doi:10.1016/j.jmb.2005.12.025 J. Mol. Biol. (2006) 356, 1263–1274

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Identification and Analysis of Deleterious Human SNPs

Peng Yue[1,2] and John Moult[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville MD 20850, USA*

[2]*Molecular and Cellular Biology Program, University of Maryland, College Park MD 20742, USA*

**Corresponding author*

We have developed two methods of identifying which non-synonomous single base changes have a deleterious effect on protein function *in vivo*. One method, described elsewhere, analyzes the effect of the resulting amino acid change on protein stability, utilizing structural information. The other method, introduced here, makes use of the conservation and type of residues observed at a base change position within a protein family. A machine learning technique, the support vector machine, is trained on single amino acid changes that cause monogenic disease, with a control set of amino acid changes fixed between species.

Both methods are used to identify deleterious single nucleotide polymorphisms (SNPs) in the human population. After carefully controlling for errors, we find that approximately one quarter of known non-synonymous SNPs are deleterious by these criteria, providing a set of possible contributors to human complex disease traits.

Keywords: single nucleotide polymorphisms (SNPs); monogenic disease; human disease; complex traits; support vector machine

Introduction

Knowledge of the human genome sequence,[1,2] together with a large number of single nucleotide polymorphisms (SNPs) present in the human population[3–6] opens the way for the development of a detailed understanding of the mechanisms by which genetic variation results in phenotype variation. In particular, it should now be possible to identify the contribution of SNPs to human disease. It is estimated that the human population has approximately one SNP with a frequency of more than 1% every 290 base-pairs, implying a total of about ten million.[7] Missense SNPs, resulting in an amino acid change in a protein, are most accessible to analysis. There are an average of about four coding region SNPs per gene, half of which are non-synonymous or missense SNPs,[8,9] implying a total of about 50,000.

Genetic mapping, especially linkage analysis,[10] has successfully mapped more than 1000 human inheritable diseases to genes. Most of those are rare monogenic (one gene/one trait) diseases, following a simple Mendelian inheritance pattern. On the other hand, common human diseases, such as hypertension, diabetes, Alzheimer, stroke, and heart disease, which follow a more complicated inheritance pattern, are proving much harder to analyze.[11–13] Difficulties are caused by incomplete penetrance (a person carrying a predisposing allele may not exhibit the disease phenotype); genetic heterogeneity (mutations on one of several genes may result in identical phenotypes); and polygenic inheritance (a trait is controlled by multiple gene interactions, such that each individual predisposing allele has a low risk factor and shows weak correlation with the disease trait). In addition, environmental factors may also play an important role in shaping disease phenotypes. Genetic mapping does not provide direct insight into the relationship between the presence of a SNP and susceptibility to a particular disease. There are a variety of mechanisms that may be involved, including effects of transcription rate, protein folding, protein function and protein half-life.

Here, we analyze non-synonymous or missense SNPs, that is, those which change an amino acid, and so may affect protein folding, function, or half-life. More than half of monogenic disease causes are these single amino acid substitutions.[10] We use two methods to identify which missense non-synonymous SNPs (nsSNPs) are deleterious to protein function. Both methods have been developed and tested on amino acid changes causative of monogenic disease, and a control set

Abbreviations used: SNP, single nucleotide polymorphism; nsSNP, non-synonymous SNP; SVM, support vector machine; ROC, receiver operating characteristic; HGMD, Human Gene Mutation Database.

E-mail address of the corresponding author: moult@umbi.umd.edu

of single residue changes fixed between closely related mammalian species.[14] One method analyzes the impact of amino acid changes on protein stability, making use of the three-dimensional structural environment.[15] We find the majority of single base changes that cause monogenic disease significantly destabilize the folded state of the protein concerned. The second method, reported here, makes use of the tendency of critical amino acids to be conserved with a protein family. The more conserved and restricted the type of amino acid at a position, the more likely that a substitution not consistent with that pattern will have a deleterious impact on protein function. This method is more general than the stability model, including all types of protein level effect. It is also more widely applicable, since it does not require knowledge of three-dimensional structure. On the other hand, it provides less direct insight into the mechanism by which a missense SNP affects protein function. The principles of sequence conservation methods have also been explored by others.[14,16–21] We have used a machine learning method, the support vector machine, trained on five simple features that capture the relative sequence conservation at each position in a multiple sequence alignment. The support vector machine allows the identification of a subset of high confidence predictions. Both methods are carefully benchmarked. The use of two separate methods provides an additional means of assessing the reliability of the conclusions.

The two methods have been used to analyze sets of non-synonymous SNPs found in the human population, extracted from the dbSNP database,[4] and a subset of those for which population frequency data are available. The subset are data from Perlegen[5] and the Hapmap consortium.[6] Using stringent criteria, we find that about one quarter of these SNPs are classified as deleterious at the same level as those causing monogeneic disease in other genes. These are very likely to have a significant impact on protein function, and so probably contribute to complex disease traits, and provide a basis for prioritization in association studies.

We have also examined a number of aspects of the relationship between monogenic disease genes and the rest. First, we have compared the occurrence of deleterious SNPs in monogenic *versus* non-monogenic disease genes. We find that, whereas in monogenic disease genes nearly all deleterious SNPs occur at low frequency in the population, in other genes a larger proportion are found at high frequencies, consistent with the idea that the effect of deleterious SNPs in other genes is buffered. Second, we have looked at the rate of sequence divergence of monogenic *versus* other genes. An interesting variation with conservation level is found. Third, we have found that there is a correlation between the phenotypic impact of mouse knockouts and whether or not the orthologous human gene is implicated in monogenic disease. Finally, we have checked to see if monogenic disease genes are less likely to have paralogs than the others, exploring the idea that paralogs sometimes can provide substitute function. No such effect was found.

Results

Training and testing data used for the classification methods

Table 1 summarizes the monogenic disease and control datasets used for training and benchmarking the sequence profile and structure stability methods. There were a total of 10,263 deleterious mutants in 731 proteins and 16,682 control substitutions in 348 proteins available. The profile model includes 92% and 71% of these respectively, since profiles can be built for most proteins. In testing, high confidence (HC, SVM score $>|0.5|$) classifications were obtained for over 80% of these. Significantly fewer data (37% and 14%, respectively) are included in the stability model, because of low structural coverage of human proteins. High confidence classifications are again obtained for about 80% of cases. The last two rows show the data for cases where both methods could be applied. The fraction of high confidence predictions is similar.

