

Identifying Functional Genetic Variants in DNA Repair Pathway Using Protein Conservation Analysis

Sevtap Savas,^{1,3} David Y. Kim,¹ M. Farhan Ahmad,¹ Mehjabeen Shariff,¹ and Hilmi Ozcelik^{1,2,3}

¹Fred A. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, ²Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Ontario, Canada and ³Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada

Abstract

The role of DNA repair in initiation, promotion, and progression of malignancy suggests that variations in DNA repair genes confer altered cancer risk. Accordingly, DNA repair gene variants have been studied extensively in the context of cancer predisposition. Single nucleotide polymorphisms (SNPs) are the most common genetic variations in the human genome. A fraction of SNPs are located within the genes, which are likely to alter the gene expression and function. SNPs that change the encoded amino acid sequence of the proteins (non-synonymous; nsSNPs) are potentially genetic disease determinant variations. However, as not all amino acid substitutions are supposed to lead to a change in protein function, it will be necessary to have *a priori* prediction and determination of the functional consequences of amino acid substitutions *per se*, and then together with

other genetic and environmental factors to study their possible association with a trait. Here we report the analysis of nsSNPs in 88 DNA repair genes and their functional evaluation based on the conservation of amino acids among the protein family members. Our analysis demonstrated that >30% of variants of DNA repair proteins are highly likely to affect the function of the proteins drastically. In this study, we have shown that three nsSNPs, which were predicted to have functional consequences (XRCC1-R399Q, XRCC3-T241M, XRCC1-R280H), were already found to be associated with cancer risk. The strategy developed and applied in this study has the potential to identify functional protein variants of DNA repair pathway that may be associated with cancer predisposition. (Cancer Epidemiol Biomarkers Prev 2004;13(5):801–7)

Introduction

Nuclear DNA is under constant DNA damage stress induced by both endogenous (such as reactive oxygen species) and exogenous sources (such as irradiation). Proper recognition and repair of the DNA damage are essential for normal homeostasis and functioning of multicellular organisms (1, 2). DNA repair activities are maintained by the presence of five different DNA damage sensor and repair mechanisms (homologous recombinational repair, non-homologous end-joining, nucleotide excision repair, base excision repair, and mismatch repair). Defects in the DNA repair pathways are often associated with excessive cell death (by apoptosis) or transformation of the cells (1, 2), and variations in DNA repair genes were hypothesized to modify individual and population cancer risk (3).

To date, much success has been obtained in the identification of high-penetrant cancer predisposition genes using linkage analysis. However, the challenge that has remained is to identify those alleles conferring low to moderate cancer risk. It is hypothesized that

genetic variation contributes to the susceptibility for complex traits such as cancer (4–6). Molecular epidemiological and genetic approaches use single nucleotide polymorphisms (SNPs) in the human genome to study disease susceptibility. Because genome-wide scans are still challenging, often candidate gene/pathway approach may prove more efficient. Due to presence of enormous number of SNPs, systematic prioritization on the basis of biological function and relevance to cancer will accelerate the identification of such susceptibility alleles (4).

The most common form of genetic variation in the human genome is the SNPs (5–8). SNPs are relatively stably inherited genomic variations with an estimated density of 1 in 1000 bp. SNPs are usually bi-allelic, their occurrence rates vary across the genomic regions, and their allelic frequencies may differ among ethnic groups. A fraction of SNPs alter the encoded amino acid sequence (non-synonymous SNPs; nsSNPs), and have the potential to affect the structure, function, and interactions of proteins. Thus, nsSNPs are excellent candidates for candidate-gene association studies (7). However, not all nsSNPs are anticipated to have functional consequences; it is essential to develop strategies to select the variations that may alter and disrupt the proper functions of the proteins. Studying the functional consequences of genetic variants has been challenging due to the enormous number of variants present in the genome. Although there is an increasing effort for establishing *in vivo* functional strategies for studying the effects of variants, it is still far from being

Received 9/24/03; revised 12/4/03; accepted 12/24/03.

Grant support: Grant (BCTR0100627) from Susan Komen Breast Cancer Foundation, USA and "CIHR Strategic Training Program Grant—The Samuel Lunenfeld Research Institute Training Program: Applying Genomics to Human Health" fellowship (S. Savas).

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.

