Research

The amino-acid mutational spectrum of human genetic disease

Dennis Vitkup*, Chris Sander†‡ and George M Church*

Addresses: *Lipper Center for Computational Genetics and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. †Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA. ‡Current address: Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA.

Correspondence: George M Church. E-mail: for_email_look@arep.med.harvard.edu

Published: 30 October 2003

Genome ***Biology*** 2003, **4:**R72

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/R72

Received: 3 July 2003
Revised: 24 September 2003
Accepted: 30 September 2003

Abstract

Background: Nonsynonymous mutations in the coding regions of human genes are responsible for phenotypic differences between humans and for susceptibility to genetic disease. Computational methods were recently used to predict deleterious effects of nonsynonymous human mutations and polymorphisms. Here we focus on understanding the amino-acid mutation spectrum of human genetic disease. We compare the disease spectrum to the spectra of mutual amino-acid mutation frequencies, non-disease polymorphisms in human genes, and substitutions fixed between species.

Results: We find that the disease spectrum correlates well with the amino-acid mutation frequencies based on the genetic code. Normalized by the mutation frequencies, the spectrum can be rationalized in terms of chemical similarities between amino acids. The disease spectrum is almost identical for membrane and non-membrane proteins. Mutations at arginine and glycine residues are together responsible for about 30% of genetic diseases, whereas random mutations at tryptophan and cysteine have the highest probability of causing disease.

Conclusions: The overall disease spectrum mainly reflects the mutability of the genetic code. We corroborate earlier results that the probability of a nonsynonymous mutation causing a genetic disease increases monotonically with an increase in the degree of evolutionary conservation of the mutation site and a decrease in the solvent-accessibility of the site; opposite trends are observed for non-disease polymorphisms. We estimate that the rate of nonsynonymous mutations with a negative impact on human health is less than one per diploid genome per generation.

Background

Several recent studies [1-6] have applied computational methods to predict potentially deleterious effects of nonsynonymous single-nucleotide polymorphisms (SNPs) in humans. SNPs represent common human alleles, usually with population frequencies greater than 1%. Both structural and evolutionary methods were used to assess potential functional effects of SNPs. It was predicted that a substantial fraction (10-30%) of human SNPs may affect protein function negatively, although the medical consequences of these SNPs remain to be established.

The main goal of the work reported here is to characterize and rationalize the overall amino-acid spectrum of disease mutations and non-disease SNPs (referred to as 'benign SNPs' below). We obtain the relative probabilities that a random

mutation (rather than an existing SNP) will cause a genetic disease while explicitly taking into account the underlying spectrum of nucleotide mutations. Such an approach will allow, in the future, the identification and characterization of highly mutable sites in the human genome which are also functionally important.

Miller and Kumar [5] performed a detailed analysis of the disease mutations and benign SNP spectra in seven human genes. While some of our results are consistent with their study, we find major differences. For example, we observe a significantly larger contribution of mutations at arginine (Arg) and glycine (Gly) to human genetic disease. We attribute the differences to the substantially larger gene set (436 genes versus 7) used in our analysis.

Results and discussion

Overall amino-acid mutational spectrum

We present the amino-acid spectra of disease mutations and polymorphisms in Figure 1. The mutations from the Mendelian Inheritance in Man (MIM) database [7] annotated in SWISS-PROT [8] were used as a source of human disease mutations. In total, 4,236 mutations from 436 genes were considered. The collection of 1,037 synonymous and nonsynonymous SNPs from the extensive analysis of haplotypes in 313 human genes [9] was used as a source of benign SNPs. There was no overlap between the disease mutations and benign sets of SNPs used in the study. The spectrum of interspecies substitutions (Figure 1d) was calculated on the basis of the PAM1 matrix [10] as described in Materials and methods.

Nearly all mutations in the current MIM database represent Mendelian disease (monogenic in etiology). It remains to be seen to what extent our results pertain to disease mutation involved in polygenic disorders. At this point, too little is known about this type of mutation, and more experimental work is required in order to understand their spectrum.

The mutation matrices in Figure 1 are sparse (that is, a large number of the matrix elements are close or equal to zero) and nonsymmetrical (in many cases the tendency of amino acid I to mutate into amino acid J is different from the tendency of amino acid J to mutate into I). The vast majority of human genetic mutations are caused by single-nucleotide changes [11,12]. Consequently, the matrices in Figure 1b,1c,1d represent amino-acid transitions resulting predominantly from single-nucleotide mutations in amino-acid codons. To rationalize the observed disease and benign spectra, we generated the expected mutation spectrum (Figure 1a) using the neighbor-dependent matrix of nucleotide mutation rates developed by Hess *et al.* [13] (see Materials and methods). The expected mutation matrix in Figure 1a represents the spectrum which would be observed if all nonsynonymous mutations were accepted (that is, there were no selection). The expected spectrum was generated for the disease genes considered and, separately, for a large collection of more than 7,000 human genes available from SWISS-PROT. These two spectra were almost identical ($R = 0.98$, $p < 0.0001$), suggesting that the expected spectrum in Figure 1a reflects general properties of all human genes (such as amino-acid codon frequencies and context-dependent nucleotide mutation frequencies). Here and throughout the paper we use the *t*-test statistics with *n*-2 degrees of freedom to estimate the significance of linear correlations. Random shuffling simulations confirmed the significance values obtained using the *t*-test.

The spectrum of disease mutations was calculated separately for membrane proteins. The program TMHMM [14] was used to detect potential transmembrane regions. The disease spectrum for membrane proteins is very similar to the all-protein disease spectrum ($R = 0.97$, $p < 0.0001$ for all disease mutations in membrane proteins, 1,598 in total; $R = 0.75$, $p < 0.0001$ for disease mutations in transmembrane regions, 372 in total). Evidently, specific properties of membrane proteins and the constraints on them are not able to significantly modify the disease spectrum common to all proteins.

