RAPID COMMUNICATION

Genomics & Informatics Vol. 4(1) 45-47, March 2006

D2GSNP: a web server for the selection of Single Nucleotide Polymorphisms within human disease genes

Hyojin Kang, Taehui Hong, Won-Hyong Chung, Younguk Kim, Jinhee Jung, Sohyun Hwang, Areum Han and Young Joo Kim*

National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoeun-dong, Yuseong-gu, Daejeon, 305-333, Korea

Abstract

D2GSNP is a web-based server for the selection of single nucleotide polymorphisms (SNPs) within genes related to human diseases. The D2GSNP is based on a relational database created by downloading and parsing OMIM, GAD, and dbSNP, and merging it with positional information of UCSC Golden Path. Totally our server provides 5,142 and 1,932 non-redundant disease genes from OMIM and GAD, respectively. With the D2GSNP web interface, users can select SNPs within genes responding to certain diseases and get their flanking sequences for further genotyping experiments such as association studies.

Keywords: SNP, disease gene, association study
Availability: D2GSNP is freely available at http://combio. kribb. re.kr:8080/D2GSNP/.

Summary

The importance of single nucleotide polymorphism (SNP) analysis has been increased continuously as the increasing requirements of numerous applications in complex genetic disease, pharmacogenomics, population genetics, and evolutionary studies (Gray *et al.*, 2000; Marnellos, G., 2003; Mooser *et al.*, 2003; Hacia *et al.*, 1999). Currently over 10 million human SNPs have been deposited into dbSNP database (Sherry *et al.*, 2001) and many companies have developed whole genome platforms for genotyping SNPs such as Affymetrix GeneChip, and Illumina BeadArray systems. Although the genotyping costs are rapidly decreasing, choosing effective SNPs within candidate genes is still important and critical to complex disease studies which needs numerous SNPs and large sample sets to maximize statistical power (Day, 2005; Hirschhorn, 2005).

We have developed a web server, D2GSNP, to support the selection of SNPs within human genes related to diseases (Fig. 1). Through the web interface, researchers are able to retrieve genes for queried diseases and SNPs within those genes intuitively. For example, a researcher who is interested in several diseases with limited resources may use D2GSNP to pick a gene-dense region of chromosomes with a few clicks and select effective SNPs for his/her disease association study. To automate the selection step, D2GSNP constructed a local relational database which integrated four public databases, OMIM (Online Mendelian Inheritance in Man), GAD (Genetic Association Database), dbSNP, and UCSC GoldenPath. The web-based interface was implemented using JavaServer Faces (JSF) technology which has an advantage of constructing a clearly defined architecture by separating application logic and presentation. It helps the rapid construction of web services and lowers the cost of maintenance.

