Short Read Sequencing Reveals Sub-Genome Structure of the Polyploid Pennate Diatom *Fragilariopsis cylindrus*

Kat Amy Hodgkinson

A Thesis for the degree of Master of Science by Research

University of East Anglia, Norwich, UK

School of Environmental Sciences

August 2018

Abstract

Fragilariopsis cylindrus is a species of pennate diatom capable of survival in the extreme, variable and fluctuating environmental conditions of polar oceans, which hints at genome plasticity. Here, we use Illumina short-read sequencing and assembly to reveal new structural details of the *F. cylindrus* genome. Our assembly methods and analysis based on k-mer spectra ensures the inclusion of as much content as possible. We highlight the level of missing content from previous assemblies using the Illumina reads and assembly as a comparison. We have confirmed this culture of *F. cylindrus* (Grunow) Krieger CCMP1102 to have a typical k-mer spectra signature of a polyploid genome. K-mer spectra plots show three clear peaks created from the frequencies of distinct k-mers found within the reads. Our sub-genome size estimate of 50.7Mb by k-mer spectra is within a mutually complimentary range of previous findings based on qPCR of target genes at 57.9Mb (± 16.9Mb). Estimates based on k-mer frequency suggest a lower-bound 36Mb total diverged content between A, B and B' sub-genomes. Local analysis of three fairly contiguous unitigs show an average of only 92% identity. Our future aim is to produce a fully-phased assembly of each sub-genome. We will use this fully phased assembly for divergence analysis, as well as more detailed analysis into gene content and transcriptome expression. We expect the assembly to be improved with scaffolding using PacBio reads. We have only sequenced one culture of *F. cylindrus,* so cannot conclude whether this is a widespread phenomenon or a product of laboratory culturing. This will become more apparent with future sequencing of *F. cylindrus* cultures. The implications of our findings may make *F. cylindrus* a very interesting model for evolutionary studies.

Table of Contents

List of Tables

List of Figures

Glossary

Genome: The full set of chromosomes or the complete set of genomic material within an organism.

Sub-genome: A diploid-pair subset of a polyploid, where the genomes of the sub-genome are identical copies; ie., genomes AA, BB and CC would be sub-genomes A, B and C. Used in the context of comparing entire diploid pairs of genomes rather than local haplotypes. As used in Bird *et al*. (2018), Ling *et al*. (2018), The International Wheat Genome Sequencing Consortium (IWGSC) *et al*. (2014).

Haplotype: A local set of haploid genes of a single chromosomal region.

Genomic plasticity: Our definition expands on the traditional alterable nature of Prokaryotic genomes by DNA exchange to include genotypic variation or chromosomal dosage changes, resulting in phenotypic diversity of Eukaryotes. Used in Leitch and Leitch (2008) and Sterkers et al. (2012).

K-mer: A word of k length within a sequence. The sequence is split into k-length fragments through its entirety, with every sequential k-mer containing k-1 of the previous nucleotides plus the next new nucleotide in sequence.

Distinct k-mer: K-mer that occurs at least once in the sequence.

Unique k-mer: K-mer that occurs only once in the sequence.

Unitig: The first stage of assembly, whereby overlapping k-mers are joined with no ambiguities. Unitigs are all unique within the assembly graph.

Contig: A consensus of unitigs with connecting edges through the assembly graph, each contig has encountered at least one level of structuring. Where nodes have multiple edges, the graph walks through one of these and discards the other based on user-defined thresholds (eg., coverage).

Scaffold: A contiguous set of contigs that contain gaps of known length in between.

"Bubble": The links formed on an assembly graph where the path splits from one node into two alternative nodes, which come together on their latter sides in another single node.

N50 value: A length of sequence such that 50% of the assembled content is included on sequences of this, or grater, length. An N50 of 1Kb means that 50% of the assembly is contained in sequences of 1Kb or larger. A high N50 indicates long contigs and a low N50 indicates a fragmented assembly. N50 is not a metric of assembly quality and is easily biased.

Acknowledgements

Firstly, I would like to thank my bioinformatics research group for their endless help and support throughout this project and their incredibly welcoming environment. Thank you to Jon, Gonza and Luis, who all provided me with invaluable computing assistance and advice from the beginning to help me get to grips with bioinformatics. Thanks to Cam and Ben for their wisdom and encouragement in completion of this MSc. And I would like to extend special thanks to Bernardo Clavijo, who has guided me through every stage of this project and who has offered endless support and advice for its duration. I could not have completed this research without all of you.

I wish to further thank my primary supervisor, Professor Cock van Oosterhout, for all his support, encouragement, understanding, and for giving me the opportunity to carry out this project. Thanks to Professor Thomas Mock for his generous donation of new data and inclusion into his lab group, and thanks to his lab group for the diatom discussions.

Finally, I would like to thank my partner, James, for his encouragement through this past year; Pwmpen, Dinari, LittleBoy and Blarney, for providing me with my will and resolve, and Sam and Neddy, for teaching me patience and galloping through my stress.

1. Introduction

1.1 The Study Organism

1.1.1 Diatoms

Diatoms are single celled, marine phytoplankton, distributed across the Northern and Southern hemisphere (Malviya *et al*., 2016). Diatoms are known to rapidly dominate all other forms of phytoplankton in nutrient-rich coastal and polar regions (Amato *et al*., 2017; Smetacek, 2012; Thomas *et al*., 1978) by explosive bloom formations when conditions become favourable. The exponential increase in biomass during these opportunistic bloom episodes allow diatoms to contribute 20% of the total primary production on earth (Field *et al*., 1998; Nelson *et al*., 1995). One key factor in the limitation of bloom formation is the restricted availability of silica, an essential nutrient for cell wall formation. This silica frustule is a characteristic feature of diatoms, responsible for variety in their distinctive shapes and patterns.

The majority of diatom reproduction is asexual (Edlund and Stoermer, 1997; Mann *et al*., 2003), allowing many species to rapidly form colonies. Continual mitotic divisions in some species are limited in number by the shrinking of the frustule in each successive generation (Edlund and Stoermer, 1997), and sexually reproduced progeny are restored to full frustule size (Armbrust and Chisholm, 1992). Cell size reaching a minimum threshold is one major factor in triggering sexual reproduction (Mann *et al*., 2003), but larger or smaller cell size triggers may be determined by whether reproduction is intra- or inter-clonal (Chepurnov and Mann, 2000). Extrinsic factors have also been shown to prompt sexual reproduction in some species, such as light exposure (Mouget *et al*., 2009). Other species are more complicated in their initiation of sexual reproduction, entering a sexual phase only when introduced to compatible strains under the correct circumstances (Fuchs *et al*., 2013) or conversely, under stress (Edlund and Stoermer, 1997).

It is thought that pennate diatoms require a dormancy stage before germination (von Stosch and Fecher, 1979). Dormancy is achieved through resting spores, which may be prompted by stress (Guppy and Withers, 2007; Lennon and Jones, 2011). These resting spores act as a seed bank for genetic diversity through time and can be germinated even after more than 100 years of dormancy (Ellegaard *et al*., 2018; Harnstrom *et al*., 2011). Little is known about how this may influence genetic diversity, but it is hypothesised that stress induced dormancy may be a mechanism for self-preservation of clonal colonies in less favourable conditions (Ellegaard *et al*., 2018).

1.1.2 Diatom genomics

Environmental stressors can change the forms of diatoms (Falasco *et al*., 2009), altering striation pattern and valve outline. Many physical and physiological observations on the effects of stress have been recorded (Bayer-Giraldi *et al*., 2010; Edlund and Stoermer, 1997; Mock and Hoch, 2005; Sarthou *et al*., 2005). However, the effects of stressors on the genome remain largely unknown, such as how stress is related to genetic diversification. Polyploidy in other organisms may infer an adaptive advantage as it is often found to occur at increased frequencies in disrupted or harsh environments (Ramsey, 2011; Van de Peer *et*

al., 2017). The production of polyploids in diatoms may confer an adaptive advantage in extreme environments, as can be seen in angiosperms (Diallo *et al*., 2016).

A number of studies have documented the chromosome number of limited diatom species (Geitler, 1973; Kociolek and Stoermer, 1989). Records gathered by Geitler (1973), and then later by Kociolek and Stoermer (1989), show many diatom species to have diploid genomes. Interestingly, the number of recorded chromosomes varies for both centric and pennate diatoms, with some intraspecific variety (although, notes Kociolek and Stoermer (1989), this may be due to difficulties in resolving chromosomes during meiotic division). However, karyotyping for diatoms is difficult to produce due to structural complications with chromosome extraction, including the presence of the protective silica frustule, which prevents access to the delicate nucleus (Geitler, 1973). With the advent of High Throughput Sequencing (HTS) came a shift from methods of observing genome organisation to gene function, so ploidy level estimation was often based on old karyotyping. In plants, historical karyotyping estimates suggested that 30-50% of angiosperms were polyploids. This was realised as a gross underestimate when modern genomic studies became available, and we now recognise almost all angiosperms to have experienced a polyploid episode in their evolutionary history (Cui *et al*., 2006).

Polyploidy has been observed many times in diatoms and speculation points to polyploidy as a key mechanism in genome divergence and evolution. Geitler (1973) observed triploid zygotes in two species, *Gomphonema parvulum* and *Cocconeis placentula*, and Mann (1994) commented on polyploid auxospores in *Navicula ulvacea*. Von Dassow *et al*. (2008) reported to have observed polyploids in the genus *Thalassiosira,* owing to aberrant meiotic and mitotic cell divisions. Furthermore, Chepurnov *et al*. (2002) has also recorded polyploid auxospores for the *Seminavis* diatom, concluding the importance of changes in ploidy as a mechanism of evolution. Additionally, polyploidy has been observed with the use of other methods; flow cytometry was used by Koester *et al*. (2010) on the *Ditylum brightwellii* genome, uncovering some two-fold differences in DNA content between populations. Many polyploids have been identified in stressful and extreme environments, lending to the hypothesis that doubling the genome confers an advantage to polyploids over their diploid progenitors (Madlung, 2013; Van de Peer *et al*., 2017).

Although we are lacking in genomic data for all diatoms, Parks *et al*. (2017) pulled together data for 37 phylogenetically diverse species for Whole Genome Duplication (WGD) event detection and concluded that WGD is a driver of diatom evolution. This is paralleled in plants by evidence of polyploid success after great extinction events (Soltis and Burleigh, 2009). WGD as a driver of evolution has long been hypothesised in plants; the resulting polyploidy is expected to be a contributing factor in major bursts of plant diversification (Soltis *et al*., 2009). A surge in the frequency of polyploid plants can be traced back to the K-T mass extinction, catalysing the survival of many plant lineages (Soltis and Burleigh, 2009). Given the stressful environments that diatoms endure, polyploidy could prime them for greater adaptive potential.

1.1.3 *Fragilariopsis cylindrus*

Fragilariopsis cylindrus is a species of highly abundant pennate diatoms native to polar oceans (Kang and Fryxell, 1992). They occupy a niche in shallow oceanic depths, facing inclusion into sea ice. In ideal conditions, *F. cylindrus* radically outcompetes other organisms in sea-ice melts (Alderkamp *et al*., 2012). Environmental adaptations have been investigated under experimental conditions and show that *F. cylindrus* is capable of survival in an inconsistent environment. This species is well documented for its persistence in the extreme polar environment, facing fluctuating temperatures, including freezing at -1.8 °C to abnormally warm at 10°C (Mock and Hoch, 2005). Additionally, it must face high salinity (Bayer-Giraldi *et al*., 2010; Mock and Hoch, 2005), typical of polar sea ice. *F. cylindrus* has been shown to maintain photosynthetic capabilities most efficiently in shallow mixed layers (Kropuenske *et al*., 2010, 2009). In this environment, irradiance is consistently high and Iron (Fe) concentrations rise above average. Experiments have shown *F. cylindrus* can even persist under overabundant UV exposure (Helbling *et al*., 1996). Furthermore, the potential to adapt to an acidifying pH (Pančić *et al*., 2015), which is occurring as a result of fossil fuel consumption, demonstrates the capability of *F. cylindrus* to acclimatise to anthropogenic changes. This adaptability to the broad, and often extreme, range of polar ocean ecosystem conditions hints at plasticity of the *F. cylindrus* genome.

1.1.4 F. cylindrus genomics

Although *F. cylindrus* is well studied for a diatom, genomic data remains limited relative to other well-studied organisms, reflecting previous difficulties in genomic observations in all diatom species. Karyotyping records for *F. cylindrus* are difficult to find, and what we speculate about the genome is largely based on pennate relatives (Geitler, 1973; Kociolek and Stoermer, 1989). Two previous assemblies have been constructed (Mock *et al*., 2017), but no genome structural information or haplotype-resolved assembly has been produced. The clarification of genomic details requires further genome structure exploration techniques to be employed. Flow cytometry has been successfully used to estimate genome structure in plants (Doležel *et al*., 2007; Xu *et al*., 2014) and in the diatom, *Ditylum brightwellii* (Koester *et al*., 2010) but has not been employed for *F. cylindrus*.

An alternative method of ploidy estimation is with analysis of whole genome sequences. Two methods are employed to estimate ploidy from high-throughput data; by sequence polymorphisms between aligning reads, or by k-mer analysis. The former method was employed by Armbrust (2004) on a single individual of a distantly related centric diatom, *Thalassiosira pseudonana*, which was found to be diploid. However, this method falls short in repetitive genomes and where sub-genomes are not divergent. Alternatively, tools such as nQuire may estimate ploidy levels. nQuire boasts the ability to distinguish diploids, triploids, tetreploids and aneuploids (Weiß *et al*., 2017), however, a reference genome is required, so is not applicable to *de novo* sequencing. Additionally, using read mapping to generate base frequencies in this way is a more convoluted approach to that of k-mer frequency analysis (see Section 1.2), but with many more fallacies from a dependence on an accurate reference and mapping heuristics. K-mer frequency analysis provides an assembly-independent method of assessing genome structure and characteristics (Liu *et al*., 2013; Mapleson *et al*., 2016; Vurture *et al*., 2017). This method allows assembly and downstream analysis to be tailored to the characteristics of the genome. However, few whole genome sequences of diatoms have been produced thus far. The difficulties in *de novo* assembly lie

with the appropriate use of available sequencing technologies and methods of genome assembly. All sequencing technologies produce errors to various degrees and are not suitable for all applications. In addition, the accuracy of any computational algorithms depend entirely on their intended purpose and the user-defined thresholds set (Batzoglou *et al*., 2002; Chin *et al*., 2016; Clavijo *et al*., 2017; Jaffe *et al*., 2003; Zerbino, 2010).

Pacific Biosciences (PacBio) produce long reads for an increase in sequence contiguity, making them advantageous in hybrid assemblies. This technology may easily resolve smaller genomes, such as the *F. cylindrus* genome (Mock *et al*., 2017), by producing reads up to 50Kb long. However, PacBio long reads are shown to contain ~15% errors (Hackl *et al*., 2014; Pootakham *et al*., 2017). Assemblies based on PacBio sequencing alone have been shown to underrepresent content when compared to hybrid assemblies (Koren *et al*., 2012; Rhoads and Au, 2015; Zimin *et al*., 2017). The success of the hybrid approach is attributed to the utilisation of the higher coverage and higher base-pair accuracy of short-reads, and expanse of read size in base pairs of long reads (Koren *et al*., 2012; Zimin *et al*., 2017). This approach provides required accuracy when used for detailed studies on structural variants; Chaisson *et al*. (2018) employed the use of both Illumina short-reads and Pac-Bio long reads as part of their multi-scale mapping and sequencing strategy to discover human genome haplotype-resolved structural variants. In another study on the Korean human genome, aimed at identifying structural variation, Seo *et al*. (2016) combined the Falcon assembler and PacBio long-reads with high coverage Illumina short-reads for error correction. This method is more reliable for *de novo* genome assembly, giving greater statistical power to later downstream analysis on gene expression than one which employs a solely PacBio assembly (Koren *et al*., 2012; Rhoads and Au, 2015; Zimin *et al*., 2017).

The first *F. cylindrus* genome was sequenced by Mock *et al*. (2017) using Sanger methods and the ARACHNE assembler, (v.20071016; Batzoglou, 2002; Jaffe, 2003). However, no structural analysis of the *F. cylindrus* genome was produced before assembly. Arachne was tested on the assemblage of diploid Eukaryote genomes, and the version used was adapted to better assemble small mammalian genomes. The algorithms it used were therefore unable to detect ploidy levels, running on the assumption of diploidy. This would allow for a loss of genomic content outside of the diploid construct. A further assembly was produced by Mock *et al*. (2017) using PacBio long reads and the FALCON assembler (v.0.3.0, Chin *et al*., 2016). However, the Falcon assembler was developed primarily for assembling the human genome and thus, the algorithms assemble a diploid construct irrespective of the content given.

Sanger sequencing is considered 99.9% accurate (Shendure and Ji, 2008) granting fosmids and original libraries used to create and validate the two *F. cylindrus* assemblies high reliability. The fosmid library was prepared from the same cultures as both Sanger and PacBio assemblies (Mock *et al*., 2017), so would be reasonably expected to confer >99.9% identity (Alexeyenko *et al*., 2014; Du *et al*., 2017). A single finished Sanger haplotyped fosmid library was used for validation of the PacBio-Falcon assembly (Paajanen *et al*., 2017), which covered 0.66% of the assembly (453Kb, 13 fosmids). Of these aligned fosmids, >50% aligned with <99.9% accuracy. A total of 15% of the 13 fosmids aligned with <99% accuracy (98.8% and 98.5%). In a study by Alexeyenko *et al*. (2014), a combined ~40Gb of the highly complex and repetitive Norway spruce genome was sampled from fosmid pools, obtaining two-fold genome coverage; Du *et al*. (2017) used a pooled total of 39.7Gb of fosmid sequence for the 390.3Mb rice genome, conferring

better sequence contiguity with hits >99% identity to the genome. These two studies demonstrate more robust quantities and percent identities of fosmid alignments used for a thorough validation. During Whole Genome Shotgun (WGS) Sanger sequencing of *F. cylindrus*, a number of other libraries were prepared, including a 2.5Kb insert size paired-end library, a 6Kb insert size paired-end library and the 35Kb fosmid library (Mock *et al*., 2017). A more extensive validation could have been achieved with the addition of these extra paired-end libraries and stricter parameters, but this stricter validation was possibly not undertaken for the PacBio-Falcon assembly because results seemed to back the diploid assumption relatively well.

The Sanger assembly estimated the haplotype-resolved genome to consist of 46Mb "collapsed" and 15.1Mb "diverged" haplotypes (totalling 61.1Mb "primary" and 15.1Mb "alternate"; Mock *et al*., 2017). The Pacbio assembly estimated 59.7Mb "primary" and 9.1Mb "alternate" contigs (Paajanen *et al*., 2017). The Falcon assembler defines "primary contigs" as a single path with suitable read support (a threshold set by the user) through the assembly graph, which is used to represent both loci in the assembly; "alternate contigs" are paths through the assembly graph that have less support (from reads) but still meet the support threshold to be considered as a possible secondary sequence for the same locus as a "primary contig". Whilst the primary estimates for both the Sanger-Arachne and PacBio-Falcon assemblies are very similar, the alternate estimates are 6Mb different (66% of the total PacBio estimate in difference). The Sanger assembly is also noted to contain partial diploidy from gene mapping results, arising from some alternate sequence being included in the haploid assembly (Mock *et al*., 2017). However, no mention is made of the level of partial diploidy in the haploid genome and it is not clear to what extent this will have affected results. A larger amount of content in the alternative Sanger contigs and the knowledge of partial diploidy in the primary contigs suggests the PacBio estimate falls short of the true level of alternative content. This may have had an effect on alternative allele identification because the full set of diverged genes may not have been identified on the assembly. Additionally, 335 out of 2 635 (12.7%) full length genes on the longest PacBio contigs, or 21 066 (1.6%) gene predictions from the haplotype resolved Sanger genome were confidently identified as having hits to both alternating haplotypes in the PacBio-Falcon assembly. These figures may underestimate alternative alleles if gene identification was not complete. Fosmids have been successfully used to detect chimeric contigs in the rice genome, which would otherwise confound detection of structural variation (Du *et al*., 2017). The existing fosmid library and Sanger paired-end libraries (Mock *et al*., 2017) can be used in this way, to ensure correct assembly of each nuclear genome and prevent the misassembly of alternative regions to the same primary haplotype.

