

Intercross population study reveals that co-mutation of *mitfa* genes in two subgenomes induces red skin color in common carp (*Cyprinus carpio wuyuanensis*)

Bi-Jun Li^{1,2}, Lin Chen^{1,2}, Meng-Zhen Yan^{1,2}, Xiao-Qing Zou^{1,2}, Yu-Lin Bai^{1,2}, Ya-Guo Xue³, Zhou Jiang^{1,3}, Bao-Hua Chen^{1,2}, Cheng-Yu Li^{1,2}, Qian He^{1,2}, Jian-Xin Feng⁴, Tao Zhou^{1,2}, Peng Xu^{1,2,*}

¹ State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian 361102, China

² Fujian Key Laboratory of Genetics and Breeding of Marine Organisms, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian 361102, China

³ College of Fisheries, Henan Normal University, Xinxiang, Henan 453007, China

⁴ Henan Academy of Fishery Science, Zhengzhou, Henan 450039, China

ABSTRACT

Common carp are among the oldest domesticated fish in the world. As such, there are many food and ornamental carp strains with abundant phenotypic variations due to natural and artificial selection. Hebao red carp (HB, *Cyprinus carpio wuyuanensis*), an indigenous strain in China, is renowned for its unique body morphology and reddish skin. To reveal the genetic basis underlying the distinct skin color of HB, we constructed an improved high-fidelity (HiFi) HB genome with good contiguity, completeness, and correctness. Genome structure comparison was conducted between HB and a representative wild strain, Yellow River carp (YR, *C. carpio haematopterus*), to identify structural variants and genes under positive selection. Signatures of artificial selection during domestication were identified in HB and YR populations, while phenotype mapping was performed in a segregating population generated by HB×YR crosses. Body color in HB was associated with regions with fixed mutations. The simultaneous mutation and superposition of a pair of homologous genes (*mitfa*) in chromosomes A06 and B06 conferred the reddish color in domesticated HB. Transcriptome analysis of common carp with different alleles of the *mitfa* mutation confirmed that gene duplication can buffer the deleterious effects of mutation in allotetraploids. This study provides new insights into genotype-phenotype associations in allotetraploid species and lays a foundation for future breeding of common carp.

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ©2023 Editorial Office of Zoological Research, Kunming Institute of Zoology, Chinese Academy of Sciences

Keywords: *Cyprinus carpio*; Artificial selection; Coloration; *mitfa*; Structural variant

INTRODUCTION

The variation in animals and plants under domestication inspired much of Darwin's theory of evolution (Darwin, 2010). Changes in species can occur through both domestication and evolutionary processes, but in different ways. Artificial selection often acquires phenotypes that are undesirable in nature, frequently resulting in intraspecific diversity but not reproductive isolation (Carneiro et al., 2014; López et al., 2019; Milla et al., 2021; Nam et al., 2019). Many plants and animals have been domesticated by humans. However, despite the importance of animal domestication in human history, significant questions regarding its evolutionary mechanisms remain unclear (Akagi et al., 2016; Carneiro et al., 2014; Qiu et al., 2015). The development of various techniques, such as sequencing, quantitative trait loci (QTL) mapping, and genome-wide association analysis, has facilitated the identification of genes associated with domestication (Ross-Ibarra et al., 2007), but verifying the molecular mechanisms underlying these causal genes remains difficult. Hybrid experiments, which create recombinations of hereditary material from parents, play important roles in genetic research, facilitating the discovery of Mendel's law of inheritance (Mendel, 1996) and the genetic mechanisms of phenotype formation (Andersson, 1997; Bakos & Gorda, 1995; Essa et al., 2021; Lexer et al., 2003; Zhou et al., 2018).

Fish domestication started much later than that of

Received: 11 December 2022; Accepted: 10 February 2023; Online: 11 February 2023

Foundation items: This work was supported by the National Key R&D Program of China (2019YFE0119000), National Natural Science Foundation of China (31872561), National Science Fund for Distinguished Young Scholars (32225049), and Alliance of International Science Organizations (ANSO-CR-PP-2021-03)

*Corresponding author, E-mail: xupeng77@xmu.edu.cn

mammals, by at least 10 000 years (Balon, 2004). Domestication plays an essential role in modulating the phenotypes of teleosts (Milla et al., 2021). Common carp are widely distributed in Eurasia and are among the earliest known domesticated fish, occurring in China by around 6 000 BC (Nakajima et al., 2019). During the long history of domestication, multiple strains (varieties) have emerged worldwide due to geographic isolation, adaptation, and natural and human selection (Xu et al., 2014), resulting in phenotypic changes in growth rate, temperature and hypoxia tolerance, body color, body shape, and scale pattern. The Hebao red carp (HB, *Cyprinus carpio wuyuanensis*) is a traditional domestic variety with an 800-year breeding history in Wuyuan, Jiangxi Province, China (Zhou et al., 2004). This variety is characterized by a short, plump, rounded body shape and bright red body color (Lou & Sun, 2001) (Figure 1A). Although it is a popular ornamental fish, it is also important in aquaculture production due to its taste and high protein content. Crossbreeding with HB provides the basis for the development of new strains with diverse aquacultural advantages (Hu et al., 2018). Although studies have indicated that its color is determined by two gene loci, the specific underlying mechanisms have not yet been investigated (Zhang & Pan, 1983). As a valuable germplasm resource, understanding the genetic mechanism behind the HB phenotype will help exploit its advantages and provide valuable genetic resources for further breeding.

Common carp is a typical allotetraploid species formed by hybridization between a *Barbus*-like species and an unidentified species approximately 12 million years ago (Xu et al., 2019b). Allopolyploidization increases the difficulty of genome assembly and genetic study in this species. A previous draft HB genome was constructed using Illumina short-read sequencing, but showed unsatisfactory quality (Xu et al., 2019b). However, the development of third-generation sequencing technology provides promising prospects for highly complicated genomes due to the capacity to span complex repetitive regions of the genome (Vollger et al., 2020; Wenger et al., 2019). In our study, a high-quality HB genome assembly was accomplished using high-fidelity long (HiFi) reads. Due to its close phylogenetic relationship with HB (Xu et al., 2014), the Yellow River carp (YR, *C. carpio haematopterus*) was adopted as a representative subspecies of wild common carp (Figure 1B) for selection signature detection of the HB strain. A HB×YR segregating population was constructed to reveal and characterize domestication genes behind the distinctive red body color of HB. This study will help better understand the genomic structure of the HB strain and provide a theoretical basis for genetic research of allotetraploid fish.

MATERIALS AND METHODS

Ethics statement

All animal experiments were conducted following the regulations of the Guide for the Care and Use of Laboratory Animals and were approved by the Committee of Laboratory Animal Experimentation at the College of Ocean and Earth Sciences, Xiamen University, China.

Sample collection and sequencing for genome assembly

A female *C. carpio wuyuanensis* individual was collected from the Breeding Station of the Henan Academy of Fishery Sciences, Zhengzhou, Henan, China. Muscle tissues were

frozen with liquid nitrogen, with genomic DNA then extracted, qualified, and quantified. High-quality DNA was sent to AnnoRoad (Wuhan, China) for polymerase chain reaction (PCR)-free SMRT bell library (CCS) construction and sequencing using the PacBio Sequel/Sequel II platform. To construct the chromosome-level genome assembly, frozen muscle tissues from the same fish were sent to Novogene (China) for high-throughput chromosome conformation capture technology (Hi-C) library construction and sequencing. Total RNA was isolated from nine tissues for RNA sequencing. Information on genome sequencing, assembly, scaffolding, and annotation is described in the Supplementary Methods.

Phylogenetic tree construction with single-copy homologous genes

To reveal the phylogenetic relationships between two subgenomes of *C. carpio wuyuanensis* and other Cyprinidae, protein sequences of representative Cyprinidae species were downloaded from Ensembl and NCBI (Supplementary Table S1), with *Danio rerio* taken as the outgroup. The gene set of allotetraploid teleosts (*C. carpio wuyuanensis*, *C. carpio haematopterus*, and *Carassius auratus*) was divided into two groups according to their distribution in the two subgenomes. For genes with more than one isoform, the longest transcript was chosen to represent that gene. Single-copy orthologous genes were identified using OrthoFinder (Emms & Kelly, 2019). Multiple sequence alignment was then performed using MAFFT (Katoh et al., 2009) and trimmed using Gblocks (Castresana, 2000). ProTest was used to select the best-fit models of amino acid replacement (Darriba et al., 2011). Phylogenetic analyses were conducted using RAXML (Stamatakis, 2014) and FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to draw the phylogenetic tree.

Nonsynonymous (Ka) and synonymous (Ks) substitution analysis

We conducted Ka/Ks analysis to compare the evolutionary rate between *Cyprinus carpio wuyuanensis* and *C. carpio haematopterus*. Annotated protein-coding genes of the two subspecies were used. Homologous gene pairs in the A and B subgenomes were identified by bidirectional free alignment (reciprocal best hit). The Ka and Ks substitution rates of each orthologous gene pair were calculated using ParaAT (Parallel Alignment and backTranslation) (Zhang et al., 2012), while Muscle was used to prepare input files for Ka/Ks_Calculator2.0 (Zhang, 2022). The Ks and Ka/Ks values of gene pairs were plotted using the ggplot2 R package.

Structural variant (SV) detection and validation

To detect SVs between *C. carpio wuyuanensis* and *C. carpio haematopterus*, the genome assembly of *C. carpio haematopterus* was aligned with that of *C. carpio wuyuanensis* using Mummer with parameters: -maxmatch -c 100 -b 500 -l 50 (Kurtz et al., 2004). The alignments were then filtered by delta-filter with parameters: -m -l 90 -l 100. Structural rearrangements and variants between genomes were identified using SyRI with default parameters (Goel et al., 2019). According to the output files from SyRI, we clustered SVs into three types: presence/absence variants (PAVs), inversions, and translocations. Absence variants in the YR genome included CPL, DEL, DUP/INVDP (loss), HDR, NOTAL, and TDM. Presence variants included CPG, INS, and

