Supplementary Materials

Supplementary Materials and Methods

Sample collection and whole-genome resequencing

In our previous study, we resequenced and analyzed six wild-type and 37 ornamental bettas, including four giant fighting fish (Wang et al., 2021). Here, we sequenced the whole genomes of 11 additional giant bettas randomly collected from different colored strains, with ~12× coverage of sequencing reads (~5.3 Gb 2×150 bp reads for each sample). All procedures for fish handling strictly followed the guidelines of the Institutional Animal Care and Use Committee (IACUC) of Temasek Life Sciences Laboratory, Singapore (Approval no. TLL (F)-16-003).

Whole-genome resequencing and variant calling

Genomic DNA was extracted using Qiagen DNeasy Blood & Tissue kits (Qiagen, Germany). DNA was then quantified using Qubit (Invitrogen, USA). DNA (2 μg) from each sample was used for library construction with 550 bp inserts for whole-genome sequencing using an Illumina DNA PCR-Free Prep Kit v2 (Illumina, USA). Libraries were sequenced inhouse using the Illumina NextSeq500 platform (Illumina, USA) for 2×150 bp paired-end reads.

Raw reads were cleaned using process_shortreads in the Stacks package v1.45 (Catchen et al., 2013) with default parameters. Cleaned reads were aligned to the fighting fish reference genome using BWA-mem v0.7 (Li & Durbin, 2010). Variant calling was conducted according to Picard/GATK v4.0 best practice workflows (DePristo et al., 2011) and our previous study (Wang et al., 2021). SNPs were filtered using the following parameters: "QD<2.0 || FS>60.0 || MQ<40.0 || MQRankSum<-12.5 || ReadPosRankSum<-8.0 || SOR>4.0", while indels with "QD<2.0 || FS>200.0 || ReadPosRankSum<-20.0 || SOR>10.0". Variants were further filtered under "--minDP 6, --max-missing 0.85, --maf 0.05" with VCFtools v0.1.15 (Danecek et al., 2011), with 3 582 429 genotypes finally retained for further analysis.

Analyzing genetic diversity and identifying selection signatures

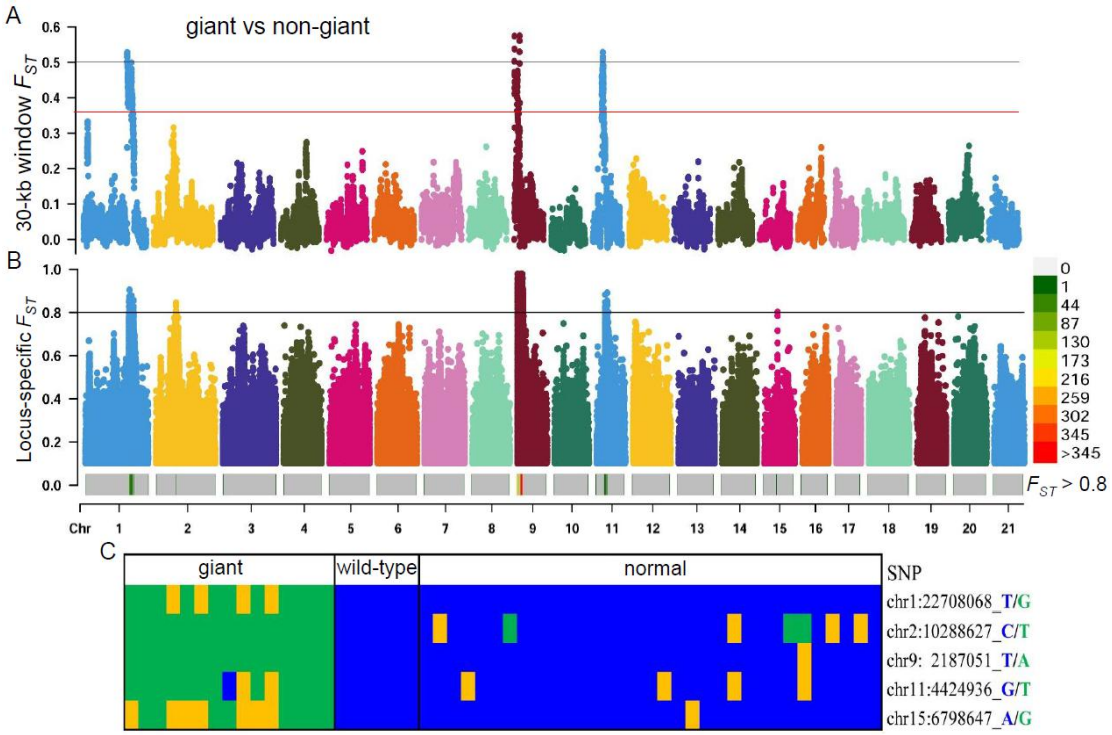
Genetic diversity, in terms of nucleotide diversity (*Pi*), was analyzed using VCFtools v0.1.15 (Danecek et al., 2011), with a 100 kb sliding window. Pairwise linkage disequilibrium (LD) between variants was assessed in terms of R^2 using PopLDdecay v1 (Zhang et al., 2019). Population structure was studied using principal component analysis (PCA) with Plink v2.0 (Purcell et al., 2007) and ADMIXTURE v1.3.0 (Alexander et al., 2009) to infer ancestral genetic clusters of the fish.