Requests for reprints: Hilmi Ozcelik, Mount Sinai Hospital Samuel Lunenfeld Research Institute, 600 University Avenue Room 992A, Toronto, ON M5G 1X5, Canada. Phone: (416) 586-4996; Fax: (416) 586-8869. E-mail: ozcelik@mshri.on.ca

available for a large number of variants of interest. Recently, several approaches have been developed and used to study the nature of the genetic variants (9–15). Among these, computational tools provide an efficient and high-throughput source for *in vivo* functional analyses and/or population studies. SIFT (Sorting Intolerant From Tolerant) (10, 11) is a powerful tool that predicts the functional importance of an amino acid based on the alignment of highly similar proteins (either orthologous or paralogous or both) with the protein of interest. The predictions rely on whether or not an amino acid is conserved (or substituted by only a similar amino acid) in the protein family, which can suggest its importance for the function/structure of the protein.

Here, using the public SNP databases, we have identified a wide range of DNA repair nsSNPs, and we have carried out a computational study to characterize the evolutionary importance of these DNA repair nsSNPs. This study has the potential to provide a pool of functional SNPs, which may play important roles in the predisposition to cancer as well as other DNA repair-associated genetic diseases.

Methods

Database Mining for SNPs. The list of DNA repair pathway genes studied was obtained from the CGAP-GAI web-site⁴ (16). A total of 88 DNA repair genes was screened for SNPs using five different public SNP databases: dbSNP⁵ (17), HGvbase⁶ (18), CGAP-GAI⁷ (16), SNP500⁸, and GeneSNP⁹. During this work, we noticed several problems concerning the specificity and accuracy of the information related to the SNPs (*i.e.*, errors on annotations of SNP locations along the protein sequence, of the nature of the SNP as a sSNP, nsSNP, etc., and the specificity of the SNP, etc.). Accordingly, we developed a strategy to standardize the uniformity of the SNP collection procedure and to validate the specificity of the SNPs by performing two independent BLAST analyses. The sequences flanking the variants (preferably 100 bp) were retrieved from SNP databases. These sequences were then aligned against the mRNA sequences in GenBank using the “BLAST against gene transcripts” tool¹⁰ to confirm the location of the nsSNPs. Once a gene-specific accession number was obtained, it was used as a reference template to locate the other variants in the same gene. Also, we have aligned the same SNP-flanking sequence against the non-redundant human genome database of NCBI using the BLAST tool¹¹ (19), and the genome view option was used to visualize the exact chromosomal location of the hit. Unigene¹² resource of

NCBI (19) was used to validate the chromosomal location of the gene of interest, and compared with that of the SNP-flanking sequence location obtained from genome view result. In those cases when a shorter than 100 bp SNP-flanking sequence was available, we checked the genome location using the “search for short nearly exact matches” option of the NCBI nucleotide BLAST choosing “Homo sapiens” as advanced blast option. The retrieved hits were then manually inspected to find the chromosomal location of the SNP sequence. We specified a SNP as gene specific if and only if the SNP-flanking sequence hits (without any other mismatch or gap) (*a*) the transcript(s) of interest but not any other gene’s or expressed pseudogene’s as a result of blast against gene transcripts, or (*b*) any other genomic sequence other than the location of the gene of interest. In case of alternatively spliced genes, only the information (such as nsSNP location, evolutionary conservation, etc.) of the amino acid sequence encoded by one alternatively spliced transcript was reported in this study. The SNPs annotated as splice site SNPs in HGvbase were excluded from this study. Whenever available, the frequency information of the nsSNPs was extracted from the SNP databases as well as from Mohrenweiser *et al.* (20). Entire list of genes and nsSNPs analyzed during this study can be found at <http://www.ozceliklab.com/savas2004a/>.

Mutation Data Set. Mutations with known functional consequences were retrieved from the SWISS-PROT database¹³ (21) using the key words “Human AND mutations AND functional” on September 2002. Following a manual inspection, a total of 231 mutations in 55 human genes constituted the final list. The mutations in this list were characterized by complete/partial loss of activity, gain of function, affecting protein-macromolecule interactions, interfering with cellular localization of the mutant protein, altering the protein stability, altering a protein-critical site, or interfering with the protein dimerization, as indicated in the feature table of each SWISS-PROT entry.