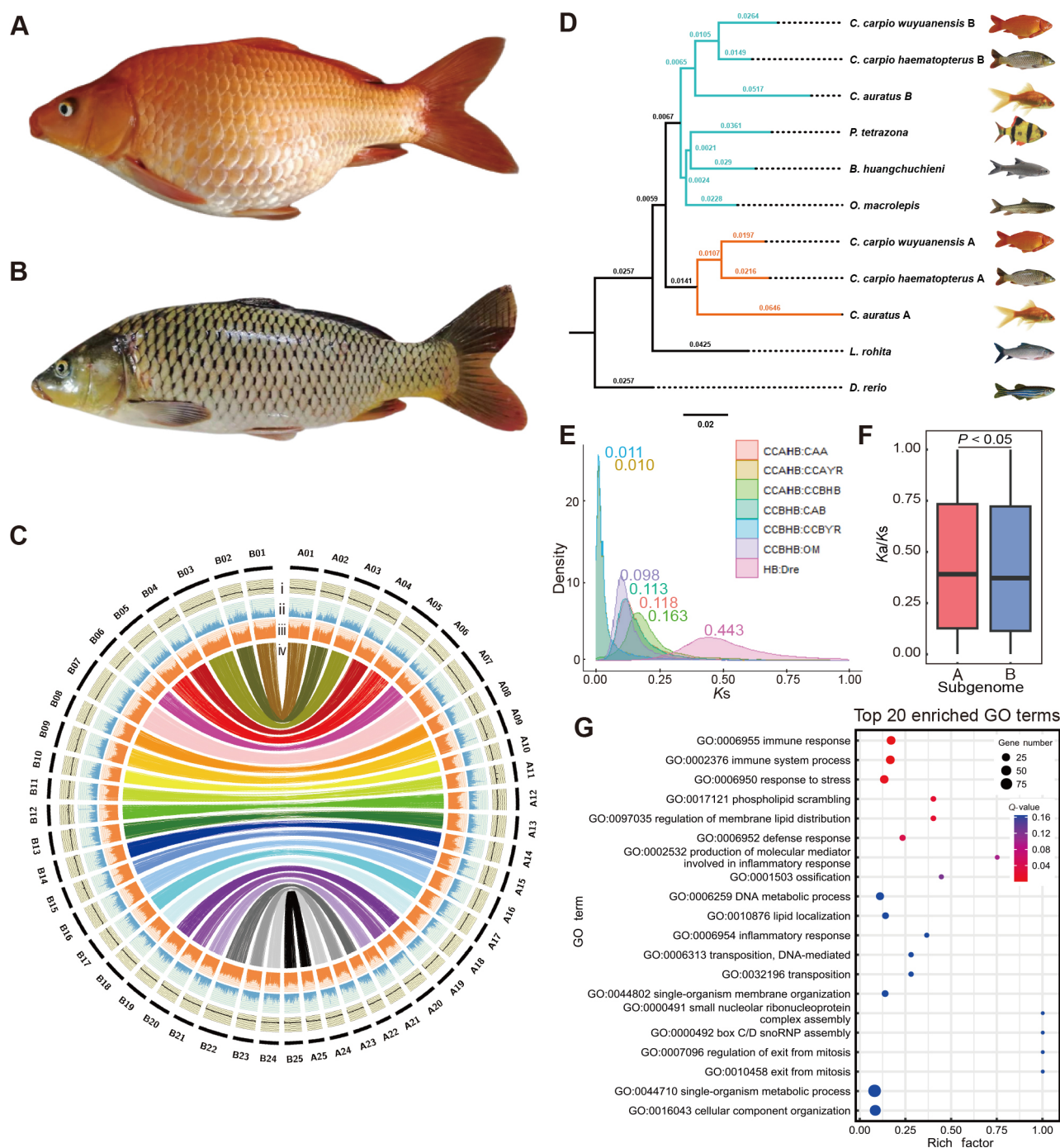


Figure 1 Genomic characteristics and *Ka/Ks* analysis of *C. carpio wuyuanensis*

A: Photo of HB (*C. carpio wuyuanensis*), taken by Bi-Jun Li. B: Photo of YR (*C. carpio haematopterus*), taken by Bi-Jun Li. C: The Genomic characteristics of *C. carpio wuyuanensis*. The 50 chromosomes are plotted and numbered on the outermost circle. The inner circles show the distribution of i) GC content, ii) gene density, iii) TE density in 100 kb sliding window and iv) the collinearity relationship of chromosomes between A and B subgenomes. D: Phylogenetic analysis between HB and representative Cyprinidae species. The tree was constructed with 1 378 shared single-copy orthologous genes. E: Distribution of *Ka* substitutions in gene pairs among *C. carpio wuyuanensis*, *C. carpio haematopterus*, *C. auratus*, *O. macrolepis*, and *D. rerio*. Allotetraploid genes were divided into A and B gene sets according to their distribution in the genome. F: *Ka/Ks* values of gene pairs between *C. carpio wuyuanensis* and *C. carpio haematopterus*. *Ka/Ks* value of genes in subgenome A (mean *Ka/Ks*=0.438) was higher than that in subgenome B (mean *Ka/Ks*=0.446). G: Biological processes related to genes under positive selection (*Ka/Ks*>1).

DUP/INVP (gain) variants, and HB sequences in HDR, NOTAL, and TDM. The functional effects of variants were annotated with snpEff. The INV variants were regarded as inversion SVs relative to YR, while TRANS and INVTR were both regarded as translocations. JCVI was applied for pairwise genome syntenic comparisons and visualizations with genes

(Tang et al., 2008).

To validate the SVs, long reads from YR (NCBI accession: PRJNA823855) were mapped to the HB genome using minimap2 and the obtained BAM files were used to display the presence regions. Long reads from HB were mapped to the YR genome to detect the absence regions. Integrative

Genomics Viewer (IGV) was used to inspect SVs in the genome (Robinson et al., 2011). For large-scale SVs, Hi-C reads from the YR genome (NCBI accession No.: PRJNA823855) were aligned to the HB assembly using Juicer. The large SVs were manually checked using heatmap in Juicebox (Durand et al., 2016).

Family construction and sample collection

Intercross segregating populations are informative for genetic analysis and may have considerable utility in QTL mapping and identification of trait loci. In our study, the experimental population was derived from a cross between HB and YR carp species. YR was selected as a wild-type source, while HB was selected as a domestic breed known for its red body color and belly bulge. The F1 generation was produced in 2017 at the Breeding Station of the Henan Academy of Fishery Research, Zhengzhou, China. HB and YR carp were crossed to generate two full-sib families. Physical characteristics of the progeny were measured at 8 and 23 months of age. Tail fins were cut and stored in absolute ethanol for further DNA extraction. The BC1F1 generation was produced from mating F1 hybrids and HB. The BC1F1 generation consisted of 3 000 carp (17 families) with body color segregation. We recorded body color and photographed each individual when they reached 6 months of age. Tail fins were cut and stored in absolute ethanol for further DNA extraction.

Whole-genome resequencing and variant calling

Eighteen HB samples and 16 YR samples were randomly selected from the natural populations and used for whole-genome sequencing for population genetic analysis. Fifty individuals were selected from one family of the BC1F1 segregating population for body color genome-wide association study (GWAS). DNA quantity and quality were examined using a Qubit fluorometer and agarose gel electrophoresis, respectively. Paired-end libraries were generated for each sample using standard procedures. The samples were sent to BGI (Qingdao, China) for genome resequencing using the MGISEQ-2000RS platform. Approximately 6 Gb of reads per individual were acquired for whole-genome genotyping. Average raw read sequence coverages of $\times 10$ and $\times 5$ were acquired for the natural and intercross populations, respectively.

Paired-end reads from each sample were trimmed and aligned to the HB reference genome using Burrows-Wheeler Aligner (BWA) with default parameters (Li & Durbin, 2010). A BAM file was generated using SAMtools and sorted by Picard (Danecek et al., 2021). Paired reads mapped to the same position on the reference genome were removed with MarkDuplicates in Picard to avoid any influence on variant detection. After mapping, variant calling was performed using HaplotypeCaller in the GATK v4.0 pipeline (Poplin et al., 2018), with the officially recommended standard applied to filter out low-quality single nucleotide polymorphism (SNPs) and insertions/deletions (indels). The output was further filtered using VCFtools (Danecek et al., 2011) to retain variants that met the following criteria: (1) minor allele count > 5 and max missing count < 2 ; (2) only two alleles. SNPs and indels across the genome were annotated using the ANNOVAR program (Wang et al., 2010).

Selective signal sweep detection and determination of candidate genomic regions

Principal component analysis (PCA) was performed based on

all SNPs using EIGENSOFT (v4.2) (Price et al., 2006). The figures were then plotted using the first and second principal components (PCs) with the ggplot2 R package (Villanueva & Chen, 2019). To detect candidate divergent regions (CDRs) between two strains, we searched the genome for regions with a high fixation index (F_{ST}) and huge difference in genetic diversity ($\log_{10}(\pi \text{ ratio})$). The π ratio and Wright's F_{ST} for each linkage group were calculated using the sliding window method in VCFtools (Danecek et al., 2011). Window width was set to 40 kb and stepwise distance was set to 10 kb. CDRs were restricted to windows with a top 0.5% value for both F_{ST} and π ratio. Finally, genes in the CDRs between HB and YR were identified. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment analyses were performed using OmicShare tools (<https://www.omicshare.com/tools>).

GWAS of body color and shape in hybrid family

Variant calling was conducted following the above GATK pipeline (Poplin et al., 2018). To preserve high-quality SNPs for downstream analysis, we filtered putative false SNPs using VariantFiltration with the parameters “QD < 2.0 || QUAL < 30.0 || SOR > 3.0 || FS > 20.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 ”. The VCF files were then filtered with the parameters “-max-alleles 2 -min-alleles 2 -maf 0.05 -max-missing 0.8” using VCFtools (v1.15) (Danecek et al., 2011). The final dataset was purged by PLINK (v1.90) (Purcell et al., 2007) with the parameters “-indep-pairwise 100 1 0.5”, reducing the redundant highly linked SNPs. The tagSNPs were retained to represent all SNP loci in the haplotype block. Missing genotypes were imputed with BEAGLE4 (Browning & Browning, 2007).

For the BC1F1 family, 1 620 715 variants and 50 samples were pass-filtered and subjected to quality control (QC) for further analysis. As body color is a qualitative trait, the trait with melanocytes was transformed to 1, while the trait without melanocytes was transformed to 0. Generalized Linear Mixed Model (GLMM) in GEMMA (Zhou & Stephens, 2012) was used for association analysis. The Wald frequentist test was chosen to test for significance, and the P -value threshold for genome-wide significance was calculated based on the Bonferroni-correction method (0.05/number of QC-filtered SNPs). Manhattan plots of $-\log_{10}(P)$ and QQ-plots were generated using the CMplot R package (Yin et al., 2021).

Validation of variants in *mitfa* genes

An insertion (1.7 kb) was detected in the *mitfa* gene in the A06 chromosome, as was a nonsynonymous SNP in the *mitfa* gene in the B06 chromosome. We evaluated the genotypes of this SV in YR, HB, F1, and BC1F1 individuals from the segregating population using IGV (Robinson et al., 2011). Input was a BAM file generated by mapping reads to the corrected genome sequence. The genotypes of each sample were identified by observation of high-quality reads and malformed reads flanking the SV. We also examined the genotypes by PCR. Primers were designed to amplify the genomic regions (Supplementary Table S2) and the PCR products from HB and YR were sequenced using Sanger sequencing technology. PCR validation was carried out in 23 HB, 15 YR, 55 F1 hybrid, and 118 BC1F1 individuals with recorded body color.

Alternative splicing analysis

Skin samples from individuals in the BC1F1 family with

different genotypes and skin color were collected at 5 months old. The samples were frozen with liquid nitrogen. Total RNA was extracted with Trizol reagent and qualified and quantified using an Equalbit RNA BR (Broad-Range) Assay Kit and agarose gel electrophoresis, respectively, then used for DNBSEQ mRNA library construction. The product was sequenced using the DNBSEQ platform to generate paired-end (150 bp) reads. The raw reads were filtered using SOAPnuke (Chen et al., 2018) with the parameters “-n 0.01 -l 20 -q 0.4 --adaMR 0.25 --ada_trim --polyX 50” to filter out adaptors and low-quality reads. Transcriptome data were analyzed using the HISAT, Stringtie, and Ballgown protocols (Pertea et al., 2016). Clean reads were aligned to the HB reference genome using HISAT2 (Kim et al., 2019) with the parameter “-dta”, then transferred into a binary file and sorted with SAMtools. Stringtie was used to conduct the assembly and quantification of transcripts (Pertea et al., 2015). Expression measurements at the exon, intron (junction), and transcript level were imported to the Ballgown R package to conduct differential expression analysis, visualization of transcript structures, and matching of assembled transcripts to annotation (Pertea et al., 2016). The coding sequences of transcripts were predicted and annotated with TransDecoder (<https://github.com/TransDecoder/TransDecoder>). The PCA structures and heatmaps of sample correlation were plotted using the Deseq2 R package (Love et al., 2014).

