International Journal of
Molecular Sciences

MDPI

Article

Comparison of Long-Read Methods for Sequencing and Assembly of Lepidopteran Pest Genomes

Tong Zhang [1,2], Weiqing Xing [1,2], Aoming Wang [1,2], Na Zhang [2], Ling Jia [1,2], Sanyuan Ma [1,2,*] and Qingyou Xia [1,2,*]

[1] State Key Laboratory of Silkworm Genome Biology, Biological Science Research Center, Southwest University, Chongqing 400715, China
[2] Chongqing Key Laboratory of Sericulture Science, Chongqing Engineering and Technology Research Center for Novel Silk Materials, Chongqing 400715, China
* Correspondence: masy@swu.edu.cn (S.M.); xiaqy@swu.edu.cn (Q.X.)

Abstract: Lepidopteran species are mostly pests, causing serious annual economic losses. High-quality genome sequencing and assembly uncover the genetic foundation of pest occurrence and provide guidance for pest control measures. Long-read sequencing technology and assembly algorithm advances have improved the ability to timeously produce high-quality genomes. Lepidoptera includes a wide variety of insects with high genetic diversity and heterozygosity. Therefore, the selection of an appropriate sequencing and assembly strategy to obtain high-quality genomic information is urgently needed. This research used silkworm as a model to test genome sequencing and assembly through high-coverage datasets by de novo assemblies. We report the first nearly complete telomere-to-telomere reference genome of silkworm *Bombyx mori* (P50T strain) produced by Pacific Biosciences (PacBio) HiFi sequencing, and highly contiguous and complete genome assemblies of two other silkworm strains by Oxford Nanopore Technologies (ONT) or PacBio continuous long-reads (CLR) that were unrepresented in the public database. Assembly quality was evaluated by use of BUSCO, Inspector, and EagleC. It is necessary to choose an appropriate assembler for draft genome construction, especially for low-depth datasets. For PacBio CLR and ONT sequencing, NextDenovo is superior. For PacBio HiFi sequencing, hifiasm is better. Quality assessment is essential for genome assembly and can provide better and more accurate results. For chromosome-level high-quality genome construction, we recommend using 3D-DNA with EagleC evaluation. Our study references how to obtain and evaluate high-quality genome assemblies, and is a resource for biological control, comparative genomics, and evolutionary studies of Lepidopteran pests and related species.

Keywords: biological control; de novo assembly; long-read sequencing; benchmarking; lepidopteran pest; assembly evaluation

Citation: Zhang, T.; Xing, W.; Wang, A.; Zhang, N.; Jia, L.; Ma, S.; Xia, Q. Comparison of Long-Read Methods for Sequencing and Assembly of Lepidopteran Pest Genomes. *Int. J. Mol. Sci.* **2023**, *24*, 649. https://doi.org/10.3390/ijms24010649

Academic Editor: Aurel Popa-Wagner

Received: 10 November 2022
Revised: 15 December 2022
Accepted: 24 December 2022
Published: 30 December 2022

1. Introduction

Lepidopteran pests have a critical impact on vegetable crop production, often with a mixture of multiple pests, with many overlapping generations each year, causing huge annual economic losses. Genome sequencing has brought Lepidopteran pest control and genomics research to a new level. Genome sequencing of *Plutella xylostella*, *Hyphantria cunea*, *Cydia pomonella*, and *Helicoverpa zea* revealed the genetics of their invasive populations, explained their host and environmental adaptations at the genetic level, provided partial evidence for the causes of their rapid invasion, and determined potential genetic targets for innovative pest management strategies and the genetic basis of Bt toxin resistance [1–5]. Whole-genome sequencing of twenty *Heliconius* butterflies revealed the complex evolutionary history of the genus, demonstrating that chromosomal structural variation due to gradual penetration is responsible for increased polymorphism in butterfly wings [6].

Since the publication of the first Lepidopteran pest genome [7], sequencing and analytical technologies have developed rapidly. The emergence of innovative subversion techniques, such as short-read sequencing, long-read sequencing, link readout, High-through chromosome conformation capture (Hi-C), optical mapping, and different assembly methods hugely promote genome assembly [8,9]. It is expensive to carry out genome assembly in non-model organisms, and the draft is usually constructed by thousands of fragmented contigs and scaffolds. Nowadays, for many genome projects, achieving high continuous and quality assembly that is close to the chromosome level is a realistic and affordable goal. More than 204 Lepidopteran pests have been sequenced at the nuclear genome level and made publicly available [10]. However, there have been disparate genome sequencing efforts in Lepidopteran pests and many orders remain without genomic representation [11]. With the advent of the pan-genomic era, more Lepidoptera will be sequenced in the future.

The continuity of de novo genome assembly was greatly improved by long-read DNA sequencing platforms, such as single molecular real-time (SMRT) sequencing, Oxford Nanopore Technologies (ONT), and Pacific Biosciences (PacBio) [12]. These technologies overcome the shortcomings of next-generation DNA sequencing (NGS), including information loss, sequence-dependent biases, and relatively short-reads [13]. Previous studies compared the genome assembly tools of ONT sequencing datasets or HiFi sequencing datasets used in Escherichia coli, viruses, pathogens, yeast, fruit flies, and rice, and predominantly used simulated datasets to construct low-quality assembly [14]. In the research of Lepidopteran pests, there has been no large-scale analysis and evaluation of genome assemblers based on high-depth third-generation sequencing (TGS) datasets, and no complete melanosome-to-melanosome genome assembly, which greatly limits the functional research and pest control of lepidopteran insects. Therefore, there remains an urgent need to choose an appropriate sequencing platform, advanced genome assembler, and sequencing depth for the investigation of Lepidopteran pests.

However, genome quality assessment is also very important. Assembly errors are not always apparent and can inadvertently lead to fictitious conclusions [15,16]. The contig and scaffold N50 were used for measuring the fragmentation degree of genome assembly, and Benchmarking Universal Single-Copy Orthologues (BUSCO) is currently used for evaluating the representation of genes [17]. Recently, new methods for assessing the quality of genome assemblies have emerged, such as QUAST-LG, Merqury, KAT, and Inspector [18]. Hi-C technology was used to study three-dimensional (3D) genomic architectures and now has been used for draft genome assembly improvement and chromosome scaffolding in large genomes [19]. Meanwhile, the quality of genome assemblies can be assessed using Hi-C interaction heat maps, the assembly errors usually appear in the chromatin interaction breakpoints. When mapping the Hi-C data to the reference genome, aberrant interaction blocks with different orientations represented different types of assembly errors. However, these methods are achieved by aligning the sequencing reads to contigs. Although Inspector has made improvements in its algorithm to reduce the runtime, it is generally not a particularly rapid method, and there is short of effective tools to accurately evaluate chromosome-level genome assembly, large structural errors in particular.

Bombyx mori (*B. mori*) is a good model for genome assemblies evaluation, as many genome datasets are readily available to benchmark the completeness and accuracy of assemblies [20]. Considering the high genome heterozygosity of field-collected Lepidopteran pests limited by time and space, the genome is at risk of degradation if it cannot be extracted in a timely manner. In this study, we performed 32 (four assemblers on eight subsets with different sequencing depths), 42 (six assemblers on seven subsets with different sequencing depths), and 12 (two assemblers on six subsets with different sequencing depths) de novo assemblies on high-coverage ONT, PacBio continuous long-reads (CLR) and HiFi datasets, respectively. These were performed for three silkworm strains D9L × N4, D9L, and P50T, corresponding to three conditions: highly heterozygous, degradation, and normal. The quality of assembly was evaluated by QUAST [21], BUSCO, Inspector, and the newly proposed EagleC [22] based on deep learning. We assessed the performance of diverse TGS

approaches in *B. mori*, focusing on how to efficiently and accurately construct and evaluate chromosome-level genome assemblies in *B. mori* and other Lepidopteran pests. We believe our results will provide valuable guidance for future Lepidopteran pest genome projects as well as improve previous genome assemblies without generating new sequencing data.

2. Results

2.1. Summary of Raw Data, Assemblies, and Benchmarks

To compare the performance of diverse TGS platforms on constructing highly contiguous genome assembly on Lepidopteran pests. We sequenced and analyzed three long-read datasets for three *B. mori* strains (Table 1 and Figure 1): (1) Silkworm D9L, PacBio CLR reads, 48 Gb data (110×, N50 = 11,722 bp, E-size = 11,909 bp), (2) Silkworm P50T, PacBio HiFi reads, 27 Gb data (60×, N50 = 15,818 bp, E-size = 16,484 bp), and (3) Silkworm D9L × N4, ONT reads, 70 Gb data (160×, N50 = 32,103 bp, E-size = 33,543 bp).

Table 1. Sequence datasets analyzed in this study.

