The Predicted Impact of Coding Single Nucleotide Polymorphisms Database

Matthew F. Rudd,[1] Richard D. Williams,[2] Emily L. Webb,[1] Steffen Schmidt,[3] Gabrielle S. Sellick,[1] and Richard S. Houlston[1]

Sections of [1]Cancer Genetics and [2]Paediatrics, Institute of Cancer Research, Sutton, Surrey, United Kingdom; and [3]Genetics Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Abstract

Nonsynonymous single nucleotide polymorphisms (nsSNP) have the potential to affect the structure or function of expressed proteins and are, therefore, likely to represent modifiers of inherited susceptibility. We have classified and catalogued the predicted functionality of nsSNPs in genes relevant to the biology of cancer to facilitate sequence-based association studies. Candidate genes were identified using targeted search terms and pathways to interrogate the Gene Ontology Consortium database, Kyoto Encyclopedia of Genes and Genomes database, Iobion's Interaction Explorer PathwayAssist Program, National Center for Biotechnology Information Entrez Gene database, and CancerGene database. A total of 9,537 validated nsSNPs located within annotated genes were retrieved from National Center for Biotechnology Information dbSNP Build 123. Filtering this list and linking it to 7,080 candidate genes yielded 3,666 validated nsSNPs with minor allele frequencies $\geq$0.01 in Caucasian populations. The functional effect of nsSNPs in genes with a single mRNA transcript was predicted using three computational tools—Grantham matrix, Polymorphism Phenotyping, and Sorting Intolerant from Tolerant algorithms. The resultant pool of 3,009 fully annotated nsSNPs is accessible from the Predicted Impact of Coding SNPs database at http://www.icr.ac.uk/cancgen/molgen/MolPopGen_PICS_database.htm. Predicted Impact of Coding SNPs is an ongoing project that will continue to curate and release data on the putative functionality of coding SNPs. (Cancer Epidemiol Biomarkers Prev 2005;14(11):2598–604)

Introduction

Much of the familial aggregation of common cancer results from inherited susceptibility, but highly penetrant mutations in known genes cannot account for most of the excess risk (1). Some of the unexplained familial risk is presumably due to high-penetrance mutations in as yet unidentified genes, but polygenic mechanisms may account for a greater proportion (1, 2). A popular hypothesis about the allelic architecture of susceptibility proposes that most of the genetic risk is caused by disease loci where there is one common variant or a restricted number of alleles (2). If true, the ''common disease-common variant'' hypothesis implies that testing for allelic association should be a powerful strategy for identifying low-penetrance alleles. This inference, coupled with technological developments, has led to a renaissance in association studies of common cancers.

The most common forms of variation in the human genome are single nucleotide polymorphisms (SNP). Currently, there are over 10 million human SNPs listed in publicly accessible databases, of which 92,000 are located within protein coding sequences (3). A fraction of these coding SNPs alter the encoded amino acid sequence [nonsynonymous SNPs (nsSNP)] and, therefore, have the potential to directly affect the structure, function, and interactions of expressed proteins. Nonsynonymous SNPs are proportionally less prevalent than synonymous SNPs, which do not alter protein sequence, presumably as a consequence of selection against the functional disruptions of amino acid variation. Although not all nsSNPs will have functional consequence, it is probable that a significant proportion of the molecular functional diversity in the human population is attributable to effects on protein function mediated through this form of genetic variation.

The types of mutations in Mendelian disease genes, coupled with issues of statistical power, provide a compelling rationale for the application of a sequence-based approach to association studies rather than complete reliance on a map of anonymous haplotypes (4). Because genome-wide scans are still financially challenging, adopting the strategy of limiting association studies to specific candidate genes or pathways has considerable use. In such analyses, it is advantageous to prioritize variants that may affect the structure or function of expressed proteins.

Missense changes can be analyzed according to the biochemical severity of the amino acid substitution and its context within the protein sequence. The Grantham matrix (5) predicts the effect of substitutions between amino acids based on chemical properties, including polarity and molecular volume. Using such criteria to classify amino acid changes, a clear relationship between the severity of replacement and the likelihood of clinical observation has been documented (4).

Recently, more sophisticated *in silico* algorithms have become available to predict the effect of amino acid substitutions on protein structure and activity. Polymorphism Phenotyping (PolyPhen; ref. 6) predicts the functional effect of substitutions by assessing the level of sequence conservation between homologous genes over evolutionary time, the physiochemical properties of the exchanged residues, and the proximity of the substitution to predicted functional domains and structural features within the protein. Sorting Intolerant from Tolerant (SIFT; ref. 7) predicts the functional importance of an amino acid substitution based on the alignment of highly similar orthologous and/or paralogous protein sequences. Predictions rely on whether or not an amino acid is conserved in the protein family, which can be indicative of its importance to the normal function or structure of the expressed protein.

Received 6/29/05; revised 8/18/05; accepted 8/31/05.