Evolutionary Conservation Analysis. Protein conservation analysis was performed using the SIFT¹⁴ software developed by Ng and Henikoff (10, 11). The SIFT algorithm predicts whether an amino acid substitution may have an impact on protein function by aligning similar proteins, and calculating a score which is used to determine the evolutionary conservation status of the amino acid of interest. It evaluates the identity (such as if only a single amino acid is observed in all proteins aligned at that position, then the alteration of it is predicted to affect the protein) and two physicochemical characteristics of amino acids, hydrophobicity and polarity (if the substituted amino acid differs in these characteristics from the wild-type amino acid and this kind of a substitution is not observed in the other proteins in the alignment at that position, it is predicted to affect the protein as well). These predictions are based on the assumption that amino acids conserved within the protein families are important for the function of the proteins. Whenever the frequency information was

⁴ Internet address: http://lpgws.nci.nih.gov/html-cgap/cgl/DNA_damage.html.

⁵ Internet address: <http://www.ncbi.nlm.nih.gov/SNP/>.

⁶ Internet address: <http://hgvdbase.cgb.ki.se/>.

⁷ Internet address: <http://lpgws.nci.nih.gov/>.

⁸ Internet address: <http://snp500cancer.nci.nih.gov/home.cfm>.

⁹ Internet address: <http://www.genome.utah.edu/genesnps/>.

¹⁰ M. Edmenson, K. Buetow. The BLAST against gene transcripts tool (unpublished). Internet address: <http://lpgws.nci.nih.gov:80/perl/blast2>.

¹¹ Internet address: <http://www.ncbi.nlm.nih.gov/BLAST/>.

¹² Internet address: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>.

¹³ Internet address: <http://us.expasy.org/sprot/>.

¹⁴ Internet address: http://blocks.fhrc.org/sift/SIFT_seq_submit2.html.

Table 1. Comparison of SIFT evolutionary conservation status of mutations versus DNA repair nsSNPs

SIFT predictions	Mutations (<i>n</i> = 230) <i>n</i> (%)	nsSNPs* (<i>n</i> = 106) <i>n</i> (%)	Validated nsSNPs [†] (<i>n</i> = 68) <i>n</i> (%)
Damaging	132 (57.39)	11 (10.38)	5 (7.36)
Possibly damaging	44 (19.13)	28 (26.41)	15 (22.06)
Possibly tolerated	13 (5.65)	39 (36.80)	30 (44.11)
Tolerated	41 (17.83)	28 (26.41)	18 (26.47)

Note: This table contains the variations (mutations and nsSNPs) for which a reliable SIFT prediction was available (≥ 6 similar proteins in the alignment).

*Includes all the nsSNPs independent of their validation status.

[†]Includes the validated SNP only.

available, this conservation analysis was performed for the common allele. As we thought they would not be reliable, in this analysis, we did not consider the SIFT predictions based on less than six proteins in the alignments. We used the default median sequence conservation in the range of 3.0. In no cases the median sequence conservation score was found ≤ 2.25 . However, there were many amino acid substitutions where the score was calculated as >3.25 . Such scores indicate that the substitution at that position might not have had time to evolve yet, and consequently, the prediction may be misleading (11). Thus, we designated the predictions with a median sequence conservation score of >3.25 as either possibly affecting or possibly tolerated. This evaluation is different from that of Ng and Henikoff (11), where such predictions were not accepted at all.

Statistical Analysis. The statistical analyses were done using a χ^2 test (22). We applied the Yates correction for approximation of 2×2 tables. The test was conducted at the $\alpha = 0.05$ level of significance. This test was applied to examine possible significant differences of the evolutionary conservation status of the amino acids altered in mutation and DNA repair nsSNP data sets, and between the rare and common DNA repair nsSNPs.

Results

We have compiled a total of over 1000 SNP entries from 88 DNA repair genes using five web-based public SNP databases (see "Methods"). Extensive manual inspection of all SNP entries have shown at least one gene-specific nsSNP in 51.1% (45 of 88) of the proteins (a total of 150 nsSNPs resulting in an amino acid substitution). Four of the nsSNPs were unique to the CGAP-GAI database. There was no nsSNP unique to the SNP500 database. Most of the nsSNPs were found in dbSNP (*n* = 128, 85.3%), GeneSNP (*n* = 105, 70.0%), and HGVbase (*n* = 89, 59.3%). The average number of nsSNPs for genes with at least one nsSNP was 3.3. Among all the genes studied, ATM was found to have the highest number of nsSNPs (*n* = 19).

