High-speed rail model reveals the gene tandem amplification mediated by short repeated sequence in eukaryote

Haidi Chen1,9, Jingwen Xue1,9, Zhenghou Zhang2,9, Geyu Zhang3,9, Xinyuan Xu1, He Li1, Ruxue Zhang2, Najeeb Ullah1, Lvxing Chen7, Amanullah1, Zhuqing Zang1, Shanshan Lai1, Ximiao He4,8,9, Wei Li6, Miao Guan1*, Jingyi Li7*, Liangbiao Chen8* & Cheng Deng1*

The occurrence of gene duplication/amplification (GDA) provide potential material for adaptive evolution with environmental stress. Several molecular models have been proposed to explain GDA, recombination via short stretches of sequence similarity plays a crucial role. By screening genomes for such events, we propose a "SRS (short repeated sequence) *N + unit + SRS*N" amplified unit under USCE (unequal sister-chromatid exchange) for tandem amplification mediated by SRS with different repeat numbers in eukaryotes. The amplified units identified from 2131 well-organized amplification events that generate multi gene/element copy amplified with subsequent adaptive evolution in the respective species. Genomic data we analyzed showed dynamic changes among related species or subspecies or plants from different ecotypes/strains. This study clarifies the characteristics of variable copy number SRS on both sides of amplified unit under USCE mechanism, to explain well-organized gene tandem amplification under environmental stress mediated by SRS in all eukaryotes.

Gene duplication/amplification (GDA) is a genetic mechanism for enhanced antibiotic resistance in bacteria1, and the same mechanism leading to copy number variations (CNVs), which was proposed to underlie many human diseases, such as mental illness, developmental disorders and cancer in humans2. Several molecular mechanisms, including nonequal homologous recombination, rolling circle replication, long-distance template switching, nonhomologous end joining (NHEJ), fork stalling and template switching (FoSTeS)/microhomology-mediated break-induced replication (MMBIR) have been proposed to give rise to GDA and CNVs2–4. All involve sequential processes of DNA double strand breaks (DSBs) and microhomology/homology-mediated recombination following by DNA template switching1. Importantly, homologous recombination, mediated by homologous sequences, plays a crucial role in DNA template switching, and thus the new junction positions created from genomic rearrangements are frequently investigated for commonalities in breakpoint junction sequences5. These studies showed that homologous sequences of variable lengths (from several bp to 2.7 kb) can mediate homologous recombination1,6. Thus far, several sequence motifs associated with recombinant hotspots have been identified in humans7. Earlier studies proposed a rolling circle mechanism without mediation by short repeated sequences in both eukaryotes and prokaryotes8. Homologous recombination can also occur via sister chromatin

---

1Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, 1 Wenyuan Rd., Nanjing 210023, China. 2The Fourth Affiliated Hospital of China Medical University, Shenyang 110032, China. 3Fourth Affiliated Hospital of China Medical University, Shenyang 110032, China. 4Department of Physiology, School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei, China. 5Center for Genomics and Proteomics Research, School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei, China. 6Huabei Key Laboratory of Drug Target Research and Pharmacodynamic Evaluation, Huazhong University of Science and Technology, Wuhan 430030, Hubei, China. 7Department of Dermatovenerology, Institutes for Systems Genetics, Rare Disease Center, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang Street, Chengdu 610041, Sichuan, China. 8Department of Dermatology and Venerology, West China Hospital of Sichuan University, No. 37 Guo Xue Lane, Chengdu 610041, China. 9Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources (Ministry of Education), Institute of Experimental Pathology, Shanghai Ocean University, Shanghai 201306, China. 10These authors contributed equally: Haidi Chen, Jingwen Xue, Zhenghou Zhang and Geyu Zhang. **email: xioniaoa8911@126.com; ljy7733@163.com; lbchen@shou.edu.cn; dengcheng2014@126.com
exchange (SCE), and unequal SCEs (USCEs), which result in the duplication or deletion of genes has been shown in several species, e.g., fly⁴, yeast⁴⁰,⁴¹, mouse⁴²,⁴³ and frog⁴. However, the study of USCE mechanism only studies a single gene of a specific species, only shows that homologous recombination sequences mediate USCE, or there are repeat sequences within these sequences. What are the characteristics of genes amplified under the USCE mechanism? Is the USCE mechanism widespread in all organisms? So far, these need to be further studied.

One extreme example of GDA and CNVs is the generation of tandemly arranged gene clusters, which are well-organized, locally and head-to-tail⁴⁵,⁴⁶. For instance, tens to hundreds of gene copies are produced in the histone and tDNA gene clusters, which are believed to fulfill the need for massive expression⁴⁷,⁴⁸. Additionally, type III antifreeze protein (AFPIII) exists in polar eelpouts and reaches 20–35 mg/ml in the blood of the Antarctic eelpout *Lycodichthys dearborni*⁴⁹. Previous screening of a genomic DNA library estimated approximately 40 similar genes⁴⁹, which are located within a single genomic locus⁴⁶,⁴⁰. Several studies⁴–⁴³ indicated that tandemly arranged repeats with massive gene copies could be generated by repeated sequence-mediated homologous recombination, but it is not known whether the molecular mechanisms among tandemly arranged repeats from genomes are the same.

How new genes evolve and functionally diversify are key questions in evolutionary biology, as new genes play vital roles in evolutionary innovation and allow organisms to adapt and increase in complexity and speciation⁵⁰. An organism can acquire new genes by three distinct routes: (1) direct horizontal acquisition from other organisms (transduction, transformation and conjugation), (2) gene duplication/amplification by recombination or retrotransposition and de novo acquisition from non-coding DNA⁵¹. Duplication/amplification mechanisms that generate new genes or gene variants are a major force in evolution⁵². Gene duplication and subsequent modification are fundamental for genetic variability and adaptation to stresses during environmental changes. Gene duplications are grouped into 5 classes: whole-genome duplication (WGD), tandem duplication (TD), proximal duplication, transposed duplication and dispersed duplication⁵³. Tandem duplicated genes account for a high proportion of eukaryotes. For example, 14–17% of genes in the human, mouse and rat genomes are duplicated genes, and nearly one-third of duplicated genes are tandemly arrayed⁵⁴. The development of 3rd-generation sequencing technologies launched a special era for the discovery of more USCE models, enabling the detection of more USCE model sequences in the genomes of diverse organisms⁵⁵,⁵⁶.

