INTRODUCTION

Transposable elements (TEs) are short stretches of DNA that proliferate within genomes, even if this activity reduces the fitness of the hosts (Hickey, 1982). An unconstrained proliferation of TEs could lead to an accumulation of deleterious TE insertions that may eventually drive host populations to extinction (Brookfield & Badge, 1997; Kofler, 2020). Hence, host organisms evolved elaborate mechanisms...
to combat the spread of TEs (Brennecke et al., 2007; Lewis et al., 2018; Yang et al., 2017). In mammals and invertebrates, the host defence against TEs is based on piRNAs, that is small RNAs with a size between 23 and 29 nt (Brennecke et al., 2007; Gunawardane et al., 2007). These piRNAs bind to PIWI-clade proteins that silence TEs at the transcriptional as well as the post-transcriptional level (Brennecke et al., 2007; Gunawardane et al., 2007; Le Thomas et al., 2013; Sienski et al., 2012). piRNAs are derived from discrete genomic loci, termed piRNA clusters, which may make up substantial portions of genomes (e.g. 3.5% in D. melanogaster). piRNA clusters play a central role in the defence against TE invasions (Bergman et al., 2006; Zanni et al., 2013). Under the currently prevailing view, the trap model, it is assumed that a newly invading TE is stopped when a copy of the TE jumps into a piRNA cluster, which triggers the production of piRNAs that silence the TE (Bergman et al., 2006; Duc et al., 2019; Gorius et al., 2014; Malone & Hannon, 2009; Ozata et al., 2019; Yamanaka et al., 2014; Zanni et al., 2013). Despite the central importance of piRNA clusters in the defence against TEs, the composition and evolution of these regions remains poorly understood. A better understanding of these regions could shed light on important open questions in TE biology, like whether or not the trap model holds (Kofler, 2019, 2020; Mohamed et al., 2020). Our lack of knowledge comes mostly from the fact that piRNA clusters are notoriously difficult to assemble. Most piRNA clusters are located in the heterochromatin and consist of highly repetitive sequences such as TEs (Asif-Laidin et al., 2017; Brennecke et al., 2007; Zanni et al., 2013). Long-read sequencing (e.g. by Pacific Biosciences or Oxford Nanopore Technology) promises to close this gap in our understanding by enabling us to obtain complete assemblies of piRNA clusters. However, it is currently not clear which assembly strategies yield reliable assemblies of piRNA clusters, since many different assembly tools, polishing strategies, sequencing data and scaffolding approaches may be used. In fact, it is not even clear on how to identify the best assemblies, as classic quality metrics such as BUSCO and NG50 are ignorant of TEs and piRNA clusters: BUSCO (Benchmarking Universal Single-Copy Orthologs) provides the fraction of correctly assembled (i.e. ‘complete’) core genes (Simão et al., 2015; Waterhouse et al., 2018) and NG50 gives the size of the smallest contig out of the largest contigs that account for 50% of the reference genome (Earl et al., 2011).

Here, we address these challenges by first introducing several novel quality metrics that assess the number and the quality of the assembled piRNA clusters. Our novel quality metrics were then used to evaluate the effect of different assembly algorithms, polishing approaches, read lengths, coverages, levels of residual polymorphisms and scaffolding methods. Based on these results, we identify strategies that generate high-quality assemblies of piRNA clusters. Using such an optimized assembly strategy, we provide novel assemblies for the Drosophila melanogaster strains Canton-S and Pi2. Additionally, we demonstrate the generality of our approach by extending our metrics to humans and A. thaliana. We provide a user-friendly pipeline, a manual and a walkthrough for assessing the quality of assembled piRNA clusters.

## 2 | MATERIALS AND METHODS

### 2.1 | Sequencing

The D. melanogaster strains Canton-S and Pi2 were obtained from the Bloomington Drosophila Stock Center (BDSC) (Canton-S = 64349; Pi2 = 2384). The reference strain, Iso-1, was kindly provided by Dr. K.A. Senti. We performed Oxford Nanopore Sequencing, Illumina paired-end sequencing and Hi-C for Canton-S and Pi2 (Table S1). The strain Iso-1 was solely sequenced using the Illumina paired-end technology.

High molecular weight DNA for Oxford Nanopore sequencing was extracted from whole bodies of 50 female virgin flies using the Phenol-Chloroform extraction protocol described by Maniatis et al. (1982) using slightly elongated incubation times (5 min). The DNA was sheared to a mean fragment length of 20–30 kb with Covaris g-TUBEs (Covaris Inc., Woburn, MA, USA). The length of the DNA was measured with a TapeStation (4200; DNA ScreenTape, Agilent Technologies). Library preparation was performed with an input of 2–5 μg of sheared DNA following the manufacturer’s protocol (kit LSK108; Oxford Nanopore Technologies; Oxford). About 1–2 μg of the libraries was run for 48–72 hours on MIN106 flow cells. The DNA concentration was measured with a Qubit fluorometer (broad-range DNA assay) (Thermo Fisher Scientific, Waltham, MA, USA), and the purity of the DNA was controlled with NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA).

DNA for Illumina paired-end sequencing was extracted from whole bodies of 20–30 virgin female flies using a salt-extraction protocol (Maniatis et al., 1982). Libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA) using 1 μg DNA. Illumina sequencing was performed by the Vienna Biocenter Core Facilities on a HiSeq2500 platform (2 × 125 bp; Illumina, San Diego, CA, USA).

Hi-C was performed following the Phase Genomics Proximo Hi-C animal Kit (Phase Genomics, Seattle, WA). About 40–50 female third instar larvae were sliced with a razor blade to obtain about 80 mg of tissue. Crosslinking and library preparation were performed according to instructions. Sequencing was performed by the Vienna Biocenter Core Facilities NGS on an Illumina HiSeq2500 platform (2 × 125 bp; Illumina, San Diego, CA, USA).

### 2.2 | Assemblies

Short-read assemblies with Illumina paired-end reads (read length 125 and mean coverage of 30x) were performed with ABySS (Simpson et al., 2009) (version 2.1.5; abyss-pe) using a k-mer size of 96.

Base calling of raw nanopore reads (fast5 format) was performed with either ALbacore (version 2.3.4; Oxford Nanopore Technologies, Oxford, GB) or Guppy (version 2.1.3; Oxford Nanopore Technologies, Oxford, GB). Summary statistics, including mean read length and the total output, were calculated with NANOPlot (De Coster et al., 2018) (version 1.20.1).
De novo assembly of the nanopore reads was performed with four different tools: canu (Koren et al., 2017) (version 1.7), miniasm (Li, 2016) (version 0.3-r179), wtdbg2 (Ruan & Li, 2020) (version 2.4) and flye (Kolmogorov et al., 2019) (version 2.8-b1674). With canu, raw nanopore reads were corrected and trimmed prior to the assembly (preset --nanopore-corrected). To generate assemblies with miniasm, we first aligned all reads against themselves with minimap2 (Li, 2018) (version 2.16-r922) using a preset for nanopore reads (-x ava-ont). We generated the assemblies with miniasm using default settings. The resulting assembly graph files (gfa) were transformed into fasta-files with awk. We launched wtdbg2 with the raw nanopore reads and a nanopore-specific preset ('preset2'). Flye was launched with the raw nanopore reads with the corresponding option (--nano-raw) and default parameters.

Polishing of long-read assemblies was carried out in two steps. We first used racon (Vaser et al., 2017) (version 1.2.1) with the raw nanopore reads mapped to the assembly (minimap2; -ax map-ont; version 2.16-r922 Li, 2018) and then pilon (Walker et al., 2014) (version 1.22) with illumina paired-end reads mapped to the assembly (bwa mem (Li & Durbin, 2009) (version 0.7.17-r1188)). The optimal number of polishing iterations was chosen based on the maximally achieved BUSCO (Benchmarking Universal Single-Copy Orthologs) values (Table S2).

Scaffolding of contigs was done with Hi-C following the SALSA2 protocol (Churye et al., 2019) (version 30. Nov.2018). Briefly, Hi-C reads were mapped to the assembly with bwa mem (Li & Durbin, 2009) (version 0.7.17-r1188), filtered (https://github.com/Arima Genomics), and duplicates were removed (picard-tools; version 2.18.23: https://broadinstitute.github.io/picard/). The mapped reads were then used for scaffolding with SALSA2 using the parameters: diploid mode (-m yes) and restriction enzyme sequence (GATC). An assembly graph was provided. Reference-guided scaffolding was performed with RgGoo (Alone et al., 2019) (version 1.1) based on release 6 of the D. melanogaster reference genome (Hoskin et al., 2015).

Random sampling of reads was performed with seqtk (https://github.com/lh3/seqtk) (version 1.3-r106). To obtain subsets of the longest reads, we sorted all reads by length and then used the appropriate number of the first reads (i.e. the longest reads). Polishing of assemblies generated with subsets of reads was carried out with the respective subsets.

For an overview of our assembly pipeline, see Figure S1.

To visualize assemblies, we generated dotplots using nucmer (Kurtz et al., 2004) (version 3.1). We aligned assemblies to the main chromosome arms (X, 2L, 2R, 3L, 3R and 4) of the D. melanogaster reference ('mumreference'; with parameters -c 1000 -l 100), created coordinate index files using DotPrep.py and visualized genome alignments with dot (https://dnanexus.github.io/dot/).

