FACILITATING PHARMACOGENETIC ASSOCIATION STUDIES USING AN EXTENSIBLE GENOTYPE INFORMATION MANAGEMENT SYSTEM

Rebecca Fletcher

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Bioinformatics
Indiana University

May 2008

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

Master's Thesis
Committee

____
Sean D. Mooney, PhD, Chair

____
Narayanan Perumal, PhD

____
David A. Flockhart, MD, PhD

Dedicated to my parents.

ACKNOWLEDGEMENTS

I would like to extend a sincere wish of gratitude to my family and friends for their support and patience.

I would especially like to thank Dr. Sean Mooney, Dr. Narayanan Perumal and Dr. David Flockhart for serving on my thesis advisory committee.

This project is due to the work of many talented individuals including:

Brandon Peters
Todd Skaar
Santosh Philips
Anne Nguyen
Jason Robarge
Lang Li
Jonathan Nowacki
Kirthi Krishnamraju
Mark Crowder
Divya Rao

I would like to express my thanks to Dr. Yaoqi Zhou for his kindness in serving as a proxy on my thesis committee during my defense.

This research is supported by NIGMS grant UO1-GM61373-05, Tamoxifen Pharmacogenetics and Clinical Effects, (Principle Investigator: David Flockhart, MD, PhD), NLM grant 5-K22-LM009135-02, Bioinformatics approaches to characterizing amino acid function, (Principle Investigator: Sean Mooney, PhD) and funds provided through an INGEN grant. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment.

ABSTRACT

Rebecca Fletcher

FACILITATING PHARMACOGENETIC ASSOCIATION STUDIES USING AN EXTENSIBLE GENOTYPE INFORMATION MANAGEMENT SYSTEM

Large-scale genome data projects employing automated, high-throughput techniques have led to a deluge of genomic data that necessitate robust informatic solutions. COBRA-DB is an integrated web-based genome information management system that provides storage for pharmacogenomic information including genotypic, phenotypic and resequencing data. The system provides an integrated solution for the acquisition, organization, storage, retrieval and analysis of pharmacogenomic data and offers a platform for genome annotation and analysis. The system also includes an export utility to automate submission of data to other bioinformatic resources and public data repositories. A web interface provides flexible data import and export options and allows users to access and download data via simple query forms. The COBRA database is dedicated to the efficient management of pharmacogenomic data with the intent to facilitate genotype-phenotype association studies and catalyze pharmacogenomic research. COBRA-DB is an internal, proprietary application in use by the Division of Clinical Pharmacology at Indiana University School of Medicine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One: **Introduction**

In the post-genomic era, large-scale scientific projects have led to a new scientific emphasis on genetic variation and related phenotypes. Focusing on the associations between sequence variation and resulting traits has in turn driven the next generation of database technologies (Frenkel 1991). In 1990 the U.S. Department of Energy and the National Institutes of Health formally launched the Human Genome Project (HGP) to identify all the genes in the human genome and to determine the specific sequence of nucleotides that comprise human DNA. This thirteen year long project represented an unprecedented scientific undertaking which helped spawn a new era of innovation not only in medicine but also in technology. Processing data in the post-genomic era presents new challenges which necessitate the engineering of novel information systems. Unlike traditional commercial or engineering datasets that were smaller, more static and less complex, genetic data are large-scale, more fluid in nature and multifaceted. As a result, investigators require robust information systems that can handle the inherent complexities of genomic data.

Genotype/Phenotype Databases

High-throughput genotyping and DNA sequencing technologies present researchers with unprecedented opportunities for performing genome analysis. However, these technical advances have made it increasingly more difficult to manage the resulting deluge of information. Public repositories for phenotype, genotype and sequence data are being developed in an effort to help manage the burgeoning data. These so-called genotype/phenotype databases store data at the level of individuals and are functional resources for investigators who are interested in studying genetic variation and associations with observable traits.

In general, phenotype is determined not only by genotype but by environmental influences, as well. For example, genotypes will not result in equal phenotypes in all environments. Instead, a genotype may produce a favorable phenotype under a certain set of conditions yet produce a negative outcome under a different set of variables. This is known as phenotypic plasticity (Price *et al*, 2003). Therefore, the interaction between genetics and environmental factors is highly significant. Unfortunately, however, accounting for environmental factors is complicated. Often times, environmental factors are difficult to measure and even when they are easily registered, recording them is another obstacle. Although it is more appropriate to describe phenotype as a function of genotype *and* the environment, with the exception of a small number of self-reported facts (*i.e.*, concomitant medications), limited environmental information is available. As a result, for the purpose of this project, environmental factors will be included where available but do not represent a major contribution to the results.

Pharmacogenomic Association Studies

Pharmacogenomics is the intersection of pharmacology and genomics and studies how an individual's genetic variation affects drug response. The discipline combines pharmacy, biochemistry and molecular biology with annotation about the human genome and genetic variation to improve our understanding of how genetic inheritance affects disease.

Genetic association studies explore the connection between genetic composition or genotype and the outward manifestation of that genetic composition or phenotype. Genome-wide association studies (GWAS) examine genetic variation across the entire human genome. In Pharmacogenomics, GWA studies have made it possible to identify genetic factors that are associated with drug response.

COBRA and Breast Cancer Pharmacogenomics

This project focuses on the development of an information management system to support the Consortium on Breast Cancer Pharmacogenomics (COBRA). COBRA is a member of the Pharmacogenomics Research Network (PGRN), a collaborative group of investigators who focus on correlating drug response phenotypes with genetic variation. COBRA's mission is to study how multiple genetic variations affect clinical pharmacology in order to accelerate breast cancer research.

More specifically, COBRA is interested in understanding how genetic polymorphisms affect normal estrogen function and how these variations impact breast cancer treatments (Ntukidem *et al*, 2008). COBRA aims to identify genetic variants in the estrogen receptors (ER-α and ER-β) and drug metabolizing enzymes that are involved in aromatase inhibitor hormone therapy. COBRA also studies genetic variants and their associations with specific phenotypes such as hot flashes or bone density in breast cancer patients (Table 1).

Table 1 List of phenotypes studied by COBRA

- Breast
  - Mammography
- Bone
  - Densitometry
  - Bone turnover markers
- Quality of Life
  - Questionnaire name
  - Rheumatologic symptoms
- Endocrine
  - Estrogens
  - Androgens
  - Thyroid markers
- Hot Flash
  - Objective monitoring
  - Diaries
- Pharmacokinetics
  - Tamoxifen
  - Letrozole
  - Exemestane
  - Anastrozole
- Cardiovascular
  - Lipids
  - Platelets
  - Inflammatory markers

COBRA uses a combined bioinformatic and direct sequencing approach to test for variants in candidate genes. Once results from high throughput techniques are generated, researchers need a straightforward method to easily access the results.

Gap in Data Management

Managing data is a major challenge. In order to be useful, data must be prepared so that it is convenient and easy to access. When we lack the proper tools to efficiently cope with data, there is a severe risk that data will be ignored or lost. In a worst case scenario, a researcher might actually find it easier to reproduce results as opposed to simply querying a database to retrieve his data. We proposed the following integrated informatics solution in an effort to eliminate the risk to data integrity. Proper data management represents a critical component of successful research and the lack thereof is often a rate limiting step in research. Data should be centralized, secure, accessible, standardized, distributable and redundant.

Description and Scope of Project

The COBRA-DB warehouses data generated as part of the *Tamoxifen Pharmacogenetics and Clinical Effects*[1] trial sponsored by the National Institute of General Medical Sciences (NIGMS). The goal of the trial is to determine how breast cancer patients respond to the cancer drug tamoxifen that affects the activity of the female hormone estrogen (Goetz *et al*, 2005). The study will test the following hypotheses.

1. "There is a relationship between genetically distinct metabolic profiles of tamoxifen and the frequency and severity of hot flashes in women on chronic tamoxifen therapy.
2. Genetically distinct metabolic profiles for tamoxifen effect lipid profile, bone turnover metabolites and bone mineral density, and coagulation factors.
3. Different genetic profiles of estrogen responsive genes influence the pharmacodynamic effects of tamoxifen in cardiovascular system".

[1] Official Title: A Pilot Trial Correlating Metabolic Profile of Tamoxifen With Pharmacogenetic Predictors and Clinical Effects

COBRA aims to sequence the estrogen receptor gene in hundreds of breast cancer patients and describe the genetic variations. Researchers plan to genotype representative tagSNPs in each of the women.

In terms of scope, the ER-β gene is 50,000bp in length. A typical sequencing reaction is 250bp long, creating a total of 200 amplicons. There are 96 samples sequenced, which generates almost 20,000 reactions. For each reaction, we align the amplicon along the gene and search for SNPs. Since SNP density in the human genome is approximately one per 200bp, we expect roughly 250 SNPs for the ER-β gene. Finally we add these 250 SNPs for each of the 96 samples, creating almost 25,000 genotype entries in our system.

Project Goals and Objectives

A computer software information management system was designed to manage genetic data in a clinical setting. The software includes a data warehouse that functions as a central repository and staging platform to collect raw data which is then locally curated before being exported to public databases for storage or other external destinations such as bioinformatic web applications for further analysis. The project goals are to implement a web-based relational database application for storage and retrieval of genotypic, phenotypic and resequencing data by furnishing a system that delivers the following functionalities:

Acquisition, Organization, Storage:

1. Collect, store and annotate genotype (polymorphism) data
2. Assemble, deposit and annotate resequencing data
3. Acquire, organize and warehouse phenotype data

Retrieval and Analysis:

4. Automate submission of data to PharmGKB
5. Format and export data to other bioinformatic applications
6. Facilitate statistical analysis of data to elucidate genotype/phenotype associations.

Chapter Two: **Background**

Currently Available Public Databases

Genome-wide association studies that combine whole genome information with phenotype data to increase our understanding of human health and disease are being carried out at an unprecedented rate. Yet the number of genotype/phenotype databases dedicated to identifying genetic factors that influence disease are relatively few. However, several such repositories have already been established; PharmGKB (Klein *et al*, 2001), PhenomicDB (Kahraman *et al*, 2005) and dbGaP (Mailman *et al*, 2007) are three currently available public genotype/phenotype databases dedicated to advancing research in genetic associations.

The database of Genotype and Phenotype (dbGaP) was developed and is operated by the National Library of Medicine's National Center for Biotechnology Information (NCBI). The database collects research data from studies that investigate the relationship between genotype and phenotype, such as genome-wide association studies (GWAS). The database offers two levels of public access: open and controlled. Open access grants the ability to retrieve summaries or studies, study documents and other related information. Controlled access can be requested and includes access to individual level genotypes. The database also includes an analysis of statistical association between genes and phenotypes.

PhenomicDB is a multi-species genotype/phenotype database that is hosted by a German bioinformatics company, Metalife. It merges data from primary databases (*e.g.*, FlyBase, WormBase, NCBI, OMIM, etc.) and includes data on numerous organisms such as human, mouse, fruit fly and C. elegans. The database also includes orthologues to allow comparison of phenotypes across many species simultaneously. RNA interference (RNAi) screen data, phenotype ontology terms and assay information is also incorporated into PhenomicDB.

PharmGKB is a knowledge base managed by Stanford University that warehouses information on drugs, diseases, phenotypes, genes, pharmacokinetics and pharmacodynamics. It also integrates variant data from a number of public repositories including dbSNP, HapMap and jSNP. It is the central databank for the PGRN and also accepts data submissions from the public. Data about the relationships between drugs, genes and diseases are collected and curated with the intent to catalyze pharmacogenomic research.

Chapter Three: **Methodology**

COBRA-DB is an integrated online database system that manages three major datasets: genotype data, phenotype data and sequencing data. We developed a relational method to integrate genotype, phenotype and sequencing information. Figure 1 shows the integration of the genomic data.

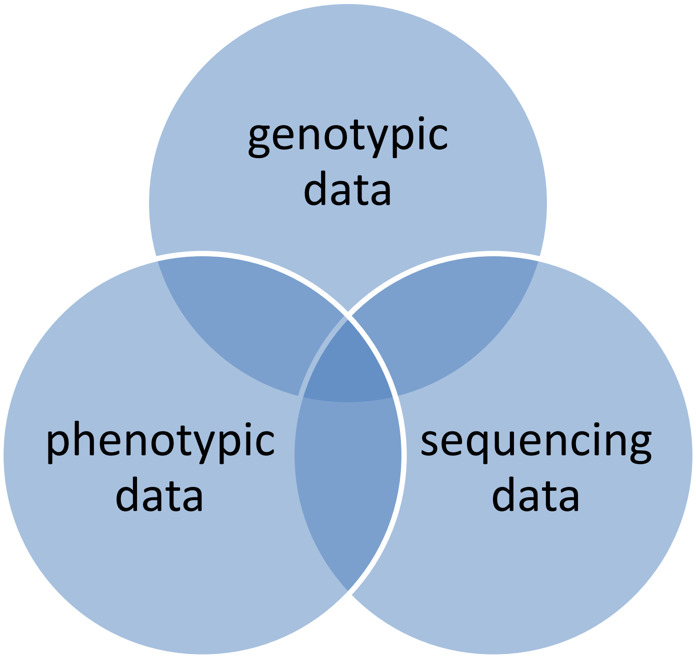


Figure 1 COBRA-DB integrates multiple datasets

COBRA Participants

The Consortium on Breast Cancer Pharmacogenomics (COBRA) is a member of the National Institutes of Health Pharmacogenomics Research Network (PGRN). The mission of the PGRN is to advance understanding of the genetic basis for variable drug responses. COBRA aims to correlate genetic variation with drug response phenotypes. The COBRA Research Network consists of the following academic institutions: Indiana University, University of Michigan, Johns Hopkins University, Baylor College of

Medicine and Mayo Clinic. Members of the PGRN, including COBRA, submit data to the Pharmacogenomics and Pharmacogenetics Knowledgebase (PharmGKB).

