https://doi.org/10.1038/s41587-023-02100-3

Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy

In the format provided by the authors and unedited

1. Context

Genome assembly involves multiple components, with software continuously evolving, driven by improvements in hardware, algorithms, and long-read sequencing[1,2]. The analysis workflows used by research labs to produce genome assemblies need to be adapting to the fast-evolving sequencing and assembly technology.

1.1 Previous work

The previously published Vertebrate Genomes Pipeline (VGP) pipeline (v1.7) used PacBio Continuous Long Reads (CLR) to generate a set of primary and alternate pseudo-haplotype contigs (or fully phased contigs when parental data were available)[3], and then used 10X Genomics (10X) linked-reads, Bionano Genomics (Bionano) optical maps, and chromatin conformation (Hi-C) for long-range information to join contigs into larger chromosome-level scaffolds. Given the lower accuracy of CLR reads compared to other sequencing data types, these scaffolds were then polished for base call errors with a combination of long and short reads (10X data).

1.2 Motivation

While the VGP has already assembled over 100 vertebrate genomes through the v1.7 pipeline, recent radical changes in sequencing technologies—including the advent of PacBio high fidelity (HiFi) reads[4], the discontinuation of 10X linked reads, as well as the parallel evolution of genome assembly algorithms—now call for a major update and enhancement of the pipeline. Particularly, PacBio HiFi reads constitute a paradigm shift over previous PacBio CLR[4], as HiFi's higher accuracy allows assembly of long, highly identical repetitive regions[5]. In combination with Oxford Nanopore (ONT) ultra-long reads, HiFi reads enabled the completion of the first telomere-to-telomere (T2T) human genome[6,7]. HiFi reads are circular consensus reads (CCS) which are nearly error-free except for homopolymers: this accelerates the assembly process, ultimately generating genome assemblies of the highest quality with minimal resources. This has been demonstrated in the human genome reference[6], and evidence is rapidly accumulating that the approach successfully applies to many other species, including those with even larger and more complex genomes[8,9].

1.3 Overview

Here we present the latest Vertebrate Genomes Project assembly pipeline (VGP-Galaxy version 2.1) and demonstrate that it delivers reference genomes at scale across a set of vertebrate species that arose over the last ~500 million years. The pipeline is versatile and combines PacBio HiFi long-reads and Hi-C-based haplotype phasing in a new graph-based paradigm. We make the pipeline freely accessible through Galaxy, accommodating researchers without local computational resources and enhancing reproducibility by democratizing the training and assembly process. We demonstrate the flexibility and reliability of the pipeline by assembling reference genomes for 51 vertebrate species from major taxonomic groups (fishes, amphibians, reptiles, birds, and mammals). These species were assembled using automatic workflows in a process at least an order of magnitude faster than past efforts that lacked such workflows.

2. Results

2.1 The VGP-Galaxy pipeline is built on 10 modular workflows

The VGP-Galaxy pipeline is divided in six stages of analyses composed of ten workflows. The first stage of the pipeline is the generation of k-mer profiles of the raw reads to estimate genome size, heterozygosity, repetitiveness, and error rate, which is necessary for parameterizing downstream workflows. The generation of k-mer counts can be done from HiFi data only (Workflow 1), or include data from parental reads for trio-based phasing (Workflow 2; trio is a combination of paternal sequencing data with that from an offspring that is being assembled). The second stage is the phased contig assembly. In addition to using only HiFi reads (Workflow 3), the contig building (contiging) step can leverage Hi-C (Workflow 4) or parental read data (Workflow 5) to produce fully-phased assemblies (hap1/hap2 or parental/maternal assigned haplotypes), using hifiasm[10]. The contiging workflows also produce a number of critical quality control (QC) metrics, such as k-mer multiplicity profiles and Benchmarking Universal Single-Copy Ortholog (BUSCO) graphs[11]. Inspection of these profiles provides information to decide whether the third stage—additional purging of false duplications—is required. Purging (Workflow 6), using purge_dups[12] identifies and resolves haplotype-specific assembly segments incorrectly labeled as primary contigs, as well as heterozygous contig overlaps. This increases continuity and the quality of the final assembly[12]. The purging stage is generally unnecessary for trio data, since reliable haplotype resolution is performed using hapmers obtained from parental reads. The fourth stage, scaffolding, produces chromosome-level scaffolds using information provided by Bionano (Workflow 7), with Bionano Solve and/or Hi-C (Workflow 8) using the YaHS scaffolding algorithms. A final stage of decontamination (Workflow 9) removes exogenous sequences (e.g., viral and bacterial sequences) as well as mitochondrial artifacts from the scaffolded assembly. Additionally, a dedicated workflow (Workflow 0) is available for mitochondrial genome assembly.

2.2 Validation in zebra finch

The zebra finch (Taeniopygia guttata) genome has been the focus of multiple in-depth analyses by the VGP[13,14]. The current zebra finch reference is bTaeGutv1.4 (ZZ with added W), a CLR-based assembly generated by the VGP 1.6 pipeline[2]. Numerous datatypes and existing benchmark assemblies are available for the individual bTaeGut2 (heterogametic, ZW), along with trio parental sequence data, making it an ideal test case for our workflows. We performed three types of assemblies using different combinations of data (Supp. Table 1): Solo assembly (Workflows 1, 3, 6, and 9) utilizes PacBio HiFi data for the single individual; Hi-C assembly (Workflows 1, 4, 8, and 9) adds Hi-C data for improved scaffolding; and lastly, Trio assembly (Workflows 2, 5, 8, and 9) adds parental (♂bTaeGut3 (homogametic, ZZ) and ♀bTaeGut4 (ZW)) Illumina data for haplotype phasing. Fig. 1 shows key quality control (QC) measures for trio assembly as automatically produced by the pipeline.

Overall, we found the HiFi data applied to different contig and scaffolding paradigms on Galaxy produced excellent assemblies directly from the workflows, with high contiguity, completeness, and accuracy. These draft assemblies were then manually curated to evaluate and resolve residual structural errors, remove contaminants, and assign sequences to chromosomes.

- **Solo assembly** : Alignment to bTaeGut1.4.pri (GCA_003957565.4) showed that the primary bTaeGut2 assembly had the expected karyotype and that the smallest of the microchromosomes (31 - 37) were highly fragmented. Curation of the solo assembly resulted in the manual correction of scaffolds resulting in 31 scaffold breaks, 90 joins and approximately 11.7Mb of sequence removed as haplotypic duplication. The curated genome yielded 39 autosomes plus Z and W, with 98.55% of the sequence being assigned to chromosomes. Retrospective alignment of this assembly to the Hi-C phased assembly from the same sample revealed 9Mb of sequence missing from the W chromosome, this missing sequence appears to have been removed by the purging process.
- **HiC-phased assemblies** : The HiC-phased assemblies were curated as individual haplotypes. Alignment of both haplotypes to the Solo and Trio assemblies of bTaeGut2 revealed poor haplotype separation particularly in relation to the sex chromosomes and the smallest of the microchromosomes. In part this may account for the difference in size between the larger Hap1 1.16Gb assembly versus the 1.06Gb Hap2 assembly. Incomplete representations of Z and W were found in both haplotypes requiring manual redistribution totalling approximately 44.7Mb of sequence in order to obtain single, complete representations for Z and W. Regarding microchromosomes 29 through 37 it was found that Hap1 contained virtually complete, albeit fragmented, representations for these plus over 60 duplicated sequences amounting to approximately 6.28Mb missing from the microchromosomes in the Hap2 assembly. It was also noted that a handful of sequences appeared to be represented only once between both data sets indicating incomplete phasing over these regions. Artificial triplication as an artifact of the phasing process led to the complete removal of 2.8Mb of sequence from the assemblies. For each of the HIC-phased assemblies manual curation ultimately resulted in the expected karyotype of 39 autosomes plus Z or W. Manual interventions for Hap 1 amounted to 37 scaffold breaks, 159 joins and a total of 9.99Mb of haplotig removals for transfer into Hap2 with 95.99% of the sequence

assigned to chromosomes. For Hap2 there were 30 scaffold breaks, 82 joins and a total of 23.7Mb of haplotig removals for transfer into Hap1, 97.93% of the sequence was assigned to chromosomes.

- **Trio assemblies :** The Trio Maternal and Paternal assemblies were curated individually. The assemblies exhibited good separation of haplotypes with each assembly achieving the expected karyotype. As observed for other iterations of this assembly the smallest chromosomes; 30-37 were fragmented. For the Maternal assembly manual correction of the assembly resulted in 16 scaffold breaks, 88 joins and 0.3Mb of sequence removed as haplotypic duplication with 99.53% of the sequence assigned to chromosomes. For the Paternal there were 32 scaffold breaks, 193 joins and 0.6Mb removed as haplotypic duplication, 99.33% of the sequence was assigned to chromosomes.

The main issues were manually fixed, including: (1) removal of false duplications in the Solo primary assembly due to incomplete purging; (2) manual rebinning of partially misbinned sex chromosomes and microchromosomes in the Hi-C-phased assemblies due to incorrect phasing of sex chromosomes in pseudoautosomal regions; and (3) reassigning some paternal-specific sequences that were mis-assigned to the maternal haplotype in the Trio assemblies, apparent in the k-mer spectra profiles (Supp. Fig. 2A and Supp. Fig. 2B). We identified the cause as contigs incorrectly assigned by Hifiasm[8] based on too few parental k-mers. To fix this, we re-ran Workflow 5 specifying the problematic reads as ambiguous. After rebinning, we estimated the k-mer duplications and genome completeness of both haplotypes and found that the k-mer duplications were higher before rebinning (0.9% paternal and 4.4% maternal) than after rebinning (paternal 0.8% and maternal 3.7%) (Supp. Fig. 2C). Moreover, the rebinned paternal assembly showed 1.2% higher k-mer completeness than the default, although a slight decrease of k-mer completeness (0.2%) was shown in the rebinned maternal assembly (hapmer-specific completeness reported in Supp. Table 2).

To identify resolved, collapsed, or spurious gene duplications, we mapped HiFi reads to each HiFi assembly, and calculated read depth within RefSeq annotations[15] (Supp. Fig. 3). We found 49 genes with multiple copies in the hap1 of the Hi-C phased assembly, and no more than one copy in each of the other assemblies. Of these genes, 41 have two copies in hap1 and no copies in the hap2 Hi-C assembly suggesting possible incorrect read phasing (Supp. Fig. 4). Notably aside from two genes in the hap2 Hi-C assembly, there are no other genes duplicated exclusively in one assembly. Genes with high copy numbers in each assembly were *ARL14EPL* variant X2, *PHF7*, and *VDAC3* with 19-43, 18-23, and 4-14 collapsed plus resolved duplications, respectively. Thus the same set of reads, assembled differently, gives a different number of genes and their copies when processed uniformly.

2.2.1 Improvements of the HiFi- over the CLR-based assembly

We also compared the CLR-based assembly and HiFi-based assemblies generated in the VGP-Galaxy pipeline on the newerindividual, bTaeGut2 (ZW). Based on K* statistics contrasting the k-mer frequencies of reads and assembly[16], the CLR-based primary genome assembly contained the highest number of k-mer expansion and collapse errors (Supp. Fig. 2E) and the highest proportion of k-mer duplications Supp. Fig. 2F). These analyses were confirmed based on analysis of whole genome alignments of CLR- and HiFi-based zebra finch assemblies and read depth calculations (see Methods), showing that the CLR-based assembly was more prone to apparent false duplications (Supp. Fig. 2G) and losses (Supp. Fig. 2H). The Trio HiFi assembly had the lowest amount of false duplications, whereas the Hi-C phased HiFi assembly had the lowest amount of false losses. The CLR-based assembly also contained more false gene gains and losses (Supp. Fig. 5). For example, the ITSN1 gene, which is associated with autism-spectrum disorders[17], was found to be partially duplicated in the CLR assembly (Supp. Fig. 2I), with 35 out of 39 coding exons found on two distinct contigs. Furthermore, for consensus quality evaluation (QV), we used k-mers from 10X short read datasets with Merqury[11], and found that the HiFi Trio assemblies had consistently higher QV values (48.2 for maternal, 48.0 for paternal; Supp. Table 3) compared to CLR-based assembly (39.4 for the primary). We also compared the phasing level between the CLR and the HiFi assemblies in the zebra finch and observed a better phasing overall for the HiFi based assemblies, with a phased block N50 of 6 Mbp for the solo HiFi assembly and 6.3 Mbp for the assembly with Hi-C (hap1), versus 1.5 Mbp in the CLR assembly (primary). The trio assembly demonstrated a higher level of phasing and contiguity compared with assemblies produced with the solo or Hi-C versions of the pipeline (Supp. Note 2.2), with phased block N50 of 29.8 Mbp and 40.7 Mbp for the paternal and maternal haplotypes respectively. It is important to note that despite the improvement of HiFi assemblies over CLR assemblies, some HiFi libraries prove difficult to assemble due to the higher sensitivity of the technology to sequencing bias in certain genomic contexts (Supp.Fig 14).