The final assemblies were based on canu using 100x of the longest reads. Misassemblies were identified based on Hi-C heatmaps and alignments of the assemblies to the reference genome (dotplots). Hi-C heatmaps were generated with juicebox (Durand et al., 2016) (version 1.7.6) using ‘Sau3A1’ as the restriction enzyme. Heatmaps were visualized and analysed with juicebox (1.11.08). Potential misassemblies identified in the Hi-C heatmaps were cross-validated with long reads that were aligned to the assemblies. Breaks in the alignment of the long reads were interpreted as support of an assembly error. Contigs with misassemblies were broken with a custom script ‘introduceBreaks2fasta.py’. Potential contamination (e.g. adaptor sequences) was removed using the standard tools implemented by NCBI.

2.3 | Quality of assemblies

busco (Waterhouse et al., 2018) (version 3.0.2) values were based on the diptera _odb9 data set (2799 genes). quast (Gurevich et al., 2013) (version 5.0.2; quast-lg) was used to compute basic assembly statistics such as NG50 and the total assembly length. As reference, we used the genome of D. melanogaster (release 6).

Computing our TE landscape metrics (abundance of TEs, number of SNPs and internal deletions (IDs) within TEs) requires Illumina raw sequencing reads (expectations) and artificial reads generated from an assembly of interest (observations). We generated artificial reads of length 125 bp starting at each position of the assembly (yielding a uniform distribution; artificial-reads-for-assembly.py). The abundance of TEs, as well as the number of SNPs and internal deletions within TEs, was estimated with Dviate (Weilguny & Kofler, 2019) (version 0.3.6) to obtain both the expected values (Illumina raw reads) and the observed values (artificial reads derived from the assembly). As reference library for Dviate, we used the consensus sequences of TE families present in D. melanogaster (v9.42; we added the sequence of Mariner: M14653) (Quesneville et al., 2005). Solely SNPs and internal deletions with a minimum frequency of 2% were considered. The GC content for each TE was calculated via a custom script (‘GC-content-calculator.py’).

The CUSCO metric relies on the annotation of piRNA clusters of D. melanogaster release 5 (Brennecke et al., 2007; Hoskin et al., 2007). From the 142 annotated piRNA clusters, we excluded clusters that were annotated at the ends of chromosomes (10) and on the highly fragmented U-chromosome (46) (as flanking sequences can not be obtained for these clusters). For the remaining 86 clusters, we identified sequences flanking the clusters at both ends. These flanking sequences were required to align uniquely to release 6 of the D. melanogaster reference genome. For two piRNA clusters that were adjacent to each other (cluster 8 and 9), we could only obtain a pair of sequences flanking both clusters. In summary, we were able to design flanking sequences for 85 piRNA clusters. These sequences had a size between 49 and 12,567 nucleotides. To compute the CUSCO, the flanking sequences were aligned to an assembly using bwa mem (Li & Durbin, 2009). The CUSCO was computed with the script ‘cusco.py’ as the fraction of complete piRNA clusters (i.e. both flanking sequences aligned to the same contig/scaffold). We furthermore distinguished between an ungapped-CUSCO and a gapped-CUSCO based on the presence of poly-N sequences between the two sequences flanking a piRNA cluster. Poly-N tracts in
assemblies were identified using the script ‘find-polyN.py’. To determine whether piRNA clusters are uniquely assembled, we tested if both flanks mapped to multiple contigs/scaffolds using the script ‘multi-cluster.py’.

To identify assembly errors in piRNA clusters, we aligned long reads to assemblies using minimap2 (Li, 2018) (version 2.16-r922). A list of complete BUSCO genes was obtained from the BUSCO pipeline (diptera_odb9; 2799 genes). Based on these data, we computed the base coverage and the soft-clip coverage along each piRNA cluster as well as the 99% quantiles of these coverages (quantiles.py). To calculate the base-coverage quality (CQ), we divided the standard deviation of base coverage in a cluster by the median of standard deviations of complete BUSCO genes. The base coverage and the CQ values were computed with the script ‘cluster-coverage-median.py’ and the parameters --min-mq 15 and --min-len 1000. To calculate the soft-clip quality (ScQ), we divided the average number of soft clipped base pairs in a cluster by the median of the average numbers of soft-clipped bases in complete BUSCO genes. The soft-clip coverage and the ScQ values were computed with the script ‘cluster-softclipcoverage-median.py’ and the parameters --min-mq 15 and --min-len 1000. The script ‘visualize.R’ was used to visualize the base coverage, the soft-clip coverage, the coverage quantiles and locations of assembly gaps (i.e. poly-N sequences) for the piRNA clusters.

2.4 | PCR validation

PCRs were performed at a volume of 20 μl, with 0.05 U/μl of Firepol polymerase (Solis BioDyne, Tartu, Estonia), 2.5 mM MgCl2, 200 μM dNTPs, 0.2 μM primer and 100 ng/μl DNA. See Table S3 for all primer pairs. We used a PCR cycler (Bio Rad CFX Connect, Hercules, CA, USA) with the following program: 5-min. denaturation at 94; 30 cycles of denaturation (30 s at 94), annealing (1 min at 58) and elongation (1 min at 72), followed by 10 min of final extension at 72. The PCR products were loaded on a 1% agarose gel and ran with 120 V for 30 min in TBE buffer. The expected length of amplicons was inferred from the assemblies. Only polymorphic TE insertions for which both breakpoints agreed with the expectations were assumed to be successfully validated.

2.5 | Data analysis

To identify heterozygous SNPs, we aligned Illumina paired-end reads to release 6 of the D. melanogaster genome with bwa mem using default parameters. Reads with a low mapping quality were removed using samtools (version 1.7; Li et al., 2009), a mpileup file was created (samtools), and allele frequency estimates were obtained using mpileup2sync (PoPoolation2; Kofler, Pandey, et al., 2011) with the parameters --fastq-type sanger --min-qual 20. The fraction of heterozygous SNPs for windows of 100 kb was computed with a custom script (polymorphicSNPs_from_sync.py). To account for sequencing errors, we solely classified SNPs with allele frequencies between 0.25 and 0.75 as segregating. Furthermore, a minimum coverage of 10 was required for each site. Windows with insufficient coverage at more than 25% of the sites were excluded. Finally, solely windows with sufficient coverage in all three samples (Pi2, Canton-S and Iso-1) were retained.

To identify the redundant contigs, we chopped assemblies into nonoverlapping fragments of 1 kb using a custom script (chopgenome.py) and aligned them to the release 6 of the D. melanogaster genome using bwa bwasw with default parameters (version 0.7.17-r1188; Li & Durbin, 2009). Ambiguously mapped reads were filtered with samtools (-q 20), and a mpileup file was generated. The mean coverage for 100 kb windows was calculated using a custom script (coverage_from_pileup.py).

We used sniffles (version 1.0.7; Seldazeck, Rescheneder, et al., 2018) to identify structural variants (SVs). Such SVs may either be present or absent in the assembly (classified as deletion and insertion, respectively). We first mapped the long reads to assemblies using ngsmlr (v0.2.7; Seldazeck, Rescheneder, et al., 2018) with the parameter -x ont (ONT data as input). SVs were identified with sniffles using the parameters --report_seq (obtain the sequence of SVs) and --genotype (report allele frequency estimates of SVs). The resulting vcf-file was filtered for SVs with a minimum length of 1kb. To obtain heterozygous SVs, we filtered for allele frequencies between 25% and 75%. To identify SVs caused by TEs, we aligned the sequences of SVs to the consensus sequences of TEs (Quesneville et al., 2005) using blastn (2.7.1+, Altschul et al., 1990).

The composition of piRNA clusters was visualized with easy-fig (v2.2.3 08.11.2016; Sullivan et al., 2011). Annotations of TEs were obtained using RepeatMasker (open-4.0.7; Smit et al., 2015) using the parameters: -no_is (skip checking for bacterial insertions), -now (skip masking low complexity regions) and D. melanogaster TE sequences (Quesneville et al., 2005) or Drosophila TE sequences (Bao et al., 2015). Synteny within piRNA clusters among the assemblies was identified with blastn (2.7.1+, Altschul et al., 1990). To avoid cluttering of the figure, we removed annotations of TEs smaller than 1 kb and blastn similarity blocks smaller than 3 kb. For D. melanogaster, we used assemblies from NCBI with following accession numbers: SIXD01000000 and SISJ02000000 (Ellison & Cao, 2020); bio-project PRJNA418342 (Chakraborty et al., 2019); GCA_002310755.1 and GCA_002310775.1 (Anreiter et al., 2017); JXOZ01000000 (Vicoso & Bachtrog, 2015); LYTF01000000 (Singhal et al., 2017); and JAQD01000000 (McCoy et al., 2014). All statistical analyses were done with R (version 3.4.3) (R Core Team, 2012), and visualizations were performed using the ggplot2 library (Wickham, 2016).

2.6 | Application in different species

To calculate the TE landscape metrics for humans, we used the repetitive sequences library for humans from RepBase (Bao et al., 2015) (version 23.10, humrep.ref) containing 1063 sequences. We
compared a short-read and a long-read-based assembly derived from the same individual (KOREF) (Cho et al., 2016; Kim et al., 2019). To obtain ‘expected’ values, short-read sequences of the KOREF individual were used (SRR2204705) (Cho et al., 2016). To obtain the ‘observed’ values, we created artificial reads for both the short-read (KOREF1.0 (Cho et al., 2016)) and the long-read assembly (KOREF PB 62x (Hi-C scaffolded) (Kim et al., 2019)). The landscape metrics were computed as described before.

