Neuroscience 145 (2007) 1273–1279

COMPUTATIONAL PREDICTION OF THE EFFECTS OF NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS IN HUMAN DNA REPAIR GENES

S. NAKKEN,[a] I. ALSETH[a] AND T. ROGNES[a,b]*

[a]*Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Rikshospitalet-Radiumhospitalet Medical Centre, NO-0027 Oslo, Norway*

[b]*Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316 Oslo, Norway*

Abstract—Non-synonymous single nucleotide polymorphisms (nsSNPs) represent common genetic variation that alters encoded amino acids in proteins. All nsSNPs may potentially affect the structure or function of expressed proteins and could therefore have an impact on complex diseases. In an effort to evaluate the phenotypic effect of all known nsSNPs in human DNA repair genes, we have characterized each polymorphism in terms of different functional properties. The properties are computed based on amino acid characteristics (e.g. residue volume change); position-specific phylogenetic information from multiple sequence alignments and from prediction programs such as SIFT (Sorting Intolerant From Tolerant) and PolyPhen (Polymorphism Phenotyping).

We provide a comprehensive, updated list of all validated nsSNPs from dbSNP (public database of human single nucleotide polymorphisms at National Center for Biotechnology Information, USA) located in human DNA repair genes. The list includes repair enzymes, genes associated with response to DNA damage as well as genes implicated with genetic instability or sensitivity to DNA damaging agents. Out of a total of 152 genes involved in DNA repair, 95 had validated nsSNPs in them. The fraction of nsSNPs that had high probability of being functionally significant was predicted to be 29.6% and 30.9%, by SIFT and PolyPhen respectively. The resulting list of annotated nsSNPs is available online (http://dna.uio.no/repairSNP), and is an ongoing project that will continue assessing the function of coding SNPs in human DNA repair genes.

Key words: SNP, non-synonymous, dbSNP, phenotypic effects.

*Correspondence to: T. Rognes, Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Rikshospitalet-Radiumhospitalet Medical Centre, NO-0027 Oslo, Norway. Tel: +47-22844787; fax: +47-22844782.
E-mail address: torbjorn.rognes@medisin.uio.no (T. Rognes).

Abbreviations: BER, base excision repair; BLAST, Basic Local Alignment Search Tool; cSNP, coding single nucleotide polymorphism (SNP in protein-coding region); dbSNP, public database of human single nucleotide polymorphisms at National Center for Biotechnology Information, USA; NCBI, National Center for Biotechnology Information; nsSNP, non-synonymous single nucleotide polymorphism; PDB, Protein Data Bank; PolyPhen, Polymorphism Phenotyping; PSSM, position-specific scoring matrix; SIFT, Sorting Intolerant From Tolerant; SNP, single nucleotide polymorphism.

The most common form of genetic variation in the human population occurs as single nucleotide polymorphisms (SNPs). It has been estimated that there exists one SNP with a minor allele frequency greater than 1% every 290 base-pairs in the human genome, implying a total of about 10 million SNPs (Kruglyak and Nickerson, 2001). The large number of SNPs, combined with a growing functional annotation of the human genome sequence, provides ample opportunity for developing improved links between genetic and phenotypic variation.

Detection of genetic variants that contribute susceptibility to a complex human disease is usually undertaken as an association analysis. In such studies, the allele frequencies of a set of polymorphisms are compared between affected cases and healthy controls. In this manner, one can identify markers that differ significantly between the two groups. A number of studies have shown association between one or a few polymorphisms and complex diseases, but most of them have been hard to replicate (Lohmueller et al., 2003). Inconsistent results may have many explanations. Often-cited reasons are improper study design, insufficient sample size and complexity of traits (Au and Salama, 2005; Newton-Cheh and Hirschhorn, 2005). The design and statistical analysis of genome-wide association studies is still a field in its infancy. The simplest and most commonly used strategy limits the polymorphic markers to the coding regions of candidate genes that are known or hypothesized to be associated with the trait of interest. By adopting such a direct approach, one is targeting a smaller number of polymorphisms that are themselves believed to be putative causal alleles. In order to increase the success rate of direct association studies, it is therefore critical to prioritize markers that have a high probability of being functional.

The most powerful way of assessing the effects of polymorphisms in coding regions is by focusing on the fraction that alter the encoded amino acid sequence (non-synonymous SNP (nsSNP) or missense change). These substitutions may directly affect the protein structure stability and efficiency of protein interactions. The biochemical severity (e.g. differences in side-chain polarity) of the substitution and the degree of evolutionary conservation at the variant site are examples of properties that indicate the degree of functional significance of an amino acid alteration. These features may act as predictors for the anticipated phenotypic effects of missense changes. A comprehensive analysis of properties of nsSNPs can as such make important contributions in the exploration of hypotheses concerning the plausible biological mechanisms that

0306-4522/07$30.00+0.00
doi:10.1016/j.neuroscience.2006.09.004

may explain the association of a gene with a specific trait. Indeed, in their in-depth-analysis of XPD (ERCC2) polymorphisms, Clarkson and Wood (2005) show that a closer investigation of the polymorphisms in question, both in silico and experimentally, is needed before phenotype/genotype association studies are performed.