Grant support: Cancer Research UK and Leukaemia Research (G.S. Sellick).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at http://www.icr.ac.uk/cancgen/molgen/MolPopGen_PICS_database.htm.

Requests for reprints: Richard S. Houlston, Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom. Phone: 44-208-722-4175; Fax: 44-208-722-4365. E-mail: richard.houlston@icr.ac.uk

doi:10.1158/1055-9965.EPI-05-0469

Here, we have computed the predicted effects of nsSNPs in a series of genes relevant to the biology of cancer. A total of 3,666 nsSNPs validated in Caucasian populations were identified across 7,080 candidate genes. The putative functional effect of each nsSNP was determined by means of the Grantham matrix, PolyPhen and SIFT, and a pool of fully annotated nsSNPs generated.

Materials and Methods

Selection of Candidate Cancer Genes. To capture as many genes with potential relevance to the development of cancer as possible, we sourced gene data from online databases using both keyword and gene pathway queries. Genes were identified by interrogating the Gene Ontology Consortium database (8, 9), Kyoto Encyclopedia of Genes and Genomes *Homo sapiens* database (10, 11), Iobion's Interaction Explorer PathwayAssist Program (12), National Center for Biotechnology Information (NCBI) Entrez Gene database (13, 14), and the CancerGene database (15). The search categories and gene pathways were as follows: amino acid metabolism; binding; biodegradation of xenobiotics; catalytic activity; cellular processes, growth, and death; development; enzyme regulator activity; folding, sorting, and degradation; ligand-receptor interaction; metabolism of cofactors and vitamins; nucleotide metabolism; physiologic processes; regulation of biological processes; replication and repair; signal transduction and signal transducer activity; transcription and transcription regulator activity; translation and translation regulator activity; and transporter activity.

Bioinformatic Analyses

Filtering of LocusLink Genes. The complete NCBI LocusLink *Homo sapiens* file (replaced March 2005 with NCBI Entrez Gene; ref. 16) of autosomal genes was downloaded from the LocusLink FTP website (November 2004). Ambiguous proteins; those entries prefaced with FLJ, HSPC, KIAA, LOC, or PRO; hypothetical proteins; or those located tentatively within chromosomal open reading frames were excluded from SNP data mining, generating a list of 21,506 annotated genes.

Ethnicity-Based Selection of SNP Validation Panels. The minor allele frequency (MAF) of many SNPs differs significantly between ethnic groups (17). For pragmatic reasons, we chose to restrict our curation of nsSNPs to those validated in Caucasian populations. Details of populations for which allele frequency data were available within the NCBI SNP database (dbSNP Build 123, NCBI Human Genome Build 35; refs. 18, 19) were reviewed to document ethnicity. Only panels based primarily on genotypes derived from Caucasian individuals were retained for further interrogation.

Database Mining for nsSNPs. We retrieved a total of 48,492 nsSNPs from dbSNP Build 123, which included the addition of 1.6 million new submissions (391,000 newly validated SNP) since Build 122, corresponding with the release of NCBI Human Genome Build 35. Complete dbSNP XML files (20) for each autosome were parsed using an in-house generated Perl script (available on request). Fields including chromosome, LocusLink ID, mRNA accession, protein accession, and the type of substitution were captured. SNP validation status and MAF were extracted from plain text (msSQL) tables obtained from the dbSNP FTP site; chromosome position and DNA strand orientation were downloaded from the University of California Santa Cruz (UCSC) Genome Browser Annotation database (UCSC Human Genome Build hg17; ref. 21, 22) FTP site (23).

For nsSNPs in genes with more than one mRNA transcript, individual entries were recorded for each unique transcript to reflect potential differences in amino acid numbering. Individual entries were also recorded where more than one allele frequency submission (termed ''ss_submission'') was available. Therefore, a nsSNP with two mRNA transcripts and three different ss_submissions resulted in a total of six separate entries.

A total of 48,492 nsSNPs were filtered to exclude variants that mapped to the genome more than once (n = 2,602; 5.4%) and variants not classified as true biallelic nsSNPs. SNPs that were insufficiently validated were also excluded; variants in dbSNP are validated by either multiple independent submissions; frequency/genotype data; alleles observed in at least two chromosomes; or submission by the HapMap Consortium (n = 26,312; 54.3%; refs. 24, 25). Entries were further filtered to remove nsSNPs that lacked MAF data regardless of the population source or genotyping method (n = 5,090; 10.5%) or SNPs located within nonannotated genes (n = 4,951; 10.2%). There were 15 nsSNPs where the reported chromosomal location differed between the dbSNP and UCSC databases; in each case, the SNP was unplaced on the relevant contig within dbSNP and UCSC coordinates were entered.

Prediction of Potential nsSNP Functionality. We applied three *in silico* algorithms—the Grantham Scale (5), the PolyPhen algorithm (6, 26), and the SIFT algorithm (7, 27)—to predict the putative effect of each nsSNP on protein function.