Correlations between the expected and the observed spectra

Close-to-diagonal mutations in Figure 1a,1b,1c,1d represent substitutions between amino acids with similar chemical properties (conservative mutations). The interspecies substitutions (Figure 1d) contain the highest fraction of conservative mutations compared to disease mutations and benign SNPs. The frequencies of the benign SNPs, disease mutations, and interspecies substitutions are plotted versus expected frequencies in Figure 2. Benner *et al.* [15] showed that the genetic code affects the amino-acid substitution spectrum at early stages of divergence, whereas chemical similarities dominate at longer evolutionary distances. The correlation between the benign and expected spectra observed in our study ($R = 0.78$, $p < 0.0001$), is an expected extension of Benner's *et al.* conclusion to even shorter evolutionary distances (variations within a population).

Interestingly, we also find a strong correlation of the disease mutation spectrum with the expected spectrum based on the genetic code ($R = 0.71$, $p < 0.0001$). The correlation of disease mutation frequencies with the chemical dissimilarities between original and mutant amino acids is apparent only after normalization by the expected frequencies (Figure 3a,3b). Consequently, in the majority of cases the comparison of amino-acid types (wild type versus mutant) will be insufficient to distinguish neutral from disease variants.

The contribution of mutations at different amino acids to the disease spectrum is highly heterogeneous (Figure 4). Interestingly, mutations at Arg residues account for almost 15% of the disease mutations. This is a direct consequence of the well-known high mutability of Arg (due to deamination of

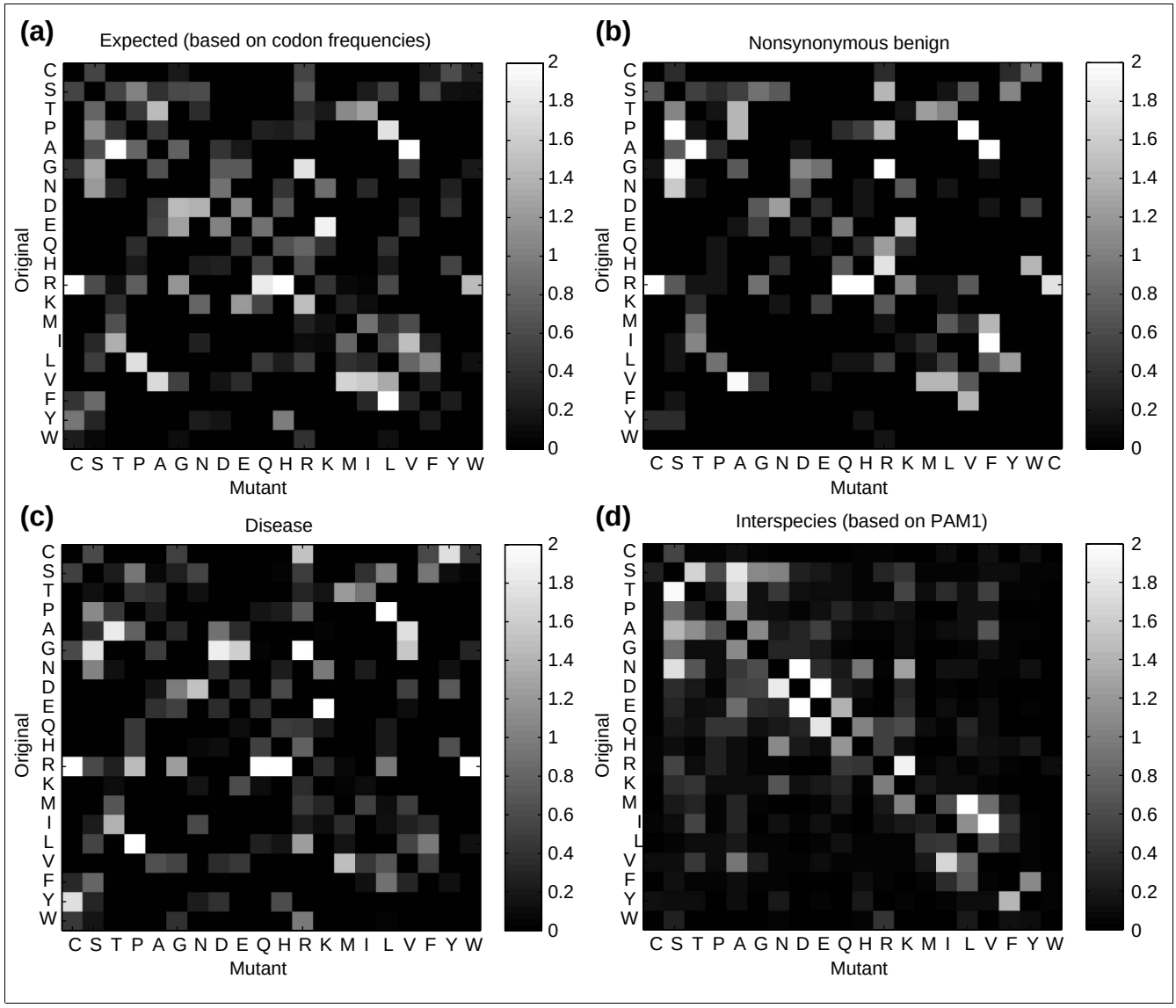

Figure 1
Amino-acid mutation frequencies in human genes. **(a)** The expected mutation frequencies based on the neighbor-dependent nucleotide mutation rates. The expected mutation matrix represents the frequencies of amino-acid transitions in the absence of selection. **(b)** The nonsynonymous benign SNP frequencies using the SNP dataset of Stephens *et al.* [9]. **(c)** The genetic disease mutation frequencies based on the Mendelian Inheritance in Man (MIM) database [7]. **(d)** The interspecies mutation frequencies based on the PAM1 matrix [10]. In the matrices, each square represents a particular amino-acid to amino-acid mutation (for example, Val → Ala). The gray level of the matrix squares is proportional to the number of observed mutations. The matrices were normalized so that the sum over all mutation frequencies for each matrix is equal to 100. The *y*-axes of the matrices represent the original (wild-type) amino acids; the *x*-axes represent the mutant amino acids (created as a result of a single-nucleotide mutation). The amino acids (given in single-letter amino-acid code) were ordered along the axes according to the side-chain chemistry [42]: (C) sulfhydryl; (STPAG) small hydrophilic; (NDEQ) acid, acid amide and hydrophilic; (HRK) basic; (MILV) hydrophobic; (FYW) large hydrophobic/aromatic. As a result of the ordering, the mutations close to the matrix diagonal tend to be more conserved.

