Human Molecular Genetics, 2003, Vol. 12, No. 24 ***3325–3330***
DOI: 10.1093/hmg/ddg359

Impact of selection, mutation rate and genetic drift on human genetic variation

Shamil Sunyaev[1,2,*], Fyodor A. Kondrashov[3,4], Peer Bork[2,5] and Vasily Ramensky[6]

[1]Genetics Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, [2]European Molecular Biology Laboratory (EMBL), Meyerhofstr. 1, 69117 Heidelberg, Germany, [3]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA, [4]Section of Evolution and Ecology, University of California at Davis, Davis, CA 95616, USA, [5]Max-Delbrueck Center for Molecular Medicine, Robert-Roessle-Strasse 10, 13122 Berlin, Germany and [6]Engelhardt Institute of Molecular Biology, Vavilova 32, 119991 Moscow, Russia

Received August 5, 2003; Revised and Accepted October 17, 2003

The accumulation of genome-wide information on single nucleotide polymorphisms in humans provides an unprecedented opportunity to detect the evolutionary forces responsible for heterogeneity of the level of genetic variability across loci. Previous studies have shown that history of recombination events has produced long haplotype blocks in the human genome, which contribute to this heterogeneity. Other factors, however, such as natural selection or the heterogeneity of mutation rates across loci, may also lead to heterogeneity of genetic variability. We compared synonymous and non-synonymous variability within human genes with their divergence from murine orthologs. We separately analyzed the non-synonymous variants predicted to damage protein structure or function and the variants predicted to be functionally benign. The predictions were based on comparative sequence analysis and, in some cases, on the analysis of protein structure. A strong correlation between non-synonymous, benign variability and non-synonymous human–mouse divergence suggests that selection played an important role in shaping the pattern of variability in coding regions of human genes. However, the lack of correlation between deleterious variability and evolutionary divergence shows that a substantial proportion of the observed non-synonymous single-nucleotide polymorphisms reduces fitness and never reaches fixation. Evolutionary and medical implications of the impact of selection on human polymorphisms are discussed.

INTRODUCTION

Since Darwin, biologists understood the importance of heritable variability for the evolution in natural populations; more recently, it has become clear that population variability is also crucial for the study of heritable diseases. Single-nucleotide polymorphisms (SNPs) are the most frequent type of genetic variation among humans and are responsible for most of the variation in human phenotypes. SNPs have been described by Kreitman in the fruit fly *Drosophila melanogaster* (1). In a subsequent study by McDonald and Kreitman (2), SNPs of different functional categories have been related to the rate of divergence between species, in order to test for the presence of natural selection. Here, we compare the densities of various functional categories of SNPs in different human genes with the divergence between these genes and their murine orthologs. This analysis allows us to identify the evolutionary forces responsible for the observed heterogeneity in polymorphism levels across human genes.

The polymorphism rate has been estimated for many human genes (3). Some genes are highly polymorphic, whereas other genes display almost no polymorphism across the human population (4–6). Theoretically, several evolutionary forces, including random drift (coalescence), heterogeneity of mutation rates, and negative and positive natural selection, can lead to this heterogeneity in the polymorphism rate. In addition to satisfying theoretical interests, understanding the heterogeneity of the polymorphism rate is important for the optimization of the design of association studies aimed at the identification of alleles responsible for human phenotypes (7–10).

Recent data demonstrated wide alternations in the polymorphism density along the genome and postulated existence of stable haplotype blocks (11,12). Recombination hot spots have been hypothesized to contribute greatly to the heterogeneity

*To whom correspondence should be addressed. Tel: +1 6175256675; Fax: +1 6177325123; Email: ssunyaev@rics.bwh.harvard.edu

in polymorphism density along the genome (11–15). Recombination hot spots split the genome into regions with different coalescent histories, i.e. the genome can be represented as a mosaic of blocks, each consisting of sites with a common genealogy, not broken by recombination in most individuals. In selectively neutral regions of the genome, differences in the coalescent history of genes are due to random genetic drift. In protein-coding regions, however, natural selection impacts genetic variation, which causes the patterns to be more complex.

