A Genome-wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs

Bryndis Yngvadottir,[1] Yali Xue,[1] Steve Searle,[1] Sarah Hunt,[1] Marcos Delgado,[1] Jonathan Morrison,[1,2] Pamela Whittaker,[1] Panos Deloukas,[1] and Chris Tyler-Smith[1,*]

Nonsense SNPs introduce premature termination codons into genes and can result in the absence of a gene product or in a truncated and potentially harmful protein, so they are often considered disadvantageous and are associated with disease susceptibility. As such, we might expect the disrupted allele to be rare and, in healthy people, observed only in a heterozygous state. However, some, like those in the *CASP12* and *ACTN3* genes, are known to be present at high frequencies and to occur often in a homozygous state and seem to have been advantageous in recent human evolution. To evaluate the selective forces acting on nonsense SNPs as a class, we have carried out a large-scale experimental survey of nonsense SNPs in the human genome by genotyping 805 of them (plus control synonymous SNPs) in 1,151 individuals from 56 worldwide populations. We identified 169 genes containing nonsense SNPs that were variable in our samples, of which 99 were found with both copies inactivated in at least one individual. We found that the sampled humans differ on average by 24 genes (out of about 20,000) because of these nonsense SNPs alone. As might be expected, nonsense SNPs as a class were found to be slightly disadvantageous over evolutionary timescales, but a few nevertheless showed signs of being possibly advantageous, as indicated by unusually high levels of population differentiation, long haplotypes, and/or high frequencies of derived alleles. This study underlines the extent of variation in gene content within humans and emphasizes the importance of understanding this type of variation.

Introduction

The theory that gene duplication is a major factor in shaping evolution was proposed many years ago by Susumu Ohno[1] and is now widely accepted. The idea that gene loss can also contribute significantly to evolution is, however, a newer one and was proposed by Maynard Olson.[2] Common sense may lead us to consider gene loss as a bad thing and to associate adaptation with genes that are somehow "better." However, as the thrifty gene theory[3] proposed, some genes that were advantageous in the past may have become a burden in modern times.

One molecular mechanism for gene loss is the introduction of a premature termination codon (PTC). This can result from a nonsense mutation, a frame-shifting indel or a splice-site mutation with the skipping of a single exon containing a number of nucleotides that cannot be divided by three (reviewed in Cartegni et al.[4]). A PTC could result in a shorter protein, but truncated proteins are likely to be deleterious and are usually eliminated by a process called nonsense-mediated mRNA decay (NMD).[5,6] If the NMD pathway is triggered, it will eliminate the production of the protein, and the gene product will be completely lost. However, if the PTC is located either in the last exon or less than 50–55 nucleotides upstream of the last exon-exon boundary, NMD can be evaded, resulting in the production of a truncated protein.[5,7]

A gene-loss event begins with a mutation within a single individual, and if the disrupted allele (hereafter referred to as the "stop allele," as opposed to the nondisrupted "normal allele") is neutral, it can either increase or decrease its frequency in a population by the random effects of genetic drift. However, if the stop allele turns out to be harmful to its carrier, it will tend to be eliminated by the forces of negative selection, whereas should it be advantageous, positive selection will act to increase its frequency. Although nonsense SNPs are common causes of genetic disease,[8] the stop alleles in the *CASP12*[9] (MIM *608633) and *ACTN3*[10] (MIM +102574) genes are found at high frequencies, are often in a homozygous state, and seem to have been advantageous in recent human evolution. Carriers of the stop allele in *CASP12* are more resistant to severe sepsis[11], and the stop allele in *ACTN3* has been associated with increased endurance in athletic performance.[12,13]

Recent studies[14,15] have provided us with important insights into the number, location within the protein, and predicted effects of nonsense SNPs in silico by using publicly available data from the dbSNP database. We have now extended these investigations to test the less-is-more hypothesis[2] by evaluating the evolutionary forces acting on nonsense SNPs as a class, genotyping 805 such SNPs in 56 worldwide populations, and resequencing a gene containing one example and its surrounding region. Our aim was to identify outliers that could potentially reveal additional contributions of gene loss to the evolution of our species.

Material and Methods

DNA Samples

DNA samples were obtained from the HapMap and extended HapMap populations[16,17] and the human genome diversity cell line panel (HGDP-CEPH),[18] from which the H1048[19] subset was used. The samples successfully genotyped were derived from 1,151 individuals from 56 geographically diverse populations

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK
[2]Present address: Cancer Research UK, Genetic Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK
*Correspondence: cts@sanger.ac.uk
DOI 10.1016/j.ajhg.2009.01.008.

(Figure S1 available online). The samples sequenced were 22 CEPH Utah residents with ancestry from northern and western Europe (CEU), 23 Yoruba in Ibadan, Nigeria (YRI), 23 Han Chinese in Beijing (CHB), and 23 Luhya in Webuye, Kenya (LWK).[16] In addition, one chimpanzee (*Pan troglodytes*) sample was included as an outgroup. All HapMap samples were purchased from the Coriell Institute for Medical Research (Camden, New Jersey, USA); the HGDP-CEPH collection[18] was kindly provided by Howard Cann (CEPH, Paris, France), and the chimpanzee sample was purchased from the ECACC (Salisbury, Wiltshire, UK).

Data Generation

Genotyping of 1,536 SNPs

Assay designs were attempted for all nonsense SNPs in dbSNP build 121; assays that failed at the design stage and others that passed the design but were known to fail from previous genotyping attempts were excluded, leaving 805 nonsense SNPs to be tested. We analyzed only nonsense SNPs that introduce stop codons (sometimes referred to as "stop gained") and excluded SNPs causing a stop codon read-through ("stop lost"). Additionally, 731 synonymous SNPs were added to provide a total of 1,536 SNPs, the number required for one bundle of an Illumina BeadArray. The synonymous SNPs were chosen to act as controls: although not perfectly neutral, they nevertheless provide an approximation of neutral variants. We selected these synonymous SNPs to roughly match the sources (submitters) of the nonsense SNPs in order to match SNPs that might have been called on the basis of poor sequencing or discovery in particular populations.

Many investigators have contributed to the discovery of the SNPs in databases by sequencing a limited (and sometimes very small) number of individuals; the sequenced regions were not consistent throughout the genome and were generally not well documented. Interpretation of the SNP genotypes in additional individuals can thus potentially be influenced by the discovery process, an effect known as "ascertainment bias." We incorporated a number of factors into our study design so that we could reach useful conclusions, despite such bias. First, by starting from the set of all SNPs in the database, we avoided the most extreme forms of ascertainment bias, and below we show examples of nonsense SNPs confined to non-European samples and demonstrate that an overtly Europe-centric bias has been avoided. Second, we compared analyses of nonsense SNPs with source-matched analyses of synonymous SNPs to ensure that the analyses were subject to the same ascertainment bias. Third, we concentrated largely on analyses less influenced by ascertainment bias.