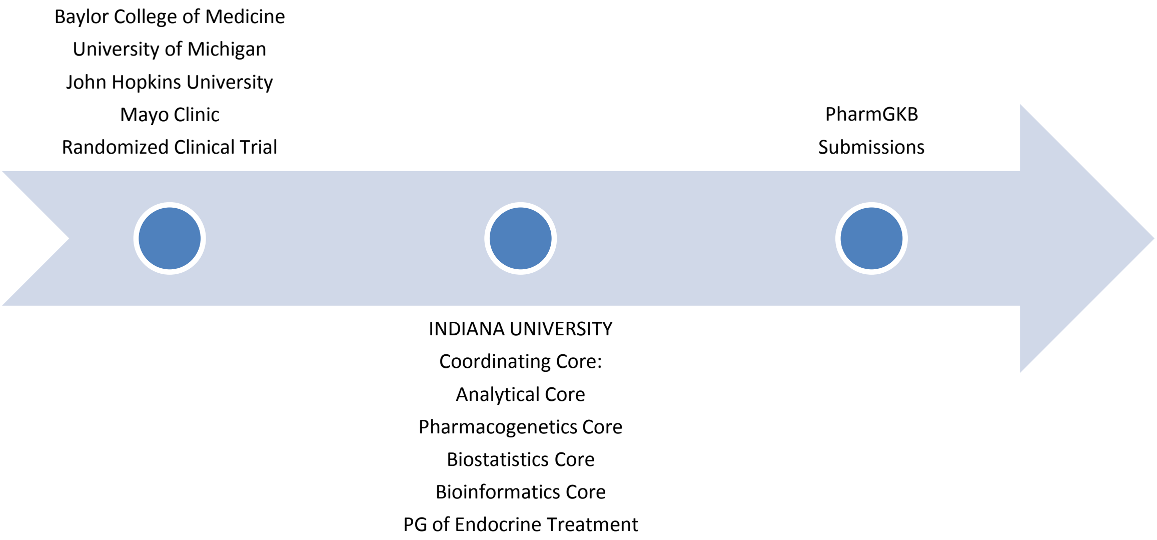


Figure 2 Flow of Information ending with PharmGKB submissions

Data Generation, Format and Entry

The following section explains the original sources of the derived data and how the data are input into the database. When genotyping test results are available, they are manually transferred from the laboratory instruments by a research technician into a standard input file that can be automatically processed. At this point in the workflow, results are verified by a laboratory supervisor acting as a data curator to ensure quality control.

Genotype data are generated in the GCRC Pharmacogenetics Core Laboratory located in the Division of Clinical Pharmacology. The Core Lab outsources DNA (re)sequencing to an external genomic service solutions company, Polymorphic DNA

Technologies (www.polymorphicdna.com). Phenotypic information is generated through patient surveys and clinical data.

Database Design

The data are stored in a relational MySQL® database running on a Linux server hosted at Dr. Sean Mooney's laboratory in the Center for Computational Biology and Bioinformatics (CCBB) at Indiana University. MySQL is a popular, industry standard open source database that provides stable performance. MySQL is installed on a Linux platform with the installed version being mysql Ver 12.22 Distrib 4.0.21, for pc-linux (i686). The database schema for this project has been revised numerous times to accommodate a growing list of user requirements.

A sample-centric approach was implemented in order to optimize the workflow. The main tables hold results for each unique sample. The tables can be grouped into categories with a few exceptions. There is a cluster of tables that represent metadata, which includes details about the clinical trials and additional information about each patient and the associated biological samples. Another group of tables that store data about the different assays (Restriction fragment length polymorphism (RFLP), Luminex®, SYBR® GreenER™, and TaqMan®) that are performed on the patient samples. The assay tables also hold metadata about the assays such as protocol descriptions. Another group of tables holds meta information about the genetic variants which the assays target. Variant data include nucleotide, locus and amino acid sequence information. Sequence data are stored in another set of tables. Other administrative information about users and sessions are also stored. Currently, the database is comprised of 57 non-redundant tables (Figure 3).

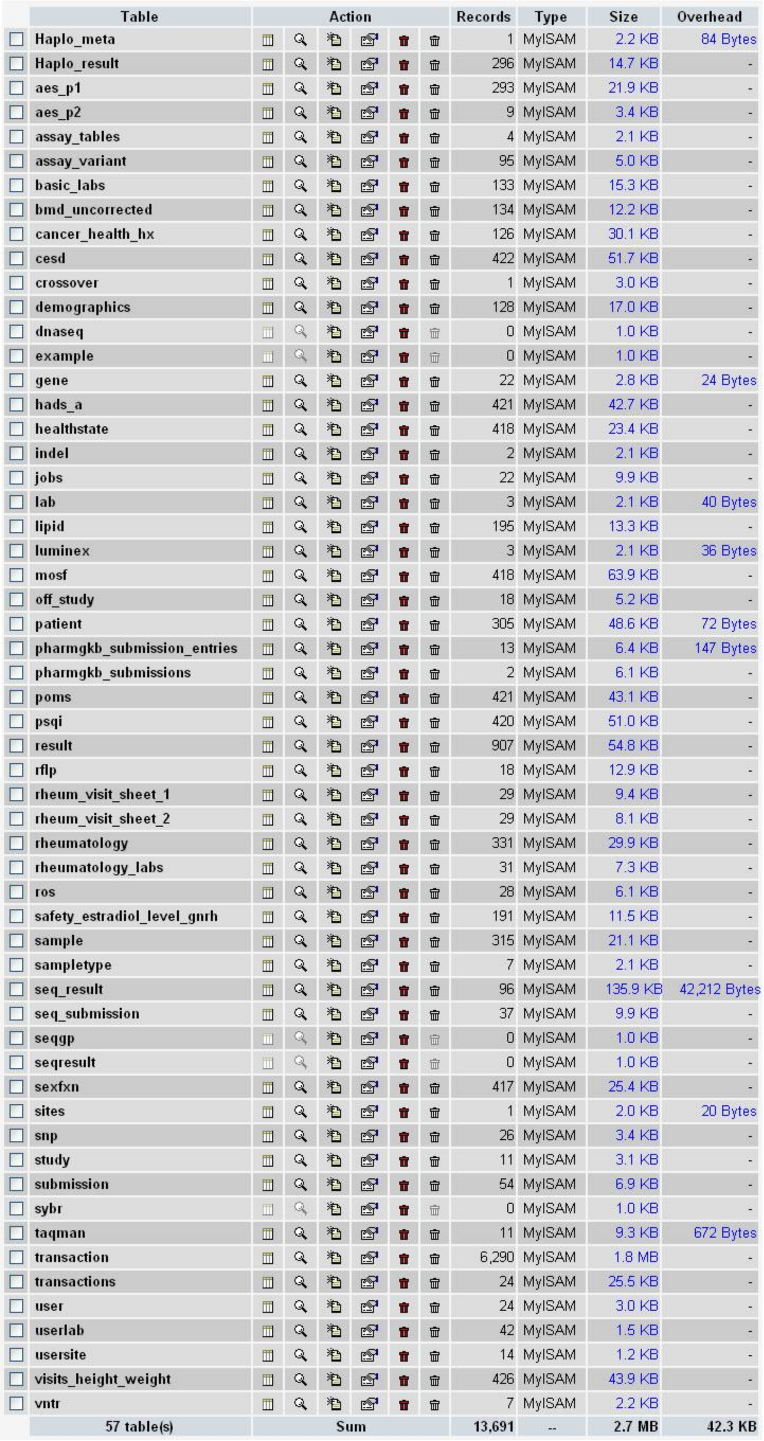


	Table	Action	Records	Type	Size	Overhead
□	Haplo_meta		1	MyISAM	2.2 KB	84 Bytes
□	Haplo_result		296	MyISAM	14.7 KB	-
□	aes_p1		293	MyISAM	21.9 KB	-
□	aes_p2		9	MyISAM	3.4 KB	-
□	assay_tables		4	MyISAM	2.1 KB	-
□	assay_variant		95	MyISAM	5.0 KB	-
□	basic_labs		133	MyISAM	15.3 KB	-
□	bmd_uncorrected		134	MyISAM	12.2 KB	-
□	cancer_health_hx		126	MyISAM	30.1 KB	-
□	cesd		422	MyISAM	51.7 KB	-
□	crossover		1	MyISAM	3.0 KB	-
□	demographics		128	MyISAM	17.0 KB	-
□	dnaseq		0	MyISAM	1.0 KB	-
□	example		0	MyISAM	1.0 KB	-
□	gene		22	MyISAM	2.8 KB	24 Bytes
□	hads_a		421	MyISAM	42.7 KB	-
□	healthstate		418	MyISAM	23.4 KB	-
□	indel		2	MyISAM	2.1 KB	-
□	jobs		22	MyISAM	9.9 KB	-
□	lab		3	MyISAM	2.1 KB	40 Bytes
□	lipid		195	MyISAM	13.3 KB	-
□	luminex		3	MyISAM	2.1 KB	36 Bytes
□	mosf		418	MyISAM	63.9 KB	-
□	off_study		18	MyISAM	5.2 KB	-
□	patient		305	MyISAM	48.6 KB	72 Bytes
□	pharmgkb_submission_entries		13	MyISAM	6.4 KB	147 Bytes
□	pharmgkb_submissions		2	MyISAM	6.1 KB	-
□	poms		421	MyISAM	43.1 KB	-
□	psqi		420	MyISAM	51.0 KB	-
□	result		907	MyISAM	54.8 KB	-
□	rflp		18	MyISAM	12.9 KB	-
□	rheum_visit_sheet_1		29	MyISAM	9.4 KB	-
□	rheum_visit_sheet_2		29	MyISAM	8.1 KB	-
□	rheumatology		331	MyISAM	29.9 KB	-
□	rheumatology_labs		31	MyISAM	7.3 KB	-
□	ros		28	MyISAM	6.1 KB	-
□	safety_estradiol_level_gnrh		191	MyISAM	11.5 KB	-
□	sample		315	MyISAM	21.1 KB	-
□	sampletype		7	MyISAM	2.1 KB	-
□	seq_result		96	MyISAM	135.9 KB	42,212 Bytes
□	seq_submission		37	MyISAM	9.9 KB	-
□	seqgp		0	MyISAM	1.0 KB	-
□	seqresult		0	MyISAM	1.0 KB	-
□	sexfxn		417	MyISAM	25.4 KB	-
□	sites		1	MyISAM	2.0 KB	20 Bytes
□	snp		26	MyISAM	3.4 KB	-
□	study		11	MyISAM	3.1 KB	-
□	submission		54	MyISAM	6.9 KB	-
□	sybr		0	MyISAM	1.0 KB	-
□	taqman		11	MyISAM	9.3 KB	672 Bytes
□	transaction		6,290	MyISAM	1.8 MB	-
□	transactions		24	MyISAM	25.5 KB	-
□	user		24	MyISAM	3.0 KB	-
□	userlab		42	MyISAM	1.5 KB	-
□	usersite		14	MyISAM	1.2 KB	-
□	visits_height_weight		426	MyISAM	43.9 KB	-
□	vntr		7	MyISAM	2.2 KB	-
	57 table(s)	Sum	13,691	--	2.7 MB	42.3 KB

Figure 3 List of tables in database SERM

Figure 4 shows the entity-relationship diagram for the database.

