G3, 2022, 12(10), jkac210
https://doi.org/10.1093/g3journal/jkac210
Advance Access Publication Date: 18 August 2022
Investigation

Recovering individual haplotypes and a contiguous genome assembly from pooled long-read sequencing of the diamondback moth (Lepidoptera: Plutellidae)

Ivy Whiteford [1],* Arjen E. van't Hof [1] Ritesh Krishna [1,2] Thea Marubbi,[3] Stephanie Widdison,[4] Ilik J. Saccheri [1] Marcus Guest [5] Neil I. Morrison [3] Alistair C. Darby [1]

[1]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
[2]IBM Research UK, STFC Daresbury Laboratory, Warrington WA4 4AD, UK
[3]Oxitec Ltd., Abingdon OX14 4RQ, UK
[4]General Bioinformatics, Jealott's Hill International Research Centre, Bracknell RG42 6EY, UK
[5]Syngenta, Jealott's Hill International Research Centre, Bracknell, RG42 6EY, UK

*Corresponding author: Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK. email: hlswhit2@liverpool.ac.uk

Abstract

The assembly of divergent haplotypes using noisy long-read data presents a challenge to the reconstruction of haploid genome assemblies, due to overlapping distributions of technical sequencing error, intralocus genetic variation, and interlocus similarity within these data. Here, we present a comparative analysis of assembly algorithms representing overlap-layout-consensus, repeat graph, and de Bruijn graph methods. We examine how postprocessing strategies attempting to reduce redundant heterozygosity interact with the choice of initial assembly algorithm and ultimately produce a series of chromosome-level assemblies for an agricultural pest, the diamondback moth, *Plutella xylostella* (L.). We compare evaluation methods and show that BUSCO analyses may overestimate haplotig removal processing in long-read draft genomes, in comparison to a k-mer method. We discuss the trade-offs inherent in assembly algorithm and curation choices and suggest that "best practice" is research question dependent. We demonstrate a link between allelic divergence and allele-derived contig redundancy in final genome assemblies and document the patterns of coding and noncoding diversity between redundant sequences. We also document a link between an excess of nonsynonymous polymorphism and haplotigs that are unresolved by assembly or postassembly algorithms. Finally, we discuss how this phenomenon may have relevance for the usage of noisy long-read genome assemblies in comparative genomics.

Keywords: pool-seq; haplotype; assembly; *Plutella xylostella*

Introduction

Technical and analytical advances in genomics have dramatically improved the achievable standard of genome projects. The amount of high molecular weight (HMW) DNA required to perform long-read sequencing has reduced significantly and has been accompanied by a steady increase in sequencing read lengths and read accuracy (Kingan *et al.* 2019). Longer reads have aided genome assembly efforts by providing information linking unique genomic sequences flanking repetitive elements, which represented a challenge to algorithms reliant on short-read data (Koren *et al.* 2017). However, early long-read methods, such as Oxford nanopore and SMRT sequencing contained higher error rates in raw sequence reads (Derrington *et al.* 2010; Chin *et al.* 2013). Whilst this technical noise can be tolerated by various means (Chin *et al.* 2013, 2016; Koren *et al.* 2017; Kolmogorov *et al.* 2019; Ruan and Li 2020), it is ultimately confounded with the real biological variation present in the underlying samples. The form this biological variation takes and the way it is distributed across the genome of an organism can influence the accuracy of a reconstructed haploid (or phased diploid) genome assembly (Kajitani *et al.* 2019). Various assembly pipelines and algorithms have been explicitly designed to overcome challenges of heterozygosity (Chin *et al.* 2016; Huang *et al.* 2017; Roach *et al.* 2018), repeat resolution (Kolmogorov *et al.* 2019), speed (Ruan and Li 2020), and integration of multiple data types (Ye *et al.* 2016; Zimin *et al.* 2017). Furthermore, all assembly software have some level of parameterization available to optimize results, yielding a huge array of possible outcomes.

Alongside these computational innovations, several experimental approaches and supplemental data types can augment existing data. Trio-sequencing can partition heterozygous variation in an F1 individual using information from parental haplotypes (Koren *et al.* 2018). Linked-reads utilize microfluidics to uniquely barcode reads that derive from discrete large DNA fragments, thereby capturing longer-range information than standard short-read preparations do not (Zheng *et al.* 2016). Chromosome conformation capture (Hi-C) data crosslinks in vivo

Received: May 06, 2022. Accepted: July 25, 2022

chromatin molecules and recovers pairs of reads that derive from these crosslinks, producing data that reflects the 3D organization of the nucleus, and also long-range cis-chromosome associations (Ghurye *et al.* 2019). In addition to the proliferation of supporting data types, the efficacy and quality of core genomic data have improved. Improvements to data quality predominantly come from platform advancements such as high-fidelity (HiFi) long reads (Nurk *et al.* 2020; Cheng *et al.* 2021) and updated nanopore proteins (Karst *et al.* 2021). Meanwhile concurrent developments in library preparation and whole-genome amplification have enabled the use of decreased input DNA amounts (Schneider *et al.* 2021). A recent study found that the best lepidopteran genome assembly available at the time utilized a combination of HiFi data and Hi-C (Ellis *et al.* 2021).

Within genome assembly, accounting for genomic variation is largely a technical consideration. However, this variation is not uniformly or randomly distributed and is shaped by a range of evolutionary and demographic processes. One particularly challenging aspect of genome assembly is the resolution of highly divergent regions (HDRs; Kajitani *et al.* 2019), which often cannot be determined as allelic within the assembly process and requires supervised analysis (Roach *et al.* 2018). Genome assembly projects often aim to pre-emptively avoid this problem by severely inbreeding the source material to increase the proportion of genome homozygosity (The Heliconius Genome Consortium 2012; Nowell *et al.* 2017). However, the previous studies indicate that high levels of heterozygosity are often counter-intuitively maintained despite multiple generations of sib–sib inbreeding (The Heliconius Genome Consortium 2012; Nowell *et al.* 2017). A candidate for such an effect is the presence of overdominant or pseudo-overdominant loci. These loci, by various mechanisms, produce severe fitness consequences in a homozygous state. In the case of pseudo-overdominance, the presence of tightly linked recessive lethal mutations on different alleles prevents either haplotype from becoming homozygous (Charlesworth and Willis 2009). Alternatively, pseudo-overdominance may be produced by multiple linked mildly deleterious alleles, of which the cumulative effect is functionally equivalent to a single recessive lethal. Whatever the fundamental cause, these phenomena can also accumulate linked neutral variation, particularly in recombination cold spots (Zhao and Charlesworth 2016). These features appear to make pseudo-overdominance blocks a plausible candidate for the HDRs known to interfere with genome assembly (Waller 2021).

