Aphidinae comparative genomics resource

Thomas C. Mathers[1*], Sam T. Mugford[1], Roland H. M. Wouters[1], Darren Heavens[2], Anna-Maria Botha[3], David Swarbreck[2], Cock Van Oosterhout[4] and Saskia A. Hogenhout[1*]

[1]Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, United Kingdom

[2]Earlham Institute, Norwich Research Park, Norwich, United Kingdom

[3]Genetics Department, Stellenbosch University, Stellenbosch, South Africa

[4]School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom

* Corresponding authors

Email: thomas.mathers@jic.ac.uk

Email: saskia.hogenhout@jic.ac.uk

Here we provide early access to 18 new genome assemblies, including 8 assembled to chromosome-scale, for aphids from the subfamily Aphidinae (**Figure 1**). Sample information and genome assembly and annotation metrics are given in **Table 1** (see attached files). A summary of all newly generated sequence data used to create the new assemblies and annotations is given in **Table 2** (see attached files). For consistency and to aid comparative analysis, all genomes have been annotated using the same repeat masking and RNA-seq-based gene prediction pipeline that we recently used to annotate the model aphid species *Myzus persicae* and *Acyrthosiphon pisum* (Mathers *et al.* 2021). Using this pipeline we also provide new annotations for our previously published genome assemblies of *Aphis glycines*, *Myzus cerasi* and *Pentalonia nigronervosa* (Mathers 2020; Mathers *et al.* 2020). The genome assemblies and annotations are made freely available without restriction, we only request that this Zenodo resource is cited when using the data. Raw sequence data upload to NCBI is underway and full details of all accessions will be given in an updated version of this resource. Manuscripts are in preparation describing the individual genome assemblies in detail and larger comparative genome analyses and we will update this resource with additional citation information as papers are published.

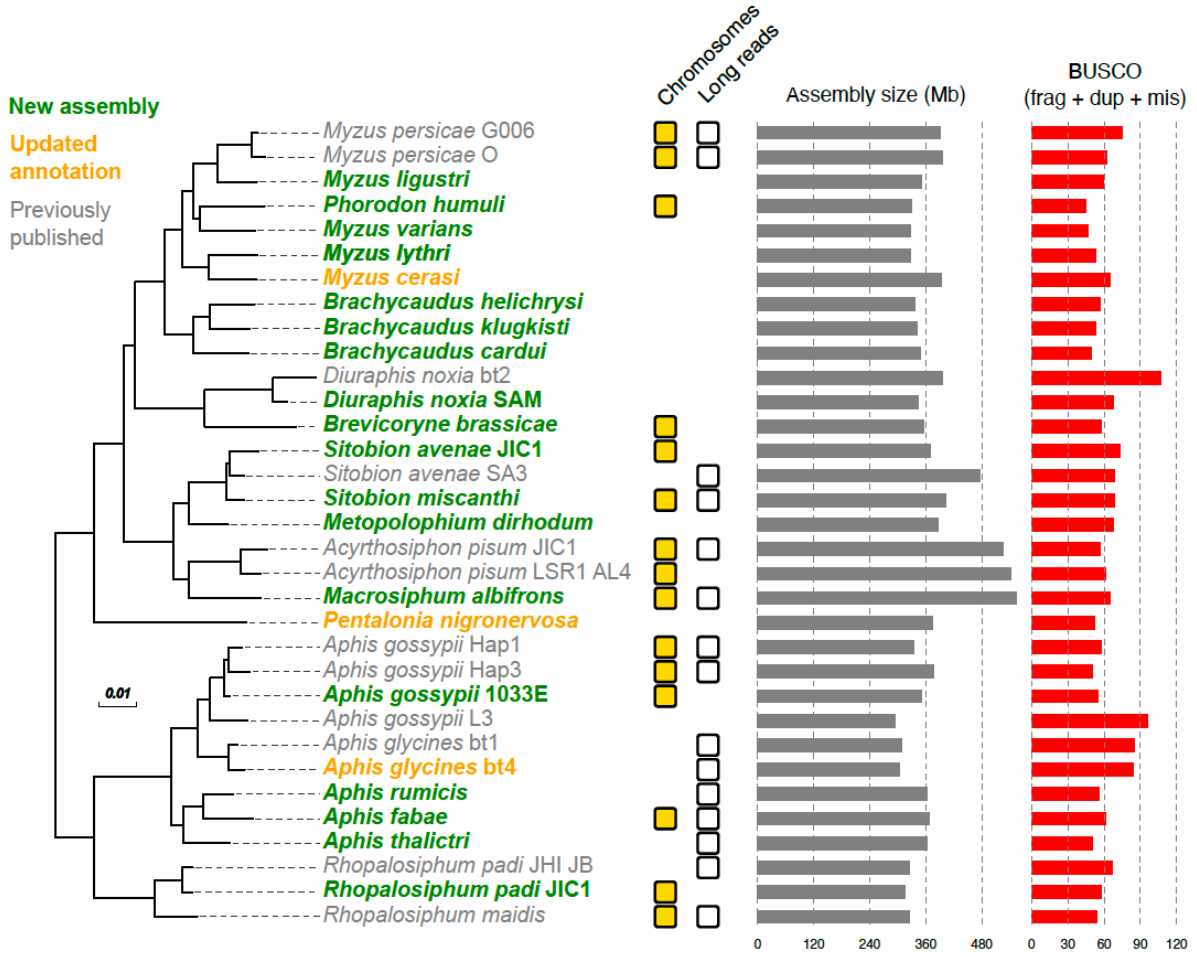


Figure 1: Maximum likelihood phylogeny of all newly generated and previously published genomes from the subfamily Aphidinae. For previously published assemblies, we include the most recent assembly version available for a given isolate. Details of the included assemblies are given in **Table 3** (see attached files). The phylogeny is based on a concatenated alignment of 928 conserved BUSCO genes. Branch lengths are given in amino acid substitutions per site. All branches received maximum support according to the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999) implemented in FastTree (Price *et al.* 2009, 2010) with 1,000 resamples. The tree was rooted with *Eriosoma lanigerum* from the subfamily Eriosomatinae (not shown). Genomes assembled to chromosome-scale with Hi-C data are annotated with a yellow square and genomes assembled with long-read sequence data are annotated with a white square. Assembly size is given inclusive of N content. The sum of fragmented (frag), duplicated (dup) and missing (mis) BUSCO Arthropoda (n=1,066) genes for each assembly is given to indicate assembly quality (lower = better).

General methodology

Sampling and species identification

Samples for DNA and RNA extraction were either taken from the John Innes Centre insectary (Norwich, UK) or collected from the wild in the UK. Species were identified based on morphology, host plant species and by DNA barcoding with the cytochrome oxidase 1 (COI) mitochondrial gene. We made use of www.InfluentialPoints.com and www.aphidsonworldsplants.info for species identification in the field.

Quality control

Regardless of the sequencing approach, we aimed to generate high-quality haploid genome assemblies, which maximise assembly completeness and minimise the inclusion of erroneously duplicated content (i.e. haplotigs). Genome assemblies were assessed by generating K-mer spectra, a procedure that involved comparing K-mer content of the raw sequencing reads to the K-mer content of the genome assembly with the K-mer analysis toolkit (KAT; [Mapleson *et al.* 2017]). We also assessed assembly completeness and duplication levels by searching for arthropod Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs; n=1,066) using BUSCO v3 (Simão *et al.* 2015; Waterhouse *et al.* 2018). Each *de novo* assembly was checked for contamination and the presence of symbiont genomes by generating taxon-annotated GC content-coverage plots (known as a BlobPlots) with BlobTools (Kumar *et al.* 2013; Laetsch and Blaxter 2017). Where Hi-C data was generated for chromosome-scale scaffolding, we carried out initial scaffolding with the 3dDNA assembly pipeline (Dudchenko *et al.* 2017), followed by manual curation with Juicebox Assembly Tools (JBAT; [Dudchenko *et al.* 2018]).

Symbiont genomes

In many cases bacterial symbiont genomes were co-assembled alongside their aphid hosts. For each species, symbiont genomes were identified in the assembly BlobPlot and removed. The symbiont genomes are included as separate assembly files as part of this data release, but we have not carried out any further curation or quality control.

Genome annotation

All genomes were annotated following Mathers *et al.* (2020b). Each genome was soft-masked with RepeatMasker v4.0.7 (Tarailo-Graovac and Chen 2009; Smit *et al.* 2015) using known Insecta repeats from Repbase (Bao *et al.* 2015) with the parameters "-e ncbi -species insecta -a -xsmall -gff". RNA-seq reads were mapped to the genomes with HISAT2 v2.0.5 (Kim *et al.* 2015). Un-stranded RNA-seq reads were aligned to the genomewith HISAT2 with the parameters "–max-intronlen 25000 –dta-cufflinks" followed by sorting and indexing with SAMtools v1.3 (Li *et al.* 2009). Strand-specific RNA-seq reads were mapped to the genomes similarly to the un-stranded ones, except for the addition of the HISAT2 parameter "--rna-strandness RF". We then ran BRAKER2 (Hoff *et al.* 2015, 2019) with UTR training and prediction enabled with the parameters "--softmasking --gff3 --UTR=on". Strand-specific RNA-seq alignments were split by forward and reverse strands and passed to BRAKER2 as separate BAM files to improve the accuracy of UTR models as recommended in the BRAKER2 documentation.

