Genomics 110 (2018) 347–354

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Functional relevance of synonymous alleles reflected in allele rareness in the population

Eu-Hyun Im[a], Yoonsoo Hahn[b], Sun Shim Choi[a,*]

[a] Division of Biomedical Convergence, College of Biomedical Science, Institute of Bioscience & Biotechnology, Kangwon National University, Chuncheon 24341, Republic of Korea

[b] Department of Life Science, Chung-Ang University, Seoul 06974, Republic of Korea

ARTICLE INFO

*Keywords:*
Synonymous codon usage bias
Common SNP
Rare SNP
Natural selection

ABSTRACT

We provide theoretical evidence supporting the non-neutrality of synonymous alleles by investigating the rareness of synonymous alleles in the population. We find a significantly greater number of synonymous rare alleles than conventional neutral alleles derived from noncoding regions. A permutation experiment shows that the rareness of synonymous alleles is not a byproduct of random statistical noise. We then compare the frequencies of synonymous rare alleles and common alleles in various functional contexts in which synonymous alleles are known to be involved. Subsequently, we perform logistic regression analysis to elucidate the effect size of each independent factor contributing to the rareness of synonymous alleles. Additionally, we show that changes in optimality caused by synonymous mutations resulting in rare SNPs in the population tend to be biased toward optimality loss. We think that our study will contribute to the development of novel strategies for identifying functional synonymous mutations.

## 1. Introduction

Codon usages are degenerate due to the redundancy of genetic codes, resulting in synonymous codons, i.e., divergent codons encoding the same amino acids in translating proteins from mRNA sequences. Synonymous sites have long been considered functionally neutral and evolve under the influence of neutral evolutionary force because mutations at synonymous sites do not affect protein structure and function. Synonymous substitution rates (dS) have thus been used as a null approximation in estimating protein evolutionary rates (dN/dS); dN/dS (also known as ω) has been used as a valid measure of the rate of protein evolution in most studies on molecular evolution [1–3]. Molecular evolutionary theory suggests that dN/dS should be equal or close to 1 if a given protein is dead or nonfunctional. On the contrary, dN/dS is less than 1 if a given protein is functional and the sites in the protein are under the influence of various functional constraints [4].

The argument for the functionality of synonymous codons is related to the controversy regarding what determines synonymous codon usage bias (SCUB). SCUB or non-random usage of synonymous codons is a pervasive phenomenon found across all living organisms from bacteria and yeasts to humans [5–7]. It remains controversial whether SCUB is due to mutational processes that depend on the nucleotide compositional bias or to natural selection acting on synonymous codons [5,8,9]. Natural selection is generally concluded to be a primary cause of SCUB in model organisms with large effective population sizes, whereas neutral evolution in mammals with small effective populations [10].

Notably, the idea that SCUB is an evolutionary consequence of natural selection is a key to elucidating the functionality of synonymous alleles. Numerous studies have provided evidence indicating that natural selection is responsible for non-random synonymous usage (i.e., the non-neutrality of synonymous codons) even in mammals, including humans, in various functional contexts [5,10–15]. In particular, synonymous mutations in mammals often cause serious problems in various biological functions, such as the regulation of mRNA splicing, transcription factor (TF) binding, and mRNA stability [16–19]. One research group showed that in humans, a number of synonymous disease variants contribute to splicing regulation, using a deep-learning algorithm [20]. Another group has reported significant enrichment of oncogene-derived synonymous mutations in human cancers [21]. Consistently, some synonymous sites are subject to evolutionary conservation and mutations that can cause disease, which leads to deposition of many disease-associated synonymous variants in databases such as dbDSM (http://bioinfo.ahu.edu.cn:8080/dbDSM/index.jsp) [22].

The functionality of synonymous codons has primarily been hinted by the correlation between the optimal codon (or preferred codon) and

* Corresponding author.
*E-mail address:* schoi@kangwon.ac.kr (S.S. Choi).

https://doi.org/10.1016/j.ygeno.2018.04.003
Received 2 November 2017; Accepted 9 April 2018
Available online 13 April 2018

the amount of its cognate tRNA in an organism. There are more cognate tRNA genes for optimal codons within the genome of an organism, and species-specific abundance of cognate tRNAs is observed [23–25]. The abundance of tRNAs is often associated with the speed or efficiency of the translation process, with the assumption that genes containing more optimal codons are translated more efficiently and rapidly [26–28]. Under this scenario, the selective pressure acting on optimal codons should be stronger for highly expressed genes than for genes expressed at low levels. In other words, synonymous mutations resulting in the replacement of optimal codons with non-optimal codons are expected to potentially be more "harmful" in highly expressed genes than in genes expressed at low levels, which is well reflected in the lower dS of highly expressed genes [12]. Mutations with harmful effects are presumed to be under the influence of purifying selection, leading to the rareness of synonymous alleles in the population.

There is no doubt that nonsynonymous alleles are more harmful and thus distributed at rarer frequencies than synonymous and other non-functional alleles in a population [29]. For this reason, researchers have primarily focused on nonsynonymous variants in searching for disease mutations in genome-wide association studies (GWASs) and studies of Mendelian diseases or cancer genomics [30–36]. Interestingly, however, a recent study showed that synonymous SNPs have a similar effect size to that of nonsynonymous SNPs in human disease association studies [37]. Reasonably, the same rule can be applied to synonymous mutations, such that synonymous alleles with rare frequencies in populations are more likely to be functional.

In the present work, we attempt to provide convincing analytical evidence of the non-neutrality of synonymous alleles in various functional contexts. The non-neutrality of synonymous alleles is investigated under the assumption that different sizes of functional constraints affecting different synonymous codon sites contribute to their rare occurrence in human populations.

