Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 49–58
GENOMICS
www.elsevier.com/locate/ygeno

Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family

Ali Torkamani [a], Nicholas J. Schork [b,*]

[a] Graduate Program in Biomedical Sciences, Department of Medicine, University of California at San Diego, La Jolla, CA 92093, USA
[b] Department of Psychiatry and Department of Family and Preventive Medicine, Center for Human Genetics and Genomics, Moores UCSD Cancer Center, California Institute of Telecommunications and Information Technology, Stein Institute for Research on Aging, University of California at San Diego, La Jolla, CA 92093, USA

Received 28 November 2006; accepted 10 March 2007
Available online 11 May 2007

Abstract

The human kinase gene family is composed of 518 genes that are involved in a diverse spectrum of physiological functions. They are also implicated in a number of diseases and encompass 10% of current drug targets. Contemporary, high-throughput sequencing efforts have identified a rich source of naturally occurring single nucleotide polymorphisms (SNPs) in kinases, a subset of which occur in the coding region of genes (cSNPs) and result in a change in the encoded amino acid sequence (nonsynonymous coding SNP; nscSNPs). What fraction of this naturally occurring variation underlies human disease is largely unknown (uDC), and much of it is assumed not to be disease causing (DC). We pursued a comprehensive computational analysis of the distribution of 1463 nscSNPs and 999 DC nscSNPs within the kinase gene family and have found that DCs are overrepresentated in the kinase catalytic domain and in receptor structures. In addition, the frequencies with which specific amino acid changes occur differ between the DCs and the uDCs, implying different biological characteristics for the two sets of human polymorphisms. Our results provide insights into the sequence and structural phenomena associated with naturally occurring kinase nscSNPs that contribute to human diseases.

Keywords: Kinase; Kinases; Disease; Single nucleotide polymorphism; Cancer; Protein domains; Amino acid sequence; Statistical study; Bioinformatics

The human protein kinase family contains 518 members, which regulate the activity of their substrates through reversible phosphorylation. As a group, they are involved in extracellular and intracellular signal transduction [1]. They are also involved in a number of other cellular processes, including metabolism, transcriptional regulation, cell cycle and apoptosis regulation, cytoskeletal rearrangements, and developmental processes [2]. Kinases, except for the atypical kinases, all contain a highly conserved catalytic core that can be complemented by a number of different regulatory domains (Fig. 1). These domains are involved in the determination of a particular kinase's specific set of substrates through a wide assortment of interactions including protein–protein, protein–membrane, and protein–carbohydrate interactions, in addition to kinase localization and response to a variety of signals including calcium, carbohydrates, and peptide hormones [3]. Alterations in protein kinase signaling play both fundamental and contributory roles in human disease [4]. In fact, kinases are the second largest family of current drug targets and are predicted to be the largest family of putative drug targets at 22% of the druggable genome [5].

An expanding body of literature and genomic databases consider single nucleotide polymorphisms (SNPs) that alter the coded amino acid sequence (nonsynonymous coding SNPs); (nscSNPs) of kinases [4,6–9]. Many of these nscSNPs are known to cause a distinct and overt disease phenotype and are classified in this study as "disease causing" (DCs) However, the majority of these nscSNPs are common and probably "neutral" variations within the human genome and are not associated with any overt clinical phenotype. We want to emphasize, however, that the functional effects of many of these SNPs have not been explored in full. As a result, we classify them as unknown as to

* Corresponding author. Fax: +1 858 822 2113.
E-mail address: nschork@ucsd.edu (N.J. Schork).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.03.006

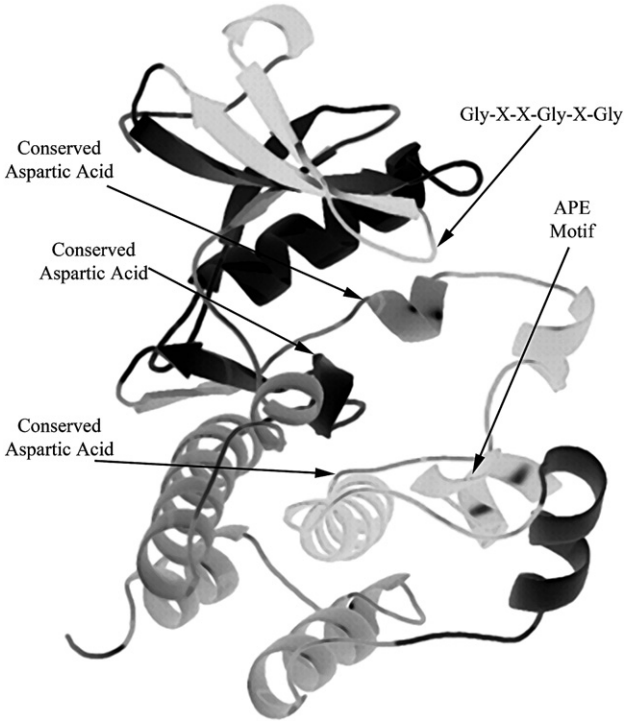

Fig. 1. Kinase catalytic core image modified from the Protein Kinase Resource (http://www.kinasenet.org/pkr/Welcome.do). A representative kinase catalytic core with some conserved motifs highlighted.

