Human Molecular Genetics, 2001, Vol. 10, No. 6 591–597

Prediction of deleterious human alleles

Shamil Sunyaev[1,2,3], Vasily Ramensky[3], Ina Koch[3], Warren Lathe III[1,2], Alexey S. Kondrashov[4] and Peer Bork[1,2,+]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany, [2]Max-Delbrueck Centre for Molecular Medicine, Robert-Roessle-Strasse 10, D-13122 Berlin, Germany, [3]Engelhardt Institute of Molecular Biology, Vavilova 32, D-117984 Moscow, Russia and [4]National Center for Biotechnology Information, National Institutes of Health, 45 Center Drive, MSC 6510, Bethesda, MD 20892-6510, USA

Received 20 November 2000; Revised and Accepted 25 January 2001

Single nucleotide polymorphisms (SNPs) constitute the bulk of human genetic variation, occurring with an average density of ~1/1000 nucleotides of a genotype. SNPs are either neutral allelic variants or are under selection of various strengths, and the impact of SNPs on fitness remains unknown. Identification of SNPs affecting human phenotype, especially leading to risks of complex disorders, is one of the key problems of medical genetics. SNPs in protein-coding regions that cause amino acid variants (non-synonymous cSNPs) are most likely to affect phenotypes. We have developed a straightforward and reliable method based on physical and comparative considerations that estimates the impact of an amino acid replacement on the three-dimensional structure and function of the protein. We estimate that ~20% of common human non-synonymous SNPs damage the protein. The average minor allele frequency of such SNPs in our data set was two times lower than that of benign non-synonymous SNPs. The average human genotype carries approximately 10^3 damaging non-synonymous SNPs that together cause a substantial reduction in fitness.

INTRODUCTION

Probably the most important question related to genetic variation is what fraction of it constitutes non-neutral allelic variants, i.e. those variants that affect phenotype and can be subject to the pressure of natural selection. Analyses of large single nucleotide polymorphism (SNP) collections show significant variations in SNP density and allele frequency distribution pointing to selection of diverse intensities in different regions of the genome, especially in protein-coding regions, due to SNPs resulting in amino acid allelic variants (non-synonymous cSNPs) (1–5). An ambitious task for human genetics is the identification of SNPs associated with various (most importantly disease) phenotypes (6,7). However, the complex nature of many phenotypes of interest and the immense number of SNPs to be analysed complicate association studies.

Linking DNA variation to variability at the level of phenotype and fitness can be facilitated by studying the impact of DNA variation on the structure and function of proteins. The importance of the functional analysis of amino acid allelic variants was appreciated (8,9) almost 10 years before the discovery of SNPs (10). However, only the recent accumulation of data on human polymorphisms, stored in databases such as HGBASE (11), dbSNP (12) and others, enabled large-scale studies aimed at linking genetic and phenotypic variation as well as estimating the key population genetic parameters.

Amino acid variants may impact folding, interaction sites, solubility or stability of the protein. These effects can be estimated from physical considerations and from the context of an amino acid replacement within the family of homologous proteins. As has been demonstrated previously (13), a significant fraction of non-synonymous cSNPs is likely to affect protein structure or function. In order to identify these SNPs, we developed a set of rules based on previous research in protein structure, interaction and evolution (14) that automatically predict whether a replacement is likely to be deleterious for the protein on the basis of three-dimensional (3D) structure and multiple alignment of homologous sequences. Mapping of amino acid replacement to the known 3D structure reveals whether the replacement is likely to destroy the hydrophobic core of a protein, electrostatic interactions, interactions with ligands or other features of a protein. Some non-synonymous cSNPs associated with human disorders and having substantial population frequencies are nevertheless known to disrupt important structural features of the affected proteins (Fig. 1).

RESULTS

To assess the possible damaging effect of amino acid substitutions we analysed the following: if the substitution is (i) in an annotated active or binding site; (ii) affects interaction with ligands present in the crystallographic structure; (iii) leads to hydrophobicity or electrostatic charge change in a buried site; (iv) destroys a disulphide bond; (v) affects the protein's solubility; (vi) inserts proline in an α-helix; or (vii) is incompatible with the profile of amino acid substitutions observed at this site in the set of homologous proteins.

