Genomics & Informatics Vol. 3(3) 68-73, September 2005

Development of a Knowledge Base for Korean Pharmacogenomics Research Network

Chan Hee Park, Su Yeon Lee, Yong Jung, Yu Rang Park, Hye Won Lee, and Ju Han Kim*

Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

Pharmacogenomics research requires an intelligent integration of large-scale genomic and clinical data with public and private knowledge resources. We developed a web-based knowledge base for KPRN (Korea Pharmacogenomics Research Network, http://kprn.snubi.org/). Four major types of information is integrated; genetic variation, drug information, disease information, and literature annotation. Eighteen Korean pharmacogenomics research groups in collaboration have submitted 859 genotype data sets for 91 disease-related genes. Integrative analysis and visualization of the large collection of data supported by integrated biomedical pathways and ontology resources are provided with a user-friendly interface and visualization engine empowered by Generic Genome Browser.

Keywords: Pharmacogenomics, Alleles, MeSH, PharmGKB, Goldenpath, Genew, Generic Genome Browser

Introduction

Human Genome map is nearly completed. Pharmacogenetics and Pharmacogenomics which deal with the genetic or genomic basis underlying variable drug responses in individuals have become the center of interest all over the world. In this situation, it is recognized that pharmacogenomic and pharmacogenetic studies form a field of study that can be applied in clinical medicine at the earliest time. And many biomedicalogists are interested in this field because they believe that it can make individualized therapy possible.

It is expected that a technique of individual therapy in genetic basis will be put to practical use in patient treatment within 10 years. Experts in this field have tried to apply the technique to a part of treatment with drugs although there still remains limitation.

At present, there is a centralized database, Pharm GKB(Klein *et al.*, 2001) (Pharmacogenomics Knowledge Base), where researchers store the outcome of the pharmacogenomics study world wide. PharmGKB provides the functions in which we can collect and search for the result of study about what drug response occurs in individuals according to genetic variation. The standard of data model for pharmacogenomics data, however, is not present, yet. Moreover, it is hard to find a database except PharmGKB which collects pharmacogenomics data. As a group of large research collaboration, KPRN felt the need of having its own repository system with integrated biomedical resources. The present study intends to deal with pharmacogenomics data repository system being developed in KPRN. Its goals include centralization of the domestic pharmacogenomics data, preventing duplication of similiar research, and information-sharing among the joint research team as well as with the general research community. It intends to provide genomic function information about sequence variance as well as expression, extending it to be knowledge-based.

Method

Referenced data schema, database, and tool

1. PharmGKB XML schema

PharmGKB is a Knowledge-base developed at the Stanford University for the function of searching and retrieving the result of research in pharmacogenomics (http://www.pharmgkb.org/). Its XML schema explains information about sequence variation, drug data, and phenotype data. It describes the information of side effects from a drug treatment that are caused by sequence variation. One can find the details of the schema at http://www.pharmgkb.org/schema/index.html/.

2. Aperon drug database

We use the drug database made in Aperon(Kaiser, 2005) to capture the name of drug. Presently, Aperon is also used by PharmGKB. It catalogs 3,885 drug names, 9,176 generic names, and 9,177 trade name.

*Corresponding author: E-mail juhan@snu.ac.kr, Tel +82-2-740-8320, Fax +82-2-747-4830

Accepted 5 September 2005

3. Medical Subjects Heading (MeSH)

MeSHMeSH(Schulman *et al.,* 2000), a thesaurus made by NLM (National Library of Medicine), is used for indexing an article in MEDLINE. It provides a consistent way for searching the information that includes different terms for the same concepts. For instance, a malignant tumor can be ordinarily named either 'neoplasm' or 'cancer'. In MeSH, however, neoplasm is adopted. Also, MeSH terms old people 'aged'.

4. USCS GoldenPat

Our knowledge base uses the GoldenPath(Kent *et al.,* 2005; Karolchik *et al.,* 2003) developed at the University of California to capture the position where a sequence variation occurred. GoldenPath enables the use of absolute coordinate system for a position of variation on Human Chromosome Map, allowing a consistent positioning of variations and annotation on the basis of the position of variation.

ChromosomeNumber : BasePosition

For example, 'chr2:23451703' indicates the 23451703rd position on the 2nd chromosome.

