GigaScience, 10, 2021, 1–12

https://doi.org/10.1093/gigascience/giab075
Data Note

DATA NOTE

De novo screening of disease-resistant genes from the chromosome-level genome of rare minnow using CRISPR-cas9 random mutation

Rong Huang[1,*,†], Mijuan Shi[1,†], Lifei Luo[1], Cheng Yang[1], Mi Ou[1], Wanting Zhang[1], Lanjie Liao[1], Yongming Li[1], Xiao-Qin Xia[1], Zuoyan Zhu[1] and Yaping Wang[1,2,*]

[1]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China and [2]Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing 100101, China

[*]**Correspondence address.** Rong Huang, 7 Donghu South Road, Wuchang District, Wuhan, Hubei Province, China. Tel: +86-027-68780081; Fax: +86-027-68780123; E-mail: huangrong@ihb.ac.cn https://orcid.org/0000-0002-1118-3450; Yaping Wang, E-mail: wangyp@ihb.ac.cn https://orcid.org/0000-0002-2691-8261
[†]These authors contributed equally to this work.

Abstract

Background: Mutants are important for the discovery of functional genes and creation of germplasm resources. Mutant acquisition depends on the efficiency of mutation technology and screening methods. CRISPR-Cas9 technology is an efficient gene editing technology mainly used for editing a few genes or target sites, which has not been applied for the construction of random mutant libraries and for the *de novo* discovery of functional genes. **Results:** In this study, we first sequenced and assembled the chromosome-level genome of wild-type rare minnow (*Gobiocypris rarus*) as a susceptible model of hemorrhagic disease, obtained a 956.05 Mb genome sequence, assembled the sequence into 25 chromosomes, and annotated 26,861 protein-coding genes. Thereafter, CRISPR-Cas9 technology was applied to randomly mutate the whole genome of rare minnow with the conserved bases (TATAWAW and ATG) of the promoter and coding regions as the target sites. The survival rate of hemorrhagic disease in the rare minnow gradually increased from 0% (the entire wild-type population died after infection) to 38.24% (F3 generation). Finally, 7 susceptible genes were identified via genome comparative analysis and cell-level verification based on the rare minnow genome. **Conclusions:** The results provided the genomic resources for wild-type rare minnow, and confirmed that the random mutation system designed using CRISPR-Cas9 technology in this study is simple and efficient and is suitable for the *de novo* discovery of functional genes and creation of a germplasm resource related to qualitative traits.

Keywords: rare minnow; genome; CRISPR-Cas9; mutant; germplasm resource

Introduction

The rare minnow (*Gobiocypris rarus*) belongs to the order Cypriniformes and family Cyprinidae, and it has the advantages of a small body, fast reproduction, and easy feeding. It is more sensitive to some pollutants compared to zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*). For example, the sensitivity of rare minnow to 17α-ethinylestradiol and pentachlorophenol is higher than that of zebrafish, and its sensitivity to ethinylestradiol is

Received: 12 May 2021; **Revised:** 5 September 2021; **Accepted:** 25 October 2021

higher than that of medaka [1–3]. Therefore, it has been widely used in genetics, physiology, biological monitoring, toxicity testing, and other fields [4].

The mortality rate of rare minnow infected with grass carp reovirus (GCRV) is 100% [5]. Grass carp (*Ctenopharyngodon idellus*), which also belongs to the family Cyprinidae, is one of the most important freshwater fishes worldwide. The mortality of grass carp hemorrhagic disease caused by GCRV infection is >s80% [6], which poses a great threat to the development of the aquaculture industry. Rare minnow, similar to grass carp, is highly sensitive to GCRV, which makes it an ideal model for studying grass carp hemorrhagic disease and exploring germplasm resources.

Research on efficient mutation methods is a prerequisite for constructing an ideal animal model. Traditional physical and chemical mutagenesis methods mainly cause genomic point mutations [7–9], which cannot be distinguished from natural single-nucleotide polymorphism (SNP) mutations, leading to considerable difficulties when performing comparative analyses of the subsequent functional genomes. Traditional transposon mutations have a strong selectivity for the mutation region of the receptor genome, and they are unable to achieve random mutations for all genes [10, 11]. Efficient and easy-to-detect mutation methods are important for obtaining mutants and for exploring new germplasm resources.

CRISPR-Cas9 technology is an efficient gene editing technology mainly used for editing a few genes or target sites [12, 13]. It is also used to study mutant libraries. In previous research on human cells and rice, the main way to construct a mutant library was to design small guide RNA (sgRNA) of all candidate genes, then mix all sgRNAs, and select target mutants after knockout [14–17]. In this method, a large number of sgRNAs needed be designed at a high cost, and it is only suitable for the construction of a mutant library with known candidate genes. To date, efficient CRISPR-Cas9 technology has not been applied for the construction of random mutant libraries and for the *de novo* discovery of functional genes.

All wild-type rare minnows die after being infected with GCRV, which provides us with an excellent mutant screening material for GCRV resistance. That is, the individual who can survive after infection is likely to be an individual with successful mutations. In this study, we assembled a high-quality genome of rare minnow, and then used CRISPR-Cas9 technology to randomly mutate the complete genome of the rare minnow and obtained a mutant population with GCRV resistance traits. Next, we obtained 7 hemorrhagic disease–susceptible genes via genome comparative analysis and experimental verification. The results not only provided genomic resources for research on rare minnow but also facilitated the establishment of a simple and feasible method for creating random genomic mutations, which are suitable for the exploration of functional genes and new germplasm resources.

Results

Genome assembly and annotation

To initially evaluate the genome of rare minnow, we obtained 124.20 Gb raw data and 121.11 Gb clean data after routine filtering. On the basis of the k-mer ($k = 21$) analysis method, the genome size was estimated to be 943.44 Mb, the heterozygosity rate was 0.41%, and the repetition rate was 35.82%. The results showed that the genome of rare minnow is a simple genome rather than a complex one.

After filtering the Pacific Biosciences data, 106.88 Gb subreads were obtained. The mean length of the subreads was 13,088.61 bp, and the N50 was 21,231.00 bp. After the subread data were self-corrected, the genome was assembled into a size of 960.27 Mb, consisting of 858 contigs with an N50 of 5.46 Mb. Using the 121.11 Gb next-generation sequencing (NGS) data obtained previously, the assembled genome was corrected again, and the final size of the corrected genome was 959.10 Mb and the contig N50 was 5.46 Mb.

We then obtained 103.47 Gb clean data from Hi-C library sequencing. After filtering and evaluating with HIC-pro, 181,735,272 pairs of uniquely mapped reads were obtained, of which 123,097,523 pairs (67.73%) were valid interaction pairs. Based on valid interaction pairs, a 924.69 Mb sequence composed of 345 contigs was assembled into 25 chromosomes, accounting for 96.41% of the total sequence length (Table 1). The completeness of the genome was 90.10% on the basis of the BUSCO evaluation. A heat map describing the contact matrix was constructed to evaluate the accuracy of the Hi-C assembly (Fig. 1a). The interaction signals obtained from the heat map could help to clearly distinguish the 25 chromosomes, indicating that the assembly effect of the genome was very good.

We annotated 43.14% of the rare minnow genome as repetitive sequences (Supplementary Table S1). In addition, 36,387 messenger RNAs were annotated, corresponding to 26,861 genes, of which 4,957 genes had alternative splicing transcripts. The average length of the longest coding sequence of all genes was 1.82 kb, which was close to the average length of zebrafish and higher than that of grass carp and blunt snout bream (*Megalobrama amblycephala*) (Supplementary Table S2).

Evolutionary analysis of the genomes

Through cluster analysis of gene families of 13 species, 20,723 gene families were obtained, among which 2,376 were shared gene families, 17,364 were shared genes, and 144 were single-copy gene families. A phylogenetic tree was constructed using all single-copy gene families (Fig. 1B). Figure 1B shows that 4 Cyprinidae species were clustered into 1 branch; the differentiation time of rare minnow and grass carp was 67.04 MYA (Fig. 1B).

Collinearity analysis showed that 18,968 similar genes were located in 97 supercontigs of grass carp. The linkage groups of the 97 supercontigs of grass carp were mapped to the genome of the rare minnow (Fig. 1C). Chromosomes 1 and 21 (LG1 and LG21) of rare minnow correspond to LG13 of grass carp, and the degree of gene collinearity of the 2 species was very high (Fig. 1C).

