농 학 박 사 학 위 논 문

Bioinformatic approach for identifying and correcting artifacts in diploid genome assemblies

이배체 유전체 조립 과정에서의 인위적 오류 식별과 교정을 위한 생물정보학적 접근

2023 년 8 월

서울대학교 대학원
농생명공학부 바이오모듈레이션 전공
고 병 준

Bioinformatic approach for identifying and correcting artifacts in diploid genome assemblies

By
Byung June Ko

Supervisor: Professor Heebal Kim

Aug, 2023

Biomodulation Major
Department of Agricultural Biotechnology
Seoul National University

이배체 유전체 조립 과정에서의 인위적 오류 식별과 교정을 위한 생물정보학적 접근

지도 교수 김 희 발

이 논문을 농학박사 학위논문으로 제출함
2023 년 7 월

서울대학교 대학원
농생명공학부 바이오모듈레이션 전공
고 병 준

고병준의 농학박사 학위논문을 인준함
2023 년 7 월

위 원 장 유 경 록 (인)

부위원장 김 희 발 (인)

위 원 정 충 원 (인)

위 원 조 서 애 (인)

위 원 이 원 석 (인)

Abstract

Bioinformatic approach for identifying and correcting artifacts in diploid genome assemblies

Byung June Ko

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Errors in genome assembly present in reference genomes can lead to errors in biological interpretation. With recent advancements in DNA sequencing technologies, large-scale genome projects are underway. The Vertebrate Genome Project (VGP), for example, aims to decode the genomes of over 66,000 vertebrate species. This project strives for high-quality reference genome construction by minimizing errors in both base and structure level in the genome assemblies. Other recent international genome projects such as the

Telomere to Telomere (T2T) Consortium and the Earth Biogenome Project (EBP) also emphasize the importance of high-quality reference genome construction, highlighting the ongoing efforts among researchers to improve genome quality.

In Chapter 2, through collaboration with VGP, false duplications resulting from assembly errors were identified in the reference genome, which was previously based on short read sequencing data, as well as in the more recent long-read and combination sequencing technologies. Hundreds to thousands of falsely duplicated genes were detected with 4 to 16% of false duplications in the reference genomes made by short read sequencing, but ~2% of false duplications were detected in long read-based reference genome assemblies. Heterozygosity and sequencing error were identified as significant factors contributing to false duplication. The result also showed that several downstream analyses can be significantly disturbed by false duplication. The findings emphasize the importance of developing more advanced assembly methods that effectively separates haplotypes and removes sequencing errors, as well as the need for careful analysis of gene gains.

In Chapter 3, a collaboration with VGP and Galaxy Project allowed for a comparison between the increasingly recognized PacBio

High-Fidelity (HiFi) sequencing method and the PacBio Continuous Long Read (CLR) method in terms of false duplications and losses in same individual of zebra finch. *K-mer* based false duplication, expansion and collapse results indicated that the CLR based assembly exhibited a higher susceptibility to both false duplication and loss. Another approach by genome-wide alignment with read coverage analysis showed that CLR based assembly had more false duplication and loss errors (1.3 and 4%, respectively) than HiFi based assemblies (~0.6 and <1%, respectively).

Chapter 4 introduces a newly developed false duplication correction software, Purge mers, which was compared to existing programs through the generation of virtual genome assemblies. The purge mers, utilizes both read depth coverage and $K*$ to detect false duplications at base-pair level. The performance of purge mers was found to be superior to existing programs when using short read or long read in some cases.

In Chapter 5, a methodology for correcting the bias caused by high GC content in the genome, resulting in underrepresentation of *k-mer* multiplicities in the read data, was proposed. Uncorrected *k-mer* measurements revealed the highest frequency of $K*$ at -1 in genomic regions with GC content over 80%. On the other hand, the

bias-corrected *k-mer* measurements presented in this study showed the highest frequency of K^* at 0 in genomic regions with GC content over 80%. These results provide confirmation that high GC content inhibits sequencing, and the underestimation of *k-mer* multiplicities can be recovered by the method suggested in this study.

In summary, the studies emphasize the importance of false duplication error correction. It proposes optimized DNA sequencing techniques, genome assembly methods to mitigate false duplication. Also, I developed a novel program to correct false duplication, and a methodology to recover *k-mer* multiplicities from GC bias.

Keywords: False duplication, phasing error, *k-mer*, assembly error, assembly curation, VGP, vertebrate Genome Project

Student number: 2018-34934

Contents

List of Tables

List of Figures

Chapter 1. General introduction

1.1 Advancing error-free genome assembly

Genome assemblies play a crucial role in understanding the species, but errors in the assembly process can lead to biological misinterpretations (Cheung et al., 2003; Kelley and Salzberg, 2010; Ko et al., 2022; Korlach et al., 2017). Recently, error-free genome assembly has become a critical task in large-scale genomics projects. One groundbreaking research publication by the Vertebrate Genome Project (VGP) (Rhie et al., 2021) has aimed to assemble the genomes of numerous vertebrate species (~66,000). This project has set out to benchmark various sequencing platforms and assembly algorithms in order to eliminate structural assembly errors, and it has gained significant recognition as a standard for diploid genome assembly. In addition, Earth Biogenome Project has made an effort to find suitable strategies for genome assembly of ~1.8 million known eukaryotic species with high quality at scale (Lewin et al., 2022). The recent assembly of the human Telomere-to-Telomere (T2T) genome represents another monumental achievement in the pursuit of generating error-free and complete sequences of a species (Nurk et al., 2022). The consensus among researchers is that producing high-quality genomes is of utmost importance; however, there is currently no universally optimized method applicable to all species even in

vertebrates. Consequently, specific methods and pipelines are currently being discussed and developed.

1.2 Structural error made by assembly artifacts

Due to technology and cost constraints, the genome should be sequenced in fragmented form, consisting of hundreds or thousands of base pairs. These assembled genome sequences frequently contain structural errors, such as redundantly duplicated nucleotide sequences from allelic divergence and sequencing errors, called as false duplication (Kelley and Salzberg, 2010; Ko et al., 2022; Rhie et al., 2021). The false duplication can lead to significant misinterpretation in genomic comparisons, particularly in the context of gene duplications and expansion (Cheung et al., 2003; Ko et al., 2022; Korlach et al., 2017). Moreover, errors in genome assembly can significantly impact various downstream analyses, including evolutionary biology, such as comparative genomics, phylogenetics, population genomics, structural variation, and copy number variation studies. For instance, false duplications can result in the removal of one-to-one orthologs between species, the formation of artificial chimeric genes, and partial gene losses within the genic region (Ko et al., 2022). In the case of a reference genome containing false

duplications, the accuracy of structural variation and copy number variation analyses can also be compromised due to the presence of redundantly inserted sequences. Previous assemblies based on short reads have revealed these types of errors, leading to practical instances of misinterpretations in published studies (Ko et al., 2022; Rhie et al., 2021).

1.3 Challenges of false duplication

False duplications commonly occur in highly heterozygous genomes, and traditional approaches to address this issue involve generating individuals with high inbreeding, which is often impractical (Koren et al., 2018; Rhie et al., 2021). To overcome these challenges, several post-processing software tools have been developed, which identify false duplications based on sequence similarity and coverage profiles of the nucleotide sequences. Purge haplotigs (Roach et al., 2018) is one of the tool collapse false duplication in contig level. The authors used the read depth coverage as an information to find false duplication because the region under false duplication should exhibit haploid-level depth coverage. Another tool Purge_dups (Guan et al., 2020) is also working for identifying false duplication based on depth coverage. The tool expands the unit of analysis for detecting false

duplications from within the contig to include the edges of the contig. This extension allows the tool to identify false duplications that overlap with the contig edges.

In addition to depth coverage, *k−mers* can also be utilized to identify structural assembly errors. *K−mer* multiplicity represents the number of identical K−length subsequences in assemblies or sequencing reads. *K−mers* have been widely used to assess genome characteristics and evaluate the quality of genome assembly (Formenti et al., 2022; Phillippy et al., 2008; Rhie et al., 2020). Based on the *k−mer* multiplicities of both read and assembly, *K** has been introduced as a metric to assess whole or part of genome regions (Formenti et al., 2022; Phillippy et al., 2008). This metric has an advantage over read depth coverage in terms of not being affected by the read mapping algorithm for regional genome evaluation. Although *K** is known to be capable of quantifying false duplication, there has been no systematic effort to utilize this metric for identifying false duplication. But the *K** is not a cure−all for identifying structural errors. It assumes consistent sequencing coverage across the genome, but it is known that GC rich regions inhibit short read sequencing (Benjamini and Speed, 2012), and there is GA drop−out in PacBio long read sequencing technology (Formenti

et al., 2022). These sequencing biases may lead to false positives in false duplication identification using $K*$.

In this study, I discovered that numerous false gene gains occurred due to false duplication in various vertebrate assemblies made by both short read and long read sequencing technologies (Chapter 2). Furthermore, I conducted a comparative analysis of the extent of structural assembly errors resulting from false duplications and losses in PacBio CLR and HiFi-based assemblies of a zebra finch, in collaboration with Galaxy and VGP (Chapter 3). Based on these findings, I proposed future directions to mitigate false duplication in diploid genome assembly. Additionally, I developed a novel tool for curating false duplication by employing both read depth coverage and $K*$ through simulation (Chapter 4). Lastly, I suggested a method for mitigating GC-bias in *k-mer* counting for short read sequencing data (Chapter 5).

This chapter was published in *Genome Biology*

as a partial fulfillment of Byung June Ko's Ph.D program.

Chapter 2. Widespread false gene gains caused by duplication errors in genome assemblies

2.1 Abstract

False duplications in genome assemblies lead to false biological conclusions. We quantified false duplications in popularly used previous genome assemblies and their new counterparts of the same species (platypus, zebra finch, Anna' s hummingbird) generated by the Vertebrate Genomes Project (VGP), of which the VGP pipeline attempted to eliminate false duplications through haplotype phasing and purging. These assemblies are among the first generated by the VGP where there was a prior chromosomal level reference assembly to compare with. Whole genome alignments revealed that 4 to 16% of the sequences were falsely duplicated in the previous assemblies, impacting hundreds to thousands of genes. These led to overestimated gene family expansions. The main source of the false duplications was heterotype duplications, where the haplotype sequences were relatively more divergent than other parts of the genome leading the assembly algorithms to classify them as separate genes or genomic regions. A minor source was sequencing errors. Ancient ATP nucleotide binding gene families had a higher prevalence of false duplications compared to other gene families. Although present in a smaller proportion, we observed false duplications remaining in the VGP assemblies that can be identified

and purged. This study highlights the need for more advanced assembly methods that better separates haplotypes and sequence errors, and the need for cautious analyses on gene gains.

2.2 Introduction

Biological misinterpretations can occur when genomic regions unknowingly have errors. But it is unclear as to the magnitude of mis-assembly errors in existing genome assemblies, generated in the transition from the fragmented DNA sequences to the assembled blueprint of a species (Cheung et al., 2003; Ekblom and Wolf, 2014; Jones et al., 2004; Kelley and Salzberg, 2010; Korlach et al., 2017; Phillippy et al., 2008; Rhie et al., 2021; Salzberg and Yorke, 2005). Followed by the first assembly of fruit fly in 2000 (Adams et al., 2000) and a human reference genome in 2003 (Venter et al., 2001), ~100 reference genomes of vertebrates were deposited in public databases by 2010 using mostly intermediate read length (~700 bp) Sanger reads. The number of genomes gradually increased to ~700 by 2018, mostly using short read-based (~35-250 bp) next generation sequencing (NGS) (Rice and Green, 2019). These genomes helped bring about discoveries in a variety of fields, including evolution, ecology, agriculture, and medicine (Church et al.,

2011; Ellegren, 2014; Huang and Han, 2014; Jarvis et al., 2014; Nakagawa and Fujita, 2018; Seehausen et al., 2014). However, with short read-based assemblies, it was difficult to resolve repeat regions longer than the read lengths (Bresler et al., 2013; Korlach et al., 2017; Luo et al., 2012; Simpson and Pop, 2015).

Preliminary studies have indicated that the longer the sequence read length, the less likely an assembly structural error (Korlach et al., 2017), which has been quantitatively validated in our companion Vertebrate Genomes Project (VGP) flagship study in 2021 (Rhie et al., 2021). An underappreciated source of mis-assembly was heterozygosity (Pryszcz and Gabaldón, 2016; Rhie et al., 2021). Mis-assignment of heterozygous genomic regions led to both copies of the partnering alleles being assembled as paralogs in the same haploid assembly (Cheung et al., 2003; Kelley and Salzberg, 2010; Rhie et al., 2021), which are called false heterotype duplications by the VGP (Rhie et al., 2021). Likewise, accumulated sequence errors in reads, particularly long reads, led to under-collapsed sequences, which were called homotype false duplications (Rhie et al., 2021). Both heterotype and homotype false duplications in genic regions can be misinterpreted as gene gains (Korlach et al., 2017; Pryszcz and Gabaldón, 2016; Schneider et al., 2017). The VGP

proposed that these false gains happen in more highly divergent regions of the genome, where assembly algorithms have difficulty distinguishing haplotype homologs from haplotype paralogs (Rhie et al., 2021), but this was not quantitatively tested in regards to the type of duplication.

Although long-read sequencing is better at resolving repetitive regions (Ameur et al., 2019; Korlach et al., 2017; Rice and Green, 2019), they alone are unable to fully resolve false duplications (Koren et al., 2018; Korlach et al., 2017; Rhie et al., 2020). One way to prevent false duplications is to make homozygous lineages through inbreeding. But this can be either impossible or very difficult under most circumstances (Koren et al., 2018; Vinson et al., 2005), especially if one were to sequence all species of a lineage, such as the goal of the VGP that aims to produce complete and error-free reference genomes for all ~70,000 vertebrate species ("A reference standard for genome biology," 2018; Genome 10K Community of Scientists, 2009; Koepfli et al., 2015). Another way to solve false duplications is to use assembly strategies for efficient haplotype phasing, some developed and applied in the VGP (Chin et al., 2016; Guan et al., 2020; Koren et al., 2018; Rhie et al., 2021). But most of the non-VGP vertebrate genomes in the public databases

as of to date have been reconstructed without haplotype phasing. A full quantitative and qualitative assessment has not been conducted on the prior versus VGP genomes to determine the extent and types of false duplications, and improvements in the VGP assemblies.

Here we performed a detailed analysis to measure the presence, magnitude and cause for false duplications in previous common reference assemblies and their VGP counterparts. We focused on three species, the platypus and zebra finch that were originally assembled using Sanger reads published in 2008 (Warren et al., 2008) and 2010 (Warren et al., 2010), respectively, and the Anna' s hummingbird that used short Illumina reads published in 2014 (Jarvis et al., 2014; Zhang et al., 2014). These are popular references, with the associated studies collectively cited over 3,600 times as of April 2021 (Google Scholar). The VGP version of the assemblies were long-read based, and used algorithms to phase haplotypes and purge false duplications at multiple steps in the assembly pipeline. We found widespread false duplications in previous assemblies that were corrected in the VGP assemblies, and also identified areas for improvement in current and future assemblies.

2.3 Materials and Methods

2.3.1 Assemblies and read data

The primary assembly of the previous and VGP version of the male zebra finch, female Anna's hummingbird, and female and male platypus were downloaded from NCBI by ftp along with their assembly statistics, gaps, repeats and annotation data (Table 2. 1). For the VGP assemblies, we included both the primary and alternate pseudo-haplotype sequences. The raw reads used for the previous assemblies of the zebra finch and platypus generated by Sanger sequencing were not available to download from the Sequencing Read Archive (SRA) on NCBI. However, the raw Sanger reads of the previous version of the platypus assembly was in the Trace Archive in NCBI (https://ftp.ncbi.nlm.nih.gov/pub/TraceDB/ornithorhynchus_anatinus/). We downloaded all '.anc' and '.fasta' files, and extracted the reads that were submitted by 'WUGSC' for platypus the assembly. These Sanger reads from the older assembly and the 10X linked reads of the new assembly were used to quantify whether the duplications were due to individual differences between previous and VGP assemblies or real false duplications. The PacBio CLR and 10X

raw reads used to generate the VGP assemblies were downloaded from the VGP Genome Ark (https://vgp.github.io/genomeark/).

Table 2. 1 Statistics of previous and VGP assemblies. Contig NG50 and Scaffold NG50 for each assembly were calculated using a source code in the VGP repository.

No.	Species	Assembly type	Date	Sequencing Technology	Assembler	Coverage	NCBI Accession	Total Length (bp)	# Sca.	# Chr.	# Gaps	Contig NG50	Scaffold NG50	Annotation	Reference
1	zebra finch	previous primary assembly	2013	Sanger + BAC cloning	PCAP	5.5x	GCF_000151805.1[a]	1,232,118,738	37,421	35	87,710	47,913	72,861,351	103	Warren et al. (2010)
2	zebra finch	VGP primary assembly	2019	PacBio RSII; 10X Genomics linked reads; Bionano Genomics DLS; Arima	VGP assembly standard pipeline[c]	88.2x	GCF_003957565.1[a]	1,058,012,133	135	33	312	11,998,827	70,430,603	104	Rhie et al. (2021)
3	zebra finch	VGP alternate pseudohaplotype	2018	PacBio RSII; 10X Genomics linked reads	VGP assembly standard pipeline[c]	88.2x	GCA_003957525.1[a]	965,644,423	5,336	0	45	2,280,982	2,280,982	-	Rhie et al. (2021)
4	Anna's hummingbird	previous primary assembly	2014	Illumina HiSeq	SOAPde novo	110.0x	GCF_000699085.1[b]	1,105,676,412	54,736	0	70,084	26,950	4,286,189	100	Zhang et al. (2014)
5	Anna's hummingbird	VGP primary assembly	2019	PacBio RSII; 10X Genomics linked reads; Bionano Genomics DLS; Arima	VGP assembly standard pipeline[c]	54.0x	GCF_003957555.1[b]	1,059,687,259	159	33	429	13,410,196	74,081,004	101	Rhie et al. (2021)
6	Anna's hummingbird	VGP alternate pseudohaplotype	2018	PacBio RSII; 10X Genomics linked reads	VGP assembly standard pipeline[c]	54.0x	GCA_003957575.1[b]	952,083,371	3,803	0	6	1,237,155	1,237,155	-	Rhie et al. (2021)
7	platypus	previous primary assembly	2011	Sanger + BAC cloning	PCAP	6.0x	GCF_000002275.2	1,995,607,322	958,970	19	243,835	11,375	1,564,930	103	Warren et al. (2008)
8	platypus	VGP primary assembly	2019	PacBio RSII; 10X Genomics linked reads; Bionano Genomics DLS; Dovetail	VGP assembly standard pipeline[c]	58.8x	GCF_004115215.1[c]	1,858,552,590	305	31	522	15,022,425	83,338,043	104	Rhie et al. (2021)
9	platypus	VGP alternate pseudohaplotype	2019	PacBio RSII	VGP assembly standard pipeline[c]	58.8x	GCA_004115175.1[c]	1,575,984,168	5,850	0	28	713,443	713,443	-	Rhie et al. (2021)

a These assemblies were made from same biosample (SAMN02981239)

b These assemblies were made from same biosample (SAMN02265252)

c Rhie et al. (2021)

2.3.2 Identifying false duplications

2.3.2.1 Candidate duplications from sequence similarity

We identified false duplication candidates by sequence similarity in whole genome alignments between the previous and VGP assemblies and self-alignment of an assembly to itself. We used Cactus (Armstrong et al., 2019; Paten et al., 2011) to generate whole genome alignment across assemblies with the default options and HAL (Hickey et al., 2013) to transform the Cactus results into a readable multiple alignment format with '--maxBlockLen 1,000,000 --noAncestors --refGenome' (VGP assembly as reference) parameters. One to many homologs between two assemblies of the same species were then considered as potential false duplication candidates. Since the Cactus alignment contained very short sequences (<20 bp) in alignment blocks, we filtered out blocks shorter than 20 bp or query sequence coverage of less than 80% to avoid spurious alignments. Self-alignment was performed with Minimap2 (Li, 2018) with the '-xasm5 -DP' option on for assembly alignment mode, after segmenting contigs by 'N' -base gaps. Purge_dups was then used to find false duplications (Guan et al., 2020) with '-2' option following the guideline of purge_dups

(https://github.com/dfguan/purge_dups). We used a purge_dups version that we asked the developers to modify ('add_loc' branch in github of purge_dups; https://github.com/dfguan/purge_dups/tree/add_loc) to output the pair-wise homologous loci for each false duplication found.

2.3.2.2 Filtering true duplications

False duplication candidates were distinguished from true haplotype specific duplications using 10X linked read alignments; it was difficult to map PacBio CLR reads to the previous assemblies, as the length of the majority of the contigs of the prior assemblies (e.g. 1~3 kbp) were shorter than the PacBio read lengths of the VGP assemblies (e.g. ~10-17 kbp). The paired-end reads from the linked reads were aligned with EMA v0.6.2 (Shajii et al., 2018) using the barcodes default option, and BWA v0.7.17 (Li and Durbin, 2009) without the barcodes with parameters '-p -M -R "@RG\tID:rg1\tSM:sample1"' options following guide line of EMA. Duplicate reads were marked by Sambamba v0.7.1. Coverage distribution across the entire assembly was extracted using samtools (Li et al., 2009). False duplication candidates from purge_dups self-alignments were further processed using the remainder of the purge_dups pipeline, which included generating coverage

distributions.

Candidates from the Cactus alignments were similarly filtered using the same read depth threshold as in purge_dups. Any duplications with lower than half the diploid read depth of coverage were further considered. We then applied two additional criteria: 1) presence of a scaffolded gap or read depth-gap between a duplicated pair; and 2) discordant read pair alignments. A depth-gap is defined as a region with no read alignments between duplicated pairs, which occurs from incorrect gap-filling or incorporation of reads with sequencing errors during assembly (**Figure 2. 1**) . A discordant read pair was defined when the insert size between the pairs is unexpectedly large (>550 bp; mean insert size of 10X read in this study) or mapped to another scaffold as in Kelley and Salzberg (Kelley and Salzberg, 2010). We required both presence of discordant reads and concordant reads to align, where one end from a discordant read pair and concordant read pair aligns to the identical flanking region (~550 bp) of a duplication, while the other end aligns to each of the homologous duplications.

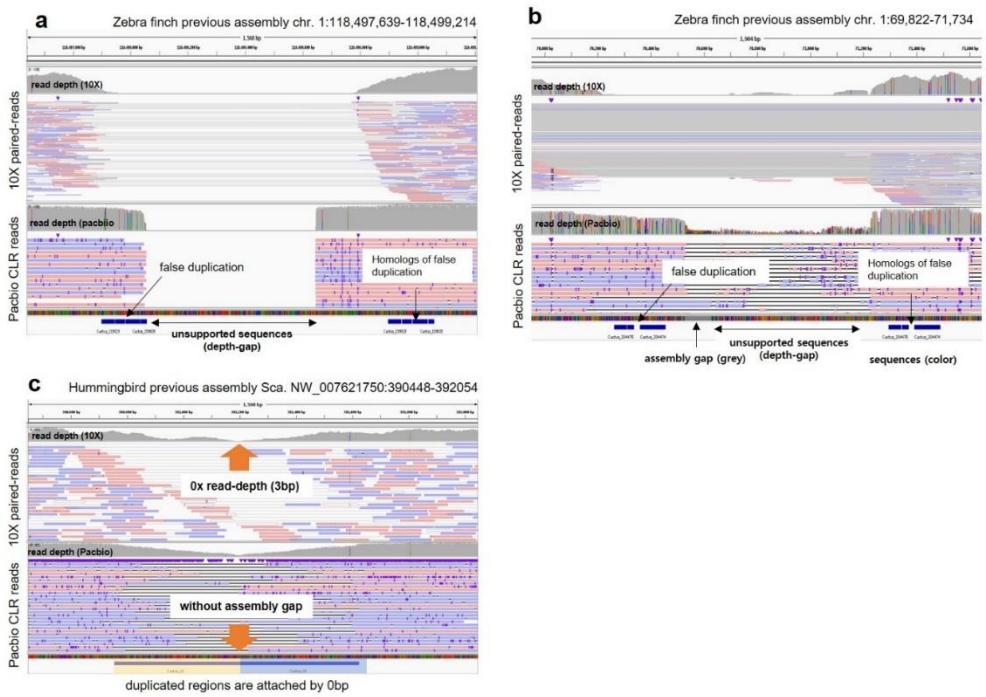

Figure 2. 1 Unsupported sequences with or without assembly gaps.

10X linked reads are shown as paired read alignments, and the PacBio CLR read alignments below them, along with the depth coverage of the respective read data. a, Unsupported sequence with a depth−gap but no assembly gap, between a false duplication. b, Unsupported sequence observed with an assembly gap, between a false duplication. c, Unsupported sequences with 0 bp between a false duplication. Unsupported sequences in the assembly were identified with 10X linked reads with no depth of coverage.

2.3.2.3 Classifying heterotype and homotype duplications

The filtered false duplications were further classified based on *k-mer* analysis (considering genome size, k = 20 for all three species). We extracted 20-mers from the assemblies and 10X linked reads using Meryl (Rhie et al., 2020) and performed Merqury (Rhie et al., 2020) analysis to obtain the *k-mer* spectrums, using the non-trio mode for pseudo-haplotype assembly. Using the *k-mer* spectrum, we defined erroneous *k-mers* as those found in the assembly with read multiplicity lower than 6x, 3x, and 18x in the previous assemblies of zebra finch, hummingbird, and the platypus, and 3x, 3x and 10x for the VGP assemblies, respectively. These are low-multiplicity *k-mers* in the *k-mer* spectrum, made by sequencing errors (Koren et al., 2018; Rhie et al., 2020). Likewise, any non-erroneous *k-mer* found once in the assembly was defined as a single-copy *k-mer*. We classified false duplications as heterotype when both of the duplicated pairs had single-copy *k-mers* with average read depth higher than sequencing error read depths, which was 5x, 8x and 22x for the previous assemblies and 2x, 2x, 9x for the VGP assemblies of the zebra finch, hummingbird, and the platypus, respectively (same principle with erroneous *k-mers* identification); otherwise as homotype duplication, which had

no single-copy *k-mer* found on either side of the duplication or one duplication of the pair had read depth below heterotype duplication levels.

2.3.3 Evaluating false duplications

PacBio CLR reads were mapped to both the previous and VGP assemblies using Minimap2 (Li, 2018) with the preset '-ax map-pb'. Sanger reads of platypus were also mapped to the previous assembly using Minimap2 with '-ax map-pb' and used for further read coverage evaluation. Since the coverage distribution of Sanger reads showed a unimodal distribution at 1x coverage, we defined the threshold of haploid-level coverage for Sanger reads as the mean depth-coverage of the total assembly*0.75. The mapped reads on each assembly were visualized with IGV (Thorvaldsdóttir et al., 2013). Duplications found in the VGP assemblies were aligned to their counterpart assembly and visualized with D-Genies (Cabanettes and Klopp, 2018). The location of false duplications in the VGP assemblies was visualized by karyoploteR (Gel and Serra, 2017).

The heterozygosity of assemblies, including of each corrected

FD and correctly assembled region, were calculated as the number of variants divided by the length of the region, with 1,000 bootstrapping replicates to generate a distribution for a Student's *t*-test between those regions. To calculate heterozygosity in the region of the introduced FD in the VGP assemblies, we masked false duplications as 'N's, then the variant was estimated from newly mapped 10X linked reads onto the masked assembly, followed by the same bootstrapping and statistical approach as used above. Samtools and bcftools were used for variant calling with the multiallelic model. We filtered-out variants with biased alleles, i.e. we only considered the locus if the proportion of major and minor alleles were in >25% and <75%. The sequence error rate of each duplicated and correct region was calculated by dividing the number of erroneous *k-mers* by the total number of *k-mers* found. The distributions of sequencing error rate for duplicated and correct regions were also generated by 1,000 bootstrapping replicates, and a Student's *t*-test was performed on those distributions.

2.3.4 Identification of false gene gain annotation errors

We calculated the number of protein coding genes affected by false duplications, defined as regions with duplicated sequences that

overlapped with the CDS regions of an assembly. The Refseq annotation of NCBI was used and only the longest CDS of all isoforms generated from each gene was used. The genes influenced by false duplications were classified into three types: 1) false gene gain (FGG) in which a gene was falsely duplicated almost entirely or partially over 50% of the CDS length; 2) false exon gain (FEG) of one or more exons within the same gene; and 3) false chimeric gain (FCG) in which duplicated exons from one gene were inserted into another gene. FGG, FEG, and FCG were included only when at least one coding exon of a gene completely overlapped the false duplication. To visualize the example cases of mis-annotation, GSDS 2.0 (Hu et al., 2015) was used. Intergenic regions were defined as the remaining regions excluding CDS and intron.

To search for possible false duplications in non-coding repetitive elements, we counted the number of LTRs, SINEs, and LINEs affected by false duplications using NCBI repeat information generated by repeatMasker (Tarailo-Graovac and Chen, 2009). Then, the relative proportion of false duplication on each genomic partition was calculated by the difference between observed and expected proportion. The observed is the proportion of each genomic partition containing false duplications (Σ feature length overlapped with false

duplication / total false duplication length) of each assembly. The expected is the normal proportion of each genome partition (Σ feature length / total assembly length). The differences between observed and expected genomic partitions were tested by one-way analysis of variance (ANOVA).

In the platypus, we also searched false gene gains of the V1R family in the same manner as above. We checked for 267 V1R genes for potential false gene gains in the previous assembly of the platypus, which included "ORNANAV1R" in the gene symbol.

2.3.5 False duplication correction using the VGP pipeline v1.7

We reassembled the zebra finch assembly using a variation of the VGP v1.0-1.6 pipelines, which we called the VGP v1.7 pipeline. Aside from software updates, the two main differences with respect to the VGP v1.0 pipeline (Rhie et al., 2021) are: 1) purge_haplotigs was replaced by purge_dups, for more effective purging of false haplotype and homotype duplications; 2) purging was done after contiging, as opposed to after scaffolding and polishing; and 3) during the final Arrow polishing step, variant calls were filtered with Merfin (https://github.com/arangrhie/merfin), to avoid introducing

erroneous *k*−mers in the assembly. This resulted in the following assembly steps: 1) FALCON−Unzip contig assembly; 2) purge_dups to purge false duplications in the primary assembly, and place them in the alternate assembly; 3) scaffolding the primary assembly with 10X linked reads and scaff10X software; 4) scaffolding with Bionano optical maps and Bionano solve software; 5) scaffolding with Arima Genomics Hi−C and Salsa v2.2 software; 6) polishing with long reads using Arrow and filtering the variant calls with Merfin; and 7) a final polishing with longranger aligner and freebayes. We added the assembled mitochondrial genome prior to the polishing steps to prevent overpolishing of NUMTS in the nuclear genome. We compared this VGP 1.7 assembly (bTaeGut1.4) with the zebra finch VGP v1.0 pipeline (bTaeGut1.0; GCF_003957565.1) by alignment using Cactus (Paten et al., 2011). Based on the regions of false duplication we found in bTaeGut1.0, the homologous regions of false duplication were extracted by Hal (Hickey et al., 2013). We calculated the uncorrected amount of false duplications in bTaeGut1.4 from each false duplication in bTaeGut1.0 as follows: Given a length of homologous sequence H of a false duplication (*FD*) from new (*v1.7*) and prior (*v1.0*) VGP zebra finch in an alignment block, an uncorrected false duplication was calculated as *uncorrected*

$FD = \Sigma H_{v1.7} - (\Sigma H_{v1.0} - FD)$. If the uncorrected false duplications were ≤ 0 bp, we regarded that false duplication was corrected in the bTaeGut1.4 assembly.