Our aim was to focus on the coding regions and study all of the human genes in public databases that had SNP information and orthologous murine sequences available, in order to: (1) analyze the impact of population history, the mutation rate, and negative (stabilizing) selection on the structure of human genetic variation, and (2) study the distribution of deleterious alleles among monomorphic and variable genes.

RESULTS

The variability of a population is shaped by an interplay of several evolutionary forces. Consequently, several not mutually exclusive explanations may account for different polymorphism rates across genes (Table 1). Mutation with heterogeneous rates, population history (coalescence) and selection may all independently affect the polymorphism rate at each individual locus. We attempted to disentangle the impacts of these factors using data on intraspecies polymorphism and interspecies divergence. The correlation of the polymorphism and evolution rates of human genes establishes the role of factors that also affect interspecies divergence, and the correlations of the polymorphism rates between different functional categories of SNPs make it possible to distinguish the effects of mutation or coalescence from that of selection.

This analysis relies on two assumptions. First, it assumes that evolutionary constraints are similar for within-population variability and between-species divergence. For example, it was shown that majority of deleterious amino acid substitutions destabilize proteins (16). Stability requirements are unlikely to differ substantially for orthologous proteins of closely related species.

The second assumption is that the mutation processes are similar for human and mouse orthologs. The mutation rate at a nucleotide site depends on a number of factors, including DNA methylation sequence contexts (17,18) and other weaker factors (19,20). The nucleotide content at a locus is not significantly different in humans and mice (21); thus, we expected that the per locus mutation rate is similar in the course of the evolution of these two species.

The mutation rate influences all functional categories of substitutions equally and affects intraspecies polymorphism and interspecies divergence in the same way (22). Therefore, if mutation rate heterogeneity were a strong force responsible for the heterogeneity of the polymorphism rate across genes, the abundance of all functional types of SNPs in a gene should be strongly correlated with the number of substitutions between this human gene and its mouse ortholog (Table 1). Also, the numbers of polymorphic sites of different functional types would be correlated with each other.

Table 1. Correlation between rates of various functional categories of SNPs in a locus with the divergence rate between species expected under different evolutionary mechanisms

	nsSNPs versus sSNPs	sSNPs versus K_S	Neutral nsSNPs versus K_A	deleterious nsSNPs versus K_A
Mutation rate differences	+	+	+	+
Gene history	+	−	−	−
Selection	−	−	+	−
Accumulation of slightly deleterious alleles	−	−	−	−

To detect the impact of mutation rate heterogeneity, we need to eliminate the effects of selection and coalescent history. The only correlation affected by heterogeneity in the mutation rate, but not by selection or coalescent history, is the correlation of neutral polymorphism and divergence rates. Thus, we analyzed the correlation of the number of synonymous substitutions K_S between human and murine genes with the number of synonymous polymorphisms. Polymorphism rate is expected to correlate with divergence in neutral sites. However, if the heterogeneity of the mutation rate among loci is within a relatively small range compared with other factors influencing the polymorphism rate, this correlation would not determine most of differences in SNP density and therefore it would not be highly significant. As shown in Figure 1, genes with high K_S show a tendency to have a higher synonymous polymorphism rate (also, a correlation coefficient of approximately 0.2 was reported recently in a related study) (23). This tendency is stronger for small values of K_S, which is probably due to the inaccuracy of the K_S values in genes with a high synonymous divergence rate. However, the dependence of number of synonymous SNPs on K_S is overall not highly significant. This suggests, in agreement with previous work (12,20,24), that mutation rate heterogeneity plays only a minor role in creating the observed heterogeneity in polymorphism rates across genes. Alternatively, fluctuations of the mutation rate along the genome are not conserved between the human and the mouse genomes.

