AN EXPLORATION OF WHOLE-GENOME COMPARATIVE GENOMIC STRATEGIES FOR POLYPLOID CROP GENOMES

by

Gillian Lucy Reynolds

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Individual Interdisciplinary PhD in Computer Science and Plant Genetics

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2022

DEDICATION

I dedicate this work to everyone who stood by me throughout my extraordinary PhD journey, especially Dr. Burcu Alptekin and Hazal Evans.

To my husband who has been there since the beginning and my daughter who appeared half way through. To my mother-in-law (bonus mum) who has provided help, support, stability and comfort throughout this PhD journey. To my little sister and best friend rolled into one amazing, supportive person. To my long-time supervisor Jennifer Lachowiec, who at times must have been at her wits end (most especially while this being written) and yet has still managed to be one of the most gently encouraging and supportive people in my life. To my second longest supervisor, Dr. Brendan Mumey, who must have thought he'd had too much of a quiet life when I asked him to be my fifth PhD supervisor, and he accepted. To my third longest supervisor and absolute inspiration as a PhD mother, Dr. Veronika Strnadova-Neeley. To Dr. Jamie Sherman, who was always kind, supportive and encouraging throughout my entire PhD, but especially during the difficult times. To Dr. Tracey Dougher, who let me talk her ear off for hours at a time, when I was going through some off the worst times of my life.

To Shannon Foley, you gave me comfort, strength and courage when I was afraid in a foreign country. To everyone at the voice centre and the OIE at MSU, especially Victor Maxon and Emily Stark. Without you I'd have left my PhD journey in the beginning, and you didn't let that happen. To Donna Neegard and all of the other wonderful people at the graduate school. You worked tirelessly to make sure my education was never interrupted despite the turbulence of my early PhD years. To Debo Chiolero and the international school who, like the graduate school, ensured I was protected and able to continue my education as I wished while always providing a safe space to be. To everyone at the Gianforte school of computing who welcomed me in to their computer science world despite me being woefully underqualified. Your enthusiasm to share your knowledge has changed my career forever. To everyone at the PSPP department who let me be as distant as I needed but still welcomed me when I reappeared. You made me feel safe again in a place I previously felt afraid.

To Mirium and Christine Cross, and Brian, who were my Montanan family when mine was so very far away.

To everyone at Worcester University, but especially Dr. Mike Wheeler, who fostered my love of bioinformatics and made sure my lack of confidence in my ability as a scientist didn't prevent me from becoming one. I wouldn't be here without you.

ACKNOWLEDGEMENTS

I am extremely grateful for the funding I received during my PhD which was obtained through grants from Montana INBRE, Montana Wheat and Barley Committee, USDA NIFA 2020-65114-30768 and NSF DBI-1759522.

I would like to extend a massive thank you to Coltran Hophan-Nichols and those at RCI for their never ending patience and support. Without your dedication and support my job as a computational biologist would have been extremely difficult.

A huge thank you also belongs to Mikhail Karasikov for all of his help with getting MetaGraph to work with all the different demands I threw at it. You're an absolute star and I am so, so grateful for all the time you put into helping me.

I would also like to extend my sincere gratitude to Dr. John Sheppard who formally introduced me to the fascinating world of machine learning and granted me access to his lab resources for my pilot experiments with PolyCracker.

I also wish to acknowledge Dr.Micha Bayer of the James Hutton institute who has been nothing but a supportive presence since our introduction. I am extremely grateful for your kindness, support and encouragement as an agricultural bioinformatician.

I am also, of course, very grateful to the friends I made at MSU. While we are now all scattered around the globe, I will forever treasure the memories we made during our studies.

VITA

Gillian Reynolds was born in 1991 in Worcester, United Kingdom. She credits her absolute love of the natural world to growing up watching BBC documentaries such as Walking with Dinosaurs, Walking with beasts, and of course, the treasure trove of documentaries narrated by Sir David Attenborough. This was fostered at her high school, Blessed Edward Oldcorne Catholic College, where she excelled as a young science student.

She studied for her Bsc(Hons) and MRes Biology at The University of Worcester where she had nothing but wonderful experiences and to whom she credits her continuation as a science research student. It was during her second-year molecular genetics module, taught by Dr. Mike Wheeler, where she discovered bioinformatics and she hasn't stopped exploring it since. Largely being self taught in the subject until her PhD she's enjoyed exploring the vast, fascinating world of bioinformatics and how it helps us uncover and explain the complexities of our natural world.

Now a wife and mother to Lyra "the-terror-tot" Reynolds, she's looking forward to the next stages of career and post-PhD life.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF FIGURES – CONTINUED

Figure Page

NOMENCLATURE

WGS	Whole genome sequencing
WES	Whole exome sequencing
GBS	Genotype-by-sequencing
SNP	single nucleotide polymorphism
$InDel$	Insertion and Deletion
DBG	De Bruijn graph
$cDBG$	coloured de Bruijn graph
MSA	multiple sequence alignment
$MUMs$	maximal unique matches

ABSTRACT

Genome comparison for large and complex polyploid crop genomes is a highly complex venture, yet it is critical. Given a rising demand for food coupled with yield-impacting resource limitations and rapidly changing global climates it has never been more important to characterise the underlying genetic variation which underpins traits of agronomic interest. In this work, the problem of polyploidy genome comparison is explored at three levels. The first chapter characterizes the sequence relationships that exist between, and within, polyploidy genomes. This is achieved by hijacking a metagenomic strategy for rapid, and efficient, genome sequence classification. The second chapter then utilizes the identified subgenome-specific k-mer profiles for recruitment of assembled contigs and scaffolds previously only recruitable via more resource intensive optical mapping strategies. This makes a greater proportion of the assembled data usable for downstream variant analysis. The third chapter then zooms into the problem of how to identify variants from large -scale sequencing data while minimizing bias and computational costs. A critical assessment of modern variant calling for crop genomes is performed and an algorithm to further extend a new, resource efficient, approach for large scale comparative genomics is presented and critically evaluated. In all, the work presented herein takes a top-down journey from genome- and subgenome-level comparative genomics all the way to identifying base-pair resolution strategies that are capable of revealing the underlying sequences responsible for keeping the world fed.

1

INTRODUCTION

Characterising genetic variation between populations, and how this variation translates into observable, phenotypic differences is a fundamental part of modern crop breeding efforts. Developing profiles of genetic variability for crops and their wild relatives has become a matter of some urgency given the projected potential for disruptions to food security induced by climate change and rising demand for crop food products[34, 102, 124, 151, 228, 324]. By 2050 it is projected food production needs to rise by a projected 35% to 56% to meet demand, yet the resources available, such a land and water, are finite, subject to competition for other uses and in some cases are becomingly increasingly scarce [19, 105, 341]. As such, there is a movement towards improving the sustainability of agricultural practices to protect food security and stave off the associated humanitarian, socio-economical and political crises that accompanies perturbations in food supply [51, 101].

By developing crop varieties that are robust to abiotic stressors such as heat, salinity, and drought, and biotic stressors, such as shifts in pest and disease ranges, the stability of food supply can be ensured in the face of climate and resource uncertainty [80, 86, 91, 289]. To achieve this, the pool of genetic variation that underpins the agronomically-valuable phentoypic variation present across crop cultivars and wild relatives needs to be characterised [84, 225, 310].

Missing Genomic Variation

Currently, there exists a plethora of methods with which genetic variation can be captured, the majority of which utilise reduced-representation, high-throughput sequencing technologies. Popular examples of these technologies include genotype-by-sequencing (GBS),

exome-sequencing (exome-seq) and RNA-sequencing (RNA-seq) all of which provide a reduced-representation view of genomic diversity by capturing only a portion of the genome. GBS is perhaps one of the most popular tools for crop breeding efforts which reduces genome complexity via the use of restriction enzymes [363]. Different combinations of restriction enzymes can preferentially capture different portions of the genome with the literature highlighting a preferential selection for protein coding region capture [137, 272, 302]. Exome-sequencing provides a reduced representation of the genome by capturing only the protein-coding portion of the genome and potentially small amounts of exome-flanking regions [70, 131, 299]. RNA-seq differs from GBS and exome-seq in that it captures sequences that are being expressed at a given time, often under experimental or control conditions, providing a window into the functional landscape of the genome [355].

While these methods are capable of providing insight into the variability between genomes, none of them are capable of capturing the full suite of genomic diversity present between individuals, cultivars or populations given their reduced-representation view. Perhaps the largest problem with such a view is that they are fairly biased towards the protein-coding portion of the genome [137, 272, 302]. However, these regions make up only a tiny portion of the genome (circa 2%) and so the vast majority of the genome, sometimes termed "genomic dark matter", remains understudied [160, 169]. This "dark matter", previously termed "junk DNA", has been revealed to contain sequences critical to the essential structure, function and maintenance of the genome and thus the organism [18, 37, 189, 227]. There has been an explosion of research into these areas, and their applications to crop improvement in the recent literature, with some of the most heavily focused regions being the impact of variants on cis-regulatory elements, pseudogenes and transposable elements.

Cis-regulatory elements

Cis-regulatory elements, such as promoters, silencers, insulators, and enhancers, are short sequences within the genome that act as binding sites for other regulatory molecules, such as transcription factors, and as such are critical for the regulation of gene expression [87, 201]. As variation in gene expression is a known driver of phenotypic traits of agronomic interest such as yield [182, 349, 364], abiotic stress responses [92, 97], biotic stress responses [352, 372], and crop development [352, 389], these largely uncharacterised sequences are an overlooked area for sources of genomic variants of functional significance [77]. For example, [182] found that rare variants in cis-regulatory regions were tied to dysregulation of gene expression in maize, which negatively affected seed weight. They also found a relationship between gene expression and the distance of the variant from the gene, with variants in proximal promoters being associated with down regulation and variants in distal promoters being associated with up regulation. This helps us understand not only which phenotypic effects are resulting from gene dysregulation, but the also the relationship between variant location and the impact on gene expression levels which could be useful for the classification of variants found within cis-regulatory regions. [294] was able to take such studies one step further by engineering novel cis-regulatory alleles for three tomato genes which controlled various fruit traits including fruit size, inflorescence architecture and plant growth, demonstrating the practical implications of characterising the impact of sequence variants on agronomic traits of interest.

Pseudogenes

Pseudogenes have been regarded as genomic relics, having once been functional genes that have since become deactivated [315, 366]. Considering these sequences are no longer functioning in their genetic capacity, they are generally subject to neutral evolution and so gather mutations over time with little to no consequence [114]. Pseudogenes have been

described in the literature as "evolutionary hotspots" for the generation of genes with potentially novel alleles [114] and as a "playground for innovation' [277]. Due to their rapid mutation accumulation, with respect to the portion of the genome under selective constraint, they are a rich source of genetic variation. However, whether such mutations are of functional significance has recently become an area of considerable interest in the literature. It does appear that mutations accumulated here do have the potential to have phenotypic consequences as not only are some pseudogenes exhibiting active expression, their upstream regions are rich in cis-regulatory elements and are the origin of a number of non-transposable-element regulatory ncRNAs [114, 366, 367]. Genetic expression does not necessarily mean such sequences are functional as they can be post-transcriptionally silenced. However, [367] argues such transcription is unlikely to be transcriptional noise, given that pseudogenes exhibit specific expression patterns including responses to abiotic stresses. This argument was recently bolstered by the results of [288] who found that the Arabidopsis epigenome responded to cold and drought stress via chromatin modification to pseudogene-rich regions. In particular, drought-stress was associated with histone modifications consistent with the opening of chromosomal regions containing pseudogenes located in the pericentric regions. Precisely which pseudogenes have functional significance, what functions they perform, and how they have obtained or maintained such activity remains to be shown. This area of research is evidently in its infancy, and so the impact of variants residing pseudogenized regions may have been overlooked. While the fruitfulness of gathering variant data for such regions is unknown, variant data is required to compliment functional data obtained via epigenomic and transcriptomic methods, to provide evidence for functionality and so is a necessary exploration. Given such data gathered from plants exhibiting a phenotype of interest, and it may be possible to infer trait-related functionality for sequences previously thought to be nothing but molecular relics.

Transposable elements

Transposable elements are genetic sequences with a bad reputation. Sometimes labeled as "selfish elements" that propagate themselves thorough the genome, either via a "copy and paste mechanism" or a "cut and paste mechanism", they were regarded as destructive elements capable of disrupting coding region sequences deleteriously [4, 69]. However, recent research has shown this reputation to be a little unfair. While all the above is true, they can be functionally useful, providing benefits to the organism in whose genome they reside. As a significant portion of crop genomes are comprised of TEs (wheat and maize 85%, barley 80%, rice 41%), their functional significance in the genome and their relation to traits of agronomic interest is of paramount interest [199, 323, 360, 361]. It appears TEs are functionally important both directly and indirectly. Directly, TEs can cause phenotypic changes by disrupting functionally important sequences, disrupting gene translation and a by causing alterations in phenotypes such as the gain of brittle rachis in semi-wild wheat populations [89, 161, 206]. Transposable elements can also contribute to phenotypes via a process known as "exadaption" or "molecular domestication", where they lose their ability to transpose and become regulatory sequences [28, 46, 69, 142, 163, 275]. These have been observed to infer phenotypes of agronomic interest in crops. For example, a type of TE known as MITES have been observed to contribute to the miRNA response of wheat during a powdery mildew infestation [275]. [163] found that exadapted TEs, particularly MULEs, were functional in response to abiotic stresses including phosphate starvation, salt stress, cold stress and arsenic toxicity identified via a phenomics and reverse genetic engineering approach in *Arabidopsis thaliana*. This study also showed that while some exadapted TEs are show stress-specific functionality, others were generalists, and it appears that TE families respond to similar stressors [163].

Indirectly, TEs are capable of causing functionally significant structural variants via their insertions and deletions as they transpose around the genome causing double-stranded

DNA breaks [140, 336]. These double-stranded breaks may be prone to erroneous repair, thus leading to the introduction of genetic mutations [28, 46, 220]. They may also cause recombination errors due to their highly repetitive nature and sequence similarity with other TEs dispersed throughout the genome, both of which can result in large-scale structural alterations including deletions, duplications, inversions and translocations [28, 46, 336]. Such large-scale variants have the capacity to cause functionally significant genomic changes including alterations to genetic dosages and locations, and modifications to gene and genome structure [109].

It appears that TEs, as the fastest evolving genomic sequences with the propensity to cause small and large-scale genomic changes are a key source of genomic variation between phylogenetically close species, and even individuals, with some TEs providing lineage- or species-specific sequences with can contribute to variations in stress responses [28, 89, 190, 202, 275]. Bizarrely, there is also evidence within plant genomes that TEs can be acquired horizontally, as opposed to the usual vertical transmission of genetic information with accompanying phenotypic alterations. Examples include the TE known as "Rider", which was horizontally transmitted from asparagus, giving Roma tomatoes their characteristic, elongated shape, SORE-1 which in soybean induces photoperiod insensitivity and Raider in maize, which is associated with the up-regulation of genes in response to UV stress [121].

Structural Variants

Unfortunately the detection of structural variations (SVs), either TE-induced or not, is out of the scope of the vast majority of the aforementioned reduced-representation approaches, which are usually only capable of identifying single nucleotide polymorphisms (SNPs) and small insertions and deletions (InDels) [150]. Structural variants are considered to be genetic variations larger than 50bp and includes variants such as copy number variants

(CNVs) and presence-absence variants (PAVs), both of which are critical for plant genomics research, especially for the characterisation of crop pan-genomes [31, 88, 122, 149, 214, 235, 331]

While structural variants are less numerous than smaller variants, they have propensity to cause much larger genomic effects because they impact many functional sequences simultaneously. Structural variants are widespread within the genomes of rice, wheat, maize, barley and other crops and have been tied to numerous agronomically important phenotypes [3, 12, 12, 109, 111, 348].

Interestingly [13] found no overlap in their mapped high confidence SNP and SV loci, and more concerning still, discovered an SV loci to be the responsible agent for a phenotype of interest that was previously thought to be tied to a SNP. Further, [90] found that SNP-based approaches are incapable of capturing TE insertion polymorphisms which as discussed, represent a significant pool of phenotypically-important genetic variation.

While there remains a need for other studies to characterise the whether such disparity between SNP and SV loci tied to phenotypic traits is present for other genomes, it does highlight a major advantage of a whole-genome-capture method; all forms of variants, across all genomic loci can be captured simultaneously.

Whole-Genome Variant Hunting

Whole genome sequencing (WGS) is the only method that is capable of capturing the full suite of genetic variations present within genomes. However, it does present with added complexities from an data analysis standpoint. WGS data is massive and contains complex genomic architectures. Given the increased costs incurred by using WGS over reduced-representation approaches it is essential that the data is used to it's fullest extent. To that end, there are two primary strategies to extract whole genome variant data from WGS: Alignment-to-reference strategies and *de novo* genome assembly.

It is important to note that the following section is written with short-read Illumina sequence data in mind. While long-read sequencing data can solve many of the issues caused by the use short-read sequencing data , it can be between 5X-12X more expensive and as such is not yet routinely implemented for whole-genome variant identification ventures [386]. Further, the vast majority of whole genome sequencing data deposited in the sequence read archive (SRA) are from short-read sequencing platforms. This freely available data presents an opportunity for re-exploration of variants using novel methods which may reveal previously uncharacterised variation.

Alignment to reference

The alignment to reference strategy involves aligning short reads to a given reference genome. Traditionally speaking, this reference genome is a single, linear, haploid representation of the genome of interest. Recently however, there has been some movement in transitioning to alignment to a reference genome augmented with known genetic variation, often using a graph structure.

The purpose of this transition is to mitigate the "reference-bias" problem. This problem refers to the fact that sequences which are absent or sufficiently divergent from the reference sequence will not align and are generally excluded from further analysis [25, 331, 338] A graph-based reference mitigates this problem by more accurately representing the known diversity of a genome and so providing more possible sites with which to align the sequencing data [284].

However, the increase in the number of sites in which to align can be problematic from a multi-mapping point of view, which is where a read aligns equally well to a number of genomic loci [176]. In this event, it is not possible to definitively determine the genomic location of the read which negatively impacts downstream variant calling analysis. Multi-mapping reads are either discarded or given a low-confidence mapping score which are often

filtered out by variant calling algorithms [72, 104]. The graphical reference genome is also problematic in that storing such a dynamically growing structure, can be computationally prohibitive, especially given large or genetically diverse genomes [176, 284]. This solution also does not entirely mitigate the reference bias problem given that there may exist novel variation in the sequence data which would not be present in the graph structure. This will be particularly true for genomes which are not well characterised.

The alignment to reference strategy is also problematic for genomes with a high repeat content, which has occurred either due to polyploidization, smaller-scale genomic duplications or expansion of repetitive sequences such as those from TEs, regardless of reference genome representation. This again is due to the multi-mapping problem. If the genome being studied has a high subgenome similarity or a particularly large repeat content, the portion of reads that align with high confidence to a single location is going to decrease. For organisms like wheat, which is both a polyploid with high subgenome similarity and contains a repeat content of over 85%, this is a particularly significant problem [361].

De Novo Assembly

De novo genome assembly is a non trivial venture for large, complex polyploids often requiring terabytes of computational memory and months of run time alongside experimental parameter tuning to optimise the resulting assembly. For example, the MaSuRCA assembler requires 30-60 days of access of up to 2Tb RAM and a minimum of 64 cores and 10 terrabytes of disk space to perform a plant genome assembly up to 30Gb in size [8, 396]. Such a large amount computational resources for such a long period are not readily available to all labs, which may prohibit the numerous whole-genome assemblies for large, and complex organisms required for assembly-based variant analysis.

Further *de novo* genome assembly is an expensive venture, requiring more than a single instance of genome sequencing. Often the genome is sequenced at high depth to correct for

errors, and using a range of insert sizes between paired-end reads so that assemble contigs can be scaffolded together to form more contiguous sequences. Additionally, long-range information such as long-read sequencing data, optical maps and genetic maps are used for ordering, orienting and anchoring sequences to their chromosomes or subgenomes of origin [11, 212].

Further, the vast majority of *de novo* assembly algorithms are not designed for polyploid graph traversal and so cannot correctly navigate the polyploid-induced subgraph structures that will be present in the assembly graph. The full suite of topological structures that may be found in a DBG containing polyploid genomes have yet to be characterised in the literature, but two major subgraph structures have been identified: superbubbles and complex bulges. Superbubbles are complex acyclic subgraphs which have only one source and sink nodes between which are highly connected nodes which have numerous solutions to their traversal [119, 251, 260]. An example can be seen in Figure 1.1. Complex bulges only differ to superbubbles in that other nodes not within the bubble or bulge structure can have edges into the structure, whereas in the superbubble definition, the superbubble is self-contained [251, 297]. This difference can be seen in figures 1.1 and 1.2. Where figure 1.2, which shows a complex bulge, has nodes coming into, and out of, the bulge structure, figure 1.1 showing a superbubble shows a self-contained complex bubble structure.

It is likely even smaller and simpler subgraph structures are inappropriately handled for polyploid genomes. For example, bubbles in assembly graphs are often popped and merged but the assembly algorithms into a single consensus sequence, under the assumption these bubbles are either caused by sequencing errors or haplotype variants for a diploid genome [251, 314, 376, 396]. Such an assumption is incorrect for polyploids, and may represent a sequence variant between subgenomes in an otherwise identically shared sequence. Such a traversal for a polyploid would mean that the subgenomic sequences are merged into one, representative sequence and are output as a single contig. What should happen here

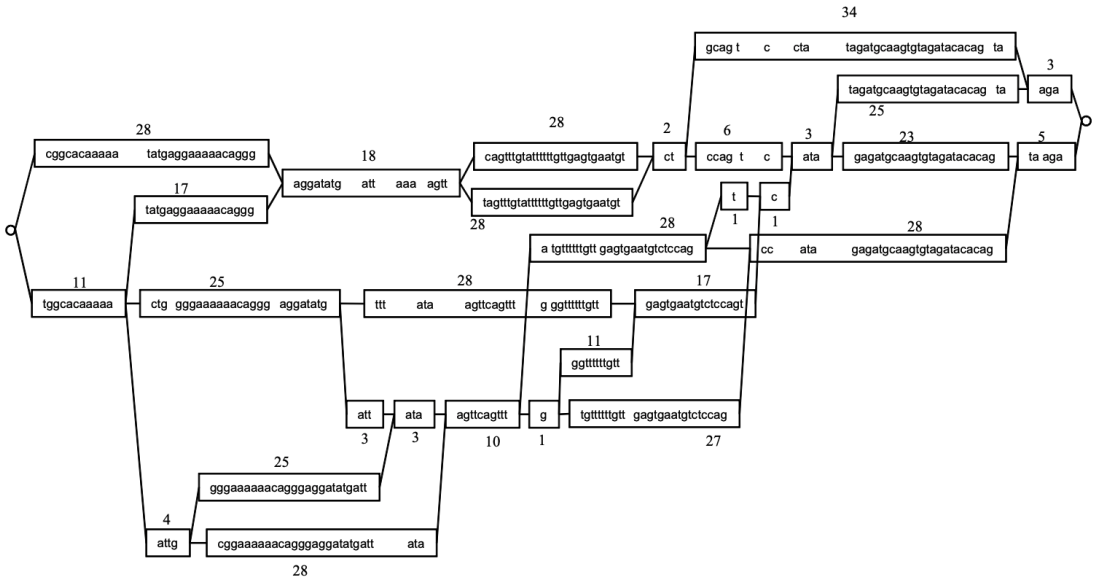


Figure 1.1: A visual representation of a superbubble structure, from [251]

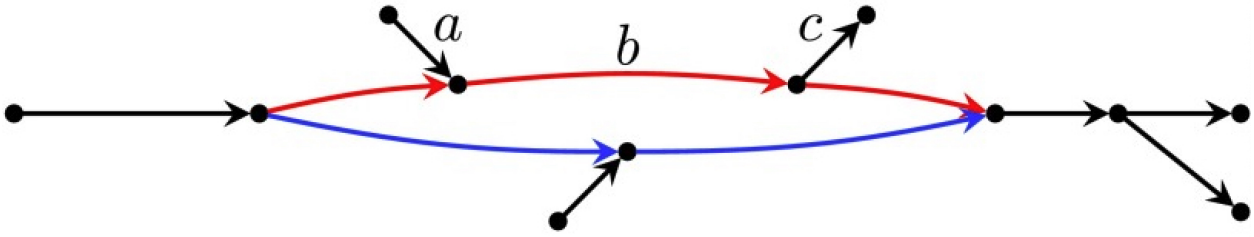


Figure 1.2: A visual representation of a complex bulge structure, modified from [297]

for a polyploid is that the graph should be navigated n times, where n is the number of subgenomes, with each subgenome traversing either a side of bubble according to coverage (*i.e* if coverage only allows one traversal of a side of the bubble). This traversal would result in n contigs which would contain either of the two variants the bubble represents, as is appropriate for a polyploid organism with n subgenomes. Failure to navigate a polyploid de Bruijn graph appropriately will lead to a heavily fragmented assembly containing misassembled fragments, particularly chimeric contigs (where disparate parts of different subgenomes are

stitched together), collapsed shared subgenomic sequences and collapsed repeat sequences [20, 22, 204]. Evidently, this is not a favourable outcome for comparative genomics facilitated variant discovery.

There does exist a single genome assembler designed for polyploid genomes, w2-rap, which was designed for the wheat genome [71]. However, it still requires significant resources and also requires specific Illumina sequencing protocol and accompanying long mate-pair data. Further it has yet to be publicly benchmarked on a range of polyploid genomes, with its performance having only been tested on human and mosquito data.

A final option does exist for polyploid assembly, which is haplotype assembly. In this approach, *de novo* haplotype assembly attempts to assemble haplotypes separately to prevent their collapse in an assembly graph [23, 115, 136, 186]. However, only one seems to be designed for non-diploid polyploids, and it has only been tested on simulated tetraploid gene fragments [136]. It seems that haplotype assembly is a great challenge using in silico approaches only. In fact, this remains true even when additional genomic data is used. The recently released phased tetraploid potato assemblies utilised a combination of genomic information including phased markers, long reads and synteny data to achieve their genomic assemblies [145]. Even given all of this genomic information a significant amount of assembled, and even chromosomally assigned, data remains unphased.

Coloured de Bruijn graph approaches

Given the pitfalls and complexities associated with both genome alignment- and assembly-based approaches, a novel approach to variant finding within polyploid crops is most certainly required. A potential solution could be the implementation of a coloured de Bruijn graph (cDBG) strategy. A coloured de Bruijn graph (cDBG) is an expansion of de Bruijn graphs which are most notably used in genome assemblies. In a classic DBG genomic information for a single individual or species, usually in the form of reads, is broken up into

k-mers. These k-mers then becomes edges in the graph, with the graph nodes representing the k-1 overlap between the k-mers. Theoretically, one could then take a Eulerian walk (visit all edges once) in the DBG to reconstruct the genome from the fragmented reads and k-mers [75].

cDBGs expand the DBG structure to enable multiple individuals or species to be included in the same graph. They achieve this by using colours to denote which individual contributes the k-mer information in the graph. As such, one could perform individual assemblies by taking a Eulerian walk through the graph for edges which are coloured to correspond to that individual. However, cDBGs are designed for variant finding rather than assembly. To achieve this, cDBG variant finding algorithms harness the colour information and variant-induced graph topologies to identify variants between the individual graph. Variant-induced topologies include bubbles induced by SNPs where a colour, or set of colours traverse the nodes at the top of the bubble, and the remaining colours traverse the bottom of the bubble [152]. An example can be seen in figure 1.3. In this simple example, the two alternative paths (orange and blue) can be traversed to assemble the variable genomic sequences. The traversal of the graph can then cease when all of the colours converge onto a single node, which means the variant has been traversed and the sequence is once again identical.

Other larger and more complex variants can form much more complex topologies in the graph, but these are still algorithmically identifiable as differentially-coloured alternative paths in the graphs.

Unfortunately cDBG-based variant finding is computationally expensive to implement due to the sheer computational overhead required to construct and store a de Bruijn graph for multiple genomes and the node colours. In fact, the node colours often require significantly more space to store than the actual graph structure as they naively require as each node or edge in the graph can have multiple colours associated with it [10, 170].

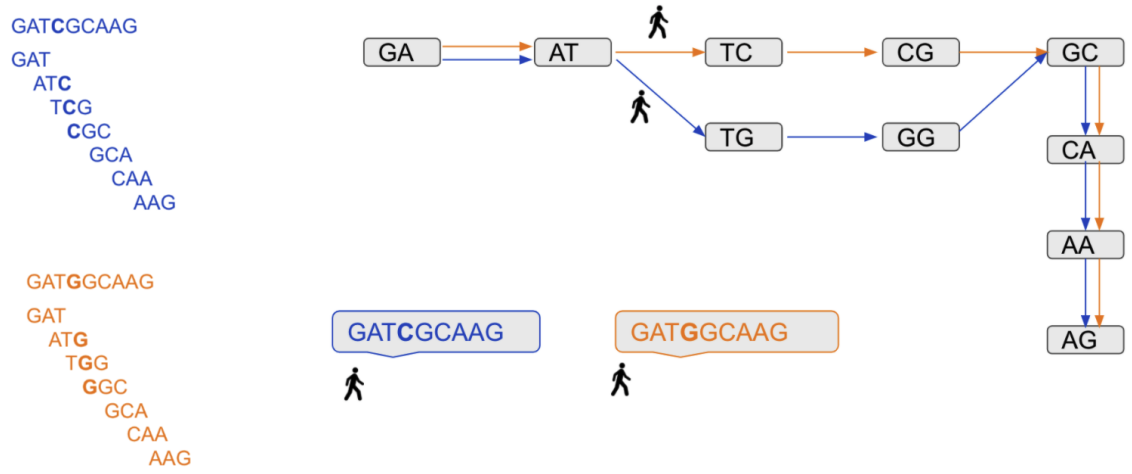


Figure 1.3: An example of coloured de Bruijn graph in which we can take two alternative paths in the graph to construct genetic variable sequences

A lot of recent research effort has gone into making coloured de Bruijn graphs more space efficient since their introduction to comparative genomes by [152], although largely the focus has been on mammalian and microbial genomic applications. One of the first improvements on Cortex, the first coloured de Bruijn graph for variant calling was Vari, which uses a succinct de Bruijn Graph Structure, known as the BOSS representation and binary matrix to store colour relationships where rows are the graph vertices (k-mers), and columns are the colours [236]. This binary matrix is then compressed by the Raman-Raman-Rao or Elias-Fano encoding, which performs very well for the sparse colour matrices (i.e. most k-mers are unique) [10, 236]. However, while this representation is efficient, Vari has to first construct an uncompressed version of the colour matrix which required terabytes of space when tested by [10]. This significant bottleneck was addressed by Rainbowfish, who developed a 0th-order entropy-based compression strategy, which reduces the memory requirements of constructing the coloured de Bruijn graph 20X over Varis implementation[10]. This was then further improved by [144] who developed Bifrost, a high-parallelized and efficient coloured de Bruijn graph constructor which uses 20X less memory than Vari. It achieves this by using a bloom

filter to represent the de Bruijn graph and one of three different colour representations to compress the colour annotations for the graph. A bloom filter is a probabilistic data structure that, when queried will either return a "no" or a "maybe". This "maybe" reply comes from the fact that being a probabilistic data structure, it is prone to false positive and as such, cannot definitively answer if data is present. The rate of false positive can be controlled, with a trade-off that a reduction in false-positives comes with an increase in data structure size [262]. The three data structures Bifrost uses for representing the colour matrices are a binary matrix, like Vari, a compressed bitmap and a roaring bitmap [144]. Precisely which data structure is used is dependent on the size of the graph and colours.

Recently an even more efficient cDBG constructor and traversal strategy was developed by [170]: Metagraph. MetaGraph was designed to construct cDBGs for petabytes of data which they demonstrated by constructing individual cDBGs for every plant, fungi and bacterial genome deposited in the sequence read archive (SRA). For the plant genome cDBG they achieved a compression ratio of 139X, reducing 576 terabytes of information into a graph 602GB in size.

Metagraph also implements a novel variant searching strategy called differential assembly. Simply put, differential assembly is a strategy that seeks to assemble only those sequences that differ between genomes. More formally, given a set of genomic sequences S = s1, s2,..sn, we can assign each genome a colour $C = c1, c2..., c3$ and construct a coloured de Bruijn graph $G = V, E, C$ [152]. From this graph we can identify which k-mers are present in one set of of genomes and absent in others. We can then extract and then assemble these k-mers using traditional, or novel, assembly algorithms. These assembled fragments, by their differential nature, the variants that are present between individuals in the graph.

This novel strategy holds excellent potential for identifying the large and complex variants that current short-read alignment-based variant calling strategies struggle to identify without the expense of assembling the whole genome that a traditional assembly-base variant

strategy requires.

Scope of this work

In this work I seek to explore novel strategies for variant calling algorithms that address the various needs outline in this introduction. This includes the need to identify variation in repetitive and complex genomic regions, alongside the need to balance such explorations with the high computational complexities associated with variant analysis.

In chapter one, I demonstrate that the metagenomic problem of sequence classification and clustering can be applied to identify repeatome-driven subgenome sequence differentiation. I extend this by implementing a metagenomic strategy to perform B-chromosome and A genome comparison sequence comparison for the Maize (*Zea mays*) and Rye *Secale cereale* genomes.

In chapter two I explore the utility of subgenome-specific k-mer profiles for previously unanchored contig recruitment. I demonstrate that global subgenomic signatures of allohexaploid bread wheat *T.aestivum* are excellent features for unassigned contig recruitment and that classification success is largely dependent on k-mer size and seemingly indifferent to contig size or false-positive mitagations strategies developed by [262].

In chapter three I perform a critical review of modern graph-based variant calling approaches for crop genomes. The advantages and disadvantages of the approaches are discussed in context of crop genome size and architecture and an assessment of currently available graph-based variant calling approaches is performed. I enhance this review by proposing a novel algorithm for differential-assembly-based variant calling approach.

17

K-MER SPECTRA FOR COMPARATIVE GENOMICS

Comparative genomics is a highly complex, yet critical venture that underpins our understanding of genome function and evolution. The complexity of such comparisons arises out of both the biology and the technology being used. Biologically speaking, complexity arises from genome evolution. Any event (i.e., mutation) which disrupts the linear order of nucleotides complicates any comparisons with another string of nucleotides as it prevents a simple nucleotide-by-nucleotide comparison [395]. Instead, the complex process of sequence evolution, which features numerous linear-ordering disrupting events, such as single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), copy number variations (CNVs) and translocations, has to be taken into consideration[5, 344, 395].

An example of such mutations can be seen in figure 2.1. In figure 1A there is an equal trinucleotide substitution that is detectable via a linear pair-wise scan of the bases in the two sequences. However figures 1B-1D, which show a trinucleotide deletion, translocation and CNV to the lower sequence respectively, show that these mutations are not detectable via the same linear pair-wise scan of the bases. In fact failing to account for non-substitution-based mutations leads to calling erroneous variants further downstream.

As such, numerous algorithms have been developed to allow for pairwise alignment of sequences with gaps. Figure 2.2 shows an alignment of the same sequences as figure 2.1, but with gaps which allow for the correct identification of the mutations. Two of the most famous are the Needleman–Wunsch [247] and Smith-Waterman algorithms for pair-wise alignment of sequences [247, 317]. However, with the added benefit of handling the biological complexity of non-substitution-based mutations comes with an increase in computational complexity. While a pairwise linear scan of two sequences will take as long as the longest length of the two sequences, known as $\mathcal{O}(N)$ using formal computer science notation, the number of possible gapped alignments of two sequences is $(2N)!/(N!)^3$ [395]. To navigate

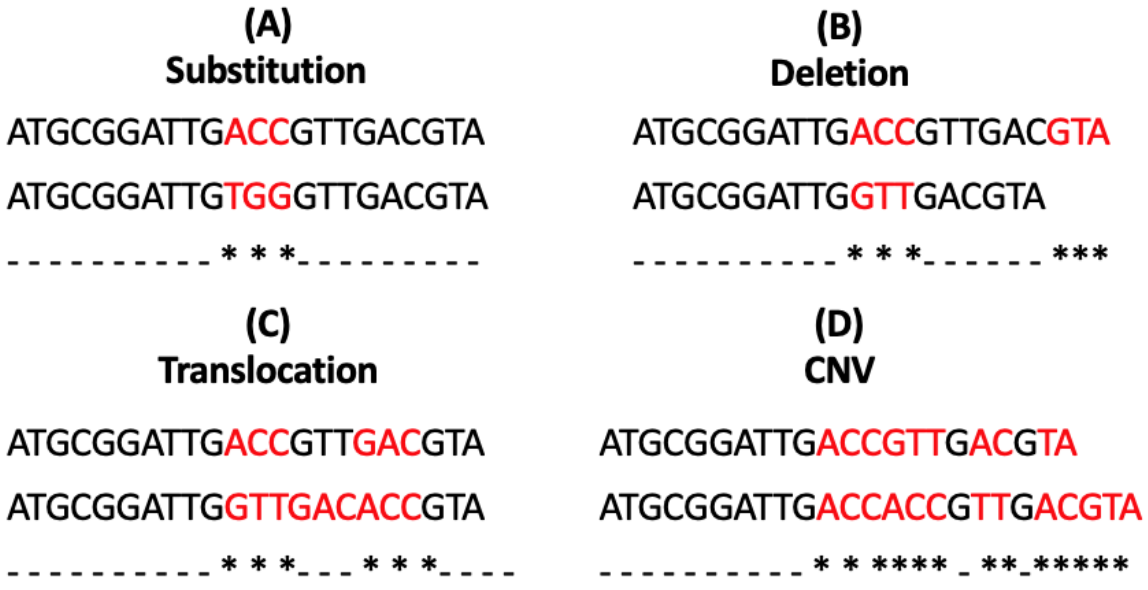


Figure 2.1: An example of how mutations can disrupt the linear ordering of nucleotides in a DNA sequence. **1A**: an equal trinucleotide substitution which causes no linear disruption. **1B**: trinucleotide deletion of the "ACC" bases the lower sequence. **1C**: trinucleotide translocation of the "ACC" bases on the lower sequence. **1D**: trinucleotide copy number variation (CNV) of the "AAC" seqence on the lower sequence. Neither **1B**, **1C** or **1D** is detectable via a linear pair-wise scan of the sequences to detect variants.

this computational complexity, algorithms like Needleman–Wunsch and Smith-Waterman implement dynamic programming strategies which guarantee to return the mathematically optimal solution while only having a computational complexity of take $\mathcal{O}((MN)+(N))$ and $\mathcal{O}(MN)$ respectively, where M and N are the length of the two sequences [24, 395]. Although numerous developments have been made to make sequence alignment less computationally costly, they still do not scale well to the size of genomes.

Should there be more than two biological sequences to compare, the problem becomes a multiple, rather than pairwise, sequence alignment problem which results in another increase

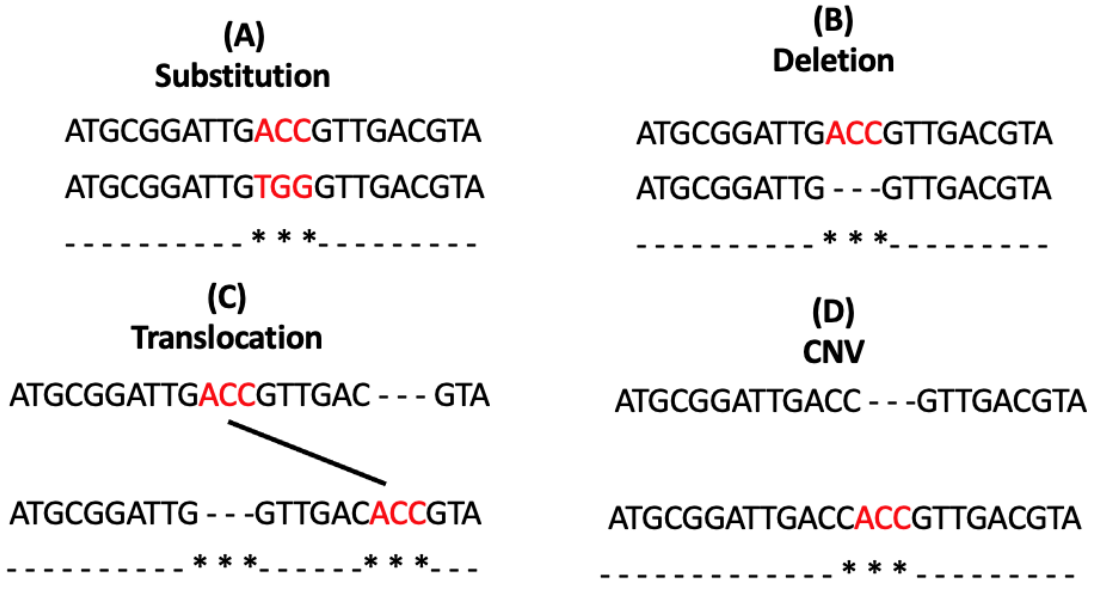


Figure 2.2: An example of how different mutations could be modeled linearly when gaps are allowed.

in complexity. For a comparison of any more than two sequences the problem becomes NP-hard and so any solution requires the use of heuristics which are not guaranteed to return a mathematical optimal solution [350]. The progressive alignment algorithms that are frequently applied to the multiple sequence alignment (MSA) problem require $\mathcal{O}((k^4)+(N^2))$ time, where K is the number of sequences to compare and N is the length of the longest sequence [24] and are prone to getting stuck in a local optimum. Given 3 genomes of *Arabidopsis thaliana* with a median genome size of 120Mb, the number of computations required would be $3^4 + 120000000^2 = 1.44 \times 10^{16}$, which if each computation took a second, would take over 456,621,005 calendar years to complete. Clustal Omega, a popular, more recent, multiple sequence alignment algorithm, has a time complexity of $\mathcal{O}((NlogN))$. Given an *Arabidopsis thaliana* genome with a median genome size of 120Mb, this would require almost a billion operations [313]. If each operation took a second to complete, that would

be 31 years of computation.

Given this prohibitive timescale, much effort was put into the development of algorithms that would enable whole-genome alignment. One of the most popular tools to emerge from this is MUMmer. MUMmer works by identifying regions between the two genomes which are both exact matches, and which occur only once in each genome [81]. These sequences are called "maximally unique matches" or "MUMs" because extending them any further left or right in either sequence, will cause them to no longer match exactly [81]. These "MUM"s act as anchors between which the sequences can then be locally aligned using a pairwise alignment algorithm to find the variants. The MUMmer program has gone through a number of different development cycles since its first release in 1999, the most recent of which, MUMmer4, is most efficient in terms of time and space [81, 82, 185, 224]. In this iteration, MUMmer uses a suffix array to construct an index of one of the two genomes to be compared. Following index construction, MUMmer "streams" the second against the suffix tree without actually adding it (which reduces space requirements for sequence storage by at most half) to find the MUMs [224]. The identified MUMs are then clustered by their coordinates in the first genome and the aforementioned Smith-Waterman algorithm is used to align the sequences which fall between the MUMs. It is from this alignment that variants can then be identified. The time complexity for this strategy is $\mathcal{O}(N)$ for suffix array construction where N is the length of the sequence used for construction, $\mathcal{O}(M)$ for the time taken to stream the second sequence against the suffix array, $\mathcal{O}(A)$ for the clustering step, there A is the number of MUM's or "anchors" and $\mathcal{O}(A(MN))$ for the local alignments between each MUM. MUMmers strategy makes it both very space and time efficient, requiring less than 4 minutes and approximately 1.8 Gb of RAM given that MUMmmr4 requires only 15 bytes per base to store the suffix array of one of the genomes to be compared[224]. For a larger genome such as wheat, the space and time requirements would be approximately 218GB and 480 minutes respectively given a genome size of 14,195,643,615bp (as reported for IWGSC

v2.0).

However, despite the impressive feats of engineering demonstrated by the authors of whole-genome alignment strategies, such as MUMmer, alignment is a flawed approach for the identification of genomic variation and determination of evolutionary distances between whole-genomic sequences [394, 395]. The major problem is that alignment strategies inherently assume sequence collinearity [85, 395]. That is, they assume sequence evolution can be modeled by a series of matches, mismatches and gaps on a linear sequence [395]. As such, they can only model smaller evolutionary changes such as substitutions and InDels. The larger, more complex mutations, such as duplications, inversions and translocations, that are seemingly inherent to genome evolution, cannot be well captured via alignment strategies. As such, models of evolution, for which phylogenetics and genomics often rely on, only consider sequence substitution changes between sequences to determine evolutionary time and distance[40, 210, 219].

This may be particularly problematic for crop genomes. Not only are crop genomes often large, which will increase the computational cost of an alignment-based evolutionary analysis, but they also often very complex. Much of the size and complexity of these genomes can be attributed to transposable elements (TEs) which make up the vast majority of many crop genomes [360]. TEs are known to undergo various expansions and radiation events, which can dramatically alter the genomic landscape [36, 45, 74]. Even when they are silenced via epigenetic modifications, they can still provide added complexity via continued expansions and rapid accumulations of mutations [187, 391]. Alignment algorithms are known to perform particularly poorly in regions of low sequence homology and given that these regions are home to a highly diverse ecosystem of TE-repeat families, alongside their propensity for collinearity-violating evolutionary events (i.e. duplications, inversions and translocations), it is reasonable to infer that such regions are unlikely to align well across genomes [5, 74, 323, 397].

Polyploidy also plays a major role in crop genome comparison complexity. Not only does having additional sets of chromosomes complicate comparisons, but the process of polyploidisation itself is capable of having profound effects on the genome at the nucleotides level. From chromosome shattering and major structural rearrangements, to fractionation-driven pseudogenization and TE radiations alongside various deletion events, the process of polyploidisation has been aptly described as a "genome collision event" [21, 43, 55, 217, 253, 298].

Given both of these factors, it seems crop genomes are particularly ill-suited to alignment-based sequence comparison and may instead benefit from alignment-free approaches to determine sequence similarity. While alignment-free approaches are not capable of returning the base-pair resolution of variants which can be identified via alignment-based approaches, they are highly scalable and are unphased by regions of poor homology or complex mutations and as such are far better suited to detecting evolutionary signals in large, complex genomes [148, 344]. With well over 100 alignment-free methods to choose from, the application of alignment-free approaches to the genomic comparison and investigation of crop genomes has the potential to become an area of major research interest [395].

K-mer based comparison of genomes is currently the leading alignment-free strategy for sequence comparison [395]. Generally speaking, these strategies work by shredding the genomes into distinct sets of overlapping sequences of length k, known as k-mers, and projecting them into vector space. This enables their analysis by robust and rapid statistical measures such as the Jaccard similarity and cosine similarity metric. The Jacccard metric is particularly intuitive as it simply enumerates the k-mers that each pair of sequences of have in common and divides it by the total number of k-mers. More formally this is described as taking the intersection of the two of k-mers and dividing by the union of the same sets, which is described in equation 2.1 [375]. The cosine similarity metric (seen in equation 2.2) is equally simplistic but can be a little more difficult to visualise. In this metric, the observed

frequencies of k-mers are stored as vectors and the cosine angle between them is taken as a measure of similarity [233, 312]. This measure has the advantage in that it is not concerned with the magnitude of the vector, which refers to the number of distinct k-mers, which is often influenced by the length of the sequence [54, 233, 312, 375]. As such, it is possible to compare sequences of varying length without introducing sequence length-based bias. The Jaccard similarity does not have this advantage and so care has to be taken to ensure the measure of similarity is not being biased by sequence length discrepancies. Both measures return a value between 0 and 1, where 0 indicates the sequences have no k-mers in common and 1 indicates they are identical.

$$J(AB) = \frac{A \cap B}{A \cup B} \tag{2.1}$$

$$similarity(AB) = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|} \tag{2.2}$$

However storing the entire set of genomic k-mers can be prohibitive given that genome space theoretically grows at a rate of 4^k. While genomic k-mer surveys have demonstrated that for sufficiently large k, genomes contain only a small portion of all possible k-mers, the number of k-mers they do contain is still sizable [54]. Given that k-mer based research is at the core of many critical computational biology applications, such as genome assembly, genome annotation, short-read sequence alignment, estimation of genome characteristics, contamination detection and genotype-phenotype linkage, much research effort has gone into developing efficient strategies for k-mer storage and analysis [75, 98, 179, 188, 207, 221, 255, 281, 293]. Despite these efforts, the analysis of whole-genomic k-mer sets for sequence similarity detection is still problematic, given the quadratic (all against all) nature of the pairwise similarity calculations required.

As such, in this work we utilise alignment-free, k-mer-based comparison of genomic

sequences. We do this by reframing the problem of polyploid subgenome-comparison as a metagenomic problem and are thus able to take advantage of the computational advances for efficient sequence comparison made in this area. Specifically we show that the rapid and highly-scalable minhashing method, which is commonly used for metagenomic-based comparison and classification of sequences, is capable of revealing subgenomic and genomic relationships previously described in the literature and as such holds promise for determination of relationships between sequences and progenitor screening for crop genomes.

Hijacking a rapid and scalable metagenomic method for crop comparative genomics highlights subgenome dynamics and evolution

Abstract

Large, and often complex, polyploid genomes present numerous problems for whole-genome comparative analysis. The sheer size of these genomes can present significant scalability issues requiring extensive computational resources. Further, their genomic complexity, in terms of their repetitive landscape and numerous collinearity-violating evolutionary events make them poor candidates for traditional linear-based whole-genome comparative strategies. In this work, the problem of intra- and inter-genome comparison for polyploid genomes is reframed as a metagenomic problem enabling the use of the rapid and scalable MinHashing approach for polyploid crop comparative genomics. This novel approach reveals literature-verified subgenomic and genomic evolutionary relationships and holds excellent promise as a tool for rapid screening of progenitor genomes. An detailed examination of the sequences responsible for the observed relationship between sequences points to the complex, and rapidly evolving, landscape of transposable elements. An extensive investigation into the MinHashing parameters reveals the MinHash generated genomic signatures are excellent approximations of sequence similarity which are little affected by the size of genomic signature generated. Further, the clustering approach used for comparison of the genomic

signatures is scrutinized to ensure applicability of the metagenomics-based method for crop genomes. In all, the MinHashing-based sequence comparison strategy implemented in this work is an excellent, novel, approach for comparative subgenomics and genomics for large and complex polyploid plant genomes.

Introduction

Whole-genome sequence comparison is a computationally complex venture, that despite decades of research, its optimisation, both computationally and biologically, is still an open problem. Biologically speaking, alignment-based comparative genomics is complicated by the complex process of genome evolution. Structural mutations, such as translocations,inversions and duplications, disrupt the linear of ordering of nucleotides within a genomes preventing a pair-wise linear nucleotide-by-nucleotide analysis [5, 344, 395]. While for a pairwise alignment of sequences mathematically optimal solutions exist, such approaches scale poorly for large sequences, necessitating the use of heuristic algorithmic approaches which do not make the same guarantees [62, 350, 395].

However crop genomes, which feature some of the largest and most complex genomes known, can stretch the scalability of whole-genome comparison strategies. This is especially true given hardware limitations to store the genomic indices many whole-genome comparison approaches require to make comparison within in a reasonable time feasible [224, 267]. Further, the complexity of crop genomes presents numerous problems for alignment-based whole genome comparisons, which assume sequence collinearity and perform poorly in areas of poor sequence homology [394, 395]. Many crop genomes are large, feature a rich and diverse landscape of transposable elements and have experienced numerous collinearity-violating events during their evolution, such as polyploidization-induced structural changes [83, 360].

