Distribution and Effects of Nonsense Polymorphisms in Human Genes

Yumi Yamaguchi-Kabata[1¤a], Makoto K. Shimada[1,2¤b], Yosuke Hayakawa[1,2], Shinsei Minoshima[3], Ranajit Chakraborty[4], Takashi Gojobori[1,5], Tadashi Imanishi[1]*

1 Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **2** Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, **3** Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan, **4** Center for Genome Information, University of Cincinnati, Cincinnati, Ohio, United States of America, **5** Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan

Abstract

Background: A great amount of data has been accumulated on genetic variations in the human genome, but we still do not know much about how the genetic variations affect gene function. In particular, little is known about the distribution of nonsense polymorphisms in human genes despite their drastic effects on gene products.

Methodology/Principal Findings: To detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in a database for single nucleotide polymorphisms (dbSNP build 125) located in the exons of 36,712 known and predicted protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8). We found a total of 252,555 single nucleotide polymorphisms (SNPs) and 8,479 insertion and deletions in the representative transcripts in these genes. The SNPs located in ORFs include 40,484 synonymous and 53,754 nonsynonymous SNPs, and 1,258 SNPs that were predicted to be nonsense SNPs or read-through SNPs. We estimated the density of nonsense SNPs to be 0.85×10^{-3} per site, which is lower than that of nonsynonymous SNPs (2.1×10^{-3} per site). On average, nonsense SNPs were located 250 codons upstream of the original termination codon, with the substitution occurring most frequently at the first codon position. Of the nonsense SNPs, 581 were predicted to cause nonsense-mediated decay (NMD) of transcripts that would prevent translation. We found that nonsense SNPs causing NMD were more common in genes involving kinase activity and transport. The remaining 602 nonsense SNPs are predicted to produce truncated polypeptides, with an average truncation of 75 amino acids. In addition, 110 read-through SNPs at termination codons were detected.

Conclusion/Significance: Our comprehensive exploration of nonsense polymorphisms showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, suggesting that nonsense mutations have more severe effects than amino acid changes. The correspondence of nonsense SNPs to known pathological variants suggests that phenotypic effects of nonsense SNPs have been reported for only a small fraction of nonsense SNPs, and that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations. These nonsense SNPs may include pathological variants that have not yet been reported. These data are available from Transcript View of H-InvDB and VarySysDB (http://h-invitational.jp/varygene/).

Citation: Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, et al. (2008) Distribution and Effects of Nonsense Polymorphisms in Human Genes. PLoS ONE 3(10): e3393. doi:10.1371/journal.pone.0003393

Editor: Alan Christoffels, University of Western Cape, South Africa

Received March 6, 2008; **Accepted** September 3, 2008; **Published** October 14, 2008

Funding: This research is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: t.imanishi@aist.go.jp

¤a Current address: Center for Genomic Medicine, RIKEN, Yokohama City, Kanagawa, Japan
¤b Current address: Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi, Japan

Introduction

Genetic variations in the human genome are maintained by a balance of mutation, selection and random genetic drift. Some of the polymorphisms cause phenotypic variations and diseases. Therefore, many studies have attempted to identify causative variants of genetic diseases and the relationships between genetic variations and phenotypic effects. Genetic variations within linked loci are inherited to the same gamete. Based on the linkage of genetic variations, loci that contain disease-causing genes have been mapped by using polymorphic markers. At present, about 14 million clusters of genetic polymorphisms have been identified in the human genome [1]. On average, two haploid genomes are estimated to differ by one single nucleotide polymorphism (SNP) in every 1200–1500 bp [2]. SNPs have been recently used to conduct genome-wide association studies to find genomic regions that are susceptible to diseases and phenotypic variations [3,4,5,6]. In this approach, usually, causative polymorphisms for diseases or phenotypic variations are identified after the identification of susceptible genomic regions by using SNP markers. Such SNPs are called landmark SNPs, and the indirect relationships between polymorphisms and phenotypic variations were examined to identify genomic regions where causative genes are located.

Another approach in finding pathological variants is to extract polymorphisms that alter amino acids in functional genes or affect gene expression or splicing, using a comprehensive set of functional elements of the human genome. Several studies have analyzed nonsynonymous SNPs to predict pathological variants [7,8,9, 10,11,12,13,14]. A large number of nonsynonymous SNPs also have been examined for associations with diseases[15,16].

Although many pathological mutations have been identified [17,18], the number of such variants is small compared to the number of known polymorphisms, and it is still unclear which polymorphisms have biological effects. In a study of consanguineous marriage [19], it was estimated that each person has deleterious alleles that are equivalent to a few lethal genes. Gene-centric SNP surveys have shown that the ratio of nonsynonymous to synonymous SNPs is significantly higher in the low frequency class than in the common frequency class [20,21,22]. These results suggest that a large fraction of the low frequency nonsynonymous SNPs are deleterious. To understand the molecular basis of the effects of human genetic variations on phenotypic variations, a prediction analysis of possible effects of polymorphisms on gene function in all human genes appears to be needed.