If HDRs can indicate regions experiencing particular forms of selection, failure to properly resolve them could impact downstream analyses, particularly the detection of balancing selection and overdominant loci. Since this bias is nonrandom, it may also affect comparative genome analyses, for example in instances of balancing selection predating speciation, or other forms of trans-species polymorphism, such as the well-studied MHC locus (Azevedo *et al.* 2015). Similarly, there may be common features of the genetic architecture that may be more likely to produce effects like overdominance or pseudo-overdominance at common ancestral regions. Nonetheless, the increasing quality of long-read sequencing, read lengths, and supporting data should help to mitigate the issue and enable evaluation of the scale of this problem across historic genome datasets. One method that is used widely in evaluating the completedness of genome assemblies is the use of highly conserved gene sequences that are consistently present as single-copy genes (Simão *et al.* 2015). However, in the context of HDRs and their putative sources, it is possible that these methods may be biased against representing genome regions that are more likely to harbor HDRs. One potential resolution to this is to find genome validation methods that do not rely on such cross-species inferences (Rhie *et al.* 2020).

Here, we investigate the complex trade-offs that are made in the choice of genome assembly algorithm using a long-read genomic dataset for the diamondback moth—*Plutella xylostella*, which was the subject of a previous major genome sequencing effort, culminating in the publication of an assembly in 2013 (GCA_000330985.1; You *et al.* 2013). The assembly strategy utilized the sequencing of fosmids, in order to mitigate the short-read lengths of Illumina sequencing. The authors report extensive structural variation based on alignments between their assembly and both the fosmids and a previously sequenced BAC (GenBank accession GU058050). The genome of *P. xylostella* therefore represents 2 distinct challenges to current long-read assembly methods, namely a large proportion of structural variation and a small amount of extractable DNA per individual. Our study includes the additional challenge of sequencing the heterogametic sex, containing the W-chromosome which has been shown to be highly repetitive and intractable to assembly (Traut *et al.* 2013).

Methods

Insect material origin and DNA extraction

Starting material was provided by Oxitec Ltd. (Abingdon, UK) from a lab colony that has been continuously cultured on artificial diet and is derived from the Vero Beach strain (Martins *et al.* 2012). Several lines were inbred in parallel by mating sib–sib pairs each generation. A 7 generation inbred family was selected for genome sequencing. DNA was extracted by phenol-chloroform from a pool of 15 sisters of the final inbred generation and a single male and female (Saccheri and Bruford 1993).

Library construction and sequencing

The pooled DNA was sheared to 7 or 10 kbp. A subset was size selected at 15 kbp on the BluePippin (Sage Science, Inc.). In total, 66 SMRT cells were sequenced with P5-C3 chemistry on the RSII platform (Pacific Biosciences, Inc.). Reads were filtered according to subread length (>50 bp), polymerase read quality (>75 bp), and polymerase read length (>50 bp). Extracted DNA from the individual male and female was sheared and used for individual libraries followed by 2× 100-bp paired-end Illumina sequencing (Illumina, Inc.).

Genome assembly parameters

We performed assembly using canu (version 2.1.1), flye (version 2.8.2-b1689), and wtdbg2 (version 2.5). These assemblies were subsequently polished with the same pacbio read-set for 2 iterations using quiver (version 2.3.3). As a preliminary step, we applied author recommended parameters for producing separated haplotypes in the presence of heterozygosity and subsequently used the resulting assembly with the highest rate of duplicated BUSCO genes (run as described below). For flye and wtdbg, we selected the default parameter result, for canu, we selected the assembly using the parameter set [genomeSize = 340m corOutCoverage = 200 correctedErrorRate = 0.040 "batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50"].

Haplotype merging

We trialed 2 postassembly haplotype merging procedures, purge_dups and Haplomerger 2. Genomes processed with Haplomerger2 were first masked using windowmasker (version 20120730). A species-specific scoring matrix was inferred at 95%

identity using the lastz_D_Wrapper.pl script included with Haplomerger2 (Huang et al. 2017). The masked genome and scoring matrix were then used to run scripts B1–B5 of the Haplomerger2 pipeline (version 20161205; Huang et al. 2017).

Scaffolding

The preparation of HiC libraries was performed by Dovetail LLC using pools of starved larvae. HiC libraries were prepared as described by Kalhor et al. (2011). Both library preparations used the restriction enzyme DpnII for digestion after proximity ligation. Scaffolding and misassembly detection were performed by running the 3D-DNA pipeline on each of the haplotype merged assembly versions.

Validation procedures

Here, we quantify the haplotype resolution processes using 2 independent methods. Firstly, we utilized the gene-based BUSCO score (version 5.0.0) with the "Lepidoptera odb10" database, consisting of 5,286 gene groups and augustus species model "heliconius_melpomene1." Secondly, we utilized a combination of the stacked k-mer coverage histograms (a.k.a. spectra-cn) plots generated with KAT (version 2.4.2) and the read-based k-mer models produced by the genomescope R script. In brief, we used the R function "pmin" to intersect the assembly copy number coverage distributions with the modeled distributions from genomescope, specifically, the error distribution, and the heterozygous and homozygous components of the unique distribution. This provided a quantitative k-mer-based comparison of the haplotype resolution processes.