5'-CpG dinucleotides in Arg codons) [16,17], the relatively high frequency of Arg in human proteins (<4%), and the fact that Arg mutates to residues with very different chemical properties (cysteine (Cys), glycine (Gly), histidine (His), lysine (Lys), leucine (Leu), methionine (Met), proline (Pro), flutamine (Gln), serine (Ser) and tryptophan (Trp)). The relative probability of a disease mutation at different amino acids (Figure 4b) was calculated by dividing the disease and expected frequencies. Accordingly, a random mutation at a Trp or Cys residue has the highest probability of causing a disease. This correlates well with the highest evolutionary conservation of exactly these two residues [10]. Both Trp and Cys residues play a prominent part in determining protein stability. In addition to Trp and Cys, the high probability of disease mutations at Gly may be related to important structural roles often played by this residue. For example, mutations at Gly,

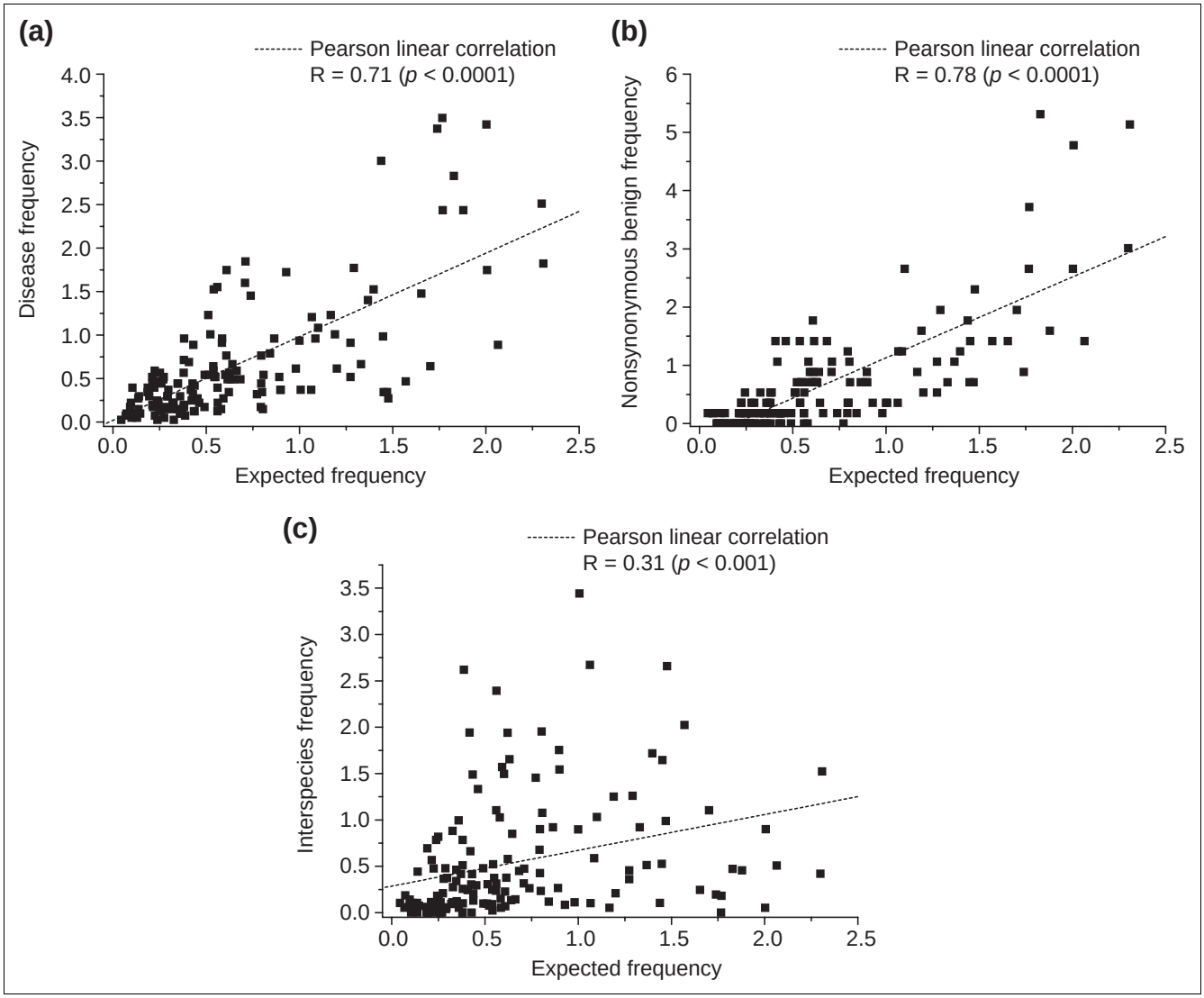

Figure 2

The expected frequencies of amino-acid to amino-acid mutations versus observed frequencies of the genetic disease mutations, nonsynonymous benign SNPs, and interspecies mutations. Comparison with **(a)** genetic disease mutations from the Mendelian Inheritance in Man (MIM) database [7], **(b)** nonsynonymous benign SNPs, based on the study by Stephens *et al.* [9], and **(c)** interspecies mutations based on the PAM1 matrix [10]. Each point in the figure represents a certain type of amino-acid to amino-acid mutation. Only amino-acid transitions resulting from single-nucleotide mutations in amino-acid codons were considered. The mutation frequencies in each class (benign, disease and interspecies) were normalized to 100.

which is frequently present at the turns of alpha-helices, might have a negative impact on protein structural stability. Our definition of the relative probability of disease mutations is similar to the relative clinical observation likelihood (RCOL) used by Cooper *et al.* in several publications (see, for example [6]). In the next section we extend the relative probabilities to interspecies comparisons.