All contemporary alleles at each locus have descended from a single allele (22). If a locus has had a long coalescent history, i.e. if the ancestral allele existed many generations ago, many sites at the locus, both functional and neutral, are expected to be polymorphic. Thus, the impact of different coalescent histories at different loci, as well as of loci-specific mutation rates, will create a correlation of densities of functional and neutral SNPs at a locus. We observe a strong and highly significant correlation between the density of synonymous and nonsynonymous SNPs in a gene (Fig. 2). Therefore, the correlation between the densities of functional and neutral SNPs cannot be explained by mutation rate heterogeneity and must be due to different coalescent histories of different loci.

Heterogeneity in the coalescent histories of loci may be caused by genetic drift and recombination, or by selection through background selection (25) or hitch-hiking (26). We assume that positive selection and, therefore, hitch-hiking is too

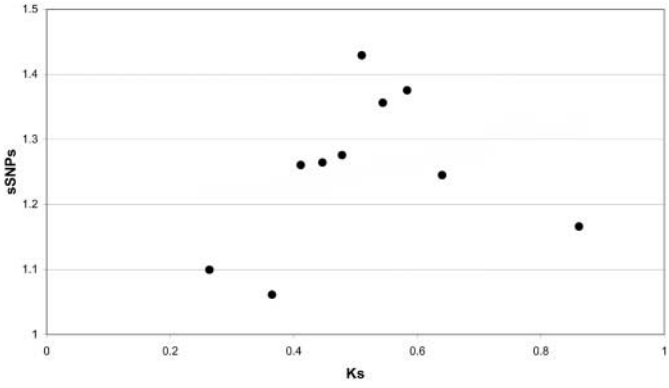

Figure 1. Dependence of the number of synonymous SNPs (sSNPs) per gene on K_S between human and murine orthologs as estimated with the Li–Pamilo–Bianchi method. The genes were binned according to K_S values so that every bin contains 250 observations. Mean number of synonymous SNPs for each bin is plotted. The dependence was not significant according to the χ^2 test (P-value of the χ^2 test is 0.088, although $P < 0.05$ if only K_S values less than 0.5 are considered). Normalization of the number of synonymous SNPs to number of synonymous sites in the gene does not change the character of the dependence. The result also does not change if only a subset of available SNPs corresponding to TSC or JSNP data is considered. The result is also robust to the choice of the method of estimation of K_S; i.e. data obtained using Li–Pamilo–Bianchi method qualitatively agree with the data obtained using Yang–Nielsen method.

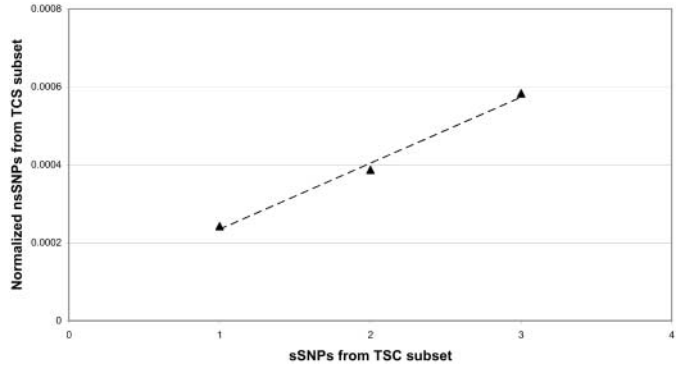

Figure 2. Correlation of the number of synonymous SNPs per gene with the density of non-synonymous SNPs. The scatter plot presents an average number of non-synonymous SNPs per gene, normalized to the total number of non-synonymous sites, versus the number of synonymous SNPs per gene. Only TSC data were used to eliminate impact of sample size. The correlation is highly significant (P-value of the χ^2 test is 8.57×10^{-5}). The contingency coefficient of 0.18 is very close to the contingency coefficient of 0.21 for the dependence of benign nsSNPs on K_A (Fig. 4).

rare to substantially alter the polymorphism rate at many loci simultaneously. Then, the cause of heterogeneous coalescent histories of human loci can be either genetic drift or differences in the strength of negative selection. If background selection has influenced polymorphism rates of modern SNPs, then genes under stronger selective constraint should have a lower density of synonymous SNP.