Haplotype divergence assessment

We quantified the divergence between duplicated genes identified by the BUSCO analysis using an alignment-based and a supporting alignment-free method. The amino-acid sequences of duplicated BUSCO gene copies were aligned with MAFFT (version v6.864b), and subsequently translated into codon-based alignments with pal2nal.pl (version 14), followed by calculation of synonymous and nonsynonymous variants using the biopython function "cal_dn_ds." For the alignment-free comparison, we used the full-genomic sequence (including introns) between the gene start and end coordinates identified by BUSCO and used the python package "alfpy" with a word size of 2 to calculate the Canberra distance (see 28).

Results

Insect materials, sequencing, and assembly

The material described in this study was inbred for 7 generations and is derived from a long-term laboratory culture, itself derived from the "Vero Beach" strain (United States). Fifteen sisters were pooled to meet minimum HMW DNA input requirements. Initial assemblies showed that substantial genetic variation was retained and was of sufficient complexity to produce multiple allelic sequence contigs from the same locus, inflating the total size way beyond the expected size of 338.7 (+/−1.1) Mbp, reported by Baxter et al. (2011). We subsequently trialed 2 approaches to resolve the redundant heterozygosity Haplomerger2 and purge_dups (Huang et al. 2017; Guan et al. 2020). After filtering 2,655,788 PacBio subreads remained (mean subread length = 7,301 bp, N50 = 10,398 bp, total bases 19.4 Gbp).

Heterozygosity assessment

All initial assembly strategies resulted in an over-inflated genome size, suggesting differing amounts of redundant haplotig sequence (Fig. 1a). We determined BUSCO results for the initial assemblies as evidence for the levels of allelic redundancy (measured as duplicated BUSCO genes) and overall completeness (Fig. 1, b and c). We also utilized k-mer-based methods. Firstly, a histogram of corrected PacBio reads, provided an initial estimation of genome heterozygosity as approximately 1.11%, which is moderately, but not exceedingly high for a North American sample [see You et al. (2020) for context] (Fig. 2a). Estimates from related individuals (nonpooled) were 0.54% for a related inbred male and 1.00% for a related inbred female (Supplementary Fig. 3). Stacked histograms colored by assembly coverage provided a qualitative assessment of genome assembly completeness and redundant allelic variation (Fig. 2b). Secondly, we intersected the modeled distributions of homozygous, heterozygous, and sequencing error content from genomescope with stacked histograms in order to make quantitative comparisons of the initial assemblies (methodology illustrated in Fig. 1a, data shown in Fig. 2c; Vurture et al. 2017).

The WTDBG2 assembly was the smallest in size (427 Mbp) and contig number (4,023) (Fig. 1a). It had the lowest number of duplicated BUSCO genes (805) and highest number of missing genes (110) (Fig. 1b). Consistent with the BUSCO results, WTDGB2 had the lowest number of homozygous k-mers duplicated in the assembly (Fig. 2b). But it also had the highest number of modeled error k-mers present (Fig. 2c). In contrast, the Flye assembly was the largest in size (494 Mbp) and contig number (5,985). It had the most duplicated BUSCO genes (1,324) and the least missing genes (88). Again the k-mer results show concordance with the highest number of homozygous k-mers present duplicated in the assembly. However, the number of modeled error k-mers present in the assembly was comparable with the canu assembly. Canu produced intermediate values in total size (448 Mbp) and contig number (5,341). Similarly, BUSCO results indicated an intermediate number of duplicated genes (1,105) and missing genes (106). k-mer results followed the same pattern except for error k-mers in the final assembly.

Both postassembly allelic redundancy approaches reduced the overall sizes of the assemblies and appear to follow the patterns observed in the initial assemblies, such that WTDBG2 still retains the lowest number of duplicated and highest number of missing genes in contrast with Flye. For each of the 3 starting assemblies, Haplomerger2 provided a greater reduction in total size and number of contigs compared to purge_dups (Fig. 1a. When applied to the canu assembly we also observe an increase in contiguity (Fig. 1a), due to a tiling effect produced when corresponding redundant heterozygous regions are merged at the ends of contigs (Supplementary Fig. 1).

Postassembly processing resolved most duplicated BUSCO genes to a single copy regardless of the initial assembly algorithm, however, in all cases, the number of missing BUSCO genes also increased (Fig. 1, b and c). We observe that purge_dups resolved some duplications that are not resolved by Haplomerger2 and vice versa (Fig. 1c). Similarly, genes that go from complete and single copy in the primary assembly to fragmented or missing after postprocessing are not necessarily the same across the 2 methods (Fig. 1c). This suggests that removal of redundancy is not simply, more or less "aggressive," and that performance varies by algorithm depending on specific sequence properties.