Genotyping was carried out by the Sanger Genotyping Platform Group via the Illumina GoldenGate assay[20] with the primers listed in Table S1, and the results were subjected to sequential quality-control filters. Each plate contained three duplicates, and SNPs with more than 33% discrepancies between duplicates were excluded. The Gene Call (GC) score, which gives the confidence of the genotype read (intensity), was then estimated. A very low value is not to be trusted. Genotypes without call, individual genotypes with a GC score less than 0.25, assays with a median GC score lower than 0.3, and assays with less than 80% data were also discarded. Additional manual assessments were also applied. First, we excluded nonsense SNPs that overlapped with Vega[21] pseudogenes. We then excluded SNPs if the ancestral state could not be inferred. Lastly, we used the Tblastx tool to search for the ORF of the sequence surrounding SNPs that had "stop lost" listed as a consequence and got rid of those for which the ancestral state (chimpanzee) was found to be the PTC and the derived state (human) was found to be a read through of the protein. The final dataset we used in the analysis consisted of 453 SNPs (169 nonsense and 284 synonymous) that passed the quality controls and were polymorphic in our samples. The genotype of each sample is provided in Table S2 (a tab-delimited .txt file).

Resequencing MAGEE2

Two ~6.5 kb fragments that cover the whole *MAGEE2* gene and an additional ~5 kb on each side of it were amplified by long-polymerase chain reaction (long-PCR). Primers are listed in Table S1. Reactions (15 μl) contained 1 × High-Fidelity PCR Buffer (Invitrogen, Paisley, UK), 2 mM $MgSO_4$, 200 μM each dNTP, 0.6 U Platinum Taq DNA Polymerase High Fidelity (Invitrogen), 0.4 μM of each primer and 125 ng genomic DNA. A touchdown protocol beginning with 2 min denaturation at 94°C, followed by 15 cycles of 94°C for 30 s, 68°C for 30 s (temperature decreased by 0.5°C each cycle), and 68°C for 6 min, then 20 cycles of 94°C for 30 s, 58°C for 30 s, and 68°C for 6 min, and finishing with extension at 68°C for 7 min was used. Nested PCR products of 500 (±15%) bp overlapping by 240 (±30%) bp were then amplified with the primers in Table S1; each 15 μl PCR contained 0.5 μl of 400× diluted long-PCR products, 0.5 U Platinum Taq (Invitrogen), 1 × buffer (Invitrogen), 1.6 mM $MgCl_2$, 10 pmol of each primer, and 200 μM of each dNTP, and the cycle conditions were 94°C for 15 min, 30 cycles of 94°C for 45 s, 61°C for 45 s, 72°C for 45 s, and finally 72°C for 7 min. Products were sequenced on both strands by the Sanger Large-Scale Sequencing Pipeline with BigDye Sanger sequencing technology and a 3730 xl DNA Analyzer (Applied Biosystems). Potential variable positions were flagged by the Mutation Surveyor v.2.0 software (SoftGenetics, PA, USA) and checked manually. Four blind duplicates were included for quality control and showed complete concordance. The SNP variation identified in *MAGEE2* is provided in Table S3 (a tab-delimited .txt file).

Data Analyses

Descriptive Statistics

We used the Table Browser on the UCSC Genome Browser website to retrieve the ancestral allele for ~98% (445/453) of the SNPs from the "snp126OrthoPanTro2RheMac2" table. The chimpanzee (*Pan troglodytes*) sequence provided the primary ancestral state, but we accepted sequences from other primates (*Macaca mulatta* or *Lagothrix lagotricha*) when the chimpanzee sequence was not available. The derived allele was then defined as the other observed human allele. We then looked manually for the ancestral state of the missing 2% (eight SNPs) by using FASTA sequences and the NCBI Blastn algorithm to find the best hit within a primate reference sequence. We obtained the derived allele frequency by direct allele counting and used a Kolmogorov-Smirnov test to evaluate the difference between the distributions of nonsense and synonymous SNPs.

We found 112 genes bearing nonsense SNPs and coding for a single transcript. The remaining 57 nonsense SNPs were found in genes undergoing alternative splicing and were reported in more than one transcript. For such SNPs we used the transcript showing the largest truncation in subsequent calculations. We estimated the proportion of protein truncation each SNP would cause as the percentage of the ancestral ORF length (100 − (SNP protein position/protein length*100)). We used the SNP2NMD database[22] to assess whether our nonsense SNPs were likely to trigger NMD according to the 50–55-nucleotide rule.[5] Approximately 63% (107) of our nonsense SNPs were in SNP2NMD, and for these we set the "NMD distance" (distance between a SNP and the 3′-most exon-exon junction) to be >50 nucleotides for the NMD pathway to be triggered. For the remaining 62 (~37%)

SNPs missing from SNP2NMD, we extracted information on the location of the nonsense SNP with respect to exon-intron boundaries from Ensembl (releases 37 and 43) and calculated the prediction for NMD manually.

We performed a gene ontology (GO)[23] term-enrichment analysis on 167 genes containing nonsense SNPs with the DAVID chart analysis tool.[24] All available GO terms were used, and all human genes (implemented in DAVID) were defined as the background. p values were calculated with the EASE score, which is a modified conservative adjustment of the one-tailed Fisher's exact test[25] and is implemented in DAVID. Terms with values below 0.05 were considered to be enriched.

Statistical Analyses

F_{ST}[26] was calculated with the R package HIERFSTAT[27] via the 37 population division (Figure S1), and differences between the distributions of nonsense and synonymous SNPs were assessed with the Kolmogorov-Smirnov test. For comparison with empirical data, we downloaded a set of 650 K publicly available SNPs genotyped by Stanford University in the HGDP-CEPH populations and calculated their F_{ST} values to find out whether our SNPs were significant outliers (i.e., lying above the 95th or 99th percentiles). Heterozygosity, the probability that any two randomly chosen samples from a population are the same, was calculated for each SNP according to Nei.[28] To estimate the strength of selection, we calculated the average selection coefficient(s) for each nonsense SNP in our set by using estimates for the number of coding nucleotides in the human genome (6.0×10^7), the average mutation rate (2.5×10^{-8}/nucleotide/generation[29]), and the fraction of mutations that can create a PTC (~1/20), together with our estimate for the average number of stop alleles per human diploid genome (46). A selection coefficient close to zero represents neutrality, and the higher the value, the more deleterious the mutation will be. The diploid genomic rate at which nonsense SNPs arise is 7.5×10^{-2}/individual/generation. On average, a nonsense SNP persists for $46/(7.5 \times 10^{-2}) = 613$ generations, implying s ~0.0016.

In order to calculate the relative extended haplotype homozygosity (REHH) statistic[30], we used the phased HapMap data (Build36), which included the majority (131 out of 169) of our nonsense SNPs; we then defined each nonsense SNP as a "core" and included 100 kb regions on each side. As controls, we chose 30 ENCODE random regions (~500 kb each), which we assumed to be neutral; this was a conservative assumption because random regions might have contained selected genes. The REHH test was performed with Sweep, and REHH was calculated with the default setting of a 0.04 marker breakdown from the core SNP. We used DnaSP[31] to calculate the summary statistics Tajima's D,[32] Fu and Li's D, D^*, F, and F^*,[33] Fu's F_S[34] and Fay and Wu's H.[35] We obtained the null distribution from simulations run by using a custom modification of the ms program[36] and incorporating the best-fit demographic model for each population,[37] and thus departures from neutrality take into account known demographic influences. Haplotypes for the resequenced data were inferred with PHASE 2.1.[38,39] Median-joining networks[40] were constructed from the inferred haplotypes with Network 4.5 and used for estimating the time to the most recent common ancestor (TMRCA) of a specified set of chromosomes, under the assumption of a time of 6.5 million years ago for the chimpanzee-human split. TMRCA was also estimated with GENETREE.[41] GENETREE employs a full maximum-likelihood method that is based on the standard coalescent[42] and assumes an infinite-sites mutational model. We estimated theta to be 5.95 by using a model of three populations (African [YRI + LKW], European [CEU], and Asian [CHB]) and performed 100,000 simulations (Ne = 15,700 with chimpanzee-human split 6.5 million years ago). Finally, with this value of theta and the populations connected by the migration rates suggested by the best-fit demographic model,[37] we estimated the TMRCA by using ten runs each of ten million simulations and chose the run with the lowest standard deviation, as recommended.[41]

