


Evaluating long-read assemblers to assemble several aphididae genomes

Nicolaas F. V. Burger , Vittorio F. Nicolis, Anna-Maria Botha*

Van der Byl Street, Genetics Department, JC Smuts Building, Faculty of AgriScience, Stellenbosch University, Stellenbosch, South Africa

*Corresponding author: Van der Byl Street, Genetics Department, JC Smuts Building, Faculty of AgriScience, Stellenbosch University, Stellenbosch, South Africa.
E-mail: ambo@sun.ac.za

Biographical note: The authors of this manuscript are active in research related to biotic and abiotic stressors of wheat, mainly focusing on plant–host interactions with the aphid *Diuraphis noxia* and how it subverts host resistance by utilizing genomic tools and techniques.

Abstract

Aphids are a speciose family of the Hemiptera compromising >5500 species. They have adapted to feed off multiple plant species and occur on every continent on Earth. Although economically devastating, very few aphid genomes have been sequenced and assembled, and those that have suffer low contiguity due to repeat-rich and AT-rich genomes. With third-generation sequencing becoming more affordable and approaching quality levels to that of second-generation sequencing, the ability to produce more contiguous aphid genome assemblies is becoming a reality. With a growing list of long-read assemblers becoming available, the choice of which assembly tool to use becomes more complicated. In this study, six recently released long-read assemblers (Canu, Flye, Hifiasm, Mecat2, Raven, and Wtdbg2) were evaluated on several quality and contiguity metrics after assembling four populations (or biotypes) of the same species (Russian wheat aphid, *Diuraphis noxia*) and two unrelated aphid species that have publicly available long-read sequences. All assemblers did not fare equally well between the different read sets, but, overall, the Hifiasm and Canu assemblers performed the best. Merging of the best assemblies for each read set was also performed using quickmerge, where, in some cases, it resulted in superior assemblies and, in others, introduced more errors. *Ab initio* gene calling between assemblies of the same read set also showed surprisingly less similarity than expected. Overall, the quality control pipeline followed during the assembly resulted in chromosome-level assemblies with minimal structural or quality artefacts.

Keywords: long-read assembly; multiple assembly comparison; aphid genomes; genome quality assessment; assembly and QC pipeline

Introduction

Aphids account for roughly 5550 species of the documented 104 000 Hemipterans [1], with at least 250 species that are considered economically significant pests [2]. Despite this, the extensive diversity of these pests is not represented in genomic sequence databases with only 79 Aphididae genome assemblies (from 36 species) available on the National Center for Biotechnology Information (NCBI; date accessed: 19 September 2024).

The lack of tools optimized to analyse the highly repetitive [3] and extremely AT-rich genomes [4] makes the analysis of aphid genomes challenging. Short reads derived from repetitive regions can align to multiple locations in the genome, leading to ambiguities that may result in these reads being classified as artefacts and consequently excluded from further analyses [5]. Assembly algorithms often struggle with accurately resolving repetitive sequences, potentially collapsing them into fewer copies than are present in the genome [6]. This can result in the inaccurate representation or underrepresentation of these regions in the final assembled genome. Consequently, the aphid genome assembled from short-read sequencing data often exhibits low contiguity due to these complexities.

With the advent of long-read DNA sequencing platforms, the contiguity of *de novo* genome assemblies has been greatly

improved as they overcome the shortcomings of next-generation DNA sequencing (NGS), including less sequence-dependent bias, more homogeneous genome coverage, and less information loss [7, 8]. As it only uses the genomic information contained within the sequenced reads, high-quality *de novo* assemblies are crucial for the study of nonmodel organisms, where they facilitate the discovery of overlooked genomic features [9]. Despite these improvements, published aphid genomes assembled with long-read sequencing data exhibit varying degrees of contiguity. For example, *Sipha maydis* has a contig number of 3570 with an N50 of 187 kb [10] and *Schlechtendalia chinensis* 378 contigs with a N50 of 4.33 Mb [11], while *Rhopalosiphum nymphaeae* has the highest reported contiguity, with only 91 contigs and an N50 of 12.7 Mb [12]. These aphid genome assemblies, as many other, mainly utilize only one assembly tool with limited associated assembly quality assessment prior to genome annotation.

Current third-generation sequencing (TGS) technologies have become a widely adopted tool in genomic studies with their lowered error rates matching that of short-read sequencing [13], improved precision across various parameters (single-nucleotide variant and insertion/deletion calling), and robust performance in difficult-to-map regions [14]. Because of these advancements, numerous genome assemblers have been developed to harness

Received: December 19, 2024. Revised: January 29, 2025. Accepted: February 25, 2025

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

the advantages of low-error long-read sequencing. However, selecting the appropriate assembly tool for achieving high-quality assembly remains a challenging and ambiguous task.

Despite the benchmarking of state-of-the-art long-read *de novo* assemblers for eukaryotic genomes [9], there has been a lack of comprehensive analysis and evaluation of genome assemblers based on TGS datasets in the field of hemipteran pests. Therefore, we attempted to assess the impact of different assemblers on the construction and evaluation of chromosome-level genome assemblies for aphid pests and how the assemblies of the various aligners compare. In this study, the outcome in terms of genome contiguity after genome assembly with six TGS assembly programmes was compared. For this, we used the data sets of four different *Diuraphis noxia* populations (i.e. biotypes/isolines) and two other aphid species, namely, *Aphis gossypii* and *Sitobion avenae*. The pipeline used in this study attempted to not only compare the assembly tools based on (i) the contiguity and quality of the resultant assemblies but also (ii) to generate the most contiguous and high-quality draft genomes for the *D. noxia* read sets.

Materials and methods

Aphid DNA extraction and sequencing

To obtain high-quality DNA for TGS sequencing, *D. noxia* populations RWA-SA1, RWA-SA5, RWA-SAM, and RWA-SAM2 were reared and DNA-extracted as previously described [15], where after library preparation was separately performed for each of aphid populations using the SMRTbell Express Template Prep Kit 2.0 kit and sequenced on the PacBio SII system using the CSS HiFi protocol with three passes (SRA accessions SRX21829916, SRX21829917, SRX21829918, SRX21829919). SRA-deposited PacBio CSS HiFi read sets were also obtained for *A. gossypii* (SRX12408027) and *S. avenae* (SRX10897632) as a comparison. Preceding further analyses, all read sets first underwent quality assessment through FASTQC v0.11.8 [16] and Trimmomatic v0.39 [17] to filter out low-quality (Q30) and short-length reads (<2000 bp), while HiFi-AdapterFilt v3.0.1 [18] was used to remove reads that contained PacBio adapter sequences.

Genome assembly and quality assessment

Prior to genome assembly, GenomeScope 2.0 v1.0 [19] and jellyfish v2.2.8 [20] were used to estimate genome size and level of heterozygosity for each of the six PacBio read sets (Supplementary Table S1), where after the genomes were assembled using the pipeline as illustrated in Fig. 1.

Assembly of each read set was performed using the default parameters of the different assemblers (Canu v2.2, Flye v2.9-b1774, Hifiasm v0.16.1-r375, MECAT2 v20190314, Raven v1.8.1, and Wtdbg2 v2.5 [21–26]). The Canu and Flye assemblies were manually filtered by removing contigs that were identified by the assemblers as circular, repetitive, or bubble elements. Where a genome size needed to be provided, the larger estimate from either GenomeScope 2.0 [19] or published genome sizes were used. The two assemblies that presented with the best assembly metrics (as seen in Table 1) were then merged, making use of the program quickmerge [27], with the most contiguous assembly used as query (or base) and the second most contiguous as reference. After genome assembly, haplotigs were removed using purge_dups [28], whereafter three misassembly correction programs, namely, Inspector v1.0.1 [29], LongStitch v1.0.5 [30], and CRAQ (Clipping Reveals Assembly Quality) v1.0.9-alpha [31], were assessed for their ability to improve overall genome quality. Finally, for the *D. noxia* read sets, RagTag [32] was used to scaffold

the assemblies between the aphid populations without patching. To establish assembly accuracy and completeness, k-mer counting was conducted using Merqury v1.3 [33] and Quast v5.2.0 [34]. For validation of the assembly accuracy and completeness, the Merqury assembly consensus quality (QV) and error rate were calculated within Merqury by dividing the number of solid k-mers found in the assembly with the number of solid k-mers found in the read set. In the case of the *D. noxia* genomes, all the genomes were compared in a reference-free comparison, whereas the assemblies of *A. gossypii* and *S. avenae* were compared to their respective NCBI references GCF_020184175.1 and GCA_019425605.1).

