

## ORIGINAL ARTICLE

# Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes

G Hu<sup>1,3,5</sup>  
B Modrek<sup>1,3,5</sup>  
HMF Riise Stensland<sup>2</sup>  
J Saarela<sup>2</sup>  
P Pajukanta<sup>2</sup>  
V Kustanovich<sup>2</sup>  
L Peltonen<sup>2</sup>  
SF Nelson<sup>2</sup>  
C Lee<sup>1,3,4,5</sup>

<sup>1</sup>Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA;

<sup>2</sup>Department of Human Genetics, University of California, Los Angeles, CA, USA; <sup>3</sup>UCLA-DOE Laboratory for Structural Biology and Molecular Medicine, University of California, Los Angeles, CA, USA; <sup>4</sup>Molecular Biology Institute, University of California, CA, USA

**Correspondence:**

C Lee, UCLA-DOE Laboratory for Structural Biology and Molecular Medicine, University of California, Los Angeles, CA 90095-1570, USA.  
Tel: +1 310 825 7374  
Fax: +1 310 267 0248  
E-mail: leec@mbi.ucla.edu

Received: 17 November 2001

Revised: 30 January 2002

Accepted: 7 March 2002

**ABSTRACT**

Single nucleotide polymorphisms in protein coding regions (cSNPs) are of great interest for their effects on phenotype and potential for mapping disease genes. We have identified 5400 novel exonic SNPs from alignments of public EST data to the draft human genome sequence, and approximately 12 000 more novel exonic SNPs from EST cluster alignments. We found 82% of the genomic-aligned SNPs and 63% of the EST-only SNPs to be detectably polymorphic in 20 Finnish DNA samples. 37% of the SNPs mapped to known protein coding regions, yielding 6500 distinct, novel cSNPs from the two datasets. These data reveal selection against mutations that alter protein structure, and distinct classes of genes under strongly positive vs. negative pressure from natural selection for amino acid replacement (detected by  $K_A/K_S$  ratio). We have searched these cSNPs for compatibility with the amino acid profile at each site and structural impact on protein core stability. *The Pharmacogenomics Journal* (2002) 2, 236–242. doi:10.1038/sj.tpj.6500109

**Keywords:** data mining; bioinformatics; functional genomics; cSNP; polymorphisms; EST

**INTRODUCTION**

There is great interest in the discovery and study of single nucleotide polymorphisms (SNPs) for disease mapping and other applications. SNPs found in coding regions (cSNPs) are especially important. They can be used for mapping disease gene mutations in regions of sufficiently large linkage disequilibrium. Moreover, the non-synonymous cSNPs themselves are likely responsible for a fraction of common human phenotype variation, including disease susceptibilities.<sup>1–3</sup>

Most high-throughput SNP discovery has been from bulk genomic sequence. However, because coding regions constitute only a small fraction of the human genome, the fraction of these data that are cSNPs is correspondingly small (1–2%). SNP identification from expressed sequences (ESTs)<sup>4–10</sup> can complement these data in a valuable way, because EST sequences primarily represent the coding regions of human genes.

EST-based SNP discovery faces difficult statistical challenges in distinguishing genuine SNPs from artifacts such as sequencing error and EST misclustering. First of all, it is essential to know which ESTs are truly from the same gene.<sup>9</sup> The UniGene clustering represents a long-term effort to group all the deposited ESTs into putative gene clusters.<sup>11</sup> While UniGene has been an immensely useful resource, its similarity-based clustering method is not sensitive enough to distinguish highly similar ESTs from paralogous genes. Sequences have been frequently found to be misclustered in UniGene.

In order to enrich the number of synonymous and non-synonymous cSNPs for association studies, we have mapped more than half of the available UniGene clusters onto finished and draft genomic sequences. The results not only

provided the precise locations and detailed gene structures for those EST clusters, but also allowed us to correct errors in the UniGene clustering by using the genomic sequence as strict reference for sorting out highly similar paralogs. This reduces the rate of false positive SNPs.

