A PERSONAL GENOMIC INFORMATION ANALYSIS AND MANAGEMENT SYSTEM FOR HEALTHCARE PURPOSES

by

Amal Adel Alzu'bi

Bachelor of Science, Jordan University of Science and Technology, 2007

Master of Science, University of Pittsburgh, 2012

Submitted to the Graduate Faculty of

School of Health and Rehabilitation Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF HEALTH AND REHABILITATION SCIENCES

This dissertation was presented

by

Amal Adel Alzu'bi

It was defended on

April 13, 2016

and approved by

Mervat Abdelhak, Ph.D., RHIA, FAHIMA, Chair and Associate Professor,
Health Information Management

Valerie Watzlaf, Ph.D., RHIA, FAHIMA, Associate Professor,
Health Information Management

M. Michael Barmada., Ph.D, Associate Professor,
Human Genetics and Biomedical Informatics

Bill Hefley, PhD., CCP, CDP, COP, Clinical Professor,
Naveen Jindal School of Management. The University of Texas at Dallas

Dissertation Advisor: Leming Zhou., PhD, DSc, Assistant Professor,
Health Information Management

A PERSONAL GENOMIC INFORMATION ANALYSIS AND MANAGEMENT SYSTEM FOR HEALTHCARE PURPOSES

Amal Adel Alzu'bi, Ph.D.

University of Pittsburgh, 2016

Currently, a large amount of personal genomic data can be generated at an affordable price in a short period of time due to the improvement in the DNA sequencing technologies. Abundant research results on genetic diseases have been published in recent years. Therefore, it is eventually possible to integrate multiple types of information together and apply them into genomic-based personalized healthcare. However, this is still a very challenging task for healthcare professionals because the desired information is hidden in highly complex and heterogeneous genomic data sets and spread in various databases, which were typically created for researchers. In this research project, a personal genomic information management and analysis system is created for healthcare professionals, especially physicians.

To properly design such a system, an exploratory survey was conducted to identify the current status of physicians in using genomics in their clinical practice and to collect their expectations about the features of a patient genomic information system. The results of this study indicated that physicians have sufficient knowledge in genomics and they are interested in incorporating genomics into their clinical practice. The results also indicated that a well-designed patient genomic information system with desired features can help physicians to incorporate genomics into their clinical practice.

Based on the survey findings, a personal genomic information system was created for the purpose of managing and analyzing patient genomic data. In this system, we first created an

integrated database, and then developed data analysis algorithms to extract clinical information from patient genetic variation data, including disease-associated genetic variations and pharmacogenomic associations. Physicians can conveniently identify the genetic reasons for diseases and determine personalized treatment options based on the information provided by the system.

A usability study was conducted to obtain physicians' feedback about the system after they use it to finish some tasks such as searching the genetic variations of one patient, determining the patient's risk of certain diseases, and identifying the corresponding pharmacogenomic results. The results of this study indicated that physicians could easily find the patient information they need and the information can be directly applied in their clinical practice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First, I would like to express sincere appreciation to my advisor, Dr. Leming Zhou, for his help, advice, and persistent encouragement throughout my Ph.D. study. Without his support and guidance, none of the work presented in this dissertation would have been possible. Moreover, I am deeply indebted to him for showing me how to do successful research and how to communicate research results effectively.

I would like to thank my committee members: Dr. Mervat Abdelhak, Dr. Valerie Watzlaf, Dr. M. Michael Barmada, and Dr. Bill Hefley. Their advice and comments were very constructive and helpful. I really appreciate their help and support. I want to acknowledge special thanks to Dr. Barmada, for his help in guiding and evaluating the analysis part of my project. I am deeply indebted to Dr. Barmada and Dr. Watzlaf for their help in recruiting physicians for the system usability study.

I would like to acknowledge PharmGKB and Stanford University for giving me access to the variant and clinical annotation files.

Finally, I would like to express my deep gratitude to my family, and especially my parents and husband, for their love and support. I am grateful beyond words for all that they have given me.

1.0 INTRODUCTION

This chapter introduces the background, significance, and the challenges of this research project.

1.1 BACKGROUND

In recent years, by using the high-throughput sequencing technologies, large amounts of personal genomic data have been generated [1]. Genomics research in the past one and a half decades has provided a better understanding of the human genome and genetic foundation of diseases. The availability of both genomic data and extensive knowledge in genomics makes it possible to determine the risk of developing certain diseases or individual response to drugs by comparing personal genomic data with published research results [2], which is the foundation of personalized medicine.

At this moment, genomic data, genomics research results, and other useful resources for healthcare are spread in many places such as in disparate databases, web services, software programs, and journal articles. For a well-trained researcher in genomics, it is not difficult to obtain gene details from Genbank [3], to perform a query in dbSNP [4], to retrieve a record from Online Mendelian Inheritance in Man (OMIM) [5], to perform a sequence alignment using BLAST [6], and to combine the collected information together to evaluate the results published in a Genome-Wide Association Study (GWAS) research project. However, these tasks are

challenging for current healthcare professionals (physicians, nurses, health IT professionals) since most of them did not receive extensive training in genomics [7]. Therefore, to make the personal genomic information easily accessible and meaningful for healthcare purposes, sophisticated information integration, data management, and data analysis systems are needed.

1.2 SIGNIFICANCE

The goal of this project is to create a sophisticated, internally consistent, and scalable system that can collect, analyze, and integrate different types of genomic data, and organize them into a structured format so that it is convenient for physicians and other healthcare professionals to use [1, 8]. Data are stored in a central database that can be used by programs in the system to perform various tasks such as variation analysis, report generation, and other tasks specific to research [9] or clinical practice [10]. This system enables physicians to utilize the available patients' personal genomic data and up-to-date reliable research results so that they can have accurate diagnosis results and create personalized treatment plans, which are the early steps of personalized medicine [11-13]. In order to keep the system up-to-date, the data extraction scripts and data process algorithms need to be updated periodically.

1.3 CHALLENGES

To create such a system, there are several challenges. The first challenge is the complexity and large size of genomic data sets. For instance, a personal genome sequence alone has three billion

base pairs and the information in this sequence is complex. Furthermore, personal genomic data are complex in nature [14] and contain heterogeneous data types such as DNA sequences, RNA sequences and structures, protein sequences and structures, gene expression profiles, and DNA methylation profiles. They are all relevant to genetic diseases or metabolic responses to drugs in a certain way.

The second challenge relates to the need for integration of different types of genomic data from various sources while each data source has its own unique format. Each type of genomic data has its own unique structure and is stored in databases in different formats [15] such as plain text sequences, matrices, information description files, tables, and diagrams. Because of some historical reasons, different databases may use different names to refer to the same gene, or report the positions of genetic variations without explicitly indicating the reference genome used. Thus, the integration of heterogeneous genomic data types from different data sources is challenging [16].

In the field of genomics, there are often revisions and changes in databases caused by new discoveries [17]. Examples of these changes include updates on the reference genome, updates on annotation of certain genes, and newly identified Single Nucleotide Polymorphisms (SNPs). This leads to the third challenge in genomic data integration, rapidly updating data sources, which often means the data extraction scripts and data process algorithms need to be updated accordingly.

Because of these challenges, it is typically difficult for healthcare professionals such as physicians to collect desired information from multiple sources and confidently apply the information into their clinical practices, especially if they have not received sufficient training in

genomics. Our system simplifies the whole process and makes the desired information easily understandable and accessible for physicians.

The system we have developed performs genomic data management and genomic data analysis; integrates data from multiple sources and organizes them into a structured format in databases; and generates easy-to-understand reports for physicians. It provides one single place for various types of information needed by physicians in personalized medicine practice. For example, the physicians can just enter a patient ID and a disease name in the search area of the system and the system automatically performs database queries, genomic data analysis, and a report that includes information about the genetic variations of the patient related to the given disease. One important advantage of the system is that it allows physicians to get their desired results without extensive training in genomic data analysis.

2.0 LITERATURE REVIEW

This chapter provides a review of the available literature about personalized medicine, genetic variation data types, genetic variation databases, genetic variation data management, and genetic variation data analysis systems.

2.1 PERSONALIZED MEDICINE

Personalized medicine utilizes personal medical information to tailor strategies and medications for diagnosing and treating diseases in order to maintain people's health [18]. In this medical practice, physicians combine results from all available patient data (such as symptoms, traditional lab results, medical history and family history, and certain personal genomic information) so that they can make accurate diagnoses and determine personalized treatment strategies accordingly [11]. Health information management (HIM) professionals will play a critical role in this personalized healthcare practice because they are responsible for managing patient data.

With the improvement of high-throughput biotechnologies and the rapid decrease in DNA sequencing cost and time, genomic-based personalized medicine has already been practiced in a number of places, including Geisinger Genomic Medicine Institute, Scripps Health, Cleveland Clinic, Vanderbilt University Medical Center, and the Medical College of

Wisconsin [1]. Doctors have ordered whole-genome sequencing for their patients, and other healthcare professionals have performed genomic data analysis to identify genetic reasons for certain diseases that cannot be determined by conventional approaches [19]. Some hospitals, such as Children's Mercy Hospital in Kansas City [20] and the University of Pittsburgh Medical Center [21] have started to consider or have already taken the first few steps toward genomic-based personalized medicine.

2.1.1 Challenges in genomic-based personalized medicine

Challenges of practicing genomic-based personalized medicine can be divided into four categories [1, 22-24]; 1) the massive volume and complexity of genomic data from high-throughput technologies; 2) difficulties in genomic data analysis, such as information extraction, reporting, and database building; 3) difficulties in applying personal genomic information into clinics; and 4) challenges in personal genomic data security and privacy.

2.1.1.1 Challenges related to genomic data management

Today, there is a rapid increase in the use of the high-throughput biotechnologies, such as next-generation DNA sequencing technologies [25] for whole genome sequencing, microarray for gene expression patterns, RNA-seq [26] for identifying all transcripts of genes and their expression patterns; ChIP-seq [27, 28] for determining all regulatory elements in a genome; and MDB-seq [29] for collecting DNA-methylation profile genome-wide. These technologies can generate huge amounts of data in a short period of time.

The Human Genome Project (HGP) [30-32] was conducted from 1990 to 2003 for the purpose of sequencing and understanding the human genome. The total cost of the HGP was

about $3 billion, and one human genome was sequenced. In recent years, next-generation sequencing (NGS) technologies have dramatically reduced the DNA sequencing cost. The price dropped sharply from $3 billion to $100 million in 2007, then to roughly $1.5 million in 2008, [33] and to roughly $6,000 in 2012 [34, 35]. By the end of 2015, one could sequence a human genome in 24 hours at a cost of roughly $1,500 [35]. The commercialization and wide application of more advanced DNA sequencing technologies currently in development will further reduce the sequencing cost and increase the speed.

Organizing and managing these raw genomic data sets can be a big challenge. One other issue is the management of complex genomic data sets. Multiple types of genomic data are generated from different technologies and platforms [17].

2.1.1.2 Challenges in personal genomic data analysis

The next category of challenges is genomic data analysis. Extracting valuable information from large data sets is a difficult task. The situation is even worse when the data sets themselves are complex [14]. Genomic data sets include different types of data [15], such as DNA sequences, RNA sequences and structures, protein sequences and structures, and gene expression profiles. Each data type has its own characteristics and none of the data sets are simple to analyze, especially when its volume is large. For instance, analyzing a DNA sequence with 300 nucleotide bases is not difficult, but identifying a short, informative DNA segment from 3 billion nucleotide bases (which is the length of a human genome) is a nontrivial task.

In parallel to the astonishing advancement of high-throughput technologies is the significant progress researchers have made in understanding the disease at the molecular level in the past decade [36]. Before the HGP, it was quite difficult and time-consuming to sequence even a small genome [37], and therefore scientists could only focus on analyzing one gene or a

few genes in their research projects. Consequently, researchers believed that most genetic diseases are caused by mutations in one gene or a few genes. However, post-HGP research has indicated that this understanding was not correct. Further investigations performed on the human genome and other species' genomes have demonstrated the complexity of these genomes [38]. The associations between disease and genomic information have been extended from one single gene to multiple genes in one or multiple chromosomes [39].

With the great advancement in research, scientists have a better idea of the association between genetic variations in genomes and the patients' risk of developing certain diseases, and patients' possible responses to some drug therapies. On the other hand, these extensive research results also make it challenging to analyze personal genomic data for healthcare purposes. For example, in the past, when doctors wanted to determine the genetic reason for sickle cell anemia in a patient, they would only check one point mutation in the hemoglobin beta gene (HBB), a mutation that leads to a change in the shape of the red blood cell [40-42]. Therefore, the data analysis procedure could be quite simple. Today, with the availability of extensive genomic data sets, researchers know that genetic variations in multiple genes might be associated with sickle cell anemia. The task of determining the precise genetic reason for one patient's condition can be much more difficult. To make the situation worse, some research results from different research groups are not consistent [43] and directly conflict with each other.

Highly skilled genomic data analysts, well-designed databases and accurate research results, sophisticated algorithms and software programs, and sufficient computational power are needed for genomic data analysis [44]. At this moment, these resources are typically not available to most physicians. Even if one can easily access excellent genomic data analysis programs (which are available for certain types of data sets) and powerful workstations, it is still

challenging for most healthcare professionals to select the best program and the correct parameters for that particular data set because they typically do not have the required training in this field.

2.1.1.3 Challenges in applying personalized medicine into clinics

For healthcare purposes, one critical piece of information is which genes and mutations are involved in the development of a disease [45]. Decades ago, scientists already could determine associations between mutations in some genes and certain diseases. Recent technological advances have made it possible to determine the role of specific genetic variations in genetic diseases in large-scale genome-wide analyses [46]. However, the newly generated research results are overwhelming (hundreds of papers have reported genetic variations associated with one disease, and many of them have reported completely different variations) and sometimes even conflict with each other. Therefore, applying them in clinical practice is another challenge in genomic-based personalized medicine.

Since the cost of sequencing a human genome has sharply dropped to an affordable level and DNA sequencing may become a routine task in hospitals in the near future, physicians already realized the potential of genomics to improve clinical practice [10]. One expensive and challenging task for obtaining clinically useful information is analyzing these large-scale sequences and other genomic data files, connecting the analysis results with research results in the literature [47], and linking the results to the EHR in a meaningful way.

2.1.1.4 Challenges in personal genomic data security and privacy

Personal genomic data are highly sensitive [48] and need to be protected properly because each record contains not just the health information about one particular patient, but potentially,

information about a large group of people who have a blood relation with the person who takes the genetic test. This impact can last for generations because the genomic information will be passed to these people's descendants. In addition, by using some algorithms and data collected from public databases and social media, it is already possible to uniquely identify the owner of a de-identified genome sequence stored in public databases [49, 50]. The privacy of the individual and his or her relatives may be threatened, and the confidentiality of the personal genomic data is lost. The threat to the individual's offspring could be even more serious because research in genomics will likely enable the discovery of more information from a human genome in the future. Therefore, a stronger and more sophisticated security measure may be needed for personal genomic data protection, and this security measure should be set up before the wide application of genomic information in clinical practice.

2.1.2 Current solutions to the challenges

2.1.2.1 Genomic data management

NGS technologies can produce an enormous amount of sequencing data in a short time, and therefore storing and managing these huge data sets can be challenging. One straightforward suggestion is to apply a compression algorithm to reduce the sizes of these sequences [51]. A basic fact about the human genome is that genome sequences of two unrelated individuals are highly similar (roughly 99 percent identical) [52]. Therefore, directly storing hundreds and thousands of human genome sequences in a database would be highly redundant. One alternative approach is to keep only one reference genome (3 billion bases) and record all the differences between other human genomes and this reference genome. This approach can significantly reduce the size of stored data. Christley and colleagues combined these two approaches

(compression and storage of differences only) and reduced the information about one human genome from 3 gigabytes to about 4 megabytes [53]. In other words, the obtained data set is 750 times smaller than its original size. This approach also has problems. For instance, the information in the data set depends on the reference genome. Once the reference genome is updated, all the information in the stored data set needs to be updated as well. When the database contains a huge number of patient genomic records, this information update process may take significant time. Table 1 summarizes the advantages and disadvantages of these approaches.

Table 1. Advantages and disadvantages of DNA sequence storage strategies

Data Storage Strategy	**Advantages**	**Disadvantages**
Saving as a plain text file	Easy to retrieve and analyze the sequence	Large file size (3 GB/genome)
Saving as a compressed file	Smaller file size (roughly 1.5 to 2 GB/genome)	Time required (can be hours) to perform compression and decompression before data analysis
Saving only differences between the genome and the reference genome and compressing the file	Very small file size (4 MB/genome)	Time required to rebuild the needed DNA sequence; need to update all the files when the reference sequence is changed

Some information systems are specifically dedicated to the management of large-scale biological information. For instance, openBIS is a distributed information system that can be used for managing DNA sequences generated by NGS technologies [54].

For the challenging task of integrating different types of genomic data and organizing them in a way that is convenient for clinicians to use in their practices, the current solution is just to store different genomic data sets in different databases and provide links between certain items within those databases. The genomic databases created by the National Center for Biotechnology Information (NCBI) [55] are an example of this solution [56]. These databases can be very useful. However, physicians and other healthcare providers face a serious challenge to fully

utilize or correctly combine useful data items from multiple databases because this information integration task requires extensive knowledge of genetics and genomics to accomplish.

2.1.2.2 Personal genomic data analysis

Because of the complexity of genomic data, many people have realized that it is necessary to enhance education in genetics and genomics [57] for future healthcare providers. Such education is apparently beneficial. On the other hand, expecting every physician to become an expert in human genetics and genomics and keep track of the research progress in these fields would not be reasonable. It is also not reasonable to expect a physician to search the scientific literature to find out all the factors related to one common genetic disease so that he or she can order the correct genetic test or genomic analysis for the patient. After all, the research literature may contain many papers about one common disease and results in those papers do not necessarily agree with one another. Physicians would face significant difficulty in determining which results are applicable to their specific patients. For similar reasons, it would be equally challenging for physicians to use the literature to determine the correct drug and dosage for their patients. This scenario is quite different from the literature searches that physicians need to do occasionally on rare diseases. In that case, the difficulty is to identify the small number of papers about the rare disease among a huge number of irrelevant articles. In genomic-based personalized care, physicians deal with common diseases (such as cancer, obesity, cardiovascular disease, and diabetes) in a different way, and the difficulty is to identify the most suitable information for each patient from a large number of highly relevant articles. Therefore, the available genomics research results should be preprocessed and organized in a certain way before they are presented to physicians for diagnosis and treatment purposes.

Although most clinical sites, especially small healthcare facilities, do not have highly sophisticated genomic data analysts, skilled programmers, or high-performance computers for large-scale data analysis, they may use resources from large sequencing centers and cloud-based computing platforms to have those tasks done with low costs [58, 59]. Large genome-sequencing centers have well-trained genomic data analysts and extensive experience in processing large-scale genomic data. Cloud-based computing facilities (e.g., Amazon Web Services) can configure their high-performance computers according to the requests from their customers and conduct intensive computation tasks. The customers do not need to hire dedicated software engineers or purchase and maintain expensive high-performance computers. Some genomic data analysis pipelines are already available for this approach. One example is the Atlas2-Cloud pipeline [60, 61]. This pipeline has been successfully implemented into Amazon Web Services.

2.1.2.3 Applying personalized medicine into clinical practice

Integrating genetic analysis into clinical practice requires a radical change in the process of data collection, management, and analysis. It's becoming increasingly important to determine the level of evidences required to assess and evaluate the genetic variations based on their effect on patient's care. For example, a variant that requires a change in diet might need less evidence than one that requires a surgical procedure [62]. The success in integrating genomic data within clinical practices and the implementation of personalized medicine depends on the ability to analyze the scalable amount of genetic variations data. Some tools and software have been developed to analyze genetic variations data such as VAAST, GenePattern, and Gemini. These tools are usually used by researchers and require a good background in genetics and genomics.

2.1.2.4 Personal genomic information security and privacy

Policies and laws play a key role in protecting individuals' genetic information. The Health Insurance Portability and Accountability Act (HIPAA) provides the basic protection of the privacy of information stored in patients' records [63]. On January 25, 2013, the Office for Civil Rights of the US Department of Health and Human Services published modifications to the privacy, security, breach notification, and enforcement rules in HIPAA under the Health Information Technology for Economic and Clinical Health (HITECH) Act. One of these modifications emphasizes that personal genetic information is protected health information and prohibits the use and disclosure of genetic information by any health plans for underwriting purposes [64]. The Genetic Information Nondiscrimination Act (GINA) is the law specifically created to protect individuals from discrimination based on their genetic test results [65]. GINA protects individuals from genetic discrimination in both health insurance and employment.

Besides these policies and laws, many technical applications aim to protect personal genomic information. For instance, Interpretome [66] is a client-side genome interpretation system developed to analyze personal genomic data on the customer's local machine. The personal genomic data and the analysis results would not leave the customer's machine in order to protect this sensitive information. Another example is the GenePING system [67]. This system provides secure storage for genome data sets and enables the sharing of personal genetic variations and gene expressions by applying the Advanced Encryption Standard (AES) encryption algorithm.

2.2 GENETIC VARIATIONS

2.2.1 Multiple types of genetic variations

Although people look quite different, all human genomes are actually highly similar. More than 99 percent of human genomes from two unrelated individuals are identical [68-70]. The differences among these genomes are genetic variations. Most of these genetic variations simply make us appear different: skin color, eye color, hair color, height, etc. Some of these genetic variations can cause diseases.

Genetic variations refer to the differences in the DNA sequences between individuals or populations [71-73]. There are several types of genetic variations such as Single Nucleotide Polymorphisms (SNPs), copy number variations (CNVs), and short insertions and deletions (INDELs).

2.2.1.1 Single Nucleotide Polymorphisms (SNPs)

SNPs are the most common type of genetic variations that represent a difference in a single nucleotide [74]. SNPs represent 90 percent of the human DNA polymorphisms [75]. Figure 1 illustrates the concept of SNPs.

Figure 1. SNP Representation [76]

SNPs can occur anywhere in a genome. When SNPs occur within a gene or within the regulatory region near/inside a gene, they can affect the gene function/expressions and may increase the risk of developing certain diseases [77, 78].

Researchers in pharmacogenomics have found that SNPs may be helpful for predicting an individual's response to certain drugs, which is critically important for personalized medicine [79]. SNPs may be useful for predicting the individual response to environmental factors [80]. SNPs can also be used to track the inheritance of disease-leading genes within families [81].

2.2.1.2 Copy Number Variations (CNVs)

The term "copy number variation" refers to an intermediate-scale genetic change. CNVs include both additional copies of sequence (duplications) and losses of genetic material (deletions) [82]. For example, A-B-C-D is a DNA sequence in a chromosome. A, B, C, and D are DNA segments. This chromosome can instead have the following sequences: A-B-C-C-D (duplication) or A-B-D (deletion). Figure 2 illustrates the concept of CNV.

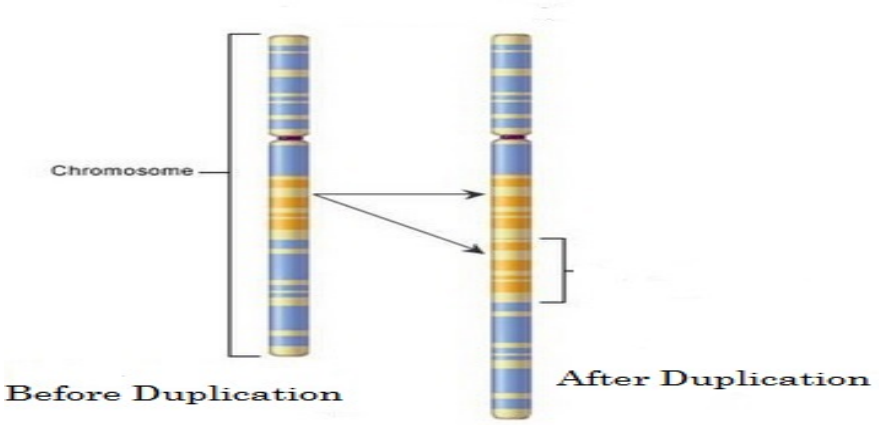

Figure 2. CNV Representation (Reproduced from [83])

Unlike a SNP that affects only one nucleotide base, the CNV ranges from about one kilobase (1,000 nucleotide bases) to several megabases in size [84, 85]. The scale of this type of genetic change may affect multiple genes [86].

There is an increasing concern about the effect of CNVs in developing complex diseases [87] since a number of CNVs overlaps with protein-coding regions [88]. CNVs were detected in genetic regions associated with complex neurological diseases [89] such as autism [90-92], Alzheimer's disease [93, 94], and schizophrenia [95-97].

2.2.1.3 Short Insertions and Deletions (INDELs)

An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly. The effect of short insertions depends on the number of base pairs inserted in the DNA strand. The insertion of one nucleotide may be more clinically significant than the insertion of 30 base pairs because the single nucleotide can result in a frame shift that can alter all the subsequent codons (a codon is a sequence of three

DNA or RNA nucleotides that corresponds with a specific amino acid or a stop signal during protein synthesis). If multiples of three base pairs are inserted, the translated protein has one or more extra amino acids [98]. Thus, its function may still be similar to the original protein.

