B. Comp Dissertation

SNP Data Integration and Analysis for Drug-Response Biomarker Discovery

By

Chen Jieqi Pauline

Department of Computer Science

School of Computing

National University of Singapore

2008/2009

B. Comp Dissertation

SNP Data Integration and Analysis for Drug-Response Biomarker Discovery

By

Chen Jieqi Pauline

Department of Computer Science

School of Computing

National University of Singapore

2008/2009

Project No: H114910

Advisor: Prof Wong Limsoon

Deliverables:

Report: 1 Volume

Abstract

Technological advances have given clinical researchers and pharmacists various methods to develop drugs targeted at various diseases. To maximize drug viability and profit, one of the greatest challenges for pharmaceuticals would be predicting the body's response to prescribed drugs, and the underlying causes for variation between people with similar illnesses. The causes may range from a patient's genetic makeup to dietary habits. Besides this, it is also an ongoing project to determine drug targets for each drug and maintaining this database. In our work, we focus on SNPs and how they affect drug response. In particular, we target ethnic diversity as an important aspect. We have developed a web-based bioinformatics tool that allows users to search for possible variations that may be significant in determining drug response without any genotype data. In addition, the tool is able to search for drug-enzyme relationships and supplementing current incomplete databases. We see our development as a prototype for future development for use in the pharmaceutical industry in helping to determine viability of drugs for the Singapore community.

Subject Descriptors:

J.3 Life and Medical Science

Keywords:

Genetic variation, Single Nucleotide Polymorphism, Allelic frequencies, Pharmacogenomics, drug response

Acknowledgements

I would like to first thank Professor Wong Limsoon for his guidance and assistance throughout the project. I would also like to thank Professor Chia Kee Seng, the director of the NUS-GIS Centre for Molecular Epidemiology, for the opportunity to work on this project and his valuable comments in the course of development of the tool and Dr YY Teo, a postdoctoral statistician at the Centre for Genomics and Global Health, from the University of Oxford, Oxford, United Kingdom for his contribution on the statistical analysis of the project. Finally, I would like to thank Mr Ku Chee Seng for his guidance in the usage of the International HapMap Project website.

Table of Contents

1. Introduction

In the recent years, there has been an exponentially increasing amount of biological data. This has brought to attention the need for solutions to interpret and utilize them for novel discoveries and inventions. Our project focuses on the field of pharmacogenomics, a promising area in the research and development of drugs. Let us first look at several important keywords.

1.1. Single Nucleotide Polymorphisms (SNPs)

Since the completion of the Human Genome Project in 2001, single nucleotide polymorphisms (SNPs) have been the focus of many researchers. It has been discovered that human beings have 99.9% similarity in genetic data, and of the 0.1% difference, more than 80% have been identified as SNPs.

An SNP is a single base mutation in the genetic sequence, for instance, a mutation from the base adenine (A) to thymine (T), at a particular position on the genome. However, not all point mutations are termed an SNP. They must occur in at least 1% of the general population[17]. There are 2 types of SNP; the biallelic SNPs, with 2 possible alleles for the mutation and the triallelic SNPs, with 3 possible alleles for the mutation. As the occurrence of triallelic SNPs are relatively small, we shall assume that the term SNP refers to biallelic SNPs for the rest of the report unless specified.

SNPs can be found in many regions of the genome, including the coding and non-coding regions of a gene, promoter regions of a gene etc. They are hence potential culprits in altering gene expression and enzyme properties. There are many ongoing efforts in discovering SNPs and their associated biological implications, the most widely known being The International HapMap Project[18]. It is collaboration between many countries to develop a public resource focused on genetic variations and its effects on various biological aspects such as diseases and drug response. Their database consists of SNP linkage disequilibrium data, allele frequency data and genotype data among many others. In Singapore, there is an ongoing project called the Singapore Genome Variation Project[23] aimed at achieving these datasets within

the Singaporean population and supplementing the International HapMap Project database.

1.2. Pharmacogenomics

Our focus, pharmacogenomics, is a discipline to discover genetic factors contributing to variations in drug response, efficiency and toxicity. As mentioned previously, SNPs have various sites of occurrence. To illustrate the importance of SNPs in pharmacogenomics, let us describe a possible situation. Take for instance a drug and its metabolizing enzymes. In the gene coding for one of these enzymes, there exists an SNP which, if takes on the mutant allele, results in a nonsense mutation. The nonsense mutation causes a premature truncation during translation of the gene and produces a non-functional enzyme. The enzyme is now unable to take part in the metabolism of the drug and the drug metabolism pathway is disrupted. The enzyme may be critical in the cycle of the drug, and the result is the lack of response. On the other hand, there is also a possibility that the enzyme does not play a major role in its metabolism, and the result is a minor drop in drug efficiency. There have already been numerous reports on the association between drug response or diseases with certain SNPs[13] and SNPs affecting drug efficacy[2].

Speculations have since been ongoing on "personalized drugs" in pharmacogenomics, drugs designed uniquely for each patient through the study of a patient's genotype[3][12] . The identification of SNP alleles and other mutations would allow drug developers to determine patient response to prescribed drugs. They may then be able to alter the drug's structure to work around any mutant enzymes and create a working drug "personalized" for this particular patient and his genotype.

1.3. Problem statement

The main issue to be tackled is the discovery of SNP biomarkers for drug response. Numerous approaches so far been successful, including candidate gene and linkage disequilibrium mapping studies[10]. Unfortunately, the former requires prior knowledge of drug metabolizing enzymes, and the latter requires high throughput genotyping techniques to be feasible.

The other issue requiring attention is the incompleteness of drug and enzyme association databases. Existing databases such as PharmGKB and DrugBank do not contain in their database such data for all drugs. Candidate gene studies base their genotyping on this information and the study cannot proceed without it.

The rest of the report will be organized as follows. We will first look at the motivation for the development of our system. Next, we will give an overview of our system including its components and functionalities. We will then discuss in details each component and the underlying concepts, design trade-offs and results if any. Finally, we will conclude with comparisons with existing systems and future improvements.

2. Motivation

Ever since the discovery of drugs, they have played an important role in life; treating minor illness such as colds and fevers to major ones such as cancer. Its development is a time and resource consuming process, requiring rigorous animal testing and clinical studies before approval by the FDA[19]. Drug developers hence put their efforts in ensuring maximal efficiency and profitability of a developing drug.

2.1. Discovery of SNPs as biomarkers

For maximum profit, the best situation for a drug would be that it works similarly for everyone regardless of ethnicity. Unfortunately, this ideal situation is not always achieved. Previously, we have demonstrated how SNPs can affect drug response and identified SNPs as potential biomarkers for drug response. Drug response can vary not only with ethnicity, but also with individuals. Let us first illustrate how ethnicity and SNPs play a part in drug response.