Acronym	Description	Reference
CLR	PacBio Continuous Long Reads; 48 Gb (110× coverage, E-size = 11,909 bp)	This study
ONT	Oxford Nanopore Technologies Reads; 70 Gb (160× coverage, E-size = 33,543 bp)	This study
HIFI	PacBio High Fidelity reads; 27 Gb (60× coverage, E-size = 16,484 bp)	This study
Hi-C	High-throughput Chromosome conformation capture sequencing; used for scaffolding	Lu et al. (2020) [20]

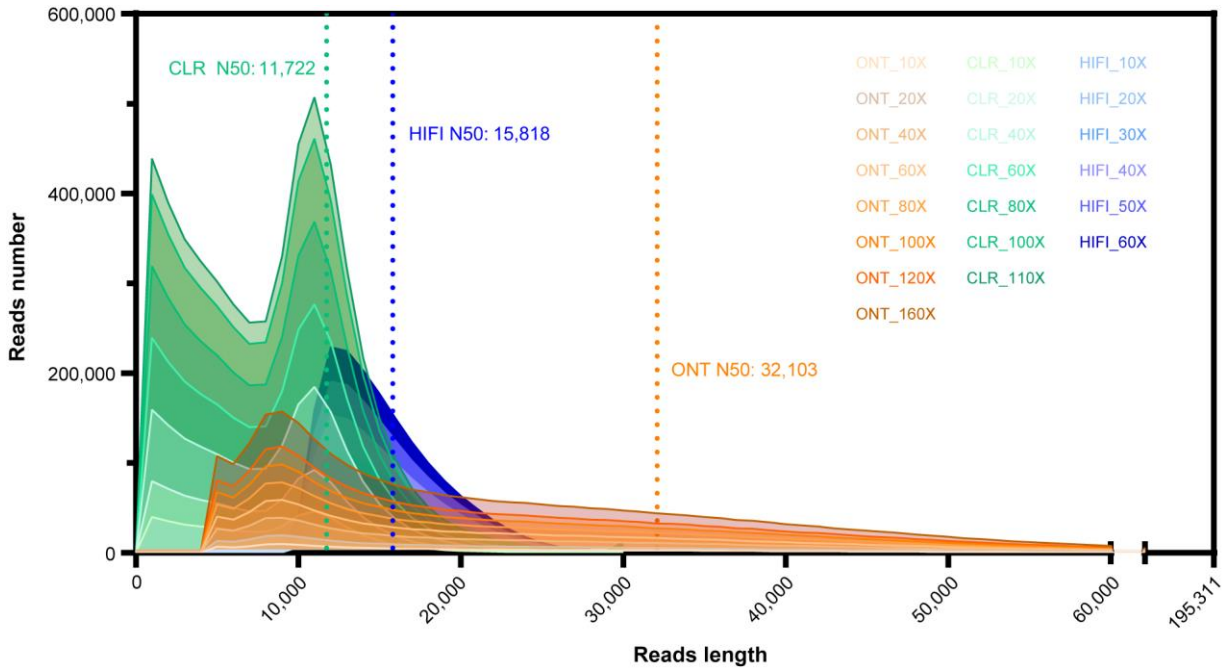

Figure 1. De novo assembly subsets with different data depths.

The genome sequencing datasets were assembled by seven different assembly tools (Figure 2). The CLR reads were assembled by Canu, NextDenovo, MECAT, and wtdbg2. Assemblies of HiFi reads were performed using HiCanu and hifiasm. ONT data were assembled using NextDenovo, NECAT, and wtdbg2.

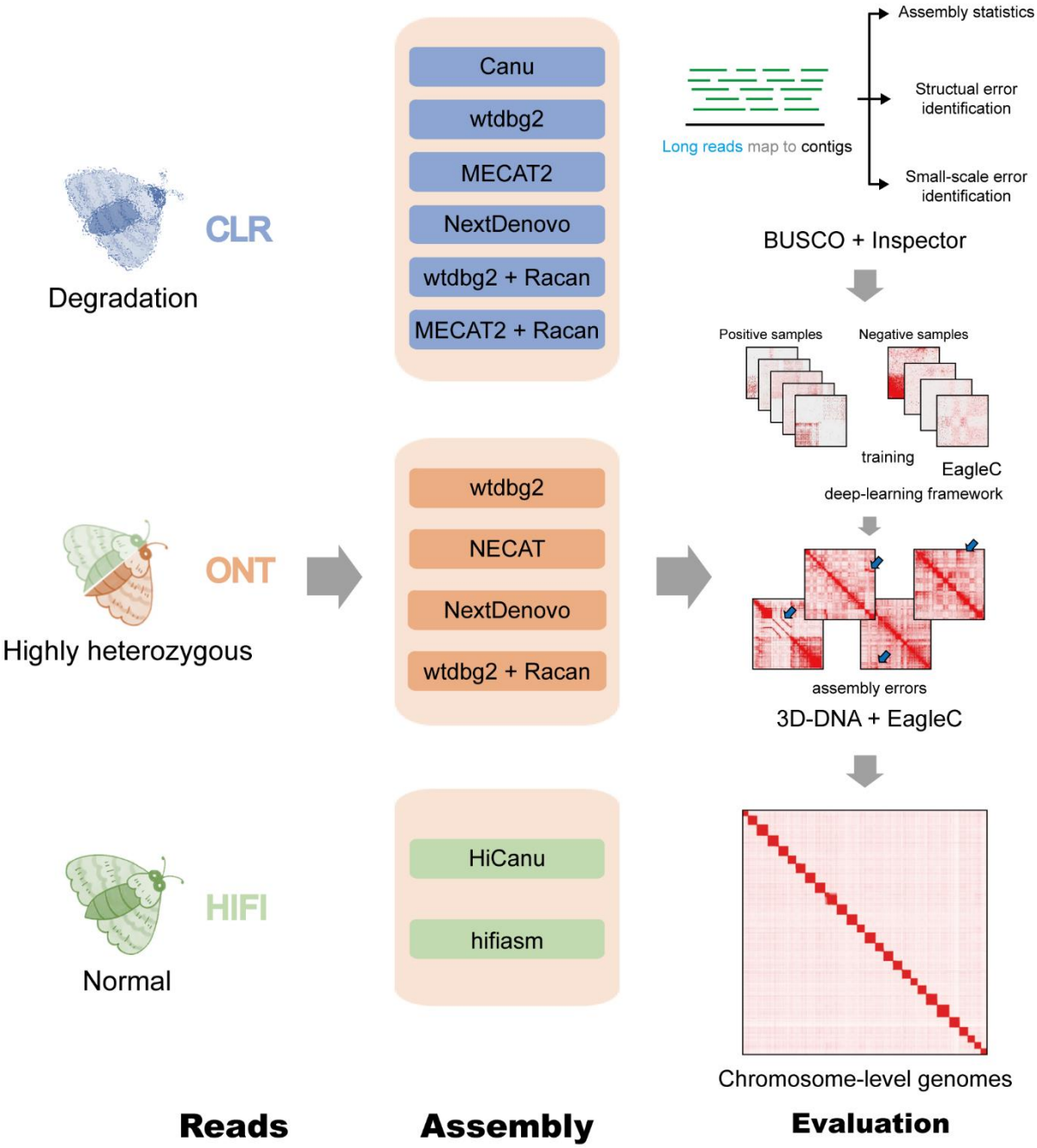

Figure 2. Summary of de novo assembly workflow and evaluation.

The assembly quality was evaluated according to the following six criteria: contig numbers (Contigs), contig N50 (N50) length, number of structural errors (Structural error), small structural errors per Mb (Small-scale error), number of BUSCO complete genes (Complete genes), Quality Value (QV) score and percentage of assembly errors (PAR) identified by EagleC for chromosome-level genomes.

2.2. ONT Genome Assembly

For investigating how to obtain high-quality haploid genome assemblies for field-caught genomically heterozygous Lepidopteran pests, we selected the silkworm D9L × N4 strain (approximately 1.11%, Figure S1a) with high genomic heterozygosity for ONT sequencing and assembly testing. The ONT sequence was assembled using three different long-read assembly tools (NextDenovo, wtdbg2, and NECAT) and eight different

subsets of various coverage (10×, 20×, 40×, 60×, 80×, 100×, 120× and 160×). Detailed statistics for each assembly are shown in Tables 2 and S1.

Table 2. The mean of metric values of different assemblers on CLR, ONT, and HIFI datasets.

	Contigs [a]	Contig N50 (kb)	Structural Error	Small-Scale Error (/Mb)	QV [b]	Complete Genes [c]
CLR-Canu	2387.9	2161	1177.3	515.3	32	989.1
CLR-wtdbg2	2347.3	1236	65.2	2386.8	26.8	883.9
CLR-wtdbg2_polished	2206.6	1246	342	765.7	30.6	1054.1
CLR-MECAT2	1992	2111	1310	5594.9	22.8	618
CLR-MECAT2_polished	1948.5	2106	265.5	490.9	32.4	1182
CLR-NextDenovo	424.2	6463	125	207.4	36.7	1237.5
ONT-wtdbg2	9410.5	442	582.4	4488	22.1	1010.1
ONT-wtdbg2_polished	6526	516	1473.3	788.6	27.4	1207.6
ONT-NECAT	742.8	2656	1939.2	1691	24.3	1264.8
ONT-NextDenovo	103.1	11,454	895.3	1093.5	27.1	1275
HIFI-HiCanu	825.5	11,854	0.3	5.6	66.1	1342.8
HIFI-hifiasm	163.7	13,247	2.7	9.4	50.6	1343.7

[a] contig numbers, [b] Quality Value (QV) score from Inspector, [c] Complete BUSCO genes numbers.