2.3.6 Duplicated *k-mers* in different hummingbird assembly approaches

We calculated *k-mer* duplications for each experimental hummingbird assembly generated by Rhie et al. (2021) for assessing the relative magnitude of introducing or removing false duplications by various assembly algorithms and steps. The assemblies are available in GenomeArk prefix on 's3://genomeark/working/release1/scaffolding/' named as 'bCalAnn1_c1.fasta.gz', 'pac_fcn_p.fasta.gz', 'pac_nano_canu.fasta.gz', 'pac_canu.fasta.gz', '10x_spnv2_hap1.fasta.gz', 'ill_soap.fasta.gz', and the primary VGP assembly of bCalAnn1.0. 10X linked reads of the hummingbird were used for calculating *k-mer* multiplicity. Meryl (Rhie et al., 2020) and Merqury (Rhie et al., 2020) were performed to obtain intermediate data points for analyzing *k-mer* duplications, with default options. *K-mer* duplications were counted by

'false_duplications.sh' in the Merqury package.

2.3.7 Gene ontology enrichment test for falsely duplicated genes

We tested gene ontology enrichment for the false gene gains, false exon gains, and false chimeric gene gains of each prior assembly. We used g:Profiler (Reimand et al., 2007) on the web (https://biit.cs.ut.ee/gprofiler/gost) for functional profiling of these genes. Because the many false gene gains were fragmentary artifacts such as '−like' gene, we converted the gene symbol of these false gene gains using the original gene product name. g:Profiler supported the zebra finch and platypus in organism parameter selection, but the hummingbird was not supported. We thus selected the organism parameter 'zebra finch' for the hummingbird by considering the closest phylogenetic distance of species listed in the database. Significance was calculated by g:SCS with a threshold of $P < 0.05$. The list of ATP−binding genes were made up by referring to vertebrate ATP−binding genes in AmiGO 2 (http://amigo.geneontology.org/amigo/). The control gene set was constructed by randomly choosing genes as the same number of

ATP-binding genes for each species. Heterozygosity of the genes were calculated in each VGP assembly using the same method above. A significant difference of heterozygosity was tested by one-sided Wilcoxon rank-sum test.

2.3.8 False duplications in emu assemblies

The assembly and raw sequenced data of emu were collected from NCBI Assembly for both previous (GCA_013396795.1) and recent (GCA_016128335.1) assemblies. The short reads generated from both individuals of the assemblies were available in NCBI SRA. We mapped the Illumina reads constructed by the 800bp paired end library (run number: SRR9946765, SRR9946766, SRR9946768, SRR9947049, SRR9994342, SRR9994343, SRR9994348, SRR9994349, SRR9994351) to the previous assembly using Minimap2 with preset '-ax sr'. We mapped the 10X linked reads of the other assembly (run number: SRR11971566) to the recent assembly with the same step for 10X linked read mapping above. We ran the same pipeline for identifying false duplications in both assemblies, with filtering true duplication as above.

2.4 Results

2.4.1 Genome assemblies and identifying false duplications

The previous Sanger-based platypus (Warren et al., 2008) and zebra finch (Warren et al., 2010) reference genomes used standard pipelines for the best reference chromosomal level genomes at the time, generated with 500-1000 bp Sanger sequence reads, BAC-based scaffolding and FISH or cytogenetic chromosome mapping and assignments. No systematic effort was made for haplotype phasing, but both the previous zebra finch and platypus assemblies were rigorously manually curated. The prior Illumina-based Anna's hummingbird reference (Jarvis et al., 2014; Zhang et al., 2014) was generated with short reads (~150 bp), and contigging and scaffolding with multiple paired-end and mate-pair libraries ranging from 200 bp to 20 kbp in size. An effort was made to remove alternate haplotypes during assembly.

The VGP assemblies of the same species was generated with PacBio-based continuous long-read (CLR) contigs (N50 read length ~17 kbp), which were scaffolded with 10X Genomics linked reads, Bionano Genomics optical maps, and Arima Genomics Hi-C chromatin interaction read pairs (Rhie et al., 2021). Systematic

attempts to prevent false duplications were made, using FALCON-Unzip to separate haplotypes after generation of contigs and purge_haplotigs (Roach et al., 2018) that search for and purged false heterotype duplications from the primary pseudo-haplotype assembly (Rhie et al., 2021). All VGP assemblies were subjected to rigorous manual curation to minimize assembly errors generated by algorithmic shortcomings. The previous and VGP assemblies of the zebra finch and Anna's hummingbird were conducted on genomic DNA from the same individuals, and thus differences would only be due to sequencing platform and assembly methods. As the platypus was from a different individual, we performed some additional steps later in the study to validate whether the issues found were due to sequence and assembly errors, and not individual biological differences.

The size of the previous assemblies of the zebra finch, hummingbird and platypus are 1.23 Gbp, 1.11 Gbp and 2.00 Gbp, respectively (Table 2. 1). They consisted of a total of 37,421, 54,736, and 958,970 scaffolds. Among the scaffolds, 35 and 19 super scaffolds were assigned to chromosomes for the zebra finch and platypus assemblies, respectively. The assemblies had 87,710, 70,084, and 243,835 gaps, and their average contig NG50s were 47.9,

27.0, and 11.4 kbp, respectively. The size of the VGP assemblies were all 0.05–0.17 Gbp smaller (Table 2. 1). They consisted of ~280 to 3,140–fold fewer scaffolds (i.e. 135, 159, and 305 total), of which 33 (now 39 in our updated version), 33, and 31, respectively, were assigned to chromosomes, including the sex chromosomes. The number of gaps likewise were ~160 to 470–fold lower, and contig NG50s were ~250 to 1,320–fold higher: 12.0, 13.4 and 15.0 Mbp for the zebra finch, hummingbird and platypus, respectively. Alternate haplotype scaffolds of 0.95–1.58 Gbp in total size were separated from the primary assembly.

We performed self–alignment of each assembly using Minimap2 (Li, 2018) as a part of the purge_dups (Guan et al., 2020) process to detect duplications independently from another assembly; purge_dups was created by members of the VGP in order to identify and purge false duplications in different contigs. Also, we aligned the previous assemblies to the new VGP assemblies of each species using the reference–free Cactus aligner (Paten et al., 2011), which allows pair–wise detection of duplicates between the previous and new assemblies at the sequence and contig levels (Figure 2. 2a, b). We distinguished false duplications from true duplications, as we found that the former had read coverage at the haploid–level, gaps

between duplications due to mis-assembly, and discordance in 10X linked read pairs mapped back to the assembly. We classified each false duplication as heterotype duplications when heterozygous *k-mers* were found, and homotype duplications when read depth coverage was lower than the haploid-level, which occurs with sequence read errors, or when heterozygous *k-mers* were not found (Figure 2. 2c).

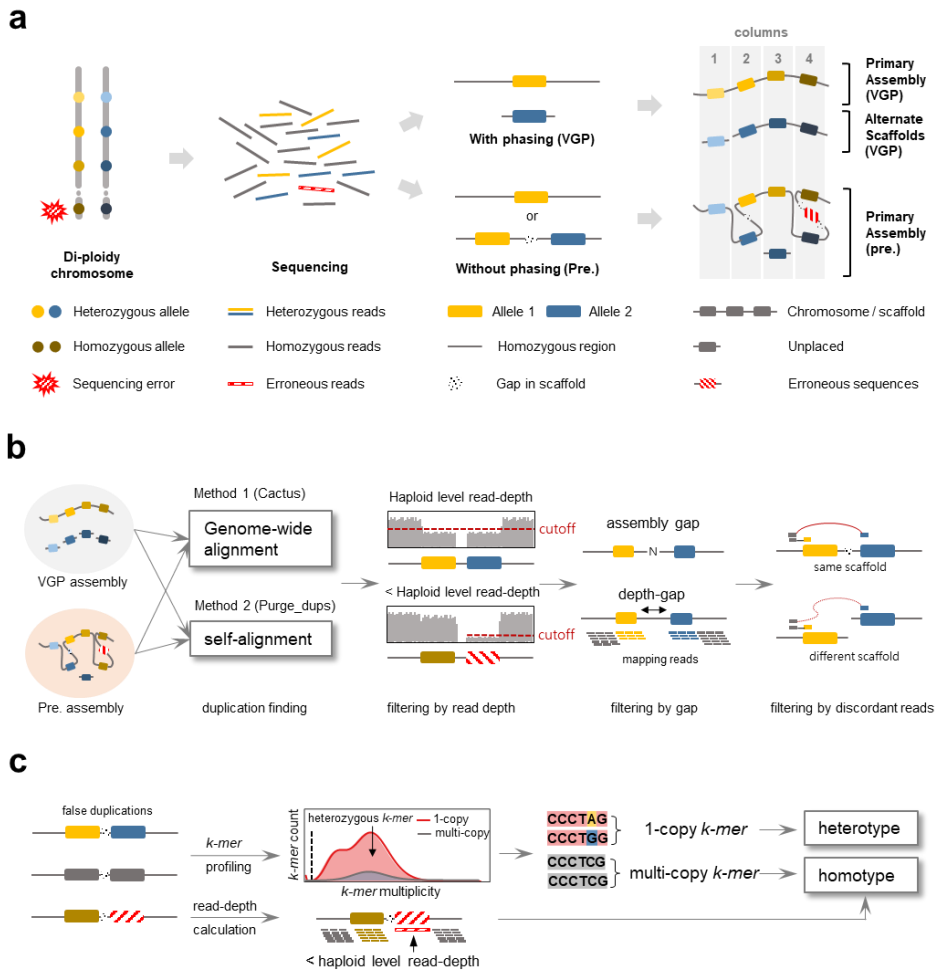

Figure 2. 2 Overview to identify false duplication. a, Mechanisms of how false assembly duplications are created. If haplotype phasing is included and correctly performed in the assembly process, there will be only one allele in the primary assembly, with the other placed in the alternate assembly (right panel, Column 1). However, without proper phasing, both alleles of heterozygous loci may be assembled in one scaffold (Column 2) or two different scaffolds (Column 3) of the primary assembly. Alternatively, randomly or systematically

piled up erroneous sequencing reads can generate false duplications (Column 4). This leads to three types of false duplications. b, Scheme to identify false duplications. Whole-genome alignment between the two assemblies using Cactus and self-alignment using purge_dups reveal candidate false duplicated regions or whole contigs. The union-set from these two independent methods is then used to find false duplications, which contain some combination of near haploid read-depth of the 10X Genomics linked reads, the presence of gaps between duplications, and discordance in read pairs between duplications. c, Scheme to classify false duplication types. Copy number and multiplicity of *k-mers* are calculated from the assembly and the 10X Genomics linked reads respectively, and used to classify false duplications as heterotype or homotype. Heterotype duplication includes haploid specific *k-mers* (i.e. 1-copy). Homotype duplication does not include haploid specific *k-mers*, but does include sequencing errors that can be detected by read-depth below the haploid-level.

2.4.2 False duplications in previous and VGP assemblies

The distributions of 10X Genomic linked read depth coverage (**Figure 2. 3**) and *k−mer* multiplicity (**Figure 2. 4**) showed that previous assemblies included significant amounts of false duplications: 16% (196 Mbp), 4% (41 Mbp), and 6% (126 Mbp) of the total length of the prior zebra finch, Anna's hummingbird, and platypus assemblies, respectively (**Figure 2. 5a, Table 2. 2**). As the 10X Genomics linked reads were generated on the new platypus individual used, we also found the Sanger raw reads of generated from the prior individual in the NCBI Trace Archive and found 104 Mbp of haploid coverage lower than the genome−wide average indicating that the vast majority of the 126 Mbp found with the 10X linked reads are not due to individual differences, but false duplications. This is a whole chromosome's worth of false duplication (6% of the genome), and thus also unlikely due to individual differences. The higher levels of false duplication found with the 10X linked reads could be due its 10−fold higher sequence coverage (60X) relative to the Sanger read coverage (6X). For all three previous assemblies, heterotype was the major source of false duplication, an order of magnitude higher than the homotype except for the previous Anna's hummingbird assembly (**Figure 2. 5a, Table 2. 2**). Of the total false duplications, 7

to 24% were on the same scaffold (Table 2. 2).

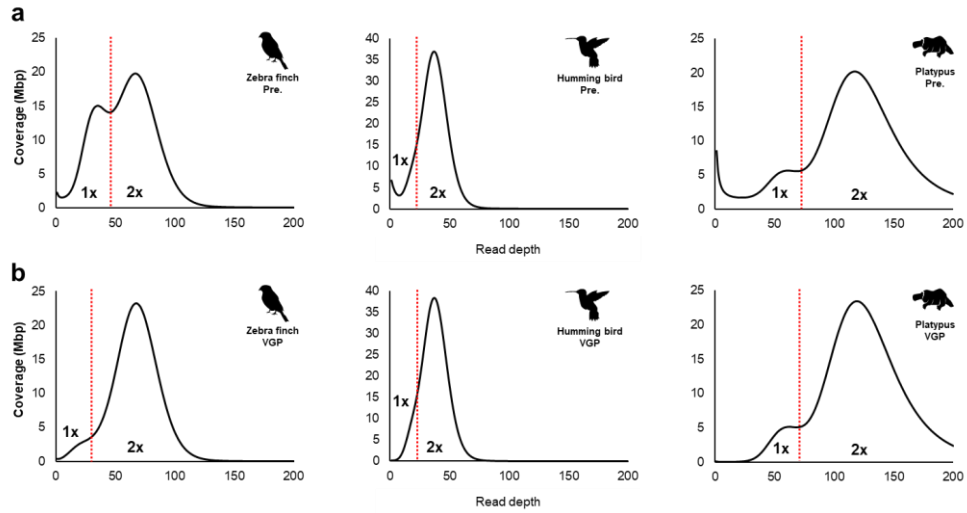

Figure 2. 3 Depth-coverage profiling of all assemblies. a, Prior assemblies. b, VGP assemblies. The 10X linked read depth-coverages of every site is summarized as a distribution. The red line shows the threshold of depth-coverage that we used to determine false duplications (to the left of the red line). The bimodal distribution in the zebra finch and the platypus assemblies are caused by highly heterozygous regions.

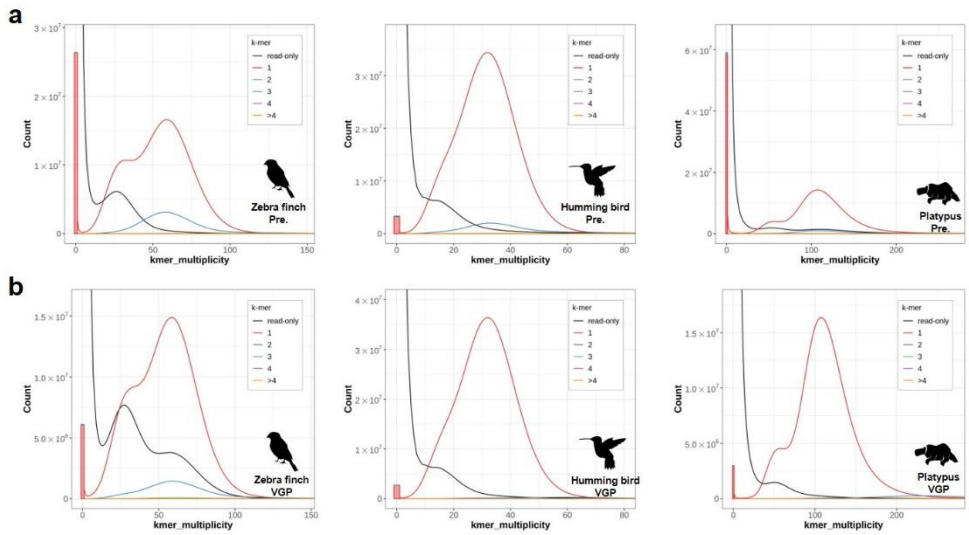

Figure 2. 4 *K−mer* profiling for all assemblies. a, Prior assemblies. b, VGP assemblies. From the sequences of 10X linked reads and assemblies, *k−mer* multiplicity was calculated. The x−axis is the *k−mer* multiplicity calculated from the raw reads, the y−axis are the counts, and the numbers in the boxes represent the *k−mer* multiplicity found in the primary pseudo−haplotype assembly. *K−mer* multiplicity of 2 copies or higher under the area of single copies (red) are overly represented as false duplications.

False duplications in the VGP assemblies were still present, but 7 to 22−fold less: 2.3% (24 Mbp), 0.5% (5.8 Mbp), and 0.3% (5.6 Mbp) of the total primary assembly in the zebra finch, hummingbird and platypus, respectively (**Figure 2. 5a, Table 2. 2**). Heterotype was also the major type of false duplication. In contrast to the prior assemblies, there was a much higher proportion of the false duplications, 61−92%, found on the same scaffold in the VGP assemblies, due to improved scaffolding using multiple long−range sequencing platforms.

Table 2. 2 False duplication statistics in previous and VGP assemblies.

		Zebra finch Pre. (Sanger)	Zebra finch VGP (bTaeGut1)	Hummingbird Pre. (Illumina)	Hummingbird VGP (bCalAnn1)	Platypus Pre. (Sanger)	Platypus VGP (mOrnAna1)
Heterozygosity	**(%)**	-	0.95	-	0.34	-	0.22
Total false duplication length	**(Mbp)**	195.6 (15.9 %)	24.1 (2.3 %)	40.9 (3.7 %)	5.8 (0.5 %)	126.0 (6.3 %)	5.6 (0.3 %)
Type	**Heterotype (Mbp)**	190.5 (15.5 %)	23.1 (2.2 %)	13.7 (1.2 %)	5.4 (0.5 %)	72.0 (3.6 %)	5.0 (0.3 %)
	Homotype (Mbp)	5.1 (0.4 %)	0.9 (0.1 %)	27.2 (2.5 %)	0.4 (<0.1 %)	54.0 (2.7 %)	0.6 (<0.1 %)
Location	**Same scaffold (Mbp)**	46.4 (3.8 %)	22.2 (2.1 %)	2.9 (0.3 %)	4.8 (0.5 %)	22.7 (1.1 %)	3.4 (0.2 %)
	Different scaffold (Mbp)	149.3 (12.1 %)	1.9 (0.2 %)	38.0 (3.4 %)	1.0 (<0.1 %)	103.3 (5.2 %)	2.2 (0.1 %)
Total assembly Length	**(Mbp)**	1232.1 (100 %)	1058.0 (100 %)	1105.7 (100 %)	1059.7 (100 %)	1995.6 (100 %)	1858.5 (100 %)

Heterozygosity (top row) was calculated as the mean number of variants in each assembly based on the 10X linked reads produced for each VGP assembly. Second row are the total Mbp (and % of genome size in brackets) that are falsely duplicated. Rows below that are the type and location of false duplications.

2.4.3 High heterozygosity and sequencing errors associated with false duplications

The heterozygosity of false duplications found in the previous assemblies that were corrected in the VGP assemblies (Corrected FD regions; **Figure 2. 5b**) were all ~1.5 to 1.8−fold higher than correctly assembled regions without false duplications in both the previous and VGP assemblies (P < 0.001; **Figure 2. 5c**). The heterozygosity of false duplications specific to the VGP assembly (Introduced FD regions; **Figure 2. 5b**) were all also higher (P < 0.001) with no specific absolute level that differed with the previous assemblies (**Figure 2. 5c**). We also found more erroneous *k−mers* in false duplications than in the correctly assembled regions in both the previous and VGP assemblies (**Figure 2. 5d**). Further, regions between the false duplications were most often separated by an assembly gap and sometimes connected by unsupported sequence read depth gaps, due to incorrect gap filling or other assembly errors (**Figure 2. 6; Figure 2. 1**). These properties were not found for true duplications, including of the acrosin (*ACR*) gene and an allele specific tandem duplication we found in the same contig with haploid level read depth coverage (**Figure 2. 7**). These findings show that

increased relative heterozygosity, especially those at the boundaries of homozygous and heterozygous sites, and sequencing errors are prone to be falsely duplicated.

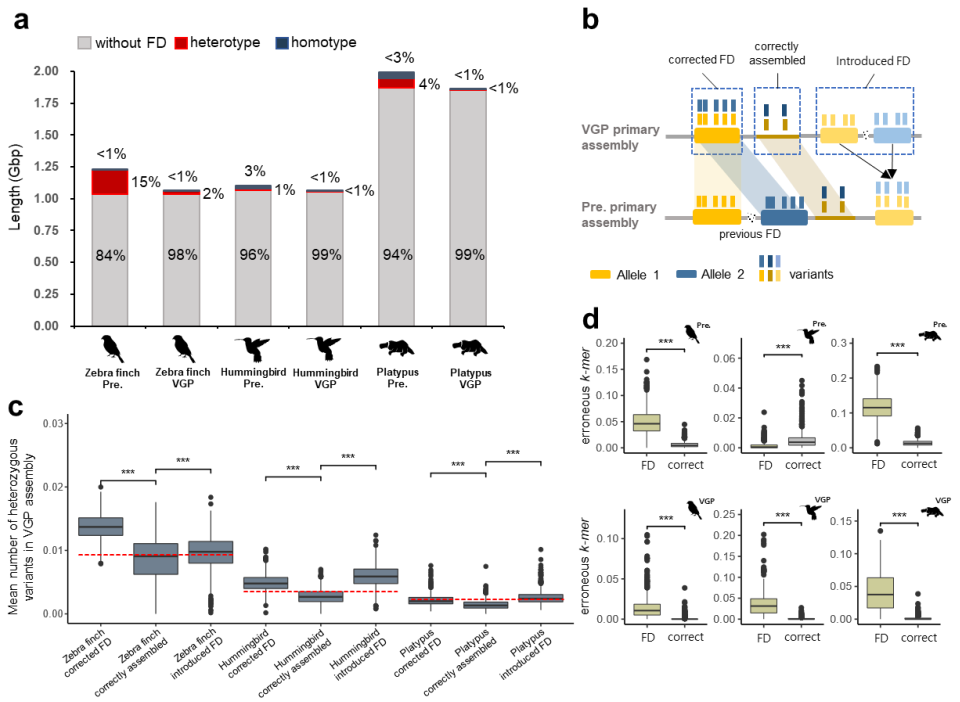

Figure 2. 5 The amount of false duplication and factors that correlate with false duplication. a, The total assembly size and the proportion that are false duplications in the previous and VGP assemblies. False duplications were classified as heterotype and homotype. b, Scheme of false duplications (FD) in the previous and VGP assemblies due to heterozygous alleles. Corrected FD are regions in the VGP assembly that are false duplications in the previous assembly. Correctly assembled are regions without any false duplication in the previous and VGP assemblies. Introduced FD are false duplications introduced in the VGP assembly that were not present in the previous assembly. c, Heterozygosity of corrected FD, correctly assembled, and

introduced FD, according to the VGP assembly haplotype data (***P < 0.001; two-sided T-test). Red dotted line, overall heterozygosity of the genome. d, The portion of erroneous *k-mers* in false duplications and correct regions of each assembly (***P < 0.001; two-sided T-test).

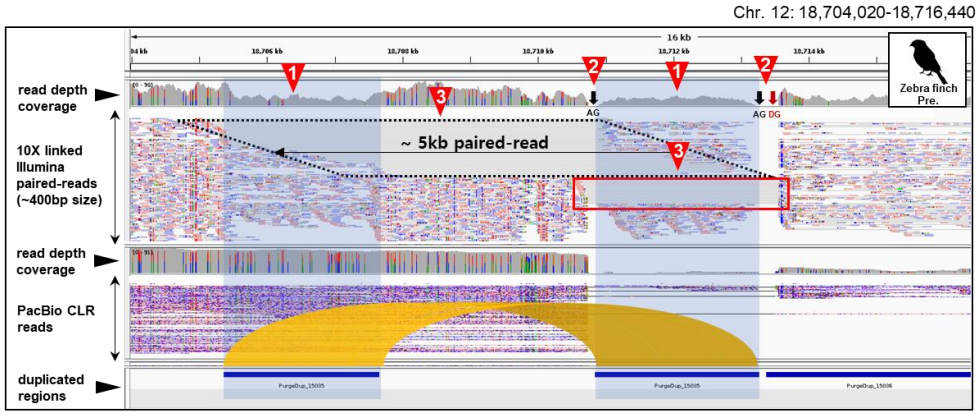

Figure 2. 6 The presence of a gap and discordant reads between false duplications. Shown is a locus in the previous zebra finch assembly with a false duplication. 10X linked read alignments are shown above the PacBio CLR read alignments, along with the depth coverage of the respective read data. Characteristics of false duplications are marked with red triangles: 1) Nearly half depth-coverage and lack of heterozygous variants – colors indicating nucleotide heterozygosity; 2) Gaps between false duplications; and 3) Discordant 10X linked reads (black dotted box). Red box, discordant reads found near the end of scaffolds that should be connected to each other. AG, assembly gap. DG, depth-gap (unsupported sequences by reads; see methods).

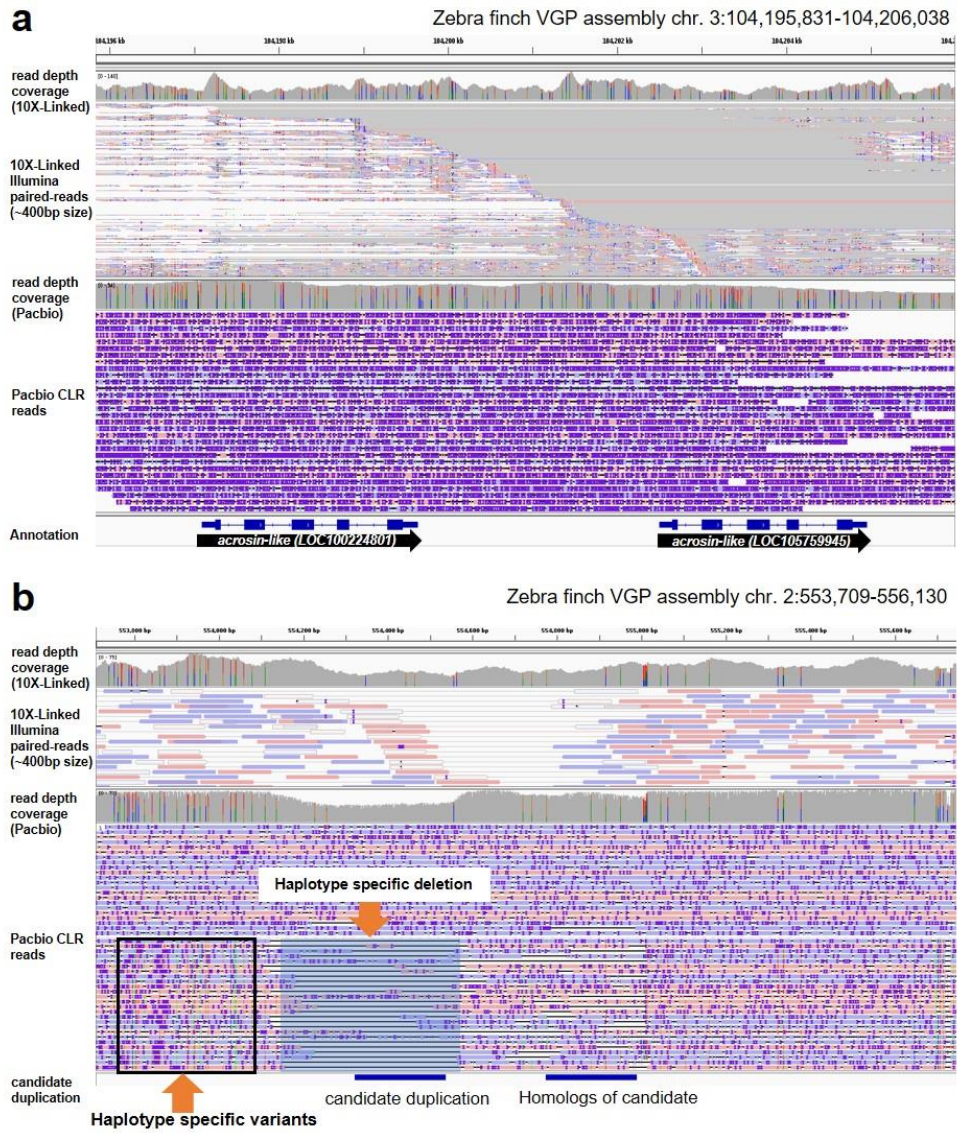

Figure 2. 7 True duplication in a VGP assembly. 10X linked reads are shown as paired read alignments above the PacBio CLR read alignments, along with the depth coverage of the respective read data. a, True gene duplication of the acrosin (*ACR*) gene in the zebra finch. The 10X linked read alignment shows discordant read alignments but has no signature of a depth-gap or decreased read-depth coverage.

PacBio CLR reads alignments connect these duplicated genes with no gaps, as single molecule reads, and no unsupported sequence. b, True haplotype specific sequence duplication with lower read depth in the zebra finch. A candidate duplication was identified as a true genomic duplication, but in one haplotype. The 10X linked read alignment shows discordant read alignments but has no signature of a depth-gap. Half of the PacBio CLR read alignments show the allele specific duplication, while the other half show a deletion on one of the two alleles.

2.4.4 False duplications cause false annotation errors

Among the false duplications, we found 4 to 24% of the coding genes were impacted in the previous assemblies, depending on species (**Figure 2. 8a**). Of these, we found three main types: 1) the majority being false gene gains [FGG] of nearly the entire coding sequence; 2) followed by false exon gains [FEG] within a gene; and 3) a minority being false chimeric gains [FCG] from a chimeric join among exons from different genes (**Figure 2. 8a, b; Table 2. 3**).

An example of a FGG included *ZBTB11* in the previous zebra finch assembly, which had 9 of the 11 coding exons falsely duplicated and annotated as *ZBTB11*-like (*LOC100218125*; **Figure 2. 8c**). The non-duplicated *ZBTB11* exon 1 was included in *ZBTB11*-like and exon 2 (red mark; 10th exon) included in *ZBTB11*, while these exons were assembled into one gene in the VGP assembly. The sequence alignment landscape of *ZBTB11* in the previous assembly showed typical characteristics of a false gene gain (**Figure 2. 9a**), whereas there was no sign of false duplications in the VGP assembly (**Figure 2. 9b**). The gamma-aminobutyric acid receptor subunit gamma 2 (*GABRG2*) was a complex example, where several false exon duplications were assembled in the same scaffold and annotated as a *GABRG2*-like (*LOC101232861*) or as another *GABRG2*-like

(LOC100229343) on another scaffold, both with presumed false exon losses after the duplication from the original gene (**Figure 2. 8d**). Because true duplications can also be annotated as gene name−like, for example *ACR* and *ACR*−like (**Figure 2. 7a**), the "like" term in the NCBI annotation can not be taken alone as evidence of a false duplication.

ALDH2, a gene with specialized upregulation in the zebra finch vocal learning nucleus HVC (Denisenko−Nehrbass et al., 2000), had three false duplicated exons that were incorporated into the adjacent *ACAD10* gene, causing a FCG for *ALDH2*−*ACAD10*, all with gaps around each of the false duplications (**Figure 2. 8e; Figure 2. 9c**), none present in the VGP assembly (**Figure 2. 9d**). The calcium voltage−gated channel subunit alpha1 H (*CACNA1H*), also a gene with specialized expression in vocal learning circuits of the zebra finch (Friedrich et al., 2019; Kurz et al., 2010, p. 1), had a FEG in the second exon (**Figure 2. 8f**). Similar examples of FGG, FCG and FEG in the previous Anna's hummingbird and platypus assemblies are shown in **Figure 2. 10**. This includes false duplications that overlap in the CDSs of *ATF3*, *PCBD1* and *VAMP4* in the previous hummingbird assembly; and of *ZP2*, *UPF2* and *HSF2* in the previous platypus assembly with haploid−level Sanger read coverage (**Figure 2. 11**).