In this study, we have used a modified interpretation of the SIFT algorithm results to define the nature of the variations (see "Methods"). To determine the sensitivity of the modified SIFT interpretation, we have used a panel of 231 missense mutations supported with functional evidence (see "Methods"; Table 1). Except one mutation, the number of proteins in all the alignments was at least six or higher (*n* = 230). Mutations in this group were

predicted as either damaging (57.39%) or possibly damaging (19.13%), whereas 17.83% and 5.65% of the mutations were predicted either tolerated or possibly tolerated, respectively. Thus, the sensitivity of the modified SIFT predictions (damaging together with possibly damaging) reported in this study was 76.52%.

We have also applied the modified SIFT predictions to study our panel of 150 nsSNPs involved in DNA repair genes. In 44 of 150 variants, the predictions were based on the alignment of less than six sequences, which was considered inconclusive (NP nsSNPs). Reliable predictions were obtained in 106 (70.6%) nsSNPs, and the results are depicted in Table 1. Within this group, 11 (10.37%) nsSNPs were predicted to be damaging the protein function. Twenty-eight of the 106 variants (26.41%) were predicted as possibly damaging, indicating that they are likely to have functional consequences as well. On the other hand, 67 nsSNPs (63.2%) were predicted either tolerated or possibly tolerated by our SIFT analysis. We have found that SIFT detects a significantly higher number of damaging alterations (including the possibly damaging alterations) in the mutation panel as compared to the DNA repair nsSNP panel ($P < 0.0001$) (Table 1).

Frequency information of 102 (68.0%) of 150 nsSNPs¹⁵ was available either in the SNP databases or in Mohrenweiser *et al.* (20) (herein called validated/proven SNPs). For 68 of the validated nsSNPs, there were reliable SIFT predictions (Table 1). We classified the nsSNPs as rare or common if the frequencies of the minor allele fell between $\leq 5\%$ and $>5\%$ ranges, respectively. There were a total of seven nsSNPs (6.86%) that were reported in independent submissions as both common and rare according to our classification. In the remaining cases, we have categorized 76.47% (78 of 102) and 16.66% (17 of 102) as rarely and commonly occurring nsSNPs, respectively. The comparisons of evolutionary conservation status of the amino acids and the frequency ranges of the nsSNPs substituting these amino acids are depicted in Table 2. In addition, there were >20 nsSNPs in our set with minor allele frequencies $\geq 10\%$ at least in one submission (see <http://www.ozceliklab.com/savas2004a/>).

In case of rare nsSNPs, we predicted 4 nsSNPs as damaging and 11 as possibly damaging (Table 3). Our

¹⁵ A few number of nsSNPs were screened in population(s) but could not be detected: we still report them as there was a chance that these nsSNPs could not be validated because they may represent either ethnic group specific or rare nsSNPs.

Table 2. Comparison of evolutionary conservation status of the rare and common nsSNPs

SIFT prediction	Rare nsSNPs (≤5%) (n = 78)* n (%)	Common nsSNPs (>5%) (n = 17)* n (%)	Rare nsSNPs (≤5%) (n = 49) [†] n (%)	Common nsSNPs (>5%) (n = 13) [†] n (%)
NP	29 (37.18)	4 (23.52)		
Damaging	4 (5.13)	0 (0.00)	4 (8.16)	0 (0.00)
Possibly damaging	11 (14.11)	3 (17.64)	11 (22.46)	3 (23.08)
Tolerated	12 (15.38)	5 (29.41)	12 (24.49)	5 (38.46)
Possibly tolerated	22 (28.20)	5 (29.41)	22 (44.89)	5 (38.46)

Note: n stands for number of SNPs. The percentages of the SIFT predictions within rare and common nsSNPs are given within parentheses.

*All validated nsSNPs regardless of their SIFT results.

[†]All validated nsSNPs with reliable SIFT predictions.

results have also shown that none of the 17 SNPs with allelic frequencies of 5% and higher were predicted to be damaging, whereas 3 of them (IGHMBP2-T671A; XRCC1-R399Q; XRCC3-T241M) were predicted to be possibly damaging (Table 3). The two nsSNPs, ERCC4-P379S (HGVBbase SNP ID: SNP000000067; Ref. 23), and XRCC1-R280H (SNP000000031/rs25489; see also GeneSNP entry) variants were predicted as damaging and possibly dam-

aging by SIFT analysis, respectively, though the reported minor allele frequencies were inconsistent (Table 3).