Grantham scores, which categorize codon replacements into classes of increasing chemical dissimilarity, were designated conservative (0-50), moderately conservative (51-100), moderately radical (101-150), or radical (≥151) according to the classification proposed by Li et al. (28).

PolyPhen predicts the functional effect of amino acid changes by considering evolutionary conservation, the physiochemical differences, and the proximity of the substitution to predicted functional domains and/or structural features. Protein sequences within which nsSNPs were identified were obtained from the NCBI human RefSeq database (29, 30). Homologous protein structures were retrieved from the protein quaternary structure database (31) to determine the proximity of the substitution to annotated ''features'' in the SwissProt database (32, 33). Conversion between SwissProt and NCBI protein accession numbers was facilitated by the UCSC Annotation database ''kgxref'' table (UCSC FTP site). PolyPhen scores were designated probably damaging (≥2.00), possibly damaging (1.50-1.99), potentially damaging (1.25-1.49), borderline (1.00-1.24), or benign (0.00-0.99) according to the classification proposed by Xi et al. (34).

SIFT predicts the functional importance of amino acid substitutions based on the alignment of orthologous and/or paralogous protein sequences (7). SIFT (7, 35), which uses NCBI PSI-BLAST (36, 37), was obtained in stand-alone form from the author's website. Wild-type protein sequences were obtained from the NCBI human RefSeq database; sequences of related proteins for SIFT comparison (UniProt/Swiss-Prot and UniProt/TrEMBL) were sourced from the UniProt Resource (38, 39). Protein sequences used for alignments were extracted from RefSeq using the EMBOSS seqret and dbifasta tools (40, 41). All data were preprocessed for SIFT input by a series of Perl scripts; an additional script coordinated sequence retrieval, SIFT automation, and data collection (available on request). SIFT scores were classified as intolerant (0.00-0.05), potentially intolerant (0.051-0.10), borderline (0.101-0.20), or tolerant (0.201-1.00) according to the classification proposed by Ng et al. (7) and Xi et al. (34).

Statistical Analyses. Concordance between the functional consequences of each nsSNP predicted by the three *in silico* methods and the relationship between putative functionality and MAF was assessed using Spearman's rank correlation coefficient ρ. Confidence intervals for correlation coefficients were calculated via Fisher's transformation, whereby the correlation ρ is converted to a z-score using the formula $z = (1 / 2) \ln [(1 + \rho) / (1 - \rho)]$. Ninety-five percent confidence limits

were then calculated and transformed back to the correlation scale via the inverse Fisher transformation. Weighted averages and SDs of MAFs were computed using the number of chromosomes in each population data set as weights. Mean MAFs for each SNP were compared between the Caucasian and non-Caucasian populations by t tests assuming unequal variances, using Satterthwaite's formula to determine the appropriate degrees of freedom for the t distribution. Results were adjusted for multiple testing using a Bonferroni correction and tested at a 95% significance level. The numbers of total counts and number of submissions were compared across all SNPs between Caucasian and other populations using a paired t test.

Results

Candidate Gene Selection and nsSNP Curation. Figure 1 shows a flow diagram depicting the various stages (I-VIII) involved in the filtering, selection, and cataloguing of nsSNPs.

Stages I to IV: Retrieval of nsSNPs from dbSNP and Filtering of Annotated LocusLink Genes. From a total of 48,492 nsSNPs listed in dbSNP, we compiled a data set of 9,537 validated biallelic nsSNPs (19.7%) with frequency data from at least one population (regardless of composition, sample size, genotyping method, or MAF) located within 1 of 21,506 annotated