## RESULTS AND DISCUSSION

### SNP Identification and Verification

To identify novel SNPs in coding regions, human EST sequences were aligned to the draft human genome sequence<sup>12</sup> employing strict matching criteria. We collected sequencing chromatograms for the aligned EST sequences and calculated quality and sequencing error probabilities for each single base mismatch. Candidate SNPs were evaluated by a lod score calculation that measures the log-odds ratio for the probability of a genuine SNP vs. sequencing error and other types of error.<sup>10</sup> We identified 12 410 high confidence SNPs from ESTs mapped to the draft genome sequence, with an average lod score of 14. We were able to compare a subset of these (8208 SNPs) with the public SNP database dbSNP.<sup>13</sup> We found that 4643 (57%) matched existing dbSNP entries (dbSNP February 2001 release, including SNPs submitted from our previous EST analysis),<sup>10</sup> while 3565 (43%) were novel. Given the prodigious growth (more than double) in the EST database since our original dbSNP

submission was generated (September 1999), this rate seems reasonable. Considering the whole set of 12 410 SNPs, this fraction implies about 5400 novel exonic SNPs.

Because more than half of the UniGene ESTs were not mapped by our procedure to an exact, unique location in the draft genome sequence<sup>14</sup> (and were therefore excluded by the above procedure), we also identified additional SNP candidates by aligning ESTs without genomic sequence, using the remaining UniGene EST clusters, as previously described.<sup>10</sup> This procedure divides an EST cluster into separate groups of ESTs when there is evidence of paralogous sequences in the alignment. We identified 40 356 candidate SNPs with high lod scores (Table 1). Because these ESTs were not unambiguously mapped to the draft genome sequence, matching these SNPs with existing data in dbSNP is uncertain. However, if the rate of novel SNP discovery is similar to the first dataset, this should include approximately 17 000 novel exonic SNPs.

To test the utility of this novel SNP dataset for population isolates commonly used in human disease gene mapping, 105 SNP candidates were randomly selected for experimental testing by sequencing DNA samples from 20 Finnish people (Figure 1). The polymorphism detection rates were 63% for the EST-only dataset, and 82% for the genome-mapped dataset (Table 1). This difference appears to be stat-

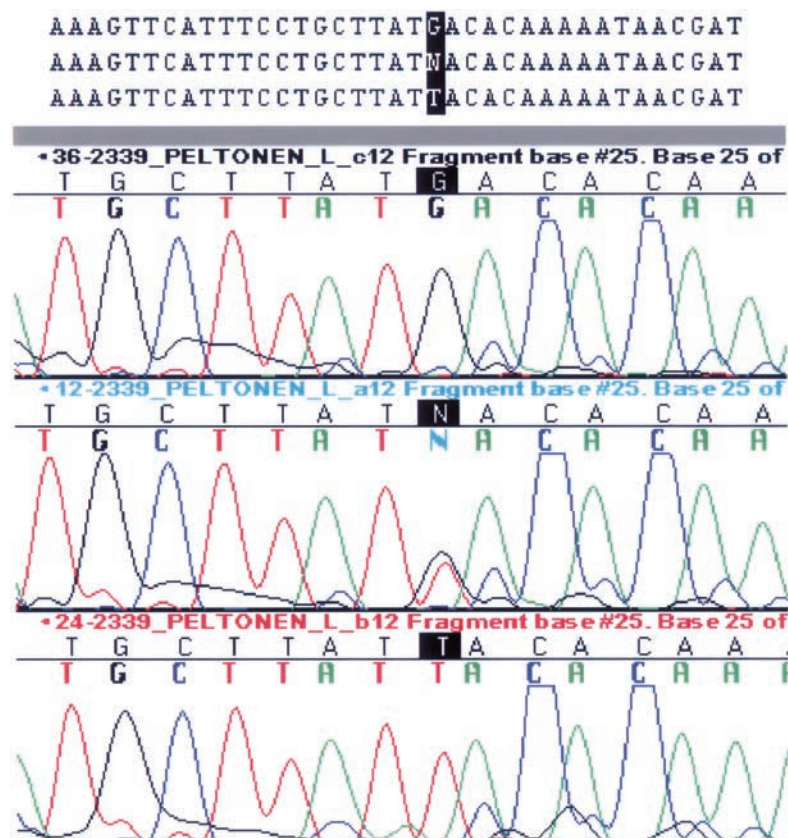


Figure 1 PCR/sequencing confirmation of a candidate SNP (allele\_id 435959 in UniGene cluster Hs.82273) from independently screened Finnish DNA samples. All three forms were observed: homozygous G/G (top panel), heterozygous G/T (middle), homozygous T/T (bottom).