2.2.2 Hi-C data enables chromosome-level phasing

We tested whether trio-based phasing with parental sequencing data generates an assembly with more complete phasing than using Hi-C data[9]. Consistent with expectations, the trio assembly had the fewest contigs with mixed hapmer content (defined here as a contig having >10% paternal and >10% maternal hapmers), with two contigs from the maternal and two contigs from the paternal assemblies being mixed. The average size of these four contigs was 152,914 bp—an order of magnitude smaller than the average contig size of ~1.5 Mbp. The Hi-C-phased assemblies had fewer contigs with mixed content (11 in hap1 and 7 in hap2), compared to the Solo haploid assemblies (26 and 1 in primary and alternate, respectively; Supp. Fig. 2D, Supp. Fig. 8). The size of contigs with mixed hapmer content was also an order of magnitude smaller in the Hi-C-phased assemblies (average size of mixed contigs being ~0.87 Mbp compared to ~1 and ~2 Mbp for hap1 and hap2, respectively) compared to the Solo assemblies (average size of mixed contigs ~6.8 Mbp compared to ~1.6 Mbp for the primary). Since Hi-C data provide relative phasing information, but do not give actual haplotype-of-origin information, it will not produce consistent phasing across separate chromosomes. Consequently, each haplotype in the Hi-C-phased assemblies has a mix of contigs from either the maternal or paternal haplotype, but the switch error rate[8] within the Hi-C phased contigs is similar to that of the trio phased contigs (~0.0016 and ~0.0008 versus ~0.0018 and ~0.0007 for hap1 and hap2 for Hi-C and Trio, respectively; Supp. Table 3). After scaffolding each set of phased contigs separately with Hi-C and Bionano data, we observe that most of the scaffolds do not contain mixed hapmer content. This indicates that the Hi-C phasing from the start succeeded in properly binning whole chromosomes (Supp. Fig. 8), confirming what was previously reported for low-heterozygosity human data[9]. Due to its higher level of phasing and contiguity, we curated the Trio assembly to establish a new reference genome for zebra finch.

At the time of publication, the VGP sequenced and assembled 51 species using the VGP-Galaxy pipeline. For these assemblies, the Hifiasm module was used for contiging and YaHS or SALSA modules were applied for scaffolding. Hi-C or parental short reads were used for haplotype phasing. These genomes were found to have a wide range of sizes (590 Mbp–8.5Gbp), repeat content (6%–73%), and heterozygosity (0.001%–1.73%). One of the species, the royal ground snake (Erythrolamprus reginae), was found during curation to be triploid, and has heterozygosity ranging from 0.422% to 3.74% depending on haplotype configuration (Supp. Table 5). To evaluate the overall quality of assembled genomes, we looked at gene- and k-mer-based completeness measures. On average, the assemblies showed ~96% gene and ~99% k-mer completeness with ~1.93% of genes appearing as duplicated. The ratio of assembly lengths to the genome sizes estimated from k-mer profiling of the HiFi reads ranged between 0.84 and 1.24, with fishes exhibiting >1 ratios and birds <1. When we performed a similar analysis using Illumina sequencing of the same individuals for two species exhibiting HiFi-based high ratios (~1.21 and 1.16 for Scomber japonicus and Podargus strigoides, respectively), the results were much closer to the observed assembly size (1.03 and 1.06, respectively). These findings suggest that some signal in the PacBio HiFi reads results in lineage species differences in k-mer estimated genome size and that this signal does not appear to be present in the Illumina k-mers estimate for genome size.

2.2.3 Bionano Scaffolding marginally improves assembly quality when using Hi-C and YaHS

To evaluate the best scaffolding strategies using the current tools, we applied different scaffolding combinations on the HiFi generated contigs of the zebra finch; the Tammar wallaby (*Macropus eugenii*), a small marsupial macropod native to Southern and Western Australia; and the Snub-nosed viper (*Vipera latastei*), a threatened viperid snake species living in the Iberian Peninsula and Northern Maghreb[18]. The wallaby is of particular interest for scaffold-level assembly due to the large size of its chromosomes (~100-730 Mbp in marsupials versus 50-250 Mbp in humans[19]). We tested combinations of Hi-C scaffolding tools (SALSA2[20] and YaHS[21]) with and without a Bionano optical mapping scaffolding tool (Bionano Solve tool), evaluating the number of gaps and scaffolds, as well as NG50 and auN[22] (Supp. Fig. 9, Supp. Table 4). The combination of Bionano Solve and SALSA2 yielded a higher assembly contiguity than scaffolding with Bionano Solve or SALSA2 independently.

However, YaHS performed similarly with and without Bionano information (auN values of 739.7 Mbp versus 732.5 Mbp for the wallaby primary assembly), and overall better than SALSA2. Scaffolding results in different taxonomy groups and genome sizes were comparable, overall leading to chromosome level assemblies (Supp. Fig. 9, Supp. Table 4). Our comparison also shows that when scaffolding Hifiasm generated contigs, the use of Bionano data only provides a marginal improvement when used with Hi-C YaHS scaffolding.

2.2.4 Contaminant analysis

The VGP-Galaxy decontamination pipeline is a new, automatic pipeline that we developed to remove exogenous sequences from assemblies. It was built and initially validated using 19 assemblies, which were chosen because they were decontaminated with the previous standard protocol (a combination of manual and automated screenings). The presence of contaminants or mitochondrial sequences was known and assumed to be ground truth (Supp. Fig. 10, Supp. Table 6). Contaminants represented a negligible fraction (<0.19%) of the assembled sequences (Supp. Fig. 10a, right side) and the contaminated scaffolds were generally small, with the largest being 0.15 Mbp, and the median number of bases removed was 61 kbp. Considering only assemblies with misclassified scaffolds (n = 5), where misclassification means the pipeline classification (assembly, contaminant or mitochondrial sequence) does not match the ground truth classification, the median size of false negative and false positive bases was 7.2 kbp (0.00051%) and 12.8 kbp (0.0012%), respectively.

The five assemblies where our decontamination results differed from the benchmark include the primary assembly of the Great White shark, the paternal haplotype of the Bottlenose dolphin, the maternal and paternal haplotypes of the Budgerigar and the primary assembly of the Maguari stork (Supp. Fig. 10a, right side). The largest incidence of misclassification was observed in the Maguari Stork, where there were two contaminant sequences (both *Neospora caninum*) in the benchmark and neither were identified, but the scaffolds only represent 0.0065% of the length of the genome (81.7Kb/1.25Gb). N. caninum and other parasites are likely not present in the Kraken2 database, and thus would not have been identified by the pipeline. In the case of the Budgerigar, two scaffolds in the maternal assembly and one scaffold in the paternal assembly were false positive classifications relative to the benchmark. The benchmark set is based on the prior VGP's established decontamination process, which involves manual and automated systems. While the results of this process are generally reliable, it is possible that true contaminants were not originally identified and thus included in the benchmark set. Based on a BLAST query against NCBI's nucleotide (nt) database, these three scaffolds are likely true contaminants with high similarity (>99% coverage) to Delftia species. The paternal assembly also contained three false negative classifications. These false negative scaffolds contain large segments of homopolymers and low-complexity repeats. These would have been masked in the assembly submitted to Kraken2, and it is possible that the remaining sequence was not long enough for reliable classification. When isolated, these non-repeat regions show high sequence similarity (>99%) to a synthetic construct from PacBio, which is a positive control sequence used for determining run success, so these are likely true contaminants that the pipeline did not identify. Interestingly, a BLAST search based on alignment coverage identified 150 contaminants from the Budgerigar, Bottlenose dolphin, Bolson tortoise and the African grass rat as almond (*P. dulcis*, accession AP021729). Further investigation revealed that this was caused by residual contamination from the PacBio internal control sequence (accession MG551957) in the published almond genome. In total, 265 contaminants were identified as the PacBio control sequences.

After validation, the decontamination pipeline was used on 32 assemblies among the 51 described in this paper (Supp. Table 7). Of the 32, three assemblies contained foreign contaminants and five contained mitochondrial sequence. Kraken2[23] uses a lowest common ancestor approach to classify contaminants, which can result in a high-level taxonomic classification. To better understand which foreign contaminants were identified, all contaminant sequences classified (from the 19 validation and additional three assemblies) were compared against the non-redundant nucleotide (nt) database, excluding one repetitive contaminant scaffold from *Erythrolamprus reginae* (rEryReg1) that was not masked by dustmasker and thus classified as a contaminant. The contaminants belonged to five different taxa, with Escherichia being the most abundant (Supp. Table 10, Supp. Material 3). Overall, these results show that our initial sequencing data had only trace amounts of contamination, and that the automated decontamination pipeline produces results similar to those as the more labor-intensive decontamination processes previously employed by the VGP.

2.2.5 Mitogenome assembly and evaluation

Assembling complete and accurate mitochondrial genomes often requires a distinct approach compared to nuclear genomes[24], but can support species identification and other analyses of mitochondrial evolution. Therefore, we implemented MitoHiFi (v. 2.2 and v. 3)[25] as part of the VGP-Galaxy pipeline v2.1 to validate species identification and provide mitochondrial references. We tested MitoHiFi on the 51 species that were used for nuclear assembly. We then used the Barcode of Life Data System (BOLD) ID Systems[26] for species identification and MITOS2 to annotate the reference as a functional evaluation (Supp. Note 3.6). We assembled 25 mitogenomes successfully and validated by using MITOS2 annotation (Supp. Table 9). Mammals had the highest success rate (11/14), while fishes, amphibians, reptiles and birds had lower fraction of mitochondrial assemblies (3/5, 1/4, 4/8, 4/15 respectively). These results are congruent to those previously noted for PacBio CLR data[24] in that the availability of mitochondrial reads appears to be a function of tissue type from specific taxonomic groups, as well as long-read library cut-off insert sizes smaller than 20 kbp (Supp. Table 9). For example, bird blood contains much fewer mitochondria than muscle, and vertebrate mitochondrial genome size is on average about 16 kbp. Several assemblies also contained well-supported mitochondrial gene duplications (particularly in the control region OH) or missing tRNA genes (Supp. Table 9), as previously reported for other species[24]. There were three cases where MitoHiFi failed to assemble a functional mitochondrion, and further investigation yielded a proper mitogenome assembly: bMorBas2 (Morus bassanus), mMusLut2 (Mustela lutreola), mDasNov1 (Dasypus novemcinctus), and mCynVol1 (Cynocephalus volans). For bMorBas2, the problem was that Morus is the name of both the genus for gannets as well as mulberries, so the program had initially used a mulberry sequence as reference. For mDasNov1 and mMusLut2, the program selected a concatamer assembled from genuine mitochondrial reads, so this was fixed by removing the duplicate sequence. For mCynVol1, the MitoHiFi-produced assembly worked for BOLD [26] identification, but MITOS2 indicated internal stops in COX1, COX2, ATP6, COB, NAD3, and NAD4, and further the mitochondrial assembly selected was a linear contig that corresponded to a 33 kbp linear unitig. Inspection of the raw unitig graph generated by the hifiasm step in MitoHiFi indicated the presence of a 16,762 bp circular unitig, a more likely candidate for a mitogenome. These results allow us to conclude that, while HiFi reads provide an excellent approach to generate complete and accurate mitogenomes, scaling up to large volumes of samples (e.g., for DNA barcoding purposes) will require careful consideration of experimental design in terms of sample selection and library preparation in order to retain mitochondrial reads.

2.3 From new assemblies to biological insights

The phylogenetic breadth of genomes assembled here provides unique opportunities for studying evolution of their structure and function. While normally that would require generating multiple genome alignments—a complex and computationally expensive process—computation power underlying global Galaxy instances provides means for rapid analysis of dozens of new genomes that are in the process of generating RefSeq/ENSEMBL gene annotations. To demonstrate this, we generated a preliminary annotation of X-box protein 1 gene (*XBP1*) across our new assemblies, a transcription factor involved in regulation of protein folding of other proteins in regulatory regions of DNA, to trace down its evolution. We chose this gene for two reasons. First, its unique structure: XBP1 mRNA is cleaved by a specific endonuclease, IRE1α, which removes a 26 bp spacer located within exon 4 (Supp. Fig. 11a) altering the phase of the reading frame downstream of the cleavage site[27]. This unique feature frequently renders this gene misannotated in genome assemblies. Second, the human genome contains an additional pseudogene copy of this locus. Our genome assemblies may shed light on when this duplication has occurred.

