An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases

Jin Ok Yang[1*], Sohyun Hwang[1, 2*], Jeongsu Oh[1], Jong Bhak[1], and Tae-Kwon Sohn[1, 3§]

[1]Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, 305-806, Korea

[2]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

[3]Department of Biochemistry, Yonsei University, Seoul, Korea

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

Jin Ok Yang: joy@kribb.re.kr

Sohyun Hwang: winnie79@biosoft.kaist.ac.kr

Jeongsu Oh: ofang@kribb.re.kr

Jong Bhak: jongbhak@yahoo.com

Tae-Kwon Sohn: loubert@naver.com

Abstract

Background

Studies on the relationship between disease and genetic variations such as single nucleotide polymorphisms (SNPs) are important. Genetic variations can cause disease by influencing important biological regulation processes. Despite the needs for analyzing SNP and disease correlation, most existing databases provide information only on functional variants at specific locations on the genome, or deal with only a few genes associated with disease. There is no combined resource to widely support gene-, SNP-, and disease-related information, and to capture relationships among such data. Therefore, we developed an integrated database-pipeline system for studying SNPs and diseases.

Results

To implement the pipeline system for the integrated database, we first unified complicated and redundant disease terms and gene names using the Unified Medical Language System (UMLS) for classification and noun modification, and the HUGO Gene Nomenclature Committee (HGNC) and NCBI gene databases. Next, we collected and integrated representative databases for three categories of information. For genes and proteins, we examined the NCBI mRNA, UniProt, UCSC Table Track and MitoDat databases. For genetic variants we used the dbSNP, JSNP, ALFRED, and HGVbase databases. For disease, we employed OMIM, GAD, and HGMD databases. The database-pipeline system provides a disease thesaurus, including genes and SNPs associated with disease. The search results for these categories are available on the web page (http://diseasome.kobic.re.kr/), and a genome browser is also available to highlight findings, as well as to permit the convenient review of

potentially deleterious SNPs among genes strongly associated with specific diseases and clinical phenotypes.

Conclusions

Our system is designed to capture the relationships between SNPs associated with disease and disease-causing genes. The integrated database-pipeline provides a list of candidate genes and SNP markers for evaluation in both epidemiological and molecular biological approaches to diseases-gene association studies. Furthermore, researchers then can decide semi-automatically the data set for association studies while considering the relationships between genetic variation and diseases. The database can also be economical for disease-association studies, as well as to facilitate an understanding of the processes which cause disease. Currently, the database contains 14,674 SNP records and 109,715 gene records associated with human diseases and it is updated at regular intervals.

Background

Many researchers have studied the relationships between disease and biological variations such as single nucleotide polymorphisms (SNPs), copy number variation, sequence repeats and genetic rearrangement [1-3]. Recently, work on genetic (SNP) variation associated with diseases has become intense, as many genetic variations are thought to affect the structure and function of proteins, as a result of amino acid substitutions [4, 5]. Significantly, SNPs, which report over 90% of genetic variation in the human genome [6], can have a major impact on how humans respond to disease, to drugs, and to other therapies. Therefore, SNP information is a great resource in biomedical studies, diagnostics, and drug development [7].

Many researchers studying disease associated SNPs require integrated information on SNPs and disease for two reasons. First, in order to capture relationships between SNPs and diseases, and then, to understand which genes cause disease and how that is impacted by SNPs. Second, disease-association researchers can save much time and effort in identifying the candidate disease-causing genes.

Despite the needs, existing servers contain insufficient information about SNP-disease associations. Because public databases for SNPs and diseases are large, complicated, and difficult to use, their integration is challenging. Therefore, we developed an integrated database-pipeline system for studying SNP and disease-association. We constructed a large database with comprehensive data on genes and SNPs associated with disease. In particular, the database-pipeline system allows biologists to retrieve integrated information on diseases, SNPs, and amino acid changes, along with functional annotation.