To determine the existence of the influence, we related the density of synonymous SNPs to the strength of negative selection, estimated as a ratio of non-synonymous to synonymous substitution rates, K_A/K_S. Correlation of the density of synonymous SNPs with K_A/K_S ratio was not observed (Fig. 3). This suggests that genetic drift, not background selection, is primarily responsible for differences in coalescent history.

The expected effect of selection on the density of SNPs is greatly dependent on their contribution to fitness. With respect to fitness, four types of changes are possible in a protein sequence: benign (neutral), slightly deleterious, strongly deleterious and beneficial. Very deleterious and beneficial SNPs are expected to be observed only at low rates in a population because very deleterious SNPs are prevented from becoming common by negative selection and beneficial substitutions achieve fixation quickly and are not generally observed in polymorphism state. Thus, the majority of the SNPs observed in a population are probably neutral or slightly deleterious.

The impact of negative selection on a gene depends on its overall contribution to fitness and on the fraction of sites (possible mutations) that are functionally important for this gene. Both the rate of non-synonymous substitution (K_A) and the number of benign SNPs are proportional to the fraction of sites that are under selective constraint, therefore, a correlation between the density of benign SNPs and K_A was expected. Any neutral variation is expected to correlate with divergence, however, if different loci show a higher heterogeneity in protein evolution rates than that in mutation rates, then they would show a more profound difference in benign amino acid variation than in synonymous variation.

Slightly deleterious mutations, those with a coefficient of selection against them below $1/4N_e$ (where N_e is the effective population size) contribute to both fixations and polymorphism (27). In contrast, substantially deleterious mutations with coefficients of selection above $1/4N_e$ almost never reach fixation. Such mutations may be present at low frequencies and, thus, contribute to polymorphism, unless they are very deleterious (with selection coefficients above, approximately $100/4N_e$). Among mutations that may affect a protein, the fraction of substantially deleterious mutations with selection coefficients within the range $1/4N_e$–$100/4N_e$ is likely to be substantial (28), because this range spans two orders of magnitude. Regardless of this fraction, we did not expect the density of substantially deleterious SNPs to correlate with K_A.

We observed a strong correlation between K_A and the density of non-synonymous SNPs predicted to be benign and only a very weak correlation between K_A and the density of SNPs predicted to be damaging for protein structure or function (Fig. 4). We predicted that some polymorphisms are damaging based on evolutionary sequence conservation and biochemical and protein structure characteristics of the substitution (see Materials and Methods). The polymorphisms that we predicted to be damaging by our routine may, nevertheless, prove not to affect fitness because what is 'damaging' from the point of view of protein structure, or even evolutionary conservation, may not have a profound effect on fitness in the modern species. However, the fact that the correlation between K_A and the density of damaging SNPs is very weak demonstrates that most of the SNPs that we predict to be damaging do have a deleterious effect on fitness strong enough to prevent their fixations (29).

Since only the density of substantially deleterious SNPs should fail to correlate with K_A, these results demonstrate that substantially deleterious alleles are present in the human

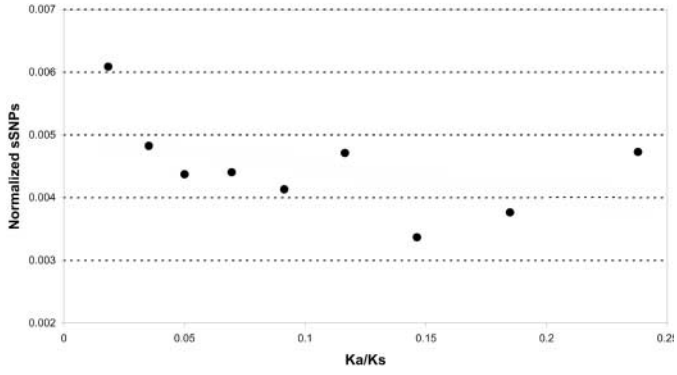

Figure 3. Number of synonymous SNPs per gene plotted versus the ratio K_A/K_S. K_A and K_S values were estimated using the Li–Pamilo–Bianchi method. The genes were binned according to K_A/K_S values so that every bin contains 250 observations. Mean number of synonymous SNPs for each bin is plotted. The impact of background selection is not detectable because conservative genes do not exhibit lower numbers of sSNPs. No significant dependence of number of sSNPs on K_A/K_S was observed (P-value of the χ^2 test is 0.073).