Results

After genotyping 1,536 SNPs in 1,151 individuals, we identified 167 genes containing 169 nonsense SNPs that were variable in our samples. A full list of the genes is given in Table S4, including a summary of all the results presented here. Two genes, *CDKL1* (MIM *603441) and *FMO2* (MIM *603955), were found with two nonsense SNPs each (*CDKL1* with rs11570829 and rs7148089; *FMO2* with rs2020866 and rs6661174) and might therefore be suspected to be pseudogenes. However, as part of our manual assessment we had excluded all genes that overlapped with the Vega set of pseudogenes[21] and because these two genes are not annotated as pseudogenes, they are included in the results.

Genotyping revealed that on average the individuals in our samples carry ~14 stop/stop homozygous SNPs and ~18 stop/normal heterozygous SNPs in their genome, a total of ~46 stop alleles per diploid genome or ~23 per haploid genome. Furthermore, these individuals were found to differ on average by 24 genes per diploid genome because of nonsense SNPs. Because the polymorphic nonsense SNPs analyzed here are only a fraction of the nonsense SNPs reported in the human genome (Figure 1), and because these in turn are only a fraction of all nonsense SNPs (but also contain some false positive calls), these estimates are lower bounds, and the actual average difference is likely to be higher. However, because the distribution of our nonsense SNPs appears random in the human genome, they can be considered to represent nonsense SNPs as a class in the following analyses.

The Consequences of Nonsense SNPs

Next, we wished to understand the effects these 169 nonsense SNPs might be having on the gene product and the carrier. At the molecular level, the stop allele could result in a truncated protein or in the complete loss of the gene product if NMD is triggered. We found that the truncations were distributed evenly throughout the polypeptide length (Figure 2). Forty-nine percent of the nonsense SNPs lead to the deletion of >50% of the amino acid sequence, an extensive truncation that might radically alter the protein structure and function. In addition, 55% of nonsense SNPs were predicted to cause transcript degradation by NMD (in at least one transcript), which would result in loss of the gene function, and the rest of the nonsense SNPs (45%) are expected to result in the production of a truncated protein (Figure 2). Either way, most of these nonsense SNPs could be having severe effects on the gene product.

Do these SNPs therefore potentially cause a recessive disorder so that they are found only in the heterozygous

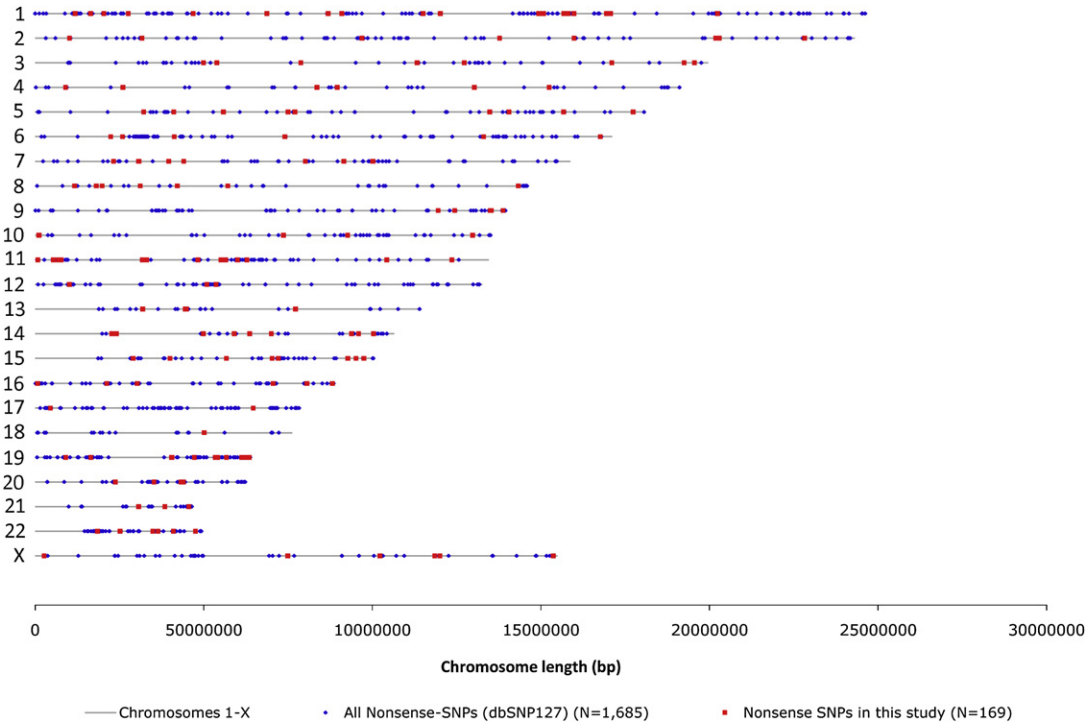

Figure 1. Genome-wide Distribution of Nonsense SNPs on Chromosomes 1 to X in the Human Genome
The nonsense SNPs that were variable in our samples are displayed in red, and all nonsense SNPs reported in the human genome (dbSNP127) are shown in blue.

state in the HapMap and HGDP-CEPH donors? For 99 nonsense SNPs (59%), at least one stop homozygous sample was found (Figure 3), showing that both copies of these genes could be truncated in our sampled individuals. We do not find unexpectedly high frequencies of heterozygotes: no significant departures from Hardy-Weinberg equilibrium were found in individual populations. In addition, only eight of the 169 nonsense SNPs were found in the Human Gene Mutation Database (HGMD) of mutations associated with human inherited disease.[43] For three of these eight