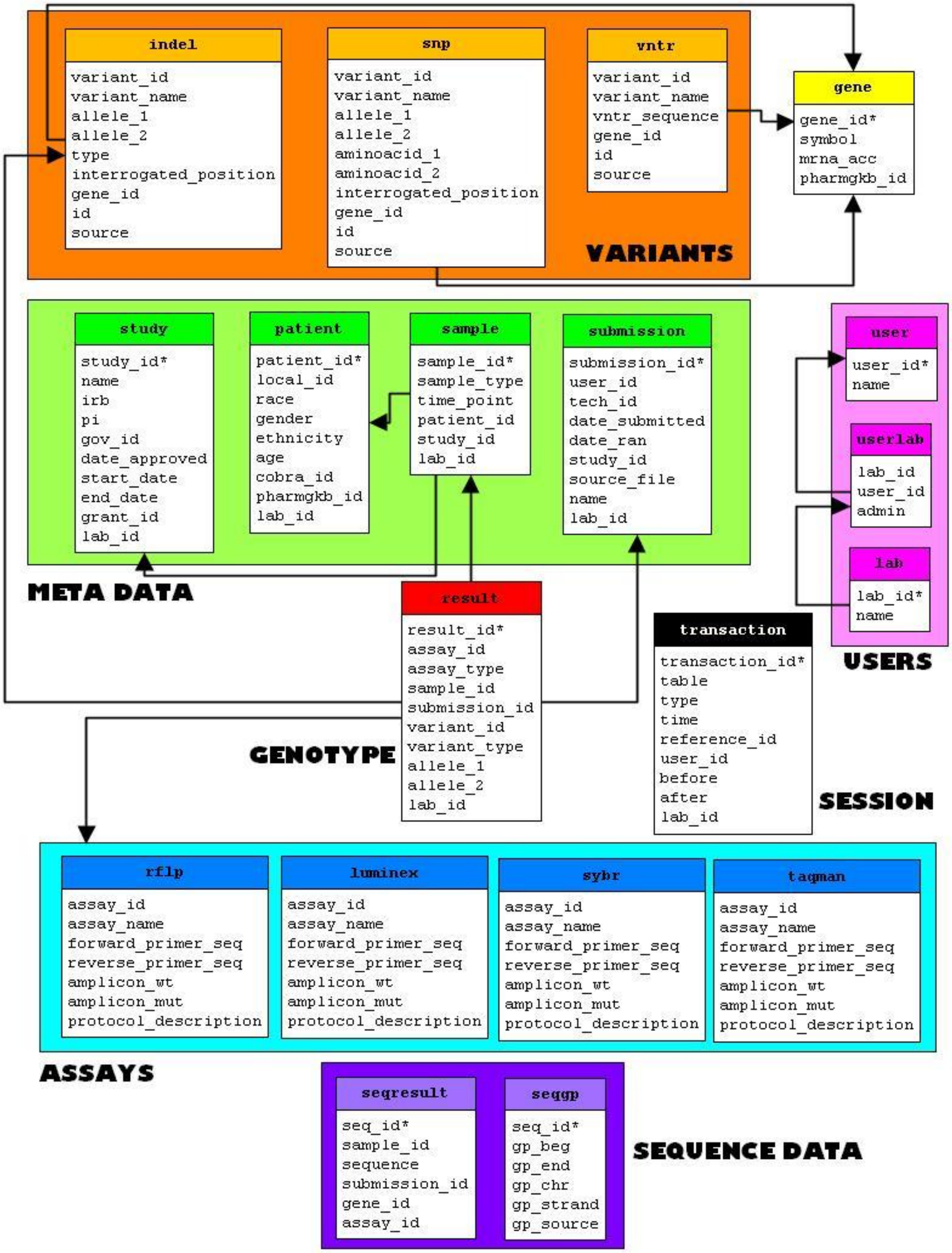


Figure 4 Entity Relationship (ER) diagram

The ER diagram conceptually represents the structure of the data in a relational database. Each entity in the model corresponds to a discrete object (e.g., gene, patient). Connections between the entities represent relationships and cardinality. For example, there is relationship between patient and sample. It is possible for a patient to have one or many associated sample (*e.g.*, from multiple time points). Similarly, a gene can have multiple associated variants.

Web Interface

The SERM database can be accessed via a website and users can query the data via a series of user-friendly web forms. We have streamlined the web-based query forms to be as simple as possible by allowing users to customize their queries.

Querying the Data

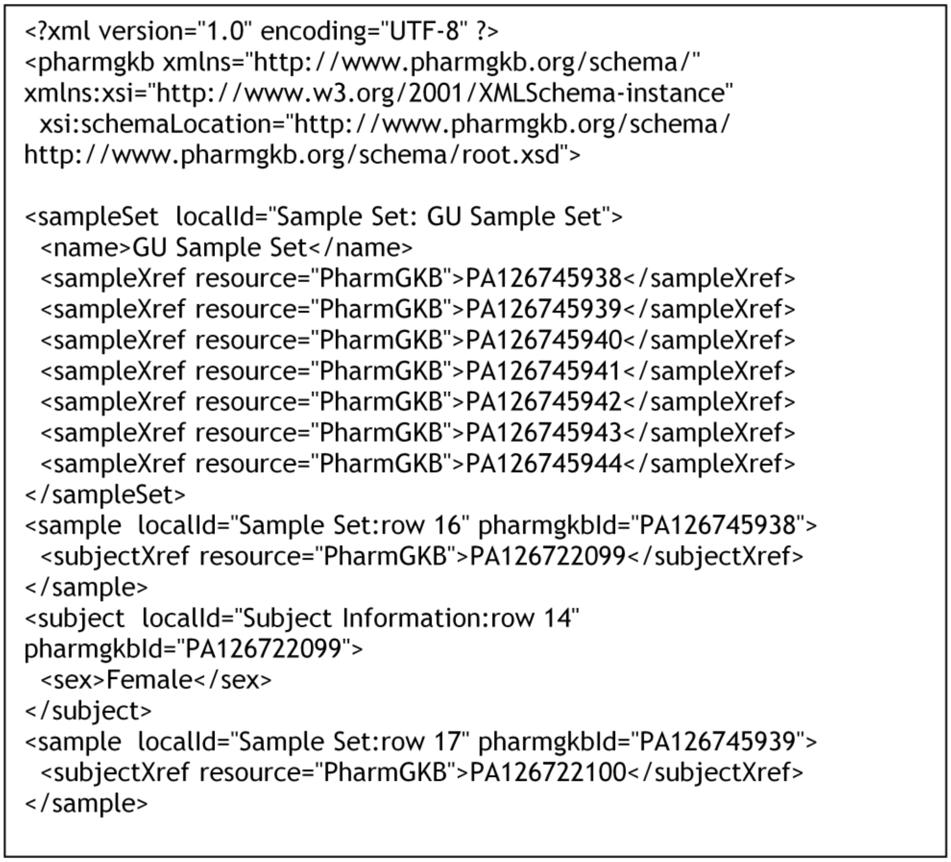
The majority of users will perform simple queries on the data, but the ability to generate more complex searches is available. By leaving query options open, users can browse the data and sort entries by category. Most queries simultaneously examine all the tables in the database and matched records are collected and a view page is dynamically generated and presented. Data can also be filtered and sorted by any attribute such as "Gene Name" shown in Table 3 below.

Table 2 Query results for "Gene"

	Gene Name ∇	mRNA Accession ∇	PharmGKB Accession ID ∇	
□	CYP2C8	NM_000770	PA125	History Edit Delete
□	CYP2C9	NM_000767	PA123	History Edit Delete
□	CYP2D6	NM_000106	PA128	History Edit Delete

Data submission to PharmGKB

COBRA data are submitted to the research network's data hub, PharmGKB, using XML specification. The data are encoded into structured documents that are machine-processable and then transported to PharmGKB. A validation process ensures that the documents are both well-formed and valid in terms of semantics. A sample XML export file shown in Figure 5 illustrates some of the various XML tags.



```
<?xml version="1.0" encoding="UTF-8" ?>
<pharmgkb xmlns="http://www.pharmgkb.org/schema/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.pharmgkb.org/schema/
http://www.pharmgkb.org/schema/root.xsd">

<sampleSet  localId="Sample Set: GU Sample Set">
  <name>GU Sample Set</name>
  <sampleXref resource="PharmGKB">PA126745938</sampleXref>
  <sampleXref resource="PharmGKB">PA126745939</sampleXref>
  <sampleXref resource="PharmGKB">PA126745940</sampleXref>
  <sampleXref resource="PharmGKB">PA126745941</sampleXref>
  <sampleXref resource="PharmGKB">PA126745942</sampleXref>
  <sampleXref resource="PharmGKB">PA126745943</sampleXref>
  <sampleXref resource="PharmGKB">PA126745944</sampleXref>
</sampleSet>
<sample  localId="Sample Set:row 16" pharmgkbId="PA126745938">
  <subjectXref resource="PharmGKB">PA126722099</subjectXref>
</sample>
<subject  localId="Subject Information:row 14"
pharmgkbId="PA126722099">
  <sex>Female</sex>
</subject>
<sample  localId="Sample Set:row 17" pharmgkbId="PA126745939">
  <subjectXref resource="PharmGKB">PA126722100</subjectXref>
</sample>
```

Figure 5 Sample XML Export File

System Security

COBRA data contain patient health information that is protected by the HIPAA privacy rules. As such, we implemented a HIPAA compliant system. The website was properly secured using SSL, a technology that uses digital certificates to authenticate users. We use a standard PHP/MySQL implementation using session variables to save authentication information. This session expires when the browser is closed. All sessions are transmitted over the SSL connection, and passwords are stored in the MySQL database using a mix of one-way hashing algorithms such as MD5 and SHA1. These sessions are checked whenever an action is to be taken on the website.

All COBRA data are kept on RAID 5 volumes ensuring that a failed hard drive will not cause data loss. The web server data are backed up, and the MySQL database is replicated to a separate machine. From there, our data are uploaded to the HPSS tape backup system hosted by Indiana University, and then are replicated to Bloomington. All of these backups occur nightly, and are incremental. Full backups occur on a weekly basis.

Bioinformatic Workflows

The following sections describe the workflow in detail. The methodological approach consists of the following steps:

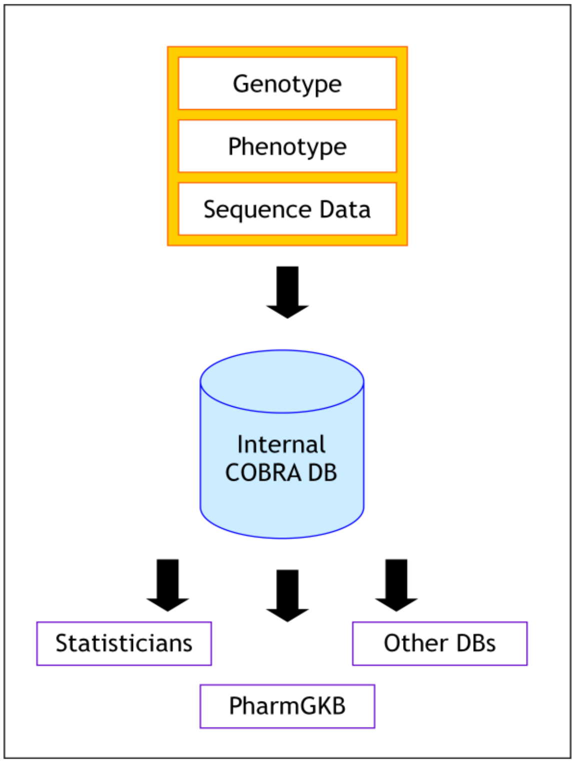
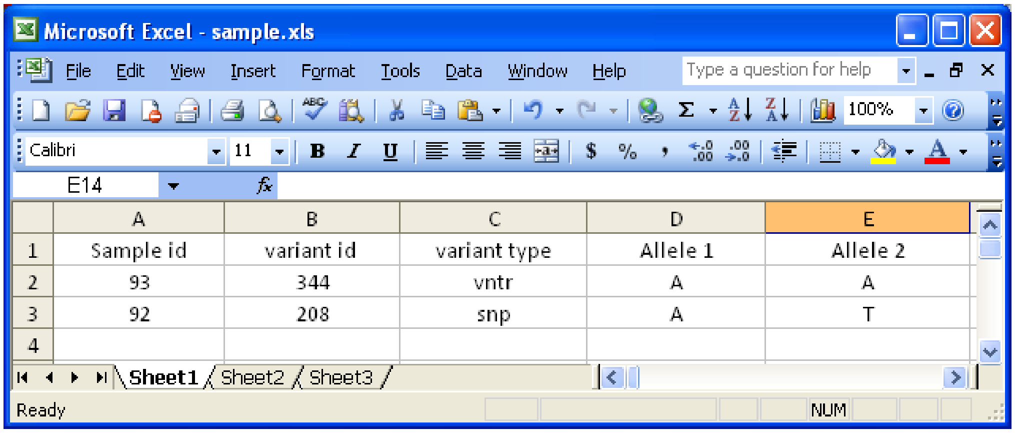


Figure 6 Data flow for submissions

Figure 6 shows information flow for this project. Genotype, phenotype and sequence data are uploaded to the internal data management system and staged for export to other entities such as statisticians and other databases including PharmGKB.