In this study, to detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in the Single Nucleotide Polymorphism Database (dbSNP) (build 125) in the exons of all 36,712 protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8)[23,24]. In summary with representative transcripts (one transcript from one gene), we detected 53,754 nonsynonymous SNPs and 1,417 SNPs causing changes between amino acids and stop codons. Among possible point mutations in ORFs, nonsense mutations cause the most drastic changes of gene products. In fact, several reports have shown that nonsense mutations cause genetic diseases [25,26,27,28]. Truncation of a polypeptide by a premature termination codon causes a drastic change in the gene product. Furthermore, it is known that a nonsense mutation can cause decay of mRNA resulting in the absence of the gene product. This process, called 'nonsense-mediated decay (NMD)' limits the synthesis of abnormal proteins[29,30,31]. On the other hand, the loss of a termination codon in a transcript also appears to cause decay of mRNA (referred to as non-stop decay) and thus to prevent translation[32,33]. In spite of the severe effects of nonsense mutations, the distribution of nonsense SNPs in human genes is little understood. In this study, we examined the density of nonsense SNPs in human genes, and showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, possibly due to the more severe effects of premature stop codons than amino acid changes. About a half of nonsense SNPs are predicted to cause NMD. The correspondence between known pathological variants and nonsense SNPs suggests that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations.

Results

Selection and classification of polymorphisms in exon regions

We analyzed 9,235,997 polymorphisms (dbSNP build 125) in the human genome with exon positions and predicted ORFs that were revealed in our annotation project of human genes (H-InvDB) (Figure 1). In all of the 36,712 protein-coding loci in the genome, we detected 252,555 SNPs and 8,479 insertions and deletions (indels) that exist in exon regions of the representative transcript (one transcript from one gene) (Table 1). The polymorphisms in the exon regions were further classified according to the predicted ORFs. We detected 96,164 SNPs within the ORFs, 51,881 SNPs in the 5′UTR regions and 104,510 SNPs in the 3′UTR regions. Among the SNPs in the ORFs, 40,484 were synonymous and 53,754 were nonsynonymous (Further analyses of nonsynonymous SNPs are described in Results S1.). Most of the indels were detected in the UTR regions. The ORF regions contained 1,258 SNPs that cause changes between amino acids and stop codons (Table S1). Of the 1,258 SNPs, 1,183 SNPs were regarded as nonsense SNPs, while 75 were found to have stop codons as ancestral alleles. We also detected 247 SNPs at termination codon sites, 88 of which were synonymous. The remaining 159 SNPs were changes between stop codons and amino acids. After checking ancestral alleles, 110 of the 159 SNPs were inferred to be read-through SNPs, while the other 49 were inferred to changes to stop codons.

Distribution of polymorphisms in exon regions

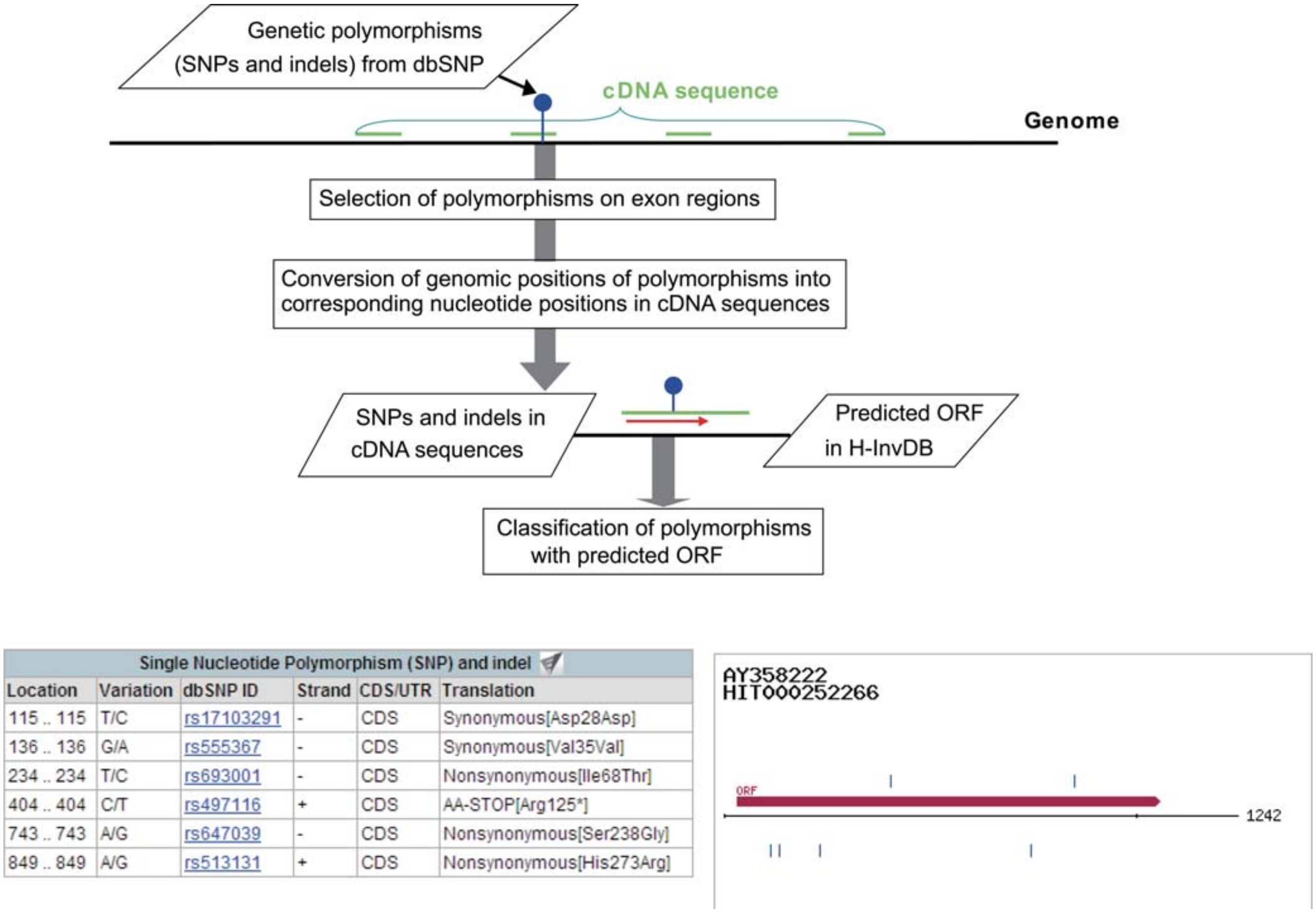
Densities of polymorphisms were estimated for 23,717 genes whose functions are clearly defined or suggested (similarity category I–III, see Materials and Methods) and genes annotated as conserved hypothetical proteins (similarity category IV). To estimate the densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in the ORFs, we calculated the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in the coding regions. The fractions of sites (%) in the coding regions for synonymous, nonsynonymous, and nonsense mutations were estimated to be 28.5%, 68.1%, and 3.4%, respectively. Of the three types of SNPs, synonymous SNPs had the highest density, 4.1×10^{-3} per synonymous site, in ORFs (Table 2). The estimated density of nonsynonymous SNP was 2.1×10^{-3} per site (Table 2). The lower density of nonsynonymous SNPs compared with synonymous SNPs (51%) is due to the functional constraint of amino acid changes, and is in agreement with previous studies [20,22,34]. However, the ratio of the numbers of nonsynonymous SNPs to synonymous SNPs per site is higher in this study compared with previous studies (32–34%) [20,21,22], which they focused on specific populations. The higher ratio of nonsynonymous SNPs in this study may be due to the fact that our study is based on pooled data from various populations world wide. This study includes many nonsynonymous SNPs that exist in relatively lower frequencies and are likely to be more population-specific in comparison to synonymous SNPs [20].