To establish flanking sequences for piRNA clusters in humans, we used annotations of 168 piRNA clusters (Sarkar et al., 2014) in the human reference genome hg19. Flanking sequences were created using the scripts ‘flankbeder.sh’ and ‘flankparser.sh’. For each piRNA cluster, the 5 kb regions flanking each cluster were first split into five regions of 1 kb each. Potential flanking sequences containing N’s were removed, and the remaining sequences were aligned back to hg19 using bwa bwasw (version 0.7.17-r1188; Li & Durbin, 2009). The potential flanking sequences were required to align back to the origin (with a tolerance of 5 kb) with a minimum mapping quality (mq) of 5. For each piRNA cluster, the most proximal pair of flanking sequence meeting these criteria was retained (136 out of 168). To calculate the CUSCO values, the flanking sequences were mapped to the respective assemblies using bwa bwasw and CUSCO was calculated as described before.

We calculated CUSCO for 11 assemblies of humans (GRCh37 GCA_000001405.1 (Church et al., 2011); GRCh38.p13 GCA_000001405.28 (Schneider et al., 2017); T2T GCA_009914755.2 (Miga et al., 2020); HG00733_Phased_Diplod GCA_003634875.1; HG00514_premil_3.0 GCA_002180035.3; Ash1.7 GCA_011064465.1 (Shumate et al., 2020); KOREF1.0 GCA_001712695.1 (Cho et al., 2016); and Hi-C scaffolded long-read assemblies of KOREF PB 30x, PB 62x, PT_27x, PT_64x (Kim et al., 2019).

To calculate CQ and ScQ, long reads (SRR9591076) of the KOREF individual were mapped to a short- and long-read assembly (KOREF1.0 (Cho et al., 2016) and KOREF PB 62x (Hi-C scaffolded) (Kim et al., 2019) with minimap2 (Li, 2018) (version 2.17-r941), using a preset for PacBio reads (-ax map-pb). Calculations of CQ and ScQ values were performed as described previously. For all human genomes, BUSCO (Seppey et al., 2019) (version 5.1.2) was computed using vertebrata_odb10.

To identify pairs of sequences flanking KEE (KNOT ENGAGED ELEMENT) regions in A. thaliana, we used the annotations of the 10 KEE regions (Grob et al., 2014) and the reference genome TAIR10. Design of flanking sequences and calculation of CUSCO were performed as described for humans. For all A. thaliana assemblies, BUSCO (version 3.0.2) was computed using embryophyta_odb10. We calculated CUSCO for eight different assemblies (TAIR10 GCA_000001735.1 (Lamesch et al., 2012); AthNd1_v1.0 GCA_001742845.1 (Pucker et al., 2016); AT9943. Cdm-0.scaffold GCA_904420315.1; AT1741.KBS-Mac-74. PacBio GCA_903064285.1; Arabidopsis_thaliana_Ler GCA_902460285.1 (Berardini et al., 2015); ONTmin_IT4 GCA_900303355.1; Ler Assembly GCA_001651475.1 (Zapata et al., 2016); and ASMB83594v1 GCA_000835945.1 (Berlin et al., 2015)).

3 | RESULTS

3.1 | Assembly quality of piRNA clusters

Here, we aim to identify strategies that enable us to generate high-quality assemblies of piRNA clusters. Since commonly used assembly quality metrics, such as BUSCO and NG50 (Earl et al., 2011; Simão et al., 2015), are ignorant of TEs, we first developed several novel quality metrics. Our novel metrics assess whether assemblies (a) accurately reproduce the abundance and diversity of TEs (i.e. the TE landscape) of an organism, (b) have complete piRNA clusters and (c) contain assembly errors within piRNA clusters.

To obtain a data set for demonstrating our novel metrics, we sequenced the D. melanogaster strain Canton-S with (a) the Oxford Nanopore long-read technology (coverage 150x, mean read length = 7 kb), (b) Illumina paired-end sequencing (coverage 30x, read length 125 bp) and (c) Hi-C (coverage 530x, read length 125 bp) (Table S1).

With our first metrics, we tested whether an assembly accurately reproduces the TE content of an organism. A good representation of the TE composition is an important quality control of assemblies and a requirement for an accurate assembly of highly repetitive regions such as piRNA clusters. With these new metrics, we do however not estimate whether TE insertion sites are correct, as this would require knowledge of the true insertion sites in an organism. Instead, we infer summary statistics of the TE landscape by measuring three different features for each TE family: (a) the abundance (in reads per million: rpm), (b) the number of SNPs and (c) the number of internal deletions. A comparison of expected and observed values for these three features allows to estimate the quality of TE representation in an assembly (Figure 1). The key idea is that the expected TE landscape can be directly inferred from the Illumina raw reads without prior need to generate an assembly. We estimate the expected TE landscape with DevaTE (Weilguny & Kofler, 2019), which aligns Illumina reads to the consensus sequences of TEs and provides estimates of the abundance (rpm) and diversity (SNPs and IDs) of each TE family (Figure 1a; Figure S2). For an assembly of interest, we compute the observed TE landscape by generating artificial reads using the assembly as template, which are then used with DevaTE to estimate abundance and diversity of TEs (Figure 1a, Figure S3). To avoid biases and sampling noise, these artificial reads should be uniformly distributed across the assembly and have the same length as the Illumina raw reads used for inferring the expected TE landscape.

To summarize the representation of TEs across all TE families (e.g. 127 TE families in D. melanogaster), we perform a linear regression between the expected and the observed values (Figure 1b; Figure S4). We propose to use the slope of each regression line as a novel quality metric (Figure 1b; Figure S4). This yields, in total, three novel quality metrics (slope of abundance, SNP count and ID count) that estimate how well an assembly captures the TE landscape. High-quality assemblies that accurately reproduce the TE landscape will have regression slopes of <1.0 for each of the three features. Assemblies that overestimate the TE abundance will have a slope
1.0 and assemblies that underestimate the TE abundance a slope <1.0. To illustrate the usage of these metrics, we generated two assemblies of Canton-S: (a) an assembly based on Illumina short reads (abyss; Simpson et al., 2009), and (b) an assembly based on ONT long reads (caNu; Koren et al., 2017) and several rounds of polishing using the long and the short reads (3x PRAIN and 3x PILON; (Vaser et al., 2017; Walker et al., 2014). For both short and long reads, the coverage was 30x. The short-read assembly poorly reproduced the TE landscape (Figure 1b; Figure S4). The abundance of many families was underestimated, and the diversity of many TE families (SNPs and IDs) was overestimated (Figure 1b; Figure S4). By contrast, the long-read assembly captured the TE landscape much more accurately, with most slopes being close to the optimum (i.e. 1; Figure 1b; Figure S4). The high quality of long-read assemblies was also observed with different assembly algorithms (Figure S5) and with unpolished assemblies (Figure S4).

Next, we developed a novel metric that allows us to assess whether piRNA clusters are completely assembled. In essence, the CUSCO value (Cluster BUSCO) estimates the fraction of completely assembled piRNA clusters (Figure 2a). Based on the reference genome of D. melanogaster, we identified pairs of flanking sequences for 85 out of the 142 annotated piRNA clusters of D. melanogaster.
Flanking sequences close to piRNA clusters were preferred. These flanking sequences are then mapped to an assembly of interest. Here, we consider a piRNA cluster to be ‘complete’ (analogous to the BUSCO terminology) when both flanking sequences align to the same contig/scaffold. We thus compute the CUSCO value as the fraction of pairs of flanking sequences aligning to the same contig/scaffold (Figure 2a). Depending on whether or not poly-N sequences (i.e. assembly gaps) are tolerated between the flanking sequences, an ungapped-CUSCO (u.CUSCO) and a gapped-CUSCO (g.CUSCO) can be computed (Figure 2a). Note that the ungapped-CUSCO is a subset of the gapped-CUSCO. The gapped-CUSCO thus includes both piRNA clusters with and without gaps. We ignored the length of piRNA clusters for computing CUSCO values as theoretical work suggests that piRNA clusters could be highly polymorphic: abundant presence/absence polymorphism of TE insertions in piRNA clusters may render the length of the clusters highly variable among individuals (Kelleher et al., 2018; Kofler, 2019). We illustrated the usage of CUSCO with the short- and the long-read assemblies of Canton-S (Figure 1b; Figure S4). CUSCO values differed substantially between short- and long-read assemblies. Although long- and short-read assemblies have similar BUSCO values, CUSCO values differ substantially. (c, d) Assembly errors in complete piRNA clusters may be identified based on (c) base-coverage heterogeneity and (d) elevated numbers of soft-clipped reads. Long reads aligned to a correct and a wrong assembly of a piRNA cluster are shown black. Red indicates not-aligned regions of long reads (i.e. soft-clipped regions). A repeat sequence is shown in blue. Example of an assembled piRNA cluster having a high (e) and low (f) quality. The clusters are from the long-read assembly of Canton-S (30x coverage). Dotted lines show the 99% quantiles of the base coverage and of the soft-clip coverage in BUSCO genes. As a rough summary of the assembly quality for individual piRNA clusters, we compute the CQ (coverage quality) and ScQ values (soft-clip quality)