However despite these challenges, comparative genomes lies at the heart of research

efforts to ensure global food security. Via comparative genomic strategies, variants which underpin traits of agronomic interest can be identified and integrated into novel crop varieties to meet 21st century needs [1, 184, 238, 346] As such it is imperative to identify methods from which sequences of interest may be identifiable. It is also critical that a better understanding of subgenome diversity, evolution and dynamics is developed. Polyploid genomes are some of the most complex to unravel, with bioinformatic analysis' complicated by their large size and sequence complexity. However, these highly dynamic genomes are responsible for some of the worlds most critical crop production and as such, it is imperative strategies are developed which enable accurate and efficient genome comparison [186, 240].

In this work, we reframe the problem of sequence comparison for polyploid genomes as a metagenomic problem which enables us to take advantage of the highly scalable, alignment-free approaches the metagenomic community has developed for rapid, but accurate, sequence comparison and classification. MinHashing is a form of locality sensitive hashing which aims to place similar objects in similar bins [223]. In terms of sequence analysis, these objects are genomic sequences.

MinHashing approaches start by k-merizing the sequences (splitting the genome into overlapping subsequences of length k) and then using Hash functions to generate a numerical representation of the k-mers. These k-mers are then down-sampled, using a range of approaches. MASH uses a bottom sketch approach in which the k-mers are sorted into desending order, and the bottom n sketches are kept to form the genomic signature, where n is a user-parameter [250]. In contrast, Sourmash uses the modulo hashing approach. In this strategy k-mers are hashed to 64-bit numbers after which those numbers which fall below the maximum 64-bit values ($2^{64} - 1$) when divided by the scale factor (i.e., $\frac{2^{64}-1}{scalefactor}$) are kept [50, 270]. Scale factor is a user-controlled parameter with a default value of 1000 for Sourmash, which would downsample the genome by keeping approximately 1 in every 1000 k-mers.

Minhashing provides two different experimental avenues: k-mer composition or k-mer frequency-based comparison of sequences. K-mer composition is concerned only with documenting the k-mer spectrum present in the sequences. K-mer composition-based sketches are suited to set analysis, and statistical approaches such as the Jaccard index are used to estimate the level of shared sequence content between the sequences [250, 270]. K-mer frequency-based sketches take this a step further and record the frequency of the observed k-mers in the sequences [270]. These frequency-based sketches are better suited to vector-based statistical strategies of comparison, such as the cosine similarity. Both strategies have been extensively applied to metagenomic-based sequence comparison but little work has been done to apply these strategies to comparative crop genomics and subgenomes [2, 139, 172, 268, 300].

In this work, we explore both the k-mer composition- and frequency-based MinHashing strategies to reduce the computational overhead of large sequence analysis, and apply it to the problem of inter- and intra-subgenomic and genomic comparison to uncover patterns of polyploid genome dynamics and evolution. We also perform a comprehensive assessment of the impact of MinHashing parameters on biological discovery and assess the suitability of the direct application of MinHashing strategies developed for metagenomics for polyploid comparative genomics.

Methods

All allopolyploid, autopolyploid and segmental allopolyploid plant genomes were checked for chromosome-scale assemblies in NCBI Genbank [249]. If multiple assemblies were available, the NCBI reference assembly was chosen. If multiple species were available but were considered the same crop with the same polyploid structure (e.g. tetraploid cotton), a single domesticated and wild (if available) genome was chosen. All genomes (Table D.1) were downloaded from NCBI [249] with the exception of *Solanum tuberosum*, which was

downloaded from spudDB [141]. Chromosomes were separated from the bulk assembly file using an in-house Python script available from `https://github.com/Glfrey`.

Sourmash version 4.0.0 [270] was used to generate MinHash k-mer signatures for each chromosome for k-mers within k-mer range 4-11 and then 21-61 in increments of 10 with scale factors of 1000, 500, 250 and 125. Sourmash was then used to perform k-mer frequency and composition signatures comparison and plot similarity scores between the chromosomes.

Subgenomic clustering The ability for chromosomes to cluster subgenomically was assessed visually. If a single or double cut to the hierarchical clustering result (representing tetraploid or hexaploid genomes respectively) could result in distinct clusters containing the chromosomes of each subgenome the clustering result was deemed to be subgenomically correct. Else, it was considered non-subgenomically clustered.

Progenitor clustering Two progenitor clustering tests were performed. The first was performed in exactly the same way as the subgenomic clustering investigations. A visual assessment of subgenomic and progenitor clustering was performed for genomes with progenitors with chromosome-scale assemblies available. This included *Triticum aestivum*, its known A and D subgenome progenitors *Triticum urartu* and *Aegilops tauschii*, and its potential B subgenome progenitor *Aegilops speltoides*. All *Brassica* genomes were involved in the subgenomic analysis alongside their progenitors, *B.napus*,*B.juncea*,*B.carinata*, *B.olecera*, *B.rapa*, and *B.napus*. The peanut genome analysis included *Arachis hypogaea* and its progenitors *Arachis duranensis* and *Arachis ipaensis*. If the subgenomes and their progenitors were clustered together such that a single cut could be made to separate each of the progenitors and their donated subgenome from the rest of the clusters they were considered to be clustered according to their hybridization history.

A second test was performed using whole genome signatures in place of chromosome signatures for the species used by [200]. Whole genome signatures were generated in the same

manner and for the same ranges as the chromosomal signatures and signature comparison and clustering performed by Sourmash in the same manner also. A visual comparison of the hierarchical clustering results against the phylogenetic tree produced by [200] was performed.

Detailed clustering analysis To investigate the causal sequences behind the clustering results, and to properly identify the optimal subgenomic clustering strategy a second, more detailed analysis was performed for those sequences with repeat-masked genomic sequences available. The repeat-masked, and non-repeat-masked chromosomes were downloaded from Ensembl Plants and repeat-masking completeness assessed via a custom Python script which enumerated the number of "N"s in the assembly file.

Sourmash version 4.0.0 [270] was used to generate MinHash k-mer signatures for all sequences for odd k-mers within k-mer range 3-61 and with scale factors of 1000, 500, 250 and 125. Sourmash was then used to perform signature comparison between the chromosomes for both k-mer frequency and composition.

The resulting csv file containing pairwise similarity scores was then downloaded and analysed using R Studio 3.6.1(2019-07-05) [296]. Hierarchical clustering was performed for all k-mers and all scale factors using hclust from the stats package for single, complete, average and Ward.D linkage schemes. The resulting dendogram that best fit the underlying similarity matrix was determined using cophenetic correlation strategy from the stats package. Final clustering with the appropriate scheme and heatmap construction was performed using ComplexHeatmap [127]. Cuts equal to those expected from the number of subgenomes was then performed using ComplexHeatmap and a visual check of chromosome membership to expected subgenomic clusters was performed. If the chromosomes were correctly clustered according to their chromosomes, the dendrogram cut height was recorded for each cut used to separate the subgenomes. For a tetraploid this would be a single cut to separate the two subgenomes, where for a hexaploid this would be two cuts.

Given that *T.aestivum* is an allohexaploid for which the hybridization timescales and evolutionary relationships are known, the subgenomic clustering correctness is also assessed by the ability for subgenomes A and B to cluster together, with the D subgenome as an outgroup.

Plots describing the relationship between k-mers and dendrogram cut height and cophenetic correlation were constructed using ggplot2 [362].

Results

Subgenomic clustering: k-mer frequency Nearly all allopolyploids show some level of subgenomic clustering across the sampled k-mer frequencies from k=7 onwards. No polyploids showed any clustering ability before k=7, with similarity scores being identical for all chromosomes. *T.dicoccoides, T. turgidum, B. carinata, G.hirsutum* and *G.tomentosum* clustered subgenomically across all sampled k-mer ranges from k=7-k=61.Figures 2.3 (A-C), 2.4 (A,B) and 2.5 (A-C) give an example of these subgenomic clustering results.

These figures show that while these species all cluster, they do so in different ways. Where the *Triticum spp* show exceptionally high intra-subgenomic similarity and high inter subgenomic similarity (figure 2.3 A-C), others exhibit far less sequence similarity (e.g. *Brassica spp*, figure 2.4 A-C), and others exhibit unequal subgenome similarity (*Brassica spp*, figure 2.4 A-C), *Gossypium spp*, figure 2.5 A-B).

T.aestivum clustered subgenomically for all sampled k-mer ranges from k=8 onwards, with k=7 showing subgenomic clustering with the exception of chromosome 4B, which formed an outgroup. *B.juncea* and *B.napus* subgenomically clustered for 6 of the sampled k-mer ranges. The ability of *B.juncea* to subgenomically cluster was interrupted by chromosome A1, which became an outgroup for k=8, 9 and 10 and chromosomes 5 and 6 which were outgroups at k=11. The ability of *B.napus* to subgenomically cluster was also interrupted by various outgrouping chromosomes. At k=7 *B.napus* also showed no subgenomic clustering

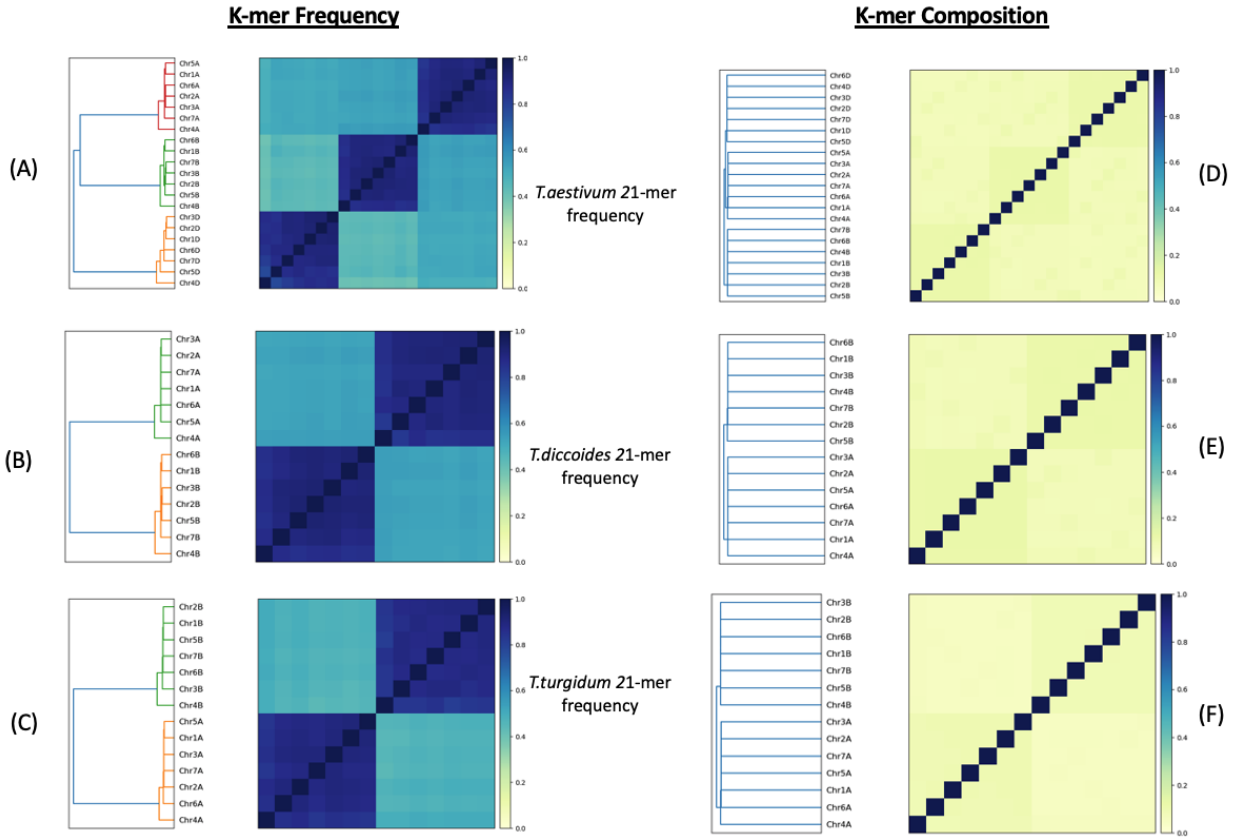


Figure 2.3: Sourmash produced dendrograms and heatmaps for *T. aestivum*, *T. diccoides* and *T. turgidum* 21-mer frequency (A-C) and composition (D-F)

due to chromosomes being placed in the wrong subgenome (chromosomes A5 and A3) and chromosome A10 acting as an outgroup. *A.hypogea* was only able to subgenomically cluster for k=21-61. Before k=21, *A.hypogeas'* subgenomic clustering ability is punctuated by chromosomes being clustered in with the wrong subgenomes and outgrouping chromosomes. Chromosome 8 was a notable outgroup, appearing at k=7-10 (Figure 2.7).

The difference between the subgenomic clustering ability of the segmental allopolyploids, *A.hypogea* and *C. arabica* can be visualised in Figure 2.7. While *A.hypogea* exhibits very high intra-subgenomic similarity, with the exception of the A8 chromosome, and high inter-subgenomic similarity, *C. arabica* exhibits a subgenomic similarity for the C subgenome

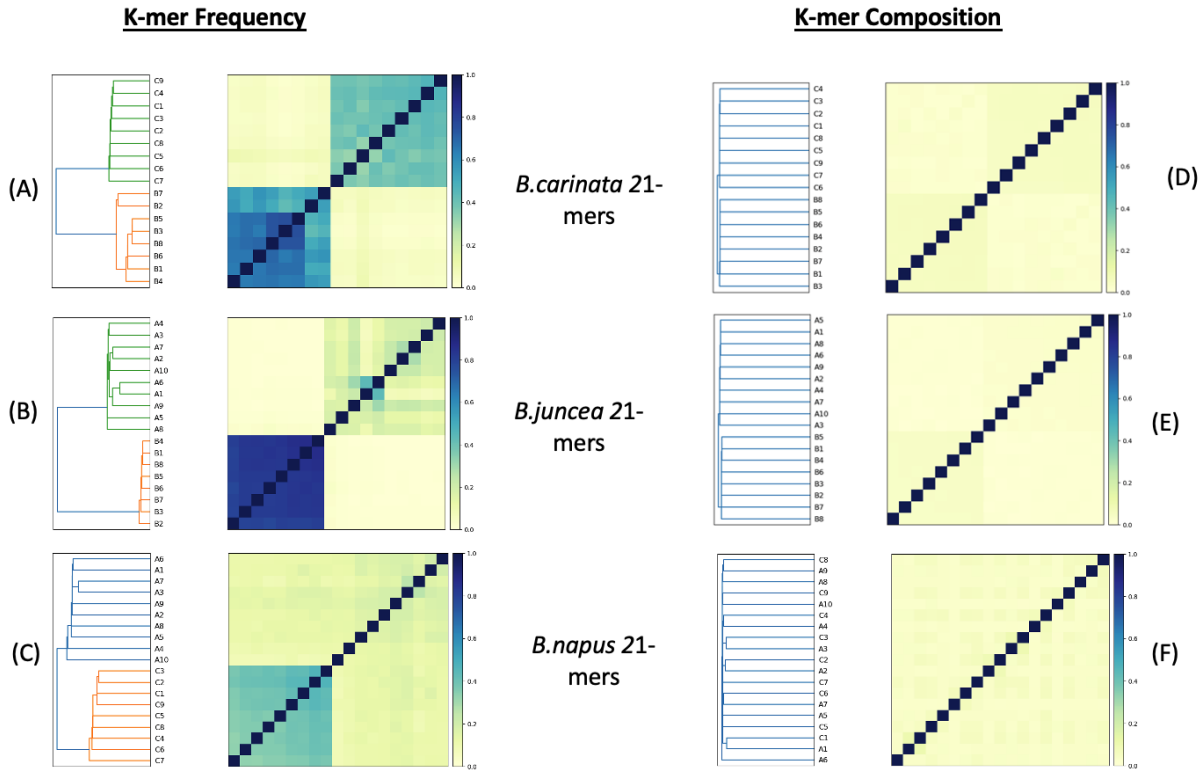


Figure 2.4: Sourmash produced dendrograms and heatmaps for *B. carinata*,*B.juncea* and *B.napus* 21-mer frequency (A-C) and composition (D-F)

that is barely perceptible against the inter subgenomic similarity. The E subgenome also show low intra-subgenomic similarity. However, a subgenomic clustering structure is present with the A and C genomes clustered separately, but this structure is interrupted by the 7E and 3E chromosomes which form outgroups.

Four of the allopolyploids and both autopolyploids showed no ability to subgenomically cluster using the visual cut criteria (Figures 2.8 and 2.9). Neither *E.tef, C.sativa*, nor *P.virgatum* or either of the autopolyploids showed any subgenomic clustering structures throughout the sampled k-mer ranges. However, while *A.sativa* did show a subgenomic clustering structure, they couldn't be cut due to the A and D subgenomes consistently

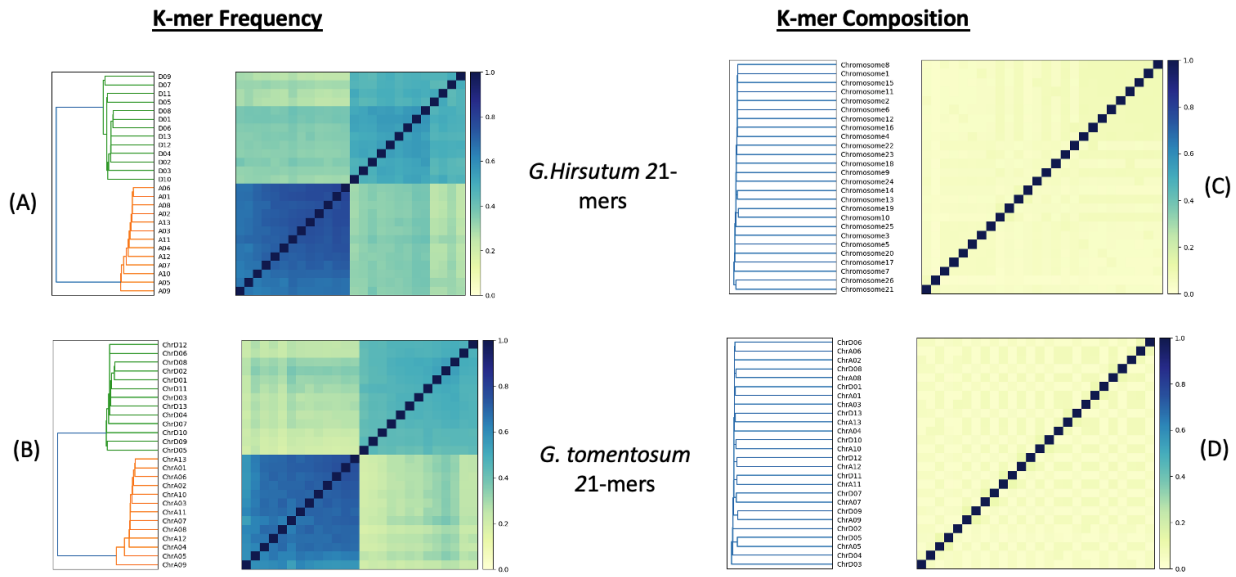


Figure 2.5: Sourmash produced dendrograms and heatmaps for *G.hirsutum* and *G.tomentosum* 21-mer frequency (A,B) and composition (C,D)

intermixing (Figure 2.9 D). Chromosome 1A was especially prone to clustering with the D subgenome throughout the sampled k-mer ranges.

Subgenomic clustering: k-mer composition Only 5 of the polyploid genomes showed any k-mer composition-based subgenomic clustering capability. These were two of the *Brassica spp* (*B. carinata* and *B. juncea*) and all of the *Triticum* species. The *Triticum spp* showed a sequence similarity of 1 for all chromosomes for k=7-11 while the two *Brassica spp* showed a non-subgenomic clustering structure for k=10 and 11. Sequence similarity was 0 for all chromosomes for all smaller k-mers. From k=21 onwards, these species showed subgenomic clustering structures as shown in figures 2.3 (D,E) and 2.4 (D-E). Both heatmaps in the figures show very low sequence similarity both intra- and inter-subgenomically.

In place of subgenomic clustering, the remaining polyploids showed homeologous chromosome clustering patterns from k=21 onwards. For k=9-11 *A. hypogea, B. napus,*

C. arabica, C. sativa, both *Gossypium spp* and *S. tuberosum* showed no homeologous or subgenomic clustering structures. The same was true but only for k=11 onwards for *P. virgatum, S. spontaneum, E. teff* and *A. sativa*. Below these k-mer values all sequences showed identical sequence similarity except for k=4-7 which showed no sequence similarity at all.

Progenitor clustering: k-mer frequency All progenitor subgenomic clustering showed results from k=7 onwards. Prior to k=7 there is no sequence similarity.

T.aestivum showed clustering results with the progenitor chromosomes consistent with its evolutionary history with the exception of the potential B subgenome progenitor *A. speltoides* [154] (figure 2.10). Throughout the sampled k-mer range only k=7 and k=10 resulted in the A and D subgenome/progenitor pairs not clustering as expected due to the presence of outgroup chromosomes, with *T. aestivum* 4A, 4B and *T. urartu* chromosome 4 outgrouping at k=7 and *T. aestivum* 5D chromosome outgrouping at k=10. Both the A and D progenitor genomes show differing clustering patterns with their donated subgenomes. *T. urartu* and *T. aestivum* A subgenome cluster together but in distinct subclusters where *A. speltoides* and *T. aestivum* D subgenome cluster together in homeologous pairings. This contrasts with the relationship between *T. aestivum* B subgenome and potential progenitor species *A. speltoides* [200] for which there are k-mer-size dependent changes in clustering patterns. For k=4-11 *A. speltoides* clusters with *T. aestivums* B subgenome only twice at k=9 and 10, although it does remain clustered in this fashion from k=21-61. However, even when clustered together it is clear from both the reduced colour in the heatmap and longer branch lengths in figure 2.10A, that they do not exhibit the same k-mer frequency similarity that other progenitor genomes and subgenomic pairs do.

All of the *Brassica* genomes showed some level of subgenomic/progenitor clustering with clearly visible unequal subgenomic similarity (figure 2.12 A-C). The relationships between

the *Brassica spp* and their progenitors can be seen in figure 2.13 which is stylized in their triangle of "U" evolutionary relationships. This relationship describes the hybridisation between three economically important diploids *B. nigra, B. rapa,B. oleracea*), to form the three tetraploid crops; *B. carinata, B. juncea,B. napus* [369].

B. carinata, its B genome progenitor *B. nigra* and its C genome progenitor *B. oleracea* form subgenomic/progenitor clustering pairs with the exception of k11 and k61 where at k=11 *B. nigra* clustered with the C genome and at k=61 the *B. carinata* C subgenome split from its progenitor and clustered in with the B subgenome and progenitor. As the k-mer size grows, the uneven sequence similarity becomes clearer with *B. carinata* showing greater sequence similarity to it's B subgenome progenitor than its C subgenome progenitor.

B. napus, which also shares the subgenome progenitor *B. oleracea* also experiences the C subgenome/progenitor split at k=61 which *B. carinata* also shows. Unlike *B. carinata* however it doesn't show subgenomic/progenitor clustering until k=11, with k=11 subgenomic/progenitor clustering interrupted by A01 and A08 outgrouping. *B. napus'* subgenome C chromosomes shows a greater sequence similarity to their progenitor *B. oleracea* than subgenome A shows to their progenitor *B. rapa*. Conversely, *B. carinata* shows lower subgenome/progenitor similarity for its C subgenome than its B subgenome and progenitor *B. nigra*.

B. juncea shares the same B subgenome progenitor as *B. carinata* and also exhibits greater B subgenome/progenitor similarity than can be seen for its A subgenome and progenitor *B. rapa*. *B. napus* also shows this lower A subgenome/progenitor relationship. *B. juncea* shows subgenomic/progenitor clustering from k=7 but this is interrupted for k=8-9, due to chromosome A01 being an outgroup and for k=31 due to the B subgenome clustering away from its progenitor *B. nigra* and with the C subgenome and progenitor instead.

The unequal subgenome/progenitor similarities for all *Brassica* species are represented in figure 2.13. The arrows coloured red which are accompanied by the minus sign indicates a

lower sequence similarity, and the arrows coloured green accompanied by a plus sign indicates the higher subgenome similarity for the 3 tetraploid genomes.

A. hypogaea showed subgenome/progenitor clustering patterns from k=21 onwards which can be seen for 21-mer frequency in figure 2.14 A. Chromosome 08 for both *A. hypogaea* and its A-genome progenitor *A. durensis* [39], are intra-subgenomic outgroups and exhibit lower intra-subgenomic similarity than the other chromosomes. For k=9-11, subgenomic clustering patterns can be seen, but subgenomic clustering is interrupted by the *A. hypogaea* A8 and *A. durensis* chromosome 8, which are outgroups for k=9-10.

Progenitor clustering: k-mer composition All *Brassica* genomes, the *A. hypogea* genome and their progenitors lose the ability to subgenomically cluster for k-mer composition, instead forming subgenome-progenitor homeologous chromosomal pairings (figures 2.12 C-F, 2.14 B).

All progenitor genomes cluster with their appropriate subgenome from *Triticum aestivum* throughout the sampled k-mer spectra from k=11 onwards. Prior to this, the sequence similarity scores are identical for all chromosomes (k=4-6=0, k=7-10=1). However, much like k-mer frequency *A. speltoides* and the B subgenome exhibit a distinctly reduced level of similarity between them, than can be seen for the other subgenome-progenitor pairs (figure 2.10 B).

All *Brassica* genomes show either clear homeologous clustering (*B. napus* and *B. juncea*) or chromosomal pair clustering (*B. carinata*) with their progenitor genomes from either k=11 (*B.napus*) or k=21 (*B.juncea*, *B. carinata*) onward (figure 2.12). *A. hypogaea* and its progenitors *A. ipaensis* and *A. durensis* show clear homeologous clustering from k=21 onward (figure 2.14 B) . *A. hypogaea* also has unequal subgenomic/progenitor similarity with the *A. hypogaea* B subgenome showing greater sequence similarity to the progenitor genome *A. ipaensis* than the A genome shows to its progenitor *A. durensis*.

Progenitor clustering: Phylogenetic tree comparison The k-mer frequency hierarchical clustering dendograms produced via the Sourmash approach do show a largely similar node structure to the maximum likelihood phylogenetic tree produced by [200] using single copy orthologous genes. However there are some important differences (figure 2.11). The first is that *A.speltoides* exhibits k-mer dependent clustering patterns, with some k-mer sizes showing it clustering with the other *Aegilops* species, others with the B subgenomes and others as a form of outgroup species within the tree structure. The second is that the clustering results do not show the D genomes clustered with the other *Aegilops* for any k-mer size. The most similar tree structure was produced for k=21 and can be seen in figure 2.11.

In contrast the k-mer composition hierarchical clustering diagrams produced via the Sourmash approach produces a phylogenetic tree that is not similar to the phylogenetic tree. While *A. speltoides* no longer exhibits k-mer dependent clustering patterns, the species designed to be outgroups (*T. elongatum, H. vulgare and B. distachyon)* are prone to clustering within the main tree structure. *T. elongatum* clusters within the tree from k=11 onward, where *H. vulgare* and *B. distachyon* cluster in from k=41 onward.

Identification of causal subgenomic signatures To understand if a particular class of sequences were driving the subgenomic clustering result an investigation into the sequences responsible was performed. Given that repetitive-elements are rapidly evolving sequences which make up a sizeable portion of most of the genomes under investigation in this work (table D.1), they were naturally under suspicion [46]. As such, a repeat of the Sourmash procedure with repeat-masked sequences was performed.

With the use of repeat-masked sequences, *T.aestivum* and *T.dicoccoides* completely lose the ability to subgenomically cluster. These results can be seen in figure 2.15, which shows clustering results for *T.aesitvum* (A, D), *T.dicoccoides* (B, E) and *T.turgidum* (C,F)

with repeat-rich and repeat-masked sequence-based clustering results. Where the repeat-rich sequences show the subgenomic clustering as previously described, the repeat-masked sequences show homeologous clustering structures in place of the subgenomic clustering structures which are present in the presence of repeats, for *T.aesitvum* and *T.dicoccoides*. The sequences also lose their sequence similarity, with barely discernible sequence similarities between the homeologous clusters which contrasts with the very high intra-subgenomic, and moderately high inter-subgenomic sequence similarities seen in the heatmap for figure 2.15 (A,B). The repeat-masked percentage for the *T.aestivum* and *T.dicoccoides* genome sequences was 86.99%, 87.64%. The repeat-masked percentage as given in the genome assembly publications can be seen in table D1.

In contrast, *T.turgidum* (figure 2.5 C,F) maintained its subgenomic clustering ability with the loss of repeat sequences, although there is a marked loss of sequence similarity both intra- and inter-subgenomic relationships. The repeat masked proportion of the *T.turgidum* was determined to be 75.45%, which contrasts with the published repeat-masked percentage give in table D1.

Characterisation of parameters on clustering structures The largest MinHashing parameter that affects subgenomic clustering is whether k-mer frequency or composition is used. This is reflected in the relationship between the use of k-mer frequency or composition-based analysis and dendrogram cut height. Dendrogram cut height is the height within the dendrogram at which the clusters are cut, with a greater height corresponding to greater sequence dissimilarity, and vice versa for smaller cut heights.

K-mer frequency shows a gradual, positive, linear relationship between k-mer size and subgenomic cut height (figure 2.16, dashed lines). Cut height represents the distance between each subgenome, as calculated by either the Jaccard (k-mer composition) or cosine (k-mer frequency) similarity metric. Where the cut height is 0, no subgenomic clustering was

possible.

The rise in subgenomic cut height is more dramatic for smaller (k<17), where after the rise continues at a steady pace for all scale factors. All species show similar cut heights as a function of k-mer size although with their own patterns. Where *T.dicoccoides* and *T.turgidum* exhibit similar patterns in cut height as k-mer size grows, *T.aestivum* shows a distinctive "zig-zag" pattern, which reflects a rising and falling of subgenomic cut height for every increment in k-mer size. *T.turgidum* repeat-masked sequences show the same positive linear relationship between k-mer size and subgenomic cut height except it is less gradual (figure 2.16 D). Where the repeat-rich k-mer frequencies reach between 0.60-0.75, the repeat-masked subgenomic cut height reaches over 0.80.

For k-mer composition (figure 2.16 solid lines), the subgenomic cut height increases more rapidly than is seen for k-mer composition, reaching around 0.8 for all species, with the exception of scale 125 for *T.dicoccoides* which loses its ability to subgenomically cluster at this k-mer size. As k-mer size grows, the cut height reaches almost 1 for all species by k=61. For *T.turgidums* repeat-masked k-mer composition, the cut height is almost 1 as soon as a subgenomic clustering pattern is shown, at k=35 (2.16 D, solid line).

For *T.aestivum*, there is an extra layer of subgenomic clustering correctness given the evolutionary relationship between its three subgenomes. The relationship between the subgenomic sequences should reflect that the A and B subgenomes hybridised some 0.3- 0.5 million years ago prior to the hybridisation with the D subgenome approximately 9,000 years ago[283]. As such, figure 2.16 A shows the subgenomic cut height in bold colours only when the subgenomes are clustered by their evolutionary history (i.e. A clustered with B with D clustered separately). For k-mer frequency, this occurred almost always with only k=17 showing incorrect clustering for all scale factors, k=47 for scale factor 500 and k=51 for scale factor 1000. For k-mer composition however, all except k=17 showed incorrect subgenomic clustering between the subgenomes.

Assessment of direct applicability of metagenomics-based minhash approaches The cophenetic correlation is a metric used to assess how faithfully the dendrogram represents the data held in the underlying similarity or dissimilarity matrix [301]. Generally, a cophenetic correlation of ≥ 0.9 is excellent where ≤ 0.7 is considered very poor [357].

A cophenetic correlation-based assessment of suitability of the single-linkage hierarchical clustering strategy implemented by Sourmash shows that while this hierarchical clustering approach is rarely the optimal strategy, the difference in cophenetic correlations between the often optimal average-linkage strategy and single-linkage are small (figure 2.17). The vast majority of cophenetic correlation scores across the k-mer spectrum (for k>7) were almost 1 for all linkage strategies for k-mer frequency and so can be considered an excellent representation of the original similarity scores. For k-mer composition there's a rapid rise to at least 0.9 for k=13, followed by a large drop for k=15 where after the cophenetic correlation rises rapidly as a function of k-mer size, reaching scores of almost 1 by k=25 for *T. dicoccoides* and *T.turgidum*, and by k=35 for *T. aestivum*. *T. dicoccoides* experiences a small drop in cophenetic correlation for k=37, which corresponds with a loss of subgenomic clustering as seen in figure 2.16 C.

There were only 6 instances, out of the 1440 clustering instances, where the single linkage strategy was identified as the optimal strategy. For *T.aestivum* these were k11 in scales 500, 250 and 125, for *T.dicoccoides* k13 in scale 1000, and k5 in scale 125 and for *T.turgidum*, k7 in scale 250. All of these were for repeat-masked k-mer frequencies and none of them produced a subgenomic clustering result.

Figure 2.18, which shows cophenetic correlations for the repeat-masked sequences show far more instability, especially for k<17. The Ward linkage strategy performs especially poorly for k-mer frequency, although this is most pronounced for *T.turgidum* and less pronounced for *T.aestivum*. For *T.dicoccoides*, the comprehensive and Ward strategies perform especially well with scores of almost 1 for k< 10, but this is only for scale factor

500. For *T.turgidum* all linkage methods achieve cophenetic correlations of almost 1 with the exception of Ward linkage. *T.aestivum* exhibits the most stable scores for k<17 although cophenetic correlation steadily declines as a function of k-mer size, which is not seen for the other two species .

The repeat-masked k-mer composition plots look similar to the repeat-rich k-mer composition plots in that there is a drop in correlation between k=11-17. However, unlike the repeat-rich sequences, cophenetic correlation declines slightly as a function of increasing k-mer size for *T.aestivum* and *T. dicoccoides*. However for *T.turgidum* the repeat-masked plot looks highly similar to the repeat-rich plot, exhibiting cophenetic correlations of almost 1 for k>27.

It is evident from the repeat-rich plots that scale factor has little impact on the cophenetic correlation of k-mer frequency and composition. For the repeat-masked sequences, there is more scale-factor dependence, especially for *T.dicoccoides* and for the smaller (k<17) k-mer sizes.

Discussion

It is clear, from the comprehensive investigation performed into the application of a whole-genome MinHash sketching approach for comparative genomics of polyploid crops, that this approach is capable of uncovering evolutionary relationships previously described in the literature. This includes not only subgenomic relationships but also the subgenome and progenitor relationships which follow their hybridisation history. Further, through the synchronous investigation of how these relationships change with the gain and loss of k-mer frequency and repeat information, this work has uncovered patterns of subgenome dynamics which provides layers of evidence as to the biological sequences responsible for the clustering patterns observed.

Start different, stay different The ability for chromosomes to subgenomically cluster evidently follows a biologically-led pattern, with no autopolyploids, half of the segmental allopolyploids, and two thirds of allopolyploids exhibiting a subgenomic clustering structure for k-mer frequency (figures 2.3(A-C), 2.4(A,B), 2.5(A-C), 2.7(A,B), 2.8(A,B)). For k-mer composition, only 5 of the allopolyploids exhibit subgenomic clustering out of all genomes surveyed (figures 2.3(D-F), 2.4(C,D), 2.5(D-F), 2.7(C,D), 2.8(C,D), 2.9(E-H)).

This pattern of subgenomic clustering across polyploidy types has been observed before by [159] who, similarly to [123] developed a subgenome-specific k-mer-based clustering method for polyploid chromosomes. Importantly, [159] also observed a number of allopolyploid genomes showed an inability to cluster chromosomes subgenomically although these were not the same ones used in this work.

Unlike [159] and [123] who implemented subgenome-specific k-mers to cluster chromosomes, this work used global chromosomal k-mer signatures. The use of global chromosomal signatures has the advantage of keeping computational costs low by not implementing k-mer set comparisons to identify subgenome-specific sequences, but as a less specific signatures, they could be impacted by noisy, non-subgenome-specific k-mers. Further, the use of k-means clustering by [159] results in distinct clusters where the hierarchical clustering strategy implemented in this work can make for more ambiguous clustering results.

As the interpretation of the number of clusters in a dendrogram can be subjective, the ability for chromosomes to subgenomically cluster was determined by the ability to make a number of cuts, equal to the number of expected subgenomes -1 (i.e. make 1 cut for a tetraploid, 2 for a hexaploid which results in the dendrogram being split 2 and 3 ways respectively). This then results in the separated clusters containing only, and all, chromosomes belonging to each subgenome. If this strategy was to be relaxed to allow slightly more cuts than the number of subgenomes, but still required that the clusters that the clusters contain only members of the same subgenome, then the other segmental

allopolyploid genome of (*C.arabica*) could also be considered capable of subgenomically clustering (figure 2.7 B). Importantly, the method implemented by[159] showed a subgenomic clustering structure for *C.arabica* which indicates that for this genome, the use of subgenome-specific k-mers in place of global chromosomal signatures is desirable.

The dependence of subgenomic clustering ability on polyploidy-type classification are perhaps unsurprising given the taxonomic definition of polyploidy, in which autopolyploidy, segmental allopolyploidy and allopolyploidy describe a genome doubling or hybridization event between increasingly distinct taxa [321]. Where autopolyploidy generally refers to an intra-species genome doubling, allopolyploidy generally results as a result of hybridization of distinct taxa [240, 321]. This subgenomic distinctness is of great importance genetically given that meiotic recombination is driven by chromosomal sequence and structural similarity [305]. As such autopolyploids are capable of polysomic inheritance facilitating exchange of genomic material between the subgenomes, where allopolyploids are capable only of disomic inheritance, with genomic material only being exchanged intra-subgenomically [240, 305, 321]. Of course, given that polyploidy is a continuum rather than classification some allopolyploids are capable of polysomic inheritance [41, 226, 305]. However, with distinct enough subgenomes, possibly in combination with certain loci (e.g. wheats *Ph* and Brassica *PrBn* genes) which ensure bivalent pairing during meiosis, distinct subgenomes can remain that way [191, 226, 321].

As such, in allopolyploids in which there is a little to no exchange of subgenomic information on a large enough scale to disrupt a chromosomal signature, the subgenomic clustering results in which the chromosomes cluster by subgenome are intuitive. They simply have more sequence in common intra-subgenomically than inter-subgenomically.

This does raise the question of how large a subgenomic exchange of information would have to be to be detectable via this MinHash clustering strategy. While this is outside the scope of this work in particular, anomalous clustering results, generated by *G.hirsutum* with

accession GCF_000987745.1 may present an interesting place to start (figure B.1). In this genome assembly accession, chromosomes 3,5 and 7, 15, 16 and 23 appear to have swapped subgenomes and chromosome 26 is a persistent outlier. It is possible that these anomalies are biological in nature, however as one of the older assemblies for *G.hirsutum*, which utilises only short-read sequencing data and only utilised linkage mapping for chromosomal placement of contigs with no manual correction of the assemblies described, it is perhaps more likely that these chromosomes feature significant misassemblies [197].

Repeat after me... While the notion of the distinctness of subgenome sequence composition explains the diminishing ability for subgenomes to cluster based on their polyploidy-type classification, it does not explain the variation observed in clustering abilities and patterns of sequence similarity for the allopolyploids in this study. A hybridisation between two distinct subgenomes explain only why they cluster separately, not why they do or do not exhibit asymmetric subgenomic similarities or why they lose the ability to subgenomically cluster when k-mer composition is used in place of k-mer frequency. Only 5 of the allopolyploids, out of all genomes studied showed any subgenomic clustering ability for k-mer composition, with all other previously subgenomic clustering genomes exhibiting some level of homeologous or other chromosomal pairing clustering structures instead (figures 2.3(D-F), 2.4(C,D), 2.5(D-F), 2.7(C,D), 2.8(C,D), 2.9(E-H)).

To investigate this further, repeat-masked genomes for the subgenomically clustering species were searched with only the *Triticum* genus having repeat-masked genomes available and only via Ensemble plants. These genomes were subject to the same Sourmash protocol as repeat-rich genomes, which revealed a distinct loss of subgenomic clustering ability for *T.aestivum* and *T.dicoccoides* for both k-mer frequency and composition (figure 2.15 (A-E). *T.turgidum* retained an ability to subgenomically cluster, albeit with reduced sequence similarity (figure 2.15 C,F). This is seemingly explainable by inadequate repeat masking. A

custom python script, which enumerates the masked percentage of the sequence genomes, revealed that while *T.aestivum* and *T. dicoccoides* has a repeat-masked percentage in line with their published repeat content (an additional 1.88% and 5.2% repeat-masked sequence for *T.aestivum* and *T. dicoccoides*), *T.turgidum* does not (6.75% loss of expected repeat-masked sequences) (table D.1). It is important to note that the actual repeat-content for *T. dicoccoides* may be higher than given in the assembly publication [21] (table D1). This is due to an improvement of the genome assembly performed by [393].

The dependence of the subgenomic clustering results on repeat sequences is understandable given the evolutionary history of *T.aestivums* transposable elements. [360] found that while the three subgenomes show conserved TE abundance at the family level, the subfamily level shows strong subgenome-specific enrichment or depletion which was inherited from their progenitor genomes. Further, for *T.aestivum*, the lack of any major TE-amplification event at the time of hexaploidisation, but an amplification shared between the A and B subgenomes event during tetraploidisation explains why the A and B subgenomes show greater similarity to each other than either do the D subgenome (figure 2.3 A).

For *T.aesitvum* k-mer composition however, the A and D subgenomes can frequently be found clustering together, with the B subgenome as an outgroup. This is visualised in figure 2.16A (solid line), where only a single point has colour, which indicates the sole position for which the subgenomes clustered according to their hybridisation history (AB, D). This may be explainable by a B genome-specific TE amplification burst. The B subgenome progenitor, which as of yet remains unknown, appears to have undergone a wave of TE amplification for the transposable element Ty1/copia 1.2 million years ago [21]. This date places it firmly before the tetraploidisation that formed wild emmer wheat which would go on, much later to form hexaploid bread wheat [263, 382].

This dependence on repeat sequences for clustering results may explain the failure of *E.teff* and *C.sativa* to subgenomically cluster (figure 2.9 A,B,E,F) as they both have low

genomic repeat-contents (20.0% and 28% respectively) [167, 342]. This is far lower than the repeat content of any of the other genomes in this study (table D.1). It may be tempting to infer that, given the *C.sativa* genome was assembled using only short-read sequencing data which is particularly prone to collapsing repeats, and potentially the lack of repeat sequences could be a technical artifact [167]. However, many other genomes in this work, including the *Triticum spp*, were assembled using only short-read sequencing data (table F4), and yet exhibit subgenomic clustering results. Further, several genomes, including *B.rapa* which was sequenced using a PacBio approach show a smaller value for genome assembly completeness (table F4). As such, the low repeat content may be biological in nature, especially as closely related species *Arabidopsis thaliana* and *Arabidopsis lyrata* genomes contain comparable proportions of repeat sequences [167]. As for the well-assembled genome (92.6%) *E.teff* genome, which used a combination of short and long read sequencing technology, this repeat content is very unlikely to be a technical artifact. Further, the repeat content is similar to that of its close relative *Eragrostis curvula* [57].

A problematically low repeat content is also the driver of the strangely clustering chromosome 8 of *A.hypogaea* and *A.durensis* (figure 2.6). These chromosomes have a much lower percentage repeat content, 49.76% and 44.32% respectively, than the rest of the chromosomes repeat content which ranges from 72.66-77.98% for the *A.hypogaeas* A subgenome (with the B subgenome being even higher) and 49.14-56.67 for *A.durensis [38].*

Further evidence for the repeatome-driven clustering results can be found for the subgenomes exhibiting asymmetric subgenome similarities. *B.napus, B.juncea, G.tomestosum* and *G.hirsutum* are all documented to have subgenomes with asymmetric transposable element (TE) content, all of which have all been reproduced here with the TE-heavy subgenomes showing greater sequence similarity [61, 63, 257, 325] (figures 2.4(B,C), 2.5(A,B)). In *B.napus*, [61] documents that the C subgenome contains the greater proportion of TEs, while for *B.juncea*, it is the B subgenome which has experienced TE expansion [256].

In both *G. tomestosum* and *G. hirsutum* the A subgenome is TE-heavy when compared their D subgenome.

Interestingly, *B. carinata* is documented to have a largely equal subgenome repeatome size, but the sequence similarity reflected in the heatmaps and dendrograms in this work show a slight dissimilarity, with the C subgenome showing a lower intra-subgenomic similarity than the B subgenome [370] (figure 2.3(A)). This is particularly curious because although their TE abundances are relatively equal (49.95% and 50.61% for B and C respectively), it is the C subgenome which has the very slightly higher TE content [370]. However, while the C subgenome is indeed more repeat-rich in general, the B subgenome contains the greater proportion of long terminal repeat (LTR) retrotransposon sequences, comprising 29.9% and 26.0% of the B and C subgenomes respectively [370]. LTR retrotransposons frequently make up the vast majority of the TE-space in plant genomes and thus the vast majority of the plant genome itself for TE-heavy genomes. LTRs are implicated as potential regulatory molecules for plant stress and have been associated with traits of agronomic interest such as rice grain width and hybrid weakness [125, 209, 239, 390]. However given that subgenome asymmetry LTR information is from a different cultivar and genome assembly accession, and that transposable element content can be highly variable even on an intra-species level, the implication of LTRs as drivers of the clustering results for *B. carinata* would require further investigation [46, 58, 308].

That being said the *A. sativa* genome also provides a layer of evidence for the LTR sequences being the molecules underpinning the clustering results. In this species clustering results, the A and D subgenomes are found frequently clustering together (figure 2.9 D), which is in conflict with their evolutionary history as, like hexaploid wheat, the C and D subgenomes first formed a tetraploid genome, before the diploid A subgenome hybridized later [65, 108]. However, the C subgenome has a 1.3 fold increase in the number of full length LTR sequences when compared to the A and D subgenome, which may be what is driving

the clustering result which conflicts with subgenome hydridisation history [168]. Importantly though, the B subgenome also has a greater number of TE sequences in general and so the general TE landscape cannot be ruled out as a the driver of these clustering results.

There's always one The lack of subgenomic clustering in *P. virgatums* is not so straightforward to explain (figure 2.9 (C,G)). The genome is larger than some of the subgenomically clustering allopolyploids and contains a high repeat content (56.9%) (table D.1). Additionally, subgenome-specific repeats were used to subgenomically assign the chromosomes to their subgenome of origin in the publication for this genome [211]. In this method, they used a k-mer size not used in the investigations performed for *P. virgatums* and so the Sourmash clustering strategy was applied at this k-mer size for scale factors 1000, 500, 250, 125 and 50, none of which showed subgenomic clustering for k-mer composition or frequency (figures C.1, D.1). As such, it may be that the subgenome-specific repeat sequence signals are being masked by other, more noisy, chromosomal k-mers. *P.virgatum* does stand out from the other allopolyploids in this study as one of the oldest allopolyploid by far (table D.3). Given that almost complete TE-turnover has been observed for plants that have diverged between 1.7-2.0 million years ago and complete turnover for plants that diverged 6.4 million years ago it may be reasonable to infer that the subgenomic landscape has completely altered since the hybridization of the *P. virgatum* subgenomes [28, 345, 361]. However the *A. sativa* hybridisation is estimated to be even older, with the tetraploidization occurring 10.6 mya and the hexaploidisation occurring 7.4 mya [65, 108]. As such, the inability for *P. virgatum* to cluster subgenomically remains a mystery

Progenitor clustering The subgenome/progenitor relationships highlighted in figures 2.10(A), 2.12 (A-C), 2.14(A,B)are further evidence that the minhash sketching approach to comparative genomics produces evolutionary sound, repeatome-driven, clustering results.

A. hypogaea shows the a subgenome/progenitor clustering in line with its evolutionary

history. *A. durensis* clusters with the *A. hypogaeas* A subgenome, and *A. ipaensis* clusters with the *A. hypogaeas* B subgenome from k=21 onward (figure 2.14A). K-mer sizes 9-11 also show subgenomic clustering patterns but with *A. durensis'* chromosome 8 and *A. hypogaeas* chromosome 8 acting as outgroups.

The *Brassica spp* exhibit subgenome/progenitor cluster relationships which are both in line with their evolutionary history and repeatome content as shown through their asymmetric sequence similarity for k-mer frequency (figure 2.12(A-C)) [369]. Given that the *Brassica spp* share subgenome progenitors, the observed asymmetrical sequence similarity has been summarised in figure 2.13. The difference in repeatome content between progenitors and subgenomes can be seen in table D.1 and shows a marked difference for all except *B. rapa* and *B. oleracea*, which together are the progenitors of *B. napus*. These two genomes show a small TE content difference of only 0.31, despite which, the subgenomes and progenitors cluster as expected. They both exhibit very different progenitor relationships, with the *B. rapa* and *B. napus* A subgenomes forming homeologous pairings from k=31-51 while the *B. napus* C subgenome and *B. oleracea* cluster separately (figure 2.12 C). The ability for *B. napus* and progenitors to cluster according to their evolutionary history, even in the presence of similar repeat content, indicates there is perhaps a distinct subgenome/progenitor sequence composition. Given that the sequence composition of transposable elements is very diverse, which allows their division into multiple types, and subtypes, this is certainly possible [390]. A comprehensive survey of the TE landscape would be able to determine if this is the case for *B. napus*.

Another interesting finding for *B. napus* is that while *B. oleracea* shows high sequence similarity in the *B. napus* dendrograms and heatmaps, the similarity is lower for the progenitor-subgenome in *B. carinata* (figure 2.12 (A,C) and figure 2.13) . It is currently thought that *B. napus* and *B. carinata* hybridised with a morphologically distinct *B. oleracea* genome which may be this factor driving the difference in cluster results [370]. However,

the *B .oleracea* genome used in this work is the kale-like type TO1000DH, which according to [370] is the type responsible for the *B. carinata* tetraploidisation, where a cabbage-type was responsible for *B. napus'* tetraploidy event. If this was the driver of the relationship difference it would be expected that the *B. carinata* C subgenome exhibits greater sequence similarity to *B. oleracea* than *B. napus* when, in fact, the opposite is true (figure 2.13). This pattern may also be driven by genome-specific evolution of the subgenomes. However, neither subgenome-pair show a different relationship between their C subgenomic sequences and *B. oleracea*, which would be expected if the *B. oleracea* genome used in this study was more similar to the one that originally form the tetraploid genomes. However, the similarity is not high for *B. napus* and *B. oleracea*, especially as k-mer size grows large (>k31). *B. napus* and *B. carinata* also share an outgrouping of *B. oleracea* at k=61. While this is inexplicable from a biological point of view, the *B. oleracea* genome is the least complete of all genomes in the study (table D4), and although it was sequenced using using long PacBio sequences, which may mean the repeat sequences are more intact than they would be with short-read data, evidently a significant amount of genomic information is missing and so technical noise cannot be ruled out.