To map the read sets onto the various assemblies, we used Minimap2 v2.24-r1150-dirty [35], while Nanoplot v1.40.2 [36] was applied to plot multiple read metrics for a comparison between the different assemblies, and read statistics were obtained from Quast and Mercury. To generate dot plots for visualizing sequence concordance between the different assemblies, we applied Dgenies v1.5.0 [37] and minimap2, while dot plots were drawn, which were limited to matches with a minimum of 50% identity with no small match filtering and sorted according to best matching contigs from the query to the target.

We applied Augustus [38] to annotate protein-coding genes in all the assemblies making use of the pea aphid training set, of which the proteins were used in a two-way DIAMOND BLASTp (Basic Local Alignment Search Tool) v2.1.7.161 [39] analysis to compare gene set similarity between the assemblies. Sequences that were unique between any of the assemblies were then compared to the NCBI's nr database (date downloaded: May 2024) using DIAMOND BLASTp. Finally, BUSCO (Benchmarking Universal Single-Copy Orthologue) v5.4.3 [40] analysis was performed to assess contiguity of the orthologous gene set using the hemiptera_odb10.2024-01-08 BUSCO dataset.

Results and discussion

Sequencing and assembly metric comparisons

Sequencing of the four *D. noxia* populations (Biotypes RWA-SA1, RWA-SA5, RWA-SAM, and RWA-SAM2) resulted in read sets of various lengths and depths, with phred scores of Q30 for at least 97% of all bases (Supplementary Table S1) with RWA-SA5 having the lowest mean sequence length at ~8300 bp and RWA-SAM2 the longest at ~17 700 bp. For the NCBI-downloaded read sets, both had phred scores of Q30 for at least 97% of bases, with *A. gossypii* having a mean read length of ~8300 bp and *S. avenae* having a mean read length of ~11 666 bp. The number of reads containing internal adapters were <0.5% of the total for all read sets and were removed before further analysis.

Analysing the read sets with GenomeScope 2.0 suggested a genome length range of 374.8–399.4 Mb for the *D. noxia* read sets, 346.3 Mb for *A. gossypii* and 433.9 Mb for *S. avenae*, which are comparable to the published genome sizes estimates for *D. noxia* (1C = 0.428 pg or 418.095 Mb, [41]), *A. gossypii* (339.6 Mb) and *S. avenae* (431.1 Mb) [42]. Since the *D. noxia* genome size estimates were lower than the reported haploid genome sizes determined via flow cytometry, we used the published genome sizes for the *D. noxia* for further analysis. All read sets fitted the GenomeScope 2.0 model for diploid species with the predicted percentage of the genomes to be repeated between 24.8% and 29.2% for *D. noxia*, 22.5% for *A. gossypii* and 31.1% for *S. avenae* (Supplementary Table S2).

When comparing the genome metrics after analyses with the different assemblers, we ordered the assemblers across all the

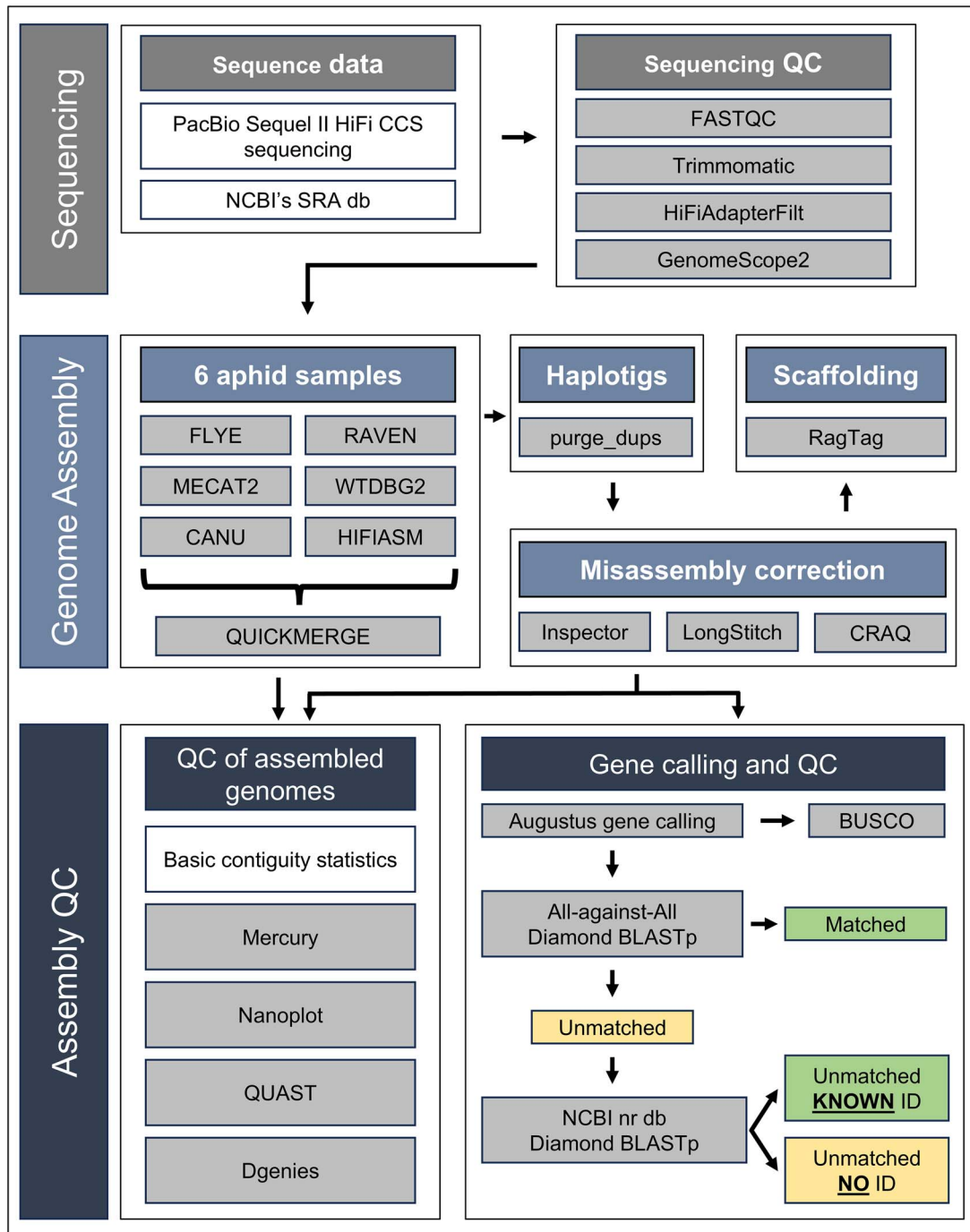


Figure 1. Pipeline applied for the assembly and quality control analysis of six PacBio HiFi read sets.

read sets consistently in terms of contiguity (N50, L50, N90, L90), number of contigs, and the single largest contig (Table 1, Fig. 2). For the *D. noxia* read sets, based on our criteria the assemblers could be ordered from most to least effective as follows: Hifiasm > Canu > Mecat2 > Wtdbg2 > Flye > Raven. For the *A. gossypii* and *S. avenae* read sets, the order only differed slightly in the third and fourth positions as follows: Hifiasm > Canu > Wtdbg2 > Mecat2 > Flye > Raven (Supplementary Table S3). When estimation of genome size was the desired outcome, we found the genome assemblies obtained from Mecat2 to be the closest to the published genome sizes for *D. noxia* (except for biotype RWA-SAM, which was Hifiasm) and *S. avenae*, while Wtdbg2 produced the closest assembly size to that published for *A. gossypii*.