Discussion

To enrich the SNP information for each gene studied, we have used five different public SNP databases. While

Table 3. List of DNA repair nsSNPs found as damaging and possibly damaging during this study

Gene symbol	SNP ID	nsSNP	Frequency range*	SIFT prediction
ATM	rs1800059	S1691R	s	possibly damaging
ATM	rs1800060	V2079I	s	possibly damaging
ATM	rs1800061	G2287A	s	possibly damaging
ATM	rs1137889	N3003D	s	possibly damaging
ERCC1	rs3188420	P77H	s	possibly damaging
ERCC3	SNP000063371/rs1805162	G402C	1	damaging
ERCC4	SNP000064450/rs2020961	A168V	1	damaging
ERCC4	SNP000000067	P379S	1/2 [†]	damaging
ERCC4	SNP000002737	I706T	1	possibly damaging
ERCC4	SNP000002795	E875G	1	possibly damaging
ERCC5	rs1047769	M254V	1	possibly damaging
FANCA	SNP000002991/rs1800282	V6D	s	possibly damaging
FANCC	SNP000003086/rs1800364	L190F	s	possibly damaging
FANCC	SNP000003087/rs1800365	D195V	s	possibly damaging
IGHMBP2	SNP000012785/rs622082	T671A	2	possibly damaging
LIG1	GAI 876498	P884R	s	damaging
LIG3	SNP000010631/rs1802880	D592V	s	damaging
LIG4	rs2232640	E461G	s	possibly damaging
MLH1	SNP000002820/rs1800149	L729V	1	possibly damaging
MLH1	SNP000064598/rs2020873	H718Y	1	possibly damaging
NTHL1	SNP000064449/rs1805378	I176T	1	damaging
NTHL1	SNP001026567	D239Y	1	damaging
PCNA	GAI 864449	Q38H	s	damaging
PCNA	rs1050525	S39R	s	damaging
POLB	E0448_302	P242R	1	possibly damaging
RAD23A	rs2242518	Q261R	s	damaging
RAD50	rs3187395	E925K	s	possibly damaging
RAD51	rs1056742	K313Q	s	damaging
TOP2A	SNP000012935/rs1804539	S1471F	s	possibly damaging
WRN	SNP001026663/rs3087414	S1079L	1	possibly damaging
XRCC1	SNP001026358/rs2307186	R7L	1	possibly damaging
XRCC1	SNP000064197/rs25496	V72A	1	possibly damaging
XRCC1	SNP001026365/rs2307191	P161L	1	possibly damaging
XRCC1	SNP000000031/rs25489	R280H	1/2 [†]	possibly damaging
XRCC1	rs2271980	V381M	s	possibly damaging
XRCC1	SNP000000032/rs25487	R399Q	2	possibly damaging
XRCC3	SNP000000060	T241M	2	possibly damaging
XRCC3	SNP000064617/rs1805380	L463F	1	possibly damaging
XRCC3	GAI 891410	P485L	s	possibly damaging

Note: SNP, rs, and GAI prefixes stand for the SNP identifiers in HGVBbase, dbSNP, and CGAP-GAI databases, respectively.

*1: nsSNPs with minor allele frequencies of ≤5%, 2: nsSNPs with minor allele frequencies of >5%.

[†]nsSNPs, the minor allele frequencies of which were reported as either ≤5% and >5% in independent SNP submissions.

dbSNP and HGVbase contained SNP information related to almost all kinds of genes, the SNP500, CGAP-GAI, and GeneSNP databases particularly focused on candidate genes/pathways that may play role in cancer susceptibility. Majority of the nsSNPs (97.33%) were found in the dbSNP, HGVbase, and GeneSNP databases. All nsSNPs reported here were curated using a highly stringent SNP extraction procedure to eliminate false annotations of the SNPs. Although SNP mining sensitivity is reduced following such a stringent procedure, we strongly suggest evaluating the SNP information using the same or similar approaches described in this study to increase the specificity of the curated data.

Among 1000 entries in the SNP databases, we have extracted a total of 150 nsSNPs resulting in an amino acid substitution from 51.1% (45 of 88) of the DNA repair genes analyzed in this study. The number of SNPs in these genes is likely to improve as more SNPs are discovered, and the SNP databases continue to be updated. Several factors may lead to underestimation of the number of SNPs in genes of interest. For example, a considerable number of SNPs in these databases is not validated to distinguish them from sequencing errors, and thus these nsSNPs represent "suspected" or "non-proven" SNPs. In terms of suspected SNPs, which are described based on the DNA/RNA sequence alignments, there may be a bias toward the genetic variations through the 3' end of the transcripts as well as for abundant transcripts, common variations, and variations in less complex regions of the genome (24–26). Therefore, sequencing of the entire coding region of the genes of interest in significant number of DNA samples may reveal additional SNPs in the genes. Sequencing might especially help to demonstrate whether these genes found to have no nsSNPs during this study are really devoid of nsSNPs or not. This information could be useful for assessing conservation status of the genes, or the different mutation/recombination rates at genomic regions containing the genes of interest (7, 8, 26).