Among random nucleotide mutations in ORFs, 3.4% would be expected to be nonsense mutations; however, the distribution of nonsense SNPs has not been evaluated or reported. The density of nonsense SNPs was estimated to be 0.85×10^{-3} per site (Table 3), which is only 21% of the density of synonymous SNPs, and 40% of the density of nonsynonymous SNPs. The reason for the lowest density of nonsense SNPs may be that premature stop codons have more severe effects than amino acid changes.

In the exons of the 36,712 loci, 8479 indels were detected, and 1,532 of them were found in ORFs. Among the latter, 1,331 are expected to cause frame shifts, resulting in drastic changes of proteins. The density of indels in ORFs was much lower than in the UTR regions (Table 4). The lower density of indels in the 5′UTRs than in the 3′UTRs suggests that functional constraint for insertions and deletions is higher in the 5′UTR regions than in the 3′UTR regions.

Nonsense SNPs

We examined the patterns and the positions of the nonsense SNPs. There are 23 possible ways to change codons into stop


Genetic polymorphisms
(SNPs and indels) from dbSNP
cDNA sequence
Genome
Selection of polymorphisms on exon regions
Conversion of genomic positions of polymorphisms into corresponding nucleotide positions in cDNA sequences
SNPs and indels in cDNA sequences
Predicted ORF in H-InvDB
Classification of polymorphisms with predicted ORF
AY358222
HIT000252266
ORF
1242

Single Nucleotide Polymorphism (SNP) and indel

Location	Variation	dbSNP ID	Strand	CDS/UTR	Translation
115 .. 115	T/C	rs17103291	-	CDS	Synonymous[Asp28Asp]
136 .. 136	G/A	rs555367	-	CDS	Synonymous[Val35Val]
234 .. 234	T/C	rs693001	-	CDS	Nonsynonymous[Ile68Thr]
404 .. 404	C/T	rs497116	+	CDS	AA-STOP[Arg125*]
743 .. 743	A/G	rs647039	-	CDS	Nonsynonymous[Ser238Gly]
849 .. 849	A/G	rs513131	+	CDS	Nonsynonymous[His273Arg]

Figure 1. Analysis of polymorphisms with gene structure. Top: Scheme of analysis pipeline of polymorphisms with gene structure. Bottom: Screen shots taken from 'Transcript View' in H-InvDB that show classified SNPs and their positions (blue bars) in the *CASP12* gene.
doi:10.1371/journal.pone.0003393.g001

codons (nine, seven and seven for the first, second and third positions, respectively), and all 23 were found (Table 5). Nonsense SNPs were more frequent at the first codon position than at the second and third positions ($p<0.005$, chi-square test). The most frequent type of nonsense mutation is the change from CGA to TGA (Table 5), which is a transitional change at CpG mutation hotspots [35]. However, it is notable that there were frequent transversional mutations such as GAA to TAA and GAG to TAG. Our analyses of nonsense polymorphisms revealed that changes between hydrophilic amino acids and termination codons by nucleotide changes at the first codon positions were very frequent.

