Biochemical and Biophysical Research Communications **287,** 288–291 (2001)
doi:10.1006/bbrc.2001.5576, available online at http://www.idealibrary.com on IDEAL®

PicSNP: A Browsable Catalog of Nonsynonymous Single Nucleotide Polymorphisms in the Human Genome

Hangil Chang*,[1] and Toshiro Fujita†
**Health Service Center and †Department of Internal Medicine, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan*

Received July 23, 2001

Recent progress in identification and mapping of single nucleotide polymorphisms (SNPs) in the human genome generates an unprecedented opportunity to explore cause–effect relationships between genetic variations and susceptibility to common diseases. For this purpose, one promising strategy would be to select a set of SNPs that potentially alter the function of proteins involved in the pathogenesis of the diseases and compare their frequencies in the affected individuals and the healthy population. In this respect, SNPs that change amino acid sequences (nonsynonymous SNPs; nsSNPs) are of particular interest, since they are more likely to affect protein functions. In this study, we have constructed a catalog of nsSNPs (PicSNP), whose unique features are (i) nsSNPs are classified according to the functions of the affected genes and are searchable under the guidance of hierarchical lists of protein functions and (ii) nsSNPs that lead to amino acid changes in the known functional sites and domains of proteins are highlighted. Out of 1,190,295 SNPs extracted from public database, we identified 3793 nsSNPs and classified them in 1247 categories of protein functions. 495 sites and domains annotated in the Swiss-Prot database were found to include nsSNPs, including 2 nsSNPs in disulfide-binding sites and 38 nsSNPs in transmembrane regions. PicSNP is available via the World Wide Web (http://picsnp.org) and would support research questing for SNPs involved in common diseases.

Key Words: **SNP; nonsynonymous SNP; database; molecular function ontology; Swiss-Prot.**

Recently, the International SNP Map Working Group had reported mapping of 1.42 million SNPs in the human genome, providing an average density of one SNP every 1.9 kilobases (6). These polymorphisms constitute the major part of human DNA sequence variation, with the rest including insertion/deletion polymorphisms and repeat length polymorphisms. The high-density SNP map generates an unprecedented opportunity to identify those genes that determine the susceptibility to common diseases but that have been eluding conventional approach with linkage analysis and positional cloning. One promising strategy is to select a limited set of SNPs that (potentially) affect functions of the proteins that are involved in the biological processes deranged in the disease and examine the association of those SNPs with the disease (association study). In this respect, a subset of SNPs that are localized in the coding regions of candidate genes and cause changes in their deduced amino acid sequences (i.e., nsSNPs) are of particular interest, since they are easily identified and likely to affect the protein functions. SNPs that affect gene regulation are also of interest, but they are difficult to identify given our limited knowledge of regulatory signals in DNA. Therefore, an information system that establishes connections between protein functions and nsSNPs would be useful in selecting target nsSNPs for association study.

In this study, we have extracted nsSNPs from public database and classified them under the categories of protein functions. Classified nsSNPs are browsable through hierarchical lists of protein functions, thus providing a comprehensive view of connections between protein functions and nsSNPs. In addition, we have compared the locations of individual nsSNPs with annotations in the Swiss-Prot database and identified nsSNPs that cause amino acid changes in functional sites and domains. These nsSNPs are highlighted in the presentation of the catalog, under the assumption that they are further likely to affect the protein functions than other nsSNPs.

[1] To whom correspondence should be addressed. Fax: +81-3-5454-4307. E-mail: hchang-tky@umin.ac.jp.

0006-291X/01 $35.00

TABLE 1

Functional Sites and Domains Annotated in Swiss-Prot That Are Affected by nsSNPs

Swiss-Prot feature	nsSNPs
CARBOHYD	3
CHAIN	3
CONFLICT	73
DISULFID	2
DNA_BIND	2
DOMAIN	216
NP_BIND	1
PEPTIDE	5
PROPEP	11
REPEAT	26
SIGNAL	5
TRANSIT	2
TRANSMEM	38
VARIANT	84
VARSPLIC	18
ZN_FING	6
Total	495

Note. Functional sites and domains annotated in FT (feature table) lines in the Swiss-Prot database that include amino acids affected by nsSNPs are listed along with the numbers of involved nsSNPs. Meanings of feature symbols are: CARBOHYD, glycosylation site; CHAIN, extent of a polypeptide chain in the mature protein; CONFLICT, different papers report different sequences; DISULFID, disulfide bond; DNA_BIND, extent of a DNA binding region; DOMAIN, extent of a domain of interest on the sequence; NP_BIND, extent of a nucleotide phosphate binding region; PEPTIDE, extent of a released active peptide; PROPEP, extent of a propeptide; REPEAT, extent of an internal sequence repetition; SIGNAL, extent of a signal sequence; TRANSIT, extent of a transit peptide; TRANSMEM, extent of a transmembrane region; VARIANT, authors report that sequence variants exist; VARPLIC, description of sequence variants produced by alternative splicing; and ZN_FING, extent of a zinc finger region.

