ZR Zoological Research

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 Xue[3], Zhou Jiang[1,3], Bao-Hua Chen[1,2], Cheng-Yu Li[1,2], Qian He[1,2], Jian-Xin Feng[4], 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

[*]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) (Abrusán 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 (Brůna 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 (Brůna 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

Abrusán 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.

Brůna 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.

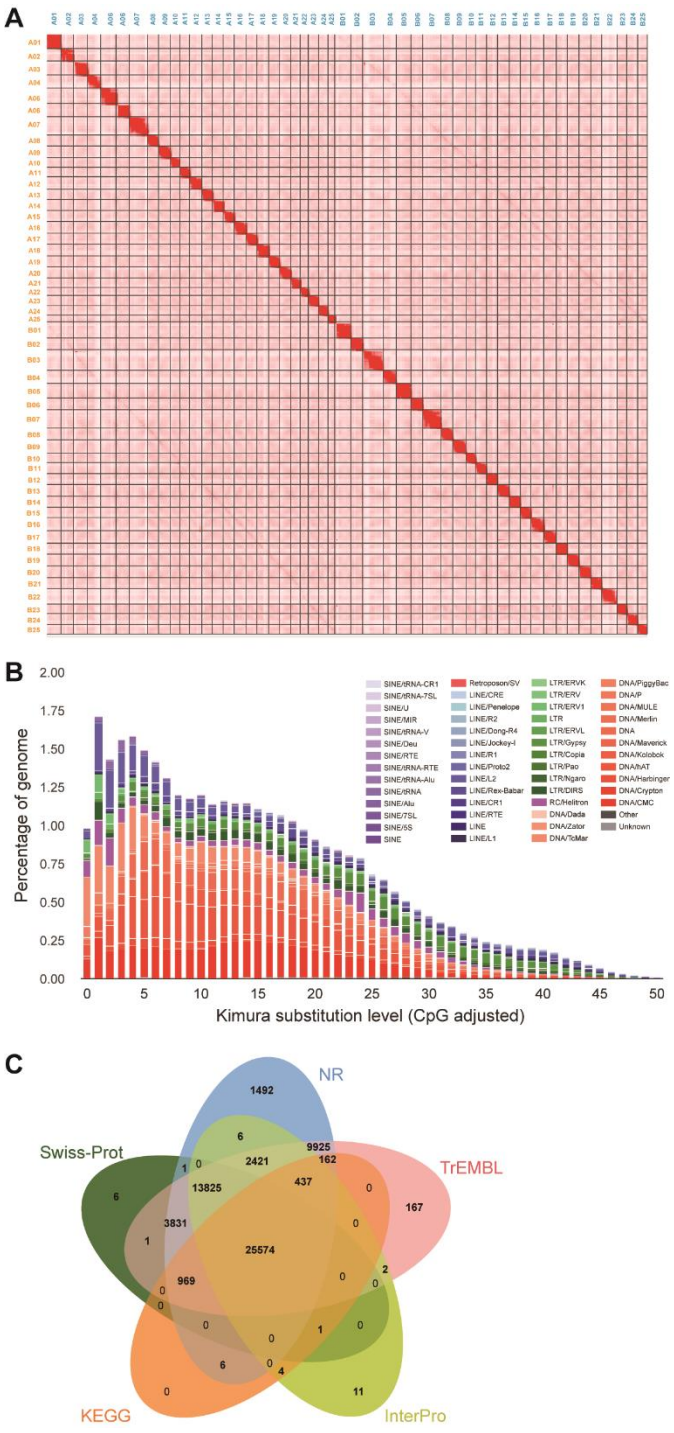
Flynn JM, Hubley R, Goubert C, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(17): 9451–9457.

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, Rødland EA, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**(9): 3100–3108.

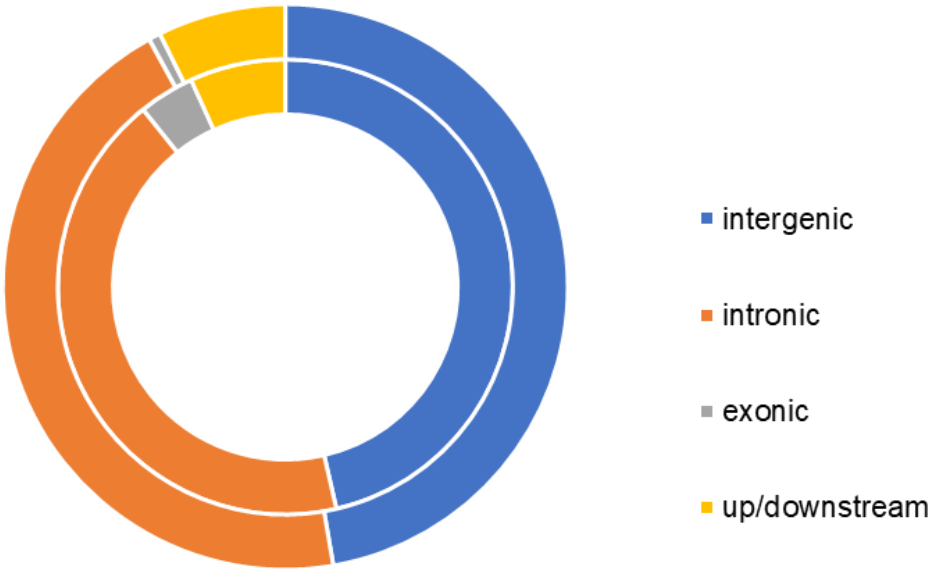
Levy C, Khaled M, Fisher DE. 2006. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine*, 12(9): 406-414.