Based on screening 568 genomes, we analyzed the sequences of 2131 USCE models and found that the characteristic of variable copy number SRS on both sides of amplification unit under USCE mechanism mined from whole genomes in nearly all taxa. We defined this amplification unit model as “SRS (short repeated sequence) *N + unit + SRS*N” structures. Unit (gene/DNA fragment under duplication) is the space between SRSs, and N (greater than or equal to 1) represents the variable copy number of SRS. The model of “SRS*N + unit + SRS*N” structure is just as high-speed rail carriage, and SRS is a connection on both sides of the carriage. Then, we proposed a high-speed rail model showing that the locally tandem amplification units are mediated by SRSs with different repeat numbers in most species.

**Results**

A high-speed rail model with an “SRS*N + unit + SRS*N” structure is universal in eukaryotes and comprises a small fraction in bacteria. To explore a universal model for tandem gene amplification, we collected 568 genomes representing eukaryotic and prokaryotic species (Supplementary Table S1). Of these 568 species, 249 have “SRS*N + unit (gene/DNA fragment under duplication) + SRS*N” structure sequences, always connected head to tail by SRSs, similar to the Chinese high-speed rail model (head to tail by junction) (Fig. 1e), including 34 (out of 41) mammals, 9 (out of 11) aves, 21 (out of 47) reptiles, 3 (out of 4) amphibians, 39 (out of 41) fishes, 4 (out of 14) echinoderms, 2 (out of 2) hemichordata, 29 (out of 91) ecphyozoa, 3 (out of 7) annelida, 4 (out of 15) mollusks, 5 (out of 12) platyhelminthes, 1 (out of 14) tunicata, 3 (out of 8) cnidaria, 45 (out of 100) fungi, 25 (out of) 51 plants, 12 (out of 57) discoba, 1 (out of 1) stramenopile and 9 (out of 50) prokaryotes (Supplementary Table S2). 79 species have no information on evolutionary time, so Fig. 1a displays only 170 species in the phylogenetic tree. A total of 1025 and 1106 high-speed rail model sequences were obtained when the SRS was selected in length of 4–200 bp and 2–3 bp, respectively (Supplementary Table S2). This result indicated that the 2–3 bp SRS is similar or slightly more active in mediating amplification than the longer 4–200 bp SRS (Supplementary Fig. S1a). Interestingly, the fish group has much more high-speed rail model sequences than other groups, which may reflect the fact that the recombination rate of fish is higher than those of mammal, amphibian, and reptile⁵⁶ (Supplementary Table S2). The median distances (in Mb) between high-speed rail models in the genome ranged from 2.66 (prokaryote group) to 912.9 (reptile group). On average, there is a high-speed rail model sequence every 250 Mb in the genomes examined (Supplementary Table S2). The presence of 2131 high-speed rail model sequences screened from 249 different species groups suggested that the high-speed rail model is a universal gene amplification model in most eukaryotic and prokaryotic species (Supplementary Table S6). Notably, the efficiency of high-speed rail model sequence screening is greatly dependent on the sequencing platform by which a genome is sequenced. First- and second-generation sequencing platforms are problematic for assembling repeated sequences⁵⁷–⁵⁹, while genomes sequenced through the 3rd generation sequencing platform can readily detect many high-speed rail sequences (Fig. 1a, Supplementary Fig. S1b). With the wide utilization of the 3rd generation platform, there must be more high-speed rail model sequences identified from various species.

The pattern of the “SRS*N + unit + SRS*N” model of high-speed rail sequences obtained was further analyzed. The median length of SRS ranged from 2 to 11 bp (Supplementary Fig. S2a), and the lengths of the majority of the ‘unit’ sequences were less than 5000 bp (Supplementary Fig. S2b). There was no significant difference in GC content between SRS and the host genome (Supplementary Fig. S2c and S2d and Supplementary Table S3). The medium number of all high-speed rail models identified ranged from 7 to 10 (Supplementary Fig. S2e). To
confirm the high-speed rail model pattern in homologous recombination, the head/tail sequences of the high-speed rail model were collected and analyzed. 0 is defined as only SRS in head/tail without any other high-speed rail model unit sequences (“SRS*N + unit … + unit + SRS*N”), and 1 is defined as 100% unit without any SRS sequence in head/tail (“unit + SRS*N … + SRS*N + unit”, start with “unit + SRS*N” or end with “SRS*N + unit”). As shown in Fig. 1d, 0 exhibits the highest frequency ratio, and other ratios between 0 and 1 are evenly distributed. This indicates that the process of gene amplification is mostly complete. Although the total frequency of the other group (0 < ratio ≤ 1) is larger than 0 group, we speculate that this is due to the large randomness of species evolutionary process, but for a single high-speed rail model, it is still more inclined to complete replication. According to head and tail data of the high-speed rail model, we proposed that the high-speed rail model pattern is homologous recombination mediated by SRS (Fig. 1c,d).
Several sequence motifs associated with recombinant hotspots have been identified in humans. After analyzing all these high-speed rail model sequences, we identified 578 conserved motifs ("AC", "CA", "TG" and "GT") from these SRSs (Supplementary Table S5). The motif "TG (AC)" was dominant with 241 high-speed rail model sequences identified to contain this motif.

The genetic nature of the amplified units. To investigate the genetic nature of the amplified units in the high-speed rail model, these units were investigated by BLASTX and BLASTN. As shown in Supplementary Table S4, S5 and S6, from the high-speed rail model sequences we identified in this study, 1190 coding genes, including 627 protein-coding genes, 214 uncharacterized genes, 349 hypothetical genes, 3 transposons, 6 retrotransposons; and 37 ribosomal RNAs and 185 non-coding RNAs (Fig. 1b), were found to be arranged into tandem duplications. Interestingly, some of these genes matched sequences from bacteria or viruses, implying horizontal gene transfer from bacterial or viral organisms to eukaryotic organisms. An example is that the hypothetical protein gene existing in the bacteria Staphylococcus aureus was detected in the genome of goats, which organized into a high-speed rail model gene cluster containing 6 duplicated units (Supplementary Table S4). There are also genes known to encode RNA sequences, including pseudogene rRNA, microRNA and long non-coding RNAs. The microRNA 430a-180 genes in zebrafish were organized in the high-speed rail model. The detailed classifications of the coding capacities of the 'unit' sequences are listed in Supplementary Table S4 and Table S5.