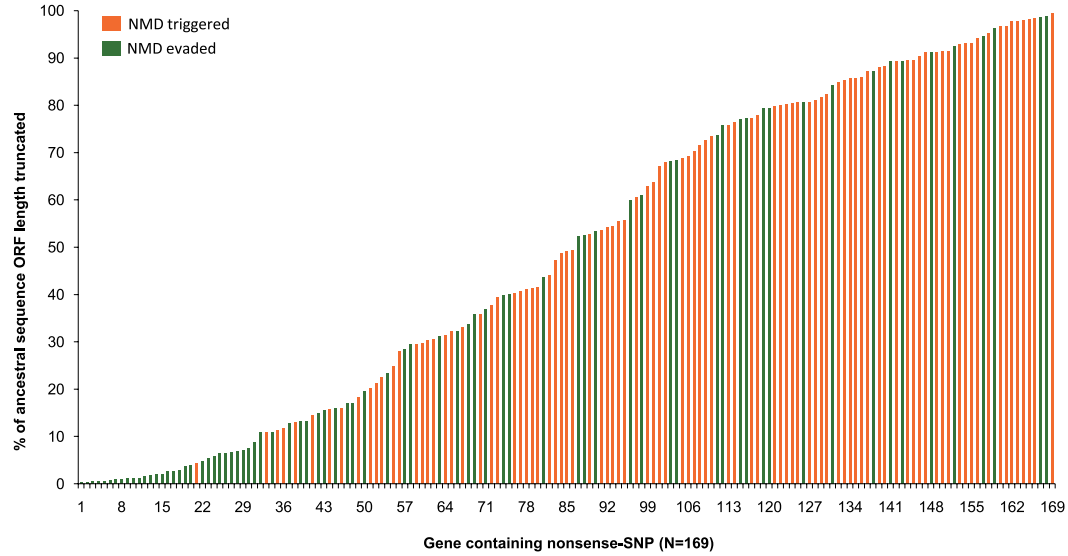

Figure 2. Even Distribution of Truncations
Truncations were calculated as the percentage of the ancestral ORF length. The 169 nonsense SNPs were sorted along the *x* axis according to the amount of peptide truncation, starting at 1 for the lowest truncation and ending at 169 for the highest truncation. The identifying number of the SNP displayed in the figure can be found in Table S4. Orange labels transcripts where NMD is predicted to be triggered with the complete loss of the gene product, whereas green refers to transcripts where NMD is evaded because the nonsense SNP is located either in the last exon or less than 50 nucleotides upstream of the last exon-exon boundary.

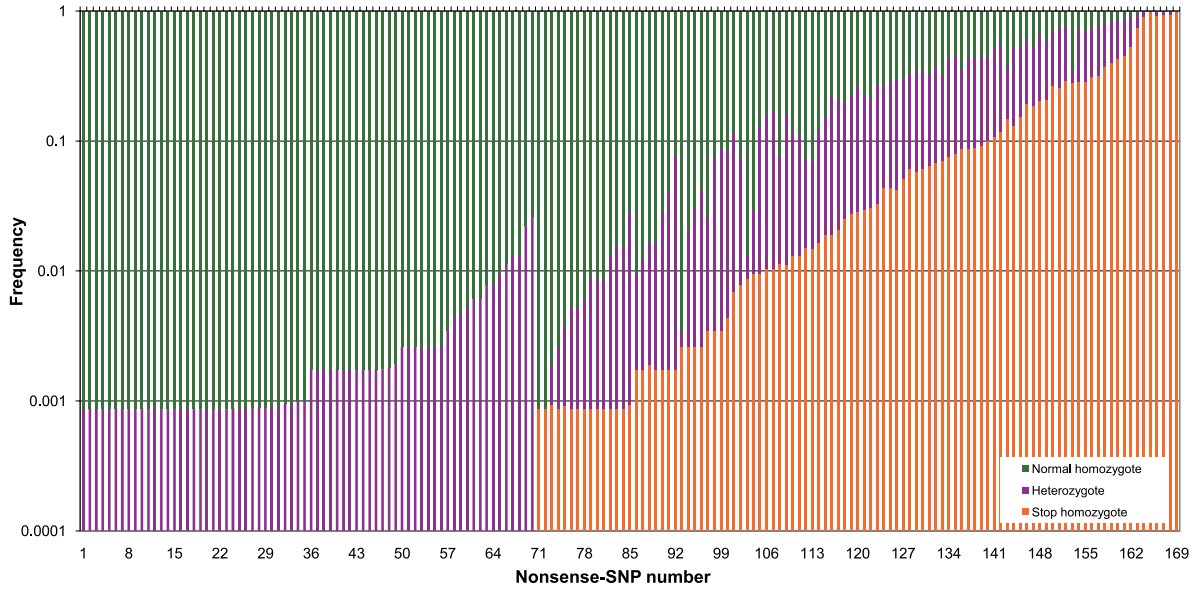

Figure 3. Frequencies of Stop Homozygotes, Normal Homozygotes, and Heterozygotes for Each Variable Nonsense SNP
The genotype frequencies of normal homozygotes (green), heterozygotes (purple), and stop homozygotes (orange) were plotted on a logarithmic scale. The nonsense SNPs were sorted along the *x* axis according to the frequency of stop homozygotes. The identifying number can be found in Table S4.

SNPs, we did not find individuals homozygous for the stop allele, but for the other five SNPs we did, and for two SNPs (in the *NPPA* (MIM *108780) and *FMO2* genes), individuals homozygous for the stop allele were found at a high frequency. It therefore appears that very few of the nonsense SNPs represent low-frequency disease-causing alleles.

Gene-Ontology Enrichment Analysis

To further investigate the functional and physiological consequences of these nonsense SNPs as a class, we used GO information to determine whether there was enrichment of any molecular function or biological process terms in these "lost" genes. The GO analysis revealed an excess of genes involved in olfactory reception and the nervous system (Table S5). The first category was expected to show up because previous studies have indicated that humans have a reduced number of active olfactory receptor genes.[44,45] Indeed, a recent study on nonprocessed pseudogenes inactivated in the human lineage reported an overrepresentation of genes involved in chemoreception (to which olfactory receptors belong) and immune response.[46] The latter, however, was not observed in our study. Finding an overrepresentation of genes involved in the nervous system was, however, unexpected because such genes have generally been shown to be very conserved.[47]

Considering the disruptive effects of nonsense SNPs, is it possible that the overrepresentation of certain GO categories largely reflects a higher number of paralogs for genes containing nonsense SNPs? If this were true, it might result from the paralogs' serving as a "backup system" for the disrupted genes and thus reducing the negative selection pressure on them. We noted that 51% of the nonsense SNP genes have at least one paralog, whereas in comparison only 35% of all human genes in Ensembl (release 50) are reported to have a paralog. This difference was found to be moderately significant (Fisher's exact test, $p < 0.05$), so it is possible that their function is "backed up" by duplicated paralogs in the human genome. However, it has been demonstrated previously, for example with the *ACTN3* gene,[12] that although a closely related gene can compensate for the function of a lost gene, the gene loss can still have significant physiological consequences.

Selective Forces on Nonsense SNPs

Although the nonsense SNPs investigated here are not overtly associated with disease, we wished to test whether they were, as a class, nevertheless mildly deleterious. Slightly deleterious alleles are subject to weak negative selection and consequently are expected show a different derived allele frequency (DAF) spectrum with an enrichment of rare derived alleles, as shown in a comparison between nonsynonymous and synonymous SNPs.[48] We therefore compared the DAF spectrum of the nonsense SNPs with that of synonymous SNPs in the same samples (Figure S2). The derived stop allele of nonsense SNPs was indeed found to be generally rarer than the derived allele of synonymous SNPs (Kolmogorov-Smirnov test, $p << 0.001$). This suggests that, as expected, weak negative selection is acting on stop alleles as a class to remove variants that are harmful over an evolutionary timescale. Indeed, we estimated the selection coefficient(s) to be ~0.0016 (see Material and Methods section), indicating that the stop alleles have on average only a slight decrease in fitness when they are compared to the normal alleles. This is lower than the value of 0.025 estimated for nonsense SNPs by Gorlov et al.[49] but similar to a value in the range of 10^{-3},

