APPLICATION NOTE Genomics & Informatics Vol. 6(3) 153-156, September 2008

StrokeBase: A Database of Cerebrovascular Disease-related Candidate Genes

Young-Uk Kim[1]*, Il-Hyun Kim[1], Ok-Sun Bang[2] and Young-Joo Kim[1]

[1]Medical Genomics Research Center, KRIBB, Daejeon 305-806, Korea, [2]Department of Medical Research, KIOM, Daejeon 305-810, Korea

Abstract

Complex diseases such as stroke and cancer have two or more genetic loci and are affected by environmental factors that contribute to the diseases. Due to the complex characteristics of these diseases, identifying candidate genes requires a system-level analysis of the following: gene ontology, pathway, and interactions. A database and user interface, termed StrokeBase, was developed; StrokeBase provides queries that search for pathways, candidate genes, candidate SNPs, and gene networks. The database was developed by using in silico data mining of HGNC, ENSEMBL, STRING, RefSeq, UCSC, GO, HPRD, KEGG, GAD, and OMIM. Forty candidate genes that are associated with cerebrovascular disease were selected by human experts and public databases. The networked cerebrovascular disease gene maps also were developed; these maps describe gene-gene interactions and biological pathways. We identified 1127 genes, related indirectly to cerebrovascular disease but directly to the etiology of cerebrovascular disease. We found that a protein-protein interaction (PPI) network that was associated with cerebrovascular disease follows the power-law degree distribution that is evident in other biological networks. Not only was in silico data mining utilized, but also 250K Affymetrix SNP chips were utilized in the 320 control/disease association study to generate associated markers that were pertinent to the cerebrovascular disease as a genome-wide search. The associated genes and the genes that were retrieved from the in silico data mining system were compared and analyzed. We developed a well-curated cerebrovascular disease-associated gene network and provided bioinformatic resources to cerebrovascular disease researchers. This cerebrovascular disease network can be used as a frame of systematic genomic research, applicable to other complex diseases. Therefore, the ongoing database efficiently supports medical and genetic research in order to overcome cerebrovascular disease.

Availability: StrokeBase is freely available at http://sysbio.kribb.re.kr:8080/stroke/.

Keywords: stroke, cerebrovascular disease, SNP, disease gene, association study

Introduction

Cerebrovascular disease is the third-leading cause of death, behind heart disease and cancer (Toole. 1999). Complex disorders such as stroke and cancer have two or more genetic loci and environmental factors that contribute to the diseases (Braunwald *et al.*, 2001; Jin *et al.*, 2008). The objective of this study is to provide a systematic approach to analyze the complex effects of genes and to represent the frame of their relationships with specific diseases. To discover candidate genes that are related to complex disease, open sources, such as Entrez Pubmed, OMIM, PathwayStudio, HPRD, and Gencards, were thoroughly utilized, resulting in a vast amount of genetic and pathway information with regard to the etiology of cerebrovascular disease. These genes were annotated to access information on transcription, translation, structural function, and relatedness to the disease. In addition to in silico data mining, 320 250K Affymetrix SNP chips (GeneChip^{TM} Human Mapping 250K Nsp Array, Affymetrix, Inc., CA) were utilized for a case/control association study to generate experimentally associated markers of cerebrovascular disease. The associated genes from the SNP chips and the genes that were retrieved from in silico data mining systems were compared and analyzed. A protein-protein network diagram that showed the integrated markers and their relationships was constructed in order to analyze the network characteristics and produce hub genes. It was found that the PPI network that was associated with cerebrovascular disease follows a power-law degree distribution, as other biological networks do (Peri *et al.*, 2003). The PathwayStudio 5.0 program (Ariadne, Inc., MD, USA) was utilized to process the natural text mining of PubMed abstracts; the use of PathwayStudio resulted in a gene-disease association network. The etiology of the disease and its related genes, which were extracted from in silico data mining and network analy-

*Corresponding author: E-mail yjkim8@kribb.re.kr
Tel +82-42-879-8127, Fax +82-42-879-8119
Accepted 9 September 2008

sis as well as from genome-wide association studies, were accumulated to the web-based portal database named StrokeBase (http://sysbio.kribb.re.kr:8080/stroke/) (Fig. 1). Information in this integrated cerebrovascular disease network may serve as a frame for systematic genomic research for cerebrovascular diseases, as well as for other complex diseases such as cancers.