To evaluate the ability of the proposed method to discriminate between deleterious and neutral amino acid variants, we first

+To whom correspondence should be addressed. Tel: +49 6221 387526; Fax: +49 6221 387517; Email: bork@embl-heidelberg.de

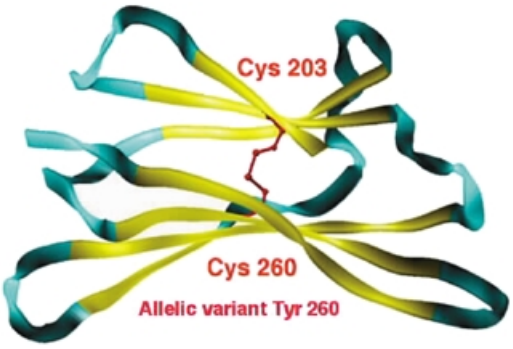

Figure 1. An example of a polymorphic variant which disrupts a critical disulphide bond. Although this variant (260 Cys→Tyr) in HLA-H protein is strongly associated with hereditary haemochromatosis (35), its frequency is as high as ~6% in Northern Europeans with up to 14% in Ireland (36).

applied it to the data on known deleterious mutations (as evident from the effect on phenotype or molecular function) and to the data on species divergence data.

The deleterious data set consists of natural replacements known to cause disease phenotypes, variants observed in individuals affected by genetic disorders and artificial replacements known to damage structure, function and/or stability of the protein (Table 1). Altogether this set comprises approximately 1550 mutations and a majority of these have been identified in patients. The method predicts that ~10–30% of these replacements are non-damaging. This estimates the rate of false-negative predictions (Table 1). Given the sources of potential inaccuracies in the data itself, this is a high accuracy of prediction. Future advances in protein databases and the prediction method will allow for more detailed analysis and improve the accuracy of the predictions (Fig. 2).

The divergence data set consists of amino acid differences between human proteins and their orthologues from other mammals. These differences can hardly have substantial negative effect on proteins, because deleterious alleles that reduce fitness by $>1/N_e$ (where N_e is effective population size) rarely reach fixation and also because long-term N_e of humans and related species is 10^4–10^6 (15,16). Our method predicts that ~9% of interspecies differences are damaging. This is an estimate of the rate of false-positive predictions (note that the possibility of fixation of some substitutions affecting protein structure or function would make the false-positive rate even lower).

The results of the application of our method to the polymorphism data set, which consists of heterogeneous SNPs from public databases (HGBASE, dbSNP, SWISS-PROT) are shown in Table 2. As many as ~30% of 245 well characterized SNPs tested were predicted to damage structure or function of proteins. Given the the rate of false-positive predictions of ~9% estimated on divergence data (Table 1), the true fraction of damaging SNPs should be ~20%. Analysis of publications cited by the HGBASE and OMIM databases for 99 cSNPs where both allele frequency and 3D structure are known revealed 11 cases where a disease association had been reported. Of these, eight (73%) were predicted to be damaging by our analysis, representing a several-fold enrichment over the rate for random non-synonymous cSNPs. Thus, an evaluation of predicted impact of non-synonymous cSNP on protein function could greatly aid the selection of candidate SNPs for direct association analysis. Given the continually increasing disparity between the multitude of known SNPs and the as yet limited ability of technology to score these in sufficient numbers, such improved prioritization of candidate SNPs will become an increasingly significant part of such investigations. Table 2 represents a list of 25 amino acid variants predicted to be

Table 1. Prediction and validation results

Control predictions	Total	Predicted	%
False-negative controls on deleterious data set			
All mutations mentioned with relation to a functional disorder	1551	1071	69
Subset of all mutations with evidence of causative effect on a disease	60	54	90
Subset of all mutations that are engineered with known effect on function	54	43	80
False-positive controls on divergence dataset			
Known between species substitutions in proteins from the deleterious dataset.	360	28	8
Known between species substitutions in proteins from polymorphism dataset.	440	41	9