For k-mer composition, all *Brassica spp*, *A. hypogaea* and their progenitors cluster homeologously rather than subgenomically. This indicates at a sequence composition level, homeologous chromosomes have more sequence in common with each other than their subgenomic counterparts but when k-mer frequency is considered, which is dominated by TEs, the intra-subgenomically similarity overrides the homeologous one. This is perhaps to be expected given that TE content, and especially LTR retrotransposons, can be species-specific, which the *Brassica spp* and *A. hypogaea* genomes would have inherited during hybridisation [46]. While the *Brassica spp* and *A. hypogaea* genomes underwent post-hybridisation repeatome alterations (table D.1, D.2), their relatively recent hybridisation (table D.3) makes it unlikely they would have experienced complete TE-turnover since

their hybridisation event and thus, would have a TE landscape more in common with their progenitors than the other subgenome which originated from a different species [28, 46, 345, 361].

Unlike the *Brassica* and *Arachis* genera, the *Triticum spp* show both homeologous pairings between progenitors and subgenomes while maintaining a subgenome-progenitor clustering structure for k-mer composition (figure 2.10B). For k-mer frequency, subgenome A clusters with *T. urartu*, with the two genomes forming separate clustering whereas the D subgenome clusters homeologously with *A. tauschii* with slightly higher sequence similarity which is reflective of its more recent hybridization into the wheat genome (Figure 2.10 A) [263, 382]. The B subgenome progenitor has not yet been uncovered although *A. speltoides* has been proposed [200]. Throughout the sampled k-mer ranges *A. speltoides* clusters either with the subgenome, or as an outgroup to the A and B subgenome and *T. urartu*. When clustered with the B subgenome, *A. speltoides* exhibits a much lower sequence similarity to the B subgenome, than *T. uratu* does to the A subgenome. This difference in sequence similarity is especially pronounced for k-mer composition (Figure 2.10 B). In their work, through more phylogenetic rigorous methods [200] determined *A. speltoides* is not the B subgenome progenitor, and the results presented herein, seem to agree given that *A. speltoides* genomes exhibits a different sequence similarity-based relationship to the B subgenome than *T. urartu* and *A. speltoides* to subgenome A and D.

To complete an investigation into the agreement of the MinHash k-mer approach with the phylogenetic approaches used by [200] an additional clustering investigation was performed for the same species as used in figure 2.11A, using genomic signatures in place of the chromosomal signatures used so far. For k-mer frequency, the Sourmash approach produces dendrograms with a topology largely in agreement with [200], with a few differences regarding the placement of the D genomes with regard to the *Aegilops* species and a k-mer dependent cluster pattern for *A. speltoides*. For k-mer composition however, the dendrograms

produced are no longer in agreement with the tree topology of [200] and the outgroup species are to be found clustering in the tree structure for larger k-mer values.

As such, while the minhash k-mer sketching approach has shown itself to be capable of producing dendrograms reflective of evolutionary relationships, its implementation as a tool of phylogenetic reconstruction requires far more research. Given that whole-genome phylogenomics for plant genomes has received a recent surge in popularity, and that TE evolution rate has been tied to species rates in mammals, the MinHash k-mer approach in this work, which is tied to genomic TE content, is certainly worth further investigation in the future [29, 229, 291, 371, 385, 388].

Impacts of MinHashing and clustering parameters To complete this work, a comprehensive assessment of the impact of MinHashing parameters (k-mer size and scale factor) and clustering parameters (linkage strategy) was performed. Ultimately it appears scale factor has little impact on the detected subgenome dissimilarity as detected via the cut height for subgenomic separation for both k-mer frequency and composition (figure 2.16). For k-mer frequency, the most pronounced differences are for the smaller k-mer range (k<17), after which they become largely identical. For k-mer composition there differences in subgenome dissimilarity for the different scale factors is even less pronounced, although there are a couple of k-mer dependent anomalies for *T. dicoccoides*.

While this has only been tested in the *Triticum spp*, and so should be tested on a greater range of species to ensure robustness, the fact that scale factor seems to have little impact on the estimates of subgenomic similarity is excellent news for keeping computational overhead to a minimum. Larger scale factors generate smaller genomic signatures which take less time, and require less resources to compute and compare.

There is a clear, linearly positive relationship between k-mer size and subgenome dissimilarity, which for k-mer frequency grows gradually with k-mer size and for k-mer

compositions rises rapidly, peaking at values close to 1. This is shown in figure 2.16 which shows subgenomic cut height (greater cut height is caused by greater dissimilarity) as a function of k-mer size for all of the *Triticum spp.* The reason behind this is that the k-mer space grows with k-mer size at a rate of 4^k. For example, for k=4, the k-mer space is $4^4 = 256$ whereas for k=31 the k-mer space is $4^31 = 4.611686 \times 10^18$. Generally speaking, the chances of encountering the same k-mer in two (or more) distinct sequences will decline as the k-mer space grows [54]. In reality, those chances will be influenced by factors such as sequence similarity and sequence length and it is important to note that no natural genome contains all possible k-mers, but the concept holds true nonetheless [54].

A final investigation to assess the optimal k-mer clustering strategy utilised cophenetic correlation, which is a measure of how faithfully the similarity or dissimilarity is represented within a constructed hierarchical clustering dendrogram [301]. Given that Sourmash was developed for microbial genomes, it is sensible to ensure that the method remains robust for plant genomes. Overall, while the single linkage strategy implemented by Sourmash was rarely the optimal linkage strategy the difference in cophenetic scores between the single linkage and the often-optimal average linkage strategy were minimal, with scores for both rarely dropping below 0.9 (figure 2.17, 2.18). As such, while a minor modification in linkage strategy would be advisable for the most accurate results, the dendrograms presented herein using Sourmash's automatic clustering steps are still very faithful representations of the underlying similarity matrix.

Conclusions

This comprehensive investigation into the implementation of MinHash-based k-mer analysis of polyploid crop genomes has demonstrated that such a strategy, as implemented via the metagenomic software package Sourmash, is capable of revealing evolutionary relationships and genome dynamics that are verified in the literature. Multiple layers of

evidence from experiments conducted herein, combined with a deep dive into published research, support the notion that subgenomic and progenitor clustering results are repeatome driven, possibly by specific TE sequence types, the LTRs. An investigation into MinHash sketching has revealed that scale factor has little impact on the biological results, where k-mer size is a large driver of sequence dissimilarity. Further investigations into hierarchical clustering linkage strategies reveal little difference between the optimal, average-linkage strategy and Sourmashs single-linkage strategy as determined by the cophenetic distance.

Overall the rapid, and highly scalable MinHash sketching method, as implemented by Sourmash, produces robust, biologically accurate results for comparative genomic analysis for even the largest and most complex allopolyploid crops.

Chapter Discussion

In this chapter the critical, yet complex, problem of comparative genomes was re-framed as a metagenomic problem. This seemingly novel take on a problem that has haunted the crop genomics community since the career of "bioinformatician" began and will hopefully provide a fresh avenue of exploration for future crop genomicists, bioinformaticians or even simply computationally-inclined biologists. The beauty of the strategy described within this chapter is that not only is it rapid and scalable, but it is also easily implemented with excellent support from the developers and wider community. Importantly, the developers of Sourmash are aware of the software being hijacked for this purpose, so they will not be blindsided by any crop genomics-related questions the community may pose for them.

Exploring polyploid sequences in this manner has granted insight into the sequences responsible for allopolyploid subgenome sequence differentiation; repeats. This is perhaps not overly surprising given their propensity for rapid sequence evolution and the fact they make up the bulk of many polyploid genomes and yet it is useful information [46]. Their removal shows that, at least for the polyploid genomes in this study, homeologous chromosomes

exhibit higher sequence similarity to each other, than to other subgenomic chromosomes, reflecting the conserved nature of the protein-coding gene space and the non-conserved nature of the repeat-element space. Given that polyploid sequence evolution can take many avenues following a polyploidization event (i.e. stable subgenomic existence versus rapid fractionation of the genome into a diploid state) this shared genome-evolution characteristic is interesting [64, 106, 298, 342].

This work also provides insight into the subgenomic relationships of the polyploid classes. While polyploidy is widely regarded as more of a continuum that discrete classes [41, 226, 305], this work highlights than subgenomes do exhibit class-specific sequence-similarity-based relationships between subgenomic chromosomes. While allopolyploids generally showed excellent subgenomic clustering, autopolyploids showed no such ability and segmental-allopolyploids subgenomic-clustering ability fell somewhere between the two extremes. Arguably, the allopolyploids which failed to show subgenomic clustering for k-mer composition have more in common with the segmental allopolypoids than their other non-segmental allopolyploid counterparts, which reflects the diversity present in allopolyploid evolution and supports the continuum-based understanding of polyploid classification.

Importantly however this work has featured a small, and unbalanced proportion of poplyploid genomes, with far more allopolyploids genomes being available than segmental allopolyploid or autopolyploid genomes in public databases. This reflects the complexity associated with polyploid genome assembly, which is particularly exacerbated by the higher sequence similarity featured within autopolyploid genomes. However, recent genome assembly efforts for a number of potato cultivars have demonstrated that excellent, phased genome assemblies are possible for autopolyploid genomes [145]. Hopefully this represents the future for polyploid genomes assemblies, which will facilitate better subgemome and genome assemblies, furthering our understanding of polyploid genome biology and evolution [381].

Presentation of work

This work has been presented to a number of conferences (cited below) where it was received with excitement from the plant genomics community. Two of the presentations are hosted on public platforms and the links for the citations below can be followed to view them.

1. **Reynolds,G**., Strnadova-Neeley,V., Mumey,B., Lachowiec,J.. Hijacking a metagenomic strategy for rapid comparative subgenomics for polyploids and their progenitors. *Botany*. 2022

2. **Reynolds,G**., Strnadova-Neeley,V., Mumey,B., Lachowiec,J.. Leveraging a metagenomic strategy to reveal patterns of accessory and subgenomic chromosomal evolution *Plant Genomes Online Conference*. 2022 **PGO link**.

3. **Reynolds,G**., Strnadova-Neeley,V., Mumey,B., Lachowiec,J.. Computational strategies reveal subgenome-specific signatures of polyploid evolution *Barker Lab Polyploidy Webinar*. 2022 **Polyploidy webinar link**.

4. **Reynolds,G**., Strnadova-Neeley,V., Lachowiec,J. MinHash k-mer sketching highlights allopolyploid subgenome sequence differentiation. *ISCB-Africa ASBCB*. 2021 .

"B" chromosome side quest

In addition to the subgenomics and comparative genomics work, additional work was also carried out on the application of Sourmash to the exploration of the "B" chromosome of crop genomes. Currently, only 2 crop genomes have had their "B" chromosomes sequenced as a compliment to their core "A" genomes which contain their usual compliment of chromosomes; *Zea mays and Secale cereale*. The "B" chromosomes are supernumerary chromosomes which are thought to exist within the genomic chromosomal compliment of 15% of all Eukaryotic species, and are especially common in grasses and cereals [6, 165].

In the past they were thought to have little to no use although when overabundant, they are implicated in reduced fitness for the plants [6, 44, 165]. Recently however, they have been observed to have actively transcribed genes, directly influence "A" chromosome gene expression, and have been associated with positive phenotypic traits under stress, including heat tolerance [44, 258, 265]. Further these sequences are capable of being transplanted to species different from their host genome where they continue to influence core gene expression [44]. In addition there is interest in their utilisation as mini chromosomes for genome engineering [44, 165]

Their genomic architecture is largely repetitive in nature although there are a number of B-essential genes (required for their propagation), protein-coding genes and pseudogenes which seem to have been commandeered from the core "A" chromosome [7, 79, 215]. Interestingly, the "B" chromosomes, while exhibiting a similar repeat content of repeats as A chromosomes, differs in repeat sequence composition [142]. As such, this presented an excellent opportunity for application of the MinHash strategy used in the main section of this work to explore the difference in sequence composition and frequency between the core "A" chromosomes, and the "B" chromosomes.

As such, the core "A" compliment of chromosomes and the corresponding "B" chromosomes were put through the same Sourmash pipeline as described in the main sections of this work and the clustering of the B chromosome and the core A chromosomes visualised. For *Z. mays*, the sequenced "B" chromosome is deposited without any corresponding "A" chromosomes. However, the same cultivar from which the "B" chromosome was extracted (Zm-B73) is available and so these sequences were used as the core "A" chromosome set. While *Z. mays'* "B" chromosome is not much smaller than its smallest "A" chromosomes, the *S. cereale* "B" chromosome is significantly smaller. As such, the *S. cereale* chromosomes were shredded to to be of equal length (250Mbp) using bbshred.sh [53]. Given that the *S. cereale* also had repeat-masked sequences for the "A" and "B" chromosomes, they were

also put through the same Sourmash pipeline. In the interest of quantifying the clustering patterns, the number of cluster results where the "B" chromosome was clustered to the "A" chromosome was enumerated.

For k-mer frequency, the two species exhibit similar clustering results. The *S. cereale* B chromosomes are capable of being "cut" from the rest of the clusters for 90%, 93%, 93% and 96% of the sampled k-mers for scale factors 1000,500,250 and 125 respectively. For *Z. mays* this is 73%, 73%, 93% and 96%. Evidently unlike the subgenome results, the scale factor parameter has a measurable different on the ability to separate the "B" chromosomes from the cote "A" chromosomes.

An example of dendrograms where the "B" chromosomes are considerable "cuttable" can be seen for *S. cereale* figure 2.19A and for *Z. mays* figure 2.19B.

For k-mer composition however, the two species show very different results (figure 2.20). *S. cereale* only shows an ability to split the "B" chromosome for k=13 across all scale factors. For all other k-mer sizes, the B chromosome segments are clustered in with the core "A" compliment of chromosomes. In contrast, *Z. mays* maintains a separate clustering structure for the B chromosome, although due to the lack of similarity between the chromosomes, and thus the shared long branch lengths, it can be difficult to ascertain if the "B" chromosome can indeed be cut from the core "A" chromosomes.

The repeat-masked sequences of *S. cereale* showed a distinct lack of ability to "cut" the B chromosomes from the dendrogram with a single cut (figure 2.21). Instead they were often clustered in with the core "A" chromosomes, often together. This was true across scale factors and for k-mer frequency and composition. This indicates that it is the repeatome driving the clustering result whereby the "B" chromosome can be cut from the core "A" compliment, much like the subgenomic and genomic clustering results in the main chapter. However, given the different in clustering patterns for the two species, it is perhaps not reasonable to infer the same sequences are driving the clustering results for *Z. mays*.

For k-mer frequency, both species also show high similarity to a select chromosome or pairs of chromosomes from the "A" compliment for given k-mer sizes. For *S. cereale* this is the 5R_0 fragment (the first 250000000bp of the shredded chromosome 5), although some similarity is also seen for 5R_1 and 5R_2 for 15-mers for scale factors 250 and 15 and 17-mers for scale factor 125. For *Z. mays* the B chromosome exhibited the most similarity to chromosomes 6 and 7 throughout the scale ranges. These relationships may be suggestive of a recent transfer of genomic material from those chromosomes to the "B" chromosome.

Overall, the clustering results are in agreement with current understandings of the mysterious "B" chromosomes, in that while they are the product of genomic material transfer from the "A" chromosomes, they also consist of B-specific sequences which enable their outgrouping in the dendrograms from the core "A" chromosomes. The difference in clustering for the two species and their "B" chromosomes may be explained by the fact that "B" chromosomes are known to undergo species-specific evolution [146].

However there are a number of short-comings for this small study. The first is that it is based on only 2 "B" chromosomes, and so the clustering patterns and relationships between chromosomes seen in these results may not be transferable to other "B" chromosomes and their "A" chromosome compliments. Hopefully in the future more "B" chromosomes will be sequenced. In addition, this study would benefit from a more statistically rigorous approaches to determine the separability of the "B" chromosomes from the "A" genome, such as the Gap statistic or silhouette coefficient [333, 387].

Further, given that the species also exhibit a scale-factor dependent relationship between k-mer size and separability of the "B" chromosomes, perhaps a MinHashing approach is not the optimal strategy for sequence comparison for this particular application. Instead, an investigation into the intersection and difference of total k-mer content may prove the most accurate approach, which may be performed using efficient k-mer counting software such as KMC3 [179].

Conclusion

Overall this chapter has consisted of exploring the limitations of MinHash-based signature-based comparative genomics. For genome and subgenome comparison it performs excellently, and as such offers computationally efficient and scalable solution for the complex game of "genomic-spot-the-difference". However, its application to comparative intra-genomics for species with "B" chromosomes is limited given the scale-dependent relationship between k-mer size and separability of the "B" chromosomes. Importantly, this is inferred from a sample size of 2, only 1 of which had repeat-masked sequences available and as such a more thorough investigation should be performed, should more "B" chromosomes become available in the future.

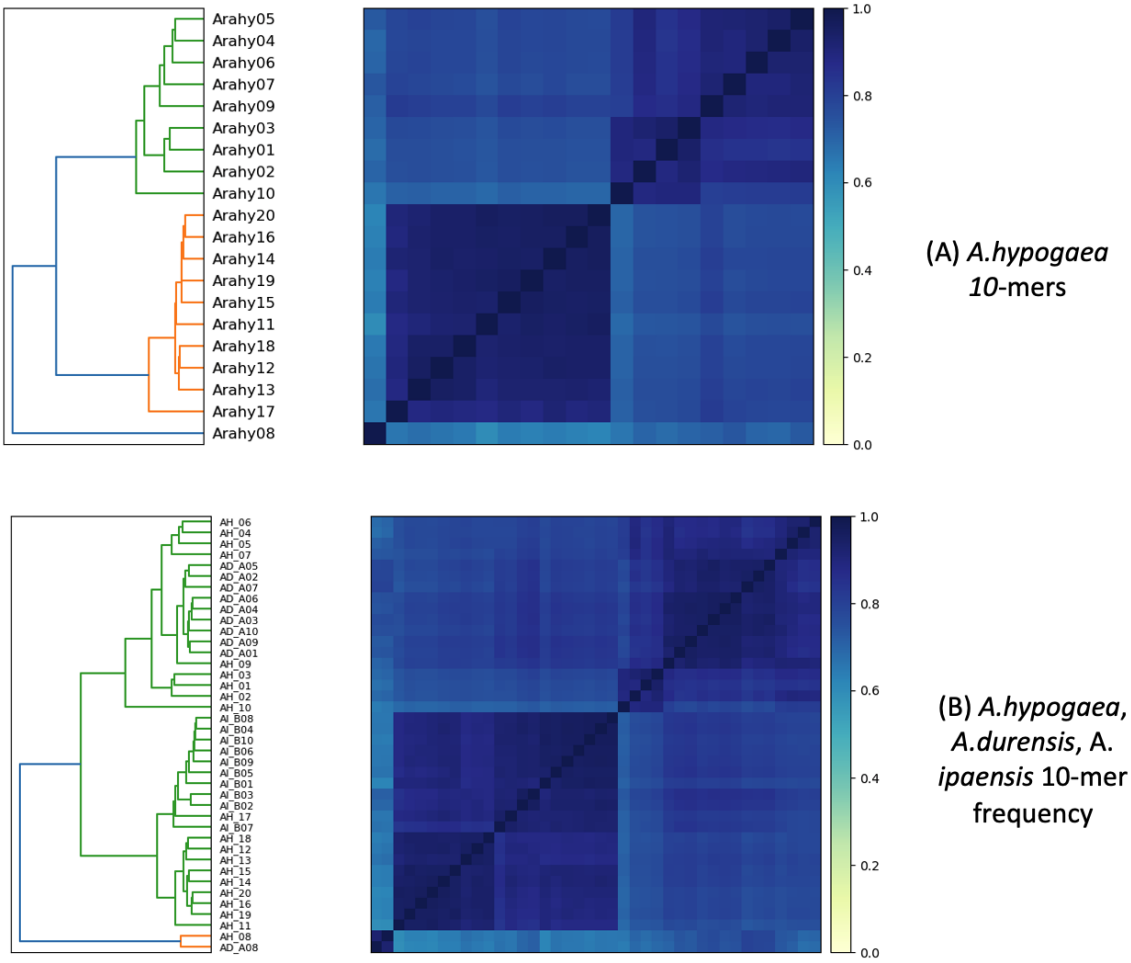


Figure 2.6: Sourmash produced dendrograms and heatmaps for *A.hypogea* (A), and *A.hypogea* and progenitors *A.durensis* and *A.ipaensis* (B) 10-mer frequency

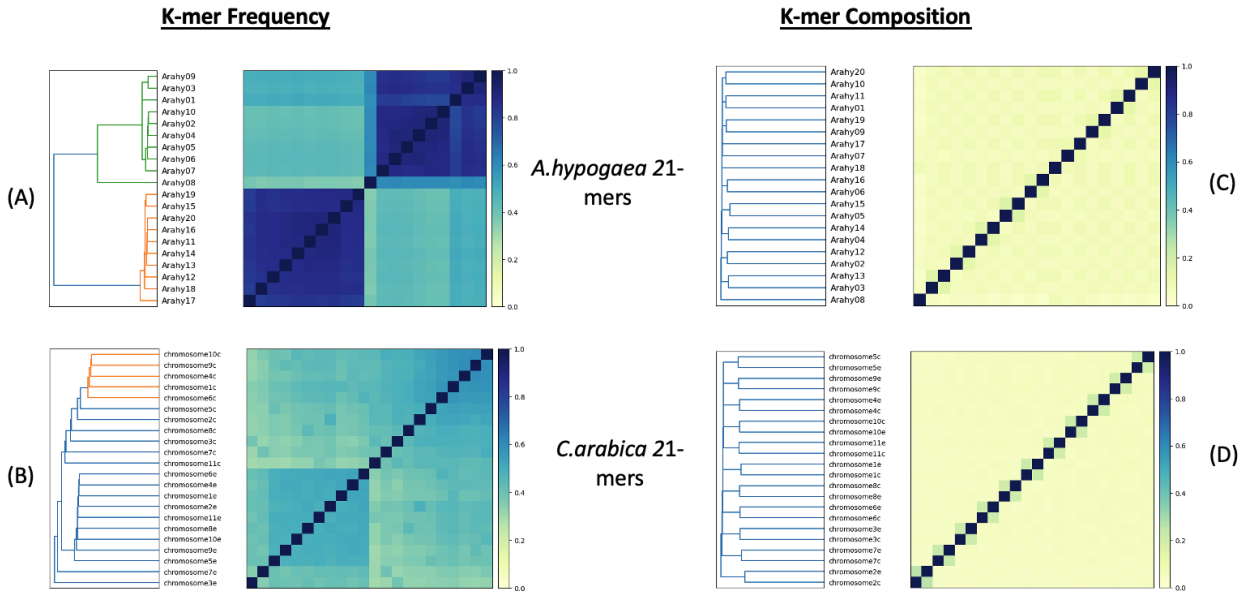


Figure 2.7: Sourmash produced dendrograms and heatmaps for *A. ahypogea* and *C. arabica* for 21-mer frequency (A,B) and composition (C,D) respectively

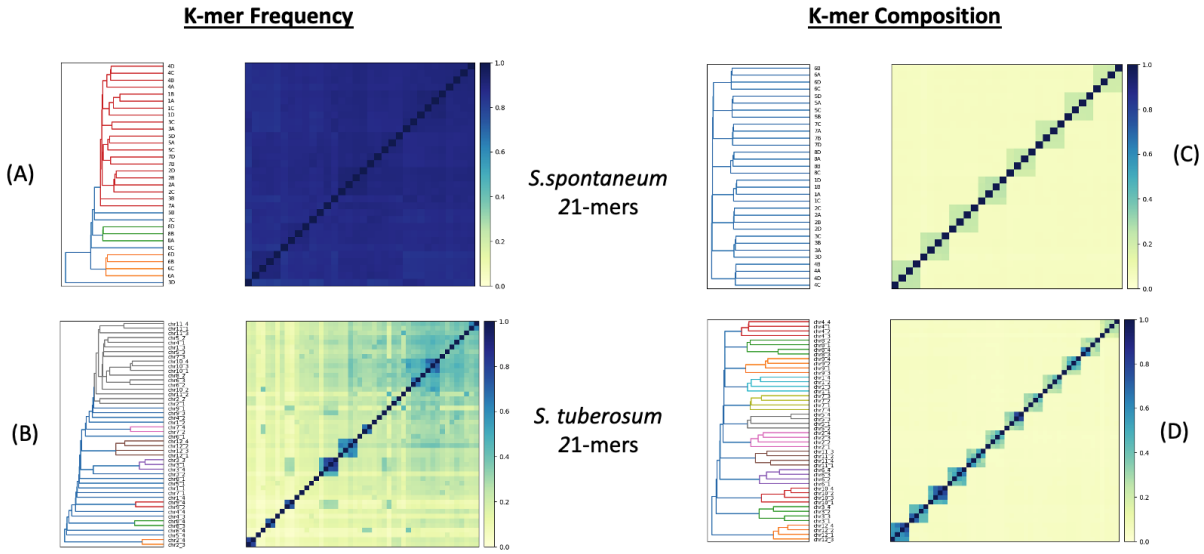


Figure 2.8: Sourmash produced dendrograms and heatmaps for *S. spontaneum* and *S. tuberosum* for 21-mer frequency (A,B) and composition (C,D)

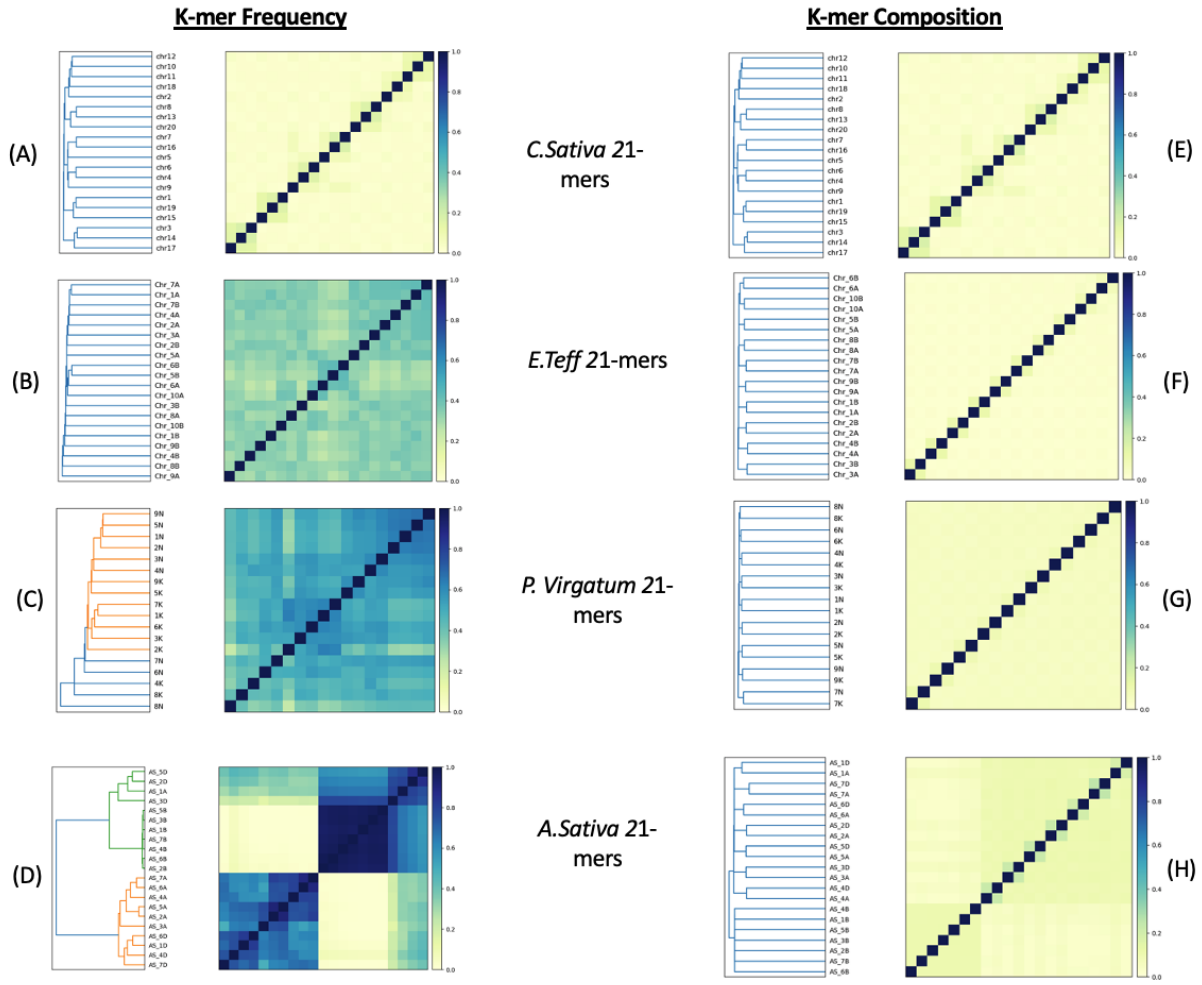


Figure 2.9: Sourmash produced dendrograms and heatmaps for *C. sativa*, *E. teff*, *P. virgatum* and *A. sativa* for 21-mer frequency (A-D) and composition (E-H)

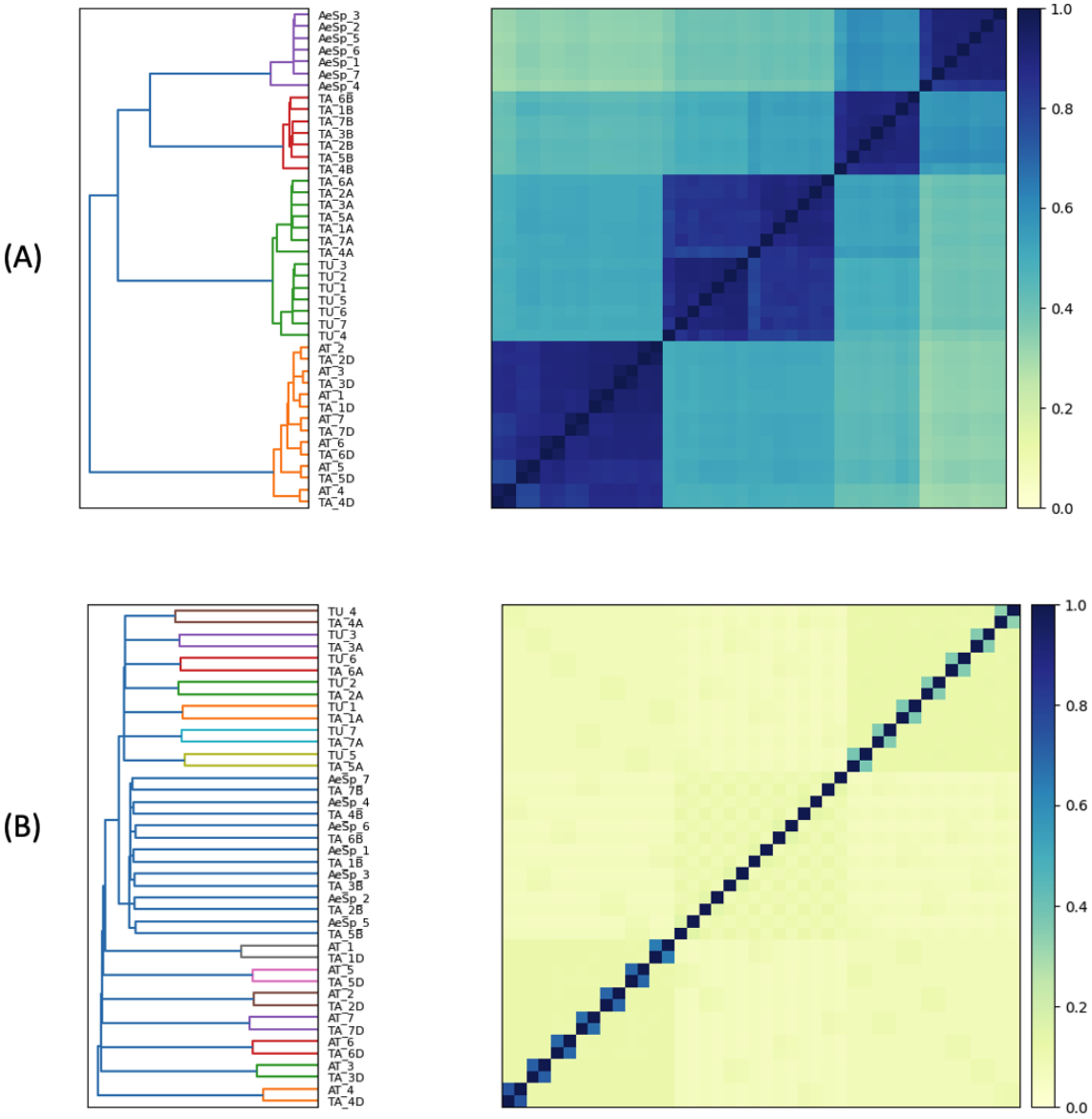
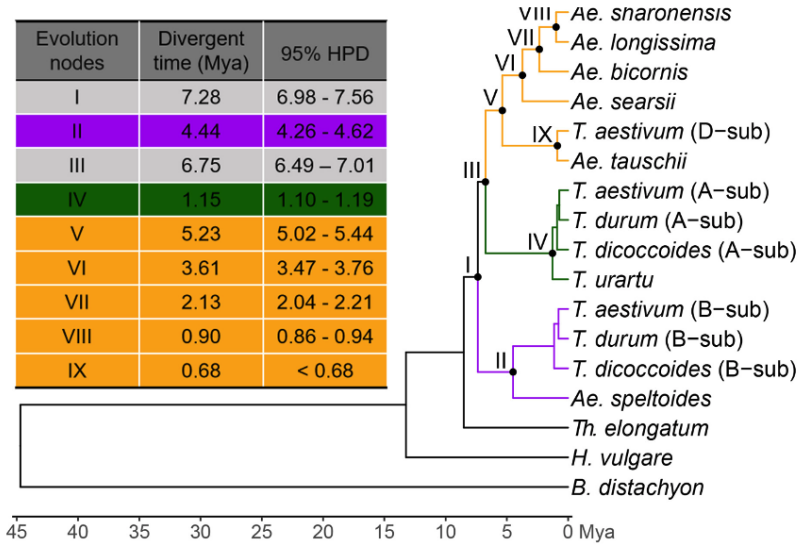


Figure 2.10: Sourmash produced dendrograms and heatmaps for *T. aestivum*, *T. urartu*, *A. tauschii* and *A. speltoides* for 21-mer frequency (A) and composition (B)

(A)

Evolution nodes	Divergent time (Mya)	95% HPD
I	7.28	6.98 - 7.56
II	4.44	4.26 - 4.62
III	6.75	6.49 – 7.01
IV	1.15	1.10 - 1.19
V	5.23	5.02 - 5.44
VI	3.61	3.47 - 3.76
VII	2.13	2.04 - 2.21
VIII	0.90	0.86 - 0.94
IX	0.68	< 0.68



(B)

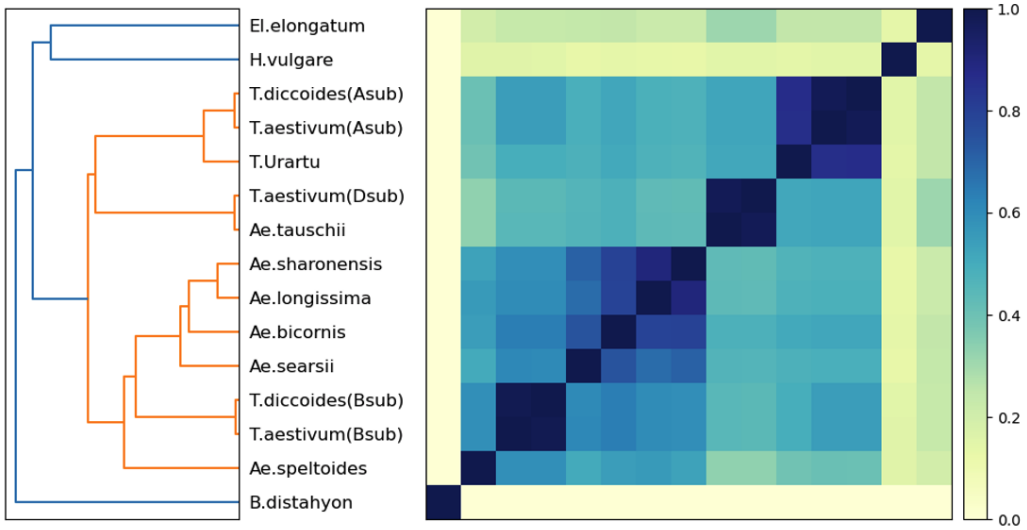


Figure 2.11: Phylogenetic tree produced by [200] (A) and a Sourmash produced dendrogram and heatmap for the same species for 21-mer frequency.

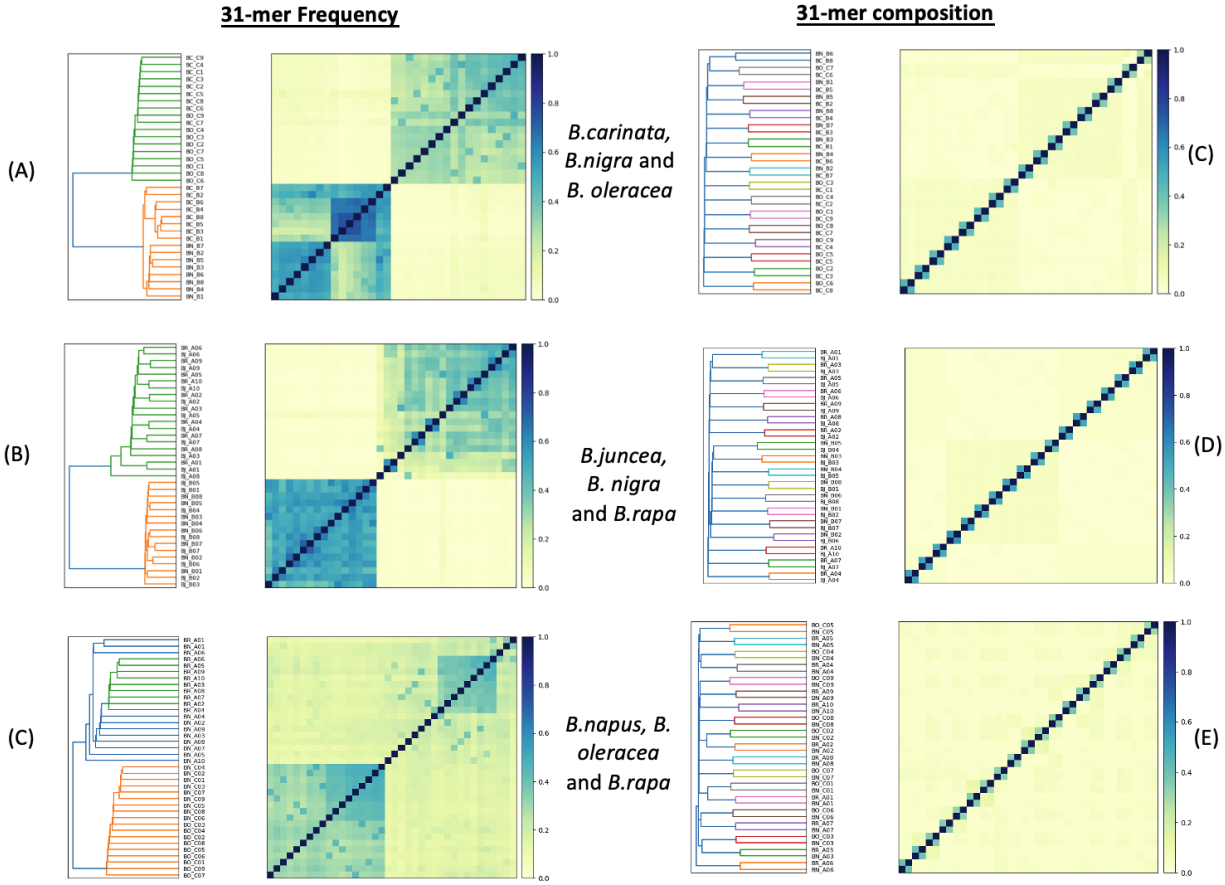


Figure 2.12: Sourmash produced dendrograms and heatmaps for *B. carinata*, *B.juncea*, *B. napus* and their progenitors *B. nigra*, *B. oleracea* and *B. rapa* for 31-mer frequency (A-C) and composition (C-E).

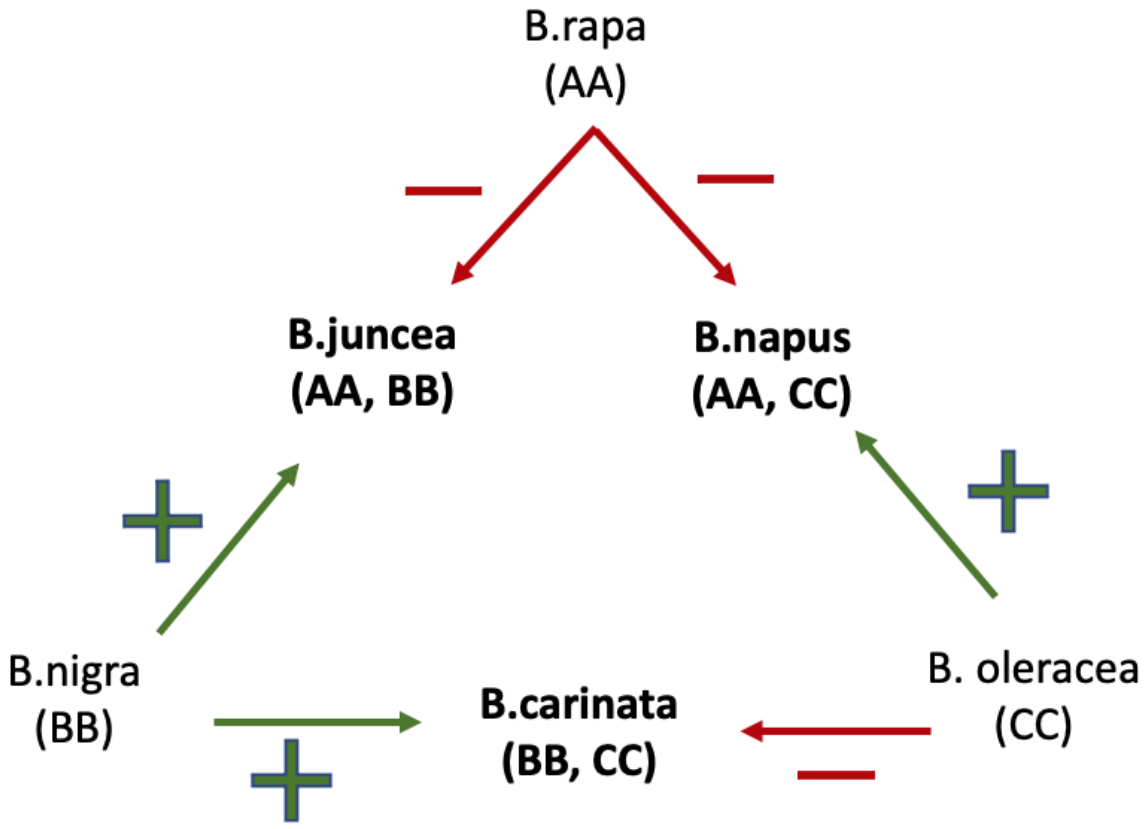


Figure 2.13: The classic triangle of "U" relationship between *B. juncea*,*B. nigra* and *B. carinata* and their progenitors *B. nigra*, *B. oleracea* and *B. rapa* augmented with information about which subgenome exhibits greater (green arrow and plus sign) and lesser (red arrow and minus sign) intra-subgenomic similarity.

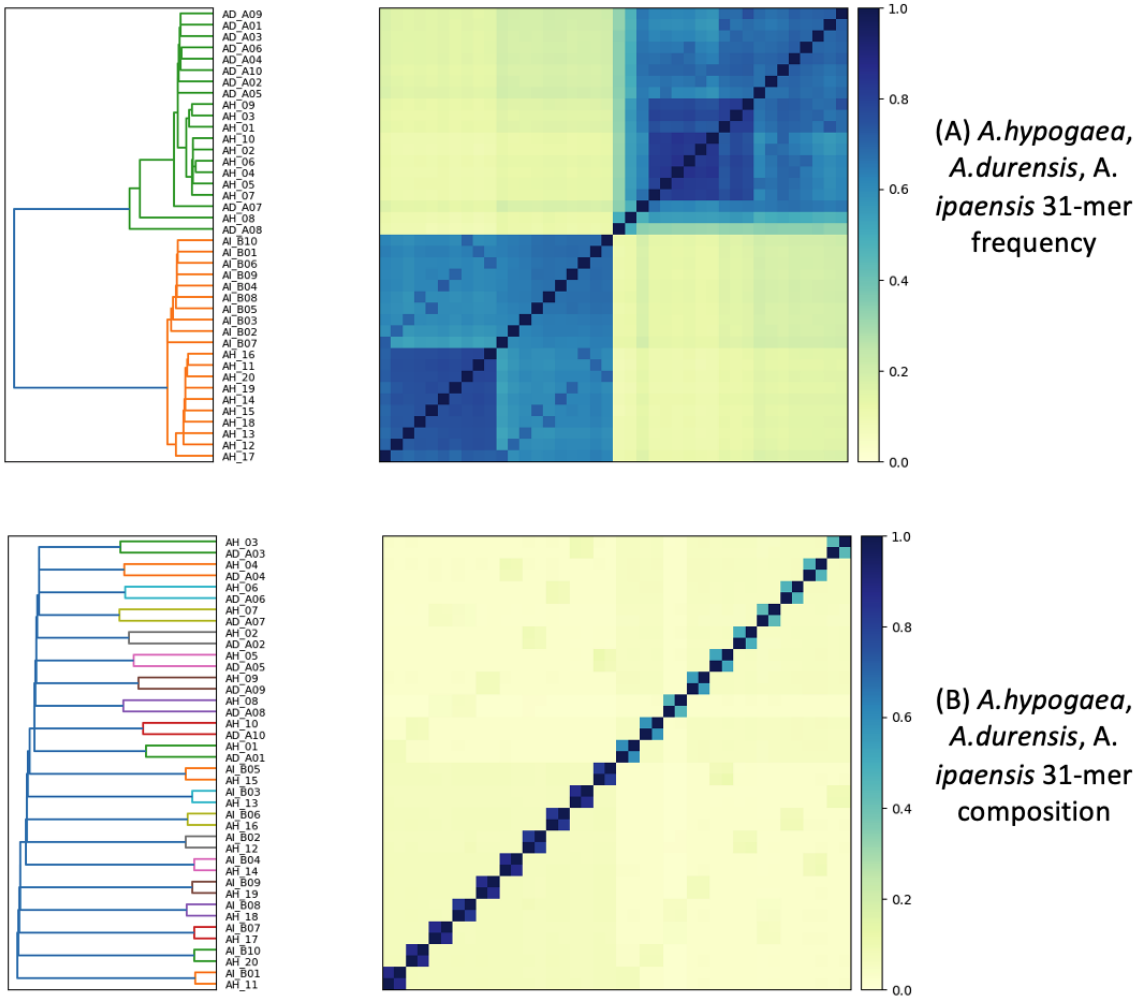


Figure 2.14: Sourmash produced dendrograms and heatmaps for *A. hypogaea* and its progenitors *A. durensis* and *A. ipaensis* 31-mer frequency (A) and composition (B).

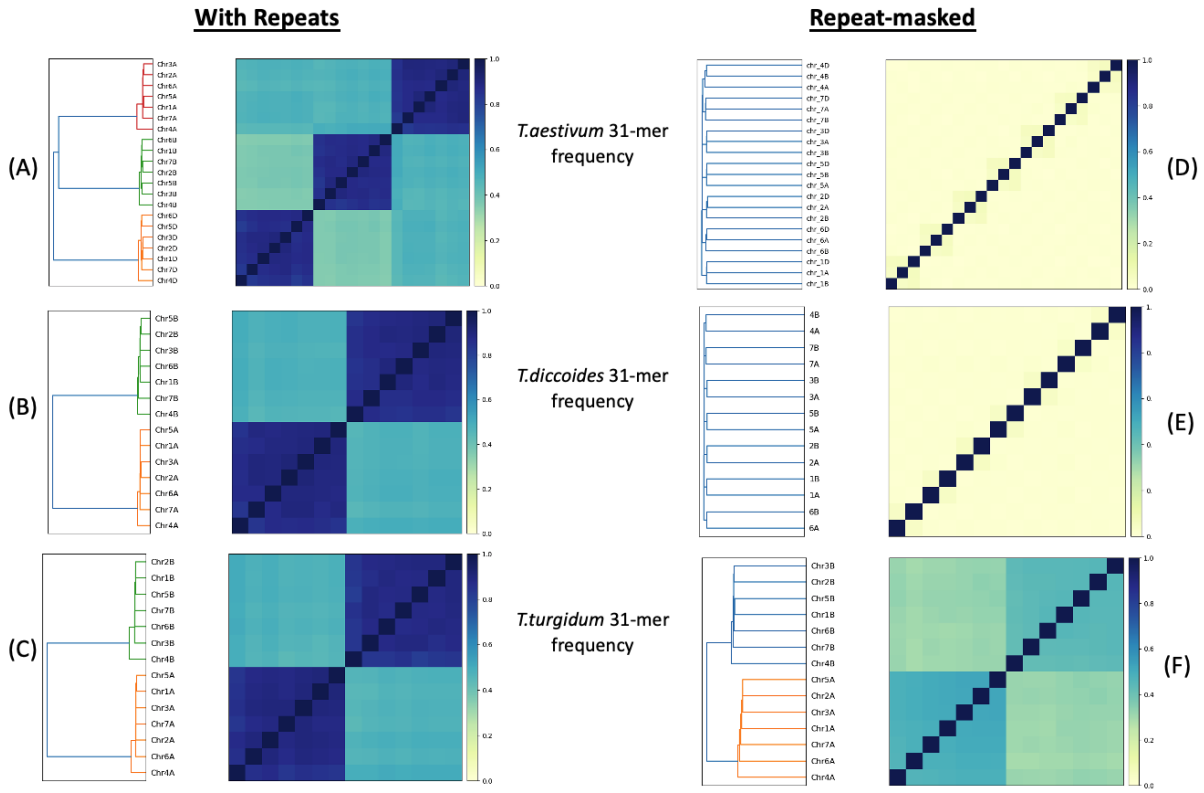


Figure 2.15: Sourmash produced dendrograms and heatmaps for *T. aestivum*, *T.dicoccoides*, *T. turgidum* for repeat-rich (A-C) and repeat-masked (D-F) sequences

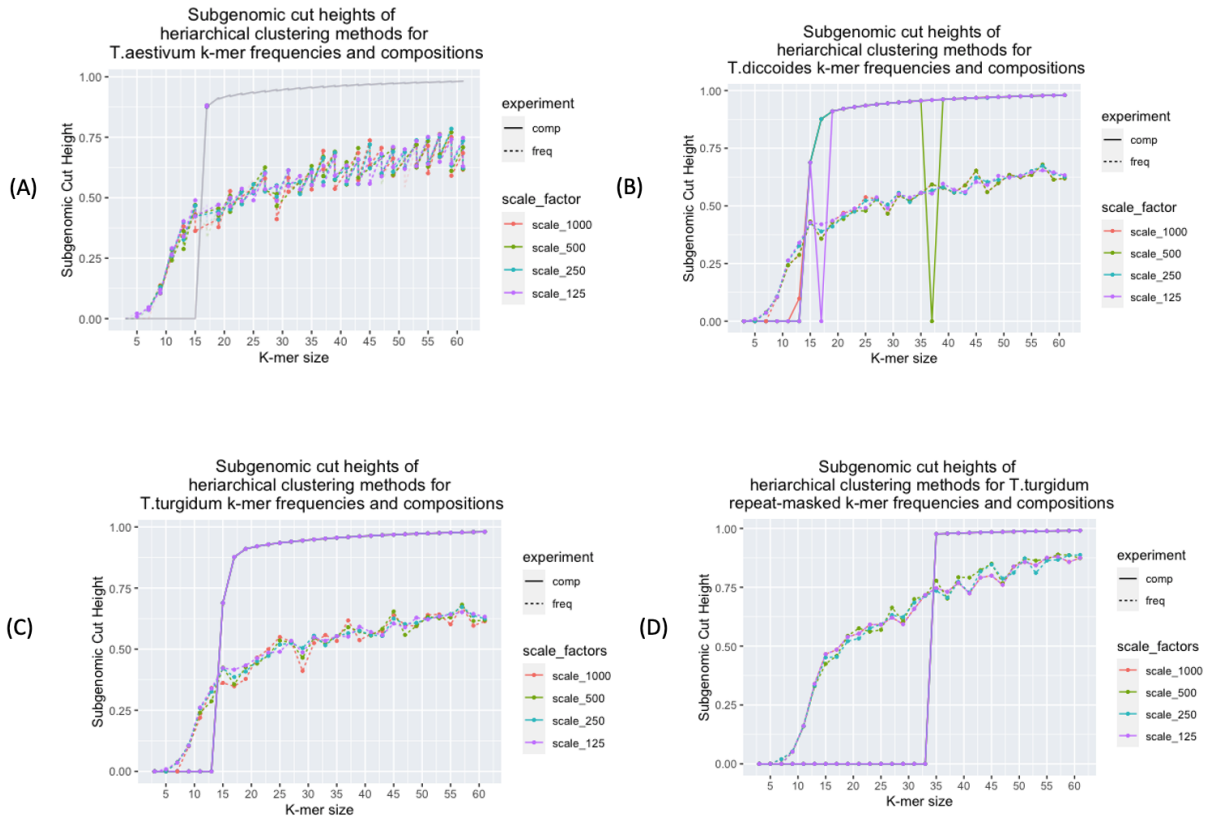


Figure 2.16: Subgenomic dendrogram cut heights for *T. aestivum* (A), *T. dicoccoides* (B) and *T. turgidum* (C) for k-mer frequency (dashed line) and composition (solid line) across scale factors 1000(red line), 500 (green line), 250 (blue line) and 125 (purple line) for repeat-rich sequences. D shows the same information for *T. turgidums* repeat-masked sequences.