istically significant ( $P < 0.05$ ). Not surprisingly, SNPs with high estimated population allele frequency in the mostly American EST population sample, were more likely to be found in the Finnish population sample. SNPs with estimated American population allele frequency of 15% or higher constituted 90% of the SNPs verified in the Finnish samples, and none of the SNPs with estimated American population allele frequency less than 5% were verified in the Finnish samples. It is possible that some of these SNPs are from ethnic groups not common in Finland. It should be noted that validating SNPs in such an independent population isolate is a more strenuous test than simply re-sequencing the same person's DNA in which the SNP candidate was originally identified, which can give validation rates of up to 95%.<sup>15</sup> It has been estimated that approximately 20% of human SNPs in the public data are rare or 'private' SNPs, and that as few as 50% may be found commonly in a given ethnic group.<sup>16</sup>

To assess the accuracy of SNP detection in a larger population sample, we examined the most polymorphic gene in our data set, HLA-C. Since HLA-C is part of a set of paralogous Major Histocompatibility Complex (MHC) genes, this is an important test of our reliability in distinguishing true SNPs within one gene vs. spurious sequence differences due to mixing of paralogous gene sequences. To validate our HLA-C SNPs, we compared them to verified HLA-C allele sequences from the IMGT/HLA Sequence Database (<http://www.ebi.ac.uk/imgt/hla>). Of the 121 cSNPs identified by our procedure in HLA-C, 107 match known HLA-C polymorphisms, yielding a verification rate of 89%, consistent with the results we reported previously.<sup>10</sup>

All of the candidate polymorphisms reported in this study are being submitted to the public dbSNP database (accession numbers 4390536–4402800 and 4403180–4442422). We have also constructed a website (<http://www.bioinformatics.ucla.edu/snp>) that allows researchers to search and analyze these integrated data for details, including the validation results, effect on protein coding, and analysis of cSNPs most likely to have functional impact (described below).

### High Yield of Coding Region SNPs

To relate these novel polymorphism data to potential impact on protein function, we searched for an exact match to a known protein sequence for each mapped gene, and mapped the amino acid effects of the SNPs. For the 6129

mapped genes in which we identified SNPs, we found matches for 1410 of these to known, curated protein sequences. Of the 3480 SNPs we identified in these genes, 37% (1281) mapped in the protein coding regions, while the rest appeared to be in 5' UTR or 3' UTR (Table 1). We have also searched for exact matches to known protein sequences for EST consensus sequences generated from our EST-only alignments. Using this approach we were able to map an additional 5673 unique cSNPs in coding regions of known proteins. Overall we have identified and characterized approximately 7000 cSNPs (Table 1).

The high yield of cSNPs from ESTs (37% of the SNPs identified are cSNPs) compares favorably with the low yield obtained by SNP discovery based on genomic sequencing. Because of the very low fraction of the human genome sequence that actually codes for protein, there is naturally a low yield of cSNPs from genomic sequence. For comparison, of the 1.7 million SNP candidates deposited by the SNP Consortium to date (October 2001), they have noted that 7724 are in exonic regions (based on the Ensembl 26k gene prediction set) and 2655 are cSNPs.

How many novel cSNPs does our dataset contribute? Considering only the 7000 cSNPs in this study that map to known, curated protein sequences, we estimate that 3000 (43%) are novel. Alternatively, applying the 37% cSNP yield to our 40356 EST-only based SNPs, indicates a total of 14930 cSNPs in all human proteins, of which about 6500 are likely to be novel. Thus, this study probably contributes 3000 to 6500 novel cSNPs. EST-based SNP discovery may also make a useful contribution to identifying SNPs in gene regions that are transcribed but not part of the coding sequence. These untranslated regions (UTR) are also of potential interest for their functional impact and strong linkage to disease mutations in genes. The analysis of 1.42 million unique SNPs by the International SNP Map Working Group annotated 12435 SNPs (from both genomic and EST sources) within UTRs of known genes. In our genome-mapped SNP set, we identified 2199 SNPs in known UTRs (Table 1).

### Analysis of Nucleotide Diversity

We have calculated the levels of nucleotide diversity, defined as heterozygosity per nucleotide site ( $\pi$ ), for different categories of cSNPs. While the nucleotide diversity of synonymous ( $1.5 \times 10^{-4}$ ) and non-synonymous cSNPs (1.8

**Table 1** SNPs identified by this study from human ESTs

	Num of identified SNPs						Total	Validation rate ( $n = 20$ , Finnish)	SNPs mapped in proteins		
	By lod score			By frequency					cSNPs	UTRs	Total
	3–6	6–20	>20	<5%	5–15%	>15%					
EST + genomic	6057	4413	1940	11	1056	11343	12410	82%	1281	2199	3480
EST alone	17122	15833	7401	106	4506	35744	40356	63%	6649 <sup>a</sup>	8695	15344

<sup>a</sup>Only 5673 are unique, which are not overlapped with the 1281 cSNPs set identified from EST + genomic sequences.