Probabilities of mutations or SNPs as a function of the mutation/SNP-site properties

To complement the analysis of the amino-acid mutation matrices, we investigated how the probabilities of benign SNPs and disease mutations depend on the properties of the mutation site. Several recent studies have focused on developing evolutionary and structural approaches to predict potentially deleterious human mutations [1-4,18,19]. Here, we focus on understanding the relative mutation probabilities (see Materials and methods). Our results are in general agreement with the previous studies. The relative probabilities of disease mutations and benign SNPs are shown in Figure 5a as a function of the interspecies evolutionary conservation of the mutation site. The conservation was characterized by the relative entropy measure using homologs with more than 30% sequence identity. The probability that a random mutation will cause a genetic disease increases monotonically with an increase in the degree of site conservation, while the

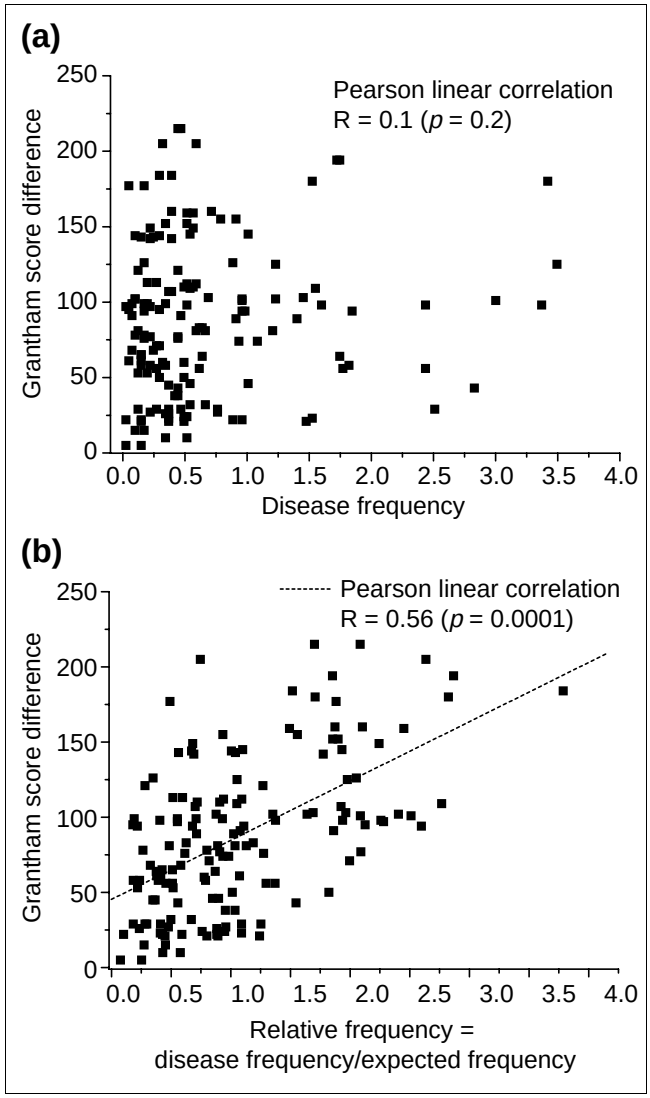

Figure 3
Chemical dissimilarities between original and mutant amino acids versus observed mutation frequencies. **(a)** The amino-acid dissimilarities versus the frequencies of disease mutations. **(b)** The amino-acid dissimilarities versus the relative frequencies of genetic disease mutations (normalized by site mutabilities). The chemical dissimilarities between amino acids were characterized using the Grantham score [22]. Each point in the figure represents a certain type of amino-acid to amino-acid mutation. Only amino-acid transitions resulting from single-nucleotide mutations in amino-acid codons were considered. The mutation frequencies in (a) were normalized to 100. The relative frequencies in (b) were determined as the ratio between the disease and the expected mutation frequencies (see Figure 1). The relative frequencies are proportional to the relative probabilities of amino-acid mutations causing a disease. No correlation between the disease mutation frequencies and chemical dissimilarities is evident (a), but there is a significant correlation between the normalized frequencies and the chemical dissimilarities (b).

probability of observing nonsynonymous benign SNPs shows the opposite trend. The synonymous benign SNPs do not change amino acids and should be predominantly neutral. As a result, their probability is uniform across sites.

The solvent accessibility of an amino-acid residue in a protein reflects the degree of the residue's exposure to the surrounding solvent in the protein structure. The relative probability of disease-causing mutations is highest in the protein interior and lowest on the protein surface (Figure 5b). The benign SNPs show the reverse trend, as their relative probability is highest on the surface and lowest in the protein interior. This is consistent with the study by Moult and co-workers [1] (see also Ferrer-Costa *et al.* [20] and Bustamente *et al.* [21]), who suggested that the dominant mechanism by which disease mutations damage protein function is a decrease in protein stability, as opposed to mutations of active-site residues (usually located on the protein surface).

Both relative entropy and solvent accessibility exclusively characterize the site of a mutation. To estimate the extent to which a given amino acid is incompatible with the residues observed at the same position in close homologs, we introduced the Grantham Ratio (GR) score based on the Grantham dissimilarity matrix [22] (see Materials and methods for a formal definition). Application of other scores, for example those based on the BLOSUM matrices, gave qualitatively similar results [3]. The GR score is the ratio of two averages - the numerator being the average dissimilarity between the mutated amino acid and the residues observed at the same site in evolution, and the denominator being the average dissimilarity within the residues observed at the site in homologous proteins. Defined in this way, a GR score smaller or close to 1 suggests that the amino acid is similar to the residues observed at the site in evolution, whereas a GR score significantly larger than 1 indicates that the amino-acid change is evolutionarily radical.