Phylogenomic analysis

To place our new genome assemblies in a phylogenetic context we gathered all (to the best of our knowledge) previously published and publicly available Aphidinae genome assemblies

and combined them with our newly generated assemblies and the genome of *Eriosoma lanigerum* (Biello *et al.* 2021). Full details of the included genome assemblies are given in **Table 3** (see attached files). For each genome, we searched for BUSCO Arthropoda (n = 1,066) genes with BUSCO v3 and extracted the annotated protein sequences. Using these data, we carried out phylogenetic analysis using the OrthoFinder v2.5.4 pipeline (Emms and Kelly 2015, 2019) with the parameters "-X -M msa -S diamond -T fasttree". The resulting species tree was manually rooted with *E. lanigerum* which belongs to the aphid subfamily Eriosomatinae (Biello *et al.* 2021). The complete OrthoFinder output including protein alignments, individual gene trees, the species tree and summary statistics is included in the Zenodo database (see attached files).

Data description

Below, brief descriptions of the genome assemblies and annotations of each species are given. The genomes in this resource have been combined from multiple projects using a variety of sequencing and assembly approaches. Species are grouped together where common methodologies were used.

Within the Zenodo database, the frozen genome assemblies and annotations are organised into separate folders for each species (see attached file "Frozen_release.zip"). In each folder, we provide a BUSCO summary of the genome and gene set along with the following figures to assess the quality of each genome assembly: (i) K-mer spectra comparing Illumina short reads to the final assembly; (ii) order-level and genus-level BlobPlot before assembly filtering to show any contamination or symbionts present; and (iii) Hi-C contact map before and after JBAT review (if Hi-C data is present for the species in question).

***Aphis fabae* JIC1**

Sampling and sequencing library preparation

Aphis fabae individuals were sampled from a clonal lineage (dubbed JIC1) maintained at the JIC insectary on *Vicia faba* (broad bean). The isolate was originally collected from *Euonymus europaeus* (spindle) at Strumpshaw Fen (Norfolk, UK) in 2012. All DNA and RNA extractions were performed on bulked individuals. We prepared DNA extractions and sequencing libraries for amplification-free Illumina short-read and Oxford Nanopore Technologies (ONT) long-read sequencing following procedures described in Mathers *et al.* (2021). We also sent bulked, snap-frozen, individuals from the same colony to Dovetail Genomics (Santa Cruz, CA) for *in vivo* Hi-C library preparation and sequencing. Hi-C libraries were prepared using the *Dpn*II restriction enzyme following a similar protocol to Lieberman-Aiden *et al.* (2009). For genome annotation, we extracted RNA from adult bulked un-winged asexual female individuals and sent it for strand-specific library preparation and sequencing at Novogene (China) following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 26.4 Gb PCR-free Illumina short reads (250 bp PE), 16.6 Gb Oxford Nanopore (ONT) long reads (2 R9.4 flow cells; read length N50 = 21 Kb) and 22.3 Gb Hi-C reads (150 bp PE).

Transcriptome sequence data: 16.7 Gb strand-specific RNA-seq reads (150 bp PE).

Genome assembly

ONT reads were base called using Guppy v2.3.1 (ONT, Oxford, UK) with the parameters "--flowcell FLO-MIN106 --kit SQK-LSK109 --qscore_filtering --min_qscore 7" and used to generate *de novo* assemblies with Flye v2.4 (Kolmogorov *et al.* 2019) using default ONT settings and wtdgb2 v2.3 (Ruan and Li 2019) with the parameters "-x ont -p 19 -k 0 -L 15000 "-x ont -p 19 -k 0 -L 15000". We then merged the Flye and wtdgb2 assemblies with quickmerge v0.3 (Chakraborty *et al.* 2016) using the parameters "-l 1983372 -ml 10000 [Flye_assembly_fasta] [wtdgb2_assembly_fasta]". We set the "-l" flag to the N50 of the least contiguous input assembly (Flye) and used the Flye assembly as the "query" sequence, because preliminary analysis showed it to be more complete than the wtdgb2 assembly. We note that quickmerge assemblies predominantly contain sequence content from the "query" assembly. The quickmerge assembly was polished with three rounds of racon v1.3.1 (Vaser *et al.* 2017) using the ONT long reads followed by three rounds of Pilon v1.22 (Walker *et al.* 2014) using the PCR-free Illumina short reads. Redundant haplotigs were removed from the polished assembly based on self-alignment and coverage estimates from the ONT reads using purge_haplotigs (Roach *et al.* 2018) with the parameters "-l 6 -m 21 -h 70 -a 90". This procedure resulted in a draft assembly containing 733 contigs with a span of 370 Mb (contig n50 = 10.3 Mb, longest contig = 43.8 Mb).

The draft assembly was scaffolded into chromosome-scale super scaffolds using Hi-C data and the 3dDNA assembly pipeline followed by manual curation with JBAT. Initial runs of the 3dDNA pipeline resulted in fragmentation of the input assembly contigs, likely due to the default parameters being optimised for short-read assemblies. Upon experimenting with different values for the "--editor-repeat-coverage" flag, the assembly generated with a value of 6 was carried forward based on assessment of the output contig N50 and inspection of the resulting Hi-C contact map. After manual curation with JBAT, we reviewed the assembly with BlobTools and removed scaffolds with ONT and Illumina short-read coverage less than 10x. We identified and removed a single high coverage (2,743x) and low GC content (17.9%) scaffold, 20.12 Kb in length that aligned to the *M. persicae* mitochondrial genome. No scaffolds that aligned to common aphid symbionts were identified in the assembly.

Summary of frozen assembly and annotation

The final assembly of *A. fabae* JIC1 contains 373 scaffolds, spans 368.3 Mb and contains 96.1% of the assembled content in four chromosome-scale super-scaffolds which correspond to the

expected *A. fabae* karyotype (2n = 8; [Marco *et al.* 2009]). We annotated 22,013 protein-coding genes (24,618 transcripts).

***Aphis gossypii* 1033E**

Sampling and sequencing library preparation

Aphis gossypii individuals were sampled from a clonal lineage (dubbed 1033E) maintained at the JIC insectary on *Gossypium sp.* (cotton). High molecular weight DNA was extracted from a single adult individual following the procedure described in Biello *et al.* (2021) and sent to Novogene (China) for 10x genomics linked-read library preparation and sequencing. As for *A. fabae*, we also sent bulked, snap-frozen, individuals from the same colony to Dovetail Genomics (Santa Cruz, CA) for *in vivo* Hi-C library preparation and sequencing. Hi-C libraries were prepared using the *Dpn*II restriction enzyme following a similar protocol to Lieberman-Aiden *et al.* (2009). For genome annotation, we extracted RNA from adult bulked un-winged asexual female individuals and sent it for strand-specific library preparation and sequencing at Novogene (China) following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 38.7 Gb 10x genomics linked-reads (150 bp PE) and 23.3 Gb Hi-C reads (150 bp PE).

Transcriptome sequence data: 19.2 Gb strand-specific RNA-seq reads (150 bp PE).

Genome assembly

We followed the procedure described in Biello *et al.* (2021) to generate a chromosome-scale genome assembly of *A. gossypii* from 10x Genomics linked-reads and *in vivo* Hi-C data. Briefly, we generated an initial *de novo* assembly with Supernova v2.1.1 (Weisenfeld *et al.* 2017) with the parameter "--maxreads=155760839". To increase contiguity of the Supernova assembly prior to Hi-C scaffolding we carried out two rounds of linked-read scaffolding with scaff10x (https://github.com/wtsi-hpag/Scaff10X) with the parameters "-longread 0 -edge 40000 -block 40000" followed by a single round of misjoin detection and scaffolding with Tigmint (Jackman *et al.* 2018) using default settings. These steps produced a draft assembly spanning 352.4 Mb with a scaffold N50 of 3.5 Mb (contig N50 = 181.9 Kb). We scaffolded the draft assembly into chromosome-scale super scaffolds using our Hi-C data and the 3dDNA assembly pipeline with default settings, followed by manual curation with JBAT. After manual curation we reviewed the assembly with BlobTools and removed scaffolds with 10x genomics linked read coverage less than 20x. The BlobTools review identified 283 short scaffolds belonging to the *Buchnera* endosymbiont which were also removed. Additionally, we identified and removed 4 high coverage (315 – 420x) and low GC content (12 – 18%) scaffolds that aligned to the *M. persicae* mitochondrial genome.

Summary of frozen assembly and annotation

The final assembly of *A. gossypii* 1033E contains 10,850 scaffolds, spans 350.5 Mb and contains 96.1% of the assembled content in four chromosome-scale super-scaffolds which correspond to the expected *A. gossypii* karyotype (2n = 8; [Samkaria *et al.* 2010]). We annotated 25,153 protein-coding genes (27,491 transcripts).

Aphis rumicis

Sampling and sequencing library preparation

Aphis rumicis individuals were sampled from *Rumex obtusifolius* (broad-leaved dock) in Tuxford (Nottinghamshire, UK) in 2018. High molecular weight DNA was extracted from a single adult individual following the procedure described in Biello *et al.* (2021) and sent to Novogene (China) for 10x genomics linked-read library preparation and sequencing. We also extracted high molecular weight DNA from a bulked sample of adult individuals and generated an ONT long-read library following procedures described in Mathers *et al.* (2021). Due to insufficient samples RNA-seq data were not generated for this species.

Sequencing results

Genome sequence data: 33.9 Gb 10x genomics linked-reads (150 bp PE) and 6.1 Gb ONT long reads (1 R9.4 flow cell; read length N50 = 49 Kb).