The existing Sanger and PacBio assemblies share a similar haploid genome size (61.1Mb Sanger; Mock *et al*., 2017; 59.7Mb PacBio; Paajanen *et al*., 2017). However, discrepancies between the two existing genome assemblies of the exact same culture suggest that they are incomplete. The Sanger based diploid assembly was estimated to be 80.5Mb (including 4.3Mb scaffolding gaps) in length, but the PacBio based diploid assembly (68.8Mb) was reported to be missing 14.5% of the Sanger assembly length. Furthermore, only 84.3% of the PacBio assembly is represented in the Sanger assembly, and only 69.8% of the Sanger assembly is accounted for in the PacBio assembly. With such discrepancies between the two existing assemblies, confirmation of genome size, structure and re-validation of gene mapping can be produced with independent Illumina short-read sequencing.

It was previously deduced from transcriptome profiling of *F. cylindrus* that the genome contains differentially expressed, highly divergent alleles (Mock *et al*., 2017), which assist the diatom in adapting to a perpetually changing environment. Differential allelic expression is difficult to validate due to a number of conditions; foremost, the assembly needs to be accurate in order to identify alleles and map reads, and enough coverage is needed to distinguish signal from noise. In other studies aimed at the detailed analysis of structural variants, success was achieved using a hybrid assembly approach (Chaisson *et al*., 2018; Seo *et al*., 2016).

The combination of high-accuracy, high-coverage short-reads with the use of long-reads grants greater statistical power to detect true variants and distinguish from noise, while retaining essential structural integrity and resolution of repetitive regions. Additionally, a low level of coverage will be a further problem if the organism was not diploid as expected, so structural characteristics need to be identified.

1.1.5 Aims

In this study we aim to construct a more robust and complete *de novo* hybrid assembly for later gene expression analysis. We expect the combination of Illumina short-reads and Sanger 8Kb paired-end library linkage to create a much more reliable assembly, with the future aim of incorporating PacBio long-reads for increased contiguity. We will contrast this against the existing assemblies using k-mer spectra analysis and provide a further genome size estimation based on our findings. We will uncover more details into the structure and divergence of the *F. cylindrus* genome. We hypothesise the incompleteness of existing assemblies has affected previous results on gene expression, and re-analysis using our improved assembly will change and enhance our understanding of the gene expression and evolution of *F. cylindrus.*

1.2 Key Themes in Bioinformatics Methods

1.2.1 K-mers and k-mer spectra

K-mer analysis is a statistically robust way of estimating genome size and determining genomic characteristics (Simpson, 2014). A k-mer is a word of k length within a sequence; words are quantified from the sequencing reads, irrespective of their location within the genome. When k-mers are computationally defined, the sequence is split into k-length fragments through its entirety. No level of assembly is necessary, so k-mer analysis can be used without a reference and escapes assembly biases (Liu *et al*., 2013; Mapleson *et al*., 2016). The number of times each distinct k-mer is counted will depend on how often it occurs in the genome, the read coverage and whether the k-mer contains a sequencing error. The net k-mers containing errors will depend on the sequencing quality, and the rate they occur will depend on the total number of k-mers. Additionally, k-mers will be observed more often if they appear on 2 or more sub-genomes than k-mers that occur on a single haploid. As such, this analysis is limited by depth of k-mer coverage and sequence error rates.

A spectrum is defined as a distribution of frequencies, therefore a k-mer spectra is a distribution of k-mer frequencies. Simpson (2014) explains the mathematical workings of the k-mer count distribution in detail; briefly, a zero-truncated Poisson distribution (because

k-mers at a rate of 0 are not observed) is a model that illustrates the probability of observing a k-mer x times for one component (Lander and Waterman, 1988; Simpson, 2014). This is visualised as a histogram (Figure I) whereby the count of words (y-axis) is plotted against their frequency (x-axis). The distribution across the graph gives indications about the structure of the genome; unique content will appear in the fundamental frequency (first peak), content that appears multiple times will appear as a harmonic series of frequencies (x * n) from the fundamental frequency (the first peak), and any highly repetitive regions will appear as a trail after the main body of the graph. Thus the graph can be used to gain initial insight into the size, ploidy levels, content structure and contamination of the genome (Liu *et al*., 2013; Mapleson *et al*., 2016; Simpson, 2014).

The k-mer spectra is a useful means of analysis and comparison when constructing *de novo* genomes (Mapleson *et al*., 2016). Post-assembly, the k-mer spectra can be used to provide a metric of comparison between the assembly and HTS reads (Figure II). This highlights loss of content, errors in copy number and misassembled junctions. Comparing the k-mer spectra between assemblies, it can be used to inform on content agreement

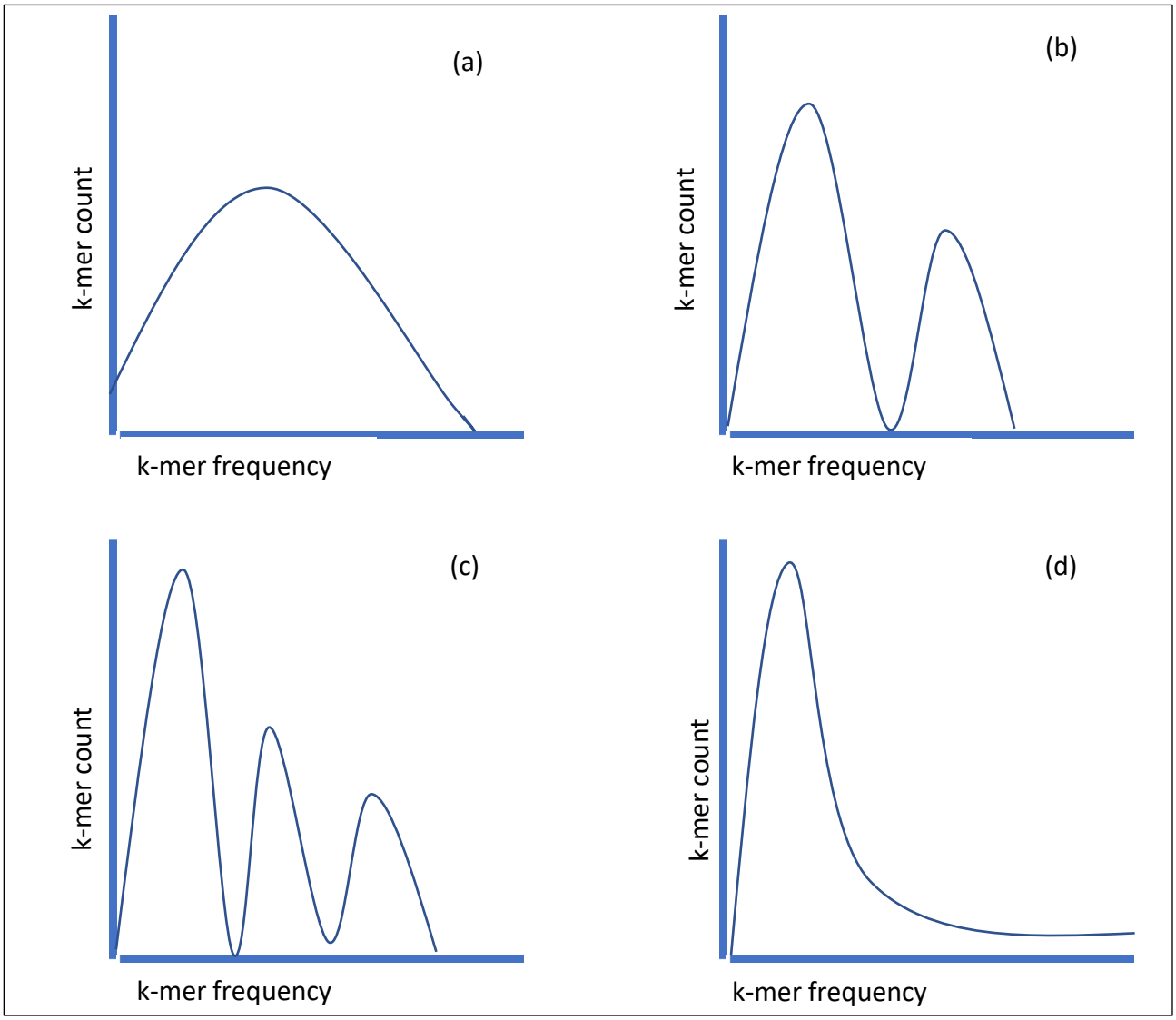

Figure I: Representation k-mer spectra histograms showing a range of distributions commonly seen in real genomic reads. The distribution of (a) illustrates a haploid genome; all the content of this genome is contained within a single symmetrical Poisson distribution. The estimated k-mer coverage is the frequency of the mode k-mer count, in this case, the centre of the peak. This represents, on average, how many times a given k-mer is seen within the distribution (coverage). An important point to address is that, since k-mer frequencies represent relative abundancies, a completely homozygous polyploid genome would produce the same signature.

The distribution of (b) gives two distinct peaks, as typical of a diploid genome. The first peak consists of unique k-mers as we saw in (a), heterozygous content that only appears once in the genome. The second peak consists of k-mers that appear in the genome twice as frequently as the first peak, representing homozygous content. This second peak will be roughly 2x the coverage of the first peak; these k-mers are seen twice as frequently. Eg, if the coverage of the first peak was 25 and the count at this point was 100K, and the coverage of the second peak was 50 with a count of 80K, we could say that 100K distinct k-mers were seen 25 times in the spectra but 80K distinct k-mers were seen 50 times in the spectra. In the WGS reads, the homozygous content will have been sampled roughly twice

Continued overleaf…

as many times as the heterozygous content because it appears in the physical sample twice as often. Continuing from the polyploid point, because the frequencies represent relative abundancies, this signature is also representative of a tetraploid, where each sub-genome pair is homozygous, but heterozygous to the other sub-genome. Each one of these peaks represents a Poisson distribution of frequencies in itself, because some parts of the physical genome will have been sampled more than others in the sequencing process; not all sections of the genome are represented equally by sequencing in a sequential order, so some content is sequenced more or less than others.

The distribution of (c) builds on this with a third peak, representing a triploid genome. The first peak, as before, represents k-mers that appear only once in the genome as heterozygous content. The second peak represents content that appears in the genome twice as often as the first peak, but is included in only two of the three haploid genomes. The third peak is entirely homozygous content that appears on all three haploid genomes. If the mode frequency of this third peak was 75 with a count of 70K, we could say that 70K distinct k-mers from the sequencing reads appear in the spectra 75 times. Again, this distribution may also be signature of a hexaploidy, where each sub-genome pair is homozygous, but heterozygous between the three sub-genomes.

The final plot (d) illustrates a haploid genome with a high level of repetitive content. We can see the same bell-shaped distribution frequency of the (a) plot, but with a long, trailing tail. Repetitive content does not appear as a distinct peak, as by its nature, repetitive content does not appear a precise number of times in the genome. In the physical DNA sample, multiple copies of an entire set of chromosomes is ploidy and will still confer a distinct distribution when plotted in a k-mer spectra, as discussed previously. Repeats are non-systematic duplications of portions of the genome and do not appear in specific repeat numbers. This causes the trailing tail in the k-mer spectra, as the low count of k-mers representing these repeats appear in high frequency but with non-specific duplicate rates. This plot could also represent a sequencing bias, where oversampled k-mers appear in the trailing tail and the Poisson distribution is skewed to the left.

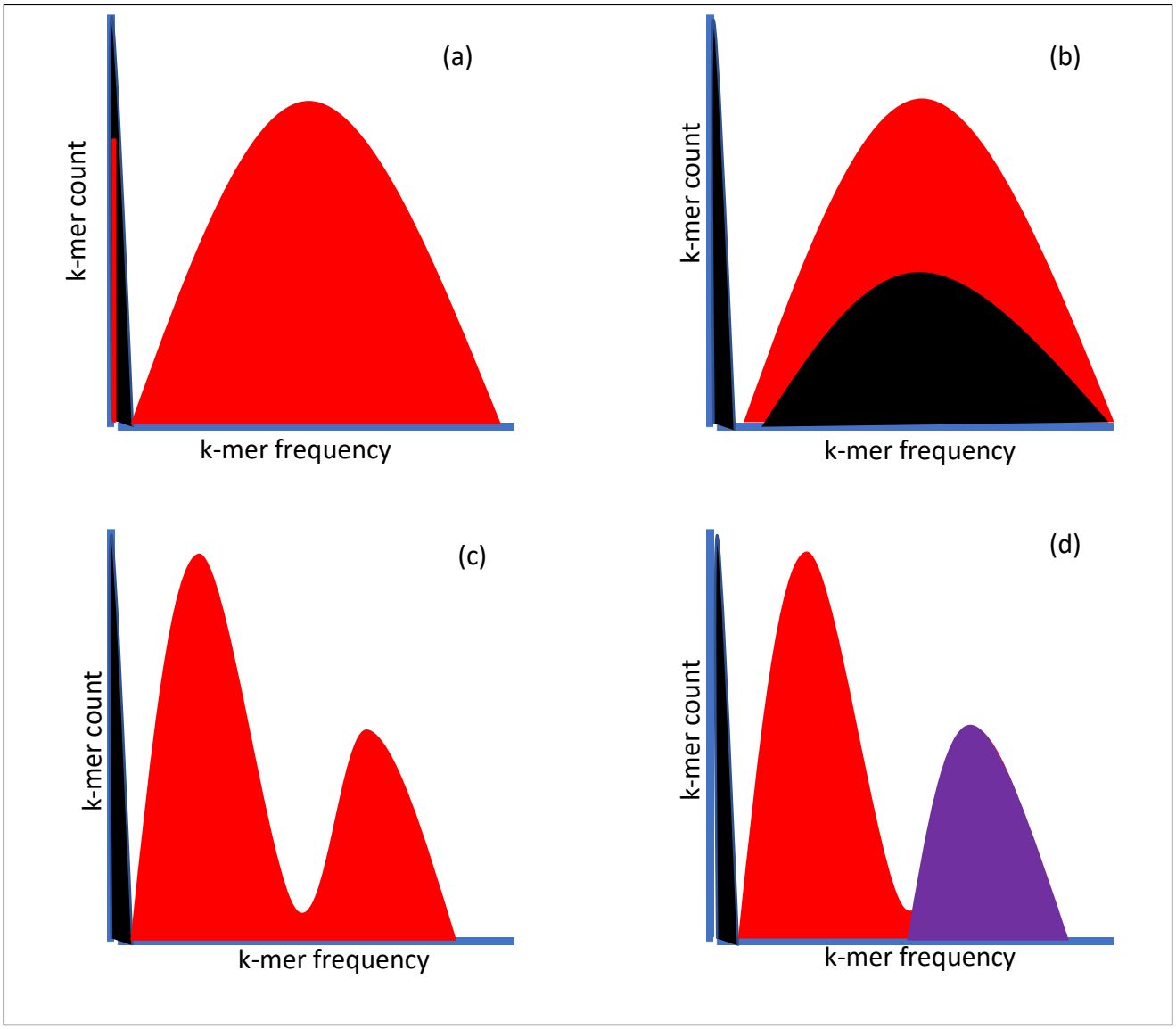

Figure II: Representation k-mer spectra histograms introducing spectra comparisons. These comparisons are generated by counting the shared number of k-mers between the two input sets of sequence. In the first plot (a) is a symmetrical Poisson distribution representing the original reads compared to an assembly. The shape is determined by the k-mer content of the reads and their frequencies, whilst the colouring is determined by the k-mer content of the assembly. At the lowest frequency on the graph is a single red line and a small black decreasing exponential. The black decreasing exponential represents erroneous reads, rare k-mers at low frequencies, which have not been included in the assembly (hence the black colour). The red line at low frequency represents misassembly; rare k-mers that appear in the assembly but not in the original reads. Ideally, we want all erroneous k-mers from the reads to be excluded from the assembly and appear as black in the spectra; we would also aim for no misassemblies, so do not want to see a red line at a low frequency like this. In plot (b) the assembled content (red) does not represent all the content of the reads. A large amount of k-mers from the reads are missing from the main distribution (black). This can be caused with too small a k-mer value on assembly, forfeiting more frequent k-mer overlaps for higher coverage. Plot (c) shows a diploid distribution whereby all the k-mers

Continued overleaf…

from the reads appear in the assembly exactly once. Here, the heterozygous content in the first peak should always be red, as these k-mers only appear once in the diploid genome.

The second peak identifies this assembly as being "collapsed," which is to contain only one copy of all homozygous k-mers. Plot (d) represents the same diploid genome but with all homozygous content included in the assembly as it would appear in the physical pair of genomes; this assembly is fully phased (haplotype-resolved) and both haploids are complete. The red content of the first peak shows that all heterozygous content from the reads is included once in the assembly. The purple second peak shows that all the homozygous content from the reads is included twice in the assembly. In reality, when we assemble a set of reads, we will usually get some level of errors from the reads, some level of heterozygous content missing, and some level of the homozygous content appearing twice but no full haplotype resolution.

1.2.2 Assembly Graphing

Some extent of read assembly is necessary for downstream analysis on sequenced data. The level of assembly required will vary depending on the final application of the data; detection of key genes may require minimal assembly whereas comparative analysis between full sets of genes will require much more contiguity from the sequence.

There are two key methods of genome assembly, which can be seen as conceptual building blocks for genome analysis; Overlap Layout Consensus (OLC) and De Bruijn Graph (DBG) assembly.

As described by Wajid & Serpedin (2012), OLC identifies overlapping patterns in reads by aligning a given read to each other read, which allows a graph of overlaps to be constructed. The key to running a successful OLC is the definition of overlap length for the reads being assembled. Once the graph of overlaps has been constructed, the assembler traverses the graph to generate contigs and a consensus sequence for each contig is determined. The read coverage is used to give confidence to the matching overlaps and any reads below x-coverage can be discounted. However, this layout can be complex and difficult to resolve, particularly in regions of high repetition. Another issue with this method is the removal of lower coverage content, which may still have value in an assembly. Although the most popular method in early genome assembly software, OLC does not scale well with the huge datasets generated by HTS as the initial all-vs-all comparison of reads is computationally impractical. Assemblers such as Arachne (Batzoglou *et al*., 2002; Jaffe *et al*., 2003), Celera (Myers *et al*., 2000) and Falcon (Chin *et al*., 2016) use this method of assembly.

The DBG method gained popularity with the advent of HTS as the process of deconstructing reads into k-mers and building the graph implicitly represents overlapping sequence without the need for an all-vs-all read comparison. DBG, as described by Compeau *et al*. (2011), splits the reads down into k-mers of a defined length. When two k-mers overlap by sharing k-1 nucleotides, they are connected in the graph. A node in a DBG represents a k-mer, which is connected to other nodes through edges. Nodes may have multiple connections in this way, allowing the resolution of repetitive regions and preventing the loss of intermediate segments of the sequence. Although these intermediate segments are not lost, they can still be difficult to resolve as paths may travel through edges an indeterminate number of times. This allows for an indeterminate number of routes from start to end, complicating the

graph. Velvet (Zerbino, 2010), ABySS (Simpson *et al*., 2009), SOAP-de novo (Luo *et al*., 2012), DISCOVAR (Love *et al*., 2016) and W2rap-contigger (Clavijo *et al*., 2017) use this method of assembly.

The effectiveness of either of these methods varies with the algorithms and sequencing methods used. OLC is generally slow to process due to the exhaustive alignments required to determine overlaps. In addition, repetitive regions are much more difficult to resolve. While OLC still has uses for low-coverage, long reads, DBG are more suited to high-coverage, short reads (Li *et al*., 2012).

2. Methods

2.1 *F. cylindrus* culture and sequencing

Fragilariopsis cylindrus (Grunow) Krieger CCMP1102 was obtained from the National Centre for Marine Algae and Microbiota (NCMA, East Boothbay, ME, USA), originally isolated from Southern Ocean seawater (64.08° S 48.7033° W, South Orkney Island Research Cruise, Station 12, 16th March 1979). Cell cultures were maintained in the Mock Lab, University of East Anglia (UEA, Norwich, Norfolk, UK) at +4 °C in f/2 medium. DNA was extracted by Thomas Mock using a cetyltrimethylammonium bromide (CTAB) based method modified from Friedl (1995).