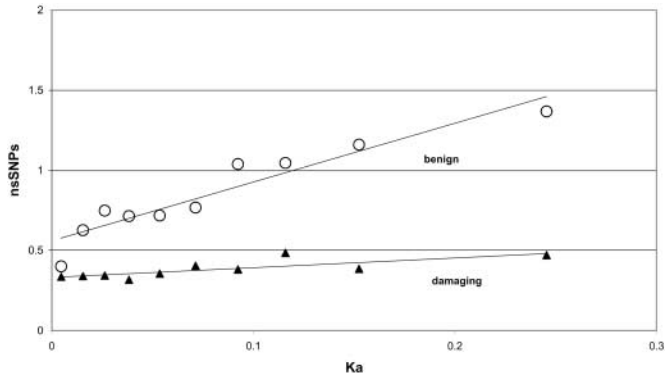

Figure 4. Dependence of the number of non-synonymous SNPs from two functional categories on K_A between human and murine orthologs. Data on SNPs predicted to be benign for protein function are shown as circles, while data on SNPs predicted to be damaging are shown as shaded triangles. The genes were grouped according to K_A values, estimated using the Li–Pamilo–Bianchi method, with 250 genes in each group. The mean number of synonymous SNPs for each group is plotted. The correlation is weak for SNPs predicted as benign (P-value of the χ^2 test is 0.046), whereas it is very strong for damaging nsSNPs (P-value of the χ^2 test is 4.21×10^{-23}). Qualitatively, the result does not depend on whether all database SNPs are considered or the analysis is limited to the particular subset of the data (TSC or JSNP data). Also, normalization of the nsSNP number to the number of non-synonymous sites in the gene does not change the character of the dependence. The application of a different model to compute K_A does not change the result.

population in large numbers; it has been suggested (16,29–31) that as many as 1/5 of all missense SNPs may be deleterious. Because the number of deleterious SNPs depends only slightly on the rate of protein evolution, estimated by interspecies sequence divergence, deleterious SNPs are evenly distributed across human genes.

DISCUSSION

Here we compared the contributions of genetic drift, mutation and selection on the distribution of human polymorphisms. Our results confirm the importance of the coalescent history, genetic drift, to the structure of human haplotypes (12). However, at least in protein coding regions, negative selection also has a profound effect on the density of polymorphisms. We found that selection lowers genetic variability by eliminating deleterious variants and the magnitude of this effect is relatively strong in comparison to genetic drift. The correlation of the non-synonymous sequence divergence to the density of damaging SNPs shows that the accumulation of deleterious SNPs also has a considerable effect on the pattern in human polymorphism such that many SNPs in the human genome appear to be substantially deleterious. While previous analyses detected high levels of deleterious SNPs in the human population (4,16,29–34), here we show that selection is strong enough to prevent their fixation.

Comparing polymorphisms and sequence divergence also has its implications for medical geneticists looking for allelic variants involved in human disorders. Sequence conservation between species has been proposed to be a useful marker to identify potentially important allelic variants. Although there is little doubt that this strategy is useful where individual amino acid sites are concerned, our study shows that the overall number of deleterious variants does not depend on the conservation of a gene as a whole (Fig. 4).

The common disease/common variant (CD/CV) hypothesis claims that genetic diseases are caused by frequent alleles (7). On the other hand, it is also possible that common multifactorial genetic diseases are mostly caused by a high number of rare alleles; this is known as the common disease/rare variant hypothesis (CD/RV) (10). From the evolutionary point of view, the CD/RV hypothesis can be explained in terms of the mutation accumulation model, which postulates that large numbers of deleterious alleles are involved in the inheritance of phenotypes. On the other hand, the CD/CV hypothesis would be most easily supported by pleiotropic models (10). Although multiple studies on specific complex phenotypes will be needed to decisively discriminate between these hypotheses, statistical analysis of allelic variation can provide indirect evidence supporting either of them. Our results support the mutation accumulation model with respect to significant number of substantially deleterious allelic variants accumulated in individual human genomes.