Methods and Results

The integrated database-pipeline system of genes and SNPs associated with diseases was developed in three parts as shown in Figure 1. By using this pipeline system, we downloaded and extracted primary information from 13 public and private databases. Next, we unified complex disease terms and gene names, and constructed an integrated database which contains the three sub-categories of diseases; genes and proteins; and SNPs.

Automatic collection and update of public resources

The integrated database-pipeline system uses file transfer protocol, hypertext transfer protocol, and JAVA-based data-extracting modules. The system also has a support

function to design the database schema and to create the modules based on a graphic user interface. The integration pipeline system checks the updated data and downloads such data automatically from 13 public and private resource servers, and then informs the system administrator by e-mail. We selected the following representative databases for the disease, SNP, and gene resources: The disease category is updated from the databases Unified Medical Language System (UMLS) [8], Online Mendelian Inheritance in Man (OMIM) [9], Gene Association Database (GAD) [10], and Human Gene Mutation Database (HGMD) [11]. The gene and protein category is updated from the databases NCBI [12], HUGO Gene Nomenclature Committee (HGNC) [13], UniProt [14], UCSC[15], and MitoDat (Mendelian Inheritance and the Mitochondrion) [16]. The genetic variation category (SNPs) is updated from the databases dbSNP [17], JSNP [18], ALFRED (Allele Frequency Database) [19], and HGVbase (Human Genome Variation database) [20]. The system is updated regularly, as the pipeline aquires data automatically.

Defining disease terms based on the UMLS

The disease terms commonly used in many research articles and several disease databases such as OMIM, GAD, and HGMD have the character of natural language; there are many synonyms and slightly different expressions which refer to the same concept. We required a unified controlled vocabulary of disease names and their synonyms, to construct a non-redundant disease database. We accomplished this by using UMLS (Release Archive 2007AA), which is a very large, multi-purpose vocabulary database containing information about biomedical and health-related concepts, the various terms used, and the relationships among them. Moreover, UMLS successfully integrates widely used clinical terms, in sub-bases such as “Systematized Nomenclature of Medicine – Clinical Terms and Medical Subject

Headings," so UMLS was an excellent resource allowing us to relate our database terms to medical informatics.

In addition, disease terms are associated with various word formations (in particular, noun modification). To solve this text stemmer problem, we defined disease terms expressed in public disease databases using four steps. First, we removed stop words employing a stop word list provided by OMIM. Second, we removed suffixes such as ", -es and -s. Third, we removed typographical errors and special characters. Finally, we mapped these processed disease terms to unique clinical concepts by comparing the terms with their several synonyms and different expressions as provided by UMLS [8].

Defining genes according to HGNC and NCBI data

To permit the exploration of SNP effects on genetic variation, we adopted various gene annotations including gene Information from NCBI, RefSeq mRNA of UCSC Table Track, protein information from UniProt, and the mitochondrial biogenesis and function criteria of MitoDat (Mendelian Inheritance and the Mitochondrion). Because the mitochondrion has a central role in cellular metabolism, the mitochondrion is involved in many human diseases [16]. We integrated gene and protein data into the SNP and diseases resources based on a gene- synonym table from HGNC and gene information at NCBI. Next, we mapped UniProt proteins onto NCBI genes by BLAST search. Finally, we added mitochondrial gene and protein data from MitoDat, because this database predominantly contains information on human nuclear-encoded mitochondrial proteins.

Integration of genes, SNPs, and diseases

To construct SNP-related information, we collected representative genetic variation (SNP), resources from dbSNP, JSNP, ALFRED, HGVbase, POLYPHEN (Polymorphism Phenotyping) [21], and SIFT (Sorting Intolerant From Tolerant) [22]. Finally, we integrated the information to show the interrelationships among SNPs located in genes, genes associated with diseases, and SNPs associated with diseases. The HGVbase database was adapted to integrate a curated resource describing human DNA variation and phenotype relationships. To predict a possible SNP impact of amino acid substitution on protein structure and function, we also linked to PolyPhen and SIFT. ALFRED contains data on allele frequencies at particular SNP loci for diverse populations, with reference to SNPs in dbSNP and JSNP. Our system can update primary databases automatically in real time. However, to integrate the various databases, we need to note the variations and partially update manually.