We next scanned the literature for reported cases of gene duplications in one or more of the three species studied here, and assessed whether they were real or false duplications. There were many cases where the gene duplications were real, but also multiple cases where they were false. An example of the later included *ADAMTS13,* related to thrombotic thrombocytopenic purpura in humans (Levy et al., 2001; Quesada et al., 2010), which was reported as duplicated in the zebra finch (Quesada et al., 2010). But we found that two of the three *ADAMTS13* genes (one *ADAMTS13* and two *ADAMTS13*–like) were falsely duplicated in the same (*LOC105760960*; **Figure 2. 9e**) and a different scaffold (*LOC101232819*; **Figure 2. 12**), respectively. These two false duplications were produced from the 5' and 3' ends of the original *ADAMTS13* gene (**Figure 2. 9f**). We confirmed that there were no additional copies of *ADAMTS13* in both the VGP zebra finch assembly and a recent VGP chicken assembly (GRCg7w; Accession# GCA_016700215.2). Another example was neurotrypsin, a gene known to be linked to neural development, which was represented as having more copies in the zebra finch than chicken (Warren et al., 2010). But we found that this extra copy of the gene (*LOC100217566* in a short unplaced scaffold ~3 kbp long) was made by a false

duplication the original neurotrypsin gene in chromosome 6 (*LOC100229828*; **Figure 2. 9g, h**), which is annotated as *NTL* in GRCg7w. This extra copy of the gene was not found in both the VGP zebra finch and GRCg7w assemblies. A third example was a platypus vomeronasal receptors (V1R) gene family expansion reported as a sensory adaptation for underwater life history (Warren et al., 2008); we found that 43 of the 267 annotated V1R genes (16%) are actually false duplications in the previous assembly (**Table 2. 4**). In examples we examined, in the VGP assemblies we found single molecule PacBio reads that crossed the assembly or read depth gaps, or contig ends, found in the prior assemblies, without the presence of a real duplication of the gene(s) (**Figure 2. 9**), experimentally validating them as false duplications in the prior assemblies, and not computational errors.

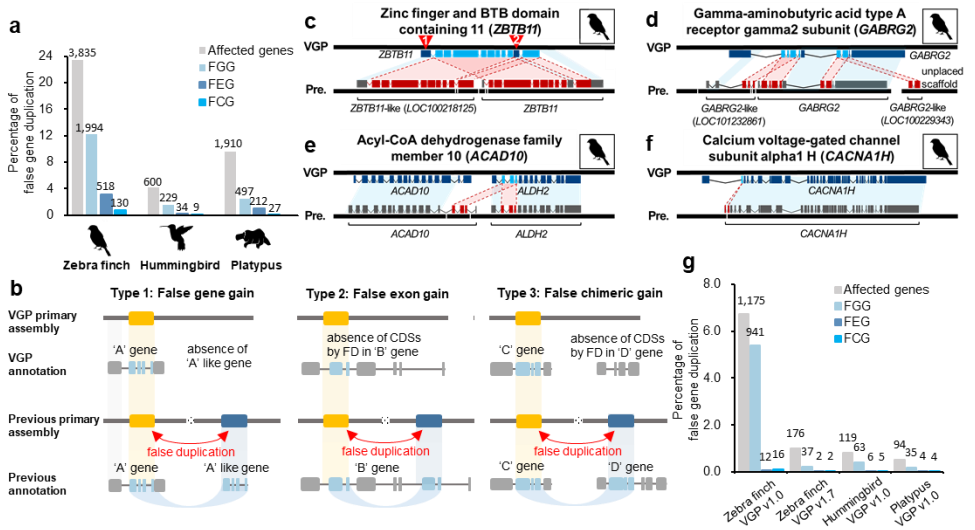

Figure 2. 8 Mis-annotations due to false duplications. a, Amount and percentage of all genes with mis-annotations caused by false duplications in the previous assemblies. The amount of genes of each type is shown on the top of each bar graph. b, Types of mis-annotations caused by false duplications. When >50% of the CDS length of a gene was falsely duplicated and annotated as another gene, and resulted in two genes with similar function (e.g. -like), we classified it to false gene gain (FGG, Type 1). When an exon within a gene was falsely duplicated, we classified it to false exon gain (FEG, Type 2). If the duplicated exon was falsely inserted to another existing gene of different function, we classified it as a false chimeric gain (FCG, Type 3). c, FGG of *ZBTB11*. d, FGG of *GABRG2*. e, FCG involving *ACAD10* and *ALDH2*. f, FEG within *CACNA1H*. The red

lines represent the connection between false duplications and the homologs in the VGP assembly. The blue boxes represent the homologous region between the VGP and previous assemblies. The white spaces in the black bars represent scaffold assembly gaps. g, Amount of false gene annotation in VGP assemblies. The zebra finch VGP v1.0 assembly had false duplications purged with purge_haplotigs after scaffolding; the zebra finch VGP v1.7 had false duplications purged with purge_dups before scaffolding.

Table 2. 3 Mis−annotations caused by false duplications in both previous and VGP assemblies. The mis−annotation cases include false gene gain (FGG), false chimeric gain (FCG), and false exon gain (FEG). If the CDS overlapped caused by the false duplication was found in both −like gene and the original genes, they were named to FGG (mixed). This occurs by false duplication without haplotype phasing. Gene IDs and gene symbols are represented based on NCBI. Data only appears up to No. 15 from Ko. et al. (2022).

Data No.	Species	Assembly	Gene ID	Gene symbol	Total No. CDSs	Total CDS length (bp)	FD overlap length %	Error type	Product name
1	Zebra finch	Previous	105758872	LOC105758872	13	1581	78.5	FGG	eukaryotic translation initiation factor 2D-like
2	Zebra finch	Previous	100217725	HERC3	25	3153	4.2	FGG (mixed)	LOW QUALITY PROTEIN: probable E3 ubiquitin-protein ligase HERC3
3	Zebra finch	Previous	100227877	LOC100227877	7	665	100	FGG (mixed)	probable E3 ubiquitin-protein ligase HERC3
4	Zebra finch	Previous	105760054	LOC105760054	16	2334	100	FGG	LOW QUALITY PROTEIN: obscurin-like
5	Zebra finch	Previous	105760339	LOC105760339	3	324	100	FGG	krev interaction trapped protein 1-like
6	Zebra finch	Previous	101233548	LOC101233548	3	619	100	FGG	LOW QUALITY PROTEIN: ATP-sensitive inward rectifier potassium channel 8-like
7	Zebra finch	Previous	100223188	LOC100223188	4	1520	100	FGG	TRPM8 channel-associated factor 2-like
8	Zebra finch	Previous	100230937	LOC100230937	6	701	100	FGG	LOW QUALITY PROTEIN: protein-lysine methyltransferase METTL21E-like
9	Zebra finch	Previous	105758939	LOC105758939	7	1155	74.9	FGG,FCG	basigin-like
10	Zebra finch	Previous	105760144	LOC105760144	2	349	100	FGG	CUB and sushi domain-containing protein 2-like
11	Zebra finch	Previous	100225740	STK4	13	1572	5.7	FEG	serine/threonine-protein kinase 4
12	Zebra finch	Previous	101234012	LOC101234012	2	249	100	FGG	NTF2-related export protein 2-like
13	Zebra finch	Previous	100224419	LOC100224419	5	438	100	FGG	pantothenate kinase 3-like
14	Zebra finch	Previous	100220232	TEX2	15	3192	1.7	FEG	LOW QUALITY PROTEIN: testis-expressed sequence 2 protein
15	Zebra finch	Previous	100225264	LOC100225264	4	676	100	FGG	dipeptidyl peptidase 1-like

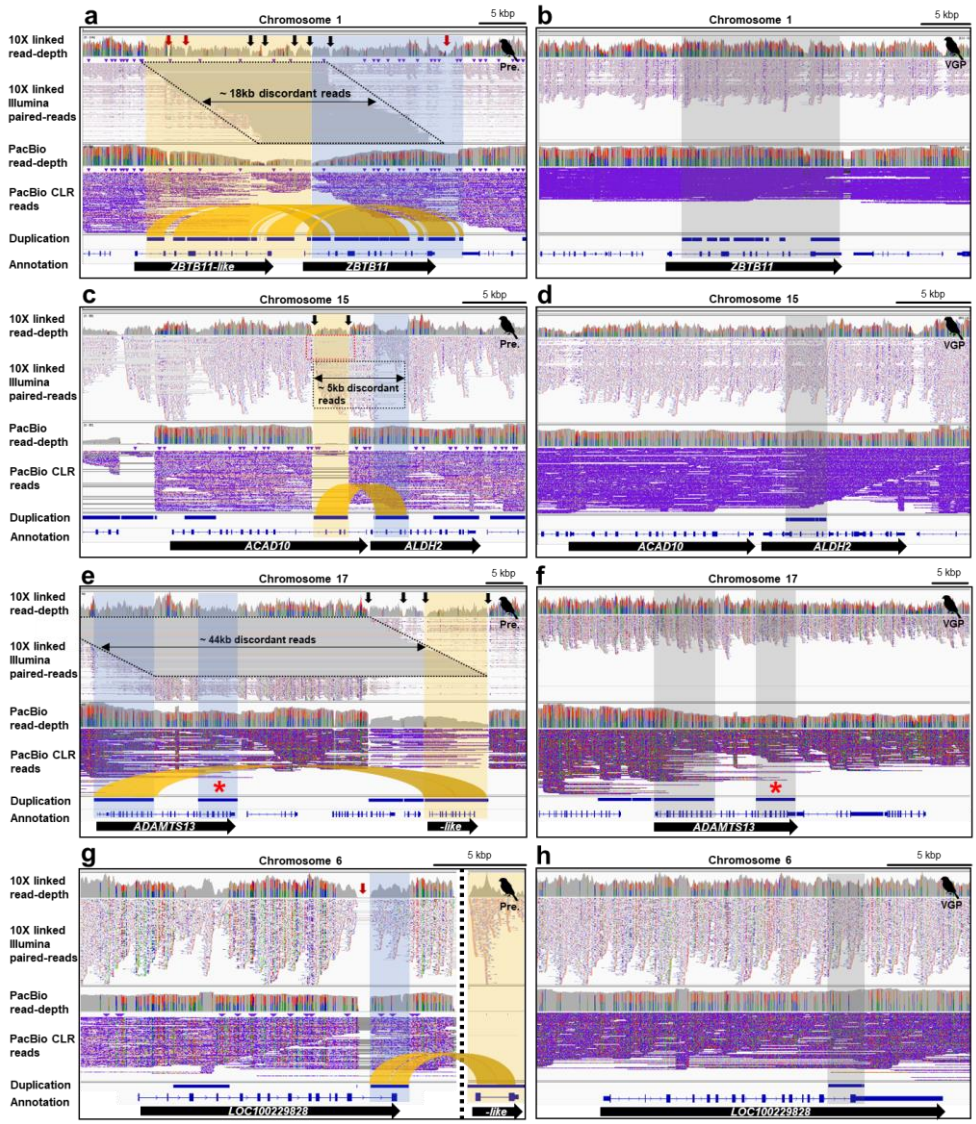

Figure 2. 9 The genome landscape of false gene gains. 10X linked read pairs are shown above the PacBio CLR reads along with the depth coverage of the respective read data. a, The genome landscape of *ZBTB11* and false *ZBTB11*-like (*LOC100218125*) genes in the previous zebra finch assembly. Most of the region of *ZBTB11* was duplicated adjacent to itself in the previous assembly (highlighted as

orange and blue) and showed typical characteristics of false heterotype duplications. Black and red arrows represent assembly gap and read depth−gap, respectively. b, Corrected gene structure in the VGP assembly (grey). c, The genome landscape of *ACAD10* and *ALDH2* genes in the previous zebra finch assembly. Three exons of *ALDH2* were inserted in *ACAD10* by a false duplication (highlighted as orange and blue). Black arrows represent assembly gaps. d, Corrected gene structure in the VGP assembly. The extrinsic three exons from *ALDH2* (grey) were not found in *ACAD10* of the VGP assembly. e, The genome landscape of *ADAMTS13* and *ADAMTS*−like (*LOC105760960*) genes in the previous zebra finch assembly. The 5' and 3' exons of *ADAMTS13* (highlighted as blue) were falsely duplicated to two *ADAMTS13*−like genes, which are assembled in the same scaffold (highlighted as orange) as *ADAMTS13*−like (*LOC105760960*) and a different scaffold (*LOC101232819*; **Figure 2. 12**). The homologous region of the *LOC101232819* gene is marked by a red star. f, Corrected gene structure in the VGP assembly (grey). The homologous region of the *LOC101232819* gene in the previous assembly is marked by a red star. g, The genome landscape of neurotrypsin (*LOC100229828*) and neurotrypsin−like genes (*LOC100217566*) in the previous zebra

finch assembly. The 3’ region of *LOC100229828* (highlighted as blue) was falsely duplicated to a different small scaffold (~3 kbp; highlighted as orange), NW_002201465. h, Corrected region in the VGP assembly (highlighted as grey). The different colors and their heights in the read depth rows are the proportion of sites in reads with haplotype variants.

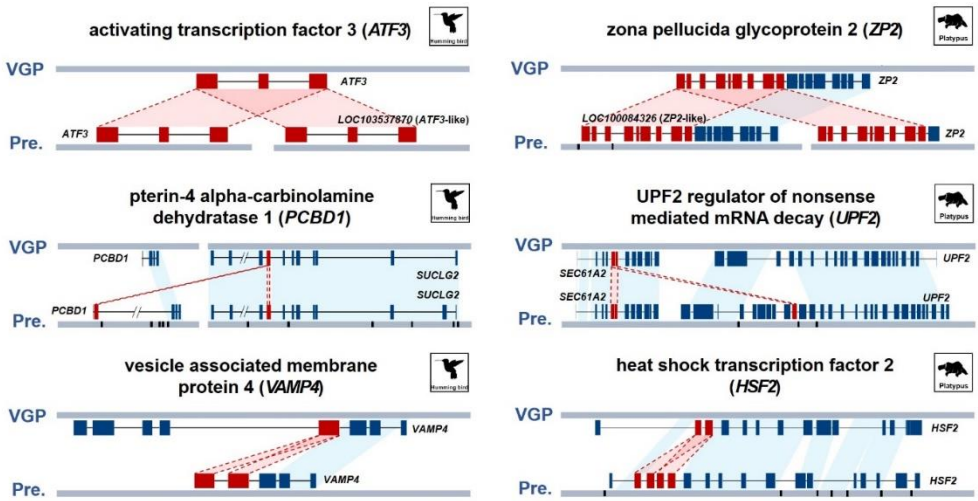

Figure 2. 10 Cases of false gene gain annotations in the prior hummingbird and platypus assemblies. Top row of each alignment shows the VGP 1.0 assembly structure and annotation. Bottom row shows the previous assembly structure and annotation. The red lines represent boundaries of the false duplicated exons in the prior assemblies that are correctly assembled in the VGP assembly. The blue boxes represent the correctly assembled exons in both the previous and VGP assemblies. The black bars represent the assembly gaps in the scaffolds.

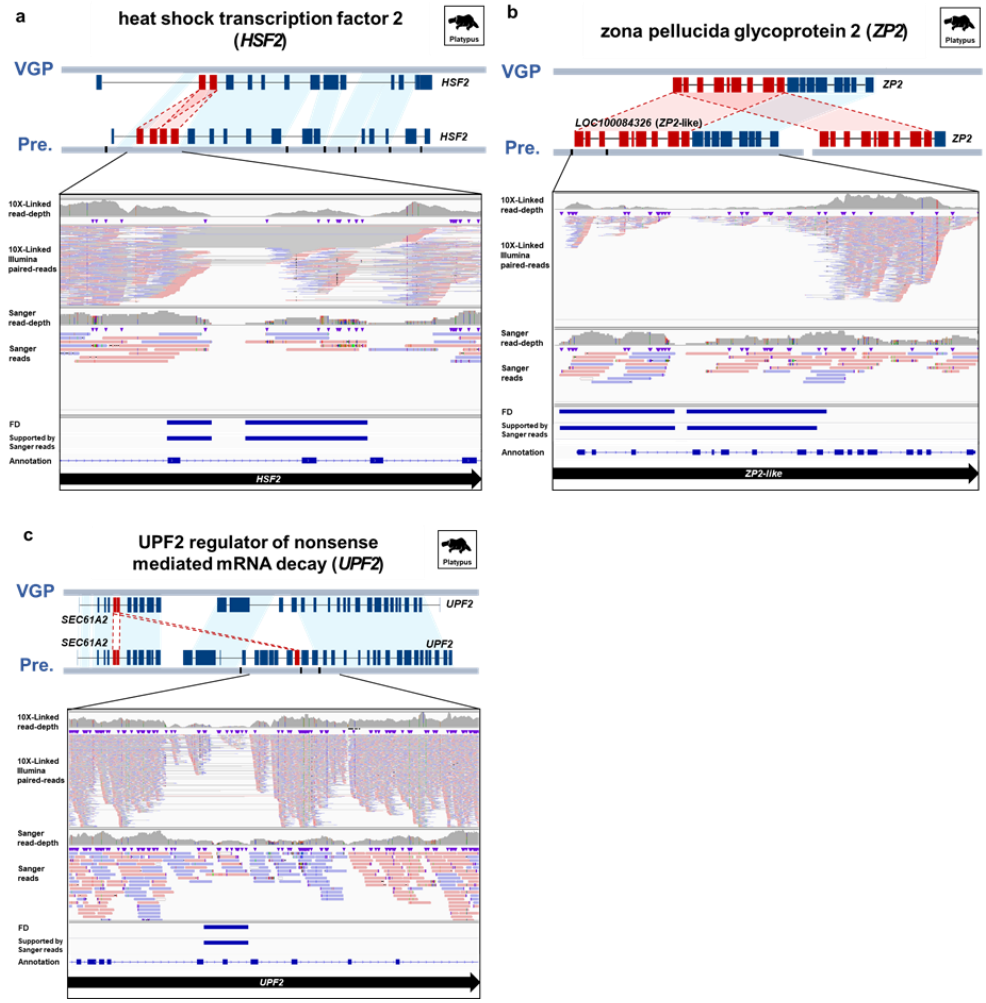

Figure 2. 11 Genome landscape of platypus assembly false duplications using Sanger reads. a, *HSF2* false duplication, b, *ZP2*–like false duplication. c, *UPF2* false duplication. 10X linked reads are shown as paired read alignments above the Sanger read alignments, along with the depth coverage of the respective read data. 'FD' is the false duplication identified without Sanger reads. The region of false duplication that was supported by Sanger reads with under haploid level low read coverage is represented below 'FD'.

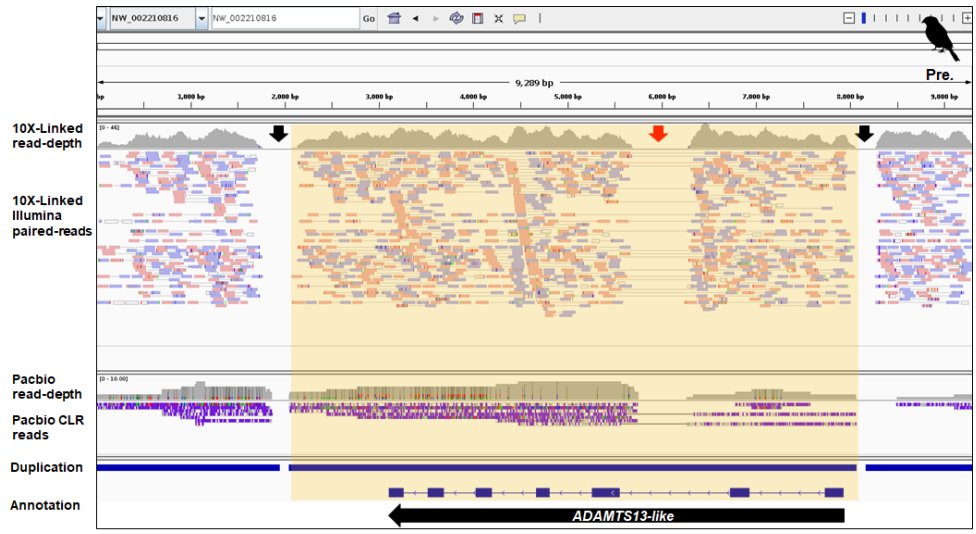

Figure 2. 12 Additional findings for the genome false duplication landscape of the *ADAMTS13*-like gene. This segment of the *ADAMTS13* gene is located in an unplaced scaffold of the previous zebra finch assembly. 10X linked reads are shown as paired read alignments above the PacBio CLR read alignments, along with the depth coverage of the respective read data. Black and red arrows represent an assembly gap and depth-gap, respectively.

Table 2. 4 False duplication of V1R family genes in the previous platypus assembly. Only V1R false duplications are listed. There were a total of 43 of 44 V1R genes with 100% of the gene sequence was a false duplication.

Data No.	Gene ID	Gene symbol	Total CDS length (bp)	FD overlap length %
1	100087537	ORNANAV1R3036	930	100%
2	100086877	ORNANAV1R3135	957	100%
3	100089239	ORNANAV1R3193	999	100%
4	100087756	ORNANAV1R3218	960	100%
5	100090986	ORNANAV1R3182	927	100%
6	100083273	ORNANAV1R3250	972	100%
7	100091981	ORNANAV1R3156	957	100%
8	100092984	ORNANAV1R3205	930	100%
9	100077569	ORNANAV1R3071	957	100%
10	100084348	ORNANAV1R3274	930	100%
11	100086320	ORNANAV1R3153	942	100%
12	100089667	ORNANAV1R3162	933	100%
13	100090040	ORNANAV1R3126	939	100%
14	100090909	ORNANAV1R3201	975	100%
15	100081224	ORNANAV1R3089	930	100%
16	100091544	ORNANAV1R3075	900	100%
17	100081065	ORNANAV1R3206	930	100%
18	100087668	ORNANAV1R3178	930	100%
19	100076831	ORNANAV1R3100	927	100%
20	100090833	ORNANAV1R3121	927	100%
21	100086008	ORNANAV1R3190	957	100%
22	100084287	ORNANAV1R3273	918	100%
23	100086540	ORNANAV1R3003	930	100%
24	100310944	ORNANAV1R3040	930	100%
25	100073803	ORNANAV1R3280	930	100%
26	100079875	ORNANAV1R3044	957	6%
27	100092791	ORNANAV1R3217	942	100%
28	100090351	ORNANAV1R3221	940	100%
29	100083451	ORNANAV1R3204	930	100%
30	100083489	ORNANAV1R3224	918	100%
31	100083601	ORNANAV1R3138	942	100%
32	100074320	ORNANAV1R3117	927	100%
33	100091869	ORNANAV1R3122	918	100%
34	100310963	ORNANAV1R3087	930	100%
35	100083242	ORNANAV1R3114	999	100%
36	100090674	ORNANAV1R3226	927	100%
37	100078712	ORNANAV1R3175	930	100%
38	100085790	ORNANAV1R3108	948	100%
39	100078109	ORNANAV1R3072	1065	100%

40	100088762	ORNANAV1R3199	927	100%
41	100082524	ORNANAV1R3165	930	100%
42	100091992	ORNANAV1R3061	933	100%
43	100086733	ORNANAV1R3223	930	100%
44	100089591	ORNANAV1R3185	996	100%

Among non-coding sequences, long terminal repeats (LTRs) sequences of the zebra finch were reported to have expanded 2.5 times more than chicken (Consortium, 2004; Warren et al., 2010) and short interspersed nuclear elements (SINEs) were reported to be highly expanded in the platypus relative to other mammals (Warren et al., 2008). However, we found 18,757 copies of LTRs (21% of the total) were false duplications in the previous zebra finch assembly and 140,279 copies of SINEs (6.1% of the total) were false duplications in the previous platypus assembly (Table 2. 5). In the previous Anna's hummingbird assembly, 3 to 5% of LTRs, SINEs, and long interspersed nuclear elements (LINEs) were false duplications (Table 2. 5).

Table 2. 5 False duplications on transposable elements in previous assemblies. Long terminal repeats (LTRs), short interspersed nuclear elements (SINEs), and long interspersed nuclear.

Data No.	Species	Repeat family	Repeat name	Location	Repeat length (bp)	FD overlap length %
1	Zebra finch	LINE/CR1	CR1-J2_Pass	NC_011474:2949508-2950059	552	100.0%
2	Zebra finch	LINE/CR1	CR1-J2_Pass	NC_011467:69043373-69043821	449	21.4%
3	Zebra finch	LTR/ERVK	TguLTRK1a	NC_011469:58492613-58492917	305	92.8%
4	Zebra finch	LINE/CR1	CR1-Z2_Pass	NW_002207001:1839-2060	222	100.0%
5	Zebra finch	LINE/CR1	CR1AVI	NW_002217199:1-177	177	100.0%
6	Zebra finch	LTR/ERVK	TguLTRK1a	NW_002197418:2500-3038	539	100.0%
7	Zebra finch	LTR/ERVL	TguERVL2a3_LTR	NC_011483:3463508-3463810	303	22.8%
8	Zebra finch	LINE/CR1	CR1-L3A_Croc	NW_002233786:2161-2332	172	100.0%
9	Zebra finch	LTR/ERVK	TguLTRK1c	NW_002233786:3136-3579	444	100.0%
10	Zebra finch	LTR/ERVK	TguLTRK1c	NW_002233786:3593-3698	106	100.0%
11	Zebra finch	LTR/ERV1	TguERV2_LTR1b	NC_011467:31683607-31684046	440	100.0%
12	Zebra finch	LINE/L2	L2-1_CPB	NC_011462:46730872-46731461	590	100.0%
13	Zebra finch	LINE/CR1	CR1-L1_Tgu	NW_002229384:5029-5428	400	100.0%
14	Zebra finch	LTR/ERVL	TguLTRL2a4	NC_011462:61492926-61493759	834	15.2%
15	Zebra finch	LTR/ERVK	TguLTRK7a	NW_002213436:1803-2192	390	100.0%
16	Zebra finch	LTR/ERVL	TguERVL1b_LTR	NC_011493:34601035-34601442	408	5.9%
17	Zebra finch	LINE/CR1	CR1-J1_Pass	NW_002200403:3686-3788	103	100.0%
18	Zebra finch	LINE/CR1	CR1-Y2_Aves	NW_002197746:2277-2464	188	100.0%
19	Zebra finch	LINE/CR1	CR1-Y2_Aves	NW_002197746:2463-2619	157	100.0%
20	Zebra finch	LINE/CR1	CR1-Y2_Aves	NW_002197746:2641-2769	129	100.0%

Data only appears up to No. 20 from Ko. et al. (2022).

2.4.5 Specific categories of genes have higher levels of false duplications

To determine if genes with false duplications belong to specific functional categories or are random in function, we performed GO enrichment analyses for the false gene lists of each species of the previous assemblies. We found 42 GO molecular function terms and 3 KEGG pathways were significantly enriched in the platypus and zebra finch falsely duplicated genes (**Figure 2. 13**). Out of these, there were 8 GO terms enriched in both species, and all 8 were nucleotide binding functions. Even though the Anna's hummingbird results did not yield GO categories at our statistical cut off ($P < 0.05$), the highest ranking categories also included nucleotide binding functions (**Figure 2. 13**). The differences in significance values between species were correlated with the number of false duplications found, where more genes lead to greater significance. This included 'ATP-binding' genes in both zebra finch and platypus, and 5 'ABC transporters' (ATP-binding cassette transporters) in platypus and 8 in zebra finch as false duplications (**Table 2. 6**). We observed the 'ATP-binding' genes tend to show higher heterozygosity than the other genes (**Figure 2. 14**). The 'ABC transporters' are known as the one of the largest and oldest

superfamilies, in diverse living organisms from prokaryotes to vertebrates, and play key roles in encoding membrane proteins that transport diverse metabolites (Dean and Annilo, 2005; Yan et al., 2021). The extensive variation in this superfamily implies a high evolutionary divergence rate (Chen et al., 2010), leading to a higher prevalence in haplotype divergence (Skibinski and Ward, 1982).

	Platypus (*n*=601)		Zebra finch (*n*=2,276)		Anna's hummingbird (*n*=256)	
	adj. *p*	# genes	adj. *p*	# genes	adj. *p*	# genes
protein binding	1.0.E+00	142	9.9.E-08	985	1.0.E+00	55
catalytic activity	1.0.E+00	104	5.0.E-06	435	1.0.E+00	42
ion binding	5.1.E-01	97	4.6.E-04	446	1.0.E+00	40
anion binding	1.4.E-01	58	4.0.E-09	227	1.0.E+00	23
small molecule binding	2.4.E-01	59	1.0.E-06	224	1.0.E+00	23
carbohydrate derivative binding	5.6.E-02	56	3.2.E-05	200	1.0.E+00	24
nucleoside phosphate binding	1.4.E-01	55	1.5.E-07	203	1.0.E+00	22
nucleotide binding	1.4.E-01	55	1.5.E-07	203	1.0.E+00	22
purine nucleotide binding	2.6.E-02	53	1.0.E-05	177	1.0.E+00	22
purine ribonucleotide binding	1.9.E-02	53	9.5.E-06	176	1.0.E+00	22
ribonucleotide binding	2.5.E-02	53	1.8.E-05	176	1.0.E+00	22
catalytic activity, acting on a protein	1.0.E+00	44	2.2.E-03	194	1.0.E+00	23
purine ribonucleoside triphosphate binding	9.3.E-03	53	2.9.E-06	173	1.0.E+00	22
adenyl nucleotide binding	3.8.E-04	50	3.1.E-09	164	1.0.E+00	18
adenyl ribonucleotide binding	3.1.E-04	50	1.6.E-09	164	1.0.E+00	18
transferase activity	1.0.E+00	41	2.4.E-02	188	1.0.E+00	16
ATP binding	1.3.E-04	50	1.9.E-10	162	1.0.E+00	18
enzyme binding	1.0.E+00	35	1.5.E-02	169	1.0.E+00	15
protein-containing complex binding	1.0.E+00	20	4.6.E-04	123	1.0.E+00	12
transferase activity, transferring phosphorus-containing	1.0.E+00	24	1.2.E-03	94	1.0.E+00	12
cytoskeletal protein binding	1.0.E+00	20	1.9.E-02	93	8.2.E-01	13
kinase activity	1.0.E+00	21	2.2.E-03	81	1.0.E+00	10
molecular function regulator	1.0.E+00	16	3.3.E-03	105	1.0.E+00	10
phosphotransferase activity, alcohol group as acceptor	1.0.E+00	20	6.7.E-03	71	1.0.E+00	10
enzyme regulator activity	1.0.E+00	14	1.2.E-04	95	1.0.E+00	9
ATPase	1.1.E-03	25	8.3.E-05	66	1.0.E+00	5
enzyme activator activity	1.0.E+00	13	1.1.E-03	77	1.0.E+00	9
protein kinase activity	1.0.E+00	17	2.6.E-03	64	1.0.E+00	9
nucleoside-triphosphatase regulator activity	1.0.E+00	11	2.2.E-02	56	1.0.E+00	7
protein serine/threonine kinase activity	1.0.E+00	13	4.1.E-04	54	1.0.E+00	6
GTPase regulator activity	1.0.E+00	8	8.2.E-03	54	1.0.E+00	7
GTPase activator activity	1.0.E+00	8	2.7.E-02	50	1.0.E+00	7
active transmembrane transporter activity	1.0.E+00	11	2.1.E-03	44	1.0.E+00	2
cell adhesion molecule binding	1.0.E+00	5	7.9.E-03	59	1.0.E+00	2
guanyl-nucleotide exchange factor activity	1.0.E+00	2	4.8.E-03	31	1.0.E+00	3
cysteine-type peptidase activity	1.0.E+00	3	2.2.E-02	28	1.0.E+00	2
ATPase-coupled transmembrane transporter activity	8.8.E-01	7	2.2.E-02	20	1.0.E+00	1
primary active transmembrane transporter activity	1.0.E+00	7	3.2.E-02	20	1.0.E+00	1
cysteine-type endopeptidase activity	1.0.E+00	2	2.5.E-02	20	1.0.E+00	1
extracellular matrix binding	1.0.E+00	2	2.8.E-03	14		
protein self-association	1.0.E+00	2	2.8.E-03	14		
flavin adenine dinucleotide binding			5.3.E-03	17		
ECM-receptor interaction	1.0.E+00	2	4.4.E-02	16	1.0.E+00	1
Human papillomavirus infection	1.0.E+00	8	7.1.E-03	43		
ABC transporters	2.9.E-02	5	1.0.E+00	8		

Figure 2. 13 Gene ontology enrichment analysis of falsely duplicated genes. The gene ontology terms are shown in the first column. The number of falsely duplicated genes in the analysis is listed below the name of each species (*n*). The number of genes for each term (# genes) are represented in each species. The significant adjusted p-values ($P < 0.05$) are highlighted. The KEGG pathway terms are shown in the bottom.