Because of the difference in ancestry, an SNP would have evolved differently in populations of different ethnicity. There is often a relationship between allele frequency and population ethnicity. Let us take for instance an SNP within an enzyme that is known to be a drug target. The presence of the mutant allele on this SNP would disrupt the structure of the enzyme and interrupt the metabolic process of the drug in question, rendering it ineffective. Let us now assume population A has a 90% frequency for the mutant allele while population B has a 10% frequency. Since the occurrence of the mutant allele is higher in population A compared to that of B, there is a high chance more people from population A would carry the mutant allele and be non responsive to the administered drug as compared to population B. In summary, the more different the populations are in the alleles of SNPs of enzymes involved in a drug's metabolism, the more likely the populations will respond differently to the drug.

In the case of Singapore, the above issue raises a concern due to her diverse ethnicity. A drug that works for the Chinese but not for Malays would greatly reduce profit. This is the motivation that drives us to develop a system that can identify these SNPs.

Ethnicity is not the only factor in variability of drug response. Within the same ethnic population, the difference in drug response may also be observed, attributing to polymorphisms between different people of the same population. Using the example as before, population A has 10% frequency for the non-mutant allele, and a 90% frequency for the mutant allele. We may also interpret this as approximately 10% of population A will respond to the drug, while 90% of population A will not. This is not limited to only people from Singapore, but encompasses individuals of all populations globally. The identification of these SNPs affecting drug response would greatly help the goal of "personalized drugs".

So far we have only seen the potential of single SNPs as biomarkers for drug response. However, single SNPs may sometimes only contribute minimally to drug response. Instead, it is often the combined presence of several mutant alleles on different SNP sites that have a greater effect on drug response. In other words, the combination of the presence of the mutant allele on SNP1, SNP2, and SNP3 and so on disrupts the drug metabolism and causes non-responsiveness. Similar to single SNPs, the identification of a combination of SNPs would also contribute to the aim of "personalized drugs".

2.2. Discovering drug-enzyme relationships

As stated in our problem statement, there is incompleteness in the databases responsible for a drug and its gene targets. The several existing drug target databases that capture the association between drug and enzyme, DrugBank[21][22] and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database[7] have been identified as our main sources for the development of the system. Unfortunately, they have only captured this information for a subset of drugs, hence resulting in our database lacking this relationship for many drugs.

This information is critical in the discovery of potential genetic mutations that affect a drug's efficacy. It is also used in candidate gene studies to determine which genes to genotype. As we have seen in our motivation for discovering SNPs as biomarkers, we must first know the drug targets as a precursor to proceed.

2.3. Integrated database

We have also considered the need for an integrated database as information is widely spread out over the net. Currently, the search for SNPs related to a drug requires access to various websites. The implementation of a combined database allows linking of information, and overcomes this need for website hopping.

The development of this system aims to help researchers in determining potential SNP biomarkers, and also allow ease of analysis on available data from various sources.

3. Overview of System

We have seen the driving forces motivation behind in the development of a system that performs SNP analysis. In this section, we will give a quick outline of the system and its components. The details of each component will be elaborated in later sections.

3.1. Software

The system developed is a browser based tool developed with the web-programming language PHP. Implementation was done using PHP 5.2.8 and MySQL version of 5.1.30. The system runs on a local server, hence an Apache server is required.

3.2. Hardware

The system was developed and run on a 2.53 GHz processor with Intel® Core™2 DUO central processing unit, 4GB Random Access Memory (RAM) on Windows Vista™ Business.

3.3. Main components

There are 2 main components to the system, each designed to tackle a specific problem statement. Let us look the general function of each component.

3.3.1. SNP drug response biomarker discovery

There are 3 subcomponents to SNP biomarker discovery. The first searches for allelic dissimilarities that may lead to a difference in drug response between different populations. The second searches for single SNPs that may be potential biomarkers given the drug response data. The third searches for a set of SNPs as potential biomarkers given genotype and drug response data.

3.3.2. Drug to enzyme association discovery

This component allows users to discover drug and enzyme associations and supplement the system's own database.

3.4. Database design and sources

We developed a database that integrated data from various sources. Let us now look at the design of the database and how each table was derived. The database was built on MySQL, a relational database management system (RDBMS).

3.4.1. Abstract articles from PubMed 2008

A total of 8059 abstracts were taken from PubMed's 2008 articles. The abstracts were obtained in a CD in .nxml format, and then parsed with a PHP script. The details of each abstract include the PubMed ID (PMID), a unique identifier for articles published in PubMed, the title of abstract, the journal the abstract is from and the abstract body. These abstracts were first obtained to test the concepts behind the discovery of enzyme and drug association and have been retained for the user's reference in confirmation of these associations.

3.4.2. Drugs

A total of 3812 drugs' data were abstracted from the PharmGKB[6] database. The information is downloadable as a text file and was parsed with a PHP script. The details of each drug include the PharmGKB ID, a unique identifier for drugs, the name of the drug, and alternative names for the drug.

3.4.3. Genes

A total of 23738 genes in the human genome was abstracted and integrated from both the PharmGKB[6] database and International HapMap[18] database. The details of each gene include the entrez ID, a unique identifier for genes determined by Entrez Gene, the gene symbol, gene name, alternate names of the

gene, alternate symbols of the gene, the start position of the gene on the chromosome, the end position of the gene on the chromosome and the chromosome the gene resides in.

3.4.4. Single Nucleotide Polymorphisms

A total of approximately 3.5 million SNPs identified by dbSNP and International HapMap[18] reside in this table. The details of each SNP include the reference cluster ID (RSID), a unique identifier for SNPs, the position of the SNP on the chromosome, the alleles for the SNP, the entrez ID of the gene the SNP resides in and also the chromosome of the SNP. It is important to note that although the chromosome can be determined with the entrez ID of the gene the SNP resides in, not all SNPs can be mapped into a gene.

3.4.5. Single Nucleotide Polymorphism frequency data

Allele frequencies determined by the International HapMap project are readily available for download. Currently, they have released this data for 11 populations, each with approximately 1.5 million genotyped SNPs. The data was parsed with JAVA and inserted into the database. This brings up to about 16 million rows of frequency data in this table. The details of each SNP frequency data include the RSID of the SNP, the population race genotyped, the allele frequencies of A, T, G and C. The RSID references to the RSID in the table for SNPs.

3.4.6. Single Nucleotide Polymorphism race data

Because of the vast amount of data for SNP frequencies, post processing was done to produce a list of distinct populations that have been genotyped to reduce time required to extract population from the frequency table.