Next, we analyzed the influence of SNP location (e.g., CDS, UTR, Intron, or Promoter) on gene structure, and the effects of synonymous or non-synonymous SNPs on genetic variation and genes associated with disease, employing BLAST [23]. We explored amino acid changes caused by codon changes, and identified the locations of the altered amino acids in proteins. In addition, we determined whether SNPs were synonymous or non-synonymous, and identified the relationships of SNPs to candidate disease-causing genes.

Web Interface

The database server was implemented in JAVA and Java Server Pages connected to MySQL. The main web interface provides two ways to explore integrated disease-related information through a tree view of disease terms and through query searching.

Users can look over all the disease terms together in the tree view. When users click on a disease term, they can obtain results consisting of targeted disease information (disease name, synonyms, and title), gene information, and SNP information directly. The web interface also allows querying with three kinds of terms: (1) a SNP identifier (rs number from dbSNP), (2) a gene ID (symbol and description), or, (3) a disease term. When the user submits a gene, the system will return a list of genes related to the query along with gene aliases and types. When selecting a specific gene, query results show the gene, gene transcripts, SNPs, and genetic variants through the genome browser [24]. The genome browser provides insights into the effects of genetic variations on transcripts. Transcripts and SNP marker information displayed in the genome browser facilitates the recognition of characteristics of disease-causing genes, especially if the SNP or genetic variation lies in the promoter region or in an intronic sequence [25]. A display of the gene, transcripts, SNPs, and disease information is shown in Figure 2.

To show a disease search, we present query results for diabetes. The results from the integrated disease- and genetic variation-related databases are more helpful to researchers than results from one database only. It provides more comprehensive information on the genes and SNP markers associated with the disease. For example, when using only one disease-related database for diabetes, researchers can obtain either disease-association study information from GAD or information on disease-related literatures from OMIM. Conversely, when using the integrated database-pipeline system, we obtain a list of genes associated with diabetes, and an SNP marker (rs1805097) associated with diabetes by making both the integrated disease and the genetic variation information available simultaneously. This integrated

information allows researchers to consider the SNP effects on the gene along with relationships between SNPs and disease. The SNP marker (rs1805097) is located in the human insulin receptor substrate-2 (IRS-2) gene, which is a primary progesterone response gene. This SNP can affect amino acid change (GLY1057ASP), which has the possible impact of an amino acid substitution on the structure and function of a human protein [26]. Because this also includes the genome locations of disease-associated genes, effects of non-synonymous SNPs at the protein level, and disease-causing risk scores, users can expect to have a better understanding of the molecular causes of the disease.

Conclusions and Future Direction

We constructed the integrated database for the study of genetic variation in disease, using an automatic integration pipeline system. Specifically, the database contains information on 124,389 disease, 12,445,925 SNP markers, and 38,597 genes, and includes 14,674 SNP records and 109,715 gene records associated with human diseases. A total of 1,319 SNPs cause amino acid changes, inevitably leading to severe disruptions of protein structure or function.

Consequently, the integrated database-pipeline system can be an indispensable resource. The system can economically facilitate disease-association studies by identifying candidate genes associated with disease, and genetic variation. It can aid the understanding of the genes which cause diseases and the impact of SNPs on diseases, by showing the relationships among genes, SNPs and diseases. The tool uses unified disease terms, which facilitates the outreach and extension of this database to various other medical sources. As the resources in this database-pipeline system are

expanding continuously, we are planning to collect validated resources used in the detection of genetic variation for comparative studies.

Authors' contributions

JOY wrote the manuscript and helped to update the website. SH also wrote the manuscript and designed the databases. JSO developed the website, and updates the system. JB directed this project and helped to draft the manuscript. TKS designed the database and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank our colleagues at KOBIC, especially Areum Han, Jung-Sun Park, and Woo-Yeon Kim. The system was co-developed as a part of Diseasome pipeline by E-Gitec Inc. This work was supported by a grant from KRIBB Research Initiative Program, and the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No.M10869030002-08N6903-00210).