Identification of specific types of high-speed rail models. In addition to those containing one amplifying "unit" per gene, some high-speed rail cases were more specific. In specific case 1, a high-speed rail model type of sequence is located within a domain of a protein, as exemplified by Mucin-5AC, in which numerous WxxW repeating units with unknown function were present. The flanking gene of high-speed rail model from 9 species is shown in Fig. 2a. Mucin-2 and Mucin-5AC, with conserved domains but unknown functions, to species-specific regulatory sequences. (Fig. 2b). The third special high-speed rail model case is that the SRS is that are different between species exemplified in the U2, U5 and U6 clusters. These sequences could contribute to species-specific regulatory sequences. (Fig. 2b). The third special high-speed rail model case is that the SRS is part of the protein coding sequence such as the keratin-associated protein genes in Capra hircus.
Figure 2. Specific types of high-speed rail structural sequences. (a) Tandem mucin-5AC with a WxxW repeating region. The specificity is that the whole high-speed rail model sequences is contained in mucin-5AC. Species with high-speed rail model sequences and their corresponding closed species without high-speed rail model sequence show the time when high-speed rail model appears. SRS and repeat units with different colors represent different SRS and repeat units, respectively. Repeat units with nearby colors (red and pink, different shades of green) indicated that their sequences are similar. (b) Tandem U2, U5 and U6 spliceosomal RNA. The specificity is that there is more than one coding gene in the repeat unit. All or two of U2, U5 and U6 spliceosomal RNAs are contained in the same repeat unit and they will be amplified by SRS at the same time. (c) Tandem keratin-associated protein cluster. The specificity is that SRS is involved in keratin proteins. The colors red, green, orange and blue represent keratin of 20–2, 4–12, 10–8 and R2D, respectively. Each vertical short line represents one copy of SRS, and the ellipsis represents the variable copy number of SRS.
sable in translation, the genes encoding the rRNA are represented in multiple copies in eukaryotic genomes. As shown in Fig. 3c, the high-speed rail model exhibits gene amplification of rDNA in various species.

High-speed rail models containing 5S ribosomal RNA coding sequences were amply found in 13 species, including Mammals (*Homo sapiens*, *Sus scrofa*, *Ovis aries*, and *Felis catus*), Aves (*Aquila chrysaetos*), Amphibians (*Xenopus tropicalis*), Fishes (*Astatotilapia calliptera*, *Oreochromis niloticus*, *Xiphophorus maculatus*, *Takifugu rubripes* and *Carassius auratus*), Insecta (*Drosophila melanogaster* and *Drosophila pseudoobscura*) and Protozoa (*Crithidia expoeki*). Amplification of 5S rRNA genes in 5 species (*Sus scrofa*, *Oreochromis niloticus*, *Carassius auratus*, *Takifugu rubripes*, and *Drosophila melanogaster*) was restricted to the high-speed rail model. However, amplification of 5S rRNA genes in another 6 species (*Xenopus tropicalis*, *Astatotilapia calliptera*, *Oreochromis niloticus*, *Carassius auratus* and *Drosophila pseudoobscura*) adopted both the high-speed rail model and other duplication mechanisms (Supplementary Table S7).
such as Homo sapiens and Xenopus tropicalis, and a nontranscribed spacer (NTS) subject to sequence variations in size and/or sequence. The units and the SRS flanking the units bear no sequence similarity among the 13 high-speed rail models, except for the 5S rRNA coding regions within these units of the high-speed rail model, which showed high similarity, and this result indicated that the 5S rRNA clusters from 13 different species were independently derived by the high-speed rail model (Fig. 3b). In addition, the 5S rRNA from 13 species also showed local tandem amplification, e.g., rolling circle model or other discrete duplicate mechanisms, which is different from the high-speed rail model (Fig. 3b and Supplementary Fig. S8). The 13 5S rRNA high-speed rail models were distributed in different chromosomal locations and had distinct flanking genes (Fig. 3b), suggesting independent origins.

The repeat high-speed rail model containing 18S, 5.8S, 28S rRNA was also found in plants (Gossypium hirsutum, Vigna unguiculata, Cucumis sativus and Ipomoea trifolia), fishes (Channa argus, Salaria fasciata, Xiphophorus helleri, Oncorhynchus kisutch, Etheostoma spectabile and Xenopus tropicalis) and fungi (Parastagonospora nodorum and Leptosphaeria maculans) (Fig. 3c). Similar to the 5S rRNA high-speed rail model described above, the SRSs and the units of the 12 high-speed rail s are different, again supporting independent origins of the 18S, 5.8S and 28S rRNA high-speed rail models in different species.

Thus, rDNA amplification in diverse taxa could have taken place in different molecular mechanisms, but the high-speed rail model appears to be the major model.

The high-speed rail model of amplification mediates expansion of the olfactory receptor 5AN1 clusters. The olfactory receptor family 5 subfamily AN number 1 (OR5AN1) is a G protein-coupled receptors (GPCR), which recognize musk and its related molecules. We found that the organization of the OR5AN1 genes followed the high-speed rail model in Sus scrofa, in which tandem repeat units encoding OR5AN1 are flanked by SRSs of “CCTT”. The same chromosomal loci in other species encode OR5AN1 as well but are not organized in the high-speed rail model (Fig. 3d). In general, pigs have more copies of functional OR5AN1 copies than other species. It is likely that the high-speed rail model of amplification resulted in further expansion of OR5AN1 in pigs which rendered pigs with higher musk sensitivity. We found that the length heterogeneity of pig OR5AN1 copies was due to variable copies of these SRSs. Meanwhile, both Bos taurus and Bubalus bubalis have 6 copies of OR5AN1 gene, Ovis aries and Canis lupus familiaris both have 4 copies of OR5AN1 gene, and none of them are amplified in an SRS-mediated manner.

