TOOLS FOR GENETIC DATA MANAGEMENT AND STRATEGIES FOR OPTIMIZED IMPUTATION OF MISSING GENOTYPES

by

Fengshen Kuo

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

School of Health Related Professions

Rutgers, the State University of New Jersey

October 2014

RUTGERS
School of Health Related Professions

Final Dissertation Approval Form

Tools for Genetic Data Management and Strategies for Optimized Imputation of Missing Genotypes

BY

Fengshen Kuo

Dissertation Committee:

Scott R. Diehl, Ph.D., Professor, Rutgers School of Dental Medicine

Masayuki Shibata, Ph.D., Associate Professor, Rutgers, SHRP

Andrea Dynder Maes, Ph.D., Associate Director, Biostatistics, Pearl Therapeutics

Approved by the Dissertation Committee:

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

ABSTRACT

This dissertation includes two main areas of research. The first focuses on the design and development of a genetic study data management and analysis system that aims to ease the burden of dealing with the very large amounts of genetic linkage and association study data from high throughput genotyping platforms and to facilitate the integration of data from multiple sources. The Genetic Study Database (GSD) system is designed to provide security in data transmission and user management, flexibility in study data management and simplicity in user interface operations.

The second area of research focuses on the imputation of inherited genetic polymorphisms or rare variants. Since 2001, with the advent of high throughput sequencing technologies, the cost of sequencing an entire human genome has dropped from 100 million dollars to less than five thousand dollars per genome. Nevertheless, it is still too costly to obtain whole genome sequencing data for every individual in a research study involving thousands of subjects. Genotype imputation, also called in-silico genotyping, is a cost-effective and efficient way to maximize genome coverage in an association study for little or no additional cost. Depending on the type of genetic study, there are two approaches for doing genotype imputation: population-based and family-based. Both are covered in the research reported here.

The population-based approach takes advantage of publicly available genotype reference panels in predicting genotypes of unobserved variants among unrelated individuals. Here, the focus will be on optimizing the post-imputation filtering strategy to find the appropriate balance in the tradeoff between accuracy and the yield of the imputation process (i.e., maximize the number of genotypes imputed). The family-based approach leverages the rich information available in a pedigree to increase power for imputing genotypes of unobserved variants among biological relatives. When performing family-based imputation, it is important to decide how many family members and which family members to select for high density variant genotyping. Their data will be used to predict genotypes of other family members. Therefore, one aim of this part of the research will be to evaluate different family-based imputation designs to identify cost-effective strategies.

This dissertation includes three chapters: 1) designing and building a sophisticated web-based genetic study data management system, 2) identifying an optimized set of genotype/SNP filters for population-based imputation, and 3) discovering the most efficient family-based imputation strategies for various pedigree structures.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my advisor, Dr. Scott Diehl, for his patience, unfailing guidance, leadership and priceless scientific training through these years. I am very fortunate to have Dr. Diehl as my advisor. Without his valuable suggestions and generous support, this work would not have come to a successful completion.

Special thanks also go to Dr. Masayuki Shibata, my SHRP supervisor, for all his valuable comments, kind help, understanding, and support through my PhD study. I also thank my thesis committee member Dr. Andrea Dynder Maes. Her comments and advice have been helpful in accomplishing my thesis project.

I also would like to thank my designated readers, Dr. Tianxia Wu and Dr. Gokce Toruner, for their valuable time to review this thesis and sharing their constructive comments.

I am deeply grateful for the various team members and friends I have encountered during my stay at the place used to be called CPCDR. I would like to thank Dr. Ching-Yu Huang, Dr. Chih-Hung Chou, Dr. Olga Korczeniewska, and Dr. Agatha Nicolaou for their support, discussion, and friendship.