References

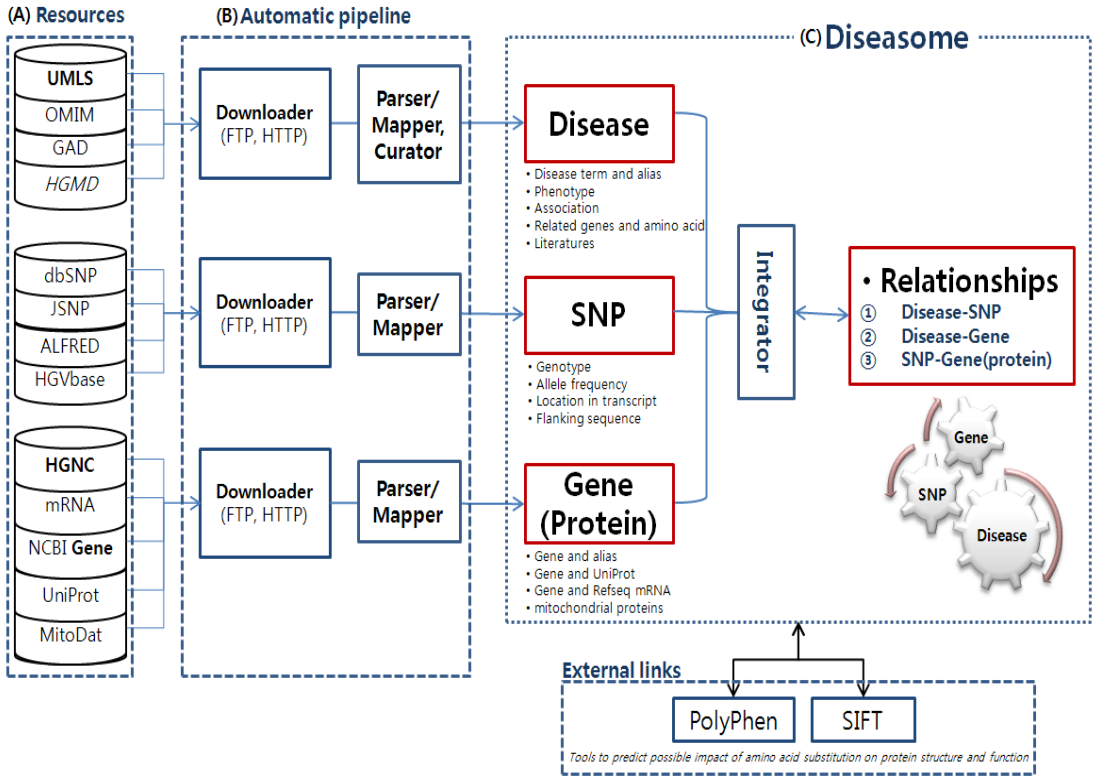
1. Matarin M, Simon-Sanchez J, Fung HC, Scholz S, Gibbs JR, Hernandez DG, Crews C, Britton A, Wavrant De Vrieze F, Brott TG *et al*: **Structural genomic variation in ischemic stroke**. *Neurogenetics* 2008, **9**(2):101-108.
2. Bae JS, Cheong HS, Kim JO, Lee SO, Kim EM, Lee HW, Kim S, Kim JW, Cui T, Inoue I *et al*: **Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population**. *Biochem. Biophys. Res. Commun.* 2008.
3. Lee JA, Lupski JR: **Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders**. *Neuron* 2006, **52**(1):103-121.
4. Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J: **SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions**. *BMC Bioinformatics* 2008, **9 Suppl 1**:S2.
5. Han A, Kang HJ, Cho Y, Lee S, Kim YJ, Gong S: **SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences**. *Nucleic Acids Res.* 2006, **34**(Web Server issue):W642-644.