To assess whether merging the first and second most contiguous assemblies would improve the final genome assembly, we applied quickmerge [27], using Hifiasm as a query and Canu as a reference, and found that merging the assemblies had improved the contiguity when compared to either the individual Hifiasm or Canu assemblies (Fig. 3, Supplementary Fig. S1.1–1.5), except for the RWA-SA5 read set whose merged assembly had a lower maximum contig size. When removing haplotigs through purge_dups, all assemblies had improved contiguity (except for the Flye assemblies of RWA-SA1 and RWA-SAM2) and the merged assembly of *A. gossypii* had an increase in two contigs. Between all the read sets, the assembly produced through Raven had the least amount of contigs removed (as a % of initial) and Wtdbg2 had

Table 1. Contiguity metrics compared between the different assemblies for the RWA-RWA-SA1 read set. Values in brackets represent the base assemblies and those without the purged assemblies.

Assembly	Number of contigs	N50 (Mb)	L50	N90 (Mb)	L90	Number of contigs >50Kbp	Assembly size (Mb)	Largest contig (Mb)	Estimated genome size/ Assembly size
RWA-SA1 Canu	84 (217)	24.75 (19.12)	6 (8)	2.39 (1.36)	25 (56)	159 (159)	433.3 (531.0)	55.77 (55.77)	4% (27%)
RWA-SA1 Flye	1814 (3162)	0.30 (0.30)	374 (660)	0.1 (0.1)	1254 (2174)	1741 (2955)	389.0 (674.2)	3.60 (3.67)	-7% (61%)
RWA-SA1 Hifiasm	45 (68)	73.90 (73.90)	3 (3)	11.10 (7.46)	8 (9)	34 (48)	434.3 (439)	82.73 (82.77)	4% (5%)
RWA-SA1 Merged	35 (54)	73.90 (73.90)	3 (3)	41.69 (11.1)	6 (7)	25 (38)	431.5 (451.6)	88.89 (88.89)	3% (8%)
RWA-SA1 MECAT2	672 (1322)	2.66 (2.46)	38 (43)	0.43 (0.23)	167 (232)	327 (466)	377.5 (402)	11.17 (11.18)	-10% (-4%)
RWA-SA1 Raven	2123 (2229)	0.27 (0.26)	412 (443)	0.08 (0.08)	1358 (1482)	1713 (1881)	364.8 (381.7)	1.44 (1.44)	-13% (-9%)
RWA-SA1 wtdbg2	742 (953)	2.04 (2.00)	45 (46)	0.31 (0.29)	215 (227)	415 (432)	344.8 (348.9)	9.34 (9.34)	-18% (-17%)

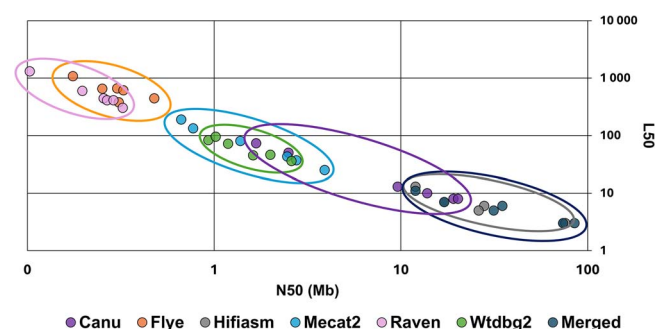


Figure 2. A dot-plot graph of the L50 and N50 values of all the read set assemblies plotted. Coloured circles denote the grouping from the various assemblers used.

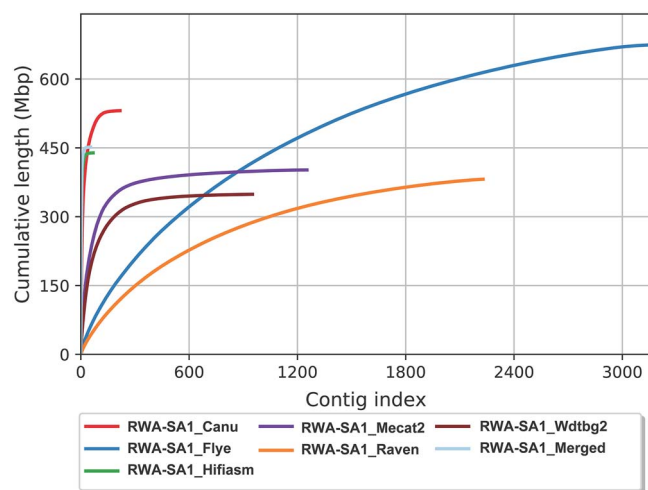


Figure 3. A line graph of the cumulative length of the seven RWA-SA1 assemblies over the number of contigs each assembly consists of.

the lowest decrease in assembly size (as a % of initial) while the assemblies from Canu and Flye had the biggest reduction in assembly size.

Assembly quality control

To assess the quality of the different assemblies beyond contiguity, several programs were utilized to assess the accuracy between the different assemblies. Nanoplot [36] was used to assess the reads incorporated into every assembly through mapping with minimap2, while Merquy [33] was used to compare *k*-mer profiles between the assemblies and the read set.

Between all the purged assemblies of all the samples, the assemblies of Canu and Hifiasm consistently had the highest fraction of bases aligned and the highest mean mapped read length, while the average identity of the mapped reads was always highest in Canu, then followed by either Flye or Hifiasm (Table 2; Supplementary Table S4). When comparing the *k*-mer sets of the reads and the assemblies, those produced by Canu presented with the lowest number of unique *k*-mers that were not found in the read sets, followed in order by Hifiasm > Wtdbg2 > Flye > Mecat2 > Raven. This was also the order that the assemblies scored for Merquy's assembly QV and error rate based on the proportion of unique *k*-mers present only in the assembly and not the read set. The assembly completeness score (as calculated by Merquy by dividing the number of solid *k*-mers found in the assembly with the number of solid *k*-mers found in the read set) of the purged assemblies was always highest for the Canu, Hifiasm, or the Merged assemblies across all read sets, except for the RWA-SA5 read set where the Merged assembly was the second lowest (Supplementary Table S4).

For most assemblies, a slight decrease (<2%) in assembly completeness was observed between the purged and original assemblies, which would indicate that some degree of overpurging occurred. The assemblies that recorded the greatest reduction in assembly completeness (>5%) after purging were those produced by Flye for the read sets from *S. avenae*, RWA-SA5 and RWA-SA1. Interestingly, when comparing the *k*-mer profiles of the Canu and Flye assemblies, where manual filtering of self-reported bubbles and repeats were removed to the base unfiltered Canu and Flye assemblies, the unfiltered assemblies had very large genome sizes and the Canu assembly still had many duplicates in the two-copy peaks (Supplementary Fig. S2).

Merquy was used to generate copy number spectrum plots of the original and purged genome assemblies to investigate missing and duplicated *k*-mers. In Fig. 4, copy number spectrum plots of the original and purged RWA-SA1 assemblies are given, while all the other read sets are presented in Supplementary Fig. S3.1–3.5. In the graphs, the coloured histograms indicate *k*-mer frequency in the assembly while the black histograms indicate *k*-mers missing from the assembly. All assemblies from all the read sets contained missing *k*-mers in the one-copy peak (heterozygous sequences unique to only one haplotype) while only Flye and Wtdbg2 contained missing *k*-mers in the two-copy peak (homozygous sequences shared between haplotypes). *K*-mers duplicated in the assemblies (blue histograms) occurred most prominently in the two-copy peaks and of Flye and Canu and less prominently in the two-copy peaks of the Hifiasm, Merged, Raven, and Mecat2 assemblies and were absent in the two-copy peak of the Wtdbg2

Table 2. RWA-SA1 read statistics produced by Nanoplot and Merquy from mapping the NGS reads onto the various assemblies.

