SNP frequencies in human genes
an excess of rare alleles and differing modes of selection

The origin and maintenance of DNA polymorphism in populations has been a subject of debate and conjecture for several decades. Strictly selectionist arguments suggest that polymorphisms must be maintained solely as a result of balancing selection, whereas neutral mutation theory states that polymorphism is maintained via mutation and stochastic mechanisms. The cause of polymorphisms is actually probably more complex and, although some small-scale studies for human polymorphism have been reported[1–3], no large-scale analyses have yet been done.

Human single nucleotide polymorphisms (SNPs) have recently become a focus of numerous studies. Apart from their importance for association studies that aim at the identification of variants that are responsible for complex diseases[4–6], SNPs can provide a key to understand many aspects of human evolution. Large-scale identification of SNPs can help to delineate the mutation process and also to detect and measure positive and negative selection.

By applying Expressed Sequence Tag- (EST-) derived data on human genetic variation, we were able to obtain qualitative inferences about the mutation and selection process in humans. Here, we report a large-scale analysis of human gene polymorphisms. This analysis looks at degenerate and non-degenerate sites within the coding regions, and variation in untranslated regions 3′ and 5′ of human genes. The pattern of allele frequency distribution for EST-based data reported here agrees qualitatively with smaller scale experimental studies in humans[1,2], suggesting this EST-based analysis is a good indicator of the true patterns of selection and mutation.

We show, as expected, that nucleotide diversity in non-degenerate sites is low compared with degenerate sites, which implies that selection is strong against non-synonymous substitutions in the coding region. We have also detected an excess of rare alleles in non-degenerate sites over fourfold degenerate sites and neutral theory prediction, as would be expected, owing to selective pressure. On the other hand, the 3′ untranslated region (3′UTR) of genes shows a very different pattern of mutation and selection. Although nucleotide diversity is low, there is no detectable excess of rare alleles. Because an excess of rare alleles is most likely to reflect a presence of slightly deleterious alleles, these alleles probably constitute a significant fraction of non-synonymous variation, whereas they are mostly absent in the UTRs. This suggests that the 3′UTR has two different types of sites, those that are under strong selection against mutations, and those that are effectively neutral. Reasons for these patterns are discussed below.

One method to detect selection is to estimate nucleotide diversity[7,8], π, which is the number of nucleotide differences per site between two sequences of the same locus chosen by random from a population. We estimated nucleotide diversity according to the following formula[8,9]

$$\pi = \frac{1}{L}\frac{n}{n-1}\sum_{i,j} x_i x_j \Pi_{ij} \qquad (1)$$

where L is the total number of nt sites considered; n is sample size (number of sequences); x_i and x_j are frequencies of variants i and j among n sequences; Π_{ij} is

TABLE 1. Nucleotide diversity

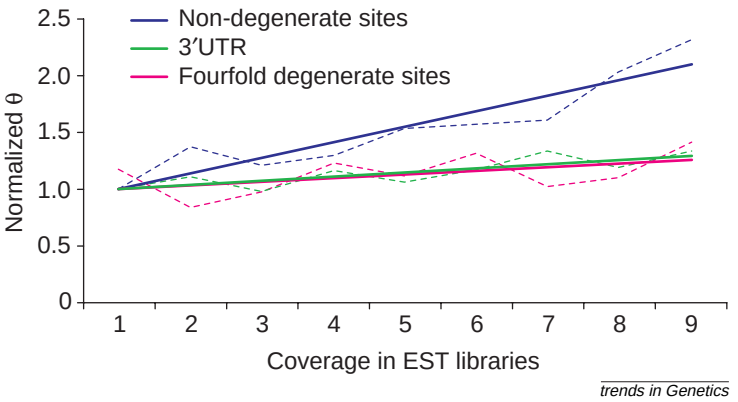
	EST data[a]	Cargill data[b]	Cargill data[b]		Halushka data[c] 'Europeans'	Halushka data 'Africans'	Halushka data All
	π[d]	θ[e]	π		θ	θ	θ
Non-degenerate sites	0.0003	0.0004	0.0003	Non-synonymous	0.0003	0.0004	0.0006
Fourfold degenerate sites	0.0009	0.0010	0.0011	Synonymous	0.0009	0.0013	0.0015
3′UTR	0.0006						0.0008
5′UTR	0.0005						0.0007
Non-coding		0.0005	0.0005		0.0005	0.0007	