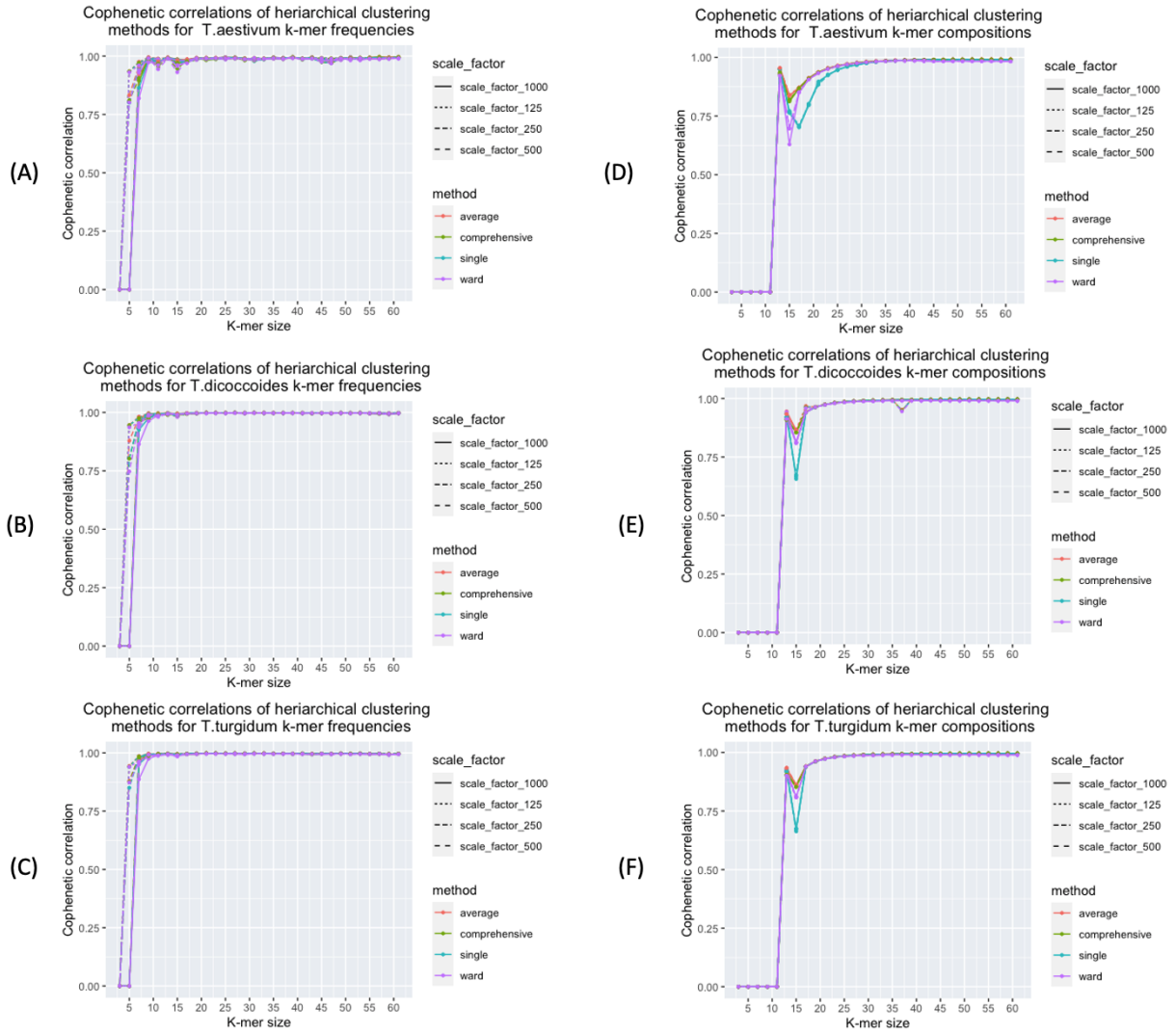


Figure 2.17: Cophenetic correlations for *T. aestivum*, *T. dicoccoides*, *T. turgidums* repeat-rich sequences for k-mer frequency (A-C) and composition (D-E).

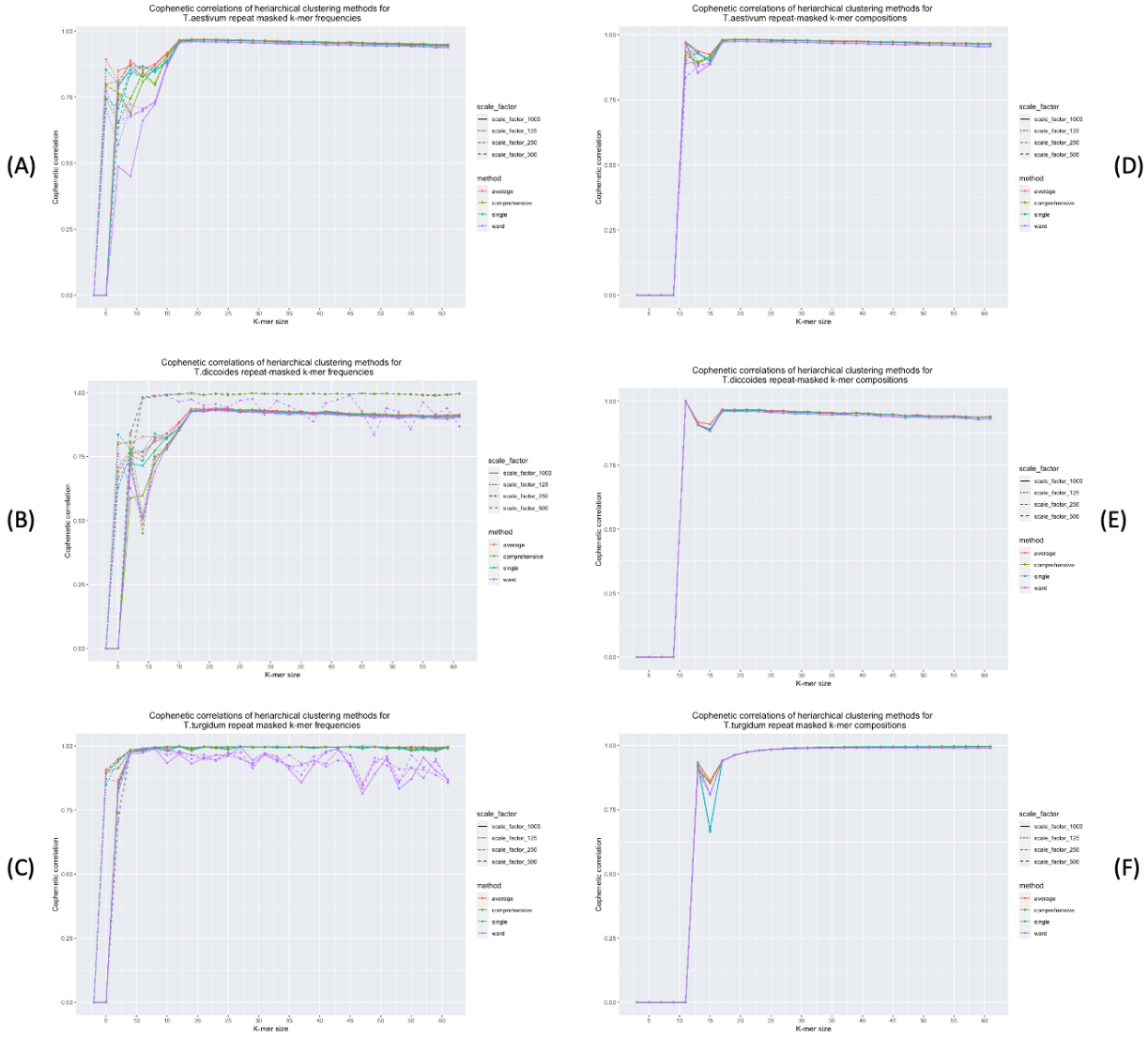


Figure 2.18: Cophenetic correlations for *T. aestivum*, *T. dicoccoides*, *T. turgidums* repeat-masked sequences for k-mer frequency (A-C) and composition (D-E).

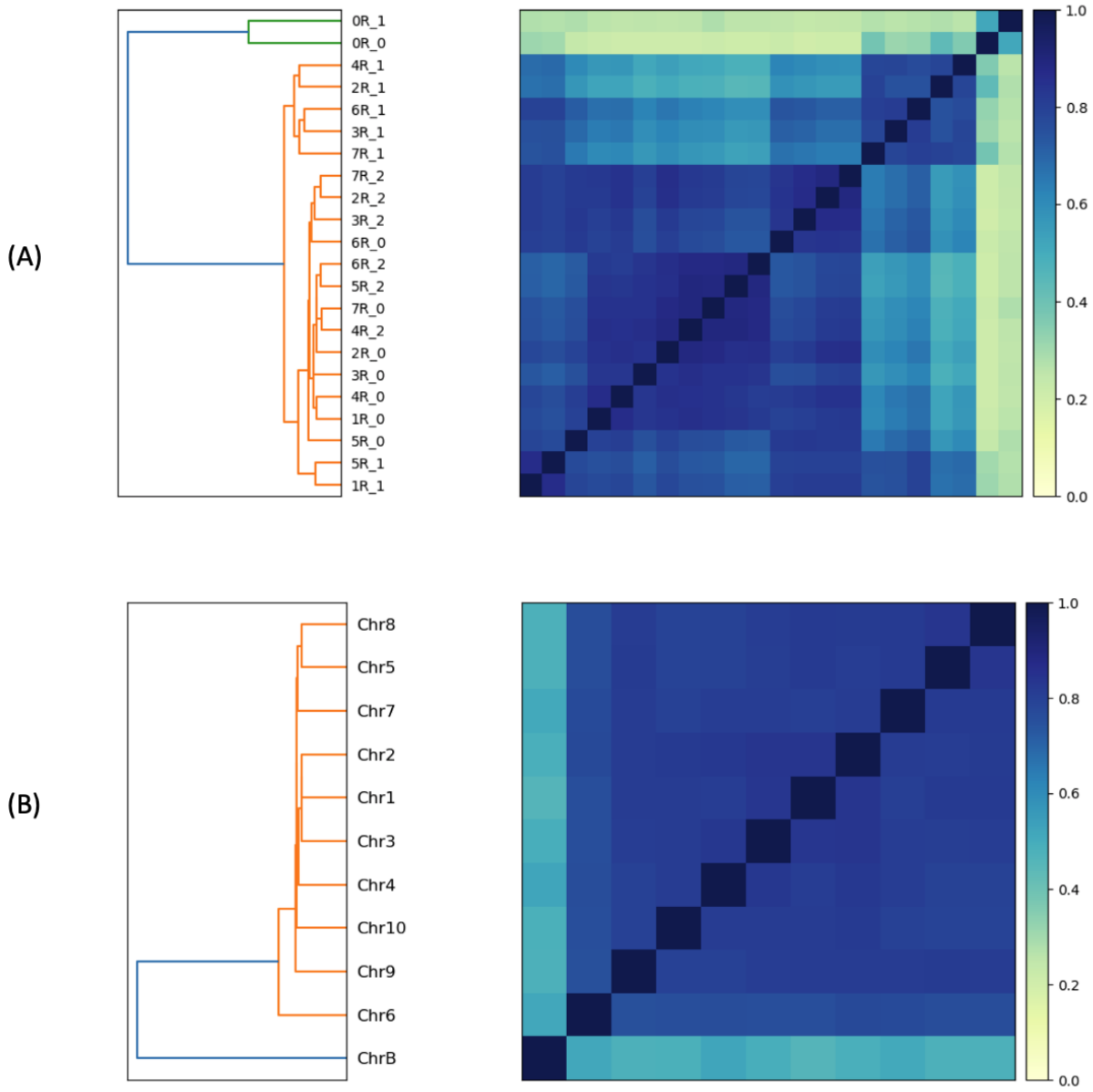


Figure 2.19: Sourmash produced dendrograms and heatmaps for *S.cereale* and *Z.mays* 31-mer frequency

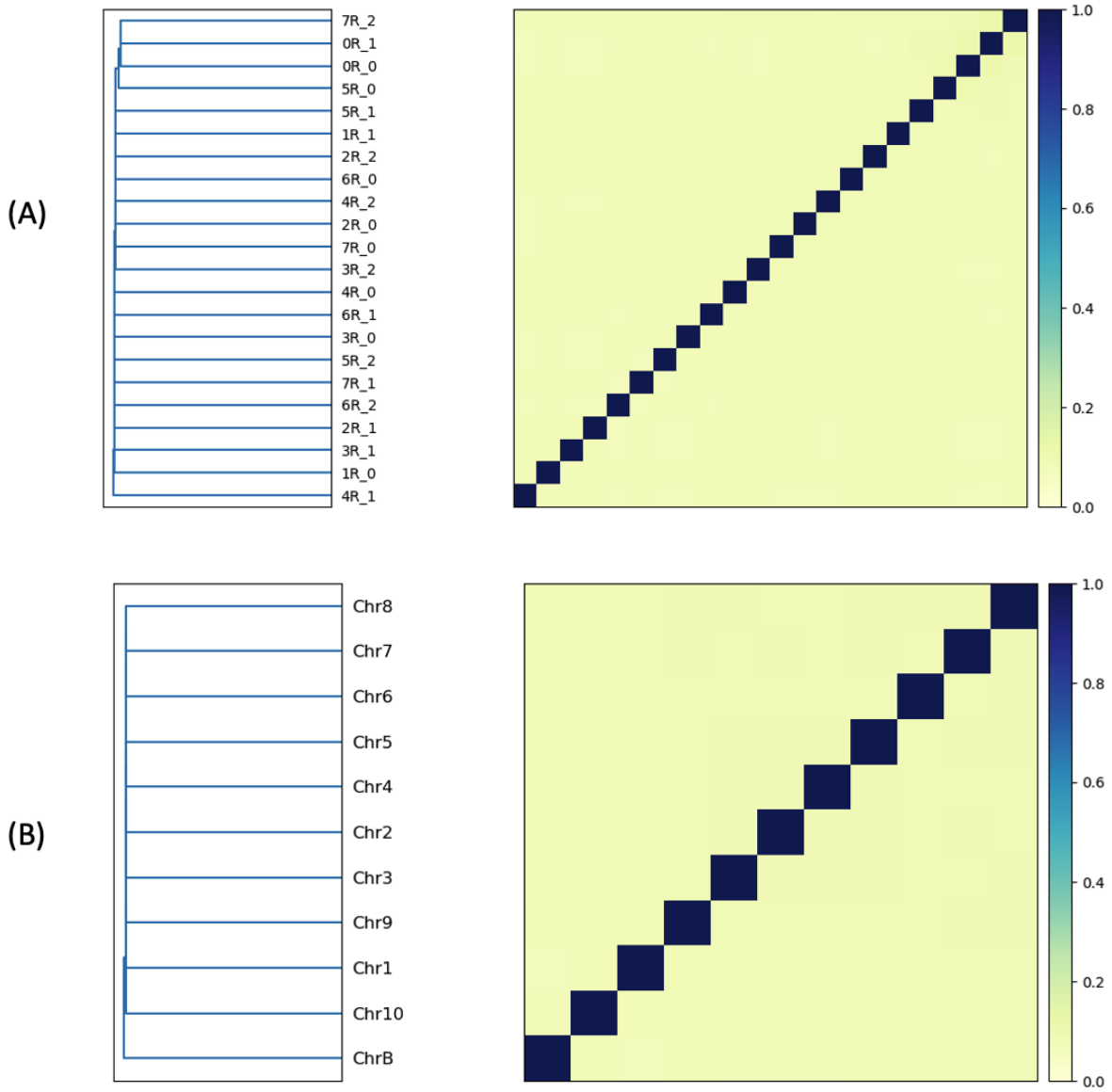


Figure 2.20: Sourmash produced dendrograms and heatmaps for *S.cereale* and *Z.mays* 31-mer composition

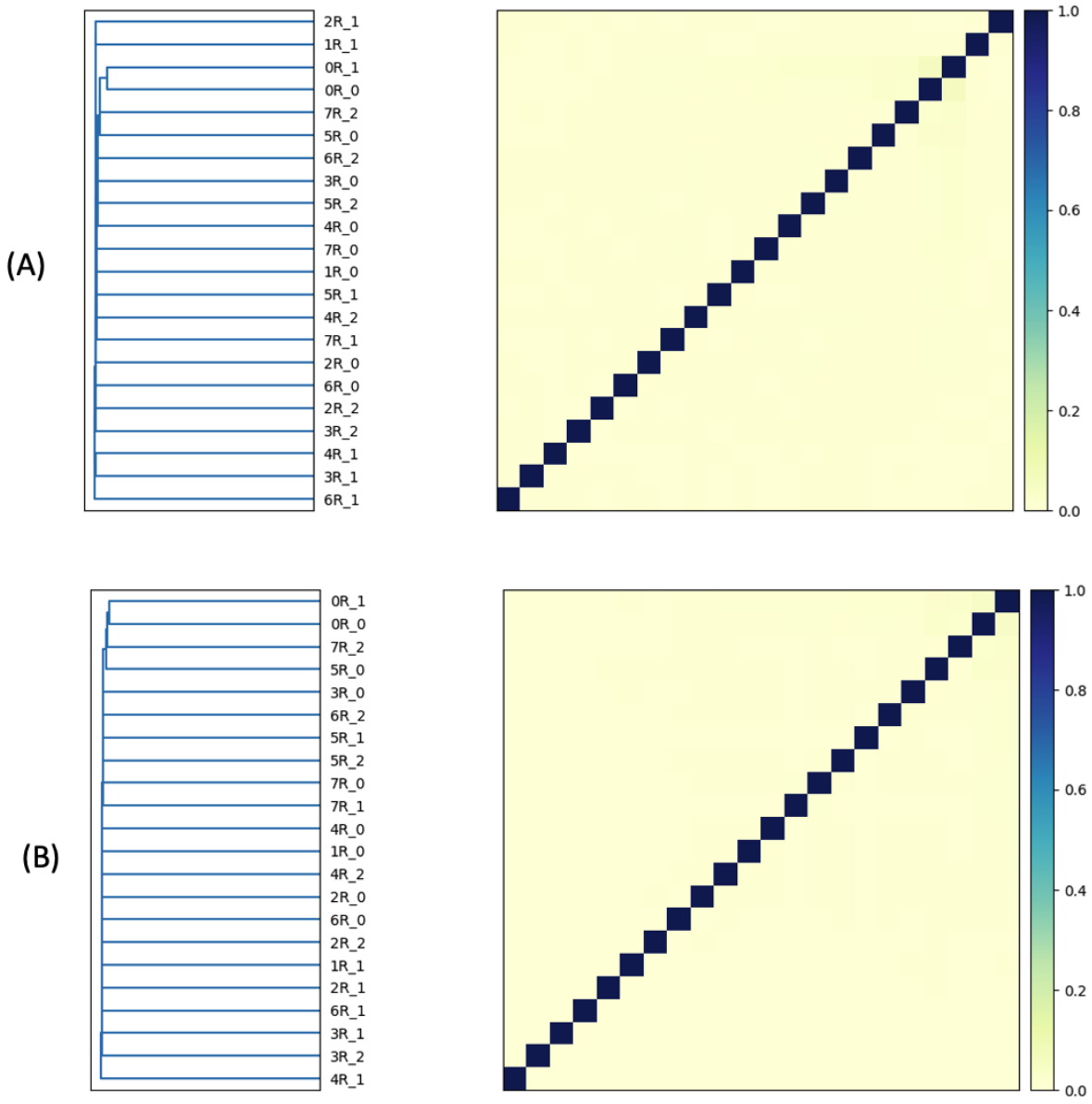


Figure 2.21: Sourmash produced dendrograms and heatmaps for *S.cereale* 31-mer frequency (A) and composition (B)

UTILISING K-MER SPECTRA FOR ASSEMBLED FRAGMENT RECRUITMENT

Polyploid genomes present numerous problems for genome assembly efforts. The first is that their generally highly repetitive nature, induced either through repeat-element expansion (not uncommon in polyploidy) or through high similar subgenome sequence content creates an array of unnavigatable topological structures in assembly graphs [48, 180, 251, 260, 343]. Coupled with the usual sequencing-platform errors and biases the resulting genome assemblies are often highly fragmented and feature a rich repertoire of misassemblies and chimeric contigs [120, 264]. While misassemblies are an unfortunate constant for any genome assembly effort, chimeric contig formation is more a problem for the fields of polyploidy and metagenomic assembly. Chimeric contigs are a mosaic of sequences from disparate subgenomes (such as homeologous sequences which have been incorrectly stitched together during the assembly process due to high sequence similarity) [22, 242, 252].

Correcting for these problems is now routine for reference-quality genome assembly efforts and is usually facilitated by the use of additional sequencing data from long-read data, genetic maps and optical mapping [11]. This complimentary data, also usually accompanied by 3D chromosome conformation data (Hi-C) can then also be used to anchor, order and orientate the sequences into the chromosomes from which they originated [11, 52].

However, given the challenges with genome assembly, it is not usual for a portion of the genome to remain unanchored, ordered or orientated. The extent of assembly completeness for the genomes used in Chapter One can be seen in table D4. The portion of assembled data which cannot be anchored are then deposited in the "unassigned" or "Un" chromosome and are rarely revisited. However, these sequences may still contain valuable genetic information and as such are worthy of re-exploration. To demonstrate this, a BLAST analysis [15] of the "Un" chromosomes from IWGSC v1.0 used by [382] was performed with the interesting matches listed below:

- Loci associated with disease resistance:
  - PM3 locus (powdery mildew resistance)
  - LR1(leaf rust)
  - PiK2 (viral)
  - RGA3
  - Sr35 region
  - Lr34 locus
- Loci associated with abiotic stress resistance:
  - ERF1 – ethylene response factor
  - Frost resistance H2-gene locus
- Grain characteristic loci:
  - Starch synthase 1 gene
  - Seed storage genes
  - Various gladin genes – gluten proteins
  - Gluten controlling genes

As such in this chapter we address the need for placement of "Un" contigs using k-mer profiles. In the previous chapter it was demonstrated that many allopolyploids exhibit global subgenome-specific k-mer signatures which in this chapter are leveraged for "Un" chromosome recruitment to their subgenome of origin. No work has previously focused on "Un" chromosome recruitment before but work has previously been carried to leverage subgenome-specific k-mer profiles for subgenomic sequence determination. Both PolyCracker [123] and SubPhaser [159] utilise subgenome-specific k-mers to recruit or phase assembled

subgenomic sequence data. This work differs to these approaches however by implementing a "wide net" approach and utilising full global subgenomic signatures in place of smaller, more specific subgenome-specific k-mer profiles.

In the following section it is demonstrated that such signatures are excellent features for the subgenomic recruitment of assembled contigs that were previously unable to be anchored within the genome. Much future work remains to be done to assess if this strategy could be extended to read data to enabled subgenomic partitioning of reads prior to genome assembly. This could massively improve the accuracy and contiguity of polyploid genome assemblies given that polyploid genomes are known to induce unnavigable topological structures within de Bruijn graphs which can cause misassembled and fragmented genome assembly results [119, 212, 251, 260, 297].

However, the utility as an "Un" chromosome recruiter is valuable nonetheless. Maximising genome assembly completeness and correctness are two of the three major tenants of genome assembly and this work contributes to those aims [332]. The third aim, which concerns maximising contiguity, is not directly met by this work, but a subgenomic assignment certainly does reduce the search space for its specific location from a genome-wide to a subgenome-wide search space. In the future, this may act as layer of evidence for definitive placement and chromosome-scale incorporation of the contigs.

Maximizing Subgenomic Assemblies; K-mer composition guided subgenome classification of "Un" contig

Abstract

Polyploid genome assembly efforts are often hampered by both genome size and complexity of the polyploid genome. Even with additional, complementary sequencing data, a significant portion of assembled data often remains unanchored to chromosomes or subgenomes and are deposited within the "unassigned " or "Un" chromosome. Given that the

position within the genome is unknown, its utility for downstream analysis is reduced. Once deposited in this "Un" chromosome the sequences are rarely revisited. However, a recent re-exploration of the assembled fragments, facilitated by optical mapping was performed by Zhu *et al*, for *T. aestivum* and *T. diccoides* [382, 393]. Aided by this additional data, hundreds of previously unanchored contigs were placed within their chromosome of origin. Given evidence from previous research that many allopolyploids exhibit subgenome-specific k-mer profiles this work explores if they can be exploited for "Un" contig recruitment in *T. aestivum*. It is demonstrated that subgenomic k-mer profiles are excellent features for "Un" contig recruitment and that there exists a strong relationship between k-mer size and classification success. Importantly classification is not strongly affected by the false positives associated with the data structures implemented to hold the k-mer profiles in working memory and that strategies to mitigate false-positive k-mer hits have little impact. As such the strategy presented herein is a robust, accurate, and memory efficient method for subgenomic recruitment of "Un" contigs for allopolyploids which exhibit subgenome-specific k-mer profiles.

Introduction

Assembling polyploid genomes is a resource- and time-intensive task [186, 304, 326]. Not only are many polyploid genomes large, but they are also architecturally complex for assembly algorithms to work with [186, 326]. This complexity comes largely from the repetitive nature of such genomes, both in terms of the extensive proportion of repeat-elements within many genomes and due to the presence of multiple homeologous chromosomes of varying similarities [94]. The presence of these shared or repetitive sequences, (especially if they are inexact, i.e. featuring variations), alongside sequencing errors, can cause highly problematic topological structures in the assembly graph (e.g. superbubbles, ultrabubbles, complex bulges, roundabouts) [48, 180, 251, 260]. Assembly

algorithms often cannot deterministically traverse such structures due to the presence of multiple, valid paths which often have to be simplified, or heuristically traversed [180, 327]. Combine these problematic structures with various sequence-platform errors and biases, and the resulting assembly is often highly fragmented while containing a number of misassembled and chimeric contigs and scaffolds [120, 264].

To correct for these problems, and to stitch these assembled fragments into their constituent chromosomes of origin, polyploid genome-assembly efforts incorporate a range information. This includes further sequencing data such as long-reads, optical mapping and Hi-C, genetic maps containing marker sequences and linkage data as well as reference-guided and reference-free placements of assembled data [11, 123]. However, even when multiple layers of additional information are used, a non-trivial portion of the genome often remains unanchored and is thus relegated to the "Un" chromosome. This chromosome, as a mosaic of contigs or scaffolds from a range of chromosomes and subgenomes is of little use for downstream analysis, except perhaps to determine if a given sequence is present within a genome and as such it is rarely discussed within the literature.

These mosaic chromosomes are rarely revisited with two notable exceptions being by Zhu *et al* [392, 393] for *T. aestivum* and *T.dicoccoides*. In this work, Zhu *et al* used optical mapping to correct misassemblies and anchor contigs and scaffolds from the "Un' chromosome to their chromosomes of origin. Optical mapping is a strategy commonly used within genome assembly efforts which is capable of fluorescently tagging specific sequence motifs [374]. These motifs are then fluoresced and imaged, resulting in a map of the motif locations throughout the chromosomes. Such maps can then be used to anchor assembled contigs and scaffolds from genome assembly efforts, giving them more biological meaning that a set of biological data with origins unknown [11, 374]. Optical mapping is considered a relatively inexpensive technology and it often accompanies many reference-quality genome assembly projects [11, 193, 374]. However it is not able to place all assembled contigs

and scaffolds, as evidenced by the less than 100% genome completeness for most genomes, including those in table D4.

For *T. diccoides*, the optical map strategy was able to incorporate over 300 previously unplaced scaffolds, correct 226 misassemblies and re-orient a further 332 fragments [393]. For *T. aestivum*, the improvement was even more dramatic with 10% of the genome being corrected regarding position and orientation, and 279 previously unplaced scaffolds being placed within the genome [392].

In the previous chapter ("Kmer spectra for comparative genomics") allopolyploid genomes were shown to contain subgenome-specific signatures which could be leveraged to cluster chromosomes by their subgenome of origin. As such in this work the possibility to leverage such signatures for "Un" chromosome recruitment is assessed.

However, working with k-mer spectra for entire genomes is well-documented to be highly resource intensive [255, 266]. In the previous chapter, MinHash k-mer sketches, generated by Sourmash, were employed to generate chromosomal signatures of k-mer spectrums in place of storing the entire k-mer set [**?**]. This allowed pairwise comparisons of chromosomes to be performed rapidly and efficiently. However in this work, the aim is not to compare but to recruit by having the full suite of k-mers available to maximise the information for each subgenome. To achieve this, given that explicit storage of the full set is prohibitive, Bloom filters are implemented for k-mer storage and set membership tests.

Bloom filters are probabilistic data structures capable of storing large amounts of information in a succinct manner [262]. They work by first initializing a bit array of a size (m) reflective of the expected number of elements (n), a desired false-positive rate (p) and the number of hash functions used (h) by filling all positions with "0s". In an ideal world, the false positive rate could be kept very low, increasing the reliability of membership queries, however, this will result in an increase in bit array size. As such, bit array storage space and the false positive rate have a trade-off. Once a bit array has been generated, sequences are

then hashed using the number of distinct hash functions (h) previously specified, to obtain a set of numbers for which they then use to set the bit array to "1s" at those positions. This process can be seen in figure 3.1.

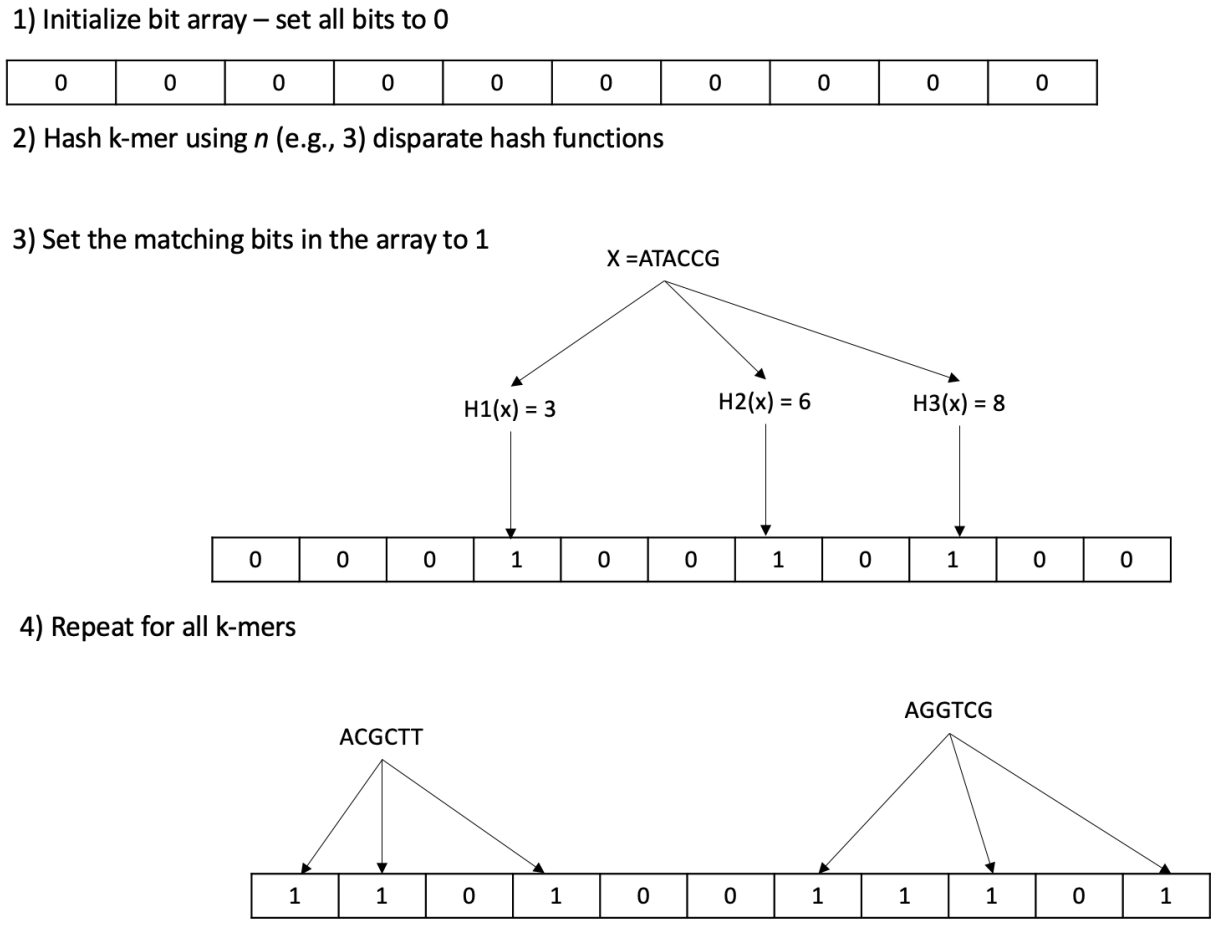


Figure 3.1: An overview of Bloom filter construction; initialization (1), query sequence preparation (2/3). and bloom filter population (3/4).

Querying the bloom filter then involves hashing the query k-mers and querying the array to see if all positions from those hash functions are set to 1. If they are, the Bloom filter check returns "True", else it returns "False" , with the former meaning the k-mer has been identified in the set and the latter meaning it is not present in the set (figure 3.2).

Bloom filters have been used extensively for biological applications including genome

assembly, sequencing read error correction, experimental sequence querying, sequence classification and k-mer counting operations [68, 132, 138, 144, 155, 178, 231, 243, 307]. To address the problem of false positive hits from Bloom filters, Pellows *et al* [262] developed a k-mer bloom filter specific strategy to reduce false positive hits without increasing the size of the Bloom filter, and thus the memory requirements. To achieve this they made use of the fact that for the vast majority of k-mers, their preceding (k-1) and succeeding (k+1) sequences will also be present in the set, given that k-mer sets are generated by sliding a window over a sequence. For their first strategy, KBF1, the Bloom filter is checked for the k-mers presence and if that returns True, its preceding **or** succeeding k-mers' presence is also queried. If that also returns True, the k-mer is considered present, else it's considered a False positive (figure 3.3).

For their second strategy, KBF2, the bloom filter is checked for the k-mer's presence and if that returns True the k-mers proceeding **and** succeeding k-mers returns True the the k-mer is considered present, else it's considered a false positive (figure 3,4).

However, for organisms without circular chromosomes the edge k-mers of the chromosomes will not have a proceeding or succeeding k-mer and so the KBF2 strategy would return a false positive. As such, k-mer on the edges of chromosomes are stored in an edge set. If a k-mers proceeding or succeeding k-mers return False, this edge set is checked to determine if it is an edge k-mer. If it is within this set, it is considered present, else it is considered not present.

It is important to note that the edge set is not stored in a Bloom filter; it is stored in a data structure that does not have false positives (e.g. list or hash table). As such, there is no need to query the edge k-mers preceding or succeeding k-mers as the k-mer is definitively present. Edge sets can be stored explicitly for fully assembled chromosomes as the size will small given that the expected number of edge k-mer is $2 \times n$, where n is the number of chromosomes.

In this work we implement the KBF-based bloom filter strategy to minimize false positives for a bloom-filter-based search of set membership for unassigned contigs subgenomic membership.

Methods

All code used in this project is available from the following github repository: `https://github.com/Glfrey/Bloom_recruit`.

The same chromosome data was used as in [392]. All *Triticum aestivum* IWGSC refseqv1.0 chromosomes, including the "Un" chromosome was downloaded from the Urgi database. The contigs that were re-assigned positions using optical mapping by [392], as detailed in their supplementary material, were extracted from the "Un" chromosome and placed into a subgenomic file using a custom Python script (extract_iterate.py).

IWGSC refseqv1.0 chromosomes were concatenated into a single subgenomic file using a custom python script (fasta_cat.py). K-mers of size 21, 31, 41, 51 and 61 were extracted for each subgenome using KMC3 [179]. Bloom filters for each subgenome were constructed using PyProbables (`https://github.com/barrust/pyprobables`) with error rates of 0.1, 0.01, 0.001 and 0.0001.

Subgenomic assignment of the extracted "Un" chromosomes was performed using the classic method of Bloom filter membership checking and also the KBF strategies as detailed in [262] using our python programs bloom_filter_check.py, KBF1_check.py, and KBF2_check.py. The k-mer edge sets required for the KBF2 strategy were constructed using edge_set_upper.py which ensures all k-mers are converted to upper case letters to match the KMC k-mer output as PyProbables Bloom filters are case sensitive. Determining which subgenome each assembled contig belongs to was determined via a vote system with each k-mer being checked for membership in subgenome Bloom filters and casting a vote for each matching Bloom filter the k-mer matched. The subgenome with the most votes won the placement of the contig.

Assessment of cluster success was performed by enumerating the number of contigs that were assigned to the correct subgenome as determined by [392].

A histogram of correctly and incorrectly assigned reads was constructed using the seaborn library [356]. Intersection upset plots were produced using the "UpSetR" library [76]. F1 and micro-F1 visualisations were constructed using ggplot2 [362] in R Studio 3.6.1(2019-07-05) [296]

An inspection into the unique k-mer content of the subgenomes as a function of k-mer size was performed using KMC3, with visualisation performed using [362].

Results

K-mer classification using subgenomic k-mer profiles shows a linear, positive relationship between k-mer size and the success metrics: F1 for assessing individual subgenome classification success, and micro-F1 for assessing combined classification success across all subgenomes (figures 3.6, 3,7 and 3.8 A-E). Figure 3.6 shows the micro F1 scores (A and D) and subgenomic F1 scores (B, C, E, F) achieved for the classic method of Bloom filter querying. It shows almost identical results for the KBF1 strategy (3.7) and KBF2 strategy (3.8) developed by [262] and as such there has been no particular benefit of implementing these strategies for this work. All Bloom filter querying strategy's (original, KBF1, KBF2) show a peak F1 scores of 0.94 for k=61 for all bloom filter error rates and bloom filter query strategies.

The strong dependence of classification success on k-mer size is evidently being driven by the growth in subgenomic k-mer distinctness as seen in figure 3.9. For the k-mer range used in this study, there is a gradual decline in the percentage of shared k-mers between the subgenomes. At k=21, the number of k-mers shared between the subgenomes makes up 12% of subgenome A, 11% of subgenome B and 12% of the subgenome D. By k=61, this falls to 2%, 2% and 3% respectively. For k=101, the percentage of shared subgenome k-mers drops

below 1% for all subgenome (0.6% to subgenome A, 0.5% for subgenome B and 0.7% for subgenome D).

The bloom filter error rate had no impact at all on the classical bloom filter querying approach (non-KBF) as seen in figure 3.6 which shows identical F1 and micro F1 scores across all k-mer values. Bloom filter error rate had a small, but measurable impact on classification performance for both KBF strategies, with the higher error rate curiously outperforming lower error rate bloom filters in many instances (figures 3.7 and 3.8).

It is evident from the intersection analysis, visualised as an upset plot in figures 3.9-3.11, that contig recruitment using subgenomic k-mer composition profiles is highly robust. These upset plots show the intersection of shared k-mers (dots) across differing parameters (seen horizontal bars) and their proportion of the data (vertical bar). Across all k-mer sizes, Bloom filter error values and KBF querying strategies the classifier consistently places the same contigs within the same genome. As k-mer size grows a greater number of contigs are clustered, which accounts for the discrepancy in the number of contigs classified across figures A-F for figure 3.10-3.12.

An analysis of the size of the correct and incorrectly classified contigs reveals overlapping size distributions and so in these instances contig sizes are not a driver of classification success (figure 3.12).

Discussion

The aim of genome assembly algorithms over the decades has been to increase the 3 C's of genome assembly; contiguity, completeness and correctness [332]. This work contributes to this effort by proposing a novel, easily-implemented and robust subgenomic classification strategy using subgenomic k-mer profiles. By being able to place assembled contigs and scaffolds within their subgenome of origin the completeness and correctness of the assembled genome grows as a greater proportion of the data becomes usable for downstream analysis.

Such subgenomic assignment of assembled data may also aid contiguity by reducing the search space for their correct location from an entire polyploid genome, to a single subgenome.

To ensure this classification strategy is robust and to assess the impact of various parameters on classification success a number of experiments were performed and assessed using F1 metrics. The F1 measure is simply the harmonic mean of precision of recall, two measures which capture complimentary information. Where precision characterises how many of the contigs were correctly classified out of all contigs $\frac{Truepositive}{Truepositive \div Falsepositive}$, recall is concerned with how many of the contigs that should have been classified as a given subgenome, were; $\frac{Truepositive}{Truepositive \div Falsenegative}$. While the F1 metric is designed for binary classification, it can be generalised to multi-class classification problems via macro- and micro-averaging [103]. Macro-averaging is simple in that it averages the F1 score across subgenomes by combining all F1 scores and dividing them by the number of classes; $\frac{F1.1,F1.2,..F1.m}{n}$, n is the number of classes. Micro-averaging take a different approach by calculating an overall F1 score given the confusion matrix for all classes $\frac{TP.1,TP.2,..TP.m}{(TP.1,TP.2,..Tp.m+TN.1,TN.2,..TN.m)}$, where TP refers to true positive instances and TN refers to true negatives. In the context of the work here, a true positive would be a contig from subgenome A which has been correctly classified as belonging to subgenome A and a true negative would be contig from subgenome B which has been correctly classified as not belonging to subgenome A. For imbalanced data sets, micro-F1 is preferred as macro-F1 weights each class (subgenome) equally even if they contain differing numbers of contigs [103]. This work contains imbalanced data, given that subgenome B contains almost twice as many contigs as subgenome D. The micro-F1 score can also be interpreted as the total probability of true positive classifications for each class [328].

Given this metric of performance, it is clear that the most impactful parameter for this method is k-mer size, with classification success growing linearly as a function of increasing k-mer size. This remains true largely regardless of Bloom filter false positive rate, or Bloom

filter querying strategy. The characterisation of shared k-mer content across the *T.aestivum* subgenomes reveals this is due the decline in the proportion of shared k-mers as a function of k-mer size. As such the use of larger, more distinct k-mers should continue to yield greater classification success up until the point where the k-mers become so large they exceed the size of the fragments being recruited or they become so distinct that they are unique within the genome and so won't match any previously unanchored sequences.

Comparing figures 3.6-3.8 highlights the minimal impact the KBF strategies developed by [262] had on the performance of subgenomic classification and in two instances were outperformed by the classic Bloom filter querying strategy. Given the positive impact of KBF strategy as demonstrated by their benchmarks in [262] this was surprising, especially for the Bloom filters with a high false positive probability (e.g 0.1). However, it may be that the contigs contain such subgenome-rich k-mer profiles that the false positive rate was not able to impact classification success in a meaningful way. Perhaps for sequences with little subgenomic k-mer information this strategy would be more impactful for reducing the impact of false positive on contig classification.

For the contigs used in this work, size has no impact on classification success, with overlapping size distributions for both correct and incorrect classified sequences (figure 3.12). However, work should be done to establish the minimum contig size that this method is capable of recruiting. Only two other subgenomic-sequence recruitment methods exist which use k-mer profiles; PolyCracker and SubPhaser of which the former requires a minimum of 1000bp and the latter has only been tested on fully assembled chromosomes [123, 159]. Unlike these methods which rely on subgenome-specific repeated k-mers the classification method developed herein uses the entire k-mer spectrum. As such, the classification strategy presented here is less specific, which could be driving the incorrectly classified instances by weighting every k-mer equally when determining where to place the contig.

This could potentially be rectified by storing a complementary data structure which

records subgenome-specific k-mers. By querying this structure first, the presence of any subgenome-specific k-mer within the contig would automatically win it a place within that subgenome. This could reduce the time take to classify those particular contigs as potentially the entire set of k-mers wouldn't need to be queried, but it would increase classification time for all other instances. Additionally, this extra data structure, if stored explicitly, will evidently grow linearly with k-mer size as shared subgenome sequence content decreases (figure 3.9). As such, intelligent data structure implementation would be required to keep memory requirements low. The largest Bloom filters for this work were found for k=61 with an error rate of 0.0001 (table E.1) but the size was still very modest; 7.5Gb for subgenome A, 8.1Gb for subgenome B and 5.8 Gb for subgenome D. Given that the entire set of 21-mers for the *T.aestivum* genome could not be stored explicitly using a hash table given 720Gb of RAM this shows an almost 10X reduction in memory requirements for the largest hash table. However this did come with a significant construction time cost, with the hash tables for k=61 taking over 7 days to construct.

Given the success of this approach for contig recruitment there remains two, very important, outstanding questions; can it be applied to other species and can this method be applied to smaller contigs or assembled data?

The answer to the first question can be found in the first chapter of this thesis: perhaps. This method relies entirely on subgenome-specific k-mer profiles which if not present, will cause the method to fail. Nearly all allopolyploids showed subgenomic clustering abilities for k-mer frequency (figures 2.3-2.7), but this method relies on k-mer composition. For this, only 5 allopolyploid species showed subgenomic clustering and so the application of this approach seems narrow. However this can be addressed. First, before implementing this method, it would be highly recommended to perform a subgenomic k-mer profile clustering investigation as performed in the first chapter. Otherwise, subgenome-specific profiles could be assessed as performed for figure 3.9 using KMC3 (supplementary method 1). Second if only k-mer

frequency shows subgenomic clustering ability, then this Bloom filter could be replaced with a counting Bloom filter to facilitate subgenomic k-mer frequency queries. However, counting Bloom filters require as much as four-times the memory than traditional Bloom filters and so knowing how Bloom filter size grows with the probability of false positives and k-mer size (table E.1), hardware constraints will determine how practical this approach could be [42, 244]. Implementation of more recent space efficient counting Bloom filters such as those developed by [244] may make k-mer frequency-based sequence recruitment more feasible.

It is also worth noting that as k-mer size grows and they become more unique, their frequency will decrease. The k-mer space theoretically grows at a rate of 4^k (although a more realistic growth of k-mer uniqueness and size can be seen in figure 3.9 and figure 4.3)), but it becomes very sparse as K grows large [276]. As such, for sufficiently large k-mers, the frequency problem is reduced to a composition problem and would be better represented using a traditional Bloom filter given the extra space required for storing a counting Bloom filter structure [42, 244].

The impact of query sequence size also requires investigation. While no size-dependency relationship existed in the contigs used in this work, there may exist a critical thresholds below which the k-mer information content drops too low to be accurately classified. For PolyCRACKER this was determined to be 1000 bases [123]. However, given that this method relies solely on subgenome-specific repeated k-mers, the threshold may be lower. While testing this could be easily implemented, a major consideration is time. Querying the Bloom filters using the previously unplaced contigs took hours (table E.2). Given a particularly large set of sequences, such as those returned from sequencing runs, this method could take weeks or even months. To address this, a pseudo-parallelization approach could be taken where the query set is split up and a number of parallel classified analysis are ran at the same time. Implementing a truly parallel querying strategy would require a different implementation of Bloom filters such as those developed by [241]. It would also be possible to construct the

Bloom filters using a pseudo-parallel approach given that PyProbables (the library used in the work) supports Bloom filter merging [30].

Should it be possible to have a more efficient run time, and should this method be capable of recruiting shorter sequenced fragments, such as read data, then this strategy has the potential to improve future polyploid genome assembly efforts. Separating the subgenomes prior to assembly could reduce the polyploid-induced complexity within the de Bruijn Graph which causes chimeric, misassembled and highly fragmented genome assemblies [48, 251, 260].

Conclusions

This work has developed a novel, space-efficient classification strategy for recruiting previously unassigned contigs using subgenomic k-mer profiles. The parameter with the most influence on classification success was k-mer size, with classification performance improving with increasing k-mer size. This corresponds to k-mer uniqueness as demonstrated via a assessment of subgenomic k-mer profiles. However there is an obvious trade-off with k-mer size and time and space requirements for bloom filter construction and storage the optimal of which will be determined by computational hardware and time available. The KBF bloom filter querying strategy's as developed by [262] did not improve classification performance and in some cases was actually outperformed by the classic bloom filter querying strategy. Given the extra time this strategy takes to perform k-mer queries, its implementation in this particular k-mer classification problem is not justified. Improvements for future implementations have been suggested, such as replacements of bloom filters with counting bloom filter and implementation of pseudo-parallelization and true parallelization strategies to broaden the applicability of the classification algorithm developed herein.