5. Genew (Human Gene Nomenclature Database)

We use the gene name provided by Genew(Wain *et al.,* 2002) to associate it with the information of a sequence variation. Genew is the precisely verified database by editors in HGNC (Human Gene Nomenclature Committee) for denominating name and symbol of human gene. This database includes gene name, gene symbol, and alias symbol.

6. Generic Genome Browser (GGB)

We use GGB(Stein *et al.,* 2002) (Generic Genome Browser) to display information of both sequence variance and Human genome which are collected. It is possible for GGB to show sequence-based information such as *genome contig, genome information, STS* (Sequence Tag Site), SNP (Single Nucleotide Polymorphism), so on and so forth by track. And also the GGB is designed to be able to be linked with other databases as each of track information. We adopted this browser to indicate the information of allele frequency entered by users.

Result

General overview

To store and search pharmacogenomics data, we defined and modeled the relationship between sequence variance information, drug information, and phenotype information that describe pharmacokinetic or pharmaco-

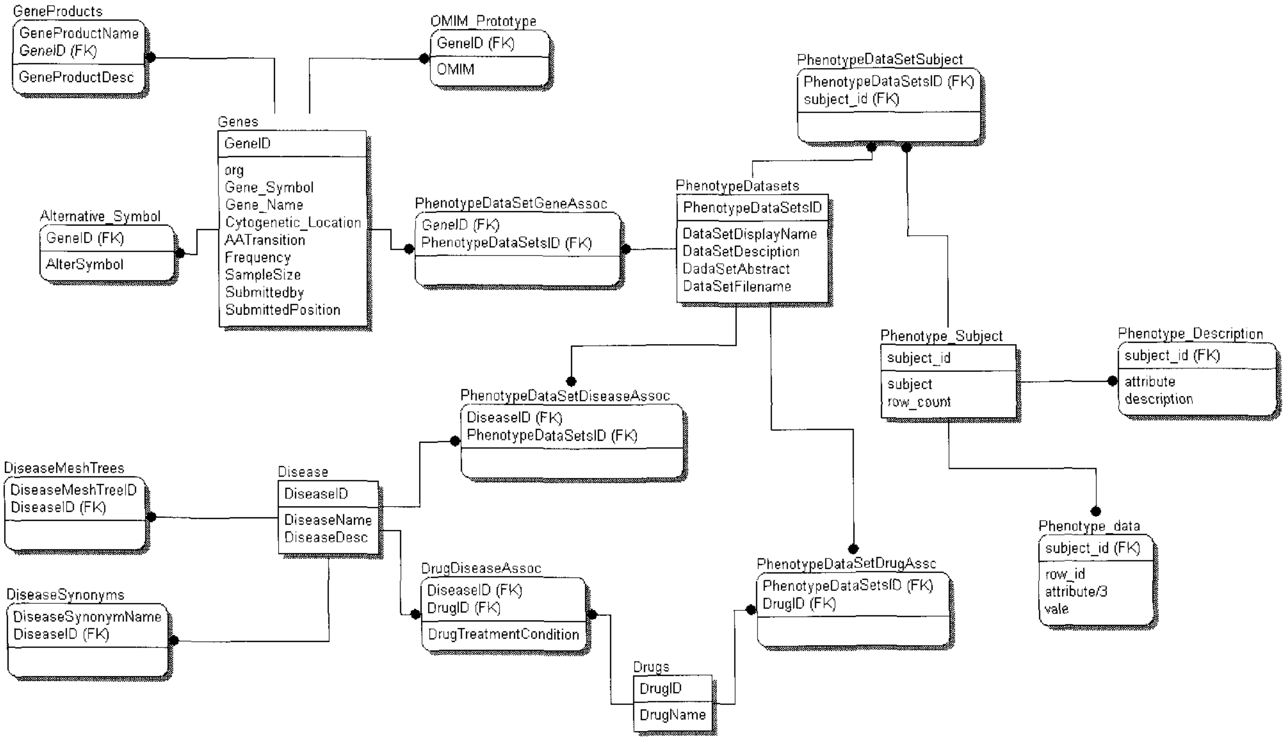
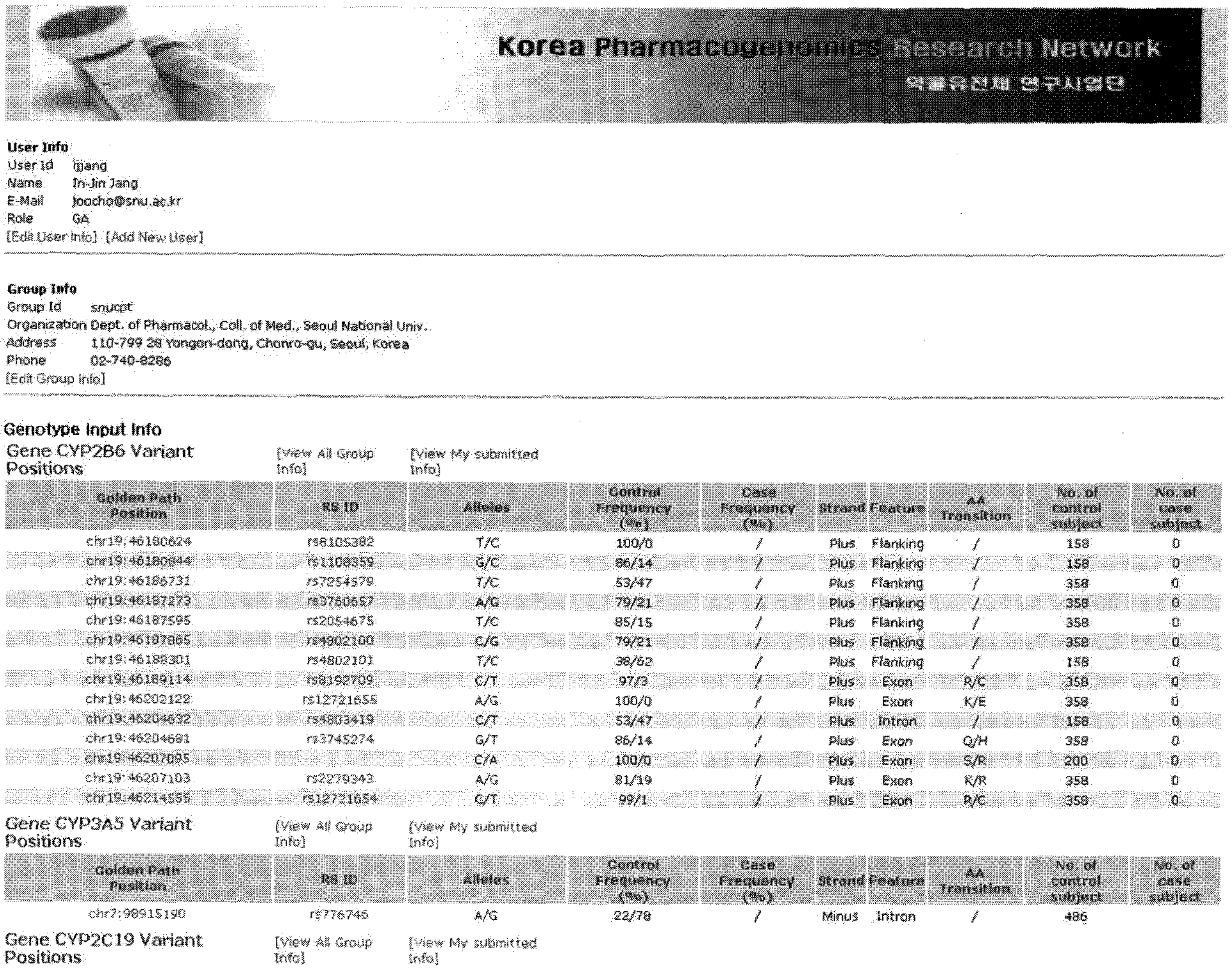


Fig. 1. Relationships among tables Databases of KPRN knowledge base



User Info
User Id ijjang
Name In-Jin Jang
E-Mail joocho@snu.ac.kr
Role GA
[Edit User Info] [Add New User]

Group Info
Group Id snucpt
Organization Dept. of Pharmacol., Coll. of Med., Seoul National Univ.
Address 110-799 28 Yongon-dong, Chonro-gu, Seoul, Korea
Phone 02-740-8286
[Edit Group Info]

Genotype Input Info

Gene CYP2B6 Variant Positions [View All Group Info] [View My submitted Info]