Workflow for Genotype

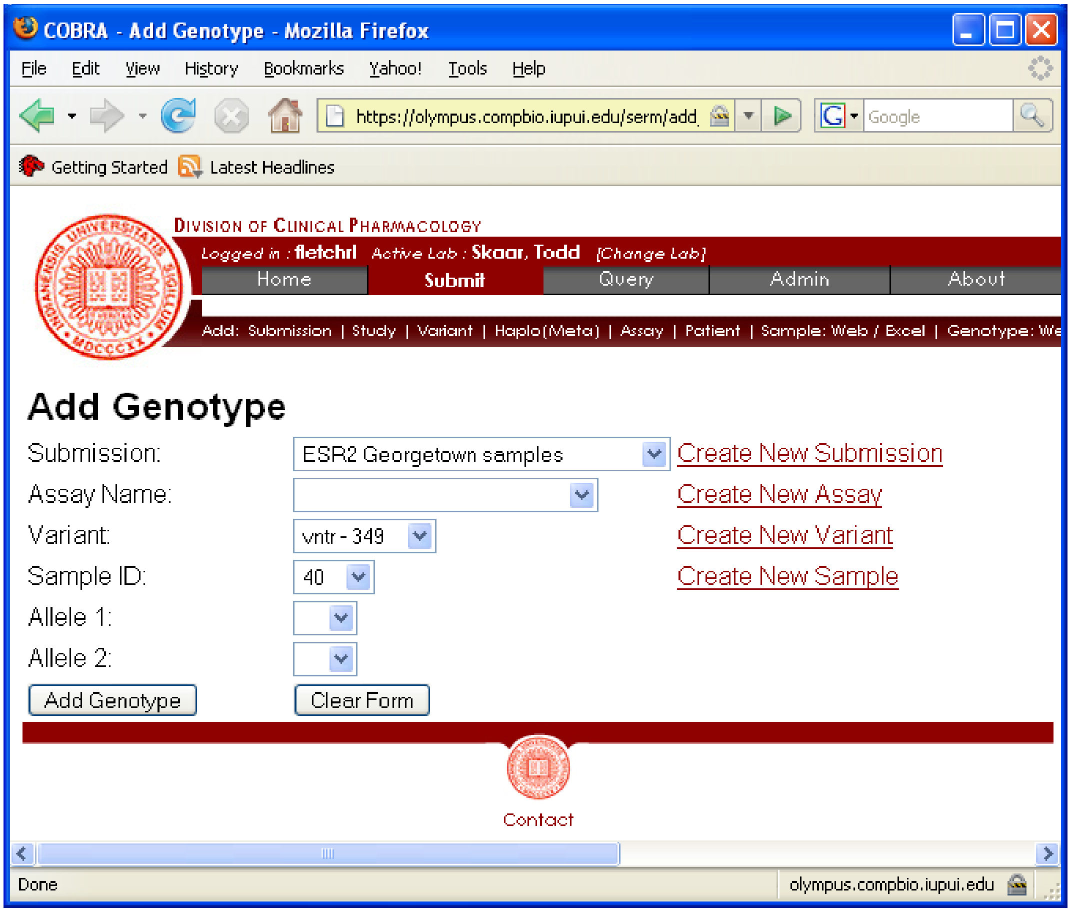
A laboratory technician performs one of several possible genotyping assays. The genotype results are obtained from the laboratory equipment. A technician or other data entry person enters the output into a spreadsheet format (Figure 7). The spreadsheet includes the primary keys for sample ID and variant IDs as well as the type of variant and the specific nucleotides for each allele.


Microsoft Excel - sample.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
100%
Calibri
11
E14
Sheet1 Sheet2 Sheet3
Ready
NUM

	A	B	C	D	E
1	Sample id	variant id	variant type	Allele 1	Allele 2
2	93	344	vntr	A	A
3	92	208	snp	A	T
4					

Figure 7 Excel input file for genotype submission

Although this approach takes considerably more time, data can also be uploaded one record at a time. This method might prove useful if a technician needs to update a single record as opposed to a bulk entry. Figure 8 shows the web form to add a single genotype record. In cases where the lab needs to repeat an experiment on a particular sample, this approach is more convenient.


COBRA - Add Genotype - Mozilla Firefox
File Edit View History Bookmarks Yahoo! Tools Help
https://olympus.compbio.iupui.edu/serm/add
Google
Getting Started
Latest Headlines
Division of Clinical Pharmacology
Logged in : fletchrl Active Lab : Skaar, Todd [Change Lab]
Home
Submit
Query
Admin
About
Add: Submission | Study | Variant | Haplo(Meta) | Assay | Patient | Sample: Web / Excel | Genotype: We

Add Genotype

Submission: ESR2 Georgetown samples

Assay Name:

Variant: vntr - 349

Sample ID: 40

Allele 1:

Allele 2:

Create New Submission
Create New Assay
Create New Variant
Create New Sample
Add Genotype
Clear Form
Contact
Done
olympus.compbio.iupui.edu

Figure 8 Web form for genotype import

Figure 9 illustrates the complete workflow for genotype data. After data is uploaded to the database server and imported into the database it is available for retrieval and analysis.

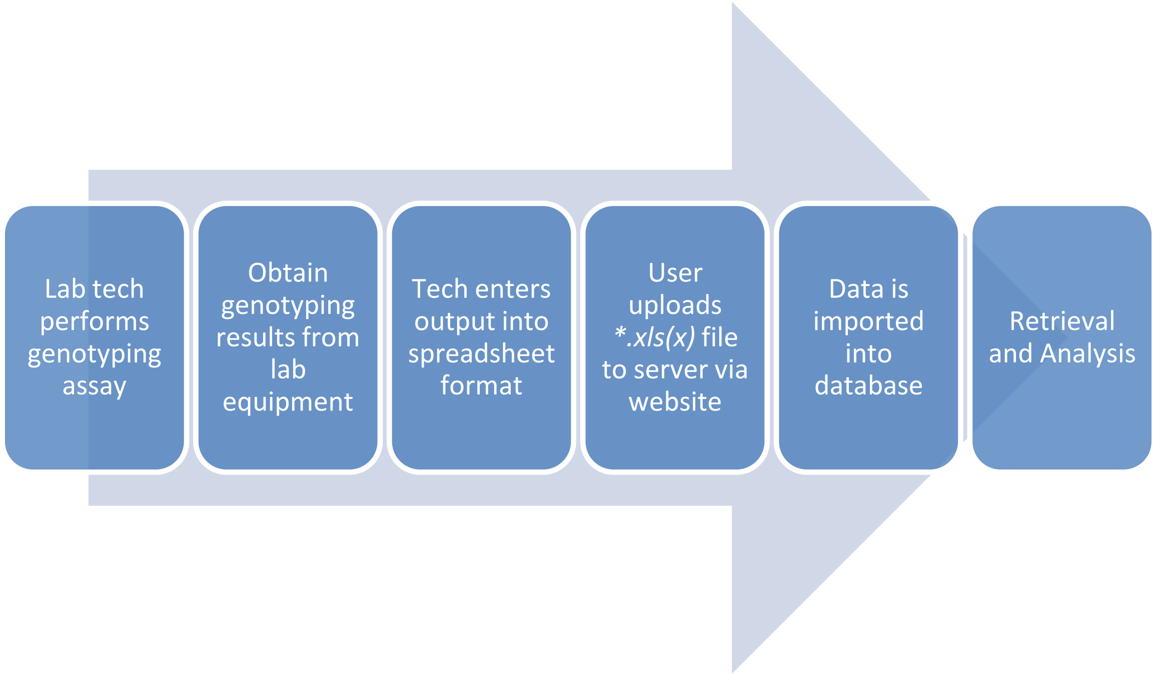

Lab tech performs genotyping assay
Obtain genotyping results from lab equipment
Tech enters output into spreadsheet format
User uploads *.xls(x) file to server via website
Data is imported into database
Retrieval and Analysis

Figure 9 Information flow for genotypic data

Workflow for Phenotype

Detailed and accurate phenotype collection is a major challenge and often a limiting factor to genetic association studies. The problem is further compounded when data are scattered geographically. Targeted recruitment and online surveys are the primary means by which data are collected for storage in the database.

Figure 10 shows the workflow for phenotype data.

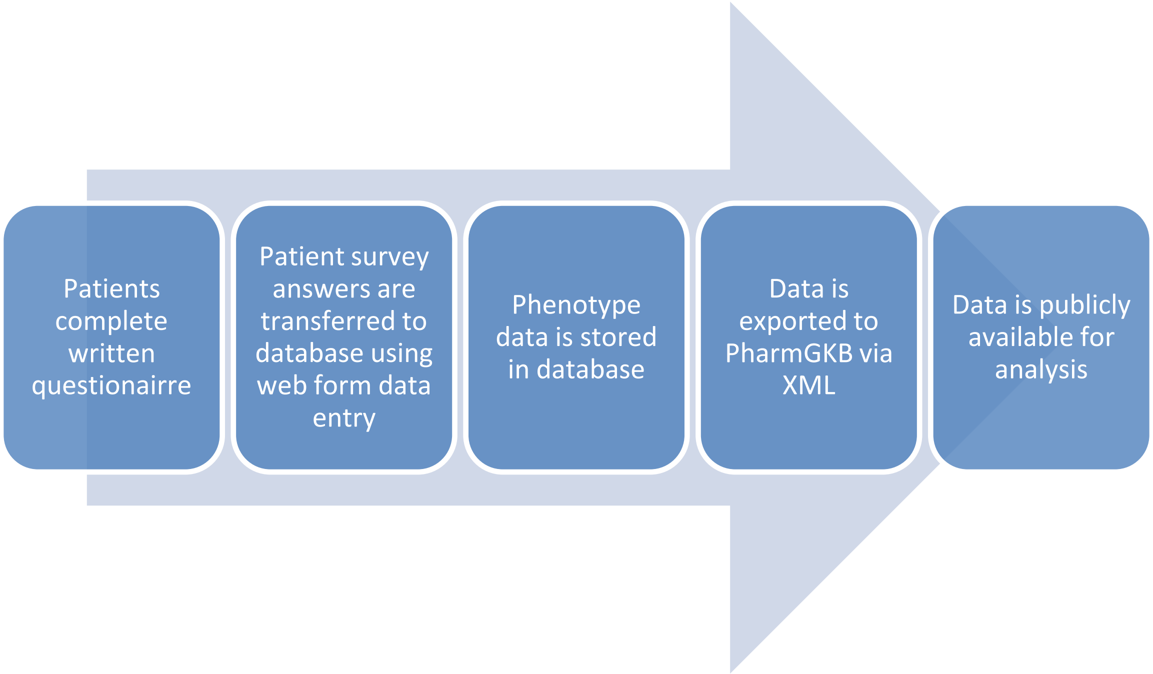
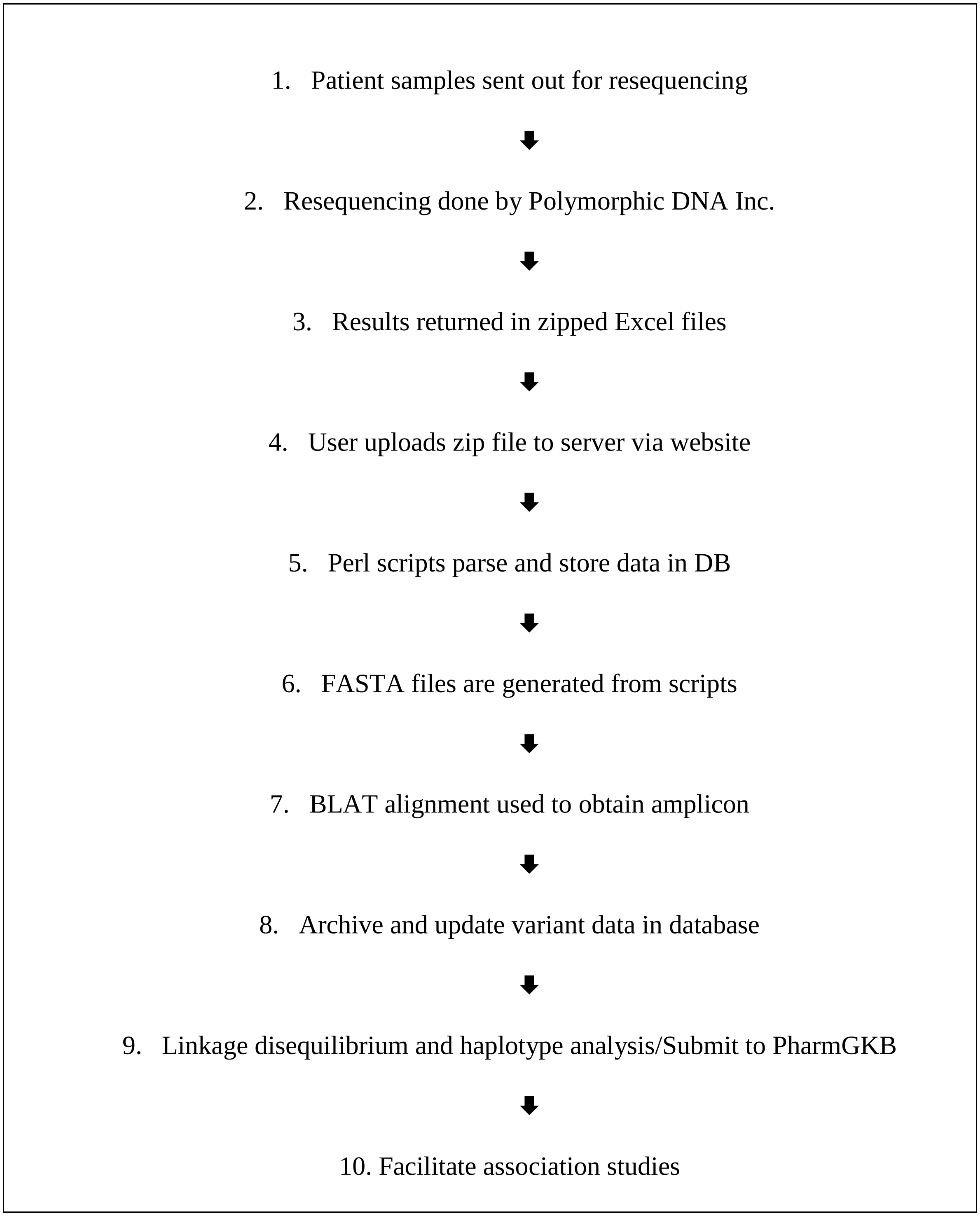

Patients complete written questionairre
Patient survey answers are transferred to database using web form data entry
Phenotype data is stored in database
Data is exported to PharmGKB via XML
Data is publicly available for analysis

Figure 10 Information flow for phenotypic data

Workflow for Sequencing

Patient samples are sent to Polymorphic DNA Inc. for resequencing. Completed resequencing data are returned in spreadsheet format. Perl code was developed to process the Excel files. Individual scripts were written to index variants and track their positions in the relative sequence. Sequence fragments were concatenated in order to generate FASTA formatted files. The BLAT alignment tool (Kent 2002) was used to determine the position of the amplicons relative to a reference sequence (*e.g.*, estrogen reference sequence). Finally, we queried SNP databases using the derived variant position to look for known SNPs. Sequencing and variant data are then stored in COBRA-DB. Figure 11 shows the workflow for sequencing data.