3.4.7. Association between drug and enzyme

Data was downloaded from PharmGKB[6] and DrugBank[21][22] and parsed with a PHP script. They contain the details of metabolic pathway enzymes and drug targets for each drug, if available. The details of each association contain the PharmGKB ID of the drug, and the entrez ID of the enzyme associated with the drug. This acts as a bridge to link drug and enzyme information. As mentioned in section 2, this database is incomplete. User confirmed drug and enzyme associations will hence supplement this database.

3.4.8. Association between drug and PubMed literature

Associations between a drug and PubMed literature PMID is stored in this table. The derivation of this table was done both through online searching and also MySQL querying of the abstracts table from section 3.4.1. using simple MySQL LIKE operation.

As for online searching, this was done with Google using a drug name as the search criteria limited to hits from the PubMed website automatically with all drugs within the database and each search result page for a drug is parsed for all PubMed literature PMID returned. The association between a drug and all its found PubMed literature PMIDs are added into this table. The literature is currently limited to only those published in February 2009.

3.4.9. Association between enzyme and abstract

Associations between an enzyme and PubMed literature PMID is stored in this table. The derivation of this table was done both through online searching and MySQL querying of the abstracts table from section 3.4.1. using simple MySQL LIKE operation.

As for online searching, this was done with Google using an enzyme name as the search criteria limited to hits from the PubMed website automatically with all genes within the database and each search result page for a gene is parsed for all

PubMed literature PMID returned. The association between a gene and all its found PubMed literature PMIDs are added into this table. The literature is currently limited to only those published in February 2009.

3.4.10. Association between drug and enzyme

Associations between a drug, an enzyme and literature abstract ID is stored in this table. This table was derived by intersecting the tables described in section 3.4.8. and 3.4.9. The basis for this intersection is that an association exists between a drug and an enzyme if both are found in the same biomedical literature. This table also keeps the status of the association of its acceptance status by the user.

4. Discovery of SNPs as biomarkers for drug response

In this section, we shall focus on how the system was implemented and its underlying concepts to discover SNPs as biomarkers for drug response.

4.1. Identifying SNPs with allele similarities or dissimilarities

We have already discussed the desirability that a drug works equally effectively on all populations. One way to check this is based on the concept earlier discussed. That is, we look at the enzymes targeted by the drug and the more different the populations are in the alleles of SNPs of these enzymes, the more likely the populations will respond differently to the drug. A commonly used statistic to test this situation is the F_{ST}, or fixation index[16]. We shall now describe the details for calculation of the statistic.

In the first step, we calculate the local expected heterozygosity (H_{exp}) or gene diversity for each population. Let us assume a biallelic SNP with alleles A and B and frequencies $F_{A,i}$ and $F_{B,i}$ respectively for a population i for all populations.

$$H_{exp,i} = 1 - (F_{A,i}^2 + F_{B,i}^2)$$

Next, we calculate $\bar{p}$ and $\bar{q}$, the frequency of allele A and B respectively over the total population. n_i represents the tested population size for population i.

$$\bar{p} = (\Sigma_i F_{A,i} \times n_i) / \Sigma_i n_i$$

$$\bar{q} = (\Sigma_i F_{B,i} \times n_i) / \Sigma_i n_i$$

In the third step, we calculate the global heterozygosity indices over subpopulations, H_S.

$$H_S = (\Sigma_i H_{exp,i} \times n_i) / \Sigma_i n_i$$

In the fourth step, we calculate the global heterozygosity indices over the total population, H_T.

$$H_T = 1 - (\bar{p}^2 + \bar{q}^2)$$

Finally, we calculate the fixation index, F_{ST}, using the following formula:

$$F_{ST} = (H_T - H_S) / H_T$$

This gives us the F_{ST} value for this SNP. There are 4 categories of interpretation for the F_{ST} value, and they are as follows:

F_{ST} = 0 to 0.05: little genetic differentiation

F_{ST} = 0.05 to 0.15: moderate genetic differentiation

F_{ST} = 0.15 to 0.25: great genetic differentiation

F_{ST} Above 0.25: very great genetic differentiation

In other words, an SNP with a F_{ST} value of 0.03 indicates high similarity of alleles for this SNP over the tested populations, and is less likely to result in different drug response between populations. On the other hand, an SNP with a F_{ST} value of 0.5 indicates high dissimilarity of alleles for this SNP over the tested populations, and is likely to cause variation in drug response between populations.

Let us look at an example to illustrate the usage of $F_{ST.}$ We consider the drug exemestane and the enzyme CYP1A1 which is involved in its metabolism. Within the CYP1A1 gene, we take the SNP rs4986879 and the frequencies of populations African, Luhya, Maasai and Yoruban (Table 1).

Race	A	T	G	C
African	0	0.962	0	0.038
Luhya	0	0.956	0	0.044
Maasai	0	0.983	0	0.017
Yoruban	0	0.96	0	0.04

Table 1: SNP allele frequencies for rs4986879

We assume equal population size for all populations. We first calculate H_{exp} for all populations:

$$H_{exp,African} = 1 - (0.962^2 + 0.038^2) = 0.0731$$

$$H_{exp,Luhya} = 0.0841$$

$$H_{exp,Maasai} = 0.0334$$

$$H_{exp,Yoruban} = 0.0768$$

Next, we calculate $\bar{p}$ and $\bar{q}$ over the total population:

$$\bar{p} = (0.962 + 0.956 + 0.983 + 0.96) / 4 = 0.9653$$

$$\bar{q} = (0.038 + 0.044 + 0.017 + 0.04) / 4 = 0.0347$$

We then calculate H_S and H_T:

$$H_S = (0.0731 + 0.0841 + 0.0334 + 0.0768) / 4 = 0.0669$$

$$H_T = 1 - (0.9653^2 + 0.0347^2) = 0.0670$$

And finally, we find F_{ST}:

$$F_{ST} = (0.0670 – 0.0669) / 0.0670 = 0.00149$$

The value of $F_{ST} < 0.05$, hence we can conclude that there is little genetic differentiation over all 4 tested populations. In other words, there is a high similarity of alleles for this SNP and is not likely to result in different drug response.

This analysis is done on all SNPs for all drug targets for a particular drug of interest. The final result is a list of SNPs and their respective F_{ST} values. The user is given a choice to reduce the scope to only those falling in 1 of the 4 categories of his interest. To conclude, the more SNPs discovered to have significant allelic dissimilarities over populations, the more likely this drug will have a variable response between them.

4.2. Identifying potential single SNP biomarkers for drug response

We have seen how the system identifies SNPs with allele dissimilarities over different populations that could be significant in variability in drug response using the F_{ST}. Let us now look as another aspect of SNP biomarker discovery, discovering single SNPs associated with provided data on drug response.