Criteria		Assembler						
		Canu	Flye	Hifiasm	Merged	Mecat	Raven	Wtdbg2
Nanoplot	Mapped reads ^a (minimap2)	908 562 (876 080)	1 019 095 (988 673)	928 390 (940 924)	941 474 (931 612)	1 027 387 (979 498)	1 160 911 (1 124 774)	1 193 439 (1 168 131)
	Bases aligned	98.3% (98.7%)	96.1% (92.7%)	97.4% (98.2%)	97.1% (97.7%)	96.2% (97.0%)	93.0% (93.8%)	90.4% (90.7%)
	Mean mapped read length	15 397 (15 407)	14 115 (13 639)	14 934 (14 966)	14 730 (14 957)	14 028 (14 621)	12 614 (12 952)	11 852 (12 083)
	Mean mapped read length StDev	5624 (5653)	6547 (7088)	6059 (6006)	6251 (6034)	6621 (6192)	7380 (7241)	7667 (7568)
	Average identity	97.6% (97.1%)	95.9% (97.1%)	96.3% (96.4%)	96.3% (96.4%)	95.4% (95.9%)	95.3% (95.5%)	93.8% (94.0%)
	Unique assembly k-mers	685 (925)	141 680 (220 587)	1729 (1844)	1473 (1599)	2,061,988 (2 228 503)	2,367,160 (2 747 602)	31,719 (32 384)
Merquy	Assembly and read set k-mers	433 297 816 (531 023 910)	388 946 023 (674 174 374)	434 251 771 (439 045 016)	431 451 300 (451 593 879)	377 511 443 (401 988 331)	364 796 921 (381 610 049)	344 789 562 (348 835 007)
	QV	71.02 (70.60)	47.40 (47.86)	67.01 (66.78)	67.68 (67.52)	35.63 (35.56)	34.88 (34.42)	53.37 (53.33)
	Error rate	7.90E-08 (8.17E-08)	1.82E-05 (1.64E-05)	1.99E-07 (2.10E-07)	1.71E-07 (1.77E-07)	2.74E-04 (2.78E-04)	3.25E-04 (3.61E-04)	4.60E-06 (4.64E-06)
	Assembly solid k-mers	295 652 086 (300 852 012)	274 044 680 (292 112 690)	296 160 873 (296 537 958)	294 759 074 (297 116 244)	293 318 300 (295 134 334)	286 575 988 (288 306 841)	277 729 843 (278 476 471)
	Read set solid k-mers	31,95,25,948	31,95,25,948	31,95,25,948	31,95,25,948	31,95,25,948	31,95,25,948	31,95,25,948
	Completeness (%)	92.5% (94.2%)	85.8% (91.4%)	92.7% (92.8%)	92.2% (93.0%)	91.8% (92.4%)	89.7% (90.2%)	86.9% (87.2%)

Greyed values in brackets represent the base assemblies and those without the purged assemblies. ^aReads mapped as primary and secondary.

assembly. Comparing the *k*-mer spectra of the purged assemblies to that of the original illustrated that all duplicated *k*-mers were removed from the two-copy peaks in the assemblies that had them and that the overpurging in the Flye assemblies (as recorded in Table 2, Supplementary Table S4) appears to be concentrated in the one-copy peak as the number of missing *k*-mers increased there after purging. When comparing the copy number spectrum plots of RWA-SA5's Canu, Hifiasm, and Merged assemblies, it was apparent that merging this read set's Canu and Hifiasm assemblies produced a more incomplete assembly with increased duplication and missing two-copy *k*-mers. Multiple attempts were made to remerge these assemblies by increasing the parameter stringency of quickmerge's -ml (minimum alignment length to consider merging), -hco (overlap cutoff used in selection of anchor contigs), and -c (overlap cutoff used for extending anchor contigs), but this did not improve the results.

To detect and correct the presence of structural misassemblies in the purged assemblies, three programs were evaluated based on their input only requiring long reads, namely, Inspector, CRAQ, and LongStitch. The corrected assemblies produced by these individual programs were then assessed by comparing the number of unique assembly *k*-mers and changes to contiguity in the corrected assemblies to that of the purged assemblies. Based on the reduction of unique assembly *k*-mers, Inspector was able to detect and correct structural errors in all the assemblies except for the RWA-SA1 Canu assembly where the number of unique assembly *k*-mers remained unchanged and the corrections slightly increased the error rate (Supplementary Table S5). All assemblies had a slight reduction in size and error rate (besides Canu), the total contig number was reduced for some assemblies (Flye, Mecat2 and Wtdbg2), and the overall percentage completeness was negligibly lower (<0.1%). CRAQ was able to reduce assembly unique *k*-mers in most of the assemblies (besides Canu,

Hifiasm, and the Merged assemblies), slightly increased the total contig number in all assemblies (except for Raven where the number of contigs increased >28%), and marginally decreased the contiguity of most of the assemblies (except for the Flye and Hifiasm assemblies that were decreased significantly). Tig-mint was able to reduce assembly unique *k*-mers in most of the assemblies (besides Canu, Hifiasm, and the Merged assemblies), greatly increased the total contig number in all assemblies, and decreased the contiguity of all the assemblies (with the Canu, Hifiasm, and Merged assemblies being decreased significantly). LongStitch, which incorporates scaffolding along with Tig-mint, decreased the total number of assembly-unique *k*-mers in some of the assemblies (Flye, Mecat, and Raven) and slightly increased them in others (Canu, Hifiasm, Merged, and Wtdbg2). The contiguity of all assemblies was greatly increased (except for the Hifiasm and Merged assemblies that were reduced), assembly QV was slightly reduced (except for Raven), and the assembly error was slightly increased (except for Mecat2). As the deciding factors for assembly correction was reduction of assembly unique *k*-mers while maintaining contiguity, Inspector performed best on the contiguous assemblies of Canu, Hifiasm, and Merged, while LongStitch performed best on the less contiguous assemblies.

Gene calling using Augustus [38] was then performed for all the read sets' purged assemblies to compare genic content captured between the different assemblers. As all the assemblies differed in size and contiguity, the number of protein-coding genes identified differed (Supplementary Table S6), and thus, the next step was to ascertain what proportion of the predicted genes were shared between the different assemblies. Protein sequences were obtained for the predicted genes and first analysed with BUSCO to estimate single copy ortholog capture and secondly through use of Diamond BLASTp by performing two-way one-to-one comparisons (Fig 5). Proteins that did not return

