Received: 14 July 2021 | Revised: 14 January 2022 | Accepted: 15 February 2022

DOI: 10.1111/1755-0998.13602

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Extensive sequence divergence between the reference genomes of two zebrafish strains, Tuebingen and AB

Yu Deng[1,2,3] | Yuting Qian[1,3] | Minghui Meng[4] | Haifeng Jiang[1,3] | Yang Dong[5] | Chengchi Fang[1,2] | Shunping He[1,2,6,7] | Liandong Yang[1,2]

[1]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China

[2]Academy of Plateau Science and Sustainability, Qinghai Normal University, Xining, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Diggers (Wuhan) Biotechnology Co., Ltd, Wuhan, China

[5]State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming, China

[6]Institute of Deep Sea Science and Engineering, Chinese Academy of Sciences, Sanya, China

[7]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

Correspondence
Shunping He and Liandong Yang, State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China.
Emails: clad@ihb.ac.cn (S.H.); yangld@ihb.ac.cn (L.Y.)

Funding information
Strategic Priority Research Program of Chinese Academy of Sciences, Grant/Award Number: XDB31000000; National Natural Science Foundation of China, Grant/Award Number: 32170480 and 31972866; Youth Innovation Promotion Association, Chinese Academy of Sciences

Handling Editor: Shotaro Hirase

Abstract

The zebrafish (*Danio rerio*) is one of the most widely used model organisms for studying vertebrate gene function and human disease, given the 70% conserved protein-coding genes between zebrafish and human. Two of the most common laboratory zebrafish strains are Tuebingen and AB. Despite the fact that the zebrafish reference genome is derived from the Tuebingen strain, the AB strain is still widely used although a high-quality genome comparable to Tuebingen is lacking. Here, we report a 1.40-Gb representative *de novo* genome assembly of the AB strain (DrAB1), with contig N50 length of 21 Mb, by integrating Illumina short-read sequencing, Nanopore long-read sequencing and HiC-based chromatin mapping. Compared with the published zebrafish Zv11 reference genome (GRCz11), this genome assembly shows considerable improvements in both contiguity and completeness. In addition, substantial structural differences and extensive sequence divergence of unprecedented resolution have been uncovered, especially with respect to 9,029,929 single nucleotide polymorphisms, 2,376,812 InDels, 32,623 insertions, 22,089 deletions and 220 inversions, which constitute ~2.6% of the DrAB1 genome. Many of these variants may have potential functional effects on phenotype, which should be considered in further experimental designs. Consequently, our study provides additional genomic resources and a high-resolution structural variation map based on whole-genome alignment for the zebrafish community, which could also be an indispensable reference genome from a model species in future research on fish phylogenetic genomics, comparative genomics and adaptive evolution.

KEYWORDS
AB strain, reference genome, structural variation, zebrafish

1 | INTRODUCTION

The zebrafish, a small, tropical, fresh-water fish, serves as a model organism for studying vertebrate development (Driever et al., 1994), genomics (Howe et al., 2013), physiology (Briggs, 2002), toxicology (Hill et al., 2005), behaviour (Deakin et al., 2019) and human pathology (Roscioli et al., 2012). Classical laboratory strains of zebrafish are commonly referred to several inbred strains such as AB (ZFIN ID: ZDB-GENO-960809–7), Tuebingen (TU; ZFIN ID: ZDB-GENO-990623–3) and WIK (ZFIN ID: ZDB-GENO-010531–2), which originated from distinct genetic backgrounds and are considered to represent a rich source of genetic and phenotypic diversity (Brown et al., 2012). Despite existing differences, few studies have incorporated the differences between strains into study design and the actual strain used is often unreported, which could potentially influence application of the results and reliable translation to human diseases in the future. Currently, the high-quality reference genome of zebrafish is from the Tuebingen strain while other high-quality whole genomes of commonly used strains such as AB and WIK are lacking, which to some extent hinders the detection and characterization of structural variations (SVs) among the strains. Previous studies have shown that these commonly strains harbour extensive genetic diversity and substructuring using array-based comparative genomic hybridization (aCGH) (Brown et al., 2012; Holden et al., 2018). However, reliance on single sequencing methods might not generate a high-quality genome showing excellence in both contiguity and accuracy due to the shortcomings of the different methods, which could decrease detection sensitivity and underestimate a huge amount of SVs such as large deletions or inversions (He et al., 2019; Ouzhuluobu et al., 2020). It is therefore crucial to take advantage of multiplatform sequencing methods to assemble a high-quality genome to comprehensively describe the SVs among the commonly used zebrafish strains.

SVs, an important class of genetic variation, are commonly defined as regions of DNA >50 bases in length exhibiting differences in genome order or DNA content between individuals, and comprise balanced SVs (inversion, insertion, translocation, fusion/fission) and unbalanced SVs (copy number variations [CNVs], polyploidy) (Alkan et al., 2011; Mérot et al., 2020; Sudmant et al., 2015). Previous studies have shown that SVs play an important role in phenotypic variation (Alonge et al., 2020), adaptive evolution (Ouzhuluobu et al., 2020), domestication (Kou et al., 2020), premating isolation (Weissensteiner et al., 2020), speciation (Berdan et al., 2019; Wellenreuther & Bernatchez, 2018) and disease (Cretu Stancu et al., 2017; Feuk et al., 2006). However, SVs remain largely unexplored despite their functional importance and the exponentially increasing accumulation of public genomic data, mainly due to the poor contiguity (fragmentation) and incompleteness (many gaps) of the most current genome assemblies. Addressing this problem first requires a high-quality genome to increase sensitivity for the identification and sequence resolution of SVs, which usually requires a combination of short-read sequencing, long-read sequencing (Nanopore and Pacbio) and multiplatform scaffolding strategies (HiC, Bionano and 10X Genomics) (Chaisson et al., 2015; Gordon et al., 2016; Kronenberg et al., 2018).