The NextDenovo assemblies were the smallest in size (approximately 449–468 Mb) with contig numbers of approximately 89–114 (Figure 3 and Table S1). NextDenovo generated the most contiguous assemblies (contig N50 approximately 10.0–13.8 Mb), with the highest number of complete (approximately 1181–1298) and single-copy (approximately 1176–1287) BUSCO genes. The wtdbg2 assemblies were the largest in size (approximately 452–794 Mb) and produced contig numbers of approximately 3273–13,714. Wtdbg2 generated the least contiguous assemblies (contig N50 0.15–0.81 Mb), and the lowest number of complete (669–1129) and single-copy (668–1083) BUSCO genes. The assembly quality of wtdbg2 for genomes with high heterozygosity was less satisfactory, but it is the only software that can generate assembly at 10× sequencing depth. The assembly quality of NECAT was between those of NextDenovo and wtdbg2. The NECAT assemblies' sizes were approximately 561–581 Mb, contig numbers were approximately 688–851, contig N50 were between 2.44 and 2.88 Mb, and complete BUSCO genes were between 1253 and 1272. Additionally, we analyzed the computational time of those assemblers and found that the wtdbg2 was the fastest assembler, followed by NextDenovo and NECAT (Figure 4b), saving between a third and a half of the time when the sequencing depth was greater than 80×.

To estimate the genome assembly accuracy, we calculated the number of Structural errors and Small-scale errors using Inspector. NextDenovo had the lowest number of Small-scale errors, and Structural error numbers just below those of wtdbg2 (Figures 5 and S2). Wtdbg2 had the highest number of Small-scale errors, while the lowest number of Structural errors. NECAT had the highest number of Structural errors and the second highest of Small-scale errors.

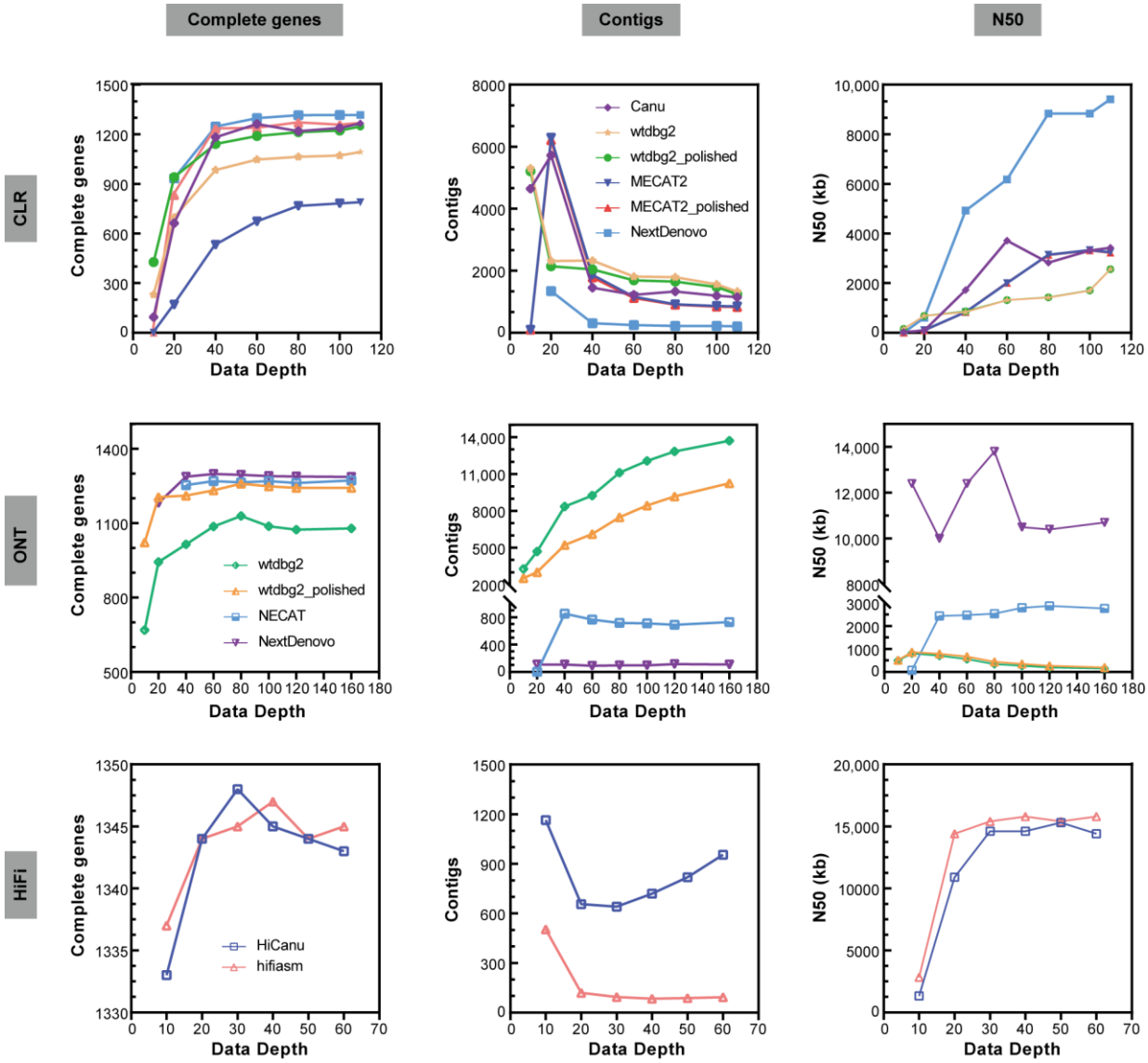

Figure 3. Main metrics of assemblies on CLR, ONT, HIFI subsets with different data depths. Complete gene number from BUSCO (Complete genes), contig number (Contigs), and N50 of contigs (N50).

The subsequent Racon long-read polishing process greatly improved the wtdbg2 draft genome assemblies' completeness as indicated by the BUSCO complete gene percentages, which increased from between a minimum range of 49% to 82.6% to a maximum range of 74.7% to 92.1% (Table S1). The assembly accuracy was also greatly improved as indicated by the number of Small-scale errors, which decreased from between 3484 and 7767 per Mbp to between 633 and 1418 per Mbp (Figure S2a).

For the purpose of investigating the effect of sequencing depth on assembly tools, we evaluated the quality of ONT assemblies on diverse sequencing depths (10×, 20×, 40×, 60×, 80×, 100×, 120× and 160×). The assembly quality on low-depth subsets (10× or 20×) varied greatly amongst different assemblers, whereas it was reliable on relatively high-depth subsets (Figure 3). According to our findings, the dataset with around 40× ONT can construct the most genomes. However, a deeper sequencing effort is required to further enhance the genome quality.

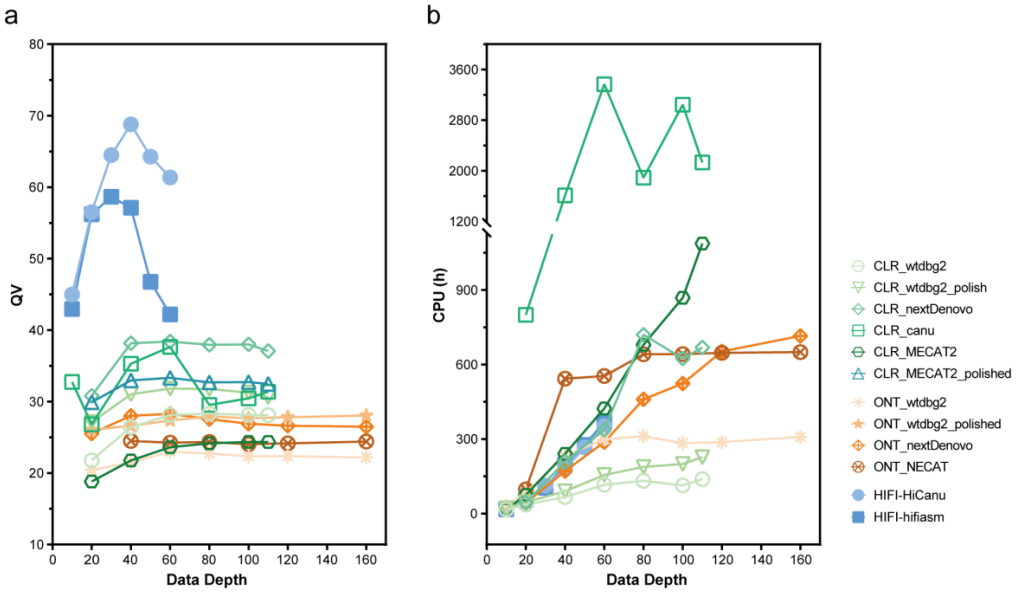

Figure 4. Quality Value (QV) score and computational time of assemblies on CLR, ONT, HIFI subsets with different data depths. (**a**) QV and (**b**) Computational time of different assemblers.

