019). Although SNPs and small InDels are easily captured by common aligners, variant callers, and shotgun sequencing (Hwang et al., 2015), a constrain with alignment and variation calling methods is that they fail to efficiently capture genomic structural variations. One way to face this problem is using long read sequencing or specific aligners which capture better large structural and copy number variations (Eggertsson et al., 2019; Handsaker et al., 2015). On the other hand, large structural variations, duplications, and large inversions are detectable only by chromosome assemblies or optical maps (Mahmoud et al., 2019; Schiessl et al., 2019).

With the advent of genome sequencing and pangenome assemblies, methods to detect structural variations are becoming common (Sibbesen et al., 2018). However, still most of them rely on alignments methods of short reads and were developed focused on other species rather than plants, which harbour a high repertoire of structural variations. An alternative to detect structural variation is to perform local alignments of shorth reads with a subsequent step of aligning them back to a reference. However, assembling raw reads is computational demanding, requires high sequencing depth, and is laborious when the number of samples is high such as those commonly used in plants studies. One option to capture and incorporate structural variations is by using *k*-mers (Gaurav et al., 2022; Voichek & Weigel, 2020).

With the deployment on sequencing at low cost, novel methods to detect variations based on *k*-mers are in progress. For example, BayesTyper uses the exact match of alignments of *k*-mers using a graph representation of a variant to detect all types of variations (Sibbesen et al 2018). In our study, we developed a novel method based on *k*-mers to detect variations represented in three

scores: "*observed_kmers*", "*variations*", and "*kmer_distance*" and we called it IBSpy. Our method integrates all types of genetic variations into a single score. In this way, when calling haplotypes and performing hapGWAS, this information is incorporated to capture similar haplotypes and detect phenotypic associations. The disadvantage, however, is that the IBSpy "*variations*" score considers a deletion, insertion, or copy number, as a "single" variation regardless of the size of the InDel or the duplicated fragment. This limitation can be partially overcome if the "*observed_kmers*" score is used instead to call haplotypes. However, in this thesis we did not explore if there are differences when using hapGWAS based on the former score. Further analysis in this direction would be worth to test for further IBSpy optimization.

5.2. IBSpy: a multi-genome approach to call haplotypes in wheat

With the progress on genome sequencing technologies, it is affordable to generate large sequencing information at relatively low cost. This large genomic data available, now faces the challenge on how to analyse and exploit it in a meaningful approach. Novel methods to integrate this information are in progress including pipelines that incorporate multi-genome references into a graph representation of a species for alignments. In our study we leveraged the availability of the recent wheat pangenome (Walkowiak et al., 2020) and > 1,000 whole genome resequencing of landraces and modern cultivars at ~12-fold coverage (here named WatSeq). This dataset is of magnitude that aligning the whole dataset against the complete wheat pangenome would be prohibited for the current computing resources and time using common aligners.

Furthermore, large portion of the WatSeq dataset are landrace genotypes which may diverge from the current genome references and deciding to use only one or a subset of the current genome references could impact in not capturing important genomic features in these landraces not present in modern cultivars. We hypothesized that these set of lines (landraces) harbour untapped genetic diversity and unique haplotypes that are worth to explore and incorporate in a compressive analysis. To embrace these challenges, we implemented a feature into IBSpy to define haplotypes integrating a multi-reference approach and using presence/absence *k*-mers instead of the routine alignment methods.

Currently, there are different approaches to define haplotypes. For example, HaploBlocker uses linkage to infer haplotypes. In other words, it defines haplotypes as a sequence of markers at a predefined minimum frequency in the population. With this method, only haplotypes with similar consecutive sequence of markers are consider to belong to the same haplotype block (Pook et al., 2019). MATILDE, a second approach uses LD for clustering contiguous SNPs (Pattaro et al., 2008).

Similarly, (Kim et al., 2018) created Big-LD which cluster LD SNPs not necessarily physically consecutive. As a result, Big-LD create larger haplotype blocks than other methods.

Depending on the species, genotyping data, computational resources, and the objective of a study, a specific method to define haplotypes may be preferred. For example, (Mayer et al., 2020) used fixed window size to define haplotypes. This would be advantageous to compare haplotypes across datasets that vary in their LD. The disadvantage, however, would be those haplotypes will be broken by the window size used and the complete haplotypes and its effect is not captured in haplotype-phenotype associations. In Chapter 2 we showed a method that uses *variations* count based on *k*-mers against multiple-genome references independently. In a followed-up step in Chapter 3, we use those *variations* counts and the syntenic regions from these multiple genome references and defined an arbitrary split window size to predict haplotypes based on clustering algorithms. These parameters are not fixed and can be modified accordingly user needs.

As briefly introduced before, with multiple reference now available (pangenomes) for several important crops, we can decide which reference from multiple options to use as a template to compare our samples. Alternatively, methods to unify multiple genomes into the so-called genome graphs or haplotype databases are available (Rakocevic et al., 2019). With these unified genome references, alignments and variant callers can exploit genome regions uniquely present in only some references and alleviate the problem of independently align to a set of references which is computing demanding and time consuming. Although this type of analysis is advanced in human genomics, great progress in several important crops is undergoing. The challenge, however, is that plant genomes are complex and repetitive, which represent a problem to solve in species with large genomes such as wheat.

In plants, the practical haplotype graph (PHG) uses haplotype representation of a pangenome (Bradbury et al., 2022). The PHG still relies on genome assemblies or WGS alignments reads to a reference to populate the initial haplotype database with a representation of enough samples. This is a constrain to align high number of samples against large genome sizes such as the wheat genome. Again, alignments to a single reference involves bias towards the reference. When using WGS data, the PGH only keeps sequences that aligns well to a single location on the reference and important information and large structural variations are missed and mostly gene content regions may be captured. Furthermore, to accomplish a good alignment, high coverage WGS is required. Once the PHG is generated skim sequence or SNPs can be used to impute haplotypes. In the haplotype database presented in our study, we have generated a haplotype database using *k*-mers with 12-fold raw reads for hexaploid wheat including landraces and modern cultivars against multiple genome assemblies. This step, would be the equivalent step in populating the PHG in the

(Bradbury et al., 2022) research with the difference, however, that the PHG uses a single genome reference. In our study, at the time of writing this thesis, we haven't reached the point to impute haplotypes with coverage below <10x. Further developments to the current version of IBSpy will take place on this direction to investigate how to overcome this challenge.

Additionally, to other crops, the PHG was applied to wheat using whole exome capture and a reduced number of 65 wheat accessions (Jordan et al., 2021). In this study of wheat, imputations of haplotypes using different datasets resulted in lower accuracy compared to sorghum (Jensen et al., 2020), maize (Franco et al., 2020), and cassava (Long et al., 2021). Given that the genome of wheat is much larger that these two species, polyploid, and highly repetitive future tests would be required using WGS or the wheat pangenome (Jordan et al., 2021) but this will be computational challenging.

Furthermore, adding a new genotype to populate the PHG would require starting the process over. In our study we used >1,000 whole genome genotypes at 12-fold to create a comprehensive database including, landraces, early breeding cultivars, and modern elite varieties, both public and from the private sector. Conversely to the PHG, to add a new genotype as a query into our haplotype database is straightforward as this will require independently run IBSpy against the pangenome references and concatenate the output information as an additional column to the initial database. Adding a new reference will require to call run IBSpy and call variations of the >1000 genotypes only to this novel reference and concatenate the already present database. Therefore, this advantage of our approach is that as we add more data, there is no need to re-run and start over from scratch. We envision our haplotype database to continue expanding until reaching a robust representation of the most important haplotypes of wheat and novel still untapped haplotypes from landraces. Then, to prevent the haplotype database becoming too large, users will be able to select only sets of samples to filter and focusing only on a particular dataset. Our haplotype database will be better described as long-range haplotypes since it can capture large haplotype blocks at 1 Mbp scale to the entire chromosomes, but it does not discriminate a few SNPs in 50 Kbp window.

In 2021 Jordan *et al.,* (Jordan et al., 2021) reported low imputation accuracy when genotypes harbouring an introgression from *Ae. ventricosa* was tested. Our findings described in Chapter 4 in this study and other reports (Keilwagen et al., 2022; Walkowiak et al., 2020) have found that the wheat pangenome, modern wheat cultivars, and landraces, harbour large induced and natural introgressions/hybridizations from wild relatives across the whole genome. Given that this type of wide crosses is becoming prevalent in wheat breeding programs to restore the genetic diversity and introgress disease resistant traits and better adaptation (Devi et al., 2019; Gaurav et al., 2022;

Grewal et al., 2020; King et al., 2022), we hypothesize that including those wild relatives in a haplotype database will be pivotal to exploit unexplored haplotypes from wheat ancestors. In our present study in addition to the wheat landraces and modern cultivars, we included publicly available accessions of wheat wild relatives, one accession of rye, known to be the donor of the chr1B introgression block (Rabanus-Wallace et al., 2021), a panel of 265 accession of *Ae. tauschii* (Gaurav et al., 2022), and 218 accessions of *T. monoccocum* (Ahmed et al., 2023). In an ongoing project, shortly we will incorporate WGS from 94 additional accessions from a wider collection of wild wheat relatives to expand the search of novel haplotypes for agronomic traits.

In the era of pangenomes, we hypothesize that, one of the reasons of lower imputation accuracies in Jordan et al., 2022 when using wheat compared to other species, may be the use of a single reference, Chinese Spring (RefSeq v.1.0), which is a landrace cultivar (Appels et al., 2018). As shown in Chapter 2 in this analysis and in (Walkowiak et al., 2020), Chinese Spring, along with Norin61, is one of the references that shares the least sequence similarity with other modern cultivars. This was also demonstrated with the low number of haplotype blocks shared in pairwise comparisons among references reported in (Brinton et al., 2020). To anticipate for this type of difficulties and to take advantage of all the available genome sequences, in our analysis we incorporated the eleven chromosome scale assemblies to build our haplotype database. As described before, this may be prohibitive for other variant caller pipelines that rely on alignment and mapping due to computational burden and the wheat genome size and complexity. Furthermore, thanks to progress on sequencing technologies, during the development of this thesis, additional chromosome-scale assemblies have been released (Athiyannan et al., 2022; Aury et al., 2022). We have used these novel references assemblies for specific analysis during this thesis and they will be added to this database to integrate a compressive set of unexplored haplotypes and strengthen our haplotype predictions.

With these advances in NGS, decline in sequencing cost, and genome assembly pipelines and improvements for long-read sequencing (Cheng et al., 2021; Hon et al., 2020; Wenger et al., 2019), the sequencing and assembly of wheat genomes is now relatively straightforward at low cost compared to five years ago. It is therefore predictable that in the forthcoming years it will be common to have access to highly accurate chromosome-scale assemblies including for wild wheat relatives. Adding ancestral wheat relatives to our haplotype database will be of importance to capture novel types of variations and genome structures coming from those unexploited genotypes with high genetic diversity. As exemplified in Chapter 4 with the detection of a novel set of landraces harbouring hybridizations from an outgroup of *Ae. tauschii* L3, adding wild relatives to the haplotype database as a reference will help to elucidate the wheat genome

evolution and domestication time and space. Importantly, at the time of writing this thesis a pangenome assembly of the *Ae. tauschii* is under progress. Efforts to assembly representative accessions from each of the lineages were recently released under Toronto agreement by the Open Wild Wheat consortium (https://openwildwheat.org). These novel genome assemblies wills help to elucidate our hypothetical introgressions from the L3 lineage detected in this thesis for further research and provide the foundation to test for the agronomical benefit and provide insights of the evolution and hybridizations of those novel haplotypes alleles introgressed into hexaploid wheat.

Haplotypes can be used to detect regions under positive or negative selection (Sabeti et al., 2002), fine mapping on multi parent populations (Druet & Georges, 2010; Islam et al., 2016), genome associations (Jiang et al., 2018), and as a dimensionality reduction for population structure (Pattaro et al., 2008). In addition, haplotypes can be used to track inherited blocks from ancestral genotypes to modern wheat cultivars and determine how breeding is shaping the wheat genome that can help to guide future selection strategies. Brinton *et al.,* 2020 demonstrated that large blocks of genome regions are maintained with low recombination in modern cultivars adapted to different environments. In their study, they evidenced how breeders have selected and maintained intact genome regions together in multiple breeding programs worldwide. In our study we investigated the haplotype diversity of wheat and linked them to agronomically important phenotypes. In Chapter 3, we used the defined haplotypes to track IBS regions back from landraces into modern wheat and pedigree relatives. We detected almost intact chromosomes passed from landraces still present in elite commercial varieties. In addition, in Chapter 4, we use our haplotype calls to identify novel introgressions into wheat and validated the donor of the *Ae. ventricosa* 2AS/2N^{v}S translocation and detected old hybridizations in landraces from *Ae. tauschii* L3 lineage not reported before.

5.3. Further applications of IBSpy

5.3.1. IBSpy to detect genome missassemblies

Now genome assemblies are becoming routine for many important and orphan crops. In the last 10 years sequencing technologies has had a huge leap in the chemistry and computing algorithms in novel sequencing methods, sequencing error corrections and high throughput impacting positively in sequencing costs (Armstrong et al., 2019; Athiyannan et al., 2022). With the release of novel genome assemblies in short periods of time, methods to detect miss assemblies quickly and reliable would be of importance.

To alleviate for this demand, methods to detect missassemblies are in the public. For example, one common method to asses and validate genome assemblies is the Optical Mapping (Udall & Dawe, 2017). While running our quality control analysis on the redundant test in Chapter 3 and querying genotypes from raw reads against its own genome assembly using IBSpy, we realized that our approach can be employed to detect genome missassemblies. For example, in a genome assembly, we can subsample raw reads from the initial data used to assemble the corresponding genome and as discussed in Chapter 2, we can run IBSpy to detect *variations* against itself. If there are major misassembles in a particular region of the assembled genome, the *variations* fingerprint will be high because those erroneous assemblies will generate *k*-mers not present in the raw reads.

Following this rationale, in a pilot test in this study we verified the quality of the two assemblies of *T. monococcum* generated as part of this collaboration discussed in Chapter 4. Consistently with our predictions, we found two main missassemblies in the two accessions assembled where the Bionano Optical map failed to validate (unpublished data). Although these results are not present in this thesis because of time, this will be a further development and documentation to integrate into upgraded versions of IBSpy.

5.3.2. IBSpy in other species and crops

One of the challenges that we embraced with the present study was the size and complexity of the wheat genome, which is 16 Gbp, polyploid, and highly repetitive. Although, IBSpy was developed and tested in hexaploidy wheat, across this study we used several others tetraploid and diploid species closely related. We observed similarities on the genome structure of these species such as high levels of *variations* at telomere region compared to centromeres and differentiable number of *variations* in 50 Kbp window among genotypes. During the development of this project several other pangenome projects for other important crops, wild relatives, and orphan crops were released. For example, the Barley pangenome was released along the wheat pangenome (Jayakodi et al., 2020), and the Mazie pangenome was released one year later (Hufford et al., 2021). Those additional genome assemblies enriched the initial repertoire of the genome assemblies. Therefore, there is a constant grow in the number of assemblies available per species and it is predicted that this trend will only increase in the following years.

In the present study, we did not show results on the functionality of IBSpy in other important crop species other than the wheat genome. However, we have run pilot tests using the barley pangenome (Jayakodi et al., 2020) and the maize pangenome (Hufford et al., 2021). The Barley pangenome showed a very similar haplotype-like structure across the genome as with wheat using

50 Kbp windows and 31-*mers* (data not shown). On the other hand, the maize genome showed much higher level of *variations* in pairwise comparisons than the wheat genome. The haplotype-blocks like when plotting the *variations* count were shorter in maize than in wheat. Interestingly, in maize, the two main heterotic groups, Siff Stalk (SS) B73 and the Non-Stiff Stalk (NSS) Mo17 (Li et al., 2022), were grouped separately in two groups when using the hierarchical clustering across the whole genome by chromosomes (data not shown). We envision the possibility that IBSpy *variations* may be of utility to develop cheap molecular markers to detected heterotic groups in a high throughput manner, however this requires further developments.

