Human SNPs resulting in premature stop codons and protein truncation

Sevtap Savas,[1,2,3] *Sukru Tuzmen*[4] *and Hilmi Ozcelik*[1,2,3]*

[1] Fred A. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, M5G 1X5, Canada
[2] Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, M5G IX5, Canada
[3] Department of Laboratory Medicine and Pathobiology, University of Toronto, 100 College Street, Toronto, ON, M5G IL5, Canada
[4] Cancer Drug Development Laboratory, Translational Genomics Research Institure, 13208 East Shea Blvd, Suite 110, Scottsdale, AZ 85259, USA
* *Correspondence to*: Tel: + 1 416 586 4996; Fax: +1 416 586 8869; E-mail: ozcelik@mshri.on.ca

Date received (in revised form): 10th November 2005

Abstract

Single nucleotide polymorphisms (SNPs) constitute the most common type of genetic variation in humans. SNPs introducing premature termination codons (PTCs), herein called X-SNPs, can alter the stability and function of transcripts and proteins and thus are considered to be biologically important. Initial studies suggested a strong selection against such variations/mutations. In this study, we undertook a genome-wide systematic screening to identify human X-SNPs using the dbSNP database. Our results demonstrated the presence of 28 X-SNPs from 28 genes with known minor allele frequencies. Eight X-SNPs (28.6 per cent) were predicted to cause transcript degradation by nonsense-mediated mRNA decay. Seventeen X-SNPs (60.7 per cent) resulted in moderate to severe truncation at the C-terminus of the proteins (deletion of >50 per cent of the amino acids). The majority of the X-SNPs (78.6 per cent) represent commonly occurring SNPs, by contrast with the rarely occurring disease-causing PTC mutations. Interestingly, X-SNPs displayed a non-uniform distribution across human populations: eight X-SNPs were reported to be prevalent across three different human populations, whereas six X-SNPs were found exclusively in one or two population(s). In conclusion, we have systematically investigated human SNPs introducing PTCs with respect to their possible biological consequences, distributions across different human populations and evolutionary aspects. We believe that the SNPs reported here are likely to affect gene/protein function, although their biological and evolutionary roles need to be further investigated.

Keywords: *SNP, premature termination codons, nonsense-mediated mRNA decay, population distribution, evolutionary selection*

Introduction

The Human Genome Project revealed the presence of a large number of genetic variations among individuals. Single nucleotide polymorphisms (SNPs) are the most common genetic variation; they occur, on average, once in every 400–1,000 base pairs along DNA.[1–4] The term 'polymorphism' traditionally refers to commonly occurring genetic variations (minor allele frequency approximately ≥ 1 per cent) in the population.[5] The density of SNPs varies among different genomic regions, and is thought to be dependent on both the mutation rate and the selective constraints on the region.[6] Currently, there is a strong interest in SNPs because they are hypothesised to contribute to differential disease risk and drug/treatment response among individuals.[7,8]

SNPs located in the coding regions of genes may have important biological consequences. For example, non-synonymous SNPs (nsSNPs) change the amino acid sequence and thus may affect protein function. Although many approaches and systematic analyses have been undertaken to identify nsSNPs with possible biological significance,[9–12] to our knowledge no large-scale systematic analysis has been carried out to identify and characterise SNPs that introduce premature termination codons (PTCs; herein called X-SNPs). Both frameshift and nonsense mutations can lead to the introduction of PTCs along the open reading frames. As a result of PTCs, the stability of transcripts or proteins may be directly affected.[13,14] Alternatively, the truncated proteins may act in a dominant-negative fashion.[15] Thus, the PTCs can lead to either loss-of-function or gain-of-function by altering the stability and function of the transcripts/proteins.

The Mendelian human diseases are associated with high-penetrant disease-causing genetic alterations that are found in very low frequencies (approximately <1 per cent) in the population, most likely due to strong selection against them.[16] In inherited human genetic disorders, approximately one-third

of mutations introduce PTCs[17] that are considered to be deleterious. Similarly, the number of SNPs introducing PTCs in the human genome is estimated to be fairly low, and a previous study suggested the presence of strong evolutionary selection against X-SNPs.[18] Therefore, disease-related or not, the PTCs are considered dramatically to affect proteins leading to potential biological abnormalities. In this study, our aim was to evaluate the polymorphisms introducing PTCs in the human genome with respect to their potential biological consequences, distributions across different human populations and minor allele frequencies. As more X-SNPs are discovered and deposited in public SNP databases, it will be possible to analyse a larger number of X-SNPs and obtain more comprehensive data. Nevertheless, our results do provide an interesting and unique catalogue of polymorphisms that deserves further biological and epidemiological disease-association studies.

Methods

SNPs

SNPs annotated 'premature termination codon SNPs' were retrieved from the dbSNP database build 120 (http://www.ncbi.nlm.nih.gov/SNP/).[19] We have annotated such SNPs as X-SNPs throughout this paper. There was a total of 977 X-SNPs in the dbSNP database; however, only 119 of them were presented with minor allele frequency information. Among these SNPs, only the ones that were found in at least two chromosomes with a sample size of ≥ 20 chromosomes were further analysed (herein annotated as validated X-SNPs). The X-SNPs that are located on the transcripts annotated as 'predictions', 'pseudogenes', 'similar to' or 'open reading frames' were excluded from this study. In total, 28 X-SNPs were in agreement with all of the above requirements.

