Supplemental Data

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

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

Figure S1. Population locations of genotyped samples
A) Geographical locations and names of the 52 HGDP-CEPH populations genotyped. The coordinates for the HGDP-CEPH populations were obtained from the CEPH website at http://www.cephb.fr/en/hgdp/diversity.php/table.php. The diameter of the orange circles is proportional to sample sizes. The HapMap populations are not shown.
B) Geographical locations of the genotyped populations as they appear in allele frequency pie charts in Figure 4A which were used in the *FST* calculations. Some related populations were clustered together to reduce population size bias, resulting in 37 populations displayed on the map. The diameter of the green circles is proportional to sample sizes. The HapMap populations(The International HapMap Consortium 2003)- CEPH Utah residents with ancestry from northern and western Europe (CEU), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT)) are inserted at the bottom of the map as they do not have geographical coordinates.

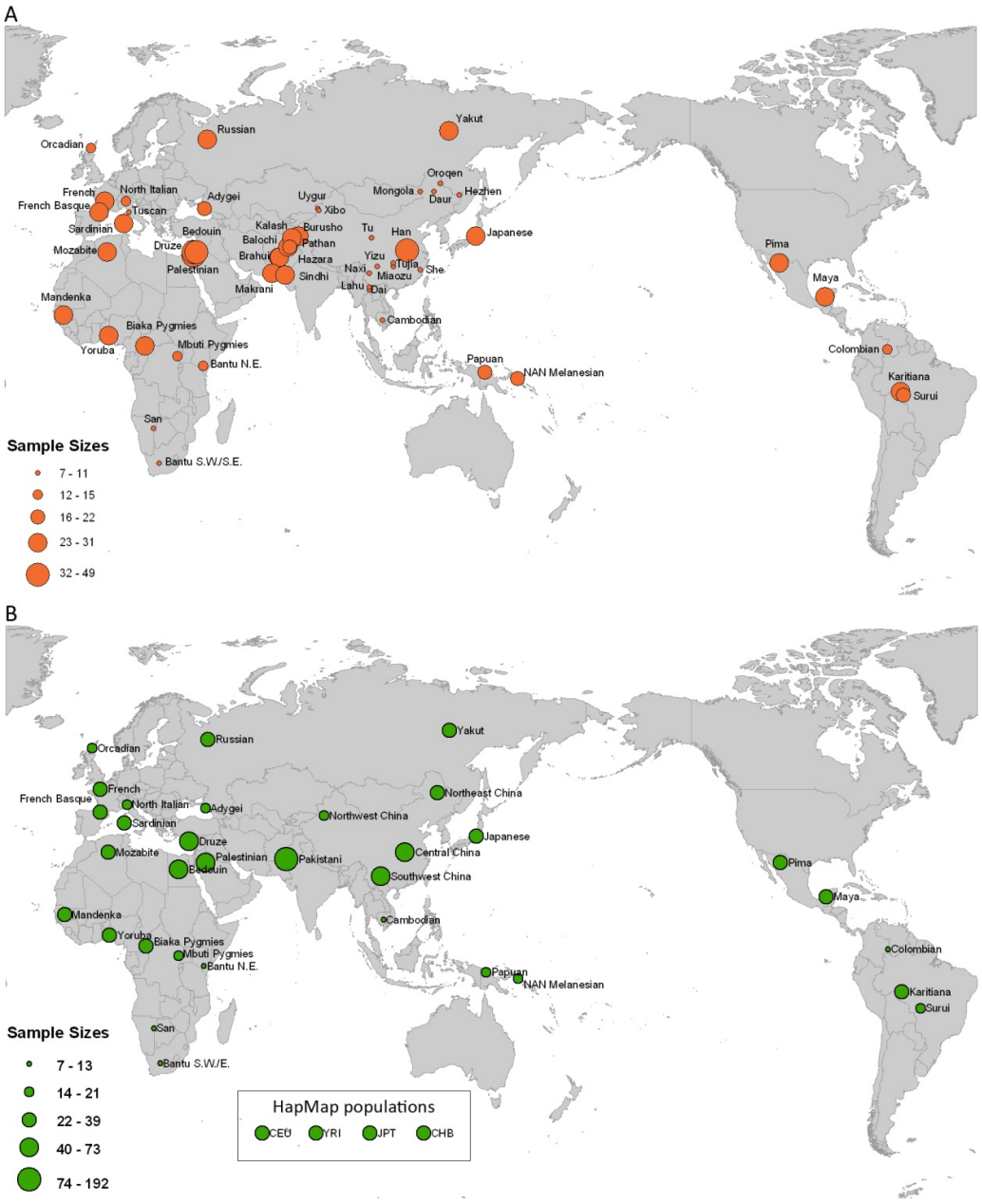
A
Orcadian
Russian
Yakut
Oroqen
Mongola
Hezhen
Daur
French
North Italian
French Basque
Tuscan
Adygei
Uygur
Xibo
Sardinian
Bedouin
Kalash
Burusho
Tu
Han
Japanese
Balochi
Pathan
Druze
Mozabite
Palestinian
Brahui
Hazara
Yizu
Naxi
Tujia
She
Miaozu
Sindhi
Lahu
Dai
Makrani
Cambodian
Pima
Maya
Mandenka
Biaka Pygmies
Yoruba
Mbuti Pygmies
Bantu N.E.
Papuan
NAN Melanesian
Colombian
Karitiana
Surui
San
Bantu S.W./S.E.
Sample Sizes
7 - 11
12 - 15
16 - 22
23 - 31
32 - 49
B
Orcadian
Russian
Yakut
French
French Basque
North Italian
Adygei
Northeast China
Northwest China
Sardinian
Druze
Japanese
Mozabite
Palestinian
Pakistani
Central China
Bedouin
Southwest China
Pima
Maya
Mandenka
Cambodian
Yoruba
Biaka Pygmies
Mbuti Pygmies
Bantu N.E.
Colombian
Papuan
NAN Melanesian
Karitiana
Surui
San
Bantu S.W./E.
Sample Sizes
7 - 13
14 - 21
22 - 39
40 - 73
74 - 192
HapMap populations
CEU
YRI
JPT
CHB