Golden Path Position	RS ID	Alleles	Control Frequency (%)	Case Frequency (%)	Strand	Feature	AA Transition	No. of control subject	No. of case subject
chr19:46180624	rs8105382	T/C	100/0	/	Plus	Flanking	/	158	0
chr19:46180844	rs1108359	G/C	86/14	/	Plus	Flanking	/	158	0
chr19:46186731	rs7254579	T/C	53/47	/	Plus	Flanking	/	358	0
chr19:46187273	rs3760657	A/G	79/21	/	Plus	Flanking	/	358	0
chr19:46187595	rs2054675	T/C	85/15	/	Plus	Flanking	/	358	0
chr19:46187865	rs4802100	C/G	79/21	/	Plus	Flanking	/	358	0
chr19:46188301	rs4802101	T/C	38/62	/	Plus	Flanking	/	158	0
chr19:46189114	rs8192709	C/T	97/3	/	Plus	Exon	R/C	358	0
chr19:46202122	rs12721655	A/G	100/0	/	Plus	Exon	K/E	358	0
chr19:46204632	rs4803419	C/T	53/47	/	Plus	Intron	/	158	0
chr19:46204681	rs3745274	G/T	86/14	/	Plus	Exon	Q/H	358	0
chr19:46207095		C/A	100/0	/	Plus	Exon	S/R	200	0
chr19:46207103	rs2279343	A/G	81/19	/	Plus	Exon	K/R	358	0
chr19:46214556	rs12721654	C/T	99/1	/	Plus	Exon	R/C	358	0

Gene CYP3A5 Variant Positions [View All Group Info] [View My submitted Info]

Golden Path Position	RS ID	Alleles	Control Frequency (%)	Case Frequency (%)	Strand	Feature	AA Transition	No. of control subject	No. of case subject
chr7:98915190	rs776746	A/G	22/78	/	Minus	Intron	/	486	

Gene CYP2C19 Variant Positions [View All Group Info] [View My submitted Info]

Fig. 2. Web form for input data

dynamic effect which is different in sequence variance for dosage. For the standardization of entering information for sequence variance, we use the Gene Symbol provided by Genew. We accept MeSH terms as the standard for disease names in phenotype data. The drug names are taken from Aperon database. The relationship of each of information is illustrated in Fig. 1. In order to add literature annotation information, we have developed a system for inputting highly curated annotation of each gene, drug, and phenotype from documents published already. Fig. 2 shows such a form through which users can input data into the system. We accomplish the visualization of information of sequence variance by integrating the functionalities supported by the Generic Genome Browser.

