UCRL-JRNL-200825

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function.

T. Xi, I. M. Jones and H. W. Mohrenweiser

Issued in 2004

Genomics (2004) 83, 970-979

Disclaimer

Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function

Tong Xi, Irene M. Jones and Harvey W. Mohrenweiser[*,a]
Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory
Livermore CA, 94550

Running Title: DNA repair gene variants
Key Words: DNA repair genes; genetic variation; polymorphisms, predicted protein activity

[*]Corresponding author - Mailing address:
Dr. Harvey W. Mohrenweiser
Epidemiology Division, 224 Irvine Hall
Department of Medicine
University of California, Irvine
Irvine, CA 92697-7550

Phone: 949-824-5781
FAX: 949-824-4773
Email: hmohrenw@uci.edu

[a]current address:
Epidemiology Division, Department of Medicine
University of California, Irvine
Irvine, CA 92697-7550

Abstract

Over 520 different amino acid substitution variants have been previously identified in the systematic screening of 91 human DNA repair genes for sequence variation. Two algorithms were employed to predict the impact of these amino acid substitutions on protein activity. Sorting Intolerant From Tolerant (SIFT) classified 226 of 508 variants (44%) as "Intolerant". Polymorphism Phenotyping (PolyPhen) classed 165 of 489 amino acid substitutions (34%) as "Probably or Possibly Damaging". Another 9-15% of the variants were classed as "Potentially Intolerant or Damaging". The results from the two algorithms are highly associated, with concordance in predicted impact observed for ~62% of the variants. Twenty one to thirty one percent of the variant proteins are predicted to exhibit reduced activity by both algorithms. These variants occur at slightly lower individual allele frequency than do the variants classified as "Tolerant" or "Benign". Both algorithms correctly predicted the impact of 26 functionally characterized amino acid substitutions in the APE1 protein on biochemical activity, with one exception. It is concluded that a substantial fraction of the missense variants observed in the general human population are functionally relevant. These variants are expected to be the molecular genetic and biochemical basis for the associations of reduced DNA repair capacity phenotypes with elevated cancer risk.

Introduction

The availability of genomic sequence has provided the infrastructure for addressing questions related to genetic variation in the human population. For example, the SNP Consortium effort identified over 1.2 million single nucleotide polymorphisms (SNPs) in screening samples from a limited number of individuals [1]. More directed efforts have identified many common missense variants in genes with potential associations with susceptibility to common human diseases, e.g. cardiovascular disease, asthma, hypertension and rheumatoid arthritis [2-6] and functionally related genes [7-11]. The genes encoding proteins with roles in the repair of damaged DNA are one family of genes where large datasets exist from the systematic screening for DNA sequence variation. An average of over five different amino acid substitution variants per gene has been observed in the sequencing of 84 DNA repair genes in ~90 generally healthy individuals [11, ref. 12 summarizes extensive data from three data bases]. Seventy percent of the missense variants in the repair genes exist at less than 2% individual frequency and only 6% of the variant alleles occur at frequencies of greater than 20%, a distribution similar to that observed for nonsynonymous variants in other sets of genes.

It is well established that variants of DNA repair genes are associated with inherited disease. Cancer syndromes associated with variation in DNA repair genes and loss of repair capacity include breast cancer [13] and colon cancer families [14]. Cancer families account for only a few percent of the cancers in the population, as observed for most common diseases. More subtle variations in capacity to repair different classes of DNA damage also exist. Studies of repair capacity and mutagen sensitivity phenotypes, where 20-40% deviations from the population mean are observed in 10-20% of the general population [15-17], clearly establish that individuals with reduced capacity to repair damaged DNA are at elevated cancer risk [16,18,19]. These highly heritable phenotypes integrate the impact of variation in the genes in the repair pathway(s) being interrogated [20,21].

At a functional level, only a few of the many missense variants in repair genes have been studied. Employing biochemical assays, four of seven of the variants of *APEX1* (the protein is commonly referred to as APE1) identified in population studies retained only 10-60% of normal or wildtype activity [22]. Concordance of predicted impact of a substitution employing molecular replacement modeling and measured enzymatic activity was observed for six of the seven variants. Additional evidence that polymorphic variants impact protein and repair pathway activity are the associations of several polymorphic variants with reduced DNA repair capacity [23-25] and modest elevations in cancer risk [26]. The impact of most variants in these genes on repair activity or susceptibility has yet to be addressed

In silico methods have been developed to predict the potential of amino acid substitutions to impact protein structure and activity [27-35]. The underpinnings for these algorithms are sequence conservation over evolutionary time, the physical and chemical properties of the exchanged residues and/or protein structural domain information. Different algorithms emphasize different

aspects of this knowledge. The SIFT (Sorting Intolerant From Tolerant) algorithm emphasizes sequence homology among related genes and domains over evolutionary time and the characteristics of the amino acid residues [27-29] in predicting the impact of amino acid substitutions. The PolyPhen (Polymorphism Phenotyping) algorithm also incorporates sequence conservation and the nature of the amino acid residues involved, but also values the location of the substitution within identified functional domains and known structures and structural features of the protein available in the annotated data base SwissPro [32,34]. These algorithms were approximately 80% successful in benchmarking studies employing amino acid substitutions assumed to have a major negative impact on the residual activity of the variant protein as the test set [28,31,32,35]. Measures of evolutionary conservation of sequence and SIFT scores predicted that 38 and 70, respectively, of 139 missense variants in exon 11 of *BRCA1* would impact function [36]. Similarly, a fraction of the variants in a series of transmembrane proteins [8], disease associated genes [10] and DNA repair genes [11] had characteristics expected of variants impacting function.

We report the results of applying the SIFT and PolyPhen algorithms to a set of 523 amino acid substitutions identified in the systematic screening of 91 DNA repair genes for sequence variation in the general population. We find that 30-50% of these variants are predicted to have a negative impact on protein activity that could result in reduced repair capacity and therefore be associated with elevated genetic susceptibility and cancer risk.