We examined the positions of 1,183 nonsense polymorphisms in the coding regions. On average, nonsense SNPs were located at 250 codons upstream of the original termination codons. To predict whether a nonsense mutation causes nonsense-mediated decay (NMD) of mRNA, we examined the locations of nonsense SNPs in the exon-intron structure of the genes (Table 6). As a result, of the 1183 nonsense SNPs, 581 were predicted to cause NMD, and thus to prevent translation. The other 602 cases of nonsense SNPs were predicted to result in truncated proteins. For the cases that truncated proteins are produced, the average truncation was estimated to be 75 amino acids.

Table 1. SNPs and indels in exon, intron and other genomic regions.

	Exon	Intron	Other genomic regions
SNPs	249,182	3,332,537	5,209,127
Indels	9,742	185,761	249,648

Polymorphisms mapped on single positions were analyzed with 36,712 protein-coding genes.
doi:10.1371/journal.pone.0003393.t001

To see which of these nonsense SNPs were known pathological mutations, we compared them with allelic variants in the Online Mendelian Inheritance in Man (OMIM) database. Only eight of 1,183 nonsense SNPs (rs17602729 in *AMPD1*, rs283413 in *ADH1C*, rs10250779 in *PGAM2*, rs17215500 in *KCNQ1*, rs497116 in *CASP12*, rs2228325 in *ACTN3*, rs3092891 in *RB1* and rs28989186 in *BUB1B*) matched the variants in the OMIM database that are known variants with phenotypic variations (Table 7). This low value suggests that the biological effects of most nonsense SNPs have not yet been reported. Interestingly, each of the eight cases that matched known pathological variants was predicted to cause NMD (Table 7).

SNPs that cause read-though of the original termination codon

Among the 247 SNPs at termination codon sites, 119 SNP-mRNA pairs were found to be read-through mutations. If the allele having the stop codon is the ancestral type, the SNP is

Table 2. Classified SNPs in exon regions.

Region	Effects on translation	Genes in category I–IV[a]	All protein-coding genes[b]
5′UTR		23454 [3.3×10^{-3}/site][c]	51881
ORF	Total	85233 [2.7×10^{-3}/site]	96164
	Synonymous	37484 [4.1×10^{-3} /site]	40484
	Nonsynonymous	46261 [2.1×10^{-3} /site]	53754
	AA↔Ter[d]	938	1258
	Unclassified[e]	398	421
Stop codon	Total	152	247
	Synonymous	63	88
	Ter↔AA[d]	89	159
3′UTR		69691 [3.3×10^{-3}/site]	104510
Total		178378	252555

[a] Representative transcripts in 23,717 genes whose function were defined or suggested (similarity category I–III) and genes annotated as conserved hypothetical proteins (similarity category IV).
[b] Representative transcripts in all protein-coding genes (36,712) including genes in similarity category I–IV plus similarity category V–VII (hypothetical protein, hypothetical short protein, and pseudogene candidate, respectively).
[c] Densities of polymorphisms are shown in brackets as average number of polymorphisms per site. The average lengths of the 5′UTR, ORF and 3′UTR regions in 23717 genes were 303.9 bp, 1343.5 bp, and 877.6 bp, respectively. The densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in coding regions. The density of nonsense SNPs is shown in Table 3.
[d] SNPs causing changes between amino acids and stop codons.

doi:10.1371/journal.pone.0003393.t002

Table 3. SNPs causing changes between amino acids and stop codons.

Region	Effects on translation	Genes in category I–IV[a]	All protein-coding genes[a]
ORF	Nonsense	910 [0.85×10^{-3}/site][d]	1183
	Read-through[b]	28	75
Stop codon	Read-through	67	110
	Nonsense[c]	22	49

[a] These two gene sets are the same as Table 2.
[b] Possible read-through SNPs in which alleles coding stop codons were ancestral type. This may be due to existence of shorter ORFs in the ancestral population.
[c] Possible nonsense SNPs in which alleles coding stop codons were derived alleles. This may be due to existence of longer ORFs in the ancestral population.
[d] The densities of nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for nonsense mutations in coding regions.

doi:10.1371/journal.pone.0003393.t003

Table 4. Insertions and deletions in exon regions.

	Genes in category I–IV[a]	All protein-coding genes[a]
5′UTR	785 [0.11×10^{-3}][b]	2005
ORF	1120 [0.035×10^{-3}]	1532
3′UTR	3323 [0.16×10^{-3}]	4942
Total	5225[c]	8479

[a] These two gene sets are the same as Table 2.
[b] Densities of polymorphisms are shown in brackets as average number of polymorphisms per site.
[c] Three indels were located on both of ORF and UTR.

doi:10.1371/journal.pone.0003393.t004

Table 5. Frequency of each type of codon change for nonsense SNPs.