Short deletions occur when a part of a DNA sequence is missing [99]. The number of deleted nucleotides ranges from a single base to a long segment of chromosome [100]. Similar to short insertions, the deletion of one single base may have a more serious impact on the function/expression of a gene than the deletion of multiple three base pairs in a gene region. Figure 3 illustrates the concept of INDELs.

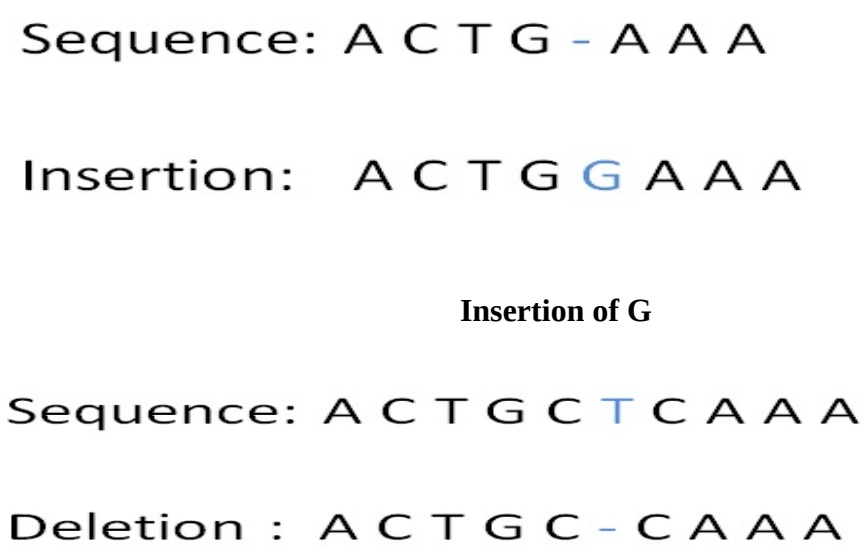

Figure 3. INDELs Representation

Approximately 36% of the short insertions and deletions are located within the promoters, introns, and exons of known genes [101] in the human genome. This means that some of these insertions and deletions can have an impact on human genes functions/expressions. Examples of diseases caused by short insertions include Huntington's disease and Myotonic

Dystrophy [98]. One example of a disease that is caused by a short deletion is cystic fibrosis [102].

2.2.2 Genetic variation databases

Computer databases become increasingly important to store and organize the growing amounts of genomic data sets. These databases help researchers and physicians to get their needed information conveniently [17]. There are several databases for genetic variations data. These databases can be divided into two categories: 1) Genetic variation databases that provide information about genetic variations only, and 2) Genetic variations and disease/phenotype databases that provide information about the associations between phenotypes and genetic variations.

2.2.2.1 Genetic variation databases

dbSNP [103] is a bioinformatics database for SNPs created by the National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine (NLM). This database stores all types of genetic variations < 50 bp [104]. This means that it includes different types of short genetic variations such as SNPs, INDELs, and microsatellite markers. dbSNP provides information about population-specific allele frequencies and genotypes and provides the validation state for each genetic variation [105].

dbVar [106] is a database that archives large-scale genetic variations for multiple species and is based on various genomic studies. Users can use dbVar to search, display, and download genomic data from submitted studies on a number of species. Genetic variation types of dbVar include INDELs and CNVs [107]. dbVar is also created and maintained by the NCBI.

The Database of Genomic Variations (DGV) [108] is an online database that provides a catalog of human structural variations, including INDELs and CNVs. The database is periodically updated according to the results reported in peer-reviewed research articles. DGV aims to catalog highest quality structural variations based on the literature in a format that is convenient to physicians, geneticists, and biologists [109]. DGV contains only data for healthy control of human samples, while dbVar accepts data from all species [110].

2.2.2.2 Genetic variations and disease/phenotype databases

Online Mendelian Inheritance in Man [111] is a comprehensive database that catalogues known diseases and their associated genetic variations. Each OMIM record has a summary of one disease and its relevant genes and variations in the human genome as reported in the literature [5]. Many links to other genetic databases such as GenBank, RefSeq, PubMed, and general and locus-specific mutation databases are also provided.

The Database of Genotypes and Phenotypes [112] is a database that archives the results of studies in the interaction between genotype and phenotype [113]. dbGaP provides an access to the large-scale genetic data sets that are needed for Genome-wide Association Studies (GWAS) designs. These data sets include public access to study documents linked to summary data on specific phenotype variables, statistical overviews of the genetic information, position of published associations in the genome, and authorized access to individual-level data [113].

The Human Gene Mutation Database (HGMD) [114] is a comprehensive collection of gene mutations that underlie or are associated with human genetic diseases. The information in HGMD is manually curated from the literature [115]. Each HGMD record includes a reference to the first literature report of a mutation, the associated disease specified in that report, the gene name, symbol, and chromosomal location.

ClinVar [116] database provides information about the medical relevance of the genetic variation. It archives the relationship between the medically significant variations and the phenotypes [117]. Clinvar is strongly related to dbSNP and dbVar since it maintains information about the location of variation on human assemblies. Unlike dbSNP and dbVar, Clinvar accepts direct submissions of structured details of phenotype, and interpretation of functional and clinical significance of the genetic variations.

SNPedia [118] is a wiki-based database. Researchers convert information presented in large-scale peer-reviewed genomic studies into machine-readable format, and then store the information in SNPedia so that the information is easily accessible to researchers. SNPedia supports personal genome annotation, interpretation, and analysis. SNPedia links the genetic variations to information about diseases or phenotypic traits published in genomic studies [119].

As a summary, short variations from multiple species can be found in dbSNP [120]. Structural genetic variations from multiple species can be found in dbvar [121]. Structural variations from healthy human beings can be searched in DGV [122]. The associations between human SNPs and disease/phenotype can be obtained from SNPedia [110]. OMIM can be used to find the association between human genetic variations and diseases, including an extensive description of relevant genes and phenotypes [123]. dbGAP is mainly used for controlled access to individual genotype/phenotype data obtained from association studies [113]. HGMD can be used to study the association between human genetic variations and genetic diseases [124]. ClinVar can be used to find the relationship between the medically significant variations and phenotypes. Table 2 presents a quick comparison between these databases.

Table 2. Comparisons between different genetic variations databases

DB Name	DB Purpose	Species
dbSNP	Stores all types of genetic variations < 50 bp.	All species
dbVar	Archives large-scale genetic variations.	All species
DGV	Provides a catalog of structural variations, including insertion, deletion, and copy number variations.	Human
OMIM	Catalogues known diseases with their genetic component.	Human
dbGAP	Archives the results of studies on the interaction between genotype and phenotype.	Human
HGMD	Includes a reference to the first literature report of a mutation, the associated disease state as specified in that report, the gene name, symbol, and chromosomal location.	Human
ClinVar	Archives the relationship between the medically significant variations and phenotypes.	Human
SNPedia	Converts the information in large-scale peer-reviewed genomic studies into machine- readable format that can be easily accessible by researchers.	Human

2.2.3 Genetic variations and disease association

There is a long history of the study on genetic diseases [125]. However, only after recent improvements in the fields of DNA sequencing [126] and SNP identification [127], researchers can now study the genetic variations in the whole genome and their association with diseases [127].

Human diseases can be classified into two categories: simple and complex. Simple (or Mendelian) diseases (such as cystic fibrosis) are caused by mutations in a single gene [128, 129]. These mutations are considered as causal mutations. Complex (or common) human diseases (such as schizophrenia) result from the combined effect of multiple genetic variations and environmental effect [130, 131]. Genetic variations associated with complex diseases do not

cause diseases but indeed influence the risk of developing these diseases [132]. The interaction between genetic variations and environmental factors can increase the risk of developing these complex genetic diseases. This means that having a genetic variation is not an absolute predictor of developing diseases. In some cases, genetic variations can reduce the risk of developing some diseases. For example, a higher copy number of CCL3L1 gene is associated with a reduced risk of HIV (Human Immunodeficiency Virus) infection [133].

Genetic variations that occur in the coding regions of the gene can affect the protein sequence, translational rate, and alternative splicing, all of which can influence the protein function and cause diseases. On the other hand, genetic variations that occur in the non-coding regions of the gene can alter the gene expression by modulating the activity of cis-regulatory elements [134].

Genome-wide association studies (GWAS) have been widely used for the discovery of genetic risk factors associated with human diseases. In GWAS, hundreds of thousands of SNPs are examined for association with a disease in hundreds or thousands of persons [46]. GWAS is typically based on a case–control design in which genetic variations from people with one disease or phenotype (cases) are compared against the ones from people without the disease of interest (controls) [135]. These studies catalogue and associate each clinical condition and phenotypic trait with SNPs [136].

Linkage analysis is a kind of study that aims to find the linkage between multiple genes. Generally, it is the tendency of genes to be inherited together because of their location near one another in the same chromosome [137]. In linkage analysis studies, researchers investigate and genotype the genetic variations that are spread throughout the genome in sets of family members in order to assess the co-segregation of alleles at any of these polymorphic variations with the

disease of interest [138]. Linkage analysis is a powerful analytical method for the discovery of genes associated with diseases [139]. Currently, linkage analysis is emerging as a useful method for the identification of rare variations associated with complex diseases. It also provides statistical evidence of the association between a genetic variation and a disease of interest.

2.2.4 Genetic variations and drugs (pharmacogenomic analysis)

Pharmacogenomic analysis is a new field of medicine that focuses on tailoring drug treatment based on individual genetic profiling. Pharmacogenomic analysis aims to improve patients' responses to drugs through the study of associations with human genetic variations. The ultimate goal of pharmacogenomics is to build some predictive models that use patients' genotypes to improve the treatment effectiveness and reduce the adverse effects of drugs [140]. The US Food and Drug Administration (FDA) produced a number of drug label revisions in order to include some relevant pharmacogenomic information. However, clinical adoption of pharmacogenomics is still slow [141]. Examples of important applications of pharmacogenomics that have been approved by the FDA [142] include warfarin and CYP2C9/VKORC1, abacavir and HLA-B*5701, HLA-B*1502, and TPMT.

In general, drugs are tested on large populations, and average individual response is reported. On the other hand, personalized medicine argues that every patient has a different response to a specific drug. Thus, if a genetic variation in a patient is associated with a specific drug response, then clinicians can use this information to make some clinical decisions such as adjusting the dose or changing the drug. Two important factors should be taken into account when studying an individual's response to a specific drug [143]: 1) how much of the drug is

required to reach its goal in the body? and 2) how well do the individual cells respond to the drug?

Despite the promising future of pharmacogenomic applications in clinical practice, there are several challenges [142, 144, 145] such as reimbursement, regulations, the need to educate and train health care providers, and the need to improve the health information infrastructure.

2.2.4.1 PharmGKB

Information about gene-drug interaction can be found in databases such as DrugBank [146] and The Therapeutic Target Database (TTD) [147, 148]. However, only The Pharmacogenomics Knowledgebase (PharmGKB) [149, 150] contains information about how human genetic variation leads to variation in drug response and drug pathways [151]. PharmGKB provides information about variation annotations, drug-centered pathway, pharmacogene summaries, clinical annotations, drug-dosing guidelines, and drug labels [152].

2.3 GENETIC VARIATIONS DATA MANAGEMENT

The goal of The Human Genome Project was to sequence the whole human genome and to identify genetic variations that are associated with diseases [153]. The information about the association between genetic variations and diseases is highly related to physicians. However, this information is very complex and inaccessible to physicians. Thus, there is a need for management systems that can integrate the diverse sets of genetic data and make them easily available and accessible to researchers and physicians.

The basic functionalities of a data management strategy include [8] 1) data collection, which maintains storage and protects security; 2) data integration, which requires the use of either standardization or metadata (systematic description of data) in order to make the data comparable; and 3) data delivery by making the data available and accessible. By applying these data management functions to personal genomic data, we can conclude that the goal of personal genomic information management is to collect and integrate different types of genomic data and organize them into a structured format (such as a database) that is convenient to use, access, and share [1]. Data can be stored in a core database that can be used by custom applications to prepare internal reports and statistics, and perform other functions that are specific to the research [9] or the clinical practice. Personal genomic information management systems allow individuals to share their genomic information with authorized physicians in order to help them in making an accurate decisions and personalized treatment options [11-13].

One simple approach for managing genetic variations data is the direct code manipulation of raw data files, where users can directly extract genetic variations data from a file, perform some necessary transformations, and write the results to another file [154]. Although this is a very simple and easy-to-use approach, it faces the scalability problem, which means that it may not be able to process the growing amount of genetic variations data. A number of applications have been developed for managing genetic variation data.

2.3.1 Genetic variation management systems

GENOME (The Georgia Tech Emory Networked Object Management Environment) is a prototype database management system (DBMS) that aims to manage large-scale and complex genomic data and to establish a network of researchable data sources [155]. GENOME can

integrate diverse sets of human genetic information from multiple data sources and share these data across the Internet. The GENOME prototype is set up as a network of data object servers that can request objects from any other server based on the object identifier. A GENOME browser is tightly associated with one particular GENOME server, which is known as a local or home server. Users can establish accounts on that server in order to obtain above-average read and write privileges on that server. In GENOME, Structured Query Language (SQL) statements can be used to help users in searching their own servers, or other servers that maintain a given type of data object. GENOME provides a dynamic interface that is designed to help users in formulating their queries. The network of data servers provides a flexible facility for scaling the database over many parallel systems.

SNPpy is open-source software for managing genotype and phenotype data from multiple Genome-Wide Association studies (GWAS) [154]. It manages and merges patients' data with the genomic SNP data from multiple studies and provides a powerful framework that facilitates the statistical analysis of SNPs data from GWAS. SNPpy consists of two parts; 1) a database to store and integrate SNPs data, and 2) a high-level interface to communicate with the database. One important feature of SNPpy is the low-level data validation [156], which is performed using the relational database that can constrain the columns' values to a fixed set of values. For example, if a record specifies that a patient's sex is 'A' (the only valid sex values are 'F' for female and 'M' for male), then the database will return an error.

TEAM (Targeted Enrichment Analysis and Management) is a web-based tool with a user interface that allows users to define, manage, and analyze panels of genes [157] in an easy-to-use environment [158]. The main goal of TEAM is to allow users to detect diagnostic variations from the sequencing data. The input data to the system consist of patient's genomic variations

that are predicted in specific genomic regions. These variations are stored in files with the standard Variant Calling Format (VCF) [159]. The VCF files contain all the genetic variations that are different from the reference genome in the sequence. The entire management of the VCF file is done locally and no patient's sequence data are sent over the internet. TEAM queries several disease-related mutation databases, such as HGMD-public [160], ClinVar [117], and COSMIC [161], in order to identify known diagnostic mutations about the disease of interest.

Cordova (Curated Online Reference Database of Variation Annotations) is an open source, web-based content management system that maintains a database of genetic variations [162]. The primary goal of the system is to help researchers to determine the clinical significance of genetic variations. Cordova integrates genetic variations with pathogenicity prediction results from popular algorithms. It provides an interface for researchers to review and organize data prior to public release. Cordova offers a platform to share reliable genetic variation data for the advancement of research. Users can search the database based on a genomic position or by a gene. In the term of variation categorization, "pathogenic" represents mutation published in the literature as causing disease, and "unknown significance" represents variation reported in dbSNP without a disease association.

SNPLims is an information system that aims to store and manage the SNP genotype data. [163]. Data are stored in a relational database. Each individual in the database is annotated with three types of data: genotypes, phenotypes, and demographics. SNPLims integrates genotype, phenotype, and demographic data from different laboratories. One goal of SNPLims is to manage large-scale genotypes for each sample and to manage a large number of samples and phenotypes in order to identify candidate genes for the disease of interest. SNPLims calculates the statistically significant association of the SNP with any measurable phenotype. The system

has been implemented as a client/server application in which users can access the data either through a command line client within a Linux server or through a web interface.

Based on the review of the previous applications for managing genomic data, we can find some features to enhance the functionality of any genomic data management system. Table 3 provides a list of these features.

Table 3. Some features in genomic data management systems

	Integration	Reports or figures	Search	Interface	Ease of use	Security
GENOME	√		√	√		√
SNPpy			√	√		
TEAM		√		√	√	√
Cordova	√	√	√	√	√	
SNPLims	√	√		√	√	√

2.4 GENETIC VARIATIONS DATA ANALYSIS

2.4.1 Analytical steps for genetic variations data

After a genome is sequenced, one critically important step is to conduct data analysis. Specific to genetic variations, there are three types of analysis [164]: variation identification, variation annotation, and variation visualization.

2.4.1.1 Variation identification

Variation identification is the process of identifying genetic variations [165]. Different methods can be used to discover and identify genetic variations such as [166] population association, case-control studies, using genome-wide markers, and DNA resequencing of candidate genes. Usually variation identification can be done through the comparison of sequenced genomes with the reference human genome (sequence alignment) and the identification of candidate sites at

which one or more samples differ from the reference sequence [167]. Variation identification is an important part of genome sequence data analysis. It identifies the relationship between genotype and phenotype [167], and therefore the results may be used to determine the risk of genetic diseases [168].

Genetic variations can be classified as either common or rare variations. Common variations have minor allele frequencies (MAF) greater than 5 percent. On the other hand, rare variations can be only found in a small fraction of sequenced samples and have MAF in the range of [0.1% to 2–3%] [169]. It is difficult to identify all variations since most of these variations are rare with population frequencies less than 1% [170]. To better understand these rare variations, more genome-wide association studies are required to examine these rare variations [171].

There are several tools for variation identification. However, there is no single tool that can identify all genetic variations [172]. This means that multiple tools need to be applied together. CRISP (Comprehensive Read Analysis for Identification of Single Nucleotide Polymorphisms (SNPs) from Pooled Sequencing) is a software program designed to detect SNPs and short INDELs from high-throughput sequencing of pooled DNA samples [173]. CRISP can detect both rare and common variations [174].

2.4.1.2 Variation annotation

Variation annotation refers to the classification and prediction of the functional impact of variations, and then filtering and prioritizing the ones that cause diseases [175]. Several tools are available for variation annotations such as ANNOVAR, VEP, and SVA. ANNOVAR is a command-line tool that provides functional annotation of single nucleotide variations (SNVs), INDELs, and CNVs [164]. ANNOVAR examines the variations' functional consequence on

genes, reports functional importance scores, and identifies variations reported in the 1000 Genomes Project and dbSNP [176]. VEP (Variation Effect Predictor) is available as a web-based tool that can be used to analyze up to 750 variations each time, as well as a downloadable script that can be used to analyze larger data sets [177]. SVA (Sequence Variation Analyzer) is a software program that provides a predicted biological function of the variations identified in the next-generation sequencing studies [178]. SVA allows visualization of the variations in the corresponding sequences by using a specified browser.

2.4.1.3 Variation visualization

Variation visualization refers to the validation and visual representation of genetic variations [179]. IGV (The Integrative Genomics Viewer) is a lightweight tool that supports the integration with clinical and phenotypic data [180]. IGV allows users to explore large-scale genomic data sets. IGV allows users to zoom in and out through the genome at any level of detail, up to a single base [181]. UCSC Genome Browser [182] is a web-based graphical viewer of genome sequences. It provides genome annotations and disease annotations [183] at various levels of detail, from base-pair level to chromosome level. Figure 4 shows visualization of a DNA sequence including genetic variations.

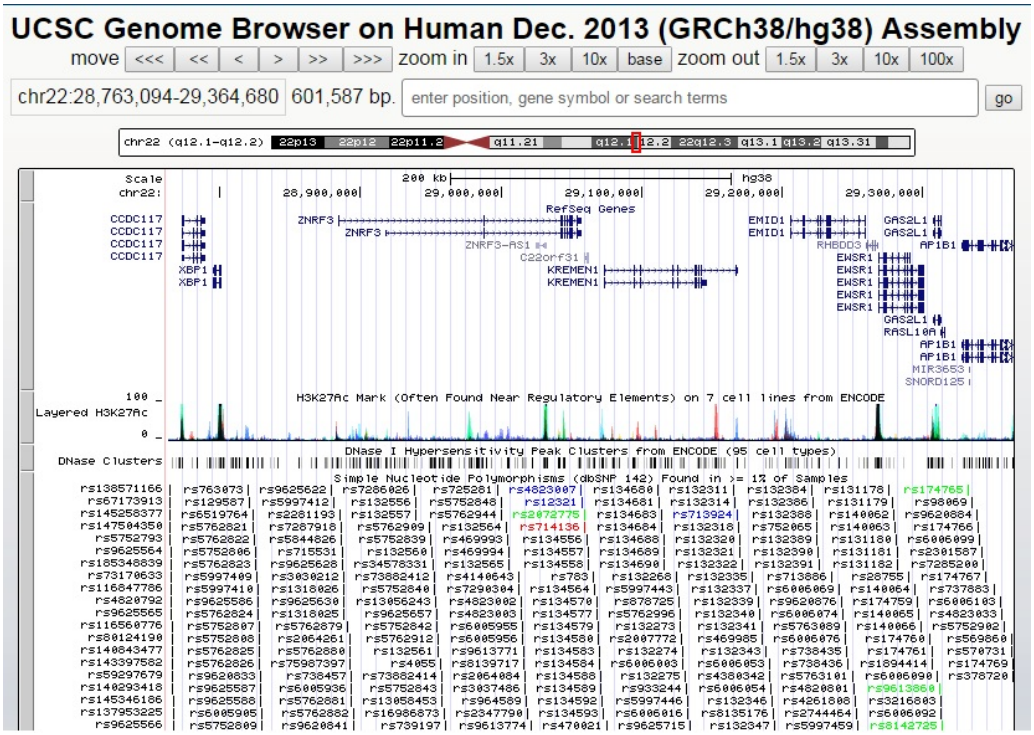

Figure 4. UCSC Visualization

2.4.2 Genetic variation analysis systems

The Genomes Management Application (GEM.app) is a set of tools that facilitate the storage, annotation, and analysis of genetic variation data. The goal of GEM.app is to manage, visualize, and analyze large genomic data sets [184]. GEM.app allows researchers to share data and perform joint analysis using an automated pipeline. GEM.app provides a powerful and user-friendly analysis and interpretation. The system is fast and can obtain results within 4 seconds across ~1,200 exomes. GEM.app is a web-based application [185] that makes genomic data available and accessible to researchers. One disadvantage of GEM.app is that it doesn't provide analysis of large chromosomal structural variations, large INDELs, or CNVs [185]. GEM.app has a flexible graphical user interface that is implemented in layers to facilitate efficient handling and querying of data. GEM.app framework has been used to identify clinically relevant

variations in a number of disorders such as inherited deafness, Charcot–Marie–Tooth (CMT) disease, and dilated cardiomyopathies [186, 187]. It also has been applied to identify novel genes [188].

VAAST (the Variation Annotation, Analysis and Search Tool) is an integrative and a probabilistic search tool that ranks DNA variations based on clinical gene importance [189]. This software is designed to screen individual genome sequences for clinically significant mutations. VAAST takes a sequence, runs it against a background database, and determines how dissimilar the sequence is to the sequences in the database. VAAST compares variations from a patient against hundreds of healthy genomes, and automatically scores the mutations in the form of a gene-by-gene ranking summary. It can identify both common and rare disease-causing variations [190]. VAAST evaluates the likelihood of observing the aggregate genotype of a feature given a background data set of control genomes. VAAST uses a generalized feature-based prioritization approach, which aggregates variations to achieve greater statistical search power. It also provides a statistically powerful means to rapidly search personal genome data for damaged genes and disease-causing variations. VAAST can score both coding and non-coding variations, and evaluate the aggregative impact of both types of SNVs simultaneously. One limitation of this tool is that it is not intended for browsing of variation and annotation data [184].

GenePattern provides a web-based interface that allows users to access a huge array of computational tools for genomic data analysis [191] such as gene expression analysis [192], proteomics analysis, RNA-seq analysis, and SNP analysis. GenePattern includes multiple user interfaces, including a web browser, application, and programmatic interfaces to make analysis modules and pipelines available to a broad range of users. One problem in GenePattern is that the Java client doesn’t always find the newly created modules or pipelines in the web

application. GenePattern provides access to more than 220 genomic analysis tools. It uses the analysis tools as building blocks to design sophisticated analysis pipelines.

Gemini (GENome MINing) is a software package for annotating and exploring genetic variations identified by large-scale whole-genome and whole-exome sequencing studies [193]. Gemini provides researchers with a standard framework for personal and medical genomics. Gemini integrates all forms of genetic variation (i.e., SNPs, INDELs, and structural variations) with diverse genome annotations databases such as dbSNP, ENCODE, UCSC, and ClinVar into a single database. When variations are uploaded, GEMINI automatically annotates them with pre-installed annotations gathered from resources such as dbSNP and ClinVar. GEMINI stores the annotated variations in a SQL database where researchers can query variations based on criteria such as sample genotypes and inheritance patterns. It also provides mechanisms for ad hoc queries and data exploration. The end result of the process is a database that researchers can query to identify variations based on the annotations or the genotypes of specific samples being studied. The architecture of GEMINI is shown in Figure 5.