Table 1. Training and testing data for the profile and stability methods

	Deleterious mutants			Control substitutions		
	Number	(%)	Proteins	Number	(%)	Proteins
All data	10,263	100	731	16,682	100	348
Profile	9468	92	693	11,778	71	336
Profile HC	7986	78	673	10,171	61	336
Stability	3768	37	243	2309	14	153
Stability HC	3046	30	229	1904	11	152
Profile+stability	3641	35	235	2141	13	148
Profile+stability HC	2501	24	216	1498	9	146

Deleterious mutants are amino acid changes that cause monogenic disease.[10] Control substitutions are amino acid differences between human proteins and closely related orthologs. HC are high confidence classifications from the support vector machine. Proteins are the number of proteins from which data are included.

Accuracy of the classification methods

Figure 1 shows the false positive (blue bars) and false negative rate (red bars) for both methods separately, on all data and for just the high confidence classifications (a support vector machine (SVM) score of greater than 0.5 for non-deleterious classifications and less than −0.5 for deleterious ones), as well as the corresponding data for the cases where the two methods agree. Bootstrapping, with 30 SVMs for each method, was used to obtain the accuracies and confidence limits. As expected, the false positive and false negative rates are highest for the individual classification methods, lower when only high confidence classifications are considered, and lowest of all when only high confidence classifications shared by both methods are included (3% false positive, 9% false negative). The false negative rate of the profile method is slightly lower than that of the stability method (20% *versus* 26% for all classifications, 16% *versus* 21% for high confidence ones, where the latter include 85% and 80% of the data, respectively). This difference is expected, since the profile method includes all effects on protein function at the amino acid level, including ligand binding, catalysis, allosteric mechanisms, and post-translational modifications, as well as stability and folding effects, whereas the stability model includes only stability and folding contributions. Less expected is the lower false positive rate for the profile method (9% *versus* 17% overall, 6% *versus* 12% for high confidence classifications). The balance between false positive and false negative rates is determined by the relative weights given to the deleterious and control datasets in training the SVM. Equal weights, taking into account the differences in data set sizes, were used. A 9/10 weighting of control to disease sets produces approximately equal false positive and false negative rates of about 17%.

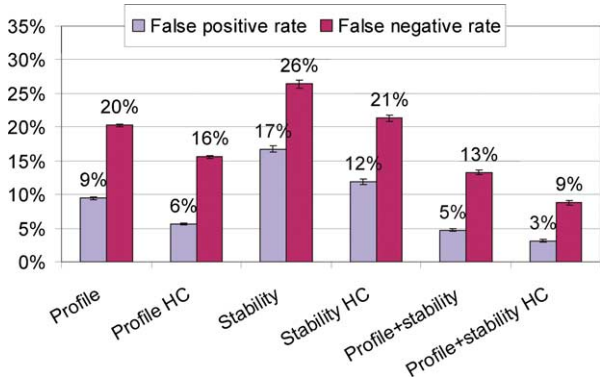

Figure 1. Evaluation of the profile and stability methods. False positive and false negative rates are shown for the two methods alone, and for cases where both can be applied and the classifications agree. Results are shown for all classifications, and for the high confidence subsets (HC, SVM score $> |0.5|$). Higher false negative rates for the stability model reflect the fact that only stability and folding effects are included, whereas the profile model includes all effects on protein function *in vivo*. Bars indicate 95% confidence limits, obtained from 30 bootstrapped runs.

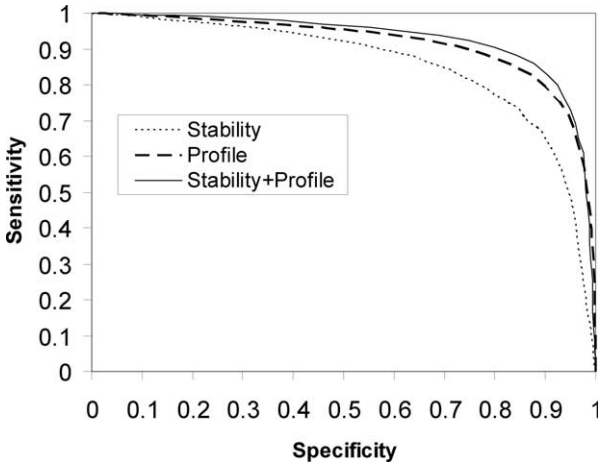

Figure 2. ROC curve showing the relationship between specificity (fraction of control set data correctly classified) and sensitivity (fraction of disease set data correctly classified), for the stability model (dotted line), the profile model (broken) and a combined method (continuous line). The profile method performs noticeably better than the stability model, and the combined method is slightly superior to the profile model alone.

Figure 2 shows a ROC (receiver operating characteristic) curve analysis of the relationship between specificity (fraction of control data correctly classified) and sensitivity (fraction of disease data correctly classified). Specificity and sensitivity were calculated for a series of SVM score thresholds, using the subset of data where both methods can be applied. Values for the combined method were obtained using the sum of the SVM scores for the two independent methods. This analysis shows that the profile model has better performance at most thresholds, and that the combined method provides a slight advantage over the profile method alone.

Causes of error

For both methods, the finite error rates reflect both the effects of approximations in the methods and the nature of the data sets. The stability method incorporates a number of approximations in modeling the structure of mutants, and uses a scenario based analysis of effects on stability.[15] As discussed later, for the profile method, the effect of a limited number of sequences in a profile is the main approximation. The Human Gene Mutation Database (HGMD) data[10] used as a disease set contains some entries that are not strictly causative of monogenic disease. For example, the mutant G15D in the α-chain of hemoglobin (HBA1) is in HGMD, but is predicted to be non-disease causing, with a confident SVM score of 2.9. The literature on

this mutation[22] gives no indication of disease. Since 1999, HGMD have added some mutants that are disease associated or risk polymorphisms. This work uses the HGMD version of 26 April 2002, which includes 152 mutants annotated as not necessarily causative of disease. The false negative rate for these is very high: 62% for the profile method and 73% for the stability method. The assumption of no deleterious effects fixed between species might contribute to a finite false positive rate. A limited check on that possibility is provided by 41 HGMD mutants where the altered amino acid is the wild-type in another species. Of the 37 classified by the profile model only four are found to be deleterious. Thus, these appear to be mostly non-disease entries in HGMD, rather than deleterious mutants fixed in other species. Overall, the data in HGMD are of high quality, and appropriate for this application. Errors in the models and the data are sufficiently small that firm conclusions about the level of deleterious SNPs in the human population can be reached, as described below.