BLAST analyses

To map the SNP sequences on transcripts, SNP-flanking sequences of X-SNPs were blasted against the transcripts in GenBank (http://www.ncbi.nih.gov/Genbank/)[20] using the BLAST against gene transcripts tool (http://lpgws.nci.nih.gov:80/perl/blast2/),[21] as explained by Savas *et al.*[22] One mismatch in the SNP-flanking sequence/transcript alignment was allowed. The SNP-flanking sequences were also blasted against the human genome using the NCBI BLAST tool (http://www.ncbi.nlm.nih.gov/BLAST/)[23] to ensure that the SNP sequences are not derived from multiple genomic regions,[24] as explained in a further paper by Savas *et al.*[25]

Alternatively spliced transcript variants (ASTVs)

Information relating to ASTVs was retrieved from the RefSeq resource of NCBI (http://www.ncbi.nlm.nih.gov/RefSeq/).[26]

Candidate transcripts for nonsense-mediated mRNA decay

Blasting the transcript sequence against the human genome identified the genomic structures of transcripts. The subsequent manual analysis of the exon–intron boundaries identified X-SNPs that can lead to nonsense-mediated mRNA decay (NMD): the transcripts with an SNP introducing a PTC located ≥ 50 nucleotides upstream of an exon–intron junction are considered candidates to undergo NMD.[13,27–29]

Results and discussion

Possible biological consequences of X-SNPs

Our systematic search of the dbSNP database[19] (build 120) yielded 28 validated X-SNPs from 28 genes (Table 1). Twenty-three genes bearing X-SNPs were found to code for a single transcript; however, the remaining five X-SNPs were found in genes undergoing alternative splicing: *DSCR8-K79X*, *HPS4-R246X*, *IL17RB-Q484X*, *OAS2-W720X* and *TAP2-Q687X*. With the exception of *HPS4-R246X*, all X-SNPs were mapped onto an ASTV coding for the longest protein isoform. For 22/28 X-SNPs, genotype information was available in the dbSNP database. As a result, for 12 X-SNPs, at least one homozygous sample was reported, suggesting that these X-SNPs do not affect the fitness *per se* (see below; Table 1). In the remaining cases, genotyping of larger sample sets may help in elucidating whether the homozygous state is deleterious (ie the homozygotes are not viable) or whether the low allele frequency makes it hard to detect the homozygotes in small populations.

We then carried out a theoretical evaluation of the possible biological consequences of the identified X-SNPs at the mRNA and protein levels. For example, NMD is a surveillance system that specifically eliminates transcripts that contain PTCs as a result of mutations in DNA or errors in RNA processing.[15] NMD usually requires a downstream intron and at least 50–55 nucleotides before the downstream exon–intron junction in order for a PTC to be recognised.[27,28] Based on the 50–55 nucleotide rule, we analysed the locations of the X-SNPs with respect to the exon–intron boundaries and predicted that eight (28.6 per cent) X-SNPs (*AGT-Q53X*, *APOC4-W47X*, *EPHX1-W97X*, *MS4A12-Q71X*, *POLE2-K443X*, *SERPINB11-E90X*, *SMUG1-Q3X* and *ZNF34-Q56X*) may potentially cause mRNA degradation via NMD. Thus, at least these eight X-SNPs are likely to result in loss of gene function. Exceptions to this rule have also been reported,[17] however, which suggests that the proportion of PTC-containing mRNAs undergoing mRNA degradation may, in fact, be larger. The reported allele frequencies of these X-SNPs ranged from rare (*AGT-Q53X* 0–5 per cent; *APOC4-W47X* 0–5 per cent; *EPHX1-W97X* 0–2.2 per cent; *POLE2-K443X* 0–1.2 per cent; *SMUG1-Q3X* 1.2–1.4 per cent) to common (*MS4A12-Q71X* 41.5–45.8 per cent;

Table 1. Validated X-SNPs in the human genome.

Gene	[a]Gene function	[b]Accession #	Location	[c]SNP ID	[d]Frequency	[e]Homo-zygosity	X-SNP	[f]Protein length (trun-cation)	[g]NMD	[h]CpG
AGT	Cell signalling; hypertension	NM_000029.1	1q42-q43	rs5039	CEPH-MULTI-NATIONAL 184 chr. G=1.000 A=0.000 HYP1-MULTI-NATIONAL 80 chr. G=0.950 A=0.050	n/a	Q53X	485 (89%)	+	+
APOC4	Lipid metabolism	NM_001646.1	19q13.2	rs5164	CEPH-MULTI-NATIONAL 184 chr. G=1.000 A=0.000 HYP1-MULTI-NATIONAL 80 chr. G=0.950 A=0.050	–	W47X	127 (63%)	+	–
CDH15	Cell adhesion; morphogenetic processes	NM_004933.2	16q24.3	rs2270416	JBIC-allele-EAST ASIA 1500 chr. G=0.826 T=0.174	n/a	Y788X	814 (3.2%)	–	+
CLCA3	Transport	NM_004921.1	1p31-p22	rs2292830	JBIC-allele -EAST ASIA 1462 chr. G=0.569 C=0.431	n/a	Y84X	262 (67.3%)	–	–
CYP2C19	Transport; drug metabolism and synthesis of lipids	NM_000769.1	10q24. 1-q24.3	rs4986893	PAC1-EAST ASIA 46 chr. G=0.913 A=0.087 CAUC1-MULTI-NATIONAL 60 chr. G=1.000 A=0.000 AFR1-MULTI-NATIONAL 48 chr. G=0.979 A=0.021 HISP1-CENTRAL/SOUTH AMERICA 44 chr. G=1.000 A=0.000 P1-MULTI-NATIONAL 198 chr. G=0.975 A=0.025	n/a	W212X	446 (52.5%)	–	–
DSCR8	Unknown	NM_032589.2	21q22.2	rs2836172	NCBI|NIHPDR-NORTH AMERICA 20 chr. A=0.900 T=0.100 AFD_EUR_PANEL-NORTH AMERICA 48 chr. A=1.000 AFD_AFR_PANEL-NORTH AMERICA 46 chr. A=0.783 T=0.217 AFD_CHN_PANEL-NORTH AMERICA 48 chr. A=0.854 T=0.146	+	K79X	91 (13.2%)	–	–