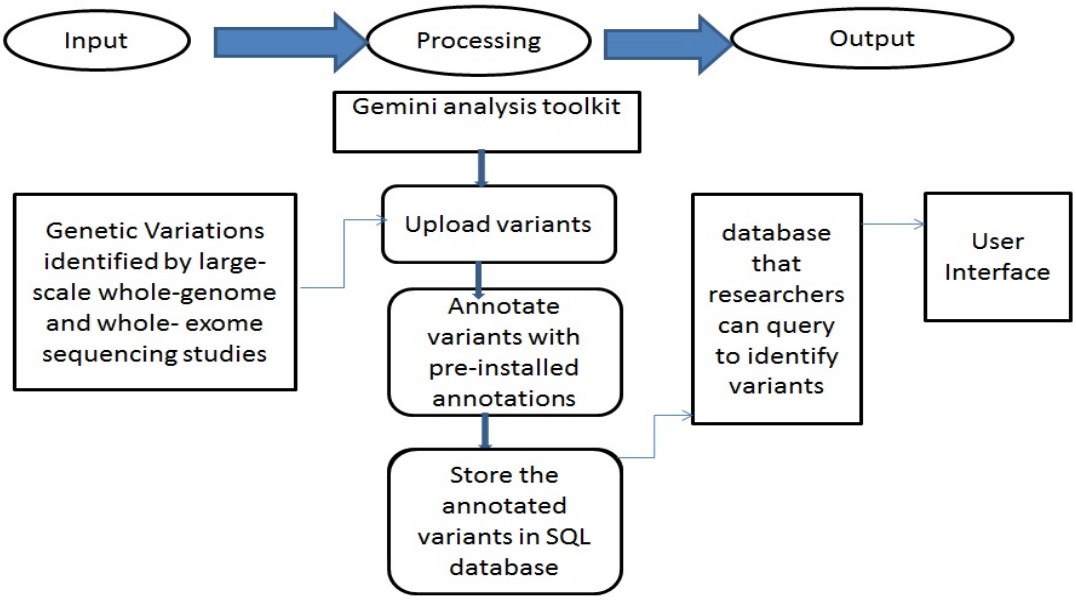

Figure 5. Architecture of GEMINI

VarSifter is a graphical software program designed to display, sort, filter, and sift variation data from parallel sequencing experiments [194]. This tool offers simple and user-friendly analyses and visualization of the extensive amount of data produced by exome sequencing. Researchers can view exome-scale sequence variation and perform sorting, filtering, and searching required to analyze these data for biological relevance. VarSifter is able to assist in the discovery of important variations linked with human disease [195, 196]. One limitation of VarSifter is that it is designed for a desktop computer and can only manage a modest amount of data [184].

VAR-MD is a software tool that analyzes genetic variations derived from exome sequencing in human pedigrees with Mendelian inheritance. VAR-MD produces a ranked list of potential disease-causing variations based on factors such as predicted pathogenicity, Mendelian inheritance models, genotype quality, and population variation frequency data. VAR-MD facilitates the diagnosis of rare diseases by improving the speed and accuracy of exome sequencing data analysis. This tool is unique as it uses family-based annotation of sequence data to enhance mutation identification [197]. VAR-MD implements a stepwise filtering algorithm to exclude variations identified as having a low potential to be disease-causing genotypes or a high potential to be false-positive genotypes. One limitation of VAR-MD is that it can work with small and simple pedigrees and a defined group of genetic models. It can’t perform as expected if there is genetic heterogeneity or incomplete phenotyping. Another limitation is that it can’t incorporate data from half-siblings and other “nonnuclear” pedigree members. Figure 6 shows the flow of information in VAR-MD system. Table 4 provides a comparison between the previously discussed genetic variations data analysis systems.

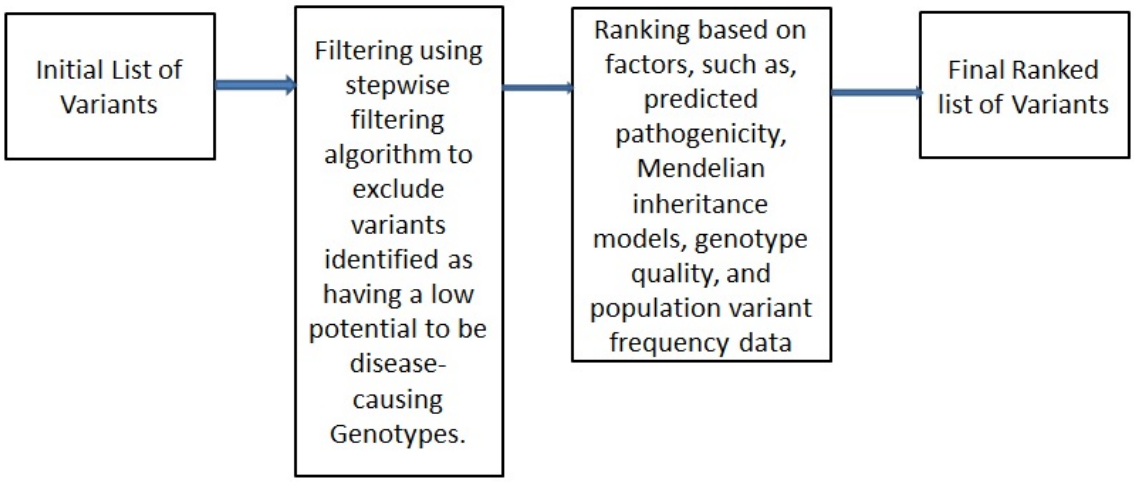

Figure 6. Flow of Information in VAR-MD System

Table 4. Comparisons between genetic variations data analysis systems

System	Advantages	Limitations
GEM.app	• Efficient for managing large genome data sets. • User-friendly analysis. • Users don't need to have computational experiences. • Simplicity and speed. • Bioinformatics skills are not required. • Secure data transmission.	It doesn't provide analysis of large chromosomal structural variations, large insertions/deletions, or copy-number variations.
VAAST	• Ranking the DNA variants based on clinical gene importance in an automatic way. • Identifying both common and rare disease-causing variants.	It is not intended for the browsing of variant and annotation data.
GenePattern	• Supporting a broad community of users at all levels of computational experience. • Allowing users to access a large array of computational tools (modules).	Java client doesn't always find the newly created modules or pipelines in the web application.
Gemini	• Annotating and exploring genetic variations. • Working as a standalone genome analysis toolkit and as a framework to build sophisticated graphical analysis and visualization tools. • GEMINI allows researchers to compose complex queries based on sample genotypes.	• The extensive time and resources required to import a VCF file and associated annotations into the GEMINI database.
VarSifter	• Offering a simple and user-friendly analyses and visualization of the huge amount of data produced by exome sequencing. • Can be used by researchers with any level of computational skills. • Able to assist in the discovery of important variants linked with human diseases.	• It is designed for desktop computers and can only manage a modest amount of data

Table 4 (Continued)

VAR-MD	• Producing a ranked list of potential disease-causing variants. • Facilitating the diagnosis of rare diseases. • Using family-based annotation of sequence data to enhance mutation identification.	• VAR-MD can work with small simple pedigrees and a defined group of genetic models. It doesn't perform as expected if there is genetic heterogeneity or incomplete phenotyping. • VAR-MD can't incorporate data from half-siblings and other "nonnuclear" pedigree members.

2.5 CONCLUSION OF LITERATURE REVIEW

Today, the study of personal genomic data is a key step toward the predictive medicine where the individual's genetic profile can be used to predict the best treatment options. Basic challenges with genetic variations data include 1) terminology is not standardized, meaning variations can be referred to as mutations, polymorphisms, or SNPs. Additionally, variation effects can be called pathogenic, deleterious, or disease-associated; 2) most variations are rare, which means that it's hard to discover and identify these variations; and 3) lack of standardization of the functional impact of genetic variations. More specifically, the literature contains many papers about one disease, and their results do not necessarily agree with one another. Thus, it would be difficult to determine which variations are significantly applicable to the patient.

Due to the availability of whole genome information, researchers have a much clearer understanding of the complete set of human genes and genetic variations. Now we can claim that most human diseases have genetic components [198], either inherited mutations from the parents

or de novo mutations in some individuals. However, the relationship between genetic variations and diseases is not that simple; we can't ensure that a person may develop a specific disease because he/she has the related genetic variation. Many challenges face the analysis of genetic variations data: tools are not readily available for biomedical researchers, tools are difficult to use, and results are difficult to interpret correctly.

Based on this review, we can find that there are few systems for managing and analyzing genetic variation data that are currently in use for research purposes. This brings us to the need for further research in the field of managing and analyzing genetic variation data to support clinical practice. HIM professionals have the basic skills to manage large-scale data, they just need education in the field of genomics to be able to understand these genomic data and pass them to genomic data analysis tools in order to provide some summary results to physicians. Thus, HIM professionals can play a key role in developing such systems for managing and analyzing large-scale genomic data and make these data available and easily accessible.

As a conclusion, the success in integrating genomic data within clinical practices and the implementation of personalized medicine depends on the ability to analyze the scalable amount of genetic variation data and the ability to integrate these genomic data with the available personal data. Performance, searchability, security, and scalability are key features that should be taken into consideration when designing a genomic data management and analysis system.

One suggested solution is to create an integrated database for clinical use. This database should be able to organize all the relevant information for each disease in one place and present the information in an easy-to-understand format. When the information is needed, it can be retrieved immediately in a single step, instead of requiring healthcare providers to come up with a data analysis procedure to query and combine pieces of information from multiple places. In

other words, an integrated and comprehensive genomic information management and analysis system for clinical use is necessary.

3.0 METHODOLOGY

This chapter discusses the specific aims of this research, the design, and research methods for each one of the specific aims.

3.1 SPECIFIC AIMS

This study aims to manage and analyze large amounts of genomic data and to enable convenient extraction of information for physicians. These aims can be achieved as follows:

Specific Aim 1: To determine, via a survey, the current status of physicians in using genomic data in their clinical practices, and their expectations about the features and characteristics of a genomic information system to support their clinical practices.

Research Hypothesis 1: This survey addresses two issues: first, whether physicians are able to, are currently using, or want to use genomics in their clinical practices; secondly, understanding physicians' expectations about the features and characteristics of a genomic information system.

Specific Aim 2: To develop a system that manages personal genomic data, such as genetic variation data, through the following steps: collecting data from multiple sources, extracting information, and integrating these data into a structured format in a central database.

Research Hypothesis 2: The system provides one single place for various types of genomic information needed by physicians, so that they can conveniently access the desired patient genetic information and current research results in one place.

Specific Aim 3: To develop data analysis algorithms and combine them with the data management system. These data analysis algorithms analyze complex personal genomic data through the following steps: identifying the genetic information related to a certain disease, analyzing VCF files, identifying patients' genetic variations related to the disease of interest, and identifying the corresponding pharmacogenomic information.

Research Hypothesis 3: The system, through the data analysis algorithms, allows physicians to screen and analyze all the genetic variations in the patient and then identify the genetic variations associated with the disease of interest. The system identifies the clinical significance of every single genetic variation in the patient. It also identifies the corresponding pharmacogenomic information for each patient.

Specific Aim 4: To generate user-friendly summary reports for physicians. These reports are well-formatted reports and provide a summary of genomic findings.

Research Hypothesis 4: The generated reports are easily understood by physicians. They include information about the patient genetic variations related to a certain disease and the corresponding pharmacogenomic information. Therefore, physicians can conveniently identify the genetic reasons for diseases and determine personalized treatment options based on the information provided in the report.

3.2 SPECIFIC AIM 1: SURVEY OF PHYSICIANS' NEEDS AND EXPECTATIONS

3.2.1 Theoretical framework

The survey was developed according to the Rogers Diffusion of Innovation (DOI) theory [199]. Rogers defines an innovation as a practice that can be perceived as new. Applying genomics into clinical practice meets Rogers' definition of innovation [200, 201]. According to the DOI theory, diffusion of innovation is led by a set of innovators who familiarize themselves with the innovation (knowledge), form attitudes about the potential of innovation to improve the current practice (attitudes/receptivity), decide whether to adopt the new practice (decision), and evaluate the adopted practice (confidence). In this survey, there are four different sections corresponding to these four domains: knowledge, attitudes/receptivity, decision, and confidence. Table 5 shows the conceptual and operational definitions for each of the domains in the DOI theory and corresponding sections in this survey.

Table 5. Conceptual and operational definitions of DOI theory

Domain (Rogers, 2003)	Conceptual Definition (Rogers, 2003)	Operational Definition	Survey Instrument Sections
Knowledge	Recognition of the innovation and evidence of understanding its function.	Knowledge of applying genomics in the clinical practice.	Knowledge in genomics
Confidence	Level of certainty that knowledge about the innovation is accurate.	Confidence in: new findings in genomics, motivations of using genomics in clinical practice, and benefits and limitations of genetic tests.	General opinions
Attitudes	The relative advantage offered by the innovations, and the recognized need for the innovation.	Perceived advantage and disadvantages of integrating genomics into clinical practice.	Expected features from a genomic information system.

Table 5 (Continued)

Decision/ adoption	Observation of use of the innovation.	Utilization of using genetic testing in clinical practice.	Specific tests

3.2.2 Design

A 31-question survey was developed and informed by relevant literature review. Theories and findings from a number of studies have been used to guide the selection of the questions in each section of the survey. Each question in the survey was reviewed to determine how well it measures the DOI domains and meets the survey goals. (Survey questions are provided in Appendix A.)

The survey has five sections: 1) the general information section is about physicians' basic information such as gender, age range, field of practice, and years of experience in the clinical practice; 2) the knowledge in genomics section corresponds to the knowledge domain of the DOI theory; 3) the general opinions section is about using genomics in the clinical practice and whether the physicians believe that applying genomics in a clinical practice can improve the quality of their practice. This section corresponds to the confidence domain of the DOI theory; 4) the specific tests section assesses physicians' willingness to order simple genetic tests for single gene disorders and sophisticated tests for multiple gene disorders. This section corresponds to the decision/adoption domain of the DOI theory; and 5) the expected features of a genomic information system section that assesses physicians' expectations to a desired genomic information system. This section corresponds to the attitude domain of the DOI theory.

A five-point Likert scale rating level of agreement and familiarity was used to assess opinions. Categorical response options were used in the background and general opinion questions. The survey was hosted online using Pitt's Qualtrics system. No personally identifiable information (such as name or address of the respondent) was collected in this study. All the collected data were kept anonymous.

3.2.3 Sample and recruitment

This survey was given to a sample of physicians who work at the University of Pittsburgh. The subject selection criteria include: 1) physicians who have at least a few years of clinical practice; 2) their ages range between 30 and 65; and 3) physicians who are fluent in English. The study protocol was reviewed and approved by the University of Pittsburgh Institutional Review Board (IRB #: PRO14070123). The study was considered to be exempt because respondents were anonymous and there was no risk to participants.

3.2.4 Statistical analysis

Respondents' answers to those survey questions were downloaded from the Qualtrics system and exported to SPSS (version 23). All survey results were tabulated and analyzed using descriptive statistics. Open-ended questions were summarized separately.

3.2.5 Survey validity

Each survey question was reviewed for content validity by four experts from relevant fields: public health, computational genomics, human genetics, and medicine. The reviewers were asked to rate the relevancy of each question relative to the study aims. A four-point score was used to determine their opinions in terms of relevancy (1: totally irrelevant, 2: somewhat not relevant, 3: somewhat relevant, and 4: highly relevant). In the first round of content validity review, three questions were regarded as irrelevant questions by two reviewers. These questions were about physicians' overall knowledge in genomics, whether they have ordered any genomic tests, and the genetic tests they have ordered in the past. These three questions are highly relevant to the study aims. We therefore discussed this with the two reviewers and they agreed with our opinion. In other words, all 31 questions in the survey are considered as relevant by all four reviewers. The four reviewers also evaluated and commented on the clarity of each question in the survey. They pointed out that the clarity level of several questions was not high. We changed the wording of those questions. In the second round of the content validity review, all four reviewers agreed that all of the survey questions are relevant and clear.

3.3 SPECIFIC AIM 2: MANAGEMENT

3.3.1 Theoretical framework

Theories and findings from a number of studies were used to guide the design of the management system. The framework of Parsons et al. for managing diverse, distributed, and

heterogeneous scientific data [202] was used to guide the management of the heterogeneous and diverse types of genomic data stored in the system. The framework emphasizes the need for continued adaptation by technology and people. It also uses some simple terms to describe the data such as discoverable, linked, useful, and safe. Table 6 reviews the conceptual and operational definitions of the framework terms.

Table 6. Conceptual and operational definitions of Parsons et al. framework

Terms	Conceptual Definition	Operational Definition
Discoverable	Data should be identified and assessed using simple tools available to the community.	Physicians can search the system by entering a key or a specific word in order to get the desired answer for their question.
Linked	Data should be interrelated and connected.	In our system we integrate data from multiple databases, combine them in a single place, and make them available to physicians.
Useful	Data should be used in different applications, by researchers and decision makers.	Our system can be used by different people, including physicians and researchers to get useful information.
Safe	Security, privacy, and confidentiality of the data should be taken into consideration.	There will be a number of security measures such as role-based access control, user authentication processes, and encryption.

The computational framework to integrate biomolecular and clinical data within a translational approach [203] was used to guide the design of the basic levels of the management system. The framework uses different levels to represent and manage the data. Our system has three basic tiers: 1) a data tier that stores information about genes, SNPs, diseases, GWAS, and pharmacogenomic information; 2) an application tier that uses Python scripts and Java codes to create, update, retrieve, and manage information in the database; and 3) a user interface tier that uses Java frames to facilitate the interactions between the user and the system. Figure 7 shows the basic tiers of the system. The relational data model was used to organize the data in an integrated way in the database [204]. Our system uses MySQL database to store data about

genes, SNPs, diseases, GWAS, and pharmacogenomic information. These data can be easily retrieved using SQL queries.

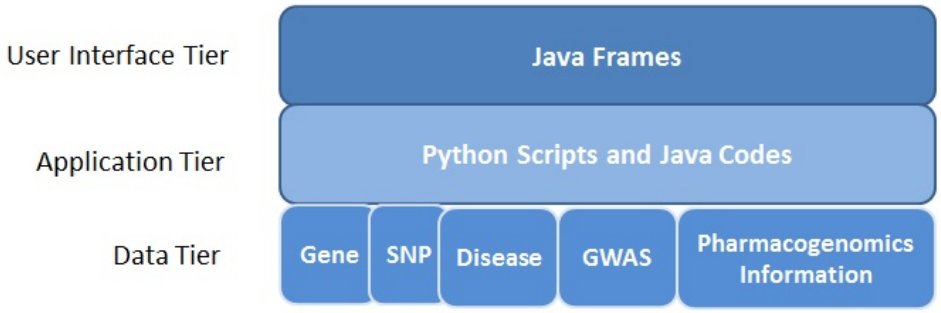

Figure 7. Basic Tiers of Our System

3.3.2 Design

Our system is designed to be used by physicians in their clinical practices, which means that the system should be easy to use and fast since physicians need the desired information readily accessible for the patient care purposes. Python scripts are used to search multiple sources and obtain all the needed information. These sources include OMIM, dbSNP, ClinVar, GenBank, GWAS catalog, and PharmGKB. The obtained information is stored in a MySQL database.

Our system provides a single place for various types of information needed by physicians in personalized medicine practice. In order to meet the system's design requirements, the system has the following essential features:

- The system is convenient to use. It collects data from multiple sources and organizes them into a structured format in the database. For example, the physician enters a patient ID and a disease name in the search area of the system, and then the system automatically

performs database queries and provides the physician with a report that includes information about the patient's genetic variations that are related to the given disease.

- The system precalculates all of the computationally intensive steps for every patient. In the clinical environment, physicians don't have time to wait. They need to get the results shortly after they input the request for patient care purposes. Therefore, it's important for the system to precalculate and save the analysis results into the central database. This precalculation feature saves a lot of time and helps the physicians to obtain their needed information and make their decisions in a timely manner (details are provided in the data analysis part).
- The system provides a number of security measures, including encryption, user authentication processes, and role-based access control. Furthermore, the central database is not accessible to the public and physicians are not allowed to insert data into the database. Dropdown lists are used to make selections.

3.4 SPECIFIC AIM 3: DATA ANALYSIS

3.4.1 Theoretical framework

The Sadedin et al. [205] study was used to guide the design of the analysis part of the system. The study identifies three basic requirements for a clinical bioinformatics analysis system: First, a clinical system should be designed with a robust and reproducible analysis. Second, genetic variations need to be assessed for their relevance to a given patient. Third, the overall flow of the analysis should be easy to understand and modify.

3.4.2 Design

Data analysis algorithms were created using Python scripts and Java codes. These algorithms aim to screen and analyze all the genetic variations in the patient and then identify the genetic variations associated with the disease of interest. The algorithms also identify the corresponding pharmacogenomic information for every patient. Data analysis is divided into four steps: (Details are provided in the implementation part.)

- Identifying the genetic information related to a certain disease.
- Analyzing VCF files.
- Identifying the patients' genetic variations related to the given disease.
- Identifying the related pharmacogenomic information.

3.5 SPECIFIC AIM 4: REPORT GENERATION

3.5.1 Theoretical framework

The design of the generated reports was guided by well-tested practices published in the literature [206, 207]. These studies provide some reporting approaches for communicating genomic findings relevant to clinical practice. Based on these studies, genomic reports should be well-formatted in order to provide a summary of genomic findings. These genomic findings enable physicians to take appropriate steps for disease diagnosis, prevention, and management for their patients. The studies also emphasize the power of user-friendly reports in reducing the

time required to explain test results, leaving more time for discussing the significance of the results for patients' care.

3.5.2 Design

The generated reports include four sections: (Details are provided in the implementation part.)

1) Patient's information

2) Result summary

4) Disease information

5) Pharmacogenomic information

4.0 IMPLEMENTATION

This chapter discusses the basic functionalities of our system and the security measures in the system.

4.1 SYSTEM FUNCTIONALITIES

The basic functionalities of the system include data collection, data integration, data analysis, and information delivery.

4.1.1 Data collection

The system collects data from multiple databases, including GenBank, dbSNP, ClinVar, OMIM, GWAS catalog, and PharmGKB. Python scripts were created to extract the required information from these heterogeneous data sources. Due to the variable nature of genomic data and the rapidly updating data sources, the data extraction scripts will be updated as needed. The system collects different types of genomic data from these data sources. The following subsections provide some examples of the collected data items.

4.1.1.1 Gene information

GenBank is a comprehensive database that contains publicly available nucleotide sequences and their protein translations for several organisms. These sequences are obtained through submissions from individual laboratories or batch submissions from large-scale sequencing projects [208].

A Python script was created and used to process and then extract gene information from the GenBank ftp site [209]. In this project our focus is on human genes. Thus, we only extracted GenBank files that are related to the human organism (*Homo sapiens*). As of December 2015, we obtained 49 GenBank files related to the human organism. These files are named "gbpri1 through gbpri49." Their sizes range from 47 MB to 245 MB. Examples of the extracted gene's information include gene symbols, aliases, chromosomes, and citations from PubMed.

4.1.1.2 Genetic variation information

Genetic variation information was extracted from both ClinVar and dbSNP databases. This project is primarily focused on the analysis of SNPs. Thus, a Python script was created and used to process and then extract the needed SNP's information from the ClinVar ftp site [210]. Examples of the extracted information include variation name, gene symbol, dbSNP ID, and clinical significance. This version of the system includes SNPs from two assemblies; GRCh37 and GRCh38. However, the results showed that some of the extracted SNPs from ClinVar missed useful details such as gene details. Therefore, the dbSNP database was used to extract the missing information. dbSNP ID was extracted and used to uniquely identify every SNP.

In the second step, a Python script was created and used to process and then extract the needed SNP's information from the dbSNP ftp site [211]. In this project, our focus is human

SNPs. Thus, we only extracted the information about human SNPs. In the dbSNP ftp site, human SNPs' information is arranged into files based on the chromosome number, in which each file belongs to one chromosome and contains information about all of the genetic variations in this chromosome. This version of the system includes SNPs from two assemblies of build (b144); b144_GRCh37p13 and b144_GRCh38p2. Examples of the extracted information include dbSNP ID, gene symbol, SNP position in the chromosome, and clinical significance.

4.1.1.3 Disease information

Disease information was extracted from the OMIM database. A download request was submitted to Johns Hopkins University in order to get access to the OMIM ftp and extract the needed information about human diseases. After getting access, a Python script was created to process the OMIM data and extract the needed disease's information. Examples of the extracted information include disease name, chromosome number, gene symbol, and dbSNP ID.

4.1.1.4 GWAS information

The GWAS catalog [136] was created by using text mining algorithms to extract information from all of the published GWAS research articles. A python script was created and used to process the GWAS catalog data [212] and then extract the needed information, such as study ID, journal information, disease name, dbSNP ID, and chromosome number.

4.1.1.5 Pharmacogenomic information

PharmGKB includes more than 26,000 genes. Only a few of them have an impact on disease or drug response [150, 151]. In our research project, only those genes in which a variant exists in

the PharmGKB variant and clinical annotation files, specifically variant-phenotype-annotation, are considered to be gold-standard pharmacogenomic information to which all patients' genetic variations are compared.