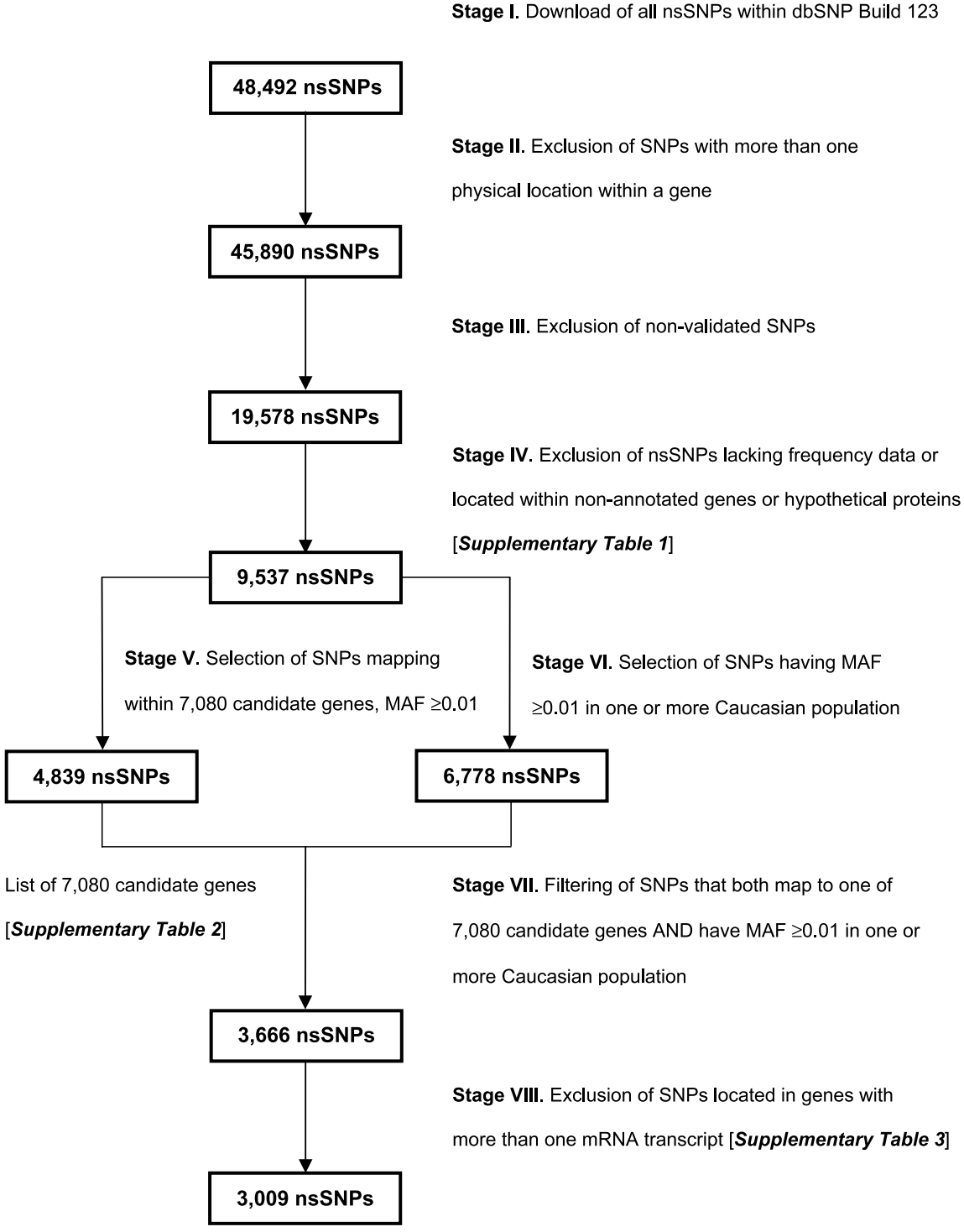

Figure 1. Flow diagram of the processes involved in selecting and filtering nsSNPs in candidate cancer genes.

Table 1. Distribution of the number of validated nsSNPs per gene

No. nsSNPs within the same gene	No. genes (total 21,506)	Cumulative no. nsSNPs (total 9,537)
0	16,574	0
1	2,797	2,797
2	1,112	5,021
3	485	6,476
4	245	7,456
5	94	7,926
6	66	8,322
7	48	8,658
8	33	8,922
9	14	9,048
10	10	9,148
11+	28	9,537

genes from the LocusLink database (Supplementary Table S1). The loss of 38,955 nsSNPs from the original download of 48,492 nsSNPs is attributable to nonvalidation or ambiguous positional data of SNPs or their location within nonannotated genes (Fig. 1). Only 4,932 of 21,506 annotated genes (23%) contained at least one validated nsSNP with associated MAF data (average of 1.9 nsSNPs per gene), suggesting that the screening and validation of nsSNPs to date has been biased toward a relatively small subset of genes. Distribution of the 9,537 nsSNPs across 21,506 annotated genes is shown in Table 1.

Stage V: Curation of Candidate Cancer Genes. Genes of potential relevance relevant to cancer *a priori* were identified by interrogation of Gene Ontology (n = 5,907), Kyoto Encyclopedia of Genes and Genomes (n = 1,754), PathwayAssist (n = 1,620), Entrez Gene (n = 1,070), and CancerGene (n = 2,527) databases. Excluding redundancy between lists yielded a total of 7,080 genes (Supplementary Table S2). The number of nsSNPs with frequency data in the 7,080 candidate genes was 5,188, with 2,608 candidate genes (37%) containing at least one nsSNP (average of 2.0 nsSNPs per gene). Stipulating a minimum MAF of ≥0.01 reduced the number of nsSNPs to 4,839 in 2,445 candidate genes (average of 2.0 nsSNPs per gene).