Cloning full-length *mitfa* cDNA

cDNA was synthesized using a PrimeScript RT reagent kit (Takara, Japan) according to the provided instructions. Primers for gene amplification were designed with the Vazyme online tools (<http://appbi.vazyme.com:8085/login>) (Supplementary Table S3). The PCR products were purified with a FastPure Gel DNA Extraction Mini Kit (Vazyme, China) and cloned into the pMD 19-T vector (Takara, Japan). The vector was transformed into DH5 α competent cells (Takara, Japan). Positive clones were selected and sent for Sanger sequencing. The functional domains of the protein sequences were predicted using the NCBI Conserved Domain Database (Lu et al., 2020).

Microscopic observation of common carp skin

Dorsal and head skin samples (3 μ m) were collected from the common carp. The samples were placed on glass slides with 100 μ L of normal saline and coverslipped lightly from the left to prevent air bubbles. Photos were taken using a camera mounted on a microscope.

RESULTS

HiFi genome of *C. carpio wuyuanensis*

Genome survey based on 21-mer frequency distribution revealed that the genome of *C. carpio wuyuanensis* was 1.65 Gb, similar to that of other allotetraploid species in Cyprinidae (Chen et al., 2020). To assemble a high-quality genome, 37.8 Gb (23X) of sequencing data generated using the PacBio Circular Consensus Sequencing platform and 211 Gb (134X) of sequencing data from two Hi-C libraries were obtained (Supplementary Table S4). We acquired a preliminary genome assembly with 809 contigs and a N50 of 19 736 777 bp. Total length was 1.61 Gb, close to the estimated value (Supplementary Table S5). The draft contigs were anchored to 50 chromosomes (Figure 1C; Supplementary Figure S1), totaling 1.499 Gb with a N50 of 29.512 Mb. In total, 93.1% of

the genome was anchored. Our assembly showed a substantial improvement compared to the previous HB assembly (Xu et al., 2019b), including contiguity, completeness, and correctness (Supplementary Table S6). Additional details on repetitive sequences and protein-coding gene annotation in the new HB genome assembly are given in the Supplementary Results (Supplementary Figure S1 and Tables S7, S8).

The HB genome was divided into subgenomes A and B by comparison with its progenitor B from Barbinae. The length of subgenome A (722.49 Mb) was smaller than that of subgenome B (776.59 Mb). The gene set was then divided into two groups (*C. carpio_A* and *C. carpio_B*) based on their distribution in the genome. The A and B subgenomes contained 27 109 and 28 540 genes, respectively. The collinearity regions of the two subgenomes were identified by sequence alignment and plotted into Circos plots, which also showed GC content, gene density, transposable element (TE) density, and collinearity relationships between the A and B subgenomes (Figure 1C).

Ka/Ks analysis with homologous gene pairs

To investigate the phylogenetic relationships of HB with other species, we compared the genomes of HB and 11 other published teleost species, including allotetraploid *C. auratus* and YR. YR is a representative subspecies of wild-like common carp, which is mainly cultivated in the Yellow River basin. Phylogenetic relationships were determined based on the 1 378 shared single-copy orthologous genes (Figure 1D). The B subgenomes of common carp and goldfish were clustered with species from the Barbinae (2n=50) subfamily (*Puntigrus tetrazona*, *Poropuntius huangchuchieni*, and *Onychostoma macrolepis*), while subgenome A showed a relatively greater distance from this cluster. Common carp and *C. auratus* exhibited a close genetic distance, in accordance with their divergence after whole-genome duplication (Chen et al., 2019; Ma et al., 2014). The tree also showed a close relationship between the two strains of common carp, YR and HB.

Ka and Ks values were calculated based on homologous gene pairs within representative species. The Ks values of each gene pair were calculated to estimate divergence time (Figure 1E). According to previous study, HB culture in China can be traced back more than 800 years (Lou & Sun, 2001). The calculated substitution rates of HB-YR orthologous genes were 0.010–0.011. Applying the previously determined molecular clock in teleosts (David et al., 2003), we estimated that HB and YR diverged ~1.4 million years ago, much earlier than the recorded history of domestication. As these two strains are distributed in the Yellow River and Yangtze River systems, respectively, geographical isolation may have promoted their genetic divergence. Previous study found that subgenome B is dominant in common carp, with genes showing higher expression levels and greater purifying selection than their homologs in subgenome A (Xu et al., 2019b). Here, to explore which subgenome was under greater selective pressure during divergence of the two strains, we identified 19 194 and 19 947 gene pairs between HB and YR in subgenomes A and B, respectively. Based on HB and YR gene pairs, the genes in subgenome B (mean Ka/Ks=0.438) had lower Ka/Ks values than those in subgenome A (mean Ka/Ks=0.446) ($P<0.05$) (Figure 1F). Thus, purifying selection was stronger in subgenome B, which may possess more

genes critical for life activities in common carp.

Genes with a *Ka/Ks* ratio greater than one are under positive selection (Hurst, 2002). Here, 4 833 HB-YR gene pairs had a *Ka/Ks* ratio greater than one and were therefore considered to be under positive selection and may be rapidly evolving and important in the divergence of the two strains. Enrichment analysis revealed that these genes were enriched in biological processes related to immune response (GO:0006955), immune system process (GO:0002376), response to stress (GO:0006950), phospholipid scrambling (GO:0017121), regulation of membrane lipid distribution (GO:0097035), and defense response (GO:0006952) (Figure 1G; Supplementary Table S9). Therefore, these genes may contribute to the characteristics of HB in cultivation, including good environmental adaptability, low disease susceptibility, and high meat quality. These results confirm that domestication induces substantial changes and improvement in fish immunity, as well as the ability to cope with aquacultural stressors (Milla et al., 2021).

Impact of SVs between HB and YR on gene expression

The genome is dynamic over certain time scales. The evolution and domestication of HB may have induced not only

specific phenotypes but also genomic changes. SVs are crucial in evolution and agriculture and may lead to changes in traits (Bertolotti et al., 2020; Du et al., 2021). However, little is known about such variants (large deletions, insertions, duplications, and chromosomal rearrangements) in polyploidy animals. To reveal genomic changes following HB domestication, we performed synteny analysis of gene pairs between HB and YR and identified rearrangements in the genome (Figure 2A). Furthermore, we compared the HB and YR genome assemblies and identified SVs (>50 bp) between the two, classified as PAVs, inversions, or translocations. Using the YR genome as a reference, we identified 187 661 PAVs, 511 inversions, and 11 472 translocations in the HB genome, affecting 348.5 Mb of genomic sequences. SVs can influence the expression of nearby genes by altering gene sequences or by perturbing regulatory sequences (Alonge et al., 2020; Chiang et al., 2017). In our study, most PAVs were located in intergenic regions, followed by intronic regions (Figure 2B). The FPKM (fragments per kilobase of exon per million mapped fragments) was calculated for genes in 12 tissues. Results indicated that genes with PAVs showed lower expression than those without PAVs (Figure 2C), consistent with previous findings (Qin et al., 2021). Sequence variants

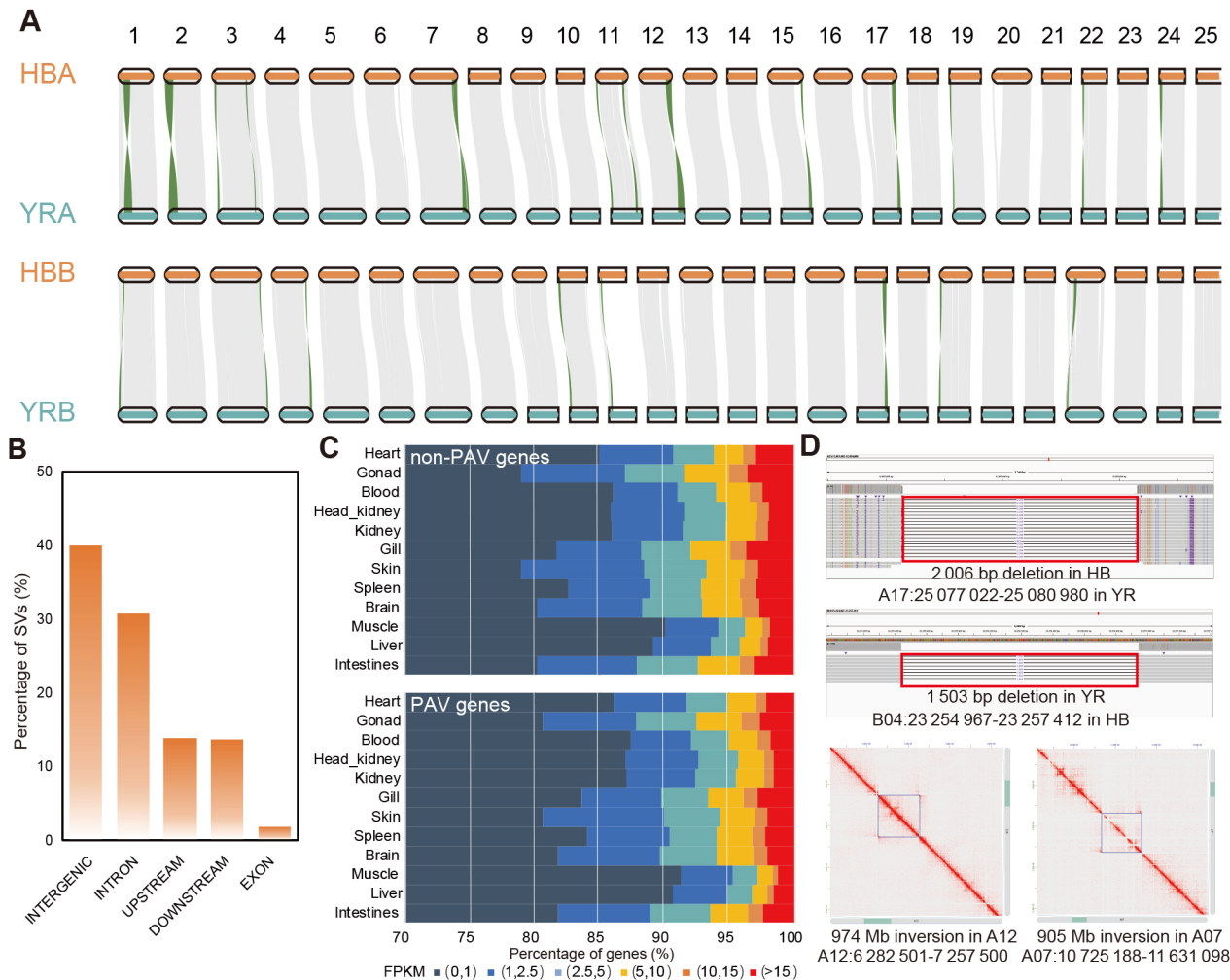


Figure 2 SVs between HB and YR

A: General view of collinearity relationship between genomes of HB and YR. Structural rearrangements are marked in green. B: Percentage of SVs overlapping different genomic regions. Most SVs were in the intergenic and intronic regions. C: Proportion comparisons between PAV-genes and non-PAV-genes with indicated expression level across 12 samples in *Cyprinus carpio*. D: Example visualization of two randomly selected SVs by mapping PacBio long-reads to genome assemblies and two by mapping Hi-C paired-end reads to genome assemblies.

between the two common carp strains were observed by mapping PacBio long reads and Hi-C paired reads to the genome assemblies (Figure 2D). SVs between the genomes of the different strains may contribute to the diversity of phenotypes in common carp.

Artificial selection leaves selection signatures in genomic regions

As one of the oldest domesticated fish species in the world, both natural and artificial selection have generated rich phenotypic variations in food and ornamental common carp strains. Due to its distinctive elliptical-shaped body and reddish skin, HB is a renowned indigenous dual-purpose strain (food and ornamental) in China. Here, to discover the genetic basis of the distinct phenotypic variations in HB, we conducted whole-genome sequencing of 34 randomly selected samples from domesticated HB and “wild-type” YR populations (18 and 16 samples, respectively), with a mean depth of 15X that of the common carp genome (Supplementary Table S10). In total, we identified 21 763 331 SNPs and 6 512 282 indels, with most variants distributed in the intergenic or intronic regions (Supplementary Figure S2).