(Brennecke et al., 2007). Flanking sequences close to piRNA clusters were preferred. These flanking sequences are then mapped to an assembly of interest. Here, we consider a piRNA cluster to be ‘complete’ (analogous to the BUSCO terminology) when both flanking sequences align to the same contig/scaffold. We thus compute the CUSCO value as the fraction of pairs of flanking sequences aligning to the same contig/scaffold (Figure 2a). Depending on whether or not poly-N sequences (i.e. gaps in the assembly) are tolerated between the flanking sequences, an ungapped-CUSCO (i.e. contig-level) and a gapped-CUSCO (i.e. scaffold-level) can be computed (Figure 2a). Note that the ungapped-CUSCO is a subset of the gapped-CUSCO. The gapped-CUSCO thus includes both piRNA clusters with and without gaps. We ignored the length of piRNA clusters for computing CUSCO values as theoretical work suggests that piRNA clusters could be highly polymorphic: abundant presence/absence polymorphism of TE insertions in piRNA clusters may render the length of the clusters highly variable among individuals (Kelleher et al., 2018; Kofler, 2019). We illustrated the usage of CUSCO with the short- and the long-read assemblies of Canton-S (Figure 1b; Figure S4). CUSCO values differed substantially between
the short- and long-read assemblies (Figure 2b). A mere 5.88% of piRNA clusters were complete with the short reads, while 60% were complete with long reads. As we did not perform scaffolding, only the ungapped-CUSCO was calculated (Figure 2b). By contrast, both assemblies show high BUSCO values, which illustrates that BUSCO is of limited use for estimating the suitability of assemblies for an analysis of piRNA clusters (Figure 2b).

However, even when both flanking sequences align to the same contig, a piRNA cluster may still be incorrectly assembled, for example if some internal sequences are missing in the assembly. Therefore, we implemented two additional metrics that allow us to identify assembly errors in complete piRNA clusters (Figure 2c-f). Long reads aligned to an assembly of interest provide two complementary pieces of information that may allow us to identify assembly errors: the base coverage (based on aligned regions of reads) and the soft-clip coverage (based on not-aligned terminal ends of reads). Assembly errors such as repeat collapse or repeat expansions lead to marked differences in the base coverage. For example, a collapsed tandem repeat will result in an elevated coverage. An elevated heterogeneity of the base coverage is thus a hallmark of assembly errors (Figure 2c). However, the base coverage varies in all contigs, including correctly assembled ones. It is therefore necessary to distinguish base-coverage heterogeneity resulting from assembly errors from background heterogeneity. Here, we propose to use the heterogeneity of the base coverage of complete BUSCO genes as null expectation (Simão et al., 2015). BUSCO relies on genes that are conserved within a certain group, for example Diptera and estimates whether these genes are 'complete', 'partial' or 'missing' in an assembly. Complete BUSCO genes provide an ideal estimate of the background heterogeneity of the base coverage based on real data as complete BUSCO genes likely (a) occur as single-copy orthologs in an assembly, and (b) have few assembly errors (since the ORFs are mostly complete). We are thus relating the base-coverage heterogeneity of repetitive heterochromatic sequences (piRNA clusters) to euchromatic, conserved single-copy genes (BUSCO genes). Relying on complete BUSCO genes is however also convenient as BUSCO values are frequently computed for assessing the quality of novel assemblies anyway and a list of complete genes is provided per default by the BUSCO pipeline. We may then visualize the base coverage along piRNA clusters compared to different quantiles of the base coverage of BUSCO genes (e.g. the 99% quantile; Figure 2e,f). These quantiles are the lower and upper boundaries of the base coverage such that a certain fraction (e.g. 99%) of the base coverage of the BUSCO genes are between these boundaries. Base coverage levels exceeding or falling below these quantiles highlight potential assembly problems in piRNA clusters (Figure 2f). As a rough summary of the base-coverage heterogeneity over the entire sequence of a cluster, we may compute the base-coverage heterogeneity (CQ) for each piRNA cluster: $CQ = \frac{\tilde{c}_{\text{busco}}}{c_{\text{cluster}}}$, where $\tilde{c}_{\text{busco}}$ is the median standard deviation of base coverages of BUSCO genes and $c_{\text{cluster}}$ the standard deviation of the base coverage of a given piRNA cluster. We used the median to guard against potential outliers in the base-coverage heterogeneity of BUSCO genes. Low CQ values (<<1.0) indicate a heterogeneous base coverage in piRNA clusters and thus highlight potential assembly problems (Figure 2e,f).

However, some assembly errors, such as deleted or misplaced sequences, might not have noticeable effects on the base coverage. These assembly problems are instead characterized by breaks in the assembly where sequences are joined in the assembly that are not joined in the genome of the organism. As a consequence, many reads spanning these breaks can only be partially aligned back to the assembly (Figure 2d). These reads are usually soft-clipped; that is, a terminal end of a read is either not aligning to any contig or aligning to an entirely different location. Soft-clipped reads can thus be used to identify assembly problems. Therefore, we propose to compute the soft-clip coverage along piRNA clusters as a complementary metric to the base-coverage heterogeneity (Figure 2d). Iterating over all reads, we compute the coverage resulting from the soft-clipped regions of reads; that is, soft-clipped regions are treated as if they were aligned to the reference (Figure 2d). Actually aligned regions of reads are ignored for computing the soft-clip coverage. As null expectation we rely on the soft-clip coverage of complete BUSCO genes. This allows us to visualize the soft-clip coverage along piRNA clusters compared to different quantiles of the soft-clip coverage based on BUSCO genes (e.g. the 99% quantile, Figure 2e,f). Note that solely an upper quantile is computed for the soft-clip coverage (a low soft-clip coverage is ideal), whereas a lower and an upper quantile is computed for the base coverage. A pronounced peak in the soft-clip coverage indicates the likely position of an assembly break (Figure 2f).

The soft-clip quality (ScQ) roughly summarizes the assembly quality of a given piRNA cluster: $\text{ScQ} = \frac{c_{\text{busco}}}{c_{\text{cluster}}}$, where $c_{\text{busco}}$ is the median of the average soft-clip coverages of BUSCO genes and $c_{\text{cluster}}$ the average soft-clip coverage of a given piRNA cluster (Figure 2e,f). Low ScQ values again highlight piRNA clusters that may contain assembly errors. To provide an estimate of quality of an assembly, we can compute the average CQ or ScQ values for all piRNA clusters in an assembly of interest. In summary, the base coverage and the soft-clip coverage can be used to estimate assembly quality at three different levels: (a) to identify errors within a piRNA clusters (e.g. elevated soft-clip coverage at a particular site), (b) to estimate the assembly quality of a particular piRNA cluster (CQ and ScQ) and (c) to estimate the overall assembly quality (average CQ and ScQ). The identification of potential assembly errors in piRNA clusters (a and b) will likely be the main application of these coverage-based metrics.

In summary, we developed novel quality metrics that enable us to estimate the assembly quality of piRNA clusters. First, the TE landscape metrics test whether an assembly accurately reproduces TE abundance and diversity (SNPs and IDs) of an organism. Second, the CUSCO estimates the fraction of complete piRNA clusters. Third, CQ and ScQ values summarize the quality of complete piRNA clusters, where the base-coverage heterogeneity and the soft-clip coverage along piRNA clusters allow us to identify the location of potential assembly problems. We made the scripts for computing our novel quality metrics and for visualizing the quality along piRNA clusters publicly available https://sourceforge.net/projects/cuscoquality/. We additionally provide the sequences flanking piRNA clusters, a manual and a walkthrough.
3.2 Optimizing the assembly strategy

Next, we aimed to identify an assembly strategy that enables us to generate high-quality assemblies of the piRNA clusters of the *D. melanogaster* strain Canton-S. At first, we evaluated the performance of four different long-read assemblers, which rely on slightly different algorithms. *Miniasm* (Li, 2016) uses the overlap among reads to build contiguous sequences. *Canu* (Koren et al., 2017) utilizes a similar approach as *miniasm*. However, to reduce the error rate, *Canu* trims reads and generates consensus sequences of reads prior to the assembly. *Wtdbg2* (Ruan & Li, 2020) uses a de Brujin graph-based assembly algorithm, where k-mers are much larger than for short reads. *Flye* (Kolmogorov et al., 2019) initially generates disjointigs (concatenations of disjoint genomic segments), builds an assembly graph and then uses reads to untangle the assembly graph. *Flye* was designed for an improved assembly of repetitive regions. Long reads usually have high error rates, and assemblies based on these reads may thus also contain an appreciable number of errors (Sović et al., 2016; Vaser et al., 2017). Following recommendations of previous works (Chakraborty et al., 2019; Ellison & Cao, 2020; Solares et al., 2018), we aimed to reduce the error rate by polishing the assembly with *Racon* (long reads) (Vaser et al., 2017) and *Pilon* (short reads) (Walker et al., 2014). Polishing algorithms align reads to an assembly and infer the consensus sequence (Vaser et al., 2017; Walker et al., 2014). Initially, we were concerned that this procedure could eliminate polymorphisms from TE sequences, such that the number of SNPs and IDs of TEs may be underestimated in polished assemblies. However, we found that polished assemblies capture the TE landscape slightly more accurately than unpolished assemblies (TE abundance: unpolished = 1.02, polished = 1.01; SNP metric: unpolished = 0.95, polished = 0.99; ID metric: unpolished = 0.93, polished = 0.99; Tables S4 and S5). Polishing thus enhances the suitability of assemblies for genomic analysis of TEs. We performed one to three rounds of polishing with *Racon* and *Pilon*, where the optimal number of iterations was selected based on the maximally attained BUSCO values (Table S2).