6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
7. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nat. Genet.* 2003, **33 Suppl**:228-237.
8. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology**. *Nucleic Acids Res.* 2004, **32**(Database issue):D267-270.
9. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res.* 2005, **33**(Database issue):D514-517.
10. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nat. Genet.* 2004, **36**(5):431-432.
11. Cooper DN, Ball EV, Krawczak M: **The human gene mutation database**. *Nucleic Acids Res.* 1998, **26**(1):285-287.
12. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2008, **36**(Database issue):D13-21.
13. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ: **The HUGO Gene Nomenclature Database, 2006 updates**. *Nucleic Acids Res.* 2006, **34**(Database issue):D319-321.
14. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase**. *Methods Mol. Biol.* 2007, **406**:89-112.
15. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A *et al*: **The UCSC genome browser database: update 2007**. *Nucleic Acids Res.* 2007, **35**(Database issue):D668-673.
16. Lemkin PF, Chipperfield M, Merril C, Zullo S: **A World Wide Web (WWW) server database engine for an organelle database, MitoDat**. *Electrophoresis* 1996, **17**(3):566-572.
17. Smigielski EM, Sirotkin K, Ward M, Sherry ST: **dbSNP: a database of single nucleotide polymorphisms**. *Nucleic Acids Res.* 2000, **28**(1):352-355.
18. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y: **JSNP: a database of common gene variations in the Japanese population**. *Nucleic Acids Res.* 2002, **30**(1):158-162.
19. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP *et al*: **ALFRED: the ALelle FREquency Database. Update**. *Nucleic Acids Res.* 2003, **31**(1):270-271.
20. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ: **HGVbase: a curated resource describing human DNA variation and phenotype relationships**. *Nucleic Acids Res.* 2004, **32**(Database issue):D516-519.
21. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y: **Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines**. *BMC Bioinformatics* 2007, **8**:450.

22. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res.* 2003, **31**(13):3812-3814.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J. Mol. Biol.* 1990, **215**(3):403-410.
24. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: **The Generic Genome Browser: A Building Block for a Model Organism System Database**. *Genome Res.* 2002, **12**:1599-1610.
25. Abu A, Frydman M, Marek D, Pras E, Stolovitch C, Aviram-Goldring A, Rienstein S, Reznik-Wolf H, Pras E: **Mapping of a gene causing brittle cornea syndrome in Tunisian jews to 16q24**. *Investigative ophthalmology & visual science* 2006, **47**(12):5283-5287.
26. Stefan N, Kovacs P, Stumvoll M, Hanson RL, Lehn-Stefan A, Permana PA, Baier LJ, Tataranni PA, Silver K, Bogardus C: **Metabolic effects of the Gly1057Asp polymorphism in IRS-2 and interactions with obesity**. *Diabetes* 2003, **52**(6):1544-1550.

Figures

Figure 1 - Overview of the integrated database-pipeline system

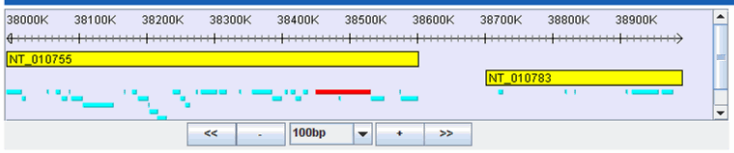


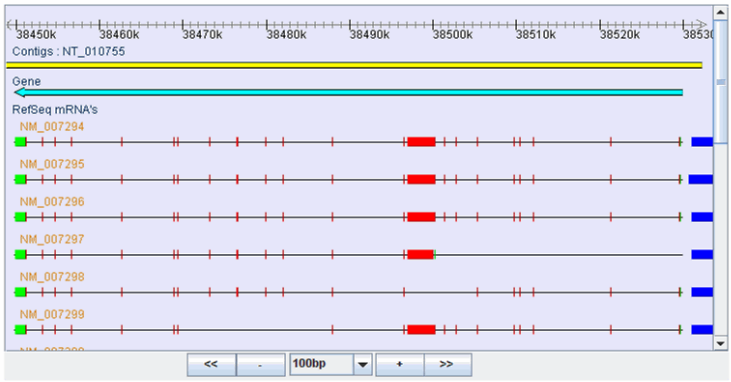
Rectangles represent computational applications, and are three in number. The Resource (A) contains gene-, SNP-, and disease-related primary resources and constructs a primary information database. The Automatic pipeline (B) retrieves

information from primary databases and extracts essential gene-, SNP-, and disease-related data. We mapped disease terms and aliases, or gene names and aliases, based on the UMLS and HGNC databases. Also, disease terms were corrected for noun modification, stop word, and suffix. SNP effects were investigated by amino acid substitution; locations are available. The Diseasome (C) is a database including three categories of information (gene, SNP, and disease), and relationships among the three categories.