Stages VI to VIII: Curation of nsSNPs by Ethnicity. Two hundred and fifty-one of the 461 populations listed within dbSNP Build 123 were used to curate data on nsSNPs. The number of nsSNPs in 21,506 annotated genes with frequency data sourced from at least one of the 251 Caucasian populations was 8,723, with 4,675 genes (22%) containing at least one nsSNP (average of 1.9 nsSNPs per gene). Specifying a MAF ≥0.01 reduced the number of nsSNPs to 6,778 in 3,928 genes (average of 1.7 nsSNPs per gene). The number of nsSNPs in the specified 7,080 candidate cancer genes with MAF ≥0.01 in at least one of the specified 251 Caucasian populations was 3,666 nsSNPs in 2,052 genes. This accounted for 8% of the 48,492 nsSNPs reported in dbSNP Build 123 (Fig. 1). For each of the 2,547 nsSNPs with MAF information represented in both Caucasian and non-Caucasian populations, 1,737 (68%) had significantly different average MAF between populations (Supplementary Table S3). The number of chromosomes genotyped and number of unique ss_submissions per nsSNP were both significantly different between the 251 Caucasian and 210 non-Caucasian populations (mean total chromosome counts in Caucasian and non-Caucasian populations were 204 and 735, respectively; $P < 0.001$; mean numbers of ss_submissions in Caucasian and non-Caucasian populations were 2.5 and 3.1, respectively; $P < 0.001$; Supplementary Table S3).

Prediction of the Functional Effect of nsSNPs and Correlation between Algorithm Predictions

Putative Functionality of Curated nsSNPs. The distribution of the putative functional effect of validated nsSNPs with MAF ≥0.01 according to Grantham, PolyPhen, and SIFT algorithms is shown in Table 2. The data do not include nsSNPs resulting in premature termination of the wild-type protein; as such, predictions cannot be generated using these algorithms. For genes where there is more than one mRNA transcript, PolyPhen and SIFT predictions were often transcript-dependent, owing to differences in the predicted protein structure or surrounding amino acids. For consistency, we restricted correlations to 3,009 nsSNPs mapping within 1,711 genes with a single mRNA transcript (Supplementary Table S3).

Table 2. Distribution of functional predictions across validated nsSNPs with MAF ≥0.01 in Caucasian populations (data obtained from stages VII and VIII) using Grantham, PolyPhen, and SIFT algorithms

Score	Prediction of functionality	Predictions for 3,666 nsSNPs	Predictions for 3,009 nsSNPs in single mRNA genes	Predictions meeting minimum inclusion criteria
Grantham				
0-50	Conservative	1,544	1,264	1,264
51-100	Moderately conservative	1,513	1,241	1,241
101-150	Moderately radical	407	332	332
≥151	Radical	202	172	172
	No prediction	—	—	—
Total		3,666	3,009	3,009
PolyPhen				
0.00-0.99	Benign	N/A	1,263	706
1.00-1.24	Borderline	N/A	349	177
1.25-1.49	Potentially damaging	N/A	381	215
1.50-1.99	Possibly damaging	N/A	427	263
≥2.00	Probably damaging	N/A	230	162
	No prediction	N/A	359	—
Total			3,009	1,523
SIFT				
0.00-0.05	Intolerant	N/A	583	324
0.051-0.10	Potentially intolerant	N/A	176	117
0.101-0.20	Borderline	N/A	262	198
0.201-1.00	Tolerant	N/A	1,304	987
	No prediction	N/A	684	—
Total			3,009	1,626

Abbreviation: N/A, not applicable.

Table 3. Correlation between Grantham and SIFT, Grantham and PolyPhen, and PolyPhen and SIFT predictions

	Grantham prediction				
	Conservative	Moderately conservative	Moderately radical	Radical	Total
SIFT prediction					
Tolerant	471 (29.0)	418 (25.7)	79 (4.9)	19 (1.2)	987 (60.7)
Borderline	94 (5.8)	66 (4.1)	29 (1.8)	9 (0.6)	198 (12.2)
Potentially intolerant	46 (2.8)	53 (3.3)	13 (0.8)	5 (0.3)	117 (7.2)
Intolerant	100 (6.2)	128 (7.9)	57 (3.5)	39 (2.4)	324 (20.0)
Total	711 (43.7)	665 (40.9)	178 (10.9)	72 (4.4)	1626 (100)
					$\rho = -0.188$; 95% CI: -0.234, -0.141
	Grantham prediction				
	Conservative	Moderately conservative	Moderately radical	Radical	Total
PolyPhen prediction					
Benign	381 (25.0)	266 (17.5)	43 (2.8)	16 (1.1)	706 (46.4)
Borderline	79 (5.2)	82 (5.4)	14 (0.9)	2 (0.1)	177 (11.6)
Potentially damaging	97 (6.4)	96 (6.3)	18 (1.2)	4 (0.3)	215 (14.1)
Possibly damaging	84 (5.5)	115 (7.6)	52 (3.4)	12 (0.8)	263 (17.3)
Probably damaging	25 (1.6)	68 (4.5)	34 (2.2)	35 (2.3)	162 (10.6)
Total	666 (43.7)	627 (41.2)	161 (10.6)	69 (4.5)	1,523 (100)
					$\rho = 0.281$; 95% CI, 0.234, 0.326
	SIFT prediction				
	Tolerant	Borderline	Potentially intolerant	Intolerant	Total
PolyPhen prediction					
Benign	359 (34.2)	57 (5.4)	25 (2.4)	37 (3.5)	478 (45.5)
Borderline	76 (7.2)	20 (1.9)	8 (0.8)	15 (1.4)	119 (11.3)
Potentially damaging	92 (8.8)	21 (2.0)	12 (1.1)	32 (3.0)	157 (15.0)
Possibly damaging	68 (6.5)	27 (2.6)	17 (1.6)	72 (6.9)	184 (17.5)
Probably damaging	32 (3.0)	10 (1.0)	10 (1.0)	60 (5.7)	112 (10.7)
Total	627 (59.7)	135 (12.9)	72 (6.9)	216 (20.6)	1050 (100)
					$\rho = -0.442$; 95% CI, -0.490, -0.392