## 2. Materials and methods

### 2.1. Obtaining SNP datasets and allele frequencies

After downloading the "dbSNP144" dataset constructed based on the "hg19" version of reference complete genome sequences through the UCSC genome browser by 'All SNP track' [38], categories such as nonsense, missense, synonymous, and introns, etc., were assigned to the SNPs designated "single" in the dataset. Rare variants of minor allele frequency (MAF) < 1% and common variants of MAF ≥ 1% were identified using the SNP information downloaded from 'Flagged SNPs track' and 'Common SNPs track', respectively. Note that the SNPs from 'Flagged SNPs track' without any information regarding allele frequencies were also considered rare alleles, because the SNPs in the 'Flagged SNPs track' are completely excluded from the 'Common SNPs track' and are clinically-associated (even though they are not necessarily risk alleles) as well. A possibility of inclusion of few common SNPs may not be problematic, because the inclusion of those common SNPs will lead our analysis to be more conservative. Subsequently, synonymous SNPs were selected from only those with annotated transcripts-related information. Reference and observed codons were annotated using the information parsed from the "CodingDbSnp track" in the 'All SNP track' of the UCSC genome browser. After excluding SNPs with two and more observed alleles or non-validated transcripts or SNPs derived from the Y chromosome, a total of 50,565 synonymous SNPs including 40,499 common synonymous SNPs (named scSNPs) and 10,066 rare synonymous SNPs (named srSNPs) were finally obtained.

These cleaned synonymous SNPs were mapped into genetic positions. For this purpose, RefSeq genes, particularly "NM_" prefixed validated mRNA sequences, were downloaded from the NCBI RefSeq database (http:/www.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/). As a result, a total of 14,297 genes corresponding to 15,292 transcripts were found to have synonymous SNPs in their coding regions; 40,499 scSNPs and 10,066 srSNPs were mapped into 13,661 and 3994 RefSeq genes, respectively.

### 2.2. Determining "highly expressed gene" group and "lowly expressed gene" group

Gene expression information was downloaded from the RNA-Seq atlas generated from 10 different healthy human tissues (http://medicalgenomics.org/rna_seq_atlas) [39]. The maximum reads per kilobase per million (RPKM) value of an mRNA detected among ten different tissues was considered the expression level of the mRNA. The top 5% and bottom 5% of expression levels were considered the "highly expressed gene" group and the "lowly expressed gene" group, respectively.

### 2.3. Obtaining CADD and phyloP scores and information on RNA structure, TFBS, splicing regulation, and GC content

Combined Annotation Dependent Depletion (CADD) was downloaded (http://cadd.gs.washington.edu/download, CADD v1.3) [40] to determine the deleteriousness of an allele. CADD scores were used as the "Scaled C-score", which is based on the rank of each variant relative to all possible 8.6 billion substitutions in the human reference genome. Evolutionary conservation of each SNP position was determined by phyloP downloaded from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/placentalMammals/). TFBS information was downloaded using the table browser (http://genome.ucsc.edu/cgi-bin/hgTables, tfbsConsSites table). All the downloaded information was mapped to the positions corresponding their synonymous sites for further analyses.

The degree of splicing regulation by synonymous SNPs was measured by the percentage splicing index (PSI) obtained from the "Set of Predicted Effects on Human Splicing Across the Entire Genome (SPIDEX)" database (https://www.deepgenomics.com/spidex/) [41]. |dPSI|, the absolute difference of PSI between a reference and altered allele, was used as the degree of splicing regulation. The effect of each SNP on the RNA secondary structure was estimated using RNAsnp software (http://rth.dk/resources/rnasnp/software) [42] by calculating the structural distances altered by each SNP based on a base-pairing probability matrix, "Mode 2", with default settings. GC and GC3 contents in srSNPs and scSNPs were estimated using codonW (Ver.1.4.2) (https://sourceforge.net/projects/codonw/).

### 2.4. Estimation of tRNA adaptation index

The tRNA adaptation index (tAI) was estimated according to the method developed by dos Reis et al. [43]. Refer to the original publication by dos Reis et al. for the detailed procedures. Briefly, after the absolute adaptiveness values $W_i$ are first estimated for each codon (or codon$_i$), the relative adaptiveness values, $w_i$, are obtained after normalizing $W_i$ values to the maximal $W_i$. Subsequently, the tAI of a gene $g$ is estimated by calculating the geometric mean of its codons as follows:

$$tAI(g) = \left(\prod_{k=1}^{lg} w_{i_{kg}}\right)^{1/lg}$$

where $i_{kg}$ and $lg$ represents the codon defined by the $k$th triplet on gene $g$ and the length of the gene after excluding stop codons, respectively. The tAI ranges from 0 to 1, which can be interpreted that a gene with a high tAI (or close to 1) has higher levels of optimal codon usage (or high levels of translation efficiency).

### 2.5. Statistical analysis

All the statistical analyses were analyzed using the R platform. For comparing srSNPs and scSNPs, Student's $t$-test was used for normally

distributed variables, otherwise Wilcoxon rank sum test (or Kolmogorov-Smirnov test) was used. All categorical variables were compared using Fisher's exact test. Functions including “wilcox.test,” “t.test,” “fisher.test,” and “ks.text” were used for the Wilcoxon rank sum test, Student's *t*-test, Fisher's exact test, and Kolmogorov-Smirnov test, respectively. Cumulative distributions were calculated using the “ecdf” function.

### 2.6. Logistic regression analysis

To predict the influential independent factors contributing to rare synonymous SNPs, we conducted a logistic regression analysis using the “glm” function. A total of five variables were used as independent variables, while two categories, i.e., common and rare, of synonymous SNPs were used as dependent variables. Optimal codons and TFBSs were treated as categorical variables, and the remaining were considered continuous variables. The odds ratios (ORs) were computed by exponentiation of the coefficients, and the *P* values and confidence intervals were obtained.

## 3. Results

### 3.1. Significantly more rare alleles in synonymous codon sites than in other nonfunctional alleles

Synonymous codons are still being ignored in studies of disease genomics based on the notion of functional neutrality, despite the fact that many studies have demonstrated otherwise, as described in the Introduction. It is assumed that there should be no significant difference between the frequencies of rare alleles derived from synonymous sites and from other putatively nonfunctional sites, such as introns, UTRs, and intergenic regions, if synonymous alleles are neutral. We first investigated this assumption by comparing the proportions of rare alleles among different genetic sites (Fig. 1A). As expected, the greatest proportion of rare alleles was in the nonsense category, followed by the missense category, whereas the lowest proportion was in the intron category. Interestingly, the proportions of rare alleles in the synonymous category were higher than those of introns, UTRs, and ncRNAs (i.e., conventional genomic regions harboring high frequencies of nonfunctional sites), suggesting that synonymous sites are not functionally neutral, as alleles in conventional nonfunctional regions might be.