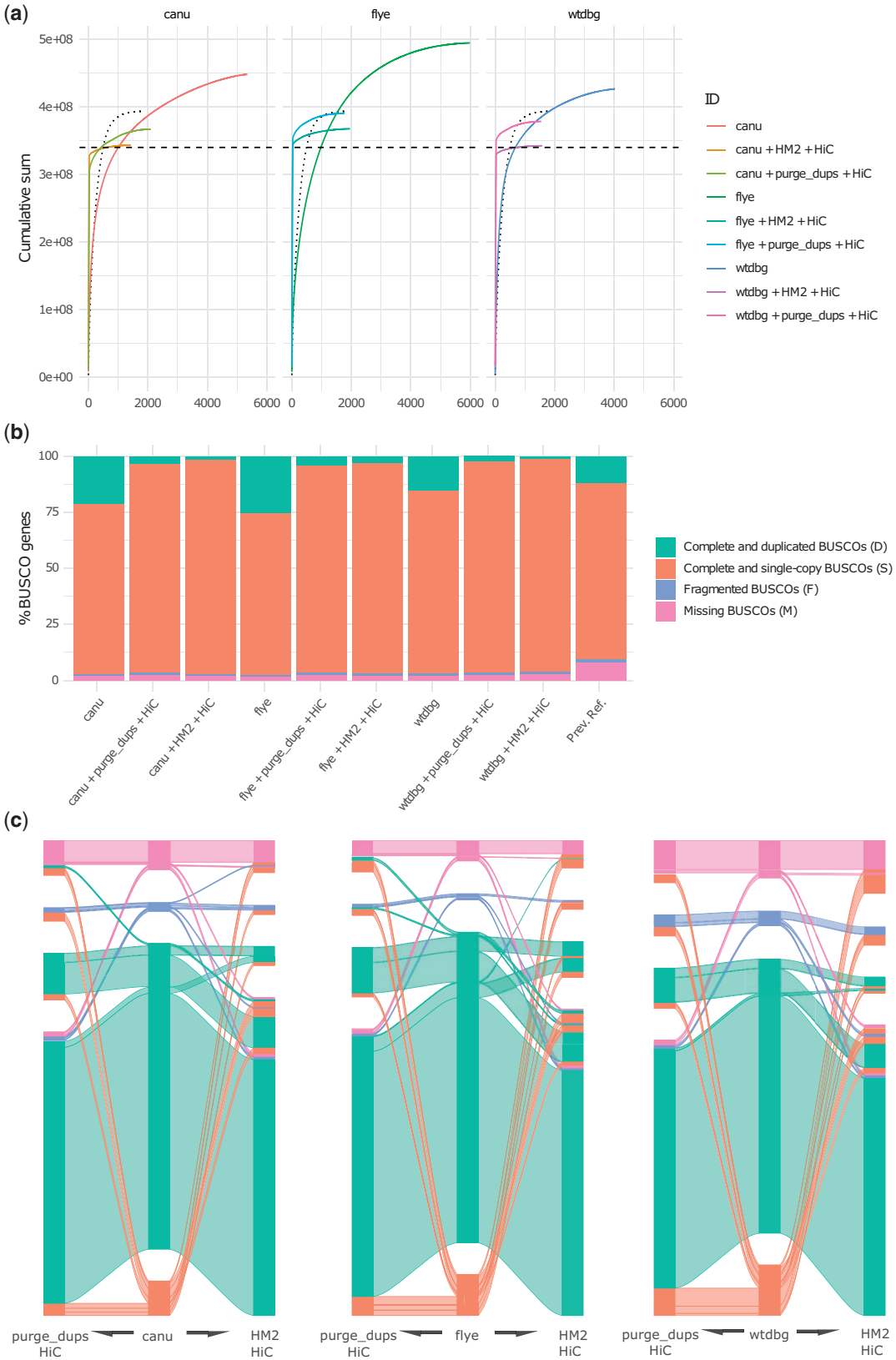

Fig. 1. Contiguity and BUSCO content and of alternative genome assembly methods and the effects of removing putative allelic redundancy. In each panel, "canu," "flye," and "wtdbg" refer to the preliminary assemblies produced by each algorithm. "+ purge_dups + HiC" refers to these same assemblies with the additional application of the purge_dups program followed by HiC scaffolding or, Haplomerger2 followed by HiC scaffolding (a) depicts the differences in overall contig size and contiguity between the different methods. The dotted curve describes a previously published reference genome (accession: GCA_000330985.1). The dashed straight line indicates the estimated genome size from an independent flow cytometry estimate (Baxter *et al.* 2011). (b) Overall BUSCO scores from a database of 5,286 genes. BUSCO scores from the aforementioned accession are also included. (C) This image details the relationships of genes within these sets. Groups of genes are colored by BUSCO score in the initial assembly. BUSCO genes that are single copy and complete in all assemblies are omitted to emphasize differences between assemblies.