Genome assembly

Limited availability of samples for this species affected are assembly strategy. We first attempted to assemble the 10x genomics linked reads that were derived from a single aphid DNA extraction. However, this resulted in an incomplete and fragmented assembly, likely caused by a degraded DNA sample (Supernova molecule size estimate ~2.5 Kb). Therefore, the remaining sampled aphids were pooled for bulk DNA extraction to generate a single ONT library which, upon sequencing, produced 6.1 Gb data. ONT reads were base called using Guppy v4.5.3 (ONT, Oxford, UK) with the Rerio model res_dna_r941_min_crf_v032.cfg (https://github.com/nanoporetech/rerio). Despite low coverage (~13x), we attempted *de novo* assembly with Flye v2.8.1 and obtained a highly contiguous assembly (scaffold N50 = 38 Mb) of similar size to the closely related species *A. fabae* (364 Mb vs 368 Mb). Alignment of the *A. rumicis* Flye assembly to our chromosome-scale assembly of *A. fabae* revealed broad agreement with the three autosomes assembled into a small number of long scaffolds, including the assembly of *A. fabae* autosome 3 (scaffold_4) into a single scaffold (**Figure 2**).

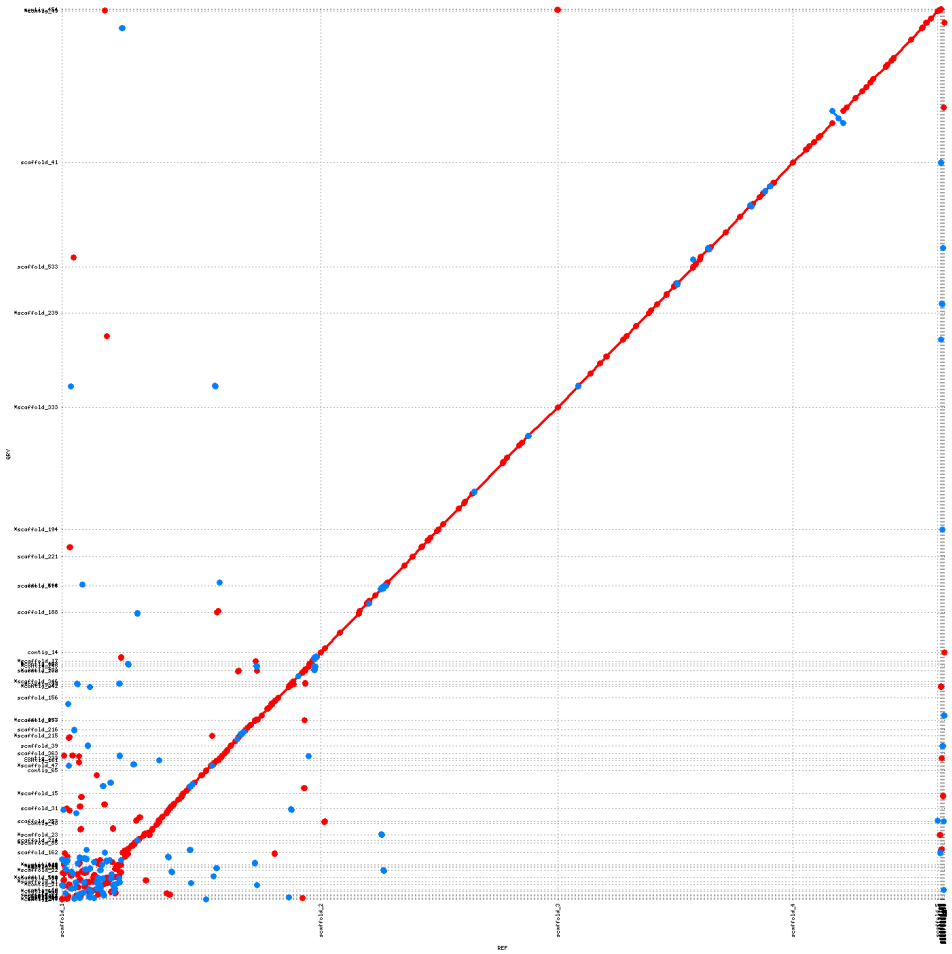


Figure 2: *Dot plot* showing mash map alignment of *A. rumicis* Flye scaffolds (*y* axis) to *A. fabae* chromosome-scale super scaffolds (*x* axis). *A. fabae* chromosomes and scaffolds are ordered from longest to shortest. The first scaffold corresponds to the X chromosome.

The *A. rumicis* Flye assembly was polished with three rounds of Pilon v1.22 using the 10x genomics linked reads. The polished assembly was checked for contamination with BlobTools. We identified and removed three scaffolds belonging to obligate endosymbiont *Buchnera aphidicola* (one 635 Kb chromosome and two plasmids) and three scaffolds belonging to *Serratia symbiotica* (one long 1.72 Mb scaffold and two shorter scaffolds of 96 Kb and 59 Kb). We also removed eight low coverage scaffolds (<15x mean linked read coverage) that were likely contamination and a single high coverage (1,934x mean linked read coverage), low GC content (16%) scaffold that aligned to the *M. persicae* mitochondrial genome.

Summary of frozen assembly and annotation

The final assembly of *A. rumicis* contains 243 scaffolds and spans 362.7 Mb with an N50 length of 38.1 Mb (contig N50 = 6.3 Mb). Alignment of the *A. rumicis* scaffolds to our *A. fabae* chromosome-scale assembly reveals that the three autosomes are contained in between 1 and 5 scaffolds each. This assembly shows that highly contiguous long-read assemblies of *Aphis* species are possible with relatively low coverage (~13x) ONT data. In this instance, the low heterozygosity (see KAT K-mer spectra in assembly release) of the wild-caught isolate likely contributed to the high contiguity of the assembly.

Aphis thalictri

Sampling and sequencing library preparation

Aphis thalictri individuals were sampled from *Thalictrum* 'Anne' (*Thalictrum rochebrunianum* x *Thalictrum flavum* ssp. *Glaucum*; ornamental meadow-rue variety) in Retford (Nottinghamshire, UK) in 2019. As for *A. fabae*, we prepared high molecular weight DNA extractions and two ONT libraries. Additionally, to generate accurate Illumina short reads for assembly polishing, DNA from a single adult asexual female was extracted and sent to Novogene for standard PCR-based Illumina sequencing. To aid genome annotation, RNA from adult bulked un-winged asexual female individuals was extracted and sent to Novogene for strand-specific library preparation and sequencing following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 20.8 Gb Illumina short reads (150 bp PE), 13.2 Gb ONT long reads (2 R9.4 flow cells; read length N50 = 22 Kb).

Transcriptome sequence data: 16.6 Gb strand-specific RNA-seq reads (150 bp PE).

Genome assembly

ONT reads were base called using Guppy v4.5.2 (ONT, Oxford, UK) with the parameters "--flowcell FLO-MIN106 --kit SQK-LSK109". We generated *de novo* ONT assemblies with Flye v2.8.1 and NECAT v0.0.1 (Chen *et al.* 2021) using default settings. The Flye and NECAT assemblies were then merged with quickmerge v0.3 using the parameters "-l 8205039 -ml 10000 [Flye_assembly_fasta] [NECAT_assembly_fasta]". We set the "-l" flag to the N50 of the least contiguous input assembly (NECAT) and used the Flye assembly as the "query" sequence as preliminary analysis showed it to be more complete than the NECAT assembly. Summary statistics calculated with abyss-fac (Jackman *et al.* 2017) for the Flye, NECAT and merged assemblies are given in the table below:

Assembly	# contigs	L50	N50 (Mb)	Longest contig (Mb)	Assembly size (Mb)
flye.fa	437	8	15.60	39.10	363.10
necat.fa	519	13	8.20	32.70	393.80
merged.fasta	387	3	37.30	100.00	363.90

Quickmerge generated a highly contiguous assembly (contig N50 = 37.3 Mb, longest contig > 100 Mb). To validate the quickmerge assembly we aligned it to our new chromosome-scale assembly of the closely related species *A. fabae* using mash map (Jain *et al.* 2018) with the parameters "-f one-to-one -s 50000" (**Figure 3**). This revealed high levels of synteny, supporting the accuracy of the *A. thalictri* quickmerge contigs. Indeed, we find that the third longest scaffold (chromosome) in the *A. fabae* assembly is assembled as a single contig in *A. thalictri* that aligns well along its full length.

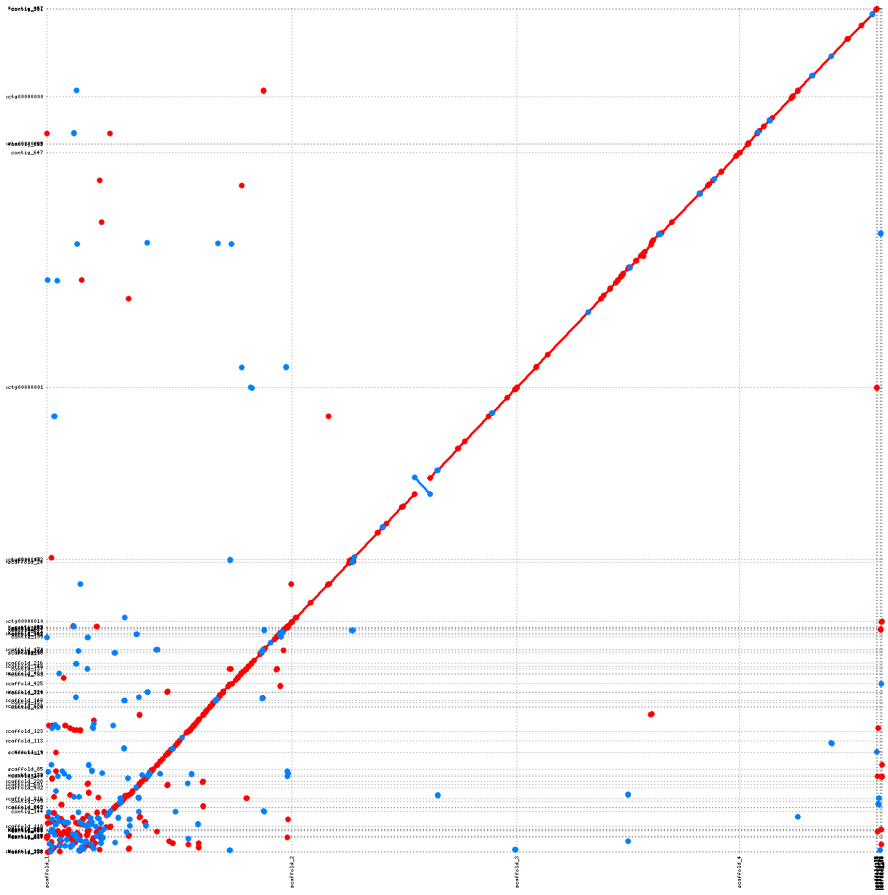


Figure 3: *Dot plot* showing mash map alignment of *A. thalictri* quickmerge contigs (*y* axis) to *A. fabae* chromosome-scale super scaffolds (*x* axis). *A. fabae* chromosomes and scaffolds are ordered from longest to shortest. The first scaffold corresponds to the X chromosome.