[a]SNP data as found by EST analysis.
[b]The authors[1] did not provide estimates for 3′UTR and 5′UTR separately. Therefore, estimates for non-coding sites are given in the table.
[c]The authors[2] did not provide estimates for different types of sites in coding regions. Therefore, estimates for non-synonymous and synonymous substitutions are given in the table.
[d]Estimate of the nucleotide diversity per nucleotide site (heterozygosity).
[e]Watterson's estimate of the population mutation parameter.
To form a dataset of potential SNPs, we had extracted 9000 human complete mRNAs from EMBL database and then aligned to a non-redundant database of human ESTs[19,20]. Only ESTs with 99% sequence identity to mRNA sequence in a window longer than 100 nt were considered. All sequences were subjected to filtering procedure in order to eliminate possible sequencing errors (based on the analysis of fluorescent traces and sequences themselves) and mismatches arising from ESTs, which belong to paralogous genes ESTs[20]. At every position only one EST per library was considered (based on library identifiers). Since, as a rule, different EST libraries come from different individuals, every EST in our experiment corresponds to a unique allele. Nucleotide diversity estimates have been computed for non-degenerate, fourfold degenerate sites and for 3′ and 5′ untranslated regions. Due to the nature of the data, it is extremely difficult to measure precision of EST-based estimates since such estimates depend on filtering schemes and applied thresholds. No EST-based methods of SNP detection can achieve 100% accuracy. Although our estimates depend on filtering scheme and applied thresholds, qualitatively the results do not change. Comparison of our estimates with experimental estimates of smaller gene sets by Halushka *et al.*[2] and Cargill *et al.*[1] show the same trend and suggest selection in non-degenerate sites and untranslated regions.

Shamil R. Sunyaev*,§,‡ sunyaev@embl-heidelberg.de
Warren C. Lathe III*,§ lathe@embl-heidelberg.de
Vasily E. Ramensky‡ ramensky@imb.ac.ru
Peer Bork*,§ bork@embl-heidelberg.de

*European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, D-69012 Heidelberg, Germany.
§Max Delbrueck Center for Molecular Medicine (MDC), Robert-Roessle-Strasse 10, D-13122 Berlin, Germany.
‡Engelhardt Institute of Molecular Biology, Vavilova 32, 117984 Moscow, Russia.

FIGURE 1. The detection of rare alleles



Watterson's estimator θ of the mutation parameter is based on the number of segregating sites, with the assumption that the neutrality expectation of θ does not depend on sample size. In the presence of selection θ grows with sample size because of the preponderance of rare alleles. We have computed θ for regions covered with various coverage depth (counted in EST libraries) and applied the permutation test of randomness against an upward trend alternative[18] to detect possible excess of rare allele. Unlike fourfold degenerate sites and the 3′UTR sites, we detect an upward trend of θ on n dependence in the case of non-degenerate sites with significance level <0.001; although we were not able to detect an upward trend in the case of fourfold degenerate sites and 3′UTR with any reasonable significance level. Data on 5′UTR and positions covered by more than 10 EST libraries have been omitted from the analysis because of very sparse data. Linear fit of θ on n dependence normalized to the first point [predicted θ (2)] is presented in the figure to illustrate our result. Analysis of SNP databases (HGBASE and dbSNP) also showed the excess of rare alleles in non-degenerate sites comparing to fourfold degenerate sites and non-coding regions (data not shown).

the number of nucleotide differences between alleles (always one if individual SNPs are considered).

As it can be seen from Table 1, our estimates confirm previously reported experimental data of much smaller sets of genes[1,2]. In the coding regions, analysis of nucleotide diversity shows that the variation rate is the highest in fourfold degenerate sites, although it is difficult to estimate if it is close to the raw mutation rate, as data on pseudogenes (where selection is believed to be absent) are not available in our analysis. The strongest selection is observed against non-synonymous substitutions, as expected. Nucleotide diversity is also significantly lower in the noncoding regions in the vicinity of genes than in fourfold degenerate sites, suggesting selective pressure in these regions, as has been proposed[10–13].