Both the direct and the indirect association studies assume that common diseases are caused by common SNPs (8). In addition, indirect association studies are dependent upon the existence of stable haplotype blocks in the genome and the association of common neutral polymorphisms with the disease causing variants in the haplotype blocks (8). If many deleterious SNPs contribute to the common complex phenotypes of medical interest, association studies might be ineffective (10) and novel strategies are to be sought to identify the genetic basis of complex diseases.

MATERIALS AND METHODS

Protein sequences for human and mouse genes were obtained by extracting all entries from the SWALL database (35) annotated as ‘Homo sapiens’ and ‘Mus musculus’ in the *organism* field. Orthologous pairs between these two species were identified via BLAST (36) as bidirectional best hits (37)

that spanned at least 80% of the length of the longest protein and had showed at least 60% amino acid sequence identity. This procedure yielded 11 597 orthologous pairs. For each gene in each orthologous pair, a nucleotide sequence was obtained by comparing the amino acid sequences of genes in our dataset with the coding sequences obtained from the complete human (38) and mouse (21) genomes and from complete mRNA sequences from GenBank. Only genes with greater than 99% similarity (excluding gaps longer than 20 nucleotides or probable alternative isoforms) were used to reconstruct the nucleotide sequence, resulting in 2592 gene pair alignments. For each orthologous gene pair, a nucleotide alignment was reconstructed from an amino acid alignment made with CLUSTAL (39) using default parameters.

We estimated the rate of synonymous (K_S) and non-synonymous (K_A) sequence divergence using the Li–Pamilo–Bianchi method (40,41). To demonstrate the independence of our results and the method of estimating K_S and K_A values, we repeated all computations using K_S and K_A estimates obtained using the Yang and Nielsen method as implemented in the PAML package (42,43) (data not shown).

SNPs from the HGVbase database, Release 13 (44) were mapped onto protein sequences using the *snp2prot* program (45) and were classified into three functional categories (synonymous, non-synonymous damaging and non-synonymous benign) using *PolyPhen* software (45). The mapping resulted in 5923 nsSNPs and 5856 sSNPs observed in 4332 human proteins.

To evaluate the statistical significance of the observed dependences, we grouped the genes according to both variables to form a 4×3 contingency table with equally populated bins. A contingency table χ^2 test of independence was applied to all statistical dependences considered. Since the data sample size and the number of degrees of freedom were kept constant, χ^2 values and corresponding *P*-values were directly comparable for all tests, except of the dependence shown in Figure 2. Figure 2 presents only TSC (the SNP Consortium) (46) data corresponding to the systematic study, which used a fixed population sample for SNP discovery. Several χ^2-based contingency measures corrected for the sample size dependence are common. We used the contingency coefficient

$$C = \sqrt{\frac{\chi^2}{(\chi^2 + N)}}$$

where N is the sample size.

We relied on χ^2 statistics on grouped data because it is distribution-free, i.e. does not depend on a specific hypothesis on the original distribution of the data. The results are robust with respect to the data grouping. We also repeated the analysis using correlation coefficient and ANOVA. The results were shown to be independent of the method for detection of statistical dependence. In ANOVA, genes were grouped according to the number of SNPs per gene. For the dependence of benign nsSNPs density on K_A, ANOVA resulted in a *P*-value of 4.62×10^{-27}($4.21 \times 10^{-23}$for the χ^2 test); the same test for damaging nsSNPs had a *P*-value of 0.03 (0.046 for the χ^2 test); and the ANOVA *P*-value for the dependence of sSNP rate on K_S was 0.64 (0.088 for the χ^2 test).