The role of purifying selection in shaping the mutation spectra is apparent from the cumulative distribution of the GR scores (Figure 6). Whereas the GR distribution for original (wild type) residues at benign sites (blue curve) is very similar to the distribution for all protein residues (black), the distribution for mutant residues at benign sites (green) clearly shows an excess of radical mutations. Importantly, the GR distribution of mutant residues at benign sites (green) is similar to the distribution for randomly generated mutations (cyan) and is quite different from the disease mutation distribution (red). Consequently, although a significant fraction of randomly arising nonsynonymous mutations are evolutionarily radical (and thus potentially deleterious) they are not, on average, as radical as the disease mutations and still have appreciable frequencies in the human population. Indeed, it was recently estimated [23] that the average reduction in evolutionary fitness due a mildly deleterious SNP with a significant frequency in the human population is in the range of 0.01-1%. The medical importance of such mildly deleterious human mutations remains to be established [24,25].

The cumulative distribution of the GR scores for disease mutations suggests that more than a half of the disease

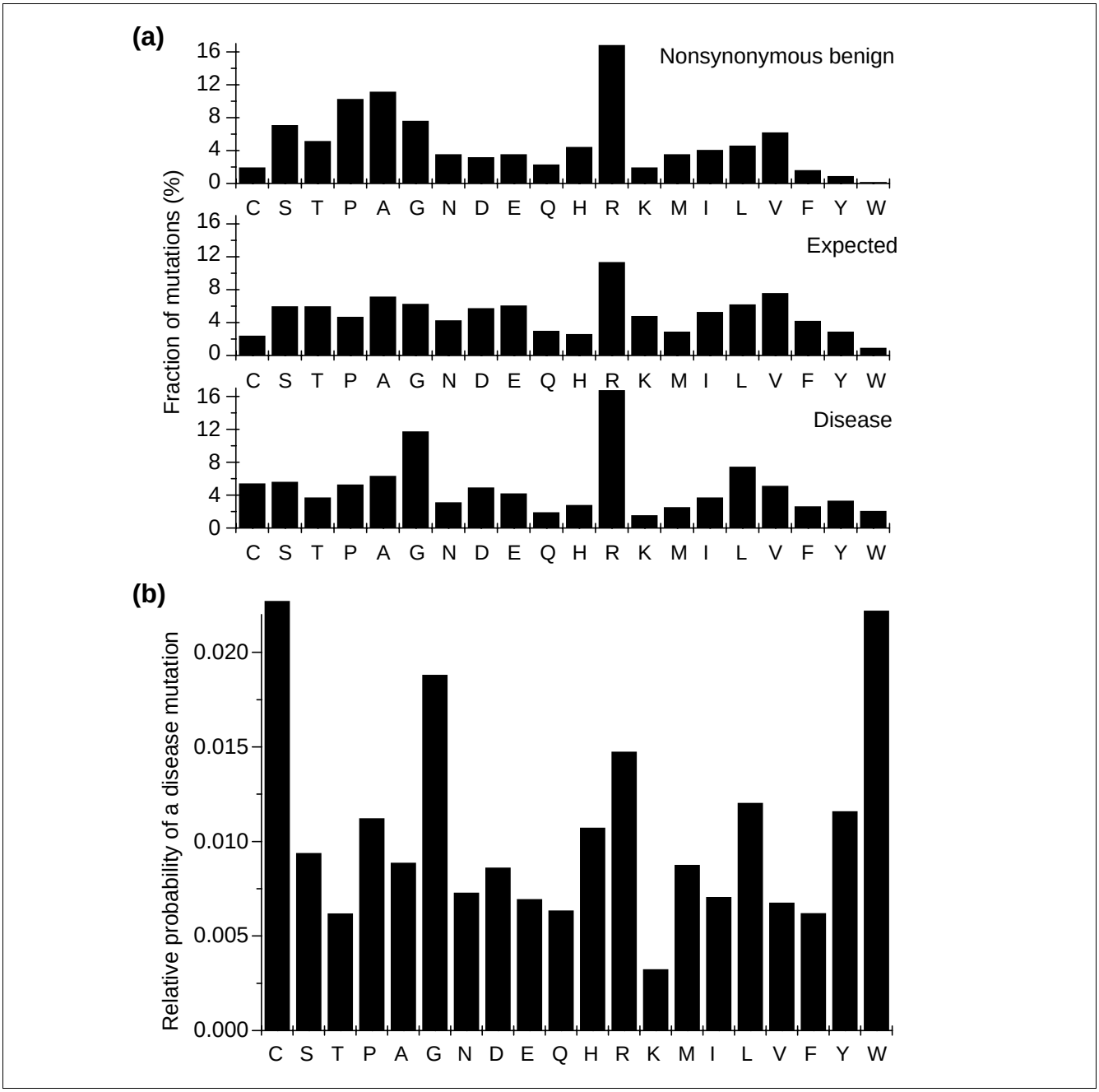

Figure 4

Contribution of mutations at different amino acids to the overall mutation spectrum. **(a)** The fraction of mutations at different amino acids. The fractions are shown separately for benign, expected and disease mutations (normalized to 100% within each class). The contribution of mutations at different amino acids to the overall spectrum is highly heterogeneous. For example, mutations at arginine (R) constitute approximately 15% of all mutations. This is a direct consequence of a high mutability of the 5'-CpG dinucleotides in the arginine codons. **(b)** The relative probability that a random mutation at different amino acids will cause a genetic disease. Importantly, because the overall probability that a random mutation will cause a genetic disease is unknown, the probabilities in (b) have only relative meaning (for example, the probability that a random mutation will cause a disease mutation at alanine (A) versus valine (V)). For display purposes it was assumed that 1 in 100 random mutations causes a genetic disease. Mutations at tryptophan (W) and cysteine (C) have the highest probability of causing a disease. This correlates with the fact that these are the most highly conserved amino acids in evolution [10].

mutations are evolutionarily radical (represented by residues with GR score greater than 2). Residues with such GR scores are almost never observed in homologous sequences (blue and black curves). It is important to note that medically damaging mutations and SNPs cannot always be rationalized in terms of evolutionary radicality. Medically harmful