$\times 10^{-4}$ ) is similar, it is depressed by a factor of greater than 3-fold for non-conservative cSNPs ( $7.8 \times 10^{-5}$ ) vs conservative cSNPs ( $2.5 \times 10^{-4}$ ). This indicates that amino acid replacements causing greater structural changes are strongly selected against in the human population, as has also been observed in previous studies of specific sets of genes.<sup>6,7,17</sup> It should be emphasized that calculations of  $\pi$  from EST data underestimate the absolute level of heterozygosity in the genome, due to the relatively small population sample represented in the EST data for most genes.

These measures of nucleotide diversity also show strong variation across genes, reflecting natural selection of their functional roles as well as meiotic recombination rates. We have identified those genes with the highest levels of polymorphism in our dataset. The Major Histocompatibility Complex genes top this list, and provide an example of the type of functional information that can be garnered from this dataset. Both class I and class II MHC (eg HLA-A, HLA-C, HLA-B, HLA-DRB5, HLA-DPB1) were detected to have the highest levels of nucleotide diversity. Moreover, the pattern of selection pressure for these genes strongly favors nonsynonymous cSNPs (2-fold higher heterozygosity than synonymous), but *disfavors* nonconservative mutations (nearly 4-fold lower than conservative mutations) even more strongly than for the average of all genes. This evident selection pressure within each HLA gene for amino acid replacements, but not large structural changes, contrasts strikingly with the pattern of amino acid differences between separate HLA loci, which favors nonconservative changes. Our data agree well with previous analyses of the unusual polymorphism patterns of these genes. It has been observed that many polymorphic features of HLA molecules are segregated between loci. When variability plots are limited to products of a single locus, values of variability are significantly reduced.<sup>18</sup>

### Analysis of cSNP Impact on Protein Structure and Function

Protein structural analysis can facilitate assessment of a cSNP's functional impact.<sup>10,19,20</sup> Recently several structure-based methods for predicting functional effects were proposed and applied for a number of cSNPs whose host protein tertiary structures are available.<sup>21–23</sup> The results suggested that about 20–32% of non-synonymous cSNPs could alter protein structure, function, stability or folding significantly. Because structures are not yet available for the majority of proteins, this type of structural modeling is not feasible for our whole data set. To provide users some information on our cSNPs' probable functional impact, we searched the ASTRAL<sup>24</sup> protein domain database for sequence similarity to the genes in which we identified cSNPs. Using this approach, we were able to map a subset of our cSNPs to a location in a known protein fold, and assessed the similarity of the cSNP mutation to this pattern of amino acid variability at this position in the set of all similar protein sequences. These data could offer researchers a significant enrichment for polymorphisms likely to cause important

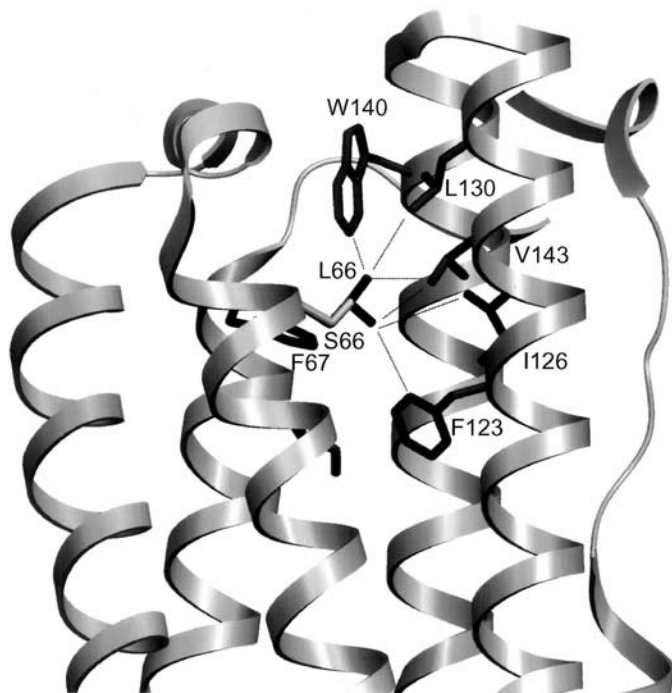
functional effects, and will be made available on our website upon publication (<http://www.bioinformatics.ucla.edu/snp>).