Results

Prediction of impact of substitutions on function and association with allele frequency

The dataset for analysis of the potential impact of common polymorphisms in DNA repair genes was 523 amino acid substitution variants identified in systematic sequencing of 91 human DNA repair and repair related genes (summarized in Methods, Table 5). No missense variants were identified in the screening of 9 genes.

PolyPhen scores were obtained for 489 missense variants in 81 genes. Table 1 presents the distribution of the variants by PolyPhen score. To provide an overview of the distribution of PolyPhen scores, the scores are placed into 8 groups. PolyPhen scores of >2.0, scores expected to be “Probably Damaging” to protein structure and function [33], account for 12.5% of the variants. An additional 21.3% of the variants exhibited PolyPhen scores of 1.99-1.50, scores indicative of variants that are “Possibly Damaging” to protein function. In total, 33.8% of the variants exhibit PolyPhen scores of greater than 1.5 and are designated as variants “Probably or Possibly Damaging” to function. We have introduced the categories “Potentially Damaging” and “Borderline” to extend the application of the algorithm further into the range of impacts that may be relevant to disease susceptibility. The focus is to capture variant proteins retaining sufficient activity to not “cause” monogenic disease, but with sufficiently reduced activity to increase risk of disease, particularly following an exposure. The variants with PolyPhen scores of 1.49-1.25 were

designated “Potentially Damaging” to function and accounted for 15.5% of the variants. A total of 49.3% of the variants have PolyPhen scores of greater than 1.25. Scores of 1.24-1.00 were designated as “Borderline”, providing a buffer between the PolyPhen scores expected to designate a variant as “Damaging” and the scores expected to designate “Benign” variants with high probability. Although no significant (inverse) correlation of variant allele frequency and PolyPhen score was observed (r=-0.13), the average allele frequency for 241 variants classified as “Probably, Possibly or Potentially Damaging” (0.031) was significantly less than the average allele frequency for the 248 variants predicted to “Borderline” or “Benign” (0.048) (Wilcoxon test; p=0.016). Elevated allele frequency is most apparent in the two groups of variants with PolyPhen scores of <0.50.

Similar data for 508 variants in 82 genes analyzed with the SIFT algorithm are in Table 2. Note the directionality of the SIFT and PolyPhen scores are opposite and the SIFT scores are limited to the range of 0.0 to 1.0, while the PolyPhen scores in this dataset ranged from 3.17 to 0.0. Twenty eight percent of the amino acid substitution variants exhibit SIFT scores of 0.0. Another 16.9% of the variants have scores between 0.01 and 0.05. Thus, 44.5% of the polymorphic amino acid substitution variants are classified [27] as “Intolerant” variants by SIFT. Again, given the additional interest in variants with more modest impact on protein activity, the variants with SIFT scores of 0.051-0.10 have been designated as “Potentially Intolerant”. The “Potentially Intolerant” group includes 45 additional variants or 8.9% of the variants scored by SIFT. No significant correlation of allele frequency and SIFT score was identified (r=0.08), but as with PolyPhen, the average allele frequency for 271 variants classified as “Intolerant” or “Potentially Intolerant” (0.038) was less than the frequency for the 237 variants classified as “Borderline” or “Tolerant” (0.045) (Wilcoxon test, p=0.049). The “Tolerant” variants with SIFT scores of >0.50 have the highest average allele frequency.

Concordance between PolyPhen and SIFT predictions

An obvious question is how concordant are the predictions of impact of the individual amino acid substitutions on function, given that the two algorithms employ different approaches and also different datasets as foundations for their analysis. Table 3 presents the relationship between SIFT and PolyPhen scores when the distribution of scores for each algorithm is reduced to three groups predicted to potentially impact function, plus one group unlikely to impact function and a group at the interface (Borderline) between the “damaging” and “tolerated” scores, as described in Tables 1 and 2. Agreement of probability of negative impact (Table 3, yellow shaded cells), that is a SIFT score of 0.00 and a PolyPhen score of >2.00, a SIFT score of 0.01-0.05 and a PolyPhen score of 2.00-1.50, etc, was observed for 35% of the 479 variants scored with both algorithms. Employing a broader definition of agreement, with the cells adjacent to the diagonal cells (Table 3, turquoise) included as being in general agreement, concordance is observed for 62% of the variants. The SIFT and PolyPhen scores are highly associated (chi-square of 27.9, one degree-of-freedom, p<0.0001). The most noticeable discrepancy involves the set of 28 variants with SIFT scores of

0.00 (high likelihood of negative impact) and PolyPhen scores of less than 1.0 (Benign). The variants in this group are not from a small number of genes and the discrepancies between the algorithms do appear to be associated with obvious aspects of data quality or quantity. With the grouping employed in Table 3, 21% of the variants are predicted by PolyPhen to be "Probably or Possibly Damaging" and "Intolerant" by SIFT (bolded numbers). Another 10.7% of the variants or a total of 32% are predicted by both algorithms to have at least the "Potential" to have a negative impact on protein activity (numbers in italics).

Benchmarking prediction algorithms with biochemically characterized variants

As a test for the ability of the SIFT and PolyPhen algorithms to identify substitutions impacting enzymatic activity of DNA repair proteins, scores were obtained for a series of 26 previously characterized amino acid substitution variants of *APEX1*. *APEX1* encodes an endonuclease (APE1) critical for removal of apurinic sites and normal processing during base excision repair [37]. The variants were biochemically characterized during either the course of site directed mutagenesis studies of the enzyme mechanism [38-41] or as part of a study to characterize amino acid substitution variants observed in the human population [22]. The biochemical activity of the APE1 variants and the SIFT and PolyPhen scores are in Table 4. The algorithms were concordant in their predictions. With the exception of the Ala substitution at residue 70, a variant that retained only 4% of wildtype activity, the algorithms correctly predicted the impact of the substitutions on protein activity. The algorithms did correctly predict the negative impact of another substitution, Arg for Glu, at residue 70, a variant also retaining 4% activity, although the score was less than the usual cutpoint for the "Possibly Damaging" designation. The six variants retaining 10-25% residual activity were scored as "Probably Damaging" by PolyPhen and "Intolerant" by SIFT. Both of the variants retaining 36% of normal activity were scored as "Possibly Damaging" and "Intolerant". The variant retaining 60% activity was classified as "Benign" and "Tolerant" (without negative impact) as were the two variants retaining wild-type activity. Two variants identified in population screening [22], 241R and 306A, retained wild-type activity, but were designated as "Potentially Intolerant" and "Borderline" substitutions respectively by SIFT and "Borderline" by PolyPhen using the extended groupings described above. Employing the two category Intolerant versus Tolerant and Probably/Possibly Damaging versus Benign scoring convention, the algorithms correctly identified the functional impact of 24 (or 25 if the variant retaining 60% residual activity is considered wildtype) of the 26 variants in the dataset, including seven variants where the substitution impaired the ability of the protein to bind to substrate and thus had a more indirect impact on enzymatic activity.