In addition to measuring relative levels of selection through nucleotide diversity as above, an excess of rare alleles can also indicate the presence of either slightly deleterious variants under pressure of negative selection, or of recent selective sweeps[7]. We propose a novel approach to detect an excess of rare alleles and thus the presence of selection. According to the infinite-site model of the neutral evolution theory, Watterson's estimate of the mutation parameter θ should be independent on the sample size in sites under neutral mutation and drift[14]. This estimate is given by[14] the equation:

$$\theta_n = \frac{1}{L}\frac{K}{\sum_{i=1}^{n-1}\frac{1}{i}} \qquad (2)$$

where K is number of segregating sites; n is sample size (number of sequences) and L is total number of nucleotide sites considered. If a site is under selective pressure or has undergone a recent selective sweep, an excess of rare alleles would then be seen as a positive correlation with sample size. As is seen in Fig. 1, distributions of allele frequencies show two different patterns. We observe a significant excess of rare alleles (i.e. correlation of θ_n with sample size) in non-degenerate sites, in contrast with 3′UTRs (5′UTRs have been excluded from the consideration because of sparse data) and with fourfold degenerate sites, which agree with neutral mutation theoretical prediction. A further analysis of SNP databases [HGBASE: http://hgbase.cgr.ki.se/ (Ref. 15) and dbSNP: http://www.ncbi.nlm.nih.gov/SNP/ (Ref. 16)] shows a similar trend. Allele frequency distributions reported by earlier experimental studies[1,2] are also consistent with our data.

One can make an interesting observation from the different results of these tests of selection, that is, that the 3′UTR exhibits selection in the first test of nucleotide diversity, while seemingly under neutral evolution in the second test, which determines selection by detecting the excess of rare alleles. It is well known that the variation rate (nucleotide diversity) in non-degenerate sites of coding regions, and in the 3′- and 5′UTRs, is much lower than in the degenerate sites of coding regions. Our extensive data confirms this trend. This is not surprising. Selection against synonymous substitutions which have no effect on proteins is either much weaker or absent in comparison with non-synonymous substitutions and, thus, fourfold degenerate sites are effectively neutral in humans. 3′UTRs and 5′UTRs also would be under purifying selection and show a reduced amount of nucleotide variation.

It is also known that purifying selection leads not only to the reduction of variation, but also to an excess of rare alleles in the remaining variation. This is why it was expected that there would be an excess of rare alleles, both in the non-degenerate and in the 3′UTR sites in comparison with fourfold degenerate sites. Both these classes of sites exhibit lower nucleotide diversity and thus a greater selective pressure than do the degenerate sites. We have tested this effect and surprisingly find an excess of rare alleles present only in the non-degenerate sites and not in the 3′UTR.

As distinct from non-degenerate sites, the frequency of rare alleles in 3′UTR agrees with the neutral model (Fig. 1). Given the detection of selection from the low nucleotide diversity of the 3′UTR, this suggests the two regions (non-degenerate coding and 3′UTR) are undergoing different types of selective pressures. Interestingly, work by Duret *et al.*[10] in Mammalia shows that much of 3′UTR conservation is due to highly conserved blocks (defined by sequences >100 bp and >70% identity). This has been shown to be associated with regulation of mRNA stability. Further research by others in orthologous sequences[13,17] suggests that this pattern of long, conserved regions in the 3′ UTR is a general phenomenon. Lipman[12] has proposed that these blocks of conservation are due to the regulatory mechanisms of anti-sense duplexes. Anti-sense transcripts of mRNA 3′UTRs have been shown to form duplexes with these regions, which are then recognized by the mRNA-degradation systems . We propose that nucleotide substitutions in such a duplex-forming region would cause its breakup. These mutations would be negatively and strongly selected against, even in heterzygotes. The observed pattern of nucleotide diversity and frequency of rare alleles in the 3′UTR is most probably a result of the process described above. The 3′UTR probably contains two classes of sites: one neutral and the other under

BOX 1. GLOSSARY

Background selection
Based on the same process as the Hitchhiking effect (see below), the elimination of neutral (or nearly so) substitutions as the result of negative selection of deleterious substitutions at linked sites.