In the present study, we present a high-quality genome assembly of the AB strain through the integration of Illumina short-read sequencing, Nanopore long-read sequencing and HiC-based chromatin mapping, which generates an additional reference genome of zebrafish and provides a platform for further comparison of intraspecific genome diversity in diverse zebrafish strains. By comparing DrAB1 and GRCz11 genomes, we discovered a total of 9,029,929 single nucleotide polymorphisms (SNPs), 2,376,811 insertion/deletion polymorphisms (InDels) and 54,932 SVs (36 Mb) in DrAB1, which constitute 2.6% of the whole genome and has the potential to affect derived phenotypes. The sequence divergence observed here first indicated the extensive intraspecific genetic variations between the two zebrafish strains by whole-genome comparison, which could provide a reference to incorporate these genetic variations into future experimental designs. Furthermore, these genome data will also provide an indispensable reference genome from a model species for further fish phylogenetic genomics, comparative genomics and adaptive evolution studies, to identify candidate genes of specialized phenotype and functional verification for those species whose experimental operation is not feasible in the laboratory.

2 | MATERIALS AND METHODS

2.1 | Ethics statement, sampling and sequencing

All animal experimental procedures were approved by the ethics committee of the Institute of Hydrobiology, Chinese Academy of Sciences. For sampling, an adult healthy zebrafish (*Danio rerio*) was collected from China Zebrafish Resource Data in Wuhan, Hubei Province. High-quality genomic DNA was extracted from muscle tissues after removal of the skin by using Qiagen Blood & Cell culture DNA Mini Kit. To assess the complexity of the genome and to correct the genome assembly, a 150-bp paired-end library with insert sizes of 250 bp was constructed and sequenced on the HiSeq X-Ten platform (Illumina). For long-read DNA sequencing data, Nanopore libraries were generated and sequenced on 13 flow-cells using the GridION X5 DNA sequencer (Oxford Nanopore). To assist genome assembly and obtain a chromosome-level genome, Hi-C libraries were constructed and sequenced using the HiSeq X-Ten platform to obtain 150-bp paired-end reads.

2.2 | *De novo* genome assembly

For the Illumina whole genome sequencing reads and Hi-C reads, adaptor sequences and PCR duplicates of the paired-end reads were first removed. Then, any of the reads having more than 30 low-quality bases or more than 5% unknown bases were removed using TRIMMOMATIC (Bolger et al., 2014). The low-quality Hi-C reads were filtered by HIC-PRO software (Servant et al., 2015). Next, NEXTDENOVO

(version 2.3) (https://github.com/Nextomics/NextDenovo) was used to correct the Nanopore long reads and assemble the genome, with default parameters. We then polished the initial draft assembly for three rounds using the raw Nanopore reads and three rounds using Illumina reads with NEXTPOLISH (version 1.01) (Hu et al., 2020). To further obtain a chromosomal-level genome, the BOWTIE2 (version 2.2.5) end-to-end algorithm (Langmead et al., 2009) was adopted to align the filtered Hi-C reads to the assembled genome, and the assembled contigs were then uniquely mapped onto the scaffolds, which were further ordered and directed onto the 25 pseudochromosomes using LACHESIS (Burton et al., 2013) with default parameters. Subsequently, the 25 predicted pseudochromosomes were cut into bins with equal length of 250 kb and used to generate a heatmap based on the interaction signals. To assess the quality of our *de novo* genome assembly, we then mapped the Illumina short reads to the newly assembled genome using BWA (version 0.7.10-r789) (Li & Durbin, 2009) to produce mapping ratio and genome coverage statistics. Furthermore, the completeness of the genome was evaluated by BUSCO (Benchmarking Universal Single-Copy Orthologs, version 3) analysis (Simao et al., 2015).

2.3 | Collinearity and gaps in DrAB1 and GRCz11

First, the two assemblies were compared via MUMMER (version 4.x) with default parameters (Kurtz et al., 2004). The syntenic regions between the two genomes were obtained by filtering those regions with length >5 kb and identity >90%. Subsequently, gaps in GRCz11 were defined as regions consisting of continuous runs of N's. The gaps in DrAB1 were the N's after the scaffolds had been ordered and directed onto the 25 pseudochromosomes.

2.4 | Genome annotation

The genome was searched for repetitive elements using both *de novo* and homology-based methods. For *de novo* predictions, a *de novo* transposable element library was generated by REPEATMODELER (version 1.0.5) and then used to predict repeats using REPEATMASKER (version 4.0.6) (Tarailo-Graovac & Chen, 2009). For homology-based analysis, REPEATMASKER (version 4.0.6) and REPEATPROTEINMASK (version 1.36) with the RepBase transposable element library (Jurka et al., 2005) were used to annotate transposable elements. Finally, TRF software (version 4.0.4) (Benson, 1999) was used to identify tandem repeats.

For protein-coding gene annotation, homology-based gene predictions, *ab initio* predictions and RNA sequencing (RNAseq)-based methods were used. For homology-based gene predictions, zebrafish, blunt snout bream, catfish and tilapia protein sequences were downloaded from ENSEMBL (release 94) (Cunningham et al., 2019) and giant stone loach protein sequences were aligned to our newly assembled genome using TBLASTN (version 2.2.26) with a cut-off e-value of 1e−5. Homologous genome sequences were then aligned against the matching proteins using GENEWISE (version 2.4.1) (Birney et al., 2004) to predict the potential gene structures on all alignments. For *ab initio* prediction, AUGUSTUS (version 3.2.1) (Stanke et al., 2006), GENEID (version 1.4) (Alioto et al., 2013), GLIMMERHMM (version 3.0.4) (Majoros et al., 2004) and SNAP (version 2013-11-29) (Korf, 2004) were employed to predict coding genes using appropriate parameters with internal gene models. For the RNAseq-based method, full-length transcripts downloaded from the SRA database (accession nos.: DRR051059, SRR8944755, SRR10199497, SRR10199501, SRR10199505, SRR10040121, SRR9849265, SRR11196606, SRR9849219, SRR9849245) were mapped to the genome sequences using BLAT (version 34) (Kent, 2002). PASA (version 2.0.2) was used to filter overlaps and predict the transcript structures. EVidenceModeler (EVM, version 1.1.1) (Haas et al., 2008) was used to combine the three gene sets to yield a comprehensive and nonredundant gene set, which was further translated into amino acid sequences and for gene functional annotations by aligning against known databases including the Non-Redundant Protein Sequence Database (NR), Gene Ontology (GO), InterPro and Kyoto Encyclopedia of Genes and Genomes (KEGG), using BLASTP (version 2.2.26) with a cutoff E value of 1e-5.