Dynamic unit numbers among closely related species, subspecies, and ecotypes/strains. The high-speed rail model of gene amplification was found to be present in four primates (Pan troglodytes, Gorilla gorilla, Pongo abelii and Nomascus leucogenys), but not in another 6 species, although all the species shared the same locus (Fig. 4a). Uniform high-speed rail sequences with the same SRS and same unit sequences were found in the DNA-directed RNA polymerase II subunit RPB1 gene clusters of the primates (Fig. 4a), the phospholipase A2 inhibitor and the Ly6/PLAUR domain-containing protein clusters in fishes (Fig. 4b) and the TSPY cluster in the DNA-directed RNA polymerase II subunit RPB1 gene clusters of the primates (Fig. 4a), the phospholipase A2 inhibitor and the Ly6/PLAUR domain-containing protein clusters in fishes (Fig. 4b) and the TSPY cluster in primates (Fig. 4c). The dynamics of the loci were observed in those species (Fig. 4). The numbers of repeated units ranged from 1 to 10, and those of SRS varied. Furthermore, the lengths of the head/tail units were also different. The interspecific variations of the same high-speed rail model indicated the high evolutionary dynamics of the high-speed rail models.

High-speed rail model dynamics are also observed among subspecies. Through genome screening, we found that the IFNα-1/2 gene locus is consists of tandem repeat 3.6-kb units (one IFNα-1/2 gene per unit) in subspecies of Canis lupus familiaris and Canis lupus dingo which belong to Canis lupus. They shared the same repeat units and SRS (Fig. 5a), suggesting the same high-speed rail model. Sequence analysis showed that this high-speed rail type of locus first appeared in Vulpes vulpes but not in Leptonychotes weddellii. We hypothesize that this high-speed rail model experienced dynamic changes in Canis lupus familiaris and Canis lupus dingo (Fig. 5a).

Interferon (IFN) is a glycoprotein, which has the antiviral functions inhibiting cell proliferation and regulating immunity and antitumor activity. As shown in Supplementary Figure S9, type I IFNs in the skin of FV3-challenged (Frog virus 3, now recognized worldwide to be an amphibian pathogen with a threatening potential to cross multiple species barriers) amphibians Xenopus laevis and Xenopus tropicalis were arranged by the high-speed rail model. Subcutaneous administration of type I IFN offered short-term protection of tadpoles against FV3, and these type I IFNs induced the expression of distinct antiviral genes in the tadpole skin.

Similarly, the heat shock protein 83 locus in the subspecies of Leishmania major was also found to be highly dynamic (Fig. 5b).

High-speed rail model genes show dynamics among different Arabidopsis thaliana ecotypes and Drosophila melanogaster strains. Genome sequences (Illumina sequencing platform) from 19 Arabidopsis thaliana ecotypes were obtained from the literature. The dynamic amplification units between ecotypes are shown in Fig. 6a, b. The number of units in the high-speed rail models from the same loci are different among subspecies of Kn-0 (Lithuania), Ler-0 (Poland, formerly Germany), Bvr-0 (Ireland), Tsu-0 (Japan), Zn-0 (Germany) and Can-0 (Canada), etc. The high-speed rail models from some of the ecotypes are identical, e.g., Ler-0, Oy-0 and Po-0 sharing a same high-speed rail model of 7 units, and the ecotypes of Can-0, C1-1, Edi-0, Hi-0, Mt-0, No-0, Sf-2, Wil-2, Ws-0 and Wu-0 share the same high-speed rail model of 12 units (Fig. 6a and Supplementary Table S8). The evolution of the high-speed rail model of gene duplication in these Arabidopsis thaliana ecotypes could be traced (Fig. 6a), although they showed different numbers of units and SRSs. The sequences at the head and tail of the high-speed rail model are highly conserved (Fig. 6b). To confirm the
**a** DNA-directed RNA polymerase II subunit RPB1

| Species          | Chr | Accession       | Start | End   | Time (MYA) |
|------------------|-----|-----------------|-------|-------|------------|
| *Homo sapiens*   |     |                 | 6.05  | 2.82  | 20.2       |
| *Pan troglodytes*|     |                 | 9.1   |       | 29.4       |
| *Gorilla gorilla*|     |                 |       |       | 15.7       |
| *Nomascus leucogenys*|   |                 | 3.88  | 9.1   | 3.69       |
| *Chlorocebus sabaeus*|   |                 | 3.88  | 9.1   | 3.69       |
| *Macaca mulatta* |     |                 |       |       | 2.68       |
| *Macaca fascicularis*|   |                 |       |       | 2.22       |
| *Maylandia zebra*|     |                 | 2.88  | 6.0   | 6.65       |
| *Pundamilia nyererei*|   |                 | 6.0   |       | 12.4       |
| *Haplochromis burtoni*|   |                 | 22.82 |      | 22.82      |
| *Astatotilapia calliprora*|   |                 | 58    |       | 58         |
| *Oreochromis niloticus*|   |                 |       |       | 2.68       |
| *Neolamprologus brevirostris*| |             |       |       | 38.9       |
| *Xiphophorus maculatus*|   |                 |       |       | 35.1       |

**b** Phospholipase A2 inhibitor and Ly6/PLAUR domain-containing protein

| Species          | Chr | Accession       | Start | End   | Time (MYA) |
|------------------|-----|-----------------|-------|-------|------------|
| *Homo sapiens*   |     |                 | 6.05  | 2.82  | 20.2       |
| *Pan troglodytes*|     |                 | 9.1   |       | 29.4       |
| *Gorilla gorilla*|     |                 |       |       | 15.7       |
| *Nomascus leucogenys*|   |                 | 3.88  | 9.1   | 3.69       |
| *Chlorocebus sabaeus*|   |                 | 3.88  | 9.1   | 3.69       |
| *Macaca mulatta* |     |                 |       |       | 2.68       |
| *Macaca fascicularis*|   |                 |       |       | 2.22       |
| *Maylandia zebra*|     |                 | 2.88  | 6.0   | 6.65       |
| *Pundamilia nyererei*|   |                 | 6.0   |       | 12.4       |
| *Haplochromis burtoni*|   |                 | 22.82 |      | 22.82      |
| *Astatotilapia calliprora*|   |                 | 58    |       | 58         |
| *Oreochromis niloticus*|   |                 |       |       | 2.68       |
| *Neolamprologus brevirostris*|   |                 |       |       | 38.9       |
| *Xiphophorus maculatus*|   |                 |       |       | 35.1       |