As discussed in our motivation, there is observed drug response variation within the same ethnic population and even between different populations. We hence wish to search for these SNPs that are responsible. One way to check this is based on the following concept. We look at the enzymes targeted by the drug. The more similar the allele frequencies of an SNP in these enzymes are with the observed drug response frequency, the more likely the allele is associated with drug response. One way to test this for this is the chi-square goodness-of-fit test[11]. Let us first look at how we can determine this specific to a single population.

We first state the null (H_0) and alternative (H_1) hypothesis:
H_0: Any deviation of the observed and expected drug response is due to chance.
H_1: Any deviation of the observed and expected drug response is not due to chance.

Let us assume an SNP, with non-mutant allele A and mutant allele B with allele frequencies F_A and F_B respectively in a population. There are 2 possible cases for association; that the non-mutant allele A does not alter drug response while mutant allele B results in non-responsiveness. Similarly, the other way round is also possible; that mutant allele B does not alter drug response while non-mutant allele A results in non-responsiveness.

Let us first assume: the non-mutant allele A does not alter drug response while mutant allele B results in non-responsiveness. Let the observed values for drug response be O_R and lack of response be O_{NR}. The expected values for drug response (E_R) and no drug response (E_{NR}) may then be calculated using F_A and F_B following the above assumption.

$$E_R = F_A * (O_R + O_{NR})$$
$$E_{NR} = F_B * (O_R + O_{NR})$$

The chi-square statistic is used to determine closeness in observed and expected values. This is calculated using O_R and O_{NR} with the following formula:

$$\chi^2 = (O_R - E_R)^2 / E_R + (O_{NR} - E_{NR})^2 / E_{NR}$$

The chi-square value is then used to calculate the p-value under the chi-square distribution. The p-value is defined as the probability that the observed deviation from the expected value can be explained by chance alone. Hence, the larger the p-value, the more likely the deviation from the expected value is insignificant and explainable by chance and thus the more likely the null hypothesis is true. The p-value is calculated with the following equation:

$$F(x;k) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} = P(k/2, x/2)$$

, where k is the degrees of freedom, $\gamma(k,z)$ is the lower incomplete Gamma function and $P(k,z)$ is the regularized Gamma function.

The user is free to choose a significance value in the tool, but it is commonly set at 5%. If the p-value is found to be greater than 5%, H_0 is then accepted. In other words, if there is greater than 5% probability that the deviation can be explained by chance alone, we conclude that the deviation between observed and expected frequency is insignificant and the SNP is selected as a potential biomarker for the individual tested population.

The calculation is done similarly for the alternative situation, where mutant allele B does not alter drug response while non-mutant allele A results in non-responsiveness.

Let us now look at a demonstration of the mathematical analysis with an example. We consider again the drug exemestane and the enzyme CYP1A1 which is involved in its metabolism. Within the CYP1A1 gene, we take again the SNP rs4986879 with alleles C and T, and the frequencies of populations African, Luhya, Maasai and Yoruban (Table 1).

For illustration, let us assume drug response data (Table 2):

Race	Response	No Response
African	900	100
Luhya	950	50
Maasai	700	300
Yoruban	40	960

Table 2: Assumed dataset supplied by user

We restate the null and alternate hypothesis:

H_0: Any deviation of the observed and expected drug response is due to chance.

H_1: Any deviation of the observed and expected drug response is not due to chance.

We first assume the allele T does not alter drug response while allele C results in non-responsiveness. Using the formula for calculation of expected response statistics, we derive the expected drug statistics for each population for each allele. For example, Africans with allele T would have an expected drug response statistic of 96.2% x (900+100) = 962 (Table 3).

	Expected	Observed
African (Allele:T), Respond to drug	962	900
African (Allele: C), No response to drug	38	100
Luhya, (Allele:T), Respond to drug	956	950
Luhya, (Allele: C), No response to drug	44	50
Maasai, (Allele:T), Respond to drug	983	700
Maasai, (Allele: C), No response to drug	17	300
Yoruban, (Allele:T), Respond to drug	960	40

Yoruban, (Allele: C), No response to drug	40	960

Table 3: Calculated expected and observed values for assumption that allele T does not alter drug response

For each population separately, we calculate χ^2 and p-value using 1 degree of freedom:

For Africans, $\chi^2 = (900-962)^2/962 + (100-38)^2/38 = 105.15$
P-value = P ($\chi^2 > 105.15$) ≈ 0
For Luhya, $\chi^2 = 0.8558$
P-value = P ($\chi^2 > 0.8558$) ≈ 0.355
For Maasai, $\chi^2 = 4792.59$
P-value = P ($\chi^2 > 4792.59$) ≈ 0
For Yoruban, $\chi^2 = 22041.67$
P-value = P ($\chi^2 > 22041.67$) ≈ 0

Using a significance level of 90%, we observe that none of the p-values calculated exceed this threshold. We hence reject the null hypothesis, and conclude that the deviation is not due to chance; hence this SNP is not a potential biomarker for any of the tested populations for drug response under the assumption that allele T does not alter drug response.

Let us now assume the alternative: that the presence of allele C does not alter drug response and similarly calculate the respective expected values. The results have been summarized:

	Expected	Observed
African, (Allele:T), No response to drug	962	100
African, (Allele: C), Respond to drug	38	900
Luhya, (Allele:T), No response to drug	956	50

Luhya, (Allele: C), Respond to drug	44	950
Maasai, (Allele:T), No response to drug	983	300
Maasai, (Allele: C), Respond to drug	17	700
Yoruban, (Allele:T), No response to drug	960	960
Yoruban, (Allele: C), Respond to drug	40	40

Table 4: Calculated expected and observed values for assumption that allele C does not alter drug response

For each population separately, we calculate χ^2 and p-value using 1 degree of freedom:

For Africans, $\chi^2 = (900-38)^2/38 + (100-962)^2/962 = 20326.18$
P-value = P ($\chi^2 > 20326.18$) ≈ 0
For Luhya, $\chi^2 = 19513.98$
P-value = P ($\chi^2 > 19513.98$) ≈ 0
For Maasai, $\chi^2 = 27915.09$
P-value = P ($\chi^2 > 27915.09$) ≈ 0
For Yoruban, $\chi^2 = 0$
P-value = P ($\chi^2 > 0$) = 1.0

Using a significance level of 90%, we discover the p-value for Yoruban exceeds the threshold. We hence accept the null hypothesis for the Yoruban population, and conclude that the deviation is due to chance; hence this SNP is a potential biomarker for drug response for the Yoruban population under the assumption that allele C does not alter drug response. In other words, we may predict that Yorubans with allele C in this SNP will respond to the drug, while Yorubans with allele T will have no response.