2.5 | Detection of SNPs, InDels and SVs

SNPs and InDels (length <50 bp) between DrAB1 and GRCz11 were identified according to the following steps. First, the reads of clean data of the DrAB1 genome were aligned to GRCz11. Alignment files were then sorted and indexed with SAMTOOLS (version 1.9) (Kurtz et al., 2004). Variants calling was conducted using GATK-4.0 (McKenna et al., 2010) (https://gatk.broadinstitute.org) with default parameters and variants were filtered by VCFTOOLS (version 0.1.13) (http://vcftools.sourceforge.net) with parameters "--min-alleles 2 --max-alleles 2 --min-meanDP 5 --maf 0.05 --max-missing 0.5." The filtered mutations were annotated by using ANNOVAR (Wang et al., 2010). Subsequently, SVs were detected according to the following steps. For long-read Nanopore data, we used the mapping software LASTAL (version 984) (Kielbasa et al., 2011) to align the error-corrected reads ("preads") from NextDenovo output to GRCz11. We then used SNIFFLES (version 1.0.12) (Sedlazeck et al., 2018) to call SVs from the bam file and we required each variant to have support from at least 10 reads. Furthermore, the two assemblies were compared by MUMMER with default parameters and SVs between the DrAB1 genome and the GRCz11 genome were called by SMARTIE. We further merged the algorithm results by using BEDTOOLS (version 2.30.0) (Quinlan & Hall, 2010).

3 | RESULTS

3.1 | *De novo* assembly of the AB strain genome

A healthy 4-month-old adult zebrafish (*Danio rerio*) was adopted and the total DNA was extracted from muscle tissue for genome sequencing and assembly. The genome of the AB strain

was sequenced by integration of Illumina short-read sequencing, Nanopore long-read sequencing and HiC-based chromatin map (Table S1). First, a genome survey was performed using a total of 414.1 Gb Illumina HiSeq X-Ten reads and the estimated genome size was about 1.35 Gb based on 17-mer analysis (Table S2). Second, we assembled the genome with the ~272 Gb of Nanopore data with contig N50 of 21 Mb, which yielded a reference genome size of 1.40 Gb, which is different from the assembly approaches used for the Tuebingen strain genome (Figures S1-S3 and Table S3). Finally, the 1.40-Gb assembly was anchored and oriented onto 25 chromosomes with the aid of the Hi-C data, with a mounting rate of up to 99.88% (Figure S4 and Table S4). Similar to the published zebrafish reference genome, a total of 55.1% repeat elements (Table S5) and 26,195 protein-coding genes (Table S6) were identified. Additionally, the Illumina short reads were mapped to the final genome assembly to assess the completeness of the final assembly, which revealed that 97.28% of the reads could be properly paired-mapped (Table S7). Furthermore, BUSCO analysis indicated ~95.5% and 96.6% of single-copy actinopterygii genes were detected in our assembly and gene set, respectively (Table S8). In total, a *de novo* assembly of the AB strain genome, named DrAB1, represents 1.40 Gb of the chromosomes with contig N50 length of ~21 Mb and scaffold N50 length of ~58 Mb (Table 1). The two genomes show high collinearity across all major chromosomes, making large-scale misassembly unlikely (Figure 1). Subsequently, the remaining gaps in the assembly version were detected (Figure 1). Compared with the published GRCz11, DrAB1 represents an improvement and supplement, which shows 15-fold higher sequence contiguity (contig N50) and less fragmentation (165 vs. 19,725 sequence contigs, >99% reduction in total contig numbers) (Figure 1 and Table 1). In addition, compared with the currently available AB strain assemblies (e.g. GCA_008692375.1, GCA_903684865.1, GCA_903798185.1), this DrAB1 also shows superior contiguity (Table S9).

3.2 | Global comparisons between AB and Tuebigen strain genomes

As generation of this high-quality assembled genome for the AB strain zebrafish provides an opportunity to unravel the genomic differences and commonalities between representatives of the two commonly used zebrafish strains, we compared this DrAB1 genome with the GRCz11 reference genome and identified SNPs, small (length, <50 bp) InDels and SVs (insertions, deletions and inversions). Between the DrAB1 genome and GRCz11 reference genome, the overall collinearities are largely conserved (Figure 1). Notably, we found that chromosome 4 in the DrAB1 genome was ~32.8 Mb shorter than in the GRCz11 reference genome. Remarkably, 32.8 Mb of deleted regions of the long arm on chromosome 4 in the DrAB1 assembly consisted of nearly 25 Mb repeat elements and more than 600 coding sequences were observed, of which the largest of these families mainly correspond to zinc-finger proteins and exhibited enrichment of nucleic acid binding among the GO terms (Table S10). These deleted genomic regions showed a higher proportion of interspersed elements than tandem repeats. In addition, the previously reported remarkable enrichment of hundreds of immune genes and a very high density of small nuclear RNAs (snRNAs) on the long arm of chromosome 4 were not found in the deleted regions. Nevertheless, chromosome 4 in the DrAB1 genome exhibits similar length when compared with other currently available AB strain assemblies (Table S9). Except for the 725.34 Mb (52.8%) of the AB strain matching in syntenic blocks with 713.54 Mb (51.0%) of the Tuebigen strain, the genome-wide alignment between AB and TU also indicated extensive intraspecific genetic variations (Table S11). More specifically, a total of 9,029,929 SNPs were identified between the two genomes with density of 6.45 SNPs per kilobase (Table S11). Additionally, SNPs in these intrasubspecific comparisons revealed that C → T and G → A transitions (Tss) were the most abundant whereas C → G and G → C transversions (Tvs) were the least abundant, with a Ts/Tv ratio of about 1.14 (Table