	TAA		TAG		TGA		Total
	Aaa→Taa	33	Aag→Tag	31	Aga→Tga	20	
1st	**Caa→Taa**	**62**	**Cag→Tag**	**162**	**Cga→Tga**	**203**	748*
	Gaa→Taa	80	Gag→Tag	125	Gga→Tga	32	
	tCa→tAa	27	tCg→tAg	19	tCa→tGa	25	
2nd			**tGg→tAg**	**80**			200
	tTa→tAa	18	tTg→tAg	18	tTa→tGa	13	
	taC→taA	25	taC→taG	25	tgC→tgA	22	
3rd					**tgG→tgA**	**85**	235
	taT→taA	19	taT→taG	27	tgT→tgA	32	
Total		264		487		432	1183

Bold letters show nucleotide changes by transition.
*P<0.005 by chi-square test.

doi:10.1371/journal.pone.0003393.t005

regarded as a change causing elongation of the polypeptide. However, an extended polypeptide would be expected only if there is an additional termination codon downstream. For 108 SNP-mRNA pairs, an additional termination codon was found in the 3′UTR region. The average extension was estimated to be 29 amino acids. Interestingly, we found five SNP-mRNA pairs that have no stop codons in the 3′UTR at all (The remaining six SNP-mRNA pairs do not have 3′UTR regions). For example, the T-to-C substitution (rs15941) in the *DDR2* gene (X74764) is predicted to be a read-through mutation (from TAG to CGA), and the transcript has no other stop codon in the 3′UTR region. The frequency of this SNP is unknown (it is monomorphic in the four populations in HapMap project [4]). However, if this polymorphism really exists, transcripts having this read-through mutation would not produce a protein. Another example is the T-to-C substitution (rs17850833) in the *MFSD3* gene (CR620962), which causes a change from TGA to CGA resulting in a change to arginine.

Functional bias of genes having nonsense SNPs

To see whether there is any functional bias in genes having nonsense SNPs, we examined the frequent biological terms in the genes having nonsense SNPs. We classified the genes having nonsense SNPs into two categories: genes with nonsense SNPs that are predicted to cause NMD and genes with nonsense SNPs that are not predicted to cause NMD. For genes having nonsense SNPs that would cause NMD (Table 8), the molecular functions that are most overrepresented included phosphorylation, ATP binding, iron/calcium ion binding, nucleotide/RNA binding and transporter activity. The localization of these genes was also biased to the cell membrane and the proteinaceous extracellular matrix. On the other hand, the genes having nonsense SNPs predicted to not cause NMD showed less bias in biological function (Table 9).

Table 6. Nonsense SNPs and prediction of NMD.

	Predicted to cause NMD[a]	Not for NMD[b]	Total
Known pathological variants	8[c]	0	8
Other nonsense SNPs	573	602	1175
Total	581	602	1183

[a]This prediction is based on that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the 3′ end of the second to last exon. Here, the nonsense SNPs located in the 5′ side of the boundary, which was set at 50 nucleotides upstream from the 3′ end of the second to last exon, were predicted to cause NMD.
[b]This number includes SNPs in genes consisting of only one exon.
[c]P = 0.0033 by Fisher's exact test.
doi:10.1371/journal.pone.0003393.t006

Table 7. Nonsense SNPs with known pathological effects.

Acc#	Chr	Gene symbol	SNP	Variation	OMIM	Biological effects
M60092	1	*AMPD1*	rs17602729	Gln12Ter	102770	AMPD deficiency
M12272	4	*ADH1C*	rs283413	Gly78Ter	103730	Parkinson disease
BC073741	7	*PGAM2*	rs10250779	Trp78Ter	261670	Myopathy
AF000571	11	*KCNQ1*	rs17215500	Arg518Ter	607542	Long QT syndrome 1
AY358222	11	*CASP12*	rs497116	Arg125Ter	608633	Sepsis susceptibility
M86407	11	*ACTN3*	rs2228325	Arg577Ter	102574	Athletic performance
L41870	13	*RB1*	rs3092891	Arg445Ter	180200	Bilateral retinoblastoma
AF068760	15	*BUB1B*	rs28989186	Arg194Ter	602860	Premature chromatid separation trait and mosaic variegated aneuploidy syndrome

doi:10.1371/journal.pone.0003393.t007

Table 8. Functional bias of genes having nonsense SNPs causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0006118	electron transport	15	4.23	3.55	5.03×10^{-5}
	0006468	protein amino acid phosphorylation	16	7.28	2.20	4.98×10^{-3}
Cellular component	0016020	membrane	41	22.55	1.82	5.57×10^{-4}
	0005578	proteinaceous extracellular matrix	8	1.21	6.62	2.17×10^{-6}
Molecular function	0005524	ATP binding	35	17.15	2.04	1.79×10^{-4}
	0004713	protein tyrosine kinase activity	16	6.46	2.48	1.56×10^{-3}
	0004674	protein serine/threonine kinase activity	16	6.78	2.36	2.51×10^{-3}
	0000166	nucleotide binding	14	5.61	2.50	2.79×10^{-3}
	0004672	protein kinase activity	16	7.15	2.24	4.21×10^{-3}
	0003723	RNA binding	10	3.11	3.22	1.82×10^{-3}
	0005506	iron ion binding	8	2.00	4.00	1.32×10^{-3}
	0005509	calcium ion binding	16	7.65	2.09	7.89×10^{-3}
	0005215	transporter activity	10	3.44	2.91	3.76×10^{-3}
	0016491	oxidoreductase activity	11	4.24	2.59	5.76×10^{-3}
	0003779	actin binding	6	1.27	4.74	2.24×10^{-3}
	0004759	carboxylesterase activity	5	0.24	20.44	4.19×10^{-6}