Anti-hemorrhagic model of rare minnow

Twenty-six sgRNAs were mixed with Cas9 protein and injected into ~8,000 single-cell embryos. Finally, 3,126 two-month-old P0 mutants were obtained. Among them, 3,000 were used in the GCRV infection experiment. The results showed that 2,993 died and 7 survived, for a survival rate of 0.23%; conversely, all 351 individuals of the control group (wild-type) died, demonstrating a survival rate of 0%. During the course of the disease, the dead individuals in the mutation and control groups exhibited a red body surface, showing obvious hemorrhagic symptoms (Fig. 2A).

To eliminate the difference in survival rate caused by experimental errors, 2 F1 families (F1-1 and F1-2) were obtained by

Table 1: Summary: statistics of the rare minnow reference genome assembly

	Contig		Scaffold	
Assembly	No.	Length (bp)	No.	Length (bp)
N50	48	5,468,461	12	36,585,240
N90	203	896,652	23	28,204,685
Max	1	25,522,336	1	53,027,249
Total	858	960,267,999	566	959,102,419
Anchored to chromosomes	694	956,050,416 (99.56%)	345	924,697,551 (96.41%)

lateral-crossing 2 surviving males from the P0 generation with wild-type females. The survival rates of F1-1 and F1-2 were 1.25% and 1.89%, respectively, and F2–F3 generation families were obtained by self-crossing, infection, and reproduction. The survival rates of the 4 F2 generation families (F2-1, F2-2, F2-3, and F2-4) were 3.33%, 4.85%, 17.02%, and 23.08%, respectively. Infection experiments of 7 F3 families (F3-1, F3-2, F3-3, F3-4, F3-5, F3-6, and F3-7) showed that the survival rates of the F3-6 and F3-7 families were 31.37% and 38.24% higher than those of the F2-4 family (the parent source of F3 families) (Fig. 2B).

During the GCRV infection, daily deaths in the F1–F3 generation mutant groups and the control group were counted and cumulative mortality curves were established (Fig. 2C–E). As shown in Fig. 2C, individuals in both the F1-2 family and the control group began to die as early as 5 days post-infection (dpi), while those in the F1-1 family began to die at 7 dpi. In 4 GCRV-infected F2 families, the mortality was higher than that of the control group at 6 dpi but was lower after 8 dpi. In addition, the survival before death of the F2 families was prolonged by 3–4 days compared with that in the control group (Fig. 2D). Among the 7 F3 generation families, 2 families (F3-1 and F3-2) died faster than the control at 5 and 6 dpi, but the death rate of all mutant families was lower than that of the control after 7 dpi. The survival before death in the F3 mutant families was prolonged by 3–10 days compared to that in the control group (Fig. 2E). Overall, compared to the control, F1, F2, and F3 mutant families exhibited delayed death induced by GCRV infection.

Screening of candidate indel loci related to hemorrhagic disease

The indel loci and genotypes of 11 datasets (C, S1, L7, T1, T2, T3, and P1–P5) were analyzed using GATK v4.1.1.0. The genotypes of the indels in the T1, T2, and T3 groups were compared with those in the same position in the control groups (C, S1, and L7), and represented by the letters T, N, and F (Fig. 3). These genotypes were then divided into 4 grades: high, moderate, low, and modified, on the basis of the contribution of these loci to gene function changes. There were 147,679 (139,632 + 1,377 + 1,971 + 1,622 + 931 + 1,125 + 1,021) indels in TTT, 2 T + 1 N, and 1 T + 2 N types (outer ring of Fig. 3). Furthermore, 147,679 loci in 5 F1 parents (P1–P5) were genotyped, and 11,668 loci with new genotypes (F0) were identified (inner ring of Fig. 3). Combined with the contribution of sites to gene function change, the contributions of 23 loci were high among the 11,668 loci (Supplementary Table S3). These 23 loci were associated with hemorrhagic diseases. The TATAWAW and ATG targets closest to these 23 loci were analyzed. The targets of 22 loci were found to be ATG, and the targets of 1 site could not be determined because both TATAWAW and ATG targets were nearby (within 20 bp) (Supplementary Table S4).

Functional verification of susceptible genes related to hemorrhagic disease

According to the genome annotation information of rare minnow, 20 genes containing 23 loci related to hemorrhagic disease were identified (Supplementary Table S5). By comparing 20 genes of rare minnow with annotation information of the grass carp genome, 23 homologous genes in grass carp were obtained (Supplementary Table S6). Small interfering RNA (siRNA) and specific primers for 23 grass carp genes were designed, and the sequences are shown in Supplementary Table S7. After the siRNAs were transfected into grass carp ovary (GCO) cells, the relative expression level of each target gene at 48 h after transfection in the siRNA-transfected cells was normalized to the expression level of the target gene at 0 h. The results indicated that 9 siRNAs had a significant inhibitory effect ($P < 0.05$) (Fig. 4A). To study the effects of siRNA knockdown on GCRV infection, these 9 siRNAs were transfected into GCO cells and infected with GCRV. RT-qPCR analysis showed that transfection of 7 siRNAs significantly reduced the copy number of GCRV in GCO cells at 32 h after transfection, compared with that in the negative control (NC) group ($P < 0.05$) (Fig. 4B). Furthermore, the titer of GCRV contained in GCO cells transfected with the 7 siRNAs was detected. It was showed that the titer decreased significantly in the GCO cell groups treated with the 7 siRNAs ($P < 0.05$) (Fig. 4C). These results suggest that these 7 genes are indeed susceptible to GCRV.

Discussion

In this study, the genome sequence and annotation information of rare minnow were obtained, providing a high-quality genome analysis platform for research and use in more fields.

In addition, we established a method for constructing a genome-wide random mutant library via the special application of CRISPR-Cas9 using rare minnow as a hemorrhagic disease–susceptible model (Supplementary Fig. S8). This method has a wide mutation range, low cost, and high efficiency and is suitable for functional genomics research and for creation of germplasm resources related to qualitative traits.

To date, some studies have used CRISPR-Cas9 technology to construct a mutant library of human cells and rice [14–17]. They designed sgRNA within the range of existing candidate genes. The advantage of this strategy is that it is helpful in detecting mutation sites; however, the disadvantage is that it requires sufficient candidate gene sequences. If there were no expected trait-related genes among the candidate genes, the expected mutant could not be obtained. The target sites in this study were designed based on the conserved bases of the gene promoter and coding region (TATAWAW and ATG) (Fig. 5), which can theoretically cover the functional region of all genes in the genome, thus increasing the abundance of mutation libraries and greatly