**c** TSPY (testis specific protein Y-linked)

| Species          | Chr | Accession       | Start | End   | Time (MYA) |
|------------------|-----|-----------------|-------|-------|------------|
| *Homo sapiens*   |     |                 | 6.05  | 2.82  | 20.2       |
| *Pan troglodytes*|     |                 | 9.1   |       | 29.4       |
| *Gorilla gorilla*|     |                 |       |       | 15.7       |
| *Nomascus leucogenys*|   |                 | 3.88  | 9.1   | 3.69       |
| *Chlorocebus sabaeus*|   |                 | 3.88  | 9.1   | 3.69       |
| *Macaca mulatta* |     |                 |       |       | 2.68       |
| *Macaca fascicularis*|   |                 |       |       | 2.22       |
| *Maylandia zebra*|     |                 | 2.88  | 6.0   | 6.65       |
| *Pundamilia nyererei*|   |                 | 6.0   |       | 12.4       |
| *Haplochromis burtoni*|   |                 | 22.82 |      | 22.82      |
| *Astatotilapia calliprora*|   |                 | 58    |       | 58         |
| *Oreochromis niloticus*|   |                 |       |       | 2.68       |
| *Neolamprologus brevirostris*|   |                 |       |       | 38.9       |
| *Xiphophorus maculatus*|   |                 |       |       | 35.1       |
The high-speed rail model shows dynamic unit numbers among closed species. (a) The species *Pan troglodytes*, *Gorilla gorilla*, *Pongo abelii*, and *Nomascus leuconyctes* all have high-speed rail model sequences that contain repeat units coding the DNA-directed RNA polymerase II subunit RPB1. Their closed species (*Chlorocebus sabaeus*, *Macaca mulatta*, *Macaca fascicularis*, and *Macaca nemestrina*) do not have complete high-speed rail model sequences. (b) The species *Maylandia zebra*, *Pandemis nyerei*, *Haplodriconis burtoni*, *Astrotelapia calliptera*, and *Oreochromis niloticus* all have high-speed rail sequences that contain repeat units coding phospholipase A2 inhibitor and Ly6/PLAUR domain-containing proteins. Their closed species (*Nolamprologus brichardi* and *Xiphophorus maculatus*) do not have complete high-speed rail model sequences. There are two insert sequences (putative RNA-directed DNA polymerase from transposon BS) in the first and second repeat units of *Maylandia zebra* and marked with black triangles. (c) The species *Homo sapiens*, *Pan troglodytes*, and *Papio Anubis* all have high-speed rail model sequences that contain repeat units coding testis-specific protein Y-linked (TSPY). Their closed species (*Macaca mulatta* and *Callithrix jacchus*) do not have complete high-speed rail model sequences. Each vertical short line represents one copy of SRS, and the ellipsis represents the variable copy number of SRS.

**Discussion**

A comparison of all gene clusters mediated by the high-speed rail model to date revealed that the high-speed rail model showed dynamics among different species/subspecies and strains. Moreover, their gene-coding sequences were characterized by high variation. Consequently, the differences in the number of copies of SRS and repeat units were most likely responsible for genetic adaptation to altered growth conditions or environmental stresses (Fig. 7). As previously studied, gene duplication amplification (GDA) is an expatiation/adaptation in response to changes in the environment [23,45,46]. Organisms need continuous genetic variations to adapt to the environment. Copy number variation (CNV) is a type of structural variant involving alterations in the number of copies of DNA [47]. The term CNV is used to describe all kinds of variations in the genome, including tandem genome repeats, duplication, and deletion [5]. CNV can be simple in structure, such as tandem gene duplication or may involve complex gains or losses of homologous sequences at multiple sites in the genome. CNVs influence gene expression, phenotype variability and adaptation by varying gene dosages and providing genetic materials of evolution.

Several studies [3–4] have indicated that tandemly arranged repeats with massive gene copies could be generated by repeated sequence-mediated homology recombination, but it is not known whether the molecular mechanisms among tandemly arranged repeats are the same. Here, we identified a high-speed rail model for tandem amplification mediated by SRSs with different repeat numbers in eukaryotes by screening 568 genomes. Amplified units from high-speed rail model encode various kinds of coding or non-coding sequences and increase gene dosage or generate gene diversity leading to adaptive evolution. The high-speed rail model shows dynamic unit numbers among closed species, subspecies, and ecotypes/strains (Supplementary Fig. S1). In conclusion, our study provides a special high-speed rail model, different from the rolling cycle model, to explain well-organized gene tandem amplification under environmental stress mediated by SRSs in all eukaryotes. The length heterogeneity of amplified units is due to variable copies of these SRSs.

**Methods**

**Screening the patterns of the high-speed rail model in genomes.** A total of 568 genomes representing all classic species of eukaryotes and prokaryotes were downloaded from the GenBank database. The reference genome of each species was chosen. Tandem repeat sequences from the whole genome were screened by Tandem Repeats Finder software [47] (version 4.09, match, mismatch and indel with the parameters of 2, 3 and 5) with the Linux system. A matrix of all short-repeated sequences (SRSs) with their information (start, end, copy number, percentage match, SRS) was derived from TRF. The program algorithm based on R language (version 4.0.3) was built with several screening conditions to screen all high-speed rail models. The set screening conditions contained the following: (1) the length of SRS was chosen as 2–200 bp; (2) the match between SRSs was larger than 90%; (3) the length of the amplified unit was 50–300,000 bp; (4) the nucleotide sequence similarity between units was larger than 90% through two sequence BLAST; and (5) the copy number of each SRS was consistent (If the copy number of each SRS is same, we will filter out these sequences). The coding sequence (protein-coding gene, uncharacterized gene, and hypothetical gene) and non-coding sequence (transposons,
Figure 5. The high-speed rail model shows dynamic unit numbers among subspecies. (a) The subspecies of *Canis lupus familiaris* and *Canis lupus dingo* have high-speed rail model sequences whose repeat unit encodes interferon alpha-1/2. (b) The subspecies of *Leishmania major*, *Leishmania infantum* and *Leishmania donovani* all have high-speed rail model sequences whose repeat unit encodes heat shock protein 83. Each vertical short line represents one copy of SRS, and the ellipsis represents the variable copy number of SRS.