Table 1. Summary of Outlier Nonsense SNPs

Gene Symbol (MIM ID)	Gene Description	SNP	Chromosome	Position (B36)	Percent Truncated	NMD candidate	DAF	F_{ST}[a]	Heterozygosity[b]	Outlier Signal
APOL3 (MIM *607253)	apolipoprotein	rs11089781	22	34886714	85.61	YES	0.022	0.258	0.043	F_{ST}
C1orf105	open reading frame	rs7532205	1	170688829	91.38	YES	0.045	0.265	0.086	F_{ST}
CASP12 (MIM *608633)	caspase	rs497116	11	104268327	63.66	YES	0.962	0.241	0.074	DAF, F_{ST}
CD36 (MIM *173510)	thrombospondin receptor	rs3211938	7	80138385	31.29	YES	0.017	0.242	0.032	REHH, F_{ST}
FMO2 (MIM *603955)	flavin-containing monooxygenase	rs6661174	1	169444714	11.78	YES	0.959	0.284	0.079	DAF, F_{ST}
HPS4 (MIM *606682)	Hermansky-Pudlak syndrome	rs3747129	22	25192041	53.50	YES	0.202	0.097	0.323	REHH
KIAA0748	KIAA0748	rs1801876	12	53630291	3.80	NO	0.364	0.240	0.463	F_{ST}[c]
LPL (MIM *609708)	lipoprotein lipase	rs328	8	19864004	0.42	NO	0.086	0.036	0.157	REHH
MAGEE2	melanoma antigen	rs1343879	X	74921254	77.10	NO	0.311	0.540	0.429	F_{ST}[c]
NPPA (MIM *108780)	natriuretic peptide precursor	rs5065	1	11828655	0.65	NO	0.848	0.145	0.259	DAF, REHH
OR1B1	olfactory receptor	rs1476860	9	124431062	39.81	NO	0.397	0.211	0.479	F_{ST}[c]
Q8N8G3_HUMAN		rs4723884	7	39615800	68.38	NO	0.225	0.225	0.349	F_{ST}
REG4 (MIM *609846)	regenerating islet-derived	rs1052972	1	120138308	8.80	NO	0.490	0.211	0.500	F_{ST}[c]
SEMA4C (MIM 604462)	semaphorin	rs12471298	2	96890515	16.91	NO	0.043	0.469	0.082	F_{ST}[c]
SIGLEC12	sialic acid binding Ig-like lectin	rs16982743	19	56696715	95.13	YES	0.198	0.221	0.317	F_{ST}
ZAN (MIM *602372)	zonadhesin	rs2293766	7	100209294	33.04	YES	0.261	0.399	0.386	F_{ST}[c]

Columns: the official gene symbol with the MIM ID (when available), gene description, SNP ID (rs number), chromosome, and position (in build 36), percent of the peptide truncation, whether or not the SNPs are predicted to trigger NMD (YES or NO), derived allele frequency (DAF), F_{ST}, level of heterozygosity, and outlier signal. The outlier signal is identified as: $F_{ST} > 0.2$, DAF > 0.8 (one example), and REHH above the 95th percentile of the control distribution.

[a] Calculated according to Weir and Cockerham[26] across the 37 populations.

[b] Calculated according to Nei.[28]

[c] F_{ST} value is significant because it is above the 99th percentile of the empirical distribution.

calculated against deleterious heterozygous SNPs segregating in the human population.[50] Because we are using a subset of the total nonsense SNPs in the human genome and the average number of nonsense SNPs is actually likely to be higher, our estimate of 0.0016 is an upper limit.

In contrast to this general trend, a few nonsense SNPs displayed a high DAF, and these include SNPs in the *CASP12*, *FMO2*, and *NPPA* genes, with DAFs at 0.962, 0.959, and 0.848, respectively (Table 1). An excess of very high-frequency derived variants has previously been observed in the normalized site-frequency spectrum and can potentially be explained by ancestral misspecification.[48] Although the ancestral state of the *CASP12* nonsense SNP is well established[9,46], this potential confounding factor might be relevant for other genes. Among the additional genes, *FMO2* codes for the precursor of atrial natriuetic peptide, and the nonsense-SNP-carrying form has been shown to be catalytically inactive.[51] Previous studies have further revealed that the derived stop allele in *FMO2* is fixed in European and Asian populations, whereas the ancestral active allele has been found in African Americans and Hispanics[51–53]; such distributions were confirmed and extended in our data (Tables S2 and S4). If carriers of the functional allele are exposed to thioureas (which are present in a wide range of industrial, household, and medical products), they are at increased risk of pulmonary toxicity.[52] Because exposure to these chemicals is now widespread, it is interesting to consider whether they might also have been present in the pre-industrial environment and whether the stop allele might have reached its high frequency because of positive selection for protection against toxicity. In addition, the stop allele in *NPPA* has previously been reported at a high frequency in human populations and was shown to be associated with a decreased risk of stroke recurrence.[54] Stroke is a disease of old age and might not itself have exerted strong selective pressure in the past, but the association with a phenotype raises the possibility that the allele might be linked to other advantageous phenotypes as well and could thus be susceptible to positive selection. These three examples show that nonsense SNPs can be associated with phenotypes that are advantageous in some environments, and so we next investigated whether a subset of the nonsense SNPs might show the evolutionary signature of such an advantage: positive selection.

Population Differentiation

Because geographically separated populations might be subject to distinctive selective environments, selection can increase population differentiation at the selected locus. We used F_{ST}[26] as a measure of population differentiation and found that when samples were grouped into 37 populations (Figure S1B), most SNPs (both nonsense and synonymous) had low F_{ST} values within the 0.00–0.19 bin (Figure S3), as might be expected for human SNPs.[17,55–57] On average, nonsense SNPs had significantly lower F_{ST} values than synonymous SNPs (Kolmogorov-Smirnov test, $p << 0.001$). This is in accordance with a recent study[55] that showed an excess of low F_{ST} values for nonsynonymous SNPs compared to other classes, such as synonymous SNPs. Furthermore, after allowing for ascertainment bias by matching the F_{ST} values to the minor allele frequency (MAF), the authors came to the conclusion that the low values observed were a signal of purifying rather than balancing selection because the excess represented an excess of rare but not intermediate variants. To test for this in our data, we plotted the F_{ST} values of nonsense SNPs against their MAF and found no significant correlation between the two. However, we also found that the majority of low F_{ST} values are in SNPs with low MAF. We therefore suspect that the excess of low F_{ST} values observed for the nonsense SNPs here is also the consequence of purifying selection acting against mildly deleterious mutations.

In order to assess the significance of the individual F_{ST} values, we compared them to the empirical-frequency-matched distribution of values in the HGDP-CEPH panel. We found 13 nonsense SNPs with F_{ST} values above 0.2, and six of these were above the HGDP-CEPH 99th percentile (in *MAGEE2*, SEM*A4C* (MIM 604462), *ZAN* (MIM *602372), *KIAA0748*, *REG4* (MIM *609846), and *ORIB1*; Table 1), when less than two would be expected by chance. Genotyping errors can be a source of unusually high F_{ST} values[55] but are unlikely to be responsible here. We found no overall correlation between F_{ST} and heterozygosity (Figure S4), but note that several of the nonsense SNPs displaying high F_{ST} values also show outlier behavior in terms of heterozygosity (Table 1). The SNPs in SEM*A4C* and *FMO2* have high F_{ST} values but a low heterozygosity, which could indicate a recent population-restricted selective sweep. The SNPs in *MAGEE2* and *ZAN*, on the other hand, have high F_{ST} values as well as high levels of heterozygosity, which could be a sign of balancing selection or an older selective sweep. It is therefore possible that several of these genes have experienced non-neutral evolution.