To investigate the influence of coverage on assembly quality, we evaluated the performance of each assembler with several different coverages. Reads were randomly subsampled to coverages ranging from 20 to 150× (Figure 3). Note that a minimum coverage of 20× was required for *Canu* and *Wtdbg2*. To assess the quality of the assemblies, we combined our novel quality metrics with classical metrics (NG50, BUSCO and assembly length) (Figure 3; Table S6). However, we noticed that BUSCO values are very similar among the evaluated coverages and assemblers, suggesting that BUSCO is of limited use for estimating the suitability of assemblies for TE research (Table S6). When considering relevant metrics (TE landscape metrics, NG50, CUSCO, assembly length, CQ and ScQ), we found that the quality of the assembly depends on the coverage but not the assembler (ANOVA comparing linear models; model1: metric, coverage, assembler; model2: metric, coverage; model3: metric, assembler; model1 versus model2 p = .47; model1 versus model3 p = .036). When solely considering NG50 and CUSCO as metrics, the assembler (but not the coverage) had a significant influence on the assembly quality (ANOVA comparing linear models; model1 vs. model2 p = .0004; model1 vs. model3 p = .11). This indicates that the quality of assemblies depends on the assembler and the coverage. Interestingly, the best assemblies were not necessarily obtained when all reads were used (Figure 3). For example, *Canu* and *Flye* yielded the largest NG50 with a coverage of 100× (Canu100x = 8.1 Mbp, Canu150x = 3.6 Mbp, Flye100x = 17.2 Mbp; Flye150x = 10.5 Mbp) and *miniasm* the best representation of TEs at a coverage of 50× (miniasm50x = 1.01, miniasm150x = 1.11). Based on our novel quality metrics (abundance, SNPs, IDs, CUSCO, CQ and ScQ), *Canu* and *miniasm* outperformed *Wtdbg2* and *Flye* at all evaluated coverages (Figure 3; Figure S6). At most coverages, *Canu* captured the TE abundance more accurately than *miniasm*, *Flye* and *Wtdbg2* (Figure 3). Assemblies generated with *Canu* mostly had the highest CUSCO values (Figure 3), where up to 80% of the piRNA clusters were contiguously assembled with coverages ranging from 100× to 150×. Furthermore, *Canu* generated the most reliable assemblies of piRNA clusters (average CQ and ScQ values; Figure S6). Although *Flye* yielded the highest NG50 values, it also generated the shortest assemblies (Figure 3). The *Canu* assemblies were the largest at most coverages and showed intermediate NG50 values (Figure 3). Overall, we conclude that *Canu* yielded the most contiguous (highest ungapped-CUSCO) and the most reliable (highest CQ and ScQ) assemblies of piRNA clusters (Figure 3). For the remainder of this manuscript, we thus relied on assemblies generated with *Canu*.

When reads are randomly sampled, large portions of the data will not be used for the assembly. These unused data may, however, still contain long reads that could be useful for improving the quality of assemblies, for example by bridging gaps between contigs. Thus, we asked if the assembly quality could be further enhanced by sampling the longest reads instead of a random subset. To test this, we sampled subsets of the longest reads with coverages ranging from 20× to 150× (Figure S7). The mean read length of these subsets ranged from 25,051 bp with 20× coverage to 7146 bp with 150× coverage (Figure S7a). *Canu* assemblies based on the longest reads usually have higher NG50 values than assemblies based on random reads (Figure S7b). The largest NG50 values were obtained when a coverage of 100× was used (Figure S7c). Interestingly, CUSCO values were consistently highest for assemblies generated with the longest reads (Figure S7c), while the coverage had little influence on the quality of the assembled piRNA clusters (average CQ and ScQ; Figure S8). The three TE landscape metrics (abundance, SNPs, IDs) revealed little differences between assemblies generated with random reads and the longest reads (Figure S9).

Finally, we were interested in whether CUSCO values could be further improved by using de-novo scaffolding with Hi-C data (Figure S7d). Scaffolding algorithms link contigs into longer sequences based on diverse information such as genetic maps, optical maps or the conformation of chromosomes (Rice & Green, 2018). One widely used approach for scaffolding, Hi-C, relies on the three-dimensional organization of chromosomes (Lieberman-Aiden et al., 2009). With Hi-C, chromatin interactions may be identified by
sequencing fragments that were physically in close proximity (Rice & Green, 2018; Sedlazeck et al., 2018). Since chromatin interactions are most often observed among neighbouring sites within chromosomes, Hi-C data can also be used for scaffolding (Rice & Green, 2018; Sedlazeck, Lee, et al., 2018).

As scaffolds usually contain gaps of unknown size between the contigs (mostly indicated by 100 ‘N’ characters), we calculated the gapped-CUSCO (Figure S7d).

Despite a substantial increase in NG50 values (145–1033%; Figure S10), scaffolding with Hi-C data only moderately improved the CUSCO values (3.5–20%; Figure S7d). This improvement was most pronounced at low coverages, where CUSCO values were quite low before scaffolding. We note that the clusters scaffolded with Hi-C contained gaps, that is missing sequences, mostly of unknown size (see below). Thus, it is crucial to distinguish between gapped- and ungapped-CUSCO to assess the quality of an assembly. As expected, other quality metrics, such as BUSCO and the three TE landscape metrics, were not influenced by Hi-C-based scaffolding (Table S5).

In summary, we found that our novel metrics are useful for assessing the quality of assemblies. Depending on the choice of the investigated regions (number and complexity), CUSCO may be a sensitive metric that identifies quality differences among assemblies not found by other metrics. With long reads and an optimized assembly strategy, up to 81% of the piRNA clusters may be contiguously assembled in D. melanogaster. Especially assemblies based on Canu and a subset of the longest reads (100x coverage) had a high quality. Finally, we found that Hi-C data were of limited use for assembling piRNA clusters.

### 3.3 Influence of segregating polymorphisms on assembly quality

In Figure 3, we showed that long reads enabled us to generate high-quality assemblies of piRNA clusters. However, Canton-S is highly isogenic, having few segregating polymorphisms...
(Figure S11a). We were interested whether piRNA clusters may also be reliably assembled for a less isogenic strain. We relied on the D. melanogaster strain Pi2, which is frequently used in TE research, for example, to assess the extent of P-element (a DNA transposon)-induced infertility in females (O’Hare et al., 1992; O’Hare & Rubin, 1983; Srivastav et al., 2019). Pi2 has substantial numbers of segregating SNPs on several chromosomes (Figure S11a). We first generated a high-quality data set for Pi2: 199x ONT long reads (mean read length = 8 kb), 40x of Illumina PE data and 260x coverage Hi-C data (Table S1). An assembly of Pi2 was generated with our previously established strategy: 100x of the longest ONT reads for heterozygous (0.25 ≤ frequency ≤ 0.75) structural variants to deal with heterozygous TE insertions. We therefore searched that the vast majority of the piRNA clusters lie in regions with few contigs). This absence of redundant clusters is likely due to the fact that the large assembly size of Pi2 (Table S5). Interestingly, we did not observe any redundant assemblies of piRNA clusters for Pi2 (we tested if both sequences flanking piRNA clusters map to multiple contigs). This absence of redundant clusters is likely due to the fact that the vast majority of the piRNA clusters lie in regions with few segregating polymorphisms in Pi2 (Figure S12). This may however not necessarily hold for other strains.

Polymorphic regions are also problematic as it is unclear on how to deal with heterozygous TE insertions. We therefore searched for heterozygous (0.25 ≤ frequency ≤ 0.75) structural variants (SVs) in our assemblies, using SNIFFLES (Sedlazeck, Rescheneder, et al., 2018). In total, we identified 9 heterozygous indels with a minimum size of 1 kb in our Canton-S assembly and 108 in our Pi2 assembly (Figure S11d). A blast search revealed that 66.67% and 84.26% of these SVs in Canton-S and Pi2, respectively, were due to TEs. Two of these heterozygous TE SVs were found in piRNA clusters of Pi2 and none in piRNA clusters of Canton-S. Due to these difficulties, we recommend to use highly isogenic strains for assembling piRNA clusters.

3.4 | Finalizing assemblies

To provide chromosome-scale assemblies of Pi2 and Canton-S to the community, we manually broke up misassemblies (Figure S13) and performed reference-based scaffolding with RaGOO (Alonge et al., 2019). Reference-based scaffolding raised the gapped-CUSCO to 95.3 for Canton-S and to 97.7 for Pi2 but had little effect on other quality metrics (Table S5). An overview of the quality of the final assembly, including the quality at the different assembly steps, can be found in Table S5. The assemblies of Canton-S and Pi2 are available at NCBI (PRJNA618654).