EPHX1	Aromatic compound catabolism; xenobiotic metabolism	NM_000120.2	1q42.1	rs4986931	PAC1-EAST ASIA 46 chr. A=0.978 G=0.022 P1-MULTI-NATIONAL 202 chr. A=0.990 G=0.010 CAUC1-MULTI-NATIONAL 62 chr. A=1.000 G=0.000 AFR1-MULTI-NATIONAL 48 chr. A=1.000 G=0.000 HISP1-CENTRAL/SOUTH AMERICA 46 chr. A=0.978 G=0.022	–	W97X	455 (78.7%)	+	–
FUT2	Carbohydrate metabolism; protein glycosylation	NM_000511.1	19q13.3	rs1800030	PAC1-EAST ASIA 48 chr. G=0.979 A=0.021 P1-MULTI-NATIONAL 202 chr. G=0.995 A=0.005 CAUC1-MULTI-NATIONAL 60 chr. G=1.000 A=0.000 AFR1-MULTI-NATIONAL 48 chr. G=1.000 A=0.000 HISP1-CENTRAL/SOUTH AMERICA 46 chr. G=1.000 A=0.000	–	W297X	346 (14.2%)	–	–
HPS4	Organelle biogenesis; protein stabilisation/ targeting	NM_152843.1	22cen--q12.3	rs3747129	JBIC-allele-EAST ASIA 1492 chr. G=0.798 A=0.202 AFD_EUR_PANEL-NORTH AMERICA 48 chr. G=0.812 A=0.188 AFD_AFR_PANEL-NORTH AMERICA 46 chr. G=0.978 A=0.022 AFD_CHN_PANEL-NORTH AMERICA 48 chr. G=0.750 A=0.250 HapMap-CEU-EUROPE 120 chr. G=0.825 A=0.175	+	R246X	528 (53.4%)	–	+
IL17RB	Immuno-regulatory activity; regulation of cell growth	NM_018725.2	3p21.1	rs1043261	JBIC-allele-EAST ASIA 1476 chr. C=0.902 T=0.098 HapMap-CEU-EUROPE 120 chr. C=0.908 T=0.092 AFD_EUR_PANEL-NORTH AMERICA 48 chr. C=0.938 T=0.062 AFD_AFR_PANEL-NORTH AMERICA 46 chr. C=0.978 T=0.022 AFD_CHN_PANEL-NORTH AMERICA 48 chr. C=0.792 T=0.208	+	Q484X	502 (3.6%)	–	+

(*continued*)

Table 1. *Continued.*

Gene	[a]Gene function	[b]Accession #	Location	[c]SNP ID	[d]Frequency	[e]Homozygosity	X-SNP	[f]Protein length (truncation)	[g]NMD	[h]CpG
KRTAP1-1	Cytoskeleton; intermediate filaments	NM_030967.2	17q12-q21	rs3213755	JBIC-allele-EAST ASIA 708 chr. C=0.617 T=0.383 HapMap-CEU-EUROPE 120 chr. G=0.800 A=0.200	+	Q51X	177 (71.2%)	–	–
LCE5A	Unknown	NM_178438.1	1q21.3	rs2282298	JBIC-allele-EAST ASIA 1504 chr. G=0.979 A=0.021 AFD_EUR_PANEL-NORTH AMERICA 48 chr. C=1.000 AFD_AFR_PANEL-NORTH AMERICA 46 chr. C=1.000 AFD_CHN_PANEL-NORTH AMERICA 48 chr. C=0.896 T=0.104	–	R79X	118 (33.1%)	–	+
LIG4	DNA repair; cell cycle	NM_206937.1	13q33-q34	rs2232636	PAC1-EAST ASIA 46 chr. G=1.000 A=0.000 P1-MULTI-NATIONAL 202 chr. G=0.995 A=0.005 CAUC1-MULTI-NATIONAL 62 chr. G=1.000 A=0.000 AFR1-MULTI-NATIONAL 48 chr. G=0.979 A=0.021 HISP1-CENTRAL/SOUTH AMERICA 46 chr. G=1.000 A=0.000	–	W46X	911 (95%)	–	–
LPL	Lipoprotein metabolism	NM_000237.1	8p22	rs328	WIAF-CSNP-MITOGPOP5-MULTI-NATIONAL 112 chr. C=0.982 G=0.018 JBIC-allele-EAST ASIA 1458 chr. C=0.860 G=0.140 CEPH-MULTI-NATIONAL 184 chr. C=0.640G=0.360 AFD_EUR_PANEL-NORTH AMERICA 44 chr. C=0.727 G=0.273 AFD_AFR_PANEL-NORTH AMERICA 42 chr. C=0.952 G=0.048 AFD_CHN_PANEL-NORTH AMERICA 46 chr. C=0.935 G=0.065	+	S474X	475 (0.2%)	–	–

MAGEE2	Unknown	NM_138703.2	Xq13.3	rs1343879	TSC_42_C-NORTH AMERICA 84 chr. C=0.950 A=0.050 C_42_A-EAST ASIA 84 chr. A=0.650 C=0.350 TSC_42_AA-NORTH AMERICA 84 chr. C=0.950 A=0.050 HapMap-CEU-EUROPE 120 chr. C=0.983 A=0.017	–	E120X	523 (77.1%)	–	–
MS4A12	Signal transduction	NM_017716.1	11q12	rs2298553	JBIC-allele-EAST ASIA 726 chr. C=0.585 T=0.415 AFD_EUR_PANEL-NORTH AMERICA 48 chr. C=0.583 T=0.417 AFD_AFR_PANEL-NORTH AMERICA 42 chr. C=0.548 T=0.452 AFD_CHN_PANEL-NORTH AMERICA 48 chr. C=0.542 T=0.458	+	Q71X	267 (73.4%)	+	–
OAS2	Immune response	NM_016817.1	12q24.2	rs15895	POOLED_CEPH-MULTI-NATIONAL 188 chr. A=0.668 G=0.332 CEPH-MULTI-NATIONAL 184 chr. C=0.670 T=0.330 SC_12_A-EAST ASIA 20 chr. G=1.000 SC_12_AA-NORTH AMERICA 24 chr. G=0.830 A=0.170 SC_12_C-NORTH AMERICA 24 chr. G=0.710 A=0.290 SC_95_C-NORTH AMERICA 184 chr. C=0.590 T=0.410 AFD_EUR_PANEL-NORTH AMERICA 48 chr. G=0.562 A=0.438 AFD_AFR_PANEL-NORTH AMERICA 46 chr. G=0.913 A=0.087 AFD_CHN_PANEL-NORTH AMERICA 48 chr. G=1.000	+	W720X	727 (1%)	–	–