Although, we tested the potential of IBSpy to call *variations* in other important crops, we did not call haplotypes by our method of AP discussed in Chapter 3. Therefore, it could be, that the optimal window size will differ for other crops with smaller genomes than the wheat genome composition, repetitiveness, and quality of the assemblies as discussed in Chapter 2. For example, for the Barley genome we efficiently detected haplotypes blocks-like with IBSpy using 50 Kbp windows as in wheat, however for maize we observed better results based on blocks-like detected using 10 Kbp window. Similarly, we tested the Brassica genome assemblies and we failed to detect blocks-like consistently (data not shown). This could be due to the still not high-quality of the brassica genome assemblies, or because the small genome sizes (Rousseau-Gueutin et al., 2020). In a recent work, our collaborators embraced a pilot work using WGS of a lettuce collection to detect old introgressions in modern cultivars and obtained promising results using different windows sizes (unpublished data). These examples are encouraging to extend the utility of IBSpy and optimize parameters in different plant species. At the time of writing this thesis, we haven't tested IBSpy in other than plant genomes.

5.3.3. Future considerations and improvements

One of the long-term goals of the present project is to create a global and stable haplotype database for wheat for breeding. Although, the information here generated can be of utility for the wheat community, we acknowledge that the current version of IBSpy requires ~12-fold coverage to efficiently call haplotypes. This coverage is still prohibited to genotype large populations commonly used in breeding programs genotyping for GS, QTL mapping, or MAS due to costs. Therefore, the next challenge to embrace is to test if we can use the current haplotype database to impute or predict haplotypes with much lower sequencing depth than <12-fold or directly call *variations* with a reduced sequencing depth.

In an initial pilot test to answer this question we attempted to call *variations* with skim sequencing at 0.2-fold. However, the analysis demonstrated that the current version of IBSpy does not allow to detect *variations* using this sequencing coverage. Similarly, as with skim sequencing, we failed to capture variation using exome-capture sequencing (data not shown) due to the large gaps in the sequencing data of the query samples. Further improvements to the pipeline will be required to address this challenge. In this project, we have not testes to impute or predict haplotypes directly from the haplotypes database generated in Chapter 3. Further modifications to the initial algorithm will be required to directly call variations from these split-sequencing data.

Finally, at the time of writing this thesis, there is a boom in the progress of computing power and improvements and development of novel predicting algorithms (Jumper et al., 2021; Silver et al., 2016). In the current version of IBSpy, because of time, we tested a limited number of algorithms and methods. However, as the data in genome sequencing is predicted to increase and keep growing, novel algorithms faster and more accurate than AP would be important to test for haplotype predictions of the current IBSpy version.

Improvements in the accuracy of haplotypes may help to use smaller haplotype windows size than <1 Mbp which would benefit hapGWAS associations explored in Chapter 3. Additionally, in the present study we did not test GS with the AP haplotype calls. This would be additional research to explore either with the current AP calls or with a method to efficiently design markers across the genome leveraging the current haplotype database created in this thesis as a starting point to strategically select targeted genome regions to deploy.

In summary, in this thesis, in Chapter 2, we provided a novel method namely IBSpy to detect genetic variations based on direct raw reads and *k*-mers instead of alignments. This method is advantageous to employ with large and complex genomes such as wheat. Using IBSpy, in Chapter 3, we defined haplotypes and created a database based on multi-genome assemblies and ~12-fold coverage raw reads of a large collection of > 1,000 wheat genotypes from the WatSeq project. This database includes landraces, modern cultivars, and wild relatives haplotypes. As a case study, we used this database to run hapGWAS for agronomically important traits and identified candidate genes for cloning for disease resistance. In Chapter 4, we used IBSpy to detect and validate known introgressions and found novel large and relatively small introgressions/hybridizations across the wheat genome either with or without genome assemblies. Finally, in Chapter 5, we describe the challenges on detecting variations in large and complex genomes in the arose of genomic data era and describe the advantages of using a multi-genome-based haplotype database. We describe the opportunities to optimise and further deploy IBSpy in other important and orphan crops with contrasting genomic sizes and complexities.

6. References

Adams, M. W. (1962). Principles of Plant Breeding. *Agronomy Journal*, *54*(4), 372-372. https://doi.org/https://doi.org/10.2134/agronj1962.00021962005400040037x

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*, e55646. https://doi.org/10.7554/eLife.55646

Adhikari, L., Shrestha, S., Wu, S., Crain, J., Gao, L., Evers, B., Wilson, D., Ju, Y., Koo, D.-H., Hucl, P., Pozniak, C., Walkowiak, S., Wang, X., Wu, J., Glaubitz, J. C., DeHaan, L., Friebe, B., & Poland, J. (2022). A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. *Scientific Reports*, *12*(1), 17583. https://doi.org/10.1038/s41598-022-19858-2

Aglawe, S. B., Verma, A. K., & Upadhyay, A. K. (2021). Bioinformatics tools and databases for genomics-assisted breeding and population genetics of plants: a review. *Current Bioinformatics*, *16*(6), 766-773.

Ahmed, H. I., Heuberger, M., Schoen, A., Koo, D.-H., Quiroz-Chavez, J., Adhikari, L., Raupp, J., Cauet, S., Rodde, N., Cravero, C., Callot, C., Lazo, G. R., Kathiresan, N., Sharma, P. K., Moot, I., Yadav, I. S., Singh, L., Saripalli, G., Rawat, N., . . . Krattinger, S. G. (2023). Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature*, *620*(7975), 830-838. https://doi.org/10.1038/s41586-023-06389-7

Akhunov, E. D., Akhunova, A. R., Anderson, O. D., Anderson, J. A., Blake, N., Clegg, M. T., Coleman-Derr, D., Conley, E. J., Crossman, C. C., Deal, K. R., Dubcovsky, J., Gill, B. S., Gu, Y. Q., Hadam, J., Heo, H., Huo, N., Lazo, G. R., Luo, M.-C., Ma, Y. Q., . . . Dvorak, J. (2010). Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics*, *11*(1), 702. https://doi.org/10.1186/1471-2164-11-702

Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K., & Goudey, B. (2019). Exploring effective approaches for haplotype block phasing. *BMC Bioinformatics*, *20*(1), 540. https://doi.org/10.1186/s12859-019-3095-8

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L., Wilkinson, P. A., Coghill, J., Waterfall, C., & Davassi, A. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (Triticum aestivum). *Plant biotechnology journal*, *15*(3), 390-401.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., . . . Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, *182*(1), 145-161.e123. https://doi.org/https://doi.org/10.1016/j.cell.2020.05.021

Altshuler, D., Donnelly, P., & The International HapMap, C. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299-1320. https://doi.org/10.1038/nature04226

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Anh, V. L., Anh, N. T., Tagle, A. G., Vy, T. T. P., Inoue, Y., Takumi, S., Chuma, I., & Tosa, Y. (2015). Rmg8, a New Gene for Resistance to Triticum Isolates of Pyricularia oryzae in Hexaploid Wheat. *Phytopathology®*, *105*(12), 1568-1572. https://doi.org/10.1094/phyto-02-15-0034-r

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A. G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., Himmelbach, A., . . . Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), eaar7191. https://doi.org/doi:10.1126/science.aar7191

Aradottir, G. I., Martin, J. L., Clark, S. J., Pickett, J. A., & Smart, L. E. (2017). Searching for wheat resistance to aphids and wheat bulb fly in the historical Watkins and Gediflux wheat collections. *Annals of Applied Biology*, *170*(2), 179-188. https://doi.org/https://doi.org/10.1111/aab.12326

Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. *Annual Review of Animal Biosciences*, *7*(1), 41-64. https://doi.org/10.1146/annurev-animal-020518-115005

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., . . . Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*(7833), 246-251. https://doi.org/10.1038/s41586-020-2871-y

Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., Johnson, R., Enk, J., Periyannan, S., Singh, N., Asyraf Md Hatta, M., Athiyannan, N., Cheema, J., Yu, G., Kangara, N., Ghosh, S., Szabo, L. J., Poland, J., Bariana, H., . . . Wulff, B. B. H. (2019). Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nature Biotechnology*, *37*(2), 139-143. https://doi.org/10.1038/s41587-018-0007-9

Athiyannan, N., Abrouk, M., Boshoff, W. H. P., Cauet, S., Rodde, N., Kudrna, D., Mohammed, N., Bettgenhaeuser, J., Botha, K. S., Derman, S. S., Wing, R. A., Prins, R., & Krattinger, S. G. (2022). Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nature Genetics*, *54*(3), 227-231. https://doi.org/10.1038/s41588-022-01022-1

Audano, P. A., Ravishankar, S., & Vannberg, F. O. (2017). Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics*, *34*(10), 1659-1665. https://doi.org/10.1093/bioinformatics/btx753

Aury, J.-M., Engelen, S., Istace, B., Monat, C., Lasserre-Zuber, P., Belser, C., Cruaud, C., Rimbert, H., Leroy, P., Arribat, S., Dufau, I., Bellec, A., Grimbichler, D., Papon, N., Paux, E., Ranoux, M., Alberti, A., Wincker, P., & Choulet, F. (2022).

Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac034

Badaeva, E. D., Dedkova, O. S., Koenig, J., Bernard, S., & Bernard, M. (2008). Analysis of introgression of Aegilops ventricosa Tausch. genetic material in a common wheat background using C-banding. *Theoretical and Applied Genetics*, *117*(5), 803-811. https://doi.org/10.1007/s00122-008-0821-4

Baenziger, P., & Principal, R. (2009). Wheat Breeding: Procedures and Strategies. In (pp. 273-308). https://doi.org/10.1002/9780813818832.ch13

Baker, M. (2010). Next-generation sequencing: adjusting to data overload. *Nature Methods*, *7*(7), 495-499. https://doi.org/10.1038/nmeth0710-495

Balfourier, F., Bouchet, S., Robert, S., De Oliveira, R., Rimbert, H., Kitt, J., Choulet, F., & Paux, E. (2019). Worldwide phylogeography and history of wheat genetic diversity. *Science Advances*, *5*(5), eaav0536. https://doi.org/doi:10.1126/sciadv.aav0536

Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, *17*(suppl_1), S22-S29. https://doi.org/10.1093/bioinformatics/17.suppl_1.S22

Bariana, H. S., & McIntosh, R. A. (1994). Characterisation and origin of rust and powdery mildew resistance genes in VPM1 wheat. *Euphytica*, *76*(1), 53-61. https://doi.org/10.1007/BF00024020

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263-265. https://doi.org/10.1093/bioinformatics/bth457

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, *6*(8), 914-920. https://doi.org/10.1038/s41477-020-0733-0

Bentley, A. R., Donovan, J., Sonder, K., Baudron, F., Lewis, J. M., Voss, R., Rutsaert, P., Poole, N., Kamoun, S., Saunders, D. G. O., Hodson, D., Hughes, D. P., Negra, C., Ibba, M. I., Snapp, S., Sida, T. S., Jaleta, M., Tesfaye, K., Becker-Reshef, I., & Govaerts, B. (2022). Near- to long-term measures to stabilize global wheat supplies and food security. *Nature Food*, *3*(7), 483-486. https://doi.org/10.1038/s43016-022-00559-y

Berger, E., Yorukoglu, D., Zhang, L., Nyquist, S. K., Shalek, A. K., Kellis, M., Numanagić, I., & Berger, B. (2020). Improved haplotype inference by exploiting long-range linking and allelic imbalance in RNA-seq datasets. *Nature Communications*, *11*(1), 4662. https://doi.org/10.1038/s41467-020-18320-z

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., & Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature*, *543*(7645), 346-354. https://doi.org/10.1038/nature22011

Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., & Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. *Communications Biology*, *4*(1), 1266. https://doi.org/10.1038/s42003-021-02782-y

Borlaug, N. E. (1983). Contributions of Conventional Plant Breeding to Food Production. *Science*, *219*(4585), 689-693. https://doi.org/doi:10.1126/science.219.4585.689

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, *32*(3), 314.

Bradbury, P. J., Casstevens, T., Jensen, S. E., Johnson, L. C., Miller, Z. R., Monier, B., Romay, M. C., Song, B., & Buckler, E. S. (2022). The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics*, *38*(15), 3698-3702. https://doi.org/10.1093/bioinformatics/btac410

Brinton, J., Ramirez-Gonzalez, R. H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Haberer, G., Spannagl, M., Walkowiak, S., Pozniak, C., Uauy, C., & Wheat Genome, P. (2020). A haplotype-led approach to increase the precision of wheat breeding. *Communications Biology*, *3*(1), 712. https://doi.org/10.1038/s42003-020-01413-2

Brown, T. A., Jones, M. K., Powell, W., & Allaby, R. G. (2009). The complex origins of domesticated crops in the Fertile Crescent. *Trends in Ecology & Evolution*, *24*(2), 103-109. https://doi.org/https://doi.org/10.1016/j.tree.2008.09.008

Brown-Guedira, G. L., Singh, S., & Fritz, A. K. (2003). Performance and Mapping of Leaf Rust Resistance Transferred to Wheat from Triticum timopheevii subsp. armeniacum. *Phytopathology®*, *93*(7), 784-789. https://doi.org/10.1094/phyto.2003.93.7.784

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J. F., Ross-Ibarra, J., Lorant, A., . . . Xu, Y. (2017). Construction of the third-generation Zea mays haplotype map. *GigaScience*, *7*(4). https://doi.org/10.1093/gigascience/gix134

Caldwell, K. S., Russell, J., Langridge, P., & Powell, W. (2006). Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. *Genetics*, *172*(1), 557-567. https://doi.org/10.1534/genetics.104.038489

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., & Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, *43*(10), 956-963. https://doi.org/10.1038/ng.911

Cao, P., Fan, W., Li, P., & Hu, Y. (2021). Genome-wide profiling of long noncoding RNAs involved in wheat spike development. *BMC Genomics*, *22*(1), 493. https://doi.org/10.1186/s12864-021-07851-4

Chapman, J. A., Mascher, M., Buluç, A., Barry, K., Georganas, E., Session, A., Strnadova, V., Jenkins, J., Sehgal, S., Oliker, L., Schmutz, J., Yelick, K. A., Scholz, U., Waugh, R., Poland, J. A., Muehlbauer, G. J., Stein, N., & Rokhsar, D. S. (2015). A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology*, *16*(1), 26. https://doi.org/10.1186/s13059-015-0582-8

Chemayek, B., Bansal, U. K., Qureshi, N., Zhang, P., Wagoire, W. W., & Bariana, H. S. (2017). Tight repulsion linkage between Sr36 and Sr39 was revealed by genetic, cytogenetic and molecular analyses. *Theoretical and Applied Genetics*, *130*(3), 587-595. https://doi.org/10.1007/s00122-016-2837-5

Chen, L., Zhu, Q.-H., & Kaufmann, K. (2020). Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. *Planta*, *252*(5), 92. https://doi.org/10.1007/s00425-020-03480-5

Chen, S., Hegarty, J., Shen, T., Hua, L., Li, H., Luo, J., Li, H., Bai, S., Zhang, C., & Dubcovsky, J. (2021). Stripe rust resistance gene Yr34 (synonym Yr48) is located within a distal translocation of Triticum monococcum chromosome 5AmL into common wheat. *Theoretical and Applied Genetics*, *134*(7), 2197-2211. https://doi.org/10.1007/s00122-021-03816-z

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170-175. https://doi.org/10.1038/s41592-020-01056-5

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*(10), 966-968. https://doi.org/10.1038/nmeth.3505

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., Leroy, P., Mangenot, S., Guilhot, N., Le Gouis, J., Balfourier, F., Alaux, M., Jamilloux, V., Poulain, J., Durand, C., . . . Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science*, *345*(6194), 1249721. https://doi.org/doi:10.1126/science.1249721

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R., & Weigel, D. (2007). Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. *Science*, *317*(5836), 338-342. https://doi.org/doi:10.1126/science.1138632