The variant-phenotype-annotation file contains associations in which the variant affects a phenotype, with or without drug information [213]. A Python script was created and used to process the file and then extract and integrate all of the related information about the pharmacogenomic associations. As a result, for each SNP, the script integrates all of the available information about the association between a SNP and a specific disease or drug.

4.1.2 Data integration

After collecting data from multiple sources, the system organizes the data into a structured format in a central database. The system uses the MySQL relational database to store the different types of genomic data, such as genes, genetic variations, diseases, GWAS results, and pharmacogenomic associations into database tables. These tables have common fields such as gene symbol and variation ID (dbSNP ID). The system uses these common fields to create relationships among the tables and link them together. SQL statements can be created to manipulate the data in the database. The following subsection illustrates the fields in the database tables.

4.1.2.1 Database tables

Tables in the database are divided into two categories:

1. Extraction output: This category includes those tables created using Python scripts; the data in these tables were collected from multiple sources and integrated into a single format.

2. Analysis output: This category includes those tables created during the system execution time using Java codes, such as the table that stores personal genetic variations for each patient. The data in these tables depend on the analysis results.

4.1.2.2 Extraction output tables

GenBank_database table. This table stores the data collected from the GenBank database. It has a total of 331,770 records. Each record has the following fields:

- Gene_Symbol, such as CFTR, BRCA.
- Accession_Number: The unique identifier for the gene in GenBank.
- Sequence_Length: Number of nucleotide base pairs (or amino acid residues) in the gene sequence.
- Sequence_Type, such as DNA, RNA, mRNA.
- Chromosome: chromosomal location of this gene.
- Gene_Synonyms: other synonyms of the gene symbol.
- PubMed_ID (PubMed Identifier): this ID provides a link to a PubMed paper about this gene.

ClinVar table. This table stores the data collected from the ClinVar database. It has a total of 199,054 records. Each record has the following fields:

- Variant_Type: In this project our focus is on the SNP. Thus, this field has a value of Single Nucleotide Polymorphism.
- Variant_Name: the name of this variant.
- Gene_Symbol: the symbol of the gene that is overlapping the SNP.
- Clinical_Significance: the reported clinical significance of this variant such as benign, likely benign, or pathogenic.
- Assembly: the name of the assembly on which locations are based such as GRCh37.
- dbSNP_ID: the unique identifier of the variant in the dbSNP database.
- Chromosome: chromosomal location of this variant.
- Reference_Allele: The allele at the location defined on the reference sequence.
- Alternative_ Allele: The difference relative to that reference.
- Disease_OMIM: the identifier of the disease reported for this variant as in OMIM database.

dbSNP_database. This table stores the data collected from the dbSNP database. It has a total of 105,113 records. Each record has the following fields:

- dbSNP_ID: the unique identifier of the variant in the dbSNP database.

- Chromosome: chromosomal location of this variant.
- Position: the position of this variant in the chromosome.
- Gene_Symbol: the gene that is overlapping the SNP.
- Significance: the reported clinical significance of this variant, including unknown, untested, non-pathogenic, probable-non-pathogenic, probable-pathogenic, pathogenic, and other.

OMIM_database. This table stores the data collected from the OMIM database. It has a total of 6,558 records. Each record has the following fields:

- Disease_Name: the name of this disorder.
- Disease_OMIM: the corresponding OMIM ID for this disease.
- Gene_Symbol: the symbol of the related gene to this disease.
- Gene_MIM_ID: the corresponding OMIM ID for the reported gene.
- Chromosome: the related chromosome number of this disease.

GWAS_Catalog table: This table stores the data collected from the GWAS catalog. It has a total of 17,832 records. Each record has the following fields:

- StudyID: this is the PubMed ID for this GWAS article.
- dbSNP_ID: the ID of the SNP reported in the GWAS. Multiple SNPs in the same GWAS have multiple records in the database.

- Chromosome: chromosomal location of the reported SNP.
- Disease_Name: the name of disease investigated in this GWAS.
- Gene_Symbol: Gene symbol related to the reported SNP.

GWAS_Study_Info. This table stores information about each GWAS article. It has a total of 2,149 records. Each record has the following fields:

- Study_ID: this is the PubMed ID for this GWAS article.
- Date_Addded_to_Catalog: when this GWAS was added to the catalog.
- Journal: the name of the journal that published this GWAS.
- Sample_Info: sample size in this GWAS.
- Population: the population studied in this GWAS.

Variant_Citation table: Each record in this table stores a PubMed ID that links to a paper about a specific SNP. These data were extracted from the ClinVar database using a Python script. The table has a total of 101,186 records. Each record has the following fields:

- dbSNP_ID: the unique identifier of the variant in the dbSNP database.
- PubMed_ID: this ID provides a link to a PubMed paper about this SNP.

Pharmacogenomics table: This table is created based on the information extracted and integrated from the PharmGKB variant-phenotype-annotation file. The table has a total of 3,602 records. Each record has the following fields:

- dbSNP_ID: the unique identifier of the variant in the dbSNP database.
- Gene_Symbol: related gene symbol.
- Drug: drug name.
- PubMed_ID: this ID links to the corresponding article in PubMed.
- Category: options include efficacy, toxicity, dosage, or other.
- Significance: yes or no, based on the significance of the association, such as the association between a SNP and a specific drug, or the association between a SNP and a specific disease [213].
- Recommendations: one sentence represents the association between the given variant and one disease or between the given variant and one drug [213].

4.1.3 Data analysis

Data analysis is divided into four steps: identifying the genetic information related to a certain disease; analyzing VCF files; identifying the patients' genetic variations related to the given disease; and identifying the corresponding pharmacogenomic information. Algorithms were created to perform each step of the data analysis.

4.1.3.1 Identifying the genetic information related to a certain disease

The system provides detailed genetic information about a specific disease. When the user enters a disease name in the search area, the system searches the OMIM database table and identifies

the related records. These OMIM records contain information such as disease name, related genes, chromosome number, and OMIM ID. Based on the OMIM ID, the system searches the ClinVar database table and returns all of the SNPs related to the disease of interest. Some of the returned SNPs from ClinVar don't have certain desired information such as gene symbol and gene ID. In that case, the system searches these SNPs in the dbSNP database table and retrieves all of the related information. After these steps are done, the system returns a database table with the detailed information related to the disease of interest, including genetic variation name, SNP ID, gene symbol, GenBank ID, chromosome number, and the clinical significance of the variation (including unknown, untested, non-pathogenic, probable-non-pathogenic, probable-pathogenic, pathogenic, and other). Additionally, the system searches the given disease name in the GWAS catalog table in the database and retrieves all of the related detailed genetic information. For every specific disease, all of these data items are stored in the MySQL database.

4.1.3.2 Analyzing VCF files

The input of the system is a Variant Call Format (VCF) file. VCF is a standardized text file format for representing genetic variations. The VCF file contains one record for every single genetic variation in one or a group of patients. There is no identifiable personal information in the VCF file.

In order to analyze the VCF file, the system first determines the number of patients in the VCF file, and then processes all of the genetic variations in the VCF. For every variation record, the system determines the detailed information, including dbSNP ID, reference allele, alternative allele, chromosome, position of the SNP in the chromosome, and the number of copies of the genetic variation in each patient. For every patient, the system extracts the genetic variations that

have one or more copies. The extracted information from the VCF file is stored in the MySQL database.

4.1.3.3 Identifying patients' variations related to the given disease

As shown in Figure 8, the system performs the following steps in order to identify patients' genetic variations related to the disease of interest:

1) The system identifies the disease of interest and performs all of the steps provided previously in the first step of data analysis (identifying the genetic information related to a certain disease).

2) The system then searches the table in the central database that stores all of the detailed genetic information of the disease of interest. Based on this table, the system retrieves a list of all the pathogenic SNPs that are related to the disease of interest.

3) The system searches the table in the central database that stores all of the patients' genetic variations that were extracted by analyzing the VCF file. From this table, the system identifies the genetic variations related to every patient. As mentioned earlier, these are the genetic variations with one or more copies.

4) The system identifies the chromosome numbers that are related to the disease of interest (the chromosome numbers can be retrieved from the OMIM table), the system then retrieves patients' genetic variations in these identified chromosomes. After that, for every genetic variation in the patient (only the genetic variations in the identified chromosomes), the system compares this genetic variation with all of the genetic variations related to the

disease of interest (which were identified in the first step of analysis). The comparison is done based on the related dbSNP ID. If this genetic variation matches any of the disease related genetic variations, the system then goes deeper and determines the specific alleles in this patient. If the alleles are matched between the patient's genetic variation and the disease-related genetic variation, the system then reports that this patient has a pathogenic variation related to the disease of interest. The process will be repeated for all of the genetic variations in the patient.

4.1.3.4 Identifying the corresponding pharmacogenomic information

The system identifies the genetic variations that have pharmacogenomic associations related to the disease of interest, pain and anesthesia (this information can be retrieved from the pharmacogenomics table in the central database). For every genetic variation in the patient, the system compares this genetic variation with all of the previously identified genetic variations that have pharmacogenomic associations. The comparison is done based on the related dbSNP ID. If this genetic variation matches any one of the genetic variations that have pharmacogenomic association, the system then goes deeper and determines the specific alleles in this patient. If the alleles are matched between the patient's genetic variation and the alleles provided in the pharmacogenomics table for this specific genetic variation, the system then reports the related pharmacogenomic associations with this genetic variation. The process will be repeated for all of the genetic variations in the patient.

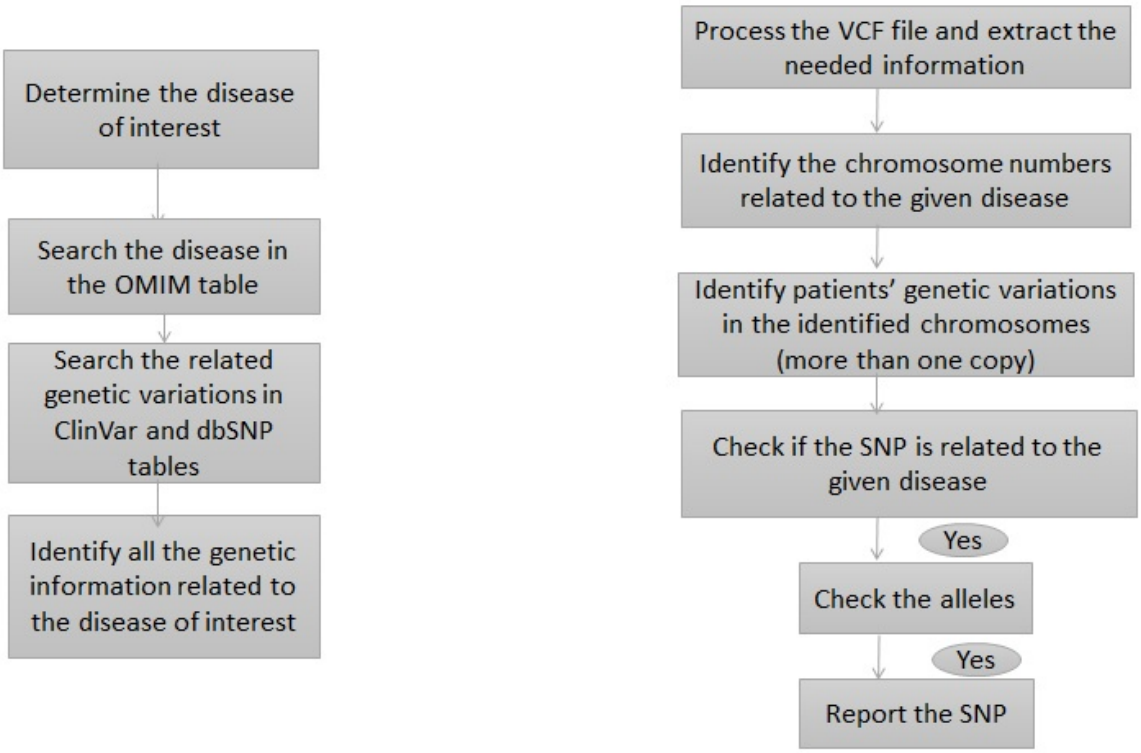

Figure 8. Information Flow in the Data Analysis

The system precalculates all the computationally intensive steps for every patient. For example, all four steps of data analysis (analyzing the VCF file, identifying the genetic variations, and identifying the related pharmacogenomic information for every patient) are pre-calculated and the results are stored in the database to be used by physicians.

4.1.4 Information delivery

Our system provides a graphical user interface (GUI) to facilitate the communication between the user and the system. The users of this system are physicians who may not have a strong background in genomics. Therefore, they need reports that are easy to read and understand

instead of typical research reports that include extensive details of analytical procedures and results. Typically, physicians don't want to see these data analysis details as long as they are convinced that the results from the system are reliable. This is how physicians treat many other types of laboratory reports; they do not need to know the specific lab technology or lab procedure that is used to obtain the results.

The system provides a search feature that helps to make the data accessible to physicians through simple queries. The system generates user-friendly summary reports after the desired information is retrieved from the central database. The generated reports include information about the patient's genetic variations related to the disease of interest. More specifically, the report has four sections: 1) patient's information such as patient ID; 2) result summary started with the word 'positive' or 'negative', which indicates whether the patient has the genetic variations related to the given disease; 3) disease information such as, relevant genes, the identified patient's genetic variations, and links to PubMed papers about the relevant genes and genetic variations; and 4) pharmacogenomic information such as information about certain medications, possible dosage, and the risk of adverse events in some cases. The report includes a link to a help page that provides a brief description of the basic parts of the report.

4.2 SYSTEM SECURITY

Based on HIPAA requirements [214-217] for database applications, our system has the following security measures:

4.2.1 Access control measures

User's access to our system is controlled by a set of security measures:

4.2.1.1 Access to the system

- Authentication and authorization service access, which means that every user needs to login with a unique username and password to gain access; only authorized people can access the data.

- Role-based access control: our system assigns roles to users in order to ensure the authorizing access to the data only when such access is appropriate based on the user's role. A role determines what a user is permitted to see and what operations a user can perform. For example, users with administrator privileges can access, retrieve, and update data. On the other hand, physicians can only view patients' genomic data without the ability to update anything.

4.2.1.2 Database access

- The central database is not accessible to the public, and physicians are not allowed to insert new information into the database. They can just upload VCF files and view the analysis results. Our system provides dropdown lists for selections. Therefore, physicians use these dropdown lists for selection.
- The central database is protected with a secure connection. Only authorized people who have a username and password can connect to the database. In our research project, only administrators can connect and access the database.

4.2.2 Encryption

The Advanced Encryption Standard (AES) algorithm is used to encrypt patients' genomic data, including patients' genetic variations, and all extracted information from a VCF file. AES is a symmetric block cipher algorithm published by the National Institute of Standards and Technology (NIST) in December 2001 [218]. AES uses four types of transformations in order to ensure data security, including substitution, permutation, mixing, and key-adding. AES provides several key lengths, including 128, 192, and 256. In our system, we used AES-128. This algorithm is simple and can be easily implemented with a minimum amount of memory and relatively low storage and hardware requirements. In the case of key transmission, the public key cryptography will be used to transmit the AES encryption key [219, 220].

5.0 RESULTS

This chapter discusses the results of our survey study about physicians' needs and expectations, as well as the results of our system.

5.1 THE SURVEY OF PHYSICIANS' NEEDS AND EXPECTATIONS

5.1.1 Study population

A total of 15 individual responses from physicians at the University of Pittsburgh were obtained. The participating physicians ranged in age from 31 to more than 60 years old. They were mostly male (80%). A large portion of the physicians have more than 30 years of experience in medical practice (40%).

5.1.2 Knowledge in genomics

As shown in Figure 9, forty percent (40%) of the participating physicians reported that they are quite familiar with genomics terms and genetic tests. Thirteen percent (13.33%) have an extensive knowledge in genomics. Twenty percent (20%) claimed to be experts in the field of genomics and felt highly confident in their abilities to deal with genomic information.

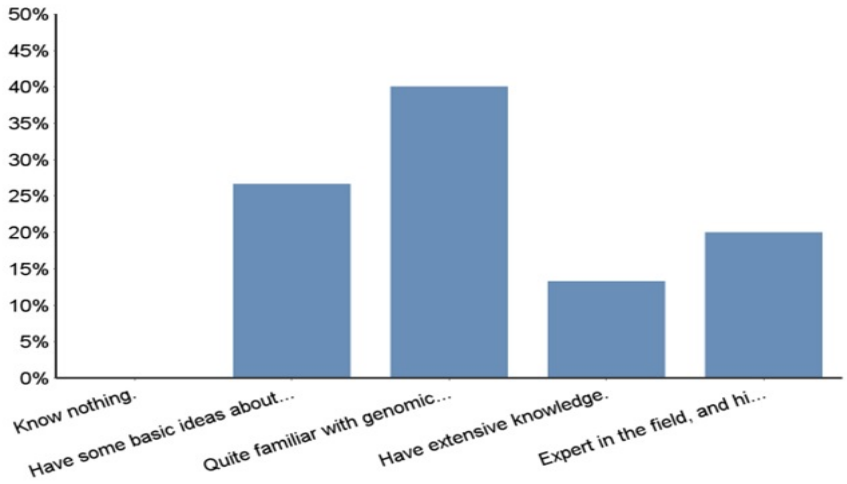

Figure 9. Overall Knowledge in Genomics

Figure 10 shows that seventy three percent (73.33%) of the participating physicians reported that they agree or strongly agree that enhancing their knowledge in genomics can be beneficial to their patients. Other physicians strongly disagreed with this claim. However, one needs to note that these respondents strongly disagreed with this claim, not the benefit of genomic knowledge to healthcare. After all, if 30 percent of the participating physicians are already experts or have an extensive knowledge in genomics, then enhancing their knowledge in genomics further will not produce much difference, and in turn, it will not produce additional benefits to their patients.

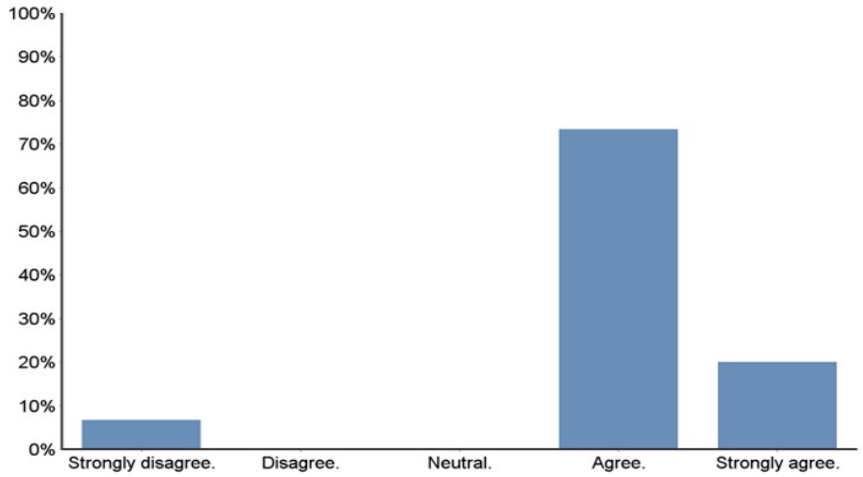

Figure 10. Enhancing Knowledge in Genomics

5.1.3 General opinions

As shown in Figure 11, the most frequently identified motivations for using genomics in clinical practice include (n=15): cancer treatment (26.76%), single gene disorder (20%), and pharmacogenomic analysis (20%). As shown in Figure 12, all of the participating physicians believed that new findings in genomics can change and improve the clinical practice (53.33% agree, and 46.67% strongly agree, n=15).

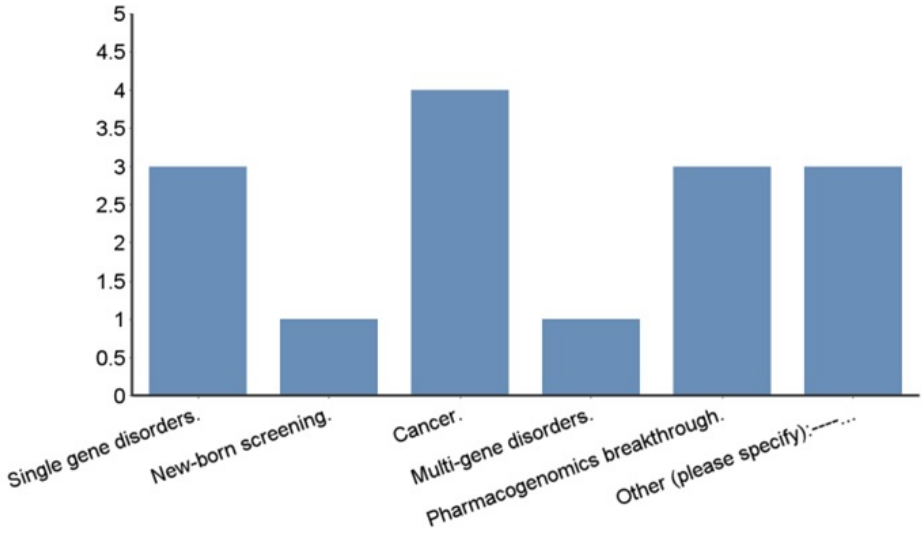

Figure 11. Motivations of Using Genomics

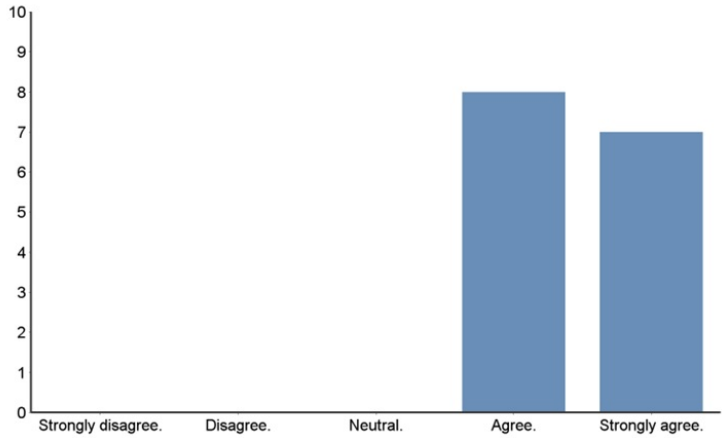

Figure 12. New Findings in Genomics

As shown in Figure 13, most of the participating physicians either reported that they agree (53.33%, n=15) or strongly agree (33.33%, n=15) that genomics should be incorporated into the clinical practice.

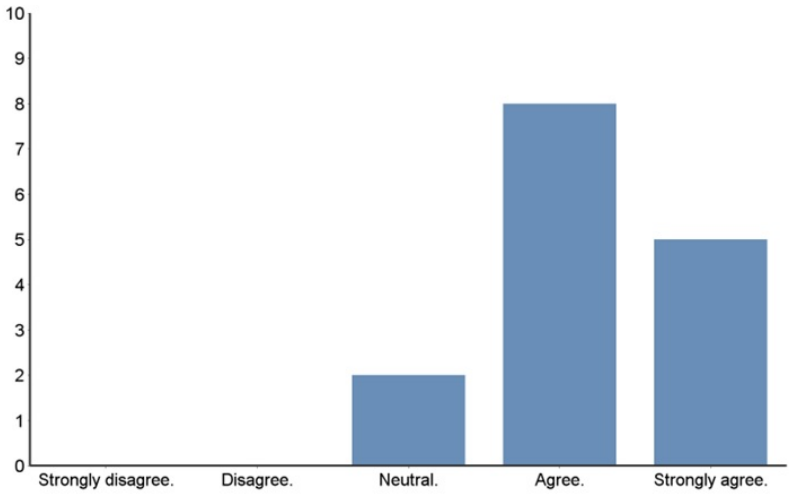

Figure 13. Incorporating Genomics in Clinical Practice

5.1.4 Specific genetic tests

The majority of the participating physicians indicated that they are either very likely to order genetic tests for single gene disorders (50%, n=14) or somewhat likely (14.29%, n=14). Twenty-one percent (21.43%, n=14) reported that they are somewhat unlikely to order genetic tests for single gene disorders. The participating physicians had different perspectives toward ordering sophisticated genetic tests for multiple gene disorders (38.46% somewhat unlikely, 15.38% neutral, 23.08% somewhat likely, 23.08% very likely, n=13).

The majority of the participating physicians agreed that genomics should be used to guide decisions about medication prescription, including dosage for each individual patient (66.67%, n=15). Fewer than seven percent (6.67%, n=15) strongly disagreed that genomics should be applied in personalized medicine. The majority of participating physicians reported that they agree that genomics can predict adverse drug reactions (64.29%, n=14); the rest of the study participants were neutral.

As shown in Figure 14, the majority of the participating physicians reported that they are confident in using genomics for personalized cancer treatment (57.14% somewhat confident, 35.71% very confident, n=14). Only one physician reported a neutral response toward the using of genomics in personalized cancer treatment.