Chapter Discussion

This chapter has expanded on the subgenomic k-mer profiling work of chapter 1 and applied it to the problem of contig anchoring to develop more correct, complete, and, potentially in the future, contiguous genome assemblies. While the major drawback is that the contigs are assigned to subgenomes rather than chromosomes, this still presents a method to increase assembly completeness and correctness for individual subgenomes.

The future directions outlined in the discussion and conclusion are certainly worthy of more exploration. Work to determine the minimum size for successful recruitment based off global subgenomic k-mer profiles is already underway, but is not include here given that a) the experiment didn't show a reliance on fragment size for classification and b) the sheer time it takes for sequence classification. This is a major roadblock in the application of this method for large-scale subgenomic fragment recruitment. A potential solution via pseudo-parallelization was discussed and it was discovered the paralellizable bloom filters have been developed and so may present as an opportunity for future exploration.

An initial experiment was performed with PolyCRACKER which was, at the time, the only subgenomic clustering strategy in the literature to assess its applicability for short-read assembly recruitment. The assembly was performed using Abyss2.0 [155] and a range of k-mer sizes which took over 500GB RAM and several days to complete even when using MPI-parallelization. Disappointingly, though not surprisingly given that PolyCRACKER requires a minimum fragment length of 1000bp, the result was poor, with over 88.21% of contigs being $>$ 200bp. Given that contigs shorted than 200bp are routinely filtered out in assembly projects, and is required for being deposited into NCBIs assembly database these were removed [194]. To experiment with contig lengths affect on PolyCRACKER, contigs of length 2500bp, 2000bp, 1500bp, 1000bp and 500bp were given to PolyCRACKER to cluster and the best results were obtained for contigs of length 2500bp, which comprised

only 0.00037% of the assembled genome. Of this data, PolyCRACKER incorrectly clustered the contigs 50% of time as determined using the same validation strategy they implemented in their paper. An analysis of the contigs given PolyCRACKER found a complete lack of the repeat k-mers that PolyCRACKER requires as features for the clustering which explains the very poor results. Evidently, PolyCRACKER is not an option for the subgenomic clustering of short-read sequence data, or even poor assemblies.

Should it be possible to ensure a more reasonable run time, then exploration of this strategy to read-recruitment has the potential to fill this gap and provide an opportunity for pre-assembly read-filtering. By separating read data prior to genome assembly, polyploidy-induced topology could be eliminated and repeat-associated structures, reduced. This has the potential to significantly improve short-read genome assembly efforts, which given that long-read sequencing data still remains too expensive or routine sequencing work, could hugely benefit the polyploid genome community.

Chapter Conclusion

This work has presented a novel, easily-implemented and highly-scalable subgenome recruitment classification algorithm for polyploids which exhibit subgenome-specific k-mer compositions. Determining whether or not a genome has subgenome-specific k-mer composition profiles can be determined via metagenomic strategies such as Sourmash or other k-mer based subgenome clustering strategies. Subgenomic clustering success by PolyCRACKER [123] or [159] is not an indicator that this strategy will work for a given polyploid genome, as demonstrated by the ability for the Subphaser approach to subgenomically cluster the segmental allopolyploid *C. arabica* [159], while the Sourmash approach was unable to correctly split the two subgenomes (figure 2.7). This is likely due to the differences in approach to utilising subgenomic k-mer profiles. While the method presented here utilises global subgenomic k-mer profiles for "Un" contig classification, [159]

and [123] use subgenome-specific repeated k-mers. A global measure of subgenome-specific k-mer composition should be used for assessment before implementing the method presented here.

The major limitation for this strategy is the time complexity associated with building and querying the Bloom filters given a large set of query sequences. A pseudo-parallelization strategy has been discussed and truly parallelized Bloom filters identified as a possibility for further implementation. The parameter with the most impact on classification accuracy is k-mer size and as such this strategy is best implemented with larger, more specific k-mers although this does require more space to store the Bloom filters. Precisely which k-mer range to explore can be determined by performing a subgenome-specific k-mer analysis as performed for figure 3.9 using KMC3 [179].

For future work, this time complexity of the Bloom filter building and querying must be addressed and so too should determining a minimum contig size threshold for accurate classification. Subgenomic recruitment of shorter sequenced fragments, such as read data, prior to genome assembly has the potential to reduce the polyploid-induced complexity within the de Bruijn Graph which result in chimeric, mis-assembled and highly fragmented genome assemblies.

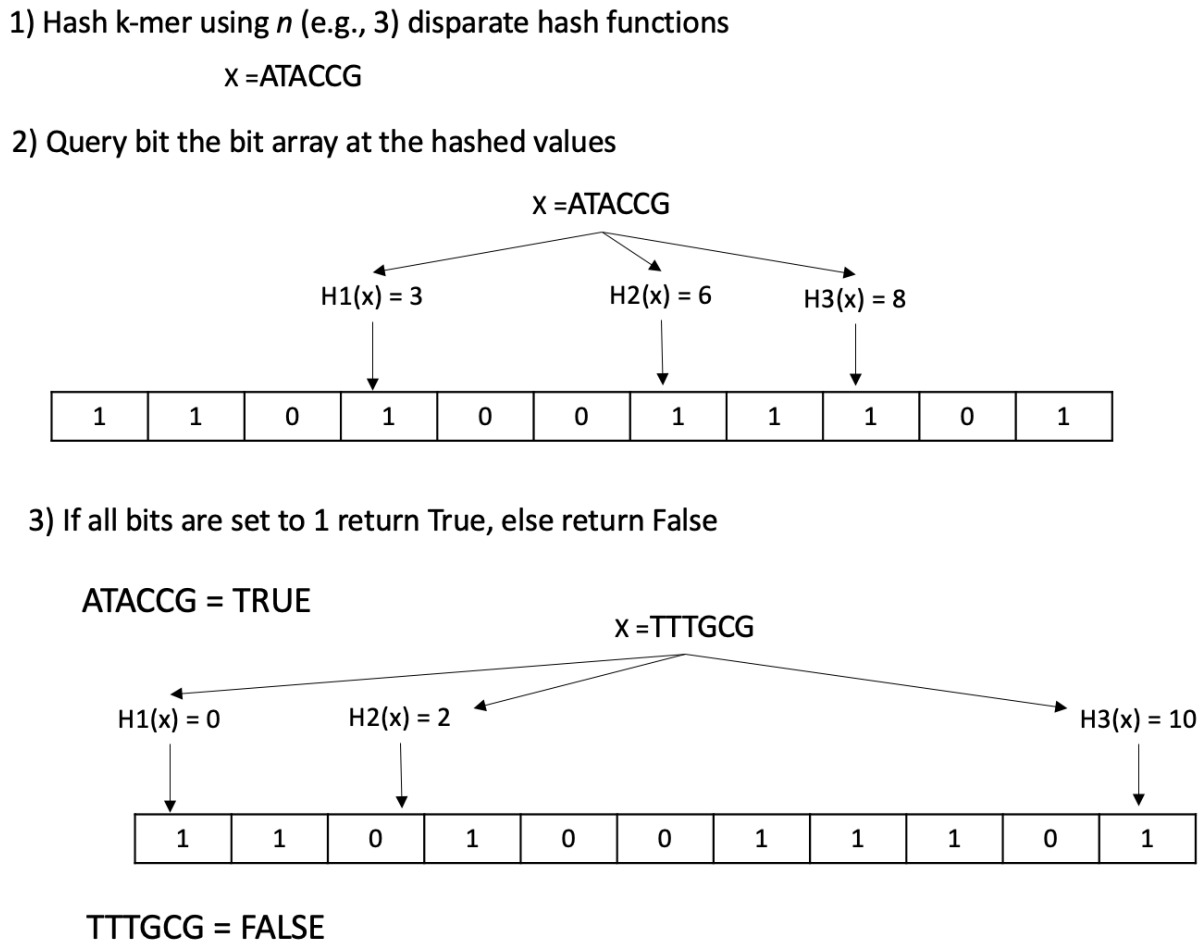


Figure 3.2: An example of Bloom filter querying. Query sequences are first k-merized using the same k-mer length that was used for Bloom filter construction (1). The Query sequence is then hashed using the same hash function used during construction (2) and the returned positions checked if the bit has been set to one, which indicates presence or if it is zero which indicates absence (3).

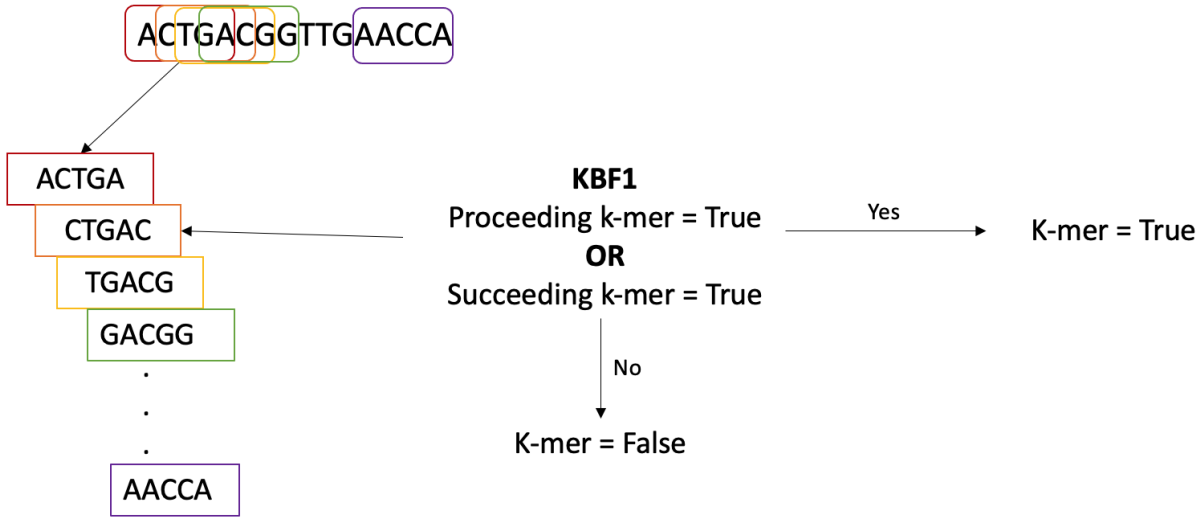


Figure 3.3: A example of the KBF1 strategy developed by [262]. A query k-mer is checked for presence within the Bloom filter after which so are its preceding (k-1) and succeeding (+k1) k-mers. If either of those k-mer also return True then the k-mer is considered present. Else, the k-mer is considered a false positive call and so it's presence is not recorded.

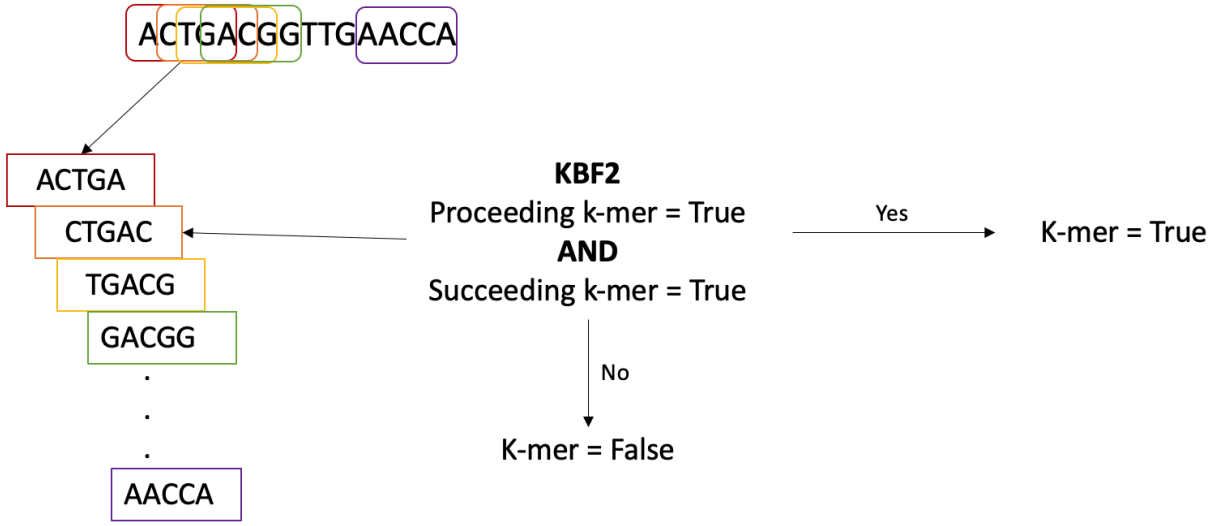


Figure 3.4: A example of the KBF2 strategy developed by [262]. A query k-mer is checked for presence within the Bloom filter after which so are its preceding (k-1) and succeeding (+k1) k-mers. If both of those k-mer also return True then the k-mer is considered present. Else, the k-mer is considered a false positive call and so it's presence is not recorded.

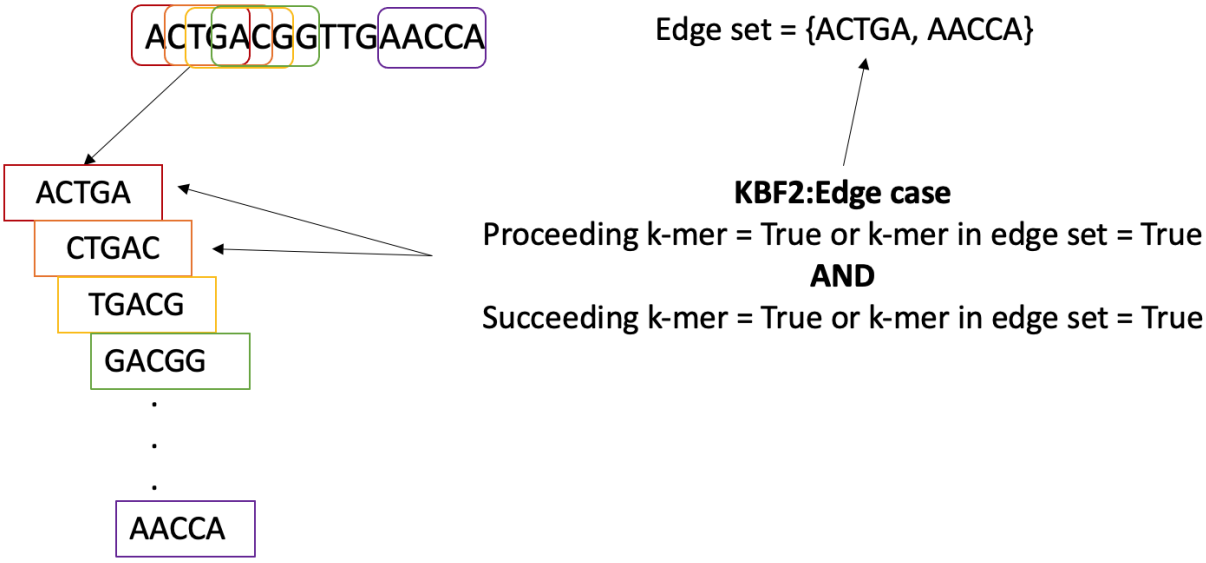


Figure 3.5: A example of the KBF2 strategy developed by [262]. A query k-mer is checked for presence within the Bloom filter after which so are its preceding (k-1) and succeeding (+k1) k-mers. If both of those k-mer also return True then the k-mer is considered present. Else, the k-mer is checked to see if it is present within the edge set which stores edge k-mers which do not have a preceding or succeeding k-mer. If this query returns True the k-mer is considered present, else it is considered as a false positive bloom filter query and its presence is not recorded..

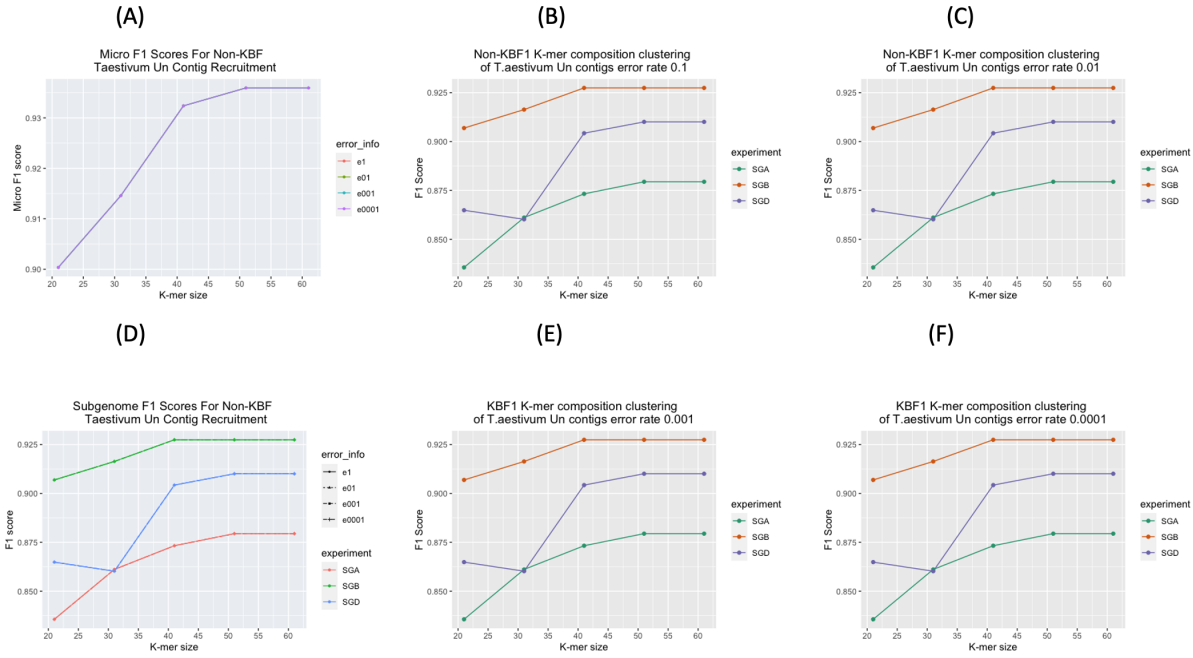


Figure 3.6: Micro F1 scores (A,D) and subgenomic F1 scores (B,C,E,D) for the classical Bloom filter querying approach as a function of k-mer size and bloom filter error rate.

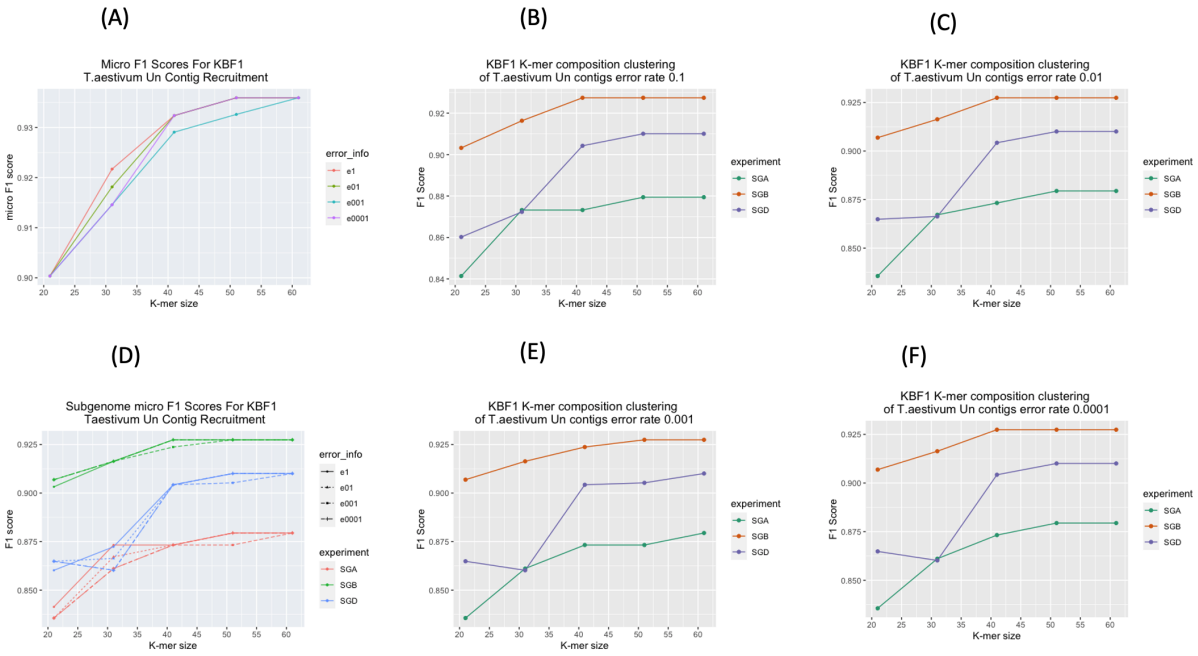


Figure 3.7: Micro F1 scores (A,D) and subgenomic F1 scores (B,C,E,D) for the KBF1 Bloom filter querying approach developed by [262] as a function of k-mer size and bloom filter error rate.

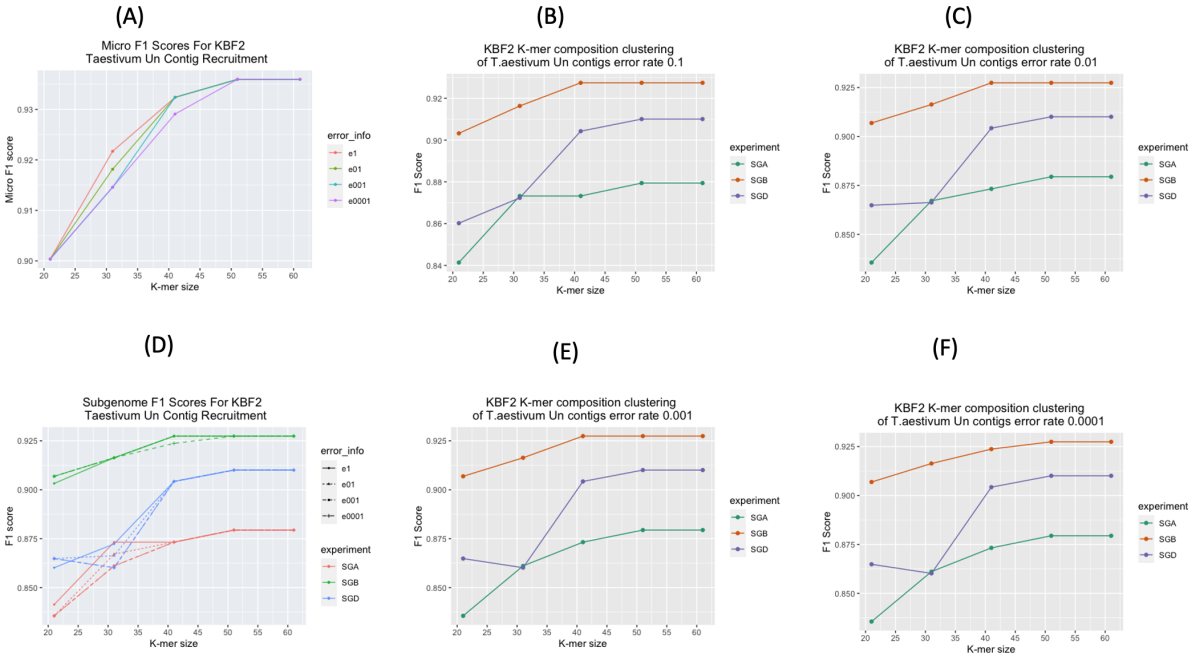


Figure 3.8: Micro F1 scores (A,D) and subgenomic F1 scores (B,C,E,D) for the KBF2 Bloom filter querying approach developed by [262] as a function of k-mer size and bloom filter error rate

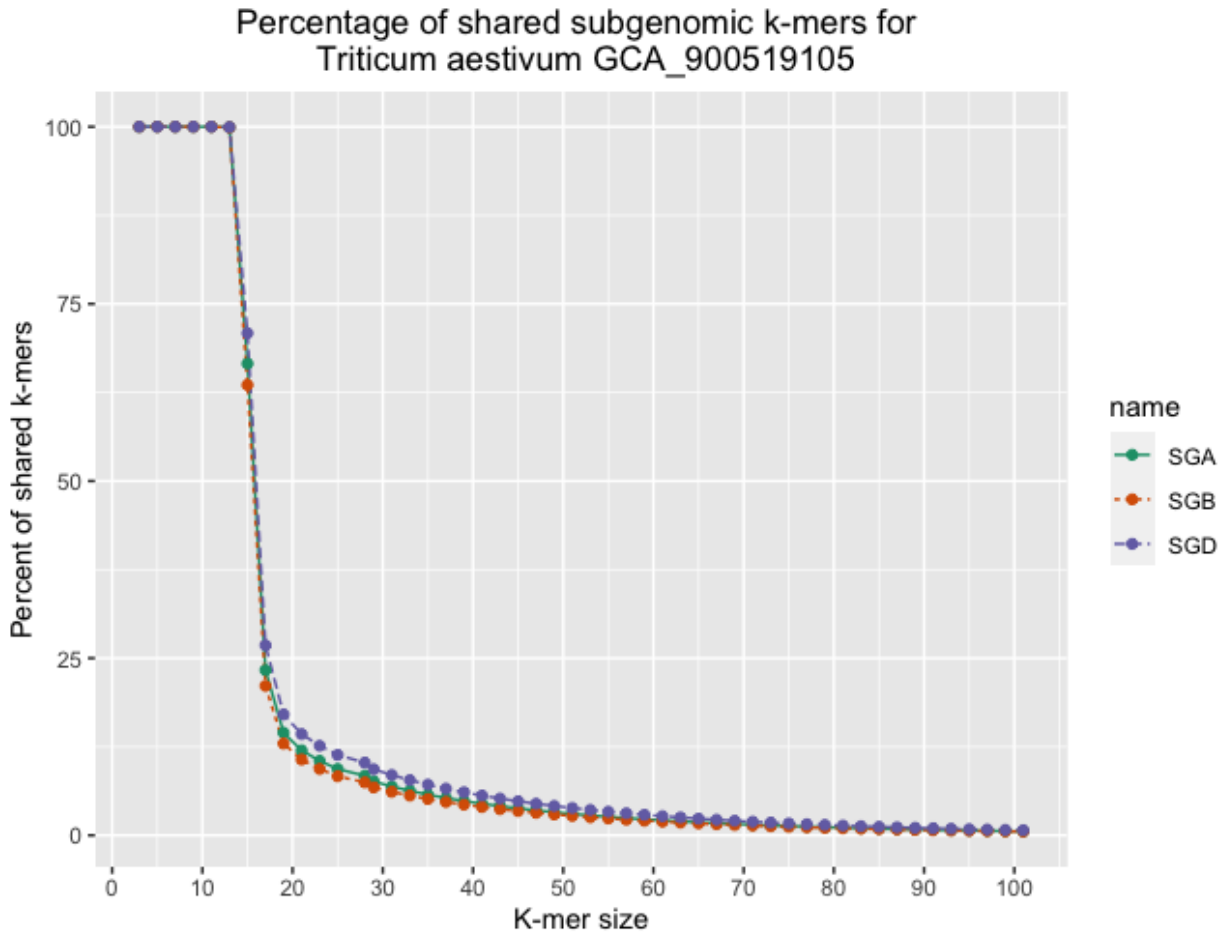


Figure 3.9: Percentage of shared subgenomic k-mers as a function of increasing k-mer size for *T.aestivum*

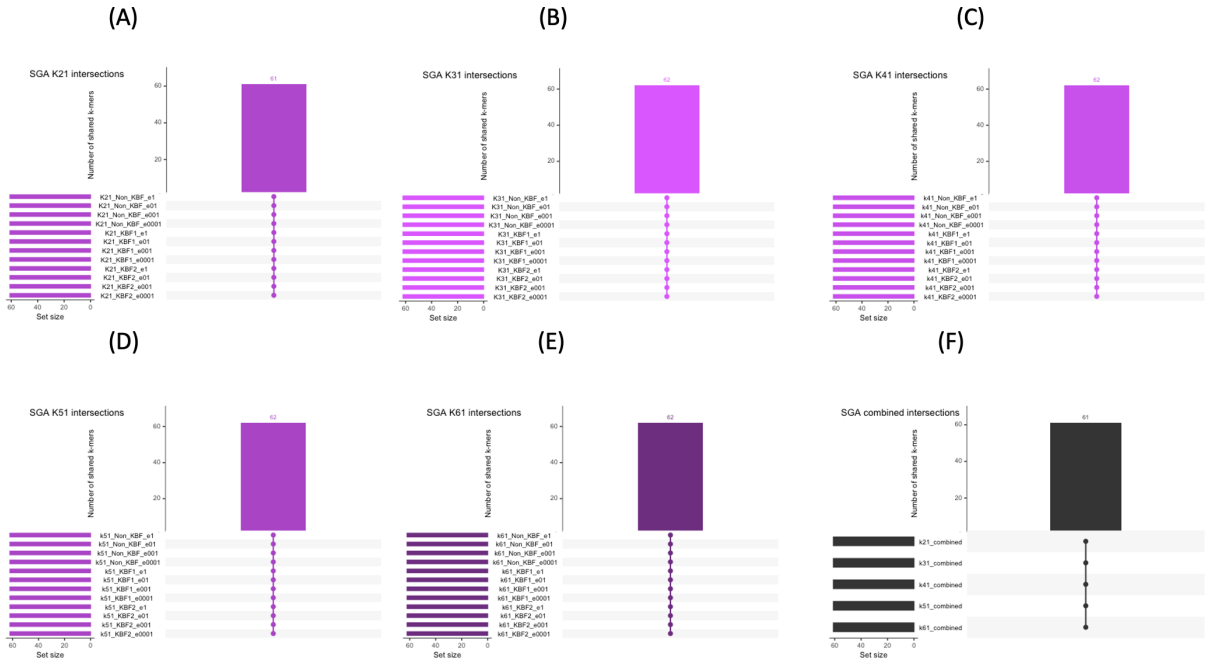


Figure 3.10: An upset plot showing intersection of contig classification across k-mer sizes and Bloom filter error rates (A-E) for subgenome A. Figure F shows the combined intersection of classified contigs across all error values for the k-mer range k21-61.

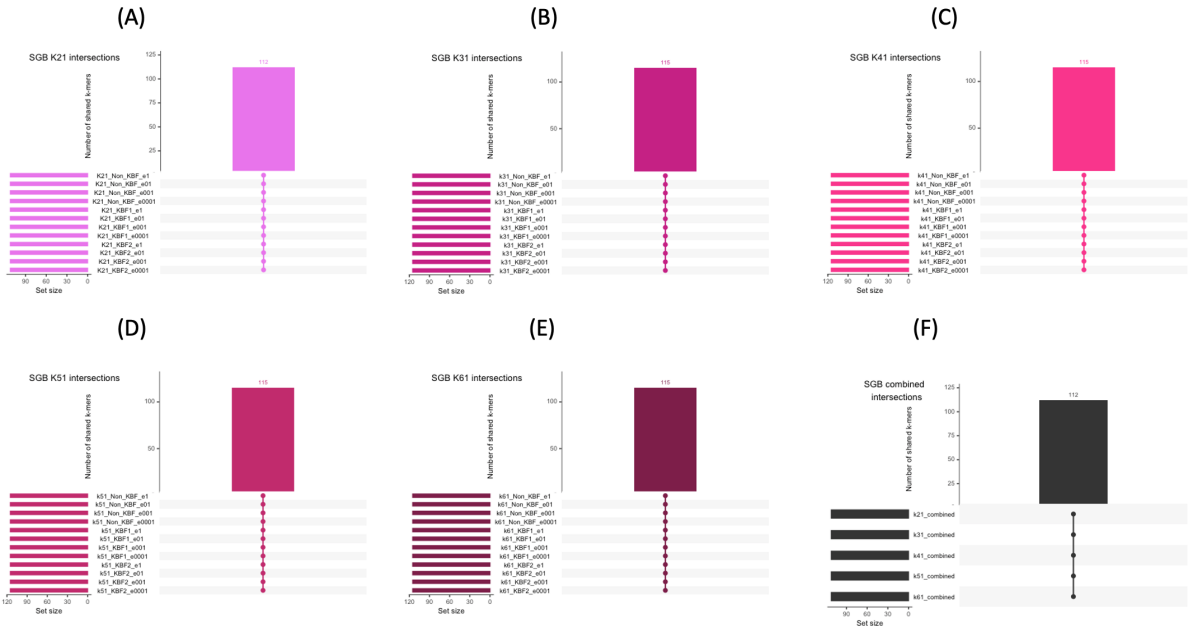


Figure 3.11: An upset plot showing intersection of contig classification across k-mer sizes and Bloom filter error rates (A-E) for subgenome B. Figure F shows the combined intersection of classified contigs across all error values for the k-mer range k21-61

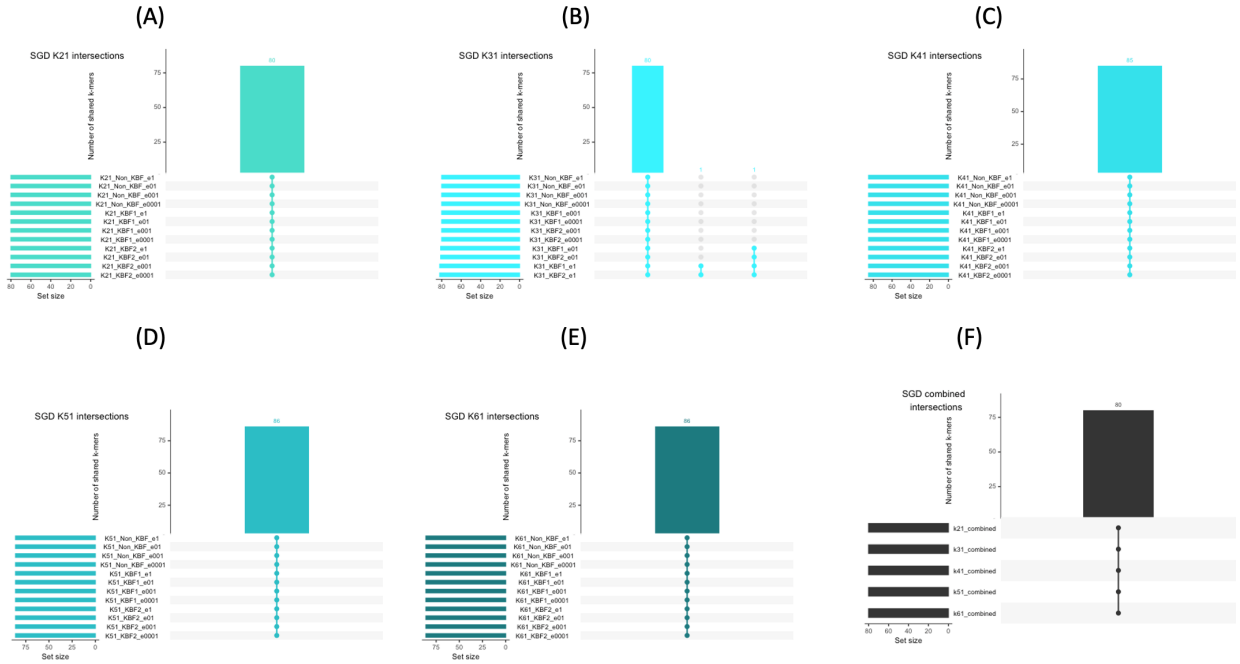


Figure 3.12: An upset plot showing intersection of contig classification across k-mer sizes and Bloom filter error rates (A-E) for subgenome D. Figure F shows the combined intersection of classified contigs across all error values for the k-mer range k21-61

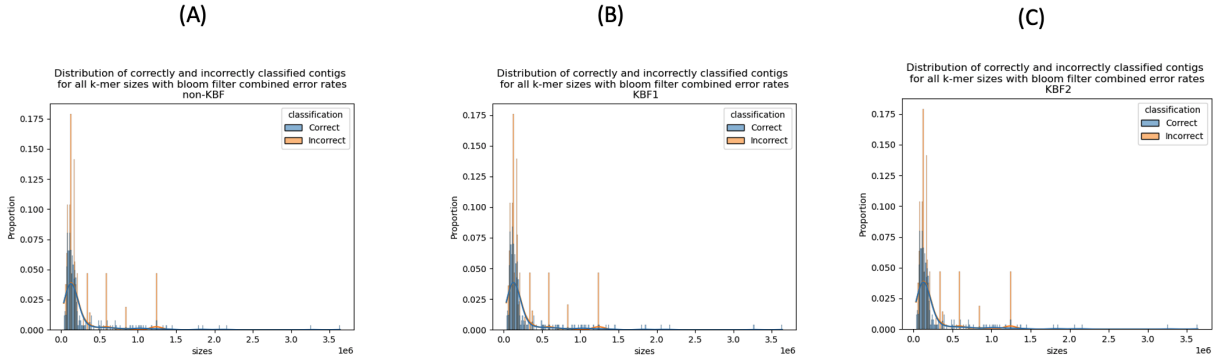


Figure 3.13: Histograms of correctly (blue) and incorrectly (orange) classified contigs as a function of contig size for the classic bloom filter querying strategy (A) and the KBF1 and KBF2 strategies developed by [262] (B,C)

MODERN VARIANT FINDING

Identifying genetic variation between genomic sequences is a fundamental problem in biology that can seem deceptively simple in theory, but is very complex in reality. The sheer size of the data and the genomic complexity of the organisms under investigation are both major complicating factors when it comes to identifying the variants that underpin agronomic traits of interest and can certainly method selection for variant identification.

Classic methods, such as alignment to a linear reference genome and then variant calling with programs such as FreeBayes and GATK offer a computationally efficient solution that is attractive for those with large genomes to analyse [117, 339]. However many large polyploid genomes, such as wheat, are mostly comprised of highly variable and repetitive transposable elements that complicate the alignment to a reference genome and reduce the effectiveness of this method for variant finding [361]. Further such methods are prone to reference-bias which inhibits the ability to discover genomic variation [331, 338].

The alternative strategy of performing a *de novo* genome assembly and then performing a comparative genomics analysis between the assembled fragments can seem an attractive option as it does not suffer from reference-bias. However, genome assembly is a non-trivial venture requiring multiple iterations of the assembly pipeline to tune parameters, with each run taking potentially months and requiring significant computational resources [8]. Further, assembly algorithms are poorly designed to handle polyploid genomes and are prone to collapsing shared subgenomic sequences and repeats, forming chimeric contigs and without additional long-range data such as optical maps, are likely to be heavily fragmented [20, 22, 204, 251, 329].

As such there is a need to develop new strategies to overcome these limitations and perform comprehensive and accurate variant calling in crop genomes. In this work I critically discuss how modern approaches to variant calling are being developed to address these

problems and how they can be applied to crop genomes. I further propose a variant calling algorithm which enhances current capabilities for a non-alignment-based variant calling strategy and highlight its strengths and its weaknesses for further development.

A Guide To Modern Variant Calling In Crop Genomes: Possibilities And Problems

Abstract

Whole genome sequencing is the only method available capable of capturing the full suite of genomic diversity present within a crop genome. However given that whole genome sequencing of crop genomes involves fragmentation, the major challenges lie in reconstructing the original genomic data. This is further exacerbated by technological limitations which limit sequenced data length and introduce various errors which if not accounted for, can result in false-positive variant identification. Currently, most whole-genome crop data is comprised of highly accurate, but short (100-250bp) short-read sequencing data. While longer-read technology exists, it is far more expensive to perform and suffers with a higher error rate. However short-read sequencing data presents major challenges for the reconstruction of the original genomic sequence.

Introduction

Variant identification is a critical venture for characterising the full suite of genomic diversity present within crop genomes, determining the genetic basis for agronomic traits of interest and understanding the working of crop genomes on a functional level [162, 163, 164, 365]. Such work is essential given the threats to 21st century food security stemming from global climate change and increased demands for crop products [34, 102, 124, 151, 228, 324].

Food security does not just encompass access to food though, but access to safe and nutritious food. Modern breeding efforts are focusing not only on developing novel crop varieties that are resilient to biotic and abiotic stresses, but also varieties which are

nutritionally enhanced, with harmful byproducts reduced or removed. This include cyanide-free cassava, rice varieties with low arsenic and cadmium contents and varieties of wheat with lower acrylamide-forming potential (a carcinogen) and phytate levels, which reduces the bioavailability of micronutrients [78, 133, 271]. Harmful effects of agriculture to the environment can also be reduced, with the development of varieties resistant to various pests and diseases, those that require less nitrogen input, and those requiring less water being possible, all of which decrease environmental damage and increase sustainability [285, 322]. The reduction of waste, either through reduced loss in the field or through enhances shelf lives is also highly beneficial to society as a whole [271, 311].

Currently a plethora of methods exists for capturing genetic variation, including whole exome sequencing, RNA-seq, genotype-by-sequencing (GBS) and other SNP panels [35, 49, 166, 337].

As present it is SNP-data, (often generated via various reference-based SNP-based panels), which dominates the understanding crop variomes [83, 177]. It also comprises the vast majority of the data used by crop breeders for investigations into genome-phenotype relationships through strategies such as genome-wise association studies (GWAS) and integration of this information into the development of novel crop varieties via strategies such as genomic selection (GS) [83, 177].

However the reliance on SNP panels has two major flaws. The first is that they are based on a single reference genome and as such are highly susceptible to reference-bias [83]. The extent of this reference bias is being slowly revealed thanks to the construction of pan-genomes for numerous crops which showcase a large number of presence/absence variants (PVs), some of which concern genes of great agronomic importance [113, 128, 129, 235, 292, 303, 309, 330]. An excellent example of the utility of pan-genomic-based markers for GWAS investigations is the pan-genomic discovery of the sugarcane mosaic virus resistance inferring gene *ZmTrxh*, which was not present in the single reference genome [112]. Given that at least

32.83% of pan-*Zea mays* genes are absent from the most commonly used Maize reference genome, it is likely more agronomically important traits are hiding in the variable set of pan-genomic genes [128].

The second major problem is that they identify a single type of genetic variant (SNPs), neglecting the full suite of genetic diversity responsible for traits of agronomic interest. The importance of including larger variants such as structural variants (SVs) in studies investigating genotype-phenotype relationships in crop genomes is being increasingly realised with an number of agronomic traits of interest being caused by SVs [83, 113, 128, 129, 135, 158, 177, 292, 303, 330]. It is also apparent that the further characterisation of SVs may reveal that SNP-based loci thought to be responsible for traits, may actually be caused by SVs as discovered for the fruit-weight loci in tomato [12]. It has also been highlighted that the failure to include a greater diversity of variant data has the potential to reduce the predictive ability of genomic selection and genotype X environment prediction studies, which therefore has a direct impact on crop breeding programs[83].

As such, it is becoming increasingly recognised that there is a need to extend variant finding efforts to include a greater diversity of variant types and the impact they may have on traits of interest. Integration of SV information into GWAS analysis is not currently performed routinely but its use in a number of species such as *Brassica napus*, *Prunus persica*, *Glycine max* and *Cucumis sativus* revealed SV-associated traits of agronomic interest [129, 208, 319, 383].

However, identification of non-SNP variant data, especially SVs, is technically challenging. Only whole-genome sequencing (WGS) data is capable of uncovering all genetic variation present within a genome and yet this data is challenging to work with. WGS produces a massive volume of highly-fragmented genomic data with individual fragments (termed "reads") representing only a tiny portion of the large genomic data. Long-read sequencing technology capable of returning fragments several kilobases in size can be applied

to whole genome sequencing, but is significantly (between 5X-12X) more expensive and as such is infrequently applied to resequencing projects for variant detection [386]. Therefore this work will focus primarily on short-read-based variome reconstruction.

Traditionally, there exists two primary strategies for short-read variant identification; Alignment to reference or *de novo* assembly. Where short-read alignment attempts to play spot-the-difference between the read data and a given reference genome, *de novo* assembly attempts to piece the reads together to reconstruct the genome first, before playing spot-the-difference with other assembled genomes. Both of these strategies have drawbacks. For read-alignment, this is primarily reference bias which limits the variants which are discoverable [331, 338]. If a read is too divergent in sequence content from the reference genome, alignment algorithms simply won't be able to align it. It is possible to compliment a read alignment with *de novo* reference assembly of the unaligned fragments, which has been termed the "pseudo-*de novo*' approach [99, 100, 143]. However caution has to be taken with this approach given that unmapped reads can consist not only of novel variants, but also low quality reads and contamination. Additionally, given that some reads from novel regions may have aligned, the resulting pseudo assembly of novel fragments may be incomplete.

For *de novo* assembly the major issue is that it's a highly complex venture which requires more data than a single whole-genome sequencing data. Most *de novo* assembly endeavours usually feature multiple whole-genome sequencing data sets, including a combination of long and short-read technologies, alongside various complimentary sequencing technologies, such as genetic maps and genome-wide chromatin data [232, 318].

However, modern variant calling strategies attempt to mitigate the issues presented by traditional single-reference alignment approaches and *de novo* assembly, enabling large-scale variant calling for short-read sequencing data. For alignment-to-reference strategies, this is via the introduction of multi-reference alignment data structures which mitigates single reference genome bias by incorporating known genomic variants to improve read-mapping

[16, 261]. For *de novo* strategies the introduction of coloured de Bruijn graphs (DBGs) has been instrumental in enabling variants to be called between genomes without the need for a reference or a genomic assembly [152, 236]. However, all of these research efforts have been performed with the human genome in mind and as such require dissection of their approaches from a crop-genome point of view to assess their suitability for whole-genome variant calling endeavours.

In this work such a dissection is performed. The following section introduces and discusses multi-reference alignment strategies and how plant genomic architecture may present some challenges to their direct application for crop genome variant finding. In section 3, cDBGs are introduced and discussed and a novel, local assembly algorithm proposed as an alternative for *de novo* genome variant calling.

Multi-reference alignment

Multi-reference alignment refers to the use of data structures to store genomic data for a set of individuals to which short-read sequencing data can be aligned [16, 261]. This is achieved using various graph based structures, where the nodes represent genomic sequences and the edges represent how the sequences connect to one another in the genome [261, 284]. A walk through the graph structure involves visiting nodes that are connected via the same edges. Performing such walks, and collecting the sequences as the walk progresses would enable reconstruction of the original genomic data. In this data structure, variants are stored a alternative walks known as paths [261, 284]. A walk along one of them will yield a different genomic sequence to a walk taken along another. An example of this can be seen in figure 4.1 where the reference sequence is coloured in blue, and known variants in orange. Nodes (rectangles) contain genomic sequence where edges (arrows) dictate the paths which can be taken.

However, given that the introduction of variants within the graph can cause an explosion

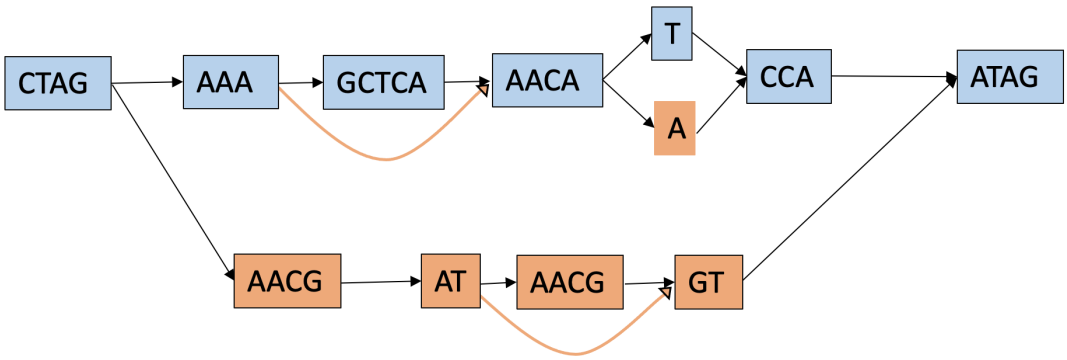
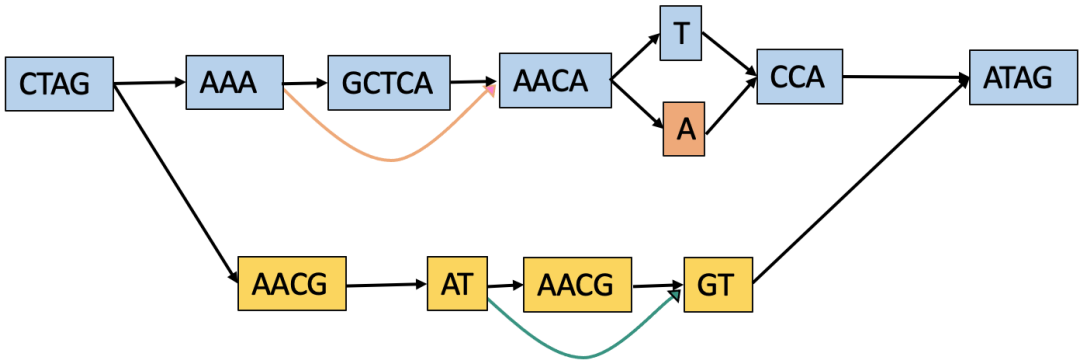


Figure 4.1: An example of a graphical representation of a single reference genome (blue nodes) augmented with known variant information (orange nodes).

of the number of paths within the graph, some multi-reference alignment tools employ haplowalks to connect variants together as previously seen within genomes [234, 316]. An example of how this works can be seen in figure 4.2 where the colours in the graph correspond to specific paths which output with the same coloured sequences. Sequences coloured black represent paths that are not supported by hapolyotype information and thus represents combinations of variants not observed in biological sequences. For example, a walk through the top nodes of the graph can multiple different combinations of nucleotides. However, the orange edge (which represents a deletion) and node (which represents a substitution) show that this deletion and substitution have been observed together and are supported by haplotype information. A path which does not include this deletion but does include the substitution has never been observed and so the resulting sequence is coloured black. Although the aim of these data structures is not to facilitate a walk but an alignment of read data, these haplowalks are still valuable layers of information for read alignment and variant calling as alignment can be restricted to paths encoded as haplowalks [316].

Alignment to these data structures is facilitated via a combination of distinct indexing and read alignment procedures. Multi-reference aligners, like their linear counterparts, follow



CTAGAAAGCTCAAACATCCACCA
CTAGAAAGCTCAAACAACCACCA
CTAGAAAAACAACCAATAG
CTAGAAAAACATCCAATAG
CTAGAACGATAACGGTATAG
CTAGAACGATGTATAG

Figure 4.2: An example of a graphical representation of a single reference genome (blue nodes) augmented with known variant and haplotype information (other colours). The sequences below the graph can be obtained by taking different paths in the graph. However, the haplotype information, encoded by colours, shows that not all possible combinations of nodes have been identified in the biological sequences to construct the graph. The sequences resulting from such paths are coloured black.

a either a hash-table-based indexing procedure or a burrows wheeler transform(BWT)-based index strategy with an accompanying FM-index [14]. These structures prevent a wholly unnecessary linear scan of a genome to identify where a read may belong. Instead, a hash-table-based strategy enables a membership check for the reads in constant time (that is, it doesn't depend on the size of the data structure, and thus the genome), denoted $\mathcal{O}(1)$ in computer science terminology. A BWT-based look up $\mathcal{O}(|P| + k)$ time to determine membership, where $|P|$ refers to the size of the query (e.g. read, or subsequence of the read) and k refers to the places where those subsequences can be found in in the FM-index (and

thus, the genome) [17].