1. Patient samples sent out for resequencing
2. Resequencing done by Polymorphic DNA Inc.
3. Results returned in zipped Excel files
4. User uploads zip file to server via website
5. Perl scripts parse and store data in DB
6. FASTA files are generated from scripts
7. BLAT alignment used to obtain amplicon
8. Archive and update variant data in database
9. Linkage disequilibrium and haplotype analysis/Submit to PharmGKB
10. Facilitate association studies

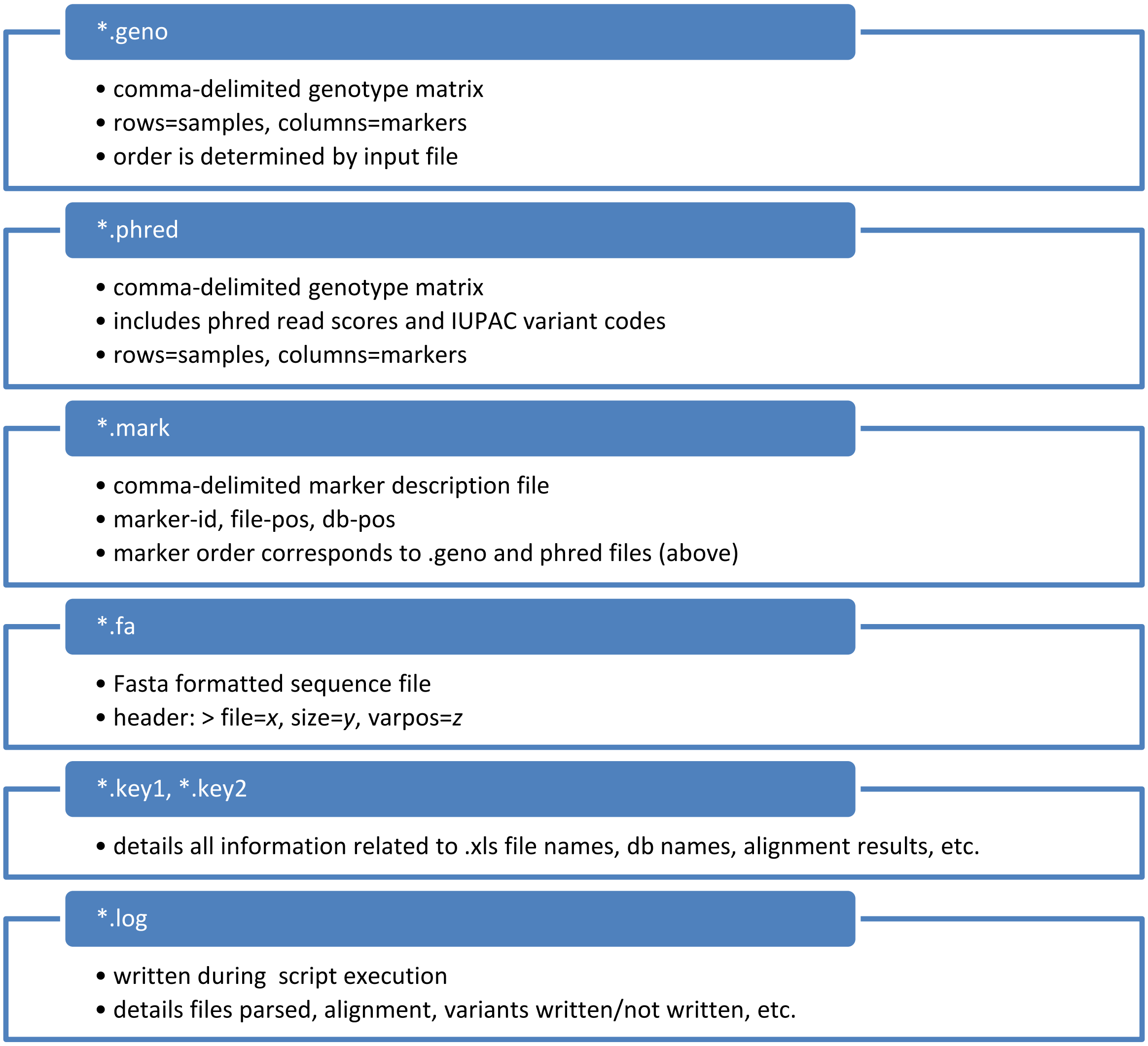
Figure 11 Scientific workflow for resequencing

Table 4 shows the main input and output files that are generated by UNIX-based scripts and then used in the sequencing workflow. `Parse_polydna.blat` parses the Excel files from Polymorphic DNA and invokes the BLAT alignment tool to find sequence matches. The script also produces two output files, *.geno and *.mark, that include comma-delimited genotype matrices and quality scores from the base-calling program Phred. `Pdna_mark2rsid` queries the NCBI dbSNP database for known variant identification numbers (*i.e.*, rsIDs) and also assigns IDs to unknown or novel SNPs. `Pdna_rsid2haplo_sort` and `Pdna_ordergeno` generate the .mark.rsid and .geno files, respectively, that are used as input for the Haploview, a program designed to simplify and expedite haplotype analysis (Barrett *et al*, 2005).

Table 3 UNIX-based Perl Scripts

script	input file	output file
Parse_polydna.blat.pl	.xls	.geno and .mark
Pdna_mark2rsid	.mark	.mark.rsid
Pdna_rsid2haplo_sort	.mark.rsid	.mark.rsid.haplo
Pdna_ordergeno	.mark.rsid and .geno	.geno.o

Figure 12 shows the collection of output files generated by the Perl scripts and provides a more detailed explanation of the contents of each file. Table 5 shows an example of actual file contents.



*.geno

- comma-delimited genotype matrix
- rows=samples, columns=markers
- order is determined by input file

*.phred

- comma-delimited genotype matrix
- includes phred read scores and IUPAC variant codes
- rows=samples, columns=markers

*.mark

- comma-delimited marker description file
- marker-id, file-pos, db-pos
- marker order corresponds to .geno and phred files (above)

*.fa

- Fasta formatted sequence file
- header: > file=*x*, size=*y*, varpos=*z*

*.key1, *.key2

- details all information related to .xls file names, db names, alignment results, etc.

*.log

- written during script execution
- details files parsed, alignment, variants written/not written, etc.

Figure 12 Perl output files and description of contents

Table 4 Screenshots of sequence output files

.geno

```
sample,SNP009_Discovery.xls_105,SNP009_Discovery.xls_178,SNP009_Discovery.xls_191,SNP009_
Discovery.xls_205,SNP009_Discovery.xls_242,SNP010_Discovery.xls_68,SNP010_Discovery.xls_1
95,SNP014_Discovery.xls_88,SNP014_Discovery.xls_252,SNP015_Discovery.xls_268,SNP017_Disco
very.xls_117,SNP017_Discovery.xls_126,SNP017_Discovery.xls_144,SNP019_Discovery.xls_138,S
NP019_Discovery.xls_148,SNP019_Discovery.xls_164,SNP023_Discovery.xls_52,SNP023_Discovery
.xls_118,SNP024_Discovery.xls_98,SNP024_Discovery.xls_139,SNP024_Discovery.xls_220,SNP026
_Discovery.xls_40,SNP026_Discovery.xls_178,SNP026_Discovery.xls_259,SNP026_Discovery.xls_
266,
AA34,G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/C,G/G,A/A,C/C,G/G,T/T,G/G,G/G,C/C,A/A,A/A,A/A,G/G,C/C
AA41,G/G,G/G,G/G,G/G,A/G,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/A,G/G,A/A,A/A,A/A,A/G,C/C
AA09,G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/G,C/G,A/A,A/A,A/A,G/G,C/C
...etc.
```

.phred

```
sample,SNP009_Discovery.xls_105,SNP009_Discovery.xls_178,SNP009_Discovery.xls_191,SNP009_
Discovery.xls_205,SNP009_Discovery.xls_242,SNP010_Discovery.xls_68,SNP010_Discovery.xls_1
95,SNP014_Discovery.xls_88,SNP014_Discovery.xls_252,SNP015_Discovery.xls_268,SNP017_Disco
very.xls_117,SNP017_Discovery.xls_126,SNP017_Discovery.xls_144,SNP019_Discovery.xls_138,S
NP019_Discovery.xls_148,SNP019_Discovery.xls_164,SNP023_Discovery.xls_52,SNP023_Discovery
.xls_118,SNP024_Discovery.xls_98,SNP024_Discovery.xls_139,SNP024_Discovery.xls_220,SNP026
_Discovery.xls_40,SNP026_Discovery.xls_178,SNP026_Discovery.xls_259,SNP026_Discovery.xls_
266,
AA34,G(58),G(69),G(63),G(55),A(34),T(55),T(65),C(54),G(46),A,C(53),G(54),T(64),G(66),G(61
),C(67),A(56),A(47),A(64),G(64),C(66),G(61),T(50),T(63),A(63),AA41,G(57),G(66),G(63),G(65
),R(20),T(64),T(63),Y(29),G(56),A,C(66),G(33),T(62),G(66),A(65),G(44),A(66),A(59),A(67),R
(32),C(62),G(59),T(53),T(55),A(44)...etc.
```

.mark

```
variant,file-pos,NC_000014.7
SNP009_Discovery.xls_105,105,63769935
SNP009_Discovery.xls_178,178,63769862
SNP009_Discovery.xls_191,191,63769849
SNP009_Discovery.xls_205,205,63769835
SNP009_Discovery.xls_242,242,63769798
SNP014_Discovery.xls_88,88,63770492...etc.
```

.fa

```
>file=SNP009_Discovery.xls nbases=247 varpos=105,178,191,205,242
TTGTCCTATGTGTCAGGCCATTGTAGGTGTGTGGTGGGACACAGAGGCTGACAAGACATCGTCCTTGCCCTTGAGCCTAAATTATCAGG
GGGAGCTGGATGCACGAGCCATGGATAAATGGGCTGGGGGAAGAGTGGGTTTAGGGGTGGGGTAGACTGGCTCTGAGCAAAGAGAGCCG
GGGAAGGCTTCGGGGTTCCTGTGGCTGCCTCGGAGGAGGGAATCTCAGCACCTTTTTGTCCCCATAGTA...etc.
```

.key1

```
file      :SNP009_Discovery.xls
blastdb   :NC_000014.7
db_start  :63770039
db_end    :63769793
ref-seq
:TTGTCCTATGTGTCAGGCCATTGTAGGTGTGTGGTGGGACACAGAGGCTGACAAGACATCGTCCTTGCCCTTGAGCCTAAATTATCAG
GGGGAGCTGGATGCACGAGCCATGGATAAATGGGCTGGGGGAAGAGTGGGTTTAGGGGTGGGGTAGACTGGCTCTGAGCAAAGAGAGCC
GGGGAAGGCTTCGGGGTTCCTGTGGCTGCCTCGGAGGAGGGAATCTCAGCACCTTTTTGTCCCCATAGTA
n-bases   :247
ins-pos   :
del-pos   :
var-pos   :105,178,191,205,242
db-var-pos:63769935,63769862,63769849,63769835,63769798...etc.
```

.log

```
PARSING FILES:
=========================================
FILE      : serm/sequence_data/testing/0125C_P1_20051020/SNP009_Discovery.xls
SHEETCOUNT: 1
SHEET     : 0125CSNP009 001-247
REFSEQ    :
TTGTCCTATGTGTCAGGCCATTGTAGGTGTGTGGTGGGACACAGAGGCTGACAAGACATCGTCCTTGCCCTTGAGCCTAAATTATCAGG
GGGAGCTG...etc.
```

Bulk download

In order to prevent users from bogging down the system with repeated queries which may lead to a potential decrease in website performance, data are available for bulk download as Excel, XML, flat files or as relational tables.

PHP Code

The server-side scripting language PHP was used to establish connectivity between the database and the web interface. Embedded within HTML, PHP was used to create dynamic web pages for the front end of the system. Programs for import and export have been written in PHP, XML and Perl.