Contreras-Soto, R. I., Mora, F., de Oliveira, M. A. R., Higashi, W., Scapim, C. A., & Schuster, I. (2017). A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. *PLOS ONE*, *12*(2), e0171105. https://doi.org/10.1371/journal.pone.0171105

Cooper, D. N., & Schmidtke, J. (1984). DNA restriction fragment length polymorphisms and heterozygosity in the human genome. *Human Genetics*, *66*(1), 1-16. https://doi.org/10.1007/BF00275182

Cruz, C. D., Peterson, G. L., Bockus, W. W., Kankanala, P., Dubcovsky, J., Jordan, K. W., Akhunov, E., Chumley, F., Baldelomar, F. D., & Valent, B. (2016). The 2NS Translocation from Aegilops ventricosa Confers Resistance to the Triticum Pathotype of Magnaporthe oryzae. *Crop Science*, *56*(3), 990-1000. https://doi.org/https://doi.org/10.2135/cropsci2015.07.0410

Cseh, A., Poczai, P., Kiss, T., Balla, K., Berki, Z., Horváth, Á., Kuti, C., & Karsai, I. (2021). Exploring the legacy of Central European historical winter wheat landraces. *Scientific Reports*, *11*(1), 23915. https://doi.org/10.1038/s41598-021-03261-4

Daud, H. M., & Gustafson, J. P. (1996). Molecular evidence for Triticum speltoides as a B-genome progenitor of wheat (Triticum aestivum). *Genome*, *39*(3), 543-548. https://doi.org/10.1139/g96-069 %M 18469915

De Coster, W., & Van Broeckhoven, C. (2019). Newest Methods for Detecting Structural Variations. *Trends in Biotechnology*, *37*(9), 973-982. https://doi.org/https://doi.org/10.1016/j.tibtech.2019.02.003

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics*, *22*(9), 572-587. https://doi.org/10.1038/s41576-021-00367-3

Delorean, E., Gao, L., Lopez, J. F. C., Mehrabi, A., Bentley, A., Sharon, A., Keller, B., Wulff, B., Steffenson, B., Steuernagel, B., Sansaloni, C. P., Liu, D.-C., Lagudah, E., Nasyrova, F., Brown-Guedira, G., Sela, H., Dvorak, J., Poland, J., Mayer, K., . . . Open Wild Wheat, C. (2021). High molecular weight glutenin gene diversity in Aegilops tauschii demonstrates unique origin of superior wheat quality. *Communications Biology*, *4*(1), 1242. https://doi.org/10.1038/s42003-021-02563-7

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., Xu, Y., Deng, X. W., He, H., & Zhang, X. (2022). A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Molecular Plant*, *15*(8), 1268-1284. https://doi.org/https://doi.org/10.1016/j.molp.2022.06.010

Denti, L., Previtali, M., Bernardini, G., Schönhuth, A., & Bonizzoni, P. (2019). MALVA: Genotyping by Mapping-free ALlele Detection of Known VAriants. *iScience*, *18*, 20-27. https://doi.org/https://doi.org/10.1016/j.isci.2019.07.011

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Devi, U., Grewal, S., Yang, C.-y., Hubbart-Edwards, S., Scholefield, D., Ashling, S., Burridge, A., King, I. P., & King, J. (2019). Development and characterisation of interspecific hybrid lines with genome-wide introgressions from Triticum timopheevii in a hexaploid wheat background. *BMC Plant Biology*, *19*(1), 183. https://doi.org/10.1186/s12870-019-1785-z

Doussinault, G., Delibes, A., Sanchez-Monge, R., & Garcia-Olmedo, F. (1983). Transfer of a dominant gene for resistance to eyespot disease from a wild grass to hexaploid wheat. *Nature*, *303*(5919), 698-700. https://doi.org/10.1038/303698a0

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.

Druet, T., & Georges, M. (2010). A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. *Genetics*, *184*(3), 789-798. https://doi.org/10.1534/genetics.109.108431

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/doi:10.1126/science.1143986

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., . . . Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537), eabf7117. https://doi.org/doi:10.1126/science.abf7117

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, *54*(4), 518-525. https://doi.org/10.1038/s41588-022-01043-w

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V., & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, *10*(1), 5402. https://doi.org/10.1038/s41467-019-13341-9

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293-314. https://doi.org/10.1093/nsr/nwt032

Feldman, M. (1966). IDENTIFICATION OF UNPAIRED CHROMOSOMES IN F1 HYBRIDS INVOLVING TRITICUM AESTIVUM AND T. TIMOPHEEV II. *Canadian Journal of Genetics and Cytology*, *8*(1), 144-151. https://doi.org/10.1139/g66-019

Franco, J. A. V., Gage, J. L., Bradbury, P. J., Johnson, L. C., Miller, Z. R., Buckler, E. S., & Romay, M. C. (2020). A maize practical haplotype graph leverages diverse NAM assemblies. *bioRxiv*.

Frey, B. J., & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, *315*(5814), 972-976. https://doi.org/doi:10.1126/science.1136800

Friebe, B., Jiang, J., Raupp, W. J., McIntosh, R. A., & Gill, B. S. (1996). Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. *Euphytica*, *91*(1), 59-87. https://doi.org/10.1007/BF00035277

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science*, *296*(5576), 2225-2229. https://doi.org/doi:10.1126/science.1069424

Gale, M. D., Scott, P. R., Law, C. N., Ainsworth, C. C., Hollins, T. W., & Worland, A. J. (1984). An α-amylase gene from Aegilops ventricosa transferred to bread wheat together with a factor for eyespot resistance. *Heredity*, *52*(3), 431-435. https://doi.org/10.1038/hdy.1984.51

Gao, L., Koo, D.-H., Juliana, P., Rife, T., Singh, D., Lemes da Silva, C., Lux, T., Dorn, K. M., Clinesmith, M., Silva, P., Wang, X., Spannagl, M., Monat, C., Friebe, B.,

Steuernagel, B., Muehlbauer, G. J., Walkowiak, S., Pozniak, C., Singh, R., . . . Poland, J. (2021). The Aegilops ventricosa 2NvS segment in bread wheat: cytology, genomics and breeding. *Theoretical and Applied Genetics*, *134*(2), 529-542. https://doi.org/10.1007/s00122-020-03712-y

Gardiner, L.-J., Brabbs, T., Akhunov, A., Jordan, K., Budak, H., Richmond, T., Singh, S., Catchpole, L., Akhunov, E., & Hall, A. (2019). Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *GigaScience*, *8*(4). https://doi.org/10.1093/gigascience/giz018

Garg, S. (2021). Computational methods for chromosome-scale haplotype reconstruction. *Genome Biology*, *22*(1), 101. https://doi.org/10.1186/s13059-021-02328-9

Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S., Peluso, P., Hatas, E., Ghurye, J., Maguire, J., Mahmoud, M., Cheng, H., Heller, D., Zook, J. M., Moemke, T., Marschall, T., Sedlazeck, F. J., Aach, J., . . . Li, H. (2021). Chromosome-scale, haplotype-resolved assembly of human genomes. *Nature Biotechnology*, *39*(3), 309-312. https://doi.org/10.1038/s41587-020-0711-0

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Gaurav, K., Arora, S., Silva, P., Sánchez-Martín, 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-Chávez, 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

Gaut, B. S., Seymour, D. K., Liu, Q., & Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nature Plants*, *4*(8), 512-520. https://doi.org/10.1038/s41477-018-0210-1

Gill, B. S., & Raupp, W. J. (1987). Direct Genetic Transfers from Aegilops squarrosa L. to Hexaploid Wheat1. *Crop Science*, *27*(3), cropsci1987.0011183X002700030004x. https://doi.org/https://doi.org/10.2135/cropsci1987.0011183X002700030004x

Goddard, R., Steed, A., Chinoy, C., Ferreira, J. R., Scheeren, P. L., Maciel, J. L. N., Caierão, E., Torres, G. A. M., Consoli, L., Santana, F. M., Fernandes, J. M. C., Simmonds, J., Uauy, C., Cockram, J., & Nicholson, P. (2020). Dissecting the genetic basis of wheat blast resistance in the Brazilian wheat cultivar BR 18-Terena. *BMC Plant Biology*, *20*(1), 398. https://doi.org/10.1186/s12870-020-02592-0

Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H., & Buckler, E. S. (2009). A First-Generation Haplotype Map of Maize. *Science*, *326*(5956), 1115-1117. https://doi.org/doi:10.1126/science.1177837

Gottlieb, L. (1981). Electrophoretic evidence and plant populations. *Progress in phytochemistry*, *7*, 1-46.

Grewal, S., Hubbart-Edwards, S., Yang, C., Devi, U., Baker, L., Heath, J., Ashling, S., Scholefield, D., Howells, C., Yarde, J., Isaac, P., King, I. P., & King, J. (2020).

Rapid identification of homozygosity and site of wild relative introgressions in wheat through chromosome-specific KASP genotyping assays. *Plant biotechnology journal*, *18*(3), 743-755. https://doi.org/https://doi.org/10.1111/pbi.13241

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., Appels, R., Peng, H., Ni, Z., & Sun, Q. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. *Nature Communications*, *11*(1), 5085. https://doi.org/10.1038/s41467-020-18738-5

Guo, Y., Li, J., Bonham, A. J., Wang, Y., & Deng, H. (2009). Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. *European Journal of Human Genetics*, *17*(6), 785-792. https://doi.org/10.1038/ejhg.2008.244

Hamazaki, K., & Iwata, H. (2020). RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLOS Computational Biology*, *16*(2), e1007663. https://doi.org/10.1371/journal.pcbi.1007663

Hammond-Kosack, M. C. U., King, R., Kanyuka, K., & Hammond-Kosack, K. E. (2021). Exploring the diversity of promoter and 5′UTR sequences in ancestral, historic and modern wheat. *Plant biotechnology journal*, *19*(12), 2469-2487. https://doi.org/https://doi.org/10.1111/pbi.13672

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, *47*(3), 296-303. https://doi.org/10.1038/ng.3200

Hao, M., Zhang, L., Ning, S., Huang, L., Yuan, Z., Wu, B., Yan, Z., Dai, S., Jiang, B., Zheng, Y., & Liu, D. (2020). The Resurgence of Introgression Breeding, as Exemplified in Wheat Improvement [Review]. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00252

Hasan, N., Choudhary, S., Naaz, N., Sharma, N., & Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *Journal of Genetic Engineering and Biotechnology*, *19*(1), 128. https://doi.org/10.1186/s43141-021-00231-1

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., Hochu, I., Poirier, S., Santoni, S., Glémin, S., & David, J. (2007). Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Molecular Biology and Evolution*, *24*(7), 1506-1517. https://doi.org/10.1093/molbev/msm077

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., Knox, R., Cuthbert, R., Pozniak, C., Akhunova, A., Morrell, P. L., Davies, J. P., Webb, S. R., Spangenberg, G., Hayes, B., . . . Akhunov, E. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nature Genetics*, *51*(5), 896-904. https://doi.org/10.1038/s41588-019-0382-2

He, S., Thistlethwaite, R., Forrest, K., Shi, F., Hayden, M. J., Trethowan, R., & Daetwyler, H. D. (2019). Extension of a haplotype-based genomic prediction model to manage multi-environment wheat data using environmental covariates. *Theoretical and Applied Genetics*, *132*(11), 3143-3154. https://doi.org/10.1007/s00122-019-03413-1

Helguera, M., Khan, I. A., Kolmer, J., Lijavetzky, D., Zhong-qi, L., & Dubcovsky, J. (2003). PCR Assays for the Lr37-Yr17-Sr38 Cluster of Rust Resistance Genes and Their Use to Develop Isogenic Hard Red Spring Wheat Lines. *Crop Science*, *43*(5), 1839-1847. https://doi.org/https://doi.org/10.2135/cropsci2003.1839

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., & Salamini, F. (1997). Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. *Science*, *278*(5341), 1312-1314. https://doi.org/doi:10.1126/science.278.5341.1312

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 399. https://doi.org/10.1038/s41597-020-00743-4

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801-811. https://doi.org/https://doi.org/10.1016/j.humimm.2021.02.012

Huang, B. E., Amos, C. I., & Lin, D. Y. (2007). Detecting haplotype effects in genomewide association studies. *Genetic Epidemiology*, *31*(8), 803-812. https://doi.org/https://doi.org/10.1002/gepi.20242

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. *Proceedings of the National Academy of Sciences*, *99*(12), 8133-8138. https://doi.org/doi:10.1073/pnas.072223799

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., Guo, Y., Lu, Y., Zhou, C., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang, Y., . . . Han, B. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature*, *490*(7421), 497-501. https://doi.org/10.1038/nature11532

Hubbard, A., Lewis, C. M., Yoshida, K., Ramirez-Gonzalez, R. H., de Vallavieille-Pope, C., Thomas, J., Kamoun, S., Bayles, R., Uauy, C., & Saunders, D. G. O. (2015). Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology*, *16*(1), 23. https://doi.org/10.1186/s13059-015-0590-8

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., . . . Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*(6555), 655-662. https://doi.org/doi:10.1126/science.abg5289

Hummel, N. F., Zhou, A., Li, B., Markel, K., Ornelas, I. J., & Shih, P. M. (2023). The trans-regulatory landscape of gene networks in plants. *Cell Systems*, *14*(6), 501-511. e504.

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*(1), 17875. https://doi.org/10.1038/srep17875

Hyten, D. L., Choi, I.-Y., Song, Q., Shoemaker, R. C., Nelson, R. L., Costa, J. M., Specht, J. E., & Cregan, P. B. (2007). Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations. *Genetics*, *175*(4), 1937-1944. https://doi.org/10.1534/genetics.106.069740

Ikeda, K., Ito, M., Nagasawa, N., Kyozuka, J., & Nagato, Y. (2007). Rice ABERRANT PANICLE ORGANIZATION 1, encoding an F-box protein, regulates meristem fate. *The Plant Journal*, *51*(6), 1030-1040. https://doi.org/https://doi.org/10.1111/j.1365-313X.2007.03200.x

Islam, M. S., Thyssen, G. N., Jenkins, J. N., Zeng, L., Delhom, C. D., McCarty, J. C., Deng, D. D., Hinchliffe, D. J., Jones, D. C., & Fang, D. D. (2016). A MAGIC population-based genome-wide association study reveals functional association of GhRBB1_A07 gene with superior fiber quality in cotton. *BMC Genomics*, *17*(1), 903. https://doi.org/10.1186/s12864-016-3249-2

Jakubosky, D., D'Antonio, M., Bonder, M. J., Smail, C., Donovan, M. K. R., Young Greenwald, W. W., Matsui, H., Bonder, M. J., Cai, N., Carcamo-Orive, I., D'Antonio, M., Frazer, K. A., Young Greenwald, W. W., Jakubosky, D., Knowles, J. W., Matsui, H., McCarthy, D. J., Mirauta, B. A., Montgomery, S. B., . . . i, Q. T. L. C. (2020). Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nature Communications*, *11*(1), 2927. https://doi.org/10.1038/s41467-020-16482-4

Jakubosky, D., Smith, E. N., D'Antonio, M., Jan Bonder, M., Young Greenwald, W. W., D'Antonio-Chronowska, A., Matsui, H., Bonder, M. J., Cai, N., Carcamo-Orive, I., D'Antonio, M., Frazer, K. A., Young Greenwald, W. W., Jakubosky, D., Knowles, J. W., Matsui, H., McCarthy, D. J., Mirauta, B. A., Montgomery, S. B., . . . i, Q. T. L. C. (2020). Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nature Communications*, *11*(1), 2928. https://doi.org/10.1038/s41467-020-16481-5

Järve, K., Peusha, H. O., Tsymbalova, J., Tamm, S., Devos, K. M., & Enno, T. M. (2000). Chromosomal location of a Triticum timopheevii - derived powdery mildew resistance gene transferred to common wheat. *Genome*, *43*(2), 377-381. https://doi.org/10.1139/g99-141 %M 10791827