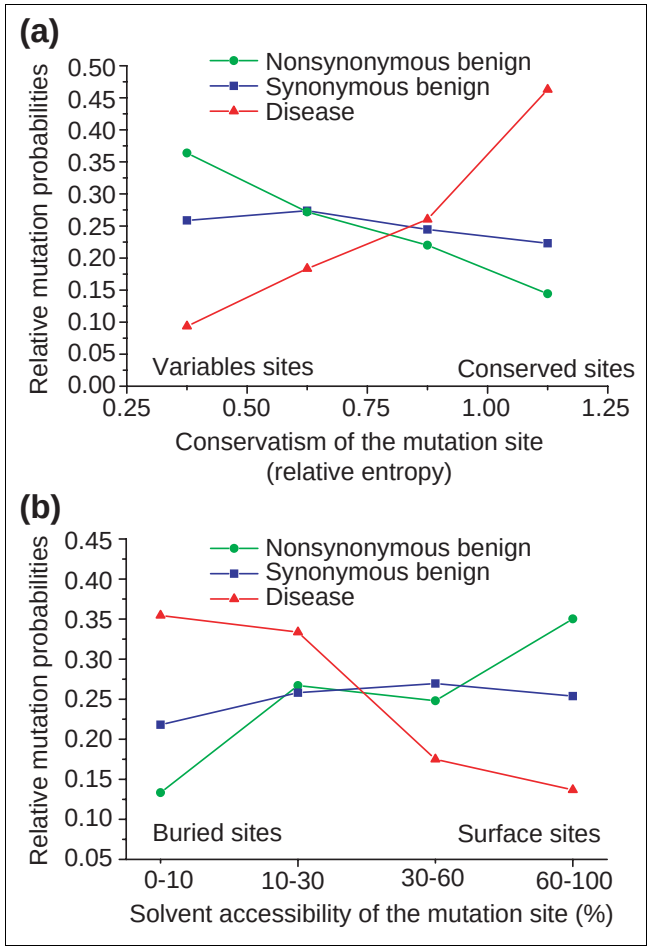

Figure 5
The relative mutation probabilities as a function of mutation site conservation and solvent accessibility. Relative mutation probability as a function of **(a)** evolutionary conservation of the mutation site (measured using relative entropy), and **(b)** solvent accessibility of the mutation site in the protein structure. Because the overall probability that a random mutation will cause a genetic disease or be observed as a polymorphism is not known, the probabilities have only relative meaning within each mutation class (disease, synonymous, nonsynonymous). To show different trends clearly, the relative probabilities were normalized to 1 within each class. Conservation of mutation sites in evolution was characterized by the relative entropy using close sequence homologs (see Materials and methods). The solvent accessibility of mutation sites was calculated using the program NACESS [41]. An increase in the degree of evolutionary conservation increases the probability of deleterious mutations and decreases the probability of nonsynonymous benign SNPs (a). An increase in the degree of solvent accessibility decreases the probability of deleterious mutations and increases the probability of nonsynonymous benign SNPs (b). Synonymous mutations do not change amino-acid sequences and are predominantly neutral. Consequently, the probability that a synonymous mutation will be deleterious is relatively constant across sites.

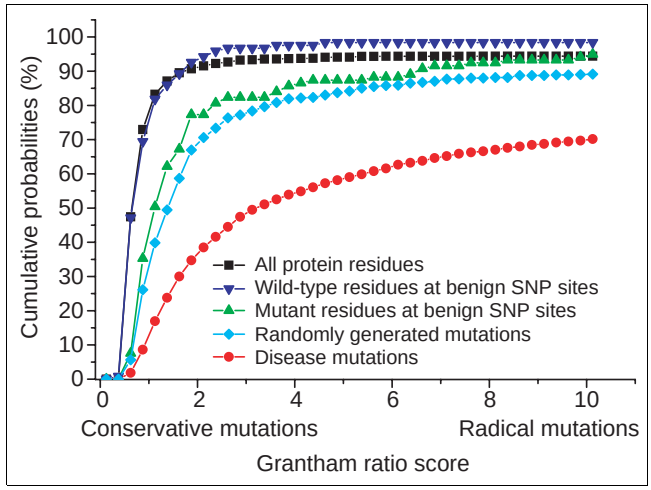

Figure 6
Cumulative probability of the Grantham ratio (GR) for different classes of residues in proteins. Black, all (wild-type) protein residues; blue, original (wild-type) residues at the sites of benign SNPs; green, mutant residues at the sites of benign SNPs; cyan, residues generated by computer simulation of random mutations based on the amino-acid mutation frequencies; red, disease-causing residues from MIM. The Grantham ratio characterizes the degree of the residue's dissimilarity to the amino acids observed at the same position in evolutionary homologs (see Materials and methods). High GR values indicate radical mutations, whereas GR values that are small or around 1 indicate conservative mutations. The GR distributions demonstrate how purifying selection affects the observed mutation spectra. Comparison of the GR scores for original residues (black and blue) and disease-causing residues shows that more than half of disease mutations are radical (GR > 2) and are almost never observed in evolution.

mutations may cause late-onset human diseases without strong selection in evolution. Alternatively, a particular amino-acid substitution can be damaging to a human protein but be relatively frequent in the homologous family due to compensatory mutations. Such substitutions may account for deleterious mutations with low GR scores.

Estimation of the maximal rate of mutations with impact on human health

From Figure 6 we can estimate the maximum rate of random mutations with significant impact on human health (that is, an impact similar to mutations currently annotated in MIM). We note that the mutation rate we estimate (a fraction of newly created deleterious mutations) is different from the fraction of existing SNPs with deleterious effects on protein function (estimated previously [1,2,23]). The comparison between the distribution of random SNP mutations (cyan) and disease mutations (red) suggests that about 10% of the randomly generated mutations have GR scores greater than 6. Such a score corresponds to approximately 40% of the disease mutations. As a result, the total rate of the disease mutations cannot be larger than one quarter of the random mutation rate. Thus, one expects, at most, 25% of random nonsynonymous mutations to be as damaging as mutations currently annotated in MIM (similar estimates are obtained using GR cutoffs larger than 6).

This estimate has a simple biochemical rationale, as mutagenesis experiments on different proteins suggest that less

than 30% of random mutations substantially damage biological function or stability of proteins [26-29].