Predictions on polymorphism data from referenced databases	Total	Predicted damage	%
All polymorphisms from databases	459	156	34
Experimentally proven polymorphisms from databases	245	79	32

Statistics on prediction results for non-synonymous cSNPs, mutations with known deleterious effect and substitutions between human proteins and their closely related mammalian orthologues. Non-synonymous cSNPs were classified according to the confidence of the SNP detection method used. Fraction of non-synonymous cSNPs predicted as deleterious is higher for unreliable cSNPs probably because of the errors in SNP discovery. Deleterious mutations were classified according to the annotation quality. Frequently, variants annotated in databases with connection to a disease are neutral variants identified in patients or variants associated with a disease phenotype due to linkage disequilibrium; therefore, data on the largest set of deleterious mutations reflect the low limit of the method's accuracy. This is why we formed a separate subset of mutations with clear evidence of their causative effect on a disorder. The accuracy of the method on this subset is significantly higher. A small subset of engineered mutations has been included in the analysis despite its distinction from a set of naturally occurring replacement mutations, because of clear evidence of functional consequences at the molecular level.

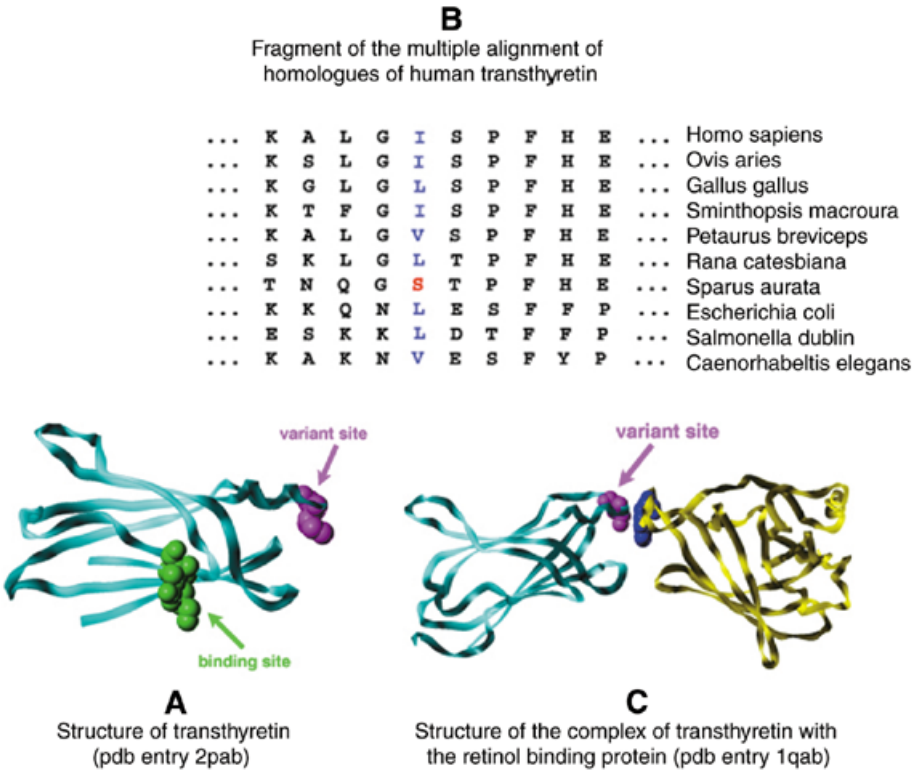

Figure 2. In some cases our straightforward rules fail to predict the deleterious effect of an amino acid substitution; however, more detailed analysis is capable of revealing the damaging character of the mutation. This suggests that further improvement of the method can significantly reduce the fraction of false-negative predictions. As an example, structural analysis (PDB entry 2pab) predicts the Ile→Ser replacement in transthyretin that causes amyloidosis type II to be a neutral substitution, because this is a substitution of a hydrophobic residue for a hydrophilic one on the surface, distant from the thyroid hormone binding site. (**A**). Comparative analysis also fails to detect the deleterious effect of the mutation because transthyretin from *Sparus aurata* (Gilthead sea bream) has a Ser residue in the corresponding site, probably as a result of a correlated mutation (**B**). However, analysis of the structure of the complex of transthyretin with the retinol binding protein (PDB entry 1qab) shows a critical role of the Ile residue for the complex formation (**C**).