Sensitivity to the number of sequences in a profile

The reliability of the PSSM and entropy values used in the profile method depends on the size of the sequence alignment. We examined the accuracy of the method as a function of the number of sequences available, after filtering out redundant and less reliably aligned sequences, as described in Materials and Methods. Profiles were divided into sets with different numbers of sequences, and the accuracy evaluated for each set. Table 2 shows the results. All sets have similar accuracy, except the set with the smallest number of sequences (2–9). This group has a similar false negative rate but a higher false positive rate (31%) than the other groups. The high false positive rate is probably a consequence of the low maximum entropy for a small number of sequences: the maximum for two sequences is approximately 1 bit, while for 20 sequences, it is 4.3 bits. Additionally, for small profiles, the PSSM is dominated by the BLOSUM scores rather than the pattern of residue use.

Comparison between BLOSUM, PSSM, and profile models

The full profile method includes the PSSM for the aligned sequences, and entropy factors. We compared the performance of a PSSM[23] alone, which takes into account which residues are observed at each position in a sequence alignment, with performance using an average substitution matrix (BLOSUM[24]), which considers only the likelihood of the substitution in all proteins at all positions. It has been suggested that the BLOSUM matrix is suitable for use in identifying damaging nsSNPs.[9,25] Since a PSSM contains information unique to each sequence family and sequence position, we would expect it to produce more accurate classifications.

BLOSUM 45 and BLOSUM 62, representing average substitution preferences between proteins with different levels of sequence identity, were tested. PSSM and BLOSUM method accuracy as a function of a score threshold were examined, and the threshold returning the lowest sum of false positives and false negatives chosen in each case. The results are shown in Table 3, with the full profile method included for comparison. The BLOSUM matrices both yield similar false positive and negative rates, of about 27% and 36%, respectively, whereas the PSSM has significantly lower values of 22% and 28%. The profile model is substantially more accurate than the PSSM alone, with false positive and false negative rates of 9% and 20%, respectively, establishing that the entropy terms do provide significant additional information.

Analysis of population SNPs: approximately a quarter of non-synonymous population SNPs are deleterious

We now use the profile and stability methods to identify deleterious non-synonymous SNPs in the human population. As described in Materials and Methods, nsSNP data were obtained from three sources: the NCBI dbSNP database,[4] the Perlegen data,[5] and the Hapmap project results.[6] dbSNP contains a wide range of data, some of which are

Table 2. Accuracy of the profile method as a function of the number of sequences in the alignment

	Deleterious			Control		
No. of sequences	FN	Proteins	Number	FP	Proteins	Number
[2–9]	0.18	60	500	0.31	16	787
[10–19]	0.17	82	1073	0.14	24	1296
[20–39]	0.18	167	1957	0.11	85	2785
[40–59]	0.22	121	1871	0.13	66	2263
[60–79]	0.19	94	804	0.10	48	1578
[>=80]	0.18	169	3263	0.10	97	3069

Accuracy is measured in terms of the false negative rate (FN) and the false positive rate (FP). The Number columns show the number of variants analyzed in each alignment size range, and Proteins are the number of human proteins included. Accuracy is approximately equal in all but the smallest alignment range, where there is a sharp rise in the false positive rate.

Table 3. Comparison of classification accuracy of BLOSUM matrices, a PSSM and the full profile method

	False positive rate (%)	False negative rate (%)
BSOSUM 45	27	38
BLOSUM 62	28	36
PSSM	22	28
Profile model	9	20

The PSSM method has substantially lower false positive and false negative rates than obtained with either BLOSUM matrix. The additional entropy information in the full profile model further improves accuracy.

based on a single observation. Both Perlegen and the Hapmap project have genotyped sets of individuals from several different populations. Since these SNPs are all verified, and have associated population frequency information, we have analyzed them as a separate data set, referred to as the Frequency set. Table 4 shows the number of data available in the full dbSNP set and the Frequency set, and the number of data that can be classified by the stability and profile methods, the combined methods, and the number of high confidence classifications in each case.

Figure 3 shows the fraction of population SNPs assigned as deleterious in dbSNP (blue bars) and the Perlegen/Hapmap data (purple bars). Results are again shown for the two methods separately, and for the combined methods, for all classifications, and those of high confidence. Deleterious classifications in both SNP sets are lowest for the most stringent conditions (high confidence classifications for the combined methods), with 33% for all dbSNP data and 17% for the Frequency subset. The highest deleterious rates are for the stability model alone, with 40% for the dbSNP data, and 31% for the Frequency subset. Deleterious SNP rates are consistently substantially lower for the Frequency subset than the full dbSNP set, presumably reflecting the effect of the unreliable single observation component in dbSNP. As a control, we also analyzed the 952 Hapmap SNPs which were found to have zero frequency, that is, are in dbSNP, but were not observed in the Hapmap population. The profile method classifies 50% of those SNPs as deleterious, a much higher value, and close to that obtained in tests introducing random mutations.

The deleterious population SNP rates in Figure 3 are distorted somewhat by the finite false positive and false negative rates of the classification methods. Distortions can occur in both directions: a high false positive rate contributes to over-estimating the deleterious SNP level, but a high false negative rate contributes to an underestimate. We correct for these effects as follows: For a given true deleterious rate D_{true}, with a false positive rate f_{p} and false negative rate f_{n}, the expected apparent deleterious rate D_{exp} is:

$$D_{\text{exp}} = D_{\text{true}} - D_{\text{true}} * f_{\text{n}} + [1 - D_{\text{true}}] * f_{\text{p}}$$

where the second term ($D_{\text{true}}{}^{*}f_{\text{n}}$) is the underestimate effect of false negatives, the third ($[1-D_{\text{true}}]^{*}f_{\text{p}}$) is the over-estimate effect of false positives. The most probable value of D_{true} is thus:

$$D_{\text{true}} = (D_{\text{exp}} - f_{\text{p}})/(1 - f_{\text{p}} - f_{\text{n}})$$

A set of apparent deleterious rates (D_{exp}) were obtained using each of the 30 SVM profile and 30 stability model SVMs, and D_{true} values estimated for each, using the corresponding f_{p} and f_{n} values. Average values of D_{true} were then calculated, together with 95% confidence limits.