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., Ens, J., Zhang, X.-Q., Angessa, T. T., Zhou, G., Tan, C., Hill, C., Wang, P., Schreiber, M., Boston, L. B., . . . Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, *588*(7837), 284-289. https://doi.org/10.1038/s41586-020-2947-8

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, *314*(6006), 67-73. https://doi.org/10.1038/314067a0

Jensen, S. E., Charles, J. R., Muleta, K., Bradbury, P. J., Casstevens, T., Deshpande, S. P., Gore, M. A., Gupta, R., Ilut, D. C., Johnson, L., Lozano, R., Miller, Z., Ramu, P., Rathore, A., Romay, M. C., Upadhyaya, H. D., Varshney, R. K., Morris, G. P., Pressoir, G., . . . Ramstein, G. P. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *The Plant Genome*, *13*(1), e20009. https://doi.org/https://doi.org/10.1002/tpg2.20009

Jiang, Y., Schmidt, R. H., & Reif, J. C. (2018). Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. *G3 Genes|Genomes|Genetics*, *8*(5), 1687-1699. https://doi.org/10.1534/g3.117.300548

Jiao, W.-B., & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, *36*, 64-70. https://doi.org/https://doi.org/10.1016/j.pbi.2017.02.002

Jordan, K. W., Bradbury, P. J., Miller, Z. R., Nyine, M., He, F., Fraser, M., Anderson, J., Mason, E., Katz, A., Pearce, S., Carter, A. H., Prather, S., Pumphrey, M., Chen, J., Cook, J., Liu, S., Rudd, J. C., Wang, Z., Chu, C., . . . Akhunov, E. D. (2021). Development of the Wheat Practical Haplotype Graph database as a resource for genotyping data storage and genotype imputation. *G3 Genes|Genomes|Genetics*, *12*(2). https://doi.org/10.1093/g3journal/jkab390

Jordan, K. W., Wang, S., Lun, Y., Gardiner, L.-J., MacLachlan, R., Hucl, P., Wiebe, K., Wong, D., Forrest, K. L., Sharpe, A. G., Sidebottom, C. H. D., Hall, N., Toomajian, C., Close, T., Dubcovsky, J., Akhunova, A., Talbert, L., Bansal, U. K., Bariana, H. S., . . . Consortium, I. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biology*, *16*(1), 48. https://doi.org/10.1186/s13059-015-0606-4

Juliana, P., He, X., Kabir, M. R., Roy, K. K., Anwar, M. B., Marza, F., Poland, J., Shrestha, S., Singh, R. P., & Singh, P. K. (2020). Genome-wide association mapping for wheat blast resistance in CIMMYT's international screening nurseries evaluated in Bolivia and Bangladesh. *Scientific Reports*, *10*(1), 15972. https://doi.org/10.1038/s41598-020-72735-8

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kale, S. M., Schulthess, A. W., Padmarasu, S., Boeven, P. H. G., Schacht, J., Himmelbach, A., Steuernagel, B., Wulff, B. B. H., Reif, J. C., Stein, N., & Mascher, M. (2022). A catalogue of resistance gene homologs and a chromosome-scale reference sequence support resistance gene mapping in winter wheat. *Plant biotechnology journal*, *20*(9), 1730-1742. https://doi.org/https://doi.org/10.1111/pbi.13843

Kearsey, M. J., & Farquhar, A. G. L. (1998). QTL analysis in plants; where are we now? *Heredity*, *80*(2), 137-142. https://doi.org/10.1046/j.1365-2540.1998.00500.x

Kehoe, D. M., Villand, P., & Somerville, S. (1999). DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends in Plant Science*, *4*(1), 38-41. https://doi.org/https://doi.org/10.1016/S1360-1385(98)01354-5

Keilwagen, J., Lehnert, H., Berner, T., Badaeva, E., Himmelbach, A., Börner, A., & Kilian, B. (2022). Detecting major introgressions in wheat and their putative origins using coverage analysis. *Scientific Reports*, *12*(1), 1908. https://doi.org/10.1038/s41598-022-05865-w

Keilwagen, J., Lehnert, H., Berner, T., Beier, S., Scholz, U., Himmelbach, A., Stein, N., Badaeva, E. D., Lang, D., Kilian, B., Hackauf, B., & Perovic, D. (2019). Detecting Large Chromosomal Modifications Using Short Read Data From Genotyping-by-Sequencing [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01133

Kesawat, M. S., & Das Kumar, B. (2009). Molecular markers: It's application in crop improvement. *Journal of Crop Science and Biotechnology*, *12*(4), 169-181. https://doi.org/10.1007/s12892-009-0124-6

Kilian, B., Dempewolf, H., Guarino, L., Werner, P., Coyne, C., & Warburton, M. L. (2021). Crop Science special issue: Adapting agriculture to climate change: A walk on the wild side. *Crop Science*, *61*(1), 32-36. https://doi.org/https://doi.org/10.1002/csc2.20418

Kilian, B., Özkan, H., Walther, A., Kohl, J., Dagan, T., Salamini, F., & Martin, W. (2007). Molecular Diversity at 18 Loci in 321 Wild and 92 Domesticate Lines Reveal No Reduction of Nucleotide Diversity during Triticum monococcum (Einkorn) Domestication: Implications for the Origin of Agriculture. *Molecular Biology and Evolution*, *24*(12), 2657-2668. https://doi.org/10.1093/molbev/msm192

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., & Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, *39*(9), 1151-1155. https://doi.org/10.1038/ng2115

Kim, S. A., Cho, C. S., Kim, S. R., Bull, S. B., & Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, *34*(3), 388-397. https://doi.org/10.1093/bioinformatics/btx609

King, J., Grewal, S., Othmeni, M., Coombes, B., Yang, C.-y., Walter, N., Ashling, S., Scholefield, D., Walker, J., Hubbart-Edwards, S., Hall, A., & King, I. P. (2022). Introgression of the Triticum timopheevii Genome Into Wheat Detected by Chromosome-Specific Kompetitive Allele Specific PCR Markers [Original Research]. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.919519

Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, *33*(17), 2759-2761. https://doi.org/10.1093/bioinformatics/btx304

Kolmer, J. A., Anderson, J. A., & Flor, J. M. (2010). Chromosome Location, Linkage with Simple Sequence Repeat Markers, and Leaf Rust Resistance Conditioned by Gene Lr63 in Wheat. *Crop Science*, *50*(6), 2392-2395. https://doi.org/https://doi.org/10.2135/cropsci2010.01.0005

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(1), 29. https://doi.org/10.1186/1746-4811-9-29

Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M., & Goto, S. (2012). GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Research*, *40*(W1), W162-W167. https://doi.org/10.1093/nar/gks459

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., Simmonds, J., Ramirez-Gonzalez, R. H., Wang, X., 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*, *114*(6), E913-E921. https://doi.org/doi:10.1073/pnas.1619268114

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Porubsky, D., Kuhn, K., Mueller, K. A., Low, W. Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R. J., Phillippy, A. M., Eichler, E. E., Williams, J. L., Smith, T. P. L., Jarvis, E. D., . . . Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, *12*(1), 1935. https://doi.org/10.1038/s41467-020-20536-y

Kumar, J., Gupta, D. S., Gupta, S., Dubey, S., Gupta, P., & Kumar, S. (2017). Quantitative trait loci from identification to exploitation for crop improvement. *Plant Cell Reports*, *36*(8), 1187-1213. https://doi.org/10.1007/s00299-017-2127-y

Kumar, K., Jan, I., Saripalli, G., Sharma, P. K., Mir, R. R., Balyan, H. S., & Gupta, P. K. (2022). An Update on Resistance Genes and Their Use in the Development of Leaf Rust Resistant Cultivars in Wheat [Mini Review]. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.816057

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

Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, *35*(9), 780-786.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Langridge, P., & Waugh, R. (2019). Harnessing the potential of germplasm collections. *Nature Genetics*, *51*(2), 200-201. https://doi.org/10.1038/s41588-018-0340-4

Lee, H., Shuaibi, A., Bell, J. M., Pavlichin, D. S., & Ji, H. P. (2020). Unique k-mer sequences for validating cancer-related substitution, insertion and deletion mutations. *NAR Cancer*, *2*(4). https://doi.org/10.1093/narcan/zcaa034

Leigh, F. J., Wright, T. I. C., Horsnell, R. A., Dyer, S., & Bentley, A. R. (2022). Progenitor species hold untapped diversity for potential climate-responsive traits for use in wheat breeding and crop improvement. *Heredity*, *128*(5), 291-303. https://doi.org/10.1038/s41437-022-00527-z

Lev-Yadun, S., Gopher, A., & Abbo, S. (2000). The Cradle of Agriculture. *Science*, *288*(5471), 1602-1603. https://doi.org/doi:10.1126/science.288.5471.1602

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, C., Guan, H., Jing, X., Li, Y., Wang, B., Li, Y., Liu, X., Zhang, D., Liu, C., Xie, X., Zhao, H., Wang, Y., Liu, J., Zhang, P., Hu, G., Li, G., Li, S., Sun, D., Wang, X., . .

. Wang, H. (2022). Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nature Plants*, *8*(7), 750-763. https://doi.org/10.1038/s41477-022-01190-2

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

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589-595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., Burgueno-Ferreira, J. A., Bhavani, S., Huerta-Espino, J., Payne, T., Sehgal, D., Wenzl, P., & Singh, S. (2015). A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics*, *16*(1), 216. https://doi.org/10.1186/s12864-015-1424-5

Li, W., & Gill, B. S. (2006). Multiple genetic pathways for seed shattering in the grasses. *Functional & Integrative Genomics*, *6*(4), 300-309. https://doi.org/10.1007/s10142-005-0015-y

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., Yu, Y., Du, H., Qi, M., Li, Y., Lu, H., Yu, H., Cui, Y., Wang, N., Chen, C., . . . Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature*, *557*(7705), 424-428. https://doi.org/10.1038/s41586-018-0108-0

Liu, F., Zhao, J., Sun, H., Xiong, C., Sun, X., Wang, X., Wang, Z., Jarret, R., Wang, J., Tang, B., Xu, H., Hu, B., Suo, H., Yang, B., Ou, L., Li, X., Zhou, S., Yang, S., Liu, Z., . . . Zou, X. (2023). Genomes of cultivated and wild Capsicum species provide insights into pepper domestication and population differentiation. *Nature Communications*, *14*(1), 5487. https://doi.org/10.1038/s41467-023-41251-4

Liu, S., Zheng, J., Migeon, P., Ren, J., Hu, Y., He, C., Liu, H., Fu, J., White, F. F., Toomajian, C., & Wang, G. (2017). Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. *Scientific Reports*, *7*(1), 42444. https://doi.org/10.1038/srep42444

Long, E. M., Bradbury, P. J., Romay, M. C., Buckler, E. S., & Robbins, K. R. (2021). Genome-wide imputation using the practical haplotype graph in the heterozygous crop cassava. *G3 Genes|Genomes|Genetics*, *12*(1). https://doi.org/10.1093/g3journal/jkab383

Lozano, R., Gazave, E., dos Santos, J. P. R., Stetter, M. G., Valluru, R., Bandillo, N., Fernandes, S. B., Brown, P. J., Shakoor, N., Mockler, T. C., Cooper, E. A., Taylor Perkins, M., Buckler, E. S., Ross-Ibarra, J., & Gore, M. A. (2021). Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nature Plants*, *7*(1), 17-24. https://doi.org/10.1038/s41477-020-00834-5

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., McGuire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van, S. L., Yuan, L., Wang, Z., Xia, Z., . . . Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature*, *551*(7681), 498-502. https://doi.org/10.1038/nature24486

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764-770. https://doi.org/10.1093/bioinformatics/btr011

Marchal, C., Project, W. G., Haberer, G., Spannagl, M., & Uauy, C. (2020). Comparative Genomics and Functional Studies of Wheat BED-NLR Loci. *Genes*, *11*(12), 1406. https://www.mdpi.com/2073-4425/11/12/1406

Marchal, C., Zhang, J., Zhang, P., Fenwick, P., Steuernagel, B., Adamski, N. M., Boyd, L., McIntosh, R., Wulff, B. B. H., Berry, S., Lagudah, E., & Uauy, C. (2018). BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. *Nature Plants*, *4*(9), 662-668. https://doi.org/10.1038/s41477-018-0236-4

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen, O.-A., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A. J., Sourdille, P., . . . Praud, S. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*(6194), 1250092. https://doi.org/doi:10.1126/science.1250092

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, *12*(2), 213-218. https://doi.org/10.1038/nprot.2016.182

Martinez-Perez, E., Shaw, P., & Moore, G. (2001). The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, *411*(6834), 204-207. https://doi.org/10.1038/35075597

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

Matias, F. I., Galli, G., Correia Granato, I. S., & Fritsche-Neto, R. (2017). Genomic Prediction of Autogamous and Allogamous Plants by SNPs and Haplotypes. *Crop Science*, *57*(6), 2951-2958. https://doi.org/https://doi.org/10.2135/cropsci2017.01.0022

Mayer, M., Hölker, A. C., González-Segovia, E., Bauer, E., Presterl, T., Ouzunova, M., Melchinger, A. E., & Schön, C.-C. (2020). Discovery of beneficial haplotypes for complex traits in maize landraces. *Nature Communications*, *11*(1), 4954. https://doi.org/10.1038/s41467-020-18683-3

MCFADDEN, E. S., & SEARS, E. R. (1946). THE ORIGIN OF TRITICUM SPELTA AND ITS FREE-THRESHING HEXAPLOID RELATIVES*. *Journal of Heredity*, *37*(3), 81-89. https://doi.org/10.1093/oxfordjournals.jhered.a105590

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., & Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.