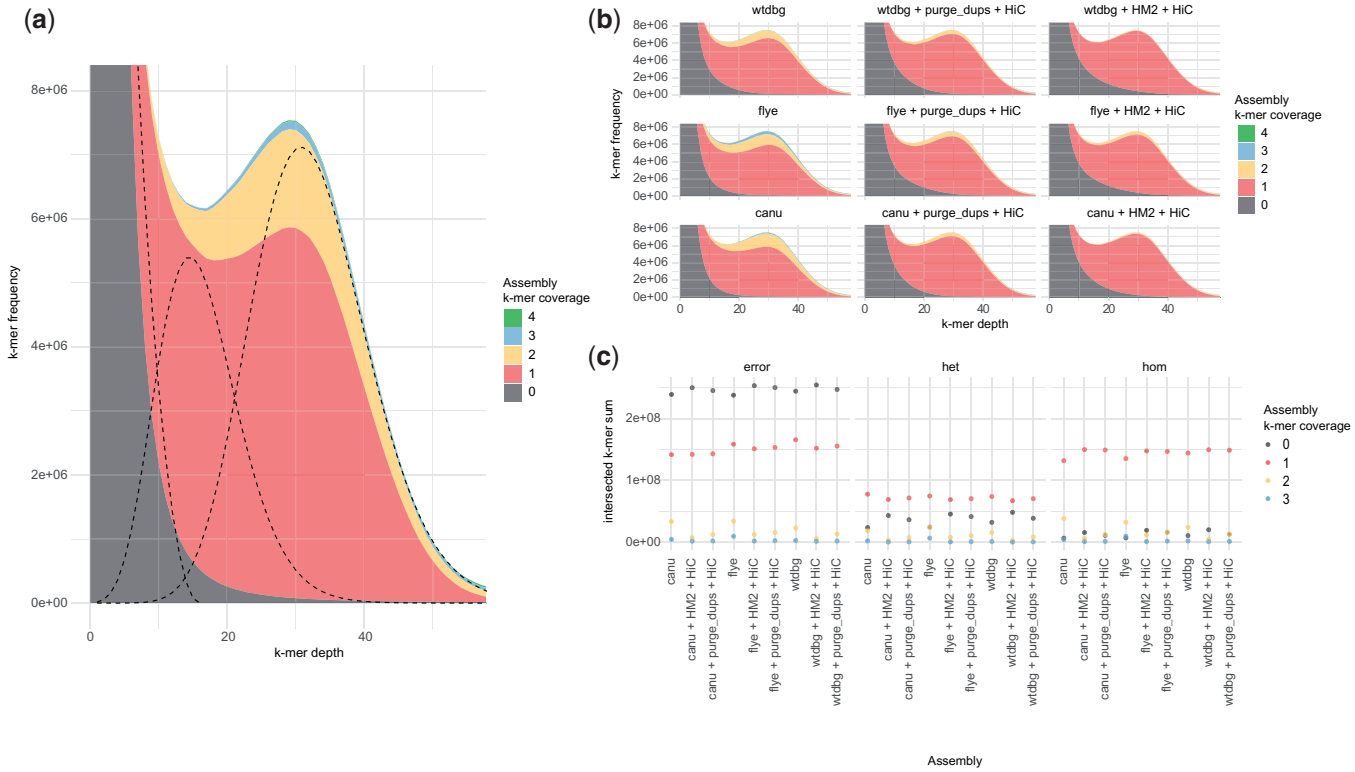

Fig. 2. A k-mer-based validation of the alternative genome assembly methods and effects of removing putative allelic redundancy. a) An example of stacked k-mer distributions subdivided by assembly representation (spectra-cn plot) and an overlay of the modeled contributions of sequencing errors, heterozygous content and homozygous content (dotted lines from left to right, respectively). b) The spectra-cn plots for each of the assembly versions (c) shows the number of k-mers present in the intersections between the modeled k-mer content distributions and individual assembly coverage categories present in the spectra-cn plots.

Comparison of heterozygosity assessments

Using the k-mer intersection approach described in Fig. 2, we produce a k-mer proxy of BUSCO genes for comparison. The proxy is calculated using the modeled homozygous k-mers (analogous to single-copy genes) and divides the occurrence of duplicated assembly k-mers by the sum of the single copy and duplicated assembly k-mers (Supplementary Table 2). Levels of percent duplication in the initial assemblies are remarkably concordant between the genic (BUSCO) and unbiased (k-mer) methods (Supplementary Table 2), with the exception of flye. This exception is likely due to the relatively higher occurrence of 3-copy redundancy observed with the flye assembly algorithm (Fig. 2b), which are not captured in our k-mer proxy measurement (BUSCO duplications do not distinguish 2 copy genes from >2 copy genes). However, after assembly postprocessing to remove redundant haplotigs, BUSCO genes appear to overestimate the efficiency of the procedures in comparison to the k-mer proxy. Across all methods and initial starting assemblies, the k-mer proxy shows consistently higher residual duplication than suggested by BUSCO genes (Supplementary Table 2).

HiC misassembly detection and scaffolding

We observed the greatest overall number of detected misassembled region candidates in "canu + purge_dups" after 2 iterations of the HiC scaffolding pipeline 3D-DNA. The least misassembled region candidates detected after 2 iterations were in 'wtdbg + HM2', followed by 'canu + HM2' (Table 1). We find the greatest disparity in total misassembled region candidates between postprocessing methods in the canu assemblies. Furthermore, we find that "canu + purge_dups" produced the lowest final N50 value, whereas all other assemblies produced very similar results, though the metric is limited by karyotype at this resolution (Table 1).

Patterns of divergence between redundant alleles

For BUSCO genes that were duplicated in the initial assembly and subsequently reduced to single copy by the postprocessing methods, we broadly describe the variation between the copies using the ratio of nonsynonymous and synonymous nucleotide diversity and an alignment-free method using the entire genomic region (Zielezinski *et al.* 2019). We observed that the distributions of genes remaining after the application of purge_dups were more heavily weighted toward a low k-mer-based distance and a $\pi N/\pi S$ ratio of 0 as compared to the distributions of genes deduplicated by Haplomerger2 (Fig. 3). We partitioned duplicated BUSCOs depending on whether they are present on the same initial assembly contig or not, as these genes may plausibly be real duplication events rather than haplotypic redundancy. The distribution of $\pi N/\pi S$ between putative tandem duplicated BUSCOs (occurring on the same assembly contig) appears to be somewhat inflated in comparison to putative haplotig BUSCO genes. This pattern is also reflected in the overall genomic DNA k-mer distances (Fig. 3).

Discussion

Plutella xylostella populations harbor large amounts of polymorphism (You *et al.* 2013, 2020). We observed a relatively low heterozygosity in our pooled data compared to other species; however, this individual should not be considered representative of the

Table 1. Number of misassembly region candidates detected within the 3D-DNA pipeline.

	Total size (initial size) Mbp	Pre-HiC N50 (Mbp)	Post-HiC N50 (Mbp)	Narrow misassemblies		Wide misassemblies		Total iteration 2 misassemblies	Total disparity
				Iteration 1	Iteration 2	Iteration 1	Iteration 2		
canu + HM2	343 (448)	1.14	11.24	719	806	134	107	913	619
canu + purge_dups	367 (448)	0.62	9.49	794	1,247	164	285	1,532	
flye + HM2	368 (494)	0.41	11.42	1,065	1,225	185	115	1,340	55
flye + purge_dups	391 (494)	0.32	11.44	835	1,187	225	208	1,395	
wtdbg + HM2	343 (427)	1.83	11.1	673	721	127	92	813	382
wtdbg + purge_dups	378 (427)	0.94	11.49	827	993	204	202	1,195	

We used the default resolution parameters "wide_res = 25,000 bp" and "narrow_res = 1,000 bp."