METHODS

SNPs were retrieved from annotations in the draft sequence of the human genome. Specifically, the draft sequence was downloaded from National Center for Biotechnology Information (NCBI) ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/, updated March 15, 2001), and SNPs were extracted as nucleotide positions that are annotated as 'variation' but do not represent insertion/deletion variations. These SNP sites were classified according to whether they are located in genes, mRNAs, and coding regions. Locations of genes, mRNAs, and coding regions were extracted from annotations of the draft sequence. nsSNPs were collected as those SNPs that are located within coding regions and also change the deduced amino acid sequence.

nsSNPs were classified under the categories of protein functions obtained from the molecular function "ontology" of Gene Ontology tool (5). Gene Ontology was constructed under the goal to "produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism," and it enumerates molecular functions in a set of well-defined categories that are structured in a hierarchical manner. We searched entries in the LocusLink database (3) that correspond to genes affected by individual nsSNPs, and from those entries followed the cross-references to Gene Ontology categories, completing the link from nsSNPs to Gene Ontology.

```
molecular_function [go|snp]
 signal transducer [go|snp]
  receptor [go|snp]
   transmembrane receptor [go|snp]
    G-protein coupled receptor [go|snp]
     peptide receptor [go]
      angiotensin II receptor [snp]
      bombesin receptor [snp]
      bradykinin receptor [snp]
      chemokine receptor [snp]
      cholecystokinin receptor [snp]
      endothelin receptor [snp]
      interleukin-8 receptor [snp]
      melanocortin receptor [snp]
      N-formyl peptide receptor [snp]
      neuropeptide receptor [snp]
      somatostatin receptor [snp]
      thrombin receptor [snp]
      vasopressin receptor [go|snp]
```

FIG. 1. A sample output of HTML page for "peptide receptor" in PicSNP.

Each coding sequence of the genes affected by nsSNPs were compared with sequences deposited in the Swiss-Prot database (release 39) (1). When exact match was encountered, annotations of the sequence data that appear in FT (feature table) lines in the database were examined whether they include the amino acids that are affected by nsSNPs.

RESULTS AND DISCUSSION

There were 1,327,856 annotated variations in the draft human genome sequence. Approximately 10% (135,783 variations) of them were variations that had ambiguous mapping and another 1778 variations were insertion/deletion variations. Excluding these variations, there remained 1,190,295 SNPs, of which 11,368 were localized in coding regions (coding-region SNPs). In 4215 cases of coding-region SNPs, deduced amino acid sequences were ambiguous (for example, due to the existence of undetermined nucleotide residue), and in other 4 cases polymorphism included an ambiguous allele (denoted by nucleotide 'N'). These SNPs were excluded from the catalog. In the remaining 7149 coding-region SNPs, 3356 SNPs were synonymous (i.e., unchanging the amino acid sequence). Consequently, we identified 3793 nsSNPs.

3793 nsSNPs were distributed among 2162 genes, and we could find corresponding Swiss-Prot entries in 627 SNPs (in 347 genes). 495 Swiss-Prot sequence

```
angiotensin II receptor
  angiotensin receptor 1
  angiotensin receptor 2
```

FIG. 2. A sample output of HTML page for "angiotensin II receptor" in PicSNP.

AGTR1 angiotensin receptor 1

dbSNP:1801021 336 T|P
dbSNP:1064533 289 C|W

```
    1 MILNSSTEDG IKRIQDDCPK AGRHNYIFVM IPTLYSIIFV VGIFGNSLVV IVIYFYMKLK
   61 TVASVFLLNL ALADLCFLLT LPLWAVYTAM EYRWPFGNYL CKIASASVSF NLYASVFLLT
  121 CLSIDRYLAI VHPMKSRLRR TMLVAKVTCI IIWLLAGLAS LPAIIHRNVF FIENTNITVC
  181 AFHYESQNST LPIGLGLTKN ILGFLFPFLI ILTSYTLIWK ALKKAYEIQK NKPRNDDIFK
  241 IIMAIVLFFF FSWIPHQIFT FLDVLIQLGI IRDCRIADIV DTAMPITICI AYFNNCLNPL
  301 FYGFLGKKFK RYFLQLLKYI PPKAKSHSNL STKMSTLSYR PSDNVSSSTK KPAPCFEVE*
```

Swiss-Prot says,
FUNCTION: RECEPTOR FOR ANGIOTENSIN II. MEDIATES ITS ACTION BY ASSOCIATION WITH G PROTEINS THAT ACTIVATE A PHOSPHATIDYLINOSITOL-CALCIUM SECOND MESSENGER SYSTEM.
SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN.
TISSUE SPECIFICITY: LIVER, LUNG, ADRENAL, AND ADRENOCORTICAL ADENOMAS.
PTM: CARBOXYL-TERMINAL SER OR THR RESIDUES MAY BE PHOSPHORYLATED.
SIMILARITY: BELONGS TO FAMILY 1 OF G-PROTEIN COUPLED RECEPTORS.