[a]Number of genes with a molecular function in the 581 genes in which nonsense SNPs causing NMD were found.
[b]Expected number of genes that have a biological function in a sample of 581 genes, assuming a proportion of genes with a molecular function in all human genes.
[c]Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.
doi:10.1371/journal.pone.0003393.t008

Table 9. Functional bias of genes having nonsense SNPs not causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0007156	homophilic cell adhesion	6	1.42	4.23	3.05×10^{-3}
	0006310	DNA recombination	3	0.19	15.50	8.25×10^{-4}
	0006414	translational elongation	3	0.34	8.85	4.48×10^{-3}
	0042254	ribosome biogenesis and assembly	2	0.15	13.77	8.68×10^{-3}
Cellular component	0005853	eukaryotic translation elongation factor 1 complex	2	0.13	15.50	6.82×10^{-3}
Molecular function	0004194	pepsin A activity	2	0.18	11.27	1.30×10^{-2}
	0003746	translation elongation factor activity	2	0.29	6.89	3.35×10^{-2}

[a]Number of genes with a molecular function in the 602 genes in which nonsense SNPs causing NMD were found.
[b]Expected number of genes that have a biological function in a sample of 602 genes, assuming a proportion of genes with a molecular function in all human genes.
[c]Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.

doi:10.1371/journal.pone.0003393.t009

Discussion

In this study, we conducted an extensive analysis of human genome polymorphisms with a comprehensive catalogue of human genes, and detected more than 50,000 polymorphisms that affect proteins. The distribution of polymorphisms showed different densities of polymorphisms among the 5′UTR, ORF and 3′UTR. The density of SNPs was lower in ORFs than in the 5′UTR and 3′UTR. The density of synonymous SNPs in the ORFs was higher than the densities of SNPs in the UTR regions. The reduction in density of SNPs in the UTR regions is consistent that there are functional constraints on nucleotide changes in UTRs related to the transcriptional and translational efficiency[22]. The density of nonsynonymous SNPs was much lower than the densities of other types of SNPs, possibly due to that the nucleotide changes with alteration of amino acids changes are under strong negative selection [36]. It was not known how nonsense SNPs are distributed in protein-coding regions. Here we showed that the density of nonsense SNPs is much lower than that of nonsynonymous SNPs. Although the biological effects of nonsense mutations appear to vary widely depending on their positions and the genes, the low density of nonsense SNPs that we found suggests that nonsense mutations have more disadvantageous effects than nonsynonymous mutations.

While nonsense mutations that cause NMD result in 'loss of function', nonsense mutations that do not cause NMD produce truncated proteins which could have the dominant effects. The proportion of predicted nonsense SNPs causing NMD in this study is in agreement with a previous study which showed that dbSNP (build 125) has 1301 nonsense SNPs, about half of which were predicted to result in NMD [37]. In order to understand the biological effects of nonsense SNPs, it is important to know whether they do or do not cause NMD, because premature stop codons in a gene can have distinct disease phenotypes depending on the positions of mutations [27,38].

The molecular functions that were overrepresented in the genes having nonsense SNPs included several molecular functions that were observed in human-specific pseudogenes[39], such as ATP binding, actin binding, calcium ion binding, extracellular matrix, nucleic acid binding and oxidoreductase. This is in accord with that nonsense mutations contribute to 'pseudogenization'. It is interesting that nonsense SNPs causing NMD were frequently found in genes that encode proteins involved in phosphorylation, cell-cell interaction, signal transduction and transport. This may be because changes in the length of polypeptides caused by nonsense mutations are under strong negative selection in the genes involved in signal transduction or transportation because abnormal translation products could cause dominant effects. Therefore, inactivation of translation by nonsense mutations in those genes could have milder effects than changes of the length of polypeptides.

Our results showed a low proportion of matches of nonsense SNPs with known pathological variants in OMIM, suggesting that the effects of most nonsense polymorphisms are unknown or not reported. Furthermore, the correspondence of the nonsense SNPs to the OMIM allelic variants (Table 6, Table 7) suggests that nonsense polymorphisms that are subject to NMD are more likely to be involved in phenotypic variations.

There is a possibility that the nonsense SNPs detected here have pathological effects, in particular, if non-dispensable genes have nonsense mutations. First, a defect in one gene by a nonsense mutation or a frame-shifting indels causing a premature termination codon could be a cause of genetic diseases including complex diseases[40]. Second, there is a possibility that nonsense mutations cause recessive lethal alleles that would not be detected as causative variant of diseases. Probably, focusing on nonsense polymorphisms observed in specific populations would be a good way of selection for finding variants with deleterious effects.

The effect of single nonsense SNPs can be compensated by the products of other genes having similar functions[41] and the other splicing isoforms of the gene [42]. Thus, single nonsense SNPs may not always cause severe phenotypic effects. In fact, some nonsense SNPs with high allele frequencies were found across populations[43]. There is a report of fixation of an inactive form of caspase 12 by a nonsense mutation (rs497116) in non-African populations[43], and this is an example supporting the 'less is more hypothesis'[44]. This example suggests that some of nonsense mutations are not disadvantageous and that the increase of frequency of a nonsense allele could be driven by positive selection.