To identify this gene within assemblies described here, we developed a workflow that finds reading frames with high similarity to exons of interest (see Methods). Interestingly, we found the gene experienced duplications in Philippine flying lemur (Cynocephalus volans), tammar wallaby (Macropus eugenii), and human (Homo sapiens; based on the analysis of the latest T2T-CHM13v2 assembly[6]). In wallaby and human, only one copy is functional with the second copy being truncated and thus pseudogenized. However, in Philippine flying lemur there are likely to be two functional copies with distinct structures (Supp. Fig. 11b) located within the same linkage group (scaffold2) and separated by over 180 Mbp. Copy 1 appears to maintain the ancestral gene structure in comparison with the human orthologue, while Copy 2 lacks introns and appears to be a transposed copy (Supp. Fig. 11c). It likely arose via duplication followed by transposition of the reverse-transcribed uncleaved form of *XBP-1* mRNA in a manner similar to the evolution of rodent insulin II gene[28,29]. Both reading frames are present and appear as contiguous segments. Copy 1 and Copy 2 are ~95%

identical at the nucleotide level with two stem-loop structures required for IRE1α cleavage[30] being absolutely conserved. The read depth in the vicinity of both copies appears to be uniform, ruling out assembly artefacts.

3. Methods

3.1 Mitogenome assembly and validation (Workflow 0)

To assemble mitogenomes from long reads, we implemented MitoHiFi in Galaxy[25] (https://github.com/marcelauliano/MitoHiFi), which included creating a Docker image for MitoHiFi In addition to generating a functional mitogenome, this step also serves as a confirmation for taxon identification. All assemblies were run with default parameters. Validation consisted of confirming the species identification as well as functional validation.

We tested MitoHiFi on 51 HiFi datasets. MitoHiFi uses a reference mitochondrial assembly to select HiFi reads that map to the reference, which is suggestive of their putative mitochondrial origin (though nuclear mitochondrial DNA — NUMTs — and other nuclear material can be a confounding factor) [24]. For species identification, we used the Barcode of Life Database (BOLD) identification engine to search the resulting mitogenome against all BOLD cytochrome c oxidase I (COI) barcode records, a standard method for species identification. We used BOLD ID Systems [26] based on a 648-bp region of the cytochrome c oxidase I (*COI*) gene. We used MITOS2's functional annotation[31] to evaluate the assembly quality with regard to missing genes or duplications. Resulting assemblies fell into three categories:

1. Complete and accurate assemblies: proper species identification in BOLD and no strong peculiarities reported by MITOS2 (25/51);
2. Further investigation (1/51);
3. No Assemblies: MitoHiFi failed to produce an assembly at all (25/51).

When MitoHiFi failed, it was usually due to a lack of true mitochondrial reads resulting in a failed assembly attempt; though there were cases when a poor quality mitochondrial assembly was generated from sequences likely belonging to NUMTs. NUMTs can be identified by their usually linear sequences longer than 16 kbp [24].

3.2 Genome profiling (Workflows 1 and 2)

In Workflow 1 (for HiFi reads only) and Workflow 2 (For HiFi reads and parental illumina reads), *k*-mers are counted using Meryl[32], and the *k*-mer profile is analyzed using GenomeScope2[33]. The input of this workflow is a collection of HiFi reads in FASTQ format[34], the *k*-mer size, and ploidy.

3.3 Phased genome assembly and duplicate purging (Workflows 3 - 6)

The logic of the pipeline (Supplemental Fig. 12) is to progressively refine and complement the initial assembly graph, taking advantage of the graph-based analysis and other functions introduced by gfastats [35]. The final product is a scaffolded assembly graph in the GFA1.2 format [36]. This approach constitutes a new conceptual framework in genome assembly, since it avoids the loss of information resulting from collapsing the assembly graph to linear sequences.

The initial contigging is done using Hifiasm[10] (https://github.com/chhylp123/hifiasm). Hifiasm workflows generate de novo genome assemblies using three potential methods of phasing. The first mode, HiFi-only (Workflow 3), uses only the HiFi reads to build a haplotype-aware assembly. The primary and alternate assemblies are then purged (Workflow 6) using purge_dups[12] (https://github.com/dfguan/purge_dups). The coverage histogram can be used to manually adjust cutoffs, if necessary. The second mode, Hifiasm-Hi-C (Workflow 4), uses Hi-C data to improve haplotype phasing. The third mode, Hifiasm-trio (Workflow 5), uses parental information to fully phase the haplotypes. All contigging workflows include quality control reports, using , BUSCO[37] (https://github.com/WenchaoLin/BUSCO-Mod), and Merqury[11] (https://github.com/marbl/merqury). For all 51 species presented in this work, we used the "vertebrata" lineage when running Busco. For non vertebrate species, the user can select a different lineage. On all test datasets Hifiasm was run with default parameters, except for the HiFi-only workflow where we turned off internal purging.

3.4 Bionano scaffolding (Workflow 7)

The Bionano scaffolding workflow uses Bionano Solve[38] for scaffolding and gfastats[35] (https://github.com/vgl-hub/gfastats) for quality control. On all test datasets, Bionano Solve was run with default parameters and without contig breaking (i.e., excluding conflicting contigs during NGS-map conflicts).

3.5 Hi-C scaffolding (Workflow 8)

The Hi-C scaffolding workflow can be used either on the primary contigs or on the scaffolds from Bionano scaffolding. Hi-C reads are aligned and prepared for scaffolding using the Arima mapping pipeline. Then YaHS[21] v1.2a.2 (https://github.com/c-zhou/yahs) is used for scaffolding. Quality control is done with gfastats, BUSCO and PretextMap (https://github.com/wtsi-hpag/PretextMap) to visualize Hi-C contacts before and after scaffolding.

3.6 Decontamination pipeline (Workflow 9)

Masking is performed with dustmasker from NCBI Blast+ v.2.6.0 using dust level 40. Kraken2 v.2.1.1 identifies non-target contaminants with a confidence level of 0.3. Mitochondrial scaffolds are identified with blastn (v.2.6.0) and the NCBI Refseq mitochondrion database (release 212). To ensure a scaffold is entirely mitochondrial and not a NUMT, a custom tool (parse_mito_blast, v. 1.0.1) takes all alignments between unique scaffold-accession number pairs and calculates the total alignment coverage taking overlap into consideration. The threshold for classifying a scaffold as mitochondrial is 95% alignment coverage. A compiled list of contaminant and mitochondrial scaffolds is passed to the gfastats (v.1.2.2) --exclude-bed function to remove these scaffolds and remaining adaptors, and generate a new FASTA file. Eleven assemblies were used to test the pipeline during development. Validation was performed on an additional eight assemblies, including three that have no contamination or mitochondrial sequence.

3.7 Manual re-binning of trio assemblies

For the female zebra finch, since the problematic contigs from the maternal assembly (hap2) were absent from the paternal haplotype (hap1), we approached this by first trying to find the contigs that contain 2-copy k-mers which were present only in the maternal haplotype using meryl and meryl databases for each haplotype:

```
meryl print difference bTaeGut2_hap2_count bTaeGut2_hap1_count
output bTaeGut2_hap2only

meryl equal-to 2 bTaeGut2_hap2_only output
bTaeGut2_hap2_only_equalto2

meryl-lookup -existence -sequence
bTaeGut2_trio.asm.dip.hap2.p_ctg.fa -mers
bTaeGut2_hap2_only_equalto2 > bTaeGut2_hap2_only_equalto2.tsv
```

This TSV (tab-separated values) file contained all the contigs in the maternal haplotype, the contig's size (in 21-mers), and how many k-mers were seen in both the contig and the database of 2-copy k-mers present only in the maternal haplotype. We used this data to calculate the percentage of each maternal contig that belonged to the database of problematic k-mers, and we focused on contigs with over 50% of their k-mers matching that database. We then used the original GFA output from Hifiasm to find the reads that were used to build these contigs. We created manual re-binning lists that assigned these reads as "ambiguous" and re-run Hifiasm with the trio data as well as these re-binning lists.

3.8 Assembly comparisons

For the zebra finch assembly comparisons, we used the three primary assemblies (CLR, Solo, and Solo w/Hi-C mode) and one rebinned paternal assembly (Trio mode) made immediately after contigging and purging (https://genomeark.s3.amazonaws.com/index.html?prefix=species/Taeniopygia_guttata/bTaeGut2/). All assemblies were masked by repeatmasker (https://www.repeatmasker.org/; [39] with default engine and commands "`-species 'Taeniopygia guttata' -xsmall -s -no_is -cutoff 255 -frag 20000`" before genome alignment. The reads produced for the assemblies by PacBio CLR, HiFi and 10X platforms were mapped to the all genome assemblies by Minimap2[40] and EMA mapper[41]. We used parameter "`-ax map-pb`" `for PacBio CLR read, and` "`-ax map-hifi`" for HiFi read mapping using Minimap2. The paired-end 10X reads were mapped using the barcodes default options of EMA[41]. The reads without barcodes were mapped using BWA[42] with parameters "`-p -M -R '@RG\tID:rg1\tSM:sample1'`" following guidelines in EMA. Intermediate BAM files produced in the read mapping step were merged by Sambamba[43]. Samtools[44] was used to sort the BAM files and to calculate read coverages of each genomic position.

We calculated k-mer duplications of each assembly using a script "`false_duplication.sh`" in Merqury[11] with optimal k-mer size of the zebra finch genome: 21. We calculated k-mer collapse and expansion with Merfin[45] using the same k-mer size. To compare the k-mer collapses and expansions of diploid assemblies, we included the maternal or alternate zebra finch sequences with the paternal or primary sequences. For optimum K* calculation, we included "lookup_table" produced by GenomeScope2[33] with the 10X reads. We included a current reference genome of zebra finch assembled from CLR reads (bTaeGut1.4; GCF_003957565.2) in GenomeScope profiling. We estimated k-mer duplications and completeness of default-mode and rebinned trio assemblies using both paternal and maternal assemblies without purging with 10X reads using Merqury.

We identified false duplications and losses using whole genome alignment with estimation of number of paralogs in alignment blocks. Firstly, we aligned the three primary assemblies (CLR, Solo, and Solo w/Hi-C mode) and the one paternal assembly (Trio mode) of the zebra finch using the Cactus alignment tool[46]. Then we extracted homologous regions to a readable multiple alignment format using HAL[47]. Because all assemblies were made from the same sample, the number of paralogs of each assembly in each alignment block should be the same, and when not the same, they are false duplications or losses occurred in one of the assemblies. To every alignment block showing this discordance of the number of paralogs between the assemblies, we calculated the likelihood of each number of paralogs to model (i.e. how many paralogous sequences will be present in the alignment block) based on summed read-coverage of the PacBio CLR, HiFi, and 10X reads. The likelihood of each model was calculated as

$$L(\theta \mid x) = \Sigma ln \frac{1}{\sigma\sqrt{2\pi}} exp(- \frac{(x-\mu)^2}{2\sigma^2})$$

where x is the sum of mean depth of each homologous sequences in an alignment block from an assembly, and μ and σ are the parameters of depth distribution estimated from each number of paralogs models. To estimate the model parameter μ and σ, we calculated the mean and variance of normal distribution from depth coverages of genomic regions that there is no multi-copy k-mer for the model that the number of paralogs is zero, then simply multiplied the mean by integer for each model, e.g. multiplying the mean by 2 for 1-paralogs model. We supposed that the variance is the same for all models. The false duplications and losses of each assembly were identified when each assembly had more or less paralogs than a best model from the likelihood estimation in each alignment block. To remove the noise for the false duplications and losses, we filtered out false duplications on contigs where false duplications occupied <50% of the contig length, and far from terminals of the contig (>20kbp), and filtered out false losses under 1kbp length. To avoid false losses include haplotype differences we calculated K* of k-mers in the region of candidate false losses after noise filtering mentioned above. We only included the candidates to false losses when a candidate has collapsed k-mers (K* >0) above 90% of the genomic sequences. Moreover, we estimated potential false gene gains and losses based on annotation data of bTaeGut2.trio (GCF_008822105.2) and the erroneous regions we identified in each assembly. We aligned the bTaeGut2.trio assembly and others together using Cactus [46], then, potential false gene gain or losses were identified when the false duplications and losses had homologous regions with any CDSs of bTaeGut2Trio annotation.

3.9 Comparative gene analysis

For each genome we annotated all open reading frames (ORFs) ≥99 bp (defined as uninterrupted runs of sense codons bound by stops) using orfipy[48]. Next, we used amino acid translations of ORFs to create a Diamond[49] database. We then queried the database using amino acid translations of human exons to identify the most likely location in assembled genomes and plotted this information using Galaxy/Jupyter integration to generate Fig 6. The Galaxy history with the step-by-step description of this process is available at https://gxy.io/GTN:T00174.