PharmGKB Submissions

As part of the PGRN network COBRA submits data to the publicly accessible knowledge database, Pharmacogenomics and Pharmacogenetics Knowledge base (PharmGKB) using XML specification. The data are encoded into structured documents that are machine-processable and relatively legible by humans and then transported to PharmGKB. A validation process ensures that the documents are both well-formed and valid in terms of semantics. Figure 13 shows a XML-formatted export file and includes numerous XML tags from the PharmGKB schema on which the code was based.

```
<?xml version="1.0" encoding="UTF-8" ?>
  <pharmgkb xmlns="http://www.pharmgkb.org/schema/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.pharmgkb.org/schema/ "
    http://www.pharmgkb.org/schema/root.xsd">
    <gene localId="GENE1">
      <altName>estrogen receptor 1</altname>
      <altSymbol>ESR1</altSymbol>
      <xref resource="PUbMed">2099</xref>
    </gene>
    <subject localId="PA126722129">
      <sex>female</sex>
      <race>
        <nihCategory>white</nihCategory>
      </race>
    <subject>
    <sample localId="SAMPLE1">
      <subjectXref resource="local">PA126722129</subjectXref>
      <timestamp>10/13/2004</timestamp>
    </sample>
    <sampleSet localId="SAMPLESET1">
      <sampleXref resource="local">SAMPLE1</sampleXref>
    </sampleSet>
    <referenceSequence>
      <geneXref resource=”local”>GENE1</geneXref>
      <dnaSequence>aaacccgggttt</dnaSequence>
      <dnaSequenceSource>genomic</dnaSequenceSource>
      <experiment localId=”EXPERIMENT1”>
        <pcrAssay localId=”PCRASSAY1”>
          <amplicon>
            <startPosition>532</startPosition>
            <stopPosition>1025</stopPosition>
          </amplicon>
          <method>
            <name>ESRI_IVSI-401</name>
            <type>Other</type>
            <templateType>Unknown</templateType>
            <multiPcrAmplificationTested>False</multiPcrAmplificationTested>
            <multiClonesTested>False</multiClonesTested>
            <description>ESRI_IVSI-401</description>
            <parameters>
              PCR Protocol: Add .5 ul of the primer working stock to each PCR tube…
            </parameters>
          </method>
        </pcrAssay>
        <sampleSetXref resource="local">SAMPLESET1</sampleSetXref>
        <genotypesInSample localId="GIS1">
          <genotypingResult localId="RESULT1">
            <assayXref resource=”local”>”PCRASSAY1</assayXref?
            <variant localId="VARIANT1">
              <position>12</position>
              <allele>T</allele>
            </variant>
            <variant localId="VARIANT2">
              <position>13</position>
              <allele>C</allele>
            </variant>
          </genotypingResult>
        <genotypesInSample>
      </expirement>
    </referenceSequence>
  </pharmgkb>
</xml>
```

Figure 13 Sample XML Export File

Data Analysis

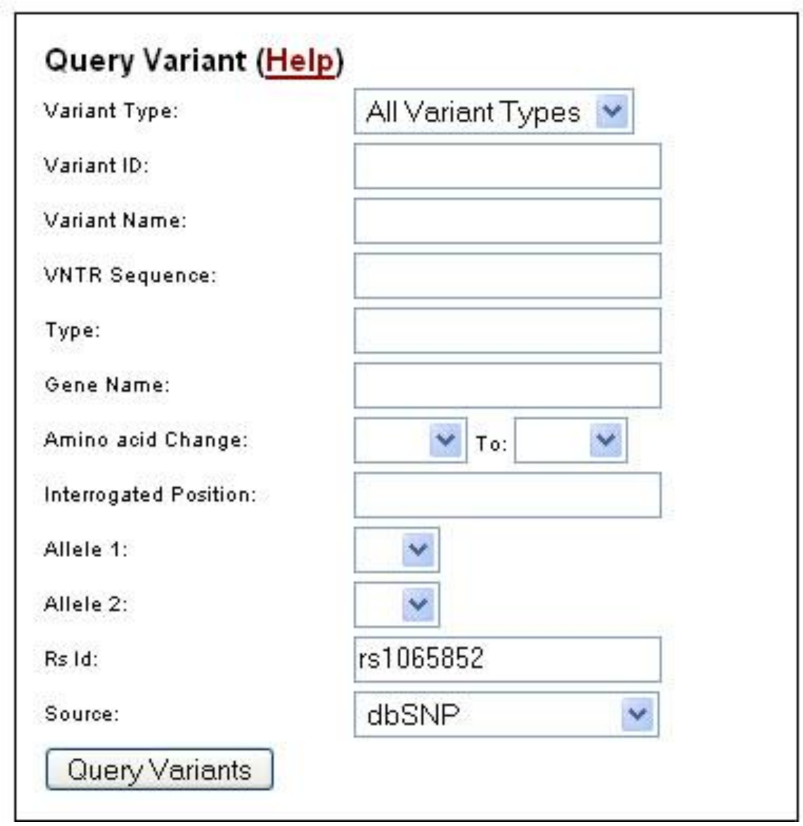
Statisticians use the data warehouse as an internal tool to conduct phenotype genotype association studies. The data can be used to help correlate drug response phenotypes with genetic variation. More specifically, the data can be used to perform a variety of statistical analysis including genetic linkage analysis (*e.g.*, linkage disequilibrium analysis, SNP haplotype reconstruction) and other statistical tests. Similarly, the data can be used as input for other in silico testing related to human disease research.

System Administration

The database includes a web-based system administration panel that allows administrators to manage system tasks such as the creation, modification or deletion of database entities such as laboratories, samples, assays and users. Likewise, administrators and group of super users can manage the dataset of results. User authentication roles have been created to establish tiered access. Administration of the database itself is accomplished over the web via the MySQL database administration tool, phpMyAdmin. Database administrators can effectively create and alter tables, manage privileges, add/edit/delete data and perform other standard admin tasks.

Chapter Four: **Results and Discussion**

The web interface was designed in order to allow users to perform queries in a straightforward manner. To that end, we have created simple web forms for querying the data. Figure 14 shows the form used to query variant data. Users can customize the search by populating the fields in the form with search criteria. Searches can be further narrowed by entering additional terms. In this example, the figure shows a custom search for a SNP with a given rsID number (*i.e*, rs1065852).

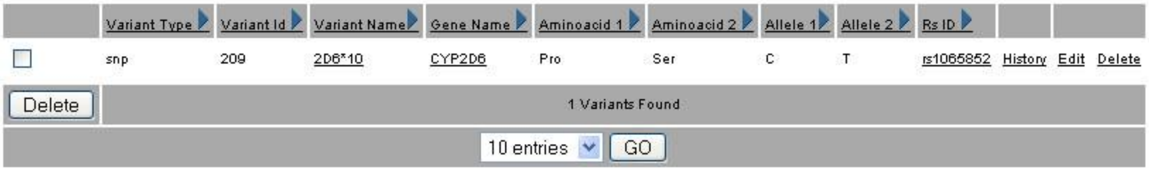


Query Variant (Help)

Variant Type:	All Variant Types
Variant ID:	
Variant Name:	
VNTR Sequence:	
Type:	
Gene Name:	
Amino acid Change:	To:
Interrogated Position:	
Allele 1:	
Allele 2:	
Rs Id:	rs1065852
Source:	dbSNP

Figure 14 Query variant form

Figure 15 shows the returned results from the above query and includes the *Variant Type* (SNP), *variant Id* (auto-incremented primary key), *Variant Name* (arbitrary), *Gene Name* (NCBI Official Symbol), *Aminoacid _1* (wildtype amino acid), *Aminoacid_2* (amino acid change due to variant), *Allele_1* (wildtype allele), *Allele_2* (mutant allele) and *RsID* (dbSNP identification number). The results also show administrative functions such as `delete` and `edit`. These options are only available when the user is logged in as an administrator.



	Variant Type	Variant Id	Variant Name	Gene Name	Aminoacid 1	Aminoacid 2	Allele 1	Allele 2	Rs ID			
□	snp	209	2D6*10	CYP2D6	Pro	Ser	C	T	rs1065852	History	Edit	Delete

Delete

1 Variants Found

10 entries GO

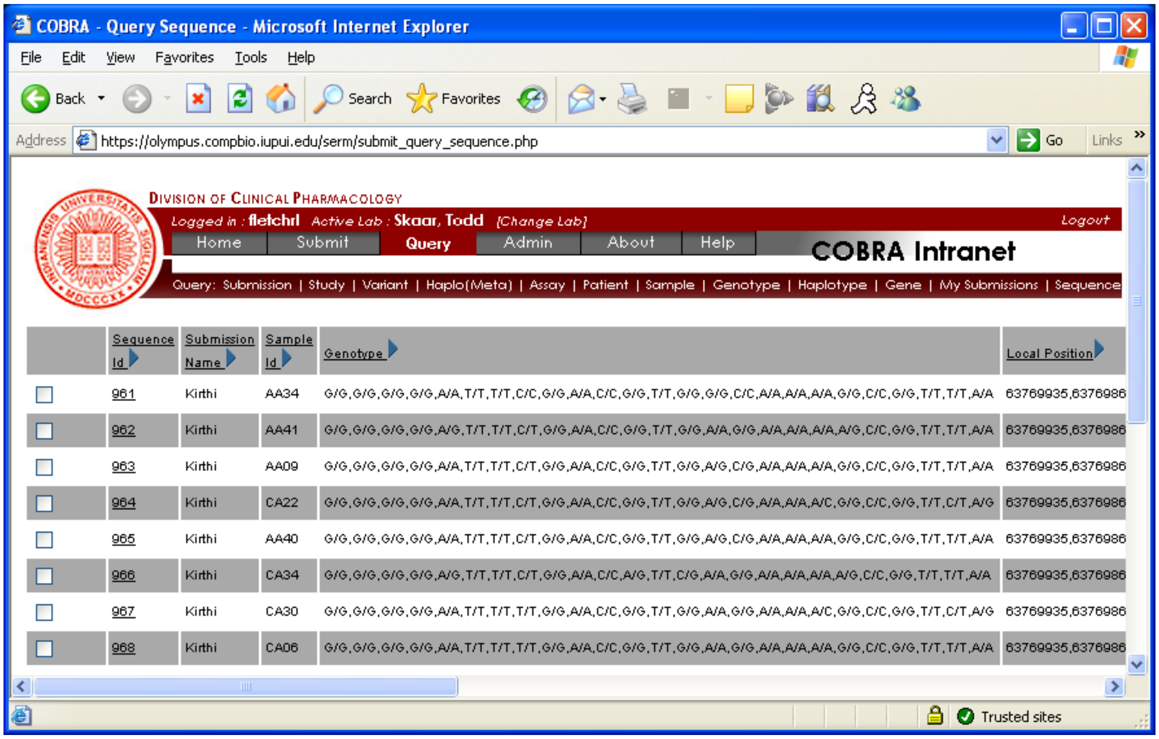
Figure 15 Search results for variant rs1065852

We are able to efficiently submit data to PharmGKB, which fulfills a major business need, which is satisfying metrics for grant renewals. Table 5 shows a summary of COBRA-DB records to date and illustrates the volume of data which this system supports.

Database Contents

Table 5 Summary of database records

Variants	33
Assays	31
Patients/Samples	305
Genotypes	890
Genes	22


COBRA - Query Sequence - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address https://olympus.compbio.iupui.edu/serm/submit_query_sequence.php
Go
Links
Division of Clinical Pharmacology
Logged in : fletchrl Active Lab : Skaar, Todd [Change Lab]
Logout
Home
Submit
Query
Admin
About
Help
COBRA Intranet
Query: Submission | Study | Variant | Haplo(Meta) | Assay | Patient | Sample | Genotype | Haplotype | Gene | My Submissions | Sequence
Trusted sites

	Sequence Id	Submission Name	Sample Id	Genotype	Local Position
□	961	Kirthi	AA34	G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/C,G/G,A/A,C/C,G/G,T/T,G/G,G/G,C/C,A/A,A/A,A/A,G/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986
□	962	Kirthi	AA41	G/G,G/G,G/G,G/G,A/G,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/A,G/G,A/A,A/A,A/A,A/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986
□	963	Kirthi	AA09	G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/G,C/G,A/A,A/A,A/A,G/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986
□	964	Kirthi	CA22	G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/G,C/G,A/A,A/A,A/C,G/G,C/C,G/G,T/T,C/T,A/G	63769935,6376986
□	965	Kirthi	AA40	G/G,G/G,G/G,G/G,A/A,T/T,T/T,C/T,G/G,A/A,C/C,G/G,T/T,G/G,A/G,C/G,A/A,A/A,A/A,G/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986
□	966	Kirthi	CA34	G/G,G/G,G/G,G/G,A/G,T/T,T/T,C/T,G/G,A/A,C/C,A/G,T/T,C/G,A/A,G/G,A/A,A/A,A/A,A/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986
□	967	Kirthi	CA30	G/G,G/G,G/G,G/G,A/A,T/T,T/T,T/T,G/G,A/A,C/C,G/G,T/T,G/G,A/A,G/G,A/A,A/A,A/C,G/G,C/C,G/G,T/T,C/T,A/G	63769935,6376986
□	968	Kirthi	CA06	G/G,G/G,G/G,G/G,A/A,T/T,T/T,T/T,G/G,A/A,C/C,G/G,T/T,G/G,A/A,G/G,A/A,A/A,A/A,G/G,C/C,G/G,T/T,T/T,A/A	63769935,6376986

Data can also be filtered and sorted by any attribute such as “Local ID” shown in Table 6 below. This table shows example results for querying the Sample table.

Table 6 Query results for "Sample"

	Local ID ∇	Sample Name ∇	Time Point ∇	
□	IU047	Blood 721	0 months	History Edit Delete
□	IU048	Blood 722	0 months	History Edit Delete
□	IU049	Blood 723	0 months	History Edit Delete

Sequence Data Files and Haploview

Users can query sequence data via the web interface and download raw data files for input into the software program Haploview developed by Dr. Mark Daly's lab at the Harvard Broad Institute (www.broad.mit.edu/mpg/haploview/) (Barrett *et al*, 2005). Users can also automatically launch Haploview (ver. 4.0) directly from the website. Haploview is a web-based haplotype analysis tool that performs linkage disequilibrium (LD) analysis and haplotype block analysis among other tests.