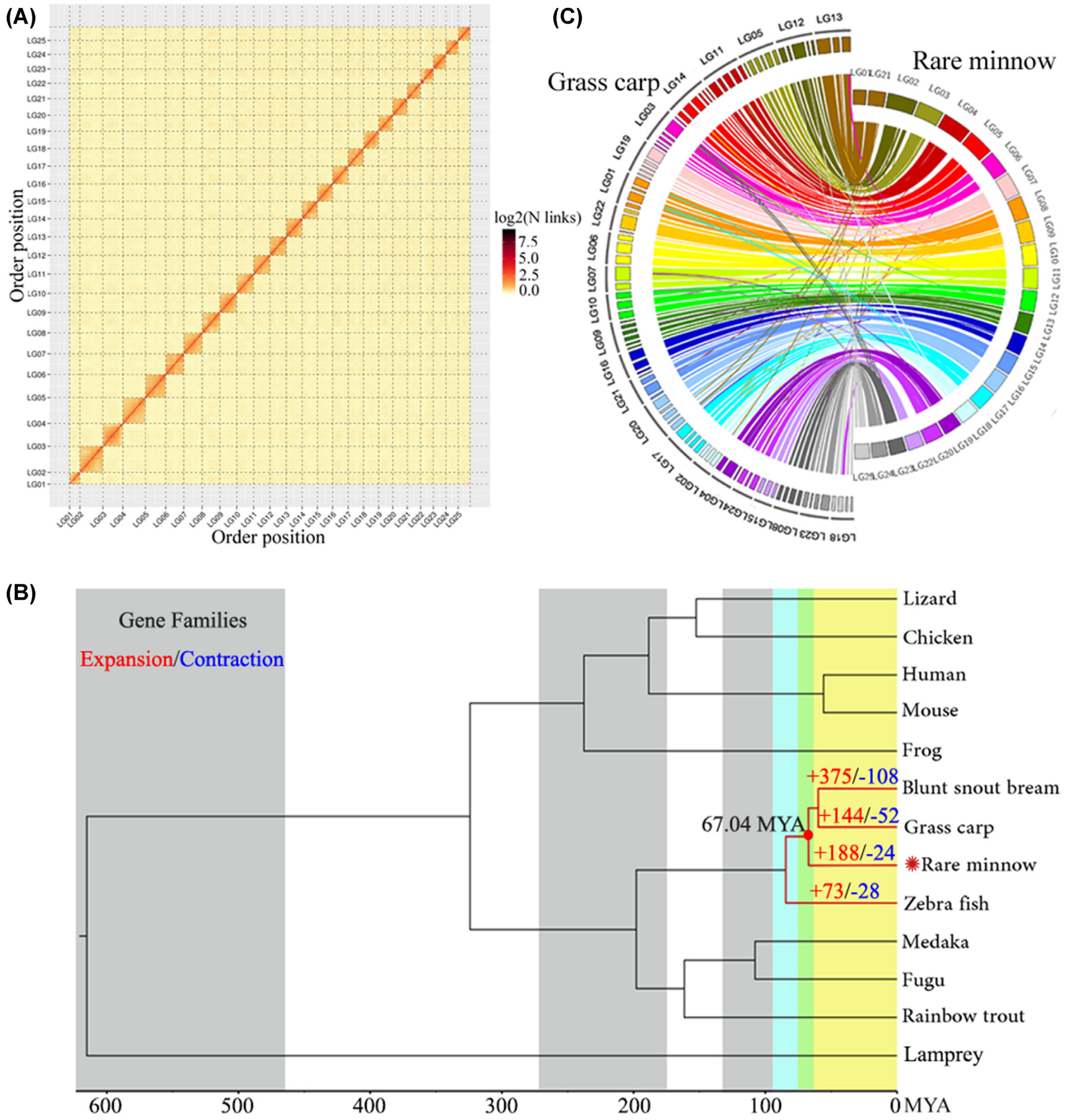


Figure 1: Evolutionary analysis of the genome of rare minnow. A. Rare minnow genome contact matrix using Hi-C data. The color bar indicates the logarithm of the contact density from red (high) to white (low) in the plot. Note that only sequences anchored on chromosomes are shown in the plot. B. A phylogenetic tree was constructed from 13 species, including the 4 Cyprinidae species. The time of divergence and the expansion and contraction of gene families of the 4 Cyprinidae species are described with a maximum-likelihood tree. The number of expansion events is indicated in red, and contraction events are indicated in blue. C. A comparative analysis of the rare minnow and grass carp genomes was performed. There was a high collinearity between the 2 species. Rare minnow LG1 and LG21 corresponded to grass carp LG13. The LG number and supercoiling number of grass carp were obtained from the study of Huang et al. [18]. MYA: million years ago.

improving the possibility of obtaining target trait mutants. In addition, the method established in this study only requires the synthesis of 26 sgRNAs, and mutants can be obtained using efficient screening methods. However, it should be noted that 23 hemorrhagic disease–associated loci were almost all produced by ATG targets (Supplementary Table S4). This may be related to the fact that only a few gene promoters contain TATA boxes. Previous studies have found that 23.85% of eukaryotic promoter sequences contain TATA boxes, and ∼20% of yeast genes contain a TATA box [19, 20]. Future studies are expected to consider only ATG as a mutation target.

Many studies have been conducted to construct plant mutant libraries by physical and chemical mutagenesis, with mutant frequency between 0.031% and 9.3% [21–24]. The efficiency of

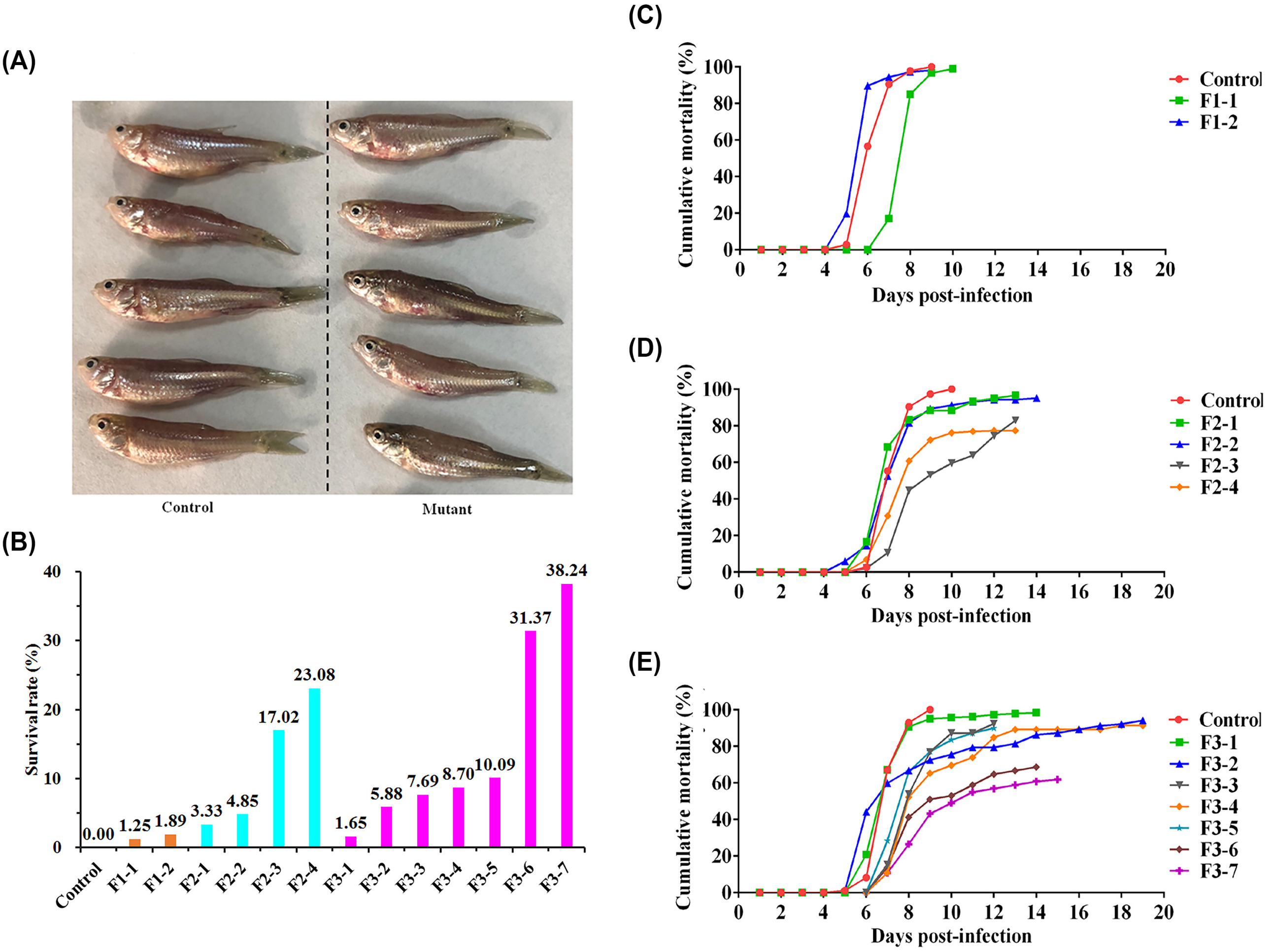


Figure 2: Establishment of an anti-hemorrhagic model of rare minnow. A. Clinical symptoms of the mutant and control groups after GCRV infection. There was no difference in the clinical phenotype between the mutant group and the control group individuals that died after GCRV infection. The body surface of the dead individuals in both the mutant and control groups was red, showing obvious symptoms of hemorrhagic disease. B. Survival rates of the F1–F3 and control groups after GCRV infection. The high-salt invasion method was used. C–E. Daily cumulative mortality for the F1, F2, F3, and control groups after GCRV infection. The number of dead fish in all groups was recorded every day. The number of the daily cumulative deaths in each group relative to the total number of individuals in each group is the daily cumulative mortality. Different colors were used to represent different families. The abscissa represents the days post-infection and the ordinate represents the cumulative mortality.