Supplementary Figures

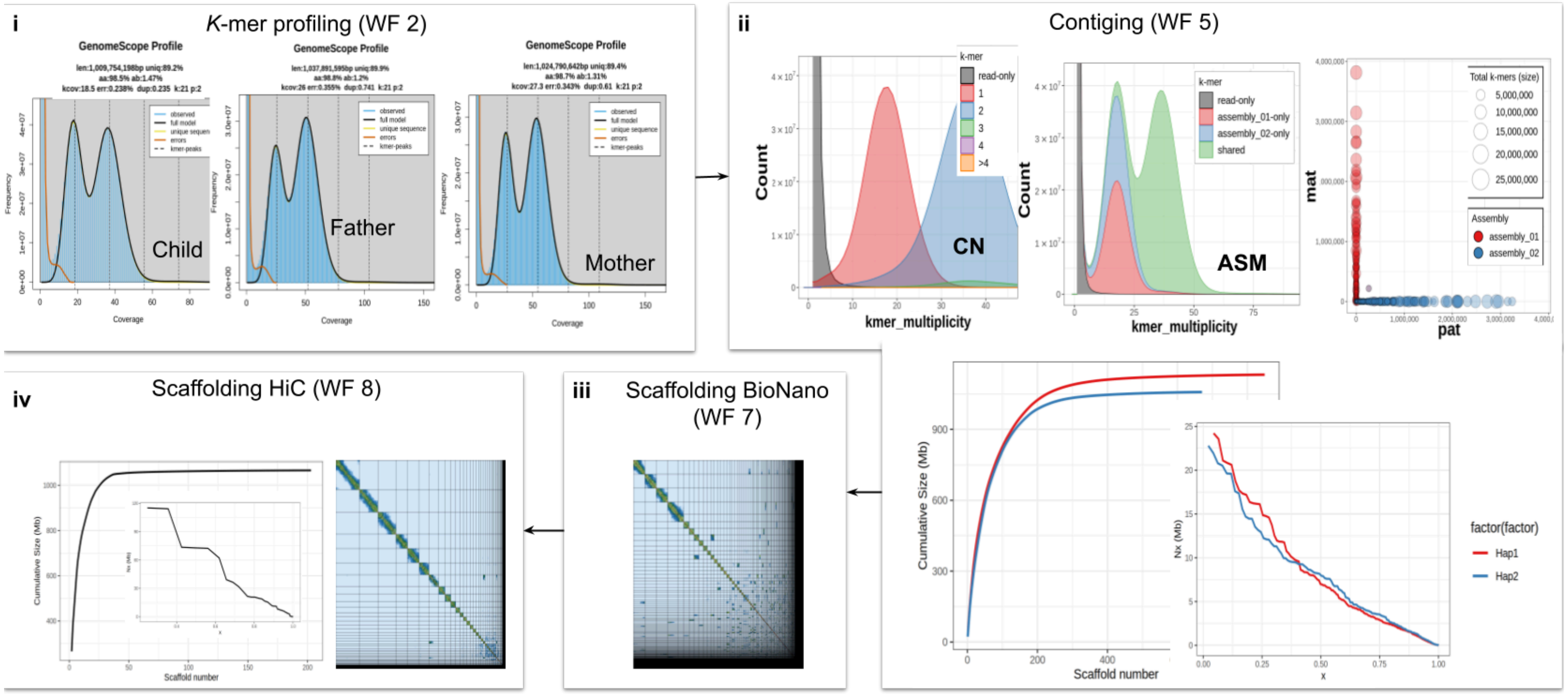

i
K-mer profiling (WF 2)
GenomeScope Profile
len:1,009,754,198bp uniq:89.2%
aa:98.5% ab:1.47%
kcov:18.5 err:0.238% dup:0.235 k:21 p:2
Child
GenomeScope Profile
len:1,037,891,595bp uniq:89.9%
aa:98.8% ab:1.2%
kcov:26 err:0.355% dup:0.741 k:21 p:2
Father
GenomeScope Profile
len:1,024,790,642bp uniq:89.4%
aa:98.7% ab:1.31%
kcov:27.3 err:0.343% dup:0.61 k:21 p:2
Mother
observed
full model
unique sequence
errors
kmer-peaks
Frequency
Coverage
ii
Contiging (WF 5)
Count
kmer_multiplicity
k-mer
read-only
1
2
3
4
>4
CN
k-mer
read-only
assembly_01-only
assembly_02-only
shared
ASM
mat
pat
Total k-mers (size)
5,000,000
10,000,000
15,000,000
20,000,000
25,000,000
Assembly
assembly_01
assembly_02
Cumulative Size (Mb)
Scaffold number
Nx (Mb)
x
factor(factor)
Hap1
Hap2
iii
Scaffolding BioNano (WF 7)
iv
Scaffolding HiC (WF 8)
Cumulative Size (Mb)
Scaffold number

Figure 1. QC generated by the pipeline using Zebra finch trio data supplemented with Bionano and Hi-C data. i. GenomeScope profiles using 21-mers on child HiFi data and parental Illumina data. ii. Merqury profiles (upper three graphs) and gfastats continuity metrics (lower two graphs) following contiging. The CN plot shows the number of copies of k-mers in both assemblies. Single copy k-mers correspond to heterozygous regions, and two copies to homozygous regions. This plot provides information about the potential need to purge the assembly if it shows the presence of k-mers in three or more copies. The ASM plot in Bii middle highlight the finding that paternal and maternal assemblies share k-mers at ~40✕ range that corresponds to the diploid coverage (also indicated by the rightmost peak in the child GenomeScope profile in panel). This is consistent with the expectation that phased assemblies will split the homozygous regions of the genome between the two haplotypes. Accordingly, heterozygous content in the genome is split mostly evenly between the two haplotypes, shown by the similar-sized assembly-specific peaks at ~20✕ (the maternal peak is slightly bigger due to the presence of the Z chromosome, which is larger than the W in the paternal assembly). The assembly blob plot (rightmost graph in Bii indicates excellent stratification of hapmers (k-mers specific to a particular haplotype) across maternal (red) and paternal (blue) assemblies. iii. Pretext map of maternal assembly after Hi-C scaffolding shows increase in continuity compared with the panel. iv. General statistics and pretext map of the maternal scaffolding after a second scaffolding step using Hi-C data. The map shows improvements compared to the previous scaffolding step, with fewer scaffolds and less physical proximity between scaffolds (non-diagonal Hi-C signal).

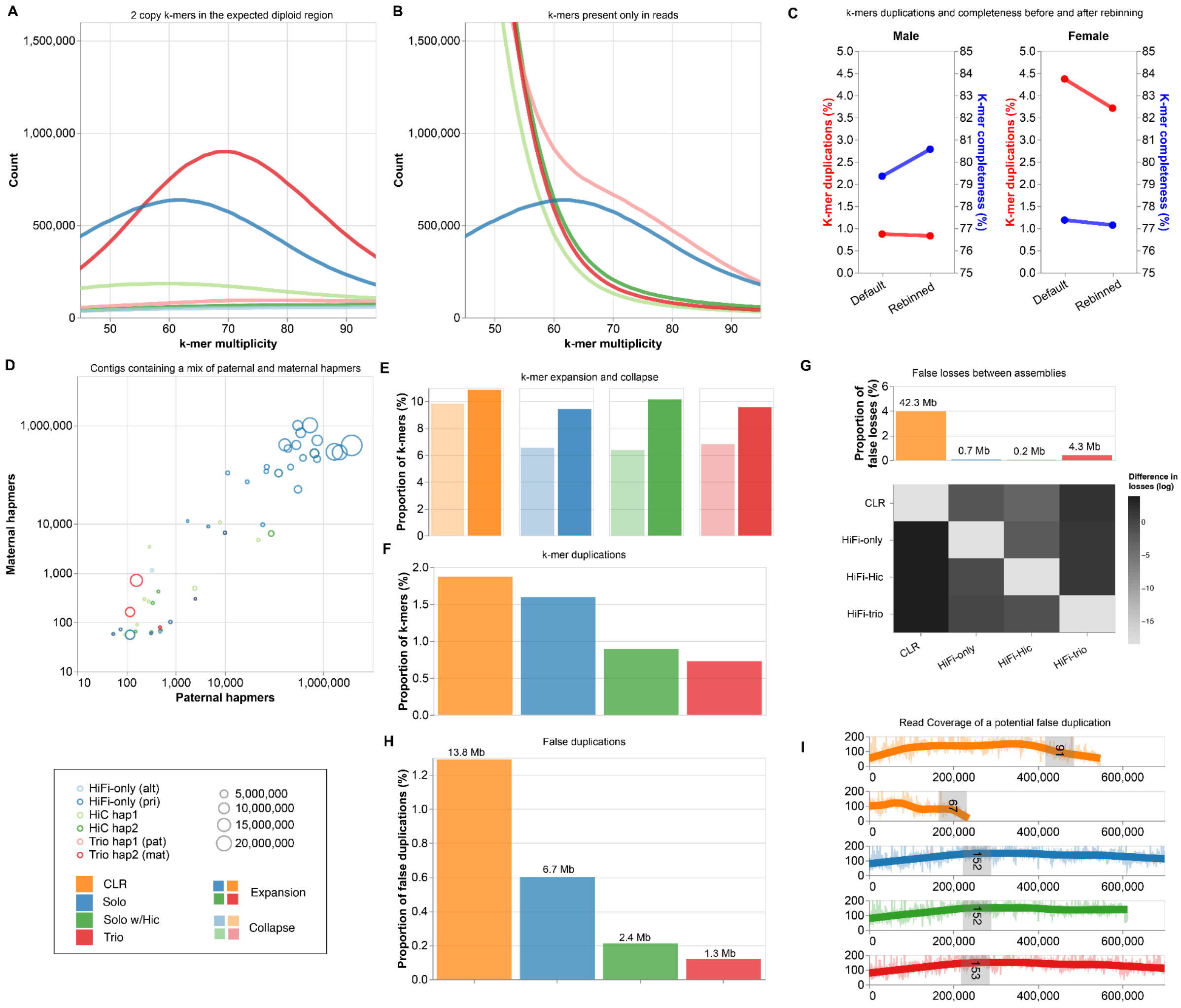


Figure 2. Extended evaluation of HiFi zebra finch assemblies using the three modes of the VGP pipeline in Galaxy: “Solo” (pseudohaplotype primary/alternate), “Solo w/Hi-C” (Hi-C-based phasing), and “Trio” (phasing using parental reads). A. shows 21-mers at 2-copy in the assembly with k-mer multiplicity between 50 and 90, a range that includes the expected diploid coverage region. B. shows k-mers that are present in only the reads, which suggests that the regions they represent (about 8 Mbp of sequence that ended up re-binned) are missing from the assembly, relative to the assembly in question. The purged Solo primary assembly has some missing regions, likely due to imperfect purging (k-mer multiplicity <5 excluded). Panels A and B are zoomed-in sections of the full k-mer spectra, to highlight regions of interest. C. Comparisons of k-mer duplication (red) and completeness (blue) between default and rebinned trio assemblies in males (left) and females (right). D. shows contigs from each assembly plotted according to how many parental hapmers are present in the contig, with contigs that were either fully phased (>95% of either parental k-mer) or lacking informative phasing information (i.e., less than 50 paternal and less than 50 maternal k-mers) excluded. The size of each bubble is proportional to the total k-mer size of the contig. Contigs along the diagonal have a mixed representation of hapmers from both parents, indicating intra-contig switch errors. Of these contigs, the Solo ones are typically larger and contain a higher amount of hapmers from both parents. E. Proportion of

k-mer expansion and collapse in each diploid bTaeGut2 assembly. F. Proportion of k-mer duplication in the bTaeGut2 assemblies. We calculated k-mer duplications from the primary assemblies (CLR, Solo, Solo w/Hi-C) and paternal assembly (Trio) from phased diploid assemblies. G. Proportion and cumulated size (in Mb) of false losses of each assembly (above), and heat map of the size (in Mb) of false losses identified between the assemblies (below) in log scale. H. Proportion and cumulated size in Mb units of false duplications of each assembly. I. A case of potential false gene gain in CLR assembly. Duplications of homologous sequences of partial ITSN1 gene was found in CLR assembly. Read depth coverage of contigs including the homologous sequences of ITSN1 gene in each bTaeGut2 assembly (highlighted in gray) is shown with a range from 0 to 200. The number in the gray highlighted region represents a mean depth coverage of ITSN1 homologous regions in each assembly.