We have just seen how SNPs specific to a single population can be determined using the chi-square goodness-of-fit test. Let us now look at how we can search for an SNP as a biomarker that is consistent over all the populations. That is, this is a general SNP that can determine drug response regardless of ethnicity. We still use the same concept; the more similar the allele frequencies are in these enzymes with the observed drug response frequency, the more likely the allele is associated with drug response. The difference between this method with that mentioned above is that the above tests for the hypothesis separately for different populations. The method we will now look at tests for the hypothesis for all populations as a single test.

We restate the null and alternate hypothesis:
H_0: Any deviation of the observed and expected drug response is due to chance.
H_1: Any deviation of the observed and expected drug response is not due to chance.

For each population with given observed data, we determine each population's expected values E_R and E_{NR} using each population's respective O_R and O_{NR}. The chi-square values for each population are then added up and the p-value is calculated. We restate that the p-value is defined as the probability that the observed deviation from the expected value can be explained by chance alone. Depending on the number of populations with available frequency or observed data (N), we determine the degrees of freedom (k) for calculation of the p-value as:

$$k = 2 * N - 1$$

Again, if the p-value is found to be greater than 5%, H_0 is then accepted. In other words, if there is greater than 5% probability that the deviation can be explained by chance alone, we conclude that the deviation between observed and expected frequency is insignificant and the SNP is selected as a consistent potential biomarker for all populations tested.

We illustrate the calculation using the same example from above. We have already calculated the expected values for all situations (Table 3 and 4). This data is used again for the calculation of chi-square values for this test.

Under the assumption that allele T does not alter drug response and with (2*4 – 1 =) 7 degrees of freedom (Table 3):

$\chi^2 = 105.15 + 22041.67 + 4792.59 + 0.8558 = 26940.265$
P-value = P ($\chi^2 > 26940.265$) ≈ 0

Under the assumption that allele C does not alter drug response and with 7 degrees of freedom (Table 4):

$\chi^2 = 67755.25$
P-value = P ($\chi^2 > 67755.25$) ≈ 0

Using a significance level of 90%, we may conclude that the SNP is not a potential general biomarker in either situations in determining drug response and we reject the null hypothesis for both.

It is important to note that although the p-value threshold is commonly set at 5%, we have set our threshold to be 90%. In other words, we are searching for SNPs such that the deviation of the expected from the observed drug response statistics has greater than 90% probability of being explained by chance. Hence this makes it a lot harder to accept the null hypothesis, and is stricter in determining SNP biomarkers.

We have seen how to determine population-specific SNPs and population-consistent SNPs as biomarkers for drug response. This approach overcomes the need for genotype data, which may not always be available, for candidate gene studies. We are currently unable to perform validation tests on the methodology due to lack of data.

There is an important issue to consider when doing this test. In the special cases where calculated expected values for drug response or non-responsiveness fall below 5, the chi-square statistic will be inaccurate as the small expected value amplifies the chi-square value. We have implemented 2 ways to work around the issue, one using the Yate's continuity correction for population-specific SNP biomarker discovery and the other, combining values such that the expected value does not fall below 5 for general SNP biomarker discovery.

4.2.1. Yate's Continuity Correction

The Yate's continuity correction aims to reduce the chi-square value and hence increases the p-value to prevent overestimation of statistical significance[15] for small expected values. The calculation of chi-square values will then use the formula:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where,

O_i = frequency of observed value

E_i = frequency of expected value

N = number of events

To illustrate this mathematical concept, let us use the previous example and focus only on the African population. However, let us reduce the number of observed data to illustrate Yate's continuity correction:

Race	Response	No Response
African	90	10

Table 5: Reduced observed data for African population

Once again, we restate the null and alternate hypothesis:

H_0: Any deviation of the observed and expected drug response is due to chance.

H_1: Any deviation of the observed and expected drug response is not due to chance.

We assume the allele T does not alter drug response while allele C results in non-responsiveness. Using the formula for calculation of expected response statistics, we derive the expected drug statistics for Africans for each allele (Table 6). For example, using the frequency values in Table 3 for Africans, allele T would have an expected drug response statistic of 96.2% x (90+10) = 96.2 (Table 3).

	Expected	Observed
African, (Allele:T), Respond to Drug	96.2	90
African, (Allele: C), No Response to Drug	3.8	10

Table 6: Calculated expected data assuming allele T does not alter drug response

The original formula for χ^2 used in section 4.2 would return the result as follows:

$\chi^2 = (90\text{-}96.2)^2/96.2 + (10 - 3.8)^2/3.8$
$= 10.515$
P-value = $P(\chi^2 > 10.515) \approx 0.0011$

However, if we use the Yate's continuity correction, the χ^2 value will then be calculated as follows:

$\chi^2 = (|10/100 - 0.038|\text{-}0.5)^2/0.038 + (|90/100 - 0.962|\text{-}0.5)^2/0.962$
$= 5.376$
P-value = $P(\chi^2 > 5.376) \approx 0.0204$

We observe that there is a significant difference between both the calculated χ^2 value and the resulting p-value. Even though both fail to pass the p-value threshold of 90%, leading to the conclusion that this SNP is not a potential biomarker, there is an increase in p-value for the chi-square value calculated under the Yate's continuity correction and potential to accept the null hypothesis.

Unfortunately, the Yate's continuity correction tends to overcorrect and result in a conclusion that does not reject the null hypothesis when it should. This issue can be solved using the randomization test for goodness-of-fit.

4.2.2. Combining response data

We have seen our approach to dealing with small expected values for determining SNP biomarkers for specific populations. We now give our approach to dealing with small expected values for determining SNP biomarkers for the general population.

Let us assume a new set of response data as follows:

Race	Response	No Response
African	90	10
Luhya	95	5
Maasai	70	30
Yoruban	4	96

Table 7: Set of drug response data

Once again, we restate the null and alternate hypothesis:
H_0: Any deviation of the observed and expected drug response is due to chance.
H_1: Any deviation of the observed and expected drug response is not due to chance.

We then calculate the expected values for all populations and summarize the results (Table 8). Let us again assume that allele T does not alter drug response.