Discussion

It is generally acknowledged that most of the burden of common disease in the population exists as sporadic or non-inherited cases and results from modest exposures in individuals with elevated

susceptibility. This is expected to be especially important in diseases with late age-of-onset [42-44]. The hypothesis that elevated susceptibility is associated with polymorphic variants in disease related genes, the “common variant-common disease hypothesis”, has been extensively discussed [43-50]. Assumptions of this hypothesis include the polygenic nature of disease susceptibility in the general population and that a (substantial) fraction of the common polymorphisms existing in the population will impact protein function. An alternative hypothesis is that common diseases, which are often late age-of-onset diseases, are associated with a large number of different low frequency polymorphisms (rare variants), the “common disease-rare variant” hypothesis [42-44,50,51]. Documenting that a substantial fraction of the amino acid substitution variants observed in the screening of the general population had a negative impact on the activity of variant proteins and the distribution of variants between the classes predicted to be functionally relevant and non-relevant may provide insight into the relative merits of the hypotheses. The large number of candidate variants already identified in DNA repair and other disease associated genes makes it infeasible to characterize all of the variants with biochemical and biological assays, thus *in silico* methods must be employed for addressing the potential functional relevance of these amino acid substitutions.

The benchmarking of the SIFT and PolyPhen algorithms with APE1 variants characterized for level of enzymatic activity confirmed that both algorithms are quite capable of discriminating between variants with minimal residual activity and variants with wild-type activity. These data are consistent with the other benchmarking studies that generally employed highly penetrant variants associated with a genetic disease [28,34,35] or impaired ability of cells to grow [31] as the test datasets. The study of the APE1 variants suggests that the “false negative” error rate is reasonably small. The few APE1 variants retaining at least 50% of wild-type activity preclude a meaningful estimation of a “false-positive” rate. In an analysis of variants associated with disease, a false positive rate of 8-9% was derived [33]. The ability of *in silico* approaches to identify variants retaining 20-50% residual activity is important. These are variants expected to be the basis for the proposed association of common variants with common disease and the focus of studies of individual susceptibility and disease risk in the general population.

From 33 to 53% of the amino acid substitutions in the DNA repair genes were predicted to negatively impact protein activity, depending upon the algorithm used and assumptions regarding scores indicative of a negative impact on activity. The 62% general concordance between the two algorithms in predicting the impact of these substitutions on function is encouraging, as the algorithms employ different approaches and types of reference data for their predictions, different scales for scoring and the boundaries for grouping the variants were somewhat arbitrary. Also, the number of sequences available for inclusion in the analysis of any variant and the quantity of data regarding the domain structure of a protein are quite variable. Non-agreement regarding specific variants could also reflect the high level of evolutionary sequence conservation of the repair genes. Over 75% of the amino acid substitutions in a subset of the genes occurred at residues that were identical in humans and mice [11]. It is notable that many processes in repairing damaged DNA require proteins to function in multiprotein complexes. Sequence conservation is expected to be

maintained in these protein interaction and communication domains, in addition to the conservation expected in the substrate binding and catalytic domains. SIFT, which emphasizes sequence conservation and classified more variants (44.5%) as having a negative impact than PolyPhen (33.8%), may be more sensitive in identifying the potential impact of variation in these important regions of the protein, regions not generally well characterized nor annotated at this time.

New and modified approaches for predicting protein folding and structure [52,53] and increased knowledge of the characteristics of variants associated with high penetrance disease [54-59] as well as the increased availability of protein structure data should provide opportunities to make enhancements to the currently available algorithms. Of special interest for the DNA repair genes will be information about amino acid residues involved in the interaction among subunits of the multiprotein complexes required for efficient repair of damaged DNA prior to replication and cell division. New methods are also required to identify variants with increased specific activity, a potential not addressed by current algorithms. Ideally, *in silico* approaches will eventually predict the quantitative impact of a substitution on protein and pathway activity [12,60].

Establishing different decision-points so as to flag variants retaining more than marginal residual activity but with still potentially meaningful reductions in activity would be helpful in identifying variants for further analysis in functional assays or molecular epidemiology studies. This group of variants is expected to be important in explaining the individual variation in repair capacity and susceptibility. A negative impact was predicted for 21% of the variants scored by both algorithms using the stringent criteria originally suggested for these algorithms. Inclusion of the next 5% of the scale for SIFT scores as "Potentially Intolerant", adds 8.9% of the variants to the affected category (Table 2), while 15.5% of the variants exist in the "Potentially Damaging" group in the PolyPhen scores (Table 1). Assuming that the variants in the "Potentially Intolerant and Damaging" category exhibit reduced activity, 31.3% of the variants are predicted by both algorithms to have at least the potential to impact activity.