Rapid querying is essential for keeping the time taken to perform a read-alignment to a minimum. This is critical for short-read sequencing data which features millions, if not billions, of individual reads to be aligned [334]. However speed is not the only issue to consider when determining which indexing structure to use. Size is also a major consideration, especially for large crop genomes. Hash-based strategies required the reference genome(s) to be fragmented and stored in a table-like structure where the keys are the sequences and the values are the positions where those sequences occur [284]. However, this data structures has a space complexity of $\Theta(nk)$ space, where n refers to the set of k-mers and k refers to the k-mer (fragment) size [66]. Precisely how many unique fragments exist for each genome will depend on a number of factors including the size of the fragment used and the repetitiveness of the genome either due to an expansion of repetitive elements or the presence of highly similar subgenomes.

Theoretically, the number of possible k-mers can theoretically grow at a rate of 4^k although for a large enough k-mer size, genomes contain only a small portion of all possible k-mers [54, 380]. Figure 4.3 shows an example of how the number of unique k-mers (fragments of length k) grows almost linearly with the the size of k-mer used for *Triticum aestivum* (supplementary method 1). It may be tempting to lower the k-mer value to reduce the resulting has table size and keep memory requirements low but this will result in an increasing number of reads not uniquely matching a position in the reference genome. This will result in a lower mapping quality score (MAPQ) which captures how confident an aligner is that it has placed the read in the correct place [56, 73]. Given that variant callers rely heavily on MAPQ scores, this could have detrimental effect on the ability to call variants with confidence. Some aligners are capable of local realignment and assembly of poorly aligned regions, but their performance has not been evaluated favourably, and there is a distinct lack of benchmarking for non-human species [73, 192].

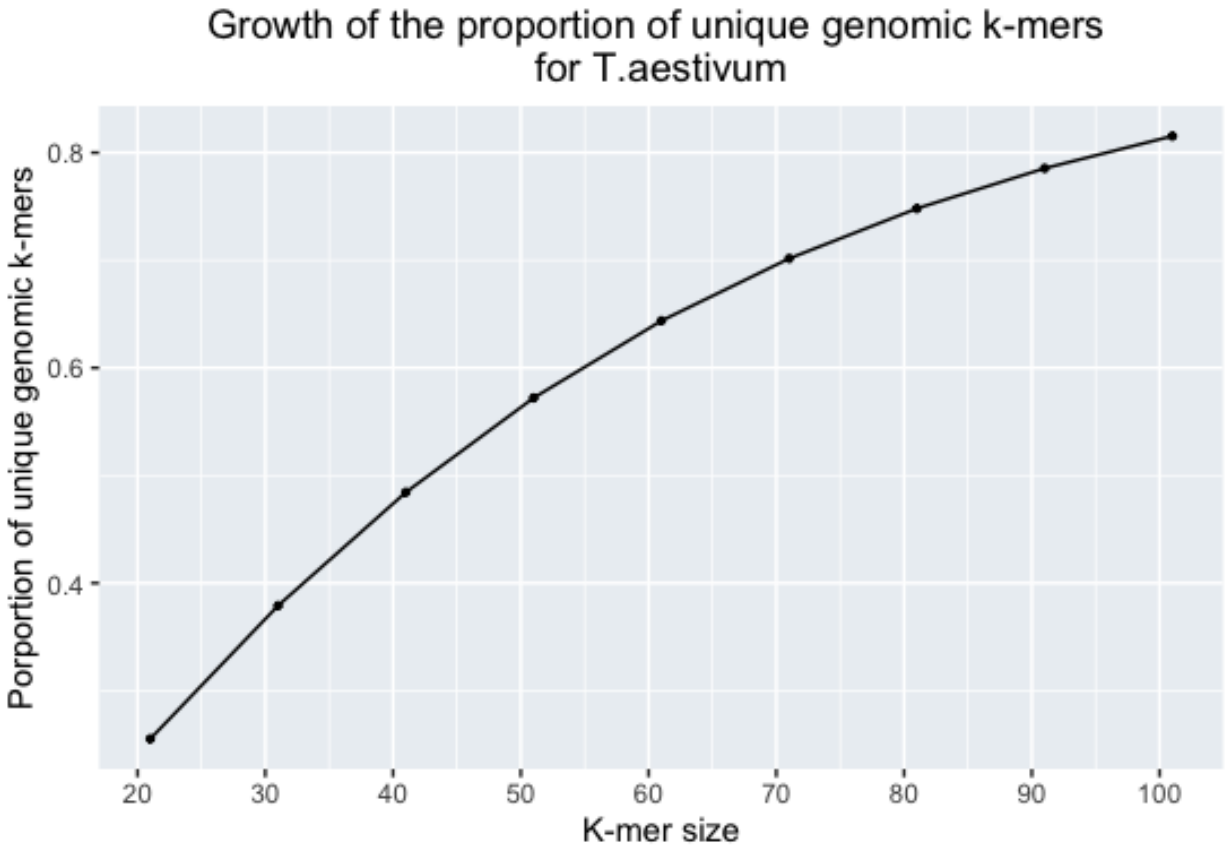


Figure 4.3: A linear relationship between k-mer size and proportion of unique genomic k-mer for *T. aestivum*.

The BWT-FM data structures however are highly succinct and are capable of compressing data close to its theoretically lower bound. Precisely how much space is required varies by BWT implementations with implementations achieving $\mathcal{O}(nlog\sigma)$ bits of space where n is the length of the genome and σ is the size of the alphabet, which for DNA is 4 (A,T,G,C) [110]. In practicality, this precise implementation required 4.84 bits per base stored, where the next-best implementation required 7.09 bits per base. For the 3.89Mbp rice (*Oryza sativa*) genome (given as median genome size by NCBI genomes), the smaller implementation would require 0.02355265Gb of space, where for the large hexaploid wheat (*T. aestivum*) genome (median 14444.8Mb) this would require 8.739104Gb [245, 246].

The differences between the two data structures in practice are captured by the amount of space used for storing 10000 human genomes by both GenomeGraph and vg. Where GenomeGraph 23.8Gb of space for the 10000 human genome project, vg took between 3.92 and 1.97Gb depending on the number of variants included [118, 284]. However, when the additional data structures required for working with the stored human genomes are considered, vg actually required 25-63Gb of space. This highlights that the underlying core data structure for storing the genomic data is not the only consideration that should be made when considering which tools to select based on hardware limitations.

It is also important to note that not all data structures are built equally. As described above, differing implementations of the BWT-FM structures can have differing memory requirements for storing the same information [110] and the same is true for hash tables. Where GenomeGraph required 23.8Gb of RAM for storing 1000 human genomes, early versions of vg which also implemented a hash table required over 300Gb to store the same data due to the extensive auxillery data structures it requires for read alignment [95, 284]. It can often be difficult to tease out of publications the precise time and space complexities of their underlying algorithms and the relationship between genome size and data structure space requirements is not always linear. As such it is not always possible to infer that a data structure that takes nGb of space for 10000 human genomes would require $n \times 5$Gb of space for a genome 5 times larger. However, size and uniqueness of a genome are major factors which can affect space complexity and these can be easily determined using k-mer counting software such as KMC3, Jellyfish or Khmer [179, 222, 380]. Other important factors include data quality, as erroneous base calls and introduce many unique k-mers within a sequencing, artificially inflating storage requirements [384].

Alignment-based Variant Discovery limitations While multi-reference genome alignments certainly have the potential to mitigate the reference bias induced via the use of a

single reference genome, there are still some outstanding limitations of any alignment-based approach. The first, and perhaps most significant for repeat-element-reach or polyploid genomes, is that a read alignment is limited by its uniqueness. That is, if a read aligns to multiple regions within the genome equally, or near-equally well, the aligner cannot confidently ascertain where it belongs (coordinates) or how it belongs (how it aligns). This is a significant limitation for all read-alignment strategies, which are especially pronounced for short-read data [60]. Research into this problem for short-read data is still active, with a novel solution being recently proposed [280]. Long read data also has the potential to mitigate these issues so long as the length is such that the read spans repetitive regions and features unique sequences on its 5' and 3' ends enabling the aligner to anchor it in the genome [14]. Recent research however has shown alignment in repetitive regions even with long reads, is still problematic [156].

Graph repetitiveness is also exacerbated by the inclusion of variant data, given that it results in more choice for the aligner [157, 176]. Multi-reference alignment accuracy has been demonstrated to decline with increasing levels of genetic variation [278]. This may be somewhat mitigated by the incorporation of haplotyping paths, which may guide the aligner to selecting paths only observed in biological sequences [234, 316]. Problematically, this in itself is a form of reference bias. Alternatively others have filtered the variants to include based on allelic frequencies and mathematical models based on variant scores [157, 176, 278]. Determining which variants to include and which to exclude to reduce reference-bias but not increase graph repetitiveness is very much an open problem.

An additional problem with short-read alignment-based approaches is with the variant calling procedures themselves. There is massive diversity within variant calling methods regarding the filters they apply to alignments, evidence they used to identify variants and statistical and mathematical strategies they implement to piece all off the evidence together to make the variant calls [27, 198, 218, 368, 373, 398]. Given such diversity, it is unsurprising

that numerous research has highlighted the lack of concordance in variant calls between methods and robustness to even the same input data [104, 173, 290, 358]. This is especially pronounced for structural variant calling, where precision is encouragingly reported to be between 70-90% but recall as low as 10-70 with a false positive rate as high as 89% in some instances [306, 373].

Given that variant calling pipelines, from sequencing, to alignment, to the variant calls themselves, are heterogeneous this lack of agreement is to be somewhat expected. To account for this, human genomics and bioinformatics experts have developed a stratified approach for benchmarking variant calling pipelines which better enables insight into the individual factors affecting variation in variant calls [183]. When applied, this revealed that variant callers are particularly prone to disagreeing about variants in regions with poor sequence coverage, extreme GC content variation and on multimapping reads [27]. However they identified that newer variant callers, especially those based on machine learning methods, are less sensitive to those factors and so make more robust variant calls [27]. Given that those particular machine learning-based variant callers were trained on mammalian data, caution should be used before applying them directly on plant genomes [107, 273].

While not all multi-reference graphs feature in built variant callers, vg and GraphGenome do and both show excellent performance [118, 284]. However, the benchmarking performed in these publications used only human data. In general, benchmarking of variant identification tools and pipelines is highly biased towards the human genome and much benchmarking work remains for plant, and especially complex crop genomes. This certainly offers an excellent opportunity for further study.

Alignment-based Variant Discovery Conclusion Multi-reference read alignment has been an excellent development for mitigating reference bias for variant discovery endeavours. The methods developed show excellent innovation and forethought for time and space

complexity. However, their ability to cope with the frequently large and complex genomes that exist in crop genomes has yet to be tested. Further, alignment-based variant calling still shows an alarming lack of concordance despite decades of research [104, 173, 290, 358]. Novel machine-learning methods for variant calling seem to show an improved ability to separate true variant signals from noise but given that they are trained on human data, may not be directly applicable to crop genomes [27].

In all, crop genomes would benefit from extensive benchmarking work to assess how novel multi-reference aligners cope with crop genomes. This would be well complimented by work benchmarking novel variant calling strategies which are seemingly better able to cope with noisy signals from problematic genomic regions. Given that multi-reference variant identification is a novel and burgeoning field, novel methods for variant identification using these structures are being released at a rapid rate (e.g. [126, 153, 196, 286, 377]). As such, now is an exciting time to begin to develop robust benchmarking strategies for multi-reference crop variomics analysis.

De novo multi-genome variant identification

The alternative to multi-reference alignment is a coloured de Bruijn graph (DBG) exploration of variants. Introduced by Iqbal [152] this variant finding strategy at its most basic is intuitive given a level of understanding of DBG structure.

A DBG, in its most basic form, is graph which represents k-mers and how they overlap each other. To construct such a graph (G), each sequence is decomposed into distinct k-mers which form nodes (also known as vertices (V)) in the graph which are then connected via edges (E) if they overlap by k-1 bases. The nodes and/or edges are then coloured by their presence in each genome. In theory if we then wanted to assemble a single genomes sequences, we could take a Eulerian walk in the path, where we walk all the edges with colours corresponding to that particular genome, piecing together the genome as we go [75].

An example of such a structure can be seen in figure 4.4 which shows how a single sequence (e.g. a genome) can be fragmented into its consistent k-mers and a represented as a DBG.

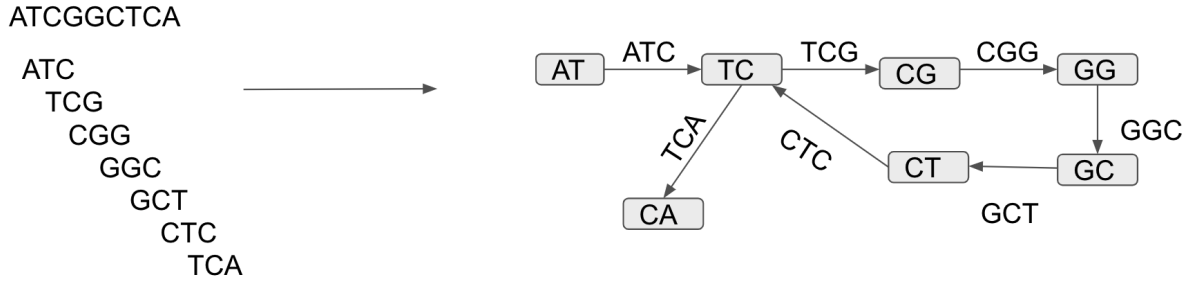


Figure 4.4: An example of how a sequence can be transformed in a de Bruijn graph, where nodes represent k-1-mers and edges represent k-mers.

A cDBG expands this structure further by incorporating sequences from more than one genome and giving them each a unique colour so it is possible to determine to which genome the k-mers belong. The same Eulerian walk can be taken through a cDBG with the exception that to reconstruct the original sequence, one would have to keep to a single coloured path [152]. Importantly, variants between such genomes in a cDBG will present themselves topologically and as such algorithms can be designed to find them. For example, a SNP will present itself as a bubble in the graph and as such a simple algorithm which detects branching paths in the graph (a node with >1 outgoing edge) will identify the start of the bubble, and the alternate paths can be followed until the algorithm finds the node where they join up again (a node with >1 incoming edges). This bubble finding algorithm was first described by [152] and later adopted by [236]. An example of this algorithm in action can be seen in figure 4.5. The initial branching node of such a structure is called a source and the node where the paths rejoin is called a sink.

Given such an intuitive way to represent variants between genomic sequences much effort has gone into the development of cDBG strategies capable of holding genomic data and their colour information for multiple genomes [9, 10, 144, 152, 170, 236]. Development

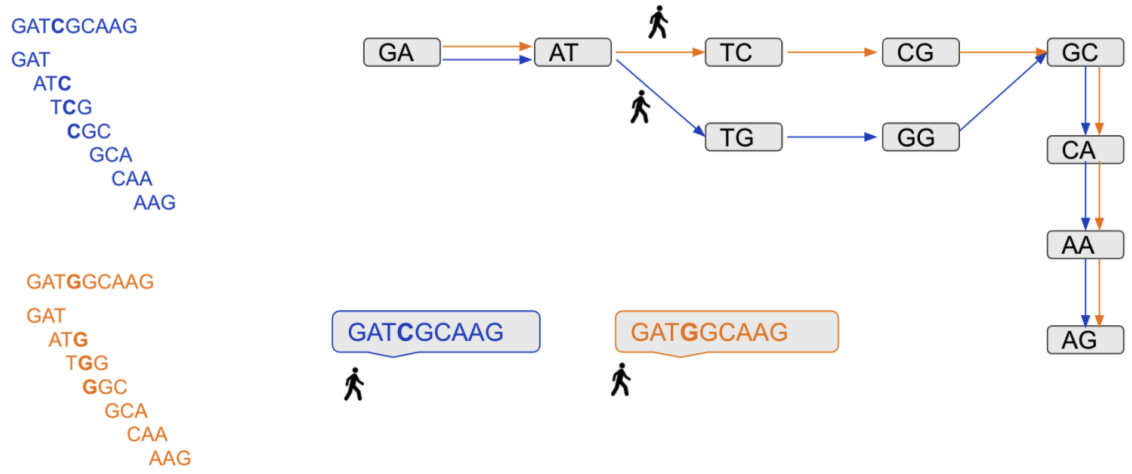


Figure 4.5: An example of coloured de Bruijn graph in which we can take two alternative paths in the graph to construct genetic variable sequences

and implementation of intelligent data structures for storing DBGs has been of paramount importance give the reliance of genome assembly on DBG structures. This saw computational requirements of DBG storage for a human genome fall from 336Gb for Abyss version 1 to 1.5Gb for an implementation developed by [67]. [236, 314]. Unfortunately, where a DBG grows linearly in size with the number of k-mers stored ($O(k)$), colouring the nodes requires magnitudes more space to store than the DBG structure itself and so cDBG implementations are far more computationally-resource heavy than underlying DBGs [10, 170].

For large plant genomes this is a major problem as there currently exists only a single cDBG-based variant finder, Vari, that has demonstrated itself as being capable of handling more than handful of mammalian genomes. In tests by [236], they were able to construct a cDBG and perform variant calling for 4 plant genomes (rice *Oryza sativa*, tomato *Solanum lycopersicum*, corn *Z.mays* and *Arabidopsis thaliana*). However, while Vari reported an impressive peak memory (RAM) usage of <4Gb for the plant genomes, [10] found that Vari requires an extensive amount of external disk space to construct an initial uncompressed DBG and colour matrix. For a single human transcriptomic data set this required terabytes

of disk space [10]. This was identified to be the result of a high number of colours relative to the number of k-mers which indicates many k-mers are sample-specific [10]. This highlights Vari's inability to cope memory-wise with a highly diverse cDBG [10].

For crop genomes, the initial construction of an uncompressed DBG and colour matrix could be very problematic. While crops are considered to have reduced genetic diversity from domestication and intensive selection, on a genomic level pan-genomes show high levels of genetic diversity which could impact the resources required for cDBG construction [116, 269].For example, the 3,000 rice genomes project uncovered millions of small variants and nearly 100,000 structural variants [354]. 721 Maize genomes showed a highly flexible core pan-genome space with almost half of the genes being dispensable and hundreds of thousands of structural variants [128], and 18 wheat cultivars showed nearly 60% of it's pangenome gene space is variable between cultivars alongside 36 million intervarietal SNPs [235]. As such available hardware limitations are going to be a major factor for implementing any cDBG software which requires uncompressed building of its colour matrix.

Other cDBG implementations have been developed since Varis release, most notably Cuttlefish, Rainbowfish (implemented in Mantis) and Bifrost [144, 174, 254, 395]. However, cuttlefish is designed for a collection of reference genomes (like a multi-reference graph) and both Mantis and BiFrost are not constructed as cDBGs for variant calling but rather for efficient querying of the graph (i.e. is a given sequence present in the graph and if so, who does it belong to?) [144, 174, 254, 395]. BiFrost can be traversed via an c++ API but no variant caller has yet been developed for this.

Metagraph

However, a novel approach to cDBG construction may yet make large-scale cDBG-based variant finding for multiple, diverse, crop genomes a reality; Metagraph [170]. Metagraph is a highly flexible cDBG tool, capable of implementing a range of cDBG construction techniques

including the highly compressed colour matrix technique developed by [10]. Through these methods, Metagraph was able to build and store a DBG of every single sequenced plant genome deposited in the sequence read archive (SRA) using 602GB of space, where the original data has a size of 576 terrabytes. While this doesn't include colour information, other species' cDBGs show an impressive level of compression. The entire 81Tb fungal SRA data set when ran through Metagraph took only 82Gb of space for the DBG and 501Gb for the corresponding colour annotations [170]. Further, Metagraph is built with the ability to work, where possible, within given RAM constraints, with temporary data able to written and read to disk (Karasikov, personal communication viewable at `https://github.com/ratschlab/metagraph/issues/407`). It is additionally able to encode genome coordinates and as such a reference genome could be used to place any identified variants within its coordinate space [170, 171].

A major roadblock to using Metagraph for variant calling is that Metagraph does not feature a variant caller. In its place however, is a differential assembly algorithm. Like the local assembly implemented in some variant callers, differential assembly assembles only certain portions of k-mers belonging to the genomic data as a whole [170, 274, 347].

For metagraph, differential assembly works by first marking k-mers present in a given list of genomes and absent in the others, both of which are specified via a JSON file format. After this, the k-mers are extracted and assembled (figure 4.6). It is important to note that metagraph can either assemble in unitig or contig mode. In unitig mode, Metagraph assembles only "simple" paths, which are paths that contain nodes with only a single incoming and outgoing edge and so can be unambiguously assembled [59]. Should metagraph come across a complex node (> 1 incoming or outgoing node), Metagraph will halt traversal and simply output the complex nodes sequence [170]. In contig mode however, Metagraph will assemble complex nodes by traversing a path until there are not further outgoing nodes or they have all been visited [170].

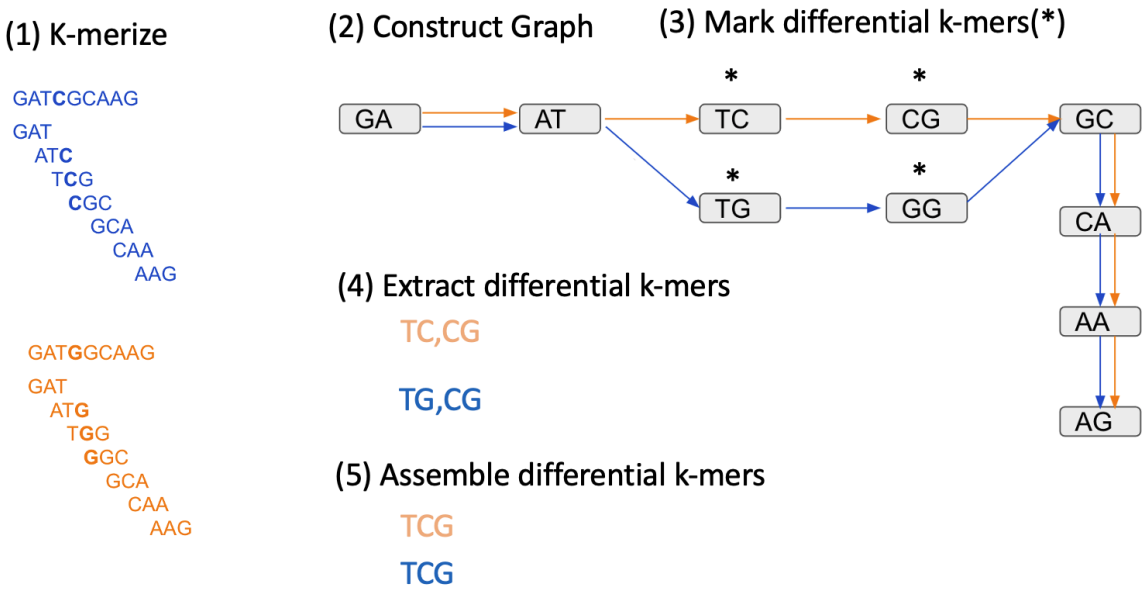


Figure 4.6: An example of how Metagraphs [170] differential assembly strategy works. First, distinct genomes are k-merized (1) and a cDBG is constructed, recording for each k-mer its genome of origin (2). Differential k-mers are then marked (3), extracted (4) and assembled (5) to give differently assembled fragments.

However, these differential assembled contigs contain no further information about their origin in the genome or the variant type they contain. As such in the following section we introduce an algorithm to anchor the k-mers in a given reference genome and assess its potential to be applied to the problem of determining the genomic origin of differentially assembled data.

MetaGraph Variant Calling

Extending differentially assembled k-mers to place them in the reference genome takes advantage of the fact that while the k-mers contained within the sequences are differential (i.e they exist in one genome but not another), their flanking sequences are not differential and so exist in both genomes. As such, simply extending the k-mers by a single k-mer to the left and right and querying for those sequences will return a position within the reference

genome given three assumptions:

1. The differentially assembled sequences are completely assembled such that an extension to the left or the right of the sequences will yield a call in the reference genome (i.e. no differential k-mers are left assembled) and their length is thus reflective off true variant size

2. The differentially assembled sequences are correctly assembled such that their first and last k-mer are correct and so any extensions of those k-mers will be correct

3. The differentially assembled k-mers do not originate from the ends of the reference sequences (i.e. the first k-mer does not originate from position 0 in the reference and the last k-mer does not originate from position n in the reference, where n is the size of the reference sequence) and as such the sequences can be anchored given a k-mer extension strategy

Given these assumptions, the extension algorithm is proposed in figure 4.7 and the procedure visualised in figure 4.8.

First, each differentially assembled sequence is extended with all possible extension k-mer extensions, followed by a querying of the graph with appended extended k-mers which enables anchoring in the reference graph. The coordinate information returned from the querying procedure is then used to determine variant size and type (e.g. equal substitution, insertion, inversion).

Proposed Variant Caller However, there are number of problems associated with this approach which can be categorised as either assembly-based or graph-based.The assembly-based issues can be summarised as so; the differentially assembled fragments, and thus any downstream variant calls, are as good as the path complexity in the graph and the assembly algorithm used. For simple paths in the graph (i.e. 1 incoming and outgoing edge per

Algorithm 1 Determine variant location and type in metagraph

Require: Assembled contigs
  for each contig **do** extend first and last k-mer
    for extended contig **do** query metagraph for coordinates
      for each returned set of coordinates **do**
        left ← coordinate[0]
        right ← coordinate[1]
        if left is not empty or right is not empty **then**
          location ← left:right
          coordinate node span ← $|right - left|$
          Contig length ← L
          Contig node span ← L-(k-1)+1
          difference ← $|coordinatenodespan - Contignodespan|$
          Inversion ← False
          if $right > left$ **then** Inversion ← True
          end if
          if $difference = 0$ **then** Variant type ← substitution
          else if $difference < 0$ **then** Variant type ← deletion
          else if $difference > 0$ **then** Variant type ← insertion
          end if **return** location, difference, variant type, inversion
        end if
      end for
    end for

Figure 4.7: The proposed MetaGraph-based variant calling algorithm

node) the assemblies will accurately reflect the sequence from where the k-mers originated as there is only a single, unambiguous way to assembled them. For complex nodes however, the assembly will be as good as the heuristic strategies used to assemble them. Given that sequence assembly is an NP-hard problem there are multiple heuristics implemented by assembly algorithms to make assembling a genome possible in polynomial time and as such the resulting assemblies can differ between tools [147, 282, 359]. This is demonstrated in figure 4.9 which shows the multiple ways a set of complex nodes can be traversed, all

which result in different assemblies of the underlying sequence. The exact structure this figure 4.9 is a superbubble, which are likely very common in variable, repetitive, and/or polyploid genomes given that such structures are induced by such sequences [48, 340]. As such, these hard-to-navigate topologies are likely to feature heavily in cDBG-based variant finding efforts for crop genomes.

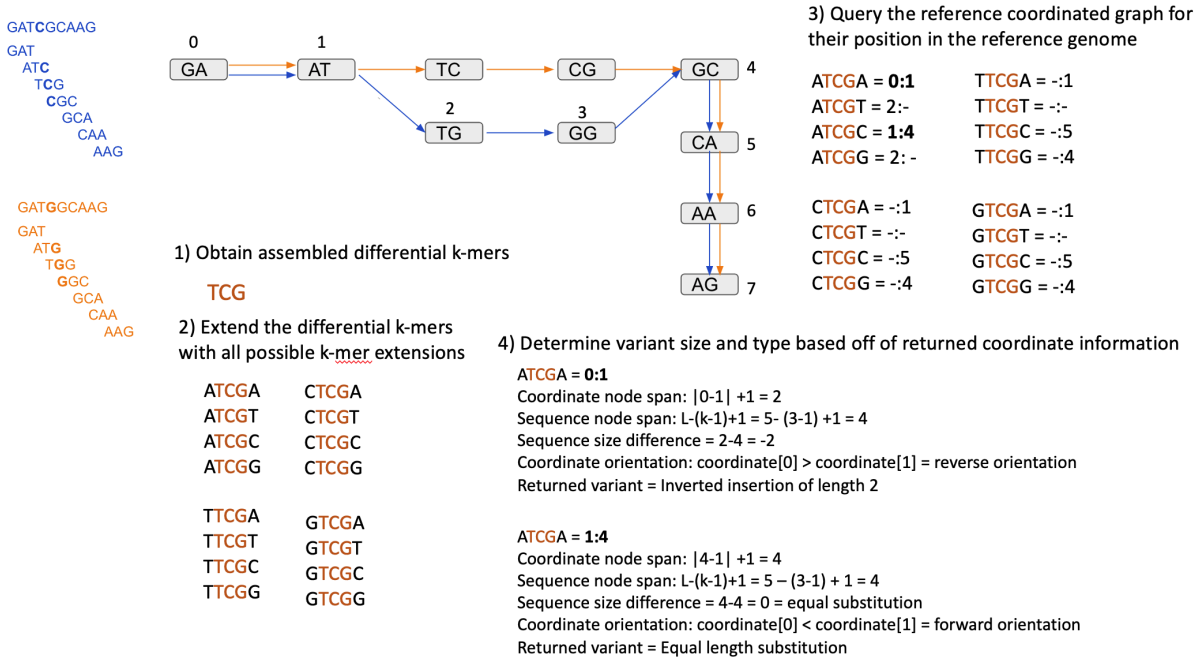


Figure 4.8: A walk through of the MetaGraph variant calling algorithm. The first step involves obtaining the differential k-mers (1) and extending each k-mer with all possible k-mer extensions to the left and right (2). The MetaGraph graph annotated with the reference genomes coordinates is then queried for every sequence (3). If query returns coordinates which match the left and right k-mer extensions their size and coordinate orientation is sued to determine variant size and type (4)

Unfortunately, the problem of complex node assembly is not easily solved here as there is no "best" assembler. Sequencing data is heterogeneous in quality and length and crop genomes contain a diverse array of genomic architectures, all of which have the potential to impact assembly performance[351, 359]. Large-scale benchmarkings of assembly algorithms have been performed, but this was for simulated and vertebrate species [47, 93]. A recent

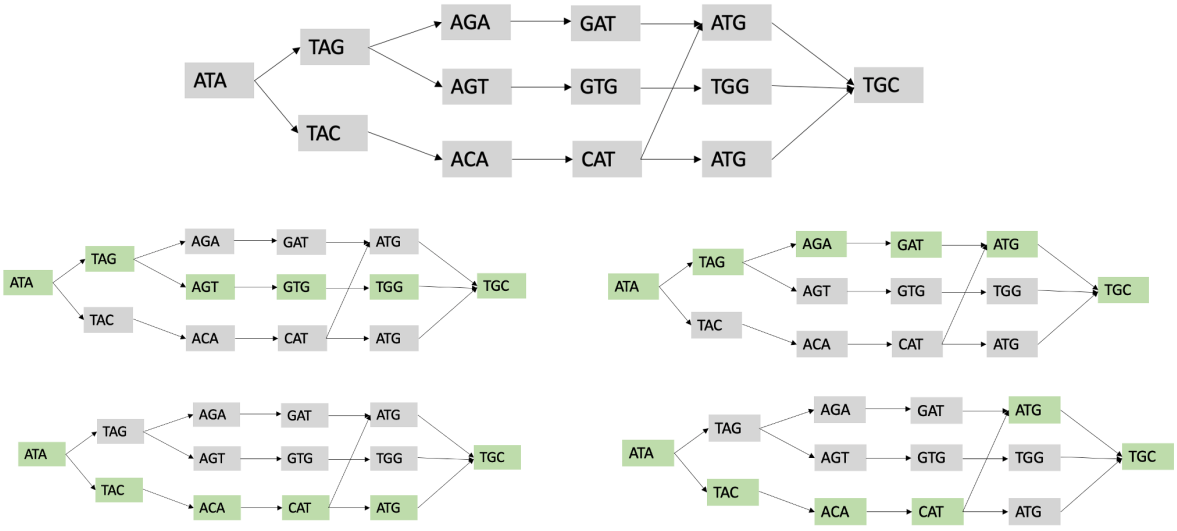


Figure 4.9: An example of an ambiguous topological structure in a de Bruijn graph which encode numerous, equally valid paths

recommendation for benchmarking assembly algorithms in plant genomes has been made [279]. However, many benchmarking assessments and assessments of assembly quality were assessed by their ability to produce contiguous rather than complete or correct assemblies [332]. As such no recommendations can be made here regarding assembly strategy to implement for the differentially assembled fragments except to recommend the testing of assembly strategies that can incorporate long-range information. Such information can aid assemblers to determine the correct paths through complex regions and piece sequences together contiguously [230]. cDBG variant calling algorithms do exist that can make use of such information (e.g. [335]), however at present MetaGraph is not one of them [170].

To take advantage of assembly algorithms capable of using long-range information (e.g. Abyss2.0 [155], MaSuRCA [396]) it would be necessary to extract the reads where the differential k-mers originate. However, this is not straightforward, with three potential options. The first is to use MetaGraph's annotation pipeline to annotate the de Bruijn

graphs with their read and k-mer ID and then for each differential k-mer, query the graph for that information and extract it from the read data. The second is to build a new DBG from the differential k-mers and query this graph with the read data directly and then extract the reads which show hits. The final, faster and less-memory intensive option is to build a bloom filter for the differential k-mers and then query the bloom filter structure for their k-mer membership in the differential k-mer set. However this method has the distinct drawback in that a bloom filter is a probabilistic data structure and so is prone to false positives, the rate of which can be tuned as a trade off against data structure size [262].

However even with the use of long-range information, genome assembly with short-read data is problematic especially for repetitive regions. [351] identified that almost 30% of a high-coverage regions for the tomato genome (indicating repeat sequences) contained misassemblies. The current consensus in the literature is that long-read sequencing data (or a hybrid of long-read and short-read data) accompanied by specialist assembly algorithms are required for better assembling data from repetitive genomic regions. [130, 134, 181]. As such accurate variant identification from these regions may require such data, although it is important to note that long reads also have limitations for repeat-resolution [203].

The graph-based issues encompass the multi-node matches of the extended k-mer set and a susceptibility to splitting variants. The multi-node matches of extended k-mers and is illustrated in figure 4.8. In this problem the extension of the differentially assembled k-mers matches more than a single pair of nodes in the graph and as such, differing types and sizes of variants can be called from the same sequence. This problem is analogous to the multi-mapping problem in read alignment and can be somewhat mitigated by using larger k-mer sizes, which tend to capture more unique sequences and thus will reduce the spurious node-matching problem. Figure 4.3 illustrates this relationship between k-mer size and uniqueness. Using large k-mers will also aid in determining the subgenome of origin of the assembled differential k-mers as the subgenomes share fewer k-mers in common as k-mer

size grow large and the k-mers become more specific in sequence content. This is illustrated in figure 2.5. However, should variants ever be called with this strategy, some form anchoring score should be implemented which captures the inability to uniquely assign and thus call the variant size and structure in much the same way as MAPQ scores are used in alignment [56, 73].

The splitting variant problem is induced when a variant contains a k-mer already present in the graph. This is illustrated in figure 4.10 and shows an inversion in the original underlying sequences. However, this variant introduce two distinct variant topologies (two distinct source and sink nodes) in the graph because the "GG" sequence is present in both the original and inverse sequence. Such sequences would then be called as separate variants rather than the single variant event they are.

A solution to this problem may be to take a longer walk through the assembly structure when multiple variants are identified within close proximity to each other. In this strategy, the walk would begin a number of base pairs upstream of the first variant and ends a number of base pairs downstream of the last variant. Using this strategy, the inverted sequence would be correctly identified from figure 4.8. A drawback to this approach is the extra time required to re-walk paths, but this may be a fair trade-off given the potential to reduce incorrect variant calls.

An addition solution to the variant splitting and multi-node matching problem is to involve long-range data, either post processing, or in the graph structure [230].For the multi-node matching solution, identification of reads mapping to the left and right flanking sequences would allow an extension past the single k-mer extension provided by the algorithm. This extra sequence data may aid in more firmly anchoring the sequence against a reference genome. For the split-variants problem the long range information could aid in joining differentially assembled sequences together that exist in close proximity. However this would require careful consideration regarding how data structures to hold this long-range

information could be implemented in a space and time efficient manner, given how massive WGS can be.

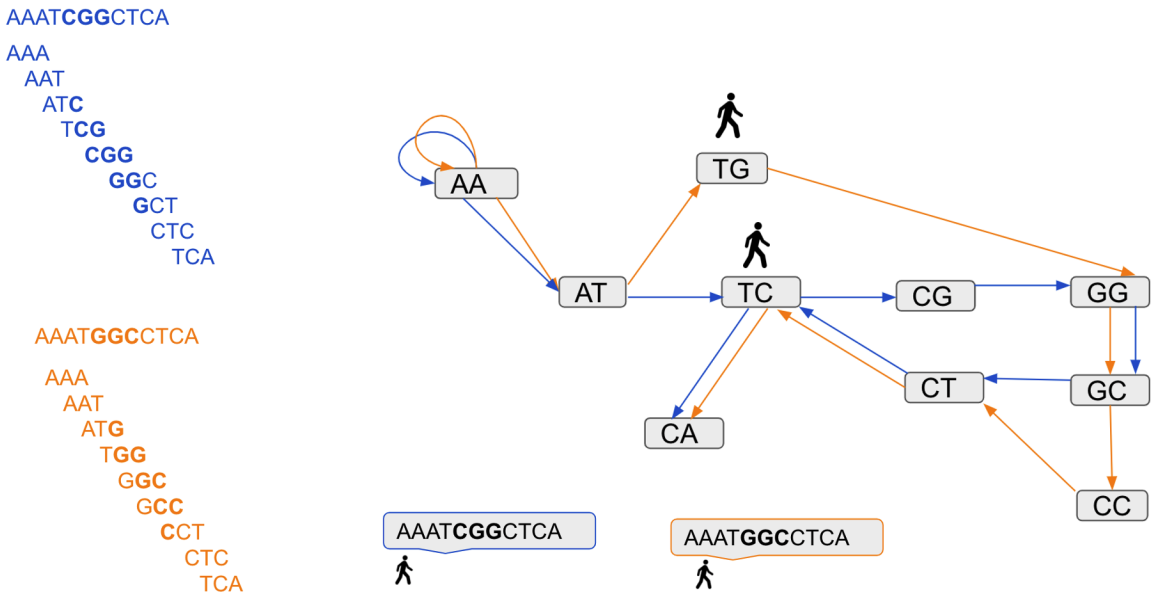


Figure 4.10: a cDBG demonstrating how certain variants can induced a split-variant topology in the graph which, if called using the algorithm in figure 4.7 would result in two, incorrect, variant calls

The two final issues with this algorithm cannot be considered graph or assembly-based in nature, but are simply inherent given given the set up of the proposed variant calling approach. The first is that any variants involving the terminal (first and last) k-mers of a reference sequence will be unable to be correctly anchored given they have no true k-mer extension. In theory, the inability to anchor the left or right extended k-mers is a signature of the differential k-mers originating from such events. However, it would not be possible to tell this event apart from a failed k-mer extension matches due to misassembly events with the algorithm in its current form. Read support may help confirm the terminal nature of such variants but as previously discussed this would require careful engineering to extract the reads from where differential k-mers originate.

The final problem is that while the algorithm can theoretically identify deletions (figure 4.7) by detecting that the coordinates of the reference genome are further apart than the differentially assembled sequence is long, in practice this approach may actually be blind to deletion. This inability to detect deletions is illustrated in figure 4.11 which shows the deletion of a "G" base in the reference (blue)genome, which does not result in any differential k-mers in the orange sequence. As such identifying deletions in this case would involve also calling for differential k-mers in the reference genome and identifying the deleted sequence in the query genome as an insertion in the reference genome. Additionally, a deletion has the potential to introduce novel k-mers via the fusion of sequences that are not joined in the reference. These have the potential to be called as variant k-mers but are not reflective of the underlying mutational event. Much theoretical exploration of how different variants are capable of presenting themselves in the graph is required.

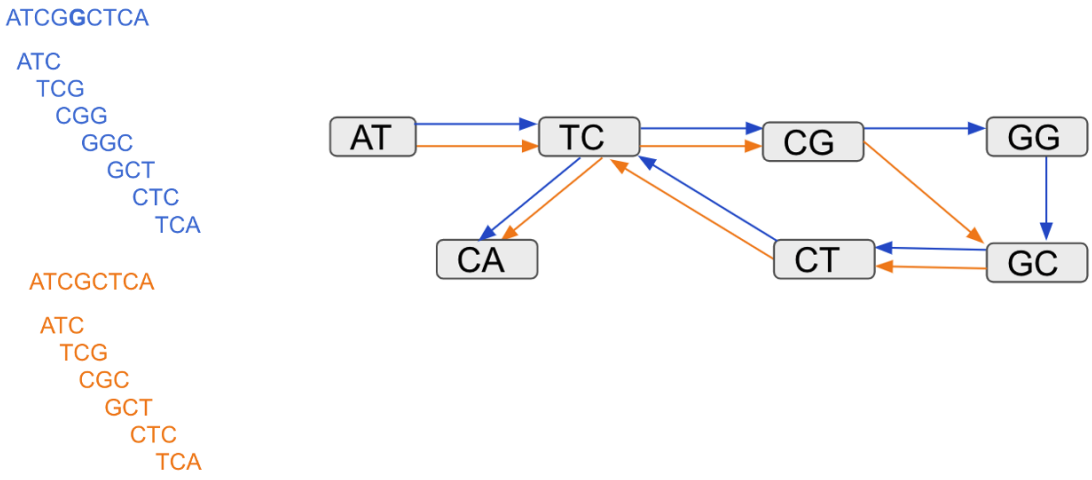


Figure 4.11: A demonstration of how deletions can be hidden in the graph structure through not creating differential k-mers.

The algorithm in its current form is also unable to detect copy number variations

(CNVs) due to the fact no weighting scheme is considered in this strategy. However, MetaGraph is capable of annotating nodes with their weights which can be returned when the k-mer is queried. This could be useful for differential sequences as it could help "uncollapse" repeat k-mers which may not have been assembled correctly. This could also be used to detect CNVs of sequences contained in all genomes (non-differential). However, detecting all CNV's via exploration of all differentially weighted k-mers could be problematic given that metagraph doesn't yet allow for directly identifying shared k-mers with different weights. Theoretically, extracting and querying shared k-mers against an graph annotated with the k-mers from differing cultivars of interest would enable the identification of k-mer-based CNVs. However this is complicated by variation if coverage induced by the sequencing process than a biological variant and the fact that the k-mers exhibiting differing copy numbers would have to be assembled to give them some genomic context [96, 295]. Overcoming these problems and implementing a variant finding strategy incorporating node weights will be essential for genotyping variants (e.g. heterozygous or homozygous) as implemented in Cortex [152].

De novo multi-genome variant identification conclusions

In all, the implementation of variant calling using cDBG-based strategies for crop genomes has yet to be formally established. On one hand, Vari presents a cDBG strategy capable of calling variants for a number of plant genomes without the need for any further algorithm development but is limited by its large computational overhead. On the other hand MetaGraph presents the most computationally efficient solution for cDBG construction and identification of differential sequences. However this solution, as illustrated in the points above, requires much more investigation and development before it could be confidently applied to variant calling for crop genomes.

Multi-reference graphs: Just add crops Overall it appears the multi-reference read alignment approach is the most "ready-to-go" for application to crop genomes for variant

detection of short read data. Multi-reference graph structures have received, and continue to receive, considerable interest in the areas of algorithm development and application to biomedical problems which paints a promising picture for their future in crop genomes. While the biomedical and crop genomic domains differ in their precise area of interest, the underlying problem of variant identification is largely the same. The only caveats are ensuring that underlying data structures and algorithms are directly applicable to crop genomic architectures. For data structures this concerns their ability to hold massive and diverse crop genomes in a space efficient manner and that their indexing structures enable rapid look up. For algorithms this concerns ensuring they are efficient given run time limitations and that they don't rely on assumptions about the genomes which may not hold true for crop genomes (e.g. ploidy). It is also worth bearing in mind that any variant finding algorithms (multi-reference or otherwise) that are based on machine learning have likely be trained on human data and so may not be directly be applicable to crop genome variant calling.

cDBG variant calling: Under construction Unlike multi-reference graphs cDBG-based variant callers for crop genomes are not really in a "plug-and-play" position. [236] presents an out-of-the-box option, but given hardware limitations it may not be possible to apply to a large number of diverse crop genomes. MetaGraph [170] is the most promising in terms of space efficiency but requires much more development as a variant caller. In this work we have addressed this by developing an algorithm for variant calling however we have also identified several major issues with it. Where possible, solutions to the problems introduced have been provided and we hope this lays the groundwork for any future developments for MetaGraphs implementation in variant identification work.

Where they both loose Both strategies have various strengths and weaknesses, especially given space and time complexities, however one region where they both fail is

repeats. Reference-based approaches (multi-reference or otherwise) fail in these regions as they are unable to confidently place the read given that it maps equally well, or near-equally well to multiple regions. cDBG-based approaches fail here for the same reason DBGs are known to collapse repetitive regions in the genome; they induce complex topologies (e.g. superbubbles, ultrabubbles, and complex bulges) which cannot be deterministically navigated [119, 212, 251, 260, 297]. This is problematic as it essentially leaves genomes with a high number of repeat elements or high subgenomic similarity largely in the dark. Repeat elements are known to harbour numerous sequences of functional importance, and a failure to determine the subgenomic location of variants will significantly hamper efforts to place variants within their genomic context, which will further hinder functional-genomic studies [18, 37, 162, 164, 189, 227].

Graph-based variant identification, either via alignment or DBG approaches have been already been acknowledged as the future of variant calling in crop genomes, but significant challenges still remain for their routine implementation [33].

Integration into crop breeding programs Of course, the identification of variants within crop genomes has limited use by itself. This variant data then needs to be placed within biological context (i.e., what impact, if any, do they have on phenotypic traits of interest). For a more abstract dive into variant impact one could construct a crop fitness map which seeks to identify which regions of the genome are under selection and as such may be regions of functional interest [162]. This approach could then be used to filter variants by locations of functional importance. However, construction of this map does require a significant amount of additional functional data and currently has only been produced for rice, *Oryza sativa* and human *Homo sapiens* [162, 164].

For a more direct application of variant data on crop breeding efforts, WGS-identified variant data could be linked to phenotype data through strategies such as GWAS. Existing

GWAS methods can be directly applied to non-SNP data with minor formatting requirements [195]. Variant data and/or candidate loci identified through GWAS can then be integrated into current genomic-selection-based predictions for crop improvement [175, 237, 248, 287].

Conclusion

It is evident, from this comprehensive discussion of modern variant calling approaches, that the only strategy that is ready for direct application to crop variomics is multi-reference genome alignment. However within this work, a novel variant calling algorithm is presented which may be applied to facilitate variant calling using an efficient cDBG approach as implemented by MetaGraph. This requires significantly more research before it is ready to be implemented formally, but it represents an exciting avenue for further exploration.

It is also clear that neither cDBG or multi-reference variant calling strategies are capable of resolving repeat sequences with additional long-read data. The inability or reference-based alignment analysis and de Bruijn graph traversals to correctly navigate repetitive sequences is well document and seemingly only addressable via additional sequencing data.

As sequencing technology continues to improve, and becomes cheaper to implement, hopefully the resolution of these sequences will be possible and variant calling strategies will be finally be able to shine a light on the dark matter of the genome. Integration of this information into the statistical strategies that underpin efforts to understand genotype-phenotype relationships and subsequent genomic selection strategies has the potential to greatly improve their predictive powers.

Chapter Discussion

This chapter has taken a deep dive into the application of modern variant-find approaches for crop genomes. In general, crop genomes are generally neglected when it comes to benchmarking bioinformatics algorithms and so the fact that Vari was tested using

4 plant genomes was highly encouraging [236]. It is evident from this work that crop genomics is crying out for robust variant calling pipelines and procedures which can only be achieved via in depth characterisation of the impact of errors and biases introduced at every level of variant calling pipeline. This also includes a characterisation of the impact of crop genome architecture on variant calling success.

Given that numerous multi-reference alignment approaches are ready for use this approach represents the most promising for reference-bias-free variant calling of crop genomes in the near future. However it is likely the implementation of these methods will incur significant computational overhead which may be a barrier given hardware limitations. Further, the problem of which variants to use to augment the reference genome is unsolved. Inclusion of all variants can massively increase space requirements, increase multi-mapping and reduce variant calling accuracy [157, 176, 278]. However, including few variants will minimise the number of variant sites the reads can align to which may result in fewer mapped reads. An additional consideration is whether or not to include HaploWalks which do also incur a further space requirement but can assist in cases where a read aligns to multiple variant paths by selecting the path that has been observed in biological sequences [234, 316]. Although this in itself is introducing a form of reference bias, it could help with the problematic explosion of potential paths in the graph that are induced through the augmentation of the graph with an increasing number of variants [234].

Given the complexity of these considerations its likely best to make the construction of a species multi-reference genome a team effort headed by experts in that particular crop and verified via comprehensive, stratified benchmarks. This would not only provide a valuable community resource but ensure that variant calling efforts for novel sequence genomes are comparable given they have gone through the same variant calling pipeline. This strategy has previously been suggested for human data and also seems sensible for crop genome applications too [261]. Indeed, a human pan-genome project is already underway to ensure

that a multi-reference structure represents human global diversity [353].

Unfortunatly, cDBG-based variant calling is not likely to see any kind of large-scale implementation for crop genomes in the near future. This is primarily due to the space and time limitations of current approaches, coupled with the lack of an experimentally-verified variant calling algorithm for the most efficient MetaGraph approach. In this work, a Metagraph-based variant finding algorithm is proposed and its shortcoming critically discussed. The hope is that this algorithm will lay the groundwork for future cDBG-based variant calling endeavours which can then be benchmarked against their multi-reference variant-calling counterparts.

Chapter Conclusion

Now is a very exciting time to be working in crop genomics given the huge developments in variant identification strategies which can minimise reference-bias and facilitate *de novo* variant identification without genome assembly. However, these approaches need benchmarking for community guidance and future algorithm development and given the pressing need for rapid variant identification for crop breeding efforts. Consideration should be paid towards making sure that any developed cDBG or multi-reference graphs are freely available to the scientific communities to prevent redundant construction of massive structures and to enable direct comparison of variant calls across cultivars.

CONCLUSION

The work presented herein represents years of efforts to make variant calling in crop genomes more efficient, more scalable and less biased towards protein-coding regions, anchored sequences, and single reference genomes. In chapter one, the subgenome-differentiation problem was re-imagined as a metagenomic problem which enabled the hijacking of efficient metagenomic approaches for rapid sequence comparison. This unusual approach has given novel insight into the sequences that drive subgenome sequence differentiation and how that can be used understand the evolutionary history of polyploid genomes and their progenitors. In chapter two, the development of a classification algorithm using subgenomic k-mer profiles was demonstrated to show excellent performance for previously unassigned contigs, making a greater proportion of the genome accessible for downstream variant analysis. In chapter three, an exploration of how modern variant finding methods may be applied to crop genomes given their various genomic sizes and complexities was performed. This work revealed that multi-reference variant callers are ready and waiting for plant variome exploration and as such represents an excellent opportunity for future study. Coloured de Bruijn graph (cDBG) approaches are not ready for implementation for crop genome variomics, but they present numerous opportunities for bioinformaticians to continue their data structure and algorithm development. We contributed to this effort via the proposal and critical analysis of a novel cDBG variant-finding algorithm, with the hope that this lays the foundation for future developments in efficient cDBG-based variant calling. In all it has been an honour to work in crop variomics research, and I hope the work presented here continues to assist agricultural genomics effort to better characterise the variation present within and between crop genomes.