Protein conservation analyses based on the alignment of similar proteins (either among species or within species) can reveal those amino acids that are important for the function and probably for the structure of the protein families. Although such analyses would not indicate newly evolved critical amino acids with a particular function, or amino acid which are under positive selection under today's conditions, it may still be critical in assigning evolutionary conserved residues along the proteins. SIFT (10, 11) is an automated tool that calculates the conservation scores of each amino acid residue along the given protein sequences. Originally, the prediction sensitivity of SIFT for damaging amino acid substitutions was found to be 69% (10). Our SIFT predictions reported in this paper differ in some aspects from what Ng and Henikoff (11) did. First, in this study, we have modified the SIFT predictions by only considering predictions that are based on at least six protein sequences in the alignment at the amino acid position of interest. Second, whenever the median sequence conservation was >3.25 , Ng and Henikoff (11) did not accept any predictions (a median sequence conservation score >3.25 indicates that the proteins in the alignment did not diverge yet, and thus the predictions would not be

reliable as much as the predictions obtained from alignment of the diverged proteins where conserved residues are more easily identified) (11). However, considering the fact that 19.03% of the mutations were also found with median sequence conservation scores of >3.25 (Table 1), we preferred to include such predictions in our results, only stating that they were either "possibly tolerated" or "possibly damaging."

The sensitivity of the modified SIFT prediction system was tested on a mutation set with experimentally determined functional consequences (see "Methods"). According to our results, it can be concluded that approximately 57.39% of the mutations occurred at amino acids that are conserved within the protein family in our set (median sequence conservation score 2.75–3.25). On the other hand, 19.03% of the mutations occurred either at regions of proteins that are highly conserved, or in the proteins for which homologous proteins from only close species were available (median sequence conservation score >3.25). Further analyses may be performed to investigate the latter possibility. The mutations that were not detected by SIFT as damaging could be those that occurred at query specific functional residues or are the variations in linkage disequilibrium with yet unidentified causative mutations (10). As far as DNA repair genes are concerned, over one third of the nsSNPs turned out to be likely to have functional consequences (*i.e.*, found damaging and possibly damaging). Eleven DNA repair nsSNPs were found damaging, suggesting that they are excellent candidates for disease-predisposition studies. Another 28 nsSNPs were predicted as possibly damaging. We suggest that along with the damaging SNPs, these possibly damaging nsSNPs may also be good candidates for functional and association studies.

We were not able to make predictions for 44 DNA repair nsSNPs, due to the lack of sufficient sequence information available from homologous proteins (<6 proteins in the alignment at the position of the nsSNPs). As these analyses are based on the availability of the similar proteins in the public databases, we believe that as the number of curated proteins increases in protein databases, the predictions will become possible for these nsSNPs, and the reliability of the predictions for other nsSNPs will also improve.

Classification of the proven (validated) nsSNPs based on allele frequencies showed that only 16.2% of the nsSNPs was presented in the population(s) with an allele frequency of $>5\%$, suggesting that most of the nsSNPs presented here are actually rare nsSNPs. These nsSNPs may be rare because they are either under negative selection, or newly evolved and thus not fixed in the population yet. None of the common nsSNPs investigated in this study were found to be truly damaging, whereas three of them were predicted to be possibly damaging (Table 3). We were unable to find any published reports regarding the analysis of the IGHMBP2-T671A variant, which was found to be possibly damaging in this study. IGHMBP2 (immunoglobulin μ binding protein 2) protein is presumably involved in a variety of cellular functions such as immunoglobulin-class switching, pre-mRNA processing, and transcription, and mutations in this protein have been shown to result in a neurodegenerative disease (27). On the other hand, the XRCC1-R399Q