The prediction that 20-50% of the large number of amino acid substitution variants observed in DNA repair genes will impact function (Tables 1-3) is consistent with results from other studies of DNA repair gene variants using these algorithms [36,61] or other approaches to identify the subset of amino acid substitution variants expected to impact protein activity in a range of genes. Fleming et al. [36] using an approach termed "ancestral sequences" that was capable of identifying 85% of known detrimental missense variants in *HBB* and *BRCA2*, predicted that 38 of 139 missense variants in *BRCA1* (from the Breast Cancer Information Core) were important for normal protein function. The SIFT algorithm classified 36 of the 38 variants as "Intolerant". SIFT also predicted an additional 34 BRCA1 variants would have reduced function. Similar analyses have been conducted of 155 missense variants identified in population screening of 24 transmembrane genes [8] using BLOSUM62 [62], SIFT [27] and Grantham [63] scores. Thirty six percent of 185 missense variants in 104 candidate disease genes were defined as non-conservative exchanges by the BLOSUM62 matrix [10]. Although the data on concordance among different approaches for

predicting the probability of a substitution impacting function is limited, the data are quite consistent in predicting that 30-50% of the substitutions will impact function.

The potential for substitutions to result in a protein with elevated specific activity is not addressed with the available algorithms. Such variants could have a protective impact, but hypermorphs can also be associated with disease [64]. Also not addressed is the impact of multiple substitutions in a subunit on the activity of the protein or the impact of substitutions in the homozygous or compound heterozygous variant individual. Preliminary analysis of the available genotype data suggest that both situations occur. For example, the linkage disequilbrium analysis infers that the three most common amino acid substitution variants in *XRCC1* each exist on a different chromosome (haplotype). In contrast, three of the four most polymorphic amino acid substitutions in *ERCC6* exhibit strong linkage disequilbrium, suggesting they exist on a single chromosome and will encode a protein subunit with three amino acid substitutions (unpublished data).

The fraction of variants in the repair genes predicted to be functionally important is also consistent with results from biochemical analyses of both repair genes and members of other gene families. Four of seven polymorphic variants of *APEX1* identified in the population retained less than 60% of normal activity [22]. Studies of other proteins arrive at similar conclusions for variants identified in population based screenings. Five of 15 amino acid substitution variants in *OCT1* exhibited reduced ability to transport organic cations [65]. All four of the polymorphic variants of *OCT2* characterized (a subset of the eight polymorphisms identified) exhibited reduced ability to transport a cationic xenobiotic [66]. Seven of 16 variants of *SLC21A6* impacted activity [67]. Functional studies of the common variants of five drug metabolizing enzymes (*CYP3A4*, *CYP2C19*, *CYP2J2*, E*PHX2*, *CYP3A5*) report that 13 of 27 variants exhibited reduced activity [68-72]. Other support for the functional importance of the variants are the reports that highly polymorphic variants of *ERCC2*, *XPC*, *XPA* and *XRCC1* can have negative impacts on measures of DNA repair activity [23-25,73,74]. These latter studies do not provide insight into the fraction of repair gene variants impacting activity, but do confirm that high allele frequency variants are associated with a negative impact on the activity of a repair pathway. Overall, these functional studies support the *in silico* prediction that 30-50% of the amino acid substitution variants identified in systematic screening of population based samples should have a measurable negative impact on activity. It can thus be projected that 1-2 functionally relevant variants will exist in a typical gene, given that 3-5 different amino acid substitution variants per gene are observed in these systematic screenings [11,12]. Particularly given that 20-40 genes have roles in each of the different repair pathways [12,75], the DNA repair gene variants predicted to negatively impact function have the collective potential to contribute substantially to the variation in repair capacity and mutagen sensitivity and thus have a role in variation in individual susceptibility.

Over 90% of the variants included in these analyses of repair gene variants are estimated to exist at individual allele frequencies of less than 10% [11,12], with 70% of the amino acid substitutions existing at allele frequencies of <0.02. The variants existing at individual frequencies of <0.10 make important contributions to the total genetic variation as they account for 34% of the

genetic variation in the general population associated with amino acid substitutions in these genes. Although the very large number of low frequency alleles in this data set limits the power to identify associations of allele frequency and predicted impact, the low frequency variants were more likely to be in the groups of variants predicted to impact protein activity. This is consistent with previous observations [6] or assumptions [9] of analyses of other genes. This observation could be considered support for the "common disease-rare variant" hypothesis [42,50,51]. But, the alleles predicted to be functionally relevant are not restricted to the lower frequency polymorphisms. Several of the amino acid substitution variants existing at allele frequencies of >0.20 have been reported to exhibit reduced functional activity or repair capacity [23-25,73,74]. Individually, these variants observed in many individuals would have a larger impact of the population risk of disease than would a low frequency variant. This is as predicted by the "common variant-common disease" hypothesis [43-50]. Although the substitutions with higher probability to impact activity exist at somewhat lower average allele frequency, a substantial fraction of all variants are predicted to impact biochemical function.

In summary, *in silico* analyses predict that 30-50% of over 500 amino acid substitution variants currently identified in the general population will exhibit reduced activity. Analyses of the distribution of relevant variants by allele frequency suggests that both the highly polymorphic and the less common variants are likely to impact protein activity and contribute to variation in disease susceptibility. Advancement in understanding the molecular and biological basis for the observed association of reduced DNA repair capacity phenotypes with elevated individual cancer risk must consider the impact of the substantial number of potentially relevant variants in a repair pathway [12].

Methods

The dataset

The majority of the variants included in this analysis were identified during the screening of 88-90 samples from the "DNA Polymorphism Discovery Resource" available at the Coriell Institute for Medical Research, Camden, NJ. This resource was established by the NIH as a set of samples available to investigators screening for common genetic variants existing in the general population of the United States [76]. The samples are from U. S. residents and the major ethnic groupings of the population. The approaches and protocols for identifying variants have been described [78,79]. Most of the data and information describing the genes and variants are available at <http://greengenes.llnl.gov/dpublic/secure/reseq/>, Lawrence Livermore National Laboratory; <http://www.genome.washington.edu/projects/egpsnps/>, Environmental Genome Project, University of Washington; and <http://egp.gs.washington.edu>, Human Genome Center, Environmental Genome Project, University of Washington. Both of the University of Washington projects were supported by the NIEHS Environmental Genome Project. Other sample sets employed in screening for variation in a small subset of the repair genes are described in [11]. Data

for several genes are from publications reporting results from more focused screening of DNA repair genes in at least 50 unrelated individuals [see ref. 12].