Figure S2. Comparison of the DAF spectrum of nonsense- and synonymous-SNPs in the sample of worldwide populations

The DAF was calculated for each SNP and sorted into ten bins. The DAF of nonsense-SNPs (red) was significantly lower than the DAF of the synonymous-SNPs (blue) (Kolmogorov-Smirnov, P<<0.001). The DAF spectrum was also viewed by separating the populations into five categories (according to K=5 in Rosenberg et al. 2002) (data not shown) and the distributions were found to be similar to the distribution observed in this figure.

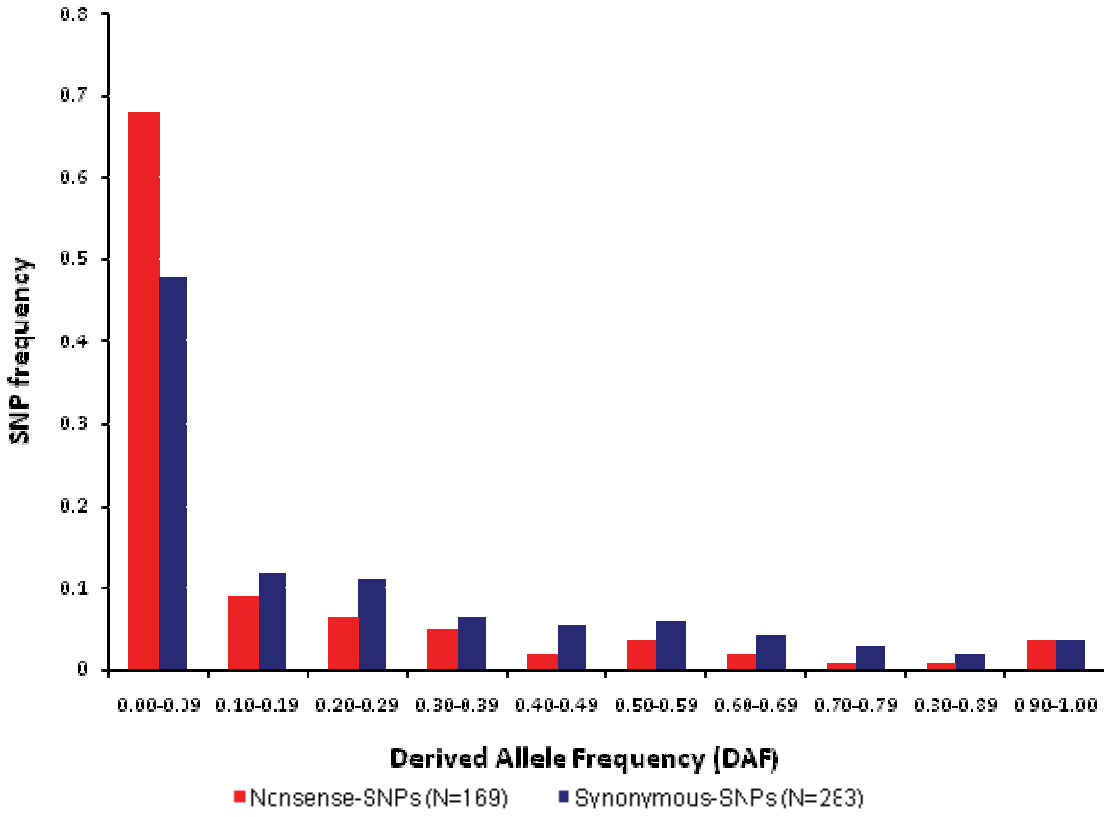

Figure S3 Comparison of *F*ST values between nonsense- and synonymous-SNPs for 37 populations. *F*ST values were calculated by conventional F-statistic methods with the HIERFSTAT(Goudet 2005) package for R using the *varcomp* function to calculate the *F*ST (theta) from Weir and Cockerham(Weir and Cockerham 1984). This F-statistic uses the allele frequencies to quantify the proportion of the total variance among the human populations. The *F*ST values were sorted into six bins and most of the SNPs (both nonsense and synonymous) fell in the lowest bin (0.00-0.19). On average, nonsense-SNPs (red) had significantly lower *F*ST values than synonymous-SNPs (blue) (Kolmogorov-Smirnov, P<<0.001) with a mean of ~0.06 and ~0.10, respectively. The highest outlier (*F*ST=0.54) was found in a nonsense-SNP (rs1343879) within the *MAGEE2* gene.