Figure 4 shows the estimated true deleterious rates for each of the method conditions, using the frequency subset. The stability model and the profile model on all data both return values of approximately 25%. Slightly lower values were obtained with the high confidence subsets, and the lowest value (15%) was obtained with the high confidence assignments common to both methods. It is expected that high confidence scores are only obtained for the more severe effects on protein function and stability. Application of the stability method to site-directed mutagenesis data where experimental folding free energies are available confirms that on average high confidence assignments have a more severe effect on protein stability (data not shown). Thus, the lower level of deleterious SNPs found for the high confidence

Table 4. Data used for identifying deleterious human SNPs

	All			Frequency set		
	Number SNPs	SNPs (%)	Number genes	Number SNPs	SNPs (%)	Number genes
dbSNP build 124	50772	100	15,710	10,403	100	6316
Profile	29,081	57	11,129	6377	61	4297
Profile HC	22,067	43	9782	4911	47	3549
Stability	5166	10	2019	885	9	624
Stability HC	3960	8	1776	681	7	509
Profile+stability	3150	6	1512	531	5	415
Profile+stability HC	2096	4	1180	370	4	304

The top line shows the number of missense SNPs available in the dbSNP database, and the subset of these with population frequency information, from Perlegen and the Hapman project. Classifications were made on the full set and the frequency set. The number of SNPs classified in each case, and the number of genes are given for the profile method, the stability method and the combined methods. In each case, values are given the full data and for the subset that are classified with high confidence (SVM score $>|0.5|$).

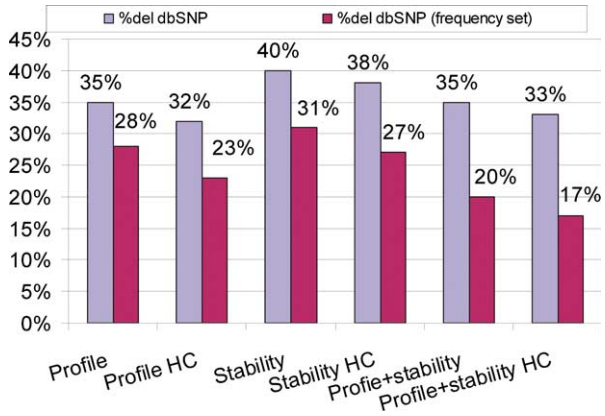

Figure 3. Estimated fraction of deleterious SNPs in the human population. Results are shown for all missense SNPs in dbSNP build 124 (blue bars), and a subset for which there are population frequency data (purple bars). Deleterious rates are calculated using the profile and stability methods, the two methods combined, and also, in each case, for high confidence (HC, SVM score $> |0.5|$) classifications only. Consistently lower rates are found for the frequency subset than for all dbSNP data, partly reflecting the effect of incorrect entries in the latter. Variations in the rate for the different classification methods reflect the differing false positive and false negative levels. Lower rates for the high confidence predictions reflect the fact that these are generally obtained only for more severe effects on protein structure and function.

score subsets are an estimate of the fraction of more severely deleterious SNPs in the population. The best estimate of the fraction of population missense SNPs that are as detrimental to protein function as those found for monogenic disease is provided by the full set of classifications for the profile and stability methods. For both, that value is close to 25%. Thus, the analysis leads to the conclusion that approximately one quarter of non-synonymous SNPs found in the population are as deleterious to protein function as single base changes known to cause monogenic disease. This value is a little lower than reports by other groups,[17,18] probably because of the effect of correcting for finite error rates in the methods.

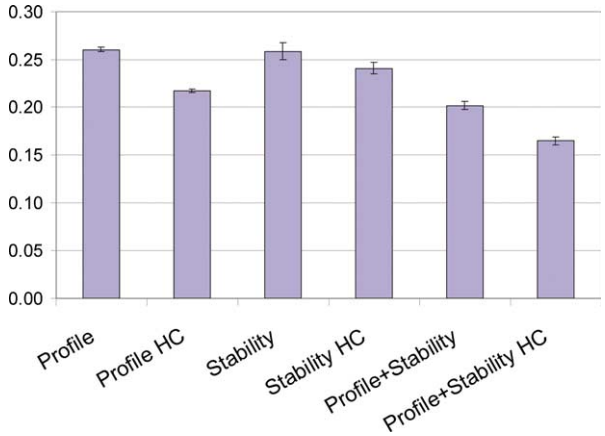

Figure 4. Estimated fraction of deleterious SNPs in the human population, corrected for the effects of false negatives and false positives. The Stability, Profile and combined methods all yield rates close to 25%. The high confidence classifications yield lower values, reflecting the fact that generally only severe effects on protein structure and function have high confidence classifications. Data are for the Frequency subset of dbSNP. The bars show 95% confidence limits.

Deleterious SNPs in monogenic disease genes

There are 4458 nsSNPs in dbSNP located in monogenic disease genes, among which 1656 are assigned as deleterious by the profile method. Only a small portion (152) is also present in HGMD as known monogenic disease mutants. The remainder might be new monogenic disease causing variants, known variants not yet entered into the HGMD, or false positives. Given a false positive rate of 10%, we only expect 446 in that category. If the additional SNPs really are disease causing, we would expect them to be predominantly at low frequencies in the population. Figure 5 shows a comparison of the population frequency distribution of the 970 of these in the frequency subset with the corresponding distribution for all other genes. As expected, there are many more low frequency SNPs in both sets. Both sets also show a higher fraction of deleterious SNPs at low frequency, compared to non-deleterious, consistent with their being selected against. That bias is stronger for the monogenic disease gene set, and only about 10% are at frequencies higher than 20%, the expected fraction of false positives. More precisely, there are 170 disease gene deleterious SNPs with a frequency of less than 5%, *versus* 723 such SNPs in non-disease genes, and 78 higher frequency deleterious SNPs in