The genes screened for variation and the number of amino acid substitution variants identified and analyzed for predicted impact on activity by SIFT and PolyPhen are listed in Table 5. In addition, no amino acid substitution variants were observed in the screening of *DDB1*, *FEN1*, *G22P1*, *PCNA*, *POLA*, *POLE2*, *POLM* , *TREX1* AND *UNG*. The gene symbols are from the HUGO Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/). Information regarding the role of the individual genes in DNA repair and localization to repair pathways are available [12,71].

Prediction of the impact of amino acid substitutions

The fundamental assumptions of the SIFT algorithm have been described by Ng and Henikoff [42-44]. The algorithms and instructions for analysis of amino acid substitutions are available at http://blocks.fhcrc.org/~pauline/SIFT.html. The data employed in the SIFT analysis were the repair gene sequences available in the NCBI nonredundant database (http://www.ncbi.nlm.nih.gov) on 7/15/03. Given the high level of evolutionary conservation generally observed for the DNA repair genes, only sequences more than 95% identical were excluded, rather than excluding sequences that were greater than the 90% identity recommended [27].

The PolyPhen algorithm and the underlying principles have been previously described [32,34]. Additional details of the algorithm and instructions for analysis of amino acid substitutions are available at http://www.bork.embl-heidelberg.de/PolyPhen/. The PolyPhen analyses utilized the data available in SwissPro on 7/18/2003. As both GenBank and SwissPro are dynamic databases, the results for the predicted impact of an amino acid substitution may change as new data become available. As seen in Table 1, insufficient data were available to predict the impact of a few amino acid substitutions in some genes, the exception being *ERCC6* where insufficient data were available to support the PolyPhen analysis for any of the variants.

Acknowledgement:
Work performed under auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory; contract No.W-7405-ENG-48 and supported by Interagency Agreement Y1-ES-8054-05 from NIEHS and NCI grant 1 U-1 CA 83180-03. We thank David Wilson 3rd (National Institute of Aging) for many discussions, especially related to the APE1 variants, and Chad Garner (University of California, Irvine) for assistance with statistical analyses. We also acknowledge the contributions of Debbie Nickerson (University of Washington) and Maynard Olson (University of Washington), both supported by the National Institute of Environmental Health Science Environmental Genome Project, for placing their sequence variation data in the public domain.

References

[1] R. Sachidanandam, et al., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, Nature 409 (2001) 928-933.

[2] F. Cambien, et al., Sequence diversity in 36 candidate genes for cardiovascular disorders, Am. J. Hum. Genet. 65 (1999) 183-191.

[3] M. Halushka, et al., Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis, Nature Genet. 22 (1999) 239-247.

[4] Y. Ohnishi, et al., Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population, Hum. Genet. 106 (2000) 288-292.

[5] R. Yamada, et al., Identification of 142 single nucleotide polymorphisms in 41 candidate genes for rheumatoid arthritis in the Japanese population, Hum. Genet. 106 (2000) 293-297.

[6] M. Cargill, et al., Characterization of single-nucleotide polymorphisms in coding regions of human genes, Nature Genet. 22 (1999) 231-238.

[7] A. Iida, et al., Catalog of 434 single-nucleotide polymorphisms (SNPs) in genes of the alcohol dehydrogenase, glutathione S-transferase, and nicotinamide adenine dinucleotide, reduced (NADH) ubiquinone oxidoreductase families, J. Hum. Genet. 46 (2001) 385-407.

[8] K.M. Small, et al., Gene and protein domain-specific patterns of genetic variability within the g-protein coupled receptor superfamily, Am. J. Pharmacogenomics 3 (2003) 65-71.

[9] M.K. Leabman, et al., Natural variation in human membrane transporter genes reveals evolutionary and functional constraints, Proc. Natl. Acad. Sci. USA. 100 (2003) 5896-5901.

[10] A. Iida, et al., Catalog of 605 single-nucleotide polymorphisms (SNPs) among 13 genes encoding human ATP-binding cassette transporters: ABCA4, ABCA7, ABCA8, ABCD1, ABCD3, ABCD4, ABCE1, ABCF1, ABCG1, ABCG2, ABCG4, ABCG5, and ABCG8, J. Hum. Genet. 47 (2002) 285-310.

[11] H.W. Mohrenweiser, T. Xi, J. Vazquez-Matias, I.M. Jones, Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1054-1064.

[12] H.W. Mohrenweiser, D.M. Wilson 3rd, I.M. Jones, Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes, Mutat. Res. 526 (2003) 93-125.

[13] P.L. Welcsh, M.C. King, BRCA1 and BRCA2 and the genetics of breast and ovarian cancer, Hum. Mol. Genet. 10 (2001) 705-713.

[14] A. Muller, R. Fishel, Mismatch repair and the hereditary non-polyposis colorectal cancer syndrome (HNPCC), Cancer Invest. 20 (2002) 102-109.

[15] L. Grossman, et al., DNA repair as a susceptibility factor in chronic diseases in human populations. In M. Dizdaroglu, A.E. Karakaya (Eds.), Advances in DNA Damage and Repair, Kluwer Academic/Plenum Publishers New York, 1999, pp 149-167.