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.

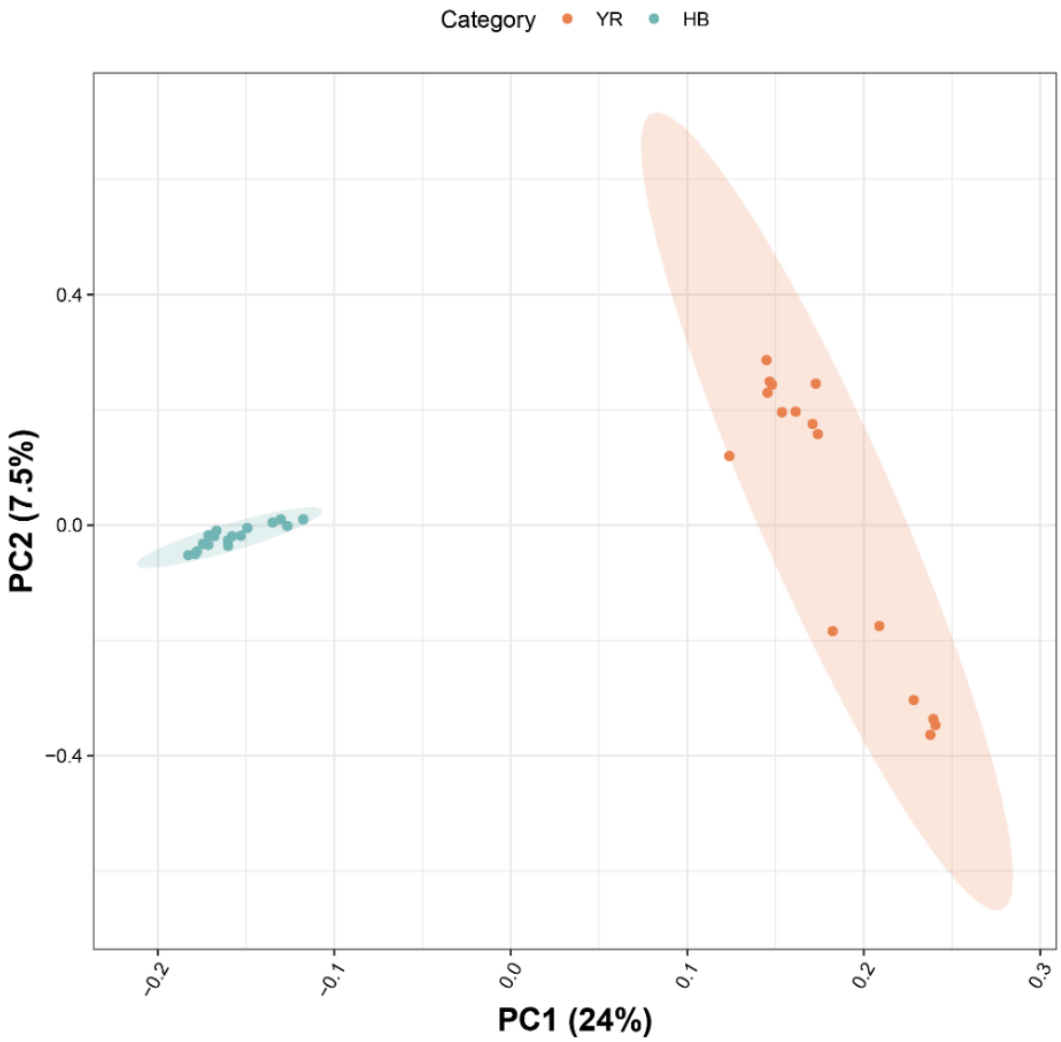


Supplementary Figure S1 Genomic characteristics of *C. carpio wuyuanensis*

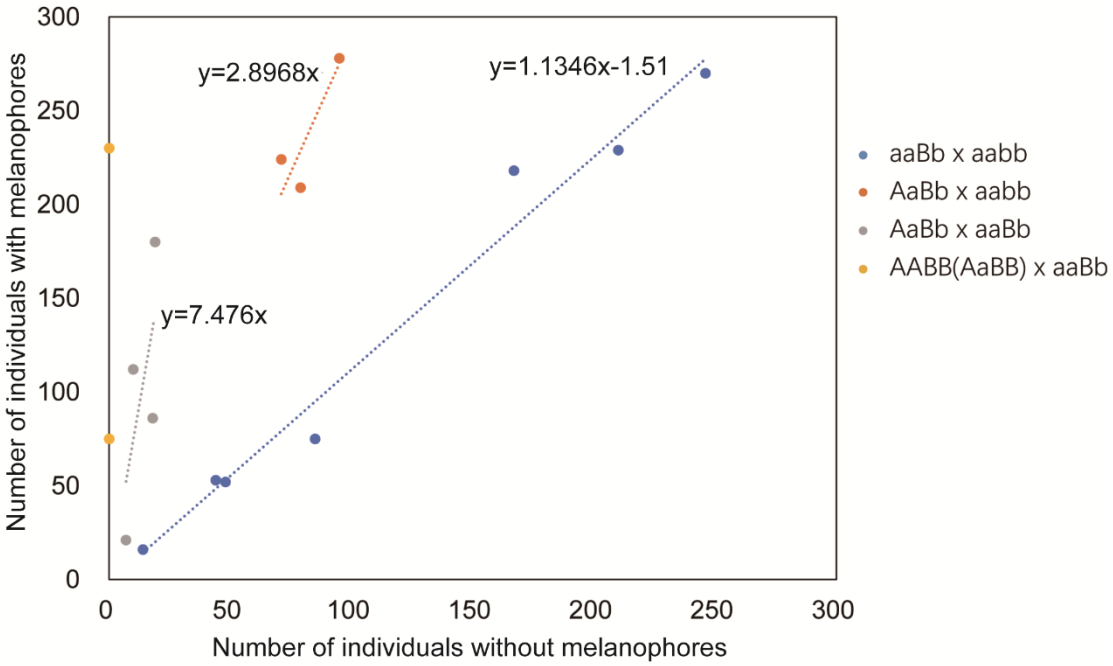
A: Heatmap of interactions of 50 chromosomes in *C. carpio wuyuanensis* genome using Hi-C data. B: Distribution of divergence rate for each type of TE in *C. carpio wuyuanensis* genome. Y-axis is percentage of TEs in genome, X-axis is Kimura divergence. Values from 0 to 50 reflect older copies (right side) and new copies (left side). C: Functional annotation of genes in different databases.