(*continued*)

Table 1. *Continued.*

Gene	[a]Gene function	[b]Accession #	Location	[c]SNP ID	[d]Frequency	[e]Homo-zygosity	X-SNP	[f]Protein length (trun-cation)	[g]NMD	[h]CpG
OVCH2	Proteolysis	NM_198185.1	11p15.4	rs4509745	HapMap-CEU-EUROPE chr. 120 T=0.658 C=0.342 HapMap-HCB-EAST ASIA 88 chr. T=0.705 C=0.295 HapMap-JPT-EAST ASIA 88 chr. T=0.614 C=0.386 HapMap-YRI-WEST AFRICA 120 chr. C=0.783 T=0.217 AFD_EUR_PANEL-NORTH AMERICA 44 chr. T=0.568 C=0.432 AFD_AFR_PANEL-NORTH AMERICA 46 chr. C=0.609 T=0.391 AFD_CHN_PANEL-NORTH AMERICA 48 chr. T=0.583 C=0.417	+	W556X	564 (1.4%)	–	–
POLE2	DNA repair	NM_002692.2	14q21-q22	rs3218790	NIHPDR-NORTH AMERICA 170 chr. A=0.988 T=0.012 HapMap-CEU-EUROPE 120 chr. A=1.000 HapMap-HCB-EAST ASIA 90 chr. A=1.000 HapMap-JPT-EAST ASIA 88 chr. A=1.000 HapMap-YRI-WEST AFRICA 120 chr. A=1.000	–	K443X	527 (15.9%)	+	–
SER-PINB11	Serine-type endopeptidase inhibitor activity	NM_080475.1	18	rs4940595	AfAm 12 chr. C=0.667 A=0.333 Caucasian 24 chr. A=0.667 C=0.333 Asian 12 chr. C=0.667 A=0.333 CEPH 12 chr. C=0.667 A=0.333 PDpanel 48 chr. A=0.521 C=0.479 AFD_EUR_PANEL-NORTH AMERICA 48 chr. T=0.625 G=0.375 AFD_AFR_PANEL-NORTH AMERICA 44 chr. G=0.545 T=0.455 AFD_CHN_PANEL-NORTH AMERICA 48 chr. G=0.771 T=0.229	+	E90X	392 (77%)	+	–

SMUG1	DNA repair	NM_014311.1	12q13.11-q13.3	rs2233919	NIHPDR-NORTH AMERICA 574 chr. C=0.986 T=0.014 PDR90 166 chr. C=0.988 T=0.012	–	Q3X	270 (98.9%)	+	–
SPTBN5	Actin cytoskeleton organisation and biogenesis	NM_016642.1	15q21	rs2271286	JBIC-allele-EAST ASIA 1482 chr. G=0.951 A=0.049	–	Q72X	3674 (98%)	–	–
TAP2	Immune response; protein transport and assembly	NM_000544.2	6p21.3	rs241448	CEPH-MULTI-NATIONAL 184 T=0.700 C=0.300 WIAF-CSNP-MITOGPOP5-MULTI-NATIONAL 48 chr. T=0.812 C=0.188	+	Q687X	703 (2.3%)	–	–
TAAR9	Signal transduction	NM_175057.1	6q23.2	rs2842899	HapMap-CEU-EUROPE 120 chr. T=0.708 A=0.292 HapMap-YRI-WEST AFRICA 120 chr. T=0.883 A=0.117 AFD_EUR_PANEL-NORTH AMERICA 48 chr. A=0.812 T=0.188 AFD_AFR_PANEL-NORTH AMERICA 46 chr. A=0.783 T=0.217 AFD_CHN_PANEL-NORTH AMERICA 48 chr. A=0.854 T=0.146	+	Q61X	348 (82.5%)	–	–
TLR5	Immune response	NM_003268.3	1q41-q42	rs5744168	D-0-NORTH AMERICA 48 chr. C=0.938 T=0.062 E-0-NORTH AMERICA 40 chr. C=0.925 T=0.075 E-1-EUROPE 6 chr. C=1.000	–	R392X	858 (54.3%)	–	+

(*continued*)

Table 1. *Continued.*

Gene	[a]Gene function	[b]Accession #	Location	[c]SNP ID	[d]Frequency	[e]Homo-zygosity	X-SNP	[f]Protein length (trun-cation)	[g]NMD	[h]CpG
TRPM1	Cation transport	NM_002420.3	15q13-q14	rs3784589	JBIC-allele-EAST ASIA 1502 chr. C=0.965 A=0.035 HapMap-CEU-EUROPE 120 chr. C=0.942 A=0.058 HapMap-HCB-EAST ASIA 90 chr. C=1.000 HapMap-JPT-EAST ASIA 88 chr. C=0.955 A=0.045 HapMap-YRI-WEST AFRICA 118 chr. C=0.958 A=0.042 AFD_EUR_PANEL-NORTH AMERICA 48 chr C=0.917 A=0.083 AFD_AFR_PANEL-NORTH AMERICA 46 chr. C=0.913 A=0.087 AFD_CHN_PANEL-NORTH AMERICA 48 chr. C=1.000	+	E1305X	1533 (14.9%)	−	−
UNC93A	Unknown	NM_018974.2	6q27	rs2235197	JBIC-allele-EAST ASIA 1484 chr. G=0.852 A=0.148	n/a	W151X	456 (66.9%)	−	−
ZNF34	Gene expression	NM_030580.2	8q24.3	rs2294120	JBIC-allele-EAST ASIA 1494 chr. C=0.729 T=0.271	n/a	Q56X	549 (89.8%)	+	+