NOTE: Percentages are shown in parentheses.
Abbreviation: 95% CI, 95% confidence interval.

Grantham scores were generated for all 3,009 nsSNPs mapping to a single mRNA transcript (Table 2). PolyPhen predictions were obtained for 2,650 of the 3,009 nsSNPs (88%) from 1,529 of the 1,711 genes (89%). We excluded predictions generated using less than six protein sequences in the alignment as these could be unreliable, yielding a total of 1,523 PolyPhen predictions used to calculate correlations (Table 2). SIFT scores could be generated for 2,325 of the 3,009 nsSNPs (77%) from 1,382 of the 1,711 genes (81%). As with PolyPhen, SIFT predictions based on fewer than six aligned sequences were not included in analyses as these have been shown to be unreliable (42). Also excluded were SNPs where the median SIFT sequence conservation score was >3.25 as such scores can indicate that the substitution is at a position that is evolving and could, therefore, be unstable (43). After these exclusions, a total of 1,626 SIFT predictions were curated (Table 2).

Correlation between Grantham, PolyPhen, and SIFT predictions. Table 3 shows the relationship between the functional consequences of nsSNPs predicted by each of the three predictive algorithms for 3,009 nsSNPs located within genes with a single mRNA transcript. Correlations were calculated from raw scores rather than the arbitrarily defined categories in Materials and Methods. There was significant correlation between the predictions obtained using Grantham and SIFT algorithms ($\rho = -0.188$; $P < 0.0001$); however, 228 nsSNPs with Grantham scores deemed at worst moderately conservative (≤100) had SIFT scores indicative of intolerant substitutions (≤0.05). SIFT predictions for these nsSNPs were based on between 7 and 400 sequences per alignment (median number, 72) across 205 genes. Concordance between the Grantham and PolyPhen predictions was stronger than that between Grantham and SIFT predictions ($\rho = 0.281$; $P < 0.0001$). The strongest concordance was observed between PolyPhen and SIFT predictions ($\rho = -0.442$; $P < 0.0001$).

Relationship between Putative Functionality of nsSNPs and MAF. To examine the relationship between putative functionality of polymorphisms and MAF, we restricted our analysis to nsSNPs for which MAF data were available within the Perlegen European American AFD_EUR_PANEL (dbSNP Population_ID 1371). This panel is based on genotypes generated from 24 unrelated individuals of European-American descent drawn from the Coriell Cell Repository. Although only alleles with frequency of at least 0.03 will have 80% probability of being represented, the panel provides the most extensive MAF data across our final nsSNP list, with submissions for 1,090 of the 2,099 (52%) nsSNPs scored by at least two of the three predictive algorithms. Moreover, using a single population panel removes potential bias associated with averaging MAF across populations that vary in size and composition.

The relationship between Grantham, PolyPhen, and SIFT predictions and MAF derived from the Perlegen European-American population panel is shown in Table 4. An association between predicted functionality derived from each of the three algorithms and MAF was observed, with correlations predicted by PolyPhen ($\rho = -0.113$; $P = 0.0019$) and SIFT ($\rho = 0.171$; $P < 0.0001$) being significant.

Discussion

Not all nsSNPs directly affect protein function. In the absence of functional data, it is advantageous to use computational

tools to identify those most likely to affect wild-type protein function. To facilitate sequence-based association studies, we have created a database of the Predicted Impact of Coding SNPs (PICS) for cancer association studies using information archived by NCBI dbSNP, a public database with an open submission policy whereby SNP data are accepted regardless of allele frequency or genotyping technique. Inevitably, there are errors in submissions, a potential problem that dbSNP has attempted to address by designating SNPs as either ''validated'' or ''unvalidated.'' To avoid including erroneous SNPs, we catalogued only validated nsSNPs with frequency submissions.