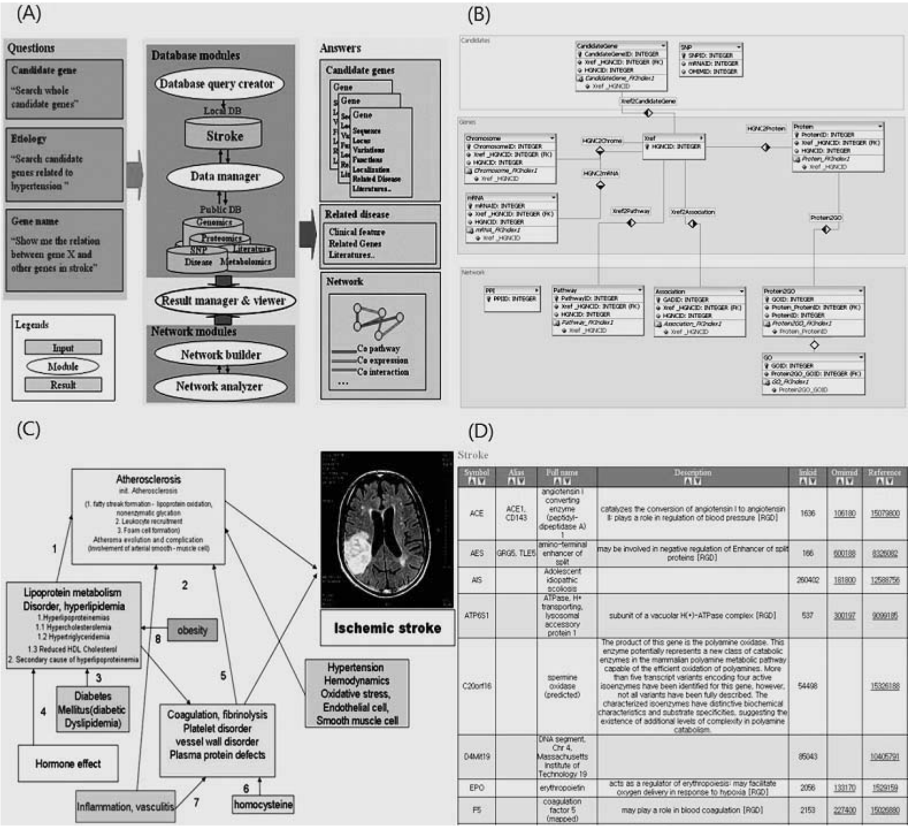
Utility and Discussion

In order to map SNPs to disease-related genes, the chromosomal position of gene boundaries was retrieved from KnownGene, a program that provides information on protein-coding genes based on proteins from Uniprot with their 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.

A database was developed by in silico data mining using HGNC, ENSEMBL, STRING, RefSeq, UCSC, GO, HPRD, KEGG, GAD, and OMIM (Fig. 1B). The Tomcat web server was utilized in StrokeBase, which simplifies building user interfaces for JSP (Java Server Page). Genes that were noted in Morbid Map (OMIM MorbidMap; http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi) and the Genetic Association Database (GAD; http://geneticassociationdb.nih.gov/) were selected (Hamosh *et al*., 2002). To cover up-to-date experimental results, a literature search also was performed by biologists. Because genes are noted by their various aliases, we mapped the gene set to the HUGO Gene Nomenclature Committee (HGNC, http://www.genenames.org/) to retain consistency of gene entries. We identified 1127 genes that were indirectly related to cerebrovascular disease but directly related to the etiology of cerebrovascular disease. The Pathway overview page has the list of signaling pathways that are related to cerebrovascular disease. This information refers to the pathway made by KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa *et al*., 2000).

The cerebrovascular disease etiology map page refers to the risk factors that are associated with cerebrovascular disease and to the network of risk factors (Fig. 1C). If one of the risk factors is clicked, a gene list that is associated with the risk factor will be displayed. Forty genes that are associated with cerebrovascular disease were selected by human experts who utilized public databases (Fig. 1D).