In order to verify that our results were not due to the comparative sequence analysis method implied, we confirmed all of the results using a set of functional predictions made on the basis of three-dimensional structure alone. To prove that the qualitative results were not seriously affected by erroneous SNPs in the database, by possible bias due to non-uniform and generally unknown sample size, all results of the analysis were repeated on subsets of the database corresponding to SNP data, annotated as 'validated' in dbSNP (47) and also on the data obtained by systematic genome-wide SNP screens in a fixed sample of individuals provided by TSC and Japanese SNP database (JSNP) (48).

The data on human–mouse sequence divergence and SNP distribution in genes are available via ftp at genetics.bwh.harvard.edu/Sunyaev/snp2div/snp2div.xls.

ACKNOWLEDGEMENTS

We are grateful to Alexey Kondrashov and Alison Wellman for the careful reading of the manuscript and providing us with their valuable comments.

REFERENCES

1. Kreitman, M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila* melanogaster. *Nature*, **304**, 412–417.
2. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
3. Sunyaev, S., Hanke, J., Brett, D., Aydin, A., Zastrow, I., Lathe, W. III, Bork, P. and Reich, J. (2000) Individual variation in protein-coding sequences of human genome. *Adv. Protein Chem.*, **54**, 409–437.
4. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.
5. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **22**, 239–247.
6. Cambien, F., Poirier, O., Nicaud, V., Herrmann, S.M., Mallet, C., Ricard, S., Behague, I., Hallet, V., Blanc, H., Loukaci, V. *et al.* (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.*, **65**, 183–191.
7. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
8. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
9. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell H.J., Eaves, I.A., Dudbridge, F., *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
10. Wright, A., Charlesworth, B., Rudan, I., Carothers, A. and Campbell, H. (2003) A polygenic basis for late-onset disease. *Trends Genet.*, **19**, 97–106.
11. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A. and Faggart, M. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
12. Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S. and Altshuler, D. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.*, **32**, 135–142.
13. Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.
14. Kauppi, L., Sajantila, A. and Jeffreys, A.J. (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.*, **12**, 33–40.

15. Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.*, **71**, 227–234.
16. Wang, Z. and Moult J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.
17. Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
18. Kondrashov, A.S. (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.*, **21**, 12–27.
19. Lercher, M.J. and Hurst, L.D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.*, **18**, 337–340.
20. Li, W.-H., Yi, S. and Makova, K. (2002) Male-driven evolution. *Curr. Opin. Genet. Dev.*, **12**, 650–656.
21. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
22. Li, W.-H. (1997) *Molecular Evolution.* Sinauer Associates, Sunderland, MA.
23. Hardison, R.C. *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, **13**, 13–26.
24. Nachman, M.W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.*, **17**, 481–485.
25. Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.
26. Maynard-Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favorable gene. *Genet. Res.*, **23**, 23–35.
27. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.
28. Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Ann. Rev. Ecol. Syst.*, **23**, 263–286.
29. 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.
30. Fay, J.C., Wyckoff, G.J. and Wu, C.I. (2001) Positive and negative selection on the human genome. *Genetics*, **158**, 1227–1234.
31. Chasman, D. and Adams, R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
32. Sunyaev, S., Lathe, W. IIIrd, Ramensky, V. and Bork, P. (2000) SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.*, **16**, 335–337.
33. Sunyaev, S., Ramensky, V. and Bork, P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, **16**, 198–200.
34. 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.
35. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.*, **31**, 365–370.
36. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.
37. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
38. International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
39. Thompson, J.D., Higgins, D.G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, **22**, 4673–4680.
40. Pamilo, P. and Bianchi, N.O. (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.
41. Li, W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.
42. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
43. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
44. Fredman, D., Siegfried, M., Yuan , Y.P., Bork, P., Lehväslaiho, H. and Brookes, A.J. (2002) HGVbase: A human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucl. Acids Res.*, **30**, 387–391.
45. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucl. Acids Res.*, **30**, 3894–3900.
46. Holden, A.L. (2002) The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*, **22-4** (suppl.), 26.
47. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.*, **29**, 308–311.
48. Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. and Nakamura, Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucl. Acids Res.*, **30**, 158–162.