TABLE 1 Comparison of genome assembly and annotation between the two zebrafish strains, Tuebingen and AB

Assembly	DrAB1	GRCz11
Assembly approach	Illumina, Nanopore, HiC	Cloned library, WGS shotgun
Number of contigs[a]	165	19,725
Contig N50 (bp)	21,088,021	1,422,317
Scaffold N50 (bp)	57,786,111	7,379,053
Gaps[b]	134	925
Total ungaped length (bp)	1,399,632,033	1,368,765,506
Total bases (bp)	1,399,645,233	1,373,454,788
GC content	36.7%	36.6%
Percentage of repeat sequences	55.1%	57.8%
Number of protein-coding genes	26,195	26,522

Note: The assembly statistics of GRCz11 were obtained from NCBI (accession no.: GCA_000002035.4).

[a]Hi-C-corrected contigs.

[b]Only N-base regions in the assembled chromosomes were counted.

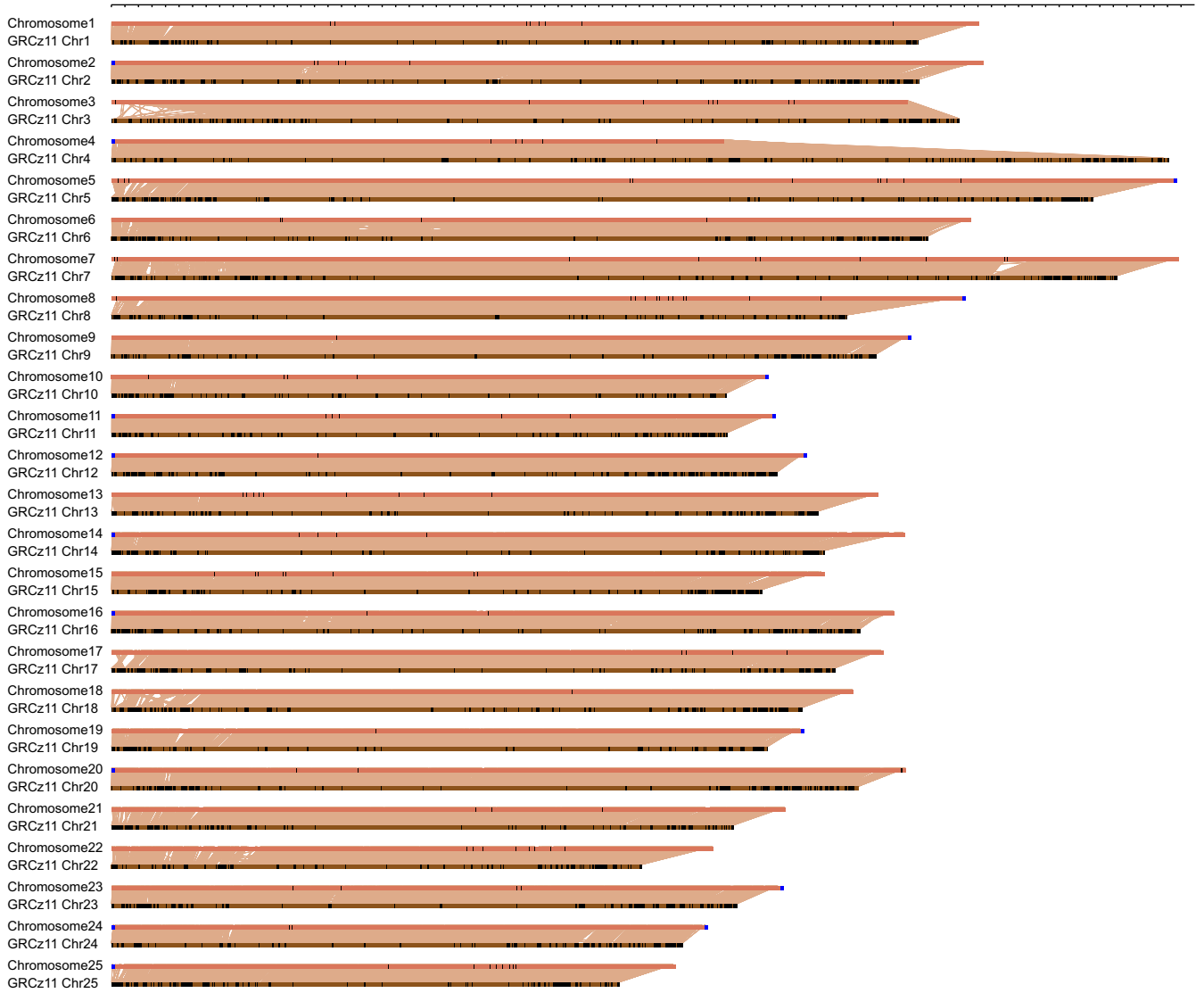

FIGURE 1 Long-read assembly of the DrAB1 genome compared with GRCz11. Twenty-five chromosomes of DrAB1 and GRCz11 were respectively aligned and represented in orange and brown, with gaps and telomeres labelled in black and blue

S12). Furthermore, we also identified 2,376,811 InDels between the two genomes with an average of 1.70 per kilobase (Table S11). These SNPs and InDels showed uneven distributions along the chromosomes (Figure 2). In addition, 54,932 AB-specific genomic SVs including insertions, deletions and inversions were present across the whole genome and accounted for 10.5, 20.5 and 5 Mb in the DrAB1 genome, respectively (Table S13). As with the distributions of SNPs and InDels, the SVs were irregularly present along the chromosomes in the genome (Figure 2). All of the comparisons above indicated that variations are abundant between the two zebrafish strains.