Specifically, a total of 59,668 SNPs were synonymous among a total of 12,924,894 SNPs (dbSNP144), approximately 22.5% of which were defined as rare (see Methods). We next tested the likelihood that 22.5% of synonymous alleles would be rare alleles under a random or neutral expectation using a permutation analysis. Briefly, proportions of rare alleles were estimated during 10,000 permutations of throwing randomly the number of mutations located in each nonfunctional category including introns, UTRs, and ncRNAs, and plotted as histograms for each category of genomic regions (Fig. 1B). It was found that none of the nonfunctional categories generated greater proportions of rare alleles than the observed proportion of synonymous rare alleles.

Additionally, we observed in the ClinVar database, i.e., a database that provides information on diseases and their causing variants, that the disease-causing srSNPs among the total srSNPs were enriched by over 6-fold compared with the disease-causing scSNPs among the total scSNPs, which indicates that srSNPs are significantly more likely to have relevant functional effects than scSNPs ($P < 2.2e\text{-}16$, OR = 6.30) (Fig. 1C).

### 3.2. Synonymous allele rareness may reflect synonymous allele functionality

We first compared evolutionary conservation of genomic sites where srSNPs and scSNPs are respectively located to see whether synonymous allele rareness can be a good indicative of synonymous allele functionality. Evolutionary conservation is an important characteristic representing the functionality of sites in genomic regions. As expected, srSNPs were found to be significantly more enriched in conserved regions defined by phyloP > 2.0 than scSNPs ($P < 2.2e\text{-}16$, OR = 2.68) (Fig. 2A).

We next investigated whether srSNPs were more deleterious than scSNPs (Fig. 2B). The deleteriousness of an allele was estimated based on a measure known as the Combined Annotation Dependent Depletion (CADD) using the “scaled C-score” [40], with higher C-scores corresponding to a greater degree of deleteriousness. As shown in Fig. 2B, srSNPs exhibited significantly higher C-scores than scSNPs, indicating that srSNPs are likely significantly more harmful than scSNPs ($P < 2.2e\text{-}16$). As expected, missense SNPs showed significantly higher C-scores than synonymous SNPs (Sup Fig. 1). These two analyses confirm that investigation of synonymous allele rareness can reveal synonymous functionality.

### 3.3. Synonymous functionality on the regulation of gene expression investigated by synonymous allele rareness

The first synonymous functional context we investigated with synonymous allele rareness was on the regulation of gene expression. Highly expressed genes are known to show a more biased preference toward cognate tRNA contents for optimal codons than genes expressed at low levels in humans [6,12,27], which we confirmed by measuring the tRNA adaptation index (tAI; see Methods) (Sup Fig. 2). Note that tAI is an index for measuring translation efficiency determined by the abundance of tRNAs for each codon (dos Reis et al., 2004). Accordingly, we found that srSNPs were significantly enriched in highly expressed genes (srSNPs: scSNPs = 12.23%: 3.31%, OR = 4.08, and $P < 2.2e\text{-}16$), while no significant difference between scSNPs and srSNPs was found in genes expressed at low levels (srSNPs: scSNPs = 5.28%: 4.14%, OR = 0.77 and $P = 1$) (Fig. 3A).

In related to this result, we wondered whether srSNPs are more enriched than scSNPs in optimal codon sites, and whether optimal codon sites are more conserved than non-optimal codon sites. To investigate this question, information on optimal codon sites was obtained from Supek et al. [21], and mapped to each SNP site. Approximately 53.9% of srSNP, but only 41.6% of scSNP were found in optimal codon sites (OR = 1.86, $P < 2.2e\text{-}16$) (Fig. 2B), which means that srSNPs, rather than scSNPs, tend to be derived from mutations in optimal codon sites. Accordingly, the conservation scores of srSNP sites in optimal codon sites (Fig. 2C) were significantly higher than those of scSNP sites, although both srSNP sites and scSNP sites showed higher conservation scores in optimal codon sites than non-optimal codon sites.

We excluded the possibility that the excess rare synonymous alleles in optimal codon sites may be due to the higher GC or GC3 contents in genes containing those rare alleles by comparing the average proportions of GC and GC3 among groups of genes classified according to different contents of srSNPs and scSNPs within genes (Sup Table S1).

Next, we investigated how the size of the optimality changes between reference alleles and alternative alleles is differed by synonymous mutations causing reference alleles either to be srSNPs or to be scSNPs. An optimality score ($C_{opt}$) was obtained for each codon from previous studies [15,44]. As expected, the size of the optimality change ($|\Delta C_{opt}|$) caused by synonymous mutations was found to be larger for srSNPs than for scSNPs (Fig. 3D), which indicates that srSNPs have higher functional impact than scSNPs on the codon optimality change. This result may suggest that the optimality changes caused by synonymous mutations may be harmful particularly in ‘highly expressed genes’ group compared with ‘lowly expressed genes’ groups, given that we confirmed the enrichment of srSNPs in ‘highly expressed genes’ group (Fig. 3A).