	Expected	Observed
African, (Allele:T), Respond to Drug	96.2	90
African, (Allele: C), No Response to Drug	3.8	10
Luhya, (Allele:T), Respond to Drug	95.6	95
Luhya, (Allele: C), No Response to Drug	4.4	5
Maasai, (Allele:T), Respond to drug	98.3	70

Maasai, (Allele: C), No Response to Drug	1.7	30
Yoruban, (Allele:T), Respond to Drug	96.0	4
Yoruban, (Allele: C), No Response to Drug	4.0	96

Table 8: Calculated expected and observed values for assumption that allele T does not alter drug response

Note that there are several cells in the expected values that fall below 5 and performing chi-square test analysis is likely to produce inaccurate results, as illustrated for single population studies illustrated in section 4.2.1. Our approach is to combine all the values for response from all populations into a single value for observed and expected values separately, and then do the same for non response values. This will remove all cells with values less than 5, and we can perform the chi-square test analysis. This means that the expected drug response statistic will be 96 + 98.3 + 95.6 + 96.2 = 386.1.

	Expected	Observed
(Allele:T), Respond to Drug	386.1	259
(Allele: C), No Response to Drug	13.9	141

Table 9: Calculated expected data assuming allele T does not alter drug response

We can then calculate the chi-square value. Note that after combining the bins, the degree of freedom is now 1:

$\chi^2 = (259\text{-}386.1)^2/386.1 + (141\text{-}13.9)^2/13.9 = 1204.028$

P-value = $P(\chi^2 > 1204.028) \approx 0$

Using a significance level of 90%, we reject the null hypothesis.

Although this is a known solution to the issue of small frequency values, there is also the issue of Simpson's paradox[5]. The issue of lurking variables, in our case it could be drug dosage or patient dietary habits, were not taken into consideration during the calculation. Hence this could have lead to a false conclusion. An alternative to prevent this is to similarly use the randomization test of goodness-of-fit, which we will elaborate in section 7.2.

4.3. Search for combination of SNPs as biomarkers

We have seen how single SNPs may affect an individual's response to a drug. We have also seen our technique developed for searching single SNPs as biomarkers for drug response. A single SNP, as we have illustrated in previous sections, has potential effects on drug response. However in some cases, a single SNP mutant in one enzyme or drug target only has little effect considering the many others involved in drug metabolism. When a combination of mutations is present over various sites in various drug targets, there is a greater probability of effect on drug metabolism and efficacy. In other words, the presence of a mutation in SNP1, SNP2, SNP3 and SNP4 in drug targets E1, E2, E3 and E4 respectively is more likely to result in a significant difference in drug response than only the presence of mutation on SNP1 in drug target E1. Let us look at how we can determine these combinations of SNPs that may be biomarkers for drug response.

We utilize the algorithm from Liu G.M, Wong L.M. and Li J.Y.'s paper on "Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns"[9]. The concept of equivalence classes, or a set of frequent itemsets that always occur together in some set of transactions, is utilized in the algorithm. An equivalence class, in SNP biomarker discovery, can be interpreted as a set of frequently present SNPs that always occur together in the same set of drug response category. The algorithm mines closed patterns, the maximal itemset of all equivalence classes, and generators, the minimal itemsets of all equivalence classes. Another important concept is that of delta-discriminative equivalence classes. An equivalence class is delta discriminative if every itemset it encompasses occurs in only one of the classes with almost 0 occurrences in other classes.

For this functionality to be available, the user must have 2 sets of information available. One set contains patient information including the patient's ID and the patient's response to a drug. The other set contains information including the patient's ID, genotyped SNP's RSID, and the presence of mutation at the SNP. Unfortunately, this information may not always be available.

Given this set of data, we can derive a file that has the following format. Each row in the file represents a patient's data, with the first column representing the response of the patient to the drug. The remaining data contains the set of RSIDs of SNPs that have been found to contain the mutant allele in the patient's genotype. This file is then fed into the algorithm for mining of SNPs as biomarkers. The user has to however provide 2 parameters, which are the minimum support threshold and a delta value. The minimum support threshold defines the lowest number of supporting evidence dataset over all classes, while the delta value defines minimum number of supporting evidence in all minority classes.

The algorithm, in brief, first mines frequent closed patterns and generators to represent all frequent equivalence classes given the dataset. Next it determines the class label distribution information, and then uses a test statistic (chi-square, risk ratio etc) to define a score for every closed pattern. Finally, it ranks all the closed patterns and outputs equivalence classes having closed pattern scores above specified significance thresholds.

The final result is a set of generators and closed patterns that satisfy the minimum support threshold and delta value. A smaller set of SNPs would be beneficial compared to a large set as this reduces genotyping site required to determine drug response. Hence, we take the file containing the set of generators and parse the result for display to the user. Each data row in the file describes the number of items in the generator, the set of SNPs in the generator and the support in each of the different classes of drug response. The user may then interpret the set of SNPs in the generator as a potential biomarker for drug response. Unfortunately, we are unable to perform validation tests due to lack of data.

5. Drug to Enzyme Association Discovery

We have seen in section 4.1 that it is desirable to determine if all the different populations will respond to a drug in the same way. We have also seen in sections 4.2 and 4.3 the techniques in discovering SNP biomarkers. To do all these 3 studies, we need to know the drug targets or metabolizing enzymes for each drug in question to determine the regions of SNPs with potential effect. Currently, we are using PharmGKB as our main source and DrugBank as our secondary source of this information. However, these databases are incomplete on this relationship and there is also no upper bound guarantee on the delay of such a drug-enzyme relationship being added into the database. It is hence desirable to develop techniques to process the latest biomedical literature to automatically identify drug-enzyme relationships as a way to supplement the information in PharmGKB and DrugBank with the latest discoveries.

We are currently utilizing the online PubMed library as a source for biomedical literature abstracts. PubMed has a large database of literature dating as far back as 1948, and is a highly reliable source.

We can assume an association between a drug and an enzyme by using the following concept. We locate all literature abstracts in PubMed found to contain the drug's name for a drug. Then, we locate all literature abstracts in PubMed found to contain the enzyme's name for an enzyme. Finally, we associate an enzyme and a drug by considering the intersection between these 2 lists on the ID of the literature abstract. In other words, an association exists between a drug and an enzyme if they are both located in the same biomedical literature abstract. We have already described the tables and their derivation in sections 3.4.8., 3.4.9. and 3.4.10. Currently, the search has only been limited to biomedical literatures published in the month of February of 2009.

To illustrate this concept, let us assume the following situation. We searched for the drug "exemestane" on Google for hits within the PubMed website, and it returned 3 literature abstracts with PMIDs 146389, 125937 and 158379. We searched for the enzyme CYP1A1 (cytochrome P450, family 1, subfamily A, polypeptide 1) similarly on Google for hits within the PubMed website, and it returned 5 literature abstracts with PMID 21748, 158379, 166994, 178503 and 103829. The intersection of these 2 results returns

the PMID 158379, and hence we can associate the drug exemestane with the enzyme CYP1A1 with support from the PubMed literature abstract with PMID 158379.

