scientific reports

OPEN

Novel start codons introduce novel coding sequences in the human genomes

He Zhang & Yang Xie

Start-gain mutations can introduce novel start codons and generate novel coding sequences that may affect the function of genes. In this study, we systematically investigated the novel start codons that were either polymorphic or fixed in the human genomes. 829 polymorphic start-gain SNVs were identified in the human populations, and the novel start codons introduced by these SNVs have significantly higher activity in translation initiation. Some of these start-gain SNVs were reported to be associated with phenotypes and diseases in previous studies. By comparative genomic analysis, we found 26 human-specific start codons that were fixed after the divergence between the human and chimpanzee, and high-level translation initiation activity was observed on them. The negative selection signal was detected in the novel coding sequences introduced by these human-specific start codons, indicating the important function of these novel coding sequences.

Genetic mutations play key roles in the evolution of genes by providing resources for novel functions. Mutations can cause genetic polymorphism in the population, and contribute to the genetic diversity of the individuals[1]. During evolution, some mutations were fixed in the species and became species-specific genetic markers that contribute to both genetic and phenotypic differences distinguishing different species[2]. Single-Nucleotide Variant (SNV) is the most common type of variation in the population, and more than 80 million SNVs have been genotyped in large-scale genetic studies[1]. Many SNVs can alter the function of the gene by changing the protein sequence and are associated with phenotype diversities and diseases[3]. Several kinds of protein-altering SNVs were reported frequently in genetic studies, including missense SNVs[4], nonsense SNVs[5], read-through SNVs[6], and splicing-relevant SNVs[7]. Besides these SNVs, start-gain SNVs[8], which was located in the 5' untranslated region (5'UTR) of the mRNA, can also change the protein sequences via converting triple-nucleotides to a novel start codon before the original start codons of the coding sequence (CDS).

The translation of coding sequence in mRNA to protein is a key step in the central dogma, and the start codon plays an important role in translation initiation[9]. The stat-gain SNV can introduce a novel CDS before the original CDS, which may alter the function of this gene, and some start-gain SNVs are associated with human diseases[10,11]. However, the start-gain SNVs were not studied as commonly as other types of protein-altering SNVs, because such SNVs were usually annotated as non-coding SNVs in 5'UTR. During evolution, a start-gain SNV can be fixed in the population and become a species-specific start codon. However, there was no systematic genome-wide study on the human species-specific start codons.

In this study, we tried to investigate the novel start codons in the human genomes at both population-level and species-level. First, we wanted to know how many potential start-gain SNVs can be found in the natural human populations, and whether they were active in translation initiation. Second, we wanted to know whether there were human-specific start codons generated and fixed in the human genome after the divergence with the chimpanzee.

Results

Start-gain SNVs exist in the natural human populations. From 62 Yoruba individuals, 110 potential start-gain SNVs were identified, and each individual possessed 13 to 26 novel start codons in the genome (Supplementary Table 1). For each start-gain SNV, the 62 Yoruba individuals can be split into two groups according to whether the individual had a start-gain allele or not. Then we examined the ribosome occupancy for the sites with the start-gain allele and without the start-gain allele. To facilitate the analysis, only 86 start-gain SNVs that passed the following three criteria were used in the ribosome occupancy analysis: (1) at least five codons between

Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. email: he.zhang@utsouthwestern.edu; yang.xie@utsouthwestern.edu

the NSC and original start codon, (2) at least three individuals possessed NSC allele, and (3) from autosomes. Based on the aggregated results across 86 start-gain SNVs, the ribosome occupancy on the novel start codons in the individuals with the start-gain allele was four times higher than the homologous sites in the individuals without the start-gain allele ($P = 0.0093$) (Fig. 1). For the sites upstream of the novel start codons, the ribosome occupancy was extremely low, which means almost no ribosome was stacked before the novel start codons. The pattern of ribosome occupancy around the novel start codons is similar to the pattern observed around the start codons of all coding transcripts in the genome (Supplementary Fig. 4). In contrast, the sites following the novel start codons showed some signal of ribosome occupancy but not as much as the novel start codons. That was a common pattern observed around the canonical start codons. Instead, the ribosome occupancy around the sites without NSS alleles didn't show a similar pattern (Fig. 1), and more uniform ribosome occupancy was observed across different positions, which may represent random background noises. This observation suggests some novel start codons indeed have the potential to accumulate ribosomes and further initiate the translation.