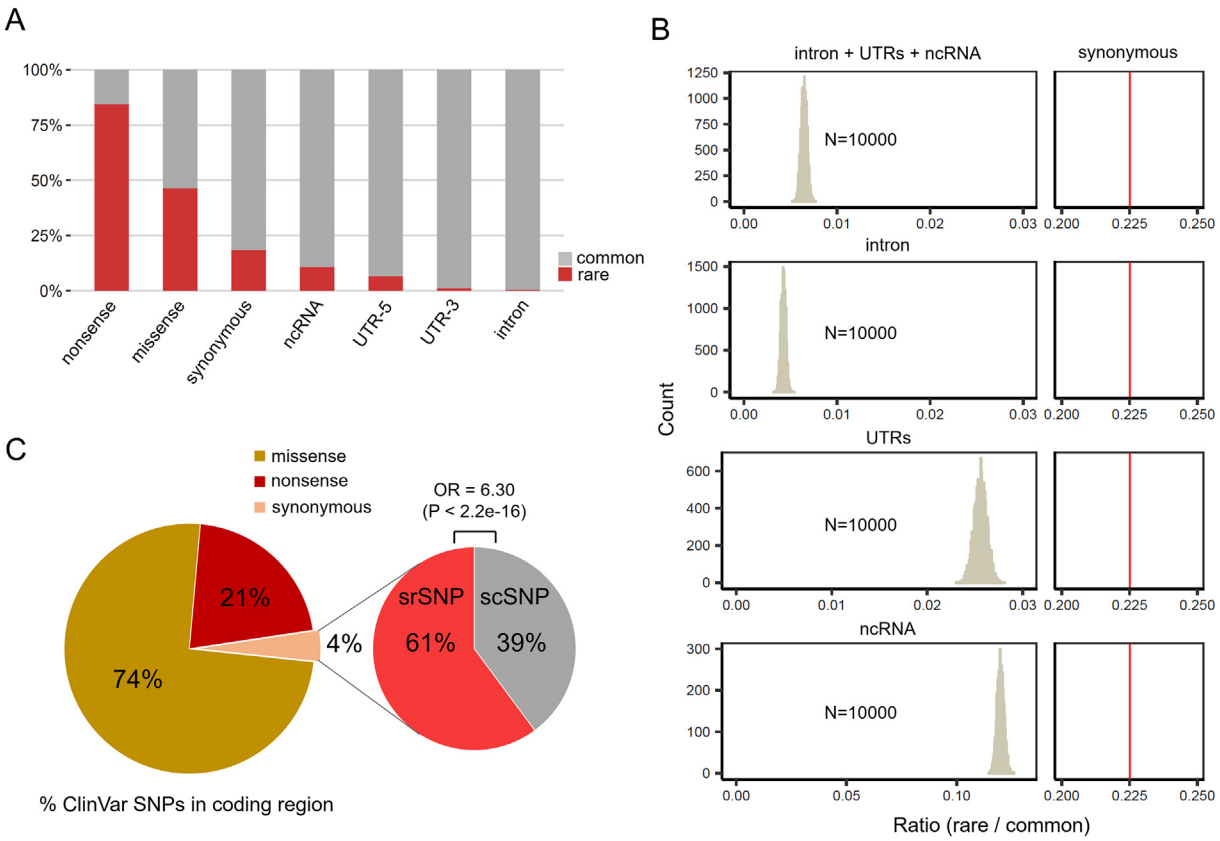


**Fig. 1.** Non-randomness of proportions of rare alleles in synonymous codon sites.
A. Stacked bar graph presents the proportions of common (gray) and rare SNPs (red) in dbSNP144. SNPs are divided into functional categories (i.e., nonsense, missense, synonymous) or by genomic sources from which they originated (i.e., 5′UTRs, 3′UTRs, ncRNAs, and introns). B. Histograms of proportions of rare alleles. The observed proportion of rare alleles in synonymous SNPs, 22.5%, is indicated by the red line in the right panel. Proportions of rare alleles for each categorical group estimated during each iteration of 10,000 permutations are plotted in the left panels. C. A pie chart representing the proportions of three types of disease-causing coding SNPs residing in the ClinVar database (see Methods). The portion of the pie representing approximately 4% disease-causing synonymous SNPs is enlarged to show how rare and common synonymous SNPs are proportioned; srSNPs: scSNPs = 61%: 39%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

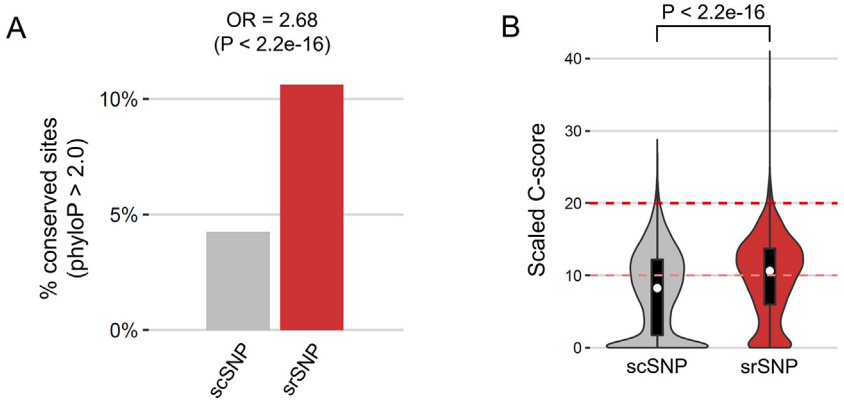


**Fig. 2.** Comparison of evolutionary conservation and allele deleteriousness between srSNPs and scSNPs.
A. The proportions of scSNPs and srSNPs were estimated in highly conserved regions with phyloP > 2.0. *P* values were estimated with a one-tailed Fisher's exact test. B. Scaled-C scores (see main text) were compared between scSNPs and srSNPs using violin plots. The dark red dotted line represents the top 1% of deleteriousness (C-score = 20), and the pink dotted line represents the top 10% of deleteriousness (C-score = 10). *P* values were calculated by Wilcoxon rank sum test. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *3.4. Other synonymous functional contexts investigated by synonymous allele rareness*

Splicing regulation has also been implicated as playing a functional role in the determination of synonymous sites. Therefore, we investigated whether srSNPs, rather than scSNPs, are linked to the sites involved in splicing mis-regulation. An index of splicing mis-regulation, |dPSI|, was assigned to each synonymous site, and the results were then compared between srSNPs and scSNPs (see Methods). As expected, the |dPSI| of srSNPs was significantly higher than that of scSNPs when the comparison was conducted only for highly conserved sites with phyloP > 2 (Fig. 4A), indicating that srSNPs have a greater impact on splicing mis-regulation. In relation to this result, we observed that srSNPs tend to be located within shorter distances from the nearest splice site than scSNPs (Sup Fig. 3).