Abbreviation: SNP = single nucleotide polymorphism.
[a] Gene functions are retrieved from the Entrez Gene database of NCBI.[30]
[b] The accession numbers onto which the SNP-flanking sequences have been located.
[c] SNP ID corresponds to the dbSNP database SNP identifiers.
[d] The frequency information is as posted in dbSNP build 124.
[e] This information indicates whether or not a homozygous sample in a sample set was reported for the corresponding X-SNP and was collected from the dbSNP database 'summary of genotypes' section: 'n/a': no information was available, '+': homozygous genotype was reported, '−': no homozygous was reported.
[f] Length of the wild-type protein products. In parentheses are the percentages of the protein truncation at the C-terminus caused by the X-SNP.
[g] SNPs that may lead to nonsense-mediated mRNA decay are annotated by '+'.
[h] SNPs occurring at CpG dinucleotides and thus can be hot spot mutations are annotated by '+'.

SERPINB11-E90X 22.9–52.1 per cent; *ZNF34-Q56X* 27 per cent) (Table 1). Individuals with a homozygous state for two of these X-SNPs, namely *MS4A12-Q71X* and *SERPINB11-E90X*, were reported in dbSNP submissions, suggesting that, in such individuals, the levels of these truncated protein products are likely to be reduced by the NMD mechanism.

If no NMD occurs and the protein products are translated, then the PTCs lead to protein truncation at the C-terminus – the consequences of which vary depending on the degree of the truncation. For example, 17 (60.7 per cent) X-SNPs led to moderate to severe truncation at the C-terminus of the proteins (deletion of >50 per cent of the amino acid sequence), which is likely radically to alter protein structure and function (Table 1, Figure 1). As an extreme example, *SMUG1-Q3X*, when translated, would yield only a two-amino acid peptide, which would presumably be non-functional (loss of function). Also, PTCs can destabilise the protein products by altering the protein-folding state or kinetics[14,31] and may cause proteolysis. In addition, they may act as dominant-negative mutations[15] or cause exon skipping and alter the open reading frame.[32] Alternatively, mRNA molecules bearing a PTC closer to the 5′ end can be still translated if an in-frame translationable AUG start codon is present downstream of the PTC.[33–35] Such N-terminal truncated proteins can be fully or partially functional. For example, in the case of *SMUG1-Q3X*, there is an in-frame AUG located at the 18th codon of the *SMUG1* gene, which can be experimentally evaluated to determine if an N-terminal truncated SMUG1 protein is produced and functional. To summarise, the stability, structure and function of the protein products or transcripts may be affected by the X-SNPs described in this study, and experimental approaches are needed to evaluate their true biological effects.

Possible evolutionary explanations of common X-SNPs

The small number of validated X-SNPs identified suggests infrequent occurrence of the PTC-introducing variations in the human genome and thus agrees with the presence of selection against them.[18] By contrast with rare PTC-introducing mutations observed in human diseases, however, X-SNPs analysed in this study represented commonly occurring variations in humans: 22 X-SNPs (78.6 per cent) were found with minor allele frequencies of ≥ 5 per cent in at least one sample panel analysed (common X-SNPs) compared with