In general, adaptive evolution or artificial selection can leave multiple selection signatures in genomic regions where the corresponding causal genes should be distributed (Ma et al., 2018). Here, PCA separated the HB and YR samples into two clusters (Supplementary Figure S3). The genome-wide population differentiation value between HB and YR was 0.14, indicating moderate differentiation between the two groups. Genomic regions with high levels of fixation in the HB strain likely include genes that were positively selected during domestication. Therefore, we performed whole-genome scanning of regions with extreme divergence in allele frequency ($F_{ST} > 0.315$) and highest differences in genetic diversity ($\log \pi$ ratio > 0.486). We identified 231 CDRs (top 5% of F_{ST} and π ratio) with significant selection signatures between the two strains, harboring 138 genes (Figure 3A; Supplementary Tables S11, S12). Among them, 13 genes were related to “response to stimulus” (GO: 0050896), 23 genes were related to “metabolic process” (GO:0008152), 12 genes were related to “phosphorus metabolic process” (GO:0006793), and two genes were related to “progesterone-mediated oocyte maturation” (Figure 3B).

Microscopic observation showed that melanocytes were abundant in the gray-brown skin of YR but were completely absent in the reddish skin of HB (Figure 3C). Of note, KEGG annotation highlighted two genes related to melanogenesis, i.e., *mitfa* gene in homologous chromosomes A06 and B06, respectively. The *mitf* gene is critical in the melanocyte and retinal pigment epithelium formation regulatory networks, acting as a “master regulator” (Levy et al., 2006). Many fish, including zebrafish, contain the *mitfa* and *mitfb* genes (Johnson et al., 2011). Notably, *mitfa* affects the development of skin pigment cells, and its mutation is related to cutaneous albinism (Johnson et al., 2011). There are two copies of *mitfa* in the two subgenomes of allotetraploid fish. YR and HB showed extreme divergence in allele frequency and the HB strain showed low genetic diversity in these two *mitfa* regions (Figure 3D). These findings indicate strong selection in the HB genome during its domestication history.

Hybrid family construction revealed inheritance of skin color

Common carp exhibit a variety of skin colors, and the

molecular mechanism associated with pigmentation is an interesting area of research (Orteu & Jiggins, 2020). Hybridization experiments play an important role in exploring genetic mechanisms (Xu et al., 2019a). To investigate the inheritance pattern and demonstrate the mutational effects of *mitfa* on skin color in the HB strain, we constructed a backcross inheritance experiment using YR and HB as parents, with the hybrid F1 generation then backcrossed with either HB or YR. The F1 individuals reached sexual maturity at three years and were used for the backcross experiments (F1 $\times$ HB or F1 $\times$ YR). After hatching and loss during the growth period, a total of 17 families were obtained and their skin color phenotypes were recorded. Offspring without melanophores were present in the intercross families in which both parents contained melanophores, suggesting that melanophore presence is dominant over their absence. We obtained four segregation ratios (7:1, 3:1, 1:1, and 1:0) in the BC1F1 family (Supplementary Figure S4 and Supplementary Table S13), which indicated the potential control of two pairs of alleles in phenotype determination. The allele for melanophore presence was dominant over the allele for melanophore absence, consistent with the observation of the breeding process in common carp (Deng, 1981).

In one family, the inheritance pattern of body color conformed to the Mendelian law in the first backcross generation, producing a 3:1 phenotype segregation in the backcross family (brown:red=278:95, $\chi^2_{df=1}=4.59e-04$) (Figure 4A). GWAS was performed using 50 individuals (Supplementary Figure S5). After pass filtering and QC of 1 620 715 variants and all samples, we identified two loci (A06: 17 668 989–23 701 042 and B06: 16 901 881–22 857 335) highly associated with skin color phenotypes ($P < 3.09e-8$) (Figure 4B; Supplementary Table S14), which overlapped with selective signals in chromosomes A06 and B06. As expected, these regions harbored the *mitfa* gene, which plays a crucial role in the melanogenesis pathway (Johnson et al., 2011).

As an allotetraploid fish, common carp exhibit complex genetic mechanisms for trait determination. While *mitf* genes generally have two paralogs (*mitfa* and *mitfb*) in diploid teleosts due to third-round genome duplication (3R WGD), we found four copies (*mitfa* (A06), *mitfa* (B06), *mitfb* (A23), *mitfb* (B23)) in the common carp. According to our research, only simultaneous loss-of-function of both copies of *mitfa* resulted in the absence of melanophores. In other words, the absence of melanocytes in the skin of HB depended on a double recessive homozygous allelic genotype (Figure 4A). After scanning gene regions by IGV, several fixed mutations were observed in the HB strain. In *mitfa* (A06), we identified a haplotype with a 240 bp deletion (A06: 23 318 912–23 318 913) and a 1 773 bp insertion (A06: 23 315 866–23 317 639), which was homozygous in the HB strain but heterozygous or without mutation in the YR strain (Figure 4C; Supplementary Figure S6A). Furthermore, a nonsynonymous single nucleotide mutation (B06: 22 467 522 (G→C)) was found in the 8th exon of *mitfa* (B06), which caused an Ala244→Pro244 substitution. Comparison among several animals revealed that the mutation was in the conserved functional domain, which may influence gene function (Figure 4D; Supplementary Figure S6B).

We performed PCR and Sanger sequencing to validate the *mitfa* (A06) and *mitfa* (B06) variants. Genotypes were identified in multiple samples, including YR, HB, F1 individuals, and backcross progeny. Results confirmed that

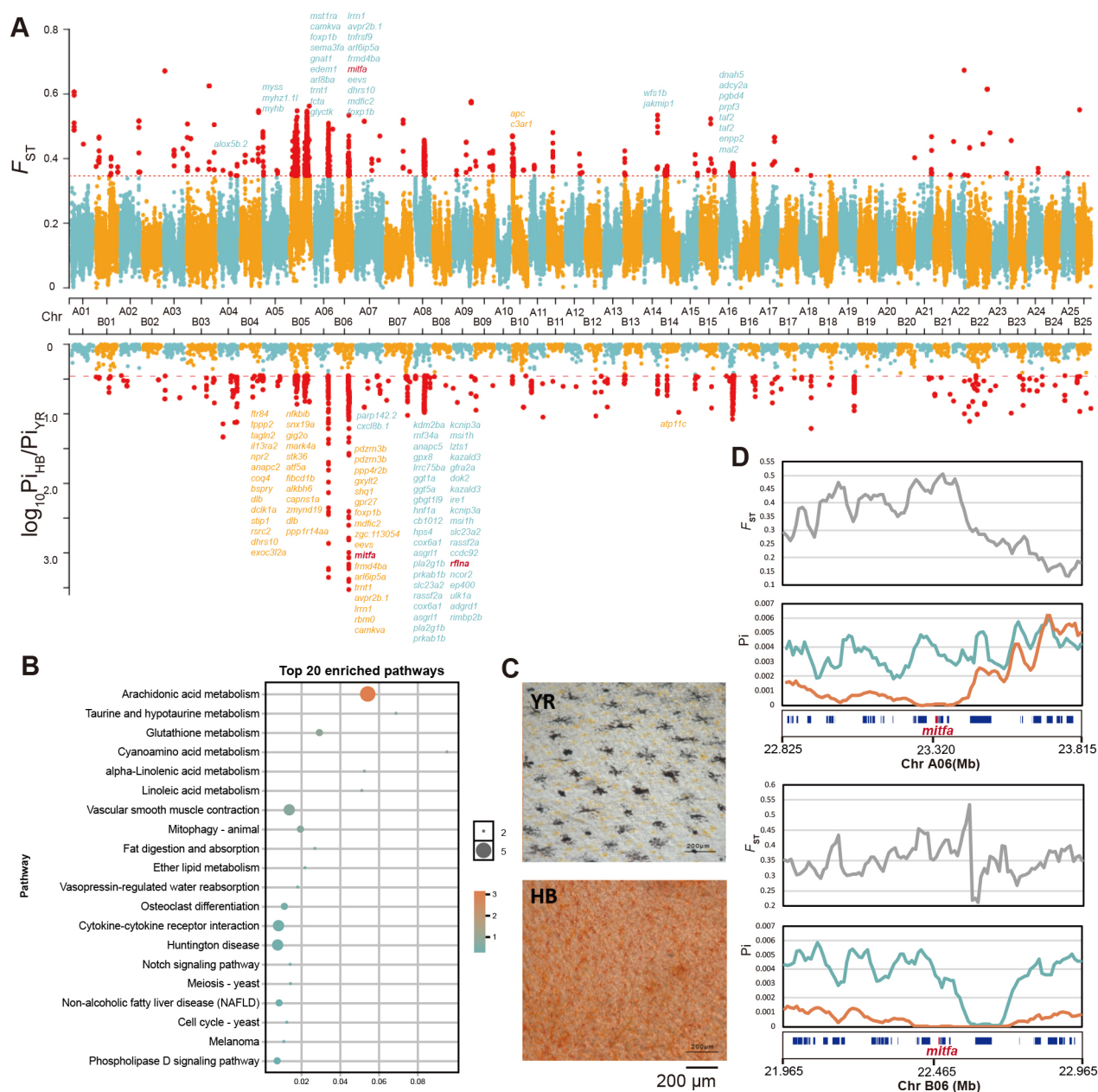


Figure 3 Selective signals identified by whole-genome analyses

A: Pairwise fixation index (F_{ST}) and ratio of nucleotide diversity (π) in 40 kb sliding windows and 10 kb sliding step length across genome between HB and YR. Top 5% value was used as cutoff. Genes located in divergent regions are indicated by gene names. B: KEGG enrichment of genes in genomic divergent regions. Top 20 pathways with highest P -values are plotted. C: Microscopic observation of skin pigment cells of YR and HB. D: Selective signals of *mitfa* regions in chromosomes A06 and B06. Gray, F_{ST} between HB and YR; blue, nucleotide diversity in YR; orange, nucleotide diversity in HB. These regions are enriched in sites nearly fixed in HB breeds during domestication process.

both mutations in *mitfa* were homozygous in HB and other individuals without melanocytes, whereas the genotypes were heterozygous or not mutated in the YR and hybrid individuals with melanophores (Supplementary Table S15). The segregation ratios of crossed families were consistent with the ratios predicted by the genotypes of parents (Supplementary Table S13). Thus, functional mutation of both copies of *mitfa* plays a prominent role in carp phenotypic evolution.