Extended Haplotypes

To gain further insight into the possible action of recent natural selection, we applied the REHH test.[30] We found no evidence of unusually extended haplotypes in the nonsense SNPs as a class, which further indicates (unsurprisingly) that the majority of these SNPs are not positively selected. Outliers above the 95th percentile (Figure S5 and Table 1) include *NPPA* (again), *LPL* (MIM *609708), which encodes lipoprotein lipase and has been implicated in disorders of lipoprotein metabolism, *CD36* (MIM *173510), which is a thrombospondin receptor, and *HPS4* (MIM *606682), which encodes the Hermansky-Pudlak syndrome 4 protein. A previous study observed a significant excess of long-range haplotypes among nonsynonymous SNPs with high F_{ST} values.[55] However, only *CD36* identified here was also reported with a high F_{ST} value ($F_{ST} = 0.24$); the others had values below 0.15. It should be noted that *MAGEE2*, our highest F_{ST} outlier, was not included in the REHH analysis because it is located on the X chromosome and appropriate controls were not available.

MAGEE2: An Example of Advantageous Gene Loss?

Finally, we investigated the nonsense SNP in *MAGEE2* further by resequencing the gene and its surrounding regions and applying sequence-based tests to determine whether the evolutionary history of the region was compatible with neutrality. This SNP displayed the highest F_{ST} value of all, resulting from the presence of the stop allele (A) at very high frequency in Asian and South-American populations but its virtual absence from European and African populations (Figure 4A). The geographical distribution suggests that the derived stop allele arose before the entry of humans into the Americas ~15–20 KYA and most likely before the exit from Africa ~50 KYA. The nonsense SNP truncated the protein by ~77%, although NMD was not predicted to be triggered (Table S4).

We resequenced the gene in 91 individuals from four HapMap populations, CEU, YRI, CHB, and LWK[16,17] and one chimpanzee. Thirty-two chromosomes were found to carry the stop allele: 1 YRI, 28 CHB, 1 CEU, and 2 LWK. These proportions are similar to the worldwide geographical distribution shown in Figure 4A. A total of 43 SNPs were detected in the *MAGEE2* gene (Table S3); the haplotypes carrying the stop allele were much less diverse than the normal ones and had a nucleotide diversity (π) of 0.8×10^{-4} compared with 3.7×10^{-4} (Table 2). This led to a higher diversity in the African populations ($\pi = 4.3 \times 10^{-4}$ in YRI and 4.7×10^{-4} in LWK) than in the CEU ($\pi = 2.9 \times 10^{-4}$) and CHB ($\pi = 1.6 \times 10^{-4}$), but this is in accordance with most other comparisons of diversity within and outside Africa.[17,58,59] The lower diversity observed for the truncated version is consistent with positive selection, and to explore this possibility further we applied additional tests. Neutrality tests (Table 2), which took into account the demography of each population, revealed two significant departures from neutral expectation. Fewer haplotypes were found in the whole sample than expected, as measured by Fu's F_s.[34] In addition, Fay and Wu's H revealed an excess of high-frequency derived alleles in the CHB, the one sample where a signal would be expected if positive selection had driven the nonsense SNP to high frequency.

A median-joining network was constructed from the inferred haplotypes (Figure 4B). As was seen in the geographical distribution of the nonsense SNP (Figure 4A),

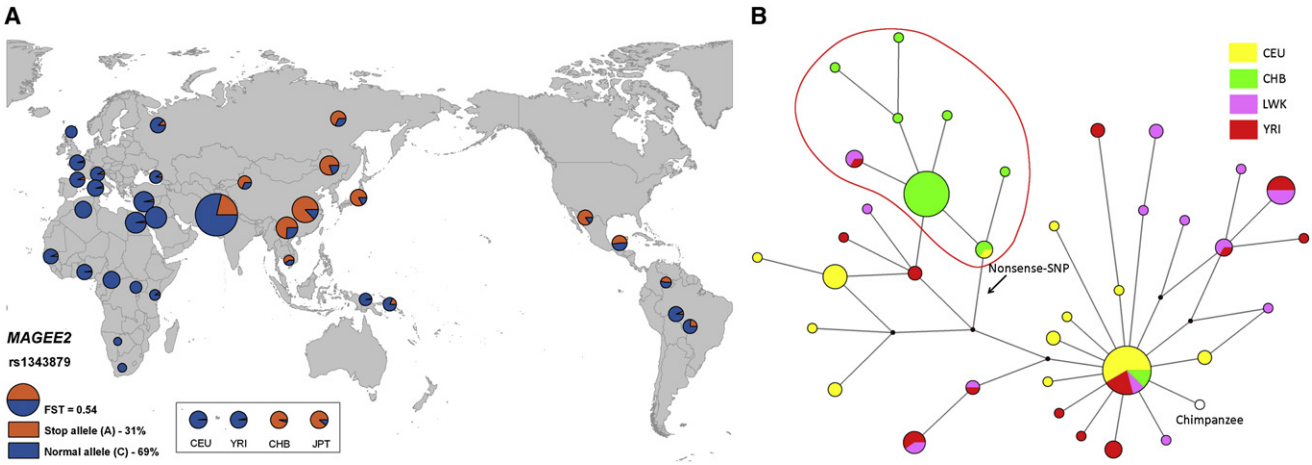

Figure 4. ***MAGEE2***

(A) Geographical distribution of stop (orange) and normal (blue) alleles in *MAGEE2*. HapMap populations are displayed separately because they do not all have precise geographic locations. Pie areas are proportional to sample sizes.

(B) Median-joining network of inferred *MAGEE2* haplotypes. Circle areas are proportional to the haplotype frequency and are color coded according to population: CEU in yellow, CHB in green, LWK in pink, and YRI in red. Lines represent mutational steps between them (one or two steps, according to length). An arrow shows the location of a nonsense SNP (rs1343879).

there is a clear east-west division for the haplotypes, reflecting the presence or absence of the nonsense SNP. All haplotypes carrying the inactive form cluster together (inside red circle in Figure 4B) such that there is one high-frequency haplotype with the other nonsense-allele haplotypes only one or two steps away. This pattern helps to explain the significantly negative value of Fay and Wu's *H* in the CHB sample by illustrating the moderately high frequency of a derived haplotype cluster specific to the CHB. The TMRCA was estimated at 69 ± 31 KY (Network) or 41 ± 6 KY (GENETREE), consistent with what would be expected on the basis of the geographical distribution.