**Genome PCR and third generation sequencing.** All the experimental materials were compiled with Regulation for the collection of genetic resources (This guideline was formulated by the Ministry of environmental protection of China and implemented in 2012). Plants should be collected according to the principle of non-injurious sampling of animals and fresh leaves tissues can be collected. All material was frozen in liquid nitrogen and stored at −80 °C before processing. High-speed rail models with different *Arabidopsis thaliana* ecotypes (Ler-0, Ws-0, Col-0 and Ws-1 wild-type) were chosen for third-generation resequencing. Library preparation was performed following the manufacturer’s instructions. Approximately 100 mg (1 cm²) of greenhouse-grown *Arabidopsis thaliana* leaves (the frozen leaves as a gift from Professor Ziqiang Zhu and Professor Weifeng Xu, the method of leaves tissues collection follows that described by Schmid et al.⁴⁸) and frozen muscle tissues of *L. dearborni* (the dead fish tissues as a gift from Professor Liangbiao Chen) was used for genomic DNA extraction. Briefly, tissues were incubated with 500 μl of homogenization buffer (0.4 M NaCl, 10 mM Tris–HCl pH 8.0, 2 mM EDTA pH 8.0, and 400 μg/ml proteinase K) at 55 °C overnight. Samples were spun down for 10 min at 12,000 g, and an equal volume of isopropanol was added to the supernatant. Samples were incubated at ~20 °C for 1 h and then centrifuged for 15 min at 4 °C and 12,000 g. The pellet was washed with 75% ethanol, dried and finally resuspended in 100 μl sterile dH₂O. Genomic DNA was amplified using a long PCR kit (M0533S, NEB)⁴⁹. Then the products were collected for sequencing (PacBio Sequel II, CCS model).

**Gene synteny analyses of 5S rRNA and other genes locus.** For the identification of 5S rRNA and other genes in other species, all annotated 5S rRNA and other genes from the GenBank database were checked for their series type. The divergence time of high-speed rail model-containing genes compared with nearest species were estimated using the real-time method, which has been shown to perform well in the calculation of divergence times for duplicated genes.

**Phylogenetic analyses.** The sequence of the high-speed rail model was subjected to BLASTX and BLASTN (https://blast.ncbi.nlm.nih.gov) analysis to confirm the gene encoded in the repeat unit, and other species with the same or similar high-speed rail model sequence were also found. Then, these species were checked until the related species without the high-speed rail model were found. When no species with the same high-speed rail
Figure 6. The high-speed rail model shows dynamic unit numbers among ecotypes or strains. (a) Dynamic unit numbers in different ecotypes of *Arabidopsis thaliana*. The names of different ecotypes with black, blue, and red colors are genomes from references of Nature, NCBI and our PCR and third-generation sequencing experiments, respectively. Arrows pointing to different ecotypes represent consistent copy numbers of SRS. The yellow box represents the TNP2 transposon. (b) Alignment of the heads and tails of different ecotypes of *Arabidopsis thaliana*. The sequences of repeat units from the head and tail of the high-speed rail model sequence and SRS are marked with red and green colors, respectively. (c) Dynamic unit numbers in strains of *Drosopila melanogaster*. (d) Alignment of the heads and tails of different strains of *Drosopila melanogaster*. The sequences of repeat units are marked in green colors. Each vertical short line represents one copy of SRS, and the ellipsis represents the variable copy number of SRS.

Figure 7. The high-speed rail model shows differences in the SRS and repeat units under different pressure. Each vertical short line represents one copy of SRS, and the ellipsis represents the variable copy number of SRS.
model were found in the first step, Genome Data Viewer (GDV) (https://www.ncbi.nlm.nih.gov/genome/gdv) was used to find related species. The genomes of these species were subjected to BLAST to determine whether they had the same model. After finding the flanking genes of the high-speed rail model, the sequences between the same flanking genes of relative species were analyzed to confirm once again whether there were similar sequences. TIMETREE29,51 (www.timetree.org) was used to determine the divergence time of the species, and to infer the generation time of the high-speed rail model. Tandem unit and AFPIII coding sequences were aligned by Clustal-W version 1.83 with default settings52. Phylogenetic trees were constructed using the neighbor-joining (NJ) algorithm with 1000 bootstrap replicates in MEGA version 453.

Sequence analysis of head/tail sequence. We used BLAST to confirm the head/tail (we artificially set a direction for each high-speed rail and marked head in the front and tail in the tail, and head means the first unit, tail means the last unit) sequence of the high-speed rail model and then calculated the sequence integrity (shown as a percentage of the complete repeat unit sequence length). 0 to 1 indicates the ratio of the length of the head/tail unit sequence to the length of the complete unit sequence in this high-speed rail. “1” is defined as 100% unit without any SRS sequence in head/tail, “0” is defined as only SRS in head/tail without any other high-speed rail model unit sequences, “0–1” means that this is an incomplete unit. To confirm whether the head/tail sequences of different strains were consistent, we aligned (CLUSTALX)44 their head/tail and nearby sequences. According to the length of the high-speed rail model of each strain, the numbers of “N” were arranged into a matrix, to analyze the model structure of different strains. Global and complete alignments were used to match the numbers, and the matched numbers were marked with the same color to show the difference more intuitively in conservation between the head/tail and the middle part of the model.

AFPIII locus assembly. BAC library construction and screening have been previously described20. Six AFPIII clones that covered most of the AFPIII locus were sequenced by shotgun library sequencing technology (Supplementary Method). Gene annotation were obtained by BLAST, and the contigs were compared with the National Center for Biotechnology Information (NCBI) database.

Chromosomal fluorescence in situ hybridization of AFPIII. The full-length digoxigenin-labeled AFPIII gene probe was hybridized to metaphase chromosomal preparations from L. dearborni kidney cells following a previously published protocol57.