Methods to predict the effect of substitutions on protein structure and/or function fall into four broad categories—those based on physiochemical differences (5, 44), protein sequence alignment (45), mapping to known protein three-dimensional structures (46), and combinations thereof (6, 7, 47-50).

Here, we assessed the effect of validated nsSNPs in candidate genes with single mRNA transcripts using three commonly used predictive methods—the Grantham matrix, PolyPhen, and SIFT algorithms—all of which can be automated to process large data sets. The Grantham scale, which scores substitutions based on assessment of chemical dissimilarity between residues, represents one of the first attempts to predict the severity of amino acid substitutions on protein structure and has been used frequently to ascertain the functional effect of SNPs (28, 48, 51, 52) or modified to underpin new predictive algorithms (44). Across the 3,666 validated nsSNPs, 6% were predicted to be radical substitutions according to the Grantham matrix, slightly higher than the 4% reported by Stephens et al. (52) in analysis of 565 nsSNPs. Although there has been some debate about the ability of the Grantham matrix to predict deleterious substitutions (48, 51), we found a significant concordance between predictions obtained using the Grantham scale and the more complex alignment-based PolyPhen and SIFT algorithms.

To date, there is relatively limited data on direct evaluations of programs, such as PolyPhen and SIFT, on protein function. Xi et al. (34) found that that PolyPhen and SIFT correctly predicted the effect of 96% of amino acid substitutions on protein activity in *APEX1*. Most data on the validity of these algorithms has, however, come from benchmarking studies based on the analysis of ''known'' deleterious substitutions annotated in databases, such as SwissProt. In such studies, PolyPhen and SIFT has been shown to successfully predict the effect of over 80% of amino acid substitutions (34, 35, 42, 53).

We obtained PolyPhen predictions for 1,523 of the 3,009 nsSNPs, with 425 (28%) predicted to have functional effects. This finding is compatible with observations made by Ramensky et al. (6), who reported 28% of validated nsSNPs in the Human Genome Variation database (54, 55) predicted to be damaging. Across genes with a single mRNA transcript, we obtained SIFT predictions for 1,626 of the 3,009 nsSNPs (54%), virtually identical to the figure obtained by Ng and Henikoff (35) from an analysis of 5,780 nsSNPs from dbSNP Build 95. In total, we identified 324 (20%) intolerant substitutions, comparable with predictions reported by Ng and Henikoff (35). We rejected SIFT predictions based on divergence scores ≥ 3.25 and those with less than six sequences in the alignment (42), which accounted for 30% of predictions. Significant concordance was observed between predictions generated by PolyPhen and SIFT, perhaps not surprising given that both are based on similar concepts.

Table 4. Relationship between functionality of nsSNPs predicted by Grantham, PolyPhen, and SIFT and MAF based on dbSNP Population 1371 (Perlegen European Americans)

	Grantham prediction				
	Conservative	Moderately conservative	Moderately radical	Radical	Total
MAF					
≤0.05	84 (7.7)	82 (7.5)	20 (1.8)	13 (1.2)	199 (18.3)
0.051-0.15	145 (13.3)	123 (11.3)	24 (2.2)	13 (1.2)	305 (28.0)
0.151-0.30	133 (12.2)	124 (11.4)	21 (1.9)	11 (1.0)	289 (26.5)
>0.30	145 (13.3)	112 (10.3)	29 (2.8)	11 (1.0)	297 (27.3)
Total	507 (46.5)	441 (40.5)	94 (8.6)	48 (4.4)	1,090
				$\rho = -0.035$; 95% CI: −0.094, 0.025	
	PolyPhen prediction				
	Benign	Borderline	Potentially damaging	Possibly damaging	Probably damaging	Total
MAF						
≤0.05	59 (7.8)	15 (2.0)	19 (2.5)	28 (3.7)	22 (2.9)	143 (18.9)
0.051-0.15	99 (13.1)	20 (2.6)	34 (4.5)	43 (5.7)	25 (3.3)	221 (29.2)
0.151-0.30	87 (11.5)	27 (3.6)	23 (3.0)	31 (4.1)	12 (1.6)	180 (23.8)
>0.30	114 (15.1)	22 (2.9)	32 (4.2)	24 (3.2)	20 (2.6)	212 (28.0)
Total	359 (47.5)	84 (11.1)	108 (14.3)	126 (16.7)	79 (10.5)	756
					$\rho = -0.113$; 95% CI: −0.183, −0.042	
	SIFT prediction				
	Tolerant	Borderline	Potentially intolerant	Intolerant	Total
MAF					
≤0.05	90 (10.6)	23 (2.7)	14 (1.7)	40 (4.7)	167 (19.7)
0.051-0.15	153 (18.1)	26 (3.1)	18 (2.1)	52 (6.1)	249 (29.4)
0.151-0.30	153 (18.1)	32 (3.8)	18 (2.1)	25 (3.0)	228 (27.0)
>0.30	143 (16.9)	19 (2.2)	13 (1.5)	27 (3.2)	202 (23.9)
Total	539 (63.7)	100 (11.8)	63 (7.4)	144 (17.0)	846
				$\rho = 0.172$; 95% CI: 0.105, 0.236	

NOTE: Percentages are shown in parentheses.