Mehrab, Z., Mobin, J., Tahmid, I. A., & Rahman, A. (2021). Efficient association mapping from k-mers—An application in finding sex-specific sequences. *PLOS ONE*, *16*(1), e0245058. https://doi.org/10.1371/journal.pone.0245058

Meuwissen, T. H. E., Odegard, J., Andersen-Ranberg, I., & Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genetics Selection Evolution*, *46*(1), 49. https://doi.org/10.1186/1297-9686-46-49

Miki, Y., Yoshida, K., Mizuno, N., Nasuda, S., Sato, K., & Takumi, S. (2019). Origin of wheat B-genome chromosomes inferred from RNA sequencing analysis of leaf transcripts from section Sitopsis species of Aegilops. *DNA Research*, *26*(2), 171-182. https://doi.org/10.1093/dnares/dsy047

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86-97. https://doi.org/https://doi.org/10.1002/widm.53

N'Diaye, A., Haile, J. K., Nilsen, K. T., Walkowiak, S., Ruan, Y., Singh, A. K., Clarke, F. R., Clarke, J. M., & Pozniak, C. J. (2018). Haplotype Loci Under Selection in Canadian Durum Wheat Germplasm Over 60 Years of Breeding: Association With Grain Yield, Quality Traits, Protein Loss, and Plant Height [Original Research]. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01589

Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., Tardieu, F., Charcosset, A., & Nicolas, S. D. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biology*, *19*(1), 318. https://doi.org/10.1186/s12870-019-1926-4

Nei, M., & Tajima, F. (1981). DNA POLYMORPHISM DETECTABLE BY RESTRICTION ENDONUCLEASES. *Genetics*, *97*(1), 145-163. https://doi.org/10.1093/genetics/97.1.145

Nordström, K. J. V., Albani, M. C., James, G. V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., & Schneeberger, K. (2013). Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotechnology*, *31*(4), 325-330. https://doi.org/10.1038/nbt.2515

Pajuste, F.-D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., & Remm, M. (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports*, *7*(1), 2537. https://doi.org/10.1038/s41598-017-02487-5

Pasquariello, M., Berry, S., Burt, C., Uauy, C., & Nicholson, P. (2020). Yield reduction historically associated with the Aegilops ventricosa 7DV introgression is genetically and physically distinct from the eyespot resistance gene Pch1. *Theoretical and Applied Genetics*, *133*(3), 707-717. https://doi.org/10.1007/s00122-019-03502-1

Pasquariello, M., Ham, J., Burt, C., Jahier, J., Paillard, S., Uauy, C., & Nicholson, P. (2017). The eyespot resistance genes Pch1 and Pch2 of wheat are not homoeoloci. *Theoretical and Applied Genetics*, *130*(1), 91-107. https://doi.org/10.1007/s00122-016-2796-x

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris,

M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., . . . Cox, D. R. (2001). Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. *Science*, *294*(5547), 1719-1723. https://doi.org/doi:10.1126/science.1065573

Pattaro, C., Ruczinski, I., Fallin, D. M., & Parmigiani, G. (2008). Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. *BMC Genomics*, *9*(1), 405. https://doi.org/10.1186/1471-2164-9-405

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867-868. https://doi.org/10.1093/bioinformatics/btx699

Peng, J. H., Sun, D., & Nevo, E. (2011). Domestication evolution, genetics and genomics in wheat. *Molecular Breeding*, *28*(3), 281-301. https://doi.org/10.1007/s11032-011-9608-4

Peng, Y., Yan, H., Guo, L., Deng, C., Wang, C., Wang, Y., Kang, L., Zhou, P., Yu, K., Dong, X., Liu, X., Sun, Z., Peng, Y., Zhao, J., Deng, D., Xu, Y., Li, Y., Jiang, Q., Li, Y., . . . Ren, C. (2022). Reference genome assemblies reveal the origin and evolution of allohexaploid oat. *Nature Genetics*, *54*(8), 1248-1258. https://doi.org/10.1038/s41588-022-01127-7

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). *G3 Genes|Genomes|Genetics*, *10*(9), 3047-3060. https://doi.org/10.1534/g3.120.401028

Plekhanova, E., Vishnyakova, M. A., Bulyntsev, S., Chang, P. L., Carrasquilla-Garcia, N., Negash, K., Wettberg, E. v., Noujdina, N., Cook, D. R., Samsonova, M. G., & Nuzhdin, S. V. (2017). Genomic and phenotypic analysis of Vavilov's historic landraces reveals the impact of environment and genomic islands of agronomic traits. *Scientific Reports*, *7*(1), 4816. https://doi.org/10.1038/s41598-017-05087-5

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F., Molnár-Láng, M., Lage, J., Kilian, B., Özkan, H., Waite, D., Dyer, S., Letellier, T., Alaux, M., Russell, J., . . . Barley Legacy for Breeding Improvement, c. (2019). Tracing the ancestry of modern bread wheats. *Nature Genetics*, *51*(5), 905-911. https://doi.org/10.1038/s41588-019-0393-z

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C. C., & Simianer, H. (2019). HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. *Genetics*, *212*(4), 1045-1061. https://doi.org/10.1534/genetics.119.302283

Pritchard, J. K., & Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics*, *69*(1), 1-14. https://doi.org/10.1086/321275

Przewieslik-Allen, A. M., Wilkinson, P. A., Burridge, A. J., Winfield, M. O., Dai, X., Beaumont, M., King, J., Yang, C.-y., Griffiths, S., Wingen, L. U., Horsnell, R., Bentley, A. R., Shewry, P., Barker, G. L. A., & Edwards, K. J. (2021). The role of gene flow and chromosomal instability in shaping the bread wheat genome. *Nature Plants*, *7*(2), 172-183. https://doi.org/10.1038/s41477-020-00845-2

Rabanus-Wallace, M. T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., Baez, M., Houben, A., Mayer, K. F. X., Guo, L., Poland, J., Pozniak, C. J., Walkowiak, S., Melonek, J., Praz, C. R., Schreiber, M., Budak, H., Heuberger, M., Steuernagel, B., . . . Stein, N. (2021). Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. *Nature Genetics*, *53*(4), 564-573. https://doi.org/10.1038/s41588-021-00807-0

Rahimi, Y., Bihamta, M. R., Taleei, A., Alipour, H., & Ingvarsson, P. K. (2019). Genome-wide association study of agronomic traits in bread wheat reveals novel putative alleles for future breeding programs. *BMC Plant Biology*, *19*(1), 541. https://doi.org/10.1186/s12870-019-2165-4

Rahman, A., Hallgrímsdóttir, I., Eisen, M., & Pachter, L. (2018). Association mapping from sequencing reads using k-mers. *eLife*, *7*, e32920. https://doi.org/10.7554/eLife.32920

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., Ji, S.-G., Demir, G., Li, L., Toptaş, B. Ç., Dolgoborodov, A., Pollex, B., Spulber, I., Glotova, I., Kómár, P., . . . Kural, D. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, *51*(2), 354-362. https://doi.org/10.1038/s41588-018-0316-4

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., Davey, M., Jacobs, J., van Ex, F., Pasha, A., Khedikar, Y., Robinson, S. J., Cory, A. T., Florio, T., Concia, L., Juery, C., Schoonbeek, H., Steuernagel, B., Xiang, D., . . . Tan, Y. (2018). The transcriptional landscape of polyploid wheat. *Science*, *361*(6403), eaar6089. https://doi.org/doi:10.1126/science.aar6089

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., & He, Z. (2017). Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. *Molecular Plant*, *10*(8), 1047-1064. https://doi.org/10.1016/j.molp.2017.06.008

Reif, J. C., Zhang, P., Dreisigacker, S., Warburton, M. L., van Ginkel, M., Hoisington, D., Bohn, M., & Melchinger, A. E. (2005). Wheat genetic diversity trends during domestication and breeding. *Theoretical and Applied Genetics*, *110*(5), 859-864. https://doi.org/10.1007/s00122-004-1881-8

Rey, M.-D., Martín, 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

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., . . . Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737-746. https://doi.org/10.1038/s41586-021-03451-0

Robbins, M. D., Sim, S.-C., Yang, W., Van Deynze, A., van der Knaap, E., Joobeur, T., & Francis, D. M. (2010). Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. *Journal of Experimental Botany*, *62*(6), 1831-1845. https://doi.org/10.1093/jxb/erq367

Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C., Falentin, C., Boideau, F., Boutte, J., Delourme, R., Deniot, G., Engelen, S., de Carvalho, J. F., Lemainque, A., Maillet, L., Morice, J., Wincker, P., Denoeud, F., Chèvre, A.-M., & Aury, J.-M. (2020). Long-read assembly of the Brassica napus reference genome Darmor-bzh. *GigaScience*, *9*(12). https://doi.org/10.1093/gigascience/giaa137

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Roussel, V., Koenig, J., Beckert, M., & Balfourier, F. (2004). Molecular diversity in French bread wheat accessions related to temporal trends and breeding programmes. *Theoretical and Applied Genetics*, *108*(5), 920-930. https://doi.org/10.1007/s00122-003-1502-y

Roussel, V., Leisova, L., Exbrayat, F., Stehno, Z., & Balfourier, F. (2005). SSR allelic diversity changes in 480 European bread wheat varieties released from 1840 to 2000. *Theoretical and Applied Genetics*, *111*(1), 162-170. https://doi.org/10.1007/s00122-005-2014-8

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832-837. https://doi.org/10.1038/nature01140

Saintenac, C., Zhang, W., Salcedo, A., Rouse, M. N., Trick, H. N., Akhunov, E., & Dubcovsky, J. (2013). Identification of Wheat Gene Sr35 That Confers Resistance to Ug99 Stem Rust Race Group. *Science*, *341*(6147), 783-786. https://doi.org/doi:10.1126/science.1239022

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., & Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nature Reviews Genetics*, *3*(6), 429-441. https://doi.org/10.1038/nrg817

Sallam, A. H., Conley, E., Prakapenka, D., Da, Y., & Anderson, J. A. (2020). Improving Prediction Accuracy Using Multi-allelic Haplotype Prediction and Training Population Optimization in Wheat. *G3 Genes|Genomes|Genetics*, *10*(7), 2265-2273. https://doi.org/10.1534/g3.120.401165

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., Campos, J., Dreher, K., Payne, T., Marshall, D., Kilian, B., Milne, I., Raubach, S., Shaw, P., Stephen, G., Carling, J., Pierre, C. S., Burgueño, J., Crosa, J., . . . Pixley, K. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nature Communications*, *11*(1), 4572. https://doi.org/10.1038/s41467-020-18404-w

Sato, K., Abe, F., Mascher, M., Haberer, G., Gundlach, H., Spannagl, M., Shirasawa, K., & Isobe, S. (2021). Chromosome-scale genome assembly of the

transformation-amenable common wheat cultivar 'Fielder'. *DNA Research*, *28*(3). https://doi.org/10.1093/dnares/dsab008
Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *3*(3), 430-439. https://doi.org/10.1038/s41559-018-0793-y
Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings in Functional Genomics*, *13*(4), 296-307. https://doi.org/10.1093/bfgp/elu016
Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H. S., & Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, *7*(2), 127-140. https://doi.org/https://doi.org/10.1016/j.cj.2018.07.006
Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., & Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, *10*(9), R98. https://doi.org/10.1186/gb-2009-10-9-r98
Scholten, O. E., van Kaauwen, M. P. W., Shahin, A., Hendrickx, P. M., Keizer, L. C. P., Burger, K., van Heusden, A. W., van der Linden, C. G., & Vosman, B. (2016). SNP-markers in Allium species to facilitate introgression breeding in onion. *BMC Plant Biology*, *16*(1), 187. https://doi.org/10.1186/s12870-016-0879-0
Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y., Beukert, U., Serfling, A., Himmelbach, A., Fuchs, J., Oppermann, M., Weise, S., Boeven, P. H. G., Schacht, J., Longin, C. F. H., Kollers, S., Pfeiffer, N., Korzun, V., . . . Reif, J. C. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nature Genetics*, *54*(10), 1544-1552. https://doi.org/10.1038/s41588-022-01189-7
Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., & Dreisigacker, S. (2020). Incorporating Genome-Wide Association Mapping Results Into Genomic Prediction Models for Grain Yield and Yield Stability in CIMMYT Spring Bread Wheat [Original Research]. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00197
Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., Li, Y., Gao, H., Wang, T., Liu, X., Zhang, H., Zhang, Y., Cao, S., Yu, X., Zhang, B., . . . Qian, Q. (2022). A super pan-genomic landscape of rice. *Cell Research*, *32*(10), 878-896. https://doi.org/10.1038/s41422-022-00685-z
Shaw, P. D., Graham, M., Kennedy, J., Milne, I., & Marshall, D. F. (2014). Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics*, *15*(1), 259. https://doi.org/10.1186/1471-2105-15-259
Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345-353. https://doi.org/10.1038/nature24286
Shimmura, K., Kato, Y., & Kawahara, Y. (2020). Bivartect: accurate and memory-saving breakpoint detection by direct read comparison. *Bioinformatics*, *36*(9), 2725-2730. https://doi.org/10.1093/bioinformatics/btaa059
Shorinola, O., Simmonds, J., Wingen, L. U., & Uauy, C. (2022). Trend, population structure, and trait mapping from 15 years of national varietal trials of UK winter

wheat. *G3 Genes|Genomes|Genetics*, *12*(2), jkab415. https://doi.org/10.1093/g3journal/jkab415
Sibbesen, J. A., Maretty, L., Krogh, A., & The Danish Pan-Genome, C. (2018). Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, *50*(7), 1054-1059. https://doi.org/10.1038/s41588-018-0145-5
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489. https://doi.org/10.1038/nature16961
Singh, D., Wang, X., Kumar, U., Gao, L., Noor, M., Imtiaz, M., Singh, R. P., & Poland, J. (2019). High-Throughput Phenotyping Enabled Genetic Dissection of Crop Lodging in Wheat [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00394
Singh, N., Wu, S., Tiwari, V., Sehgal, S., Raupp, J., Wilson, D., Abbasov, M., Gill, B., & Poland, J. (2019). Genomic Analysis Confirms Population Structure and Identifies Inter-Lineage Hybrids in Aegilops tauschii [Original Research]. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00009
Sinha, P., Singh, V. K., Saxena, R. K., Khan, A. W., Abbai, R., Chitikineni, A., Desai, A., Molla, J., Upadhyaya, H. D., Kumar, A., & Varshney, R. K. (2020). Superior haplotypes for haplotype-based breeding for drought tolerance in pigeonpea (Cajanus cajan L.). *Plant biotechnology journal*, *18*(12), 2482-2490. https://doi.org/https://doi.org/10.1111/pbi.13422
Smith, J. S. C., & Smith, O. S. (1992). Fingerprinting Crop Varieties. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 47, pp. 85-140). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2113(08)60489-7
Sohn, J.-i., & Nam, J.-W. (2016). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23-40. https://doi.org/10.1093/bib/bbw096
Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, *38*, 1409-1438.
Soleimani, B., Lehnert, H., Babben, S., Keilwagen, J., Koch, M., Arana-Ceballos, F. A., Chesnokov, Y., Pshenichnikova, T., Schondelmaier, J., Ordon, F., Börner, A., & Perovic, D. (2022). Genome wide association study of frost tolerance in wheat. *Scientific Reports*, *12*(1), 5275. https://doi.org/10.1038/s41598-022-08706-y
Standage, D. S., Brown, C. T., & Hormozdiari, F. (2019). Kevlar: A Mapping-Free Framework for Accurate Discovery of De Novo Variants. *iScience*, *18*, 28-36. https://doi.org/https://doi.org/10.1016/j.isci.2019.07.032
Suchecki, R., Sandhu, A., Deschamps, S., Llaca, V., Wolters, P., Watson-Haigh, N. S., Pallotta, M., Whitford, R., & Baumann, U. (2019). LNISKS: Reference-free mutation identification for large and complex crop genomes. *bioRxiv*, 580829.
Sukumaran, S., Lopes, M., Dreisigacker, S., & Reynolds, M. (2018). Genetic analysis of multi-environmental spring wheat trials identifies genomic regions for locus-specific trade-offs for grain weight and grain number. *Theoretical and Applied Genetics*, *131*(4), 985-998. https://doi.org/10.1007/s00122-017-3037-7

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N., & Chen, F. (2020). The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant biotechnology journal*, *18*(6), 1354-1360. https://doi.org/https://doi.org/10.1111/pbi.13361

Tagle, A. G., Chuma, I., & Tosa, Y. (2015). Rmg7, a New Gene for Resistance to Triticum Isolates of Pyricularia oryzae Identified in Tetraploid Wheat. *Phytopathology®*, *105*(4), 495-499. https://doi.org/10.1094/phyto-06-14-0182-r

Tanno, K.-i., & Willcox, G. (2006). How Fast Was Wild Wheat Domesticated? *Science*, *311*(5769), 1886-1886. https://doi.org/doi:10.1126/science.1124635

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. *Nature*, *618*(7965), 616-624. https://doi.org/10.1038/s41586-023-06139-9

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1), e20077. https://doi.org/https://doi.org/10.1002/tpg2.20077

Udall, J. A., & Dawe, R. K. (2017). Is It Ordered Correctly? Validating Genome Assemblies by Optical Mapping. *The Plant Cell*, *30*(1), 7-14. https://doi.org/10.1105/tpc.17.00514

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666-681. https://doi.org/https://doi.org/10.1016/j.tig.2018.05.008

van Heyningen, V. (2019). Genome sequencing—the dawn of a game-changing era. *Heredity*, *123*(1), 58-66. https://doi.org/10.1038/s41437-019-0226-y

Vikram, P., Franco, J., Burgueño-Ferreira, J., Li, H., Sehgal, D., Saint Pierre, C., Ortiz, C., Sneller, C., Tattaris, M., Guzman, C., Sansaloni, C. P., Ellis, M., Fuentes-Davila, G., Reynolds, M., Sonder, K., Singh, P., Payne, T., Wenzl, P., Sharma, A., . . . Singh, S. (2016). Unlocking the genetic diversity of Creole wheats. *Scientific Reports*, *6*(1), 23092. https://doi.org/10.1038/srep23092

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5), 534-540. https://doi.org/10.1038/s41588-020-0612-7

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669-686. https://doi.org/10.1007/s00122-018-3270-8

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., & Thambugala, D. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, *588*(7837), 277-283.