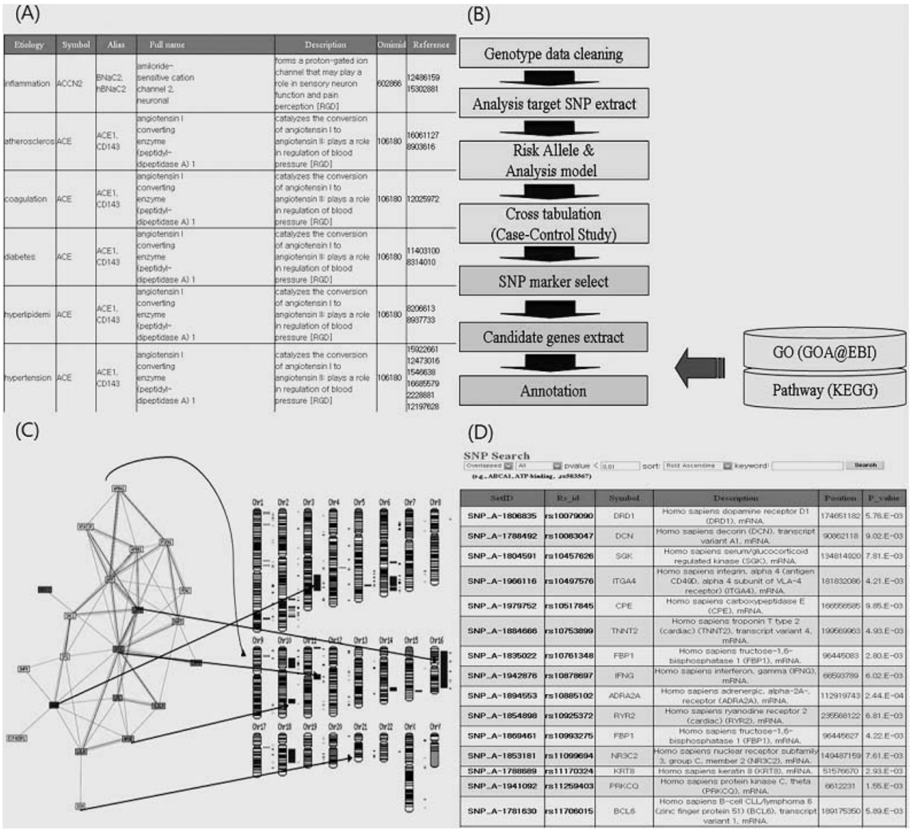
250K Affymetrix SNP chips were utilized in the 320 control/disease association study in order to generate associated markers for cerebrovascular disease in a genome-wide search. Some of the candidate genes that



Stroke

Symbol	Alias	Full name	Description	Inkid	Omimid	Reference
ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II: plays a role in regulation of blood pressure [RGD]	1636	106180	15079800
AES	GRG5, TLE5	amino-terminal enhancer of split	may be involved in negative regulation of Enhancer of split proteins [RGD]	166	600188	8326082
AIS		Adolescent idiopathic scoliosis		260402	181800	12588756
ATP6S1		ATPase, H+ transporting, lysosomal accessory protein 1	subunit of a vacuolar H(+)-ATPase complex [RGD]	537	300197	9099185
C20orf16		spermine oxidase (predicted)	The product of this gene is the polyamine oxidase. This enzyme potentially represents a new class of catabolic enzymes in the mammalian polyamine metabolic pathway capable of the efficient oxidation of polyamines. More than five transcript variants encoding four active isoenzymes have been identified for this gene, however, not all variants have been fully described. The characterized isoenzymes have distinctive biochemical characteristics and substrate specificities, suggesting the existence of additional levels of complexity in polyamine catabolism.	54498		15326188
D4Mit19		DNA segment, Chr 4, Massachusetts Institute of Technology 19		85043		10405791
EPO		erythropoietin	acts as a regulator of erythropoiesis: may facilitate oxygen delivery in response to hypoxia [RGD]	2056	133170	1529159
F5		coagulation factor 5 (mapped)	may play a role in blood coagulation [RGD]	2153	227400	15026880

Fig. 1. Screenshots for StrokeBase (http://sysbio.kribb.re.kr:8080/stroke/). (A) The system structure for data manager, result viewer, and network analyzer. (B) The physical schema of the database. (C) Etiology map summarizing cerebrovascular disease risk factors. (D) Gene list associated with risk factors of cerebrovascular disease. Kim *et al*.



Etiology	Symbol	Alias	Full name	Description	Omimid	Reference
inflammation	ACCN2	BNaC2, hBNaC2	amiloride-sensitive cation channel 2, neuronal	forms a proton-gated ion channel that may play a role in sensory neuron function and pain perception [RGD]	602866	12486159 15302881
atherosclerosis	ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II; plays a role in regulation of blood pressure [RGD]	106180	16061127 8903616
coagulation	ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II; plays a role in regulation of blood pressure [RGD]	106180	12025972
diabetes	ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II; plays a role in regulation of blood pressure [RGD]	106180	11403100 8314010
hyperlipidemi	ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II; plays a role in regulation of blood pressure [RGD]	106180	8206613 8937733
hypertension	ACE	ACE1, CD143	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	catalyzes the conversion of angiotensin I to angiotensin II; plays a role in regulation of blood pressure [RGD]	106180	15922661 12473016 1546638 16685579 2228881 12197628