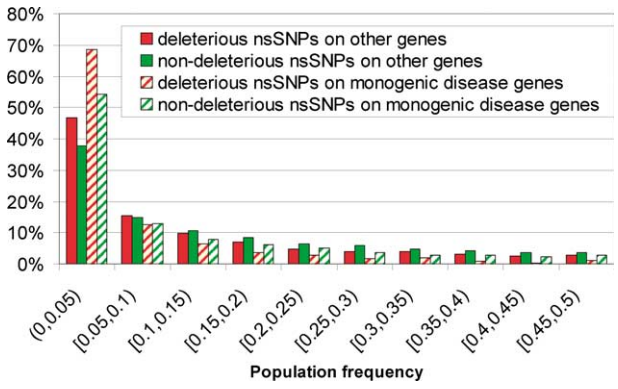

Figure 5. Distribution of SNP frequencies in the human population. Filled red bars show the fraction of all deleterious missense SNPs in each frequency range, for all non-monogenic disease genes. The hashed red bars show the same data for monogenic disease genes; green bars show the corresponding data for non-deleterious SNPs. As expected low frequency SNPs are the most common for all categories. Deleterious SNPs are biased towards low frequencies in both sets, but the effect is considerably stronger for monogenic disease genes.

monogenic disease genes *versus* 827 in the others. A χ^2 test on these values shows that the probability of the excess low frequency deleterious SNPs in monogenic disease genes occurring by chance is 2.3×10^{-10}. Thus, there is more selection against deleterious SNPs in the monogenic disease genes. Many of the low frequency disease gene deleterious SNPs are likely involved in monogenic disease. As noted above, some may have already been known, but have not been entered into the HGMD. Others are likely previously unrecognized contributors to monogenic disease.

To investigate the possibility that some of the additional deleterious missense SNPs in monogenic disease genes are in fact disease causing, we examined the subset of 18, in 15 different genes, which are assigned as deleterious with high confidence by both classification methods. Table 5 summarizes the data for these SNPs. Five are already in the HGMD, but given the very low false positive rate for this subset (3%), the others are candidate mutants for monogenic disease. Two of these have surprisingly high population frequencies for monogenic disease mutants: SERPINA7 L303F, at 20%; and AMACR G175D with a frequency of 34%. SERPINA7 belongs to a family of serine protease inhibitors, but also functions as a thyroid binding–globulin (TBG). There are many mutations associated with TBG deficiency, and many of these also have a high population frequency.[26,27] These mutants alone are not sufficient to cause disease, since the resulting tendency for hyperthyroid is usually reversed by reduced thyroid hormone secretion. The high frequency is thus likely a consequence of a second factor being required for disease. There is no obvious explanation for the high frequency of the AMACR SNP.

Divergence rates of monogenic disease-associated proteins

Figure 6 shows a comparison of divergence rates of monogenic disease proteins *versus* all others. A larger proportion of the most highly conserved proteins are non-disease, whereas at moderate to high conservation, a higher proportion is disease. At the lower conservation levels, non-disease proteins are slightly more common: There are 278 disease genes and 3374 others with sequence identities less than or equal to 75%, and 1258 disease genes and 12,337 others with identities higher than 75%. A χ^2 test yields a probability of 0.0022 of this bias occurring by chance. This pattern can be rationalized as follows. Damage to the most conserved proteins is more likely to be lethal, and thus, not identified as disease causing. The lowest conserved proteins are likely buffered against deleterious changes in some way, and so are also not involved in monogenic disease. It is the more moderately to highly conserved genes where deleterious SNPs are likely to lead to disease, but not to be lethal. Other reports,[28,29] using only average values, and separately analyzed K_s and K_a rates, come to contradictory conclusions. With more genomes becoming available, further study will be worthwhile.

Comparison with mouse knockout data

The profile and stability models detect SNPs that reduce the level of protein function *in vivo*. The limit of reduced function is the absence of the gene. Thus, we would expect a relationship between the response of the human phenotype to deleterious SNPs, and the response of mice to knockout of the

Table 5. Very high confidence classifications of deleterious population SNPs in monogenic disease genes

Gene	SNP_ID	SVM stability	SVM Profile	Substitution	Freq.	Source	Population	HGMD
CFTR	766874	−0.88	−1.75	S605F	0.002	Hapmap	afr,eur,chn,jap	
FCER1A	2298805	−0.73	−2.67	S101N	0.007	Perlegen	afr,eur,chn	
NTRK1	6336	−1.06	−1.17	H604Y	0.011	Hapmap	afr,eur,chn,jap	CM990977
DNASE1	1799891	−0.54	−0.77	P154A	0.011	Hapmap	afr,chn,jap	
CFTR	1800100	−1.06	−2.12	R668C	0.014	Perlegen	afr,eur,chn	CM950247
LYZ	1800973	−0.92	−0.72	T88N	0.015	Hapmap	afr,eur,chn,jap	
CHAT	8178990	−0.82	−1.26	L125F	0.021	Hapmap	afr,eur,chn,jap	
EPX	2302311	−1.32	−0.81	M572Y	0.027	Hapmap	afr,eur,chn,jap	
HFE	1800562	−1.00	−1.77	C194Y	0.028	Perlegen	afr,eur,chn	CM960828
TAP1	1057149	−0.80	−1.94	R708Q	0.029	Perlegen	afr,eur,chn	
CYP2A6	17791931	−0.81	−1.99	L160H	0.035	Perlegen	afr,eur,chn	CM980517
KLK3	17632542	−0.58	−1.67	I179T	0.036	Perlegen	afr,eur,chn	
PTGS2	5272	−1.40	−1.42	E488G	0.056	Hapmap	afr,eur,chn,jap	
HFE	1799945	−0.70	−0.66	H63D	0.085	Hapmap	afr,eur,chn,jap	CM960827
CYP2A6	5031017	−1.57	−1.61	G479V	0.125	Hapmap	afr	
OTOR	6135876	−0.91	−0.96	L31P	0.141	Perlegen	afr,eur,chn	
SERPINA7	1804495	−1.28	−1.38	L303F	0.203	Perlegen	afr,eur,chn	
AMACR	10941112	−1.51	−2.35	G175D	0.341	Hapmap	afr,eur,chn,jap	

SVM stability and SVM profile are the scores assigned by the two classification methods. A score < -0.5 is high confidence. The Freq. column gives the mean frequency of each SNP over the populations. The Population column lists the populations in which each SNP has been genotyped: afr, African; eur, European; chn, Chinese, jap, Japanese populations. Only five of these SNPs are in the HGMD database of disease causing mutations (IDs in the last column).