Sequencing was carried out at Earlham Institute (Norwich Research Park, Norwich, Norfolk, UK) using Illumina MiSeq and a PCR-free protocol.

PacBio assembly and reads obtained from the European Nucleotide Archive, accession PRJEB15040. Sanger assembly obtained from GenBank, accession AC275662. Whole genome shotgun project available from GenBank, accession LFJG00000000.

2.2 Initial assessment of Illumina reads

2.2.1 Initial spectra and assembly

An initial histogram of k-mer frequencies of the Illumina reads was produced to visualise the read content using KAT (v. 2.3.4, Mapleson *et al*., 2016) and the `kat hist` command with default options. This gave a graphical representation of the frequencies of unique k-mers in the unassembled reads, which avoided assembly bias. From this, we ascertained the rough genome size and structure from totalling the k-mers in the distribution and by observation of the graphical output. We also acknowledged there may be some contamination within the sequenced sample but note there was no evidence of sequencing bias. This provided a knowledge base for us to proceed with suitable tools and methods.

The initial assembly was built using W2RAP-CONTIGGER (v. 0.1, Clavijo *et al*., 2017) using a K of 260. K 260 provided a small enough k-mer to capture the greatest range of content without compromising k-mer overlap necessary for inclusion into an assembly. The W2RAP output contained a DBG of unitigs.

A k-mer spectra plot was constructed using `kat comp` to visualise the assembled content against the reads. This provided us with a metric to ascertain how much of the original content was being assembled.

2.2.2 Identification and removal of contaminants

We isolated the bacterial contaminant content from the assembly and performed a BLAST (v. 2.6.0+, AltschuP *et al*., 1990) search for identity. As the W2RAP-CONTIGGER only assembles content with a frequency >4, we avoided incomplete removal of contamination at the lower frequencies by using `kat sect` stats to pull out only the *F. cylindrus* content with a coverage >30. All top BLAST alignments (>=95% identity and >=95% sequence alignment) deviant from *F. cylindrus* were also removed. Small contigs <1000bp were removed to avoid inclusion of single-read based constructs and keep content that is both

long enough to represent cleanly assembled content, and to provide good anchor for long read mapping. A further `kat comp` spectra plot of the isolated *F. cylindrus* content against the reads was constructed to ensure no bacterial contamination remained.

2.2.3 Estimating genome size

Polyploid genome size was estimated using the k-mer spectra matrix file (from the command `kat comp`) of the reads vs Illumina assembly by the following equation:

$$\frac{\text{total shared kmers from reads} + \text{missing content}}{\text{k} - \text{mer coverage}}$$

This estimate is based on the total shared k-mers between the reads and the assembly. K-mer coverage was obtained from the mode frequency of the highest k-mer count (x=53) in the spectra. This method avoids assembly bias by accounting for collapsed regions (Vurture *et al*., 2017). Missing content was based on the number of k-mers not shared between the reads and the assembly, which was obtained from the same k-mer spectra.

2.3 Assessing previous assemblies

K-mer spectra plots were made of the PacBio-Falcon and Sanger-Arachne assemblies vs Illumina reads using `kat comp` with default options. The spectra matrices outputted produced values for the frequencies of shared k-mers to assess how much of the original content was assembled. It also produced values for the amount of content contained in the reads that was not assembled. The spectra plot gave an indication of which haploid genomes had been included in the assemblies by the frequencies the content was presented at. This method of comparison with reads kept the content assessment free of assembly bias (Liu *et al*., 2013).

The Illumina assembly was aligned to the Sanger-Arachne and Pacbio-Falcon assemblies to produce dotplots using MUMMER (v. 3, Vurture *et al*., 2017) for direct assembly comparison. QUAST (v. 4.6.3, Mikheenko *et al*., 2016) was used to further compare the Illumina assembly with the existing PacBio-Falcon and Sanger-Arachne assemblies.

2.4 Analysing genome structure

2.4.1 Haplotype Identification

The initial k-mer frequency plots showed three distinct peaks, representing a possible three sub-genomes (which we termed A, B and B'; see Section 3.4). To confirm the presence of the three sub-genomes and the distinctness of the A sub-genome, we manually separated out A and B/B' as two entities; we used `kat sect` to create a k-mer coverage file for the entire assembly and constructed a DBG annotated with these coverage values. BANDAGE (v. 0.8.1, Wick *et al*., 2015) was used to view the DBG for manual analysis. We had not split the B and B' sub-genomes in this study due to the complexities in fully-phasing this polyploid genome, but this is an achievable aim in the future.

We used a custom python script to separate out the unitigs as A and B/B' based on the k-mer coverages produced from the coverage file (Appendix A). The script first assigned the variance in k-mer frequencies to three distinct categories based on the distribution of peaks in the spectra graph; 0-79, 80-129 and >130. The script then assigned each unitig a sub-genome based on the percentages of k-mers contained in each of those three categories; the A subgenome, being largely unique, contained most content from the 0-79 or >130 range; the B and B' genomes, being highly similar, contained most content from the 80-129 and >130 range. Any unitigs with the majority of content appearing in the >130 category are shared between the three genomes and were assigned to a separate category. The DBG was annotated with this classification and manually checked in BANDAGE. These easily resolved local areas of the graph manually confirm the script is assigning sub-genomes correctly on the majority. This sub-genome assignment was confirmed by independent K-mer Compression Index (KCI) using in-house tools, which sorted the sub-genomes based on:

$$\frac{\text{read coverage}}{(\text{kmer coverage} * \text{assembly coverage})}$$

2.4.2 Comparing homologous regions

The identified A and B/B' unitigs were cut from the assembly into separate files. Each set of unitigs was compared to the Illumina reads by a KAT comp spectra plot to confirm the separation of the two sub-genome types. The values from the spectra matrix were used to estimate the amount of diverged content. QUAST compared the A sub-genome against the collapsed B/B' sub-genomes.

These separated A and B/B' unitigs were aligned using MUMMER (command `nucmer --mum -L500).` The coordinate output of MUMMER was used in Genome Ribbon (Nattestad *et al*., 2016) for visualisation. Three of the more contiguous sections of the sub-genomes were selected for more in-depth analysis, which were further extracted and trimmed using a custom Python script (Appendix B). These unitigs matched in bp size exactly and were large enough for local-level analysis at 84Kb, 121Kb and 98Kb. The trimmed A unitigs were aligned to the trimmed B/B' unitigs using MUMMER (command `nucmer --mum -L500`).

2.4.3 Sub-genome longer scale linkage

A Sanger 8Kb paired-end insert library was used for linkage in the Illumina assembly using in-house tools. Links are formed when paired-ends align to contigs greater than 1Kb and are confirmed by the alignment of 4 or more paired-ends forming the same link. This alignment was used to confirm sub-genome contiguity and identification.

The Sanger 8Kb paired-end library was aligned to the isolated sub-genome unitigs using BWA (v. 0.7.12-r1126, Li and Durbin, 2009). The output SAM file was filtered for exact hits on both paired-ends.

2.5 Structure and missing content in the previous assemblies

The three trimmed unitigs were concatenated and aligned to the PacBio-Falcon and Sanger-Arachne assemblies using MUMMER (command `nucmer --mum -L500`). These were visualised in Genome Ribbon to assess the genome completeness and inclusion of the sub-genome types in the two previous assemblies.

3. Results

3.1 Sequencing

A total of 24 308 332 paired end reads of 300bp in length were sequenced in 1 library with a 600bp insert size. This represents 7 292 499 600bp total sequenced content with 47x read coverage (for a 152Mb polyploid estimate; see Section 3.2.3).

3.2 Initial assessment of Illumina reads

3.2.1 Initial spectra and assembly

The initial k-mer frequency histogram (Figure 1) shows five distinct peaks, two of which (x~15 and x~275) were not consistent with multiple sub-genomes in the assembly, while the other three were (x=53, x~106, x~159). These secondary peaks are of inconsistent harmonic frequencies so are not part of the genome (Mapleson *et al*., 2016). The three main peaks are in harmonic frequencies with the fundamental frequency (x=53), implying a triple genome structure. K-mers that occur in the first peak (x=53) represent heterozygous content, unique to all sub-genomes. The second peak k-mers (x~106) appear twice as frequently as the heterozygous content, so are shared across two sub-genomes. The third peak k-mers (x~159) represent homozygous content, which exist exactly three times in the reads. Chloroplasts and mitochondria occur in low counts (relative to the *F.cylindrus* genomic content) and at a high frequency, so are not visible in the graphical spectra representation.

We created a provisional assembly of the Illumina reads for downstream analysis but note that no fully scaffolded assembly was necessary for our purposes. Our main aim was to include as much of the original content from the reads as possible in the assembly and determine genome structure, which had not been successfully produced previously. Of the content in the reads, our k-mer spectra plots show that the majority was assembled successfully (Figure 2 (a)). Only 5.2Mb (3%) of content was missing from the assembly. This assembly represents a collapsed genome and does not include all the content of a haplotype resolved assembly. Figure 2 (b) is a synthesised illustration of how an ideal haplotype resolved assembly would look including all completed haploid genomes. The assembler collapses identical contiguous k-mer sequences in the DBG, which appears as a single iteration of the sequence in the assembly. These nodes may have two or three times the coverage of the heterozygous content, because they appear two or three times as frequently, but are only represented in the graph once. In total, 60Mb (40%) of content is missing from the entire phased assembly, but phased components can be reconstructed from the DBG.

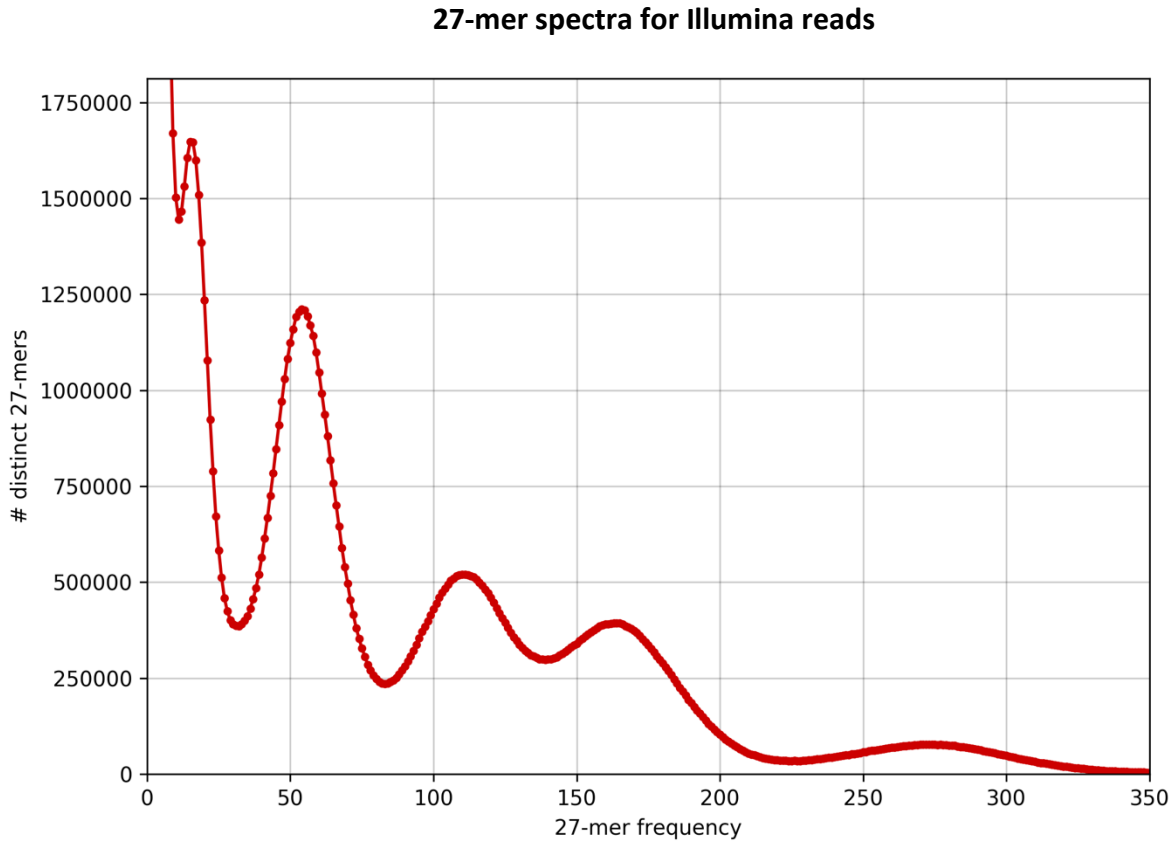

Figure 1: Initial K-mer histogram showing the frequency of distinct k-mers. Three key peaks at x = 53, x ~ 106 and x ~ 159 were confirmed as *F. cylindrus* in a BLAST alignment. The triple peaks of *F. cylindrus* occur at precise multiples in harmonic frequencies of the fundamental frequency (x=53), showing a polyploid structure of the genome. The two peaks at x~15 and x~275 are not in harmonic frequency, suggesting these are contamination. These peaks were confirmed as bacterial contamination by a BLAST alignment and were removed.

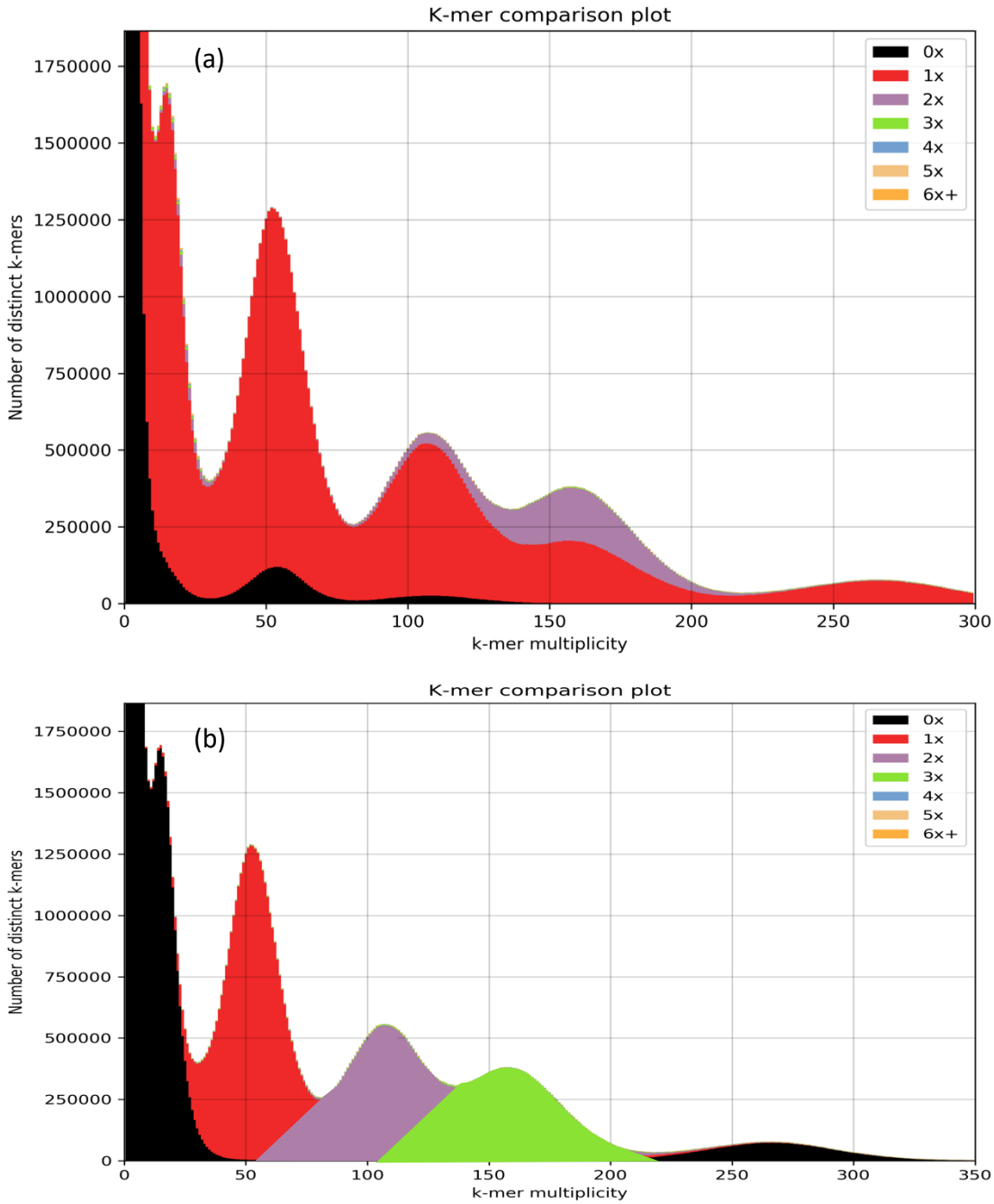

Figure 2: (a) Initial Illumina assembly using default W2RAP-CONTIGGER options and a K of 260, compared with the original reads, and (b) synthesised ideal representation of a fully haplotype resolved genome k-mer spectra. (a) The bacterial contamination is still present in this assembly (x~15 and x~275) but has been assembled into two distinct peaks, so can be removed. The majority of the *F. cylindrus* content in the reads has been assembled with a small amount of content missing at x ~ 53 and x ~ 106. This content will not have been assembled due to the small k-mer size, which gives the assembly greater coverage at the expense of a small amount of content unable to overlap with other k-mers. We can see that the majority of the content is included in this assembly, but the assembly is collapsed. The purple area shows content that was assembled twice from the reads; in an ideal phased assembly (b), the x~106 peak would be uniformly purple and the x ~ 159 peak, uniformly green. This plot a representation for illustrative purposes.

3.2.2 Identification and removal of contaminants

BLAST confirmed the contaminant peaks (Figure 2 (a), x~15 and x~275) to be bacterial, with the peak at x~15 showing best hits to *Methylophaga nitratireducenticrescens*, *Octadecabacter arcticus* and *Paraglaciecola psychrophila,* and the peak x~275 showing best hits to *Colwellia sp.* Contamination was removed and a further spectra plot of the assembly confirmed the absence of contamination (Figure 3). A further BLAST of the three key peaks confirmed *F. cylindrus* identity.

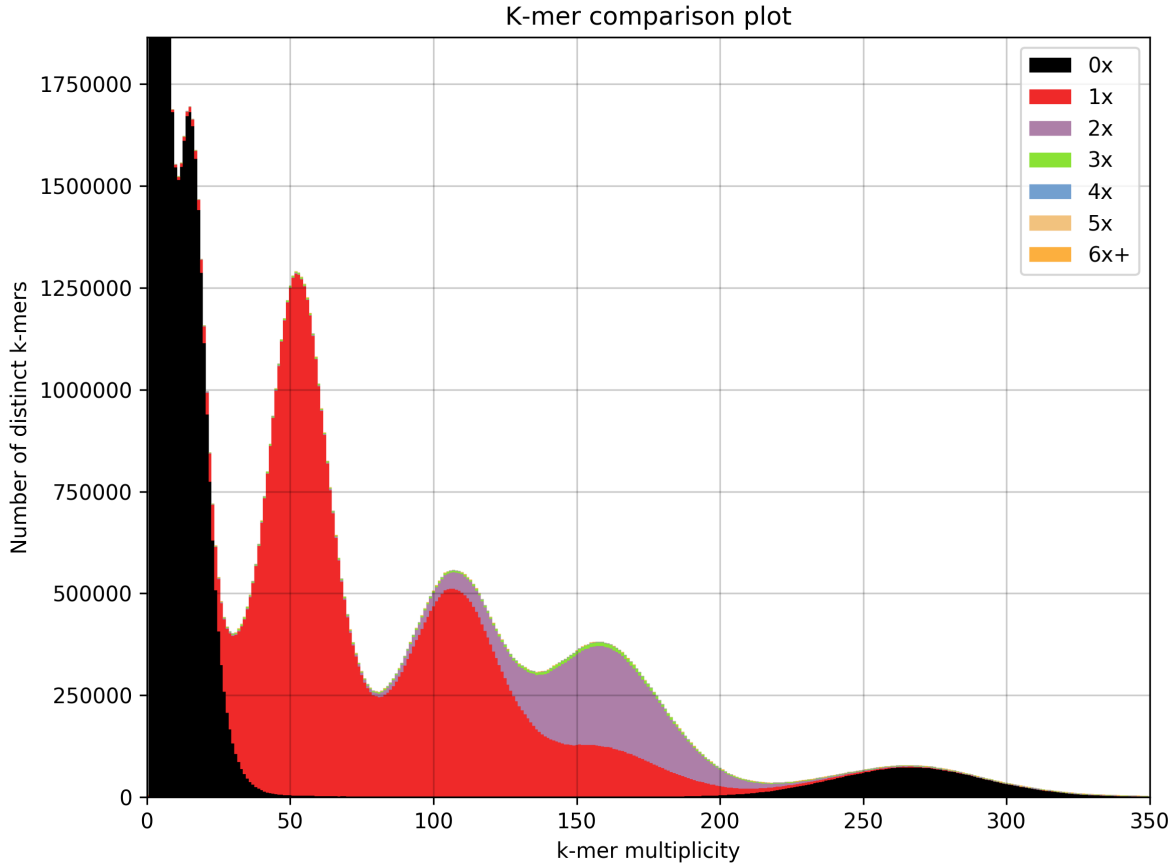

Figure 3: The assembled unitigs of the *F. cylindrus* genome using parameters K 260, with bacterial content removed. The entire contamination peaks (x~15 and x~275) are colourless, showing all the contaminant content has been extracted.