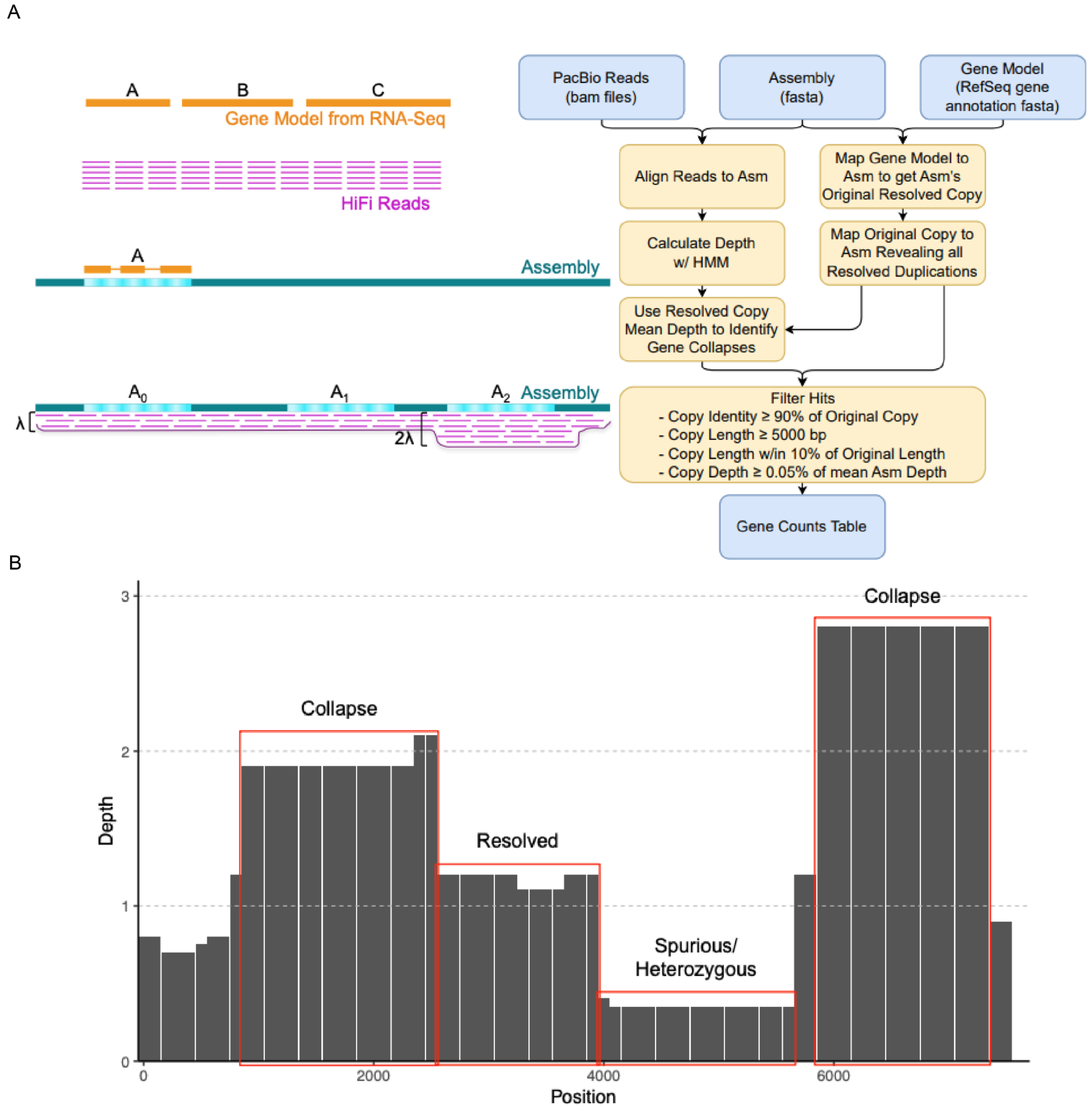

A
A
B
C
Gene Model from RNA-Seq
HiFi Reads
A
Assembly
A_0
A_1
A_2
Assembly
λ
2λ
PacBio Reads (bam files)
Assembly (fasta)
Gene Model (RefSeq gene annotation fasta)
Align Reads to Asm
Map Gene Model to Asm to get Asm's Original Resolved Copy
Calculate Depth w/ HMM
Map Original Copy to Asm Revealing all Resolved Duplications
Use Resolved Copy Mean Depth to Identify Gene Collapses
Filter Hits
- Copy Identity ≥ 90% of Original Copy
- Copy Length ≥ 5000 bp
- Copy Length w/in 10% of Original Length
- Copy Depth ≥ 0.05% of mean Asm Depth
Gene Counts Table
B
Depth
0
1
2
3
Collapse
Resolved
Spurious/
Heterozygous
Collapse
0
2000
4000
6000
Position

Figure 3. A. Method to obtain gene count tables of duplicated and collapsed genes. We aligned the PacBio HiFi reads to each assembly and used an hmm algorithm to calculate the read depth in sliding windows (Figure S1). We mapped the RefSeq gene Model to each assembly to get the assembly's original resolved copy. We then used the hmm depth to calculate the mean depth of resolved copies to identify gene collapses. We filtered the genes showing discordant depth to keep genes where the copy identity is above 90% of the original copy, the copy is longer than 5 kbp and within 10% of the original length, and the discordance in read depth is higher than 0.05% of the average assembly depth. **B**. Three types of regions are identified by the hmm algorithm: i) Resolved regions if the read depth is close to the assembly average depth, ii) Spurious duplications or heterozygous regions if the read depth is below the average, iii) Collapsed duplications if the read depths is twice the average or above.

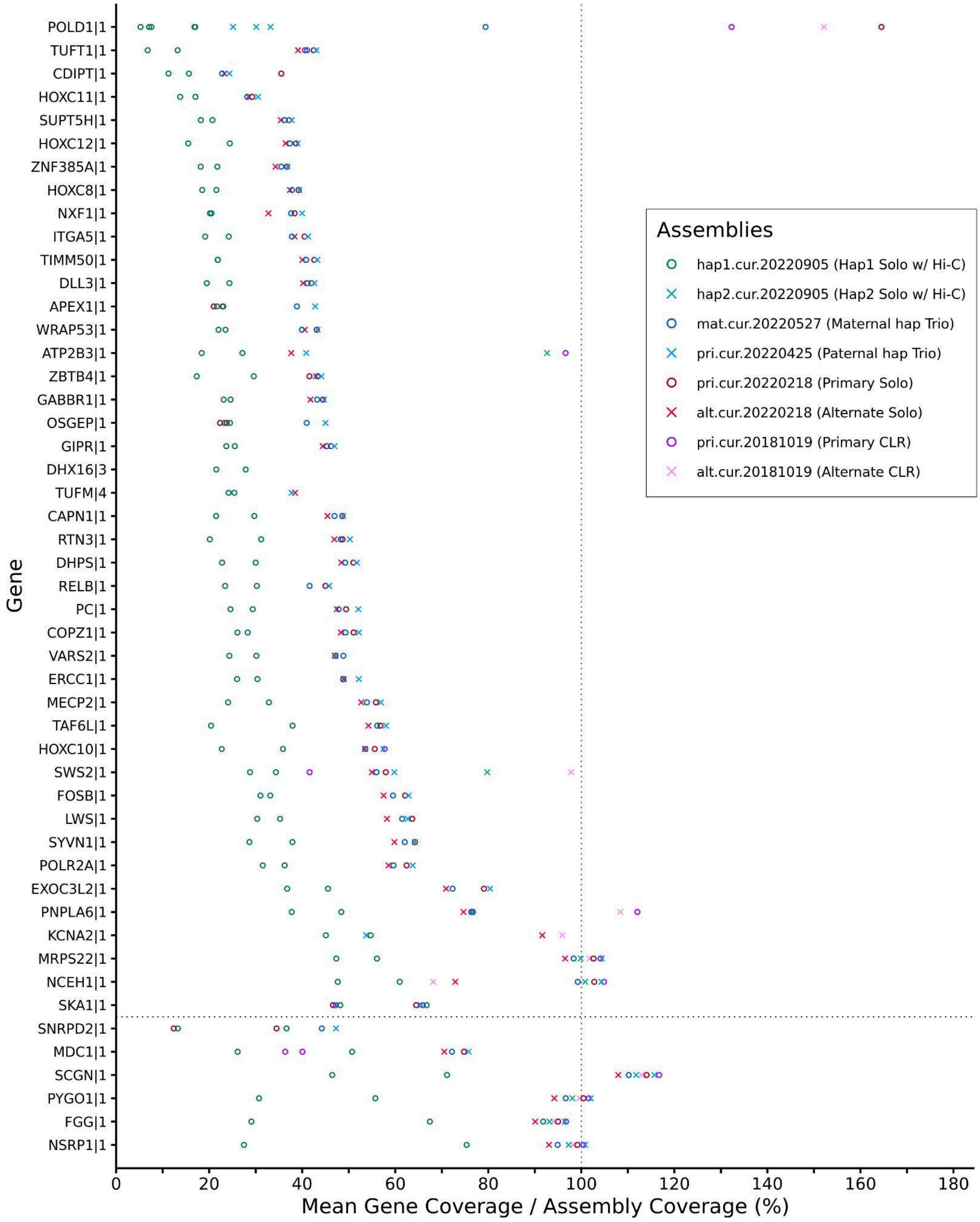


Figure 4. Spurious duplications in the Hi-C phased assembly of the zebra finch. Each data point of this graph represents a copy of the gene in the Y axis. The X axis represents the average read coverage relative to average coverage across the assembly. This graph shows that most of the spurious duplications occur in low coverage regions, and we can see that the two copies present in hap1 of the Hi-C assembly (green circles) have on average half the read coverage compared to the single copy in the other assemblies.

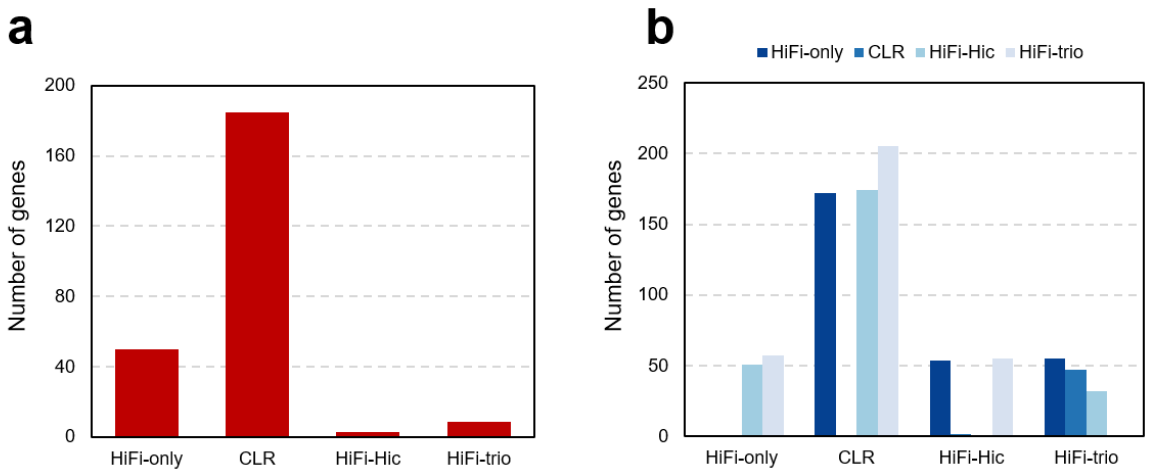


Figure 5. A. Number of potential false gene gains in each zebra finch (bTaeGut2) assembly. Potential false gene gains are calculated for individual genomes based on false duplication identified by read coverage of each assembly. B. Number of potential false gene losses in each bTaeGut2 assembly (X-axis) estimated from other bTaeGut2 assemblies (colours). False losses are estimated by comparing missing regions between two assemblies from whole genome alignment.