Wang, J., Luo, M.-C., Chen, Z., You, F. M., Wei, Y., Zheng, Y., & Dvorak, J. (2013). Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of

hexaploid wheat. *New Phytologist*, *198*(3), 925-937. https://doi.org/https://doi.org/10.1111/nph.12164

Wang, L., Zhu, T., Rodriguez, J. C., Deal, K. R., Dubcovsky, J., McGuire, P. E., Lux, T., Spannagl, M., Mayer, K. F. X., Baldrich, P., Meyers, B. C., Huo, N., Gu, Y. Q., Zhou, H., Devos, K. M., Bennetzen, J. L., Unver, T., Budak, H., Gulick, P. J., . . . Dvorak, J. (2021). Aegilops tauschii genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. *G3 (Bethesda)*, *11*(12). https://doi.org/10.1093/g3journal/jkab325

Wang, S., Asuke, S., Vy, T. T. P., Inoue, Y., Chuma, I., Win, J., Kato, K., & Tosa, Y. (2018). A New Resistance Gene in Combination with Rmg8 Confers Strong Resistance Against Triticum Isolates of Pyricularia oryzae in a Common Wheat Landrace. *Phytopathology®*, *108*(11), 1299-1306. https://doi.org/10.1094/phyto-12-17-0400-r

Wang, S., Qian, Y.-Q., Zhao, R.-P., Chen, L.-L., & Song, J.-M. (2022). Graph-based pan-genomes: increased opportunities in plant genomics. *Journal of Experimental Botany*, *74*(1), 24-39. https://doi.org/10.1093/jxb/erac412

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., Consortium, I. W. G. S., Lillemo, M., Mather, D., . . . Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant biotechnology journal*, *12*(6), 787-796. https://doi.org/https://doi.org/10.1111/pbi.12183

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., . . . Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, *557*(7703), 43-49. https://doi.org/10.1038/s41586-018-0063-9

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348-1365. https://doi.org/10.1038/s41587-021-01108-x

Wei, T., van Treuren, R., Liu, X., Zhang, Z., Chen, J., Liu, Y., Dong, S., Sun, P., Yang, T., Lan, T., Wang, X., Xiong, Z., Liu, Y., Wei, J., Lu, H., Han, S., Chen, J. C., Ni, X., Wang, J., . . . Liu, H. (2021). Whole-genome resequencing of 445 Lactuca accessions reveals the domestication history of cultivated lettuce. *Nature Genetics*, *53*(5), 752-760. https://doi.org/10.1038/s41588-021-00831-0

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155-1162. https://doi.org/10.1038/s41587-019-0217-9

Wilhelm, E. P., Boulton, M. I., Barber, T. E. S., Greenland, A. J., & Powell, W. (2013). Genotype analysis of the wheat semidwarf Rht-B1b and Rht-D1b ancestral lineage. *Plant Breeding*, *132*(6), 539-545. https://doi.org/https://doi.org/10.1111/pbr.12099

Wimalanathan, K., & Lawrence-Dill, C. J. (2021). Gene Ontology Meta Annotator for Plants (GOMAP). *Plant Methods*, *17*(1), 54. https://doi.org/10.1186/s13007-021-00754-1

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L. A., Benbow, H. R., Wilkinson, P. A., Coghill, J., Waterfall, C., Davassi, A., Scopes, G., Pirani, A., Webster, T., Brew, F., Bloor, C., King, J., West, C., Griffiths, S., King, I., Bentley, A. R., & Edwards, K. J. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant biotechnology journal*, *14*(5), 1195-1206. https://doi.org/https://doi.org/10.1111/pbi.12485

Winfield, M. O., Allen, A. M., Wilkinson, P. A., Burridge, A. J., Barker, G. L. A., Coghill, J., Waterfall, C., Wingen, L. U., Griffiths, S., & Edwards, K. J. (2018). High-density genotyping of the A.E. Watkins Collection of hexaploid landraces identifies a large molecular diversity compared to elite bread wheat. *Plant biotechnology journal*, *16*(1), 165-175. https://doi.org/https://doi.org/10.1111/pbi.12757

Wingen, L. U., Orford, S., Goram, R., Leverington-Waite, M., Bilham, L., Patsiou, T. S., Ambrose, M., Dicks, J., & Griffiths, S. (2014). Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat. *Theoretical and Applied Genetics*, *127*(8), 1831-1842. https://doi.org/10.1007/s00122-014-2344-5

Wingen, L. U., West, C., Leverington-Waite, M., Collier, S., Orford, S., Goram, R., Yang, C.-Y., King, J., Allen, A. M., Burridge, A., Edwards, K. J., & Griffiths, S. (2017). Wheat Landrace Genome Diversity. *Genetics*, *205*(4), 1657-1676. https://doi.org/10.1534/genetics.116.194688

Won, S., Park, J.-E., Son, J.-H., Lee, S.-H., Park, B. H., Park, M., Park, W.-C., Chai, H.-H., Kim, H., Lee, J., & Lim, D. (2020). Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium [Original Research]. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00134

Wulff, B. B. H., & Moscou, M. J. (2014). Strategies for transferring resistance into wheat: from wide crosses to GM cassettes [Review]. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00692

Würschum, T., Langer, S. M., Longin, C. F. H., Tucker, M. R., & Leiser, W. L. (2017). A modern Green Revolution gene for reduced height in wheat. *The Plant Journal*, *92*(5), 892-903. https://doi.org/https://doi.org/10.1111/tpj.13726

Würschum, T., Weiß, T. M., Renner, J., Friedrich Utz, H., Gierl, A., Jonczyk, R., Römisch-Margl, L., Schipprack, W., Schön, C.-C., Schrag, T. A., Leiser, W. L., & Melchinger, A. E. (2022). High-resolution association mapping with libraries of immortalized lines from ancestral landraces. *Theoretical and Applied Genetics*, *135*(1), 243-256. https://doi.org/10.1007/s00122-021-03963-3

Xue, S., Kolmer, J. A., Wang, S., & Yan, L. (2018). Mapping of Leaf Rust Resistance Genes and Molecular Characterization of the 2NS/2AS Translocation in the Wheat Cultivar Jagger. *G3 Genes|Genomes|Genetics*, *8*(6), 2059-2065. https://doi.org/10.1534/g3.118.200058

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., Huang, J., Deng, T., Luo, J., He, L., Wang, Y., Xu, P., Peng, Y., Shi, Z., Lan, L., Ma, Z., Yang, X., Zhang, Q., Bai,

M., . . . Yan, J. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature Genetics*, *51*(6), 1052-1059. https://doi.org/10.1038/s41588-019-0427-6

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.-c., Hu, L., Yamasaki, M., Yoshida, S., Kitano, H., Hirano, K., & Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, *48*(8), 927-934. https://doi.org/10.1038/ng.3596

Zhang, J., Debernardi, J. M., Burguener, G. F., Choulet, F., Paux, E., O'Connor, L., Enk, J., & Dubcovsky, J. A second-generation capture panel for cost-effective sequencing of genome regulatory regions in wheat and relatives. *The Plant Genome*, *n/a*(n/a), e20296. https://doi.org/https://doi.org/10.1002/tpg2.20296

Zhang, J., Gizaw, S. A., Bossolini, E., Hegarty, J., Howell, T., Carter, A. H., Akhunov, E., & Dubcovsky, J. (2018). Identification and validation of QTL for grain yield and plant water status under contrasting water treatments in fall-sown spring wheats. *Theoretical and Applied Genetics*, *131*(8), 1741-1759. https://doi.org/10.1007/s00122-018-3111-9

Zhang, W., Olson, E., Saintenac, C., Rouse, M., Abate, Z., Jin, Y., Akhunov, E., Pumphrey, M., & Dubcovsky, J. (2010). Genetic Maps of Stem Rust Resistance Gene Sr35 in Diploid and Hexaploid Wheat. *Crop Science*, *50*(6), 2464-2474. https://doi.org/https://doi.org/10.2135/cropsci2010.04.0202

Zhao, X., Guo, Y., Kang, L., Bi, A., Xu, D., Zhang, Z., Zhang, J., Yang, X., Xu, J., & Xu, S. (2022). Population genomics unravels the Holocene history of Triticum-Aegilops species. *bioRxiv*.

Zhou, Y., Bai, S., Li, H., Sun, G., Zhang, D., Ma, F., Zhao, X., Nie, F., Li, J., Chen, L., Lv, L., Zhu, L., Fan, R., Ge, Y., Shaheen, A., Guo, G., Zhang, Z., Ma, J., Liang, H., . . . Song, C.-P. (2021). Introgressing the Aegilops tauschii genome into wheat as a basis for cereal improvement. *Nature Plants*, *7*(6), 774-786. https://doi.org/10.1038/s41477-021-00934-w

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., & Gaut, B. S. (2019). The population genetics of structural variants in grapevine domestication. *Nature Plants*, *5*(9), 965-979. https://doi.org/10.1038/s41477-019-0507-8

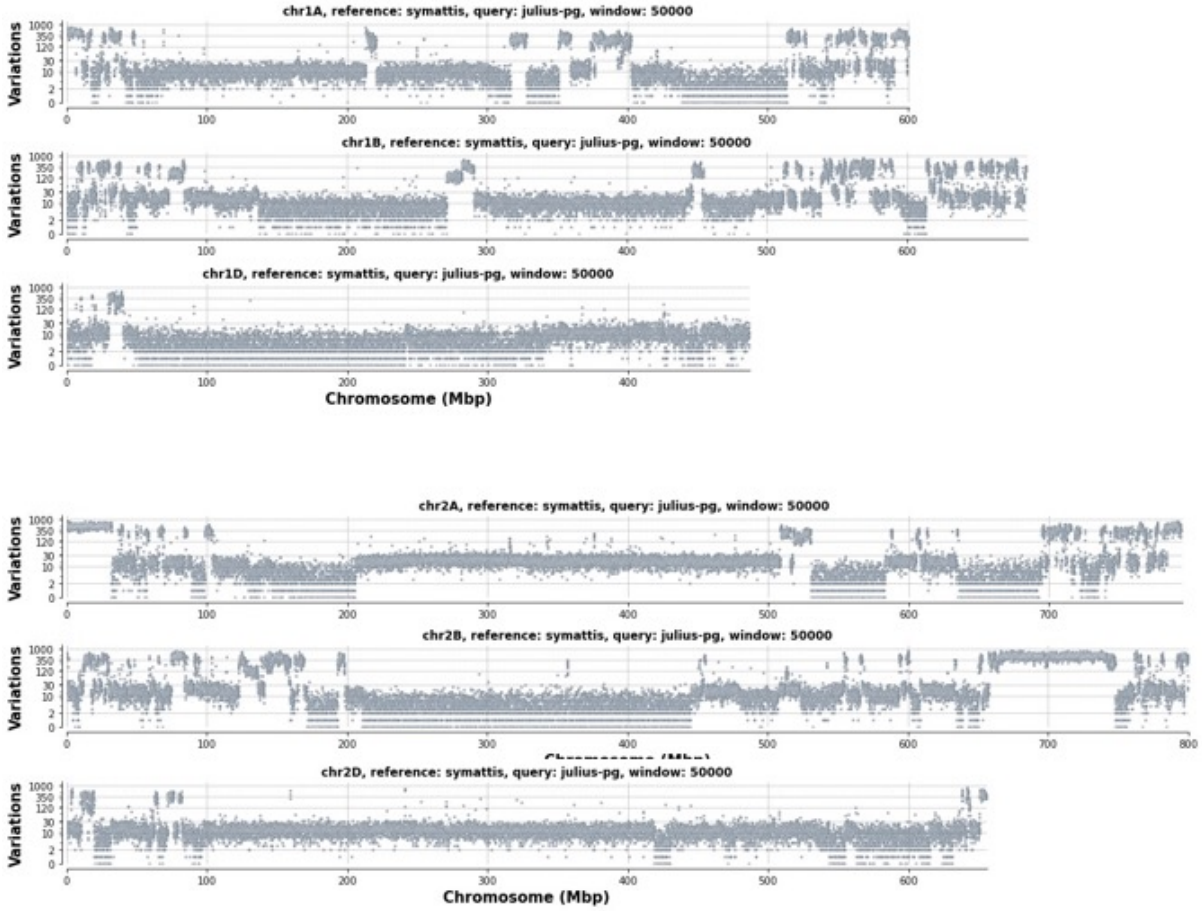
Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y.-g., Liu, S., Jiao, C., Lu, H., Wang, J., Yin, C., Jiao, Y., & Lu, F. (2020). Triticum population sequencing provides insights into wheat adaptation. *Nature Genetics*, *52*(12), 1412-1422. https://doi.org/10.1038/s41588-020-00722-w

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., . . . Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*(4), 408-414. https://doi.org/10.1038/nbt.3096

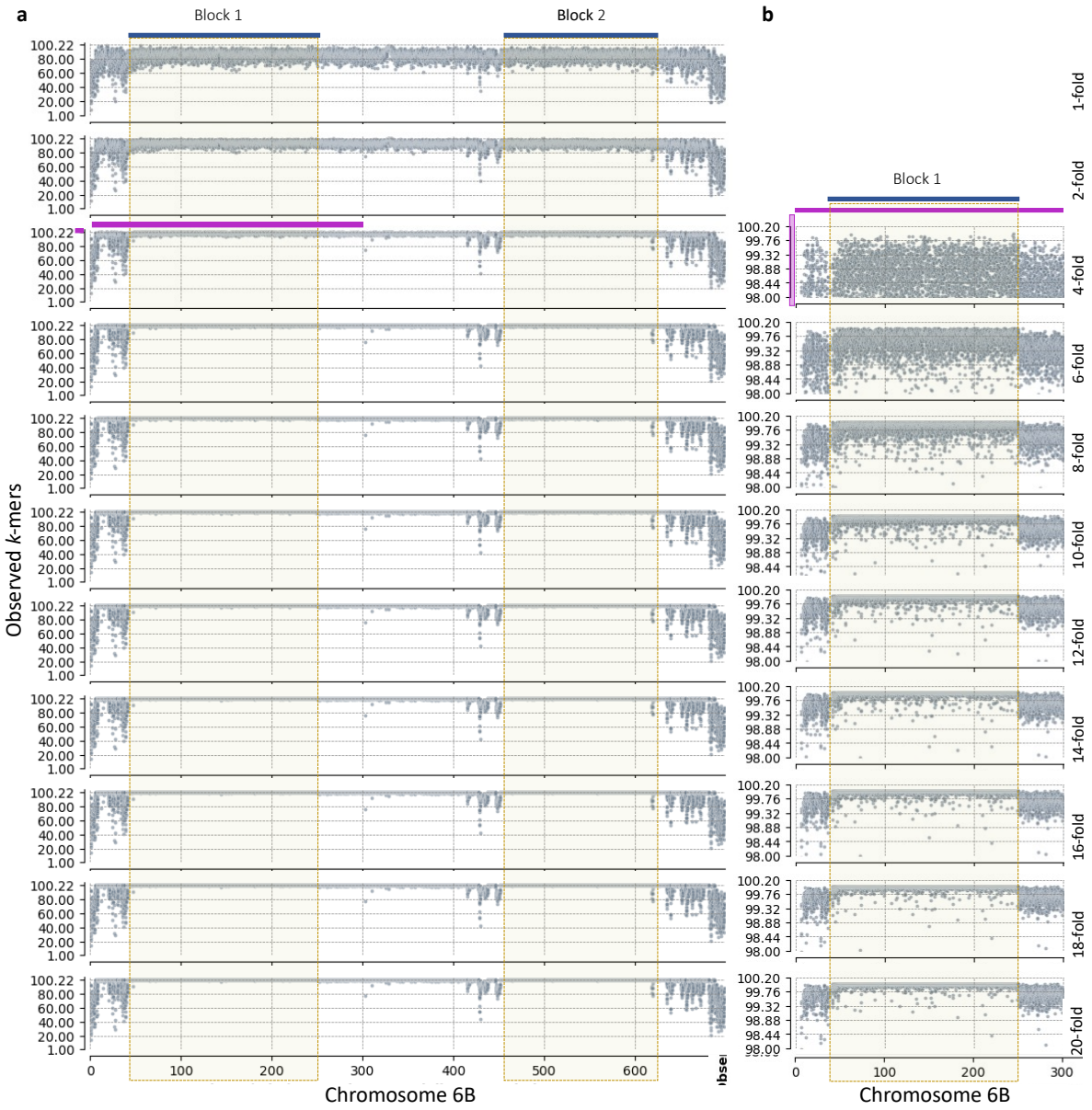
Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., Lau, A. K., Röhling, S., Choi, J. J., Waterman, M. S., Comin, M., Kim, S.-H., Vinga, S., Almeida, J. S., Chan, C. X., James, B. T., Sun, F., Morgenstern, B., & Karlowski,