Mapping diseases to genes

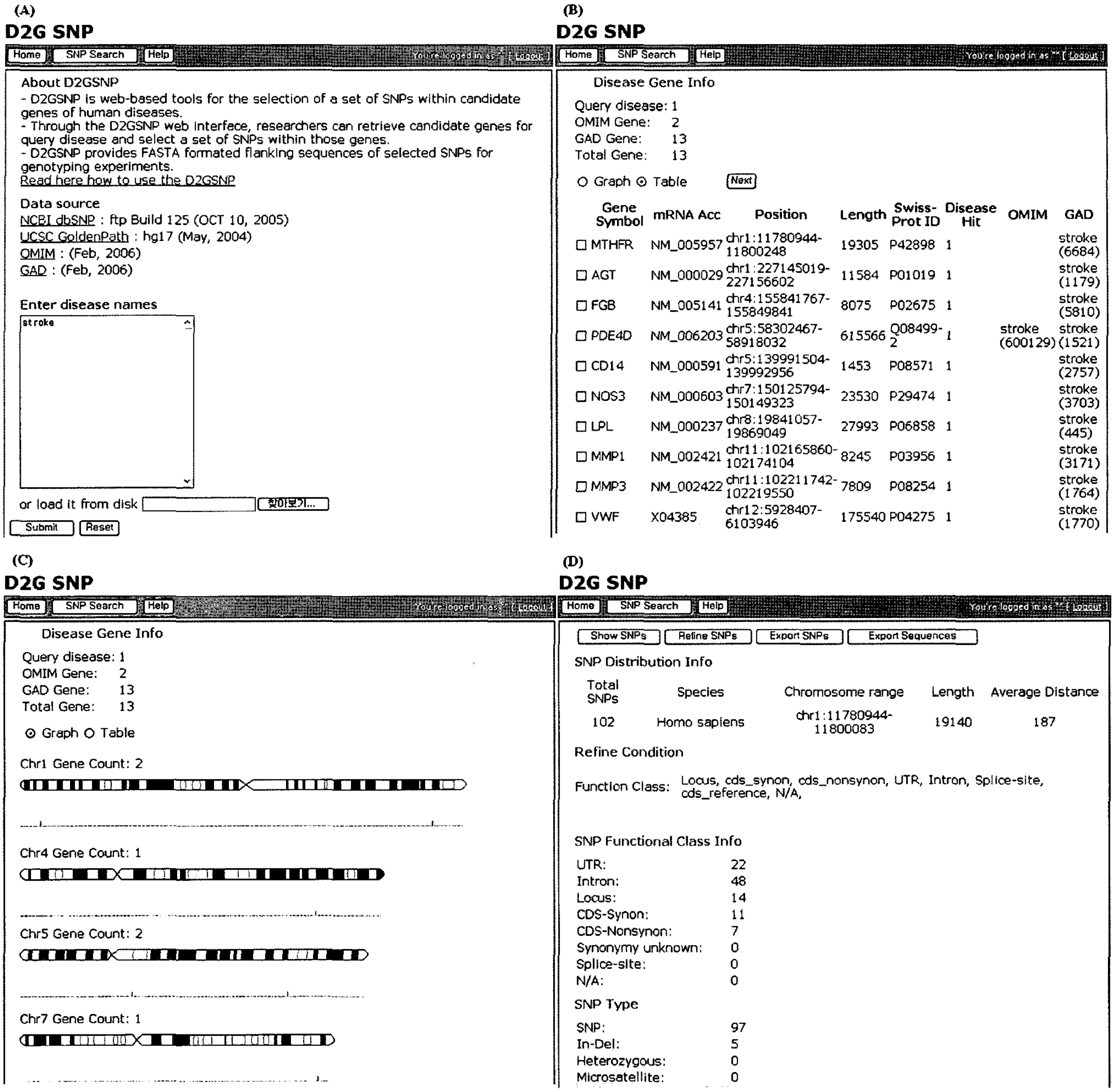
Previously, there have been many alternative methods to identify human genes related to monogenic or complex diseases (Perez-Iratxeta *et al.*, 2002; Lopez-Bigas *et al.*, 2004; Tiffin *et al.*, 2005). We constructed highly accurate gene maps of diseases based on two manually curated databases including OMIM and GAD. OMIM is a catalog of human genes and genetic disorders (Hamosh *et al.*, 2002) and GAD is an archive of human genetic association studies of complex diseases and disorders (Becker *et al.*, 2004). In detail, we use OMIM Morbid Map which is the cytogenetic map location of disease genes described in OMIM. Among GAD genes, we used genes which showed at least one positive association to a disease. As a result, the number of disease-gene records and non-redundant gene counts in OMIM were 4,058 and 5,142, respectively. There were 8,179 disease-gene records and 1,932 non-redundant genes in GAD.

Mapping genes to SNPs

In order to map SNPs to disease genes, chromosomal position of gene boundaries were retrieved from knownGene,

*Corresponding author: E-mail yjkim8@kribb.re.kr, Tel +82-42-879-8540, Fax +82-42-879-8519

Accepted 28 February 2006



(A)

About D2GSNP

- D2GSNP is web-based tools for the selection of a set of SNPs within candidate genes of human diseases.
- Through the D2GSNP web interface, researchers can retrieve candidate genes for query disease and select a set of SNPs within those genes.
- D2GSNP provides FASTA formated flanking sequences of selected SNPs for genotyping experiments.

Read here how to use the D2GSNP

Data source
NCBI dbSNP : ftp Build 125 (OCT 10, 2005)
UCSC GoldenPath : hg17 (May, 2004)
OMIM : (Feb, 2006)
GAD : (Feb, 2006)

Enter disease names

stroke

or load it from disk

(B)

Disease Gene Info

Query disease: 1
OMIM Gene: 2
GAD Gene: 13
Total Gene: 13

○ Graph ● Table

Gene Symbol	mRNA Acc	Position	Length	Swiss-Prot ID	Disease Hit	OMIM	GAD
□ MTHFR	NM_005957	chr1:11780944-11800248	19305	P42898	1		stroke (6684)
□ AGT	NM_000029	chr1:227145019-227156602	11584	P01019	1		stroke (1179)
□ FGB	NM_005141	chr4:155841767-155849841	8075	P02675	1		stroke (5810)
□ PDE4D	NM_006203	chr5:58302467-58918032	615566	Q08499-2	1	stroke (600129)	stroke (1521)
□ CD14	NM_000591	chr5:139991504-139992956	1453	P08571	1		stroke (2757)
□ NOS3	NM_000603	chr7:150125794-150149323	23530	P29474	1		stroke (3703)
□ LPL	NM_000237	chr8:19841057-19869049	27993	P06858	1		stroke (445)
□ MMP1	NM_002421	chr11:102165860-102174104	8245	P03956	1		stroke (3171)
□ MMP3	NM_002422	chr11:102211742-102219550	7809	P08254	1		stroke (1764)
□ VWF	X04385	chr12:5928407-6103946	175540	P04275	1		stroke (1770)