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

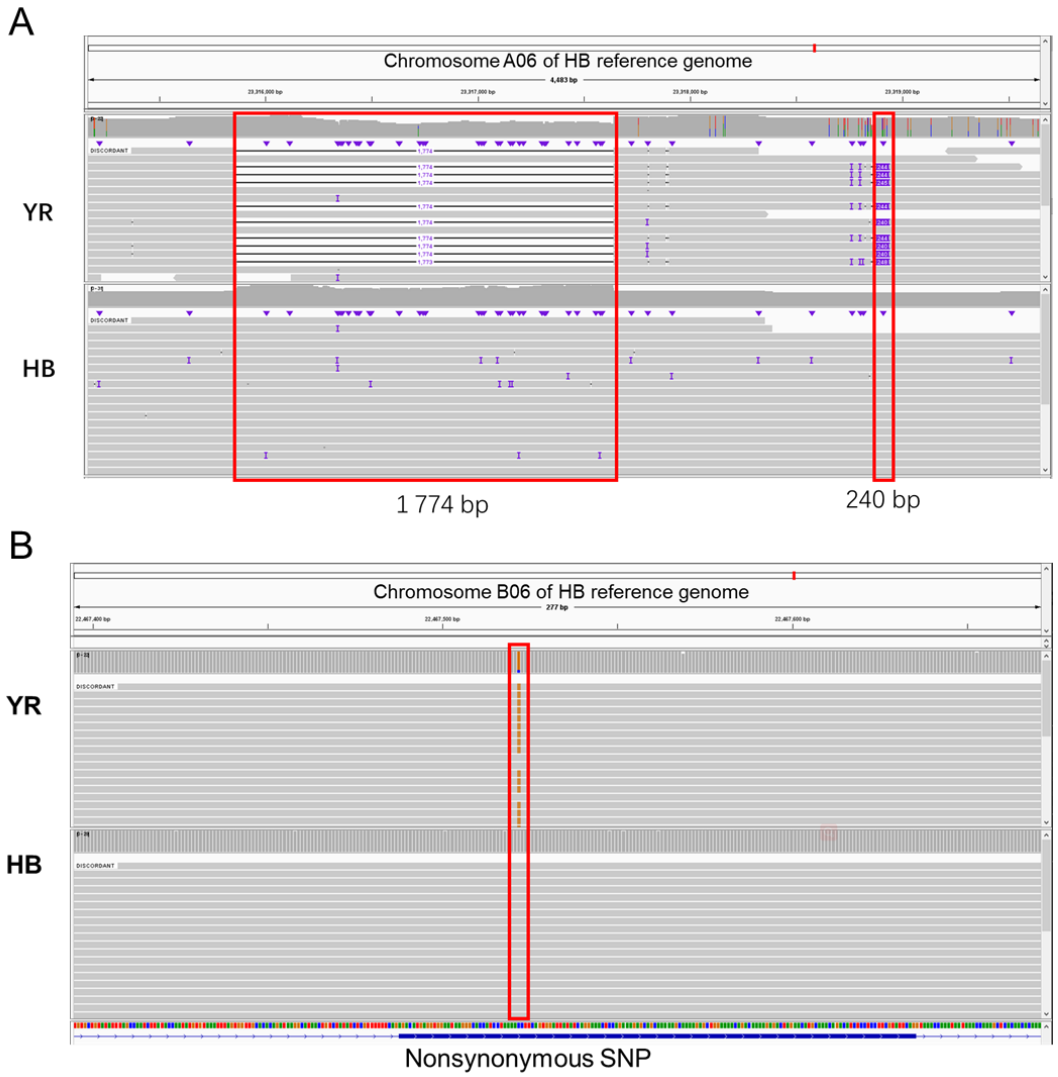


Supplementary Figure S3 PCA structure of natural Hebao red carp (HB) and Yellow River carp (YR) used in whole-genome resequencing