Elongation of polypeptides by read-through mutations can affect protein folding and aggregation of proteins, which could affect phenotypic variations. Furthermore, a read-through mutation can cause more severe effects on translation when no additional stop codon follows. Such mutations are subject to 'non-stop decay' [32,33], and would result in no gene product. It has been suggested that non-stop decay and NMD serve to remove toxic, aberrant proteins [29]. It is unclear how frequently such mutations prevent mRNA from producing proteins. Therefore, it would be quite useful to be able to predict the effects of various types of genetic changes on mRNA.

Although the present results are based on representative transcripts (one transcript for one gene), the total number of

SNPs causing changes between amino acids and stop codons in all the splicing isoforms was much larger (2,234). These variations, which cause changes in the length of a polypeptide or which determine whether a protein is translated, may include pathological variants that have yet not been reported. Therefore, it is important to examine their presence in human populations.

Materials and Methods

Data of human genetic polymorphisms

As data of genetic polymorphisms of human genome, single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) in dbSNP [1] were used in this study. The whole data of human SNPs and indels were downloaded from dbSNP (build 125). We used all SNPs and indels that were mapped on single position in the genome, except for 'large insertions' in dbSNP.

Data of human genes

The data of human gene structure were obtained from H-InvDB ver3.8 (http://www.h-invitational.jp/), created by the annotation project of human genes (H-Invitational project) [23,45]. Our analysis of all human genes that corresponds to H-InvDB (ver 3.8) predicted 36,712 protein coding loci. All protein-coding genes were annotated and classified based on similarity to known genes as follows; Category I, Identical to known human protein; Category II, Similar to known protein; Category III, IPR domain containing protein; Category IV, Conserved hypothetical protein; Category V, Hypothetical protein; Category VI, Hypothetical short protein; Category VII, pseudogene candidate. We used the following three kinds of data of the gene structure: 1) genomic location of exons to the human genome (build 35), 2) predicted ORF regions in transcripts, and 3) original and curated cDNA sequences.

Analysis

1. Analysis of polymorphism with exons and predicted ORFs. Selection of polymorphisms on exon regions. We selected polymorphisms in exon regions by comparing the genomic positions of polymorphisms and the start and end positions of exons that were obtained from mapping cDNA sequences to the human genome (Figure 1). Polymorphisms in introns were also selected in a same way.

Conversion of genomic position of polymorphism into nucleotide position in cDNA sequence. To analyze polymorphisms with a predicted ORF, nucleotide positions of polymorphisms in the human genome sequences were converted into the nucleotide positions in cDNA sequences. Because there could be gaps in the alignment of cDNA sequence and the human genome sequence, the nucleotide position was converted considering possible gaps in the alignment. When the cDNA sequence was corrected in ORF prediction because of frame-shifting and remaining intron, the nucleotide position of SNP was modified based on addition or deletion of nucleotides. For a quality control of polymorphism data used for classification, we conformed that one of the nucleotides in each pair of SNP alleles was the same nucleotide at the corresponding position in the cDNA sequence.

Classification of polymorphisms with predicted ORF. Polymorphisms within ORF were classified according to their effect on ORF. For SNPs with two alleles, alleles in nucleotide were converted into 'alleles in codon' by adding two other nucleotides in the codon from cDNA sequence. When a cDNA sequence was corrected in the annotation process by removing a remaining intron or by correcting a frameshift error, the corrected cDNA sequence was used. If these alleles in codon do not contain any stop codon, the alleles were classified into synonymous and nonsynonymous. In case a stop codon is included in the alleles in codon, they were classified into 1) premature termination (nonsense) codon, 2) read-through of original stop codon, and 3) synonymous at stop codon site, by assuming that the cDNA sequence has an ancestral allele. Indels were classified based on whether they are located in ORF. The indels within ORF were further classified by whether the insertion or deletion causes frame shifting in translation.

Inference of direction of nonsense and read-through mutations. Ancestral alleles were obtained from dbSNP (build 128) to check direction of mutations for SNPs causing changes between amino acids and stop codons. For nonsense SNPs in protein-coding regions, we checked whether the ancestral allele codes amino acids. In case that the ancestral allele codes stop codon, we do not regard this SNP as nonsense SNP, but is a read-through mutation assuming that there was a variant having a shorter ORF. For read-though SNPs at termination codon site, we checked whether the ancestral allele codes stop codon. In case that the ancestral allele codes amino acids, we regard this SNP not as a read-through mutation, but as a nonsense mutation in a variant having a longer ORF.

Number of sites for synonymous, nonsynonymous and nonsense mutations. To estimate densities of synonymous, nonsynonymous and nonsense SNPs, the numbers of potential synonymous, nonsynonymous and nonsense sites by single nucleotide changes were estimated for the ORF sequences. This is an extension of estimation of the numbers of synonymous and nonsynonymous sites[46]; the number of synonymous sites is calculated as the number of four-fold degenerate sites plus one-third of the number of two-fold degenerate sites. For 61 codons encoding amino acids, the numbers of nucleotide sites that would cause synonymous, nonsynonymous and nonsense mutations by a single nucleotide change were estimated with a model of nucleotide change. Here, the relative occurrence of a transitional mutation versus a transversional mutation (r) was set to be 4.0 (the expected ratio in the numbers of transitional and transversional mutations was 2.0). For example of the TTA codon for leucine, the number of nonsense sites was estimated to be $2.0/(r+2.0)$, because two types of transversional mutations at the second position cause nonsense mutations.