Alternative splicing induced by intronic SVs in *mitfa*

Based on alternative promoters, a single *mitf* gene can produce more than one isoform, with Mitf-M found to be specific in skin and melanophores (Hartman & Czyz, 2015). Here, SVs in the intronic region of *mitfa* (A06) were associated

with red skin color in the common carp, as homozygous individuals were completely red, whereas heterozygous individuals or those without SVs were brown. Intronic SVs in the genome not only influence expression levels but also regulation through aberrant splicing patterns of genes (Stern et al., 2017; Varagona et al., 1992; Zhou et al., 2018). The hybrid common carp populations offered the opportunity to explore the influence of intronic SVs in *mitfa* (A06). A 1 773 bp insertion in *mitfa* (A06) was detected in the intron between exon 1-2 and exon 2. Skin samples from 20 individuals (one family) with different haplotypes were collected for RNA sequencing (RNA-seq). After assembling the transcripts, we found that the gene possessed two transcripts with alternative first exons (exon 1-1 and exon 1-2) in the common carp skin

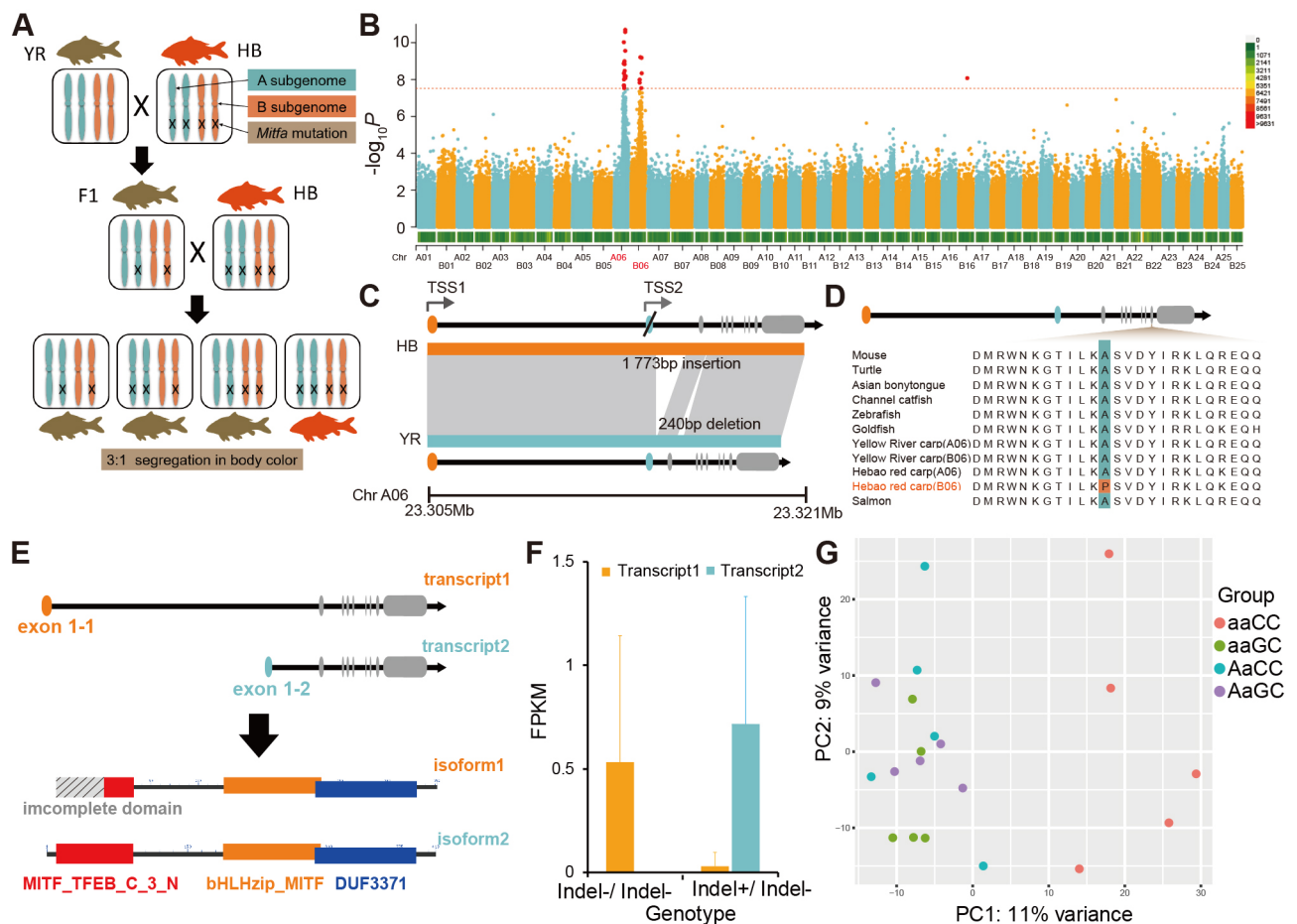


Figure 4 Deficiency in both homologous *mitfa* genes responsible for reddish color in domesticated HB

A: Diagram of BC1F1 population generated by intercross between HB and YR. Body color trait segregated in one backcross family with a 3:1 separation ratio. **B:** GWAS of carp body color, including 30 individuals with melanocytes and 30 individuals without melanocytes. Horizontal dashed lines indicate Bonferroni significance threshold of GWAS ($P < 3.09 \times 10^{-8}$). Two loci (A06: 17 668 989–23 701 042 and B06: 16 901 881–22 857 335) were highly associated with skin color phenotypes. **C:** Diagram of mutation of *mitfa* gene in chromosome A06. A 243 bp deletion and 1 773 bp insertion were detected in intronic regions of HB genome. **D:** Nonsynonymous mutation (B06: 22 381 071 (G→C)) of *mitfa* gene in B06 chromosome. Mutation caused an Ala244→Pro244 substitution in conserved region among multiple species. **E:** Two transcript variants with specific first exons of *mitfa* (A06) and functional domain of their isoforms. **F:** Expression level of two transcripts of *mitfa* (A06). Transcript 2 showed no expression in skin with homozygous haplotype of HB.

(Figure 4E). The expression patterns of the two transcripts in different haplotypes differed significantly. The transcript that expressed exon 1-2 was not expressed in the skin when the insertion existed (Figure 4F). The transcript, which was not expressed in HB, may be necessary for the development of melanophores, with its absence resulting in red skin. After cloning the full length of both transcripts, we found that alternative splicing resulted in a loss of 60 amino acids and deficiency of the functional domain “MITF-TFEB-C-3-N” in Mitfa (Figure 4E), which may be necessary for the generation of melanocytes. Previous research reported that a large intronic insertion in Pekin ducks induces white plumage by causing splicing changes in *MITF* (Zhou et al., 2018). Thus, the large intronic insertion in HB may induce red skin by causing abnormal splicing in Mitfa

DISCUSSION

Based on population genetic analysis, GWAS of body color traits in backcross populations, and transcript analysis, our results indicated that co-mutation of *mitfa* genes in the two subgenomes of allotetraploid common carp induced skin

melanophore loss. The number and recessiveness of alleles in HB body color determination are consistent with previous studies (Deng, 1981; Zhang & Pan, 1983). In our backcross family, melanophore disappearance was accompanied by red color enhancement, suggesting the potential role of *mitfa* in red coloration in fish. However, further research is needed to clarify this possibility.

For allotetraploids, duplication of the genome can result in neofunctionalization and subfunctionalization of genes. In our study, based on the transcriptomes of different genotypes, the expression patterns of individuals with mutations in either *mitfa* copy in the two subgenomes (Aabb and aaBb) were similar to those with two functional *mitfa* (AaBb) genes (Figure 4G; Supplementary Figure S7), indicating that the two copies of *mitfa* exert similar (same) functions in common carp. Furthermore, dysfunction of a gene in one subgenome can be repaired by the other. Genome duplication in common carp may lead to higher fault tolerance in the polyploid genome and greater opportunities for genomic evolution. Thus, this suggests that genome duplication may contribute to species diversity and phenotypic diversification (Crow & Wagner, 2006).

In our study, high-quality HB genome assembly was accomplished using long HiFi reads. The long domestication history of HB has induced considerable genomic changes and selection signatures were detected in the HB genome. Functional mutations in the *mitfa* copies in the two subgenomes played a prominent role in carp body color determination. Our results will help better understand the genomic structure of the HB strain and provide a theoretical basis for genetic study in allotetraploid fish.

DATA AVAILABILITY

All sequencing data generated in this study were deposited at the Sequence Read Archive (<https://ncbi.nlm.nih.gov/sra>) under BioProjectID PRJNA510861, PRJNA824207, PRJNA853462, and PRJNA853504, Genome Sequence Archive (GSA) database of the National Genomics Data Center (NGDC) under accession number CRA009294, and Science Data Bank (doi: 10.57760/sciencedb.j00139.00045).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

P.X. conceived and designed the research. B.J.L., L.C., and M.Z.Y. performed the experiments. B.J.L., X.Q.Z., Y.L.B., Q.H., and C.Y.L. analyzed the data. B.J.L. wrote the manuscript. Z.J., Y.G.X., and J.X.F. helped with the construction of fish families and fish farming. B.H.C., T.Z., and P.X. revised the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We would like to express our thanks to Yan-Hui Wang and staff at the Henan Academy of Fishery Science and Professor Xue-Jun Li at Henan Normal University for help during fieldwork, and Yu-Xin Chen for providing pictures used in the manuscript.

REFERENCES

Akagi T, Hanada T, Yaegaki H, et al. 2016. Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. *DNA Research*, **23**(3): 271–282.

Alonge M, Wang XG, Benoit M, et al. 2020. Major impacts of widespread structural variation on gene expression and crop Improvement in tomato. *Cell*, **182**(1): 145–161.e23.

Andersson L. 1997. The use of a wild pig×domestic pig intercross to map phenotypic trait loci. *Journal of Heredity*, **88**(5): 380–383.

Bakos J, Gorda S. 1995. Genetic improvement of common carp strains using intraspecific hybridization. *Aquaculture*, **129**(1–4): 183–186.

Balon EK. 2004. About the oldest domesticates among fishes. *Journal of Fish Biology*, **65**(S1): 1–27.

Bertolotti AC, Layer RM, Gundappa MK, et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, **11**(1): 5176.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**(5): 1084–1097.

Carneiro M, Rubin CJ, Di Palma F, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science*, **345**(6200): 1074–1079.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*,

17(4): 540–552.

Chen D, Zhang Q, Tang WQ, et al. 2020. The evolutionary origin and domestication history of goldfish (*Carassius auratus*). *Proceedings of the National Academy of Sciences of the United States of America*, **117**(47): 29775–29785.

Chen YX, Chen YS, Shi CM, et al. 2018. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience*, **7**(1): gix120.

Chen ZL, Omori Y, Koren S, et al. 2019. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Science Advances*, **5**(6): eaav0547.

Chiang C, Scott AJ, Davis JR, et al. 2017. The impact of structural variation on human gene expression. *Nature Genetics*, **49**(5): 692–699.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity?. *Molecular Biology and Evolution*, **23**(5): 887–892.

Danecek P, Auton A, Abecasis G, et al. 2011. The variant call format and VCFtools. *Bioinformatics*, **27**(15): 2156–2158.

Danecek P, Bonfield JK, Liddle J, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience*, **10**(2): giab008.

Darriba D, Taboada GL, Doallo R, et al. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**(8): 1164–1165.

Darwin C. 2010. The Variation of Animals and Plants Under Domestication. New York: Cambridge University Press.

David L, Blum S, Feldman MW, et al. 2003. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Molecular Biology and Evolution*, **20**(9): 1425–1434.

Deng ZJ. 1981. Study on body shape formation and body color inheritance of Hebao red Carp in Wuyuan, Jiangxi province. *Freshwater Fisheries*, **6**(2): 14,22. (in Chinese)

Du H, Zheng XR, Zhao QQ, et al. 2021. Analysis of structural variants reveal novel selective regions in the genome of Meishan pigs by whole genome sequencing. *Frontiers in Genetics*, **12**: 550676.

Durand NC, Robinson JT, Shamim MS, et al. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, **3**(1): 99–101.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**(1): 238.

Essa BH, Suzuki S, Nagano AJ, et al. 2021. QTL analysis for early growth in an intercross between native Japanese Nagoya and White Plymouth Rock chicken breeds using RAD sequencing-based SNP markers. *Animal Genetics*, **52**(2): 232–236.

Goel M, Sun HQ, Jiao WB, et al. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, **20**(1): 277.

Hartman ML, Czyz M. 2015. MITF in melanoma: mechanisms behind its expression and activity. *Cellular and Molecular Life Sciences*, **72**(7): 1249–1260.

Hu XS, Ge YL, Li CT, et al. 2018. Developments in common carp culture and selective breeding of new varieties. In: Gui JF, Tang QS, Li ZJ, Liu JS, De Silva SS. Aquaculture in China. Chichester: John Wiley & Sons Ltd, 125–148.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, **18**(9): 486–487.