3.2.3 Estimating genome size

We estimate the whole polyploid genome size to be 152Mb, with each individual sub-genome as 50.7Mb. We note that we present no evidence to suggest that this genome is not hexaploid. Altogether, our evidence concludes that this *F. cylindrus* (Grunow) Krieger CCMP1102 genome has three sub-genomes, which may or may not be two identical 3N genome copies. The Illumina assembly spans 100Mb, with the possibility of recovering collapsed regions to complete haploids. This consists of 8753 unitigs with the largest at 660Kb in length, and an N50 of 48Kb (Table 1). Our polyploid assembly contains 19% more genomic content than the Sanger-Arachne diploid assembly.

3.3 Assessing previous assemblies

Based on k-mer spectra plots and matrix, we found the same bacterial content present on both the PacBio and Illumina reads. From the plots and matrix we could also see PacBio reads were missing content from the Illumina reads, and contained a high error rate as a decreasing exponential at low frequency (Figure 4 (a)). The reads, although containing the same sample content as Illumina, were missing sequence that the Illumina reads contained. While this could theoretically be resolved during error correction where it may be plausible to correct nearby errors, it is difficult to do so from low coverage. Therefore, the PacBio assembly would likely not have been assembled to completion based on these reads.

Our assessment of the PacBio assembly indicated it had not assembled the bacterial contamination from the reads. Large amounts of the original content was also missing from the assembly (Table 1; Figure 4(b); Figure 6 (a) and (b)). The PacBio diploid assembly was missing 58% of the total haplotype resolved genomic content (80Mb), and was 31% shorter than the Illumina assembly (Table 1). Mummerplots (Figure 6 (a)) at a large scale show consistency between the PacBio-Falcon assembly and the Illumina assembly. However, based on misalignments between the PacBio-Falcon and Illumina assemblies, the PacBio-Falcon assembly appears to not be contiguous, containing a number of insertions, deletions and translocations (Figure 6 (b)).

The Sanger assembly did not contain any of the bacterial contamination. However, the diploid assembly was missing 46% of the total haplotype resolved genomic content (72Mb) and was 19% shorter than the Illumina assembly (Figure 5, Table 1). The smaller assembly length can be attributed to the loss of heterozygous content, which will have been discarded by the OLC diploid assembler due to low coverage. Mummerplots (Figure 6 (c) and (d)) show more consistency between the Sanger-Arachne assembly and the Illumina assembly; it is more contiguous than the PacBio-Falcon assembly, but contains some amount of repeated content.

Table 1: Overview of the Illumina, Pacbio-Falcon and Sanger-Arachne assembly stats. Although the Illumina assembly appears to be less contiguous, it contains a much higher proportion of overall content.

Assembly	Illumina	Pacbio-Falcon	Sanger-Arachne
Total length	99854374	68972051	80540407
Length percentage based on the Illumina assembly	100%	69.1%	80.7%
Total length (>= 1000 bp)	97311974	68964007	80540407
Number unitigs/contigs	8753	1021	271
Number unitigs/contigs (>= 1000 bp)	4636	1011	271
Largest unitig/contig	659863	1265703	5926375
GC (%)	38.91	38.60	38.51
N50	47827	190760	1295603
N75	23633	64613	550327
L50	550	84	16
L75	1289	251	40
# N's per 100 kbp	0.00	0.00	5384.58

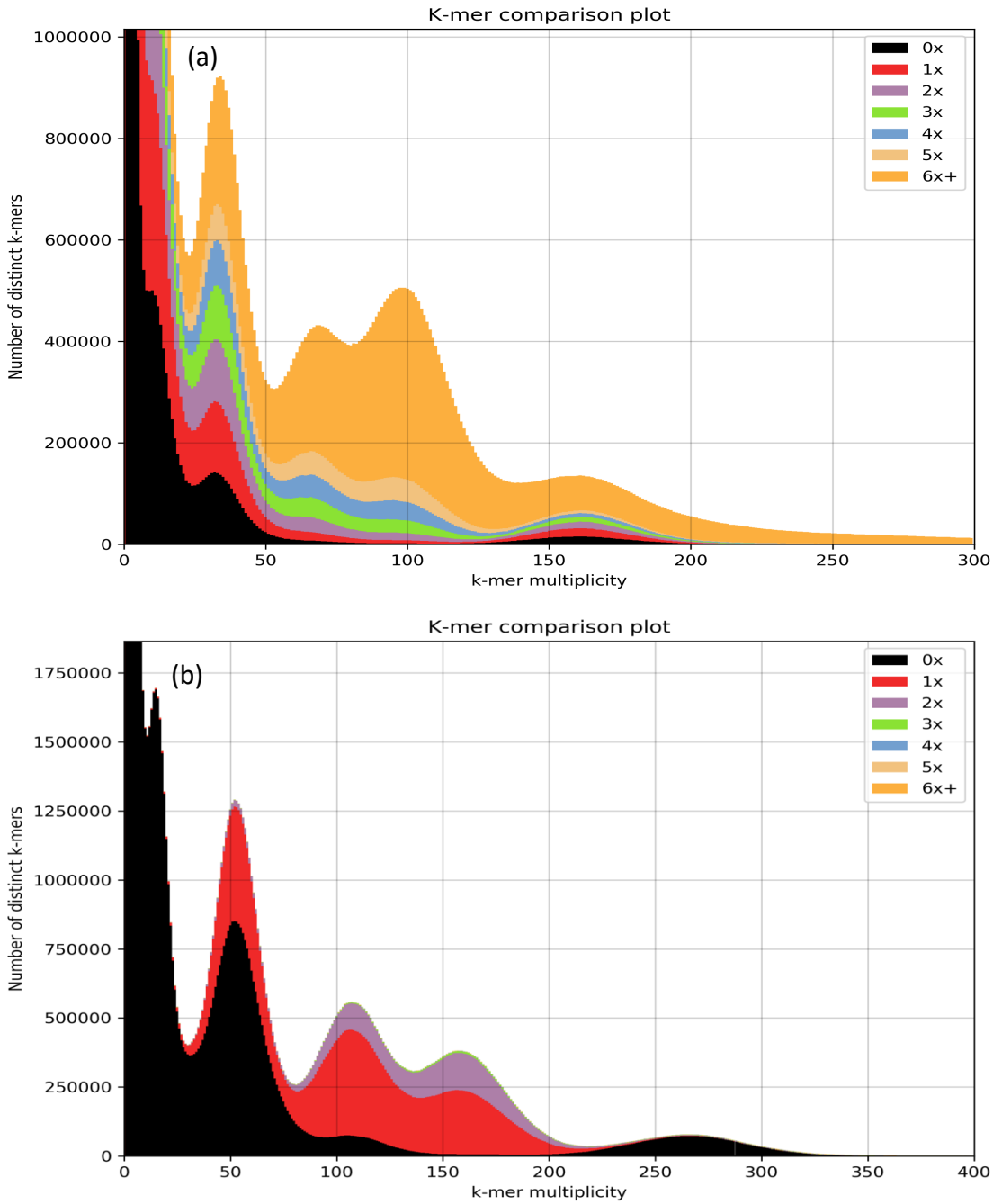

Figure 4: K-mer spectra analysis of PacBio (a) reads and (b) assembly compared to Illumina reads. The same number of peaks between PacBio and Illumina reads, including the contamination, can be seen in (a), showing PacBio to have captured the same range of content as Illumina. This plot also shows some missing content on the first distribution (black), which is present in the Illumina reads. While this could theoretically be resolved during error correction where it may be plausible to correct errors that are isolated and in close proximity to perfectly correct sequence, it is difficult to do so from low coverage. The steep decreasing exponential at low frequencies (which is too large to plot, >1 000 000 counts) represents the high error rate of the reads. Plot (b) shows an absence of k-mer content in the black peaks at x~15 and x~275, so the assembly has not incorporated bacterial contamination. However, the assembly is also missing large amounts of unique content in peaks x ~ 53, x ~ 106 and x ~ 159, amounting to 36Mb. This assembly is missing large amounts of unique, low-coverage content, which is typically discarded in the assembly of a diploid construct.

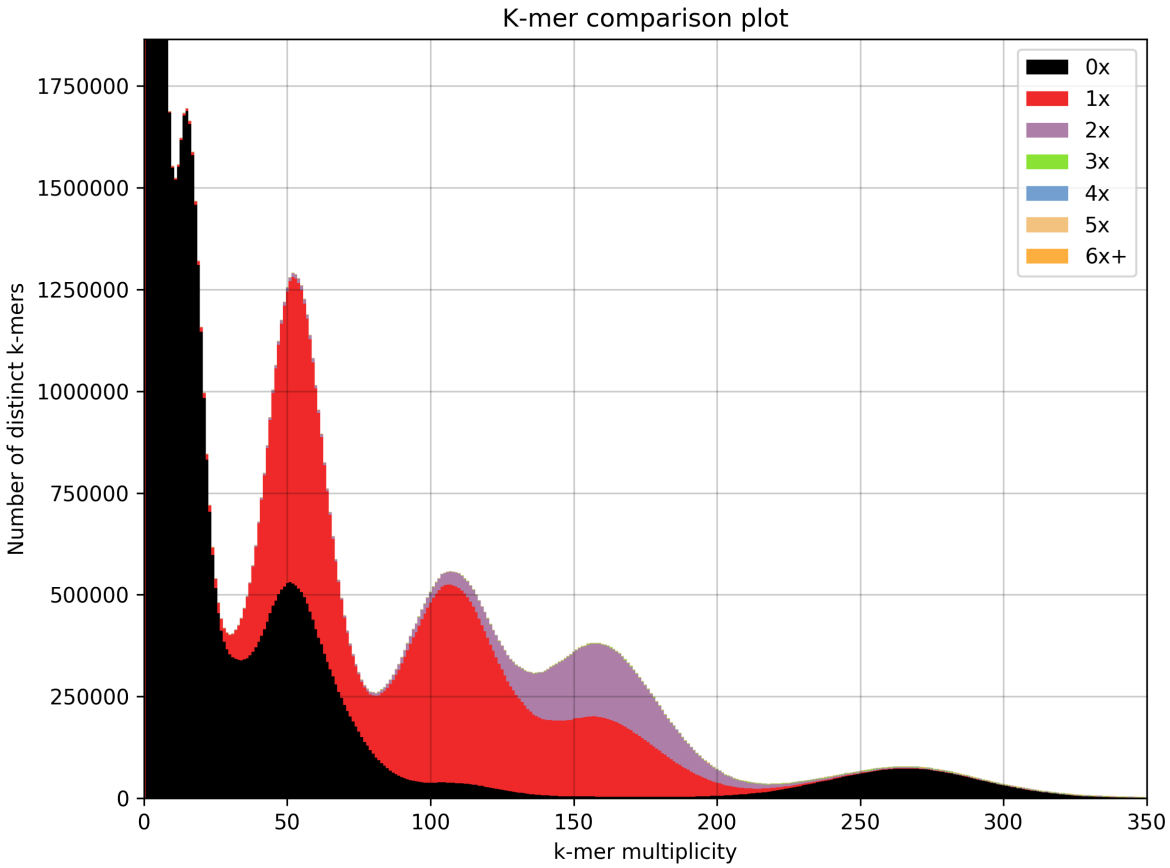

Figure 5: K-mer spectra analysis of Sanger-Arachne assembly compared to the Illumina reads. The absence of k-mer content in the black peaks at x~15 and x~275 show there is no bacterial contamination present in the Sanger-Arachne assembly. Black in the x ~ 53 and x ~ 106 peaks represents content missing from the assembly and amounts to 24Mb. This assembly is missing large amounts of unique, low-coverage content, which is typically discarded in the assembly of a diploid construct.

Dotplots: PacBio-Falcon assembly vs Illumina assembly; Sanger-Arachne assembly vs Illumina assembly

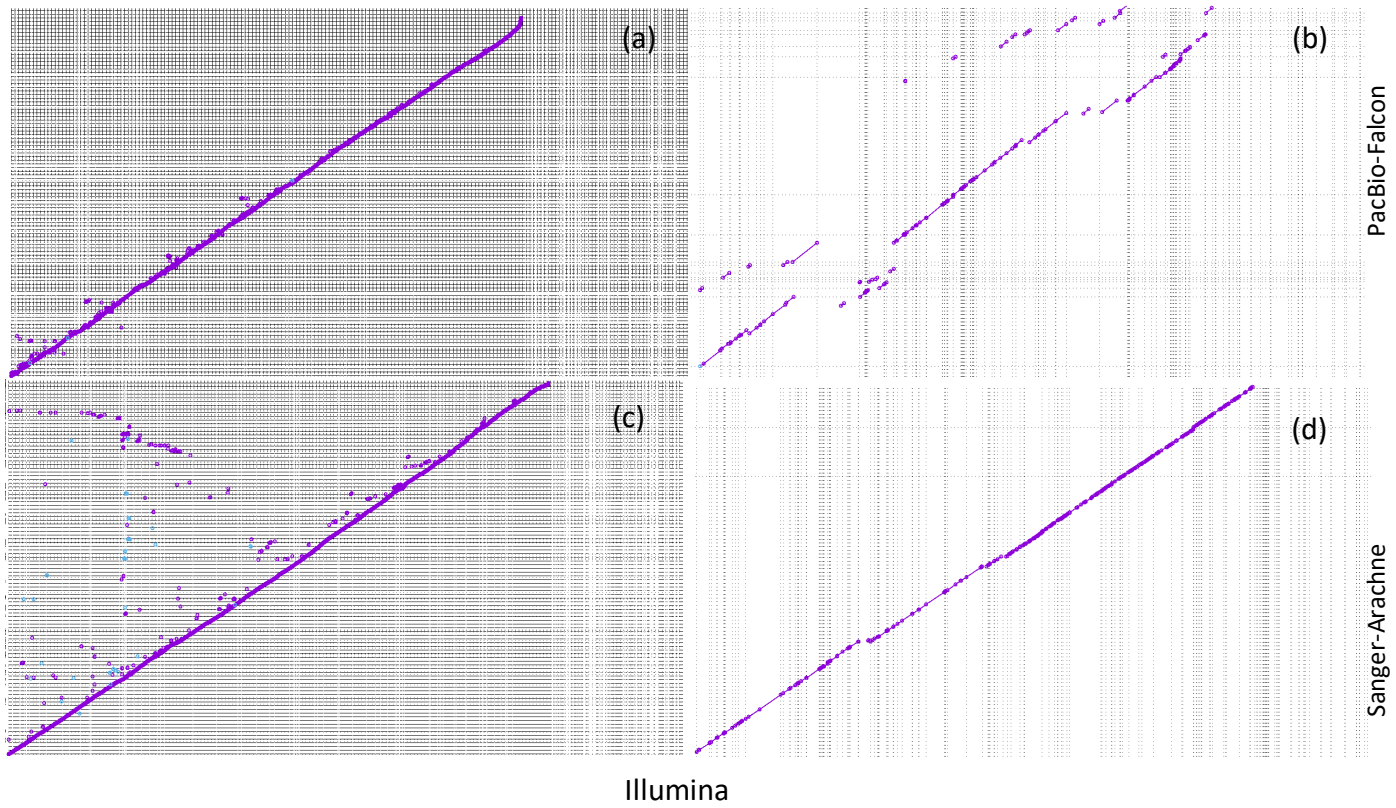

Figure 6: MUMMER dotplots of the (a) PacBio-Falcon assembly vs Illumina assembly, (b) close up of the PacBio-Falcon assembly vs Illumina assembly from the bottom left corner, (c) Sanger-Arachne assembly vs Illumina assembly, and (d) close up of the in Sanger-Arachne assembly vs Illumina assembly from the bottom left corner. Purple dots and lines represent forward alignment and blue dots represent reverse alignment. Most of the PacBio-Falcon content is represented in the Illumina content (a). However, the PacBio-Falcon assembly is clearly disjointed on closer inspection (b), showing contiguity differences and a number of misassembled regions, insertions, deletions and translocations when compared to the Illumina assembly. The PacBio-Falcon assembly is also missing a large portion of the content that the collapsed Illumina assembly contains, shown by the diagonal that does not terminate at the upper right corner (a). The Sanger-Arachne assembly follows the same alignment as the Illumina assembly (c), and also shows a great deal of contiguity in the close view (d). However, the Sanger-Arachne assembly is also missing a large amount of the content that the Illumina assembly contains, shown by the termination of the diagonal a distance in from the upper right corner (c). The dots above the diagonal show repeated content in the Sanger assembly.

3.4 Analysing genome structure

3.4.1 Haplotype Identification

Manual analysis of the DBG appeared to show a clear division between the sub-genomes; one sub-genome that appeared very diverged and two that appeared very similar. Nodes of collapsed content had a k-mer coverage three times the mode of the unique content, with k-mers appearing exactly three times in these regions. From these nodes branched two alternative linking nodes in a "bubble", which distinctly carried either single or double amounts of k-mer coverage (Figure 7 & 8 (a)). K-mer spectras of these nodes show the same signature as the spectra of the entire assembly, with one distinct frequency peak for one side of the "bubble" and two distinct frequency peaks for the alternate side. There were also distinctions between B and B' (Figure 8, (b)), which stemmed from a node containing double coverage (x~106 on the k-mer spectra) into low-coverage unique content on alternating sides (x~53 on the k-mer spectra).

Automatic labelling of the DBG by two methods based on unitig coverage percentage and k-mer compression index of the DBG resulted in the same sub-genome assignment. Nodes that represented low-coverage unique content were labelled as "A" and nodes that represented the highly similar, collapsed double content were labelled "B/B'". These unitigs continued to show an alternating pattern on the DBG (Figure 8) where A and B/B' stemmed from a collapsed triple coverage node. K-mer spectra plots of the A and B/B' unitigs confirmed sub-genome separation based on coverage (Figure 9). Sub-genome A k-mers originate in the x=53 peak and x~159 peaks of the spectra as content from this sub-genome is highly unique, with k-mers also in common with all three sub-genomes. Sub-genomes B/B', being highly similar, share most of their k-mer content in the x~106 peak of the spectra, with some k-mers commonly shared amongst all three. Missing k-mer content is from unitigs that represent all three sub-genomes, or remained undefined.

We identified 125 "bubbles" in the DBG between A and B/B' haplotypes, of which we could be sure that the "bubbles" commenced and terminated at collapsed triple coverage nodes (Figure 8 (a)). These defined "bubbles" account for 3% of the polyploid size and confirmed that the A and B/B' haplotypes were alternate. Interestingly, the "bubbles" accounted for such little overall content because, more often, we found large nodes of up to ~1Mb where A was alternate from the collapsed B/B'. These long unitigs were connected in the assembly graph by a collapsed triple coverage node, but contained either complex repeat regions at the latter end, or had no further links. These highly contiguous nodes suggest high divergence between the A and collapsed B/B' sub-genomes, because shared regions would have been collapsed by the assembly algorithm. Furthermore, these highly dissimilar areas also contained rare divergences between B and B', where the double coverage node branched into two single coverage nodes on alternating sides (Figure 8 (b)).