Finally, to my family, thank you all for unconditionally supporting behind me through these years. I couldn't have done it without you. This thesis is especially dedicated to all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: STUDY OVERVIEW

1.1 Challenges to genetic study management in the Post-Genome era

Despite successes in mapping disease causing genes in rare Mendelian disorders like Cystic fibrosis [1] and Huntington disease [2], in recent years profound understanding of human genome and advent in new genotyping/sequencing technology have made it possible to identify millions of informative single nucleotide polymorphisms (SNPs) capturing much of the human genome common variation across different populations and can be used to carry out genome-wide association studies (GWASs) [3,4] in dissecting common diseases. The GWAS approach has been successful in identifying SNPs that increase susceptibility to common disorders such as diabetes [5-7] and Crohn's disease [8,9]. Nevertheless, the GWAS design is based on so-called common disease/common variant hypothesis [10]. Based on this hypothesis, common genetic variant with small genetic effect is likely to be responsible for the genesis of common disorders which show heritability in the population. Moreover, since each genetic variant has small effect for common disorders, to account for all the genetic heritability there must be multiple common variations influencing disease susceptibility. To that extent, the traditional family-based genetic studies which typically have hundreds of samples and genetic markers for testing are not likely to be

successful for common disorders/traits. Therefore a shift toward population-based studies, GWAS for example, has been seen in recent years. Unlike the small sample size found in single gene or candidate genes approach used in mapping rare diseases, it is estimated at least 2000 samples are needed in order to gain 80% power in studying disease with relative risk of 1.5 through GWAS approach [11,12]. Under this circumstance, it is not feasible to rely on text editor or spreadsheet program for integrating and managing study data. Thus a daunting challenge in studying complex diseases through GWAS approach is designing a genetic study management system for efficiently managing and analyzing data.

In addition to the genotypic data, a genetic study management system should also manage study subject data including, population, family information and individual annotation information, genetic marker data including, mapping chromosomal position, gene loci ID, and other annotation information, pedigree information if available, disease or trait definitions, genetic models used in testing, phenotypic data as well as risk factor data. The genetic study management system should be able to efficiently integrate these data coming from different study sites and provide easy means to manage them for downstream statistical analyses. In the Post-Genome era, owing to the need for large sample size in GWAS it's becoming more common to have studies including researchers

around the globe. The study subject recruitment could happen in multiple sites and the genotyping processes could be done in not only one laboratory but also many laboratories and same goes to the phenotype and demographic data collection. Therefore, the need of a genetic study management system which suits for collaborative studies is more urgent than before. The ideal system should also allow multiple logon sessions and secure encrypted data transferring method among sites with ability to record audit trail information. In addition, it is important to handle the dynamic changes happened during the study life cycle. Each genetic study is evolving according to the development in the fields and laboratories. At different study developing stages, for example raw data, data cleaning, and final data stages, there might be important stage specific data existing. An ideal genetic study data management system should be able to manage and preserve the dynamic changes happened during the study life cycle.

Among different studies, there might be difference in terms of study subject and genetic marker annotation data. The annotation data is the kind of data can be used to describe the characteristics of the study subject and genetic marker like subject population or genetic marker type. In most cases, substantial annotation data types are shared among different studies. But in some cases, study specific annotation data types exist in certain study only. For example, the birth order of the study subject

and the gene locus which a genetic marker is mapped onto may be interested to one study but not for the other study. In this case, it does not make sense to dictate every study in the system having the same list of annotation data types. The ideal system should have efficient design in handling this kind of heterogeneity in study subject and genetic marker annotation data between studies and yet provide flexibility in creating any number of annotation data types according to the needs of a study. Similarly, in terms of genetic marker type, the ideal system should not be restricted to handle only certain kind of genetic marker type, for example SNP which is a bi-allele marker, and not able to handle other type of genetic marker, for example Short Tandem Repeat (STR) marker which is a multiple-allele marker. Before SNP getting popular, most of the legacy Genome Scan Linkage studies were done through genotyping STR markers. Therefore, the ideal genetic data management system should be able to handle this kind of genetic marker type heterogeneity within or between studies. Another important functionality in managing genetic study data is to be able to freely subgroup study subjects and genetic markers according to data analysis plans. The ideal system should provide easy means to select and subgroup study subjects and genetic marker for further data management. For example, a study may perform different statistical analysis method according to different ethnic background. Therefore the ideal system should allow user easily select