Arabidopsis thaliana mutants obtained through transposon mutagenesis was 0.091% and 1% [25, 26]. In animals, the chemical mutagen ethyl nitrosouria is mainly used in relevant studies in some species, such as *Caenorhabditis elegans* [27, 28], zebrafish [29, 30], mouse (*Mus musculus*) [31, 32], grass carp [33], and pig (*Sus scrofa*) [34], and the mutant frequency is generally not more than 0.03%. Compared with existing studies, the mutant frequency of this method (0.23%) is similar to that in plants, but ~10 times higher than that in animals. In addition, another important reason why we successfully obtained resistant mutants by this method is that the selected traits were qualitative traits. The entire wild-type population of rare minnow died after hemorrhagic disease; individuals who survived after infection were mutated individuals, who could be easily and efficiently identified. However, it must be pointed out that this method requires whole-genome sequence information and may not be feasible in the screening of quantitative traits or some species with long generation times.

Methods

Sources of experimental fish, viruses, and cells

Rare minnow (*Gobiocypris rarus*; NCBI:txid143606) samples were collected from the Liusha River, Hanyuan County, Sichuan Province, China, by the ichthyology laboratory at the Institute of Hydrobiology, Chinese Academy of Sciences. GCRVs were isolated and preserved in our laboratory. GCO cells were presented by Li Shun, Associate Professor at the Institute of Hydrobiology, Chinese Academy of Sciences. *Ctenopharyngodon idellus* kidney (CIK) cells were purchased from China Center for Type Culture Collection (CCTCC) (Wuhan, Hubei, China).

Experiments involving rare minnows in this study were carried out in accordance with the Guide for the Care and Use of Laboratory Animals (Ministry of Science and Technology of China, 2006), and the protocol was approved by the Committee of the Institute of Hydrobiology, Chinese Academy of Sciences. The reference number obtained was Y9110306.

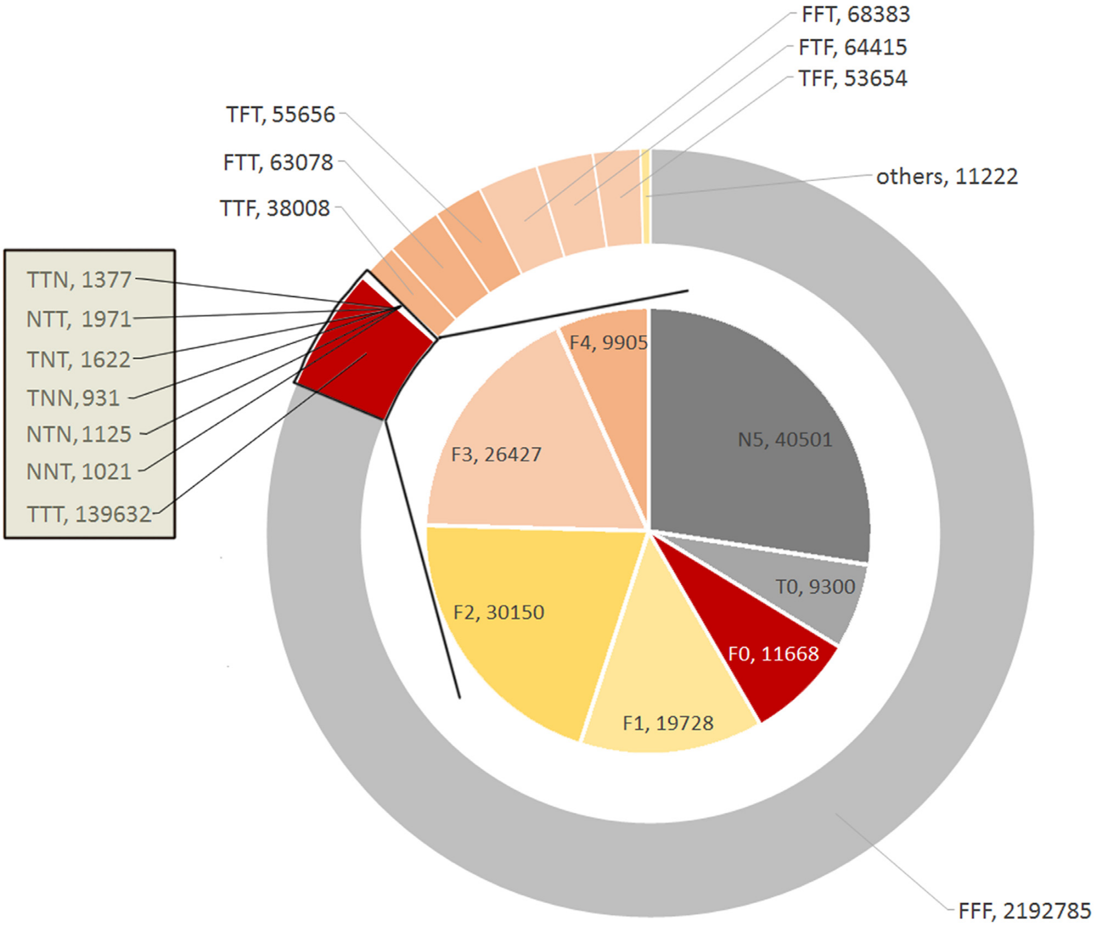


Figure 3: Screening of candidate indels related to hemorrhagic disease. T denotes new genotypes compared with those in the control group, N denotes no results of genotyping, and F indicates that the genotypes also appear in the control group. The outer ring shows the statistics of the genotyping results of 3 F2 families; TTN indicates that the first and the second families have new genotypes and the third family has no genotyping results; TFT indicates that the first and third families have new genotypes, and the second family's genotypes also appear in the control group, and so on. The inner ring indicates the genotyping results of 147,679 loci (the sum of TTN, NTT, TNT, TNN, NTN, NNT, and TTT type in the outer ring) in 5 parents (P1–P5). F0 means that there are either new genotypes or no genotyping results among the 5 parents compared with that in the control group. F1 indicates that 1 of the 5 parents has the same genotypes as the control group, while the other 4 parents have either new genotype or no genotyping results, and so on.

Genome sequencing and assembly

A sexually mature female rare minnow was selected for this study. Part of the muscle tissue was frozen in liquid nitrogen and genomic DNA was extracted from the other part. The cetyltrimethylammonium bromide method was used to extract DNA. NGS was performed on an Illumina HiSeq X Ten platform using paired-end reads (PE) of 150 bp, and the sequencing fragments were 350 ± 50 bp. After conventional filtering, a k-mer frequency distribution map was drawn based on the k-mer (k = 21) analysis method, and genome size, heterozygosity, and repetition rate were evaluated.

The Pacific Biosciences Sequel system was used for third-generation sequencing. Subreads were obtained using signal-to-noise ratio filtering. After using Canu v1.9 (Canu, RRID:SCR_015880) [35] to self-correct subreads, WTDBG v1.2.8 (WTDBG, RRID:SCR_017225) [36] was used for sequence assembly. Based on previous NGS data used for genome evaluation, the assembled genome sequence was corrected using Pilon v1.23 (Pilon, RRID:SCR_014731) [37].

The muscle tissue cryopreserved in liquid nitrogen was fixed and cross-linked with formaldehyde, and a Hi-C library was constructed. NGS was performed using an Illumina HiSeq X Ten platform. Clean data were obtained after routine filtration and compared with assembled genome sequences. The comparison results were filtered using HIC-Pro v2.11.1 (HIC-Pro, RRID:SCR_017643) [38] to obtain valid interaction pairs. Based on valid interaction pairs, the genome assembled in the previous step was divided, sorted, and oriented using LACHESIS (LACHESIS, RRID:SCR_017644) [39], and the assembly sequence at the chromosome level was obtained. The completeness of the genome was evaluated through BUSCO v5.2.2 (BUSCO, RRID:SCR_015008) [40] using the gene set of actinopterygii_odb10. Then the number of Hi-C read pairs covering any 2 bins was used as the intensity signal of the interaction between the 2 bins, and a heat map was drawn to evaluate the Hi-C assembly results.