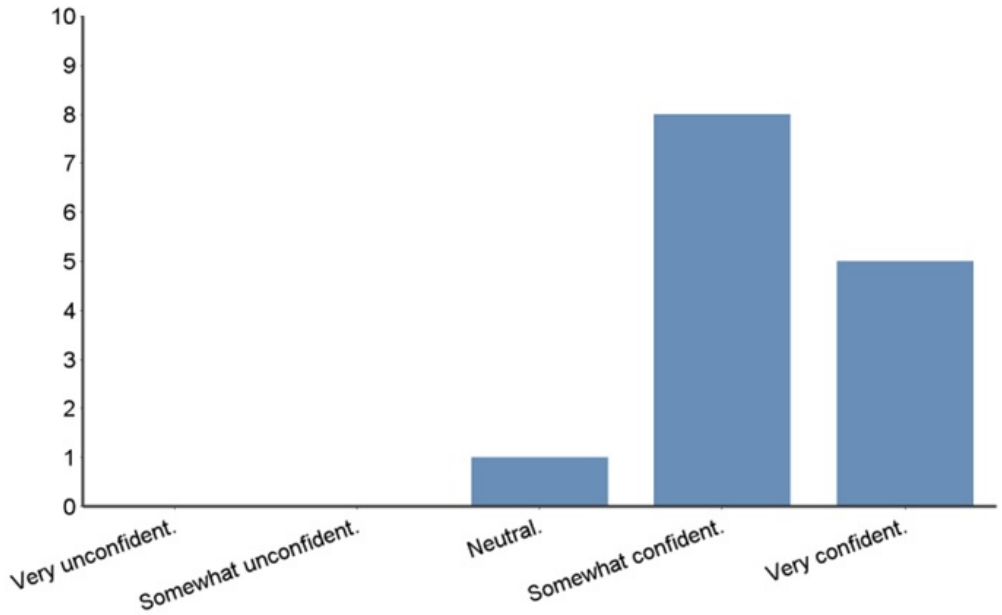

Figure 14. Personalized Cancer Treatment

Physicians have very high standards in terms of the accuracy of the genetic analysis results in a clinical report. As shown in Figure 15, sixty-six percent (66.37%, n=15) of the participating physicians reported that they can tolerate a small number of errors or uncertainty in the report, and twenty percent (20%) reported that they expect an absolutely correct report (no error at all in a report, no matter how complicated the tests or analyses can be).

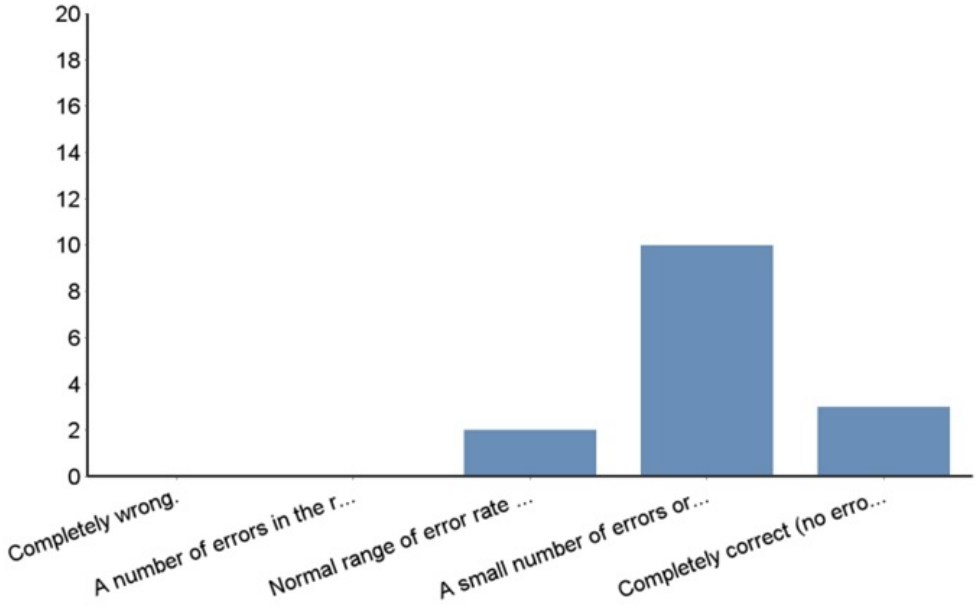

Figure 15. Desired Levels of Accuracy of Genetic Analysis Results

Different types of genetic tests were reported by the participating physicians as the most frequently ordered tests. Table 7 shows the reported genetic tests and a brief description.

Table 7. The reported genetic tests

Gene Testing	**Disorder Testing**	**Other Testing**
RAS family members (A family of genes that may cause cancer when they are mutated. These genes include Kras, Hras, and Nras)	CDH (a test for congenital diaphragmatic hernia defect)	Foundation One (a company provides a pan cancer test for a number of genes such as BRAF, BRCA1, and BRCA2)
BRCA (related to breast cancer)	cystic fibrosis	WES (Whole exome sequencing)
MMR (related to colon cancer)	BCR-ABL (the presence of this gene sequence confirms the diagnosis of chronic myelogenous leukemia and a form of acute lymphoblastic lymphoma)	

Table 7 (Continued)

PALB2 (related to breast cancer)	factor V Leiden (inherited disorder of blood clotting)	
MUTYH (related to colorectal cancer)	prothrombin gene variant 20210 (prothrombin is a bleeding disorder that slows the blood clotting process)	
ALK (related to lung cancer)	HFE genes for hemochromatosis (a disorder that causes the body to absorb too much iron from the diet)	
p53 (related to breast cancer)	HLA-B27 (a blood test to look for a protein that is found on the surface of white blood cells)	
PRSS1 screening (related to hereditary pancreatitis)	alpha-1 anti-trypsin (an inherited disorder that may cause lung disease and liver disease)	
DNA repair gene mutations	MSI PCR (which is related to lynch syndrome)	
	Hereditary gene panels (a test of multiple genes panel that identifies an elevated risk for important cancers).	

5.1.5 Expected features from a genomic information system

Based on the responses to the survey, we identified a set of desired features of a patient genomic information system. First, the system needs to be easy to search (93.33%, n=15); second, the information in the system needs to be updated periodically (93.34%, n=15); third, the data in the system need to be comprehensive and include complete information about diseases, related genetic variations and genes, and pharmacogenomic information (86.66%, n=15); fourth, the system needs to be easy to use and interpret results (100%, n=15); fifth, the system needs to be convenient to access at any time (100%, n=15); and finally, the system needs to be secure (53.33%, n=15). Table 8 lists the desired features of a genomic information system.

Table 8. Desired features of a genomic information system

Desired Features	Physicians Percent of Agreement
Easy to search	93.33%
Updated periodically	93.34%
Comprehensive	86.66%
Easy to use and interpret results	100%
Convenient to access	100%
Secure	53.33%

The majority of the participating physicians indicated that they want the genomic report to be stored in the EHR along with other patient information (85.71%, n=14). The majority of the participating physicians preferred to store the genetic analysis report for more than three years (92.86%, n=14). The majority of the participating physicians believed that genomic information systems should provide explanations for genomic test results interpretation (85.72%, n=14). Table 9 lists physicians' suggestions for a genomic information system.

Table 9. Physicians' suggestions for a genomic information system

Suggestions	Physicians' Percent of Agreement
Genomic reports need to be stored in the EHR	85.71%
Genetic analysis reports need to be stored for more than three years.	92.86%
Genomic information systems should provide explanations for genomic test results interpretation.	85.72%

As shown in Figure 16, the most frequently identified problems for applying genomics in the clinical practice include the usefulness of the genetic analysis result to physicians' treatment plan (53.33%, n=15) and insurance coverage for genetic tests (13.33%, n=15).

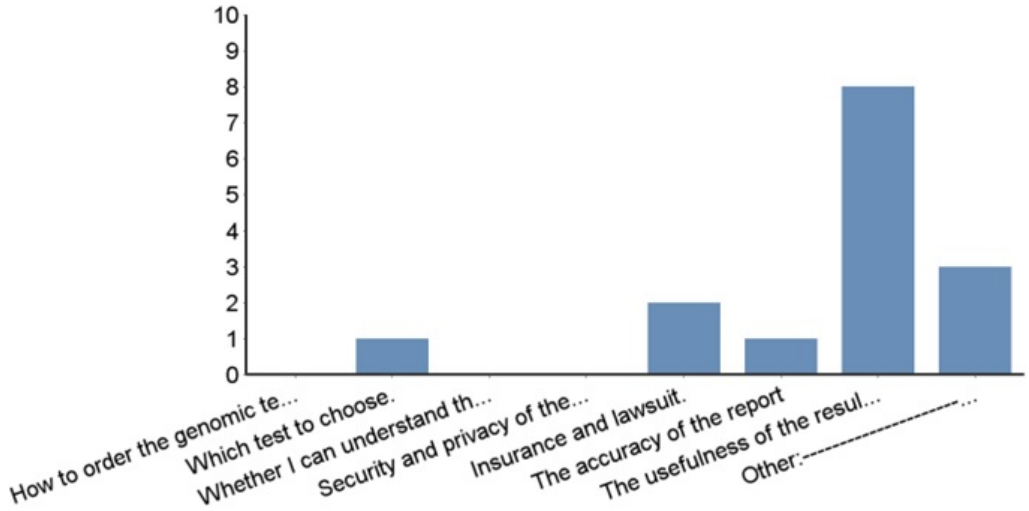

Figure 16. Problems of Applying Genomics in Clinical Practice

5.2 SYSTEM RESULTS

5.2.1 System data

Our system is mainly focused on the analysis of SNPs, which are the most common type of genetic variations in the human genome [74]. Thus, the system can screen and analyze all human SNPs available in ClinVar and dbSNP databases. Additionally, the system can search and analyze all the diseases available in the OMIM database and GWAS catalog.

The extraction scripts were re-executed on December 2015 to extract all the needed information from OMIM, dbSNP, ClinVar, GWAS catalog, PharmGKB, and GenBank databases. As a result, the current version of the system includes:

- 6,558 diseases such as Alzheimer's, Pancreatitis, and Autism.

- 199,054 human SNPs from ClinVar and 105,113 human SNPs from dbSNP, such as rs 150829393, rs 200401432, and rs 111033557.
- 331,770 human genes such as BRCA1, BRCA2, and PRSS1.
- 2,320 GWAS articles.
- 3,602 different pharmacogenomic associations.

5.2.2 Use case scenario

In this use case, the system is used to screen the personal genetic variations in 140 patients and to identify all of the genetic information related to Pancreatitis. At the beginning, the user needs to create an account in the system and then login to the system. Figure 17 shows the sign-up page. Figure 18 shows the login page.

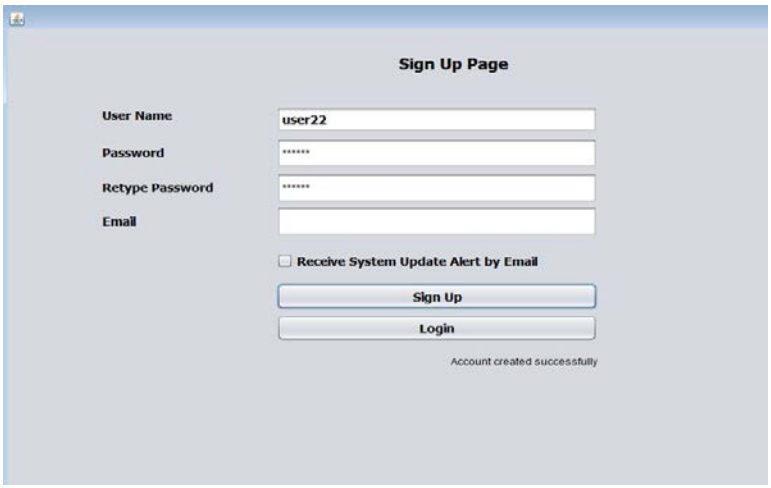

Figure 17. Sign-Up Page

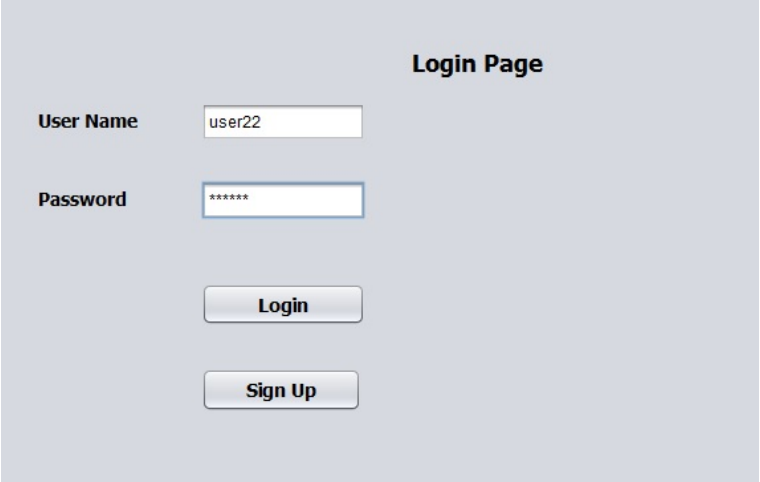

Figure 18. System Login Page

Then, the user gets into the system home page as shown in Figure 19. This page provides links to all of the pages in the system, including search for disease, analyze VCF files, show group reports, and show individual reports.

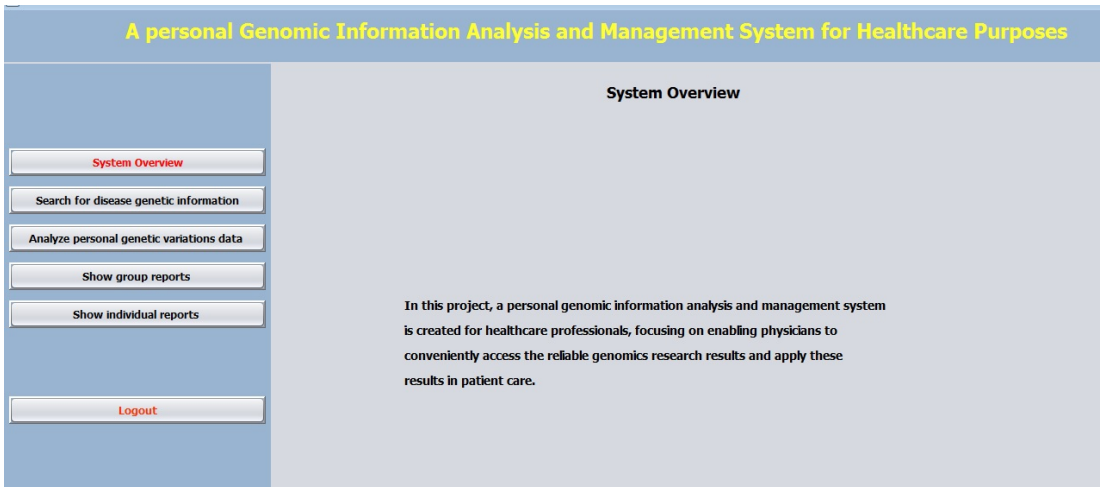

Figure 19. System Home Page

In the “search for disease” page, the user can search for any disease of interest. In this use case, the user needs to identify the genetic information associated with “Pancreatitis”. The user chooses the disease name from the dropdown list and then hits the button “search”. The system

then automatically searches the disease name in the OMIM database table and the GWAS catalog table and displays the results into a table as shown in Figure 20.

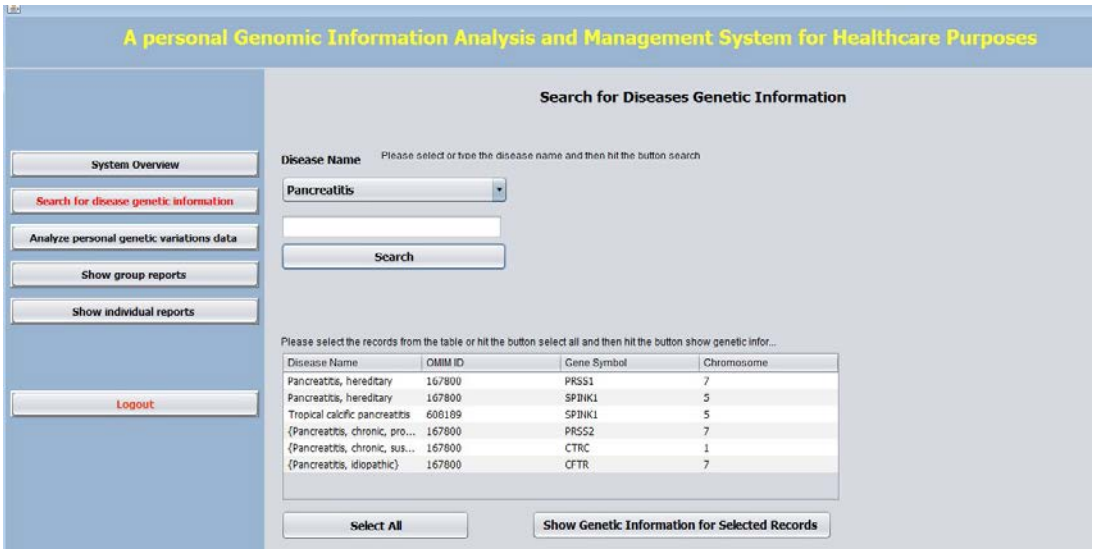

Figure 20. Search Page

The user can then select any of the specific kinds of Pancreatitis that are displayed in the OMIM table, and then hit the button “show genetic information” in order to get detailed genetic information about this specific disease. The user can also click the button “select all” in order to select all of the specific kinds of Pancreatitis.

In this use case, the user clicks the button “select all” and then clicks the button “show genetic information”. The system then gets the OMIM IDs and searches them into the ClinVar table in the central database in order to get all of the related genetic variations. The system also gets detailed genetic information for every genetic variation such as variant name, dbSNP ID, chromosome, clinical significance, reference allele, and alternative allele. Figure 21 shows the detailed genetic information related to Pancreatitis. The clinical significance types of the

identified SNPs include benign, likely benign, risk factor, uncertain significance, likely pathogenic, and pathogenic.

Detailed Genetic Information

Variant_Type	Variant_Name	Gene_Symbol	Clinical_Significance	dbSNP_ID	Chromosome	Reference_Allele	Alternative_Allele	Cytogenetic
single nucleotide variant	NM_007272.2(CTRC):c....	CTRC	Pathogenic;risk factor	121909293	1	C	T	1p36.21
single nucleotide variant	NM_007272.2(CTRC):c....	CTRC	Pathogenic;risk factor	121909294	1	G	A	1p36.21
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033565	7	G	A	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033566	7	A	T	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033567	7	A	G	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033564	7	G	A	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	144422014	7	A	G	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033568	7	C	T	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Uncertain significance	199422123	7	G	A	7q34
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	104893938	5	A	G	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	104893939	5	A	G	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	104893939	5	A	C	5q32
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	387906698	7	C	T	7q34
single nucleotide variant	NM_000492.3(CFTR):c....	CFTR	Benign	1800094	7	A	G	7q31.2
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Likely pathogenic	193922655	7	C	T	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Likely pathogenic	193922656	7	C	G	7q34
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Benign;Pathogenic	111966833	5	G	A	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Uncertain significance	141634296	5	C	T	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Likely benign	35877720	5	C	G	5q32
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	397507439	7	T	C	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	397507440	7	T	A	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	202003805	7	C	T	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	397507442	7	A	G	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Pathogenic	111033566	7	A	C	7q34
single nucleotide variant	NM_000492.3(CFTR):c....	CFTR	Benign	1800095	7	G	A	7q31.2
single nucleotide variant	NM_000492.3(CFTR):c....	CFTR	Benign;Likely benign	1800130	7	A	G	7q31.2
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	148954387	5	A	G	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	515726206	5	A	C	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	515726207	5	A	G	5q32
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Pathogenic	515726208	5	G	A	5q32
single nucleotide variant	NM_007272.2(CTRC):c....	CTRC	Pathogenic	515726209	1	G	A	1p36.21
single nucleotide variant	NM_003122.4(SPINK1):...	SPINK1	Benign	35523678	5	C	T	5q32
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Likely benign	606231344	7	T	A	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Likely pathogenic	199769221	7	G	C	7q34
single nucleotide variant	NM_002769.4(PRSS1):...	PRSS1	Uncertain significance	606231346	7	A	C	7q34

Figure 21. Detailed Genetic Information Related to Pancreatitis

In the “Analyze personal genetic variation data” page, the user can upload a VCF file and analyze it. In this use case, the user uploads a VCF file of 140 patients into the system. The system then processes the VCF file and identifies the SNP copies, reference, and alternative alleles for every patient. In the next step, the system executes the data analysis algorithms. As shown in Figure 22, the user can also see the recently analyzed VCF files.

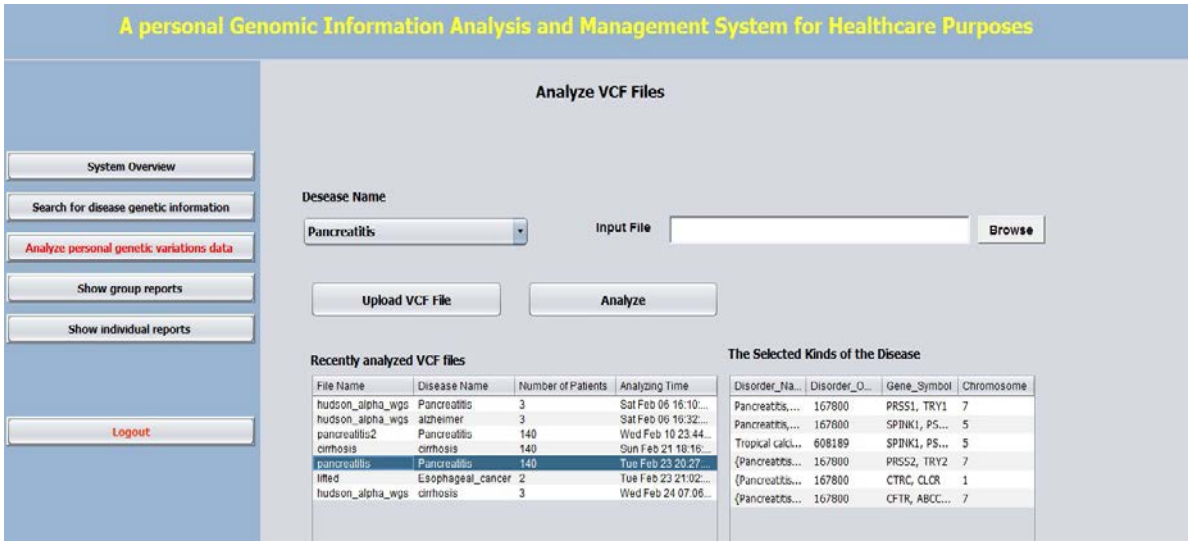

Figure 22. Analysis Page

In the “Show group reports” page, the user can select the disease of interest from the dropdown list, and the name of the VCF file from the list of recently analyzed files, and then click the button “Show report” or “Show pharmacogenomic report”. The user needs to enter a key in order to be able to show these reports as shown in Figure 23. Figure 24 shows the final group report.

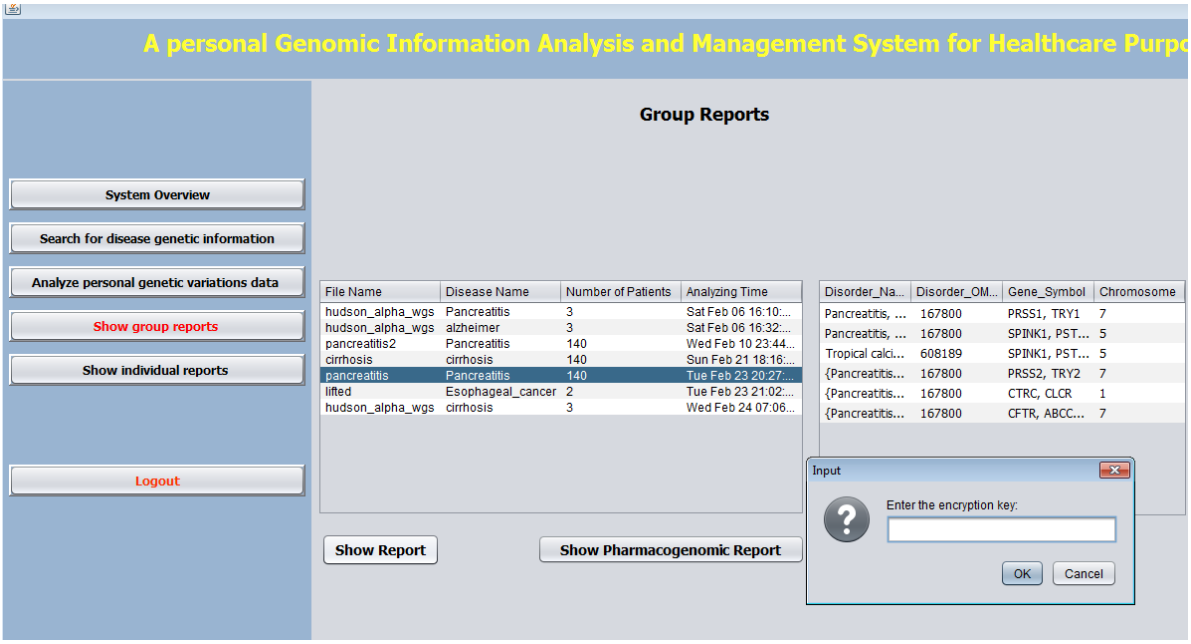

Figure 23. Show Group Report Page

Patient ID	SNP ID	Alleles	Variant Name	Gene Symbol	Chromosome
1	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
2	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
4	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
5	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
6	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
7	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
8	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
9	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
10	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
11	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
12	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
13	rs1800076	G/A	NM_000492.3(CFTR):c.224G>A (p.Arg75Gln)	CFTR	7
13	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
14	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
16	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
18	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
19	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
20	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
21	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
23	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
24	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
25	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
26	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7
27	rs111033566	A/T	NM_002769.4(PRSS1):c.86A>T (p.Asn29Ile)	PRSS1	7

Figure 24. Group Report

In the “Show individual report” page, the user can select the VCF file from the recently analyzed VCF files and enter the patient number. The user can then click the button “Show report” in order to view the final summary report as shown in Figure 25.