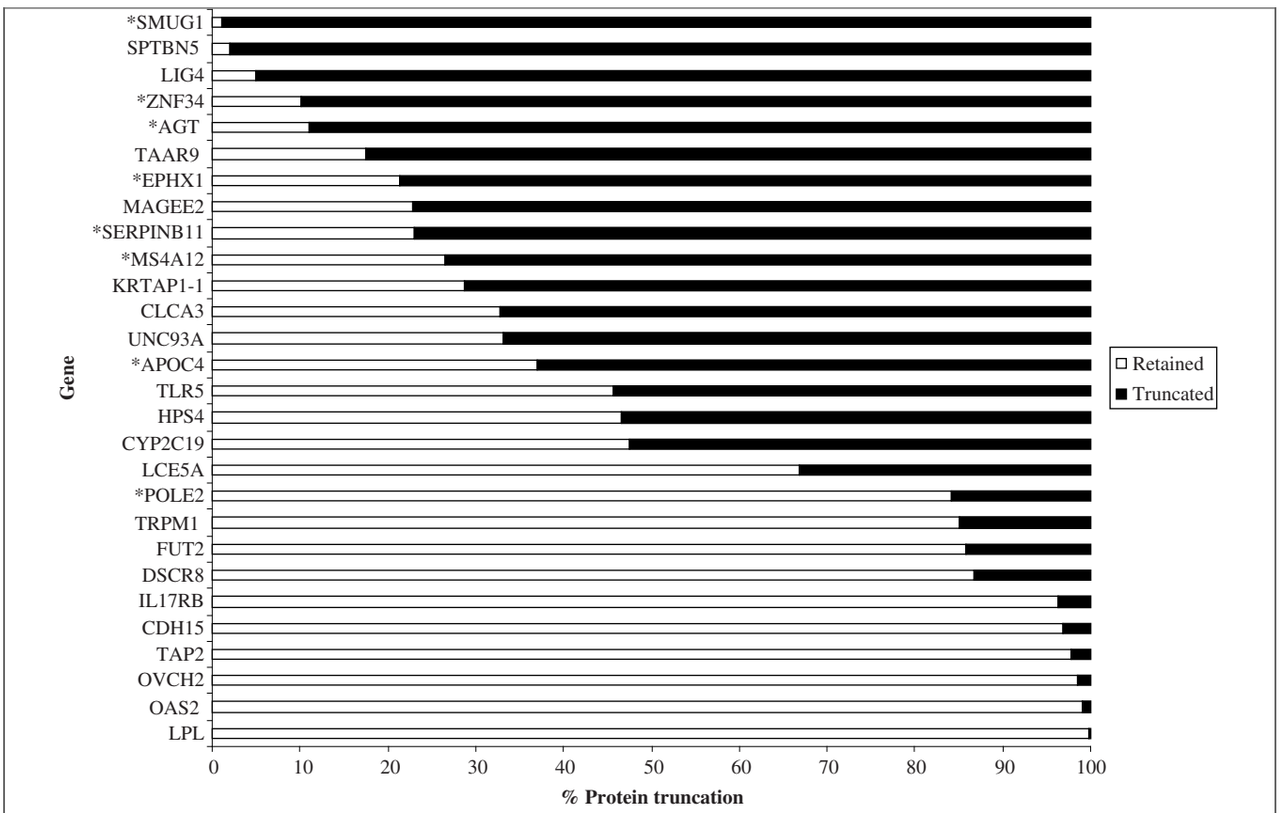

Figure 1. Percentage truncation induced by the X-SNPs in the corresponding proteins. The X-SNPs that can cause mRNA degradation via NMD are indicated by * in the left margin of the histogram.

only six rare X-SNPs (with <5 per cent minor allele frequencies).

How can we explain the abundance of such common (and perhaps deleterious) X-SNPs in the human population? Possible scenarios are summarised in Figure 2. For example, one explanation could be that the truncated protein product may still be functional in the presence of the X-SNP. For instance, *LPL-S474X* was located only one amino acid prior to the natural termination codon; thus, it may not really alter the protein properties and thus may not be deleterious to cell function at all. Alternatively, the protein may not be essential for the fitness of human beings; in this case, the evolutionary pressure is relieved, which can lead to toleration of an increase in allele frequency of premature stop codons in human populations.

Another possibility is that X-SNPs may be capable of affecting protein function/the organism *per se*, but other factors might modify their effects. Here, we will assume that these PTCs represent both the strongly deleterious mutations that are a result of selection and quickly removed from the populations, as well as the slightly deleterious mutations that are subject to both selection and drift.[36] For example, these X-SNPs may be hot-spot mutations, where the new mutations introduce (slightly) deleterious alleles and thus increase the allele frequency, despite the selection. In order to assess whether some of these X-SNPs might in fact represent the hot-spot mutations, we analysed the immediate flanking sequences of each X-SNP. As a result, we found that 25 per cent (7/28) of X-SNPs (all common) had occurred at CpG dinucleotides (Table 1). These data suggest that these X-SNPs might have arisen from spontaneous deamination of methyl-cytosine leading to a thymine, and thus may represent hot-spot mutations.[37]

Additionally, diploidy was suggested to relieve the tension of purifying selection and increase the tolerance for PTCs,[38] which predicts a recessive effect or loss of function. All but one gene (*MAGEE2*) in Table 1 were located in autosomal chromosomes, which may also help to explain the frequency of the naturally occurring PTC polymorphisms in humans. Moreover, it is also likely that, even though (slightly) deleterious in a homozygous state, some X-SNPs can confer selective advantage to heterozygotes.[39] Alternatively, epistatic interactions of additional mutations, either on the same or different genes, may compensate for the (slightly) deleterious effects of the X-SNP.[16,40] Furthermore, X-SNPs may be beneficial at present conditions, which may favour the positive selection of the X-SNPs and increase their allele frequencies. Moreover, if a PTC is located at the 5′ end of a gene and there is a nearby in-frame initiation codon after that PTC, then the protein translation can re-initiate and a peptide with amino-truncation may be produced.[33–35] Depending on the nature and extent of the truncation, the truncated peptide can fully or partially function and thus can, completely or to some extent, rescue the phenotype. There is a need for further studies to elucidate the molecular basis of the discrepancy and the determination of the biological differences between human disease-related mutations and naturally occurring stop codon-creating polymorphisms.

Frequency spectrum of X-SNPs in different human populations

Comparison of the population(s) and the minor allele frequencies of X-SNP entries in the dbSNP database[19] presented great variability across different human populations, at least in some cases (Table 1). For example, *HSP4-R246X*, *IL17RB-Q484X*, *LPL-S474X*, *MS4A12-Q71X*, *OVCH2-W556X*, *SERPINB11-E90X*, *TAAR9-Q61X* and *TRPM1-E1305X* were detected in samples from African, Asian and Caucasian backgrounds. This might mean that either these X-SNPs have been inherited from a common ancestor or they represent hot-spot mutations (*HSP4-R246X* and *IL17RB-*

High allelic frequency:

1. X-SNP is not deleterious because the protein product is still functional
2. X-SNP is deleterious but also is a hot spot mutation
3. X-SNP is deleterious but (mildly) tolerated because of diploidy/heterozygote advantage
4. X-SNP is beneficial
5. X-SNP is deleterious to protein function, but the protein is not required for the fitness of the organism
6. X-SNP is deleterious to protein function, but protein function is compensated by other protein(s)
7. X-SNP is deleterious but compensated by other mutation(s) either in the same or in other gene(s)

Low allelic frequency:

1. X-SNP is relatively new in the human population
2. X-SNP is deleterious and thus subject to purifying selection
3. Technical issues (sample size is not large enough to draw conclusions from, errors in genotyping, population specificity etc)

Figure 2. How can we explain the allele frequencies of the X-SNPs? This figure presents a summary of possible biological consequences of X-SNPs. For simplicity, both deleterious and slightly deleterious variations are annotated as deleterious.