Next, we polished the quickmerge assembly with three rounds of Pilon v1.22 using Illumina short reads. The polished assembly was checked for contamination with BlobTools. We identified and removed six contigs belonging to the *Buchnera* endosymbiont and five high coverage (360 – 784x) and low GC content (15 – 18%) contigs that aligned to the *M. persicae* mitochondrial genome. We also removed short low coverage (<10x Illumina) and high GC content (>50%) contigs that likely correspond to contamination.

Summary of frozen assembly and annotation

The final assembly of *A. thalictri* contains 351 contigs and spans 363 Mb with an N50 length of 37 Mb. Alignment of the *A. thalictri* contigs to our *A. fabae* chromosome-scale assembly reveals that the three autosomes are contained in between 1 and 3 contigs each. We annotated 20,364 protein-coding genes (22,591 transcripts).

***Brevicoryne brassicae* JIC1**

Sampling and sequencing library preparation

Brevicoryne brassicae individuals were sampled from a clonal lineage (dubbed JIC1) maintained at the JIC insectary on *Brassica rapa* (Chinese cabbage). Following the procedure described in Mathers *et al.* (2020), DNA was extracted from a single individual and sent to Novogene where an amplification-free Illumina sequencing library was prepared with a target insert size of 500 – 1000 bp. We also extracted high molecular weight DNA from a second individual following the procedure described in Biello *et al.* (2021) and sent it to Novogene where a 10x genomics link-read library was prepared. Additionally, bulked, snap-frozen, individuals from the same colony were sent to Dovetail Genomics (Santa Cruz, CA) for *in vivo* Hi-C library preparation and sequencing. Hi-C libraries were prepared using the *Dpn*II restriction enzyme following a similar protocol to Lieberman-Aiden et al. (2009).

Sequencing results

Genome sequence data: 32.3 Gb PCR-free Illumina short reads (250 bp PE), 37.7 Gb 10x genomics linked-reads (150 bp PE) and 21.1 Gb Hi-C reads (150 bp PE).

Transcriptome sequence data: The previously published high coverage un-stranded RNA-seq library SRR5030749 [Drurey *et al.* 2017]).

Genome assembly

We generated an initial *de novo* assembly of the *B. brassicae* genome using our PCR-free Illumina short reads following the procedure described in Mathers *et al.* (2020). Briefly, we assembled the PCR-free Illumina reads with Discovar *de novo* (https://software.broadinstitute.org/software/discovar/blog/), removed haplotigs based on K-mer analysis as described in Mathers *et al.* (2020) and scaffolded the de-duplicated assembly using our RNA-seq data with P_RNA_scaffodler (Zhu *et al.* 2018). This initial draft assembly was then screened for contamination with BlobTools. We identified 3 scaffolds corresponding to the obligate endosymbiont *Buchnera aphidicola* (one long 643 Kb scaffold flagged as circular by Discovar, and two shorter scaffolds of 7.9 Kb and 3.7 Kb). Additional low coverage scaffolds were identified as probably derived from Poales by BlobTools and were removed as well as other low coverage scaffolds (<18x PCR-free Illumina coverage). Finally, we removed nine high coverage (1342x – 3469x PCR-free Illumina coverage) and low GC content (11% - 20%) scaffolds that aligned to the *M. persicae* mitochondrial genome. Following removal of putative contamination, the assembly spans 355 Mb with a scaffold N50 length of 96 Kb (contig N50 = 40 Kb).

Following the initial assembly based on PCR-free Illumina reads, we also generated 10x Genomics linked reads from a second individual derived from the same JIC insectary colony. We attempted *de novo* assembly of these data with Supernova but this resulted in an

assembly with excessive duplicated content. We therefore used the 10x genomics linked-reads to further scaffold our draft assembly. Two rounds of linked-read scaffolding with scaff10x with the parameters "-longread 0 -edge 40000 -block 40000" were carried out followed by a single round of mis-join detection and scaffolding with Tigmint using default settings. These steps increased scaffold N50 to 725 Kb, with the longest scaffold spanning 6.4 Mb. Finally, this assembly was further scaffolded into chromosome-scale super scaffolds using our Hi-C data (also derived from the same JIC insectary colony) with the 3dDNA assembly pipeline with default settings, followed by manual curation with JBAT.

Summary of frozen assembly and annotation

The final assembly of *B. brassicae* JIC1 contains 15,790 scaffolds, spans 355.8 Mb and contains 87.3% of the assembled content in eight chromosome-scale super-scaffolds which correspond to the expected *B. brassicae* karyotype (2n = 16; [Giannini *et al.* 2003]). We annotated 20,219 protein-coding genes (22,367 transcripts).

Diuraphis noxia* SAM and *Metopolophium dirhodum

Sampling and sequencing library preparation

Diuraphis noxia individuals were sampled from a clonal lineage (dubbed Biotype SAM) maintained at the Stellenbosch University insectary on *Triticum aestivum* (bread wheat cultivar 'TugelaDN'). The colony was originally collected from dryland wheat near Bethlehem in the Free State Province, South Africa, in 2001. *Metopolophium dirhodum* individuals were sampled from a clonal lineage (dubbed UK035) maintained at the JIC insectary on *Avena sativa* (oats). The *M. dirhodum* colony was originally collected from a rose bush in Norwich in 2015. For each species, following the procedure described in Mathers *et al.* (2020), DNA was extracted from a single individual and sent to Novogene (China) where an amplification-free Illumina sequencing library was prepared with a target insert size of 500 – 1000 bp. RNA was extracted from adult bulked un-winged asexual female individuals of *M. dirhodum* and sent for strand-specific library preparation and sequencing at Novogene (China) following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 26.9 Gb and 23.9 GB PCR-free Illumina short reads (250 bp PE) for *D. noxia* SAM and *M. dirhodum*, respectively.

Transcriptome sequence data: For *M. dirhodum*, 15.6 Gb strand-specific RNA-seq reads (150 bp PE). For *D. noxia* SAM, previously published un-stranded RNA-seq data (Nicholson *et al.* 2015).

Genome assembly

We generated *de novo* assemblies of *D. noxia* SAM and *M. dirhodum* using our PCR-free Illumina short reads following the procedure described in Mathers *et al.* (2020) and as described above for *B. brassicae*. For both species, we assembled the obligate endosymbiont *Buchnera aphidicola* into a single scaffold flagged as circular by Discovar *de novo* (636 Kb long for *D. noxia* SAM and 642 Kb long for *M. dirhodum*) plus additional scaffolds corresponding to known *Buchnera* plasmids. For *M. dirhodum*, we also identified scaffolds corresponding to the secondary symbiont *Regiella insecticola*. *M. dirhodum R. insecticola* scaffolds spanned ~2.5 Mb which is similar to the reported genome size of an *R. insecticola* isolate from pea aphid (Nikoh *et al.* 2020), indicating we have likely assembled the complete genome of this bacterium. For both species, we removed symbiont scaffolds, additional low coverage contamination and high coverage, low GC content scaffolds that aligned to the *M. persicae* mitochondrial genome. Following removal of contamination, we carried out additional haplotig removal on both assemblies using purge_dups (Guan *et al.* 2020) with scaffold coverage estimated from mapping our PCR-free Illumina libraries to each species with BWA mem v0.7.7 (Li 2013) and assembly self-alignment with minimap v2.16 (Li 2018) with the parameters "-xasm5 -DP". For *D. noxia* SAM, scaffold coverage cutoffs for purge_dups were estimated automatically using the recommended pipeline. For *M. dirhodum*, we set manual coverage cutoffs with the calcuts script supplied with purge_dups with the parameters "-l 5 -m 25 -u 90 ".

Summary of frozen assembly and annotation

The final assembly of *D. noxia* SAM contains 22,513 scaffolds and spans 344 Mb with a scaffold N50 length of 59 Kb (contig N50 = 32 kb). The final assembly of *M. dirhodum* contains 24,973 scaffolds and spans 387 Mb with a scaffold N50 length of 52 Kb (contig N50 = 36 Kb). In total, we annotated 20,717 genes (23,246) in *D. noxia* SAM and 22,349 genes (24,826) in *M. dirhodum.*

We note that a different isolate of *D. noxia* has previously been assembled using a combination of Illumina short reads and Illumina mate-pair libraries (Nicholson *et al.* 2015). However, this assembly has a low contig N50 (14 Kb) and contains 25% Ns. Furthermore, K-mer analysis vs the Illumina short reads used for the assembly (SRR1214581 - SRR1214584) reveals substantial missing genome content (**Figure 4a**). Although still highly fragmented, our new assembly of *D. noxia* SAM is substantially more contiguous at the contig level and is highly complete (**Figure 4b**).