Genome annotation

Genome annotation was performed in 2 parts: repetitive sequence annotation and coding gene annotation. RepeatModeler v1.0.11 (RepeatModeler, RRID:SCR_015027) was used to construct a repetitive sequence library of the genome, and RepeatMasker v4.0.9 (RepeatMasker, RRID:SCR_012954) was used to mark the

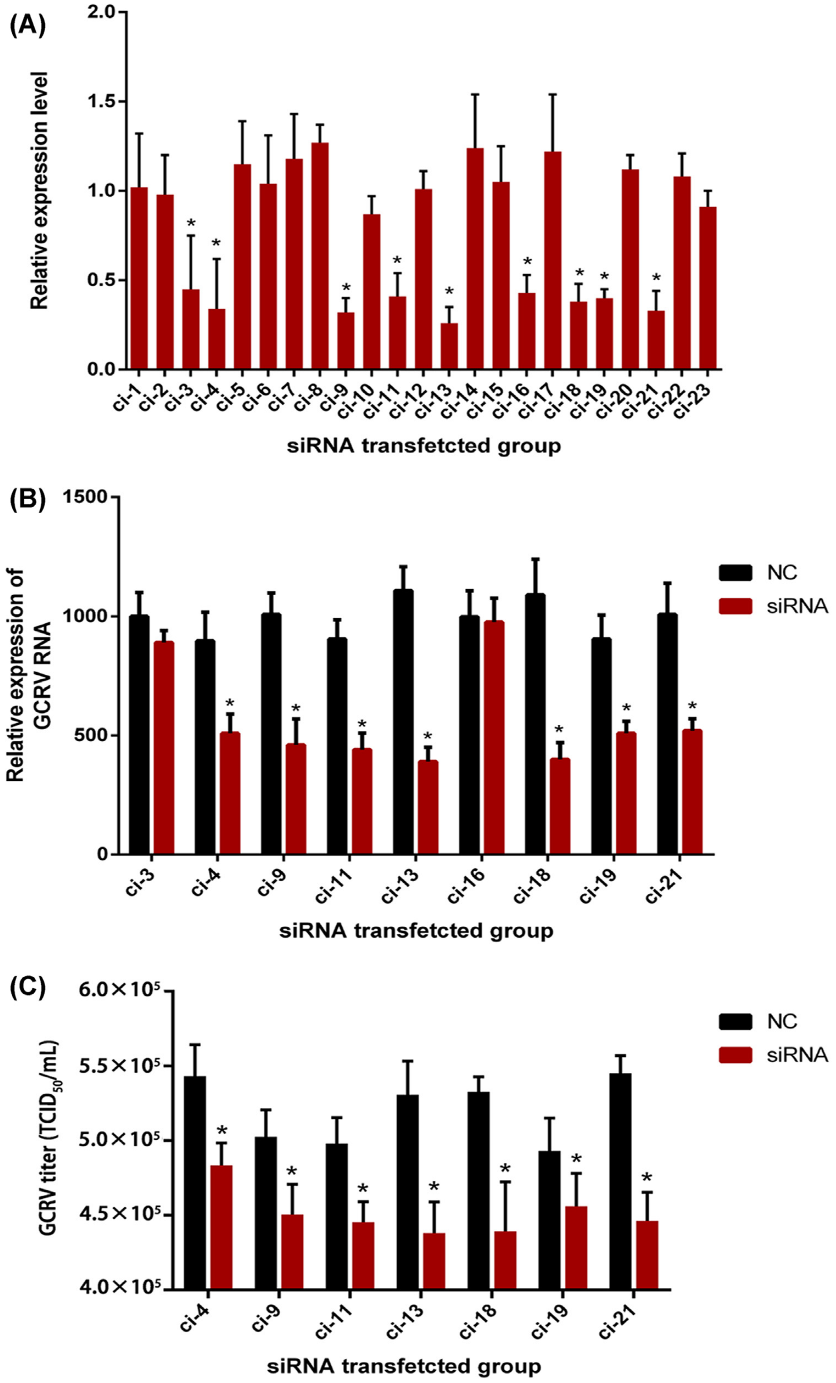
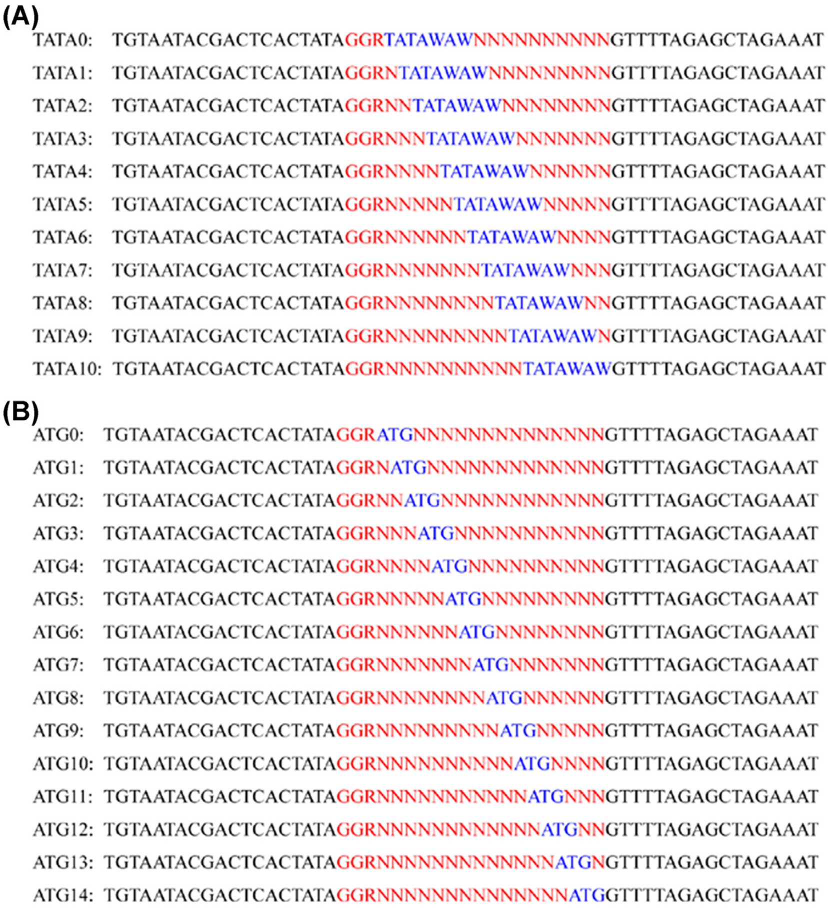


Figure 4: Effects of siRNAs on screened target genes and GCRV. A. The GCO cells were cultured in 24-well plates. Each siRNA for grass carp genes was transfected into the cells. The cellular total RNA was extracted at 0 and 48 h post-transfection. q-PCR was used to detect the relative expression of 23 grass carp genes using β actin as the internal reference gene and the $2^{-\Delta\Delta Ct}$ method. The ratios of 48 h/0 h of each group were then calculated. B. The 9 selected siRNAs were transfected into the GCO cells. The siRNA negative control (NC) was used in each group. At 16 h post-transfection, the medium was removed, and the cells were infected with GCRV with MOI = 5. The cells were collected at 32 h post-infection. Using β actin as the internal reference gene, the relative expression of GCRV RNA relative to NC was detected by the $2^{-\Delta\Delta Ct}$ method. C. CIK cells were seeded into 96-well plates. Then, the cells were infected with viral samples (from GCO cells transfected and infected as in [B] for 3 days). Cytopathic effect was then observed under the microscope, and the titer was determined using the Reed-Muench formula. Data represent results of 3 independent experiments, and error bars indicate mean ± SD. Statistical analyses were performed using multiple *t*-tests (n = 3). $^*P < 0.05$.

repetitive sequences based on the repetitive sequence library; the parameter was –nolow –div 20 –GC 39 [41]. Finally, previous results were further annotated using the existing repeat sequences of rare minnow in the Repbase database (Repbase, RRID:SCR_021169); the parameter was –nolow –div 20 –GC 39 [42].