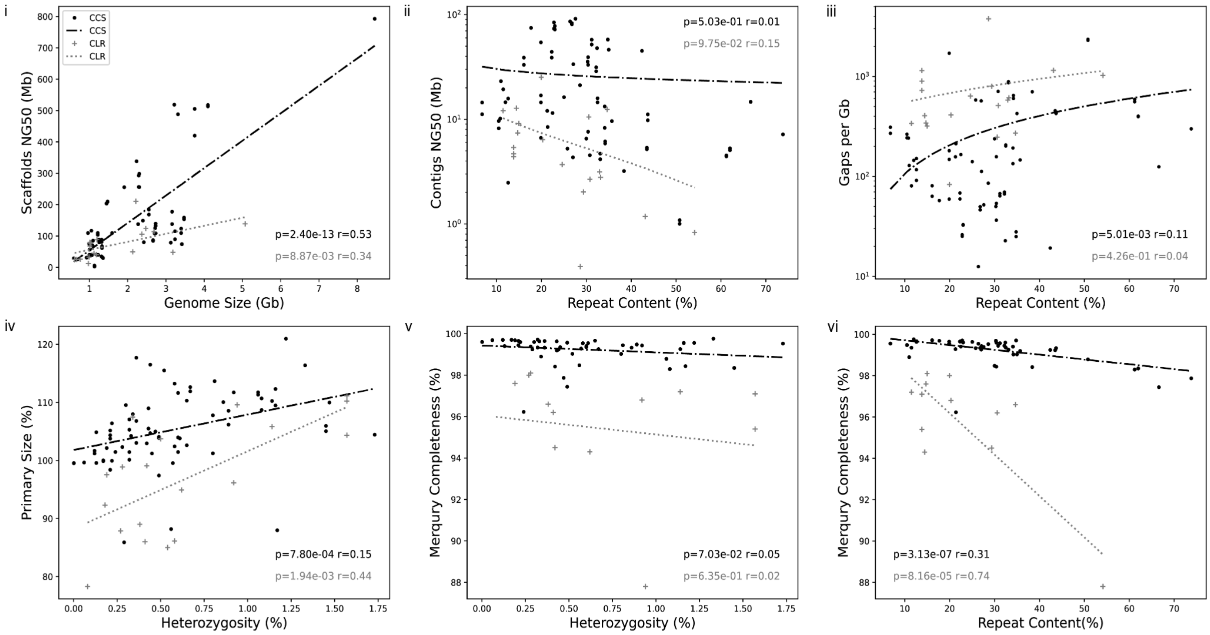


Figure 6. Comparison of assembly statistics between sequencing technologies and scaffolding modes. Panels i to vi compare assembly statistics between HiFi technology in black, used in this study, and CLR technology in grey, used in the previous version of the VGP assembly pipeline. Each dot represents an assembled species (Primary or Hap1), and the lines represent the linear regression for each technology. i. Scaffold NG50 in Mb in relation to Genome size in Gb. ii. Contigs NG50 in relation to repeat content, iii. Gaps per GB in relation with repeat content, iv. Size of the primary assembly in percent of the

estimated genome size in relation with the heterozygosity rate of the genome, v. Genome completeness estimated by Merqury (both haplotypes together) in relation with the heterozygosity rate (outlier *Amblyraja radiata*), vi. Genome completeness in relation with repeat content.

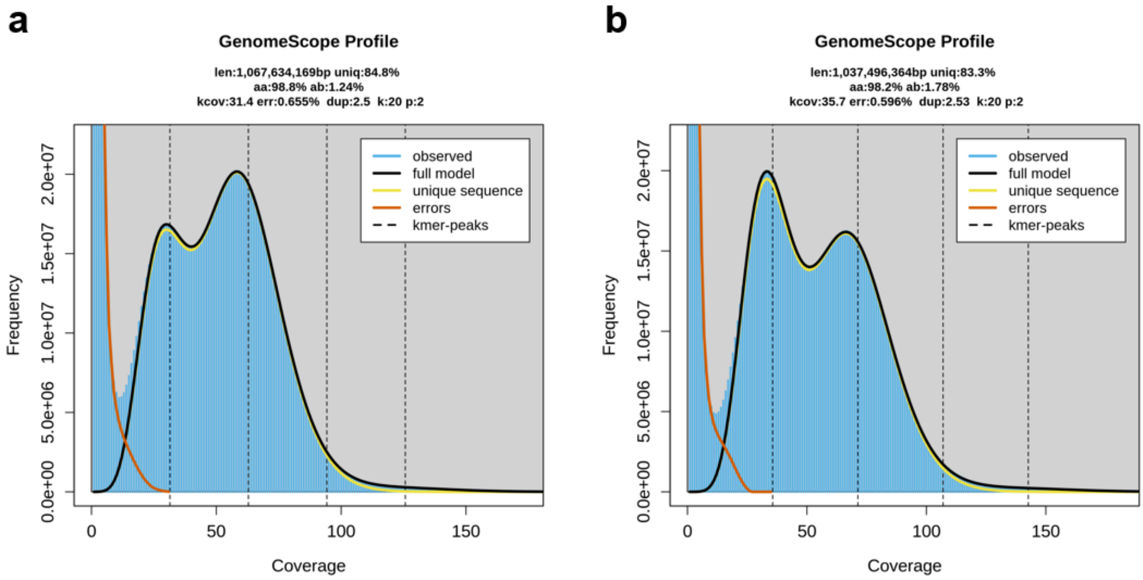


Figure 7. Genomescope profile of zebra finch assemblies calculated from 10X-Linked reads of. A. bTaeGut1.4 assembly and B. bTaeGut2 assembly.

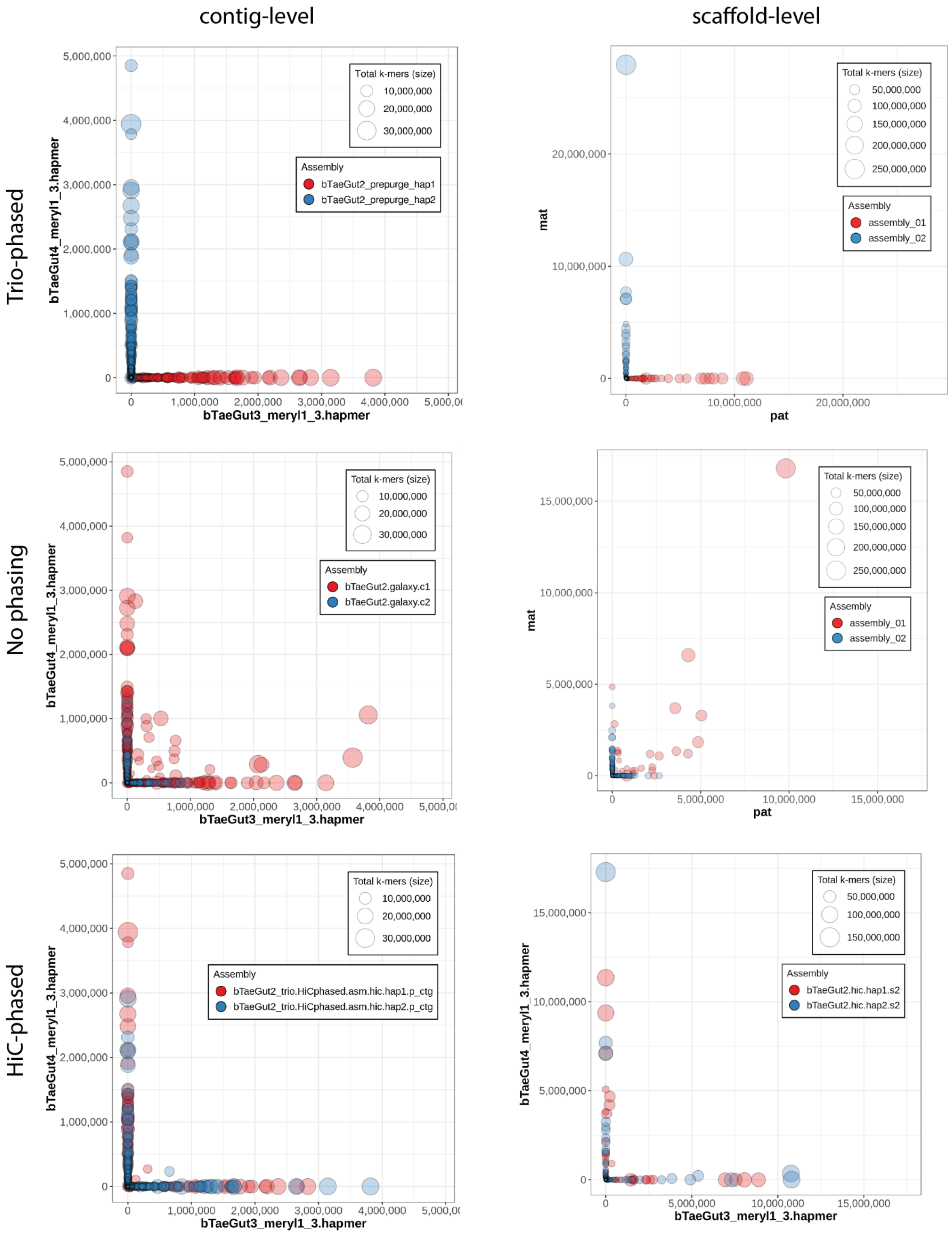


Figure 8. Merqury blob plots for Taeniopygia guttata (bTaeGut2) contigs with various hifiasm modes: top, trio-phased assembly; middle, pseudohaploid ("solo") assembly; and bottom, HiC-phased assembly. Each circle represents a contig, and the size of the circle corresponds to the overall size of the contig, while the color of the circle represents which assembly the contig is in, and the position of the circle along the X- and Y-axes corresponds to parental hapmer content. Contigs that are considered to be properly phased

(e.g., no switch errors) will be flush with either the X- or Y-axis depending on which parental haplotype the contig corresponds to, since this means that the contig contains a number of hapmers from one parent and none from the other parent. This is illustrated in the trio-phased plots (top row), where each contig has been properly phased using parental data. The pseudohaplotype contigs/scaffolds (middle row) shows numerous sequences with a mix of parental hapmer content, as they are not flush to either of the axes. The Hi-C-phased (bottom row) sequences are largely properly phased, with a few contigs off the axes.

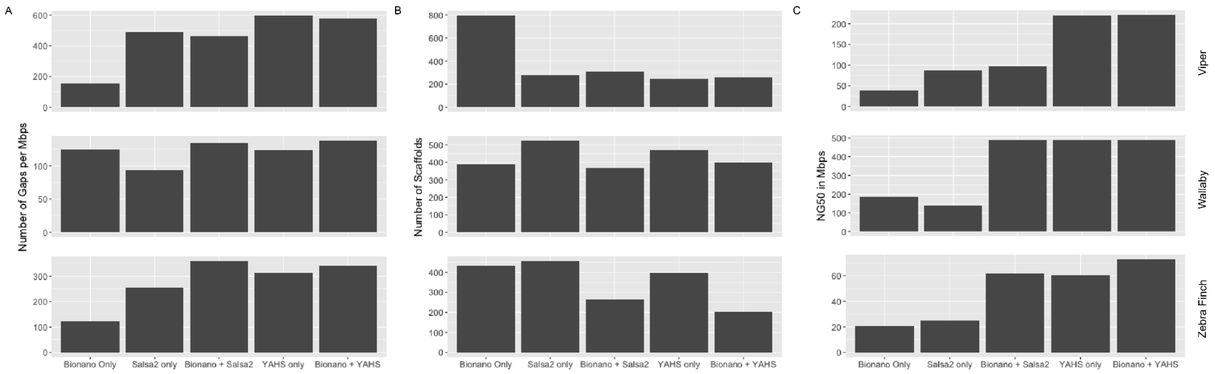
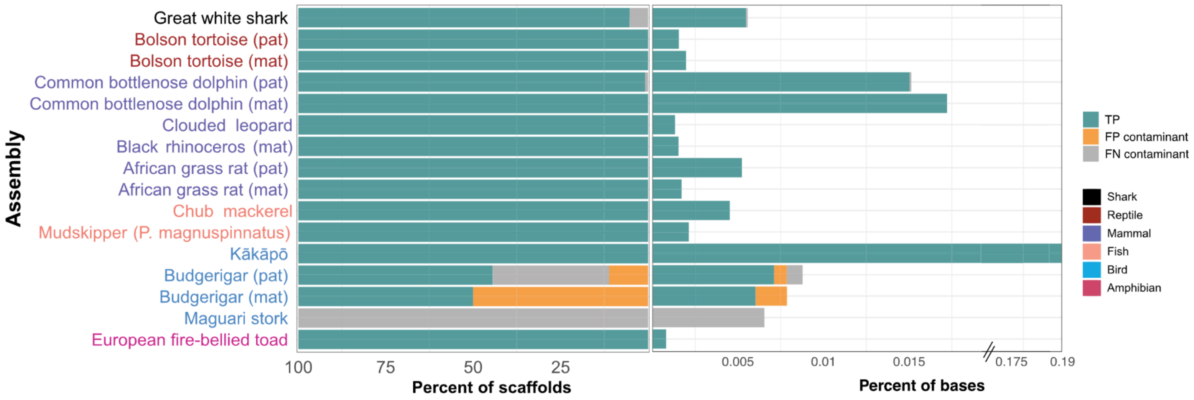


Figure 9. Comparison of scaffolding methods. Comparison of assembly quality between Bionano scaffolding only, Salsa scaffolding only, Bionano and Salsa scaffolding, Yahs scaffolding only, and Bionano and Yahs scaffolding. The comparison has been made for viper, wallaby, and zebra finch assemblies. a. Number of gaps per Mbps; b. Number of scaffolds; c. NG50 in Mbp