and XRCC3-T241M variants were intensively studied in the context of cancer association. XRCC1-R399Q SNP was shown to be associated with altered breast (28, 29) and lung (30) cancer risk. XRCC3-M241T has also been shown to confer increased risk to breast cancer¹⁶ (31), bladder cancer (32), and melanoma (33). Both of the XRCC1-R399 and XRCC3-T241 residues were conserved in mammalian orthologues, suggesting that they may be important for the function of these proteins.¹⁷ There were two nsSNPs (ERCC4-P379S and XRCC1-R280H) for which the minor allele frequencies were reported as both lower and higher than 5% cutoff. The ERCC4-P379S variation was reported in SNP databases as well as in the literature (23) as both rare and common in different sample sets. Our SIFT analysis showed that ERCC4-P379 was damaging. To our knowledge, the association of this SNP with cancer risk has not been studied yet. On the other hand, XRCC1-R280H nsSNP was predicted possibly damaging by SIFT analysis and was already found to be associated with nasopharyngeal carcinoma (34), prostate cancer (35), and lung cancer (36), and was found to have a role in mutagen sensitivity (37). There were 4 and 11 nsSNPs which were both rare and either damaging or possibly damaging, respectively (Table 3). Literature search for these nsSNPs did not reveal any association of them with cancer risk. To sum up, our strategy has the ability to select the potentially disease-related SNPs, and we propose that the other nsSNPs found as evolutionary conserved during this study are good candidates for further cancer-association studies.

Mutations that reduce the fitness of the individuals will be subject to purifying selection that eventually eliminate the mutations from the gene pool of a population, and thus never reach high frequencies (38), unless they confer a selective advantage because of a disease resistance in carriers of such mutations (39). Therefore, we analyzed the common and rare DNA nsSNPs for their conservation status. As a result, we could not detect any statistically significant difference ($P < 0.0001$, Table 3). Thus, it is tempting to speculate that some deleterious nsSNPs with moderate-high frequencies do not reduce the fitness of the individuals. In this context, the nature of such proteins with deleterious variations can be explained by either (a) the protein's function can be compensated by other proteins, (b) the protein's function is required only under certain environmental exposures/conditions, or (c) the protein is a rapidly evolving one, thus accumulating more mutations without affecting the fitness of the individual. Alternatively, these new substitutions may be either neutral or even positively selected. Analysis of a much larger data set will be helpful to fully characterize frequency-conservation status relation of genetic variations.

Genetic variation has been suggested to alter disease-susceptibility risk. SNPs being the most common variation in the human genome have been extensively studied in the context of disease predisposition. SNPs

that alter important molecular features such as the expression, function, structure, stability, and interaction of candidate proteins are excellent candidates to study a possible association/direct involvement of a SNP and a phenotypic expression. However, both the presence of an enormous number of SNPs and the search for biologically relevant SNPs in candidate gene approaches require the application of reliable and logical selection systems. Here we presented results obtained using a highly stringent SNP mining strategy and a modified version of the previously developed SIFT tool to select DNA repair nsSNPs that are conserved within the protein family. Our results suggest that more than one third of the nsSNPs in the DNA repair genes are likely to have functional consequences. These nsSNPs are excellent candidates for cancer association as well as for experimental functional studies. In addition, these genetic variations are likely to be critical in studies aiming to elucidate the disparity in cancer-treatment responses among patients as well as to improve the effectiveness of the cancer treatments (40).

Acknowledgments

We thank the groups that have developed the databases and the web-based tools used in this study. We are indebted to Michael Edmonson and Pauline Ng for their invaluable assistance with the Blast against gene transcripts and SIFT tools, respectively.

References

- Bernstein C, Bernstein H, Payne CM, Garewal H. DNA repair/proapoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis. *Mutat Res* 2002;511:145-78.
- Rouse J, Jackson SP. Interfaces between the detection signaling and repair of DNA damage. *Science* 2002;297:547-51.
- Mohrenweiser HW, Wilson DM III, Jones IM. 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* 2003;526:93-125.
- Daly AK, Day CP. Candidate gene case-control association studies: advantages and potential pitfalls. *Br J Clin Pharmacol* 2001;52:489-99.
- Miller RD, Kwok P-Y. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* 2001;10:2195-8.
- Taylor JG, Choi EH, Foster CB, Chanock SJ. Using genetic variation to study human disease. *Trends Mol Med* 2001;7:507-12.
- Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 2000;9:2403-8.
- Shastry BK. SNP alleles in human disease and evolution. *J Hum Genet* 2002;47:561-6.
- 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.
- Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863-74.
- Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 2002;12:436-46.
- Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;10:591-7.
- Wang Z, Moul J. SNPs, protein structure, and disease. *Hum Mutat* 2001;17:263-70.
- Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol* 2002;315:771-86.
- Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;30:3894-900.
- Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH. Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. *Genome Res* 2000;10:1259-65.