The annotation of the coding genes integrated the results of *ab initio* gene prediction, protein sequence alignment,



(A)

```
TATA0:  TGTAATACGACTCACTATAGGRTATAWAWNNNNNNNNNNGTTTTAGAGCTAGAAAT
TATA1:  TGTAATACGACTCACTATAGGRNTATAWAWNNNNNNNNNGTTTTAGAGCTAGAAAT
TATA2:  TGTAATACGACTCACTATAGGRNNTATAWAWNNNNNNNNGTTTTAGAGCTAGAAAT
TATA3:  TGTAATACGACTCACTATAGGRNNNTATAWAWNNNNNNNGTTTTAGAGCTAGAAAT
TATA4:  TGTAATACGACTCACTATAGGRNNNNTATAWAWNNNNNNGTTTTAGAGCTAGAAAT
TATA5:  TGTAATACGACTCACTATAGGRNNNNNTATAWAWNNNNNGTTTTAGAGCTAGAAAT
TATA6:  TGTAATACGACTCACTATAGGRNNNNNNTATAWAWNNNNGTTTTAGAGCTAGAAAT
TATA7:  TGTAATACGACTCACTATAGGRNNNNNNNTATAWAWNNNGTTTTAGAGCTAGAAAT
TATA8:  TGTAATACGACTCACTATAGGRNNNNNNNNTATAWAWNNGTTTTAGAGCTAGAAAT
TATA9:  TGTAATACGACTCACTATAGGRNNNNNNNNNTATAWAWNGTTTTAGAGCTAGAAAT
TATA10: TGTAATACGACTCACTATAGGRNNNNNNNNNNTATAWAWGTTTTAGAGCTAGAAAT
```

(B)

```
ATG0:  TGTAATACGACTCACTATAGGRATGNNNNNNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG1:  TGTAATACGACTCACTATAGGRNATGNNNNNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG2:  TGTAATACGACTCACTATAGGRNNATGNNNNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG3:  TGTAATACGACTCACTATAGGRNNNATGNNNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG4:  TGTAATACGACTCACTATAGGRNNNNATGNNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG5:  TGTAATACGACTCACTATAGGRNNNNNATGNNNNNNNNNGTTTTAGAGCTAGAAAT
ATG6:  TGTAATACGACTCACTATAGGRNNNNNNATGNNNNNNNNGTTTTAGAGCTAGAAAT
ATG7:  TGTAATACGACTCACTATAGGRNNNNNNNATGNNNNNNNGTTTTAGAGCTAGAAAT
ATG8:  TGTAATACGACTCACTATAGGRNNNNNNNNATGNNNNNNGTTTTAGAGCTAGAAAT
ATG9:  TGTAATACGACTCACTATAGGRNNNNNNNNNATGNNNNNGTTTTAGAGCTAGAAAT
ATG10: TGTAATACGACTCACTATAGGRNNNNNNNNNNATGNNNNGTTTTAGAGCTAGAAAT
ATG11: TGTAATACGACTCACTATAGGRNNNNNNNNNNNATGNNNGTTTTAGAGCTAGAAAT
ATG12: TGTAATACGACTCACTATAGGRNNNNNNNNNNNNATGNNGTTTTAGAGCTAGAAAT
ATG13: TGTAATACGACTCACTATAGGRNNNNNNNNNNNNNATGNGTTTTAGAGCTAGAAAT
ATG14: TGTAATACGACTCACTATAGGRNNNNNNNNNNNNNNATGGTTTTAGAGCTAGAAAT
```

Figure 5: Design of the upstream primers for the 26 target sites. A. Eleven forward sgRNA primer sequences designed with TATAWAW as the target. B. Fifteen forward sgRNA primer sequences designed with ATG as the target, where M = A/C, R = A/G, W = A/T, Y = C/T, N = A/T/C/G; Red and blue sequences represent the whole target sequence region, which starts with GGR to improve transcription efficiency. Blue sequences represent the conservative sequences of the promoter region (A) and the start codon sequences (B).

and transcriptional assembly. For *ab initio* gene prediction, AUGUSTUS v3.3.3 (AUGUSTUS, RRID:SCR_008417) [43], GlimmerHMM v3.0.4 (GlimmerHMM, RRID:SCR_002654) [44], geneID v1.4 (geneID, RRID:SCR_021639) [45], and SNAP v2006-07-28 (SNAP, RRID:SCR_002127) [46] were used. Whole-genome protein sequences of 3 related species—common carp (*Cyprinus carpio*), goldfish (*Carassius auratus*), and zebrafish (*Danio rerio*)—were used for homologous protein sequence alignment prediction using Gemoma v1.6.4 (Gemoma, RRID:SCR_017646) [47, 48]. Two strategies were used for transcriptional assembly: with and without a reference genome. The strategy with reference genome involved using hisat2 v2.1.0 (hisat2, RRID:SCR_015530) [49] for alignment and StringTie v1.3.5 (StringTie, RRID:SCR_016323) [50] for assembly. The strategy without the reference genome involved the assembly of Trinity v2.8.5 (Trinity, RRID:SCR_013048) [51]. The transcripts from the 2 sources were processed using the PASA pipeline (PASA, RRID:SCR_014656) [52], including sequence filtering and realignment analysis. Finally, the results of the 3 sources were evaluated with EVM (EVM, RRID:SCR_014659) [53] to obtain the gene coding regions, and the untranslated region of the gene was annotated using the PASA pipeline and transcriptome data. In AUGUSTUS, "zebrafish" was selected as the training set for prediction, and default parameters were used for all other software.

Evolutionary analysis of the genomes

Thirteen vertebrate genomes were collected for evolutionary analysis. Using Orthofinder v2.4.0 (Orthofinder, RRID:SCR_017118) [54], the protein sequences of the 13 species were classified (the DIAMOND alignment program was used, with an e-value of 0.001) to obtain shared gene families, shared genes, and single-copy gene families among species. Then the gene families obtained were annotated using the PANTHER database (PANTHER, RRID:SCR_004869) [55]. The obtained single-copy gene family was sorted by MAFFT v7.471 (MAFFT, RRID:SCR_011811) [56], and a phylogenetic tree was constructed using RAxML-NG v0.9.0 (RAxML-NG, RRID:SCR_006086) [57] and the maximum likelihood (ML) method, for which the number of bootstraps was set to 1,000. Combined with fossil evidence, r8s v1.81 (r8s, RRID:SCR_021161) was used to construct a phylogenetic tree with divergence time.

Using CAFE v4.2 (CAFÉ, RRID:SCR_018924) [58] and the results of the phylogenetic tree with divergence time and gene family clustering, we estimated the number of gene family members in the ancestors of the 4 Cyprinidae fish species using the birth mortality model, and predicted the contraction and expansion of the gene family of the 4 Cyprinidae fish species relative to their ancestors (the criterion for contraction and expansion was $P < 0.05$).

Because the grass carp genome is at the supercontig level, 99 large supercontigs attached by a published genetic linkage map of grass carp were used for collinearity analysis [18]. JCVI v0.18 (JCVI, RRID:SCR_021641) [59] was used to perform protein sequence alignment between rare minnow and grass carp. Finally, a collinearity graph was drawn using Circos v0.69 (Circos, RRID:SCR_011798).

Establishment of an anti–hemorrhagic disease model

The promoter and coding regions were selected as the main mutation regions. While designing the mutation target site, among the 20 bases starting from GGR (requirements of T7 promoter, R for A/G), the conserved sequence TATAWAW (W for A/T) in the TATA frame and the start codon ATG were gradually shifted backward, and N was used as a supplement. The 26 primers upstream of the target site are shown in Fig. 5, and the primer downstream of the target site was AAAAAAAGCACCGACTCGGTGCCACT. After PCR amplification using the pMD-19T-gRNA plasmid as a template, 26 sgRNAs were transcribed using a TranscriptAid T7 High Yield Transcription Kit (Thermo Scientific, USA).

Twenty-six sgRNAs were mixed with Cas9 protein (Invitrogen, USA) at final concentrations of 400 and 100 ng/μL. Each sgRNA was injected into ~300 rare minnow embryos, which constituted the P0 generation. At 2 months of age, a high-salt invasion method was used for GCRV infection. The method was as follows: the fish were soaked in 6% NaCl solution for 2 min and then quickly transferred to GCRV suspension (virus titer: 2.75 × 10^8 $TCID_{50}$/mL) for 30 min. The wild-type mixed population used as a control group was infected in the same manner. The number of dead fish in each group was recorded daily.