With the selection of this function by the user, the system returns 20 items at a time from the list of unconfirmed associations. Each item specifies the drug name, the enzyme name, the PMID of the supporting biomedical literature abstract from PubMed, the actual abstract (if available in the local database), and a choice of whether to accept or reject this association. The user is given the responsibility to determine the correctness of the association through study of the given supporting literature abstracts. Choosing to accept the association would supplement the drug and gene pair into the drug target database. Choosing to reject the association would remove it from the list of unconfirmed associations.

As there are numerous possible combinations of drug and gene relationships, we have come up with several tactics to reduce the list to conform to the user's interest. In the event the user wants to only view associations of a drug or gene of interest, they are given a search function to refine the list. In addition, the user may choose to search for drug and gene associations that have a specified minimum number of literature supports that could indicate extensive study and higher chance of association the threshold is high.

For greater association strength and reduction of scope, we have also taken into consideration the Jaccard index of an association. The Jaccard index is a statistic used to compare similarity and diversity of sample sets. It is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

For calculation of the Jaccard index for drug and gene associations, we take A to be the set of drug and literature links for a particular drug, while B to be the set of gene and literature links for a particular gene. The intersection of A and B returns the list of PMIDs that relates the drug and gene, while the union returns the total list of distinct PMIDs that captures any association for the drug and for the enzyme. The ratio of their cardinality then returns the Jaccard index of the particular drug and gene association. We may interpret the Jaccard index as the strength of association between a drug and gene, and the

higher the Jaccard index, the higher the probability of correctness for the association. For example, in the previous example we saw how intersecting 2 sets gave an association linking exemestane and CYP1A1. The Jaccard index for this association is J(A,B) = 1 / 7, as there is only 1 intersection object, and 7 distinct total PMIDs from both sets.

Let us now look at some results in our identification of drug-enzyme relationships. To verify the validity of our association mining, we have taken the list of associations existing in current databases and compared them to those we have found. There are currently 14 drug and gene associations overlapping. We take all the overlapped associations and reviewed the literatures supporting their discovery. Here are the results:

Drug	Gene	Total Supporting Literature	Accurate Supporting Literature
Warfarin	CYP2C9	1	1
Warfarin	VKORC1	2	2
Azathiprine	TPMT	1	1
6-mercaptopurine	TPMT	3	3
Acetylcholine	CHAT	2	0
Azathioprine	TPMT	1	1
6-thioguanine	TPMT	1	0
Glucocorticoids	IL-8	1	0
Simvastatin	LPL	1	1
Rivastigmine	BCHE	1	1
Norepinephrine	TH	1	0
Levodopa	MAOB	1	0
Dopamine	COMT	1	1
Dopamine	MAOA	2	0
Dopamine	MAOB	2	0

Table 10. Results of known gene and drug associations

In summary, there are a total of 8 accurately discovered associations. This is a small sample size as we have only considered literature from February 2009. However, review

of the supporting literature has proven that this is a potential approach to abstracting all associations between a drug and a gene.

Let us now look at drug and gene associations that are not in the major databases. Here are a few examples:

Drug	Gene	Total Supporting Literature	Accurate Supporting Literature
Valproic Acid	HDAC6	2	1
Valproic Acid	HDAC1	4	3
5-fluorouracil	CDK4	1	1
Acetylcholine	IL1R1	1	0

Table 11. Accuracies of unknown gene and drug associations

Unfortunately, the correctness of the association discovered is subjective for gene-enzyme relationships not in the existing databases. User interpretation of biomedical literature abstracts may vary between individuals. In our testing, we assume the supporting evidence to be accurate as long as a relationship can be inferred between the enzyme and a drug.

This implementation has allowed user intervention for augmenting data collected from the DrugBank and PharmGKB website. Users will not have to manually go through all recently published literature to discover drug-enzyme relationships not yet updated in the database without knowing the location of relevant drug or enzyme information in the literatures. It is important to note that, however, the update is done on the local database, and not on the actual PharmGKB database.

However, there are several drawbacks for this implementation. The first is the mining of false associations when the gene and drug have the same name or overlapping names. As we rely on the name as criteria for our search, the search on both the drug and gene would hit the same literature abstracts and an association would definitely be formed for such pairs of genes and drugs. For example, the drug with the name ‘Statin’ would be associated with the gene ‘Statin’. Drugs with common names such as “amino acids” have also proved to be a problem since it is not a unique drug name like exemestane. These

common terms return a list of search hits highly unlikely to be related to its function as a drug. The way around this was to post process the list of associations and remove these instances.

Another drawback of this implementation is false search hits from Google as a result of links on the PubMed website. When searching for a term, Google returns a hit on a page if the term can be found somewhere in the page. The pages for PubMed for literature abstracts not only display the abstract body, but also have a feature that displays a list of related articles to the current article. Google returns the page when the search criteria can be found in the names for these links even when it's not in the abstract body, resulting in false associations when we intersect these lists.

6. Related work

The Drug-SNP DB[8] project was developed with the aim of integration of data to allow users to search for SNPs related to a drug. The database utilizes information from dbSNP[14] and gene information from GenBank[1]. The basis of the development is similar to our system, as both allow users to search for SNPs related to a drug. Their system returns the full list of SNPs within the enzymes that interact with the drug of interest as for drug response study. In the process of searching for significant SNPs, the user may choose to view all SNPs within each enzyme for our system, providing a similar function to that of Drug-SNP DB. Our system, however, further takes into consideration actual drug response data and reducing the list of all SNPs to those found to be potential biomarkers and has greater functionality.

7. Improvements and expansions

We have seen the motivations behind our development for the system and its potential contributions in the field of pharmacogenomics. After which we've described and illustrated our approach to answer the issues prevalent in the field and discussed their results and drawbacks. The system will possibly be an ongoing project to finally develop a locally suited prototype, hence there are many aspects for potential development.

7.1. Supplementing drug enzyme relationship database

We have seen in section 5 how we have discovered drug-enzyme relationships for augmenting the current database using online biomedical literature. The concept is based on simple association, which, as discussed, has lead to several false associations. The feature can be improved by performing a sentence based search; that is, only associate a gene and a drug when they appear in the same sentence. This would remove the drug or gene search hits resulting from links to other abstracts in the PubMed website and hence higher confidence in association.

Also mentioned was that these associations have been discovered from literature in the second half of 2008 and that published within February 2009. The search may now be extended to all literatures to include earlier literatures. However, due to our dependence on Google, this should be done in phases instead of in a single shot to avoid Google's protective feature while continuously searching for all drugs and all genes. An additional feature that searches for these associations automatically every month can be implemented so that more recent biomedical discoveries can be added.