A previous study showed that synonymous codon usage is related to exonic TF binding in the human genome [19]. Therefore, we investigated whether srSNPs are more enriched in exonic TFBSs than scSNPs. In fact, we observed that a slightly higher proportion of srSNPs (28.6%) than scSNPs (25.4%) were located in exonic TFBSs (OR = 1.17

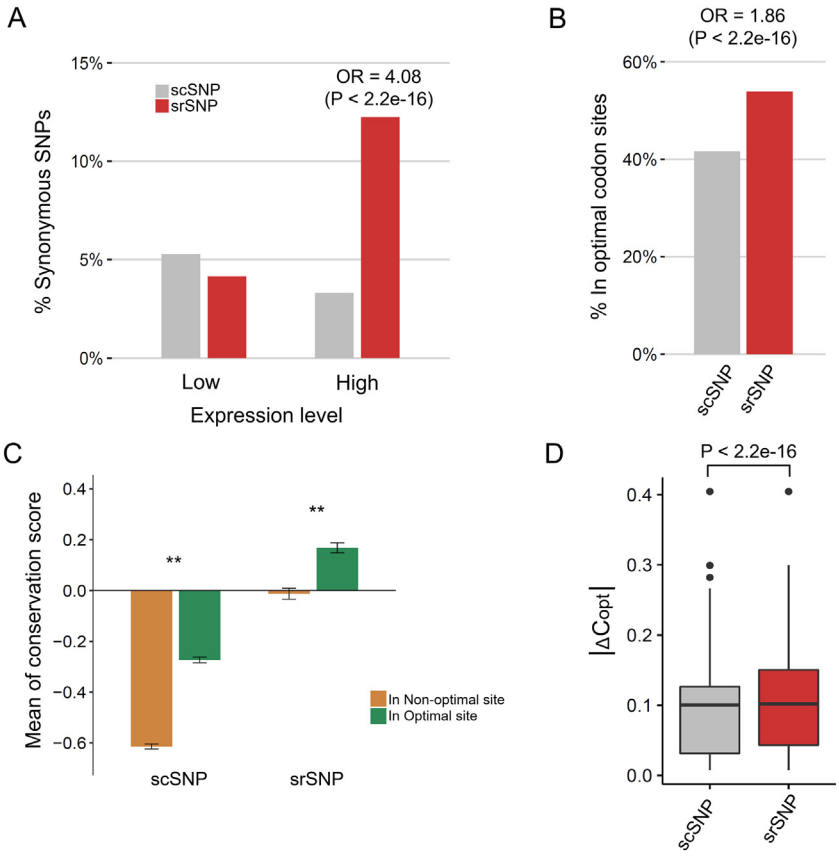


**Fig. 3.** Comparisons of srSNPs and scSNPs using factors associated with the regulation of gene expression.
A. Proportions of srSNPs and scSNPs are estimated for the 'lowly expressed gene (Low)' group and the 'highly expressed gene (High)' group. Please refer to the Methods for the description of how expression levels of genes were estimated. The bottom 5% and the top 5% were considered "Low" and "High", respectively. *P* values were calculated using one-tailed Fisher's exact tests. B. Proportions of scSNPs and srSNPs residing at optimal codon sites are plotted, referring the information on optimal codons reported by Supek et al. *P* values are estimated by one-tailed Fisher's exact test. C. Degrees of evolutionary conservation determined by phyloP scores are compared between optimal codon sites and non-optimal codon sites for scSNPs and srSNPs. $^{**}P < .0001$ by one-tailed Student's *t*-tests, and error bars indicate the standard deviations. D. $|\Delta C_{opt}|$ values were plotted as a boxplot for srSNPs and scSNPs. *P* values were calculated using the Wilcoxon rank sum test.

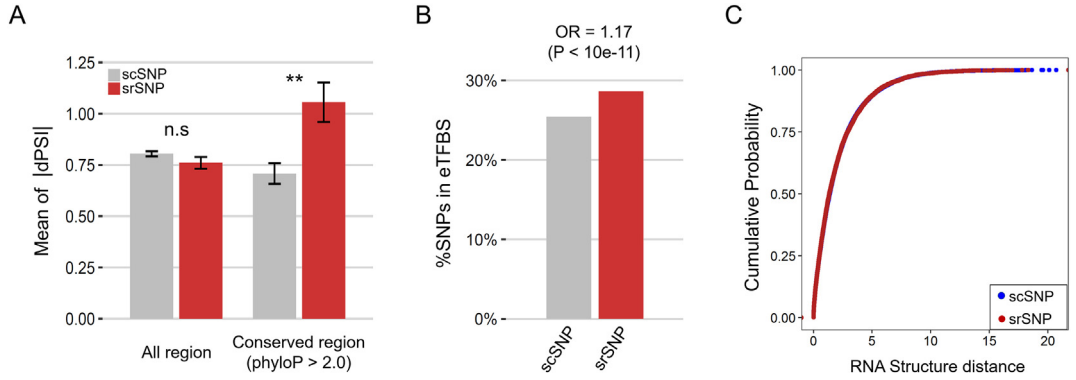


**Fig. 4.** Comparisons of srSNPs and scSNPs using other functional factors.
A. Investigation of the splicing effect of synonymous mutations, considering either all sites or only highly conserved sites. Each bar represents the mean |dPSI| for scSNPs or srSNPs. $^{**}P < 0.01$ by a two-tailed Student's *t*-tests. B. Proportions of SNPs located in TFBSs were compared between scSNPs and srSNPs. *P* values were estimated by one-tailed Fisher's exact tests. C. Plot of the cumulative probabilities of RNA structure distances for scSNPs and srSNPs. Statistical differences were estimated with the Kolmogorov-Smirnov test ($P = 0.70$). RNAsnp software was used to calculate the effect of each SNP on RNA secondary structural information (see Methods).

and $P <$ 10e-11) (Fig. 4B).

Additionally, we investigated the implicated role of synonymous alleles in the regulation of RNA secondary structure and found no significant difference between scSNPs and srSNPs (Fig. 4C). This result may indicate that synonymous alleles may play minimal roles in RNA secondary structure, at least in human populations.