Johnson SL, Nguyen AN, Lister JA. 2011. *mitfa* is required at multiple stages of melanocyte differentiation but not to establish the melanocyte stem cell. *Developmental Biology*, **350**(2): 405–413.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. In: Posada D. Bioinformatics for DNA Sequence Analysis. New York: Humana Press, 39–64.

- Kim D, Paggi JM, Park C, et al. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**(8): 907–915.
- Kurtz S, Phillippy A, Delcher AL, et al. 2004. Versatile and open software for comparing large genomes. *Genome Biology*, **5**(2): R12.
- Lexer C, Randell RA, Rieseberg LH. 2003. Experimental hybridization as a tool for studying selection in the wild. *Ecology*, **84**(7): 1688–1699.
- Levy C, Khaled M, Fisher DE. 2006. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine*, **12**(9): 406–414.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5): 589–595.
- López ME, Benestan L, Moore JS, et al. 2019. Comparing genomic signatures of domestication in two Atlantic salmon (*Salmo salar* L.) populations with different geographical origins. *Evolutionary Applications*, **12**(1): 137–156.
- Lou YD, Sun JC. 2001. Progress on studies of origin and genetic diversity of three breeds of red carp in Jiangxi Province. *Journal of Fisheries of China*, **25**(6): 570–575. (in Chinese)
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**(12): 550.
- Lu SN, Wang JY, Chitsaz F, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*, **48**(D1): D265–D268.
- Ma W, Zhu ZH, Bi XY, et al. 2014. Allopolyploidization is not so simple: evidence from the origin of the Tribe Cyprinini (Teleostei: Cypriniformes). *Current Molecular Medicine*, **14**(10): 1331–1338.
- Ma YL, Zhang SX, Zhang KL, et al. 2018. Genomic analysis to identify signatures of artificial selection and loci associated with important economic traits in Duroc pigs. *G3 Genes Genomes Genetics*, **8**(11): 3617–3625.
- Mendel G. 1996. Experiments in plant hybridization (1865).
- Milla S, Pasquet A, El Mohajer L, et al. 2021. How domestication alters fish phenotypes. *Reviews in Aquaculture*, **13**(1): 388–405.
- Nakajima T, Hudson MJ, Uchiyama J, et al. 2019. Common carp aquaculture in Neolithic China dates back 8, 000 years. *Nature Ecology & Evolution*, **3**(10): 1415–1418.
- Nam BH, Yoo D, Kim YO, et al. 2019. Whole genome sequencing reveals the impact of recent artificial selection on red sea bream reared in fish farms. *Scientific Reports*, **9**(1): 6487.
- Orteu A, Jiggins CD. 2020. The genomics of coloration provides insights into adaptive evolution. *Nature Reviews Genetics*, **21**(8): 461–475.
- Pertea M, Kim D, Pertea GM, et al. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, **11**(9): 1650–1667.
- Pertea M, Pertea GM, Antonescu CM, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3): 290–295.
- Poplin R, Ruano-Rubio V, DePristo MA, et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.
- Price AL, Patterson NJ, Plenge RM, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8): 904–909.
- Purcell S, Neale B, Todd-Brown K, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3): 559–575.
- Qin P, Lu HW, Du HL, et al. 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**(13): 3542–3558.e16.
- Qiu Q, Wang LZ, Wang K, et al. 2015. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nature Communications*, **6**: 10283.
- Robinson JT, Thorvaldsdóttir H, Winckler W, et al. 2011. Integrative genomics viewer. *Nature Biotechnology*, **29**(1): 24–26.
- Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(S1): 8641–8648.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9): 1312–1313.
- Stern DL, Ding Y, Berrocal A, et al. 2017. Natural courtship song variation caused by an intronic retroelement in an ion channel gene. *Integrative and Comparative Biology*, **57**: E419–E419.
- Tang HB, Bowers JE, Wang XY, et al. 2008. Synteny and collinearity in plant genomes. *Science*, **320**(5875): 486–488.
- Varagona MJ, Purugganan M, Wessler SR. 1992. Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. *The Plant Cell*, **4**(7): 811–820.
- Villanueva RAM, Chen ZJ. 2019. ggplot2: elegant graphics for data analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives*, **17**(3): 160–167.
- Vollger MR, Logsdon GA, Audano PA, et al. 2020. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*, **84**(2): 125–140.
- Wang K, Li MY, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16): e164.
- Wenger AM, Peluso P, Rowell WJ, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**(10): 1155–1162.
- Xu CX, Li Q, Yu H, et al. 2019a. Inheritance of shell pigmentation in Pacific oyster *Crassostrea gigas*. *Aquaculture*, **512**: 734249.
- Xu P, Xu J, Liu GJ, et al. 2019b. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nature Communications*, **10**(1): 4625.
- Xu P, Zhang XF, Wang XM, et al. 2014. Genome sequence and genetic diversity of the common carp. *Cyprinus carpio*. *Nature Genetics*, **46**(11): 1212–1219.
- Yin LL, Zhang HH, Tang ZS, et al. 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics*, **19**(4): 619–628.
- Zhang JS, Pan GB. 1983. Body form and body colour in hybrids of *Cyprinus carpio*. *Journal of Fisheries of China*, **7**(4): 301–312. (in Chinese)
- Zhang Z. 2022. KaKs_calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics, Proteomics & Bioinformatics*, **20**(3): 536–540.
- Zhang Z, Xiao JF, Wu JY, et al. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications*, **419**(4): 779–781.
- Zhou J, Wu Q, Wang Z, et al. 2004. Genetic variation analysis within and among six varieties of common carp (*Cyprinus carpio* L.) in China using microsatellite markers. *Russian Journal of Genetics*, **40**(10): 1144–1148.
- Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**(7): 821–824.
- Zhou ZK, Li M, Cheng H, et al. 2018. An intercross population study reveals genes associated with body size and plumage color in ducks. *Nature Communications*, **9**(1): 2648.

Supplementary Materials

Intercross population study reveals that co-mutation of *mitfa* genes in two subgenomes induces red skin color in common carp (*Cyprinus carpio wuyuanensis*)

Bi-Jun Li^{1,2}, Lin Chen^{1,2}, Meng-Zhen Yan^{1,2}, Xiao-Qing Zou^{1,2}, Yu-Lin Bai^{1,2}, Ya-Guo Xue³, Zhou Jiang^{1,3}, Bao-Hua Chen^{1,2}, Cheng-Yu Li^{1,2}, Qian He^{1,2}, Jian-Xin Feng⁴, Tao Zhou^{1,2}, Peng Xu^{1,2,*}

¹State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian 361102, China

²Fujian Key Laboratory of Genetics and Breeding of Marine Organisms, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian 361102, China

³College of Fisheries, Henan Normal University, Xinxiang, Henan 453007, China

⁴Henan Academy of Fishery Science, Zhengzhou, Henan 450039, China

*Corresponding author, E-mail: xupeng77@xmu.edu.cn

Supplementary Materials and Methods

Sample collection and sequencing for genome assembly

A female *C. carpio wuyuanensis* individual was collected from the Breeding Station of the Henan Academy of Fishery Sciences, Zhengzhou, Henan, China. Muscle tissues were frozen with liquid nitrogen for further DNA extraction. Genomic DNA was extracted using the phenol-chloroform protocol, then qualified with a NanoDrop 2000 spectrophotometer and quantified with a Qubit fluorometer. High-quality DNA was sent to AnnoRoad (Wuhan, China) for polymerase chain reaction (PCR)-free SMRT bell library (CCS) construction and sequencing using the PacBio Sequel/Sequel II platform. CCS software (<https://ccs.how/>) was applied to generate highly accurate single-molecule consensus reads from subreads, yielding 37.8 Gb of CCS reads. Average subread length was 15 321 bp and number of reads was 34.7 million.

For chromosome-level *C. carpio* genome assembly, frozen muscle tissues from the same fish were sent to Novogene (Tianjin, China) for high-throughput chromosome conformation capture technology (Hi-C) library construction and sequencing. Two libraries were constructed using *Mbo*I restriction endonuclease and sequencing was performed on the Illumina NovaSeq 6000 platform. In total, 200 Gb of Hi-C data were obtained. Total RNA was isolated from nine tissues of *C. carpio wuyuanensis*, including gill, liver, heart, blood, muscle, spleen, brain, skin, and intestines. RNA sequencing was performed following the protocols of a PureLink™ RNA Mini Kit (Invitrogen, Shanghai, China). The library was constructed using Illumina standard protocols (San Diego, CA, USA) and sequenced on the Illumina HiSeq 6000 platform.

Assembly and scaffolding

Due to the large size and complexity of the allotetraploid genome, the assembly algorithm for the PacBio HiFi read data was used to perform Hi-C-integrated assembly with both CCS reads and paired-end Hi-C reads. Contigs were generated with Hifisam in Hi-C integration mode (Cheng et al., 2021). The Hi-C interactions were used in the scaffolding of contigs. Clean reads were aligned to the draft genome using Juicer and the 3D-DNA pipeline (Dudchenko et al., 2017) was run to scaffold the genome into chromosomes. Juicebox was used for manual adjustment.

Assembly completeness and correctness were assessed using four methods. 1) Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was performed by searching against the BUSCO database (actinopterygii_odb10). 2) Illumina short reads were mapped to the genome using BWA. 3) HiFi reads were mapped to the genome with minimaps. 4) RNA-seq and iso-seq reads were mapped to the genome using hisat2 and minimap2, respectively.

Repeat sequence annotation

Repeat sequence annotation was conducted with homology- and *de novo*-based methods. Repeat sequences in the genome were detected with RepeatModeler (v2.0.1) (Flynn et al., 2020) and LTR_Finder (v1.07) (Xu & Wang, 2007). First, a repeat sequence library was built using the Repbase database (Bao et al., 2015) to search and classify repeats. Unclassified repeats were annotated with TEclass (v2.1.3) (Abrusn et al., 2009). The script buildSummary.pl in RepeatMasker (v4.1.0) (Tarailo-Graovac & Chen, 2009) was adopted to determine and summarize annotation results. Tandem Repeats Finder (v4.09) (Benson, 1999) was used to identify tandem repeats. For protein-coding gene structure annotation, all repetitive regions except tandem repeats were soft-masked.

Gene structure annotation and functional annotation

RNA-seq data were used in gene structure annotation. Homology-based prediction, *ab initio* annotation, and transcriptome-based prediction were applied for protein-coding gene structure annotation. For homology-based prediction, the protein sequences of *Danio rerio*, *Carassius auratus*, *Onychostoma macrolepis*, *Labeo rohita*, and *Poropuntius huangchuchieni* were downloaded from an online database (**Supplementary Table S1**). These sequences were input into Braker2 (v2.1.5) software (Bruna et al., 2021), with genes in the genome then annotated with the close homologous protein model. For *ab initio* annotation, Trinity-assembled transcripts and corrected reads of iso-seq were aligned to the genome and open reading frames (ORFs) were predicted using PASA software (Haas et al., 2008). For transcriptome-based prediction, RNA-seq reads from nine tissues were mapped to the genome using hisat2, then assembled into gene models using Stringtie. TransDecoder was used to predict CDS regions. Finally, evidence from the three methods was provided to EvidenceModeler and integrated (Bruna et al., 2021).

For gene function annotation, Diamond was used to BLAST search against the protein sequences in the Swiss-Prot (<http://www.uniprot.org/>), TrEMBL (<http://www.uniprot.org/>), and NR protein databases. In addition, GO annotation and protein family annotation were conducted using InterProScan (<https://www.ebi.ac.uk/interpro/>). KO terms for each gene were assigned using KAAS (<https://www.genome.jp/tools/kaas/>). The tRNAscan-SE (Chan et al., 2021) and RNAmmer programs (Lagesen et al., 2007) were used to predict transfer RNA (tRNA) and ribosomal RNA (rRNA), respectively. Other non-coding RNA (ncRNA) was identified by searching against the Rfam database (<http://eggnogdb.embl.de/>).