Then we investigated the distribution of novel start codons in the natural populations using the dataset from the 1000 Genomes Project. Across all 2,504 individuals from 26 populations, 829 SNVs can introduce a potential novel start codon, and the frequencies of the start-gain alleles were from 0.019 to 99.94% (Supplementary Table 2). Although around 78.5% of these start-codon SNVs were rare (allele frequency less than 0.1%) in the populations, there were still 69 start-gain SNVs with a frequency greater than 1%, and 32 of them were common (> 5%) in the populations. On average, each individual carried 23 candidate novel start codons in the genome, and that number was 15 if we only considered the common start-gain SNVs with allele frequency greater than 5% (Supplementary Table 3). The median length of the novel coding sequences between the novel start codons and the original canonical start codons was 48nt (16 codons), and some of them can even reach hundreds of codons (Supplementary Table 2).

Interestingly, 12 start-gain SNVs were reported associated with diseases or phenotypes according to the records in the ClinVar database[12], but all of these SNVs were annotated as non-coding SNVs in 5'UTR (Supplementary Table 4). Based on the finding in this study, a potential explanation can be provided for the effect of these SNVs since these potential start-gain SNVs may introduce novel start codons and novel peptides which may affect the function of the protein products.

Novel start codons were generated and fixed in the human genome after the divergence between the human and chimpanzee. Start-gain SNPs are polymorphic in the human populations, and can only be found in some individuals. Next, we wanted to investigate whether some novel start codon mutations have been fixed in human populations. By a comparative genomics analysis of the human, chimpanzee, gorilla, and orangutan, 26 human-specific start codons were identified in the human genome (Supplementary Table 5). These start codons were fixed in the human populations but not observed in the chimpanzee, gorilla, and orangutan genomes, which means they were fixed in the human genome after the divergence between humans and chimpanzees. The length of the novel coding sequence between the human-specific start codon (hSSC) and the ancestral start codons (hASC) varied from 3 to 186nt, with a median value of 30nt (10 codons). To test whether the human-specific start codons initiate the translation, ribosome occupancy was evaluated for each human-specific start codon and their orthologous sites in the chimpanzee genome. 16 of 26 human-specific start codons had at least five codons between the human-specific start codon (hSSC) and the ancestral start codon (hASC), and they were used in the ribosome occupancy analysis.

For the ribosome-profiling dataset from 62 Yoruba individuals, the aggregated ribosome occupancy at the human-specific start codons (hSSC) was more than 4.9 times higher than the ancestral start codons (hASC)

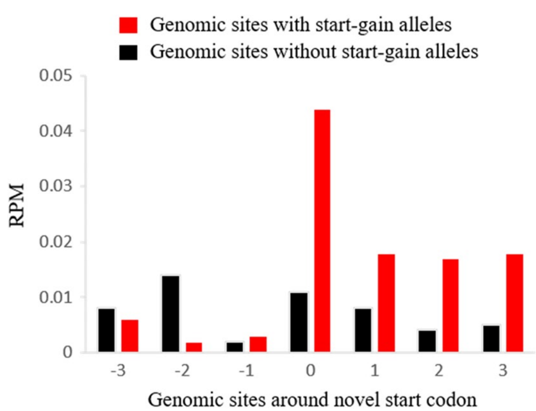

Figure 1. Ribosome occupancy around the start-gain SNVs in the 62 Yoruba individuals. For each polymorphic start-gain SNV, some individuals (red) possessed the start-gain allele introducing a novel start codon, and the other individuals (black) possessed an allele not introducing a start codon. Each position represents a codon or a segment of triple-nucleotides. '0' represents the novel start codons (AUG) for individuals with the start-gain allele or the corresponding triple-nucleotides for the individuals without the start-gain allele. Y-axis represents the mean RPM of ribosome occupancy aggregated across all start-gain SNVs.