REFERENCES CITED

[1] Genomics and our future food security. *Nature Genetics*, 51(2):197–197, 2019.

[2] Kaleb Abram, Zulema Udaondo, Carissa Bleker, Visanu Wanchai, Trudy M Wassenaar, Michael S Robeson, and David W Ussery. Mash-based analyses of escherichia coli genomes reveal 14 distinct phylogroups. *Communications biology*, 4(1):1–12, 2021.

[3] Nikolai M Adamski, James Simmonds, Jemima F Brinton, Anna E Backhaus, Yi Chen, Mark Smedley, Sadiye Hayta, Tobin Florio, Pamela Crane, Peter Scott, et al. Ectopic expression of triticum polonicum vrt-a2 underlies elongated glumes and grains in hexaploid wheat in a dosage-dependent manner. *The Plant Cell*, 2021.

[4] J Arvid Ågren and Andrew G Clark. Selfish genetic elements. *PLoS genetics*, 14(11):e1007700, 2018.

[5] Guillermin Agüero-Chapin, Deborah Galpert, Reinaldo Molina-Ruiz, Evys Ancede-Gallardo, Gisselle Pérez-Machado, Gustavo A De la Riva, and Agostinho Antunes. Graph theory-based sequence descriptors as remote homology predictors. *Biomolecules*, 10(1):26, 2019.

[6] Syed F Ahmad, Maryam Jehangir, Adauto L Cardoso, Ivan R Wolf, Vladimir P Margarido, Diogo C Cabral-de Mello, Rachel O'Neill, Guilherme T Valente, and Cesar Martins. B chromosomes of multiple species have intense evolutionary dynamics and accumulated genes related to important biological processes. *BMC genomics*, 21(1):1–25, 2020.

[7] Syed Farhan Ahmad and Cesar Martins. The modern view of b chromosomes under the impact of high scale omics analyses. *Cells*, 8(2):156, 2019.

[8] Alekseyzimin. alekseyzimin/masurca.

[9] Fatemeh Almodaresi, Prashant Pandey, Michael Ferdman, Rob Johnson, and Rob Patro. An efficient, scalable, and exact representation of high-dimensional color information enabled using de bruijn graph search. *Journal of Computational Biology*, 27(4):485–499, 2020.

[10] Fatemeh Almodaresi, Prashant Pandey, and Rob Patro. Rainbowfish: a succinct colored de bruijn graph representation. In *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[11] Michael Alonge, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J Sedlazeck, Zachary B Lippman, and Michael C Schatz. Ragoo: fast and accurate reference-guided scaffolding of draft genomes. *Genome biology*, 20(1):1–17, 2019.

[12] Michael Alonge, Xingang Wang, Matthias Benoit, Sebastian Soyk, Lara Pereira, Lei Zhang, Hamsini Suresh, Srividya Ramakrishnan, Florian Maumus, Danielle Ciren, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, 182(1):145–161, 2020.

[13] Saleh Alseekh, Federico Scossa, and Alisdair R Fernie. Mobile transposable elements shape plant genome diversity. *Trends in Plant Science*, 25(11):1062–1064, 2020.

[14] Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D Singer, et al. Technology dictates algorithms: recent developments in read alignment. *Genome biology*, 22(1):1–34, 2021.

[15] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[16] Adam Ameur. Goodbye reference, hello genome graphs. *Nature biotechnology*, 37(8):866–868, 2019.

[17] Tim Anderson and Travis J Wheeler. An optimized fm-index library for nucleotide and amino acid search. *Algorithms for Molecular Biology*, 16(1):1–13, 2021.

[18] Federico D Ariel and Pablo A Manavella. When junk dna turns functional: transposon-derived non-coding rnas in plants. *Journal of Experimental Botany*, 72(11):4132–4143, 2021.

[19] Naveen Kumar Arora. Impact of climate change on agriculture production and its sustainable solutions, 2019.

[20] Laila Sara Arroyo Mühr, Camilla Lagheden, Sadaf Sakina Hassan, Sara Nordqvist Kleppe, Emilie Hultin, and Joakim Dillner. De novo sequence assembly requires bioinformatic checking of chimeric sequences. *Plos one*, 15(8):e0237455, 2020.

[21] Raz Avni, Moran Nave, Omer Barad, Kobi Baruch, Sven O Twardziok, Heidrun Gundlach, Iago Hale, Martin Mascher, Manuel Spannagl, Krystalee Wiebe, et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, 357(6346):93–97, 2017.

[22] Martin Ayling, Matthew D Clark, and Richard M Leggett. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, 21(2):584–594, 2020.

[23] Jasmijn A Baaijens and Alexander Schönhuth. Overlap graph-based generation of haplotigs for diploids and polyploids. *Bioinformatics*, 35(21):4281–4289, 2019.

[24] Shakuntala Baichoo and Christos A Ouzounis. Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Biosystems*, 156:72–85, 2017.

[25] Sara Ballouz, Alexander Dobin, and Jesse A Gillis. Is it time to change the reference genome? *Genome biology*, 20(1):1–9, 2019.

[26] Zhigui Bao, Canhui Li, Guangcun Li, Pei Wang, Zhen Peng, Lin Cheng, Hongbo Li, Zhiyang Zhang, Yuying Li, Wu Huang, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant*, 15(7):1211–1226, 2022.

[27] Yury A Barbitoff, Ruslan Abasov, Varvara E Tvorogova, Andrey S Glotov, and Alexander V Predeus. Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC genomics*, 23(1):1–17, 2022.

[28] Inbar Bariah, Danielle Keidar-Friedman, and Khalil Kashkush. Where the wild things are: Transposable elements as drivers of structural and functional variations in the wheat genome. *Frontiers in Plant Science*, 11:1477, 2020.

[29] Craig F Barrett, Christine D Bacon, Alexandre Antonelli, Ángela Cano, and Tobias Hofmann. An introduction to plant phylogenomics with a focus on palms. *Botanical Journal of the Linnean Society*, 182(2):234–255, 2016.

[30] Tyler Barrus. PyProbables, 3 2022.

[31] Jacqueline Batley and David Edwards. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Current opinion in plant biology*, 30:78–81, 2016.

[32] Yves Bawin, Tom Ruttink, Ariane Staelens, Annelies Haegeman, Piet Stoffelen, Jean-Claude Ithe Mwanga Mwanga, Isabel Roldán-Ruiz, Olivier Honnay, and Steven B Janssens. Phylogenomic analysis clarifies the evolutionary origin of coffea arabica. *Journal of Systematics and Evolution*, 59(5):953–963, 2021.

[33] Philipp E Bayer, Agnieszka A Golicz, Armin Scheben, Jacqueline Batley, and David Edwards. Plant pan-genomes are the new reference. *Nature plants*, 6(8):914–920, 2020.

[34] Daniel P Bebber and Sarah J Gurr. Crop-destroying fungal and oomycete pathogens challenge food security. *Fungal Genetics and Biology*, 74:62–64, 2015.

[35] Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, 112(17):5473–5478, 2015.

[36] A Belyayev. Bursts of transposable elements as an evolutionary driving force. *Journal of evolutionary biology*, 27(12):2573–2584, 2014.

[37] Giorgio Bernardi. The genomic code: A pervasive encoding/molding of chromatin structures and a solution of the "non-coding dna" mystery. *BioEssays*, 41(12):1900106, 2019.

[38] David J Bertioli, Jerry Jenkins, Josh Clevenger, Olga Dudchenko, Dongying Gao, Guillermo Seijo, Soraya CM Leal-Bertioli, Longhui Ren, Andrew D Farmer, Manish K Pandey, et al. The genome sequence of segmental allotetraploid peanut arachis hypogaea. *Nature genetics*, 51(5):877–884, 2019.

[39] David John Bertioli, Steven B Cannon, Lutz Froenicke, Guodong Huang, Andrew D Farmer, Ethalinda KS Cannon, Xin Liu, Dongying Gao, Josh Clevenger, Sudhansu Dash, et al. The genome sequences of arachis duranensis and arachis ipaensis, the diploid ancestors of cultivated peanut. *Nature genetics*, 48(4):438–446, 2016.

[40] Niklas Birth, Thomas Dencker, and Burkhard Morgenstern. Insertions and deletions as phylogenetic signal in an alignment-free context. *PLoS computational biology*, 18(8):e1010303, 2022.

[41] Paul D Blischak, Mathews Sajan, Michael S Barker, and Ryan N Gutenkunst. Demographic history inference and the polyploid continuum. *bioRxiv*, 2022.

[42] Flavio Bonomi, Michael Mitzenmacher, Rina Panigrahy, Sushil Singh, and George Varghese. An improved construction for counting bloom filters. In *European Symposium on algorithms*, pages 684–695. Springer, 2006.

[43] Philippa Borrill, Nikolai Adamski, and Cristobal Uauy. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist*, 208(4):1008–1022, 2015.

[44] Anastassia Boudichevskaia, Anne Fiebig, Katrin Kumke, Axel Himmelbach, and Andreas Houben. Rye b chromosomes differently influence the expression of a chromosome–encoded genes depending on the host species. *Chromosome Research*, pages 1–15, 2022.

[45] Yann Bourgeois and Stéphane Boissinot. On the population dynamics of junk: a review on the population genomics of transposable elements. *Genes*, 10(6):419, 2019.

[46] Guillaume Bourque, Kathleen H Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L Levin, Todd S Macfarlan, et al. Ten things you should know about transposable elements. *Genome biology*, 19(1):1–12, 2018.

[47] Keith R Bradnam, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan

Chikhi, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):2047–217X, 2013.

[48] Ljiljana Brankovic, Costas S Iliopoulos, Ritu Kundu, Manal Mohamed, Solon P Pissis, and Fatima Vayani. Linear-time superbubble identification algorithm for genome assembly. *Theoretical Computer Science*, 609:374–383, 2016.

[49] Jean-Simon Brouard, Flavio Schenkel, Andrew Marete, and Nathalie Bissonnette. The gatk joint genotyping workflow is appropriate for calling variants in rna-seq experiments. *Journal of animal science and biotechnology*, 10(1):1–6, 2019.

[50] C Titus Brown and Luiz Irber. sourmash: a library for minhash sketching of dna. *Journal of Open Source Software*, 1(5):27, 2016.

[51] Tilman Brück and Marco d'Errico. Food security and violent conflict: Introduction to the special issue. *World development*, 117:167–171, 2019.

[52] Joshua N Burton, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman, and Jay Shendure. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology*, 31(12):1119–1125, 2013.

[53] Brian Bushnell. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[54] Yuval Bussi, Ruti Kapon, and Ziv Reich. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PloS one*, 16(10):e0258693, 2021.

[55] Xu Cai, Lichun Chang, Tingting Zhang, Haixu Chen, Lei Zhang, Runmao Lin, Jianli Liang, Jian Wu, Michael Freeling, and Xiaowu Wang. Impacts of allopolyploidization and structural variation on intraspecific diversification in brassica rapa. *Genome biology*, 22(1):1–24, 2021.

[56] Stefan Canzar and Steven L Salzberg. Short read mapping: an algorithmic tour. *Proceedings of the IEEE*, 105(3):436–458, 2015.

[57] José Carballo, BACM Santos, D Zappacosta, Ingrid Garbus, Juan Pablo Selva, Cristian Andrés Gallo, A Díaz, Emiliano Albertini, M Caccamo, and V Echenique. A high-quality genome of eragrostis curvula grass provides insights into poaceae evolution and supports new strategies to enhance forage quality. *Scientific reports*, 9(1):1–15, 2019.

[58] Nathan S Catlin and Emily B Josephs. The important contribution of transposable elements to phenotypic variation and evolution. *Current Opinion in Plant Biology*, 65:102140, 2022.

[59] Bastien Cazaux, Thierry Lecroq, and Eric Rivals. Linking indexing data structures to de bruijn graphs: Construction and update. *Journal of Computer and System Sciences*, 104:165–183, 2019.

[60] Monika Cechova. Probably correct: rescuing repeats with short and long reads. *Genes*, 12(1):48, 2020.

[61] Boulos Chalhoub, France Denoeud, Shengyi Liu, Isobel AP Parkin, Haibao Tang, Xiyin Wang, Julien Chiquet, Harry Belcram, Chaobo Tong, Birgit Samans, et al. Early allopolyploid evolution in the post-neolithic brassica napus oilseed genome. *science*, 345(6199):950–953, 2014.

[62] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

[63] Z Jeffrey Chen, Avinash Sreedasyam, Atsumi Ando, Qingxin Song, Luis M De Santiago, Amanda M Hulse-Kemp, Mingquan Ding, Wenxue Ye, Ryan C Kirkbride, Jerry Jenkins, et al. Genomic diversifications of five gossypium allopolyploid species and their impact on cotton improvement. *Nature genetics*, 52(5):525–533, 2020.

[64] Feng Cheng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. Gene retention, fractionation and subgenome differences in polyploid plants. *Nature plants*, 4(5):258–268, 2018.

[65] Paul Chew, Kendra Meade, Alec Hayes, Carlos Harjes, Yong Bao, Aaron D Beattie, Ian Puddephat, Gabe Gusmini, and Steven D Tanksley. A study on the genetic relationships of avena taxa and the origins of hexaploid oat. *Theoretical and Applied Genetics*, 129(7):1405–1415, 2016.

[66] Rayan Chikhi, Jan Holub, and Paul Medvedev. Data structures to represent a set of k-long dna sequences. *ACM Computing Surveys (CSUR)*, 54(1):1–22, 2021.

[67] Rayan Chikhi, Antoine Limasset, Shaun Jackman, Jared T Simpson, and Paul Medvedev. On the representation of de bruijn graphs. In *International conference on Research in computational molecular biology*, pages 35–55. Springer, 2014.

[68] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology*, 8(1):1–9, 2013.

[69] Edward B Chuong, Nels C Elde, and Cédric Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86, 2017.

[70] Michael J Clark, Rui Chen, Hugo YK Lam, Konrad J Karczewski, Rong Chen, Ghia Euskirchen, Atul J Butte, and Michael Snyder. Performance comparison of exome dna sequencing technologies. *Nature biotechnology*, 29(10):908–914, 2011.

[71] Bernardo J Clavijo, Gonzalo Garcia Accinelli, Jonathan Wright, Darren Heavens, Katie Barr, Luis Yanes, and Federica Di-Palma. W2rap: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *BioRxiv*, page 110999, 2017.

[72] Josh Clevenger, Carolina Chavarro, Stephanie A Pearl, Peggy Ozias-Akins, and Scott A Jackson. Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations. *Molecular plant*, 8(6):831–846, 2015.

[73] Eliot Cline, Nuttachat Wisittipanit, Tossapon Boongoen, Ekachai Chukeatirote, Darush Struss, and Anant Eungwanichayapant. Recalibration of mapping quality scores in illumina short-read alignments improves snp detection results in low-coverage sequencing data. *PeerJ*, 8:e10501, 2020.

[74] Nunzia Colonna Romano and Laura Fanti. Transposable elements: Major players in shaping genomic and evolutionary patterns. *Cells*, 11(6):1048, 2022.

[75] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011.

[76] Jake R Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 2017.

[77] Peter A Crisp, Pooja Bhatnagar-Mathur, Penny Hundleby, Ian D Godwin, Peter M Waterhouse, and Lee T Hickey. Beyond the gene: epigenetic and cis-regulatory targets offer new breeding potential for the future. *Current Opinion in Biotechnology*, 73:88–94, 2022.

[78] Tanya Y Curtis and Nigel G Halford. Reducing the acrylamide-forming potential of wheat. *Food and Energy Security*, 5(3):153–164, 2016.

[79] Elena Dalla Benetta, Omar S Akbari, and Patrick M Ferree. Sequence expression of supernumerary b chromosomes: Function or fluff? *Genes*, 10(2):123, 2019.

[80] Romain De Oliveira, Hélène Rimbert, François Balfourier, Jonathan Kitt, Emeric Dynomant, Jan Vrána, Jaroslav Doležel, Federica Cattonaro, Etienne Paux, and Frédéric Choulet. Structural variations affecting genes and transposable elements of chromosome 3b in wheats. *Frontiers in genetics*, 11:891, 2020.

[81] Arthur L Delcher, Simon Kasif, Robert D Fleischmann, Jeremy Peterson, Owen White, and Steven L Salzberg. Alignment of whole genomes. *Nucleic acids research*, 27(11):2369–2376, 1999.

[82] Arthur L Delcher, Adam Phillippy, Jane Carlton, and Steven L Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research*, 30(11):2478–2483, 2002.

[83] Rafael Della Coletta, Yinjie Qiu, Shujun Ou, Matthew B Hufford, and Candice N Hirsch. How the pan-genome is changing crop genomics and improvement. *Genome biology*, 22(1):1–19, 2021.

[84] Hannes Dempewolf, Gregory Baute, Justin Anderson, Benjamin Kilian, Chelsea Smith, and Luigi Guarino. Past and future use of wild relatives in crop breeding. *Crop science*, 57(3):1070–1082, 2017.

[85] Colin N Dewey. Whole-genome alignment. *Evolutionary Genomics*, pages 237–257, 2012.

[86] Om Parkash Dhankher and Christine H Foyer. Climate resilient crops for improving global food security and safety, 2018.

[87] Yu Ding, Jiannan Zhu, Dongsheng Zhao, Qiaoquan Liu, Qingqing Yang, and Tao Zhang. Targeting cis-regulatory elements for rice grain quality improvement. *Frontiers in Plant Science*, 12, 2021.

[88] Aria Dolatabadian, Dhwani Apurva Patel, David Edwards, and Jacqueline Batley. Copy number variation and disease resistance in plants. *Theoretical and Applied Genetics*, 130(12):2479–2490, 2017.

[89] Katherine Domb, Danielle Keidar-Friedman, and Khalil Kashkush. A novel miniature transposon-like element discovered in the coding sequence of a gene that encodes for 5-formyltetrahydrofolate in wheat. *BMC plant biology*, 19(1):1–11, 2019.

[90] Marisol Domínguez, Elise Dugas, Médine Benchouaia, Basile Leduque, José M Jiménez-Gómez, Vincent Colot, and Leandro Quadrana. The impact of transposable elements on tomato diversity. *Nature communications*, 11(1):1–11, 2020.

[91] Thomas Dresselhaus and Ralph Hückelhoven. Biotic and abiotic stress responses in crop plants, 2018.

[92] Karolina Dudziak, Magdalena Zapalska, Andreas Börner, Hubert Szczerba, Krzysztof Kowalczyk, and Michał Nowak. Analysis of wheat gene expression related to the oxidative stress response and signal transduction under short-term osmotic stress. *Scientific reports*, 9(1):1–14, 2019.

[93] Dent Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R Zerbino, Mark Diekhans, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, 21(12):2224–2241, 2011.

[94] David Edwards, Jacqueline Batley, and Rod J Snowdon. Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics*, 126(1):1–11, 2013.

[95] Jordan M Eizenga, Adam M Novak, Emily Kobayashi, Flavia Villani, Cecilia Cisar, Simon Heumos, Glenn Hickey, Vincenza Colonna, Benedict Paten, and Erik Garrison. Efficient dynamic variation graphs. *Bioinformatics*, 36(21):5139–5144, 2021.

[96] Robert Ekblom, Linnéa Smeds, and Hans Ellegren. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC genomics*, 15(1):1–9, 2014.

[97] Mohamed A El-Esawi, Abdullah A Al-Ghamdi, Hayssam M Ali, and Margaret Ahmad. Overexpression of atwrky30 transcription factor enhances heat and drought stress tolerance in wheat (triticum aestivum l.). *Genes*, 10(2):163, 2019.

[98] Marius Erbert, Steffen Rechner, and Matthias Müller-Hannemann. Gerbil: a fast and memory-efficient k-mer counter with gpu-support. *Algorithms for Molecular Biology*, 12(1):1–12, 2017.

[99] Joshua J Faber-Hammond and Kim H Brown. Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads. *Human genetics*, 135(7):727–740, 2016.

[100] Joshua J Faber-Hammond and Kim H Brown. Pseudo-de novo assembly and analysis of unmapped genome sequence reads in wild zebrafish reveal novel gene content. *Zebrafish*, 13(2):95–102, 2016.

[101] FAO. Food security and conflict. empirical challenges and future opportunities for research and policy making on food security and conflict., 2018.

[102] FAO. The future of food and agriculture - alternative pathways to 2050 - summary version, 2018.

[103] Bruno Fernandes, Alfonso González-Briones, Paulo Novais, Miguel Calafate, Cesar Analide, and José Neves. An adjective selection personality assessment method using gradient boosting machine learning. *Processes*, 8(5):618, 2020.

[104] Can Firtina and Can Alkan. On genomic repeats and reproducibility. *Bioinformatics*, 32(15):2243–2247, 2016.

[105] Martina Flörke, Christof Schneider, and Robert I McDonald. Water competition between cities and agriculture driven by climate change and urban growth. *Nature Sustainability*, 1(1):51–58, 2018.

[106] Michael Freeling, Michael J Scanlon, and John E Fowler. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development*, 35:110–118, 2015.

[107] Sam Friedman, Laura Gauthier, Yossi Farjoun, and Eric Banks. Lean and deep models for more accurate filtering of snp and indel variant calls. *Bioinformatics*, 36(7):2060–2067, 2020.

[108] Yong-Bi Fu. Oat evolution revealed in the maternal lineages of 25 avena species. *Scientific Reports*, 8(1):1–12, 2018.

[109] Roven Rommel Fuentes, Dmytro Chebotarov, Jorge Duitama, Sean Smith, Juan Fernando De la Hoz, Marghoob Mohiyuddin, Rod A Wing, Kenneth L McNally, Tatiana Tatarinova, Andrey Grigoriev, et al. Structural variants in 3000 rice genomes. *Genome research*, 29(5):870–880, 2019.

[110] José Fuentes-Sepúlveda, Gonzalo Navarro, and Yakov Nekrich. Space-efficient computation of the burrows-wheeler transform. In *2019 Data Compression Conference (DCC)*, pages 132–141. IEEE, 2019.

[111] Iulian Gabur, Harmeet Singh Chawla, Rod J Snowdon, and Isobel AP Parkin. Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, 132(3):733–750, 2019.

[112] Joseph L Gage, Brieanne Vaillancourt, John P Hamilton, Norma C Manrique-Carpintero, Timothy J Gustafson, Kerrie Barry, Anna Lipzen, William F Tracy, Mark A Mikel, Shawn M Kaeppler, et al. Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *The plant genome*, 12(2):180069, 2019.

[113] Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M Tieman, Elizabeth A Burzynski-Chang, Tara L Fish, Kaitlin A Stromberg, Gavin L Sacks, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, 51(6):1044–1051, 2019.

[114] Naina Garewal, Neetu Goyal, Shivalika Pathania, Jagdeep Kaur, and Kashmir Singh. Gauging the trends of pseudogenes in plants. *Critical Reviews in Biotechnology*, pages 1–16, 2021.

[115] Shilpa Garg. Computational methods for chromosome-scale haplotype reconstruction. *Genome biology*, 22(1):1–24, 2021.

[116] Sarah Garland and Helen Anne Curry. Turning promise into practice: Crop biotechnology for increasing genetic diversity and climate resilience. *PLoS Biology*, 20(7):e3001716, 2022.

[117] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

[118] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, 36(9):875–879, 2018.

[119] Fabian Gärtner, Lydia Müller, and Peter F Stadler. Superbubbles revisited. *Algorithms for Molecular Biology*, 13(1):1–17, 2018.

[120] Jean-François Gibrat. On the use of algebraic topology concepts to check the consistency of genome assembly. *Biophysics and Physicobiology*, 16:444–451, 2019.

[121] Clément Gilbert and Cédric Feschotte. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current opinion in genetics & development*, 49:15–24, 2018.

[122] Agnieszka A Golicz, Jacqueline Batley, and David Edwards. Towards plant pangenomics. *Plant Biotechnology Journal*, 14(4):1099–1105, 2016.

[123] Sean P Gordon, Joshua J Levy, and John P Vogel. Polycracker, a robust method for the unsupervised partitioning of polyploid subgenomes by signatures of repetitive dna evolution. *BMC genomics*, 20(1):580, 2019.

[124] R Quentin Grafton, Carsten Daugbjerg, and M Ejaz Qureshi. Towards food security by 2050. *Food Security*, 7(2):179–183, 2015.

[125] Marie-Angèle Grandbastien. Ltr retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1849(4):403–416, 2015.

[126] Ivar Grytten, Knut Dagestad Rand, and Geir Kjetil Sandve. Kage: Fast alignment-free graph-based genotyping of snps and short indels. *Genome biology*, 23(1):1–15, 2022.

[127] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

[128] Songtao Gui, Wenjie Wei, Chenglin Jiang, Jingyun Luo, Lu Chen, Shenshen Wu, Wenqiang Li, Yuebin Wang, Shuyan Li, Ning Yang, et al. A pan-zea genome map for enhancing maize improvement. *Genome biology*, 23(1):1–22, 2022.

[129] Jian Guo, Ke Cao, Cecilia Deng, Yong Li, Gengrui Zhu, Weichao Fang, Changwen Chen, Xinwei Wang, Jinlong Wu, Liping Guan, et al. An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome biology*, 21(1):1–19, 2020.

[130] Rui Guo, Yan-Ran Li, Shan He, Le Ou-Yang, Yiwen Sun, and Zexuan Zhu. Replong: de novo repeat identification using long read sequencing data. *Bioinformatics*, 34(7):1099–1107, 2018.

[131] Yan Guo, Jirong Long, Jing He, Chung-I Li, Qiuyin Cai, Xiao-Ou Shu, Wei Zheng, and Chun Li. Exome sequencing generates high quality data in non-target regions. *BMC genomics*, 13(1):194, 2012.

[132] Gaurav Gupta, Minghao Yan, Benjamin Coleman, Bryce Kille, RA Leo Elworth, Tharun Medini, Todd Treangen, and Anshumali Shrivastava. Fast processing and querying of 170tb of genomics data via a repeated and merged bloom filter (rambo). In *Proceedings of the 2021 International Conference on Management of Data*, pages 2226–2234, 2021.

[133] Raj Kishor Gupta, Shivraj Singh Gangoliya, and Nand Kumar Singh. Reduction of phytic acid and enhancement of bioavailable micronutrients in food grains. *Journal of food science and technology*, 52(2):676–684, 2015.

[134] Ehsan Haghshenas, Hossein Asghari, Jens Stoye, Cedric Chauve, and Faraz Hach. Haslr: Fast hybrid assembly of long reads. *Iscience*, 23(8):101389, 2020.

[135] Tuomas Hämälä, Eric K Wafula, Mark J Guiltinan, Paula E Ralph, Claude W dePamphilis, and Peter Tiffin. Genomic structural variants constrain and facilitate adaptation in natural populations of theobroma cacao, the chocolate tree. *Proceedings of the National Academy of Sciences*, 118(35):e2102914118, 2021.

[136] Dan He, Subrata Saha, Richard Finkers, and Laxmi Parida. Efficient algorithms for polyploid haplotype phasing. *BMC genomics*, 19(2):171–180, 2018.

[137] Jiangfeng He, Xiaoqing Zhao, André Laroche, Zhen-Xiang Lu, HongKui Liu, and Ziqin Li. Genotyping-by-sequencing (gbs), an ultimate marker-assisted selection (mas) tool to accelerate plant breeding. *Frontiers in plant science*, 5:484, 2014.

[138] Yun Heo, Xiao-Long Wu, Deming Chen, Jian Ma, and Wen-Mei Hwu. Bless: bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics*, 30(10):1354–1362, 2014.

[139] Julie E Hernández-Salmerón and Gabriel Moreno-Hagelsieb. Fastani, mash and dashing equally differentiate between klebsiella species. *PeerJ*, 10:e13784, 2022.

[140] Alison B Hickman, Andrea Regier Voth, Hosam Ewis, Xianghong Li, Nancy L Craig, and Fred Dyda. Structural insights into the mechanism of double strand break formation by hermes, a hat family eukaryotic dna transposase. *Nucleic acids research*, 46(19):10286–10301, 2018.

[141] Cory D Hirsch, John P Hamilton, Kevin L Childs, Jason Cepela, Emily Crisovan, Brieanne Vaillancourt, Candice N Hirsch, Marc Habermann, Brayden Neal, and C Robin Buell. Spud db: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome*, 7(1):plantgenome2013–12, 2014.

[142] Douglas R Hoen and Thomas E Bureau. Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Molecular biology and evolution*, 32(6):1487–1506, 2015.

[143] Lindsay A Holden, Meharji Arumilli, Marjo K Hytönen, Sruthi Hundi, Jarkko Salojärvi, Kim H Brown, and Hannes Lohi. Assembly and analysis of unmapped genome sequence reads reveal novel sequence and variation in dogs. *Scientific reports*, 8(1):1–11, 2018.

[144] Guillaume Holley and Páll Melsted. Bifrost: highly parallel construction and indexing of colored and compacted de bruijn graphs. *Genome biology*, 21(1):1–20, 2020.

[145] Genevieve Hoopes, Xiaoxi Meng, John P Hamilton, Sai Reddy Achakkagari, Fernanda de Alves Freitas Guesdes, Marie E Bolger, Joseph J Coombs, Danny Esselink, Natalie R Kaiser, Linda Kodde, et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular Plant*, 15(3):520–536, 2022.

[146] Andreas Houben, Ali Mohammad Banaei-Moghaddam, Sonja Klemme, and Jeremy N Timmis. Evolution and biology of supernumerary b chromosomes. *Cellular and Molecular Life Sciences*, 71(3):467–478, 2014.

[147] Mark Howison, Felipe Zapata, and Casey W Dunn. Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics*, 29(23):2959–2963, 2013.

[148] Yana Hrytsenko, Noah M Daniels, and Rachel S Schwartz. Sensitivity of a frequency-based alignment-free approach for phylogeny reconstruction. *Research Square*, 2022.

[149] Bhavna Hurgobin, Agnieszka A Golicz, Philipp E Bayer, Chon-Kit Kenneth Chan, Soodeh Tirnaz, Aria Dolatabadian, Sarah V Schiessl, Birgit Samans, Juan D Montenegro, Isobel AP Parkin, et al. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid brassica napus. *Plant biotechnology journal*, 16(7):1265–1274, 2018.

[150] Illumina. genotyping by sequencing sequence-based genotyping methods. `https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/genotyping-by-sequencing.html`, 2021.

[151] IPCC. Climate change and land. an ipcc special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. summary for policymakers., 2020.

[152] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226–232, 2012.

[153] Pesho Ivanov, Benjamin Bichsel, Harun Mustafa, André Kahles, Gunnar Rätsch, and Martin Vechev. Astarix: Fast and optimal sequence-to-graph alignment. In *International Conference on Research in Computational Molecular Biology*, pages 104–119. Springer, 2020.

[154] International Wheat Genome Sequencing Consortium (IWGSC), Rudi Appels, Kellye Eversole, Nils Stein, Catherine Feuillet, Beat Keller, Jane Rogers, Curtis J Pozniak, Frederic Choulet, Assaf Distelfeld, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403):eaar7191, 2018.

[155] Shaun D Jackman, Benjamin P Vandervalk, Hamid Mohamadi, Justin Chu, Sarah Yeo, S Austin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, Rene L Warren, et al. Abyss 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome research*, 27(5):768–777, 2017.

[156] Chirag Jain, Arang Rhie, Nancy F Hansen, Sergey Koren, and Adam M Phillippy. Long-read mapping to repetitive reference sequences using winnowmap2. *Nature Methods*, pages 1–6, 2022.

[157] Chirag Jain, Neda Tavakoli, and Srinivas Aluru. A variant selection framework for genome graphs. *Bioinformatics*, 37(Supplement_1):i460–i467, 2021.

[158] Murukarthick Jayakodi, Sudharsan Padmarasu, Georg Haberer, Venkata Suresh Bonthala, Heidrun Gundlach, Cécile Monat, Thomas Lux, Nadia Kamal, Daniel Lang, Axel Himmelbach, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 588(7837):284–289, 2020.

[159] Kai-Hua Jia, Zhao-Xuan Wang, Longxin Wang, Guang-Yuan Li, Wei Zhang, Xiao-Ling Wang, Fang-Ji Xu, Si-Qian Jiao, Shan-Shan Zhou, Hui Liu, et al. Subphaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytologist*, 2022.

[160] Jiming Jiang. The 'dark matter'in the plant genomes: non-coding and unannotated dna sequences associated with open chromatin. *Current opinion in plant biology*, 24:17–23, 2015.

[161] Yun-Feng Jiang, Qing Chen, Yan Wang, Zhen-Ru Guo, Bin-Jie Xu, Jing Zhu, Ya-Zhou Zhang, Xi Gong, Cui-Hua Luo, Wang Wu, et al. Re-acquisition of the brittle rachis trait via a transposon insertion in domestication gene q during wheat de-domestication. *New Phytologist*, 224(2):961–973, 2019.

[162] Zoé Joly-Lopez, Jonathan M Flowers, and Michael D Purugganan. Developing maps of fitness consequences for plant genomes. *Current opinion in plant biology*, 30:101–107, 2016.

[163] Zoé Joly-Lopez, Ewa Forczek, Emilio Vello, Douglas R Hoen, Akiko Tomita, and Thomas E Bureau. Abiotic stress phenotypes are associated with conserved genes derived from transposable elements. *Frontiers in plant science*, 8:2027, 2017.

[164] Zoé Joly-Lopez, Adrian E Platts, Brad Gulko, Jae Young Choi, Simon C Groen, Xuehua Zhong, Adam Siepel, and Michael D Purugganan. An inferred fitness consequence map of the rice genome. *Nature plants*, 6(2):119–130, 2020.

[165] Neil Jones and Alevtina Ruban. Are b chromosomes useful for crop improvement? *Plants, People, Planet*, 1(2):84–92, 2019.

[166] Sateesh Kagale, Chushin Koh, Wayne E Clarke, Venkatesh Bollina, Isobel AP Parkin, and Andrew G Sharpe. Analysis of genotyping-by-sequencing (gbs) data. In *Plant Bioinformatics*, pages 269–284. Springer, 2016.

[167] Sateesh Kagale, Chushin Koh, John Nixon, Venkatesh Bollina, Wayne E Clarke, Reetu Tuteja, Charles Spillane, Stephen J Robinson, Matthew G Links, Carling Clarke, et al. The emerging biofuel crop camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature communications*, 5(1):1–11, 2014.

[168] Nadia Kamal, Nikos Tsardakas Renhuldt, Johan Bentzer, Heidrun Gundlach, Georg Haberer, Angéla Juhász, Thomas Lux, Utpal Bose, Jason A Tye-Din, Daniel Lang, et al. The mosaic oat genome gives insights into a uniquely healthy cereal crop. *Nature*, 606(7912):113–119, 2022.

[169] Veronika Kapustová, Zuzana Tulpová, Helena Toegelová, Petr Novák, Jiří Macas, Miroslava Karafiátová, Eva Hřibová, Jaroslav Doležel, and Hana Šimková. The dark matter of large cereal genomes: Long tandem repeats. *International journal of molecular sciences*, 20(10):2483, 2019.

[170] Mikhail Karasikov, Harun Mustafa, Daniel Danciu, Marc Zimmermann, Christopher Barber, Gunnar Ratsch, and André Kahles. Metagraph: Indexing and analysing nucleotide archives at petabase-scale. *bioRxiv*, 2020.

[171] Mikhail Karasikov, Harun Mustafa, Gunnar Rätsch, and André Kahles. Lossless indexing with counting de bruijn graphs. In *International Conference on Research in Computational Molecular Biology*, pages 374–376. Springer, 2022.

[172] Lee S Katz, Taylor Griswold, Shatavia S Morrison, Jason A Caravas, Shaokang Zhang, Henk C den Bakker, Xiangyu Deng, and Heather A Carleton. Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software*, 4(44):1762, 2019.

[173] Pınar Kavak, Bayram Yüksel, Soner Aksu, M Oguzhan Kulekci, Tunga Güngör, Faraz Hach, S Cenk Şahinalp, Turkish Human Genome Project, Can Alkan, and Mahmut Şamil Sağıroğlu. Robustness of massively parallel sequencing platforms. *PLoS One*, 10(9):e0138259, 2015.

[174] Jamshed Khan and Rob Patro. Cuttlefish: fast, parallel and low-memory compaction of de bruijn graphs from large-scale genome collections. *Bioinformatics*, 37(Supplement_1):i177–i186, 2021.

[175] Maguta Kibe, Christine Nyaga, Sudha K Nair, Yoseph Beyene, Biswanath Das, Jumbo M Bright, Dan Makumbi, Johnson Kinyua, Michael S Olsen, Boddupalli M Prasanna, et al. Combination of linkage mapping, gwas, and gp to dissect the genetic basis of common rust resistance in tropical maize germplasm. *International journal of molecular sciences*, 21(18):6518, 2020.

[176] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

[177] Kyung Do Kim, Yuna Kang, and Changsoo Kim. Application of genomic big data in plant breeding: Past, present, and future. *Plants*, 9(11):1454, 2020.

[178] Marius Knaust, Enrico Seiler, Knut Reinert, and Thomas Steinke. Co-design for energy efficient and fast genomic search: Interleaved bloom filter on fpga. In *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 180–189, 2022.

[179] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

[180] Mikhail Kolmogorov, Derek M Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy PL Smith, et al. metaflye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, 2020.

[181] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546, 2019.

[182] Karl AG Kremling, Shu-Yun Chen, Mei-Hsiu Su, Nicholas K Lepak, M Cinta Romay, Kelly L Swarts, Fei Lu, Anne Lorant, Peter J Bradbury, and Edward S Buckler. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, 555(7697):520–523, 2018.

[183] Peter Krusche, Len Trigg, Paul C Boutros, Christopher E Mason, Francisco M De La Vega, Benjamin L Moore, Mar Gonzalez-Porta, Michael A Eberle, Zivana Tezak, Samir Lababidi, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology*, 37(5):555–560, 2019.

[184] Ajay Kumar, Reyazul Rouf Mir, Deepmala Sehgal, Pinky Agarwal, and Arron Carter. Genetics and genomics to enhance crop production, towards food security. *Frontiers in Genetics*, 12, 2021.

[185] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):1–9, 2004.

[186] Maria Kyriakidou, Helen H Tai, Noelle L Anglin, David Ellis, and Martina V Strömvik. Current strategies of polyploid plant genome sequence assembly. *Frontiers in plant science*, 9:1660, 2018.

[187] Sophie Lanciano and Marie Mirouze. Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Current opinion in genetics & development*, 49:106–114, 2018.

[188] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[189] Peter A Larsen. Transposable elements and the multidimensional genome. *Chromosome Research*, 26(1):1–3, 2018.

[190] Debbie Laudencia-Chingcuanco and D Brian Fowler. Genotype-dependent burst of transposable element expression in crowns of hexaploid wheat (triticum aestivum l.) during cold acclimation. *Comparative and functional genomics*, 2012, 2012.

[191] SC Le Comber, ML Ainouche, A Kovarik, and AR Leitch. Making a functional diploid: from polysomic to disomic inheritance. *New Phytologist*, 186(1):113–122, 2010.

[192] Messaoud Lefouili and Kiwoong Nam. The evaluation of bcftools mpileup and gatk haplotypecaller for variant calling in non-human species. *Scientific Reports*, 12(1):1–8, 2022.

[193] Miika Leinonen and Leena Salmela. Optical map guided genome assembly. *BMC bioinformatics*, 21(1):1–19, 2020.

[194] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[195] Marc-André Lemay and Sidiki Malle. A practical guide to using structural variants for genome-wide association studies. In *Genome-Wide Association Studies*, pages 161–172. Springer, 2022.

[196] Brice Letcher, Martin Hunt, and Zamin Iqbal. Gramtools enables multiscale variation analysis with genome graphs. *Genome biology*, 22(1):1–27, 2021.

[197] Fuguang Li, Guangyi Fan, Cairui Lu, Guanghui Xiao, Changsong Zou, Russell J Kohel, Zhiying Ma, Haihong Shang, Xiongfeng Ma, Jianyong Wu, et al. Genome sequence of cultivated upland cotton (gossypium hirsutum tm-1) provides insights into genome evolution. *Nature biotechnology*, 33(5):524–530, 2015.

[198] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, 2014.

[199] Kui Li, Wenkai Jiang, Yuanyuan Hui, Mengjuan Kong, Li-Ying Feng, Li-Zhi Gao, Pengfu Li, and Shan Lu. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Molecular Plant*, 14(10):1745–1756, 2021.

[200] Lin-Feng Li, Zhi-Bin Zhang, Zhen-Hui Wang, Ning Li, Yan Sha, Xin-Feng Wang, Ning Ding, Yang Li, Jing Zhao, Ying Wu, et al. Genome sequences of five sitopsis species of aegilops and the origin of polyploid wheat b subgenome. *Molecular plant*, 15(3):488–503, 2022.

[201] Yifeng Li, Chih-yu Chen, Alice M Kaye, and Wyeth W Wasserman. The identification of cis-regulatory elements: A review from a machine learning perspective. *Biosystems*, 138:6–17, 2015.

[202] Zhikai Liang, Sarah N Anderson, Jaclyn M Noshay, Peter A Crisp, Tara A Enders, and Nathan M Springer. Genetic and epigenetic variation in transposable element expression responses to abiotic stress in maize. *Plant Physiology*, 186(1):420–433, 2021.

[203] Xingyu Liao, Min Li, You Zou, Fang-Xiang Wu, Jianxin Wang, et al. Current challenges and solutions of de novo assembly. *Quantitative Biology*, 7(2):90–109, 2019.

[204] Antoine Limasset, Bastien Cazaux, Eric Rivals, and Pierre Peterlongo. Read mapping on de bruijn graphs. *BMC bioinformatics*, 17(1):1–12, 2016.

[205] Hong-Qing Ling, Bin Ma, Xiaoli Shi, Hui Liu, Lingli Dong, Hua Sun, Yinghao Cao, Qiang Gao, Shusong Zheng, Ye Li, et al. Genome sequence of the progenitor of wheat a subgenome triticum urartu. *Nature*, 557(7705):424–428, 2018.

[206] Damon Lisch. How important are transposons for plant evolution? *Nature Reviews Genetics*, 14(1):49–61, 2013.

[207] Binghang Liu, Yujian Shi, Jianying Yuan, Xuesong Hu, Hao Zhang, Nan Li, Zhenyu Li, Yanxiang Chen, Desheng Mu, and Wei Fan. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012*, 2013.

[208] Yucheng Liu, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo-An Zhou, Haikuan Zhang, Zhi Liu, Miao Shi, et al. Pan-genome of wild and cultivated soybeans. *Cell*, 182(1):162–176, 2020.

[209] Zhen Liu, Yuling Liu, Fang Liu, Shulin Zhang, Xingxing Wang, Quanwei Lu, Kunbo Wang, Baohong Zhang, and Renhai Peng. Genome-wide survey and comparative analysis of long terminal repeat (ltr) retrotransposon families in four gossypium species. *Scientific reports*, 8(1):1–10, 2018.

[210] Gil Loewenthal, Dana Rapoport, Oren Avram, Asher Moshe, Elya Wygoda, Alon Itzkovitch, Omer Israeli, Dana Azouri, Reed A Cartwright, Itay Mayrose, et al. A probabilistic model for indel evolution: differentiating insertions from deletions. *Molecular biology and evolution*, 38(12):5769–5781, 2021.

[211] John T Lovell, Alice H MacQueen, Sujan Mamidi, Jason Bonnette, Jerry Jenkins, Joseph D Napier, Avinash Sreedasyam, Adam Healey, Adam Session, Shengqiang Shu, et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, 590(7846):438–444, 2021.

[212] Junwei Luo, Jianxin Wang, Zhen Zhang, Fang-Xiang Wu, Min Li, and Yi Pan. Epga: de novo assembly using the distributions of reads and insert size. *Bioinformatics*, 31(6):825–833, 2015.

[213] Ming-Cheng Luo, Yong Q Gu, Daniela Puiu, Hao Wang, Sven O Twardziok, Karin R Deal, Naxin Huo, Tingting Zhu, Le Wang, Yi Wang, et al. Genome sequence of the progenitor of the wheat d genome aegilops tauschii. *Nature*, 551(7681):498–502, 2017.

[214] Zoe N Lye and Michael D Purugganan. Copy number variation in domestication. *Trends in plant science*, 24(4):352–365, 2019.

[215] Wei Ma, Tobias Sebastian Gabriel, Mihaela Maria Martis, Torsten Gursinsky, Veit Schubert, Jan Vrána, Jaroslav Doležel, Heidrun Grundlach, Lothar Altschmied, Uwe Scholz, et al. Rye b chromosomes encode a functional argonaute-like protein with in vitro slicer activities similar to its a chromosome paralog. *New Phytologist*, 213(2):916–928, 2017.

[216] Marco Maccaferri, Neil S Harris, Sven O Twardziok, Raj K Pasam, Heidrun Gundlach, Manuel Spannagl, Danara Ormanbekova, Thomas Lux, Verena M Prade, Sara G Milner, et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature genetics*, 51(5):885–895, 2019.

[217] Jennifer Mach. Camelina: a history of polyploidy, chromosome shattering, and recovery, 2019.

[218] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20(1):1–14, 2019.

[219] Massimo Maiolo, Xiaolei Zhang, Manuel Gil, and Maria Anisimova. Progressive multiple sequence alignment with indel evolution. *BMC bioinformatics*, 19(1):1–8, 2018.

[220] Vasilissa Manova and Damian Gruszka. Dna damage and repair in plants–from models to crops. *Frontiers in plant science*, 6:885, 2015.

[221] Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J Clavijo. Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, 33(4):574–576, 2017.

[222] G Marcais and C Kingsford. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais*, 1:1–8, 2012.

[223] Guillaume Marçais, Dan DeBlasio, Prashant Pandey, and Carl Kingsford. Locality-sensitive hashing for the edit distance. *Bioinformatics*, 35(14):i127–i135, 2019.

[224] Guillaume Marçais, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin. Mummer4: a fast and versatile genome alignment system. *PLoS computational biology*, 14(1):e1005944, 2018.

[225] Jacob I Marsh, Haifei Hu, Mitchell Gill, Jacqueline Batley, and David Edwards. Crop breeding for a changing climate: Integrating phenomics and genomics with bioinformatics. *Theoretical and Applied Genetics*, 134(6):1677–1690, 2021.

[226] Annaliese S Mason and Jonathan F Wendel. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Frontiers in Genetics*, 11:1014, 2020.

[227] Florian Maumus and Hadi Quesneville. Deep investigation of arabidopsis thaliana junk dna reveals a continuum between repetitive elements and genomic dark matter. *PLoS One*, 9(4):e94101, 2014.

[228] Ultan Mc Carthy, Ismail Uysal, Ricardo Badia-Melis, Samuel Mercier, Colm O'Donnell, and Anastasia Ktenioudaki. Global food security–issues, challenges and technological solutions. *Trends in Food Science & Technology*, 77:11–20, 2018.

[229] Michael R McKain, Matthew G Johnson, Simon Uribe-Convers, Deren Eaton, and Ya Yang. Practical considerations for plant phylogenomics. *Applications in plant sciences*, 6(3):e1038, 2018.

[230] Paul Medvedev, Son Pham, Mark Chaisson, Glenn Tesler, and Pavel Pevzner. Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, 18(11):1625–1634, 2011.

[231] Pall Melsted and Jonathan K Pritchard. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC bioinformatics*, 12(1):1–7, 2011.

[232] Todd P Michael and Robert VanBuren. Building near-complete plant genomes. *Current Opinion in Plant Biology*, 54:26–33, 2020.

[233] Marko Mihajlovic and Ning Xiong. Finding the most similar textual documents using case-based reasoning. *arXiv preprint arXiv:1911.00262*, 2019.

[234] Tom Mokveld, Jasper Linthorst, Zaid Al-Ars, Henne Holstege, and Marcel Reinders. Chop: haplotype-aware path indexing in population graphs. *Genome biology*, 21(1):1–16, 2020.

[235] Juan D Montenegro, Agnieszka A Golicz, Philipp E Bayer, Bhavna Hurgobin, HueyTyng Lee, Chon-Kit Kenneth Chan, Paul Visendi, Kaitao Lai, Jaroslav Doležel, Jacqueline Batley, et al. The pangenome of hexaploid bread wheat. *The Plant Journal*, 90(5):1007–1013, 2017.

[236] Martin D Muggli, Alexander Bowe, Noelle R Noyes, Paul S Morley, Keith E Belk, Robert Raymond, Travis Gagie, Simon J Puglisi, and Christina Boucher. Succinct colored de bruijn graphs. *Bioinformatics*, 33(20):3181–3187, 2017.

[237] Pankaj S Mundada, Swapnil B Kadam, Anupama A Pable, and Vitthal T Barvkar. Recent advances and applicability of gbs, gwas, and gs in millet crops. *Genotyping by Sequencing for Crop Improvement*, pages 270–294, 2022.

[238] Mehanathan Muthamilarasan, P Theriappan, and Manoj Prasad. Recent advances in crop genomics for ensuring food security. *Current Science*, 105(2):155–158, 2013.

[239] Sadia Nadir, Wei Li, Qian Zhu, Sehroon Khan, Xiao-Ling Zhang, Hui Zhang, Zhen-Fei Wei, Meng-Ting Li, Li Zhou, Cheng-Yun Li, et al. A novel discovery of a long terminal repeat retrotransposon-induced hybrid weakness in rice. *Journal of experimental botany*, 70(4):1197–1207, 2019.

[240] Brian Nadon and Scott Jackson. The polyploid origins of crop genomes and their implications: A case study in legumes. *Advances in Agronomy*, 159:275–313, 2020.

[241] Bernardo Najlis. parallel-bloom-filters.

[242] Lars Nauheimer, Nicholas Weigner, Elizabeth Joyce, Darren Crayn, Charles Clarke, and Katharina Nargar. Hybphaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in plant sciences*, 9(7), 2021.

[243] Sabuzima Nayak and Ripon Patgiri. A review on role of bloom filter on dna assembly. *IEEE Access*, 7:66939–66954, 2019.

[244] Sabuzima Nayak and Ripon Patgiri. countbf: A general-purpose high accuracy and space efficient counting bloom filter. In *2021 17th International Conference on Network and Service Management (CNSM)*, pages 355–359. IEEE, 2021.

[245] NCBI. Oryza sativa (asian cultivated rice). `https://www.ncbi.nlm.nih.gov/genome/?term=Oryza+sativa`, 2022.

[246] NCBI. Triticum aestivum (bread wheat). https://www.ncbi.nlm.nih.gov/genome/?term=wheat, 2022.

[247] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[248] Firuz Odilbekov, Rita Armoniené, Alexander Koc, Jan Svensson, and Aakash Chawade. Gwas-assisted genomic prediction to predict resistance to septoria tritici blotch in nordic winter wheat at seedling stage. *Frontiers in genetics*, 10:1224, 2019.

[249] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[250] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

[251] Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Detecting superbubbles in assembly graphs. In *International workshop on algorithms in bioinformatics*, pages 338–348. Springer, 2013.