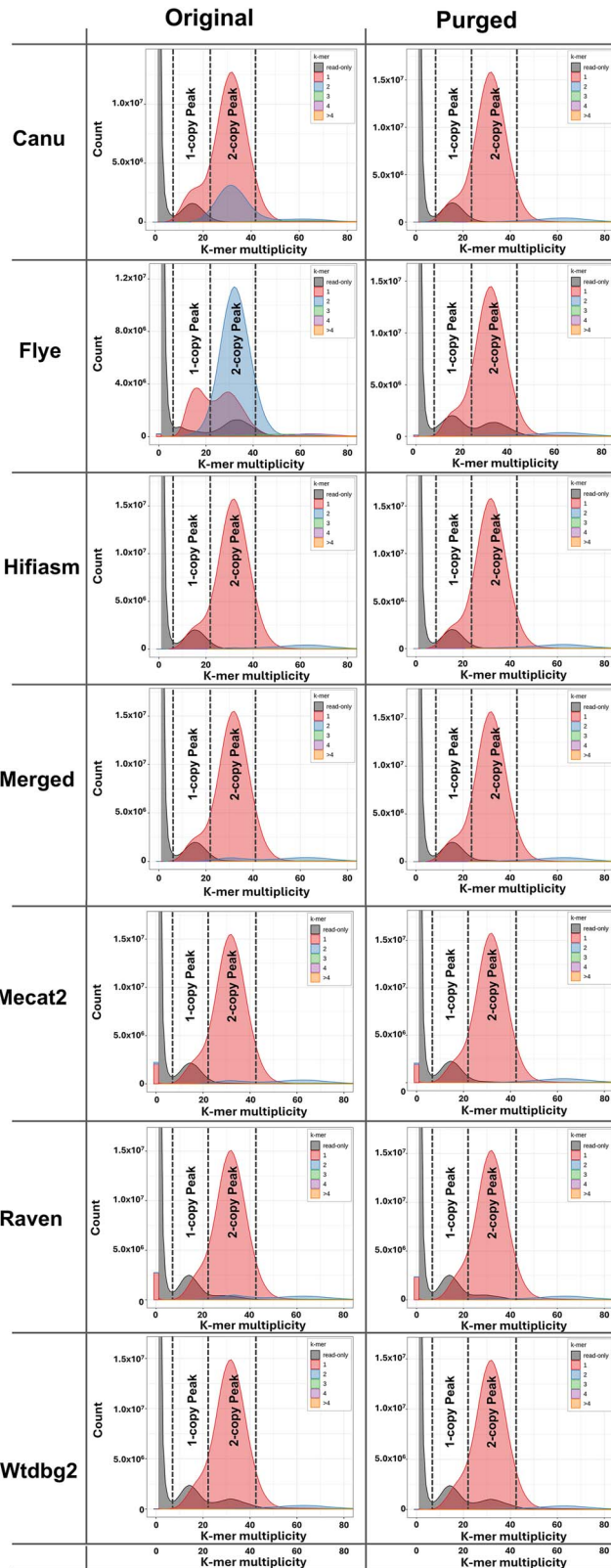


Figure 4. K-mer copy graphs of the RWA-SA1 assemblies as produced by Merqury. The black curve indicates k-mers missing from the assembly, but present in the read set. The coloured curves indicate k-mer multiplicity. The dotted lines represent the boundaries of the 1- and 2-copy k-mer peaks representing the heterozygous (present in only one haplotype) and homozygous (present in both haplotypes) components, respectively.

at least a 95% query coverage and 95% identity match were subjected to a Diamond BLASTp search against the nr database of the NCBI (date downloaded 8 August 2024) and limited to sequences from the Arthropoda phylum (using a 70% query coverage and 70% identity threshold) to determine if these unmatched genes had orthologs on the NCBI. When viewing the results from the RWA-SA1 assemblies (Fig. 5), all assemblies had relatively similar BUSCO results (between 4.1% and 5.6% fragmented BUSCOs and <1% missing) except for the RWA-SA1 Wtdbg2 assembly, which had 11.3% BUSCOs fragmented. This trend was similar for all the other assemblies as well (Supplementary Fig. S4.1–4.5).

Comparing the number of predicted proteins matched between the assemblies revealed that roughly between 9% and 30% were unique to either assembly in the one-to-one comparisons (Fig. 5, Supplementary Table S7). Proteins predicted from the Mecat2 assembly shared the least overall similarity when compared to the other assemblies and presented with the most predicted proteins that did not have orthologs on the NCBI. The Merged assembly produced the overall best matching-predicted protein set to the other assemblies. Given its origin in the Hifiasm and Canu assemblies, this is not surprising. The Merged assembly-predicted protein set also had results comparable to its base assemblies, illustrating that the merging of the assemblies did not introduce any anomalies. Excluding the Merged assembly from the comparison, the Wtdbg2 assembly shared the most predicted protein similarity with the other assemblies. When investigating the results from the other aphids' assemblies (Supplementary Figs. S4.1–4.5), the Merged assemblies from the RWA-SA5 and *A. gossypii* read sets performed worse than their base assemblies, while the other assemblies performed comparable to that of RWA-SA1.

To visualize whole assembly alignments as dot-plot graphs across the assemblies obtained from the minimap2 alignments, we applied Dgenies. When two assemblies share collinearity, a diagonal line should form on the dot-plot graph, while all other dots plotted indicate mapping to multiple contigs (minimap2 default parameter set to 5 secondary alignments). The latter may be due to low sequence complexity, tandem repeats, duplications, or lack of contig uniqueness. As can be seen in Fig. 6, the dot-plot graphs of the RWA-SA1 assembly self-alignments produced a diagonal with the least number of alternative mappings for every target (or reference) assembly. The self-alignments of the more discontinuous assemblies (namely, Flye, Mecat2, Raven, and Wtdbg2) unexpectedly had bases that did not match during the alignment. These unmapped bases could either have originated from the more prevalent secondary alignments or, as stated by the authors of minimap2 in its manual, that long stretches of low-complexity regions are prone to suboptimal seeding, which would produce inferior alignments. When the more contiguous assemblies (from Canu, Hifiasm, and Merged) acted as query, they were unable to map accurately to the more discontinuous assemblies as can be seen from the lack of a diagonal line forming. When these assemblies acted as targets during the alignments, they produced dot plots with the clearest diagonals and least alternative mappings, followed closely by the Mecat2 assembly, while the target assemblies with the least defined diagonal and most alternate mappings were Flye, followed by Raven and lastly Wtdbg2. When viewing the dot-plot graphs of Hifiasm and Merged assemblies mapped against the Canu assembly, the lack of a continuous diagonal line appears to be due to a sorting issue during graphing as no lines overlap vertically or horizontally. The same appears to be the case when viewing the dot plot of the