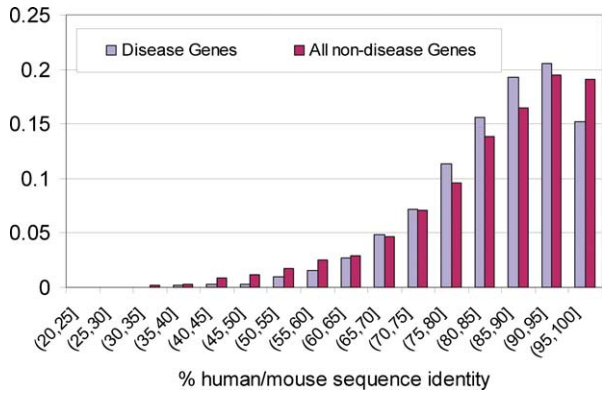

Figure 6. Protein sequence divergence rates for human monogenic disease proteins (blue bars) and all others (purple bars). Rates are expressed in terms of the sequence identity between each human protein and its mouse ortholog. Disease proteins have a larger proportion of high sequence identity mouse orthologs showing that, on average, their sequences diverge more slowly than those of other proteins.

corresponding orthologs.†. In this database, genes are clustered into four knockout phenotype groups. The first group is of genes where the knockout is compatible with viability. This group is further subdivided into cases where there is a detectable effect on the phenotype, and cases where the phenotype is apparently unaffected. The other three groups are of genes where knockout causes post-, peri- and prenatal mortality.

Table 6 shows the fraction of monogenic disease genes in each of the mouse knockout groups. The lowest fraction of monogenic genes is for the no-effect group of knockouts (8%), consistent with fully buffered genes generally not contributed to monogenic disease. The next lowest fraction is for the prenatal mortality set (28%), consistent with defects in these human genes probably resulting in a non-viable fetus, and so not classified as disease associated. Approximately half of the other knockout groups have equivalent monogenic disease genes, consistent with non-lethal but significant impact on the phenotype often being classified as monogenic disease. In all, there are nine disease genes in the two categories not consistent with monogenic disease (no effect and prenatal lethal), and 33 non-disease genes. For the categories consistent with monogenic disease (affected phenotype and post-natal lethal) the corresponding numbers are 93 disease genes and 105 others. A χ^2 test yields a low probability of the correlation occurring by chance (0.004), but the correlation is not as high as might be expected. There are several possible reasons for that. As more mouse knockout data becomes available, a fuller analysis will be possible.

† Mouse knockout data were obtained from http://www.bioscience.org/knockout/knochome.htm

Table 6. Relationship between mouse knockout phenotypes and human monogenic disease genes

Phenotype		Total genes	Disease genes	Fraction genes(%)
Compatible with viability	No apparent effect	13	1	8
	With effect	147	71	48
Postnatal or perinatal mortality		51	22	43
Prenatal mortality	.	29	8	28

Total genes are the number of mouse knockouts in each phenotype category, and Disease genes are the number for which the human ortholog is a monogenic disease gene.

Frequency of paralogs for monogenic disease and other genes

A possible distinguishing feature between monogenic disease genes and the rest is that the phenotype is robust to reduced function on the latter because of redundancy of function, other genes can at least partly compensate for reduced activity. Full identification of possible substitute genes requires a detailed knowledge of human protein networks, not yet available. However, it might be expected that paralogs would often perform this role, and a number of such cases are known. For example, E-selectin and P-selectin are paralogous, with 40% protein sequence identity. Single gene knock-out mice show mild phenotypes, while the double knock-out mice have a severe disease phenotype, consistent with overlapping function.[30] On the other hand, there are many cases where paralogous genes are involved in different biological processes, for example malate and lactate dehydrogenases.

Paralogs were identified by searching each human protein sequence against all others, selecting relatives with a BLAST *E*-score of 10^{-3} or better. Table 7 shows the fraction of monogenic and other genes with at least one paralog. There is no difference between the two types of gene; in both cases about 87% have paralogs. We conclude from this that buffering mechanisms are more varied than just the use of paralogs.

Discussion

The main conclusion of this study is that about one quarter of the known missense SNPs in the human population are significantly deleterious to protein function *in vivo*. Others have reported a figure of about one third.[17,18] It has also been suggested that the fraction is much lower,[16] with false positives, errors in dbSNP, and known monogenic disease mutations inflating the apparent value. We have taken into account the effect of false positives and false negatives to obtain a corrected value for deleterious rate. We have also examined

Table 7. Fraction of monogenic and other genes that have paralogs

	Monogenic disease genes		Other genes	
	Count	(%)	Count	(%)
No paralogs	227	13	705	13
Paralogs	1460	87	4887	87

Monogenic disease data are from HGMD.[10] Other genes are other human genes containing at least one SNP classified as deleterious. There is no difference in the fraction with paralogs for the two sets, suggesting that other mechanisms are dominant in shielding the phenotype from the adverse effects of deleterious SNPs in the non-monogenic disease genes.

the difference in apparent deleterious rate for all of dbSNP and a validated subset. There is indeed a higher value of about one third for all dbSNP, but the value of a quarter is obtained on reliable data. Some of the deleterious SNPs are in known monogenic disease genes, but about 80% of the dbSNP ones, and 70% of the validated set, are not.

Some of the new deleterious SNPs in monogenic disease genes are candidates for previously unrecognized disease causes. The deleterious SNPs in non-monogenic disease genes are candidates for contributing to complex disease traits. Presumably, the network environment of the proteins concerned buffers the effect on the phenotype. This view is supported by the analysis of the relationship between monogenic disease genes and mouse knockout phenotypes; knockouts with intermediate impact on the phenotype are more likely to be orthologs of human monogenic disease genes. A simple form of buffering is overlapping function with paralogous proteins. For example, a T cell mediated immune response will involve many different T-cell receptors. We have found deleterious SNPs in some of these proteins,[31] but redundancy through paralogs will provide buffering. Surprisingly, we did not find that monogenic disease genes are less likely to have paralogs than others, so this mechanism is probably only one of a number. A proper understanding of these buffering processes will require a detailed knowledge of the relationship between protein function and network behavior.