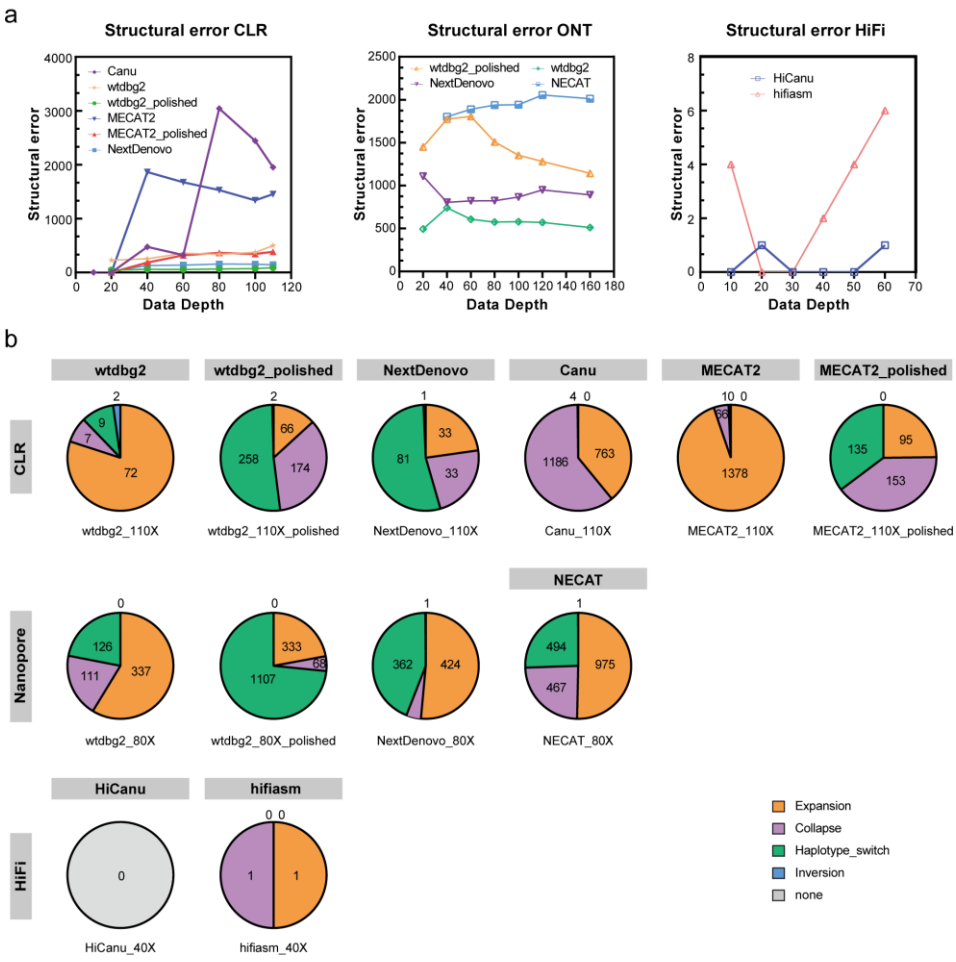

Figure 5. Structural error of assemblies on CLR, ONT, HIFI subsets. (**a**) Structural errors of assemblies with different data depths. (**b**) Pie graphs displaying the percentage of four types of structural errors discovered by the Inspector. The number of assembly errors is shown in each sector.

2.3. CLR Genome Assembly

In order to investigate how to generate high-quality haploid genome assemblies from wild harvested genome degradation samples, we selected the slightly poorer quality genomes extracted from silkworm D9L samples for testing (N50 = 11,722 bp, E-size = 11,909 bp, Table 1 and Figure 1). The assembly of the CLR reads was conducted using four different long-read assembly tools (NextDenovo, Canu, wtdbg2, and MECAT2). Due to lower genomic heterozygosity, the CLR assemblies showed much smaller differences than ONT in genome size (426–506 Mb, excluding the 10× and 20× results, Table S1). When a certain sequencing depth is satisfied (>= 40×), the difference in the number of contigs for each genome assembly is not significant, and the result of NextDenovo remains the best. The continuity of all the assemblies (N50 of contigs) increased by the sequencing depth, the NextDenovo assembly increased the most pronounced (Figure 3).

The CLR assemblies showed similar complete BUSCO gene numbers as the ONT assemblies (excluding MECAT2). Wtdbg2 generated the lowest number of Structural errors, followed by NextDenovo (Figure 5). NextDenovo generated the lowest number of Small-scale errors followed by Canu (Figure S2). The NextDenovo assembly showed the highest contiguous (contig N50 = 9.41 Mb), smallest size (477 Mb), and the least contigs (n = 205) (Table S1). The Canu assembly was the largest (506 Mb) but contained a high degree of duplication as indicated by the percentage of duplicated BUSCOs (2.9%). Therefore, as was recently discovered, the assembly of Canu probably contains uncollapsed haplotypes corresponding to artifactually duplicated areas [23]. The assembly quality of four assemblers was assessed by the metric mean of the six different subsets (Table 2). NextDenovo shows the best overall performance, followed by Canu. Though Canu needs the longest CPU hours and generates fragmented assemblies, the accuracy is excellent (Figure 4b).

Before polishing, the wtdbg2 assembly was the most fragmented (contig N50: 0.154–2.56 Mb) and the MECAT2 assembly was the least complete (12.5–57.8% complete BUSCOs) (Table S1). Subsequently, we polished the wtdbg2 and MECAT2 assemblies using the CLR long-reads. The Racon polishing steps greatly improved the wtdbg2 and MECAT2 draft assemblies' genome completeness (Figure 3). As expected, a reduced number of Small-scale errors and Structural errors were identified in CLR assemblies when compared with ONT assemblies (Figures 5 and S2). The long-read polishing process resulted in the percentage of single-copy BUSCOs increasing significantly and the number of Small-scale errors for the MECAT2 assembly reduced sharply, while the number of Small-scale errors of wtdbg2 assembly reduced modestly (Table S1). Interestingly, the long-read polishing process did not improve the integrity of the NextDenovo and Canu assemblies. In this part, we found that the dataset with roughly 40× CLR can construct most genomes, increase the sequencing depth could improve the genome quality, and need a polishing strategy or not depending on which assemblers are used.

2.4. HiFi Genome Assembly

Purposed to testing the contribution of the new technology to high-quality genome assembly, we performed PacBio HiFi sequencing on P50T silkworm whose genomes had been previously sequenced. HiFi reads were assembled using HiCanu and hifiasm. Compared with CLR and ONT assemblies, the genomic continuity and integrity of HiFi assemblies were significantly superior. There were no significant differences in the size, continuity, and completeness of the HiFi genome assemblies. The greatest difference is reflected in the contig numbers, which are much smaller in hifiasm assemblies than those in HiCanu assemblies (Figure 3 and Table S1). We polished the HiFi assemblies using the HiFi long-read sequences. Just as expected, the polished HiFi assembly showed a similar percentage of complete BUSCO genes compared to the raw HiFi assembly (Table S1). When compared with the ONT and CLR assembly, the HiFi assembly contained the fewest Structural errors and Small-scale errors (Figures 5, S2 and S3).

Furthermore, we evaluated the quality of HiFi assemblies on datasets with different sequencing depths (10×, 20×, 30×, 40×, 50×, and 60×) to investigate the effect of data depth on assemblers. For low-depth subsets (as the 10× subset depicted in Figure 3), the assembly quality was highly varied amongst assemblers, while on relatively high-depth subsets, it was resilient. When exceeding a certain threshold, high coverage subsets do not significantly improve the quality of assembly, either. However, higher depths will require more computing resources and instantiation times. Therefore, choosing an appropriate depth is crucial. According to our findings, the dataset with about 20× HiFi data was able to create the most genomes. Since HiFi only requires a sequencing depth of 20× or more to build most of the genome and does not require a subsequent polish process, the time used for genome assembly is much less than that of ONT and CLR, especially when using Canu.

Compared with the other two sequencing methods, HiFi assembly shows the best assembly quality, the lowest contig number, and the highest continuity, accuracy, and completion, without relying on other scaffolding. It also requires the least amount of time and computer memory and can be considered the optimal sequencing method for future Lepidopteran pest genomes.