We used the ASTRAL database of protein domain sequences,<sup>24</sup> based on the SCOP structural domain database, to map our cSNPs onto protein domains whose three-dimensional structures have been solved. Starting with 688 protein sequences in which we have identified non-synonymous cSNPs (688 from the 1281 cSNP subset), we searched against ASTRAL both to identify their domain structures and amino acid variability of aligned ASTRAL protein family members at each cSNP location. We mapped 272 cSNPs (40%) onto ASTRAL domain sequences. cSNPs in highly variable positions are less likely to have strong functional effects, while non-conservative amino acid changes in positions with very low variability may be highly enriched for phenotypic impact. To assess this, we calculated for each non-synonymous cSNP location the amino acid variability within the homologous domain sequences from ASTRAL.

Structural modeling can provide additional predictors of functional impact. As an illustrative example from our database, cSNP T/C (allele\_id=1009925) in UniGene cluster Hs.1510 causes an amino acid substitution of Leu 66 → Ser in interferon  $\alpha 4$ . The tertiary structure of interferon  $\alpha 4$  showed that Leu 66 is involved in the hydrophobic core packing of the four helix bundle. We identified twelve closely related domain sequences in ASTRAL, with two hydrophobic, conservative variations (Val, Ile) at this location, consistent with close packing with other hydrophobic residues in the core. Based on structural modeling,<sup>25</sup> replacement of leucine by serine at this site is likely to destabilize the helix bundle because of its smaller size and polar character (Figure 2).

### Patterns of Polymorphism: Identifying Distinct Classes of Highly Polymorphic Genes

The level and pattern of coding region polymorphism detected in human genes can provide interesting insights into their function. For example, unusually high levels of polymorphism reveal unusual pressure from natural selection. Our initial analysis of the 50 most polymorphic genes highlights several classes of genes (Table 2). Two very different patterns of polymorphism are evident in these data, based on the  $K_A/K_S$  ratio measuring the normalized ratio of amino acid changes vs. synonymous substitutions observed in a gene: *negative selection*, indicated by  $K_A/K_S$  ratio much less than 1; versus *positive selection* indicated by a  $K_A/K_S$  ratio of at least 1 (or 0.6 with relaxed functional constraints).<sup>30</sup> Among genes with large  $K_A/K_S$  ratios (greater than 0.6), the most conspicuous group consists of proteins that directly interact with pathogen molecules. This includes the MHC family, genes containing immunoglobulin-like domains, pregnancy-specific glycoprotein (*PSG4*), placental alkaline phosphatase (*ALPP*), and natural killer cell associated transcript 5. MHC class I and II loci comprised seven of the top eight most polymorphic genes, consistent with our previous finding.<sup>10</sup> These proteins present peptide antigens to T cells, initiating an immune response and clearance of foreign pathogens. The hypothesis of overdominant selection



**Figure 2** Superposition of wild type interferon  $\alpha 4$  crystal structure and Leu66  $\rightarrow$  Ser mutant model. In the wild type, Leu66 (black) is critical in maintaining a stable helix bundle via hydrophobic and van der Waals interactions with surrounding hydrophobic residues. The Leu66  $\rightarrow$  Ser (in grey) substitution caused by a non-conservative cSNP (allele 1009925 in UniGene cluster Hs.1510) is predicted to destabilize the host protein interferon  $\alpha 4$  due to the smaller size and polar property of Ser, although the backbone structure of the mutant is essentially unchanged from that of the wild type.

(heterozygote advantage) of the MHC proposes that individuals heterozygous at MHC loci are able to present a greater variety of antigenic peptides than individuals homozygous at these loci, resulting in broader immune response to a diverse array of pathogens.<sup>31</sup> Therefore, selective forces such as infectious disease morbidity act not to conserve the sequence, but to maintain a high level of diversity in the human population. These patterns are easily detected in our data, across a large number of genes known to be involved in pathogen interactions. Our data suggest that genome-wide cSNP analysis can potentially indicate new genes that have important interactions with pathogens.