3.3 | Extensive SVs between the AB and Tuebigen strain

Given the increased sensitivity in detection of SVs affecting gene expression and phenotype following improvements in genome sequencing strategies, the putatively intraspecific SVs were investigated by mapping both long reads against GRCz11 and comparing the whole genome with that of GRCz11, for which a total of 54,932 SVs including 22,089 deletions, 32,623 insertions and 220 inversions overlapped and were identified in this DrAB1 assembly (Figure 3a,b; Table S13). In total, SVs identified in the DrAB1 assembly cover 36 Mb, collectively impacting 2.6% of sequence in the entire reference genome. Median lengths and total length for deletions and insertions were 156 bp, 111 bp, 20.51 Mb and 10.46 Mb, respectively (Figure 3c). Furthermore, the distribution of these SV events shows no chromosome preference and spread over all the chromosomes (Figure 3a). We also analysed the intersection of SVs (deletions and insertions) with exons, introns, untranslated regions (UTRs) and intergenic regions to speculate on their functions. The results indicated that 0.2% (24 insertions and 112 deletions) partially or entirely overlapped with exons, three insertions and one deletion with splicing regions, 50% (16,219 insertions and 10,139 deletions)

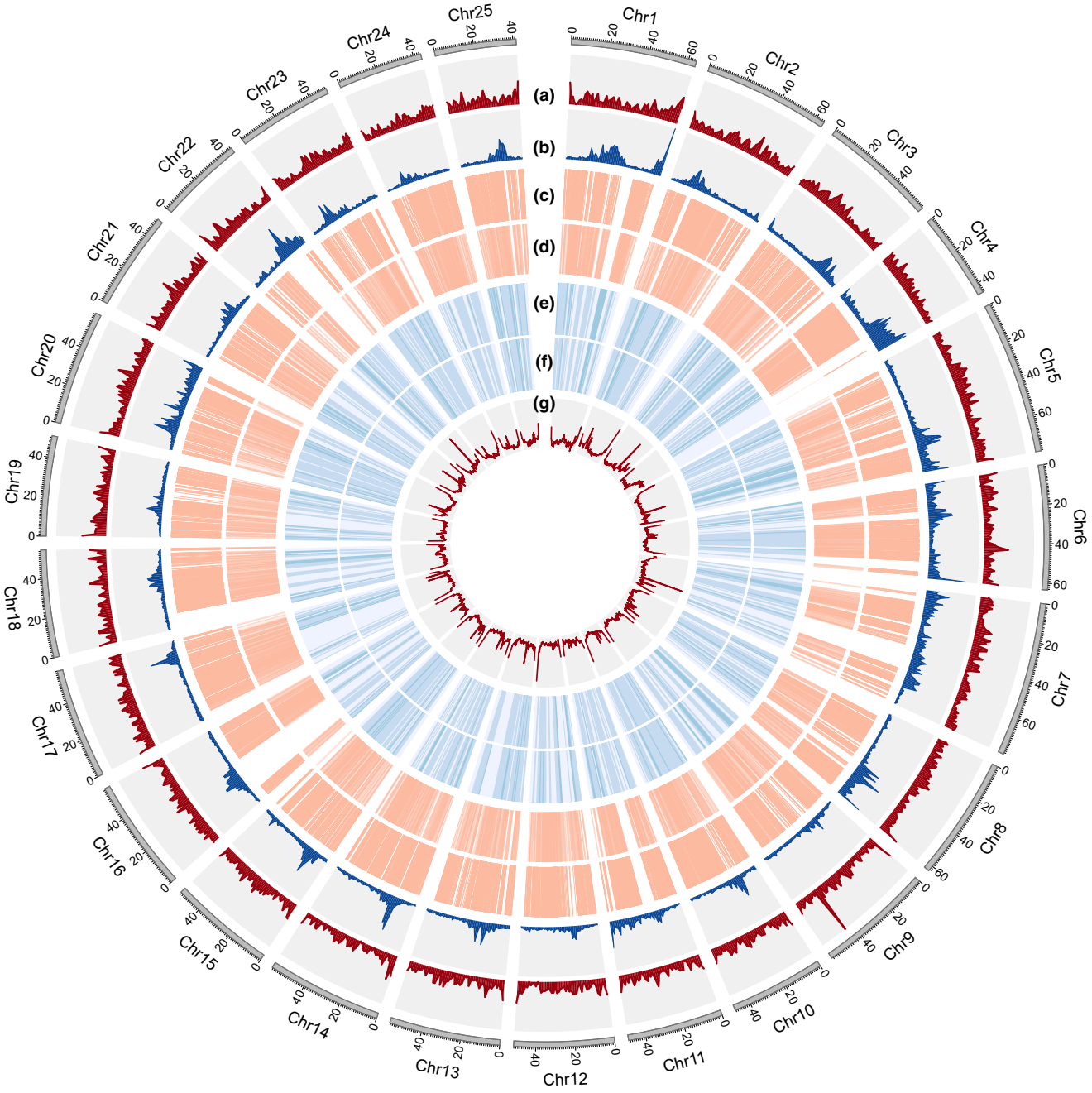

FIGURE 2 Whole-genome comparison between DrAB1 and GRCz11. Circos diagram showing (outer to inner): (a) gene density (genes per 1-Mb window); (b) repeat density (repetitive sequence per 1-Mb window); (c) the numbers and distribution of deletions in sliding windows of 1 Mb across each chromosome; (d) the numbers and distribution of insertions in sliding windows of 1-Mb across each chromosome; (e) the numbers and distribution of InDels in sliding windows of 1 Mb across each chromosome; (f) the numbers and distribution of SNPs in sliding windows of 1 Mb across each chromosome; and (g) the GC content in sliding windows of 1 Mb across each chromosome

with introns, 2.8% (959 insertions and 559 deletions) with UTRs and the remaining 46.2% (15,418 insertions and 9,970 deletions) in intergenic regions (Figure 3d). Among the coding sequences affected by the candidate SVs, some genes were predicted or reported to be expressed in the nervous system or be associated with brain development and photoreceptor cell development (Table S14). Given the high repeat content in zebrafish, SVs involved in repeat sequences were also classified into five categories including DNA transposons, SINEs, LINEs, LTRs and "Other" (Figure 3e). Annotation of these SVs shows that 71.0% of deletions and 53.6% of insertions were larger than 100 bp. Furthermore, 220 inversions (~10.5 Mb in total) ranging from 0.2 to 454 kb were also detected by comparing the two genomes and many inversions were predicted to affect many genes involved in eye development and nervous system development. Overall, these genome-wide comparisons demonstrated extensive intraspecific genome diversity in different zebrafish strains.