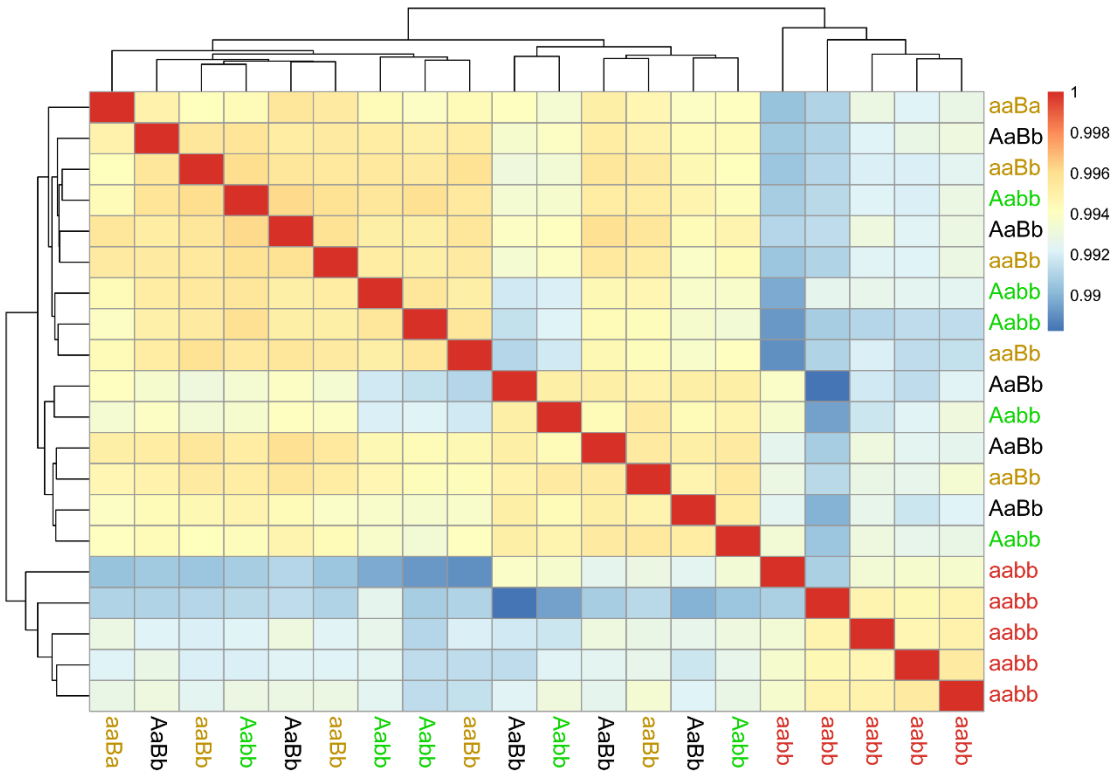
Supplementary Figure S4 Number of individuals with and without melanophores and segregation ratio of backcross family

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.