damaging (from the set of 99 with known allele frequencies). The complete list of the predictions can be downloaded from www.bork.embl-heidelberg.de/SNPs and has been submitted to HGBASE. This list is expected to grow rapidly as many new cSNP discoveries continue to be added to these databases (11).

Further analysis of the 99 SNPs reveals that non-synonymous SNPs predicted to be damaging have lower allele frequencies compared with benign SNPs. The difference in allele frequencies is statistically significant (*P*-value of the Exact test for independence is 0.004; this is not an artefact of the mixed sample, since the difference remains significant even if large subsamples from a single survey are considered). This implies that damaging SNPs reduce fitness and, thus, are under negative selection. The 15% excess of rare alleles for non-synonymous cSNPs in comparison with synonymous SNPs reported previously (1) suggests that the majority, if not all, damaging SNPs are under negative selective pressure.

DISCUSSION

Estimation of the number of deleterious SNPs in the average genotype

The estimates of diversity (π) for amino acid replacing nucleotide substitutions imply that an average diploid human genotype contains approximately one non-synonymous deviation from the human consensus per 1000 encoded amino acids (0.0003–0.0004 in nucleotide base pairs; 1,2,5,17). Thus, assuming that the human genome encodes 45 000 proteins (current estimates are in the range of 30 000–120 000; 18) of average length 500 amino acids, a diploid human carries approximately 20 000 amino acids that differ from the consensus. Since ~20% of SNPs damage the protein, and since their allele frequencies in our data set are on average approximately two times less than the benign SNPs, ~10% of amino acid deviations from the consensus are damaging. Thus, an average human carries approximately 10^3 (more precisely approximately 2000) deleterious amino acid variants.

We note that although both estimates of the number of deleterious variants among all SNPs and of the average allele frequency may depend on sample size, for statistical reasons, the product of the two gives an unbiased estimate of the number of such variants in the average sequence.

Identification of many of these deleterious variants should be feasible in the near future owing to incoming data from SNP consortia and structural genomics projects.

Mutation selection balance

Recently, the genomic rate of *de novo* deleterious amino acid replacements in humans U_r was estimated to be approximately two per generation (19). This estimate includes only those

Table 2. Amino acid variants predicted to affect protein structure/function

Protein	SWISS-PROT ID	SWISS-PROT amino acid position	Amino acid substitution	PDB reference	PDB amino acid position	Proven SNP	Known disease association
E-selectin	P16581	130	C→W	1esl	109_	No	
Haemochromatosis gene product	O75931	180	C→Y	1a6z	260C	Yes	Yes
Acrosomal serine protease inhibitor	Q07616	38	N→S	1pai	49A	Yes	
Apolipoprotein I	P02647	184	R→P	1gw3	19_	No	
Annexin III	P12429	251	P→L	1axn	252_	Yes	
Beta-1 arenergic receptor	P08588	389	G→R	1dep	11_	Yes	Yes
Hydrogenase expression/formation protein	P26441	182	H→R	1cnt	1821_	Yes	
Placental ribonuclease inhibitor	P13489	169	P→L	1a4y	169A	Yes	
Apolipoprotein E	P02649	176	R→C	1le2	158_	Yes	Yes
Placental growth hormone	O14643	90	S→C	1bp3	79A	Yes	
Placental growth hormone	O14643	75	R→W	1bp3	64A	No	
Aldose reductase	P15121	203	G→S	1ads	203_	No	
Glandular kallikrein 1	P06870	77	R→H	1sgf	70G	No	
Prostaglandin G/H synthase 2	P35354	488	E→G	5cox	502B	No	
Insulin receptor substrate-1	P35568	158	P→R	1irs	158A	Yes	
Insulin receptor substrate-1	P35568	209	M→T	1irs	209A	Yes	Yes
Apolipoprotein E	P02649	130	C→R	1le2	112_	Yes	Yes
Cholinesterase	P06276	98	D→G	1mah	74A	Yes	Yes
Alpha-1 anti-trypsin	Q13747	147	E→V	1kct	264_	Yes	
Plasma serine protease inhibitor	P05154	217	G→R	1pai	202A	Yes	
Prion protein	P78446	212	E→K	1ag2	219_	Yes	
Guanine nucleotide-binding protein beta subunit-1	P04901	339	W→L	1gp2	339B	No	
Chymase	P23946	66	H→R	1pjp	57A	No	
Alcohol dehydrogenase beta chain	P00325	47	R→H	1hdx	47A	Yes	Yes
Alcohol dehydrogenase beta chain	P00325	369	R→C	1hdx	369A	Yes	Yes