Discussion

The analyses described here have identified the general characteristics of the class of human nonsense SNPs and have also pinpointed a small number of nonsense SNPs that appeared to be exceptional. Previous studies[14,15,46] have been largely restricted to in silico investigations but have revealed an abundance of nonsense SNPs in the human genome and the potential impact of gene loss on the human lineage after the split from the chimpanzee[46]. As a consequence of the accumulation of nonsense SNPs, functions such as chemoreception and immune response display species-specific features in humans. The current investigation focused on mutations that are still segregating in the human population and reveals that nonsense SNPs are surprisingly prevalent in the general human population, in contrast to previous reports that such SNPs are infrequent in the human genome.[60] Although our estimate is a lower bound, we found that the sampled individuals differ, on average, by 24 genes, or more than 0.05% of their gene number, because of these nonsense SNPs. Only three out of the 169 confirmed variable nonsense SNPs showed the pattern expected in the healthy population

Table 2. Summary Statistics for *MAGEE2*

	Sample Characteristics			Allele-Frequency Distribution Tests						Haplotype Test
Sample	Sample Size (chromosomes)	Number of Polymorphic Sites	Nucleotide Diversity (π) ($\times 10^4$)	Tajima's *D*	Fu & Li's *D*	Fu & Li's *D**	Fu & Li's *F*	Fu & Li's *F**	Fay & Wu's *H*	Fu's *Fs*
Worldwide[b]	111	43	4.2	−1.24	−2.20	−2.28	−2.16	−2.23	0.42	−27.03[a]
YRI	26	22	4.3	−0.49	−0.07	0.00	−0.26	−0.18	3.10	−4.25
LWK	21	21	4.7	−0.24	0.26	0.29	0.12	0.16	2.93	−4.15
CEU	33	17	2.9	−0.68	−1.58	−1.35	−1.54	−1.34	1.05	−2.36
CHB	31	11	1.6	−1.10	−0.14	−0.58	−0.54	−0.87	−8.32[a]	−3.54
Active (all[b])	79	36	3.7							
Inactive (all[b])	32	8	0.9							
Inactive (CHB)	28	7	0.8							

[a] $p < 0.01$ (one-sided tests, simulated distribution from the best-fit model).

[b] All samples (YRI, LWK, CEU, and CHB).

for known recessive-disease-causing alleles; namely, this pattern is being listed in the HGMD and being present in our samples only as heterozygotes. The remaining 67 nonsense SNPs that were found only as heterozygotes could represent novel recessive-disease alleles, or they could simply represent variants found at low frequency by chance.

Ninety-nine nonsense SNPs were found in our population samples in the homozygous (or hemizygous) state. The samples are from anonymous individuals with no phenotype information beyond sex and ethnicity, but the ethical considerations guiding the sampling required the donors to be adults competent to provide informed consent, and so it is likely that the donors were overtly healthy at the time of sampling. Truncation or loss of these 99 genes is therefore compatible with normal adult life and cannot be strongly disadvantageous. Confirmation of this expectation is found in the presence of 18 of the 169 nonsense SNPs in the Venter genome[61]; all 18 of these were present in the homozygous state in HapMap or HGDP-CEPH individuals. Nevertheless, population-genetic tests suggested that nonsense SNPs are generally mildly deleterious and subject to weak negative selection (s ~0.0016), which is reflected in the fact that their frequencies and levels of population differentiation are lower than those of synonymous SNPs.

One additional factor to consider is whether a significant proportion of the genes harboring nonsense SNPs might in fact be pseudogenes already inactivated by regulatory, missense, or other mutations. Known (Vega) pseudogenes were excluded from our study, and the genes examined in more detail showed evidence of an active form, so the proportion of pseudogenes seems likely to be low.

Direct insights into the phenotypic consequences of nonsense SNPs could potentially be obtained by future detailed studies of individuals of known genotype and phenotype, from the inclusion of these SNPs in association surveys, or from model organisms. Indirect insights come from the patterns of variation in the population; such patterns point to possible advantages associated with the loss of individual genes such as *MAGEE2*, *NPPA*, *FMO2*, *LPL*, and *HPS4*.

We chose to resequence the *MAGEE2* gene to provide more detailed insight into its evolutionary history. This gene displayed limited but significant evidence for a departure from neutrality and thus for positive selection favoring the truncated version in the CHB; it most likely originated shortly before the expansion out of Africa but had a selective advantage restricted to East Asia and the Americas. The *MAGEE2* gene is a melanoma-associated antigen that belongs to a family of *MAGE* genes found predominantly on the X chromosome. Several members of the *MAGE* gene family (including *MAGEE2*) are expressed in tumor cells but are silent in normal adult tissues except in the male germ line, leading to an alternative name for these genes, cancer-testis genes. Because of their specific expression on tumor cells, these antigens are potential targets for cancer immunotherapy[62,63], but their normal function is completely unknown and merits further investigation. Other genes of particular evolutionary interest include *SIGLEC12*, a member of a family of sialic acid-binding genes showing rapid evolutionary change, including the deletion of *SIGLEC13*, in humans.[64]

To conclude, we see the set of nonsense SNPs documented here as being particularly significant for three areas of genetics and medicine. First, sequencing is starting to be used to survey genes or genomes for disease-associated variants and to inform genetic-risk counseling, including whole-genome resequencing for personalized medicine.[61] Nonsense SNPs discovered in such studies would merit particular attention, but at least the 99 found here in the homozygous state are not associated with mendelian disorders, have no overt influence on the phenotype, and are compatible with healthy life. Second, there are nevertheless some situations in which generally neutral differences in gene content have medical consequences: for example, in allogeneic hematopoietic stem cell transplantation, a donor lacking a gene can mount an immune reaction against the tissues of a recipient with that gene, leading to graft-versus-host disease.[65] Donors and recipients should be screened for potential gene differences, including those resulting from these nonsense SNPs. Third, a general treatment for a wide variety of genetic disorders caused by nonsense SNPs has been proposed: administration of the drug PTC124, which promotes read-through of premature but not normal termination codons.[66] Such treatment would, if effective, also promote the expression of endogenous nontarget genes carrying nonsense SNPs, and the consequences of this should be evaluated. We need to understand the full extent of human genetic variation in order to reap the full benefits of present and future medicine.

Supplemental Data

Supplemental data include five tables and five figures and can be found with this article online at http://www.ajhg.org/.

Acknowledgments

We thank all the sample donors for making this work possible, Howard M. Cann for providing the HGDP-CEPH DNA panel, John Burton, Alison J Coffey, Sanjeev Bhaskar, and Jonathan Bailey at the Sanger Large-Scale Sequencing Pipeline Group for generating the sequence data, William C. Amos, Alex Bateman, and Matthew E. Hurles for input throughout the study, Alex Kondrashov for his suggestions about calculating the selection coefficient, Matthew E. Hurles for the template used to prepare Figure 1, and an anonymous reviewer for helpful comments. This work was supported by The Wellcome Trust.

Received: November 28, 2008
Revised: January 10, 2009
Accepted: January 14, 2009
Published online: February 5, 2009

Web Resources

The URLs for data presented herein are as follows:

DAVID chart analysis tool, http://david.abcc.ncifcrf.gov/summary.jsp
Ensembl Genome Browser, http://www.ensembl.org/index.html
Human Gene Mutation Database, www.hgmd.org
Network, http://www.fluxus-engineering.com/sharenet.htm
NCBI Blastn and Tblastx, http://blast.ncbi.nlm.nih.gov/Blast.cgi
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim
SNP2NMD, http://bioportal.kobic.re.kr/SNP2NMD
Sweep, http://www.broad.mit.edu/mpg/sweep/download.html
Stanford University HGDP-CEPH SNP Genotyping Data, ftp://ftp.cephb.fr/hgdp_supp1/
UCSC Genome Browser, http://genome.ucsc.edu/cgi-bin/hgTables