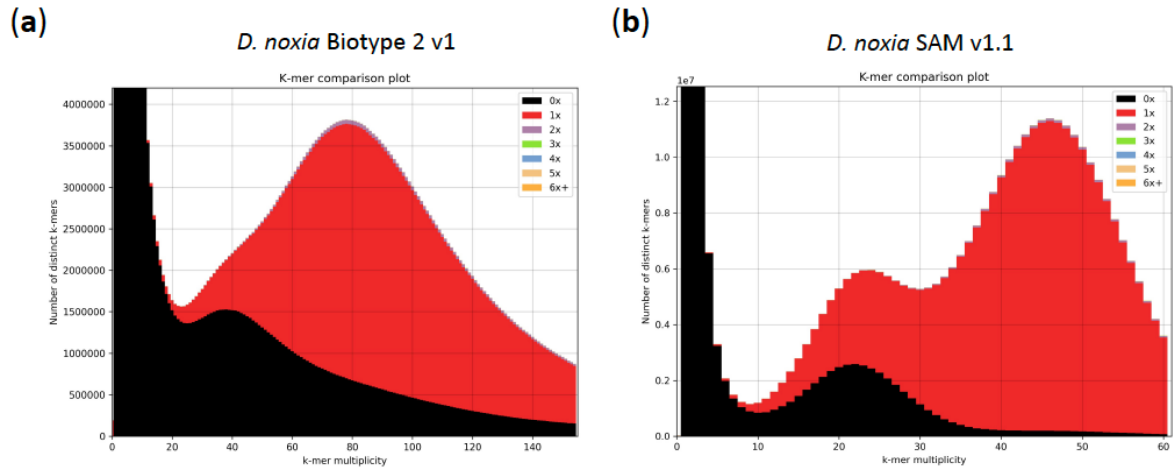


Figure 4: KAT k-mer spectra comparing K-mer content of *D. noxia* genome assemblies to K-mer content of the Illumina short reads used for each assembly. (**a**) *D. noxia* biotype 2 v1 vs SRR1214581 - SRR1214584. (**b**) *D. noxia* SAM v1.1 vs PCR-free Illumina short reads generated for this resource.

Myzus, *Brachycaudus* and *Phorodon* genomes

Sampling and sequencing library preparation

We sampled seven species from the genera *Myzus, Brachycaudus and Phorodon* which are all closely related to the model aphid generalist crop pest *Myzus persicae*. Sampled species are summarised below:

Speiceis	Common name	Sampling location	Plant host
Brachycaudus cardui	Plum-thistle aphid	Norwich, UK	Jacobaea vulgaris
Brachycaudus helichrysi	Leaf-curling plum aphid	Bowthorpe marsh, Norwich, UK	Myosotis scorpioides
Brachycaudus klugkisti		Norwich, UK	Silene dioica
Myzus ligustri	Privet aphid	Norwich, UK	Ligustrum vulgare
Myzus lythri	Purple loosestrife aphid	Bowthorpe marsh, Norwich, UK	Lythrum salicaria
Myzus varians	Peach-clematis aphid	Egmanton, Nottinghamshire, UK	Clematis 'perle d'Azur'
Phorodon humuli	Damson-hop aphid	East Tuddenham, Norfolk, UK	Humulus lupulus

We extracted DNA from single individuals of each species following the procedure described in Mathers *et al.* (2020). Using these DNA extractions, we generated amplification-free Illumina sequencing libraries with a procedure modified for low DNA input and to give an insert size of between 500 bp and 1,000 bp. A minimum of 56 ng of DNA was sheared in a 30 µl volume on a Covaris S2 (Covaris, Massachusetts, USA) for 1 cycle of 45 seconds with a duty cycle of 10%, cycles per burst of 200 and intensity of 5. Post fragmentation, 30 µl 10mM Tris- HCl, pH 8 was added to the DNA, and then 55 µl of this mixture added to 35 µl of Kapa beads (Kapa Biosystems, London, UK) and mixed to precipitate the DNA onto the beads with many molecules <350 bp remaining in solution. DNA was then washed twice with 70% ethanol and then eluted in 25µl Nuclease free water (Qiagen, Manchester, UK). The fragmented and size selected DNA molecules were end repaired and A tailed in 30 µl volume using the Kapa Hyper Prep Kit (Kapa) incubating the reaction at 20 ^{o}C for 30 minutes then 65 ^{o}C for 30 minutes. To the A tailed library molecules 1µl of an appropriate Illumina TruSeq Index adapter

(Illumina, San Diego, USA) diluted 1 in 4 with an oligo annealing buffer (10mM Tris- HCl, 1mM EDTA, 50mM NaCl) was added and mixed then 15 µl of Kapa Hyper Prep ligation buffer, 5 µl of ligase and 4 µl of nuclease free water added and the mixture incubated at 20^{o}C for 20m.

Following adapter ligation, two 1x Kapa bead clean ups were performed to remove any adapter dimer molecules that may have formed during the adapter ligation step with the final library being eluted in 25µl 10mM Tris- HCl, pH 8. Library QC was performed by running a 1µl aliquot on a High Sensitivity BioAnalyser chip (Agilent, Stockport, UK) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher, Cambridge, UK). To determine the number of viable library molecules the library was subjected to quantification by the Kappa qPCR Illumina quantification kit (Kapa) and then sequenced at Earlham Institute (Norwich, UK) on the Hiseq2500 (Illumina) with a 2x250bp read metric loading at 9 pM based on the qPCR quantification and a predicted fragment size of 600 bp.

For *Phorodon humulli*, we also sent bulked, snap-frozen, individuals to Dovetail Genomics (Santa Cruz, CA) for *in vivo* Hi-C library preparation and sequencing. Hi-C libraries were prepared using the *Dpn*II restriction enzyme following a similar protocol to Lieberman-Aiden *et al.* (2009).

For genome annotation, we extracted RNA from adult bulked un-winged asexual female individuals for each species and sent it for un-stranded RNA-seq library preparation and sequencing at Earlham Institute (Norwich, UK) following the procedure described in Mathers *et al.* (2019) with the exception that the libraries were sequenced on an Illumina Hiseq2500 (Illumina) with the 2x125bp read metric.

Sequencing results

Genome sequence data: Between 18.5 and 49.4 Gb PCR-free Illumina short reads (250 bp PE) for each species and 29.4 Gb Hi-C reads for *P. humulli* (150 bp PE).

Transcriptome sequence data: 9.2 - 17.3 Gb un-stranded RNA-seq reads (125 bp PE) for each species.

Full details for each species are given in **Table 2** (see attached files).

Genome assembly

We followed the procedure described in Mathers *et al.* (2020) to generate *de novo* assemblies of each species. Briefly, we assembled the PCR-free Illumina reads with Discovar *de novo* (https://software.broadinstitute.org/software/discovar/blog/), removed haplotigs based on K-mer analysis as described in Mathers *et al.* (2020) and scaffolded the de-duplicated assembly using our RNA-seq data with P_RNA_scaffolder (Zhu *et al.* 2018). The assemblies were then checked for contamination with BlobTools. In all cases we assembled the genome of the obligate endosymbiont *Buchnera aphidicola* into a single scaffold flagged as circular by Discovar *de novo* (635 – 645 Kb long) plus additional scaffolds corresponding to known

Buchnera plasmids. In several cases we also assembled secondary symbiont genomes. For each species we removed symbiont scaffolds, additional low coverage contamination and high coverage, low GC content scaffolds that aligned to the *M. persicae* mitochondrial genome. For *P. humuli*, we further scaffolded the filtered draft assembly into chromosome-scale super scaffolds using *in vivo* Hi-C data.

Summary of frozen assembly and annotation

Assembly and annotation statistics are summarised in the table below. For *P. humuli*, 92% of the assembled genome content is contained in six chromosome-scale super scaffolds that correspond to the expected karyotype based on the closely related species *Phorodon cannabis* (2n =12; [Khagta and Gautam 2016]).

Species	Assembly size (Mb)	Number of scaffolds	Scaffold N50 (Mb)	Contig N50 (Mb)	Number of genes	Number of transcripts
Brachycaudus cardui	347.4	15,055	0.19	0.1	19,443	21,768
Brachycaudus helichrysi	336.1	14,857	0.17	0.11	17,556	19,585
Brachycaudus klugkisti	342.5	15,105	0.1	0.05	22,328	24,739
Myzus ligustri	350	17,355	0.07	0.04	20,046	22,100
Myzus lythri	327.8	10,894	0.23	0.15	16,567	18,796
Myzus varians	328	8,197	0.24	0.12	17,019	19,250
Phorodon humuli	329.9	8,449	59.7	0.11	17,068	19,444

Macrosiphum albifrons

Sampling and sequencing library preparation

Macrosiphum albifrons individuals were sampled from *Lupinus* 'Masterpiece' in Tuxford (Nottinghamshire, UK) in 2019. High molecular weight DNA was extracted from a single adult individual following the procedure described in Biello *et al.* (2021) and used to generate a TELL-seq linked-read library (Chen *et al.* 2020) following the manufacturers protocol (https://www.universalsequencing.com/protocol-gate). We also extracted high molecular weight DNA from bulked samples of individuals from the same colony and prepared four ONT libraries as per *A. fabae* above. Finally, to enable chromosome-scale scaffolding we sent a further bulked sample of individuals from the same colony to Phase Genomics (Seattle, USA) who prepared and sequenced an *in vivo* Hi-C library. For genome annotation, we extracted RNA from adult bulked un-winged asexual female individuals and sent it for strand-specific library preparation and sequencing at Novogene (China) following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 37.2 Gb TELL-seq linked reads (150 bp PE), 22.6 Gb Oxford Nanopore (ONT) long reads (4 R9.4 flow cells; read length N50 = 22 Kb) and 34.4 Gb Hi-C reads (150 bp PE).