Among genes with low  $K_A/K_S$  ratios ( $K_A/K_S < 0.6$ ) but large numbers of detected polymorphisms, a very different functional pattern is observed. These are typically genes with ubiquitous, high expression. Because these genes are represented by far more ESTs per gene than for an average gene, we can detect many polymorphisms. However, their pattern reveals strongly negative selection. For example, although we identified 17 cSNPs in  $\alpha$  tubulin (its ubiquitous expression has caused it to be re-sequenced repeatedly in hundreds of different ESTs and libraries), 82% (14) are synonymous, and the remaining three are conservative. The calculated  $K_A/K_S$  value is only 0.09. This gene is well known to

be ubiquitous in eukaryotes and highly conserved in evolution. Ubiquitin C (Hs.183704) is also widely expressed, and shows a similar cSNP profile with a  $K_A/K_S$  value of 0. We detect 19 cSNPs in ubiquitin C; all 19 are synonymous. Additional examples include keratin 6B and tubulin  $\beta$ .

**Conclusion:** This study's integration of expressed sequence data, genomic mapping and gene structure identification and protein coding consequences provides a valuable resource for biologists seeking additional functionally interesting polymorphisms. Although the numbers of novel SNPs presented in this study are small (3000–6500 novel cSNPs) relative to total human SNP discovery to date, they are a significant subset of SNPs in protein coding regions. Of the 1.42 million human SNPs in dbSNP (February 2001<sup>15</sup>), only 10 016 (from both genomic and EST sources) were annotated as cSNPs in protein coding regions (this number should be higher now; as of October 2001, dbSNP contains 1.98 million distinct, mappable SNPs). Analysis of a set of 68 well-characterized genes indicated that our previous EST-based SNP submission<sup>10</sup> constituted 35% of all cSNPs deposited in dbSNP for these genes,<sup>32</sup> despite the enormously larger total numbers of SNPs generated from genomic sequencing methods. Finally, the high yield of cSNPs from EST data in our study suggests that the continuing rapid growth of EST sequence databases will make an important contribution to cSNP-based mapping efforts. Current estimates suggest a total of 60 000 common exonic SNPs in the human population,<sup>15</sup> of which 30 000 would likely be cSNPs. Thus, additional cSNP discovery is needed, and our data demonstrate that EST-based approaches can produce a particularly rich yield.

## MATERIALS AND METHODS

### SNP Candidate Identification

To find candidate genomic sequences for each UniGene cluster, we searched the draft human genome sequence for matches to each cluster's consensus sequence as described previously.<sup>14</sup> ESTs were aligned to the draft genomic sequence using the program POA.<sup>33</sup> We eliminated ESTs and mRNAs with less than 95% identity (excluding poly-A tails) to the genomic sequence, or insertions or deletions greater than 3 bp (EST gaps of greater than 10 bp were allowed, as potential introns). Using the genomic-aligned EST sequences, SNP identification, scoring, and allele frequency estimation was performed as previously described.<sup>10</sup>

### SNP Validation

Primers flanking the predicted SNPs were designed for PCR amplification of genomic DNA. The 70–200 bp PCR products were sequenced on an ABI 377 sequencer following standard procedures, and sequences were analyzed using the Sequencher™ software (Gene Codes Corporation, 1999).

### Heterozygosity Calculation

The nucleotide diversity calculation is limited to 1.6 million bp of coding region from which 1281 cSNPs were identified. Assuming that each SNP is biallelic, nucleotide diversity ( $\pi$ ) is calculated using:

**Table 2 Fifty most polymorphic genes**

UniGene cluster	Gene name	Number of cSNPs			$K_a/K_s$
		Total	Non-synonymous	Non-conservative	
Hs.277477	MHC, class I, C	121	85	35	0.99
Hs.181244	MHC, class I, A	108	75	27	0.92
Hs.180255	MHC, class II, DR $\beta$ 1	93	67	24	1.11
Hs.181125	Immunoglobulin $\lambda$	90	42	14	0.40
Hs.77961	MHC, class I, B	86	59	25	0.92
Hs.198253	MHC, class II, DQ $\alpha$ 1	85	52	21	0.66
Hs.181366	MHC, class II, DR $\beta$ 5	76	53	21	0.98
Hs.73931	MHC, class II, DQ $\beta$ 1	74	43	17	0.58
Hs.173609	Pregnancy specific $\beta$ -1-glycoprotein 1	61	41	12	0.86
Hs.140	Immunoglobulin heavy constant $\gamma$ 3	55	34	9	0.63
Hs.156110	Immunoglobulin $\kappa$ variable 1D-8	53	29	11	0.50
Hs.469	Flavoprotein	48	26	8	0.47
Hs.104991	P 845O24	46	31	16	0.91
Hs.105928	Leukocyte immunoglobulin-like receptor	41	32	13	1.60
Hs.108380	Rhesus blood group, D antigen	41	36	19	3.24
Hs.111758	keratin 6B	33	8	2	0.14
Hs.198287	Pregnancy specific $\beta$ -1-glycoprotein 11	33	22	9	0.85
Hs.218329	Hypothetical protein	31	19	10	0.62
Hs.89925	Calcium channel $\alpha$ 1 C	31	15	5	0.37
Hs.180266	Tropomyosin 2	30	14	4	0.29
Hs.814	MHC, class II, DP $\beta$ 1	30	21	7	0.94
Hs.11611	KIAA1424	29	17	5	0.61
Hs.181246	Glucosidase $\beta$	29	20	10	0.91
Hs.258612	Killer cell immunoglobulin-like receptor	28	19	8	0.87
Hs.5233	Glutathione S-transferase M4	28	18	3	0.64
Hs.101047	Transcription factor 3	27	11	3	0.33
Hs.278581	Fibroblast growth factor receptor 2	25	16	6	0.67
Hs.84389	Synaptosomal-associated protein	25	7	2	0.13
Hs.109045	FLJ10498	24	12	7	0.35
Hs.274601	Killer cell immunoglobulin-like receptor	24	19	12	1.62
Hs.89552	Glutathione S-transferase A2	24	14	5	0.53
Hs.96253	Calcium channel $\alpha$ 1 A	23	12	5	0.42
Hs.118174	Tetratricopeptide repeat domain 3	22	10	5	0.31
Hs.167835	Acyl-Coenzyme A oxidase	21	7	2	0.19
Hs.183418	Cell division cycle 2-like 1	21	9	5	0.29
Hs.234573	TL132	21	12	6	0.53
Hs.171391	C-terminal binding protein 2	20	13	5	0.84
Hs.182280	MADS box transcription enhancer factor 2	20	11	3	0.53
Hs.119076	Tubulin $\beta$	19	2	1	0.04
Hs.183704	Ubiquitin C	19	0	0	0.00
Hs.9006	VAMP-associated protein A	19	11	6	0.52
Hs.99526	Odorant-binding protein 2B	19	19	8	>1
Hs.1510	Interferon $\alpha$ 10	18	13	6	0.88
Hs.201967	Aldo-keto reductase family 1	18	8	2	0.30
Hs.250700	Tryptase $\beta$	18	10	4	0.45
Hs.11900	Actinin $\alpha$ 1	17	13	3	1.12
Hs.182982	golgin-67	17	14	4	1.90
Hs.239189	KIAA0838	17	0	0	0
Hs.278242	Ubiquitous tubulin $\alpha$	17	3	0	0.09
Hs.278552	Immunoglobulin $\lambda$ joining 3	17	11	1	0.80

$$\pi = \frac{1}{L} \sum_{i=1}^L [1 - (f_{im}^2 + f_{iM}^2)] = \frac{N\bar{H}}{L}$$

where  $N$  is the total number of cSNPs (1281);  $f_{im}$ ,  $f_{iM}$  are the minor and major allele frequency of SNP  $i$ , respectively;  $\bar{H}$  is mean heterozygosity;  $L$  is the total combined length of genomic sequence from which cSNPs were identified.

### SNP Mapping to Protein Sequence and Structure

For UniGene clusters with a full length, human-curated mRNA sequence, we extracted the protein sequence from GenBank, and aligned it to the EST consensus sequence using POA's three-frame alignment mode. We also searched it for matches to the ASTRAL structural domain sequence database,<sup>24</sup> using BLAST expectation value ( $10^{-4}$ ) and identity (30%) cutoffs to choose good candidates for further analysis, including 'variability' calculation, mutant modeling using SCEO,<sup>25</sup> and automated annotation (domain, homology, polymorphism, disease association, etc.) using GeneMine.<sup>34</sup> We also calculated nucleotide diversity data from a total of 1.6 million bp of mapped coding regions for known proteins, in which we identified 1281 cSNPs (one cSNP per 1.2 kb).