Many of the deleterious SNPs in non-monogenic disease genes are relatively rare. In one sense, this is expected, since overall, there are many more rare SNPs than common ones. The low frequency deleterious SNPs may contribute to relatively rare complex traits, or they may contribute in many combinations to produce common traits.[32,33]

For complex diseases, variation in a single gene only marginally increases risk, and as a consequence, most association studies present weak and sometimes inconsistent results.[34] The deleterious SNPs found in this and other analyses provide additional information that can be used to select SNPs for inclusion in association studies, or, in larger scale studies, to provide prior probabilities that can be incorporated into the statistical model.

The analysis of human SNPs was done using a previously developed method, based on protein structure/stability factors,[15] and a new, sequence profile based method. The sequence method has a larger coverage of missense SNPs because it does not require knowledge of three-dimensional structure. Also, since sequence methods are based on evolutionary selection information extracted from multiple sequence alignments, they are not limited by current knowledge of protein function and structure, and so include a wider range of effects. On the other hand, the sequence method assumes that deleterious SNPs will eventually be removed during evolution. While this assumption may be true for those genes associated with monogenic disease or serving as major contributors to complex diseases, it may not be as true for those with only subtle effects on the phenotype. For this reason, it is desirable to develop broadly based mechanistic models of SNP impact†.

Materials and Methods

Construction of the deleterious variant dataset

The deleterious variants are a set of single amino acid substitutions known to cause monogenic disease. Genes associated with monogenic disease were identified by checking all 16,220 human gene names in the NCBI Locuslink[35] database (as of 26 April 2002) against the Human Gene Mutation Database[10] (HGMD) (as of 9 February 2002). HGMD contains the most comprehensive collection of mutations related to monogenic disease. Most cause monogenic disease, although a few may be associated with disease as a result of linkage disequilibrium rather than directly causative, or contribute to a complex trait disease. Later versions of HGMD include more of the latter class, and so the earlier version was preferred. A total of 731 genes containing 10,263 single residue variations were identified.

Identification of a set of non-deleterious single residue variants

We also required a control set of mutants, not causative of disease. It is not known which base variants in the human population contribute to complex trait disease, and so it is not possible to use these. Following others,[14] we used non-synonymous base differences between human proteins and closely related proteins in other mammals. The justification here is that almost all variants that are fixed between species are essentially neutral and non-deleterious. To maintain compatibility between the deleterious and control sets, the same 731 monogenic disease proteins were used. The protein sequences of these genes were compared to all other mammalian protein sequences in Swiss-Prot,[36] using BLAST.[23] Proteins with at least 90% sequence identity over at least 80% of the full length were selected. Single residue differences in these alignments were used as a set of pseudo mutations. A total of 348

† http://www.snps3d.org

proteins containing 16,682 such single-residue differences to the human disease set were obtained.

Source of human population missense SNPS

SNPs were obtained from NCBI dbSNP, build 124. Many of the dbSNP entries are not verified (are based on single observations, or population frequency data have not been deposited). A confirmed SNP set was built from data in Perlegen (as of May 2005) and the Haplotype genotyping projects (Phase I, as of May 2005). Files containing SNP and frequency information were downloaded from Perlegen and Hapmap project websites‡. These two datasets were processed as follows. (1) Both datasets were mapped to dbSNP RefSNP clusters. Hapmap provides a link from each record to a RefSNP ID; the Perlegen submission SNP ID and the mapping table, SNPSubSNPLink, between submission SNP IDs and RefSNP IDs in dbSNP build 124 were used to link each Perlegen record to the related RefSNP cluster. (2) For each RefSNP entry, mean frequencies were calculated from the three Perlegen populations, and from the available Hapmap populations. (3) In cases where data are available for both sources, the Hapmap information was discarded. dbSNP links were used to map each SNP to the corresponding amino acid substitution.

Construction of sequence profiles

Each human protein sequence was searched against the NR (Non-redundant Protein Database) using PSIBLAST[23] with an *E*-score cutoff of 10^{-3} and three search rounds. The PSIBLAST sequence alignment (profile) and the position specific scoring matrix (PSSM) were retained for further use. Profiles were filtered as follows.

(1) Closely related proteins were removed: if a pair of proteins had more than 90% sequence identity in PSIBLAST, one was eliminated from the profile.
(2) Less reliably aligned proteins were removed: any protein with less than 30% sequence identity to the query human sequence was removed.
(3) Regions of the alignment where more than 50% of the sequences have a gap were removed.

Features for the support vector machine

The following five features were used for the SVM.

(1) The probability of substituting the variant residue type a at position j in the sequence alignment, $P(a,j)$, taken from the corresponding matrix element in the PSSM.
(2) The entropy at each position j in the alignment, calculated using the Shannon[37] entropy formula:

$$S_j = -\Sigma P_i \log_2 P_i$$

where the sum is over the 20 possible amino acids, and P_i is the probability of a particular residue type i at this position. Probabilities are calculated from the filtered alignment profile.

(3) The mean entropy $\langle S \rangle$ over the sequence, calculated by averaging over all sequence positions.
(4) The standard deviation of the entropy over all positions, calculated as:

$$\sigma(S) = [(\sum_i (S_i - \langle S \rangle)^2)/(N-1)]^{1/2}$$

where the sum is over all sequence positions, S_i is the entropy at a particular position, and N is number of sequence positions.
(5) The entropy at each position j, expressed as a Z score:

$$Z_j = (S_j - \langle S \rangle)/\sigma(S)$$

Support vector machine (SVM)

The five parameters described above: probability of accepting that amino acid substitution, entropy, mean entropy, standard deviation of the entropy and the entropy Z score, were used as features to train a SVM. The deleterious variant set consisted of these values for all the monogenic disease causing residue positions, and the control set were the values for the inter-species amino acid differences. SVM^{light}§, an implementation of SVM in C, was used, with a linear kernel. Weights were assigned to the disease and control data sets to compensate for their different sizes, such that they contributed equally to determining the partitioning surface. The distance of a data point from the partitioning surface provides an approximate measure of confidence in a classification. Bootstrapped datasets were used for training and accuracy assessment. That is, each SVM was trained on data points drawn randomly from the disease and control sets, with the total number of points equal to the size of each set. The training and testing procedure was repeated 30 times. For each trial, the false negative rate (the fraction of deleterious variations mis-classified as non-deleterious) and false positive rate (the fraction of non-deleterious variations mis-classified as deleterious) in the test dataset (those points not included in training) were calculated. The average false positive and false negative rates provide the measure of the classification accuracy. The 95% confidence intervals were also obtained from the distribution of false positive and false negative values. A similar procedure was used to obtain confidence intervals for the fraction of deleterious SNPs in the population data sets.