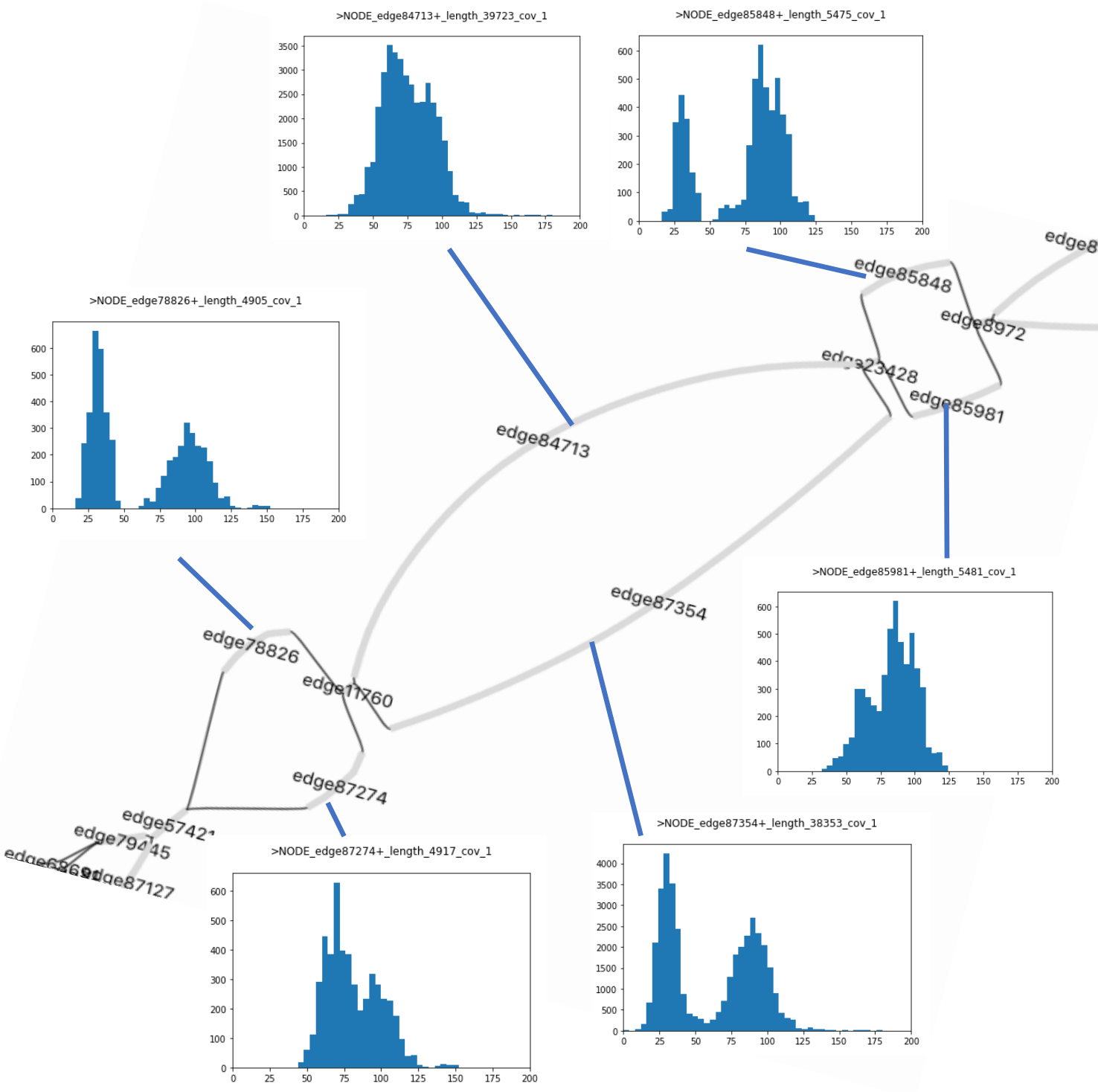

Figure 7: De Bruijn Graph "bubbles" in the Illumina assembly with corresponding k-mer spectra of each node. The pattern of the graph shows a typical diploid arrangement whereby a collapsed node (edge 11760) branches out into two diverged nodes (edge 84713 and edge 87354). However, when we view the corresponding k-mer spectras, we can see that edge 84713 contains a single k-mer peak (the A haplotype) but edge 87354 contains two distinct k-mer peaks (haplotypes B and B'). This is reflected across the other nodes in the graph and is evidence that two very similar haplotypes are collapsed into the nodes that form the double-peaked spectras.

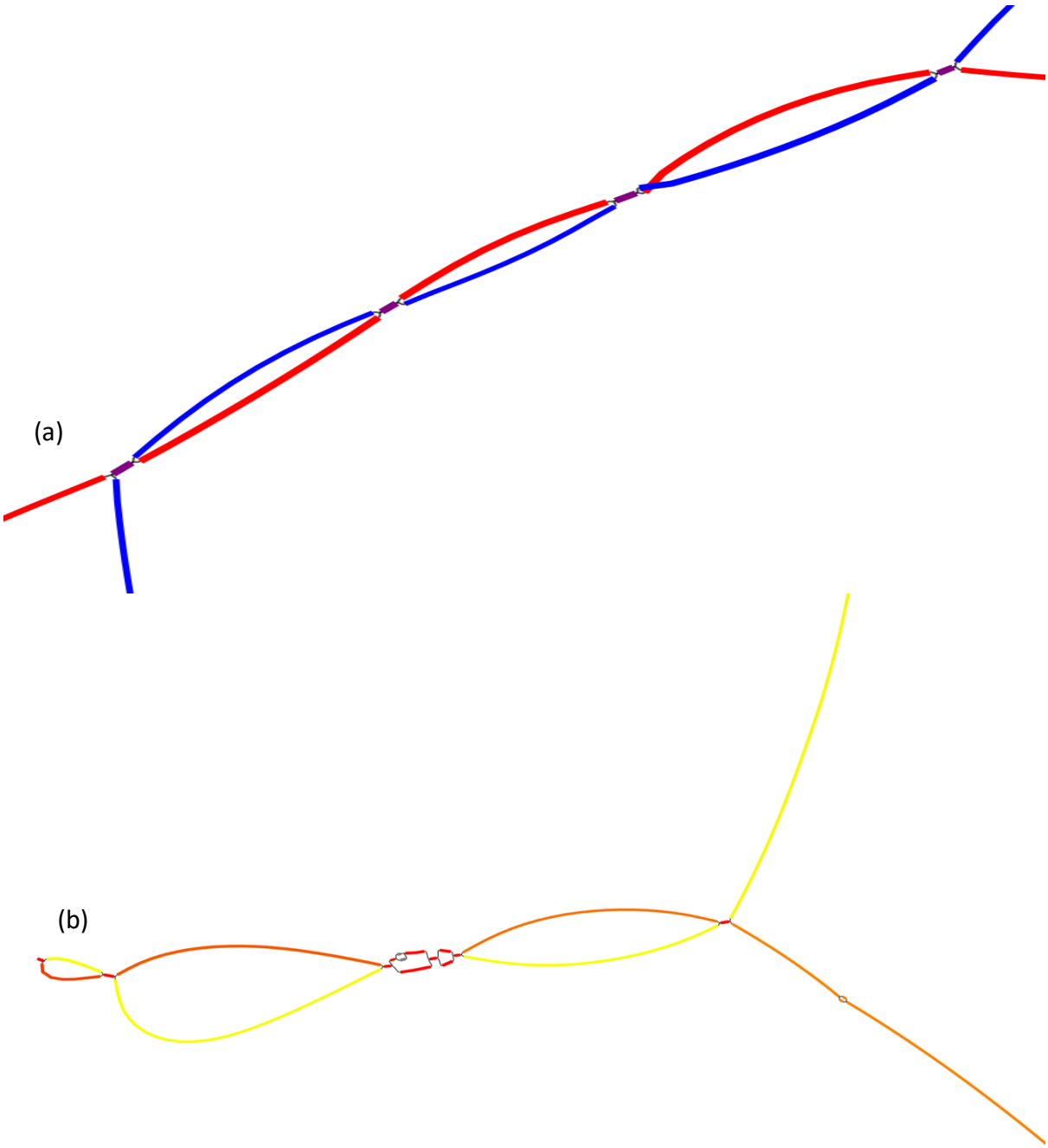

Figure 8: Sub-genome classified De Bruijn Graphs of the Illumina assembly based on coverage. Each "bubble" in (a) originates and terminates at a collapsed triple coverage node (purple). The A sub-genome (blue) alternates from the triple collapsed nodes with the B/B' sub-genomes (red). The coverage of A haplotypes is within the x~53 range, the B/B' coverage within the x~106 range and the triple coverage within the x~159 range of the k-mer spectra in Figure 1. There is an estimated 29Mb diverged content between the A and B/B' sub-genomes. The red nodes in (b) are from the same x~106 range, but the B and B' genomes split here into two alternating unique unitigs from the x~53 range (yellow and orange). These divergent nodes account for 7Mb in kmers. The alternating red nodes are likely errors as they contain the same x~106 coverage.

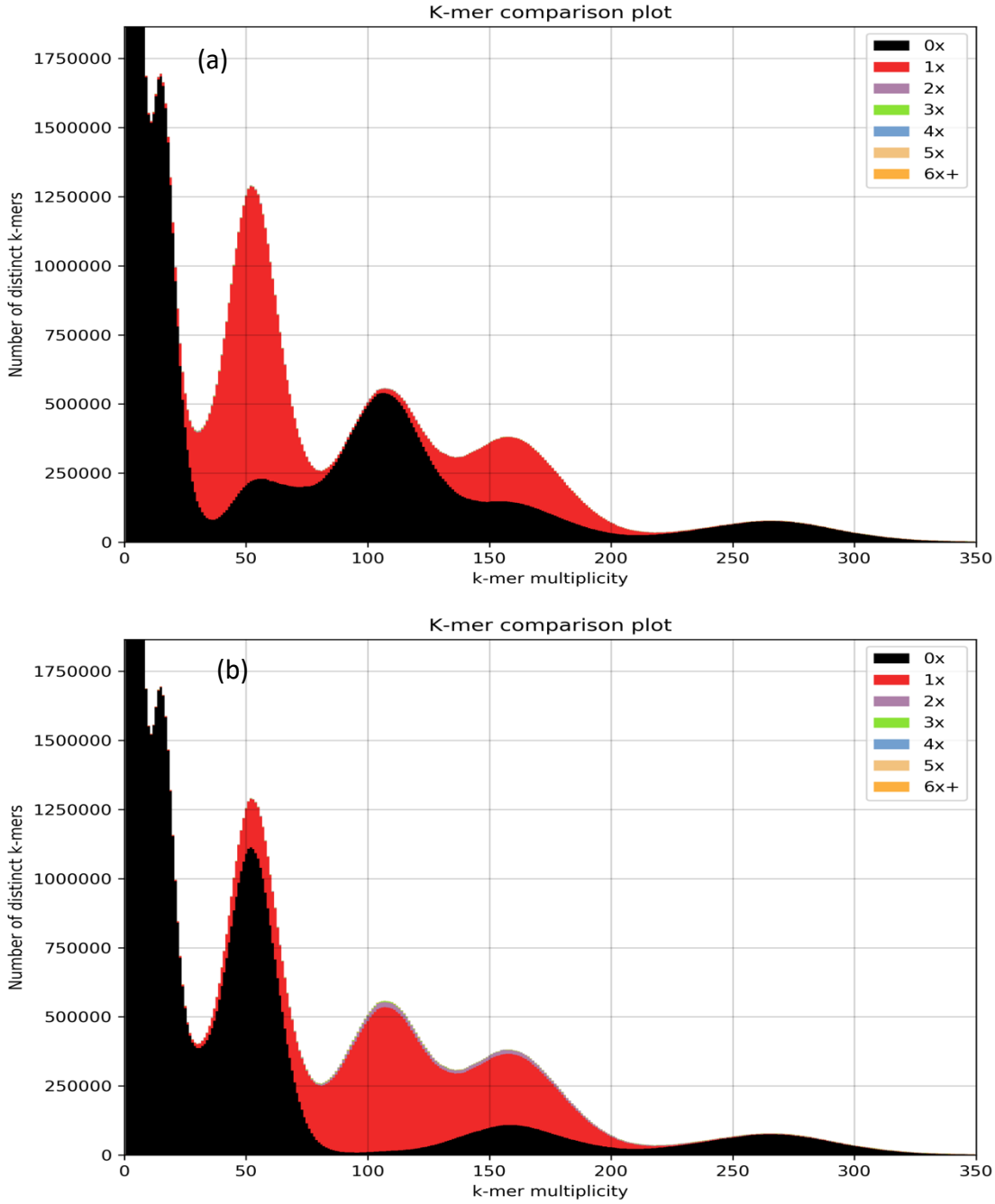

Figure 9: Isolated A (a) and B/B' (b) sub-genome unitigs based on coverage. Sub-genome A (a) includes many unique k-mers (red, x=53) and k-mers that appear three times in the assembly (red, x~159), but almost no k-mers that appear just twice (black, x~106). This confirms separation of the diverged A sub-genome unitigs, containing high levels of unique low coverage and collapsed triple coverage content. This plot shows 29Mb of content unique to this sub-genome (figure attained from the k-mer spectra matrix file, which is used to plot the graph). Sub-genomes B and B' (b) include very few unique k-mers where they diverge from each other (small red content at x=53) and many k-mers that appear three times (red, x~159), but mostly consists of k-mers that appear twice in the assembly (red, x~106). This isolation contains two collapsed sub-genomes that are almost identical in content, but still contain 7Mb of content unique between B and B'.

3.4.2 Comparing homologous regions

A total of 45Mb was isolated for the B/B' sub-genomes and 44Mb for the A sub-genome. The A sub-genome was more contiguous with an N50 of 68Kb and largest unitig of 660Kb; the B/B' N50 was 42Kb and the largest unitig was half of that of the A sub-genome at 309Kb (Table 2).

A total of 29Mb (57% of the estimated sub-genome size) in diverged content was found between A and B/B' (Figure 9 (a)) in the isolated subset (Table 2). Between the B and B' sub-genomes, we found 7Mb (14% of the estimated sub-genome size) diverged content and 23Mb (45% of the estimated sub-genome size) content shared exclusively between them (diverged from the A sub-genome; Figure 9 (b)). The remaining content from the isolations (Table 2) are k-mers found in all three sub-genomes. Based on these estimates alone, we would expect 20.7Mb (41% of the estimated sub-genome size) of shared content between all three sub-genomes. This estimate will be biased based on the accuracy of sub-genome separation, which is still in progress of refinement, but provides a rough guide as to the level of divergence between the three sub-genomes.

Alignment of the A and B/B' sub-genome unitigs showed high divergence. Using Genome Ribbon, we can see many occurrences of insertions, deletions, translocation and inversions (Figure 10). We used an isolated, less diverged selection of regions for local analysis (Figure 11), which contained an average of 92% identity across three unitigs (84Kb, 121Kb and 98Kb).

Table 2: Overview of isolated A and B/B' sub-genomes from the Illumina assembly. The B and B' content is collapsed (not haplotype resolved) so represents the size in bp of the content of one sub-genome. The size difference between the two sub-genomes may be due to extra content included in the B/B' selection where the two sub-genomes differentiate. The less contiguous selection of unitigs from B/B' when compared with A is because of small divergences between the B and B' sub-genomes, which fragments the selection of unitigs more than A.

Assembly	**A sub-genome**	**B and B' Sub-genomes**
Total length	43536951	45366826
Number unitigs	1263	2414
Largest unitig	659863	309215
GC (%)	39.69	38.13
N50	68254	41911
N75	32042	21952
L50	163	318
L75	402	690

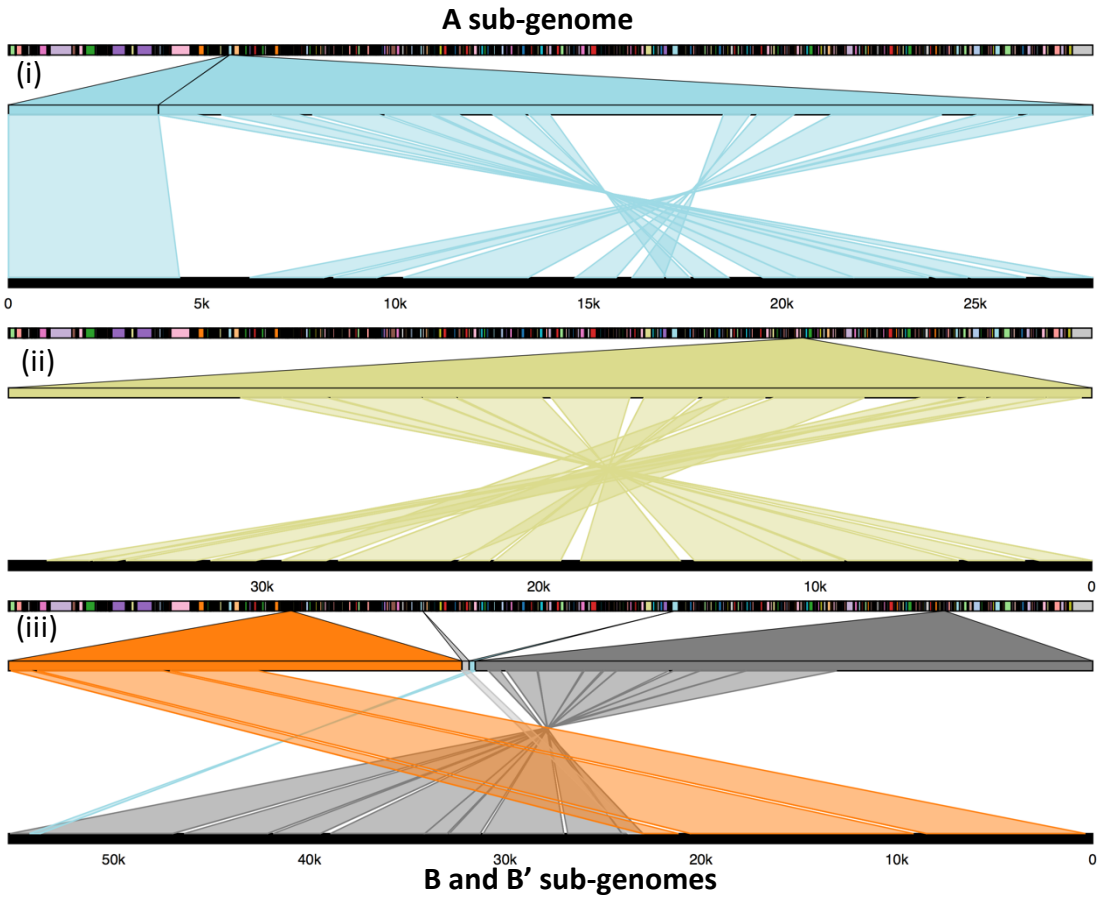

Figure 10: Genome Ribbon plots of the A sub-genome aligned against the B and B’ sub-genomes in complex regions. The plots (i), (ii) and (iii) represent three unitig alignments of A against the B and B’ sub-genomes. A high rate of divergence is clear between the three sub-genomes, with many insertions, deletions, inversions and translocations seen.