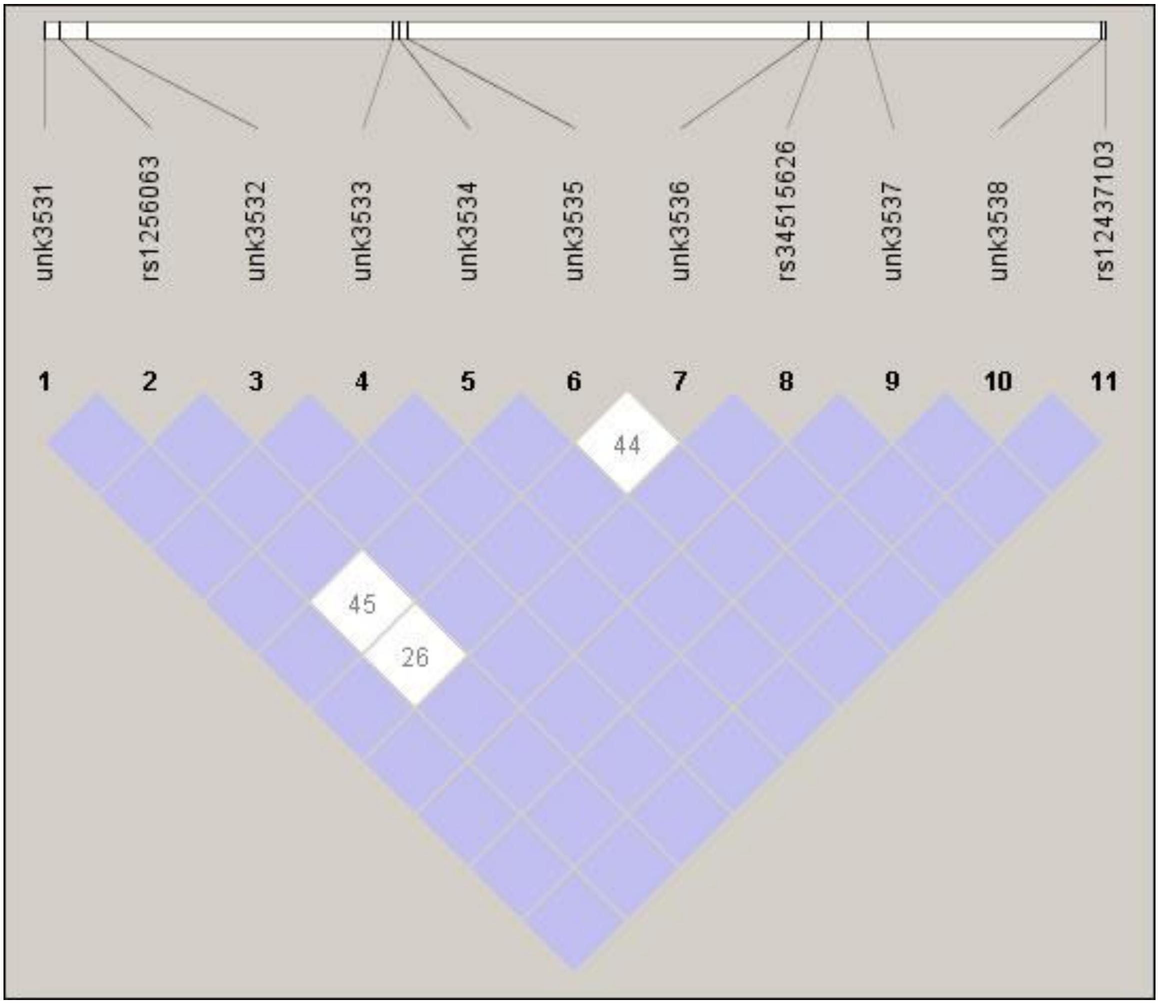
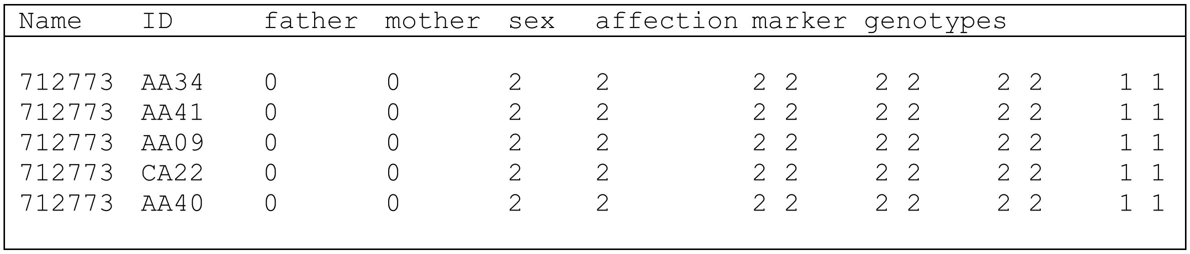


Figure 16 Haploview LD Plot

Figure 16 shows a typical LD plot generated in Haploview. The plot is used to analyze and visualize patterns of linkage disequilibrium in genetic data. The top of the figure includes a map that shows the locations of variants followed by labels for those same variants. Variants are labeled with the prefix unk- or rs- depending on whether they are novel or previously unknown or already exist in NCBI's dbSNP database, respectively.

Haploview accepts input data in five formats. Figure 17 is a sample file containing linkage data in standard linkage format. The last eight columns are paired (one column for each allele) and coded as 1-4 where 1=A, 2=C, 3=G and T=4 (0 indicates missing data).



```
Name    ID      father  mother  sex   affection marker genotypes

712773  AA34    0       0       2     2         2 2     2 2     2 2     1 1
712773  AA41    0       0       2     2         2 2     2 2     2 2     1 1
712773  AA09    0       0       2     2         2 2     2 2     2 2     1 1
712773  CA22    0       0       2     2         2 2     2 2     2 2     1 1
712773  AA40    0       0       2     2         2 2     2 2     2 2     1 1
```

Figure 17 Linkage data flat file

The flat file *pdna.mark.rsid.haplo* includes a list of known (rs) and unknown (unk) SNPs in the first column, followed by each of their chromosomal positions per the respective NCBI genomic contig (Figure 18).

```
unk3531        63771894
rs1256063      63771970
unk3532        63772099
unk3533        63773569
unk3534        63773596
unk3535        63773636
unk3536        63775566
rs34515626     63775623
unk3537        63775854
unk3538        63776968
rs12437103     63776987
```

Figure 18 Variant ID data file

PharmGKB Submissions

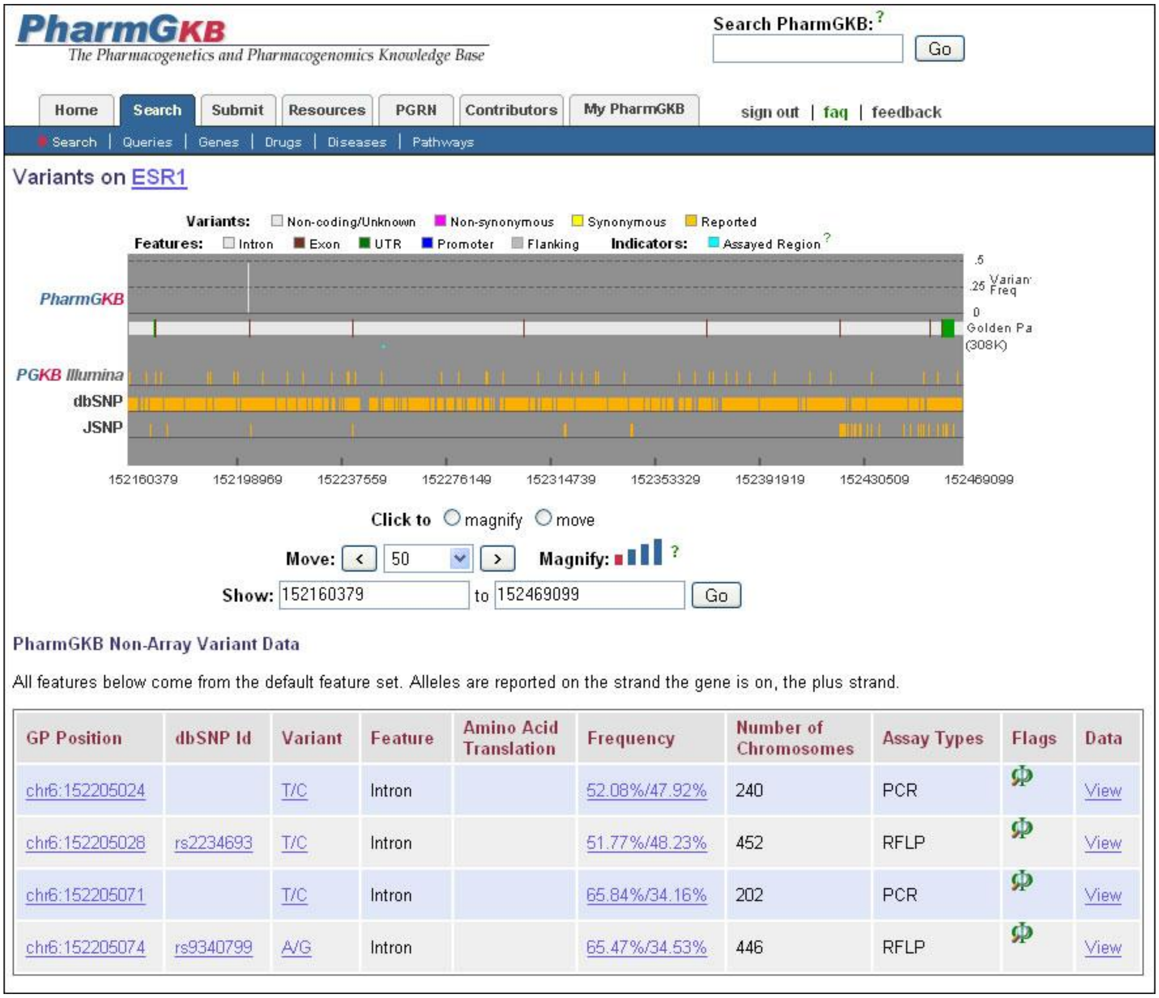
The following tables (Tables 8 and 9) illustrate COBRA data submissions to PharmGKB as of May 2008.

Table 7 PharmGKB Genotype Submissions by COBRA

Genotype Submissions	Number of Variants
ABCB1	79
CYP19A1	98
CYP2A7P1	26
CYP2B6	26
CYP2C19	27
CYP2C9	22
CYP2D6	23
CYP3A	304
CYP3A5	64
ESR1	4
ESR2	51
HTR2A	4
NOS3	120
SULT1A1	45
SULT1A2	53

Table 8 PharmGKB Phenotype Submissions by COBRA

Phenotype Submissions	Number of Patients (n)
Patient responses to tamoxifen	30
Lipid measurements in tamoxifen study	61
Patient responses to tamoxifen (dataset 2)	61
Lipid measurements in tamoxifen study (dataset 2)	104
Thyroid binding globulin in tamoxifen patients	60
Hot flashes in tamoxifen patients	169
Pharmacokinetics of tamoxifen at 4 months	54


PharmGKB
The Pharmacogenetics and Pharmacogenomics Knowledge Base
Search PharmGKB: ?
Go
Home
Search
Submit
Resources
PGRN
Contributors
My PharmGKB
sign out | faq | feedback
Search | Queries | Genes | Drugs | Diseases | Pathways
Variants: □ Non-coding/Unknown ■ Non-synonymous ■ Synonymous ■ Reported
Features: □ Intron ■ Exon ■ UTR ■ Promoter ■ Flanking Indicators: ■ Assayed Region ?
PharmGKB
PGKB Illumina
dbSNP
JSNP
.5
.25 Varian Freq
0
Golden Pa (308K)
152160379
152198969
152237559
152276149
152314739
152353329
152391919
152430509
152469099
Click to ○ magnify ○ move
Move: < 50 >
Magnify: ?
Show: 152160379 to 152469099 Go

Variants on ESR1

PharmGKB Non-Array Variant Data

All features below come from the default feature set. Alleles are reported on the strand the gene is on, the plus strand.

GP Position	dbSNP Id	Variant	Feature	Amino Acid Translation	Frequency	Number of Chromosomes	Assay Types	Flags	Data
chr6:152205024		T/C	Intron		52.08%/47.92%	240	PCR	Φ	View
chr6:152205028	rs2234693	T/C	Intron		51.77%/48.23%	452	RFLP	Φ	View
chr6:152205071		T/C	Intron		65.84%/34.16%	202	PCR	Φ	View
chr6:152205074	rs9340799	A/G	Intron		65.47%/34.53%	446	RFLP	Φ	View

Figure 19 COBRA ESR1 PharmGKB Submission

The previous figure (Figure 19) represents one of many COBRA submissions to PharmGKB.

Submitting data efficiently has been a major improvement. As described previously, data submission was a major problem, traditionally a laborious task, and caused a bottle neck effect in the work flow, which ultimately hindered research and negatively impacted grant renewals.

Chapter Five: **Conclusions**

COBRA-DB is an online information management tool for the storage, organization and export of pharmacogenomic data related to breast cancer and drug responses. A relational model provides integration of multiple data sets while a web interface supports trouble-free data import and export. Simple query forms provide users with a tool to perform uncomplicated searches. Data annotation provides information to facilitate genotype/phenotype association studies that will help advance pharmacogenomic breast cancer research.