Table 2. 6 Gene ontology enrichment analysis for the false gene gains, false chimeric gains and false exon gains in previous assemblies. Adjusted p value was calculated by g:SCS

Data No.	Species	Source	Term name	Term ID	Adjusted p value	Term size	Query size	Intersection size
1	Zebra finch	GO:BP	cellular component organization or biogenesis	GO:0071840	1.51E-11	6880	1125	551
2	Zebra finch	GO:BP	cellular component organization	GO:0016043	3.65E-11	6661	1125	535
3	Zebra finch	GO:MF	ATP binding	GO:0005524	1.89E-10	1500	1144	162
4	Zebra finch	GO:MF	adenyl ribonucleotide binding	GO:0032559	1.56E-09	1564	1144	164
5	Zebra finch	GO:MF	adenyl nucleotide binding	GO:0030554	3.07E-09	1577	1144	164
6	Zebra finch	GO:MF	anion binding	GO:0043168	3.96E-09	2413	1144	227
7	Zebra finch	GO:MF	protein binding	GO:0005515	9.89E-08	14767	1144	985
8	Zebra finch	GO:BP	localization	GO:0051179	1.09E-07	6938	1125	535
9	Zebra finch	GO:MF	nucleotide binding	GO:0000166	1.45E-07	2176	1144	203
10	Zebra finch	GO:MF	nucleoside phosphate binding	GO:1901265	1.51E-07	2177	1144	203
11	Zebra finch	GO:MF	small molecule binding	GO:0036094	1.0023E-06	2516	1144	224
12	Zebra finch	GO:MF	purine ribonucleoside triphosphate binding	GO:0035639	2.9305E-06	1846	1144	173
13	Zebra finch	GO:BP	cell projection organization	GO:0030030	4.2519E-06	1636	1125	161
14	Zebra finch	GO:MF	catalytic activity	GO:0003824	5.0178E-06	5682	1144	435
15	Zebra finch	GO:BP	cell morphogenesis	GO:0000902	5.1474E-06	1045	1125	114
16	Zebra finch	GO:BP	cellular localization	GO:0051641	6.3623E-06	3519	1125	297
17	Zebra finch	GO:BP	establishment of localization	GO:0051234	6.779E-06	5364	1125	422
18	Zebra finch	GO:BP	plasma membrane bounded cell projection organization	GO:0120036	7.2484E-06	1596	1125	157
19	Zebra finch	GO:MF	purine ribonucleotide binding	GO:0032555	9.5292E-06	1917	1144	176
20	Zebra finch	GO:MF	purine nucleotide binding	GO:0017076	1.004E-05	1932	1144	177

Data only appears up to No. 20 from Ko. et al. (2022). Intersected genes were omitted in this table.

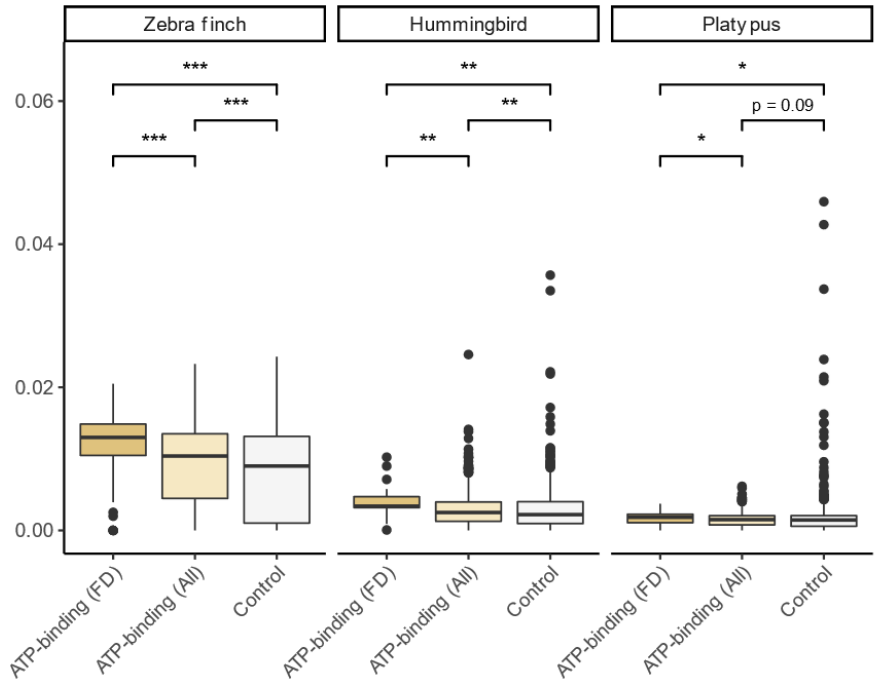

Figure 2. 14 Heterozygosity of ATP-binding genes with or without false duplications. 'Control' genes were randomly chosen for each species without ATP-binding genes. Box plots show median, first and third quartiles, range, and outliers as dots. One-sided Wilcoxon rank sum test was used to calculate significance between heterozygosity levels (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$; one-sided).

2.4.6 False duplication and annotation errors remaining in VGP assemblies

Although the amount of false duplications in the VGP assemblies was drastically lower than previous assemblies, here we found 74 to 119 scaffolds included false duplications, of which 5 to 34 (3–11% of the total number of scaffolds) were complete scaffold duplications (**Figure 2. 15**). From this error, we observed 1,175, 119, and 94 genes of the zebra finch, hummingbird and platypus were total or partial false duplications in the VGP v1.0 assemblies (**Figure 2. 8**). False duplications were observed within both named chromosomes and unplaced scaffolds, with no discernable patterns in terms of chromosome (**Figure 2. 16a, c, e**). However, for some small unplaced scaffolds (< 50 kbp) the proportion of their scaffolds as false duplications were large, with some cases where the entire scaffold was a false duplication (**Figure 2. 16b, d, f**). This indicates that for the VGP assemblies, some unplaced scaffolds are simply the other haplotype or a homotype duplication the length of a raw read (1 to 50 kbp) with sequence errors.

We manually verified examples, and found some of the same type of errors seen in the previous assemblies, except the duplications were larger, presumably due to the longer read lengths

and long optical maps of the VGP assemblies. An example was a false gene gain of *NPNT*, called *NPNT*−like (*LOC100218132*), on chromosome 4, named as such by the NCBI annotation pipeline applied to the VGP zebra finch 1.0 assembly (**Figure 2. 17**). However, the false duplication structure caused 4 missing exons in the 5' region of *NPNT* and 3 missing exons in the 3' region of *NPNT*−like. Characteristic of the previous assembly, the false duplications were separated by an assembly gap, with discordant 10X linked reads and at haploid depth coverage. Other examples included those that contained non−coding sequence (**Figure 2. 18a**), and those that contained false chimeric PacBio palindromic sequence the length or raw reads (7−17 kbp), both with 10X linked read depth gaps (**Figure 2. 18b, c**). A case of a large duplication was on zebra finch chromosome 29, where 4 segments adding up to ~1.9 Mbp total were classified as false duplications using our criterion, making up ~45% of the assembled 4.2 Mbp microchromosome (**Figure 2. 17b**).

To verify whether many of these false duplications are due to false haplotype separation, we examined a VGP trio based assembly of another zebra finch individual (Rhie et al., 2021). This trio based approach was recently developed with the goal of using parental short reads to separate out haplotype sequences of the child long reads,

before contig assembly and scaffolding (Koren et al., 2018; Rhie et al., 2021). We found that both the local *NPNT* (**Figure 2. 17a**) and the large ~1.9 Mbp of duplications of chromosome 29 were prevented in the trio-based assembly (**Figure 2. 17c**).

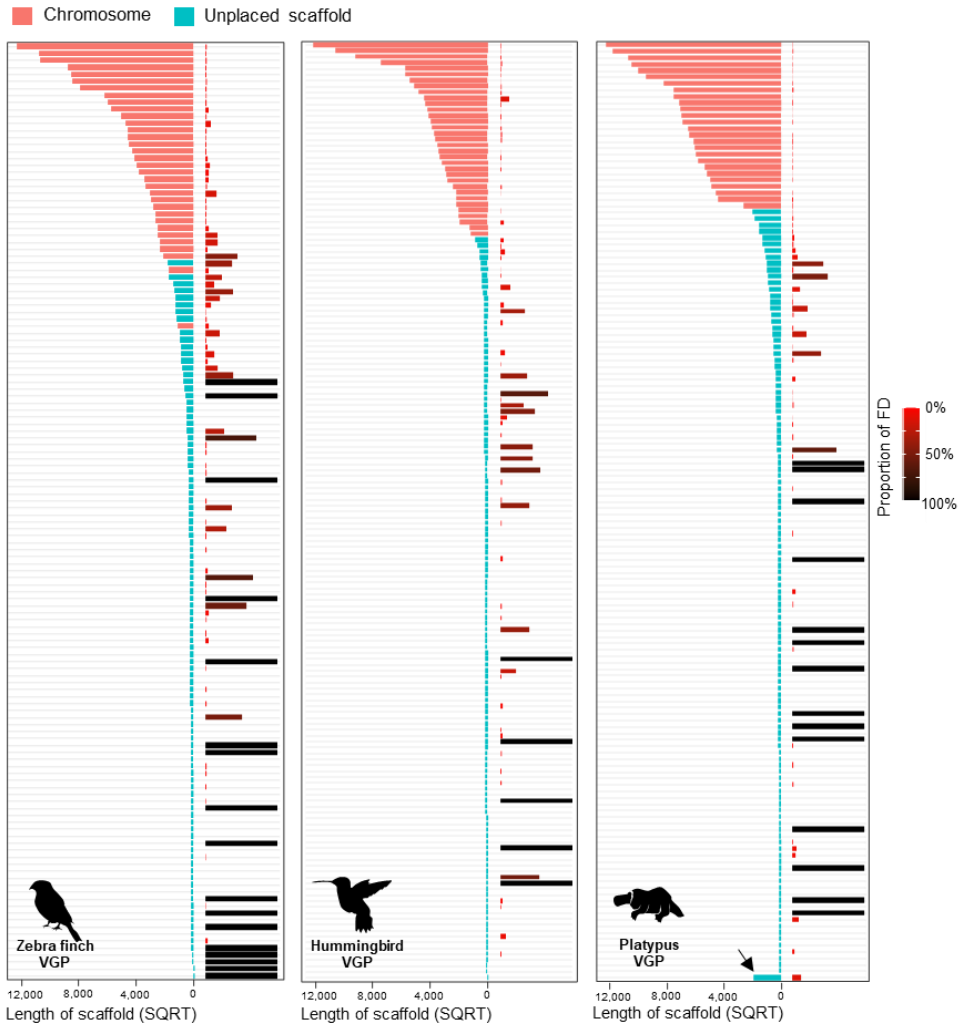

Figure 2. 15 False duplications left in VGP assemblies. The left side of each graph shows the scaffold length of named chromosomes (pink) and unplaced scaffolds (turquoise). The right side shows the proportion of each scaffold that is falsely duplicated either within the same or different scaffolds, with color intensity indicating 0% (in red) to 100% (in black) falsely duplicated. Arrow: for the platypus, scaffolds < 40 kbp were concatenated into the one scaffold, where

we found 20 scaffolds were completely duplicated among them.

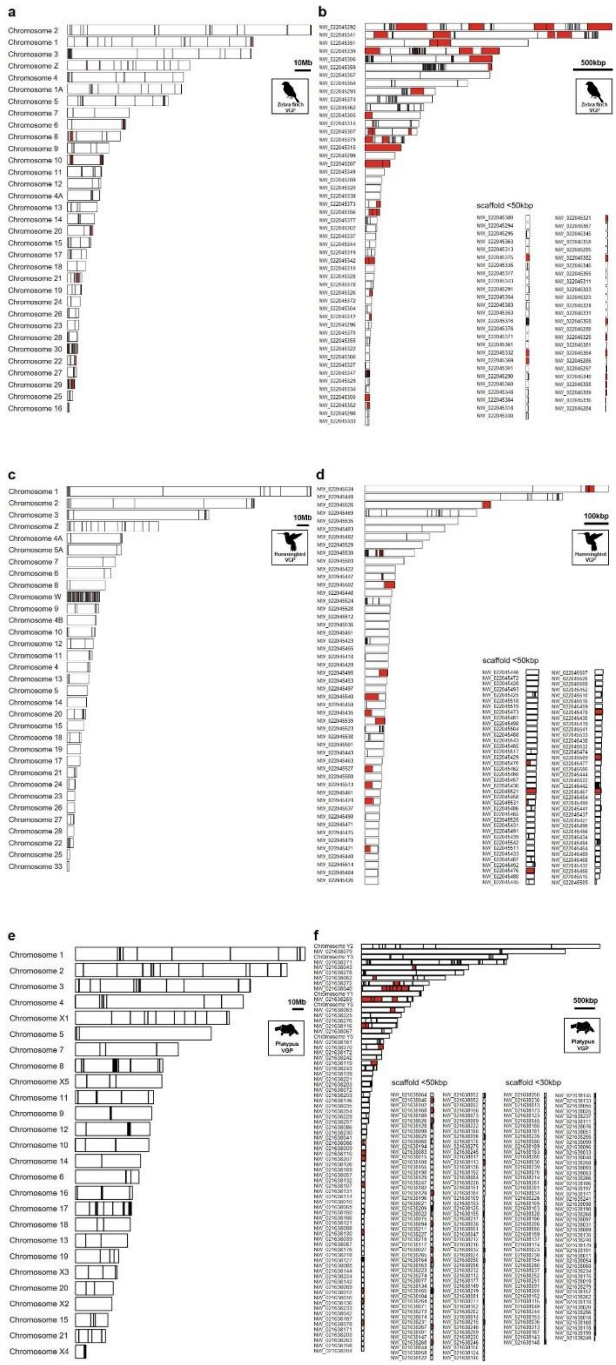

Figure 2. 16 Chromosomal location of false duplications in the VGP assemblies. False duplications are marked as small (black, <1kbp) or large (≥ 1kbp, red) bars, in each named chromosome (a, c, e) or

unplaced scaffold (b, c, d) for the zebra finch (a, b), hummingbird (c, d) and platypus (e, f).

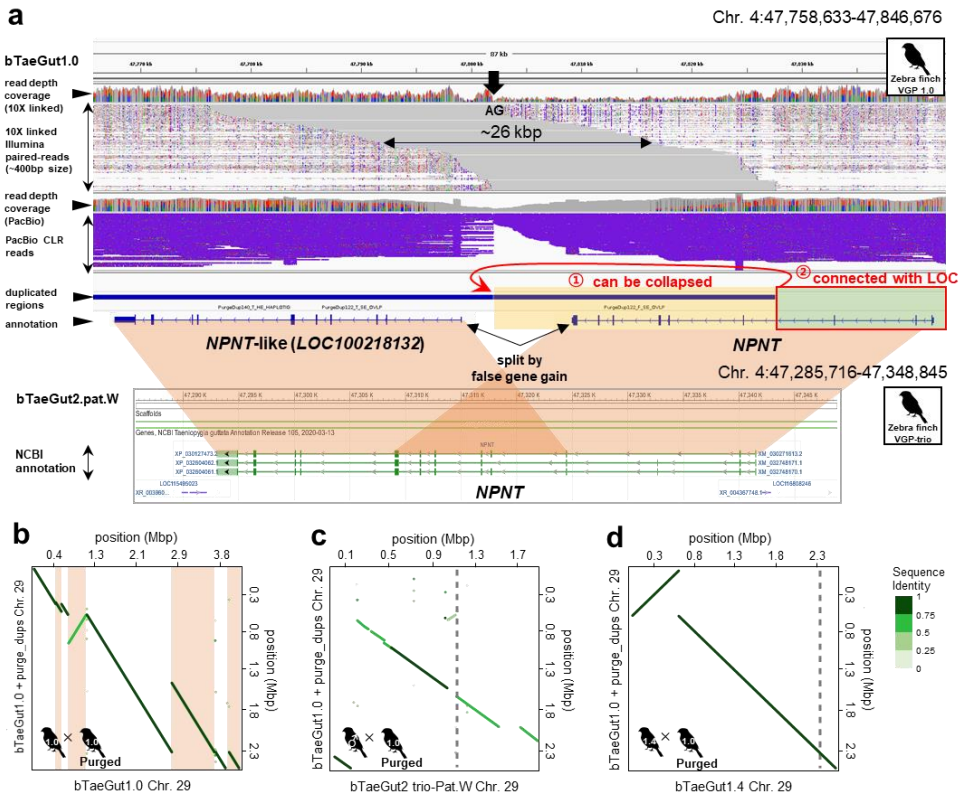

Figure 2. 17 False duplications and their correction in the VGP zebra finch assembly. a, The *NPNT* gene in the VGP zebra finch v1.0 assembly bTaeGut1.0 (first release) has the *NPNT*-like gene adjacent to it with an assembly gap (AG) and discordant 10X linked reads in this region. In contrast, the trio-based assembly (bTaeGut2.pat.W) had no *NPNT*-like gene, suggesting a false gene gain in bTaeGut1.0. The false duplication we found in this region was collapsed by purge_dups, and the falsely segmented gene structure was recovered. The VGP assembly v1.7 pipeline with purge_dups conducted before scaffolding prevented this false duplication (**Figure**

2. 19). b, Dot plot of alignment showing large ~1.9 Mbp false duplication of chromosome 29 (apricot) in the zebra finch VGP v1.0 pipeline assembly, bTaeGut1.0. c, The large ~1.9 Mbp of duplications of chromosome 29 in bTaeGut1.0 were prevented in the trio-based assembly. d, The 1.8 Mbp duplication was prevented with purging pre-scaffolding in bTaeGut1.4 using the VGP v1.7 pipeline. The boundaries of the scaffolds are represented as grey dashed lines.

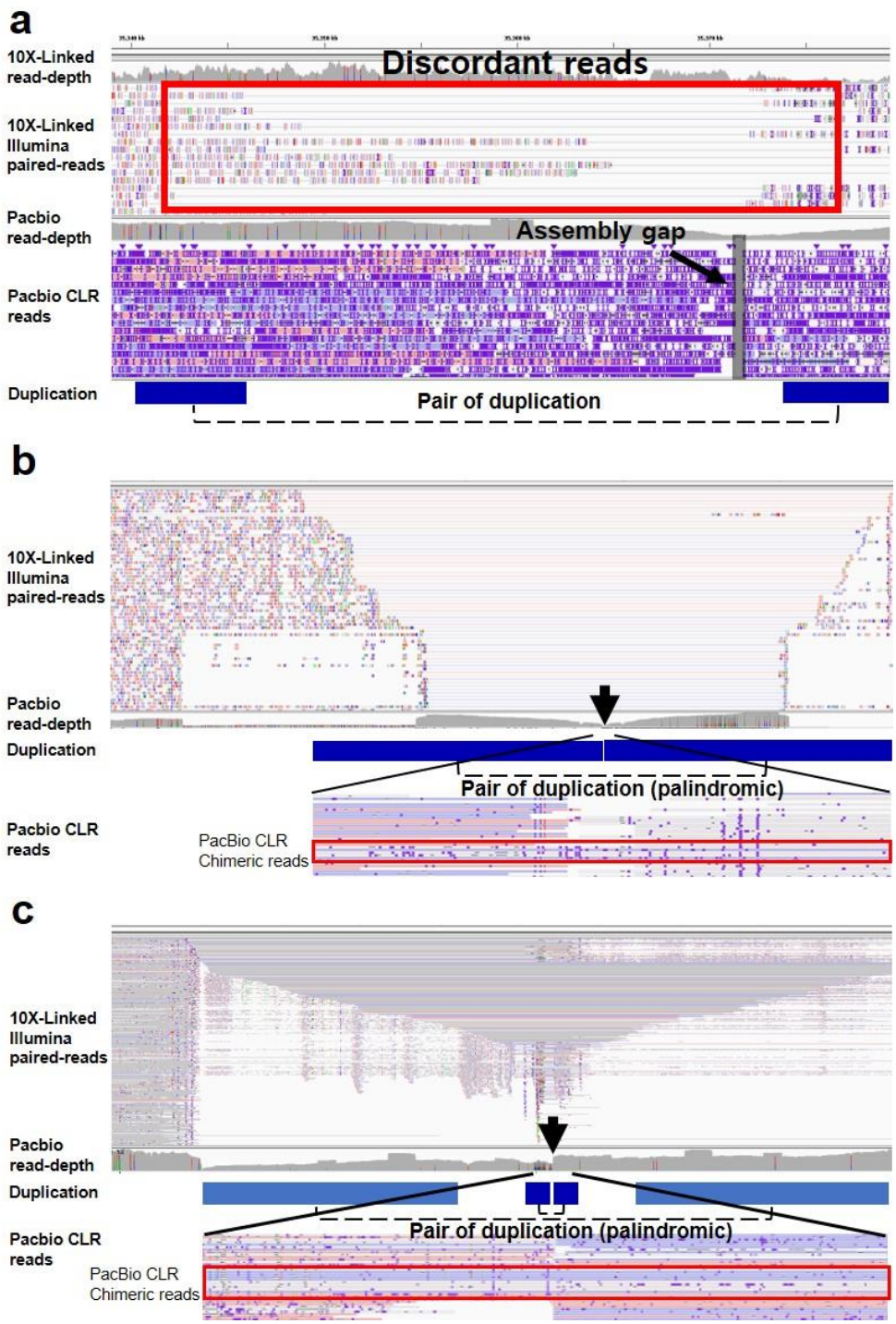

Figure 2. 18 Example cases of false duplications in the VGP assemblies.

a, A false duplicated region on zebra finch chromosome 6, ~7 kbp long, with an assembly gap and discordant 10X linked reads around the duplication (red box). 10X linked reads are shown as paired read alignments above the PacBio CLR read alignments, along with the depth coverage of the respective read data. b, False duplication in zebra finch scaffold NW_022045321 caused by a PacBio sequence read chimera. The ~10 kbp of palindromic sequence was duplicated without an assembly gap. But this region includes a sequence depth-gap with 10X linked reads (black arrow), signifying sequencing artifacts connecting the two regions. The symmetric reduction of insert sizes of 10X linked reads signifies the duplication of palindromic sequence. Near the center of the duplication, the six PacBio reads containing the chimeric sequences overlapped 10X depth-gap connecting the two duplicated sequences (red box). c, A duplicated region on chromosome 17 of the platypus, similar to the chimeric type in (b) except the PacBio read depth has a more pyramid structure in the palindromic duplicated regions.

We sought a means to further prevent false duplications in non-trio based assemblies, as the individuals used in this study do not have available parental data. The VGP assemblies used in this study were produced with the VGP v1.0 pipeline (Rhie et al., 2021), where heterotype duplications were removed by the purge_haplotigs algorithm (Roach et al., 2018) after scaffolding. In addition, many false duplications were detected and removed during manual curation (Rhie et al., 2021). As done for some later VGP assemblies of other species (Rhie et al., 2021), but not directly tested on the same individual, we reassembled the zebra finch individual used for the v1.0 assembly here, but performed purge_dups before scaffolding contigs in the VGP v1.6 pipeline rather than afterwards in the VGP v1.0 pipeline. We also added a new tool called Merfin, to polish the assembly with long reads (https://github.com/arangrhie/merfin), a step that does not influence false duplications but improves base level accuracy. We called the update the VGP v1.7 pipeline. After reassembly, the *NPNT* and other false duplications were prevented (**Figure 2. 19; Table 2. 7**). The 1.8 Mbp of false duplications on chromosome 29 were prevented, resulting in a smaller chromosome 29 consistent with removing false duplications manually (**Figure 2. 17d**). Overall, we observed a reduction from 1,175 genes to 176

genes with false duplications (**Figure 2. 8g**), and reduction of 16 entire false duplicated scaffolds to 5 (**Table 2. 7**). These findings show that false duplications are still prevalent in some of the best assemblies, but have potential to be removed with improved haplotype phasing.

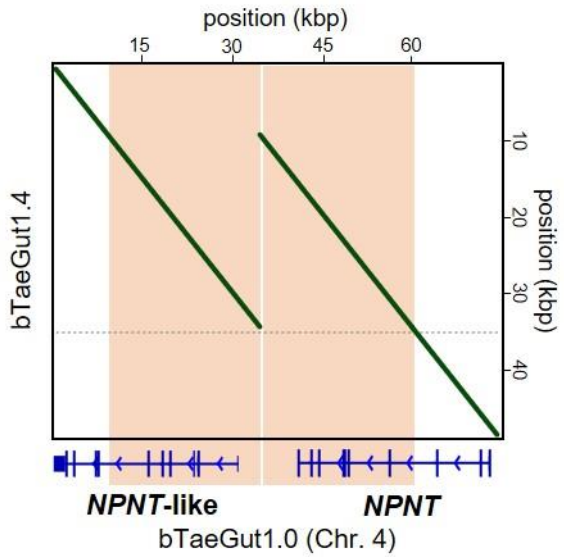

Figure 2. 19 Correction of the *NPNT* gene in VGP v1.7 pipeline assembly. Alignment dot-plot shows that the region with six duplicated exons of the *NPNT* gene in VGP v1.0 pipeline assembly (bTaeGut1.0) was prevented in the VGP v1.7 pipeline assembly (bTaeGut1.4). The blue bars represent exons of *NPNT* and *NPNT*-like genes in bTaeGut1.0.

Table 2. 7 Reduction of false duplications in the reassembled bTaeGut1.4 zebra finch genome with the VGP v1.7 pipeline. Amount of uncorrected false duplication in bTaeGut1.4 was calculated by ΣHv1.7 − (ΣHv1.0 − FD), where the H is the length of homologous sequence of the false duplication in each assembly (v1.7 and v1.0 pipelines). A negative value in column G represents the lack of homologous sequences than expected, after false duplication correction (ΣHv1.0 − FD).

Chromosome Name	Scaffold Name	Total length of Zebra finch VGP v1.0	FD lenth in Zebra finch VGP v1.0	FD length % of total length	*uncorrected FD*[a]	*Uncorrected FD* % in Zebra finch VGP v1.7
1	NC_044211.1	1.15E+08	728796	0.6%	125362	0.1%
1A	NC_044212.1	70430603	141131	0.2%	141537	0.2%
2	NC_044213.1	1.51E+08	763794	0.5%	164203	0.1%
3	NC_044214.1	1.13E+08	559728	0.5%	43356	0.0%
4	NC_044215.1	71552918	650200	0.9%	466	0.0%
4A	NC_044216.1	19824313	438158	2.2%	2371	0.0%
5	NC_044217.1	62005366	597833	1.0%	103403	0.2%
6	NC_044218.1	35665034	797224	2.2%	32206	0.1%
7	NC_044219.1	38060014	54270	0.1%	53335	0.1%
8	NC_044220.1	32610028	1742655	5.3%	71663	0.2%
9	NC_044221.1	25575665	111298	0.4%	-13846	-0.1%
10	NC_044222.1	22017011	1790758	8.1%	23743	0.1%
11	NC_044223.1	21012354	264285	1.3%	26823	0.1%
12	NC_044224.1	20435652	63029	0.3%	864	0.0%
13	NC_044225.1	18281482	304821	1.7%	5098	0.0%
14	NC_044226.1	16673467	420675	2.5%	1021	0.0%
15	NC_044227.1	14326562	581515	4.1%	1382	0.0%
16	NC_044228.1	1219933	49980	4.1%	8	0.0%
17	NC_044229.1	11687202	585838	5.0%	8200	0.1%
18	NC_044230.1	11044519	26319	0.2%	25595	0.2%

[a] ΣHv1.7 − (ΣHv1.0 − FD)

Data only appears up to 20 rows from Ko. et al. (2022).

2.4.7 Specific partitions of the genome with greater false duplications

Our above analyses focused on protein coding genes. Here we calculated the proportion of each genomic partition that was falsely duplicated. We found that in the previous and VGP assemblies, the intergenic regions had higher than expected false duplications based on the intergenic proportion of the genome, the introns lower than expected, and the exons no different than expected (**Figure 2. 20a and Figure 2. 21a,c**). An exception was the VGP zebra finch assembly, where the introns and exons were higher than expected (**Figure 2. 20a**). Among repetitive elements (LINEs, SINEs, LTRs, DNA, RNA, satellites,...), there were smaller differences relative to the expected proportions, with satellite repeats showing the highest above expected in some cases (**Figure 2. 20 and Figure 2. 21**). These findings are consistent with the common knowledge that intergenic regions diverge at a greater rate than genic regions, thus having higher heterozygosity. We were surprised to find that introns have a lower proportion relative to exons, given their higher divergence as well.

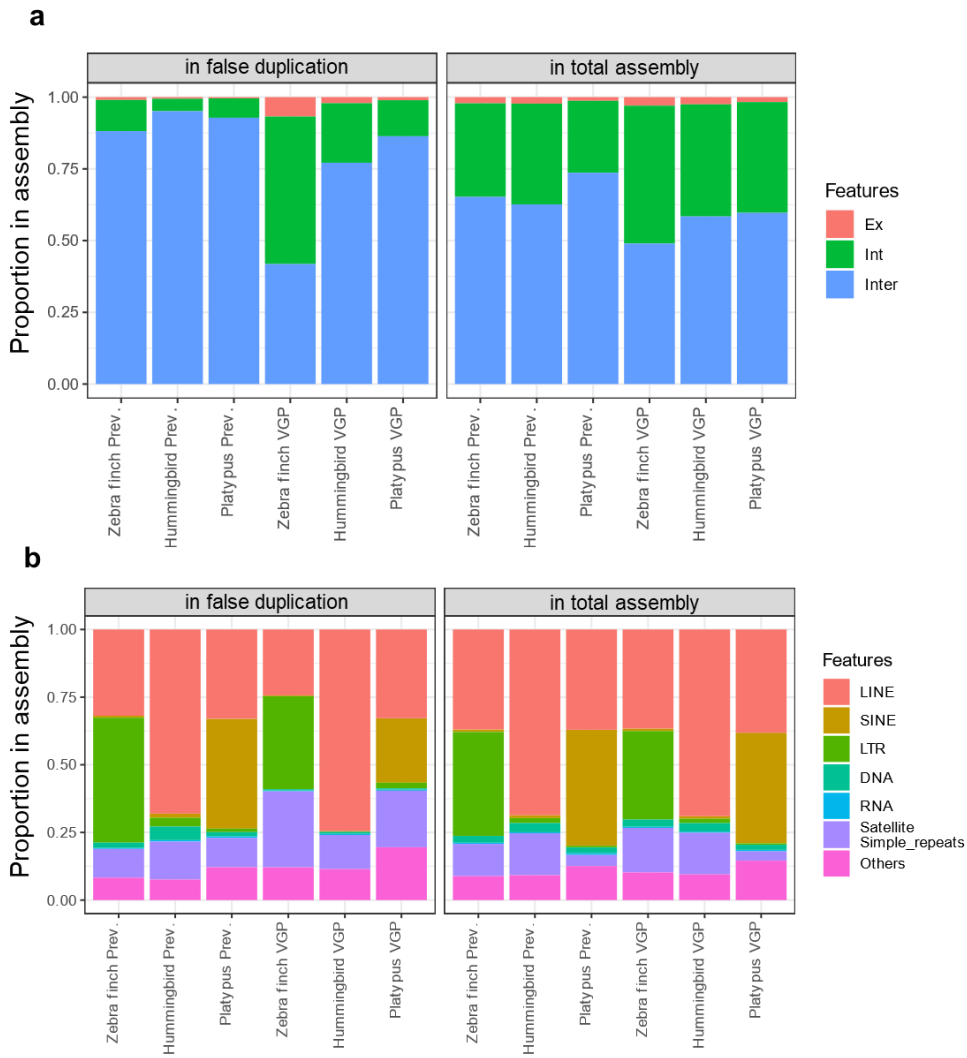

Figure 2. 20 Proportions of genomic partitions represented among the falsely duplicated regions. a, Proportion of false duplications among exon (Ex), intron (Int), intergenic (Inter) regions. b, Proportion of false duplications among different types of repetitive elements. Left panels show proportions among the false duplicated sequences; Right panels show proportions of all sequence types relative to the whole genome size.