### $K_A/K_S$ Calculation

$K_A/K_S$  was calculated for each gene by dividing the number of non-synonymous cSNPs by the number of synonymous cSNPs, and then normalizing it with the ratio of non-synonymous over synonymous sites ( $N_{ratio}$ ) in the whole coding region of the gene sequence.<sup>27</sup>  $N_{ratio}$  was calculated by using the corresponding protein sequence, and a constructed profile containing the numbers of non-synonymous and synonymous substitutions in the genetic codon for each amino acid.

### ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant No. 0082964. CL was supported by Department of Energy grant DEFG0387ER60615 and a grant from the Searle Scholars Program. VK was supported by USPHS National Research Service Award GM07104. SN was supported by NIH/NIMH R01 grant MH58277. HMFERS was supported by a post-doctoral fellowship from the Research Council of Norway (#129517/310). BM is a predoctoral trainee supported by NSF IGERT Award #DGE-9987641. LP was supported by NIH/NHLBI grant #PO1HL28481, National Multiple Sclerosis Society grant #RG3050-A-IT and a grant from The Academy of Finland.

### DUALITY OF INTEREST

None declared.

### REFERENCES

- Malkin D, Li FP, Strong LC, Fraumeni JF, Nelson CE, Kim DH et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 1990; **250**: 1233–1238.
- Collins FS. Cystic fibrosis: molecular biology and therapeutic implications. *Science* 1992; **256**: 774–779.
- Dunning AM, Chiano M, Smith NR, Dearden J, Gore M, Oakes S et al. Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. *Hum Mol Genet* 1997; **6**: 285–289.
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998; **280**: 1077–1082.
- Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA et al. Mining SNPs from EST databases. *Genome Res* 1999; **9**: 167–174.
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet* 1999; **22**: 231–238.
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet* 1999; **22**: 239–246.
- Buetow KH, Edmonson MN, Cassidy AB. Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet* 1999; **21**: 323–325.
- Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H et al. A general approach to single nucleotide polymorphism discovery. *Nature Genet* 1999; **23**: 452–456.
- Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W et al. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet* 2000; **26**: 233–236.
- Schuler G. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 1997; **75**: 694–698.
- Consortium. IHGS. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
- Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 1999; **9**: 677–679.
- Modrek B, Resch A, Grasso C, Lee C. Genome-wide analysis of alternative splicing using human expressed sequence data. *Nucleic Acids Res* 2001; **29**: 2850–2859.
- Group ISMW. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409**: 928–933.
- Marth GT, Yeh R, Minton M, Donaldson R, Li Q, Duan S et al. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet* 2001; **27**: 371–372.
- Sunyaev S, Lathe WC, Ramensky V, Bork P. SNP frequencies in human genes; an excess of rare alleles and differing modes of selection. *Trends Genet* 2000; **16**: 335–337.
- Parham P, Lomen CE, Lawlor DA, Ways JP, Holmes N, Coppin HL et al. Nature of polymorphism in HLA-A, -B, and -C molecules. *Proc Natl Acad Sci USA* 1988; **85**: 4005–4009.
- Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* 1991; **352**: 448–451.
- Lee C. Testing homology modeling on mutant proteins: the A98V mutants of T4 lysozyme. *Folding & Design* 1996; **1**: 1–12.
- Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 2001; **307**: 683–706.
- Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* 2001; **10**: 591–597.
- Wang Z, Moulton J. SNPs, protein structure, and disease. *Hum Mut* 2001; **17**: 263–270.
- Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 2000; **28**: 254–256.
- Lee C. Predicting protein mutant energetics by self-consistent ensemble optimization. *J Mol Biol* 1994; **236**: 918–939.
- Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 1985; **2**: 150–174.
- Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 1993; **36**: 96–99.
- Wyckoff GJ, Wang W, Wu C-I. Rapid evolution of male reproductive genes in the descent of man. *Nature* 2000; **403**: 304–309.
- Fay JC, Wyckoff GJ, Wu C-I. Positive and negative selection on the human genome. *Genetics* 2001; **158**: 1227–1234.
- Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. The Adaptive Evolution Database (TAED). *Genome Biol* 2001; **2**: 1–6.
- Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet* 1998; **32**: 415–435.
- Irizarry K, Hu G, Wong ML, Licinio J, Lee C. Single nucleotide polymorphism identification in candidate gene systems of obesity. *Pharmacogenomics* 2001; **1**: 193–203.
- Lee C, Grasso C, Sharlow M. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**: 452–464.
- Lee C, Irizarry K. The GeneMine system for genome/proteome annotation and collaborative data-mining. *IBM Systems J* 2001; **40**: 592–603.