3.5 | Composition of piRNA clusters

Next, we investigated the quality and composition of the assembled piRNA clusters in more detail. We compared piRNA clusters between our chromosome-scale assemblies of Canton-S and Pi2 to the reference genome. Assembly errors, but also presence/absence polymorphism of TE insertions in piRNA clusters, could lead to vast size differences of clusters among assemblies. Thus, we first investigated the length of the piRNA clusters (i.e. the distance between the two sequences flanking each cluster). The length of ungapped clusters in both assemblies is very similar to the length in the reference genome (release 6; paired Wilcoxon rank-sum test; CS: V = 951, p = .55; Pi2: V = 1259.5, p = .45; Figure 4a). Solely 19 clusters in Pi2 and 25 clusters in Canton-S deviated in length by more than 20% from the length of the clusters in the reference genome. Some of this size variation (11 in Pi2 and 11 in Canton-S) was due to clusters with a gap in the assembly (recognized by several ‘N’ characters; Figure 4a, coloured dots). An analysis of gapped clusters revealed a significant length difference in Canton-S, indicating that length estimates of clusters with gaps might not be reliable (paired Wilcoxon rank-sum test; CS: V = 995, p = .007; Pi2: V = 1520.5, p = .51).

When we estimated the quality of the assembled piRNA clusters using CQ and ScQ, we observed considerable differences among the clusters in both assemblies (Figure 4b,c). As expected, piRNA clusters with assembly gaps have low CQ values (Figure 4b). By contrast, assembly gaps had little impact on the ScQ values (Figure 4c). Investigating the base coverage and the soft-clip coverage along each position of some clusters with low and high ScQ values revealed potential assembly issues at some positions of clusters with a low ScQ but not in the clusters with a high ScQ (Figure S14). Taken together, this illustrates that both CQ and ScQ values help to identify clusters with potential assembly issues. However, solely an analysis of the base coverage and the soft-clip coverage along clusters will provide detailed information about the abundance and position of potential assembly problems. For comparing the composition of piRNA clusters, it is therefore necessary to consider the annotations of the clusters as well as the quality along clusters. We illustrate this approach with 42AB, one of the largest contiguously assembled clusters in D. melanogaster (Figure 5). We computed the...
base coverage and the soft-clip coverage for 42AB in Canton-S and Pi2 (Figure 5). We also annotated TEs with **RepeatMasker** (Smit et al., 2015) and identified sequence similarity between our assemblies and the reference genome with **BLAST** (Figure 5; Altschul et al., 1990). The base coverage and the soft-clip coverage of 42AB in Canton-S are mostly within the 99% quantiles of BUSCO genes, which suggests that this assembly is of high-quality (Figure 5). However, the soft-clip coverage and to a lesser extent also the base coverage of 42AB in Pi2 is elevated at the end and in the central simple-repeat region, indicating potential assembly problems in these regions (Figure 5). When searching for causes of these potential assembly problems with **sNiffles**, we found a **P**-element insertion with a frequency of 100% at a site of the elevated soft-clip coverage in Pi2 (Figure 5), demonstrating the utility of our novel quality metrics. We did not find a cause for the elevated soft-clip coverage in the central regions of 42AB in Pi2. Most TE insertions are shared between the three strains, and large synteny blocks, frequently involving several TE insertions, can be found (Figure 5). Nevertheless, we also found differences among the three strains (Figure 5). Most notably, a 26-kb region – involving the X-element, GATE, Max-element and rover – was duplicated in Pi2 (Figure 5). Relative to the reference genome, we also found several TE presence/absence polymorphism in both strains (7 in Pi2 and 11 in Canton-S; Figure 5). Interestingly, most of these polymorphic TEs show little divergence from the consensus sequence (<1%; Figure 5), which suggests that these polymorphisms are due to recent TE insertions into 42AB. These polymorphisms are largely in regions with inconspicuous base coverage and soft-clip coverage, which suggests that they are not due to assembly mistakes. Apart from Chimpo, which was identified using RepBase (Bao et al., 2015), all TEs identified in the cluster 42AB were present in the consensus sequences of TEs in *D. melanogaster* (version 10.01; Quesneville et al., 2005).

Finally, we validated several of the polymorphic TE insertions in piRNA clusters with PCR. In the cluster 42AB, we confirmed 11 out of the 14 tested polymorphic TE insertions, including the missing **P**-element insertion and the large duplication in Pi2 (7 present in Pi2; 3 present in Canton-S; 7 present in Iso-1 of which three are shared with Pi2; Figure S15; Table S3). In other piRNA clusters, we confirmed 20 out of the 22 tested polymorphic TE insertions (12 present in Pi2; 10 present in Canton-S; Figure S15; Table S3).

We conclude that our assembly strategy yields contiguous sequences of many piRNA clusters. Furthermore, our novel quality metrics may be used to identify the location of potential assembly problems in piRNA clusters.
To demonstrate the generality of our approach, we extended our metrics to different species. We first tested the TE landscape metrics with a short- and a long-read based assembly (observations) of the same human individual (Korean reference genome: KOREF1.0 (Cho et al., 2016) and PB_62x (Kim et al., 2019)). The expected TE abundance and diversity was derived from the short-read data (Cho et al., 2016). The TE landscape metrics are based on 1063 TE families. We did not compute the ID metric as solely 39 TE families possessed IDs in the 'expected' data set. Similarly to Drosophila, the long-read assembly of humans captures the abundance and diversity of TEs better than the short-read assembly (abundance: long-read = 0.898, short-read = 0.824; SNPs: long-read = 1.051, short-read = 1.056; Figure S16).

To extend CUSCO to humans, we designed flanking sequences for 168 piRNA clusters (Sarkar et al., 2014) and obtained unique flanking sequences for 136 of them. We applied CUSCO to 11

FIGURE 5 The sequence of the cluster 42AB in our assemblies compared to the reference genome. The TE annotation (yellow-red gradient indicates similarity to the consensus sequence of the TE) and sequence similarity to the reference genome (grey gradient indicates the degree of similarity) are shown. PCR validated presence/absence polymorphisms of TEs or SVs are marked with a ‘*’. A TE insertion missed in the assembly is shown in green. The base coverage and soft-clip coverage are shown for Canton-S (top) and for Pi2 (bottom). The 99% quantiles based on BUSCO genes are shown as dotted lines. Note that the soft-clip coverage and to a lesser extent the coverage is elevated at the site of the missing TE insertion and in the simple-repeat region, indicating possible assembly problems.

3.6 Extending our approach to different species

We did not compute the ID metric as solely 39 TE families possessed IDs in the ‘expected’ data set. Similarly to Drosophila, the long-read assembly of humans captures the abundance and diversity of TEs better than the short-read assembly (abundance: long-read = 0.898, short-read = 0.824; SNPs: long-read = 1.051, short-read = 1.056; Figure S16).

To extend CUSCO to humans, we designed flanking sequences for 168 piRNA clusters (Sarkar et al., 2014) and obtained unique flanking sequences for 136 of them. We applied CUSCO to 11
publicly available human assemblies, where five are different versions of the same Korean individual (Cho et al., 2016; Kim et al., 2019). Although BUSCO values were nearly identical among the assemblies the CUSCO values showed more variation, where especially the ungapped-CUSCO revealed marked differences among the assemblies (Kolmogorov-Smirnoff test; u.CUSCO vs. BUSCO p = .006; Figure 6a). The lowest ungapped-CUSCO value was obtained with the short-read assembly (Figure 6a KOREF1.0; u. CUSCO = 10.29), while all long-read assemblies had markedly higher ungapped-CUSCO values (between 38.24 for HG00733 and 98.53 for T2T). This shows that CUSCO is a sensitive metric in humans, and can be used to identify assemblies with a high fraction of assembled piRNA clusters.

Next, we asked if the base-coverage heterogeneity and the soft-clip coverage can be used in humans to identify clusters with potential assembly errors. We investigated the cluster chr4.117 in different versions of the Korean reference genome (KOREF). This cluster has an apparent polymorphism in the short-read assembly (Figure 6c KOREF1.0). However, the base coverage and soft-clip coverage reveal that this polymorphism is likely an assembly error (Figure 6c). Accordingly, this cluster has high ScQ and CQ values in the long-read assembly (KOREF PB_62x) but low values in the short-read assembly (Figure 6c). We thus argue that the base-coverage heterogeneity and the soft-clip coverage will be useful to identify potential assembly problems in human piRNA clusters.

So far we used the CUSCO solely with piRNA clusters. However, our approach where sequences flanking piRNA clusters are aligned to assemblies can be extended to any regions of interest, such as heterochromatic regions or rDNA clusters. This would also enable extending the CUSCO approach to species not having piRNA clusters such as plants. To test whether our CUSCO approach can be used with such alternative regions, we designed flanking sequences for the 10 KEE regions forming the KNOT region in A. thaliana (Grob et al., 2014). These KEE regions are thought to be involved in control of TEs (Grob et al., 2014). Although the assemblies had similarly high BUSCO values, the CUSCO (i.e. the fraction number of complete KEE regions) differed significantly (Kolmogorov–Smirnoff test; u.CUSCO vs. BUSCO p = .004) among the assemblies (Figure 6c). The short-read assembly (AthNd1_v1.0) again had the lowest CUSCO value (10.0) (Figure 6c). However, the resolution with solely 10 KEE regions is rather coarse as compared to humans (136 clusters) and D. melanogaster (85 clusters). CUSCO values will likely be most informative if they are based on many regions.