List of non-synonymous cSNPs with predicted deleterious effect on the structure or function of the protein. The list contains only SNPs with reported allele frequency. The complete list of predictions is available at www.bork.embl-heidelberg.de/SNPs and has been submitted to HGBASE.

replacements that are deleterious enough not to be fixed in human and chimp lineages since their divergence from the common ancestor. Since our method has been tested against a set of substitutions between human proteins and their orthologues in other mammals, we will treat $U_r = 2$ as an estimate of genomic mutation rate towards damaging replacements, as defined above. We can make several inferences from our estimate of $N \approx 10^3$ (more precisely approximately 2000) deleterious amino acid allelic variants in the genome of the average human individual, assuming that the modern human population is close to deterministic mutation-selection equilibrium. This assumption is justified because we consider only alleles with selection coefficient $>1/N_e$, and the impact of random drift on the frequencies of such alleles is small and they reach mutation-selection equilibrium rapidly.

At mutation-selection equilibrium, for every generation the total decline in the number of deleterious variants (N) due to selection is equal to its increase due to mutation. Contribution of the ith nucleotide site into this decline is $x_i \cdot k_i = u_i$, where $x_i << 1$ is the frequency of deleterious replacement at this site and k_i is the coefficient of selection against heterozygous deleterious replacement (neglecting the elimination of homozygotes; 20). Contribution of the ith site into the increase of N is u_i, the mutation rate towards the deleterious replacement (neglecting back mutations). Thus, $\Sigma x_i \cdot k_i = \Sigma u_i$, where summation is over all sites where mutations cause deleterious amino acid replacements (21,22). Since $\Sigma x_i \cdot k_i = U_r = 2$ and $N = \Sigma x_i = \sim 10^3$, the arithmetic mean $K = (\Sigma x_i \cdot k_i)/(\Sigma x_i)$ of coefficients of selection against deleterious heterozygous SNPs segregating in the human population must be $U_r/N \approx 10^{-3}$. Conversely, the mean persistence time of deleterious SNPs is $N/U_r \approx 1000$ generations, substantially higher than that of mildly deleterious mutations in *Drosophila melanogaster* (21).

Under a realistic assumption that there is no covariance across sites between the mutation rate and the coefficient of selection against mutants (23), $\Sigma(u_i/k_i) = (\Sigma u_i) \times M[1/k_i]$. Since $N = \Sigma x_i = \Sigma(u_i/k_i)$ (the expected value of $x_i = u_i/k_i$) and $\Sigma u_i = U_r$, we can conclude that $1/M[1/k_i] = U_r/N$. In other words, the harmonic mean coefficient of selection against a new damaging non-synonymous cSNP is also approximately 10^{-3}. Of course, the arithmetic mean coefficient of selection against new mutations is higher than that against segregating ones, because strongly deleterious mutations are eliminated first. Since a lot of damaging SNPs, nevertheless, reach high frequencies, the variance in coefficients of selection against them must be high, so that the arithmetic mean is much higher than the harmonic mean.