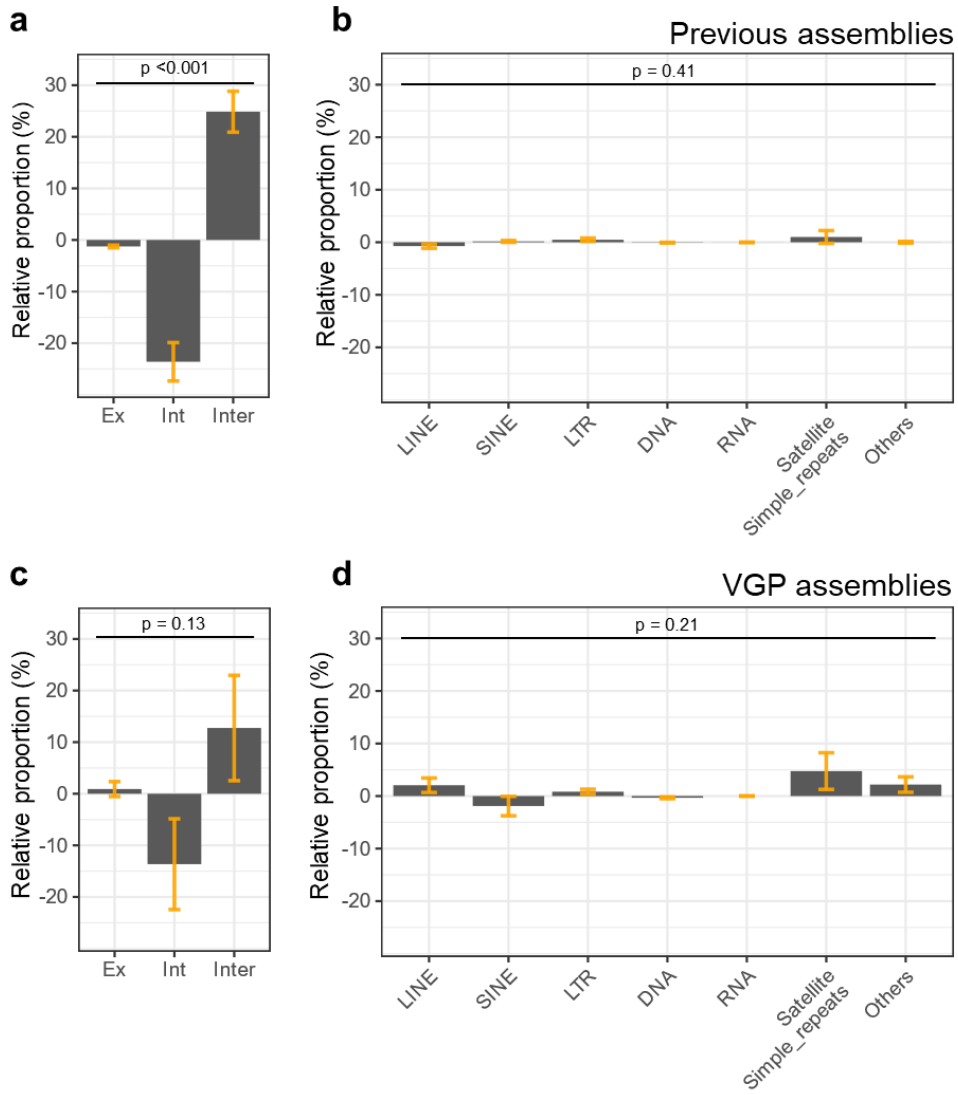

Figure 2. 21 Difference of proportion of each genomic partition containing false duplications relative to expected frequency. a and b, Relative proportions of false duplications in exon, intron, intergenic, and repeat regions in the previous assemblies. c and d, Relative values in the VGP assemblies. Error bars are standard error (n = 3 species each). The difference of relative proportion between

genomic partitions was tested by one-way analysis of variance (ANOVA).

2.4.8 Assembly methods to minimize false duplication

We further investigated the false duplication rates generated by different assembly algorithms and steps that went into testing and developing the VGP assemblies, using the hummingbird data (Rhie et al., 2021). We calculated the proportion of *k−mer* duplications as a proxy for false duplications. As expected, the fully scaffolded, haplotype purged assembly had the least *k−mer* duplications (0.6% of the genome; Table 2. 8). For the specific steps, PacBio CLR assembly with FALCON−Unzip showed < 0.7% *k−mer* duplication (Table 2. 8). FALCON alone on PacBio CLR reads resulted in more *k−mer* duplications (1.4%). The Canu contig assembler on CLR generated the most *k−mer* duplications (5.4%). A hybrid Canu assembly of PacBio and Oxford Nanopore reads showed better performance (2.1%) than Canu alone. This suggests that although both FALCON and Canu are diploid−aware assemblers, the haplotype resolving algorithm in FALCON−Unzip has a greater advantage in preventing false duplications. The Illumina short read assemblies generated with the 10X Genomics linked reads Supernova2.2 assembler and paired end reads with SOAPdenovo produced high *k−mer* duplications of 10.1% and 5.2%, respectively, even though both algorithms attempt to phase haplotypes (Luo et al., 2012; Weisenfeld

et al., 2017). In general, the scaffolding steps in the VGP pipeline with 10X linked reads, optical mapping, and Hi-C reads did not suppress false duplications, whereas purge_haplotigs and purge_dups were more effective to eliminate false duplications before scaffolding. False duplications of Bionano maps could also be introduced. But, most false duplications occur in the contigging step, which are then they are propagated in the scaffolding steps. Therefore, supporting our prior conclusions (Rhie et al., 2021), there is a need for the haplotype resolving early in the assembly process, in contigging step.

We also further investigated the presence of false duplications to the other recent assemblies produced originally outside of the VGP group (**Figure 2. 22**). We analyzed the emu, which included one assembly generated with Illumina short reads (ASM1339679v1) (Feng et al., 2020) and another generated with PacBio CLR reads, and scaffolded with 10X and Hi-C reads (ZJU1.0) (Liu et al., 2021). Surprisingly, we found more false duplications in recent long-read assembly made by FALCON-Unzip and purge_haplotigs (14Mbp; 1.1% of the assembly) than the short-read one (1Mbp; 0.1% of the assembly) made by AllPaths-LG (Gnerre et al., 2011). *K-mer* profiles of these assemblies show that the heterozygosity is not significantly different between the individuals (**Figure 2. 23**), and thus

this can not be the explanation for the differences in the assemblies. We know that purge_haplotigs only removes false duplications that are on different contigs, whereas purge_dups also removes false duplications within contigs/scaffolds. We are working with the developers of the emu assembly to clean up these false duplications potentially with purge_dups. Overall these findings show that a combination of assembly methods and level of heterozygosity are key factors contributing to and preventing false duplications.

Table 2. 8 Proportion of k-mer duplication measured for each assembly strategy. The assemblies of Anna's hummingbird produced for benchmarking in Rhie et al. (2021) were analyzed. Assembly method represents the type of sequencing platform used including single or mixed data. PacBio continuous long reads (CLR), 10X Genomics linked read (10X), Bionano optical mapping (Opt3), Arima-Hi-C (Hi-C) and Oxford Nanopore technology (ONT) were used for benchmarking the assemblies. Purge_haplotigs was run on the primary contigs for reducing false duplication. The k-mers duplication was calculated by 'false_duplications.sh' in Merqury (Rhie et al. 2020.

assembly method	assembler	coverage (x)*	# contigs	k-mer duplication (%)
CLR + 10X+ Opt3 + Hi-C (v1.0) + purge_haplotigs	FALCON-Unzip + Scaff10x2.0 + Solve3.2.1 + Salsa2.2	CLR (69x)	585	0.6
CLR	FALCON-Unzip	CLR (69x)	680	0.7
CLR	FALCON	CLR (69x)	2,091	1.4
CLR + ONT	Canu	CLR (69x); ONT (26x)	1,461	2.1
CLR	Canu	CLR (69x)	2,600	5.4
10X	Supernova2.2	10X (50x)	36,557	10.1
SR	Soap de novo	SR (155x)*	124,820	5.2

*SRR943143 ~ SRR943153 (Zhang et al. 2014)

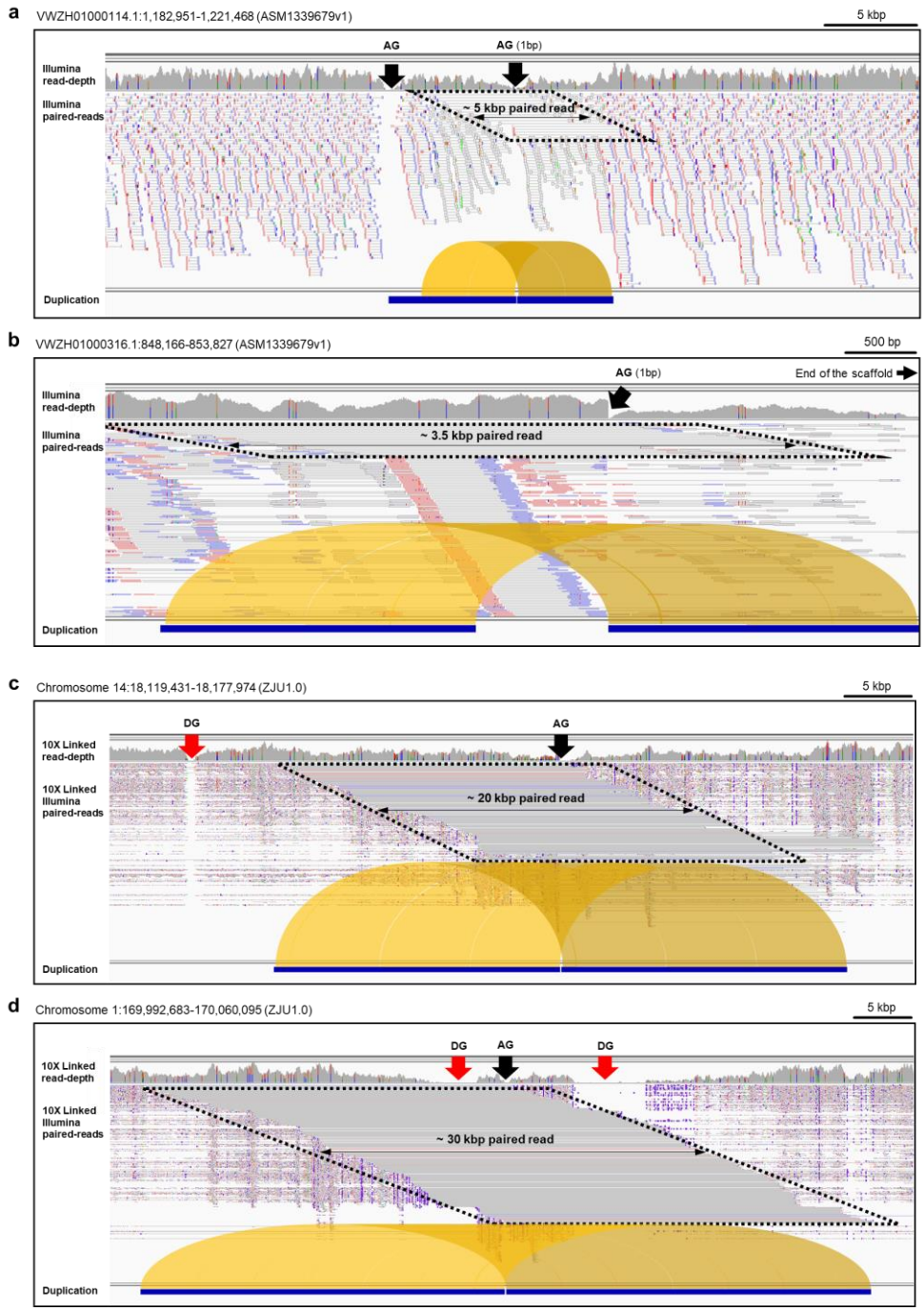

Figure 2. 22 The genome landscape of false duplications in emu assemblies. a and b, Duplicated region in the previous short-read based assembly. c and d, Duplicated region in the recent long-read

based assembly. 10X linked reads and Illumina reads are shown as paired read alignments above the duplication region, along with the depth coverage of the respective read data. AG, assembly gap. DG, depth−gap.

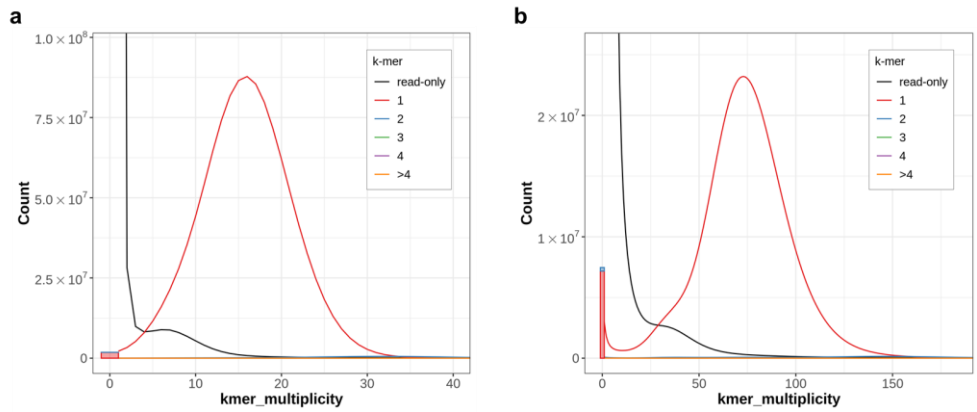

Figure 2. 23 *K−mer* profiling for emu assemblies. a, Previous short−read based assembly. b, Recent long−read based assembly. From the sequences of reads (Illumina shotgun for previous, 10X linked read for recent) and assemblies, *k−mer* multiplicity was calculated. The x−axis is the *k−mer* multiplicity calculated from the reads, and the numbers in the box represent the *k−mer* multiplicity found in the primary pseudo−haplotype assembly. *K−mer* multiplicity of 2 copies or higher under the area of single copies (red) are overly represented as false duplications.

2.5 Discussion

In this study, we found that the primary characteristics of false duplications include: 1) half depth of read coverage for heterotype duplications, or very low depth for homotype duplications; 2) presence of gaps between duplicated pairs on the same scaffold; 3) discordant or spanned linked read pairs used for scaffolding, whenever 10X or other types of paired reads of a DNA fragment were used; and 4) 1 copy *k-mers* for heterotype duplications. Some of these characteristics have been reported in other studies prior to the VGP effort (Kelley and Salzberg, 2010; Rhie et al., 2021; Roach et al., 2018; Schneider et al., 2017), but not in a systematic manner of comparing previous and new assemblies that attempted to remove false duplications as reported here. The false duplications were highest in the previous Sanger-based assemblies and lowest in the VGP PacBio-based long-read assemblies that purged them before scaffolding or in a VGP trio PacBio-based long-read assembly that sorted haplotype reads before contig and scaffold generation. One major source of the false duplications was a near doubling in the level of heterozygosity in the false duplicated regions compared to the rest of the genome. Further, the species with the highest heterozygosity, the zebra finch, had the highest proportion of false duplications in the

previous and VGP assemblies. Another major source was sequencing error, in both previous and VGP assemblies.

These false duplications led to mis-annotations as false gene, exon, and chimeric gene gains. When the duplication is created, the inserted allelic sequence results in annotation of two similar genes or one original and several fragmented genes. These types of false gains were made in genes involved in important phenotypes, leading to misinterpretations in downstream analysis (Pryszcz and Gabaldón, 2016). For example, false gene gains reduce one-to-one orthologs, which are used in comparative genomics and phylogeny. When false gains occur in an expanded gene family of closely related genes, this leads to false positive cases of gene family expansions and gene duplications as we report here, others previously (Han et al., 2013), and in a companion study on the oxytocin family of receptors (Theofanopoulou et al., n.d., p.). For phylogeny, these duplications create false orthologs or indels in genes that weaken gene- and species-inferred relationships. This can be made worse with multiple false duplications of genes with closely related paralogs, such as the overestimated LTR expansion in the zebra finch (Consortium, 2004; Warren et al., 2010), and ATP-binding gene family across species. Our findings indicate that caution should be

taken when interpreting gene family expansion in assemblies generated without haplotype phasing and checking for false duplications.

Our findings that heterotype false duplications are much higher than homotype, indicates that proper haplotype separation is still a current problem in genome assembly, even when they have been greatly reduced in the VGP assemblies. The VGP 1.6 trio pipeline removes more heterotype false duplications (Rhie et al., 2021), but it requires parental sequence data to sort haplotypes, and parents will not be available for all individuals. Scanning regions around gaps with reads and *k−mer* profiling, and discordantly mapped Illumina linked short reads or disconnected PacBio long reads should be helpful in identifying false duplications in any assembly. However, the best way to prevent these we propose would be to improve haplotype phasing of raw reads without parental data, remove reads with sequence errors before assembly, and generate complete diploid genome assemblies.

The VGP group is constantly updating its sequencing and assembly pipeline to create a genuine blueprint for assembly of complex and large genomes as found among vertebrates. Doing so requires in depth evaluation of assemblies, as done in this study. In

the VGP assembly pipeline, the CLR data type of PacBio sequencing was recently replaced in 2021–2022 with the closed circular sequence (CCS) high fidelity (HiFi) read data type (Rhie et al., 2021; Wenger et al., 2019), which reduces the base–pair error rate without the need for short–read Illumina polishing. We expect these new HiFi reads to also reduce the false duplications due to sequence errors, and it may allow better separation of haplotypes; promising alternatives include recent assemblers that use Hi–C data to phase haplotypes before or during contig assembly, FALCON Phase (Kronenberg et al., 2021) and hifiasm (Hi–C) (Cheng et al., 2021). The HiFi sequence read lengths, however, are currently ~20% shorter (15–20 kbp) than CLR, and thus may lead to less contiguity across real duplications longer than the read lengths. Our findings emphasize that creating haplotype–phased reference genome assemblies free of false duplications should be a fundamental requirement of future genomics and biology.

This chapter will be submitted to *Nature Methods*

as a partial fulfillment of Byung June Ko's Ph.D program.

Chapter 3. Automated HiFi-Based Genome Assemblies Reveal Lower Assembly Errors than Current Long-Read-Based Assembly.

3.1 Abstract

Accessible, fully automated, and validated assembly pipelines can provide availability and efficiency for individual researchers or small laboratories in genome assembly. The Galaxy Project has implemented a web-based, open assembly pipeline based on PacBio High-Fidelity (HiFi) reads through collaboration with the Vertebrate Genome Project. However, the assembly methods in the pipeline have not been validated in terms of the extent to which HiFi reads can correct errors compared to PacBio CLR reads. As a part of collaboration with the projects, I quantified potential assembly errors using *k-mer* profiling and whole-genome alignment for assemblies generated from an individual zebra finch using CLR and HiFi reads. The *K** metric revealed that HiFi-based assemblies had fewer errors compared to CLR assemblies. Furthermore, HiFi-based assemblies exhibited fewer structural errors such as false duplications and losses compared to CLR assembly. Among the different assembly modes, the HiFi-Trio mode produced the most stable assembly with respect to false duplication errors, while the HiFi-HiC mode resulted in the lowest amount of false losses. I propose that the optimal method is the HiFi-Trio mode, and the

HiFi-HiC mode is also effective when trio data is unavailable in the VGP assembly pipeline of the Galaxy Project.

3.2 Introduction

To establish high-quality reference genome assemblies, a well-constructed automatic computational pipeline has been recognized as one of the most important foundations for recent mega-scale genome projects (Giani et al., 2020; Lewin et al., 2022; Nurk et al., 2022; Rhie et al., 2021). For optimization, a variety of sequencing technologies and assembly algorithms have been tested (Bradnam et al., 2013; Giani et al., 2020; Mc Cartney et al., 2022). Recently, a groundbreaking study guiding the future direction of genome assembly was published. (Giani et al., 2020; Lewin et al., 2022; Nurk et al., 2022; Rhie et al., 2021) developed an efficient and stable assembly pipeline for vertebrate species by benchmarking various short-read and long-read platforms and assembly algorithms together. The authors suggested that using PacBio continuous long reads (CLR) in diploid genome assembly offers numerous advantages over short reads, as it helps avoid both structural assembly errors and missing sequences. However, the base accuracy of this technology is much lower (<99%) compared to short reads (Carneiro et al., 2012; Giani et al., 2020). It incurs additional costs for producing short reads separately and requires time for polishing long-read sequences.

To overcome the drawbacks of CLR reads, PacBio High-Fidelity sequencing technology (HiFi) utilizing circular consensus sequencing (CCS) has been developed (Wenger et al., 2019). The base accuracy of HiFi reads is significantly higher (99.9%) than that of CLR reads, eliminating the need for additional short-read production. Consequently, recent genome projects have been transitioning from CLR to HiFi for assembly construction (Cheng et al., 2021; Nurk et al., 2022). The Galaxy Project (The Galaxy Community, 2022), a web-based platform for bioinformatic analysis, in collaboration with the Vertebrate Genome Project (VGP), aims to create a fully automatic and scalable genome assembly platform with an intuitive web-based interface for widespread use (https://galaxyproject.org/). Although a recent HiFi-based VGP assembly pipeline has been integrated into the Galaxy Project recently, there has been no systematic examination to determine whether this automated HiFi-based assembly has advantages over previous CLR assembly.

As part of the collaboration between the Galaxy Project and VGP, I investigated the extent of false duplications (Ko et al., 2022) and loss (Kim et al., 2022) errors between CLR assembly generated by the previous VGP pipeline and automated HiFi

assemblies generated by the recent VGP pipeline in the Galaxy Project. All HiFi assemblies were produced using Hifiasm (Cheng et al., 2021), but with different data and methods, as described below: 1) standard method of Hifiasm using HiFi reads exclusively, 2) Hi-C integration of Hifiasm using both HiFi reads and Hi-C reads (Jarvis et al., 2022), and 3) Trio binning (Koren et al., 2018) of Hifiasm using fully phased HiFi reads based on maternal and paternal *k-mers*. In this study, I propose the best strategy to mitigate structural assembly errors and provide alternatives when the optimal strategy is not feasible, based on the quantification of assembly errors using genome-wide *k-mer* profiling and whole-genome alignment.

3.3 Materials and Methods

3.3.1 Used data

For the comparison, I used the three primary assemblies (CLR, HiFi-only, and HiFi-HiC mode) and one rebinned paternal assembly (HiFi-Trio mode) of zebra finch 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/; with default engine and commands "-species 'Taeniopygia guttata' -xsmall -s -no_is -cutoff 255 -frag 20000" before genome alignment.

3.3.2 Read mapping and coverage calculation

The reads generated for the assemblies by Pacbio CLR, HiFi and 10X platforms were mapped to the all genome assemblies using Minimap2 (Li, 2018) and EMA mapper (Shajii et al., 2018). For PacBio CLR and HiFi read mapping with Minimap2, the parameters "-ax map-pb" and "-ax map-hifi" were used, respectively. The paired-end 10X reads with barcodes were mapped using default

options of EMA (Shajii et al., 2018), while the reads without barcodes were mapped using BWA (Li and Durbin, 2009) with parameters "-p -M -R '@RG\tID:rg1\tSM:sample1'", following the guidelines in EMA. Sambamba was used to merge the intermediate BAM files produced during the read mapping step, and Samtools (Li et al., 2009) was used for sorting the BAM files and calculating read coverages for each genomic position.

3.3.3 *K-mer* counting and *K** calculation

To calculate *k-mer* duplications for each assembly, I employed a script "false_duplication.sh" in Merqury (Rhie et al., 2020), using the optimal *k-mer* size of 20 for the zebra finch genome. *K-mer* collapse and expansion were calculated by Merfin (Formenti et al., 2022) with the same *k-mer* size. For the *k-mer* collapse and expansion calculations of diploid assembly, I included the alternate haploid and maternal zebra finch sequences generated with each primary and paternal assembly used in the comparison. For the optimal *K** calculation, I incorporated the "lookup_table" produced by GenomeScope2 (Ranallo-Benavidez et al., 2020) with the 10X reads. In this analysis, I also included the current reference genome of zebra

finch assembled from CLR reads (bTaeGut1.4; GCF_003957565.2). *K-mer* duplications and the completeness of default-mode and rebinned trio assemblies were estimated from both paternal and maternal assemblies without purging, using Merqury (Rhie et al., 2020) along with the 10X reads.

3.3.4 False duplication and loss identification

I identified false duplications and losses by performing whole-genome alignment with an estimation of the number of paralogs in alignment blocks. Firstly, I aligned the three primary assemblies (CLR, HiFi, and HiFi-HiC mode) and the paternal assembly (HiFi-Trio mode) of the zebra finch using the Cactus alignment tool (Paten et al., 2011). Then, I extracted homologous regions to a readable multiple alignment format using Hal (Hickey et al., 2013). Since all assemblies were derived from the same sample, the number of paralogs in each assembly within each alignment block should be the same, except in cases where false duplications or losses occurred on the homologs.

For each alignment block that displayed a discrepancy in the number of paralogs between the assemblies, I calculated the

likelihood of each paralog model (i.e., the number of paralogous sequences present in the alignment block) based on the 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. where x represents the sum of the mean depth of each homologous sequence in an alignment block from an assembly, and and are the parameters of the depth distribution estimated from each paralog model. To estimate the model parameters μ and σ, I calculated the mean and variance of the normal distribution from the depth coverages of genomic regions where there are no multi-copy k-*mers* for the model with zero paralogs. I then multiplied the mean by an integer for each model (e.g., multiplying the mean by 2 for the 1-paralog model). I assumed that the variance is the same for all models.

False duplications and losses of each assembly were identified when an assembly had more or fewer paralogs than the best model determined from the likelihood estimation in each alignment block. To reduce noise in the identified false duplications, I filtered out cases where false duplications occupied less than 50% of the contig length

and were far from the terminals of the contig (>20 kbp). Additionally, I filtered out false losses that were less than 1 kbp in length. To avoid including haplotype differences as false losses, I calculated $K*$ values (Formenti et al., 2022) for *k-mers* in the regions of candidate false losses after the mentioned noise filtering. I only considered candidates for false losses when they had collapsed *k-mers* ($K* > 0$) covering over 90% of the genomic sequences.

Furthermore, I estimated potential false gene gains and losses (Kim et al., 2022; Ko et al., 2022) based on the annotation data of bTaeGut2.trio (GCF_008822105.2) and the erroneous regions identified in each assembly. I aligned the bTaeGut2.trio assembly and other assemblies together using Cactus (Paten et al., 2011). Potential false gene gains or losses were identified when the false duplications and losses had homologous regions with any coding sequences (CDSs) of the bTaeGut2.trio annotation.

3.4 Results and Discussion

3.4.1 *K−mer* profiles of CLR and HiFi−based assemblies

I calculated the amount of expansion and collapse errors in the assemblies based on the *K** metric (Formenti et al., 2022) using *k−mers* from the reads and diploid assembly data. I found the CLR assembly had the highest amount of both expansion and collapse errors (11% and 10%, **Figure 3. 1b**), whereas HiFi−only mode and HiFi−HiC mode had the lowest amount of expansion (9%) and collapse (6%). Because the *K** metric does not consider false duplication caused by phasing error, I also estimated the amount of false duplication based on duplicated *k−mers* from the reads and haploid assembly data (**Figure 3. 1a**). This showed the highest amount of *k−mer* duplication was in the CLR assembly (1.87%), followed by the HiFi−only (1.60%) and HiFi−HiC (0.89%) assemblies, and the lowest amount in the HiFi−Trio assembly (0.73%). This result is consistent with the fact that the Trio method is the most reliable approach for preventing false duplications caused by phasing errors.

The VGP group found that there were phasing errors between the haplotype of paternal and maternal assembly made by HiFi−Trio

mode. I checked the manual rebinning they did for the trio assembly was effective. I calculated the *k−mer* duplications and genome completeness of the both haplotypes (**Figure 3. 1c**), and found that the *k−mer* duplications were higher before rebinning (0.9 and 4.4% from both paternal and maternal, respectively) than both paternal (0.8%) and maternal (3.7%) rebinned assemblies. Moreover, the rebinned paternal assembly showed 1.2% higher *k−mer* completeness than the assembly before rebinning, although a slight decrease of *k−mer* completeness (0.2%) was shown in the rebinned maternal assembly. Therefore, manual rebinning by assembly experts would be effective when there is a clear indication of phasing errors observed from automated triobinning.

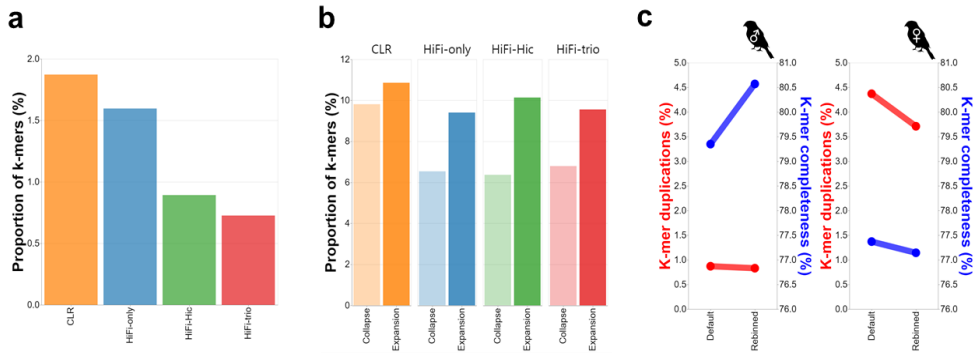

Figure 3. 1 *K−mer* evaluation of zebra finch assemblies made by PacBio CLR and HiFi reads. a, Proportion of *k−mer* duplication in the bTaeGut2 assemblies. *K−mer* duplications were calculated from the primary assemblies (CLR, HiFi−only, HiFi−HiC) and paternal assembly (Trio) from phased diploid assemblies. b. Proportion of *k−mer* expansion and collapse in each diploid bTaeGut2 assembly. c, Comparisons of *k−mer* duplication (red) and completeness (blue) between default and rebinned trio assemblies in males (left) and females (right).

3.4.2 The amount of structural assembly errors in CLR and HiFi-based assemblies

To further investigate structural assembly errors, I identified false duplication and loss in the assemblies based on the whole genome alignment of the bTaeGut2 assemblies followed by a likelihood calculation from read-depth coverages to determine whether the duplication and losses are true errors or statistical noise. From this approach, 1.3 to 13.8 Mbp of false duplication (**Figure 3. 2a**) and 0.2 to 42.3 Mbp of mean losses were found in the four assemblies (**Figure 3. 2b**). Notably, I observed that the assembly made by PacBio CLR reads is most prone to have both duplication (13.8 Mbp) and loss (42.3 Mbp) errors. The assembly with the lowest amount of false duplication was the Trio assembly, but the HiFi-Hic assembly had the lowest amount of false losses. These results showed concordances with the result of *k-mer* collapses and duplication (**Figure 3. 1a, b**). It has been known that one of the main sources of false duplication is sequencing errors, therefore, the lower accuracy of PacBio CLR reads compared to HiFi reads would be a cause of false duplication.

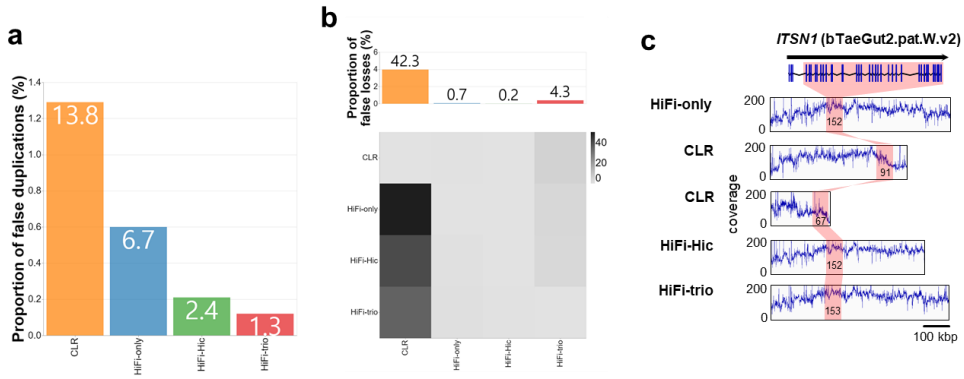

Figure 3. 2 Amount of false duplication and losses in zebra finch assemblies made by PacBio CLR and HiFi reads. a, Proportion and cumulated size (in Mb) of false duplications of each assembly. b, Proportion and cumulated size (in Mb) of false losses of each assembly (upper), and heat map of the size (in Mb) of false losses identified between the assemblies (below) in log scale. c, 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 grey) 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.