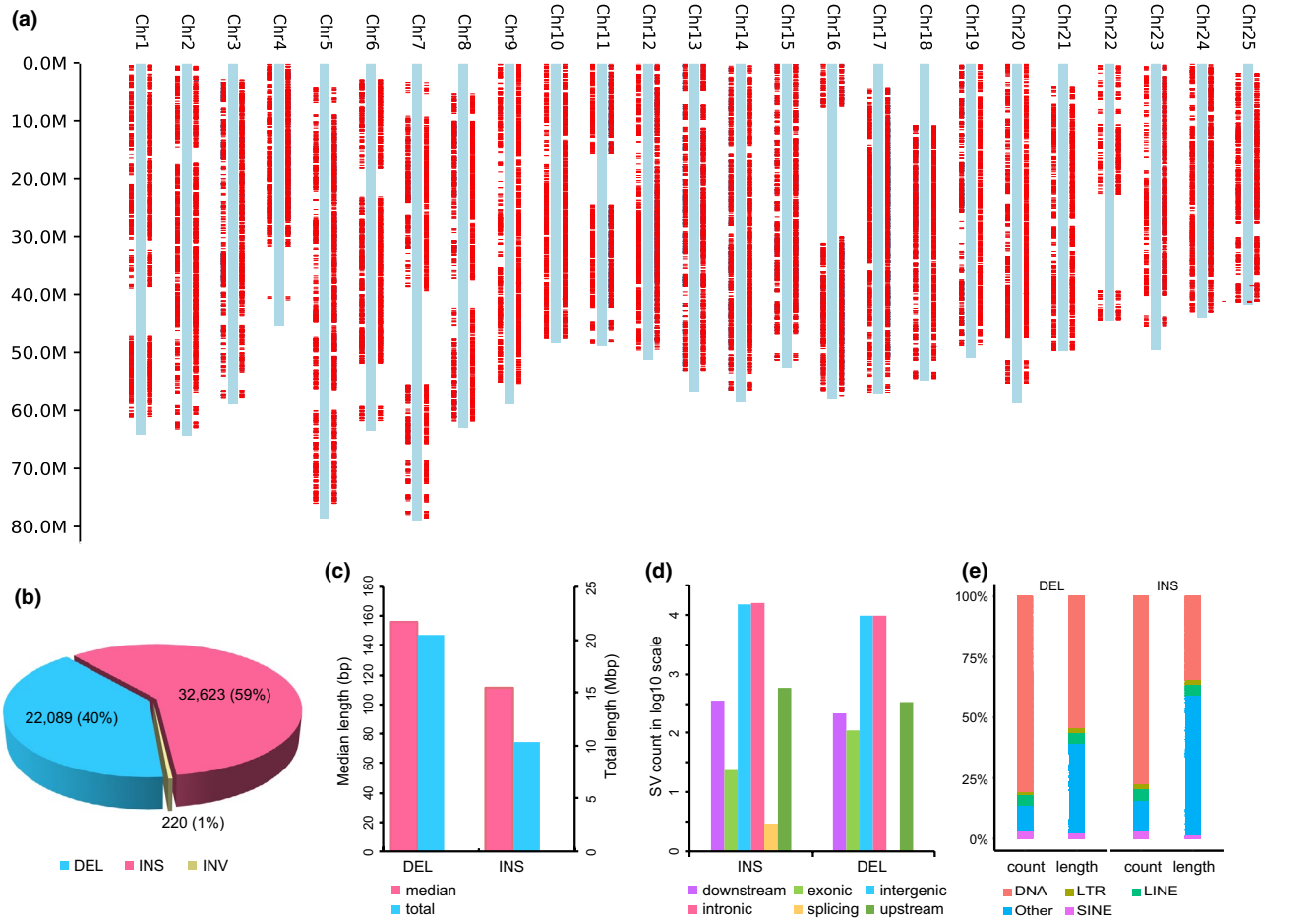

FIGURE 3 Structural variations in the DrAB1 genome compared with GRCz11. (a) The distribution of deletions (DELs) (left) and insertions (INSs) (right) (>500 bp) across the 25 chromosomes in DrAB1. (b) The proportions of DELs, INSs and inversions (INVs) detected in DrAB1. (c) The length statistics of the DELs and INSs in DrAB1. (d) The numbers of SVs assigned to exons, introns and UTRs. (e) The distribution of SVs assigned to repeat sequences

4 | DISCUSSION

The zebrafish serves as an indispensable species for understanding vertebrate development (Driever et al., 1994), genomics (Howe et al., 2013), physiology (Briggs, 2002), toxicology (Hill et al., 2005) and behaviour (Deakin et al., 2019), and represents a transformative resource for human genetics and disease (Roscioli et al., 2012). Up to now, a high-resolution SV map based on whole-genome alignment remained unexplored for commonly used zebrafish strains although differences between the two strains exist. At the same time, integration of Nanopore sequencing with additional technologies such as optical mapping or Hi-C sequencing is becoming an increasingly popular approach to improve assembly contiguity for further genomic research (Li et al., 2020; Weissensteiner et al., 2020). Here, we *de novo* assembled a high-quality genome assembly of the AB strain zebrafish (DrAB1) by integrating Illumina short-read sequencing, Nanopore long-read sequencing and Hi-C technologies, which is thus more complete and will enable new zebrafish basic research.