Loci were screened using a genome-wide F_{ST} scan between giant mutants (15) and non-giant bettas, including wild-type fish (39 individuals), at the individual variant level and with a 30 kb window size and 15 kb step, according to our previous study (Wang et al., 2021). Genome-wide selection signatures were examined using Tajima's *D* (Tajima, 1989). As selective sweeps eliminate rare alleles, genomic regions under selection show a reduced Tajima's *D* value. Genome-wide hard and soft

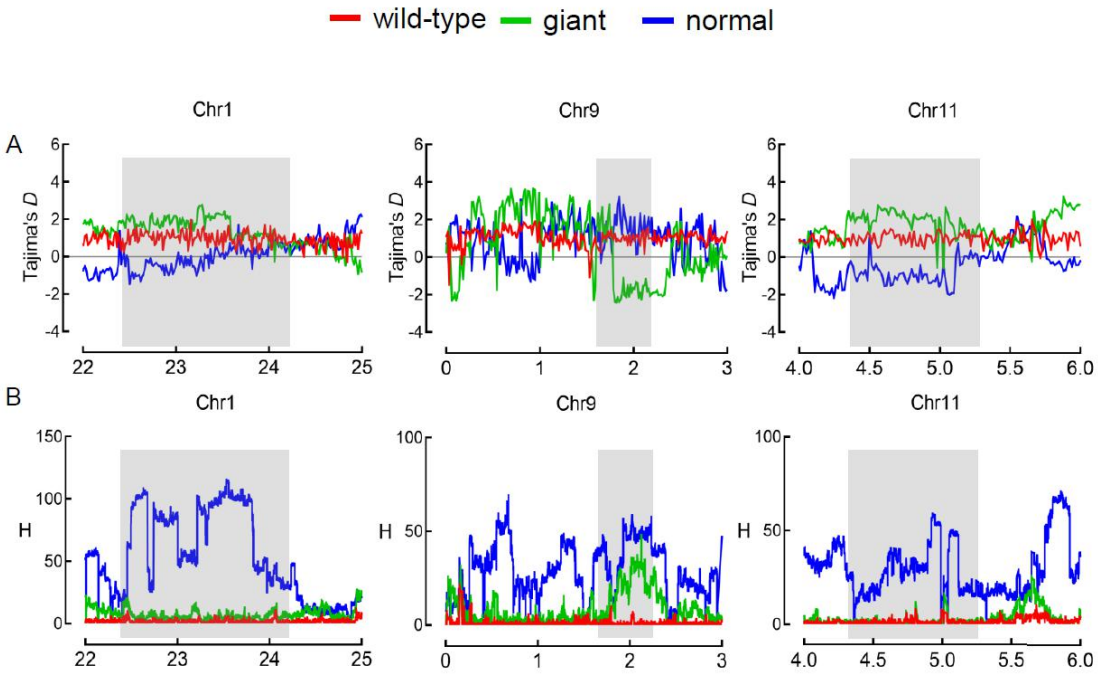
selective sweeps were further screened by H-scan (Schlamp et al., 2016) using default parameters. Variants within the genomic regions of selective sweeps were annotated based on the whole-genome annotations of the fighting fish *Betta splendens* (Wang et al., 2021) and analyzed in combination with expression data.