### *3.5. Estimation of the effect size of independent factors contributing to the rareness of synonymous alleles*

Thus far, we have provided evidence supporting the divergent usage of synonymous codons in various functional contexts, including the regulation of translational efficiency, TF binding, and splicing regulation, as shown in several previous studies [6,17–19,27,45]. It is expected that mutations that occur in synonymous codons located in different functional contexts may have different degrees of functional effects against natural selection, such that all these functional contexts influence with different degrees on the rareness of synonymous alleles in an intermingled manner. Hence, we sought to discriminate an independent effect size of each factor influencing the rareness of synonymous alleles. We therefore chose five different factors that are associated with these functional contexts in which synonymous codons might play roles, including "$|\Delta C_{opt}|$", "optimal codon", "|dPSI|", "TF binding", and "RNA structure". Logistic regression analysis was then conducted using a generalized linear model (see Methods) to determine how much each factor contributes to srSNPs.

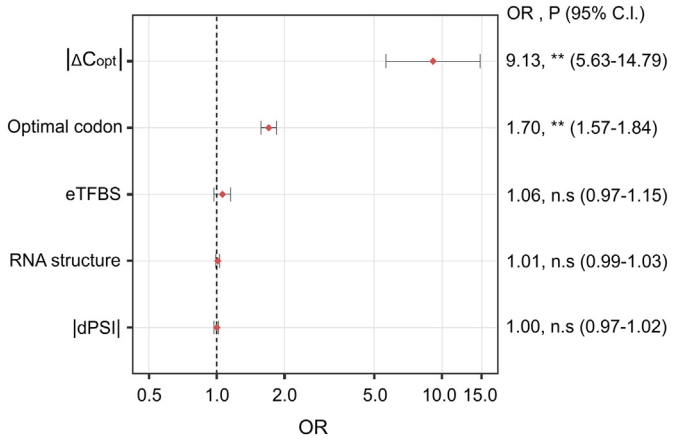


**Fig. 5.** Logistic regression analysis for five independent variables involved in the rareness of synonymous alleles.
Logistic regression analysis using five independent factors associated with srSNPs (see Methods). Error bars indicate the 95% confidence intervals (CIs). $^{**}P < 0.0001$. OR means odds ratio.

We found that the size of the optimality change ($|\Delta C_{opt}|$) corresponded to the highest odds ratio, ranging from 5.63 to 14.79, which indicates up to 14.8-fold-greater odds of a SNP being an srSNP compared with an scSNP (Fig. 5). Interestingly, according to this analysis, "$|\Delta C_{opt}|$" exhibited a greater effect size than "optimal codon", indicating that mutations influencing the size of optimality change have a greater functional impact, on the synonymous allele rareness, than mutations influencing optimality itself. As expected, |dPSI|, eTFBS and RNA structure showed relatively small effects on the rareness of synonymous alleles (Fig. 5).

### 3.6. A biased direction of optimality changes caused by synonymous mutations

Given that the size of the codon optimality changes was found to be most strongly influenced by synonymous mutations, we decided to perform an in-depth investigation of the positive or negative changes in optimality at the codon level. The direction of optimality changes was estimated by log2 transformation of the ratio between $C_{opt}$ scores of altered alleles and reference alleles as follows: $log_2Op = log_2[C_{opt}$ of altered allele/$C_{opt}$ of reference allele] (Sup Fig. 4). As a result, codons were primarily found to end in G or C (mostly with high optimality scores, Sup Table S2) and were all associated with optimality losses (i.e., $log_2Op < 0$), regardless of whether they were srSNPs or scSNPs (Fig. 6). Notably, the proportion of srSNPs responsible for optimality losses in these codon sites was greater than that of scSNPs (Sup Table S2). In contrast, mutations in primarily A- or T-ending codons (corresponding to mostly low optimality scores, Sup Table S2) were linked to optimality gains (i.e., $log_2Op > 0$) (Fig. 5), and the proportion of srSNPs responsible for optimality gains was less than that of scSNPs (Sup Table S2). It is also worth noting that scSNPs were relatively evenly distributed across all synonymous codon sites, while srSNPs exhibited highly biased distributions in sites of optimality losses (Fig. 6).

Consistently, the $C_{opt}$ values of srSNPs were found to be significantly lower overall than those of scSNPs (Sup Fig. 5), indicating that mutations resulting in srSNPs tend to be associated with optimality loss rather than optimality gain.

## 4. Discussion

In the present study, we confirmed the non-neutrality of synonymous alleles by showing that synonymous alleles that are linked to known functional contexts are existed at a low frequency in the population. More importantly, we estimated an independent effect size of each factor representing for each functional context influencing the rareness of synonymous alleles and found that the size of the optimality change, $|\Delta C_{opt}|$, is the most significant factor affecting the rareness of synonymous alleles. We designed a novel theoretical strategy for investigating the functionality of synonymous codons based on the rareness of synonymous alleles in the population and showed that the strategy can be successfully applied for such investigations.

According to the theory of population genetics, high-impact mutant alleles cannot increase their frequency in a population. We challenged the notion of neutrality of synonymous alleles by asking why synonymous alleles are significantly rarer than other neutral alleles in populations. Approximately 22.5% of the synonymous alleles deposited in dbSNP are rare, and we showed here that this proportion cannot be a byproduct of random noise (Fig. 1B).

We hypothesized that synonymous alleles should be rare in the population if synonymous alleles are truly functional in any context and are under the influence of purifying selection. Basically, we compared the frequencies of srSNPs and scSNPs to determine whether srSNPs are significantly more enriched in functional regions than scSNPs. Through this analysis, we confirmed that synonymous alleles are functional in various contexts, including the regulation of translation efficiency by codon optimality, splicing regulation, and TF binding regulation. Moreover, we were able to measure the magnitude of the impact of synonymous mutations involved in several functional contexts on the rareness of synonymous alleles, and we revealed that synonymous mutations affecting codon optimality have the largest impact (Fig. 5). This result gives rise to another question, of whether synonymous mutations affecting changes in optimality are more harmful than other synonymous mutations occurring in different functional contexts, which should be thoroughly addressed by experiments in the future.