From the surviving individuals of the P0 generation, male individuals were selected and lateral-crossed with wild-type female individuals to obtain F1 full-sibling families. GCRV infection was performed at 2 months of age. The surviving individuals in an F1 full-sibling family with the highest survival rate were self-crossed to construct F2 full-sibling families. The F3 generation was obtained by self-crossing in the same way and was infected with GCRV. The wild-type mixed population was used as a control group for infection. The number of deaths in the F1–F3 population and the wild-type population were counted every day after infection, and those individuals who did not die for 2 consecutive weeks were considered survivors. Cumulative mortality curves were drawn, and the survival rate of each family was calculated.

Screening of candidate indels associated with hemorrhagic disease

Three surviving individuals were randomly selected from 3 families (F2-2, F2-3, and F2-4) with high disease resistance in the F2 generation. Three wild-type female and 3 wild-type male individuals were selected. Genomic DNA was extracted from 15 fish using the high-salt method. Sequencing libraries T1, T2, and T3 were constructed by mixing the DNA of 3 fish in F2-2, F2-3, and F2-4, and sequencing library C was constructed by mixing the DNA of 6 wild-type individuals. The inserted fragment size was 350 ± 50 bp, and NGS was performed on the BGI MGISEQ-2000 platform with a PE 150. Five parents (F1 survival mutant P1–P5) of the F2 families were sequenced in the same manner. In addition, the NGS data (S1 and L7) of the 2 groups of wild-type were collected from our lab to increase the information richness of the control group. L7 was obtained from a wild-type female and a wild-type male mixed sample, and S1 was from a wild-type male sample.

Clean data were obtained by filtering the raw data of all samples. Using Bowtie2 v2.3.5 (Bowtie2, RRID:SCR_005476) [60], 11 datasets were compared with the reference genome of rare minnow assembled above. Then, the HaplotypeCaller of GATK v4.1.1.0 (GATK, RRID:SCR_001876) [61] was used for indel calling. Library C had 6 mixed samples, and the parameter was set to –sample ploidy 12; T1, T2, and T3 had 3 mixed samples, and the parameter was set to –sample ploidy 6. The indel filters of all samples were hard filtered with QD < 2, FS > 100, read position Mann –Whitney Rank-Sum < 20, and SOR > 10. Finally, VCF files were used to record the indel loci and genotypes of 11 datasets; snpEff (snpEff, RRID:SCR_005191) [62] was used to annotate the VCF files.

The genotypes of each indel locus in 3 samples (C, S1, and L7) were combined as controls and compared with corresponding indels in 8 samples (T1, T2, T3, and P1–P5). Among the 8 samples, the locus with the new genotype was recorded as "T," the locus without a genotyping result was recorded as "N," and the locus with a genotyping result but without a new genotype was recorded as "F." The contribution of these loci to gene function changes was used to distinguish the SnpEff annotation results, which can be divided into 4 levels: high, moderate, low, and modified [62]. Next, there were 3 steps in the screening process: the first step was to screen the loci that were not "F" type in T1, T2, and T3; the second step was to screen the loci that were not "F" type in the 5 parents from the results of the first step; and the third step was to screen the loci with "high" contribution. Finally, the candidate loci associated with hemorrhagic disease were identified.

Functional verification of susceptible genes related to hemorrhagic disease

The genome annotation information of candidate loci of rare minnow was used to obtain the genes corresponding to these sites. Then, the complementary DNA sequences of these genes were compared with the annotated information of the grass carp genome [63], and homologous genes in grass carp were selected. To preliminarily and quickly verify the function of these candidate loci, we carried out relevant studies at the cell level of grass carp using knockdown technology. For each homologous grass carp gene, siRNA was designed and synthesized by RiboBio Co. (Guangzhou, China). qPCR primers for homologous grass carp genes were designed to confirm the knockdown effect of the siRNA.

A monolayer of GCO cells was subcultured in 24-well plates. When the cells reached 80% confluence at the bottom of the well, siRNA was transfected into the cells using FishTransH (Meisent Co., Wuhan, China). The dosage of siRNA (concentrated at 20 μmol/L) was 40 pmol per well. Cells were collected at 0 and 48 h post-transfection, and total RNA was extracted using TRIzol (Life Technologies). RT-qPCR was used to detect the expression of 23 grass carp genes at 48 h relative to 0 h post-transfection. siRNAs with inhibitory effects were selected for the subsequent experiments.

The GCO cells were subcultured in 24-well plates. When the cells reached ~80% confluence at the bottom of the well, the selected siRNA was transfected into the cells using FishTransH. siRNA-NC (RiboBio Co.) was used as an NC in each group, and the dosage of siRNA was 40 pmol per well. At 16 h post-transfection, the medium was removed and the cells were infected with GCRV at a MOI of 5. The cells were collected 32 h after infection.

Total RNA was extracted, and RT-qPCR was performed to detect the relative changes of GCRV RNA relative to the NC. GCO cells transfected and infected in the same way were removed to −70 °C, and frozen and thawed 2 times for collecting viral samples. Then, CIK cells were seeded into 96-well plates, 5,000 cells per well. After 24 h, the cells per well were infected with 100 μL viral samples of 10-fold serial dilutions in culture medium and incubated for 3 days. Cytopathic effect was then observed under the microscope, and the titer was determined using the Reed-Muench formula [64] and expressed as $TCID_{50}$/mL.

Data Availability

Raw sequences for genome assembly including Illumina, Pacific Biosciences, and Hi-C reads are available in NCBI and can be accessed with accession No. PRJNA732062. Sequencing data for screening of candidate indels associated with hemorrhagic disease are also available in NCBI under accession Nos. PRJNA732511 and PRJNA613868. Other supporting data, including the genome assembly and annotation files of rare minnow, are available via the GigaScience database, GigaDB [65].

Additional Files

Supplementary Table S1. Statistics of repeat elements.
Supplementary Table S2. Gene annotation statistics for 4 Cyprinidae fish species.
Supplementary Table S3. Twenty-three loci associated with hemorrhagic diseases.
Supplementary Table S4. Distance from 23 loci to target sites.
Supplementary Table S5. Twenty genes associated with hemorrhagic diseases.
Supplementary Table S6. Twenty-three homologous genes in grass carp.
Supplementary Table S7. siRNAs and specific primer sequences for the 23 grass carp genes.
Supplementary Figure S8. Pipeline of random genome editing.

Abbreviations

bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; CCTCC: China Center for Type Culture Collection; CIK: *Ctenopharyngodon idellus* kidney; CRISPR: clustered regularly interspersed short palindromic repeats; dpi: days post-infection; GATK: Genome Analysis Toolkit; Gb: gigabase pairs; GCO: grass carp ovary; GCRV: grass carp reovirus; kb: kilobase pairs; MAFFT: Multiple Alignment using Fast Fourier Transform; Mb: megabase pairs; ML: maximum likelihood; MOI: multiplicity of infection; NC: negative control; NCBI: National Center for Biotechnology Information; NGS: next-generation sequencing; PASA: Program to Assemble Spliced Alignments; sgRNA: small guide RNA; SNP: single-nucleotide polymorphism.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (31972788) and the State Key Laboratory of Freshwater Ecology and Biotechnology (2019FBZ05, 2021FB11).

Authors' Contributions

R.H. and Y.W. conceived and designed the experiments. L.Luo, R.H., M.O., and Y.L. performed the experiments. M.S., C.Y., W.Z., and X.X. analyzed the genome data. R.H., M.S., and L.Luo. drafted and revised the manuscript. Y.W., L.Liao, and Z.Z. provided advice on manuscript writing. All authors reviewed and declare that they have no conflict of interest.