2. Correspondence to known pathological variants. To check whether the polymorphisms that alter proteins are known pathological variants with phenotypic effect, we examined correspondence of SNPs with data of known pathological variants. We used data of 'allelic variant' in the Online Mendelian Inheritance in Man (OMIM) database [18] as information of variants with phenotypic effect. For nonsynonymous and nonsense SNPs, their effects on translation and positions in ORF were compared with the 'list of alleles' in OMIM (e.g. described as "TRP324TER" or "ALA279THR" for the *NGAS* gene).

3. Prediction of nonsense SNPs causing NMD. Some of nonsense mutations cause nonsense-mediated decay (NMD), resulting in prevention of translation. It has been reported that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the end of the second to last exon [30,31]. To predict whether a nonsense SNP causes NMD, we examined whether a nonsense SNP is located in the 3′ side of the boundary, which was set at 50 nucleotides upstream from the end of the second to last exon, in the exon-intron structure. This method is the same as the method in SNP2NMD [37] when 'NMD distance' is 50 nucleotides.

5. Functional bias of genes with nonsense SNPs. For each biological term from Gene Ontology (www.geneontology.org), a proportion of genes with the biological function in the genes having nonsense SNPs was compared with that in all human genes (representative transcripts in all human genes in H-InvDB ver 5.0), and the significance of over representation of a molecular function in the genes having nonsense SNPs was evaluated as the upper probability of the hypergeometric distribution.

Supporting Information

Results S1 Supplementary results and a table for analyses of nonsynonymous SNPs.
Found at: doi:10.1371/journal.pone.0003393.s001 (0.70 MB DOC)

Table S1 Nonsense SNPs and read-through SNPs on representative transcripts.
Found at: doi:10.1371/journal.pone.0003393.s002 (4.24 MB DOC)

Acknowledgments

We thank Drs. Craig Gough, Norikazu Yasuda, Shuhei Mano, Naoki Nagata, Yoshiyuki Suzuki and Hisakazu Iwama for helpful discussion. We also thank Ryuzou Matsumoto, Seigo Hosono, and all the members in the Integrated Database Team of BIRC, AIST for their technical assistance and providing data of gene structure and annotation.

Author Contributions

Conceived and designed the experiments: YYK SM RC TG TI. Analyzed the data: YYK MS YH. Wrote the paper: YYK MS TI.

References

1. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
3. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
4. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
5. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32: 650–654.
6. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
7. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527–532.
8. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12: 436–446.
9. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33: W480–482.
10. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591–597.
11. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176–3178.
12. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
13. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520–522.
14. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21: 2814–2820.
15. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329–1337.
16. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, et al. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 39: 207–211.
17. Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, et al. (2001) The KMDB/MutationView: a mutation database for human disease genes. Nucleic Acids Res 29: 327–328.
18. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30: 52–55.
19. Morton NE, Crow JF, Muller HJ (1956) An Estimate of the Mutational Damage in Man from Data on Consanguineous Marriages. Proc Natl Acad Sci U S A 42: 855–863.
20. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238.
21. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.
22. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, et al. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22: 239–247.
23. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162.
24. Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, et al. (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36: D793–799.
25. Chang JC, Kan YW (1979) beta 0 thalassemia, a nonsense mutation in man. Proc Natl Acad Sci U S A 76: 2886–2889.
26. Rosenfeld PJ, Cowley GS, McGee TL, Sandberg MA, Berson EL, et al. (1992) A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet 1: 209–213.
27. Inoue K, Khajavi M, Ohyama T, Hirabayashi S, Wilson J, et al. (2004) Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. Nat Genet 36: 361–369.
28. Mimori A, Hidaka Y, Wu VC, Tarle SA, Kamatani N, et al. (1991) A mutant allele common to the type I adenine phosphoribosyltransferase deficiency in Japanese subjects. Am J Hum Genet 48: 103–107.
29. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE (2004) Nonsense-mediated decay approaches the clinic. Nat Genet 36: 801–808.
30. Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, et al. (1998) Binary specification of nonsense codons by splicing and cytoplasmic translation. Embo J 17: 3484–3494.
31. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE (1998) At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. Mol Cell Biol 18: 5272–5283.
32. Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R, et al. (2002) An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science 295: 2258–2261.
33. van Hoof A, Frischmeyer PA, Dietz HC, Parker R (2002) Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science 295: 2262–2264.
34. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, et al. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100: 15754–15757.
35. Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. Science 212: 1350–1357.
36. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
37. Han A, Kim WY, Park SM (2007) SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23: 397–399.
38. Thein SL, Hesketh C, Taylor P, Temperley IJ, Hutchinson RM, et al. (1990) Molecular basis for dominantly inherited inclusion body beta-thalassemia. Proc Natl Acad Sci U S A 87: 3924–3928.
39. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4: e52.
40. Senee V, Chelala C, Duchatelet S, Feng D, Blanc H, et al. (2006) Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. Nat Genet 38: 682–687.
41. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
42. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, et al. (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Res 34: 3917–3928.
43. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, et al. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78: 659–670.
44. Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64: 18–23.
45. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, et al. (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). Gene 364: 99–107.
46. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.