Importantly, among the false duplications, I found whole exon regions of 3 to 185 genes were affected (**Figure 3. 3a**). The CLR-based assembly was also more prone to false gene gains (185 genes), compared to the HiFi assemblies with 50, 3 and 9 genes for HiFi-only, HiFi-HiC and Trio assembly, respectively. For example, the *ITSN1* gene, which is strongly associated with autism-spectrum disorders (Feliciano et al., 2019), was found with high likelihood to be a false gene gain in the CLR assembly (**Figure 3. 2c**). Homologous sequences of 35 CDSs of total 38 CDSs of the *ITSN1* were found in two different contigs (000339F and 000509F) in the CLR assembly. Read coverage profiles of these duplicated regions showed the signature pattern for false duplications: haploid-level read-coverages were extensively observed in the duplicated regions. Similarly, I estimated 36 to 184 genes can be under false gene losses, and CLR-based assembly was more error-prone than HiFi based assemblies (**Figure 3. 3b**). Therefore, structural assembly errors from low accuracy long reads can propagate to the annotation level, resulting in significant misinterpretation.

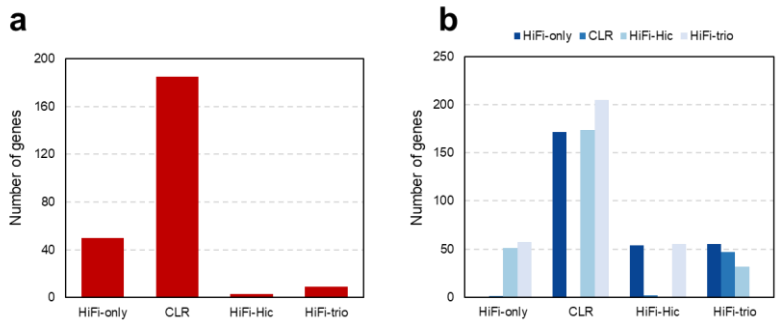

Figure 3. 3 Number of genes affected by false duplication (a) and losses (b).

3.4.3 Assembly errors in the current reference genome and the optimal strategy for assembly

Finally, I compared these zebra finch assemblies to the current reference for the zebra finch, which is based on a different individual (bTaeGut1.4; GCF_003957565.2), and found that more *k−mer* collapses (15%), but slightly less or more *k−mer* duplications (0.8%) in the bTaeGut1.4 compared to all four bTaeGut2 assemblies (**Figure 3. 4**). This may be, in part, because bTaeGut2 has 1.5−fold higher heterozygosity than bTaeGut1.4 (**Figure 3. 5**). Nevertheless, the comparison highlights that assemblies made by HiFi reads have the advantages to fewer false duplications and loss errors compared to the assembly made by CLR reads. Moreover, I found Trio data to be the best strategy to avoid false duplication when possible, although HiFi−HiC is also effective when trio data are not available (**Figure 3. 1**).

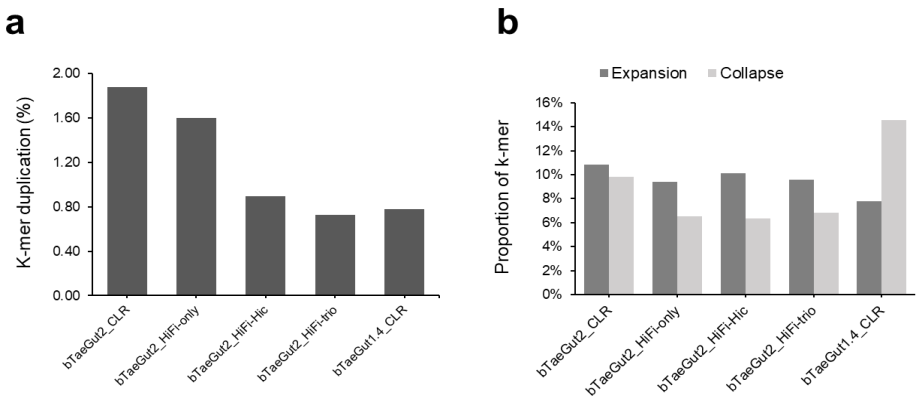

Figure 3. 4 *K−mer* evaluation between bTaeGut2 and bTaeGut1.4. a, Proportion of *k−mer* duplication in the bTaeGut2 assemblies and bTaeGut1.4 assembly. *K−mer* duplications were calculated from the primary assemblies (bTaeGut2 CLR, bTaeGut2 HiFi−only, bTaeGut2 HiFi−HiC and bTaeGut1.4 CLR) and paternal assembly (Trio) from phased diploid assemblies. b. Proportion of *k−mer* expansion and collapse in each diploid bTaeGut2 and bTaeGut1.4 assembly.

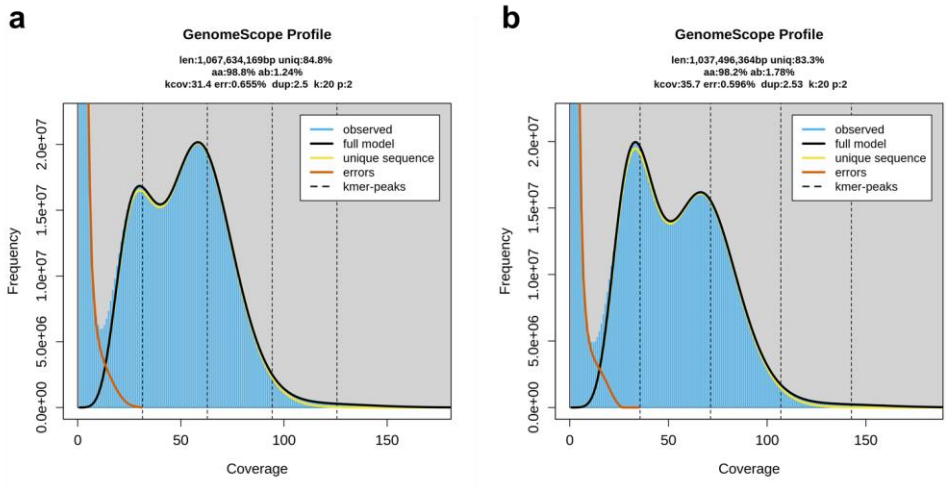

Figure 3. 5 Genome characteristics profile of zebra finch assemblies, bTaeGut2 (a) and bTaeGut1.4 (b) assemblies estimated from GenomeScope.

This chapter will be published elsewhere

as a partial fulfillment of Byung June Ko's Ph.D program.

Chapter 4. Purge mers: A New False Duplication Curation Tool Based on Sequencing Read and *K-mers* for Diploid Genome Assembly

4.1 Abstract

To achieve error-free genome assembly, it is essential to have a post-processing step for eliminating false duplications in diploid genome sequences. Existing tools for false duplication curation primarily rely on read coverage of genomic regions where sequencing reads are mapped to the assembly. In this study, I developed a new false duplication identification tool, Purge mers, which incorporates both read coverage and $K*$ metrics from *k-mer* multiplicity. Purge mers estimates model parameters for false duplication using both read coverage and $K*$ and employs maximum likelihood estimation to identify false duplications from whole or regional parts of a contig. I conducted simulations of two vertebrate species, zebra finch and human, using PacBio HiFi reads to evaluate the performance of Purge mers. The results showed that using short reads is more effective than using long reads for identifying false duplications. Purge mers exhibited better performance compared to Purge_dups in short read-based false duplication identification. Additionally, I propose a new $K*$ metric that offers advantages for model-based statistical inference.

4.2 Introduction

Error-free genome assembly is a most important task in recent large-scale genome projects (Lewin et al., 2022; Nurk et al., 2022; Rhie et al., 2021). Due to cost considerations, the genetic information of a species is fragmented into hundreds or thousands of base pairs during sequencing. However, the assembled genome sequence often contains significant structural errors, including redundant sequence duplications arising from allelic divergence or sequencing errors, known as false duplications (Guan et al., 2020; Kelley and Salzberg, 2010; Ko et al., 2022; Rhie et al., 2021).

False duplications resemble paralogs located in different loci from the original sequence but are actually alleles or artifacts by sequencing. These errors can lead to severe misinterpretation in comparative genomics, particularly regarding gene gain and gene family expansion (Ko et al., 2022; Korlach et al., 2017). False duplications frequently occur in genomes with high heterozygosity (Ko et al., 2022; Rhie et al., 2021). While one traditional approach to addressing this problem is to create highly inbred lineages (Koren et al., 2018), this is often not feasible. To overcome this issue computationally, post-processing software has been developed (Guan et al., 2020; Roach et al., 2018). These tools

identify false duplications based on sequence similarity and read coverage profiles of genomic regions.

Additionally, *k−mer*, which refers to a substring of a sequence with a specific length K, is commonly used to evaluate and quantify assembly quality (Rhie et al., 2020). It has been observed that the difference in *k−mer* copy numbers (or multiplicities) between sequencing reads and the genome assembly can be indicative of false duplications (Formenti et al., 2022). For instance, the multiplicity of a unique *k−mer* in a genome counted from sequencing reads should approximately match the mean read coverage of the genome. If the copy number of a *k−mer* in the genome sequences is unexpectedly higher than the multiplicity from sequencing reads, it suggests the presence of false duplication (Formenti et al., 2022; Phillippy et al., 2008).

While false duplication identification based on read coverage is widely used in current genome assembly (Cheng et al., 2021; Rhie et al., 2021), this method can be sensitive to the choice of mapping algorithm and repetitive regions. Long reads are preferable for avoiding disruptions in repetitive elements (Giani et al., 2020), but read coverage of small contigs shorter than the reads may not be accurately calculated due to the penalty of clipping. Merfin was

developed to quantify genome quality using the $K*$ metric at the base pair level (Formenti et al., 2022). However, there is currently no systematic software available for utilizing $K*$ to detect false duplications.

In this study, I developed a false duplication identification tool, purge mers, operating on both read coverage and $K*$ profile in genome-wide scale with base-pair level resolution. The tool estimates model parameters for false duplications and non-error regions using a bivariate Gaussian Mixture Model (GMM) based on both read coverage and $K*$. Subsequently, purge mers identifies false duplication from self-aligned homologs through maximum likelihood estimation with the models. To evaluate the performance of purge mers, I conducted simulations using two PacBio HiFi reads generated from vertebrate genome assemblies of a zebra finch and a human. I compared the results of purge mers with those of other existing false duplication identification tools. In the final part of the paper, I provide suggestions for the best strategy to identify false duplications based on the findings and results obtained from the study.

4.3 Materials and Methods

4.3.1 Generating simulation data

To generate simulated assembly and sequencing reads, the fully phased haplotype sequences of zebra finch (bTaeGut2) and human (mHomSap) made by Vertebrate Genome Project (Rhie et al., 2021) were collected from GenomeArk (https://genomeark.github.io/) and NCBI assembly (https://www.ncbi.nlm.nih.gov/assembly) for both paternal (bTaeGut2_trio.rebinned.hap1.s2.fasta and mHomSap3.pat) and maternal (bTaeGut2_trio.rebinned.hap2.s2.fasta and mHomSap3.mat), respectively (Table 4. 1). ReSeq (Schmeing and Robinson, 2021) was used to generate simulations of 60X coverage of diploid Illumina paired-end reads. To model the short read accurately, ReSeq required read mapping to reference first. 10X-Linked reads of each species were collected from GenomeArk and mapped to paternal assembly of zebra finch, and maternal assembly of human by EMA (Shajii et al., 2018) mapper and BWA (Li and Durbin, 2009) with standard EMA pipeline. Each mapping for a species was multi-processed by Parallel (Tange, 2011). A final merged and sorted read mapping file (.bam) was produced by

Sambamba and Samtools (Li et al., 2009). A variants file (.vcf) containing haplotype differences called by BCFTools with parameter "–mv –Ov –V indels" was used together for modeling diploid reads in ReSeq (Schmeing and Robinson, 2021). For simulating 40X coverage PacBio HiFi reads of diploid genome, Pbsim3 (Ono et al., 2022) was used for generating PacBio CLR (continuous long read) reads by multi–pass sequencing with command "–strategy wgs ––method qshmm ––qshmm QSHMM–RSII.model –pass–num 10". Then, CCS software was used to merge CLR reads to HiFi (Ono et al., 2022) for both paternal and maternal haplotype of the genomes, with a filtering threshold of "––min–rq 0.995". The intermediate output from Pbsim3 was converted and merged using Samtools.

Table 4. 1 Statistics of original zebra finch and human assemblies in this study.

Species	Haplotype	Length (bp)	No. Scaffold	Scaffold N50	Path to Access
Zebra finch	Maternal	1,031,735,314	279	71,356,113	https://genomeark.s3.amazonaws.com/index.html?prefix=species/Taeniopygia_guttata/bTaeGut2/assembly_vgp_trio_2.0/bTaeGut2_trio.rebinned.hap2.s2.fasta.gz
Zebra finch	Paternal	1,101,737,721	360	71,086,444	https://genomeark.s3.amazonaws.com/index.html?prefix=species/Taeniopygia_guttata/bTaeGut2/assembly_vgp_trio_2.0/bTaeGut2_trio.rebinned.hap1.s2.fasta.gz
Human	Maternal	2,902,214,236	950	150,534,096	https://www.ncbi.nlm.nih.gov/assembly/GCA_016695395.2
Human	Paternal	2,744,463,177	994	142,230,338	https://www.ncbi.nlm.nih.gov/assembly/GCA_016700455.2

The simulated Pacbio HiFi reads were used for generating assemblies of each zebra finch and human genomes using Hifiasm (Cheng et al., 2021). The "--primary -l2" parameters were used to obtain primary assembly. The Illumina paired-end read and PacBio HiFi read generated by simulation were mapped to the primary assembly from the simulation. BWA (Li and Durbin, 2009) was used for short read mapping with command "-M -R \"@RG\\tID:rg1\\tSM:sample1\" ", and Minimap2 (Li, 2018) was used for HiFi read mapping with a preset "map-hifi" with CIGAR string printing option "-c" for PAF format, and "-a" for generating .sam format.

Meryl software (Rhie et al., 2020) was used to count the *k-mers* from the short reads and long reads generated during the simulation process. The *k-mers* were also counted in the assemblies, and organized along the genomic positions for each zebra finch and human. The revised $K*$ in this study was calculated for each genomic position as:

$K*_{new} = (K_R - K_c) / \min(K_r, K_C) * \text{diploid peak}$

K_R = *k-mer* count from the reads

K_c = expected *k-mer* count of the reads estimated from the *k-mer* copy number in assembly (K_C in below).

K_r = expected copy number of the *k−mer* estimated from the read

K_C = observed *k−mer* copy number in the assembly.

diploid peak = mode of diploid *k−mer* model estimated by Genomescope2.0

K_r (Formenti et al., 2022) was probabilistically binned K_R by Genomescope2.0 (Ranallo−Benavidez et al., 2020) with "−−fitted_hist" option for K_r <5. The others of K_r>=5 were binned by rounding the K_R/diploid peak as same as Merfin (Formenti et al., 2022). The histogram of K_R was calculated by Merqury (Rhie et al., 2020), and set as input of Genomscope2.0.

4.3.2 Parameter estimation

In order to estimate the parameters of the models for false duplication and non−error, the following steps were performed for each dataset (zebra finch and human) and sequencing type (short−read and long−read). First, mean depth coverage and mean revised $K*$ of 250bp non−overlapping windows were calculated from contigs. For the short−read based analysis, contigs ranging to 2−300 kbp in length were considered, and for the long−read based

analysis, contigs ranging from 20 kbp to 500 kbp in length were included. Second, By the observation of the windows, the mean and covariance of read coverage and $K*$ were estimated using a bivariate GMM fitting with the expectation−maximization (EM) algorithm. The Mclust (Scrucca et al., 2016) package was used for GMM fitting with G=3 (false duplication, non−error and the other including noise and repeats). Mclust is also used to fit univariate GMM models for $K*$ alone with G=4. This univariate model was used for the candidate of false duplication composed by haploid specific sequences. In this case, an additional model for false duplication distributed on $K*=-2$ is considered.

4.3.3 False duplication identification

To identify false duplication, candidates were searched for using self−alignment of assembly with Minimap2 (Li, 2018) and "DP −cx asm5" for PAF formatting, which includes CIGAR string information of alignment. Many targets in alignments for each query were resolved as one−to−one best hit based on maximizing DP score (MS), chaining score (S1), alignment score (AS), and minimizing the total number of mismatches (NM) sequentially. In

case there was an overlap between the queries, the region of overlap was eliminated from the query that had lower scores compared to the other query. If the alignments obtained through self-alignment and the subsequent one-to-one resolution process covered more than 80% of the genomic region within a contig, the corresponding candidate for false duplication was categorized as a "haplotig" (indicating a potential haplotype duplication). Alternatively, if there were any queries located within a distance of 10 kbp from the contig terminals, alignments were chained using whole alignment information to find representative chains in a contig. Purge mers builds a direct-acyclic-graph (DAG) and finds a local optimal path to search the chains, similar to how purge_dups operates with 10kbp gap size threshold and filtration cutoff of 5,000 match score in DAG building (Guan et al., 2020). The alignment chains discovered through this process within a contig were categorized as "OVLP".

To identify false duplications from the candidates classified as both Haplotig and OVLP, Purge mers calculated the depth coverage and $K*$ of each non-overlapping window of 250 bp within the regions. Then, log-likelihoods of two models were calculated from the smallest to the largest contig. If the likelihood of the false

duplication model was higher than that of the non-error model, the candidate region was identified as a false duplication. For regions where more than 50% were comprised of haploid *k-mers*, univariate *K** models were used for maximum likelihood estimation. Regions where more than 50% of *k-mers* were composed of $K_r > 4$ were considered as repetitive elements. If the number of negative *K** values was larger than the number of positive *K** values, the candidate was identified as a false duplication. One notable algorithm included in Purge mers to reduce false positives is the recursive adjustment of depth coverage and *K** values for every identified false duplication candidate. When a false duplication is identified, it indicates that the region should be removed from the original assembly, and the depth coverage and *k-mer* count of the remaining homologs should be updated accordingly. Purge mers recursively recalculates the read coverages and *K** values of the homologs affected by the identified false duplication using the modified KC values in each maximum likelihood calculation.

Purge mers also identified artifactual contigs in cases where a contig exhibited erroneous *k-mers* exceeding 10% or a mean depth coverage below 0.25 times the haploid depth. To identify regional sequencing errors within contigs exhibiting erroneous *k-*

mers (5–10% of erroneous *k–mers* in a contig) meet these threshold the following steps were performed: 1) Merging regions with erroneous *k–mers* within a 50 bp distance. 2) Filtering the merged regions based on a size threshold of 200 bp, discarding regions smaller than 200 bp. 3) Further merging the remaining regions in 5 kbp distance. 4) Filtering the merged regions based on a size threshold of 10 kbp. Bedtools (Quinlan and Hall, 2010) was extensively used for handling genomic regions.

4.3.4 Performance assessment

To evaluate the performance by the tools and sequencing platform, I generated simulation data and compared the original assembly with the simulated one. Determining true false duplications in a simulation assembly is theoretically possible. However, it is challenging due to the complexity of identifying continuous one–to–one homologs through whole–genome alignment. If there is false duplication in repetitive regions, it is also hard to find true false duplication at the whole genome level by alignment.

Instead of using alignment-based methods to determine the true set, I employed a different approach that analyzes the *k-mer* copy numbers (K_C) in both the original and simulated assemblies. I quantified the difference in K_C (delta K_C; K_C in simulation – max[K_C in haploid1, K_C in haploid2] in original assembly) for each *k-mer* observed in the simulated assembly. If the delta K_C is greater than 0, it indicates the presence of false duplication for that *k-mer*. It's important to note that a *k-mer* copy number is calculated based on every genomic position that contains the *k-mer*. Consequently, it is not possible to pinpoint the exact region where the false duplication occurred in simulation (i.e. true false duplication). However, this limitation does not hinder the evaluation of performance at the genome-wide level because every *k-mers* are counted from the whole genome.

In this study, classification performance of the methods was evaluated by examining true positives, true negatives, false positives, and false negatives for each *k-mer* observed in the simulation assembly. The evaluation process is defined as follows:

1) True positive: min(*FCN*, delta K_C)

2) True negative: K_C – (True positive + False positive + False negative)

3) False positive: when delta K_C < FCN, FCN − delta K_C

4) False negative: when delta K_C > FCN, delta K_C − FCN

where K_C is the copy number of a $k-mer$ in simulation assembly, FCN is the copy number of a $k-mer$ identified as false duplication. In this evaluation process, the study does not take into account the effect of false loss, and a minimum delta K_C value of 0 is fixed.

4.4 Results

4.4.1 Model parameter estimation

Purge mers employs a model-based approach to identify false duplications by calculating base pair-level read coverage and $K*$ (Formenti et al., 2022). This structural error will exhibit specific erroneous signatures such as a haploid state of depth coverage (1X) and -1 of $K*$ on the primary assembly (**Figure 4. 1**). On the other hand, paralogous regions, which do not involve allelic differences, display a diploid state of depth coverage (2X) and no evidence of duplication or expansion in $K*$ ($K* = 0$). Although the original $K*$ metric effectively represents false duplication (or expansion with diploid sequences) and losses at the base-pair level, the current $K*$ collapses the variance of *k-mer* copy number from sequencing reads. Consequently, it is not suitable for the Gaussian Mixture Model (GMM). Therefore, I propose a modified version of the original $K*$ equation, which preserves the variance of *k-mer* copy numbers while still allowing estimation of false duplications ($K* = -1$) and non-errors ($K* = 0$; **SEE METHODS**).

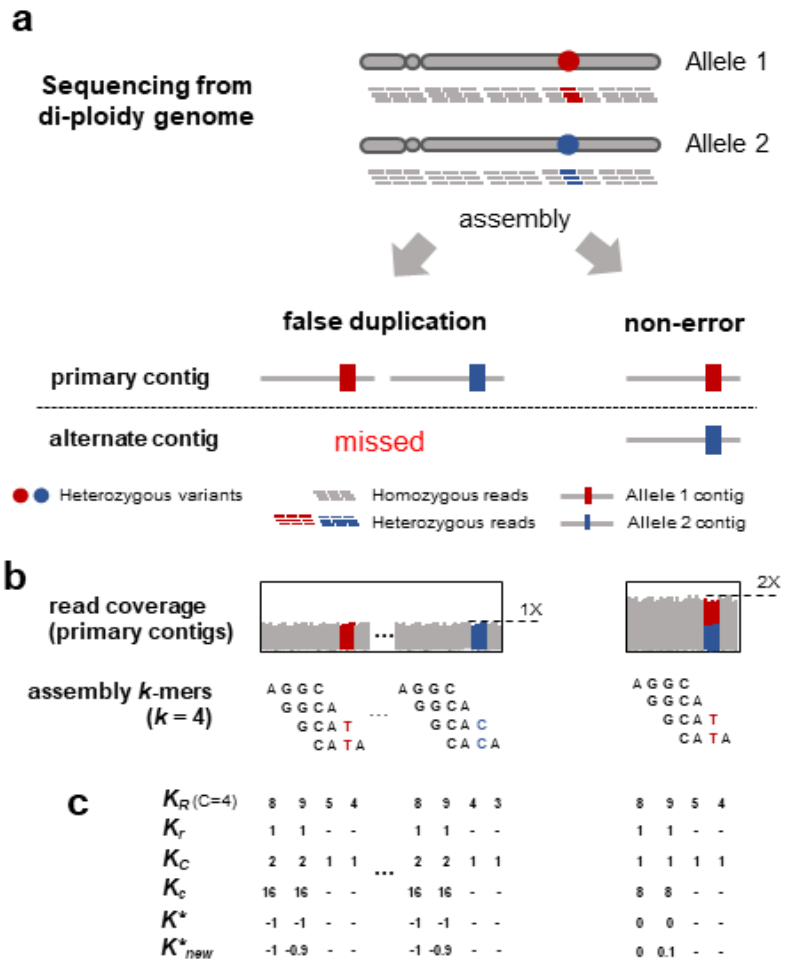

Figure 4. 1 Overview of identifying false duplication on both read coverage and *K**. a, False duplication occurred on heterozygous position. Two allelic sequences in diploid species can be located in a pseudo-haplotype assembly (false duplication). Proper phasing algorithm inhibit this, allocate them to each primary and alternate assembly (non-error). b, read coverage characteristics on false duplication and non-error regions. Homologs with false duplication typically indicates haploid-level depth coverage by heterozygous

read mapping. c, Various *k−mer* statistics on false duplication and non−error. K_R is *k−mer* multiplicities counted from sequencing reads. K_r is expected copy number of the *k−mer* in genome assembly estimated from the reads. K_C is observed *k−mer* copy number in the assembly. $K*$ is $(K_r - K_C) / \min(K_r, K_C)$. $K*_{New}$ is $(K_R - K_c) / \min(K_r, K_C) * \text{diploid peak}$, where K_c is expected multiplicity of the *k−mer* in reads estimated from the assembly, and diploid peak is a mode of diploid *k−mer* distribution.

The first step in the operation of purge mers is to estimate the parameters for both the false duplication and non-error models using read depth and *K**. Purge mers employs a bivariate Gaussian mixture model to estimate the mean and covariance of both models. For parameter estimation, purge mers samples genomic regions by non-overlapping 250bp windows within a certain size of contigs (SEE METHODS). The mean depth coverage and *K** values are calculated for each window. By constructing a bivariate distribution using the collected samples, the parameters for both the false duplication and non-error models can be estimated using the Gaussian mixture model. Purge mers fits the Gaussian mixture models using the Expectation-Maximization (EM) algorithm based on the distributions obtained from window sampling. Given the estimated parameters for each bivariate Gaussian distribution of both models, purge mers identifies false duplications through maximum likelihood estimation from candidate duplications by self-alignment.

4.4.2 False duplication candidate identification algorithm

After parameter estimation, purge mers proposes candidates of false duplication by sequence similarity between contigs by self-alignment, similar to purge_dups (Guan et al., 2020). Based on the composition pattern of the candidates of false duplication within a contig, purge mers classifies them into two distinct categories: 1) haplotigs and 2) overlaps (SEE METHOD). Each non-overlapping window-binned genomic region (250bp) within the haplotigs or overlaps serves as an observation unit for Maximum Likelihood Estimation (MLE) of the bivariate Gaussian model parameters for mean read coverage and $K*$ when the region is primarily composed of diploid and non-repeat *k-mers* (i.e., $K_r < 5$). However, purge mers employs different strategies for identifying false duplications when the region is predominantly composed of haploid-specific sequences or repeats (i.e., $K_r >= 5$). In addition, purge mers considers unaligned artifactual contigs as candidates for false duplications. During the assembly process, sequencing reads may accumulate base calling errors, leading to artifacts at the contig level (Ko et al., 2022). If a whole contig or a specific region contains a high number of erroneous *k-mers* beyond a certain threshold, or if they exhibit lower read coverage than a specified

threshold, the contig is identified as a false duplication caused by sequencing errors. In cases where the erroneous regions are clustered together by certain distance, the regions are merged.

4.4.3 Simulation statistics

To evaluate the performance of the proposed method, simulations were conducted using the original long-read-based assemblies bTaeGut2 and mHomSap3.mat, which were generated by the Vertebrate Genomes Project (VGP) group (**Table 4. 1**). The zebra finch assembly exhibited high heterozygosity (1.78%), while the human assembly had relatively low heterozygosity (0.32%) (**Figure 4. 2**). This allowed us to assess the performance of the method under both high and low heterozygosity conditions. Furthermore, the original assemblies bTaeGut2 and mHomSap3 were created using the VGP trio pipeline, which enabled the generation of simulated reads representing both long reads and short reads, considering the diploid of the genome.

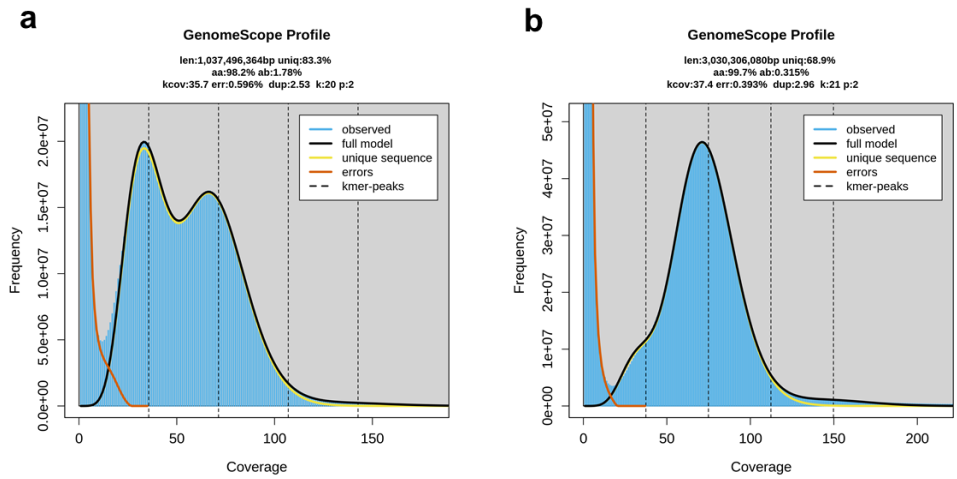

Figure 4. 2 Genome characteristics of zebra finch (a) and human (b) assemblies estimated by GenomeScope2. 10X-Linked reads produced in each species were used for estimation of genome characteristics.

During the simulation, 40X Pacbio HiFi reads were generated for both zebra finch and human diploid genomes. Initially, CLR reads generated by multi-pass sequencing (Ono et al., 2022) were merged to total 38.8 and 103.3 Gbp HiFi reads (4,305,410 and 11,475,896 sequences; Table 4. 1 and Table 4. 2) for each zebra finch and human. Additionally, 60X paired-end Illumina short reads were produced based on the assembly sequences. Total 66.3 and 174.6 Gbp reads (478,842,214 and 1,260,772,024 sequences) were generated for zebra finch and human, respectively. The 40X Pacbio HiFi reads for each species were used for diploid genome assembly. As a result, primary assemblies were obtained for zebra finch and human, consisting of 1.17 and 2.92 Gbp, 1,210 and 1,358 contigs, and possessing an N50 value of 15.8 and 38.4 Mbp, respectively (Table 4. 2). Heterozygous sequences separated from primary assembly were assigned to alternate contigs, resulting in 0.95 and 2.67 Gbp, 3,333 and 44,083 contigs, and an N50 value of 2.60 and 0.17 Mbp for zebra finch and human, respectively (Table 4. 2). The zebra finch assembly made by simulation showed high heterozygosity, but the simulation of human assembly represented relatively lower heterozygosity (Figure 4. 3).

Table 4. 2 Statistics of simulation data.