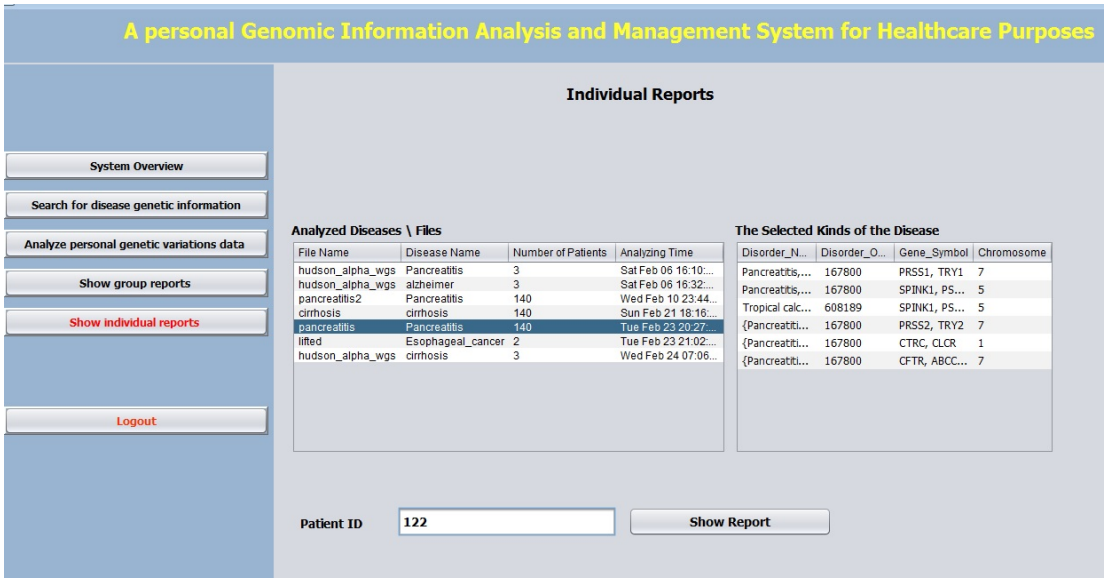

Figure 25. Show Individual Report Page

In this use case, the user enters the patient number 122 and then clicks the button “Show report”. The generated report is shown in Figure 26.

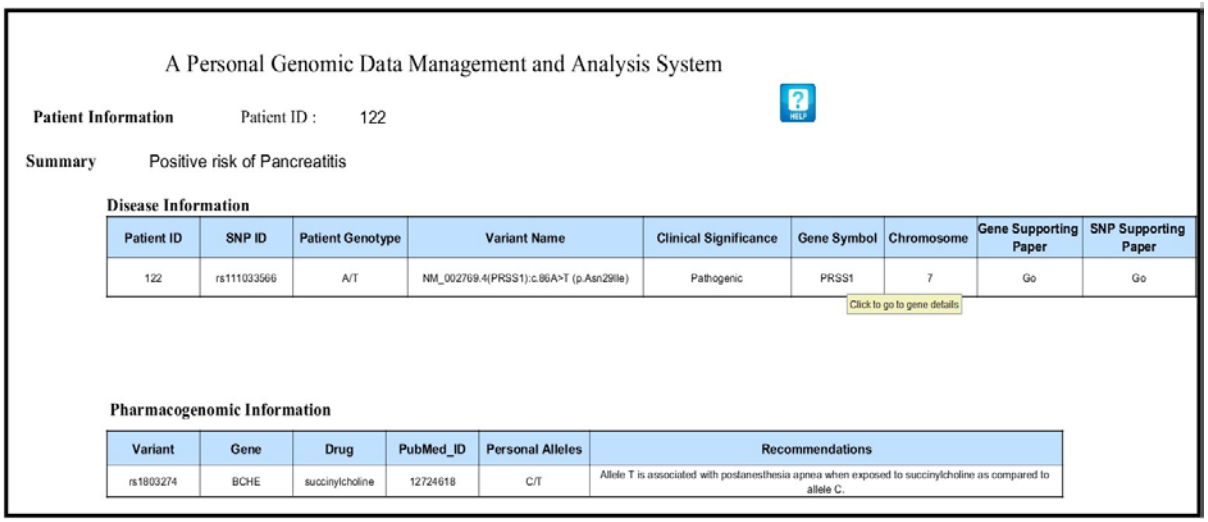

Figure 26. Final Report

As shown in Figure 26, the final report provides information about the patient genetic variations related to the disease of interest and the corresponding pharmacogenomic information.

In this use case, patient 122 has one SNP related to Pancreatitis. Thus, the patient has a positive risk of Pancreatitis. Additionally, the final report provides links to the dbSNP and GenBank databases, and links to supporting papers about the SNP, the gene, and the corresponding pharmacogenomic information. Figure 27 shows the related dbSNP and GenBank records, which were obtained by clicking on the SNP ID (rs111033566) and the gene symbol (PRSS1) in the final report shown in Figure 27.

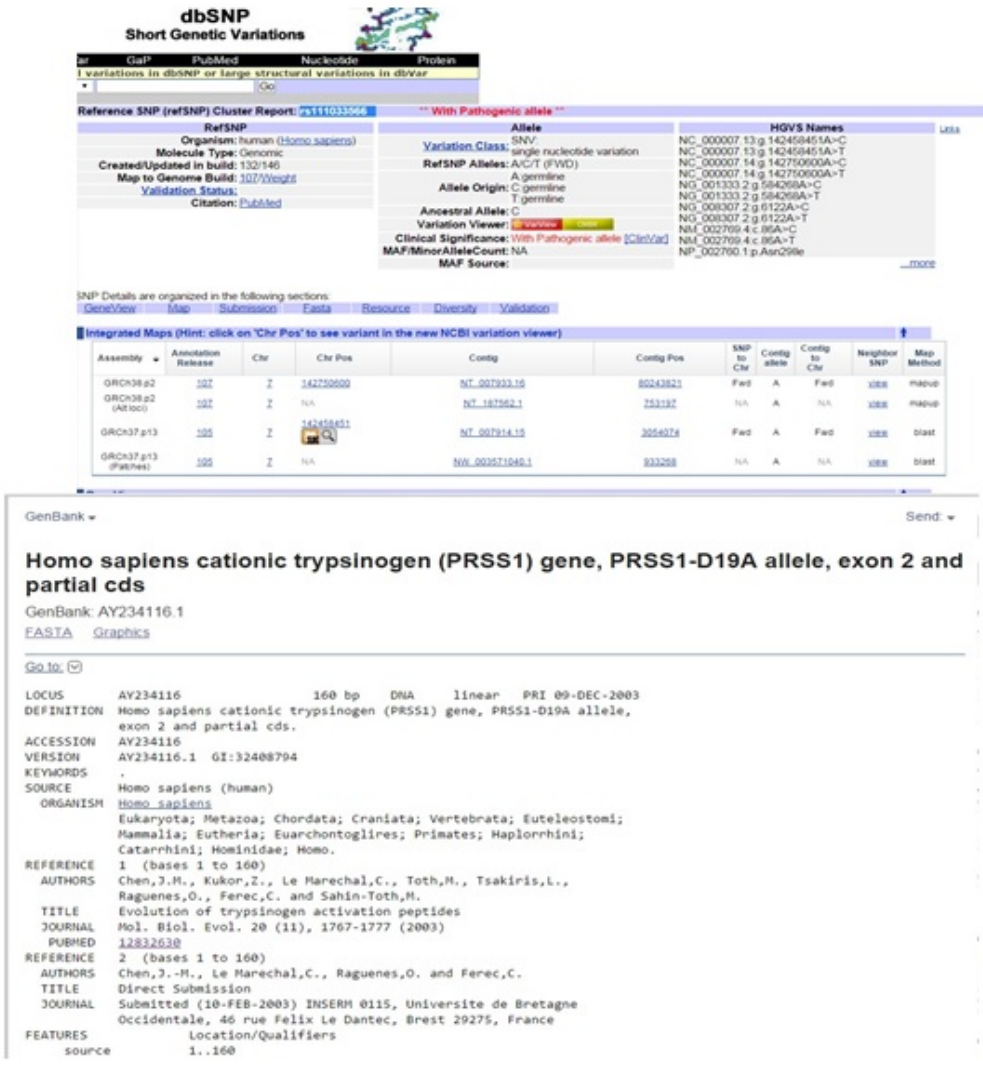

Figure 27. Related dbSNP and GenBank Records

5.2.3 System performance

The system was developed and tested using a Windows 7 machine with the following features: Intel core i7, CPU @ 2.20 GHz, and 8 GB RAM. MySQL community server version 5.1.73 was used to store different types of genomic data.

Several factors affect the time required to analyze patients' genetic variation data using our system, including the size of the VCF file, the number of patients in the VCF file, and the number of SNPs in the VCF file. We used our system to analyze three VCF files and we reported the time required to perform the analysis of every file. Table 10 provides the number of patients, number of SNPs, and the size of each file. Figure 28 shows the total analysis time for every VCF file.

Table 10. Multiple VCF files

File Name	No of Patients	No of SNPs	Size (GB)
File 1	140	552782	2.53
File 2	3	1685578	2.24
File 3	629	178695	6.4

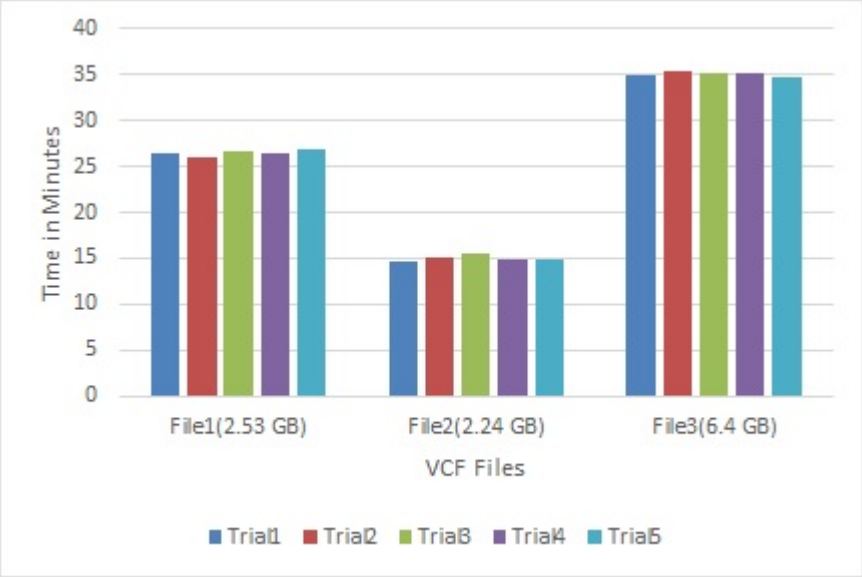

Figure 28. Total Analysis Time of Multiple VCF Files

.

6.0 EVALUATION

This chapter discusses the steps we have considered to evaluate the accuracy of the information and the analysis results provided by our system.

6.1 ALGORITHMS EVALUATION

Several steps were taken to evaluate the results of the extraction algorithms, which were used to extract the needed information from multiple data sources. The first step was to check the accuracy of the extracted data, which was done by performing SQL queries from the central database and comparing the results with the results from the existing databases such as dbSNP and ClinVar. The second step was to determine the integrity of the data in order to ensure that all genomic data stored in the database tables are consistent and meaningful.

In order to evaluate the results of extracting the needed information from VCF files, we compared the output of our VCF information extraction algorithm with the output from the VCF-Miner tool [221]. VCF-Miner is an existing tool for extracting genetic variation stored in the VCF file. We used our algorithm and the VCF-Miner to process and extract information from the same VCF file. We got the same number of genetic variations and the same number of samples for every variation. We also got the same detailed information about each variant such as chromosome, reference allele, alternative allele, and position in the chromosome. However, our

algorithm is able to extract the number of copies and the corresponding code of each variant in each patient. We need this specific information to be used in the data analysis algorithms in order to identify the patients with disease-associated genetic variations and pharmacogenomic associations. We customized the output of our algorithm to meet the format of the VCF-Miner output, so that we can easily compare the results. Figure 29 shows the output of our VCF information extraction algorithm and the output of VCF-Miner.

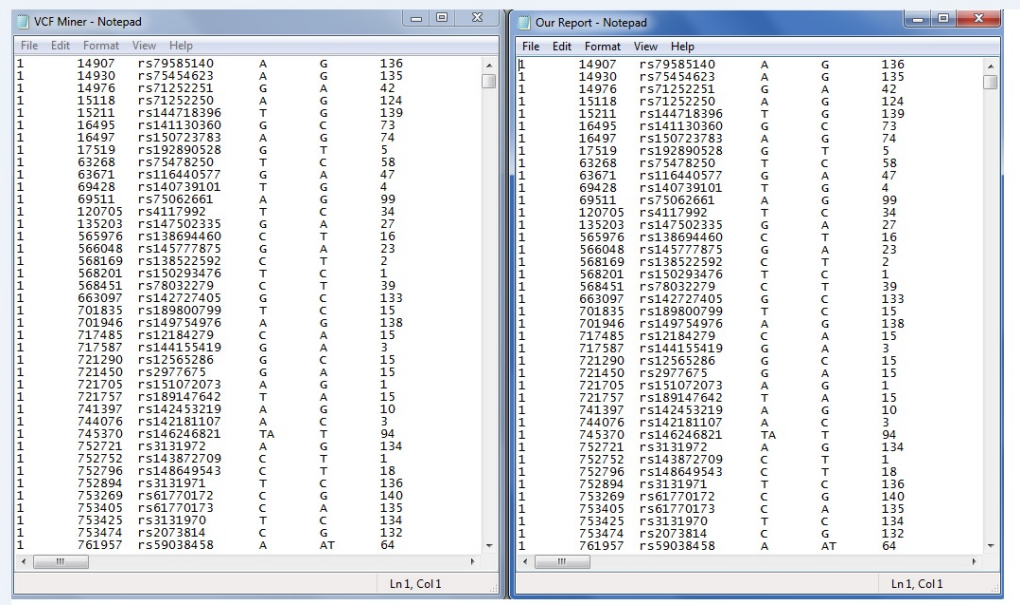

Figure 29. VCF File Information Extraction

In order to evaluate the data analysis algorithm that can identify patients' genetic variations related to a disease of interest, our system was used to analyze multiple VCF files, such as VCF files for Pancreatitis, Cirrhosis, and Esophageal cancer. Each file contains genetic variations data from real patients who were diagnosed with one of the aforementioned diseases. Our system was able to screen these VCF files, analyze them correctly, and detect the patients who have genetic variations related to the disease of interest.

6.2 SYSTEM USABILITY STUDY

6.2.1 Usability study methodology

The purpose of this usability study was to evaluate the usability of the stand-alone system we have developed for managing and analyzing personal genomic data. The participating physicians were asked to use the system and fill out a questionnaire. The participating physicians were asked to do a few tasks, such as creating a user account, searching genetic variation information for a few diseases, viewing genomic analysis results, and viewing the final reports. The participating physicians were asked to fill out a paper-based questionnaire about the critical features of the system such as, security of the system, ease of use, and ease of accessing results. This paper-based questionnaire was distributed to the participating physicians before they perform the tasks on the system. The questionnaire has questions about specific tasks performed on the system. This usability study was conducted in the participating physicians' own offices or in the School of Health and Rehabilitation Sciences. (The usability study questionnaire is provided in Appendix B.)

The study protocol was reviewed and approved by the University of Pittsburgh Institutional Review Board (IRB #: PRO15110267). The study was considered to be exempt because respondents were anonymous and there was no risk to participants.

6.2.2 Usability study design

The design of our usability study questionnaire was guided by the IBM post-study system usability questionnaire (PSSUQ) [222]. This PSSUQ aimed to address five system usability

characteristics: quick completion of work, ease of learning, high-quality documentation, functional adequacy, and rapid acquisition of usability experts [223].

The usability study questionnaire is divided into five parts; four of them are related to four different tasks, and the last part includes post-task overall questions. The four tasks are as follows: 1) creating accounts in the system; 2) searching for diseases; 3) displaying the detailed genetic information related to the disease of interest; and 4) displaying the final genomic analysis report. The post-task overall questions assess physicians' overall satisfaction with the system and the information provided by the system.

6.2.3 Usability study results

SPSS (Version 23) was used to analyze the result of the usability study questionnaire. All questionnaire results were tabulated and analyzed using descriptive statistics.

6.2.3.1 Study population

A total of six physicians from the University of Pittsburgh participated in this usability study.

6.2.3.2 Task 1: Creating own account in the system

All of the participating physicians reported that it is easy to create a username and password in the system (100%, n=6). Furthermore, all the participating physicians considered the system to be a secure system (100%, n=6).

6.2.3.3 Task 2: Searching for diseases

All of the participating physicians reported that it is easy to search the genetic information related to a certain disease in the system (100%, n=6). Additionally, all of them reported that the search results are easy to read and understand (100%, n=6).

6.2.3.4 Task 3: Displaying the detailed genetic information related to the given disease

As shown in Figure 33, eighty four percent (84%, n=6) of the participating physicians reported that the system provides sufficient genetic information about the disease of interest. Additionally, all of them reported that this detailed genetic information is easy to read (100%, n=6).

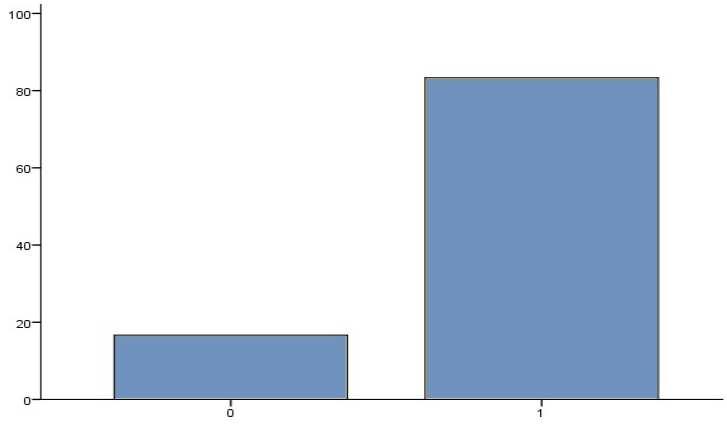

Figure 30. Genetic Information Provided in the System

6.2.3.5 Task 4: Displaying the final genomic analysis report

As shown in Figure 32, eighty four percent of the participating physicians reported that the report is well-formatted (84%, n=6). All of the participating physicians reported that the final genomic analysis report is easy to read and understand (100%, n=6). All of the participating physicians

reported that it is useful to include links to other important genomic databases such as dbSNP and GenBank (100%, n=6).

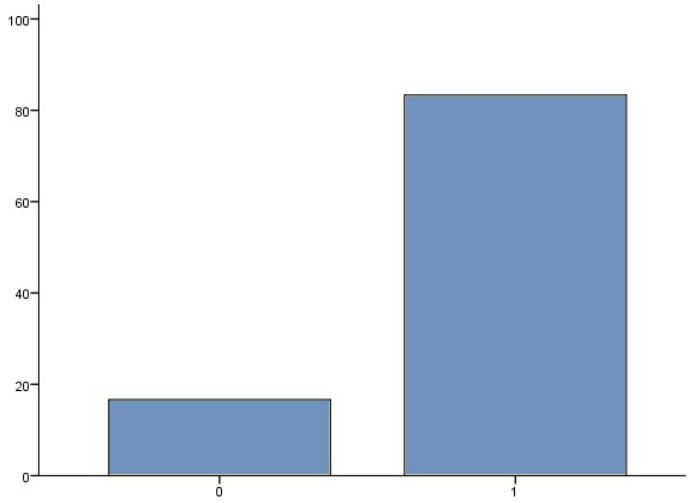

Figure 31. The Final Report Format

As shown in Figure 32, sixty seven percent (67%, n=6) of the participating physicians reported that the final report provides useful guidelines about each section in the report. One of the participating physicians suggested that the report can be improved by adding titles to each part of the report. Additionally, it is important to allow the user to hover over each part of the report in order to get a brief description.

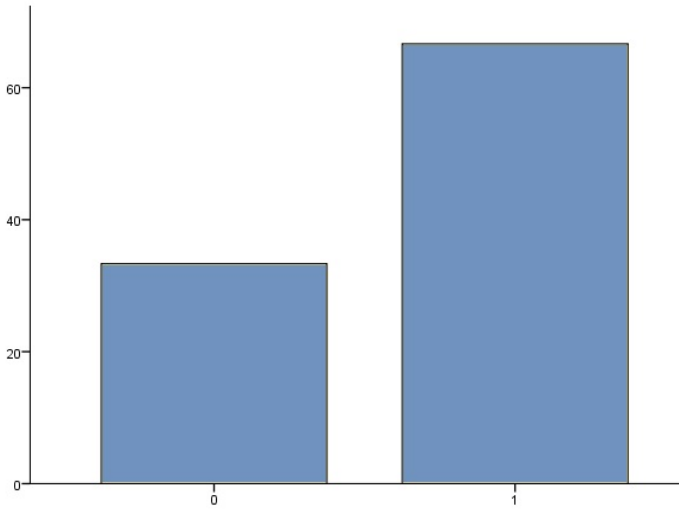

Figure 32. Final Report Guidelines

6.2.3.6 Post-task overall questions

Overall, the participating physicians were satisfied with the system and the information provided by the system. Table 11 provides physicians' responses to the post-tasks overall questions.

Table 11. Post-tasks overall questions

Question	Yes	No	Number of Participants	Percent Agreement
Overall satisfaction with how easy is it to use the system.	6	0	6	100%
It was simple to use the system.	6	0	6	100%
I was able to complete the tasks quickly.	6	0	6	100%
I felt comfortable using the system.	6	0	6	100%
It was easy to learn to use the system.	6	0	6	100%
It was easy to find the information I need.	6	0	6	100%

Table 11 (Continued)

The information provided in the system is easy to understand.	6	0	6	100%
I liked using the interface of the system.	6	0	6	100%
Overall, I am satisfied with using the system.	6	0	6	100%

6.2.3.7 Recommendations

Based on the information provided by the participating physicians in the usability study, we have considered certain changes in the system interface. Table 12 reviews the recommended changes and provides a brief justification of each change.

Table 12. Recommended change**s**

Change	Justification
Add titles for every part of the final report.	These titles can increase the report's usability by making it more scannable.
Add tooltips (a message that appears when a cursor is positioned over a text) for every part of the final report.	These tooltips can provide additional descriptions of the major parts of the report and help physicians to understand the report.

7.0 DISCUSSION

This chapter provides a discussion of the results and limitations of our survey study about physicians' needs and expectations. The chapter also discusses the results of the system usability study and the limitations of the system.

7.1 THE SURVEY OF PHYSICIANS' NEEDS AND EXPECTATIONS

Two alternative methods other than a survey could be used to collect qualitative data about the current status of using genomics in clinical practice and the expected features of a patient genomic information system. The first is an interview. In an interview, researchers can have a structured conversation with study subjects. Interviews can capture subjects' points of view about a specific topic [224, 225] in depth. The second is a focus group. During a focus group study, a group of 6 to 12 [226] people are brought together by a researcher in a social space where the participants interact, discuss, and acquire knowledge about a specific topic [227, 228]. The researcher (also the moderator) can collect desired information by taking notes.

Each method has its own advantages and disadvantages. Interviews provide an opportunity to explore topics in depth. The researcher (interviewer) can explain and clarify questions, which may increase the likelihood of receiving useful responses. The researcher can also ask further questions during an interview to gain more in-depth information. On the other

hand, the result of an interview is influenced by personalities, moods, and interpersonal dynamics between the interviewer and the interviewees. It can be difficult to arrange a suitable place and time between researcher and interviewee, which usually results in a small number of people interviewed in a given period of time.

A focus group generates quick and effective results. It is faster than an interview or survey [229]. It can produce high-quality data because the moderator can respond to questions and ask for clarification and more detailed responses [226]. On the other hand, focus group study has some disadvantages such as susceptibility to bias since the results can be affected by the researcher's point of view [226]. Data from focus groups are difficult to analyze since the discussions should be audio-taped or videotaped in addition to the notes taken during the discussion. All these data should be transcribed verbatim. Although there are qualitative software programs such as NVivo that can review the transcribed reports and provide themes and graphics, the large volume of qualitative data may be difficult to analyze [230]. Furthermore, the results from focus groups are hard to be generalized to a larger population since they strongly depend on the participants' selection and the small number of participants [230].