Q484X occurred at CpG dinucleotides and thus might in fact be hot-spot mutations; see Table 1). By contrast, *CYPC19-W212X* (African and Asian), *EPHX1-W97X* (Asian and Hispanic) and *OAS2-W720X* (African and European) were detected in some populations but not in others. In addition, there were three X-SNPs that were found exclusively in one population: *FUT2-R297X* and *LCE5A-R79X* in Asian and *LIG4-W46X* in African samples. Either different selection in different populations or the occurrence of founder effect/genetic drift may explain the population spectrum of these SNPs.[16,41]

Conclusion

In conclusion, we have evaluated SNPs that introduce PTCs in the human genome that can potentially affect the stability of transcripts and their protein products. Although there is considerable information regarding the PTC-creating mutations in human genetic diseases, to date, there has been no systematic study reporting on the PTC-causing polymorphisms in the human genome and their evolutionary and biological roles in humans. Our results indicated that the allelic frequencies of the disease-causing PTC-creating mutations and polymorphisms display a marked difference. These X-SNPs were found in a variety of proteins with different cellular functions (signal transduction, DNA repair, transcription, immune response, drug metabolism etc; Table 1). A search of literature reports and the Human Gene Mutation Database[42] showed that a fraction of these genes have already been implicated in human diseases: *AGT* in essential hypertension;[43] *HPS4* in Hermansky–Pudlak syndrome type 4;[44] *LPL* in disorders of lipoprotein metabolism;[45] and *TLR5* in pneumonia caused by *Legionella pneumophila*.[46] In the latter case, the *TLR5-R392X* SNP was functionally characterised and found to be defective in flagellin signalling and associated with the pneumonia susceptibility.[46] In the case of the *TAP2-Q687X* SNP, *TAP2-Q687* was reported to be a part of a haplotype associated with a reduced risk of insulin-dependent diabetes mellitus in a small sample set.[47] Our data suggest a potential deleterious effect for X-SNPs identified in this study; however, their true biological consequences and potential roles in human disease and health have yet to be experimentally verified and identified.

Acknowledgments

The authors thank Baris Tuncertan and Mehjabeen Shariff for automatic retrieval of the SNPs from the dbSNP database and Stewart Cho for editing the manuscript. This work was supported by a grant (BCTR0100627) from the Susan Komen Breast Cancer Foundation, USA. Sevtap Savas is supported, in part, by a "CIHR Strategic Training Program Grant - The Samuel Lunenfeld Research Institute Training Program: Applying Genomics to Human Health" fellowship.

Electronic database information

BLAST: http://www.ncbi.nlm.nih.gov/BLAST/.
BLAST against gene transcripts: http://lpgws.nci.nih.gov:80/perl/blast2/.
dbSNP: http://www.ncbi.nlm.nih.gov/SNP/.
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD = search&DB = gene.
GenBank: http://www.ncbi.nih.gov/Genbank/.
Human Gene Mutation Database: http://www.hgmd.cf.ac.uk/hgmd0.html.

References

1. Gray, I.C., Campbell, D.A. and Spurr, N.K. (2000), 'Single nucleotide polymorphisms as tools in human genetics', *Hum. Mol. Genet.* Vol. 9, pp. 2403–2408.
2. Miller, R.D. and Kwok, P-Y. (2001), 'The birth and death of human single-nucleotide polymorphisms: New experimental evidence and implications for human history and medicine', *Hum. Mol. Genet.* Vol. 10, pp. 2195–2198.
3. Taylor, J.G., Choi, E.H., Foster, C.B. and Chanock, S.J. (2001), 'Using genetic variation to study human disease', *Trends Mol. Med.* Vol. 7, pp. 507–512.
4. Shastry, B.K. (2002), 'SNP alleles in human disease and evolution', *J. Hum. Genet.* Vol. 47, pp. 561–566.
5. Brookes, A.J. (1999), 'The essence of SNPs', *Gene* Vol. 234, pp. 177–186.
6. Lercher, M.J. and Hurst, L.D. (2002), 'Human SNP variability and mutation rate are higher in regions of high recombination', *Trends Genet.* Vol. 18, pp. 337–340.
7. Chakravarti, A. (1999), 'Population genetics — Making sense out of sequence', *Nat. Genet.* Vol. 21(1), pp. 56–60.
8. Thomas, F.J., McLeod, H.L. and Watters, J.W. (2004), 'Pharmacogenomics: The influence of genomic variation on drug response', *Curr. Top. Med. Chem.* Vol. 4, pp. 1399–1409.
9. Sunyaev, S., Ramensky, V., Koch, I. *et al.* (2001), 'Prediction of deleterious human alleles', *Hum. Mol. Genet.* Vol. 10, pp. 591–597.
10. Wang, Z. and Moult, J. (2001), 'SNPs, protein structure, and disease', *Hum. Mutat.* Vol. 17, pp. 263–270.
11. Ng, P.C. and Henikoff, S. (2002), 'Accounting for human polymorphisms predicted to affect protein function', *Genome Res.* Vol. 12, pp. 436–446.
12. Ramensky, V., Bork, P. and Sunyaev, S. (2002), 'Human non-synonymous SNPs: Server and survey', *Nucleic Acids Res.* Vol. 30, pp. 3894–3900.
13. Byers, P.H. (2002), 'Killing the messenger: New insights into nonsense-mediated mRNA decay', *J. Clin. Invest.* Vol. 109, pp. 3–6.
14. Gregersen, N., Bross, P., Jorgensen, M.M. *et al.* (2000), 'Defective folding and rapid degradation of mutant proteins is a common disease mechanism in genetic disorders', *J. Inherit. Metab. Dis.* Vol. 23, pp. 441–447.
15. Schell, T., Kulozik, A.E. and Hentze, M.W. (2002), 'Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway', *Genome Biol.*, Vol. 3, pp. REVIEWS1006.
16. Fay, J.C. and Wu, C.I. (2003), 'Sequence divergence, functional constraint, and selection in protein evolution', *Annu. Rev. Genomics Hum. Genet.* Vol. 4, pp. 213–235.
17. Frischmeyer, P.A. and Dietz, H.C. (1999), 'Nonsense-mediated mRNA decay in health and disease', *Hum. Mol. Gen.* Vol. 8, pp. 1893–1900.
18. Sawyer, S.L., Berglind, L.C. and Brookes, A.J. (2003), 'Negligible validation rate for public domain stop-codon SNPs', *Human Mut.* Vol. 22, pp. 252–254.
19. Sherry, S.T., Ward, M.H., Kholodov, M. *et al.* (2001), 'dbSNP: The NCBI database of genetic variation', *Nucleic Acids Res.* Vol. 29, pp. 308–311.
20. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. *et al.* (2004), 'GenBank: Update', *Nucleic Acids Res.* Vol. 32, pp. D23–D26.
21. Clifford, R.J., Edmonson, M.N., Nguyen, C. and Buetow, K.H. (2004), 'Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms', *Bioinformatics* Vol. 20, pp. 1006–1014.