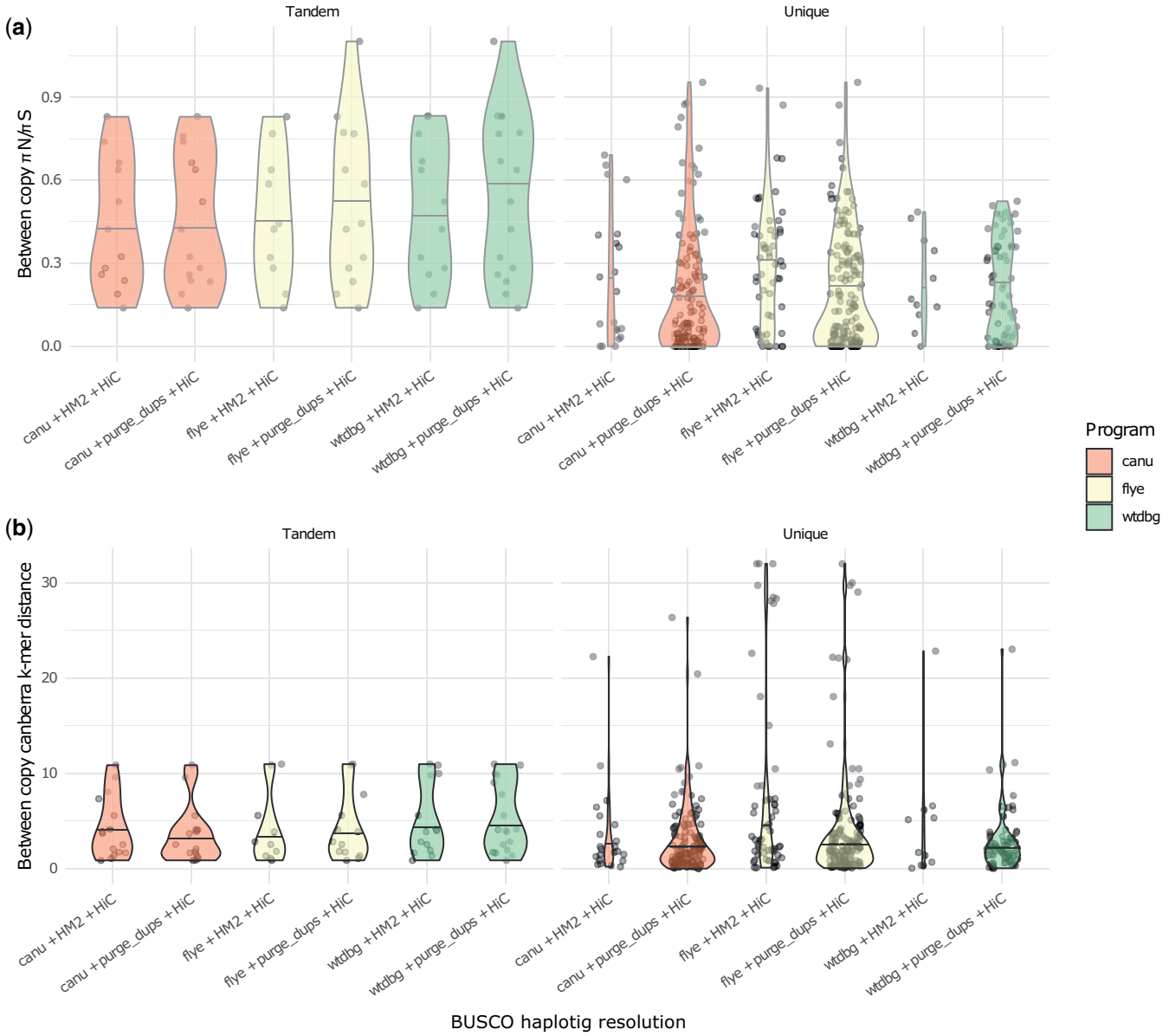

Fig. 3. Quantifying divergence between duplicated BUSCO genes. a) shows the distribution of πN/πS scores for duplicated (N. copies = 2) BUSCO genes remaining after the application of purge_dups or Haplomerger2. b) This image shows an alignment-free quantification of the dissimilarity of intronic and exonic sequence between the same duplicated BUSCO genes (see *Methods* for details). Panels labeled "Tandem" indicate that the BUSCO copies were found on the same assembly contig, whereas "Unique" indicates that the copies were found on different assembly contigs.

wild population due to severe inbreeding and prior laboratory domestication. Despite this apparently reduced heterozygosity, a large amount of redundant sequence remains after genome assembly, suggesting that the heterozygosity is largely colocalized in highly divergent alleles. This pattern may suggest regions of low recombination, enabling haplotypes to accumulate linked

neutral variation and persist through drift. Alternatively, in the case of associative overdominance, neutral variation can accumulate alongside linked overdominant or pseudo-overdominant loci (linked deleterious recessives with opposing phase; Ohta and Kimura 1970). It is important that such regions are represented appropriately in genome assemblies, as downstream analyses involving mapping reads rely on both overall completeness and regions being present in a haploid state, although see (Armstrong et al. 2020) for how this is changing.

We tested 2 postassembly redundancy reduction procedures (Haplomerger2 and purge_dups) and found that Haplomerger2 generally appears to "resolve" more redundant sequence, at the expense of erroneous removal of nonredundant genome content and erroneous scaffolding of overlapping divergent regions. Both programs utilize a self-alignment step to detect haplotigs, purge_dups then implements a further QC step to these results by assessing the coverage of the identified haplotigs. For self-alignment, Haplomerger2 utilizes LASTZ and enables users to calculate and use a sample-specific scoring matrix, whilst purge_dups utilizes minimap2 with a fixed intraspecies scoring parameter (asm5). The parameterization reflects a balance in differentiating intralocus divergence, from interlocus paralog similarity. To give specific examples; ancient balancing selection vs relatively recent gene duplication or ancient balancing selection vs genetic convergence. The idealized genome assembly or redundancy removal pipeline can accurately differentiate these effects.