which were the orthologous sites of the start codons of the chimpanzee genome (Wilcoxon signed-rank test, $P = 7.4 \times 10^{-16}$) (Fig. 2a–c). A similar pattern was observed for four Yoruba individuals from the other independent ribosome-profiling dataset ($P = 0.0043$) (Supplementary Fig. 1a, 1b), confirming the pattern was not due to the random effect in different experiments. That suggests more ribosomes were accumulated at the human-specific start codons (hSSC) than the ancestral start codons (hASC) in some genes, and the human-specific start codons (hSSC) were more active in the translation initiation than the ancestral start codons (hASC). Then we examined the ribosome occupancy in the chimpanzee genomes, and the ribosome occupancy at the orthologous sites of human-specific start codons (cOSC) was not higher than the start codons (cSC) (Fig. 2d, e). Combining the pattern of ribosome occupancy for both the human and chimpanzee, we can conclude that the orthologous sites of the human-specific start codons in the chimpanzee genome (cOSC) were not as active as human-specific start codons (hSSC) in recruiting ribosomes to initiate the translation.

In the chimpanzee genome, the ribosome occupancy at the orthologous sites (cOSC) of human-specific start codons still was higher than the nearby region, although it was lower than the canonical AUG start codons (cSC). We found that the ribosome occupancy at cOSC sites was majorly contributed by the CUG codon, which was known as the non-AUG start codon in mammalian genomes[13]. Considering that situation, we excluded two novel AUG start codons converted from CUG in the next round of analysis of ribosome occupancy. For the remaining 14 human-specific start codons, the ribosome occupancy in the human-specific start codons (hSSC) was still higher with the ancestral start codons (hASC) (Wilcoxon signed-rank test, $P = 1.3 \times 10^{-5}$) (Fig. 2f, g; Supplementary Fig. 1c, 1d), but the ribosome occupancy at the orthologous sites of the human-specific start codons in chimpanzee (cOSC) was much lower than that on the canonical start codons (cSC) (Wilcoxon signed-rank test, $P = 2.4 \times 10^{-4}$) (Fig. 2h, i). That suggests the non-CUG orthologous sites of human-specific start codons (cOSC) have very low activity of translation initiation, while the human-specific start codons (hSSC) showed