[16] M.R. Spitz, Q. Wei, Q. Dong, C.I. Amos, X. Wu, Genetic susceptibility to lung cancer: the role of DNA damage and repair, Cancer Epidemiol. Biomarkers Prev. 12 (2003) 689-698.
[17] X. Wu, et al., A parallel study of in vitro sensitivity to benzo[a]pyrene diol epoxide and bleomycin in lung cancer cases and controls, Cancer 83 (1998) 1118-1127.
[18] M. Berwick, P. Vineis, Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review, J. Natl. Cancer Inst. 92 (2000) 874-897.
[19] M. Berwick, G. Matullo, P. Vineis, Studies of DNA repair and human cancer: an update, in: S.H. Wilson, W.A. Suk, (Eds.), Biomarkers of Environmentally Associated Disease: Technologies, Concepts and Perspectives, Lewis Publishers, Boca Raton, 2002, pp. 84–105.
[20] J. Cloos, et al., Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripheral blood lymphocytes, J. Natl. Cancer Inst. 91 (1999) 1125-1130.
[21] S.A. Roberts, et al., Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer, Am. J. Hum. Genet. 65 (1999) 784-794.
[22] M.Z. Hadi, M.A. Coleman, K. Fidelis, H.W. Mohrenweiser, D.M. Wilson, III, Functional characterization of Ape1 variants identified in the human population, Nucleic Acids Res. 28 (2000) 3871–3879.
[23] Y. Qiao, et al. Modulation of repair of ultraviolet damage in the host-cell reactivation assay by polymorphic XPC and XPD/ERCC2 genotypes, Carcinogenesis 23 (2002) 295-299.
[24] Y. Qiao, et al., Rapid assessment of repair of ultraviolet DNA damage with a modified host-cell reactivation assay using a luciferase reporter gene and correlation with polymorphisms of DNA repair genes in normal human lymphocytes, Mutat. Res. 509 (2002) 165-174.
[25] Y. Wang, et al., From genotype to phenotype: correlating XRCC1 polymorphisms with mutagen sensitivity, DNA Repair 2 (2003) 901-908.
[26] E.L. Goode, C.M. Ulrich, J.D. Potter, Polymorphisms in DNA repair genes and associations with cancer risk, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1513-1530.
[27] P.C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions, Genome Res. 11 (2001) 863-874.
[28] P.C. Ng, S. Henikoff, Accounting for human polymorphisms predicted to affect protein function, Genome Res. 12 (2002) 436-446.
[29] P.C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function, Nucleic Acids Res. 3 (2003) 3812-3814.
[30] Z. Wang, J. Moult. SNPs, protein structure, and disease, Hum. Mutat. 17 (2001) 263-270.
[31] D. Chasman, R.M. Adams, Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation, J. Mol. Biol. 307 (2001) 683-706.
[32] S. Sunyaev, et al., Prediction of deleterious human alleles, Hum. Mol. Genet. 10 (2001) 591-597.
[33] S. Sunyaev,V. Ramensky, P. Bork , Towards a structural basis of human non-synonymous single nucleotide polymorphisms, Trends Genet. 16 (2000) 198-200.

[34] V. Ramensky, P. Bork, S. Sunyaev, Human non-synonymous SNPs: server and survey, Nucleic Acids Res. 30 (2002) 3894-3900.
[35] S.R. Sunyaev, W.C. Lathe, 3rd, V.E. Ramensky, P. Bork, SNP frequencies in human genes, an excess of rare alleles and differing modes of selection, Trends Genet. 16 (2000) 335-337.
[36] M.A. Fleming, J.D. Potter, C.J. Ramirez, G.K. Ostrander, E.A. Ostrander, Understanding missense mutations in the BRCA1 gene: an evolutionary approach, Proc. Natl. Acad. Sci. USA. 100 (2003) 1151-1156.
[37] D.M. Wilson3rd, T.M. Sofinowski, D.R. McNeill, Repair mechanisms for oxidative DNA damage, Front. Biosci. 8 (2003) d963-981.
[38] J.P. Erzberger, D.M. Wilson III, The role of Mg^{2+} and specific amino acid residues in the catalytic reaction of the major human abasic endonuclease: new insights from EDTA-resistant incision of acyclic abasic site analogs and site-directed mutagenesis, J. Mol. Biol. 290 (1999) 447-457.
[39] J.P. Erzberger, D. Barsky, O.D. Scharer, M.E. Colvin, D.M. Wilson III, Elements in abasic site recognition by the major human and Escherichia coli apurinic/apyrimidinic endonucleases, Nucleic Acids Res. 26 (1998) 2771-2778.
[40] L.H. Nguyen, D. Barsky, J.P. Erzberger, D.M. Wilson III, Mapping the protein-DNA interface and metal binding site of the major human apurinic/apyrimidinic endonuclease, J. Mol. Biol. 298 (2000) 447-459.
[41] M.Z. Hadi, K. Ginalski, L.H. Nguyen, D.M. Wilson III, Determinants in nuclease specificity of Ape1 and Ape2, human homologues of Escherchia coli exonuclease III, J. Mol. Biol. 316 (2000) 853-866.
[42] A. Wright, B. Charlesworth, I. Rudan, A. Carothers, H. Campbell, A polygenic basis for late-onset disease, Tends Genet. 19 (2003) 97-106.
[43] J.L. Badano, N. Katsanis, Beyond Mendel: an evolving view of human genetic disease transmission, Nat. Rev. Genet. 3 (2002) 779-789.
[44] M.E. Zwick, D.J. Cutler, A. Chakravarti, Patterns of genetic variation in Mendelian and complex traits, Annu. Rev. Genomics Hum. Genet. 1 (2000) 387-407.
[45] N. Risch, K. Merikangas, The future of genetic studies of complex human diseases, Science 273 (1996) 1516-1517.
[46] E.S. Lander, The new genomics: global views of biology, Science 274 (1996) 536-539.
[47] N.J. Risch, Searching for genetic determinants in the new millennium, Nature 405 (2000) 847-856.
[48] D.E. Reich, E.S. Lander, On the allelic spectrum of human disease, Trends Genet. 17 (2001) 502-510.
[49] M. Cargill, G.Q. Daley, Mining for SNPs: putting the common variants--common disease hypothesis to the test, Pharmacogenomics 1 (2000) 27-37.
[50] N.H. Barton, P.D. Keightley, Understanding quantitative genetic variation, Nat. Rev. Genet. 3 (2002) 11-21.