Although recent studies have provided genetic and empirical evidence of the functionality of synonymous alleles acting in various contexts, the effect of codon optimality in the regulation of translation efficiency has been the best studied [25,27,46,47]. In particular, Presnyak et al. [48] recently provided direct empirical evidence of the functional impact of synonymous mutations linked to changes in codon optimality. They showed that converting optimal codons into non-optimal codons causes mRNA destabilization, which is consistent with our finding from the present study that srSNPs residing in optimal codon sites are biased toward "optimality loss" (Fig. 6 and Sup Fig. 5). In addition, they showed that codon optimality is associated with the translational elongation rate, which is consistent with our results showing that srSNPs are significantly enriched in highly expressed genes (Fig. 3A).

Taken together, the evidence presented herein shows that synonymous variants cannot be ignored when searching for human disease-associated or disease-causing alleles. However, it remains the case that most synonymous variants are excluded and ignored in further functional validation steps, partly because no good strategy for exploring synonymous function has yet been developed.

We believe that the present study will contribute not only to understanding the molecular characteristics of synonymous alleles but also to the development of strategies for exploring their functionality in the future.

## Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2016R1D1A1B03930411).

## Author's contributions

SSC conceived the study. EI prepared all figures. SSC and EI wrote

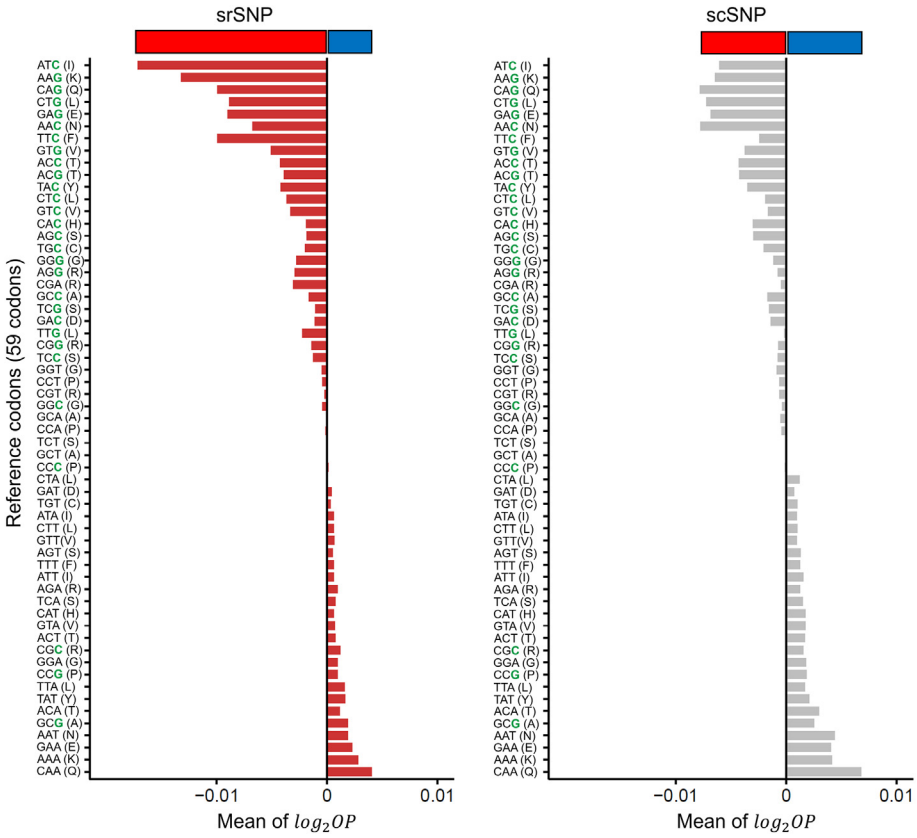


**Fig. 6.** Comparison of the optimality change caused by synonymous mutations between srSNPs and in scSNPs.
Each bar represents the size and direction of optimality change estimated by the average of $log_2Op$ values. $log_2Op > 0$ and $log_2Op < 0$ indicate "optimality gain" and "optimality loss", respectively. The arithmetic averages of the $log_2Op$ values and the 59 codons are shown on the x- and y-axes, respectively. The red bar and blue bar at the top of both plots represent the maximum range of the negative and positive average values of $log_2Op$, sequentially. The G or C ending of each codon is indicated in green face. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the manuscript. YH participated in data discussion. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.04.003.

## References

[1] J.L. King, T.H. Jukes, Non-darwinian evolution, Science 164 (1969) 788–798.
[2] M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, Mol. Biol. Evol. 3 (1986) 418–426.
[3] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.
[4] Z. Yang, J.P. Bielawski, Statistical methods for detecting molecular adaptation, Trends Ecol. Evol. 15 (2000) 496–503.
[5] J.B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42.
[6] A. Doherty, J.O. McInerney, Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates, Mol. Biol. Evol. (2013) mst128.
[7] A.A. Komar, The Yin and Yang of codon usage, Hum. Mol. Genet. (2016) ddw207.
[8] S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 3480–3485.
[9] M. Nabiyouni, A. Prakash, A. Fedorov, Vertebrate codon bias indicates a highly GC-rich ancestral genome, Gene 519 (2013) 113–119.
[10] J.V. Chamary, J.L. Parmley, L.D. Hurst, Hearing silence: non-neutral evolution at synonymous sites in mammals, Nat. Rev. Genet. 7 (2006) 98–108.
[11] D.C. Shields, P.M. Sharp, D.G. Higgins, F. Wright, "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons, Mol. Biol. Evol. 5 (1988) 704–716.
[12] D.A. Drummond, C.O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution, Cell 134 (2008) 341–352.
[13] Z. Yang, R. Nielsen, Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage, Mol. Biol. Evol. 25 (2008) 568–579.
[14] S.A. Shabalina, N.A. Spiridonov, A. Kashina, Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity, Nucleic Acids Res. 41 (2013) 2073–2094.
[15] W. Gu, C.I. Gurguis, J.J. Zhou, Y. Zhu, E.A. Ko, J.H. Ko, T. Wang, T. Zhou, Functional and structural consequence of rare Exonic single nucleotide polymorphisms: one story, two Tales, Genome Biol. Evol. 7 (2015) 2929–2940.
[16] J.V. Chamary, L.D. Hurst, Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals, Genome Biol. 6 (2005) R75.
[17] J.L. Parmley, J.V. Chamary, L.D. Hurst, Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers, Mol. Biol. Evol. 23 (2006) 301–309.
[18] M.F. Lin, P. Kheradpour, S. Washietl, B.J. Parker, J.S. Pedersen, M. Kellis, Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes, Genome Res. 21 (2011) 1916–1928.
[19] A.B. Stergachis, E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E.M. LeProust, J.M. Akey, Exonic transcription factor binding directs codon choice and affects protein evolution, Science 342 (2013) 1367–1372.
[20] H.Y. Xiong, B. Alipanahi, L.J. Lee, H. Bretschneider, D. Merico, R.K. Yuen, Y. Hua, S. Gueroussov, H.S. Najafabadi, T.R. Hughes, The human splicing code reveals new insights into the genetic determinants of disease, Science 347 (2015) 1254806.
[21] F. Supek, B. Miana, J. Valcrcel, T. Gabaldn, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers, Cell 156 (2014) 1324–1335.
[22] P. Wen, P. Xiao, J. Xia, dbDSM: a manually curated database for deleterious synonymous mutations, Bioinformatics (2016) btw086.
[23] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, T. Ikemura, Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis, J. Mol. Evol. 53 (2001) 290–298.
[24] W. Ran, P.G. Higgs, The influence of anticodon–codon interactions and modified bases on codon usage bias in bacteria, Mol. Biol. Evol. 27 (2010) 2129–2140.
[25] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.
[26] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, Y. Pilpel, An evolutionarily conserved mechanism for controlling the efficiency of protein translation, Cell 141 (2010) 344–354.
[27] Y.Y. Waldman, T. Tuller, T. Shlomi, R. Sharan, E. Ruppin, Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages, Nucleic Acids Res. 38 (2010) 2964–2974.
[28] T.E. Gorochowski, Z. Ignatova, R.A. Bovenberg, J.A. Roubos, Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate, Nucleic Acids Res. (2015) gkv199.