(C)

Disease Gene Info

Query disease: 1
OMIM Gene: 2
GAD Gene: 13
Total Gene: 13

● Graph ○ Table

Chr1 Gene Count: 2

Chr4 Gene Count: 1

Chr5 Gene Count: 2

Chr7 Gene Count: 1

(D)

SNP Distribution Info

Total SNPs	Species	Chromosome range	Length	Average Distance
102	Homo sapiens	chr1:11780944-11800083	19140	187

Refine Condition

Function Class: Locus, cds_synon, cds_nonsynon, UTR, Intron, Splice-site, cds_reference, N/A,

SNP Functional Class Info

UTR:	22
Intron:	48
Locus:	14
CDS-Synon:	11
CDS-Nonsynon:	7
Synonymy unknown:	0
Splice-site:	0
N/A:	0

SNP Type

SNP:	97
In-Del:	5
Heterozygous:	0
Microsatellite:	0

Fig. 1. A web interface of D2GSNP. (A) Disease name input page. Users can input multiple disease names through input area, or upload a file which contains a list of disease names (B) Search result showing disease related genes with a table view. (C) Chromosomal distribution of disease related genes with a graphic view. The bars indicate the locus of the disease genes. (D) SNP search result page. Users are able to navigate or refine a set of SNPs with various conditions, and download their flanking sequences.

which provides the information of protein coding genes based on proteins from Uniprot andtheir corresponding mRNAs from GenBank. To link knownGenes with gene symbols from OMIM and GAD, UCSC's table kgXref (known genes to external reference) was used. However, a few gene symbols from OMIM and GAD did not match with those from kgXref. Some of these orphan gene symbols were finally linked through gene alias table based on Entrez gene which covers more gene symbols. Our system is based on hg17 in UCSC GoldenPath and Build 125 in dbSNP. Note that the number of SNPs in our database is lower than the total number of SNPs in dbSNP, because some of them have not been mapped to a unique position on the human genome yet.

SNPs filtering and Data export

The general purpose of D2GSNP is to select SNPs within human genes related to user-queried disease. SNPs can be filtered through their variation types, validation status, minor allele frequency, and functional classes. D2GSNP also provides flanking sequences of selected SNPs for genotyping experiments. Flanking sequences of SNPs with user-defined length are provided in a FASTA format file. Fig. 1D shows an example of filtered SNPs. Usage details of the D2GSNP are available in the on-line help page.

Acknowledgements

This research was supported by a grant from KRIBB Research Initiative Program and in part by the Korean HapMap Project from the Korea Ministry of Science & Technology.

References

Day, C. P. (2005). Genetic studies to identify hepatic fibrosis genes and SNPs in human populations. *Methods Mol. Med.* 117, 315-331.

Gray, I. C., Campbell, D. A., and Spurr, N. K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Gene.* 9, 2403-2408.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M. *et al.* (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* 22, 164-167.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V.A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52-55.

Hirschhorn, J.N. (2005). Genetic approaches to studying common diseases and complex traits. *Pediatr. Res.* 57 (5 Pt 2), 74R-77R.

Lopez-Bigas, N. and Ouzounis, C.A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* 32, 3108-3114.

Marnellos, G. (2003). High-throughput SNP analysis for genetic association studies. *Curr. Opin. Drug Discov. Devel.* 6, 317-321.

Mooser, V., Waterworth, D.M., Isenhour, T., and Middleton, L. (2003). Cardiovascular pharmacogenetics in the SNP era. *J. Thromb. Haemost.* 1, 1398-1402.

Perez-Iratxeta, C., Bork, P., and Andrade, M.A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genet.* 31, 316-319.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311.

Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., and Hide, W.A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* 33, 1544-1552.

Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The Genetic Association Database. *http://geneticassociationdb.nih.gov/.*

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *http://genome.ucsc.edu/.*

McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). (2000). Online Mendelian Inheritance in Man, OMIM (TM). *http://www.ncbi.nlm.nih.gov/omim/.*

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *http://www.ncbi.nlm.nih.gov/SNP/.*