[252] Weihua Pan and Stefano Lonardi. Accurate detection of chimeric contigs via bionano optical maps. *Bioinformatics*, 35(10):1760–1762, 2019.

[253] Yuxin Pan, Fanbo Meng, and Xiyin Wang. Sequencing multiple cotton genomes reveals complex structures and lays foundation for breeding. *Frontiers in Plant Science*, 11:560096, 2020.

[254] Prashant Pandey, Fatemeh Almodaresi, Michael A Bender, Michael Ferdman, Rob Johnson, and Rob Patro. Mantis: A fast, small, and exact large-scale sequence-search index. *Cell systems*, 7(2):201–207, 2018.

[255] Prashant Pandey, Michael A Bender, Rob Johnson, and Rob Patro. Squeakr: an exact and approximate k-mer counting system. *Bioinformatics*, 34(4):568–575, 2018.

[256] Kumar Paritosh, Akshay Kumar Pradhan, and Deepak Pental. A highly contiguous genome assembly of brassica nigra (bb) and revised nomenclature for the pseudochromosomes. *BMC genomics*, 21(1):1–12, 2020.

[257] Kumar Paritosh, Satish Kumar Yadava, Priyansha Singh, Latika Bhayana, Arundhati Mukhopadhyay, Vibha Gupta, Naveen Chandra Bisht, Jianwei Zhang, David A Kudrna, Dario Copetti, et al. A chromosome-scale assembly of allotetraploid brassica

juncea (aabb) elucidates comparative architecture of the a and b genomes. *Plant biotechnology journal*, 19(3):602–614, 2021.

[258] Doori Park, Jong-Hwa Kim, and Nam-Soo Kim. De novo transcriptome sequencing and gene expression profiling with/without b-chromosome plants of lilium amabile. *Genomics & informatics*, 17(3), 2019.

[259] Isobel AP Parkin, Chushin Koh, Haibao Tang, Stephen J Robinson, Sateesh Kagale, Wayne E Clarke, Chris D Town, John Nixon, Vivek Krishnakumar, Shelby L Bidwell, et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid brassica oleracea. *Genome biology*, 15(6):1–18, 2014.

[260] Benedict Paten, Jordan M Eizenga, Yohei M Rosen, Adam M Novak, Erik Garrison, and Glenn Hickey. Superbubbles, ultrabubbles, and cacti. *Journal of Computational Biology*, 25(7):649–663, 2018.

[261] Benedict Paten, Adam M Novak, Jordan M Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.

[262] David Pellow, Darya Filippova, and Carl Kingsford. Improving bloom filter performance on sequence data using k-mer bloom filters. *Journal of Computational Biology*, 24(6):547–557, 2017.

[263] Junhua H Peng, Dongfa Sun, and Eviatar Nevo. Domestication evolution, genetics and genomics in wheat. *Molecular Breeding*, 28(3):281–301, 2011.

[264] Valentina Peona, Mozes PK Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, Tri Haryoko, Knud A Jønsson, Qi Zhou, et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular ecology resources*, 21(1):263–286, 2021.

[265] H Sofia Pereira, Margarida Delgado, Wanda Viegas, João M Rato, Augusta Barão, and Ana D Caperta. Rye (secale cereale) supernumerary (b) chromosomes associated with heat tolerance during early stages of male sporogenesis. *Annals of botany*, 119(3):325–337, 2017.

[266] Nelson Pérez, Miguel Gutierrez, and Nelson Vera. Computational performance assessment of k-mer counting algorithms. *Journal of Computational Biology*, 23(4):248–255, 2016.

[267] Esteban Pérez-Wohlfeil, Sergio Diaz-del Pino, and Oswaldo Trelles. Ultra-fast genome comparison for large-scale genomic experiments. *Scientific reports*, 9(1):1–10, 2019.

[268] Amandine Perrin and Eduardo PC Rocha. Panacota: a modular tool for massive microbial comparative genomics. *NAR genomics and bioinformatics*, 3(1):lqaa106, 2021.

[269] Jakob Petereit, Philipp E Bayer, William JW Thomas, Cassandria G Tay Fernandez, Junrey Amas, Yueqi Zhang, Jacqueline Batley, and David Edwards. Pangenomics and crop genome adaptation in a changing climate. *Plants*, 11(15):1949, 2022.

[270] N Tessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, and C Titus Brown. Large-scale sequence comparisons with sourmash. *F1000Research*, 8, 2019.

[271] Kevin V Pixley, Jose B Falck-Zepeda, Robert L Paarlberg, Peter WB Phillips, Inez H Slamet-Loedin, Kanwarpal S Dhugga, Hugo Campos, and Neal Gutterson. Genome-edited crops for improved food security of smallholder farmers. *Nature Genetics*, 54(4):364–367, 2022.

[272] Wirulda Pootakham, Chutima Sonthirod, Chaiwat Naktang, Nukoon Jomchai, Duangjai Sangsrakru, and Sithichoke Tangphatsornruang. Effects of methylation-sensitive enzymes on the enrichment of genic snps and the degree of genome complexity reduction in a two-enzyme genotyping-by-sequencing (gbs) approach: a case study in oil palm (elaeis guineensis). *Molecular Breeding*, 36(11):1–7, 2016.

[273] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

[274] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178, 2018.

[275] Manuel Poretti, Coraline Rosalie Praz, Lukas Meile, Carol Kälin, Luisa Katharina Schaefer, Michael Schläfli, Victoria Widrig, Andrea Sanchez-Vallet, Thomas Wicker, and Salim Bourras. Domestication of high-copy transposons underlays the wheat small rna response to an obligate pathogen. *Molecular biology and evolution*, 37(3):839–848, 2020.

[276] Natapol Pornputtapong, Daniel A Acheampong, Preecha Patumcharoenpol, Piroon Jenjaroenpun, Thidathip Wongsurawat, Se-Ran Jun, Suganya Yongkiettrakul, Nipa Chokesajjawatee, and Intawat Nookaew. Kitsune: A tool for identifying empirically optimal k-mer length for alignment-free phylogenomic analysis. *Frontiers in bioengineering and biotechnology*, 8:556413, 2020.

[277] Verena M Prade, Heidrun Gundlach, Sven Twardziok, Brett Chapman, Cong Tan, Peter Langridge, Alan H Schulman, Nils Stein, Robbie Waugh, Guoping Zhang, et al. The pseudogenes of barley. *The Plant Journal*, 93(3):502–514, 2018.

[278] Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. Forge: prioritizing variants for graph genomes. *Genome biology*, 19(1):1–16, 2018.

[279] Boas Pucker, Iker Irisarri, Jan de Vries, and Bo Xu. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, 3, 2022.

[280] Wei Quan, Dengfeng Guan, Guangri Quan, Bo Liu, and Yadong Wang. Short read alignment based on maximal approximate match seeds. *Frontiers in Molecular Biosciences*, 7:572934, 2020.

[281] Atif Rahman, Ingileif Hallgrímsdóttir, Michael Eisen, and Lior Pachter. Association mapping from sequencing reads using k-mers. *Elife*, 7:e32920, 2018.

[282] Atif Rahman and Lior Pachter. Cgal: computing genome assembly likelihoods. *Genome biology*, 14(1):1–10, 2013.

[283] Shanjida Rahman, Shahidul Islam, Zitong Yu, Maoyun She, Eviatar Nevo, and Wujun Ma. Current progress in understanding and recovering the wheat genes lost in evolution and domestication. *International journal of molecular sciences*, 21(16):5836, 2020.

[284] Goran Rakocevic, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J Johnson, Vladan Arsenijevic, Jelena Nadj, Kaushik Ghose, Maria C Suciu, et al. Fast and accurate genomic analyses using genome graphs. *Nature genetics*, 51(2):354–362, 2019.

[285] Cláudia Rato, Miguel F Carvalho, Cristina Azevedo, and Paula Rodrigues Oblessuc. Genome editing for resistance against plant pests and pathogens. *Transgenic research*, 30(4):427–459, 2021.

[286] Mikko Rautiainen and Tobias Marschall. Graphaligner: rapid and versatile sequence-to-graph alignment. *Genome biology*, 21(1):1–28, 2020.

[287] Waltram Ravelombola, Jun Qin, Ainong Shi, Qijian Song, Jin Yuan, Fengmin Wang, Pengyin Chen, Long Yan, Yan Feng, Tiantian Zhao, et al. Genome-wide association study and genomic selection for yield and related traits in soybean. *Plos one*, 16(8):e0255761, 2021.

[288] Vivek Kumar Raxwal, Sourav Ghosh, Somya Singh, Surekha Katiyar-Agarwal, Shailendra Goel, Arun Jagannath, Amar Kumar, Vinod Scaria, and Manu Agarwal. Abiotic stress-mediated modulation of the chromatin landscape in arabidopsis thaliana. *Journal of Experimental Botany*, 71(17):5280–5293, 2020.

[289] Ali Raza, Ali Razzaq, Sundas Saher Mehmood, Xiling Zou, Xuekun Zhang, Yan Lv, and Jinsong Xu. Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. *Plants*, 8(2):34, 2019.

[290] Antonio Ribeiro, Agnieszka Golicz, Christine Anne Hackett, Iain Milne, Gordon Stephen, David Marshall, Andrew J Flavell, and Micha Bayer. An investigation of

causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC bioinformatics*, 16(1):1–16, 2015.

[291] Marco Ricci, Valentina Peona, Etienne Guichard, Cristian Taccioli, and Alessio Boattini. Transposable elements activity is positively related to rate of speciation in mammals. *Journal of molecular evolution*, 86(5):303–310, 2018.

[292] Habib Rijzaani, Philipp E Bayer, Mathieu Rouard, Jaroslav Doležel, Jacqueline Batley, and David Edwards. The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*, 15(1):e20100, 2022.

[293] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. Dsk: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653, 2013.

[294] Daniel Rodríguez-Leal, Zachary H Lemmon, Jarrett Man, Madelaine E Bartlett, and Zachary B Lippman. Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, 171(2):470–480, 2017.

[295] Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):1–20, 2013.

[296] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[297] Yana Safonova, Anton Bankevich, and Pavel A Pevzner. dipspades: assembler for highly polymorphic diploid genomes. *Journal of Computational Biology*, 22(6):528–545, 2015.

[298] Birgit Samans, Boulos Chalhoub, and Rod J Snowdon. Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species brassica napus. *The plant genome*, 10(3):plantgenome2017–02, 2017.

[299] David C Samuels, Leng Han, Jiang Li, Sheng Quanghu, Travis A Clark, Yu Shyr, and Yan Guo. Finding the lost treasures in exome sequencing data. *Trends in Genetics*, 29(10):593–599, 2013.

[300] Ayixon Sánchez-Reyes and Maikel Gilberto Fernández-López. Mash sketched reference dataset for genome-based taxonomy and comparative genomics. 2021.

[301] Sinan Saraçli, Nurhan Doğan, and İsmet Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of inequalities and Applications*, 2013(1):1–8, 2013.

[302] Armin Scheben, Jacqueline Batley, and David Edwards. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant biotechnology journal*, 15(2):149–161, 2017.

[303] Sarah-Veronica Schiessl, Elvis Katche, Elizabeth Ihien, Harmeet Singh Chawla, and Annaliese S Mason. The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, 7(2):127–140, 2019.

[304] Mona Schreiber, Nils Stein, and Martin Mascher. Genomic approaches for studying crop evolution. *Genome biology*, 19(1):1–15, 2018.

[305] Alison Dawn Scott, Jozefien D Van de Velde, and Polina Yu Novikova. Inference of polyploid origin and inheritance mode from population genomic data. *bioRxiv*, 2021.

[306] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt Von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6):461–468, 2018.

[307] Enrico Seiler, Svenja Mehringer, Mitra Darvish, Etienne Turc, and Knut Reinert. Raptor: A fast and space-efficient pre-filter for querying very large collections of nucleotide sequences. *Iscience*, 24(7):102782, 2021.

[308] Antonio Serrato-Capuchina and Daniel R Matute. The role of transposable elements in speciation. *Genes*, 9(5):254, 2018.

[309] Lianguang Shang, Xiaoxia Li, Huiying He, Qiaoling Yuan, Yanni Song, Zhaoran Wei, Hai Lin, Min Hu, Fengli Zhao, Chao Zhang, et al. A super pan-genomic landscape of rice. *Cell research*, 32(10):878–896, 2022.

[310] Robert E Sharwood, W Paul Quick, Demi Sargent, Gonzalo M Estavillo, Viridiana Silva-Perez, and Robert T Furbank. Mining for allelic gold: finding genetic variation in photosynthetic traits in crops and wild relatives. *Journal of Experimental Botany*, 73(10):3085–3108, 2022.

[311] Emma N Shipman, Jingwei Yu, Jiaqi Zhou, Karin Albornoz, and Diane M Beckles. Can gene editing reduce postharvest waste and loss of fruit, vegetables, and ornamentals? *Horticulture research*, 8, 2021.

[312] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014.

[313] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

[314] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009.

[315] Rajesh Kumar Singh, Divya Singh, Arpana Yadava, and Akhileshwar Kumar Srivastava. Molecular fossils "pseudogenes" as functional signature in biological system. *Genes & genomics*, 42(6):619–630, 2020.

[316] Jouni Sirén, Erik Garrison, Adam M Novak, Benedict Paten, and Richard Durbin. Haplotype-aware graph indexes. *Bioinformatics*, 36(2):400–407, 2020.

[317] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[318] Jang-il Sohn and Jin-Wu Nam. The present and future of de novo whole-genome assembly. *Briefings in bioinformatics*, 19(1):23–40, 2018.

[319] Jia-Ming Song, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, Bo Wang, Shaoping Lu, Run Zhou, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of brassica napus. *Nature Plants*, 6(1):34–45, 2020.

[320] Xiaoming Song, Yanping Wei, Dong Xiao, Ke Gong, Pengchuan Sun, Yiming Ren, Jiaqing Yuan, Tong Wu, Qihang Yang, Xinyu Li, et al. Brassica carinata genome characterization clarifies u's triangle model of evolution and polyploidy in brassica. *Plant Physiology*, 186(1):388–406, 2021.

[321] Jonathan P Spoelhof, Pamela S Soltis, and Douglas E Soltis. Pure polyploidy: closing the gaps in autopolyploid research. *Journal of Systematics and Evolution*, 55(4):340–352, 2017.

[322] Saurabh Srivastav, Mahino Fatima, Mir Hilal Ahmad, and Amal Chandra Mondal. Impact of crispr-based gene editing in environmental biotechnology. In *Emerging Trends in Environmental Biotechnology*, pages 131–142. CRC Press, 2022.

[323] Michelle C Stitzer, Sarah N Anderson, Nathan M Springer, and Jeffrey Ross-Ibarra. The genomic ecosystem of transposable elements in maize. *PLoS genetics*, 17(10):e1009768, 2021.

[324] Krishna V Subbarao, George W Sundin, and Steven J Klosterman. Focus issue articles on emerging and re-emerging plant diseases. *Phytopathology*, 105(7):852–854, 2015.

[325] Fengming Sun, Guangyi Fan, Qiong Hu, Yongming Zhou, Mei Guan, Chaobo Tong, Jiana Li, Dezhi Du, Cunkou Qi, Liangcai Jiang, et al. The high-quality genome of brassica napus cultivar 'zs 11'reveals the introgression history in semi-winter morphotype. *The Plant Journal*, 92(3):452–468, 2017.

[326] Yanqing Sun, Lianguang Shang, Qian-Hao Zhu, Longjiang Fan, and Longbiao Guo. Twenty years of plant genome sequencing: Achievements and challenges. *Trends in Plant Science*, 2021.

[327] Wing-Kin Sung, Kunihiko Sadakane, Tetsuo Shibuya, Abha Belorkar, and Iana Pyrogova. An o(m log m)-time algorithm for detecting superbubbles. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4):770–777, 2014.

[328] Kanae Takahashi, Kouji Yamamoto, Aya Kuchiba, and Tatsuki Koyama. Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied Intelligence*, 52(5):4961–4972, 2022.

[329] Haibao Tang, Eric Lyons, and Christopher D Town. Optical mapping in plant comparative genomics. *GigaScience*, 4(1):s13742–015, 2015.

[330] Yongfu Tao, Hong Luo, Jiabao Xu, Alan Cruickshank, Xianrong Zhao, Fei Teng, Adrian Hathorn, Xiaoyuan Wu, Yuanming Liu, Tracey Shatte, et al. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*, 7(6):766–773, 2021.

[331] Yongfu Tao, Xianrong Zhao, Emma Mace, Robert Henry, and David Jordan. Exploring and exploiting pan-genomics for crop improvement. *Molecular plant*, 12(2):156–169, 2019.

[332] Adam Thrash, Federico Hoffmann, and Andy Perkins. Toward a more holistic method of genome assembly assessment. *BMC bioinformatics*, 21(4):1–8, 2020.

[333] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[334] Cole Trapnell and Steven L Salzberg. How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5):455–457, 2009.

[335] Isaac Turner, Kiran V Garimella, Zamin Iqbal, and Gil McVean. Integrating long-range connectivity information into de bruijn graphs. *Bioinformatics*, 34(15):2556–2565, 2018.

[336] Charles J Underwood and Kyuha Choi. Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. *Chromosoma*, 128(3):279–296, 2019.

[337] Sandra Unterseer, Eva Bauer, Georg Haberer, Michael Seidel, Carsten Knaak, Milena Ouzunova, Thomas Meitinger, Tim M Strom, Ruedi Fries, Hubert Pausch, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k snp genotyping array. *BMC genomics*, 15(1):1–15, 2014.

[338] Daniel Valenzuela, Tuukka Norri, Niko Välimäki, Esa Pitkänen, and Veli Mäkinen. Towards pan-genome read alignment to improve variation calling. *BMC genomics*, 19(2):123–130, 2018.

[339] Geraldine A Van der Auwera and Brian D O'Connor. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, 2020.

[340] Lucas van Dijk. A bayesian approach to haplotype-aware de novo genome assembly for polyploid organisms. 2017.

[341] Michiel van Dijk, Tom Morley, Marie Luise Rau, and Yashar Saghai. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nature Food*, 2(7):494–501, 2021.

[342] Robert VanBuren, Ching Man Wai, Xuewen Wang, Jeremy Pardo, Alan E Yocca, Hao Wang, Srinivasa R Chaluvadi, Guomin Han, Douglas Bryant, Patrick P Edger, et al. Exceptional subgenome stability and functional divergence in the allotetraploid ethiopian cereal teff. *Nature communications*, 11(1):1–11, 2020.

[343] Carlos M Vicient and Josep M Casacuberta. Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, 2017.

[344] Susana Vinga. Alignment-free methods in computational biology, 2014.

[345] C Vitte and O Panaud. Ltr retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and genome research*, 110(1-4):91–107, 2005.

[346] Kai Peter Voss-Fels, Mark Cooper, and Ben John Hayes. Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3):669–686, 2019.

[347] Jeremiah A Wala, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, et al. Svaba: genome-wide detection of structural variants and indels by local assembly. *Genome research*, 28(4):581–591, 2018.

[348] Sean Walkowiak, Liangliang Gao, Cecile Monat, Georg Haberer, Mulualem T Kassa, Jemima Brinton, Ricardo H Ramirez-Gonzalez, Markus C Kolodziej, Emily Delorean, Dinushika Thambugala, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588(7837):277–283, 2020.

[349] Lei Wang and Qifa Zhang. Boosting rice yield by fine-tuning spl gene expression. *Trends in Plant Science*, 22(8):643–646, 2017.

[350] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.

[351] Peipei Wang, Fanrui Meng, Bethany M Moore, and Shin-Han Shiu. Impact of short-read sequencing on the misassembly of a plant genome. *BMC genomics*, 22(1):1–18, 2021.

[352] Qing Wang, Beiqi Shao, Fayaz Imamrasul Shaikh, Wolfgang Friedt, and Sven Gottwald. Wheat resistances to fusarium root rot and head blight are both associated with deoxynivalenol-and jasmonate-related gene expression. *Phytopathology*, 108(5):602–616, 2018.

[353] Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A Lawson, Julian K Lucas, Adam M Phillippy, Alice B Popejoy, Mobin Asri, Caryn Carson, Mark JP Chaisson, et al. The human pangenome project: a global resource to map genomic diversity. *Nature*, 604(7906):437–446, 2022.

[354] Wensheng Wang, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, Tianqing Zheng, Roven Rommel Fuentes, Fan Zhang, et al. Genomic variation in 3,010 diverse accessions of asian cultivated rice. *Nature*, 557(7703):43–49, 2018.

[355] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[356] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[357] Andrew Watts. *Molecular Breeding of Forage Crops: Proceedings of the 2nd International*. Springer, 2010.

[358] Stephan Weißbach, Stanislav Sys, Charlotte Hewel, Hristo Todorov, Susann Schweiger, Jennifer Winter, Markus Pfenninger, Ali Torkamani, Doug Evans, Joachim Burger, et al. Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines. *BMC genomics*, 22(1):1–15, 2021.

[359] Alejandro Hernandez Wences and Michael C Schatz. Metassembler: merging and optimizing de novo genome assemblies. *Genome biology*, 16(1):1–10, 2015.

[360] Thomas Wicker, Heidrun Gundlach, Manuel Spannagl, Cristobal Uauy, Philippa Borrill, Ricardo H Ramírez-González, Romain De Oliveira, Klaus FX Mayer, Etienne Paux, and Frédéric Choulet. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome biology*, 19(1):1–18, 2018.

[361] Thomas Wicker, Alan H Schulman, Jaakko Tanskanen, Manuel Spannagl, Sven Twardziok, Martin Mascher, Nathan M Springer, Qing Li, Robbie Waugh, Chengdao Li, et al. The repetitive landscape of the 5100 mbp barley genome. *Mobile DNA*, 8(1):1–16, 2017.

[362] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[363] Daniel P Wickland, Gopal Battu, Karen A Hudson, Brian W Diers, and Matthew E Hudson. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, gb-easy. *BMC bioinformatics*, 18(1):1–12, 2017.

[364] Jingrui Wu, Shai J Lawit, Ben Weers, Jindong Sun, Nick Mongar, John Van Hemert, Rosana Melo, Xin Meng, Mary Rupe, Joshua Clapp, et al. Overexpression of zmm28 increases maize grain yield in the field. *Proceedings of the National Academy of Sciences*, 116(47):23850–23858, 2019.

[365] Xing Wu, Christopher Heffelfinger, Hongyu Zhao, and Stephen L Dellaporta. Benchmarking variant identification tools for plant diversity discovery. *BMC genomics*, 20(1):1–15, 2019.

[366] Jianbo Xie, Sisi Chen, Weijie Xu, Yiyang Zhao, and Deqiang Zhang. Origination and function of plant pseudogenes. *Plant signaling & behavior*, 14(8):1625698, 2019.

[367] Jianbo Xie, Ying Li, Xiaomin Liu, Yiyang Zhao, Bailian Li, Pär K Ingvarsson, and Deqiang Zhang. Evolutionary origins of pseudogenes and their association with regulatory sequences in plants. *The Plant Cell*, 31(3):563–578, 2019.

[368] Peng Xu, Yu Chen, Min Gao, and Zechen Chong. Clipsv: improving structural variation detection by read extension, spliced alignment and tree-based decision rules. *NAR genomics and bioinformatics*, 3(1):lqab003, 2021.

[369] Jia-Yu Xue, Yue Wang, Min Chen, Shanshan Dong, Zhu-Qing Shao, and Yang Liu. Maternal inheritance of u's triangle and evolutionary process of brassica mitochondrial genomes. *Frontiers in plant science*, 11:805, 2020.

[370] Won Yim, Mia Swain, Dongna Ma, Hong An, Kevin A Bird, David Curdie, Samuel Wang, Agusto Luzuriaga-Neira, Jay S Kirkwood, Manhoi Hur, et al. The final piece of the triangle of u: Evolution of the tetraploid brassica carinata genome. 2022.

[371] Xiangqin Yu, Dan Yang, Cen Guo, and Lianming Gao. Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. *Plant diversity*, 40(4):158–164, 2018.

[372] Hongjun Yuan, Xingquan Zeng, Qiaofeng Yang, Qijun Xu, Yulin Wang, Dunzhu Jabu, Zha Sang, and Nyima Tashi. Gene coexpression network analysis combined with metabonomics reveals the resistance responses to powdery mildew in tibetan hulless barley. *Scientific reports*, 8(1):1–13, 2018.

[373] Yuxuan Yuan, Philipp E Bayer, Jacqueline Batley, and David Edwards. Current status of structural variation studies in plants. *Plant Biotechnology Journal*, 19(11):2153–2163, 2021.

[374] Yuxuan Yuan, Claire Yik-Lok Chung, and Ting-Fung Chan. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal*, 18:2051–2062, 2020.

[375] Lisna Zahrotun. Comparison jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method. *Computer Engineering and Applications Journal*, 5(1):11–18, 2016.

[376] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[377] Haowen Zhang, Shiqi Wu, Srinivas Aluru, and Heng Li. Fast sequence to graph alignment using the graph wavefront algorithm. *arXiv preprint arXiv:2206.13574*, 2022.

[378] Jisen Zhang, Xingtan Zhang, Haibao Tang, Qing Zhang, Xiuting Hua, Xiaokai Ma, Fan Zhu, Tyler Jones, Xinguang Zhu, John Bowers, et al. Allele-defined genome of the autopolyploid sugarcane saccharum spontaneum l. *Nature genetics*, 50(11):1565–1573, 2018.

[379] Lei Zhang, Xu Cai, Jian Wu, Min Liu, Stefan Grob, Feng Cheng, Jianli Liang, Chengcheng Cai, Zhiyuan Liu, Bo Liu, et al. Improved brassica rapa reference genome by single-molecule sequencing and chromosome conformation capture technologies. *Horticulture research*, 5, 2018.

[380] Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, and C Titus Brown. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PloS one*, 9(7):e101271, 2014.

[381] Xingtan Zhang, Ruoxi Wu, Yibin Wang, Jiaxin Yu, and Haibao Tang. Unzipping haplotypes in diploid and polyploid genomes. *Computational and structural biotechnology journal*, 18:66–72, 2020.

[382] Yingxin Zhang, Chengming Fan, Yuhong Chen, Richard R-C Wang, Xiangqi Zhang, Fangpu Han, and Zanmin Hu. Genome evolution during bread wheat formation unveiled by the distribution dynamics of ssr sequences on chromosomes using fish. *BMC genomics*, 22(1):1–15, 2021.

[383] Zhonghua Zhang, Linyong Mao, Huiming Chen, Fengjiao Bu, Guangcun Li, Jinjing Sun, Shuai Li, Honghe Sun, Chen Jiao, Rachel Blakely, et al. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *The Plant Cell*, 27(6):1595–1604, 2015.

[384] Liang Zhao, Jin Xie, Lin Bai, Wen Chen, Mingju Wang, Zhonglei Zhang, Yiqi Wang, Zhe Zhao, and Jinyan Li. Mining statistically-solid k-mers for accurate ngs error correction. *BMC genomics*, 19(10):1–10, 2018.

[385] Tao Zhao, Arthur Zwaenepoel, Jia-Yu Xue, Shu-Min Kao, Zhen Li, M Eric Schranz, and Yves Van de Peer. Whole-genome microsynteny-based phylogeny of angiosperms. *Nature Communications*, 12(1):1–14, 2021.

[386] Xuefang Zhao, Ryan L Collins, Wan-Ping Lee, Alexandra M Weber, Yukyung Jun, Qihui Zhu, Ben Weisburd, Yongqing Huang, Peter A Audano, Harold Wang, et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*, 108(5):919–928, 2021.

[387] Hong Bo Zhou and Jun Tao Gao. Automatic method for determining cluster number based on silhouette coefficient. In *Advanced materials research*, volume 951, pages 227–230. Trans Tech Publ, 2014.

[388] Huijuan Zhou, Yiheng Hu, Aziz Ebrahimi, Peiliang Liu, Keith Woeste, Peng Zhao, and Shuoxin Zhang. Whole genome based insights into the phylogeny and evolution of the juglandaceae. *BMC ecology and evolution*, 21(1):1–16, 2021.

[389] Peng Zhou, Candice N Hirsch, Steven P Briggs, and Nathan M Springer. Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Molecular plant*, 12(3):410–425, 2019.

[390] Shan-Shan Zhou, Xue-Mei Yan, Kai-Fu Zhang, Hui Liu, Jie Xu, Shuai Nie, Kai-Hua Jia, Si-Qian Jiao, Wei Zhao, You-Jie Zhao, et al. A comprehensive annotation dataset of intact ltr retrotransposons of 300 plant genomes. *Scientific Data*, 8(1):1–9, 2021.

[391] Wanding Zhou, Gangning Liang, Peter L Molloy, and Peter A Jones. Dna methylation enables transposable element-driven genome expansion. *Proceedings of the National Academy of Sciences*, 117(32):19359–19366, 2020.

[392] Tingting Zhu, Le Wang, Hélène Rimbert, Juan C Rodriguez, Karin R Deal, Romain De Oliveira, Frédéric Choulet, Gabriel Keeble-Gagnère, Josquin Tibbits, Jane Rogers, et al. Optical maps refine the bread wheat triticum aestivum cv. chinese spring genome assembly. *The Plant Journal*, 107(1):303–314, 2021.

[393] Tingting Zhu, Le Wang, Juan C Rodriguez, Karin R Deal, Raz Avni, Assaf Distelfeld, Patrick E McGuire, Jan Dvorak, and Ming-Cheng Luo. Improved genome sequence of wild emmer wheat zavitan with the aid of optical maps. *G3: Genes, Genomes, Genetics*, 9(3):619–624, 2019.

[394] Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna Katharina Lau, Sophie Röhling, Jae Jin Choi, Michael S Waterman, et al. Benchmarking of alignment-free sequence comparison methods. *Genome biology*, 20(1):1–18, 2019.

[395] Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18(1):1–17, 2017.

[396] Aleksey V Zimin, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L Salzberg, and James A Yorke. The masurca genome assembler. *Bioinformatics*, 29(21):2669–2677, 2013.

[397] Aleksey V Zimin, Daniela Puiu, Richard Hall, Sarah Kingan, Bernardo J Clavijo, and Steven L Salzberg. The first near-complete assembly of the hexaploid bread wheat genome, triticum aestivum. *Gigascience*, 6(11):gix097, 2017.

[398] Stepanka Zverinova and Victor Guryev. Variant calling: Considerations, practices, and developments. *Human mutation*, 43(8):976–985, 2022.

APPENDICES

APPENDIX A

K-MER SPECTRA FOR COMPARATIVE GENOMICS

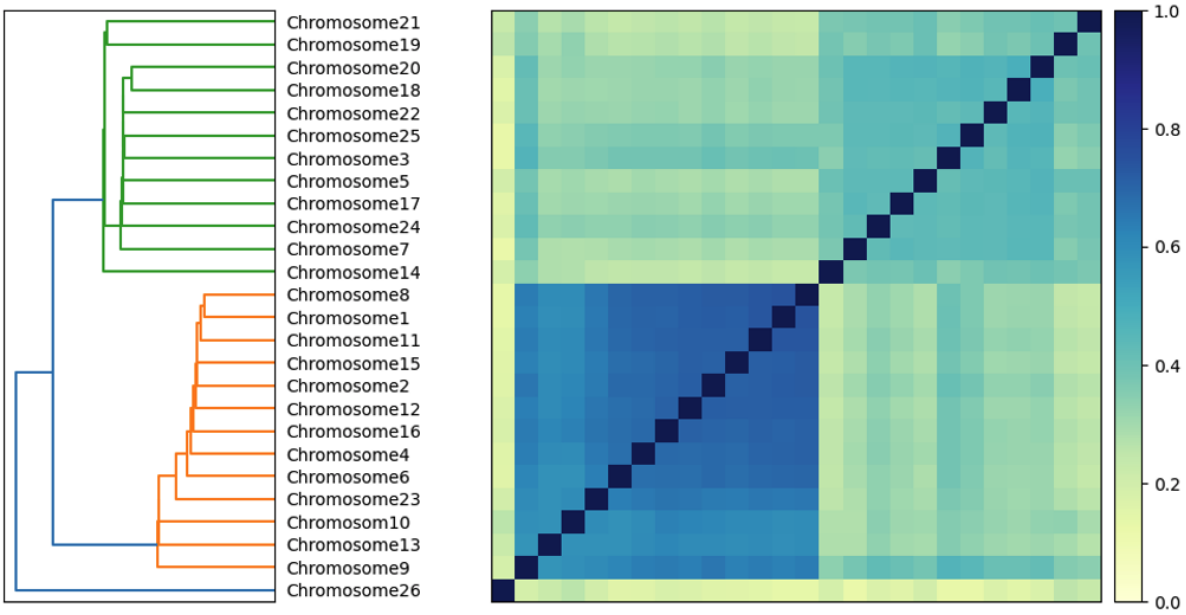


Figure A.1: SourMash produced dendrograms and heatmaps for *G. hirsutum* accession GCF_000987745.1

Table A.1: Whole genome repeat information for all genomes as given in their published works. "NP" means "not provided".

Whole Genome Repeat information table					
Species	NCBI accession	Total repeat percentage	TE percentage	Total LTR percentage	Source
A.hypogaea	GCF_003086295.2	74.3	NP	64.43	[38]
A.durensis	GCF_000817695.2	61.73	NP	52.68	[39]
A. ipaensis	GCF_000816755.2	68.5	NP	57.58	[39]
A.sativa	GCA_022788535.12	64	61.90	61.90	[168]
A.speltoides	GCA_021437245.1	NP	86.13	NP	[200]
A.tauschii	GCA_002575655.1	NP	84.4	65.9	[213]
B.carinata	GCA_016771965.1	58.34	NP	NP	[320]
B.juncea	GCA_015484525.1	45.8	45.8	NP	[257]
B.napus	GCF_000686985.2	49.78	49.78	NP	[325]
B.nigra	GCA_016432835.1	47.12	37.46	29.38	[256]
B.oleracea	GCF_000695525.1	37.20	37.20	8.29	[259]
B.rapa	GCF_000309985.2	37.93	37.51	16.32	[379]
C.arabica	GCF_003713225	NP	NP	NP	NP
C.sativa	GCA_000633955.1	28	22.53	NP	[167]
E.teff	GCA_024500355.1	NP	25.6	20.0	[57]
Ghirsutum	GCA_007990345.1	73	25.6	NP	[63]
Ghirsutum	GCF_000987745.1	67.2	66	NP	[197]
G.tomentosum	GCA_007990485.1	73	NP	NP	[63]
P.virgatum	GCF_016808335.1	NP	56.9	NP	[211]
S.spontaneum	GCA_003544955.1	58.65	54.93	45.62	[378]
Stuberosum	Cooperation-88	NP	NP	NP	[26]
T.aestivum	GCA_900184675.1	NP	85	NP	[154]
T.dicoccoides	GCA_900231445.1	NP	82.2	69.9	[21]
T.turgidum	GCA_003073215.1	NP	82.2	70.0	[216]
T.urartu	GCA_003073215.1	81.42	79.23	42.71	[205]

Table A.2: Subgenome repeat information for those genomes whose publications describe it. Subgenome order in the table is alphabetic.

Subgenome Repeat information table					
Species	NCBI accession	Subgenome repeat percentage			source
A.hypogaea	GCF_003086295.2	73.65	77.71	N/A	[38]
B.juncea	GCA_015484525.1	33.9	51	N/A	[257]
T.aestivum	GCA_002220415.3	85.9	84.7	83.1	[154]

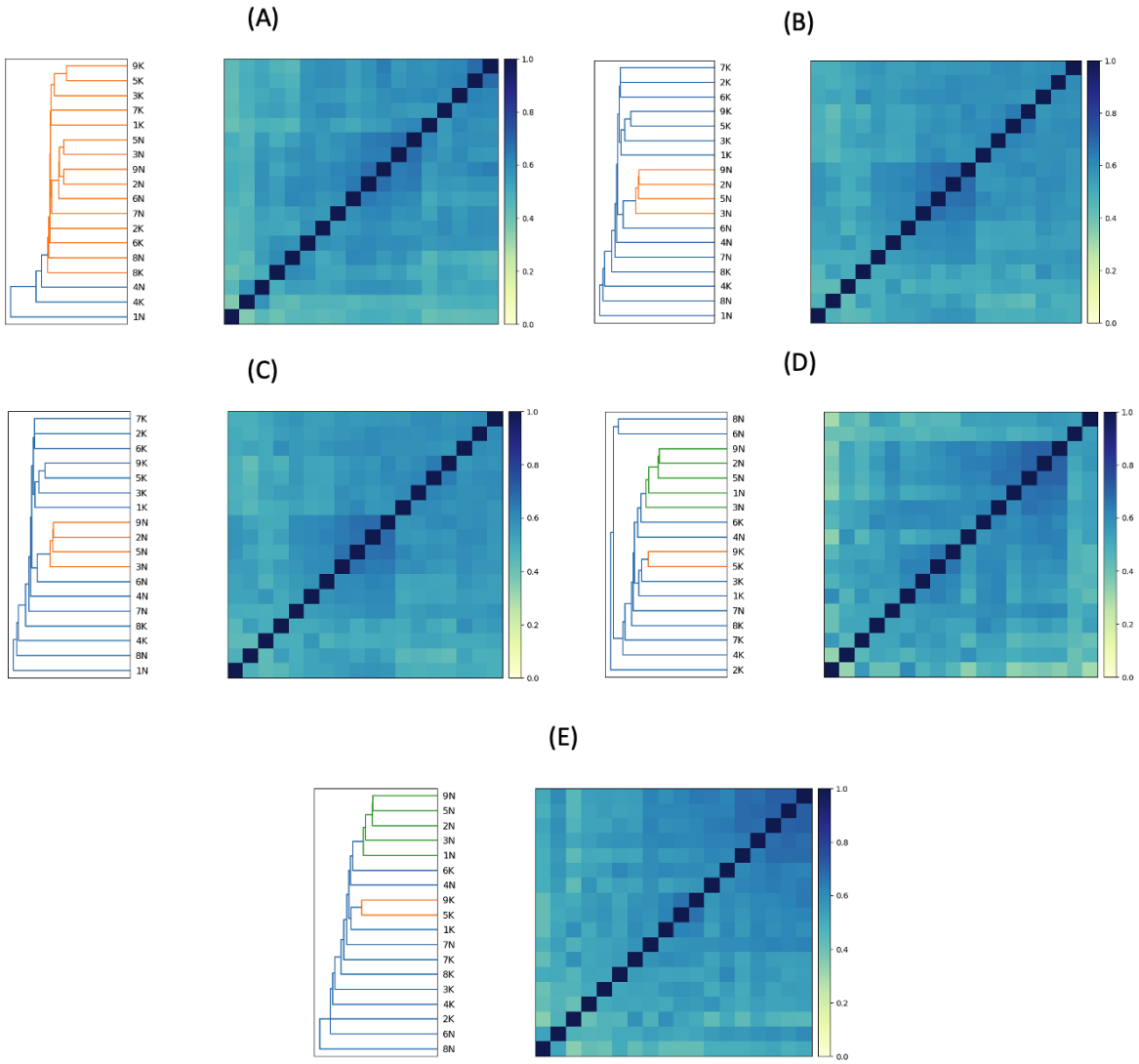


Figure A.2: SourMash produced dendrograms and heatmaps for *P. virgatum* 15-mer frequency for scale factors 1000 (A), 500 (B), 250 (C), 125 (D) and 50 (E).

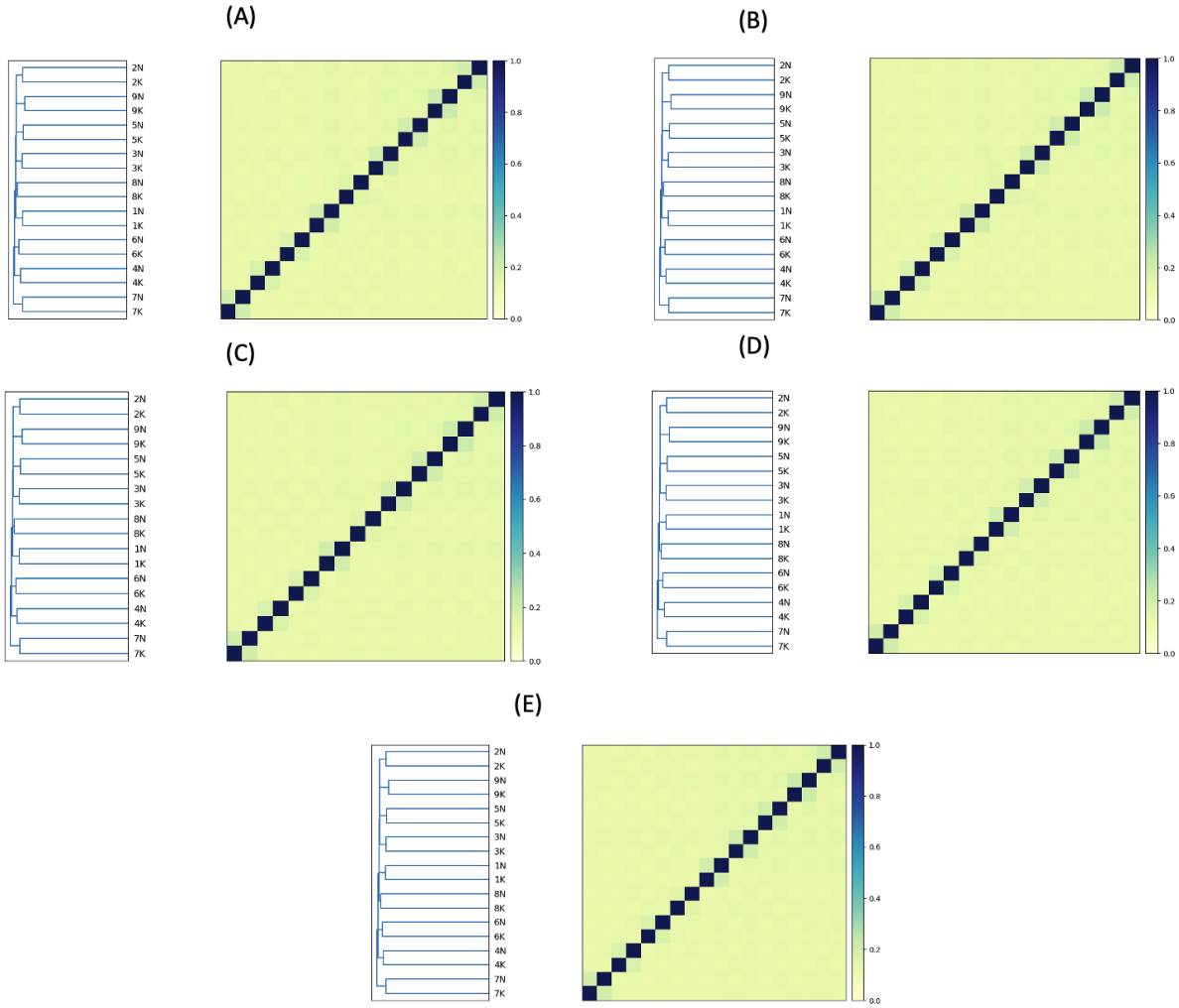


Figure A.3: SourMash produced dendrograms and heatmaps for *P. virgatum* 15-mer composition for scale factors 1000 (A), 500 (B), 250 (C), 125 (D) and 50 (E).

Table A.3: Subgenome hybridisation information.

Subgenome Hybridization information table			
Species	Subgenome hybridization dates		source
A.hypogaea	9,400	N/A	[38]
A.sativa	10.6 mya	7.4 mya	[65, 108]
B.carinata	11,000–29,900	N/A	[370]
B.juncea	11,000–28,600	N/A	[370]
B.napus	5,600–7,000	N/A	[370]
C.arabica	934,000	720,000	[32]
E.teff	1.1mya	5.0mya	[57]
Ghirsutum	4.7–5.2Mya	1.5Mya	[197]
G.tomentosum	4.7–5.2Mya	1.5Mya	[197]
P.virgatum	6.7mya	4.6Mya	[211]
T.aestivum	0.5–0.36Mya	4.6Mya	[263, 382]
T.dicoccoides	0.5–0.36Mya	N/A	[263, 382]
T.tugidum	0.5–0.36Mya	N/A	[263, 382]

Table A.4: Genome sequencing technology and completeness for all species investigated for subgenomic clustering ability. UP refers to unpublished genomes which means sequencing and completeness information cannot be gathered.

Genome sequencing technology and completeness				
Species	Accession	Technology	Completeness	Source
A.hypogaea	GCF_003086295.2	Illumina, Pacbio	99.3	[38]
A.durensis	GCF_000817695.2	Illumina	82 - 95	[39]
A.durensis	GCF_000816755.2	Illumina	86 - 98	[39]
A.durensis	GCF_000816755.2	Illumina	86 - 98	[39]
A.sativa	GCA_021437245.1	Illumina, Nanopore	NP	[168]
A.speltoides	GCA_021437245.1	Illumina, Nanopore	89.34	[200]
A.tauschii	GCA_002575655.1	Illumina, PacBio	1	[213]
B.carinata	GCA_016771965.1	Illumina, Nanopore	94.44	[320]
B.juncea	GCA_015484525.1	PacBio	98.5	[257]
B.napus	GCF_000686985.2	llumina HiSeq2000	82	[325]
B.nigra	GCA_016432835.1	Nanopore	98.73	[256]
B.oleracea	GCF_000695525.1	Illumina, 454	75	[259]
B.rapa	GCF_000309985.2	PacBio	79.73	[379]
C.arabica	GCF_003713225.1	UP	UP	UP
C.sativa	GCA_000633955.1	Illumina, 454	82	[167]
E.teff	GCA_024500355.1	Illumina, PacBio	92.6	[57]
G.hirsutum	GCF_000987745.1	Illumina, PacBio	99.99	[63]
G.hirsutum	GCA_007990345.1	Illumina	89.6 - 96.7	[197]
G.tomentosum	GCA_007990485.1	PacBio, Illumina	1	[63]
P.virgatum	GCF_016808335.1	PacBio, Illumina	99.58	[211]
S.spontaneum	GCA_003544955.1	PacBio, Illumina	93.15	[378]
S.tuberosum	Cooperation-88	PacBio	1	[26]
T.aestivum	GCA_002220415.3	Illumina	94	[154]
T.turgidum	GCA_003073215.1	Illumina	87.08	[216]
T.urartu	GCA_003073215.1	Illumina	98.38	[205]

APPENDIX B

SUBGENOME CLASSIFICATION

Table B.1: Subgenome bloom filter size across error values given in gigabytes (Gb)

Bloom filter size in gigabyes (Gb)			
	Subgenome A	Subgenome B	Subgenome D
K21 e1	0.826	0.923	0.687
K21 e01	1.7	1.9	1.4
K21 e001	2.5	2.8	2.1
K21 e0001	3.3	3.7	2.7
K31 e1	1.2	1.3	0.94
K31 e01	2.3	2.6	1.9
K31 e001	3.5	3.9	2.8
K31 e0001	4.6	5.1	3.7
K41 e1	1.5	1.6	1.2
K41 e01	2.9	3.2	2.3
K41 e001	4.3	3.0	3.4
K41 e0001	5.7	6.4	4.6
K51 e1	1.7	1.9	1.4
K51 e01	3.4	3.7	2.7
K51 e001	5.0	5.5	4.0
K51 e0001	6.7	7.3	5.3
K61 e1	1.9	2.1	1.5
K61 e01	3.8	4.1	2.9
K61 e001	5.6	6.1	4.4
K61 e0001	7.5	8.1	5.8

Table B.2: Subgenomic contig set query time in seconds

	Subgenome A	**Subgenome B**	**Subgenome D**
K21 e1 Non-KBF	1021	1999	3095
K21 e01 Non-KBF	2293	4298	6806
K21 e001 Non-KBF	3790	7271	11561
K21 e0001 Non-KBF	4817	9277	13084
K21 e1 KBF1	1061	2362	3358
K21 e01 KBF1	2303	4530	6970
K21 e001 KBF1	3338	7234	11004
K21 e0001 KBF1	4535	9446	14269
K21 e1 KBF2	1615	3157	4950
K21 e01 KBF2	3302	6520	10288
K21 e001 KBF2	5268	10550	15794

Continued on next page

Table B.2 – continued from previous page

	Subgenome A	Subgenome B	Subgenome D
K21 e0001 KBF2	6439	14547	21958
K31 e1 Non-KBF	1437	2740	4303
K31 e01 Non-KBF	3846	7528	11863
K31 e001 Non-KBF	5637	10311	16257
K31 e0001 Non-KBF	6809	13345	20270
K31 e1 KBF1	1413	2766	4933
K31 e01 KBF1	3028	5954	9395
K31 e001 KBF1	4267	8304	13154
K31 e0001 KBF1	5375	10294	16425
K31 e1 KBF2	2193	4211	6599
K31 e01 KBF2	4149	8002	12925
K31 e001 KBF2	5907	11138	17940
K31 e0001 KBF2	17940	4211	6599
K41 e1 Non-KBF	1850	348	5793
K41 e01 Non-KBF	4079	7718	13278
K41 e001 Non-KBF	5788	11145	18699
K41 e0001 Non-KBF	8256	15695	24215
K41 e1 KBF1	1741	3434	5642
K41 e01 KBF1	3871	7542	12069
K41 e001 KBF1	5523	10193	16713
K41 e0001 KBF1	6919	13801	21731
K41 e1 KBF2	2675	5910	9802
K41 e01 KBF2	5245	10340	16732
K41 e001 KBF2	7645	14539	23722
K41 e0001 KBF2	9849	19433	30767
K51 e1 Non-KBF	2787	4611	7558
K51 e01 Non-KBF	5496	11279	17719
K51 e001 Non-KBF	8056	14350	23638
K51 e0001 Non-KBF	9920	19049	30378
K51 e1 KBF1	2131	3908	6553
K51 e01 KBF1	4218	8171	13495
K51 e001 KBF1	6802	11348	18633
K51 e0001 KBF1	7857	16390	25437
K51 e1 KBF2	3585	6484	11308
K51 e01 KBF2	6526	11792	19737
K51 e001 KBF2	8384	16511	27289
K51 e0001 KBF2	10985	21730	35552
K61 e1 Non-KBF	2710	5644	8993
K61 e01 Non-KBF	6376	11989	19821
K61 e001 Non-KBF	9106	17521	28170

Continued on next page

Table B.2 – continued from previous page

	Subgenome A	**Subgenome B**	**Subgenome D**
K61 e0001 Non-KBF	11739	22350	35833
K61 e1 KBF1	2642	5386	9072
K61 e01 KBF1	4683	9039	14562
K61 e001 KBF1	6880	12666	21872
K61 e0001 KBF1	10448	16442	7492
K61 e1 KBF2	3716	7429	12199
K61 e01 KBF2	6565	14410	23780
K61 e001 KBF2	10684	17083	27956
K61 e0001 KBF2	11638	22051	36445

APPENDIX C

VARIANT IDENTIFICATION

Supplementary method 1 Figure 4.3 was generated using *Triticum aestivums* IWGSC version 2 downloaded from NCBI genbank. K-mers were counted using KMC3 [179] and the unique proportion was calculated using the reported distinct k-mers and total counted k-mers output by the KMC counting procedure. The relationship between the proportion of unique genomic k-mers and k-mer size was plotting using ggplot2 [362] on R Studio 3.6.1(2019-07-05) [296].