Selection pressure analysis of AFPIII. A wrapper tool named EasycodeML56 were used for AFPIII selection pressure analysis. Branch site model (BSM) can be used to identify signals of episodic selection occurring along a specified branch after gene duplication57. The output of the BSM model is a table shown the model, Ln L, estimates of parameters and p-value.

Data accessibility
The sequences reported in this paper have been deposited in the GenBank database (LdBAC001 accession no. JX844826, LdBAC003 accession no. JX844828, LdBAC005-004 accession no. JX844827, and LdBAC007 accession no. JX844825). All sequences used in data analysis are available on NCBI at https://www.ncbi.nlm.nih.gov/ and UCL at http://mntweb.cs.ucl.ac.uk/mus/www/19genomes/19genomes.htm.

Received: 1 July 2021; Accepted: 24 January 2022
Published online: 10 February 2022

References
1. Sandegren, L. & Andersson, D. I. Bacterial gene amplification: Implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. 7, 578–588. https://doi.org/10.1038/nrmicro2174 (2009).
2. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. Nat. Rev. Genet. 10, 551–564. https://doi.org/10.1038/nrg2593 (2009).
3. Zhang, F. et al. The DNA replication fork/TE/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. 41, 849–853. https://doi.org/10.1038/ng.399 (2009).
4. Slack, A., Thornton, P. C., Magner, D. B., Rosenberg, S. M. & Hastings, P. J. On the mechanism of gene amplification induced under stress in Escherichia coli. PLoS Genet. 2, e48. https://doi.org/10.1371/journal.pgen.0020048 (2006).
5. Chen, L. et al. CNV instability associated with DNA replication dynamics: Evidence for replicative mechanisms in CNV mutagenesis. Hum. Mol. Genet. 24, 1574–1583. https://doi.org/10.1093/hmg/ddu572 (2015).
6. Andersson, D. I. & Hughes, D. Gene and adaptive evolution in bacteria. Annu. Rev. Genet. 43, 167–195. https://doi.org/10.1146/annurev-genet-102108-134805 (2009).
7. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. 40, 1124–1129. https://doi.org/10.1038/ng.213 (2008).
8. Cheung, A. K. Rolling-circle replication of an animal circovirus genome in a theta-replicating bacterial plasmid in Escherichia coli. J. Virol. 80, 8686–8694. https://doi.org/10.1128/JVI.00655-06 (2006).
9. Tartof, K. D. Unequal mitotic sister chromatid exchange as the mechanism of ribosomal RNA gene magnification. Proc. Natl. Acad. Sci. U. S. A. 71, 1272–1276. https://doi.org/10.1073/pnas.71.4.1272 (1974).
10. Szostak, J. W. & Wu, R. Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. Nature 284, 426–430. https://doi.org/10.1038/324426a0 (1980).
11. Oling, D., Masoom, R. & Kivnt, K. Loss of U1bp3 increases silencing, decreases unequal recombination in rDNA, and shortens the replicative life span in Saccharomyces cerevisiae. Mol. Biol. Cell 25, 1916–1924. https://doi.org/10.1091/mbc.E13-10-0591 (2014).
12. Harbers, K., Soriano, P., Muller, U. & Jaenisch, R. High frequency of unequal recombination in pseudoautosomal region shown by proviral insertion in transgenic mouse. Nature 324, 682–685. https://doi.org/10.1038/324682a0 (1986).
13. Tilley, S. A. & Birschtein, B. K. Unequal sister chromatid exchange: A mechanism affecting Ig gene arrangement and expression. J. Exp. Med. 162, 675–694. https://doi.org/10.1084/jem.162.2.675 (1985).

14. Wellauer, P. K., Davids, I. B., Brown, D. D. & Reeder, R. H. The molecular basis for length heterogeneity in ribosomal DNA from Xenopus laevis. J. Mol. Biol. 105, 461–486. https://doi.org/10.1016/0022-2836(76)90229-1 (1976).

15. Shoji, V. & Zhang, L. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol. Biol. Evol. 23, 2134–2141. https://doi.org/10.1093/molbev/msa085 (2006).

16. Wang, X., DeVries, A. L. & Cheng, C. H. Antifreeze peptide heterogeneity in an antarctic eel pout includes an unusually large major variant comprised of two 7 kDa type III AFPs linked in tandem. Biochim. Biophys. Acta 1247, 163–172. https://doi.org/10.1016/0167-4838(94)00205-u (1995).

17. Tekel, S. J. et al. Tandem histone-binding domains enhance the activity of a synthetic chromatin effector. ACS Synth. Biol. 7, 842–852. https://doi.org/10.1021/acssynbio.7b00281 (2018).

18. Nakajima, R. T., Cabral-de-Mello, D. C., Valente, G. T., Venere, P. C. & Martins, C. Evolutionary dynamics of rRNA gene clusters in cichlid fish. BMC Evol. Biol. 12, 198. https://doi.org/10.1186/1471-2148-12-198 (2012).

19. Cheng, C. C., Cziko, P. A. & Evans, C. W. Nonhepatic origin of notothenioid antifreeze reveals pancreatic synthesis as common mechanism in polar fish freezing avoidance. Proc. Natl. Acad. Sci. U.S.A. 103, 10491–10496. https://doi.org/10.1073/pnas.0603796103 (2006).

20. Deng, C., Cheng, C. H., Ye, H., He, X. & Chen, L. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc. Natl. Acad. Sci. U.S.A. 107, 21593–21598. https://doi.org/10.1073/pnas.1007883107 (2010).

21. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. 9, 938–950. https://doi.org/10.1038/nrg2482 (2008).

22. Andersson, D. I., Jerstrom-Hultqvist, J. & Nasvall, J. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb. Perspect. Biol. https://doi.org/10.1101/cshperspect.a017996 (2015).

23. Panchy, N., Lehti-Shiu, M. & Shiu, S. H. Evolution of gene duplication in plants. Science 347, 1812–1819. https://doi.org/10.1126/science.aau5656 (2019).

24. Chaisson, M. J. P. Multiple hybridization and mapping studies of the eukaryotic genome. Philos. Trans. R. Soc. Lond. B Biol. Sci. 366, 1213–1218. https://doi.org/10.1098/rstb.2010.0299 (2010).