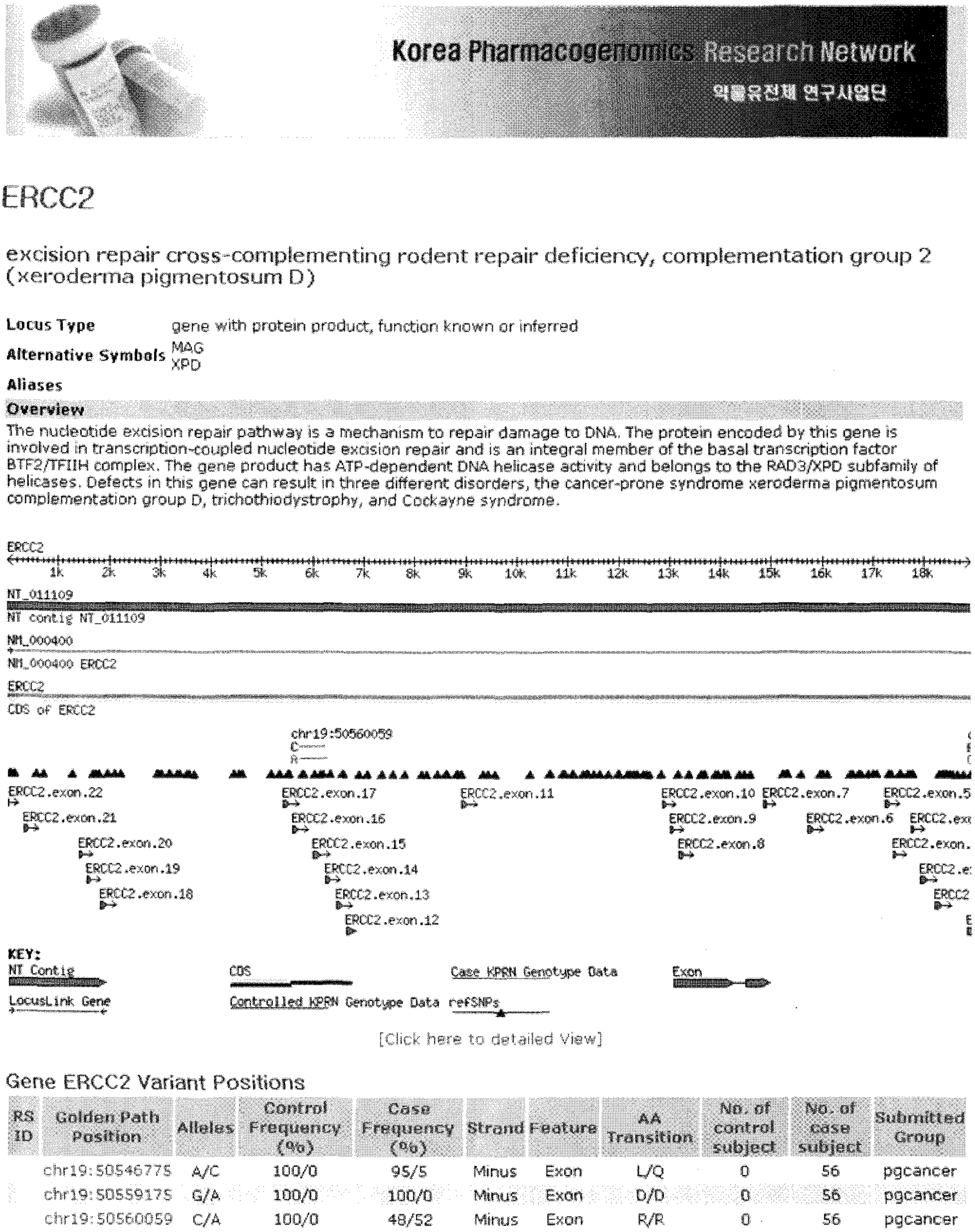
Detail explanation of the Model

1) Information of Sequence variance

This part describes information of experiment about sequence information genotyped (i.e. allele position, alleles, allele frequency, and amino acid transition and Gene symbol). Actually, the marking rule in indicating allele position exists already. For example, "CYP2C9, Arg144Cys" means that amino acid transition occurs at the 144th position of CYP2C9 gene, changing Arg to Cys. But the result may vary depending where the experimenter count as the 1st position of the gene. Thus we adopt an absolute coordinate system by using GoldenPath position. It prevents ambiguity which is caused by using relative position information and we need it to combine our database with the others. To mark alleles, we adopted the format described below.

The notation of amino acid transition is designed to understand one or three-letter codes. There are two categories in allele frequency, control frequency and case frequency. The former corresponds to searching for frequency in healthy volunteers, the latter in patients.

We used the Genew system to check what gene the



ERCC2

excision repair cross-complementing rodent repair deficiency, complementation group 2 (xeroderma pigmentosum D)

Locus Type gene with protein product, function known or inferred

Alternative Symbols MAG
XPD

Aliases

Overview

The nucleotide excision repair pathway is a mechanism to repair damage to DNA. The protein encoded by this gene is involved in transcription-coupled nucleotide excision repair and is an integral member of the basal transcription factor BTF2/TFIIH complex. The gene product has ATP-dependent DNA helicase activity and belongs to the RAD3/XPD subfamily of helicases. Defects in this gene can result in three different disorders, the cancer-prone syndrome xeroderma pigmentosum complementation group D, trichothiodystrophy, and Cockayne syndrome.