[29] L.B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans, Nat. Genet. 40 (2008) 340–345.
[30] V. Ramensky, P. Bork, S. Sunyaev, Human non-synonymous SNPs: server and survey, Nucleic Acids Res. 30 (2002) 3894–3900.
[31] R.K. Thomas, A.C. Baker, R.M. DeBiasi, W. Winckler, T. LaFramboise, W.M. Lin, M. Wang, W. Feng, T. Zander, L.E. MacConaill, High-throughput oncogene mutation profiling in human cancer, Nat. Genet. 39 (2007) 347–351.
[32] J. Hampe, A. Franke, P. Rosenstiel, A. Till, M. Teuber, K. Huse, M. Albrecht, G. Mayr, La Vega De, M. Francisco, J. Briggs, A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1, Nat. Genet. 39 (2007) 207–211.
[33] S. Romeo, J. Kozlitina, C. Xing, A. Pertsemlidis, D. Cox, L.A. Pennacchio, E. Boerwinkle, J.C. Cohen, H.H. Hobbs, Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease, Nat. Genet. 40 (2008) 1461–1465.
[34] R. Calabrese, E. Capriotti, P. Fariselli, P.L. Martelli, R. Casadio, Functional annotations improve the predictive score of human disease-related mutations in proteins, Hum. Mutat. 30 (2009) 1237–1244.
[35] S.B. Ng, K.J. Buckingham, C. Lee, A.W. Bigham, H.K. Tabor, K.M. Dent, C.D. Huff, P.T. Shannon, E.W. Jabs, D.A. Nickerson, Exome sequencing identifies the cause of a mendelian disorder, Nat. Genet. 42 (2010) 30–35.
[36] M. Li, J.S. Kwan, S. Bao, W. Yang, S. Ho, Y. Song, P.C. Sham, Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies, PLoS Genet. 9 (2013) e1003143.
[37] R. Chen, E.V. Davydov, M. Sirota, A.J. Butte, Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association, PLoS One 5 (2010) e13574.
[38] M.L. Speir, A.S. Zweig, K.R. Rosenbloom, B.J. Raney, B. Paten, P. Nejad, B.T. Lee, K. Learned, D. Karolchik, A.S. Hinrichs, The UCSC genome browser database: 2016 update, Nucleic Acids Res. 44 (2016) D725.
[39] M. Krupp, J.U. Marquardt, U. Sahin, P.R. Galle, J. Castle, A. Teufel, RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing, Bioinformatics 28 (2012) 1184–1185.
[40] M. Kircher, D.M. Witten, P. Jain, B.J. O'roak, G.M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants, Nat. Genet. 46 (2014) 310–315.
[41] H.Y. Xiong, B. Alipanahi, L.J. Lee, H. Bretschneider, D. Merico, R.K. Yuen, Y. Hua, S. Gueroussov, H.S. Najafabadi, T.R. Hughes, The human splicing code reveals new insights into the genetic determinants of disease, Science 347 (2015) 1254806.
[42] R. Sabarinathan, H. Tafer, S.E. Seemann, I.L. Hofacker, P.F. Stadler, J. Gorodkin, RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs, Hum. Mutat. 34 (2013) 546–556.
[43] M. dos Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection, Nucleic Acids Res. 32 (2004) 5036–5044.
[44] T. Zhou, M. Weems, C.O. Wilke, Translationally optimal codons associate with structurally sensitive sites in proteins, Mol. Biol. Evol. 26 (2009) 1571–1580.
[45] Y.Y. Waldman, T. Tuller, A. Keinan, E. Ruppin, Selection for translation efficiency on synonymous polymorphisms in recent human evolution, Genome Biol. Evol. 3 (2011) 749–761.
[46] Z. Zhou, Y. Dang, M. Zhou, L. Li, C. Yu, J. Fu, S. Chen, Y. Liu, Codon usage is an important determinant of gene expression levels largely through its effects on transcription, Proc. Natl. Acad. Sci. 113 (2016) E6125.
[47] Y. Lavner, D. Kotlar, Codon bias as a factor in regulating expression via translation rate in the human genome, Gene 345 (2005) 127–138.
[48] V. Presnyak, N. Alhusaini, Y. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K.E. Baker, B.R. Graveley, Codon optimality is a major determinant of mRNA stability, Cell 160 (2015) 1111–1124.