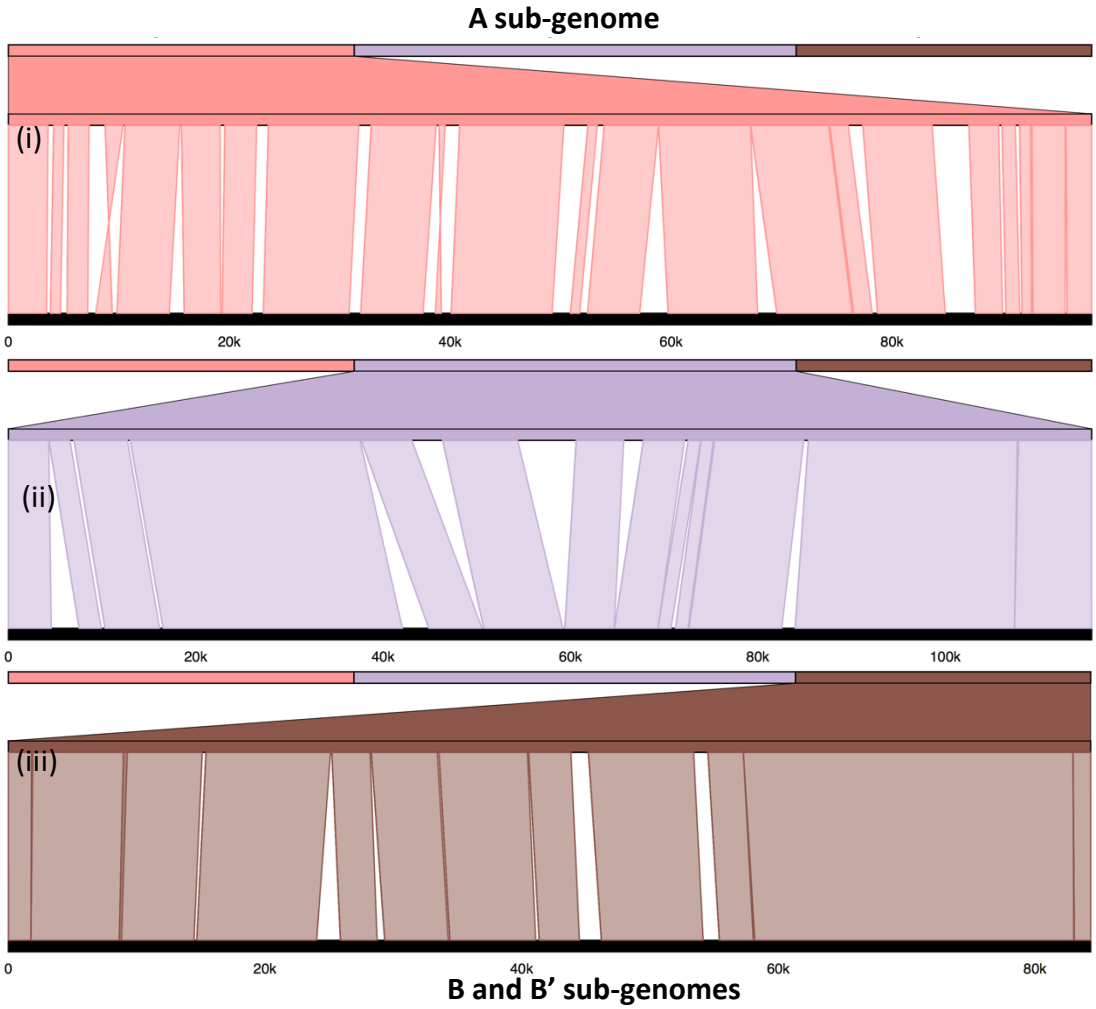

Figure 11: Genome Ribbon representation of three contiguous unitigs of the A vs B and B' sub-genomes. The plots (i), (ii) and (iii) represent three unitig alignments of A against the B and B' sub-genomes. These three aligned unitigs are of the more contiguous regions of the three sub-genomes. A number of insertions, deletions and inversion can be seen between them, lending to a relatively low average identity of 92% (figure attained from the average alignment identity of the nucmer output).

3.4.3 Sub-genome longer scale linkage

Links between unitigs were defined when 4 or more Sanger 8Kb paired-ends aligned between two unitigs. A total of 4 230 links were made by paired-ends on the DBG and 94% confirmed correct contiguity. Of these, 73% confirmed A or B/B' contiguity; 830 links confirmed A contiguity and more than double the number, at 2236 links, confirmed B/B' contiguity. This is because the B and B' sub-genomes contained content represented twice as often as the A sub-genome in the sample pool at sequencing, so will have twice the coverage of the A sub-genome. The remainder contained links to commonly shared unitigs. Of the 6% links that did not confirm correct contiguity, 140 were undefined and 230 (5%) were incorrect links.

Sanger 8Kb paired-ends aligned twice as often to B/B' as they did to A (Table 3). A total of 144 745 paired-ends landed on the same unitig for the B/B' sub-genomes, and half as many at 71 383 paired-ends landed on the same A sub-genome unitig. More than twice as many paired-ends linked B/B' unitigs at 2083 compared to linked A unitigs at 563. This is due to a

higher proportion (A 1:2 B) of paired-ends produced from the B sub-genomes because most of their content appears twice as often as the A sub-genome. The high rate of alignment we see is despite two layers of potential errors; genuine misidentification of unitigs and misidentification based on the separation of B and B', making them appear as A unitigs in the assembly graph.

Interestingly, we found a high rate of Sanger 8Kb paired-end alignments linking A to B/B', which increased with an alignment ID score decrease and was not consistent with our expectations. When thresholds are set to exact alignments on both paired-ends, same-same sub-genome unitigs are aligned 69% for B/B', 19% for A and 12% for A to B/B'. However, when this threshold is decreased to include an exact alignment on the first of the pair, and all alignments for the second of the pair, we see 61% for B/B', 18% for A and 21% for A to B/B'. This was likely a product of missing A or B/B' content, where the best hit for the second paired-end is an imperfect alignment with the opposite sub-genome. The aforementioned errors may also have affected these alignments, especially where B and B' are divergent.

Table 3: Sanger 8Kb paired-end linkage of the A and B/B' sub-genomes summary.

Sub-genome	**A**	**B&B'**
Total aligned paired-ends	71383	144745
Links confirming contiguity	830	2236
Number paired-ends linking unitigs	563	2083

3.5 Structure and missing content in the previous assemblies

The PacBio-Falcon assembly had assembled the B/B' sub-genomes on one of the Illumina unitig pairs in a local alignment (Figure 12, (i)). However, the corresponding A sub-genome has not been completely assembled when compared with the Illumina unitig, borrowing large amounts of content from the same scaffold as the B/B' sub-genomes. Further plots show a large amount of content from the unitig pairs scattered throughout the PacBio assembly, with many repeats, inversions, insertions and deletions. Most of the plots show that neither of the sub-genomes has been fully assembled when compared with the Illumina unitig, but Figure 12 (iii) shows an assembling of most of haplotype A with a number of repeats on other scaffolds and partial assembly of the corresponding B and B' haplotypes. All of these alignments correspond to "primary" contigs on the PacBio assembly, and none of them correspond to "alternate" contigs as expected. This will have amplified the haploid genome estimate.

The local unitig alignment shows that the Sanger-Arachne assembly has a good level of sub-genome assembly for one of each of the unitig pairs (Figure 13). Of the three unitig pairs, B/B' sub-genomes of Figure 13 (i) and (ii) are almost fully constructed on the Sanger-Arachne assembly and the A sub-genome of Figure 13 (iii) is almost fully constructed. The corresponding unitig pair is often fragmented, however, with unitig pair Figure 13 (ii) showing the most complete reconstruction of both sub-genomes.

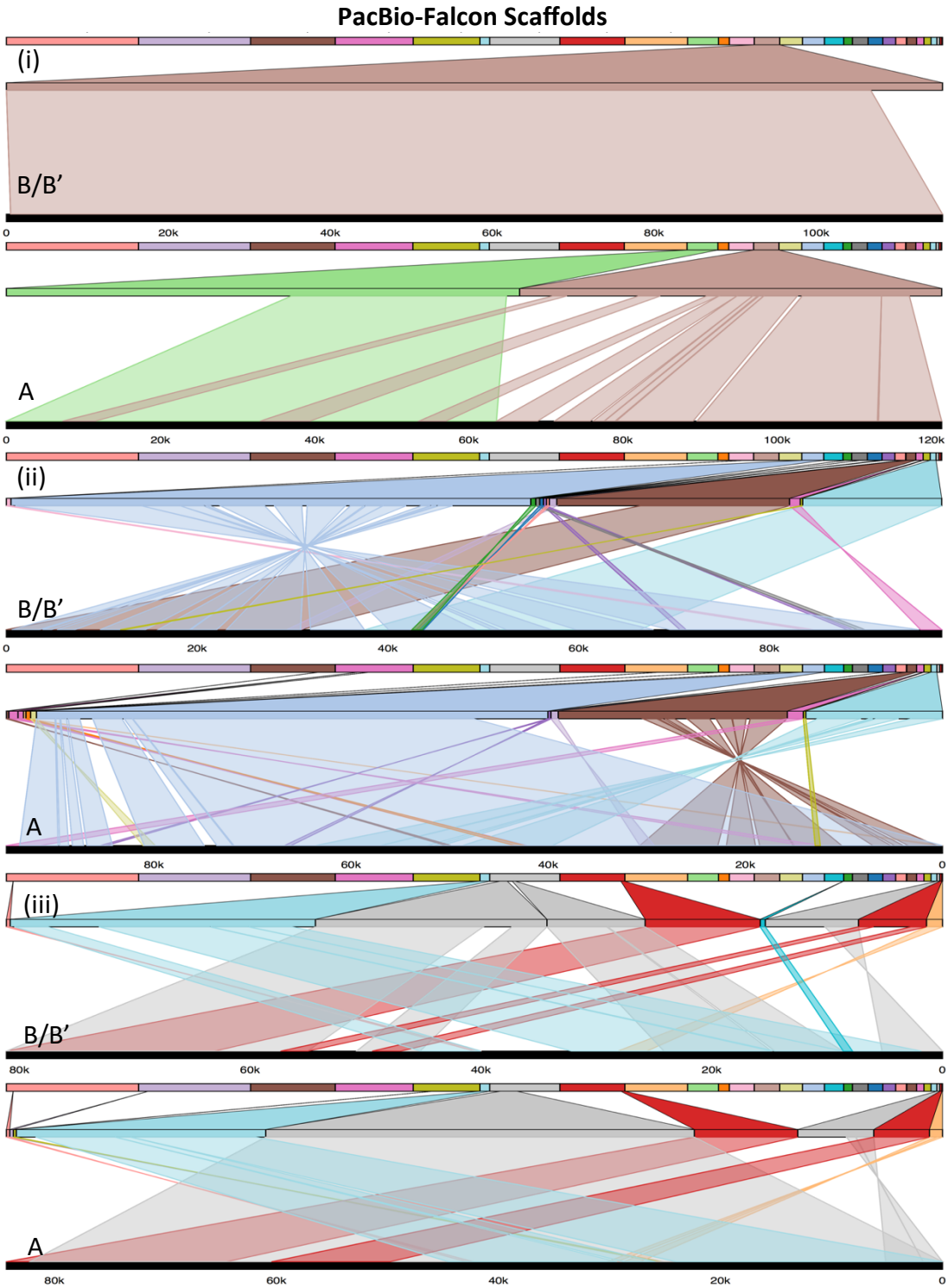

Figure 12: Genome Ribbon representation of the PacBio-Falcon assembly vs three Illumina unitigs from the three sub-genomes. The plot pairs (i), (ii) and (iii) represent three unitig alignments of the PacBio-Falcon scaffolds against the Illumina A or B/B' sub-genomes. Less diverged unitigs between A and B/B' were selected from the Illumina assembly as a standard for this comparison. The first PacBio scaffold ((i), brown) has contiguous alignment to the B/B' sub-genomes; these sub-genomes contain double the coverage, making them more likely to be assembled. However, the corresponding A sub-genome ((i), green) is not included fully in the PacBio assembly, with much of the content aligning to this unitig being represented by the B/B' sub-genome (brown). The other four unitig alignments show how inconsistent the PacBio assembly is, containing bits of each unitig across many scaffolds and reconstructing no full unitig without repeats (overlapping colours) and inversions elsewhere in the assembly. These alignments have a minimum of 77% identity to the Illumina unitigs.

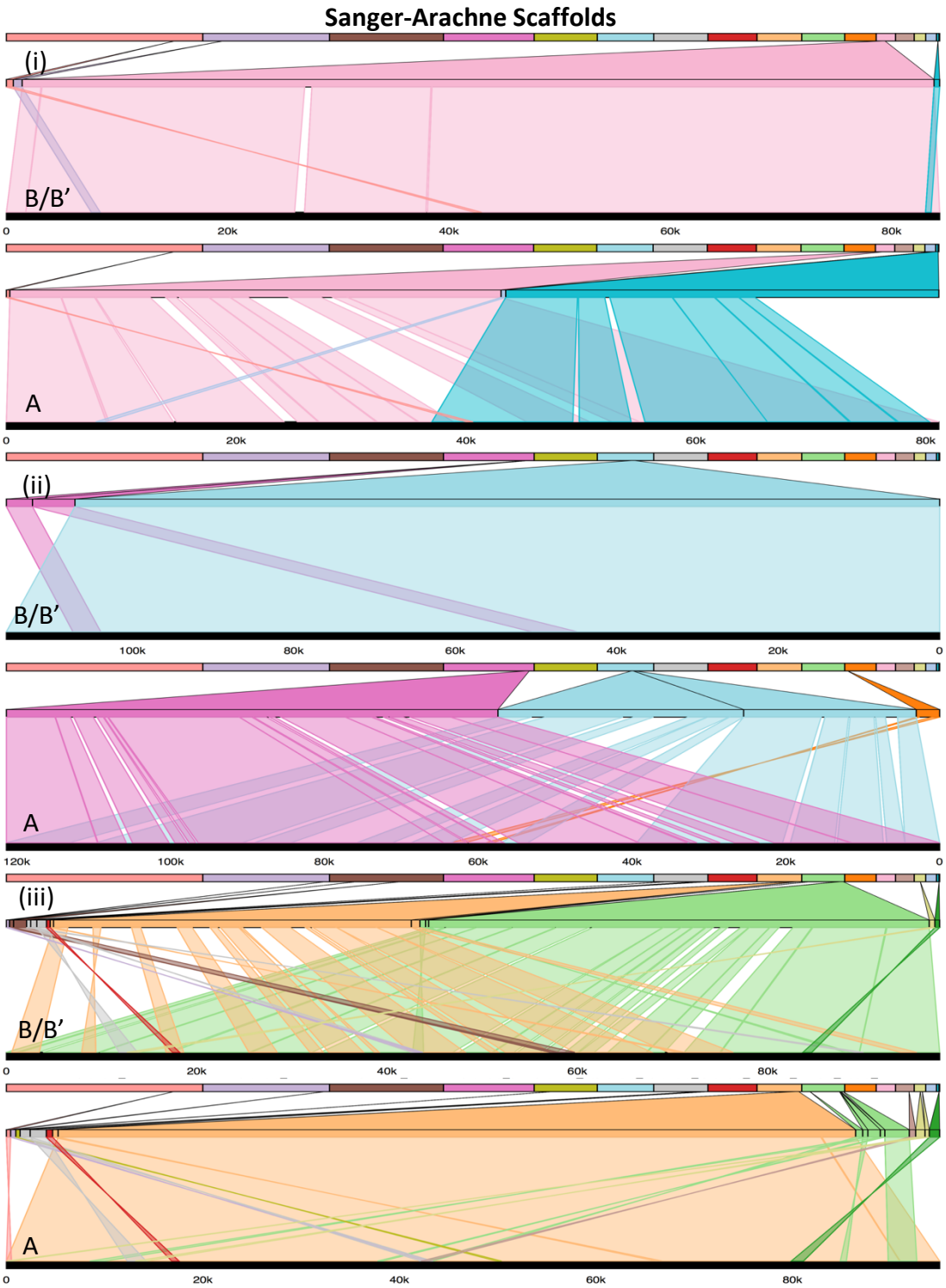

Figure 13: Genome Ribbon representation of the Sanger-Arachne assembly vs three Illumina unitigs from the three sub-genomes. The plot pairs (i), (ii) and (iii) represent three unitig alignments of the Sanger-Arachne scaffolds against the Illumina A or B/B' sub-genomes. Less diverged unitigs between A and B/B' were selected from the Illumina assembly as a standard for this comparison. No two unitig pairs fully represent the Illumina unitigs in the Sanger-Arachne assembly, but (ii) closely represents both B/B' and A unitigs. At least one of each pair is (almost) fully represented. There are also a number of repeated regions (overlapping colours) on the Sanger assembly. These alignments have a minimum 82% identity to the Illumina unitigs.

4. Discussion

4.1 Polyploidy

We have confirmed this culture of *F. cylindrus* (Grunow) Krieger CCMP1102 to have a typical k-mer spectra signature of a polyploid genome and not diploid, as previously thought. K-mer spectra plots show three clear peaks created from the frequencies of distinct k-mers found within the reads (Figure 1). The first (x=53) and second (x~106) peaks of the spectra show unique, heterozygous content and content that appears twice, as would be expected of a diploid organism. However, the third peak shows homozygous content shared between all three sub-genomes, which appears exactly three times; this third spectra peak is distinct, with a mode at a harmonic frequency (x~159) to that of the fundamental frequency (x=53). These k-mers cannot be confused with repetitive content, which would appear at varying frequencies above the triplicated mark and display a trailing tail on the spectra graph (Section 1.2.1). Furthermore, the DBG shows distinct "bubbles" as would be expected of a diploid organism, but alternating sides contain coverage signatures of haploid and diploid genomes (Figure 7). This strongly infers one half of the "bubble" contains twice the k-mer content of the alternative side (but has been collapsed by the graph into a single node). Additionally, we have found twice as many Sanger 8Kb paired-ends mapped to the B/B' sub-genome than to the A sub-genome (Table 3). This is further reinforced by the paired-ends creating twice as many links within and between B/B' than A, despite probable identification errors in the DBG between diverged unitigs of B and B'. *F. cylindrus* has been shown to survive in a range of environmental conditions, from extreme temperatures, high salinity, high Fe, increased pH and even overexposure to UV (Bayer-Giraldi *et al*., 2010; Helbling *et al*., 1996; Mock and Hoch, 2005; Pančić *et al*., 2015). A polyploid genome will lend to persistence and adaptability, as has been shown many times in plants, including through great extinction events (Diallo *et al*., 2016; Madlung, 2013; Ramsey, 2011; Soltis and Burleigh, 2009; Van de Peer *et al*., 2017).

We have shown this culture of *F. cylindrus* (Grunow) Krieger CCMP1102 to have a higher total haploid-specific genomic content than expected at an estimate of 152Mb (Sanger:122.2Mb, PacBio: 119.4). The increase in estimated genomic content further enforces a polyploid hypothesis. Our sub-genome size estimate of 50.7Mb (or haploid estimate on the assumption of triploidy) by k-mer spectra is within a mutually complimentary range (Kim *et al*., 2014) of previous findings based on qPCR of target genes at 57.9Mb (± 16.9Mb; Mock *et al*., 2017). Our findings are a 10Mb decrease in estimated haploid genome size compared to the Sanger-Arachne (61.1Mb) and PacBio-Falcon (59.7Mb) assemblies. Briefly, methods of attaining a haploid genome size, used in assemblers by Mock *et al*. (2017) and Paajanen *et al*. (2017), use alignments of the assembly against itself to identify highly similar "alternative" contigs. These are extracted and the "primary" contigs are assumed to be the size of the haploid genome. However, all unique content will be included in the haploid, including contigs that contain highly divergent A or B/B' content. This is demonstrated in Figure 12 (i), where the whole B/B' unitig (brown) has been scaffolded as part of the PacBio-Falcon haploid genome, but some highly divergent content of A (green) has also been scaffolded on the haploid genome. This mosaic haploid genome structure will confound gene functional assignment and structural variant detection, biasing detection of true alleles, divergence and gene expression. Additionally, previous *de novo* assemblers have made the assumption of diploidy, excluding up to 36Mb (PacBio-Falcon assembly) of low-coverage content (Figures 4, 5 and 6) and losing large

proportions of the A sub-genome. This may have biased findings on allele identification and transcript mapping, as transcripts with a missing A target will map to the next closest B or B' target. In this study, we have used k-mer spectra analysis for an independent genome size estimation, which has the statistical power to grant an assembly independent method of characteristic assessment of the genome (Liu *et al*., 2013; Mapleson *et al*., 2016; Vurture *et al*., 2017).