In summary, we argue that our quality metrics can be readily extended to diverse species and that CUSCO in particular is a sensitive metric detecting differences in assembly quality that are not easily detected by classic metrics such as BUSCO.

4 | DISCUSSION

Here, we showed that long-read sequencing technologies enable us to generate high-quality assemblies of piRNA clusters. With an optimized assembly strategy, more than 80% of the piRNA clusters in D. melanogaster may be assembled, which can be increased up to 98% with scaffolding approaches.

4.1 | Novel quality metrics

Since current metrics of assembly quality largely ignore TEs and piRNA clusters, we introduced several novel quality metrics.

With three metrics, we first estimate whether an assembly accurately captures the TE landscape (abundance, SNPs and IDs) of an organism. These metrics may be viewed as a general control, since it is unlikely that repeat rich regions, like piRNA clusters, have been accurately assembled when TEs are poorly represented in the assembly. Unfortunately, the real TE landscape is not known for any organism. However, we argue that illumina raw reads may be used to derive a useful approximation of the expected TE landscape. Assuming that reads are more or less randomly distributed over the genome, and that sequencing errors are largely random within reads, this assumption should mostly be valid. Sequencing errors can be largely eliminated from the analysis of the abundance of SNPs and IDs by using a minimum allele frequency (Kofler, Orozco-terWengel, et al., 2011). Here, we used a minimum allele frequency of 2% for SNPs and IDs. In case more stringent criteria are required, a higher threshold may be used. The coverage will fluctuate over the genome, which could affect estimates of TE abundance. Especially, the GC-bias, where regions with a high GC content have an elevated coverage (Minoche et al., 2011), could lead to overestimating the expected abundance of TEs with a high GC content. However, since we sum the average coverage over many different insertions of a TE family, with insertion sites in diverse genomic backgrounds (with varying GC contents), the influence of the GC-bias and of stochastic coverage fluctuations should be minimized by our approach. Furthermore, since we rely on the slope between the expected and the observed TE abundance, which is based on many TE families with different GC contents, the influence of the GC bias should be further reduced. In agreement with this, we did not find any correlation between GC content and TE abundance in the raw reads (Figure S17). Moreover, we solely found a small but nonsignificant difference in GC content among TEs that are well-represented in genomes compared to TEs that are not well-represented (Figure S17). Finally, it is reassuring that an assembly based on long reads captures the expected TE landscape more accurately than an assembly based on the short reads, which have been used for estimating the expected TE landscape (Figure 1).

Apart from sequencing biases, also biases occurring during data analysis, such as mapping and quantifying of reads may occur. Since we use the same pipeline for the raw reads (expectations) and the artificial reads derived from an assembly (observations), these biases should largely be eliminated. Taken together, we think that Illumina raw reads provide a useful approximation of the expected TE landscape. Since computing the TE landscape metrics only requires Illumina short reads for an organism (which are often generated
for the polishing of assemblies anyway) and consensus sequences of TEs, these metrics may thus be used for model and nonmodel organisms (assuming some TE sequences are available or identified de novo).

The CUSCO value estimates the fraction of contiguously assembled piRNA clusters based on an alignment of unique sequences flanking the clusters. piRNA clusters are of central importance for TE biology as they are thought to act as genomic traps that stop TE invasions (Bergman et al., 2006; Duc et al., 2019; Goriaux et al., 2014; Malone & Hannon, 2009; Ozata et al., 2019; Yamanaka et al., 2014; Zanni et al., 2013). However, the CUSCO may generally be a useful metric for assessing the quality of assemblies. An increased CUSCO indicates a more contiguous and thus generally more complete assembly. Furthermore, CUSCO allows us to differentiate

**Figure 6** Extending the quality metrics to different species. (a) CUSCO and BUSCO values of different human assemblies. (b) Coverage heterogeneity and soft-clip coverage for a short- and a long-read assembly of the KOREF individual. Note that our metrics reveal misassemblies at both ends of the short-read assembly. (c) CUSCO and BUSCO values for different A. thaliana assemblies. CUSCO values are based on flanking sequences of the 10 KEE regions. Short-read assemblies are labeled in blue.
between assemblies of very different qualities, as the difficulty of assembling a piRNA cluster varies substantially among the clusters. Long clusters may, for example, be much more challenging to assemble than short ones. This broad range of CUSCO values is demonstrated by our assemblies of Canton-S, where the CUSCO ranges from 5.88% (short reads, ungapped CUSCO), over 81.18% (long reads, ungapped CUSCO) to 95.29% (scaffolding, gapped CUSCO). Also, results in humans and A. thaliana support the broad range and general applicability of CUSCO (Figure 6). Depending on the choice of the repetitive region (number and complexity), CUSCO may thus be a sensitive quality metric capable of differentiating among assemblies of diverse qualities, even when assemblies have a similar quality according to other metrics such as BUSCO (Figure 3; Figure 6; Table S6).

It is important to distinguish between ungapped- and gapped-CUSCO values. Clusters containing gaps likely miss some sequences, including TE insertions, which prevents a comprehensive analysis of the composition of clusters. It is thus most important to maximize ungapped-CUSCO values. However, scaffolding algorithms, which introduce gaps between adjacent contigs, have been used to generate most publicly available assemblies (Figure S18). In these assemblies, many piRNA clusters may contain gaps. To gain a complete picture of piRNA clusters in an assembly, we thus recommend evaluating both CUSCO values (our script computes both).

Identification of the sequences flanking piRNA clusters requires a reference genome. Hence, CUSCO can only be used with species with a reference genome and an annotation of piRNA clusters. But even for species with a reference genome, it will not be feasible to identify suitable flanking sequences for all piRNA clusters (e.g. clusters at terminal ends of contigs/chromosomes).

One limitation of CUSCO is that the sequences flanking piRNA clusters need to be identified for each species separately. However, once sequences flanking piRNA clusters are identified, CUSCO values can be readily computed for many different assemblies (Figure 6; Figure S18). Although we primarily designed CUSCO for species with piRNA clusters, we showed that the CUSCO approach can be extended to any regions of interest such as KEE regions in A. thaliana (Figure 6c).

Since CUSCO ignores the actual sequence within the piRNA clusters, complete clusters may yet contain assembly errors, for example if internal regions are missing in the assembly. Therefore, we suggested that the base-coverage heterogeneity and the soft-clip coverage are useful metrics to identify potential assembly problems in piRNA clusters (Figures 5 and 6b). To derive the null expectations for these two metrics, we relied on complete BUSCO genes. Complete BUSCO genes are ideal for this task: first, BUSCO genes are conserved single copy genes, which makes them relatively easy to assemble, even with short reads and a low coverage (Figure 2b). Second, complete BUSCO genes provide a high-confidence set of genes that contain no or few assembly errors (since the ORFs are mostly complete). Third, BUSCO values are usually computed as a standard metric to assess the quality of novel assemblies. The list of complete BUSCO genes is provided as an output of the BUSCO pipeline. Based on the base coverage and the soft-clip coverage, potential assembly errors in a cluster can be identified by coverage values transgressing the quantiles computed from the BUSCO genes (Figure 5; Figure S14). To roughly summarize the assembly quality of each piRNA cluster with representative numbers, we introduced the ScQ and CQ values.

Although the soft-clip coverage of many piRNA clusters approaches the soft-clip coverage of BUSCO genes, the base-coverage heterogeneity of piRNA clusters is always higher than of BUSCO genes, which explains why the CQ values are usually smaller than ScQ values and rarely approach optimal values (i.e. ≥1.0; Figure 4; Figure S19). Repetitive regions, such as found within piRNA clusters, usually lead to alignment problems that may be responsible for the high base-coverage heterogeneity of piRNA clusters. Computing the base-coverage heterogeneity and the soft-clip coverage along piRNA clusters requires long reads (that are then mapped to the assembly), which are usually available anyway when assembling repetitive regions such as piRNA clusters. Furthermore, the CQ and ScQ values depend on complete BUSCO genes to derive the null expectation. In case few BUSCO genes are assembled (i.e. low BUSCO values), the CQ and ScQ values should be interpreted with caution, that is an assembly with high CQ/ScQ values but a low BUSCO is likely of low quality. This emphasizes that our metrics should not be interpreted in isolation but rather be used in combination with classic metrics such as BUSCO and NG50. However, we think the main use of CQ and ScQ values is to identify clusters with potential assembly problems within a given assembly. Finding such outlier clusters is robust to varying numbers of BUSCO genes as the null expectation for computing CQ and ScQ is identical for all clusters within an assembly.

Our novel quality metrics may not only be used to compare the quality of available assemblies but may also serve as a guide during the assembly procedure, for example, to identify the most suitable assembly algorithm. Our metrics should thus help to generate and to identify assemblies having a high fraction of correctly assembled piRNA clusters (or other regions of interest). In unison with standard assembly metrics such as NG50, BUSCO and the total size of assemblies, our metrics should help to generate and identify assemblies with high contiguity and reliability.