Transcriptome sequence data: 16 Gb strand-specific RNA-seq reads (150 bp PE).

Genome assembly

ONT reads were base called using Guppy v4.5.2 (ONT, Oxford, UK) with the parameters "--flowcell FLO-MIN106 --kit SQK-LSK109 --qscore_filtering --min_qscore 7" and used to generate an initial *de novo* assembly with Flye v2.8.1 with default parameters. The Flye assembly was polished with three rounds of Pilon v1.22 using the TELL-seq linked reads. We removed haplotigs from the polished assembly using purge_dups with scaffold coverage estimated from mapping our ONT long reads with minimap v2.16 using the parameter "-x map-ont" and assembly self-alignment with minimap v2.16 (Li 2018) with the parameters "-xasm5 -DP". Scaffold coverage cutoffs for purge_dups were estimated automatically using the recommended pipeline. We then scaffolded the draft assembly into chromosome-scale super scaffolds using our in vivo Hi-C data with the 3dDNA assembly pipeline with default settings, followed by manual curation with JBAT. The assembly was checked for contamination with BlobTools. We identified and removed six scaffolds belonging to obligate endosymbiont *Buchnera aphidicola* (spanning 202 Kb) and single scaffold belonging to the secondary symbiont *Hamiltonella defensa* (spanning 2.2 Mb). We also removed a short (~1 Kb) scaffold that aligns to the *Hamiltonella* bacteria phage APSE-1 (likely a partial assembly) and 29 scaffolds with low coverage (<10x mean linked read coverage) and / or high GC content (>50%). No mitochondria derived scaffolds were identified in the assembly.

Summary of frozen assembly and annotation

The final assembly of *M. albifrons* contains 1,309 scaffolds (scaffold N50 = 86.8 Mb; contig N50 = 1.1 Mb), spans 553.6 Mb and contains 97.6% of the assembled content in five chromosome-scale super-scaffolds which correspond to the expected *M. albifrons* karyotype (2n = 10; [Blackman 1980]). We annotated 31,036 protein-coding genes (33,505 transcripts).

***Rhopalosiphum padi* JIC1 and *Sitobion avenae* multi-species Hi-C assembly**

Sampling and sequencing library preparation

For the chromosome-scale assemblies of *Rhopalosiphum padi* and *Sitobion avenae,* we experimented with using mixed species Hi-C data to reduce library preparation costs. We pooled *R. padi* and *S. avenae* individuals from clonally reproducing colonies maintained at the JIC Insectary and sent the resulting sample to Dovetail Genomics (Santa Cruz, CA) for Hi-C library preparation. As nuclei are cross-linked *in vivo* during Hi-C library preparation species-specific chromatin conformation information is maintained for each species allowing a single library to be used to scaffold multiple species (Marbouty *et al.* 2014). Furthermore, high sequence divergence between *S. avenae* and *R. padi* (Aphidini vs Macrosyphini synonymous site divergence = ~34% [Mathers *et al.* 2021]) minimises the chance of Hi-C reads mismapping between the two species.

For each species we also generated Illumina sequence libraries to generate *de novo* assemblies to scaffold with our multi-species Hi-C data. Procedures described in Biello *et al.*

(2021) were followed by extracting high molecular weight DNA from a single individual and sending this to Novogene for 10x genomics link-read sequencing. Procedures described in Mathers *et al.* (2020) were used for *S. avenae* by extracting DNA from a single individual and sending this to Novogene for amplification-free Illumina sequencing with a library target insert size of 500 – 1000 bp.

For *S. avenae* genome annotation*,* RNA extracted from adult bulked un-winged asexual female individuals was sent for strand-specific library preparation and sequencing at Novogene following the procedure described in Mathers *et al.* (2020).

Sequencing results

Genome sequence data: 20.7 Gb multi-species Hi-C data (150 bp PE); 45.8 Gb 10x genomics linked-reads (150 bp PE) for *R. padi*; 30.1 Gb PCR-free Illumina short reads (250 bp PE) for *S. avenae*.

Transcriptome sequence data: 14.8 Gb strand-specific RNA-seq reads (150 bp PE) for *S. avenae*. Publicly available un-stranded RNA-seq data from Thorpe *et al.* (2016) for *R. padi* genome annotation.

Genome assembly

For *R. padi*, we generated an initial *de novo* assembly with Supernova v2.1.1 with the parameter "--maxreads=143797524". To increase contiguity of the Supernova assembly prior to Hi-C scaffolding we carried out two rounds of linked-read scaffolding with scaff10x with the parameters "-longread 0 -edge 45000 -block 45000" followed by a single round of misjoin detection and scaffolding with Tigmint using default settings. These steps produced a draft assembly spanning 317 Mb with a scaffold N50 of 8.7 Mb (contig N50 = 256 Kb).

For *S. avenae*, we generated an initial *de novo* assembly using our PCR-free Illumina short reads following the procedure described in Mathers *et al.* (2020) and as described above for *B. brassicae*. We screened the short-read assembly for contamination with BlobTools and identified three circular scaffolds corresponding to the chromosome (636 Kb long) and two plasmids of the obligate endosymbiont *Buchnera aphidicola*. We also, removed low coverage contamination (scaffolds < 15x coverage). These steps produced a draft assembly spanning 397.2 Mb with a scaffold N50 of 56 Kb (contig N50 = 37 Kb).

We scaffolded both assemblies using the same multi-species Hi-C library with the 3dDNA assembly pipeline (with default settings) followed by manual curation with JBAT. For *R. padi*, we screened the resulting chromosome-scale assembly for contamination with BlobTools and identified a fragmented assembly (159 short scaffolds) of the obligate endosymbiont *Buchnera* which was removed from the final assembly. We also removed low coverage contamination (< 20x coverage) and two short high coverage, low GC content fragments of the mitochondrial genome. For *S. avenae*, following manual curation of the 3dDNA assembly,

we removed additional duplicated content (haplotigs) from the assembly with the purge_dups pipeline with scaffold coverage estimated from mapping the PCR-free Illumina library with BWA mem v0.7.7 and assembly self-alignment with minimap v2.16 with the parameters "-xasm5 -DP", coverage cutoffs were estimated automatically using the calcuts script.

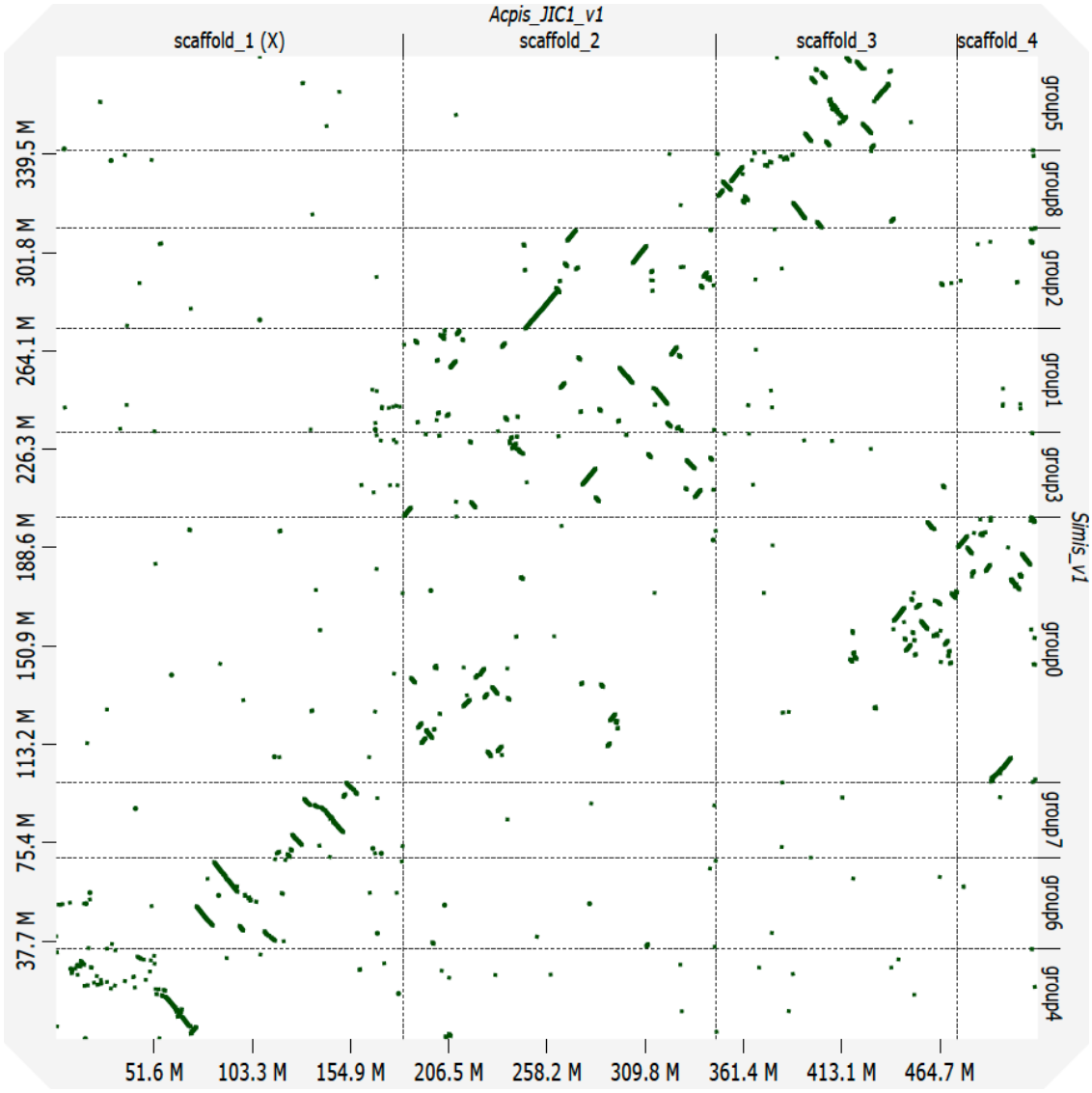
Summary of frozen assemblies and annotations

The final assembly of *R. padi* spans 316 Mb with a scaffold N50 size of 89 Mb (contig N50 = 257. Kb) and contains 95% of the assembled content in four chromosome-scale super scaffolds which correspond to the expected *R. padi* karyotype (2n = 8 [Monti *et al.* 2010]). In total, we annotated 16,977 protein-coding genes (19,137 transcripts).