Functional effect of deleterious amino acid variants

Simple consideration helps to demonstrate that most of the deleterious non-synonymous cSNPs are not associated with complete function loss. In the *Drosophila* genome ~25% of all genes are essential, in the sense that homozygous loss of function of such a gene is lethal (24). If the corresponding number for humans is about the same order of magnitude, a significant fraction of amino acid variants, which lead to function loss, would be recessive lethals. Moreover, if most deleterious amino acid variants could cause the complete loss of function, the average human individual would carry hundreds of recessive lethals. If the number of recessive lethals in the genome of the average individual equals H, a person whose parents are first cousins will be, on average, homozygous by a recessive lethal at $H/32$ loci. The assumption of hundreds of recessive lethals per individual would mean a high number (greater than approximately 10) of these loci.

This is clearly not the case. Fecundity of first cousin marriages is below average but it is very far from being 0. In outbred marriages, >30% of conceptions result in spontaneous abortion (which would be the consequence of even one early acting homozygous recessive lethal in the fetus) (25). The corresponding figure for marriages between first cousins is unknown, but it can hardly be >60–80% (26,27). Thus, an offspring from a marriage between first cousins cannot, on average, carry more than approximately two homozygous recessive lethals. Even currently very rare brother–sister marriages which produce offspring homozygous by a recessive lethal at $H/8$ loci have a reasonable fecundity since they constitute >30% of all marriages in some societies (28).

We can conclude that only a small fraction of deleterious amino acid-altering SNPs segregating in human population lead to total loss of function of the affected protein, and the rest must have relatively mild effects.

MATERIALS AND METHODS

In order to identify SNPs affecting molecular phenotypes, we limited our analysis to amino acid replacement in human proteins with experimentally known 3D structure or proteins, for which 3D structure can be assessed with high confidence by comparison with homologues. Of 74 125 human protein sequences annotated in the SWALL database, 8510 had a link to the Protein Data Bank (PDB) or HSSP (29) databases. PDB is a database which contains the data on protein 3D structures. HSSP is a database of proteins for which 3D structure can be modelled on the basis of sequence similarity. Proteins with <50% sequence identity to a protein with experimentally determined 3D structure were excluded from this data set because of low conservation of structural characteristics (30,31), so that amino acid substitutions in the remaining 5389 proteins were considered for further analysis.

Prediction rules

An amino acid variant is predicted to affect function or structure of the protein if one of the following conditions is satisfied. These conditions are empirical but are based on current knowledge of protein structure, interactions and evolution (14).

(i) The variant is located in a site defined in the SWISS-PROT database as binding site, active site, site involved in a disulphide bond etc. (SWISS-PROT annotations: ACT_SITE, BINDING, MOD_RES, SITE, LIPID, METAL, DISULFID).

(ii) The variant is not compatible with the context of amino acid substitutions at the position in the family of homologous proteins. To quantify this rule we collected homologous sequences for all proteins of our data, which can be reliably aligned using the BLAST (32) software. Only sequences with sequence identity of at least 30% were considered. On the basis of the alignment of these homologous sequences, profile scores (elements of the position-specific substitution matrix) were computed for both allelic variants. In this study we used a new version of the PSIC (33) profile analysis program (http://combio.imb.ac.ru/psic/formprf.html). Profile scores are logarithmic ratios of the likelihood of given amino acid occurring at a particular site to the likelihood of this amino acid occurring at any site (background frequency). A variant is predicted to be damaging if an absolute value of the difference between profile scores of the two amino acid variants is >1.7. In the case of substitutions between species used to estimate the rate of false-positive predictions, it is not clear which amino acid corresponds to an ancestral variant and where in the lineage the substitution occurred. To ensure that the profile is never affected by the substitution to be analysed, we used an absolute value of the difference in profile scores (instead of the difference itself) and discarded all sequences with an identity ≥95%. Due to these precautions, our threshold appears to be very conservative.