SetID	Rs_id	Symbol	Description	Position	P_value
SNP_A-1806835	rs10079090	DRD1	Homo sapiens dopamine receptor D1 (DRD1), mRNA.	174651182	5.78.E-03
SNP_A-1788492	rs10083047	DCN	Homo sapiens decorin (DCN), transcript variant A1, mRNA.	90862118	9.02.E-03
SNP_A-1804591	rs10457626	SGK	Homo sapiens serum/glucocorticoid regulated kinase (SGK), mRNA.	134814920	7.81.E-03
SNP_A-1966116	rs10497576	ITGA4	Homo sapiens integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor) (ITGA4), mRNA.	181832086	4.21.E-03
SNP_A-1979752	rs10517845	CPE	Homo sapiens carboxypeptidase E (CPE), mRNA.	166556585	9.85.E-03
SNP_A-1884666	rs10753899	TNNT2	Homo sapiens troponin T type 2 (cardiac) (TNNT2), transcript variant 4, mRNA.	199569963	4.93.E-03
SNP_A-1835022	rs10761348	FBP1	Homo sapiens fructose-1,6-bisphosphatase 1 (FBP1), mRNA.	96445083	2.80.E-03
SNP_A-1942876	rs10878697	IFNG	Homo sapiens interferon, gamma (IFNG), mRNA.	66593789	6.02.E-03
SNP_A-1894553	rs10885102	ADRA2A	Homo sapiens adrenergic, alpha-2A-, receptor (ADRA2A), mRNA.	112919743	2.44.E-04
SNP_A-1854898	rs10925372	RYR2	Homo sapiens ryanodine receptor 2 (cardiac) (RYR2), mRNA.	235568122	6.81.E-03
SNP_A-1869461	rs10993275	FBP1	Homo sapiens fructose-1,6-bisphosphatase 1 (FBP1), mRNA.	96445627	4.22.E-03
SNP_A-1853181	rs11099694	NR3C2	Homo sapiens nuclear receptor subfamily 3, group C, member 2 (NR3C2), mRNA.	149487159	7.61.E-03
SNP_A-1788689	rs11170324	KRT8	Homo sapiens keratin 8 (KRT8), mRNA.	51576670	2.93.E-03
SNP_A-1941092	rs11259403	PRKCQ	Homo sapiens protein kinase C, theta (PRKCQ), mRNA.	6612231	1.55.E-03
SNP_A-1781630	rs11706015	BCL6	Homo sapiens B-cell CLL/lymphoma 6 (zinc finger protein 51) (BCL6), transcript variant 1, mRNA.	189175350	5.89.E-03

Fig. 2. Screenshots for the candidate genes identified in StrokeBase. (A) Search result table showing etiology candidate genes. (B) SNP chip data analysis flow. (C) Candidate gene network with chromosome mapping. (D) Search table showing associated SNPs by 320 250K Affymetrix SNP chips. Kim *et al*.

were identified in the chip analyses are shown (Fig. 2A). The SNP chip data analysis flow is shown (Fig. 2B). SNP-markers that were selected by the experiments were annotated and chromosomally mapped to reveal a comprehensive, networked gene map (Fig. 2C). We found overlaps between the gene markers that were annotated by experts and by the SNP experiment. Each SNP ID is linked to more detailed information. Selected SNPs with a given p-value are available in StrokeBase (Fig. 2D).

Conclusion

StrokeBase is an online database for cerebrovascular disease researchers. StrokeBase makes it possible for scientists to provide cerebrovascular disease association markers by in silico and whole-genome SNP arrays. Using data and tools that currently are available in StrokeBase, a scientist is able to analyze the complex effect of genes and to represent the frame of their relationships associated with cerebrovascular disease. The etiology of cerebrovascular disease was constructed, and 40 genes directly related to cerebrovascular disease were identified. Also, 1127 genes that were indirectly related to cerebrovascular disease but directly related to the etiology of cerebrovascular disease were identified. The SNP chip-associated genes and those genes that were retrieved from in silico and data mining systems were compared. We developed a well-curated cerebrovascular disease-associated gene network and provided bioinformatic resources to cerebrovascular disease researchers. StrokeBase will thus serve as a novel and powerful resource for reviewing and interpreting cerebrovascular disease association data.

Acknowledgments

This research was supported by the Cerebrovascular Disease Project of the Ministry of Education, Science and Technology (MEST) and in part by the KRIBB Intramural Research Program.

References

Braunwald, E., Fauci, A.S., Kasper, D.L., Hauser, S.L., Longo, D.L., and Jameson, J.L. (2001). *Harrison's principles of internal medicine* (15th int. ed), McGraw-Hill.

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.

Jin, H., Kim, S.H., Kim, Y.U., Park, Y.K., Ji, M., and Kim, Y.J. (2008). Development of KHapmap Browser using DAS for Korean HapMap Research. *Genomics & Informatics* 6, 57-63.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27-30.

Peri, S. *et al.* (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research* 13, 2363-2371.

Toole, J.F. (1999). *Chapter 13: Strokes in the young. Cerebrovascular disorders.* Edition 5, Lippincott Williams & Wilkins, Philadelphia, pp.283-316.

Websites

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/.*

Nucleus Medical Art: 3D Medical Animations. (2008). *http://www.nucleusinc.com.*

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/.*

xPharm: The Comprehensive Pharmacology Reference. (2008). *http://www.xpharm.com/citation?Article_ID=1050.*