Transcriptome analysis in brain and muscle

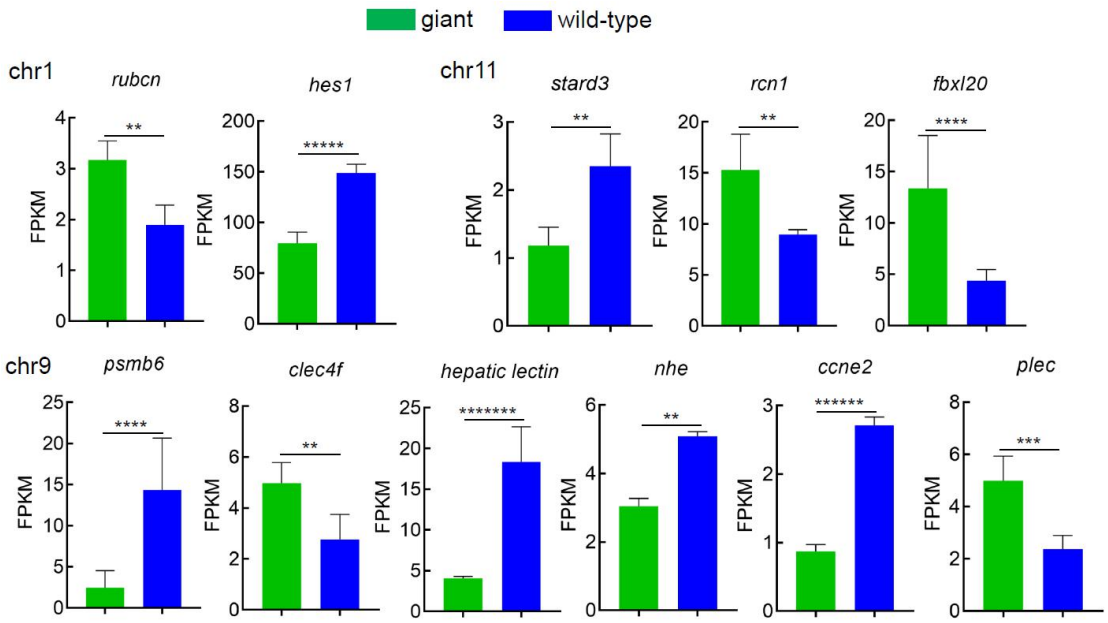
Total RNA from brain and muscle tissue was separately isolated from three three-month-old giant bettas and three non-giant ornamental bettas using TRIzol reagent (Invitrogen, USA). We observed a much faster growth rate in the giant bettas than in the non-giant bettas at this developmental stage. RNA from each fish was assessed using gel electrophoresis and quantified with NanoDrop (Thermo Fisher Scientific, USA). Total RNA (1 μg) from each sample was used for mRNA library construction with a Illumina TruSeq RNA Library Prep Kit v2 (Illumina, USA), according to the manufacturer's instructions. Libraries were quantified using a KAPA Library Quantification Kit (Roche, Switzerland) and then sequenced inhouse (2×75 bp reads) using the Illumina NextSeq500 platform (Illumina, USA). Raw sequencing reads were cleaned using the *process_shortreads* program in the Stacks package v1.45 (Catchen et al., 2013), with the following parameters (-c -q). Average number of clean reads for each sample was ~45 million (Supplementary Table S1). The clean reads were aligned to the reference genome sequences (Wang et al., 2021) using STAR v2.5.2b (Dobin et al., 2013), with default parameters. The program HTSeq-count v0.9.1 (Anders et al., 2015) was then used to quantify the expression levels of annotated protein-coding genes. The program EdgeR v3.14 (Robinson et al., 2010) was used to normalize the relative expression levels of transcripts across samples. Genes with FPKM (fragments per kilobase of transcript per million mapped reads) <1 were removed from further analysis. Transcripts with fold-change >1.5 and significance value $P<0.01$ for exact tests were considered as DEGs.