Compared with the published GRCz11, the DrAB1 assembly exhibited substantially improved quality with longer contig and scaffold N50 sizes and with a total size of 1.40 Gb. In the DrAB1 genome, 55.1% repeat content and 26,195 annotated protein-coding genes were comparable to those for GRCz11. Notably, 32.8 Mb of deleted regions of the long arm on chromosome 4 in the DrAB1 assembly mainly consisted of repeat elements and zinc-finger proteins. Given the wide range of molecular functions of zinc-finger proteins (Cassandri et al., 2017), these data provide a reference for zebrafish strains in the future. Further direct comparison between the two zebrafish genomes revealed extensive genetic diversity including SNPs, InDels, as well as a large extent of SVs that encompasses 2.6% (36 Mb) of the zebrafish genome. Our data first elucidated 54,932 SVs previously unknown in the two zebrafish strain genomes according to whole-genome comparison, which probably resulted from diverse genetic backgrounds (Mullins et al., 1994). These SVs may play vital roles in some lineage-specific functions and should be taken into consideration in translation of future experimental results. Specifically, many genes affected by these identified SVs are expressed in the nervous system or are associated with brain and eye development, which could provide an explanation for the behavioural differences in assays of a light/dark challenge and

contrasted circadian rhythms (Vignet et al., 2013). In addition, SVs partially or completely overlapping 5′ or 3′ UTRs could potentially affect gene expression.

Although two different methods were used to detect SVs based on both read-mapping and whole-genome alignment approaches and the SVs obtained were compared from different data set to make the results more reliable, the accurate calling of SVs still depends on, for example, sequencing methods, assembly strategies and detection algorithms (Kosugi et al., 2019). Thus, sequencing methods and assembly strategies provide the essential basis for the identification of SVs, which could have direct impacts on the completeness, contiguity and base-pair accuracy of the resulting genome sequence due to choices regarding read type and sequencing depth (Minoche et al., 2011; Rhoads & Au, 2015; Sedlazeck et al., 2018). A previous study reported the effects of different genome assemblies built from the same data on gene annotation and SNP annotations, which showed the importance of a high-quality genome assembly and consensus in the data inputs for downstream analysis (Florea et al., 2011). Another study on acute respiratory syndrome coronavirus 2 (SARS-CoV-2) also indicated that the choice of assemblers plays a significant role in the detection of variants and a number of variants present in assemblies are unique to the assembly methods (Islam et al., 2021). Hence, as sequencing methods are diversifying and increasingly new assembly algorithms are being developed, future genome downstream analysis such as detection of SVs should take the same data input into consideration to eliminate differences resulting from the data and assembly methods.

High-quality genome assemblies are a prerequisite for the identification and characterization of extensive intraspecific genome diversity. Although genetic variations can arise at both the level of individuals and populations, the extensive genetic variations found in our work suggest that the two different zebrafish strains may harbour a dramatically different complement of proteins and regulatory sequences, and strain-specific genetic variations should be taken into consideration when designing experiments for confounding studies for data applications and translations.

ACKNOWLEDGEMENTS

This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB31000000), the National Natural Science Foundation of China (32170480 and 31972866), and Youth Innovation Promotion Association, Chinese Academy of Sciences (http://www.yicas.cn). This research was supported by the Wuhan Branch, Supercomputing Center, Chinese Academy of Sciences, China.

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

S.H. led the project. L.Y. and Y.D designed the study. L.Y., Y.D., Y.Q., M.M. and C.F., performed the bioinformatic analysis. L.Y. and H.J. collected the samples. Y.D. analysed the results and wrote the manuscript with inputs from the other authors. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The raw sequencing data (Nanopore, Illumina and Hi-C) have been deposited in the NCBI BioProject database with accession no. PRJNA734711. The genome assembly file is under accession JAHVXS000000000. The annotation files of the zebrafish genome are deposited in github (https://github.com/dana0201/zebrafish-AB-genome-annotation).

ORCID

Yu Deng https://orcid.org/0000-0002-6354-6647
Yang Dong https://orcid.org/0000-0001-6212-3055
Liandong Yang https://orcid.org/0000-0001-7570-0296

REFERENCES

Alioto, T., Picardi, E., Guigo, R., & Pesole, G. (2013). ASPic-GeneID: A lightweight pipeline for gene prediction and alternative isoforms detection. *BioMed Research International*, *2013*, 502827, https://doi.org/10.1155/2013/502827

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., … Lippman, Z. B. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*(1), 145–161.e123. https://doi.org/10.1016/j.cell.2020.05.021

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Berdan, E. L., Blanckaert, A., Butlin, R. K., & Bank, C. (2019). Muller's ratchet and the long-term fate of chromosomal inversions. *bioRxiv*. 606012. https://doi.org/10.1101/606012

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research*, *14*(5), 988–995. https://doi.org/10.1101/gr.1865504

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Briggs, J. P. (2002). The zebrafish: A new model organism for integrative physiology. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, *282*(1), R3–9. https://doi.org/10.1152/ajpregu.00589.2001

Brown, K. H., Dobrinski, K. P., Lee, A. S., Gokcumen, O., Mills, R. E., Shi, X., Chong, W. W. S., Chen, J. Y. H., Yoo, P., David, S., Peterson, S. M., Raj, T., Choy, K. W., Stranger, B. E., Williamson, R. E., Zon, L. I., Freeman, J. L., & Lee, C. (2012). Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(2), 529–534. https://doi.org/10.1073/pnas.1112163109

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R. L., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), 1119–1125. https://doi.org/10.1038/Nbt.2727

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., & Raschellà, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discov*, *3*, 17071. https://doi.org/10.1038/cddiscovery.2017.71

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., Korzelius, J., de Bruijn, E., Cuppen, E., Talkowski, M. E., Marschall, T., de Ridder, J., & Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*, *8*(1), 1326. https://doi.org/10.1038/s41467-017-01343-4