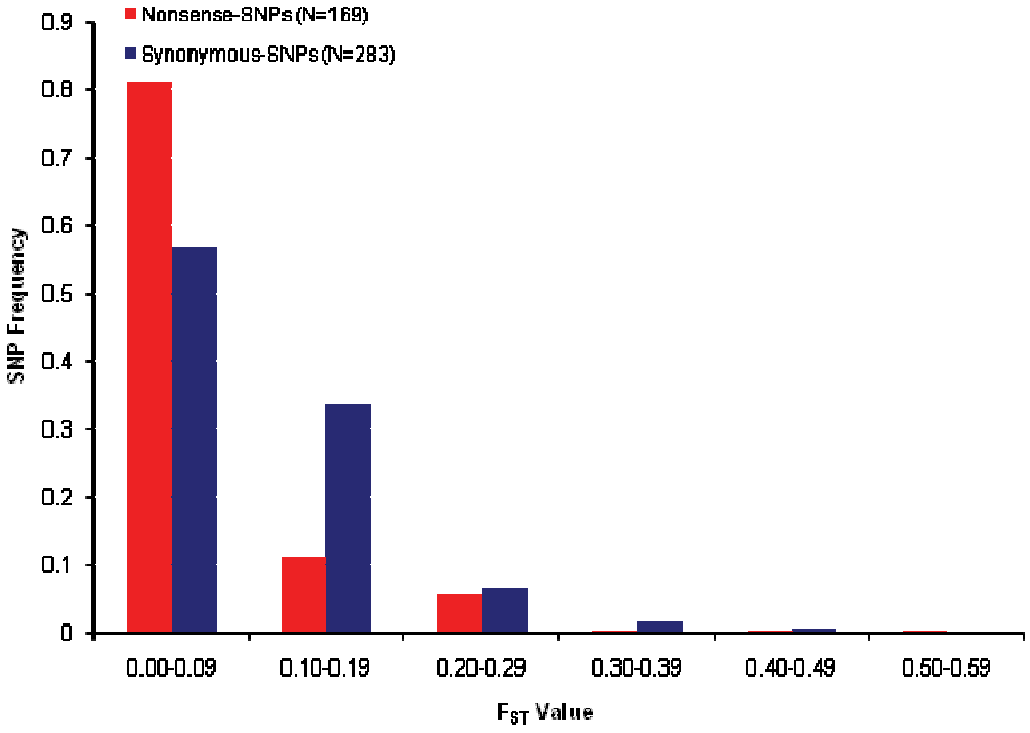

Figure S4 F_{ST} versus heterozygosity.
Synonymous-SNPs are plotted in blue and nonsense-SNPs in red and no linear correlation is observed. Some nonsense-SNP gene outliers are labelled.

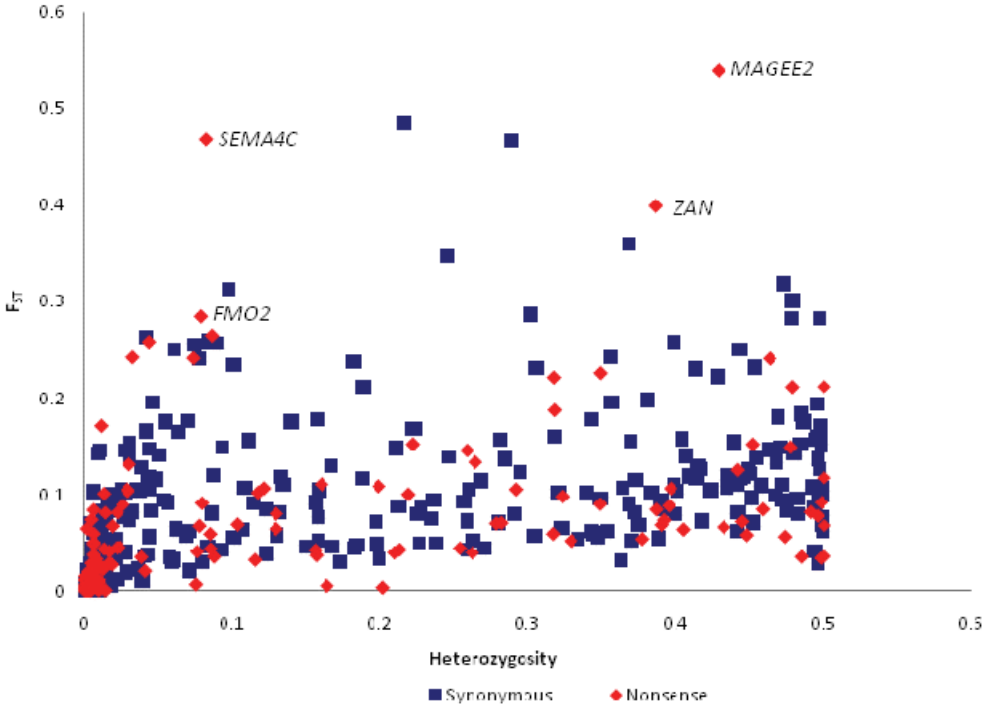

Figure S5 Relative extended haplotype homozygosity (REHH) versus frequency distribution
REHH is plotted against the frequency of each SNP in each HapMap sample **A)** CEU, **B)** JPT+CHB and **C)** YRI and some outliers are labelled**.** The grey dots represent the controls (30 ENCODE random regions) while the red dots are the stop alleles. Green and blue lines represent the 95th and 99th percentiles, respectively.