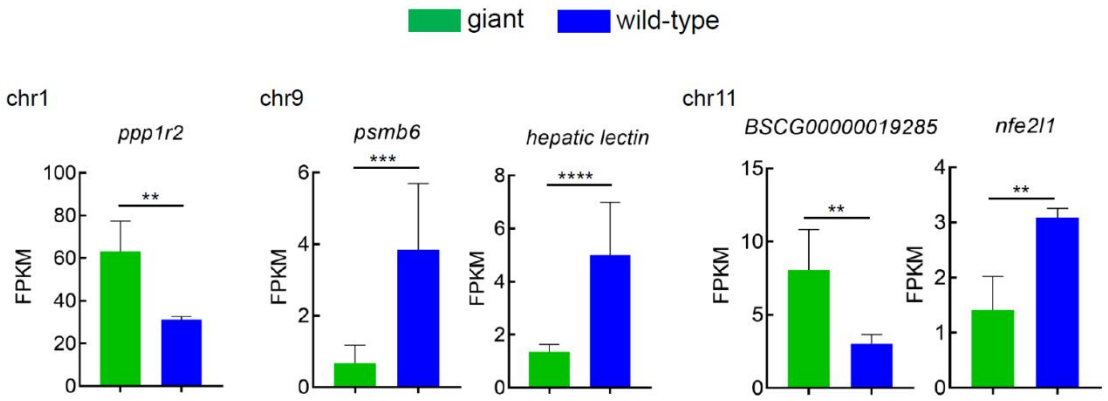
Supplementary Figure S1 Genome scan of giant (sized) loci between giant and non-giant bettas. **A**, F_{ST} scan based on 30 kb window size between giant and non-giant bettas, with F_{ST} cutoff value in upper 1% percentile of null distribution (0.36) and at 0.5 level indicated with red and gray lines, respectively. **B**, F_{ST} scan based on individual variants between giant and non-giant bettas, with cutoff value of 0.8 denoted with a gray line. **C**, Allele frequency in giant, wild-type, and normal bettas for peak SNPs within five highly differentiated regions, as revealed by F_{ST} scan based on individual variants. Green and blue indicate homozygous allele, yellow indicates heterozygous alleles.



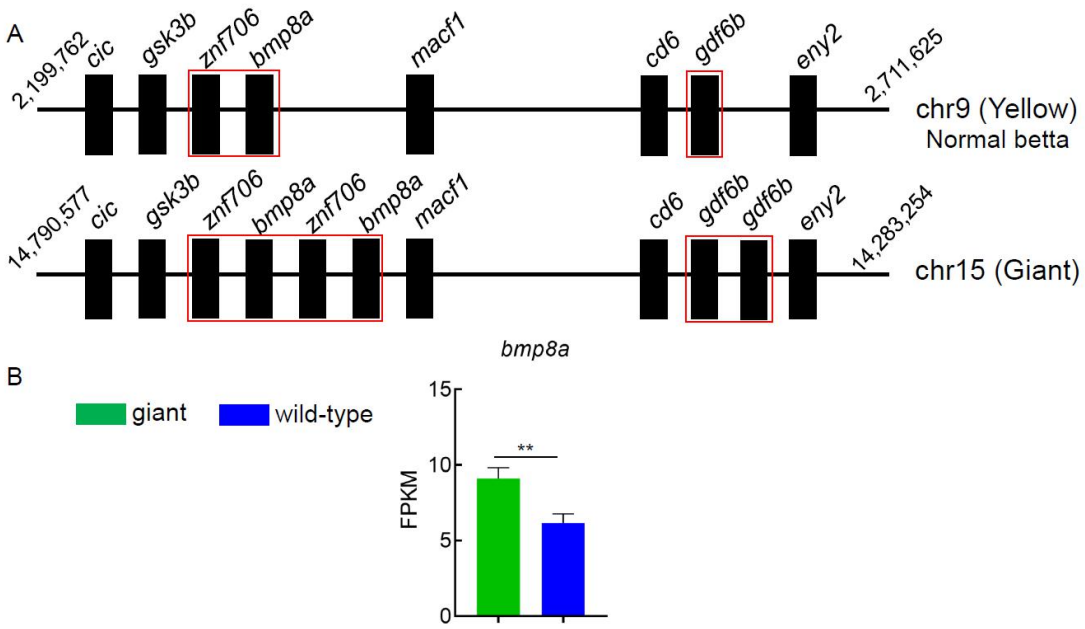
Supplementary Figure S2 Selection signatures at three major genomic regions highly associated with giant body size in fighting fish. **A**, Distribution of Tajima's *D* across three genomic regions at chr1, chr9, and chr11, respectively, with highly differentiated regions between giant and non-giant bettas highlighted in gray. **B**, Selective sweeps identified by H-scan across three genomic regions at chr1, chr9, and chr11, respectively, with highly differentiated regions between giant and non-giant bettas highlighted in gray.