Limitations

In terms of quality control, it is important to note that the information system does not interface directly with laboratory instruments. Unfortunately, a human operator is needed to transfer output from lab devices and perform an initial formatting of the data. At this point in the workflow, the system is vulnerable to human error and this issue requires attention on a quality control level.

Haplotype data also presented a challenge. Initially, the database was not designed to store haplotype data. However, the schema was updated in an attempt to include haplotype results, but the database has not been completely populated due to an unresolved and still outstanding issue relating to nomenclature used to describe variants in the Cytochrome P450 system, a group of drug-metabolizing enzymes. Once a consensus on nomenclature is reached, efforts to fully incorporate haplotype results will be resumed.

Similarly, the inherent fluidity present in biological data presented a related challenge. Developing a system that can be easily adapted to concepts which are highly susceptible to change presents a technical challenge. As new data are collected or as data evolve, the database must be extended to accommodate the new information.

Finally, providing an accurate estimate of confidence in the data was difficult. Statistical quality measures were lacking from the quality control process. We aim to address this and the other limitations of this study in our future work.

Future Enhancements

In accordance with budget and time constraints, we propose to upgrade the system by incorporating the following improvements.

Multiplex Assay Data

We intend to extend the current database model to support new data generated by multiplex SNP detection assays. Currently, the schema is configured to handle one unique variation per assay and would need to be expanded to include storage for additional results.

Web-based Reporting

User requirements have indicated that the addition of web-based reporting tools would be beneficial. Web-based reports can be used to generate data summaries to help track data and deliver a complete data picture. Reports can also generate effective charts and graphs to help manage project progress.

Haplotype Data

Future version of the database will include updates to feature haplotype information. Presently, variation data can be visualized as individual genotypes.

However, an aggregated view of these variations would be helpful for linkage disequilibrium (LD) and haplotype analysis.

Software distribution

Finally, we have started the process to bundle the software for distribution to other parties. Other members within the PGRN work with similar datasets and struggle with similar data management issues and have expressed an interest in our system. We plan to share the software with COBRA collaborators and offer the system to PharmGKB for distribution on their website to a broader audience.

Summary

We have created a web-based information system to manage genotype, phenotype and sequence data for COBRA. The integration of this data is an essential step for performing holistic pharmacogenomic analysis. Assembly of data is a first step to understanding the associations between genetic variation and phenotype response and expanding our knowledge of drug response in individuals. The relational database serves as a staging ground to organize and annotate the data before it is transported to public repositories such as PharmGKB or other bioinformatic applications for further such analysis. By combining these datasets via a relational database, investigators are able to access the information and perform research in a straightforward and simple manner. The system also ensures data integrity by adhering to best practices in security. As the amount of information from genomic studies rapidly increases, COBRA-DB can be extended to incorporate new data.

References

Abdullah Kahraman, Andrey Avramov, Lyubomir Nashev, Dimitar Popov, Rainer Ternes, Hans-Dieter Pohlenz, and Bertram Weiss. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics.

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005 Jan 15.

Frenkel, KA. The Human Genome Project and Informatics. Communications of the ACM. 1991 Nov; 34(11):41-51.

Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, Gong L, Owen R, Gong M, Gor W, Liu F, Truong C, Whaley R, Woon M, Zhou T, Altman RB, Klein TE. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Res. 2008 Jan;36(Database issue):D913-8.

Giacomini KM, Brett CM, Altman RB, Benowitz NL, Dolan ME, Flockhart DA, Johnson JA, Hayes DF, Klein T, Krauss RM, Kroetz DL, McLeod HL, Nguyen AT, Ratain MJ, Relling MV, Reus V, Roden DM, Schaefer CA, Shuldiner AR, Skaar T, Tantisira K, Tyndale RF, Wang L, Weinshilboum RM, Weiss ST, Zineh I; Pharmacogenetics Research Network. The pharmacogenetics research network: from SNP discovery to clinical drug response. Clin Pharmacol Ther. 2007 Mar;81(3):328-45. Review.

Goetz MP, Rae JM, Suman VJ, Safgren SL, Ames MM, Visscher DW, Reynolds C, Couch FJ, Lingle WL, Flockhart DA, Desta Z, Perez EA, Ingle JN. Pharmacogenetics of tamoxifen biotransformation is associated with clinical outcomes of efficacy and hot flashes. J Clin Oncol. 2005 Dec 20;23(36):9312-8.

Goh CS, Gianoulis TA, Liu Y, Li J, Paccanaro A, Lussier YA, Gerstein M. Integration of curated databases to identify genotype-phenotype associations. BMC Genomics. 2006 Oct 12;7:257.

Groth P, Pavlova N, Kalev I, Tonov S, Georgiev G, Pohlenz HD, Weiss B. PhenomicDB: a new cross-species genotype/phenotype resource. Nucleic Acids Res. 2007 Jan;35(Database issue):D696-9.

Hernandez-Boussard T, Woon M, Klein TE, Altman RB. Integrating large-scale genotype and phenotype data. OMICS. 2006 Winter;10(4):545-54.

Hewwett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res. 2002 Jan 1;30(1):163-5.

Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, Pohlenz HD, Weiss B. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics. 2005 Feb 1;21(3):418-20.

Kent, JW. BLAT – The BLAST-Like Alignment Tool. Genome Research. 2002 Apr 12(4):656-664.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007 Oct;39(10):1181-6.

Ntukidem NI, Nguyen AT, Stearns V, Rehman M, Schott A, Skaar T, Jin Y, Blanche P, Li L, Lemler S, Hayden J, Krauss RM, Desta Z, Flockhart DA, Hayes DF;

Consortium on Breast Cancer Pharmacogenomics. Estrogen receptor genotypes, menopausal status, and the lipid effects of tamoxifen. Clin Pharmacol Ther. 2008 May;83(5):702-10.

Owen RP, Altman RB, Klein TE. PharmGKB and the International Warfarin Pharmacogenetics Consortium: the changing role for pharmacogenomic databases and single-drug pharmacogenetics. Hum Mutat. 2008 Apr;29(4):456-60.

Price TD, Qvarnström A, Irwin DE. The role of phenotypic plasticity in driving genetic evolution. Proc Biol Sci. 2003 Jul 22;270(1523):1433-40.

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

Whirl-Carrillo M, Woon M, Thorn CF, Klein TE, Altman RB. An XML-based interchange format for genotype-phenotype data. Hum Mutat. 2008 Feb;29(2):212-9.

Websites

1. www.broad.mit.edu/mpg/haploview
2. www.ncbi.nlm.nih.gov
3. www.pharmgkb.org
4. www.phenomicdb.de
5. www.polymorphicdna.com

Appendices

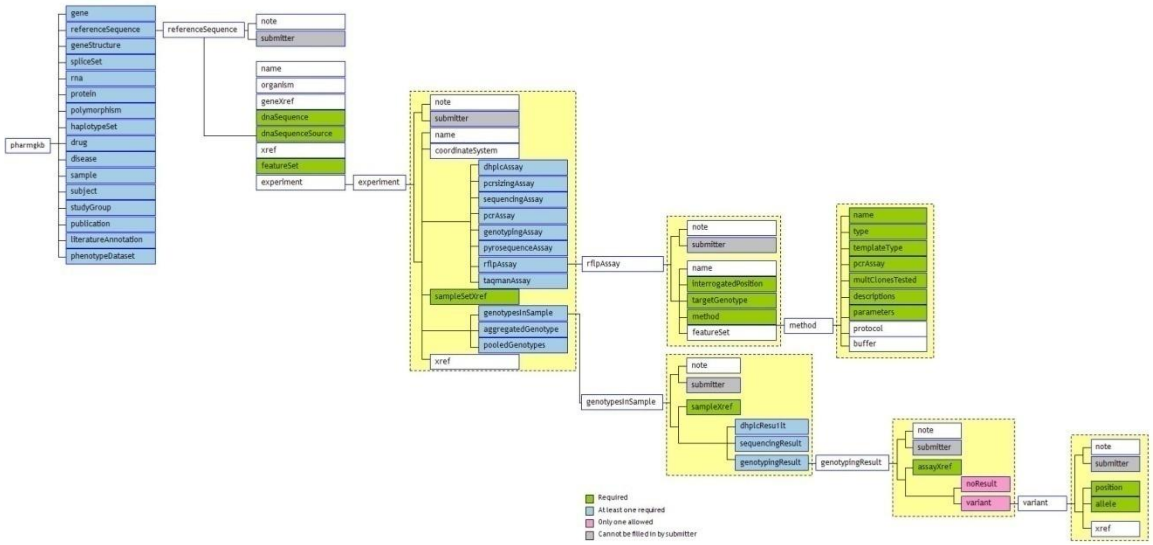
Appendix A: Representative Code

The following PHP code accesses and displays contents of database tables.

```
/*  Filename: view_gene.php
    Last Update: Oct. 26, 2007
    Description: Displays gene table
*/

<?php
        include("include/function.inc.php");
        session_start();
        check_session();

        db_connect();
        if(isset($_GET['gene_id']))
        {
                $gene_id = $_GET['gene_id'];

                $sql = "SELECT * FROM gene WHERE gene_id = '$gene_id';";
                $result = mysql_query($sql);
                if(!$result || mysql_num_rows($result) < 1)
                {
                        error("There are no genes that exist with that gene id");
                        exit(-1);
                }
                else
                {

                        $gene_id = mysql_result($result, 0, "gene_id");
                        $symbol = mysql_result($result, 0, "symbol");
                        $mrna_acc = mysql_result($result, 0, "mrna_acc");
                        $pharmgkb_id = mysql_result($result,0,"pharmgkb_id");


                }
        }
        else
        {
                error("You must specify an gene_id");
                exit(-1);
        }
        db_close();
        page_start("query", "View Gene");
?>

<table width="50%">
        <tr class="header"><td>View Gene</td></tr>
        <tr><td class="field">Gene Id:</td><td><?php print("$gene_id"); ?></td></tr>
        <tr><td class="field">Gene Symbol:</td><td><?php print($symbol); ?></td></tr>
        <tr><td class="field">mRNA_Accession:</td><td><?php print($mrna_acc);
?></td></tr>
        <tr><td class="field">pharmGKB_Id:</td><td><?php print($pharmgkb_id);
?></td></tr>

</table>

<?php
        page_stop();
?>
```

Appendix B: PharmGKB XML Schema

NOTE: This figure only includes the portions of the schema relevant to COBRA submissions.



Appendix C: Definitions of Database Terms

The following terms are linked to web forms associated with importing data:

Technician(Id/Name)	User ID of laboratory research analyst
Date Ran	Date an assay was performed
Study Name	Official name of clinical trial
Submission Name	Arbitrary name to describe an import event
IRB Number	Internal Review Board approval number
PI Name	Real name of Principal Investigator
Clinical Trials Gov ID	Id number from Clinical Trial (clinicaltrials.gov)
Date Approved	Data the study was approved
Grant ID	Id number for grant
Gene Symbol	NCBI Official Symbol
mRNA Accession	NCBI Accession number
PharmGKB ID	Accession ID for PharmGKB
Variant Type	Type of mutation (SNP, indel, VNTR)
Variant Name	Arbitrary name to describe variation
VNTR Sequence	Actual DNA sequence of tandem repeat
Amino Acid Change	3-letter code changes for amino acids
Interrogated Position	Locus of variation (bp)
Allele 1	Specifies A,C,G,T
Allele 2	Specifies A,C,G,T
RS Id	dbSNP ID number
Source	Source of SNP (dbSNP, HapMap, jSNP)
Assay Type	Type of assay (RFLP, TaqMan, Sybr, Luminex)
Assay Name	Arbitrary name to describe assay
Forward Primer Sequence	DNA sequence of forward primer
Reverse Primer Sequence	DNA sequence of reverse primer
Amplicon Wild Type (bp)	Size of wildtype amplicon
Amplicon Mutation Type	Size of mutation amplicon
Variant	Arbitrary name of variant
Interrogated Position	Locus of variant
Protocol Description	Description of assay protocol
Assay Type Name	Arbitrary name of assay
Local Id	Clinical Pharmacology Id for subject
Race	Race as specified by NIH
Gender	Male or female
Ethnicity	Ethnicity as specified by NIH
Age	Patient's age at start of trial
Sample Type	Type of biological sample (blood, saliva)
TimePoint	Data time point for collection of samples/data
Patient	De-identified number

Appendix D: Curriculum Vitae

curriculum vitae | **REBECCA FLETCHER**

becalyna@gmail.com
(317) 965-0962
4502 Lakeridge Drive
Indianapolis, IN, 46222 USA

Education

Master of Science in Bioinformatics, May 2008
School of Informatics, Indiana University Purdue University at Indianapolis (IUPUI)
Thesis: Genotype Information Management System
Advisor: Sean D. Mooney, PhD

Bachelor of Arts in Spanish for the Professions, May 1998
Marquette University, USA

Research Interests

- Pharmacogenomics
- Genotype/phenotype associations
- Integrating large-scale genomic data

Professional Experiences

Bioinformatics Manager, Division of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN
Aug 2003 – Present

Internet Security Engineer, Symantec Corporation
Nov 2002 – Aug 2003

Web Hosting Engineer, UUNET Technologies, Inc.
Feb 2000 – Nov 2002

Research Assistant, Japan Bank for International Cooperation
Feb 1999 – Feb 2000

Translator/Interpreter, ESL educator, Center for Professional Translation and Interpretation (CETIP), Mexico
May 1998 – Sep 1998