DNA repair genes play a critical role in the maintenance of genome integrity. Variation in these genes may modulate the repair capacity, which in effect may lead to elevated risk of complex disease (e.g. various types of cancer) (Berwick and Vineis, 2000; Spitz et al., 2003; Hung et al., 2005). An overview and functional analysis of validated nsSNPs in the coding regions of 152 human DNA repair genes is presented here. Although restricted to repair genes in our analysis, the general strategy is applicable to any functional set of genes. The aim of the study is to assess the isolated phenotypic effects of all coding-region nsSNPs in repair genes, and as such contribute in the process of selecting target SNPs for direct association studies involving repair genes. Our selection of repair genes is based on the list provided by Wood and colleagues (Wood et al., 2001, 2005), accessible at http://www.cgal.icnet.uk/DNA_Repair_Genes.html. Two other genes of relevance to DNA repair are also included; TFAM (mitochondrial transcription factor A), responsible for maintaining mitochondrial DNA (Kang and Hamasaki, 2005), and SIRT3, a mitochondrial NAD-dependent deacetylase which may have a role in human longevity (Rose et al., 2003).

Initially, we outline a brief procedure for obtaining reliable SNP data. Further, we review the various computational approaches for in silico predictions of nsSNPs. Finally, we report the results of applying a set of these predictors on nsSNPs in repair genes. Previously, similar studies have been reported (Savas et al., 2004; Xi et al., 2004; Zhu et al., 2004; Rudd et al., 2005), with different types of data material and different subsets of repair genes. We provide an updated, comprehensive and publicly available resource on known polymorphisms within this functional category of genes.

SNP MINING

The largest repository of SNP data is located within the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA) dbSNP database (public database of human single nucleotide polymorphisms at National Center for Biotechnology Information, USA; http://www.ncbi.nlm.nih.gov/SNP). The bidirectional data exchange between dbSNP and other large SNP efforts such as HGVbase (Fredman et al., 2002) and TSC (Holden, 2002) has ensured its position as the main public resource for SNP mining. However, an unknown fraction of the submissions to dbSNP may not be true polymorphisms, but rather examples of sequencing errors or paralogous sequence variants (Fredman et al., 2004). Large-scale verification studies of putative polymorphic loci in dbSNP have reported a monomorphic rate of 17–48% (different rates are likely to stem from different technologies and protocols) (Carlson et al., 2003; Reich et al., 2003; Nelson et al., 2004). For the purpose of selecting proper SNPs for an association study it is therefore essential to exclude entries that have a high probability of being artifacts or monomorphic in the relevant population.

At first, a verification of the genomic location of a SNP should be performed by a BLAST (Basic Local Alignment Search Tool; Altschul et al., 1997) search with the flanking sequence of the SNP as the query (Savas et al., 2004). Neighboring sequences of SNPs that match multiple regions in the genome are then filtered out as unreliable. Furthermore, a reliability measure is provided by NCBI in the form of a record validation status; a polymorphism is *validated* by either independent submissions, frequency/genotype data, observed alleles in at least two chromosomes or submitted allele frequencies estimated by the HapMap project (International HapMap Consortium, 2003). The use of validated entries, preferably with allele frequency information, has shown to increase the genotype success rate significantly (Nelson et al., 2004; Edvardsen et al., 2006), indicating that these entries are more likely true polymorphisms. As of April 2006, dbSNP contains approximately 4.9 million validated biallelic positions in the human genome.

In our analysis of nsSNPs in repair genes, we have excluded non-validated polymorphisms and entries that map to multiple locations in the human genome. Since some repair genes encode several protein isoforms, a polymorphism is included several times if it occurs within more than one RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq) gene product. Whenever available, we annotate the polymorphisms with estimated population-specific allele frequencies from two different sources, the HapMap project and Perlegen (http://genome.perlegen.com) (Hinds et al., 2005).

PREDICTION OF FUNCTIONAL MISSENSE CHANGES

Allelic variants that alter the amino acid sequence of a gene product may affect the cellular phenotype at various levels. They may directly influence the stability of the native protein structure and the folding rate, resulting in a reduced concentration of the protein (Karchin et al., 2005). Polymorphisms residing in ligand-binding and catalytic sites may further affect protein interactions and other biochemical activities inside the cell (Sunyaev et al., 2001). Effects at the level of transcription, translation as well as post-translational modification can also occur, but these are relatively poorly characterized (Wang and Moult, 2001).

Following the steadily increasing number of known human nsSNPs, there has been growing interest in the identification of the subset that may affect the cellular phenotype. The ultimate goal is a separation between neutral, non-functional nsSNPs and the ones that are functional, providing a damaging potential to the encoded protein. The proposed approaches for this problem use various types of features for prediction, mainly physical and chemical properties of the amino acids, structural properties of the encoded protein and evolutionary properties derived from sequence alignments of homologous proteins