7.2. Randomization test of goodness-of-fit[11]

As mentioned in sections 4.2.1 and 4.2.2., Yate's continuity correction and the combination of values could lead to false conclusions. A more appropriate method to implement would be to use the randomization test of goodness-of-fit. The test works by randomly generating possible combinations of observed data and determining the chi-square statistic for each result. Using the original chi-square statistic calculated from the original observed statistics, we do the following test: if the randomization

test produces a chi-square value greater than or equal to the original chi-square less than α% of the time, we can reject the null hypothesis at level of significance α. It is important to note that the p-values for each result vary because of randomization, and the accuracy increases as the number of randomized samples taken increases.

7.3. Discovering combinations of SNPs as biomarkers

As mentioned in section 4.3, it is usually a combination of SNPs that contribute greatly to drug response. However, our project has so far mainly dealt with single SNPs as a potential biomarker when given only drug response data. In addition, the assumption of absolute correlation between the drug response and a single allele is insufficient. Using the allele frequencies available, we could derive the expected frequencies of various combinations of SNP alleles. Once again, using the chi-square statistic, we could then determine if such a combination could be significant in determining drug response. Unfortunately, SNPs are not independent of one another. However, with consideration of linkage disequilibrium data, we could newly define how to calculate predicted genotype frequencies.

7.4. Integration of SGVP

We have been informed that the data from SGVP will be released soon and can be integrated into the system. With this data, we can tune the tool to be locally feasible. In addition, it would allow validation of our methodologies.

7.5. Linkage Disequilibrium studies

Expansion could be made to include linkage disequilibrium studies. Suppose the user has been informed of an SNP that has effects on drug response for a discovered population. The user would like to know if this SNP would have a similar effect for drug response in the local population by observing linkage disequilibrium similarities between SGVP populations and the discovered population. If the SNP is located in SGVP data, the similarity can be computed directly. Else if the SNP is not within the SGVP data, we hope to utilize data from International HapMap to determine the best surrogate SNP that can be found in SGVP using r^2 data.

8. Conclusion

We have seen the importance of SNPs and the part it plays in determining drug response. We have also determined the necessity for supplementing current drug target databases and the need to have an integrated database. We then gave an overview and subsequently details of our approach to these issues and how some of them have fared. In conclusion, we have not only combined data from various public access websites and building a bridge between them for ease of information gathering, but also developed techniques for discovering potential SNPs biomarkers for drug response, and devised an approach to augment the existing drug target databases. We hope to further improve on the system to become a useful tool in local pharmacogenomic studies as an ongoing project with the NUS-GIS Centre for Molecular Epidemiology.

References

1. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. (2008) GenBank. Nucleic Acids Research, Volume 36, January 2008, D25-30
2. Cooper G.M., Johnson J.A., Langaee T.Y., Feng H., Stanaway I.B., Schwarz U.I., Ritchie M.D., Stein C.M., Roden D.M., Smith J.D., Veenstra D.L., Rettie A.E., Rieder M.J. (2008) A Genome-Wide Scan for Common Genetic Variants With a Large Influence on Warfarin Maintenance Dose. Blood, Volume 112, No. 4, pp1022-7
3. Evans W.E., Relling M.V. (2004) Moving towards individualized medicine with pharmacogenomics. Nature, Volume 429, pp. 464–8
4. Hooven L.A., Butler J., Ream L.W., Whanger P.D. (2006) Microarray Analysis of Selenium-depleted and Selenium-supplmented Mice. Biological Trace Element Research, Volume 109, No. 2, pp. 173-79
5. Intuitor.com (2001) *Simpson's Paradox.* Online. Available at: http://www.intuitor.com/statistics/SimpsonsParadox.html [20 January 2009]
6. 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 Journal, Volume 1, pp. 167-170.
7. Kyoto Encyclopedia of Genes and Genomes (KEGG) (2009) *KEGG PATHWAY Database.* Online. Available at: http://www.genome.jp/kegg/pathway.html [10 March 2009]
8. Lee J.Y., Koh I.S. (2001) Drug to SNP: A Pharmacogenomics Database for Linkind Drug Response to SNPs. Genome Informatics, Volume 12, pp. 482-483
9. Liu G.M. Wong L.S., Li J.Y. (2007). Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns. In Proceedings of 13th International Conference on Knowledge Discovery and Data Mining, (San Jose, California, 12-15 August 2007)
10. McCarthy J.J. (2002) Pharmacogenomics: The Search for Individualized Therapies. Wiley-VCH Verlag GmbH & Co. KGaA, Germany, 2002.
11. McDonald, J.H. (2008) Handbook of Biological Statistics. Sparky House Publishing, Baltimore, Maryland, 2008.
12. Meyer U.A. (2004) Pharmacogenetics – five decades of therapeutic lessons from genetic diversity. Nature Reviews Genetics, Volume 5, pp. 669–76.
13. National Center for Biotechnology Information (NCBI) (2007) *SNPs: Variations on a Theme.* Online. Available at: http://www.ncbi.nlm.nih.gov/About/primer/snps.html [8 November 2008]
14. National Center for Biotechnology Information (NCBI) (2007) *Single Nucleotide Polymorphism.* Online. Available at: http://www.ncbi.nlm.nih.gov/projects/SNP/ [9 January 2009]
15. Oklahoma State University (1997) *Chi Square.* Online. Available at: http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage28.htm [10 November 2008]
16. Pasternak J.J. (2005) An Introduction to Human Molecular Genetics: Mechanisms of Inherited Diseases. 2nd Edition, Wiley-IEEE, 2005.

17. PerkinElmer, Inc (2005) *Single Nucleotide Polymorphisms – SNPs.* Online. Available from: http://las.perkinelmer.com/content/snps/genotyping.asp [15 March 2009]
18. The International HapMap Consortium. (2003) The International HapMap Project. Nature. Volume 426, No. 6968, pp.789-96
19. United States Food and Drug Administration (FDA) (2008) The New Drug Development Process. Online. Available at: http://www.fda.gov/CDER/HANDBOOK/ [10 March 2009]
20. United States Food and Drug Administration (FDA) (2005) Whole Genome Association Study and Risk Assessment. Online. Available at: http://www.fda.gov/cder/Offices/OCPB/Cox2005.pdf [10 March 2009]
21. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research, January 2008, Volume 36, D901-6
22. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research, January 2006, Volume 34, D668-72
23. Yong Loo Lin School of Medicine (National University of Singapore) (2008) The NUS-GIS Centre for Molecular Epidemiology. Online. Available from: http://www.med.nus.edu.sg/cof/cme.html [20 January 2009]