Figure 2 - Query table results and graphic viewer

Genome Browser





Gene Information

Gene Symbol	BRCA1
Gene Aliases	BRCAI, BRCC1, IRIS, PSCP, RNF53
Gene Name	breast cancer 1, early onset
Gene ID	672
Cytogenetic Band	17q21
Gene Type	protein-coding
mitoDat ID	-

Disease Information

Disease Hit	9
Disease Name	UGATZ TUMORE GAIZTOA BRCA 1 Syndrome Ovarian Carcinoma Malignant Neoplasms Endometrial Carcinoma Colorectal Cancer Germ-Line Mutation Ondartet svulst i prostata Fallopian Tube Carcinoma
Others	no cui diseases

Transcription Information

No	mRNA Accession	Chromosome Position	Strand	Exon Count	Promoter SNP Count	5'UTR SNP Count	CDS SNP Count	3'UTR SNP Count	Intron SNP Count
1	NM_007294	chr17:38449840~38530994	-	23	22	3	71	11	440
2	NM_007295	chr17:38449840~38530657	-	23	23	1	71	11	438
3	NM_007296	chr17:38449840~38530994	-	23	22	3	71	11	440
4	NM_007297	chr17:38449840~38530994	-	15	22	6	62	11	446
5	NM_007298	chr17:38449840~38530994	-	20	22	3	29	11	482
6	NM_007299	chr17:38449840~38530994	-	19	22	3	59	11	452
7	NM_007300	chr17:38449840~38530994	-	18	22	3	58	11	453
8	NM_007302	chr17:38449840~38530994	-	21	22	3	70	11	441
9	NM_007303	chr17:38449840~38530994	-	22	22	3	30	11	481
10	NM_007304	chr17:38449840~38530994	-	23	22	3	32	11	479
11	NM_007305	chr17:38449840~38530994	-	21	22	3	31	11	480

SNP Information

SNP ID	Chromosome	Chromosome Position	Strand	Allele	Fuction Class
rs799905	17	38530713	+	C/G	Intron, Promoter
rs34191881	17	38530897,,38530898	+	-/A	5UTR, Promoter
rs35436937	17	38530928,,38530929	+	-/T	5UTR, Promoter
rs8176075	17	38530940	-	-/T	5UTR, Promoter
rs8176074	17	38531291	-	A/G	Promoter
rs8176073	17	38531359	-	A/G	Promoter
rs8176072	17	38531470	-	A/T	Promoter
rs3092986	17	38531522	-	A/G	Promoter
rs34085552	17	38531530,,38531531	+	-/GT	Promoter
rs8176071	17	38531531,,38531532	-	-/ACA	Promoter
rs799906	17	38531642	+	C/T	Promoter
rs11655505	17	38531903	+	A/G	Promoter
rs799907	17	38532251	+	C/G	Promoter
rs799908	17	38532442	+	A/G	Promoter
rs799909	17	38532753	+	A/G	Promoter

The retrieval page of the integrated gene, SNP, and diseases database. The information on diseases, genes, and SNP markers found as result of a query (e.g., BRCA1) are shown. When a user queries a gene symbol, the system retrieves the Gene Information table, which shows various gene annotations, disease information related to the queried gene, transcript information including the number of SNPs located in each transcript, and SNP information associated with the queried gene. In addition, the user can explore the data on gene-related transcripts, SNPs, and disease information, using the genome browser**.** If a user requires more specific information on any item, the user can click on a disease term, a gene ID, or a genetic variation number (SNP rs number).