References

1. Liao T, Xu Y, Zhong XP, et al. Comparative vitellogenic responses in zebrafish (*Brachydanio rerio*) and rare minnow (*Gobiocypris rarus*) exposed to 17α-ethinylestradiol. Acta Hydrobiol Sin 2005;**29**(5):513–7.
2. Jiang FQ. Methods of acute and subchronic toxicity test using *Gobiocypris rarus* and their application in toxicity study of sediment in Donghu Lake. Beijing: Graduate School of Chinese Academy of Sciences; 2006.
3. Zha JM, Sun LW, Spear PA, et al. Comparison of ethinylestradiol and nonylphenol effects on reproduction of Chinese rare minnows (*Gobiocypris rarus*). Ecotoxicol Environ Saf 2008;**71**(2):390–9.
4. Wang J, Cao W. *Gobiocypris rarus* as a Chinese native model organism: history and current situation. Asian J Ecotoxicol 2017;**12**:20–33.
5. Wang T, Liu P, Chen H, et al. Preliminary study on the susceptible of *Gobiocypris rarus* to hemorrhagic virus of grass carp (GCHV). Acta Hydrobiol Sin 1994;**2**:144–9.
6. Zhang L, Luo Q, Fang Q, et al. An improved RT-PCR assay for rapid and sensitive detection of grass carp reovirus. J Virol Methods 2010;**169**(1):28–33.
7. Sega GA. A review of the genetic effects of ethyl methanesulfonate. Mutat Res 1984;**134**(2-3):113–42.
8. McCallum CM, Comai L, Greene EA, et al. Targeted screening for induced mutations. Nat Biotechnol 2000;**18**(4):455–7.
9. Till BJ, Reynolds SH, Greene EA, et al. Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Res 2003;**13**(3):524–30.
10. Singh M, Lewis PE, Hardeman K, et al. Activator mutagenesis of the pink scutellum1/viviparous7 locus of maize. Plant Cell 2003;**15**(4):874–84.
11. Bai L, Singh M, Pitt L, et al. Generating novel allelic variation through Activator insertional mutagenesis in maize. Genetics 2007;**175**(3):981–92.
12. Ran FA, Hsu PD, Wright J, et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 2013;**8**(11):2281–308.
13. Mehravar M, Shirazi A, Nazari M, et al. Mosaicism in CRISPR/Cas9-mediated genome editing. Dev Biol 2019;**445**(2):156–62.
14. Zhou Y, Zhu S, Cai C, et al. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 2014;**509**(7501):487–91.
15. Kim HS, Lee K, Bae S, et al. CRISPR/Cas9-mediated geneknockout screens and target identification via whole genome sequencing uncover host genes required for picornavirus infection. J Biol Chem 2017;**292**(25):10664–71.
16. Lu Y, Ye X, Guo R, et al. Genome-wide targeted mutagenesis in rice using the CRISPR/Cas9 system. Mol Plant 2017;**10**(9):1242–5.
17. Meng X, Yu H, Zhang Y, et al. Construction of a genome-wide mutant library in rice using CRISPR/Cas9. Mol Plant 2017;**10**(9):1238–41.

18. Huang X, Jiang Y, Zhang W, et al. Construction of a high-density genetic map and mapping of growth related QTLs in the grass carp (*Ctenopharyngodon idellus*). BMC Genomics 2020;**21**(1):313.
19. Yang C, Bolotin E, Jiang T, et al. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 2007;**389**(1):52–65.
20. Zhang XH, Qi YX. Analysis on TATA-box, GC-box and CAAT-box in eukaryotic promoters. J Anhui Agric Sci 2008;**36**(4):1380–95.
21. Ren T, Ren H, Du H, et al. Construction of EMS mutant library and screening of total flavonoid content mutants of *Apocynum venetum*. J Plant Genet Res 2020;**21**(3):655–62.
22. Shan C, Shu Q, Wu D. Preliminary study on adding to leaf color marker for rice cytoplasmic male sterile (cms) line long-te-fu A by mutation technology. J Zhejiang Univ 1999;**25**:569–72.
23. Xin Z, Wang ML, Barkley NA, et al. Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. BMC Plant Biol 2008;**8**(1):103.
24. Julio E, Laporte F, Reis S, et al. Reducing the content of nornicotine in tobacco via targeted mutation breeding. Mol Breed 2008;**21**(3):369–81.
25. Wilson K, Long D, Swinburne J, et al. A dissociation insertion causes a semidominant mutation that increases expression of *TINY*, an arabidopsis gene related to *APETALA2*. Plant Cell 1996;**8**(4):659–71.
26. Marsch-Martinez N, Greco R, Van Arkel G, et al. Activation tagging using the *En-I* maize transposon system in *Arabidopsis*. Plant Physiol 2002;**129**(4):1544–56.
27. De Stasio EA, Dorman S. Optimization of ENU mutagenesis of *Caenorhabditis elegans*. Mutat Res 2001;**495**(1-2):81–8.
28. Epstein HF, Shakes DC. *Caenorhabditis elegans*: Modern Biological Analysis of an Organism. 1st ed. Academic; 1995.
29. Driever W, Solnica-Krezel L, Schier AF, et al. A genetic screen for mutations affecting embryogenesis in zebrafish. Development 1996;**123**(1): 37–46.
30. Geisler R, Rauch GJ, Geiger-Rudolph S, et al. Large-scale mapping of mutations affecting zebrafish development. BMC Genomics 2007;**8**(1):11.
31. Hrabe de Angelis MH, Flaswinkel H, Fuchs H, et al. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. Nat Genet 2000;**25**(4):444–7.
32. Concepcion D, Seburn KL, Wen G, et al. Mutation rate and predicted phenotypic target sizes in ethylnitrosourea-treated mice. Genetics 2004;**168**(2):953–9.
33. Jiang XY, Sun CF, Zhang QG, et al. ENU-induced mutagenesis in grass carp (*Ctenopharyngodon idellus*) by treating mature sperm. PLoS One 2011;**6**(10):e26475.
34. Hai T, Cao C, Shang H, et al. Pilot study of large-scale production of mutant pigs by ENU mutagenesis. Elife 2017;**6**:e26248.
35. Koren S, Walenz B P, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;**27**(5):722–36.
36. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020;**17**(2):155–8.
37. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;**9**(11):e112963.
38. Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015;**16**(1):259.
39. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013;**31**(12):1119–25.
40. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.
41. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 2004;**4**(4):10.
42. Bao WD, Adey A, Patwardhan RP. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015;**6**(1):11.
43. Stanke M, Diekhans M, Baertsch R, et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 2008;**24**(5):637–44.
44. Allen JE, Majoros WH, Pertea M, et al. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. Genome Biol 2006;**7**(Suppl 1):doi:10.1186/gb-2006-7-s1-s9
45. Blanco E, Abril JF. Computational gene annotation in new genome assemblies using GeneID. Methods Mol Biol 2009;**537**(537):243–61.
46. Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;**5**(1):59.
47. Keilwagen J, Wenk M, Erickson JL, et al. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 2016;**44**(9):e89.
48. Keilwagen J, Hartung F, Paulini M, et al. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics 2018;**19**(1):189.
49. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;**12**(4):357–60.
50. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;**33**(3):290–5.
51. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 2011;**29**(7):644–52.
52. Campbell MA, Haas BJ, Hamilton JP, et al. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics 2006;**7**(1):327.
53. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 2008,**9**(1):R7.
54. Emms D M, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 2019;**20**(1):238.
55. Mi H, Muruganujan A, Ebert D, et al. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 2019;**47**(D1):D419–26.
56. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 2009;**537**:39–64.
57. Kozlov AM, Darriba D, Flouri T, et al. RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 2019;**35**(21):4453–5.
58. Han MV, Thomas GW, Lugo-Martinez J, et al. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 2013;**30**(8):1987–97.
59. Tang H, Bowers JE, Wang X. Synteny and Collinearity in Plant Genomes. Science 2008;**320**(5875):486–8.

60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.
61. Mckenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;**20**(9):1297–303.
62. Cingolani P,Platts A,Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 2012;**6**(2):80–92.
63. Wang Y, Lu Y, Zhang Y, et al. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides genomic insights into its evolution and vegetarian diet adaptation. Nat Genet 2015;**47**(6):625–31.
64. Lindenbach BD. Measuring HCV infectivity produced in cell culture and in vivo. Methods Mol Biol 2009;**510**: 329–36.
65. Huang R, Shi M, Luo L, et al. Supporting data for “De novo screening of disease-resistant genes from the chromosome-level genome of rare minnow using CRISPR-cas9 random mutation.” GigaScience Database 2021. http://dx.doi.org/10.5524/100937.