Using the recent estimate of the human mutation rate of 175 mutations per diploid genome per generation [30] (corresponding to approximately two to three nonsynonymous mutations), we conclude that the rate of nonsynonymous mutations with serious impact on human health should be less than one per diploid genome per generation. This is probably a substantial overestimation of the rate because we assume that all human genes are as important for human health as the well-annotated disease genes currently in the MIM database. We emphasize that the rate of health-damaging nonsynonymous mutations is smaller than the total rate of deleterious human mutations, which is estimated to be larger than one [30,31].

The present analysis, together with other recent studies [1-4,23], establishes the basis for understanding the spectrum of deleterious human mutations. The amino-acid substitution matrices, such as PAM [10] and BLOSUM [32], apart from playing a fundamental role in sequence alignment, qualitatively characterize the evolutionary interchangeability of amino acids averaged over many protein families. The disease spectrum, characterized by our analysis, explores another important aspect of evolution, namely the generation of deleterious mutations. Because of all mammalian species have a broadly similarity physiology, the properties of the disease spectrum should be general, at least for mutations leading to early-onset diseases. We anticipate that understanding the disease spectrum will allow one to predict, in advance, the rates and potential medical consequences of all possible single-nucleotide mutations in the human genome.

Materials and methods

Calculation of mutation spectra

The spectrum of expected amino-acid mutation frequencies (Figure 1a) was generated using the matrix of neighbor-dependent nucleotide mutation rates obtained by Hess *et al.* [13] (Additional data file 1). The neighbor-dependent mutation matrix was calculated by Hess *et al.* on the basis of 20,200 substitutions in aligned gene/pseudogene human sequences; the relative mutation rates were calculated for the four nucleotides in all 16 possible 5' and 3' neighborhoods. To obtain the expected amino-acid mutation frequencies for a given collection of genes, we simulated all possible single-nucleotide mutations with appropriate rates, and recorded the corresponding amino-acid changes. The nucleotide mutational spectrum of individual genes may be affected by the presence of so-called mutation hot spots [33-35]. However, on average, there is only a small influence of the surrounding DNA sequence (beyond nearest 5' and 3' neighbors) on the relative nucleotide mutation rates [17].

The interspecies spectrum of amino-acid mutation frequencies (Figure 1d) was calculated on the basis of Dayhoff's PAM1 matrix. The original PAM1 matrix [10] gives the probabilities of amino-acid substitutions over small evolutionary distances. These probabilities were multiplied by the amino-acid frequencies in human genes for direct comparison with the expected, disease, and benign SNPs matrices.

Structural and evolutionary analysis of mutations

The list of disease genes obtained from SWISS-PROT was filtered using the program PSEG [36] to exclude genes with a significant fraction of low-complexity regions. As a result of the filtering, six genes for collagen proteins were excluded from the original set of 436 genes. Mutations at Gly residues constitute more than 50% of the collagen disease mutations (due to the collagen structural motif). Because of this bias, the collagen mutations were excluded from all calculations. If the collagen mutations are included, the total fraction of disease mutations at Gly (Figure 4a) increases from 12% to 15%.

Membrane proteins and transmembrane protein regions were detected using the program TMHMM [14] with standard parameters. Out of 430 disease genes, 105 (24%) were classified as membrane proteins on the basis of the presence of at least two distinct transmembrane domains. To characterize the evolutionary conservation of mutation sites we used BLASTGP to search the nrdb90 database [37] for homologs with greater than 30% sequence identity. The nrdb90 database constitutes a nonredundant merge of sequence and structural databases, which is filtered so that no pair of sequences has greater than 90% sequence identity. The homologs to each human protein were subsequently aligned using the program CLUSTALW [38] with default parameters. Only mutation sites covered by more than 10 homologous sequences (excluding gaps) were used in the evolutionary analysis. The multiple sequence alignments obtained using CLUSTALW were used to characterize the relative entropy (Kullback-Leibler distance) of the benign and disease mutation sites. The relative entropy was calculated according to the formula:

$$\text{Relative Entropy} = \sum_{n} P(n) \log \frac{P(n)}{Q(n)}$$

where the summation is over all amino-acid types n in the alignment; P(n) is the probability of the amino acid n in the column corresponding to mutation; Q(n) is the probability of the amino acid n in all columns of the multiple sequence alignment.

The multiple sequence alignments were also used to calculate the Grantham ratio (GR) score according to the formula:

$$GR = \frac{\frac{1}{n}\sum_i D(Human\_RES, RES(i))}{\frac{2}{n(n-1)}\sum_i \sum_{j>1} D(RES(i), RES(j))}$$

where $D(A,B)$ is the Grantham measure of chemical dissimilarities between amino-acid residues A and B, *Human_RES* is the human residues at the mutation site, *RES(i)* is the amino acid from the ith aligned sequence homolog at the mutation site, and n is the number of aligned sequences. Qualitatively, the GR score is a measure of dissimilarity between a human amino acid and the residues seen at the same site in homologs. In total, the relative entropy and Grantham ratio were calculated for 258 benign SNPs and 2,636 disease mutations.

To characterize the structural location of disease mutations and benign SNPs, BLASTGP [39] was used to search the Protein Data Bank (PDB) [40] for sequences homologous to known structures. Only sequences with greater than 30% identity to human sequences over the entire length of the alignment were considered. In total, the solvent accessibilities were calculated for 110 benign SNPs and 840 disease mutations. The solvent accessibility of mutation sites was determined by the program NACCESS [41] using the water-sphere radius of 1.4 Å. The solvent accessibility represents the relative exposure of a residue X in a protein structure compared to its exposure in the tripeptide Ala-X-Ala.