[51] G.K. Wong, et al., A population threshold for functional polymorphisms, Genome Res. 13 (2003) 1873-1879.
[52] J.D. Wright, C. Lim, A fast method for predicting amino acid mutations that lead to unfolding, Protein Eng. 14 (2001) 479-486.
[53] C. Machicado, M. Bueno, J. Sancho, Predicting the structure of protein cavities created by mutation, Protein Eng. 15 (2002) 669-675.
[54] C.T. Saunders, D. Baker, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J. Mol. Biol. 322 (2002) 891-901.
[55] J. Majewski, J. Ott, Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms, Gene 305 (2003) 167-173.
[56] C. Ferrer-Costa, M. Orozco, X. de la Cruz, Characterization of disease associated single amino acid polymorphisms in terms of sequence and structure properties, J. Mol. Biol. 31 (2002) 771-786.
[57] N.O. Stitziel, et al., Structural location of disease-associated single-nucleotide polymorphisms, J. Mol. Biol. 327 (2003) 1021-1030.
[58] M.F. Santibanez Koref, et al., A phylogenetic approach to assessing the significance of missense mutations in disease genes, Hum. Mutat. 22 (2003) 51-58.
[59] M.P. Miller, S. Kumar, Understanding human disease mutations through the use of interspecific genetic variation, Hum. Mol. Genet. 10 (2001) 2319-2328.
[60] B.A. Sokhansanj, G.R. Rodrigue, J.P. Fitch, D.M. Wilson 3rd, A quantitative model of human DNA base excision repair. I. Mechanistic insights, Nucleic Acids Res. 30 (2002) 1817-1825.
[61] S. Savas, J.A. Knight, L. Briollais, H. Ozcelik, Prediction of deleterious single nucleotide polymorphisms (SNPs) in DNA repair genes for breast cancer association studies, Cancer Epidemiol. Biomarkers Prev. 11 (2002) 1178s (abs).
[62] S. Henikoff, J.G. Henikoff, Performance evaluation of amino acid substitution matrices, Proteins 17 (1993) 49-61.
[63] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862-864.
[64] E.G. Chottiner, T.E. Gribbin, D.Ginsburg, B.S. Mitchell, Erythrocyte-specific overproduction of adenosine deaminase: molecular genetic studies, Prog. Clin. Biol. Res. 319 (1989) 55-64.
[65] Y. Shu, et al., Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1, Proc. Natl. Acad. Sci. USA. 100 (2003) 5902-2907.
[66] M.K. Leabman, et al., Polymorphisms in a human kidney xenobiotic transporter, OCT2, exhibit altered function, Pharmacogenetics 12 (2002) 395-405.
[67] R.G. Tirona, B.F. Leake, G. Merino, R.B. Kim,Polymorphisms in OATP -C: identification of multiple allelic variants associated with altered transport activity among European- and African-Americans, J. Biol. Chem. 276 (2001) 35669-35675.
[68] S.J. Lee, et al., Genetic findings and functional studies of human CYP3A5 single nucleotide polymorphisms in different ethnic groups, Pharmacogenetics 13 (2003) 461-472.

[69] B.D. Przybyla-Zawislak, et al., Polymorphisms in human soluble epoxide hydrolase, Mol. Pharmacol. 64 (2003) 482-490.
[70] J. Blaisdell, et al., Identification and functional characterization of new potentially defective alleles of human CYP2C19, Pharmacogenetics 12 (2002) 703-11.
[71] L.M. King, et al., Cloning of CYP2J2 gene and identification of functional polymorphisms, Mol. Pharmacol. 61 (2002) 840-852.
[72] D. Dai, et al., Identification of variants of CYP3A4 and characterization of their abilities to metabolize testosterone and chlorpyrifos, J. Pharmacol. Exp. Ther. 299 (2001) 825-831.
[73] H. Seker H, et al., Functional significance of XPD polymorphic variants: attenuated apoptosis in human lymphoblastoid cells with the XPD 312 Asp/Asp genotype, Cancer Res. 61 (2001) 7430-7434.
[74] R. Pastorelli, A. Cerri, M. Mezzetti, E. Consonni, L. Airoldi, Effect of DNA repair gene polymorphisms on BPDE-DNA adducts in human lymphocytes, Int. J. Cancer 100 (2002) 9-13.
[75] C. Bernstein, H. Bernstein, C.M. Payne, H. Garewal, DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis, Mutat. Res. 511 (2002) 145–178.
[76] F.S. Collins, L.D. Brooks, A. Chakravarti, A DNA polymorphism discovery resource for research on human genetic variation, Genome Res. 8: (1998) 1229-1231.
[77] D.A. Nickerson, et al., DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene, Nature Genet. 19 (1998) 233-240.
[78] M.R. Shen, I.M. Jones, H. Mohrenweiser,Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans, Cancer Res. 58 (1998) 604-608.

Table 1
Distribution of variants by PolyPhen scores

PolyPhen Scores	**Impact**	**# of variants**	**%**	**average allele freq**
>2.00	Probably Damaging	61	12.5	0.033
1.99-1.75	Possibly Damaging	48	9.8	0.027
1.74-1.50	Possibly Damaging	56	11.5	0.031
1.49-1.25	Potentially Damaging	76	15.5	0.033
1.24-1.00	Borderline	62	12.7	0.042
0.99-0.50	Benign	91	18.6	0.036
0.49-0.01	Benign	88	18	0.064
0.00	Benign	7	1.4	0.090
Total		489	100	

Table 2
Distribution of variants by SIFT scores

SIFT Scores	Impact	# of variants	%	average allele freq
0.00	Intolerant	140	27.6	0.037
0.01-0.05	Intolerant	86	16.9	0.036
0.051-0.10	Potentially Intolerant	45	8.9	0.044
0.101-0.20	Borderline	64	12.6	0.036
0.201-0.50	Tolerant	84	16.5	0.041
0.501-0.99	Tolerant	51	10	0.070
1.00	Tolerant	38	7.5	0.034
Total		508	100	

Table 3
Concordance of the SIFT and PolyPhen predictions of the impact of amino acid substitutions on activity of DNA repair proteins