2.5. Construction and Quality Assessment of Hi-C-Based Chromosome-Level Genomes

The quality of the three long-read sequencing assemblies was significantly superior compared with short-read sequencing. However, none of the HiFi assemblies completed the assembly of all the chromosomes. We selected the best genome assembly for each sequencing method using 3D-DNA for genome construction at the chromosome level. Using default parameters, 3D-DNA achieved clustering of most of the chromosomes. However, there remained some chromosome clustering errors and contig translocations and inversions, which were identified using the Hi-C map (Figure S4). Further manual adjustment by Juicebox can fix these organizational errors. However, this individualized manual adjustment often does not conform to a uniform standard. We then designed a quality assessment standard for chromosome-level genome assembly based on EagleC. This can identify organizational errors rapidly and accurately and is able to report the percentage of misassemblies in the genome assembly in the form of a table to facilitate the correction of these assembly errors (Figure 6c and Table 3). Based on EagleC's recommendations, we completed the adjustment of the genome assemblies and performed polishing using Racon and gap filling using TGS-GapCloser. Finally, using five-base telomere repeats ('*TTAGG*') [24] as a sequence query, we identified 50 telomeres and constructed 28 pseudomolecules (25 of 28 were represented by a single large contig, and the remaining three were assembled from two main contigs) for the silkworm (P50T-HiFi) genome (Figure 6a,c). Compared with the SilkBase reference genome (P50T-SilkBase), the P50T-HiFi assembly filled 30 gaps that were found in the SilkBase assembly. These gaps ranged from 99 to 75,391 bp and were distributed throughout the genome. The parallel plots showed that the P50T-HiFi assembly displayed good collinearity with the P50T-SilkBase assembly in most of the chromosomes, however, we also found some differences (Figure 6c). According to the EagleC report, these discrepant regions are caused by several Mb-level assembly errors, for example Chr24 (Figure 6e). The assembly mistake in the P50T-SilkBase assembly is also confirmed by the Chr19 parallel plots of five silkworm genome assemblies (Figure 6d). Furthermore, the final remaining three gaps in the P50T-HiFi assembly were filled with P50T-SilkBase assembly, resulting in a gap-free silkworm genome assembly (P50T). This is the first nearly complete telomere-to-telomere reference genome of silkworm (P50T). Although the genome assembly quality of CLR and ONT is not as good as that of HiFi, both completed very high consecutive and complete chromosome-level genome assemblies after treatment with EagleC and 3D-DNA, which is based on Hi-C (Figure 6b and Table 3).

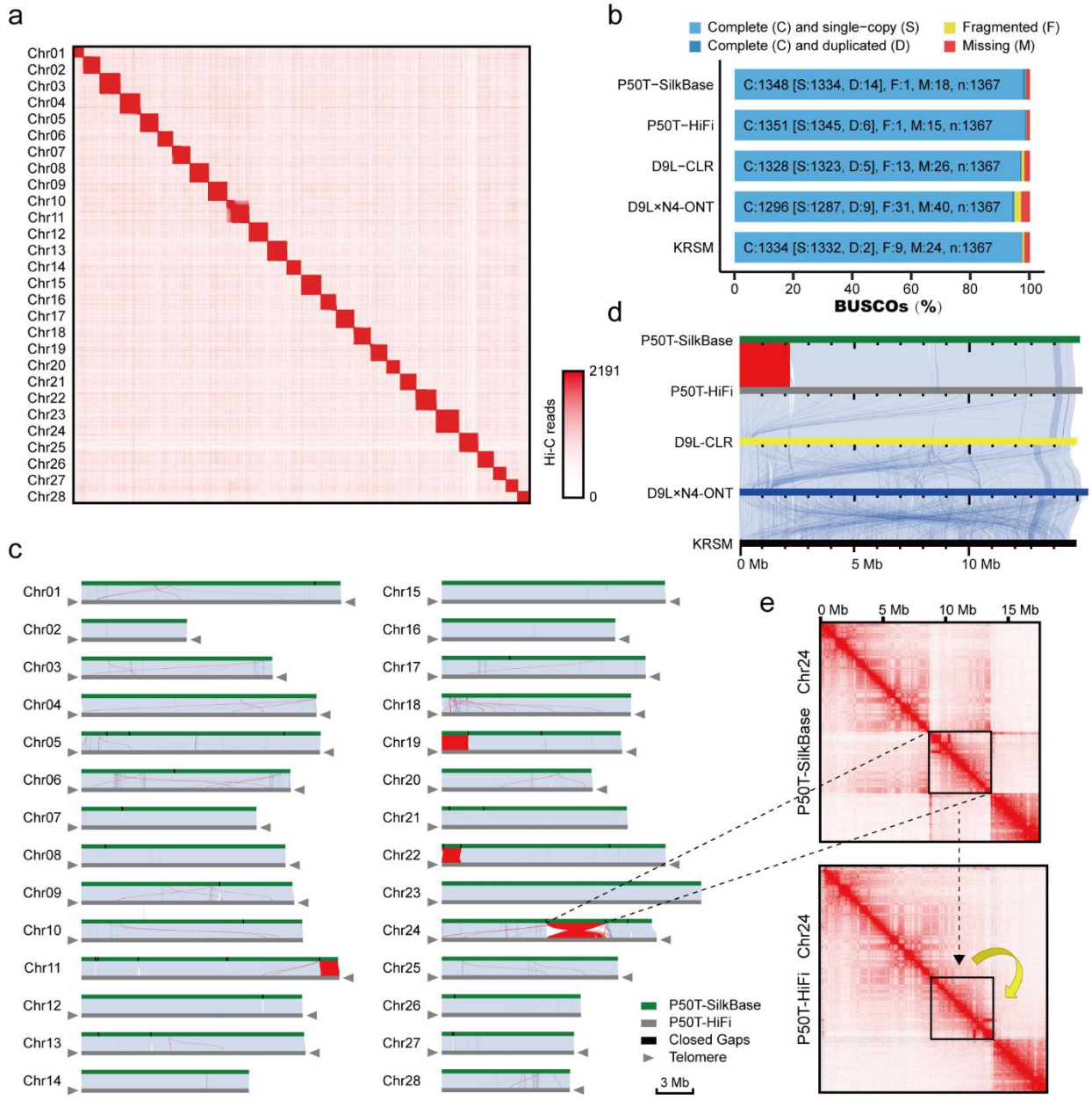

Figure 6. Summary of different silkworm strains chromosome-level genome assembly. (**a**) Hi-C genome-wide interaction map of the silkworm (P50T-HiFi) assembly. (**b**) BUSCO analysis of chromosome-level genome assemblies using the insect odb10 (1367 genes). (**c**) Collinearity between the silkworm P50T-SilkBase and P50T-HiFi genomes. The synteny blocks are shown by light blue lines. The inversions are indicated by red lines. The telomere sequence repeats are marked by gray triangles. All the P50T-SilkBase gap regions closed in P50T-HiFi are shown as black blocks. The gray triangles indicate the presence of telomere sequence repeats. (**d**) Collinearity of the P50T-SilkBase, P50T-HiFi, D9L-CLR, D9L × N4-ONT and Korean silkworm (KRSM) Chromosome 19. Synteny blocks are shown by light blue lines. The inversions are indicated by red lines. (**e**) Hi-C interaction map of silkworm chromosome 24. The black boxed area indicates the error inversion in P50T-SilkBase.

Table 3. Statistics of different silkworm chromosome-level genome assemblies.

	N9L-CLR [a]	D9L × N4-ONT [b]	P50T-HiFi [c]	KRSM [d]
Size (Mb)	446.7	454.6	456.6	446.2
Scaffold N50 (Mb)	16.92	17.48	16.92	16.89
Contigs	51	29	29	33
Contig N50 (Mb)	14.38	17.48	16.92	16.89
Max contig [e] (Mb)	21.50	21.92	21.53	21.51
Structural error	193	852	2	34
Small-scale error (/Mb)	88.47	980.86	1.79	55.13
QV	36.84	27.37	56.84	41.75
BUSCOs [f]	97.2%	94.8%	98.8%	97.5%
PAR [g]	0.11%	0.03%	0.03%	0.05%

[a–d] chromosome-level genome assemblies of silkworm N9L, D9L × N4, P50T and KRSM, [e] Maximum length of contigs, [f] Percentage of Complete BUSCO genes, [g] Percentage of assembly errors identified by EagleC.

2.6. Case

Whether our assembly process is applicable to the genome assembly of other Lepidopteran pests, and whether it can help optimize the genome assembly of published genomes, we selected the genome sequencing data of Korean silkworm (KRSM) [25] and *Dendrolimus punctatus* (*D. punctatus*) [26] for testing. Based on the results of the above comparison, an optimal pipeline was selected for those Lepidopteran insects.