The final assembly of *S. avenae* spans 369 Mb with a scaffold N50 of 29 Mb (contig N50 = 40 Kb) and contains 74% of the assembled content in nine chromosome-scale super-scaffolds which correspond to the expected *S. avenae* karyotype (2n = 18; [Celis *et al.* 1997]). In total, we annotated 19,919 protein-coding genes (22,368 transcripts).

Reassembly of *Sitobion miscanthi*

We found evidence of substantial scaffolding errors in the published *Sitobion miscanthi* genome. In particular, alignment of the *S. miscanthi* v1 assembly with chromosomes of the closely related model aphid *A. pisum* (Mathers *et al.* 2021) reveals fragmentation of the X chromosome into three chunks as well as substantial rearrangement of the autosomes (**Figure 5**). This is unexpected as we, and others, have recently shown long-term conservation of aphid X chromosome structure across divergent lineages (Li *et al.* 2020; Mathers *et al.* 2021). Visual inspection of the Simis_v1 Hi-C contact map shows multiple off diagonal Hi-C contacts (**Figure 6**) that are indicative of large-scale assembly error within scaffolds (Dudchenko *et al.* 2018; Howe *et al.* 2021). Furthermore, several scaffolds show regions that have very low contact frequencies with adjacent sequence in the scaffold, potentially indicating incorrect assignment of scaffold start and end points. Taken together, these analyses suggest that the assembled chromosomes of Simis_v1 are likely to be inaccurate.



Supplementary Figure 1: *Dot plot* showing whole genome alignment of the *Sitobion miscanthi* v1 (y-axis; Simis_v1) and *Acyrthosiphon pisum* JIC1 v1 (x-axis; Acpis_JIC1_v1) genome assemblies. Simis_v1 scaffolds are ordered along the Acpis_JIC1_v1 assembly. For clarity, only chromosome scale genomic scaffolds are aligned. The x- and y-axis show cumulative scaffold length in Mb. The alignment was carried out with MashMap (Jain *et al.* 2018) and visualised with D-Genies (Cabanettes and Klopp 2018).

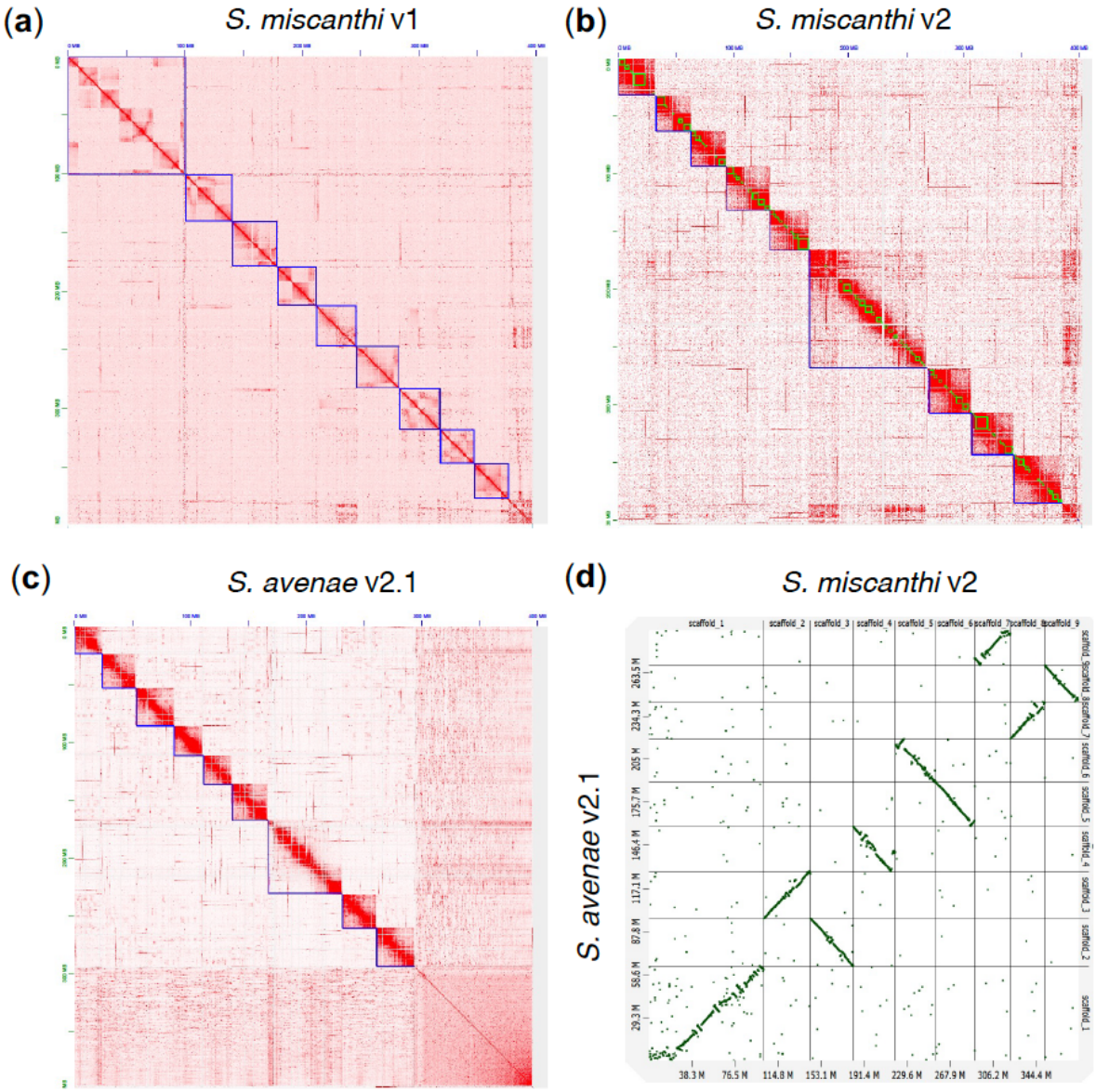


Figure 6: New chromosome-scale assemblies of *S. miscanthi* and *S. avenae* show conserved synteny. (**a**) Hi-C contact map for the *S. miscanthi* v1 genome assembly. Blue lines show chromosome-scale super scaffolds. Genomic scaffolds are ordered from longest to shortest with the x- and y-axis showing cumulative length in millions of base pairs (Mb). (**b**) Hi-C contact map for the *S. miscanthi* v2 genome assembly. Green lines show contigs. (**c**) Hi-C contact map for the *S. avenae* v2.1 genome assembly. (**d**) *Dot plot* showing a MashMap (Jain *et al.* 2018) whole genome alignment between the *S. miscanthi* v2 and *S. avenae* v2.1 genome assemblies. For clarity, only chromosome-scale scaffolds are included. The x- and y-axis show cumulative scaffold length in Mb.

Genome assembly

We reassembled the *S. miscanthi* genome using the published sequence data (Jiang *et al.* 2019). These data include PacBio long-reads (85x coverage), Illumina short-reads (105x coverage) and *in vivo* Hi-C data (76x coverage). We generated *de novo* assemblies of the PacBio long reads with Flye v2.8.1 using default PacBio settings ("--pacbio-raw") and wtdgb2 v2.3 with default parameters. The wtdgb2 assembly was subjected to a single round of long-read polishing with wtpoa-cns based on minimap v2.14 PacBio read alignments carried out

with the parameter "-ax map-pb". The Flye and wtdgb2 assemblies were then merged with quickmerge v0.3 using the parameters "-l 1256119 -ml 10000 [Flye_assembly_fasta] [wtdgb2_assembly_fasta]". The "-l" flag was conservatively set to the N50 of wtdbg2 assembly as the Flye assembly had an N50 below 1 Mb (scaffold N50 = 583 Kb) and low values of "-l" may lead to increased misjoins. We used the Flye assembly as the "query" sequence because preliminary analysis showed it to be more complete than the wtdgb2 assembly and quickmerge assemblies predominantly contain sequence content from the "query" assembly. We then polished the quickmerge assembly with a single round of long-read polishing using the Flye assembly polisher and PacBio reads followed by three rounds of Pilon v1.22 using the Illumina short reads. Redundant haplotigs were removed from the polished assembly using purge_dups. For purge_dups, scaffold coverage was estimated by mapping the PacBio long reads with mininmap v2.16 with the parameter "-x map-pb" and assembly self-alignment was carried out with minimap v2.16 with the parameter "-xasm5 -DP". Scaffold coverage cutoffs for purge_dups were estimated automatically using the calcuts script.

We scaffolded the draft assembly into chromosome-scale super scaffolds using the published Hi-C data with the 3dDNA assembly pipeline with default settings, followed by manual curation with JBAT. We found that the Hi-C library had low resolution resulting in sub-optimal scaffolding performance by 3dDNA. However, 3dDNA first assembles the input assembly into a single super-scaffold before breaking the assembly into putative chromosome-scale fragments. Inspection of the initial round of scaffolding revealed sufficient signal to manually assemble the *S. miscanthi* chromosomes in JBAT (**Figure 6b**). We screened the resulting assembly for contamination with BlobTools and identified and removed two scaffolds belonging to the obligate *Buchnera* endosymbiont and additional scaffolds with low coverage (<30x Illumina short-read converge).