References

1. Ohno, S. (1970). Evolution by Gene Duplication (Berlin: Springer).
2. Olson, M.V. (1999). When less is more: Gene loss as an engine of evolutionary change. Am. J. Hum. Genet. *64*, 18–23.
3. Neel, J.V. (1962). Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"? Am. J. Hum. Genet. *14*, 353–362.
4. Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat. Rev. Genet. *3*, 285–298.
5. Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat. Rev. Mol. Cell Biol. *5*, 89–99.
6. Nagy, E., and Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. Trends Biochem. Sci. *23*, 198–199.
7. Mort, M., Ivanov, D., Cooper, D.N., and Chuzhanova, N.A. (2008). A meta-analysis of nonsense mutations causing human genetic disease. Hum. Mutat. *29*, 1037–1047.
8. Frischmeyer, P.A., and Dietz, H.C. (1999). Nonsense-mediated mRNA decay in health and disease. Hum. Mol. Genet. *8*, 1893–1900.
9. Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., Chen, Y., Stalker, J., Huckle, E., 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.
10. MacArthur, D.G., and North, K.N. (2004). A gene for speed? The evolution and function of alpha-actinin-3. Bioessays *26*, 786–795.
11. Saleh, M., Vaillancourt, J.P., Graham, R.K., Huyck, M., Srinivasula, S.M., Alnemri, E.S., Steinberg, M.H., Nolan, V., Baldwin, C.T., Hotchkiss, R.S., et al. (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature *429*, 75–79.
12. Yang, N., MacArthur, D.G., Gulbin, J.P., Hahn, A.G., Beggs, A.H., Easteal, S., and North, K. (2003). *ACTN3* genotype is associated with human elite athletic performance. Am. J. Hum. Genet. *73*, 627–631.
13. MacArthur, D.G., Seto, J.T., Raftery, J.M., Quinlan, K.G., Huttley, G.A., Hook, J.W., Lemckert, F.A., Kee, A.J., Edwards, M.R., Berman, Y., et al. (2007). Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nat. Genet. *39*, 1261–1265.
14. Savas, S., Tuzmen, S., and Ozcelik, H. (2006). Human SNPs resulting in premature stop codons and protein truncation. Hum. Genomics *2*, 274–286.
15. Yamaguchi-Kabata, Y., Shimada, M.K., Hayakawa, Y., Minoshima, S., Chakraborty, R., Gojobori, T., and Imanishi, T. (2008). Distribution and effects of nonsense polymorphisms in human genes. PLoS ONE *3*, e3393.
16. The International HapMap Consortium. (2003). The International HapMap Project. Nature *426*, 789–796.
17. The International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
18. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
19. Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. *70*, 841–847.
20. Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P., et al. (2003). Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. *68*, 69–78.
21. Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., et al. (2005). The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res. *33*, D459–D465.
22. Han, A., Kim, W.Y., and Park, S.M. (2007). SNP2NMD: A database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics *23*, 397–399.
23. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.
24. Dennis, G. Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, 3.
25. Hosack, D.A., Dennis, G. Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.
26. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.
27. Goudet, J. (2005). HIERFSTAT, a package for R to compute and test variance components and *F*-statistics. Mol. Ecol. Notes *5*, 184–186.
28. Nei, M. (1987). Molecular Evolutionary Genetics (New York: Colombia University Press).
29. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.
30. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.
31. Rozas, J., Sánchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics *19*, 2496–2497.
32. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

33. Fu, Y.-X., and Li, W.-H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.
34. Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics *147*, 915–925.
35. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.
36. Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.
37. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.
38. Stephens, M., and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. *73*, 1162–1169.
39. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
40. Bandelt, H.J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. *16*, 37–48.
41. Bahlo, M., and Griffiths, R.C. (2000). Inference from gene trees in a subdivided population. Theor. Popul. Biol. *57*, 79–95.
42. Kingman, J.F.C. (1982). The coalescent. Stochastic Process. Appl. *13*, 235–248.
43. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. *21*, 577–581.
44. Gilad, Y., Man, O., Paabo, S., and Lancet, D. (2003). Human specific loss of olfactory receptor genes. Proc. Natl. Acad. Sci. USA *100*, 3324–3327.
45. Gilad, Y., and Lancet, D. (2003). Population differences in the human functional olfactory repertoire. Mol. Biol. Evol. *20*, 307–314.
46. Wang, X., Grus, W.E., and Zhang, J. (2006). Gene losses during human origins. PLoS Biol. *4*, e52.
47. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153–1157.
48. Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C.D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA *102*, 7882–7887.
49. Gorlov, I.P., Kimmel, M., and Amos, C.I. (2006). Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Hum. Mol. Genet. *15*, 1143–1150.
50. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W. 3rd, Kondrashov, A.S., and Bork, P. (2001). Prediction of deleterious human alleles. Hum. Mol. Genet. *10*, 591–597.
51. Dolphin, C.T., Beckett, D.J., Janmohamed, A., Cullingford, T.E., Smith, R.L., Shephard, E.A., and Phillips, I.R. (1998). The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein. J. Biol. Chem. *273*, 30599–30607.
52. Veeramah, K.R., Thomas, M.G., Weale, M.E., Zeitlyn, D., Tarekegn, A., Bekele, E., Mendell, N.R., Shephard, E.A., Bradman, N., and Phillips, I.R. (2008). The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. Pharmacogenet. Genomics *18*, 877–886.
53. Krueger, S.K., Siddens, L.K., Martin, S.R., Yu, Z., Pereira, C.B., Cabacungan, E.T., Hines, R.N., Ardlie, K.G., Raucy, J.L., and Williams, D.E. (2004). Differences in FMO2*1 allelic frequency between Hispanics of Puerto Rican and Mexican descent. Drug Metab. Dispos. *32*, 1337–1340.
54. Rubattu, S., Stanzione, R., Di Angelantonio, E., Zanda, B., Evangelista, A., Tarasi, D., Gigante, B., Pirisi, A., Brunetti, E., and Volpe, M. (2004). Atrial natriuretic peptide gene polymorphisms and risk of ischemic stroke in humans. Stroke *35*, 814–818.
55. Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat. Genet. *40*, 340–345.
56. Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.
57. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., and Hill, W.G. (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Res. *15*, 1468–1476.
58. Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Curr. Biol. *15*, R159–R160.
59. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.
60. Sawyer, S.L., Berglind, L.C., and Brookes, A.J. (2003). Negligible validation rate for public domain stop-codon SNPs. Hum. Mutat. *22*, 252–254.
61. Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L., and Venter, J.C. (2008). Genetic variation in an individual human exome. PLoS Genet. *4*, e1000160.
62. Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., et al. (2005). The DNA sequence of the human X chromosome. Nature *434*, 325–337.
63. Chomez, P., De Backer, O., Bertrand, M., De Plaen, E., Boon, T., and Lucas, S. (2001). An overview of the MAGE gene family with the identification of all human members of the family. Cancer Res. *61*, 5544–5551.
64. Angata, T., Margulies, E.H., Green, E.D., and Varki, A. (2004). Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. Proc. Natl. Acad. Sci. USA *101*, 13251–13256.
65. Murata, M., Warren, E.H., and Riddell, S.R. (2003). A human minor histocompatibility antigen resulting from differential expression due to a gene deletion. J. Exp. Med. *197*, 1279–1289.
66. Welch, E.M., Barton, E.R., Zhuo, J., Tomizawa, Y., Friesen, W.J., Trifillis, P., Paushkin, S., Patel, M., Trotta, C.R., Hwang, S., et al. (2007). PTC124 targets genetic disorders caused by nonsense mutations. Nature *447*, 87–91.