We evaluated the metrics of the final assemblies and demonstrated these in Table 3 and Table S2. The pipeline can build the high-completeness genome assembly in KRSM with approximately 16.89 Mb scaffold N50, 97.5% complete BUSCO genes (Figure 6b), 55.13 small-scale errors per Mb, 34 structural errors and the value of PAR is 0.05%, demonstrating the accuracy of this genome assembly. On the other hand, misassemblies in the genome assembly of *D. punctatus* were identified and optimized using the EagleC evaluation process, significantly improving the quality of the optimized genome (Figure 7b–d). The circle plots showed that the genome of *D. punctatus* we assembled here shared good collinearity with *Dendrolimus kikuchii (D. kikuchii)*, and confirmed the assembly errors published in previous studies (Figure 7e). This demonstrates the compelling potential of the EagleC evaluation process to assess and optimize the quality of published genome assemblies.

For genome sequencing of Lepidopteran pests, we recommend HiFi and Hi-C sequencing followed by hifiasm and 3D-DNA for assembly and chromosome mounting, which achieves the best haploid genome assembly. For species already sequenced by ONT or CLR, we recommend NextDenovo, 3D-DNA, and EagleC for chromosome-level genome optimization.

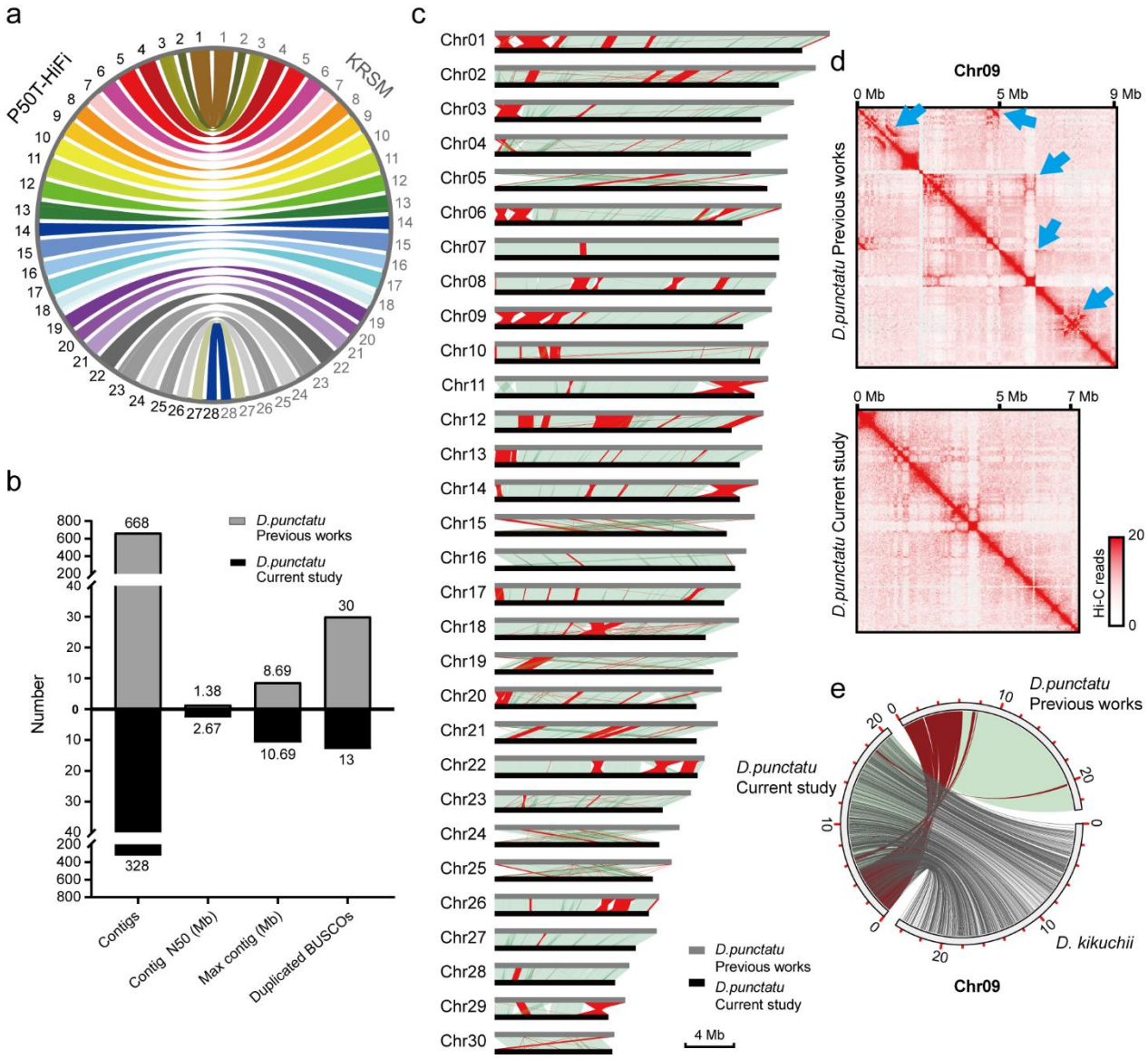

Figure 7. Summary of KRSM and *Dendrolimus punctatus* (*D. punctatus*) assemblies. (**a**) The synteny between silkworm P50T-HiFi and KRSM genomes. (**b**) Comparative analysis of the metric values of *D. punctatus* genome assemblies between the previous studies and this study. Contig numbers (Contigs), N50 of contigs (Contig N50), the maximum length of contigs (Max contig) and duplicated complete BUSCO gene number (Duplicated BUSCOs). (**c**) Collinearity of *D. punctatus* genome assemblies in previous research and this study. The synteny blocks are shown by light green lines. The inversions are indicated by red lines. (**d**) Hi-C interaction map of *D. punctatus* chromosome 9. The assembly errors in previous research were marked by blue arrows. (**e**) Collinearity circle plots. The synteny blocks are shown by light blue and gray lines. The inversions are indicated by red lines. It showed that the genome of *D. punctatus* (in this study) shared good collinearity with *Dendrolimus kikuchii* (*D. kikuchii*), and confirmed the assembly errors in previous studies.

3. Discussion

Lepidoptera is the second largest order of insects, some of which are severe pests of agriculture and forests and cause significant economic losses each year. Genome sequencing of Lepidopteran pests has contributed greatly to their control. The accumulating genomic resources have been a crucial source for major breakthroughs in life science innovations and discoveries. Seventy-six arthropod genome assemblies were used to characterize the changes in genes and protein contents for a better understanding of 500 million years of

evolution [27]. A study of 195 insect genomes revealed a high diversity of transposable elements across insects with varying degrees of conservation depending on phylogenetic position [28]. Horizontal gene transfer (HGT) events in 218 insects acquired from non-metazoan sources provide insight into the adaptation of GTs in insects [29]. The breadth of Lepidopteran pest genome sequencing spans approximately 300 million years of evolution and roughly two orders, with genome sizes ranging from the tiny 229.9 Mb genome of *Papilio polytes* to the massive genome of *Parnassius apollo* at 1392 Mb [10]. However, this represents a mere 0.08% of the approximately 160,000 described insects [11], with many orders remaining without genomic representation. In the future, more Lepidopteran pest genomes will be sequenced. There are a number of questions that need to be addressed at that time: What sequencing strategy and which assembler should be chosen? How can we complete the genome assembly and quality assessment faster and more accurately?

In this study, we use the silkworm as a model to compare and analyze the mainstream TGS strategies, including sequencing platform, sequencing depth, and assemblers. Using the silkworm high-quality genome assembly as a reference, we compared BUSCO, Inspector, and EagleC to find the most appropriate long-read sequencing strategy under different conditions. We performed 32, 42, and 12 de novo genome assemblies of silkworms on high-depth ONT, CLR, and HiFi reads, correspondingly to comprehensively assess the effect of assemblers and sequencing coverage on Lepidopteran insect genome assembly. Each sequencing strategy has its own advantages: the ONT genomic library has the largest fragment size, the HiFi data has the highest accuracy, and the CLR is in between. The focus of different assemblers is different.

On graft genome assembly, the NextDenovo assembler shows the best performance on both CLR and ONT sequencing datasets, and hifiasm is better for HiFi datasets. We recommend the use of 3D-DNA in combination with the EagleC evaluation to complete the construction of chromosome-level genomes. The polishing process is completed without the original assembly, you can finish the scaffolding and polish it afterward, the quality of the assembly is similar, and it is timesaving. In the case application, the KRSM genome was obtained from CLR and Hi-C data with excellent continuity and integrity. Other factors, such as throughput, convenience, and price should also be reasons for considering which genome sequencing platform to choose.

Here, using various techniques and ultra-high-depth datasets, we evaluated the effects of sequencing coverage on Lepidopteran pest genomes assembly. The quality of the genome tends to stabilize after 40× on the ONT and CLR datasets, and at 20× on HiFi datasets, and improves as the sequencing depth increases. The sequencing depth of 20× is the minimum for genome construction, however, the higher the sequencing depth, the better the assembly effect is not necessary, and too high a sequencing depth will cause excessive consumption of computing resources. Especially for large genome projects, such as pan-genome, you must select an appropriate sequencing depth to efficiently reduce the burden of computing resources and the cost of time and money.