Given these disadvantages of interview and focus group methods, we decided to use the survey method to collect the information we need from physicians. Several reasons can justify our selection: 1) a survey can be administered remotely via the internet, which can reduce or prevent geographical dependence; 2) a survey can collect data from a large number of respondents; 3) numerous questions can be asked about a subject, which provides extensive flexibility in data analysis; and 4) some statistical techniques allow the researcher to analyze survey data and determine validity, and statistical significance. Surveys still have some disadvantages: respondents may not provide accurate answers and they may not feel comfortable

providing answers that present themselves in an unfavorable manner. Additionally, in the case of convenience sampling, the sample may not represent the population as a whole. Thus, the results of the survey can be biased.

The survey study was a small exploratory study. However, analysis of the responses indicates that the majority of the participating physicians (73.33%) claimed to have sufficient knowledge in genomics (40% of them are quite familiar with genomics, 13.33% of them have extensive knowledge in genomics, and 20% of them are experts in the field). Additionally, the majority of them (86.67%) believed that genomics should be incorporated into the clinical practice, and they all agreed that genomics can improve clinical practice. Moreover, the majority of the participating physicians (92.85%) felt confident toward using genomics in personalized cancer treatment, and 64.29 percent of them indicated they are likely to order tests for a single-gene-disorder. However, only 46.16 percent of them indicated they are likely to order tests for multiple-gene-disorder.

At first glance, these numbers seem to be showing conflicting results: the majority of the physicians have sufficient knowledge in genomics and strong motivation to apply genomics into their clinical practice, on the other hand, close to half or more than half of them are unlikely to order tests for single- or multiple-gene-disorders. One reason that may justify this finding is that there are many conflicting research results reported in the current genetic and genomic research on diseases. Physicians need support from the research community to make selections on the most reliable research results. Before that support is available to some physicians, their best choice is probably not to order those genetic tests or genomic analyses [231].

Based on the survey findings, we can find that physicians are willing and ready to incorporate genomics into their clinical practice. However, they need help or tools to do that.

One approach is to help them to make the appropriate selection on research results. Continued education is one of the strategies to enhance the physicians' skill in selecting results from the latest genomics research. The other approach is to provide a patient genomic information system with their desired features: easy to use, secure, easy to access, and providing useful genetic information.

A well-designed patient genomic information system for clinical purposes can help physicians to easily incorporate genomics into their clinical practice and to make their tasks more convenient. Our study identified physicians' expectation about the features that motivate them to use a genomic information system. They should prove very helpful when people implement such systems in the future.

This study has a number of limitations. First, the survey doesn't assess all DOI domains that may affect the innovation diffusion such as personal factors or the environment in which the physician works. Second, our sample is a convenience sample of physicians at the University of Pittsburgh and the number of responders is small, which may narrow the confidence and generalizability of the findings. Third, the responses represent physicians' self-reported data, which means that in some cases physicians might not provide accurate information.

7.2 THE SYSTEM

According to the current understanding between genetic variations and diseases, for most common diseases, the genetic variations identified in GWAS articles can increase the risk of developing the corresponding disease(s). However, patients who have some of those variations may not develop those diseases during their lifetime. In addition, these genetic variations may be

helpful for diagnostic purposes, though they may not be very useful for the treatment of diseases if they do not have any pharmacogenomic consequences. After all, at this moment, technologies such as gene therapy are still not reliable. Therefore, physicians may not have a very strong motivation to use genomic information in their clinical practice [10]. The complexity of the genomic data and molecular biology databases also keeps physicians away from accessing genomic information. Our system is one step to help physicians utilize genomic information in their clinical practice. It also facilitates the translation of results from the study of human genetic variation into clinical practice, which is a desired goal of the current research and is likely to have the most immediate impact through pharmacogenomics studies [141] that aim to maximize the treatment benefit for the patient while minimizing the risk of adverse effects. As the understanding of genetic risk and its relationship to other risk factors becomes clearer, the opportunities for personalized medicine will increase.

One limitation of our system is that it mainly uses the genetic variation results from the OMIM database and the GWAS method. Even though this approach can cover most available results in variation and disease association studies, it does miss results identified with other methods, which are not stored in the OMIM database.

Another limitation of the system is that it does not take into account family relationships. It is actually quite common for multiple members of the same family to see the same doctor or go to the same hospital. Therefore, if their genomic information is available, it is likely to be stored in the same system. In this case, the system should be able to perform linkage analysis [139] and family-based analysis [232]. After all, genome sequences in family members are highly similar and the association between genetic variation and certain diseases can be more confidently determined than in a population samples with independent individuals.

The results of our system usability study indicate that physicians could easily find the patient information they need and the information can be directly applied in their clinical practice. Although the number of participating physicians in this usability study is small, it is clear that physicians are satisfied with the system and the information provided by the system. They appear to be willing to use the system in their clinical practices. Additionally, they are interested in seeing how the results from this system can change their diagnosis and treatment plans for their patients. One reason that may justify physicians' satisfaction with our system is that the system was properly designed based on the desired features and suggestions obtained from the participating physicians in our survey study. These features help to motivate physicians to use a genomic information system in their clinical practice.

8.0 FUTURE WORK

This research project provides a new resource to facilitate precision medicine, which is an emerging approach for disease treatment and prevention taking into account individual variability in genes, lifestyle, and environment of each person [12]. One future plan of this project is to integrate and connect the system with other systems, such as Electronic Health Records (EHRs). The information provided in our system can create great opportunities to influence clinical care. However, applications are limited by the current EHR designs. According to the Vanderbilt University Medical Center (VUMC), the design of a Pharmacogenomics-Enabled EHR should provide some basic functions [233] such as, providing timely access to clinically significant genetic variations; providing a preemptive identification of patients who are expected to benefit from the knowledge of genetic variation analysis in order to tailor future treatment options; facilitating genotyping of patients with immediate clinical needs; and rapidly distributing genetic analysis results to laboratory, patient portal, inpatient, and outpatient prescribing environments.

The current EHR design requires significant modifications to incorporate genetic data in an actionable format [234, 235]. One limitation is the lack of structured data fields for storing the genetic test results in the EHR, which can be used for clinical decision support [236]. Even if these structured data fields are available, there is still uncertainty about how to report and integrate the genetic results with the existing clinical vocabulary.

Given the current limitations of EHR design, our system can act as an ancillary system for the EHR. This ancillary system has many advantages, such as reducing the storage requirements in the EHR since all of the patients' genetic information will be stored in the ancillary system, and allowing for ongoing review and interpretation of the results as needed using algorithms that are not available in the EHR.

The big data approach [237] can be used to integrate information from patients' EHRs, mobile health apps, which can collect many types of patient information such as physical activities and diet, and the genomic information stored in this system. Integrating all these types of data together can provide a comprehensive understanding about patients. This integration can lead to a significant improvement in the quality of health care.

One future plan is to incorporate our system into a specific clinic such as a liver diseases clinic. In this case, the physician needs to identify the disease of interest and uploads a VCF file containing the genetic variations of his/her patients. The system will then analyze the VCF file, screen all of the genetic variations in this group of patients, and then identify the ones who have genetic variations associated with the disease of interest. Our plan is to follow the patients for a specific period of time, such as six months, and based on the analysis results provided by the system, we can determine how the system can affect the diagnosis and treatment options for patients.

At this point, our system can handle and analyze uncompressed or compressed (*.gz) VCF files. Additional formats can be considered in the future. It is important to ensure that these VCF files are obtained from CLIA-certified laboratories, which comply with the federal Clinical Laboratory Improvements Amendments (CLIA) certification as administered by the Centers for

Medicare and Medicaid Services. CLIA creates quality standards to ensure the accuracy, reliability, and timeliness of patient test results [238].

The pharmacogenomic information provided by our system is extracted from the PharmGKP database. Thus, the resulted pharmacogenomic associations are applied to the literature curated by PharmGKB. There may be more literature in the public domain to support or contradict the resulted associations. At this point, pharmacogenomic information provided by our system can be used for research purposes only. Thus, it can’t be applied to the clinical practice.

9.0 CONCLUSION

Our survey study evaluated physicians' status in using genomics in their clinical practice. Based on the survey findings, we realize that several factors affect the success in integrating genomics in clinical practices and the implementation of personalized medicine such as desired support or tools for physicians. Although physician's knowledge in genomics has been enhanced in recent years, they still need support from the genomics community to make selections on the latest research results. The desired genomic information management system should have the ability to make all physician-desired data readily accessible to physicians in a convenient manner.

In this research project, we have developed a system to manage and analyze large amounts of genomic data and to enable convenient extraction of genetic information for physicians. Our system provides one encapsulated and presentable place for various types of genomic information needed by physicians, so that they can conveniently access the desired patient's genetic information and current research results at a single place. The system is able to screen and analyze all of the genetic variations in the patient and then identify the genetic variations that are associated with the disease of interest. The system identifies the clinical significance of every single genetic variation in the patient. It also provides the corresponding pharmacogenomic information for every patient. One important advantage of the system is that it allows physicians to get their desired results without extensive training in genomics. The results of our system usability study indicate that physicians are satisfied with the system and the

information provided by the system. They appear to be willing to use the system in their clinical practices. Additionally, they are interested in seeing how the results from this system can change the diagnosis and treatment plans of their patients.

Appendix A

SURVEY QUESTIONS

General Questions

1) Please indicate your gender:

A: M.

B: F.

2) Which age range are you in?

A: <= 30.

B: 31-35.

C: 36-40.

D: 41-45.

E: 46-50.

F: 51-55.

G: 56-60.

H: >60.

3) Please indicate your field of practice (listed alphabetically):

A: Cardiology.

B: Dentistry.

C: Dermatology.

D: Endocrinology.

E: Family.

F: Gastroenterology.

G: General.

H: Hepatology.

I: Nephrology.

J: Neurology.

K: OB/GYN

L: Oncology.

M. Ophthalmology.

N: Otolaryngology.

O: Pediatrics.

P: Pulmonology.

Q: Rehabilitation.

R: Rheumatology.

S: Other (please specify):----------

4) How many years of experience in medical practice do you have?

A: 1-5.

B: 6-10.

C: 11-15.

D: 16-20.

E: 21-25.

F: 26-30.

C: More than 30 years.

Knowledge

5) How would you rate your overall knowledge in genetics?

A: I know nothing about genetics or genetic tests.

B: I have some basic understanding about the principles of genetics.

C: I am quite familiar with genetics terms and tests.

D: I have extensive knowledge.

E: I am an expert in the field, and highly confident in dealing with genetic information.

6) Do you agree that enhancing your knowledge in genetics may be beneficial to your patients?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

General Opinions

7) Do you believe that genetic testing should be incorporated into clinical practice?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

8) Which of the following reasons would motivate you to seek genetic information for patient care?

A: Single-gene disorders.

B: Newborn screening.

C: Cancer.

D: Multi-gene disorders.

E: Pharmacogenomic breakthrough.

F: Other (please specify):----------

9) Do you believe that new findings in genetics can change clinical practice, even though it may or may not be true in your particular field?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

Specific Tests

10) How likely are you to order tests for common single-gene disorders?

(For example, the BRCA1 or BRCA2 mutations linked to hereditary breast cancers. Many cases of cystic fibrosis can be traced back to a mutation in the CFTR gene.)

A: Very unlikely.

B: Somewhat unlikely.

C: Neutral.

D: Somewhat likely.

E: Very likely.

F. Not Applicable

11) How likely are you to order sophisticated genetic tests for complex multiple-gene disorders? (For example, asthma, Alzheimer's, cancer, cardiovascular disease, obesity, and restless leg syndrome (RLS).)

A: Very unlikely.

B: Somewhat unlikely.

C: Neutral.

D: Somewhat likely.

E: Very likely.

F. Not Applicable

12) Do you agree that patients' genetic information should be used to guide decisions about medication utilization or dosage?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

F. Not Applicable

13) Do you believe that genetics can be used to predict adverse drug reactions for some patients?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

F. Not Applicable

14) How confident are you that genetics will be used as a method for personalized cancer treatment for patients in the next 10 years?

A: Very unconfident.

B: Somewhat unconfident.

C: Neutral.

D: Somewhat confident.

E: Very confident.

F. Not Applicable

15) Please indicate your level of confidence in your ability to make medical recommendations based on genetic data obtained from genetic testing:

A: Very unconfident.

B: Somewhat unconfident.

C: Neutral.

D: Somewhat confident.

E: Very confident.

16) How likely are you to need to consult with genetic counselors or specialists to interpret the result of a genetic test?

A: Very unlikely.

B: Somewhat unlikely.

C: Neutral.

D: Somewhat likely.

E. Very likely.

F. Not Applicable

17) What is the lowest level of accuracy that you would accept from genetic testing reports?

A: Completely wrong.

B: A number of errors in the report.

C: Normal range of error rates in a medical report.

D: A small number of errors or uncertainty in the report.

E: Completely correct (no errors at all).

F. Not Applicable

18) How often do you believe that genetic testing reports need to be updated?

A: Every week.

B: Every month.

C: Every 6 months.

D: Once a year.

E: Every two years.

G: Every ten years.

H: No need to make update on the report.

19) How likely would you be to use a patient's family history to support clinical decisions?

A. Very unlikely

B: Somewhat unlikely.

C: Neutral.

D: Somewhat likely.

E: Very likely.

Expected Features from a Patient Genome Information Management System

(A patient genomic information system aims to integrate genetic information from multiple sources, integrate them, and organize them into standard format.)

20) Do you agree that genetic information management systems should be easy to search based on keywords such as disease name?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

21) Do you agree that the information used for interpretation in a genetic information management system should be updated periodically (for example, every month, twice a year, or once a year) to keep pace with new clinical observations and validated research findings?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

22) Do you agree that genetic information management systems should be comprehensive (including complete information about diseases, related genetic variations and genes, and pharmacogenomic information)?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

23) Do you agree that genetic information management systems should be easy to use and provide information/recommendations in an easy to interpret format?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

24) Do you agree that genetic information management systems should provide some explanation for genomic test result interpretations?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

25) Do you agree that genetic information management systems should be easily accessible (available anytime and anywhere, i.e, through an online portal)?

A: Strongly disagree.

B: Disagree.

C: Neutral.

D: Agree.

E: Strongly agree.

26) If you were to get a genetic testing report for your patient from a personal genome information management system, where would you prefer to store it?

A: In the EHR system I use to store all other patient information.

B: In a separate database.

C: On my local hard drive or an external hard disk.

D: On a CD/DVD.

E: On a flash drive.

F: Other:----------

27) How long do you believe genetic test reports should be stored?

A: 1 - 3 months.

B: 4 - 6 months.

C: 7 months- 1 year.

D: 1 year- 3 years.

E: More than 3 years.

F. Whatever hospital administrators decide is an acceptable length of time.

28) What is the biggest hindrance to using genetic information in your clinical practice?

A: How to order the genomic tests.

B: Which test to choose.

C: Whether or not I can understand the report.

D: Security and privacy of patient's genetic information.

E: Insurance.

F: The accuracy of the report

G: The usefulness of the result to my treatment plan

H: Other:----------

29) Do you believe genetic information should be encrypted?

(Encryption/Decryption slows down the performance of a genetic information management system, but encrypted information is secure as long as the encryption key is protected properly.)

A: Yes.

B: No.

30) Please indicate whether you have ordered any genetic tests before you completed this survey

A: Yes.

B: No.

31) If your answer in question 30 is yes, please list the tests that you most frequently ordered.

Appendix B

USABILITY STUDY QUESTIONNAIRE

Guideline: This questionnaire is divided into five parts, which are corresponding to the four tasks you will perform in this study. You will be asked to perform each task and then answer the questions related to that task. The last part is for overall questions about the system.

Task 1: Create your own username and password in the system

Q1) Is it easy to set up an account in this system?

A. Yes.
B. No

Q2) If your answer is No, which part is difficult to use?

Q3) Do you consider the system secure?

A. Yes.
B. No

Q4) If your answer is No, please indicate the part you believe is not secure.

Task 2: In the "Search" page, type "Pancreatitis" in the search area and then click on the search button.

Q5) Is it easy to search genetic information related to a disease in this system?

A. Yes
B. No.

Q6) If your answer is No, please indicate the part that is hard to use.

Q7) Is the information provided in the system accurate?

A. Yes
B. No.

Q8) If your answer is No, please indicate the part you believe is not accurate.

Q9) Do you believe the search results provided in the system are easy to read and understand?

A. Yes
B. No.

Q10) If your answer is No, please suggest an alternative way of show the information.

Task 3: Click on the first record in the search result table in order to access a complete set of detailed genetic information related to the disease.

Q11) Do you believe the system provides sufficient genetic information about the disease for physicians to order genetic tests and to understand the genetic foundation of the disease?

A. Yes
B. No.

Q12) If your answer is No, please suggest another type of genetic information that needs to be provided by the system.

Q13) Do you believe the detailed genetic information provided from the search is easy to read?

A. Yes
B. No.

Q14) If your answer is No, please suggest an alternative way of show the detailed information.

Task 4: In the "Show Report" page. Select the disease "Pancreatitis" from the drop down list, and type 122 in the patient ID textbox, and then click on the "Show report" button.

Q15) Is the final genomic analysis report easy to read?

A. Yes.
B. No.

Q16) If your answer is No, please point out the specific parts of the analysis that are hard to read.

Q17) Do you believe the report is well-formatted?

A. Yes.
B. No.

Q18) If your answer is No, please point out the specific parts of the report that are not well-formatted.

Q19) Do you believe the report provided in the system is easy to understand?

A. Yes.
B. No.

Q20) If the answer is No, please point out the specific parts that are hard to understand.

Q21) Does the report provide useful guidelines about each section in the report

A. Yes.

B. No.

Q22) If your answer is No, please point out the specific parts of the report that need more guidelines.

Q23) Do you believe it is useful to include links to other important genomic databases such as dbSNP and GenBank in the report?

A. Yes.
B. No.

Q24) If your answer is No, please suggest the resources you want in the report, or indicate that you do not need or use any of these links.

Overall Questions

Q25) Overall, I am satisfied with how easy it is to use this system

A. Yes.
B. No.

Q26) It was simple to use this system.

A. Yes.
B. No.

Q27) I was able to complete the tasks and scenarios quickly using this system.

A. Yes.
B. No.

Q28) I felt comfortable using this system.

A. Yes.
B. No.

Q29) It was easy to learn to use this system.

A. Yes.
B. No.

Q30) It was easy to find the information I needed.

A. Yes.
B. No.

Q31) The information provided for the system was easy to understand.

A. Yes.
B. No.

Q32) I liked using the interface of this system.

A. Yes.
B. No.

Q33) Overall, I am satisfied with this system.

A. Yes.
B. No.

BIBLIOGRAPHY

1. Alzu'bi, A., L. Zhou, and V. Watzlaf, *Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals.* Perspect Health Inf Manag, 2014. **11**: p. 1c.
2. Snyder, M., J. Du, and M. Gerstein, *Personal genome sequencing: current approaches and challenges.* Genes Dev, 2010. **24**(5): p. 423-31.
3. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2015. **43**(Database issue): p. D30-5.
4. NCBI Resource Coordinators, *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D6-17.
5. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2005. **33**(Database issue): p. D514-7.
6. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
7. Korf, B.R., *Genetics and genomics education: the next generation.* Genet Med, 2011. **13**(3): p. 201-2.

8. Wruck, W., M. Peuker, and C.R. Regenbrecht, *Data management strategies for multinational large-scale systems biology projects.* Brief Bioinform, 2014. **15**(1): p. 65-78.

9. Zimmerman, M.D., et al., *Data management in the modern structural biology and biomedical research environment.* Methods Mol Biol, 2014. **1140**: p. 1-25.

10. Raghavan, S. and J.L. Vassy, *Do physicians think genomic medicine will be useful for patient care?* Per Med, 2014. **11**(4): p. 424–433.

11. Abrahams, E., G.S. Ginsburg, and M. Silver, *The Personalized Medicine Coalition: goals and strategies.* Am J Pharmacogenomics, 2005. **5**(6): p. 345-55.

12. Abrams, J., et al., *National Cancer Institute's Precision Medicine Initiatives for the new National Clinical Trials Network.* Am Soc Clin Oncol Educ Book, 2014: p. 71-6.

13. Wilson, B.J. and S.G. Nicholls, *The Human Genome Project, and recent advances in personalized genomics.* Risk Manag Healthc Policy, 2015. **8**: p. 9-20.

14. West, M., et al., *Embracing the complexity of genomic data for personalized medicine.* Genome Res, 2006. **16**(5): p. 559-66.

15. Cao, H., et al., *Identification of genes for complex diseases using integrated analysis of multiple types of genomic data.* PLoS One, 2012. **7**(9): p. e42755.

16. Ray, P., et al., *Bayesian joint analysis of heterogeneous genomics data.* Bioinformatics, 2014. **30**(10): p. 1370-6.

17. Lathe III, W., et al., *Genomic Data Resources: Challenges and Promises.* Nature Education, 2008. **1**(3).

18. Kawamoto, K., et al., *A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine.* BMC Med Inform Decis Mak, 2009. **9**: p. 17.

19. Kerschner, J.E., *Clinical implementation of whole genome sequencing a valuable step toward personalized care.* WMJ, 2013. **112**(5): p. 224-5.

20. Karow, J., *Children's Mercy Hospital Explores Rapid WGS for Newborn Screening, Disease Diagnosis.* Genomeweb, 2013.

21. *http://mgp.upmc.com/Applications/mgp/*. 2015.

22. Louca, S., *Personalized medicine--a tailored health care system: challenges and opportunities.* Croat Med J, 2012. **53**(3): p. 211-3.

23. Mendoza, M.C., *HIM and the path to personalized medicine.* J AHIMA, 2010. **81**(11): p. 38-42; quiz 43.

24. Hamburg, M.A. and F.S. Collins, *The path to personalized medicine.* N Engl J Med, 2010. **363**(4): p. 301-4.

25. Mardis, E.R., *Next-generation sequencing platforms.* Annu Rev Anal Chem, 2013. **6**(1): p. 287-303.

26. Zhao, Q., et al., *Tracing the transcriptomic changes in synthetic Trigenomic allohexaploids of Brassica using an RNA-Seq approach.* PLoS One, 2013. **8**(7): p. e68883.

27. Furey, T.S., *ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.* Nat Rev Genet, 2012. **13**(12): p. 840-52.

28. Pepke, S., B. Wold, and A. Mortazavi, *Computation for ChIP-seq and RNA-seq studies.* Nat Methods, 2009. **6**(11 Suppl): p. S22-32.

29. Li, N., et al., *Whole genome DNA methylation analysis based on high throughput sequencing technology*. Methods, 2010. **52**(3): p. 203-12.

30. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-90.

31. Wiechers, I.R., N.C. Perin, and R. Cook-Deegan, *The emergence of commercial genomics: analysis of the rise of a biotechnology subsector during the Human Genome Project, 1990 to 2004.* Genome Med, 2013. **5**(9): p. 83.

32. Hood, L. and L. Rowen, *The Human Genome Project: big science transforms biology and medicine.* Genome Med, 2013. **5**(9): p. 79.

33. Wadman, M., *James Watson's genome sequenced at high speed.* Nature, 2008. **452**(7189): p. 788.

34. Bonetta, L., *Whole-genome sequencing breaks the cost barrier.* Cell, 2010. **141**(6): p. 917-9.

35. Wetterstrand, K. *seq_cost 2015, https://www.genome.gov/sequencingcosts/*. 2016 3/13/2016].

36. Durmaz, A.A., et al., *Evolution of genetic techniques: past, present, and beyond.* Biomed Res Int, 2015. **2015**: p. 461524.

37. Timmerman, V., A.V. Strickland, and S. Zuchner, *Genetics of Charcot-Marie-Tooth (CMT) Disease within the Frame of the Human Genome Project Success.* Genes (Basel), 2014. **5**(1): p. 13-32.

38. Ralston, A., *Gene Interaction and Disease.* Nature Education 2008. **1**(1).

39. Chial, H., *Rare Genetic Disorders: Learning About Genetic Disease Through Gene Mapping, SNPs, and Microarray Data.* Nature Education, 2008. **1**(1).

40. Chen, S.Y., et al., *The genomic analysis of erythrocyte microRNA expression in sickle cell diseases.* PLoS One, 2008. **3**(6): p. e2360.

41. Romana, M., et al., *Thrombosis-associated gene variants in sickle cell anemia.* Thromb Haemost, 2002. **87**(2): p. 356-8.

42. Singh, J., et al., *Dental and periodontal health status of Beta thalassemia major and sickle cell anemic patients: a comparative study.* J Int Oral Health, 2013. **5**(5): p. 53-8.

43. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.

44. Tsiknakis, M., et al., *Guest editorial: Computational solutions to large-scale data management and analysis in translational and personalized medicine.* IEEE J Biomed Health Inform, 2014. **18**(3): p. 720-1.

45. Knight, J.C., *Understanding human genetic variation in the era of high-throughput sequencing.* EMBO Rep, 2010. **11**(9): p. 650-2.

46. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.