[Click here to detailed View]

Gene ERCC2 Variant Positions

RS ID	Golden Path Position	Alleles	Control Frequency (%)	Case Frequency (%)	Strand	Feature	AA Transition	No. of control subject	No. of case subject	Submitted Group
	chr19:50546775	A/C	100/0	95/5	Minus	Exon	L/Q	0	56	pgcancer
	chr19:50559175	G/A	100/0	100/0	Minus	Exon	D/D	0	56	pgcancer
	chr19:50560059	C/A	100/0	48/52	Minus	Exon	R/R	0	56	pgcancer

Fig. 3. Visualization of sequence information using Generic Genome Browser

Table 1. Marking Rule of Allele

Mark	Attribute
A/-	Deletion of A
A/G	Single nucleotide change
T/TCGGT	CGGT insertion
G/TA(3)	TA repeats 3 times instead of G

Table 2. The Category of Phenotype data

Abbreviation	Attribute
CO	Clinical Outcome
FA	Molecular & Cellular Functional Assays
GE	Genotype
PD	Pharmacodynamics & Drug Response
PK	Pharmacokinetics

information of a sequence variation be included by. For a gene, more than one symbols may exist. In case of an alpha-1-B glycoprotein, alias symbols are A1B, ABG, GAB, HTSY2477 and so on. So, we need an official gene symbol that represents them for preventing ambiguity caused by redundancy. We use 23,449 official symbols now.

2) Drug information

We use the drug database made in Aperon to input the name of drug. Presently, it is also used by PharmGKB. The Aperon drug database consists of 3,885 drug names, 9,176 generic names, and 9,177 trade names. When users search or retrieve drug name, generic name, and trade name simultaneously, it gives proper drug name to users to use.

3) Phenotype information

The phenotype that the gene in point shows when a drug is administered into normal group/patient group is described in phenotype information. We use the MeSH term to describe disease name used for input and classify phenotype into five categories used in PharmGKB.

This system uses EAV (Entity-Attribute-Value) model(Anjoy *et al.*, 2003) because each phenotype information is input in each category but cannot be input as unified or abbreviated form.

4) Literature Annotation

We developed a Literature Annotation system to search for the information of pharmacogenomics that is published already because we intend to get literature information which is highly curated by researchers. The Literature Annotation consists of PubMed information, keyword, abstract, related gene, drug, and disease information that can be input and retrieved.

5) Visualization using Generic Genome Browser

We used Generic Genome Browser (GGB) to visualize information of sequence variance. It is possible for GGB to show sequence-based information by tracks. Also it is designed to be able to be linked with other databases as each of track information. We modified this browser to indicate the information of allele frequency entered by users (Fig. 3).

Discussion

This paper introduces the knowledge base for KPRN we have developed for storing and searching for pharmacogenomics data with integrated biomedical resources. The system includes sequence information from pharmacogenomic experiment, drug data, phenotype data coincided with the drug data, literature annotation, and visualization of sequence variance information. We used thesaurus in MeSH, Genew and many standard systems for building the knowledge base to describe gene information, drug information, phenotype information. The system enables to link with other public databases for input information and helps understand pharmacogenomics data by means of visualization. At present, 859 genotype data sets are integrated for 91 genes by the 18 research groups. The number of input data will exponentially increase in the near future.

Except for the PharmGKB in Stanford, there has been no standard for storing pharmacogenomics data. But we think that ours is a valuable effort to create our own knowledge base for understanding and using pharmacogenomics data. We hope it will server for the development of a standard for collaborative and integrative pharmacogenomics research.

Acknowledgments

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (03-PJ10-PG13-GD01-0002)

References

Klein, T.E., Chang, J.T., Cho, M.K., Easton, K.L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D.E., Rubin, D.L., Shafa, F., Stuart, J.M., and Altman, R.B. (2001). "Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project", *The Pharmacogenomics J.* 1, 167-170.

Kaiser, J. (2005).Aperon Biosystems. *http://www.aperon.com/index.htm*

Schulman, J.L. (2000). Using Medical Subject Headings (MeSH) to examine patterns in American medicine. [course paper, STS 5206]. Falls Church: Virginia

Polytechnic Institute and State University, Northern Virginia Center

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. *Genome Res. 12*, 996-1006.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., and Kent, W.J. (2003).The UCSC Genome Browser Database. *Nucleic Acids Res.* 31. 51-54.

Wain, H.M., Lush, M., Ducluzeau, F., and Povey, S. (2002). Genew: the Human Gene Nomenclature Database. *Nucleic Acids Res.* 30, 1169-1171

Stein , L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S.. (2002).The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Res.* 12, 1599-1610

Anhoj, J. (2003). Generic Design of Web-Based Clinical Databases. *J. Med. Internet Res.* 5, 27