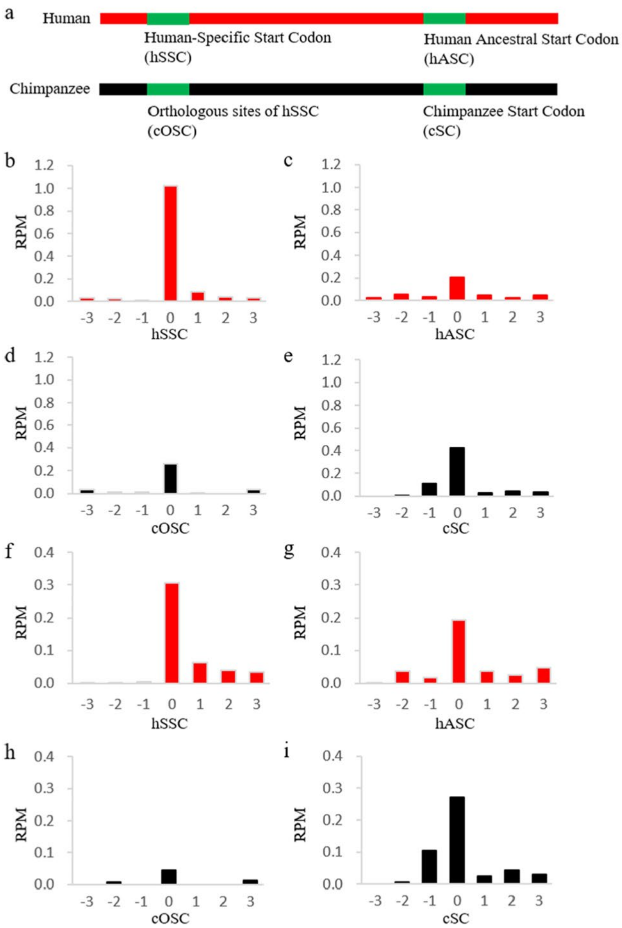

Figure 2. Ribosome occupancy for the human-specific start codons and adjacent positions in the 62 Yoruba individuals and 5 chimpanzee individuals. (**a**) The diagram shows the sites evaluated for ribosome occupancy. The mean ribosome occupancy was calculated across (**b**) 16 human-specific start codons (hSSC), (**c**) the corresponding downstream human ancestral start codons (hASC), (**d**) the orthologous sites in the chimpanzee (cOSC), and (**e**) start codons in the chimpanzee (cSC). After excluding human-specific start codons originated from CUG, the mean ribosome occupancy was calculated across (**f**) 14 non-CUG originated human-specific start codons (hSSC), (**g**) the corresponding downstream human ancestral start codons (hASC), (**h**) the orthologous sites in the chimpanzee (cOSC) and i) start codons in the chimpanzee.

a strong activity comparable with ancestral start codons (hASC). These observations supported the hypothesis that some human-specific start codons acquired significant activity to initiate the translation after mutating from their orthologous sites in the chimpanzee genome.

Negative selection acts on the new coding sequences introduced by the human-specific start codons. Since we observed significant translation initiation activity in the human-specific start codons, the novel peptides may be generated by the novel coding sequences between the human-specific start codons (not included) and the ancestral start codons (not included). If these novel peptides were generated in the translation elongation and have a function in the protein product, the evolution constraints can be expected on these novel peptides. To test that hypothesis, we examined the negative selection signals on these novel CDSs.

For each novel CDS following the human-specific start codons, the alignment with its orthologous sequences in the chimpanzee genome and gorilla genome was extracted from the genome alignment. Since the number of mutations observed in each CDS was limited, alignments for 15 CDSs, which is longer than 15nt, were combined to make a more reliable estimation of the mutation rates among branches. Although the orthologous sequences of these novel CDSs probably are not coding sequences in the chimpanzee genome and gorilla genome, we still treated them as 'coding sequences' to inspect the possible selection during evolution.

Branch models from PAML[14] were used to test whether the negative selection signals can be observed on the CDSs introduced by the human-specific start codons. In the first model, a fixed Ka/Ks value (ω_0) was assumed across all branches (Fig. 3a) to represent the null hypothesis that the Ka/Ks value was the same for each branch. In the second model, a foreground Ka/Ks value (ω_1) was assigned to the human branch, and the other branches shared the same background Ka/Ks value (ω_0) (Fig. 3b). In the second model, the estimated ω_1 was 0.36 for the human branch (Fig. 3c), while the ω_0 was 2.74 for other branches (Fig. 3d). The Chi-squared test ($P=0.028$) showed that the second model with two Ka/Ks values is significantly better than the first model with a fixed Ka/Ks ratio ($\omega_0=0.97$) across all branches. That indicates purifying selection has been operating in these human-specific CDSs to exclude nonsynonymous mutations altering the protein sequence. To check whether we can observe a similar pattern in the chimpanzee and gorilla branches, we also tested the third model in which the chimpanzee branch was assigned a foreground ω_1 and the fourth model in which the gorilla branch was assigned a foreground ω_1. In both models, the ω_1 was greater than one (Supplementary Figs. 2c and 2d), indicating no overall negative selection signal was observed on the chimpanzee or gorilla branch. That suggests some of the peptides generated from the CDSs introduced by human-specific start codons may play important functions in the final protein product, and the nonsynonymous mutations that change the protein sequences were not favored during human evolution. That provides another layer of evidence that human-specific start codons are functional in the translation initiation, and the novel CDSs following the novel start codons are functional.