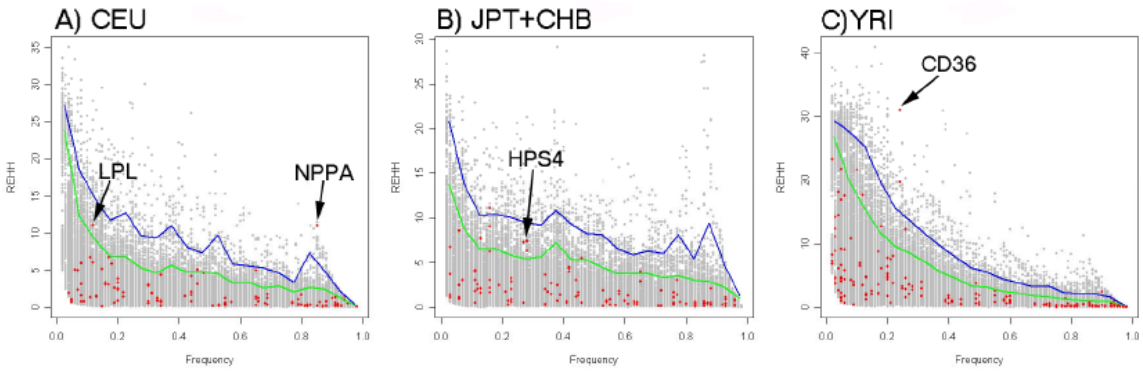

Table S5. GO terms strongly enriched (P<0.05) in the set of nonsense-SNP genes
The table displays the enriched terms associated with the list of nonsense-SNP genes, the number of genes involved in the term, the percentage (involved genes/total genes) and the P-value (modified Fisher-Exact, EASE score).

Term	Count	%	P-value
Biological Process			
GO:0007608~sensory perception of smell	12	8.11%	0.0002
GO:0007606~sensory perception of chemical stimulus	12	8.11%	0.0005
GO:0032501~multicellular organismal process	40	27.03%	0.0014
GO:0003008~system process	20	13.51%	0.0032
GO:0050877~neurological system process	16	10.81%	0.0114
GO:0019320~hexose catabolic process	4	2.70%	0.0178
GO:0046365~monosaccharide catabolic process	4	2.70%	0.0184
GO:0046164~alcohol catabolic process	4	2.70%	0.0196
GO:0007166~cell surface receptor linked signal transduction	21	14.19%	0.0202
GO:0007186~G-protein coupled receptor protein signalling pathway	15	10.14%	0.0203
GO:0007154~cell communication	38	25.68%	0.0228
GO:0009056~catabolic process	11	7.43%	0.0263
GO:0007600~sensory perception	12	8.11%	0.0265
GO:0022610~biological adhesion	11	7.43%	0.0327
GO:0007155~cell adhesion	11	7.43%	0.0327
GO:0044275~cellular carbohydrate catabolic process	4	2.70%	0.0368
GO:0006118~electron transport	8	5.41%	0.0406
GO:0016052~carbohydrate catabolic process	4	2.70%	0.0430
GO:0016337~cell-cell adhesion	6	4.05%	0.0434
GO:0007165~signal transduction	34	22.97%	0.0436
GO:0050878~regulation of body fluid levels	4	2.70%	0.0486
Molecular Function			
GO:0004984~olfactory receptor activity	12	8.11%	0.0003
GO:0030246~carbohydrate binding	10	6.76%	0.0005
GO:0004872~receptor activity	28	18.92%	0.0011
GO:0004499~flavin-containing monooxygenase activity	3	2.03%	0.0013
GO:0004888~transmembrane receptor activity	20	13.51%	0.0033
GO:0005529~sugar binding	7	4.73%	0.0037
GO:0060089~molecular transducer activity	30	20.27%	0.0045
GO:0004871~signal transducer activity	30	20.27%	0.0045
GO:0001584~rhodopsin-like receptor activity	13	8.78%	0.0126
GO:0050661~NADP binding	3	2.03%	0.0150
GO:0004930~G-protein coupled receptor activity	14	9.46%	0.0151
GO:0016709~oxidoreductase activity	3	2.03%	0.0196
GO:0046983~protein dimerization activity	7	4.73%	0.0271
GO:0042803~protein homodimerization activity	5	3.38%	0.0287

Supplemental References

Goudet J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5:184-186.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. Science 298(5602):2381-2385.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426(6968):789-796.

Weir BS, and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.