Random surveys of SNPs have shown a nonuniform distribution of alleles, with the numbers of SNPs increasing with decreasing MAF (56). This has been hypothesized to provide insight into the allelic architecture of disease susceptibility with functional SNPs skewed toward the lower end of the frequency distribution (4). Alleles that are functionally deleterious will tend to be selected against and thus underrepresented at high frequencies. Such a tenet is supported by our analysis, which showed a relationship between putative functionality and MAF, concordant with the observations made by Leabman et al. (51).

The sequence-based approach to association analyses makes use of SNPs, which may directly affect protein structure and/or function, be they missense or nonsense mutations that alter protein sequence, splice-site variants, or those located within putative promoter regions and transcription sites. This pool of 3,009 fully annotated nsSNPs is accessible from the PICS database at http://www.icr.ac.uk/cancgen/molgen/MolPopGen_PICS_database.htm. Curation and release of updated versions of the PICS database will continue as new data becomes available.

References

1. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. Oncogene 2004;23:6471–6.
2. Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer 2004;4:850–60.
3. Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33:D39–45.
4. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 2003;33 Suppl:228–37.
5. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–4.
6. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30:3894–900.
7. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res 2001;11:863–74.
8. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
9. Available from: http://www.geneontology.org/ [Release go_200411].
10. Kanehisa M, Goto S, Kawashima S, et al. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277–80.
11. Available from: http://www.genome.jp/kegg/ [Release 32.0].
12. Available from: www.iobion.com/news/hotnews.html?cmd=Retrieve&dopt=Abstract [Release 2.5].
13. Maglott D, Ostell J, Pruitt KD, et al. Entrez gene: gene-centered information at NCBI database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33 Database Issue:D54–8.
14. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene [Accessed 2004 November].
15. Available from: http://caroll.vjf.cnrs.fr/cancergene/HOME.html [Accessed 2004 July].
16. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res 2000;28:126–8.
17. Romualdi C, Balding D, Nasidze IS, et al. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 2002;12:602–12.
18. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
19. Available from: http://www.ncbi.nlm.nih.gov/SNP/ [Build 123].
20. Available from: ftp://ftp.ncbi.nih.gov/snp/human/XML/ [Build 123].
21. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006.
22. Available from: http://genome.ucsc.edu/index.html [Human Genome Build hg17].
23. Available from: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/snp.txt.gz [Human Genome Build hg17].
24. The International HapMap Project. Nature 2003;426:789–96.
25. Available from: http://www.hapmap.org/.
26. Available from: http://www.bork.embl-heidelberg.de/PolyPhen/ [Accessed 2004 December].
27. Available from: http://blocks.fhcrc.org/sift//SIFT.html [Version 2.1].
28. Li WH, Wu CI, Luo CC. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 1984;21:58–71.
29. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005;33:D501–4.
30. Available from: http://www.ncbi.nlm.nih.gov/RefSeq/ [Accessed 2004 December].
31. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–61.
32. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank: current status. Nucleic Acids Res 1994;22:3578–80.
33. Available from: http://www.ebi.ac.uk/swissprot/ [Accessed 2004 December].
34. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics 2004;83:970–9.
35. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res 2002;12:436–46.
36. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
37. Available from: http://www.ncbi.nlm.nih.gov/blast/ [Accessed 2004 December].
38. Bairoch A, Apweiler R, Wu CH, et al. The universal protein resource (UniProt). Nucleic Acids Res 2005;33:D154–9.
39. Available from: http://www.ebi.uniprot.org/ [Accessed 2004 December].
40. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–7.
41. Available from: http://emboss.sourceforge.net/.
42. Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. Cancer Epidemiol Biomarkers Prev 2004;13:801–7.
43. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
44. Krawczak M, Ball EV, Cooper DN. Neighbouring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 1998;63:474–88.
45. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–9.
46. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat 2001;17:263–70.
47. Sunyaev S, Ramensky V, Koch I, Lathe W, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591–7.
48. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 2001;10:2319–28.
49. 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.
50. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. Structural location of disease-associated single-nucleotide polymorphisms. J Mol Biol 2003;327:1021–30.
51. Leabman MK, Huang CC, DeYoung J, et al. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci U S A 2003;100:5896–901.
52. Stephens JC, Schneider JA, Tanguay DA, et al. Haplotype variation and linkage disequilibrium in 313 human genes. Science 2001;293:489–93.
53. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 2000;16:198–200.
54. Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 2002;30:387–91.
55. Available from: http://hgvbase.cgb.ki.se/.
56. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999;22:231–8.