Discussion

In this study, we investigated the novel start codon at the species level and population level. Novel start codons were generated continuously during evolution. 26 novel start codons have been fixed in the human genomes after the divergence with the chimpanzee and became a part of genetic mutations distinguishing between the human and other species. Besides that, we observed 829 polymorphic SNVs that potentially introduced novel start codons in genomes of different modern human populations. These SNVs were not fixed in the human genomes and contributed to the genetic diversity of human populations. Ribosome-profiling results strongly indicate that some of the novel start-codon SNVs in human populations and the fixed human-specific start-codons were active in translation initiation.

Mutations are important sources to generate novel phenotypes or functions in evolution. Novel start-codon SNV can change the sequence of the protein product directly by adding an extra peptide to the N-terminal of the original protein sequence, which may alter the function of the gene directly. Interestingly, several start-gain SNVs were reported to be associated with phenotypes or diseases, and these SNVs were annotated as non-coding

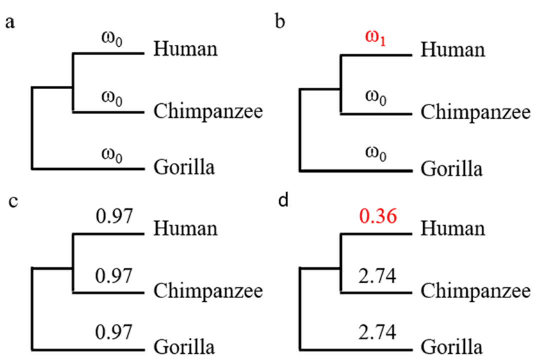

Figure 3. Different branch models were compared to address whether the human branch had a lower ω value than the other branches. (**a**) The first model assumed a fixed ω value (ω0) across all branches. (**b**) the second model assumes the human branch had an independent ω value (ω1) and the other branches shared the same ω value (ω0). The values of ω were estimated for (**c**) the first model and d) the second model using PAML.

mutations in 5'UTR. By assuming that these SNVs may introduce a novel start codon and an extra peptide that changed or disrupted the function of these genes, we got a candidate interpretation of the genetic mechanism for the association between these SNVs and phenotypes. That's another layer of evidence that the start-gain SNVs affected the function of the genes.

In the human genome, a lot of transcripts are involved in the nonsense-mediated mRNA decay (NMD) process, thus a novel start codon may not affect the function if it is located in a transcript with NMD. Among all 829 transcripts with a novel start codon, only seven of them have stop codons located more than 50 nucleotides upstream of the 3' most splice-generated exon-exon junctions (Supplementary Fig. 5). That indicates most of the start-codon gain transcripts are not likely undergoing NMD[15]. Then we checked the expression pattern of these 829 transcripts using the GTEx V8 (https://gtexportal.org). For each transcript, the maximum expression level across all tissues was used to represent the expression level of this transcript. Among 829 transcripts, 605 have an expression level greater than 1 TPM in at least one tissue (Supplementary Fig. 6). The expression levels of the 829 transcripts were significantly higher than the other transcripts in the human transcriptome (Wilcoxon rank sum test P-value $= 2.07 \times 10^{-83}$).

It's helpful to detect potential evolution selection signal to infer the potential role of these novel peptides generated by the coding sequences following the novel start codons. Although we cannot assay the evolution signal for each novel coding sequence due to the limited mutations observed after the divergence between the human and chimpanzee, we still observed the negative selection signal if the novel coding sequences from different human-specific start codons were combined. That suggests at least some of the novel coding sequences may play an important role in corresponding genes. However, some of the novel coding sequences may have a role in the regulation of gene expression, although they were annotated as coding sequences. We can not exclude the possibility that the negative selection signal observed on the human-specific novel coding sequences was caused by the regulatory element on these sequences.