A



B

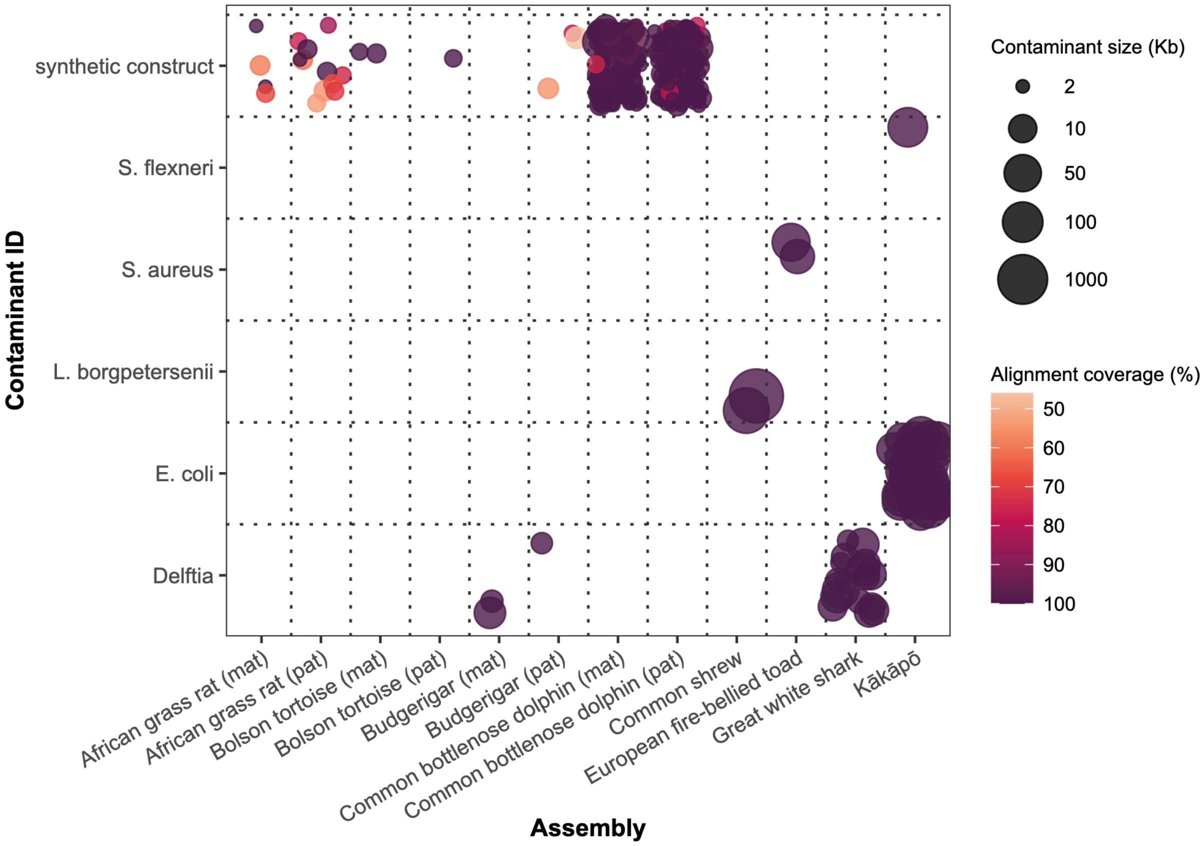


Figure 10. Decontamination pipeline. A. Comparison of foreign contaminants classified by VGP-Galaxy Decontamination Pipeline to the ground truth set. Percent of ground truth contaminant scaffolds that were true positive, false negative and false positive classifications, relative to percent of bases these scaffolds represent from the whole assembly. B. Species of contaminants identified by decontamination. Blast results of the foreign contaminants including coverage from the alignment and the size of the contaminant sequence. 323 of the 325 contaminant sequences identified are represented here; the two excluded contaminants were identified as killer whale and are false classifications. Two hundred sixty five of the contaminant sequences are synthetic constructs. The points are jittered to capture density.

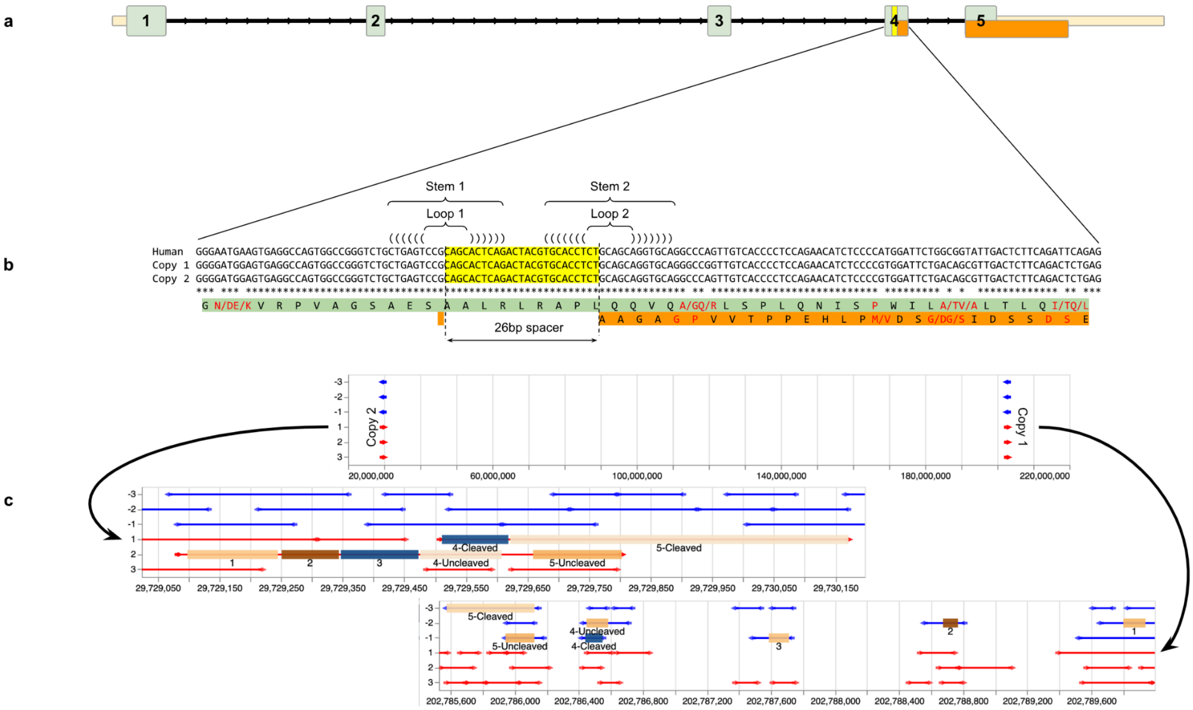

a
1
2
3
4
5
b
Stem 1
Loop 1
Stem 2
Loop 2
Human GGGAATGAAGTGAGGCCAGTGGCCGGGTCTGCTGAGTCCGCAGCACTCAGACTACGTGCACCTCTGCAGCAGGTGCAGGCCCAGTTGTCACCCCTCCAGAACATCTCCCCATGGATTCTGGCGGTATTGACTCTTCAGATTCAGAG
Copy 1 GGGGATGGAGTGAGGCCAGTGGCCGGGTCTGCTGAGTCCGCAGCACTCAGACTACGTGCACCTCTGCAGCAGGTGCAGGGCCGGTTGTCACCCCTCCAGAACATCTCCCCGTGGATTCTGACAGCGTTGACTCTTCAGACTCTGAG
Copy 2 GGGGATGGAGTGAGGCCAGTGGCCGGGTCTGCTGAGTCCGCAGCACTCAGACTACGTGCACCTCTGCAGCAGGTGCAGGCCCAGTTGTCACCCCTCCAGAACATCTCCCCGTGGATTCTGACAGCGTTGACTCTTCAGACTCTGAG
G N/DE/K V R P V A G S A E S A A L R L R A P L Q Q V Q A/GQ/R L S P L Q N I S P W I L A/TV/A L T L Q I/TQ/L
A A G A G P V V T P P E H L P M/V D S G/DG/S I D S S D S E
26bp spacer
c
Copy 2
Copy 1
20,000,000
60,000,000
100,000,000
140,000,000
180,000,000
220,000,000
4-Cleaved
5-Cleaved
4-Uncleaved
5-Uncleaved
29,729,050
29,729,150
29,729,250
29,729,350
29,729,450
29,729,550
29,729,650
29,729,750
29,729,850
29,729,950
29,730,050
29,730,150
202,785,600
202,786,000
202,786,400
202,786,800
202,787,200
202,787,600
202,788,000
202,788,400
202,788,800
202,789,200
202,789,600

Figure 11. Duplication of XBP-1 locus in Cynocephalus volans and comparison to the human locus. a. The structure of the orthologous human XBP1 gene. Exon 4 contains a 26 bp spacer (yellow) that is excised from mature mRNA by endonuclease IRE1α. When the transcript is not cleaved, the green reading frame is translated. When the spacer is removed the reading frame switches to orange downstream of the cleavage site. b. A nucleotide-level representation of exon 4 and corresponding translations for human and two flying lemur copies (Copy 1 and Copy 2). Single letter amino acid identifiers are centered at the second codon position. Red amino acids highlight sites with nucleotide changes. Two amino acids separated by "/" indicate amino acid replacement from one preceding slash to the trailing one: Q/R = change from Q to R. Single red amino acid indicates no change (synonymous substitution). Two stem-loop structures are critical secondary structure elements of the IRE1α cleavage site. c. Structure of two XBP-1 loci in Cynocephalus volans. Top panel shows the relative position of the two copies within scaffold 2. Copy 1 retains exon/intron structure identical to that of the human gene. Copy 2 lacks introns completely. Arrows indicate all possible reading frames (STOP-to-STOP) in the vicinity of each exon. Red = + strand; Blue = – strand.

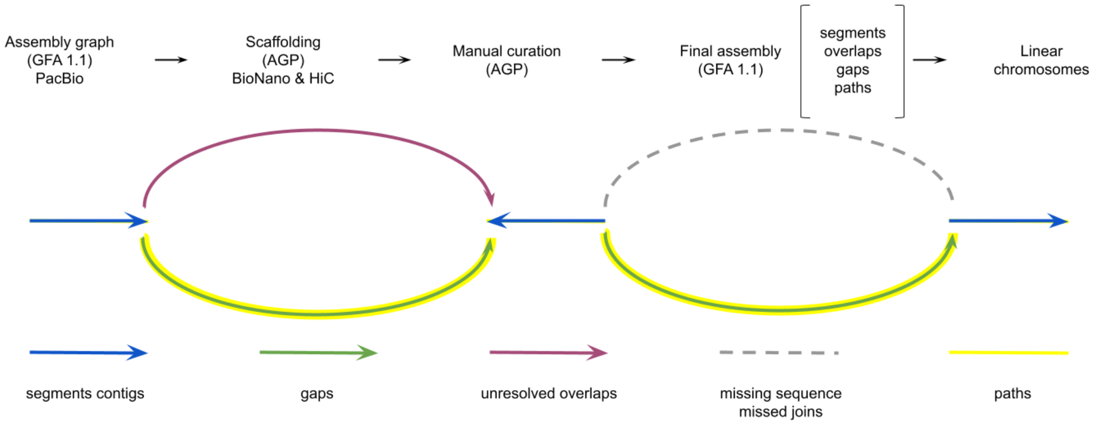
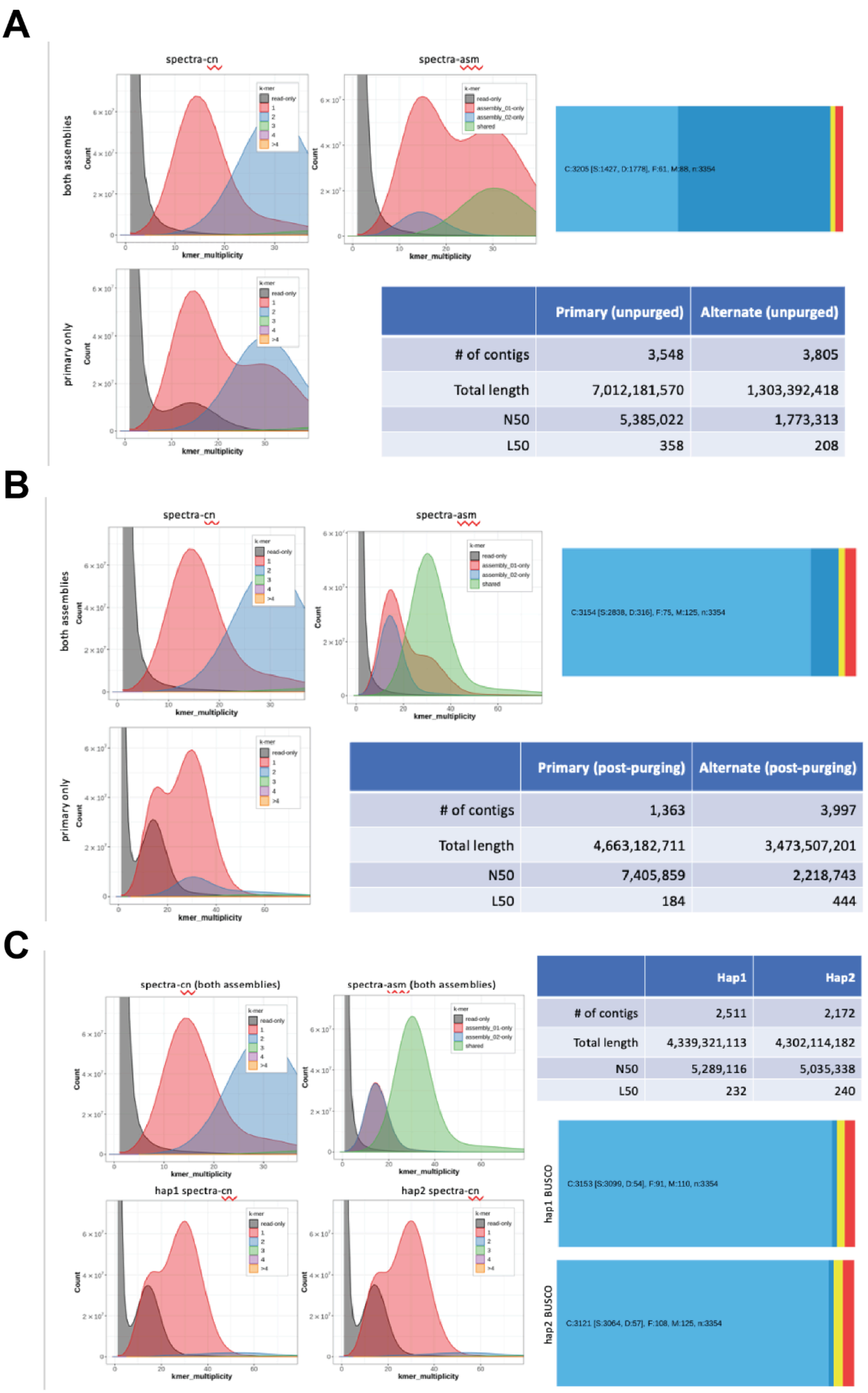