Summary of frozen assembly and annotation

The final assembly of *S. miscanthi* spans 403 Mb with a scaffold N50 of 37.5 Mb (contig N50 = 1.9 Mb) and contains 95% of the assembled content in nine chromosome-scale super-scaffolds which correspond to the expected *S. avenae* karyotype (2n = 18; [Celis *et al.* 1997]). Whole genome alignment of our new *S. miscanthi* assembly and the chromosome-scale assembly of *S. avenae* reveals broad structural agreement (**Figure 6d**). In total, we annotated 21,798 protein-coding genes (23,875 transcripts).

Acknowledgments

We are grateful to Ian Bedford, Victor Soria-Carrasco, Darrell Bean, Mike Darrington, Susannah Gill, William Hampton, Anna Jordan and Jake Stone at the JIC Entomology and Insectary Platform for aphid rearing and sampling. We thank Janet Higgins, Irene Lewis, Mick Mathers and Sandra Mathers for assistance with aphid sampling. Anthony Dixon (University of East Anglia) provided useful advice on locating and identifying aphids from the *Myzus* and *Brachycaudus* genera. This research was supported in part by the NBI Research Computing

Group, which provides technical support and maintenance to the John Innes Centre's high-performance computing cluster and storage systems. We also received technical support from the JIC Informatics team.

Funding Statement

This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) Future Leaders Fellowship to T.C.M. (BB/R01227X/1), Earlham Institute Development grant to T.C.M. (IDG2015-09), the CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS) grant (17.03.2) to S.A.H and BBSRC Industrial Partnership Award (BB/R009481/1) awarded to S.A.H, D.S and C.V.O. R.H.M.W. was funded from the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership Award (BB/M011216/1). Additional support was received from the BBSRC Institute Strategy Programme (BBS/E/J/000PR9798) and the John Innes Foundation.

References

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6: 4–9.

Biello, R., A. Singh, C. J. Godfrey, F. F. Fernández, S. T. Mugford *et al.*, 2021 A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae). Mol. Ecol. Resour. 21: 316–326.

Blackman, R., 1980 Chromosome numbers in the Aphididae and their taxonomic significance. Syst. Entomol. 5: 7–25.

Cabanettes, F., and C. Klopp, 2018 D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. PeerJ 2018:.

Celis, V. E., D. N. Gassen, M. C. Santos-Colares, A. K. Oliveira, and V. L. S. Valente, 1997 Chromosome studies in southern Brazilian wheat pest aphids *Sitobion avenae*, *Schizaphis graminum,* and *Methopolophium dirhodum* (Homoptera: Aphididae). Brazilian J. Genet. 20: 415–419.

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson, 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44: 1–12.

Chen, Y., F. Nie, S. Q. Xie, Y. F. Zheng, Q. Dai *et al.*, 2021 Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. 12: 1–10.

Chen, Z., L. Pham, T. C. Wu, G. Mo, Y. Xia *et al.*, 2020 Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res. 30: 898–909.

Drurey, C., T. C. Mathers, D. C. Prince, C. Wilson, C. Caceres-Moreno *et al.*, 2017 Chemosensory proteins in the CSP4 clade evolved as plant immunity suppressors before two suborders of plant-feeding hemipteran insects diverged. bioRxiv 173278.

Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger *et al.*, 2017 De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science (80-.). 10:.

Dudchenko, O., M. S. Shamim, S. Batra, N. C. Durand, N. T. Musial *et al.*, 2018 The Juicebox

Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797.

Emms, D. M., and S. Kelly, 2019 OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20: 1–14.

Emms, D. M., and S. Kelly, 2015 OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16: 157.

Giannini, S., G. C. Manicardi, D. Bizzaro, and M. Mandrioli, 2003 Cytogenetic analysis on the holocentric chromosomes of the cabbage aphid *Brevicoryne brassicae*. Caryologia 56: 143–147.

Guan, D., D. Guan, S. A. McCarthy, J. Wood, K. Howe *et al.*, 2020 Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36: 2896–2898.

Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2015 BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769.

Hoff, K. J., A. Lomsadze, M. Borodovsky, and M. Stanke, 2019 Whole-genome annotation with BRAKER, pp. 65–95 in *Gene Prediction: Methods and Protocols*, edited by M. Kollmar. Springer New York, New York, NY.

Howe, K., W. Chow, J. Collins, S. Pelan, D. L. Pointon *et al.*, 2021 Significantly improving the quality of genome assemblies through curation. Gigascience 10: 1–9.

Jackman, S. D., L. Coombe, J. Chu, R. L. Warren, B. P. Vandervalk *et al.*, 2018 Tigmint: Correcting assembly errors using linked reads from large molecules. BMC Bioinformatics 19: 1–10.

Jackman, S. D., B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo *et al.*, 2017 ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27: 768–777.

Jain, C., S. Koren, A. Dilthey, A. M. Phillippy, and S. Aluru, 2018 A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics 34: i748–i756.

Jiang, X., Q. Zhang, Y. Qin, H. Yin, S. Zhang *et al.*, 2019 A chromosome-level draft genome of the grain aphid *Sitobion miscanthi*. Gigascience 8: giz101.

Khagta, P., and D. C. Gautam, 2016 Chromosomal studies on six species of crop aphids from Himachal Pradesh, India. Nucl. 59: 137–140.

Kim, D., B. Langmead, and S. L. Salzberg, 2015 HISAT: A fast spliced aligner with low memory requirements. Nat. Methods 12: 357–360.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37: 540–546.

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter, 2013 Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4: 1–12.

Laetsch, D. R., and M. L. Blaxter, 2017 BlobTools: Interrogation of genome assemblies. F1000Research 6: 1287.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv arXiv:1303.3997v2.

Li, H., 2018 Sequence analysis Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, Y., B. Zhang, and N. A. Moran, 2020 The aphid X chromosome is a dangerous place for

functionally important genes: Diverse evolution of hemipteran genomes based on chromosome-level assemblies. Mol. Biol. Evol. 37: 2357–2368.

Lieberman-Aiden, E., N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). 33292: 289–294.

Mapleson, D., G. G. Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, 2017 KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33: 574–576.

Marbouty, M., A. Cournac, J. F. Flot, H. Marie-Nelly, J. Mozziconacci *et al.*, 2014 Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife 3: e03318.

Marco, R., S. Cassanelli, E. Mazzoni, D. Bizzaro, and G. C. Manicardi, 2009 Heterochromatin and rDNA localization on the holocentric chromosomes of black bean aphid, *Aphis fabae* Scop. (Hemiptera: Aphididae). Caryologia 62: 341–346.

Mathers, T. C., 2020 Improved genome assembly and annotation of the soybean aphid (*Aphis glycines* Matsumura). G3 Genes, Genomes, Genet. 10: 899–906.

Mathers, T. C., S. T. Mugford, S. A. Hogenhout, and L. Tripathi, 2020 Genome sequence of the banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts. G3 Genes, Genomes, Genet. 10: 4315–4321.

Mathers, T. C., S. T. Mugford, L. Percival-Alwyn, Y. Chen, G. Kaithakottil *et al.*, 2019 Sex-specific changes in the aphid DNA methylation landscape. Mol. Ecol. 28: 4228–4241.

Mathers, T. C., R. H. M. Wouters, S. T. Mugford, D. Swarbreck, C. van Oosterhout *et al.*, 2021 Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. Mol. Biol. Evol. 38: 856–875.

Monti, V., G. C. Manicardi, and M. Mandrioli, 2010 Distribution and molecular composition of heterochromatin in the holocentric chromosomes of the aphid *Rhopalosiphum padi* (Hemiptera: Aphididae). Genetica 138: 1077–1084.

Nicholson, S. J., M. L. Nickerson, M. Dean, Y. Song, P. R. Hoyt *et al.*, 2015 The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC Genomics 16: 429.

Nikoh, N., T. Tsuchida, R. Koga, K. Oshima, M. Hattori *et al.*, 2020 Genome analysis of "*Candidatus Regiella insecticola*" strain TUt, facultative bacterial symbiont of the pea aphid *Acyrthosiphon pisum*. Microbiol. Resour. Announc. 9: e00598-20.

Price, M. N., P. S. Dehal, and A. P. Arkin, 2009 FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26: 1641–1650.

Price, M. N., P. S. Dehal, and A. P. Arkin, 2010 FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

Roach, M. J., S. A. Schmidt, and A. R. Borneman, 2018 Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19: 1–10.

Ruan, J., and H. Li, 2019 Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17:.

Samkaria, R., J. Bala, and D. C. Gautam, 2010 Karyotype studies on some commonly occurring aphid species. Nucl. 53: 55–59.

Shimodaira, H., and M. Hasegawa, 1999 Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16: 1114–1116.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy

orthologs. Bioinformatics 31: 3210–3212.
Smit, A. F. A., R. Hubley, and P. Green, 2015 RepeatMasker Open-4.0. 2013--2015.
Tarailo-Graovac, M., and N. Chen, 2009 Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma. 1–14.
Thorpe, P., P. J. A. Cock, and J. Bos, 2016 Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC Genomics 17: 172.
Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746.
Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.
Waterhouse, R. M., M. Seppey, F. A. Simao, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35: 543–548.
Weisenfeld, N. I., V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe, 2017 Direct determination of diploid genome sequences. Genome Res. 27: 757–767.
Zhu, B. H., J. Xiao, W. Xue, G. C. Xu, M. Y. Sun *et al.*, 2018 P_RNA_scaffolder: A fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics 19: 1–13.