(iii) The variant is likely to destroy the hydrophobic core of the protein. Technically, the variant is located in the site with a solvent accessible surface area <25% and difference in accessible surface propensities of the two amino acids >0.75. Accessible surface propensities have been estimated on the independent set of unrelated monomeric proteins with the 3D structure determined with high resolution (34). Accessible surface propensities (knowledge-based hydrophobic 'potentials') are logarithmic ratios of the likelihood of given amino acid occurring at a site with a particular accessibility to the likelihood of this amino acid occurring at any site (background frequency). Accessible surface area data were taken from the HSSP (29).

(iv) The variant in the buried site (with a solvent accessible surface area <25%) displays a change in the electrostatic charge. Substitutions of positively charged residues to histidine were not considered as charge changes.

(v) The variant is predicted to affect solubility of the protein, i.e. it is located in the exposed site (>50% of the accessible surface area) and the difference in accessible surface propensities is >2.
(vi) The variant involves a proline residue in the α-helix. Secondary structure assignments were extracted from the HSSP database.
(vii) The variant was predicted to affect protein–ligand interactions, i.e. the minimal distance between any of the non-water hetero-atoms reported in the PDB file and any atom of the amino acid residue at the variation site is <6 Å, whereas the difference in the profile scores of the two amino acid variants is >1.

Benchmark

In order to estimate the number of false-negative predictions, the method was tested on a set of mutations which are likely to affect the 'molecular phenotype' of the protein (deleterious data set). We extracted all variants annotated in the SWISS-PROT database pointing to a disease phenotype or known effect on function, structure or stability of the protein. Among a total of 1551 amino acid substitutions of this type, 60 were clearly annotated as disease causing and 54 were results of *in vitro* site-directed mutagenesis with a known damaging effect on the protein (Table 1). In many cases false-negatives could be explained by more detailed analysis; an example is given in Figure 2.

The fraction of false-positives can be estimated by the analysis of amino acid substitutions without negative effect on protein structure or function (divergence data set). Using the BLASTP search we have collected a set of substitutions between human proteins and closely related mammalian orthologues. Only isolated substitutions from proteins with at least 95% identity to the human homologue were considered. To ensure that possible paralogues with different function were not included in the analysis, only the best hit per species was taken into account. To exclude possible bias due to particular protein classes, we limited our analysis of amino acid substitutions between human proteins and closely related mammalian homologues to two protein sets: (i) proteins from the set of non-synonymous SNPs; and (ii) proteins from the set of damaging mutations described above. Data on the estimates of the rate of false-positives are given in Table 1.

Extraction of non-synonymous SNP data sets

All allelic variants in the proteins from our data set annotated in HGBASE (11) (release 7.3) and dbSNP (12) (release of May 2000) have been identified via BLASTX search. Synonymous SNPs were not considered. Additional non-synonymous SNPs were extracted via analysis of the VARIANT field in the corresponding SWISS-PROT entries. We have extracted all amino acid variants annotated in the entries with the keyword 'polymorphism' but without the keyphrase 'disease mutations'. In the rest of the entries only variants clearly annotated as polymorphisms were considered.

The 459 non-synonymous SNPs from the resulting data set have been further classified according to the reliability of the SNP discovery method used (Table 2). The following entries were considered as reliable: (i) all HGBASE entries annotated as 'proven'; (ii) dbSNP entries confirmed by direct sequencing or VDA data annotated as 'certain'; and (iii) all entries from SWISS-PROT. Information on allele frequencies was available for 99 variants, 78 of these were classified as reliable.

In some cases allelic variants annotated in the SNP databases have been identified in individuals with genetic disorders. These erroneous annotations probably lead to a slight overestimation of the fraction of deleterious cSNPs, although literature analysis of several randomly chosen entries did not reveal frequent misannotations of this sort, so we believe that the effect is minor.