Stability model

Full details are available in an earlier paper.[15] A summary is provided here. Eleven contributions to the energy and entropy of protein stability are considered. There are four classes of electrostatic interaction: reduction of charge-charge, charge-polar or polar-polar energy, or introduction of electrostatic repulsion; three solvation effects: burying of charge or polar groups, and reduction in non-polar area buried on folding; and two terms representing steric strain: backbone strain and overpacking. The other two contributions considered are cavity formation (affecting van der Waals energy), and loss of a disulfide bridge. Surface accessibility of the mutated residue relative to the unfolded state is also included, as well as three parameters related to the C^{α} temperature factor of the mutated residues, so that in all there are 15 parameters. The same SVM software as for the profile model was used to determine the partition-

‡ http://genome.perlegen.com and http://www.hapmap.org/

§ http://svmlight.joachims.org

ing surface between the disease and non-disease data in the 15-dimensional parameter space. Continuous variables were normalized in the form of a Z score (Z= (value-mean)/standard-deviation). A radial basis kernel with a γ value of 0.2 was used.

Estimate of protein divergence rate from human and mouse orthologous genes

Mouse orthologs were taken from the NCBI HomoloGene database.[38] For each orthologous pair, the BLAST sequence identity was calculated between all refseq mouse protein sequences and those of all the corresponding human refseq entries, and the highest value was used. (This procedure is necessary, since each gene may have multiple protein isoforms.)

Matching of mouse knockouts with human genes

The OMIM ID of each available mouse knockout gene was extracted and matched to the NCBI locuslink database, to identify the corresponding human gene name. Human curation was used to match remaining mouse genes and verify each link. The matched human genes were compared to those in the HGMD database, to find the subset involved in monogenic disease.

Acknowledgements

This work was supported by grant LM07174 from the National Library of Medicine. We thank Eugene Melamud for help with the database infrastructure, and many useful discussions.

References

1. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al*. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.
2. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G. *et al*. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
4. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.* **29**, 308–311.
5. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G. *et al*. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
6. The International Hapmap Consortium. (2003). The International HapMap Project. *Nature*, **426**, 789–796.
7. Kruglyak, L. & Nickerson, D. A. (2001). Variation is the spice of life. *Nature Genet.* **27**, 234–236.
8. Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A. *et al*. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247.
9. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N. *et al*. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238.
10. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. *et al*. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581.
11. Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.
12. Botstein, D. & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genet.* **33**, 228–237.
13. Emahazion, T., Feuk, L., Jobs, M., Sawyer, S. L., Fredman, D., St Clair, D. *et al*. (2001). SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends Genet.* **17**, 407–413.
14. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597.
15. Yue, P., Li, Z. & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473.
16. Ng, P. C. & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucl. Acids Res.* **31**, 3812–3814.
17. Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucl. Acids Res.* **30**, 3894–3900.
18. Chasman, D. & 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.
19. Krishnan, V. G. & Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, **19**, 2199–2209.
20. Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N. *et al*. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucl. Acids Res.* **31**, 334–341.
21. Verzilli, C. J., Whittaker, J. C. & Stallard, N. D. C. (2005). A hierarchical Bayesian model for predicting the functional consequences of amino acid polymorphisms. *J. Roy. Statist. Soc. C*, **54**, 191–207.
22. Molchanova, T. P., Pobedimskaya, D. D. & Postnikov, Yu. V. (1994). A simplified procedure for sequencing amplified DNA containing the alpha 2- or alpha 1-globin gene. *Hemoglobin*, **18**, 251–255.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
24. Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Funct. Genet.* **17**, 49–61.
25. Ferrer-Costa, C., Orozco, M. & de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315**, 771–786.

26. Mori, Y., Takeda, K., Charbonneau, M. & Refetoff, S. (1990). Replacement of Leu227 by Pro in thyroxine-binding globulin (TBG) is associated with complete TBG deficiency in three of eight families with this inherited defect. *J. Clin. Endocrinol. Metab.* **70**, 804–809.
27. Waltz, M. R., Pullman, T. N., Takeda, K., Sobieszczyk, P. & Refetoff, S. (1990). Molecular basis for the properties of the thyroxine-binding globulin-slow variant in American blacks. *J. Endocrinol. Invest.* **13**, 343–349.
28. Huang, H., Winter, E. E., Wang, H., Weinstock, K. G., Xing, H., Goodstadt, L. *et al*. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. *Genome Biol.* **5**, R47.
29. Smith, N. G. & Eyre-Walker, A. (2003). Human disease genes: patterns and predictions. *Gene*, **318**, 169–175.
30. Frenette, P. S., Mayadas, T. N., Rayburn, H., Hynes, R. O. & Wagner, D. D. (1996). Susceptibility to infection and altered hematopoiesis in mice deficient in both P- and E-selectins. *Cell*, **84**, 563–574.
31. Wang, Z. & Moult, J. (2003). Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins: Struct. Funct. Genet.* **53**, 748–757.
32. Smith, D. J. & Lusis, A. J. (2002). The allelic structure of common disease. *Hum. Mol. Genet.* **11**, 2455–2461.
33. Pritchard, J. K. & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant···or not? *Hum. Mol. Genet.* **11**, 2417–2423.
34. Prince, J. A., Feuk, L., Sawyer, S. L., Gottfries, J., Ricksten, A., Nagga, K. *et al*. (2001). Lack of replication of association findings in complex disease: an analysis of 15 polymorphisms in prior candidate genes for sporadic Alzheimer's disease. *Eur. J. Hum. Genet.* **9**, 437–444.
35. Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L. *et al*. (2004). Database resources of the National Center for Biotechnology Information: update. *Nucl. Acids Res.* **32**, D35–D40.
36. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al*. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
37. Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.
38. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M. *et al*. (2005). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **33**, D39–D45.

Edited by J. Karn

(Received 6 July 2005; received in revised form 4 December 2005; accepted 8 December 2005)
Available online 27 December 2005