¹⁶J.C. Figueiredo, J.A. Knight, L. Briollais, I.L. Andrusis, H. Ozcelik. Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the breast cancer family registry, in press.

¹⁷J.C. Figueiredo, N. Diaz-Granados, J.A. Knight, S. Savas, L. Briollais, H. Ozcelik. XRCC1-R399Q and XRCC3-T241M: a systematic review of biological importance and role in cancer, in preparation.

17. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308-11.
18. 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.
19. Wheeler DL, Church DM, Lash AE, et al. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* 2002;30:13-6.
20. Mohrenweiser HW, Xi T, Vazquez-Matias J, Jones IM. Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. *Cancer Epidemiol Biomark Prev* 2002;11(10 Pt 1): 1054-64.
21. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;3:275-84.
22. Pagano M, Gauvreau K. Principles of biostatistics. 2nd ed. Pacific Grove, CA: Duxbury; 2000. p. 342-6.
23. Shen MR, Jones IM, Mohrenweiser H. Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans. *Cancer Res* 1998;58:604-8.
24. Gu Z, Hillier L, Kwok PY. Single nucleotide polymorphism hunting in cyberspace. *Hum Mutat* 1998;12:221-5.
25. Cox DG, Boillot C, Canzian F. Data mining: efficiency of using sequence databases for polymorphism discovery. *Hum Mutat* 2001; 17:141-50.
26. Lercher MJ, Hurst LD. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* 2002;18: 337-340.
27. Grohmann K, Schuelke M, Diers A, et al. Mutations in the gene encoding immunoglobulin μ -binding protein 2 cause spinal muscular atrophy with respiratory distress type 1. *Nat Genet* 2001;29:75-7.
28. Duell EJ, Millikan RC, Pittman GS, et al. Polymorphisms in the DNA repair gene *XRCC1* and breast cancer. *Cancer Epidemiol Biomark Prev* 2001;10:217-22.
29. Kim SU, Park SK, Yoo KY, et al. *XRCC1* genetic polymorphism and breast cancer risk. *Pharmacogenetics* 2002;12:335-8.
30. Divine KK, Gilliland FD, Crowell RE, et al. The *XRCC1* 399 glutamine allele is a risk factor for adenocarcinoma of the lung. *Mutat Res* 2001; 461:273-8.
31. Kuschel B, Auranen A, McBride S, et al. Variants in DNA double-strand break repair genes and breast cancer susceptibility. *Hum Mol Genet* 2002;11:1399-407.
32. Matullo G, Guarrera S, Carturan S, et al. DNA repair gene polymorphisms, bulky DNA adducts in white blood cells and bladder cancer in a case-control study. *Int J Cancer* 2001;92:562-7.
33. Winsey SL, Haldar NA, Marsh HP, et al. A variant within the DNA repair gene *XRCC3* is associated with the development of melanoma skin cancer. *Cancer Res* 2000;60:5612-6.
34. Cho EY, Hildesheim A, Chen CJ, et al. Nasopharyngeal carcinoma and genetic polymorphisms of DNA repair enzymes *XRCC1* and *hOGG1*. *Cancer Epidemiol Biomark Prev* 2003;12:1100-4.
35. van Gils CH, Bostick RM, Stern MC, Taylor JA. Differences in base excision repair capacity may modulate the effect of dietary anti-oxidant intake on prostate cancer risk: an example of polymorphisms in the *XRCC1* gene. *Cancer Epidemiol Biomark Prev* 2002;11:1279-84.
36. Ratnasinghe D, Yao SX, Tangrea JA, et al. Polymorphisms of the DNA repair gene *XRCC1* and lung cancer risk. *Cancer Epidemiol Biomark Prev* 2001;10:119-23.
37. Tuimala J, Szekely G, Gundy S, Hirvonen A, Norppa H. Genetic polymorphisms of DNA repair and xenobiotic-metabolizing enzymes: role in mutagen sensitivity. *Carcinogenesis* 2002;23:1003-8.
38. Graur D, Li W-H. Dynamics of genes in populations. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinaur Associates, Inc; 2000. p. 41.
39. Dean M, Carrington M, O'Brien SJ. Balanced polymorphism selected by genetic *versus* infectious human disease. *Annu Rev Genomics Hum Genet* 2002;3:263-92.
40. Martin NMB. DNA repair inhibition and cancer therapy. *J Photochem Photobiol* 2001;63:162-70.