Although non-AUG start codons were discovered in the mammalian genomes, we still only consider AUG as the start codon in this study since non-AUG start codons are rare in mammalian genomes. We did find that the CUG start codon had considerable activity in the translation initiation of some genes, but the other potential non-AUG start codons showed extremely low-level activity. Even in the situation where the AUG was converted from CUG, the translation ignition activity on AUG was still much higher than the CUG as a start codon. Thus it's reasonable to consider AUG converted from other triple nucleotides as a candidate novel start codon in most situations. If non-AUG triple nucleotides became AUG, that doesn't necessarily mean it became an active start codon, because usually some other sequence features such as Kozak sequence[16] may be required to promote the translation initiation. However, it's not clear what features are essential to initiate the translation. Thus we didn't require the existence of any of those features to define a novel start codon. That would improve the sensitivity to identify the candidate novel start codons, although some of the false positive 'novel start codons' may be identified mistakenly.

Methods

Identification of candidate novel start codon SNVs in the human populations. All of the genome positions in this study were based on the human reference genome GRCh38[17]. Genotypes were get from the 1000 Genomes Project[1]. The annotation of the human genome was from GENCODE v30[18]. Firstly, we identified all SNVs that changed non-AUG triple nucleotides to AUG in the 5'UTR of a transcript. Then an SNV can be kept only if the AUG was in the same reading frame with the downstream coding sequence and no stop codon was observed between the AUG and downstream original start codon. Furthermore, SNVs located in any known coding sequences were excluded. The remaining SNVs were considered as the start-gain SNVs.

Identification of human-specific start codons. Pairwise genome alignments (Syntenic Net) between the human (hg38) and chimpanzee (panTro6), gorilla (gorGor4), and orangutan (ponAbe3) were downloaded from the USCS genome browser[19]. The annotation of the human genome was from GENCODE v30[18], and the annotation for the chimpanzee genome (NCBI.105) was from NCBI RefSeq[20].

To reduce the possible false positives caused by misalignment and non-orthologous alignment, only the 'one-to-one' alignments (reciprocal single hit) were kept. Based on the filtered genome alignments, we identified all mutations which have different alleles between the human and chimpanzee genomes. If a mutation was located in the annotated start codon of a transcript and its homologous sequence in the chimpanzee genome was within the 5'UTR of a transcript, it was identified as a candidate mutation that introduced a human-specific start codon. Then the alleles at the mutation position in the human genomes were compared with their orthologous alleles in gorilla and orangutan, and a mutation was dropped if either gorilla or orangutan had the same allele as the human. Then we examined the remaining mutations in 2,504 individuals from the 1000 Genomes Project, and only the ones that were monomorphic in all of the 2,504 individuals were considered as fixed in the human populations. Finally, only the mutations fixed in the human populations were kept, and the start codons generated by those remaining mutations were considered human-specific start codons.

Processing of ribosome profiling dataset. The ribosome profiling data of 62 Yoruba individuals were from dataset GSE61742[21] of the NCBI GEO database. The ribosome profiling data of four Yoruba individuals and five chimpanzee samples were from dataset GSE71808[22] of the NCBI GEO database.

All reads were mapped to the curated non-coding RNA (including rRNAs, tRNAs, and snRNAs)[23–25] databases using STAR (2.7.1a)[26], and the reads mapped to any non-coding RNA were dropped. Then, the remaining reads were mapped to the reference genome using STAR, and only the uniquely mapped reads with a length between 26 and 32nt were kept. The starting position of the P-site for each read was estimated using the 12th nucleotide from 5' end for the reads with a length between 26 and 29nt and the 13th nucleotide from 5'end for the reads with a length between 30 and 32nt. Read counts at P-sites were normalized to the unit 'reads per million' (RPM) across the genome to represent the ribosome occupancy at every single nucleotide. Obvious periodicity was observed for the normalized read count at coding sequences (Supplementary Fig. 3). To calculate the ribosome occupancy for each codon, the RPMs at the first nucleotide of the codon, the second nucleotide of the codon, and the nucleotide before the first nucleotide of the codon were added up to tolerate error in the cleavage of ribosome footprint.