The currently available mainstream genome quality assessment markers are N50, BUSCO, Merqury, QUAST-LG, and Inspector. However, the N50 is just a simple continuous statistic, BUSCO can only evaluate conserved genomic regions, Mercury requires users to input high-precision reads and is not suitable for long-read data, and QUAST-LG relies excessively on existing reference genomes [18]. Several software programs have been developed to implement scaffolding based on Hi-C data, HiRISE, LACHESIS, SALSA, 3D-DNA, and ALLHiC [30]. EagleC combines deep-learning and ensemble-learning strategies can predict the whole range of SVs with a Hi-C map very quickly and accurately [22]. SVs can induce de novo chromatin interactions across the breakpoints, which are similar to assembly errors, both show aberrant interaction blocks. We then designed a quality assessment standard for chromosome-level genome assembly based on EagleC. The EagleC process that we developed is different from the previous evaluation work. It is a deep learning-based process used to accurately and rapidly evaluate the quality of chromosome-level genome assemblies and direct the repair of assembly errors. It can also be used for the

optimization of published chromosome-level genome assemblies. The quality of de novo genome assemblies has a great significant impact on gene annotation and comparative genomic research [31]. At the same time, we noticed that although we have developed good quality assessment standards for chromosome-level genome assembly, it is difficult for novices to complete the assessment. Building an online database that can generate results reports with one click can greatly solve this problem, and facilitate the widespread use of scholars with different research backgrounds, which is what we are currently doing.

In this study, we take into account that practical issues are faced in Lepidopteran pest genome sequencing projects, including high genome heterozygosity, poor quality genome libraries with short fragments, and assembly approaches tailored to various scenarios. Additionally, it has demonstrated how to improve an existing genome assembly based on the findings of a genome evaluation without producing new sequencing data. This benchmark work offers insights for other eukaryote genomes such as mildew and microalgae, and even complicated human genomes, in addition to helping to build the high-quality genomes of Lepidopteran pests.

4. Materials and Methods

4.1. Insect Material

B. mori strains (P50T, D9L, and D9 × N4) were sourced from the Center for Frontier Interdisciplinary Biology, Southwestern University, China. Silkworms were reared on mulberry leaves under 12-h light and 12-h dark photoperiod at 28 °C from the 1st to 4th instars, and 25 °C after the 4th ecdysis.

4.2. Genome Sequencing and Creation of Subsets

Genomic DNA of P50T, D9L, and D9 × N4 was extracted, detected, and sequenced for generating PacBio HiFi, PacBio CLR and ONT reads at Frasergen (Wuhan, China), separately. Among them, D9L and D9 × N4 have not been previously sequenced.

Supplementary Table S1 provides a summary of the statistics for each sequenced dataset. In order to investigate the dependence of assemblers on different sequencing depths and its influence on the quality of assembly, we used Seqtk (v1.2) and randomly selected eight subsets with divergence sequence depths (10×, 20×, 40×, 60×, 80×, 100×, 120×, 160×) in ONT data, seven subsets (the depths were 10×, 20×, 40×, 60×, 80×, 100×, 110×) of CLR data and six subsets (the depths were 10×, 20×, 30×, 40×, 50×, 60×) in HiFi data. Each subset shared similar read length distribution and coincident read length density (Figure 1).

4.3. De Novo Genome Assembly Workflow

De novo genome assembly and polishing workflow are displayed in Figure 2. For ONT and CLR subsets, we used four long-read assembly tools with default parameters: Canu (v1.9) [32], wtdbg2 (v2.5) [33], NECAT (v20200803) [34] /MECAT2 (v20200228) [35], and NextDenovo (v2.5.0) (https://github.com/Nextomics/NextDenovo, accessed on 10 June 2022). For HiFi subsets, we performed two assemblers, HiCanu (v2.2) [36] and hifiasm (v0.16.1) [37] using the default parameters. The long-reads were mapped to the graft assemblies with Minimap2 (v2.17) [38] and then polished using Racon (v1.5.0) [39].

4.4. Hi-C Scaffolding and Gap Filling

The Hi-C raw data of silkworm (P50T) was used to scaffold the genome assembly to the chromosomal level. Low-quality Hi-C raw reads were filtered out using Trimmomatic (v0.39) [40] (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50 CROP:50). The clean paired-end reads were mapped to the draft assembly by bwa (0.7.17) [41], and then analyzed by juicer (v1.6) [42]. Following this, 3D-DNA (v 180419) [43] and juicerbox (v1.11.08) [44] was applied to produce the chromosome-level assembly for silkworm. TGS-GapCloser (v 1.1.1) [45] was used to close gaps in the genome assemblies by long-reads.

4.5. Genome Assembly Evaluation

QUAST (v5.0.2) and Inspector were used to assess the assembly quality generated by different assembly tools. BUSCO(v4) was used to assess the completeness of the genome assembly with the insect (odb10) protein set. We selected the number and N50 of contigs, complete genes number from BUSCO, Small-scale errors per Mb, the number of Structural errors, and Quality Value (QV) score from Inspector to visualize in the main text. The QV was calculated on the base of the identified structural and small-scale errors in the assemblies [18].

Furthermore, we designed a quality assessment standard for chromosome-level genome assembly based on EagleC, a deep-learning framework for detecting a full range of assembly errors from Hi-C contact maps that were used to identify both small-scale and large-scale assembly errors, accurately. It reports the percentage (Equation (1)), type, and locus of specific assembly errors, and provides solutions on how to fix these assembly errors.

$$\text{Percentage of assembly errors (PAR)} = \frac{\text{Total length of the assembly errors}}{\text{Total length of the assembly}} \quad (1)$$

4.6. Chromosomal Synteny Analysis

Genome comparisons were completed using NUCmer (v4) [46] with default parameters. NUCmer's alignment file was filtered using delta-filter(-i 85 -l 8000 -o 85 -1). Mummerplot was used to create a dot plot. TBtools [47] was used to create a circos plot.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24010649/s1.

Author Contributions: Conceptualization, T.Z., S.M. and Q.X.; Formal Analysis, Methodology, Software, T.Z. and W.X.; Data Curation, L.J.; Investigation, A.W.; Resources, N.Z.; Writing—Original Draft Preparation, T.Z. and S.M.; Writing—Review and Editing, S.M. and Q.X.; Visualization, T.Z. and W.X.; Supervision, Q.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the National Natural Science Foundation of China (no. 32122084, 32030103), the Natural Science Foundation of Chongqing (no. cstc2020jcyj-cxttX0001), and the Fundamental Research Funds for the Central Universities (no. SWU-KT22042).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All raw sequencing data are available from the NCBI database with BioProject accession PRJNA898773. The genome assemblies of silkworm strain P50T, D9L and D9L × N4 are available from the NCBI database with BioProject accession PRJNA912164, PRJNA912165 and PRJNA912157.

Conflicts of Interest: The authors declare that they have no competing interest with this article's content.

References

1. You, M.; Ke, F.; You, S.; Wu, Z.; Liu, Q.; He, W.; Baxter, S.W.; Yuchi, Z.; Vasseur, L.; Gurr, G.M.; et al. Variation among 532 genomes unveils the origin and evolutionary history of a global insect herbivore. *Nat. Commun.* **2020**, *11*, 2321. [CrossRef]
2. Wu, N.; Zhang, S.; Li, X.; Cao, Y.; Liu, X.; Wang, Q.; Liu, Q.; Liu, H.; Hu, X.; Zhou, X.J.; et al. Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nat. Ecol. Evol.* **2019**, *3*, 105–115. [CrossRef] [PubMed]
3. Chen, Q.; Zhao, H.; Wen, M.; Li, J.; Zhou, H.; Wang, J.; Zhou, Y.; Liu, Y.; Du, L.; Kang, H.; et al. Genome of the webworm Hyphantria cunea unveils genetic adaptations supporting its rapid invasion and spread. *BMC Genom.* **2020**, *21*, 242. [CrossRef]
4. Wan, F.; Yin, C.; Tang, R.; Chen, M.; Wu, Q.; Huang, C.; Qian, W.; Rota-Stabelli, O.; Yang, N.; Wang, S.; et al. A chromosome-level genome assembly of Cydia pomonella provides insights into chemical ecology and insecticide resistance. *Nat. Commun.* **2019**, *10*, 4237. [CrossRef] [PubMed]
5. Benowitz, K.M.; Allan, C.W.; Degain, B.A.; Li, X.; Fabrick, J.A.; Tabashnik, B.E.; Carrière, Y.; Matzkin, L.M. Novel genetic basis of resistance to Bt toxin Cry1Ac in Helicoverpa zea. *Genetics* **2022**, *221*, iyac037. [CrossRef]