The Sanger assembly estimated the haplotype-resolved genome to consist of 15.1Mb diverged haplotypes (Mock *et al*., 2017) and the Pacbio assembly estimated 9.1Mb alternate contigs (Mock *et al*., 2017; Paajanen *et al*., 2017). In this study we have shown this level of alternative haplotypes to be a huge underestimate, with a lower-bound estimate of 36Mb (between the A, B and B' sub-genomes) diverged content. Independent Sanger 8kb paired-end library alignment to both A and B/B' sub-genomes in the expected amounts confirms that all sub-genomes were present from the initial Sanger sequencing. Consideration of the genomic characteristics is a fundamental necessity to ensure reliable assembly (Simpson, 2014). The previous assemblers used (ARACHNE; Batzoglou, 2002; Jaffe, 2003 and FALCON; Chin *et al*., 2016) were intended to assemble diploid organisms. This resulted in the algorithms excluding up to 36Mb (PacBio-Falcon assembly) of low-coverage content. Low-coverage content represents diverged sequence seen uniquely between the three sub-genomes (Figure 1). Estimates based on k-mer frequency (Figure 9) suggests 29Mb diverged content between A and B/B', and 7Mb diverged content between B and B'. A graphical comparison of the A and B/B' sub-genomes highlights the vast amount of insertions, deletions and inversions between unitigs (Figure 10 & 11). Local analysis of three fairly contiguous unitigs show an average of only 92% alignment identity, although a true estimate over the whole genome is yet to be produced. Additionally, the A sub-genome (N50 69Kb) is much more contiguous than the B/B' sub-genomes (N50 42Kb); the B/B' sub-genomes are represented in shorter unitigs than the A sub-genome because of mismatches between the B and B' sequences. We predict this diverged genomic content to be a key component in the plasticity of the *F. cylindrus* genome. However, more precise estimations will be attainable with our future work to phase and fully assemble all three sub-genomes. Further transcriptional analysis may lead to a better understanding on how the divergences between A, B and B' lend to *F. cylindrus's* ability to cope under stressful conditions. The implications of this level of diversity may make *F. cylindrus* a very interesting model for evolutionary studies.

Differences between A and B/B' unitig size and contiguity, and misidentification of unitigs in the DBG, may have biased Sanger 8Kb paired-end alignment. Some A content was lost in assembly due to low coverage and the B/B' sub-genome contains a much more fragmented sequence due to mismatches between B and B'. Thus, where a Sanger paired-end would align perfectly to the B/B' genome, but the second site of alignment is missing, the paired-end will align to the next highest identity sequence – the A sub-genome. If alignment thresholds are set to include only the highest hits for both Sanger paired-end alignment sites, these numbers become within the expected range because fewer low quality misalignments to the unexpected sub-genome are being made. Linkage between the paired-ends shows consistent alignment of contiguous sub-genome sequence when links are filtered for quality. These links are formed when Sanger paired-ends align to contigs greater than 1Kb and are confirmed by the alignment of 4 or more paired-ends forming the same link. We expect that these unexpected results are therefore a by-product of unequal sequence representation on the unitigs. There are also some areas of the DBG that have

been misidentified due to the coverage signature of the diverged B and B' sub-genomes being similar to that of the A sub-genome; where a double coverage B/B' node branches into the diverged B and B', the constituent k-mers are unique and are represented by low coverage. In this study we only split the A from the B and B' sub-genomes because they were more readily distinguished in manual identification. The figures presented in Table 2 do not include collapsed content from all three sub-genomes, which was filtered separately. Due to their similarities, separation of the B and B' sub-genomes would have given little content for comparison without full haplotype resolution of the whole genome. In the future, we hope to resolve all three haplotypes, but it is unclear as to whether we would be able to fully resolve whole chromosome-lengths for each haploid. We have observed a number of instances where there is distinction between A, B and B', but the current assembly graph is collapsing large sections of B and B'. Nonetheless, we will be able to use a partially-phased assembly for divergence analysis, as well as more detailed analysis into gene content and transcriptome expression. Futhermore, we expect Sanger paired-end alignment to be improved with scaffolding using PacBio reads, which will bridge gapped regions.

It is unclear from our study how this *F. cylindrus* culture came to be polyploid. However, aneuploidy or Whole Genome Duplication (WGD) events may explain an autopolyploid origin for the ploidy signature we have observed. During past karyotyping, aneuploidy has been recorded in a number of diatom auxospores (Geitler, 1973; Kociolek and Stoermer, 1989). However, aneuploidy of a whole genome is improbable. Alternatively, WGD events are suggested as a driver in diatom evolution through phylogenomic analyses (Parks *et al*., 2017). This has long been the case in plants, which are known to gain an adaptive advantage as polyploids (Diallo *et al*., 2016; Madlung, 2013; Ramsey, 2011; Van de Peer *et al*., 2017). Polyploidisation may occur by meiotic non-reduction, whereby two 2n gametes are produced instead of the expected four 1n gametes. In this scenario, a triploid zygote could have arisen from the fusion of a non-reduced gamete with a reduced gamete. Interestingly, triploid zygotes have also been directly recorded in diatoms (Geitler, 1973), formed by the fusion of three gametes. The divergence between the A and B/B' subgenomes may indicate that they originated from populations that have been isolated for a length of time. Alternatively, hybridisation and allopolyploidy could also lend to the apparent extensive divergence we observed between the A and B/B' sub-genomes. This may indicate an historic hybridisation event between the two origin species. In this scenario, a non-reduced species B genome may have fused with a reduced species A genome to form a triploid zygote. We have not provided any evidence to confirm these speculations, but anticipate that future work into this area will provide essential insights into diatom evolution. In addition, we have only sequenced one culture of *F. cylindrus* and have thus far produced no divergence data, so cannot conclude whether this is a widespread or recent phenomenon, or a product of laboratory culturing. This will become more apparent with future sequencing of *F. cylindrus* cultures, and further phylogenomic and divergence analyses to ascertain the origins of the A and B/B' genomes.

We are unable to determine from our results whether this culture of *F. cylindrus* (Grunow) Krieger CCMP1102 is triploid or hexaploid. As described in Griffiths, *et al.* (2000), problems arise in the pairing of triploid chromosomes during meiosis, resulting in an uneven segregation at either pole. This can result in aneuploidy, which may be deleterious due to extra gene dosage. As *F. cylindrus* is thought to sexually reproduce, albeit rarely, hexaploidy would potentially produce more viable gametes. However, evidence suggests that the

majority of reproduction in diatoms is asexual (Chepurnov et al., 2002; Edlund and Stoermer, 1997; Godhe et al., 2014). The persistence of triploids in populations of plants has been seen many times in clonal species (Bai et al., 2011; Mock et al., 2012). Diatoms are suggested to have a high diversification rate (Bowler et al., 2008), which will persist within asexually reproduced progeny. Additionally, due to this diversification rate, we would expect to see more variation between *F. cylindrus* sub-genome pairs in a hexaploid, which would be reflected in the spectra by further distinct peaks. Instead, we see the same spectra signature from three independent sequencings over a span of 10 years, suggesting triploidy. This ambiguity could be definitively answered using the old karyotyping techniques of Geitler (1973) or flow cytometry as used by von Dassow et al. (2008).

Polyploidy has long been recognised as a powerful driver of evolution in plants, conferring greater adaptive potential when compared to their progenitors (Diallo *et al*., 2016; Madlung, 2013; Ramsey, 2011; Soltis and Burleigh, 2009; Van de Peer *et al*., 2017). In particular, polyploidy has an association with disruptive, new and harsh environments (Ramsey, 2011; Soltis and Burleigh, 2009; Van de Peer *et al*., 2017). It should come as no surprise that diatoms, which typically inhabit extreme environments, have long been recorded as polyploids (Chepurnov *et al*., 2002; Geitler, 1973; Koester *et al*., 2010; Mann, 1994; von Dassow *et al*., 2008). Many instances of ploidy have been recorded in diatom gametes (Geitler, 1973; Mann, 1994; von Dassow *et al*., 2008), but this study backs Koester *et al*. (2010) in confirming ploidy in a mature clonal culture.

4.2 Improved methodology

We have demonstrated the value of a hybrid assembly using Illumina short reads and a Sanger 8Kb paired-end library in creating contiguous assemblies; the benefits of utilising existing Sanger data, when paired with re-sequencing for modern assembly, should not be overlooked. The sensitive base pair accuracy of Illumina short reads gives greater confidence to gene identification and transcriptome mapping for further down-stream analysis (Chaisson *et al*., 2018; Koren *et al*., 2012; Rhoads and Au, 2015; Seo *et al*., 2016; Zimin *et al*., 2017). Coupled with the alignment of the Sanger 8Kb paired-end library, greater contiguity can be achieved. Future utilisation of PacBio long reads in this assembly will scaffold unitigs and close gaps for greater improvement; while PacBio reads do not present the accuracy (Hackl *et al*., 2014; Pootakham *et al*., 2017) required for detailed analyses (such as divergence), longer read-lengths have been shown to aid scaffolding of high base-pair accuracy short reads (Koren *et al*., 2012; Rhoads and Au, 2015; Zimin *et al*., 2017).

A k-mer analysis based approach has proven invaluable for character assessment of the genome and assemblies. This method provides an assembly bias free way of assessing initial characteristics of unprocessed reads (Liu *et al*., 2013; Mapleson *et al*., 2016). Additionally, k-mer analysis provides a crucial insight during assembly and content comparisons by leveraging greater statistical power (Mapleson *et al*., 2016). The use of the DBG allows for all content to be retained and manually checked, however, it complicates the production of contiguous sequences from unitigs. The complication of the assembly graph prevents a single walk through the nodes, so no single consensus is produced with ease. Despite minor shortcomings, this information has allowed us to tailor further assembly and analysis to the specifics of this genome. We recommend further HTS analyses on diatoms to use this approach for initial character and structural assessments of the genome before assembly.

This study intended to address discrepancies between previous assemblies by sequencing of a new, hybrid assembly. Illumina short-reads are used due to their high base-pair accuracy (Chaisson *et al*., 2018; Seo *et al*., 2016), which we have partially assembled for future downstream analyses. Our findings highlight the need for old work to be re-analysed using this new assembly. Further work will re-map RNA-seq reads to understand expression patterns for this potentially polyploid genome. We would also hope to see a broader variety of *F. cylindrus* and other diatoms re-sequenced to understand how widespread polyploidy is, particularly between cultures, stresses and life-cycle changes. The sequencing of further *F. cylindrus* will also broaden the prospects of genomic population analyses, currently unachievable due to limited sequence availability.

4.3 Conclusions

We have shown, with multiple lines of evidence, using independent methods that this culture of *F. cylindrus* (Grunow) Krieger CCMP1102 has genomic characteristics compatible with those of a polyploid genome. We highlight three distinct sub-genomes, the A sub-genome, which appears highly diverged from the B and B' sub-genomes. Using k-mer spectra and KCI on new Illumina reads, and reusing a Sanger 8Kb paired-end library, we have combined the best of the sequencing data available to us to produce the start of a reliable genome assembly. We suggest the previous allelic expression will have been compromised due to the unknown complexities of this genome. Our fit-for-purpose assembly will be used for future analysis into gene mapping and expression of *F. cylindrus* to uncover the genomic response to environmental stresses.

High levels of intraspecific morphological variability have been found in the diatom species *Phaeodactylum tricornutum* (Martino *et al*., 2007), but no population studies have thus far been performed for genomic variability in any diatom species. The deficit of genomic information on such an environmentally important clade of organisms hinders our understanding of their adaptability and evolution. Furthermore, our study has underpinned a necessity to cast out historical assumptions and rethink the survival mechanisms and evolution of the unorthodox diatom genome.

References:

Alderkamp, A.-C., Kulk, G., Buma, A.G.J., Visser, R.J.W., Van Dijken, G.L., Mills, M.M., Arrigo, K.R., 2012. The effect of iron limitation on the photophysiology of *Phaeocystis antarctica* (Prymnesiophyceae) and *Fragilariopsis cylindrus* (Bacillariophyceae) under dynamic irradiance. *Journal of Phycology,* 48, pp. 45–59. doi: https://doi.org/10.1111/j.1529-8817.2011.01098.x

Alexeyenko, A., Nystedt, B., Vezzi, F., Sherwood, E., Ye, R., Knudsen, B., Simonsen, M., Turner, B., de Jong, P., Wu, C.-C., Lundeberg, J., 2014. Efficient de novo assembly of large and complex genomes by massively parallel sequencing of Fosmid pools. *BMC Genomics,* 15(439). doi: https://doi.org/10.1186/1471-2164-15-439

AltschuP, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology,* 215, pp. 403–410.

Amato, A., Dell'Aquila, G., Musacchia, F., Annunziata, R., Ugarte, A., Maillet, N., Carbone, A., Ribera d'Alcalà, M., Sanges, R., Iudicone, D., Ferrante, M.I., 2017. Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions. *Scientific Reports,* 7. doi: https://doi.org/10.1038/s41598-017-03741-6

Armbrust, E.V., Chisholm, S.W., 1992. Patterns of Cell Size Change in Marine Centric Diatom: Variability Evolving from Clonal Isolates. *Journal of Phycology,* 28, pp. 146–156.

Bai, Z., Liu, F., Li, J., Yue, G.H., 2011. Identification of Triploid Individuals and Clonal Lines in *Carassius Auratus* Complex Using Microsatellites. *International Journal of Biological Sciences,* 7, pp. 279–285. doi: https://doi.org/10.7150/ijbs.7.279

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., Lander, E.S., 2002. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research,* 12, pp. 177–189. doi: https://doi.org/10.1101/gr.208902

Bayer-Giraldi, M., Uhlig, C., John, U., Mock, T., Valentin, K., 2010. Antifreeze proteins in polar sea ice diatoms: diversity and gene expression in the genus *Fragilariopsis*: Cold adaptation in the polar genus *Fragilariopsis*. *Environmental Microbiology,* 12, pp. 1041–1052. doi: https://doi.org/10.1111/j.1462-2920.2009.02149.x

Bird, K.A., VanBuren, R., Puzey, J.R., Edger, P.P., 2018. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist*. doi: https://doi.org/10.1111/nph.15256

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O., Guo, L., Collins, R.L., Fan, X., Wen, J., Handsaker, R.E., Fairley, S., Kronenberg, Z.N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A.M., Hastie, A., Antaki, D., Audano, P., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C.-S., Chong, Z., Chuang, N.T., Lambert, C.C., Church, D.M., Clarke, L., Farrell, A., Flores, J., Galeev, T., Gorkin, D., Gujral, M., Guryev, V., Haynes Heaton, W., Korlach, J., Kumar, S., Kwon, J.Y., Lee, J.E., Lee, J., Lee, W.-P., Lee, S.P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K.M., Navarro, F., Nelson, B.J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A., Qiu, Y., Rosanio, G., Ryan, M., Stutz, A., Spierings, D.C.J., Ward, A., Welch, A.E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C.L., Ren, B., Flicek, P., Chen, K., Gerstein, M.B., Kwok, P.-Y., Lansdorp, P.M., Marth, G., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S.E., Talkowski, M., Mills, R.E., Marschall, T., Korbel, J.O., Eichler, E.E., Lee, C., 2018. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *BioRxiv,* doi: https://doi.org/10.1101/193144

Chepurnov, V., Mann, D., 2000. Variation in the sexual behaviour of *Achnanthes longipes* (Bacillariophyta). *European Journal of Phycology,* 35, pp. 213–223. doi: https://doi.org/10.1080/09670260010001735821

Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K., Danielidis, D.B., 2002. Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology,* 38, pp. 1004–1019. doi: https://doi.org/10.1046/j.1529-8817.2002.t01-1-01233.x

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M., Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., Schatz, M.C., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods,* 13, pp. 1050–1054. doi: https://doi.org/10.1038/nmeth.4035

Clavijo, B., Garcia, G.A., Wright, J., Heavens, D., Barr, K., Yanes, L., Di Palma, F., 2017. W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *BioRxiv,* doi: https://doi.org/10.1101/110999

Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology,* 29, pp. 987–991. doi: https://doi.org/10.1038/nbt.2023

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., dePamphilis, C.W., 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research,* 16, pp. 738–749. doi: https://doi.org/10.1101/gr.4825606

Diallo, A.M., Nielsen, L.R., Kjær, E.D., Petersen, K.K., Ræbild, A., 2016. Polyploidy can Confer Superiority to West African *Acacia senegal* (L.) Willd. Trees. *Frontiers in Plant Science,* 7. doi: https://doi.org/10.3389/fpls.2016.00821

Doležel, J., Greilhuber, J., Suda, J., 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols,* 2, pp. 2233–2244.

Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Li, Y., Zhao, X., Wang, J., Liu, K., Qin, P., Yang, X., Zhu, L., Li, S., Liang, C., 2017. Sequencing and de novo assembly of a near complete *indica* rice genome. *Nature Communications,* 8(15324). doi: https://doi.org/10.1038/ncomms15324

Edlund, M.B., Stoermer, E.F., 1997. Ecological, evolutionary, and systematic significance of diatom life histories. *Journal of Phycology,* 33, pp. 897–918. doi: https://doi.org/10.1111/j.0022-3646.1997.00897.x

Ellegaard, M., Godhe, A., Ribeiro, S., 2018. Time capsules in natural sediment archives-Tracking phytoplankton population genetic diversity and adaptation over multidecadal timescales in the face of environmental change. *Evolutionary Applications,* 11, pp. 11–16. doi: https://doi.org/10.1111/eva.12513

Falasco, E., Bona, F., Badino, G., Hoffmann, L., Ector, L., 2009. Diatom teratological forms and environmental alterations: a review. *Hydrobiologia,* 623, pp. 1–35. doi: https://doi.org/10.1007/s10750-008-9687-3

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science,* 281, pp. 237–240. doi: https://doi.org/10.1126/science.281.5374.237

Friedl, T., 1995. Inferring taxonomic positions and testing genus level assignments in coccoid green lichen algae: A phylogenetic analysis of 18s ribosomal RNA sequences from *Dictyochloropsis reticulata* and from members of the genus *Myrmecia* (Chlorophyta, Trebouxiophyceae cl. nov.). *Journal of Phycology,* 31, pp. 632–639. doi: https://doi.org/10.1111/j.1529-8817.1995.tb02559.x

Fuchs, N., Scalco, E., Kooistra, W.H.C.F., Assmy, P., Montresor, M., 2013. Genetic characterization and life cycle of the diatom *Fragilariopsis kerguelensis*. *European Journal of Phycology,* 48, pp. 411–426. doi: https://doi.org/10.1080/09670262.2013.849360

Geitler, L., 1973. Auxosporenbildung und Systematik bei pennaten Diatomeen und die Cytologie yon *Cocconeis*-Sippen. *Osterreichische Botanische Zeitschrif,* 122, pp. 299-321.

Griffiths A. J. F., Miller J. H., Suzuki D. T., *et al.*, 2000. *An Introduction to Genetic Analysis.* 7th edition. New York: W. H. Freeman. Aberrant euploidy.

Guppy, M., Withers, P., 2007. Metabolic depression in animals: physiological perspectives and biochemical generalizations. *Biological Reviews,* 74, pp. 1–40. doi: https://doi.org/10.1111/j.1469-185X.1999.tb00180.x

Hackl, T., Hedrich, R., Schultz, J., Förster, F., 2014. *proovread* : large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics,* 30(21), pp. 3004–3011. doi: https://doi.org/10.1093/bioinformatics/btu392

Harnstrom, K., Ellegaard, M., Andersen, T.J., Godhe, A., 2011. Hundred years of genetic structure in a sediment revived diatom population. *Proceedings of the National Academy of Sciences,* 108, pp. 4252–4257. doi: https://doi.org/10.1073/pnas.1013528108

Helbling, E.W., Chalker, B.E., Dunlap, W.C., Holm-Hansen, O., Villafañe, V.E., 1996. Photoacclimation of antarctic marine diatoms to solar ultraviolet radiation. *Journal of Experimental Marine Biology and Ecology,* 204, pp. 85–101. doi: https://doi.org/10.1016/0022-0981(96)02591-9

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., Lander, E.S., 2003. Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Research,* 13, pp. 91–96. doi: https://doi.org/10.1101/gr.828403

Kang, S.-H., Fryxell, G., 1992. *Fragilariopsis cylindrus* (Grunow) Krieger: The most abundant diatom in water column assemblages of Antarctic marginal ice-edge zones. *Polar Biology,* 12. doi: https://doi.org/10.1007/BF00236984

Kim, J., Roh, J., Kwon, D., Kim, Y., Yoon, K.A., Yoo, S., Noh, S.-J., Park, J., Shin, E., Park, M.-Y., Lee, S., 2014. Estimation of the genome sizes of the chigger mites *Leptotrombidium pallidum* and *Leptotrombidium scutellare* based on quantitative PCR and k-mer analysis. *Parasites & Vectors,* 7(279). doi: https://doi.org/10.1186/1756-3305-7-279

Kociolek, J.P., Stoermer, E.F., 1989. Chromosome numbers in diatoms: a review. *Diatom Research,* 4, pp. 47–54. doi: https://doi.org/10.1080/0269249X.1989.9705051

Koester, J.A., Swalwell, J.E., von Dassow, P., Armbrust, E.V., 2010. Genome size differentiates co-occurring populations of the planktonic diatom *Ditylum brightwellii* (Bacillariophyta). *BMC Evolutionary Biology,* 10(1). doi: https://doi.org/10.1186/1471-2148-10-1

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology,* 30, pp. 693–700. doi: https://doi.org/10.1038/nbt.2280

Kropuenske, L.R., Mills, M.M., van Dijken, G.L., Alderkamp, A.-C., Mine Berg, G., Robinson, D.H., Welschmeyer, N.A., Arrigo, K.R., 2010. Strategies and rates of photoacclimation in two major southern ocean phytoplankton taxa: *Phaeocystis antarctica* (Haptophyta) and *Fragilariopsis cylindrus* (Bacillariophyceae). *Journal of Phycology,* 46, pp. 1138–1151. doi: https://doi.org/10.1111/j.1529-8817.2010.00922.x

Kropuenske, L.R., Mills, M.M., van Dijken, G.L., Bailey, S., Robinson, D.H., Welschmeyer, N.A., Arrigoa, K.R., 2009. Photophysiology in two major Southern Ocean phytoplankton taxa: Photoprotection in *Phaeocystis antarctica* and *Fragilariopsis cylindrus*. *Limnology and Oceanography,* 54, pp. 1176–1196. doi: https://doi.org/10.4319/lo.2009.54.4.1176

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. *Genome Biology,* 9.