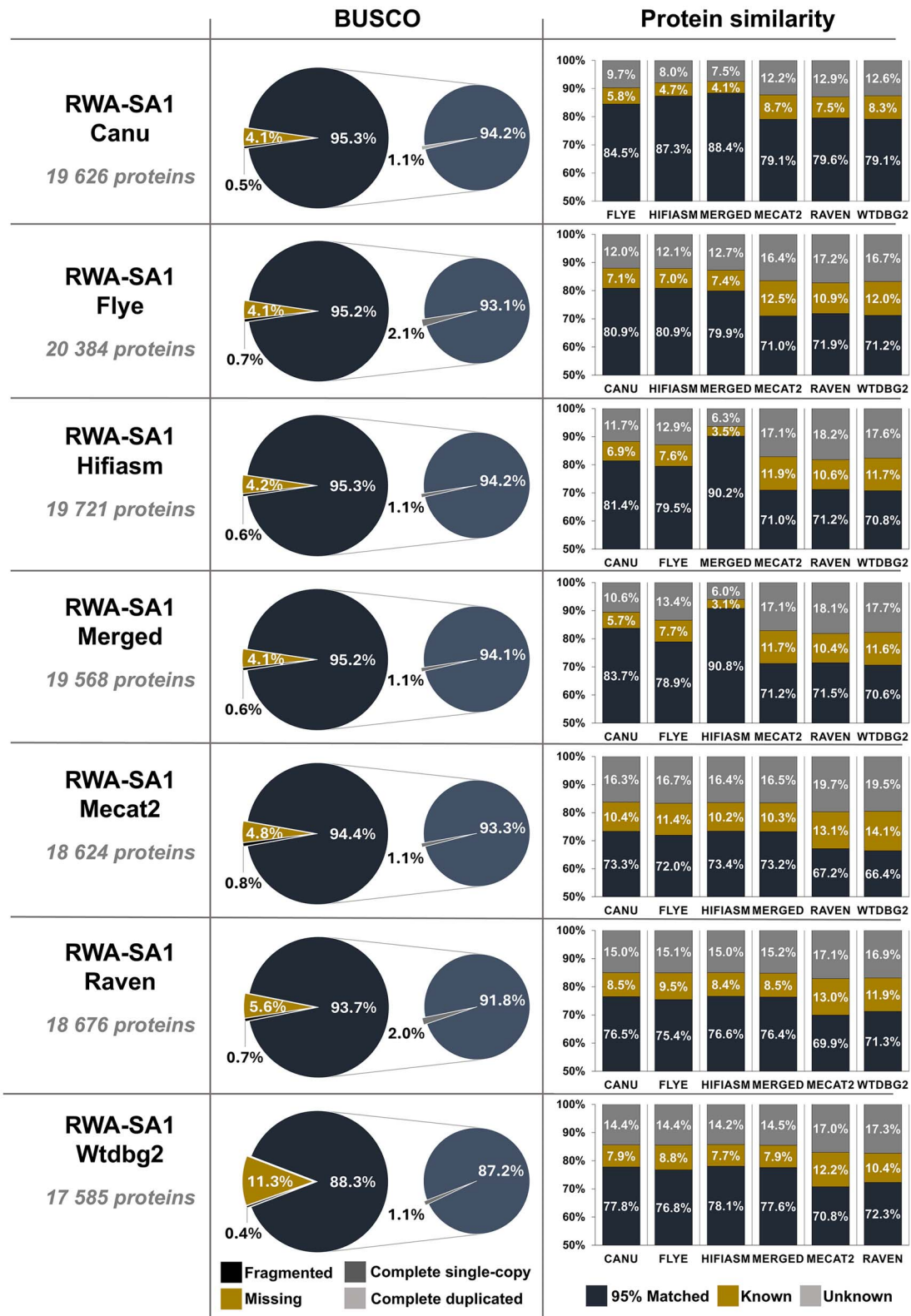


Figure 5. A combined graph showing the BUSCO results obtained for the various RWA-SA1 assemblies in pie charts and the number of proteins shared between the different assemblies as determined through Diamond-BLASTp in bar charts.

Merged assembly mapped against the Hifiasm assembly, illustrating that for the RWA-SA1 set, no anomalies were introduced during merging. The other read sets performed similarly to that of the RWA-SA1 set (Supplementary Fig. S5.1–S5.5), except for the Merged assemblies of the RWA-SA5 and RWA-SAM read sets and

the Canu assemblies of the *A. gossypii* and *S. avenae* read sets. The Merged assembly of the RWA-SA5 and RWA-SAM read sets had increased alternative mappings and several duplications when mapped to the target Canu and Hifiasm assemblies, as well as a duplication in the Merged self-alignments. The Canu assembly

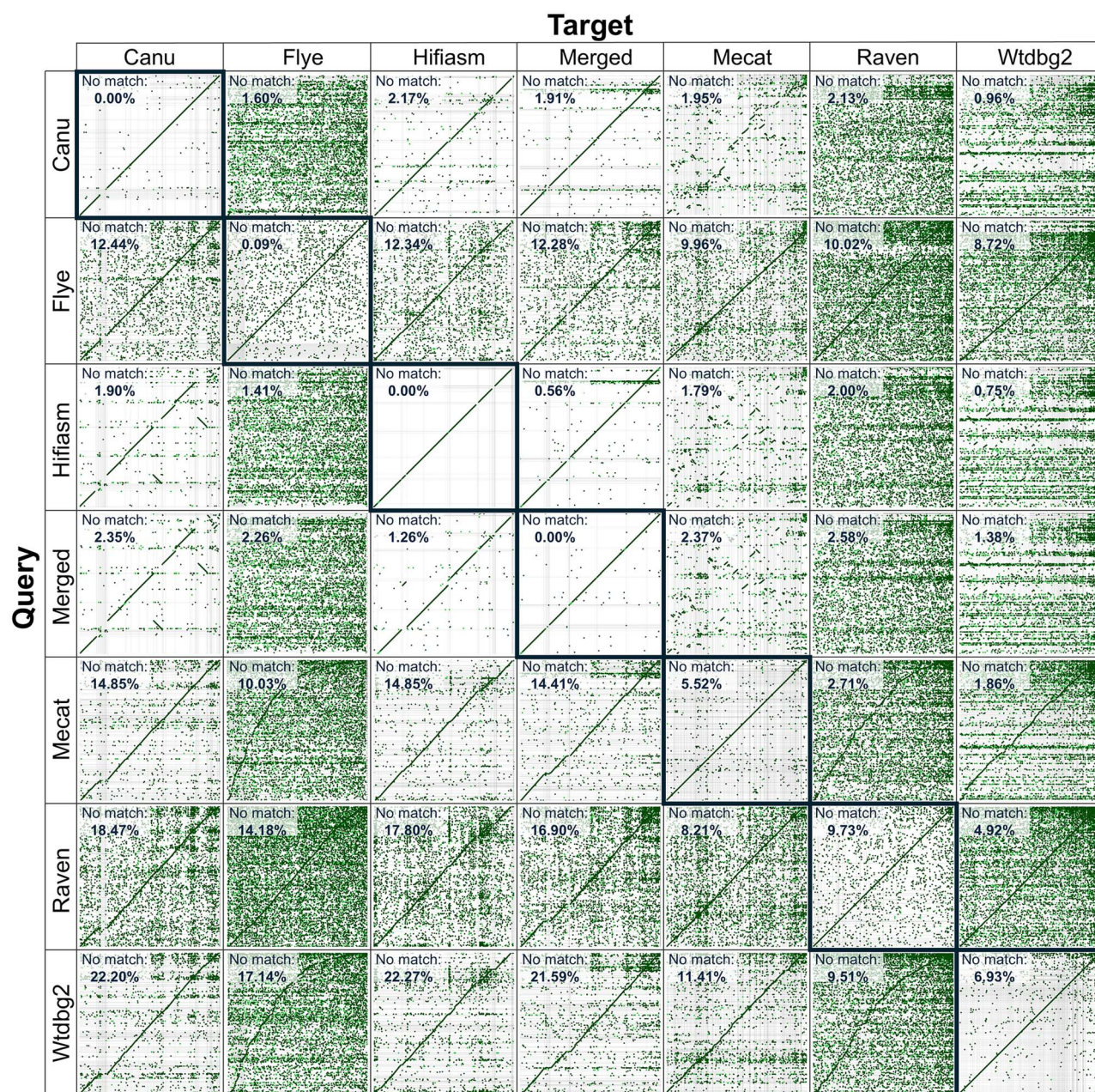


Figure 6. Dot-plot graphs of whole assembly alignments of RWA-SA1 as performed by minimap2 and graphed with Dgenies. Noted on the graphs are the percentage of bases that did not align between the respective assemblies.

of the *A. gossypii* and *S. avenae* read sets were too discontinuous to properly act as targets when the more contiguous Merged and Hifiasm assemblies were mapped against them. Considering the query assembly that had the fewest bases not matching when mapped as a query to the target assemblies (no match percentage) for the RWA-SA1 assemblies, Hifiasm performed the best, followed by Canu and then the Merged assembly. The query assemblies that had the greatest number of bases not matching the target assemblies were Wtdbg2, followed by Raven, Flye, and lastly Mecat2. This trend was mostly similar for the other read sets, except for the *A. gossypii* read set where the order of assemblies with the least number of bases not matching when acting as query were Hifiasm, followed by Canu and then the Flye assembly.

Ranking the qualitative performance of each assembly

All assemblies were ranked according to the various metrics assessed in the pipeline (Table 3; Supplementary Table S8) to compare the overall performance of each assembler for each read set. As the main aim of this study was the construction of the most contiguous genome assembly, while not compromising on quality or the introduction of misassembly artefacts, the Merged assemblies displayed the overall highest rank for the RWA-SA1, RWA-SAM2, *A. gossypii*, and *S. avenae* read sets. For the RWA-SA5 and RWA-SAM read sets, the highest scoring assemblies were from the Hifiasm assembler, as the Merged assembly from the RWA-SAM read set introduced artificial duplications (as seen from Supplementary Fig. S5.1 and S5.2) and the Merged assembly from

Table 3. All assemblies from the RWA-SA1 read set were ranked according to various metrics obtained from the programs utilized to assess assembly quality.