Genic analyses of assembly completeness such as BUSCO are widely used and relatively straightforward to apply, however, by definition they are limited to genomic regions containing coding sequences (Simão et al. 2015). The genes are highly conserved at the amino-acid level, suggesting that nonsynonymous substitutions are largely deleterious. Because of this, the surrounding genomic region (including noncoding variation) may be likely to harbor less variation than a neutral region, due to the action of background selection (Gilbert et al. 2020). In short, BUSCO genes are likely to inhabit (and help maintain) conserved genomic regions. The practical implication is that BUSCO genes, when utilized to assess the removal of redundant haplotigs, may systematically overestimate the effectiveness of the procedure, as they are unlikely to represent HDRs. Indeed, our results support the notion that before and after BUSCO duplication scores overestimate the removal of redundant haplotig sequences when compared to an analogous k-mer estimator. If BUSCO duplication results are liable to overestimate the haploid nature of a given draft genome assembly, it may hamper comparative genomic efforts to identify balancing selection or overdominance (which may have either common or independent origins).

Despite this potential limitation, BUSCO scores are still useful as a guide to assembly completeness. BUSCO scores also provided insights into assembly postprocessing, showing that, despite resolving more duplications than purge_dups, Haplomerger2 results do not completely overlap those of purge_dups. This indicates that both underlying methods are suboptimal and the results may be complementary. We also note that BUSCO results, particularly missing genes, are dependent on the optimization of input parameters. For example, the "–long" parameter can increase sensitivity at the cost of greater runtime. Similarly, the detection of BUSCO genes may differ between haplotypes, thereby underestimating the number of duplicated genes.

We demonstrate a supporting validation method, providing relative quantification of assembly accuracy, using overlaps between modeled k-mer distributions and the k-mer frequency histogram subdivided by numerical representation in the final assembly version. Whilst this assessment applies to any nonrepetitive genome region it only offers a general comparative measure between assemblies from the same read set and cannot determine appropriate representation at specific sites, due to stochasticity in read coverage. However, the ability to confidently determine truly heterozygous k-mers from homozygous will be increased in low-error, high-coverage read datasets, such as those currently being generated by projects like DToL (Darwin Tree of Life). This would offer an independent and unbiased validation method, but only with sufficiently high coverage to accurately partition the different k-mer peaks.

Our initial expectation was that a greater reduction in redundancy may correspond with an increase in detected misassembled region candidates, particularly in the case of Haplomerger2, which can join overlapping contigs (Supplementary Fig. 1). Instead, we find the opposite pattern, though this does not necessarily imply more accurate assembly representation, since complex regions may be absent from a final assembly altogether. For example, the lowest number of misassemblies ("wtdbg + HM2"), occurred alongside both the lowest putative allelic redundancy, but also the greatest values for missing BUSCO genes and het/hom modeled k-mers with 0× assembly coverage.

After an appraisal of the results of haplotig resolution, we compared the overall divergence of duplicated genes from the 2 methods. The results of purge_dups retained a greater proportion of low divergence haplotigs and this also corresponded to genes with a lower proportion of nonsynonymous substitutions. The remaining genes in both sets suggest that both methods did not resolve more greatly diverged sequences and that this divergence corresponded to elevated nonsynonymous substitutions relative to synonymous substitutions. Taken together with the high levels of coding sequence conservation intrinsic to BUSCO genes, this pattern would appear consistent with pseudo-overdominant regions generated by the linked arrangement of multiple deleterious nonsynonymous substitutions. However, additional investigations with supplementary data will be required to establish this with confidence and determine the processes responsible for these patterns.

Conclusions

Highly divergent alleles can pose a challenge to accurate haploid reconstruction from noisy long-read data. Postprocessing can mitigate these problems somewhat to produce mosaic resolved sequences for reference purposes, however, results in our case are largely imperfect and present a set of complex trade-offs between assembly completeness, redundancy, and misassembly. Researchers producing or using genomes should be aware of these issues when using genome assembly data derived from noisy long reads, especially when investigating genomic regions likely to harbor significant linked variation. Our results lead us to the conclusion that unresolved HDRs may be widespread in draft genomes assembled from noisy long-read data and that BUSCO analyses may overestimate their resolution by postprocessing methods. Plausible causes include loci experiencing balancing-selection or overdominance effects that originated prior to speciation events, or that exist within a genetic architecture liable to parallel origins of these processes. This may impact comparative genomic studies that aim to identify or describe these evolutionary processes, however, further investigation is required to examine this.

Finally, as a recommendation to researchers utilizing similar data, we suggest that the optimal strategy is research question

dependent. Our data show that there are complex trade-offs between gene set completeness, the presence and abundance of redundant haplotigs, and overall genome contiguity. We list some recommendations resulting from our dataset: (1) For comparative analyses of large-scale shared-synteny or chromosome-level structures, we would recommend wtdbg2 followed by Haplomerger2, however, it should be stressed that for this type of analysis researchers should supplement long-read data with HiC, both to extend contigs into larger-scale scaffolds, and also to correct any erroneous assembly, or postprocessing misjoins. When HiC data are available, the contiguity differences between different assemblers become less important, however, wtdbg2 followed by Hapolomerger2 should still reduce misleading interspecific alignment signals produced by residual haplotigs. (2) For comparative analysis of orthologs, the decision is complicated, as there is a trade-off between false-positive paralogs due to redundant haplotigs, vs false-negative missing genome content that is eliminated by redundancy removal procedures. (3) For analysis of a particular gene of interest, researchers can assemble their data with flye or canu, and process the resulting assembly with purge_dups. Reference to both the initial assembly and the postprocessed data should enable researchers to recover their gene of interest and examine whether any allelic variation is present. (4) For researchers wishing to produce a multipurpose reference assembly, with no specific research question, we would suggest producing detailed and transparent methods, such that subsequent users can understand the limitations and reanalyze for their specific purpose if necessary.

Data availability

The read datasets generated during this study are available in the ENA database under accession PRJEB34571 (see Supplementary Table 1). All assembly versions are available at 10.5281/zenodo.5647466. Code provided at https://github.com/swomics/Plutella_genomes

Supplemental material is available at G3 online.

Acknowledgments

PacBio and Illumina sequencing were performed at the Centre for Genomics Research, University of Liverpool by Margaret Hughes and Lucille Rainbow. We thank Nerys Humphrey-Jones and Adam S. Walker for their work maintaining insect strains. We also thank Matthew R. Gemmell for providing a template script that was modified for the quiver polishing procedure.