Lander, E.S., Waterman, M.S., 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics,* 2, pp. 231–239. doi: https://doi.org/10.1016/0888-7543(88)90007-9

Leitch, A.R., Leitch, I.J., 2008. Genomic Plasticity and the Diversity of Polyploid Plants. Science, 320, pp. 481–483. doi: https://doi.org/10.1126/science.1153585

Lennon, J.T., Jones, S.E., 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology,* 9, pp. 119–130. doi: https://doi.org/10.1038/nrmicro2504

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

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., Fan, W., 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics,* 11, pp. 25–37. doi: https://doi.org/10.1093/bfgp/elr035

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Ye, Yu, Y., Du, H., Qi, M., Li, Yan, Lu, H., Yu, H., Cui, Y., Wang, N., Chen, C., Wu, H., Zhao, Y., Zhang, J., Li, Yiwen, Zhou, W., Zhang, B., Hu, W., van Eijk, M.J.T., Tang, J., Witsenboer, H.M.A., Zhao, S., Li, Z., Zhang, A., Wang, D., Liang, C., 2018. Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature,* 557, pp. 424–428. doi: https://doi.org/10.1038/s41586-018-0108-0

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., Fan, W., 2013. Estimation of genomic characteristics by analyzing k- mer frequency in de novo genome projects.

Love, R.R., Weisenfeld, N.I., Jaffe, D.B., Besansky, N.J., Neafsey, D.E., 2016. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics,* 17. doi: https://doi.org/10.1186/s12864-016-2531-7

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1. doi: https://doi.org/10.1186/2047-217X-1-18

Madlung, A., 2013. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity,* 110, pp. 99–104. doi: https://doi.org/10.1038/hdy.2012.79

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A., Bowler, C., 2016. Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences,* 113, pp. E1516–E1525. doi: https://doi.org/10.1073/pnas.1509523113

Mann, D.G., 1994. Auxospore formation, reproductive plasticity and cell structure in *Navicula ulvacea* and the resurrection of the genus *Dickieia* (Bacillariophyta).

European Journal of Phycology, 29, pp. 141–157. doi: https://doi.org/10.1080/09670269400650591

Mann, D.G., Chepurnov, V.A., Idei, M., 2003. Mating system, sexual reproduction, and auxosporulation in the anomalous raphid diatom *Eunotia* (Bacillariophyta). *Journal of Phycology,* 39, pp. 1067–1084. doi: https://doi.org/10.1111/j.0022-3646.2003.03-011.x

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., Clavijo, B.J., 2016. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics,* 33(4), pp. 574-576. doi: https://doi.org/10.1093/bioinformatics/btw663

Martino, A.D., Meichenin, A., Shi, J., Pan, K., Bowler, C., 2007. Genetic and phenotypic characterization of *Phaeodactylum tricornutum* (Bacillariophyceae). *Journal of Phycology,* 43, pp. 992–1009. doi: https://doi.org/10.1111/j.1529-8817.2007.00384.x

Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V., Gurevich, A., 2016. Icarus: visualizer for de novo assembly evaluation. *Bioinformatics,* 32(21), pp. 3321–3323. doi: https://doi.org/10.1093/bioinformatics/btw379

Mock, K.E., Callahan, C.M., Islam-Faridi, M.N., Shaw, J.D., Rai, H.S., Sanderson, S.C., Rowe, C.A., Ryel, R.J., Madritch, M.D., Gardner, R.S., Wolf, P.G., 2012. Widespread Triploidy in Western North American Aspen (*Populus tremuloides*). *PLoS ONE,* 7(10). doi: https://doi.org/10.1371/journal.pone.0048406

Mock, T., Hoch, N., 2005. Long-Term Temperature Acclimation of Photosynthesis in Steady-State Cultures of the Polar Diatom *Fragilariopsis cylindrus*. *Photosynthesis Research,* 85, pp. 307–317. doi: https://doi.org/10.1007/s11120-005-5668-9

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E.A., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J.A., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I.V., 2017. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature,* 541, pp. 536–540. doi: https://doi.org/10.1038/nature20803

Mouget, J.-L., Gastineau, R., Davidovich, O., Gaudin, P., Davidovich, N.A., 2009. Light is a key factor in triggering sexual reproduction in the pennate diatom *Haslea ostrearia*: Light induction of sexual reproduction in diatoms. *FEMS Microbiology Ecology,* 69, pp. 194–201. doi: https://doi.org/10.1111/j.1574-6941.2009.00700.x

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.-H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D., Venter, J.C., 2000. A Whole-Genome Assembly of *Drosophila*. *Science,* 287, pp. 2196–2204. doi: https://doi.org/10.1126/science.287.5461.2196

Nattestad, M., Chin, C.-S., Schatz, M.C., 2016. Ribbon: Visualizing complex genome alignments and structural variation. *BioRxiv,* doi: https://doi.org/10.1101/082123

Nelson, D.M., Tréguer, P., Brzezinski, M.A., Leynaert, A., Quéguiner, B., 1995. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles,* 9, pp. 359–372. doi: https://doi.org/10.1029/95GB01070

Paajanen, P., Strauss, J., van Oosterhout, C., McMullan, M., Clark, M.D., Mock, T., 2017. Building a locally diploid genome and transcriptome of the diatom *Fragilariopsis cylindrus*. *Scientific Data,* 4(170149). doi: https://doi.org/10.1038/sdata.2017.149

Pančić, M., Hansen, P.J., Tammilehto, A., Lundholm, N., 2015. Resilience to temperature and pH changes in a future climate change scenario in six strains of the polar diatom *Fragilariopsis cylindrus*. *Biogeosciences Discussions,* 12, pp. 4627–4654. doi: https://doi.org/10.5194/bgd-12-4627-2015

Parks, M., Nakov, T., Ruck, E.C., Wickett, N.J., Alverson, A.J., 2017. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *BioRxiv,* doi: doi: https://doi.org/10.1101/181115

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Sonthirod, C., Naktang, C., Thongtham, N., Tangphatsornruang, S., 2017. High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Scientific Reports,* 7. doi: https://doi.org/10.1038/s41598-017-03139-4

Ramsey, J., 2011. Polyploidy and ecological adaptation in wild yarrow. Proceedings of the *National Academy of Sciences,* 108, pp. 7096–7101. doi: https://doi.org/10.1073/pnas.1016631108

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics & *Bioinformatics,* 13, pp. 278–289. doi: https://doi.org/10.1016/j.gpb.2015.08.002

Sarthou, G., Timmermans, K.R., Blain, S., Tréguer, P., 2005. Growth physiology and fate of diatoms in the ocean: a review. *Journal of Sea Research,* 53, pp. 25–42. doi: https://doi.org/10.1016/j.seares.2004.01.007

Seo, J.-S., Rhie, A., Kim, Junsoo, Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, Jihye, Kuk, J., Park, G.H., Kim, Juhyeok, Ryu, H., Kim, Jongbum, Roh, M., Baek, J., Hunkapiller, M.W., Korlach, J., Shin, J.-Y., Kim, C., 2016. De novo assembly and phasing of a Korean human genome. *Nature,* 538, pp. 243–247. doi: https://doi.org/10.1038/nature20098

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology,* 26, pp. 1135–1145. doi: https://doi.org/10.1038/nbt1486

Simpson, J.T., 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics,* 30(9), pp. 1228–1235. doi: https://doi.org/10.1093/bioinformatics/btu023

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, I., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research,* 19, pp. 1117–1123. doi: https://doi.org/10.1101/gr.089532.108

Smetacek, V., 2012. Making sense of ocean biota: How evolution and biodiversity of land organisms differ from that of the plankton. *Journal of Biosciences,* 37, pp. 589–607. doi: https://doi.org/10.1007/s12038-012-9240-4

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., dePamphilis, C.W., Wall, P.K., Soltis, P.S., 2009. Polyploidy and angiosperm diversification. *American Journal of Botany,* 96, pp. 336–348. doi: https://doi.org/10.3732/ajb.0800079

Soltis, D.E., Burleigh, J.G., 2009. Surviving the K-T mass extinction: New perspectives of polyploidization in angiosperms. *Proceedings of the National Academy of Sciences,* 106, pp. 5455–5456. doi: https://doi.org/10.1073/pnas.0901994106

Sterkers, Y., Lachaud, L., Bourgeois, N., Crobu, L., Bastien, P., Pagès, M., 2012. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in *Leishmania*: Mosaic

aneuploidy in *Leishmania*. *Molecular Microbiology*, 86, pp. 15–23. doi: https://doi.org/10.1111/j.1365-2958.2012.08185.x

The International Wheat Genome Sequencing Consortium (IWGSC), Mayer, K.F.X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., Endo, T.R., Kubalakova, M., ihalikova, J., Dubska, Z., Vrana, J., perkova, R., imkova, H., Febrer, M., Clissold, L., McLay, K., Singh, K., Chhuneja, P., Singh, N.K., Khurana, J., Akhunov, E., Choulet, F., Alberti, A., Barbe, V., Wincker, P., Kanamori, H., Kobayashi, F., Itoh, T., Matsumoto, T., Sakai, H., Tanaka, T., Wu, J., Ogihara, Y., Handa, H., Maclachlan, P.R., Sharpe, A., Klassen, D., Edwards, D., Batley, J., Olsen, O.-A., Sandve, S.R., Lien, S., Steuernagel, B., Wulff, B., Caccamo, M., Ayling, S., Ramirez-Gonzalez, R.H., Clavijo, B.J., Wright, J., Pfeifer, M., Spannagl, M., Martis, M.M., Mascher, M., Chapman, J., Poland, J.A., Scholz, U., Barry, K., Waugh, R., Rokhsar, D.S., Muehlbauer, G.J., Stein, N., Gundlach, H., Zytnicki, M., Jamilloux, V., Quesneville, H., Wicker, T., Faccioli, P., Colaiacovo, M., Stanca, A.M., Budak, H., Cattivelli, L., Glover, N., Pingault, L., Paux, E., Sharma, S., Appels, R., Bellgard, M., Chapman, B., Nussbaumer, T., Bader, K.C., Rimbert, H., Wang, S., Knox, R., Kilian, A., Alaux, M., Alfama, F., Couderc, L., Guilhot, N., Viseux, C., Loaec, M., Keller, B., Praud, S., 2014. A chromosome-based draft sequence of the hexaploid bread wheat *(Triticum aestivum)* genome. *Science,* 345, pp. 1251788–1251788. doi: https://doi.org/10.1126/science.1251788

Thomas, W.H., Dodson, A.N., Reid, F.M.H., 1978. Diatom productivity compared to other algae in natural marine phytoplankton assemblages. *Journal of Phycology,* 14, pp. 250–253. doi: https://doi.org/10.1111/j.1529-8817.1978.tb00294.x

Van de Peer, Y., Mizrachi, E., Marchal, K., 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics,* 18, pp. 411–424. doi: https://doi.org/10.1038/nrg.2017.26

von Dassow, P., Petersen, T.W., Chepurnov, V.A., Virginia Armbrust, E., 2008. Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus *Thalassiosira* (Bacillariophyceae). *Journal of Phycology,* 44, pp. 335–349. doi: https://doi.org/10.1111/j.1529-8817.2008.00476.x

von Stosch, H., Fecher, K., 1979. “Internal thecae” of *Eunotia soleirolii* (Bacillariophyceae): Development, structure and function as resting spores. *Journal of Phycology, 15,* pp. 233-243.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., Schatz, M.C., 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics,* 33(14), pp. 2202–2204. doi: https://doi.org/10.1093/bioinformatics/btx153

Wajid, B., Serpedin, E., 2012. Review of General Algorithmic Features for Genome Assemblers for Next Generation Sequencers. *Genomics, Proteomics & Bioinformatics,* 10, pp. 58–73. doi: https://doi.org/10.1016/j.gpb.2012.05.006

Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S., Burbano, H.A., 2017. nQuire: A Statistical Framework For Ploidy Estimation Using Next Generation Sequencing. *BMC Bioinformatics.* doi: https://doi.org/10.1101/143537

Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E., 2015. Bandage: interactive visualization of *de novo* genome assemblies *Bioinformatics,* 31(20), pp. 3350–3352. doi: https://doi.org/10.1093/bioinformatics/btv383

Xu, H., Zhang, W., Zhang, T., Li, J., Wu, X., Dong, L., 2014. Determination of Ploidy Level and Isolation of Genes Encoding Acetyl-CoA Carboxylase in Japanese Foxtail *(Alopecurus japonicus)*. *PLoS ONE,* 9(12). doi: https://doi.org/10.1371/journal.pone.0114712

Zerbino, D.R., 2010. Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies, in: Baxevanis, A.D., Davison, D.B., Page, R.D.M., Petsko, G.A., Stein, L.D., Stormo, G.D. (Eds.), *Current Protocols in Bioinformatics.* John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: https://doi.org/10.1002/0471250953.bi1105s31

Zimin, A.V., Puiu, D., Hall, R., Kingan, S., Clavijo, B.J., Salzberg, S.L., 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience,* 6, pp. 1–7. doi: https://doi.org/10.1093/gigascience/gix097

Appendices:

Appendix A: Python script to assign haplotypes based on coverage.

```
import sys

# Usage: python colour-by-haplotype.py tsv_file gfa_file cvg_file >
output_filename

gfa_file = sys.argv[2]        # colour empty gfa file to add colours to
cvg_file = sys.argv[3]          #cvg t

ABC_hash={}
C_hash={}
AB_hash={}
undefined={}
with open(cvg_file, 'r') as cvg:
        for line in cvg:
                sline=line.split("\t")
                name = sline[0]
                seq = sline[1]
                list_0=[]
                list_a=[]
                list_b=[]
                list_c=[]
                for number in seq.split():
                        if int(number) >=0 and int(number) <80:
                                list_a.append (number)
                        elif int(number)>=80 and int(number)<130:
                                list_b.append (number)
                        elif int(number)>=130:
                                list_c.append (number)
                        else:
                                list_0.append (number)

                contam = len(list_0)
                a = len(list_a)
                b = len(list_b)
                c = len(list_c)
                abc = float(a+b+c)

                a_pcnt = float((a/abc)*100)
                b_pcnt = float((b/abc)*100)
                c_pcnt = float((c/abc)*100)
```

```
                if (b_pcnt < float(10) and a_pcnt < float(10)) or c_pcnt
== float(100):
                        ABC_hash[name]=name

                elif ((a_pcnt + c_pcnt) > b_pcnt) and (a_pcnt > 30) and
(b_pcnt < float(15))and (c_pcnt != float(100)):
                        C_hash[name]=name

                elif ((c_pcnt > a_pcnt) and (c_pcnt != float(100))) or
((((a_pcnt + b_pcnt) > float(90)) and b_pcnt > float(10))):
                        AB_hash[name]=name

                else:
                        undefined[name]=name

# open gfa file
with open(gfa_file, 'r') as gfa_in:
        next(gfa_in)
        for line in gfa_in:
                line_stripped = line.strip()
                fields = line_stripped.split()
                edge = fields[1]
                if line.startswith("S"):
                        if edge in C_hash.keys():
                                print
"{0}\t{1}\t{2}\t{3}\tCL:Z:blue\t61_260_contigs_raw.good2.fasta\tKC:i:{4}".
format(fields[0],fields[1],fields[2],fields[3])
                        elif edge in AB_hash.keys():
                                print
"{0}\t{1}\t{2}\t{3}\tCL:Z:red\t61_260_contigs_raw.good2.fasta\tKC:i:{4}".f
ormat(fields[0],fields[1],fields[2],fields[3])
                        elif edge in ABC_hash.keys():
                                print
"{0}\t{1}\t{2}\t{3}\tCL:Z:purple\t61_260_contigs_raw.good2.fasta\tKC:i:{4}
".format(fields[0],fields[1],fields[2],fields[3])
                        else:
                                print
"{0}\t{1}\t{2}\t{3}\tCL:Z:grey\t61_260_contigs_raw.good2.fasta\tKC:i:{4}".
format(fields[0],fields[1],fields[2],fields[3])
                else:
                        print(line)
```

Appendix B: Python script to cut selected unitigs from sub-genomes.

```
import sys
from Bio import SeqIO

if len(sys.argv) != 6:
   print "Usage: {0} exonerate_file query_fasta subject_fasta query_out
subject_out".format(sys.argv[0])
   sys.exit()

EXONERATE = sys.argv[1]
Q_FASTA = sys.argv[2]
S_FASTA = sys.argv[3]
Q_OUT = sys.argv[4]
S_OUT = sys.argv[5]

seqiter_q = SeqIO.parse(open(Q_FASTA), 'fasta')
q_record_dict = SeqIO.to_dict(SeqIO.parse(Q_FASTA, "fasta"))
print "Loaded {0} seqs from {1}".format(len(q_record_dict), Q_FASTA)

seqiter_s = SeqIO.parse(open(S_FASTA), 'fasta')
s_record_dict = SeqIO.to_dict(SeqIO.parse(S_FASTA, "fasta"))
print "Loaded {0} seqs from {1}".format(len(s_record_dict), S_FASTA)

q_out = open(Q_OUT, 'w')
s_out = open(S_OUT, 'w')

count = 1

print "Reading file {0}".format(EXONERATE)
with open(EXONERATE, 'r') as f:
 for line in f:
   line = line.strip()
   (query_bit, subject_bit, pid, score) = line.split("\t")

   (qid, tmp2, qstrand, tmp3) = query_bit.split(":")
   (qstart, qend) = tmp2.split("-")
   (qalen, qlen) = tmp3.split("-")

   (sid, tmp4, sstrand, tmp5) = subject_bit.split(":")
   (sstart, send) = tmp4.split("-")
   (salen, slen) = tmp5.split("-")

   if int(qalen) > 1000:

     print "Parsing line: {0}".format(line)
     print "getting subseq {0}-{1} from {2}".format(qstart, qend, qid)
     print "getting subseq {0}-{1} from {2}".format(sstart, send, sid)
```

```
     qseq = q_record_dict[qid]
     trim_qseq = qseq[int(qstart):int(qend)]
     trim_sseq = sseq[int(sstart):int(send)]


      if int(sstart) > int(send):
        tmp = sstart
        sstart = send
        send = tmp

     sseq = s_record_dict[sid]
     trim_sseq = sseq[int(sstart):int(send)]

print "id={0}_{1}".format(qid, count)
     print trim_qseq.seq

     print "id={0}_{1}".format(sid, count)
     print trim_sseq.seq

     print "writing to q file as {0}_{1}".format(qid, count)
     print "writing to s file as {0}_{1}".format(sid, count)
     q_out.write(">{0}_{1}\n{2}\n".format(qid, count, trim_qseq.seq))
     s_out.write(">{0}_{1}\n{2}\n".format(sid, count, trim_sseq.seq))
     count = count + 1
     print "**"

q_out.close()
s_out.close()

print "{0} sequences written.".format(count-1)
print "DONE"
```