6. Edelman, N.B.; Frandsen, P.B.; Miyagi, M.; Clavijo, B.; Davey, J.; Dikow, R.B.; García-Accinelli, G.; Van Belleghem, S.M.; Patterson, N.; Neafsey, D.E.; et al. Genomic architecture and introgression shape a butterfly radiation. *Science* **2019**, *366*, 594–599. [CrossRef]
7. Xia, Q.; Li, S.; Feng, Q. Advances in silkworm studies accelerated by the genome sequencing of Bombyx mori. *Annu. Rev. Entomol.* **2014**, *59*, 513–536. [CrossRef]
8. Kumar, K.R.; Cowley, M.J.; Davis, R.L. Next-Generation Sequencing and Emerging Technologies. *Semin. Thromb. Hemost.* **2019**, *45*, 661–673. [CrossRef]
9. Sohn, J.I.; Nam, J.W. The present and future of *de novo* whole-genome assembly. *Brief Bioinform.* **2018**, *19*, 23–40.
10. Mei, Y.; Jing, D.; Tang, S.; Chen, X.; Chen, H.; Duanmu, H.; Cong, Y.; Chen, M.; Ye, X.; Zhou, H.; et al. InsectBase 2.0, a comprehensive gene resource for insects. *Nucleic Acids Res.* **2022**, *50*, D1040–D1045. [CrossRef]
11. Triant, D.A.; Cinel, S.D.; Kawahara, A.Y. Lepidoptera genomes, current knowledge.; gaps and future directions. *Curr. Opin. Insect. Sci.* **2018**, *25*, 99–105. [CrossRef] [PubMed]
12. 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.; et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **2019**, *37*, 1155–1162. [CrossRef] [PubMed]
13. van Dijk, E.L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **2018**, *34*, 666–681. [CrossRef]
14. Zhang, X.; Liu, C.G.; Yang, S.H.; Wang, X.; Bai, F.W.; Wang, Z. Benchmarking of long-read sequencing.; assemblers and polishers for yeast genome. *Brief Bioinform.* **2022**, *23*, bbac146. [CrossRef]
15. Kim, J.; Lee, C.; Ko, B.J.; Yoo, D.A.; Won, S.; Phillippy, A.M.; Fedrigo, O.; Zhang, G.; Howe, K.; Wood, J.; et al. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. *Genome Biol.* **2022**, *23*, 204. [CrossRef] [PubMed]
16. Ko, B.J.; Lee, C.; Kim, J.; Rhie, A.; Yoo, D.A.; Howe, K.; Wood, J.; Cho, S.; Brown, S.; Formenti, G.; et al. Widespread false gene gains caused by duplication errors in genome assemblies. *Genome Biol.* **2022**, *23*, 205. [CrossRef]
17. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO, assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
18. Chen, Y.; Zhang, Y.; Wang, A.Y.; Gao, M.; Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **2021**, *22*, 312. [CrossRef]
19. Yamaguchi, K.; Kadota, M.; Nishimura, O.; Ohishi, Y.; Naito, Y.; Kuraku, S. Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Mol. Ecol.* **2021**, *30*, 5923–5934. [CrossRef]
20. Lu, F.; Wei, Z.; Luo, Y.; Guo, H.; Zhang, G.; Xia, Q.; Wang, Y. SilkDB 3.0, visualizing and exploring multiple levels of data for silkworm. *Nucleic Acids Res.* **2020**, *48*, D749–D755. [CrossRef]
21. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST, quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef] [PubMed]
22. Wang, X.; Luan, Y.; Yue, F.; Eagle, C. A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* **2022**, *8*, eabn9215. [CrossRef]
23. Murigneux, V.; Rai, S.K.; Furtado, A.; Bruxner, T.J.C.; Tian, W.; Harliwong, I.; Wei, H.; Yang, B.; Ye, Q.; Anderson, E.; et al. Comparison of long-read methods for sequencing and assembly of a plant genome. *Gigascience* **2020**, *9*, giaa146. [CrossRef] [PubMed]
24. Nichuguti, N.; Fujiwara, H. Essential factors involved in the precise targeting and insertion of telomere-specific non-LTR retrotransposon.; SART1Bm. *Sci. Rep.* **2020**, *10*, 8963. [CrossRef] [PubMed]
25. Kim, S.W.; Kim, M.J.; Kim, S.R.; Park, J.S.; Kim, K.Y.; Kim, K.H.; Kwak, W.; Kim, I. Whole-genome sequences of 37 breeding line Bombyx mori strains and their phenotypes established since 1960s. *Sci. Data* **2022**, *9*, 189. [CrossRef] [PubMed]
26. Zhang, S.; Shen, S.; Peng, J.; Zhou, X.; Kong, X.; Ren, P.; Liu, F.; Han, L.; Zhan, S.; Huang, Y.; et al. Chromosome-level genome assembly of an important pine defoliator.; Dendrolimus punctatus (Lepidoptera; Lasiocampidae). *Mol. Ecol. Resour.* **2020**, *20*, 1023–1037. [CrossRef]
27. Thomas, G.W.C.; Dohmen, E.; Hughes, D.S.T.; Murali, S.C.; Poelchau, M.; Glastad, K.; Anstead, C.A.; Ayoub, N.A.; Batterham, P.; Bellair, M.; et al. Gene content evolution in the arthropods. *Genome Biol.* **2020**, *21*, 15. [CrossRef]
28. Peccoud, J.; Loiseau, V.; Cordaux, R.; Gilbert, C. Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4721–4726. [CrossRef]
29. Li, Y.; Liu, Z.; Liu, C.; Shi, Z.; Pang, L.; Chen, C.; Chen, Y.; Pan, R.; Zhou, W.; Chen, X.X.; et al. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **2022**, *185*, 2975–2987.e10. [CrossRef]
30. Zhang, X.; Zhang, S.; Zhao, Q.; Ming, R.; Tang, H. Assembly of allele-aware.; chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **2019**, *5*, 833–845. [CrossRef]
31. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu, scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [CrossRef] [PubMed]
32. Li, M.; Sun, C.; Xu, N.; Bian, P.; Tian, X.; Wang, X.; Wang, Y.; Jia, X.; Heller, R.; Wang, M.; et al. De Novo Assembly of 20 Chicken Genomes Reveals the Undetectable Phenomenon for Thousands of Core Genes on Microchromosomes and Subtelomeric Regions. *Mol. Biol. Evol.* **2022**, *39*, msac066. [CrossRef] [PubMed]
33. Ruan, J.; Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **2020**, *17*, 155–158. [CrossRef] [PubMed]

34. Chen, Y.; Nie, F.; Xie, S.Q.; Zheng, Y.F.; Dai, Q.; Bray, T.; Wang, Y.X.; Xing, J.F.; Huang, Z.J.; Wang, D.P.; et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **2021**, *12*, 1–10. [CrossRef] [PubMed]
35. Xiao, C.L.; Chen, Y.; Xie, S.Q.; Chen, K.N.; Wang, Y.; Han, Y.; Luo, F.; Xie, Z. MECAT: Fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat. Methods* **2017**, *14*, 1072–1074. [CrossRef]
36. Nurk, S.; Walenz, B.P.; Rhie, A.; Vollger, M.R.; Logsdon, G.A.; Grothe, R.; Miga, K.H.; Eichler, E.E.; Phillippy, A.M.; Koren, S. HiCanu, accurate assembly of segmental duplications.; satellites.; and allelic variants from high-fidelity long reads. *Genome Res.* **2020**, *30*, 1291–1305. [CrossRef]
37. Cheng, H.; Concepcion, G.T.; Feng, X.; Zhang, H.; Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **2021**, *18*, 170–175. [CrossRef]
38. Li, H. Minimap2, pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef]
39. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737–746. [CrossRef]
40. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic, a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
41. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
42. Durand, N.C.; Shamim, M.S.; Machol, I.; Rao, S.S.; Huntley, M.H.; Lander, E.S.; Aiden, E.L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **2016**, *3*, 95–98. [CrossRef]
43. Dudchenko, O.; Batra, S.S.; Omer, A.D.; Nyquist, S.K.; Hoeger, M.; Durand, N.C.; Shamim, M.S.; Machol, I.; Lander, E.S.; Aiden, A.P.; et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **2017**, *356*, 92–95. [CrossRef] [PubMed]
44. Durand, N.C.; Robinson, J.T.; Shamim, M.S.; Machol, I.; Mesirov, J.P.; Lander, E.S.; Aiden, E.L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **2016**, *3*, 99–101. [CrossRef]
45. Xu, M.; Guo, L.; Gu, S.; Wang, O.; Zhang, R.; Peters, B.A.; Fan, G.; Liu, X.; Xu, X.; Deng, L.; et al. TGS-GapCloser, A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **2020**, *9*, giaa094. [CrossRef] [PubMed]
46. Marçais, G.; Delcher, A.L.; Phillippy, A.M.; Coston, R.; Salzberg, S.L.; Zimin, A. MUMmer4, A fast and versatile genome alignment system. *PLoS Comput. Biol.* **2018**, *14*, e1005944. [CrossRef] [PubMed]
47. Chen, C.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.; Xia, R. TBtools, An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* **2020**, *13*, 1194–1202. [CrossRef] [PubMed]