REFERENCES

1. 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. *Nature Genet.*, **22**, 231–238.
2. 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. *Nature Genet.*, **22**, 239–247.
3. Cambien, F., Poirier, O., Nicaud, V., Hermann, S.M., Mallet, C., Ricard, S., Beague, I., Hallet, V., Blanc, H., Loucaci, V. *et al.* (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.*, **65**, 183–191.
4. Goddard, K.A., Hopkins, P.J., Hall, J.M. and Witte, J.S. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.*, **66**, 216–234.
5. Sunyaev, S.R., Lathe III, W.C., Ramensky, V.E. and Bork, P. (2000) SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. *Trends Genet.*, **16**, 335–337.
6. Collins, F.S., Guyer, M.S. and Charkravarti, A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
7. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
8. Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York, NY.
9. Eanes, W.F. (1999) Analysis of selection on enzyme polymorphisms. *Annu. Rev. Ecol. Syst.*, **30**, 301–326.
10. Kreitman, M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, **304**, 412–417.
11. Brookes, A.J., Lehvaslaiho, H., Sigfried, M., Boehm, J.G., Yuan, Y.P., Sarkar, C.M, Bork, P. and Ortigao, F. (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res.*, **28**, 356–360.
12. Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
13. Sunyaev, S., Ramensky, V. and Bork, P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, **16**, 198–200.
14. Eisenhaber, F. and Bork, P. (1998) Sequence and structure of proteins. In Rehm, H.-J. and Reed, G. (eds), *Biotechnology*. Wiley-VCH, Weinheim Vol. 5a.
15. Harpending, H.C., Batzer, M.A., Gurven, M., Jorde, L.B., Rogers, A.R. and Sherry, S.T. (1998) Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA*, **95**, 1961–1967.
16. Nachman, M.W. (1997) Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics*, **147**, 1303–1316.
17. Li, W.H. and Sadler, L.A. (1991) Low nucleotide diversity in man. *Genetics*, **129**, 513–523.
18. Aparicio, S.A. (2000) How to count human genes. *Nature Genet.*, **25**, 129–130.
19. Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344–347.
20. Crow, J.F. (1979) Minor viability mutants in *Drosophila*. *Genetics*, **92** (suppl.), s165–s172.
21. Crow, J.F. (1970) Genetic loads and the cost of natural selection. In Kojima K.-i. (ed.), *Mathematical Topics in Population Genetics*. Springer, New York, NY, pp. 128–177.

22. Kondrashov, A.S. (1995) Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.*, **175**, 583–594.
23. Morton, N.E., Crow, J.F. and Muller, H.J. (1956) An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl Acad. Sci. USA*, **42**, 855–863.
24. Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., Misra, S. and Rubin, G.M. (1999) The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics*, **153**, 135–177.
25. Wilcox, A.J., Weinberg C.R. and Baird D.D. (1995) Timing of sexual intercourse in relation to ovulation – effects on the probability of conception, survival of the pregnancy, and sex of the baby. *N. Engl. J. Med.*, **333**, 1517–1521.
26. Bittles, A.H. and Neel, J.V. (1994) The costs of human inbreeding and their implications for variations at the DNA level. *Nature Genet.*, **8**, 117–121.
27. Grant, J.C. and Bittles, A.H. (1997) The comparative role of consanguinity in infant and childhood mortality in Pakistan. *Ann. Hum. Genet.*, **61**, 143–149.
28. Scheidel, W. (1997) Brother-sister marriage in Roman Egypt. *J. Biosoc. Sci.*, **29**, 361–371.
29. Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.
30. Flores, T.P., Orengo, C.A., Moss, D.S. and Thornton, J.M. (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.*, **2**, 1811–1826.
31. Russell, R.B. and Barton, G.J. (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.*, **244**, 332–350.
32. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
33. Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G. and Kuznetsov, E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.
34. Sunyaev, S.R., Eisenhaber, F., Argos, P., Kuznetsov, E.N. and Tumanyan, V.G. (1998) Are knowledge-based potentials derived from protein structure sets discriminative with respect to amino acid types? *Proteins*, **31**, 225–246.
35. Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo Jr, R., Ellis, M.C., Fullan, A. *et al.* (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.*, **13**, 399–408.
36. Lucotte, G. **(1998)** Celtic origin of the C282Y mutation of hemochromatosis. *Blood Cells Mol. Dis.*, **24**, 433–438.