47. Ashley, E.A., et al., *Clinical evaluation incorporating a personal genome.* Lancet, 2010. **375**(9725): p. 1525–1535.

48. Shabani, M. and P. Borry, *Challenges of web-based personal genomic data sharing.* Life Sci Soc Policy, 2015. **11**(1): p. 22.

49. Claerhout, B. and G.J. DeMoor, *Privacy protection for clinical and genomic data. The use of privacy-enhancing techniques in medicine.* Int J Med Inform, 2005. **74**(2-4): p. 257-65.

50. Hayden, E.C., *Privacy protections: The genome hacker*. Nature, 2013. **497**(7448): p. 172-4.

51. Chen, X., et al., *DNACompress: fast and effective DNA sequence compression.* Bioinformatics, 2002. **18**(12): p. 1696-8.

52. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease.* Annu Rev Med, 2012. **63**: p. 35-61.

53. Christley, S., et al., *Human genomes as email attachments.* Bioinformatics, 2009. **25**(2): p. 274–275.

54. Bauch, A., et al., *openBIS: a flexible framework for managing and analyzing complex data in biology research.* BMC Bioinformatics, 2011. **12**: p. 468.

55. *http://www.ncbi.nlm.nih.gov*. 2016.

56. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2014. **42**(Database issue): p. D7-17.

57. McCarthy, J.J., H.L. McLeod, and G.S. Ginsburg, *Genomic medicine: a decade of successes, challenges, and opportunities.* Sci Transl Med, 2013. **5**(189): p. 189sr4.

58. Schadt, E.E., et al., *Computational solutions to large-scale data management and analysis.* Nat Rev Genet, 2010. **11**(9): p. 647-57.

59. Langmead, B., et al., *Searching for SNPs with cloud computing.* Genome Biol, 2009. **10**(11): p. R134.

60. Evani, U.S., et al., *Atlas2 Cloud: a framework for personal genome analysis in the cloud.* BMC Genomics, 2012. **13 Suppl 6**: p. S19.

61. *http://sourceforge.net/projects/atlas2cloud*. 2016.

62. Ginsburg, G., *Medical genomics: Gather and use genetic data in health care.* nature, 2014. **508**: p. 451–453.

63. Lindberg, D.A., *Biomedical informatics: precious scientific resource and public policy dilemma.* Trans Am Clin Climatol Assoc, 2003. **114**: p. 113-20; discussion 121.

64. Broccolo, B.M., *OCR Issues Final Modifications to the HIPAA Privacy, Security, Breach Notification and Enforcement Rules to Implement the HITECH Act.* McDermott Will & Emery, 2013.

65. Feldman, E.A., *The Genetic Information Nondiscrimination Act (GINA): public policy and medical practice in the age of personalized medicine.* J Gen Intern Med, 2012. **27**(6): p. 743-6.

66. Adida, B. and I.S. Kohane, *GenePING: secure, scalable management of personal genomic data.* BMC Genomics, 2006. **7**: p. 93.

67. Karczewski, K.J., et al., *Interpretome: a freely available, modular, and secure personal genome interpretation engine.* Pac Symp Biocomput, 2012: p. 339-50.

68. Shafer, A., *Understanding genetics*, in *The Tech*. 2006, Stanford University.

69. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

70. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

71. Bergstrom, A., et al., *A high-definition view of functional genetic variation from natural yeast genomes.* Mol Biol Evol, 2014. **31**(4): p. 872-88.

72. Salzano, F.M., *Permanence or change? The meaning of genetic variation.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5317-21.

73. Gregorius, H.R., *The meaning of genetic variation within and between subpopulations.* Theor Appl Genet, 1988. **76**(6): p. 947-51.

74. Altshuler, D., et al., *An SNP map of the human genome generated by reduced representation shotgun sequencing.* Nature, 2000. **407**(6803): p. 513-6.

75. Varela, M.A. and W. Amos, *Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence.* Genomics, 2010. **95**(3): p. 151-9.

76. *http://www.genome.gov/glossary/index.cfm?p=viewimage&id=185*. 2015.

77. Xiong, Q., et al., *Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets.* Genome Res, 2012. **22**(2): p. 386-97.

78. Talseth-Palmer, B.A. and R.J. Scott, *Genetic variation and its role in malignancy.* Int J Biomed Sci, 2011. **7**(3): p. 158-71.

79. Carlson, B., *SNPs—A Shortcut to Personalized Medicine.* Genetic Engineering & Biotechnology News, 2008. **28**(13).

80. Fareed, M. and M. Afzal, *Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service.* Egyptian Journal of Medical Human Genetics, 2013. **14**: p. 123-134.

81. Risch, N.J., *Searching for genetic determinants in the new millennium.* Nature, 2000. **405**(6788): p. 847-56.

82. Eichler, E.E., *Copy Number Variation and Human Disease.* Nature Education 2008. **1**(3).

83. *http://www.bio.miami.edu/dana/107/107F13_13.html* 2015.

84. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annu Rev Med, 2010. **61**: p. 437-55.

85. Pollex, R.L. and R.A. Hegele, *Copy number variation in the human genome and its implications for cardiovascular disease.* Circulation, 2007. **115**(24): p. 3130-8.

86. Check, E., *Human genome: patchwork people.* Nature, 2005. **437**(7062): p. 1084-6.

87. Beckmann, J.S., X. Estivill, and S.E. Antonarakis, *Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability.* Nat Rev Genet, 2007. **8**(8): p. 639-46.

88. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.

89. Girirajan, S., C.D. Campbell, and E.E. Eichler, *Human copy number variation and complex genetic disease.* Annu Rev Genet, 2011. **45**: p. 203-26.

90. Pinto, D., et al., *Functional impact of global rare copy number variation in autism spectrum disorders.* Nature, 2010. **466**(7304): p. 368-72.

91. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.

92. Gai, X., et al., *Rare structural variation of synapse and neurotransmission genes in autism.* Mol Psychiatry, 2012. **17**(4): p. 402-11.

93. Villela, D., et al., *A microdeletion in Alzheimer's disease disrupts NAMPT gene.* J Genet, 2014. **93**(2): p. 535-7.

94. Zheng, X., et al., *Genome-wide copy-number variation study of psychosis in Alzheimer's disease.* Transl Psychiatry, 2015. **5**: p. e574.

95. Cook, E.H., Jr. and S.W. Scherer, *Copy-number variations associated with neuropsychiatric conditions.* Nature, 2008. **455**(7215): p. 919-23.

96. Singh, S.M., C.A. Castellani, and R.L. O'Reilly, *Copy number variation showers in schizophrenia: an emerging hypothesis.* Mol Psychiatry, 2009. **14**(4): p. 356-8.

97. St Clair, D., *Copy number variation and schizophrenia.* Schizophr Bull, 2009. **35**(1): p. 9-12.

98. Bergeron, B., *Case Studies in Genes and Disease: A Primer for Clinicians*. 2004.

99. Gibson, W.T., *Key concepts in human genetics: understanding the complex phenotype.* Med Sport Sci, 2009. **54**: p. 1-10.

100. Clancy, S., *DNA Deletion and Duplication and the Associated Genetic Disorders.* Nature Education, 2008. **1**(1).

101. Mills, R.E., et al., *An initial map of insertion and deletion (INDEL) variation in the human genome.* Genome Res, 2006. **16**(9): p. 1182-90.

102. Collins, F.S., et al., *Construction of a general human chromosome jumping library, with application to cystic fibrosis.* Science, 1987. **235**(4792): p. 1046-9.

103. *http://www.ncbi.nlm.nih.gov/SNP/*. 2016.

104. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2007. **35**(Database issue): p. D5-12.

105. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2013. **41**(Database issue): p. D8-D20.

106. *http://www.ncbi.nlm.nih.gov/dbvar*. 2016.

107. Lappalainen, I., et al., *DbVar and DGVa: public archives for genomic structural variation.* Nucleic Acids Res, 2013. **41**(Database issue): p. D936-41.

108. *http://dgv.tcag.ca/dgv/app/home*. 2016.

109. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.

110. Capriotti, E., et al., *Bioinformatics for personal genome interpretation.* Brief Bioinform, 2012. **13**(4): p. 495-512.

111. *http://www.ncbi.nlm.nih.gov/omim*. 2016.

112. *http://www.ncbi.nlm.nih.gov/gap*. 2016.

113. Mailman, M.D., et al., *The NCBI dbGaP database of genotypes and phenotypes.* Nat Genet, 2007. **39**(10): p. 1181-6.

114. *http://www.hgmd.cf.ac.uk/ac/index.php*. 2016.

115. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.

116. *http://www.ncbi.nlm.nih.gov/clinvar/*. 2016.

117. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.

118. *http://www.snpedia.com/index.php/SNPedia*. 2016.

119. Cariaso, M. and G. Lennon, *SNPedia: a wiki supporting personal genome annotation, interpretation and analysis.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1308-12.

120. Futreal, P., L. Coin, and M. Marshall, *A census of human cancer gene.* Nat Rev Cancer, 2004. **4**: p. 177–183.

121. Church, D.M., et al., *Public data archives for genomic structural variation.* Nat Genet, 2010. **42**(10): p. 813-4.

122. Zhang, J., et al., *Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome.* Cytogenet Genome Res, 2006. **115**(3-4): p. 205-14.

123. Amberger, J., et al., *McKusick's Online Mendelian Inheritance in Man (OMIM).* Nucleic Acids Res, 2009. **37**(Database issue): p. D793-6.

124. Stenson, P.D., et al., *The Human Gene Mutation Database: 2008 update.* Genome Med, 2009. **1**(1): p. 13.

125. Sturtevant, A., *The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association.* Journal of Experimental Biology, 1913. **14**: p. 43-59.

126. Bras, J., R. Guerreiro, and J. Hardy, *Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease.* Nat Rev Neurosci, 2012. **13**(7): p. 453-64.

127. Le Morvan, V., et al., *Identification of SNPs associated with response of breast cancer patients to neoadjuvant chemotherapy in the EORTC-10994 randomized phase III trial.* Pharmacogenomics J, 2014.

128. Chial, H., *Mendelian Genetics: Patterns of Inheritance and Single-Gene Disorders.* Nature Education 2008. **1**(1).

129. Hardy, J. and A. Singleton, *Genomewide association studies and human disease.* N Engl J Med, 2009. **360**(17): p. 1759-68.

130. Craig, J., *Complex Diseases: Research and Applications.* Nature Education, 2008. **1**(1).

131. Thornton-Wells, T.A., J.H. Moore, and J.L. Haines, *Genetics, statistics and human disease: analytical retooling for complexity.* Trends Genet, 2004. **20**(12): p. 640-7.

132. Auer, P.L. and G. Lettre, *Rare variant association studies: considerations, challenges and opportunities.* Genome Med, 2015. **7**(1): p. 16.

133. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.* Science, 2005. **307**(5714): p. 1434-40.

134. Floudas, C.S., et al., *Identifying genetic interactions associated with late-onset Alzheimer's disease.* BioData Min, 2014. **7**(1): p. 35.

135. Spencer, C.C., et al., *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genet, 2009. **5**(5): p. e1000477.

136. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

137. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.

138. Kho, A.N., et al., *Practical challenges in integrating genomic data into the electronic health record.* Genet Med, 2013. **15**(10): p. 772-8.

139. Ott, J., J. Wang, and S.M. Leal, *Genetic linkage analysis in the age of whole-genome sequencing.* Nat Rev Genet, 2015. **16**(5): p. 275-84.

140. Jack, J., D. Rotroff, and A. Motsinger-Reif, *Lymphoblastoid cell lines models of drug response: successes and lessons from this pharmacogenomic model.* Curr Mol Med, 2014. **14**(7): p. 833-40.

141. Scott, S.A., *Personalizing medicine with clinical pharmacogenetics.* Genet Med, 2011. **13**(12): p. 987-95.

142. Schwab, M. and E. Schaeffeler, *Pharmacogenomics: a key component of personalized therapy.* Genome Med, 2012. **4**(11): p. 93.

143. Adams, J., *Pharmacogenomics and Personalized Medicine.* Nature Education, 2008. **1**(1).

144. Hulot, J.S., *Pharmacogenomics and personalized medicine: lost in translation?* Genome Med, 2010. **2**(2): p. 13.

145. Vogenberg, F.R., C.I. Barash, and M. Pursel, *Personalized medicine: part 3: challenges facing health care plans in implementing coverage policies for pharmacogenomic and genetic testing.* P T, 2010. **35**(12): p. 670-5.

146. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.

147. Chen, X., Z.L. Ji, and Y.Z. Chen, *TTD: Therapeutic Target Database.* Nucleic Acids Res, 2002. **30**(1): p. 412-5.

148. Zhu, F., et al., *Update of TTD: Therapeutic Target Database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D787-91.

149. Thorn, C.F., T.E. Klein, and R.B. Altman, *PharmGKB: the Pharmacogenomics Knowledge Base.* Methods Mol Biol, 2013. **1015**: p. 311-20.

150. Whirl-Carrillo, M., et al., *Pharmacogenomics knowledge for personalized medicine.* Clin Pharmacol Ther, 2012. **92**(4): p. 414-7.

151. Funk, C.S., L.E. Hunter, and K.B. Cohen, *Combining heterogenous data for prediction of disease related and pharmacogenes.* Pac Symp Biocomput, 2014: p. 328-39.

152. Zhu, Q., et al., *Exploring the pharmacogenomics knowledge base (PharmGKB) for repositioning breast cancer drugs by leveraging Web ontology language (OWL) and cheminformatics approaches.* Pac Symp Biocomput, 2014: p. 172-82.

153. Chiche, J.D., A. Cariou, and J.P. Mira, *Bench-to-bedside review: fulfilling promises of the Human Genome Project.* Crit Care, 2002. **6**(3): p. 212-5.

154. Mitha, F., et al., *SNPpy--database management for SNP data from genome wide association studies.* PLoS One, 2011. **6**(10): p. e24982.

155. Kogelnik, A., S. Navathe, and D. Wallace, *GENOME: A Networked Database Environment for Human Genome Data.* Genome Informatics 2011. **8**(1997): p. 207-214.

156. Mitha, F., *Managing large SNP datasets with SNPpy.* Methods Mol Biol, 2013. **1019**: p. 99-127.

157. Rehm, H.L., *Disease-targeted sequencing: a cornerstone in the clinic.* Nat Rev Genet, 2013. **14**(4): p. 295-300.

158. Aleman, A., et al., *A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W83-7.

159. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

160. Stenson, P.D., et al., *The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 13.

161. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic Acids Res, 2011. **39**(Database issue): p. D945-50.

162. Ephraim, S.S., et al., *Cordova: web-based management of genetic variation data.* Bioinformatics, 2014. **30**(23): p. 3438-9.

163. Orro, A., et al., *SNPLims: a data management system for genome wide association studies.* BMC Bioinformatics, 2008. **9 Suppl 2**: p. S13.

164. Pabinger, S., et al., *A survey of tools for variant analysis of next-generation genome sequencing data.* Brief Bioinform, 2014. **15**(2): p. 256-78.

165. Stitziel, N.O., A. Kiezun, and S. Sunyaev, *Computational and statistical approaches to analyzing variants identified by exome sequencing.* Genome Biol, 2011. **12**(9): p. 227.

166. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.

167. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

168. Piskol, R., G. Ramaswami, and J.B. Li, *Reliable identification of genomic variants from RNA-seq data.* Am J Hum Genet, 2013. **93**(4): p. 641-51.

169. Pierre, A. and E. Genin, *How important are rare variants in common disease?* Briefings in Functional Genomics, 2014.

170. Ionita-Laza, I., et al., *Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism.* PLoS Genet, 2014. **10**(12): p. e1004729.

171. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

172. Quintánsa, B., et al., *Medical genomics: The intricate path from genetic variant identification to clinical interpretation.* Appl. Transl. Genomic, 2014.

173. Altmann, A., et al., *vipR: variant identification in pooled DNA using R.* Bioinformatics, 2011. **27**(13): p. i77-84.

174. Bansal, V., *A statistical method for the detection of variants from next-generation resequencing of DNA pools.* Bioinformatics, 2010. **26**(12): p. i318-24.

175. McCarthy, D.J., et al., *Choice of transcripts and software has a large effect on variant annotation.* Genome Med, 2014. **6**(3): p. 26.

176. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.

177. McLaren, W., et al., *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor*. Bioinformatics, 2010. **26**(16): p. 2069-70.

178. Ge, D., et al., *SVA: software for annotating and visualizing sequenced human genomes.* Bioinformatics, 2011. **27**(14): p. 1998-2000.

179. Nielsen, C.B., et al., *Visualizing genomes: techniques and challenges.* Nat Methods, 2010. **7**(3 Suppl): p. S5-S15.

180. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

181. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

182. *https://genome.ucsc.edu*. 2016.

183. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011*. Nucleic Acids Res, 2012. **40**(Database issue): p. D918-23.

184. Gonzalez, M.A., et al., *GEnomes Management Application (GEM.app): a new software tool for large-scale collaborative genome analysis.* Hum Mutat, 2013. **34**(6): p. 842-6.

185. Gonzalez, M.A., et al., *Whole Genome Sequencing and a New Bioinformatics Platform Allow for Rapid Gene Identification in D. melanogaster EMS Screens.* Biology (Basel), 2012. **1**(3): p. 766-77.

186. Diaz-Horta, O., et al., *Whole-exome sequencing efficiently detects rare mutations in autosomal recessive nonsyndromic hearing loss.* PLoS One, 2012. **7**(11): p. e50628.

187. Montenegro, G., et al., *Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family.* Ann Neurol, 2011. **69**(3): p. 464-70.

188. Martin, E., et al., *Loss of function of glucocerebrosidase GBA2 is responsible for motor neuron defects in hereditary spastic paraplegia.* Am J Hum Genet, 2013. **92**(2): p. 238-44.

189. Yandell, M., et al., *A probabilistic disease-gene finder for personal genomes.* Genome Res, 2011. **21**(9): p. 1529-42.

190. Kennedy, B., et al., *Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data.* Curr Protoc Hum Genet, 2014. **81**: p. 6 14 1-6 14 25.

191. Reich, M., et al., *GenePattern 2.0.* Nat Genet, 2006. **38**(5): p. 500-1.

192. Kuehn, H., et al., *Using GenePattern for gene expression analysis.* Curr Protoc Bioinformatics, 2008. **Chapter 7**: p. Unit 7 12.

193. Paila, U., et al., *GEMINI: integrative exploration of genetic variation and genome annotations.* PLoS Comput Biol, 2013. **9**(7): p. e1003153.

194. Teer, J.K., et al., *VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer.* Bioinformatics, 2012. **28**(4): p. 599-600.

195. Lindhurst, M.J., et al., *A mosaic activating mutation in AKT1 associated with the Proteus syndrome.* N Engl J Med, 2011. **365**(7): p. 611-9.

196. Wei, X., et al., *Exome sequencing identifies GRIN2A as frequently mutated in melanoma.* Nat Genet, 2011. **43**(5): p. 442-6.

197. Sincan, M., et al., *VAR-MD: a tool to analyze whole exome-genome variants in small human pedigrees with mendelian inheritance.* Hum Mutat, 2012. **33**(4): p. 593-8.

198. *Understanding Human Genetic Variation*. 2007, Bethesda (MD): National Institutes of Health (US).

199. Rogers, E., *Diffusion of innovations (5 th ed.).* New York: The Free Press, 2003.

200. Calzone, K.A., et al., *National nursing workforce survey of nursing attitudes, knowledge and practice in genomics.* Per Med, 2013. **10**(7).

201. Calzone, K.A., et al., *Survey of nursing integration of genomics into nursing practice.* J Nurs Scholarsh, 2012. **44**(4): p. 428-36.

202. Parsons, M.A., et al., *A conceptual framework for managing very diverse data for complex, interdisciplinary science.* Journal of Information Science (JIS), 2011. **37**(7): p. 555–569.

203. Miyoshi, N.S., et al., *Computational framework to support integration of biomolecular and clinical data within a translational approach.* BMC Bioinformatics, 2013. **14**: p. 180.

204. Rubin, D.L., et al., *Representing genetic sequence data for pharmacogenomics: an evolutionary approach using ontological and relational models.* Bioinformatics, 2002. **18 Suppl 1**: p. S207-15.

205. Sadedin, S.P., et al., *Cpipe: a shared variant detection pipeline designed for diagnostic settings.* Genome Med, 2015. **7**(1): p. 68.

206. Haga, S.B., et al., *Developing patient-friendly genetic and genomic test reports: formats to promote patient engagement and understanding.* Genome Med, 2014. **6**(7): p. 58.

207. McLaughlin, H.M., et al., *A systematic approach to the reporting of medically relevant findings from whole genome sequencing.* BMC Med Genet, 2014. **15**: p. 134.

208. Clark, K., et al., *GenBank.* Nucleic Acids Res, 2016. **44**(D1): p. D67-72.

209. *ftp://ftp.ncbi.nlm.nih.gov/genbank*. 2016.

210. *ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/*. 2016.

211. *ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b144_GRCh38p2/*. 2016.

212. *https://www.genome.gov/26525384*. 2016.

213. *https://www.pharmgkb.org/page/downloadVariantAnnotationsHelp*. 2016.

214. Fernandez-Aleman, J.L., et al., *Security and privacy in electronic health records: a systematic literature review.* J Biomed Inform, 2013. **46**(3): p. 541-62.

215. Carrion Senor, I., J.L. Fernandez-Aleman, and A. Toval, *Are personal health records safe? A review of free web-accessible personal health record privacy policies.* J Med Internet Res, 2012. **14**(4): p. e114.

216. Kwon, J. and M.E. Johnson, *Security practices and regulatory compliance in the healthcare industry.* J Am Med Inform Assoc, 2013. **20**(1): p. 44-51.

217. Petrovic, M. *How to meet requirements of HIPAA compliance as a part of a SQL Server audit*. [cited 2015 2/7/2015]; Available from: http://solutioncenter.apexsql.com/how-to-meet-requirements-of-hipaa-compliance-as-part-of-sql-server-audit/.

218. Stallings, W., *Cryptography and Network Security: Principles and Practice* 2013: Pearson.

219. Neame, R., *Effective sharing of health records, maintaining privacy: a practical schema.* Online J Public Health Inform, 2013. **5**(2): p. 217.

220. Paar, C. and J. Pelzl, *Introduction to Public-Key Cryptography*, in *Understanding Cryptography*. 2009, Springer.

221. Hart, S.N., et al., *VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files.* Brief Bioinform, 2016. **17**(2): p. 346-51.

222. Lewis, J.R., *Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies.* International Journal of Human-Computer Interaction, 2002. **14**(3): p. 463-488.

223. *Post-Study System Usability Questionnaire (PSSUQ).* 2010 [cited 2016 1/21/2016]; Available from: http://www.conetrees.com/2010/12/ux-glossary/post-study-system-usability-questionnaire-pssuq/.

224. McNair, R., A. Taft, and K. Hegarty, *Using reflexivity to enhance in-depth interviewing skills for the clinician researcher.* BMC Med Res Methodol, 2008. **8**: p. 73.

225. Gill, P., et al., *Methods of data collectionin qualitative research:interviews and focus groups.* British Dental Journal, 2008. **204**(6): p. 291-295.

226. Wong, L.P., *Focus group discussion: a tool for health and medical research.* Singapore Med J, 2008. **49**(3): p. 256-60; quiz 261.

227. Lehoux, P., B. Poland, and G. Daudelin, *Focus group research and "the patient's view".* Soc Sci Med, 2006. **63**(8): p. 2091-104.

228. Traynor, M., *Focus group research.* Nurs Stand, 2015. **29**(37): p. 44-8.

229. Antoniou, P.E., et al., *Exploring design requirements for repurposing dental virtual patients from the web to second life: a focus group study.* J Med Internet Res, 2014. **16**(6): p. e151.

230. Leung, F.H. and R. Savithiri, *Spotlight on focus groups.* Can Fam Physician, 2009. **55**(2): p. 218-9.

231. Carroll, J., *Genetic testing: Counselors Desperately Needed.* Biotechnol Healthc, 2009. **6**(2): p. 14-22.

232. Kazma, R. and J.N. Bailey, *Population-based and family-based designs to analyze rare variants in complex diseases.* Genet Epidemiol, 2011. **35 Suppl 1**: p. S41-7.

233. Peterson, J.F., et al., *Electronic health record design and implementation for pharmacogenomics: a local perspective.* Genet Med, 2013. **15**(10): p. 833-41.

234. Murdoch, T.B. and A.S. Detsky, *The inevitable application of big data to health care.* JAMA, 2013. **309**(13): p. 1351-2.

235. Kannry, J.L. and M.S. Williams, *Integration of genomics into the electronic health record: mapping terra incognita.* Genet Med, 2013. **15**(10): p. 757-60.

236. Mitchell, D.R. and J.A. Mitchell, *Status of clinical gene sequencing data reporting and associated risks for information loss.* J Biomed Inform, 2007. **40**(1): p. 47-54.

237. Frey, L.J., L. Lenert, and G. Lopez-Campos, *EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group.* Yearb Med Inform, 2014. **9**: p. 206-11.

238. Cimino, J.J., *Improving the electronic health record--are clinicians getting what they wished for?* JAMA, 2013. **309**(10): p. 991-2.