Estimation of evolution rate. PAML (4.9j)[14] was used to estimate the ratio between the nonsynonymous mutation rate and synonymous mutation rate, which is known as Ka/Ks or ω. Different branch models were compared to address whether the human branch had a higher Ka/Ks value than the other branches. The first model assumed a fixed Ka/Ks value (ω_0) across all branches (Fig. 3a), and the second model assumes the human branch had an independent Ka/Ks value (ω_1) and the other branches shared the same Ka/Ks value (ω_0) (Fig. 3b). The third and fourth models assumed the chimpanzee branch or the gorilla brand had independent Ka/Ks values correspondingly (Supplementary Fig. 2a and 2b). The significance of the difference between the likelihoods of models was evaluated using the Chi-squared test.

Data availability

The SNPs analyzed during the current study are available in the 1000 Genomes Project (https://www.internationalgenome.org/). The expression data used in this study are available in the GTEx Portal (https://gtexportal.org).

Received: 31 December 2022; Accepted: 7 May 2023
Published online: 19 May 2023

References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Varki, A. & Altheide, T. K. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res.* **15**, 1746–1758 (2005).
3. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. Pal, L. R. & Moult, J. Genetic basis of common human disease: insight into the role of missense SNPs from genome-wide association studies. *J. Mol. Biol.* **427**, 2271–2289 (2015).
5. Yngvadottir, B. *et al.* A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am. J. Hum. Genet.* **84**, 224–234 (2009).
6. Shibata, N. *et al.* Degradation of stop codon read-through mutant proteins via the ubiquitin-proteasome system causes hereditary disorders. *J. Biol. Chem.* **290**, 28428–28437 (2015).
7. Kurmangaliyev, Y. Z., Sutormin, R. A., Naumenko, S. A., Bazykin, G. A. & Gelfand, M. S. Functional implications of splicing polymorphisms in the human genome. *Hum. Mol. Genet.* **22**, 3449–3459 (2013).
8. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
9. Clark, B. F. & Marcker, K. A. The role of N-formyl-methionyl-sRNA in protein biosynthesis. *J. Mol. Biol.* **17**, 394–406 (1966).
10. Semler, O. *et al.* A mutation in the 5'-UTR of IFITM5 creates an in-frame start codon and causes autosomal-dominant osteogenesis imperfecta type V with hyperplastic callus. *Am. J. Hum. Genet.* **91**, 349–357 (2012).
11. von Bohlen, A. E. *et al.* A mutation creating an upstream initiation codon in the SOX9 5' UTR causes acampomelic campomelic dysplasia. *Mol. Genet. Genomic Med.* **5**, 261–268 (2017).
12. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
13. Starck, S. R. *et al.* Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. *Science* **336**, 1719–1723 (2012).
14. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
15. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* **23**, 198–199 (1998).
16. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
17. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
18. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).
19. Lee, C. M. *et al.* UCSC Genome Browser enters 20th year. *Nucleic Acids Res.* **48**, D756–D761 (2020).
20. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
21. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–7 (2015).
22. Wang, S. H., Hsiao, C. J., Khan, Z. & Pritchard, J. K. Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol.* **19**, 83 (2018).
23. Kuksa, P. P. *et al.* DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. *Bioinformatics* **35**, 1033–1039 (2019).
24. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184-9 (2016).
25. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
26. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

Acknowledgements

This work is supported by the National Institutes of Health [P30CA142543, R35GM136375, 1R01GM115473], and the Cancer Prevention and Research Institute of Texas [RP180805].

Author contributions
H.Z. designed the study. H.Z. analyzed and interpreted the data. H.Z. and Y.X. write the manuscript. All authors read and approved the final 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/s41598-023-34770-z.

Correspondence and requests for materials should be addressed to H.Z. or Y.X.

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.