Funding

NIM and ACD received funding from the BBSRC and Innovate UK grants BB/M001512/1 and BB/M503472/1.

Conflicts of interest

IW's studentship was cofunded by Oxitec. TM and NIM are employees of Oxitec. RK is an employee of IBM U.K. StW is an employee of General Bioinformatics. MG is an employee of Syngenta. ACD has held multiple grants with Oxitec.

Author contributions

IW, NIM, and ACD designed and planned the project. IW performed analyses with ACD. AH, SH, IS, and MG prepared material for sequencing. The manuscript was written by IW, IJS, and ACD with input from all authors. StWi and MG provided additional data. RK performed preliminary analyses on the data.

Literature cited

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587(7833):246–251. doi:10.1038/s41586-020–2871-y.

Azevedo L, Serrano C, Amorim A, Cooper DN. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. Hum Genomics. 2015;9(1):4–9. doi:10.1186/s40246-015–0043-1.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS One. 2011;6(4):e19315. doi:10.1371/journal.pone.0019315.

Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. 2009;10(11):783–796. doi:10.1038/nrg2664.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–175. doi:10.1038/s41592-020–01056-5.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–569. doi:10.1038/nmeth.2474.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–1057.

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH. Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A. 2010;107(37):16060–16065. doi:10.1073/pnas.1001831107.

Ellis EA, Storer CG, Kawahara AY. De novo genome assemblies of butterflies. GigaScience. 2021;10(6):1–8. doi:10.1093/gigascience/giab041.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):e1007273.

Gilbert KJ, Pouyet F, Excoffier L, Peischl S. Transition from background selection to associative overdominance promotes diversity in regions of low recombination. Curr Biol. 2020;30(1):101–107.e3. doi:10.1016/j.cub.2019.11.063.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–2898. doi:10.1093/bioinformatics/btaa025.

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B, *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21(1):35. doi:10.1101/654566.

Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly.

Bioinformatics. 2017;33(16):2577–2579. doi:10.1093/bioinformatics/btx220.

Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, Itoh T, *et al.* Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. Nat Commun. 2019;10(1):1702. doi:10.1038/s41467-019–09575-2.

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2011;30(1):90–98. doi:10.1038/nbt.2057.

Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods. 2021;18(2):165–169. doi:10.1038/s41592-020–01041-y.

Kingan S, Heaton H, Cudini J, Lambert C, Baybayan P, Galvin B, Durbin R, Korlach J, Lawniczak M. A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. Genes. 2019;10(1):62. doi:10.3390/genes10010062.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–546. doi:10.1038/s41587-019–0072-8.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM, *et al.* De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;36(12):1174–1182.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–736. doi:10.1101/gr.215087.116.Freely.

Martins S, Naish N, Walker AS, Morrison NI, Scaife S, Fu G, Dafa'alla T, Alphey L. Germline transformation of the diamondback moth, *Plutella xylostella* L., using the piggyBac transposable element. Insect Mol Biol. 2012;21(4):414–421.

Nowell RW, Elsworth B, Oostra V, Zwaan BJ, Wheat CW, Saastamoinen M, Saccheri IJ, Van't Hof AE, Wasik BR, Connahs H *et al.* A high-coverage draft genome of the mycalesine butterfly *Bicyclus anynana*. Gigascience. 2017;6:1–7.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S, *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30(9):1291–1305. doi:10.1101/GR.263566.120.

Ohta T, Kimura M. Development of associative overdominance through linkage disequilibrium in finite populations. Genet Res. 1970;16(2):165–177. doi:10.1017/S0016672300002391.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):1–27. doi:10.1186/s13059-020–02134-9.

Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–158. doi:10.1038/s41592-019–0669-3.

Saccheri IJ, Bruford MW. DNA fingerprinting in a butterfly, *Bicyclus anynana* (Satyridae). J Heredity. 1993;84(3):195–200. doi:10.1093/oxfordjournals.jhered.a111316.

Schneider C, Woehle C, Greve C, D'Haese CA, Wolf M, Hiller M, Janke A, Bálint M, Huettel B. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). Gigascience. 2021;10(5):1–12. doi:10.1093/gigascience/giab035.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012;487(7405):94–98. doi:10.1038/nature11041.

Traut W, Vogel H, Glöckner G, Hartmann E, Heckel DG. High-throughput sequencing of a single chromosome: a moth W chromosome. Chromosome Res. 2013;21(5):491–505. doi:10.1007/s10577-013–9376-6.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202–2204. doi:10.1093/bioinformatics/btx153.

Waller DM. Addressing Darwin's dilemma: can pseudo-overdominance explain persistent inbreeding depression and load? Evolution. 2021;75(4):779–793. doi:10.1111/evo.14189.

Ye C, Hill CM, Wu S, Ruan J, Ma Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep. 2016;6(1):1–9. doi:10.1038/srep31900.

You M, Ke F, You S, Wu Z, Liu Q, He W, Baxter SW, Yuchi Z, Vasseur L, Gurr GM, *et al.* Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. Nat Commun. 2020;11(1):2321. doi:10.1038/s41467-020–16178-9.

You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, *et al.* A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet. 2013;45(2):220–225. doi:10.1038/ng.2524.

Zhao L, Charlesworth B. Resolving the conflict between associative overdominance and background selection. Genetics. 2016;203(3):1315–1334. doi:10.1534/genetics.116.188912.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34(3):303–311. doi:10.1038/nbt.3432.

Zielezinski A, Girgis HZ, Bernard G, Leimeister C-A, Tang K, Dencker T, Lau AK, Röhling S, Choi JJ, Waterman MS, *et al.* Benchmarking of alignment-free sequence comparison methods. Genome Biol. 2019;20(1):144. doi:10.1101/611137.

Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017;27(5):787–786. doi:10.1101/gr.213405.116.27.

Communicating editor: P. Ingvarsson