DOMAIN	1	27	EXTRACELLULAR (POTENTIAL).
TRANSMEM	28	52	1 (POTENTIAL).
DOMAIN	53	64	CYTOPLASMIC (POTENTIAL).
TRANSMEM	65	87	2 (POTENTIAL).
DOMAIN	88	102	EXTRACELLULAR (POTENTIAL).
TRANSMEM	103	124	3 (POTENTIAL).
DOMAIN	125	142	CYTOPLASMIC (POTENTIAL).
TRANSMEM	143	162	4 (POTENTIAL).
DOMAIN	163	192	EXTRACELLULAR (POTENTIAL).
TRANSMEM	193	214	5 (POTENTIAL).
DOMAIN	215	240	CYTOPLASMIC (POTENTIAL).
TRANSMEM	241	262	6 (POTENTIAL).
DOMAIN	263	275	EXTRACELLULAR (POTENTIAL).
TRANSMEM	276	296	7 (POTENTIAL).
DOMAIN	297	359	CYTOPLASMIC (POTENTIAL).
CARBOHYD	4	4	N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	176	176	N-LINKED (GLCNAC...) (POTENTIAL).
CARBOHYD	188	188	N-LINKED (GLCNAC...) (POTENTIAL).
DISULFID	101	180	BY SIMILARITY.
LIPID	355	355	PALMITATE (POTENTIAL).

FIG. 3. A sample output of HTML page for non-synonymous SNPs in type 1 angiotensin II receptor.

annotations included nsSNPs. The types of annotations along with the numbers of corresponding nsSNPs are listed in Table 1. Admittedly, localization of nsSNPs in the annotations does not directly imply that these nsSNPs are critically involved in the function of these proteins, since many of the annotations specify a broad range of the sequence such as "cytoplasmic domain." However, we identified 2 nsSNPs in disulfide-binding sites and 38 nsSNPs in transmembrane regions. These nsSNPs can be regarded as more likely to affect protein function than other nsSNPs. Additionally, 73 and 84 nsSNPs were localized in locations annotated as "CONFLICT" and "VARIANT," respectively, suggesting that these nsSNPs are the underlying causes of 'conflicts' and 'variations' in these amino acid sequences.

2826 nsSNPs (distributed among 1506 genes) could be connected to 1247 Gene Ontology categories. Hierarchical lists of these categories, together with associated nsSNPs, were implemented in a set of HyperText Markup Language (HTML) pages that can be explored dynamically via World Wide Web browsers. As an example, we present in Fig. 1 an HTML page that shows a Gene Ontology category of peptide receptor. One can reach this page by browsing pages for higher level categories such as "signal transducer" and "transmembrane receptor." From Fig. 1, it can be seen that angiotensin II receptor, which has an important role in sodium homeostasis and blood pressure regulation, has known nsSNPs. Following one more page (Fig. 2) that lists all the genes (type 1 and type 2 angiotensin II receptors) relevant to the category "angiotensin II receptor," one can reach a page that enumerates all nsSNPs located in (type 1) angiotensin II receptor (Fig. 3). This page shows two nsSNPs that affect 289th and 336th amino acid, respectively, along with their links to the dbSNP database (a central repository for SNPs in NCBI; Ref. 4); the complete amino acid sequence

with nsSNP sites highlighted (underlined in this figure, but displayed in a different color in the actual HTML page); comments about function and tissue specificity that appear in Swiss-Prot entry; and locations of sites and domains annotated in Swiss-Prot entry. We can see that one of the nsSNPs is localized in (potential) cytoplasmic domain, and another nsSNP is in the seventh (potential) transmembrane region.

Lastly, we notice that all the processes of constructing the catalog, from downloading the public database through creating the HTML pages, are fully automated by locally developed computer programs. Therefore, updating the catalog is straightforward and can be conducted without human supervision, which is an important feature in the face of expected explosion of publicly available SNPs in the coming years (2).

REFERENCES

1. Bairoch, A., and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.* **28,** 45–48.
2. Kruglyak, L., and Nickerson, D. A. (2001) Variation is the spice of life. *Nat. Genet.* **27,** 234–236.
3. Pruitt, K. D., and Maglott, D. R. (2001) RefSeq and Locus-Link: NCBI gene-centered resources. *Nucleic Acids Res.* **29,** 137–140.
4. Sherry, S. T., Ward, M., and Sirotkin, K. (1999) dbSNP—Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9,** 677–679.
5. The Gene Ontology Consortium (2000) Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25,** 25–29.
6. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933.