W. M. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, *20*(1), 144. https://doi.org/10.1186/s13059-019-1755-7

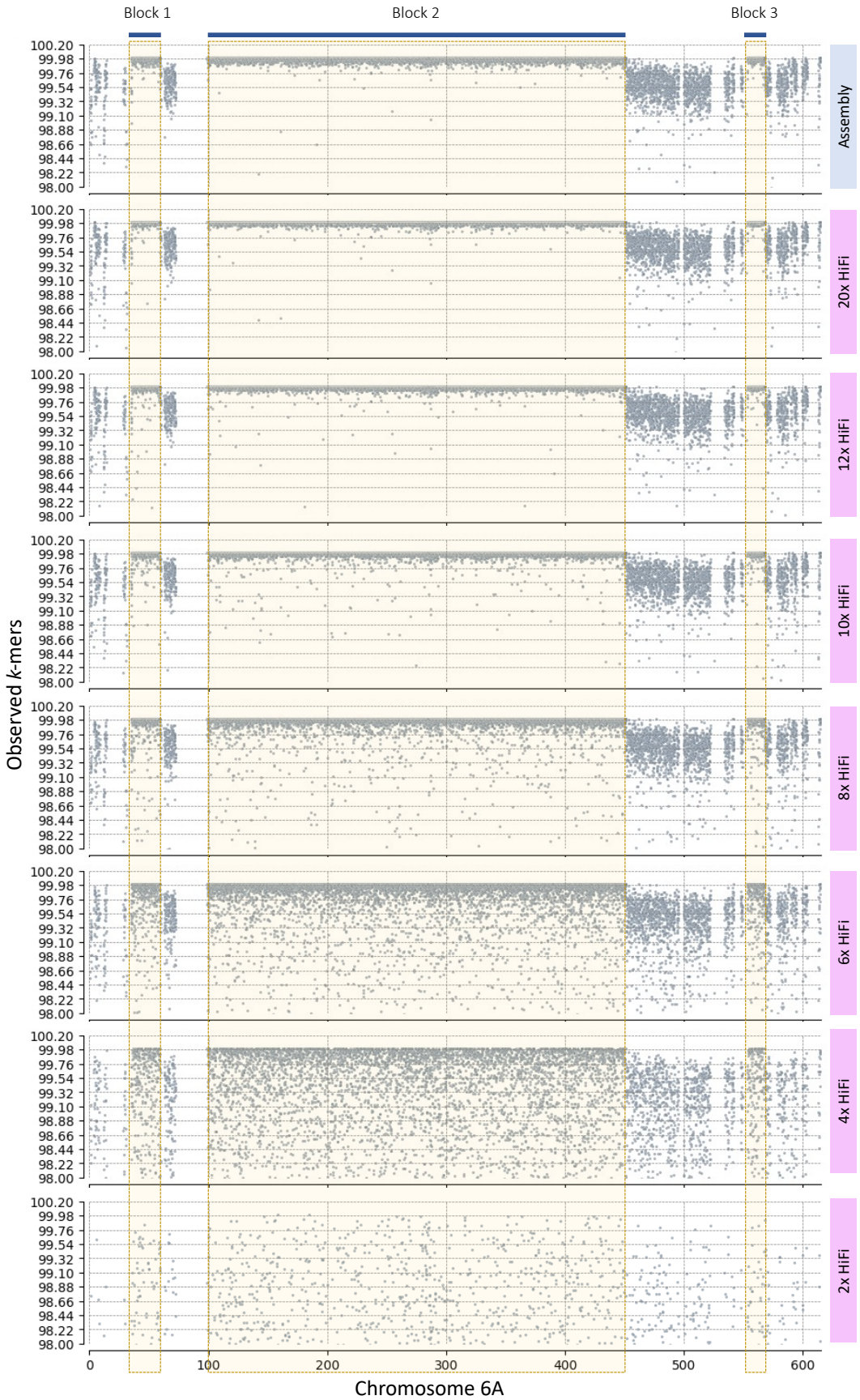
7. Supplementals



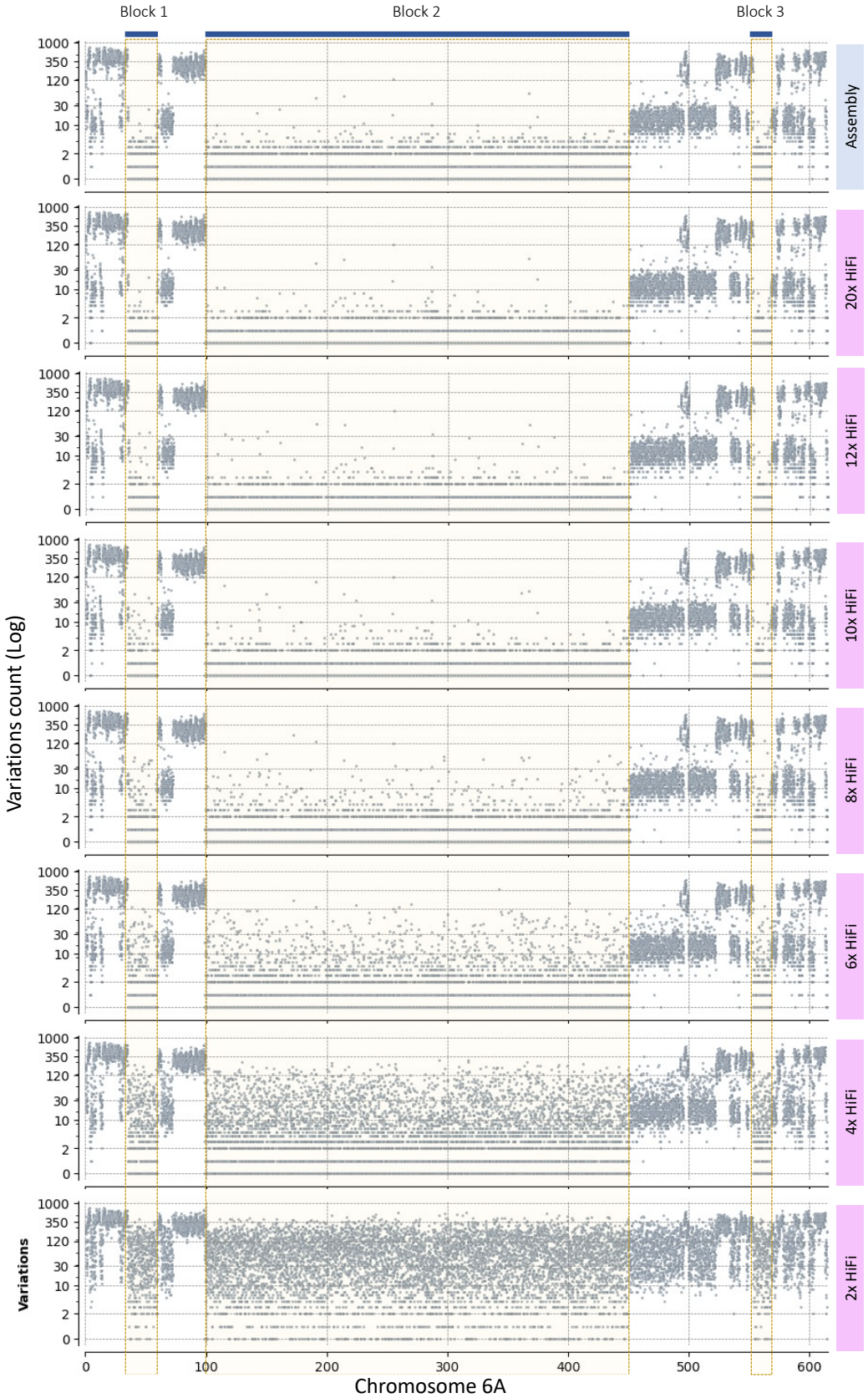
Supplemental Fig. S2. 1. Mattis vs Julius IBSpy *variations fingerprint* across the whole genome.



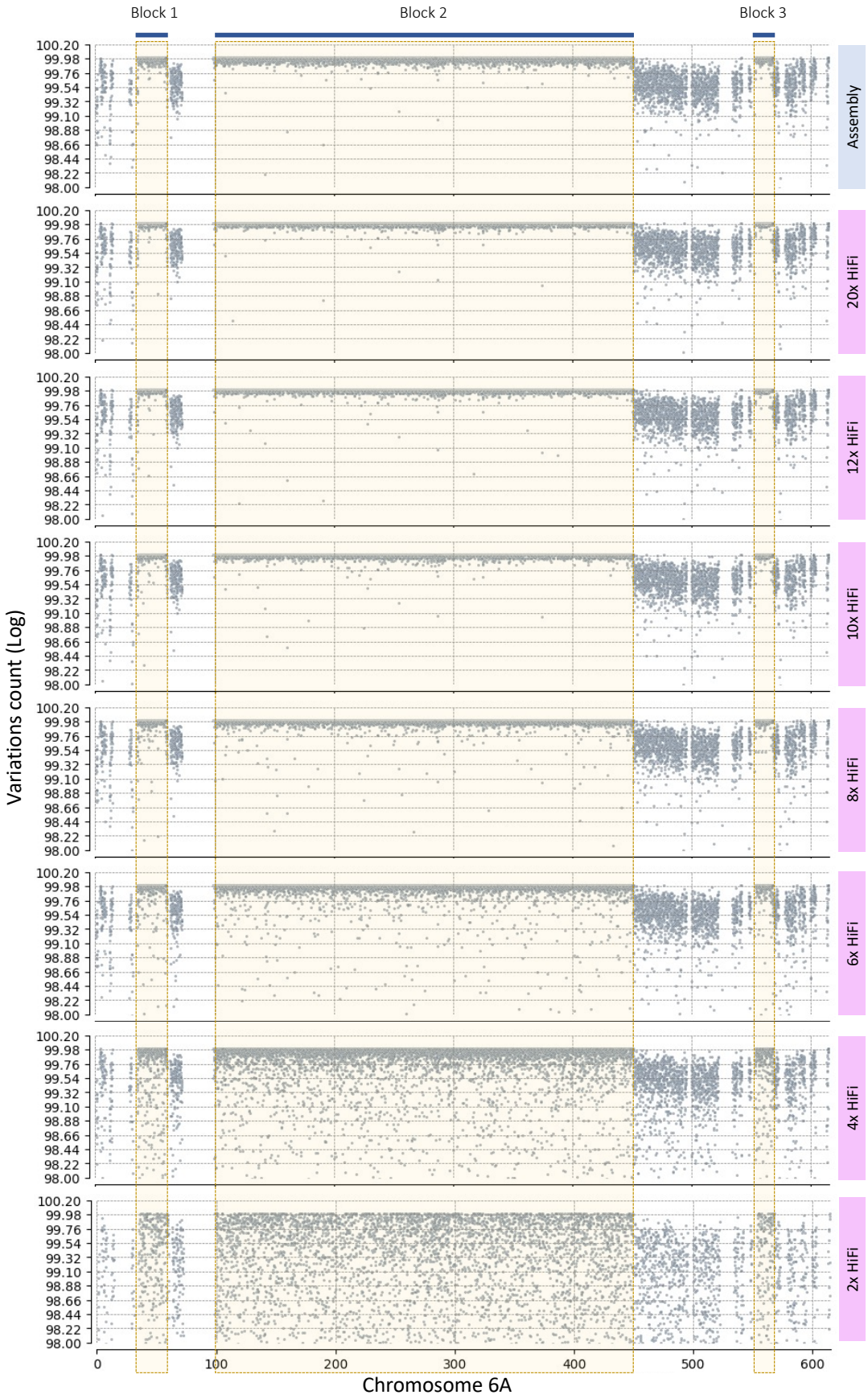
Supplemental Fig. S2. 2. Observed *k*-mers keeping unique *k*-mers.


Block 1
Block 2
Block 3
Assembly
20x HiFi
12x HiFi
10x HiFi
8x HiFi
6x HiFi
4x HiFi
2x HiFi
Observed k-mers
100.20
99.98
99.76
99.54
99.32
99.10
98.88
98.66
98.44
98.22
98.00
0
100
200
300
400
500
600
Chromosome 6A

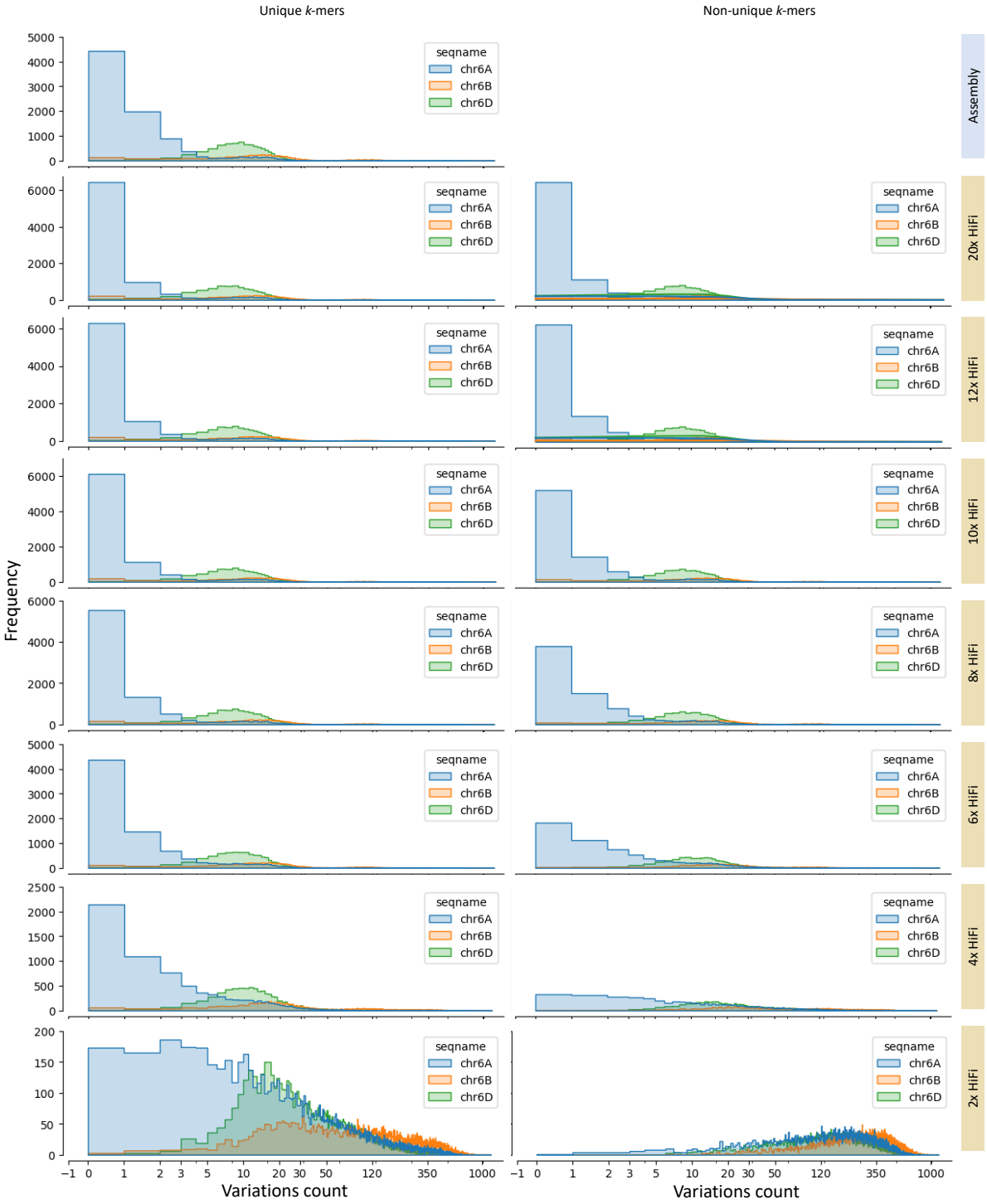
Supplemental Fig. S2. 3. Observed *k*-mers removing unique *k*-mers using HiFi reads at different sequencing coverage.