22. Savas, S., Ahmad, F., Kim, D.Y. *et al.* (2005), 'Candidate nsSNPs that can affect the functions and interactions of the cell cycle genes', *Proteins* Vol. 58, pp. 697–705.
23. Wheeler, D.L., Church, D.M., Edgar, R. *et al.* (2004), 'Database resources of the National Center for Biotechnology Information: Update', *Nucleic Acids Res.* Vol. 32, pp. D35–D40.
24. Estivill, X., Cheung, J., Pujana, M.A. *et al.* (2002), 'Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome', *Hum. Mol. Genet.* Vol. 11, pp. 1987–1995.
25. Savas, S., Kim, D.Y., Ahmad, M.F. *et al.* (2004), 'Identifying functional genetic variants in DNA repair pathway using protein conservation analysis', *Cancer Epidemiol. Biomarkers Prev.* Vol. 13, pp. 801–807.
26. Pruitt, K.D. and Maglott, D.R. (2001), 'RefSeq and LocusLink: NCBI gene-centered resources', *Nucleic Acids Res.* Vol. 29, pp. 137–140.
27. Thermann, R., Neu-Yilik, G., Deters, A. *et al.* (1998), 'Binary specification of nonsense codons by splicing and cytoplasmic translation', *EMBO J.* Vol. 17, pp. 3484–3494.
28. Zhang, J., Sun, X., Qian, Y. *et al.* (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.* Vol. 18, pp. 5272–5283.
29. Baker, K.E. and Parker, R. (2004), 'Nonsense-mediated mRNA decay: Terminating erroneous gene expression', *Curr. Opin. Cell Biol.* Vol. 16, pp. 293–299.
30. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005), 'Entrez Gene: Gene-centered information at NCBI', *Nucleic Acids Res.* Vol. 33, pp. D54–D58.
31. Williams, R.S., Chasman, D.I., Hau, D.D. *et al.* (2003), 'Detection of protein folding defects caused by BRCA1-BRCT truncation and missense mutations', *J. Biol. Chem.* Vol. 278, pp. 53007–53016.
32. Liu, H.X., Cartegni, L., Zhang, M.Q. and Krainer, A.R. (2001), 'A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes', *Nat. Genet.* Vol. 27, pp. 55–58.
33. Ozisik, G., Mantovani, G., Achermann, J.C. *et al.* (2003), 'An alternate translation initiation site circumvents an amino-terminal DAX1 nonsense mutation leading to a mild form of X-linked adrenal hypoplasia congenita', *J. Clin. Endocrinol. Metab.* Vol. 88, pp. 417–423.
34. Heppner Goss, K., Trzepacz, C., Tuohy, T.M. and Groden, J. (2002), 'Attenuated APC alleles produce functional protein from internal translation initiation', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 8161–8166.
35. Howard, M.T., Malik, N., Anderson, C.B. *et al.* (2004), 'Attenuation of an amino-terminal premature stop codon mutation in the *ATRX* gene by an alternative mode of translational initiation', *J. Med. Genet.* Vol. 41, pp. 951–956.
36. Ohta, T. (2002), 'Near-neutrality in evolution of genes and gene regulation', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 16134–16137.
37. Tomso, D.J. and Bell, D.A. (2003), 'Sequence context at human single nucleotide polymorphisms: Overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands', *J. Mol. Biol.* Vol. 327, pp. 303–308.
38. Xing, Y. and Lee, C.J. (2004), 'Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy', *Trends Genet.* Vol. 20, pp. 472–475.
39. Dean, M., Carrington, M. and O'Brien, S.J. (2002), 'Balanced polymorphism selected by genetic versus infectious human disease', *Annu. Rev. Genomics Hum. Genet.* Vol. 3, pp. 263–292.
40. Cordell, H.J. (2002), 'Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans', *Hum. Mol. Genet.* Vol. 11, pp. 2463–2468.
41. Cavalli-Sforza, L.L. and Feldman, M.W. (2003), 'The application of molecular genetic approaches to the study of human evolution', *Nat. Genet.* Vol. 33, pp. 266–275.
42. Stenson, P.D., Ball, E.V., Mort, M. *et al.* (2003), 'Human Gene Mutation Database (HGMD): 2003 update', *Hum. Mutat.* Vol. 21, pp. 577–581.
43. Jeunemaitre, X., Soubrier, F., Kotelevtsev, Y.V. *et al.* (1992), 'Molecular basis of human hypertension: Role of angiotensinogen', *Cell* Vol. 71, pp. 7–20.
44. Anderson, P.D., Huizing, M., Claassen, D.A. *et al.* (2003), 'Hermansky–Pudlak syndrome type 4 (HPS-4): Clinical and molecular characteristics', *Hum. Genet.* Vol. 113, pp. 10–17.
45. Otarod, J.K. and Goldberg, I.J. (2004), 'Lipoprotein lipase and its role in regulation of plasma lipoproteins and cardiac risk', *Curr. Atheroscler. Rep.* Vol. 6, pp. 335–342.
46. Hawn, T.R., Verbon, A., Lettinga, K.D. *et al.* (2003), 'A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires disease', *J. Exp. Med.* Vol. 198, pp. 1563–1572.
47. Clonna, M., Bresnahan, M., Bahram, S. *et al.* (1992), 'Allelic variants of the human putative peptide transporter involved in antigen processing', *Proc. Natl. Acad. Sci. USA* Vol. 89, pp. 3932–3936.