	PolyPhen predicted impact (scores)				
SIFT predicted impact (scores)	Probable (>2.00)	Possible (1.99-1.50)	Potential (1.49-1.25)	Borderline (1.24-1.00)	Benign (<1.00)
Intolerant (0)	**26 (5.4)**[a,b]	**34 (7.1)**	*18 (3.8)*[c]	14 (2.9)	28 (5.8)
Intolerant (0.01-0.05)	**16 (3.3)**	**25 (5.2)**	*13 (2.7)*	12 (2.5)	14 (2.9)
Potential (0.051-0.10)	*4 (0.8)*	*7 (1.5)*	*7 (1.5)*	9 (1.8)	18 (3.8)
Borderline (0.101-0.20)	5 (1.0)	17 (3.4)	13 (2.7)	7 (1.5)	22 (4.6)
Tolerant (>0.20)	10 (2.1)	17 (3.4)	23 (4.8)	19 (4.0)	100 (20.9)

[a]number (%) of variants
[b]bold indicates variants where both SIFT and PolyPhen predict negative impact
[c]italics indicate variants where both SIFT and PolyPhen predict at least a potential negative impact

Table 4
SIFT and PolyPhen scores for functionally characterized APE1 variants

Variants	binding (%)	incision (%)	Sift score[b]	PolyPhen score	PolyPhen Impact[c]
Wild Type	100.0	100.0			
N68A[d]	100.0	0.02	0.00	3.05	PRB
N68D	100.0	0.05	0.00	2.37	PRB
D70A	100.0	4.0	0.41	0.44	Ben
D70R	100.0	4.0	0.05	1.37	PRB
E96A	100.0	0.02	0.00	2.59	PRB
E96Q	100.0	0.02	0.00	1.92	PRB
Y128A	<1.00	25.0	0.02	3.17	PRB
R156Q	<1.00	1.0	0.00	2.34	PRB
Y171F	100.0	0.02	0.00	2.18	PRB
Y171H	100.0	0.01	0.00	2.40	PRB
Y171Q	7.0	nd[a]	0.00	3.08	PRB
D210A	100.0	<0.004	0.00	2.91	PRB
D210N	100.0	<0.004	0.00	2.23	PRB
D210H	100.0	0.06	0.00	2.68	PRB
F266A	15.0	15.0	0.01	2.23	PRB
D283N	nd	10.0	0.01	2.23	PRB
D308A	50.0	20.0	0.00	2.91	PRB
D308S	100.0	20.0	0.00	2.46	PRB
H309S	10.0	<0.004	0.00	3.51	PRB
L104R	50.0	36.0	0.02	1.53	POS
E126D	75.0	60.0	0.30	0.86	Ben
D148E	100.0	94.0	1.00	0.19	Ben
R1237A	100.0	36.0	0.05	1.93	POS
G241R	100.0	106.0	0.09	1.13	Ben
D283G		~10	0.01	2.68	PRB
G306A	100.0	107.0	0.20	1.18	Ben

[a]As with the other variants with low binding efficiency and other substitutions at this residue, this variant is assumed to exhibit low incision activity, although the incision activity was not determined.
[b]SIFT scores of 0.05 or less are classified as “Intolerant” while scores above 0.05 are designated as “Tolerated” variants
[c] PRB, “Probably damaging”; POS, “Possibly damaging; Ben, “Benign” variants
[d] most common amino acid residue-residue number-variant amino acid

Table 5
DNA repair genes screened for variation, number of missense variants identified and number of variants scored by SIFT and PolyPhen algorithms

Gene	# of variants identified	# of variants scored by SIFT	# of variants scored by PolyPhen	# of variants scored by both algorithms
ADPRT1	5	5	5	5
ADPRTL2	5	5	5	5
APEX1	4	4	4	4
ATM	45	38	41	38
BRCA1	10	10	10	10
BRCA2	35	31	34	30
CHEK1	1	1	1	1
CKN1	1	1	1	1
DDB2	3	3	3	3
ERCC1	1	1	1	1
ERCC2	5	5	5	5
ERCC3	4	4	4	4
ERCC4	9	9	9	9
ERCC5	15	15	15	15
ERCC6	16	16	0	0
EXO1	18	18	18	18
FANCC	3	3	3	3
FANCG	4	4	4	4
GTF2H1	3	3	3	3
GTF2H3	1	1	1	1
GTF2H4	1	1	1	1
HCNP	3	3	3	3
HUS1	4	4	4	4
LIG1	11	11	11	11
LIG3	3	3	3	3
LIG4	2	2	2	2
MBD4	6	6	6	6
MGMT	7	6	6	5
MLH1	9	9	9	9
MPG	7	7	7	7
MRE11A	2	2	2	2
MSH2	6	6	6	6
MSH3	10	10	9	9
MSH4	5	5	5	5
MSH6	8	7	8	7
MUTY	8	8	8	8

NBS1	6	6	6	6
NEIL1	4	4	4	4
NEIL2	5	4	4	3
NTHL1	6	6	6	6
OGG1	4	4	4	4
PMS1	6	6	6	6
PMS2	6	6	6	6
PNKP	5	5	5	5
POLB	3	3	3	3
POLD1	7	7	6	6
POLD2	1	1	1	1
POLE	19	19	19	19
POLG	8	8	8	8
POLH	1	1	1	1
POLI	7	7	7	7
POLK	2	2	2	2
POLL	2	2	2	2
POLQ	11	11	7	7
PRKDC	19	19	19	19
RAD1	5	4	4	4
RAD9	3	3	3	3
RAD23A	2	2	2	2
RAD23B	1	1	1	1
RAD50	3	3	3	3
RAD51	1	1	1	1
RAD52	3	3	3	3
RAD54L	5	5	5	5
RECQL1	9	9	7	7
REV1L	11	11	11	11
REV3L	17	17	17	17
RFC1	2	2	2	2
RFC2	1	1	1	1
RFC3	1	1	1	1
RFC4	1	1	1	1
RFC5	1	1	1	1
RPA4	1	1	1	1
SMUG1	3	3	3	3
TDG	2	2	2	2
WRN	4	4	4	4
XPA	4	4	4	4
XPC	13	13	11	11
XRCC1	16	16	16	16
XRCC2	3	3	3	3

XRCC3	2	2	2	2
XRCC4	5	5	5	5
XRCC5	2	2	2	2
TOTAL	523	508	489	479