Species	Data	Total length (bp)	No. of sequences (reads or contigs)	Mean length of reads (bp)	Generated read coverage	N50
Zebra finch	PacBio HiFi	38,767,427,306	4,305,410	9004.4	40X	-
	Illumina	66,319,646,639	478,842,214	138.5	60X	-
	Primary assembly	1,165,444,248	1,210	-	-	15,771,177
	Alternate assembly	953,141,209	3,333	-	-	2,599,748
Human	PacBio HiFi	103,290,644,408	11,475,896	9000.7	40X	-
	Illumina	174,616,925,324	1,260,772,024	138.5	60X	-
	Primary assembly	2,917,922,356	1,358	-	-	38,390,416
	Alternate assembly	2,655,548,067	44,083	-	-	173,454

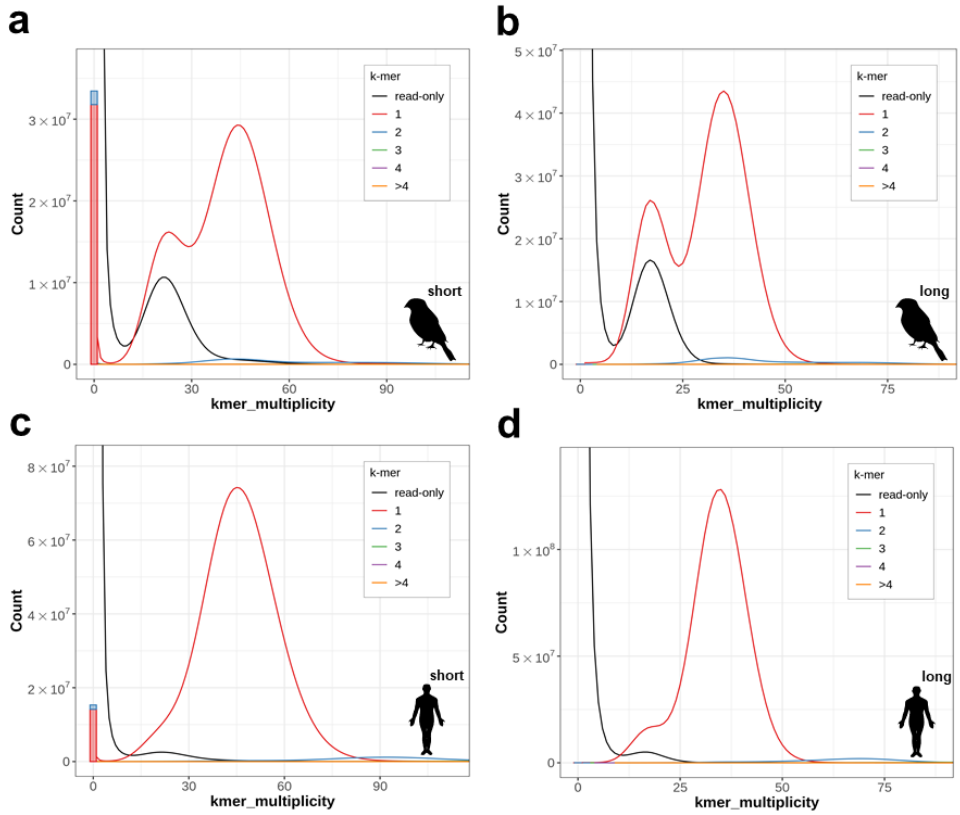

Figure 4. 3 *K-mer* profiles of simulated assemblies. Illumina reads (short; a and c) and PacBio HiFi (long; b and d) reads were used for drawing multiplicity distribution for zebra finch (a and b) and human (c and d) assemblies.

Both the long and short reads generated from the simulations were mapped to the primary assemblies of each species, and the base-level genome-wide read coverage and $K*$ were calculated. For zebra finch, the mean read coverage and $K*$ values of the false duplication model in the short-read simulations were 30.9 and -0.96 while the parameters of the non-error model were 60.3 and -0.04. In the long-read simulations, the estimated parameters for $K*$ exhibited little difference (-0.84 and 0.01 for false duplication and non-error model, respectively) compared to the short-read simulations. However, the parameters for read coverage were 22.3 and 40.9 by different data sizes of generated reads. Similarly, for human, the mean read coverage and $K*$ values of the false duplication model in the short-read simulations were estimated as 27.6 and -1.10, respectively, while the parameters of the non-error model were 59.6 and -0.02. In the long-read simulations, the parameters for read coverage and $K*$ were 19.0 and -0.71 for the false duplication model, and 37.0 and 0 for the non-error model, respectively.

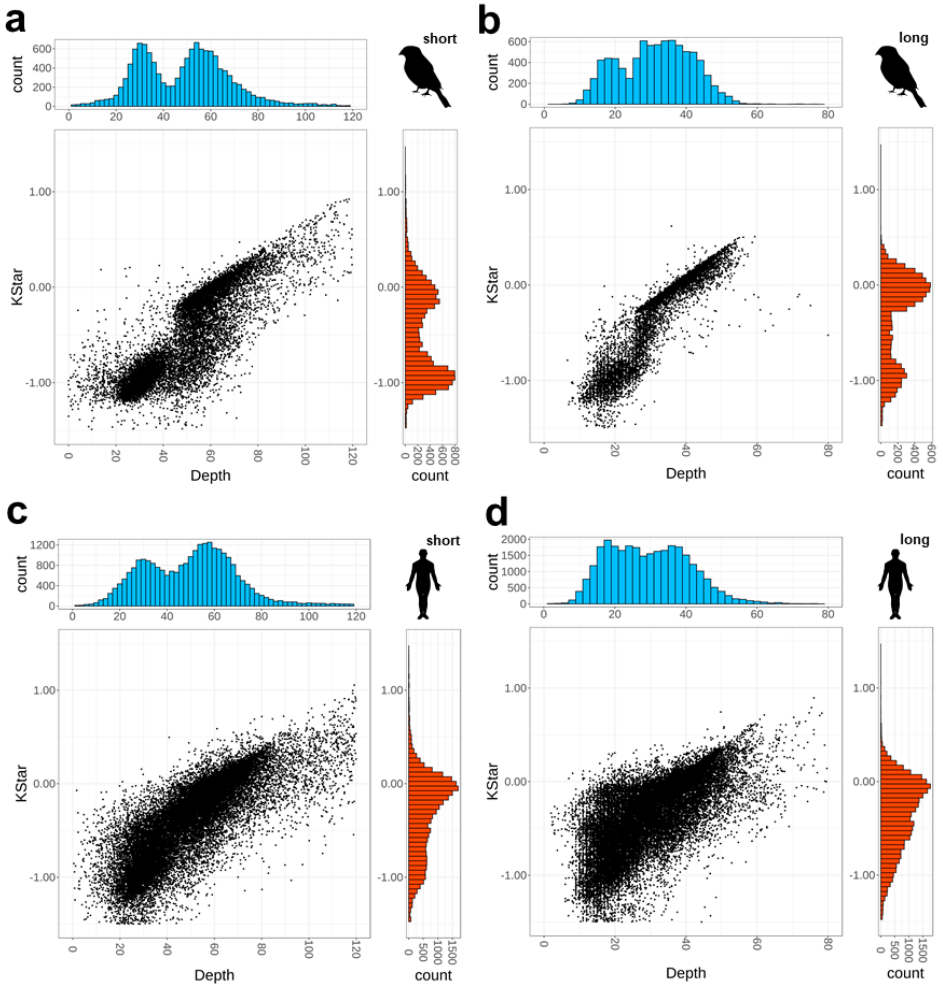

Figure 4. 4 Bivariate distributions of read coverage and *K of each zebra finch and human assembly.** Illumina reads (a and c) and PacBio HiFi reads (b and d) were mapped to zebra finch (a and b) and human (c and d) assemblies, respectively. Mean values of read coverage and *K** were calculated from non-overlapping 250bp window.

4.4.4 Performance assessment

To assess the performance of purge mers in identifying false duplications, I measured the precision, recall, accuracy and F1 score (harmonic mean of precision and recall combination) based on *k−mer* profiles between the template and simulation assembly. This evaluation was conducted separately using short read and long read data, and the results were compared with those obtained from purge_dups (Guan et al., 2020) for benchmarking purposes. Using short read data, total 79.7 and 36.7 Mbp of genomic regions (6.8% and 1.3% size of primary assembly) were identified as false duplication in zebra finch and human by purge mers (**Table 4. 3; Figure 4. 5a**). In comparison, purge_dups identified 93.2 and 81.8 Mbp of false duplication (8.0% and 2.8% of primary assembly size) in these species using short read. The benchmark results using short read data indicated that purge_dups outperformed purge mers for zebra finch. The precision, recall, accuracy and F1 scores for purge mers were 45%, 60%, 95% and 52%, respectively, based on the short read false duplication identification, whereas purge_dups achieved 47%, 84%, 95% and 60% for those metrics (**Figure 4. 5b**). In the human assembly, purge mers showed better performance than purge_dups except recall, different from zebra finch (**Figure 4.**

5c). In Particular, purge mers showed greater difference in Precision and F1 up to 17% and 16% than purge_dups.

Table 4. 3 The amount of false duplication in each assembly calculated by each sequencing technology.

Species	Read	Tool	Length (assembly)	Length (false duplication)	Proportion (false duplication)
Zebra finch	Short read	Purge_dups	1,165,444,248	93,176,878	8.0%
Zebra finch	Long read	Purge_dups	1,165,444,248	66,306,741	5.7%
Human	Short read	Purge_dups	2,917,922,356	81,845,196	2.8%
Human	Long read	Purge_dups	2,917,922,356	67,238,041	2.3%
Zebra finch	Short read	Purge mers	1,165,444,248	79,689,225	6.8%
Zebra finch	Long read	Purge mers	1,165,444,248	32,859,334	2.8%
Human	Short read	Purge mers	2,917,922,356	36,690,568	1.3%
Human	Long read	Purge mers	2,917,922,356	38,694,030	1.3%

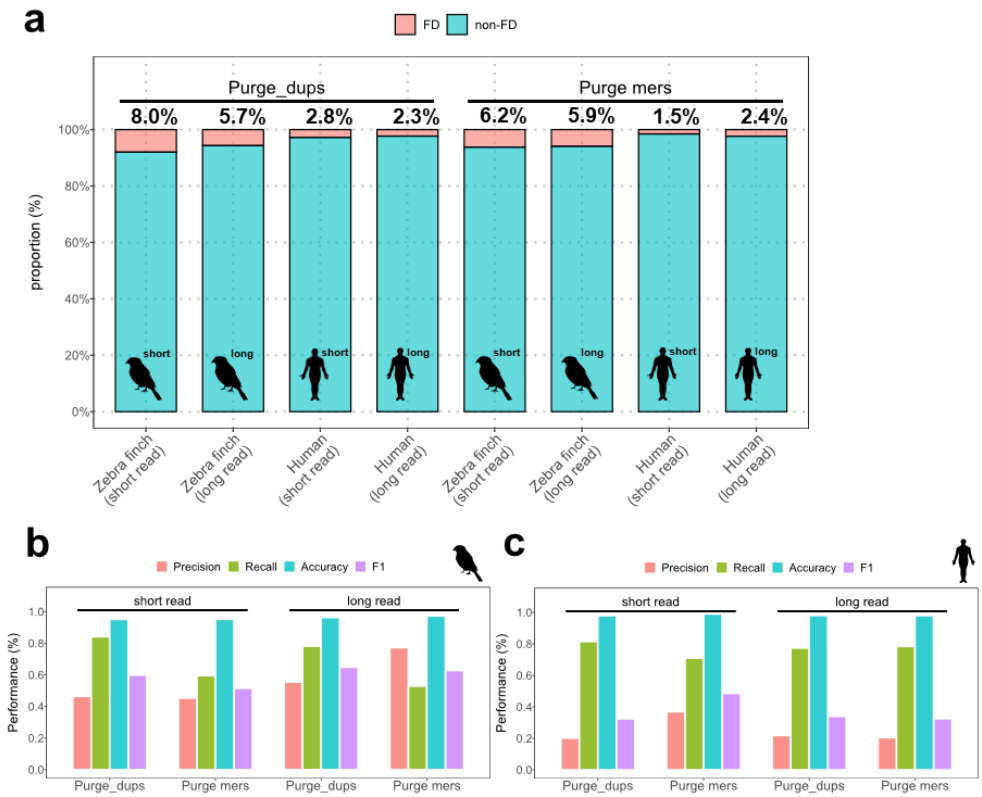

Figure 4. 5 The proportion of false duplications and performance assessment. a, proportion of false duplication in each species and sequencing technology estimated by purge_dups and purge mers. *K-mer* based performance of each method and sequencing technology in zebra finch (b) and human (c).

When using long read data from PacBio HiFi reads, purge mers identified 32.9 and 38.7 Mbp (2.8% and 1.3% of the primary assembly size) as false duplications in zebra finch and human genomes. In comparison, purge_dups detected 66.3 and 67.2 Mbp of false duplications (5.7% and 2.3% of the primary assembly size) in these species. In contrast to the results obtained with short reads, purge mers demonstrate better performance in terms of precision (77%), accuracy (97%) compared to purge_dups (56%, 96%) in the zebra finch assembly. In the human assembly, purge mers did not showed better performance except recall (79%) than purge dups (77%). Both purge mers and purge_dups showed higher performance with long read data compared to short read data in the benchmark evaluation on zebra finch, but purge_dups only showed this pattern in the human data.

4.5 Discussion

Previous methods for removing false duplications relied solely on identifying regions with low read depth coverage in the genome assembly (Guan et al., 2020; Roach et al., 2018). This approach is generally reliable since most false duplications exhibit haploid-level depth coverage or lower. However, it can be influenced by the mapping algorithm used, and a simple threshold-based method is susceptible to both false positives and negatives due to local deviations in read mapping and the sequencing process. Although long reads can be used as an alternative to mitigate fluctuations in read coverage, they may not accurately evaluate small contigs shorter than the reads themselves (e.g., <20kbp) since the reads may not be mapped to those contigs. In contrast, purge mers supplements these limitations by incorporating not only read coverage but also the $K*$ metric, which is commonly used for regional assembly evaluation (Formenti et al., 2022). Furthermore, purge mers operates on a model-based approach using maximum likelihood estimation (MLE) with automatic parameter estimation from the data itself. It partially exhibits better performance compared to previous tools when long read data from the assembly is available. This study also proposes a new revised $K*$ metric,

which incorporates the variance of *k−mer* copy numbers from sequencing reads. This metric determines whether a *k−mer* is falsely duplicated in the assembly or not, as same as the previous *K** metric, but it is more suitable for statistical inference for probabilistic analysis.

The benchmark results indicated that long reads outperformed short reads in identifying false duplications for a high heterozygous species, but this was deviated by tools in a low heterozygous species. There could be several reasons that using short read data partially outperformed long read. Firstly, the simulated short reads in this study had a higher coverage (60X) compared to the long reads (40X). Having higher read coverage provides an advantage in accurate parameter estimation with lower variance. Another factor that contributed to the better performance of short reads is the deficiency in read mapping to small contigs by long reads. The simulated assembly included a significant number of small contigs (<2kbp) in both zebra finch and human assemblies. However, the long reads were not properly mapped to these small contigs, leaving no room to determine if the small contigs were not false duplications.

While purge mers offers a more informative and sophisticated approach for identifying false duplications, performance of purge mers is heavily sensitive to the accuracy of parameter estimation and model fitting. Sampling bias is an important issue in Gaussian Mixture Models (GMM). Purge mers samples non-overlapping windows in contigs smaller than a specific threshold, but this threshold may vary across different datasets. Moreover, if falsely duplicated regions in the assembly are too small, the GMM may fail to identify the model for false duplication. Finding the optimal number of model components (K) in GMM can be challenging in the presence of complex assembly features. For example, a significant amount of repeats in the genome can disrupt the mixtures, making it hard to decide the K. Sequencing biases, such as GC-rich regions (Ross et al., 2013) or GA dropout (Formenti et al., 2022), can also lead to inaccurate false duplication identification. These regions can affect the estimation of *k-mer* counts from sequencing reads, leading to underestimation. As a result, the $K*$ metric may shift to negative values, causing false positives in false duplication identification.

In this study, purge mers showed better performance in partial data set and criteria. The cause of these discordance

between tools and data sets was not fully explained, it should be investigated in further research.

This chapter will be published elsewhere

as a partial fulfillment of Byung June Ko's Ph.D program.

Chapter 5. A *K-mer* Counting Method Minimizing GC bias in Sequencing Reads

5.1 Abstract

K−mers, substrings of length k derived from the genomic sequences, have been used to assess the quality of genome assemblies and quantify assembly errors. The multiplicity of *k−mers* obtained from sequencing reads provides information about the expected number of occurrences of a particular *k−mer* in an assembly. However, the underestimation of multiplicity due to sequencing suppression in GC−rich regions can lead to a significant disparity between the true multiplicity of *k−mers* in a genome and the measured multiplicity based on short−read *k−mer* analysis. I propose a method that reduces the bias in *k−mer* multiplicities derived from sequencing reads by applying a weighting approach using a bias function. The bias function is constructed using read coverages of non−erroneous, single−copy genomic regions from zebra finch and human assemblies, then used to weight the *k−mer* multiplicities in the *K** calculation for each species. As a result, the GC−rich regions in both assemblies exhibited a biased distribution of *K** values prior to bias removal. However, after applying the proposed method to remove the bias, the *K** distribution in these regions was corrected.

5.2 Introduction

K−mers, which are substrings of genome sequences, have been widely used to quantify genome characteristics and evaluate genome assemblies (Nurk et al., 2022; Rhie et al., 2021, 2020). These applications rely on information about the count or multiplicity of *k−mers*, which indicates the number of occurrences of identical *k−mers* in both the assembly and sequencing reads. *K** was introduced as a metric to assess whole or local regions of a genome assembly by quantifying erroneous regions, using the *k−mer* multiplicities from both the assembly and reads (Formenti et al., 2022; Phillippy et al., 2008).

In Chapter 4, I employed the *K** metric of genomic regions as a correlated variable to identify false duplications (Ko et al., 2022). However, the evaluation using *K** assumes uniform sequencing coverage along the genome, which is often not the case due to known *k−mer* biases (Formenti et al., 2022). Several sequencing biases have been reported from both short read and long read sequencing platforms. For instance, Ross et al. (2013) observed significant deviations from a uniform read coverage distribution along the GC proportion of the human genome for Illumina HiSeq and Ion Torrent PGM platforms. They showed significant drops in

read coverage for genomic regions with extremely low and high GC proportions. This GC bias was also observed in several vertebrate species, and to be a cause of false sequence losses of promoter regions in the assemblies (Kim et al., 2022). PacBio platform was also reported about read coverage dropouts in GA-rich sequences in HiFi reads of human T2T assembly (Formenti et al., 2022; Nurk et al., 2022).

Several suggestions have been proposed to address this issue. For example, Formenti et al. (2022) demonstrated that long reads can help to mitigate bias from short reads as their complementary. Benjamini and Speed (2012) suggested a direct method to correct Illumina sequencing read coverage using GC-bias models estimated from human tissues. However, combining both short and long reads is limited to researchers who have access to both types of reads. Furthermore, the models need to be updated to accommodate advancements in library preparation or sequencing platforms.

In this study, I estimated the sequencing bias of Illumina reads using the 10X-Linked platform in two vertebrate species, zebra finch and human. I then corrected *k-mer* multiplicities using a bias function derived from these estimates. I calculated the K^*

metric using both the normal *k−mer* count and the GC−bias corrected *k−mer* count at the whole−genome level, demonstrating the impact of the GC−bias removal method on *K** calculations.

5.3 Materials and Methods

5.3.1 Bias function estimation

If we have information about the degree of suppression caused by GC contents in sequencing data, we can use this information as weights for removing GC bias removal in $k-mer$ counting. For calculating GC bias-removed $k-mer$ counts, we can begin with a simple read coverage model of a locus i.

$$Read\ coverage\ of\ locus\ i = C \times f(GC_i) + \mathcal{E}_i$$

Where C is a constant of mean genome coverage of sequencing reads, and $f(GC)$ is a function that describes the inhibitor effect caused by GC bias. "$\mathcal{E}$" is a random variation of read producing in sequencing process which is not related with GC bias. Based on the read coverage model, we can propose a simple approximate model for the copy number (or multiplicity) of a $k-mer$ counted from sequencing reads in n loci as follows:

$$Kmer\ count \approx \sum_{i=1}^{n} C \times f(GC_i) + \mathcal{E}_i$$

$$\approx \sum_{i=1}^{n} C \times f(GC_i) + \mathcal{E}_i$$

As the sum of random variations "$\mathcal{E}$" is converged to zero when the count is enoughly large, the effects of GC inhibitions can be canceled-out by multiplying the reciprocal of the bias function value for each GC value of the reads that include the *k-mer*. To estimate GC biases, we can consider a model of read depth coverage obtained by mapping reads to the genome assembly in a locus with a GC proportion of x first.

$$Depth\ coverage_x = C \times f(GC_x) + \mathcal{E}_x + r_{fp} + r_{fn}$$

Where r_{fp} is false positive in read mapping, and r_{fn} is false negative. A difference between the read coverage model described above and the depth coverage here is that the depth coverage model should be considered to include the false positive and negative from the read mapping process. If we can calculate depth coverages from regions that r_{fp}=0 and r_{fn}=0, we can estimate the function value of *f(GCx)* by summing of the depth coverages from all *n* loci that share the same GC content *(GCx)* as follows:

$$\sum_{i=1}^{n} Depth\ coverage_{xi} = \sum_{i=1}^{n} C \times f(GC_{xi}) + \sum_{i=1}^{n} \mathcal{E}_i$$

$$\therefore f(GC_x) = \frac{\sum_{i=1}^{n} Depth\ coverage_{xi}}{C\ n}$$

In this study, we selected regions that consisted of 1−copy $k-mers$ and had $K* = 0$ in the assembly to obtain regions with $r_{fp} = 0$ and $r_{fn} = 0$. Using the equation mentioned above, we can estimate the function values for specific GC proportions, representing the degree of inhibition caused by GC contents during producing sequencing reads. The genome coverage constant C was approximated from the data by maximizing $C*f(GCx)$ across various GC proportions, considering that GC content always contributes to suppression. Finally, a new $k-mer$ count for a specific $k-mer$ from reads m, incorporating the bias function $f(GC)$, is calculated as follows:

$$GC\ Bias\ removed\ Kmer\ count = \sum_{j=1}^{m} \frac{1}{f(GC_j)}$$

5.3.2 Used data

Two assemblies of vertebrate species, primary assembly of zebra finch (bTaeGut2) and human (maternal of mHomSap), were used for bias function estimation. The zebra finch assembly was assembled by VGP 2.0 pipeline with PacBio HiFi reads. The Human assembly was fully phased by VGP trio pipeline, and assembled using PacBio CLR reads. The bias functions were calculated from

depth-coverage profiles of primary assemblies by 10X Linked reads of the species mapping to each assembly. EMA (Shajii et al., 2018) was used to map 10X Linked reads with barcode sequence. BWA (Li and Durbin, 2009) was used for read mapping without barcode following EMA standard pipeline. The mapping was multi-processed by Parallel (Tange, 2011). Merging and sorting the read mapping file (.bam) was done by Sambamba and Samtools (Li et al., 2009).

5.3.3 *K** calculation

Meryl (Rhie et al., 2020) was used for genome-wide *k-mer* counting from assembly and 10X Linked reads. For calculating weighted *k-mer* counts with bias function, a new source code was developed with saving information of GC proportion of each read and each *k-mer*. *K** was calculated from both normal *k-mer* counting (Formenti et al., 2022) and GC-bias removed *k-mer* counting methods. Diploid genome sequences by combining both paternal and maternal sequences were used for *K** calculation. To calculate expected copy number of *k-mer*, Genomescope2.0 (Ranallo-Benavidez et al., 2020) was used for generating a

probability model distribution of *k−mer* multiplicities. Merqury (Rhie et al., 2020) was used for generating histogram of *k−mer* multiplicity from reads for Genomescope2.0. Bedtools (Quinlan and Hall, 2010) was used for single copy region merging and sequence extraction.

5.4 Results and Discussion

5.4.1 Bias function estimation

For both the zebra finch and human assemblies, genomic regions spanning 100 and 764 Mbp, respectively, consisting of single-copy *k-mers* and exhibiting no signs of duplication or loss errors, were extracted. The read coverage profiles of 10X Linked reads mapped to these regions were utilized to estimate the function values of GC bias. In the case of the zebra finch assembly, the function values of the bias function ranged from 1.0 to 5.1, with the highest value observed at >88% GC proportion (**Figure 5. 1a**). On the other hand, in the human assembly, the function values ranged from 1.0 to 7.6, with >90% GC proportion showing the highest value (**Figure 5. 1b**). These findings align with previous knowledge of GC bias in short-read sequencing (Ross et al., 2013). It is known that there is a moderate inhibition of sequencing in regions with GC proportions higher than 40, and the inhibitory effect gradually increases as the GC proportion reaches extremely high levels.

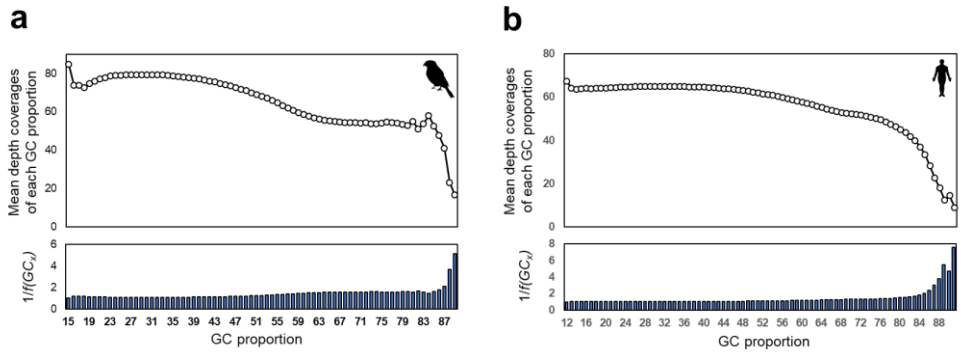

Figure 5. 1 Mean depth coverage and bias function with along GC proportions for a, zebra finch, and b, human assemblies.

5.4.2 *K** distributions along GC proportions

*K** is a metric that can be used for evaluating assembly in both genome-wide and regional level. A positive value of *K** indicates the presence of false collapses in the assembly, while a negative value suggests false expansions (Formenti et al., 2022). The effectiveness of *K** relies on the assumption that sequencing reads are uniformly produced throughout the genome. However, a negative *K** value can also indicate unexpectedly low expected copy numbers of *k-mers* in the assembly, implying the possibility of sequencing suppression affecting a particular *k-mer*.

In this study, I calculated *K** distributions based on the mean GC proportion of *k-mers*, categorized into three levels: moderate, high, and extremely high GC proportions. The distribution of *k-mer* counts without GC bias removal revealed that the highest frequency of *K** values peaked at 0 for both moderate and high GC level *k-mers* (**Figure 5. 2a, c**). However, in the distribution of extremely high GC *k-mers*, the peak value was -1. This indicates that *k-mers* derived from sequencing reads in high GC regions were affected by either false duplication or sequencing suppression. While false losses are commonly associated with GC-rich regions (Kim et al., 2022), it is less known that false duplications correlate

with such regions. Therefore, the *k-mers* in the extremely high GC category seem to be affected by suppression in the sequencing step rather than false duplication. These patterns were also represented in K^* distribution of human genome assembly.

On the contrary, when using GC-bias removed *k-mers*, K^* distributions of all GC categories in both zebra finch and human assemblies exhibited a peak at 0 for the highest frequency (**Figure 5. 2b, d**). This indicates that the K^* metrics for extremely high GC regions were corrected, eliminating the unexpected -1 values in the assemblies. Although I did not calculate K^* using long-read sequencing data such as PacBio HiFi, it has been reported that unexpected dropouts in read coverage profiles occur in GA-rich regions of the human T2T assembly (Formenti et al., 2022; Nurk et al., 2022). Therefore, for more extensive application and testing, it would be valuable to include various sequencing data from other platforms such as PacBio and Oxford Nanopore Technology in future research. In recent mega-scale genome projects, *k-mer* counting has become a common method for assessing assembly quality and conducting regional evaluations (Nurk et al., 2022; Rhie et al., 2021). In Chapter 4, I used the K^* metric to identify false duplications. However, these evaluations assume that there is no

sequencing bias across the entire genome. Similarly, other popular assembly evaluation tools also depend on this assumption (Formenti et al., 2022; Rhie et al., 2020). The GC-bias removed *k-mer* counting method proposed in this study can serve as a solution to prevent biological misinterpretation that may occur in high GC regions of genome assemblies.

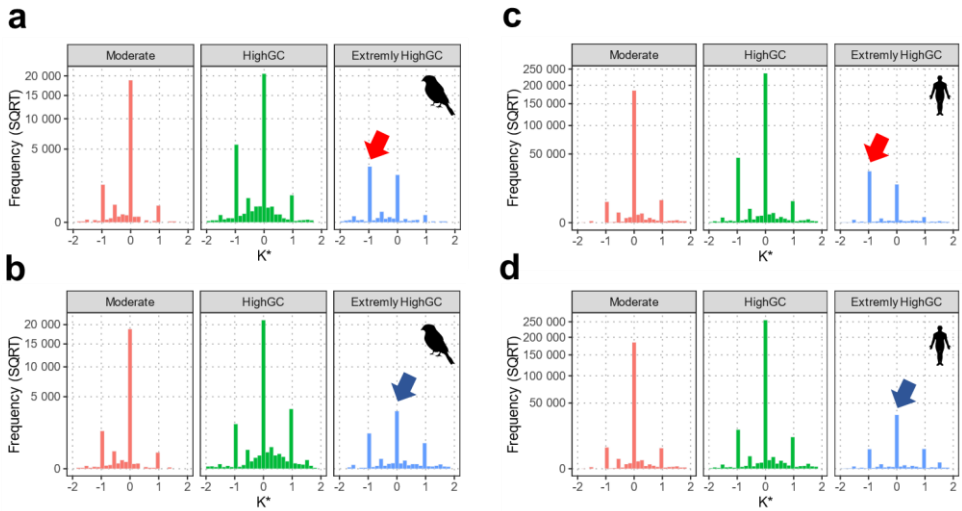

Figure 5. 2 *K distribution across GC proportion categories.** (a) *K** distributions without GC-bias removal in the zebra finch assembly. (b) *K** distributions with GC-bias removal in the zebra finch assembly. (c) *K** distribution with GC-bias in the human assembly. (d) *K** distributions with GC-bias removal in the human assembly. GC proportion categories: moderate (20%-40%), high GC (50%-80%), extremely high GC (>80%).

General discussion

In this study, I investigated the primary characteristics of false duplications in genome assemblies and their impact on gene annotations and downstream analyses. I found that false duplications exhibit specific patterns such as lower read coverage, presence of gaps between duplicated pairs, and discordant linked read pairs. These false duplications were more prevalent in previous Sanger-based assemblies compared to the VGP PacBio-based long-read assemblies. Heterozygosity levels and sequencing errors were identified as major sources of false duplications. These false duplications resulted in mis-annotations of genes, exons, and chimeric gene gains, leading to misinterpretations in comparative genomics and genome assembly-based research. Haplotype phasing and careful evaluation of assemblies were recommended to mitigate false duplications. The VGP assembly pipeline has been updated to incorporate new sequencing technologies, such as HiFi reads, which are better to reduce false duplications caused by sequencing errors. Future genomic studies should prioritize haplotype-phased assemblies free of false duplications.

The findings also revealed that heterotype false duplications are more prevalent than homotype duplications, highlighting the need for

improved haplotype separation in genome assemblies. Scanning regions around gaps and utilizing various profiling techniques can help identify false duplications. Trio data were found to be effective in reducing false duplications, but the availability of parental data is a limiting factor. HiFi-HiC assembly can be one of the alternatives for the limitation.

Previous false duplication identification methods relied on read coverage alone, but incorporating additional metrics such as *k-mer* analysis can improved the accuracy of false duplication identification. Long read platform is more prevalently used in genome assembly projects recently, but short reads performed better than long reads in identifying false duplications in human data with purge mers. Furthermore, this study highlighted the impact of GC bias in *k-mer* counting. The proposed GC-bias removed *k-mer* counting method provides a solution to mitigate underestimation of counting in high GC regions of genome assemblies. Continuous improvement of sequencing and assembly pipelines, including the integration of new technologies and refined evaluation methods, is crucial for generating accurate reference genome assemblies with objective evaluation.

In summary, this study emphasized the significance of addressing false duplications in genome assemblies to ensure reliable gene

annotations and accurate downstream analyses. It highlighted the characteristics of false duplications, identified their sources, and proposed strategies for their detection and prevention. These findings contribute to the ongoing efforts in creating high-quality reference genomes and emphasize the importance of haplotype-phased assemblies free of false duplications in genomics research.