		Canu	Flye	Hifiasm	Merged	Mecat2	Raven	Wtdbg2
Contiguity	Closest to expected genome size	3	4	2	1	5	6	7
	Lowest number of contigs	3	6	2	1	4	7	5
	Highest N50	2	5	1	1	3	6	4
	Highest N90	3	6	2	1	4	7	5
	Largest contig	3	6	2	1	4	7	5
Nanoplot	Largest fraction of bases aligned	1	5	2	3	4	6	7
	Largest mean mapped read length	1	4	2	3	5	6	7
	Highest average identity	1	3	2	2	4	5	6
Merquy	Lowest unique assembly k-mers	1	5	3	2	6	7	4
	Highest QV	1	5	3	2	6	7	4
	Lowest error rate	1	5	3	2	6	7	4
	Completeness	2	7	1	3	4	5	6
	Missing two-copy k-mers	1	7	1	1	1	7	7
BUSCO	Highest number of single-copy orthologs	1	4	2	3	5	6	7
	Lowest number of duplicated orthologs	2	5	1	2	2	4	3
	Lowest number of fragmented orthologs	2	5	3	4	6	5	1
	Lowest number of missing orthologs	2	1	3	2	4	5	6
BLASTp	Highest number of predicted proteins	3	1	2	4	6	5	7
	Highest matched average identity	3	4	2	1	7	6	5
	Lowest average NCBI unmatched proteins	3	3	2	1	6	5	4
Dgenies	Lowest number of bases not matched	3	7	1	2	4	5	6
	Presence of duplications	1	1	1	1	1	1	1
TOTAL		43	99	43	43	97	125	111

Table 4. Contiguity metrics of *D. noxia* assemblies before and after homology-based scaffolding using the RWA-SAM2 Merged assembly.

	Number of contigs/ scaffolds	N50 (Mb)	L50	N90 (Mb)	L90	Number of contigs/ scaffolds >50Kbp	Assembly size (Mb)	Largest contig/ scaffold (Mb)
RWA-SA1 Merged contigs	35	73.900	3	41.690	6	25	431.5	88.886
RWA-SA1 Merged RagTag scaffold	26	73.900	3	41.690	6	22	433.7	88.886
RWA-SA5 Hifiasm contigs	385	27.932	6	2.828	23	255	430.2	57.765
RWA-SA5 Hifiasm RagTag scaffold	200	27.932	6	2.449	23	180	454.2	57.765
RWA-SAM Hifiasm contigs	69	26.102	5	3.900	19	45	381.7	70.217
RWA-SA1 Hifiasm RagTag scaffold	25	37.721	4	8.913	13	25	408.6	76.097

the RWA-SA5 read set had lower contiguity than its base assemblies (as seen in [Supplementary Table S3](#)), increase in missing two-copy *k*-mers (as seen in [Supplementary Fig. S3.1](#)), and introduction of artificial duplications (as seen in [Supplementary Fig. S5.1](#)).

To understand the large difference in contiguity between the best assemblies of the four *D. noxia* read sets, Nanoplot was used to visualize the sequenced read lengths compared to the mapped read lengths and mapped percentage identity. The overall mapped identity between the *D. noxia* read sets appears to be relatively equal ([Supplementary Fig. S6](#)), which would indicate that sequencing quality wasn't a determining factor for the difference in obtained contiguity. Sequencing depth also appears not to have affected the maximum available contiguity as the RWA-SA1 read set had the lowest sequencing coverage ($\times 33$) but the second highest contiguity, while the RWA-SAM read set had the highest sequencing coverage ($\times 48$) but the second lowest contiguity ([Table 1](#); [Supplementary Table S3](#)). Collectively, findings suggested the mean sequence length was the greatest contributor to assembly contiguity as the order for the largest mapped read length correlates to the highest contiguity (i.e. RWA-SAM2 > RWA-SA1 > RWA-SAM > RWA-SA5) ([Supplementary Fig. S7](#); [Supplementary Table S4](#)).

To improve further on the contiguity of the selected *D. noxia* assemblies, RagTag [32], a homology-based scaffolder, was used to scaffold the selected RWA-SA1, RWA-SA5, and RWA-SAM assemblies against that of the RWA-SAM2 Merged assembly as it was by far the most contiguous. After scaffolding, it was found that RagTag had also unexpectedly incorporated gaps of exactly 100 bp, although the minimum gap length for scaffolding was set to 30 kbp. All 100 bp gaps were removed by breaking the scaffolds at these positions, leaving only gaps larger than 30 kbp. RagTag was able to scaffold six contigs in the RWA-SA1 Merged assembly, 64 contigs in the RWA-SAM Hifiasm assembly, and 248 contigs in the RWA-SA5 Hifiasm assembly. Homology-based scaffolding resulted in a limited increase in contiguity of RWA-SA1, moderate improvement for RWA-SA5, and the most improvement for RWA-SAM ([Table 4](#)).

As one of the main goals for this study was an improved genome assembly for *D. noxia*, the most contiguous assembly from this study (RWA-SAM2 with an N50 of ~ 85 Mb) was compared to the two available *D. noxia* genomes (namely, the RWA-US2 and original RWA-SAM short-read genome assemblies [4, 43]). The RWA-SAM2 assembly N50 represents an increase of $\sim 6400\times$ (or $214\times$ when considering scaffolds) and $\sim 14\ 600\times$ compared to

the short-read RWA-US2 and RWA-SAM assemblies, respectively, while incorporating 47% and 15% more ungapped bases than the short-read RWA-US2 and RWA-SAM assemblies, respectively (Supplementary Table S9).

Conclusion

Typically, the main objective during genome assembly is to obtain the most contiguous genome, while little attention is given to the underlying quality of such an assembly. By following the pipeline outlined in Fig. 1, TGS read sets from four different *D. noxia* biotypes, as well as *A. gossypii* and *S. avenae*, delivered genome assemblies close to chromosome level for these hemipteran species. The genomes assembled during this study showed a significant increase in both contiguity and overall assembly size when compared with the two earlier *D. noxia* genome assemblies. Having sequencing read sets available for multiple distinct populations of one species, further allowed for a comparative view of how these assemblers handled TGS read sets of various lengths and sequencing depths. Collectively, findings suggest that the sequence length contributed more to assembly contiguity than sequence depth. The use of quickmerge improved the contiguity and decreased errors when compared to its base assemblies. However, it also introduced anomalies in two of the assemblies, indicating that it should be used with caution and that verification of genome quality needs to be assessed after its use.

Key Points

- Chromosome-level assemblies of high quality were obtained for all the *Diuraphis noxia* biotypes in this study.
- Longer read lengths had a much greater effect on contiguity than coverage.
- Haplotig purging of the assemblies decreased the overall error rate and unique assembly *k*-mers.
- Gene prediction between assemblies of the same read set differed greatly and showed lower than expected similarity to each other.
- BUSCO analysis of conserved single-copy orthologs, although important to estimate gene capture in assemblies, does not effectively discriminate assembly quality.

Author contributions

N.F.V.B. and A.M.O. conceptualized the experimental design, while N.F.V.B. wrote the manuscript. N.F.V.B. and V.F.N. performed the formal analyses of the data. A.M.O. and V.F.N. edited the manuscript before submission. All authors read and approved the final manuscript.

Supplementary data

Supplementary data are available at Briefings in Bioinformatics online.

Conflict of interest: None declared.

Funding

This work was supported by grants from the National Research Foundation of South Africa (Grant No. CSRU180414320893) and the South African Winter Cereal Industry Trust (Grant No. WCT2001/02).