whether they cause disease (uDC). In this study, we have analyzed the distribution of nscSNPs in kinase domains and the frequency of specific amino acid transitions to predict and characterize the likely functional effects of nscSNPs in kinases. In this light, we pursued a number of different analyses that addressed the properties associated with kinase uDCs and DCs. These included (1) an analysis of the evolutionary conservation of the amino acids implicated in kinase nscSNPs as derived from the Panther database and analysis tools (http://www.pantherdb.org/), (2) an analysis of the distribution of nscSNPs (both uDCs and DCs) within different kinase groups, (3) an analysis of the domain distribution of the SNPs, (4) an analysis of amino acid distributions, (5) an analysis of amino acid changes induced by the nscSNPs, (6) an analysis of the nucleotides implicated in nscSNPs, and (7) a comprehensive and integrated analysis in which we tried to predict which groups, domains, etc. and their potential interactions differentiate uDCs from DCs. We also considered the comparison of mouse kinase SNPs and human kinase SNPs.

Results

SNP identification

Using public sources, we have compiled an extensive record of nscSNPs in kinases [10–13]. nscSNPs resulting in premature stop codons were excluded as these represent a rare, special class of nscSNPS that are very likely to be disease causing. In total, 999 DCs (41% of total nscSNPs identified) in 52 kinases and 1463 uDCs (59% of total nscSNPs identified) in 393 kinases were cataloged. Most kinases in the DC set had 20 or fewer DCs, while a few, BTK and RET, had over 100 DCs. All DCs were from published literature compiled in OMIM [7] (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), KinMutBase [4] (http://bioinf.uta.fi/KinMutBase/main_frame.html), and the Human Gene Mutation Database (HMGD) [9] (http://www.hgmd.cf.ac.uk/ac/index.php). The DCs that we identified were associated with a vast spectrum of inherited diseases including cancers, metabolic disorders, developmental diseases, and endocrine-related diseases. We obtained uDCs from dbSNP [6] (http://www.ncbi.nlm.nih.gov/projects/SNP/) and through the use of the PupaSNP [8] server (http://pupasnp.bioinfo.ochoa.fib.es/) to compile a list of SNPs that have not been functionally characterized. The wild-type or major amino acid was assumed to be the corresponding amino acid from published sequences in Kinbase (http://kinase.com/human/kinome/). nscSNP domain distribution was determined by using InterProScan [14] (http://www.ebi.ac.uk/InterProScan/) using mainly Prosite [15] and Pfam [16] domain determinations. Domains were then classified into more general categories including kinase catalytic (kinase; kin), extracellular receptor (receptor; recp), src homology (SH), pleckstrin homology (PH), fibronectin (FN), protein–protein interaction (PPI), protein–membrane interaction (PMI), carbohydrate binding (CB), immunoglobulin-like (IGL) domains that do not function as receptors, cytoskeletal interaction (CI), G-protein and GTPase interaction (GPI), and nucleic acid interaction (NAI) domains. nscSNPs in domains that did not clearly belong to one of the following categories were rare and grouped with nscSNPs outside of any functional domains (Table 2).

Evolutionary analysis via the panther database

We considered the use of the suite of analysis tools on the Panther database website to assess the conservation of the positions of the kinase nscSNPs. Using the substitution position-specific evolutionary conservation score, “subPSEC” (http://www.pantherdb.org/tools/csnpScoreForm.jsp), we were able to differentiate between uDCs (mean=−2.3125±0.04964) and DCs (mean=−4.1870±0.06830) by the Wilcoxon test ($p<0.0001$). The subPSEC score is derived from aligning a test protein against a library of hidden Markov models representing distinct protein families. The score is defined as $-|\ \ln(P_{aij}/P_{bij})\ |$, where P_{aij} is the probability of observing amino acid a at position i in HMM j. According to the Panther website, a score of −3 corresponds to a 50% probability that the SNP is disease-causing. This result suggests that the DCs in kinases occupy positions in DNA sequences that are more highly conserved across species than uDCs in kinases. We acknowledge that such an analysis has its limitations, since neighboring amino acids may influence the functional effects of the amino acid affected by an nscSNP. However, this fact would tend to bias the results toward the null hypothesis of no differences between DCs and uDCs; thus our observation of conservation differences is compelling given the conservative nature of the analysis.