25. Warburton, P. E., Waye, J. S. & Willard, H. F. Nonrandom localization of recombination events in human alpha satellite repeat unit variants: Implications for higher-order structural characteristics within centromeric heterochromatin. Mol. Cell Biol. 13, 6520–6529. https://doi.org/10.1128/mcb.13.10.6520 (1993).

26. Yu, A. et al. Comparison of human genetic and sequence-based physical maps. Nature 409, 951–953. https://doi.org/10.1038/35057185 (2001).

27. Hollingworth, M. A. & Swanson, B. J. Mcincus in cancer: Protection and control of the cell surface. Nat. Rev. Cancer 4, 45–60. https://doi.org/10.1038/nrc1251 (2004).

28. Huang, Q., Hu, R., Peng, C. & Chen, L. Expression of multi-domain type III antifreeze proteins from the Antarctic eelpout (Lycodictys dearborni) in transgenic tobacco plants improves cold resistance. Aquacult. Fisheries https://doi.org/10.1016/j.aaf.2019.11.006 (2019).

29. Pisors, D. et al. Chromosomal mapping of repetitive sequences in Hyphessobrycon equus (Characiformes, Characidae): a special case of the spreading of 55 rDNA clusters in a genome. Genetica 148, 25–32. https://doi.org/10.1007/s10709-000-00086-3 (2020).

30. Vienna, J., Wehner, S., Honerzu Siederdissen, C., Martinez-Lage, A. & Marz, M. Systematic analysis and evolution of 5S ribosomal DNA in metazoans. Heredity 111, 410–421. https://doi.org/10.1038/hdy.2013.63 (2013).

31. Ahmad, L. et al. Molecular mechanism of activation of human musk receptor OML16 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. Proc. Natl. Acad. Sci. U.S.A. 115, E9590–E9598. https://doi.org/10.1073/pnas.1710261115 (2018).

32. Zhou, X. et al. Interferon induced IFIT family genes in host antiviral defense. Int. J. Biol. Sci. 9, 200–208. https://doi.org/10.7150/ijbs.5613 (2013).

33. Grayfer, L., De Jesus Andino, F. & Robert, J. The haemoglobin (Xenopus laevis) type I interferon response to frog virus 3: new insight into ranavirus pathogenicity. J. Virol. 88, 5766–5777. https://doi.org/10.1128/JVI.02633-14 (2014).

34. Jancovich, J. K., Bremont, M., Touchman, J. W. & Jacobs, B. L. Evidence for multiple recent host species shifts among the Ranavirus (family Iridoviridae). J. Virol. 84, 2636–2647. https://doi.org/10.1128/JVI.01991-09 (2010).

35. Grayfer, L., De Jesus Andino, F. & Robert, J. Prominent amphibian (Xenopus laevis) tadpole type III interferon response to the frog virus 3 ranavirus. J. Virol. 89, 5072–5082. https://doi.org/10.1128/JVI.00051-15 (2015).

36. Gao, X. et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nucleic Acids Res. 47, 419–423. https://doi.org/10.1093/nature10141 (2011).

37. Scavetta, R. J. & Tauxe, D. Copy number changes of CNV regions in intersubspecific crosses of the house mouse. Mol. Biol. Evol. 27, 1845–1856. https://doi.org/10.1093/molbev/msq064 (2010).

38. Rice, A. M. & Mclysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. Nature 508, 1436–1436 (2017).

39. Weber, J. N. & Tong, W. Jumping gene gave fish a freshwater start. Science 364, 831–832. https://doi.org/10.1126/science.aax7936 (2019).

40. Ishikawa, A. et al. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. Science 364, 888–889. https://doi.org/10.1126/science.aau5656 (2019).

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

42. Schmid, M. et al. A gene expression map of Arabidopsis thaliana development. Nat. Genet. 37, 501–506. https://doi.org/10.1038/ng1534 (2003).

43. Jia, H., Guo, Y., Zhao, W. & Wang, K. Long-range PCR in next-generation sequencing: Comparison of six enzymes and evaluation on the MiSeq sequencer. Sci. Rep. 4, 5737. https://doi.org/10.1038/srep05737 (2014).

44. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34, 1812–1819. https://doi.org/10.1093/molbev/msx116 (2017).
51. Letunic, I. & Bork, P. Interactive tree of life (iTOl) v4: Recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. https://doi.org/10.1093/nar/gkz239 (2019).

52. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404 (2007).

53. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24, 1596–1599. https://doi.org/10.1093/molbev/msm092 (2007).

54. Kohli, D. K. & Bachhawat, A. K. CLOURE: Clustal output reformatter, a program for reformatting ClustalX/ClustalW outputs for SNP analysis and molecular systematics. *Nucleic Acids Res.* 31, 3501–3502. https://doi.org/10.1093/nar/gkg502 (2003).

55. Jiang, J., Gill, B. S., Wang, G. L., Ronald, P. C. & Ward, D. C. Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* 92, 4487–4491. https://doi.org/10.1073/pnas.92.10.4487 (1995).

56. Gao, F. et al. EasyCodeML: A visual tool for analysis using CodeML. *Ecol. Evol.* 9, 3891–3898. https://doi.org/10.1002/ece3.5015 (2019).

57. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. https://doi.org/10.1093/molbev/msi237 (2005).

Acknowledgements
We thank Professor Ziqiang Zhu (Nanjing Normal University), Weifeng Xu (Fujian Agriculture and Forestry University), Qiang Qiu (Northwestern Polytechnical University) and Liyu Chen (Fujian Agriculture and Forestry University) for materials support. This work was supported by the National Natural Science Foundation of China (31970388, 32170498) and National Key Research and Development Program of China (2018YFD0900602, 2021YFF0702000), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (ZJC21050), Natural Science Foundation of the Jiangsu Higher Education Institutions (grant no. 19KJB180003).

Author contributions
C.D. designed the study. C.D., H.C., M.G. and L.C. wrote the paper designed. H.C. and J.X. performed and analyzed the experiments. C.D., M.G., H.C., Z.Z., J.X., G.Z., X.X., H.L., R.Z., N.U., L.C., A.U., Z.Z., S.L., X.H., W.L. and J.L. analyzed the data. All authors reviewed the results and approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-06250-3.

Correspondence and requests for materials should be addressed to M.G., J.L., L.C. or C.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022