Supplementary Results

HiFi genome of *C. carpio wuyuanensis*

Genome survey based on 21-mer frequency distribution revealed that the *C. carpio wuyuanensis* genome was 1.65 Gb, similar to that reported in other allotetraploid species in Cyprinidae (Chen et al., 2020). To assemble a high-quality genome, 37.8 Gb (23X) of sequencing data generated using the PacBio Circular Consensus Sequencing platform and 211 Gb (134X) of sequencing data from two Hi-C libraries (read length 150 bp) were obtained (**Supplementary Table S4**). Average length of the HiFi reads was 15 321 bp. We acquired a preliminary genome assembly with 809 contigs and a N50 of 19 736 777 bp. Total length was 1.610 Gb, close to the estimated value (**Supplementary Table S5**). The draft contigs were then anchored and oriented into a chromosome-scale assembly using the 3D-DNA pipeline. After scaffolding, the genome was anchored to 50 chromosomes (**Figure 1C, Supplementary Figure S1**), totaling 1.499 Gb with a N50 of 29.512 Mb. In total, 93.1% of the genome was anchored. To determine genome assembly completeness, we performed BUSCO assessment, resulting in complete BUSCOs (C) of 98.6% (**Supplementary Table S6**).

The previous HB assembly was developed in 2019 (Xu et al., 2019). Three key dimensions were employed to compare the quality of the two assemblies, i.e., contiguity, completeness, and correctness. First, contig N50 increased substantially from 20.68 kb to 19.74 Mb. Second, genome completeness based on BUSCO improved from 96.5% to 98.60%. Third, DNA and RNA sequencing reads of HB were remapped to the genome to evaluate correctness with the mapping ratio. The mapping ratios of genomic Illumina short reads and HiFi long reads improved from 96.64% and 99.91% in the previous assembly to 99.18% and 99.93% in the new assembly. In addition, the RNA-seq read mapping rate improved from 82.31% to 88.66%, while the iso-seq read mapping rate improved from 99.28% to 99.93%. Furthermore, 93.1% of contigs were anchored to chromosomes in the new assembly compared to 82.2% in the previous assembly. These data show a substantial improvement in the new assembly compared to the previous version.

Transposable element (TE) and gene annotation of genome

We identified a total of 664 Mb of repetitive sequences, accounting for 44.30% of the HB genome. The percentage of repetitive sequences in the HB genome was larger than that of the previous version (557.87 Mb, 36.94%) (Xu et al., 2019). Among the sequences, DNA transposons were the most abundant repetitive elements, totaling 328 Mb and accounting for 21.87% of the entire genome. In addition, we identified 69.56 Mb of long interspersed elements (LINEs; 4.64 %), 6.75 Mb of short interspersed nuclear elements (SINEs; 0.45%), and 75.41 Mb of long terminal repeats (LTRs; 5.03%) (**Supplementary Table S7**). L2 was the most abundant LINE, while hobo-Activator and Tc1-IS630-Pogo were the most abundant DNA transposons. TE divergence analysis suggested recent DNA transposon and LINE activity in the genome (**Supplementary Figure S1B**).

Combining *de novo*, homology, and transcriptome-assisted predictions, a total of

58 842 protein-coding genes were annotated in the genome. In total, 75.1%, 46.1%, 71.9%, 97.4%, and 99.7% of genes were functionally annotated in the Swiss-Prot, KEGG, GO, Pfam, and NCBI non-redundant nucleotide databases (**Supplementary Figure S1C, Supplementary Table S8**). BUSCO evaluation of the protein-coding annotations revealed that 94.7% of *Actinopterygii* genes were annotated, including 1 431 (39.3%) complete and single-copy and 2 016 (55.4%) complete and duplicated BUSCOs (D) (**Supplementary Table S6**). The high D value may be characteristic of tetraploids, as observed in *C. auratus* (Chen et al., 2019, 2020). Furthermore, 5 100 microRNAs (miRNAs), 27 539 tRNAs, 11 178 rRNAs, and 4 412 small nuclear RNAs (snRNAs) were identified in the HB genome.

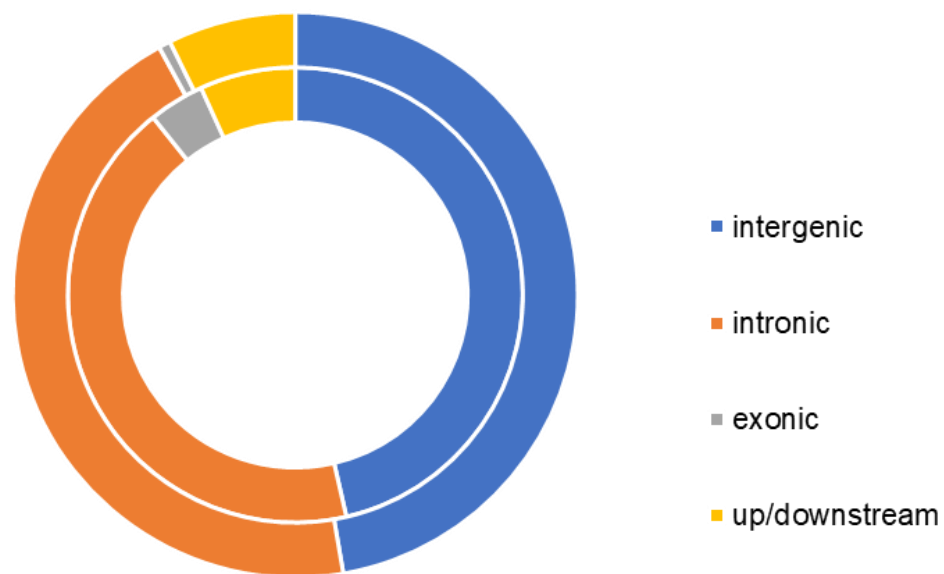
References

- Abrusn G, Grundmann N, DeMester L, et al. 2009. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**(10): 1329–1330.
- Bao WD, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 2015, **6**: 11.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**(2): 573–580.
- Bruna T, Hoff KJ, Lomsadze A, et al. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP⁺ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, **3**(1): lqaa108.
- Chan PP, Lin BY, Mak AJ, et al. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, **49**(16): 9077–9096.
- Chen D, Zhang Q, Tang WQ, et al. 2020. The evolutionary origin and domestication history of goldfish (*Carassius auratus*). *Proceedings of the National Academy of Sciences of the United States of America*, **117**(47): 29775–29785.
- Chen ZL, Omori Y, Koren S, et al. 2019. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Science Advances*, **5**(6): eaav0547.
- Cheng HY, Concepcion GT, Feng XW, et al. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, **18**(2): 170–175.
- Dudchenko O, Batra SS, Omer AD, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**(6333): 92–95.
- Flynn JM, Hubley R, Goubert C, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(17): 9451–9457.
- Haas BJ, Salzberg SL, Zhu W, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, **9**(1): R7.
- Lagesen K, Hallin P, Rodland EA, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**(9): 3100–3108.
- Levy C, Khaled M, Fisher DE. 2006. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine*, **12**(9): 406–414.

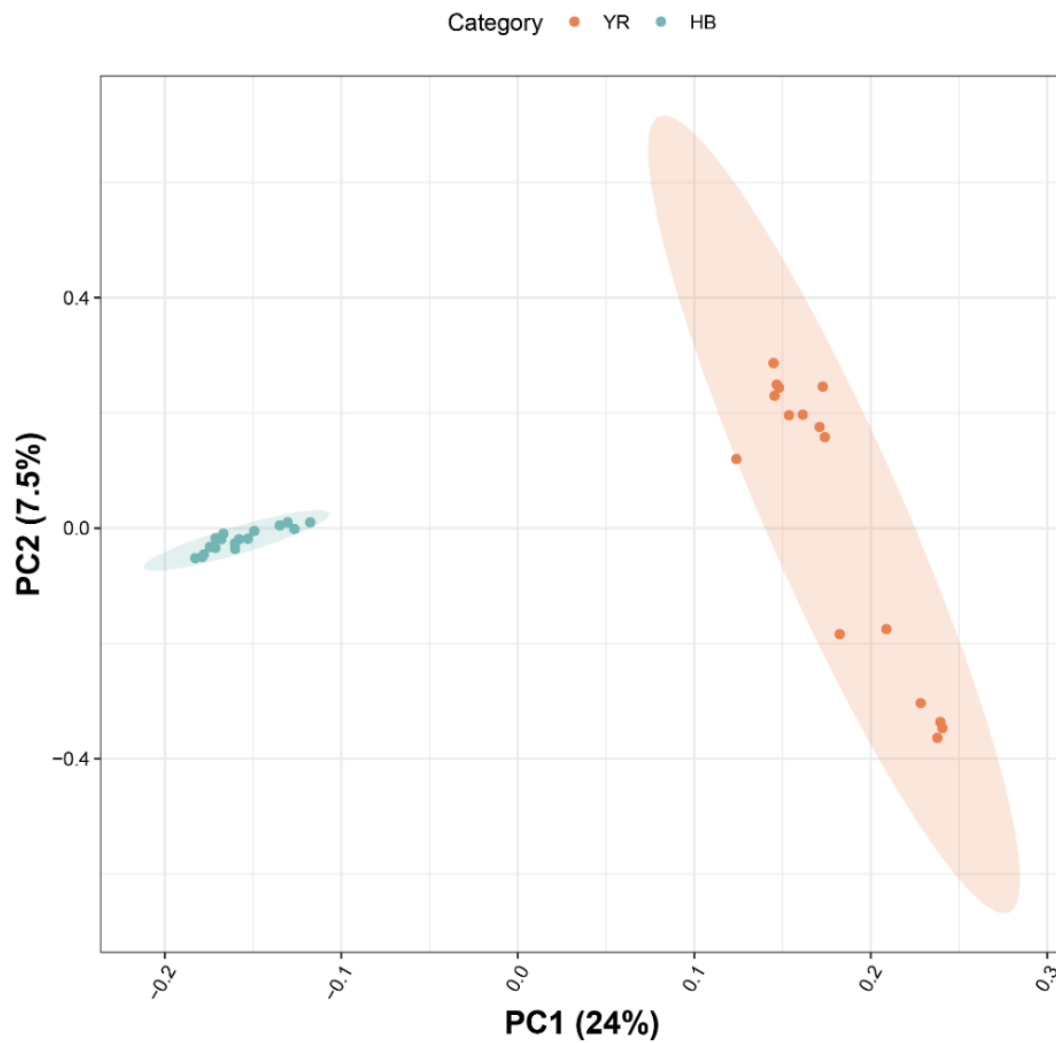
Tarailo-Graovac M, Chen NS. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, **25**(1): 4–10.