Supplementary Figure S3 Differentially expressed genes (DEGs) identified in brain transcriptomes between giant (n=3) and normal ornamental bettas (wild-type, n=3). Two, three, and six genes located in genomic regions under selection on chr1, chr11, and chr9, respectively, were identified as DEGs. *P*-values for exact test examined using EdgeR are provided (**, $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$; *****, $P<0.00001$; ******, $P<0.000001$).



Supplementary Figure S4 Differentially expressed genes (DEGs) identified in muscle transcriptomes between giant (n=3) and normal ornamental bettas (wild-type, n=3). One, two, and two genes located in genomic regions under selection on chr1, chr9, and chr11, respectively, were identified as DEGs. *P*-values for exact test examined using EdgeR are provided (**, $P<0.01$; ***, $P<0.001$; ****, $P<0.0001$).



Supplementary Figure S5 Genomic structural variations between giant and normal ornamental bettas (wild-type). **A**, Two segmental chromosomal duplications were identified within genomic region on chr9, which were under selection and associated with giant body size. One duplication contained two genes, *znf706* and *bmp81*, the other duplication contained *gdf6b*. Gene duplications were determined by genomic synteny between normal ornamental (wild-type) and giant betta (corresponding to chr15 in genome assembly) genome assemblies. Duplicated genes are highlighted in red. **B**, Overall expression of *bmp8a* transcripts in brain between giant (n=3) and normal ornamental bettas (wild-type, n=3), with fold-change of 1.48. *P*-values for exact test examined using EdgeR are provided (**, $P<0.01$).

Supplementary Table S1 Summary statistics of clean reads and mapping efficiency for brain and muscle mRNA sequencing samples from giant and normal bettas.

Betta	Sample	Clean reads	Uniquely mapped reads
Giant	Brain_1	49,208,605	88.71%
Giant	Brain_2	43,627,921	86.22%
Giant	Brain_3	47,145,011	84.49%
Giant	Muscle_1	44,736,862	92.40%
Giant	Muscle_2	40,277,879	90.81%
Giant	Muscle_3	42,324,009	91.28%
Normal	Brain_1	42,219,316	87.73%
Normal	Brain_2	57,746,586	70.38%
Normal	Brain_3	45,613,897	81.82%
Normal	Muscle_1	45,417,771	92.42%
Normal	Muscle_2	44,864,062	86.34%
Normal	Muscle_3	45,310,861	85.41%

REFERENCES

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(9): 1655–1664.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2): 166–169.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**(11): 3124–3140.

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

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**(5): 491–498.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1): 15–21.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**(5): 589–595.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, 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.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1): 139–140.

Schlamp F, Van Der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, **25**(1): 342–356.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**(3): 585–595.

Wang L, Sun F, Wan ZY, Ye BQ, Wen YF, Liu HM, et al. 2021. Genomic basis of striking fin shapes and colors in the fighting fish. *Molecular Biology and Evolution*, **38**(8): 3383–3396.

Zhang C, Dong SS, Xu JY, He WM, Yang TL. 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **35**(10): 1786–1788.