4.2 Assembly strategy

We showed that high-quality assemblies of piRNA clusters can be obtained if: (a) the sequenced strains are isogenic; (b) long reads are available; (c) suitable assemblers, such as Canu are used with an optimized coverage and read length; (d) assemblies are polished using short and long reads; and (e) a scaffolding approach is used. Isogenic strains are necessary to avoid redundant contigs and error-prone assemblies of piRNA clusters (Figure S11; Figure 5). However, it is possible that future tools generate high-quality assemblies of nonisogenic strains. For example, phased assemblers, such as FALCON-D PHASE (Kronenberg et al., 2018), may yield a
separate contig for each homologous chromosome. For these algorithms, segregating polymorphism could even be an advantage as polymorphisms may help to distinguish between homologous chromosomes. We also found that long reads allow us to generate high-quality assemblies of piRNA clusters. Assemblies generated by any of the two major long-read technologies, ONT and PacBio, have a high quality (Figure S18).

We found that Canu yields high-quality assemblies of piRNA clusters and that the TE landscape is most accurately reproduced. The high quality of assemblies generated by Canu was also noticed in several previous works (Jayakumar & Sakakibara, 2017; de Lannoy et al., 2017; Solares et al., 2018; Wick & Holt, 2019). The best assemblies were obtained when solely a subset of the long reads was used for an assembly with Canu, that is 100x coverage with the longest reads. We suspect that this may be related to an algorithmic assumption about the corrected error rate, which is coverage-dependent and governs the overlap among reads (see Canu manual https://canu.readthedocs.io/en/latest/algorithm-reference.html).

Since long reads have a high error rate, polishing of assemblies using long or short reads is usually recommended (Rice & Green, 2018; Sedlazeck, Lee, et al., 2018). Interestingly, polishing also increased the fraction of contiguously assembled piRNA clusters as well as the representation of the TE abundance and diversity (Tables S4 and S5). Scaffolding with Hi-C slightly increased the number of assembled piRNA clusters (using gapped-CUSCO) but had little influence on the representation of the TE landscape (Figure S7 Table S5). Nevertheless, scaffolding approaches may still be useful for TE research, since scaffolding enables generating chromosome-sized sequences, which could be important when the genomic context of a TE insertion is relevant (e.g. whether a TE or piRNA cluster is close to a telomere).

Despite our optimized assembly strategy, about 19% of the piRNA clusters were not contiguously assembled (Table S5, after polishing). Additionally, manual curation of the final assemblies was necessary to avoid misassemblies (Figure S13). This demonstrates that assembly strategies may still be improved. Especially, promising may be further advances in the length of reads (e.g. by improvements in library preparation protocols), their accuracy (e.g. long high-fidelity reads (Wenger et al., 2019)) and in algorithms generating phased assemblies, which could yield a separate contig for each homologous chromosome. Phased assembly algorithms may even allow us to use outbred strains. Furthermore, such phase assemblers avoid the central problem of assemblies of diploid organisms; that is, that two potentially distinct sequences (i.e. the homologous chromosomes) need to be represented as a single one.

Our novel quality metrics may be used to generate high-quality assemblies of piRNA clusters and thus allow us to address some of the central open questions in TE biology, such as the evolutionary dynamics of piRNA clusters.

ACKNOWLEDGEMENTS

We thank Kirsten-André Senti for advice and providing the D. melanogaster strain Iso-1, Christos Vlachos for sharing scripts, Elisabeth Salbaba for technical support and all members of the Institute of Population Genetics for feedback and support. This work was supported by the Austrian Science Foundation (FWF) grants P30036-B25 to RK and W1225.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

RK, FS and FW conceived this work. FS and OC generated the data. FW performed PCR. FS and FW analysed the data. RK and FW provided software. RK, FS and FW wrote the manuscript.

DATA AVAILABILITY STATEMENT

Scripts for computing our quality metrics, including a manual and a walkthrough, are available at https://sourceforge.net/projects/cuscoquality/. We recommend to obtain the scripts via subversion (using the command ‘svn checkout https://svn.code.sf.net/p/cuscoquality/code/cuscoquality/’). The assemblies of Canton-S and Pi2 and the reads are available at NCBI (PRJNA618654). Tables showing the positions of piRNA clusters and the flanking sequences are available at https://sourceforge.net/projects/cuscoquality/files/CUSCO-data/. The positions of piRNA clusters in our assemblies of Canton-S and Pi2 are available at https://sourceforge.net/projects/cuscoquality/files/publicationdata/piRNA-cluster/. All other scripts used in this work are available at https://sourceforge.net/projects/cuscoquality/files/publicationdata/scripts/.

ORCID

Filip Wierzbicki https://orcid.org/0000-0002-6171-2461
Florian Schwarz https://orcid.org/0000-0002-3683-3974
Robert Kofler https://orcid.org/0000-0001-9960-7248

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. Genome Biology, 20(1), 224.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.
Anreiter, I., Kramer, J. M., & Sokolowski, M. B. (2017). Epigenetic mechanisms modulate differences in Drosophila foraging behavior. Proceedings of the National Academy of Sciences of the United States of America, 114(47), 12518–12523.
Asif-Laidin, A., Delmarre, V., Laurentie, J., Miller, W. J., Ronsseray, S., & Teysset, L. (2017). Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in Drosophila. DNA Research, 24(5), 1–14.
Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6(1), 11.
Berardini, T. Z., Reiser, L., Li, D., Mezheriti, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. Genome, 58(8), 474–485.
Bergman, C. M., Quesneville, H., Anxolabéhère, D., & Ashburner, M. (2006). Recurrent insertion and duplication generate networks of
against transposable elements. *Nature Ecology and Evolution*, 2(1), 174–181.

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14), 2103–2110.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMTools. *Bioinformatics*, 25(16), 2078–2079.

Lieberman-Aiden, E., Berkum, N. L. V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, B. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289–293.

Malone, C. D., & Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell*, 138(6), 656–668.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982). *Molecular cloning: A laboratory manual* (Vol. 545). Cold Spring Harbor Laboratory.

McCoy, R. C., Taylor, R. W., Blauwkamp, T. A., Kelley, J. L., Kertesz, M., Ozata, D. M., Gainetdinov, I., Zoch, A., O’Carroll, D., & Zamore, P. D. (2012). Evaluation of the RCH3e and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5), 849–864.

Sedlacek, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6), 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sedlacek, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6), 461–468. https://doi.org/10.1038/s41592-018-0011-7

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, 162, 227–245.

Shumate, A., Zimin, A. V., Sherman, R. M., Puiu, D., Wagner, J. M., Olson, N. D., Pertea, M., Salit, M. L., Zook, J. M., & Salzberg, S. L. (2020). Assembly and annotation of an Ashkenazi human reference genome. *Genome Biology*, 21(1), 1–18.

Sienski, G., Dönnerts, D., & Brennecle, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, 151(5), 946–980.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212.

Simpson, J. T., Wang, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), 1117–1123.

Singhal, K., Khanna, R., & Mohanty, S. (2017). Is Drosophila-microbe interaction species-specific or region specific? A study undertaken involving six Indian Drosophila species. *World Journal of Microbiology and Biotechnology*, 33(6), 103.

Smit, A. F. A., Hubley, R., & Green, P. (2013–2015). RepeatMasker Open-4.0.

Solares, E. A., Chakraborty, M., Miller, D. E., Kalsov, S., Hall, K., Perera, A. G., Emerson, J. J., & Hawley, R. S. (2018). Rapid low-cost assembly of the Drosophila melanogaster reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, 8(10), 3143–3154.

Sović, I., Kržanović, K., Skala, K., & Šikić, M. (2016). Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics*, 32(17), 2582–2589.

Strivastav, S. P., Rahman, R., Ma, Q., Pierre, J., Bandyopadhyay, S., & Lau, N. C. (2019). Har-P, a short P-element variant, weaponizes p-transposase to severely impair Drosophila development. *eLife*, 8, e49948.

Sullivan, M. J., Petty, N. K., & Beaton, S. A. (2011). Easyfig: A genome comparison visualizer. *Bioinformatics*, 27(7), 1009–1010.

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5), 737–746.

Vicoso, B., & Bachtrog, D. (2015). Numerous transitions of sex chromosomes in diptera. *PLoS Biology*, 13(4), e1002078.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelleil, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial
variant detection and genome assembly improvement. PLoS One, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution, 35(3), 543–548.

Weilguny, L., & Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. Molecular Ecology Resources, 19(5), 1346–1354.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, J. R., ... Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology, 37(10), 1155–1162.

Wick, R. R., & Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research, 8, 2138.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer Nature.

Yamanaka, S., Siomi, M. C., & Siomi, H. (2014). piRNA clusters and open chromatin structure. Mobile DNA, 5(1), 22.

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. Trends in Genetics, 33(11), 871–881.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., & Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proceedings of the National Academy of Sciences of the United States of America, 110(49), 19842–19847.

Zapata, L., Ding, J., Willing, E. M., Hartwig, B., Bezdan, D., Jiao, W. B., Patel, V., James, G. V., Koornneef, M., Ossowski, S., & Schneeberger, K. (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proceedings of the National Academy of Sciences of the United States of America, 113(28), E4052–E4060.

### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

---

**How to cite this article:** Wierzbicki, F., Schwarz, F., Cannalonga, O., & Kofler, R. (2022). Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. Molecular Ecology Resources, 22, 102–121. [https://doi.org/10.1111/1755-0998.13455](https://doi.org/10.1111/1755-0998.13455)