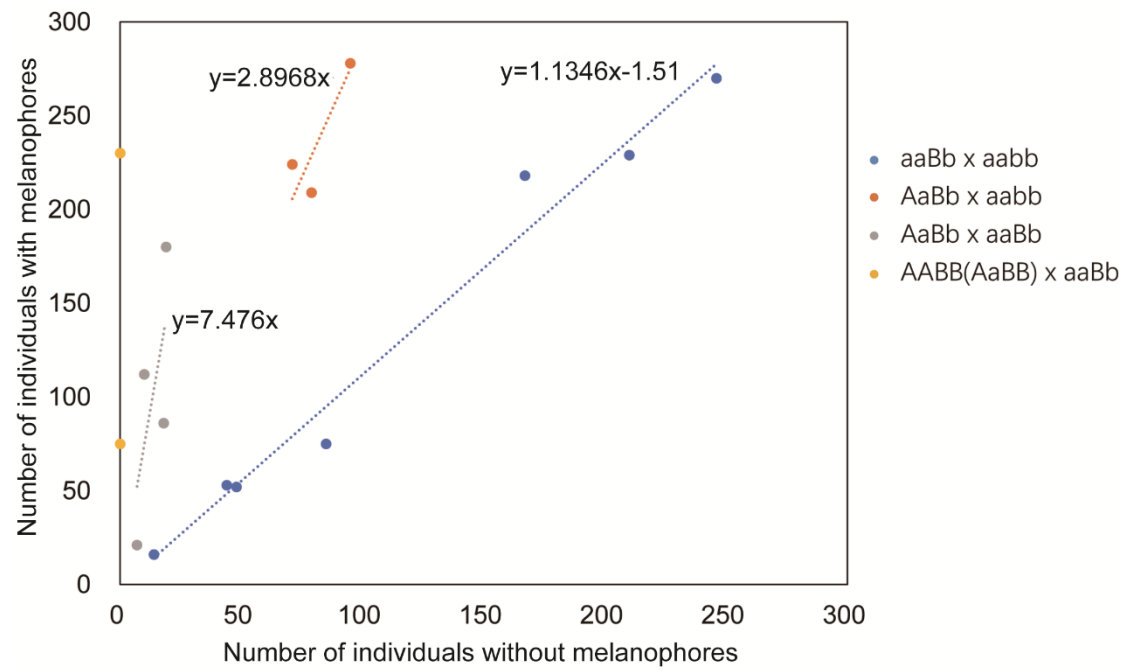
Xu P, Xu J, Liu GJ, et al. 2019. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nature Communications*, **10**(1): 4625.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**(S2): W265–W268.



Supplementary Figure S2 Distribution of SNPs and indels in genome
Outer circle is distribution of indels, inner circle is distribution of SNPs.

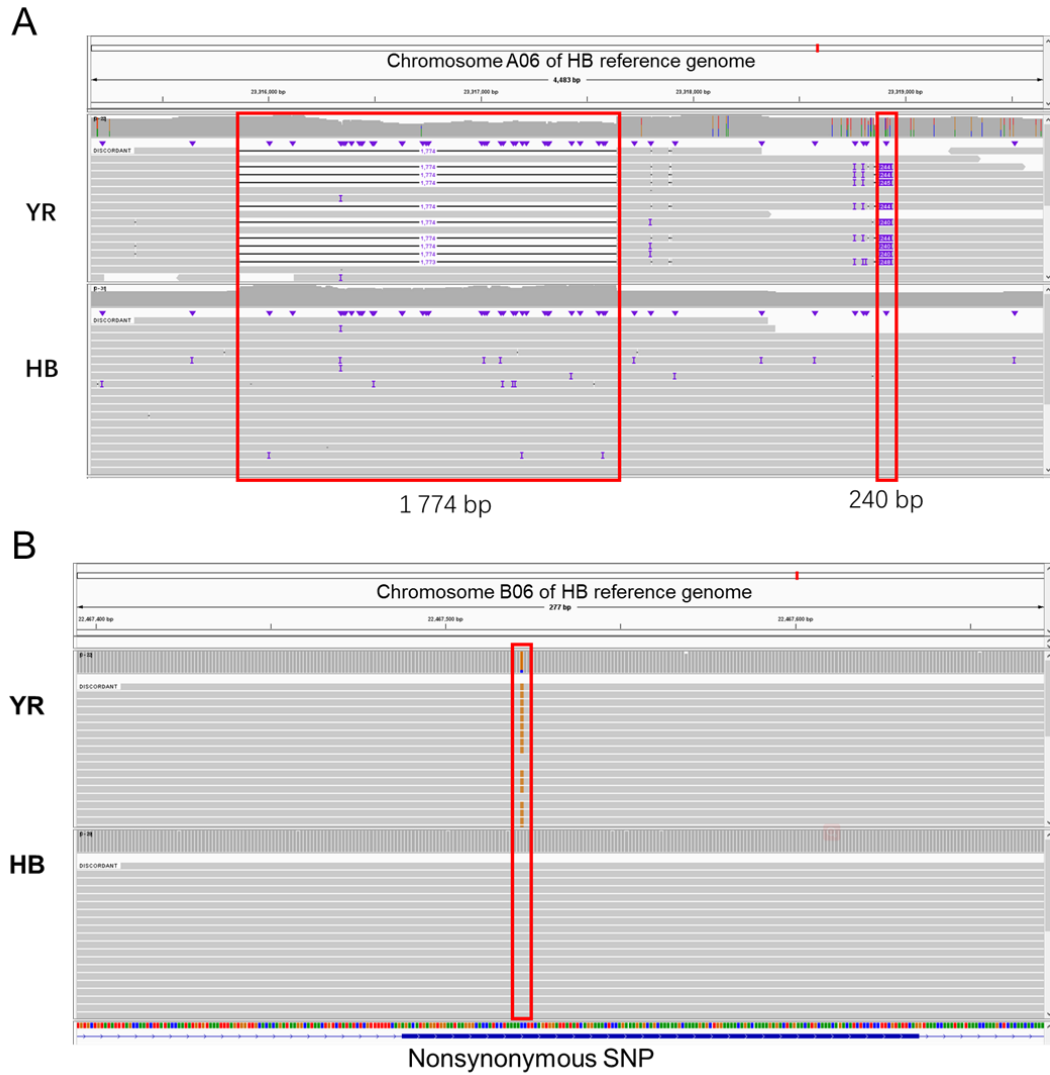




Supplementary Figure S4 Number of individuals with and without melanophores and segregation ratio of backcross family

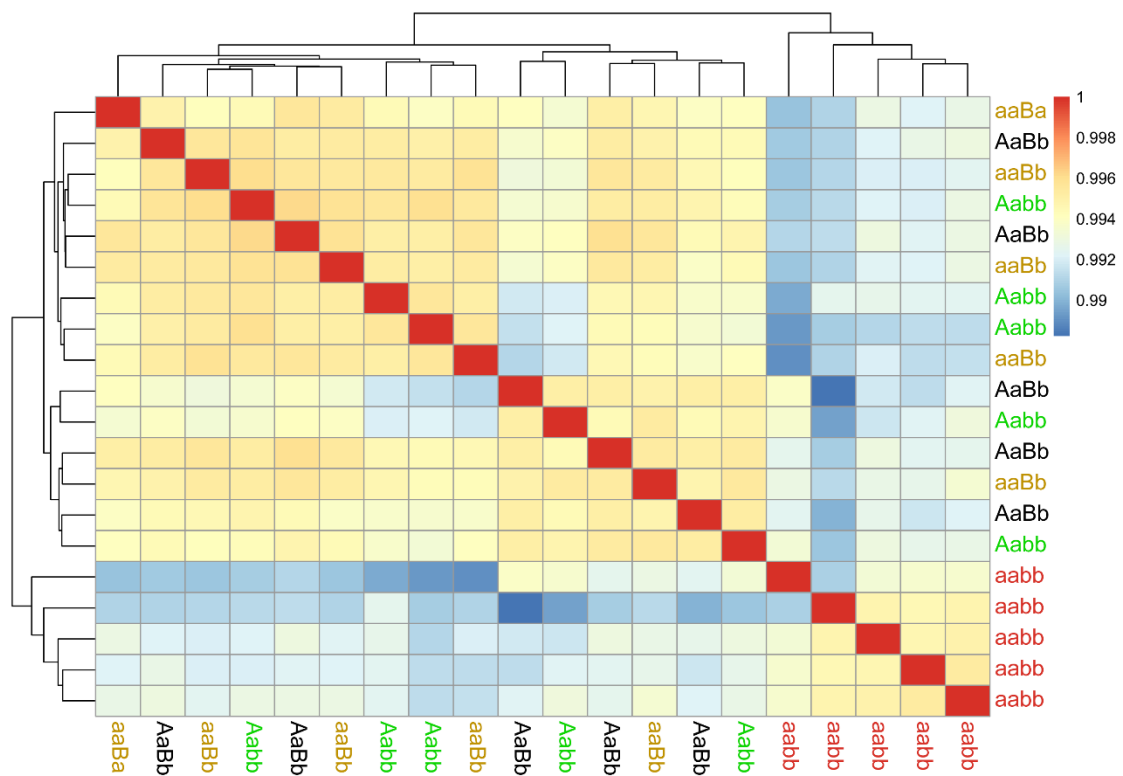


Supplementary Figure S5 Backcross family at 1 month old (A) and 7 months old (B)



Supplementary Figure S6 Visualization of *mitfa* mutation by IGV

Input was a BAM file generated by mapping PacBio reads of HB and YR to corrected genome sequence. A: Larger-scale SVs in *mitfa* of chromosome A06. B: Nonsynonymous SNPs in *mitfa* of chromosome B06.



Supplementary Figure S7 Heatmap of expression patterns with clustering trees
 Genotype of each sample is labeled. Red to blue indicates correlation among samples.

Supplementary Tables

Supplementary Table S1 The resource of species used in the phylogenetic relationship analysis

Supplementary Table S2 Primers used to amplify the divergent genomic regions between YR and HB

Supplementary Table S3 Gene specific primers used to amplify the whole-length of mitfa_A06

Supplementary Table S4 Sequencing data of *C. carpio wuyuanensis* for genome assembly and annotation

	Clean data(Gb)	Depth	Mean read length(bp)
HiFi reads	37.8	23X	15,321
Hi-C reads	221	134X	150

Supplementary Table S5 Statistics of *C. carpio wuyuanensis* genome assembly

Statistic type	Hifi assembly	Hi-C assembly	Chromosome
N50	19,736,777	29,344,324	29,512,493
N90	3,696,235	21,775,851	23,884,888
number of contigs	809	957	50
Maximum length	33,080,274	48,996,568	48,996,568
Mean length	58,087	40,500	20,014,000
Total length	1,609,897,486	1,610,108,486	1,499,080,759

Supplementary Table S6 Assessment of genome completeness by BUSCO

Type	Genome number	Percentage (%)	Proteins number	Percentage (%)
Complete BUSCOs (C)	3589	98.60	3447	94.70
Complete and single-copy BUSCOs (S)	1196	32.90	1431	39.30
Complete and duplicated BUSCOs (D)	2393	65.70	2016	55.40
Fragmented BUSCOs (F)	22	0.60	85	2.30
Missing BUSCOs (M)	29	0.80	108	3.00
Total BUSCO groups searched	3640		3640	

Results from dataset actinopterygii_odb10

Supplementary Table S7 Statistics of repeat elements in genome

Type	Number of elements	Length (bp)	Percentage of sequence (%)
DNA transposons	2 322 310	327 872 931	21.87
SINEs:	40 667	6 748 784	0.45
LINEs:	222 543	69 556 112	4.64
LTR elements:	250 998	75 412 257	5.03
Rolling-circles	96 777	27 542 134	1.84
Unclassified:	735 949	134 214 657	8.95
Satellites	126 136	22 814 998	1.52
Total			44.30

Supplementary Table S8 Gene number annotated in different databases

Database	Gene number
Swiss-Prot	44 209
NR	58 650
TrEMBL	57 315
InterProScan /GO	42 282
KEGG	27 154
All	58 842

Supplementary Table S9 GO term enrichment of genes under positive selection

Supplementary Table S10 Mapping details of 34 resequencing samples

Supplementary Table S11 Divergent regions between YR and HB

Supplementary Table S12 Annotation of genes in divergent regions

Supplementary Table S13 Segregation ratio of BC1F1 families

Supplementary Table S14 Significant SNPs in GWAS of body color

Supplementary Table S15 Genotype of individuals in natural, F1, and BC1F1 populations

Supplementary Tables S1, S2, S3, S9-S15 are listed as a separate excel file due to their large size.