Data availability

The *Diuraphis noxia* genomes constructed in this study have been uploaded to the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) with the accessions GCA_037042775.1 (RWA-SA1), GCA_037042715.1 (RWA-SA5), GCA_037042725.1 (RWA-SAM), and GCA_037042655.1 (RWA-SAM2).

References

1. Stork NE. How many species of insects and other terrestrial arthropods are there on earth? *Annu Rev Entomol* 2018;**63**:31–45. <https://doi.org/10.1146/annurev-ento-020117-043348>
2. Mille C, Jourdan H, Cazères S. et al. New data on the aphid (Hemiptera, Aphididae) fauna of New Caledonia: Some new biosecurity threats in a biodiversity hotspot. *ZooKeys* 2020;**943**:53. <https://doi.org/10.3897/zookeys.943.47785>
3. Mathers TC, Wouters RH, Mugford ST. et al. Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. *Mol Biol Evol* 2021;**38**:856–75. <https://doi.org/10.1093/molbev/msaa246>
4. Burger NFV, Botha AM. Genome of Russian wheat aphid an economically important cereal aphid. *Stand Genomic Sci* 2017;**12**: 1–12.
5. Liao X, Li M, Zou Y. et al. Current challenges and solutions of *de novo* assembly. *Quant Biol* 2019;**7**:90–109. <https://doi.org/10.1007/s40484-019-0166-9>
6. Kong W, Wang Y, Zhang S. et al. Recent advances in assembly of complex plant genomes. *Genomics Proteomics Bioinformatics* 2023;**21**:427–39. <https://doi.org/10.1016/j.gpb.2023.04.004>
7. Espinosa E, Bautista R, Larrosa R. et al. Advancements in long-read genome sequencing technologies and algorithms. *Genomics* 2024;**116**:110842. <https://doi.org/10.1016/j.ygeno.2024.110842>
8. Van Dijk EL, Jaszczyzyn Y, Naquin D. et al. The third revolution in sequencing technology. *Trends Genet* 2018;**34**:666–81. <https://doi.org/10.1016/j.tig.2018.05.008>
9. Cosma BM, Shirali Hossein Zade R, Jordan EN. et al. Evaluating long-read *de novo* assembly tools for eukaryotic genomes: Insights and considerations. *GigaScience* 2023;**12**:giad100. <https://doi.org/10.1093/gigascience/giad100>
10. Renoz F, Parisot N, Baa-Puyoulet P, et al. PacBio hi-fi genome assembly of *Sipha maydis*, a model for the study of multipartite mutualism in insects. *Sci Data* 2024;**11**:450. <https://doi.org/10.1038/s41597-024-03297-x>
11. Wei HY, Ye YX, Huang HJ. et al. Chromosome-level genome assembly for the horned-gall aphid provides insights into interactions between gall-making insect and its host plant. *Ecol Evol* 2022;**12**:e8815. <https://doi.org/10.1002/ece3.8815>
12. Wang Y, Xu S. A high-quality genome assembly of the waterlily aphid *Rhopalosiphum nymphaeae*. *Sci Data* 2024;**11**:194. <https://doi.org/10.1038/s41597-024-03043-3>
13. Foox J, Tighe SW, Nicolet CM. et al. Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. *Nat Biotechnol* 2021;**39**:1129–40. <https://doi.org/10.1038/s41587-021-01049-5>
14. Olson ND, Wagner J, McDaniel J. et al. PrecisionFDA truth challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell genom* 2022;**2**:100129. <https://doi.org/10.1016/j.xgen.2022.100129>
15. Burger NF, Nicolis VF, Botha AM. Host-specific co-evolution likely driven by diet in *Buchnera aphidicola*. *BMC Genomics* 2024;**25**:153. <https://doi.org/10.1186/s12864-024-10045-3>

16. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
17. Bolger AM, Lohse M. Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**: 2114–20. <https://doi.org/10.1093/bioinformatics/btu170>
18. Sim SB, Corpuz RL, Simmonds TJ. et al. HiFiAdapterFilter, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* 2022;**23**:157. <https://doi.org/10.1186/s12864-022-08375-1>
19. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 2020;**11**:1432. <https://doi.org/10.1038/s41467-020-14998-3>
20. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;**27**:764–70. <https://doi.org/10.1093/bioinformatics/btr011>
21. Cheng H, Concepcion GT, Feng X. et al. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifi-asm. *Nat Methods* 2021;**18**:170–5. <https://doi.org/10.1038/s41592-020-01056-5>
22. Kolmogorov M, Yuan J, Lin Y. et al. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**:540–6. <https://doi.org/10.1038/s41587-019-0072-8>
23. Nurk S, Walenz BP, Rhie A. et al. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* 2020;**30**:1291–305. <https://doi.org/10.1101/gr.263566.120>
24. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**:155–8. <https://doi.org/10.1038/s41592-019-0669-3>
25. Vaser R, Šikić M. Time-and memory-efficient genome assembly with raven. *Nat Comput Sci* 2021;**1**:332–6. <https://doi.org/10.1038/s43588-021-00073-4>
26. Xiao CL, Chen Y, Xie SQ. et al. MECAT: Fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat Methods* 2017;**14**:1072–4. <https://doi.org/10.1038/nmeth.4432>
27. Chakraborty M, Baldwin-Brown JG, Long AD. et al. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* 2016;**44**:e147. <https://doi.org/10.1093/nar/gkw654>
28. Guan D, McCarthy SA, Wood J. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 2020;**36**:2896–8. <https://doi.org/10.1093/bioinformatics/btaa025>
29. Chen Y, Zhang Y, Wang AY. et al. Accurate long-read *de novo* assembly evaluation with inspector. *Genome Biol* 2021;**22**:1–21. <https://doi.org/10.1186/s13059-021-02527-4>
30. Coombe L, Li JX, Lo T. et al. LongStitch: High-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* 2021;**22**:1–13. <https://doi.org/10.1186/s12859-021-04451-7>
31. Li K, Xu P, Wang J. et al. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat Commun* 2023;**14**:6556. <https://doi.org/10.1038/s41467-023-42336-w>
32. Alonge M, Lebeigle L, Kirsche M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* 2022;**23**:258. <https://doi.org/10.1186/s13059-022-02823-7>
33. Rhie A, Walenz BP, Koren S. et al. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 2020;**21**:1–27. <https://doi.org/10.1186/s13059-020-02134-9>
34. Gurevich A, Saveliev V, Vyahhi N. et al. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5. <https://doi.org/10.1093/bioinformatics/btt086>
35. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. <https://doi.org/10.1093/bioinformatics/bty191>
36. De Coster W, Rademakers R. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;**39**:btad311. <https://doi.org/10.1093/bioinformatics/btad311>
37. Cabanettes F, Klopp C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 2018;**6**:e4958. <https://doi.org/10.7717/peerj.4958>
38. Stanke M, Morgenstern B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nuc Acids Res* 2005;**33**:W465–7. <https://doi.org/10.1093/nar/gki458>
39. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60. <https://doi.org/10.1038/nmeth.3176>
40. Simão FA, Waterhouse RM, Ioannidis P. et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**:3210–2. <https://doi.org/10.1093/bioinformatics/btv351>
41. Novotná J, Havelka J, Starý P. et al. Karyotype analysis of the Russian wheat aphid, *Diuraphis noxia* (Kurdjumov) (Hemiptera: Aphididae) reveals a large X chromosome with rRNA and histone gene families. *Genetica* 2011;**139**:281–9. <https://doi.org/10.1007/s10709-011-9546-4>
42. Wenger JA, Cassone BJ, Legeai F. et al. Whole genome sequence of the soybean aphid. *Aphis glycines*. *Insect Biochem Mol Biol* 2020;**123**:102917. <https://doi.org/10.1016/j.ibmb.2017.01.005>
43. Nicholson SJ, Nickerson ML, Dean M. et al. The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC Genomics* 2015;**16**:1–16. <https://doi.org/10.1186/s12864-015-1525-1>