Balancing selection
Heterozygote advantage. The fitness of a heterozygote is greater than either possible homozygote and thus selection maintains both alleles within a population.

Hitchhiking effect – selective sweep
When selection drives an advantageous substitution at one site to fixation, neutral variation is eliminated at linked sites.

Neutral mutation
Substitutions that have no selective advantage or disadvantage.

Neutral theory
The neutral theory of molecular evolution proposes that the majority of nucleotide substitutions and polymorphisms are the result of selectively neutral mutations. It suggests that the fate of alleles are determined mainly by random genetic drift.

Non-degenerate, twofold and fourfold degenerate
A site (or position in a codon) is non-degenerate if any possible subsitutions at a position is non-synonymous. If one of three possible changes is synonymous, the site is characterized as twofold degenerate and fourfold degenerate if any substitution is synonymous.

Nucleotide diversity
The number of nucleotide differences per site between two randomly selected sequences (from a single locus) from a population.

Polymorphism
A gene locus is polymorphic if there are two or more alleles within a population.

Purifying selection
Many mutations are deleterious to the fitness of an organism. These will be selected against and eventually lost from the population. This type of selection is termed negative or 'purifying' selection.

'Slightly' deleterious alleles
An allelic variant is called slightly deleterious if it is subject to purifying selection but the selection coefficient is relatively low. The frequency of a slightly deleterious allele in a population is subject both to the stochastic fluctuations of genetic drift, depending on population size, and to a very weak negative selection. Unlike strongly deleterious alleles, which are quickly eliminated by selection, slightly deleterious alleles can be kept in a population for a long time owing to drift. Due to selective pressure, they are predominantly observed at low frequencies (in comparison to purely neutral alleles).

SNP
A single nucleotide polymorphism is a single site in a nucleotide sequence that contains two to four allelic variations within a population at relatively high frequencies (>1.0% by convention).

Synonymous and non-synonysmous
Substitutions in coding regions that result in the same amino acid are synonymous (eg. TTT-Phe to TTC-Phe). Those that result in a different amino acid are non-synonymous (eg. TTT-Phe to TTA-Leu)

3′- and 5′- UTR
Regions flanking (3' and 5') the protein-coding region of a gene that are transcribed but not translated (stands for untranslated region).

strong selection. Sites in the 3′UTR under strong selective pressure are responsible for a low level of variation, and the variation that is observed is mainly due to neutral sites. This would explain the absence of an excess of rare alleles.

This first large-scale analysis of human polymorphisms suggests a complex, yet discernible, pattern of selection and mutation in human genes. Further research on EST-based SNP data will assist us in greater quantitative and qualitative analysis of the processes of selection and mutation in human genes.

References
1 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238
2 Halushka, M.K. *et al.* (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247
3 Li, W.H. and Sadler, L.A. (1991) Low nucleotide diversity in man. *Genetics* 129, 513–523
4 Lander, E.S. (1996) The new genomics: global views of biology. *Science* 274, 536–539
5 Collins, F.S. *et al.* (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580–1581
6 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
7 Gillespie, J.H. (1998) *Population Genetics: A Concise Guide*, Johns Hopkins University Press
8 Li, W.H. (1997) *Molecular Evolution*, Sinauer Associates
9 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press
10 Duret, L. *et al.* (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365
11 Koop, B.F. (1995) Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends Genet.* 11, 367–371
12 Lipman, D.J. (1997) Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* 25, 3580–3583
13 Shabalina, S.A. and Kondrashov, A.S. (1999) Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genetic Res.* 74, 23–30
14 Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276
15 Brookes, A.J. *et al.* (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res.* 28, 356–360
16 Smigielski, E.M. *et al.* (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* 28, 352–355
17 Makalowski, W. and Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9407–9412
18 Kendall, M. and Stuart, A. (1979) *The Advanced Theory of Statistics*. Charles Griffin and Co.
19 Sunyaev, S.R. *et al.* (2000) Individual variation in protein-coding sequences of the human genome. *Adv. Protein Chem.* 54, 409–437
20 Sunyaev, S.R. *et al.* (1999) Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes. *J. Mol. Med.* 77, 754–760