Calculation of relative mutation probabilities

The relative mutation probabilities shown in Figures 4b, 5a, and 5b represent conditional probabilities. Specifically, the conditional probability P(disease|descriptor), that a mutation will cause a genetic disease given a certain property (descriptor) of the mutation site was calculated according to the formula:

$$P(disease \mid descriptor) = \frac{P(descriptor \mid disease)}{P(descriptor)} P(disease)$$

where 'descriptor' represents solvent accessibility or evolutionary conservation of the mutation site, P(descriptor|disease) is the probability that a disease mutation has a given descriptor value, P(descriptor) is the probability that a random mutation (disease or non-disease) has a given descriptor value, and P(disease) is the probability that a random mutation will cause a genetic disease. Importantly, because P(disease) is unknown, we can only estimate P(disease|descriptor) up to a constant (assuming certain P(disease) value). Consequently, we refer to P(disease|descriptor) as relative mutation probabilities. The probability that a random mutation has a given descriptor value P(descriptor) was estimated by simulating random single-nucleotide mutations using the expected amino-acid mutation frequencies (Figure 1a).

Additional data files

The following additional data are included with the online version of this article: a list of relative mutation rates (Additional data file 1), a list of disease mutations (Additional data file 2), a list of disease mutation genes (Additional data file 3), a list of SNPs used in the analysis (Additional data file 4), and the Grantham ratio scores (Additional data file 5).

Acknowledgements

We thank Jay Shendure, John Aach, Patrik D'haeseleer, Daniel Segre, Peter Kharchenko, and Tzachi Pilpel for discussions. This work was supported in part by research grants from the US Department of Energy through the grant DOE DE-FG02-87-ER60565.

References

1. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Hum Mutat* 2001, **17:**263-270.
2. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P: **Prediction of deleterious human alleles.** *Hum Mol Genet* 2001, **10:**591-597.
3. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11:**863-874.
4. Chasman D, Adams M: **Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation.** *J Mol Biol* 2001, **307:**683-706.
5. Miller MP, Kumar S: **Understanding human disease mutations through the use of interspecific variation.** *Hum Mol Genet* 2001, **10:**2319-2328.
6. Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M: **Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease.** *Hum Mutat* 2002, **20:**98-109.
7. McKusick VA: *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders* 12th edition. Baltimore: John Hopkins University Press; 1998.
8. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TrEMBL.** *Nucleic Acids Res* 1996, **24:**21-25.
9. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al.*: **Haplotype variation and linkage disequilibrium in 313 human genes.** *Science* 2001, **293:**489-493.
10. Dayhoff MO: **A model of evolutionary change in proteins.** In *Atlas of Protein Sequence and Structure* Edited by: Silver Spring: National Biomedical Research Foundation. Dayhoff MO; 1978:345-352.
11. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.** *Nat Genet* 1999, **22:**239-247.
12. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, *et al.*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22:**231-238.
13. Hess ST, Blake JD, Blake RD: **Wide variations in neighbor-dependent substitution rates.** *J Mol Biol* 1994, **236:**1022-1033.
14. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-182.
15. Benner SA, Cohen MA, Gonnet GH: **Amino acid substitution during functionally constrained divergent evolution of protein sequences.** *Protein Eng* 1994, **7:**1323-1332.
16. Cooper DN, Youssoufian H: **The CpG dinucleotide and human genetic disease.** *Hum Genet* 1988, **78:**151-155.
17. Krawczak M, Ball EV, Cooper DN: **Neighboring-nucleotide effects on the rates of germ-line single base-pair substitution in human genes.** *Am J Hum Genet* 1998, **63:**474-488.
18. Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function.** *Genome Res* 2002, **12:**436-446.
19. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

20. Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.** *J Mol Biol* 2002, **315:**771-786.
21. Bustamante CD, Townsend JP, Hartl DL: **Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*.** *Mol Biol Evol* 2000, **17:**301-308.
22. Grantham R: **Amino acid difference formula to help explain protein evolution.** *Science* 1974, **185:**862-864.
23. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome.** *Genetics* 2001, **158:**1227-1234.
24. Terwilliger JD, Haghighi F, Heikkalinna TS, Goring HH: **A biased assessment of the use of SNPs in human complex traits.** *Curr Opin Genet Dev* 2002, **12:**726-734.
25. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: **Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.** *Nat Genet* 2003, **33:**177-182.
26. Olins PO, Bauer SC, Braford-Goldberg S, Sterbenz K, Polazzi JO, Caparon MH, Klein BK, Easton AM, Paik K, Klover JA, *et al.*: **Saturation mutagenesis of human interleukin-3.** *J Biol Chem* 1995, **270:**23754-23760.
27. Huang W, Petrosino J, Hirsch M, Shenkin PS, Palzkill T: **Amino acid sequence determinants of beta-lactamase structure and activity.** *J Mol Biol* 1996, **258:**688-703.
28. Pakula AA, Sauer RT: **Genetic analysis of protein stability and function.** *Annu Rev Genet* 1989, **23:**289-310.
29. Matthews BW: **Structural and genetic analysis of the folding and function of T4 lysozyme.** *FASEB J* 1996, **10:**35-41.
30. Nachman MW, Crowell SL: **Estimate of the mutation rate per nucleotide in humans.** *Genetics* 2000, **156:**297-304.
31. Eyre-Walker A, Keightley PD: **High genomic deleterious mutation rates in hominids.** *Nature* 1999, **397:**344-347.
32. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
33. Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF: **Recombinational and mutational hotspots within the human lipoprotein lipase gene.** *Am J Hum Genet* 2000, **66:**69-83.
34. Zavolan M, Kepler TB: **Statistical inference of sequence-dependent mutation rates.** *Curr Opin Genet Dev* 2001, **11:**612-615.
35. Rogozin I, Kondrashov F, Glazko G: **Use of mutation spectra analysis software.** *Hum Mutat* 2001, **17:**83-102.
36. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-571.
37. Holm L, Sander C: **Removing near-neighbour redundancy from large protein sequence collections.** *Bioinformatics* 1998, **14:**423-429.
38. Higgins DG, Thomposon JD, Gibson TJ: **Using CLUSTAL for multiple sequence alignments.** *Methods Enzymol* 1996, **266:**383-402.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acid Res* 1997, **25:**3389-3402.
40. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: A computer based archival file for macromolecular structures.** *J Mol Biol* 1977, **112:**535-542.
41. Hubbard SJ, Thornton JM: *NACCESS Computer Program* London: Department of Biochemistry and Molecular Biology, University College London; 1993.
42. Mount DW: *Bioinformatics* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2001.