Figure 12. Schematic of assembly graph propagation through the assembly pipeline using the graphical fragment assembly (GFA) format. Initially, PacBio contigs are generated in GFA format. The scaffolding information from Bionano and/or Hi-C is added to the graph through (A Golden Path) AGP intermediates [50]. Manual curation can be integrated in the graph using AGP files. The final assembly is a collection of segments and unresolved overlaps from the original graph, with gaps and paths representing the scaffolding information. Paths can be converted to linear FASTA sequences for downstream analyses. During the scaffolding process gaps (green arrows) are added as jump (J) lines between the segments (blue arrows) to the GFA. This allows the information on unresolved overlaps (purple arrows) to be maintained while missed joins (dashed arrows) inferred from the scaffolding information are added to the graph. The final linear sequences are represented as paths in the graph (yellow highlight).



	Primary (unpurged)	Alternate (unpurged)
# of contigs	3,548	3,805
Total length	7,012,181,570	1,303,392,418
N50	5,385,022	1,773,313
L50	358	208

	Primary (post-purging)	Alternate (post-purging)
# of contigs	1,363	3,997
Total length	4,663,182,711	3,473,507,201
N50	7,405,859	2,218,743
L50	184	444

	Hap1	Hap2
# of contigs	2,511	2,172
Total length	4,339,321,113	4,302,114,182
N50	5,289,116	5,035,338
L50	232	240

Figure 13. Quality control for several Gastrophryne carolinensis (aGasCar1) assemblies: a. pseudohaplotype assembly pre-purging, b. pseudohaplotype post-purging, and c. Hi-C-phased contigs. The quality control metrics shown are merqury k-mer spectra graphs, assembly statistics, and BUSCO genes for Vertebrata. The merqury plots show the partitioning of k-mers from the readset across the two assemblies, which can signal that the

two assemblies are unbalanced such as in S8b, where there are regions with diploid coverage present only in the primary assembly (spectra-asm plot in panel B), and at 2-copy (seen in the primary-only merqury plot in panel B and the BUSCO image for the same panel). Panel C shows the spectra-cn plot for hap1 and hap2 individually, and there are much fewer 2-copy k-mers at diploid coverage.

The pre-purging pseudohaplotype assemblies show that the assemblies are largely unbalanced, with many diploid regions being retained twice in the primary assembly. Purging addresses this problem, but unsatisfactorily, as there are still numerous BUSCO duplicates, as well as unevenness between the two haplotypes. In contrast, the Hi-C-phased contigs show proper haplotype resolution from the start, with minimal BUSCO duplicates and the two haplotypes resolved on a k-mer level from the start.

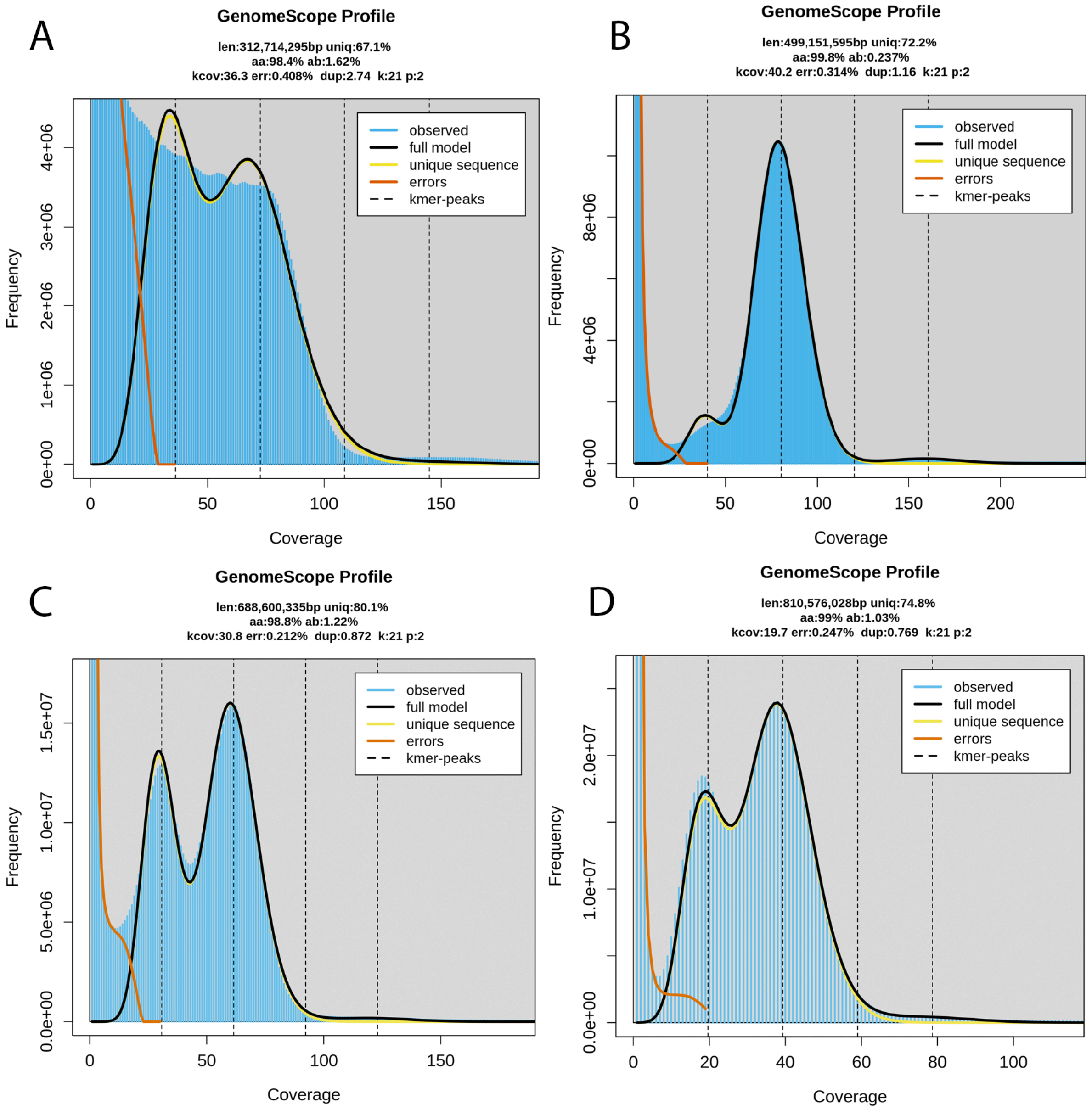


Figure 14. Examples of GenomeScope 2.0 k-mer plots for species that did not perform well with HiFi. a. HiFi k-mer-based plot for *Osmerus mordax*, the rainbow smelt. b. Illumina k-mer-based plot for the same *O. mordax* individual. Due to sequencing bias, the HiFi k-mer distribution is far from a regular Poisson distribution, leading to poor fitting of the GenomeScope 2.0 model, whereas the Illumina-based k-mer spectra has a defined coverage peak. c. HiFi k-mer-based GenomeScope 2.0 plot for *Scomber japonicu*s, the chub mackerel. d. Illumina k-mer-based spectrum for the same *S. japonicu*s individual. The Illumina-based spectrum has a comparatively lower error peak (orange line).

References

1. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).
2. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
3. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
4. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
5. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
6. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
7. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *bioRxiv* 2022.12.01.518724 (2022) doi:10.1101/2022.12.01.518724.
8. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
9. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01261-x.
10. Sharma, P., Masouleh, A. K., Topp, B., Furtado, A. & Henry, R. J. De novo chromosome level assembly of a plant genome from long read sequence data. *Plant J.* (2021) doi:10.1111/tpj.15583.
11. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
12. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

13. Ko, B. J. *et al.* Widespread false gene gains caused by duplication errors in genome assemblies. *bioRxiv* 2021.04.09.438957 (2021) doi:10.1101/2021.04.09.438957.

14. Kim, J. *et al.* False gene and chromosome losses affected by assembly and sequence errors. *bioRxiv* 2021.04.09.438906 (2021) doi:10.1101/2021.04.09.438906.

15. Toh, H. *et al.* A haplotype-resolved genome assembly of the Nile rat facilitates exploration of the genetic basis of diabetes. *BMC Biol.* **20**, 245 (2022).

16. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).

17. Feliciano, P. *et al.* Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genom Med* **4**, 19 (2019).

18. IUCN. Vipera latastei: Jose Antonio Mateo Miras, Marc Cheylan, M. Saïd Nouira, Ulrich Joger, Paulo Sá-Sousa, Valentin Pérez-Mellado, Iñigo Martínez-Solano. *IUCN Red List of Threatened Species* (2008) doi:10.2305/iucn.uk.2009.rlts.t61592a12503848.en.

19. Kasai, F., O'Brien, P. C. M., Pereira, J. C. & Ferguson-Smith, M. A. Marsupial chromosome DNA content and genome size assessed from flow karyotypes: invariable low autosomal GC content. *R Soc Open Sci* **5**, 171539 (2018).

20. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).

21. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, (2023).

22. Li, H. auN: a new metric to measure assembly contiguity. https://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity.

23. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

24. Formenti, G. *et al.* Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* **22**, 120 (2021).

25. Uliano-Silva, M., Nunes, J. G. F., Krasheninnikova, K. & McCarthy, S. A. *marcelauliano/MitoHiFi: mitohifi_v2.0*. (2021). doi:10.5281/zenodo.5205678.

26. Ratnasingham, S. & Hebert, P. D. N. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364 (2007).

27. Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S. K. & Nekrutenko, A. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* **3**, e91 (2007).

28. Soares, M. B. *et al.* RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol. Cell. Biol.* **5**, 2090–2103 (1985).

29. Wentworth, B. M., Schaefer, I. M., Villa-Komaroff, L. & Chirgwin, J. M. Characterization of the two nonallelic genes encoding mouse preproinsulin. *J. Mol. Evol.* **23**, 305–312 (1986).

30. Calfon, M., Zeng, H., Urano, F. & Till, J. H. IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. … (2002).

31. Donath, A. *et al.* Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Res.* **47**, 10543–10552 (2019).

32. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).

33. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

34. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).

35. Formenti, G. *et al.* Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *bioRxiv* 2022.03.24.485682 (2022) doi:10.1101/2022.03.24.485682.

36. The GFA Format Specification Working Group. Graphical Fragment Assembly (GFA) Format Specification. *GFA-spec* http://gfa-spec.github.io/GFA-spec/GFA1.html (2022).

37. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and

Annotation Completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

38. Bocklandt, S., Hastie, A. & Cao, H. Bionano Genome Mapping: High-Throughput, Ultra-Long Molecule Genome Analysis System for Precision Genome Assembly and Haploid-Resolved Structural Variation Discovery. *Adv. Exp. Med. Biol.* **1129**, 97–118 (2019).
39. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. Preprint at (2015).
40. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
41. Shajii, A., Numanagić, I. & Berger, B. Latent Variable Model for Aligning Barcoded Short-Reads Improves Downstream Analyses. *Res. Comput. Mol. Biol.* **10812**, 280–282 (2018).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
46. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
47. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
48. Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
49. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using

DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

50. AGP Specification v2.1. https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/.