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113

Deakin, A. G., Buckley, J., AlZu'bi, H. S., Cossins, A. R., Spencer, J. W., Al'Nuaimy, W., Young, I. S., Thomson, J. S., & Sneddon, L. U. (2019). Automated monitoring of behaviour in zebrafish after invasive procedures. *Scientific Reports*, *9*. Artn 9042. https://doi.org/10.1038/S41598-019-45464-W

Driever, W., Stemple, D., Schier, A., & Solnica-Krezel, L. (1994). Zebrafish: Genetic tools for studying vertebrate development. *Trends in Genetics*, *10*(5), 152–159. https://doi.org/10.1016/0168-9525(94)90091-4

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85–97. https://doi.org/10.1038/nrg1767

Florea, L., Souvorov, A., Kalbfleisch, T. S., & Salzberg, S. L. (2011). Genome assembly has a major impact on gene content: A comparison of annotation in two bos taurus assemblies. *PLoS One*, *6*(6), e21400. ARTN e21400. https://doi.org/10.1371/journal.pone.0021400

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., & Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, *352*(6281), aae0344. https://doi.org/10.1126/science.aae0344

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

He, Y., Luo, X., Zhou, B., Hu, T., Meng, X., Audano, P. A., Kronenberg, Z. N., Eichler, E. E., Jin, J., Guo, Y., Yang, Y., Qi, X., & Su, B. (2019). Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nature Communications*, *10*(1), 4233. https://doi.org/10.1038/s41467-019-12174-w

Hill, A. J., Teraoka, H., Heideman, W., & Peterson, R. E. (2005). Zebrafish as a model vertebrate for investigating chemical toxicity. *Toxicological Sciences*, *86*(1), 6–19. https://doi.org/10.1093/toxsci/kfi110

Holden, L. A., Wilson, C., Heineman, Z., Dobrinski, K. P., & Brown, K. H. (2018). An interrogation of shared and unique copy number variants across genetically distinct zebrafish strains. *Zebrafish*, *16*(1), 29–36. https://doi.org/10.1089/zeb.2018.1644

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics (Oxford, England)*, *36*(7), 2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Islam, R., Raju, R. S., Tasnim, N., Shihab, I. H., Bhuiyan, M. A., Araf, Y., & Islam, T. (2021). Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Briefings in Bioinformatics*, *22*(5). ARTN bbab102. https://doi.org/10.1093/bib/bbab102

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, *110*(1–4), 462–467. https://doi.org/10.1159/000084979

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, Artn 59. https://doi.org/10.1186/1471-2105-5-59

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 117. https://doi.org/10.1186/s13059-019-1720-5

Kou, Y., Liao, Y., Toivainen, T., Lv, Y., Tian, X., Emerson, J. J., Gaut, B. S., & Zhou, Y. (2020). Evolutionary genomics of structural variation in Asian rice (Oryza sativa) domestication. *Molecular Biology and Evolution*, *37*(12), 3507–3524. https://doi.org/10.1093/molbev/msaa185

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., … Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science (New York, N.Y.)*, *360*(6393), eaar6343. https://doi.org/10.1126/science.aar6343

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), Artn R12. https://doi.org/10.1186/Gb-2004-5-2-R12

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Q., Ramasamy, S., Singh, P., Hagel, J. M., Dunemann, S. M., Chen, X., Chen, R., Yu, L., Tucker, J. E., Facchini, P. J., & Yeaman, S. (2020). Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. *Nature Communications*, *11*(1), 1190. https://doi.org/10.1038/s41467-020-15040-2

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*(16), 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural

genomic variation. *Trends in Ecology & Evolution*, *35*(7), 561–572. https://doi.org/10.1016/j.tree.2020.03.002

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, *12*(11), R112. https://doi.org/10.1186/gb-2011-12-11-r112

Mullins, M. C., Hammerschmidt, M., Haffter, P., & Nusslein-Volhard, C. (1994). Large-scale mutagenesis in the zebrafish: In search of genes controlling development in a vertebrate. *Current Biology*, *4*(3), 189–202. https://doi.org/10.1016/s0960-9822(00)00048-8

Ouzhuluobu, He, Y., Lou, H., Cui, C., Deng, L., Gao, Y., Zheng, W., Guo, Y., Wang, X., Ning, Z., Li, J., Li, B., Bai, C., Baimakangzhuo, Gonggalanzi, Dejiquzong, Bianba, Duojizhuoma, Liu, S., … Su, B. (2020). De novo assembly of a Tibetan genome and identification of novel structural variants associated with high-altitude adaptation. *National Science Review*, *7*(2), 391–402. https://doi.org/10.1093/nsr/nwz160

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Roscioli, T., Kamsteeg, E. J., Buysse, K., Maystadt, I., van Reeuwijk, J., van den Elzen, C., … van Bokhoven, H. (2012). Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of alpha-dystroglycan. *Nature Genetics*, *44*(5), 581–585. https://doi.org/10.1038/ng.2253

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*, 259. https://doi.org/10.1186/s13059-015-0831-x

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server), W435–W439. https://doi.org/10.1093/nar/gkl200

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … The Genomes Project, C (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1). https://doi.org/10.1002/0471250953.bi0410s25

Vignet, C., Begout, M. L., Pean, S., Lyphout, L., Leguay, D., & Cousin, X. (2013). Systematic screening of behavioral responses in two zebrafish strains. *Zebrafish*, *10*(3), 365–375. https://doi.org/10.1089/zeb.2013.0871

Wang, K., Li, M. Y., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, *11*(1), 3403. https://doi.org/10.1038/s41467-020-17195-4

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Deng, Y., Qian, Y., Meng, M., Jiang, H., Dong, Y., Fang, C., He, S., & Yang, L. (2022). Extensive sequence divergence between the reference genomes of two zebrafish strains, Tuebingen and AB. *Molecular Ecology Resources*, *00*, 1–10. https://doi.org/10.1111/1755-0998.13602