Block 1
Block 2
Block 3
Assembly
20x HiFi
12x HiFi
10x HiFi
8x HiFi
6x HiFi
4x HiFi
2x HiFi
Variations count (Log)
Variations
1000
350
120
30
10
2
0
0
100
200
300
400
500
600
Chromosome 6A

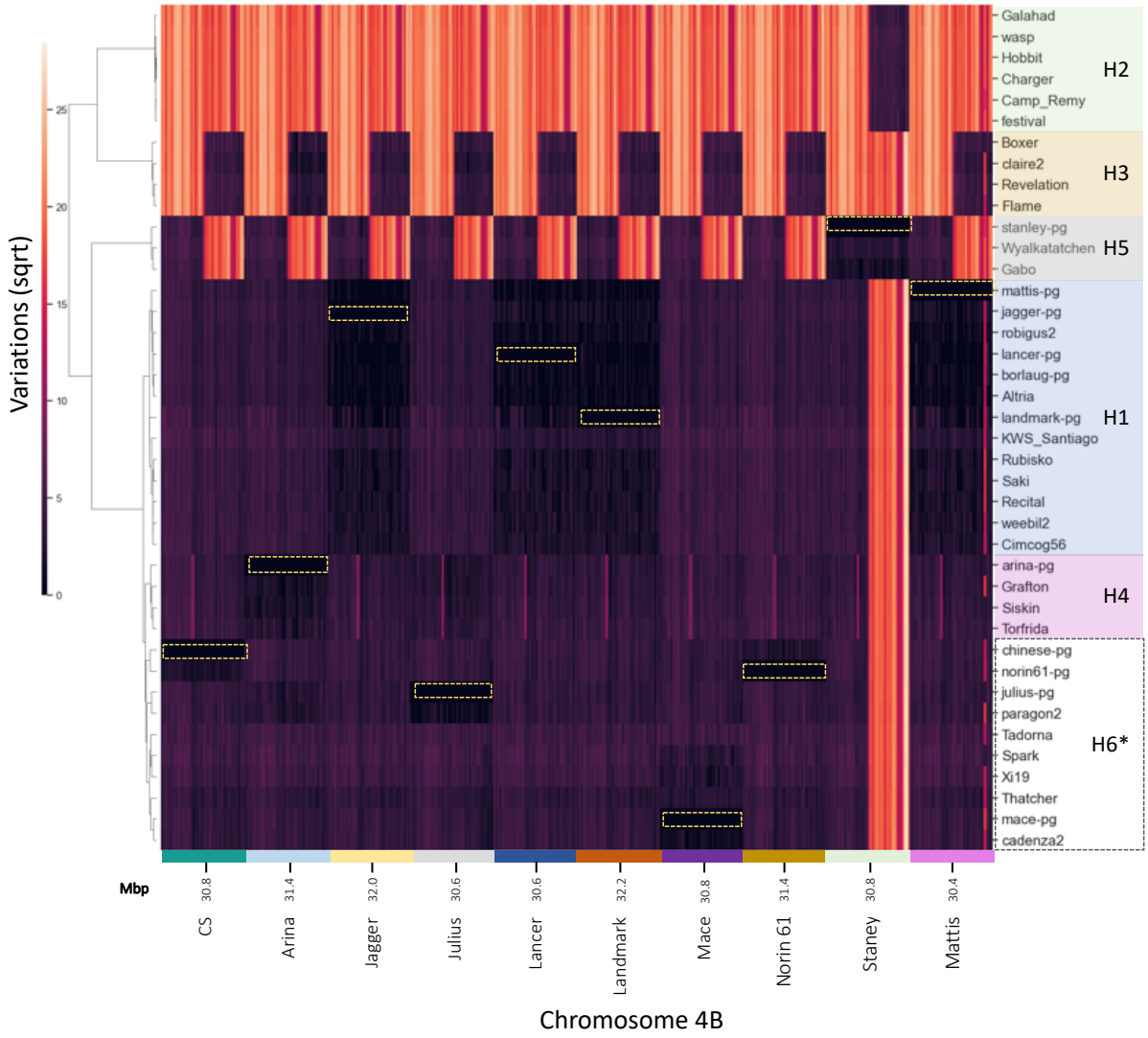
Supplemental Fig. S2. 4. HiFi reads at different sequence coverage Mattis and Kariega keeping unique *k*-mer.


Block 1
Block 2
Block 3
Assembly
20x HiFi
12x HiFi
10x HiFi
8x HiFi
6x HiFi
4x HiFi
2x HiFi
Variations count (Log)
100.20
99.98
99.76
99.54
99.32
99.10
98.88
98.66
98.44
98.22
98.00
0
100
200
300
400
500
600
Chromosome 6A

Supplemental Fig. S2. 5. Observed *k*-mers of HiFi reads (keeping unique *k*-mers) at different sequence coverage.



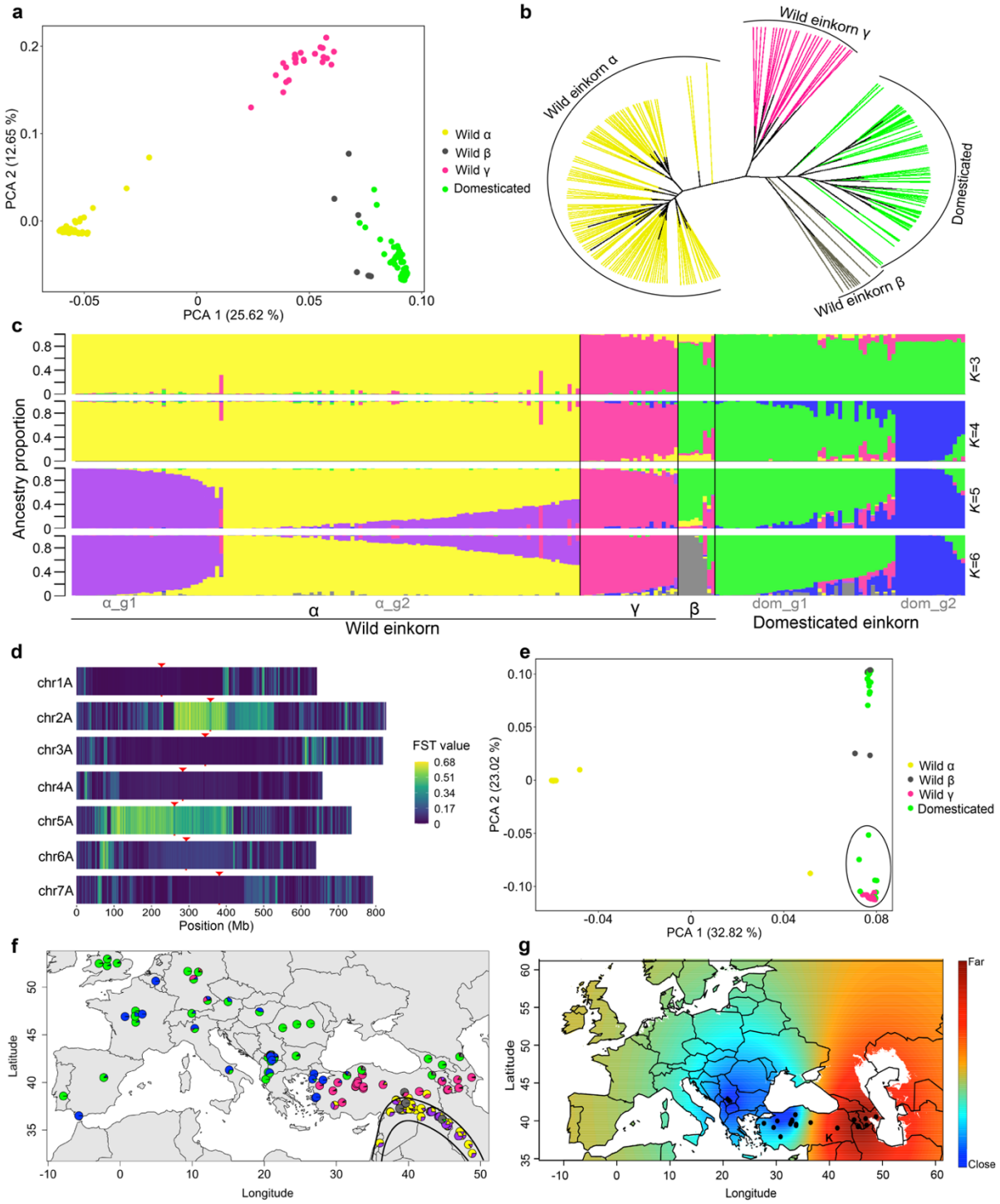
Supplemental Fig. S2. 6. IBSpy variations distributions from HiFi raw reads at different sequence coverage from Kariega vs Mattis reference comparison including or removing unique *k*-mers.



Supplemental Fig. S3. 1. Rht_B1 Multi-reference cluster map of the Rht-B1 locus ± 1 Mbp.

```
TraesSYM2B03G01095480  MEHKASTARSDLEQMLVDEKIEPKALPLSLLKDITDDFSDDREIGRGGFAVVYKGILGDR  60
                       MEHK ST +SDLEQ+L+ E  EPKALPLSLL+DIT+ FSDD+EIGRGGFAVVYKG L +R
TraesSYM2A03G00828360  MEHKTSTTQSDLEQILLSETAEPKALPLSLLEDITNGFSDDQEIGRGGFAVVYKGTLCNR  60

TraesSYM2B03G01095480  LIAVKKLSKAYMHETEFDREIECLMRAKHRNVVRFLGYCDDRQRSAKTYDGKLIMADVQQ  120
                        +AVK++S A M ET+F RE+ECLM+ KH+NVVRFLGYC DRQ +   Y+GKL+MADV Q
TraesSYM2A03G00828360  AVAVKRMSNALMDETKFHREVECLMQVKHKNVVRFLGYCADRQGNMARYNGKLVMADVHQ  120

TraesSYM2B03G01095480  RLLCFEYIPKGSLDLYLTDAHREWDTCYKIIKGICHGLQYLHDNRIIHLDLKPANILLDN  180
                       RLLCFEYIPKG LD Y+++A+REW TCYKIIK IC GLQYLHDN IIHLDLKPANILLD+
TraesSYM2A03G00828360  RLLCFEYIPKGGLDKYISNANREWGTCYKIIKAICEGLQYLHDNHIIHLDLKPANILLDD  180

TraesSYM2B03G01095480  DMVPKITDFGLSRCLDENQSQVLTKNISGTTGYLAPERYEGSGITRSGDLYSLGIIIMEI  240
                       +M PKI DFGLSRC DENQS+ +T+ I GT GYLAPE  EG  I RS DLYSLG+II+EI
TraesSYM2A03G00828360  NMEPKIADFGLSRCFDENQSRDITETILGTMGYLAPEVREGGVIARSADLYSLGVIILEI  240

TraesSYM2B03G01095480  LTGQKGHQTSEDVLESWSDRLERSQRDTLYEQIRVCYEIALNCIQFNPKDRPASARDMID  300
                       LTGQKG+Q   +VL SWSDRLERSQRDTL EQI+VCYE AL C  FNPK RPASARD+I
TraesSYM2A03G00828360  LTGQKGYQDIGEVLRSWSDRLERSQRDTLCEQIQVCYETALECRDFNPKKRPASARDIIG  300

TraesSYM2B03G01095480  SLHQMENIQKLRK  313
                        LH+ME IQ   K
TraesSYM2A03G00828360  RLHKMEGIQVFSK  313
```

Supplemental Fig. S3. 2. Protein pairwise alignment of *TraesSYM2A03G00828360* vs *TraesSYM2B03G01095480*. The candidate genes for the SRA blast resistant phenotype in Mattis.



Supplemental Fig. S4. 1. Population analyses of an einkorn diversity panel. Fig. 3 from Hamed et al., 2022 (under revision).

a, Principal component analysis (PCA) of 218 einkorn accessions using all (121,459,674) SNPs. **b**, Unrooted neighbor-joining tree constructed using a randomly selected subset of SNPs (5,318,268). **c**, Population structure (from K = 3 to K = 6) using the same SNPs subset as in (**b**). The split into two domesticated einkorn groups appears at K = 4. Each vertical bar represents one accession, and the bars are filled with colors representing the proportion of each ancestry.

Einkorn groups were assigned considering K= 6 (K with the lowest cross-entropy value) based on the maximal local contribution of ancestry except for β (all β genetic groups were assigned as one group regardless of the contribution of an ancestry). The α group 1 (α_g1, n = 37) is in purple, α group 2 (α_g2, n = 87) is in yellow, γ (n = 24), β (n = 9), domesticated einkorn group 1 (dom_g1, n = 44) is in green, and domesticated einkorn group 2 (dom_g2, n = 17) is in blue. A list of accessions in each group is provided in Supplementary Table 11. d, Heat map showing the mean fixation index (FST) between the two domesticated einkorn groups calculated in 1 Mb sliding windows. Only accessions with 80% ancestry threshold at K= 4 were considered. Centromere midpoints are indicated by red arrowheads. e, PCA using only variants present on the introgressed segment on chromosome 5A. Each point shows an individual accession, colored according to the structure analysis in panel (c). Circled accessions include wild γ accessions and some domesticated einkorn accessions. f, Geographic location of einkorn collection sites. Colors in pie charts correspond to the ancestry at K = 6. The Fertile Crescent is indicated by black lines. Only accessions with known collection sites are shown. g, Geographical projection of the first PCA axis for γ accessions based on the introgressed segment on chromosome 2A (this analysis was done excluding α and β accessions). Black dots represent the location of each γ accessions. Blue color represents the collection sites of γ accessions that were genetically the least diverged from the γ introgression found in domesticated einkorn. The letter K on the map refers to the Karacadağ mountains.

7.1.1. Supplemental tables links

Supplemental tables are in the following link and folder "Supplemental_tables":

https://github.com/quirozcj/PhD_thesis_JQCH_2022

Supplemental Table S2. 1. Whole Genome Sequencing of the (WatSeq project).
Supplemental Table S2. 2. *Ae. tauschii* collection (Gaurav et al., 2022).
Supplemental Table S2. 3. *T. monococcum* from Hamed et al., 2022 (under revision).
Supplemental Table S2. 4. Other wild relatives accessions publicly available.
Supplemental Table S2. 5. Pangenome *k*-mer sizes.
Supplemental Table S2. 6. *k*-mer histograms from WatSeq sequencing samples.
Supplemental Table S3. 1. WatSeq metadata with IBSpy variations information.
Supplemental Table S3. 2. *Ae. tauschii* redundancy test.
Supplemental Table S3. 3. Parent-child test analysis.
Supplemental Table S3. 4. Spikelet number phenotype.
Supplemental Table S3. 5. Max floret number.
Supplemental Table S3. 6. rust resistance phenotypes.
Supplemental Table S3. 7. SRA blast resistant phenotypes.
Supplemental Table S3. 8. Additional modern cultivars samples tested against SRA blast isolated.
Supplemental Table S4. 1. Introgression blocks stitching 50 Kbp with < 30 variations separated less than 10 50 kbp window.
Supplemental Table S4. 2. *Ae. tauschii* lineage specific into the wheat D sub genome.