References

A reference standard for genome biology, 2018. . Nat. Biotechnol. 36, 1121. https://doi.org/10.1038/nbt.4318

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.-H.C., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Miklos, Abril, J.F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., Pablos, B. de, Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W.,

Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D.C., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.-Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye,

J., Yeh, R.-F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter, J.C., 2000. The Genome Sequence of Drosophila melanogaster. Science 287, 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Ameur, A., Kloosterman, W.P., Hestand, M.S., 2019. Single-Molecule Sequencing: Towards Clinical Applications. Trends Biotechnol. 37, 72–85. https://doi.org/10.1016/j.tibtech.2018.07.013

Armstrong, J., Fiddes, I.T., Diekhans, M., Paten, B., 2019. Whole-Genome Alignment and Comparative Annotation. Annu. Rev. Anim. Biosci. 7, 41–64. https://doi.org/10.1146/annurev-animal-020518-115005

Benjamini, Y., Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40, e72. https://doi.org/10.1093/nar/gks001

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca,

N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., Korf, I.F., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2. https://doi.org/10.1186/2047-217X-2-10

Bresler, G., Bresler, M., Tse, D., 2013. Optimal assembly for high throughput shotgun sequencing. BMC Bioinformatics 14, S18. https://doi.org/10.1186/1471-2105-14-S5-S18

Cabanettes, F., Klopp, C., 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6, e4958. https://doi.org/10.7717/peerj.4958

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., DePristo, M.A., 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13, 375. https://doi.org/10.1186/1471-2164-13-375

Chen, F.-C., Chen, C.-J., Li, W.-H., Chuang, T.-J., 2010. Gene Family Size Conservation Is a Good Indicator of Evolutionary Rates. Mol. Biol. Evol. 27, 1750–1758. https://doi.org/10.1093/molbev/msq055

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175. https://doi.org/10.1038/s41592-020-01056-5

Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.-C., Scherer, S.W., 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. 4, R25. https://doi.org/10.1186/gb-2003-4-4-r25

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O' Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M., Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., Schatz, M.C., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054. https://doi.org/10.1038/nmeth.4035

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E.E., Weinstock, G., Mardis, E.R., Wilson, R.K., Howe, K., Flicek, P., Hubbard, T., 2011. Modernizing Reference Genome Assemblies. PLOS Biol. 9, e1001091. https://doi.org/10.1371/journal.pbio.1001091

Consortium, I.C.G.S., 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.

https://doi.org/10.1038/nature03154

Dean, M., Annilo, T., 2005. Evolution of the Atp-Binding Cassette (abc) Transporter Superfamily in Vertebrates. Annu. Rev. Genomics Hum. Genet. 6, 123-142. https://doi.org/10.1146/annurev.genom.6.080604.162122

Denisenko-Nehrbass, N.I., Jarvis, E., Scharff, C., Nottebohm, F., Mello, C.V., 2000. Site-Specific Retinoic Acid Production in the Brain of Adult Songbirds. Neuron 27, 359-370. https://doi.org/10.1016/S0896-6273(00)00043-X

Ekblom, R., Wolf, J.B.W., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7, 1026-1042. https://doi.org/10.1111/eva.12178

Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29, 51-63. https://doi.org/10.1016/j.tree.2013.09.008

Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T.N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L.G., Muzny, D.M., Sabo, A., Gibbs, R.A., Eichler, E.E., O' Roak, B.J., Michaelson, J.J., Volfovsky, N., Shen, Y., Chung, W.K., 2019. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. Npj Genomic Med.

4, 1-14. https://doi.org/10.1038/s41525-019-0093-8

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A.H., Xie, D., Chen, G., Guo, C., Faircloth, B.C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J.T., Parent, C., Pacheco, G., Sinding, M.-H.S., Puetz, L., Cavill, E., Ribeiro, Â.M., Eckhart, L., Fjeldså, J., Hosner, P.A., Brumfield, R.T., Christidis, L., Bertelsen, M.F., Sicheritz-Ponten, T., Tietze, D.T., Robertson, B.C., Song, G., Borgia, G., Claramunt, S., Lovette, I.J., Cowen, S.J., Njoroge, P., Dumbacher, J.P., Ryder, O.A., Fuchs, J., Bunce, M., Burt, D.W., Cracraft, J., Meng, G., Hackett, S.J., Ryan, P.G., Jønsson, K.A., Jamieson, I.G., da Fonseca, R.R., Braun, E.L., Houde, P., Mirarab, S., Suh, A., Hansson, B., Ponnikas, S., Sigeman, H., Stervander, M., Frandsen, P.B., van der Zwan, H., van der Sluis, R., Visser, C., Balakrishnan, C.N., Clark, A.G., Fitzpatrick, J.W., Bowman, R., Chen, N., Cloutier, A., Sackton, T.B., Edwards, S.V., Foote, D.J., Shakya, S.B., Sheldon, F.H., Vignal, A., Soares, A.E.R., Shapiro, B., González-Solís, J., Ferrer-Obiol, J., Rozas, J., Riutort, M., Tigano, A., Friesen, V., Dalén, L., Urrutia, A.O., Székely, T., Liu, Y., Campana, M.G., Corvelo,

A., Fleischer, R.C., Rutherford, K.M., Gemmell, N.J., Dussex, N., Mouritsen, H., Thiele, N., Delmore, K., Liedvogel, M., Franke, A., Hoeppner, M.P., Krone, O., Fudickar, A.M., Milá, B., Ketterson, E.D., Fidler, A.E., Friis, G., Parody-Merino, Á.M., Battley, P.F., Cox, M.P., Lima, N.C.B., Prosdocimi, F., Parchman, T.L., Schlinger, B.A., Loiselle, B.A., Blake, J.G., Lim, H.C., Day, L.B., Fuxjager, M.J., Baldwin, M.W., Braun, M.J., Wirthlin, M., Dikow, R.B., Ryder, T.B., Camenisch, G., Keller, L.F., DaCosta, J.M., Hauber, M.E., Louder, M.I.M., Witt, C.C., McGuire, J.A., Mudge, J., Megna, L.C., Carling, M.D., Wang, B., Taylor, S.A., Del-Rio, G., Aleixo, A., Vasconcelos, A.T.R., Mello, C.V., Weir, J.T., Haussler, D., Li, Q., Yang, H., Wang, J., Lei, F., Rahbek, C., Gilbert, M.T.P., Graves, G.R., Jarvis, E.D., Paten, B., Zhang, G., 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature 587, 252-257. https://doi.org/10.1038/s41586-020-2873-9

Formenti, G., Rhie, A., Walenz, B.P., Thibaud-Nissen, F., Shafin, K., Koren, S., Myers, E.W., Jarvis, E.D., Phillippy, A.M., 2022. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. Nat. Methods 19, 696-704.

https://doi.org/10.1038/s41592-022-01445-y

Friedrich, S.R., Lovell, P.V., Kaser, T.M., Mello, C.V., 2019. Exploring the molecular basis of neuronal excitability in a vocal learner. BMC Genomics 20, 629. https://doi.org/10.1186/s12864-019-5871-2

Gel, B., Serra, E., 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33, 3088–3090. https://doi.org/10.1093/bioinformatics/btx346

Genome 10K Community of Scientists, 2009. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. J. Hered. 100, 659–674. https://doi.org/10.1093/jhered/esp086

Giani, A.M., Gallo, G.R., Gianfranceschi, L., Formenti, G., 2020. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput. Struct. Biotechnol. J. 18, 9–19. https://doi.org/10.1016/j.csbj.2019.11.002

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B.,

2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 108, 1513-1518. https://doi.org/10.1073/pnas.1017351108

Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., Durbin, R., 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36, 2896-2898. https://doi.org/10.1093/bioinformatics/btaa025

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., Hahn, M.W., 2013. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. Mol. Biol. Evol. 30, 1987-1997. https://doi.org/10.1093/molbev/mst100

Hickey, G., Paten, B., Earl, D., Zerbino, D., Haussler, D., 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29, 1341-1342. https://doi.org/10.1093/bioinformatics/btt128

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., Gao, G., 2015. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics 31, 1296-1297. https://doi.org/10.1093/bioinformatics/btu817

Huang, X., Han, B., 2014. Natural Variations and Genome-Wide

Association Studies in Crop Plants. Annu. Rev. Plant Biol. 65, 531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M.R., Porubsky, D., Cheng, H., Asri, M., Logsdon, G.A., Carnevali, P., Chaisson, M.J.P., Chin, C.-S., Cody, S., Collins, J., Ebert, P., Escalona, M., Fedrigo, O., Fulton, R.S., Fulton, L.L., Garg, S., Gerton, J.L., Ghurye, J., Granat, A., Green, R.E., Harvey, W., Hasenfeld, P., Hastie, A., Haukness, M., Jaeger, E.B., Jain, M., Kirsche, M., Kolmogorov, M., Korbel, J.O., Koren, S., Korlach, J., Lee, J., Li, D., Lindsay, T., Lucas, J., Luo, F., Marschall, T., Mitchell, M.W., McDaniel, J., Nie, F., Olsen, H.E., Olson, N.D., Pesout, T., Potapova, T., Puiu, D., Regier, A., Ruan, J., Salzberg, S.L., Sanders, A.D., Schatz, M.C., Schmitt, A., Schneider, V.A., Selvaraj, S., Shafin, K., Shumate, A., Stitziel, N.O., Stober, C., Torrance, J., Wagner, J., Wang, J., Wenger, A., Xiao, C., Zimin, A.V., Zhang, G., Wang, T., Li, H., Garrison, E., Haussler, D., Hall, I., Zook, J.M., Eichler, E.E., Phillippy, A.M., Paten, B., Howe, K., Miga, K.H., 2022. Semi-automated assembly of high-quality

diploid human reference genomes. Nature 611, 519–531. https://doi.org/10.1038/s41586-022-05325-5

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., Suh, A., Weber, C.C., Fonseca, R.R. da, Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H., Brumfield, R.T., Mello, C.V., Lovell, P.V., Wirthlin, M., Schneider, M.P.C., Prosdocimi, F., Samaniego, J.A., Velazquez, A.M.V., Alfaro-Núñez, A., Campos, P.F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman, P., Driskell, A.C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F.E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F.K., Jønsson, K.A., Johnson, W., Koepfli, K.-P., O' Brien, S., Haussler, D., Ryder, O.A.,

Rahbek, C., Willerslev, E., Graves, G.R., Glenn, T.C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S.V., Stamatakis, A., Mindell, D.P., Cracraft, J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P., Zhang, G., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331. https://doi.org/10.1126/science.1253451

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., Davis, R.W., Scherer, S., 2004. The diploid genome sequence of Candida albicans. Proc. Natl. Acad. Sci. U. S. A. 101, 7329–7334. https://doi.org/10.1073/pnas.0401648101

Kelley, D.R., Salzberg, S.L., 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 11, R28. https://doi.org/10.1186/gb-2010-11-3-r28

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., Frith, M.C., 2011. Adaptive seeds tame genomic sequence comparison. Genome Res. 21, 487–493. https://doi.org/10.1101/gr.113985.110

Kim, J., Lee, C., Ko, B.J., Yoo, D.A., Won, S., Phillippy, A.M.,

Fedrigo, O., Zhang, G., Howe, K., Wood, J., Durbin, R., Formenti, G., Brown, S., Cantin, L., Mello, C.V., Cho, S., Rhie, A., Kim, H., Jarvis, E.D., 2022. False gene and chromosome losses in genome assemblies caused by GC content variation and repeats. Genome Biol. 23, 204. https://doi.org/10.1186/s13059-022-02765-0

Ko, B.J., Lee, C., Kim, J., Rhie, A., Yoo, D.A., Howe, K., Wood, J., Cho, S., Brown, S., Formenti, G., Jarvis, E.D., Kim, H., 2022. Widespread false gene gains caused by duplication errors in genome assemblies. Genome Biol. 23, 205. https://doi.org/10.1186/s13059-022-02764-1

Koepfli, K.-P., Paten, B., and O' Brien, S.J., 2015. The Genome 10K Project: A Way Forward. Annu. Rev. Anim. Biosci. 3, 57-111. https://doi.org/10.1146/annurev-animal-090414-014900

Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., Phillippy, A.M., 2018. De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol. 36, 1174-1182. https://doi.org/10.1038/nbt.4277

Korlach, J., Gedman, G., Kingan, S.B., Chin, C.-S., Howard, J.T.,

Audet, J.-N., Cantin, L., Jarvis, E.D., 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience 6. https://doi.org/10.1093/gigascience/gix085

Kronenberg, Z.N., Rhie, A., Koren, S., Concepcion, G.T., Peluso, P., Munson, K.M., Porubsky, D., Kuhn, K., Mueller, K.A., Low, W.Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R.J., Phillippy, A.M., Eichler, E.E., Williams, J.L., Smith, T.P.L., Jarvis, E.D., Sullivan, S.T., Kingan, S.B., 2021. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat. Commun. 12, 1935. https://doi.org/10.1038/s41467-020-20536-y

Kurz, A., Wöhr, M., Walter, M., Bonin, M., Auburger, G., Gispert, S., Schwarting, R.K.W., 2010. Alpha-synuclein deficiency affects brain Foxp1 expression and ultrasonic vocalization. Neuroscience 166, 785–795. https://doi.org/10.1016/j.neuroscience.2009.12.054

Levy, G.G., Nichols, W.C., Lian, E.C., Foroud, T., McClintick, J.N., McGee, B.M., Yang, A.Y., Siemieniak, D.R., Stark, K.R., Gruppo, R., Sarode, R., Shurin, S.B., Chandrasekaran, V.,

Stabler, S.P., Sabio, H., Bouhassira, E.E., Upshaw, J.D., Ginsburg, D., Tsai, H.-M., 2001. Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura. Nature 413, 488-494. https://doi.org/10.1038/35097008

Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S.-M., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., DiPalma, F., Durbin, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Gilbert, M.T.P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C.J., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O' Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C.,

Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D., Zhang, G., 2022. The Earth BioGenome Project 2020: Starting the clock. Proc. Natl. Acad. Sci. 119, e2115635118. https://doi.org/10.1073/pnas.2115635118

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, Jing, Wang, Z., Li, J., Xu, L., Liu, Jiaqi, Feng, S., Guo, C., Chen, S., Ren, Z., Rao, J., Wei, K., Chen, Y., Jarvis, E.D., Zhang, G., Zhou, Q., 2021. A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian

chromosomes. Genome Res. 31, 497–511. https://doi.org/10.1101/gr.271569.120

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1. https://doi.org/10.1186/2047-217X-1-18

Mc Cartney, A.M., Shafin, K., Alonge, M., Bzikadze, A.V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G.A., Miga, K.H., Mikheenko, A., Paten, B., Shumate, A., Soto, D.C., Sović, I., Wood, J.M.D., Zook, J.M., Phillippy, A.M., Rhie, A., 2022. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat. Methods 19, 687–695. https://doi.org/10.1038/s41592-022-01440-3

Nakagawa, H., Fujita, M., 2018. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. 109, 513–522. https://doi.org/10.1111/cas.13505

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S.L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A.,

Gerton, J.L., O’ Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376, 44–53. https://doi.org/10.1126/science.abj6987

Ono, Y., Hamada, M., Asai, K., 2022. PBSIM3: a simulator for all types of PacBio and ONT long reads. NAR Genomics Bioinforma. 4, lqac092. https://doi.org/10.1093/nargab/lqac092

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D., 2011. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 21, 1512–1528. https://doi.org/10.1101/gr.123356.111

Phillippy, A.M., Schatz, M.C., Pop, M., 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 9, R55. https://doi.org/10.1186/gb-2008-9-3-r55

Pryszcz, L.P., Gabaldón, T., 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44, e113–e113. https://doi.org/10.1093/nar/gkw294

Quesada, V., Velasco, G., Puente, X.S., Warren, W.C., López-Otín, C., 2010. Comparative genomic analysis of the zebra finch degradome provides new insights into evolution of proteases

in birds and mammals. BMC Genomics 11, 220. https://doi.org/10.1186/1471-2164-11-220

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842. https://doi.org/10.1093/bioinformatics/btq033

Ranallo-Benavidez, T.R., Jaron, K.S., Schatz, M.C., 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. 11, 1432. https://doi.org/10.1038/s41467-020-14998-3

Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J., 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35, W193-W200. https://doi.org/10.1093/nar/gkm226

Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B.J., Chaisson, M., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S.C., Lama, T.M., Grutzner, F., Warren, W.C., Balakrishnan, C.N., Burt, D., George, J.M., Biegler, M.T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T.,

Wilkinson, M., Turner, G., Meyer, A., Kautt, A.F., Franchini, P., Detrich, H.W., Svardal, H., Wagner, M., Naylor, G.J.P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B.T., Pesout, T., Houck, M., Misuraca, A., Kingan, S.B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R.E., Putnam, N.H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R.E., Guan, D., London, S.E., Clayton, D.F., Mello, C.V., Friedrich, S.R., Lovell, P.V., Osipova, E., Al-Ajli, F.O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R.S., Makova, K.D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, Kevin, Flicek, P., Walenz, B.P., Kwak, W., Clawson, H., Diekhans, M., Nassar, L., Paten, B., Kraus, R.H.S., Crawford, A.J., Gilbert, M.T.P., Zhang, G., Venkatesh, B., Murphy, R.W., Koepfli, K.-P., Shapiro, B., Johnson, W.E., Di Palma, F., Marques-Bonet, T., Teeling, E.C., Warnow, T., Graves, J.M., Ryder, O.A., Haussler, D., O'Brien, S.J., Korlach, J., Lewin, H.A., Howe, Kerstin, Myers, E.W., Durbin, R., Phillippy, A.M., Jarvis, E.D., 2021. Towards complete and error-free genome assemblies of all

vertebrate species. Nature 592, 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M., 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21, 245. https://doi.org/10.1186/s13059-020-02134-9

Rice, E.S., Green, R.E., 2019. New Approaches for Genome Assembly and Scaffolding. Annu. Rev. Anim. Biosci. 7, 17–40. https://doi.org/10.1146/annurev-animal-020518-115344

Roach, M.J., Schmidt, S.A., Borneman, A.R., 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19, 460. https://doi.org/10.1186/s12859-018-2485-7

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., Jaffe, D.B., 2013. Characterizing and measuring bias in sequence data. Genome Biol. 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Salzberg, S.L., Yorke, J.A., 2005. Beware of mis-assembled genomes. Bioinformatics 21, 4320–4321. https://doi.org/10.1093/bioinformatics/bti769

Schmeing, S., Robinson, M.D., 2021. ReSeq simulates realistic

Illumina high-throughput sequencing data. Genome Biol. 22, 67. https://doi.org/10.1186/s13059-021-02265-7

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., Fulton, R.S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K.M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J.T., Threadgold, G., Torrance, J., Wood, J.M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A.M., Durbin, R., Wilson, R.K., Flicek, P., Eichler, E.E., Church, D.M., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864. https://doi.org/10.1101/gr.213611.116

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. 8, 289–317.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C., Brännström, Å., Brelsford, A., Clarkson, C.S., Eroukhmanoff, F., Feder, J.L., Fischer, M.C., Foote, A.D., Franchini, P.,

Jiggins, C.D., Jones, F.C., Lindholm, A.K., Lucek, K., Maan, M.E., Marques, D.A., Martin, S.H., Matthews, B., Meier, J.I., Möst, M., Nachman, M.W., Nonaka, E., Rennison, D.J., Schwarzer, J., Watson, E.T., Westram, A.M., Widmer, A., 2014. Genomics and the origin of species. Nat. Rev. Genet. 15, 176–192. https://doi.org/10.1038/nrg3644

Shajii, A., Numanagić, I., Berger, B., 2018. Latent Variable Model for Aligning Barcoded Short-Reads Improves Downstream Analyses. Res. Comput. Mol. Biol. Annu. Int. Conf. RECOMB Proc. RECOMB Conf. 2005– 10812, 280–282.

Simpson, J.T., Pop, M., 2015. The Theory and Practice of Genome Sequence Assembly. Annu. Rev. Genomics Hum. Genet. 16, 153–172. https://doi.org/10.1146/annurev-genom-090314-050032

Skibinski, D.O.F., Ward, R.D., 1982. Correlations between heterozygosity and evolutionary rate of proteins. Nature 298, 490–492. https://doi.org/10.1038/298490a0

Tange, O., n.d. GNU Parallel: The Command-Line Power Tool.

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinforma. 25, 4.10.1–4.10.14.

https://doi.org/10.1002/0471250953.bi0410s25

The Galaxy Community, 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 50, W345–W351. https://doi.org/10.1093/nar/gkac247

Theofanopoulou, C., Gedman, G.L., Cahill, J.A., Boeckx, C., Jarvis, E.D., n.d. Universal nomenclature for oxytocin–vasotocin ligand and receptor families. Nature.

Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high–performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. https://doi.org/10.1093/bib/bbs017

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A.,

Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L.,

Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell,

M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The Sequence of the Human Genome. Science 291, 1304–1351. https://doi.org/10.1126/science.1058040

Vinson, J.P., Jaffe, D.B., O' Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J.E., Lander, E.S., 2005. Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. Genome Res. 15, 1127–1135. https://doi.org/10.1101/gr.3722605

Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Künstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., Heger, A., Kong, L., Ponting, C.P., Jarvis, E.D., Mello, C.V., Minx, P., Lovell, P., Velho, T.A.F., Ferris, M., Balakrishnan, C.N., Sinha, S., Blatti, C., London, S.E., Li, Y., Lin, Y.-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A.R., Howard, J., Völker, M., Skinner, B.M., Griffin, D.K., Ye, L., McLaren, W.M., Flicek, P.,

Quesada, V., Velasco, G., Lopez-Otin, C., Puente, X.S., Olender, T., Lancet, D., Smit, A.F.A., Hubley, R., Konkel, M.K., Walker, J.A., Batzer, M.A., Gu, W., Pollock, D.D., Chen, L., Cheng, Z., Eichler, E.E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S.V., Yang, S.-P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E.R., Wilson, R.K., 2010. The genome of a songbird. Nature 464, 757-762. https://doi.org/10.1038/nature08819

Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., Yang, S.-P., Heger, A., Locke, D.P., Miethke, P., Waters, P.D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X.S., López-Otín, C., Ordóñez, G.R., Eichler, E.E., Chen, L., Cheng, Z., Deakin, J.E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A.T., Wakefield, M.J., Olender, T., Lancet, D., Huttley, G.A., Smit, A.F.A., Pask, A., Temple-Smith, P., Batzer, M.A., Walker, J.A., Konkel, M.K., Harris, R.S., Whittington, C.M., Wong, E.S.W., Gemmell, N.J., Buschiazzo, E., Vargas Jentzsch, I.M., Merkel, A., Schmitz, J.,

Zemann, A., Churakov, G., Ole Kriegs, J., Brosius, J., Murchison, E.P., Sachidanandam, R., Smith, C., Hannon, G.J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefèvre, C.M., Sharp, J.A., Nicholas, K.R., Ray, D.A., Kube, M., Reinhardt, R., Pringle, T.H., Taylor, J., Jones, R.C., Nixon, B., Dacheux, J.-L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M.N., Pohl, C.S., Smith, S.M., Hou, S., Renfree, M.B., Mardis, E.R., Wilson, R.K., A list of authors and their affiliations appears at the end of the paper, 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453, 175-183. https://doi.org/10.1038/nature06936

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B., 2017. Direct determination of diploid genome sequences. Genome Res. 27, 757-767. https://doi.org/10.1101/gr.214874.116

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J.,

Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155-1162. https://doi.org/10.1038/s41587-019-0217-9

Yan, L., Zhang, J., Chen, H., Luo, H., 2021. Genome-wide analysis of ATP-binding cassette transporter provides insight to genes related to bioactive metabolite transportation in Salvia miltiorrhiza. BMC Genomics 22, 315. https://doi.org/10.1186/s12864-021-07623-0

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J., Meredith, R.W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W., Hu, J., Xiao, J., Yang, Z., Liu, Y., Xie, Q., Yu, H., Lian, J., Wen, P., Zhang, F., Li, H., Zeng, Y., Xiong, Z., Liu, S., Zhou, L., Huang, Z., An, N., Wang, Jie, Zheng, Q., Xiong, Y., Wang, G., Wang, B., Wang, Jingjing, Fan, Y., Fonseca, R.R. da,

Alfaro-Núñez, A., Schubert, M., Orlando, L., Mourier, T., Howard, J.T., Ganapathy, G., Pfenning, A., Whitney, O., Rivas, M.V., Hara, E., Smith, J., Farré, M., Narayan, J., Slavov, G., Romanov, M.N., Borges, R., Machado, J.P., Khan, I., Springer, M.S., Gatesy, J., Hoffmann, F.G., Opazo, J.C., Håstad, O., Sawyer, R.H., Kim, H., Kim, K.-W., Kim, H.J., Cho, S., Li, N., Huang, Y., Bruford, M.W., Zhan, X., Dixon, A., Bertelsen, M.F., Derryberry, E., Warren, W., Wilson, R.K., Li, S., Ray, D.A., Green, R.E., O' Brien, S.J., Griffin, D., Johnson, W.E., Haussler, D., Ryder, O.A., Willerslev, E., Graves, G.R., Alström, P., Fjeldså, J., Mindell, D.P., Edwards, S.V., Braun, E.L., Rahbek, C., Burt, D.W., Houde, P., Zhang, Y., Yang, H., Wang, Jian, Consortium, A.G., Jarvis, E.D., Gilbert, M.T.P., Wang, Jun, 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346, 1311-1320. https://doi.org/10.1126/science.1251385

국문 초록

이배체 유전체 조립 과정에서의 인위적 오류 식별과 교정을 위한 생물정보학적 접근

고 병 준

농생명공학부 바이오모듈레이션 전공

서울대학교 대학원

참조 유전체에 존재하는 조립 오류는 생물학적 해석의 오류로 이어진다. 최근 염기서열 해독 기술의 발전과 더불어 대규모 유전체 프로젝트가 진행중이다. 척추동물 유전체 프로젝트(VGP)의 경우 척추동물 6 만 6 천여종의 염기서열을 해독하는 것을 목표로 한다. 또한, 유전체 해독 시 염기서열 분석 오류와 유전체 조립을 최소화하는 고품질 표준유전체 구축을 추구한다. 최근의 Telomere to Telomere 컨소시엄, Earth Biogenome Project 등 국제 규모 유전체 프로젝트 역시 VGP 가 추구하는 고품질 표준유전체 구축의 중요성을 강조하는 등 유전체의 오류를 개선하기 위한 노력이 연구자들 사이에서 활발하다. 제 2 장에선 VGP 와의 협업을 통해 유전체 조립오류에서 발생하는 허위 복제 오류를 과거 짧은 염기서열

길이를 기반으로 구축된 유전체와 최근의 긴 길이 염기서열 분석기법을 통해 구축된 참조 유전체 내에서 발견하였다. 그 결과, 짧은 길이 기반의 염기서열 분석법에서 수천개에 달하는 허위 복제 유전자를 발견하였다. 또한 이형접합성 및 염기서열 분석 오류가 허위 복제 오류를 발생시키는 중요한 요인으로 작용 한다는 것을 확인하였으며, 이를 통해 향후 참조유전체 구축 시 허위 복제 오류를 감소시키기 위한 방향을 제시하였다. 뿐만 아니라 허위 복제가 포함된 표준유전체를 바탕으로 이루어진 연구사례를 제시하여, 허위 복제 교정의 중요성을 강조하였다. 제 3 장에선 VGP 및 Galaxy Project 와의 협업을 통해 최근 각광받는 PacBio HiFi 염기서열 분석법의 이점을 허위 복제 및 손실 두가지 측면에서 PacBio CLR 염기서열 분석방법과 비교하였다. 제 4 장에선 허위 복제 교정 프로그램을 새롭게 개발하였으며, 가상의 유전체를 생산을 통해 기존 허위 복제 교정 프로그램과의 성능을 비교하였다. 새롭게 개발된 프로그램 Purge mers 는 기존의 염기서열 리드 깊이(depth)기반 분석법과 더불어 유전체상의 허위복제 및 손실 여부를 $k-mer$ 단위에서 알 수 있는 $K*$를 허위 복제 탐색에 이용한다. 그 결과, Purge mers 의 성능이 기존의 프로그램보다 뛰어난 몇몇 경우를 발견하였다. 제 5 장에선 유전체상의 높은 GC 비율에 의해 염기서열 리드에서 계산된 $k-mer$ 의 빈도가 적게 측정되는 편향을 보정하는 방법론을 제시하였다. 편향이 제거되지 않은 $k-mer$ 측정결과는 GC 비율이 80%이상인 유전체 지역에서 $K*$가 -1 일때의 빈도가 가장 높은 결과를 나타냈다. 반면, 이 연구에서 제시한 편향이 제거된 $k-mer$ 측정결과는 GC 비율이 80%이상인 유전체

지역에서 K^*가 0 일때의 빈도가 가장 높은 결과를 나타냈다. 앞선 연구결과들을 종합하여 정리하자면 이 연구에서는 허위 복제 오류 교정의 중요성을 강조하였으며, 최적화된 염기서열 해독 기법 및 유전체 구축 방법 제시, 프로그램 및 방법론 개발 등을 통해 표준유전체 내 허위 복제 오류 해결방법을 제안하였다.

주요어 : 허위 복제, 페이징 오류, 케이머, 유전체 조립 오류, 조립 유전체 정제, 척추동물유전체프로젝트

학번 : 2018-34934

감사의 글

박사과정 동안 많은 분들께 도움을 받았기에 무사히 학위과정을 마칠 수 있었습니다. 이 자리를 통해 부족하지만 조금이나마 감사의 마음을 전합니다.

가장 먼저 저에게 박사연구자의 길을 열어 주신 김희발 선생님께 감사드립니다. 선생님께서 저를 믿고 기다려 주시고 이끌어 주신 덕분에 무사히 연구를 마무리할 수 있었습니다. 첫 논문을 작성할 때에도, 마지막 마무리 논문을 작성할 때에도 선생님의 진심 어린 조언과 지도가 있었기에 여기까지 올 수 있었습니다. 감사합니다.

저의 박사학위 심사에 참여해주신 유경록 교수님, 정충원 교수님, 조서애 박사님, 이원석 박사님께 감사드립니다. 저의 학위논문과 향후 연구 방향에 반드시 필요한 조언을 아낌없이 받을 수 있었습니다.

어머니, 무슨 말이 더 필요 할까요. 항상 제 곁에 있어 주셔서 감사합니다. 한사람의 박사로서 저의 가치가 크지는 않을지라도 제가 해왔던, 그리고 이루었던 모든 것 들, 그리고 앞으로 이룰 모든 것 들이 어머니가 있었기에 가능했고, 의미가 있는 일일 것입니다. 사랑합니다.

저의 박사 연구와 출판에 큰 기회와 도움을 주신 Erich D. Jarvis 교수님께 감사드립니다. 또한 길을 잃고 헤맸던 저의 박사 연구에 올바른 방향을 제시해주신 이아랑 박사님께 감사드립니다. 제가 연구실에서 맡은 긴꼬리닭 프로젝트에 큰 도움을 주신 채한화 연구사님께 감사드립니다. 저에게 처음 연구자의 길을 알려주시고 지도해 주셨던 어수형 교수님께 감사드립니다.

지난 학위기간 동안 연구실에서 같이 머리를 싸맸던 동료들 덕분에 무사히

졸업할 수 있었습니다. 철이 형, 형과 같이 일했던 덕분에 여기까지 올 수 있었습니다. 주완이, 매주 화요일 새벽 2 시까지 같이 남아서 미팅했던 덕분에 그 시간을 버틸 수 있었습니다. 고맙습니다. 긴꼬리닭 프로젝트를 진행하며 믿고 의지할 수 있었던 동혁이, 봉상이, Clémentine, 동재, 박사 학위 심사를 같이 준비했던 지성이, 본문 수정과 커멘트를 준 동안이, 방법론에 조언을 해준 소영이, 그 외 학위과정에서 같이 시간을 보낸 친구들 모두 고맙습니다.