subjects from same ethnic group through pre-defined subject cluster based on ethnic background instead of going over through all subjects for selecting subjects. Similarly, the user should be able to easily select genetic markers which, for example, are mapped onto the same gene for further data management instead of going over many thousands of genetic marker to identify the markers in interest. One of advantage of family study is the rich information provided by the family pedigree structure. Often, showing a pedigree drawing is more straightforward and informative to the researcher than tabulated data. Therefore it's essential for a genetic study management system to provide a comprehensive means in drawing pedigree along with genotypic, phenotypic, demographic, and risk factor data to help researchers obtain understanding of the study families. For example, through reviewing the affection status of each family member in the pedigree, researchers may be able to infer the genetic model of the disease interested and implement the model in the statistical analyses. Owing to the heterogeneity in disease classification, the ideal genetic data management system should also allow multiple affection status definitions existing in the study. It should provide flexibility in defining the affection status, which may be based on the value of certain phenotype/trait or a value combination of a number of phenotypes and/or traits. Therefore, the affection status of each study subject is decided at the run time based on the current phenotype

values instead of fixed and pre-defined affection status value. One of merits of employing a genetic study data management system is being able to integrate massive amount of genotype data and phenotype data including possible risk factors which are modeled as the covariates in the statistical analyses. The ideal system should allow various type of phenotypic variable, for example numeric type, string type or categorical type, and have no restriction over the number of phenotypic variables can be created in a study. Finally, the ultimate goal of genetic study is to perform association analyses and/or linkage analyses on the integrated data. Therefore the ideal system should provide data analysis functionality and data export functionality to facilitate downstream statistical analyses. Its data export formats must support common and popular software packages in the field, for example MERLIN [13] and PLINK [14].

Although there are several software tools existing that assist researchers in managing genome-wide association studies and/or legacy genome-scan linkage studies, they tend to only cover some but not all aspects mentioned previously. Therefore an urgent need of a data management tool in the post-Genome era is a single genetic study management tool with utilities that integrate massive amount of genotype and phenotype data as well as genetic variant annotation data and perform multiple statistical analyses for dissecting disease causing genetic variants for large, collaborative genetic studies. In order to address this

need, I have developed a robust, easy-to-use data management system named Genetic Study Database (GSD) system.

1.2 Gaining more study power through genotype imputation

To infer the information of larger set of unobserved genetic variants among related individuals through genotyping a modest set of informative genetic variants has been the central theory of genetic linkage studies and of haplotype mapping approaches [15-22]. The same idea has been extended to genotype imputation which uses the stretches of shared haplotype identified among unrelated individuals to estimate the effect of many variants that are not directly genotyped. Ever since the HapMap Consortium database [23,24] was released to the public, it has been widely used in studies for GWAS microarray design and genotype imputation for samples that have ancestry close to HapMap panel populations. After the 1000 Genomes Project Phase I [25] database was released to the public, the interest of genotype imputation has grown significantly. It has become a standard practice to perform genotype imputation in a GWAS.

The identification of underlying genes for a complex disease is achieved by studying the association between the human diseases or traits and genetic variants, mainly genotypes. In human genetics, one of the central objectives is to dissect the identity and characteristics of genetic variants underlying human traits, including disease susceptibility

and variability in all kinds of biological measures. Ultimately this will be achieved by examining the possible association between trait and genetic variants discovered by thoroughly surveying the entire genome of study samples. However, given the cost of whole human genome sequencing at this time, it is still not feasible to sequence thousands of individuals. Instead, geneticists have long recognized the theory of using observed genetic variants to predict or impute the genotypes of unobserved genetic variants.

Genotype imputation refers to the use of a reference panel of haplotypes from a dense set of SNPs to impute the dense SNP genotypes of individuals whom have been genotyped at a subset of the SNPs [26-30]. The typical number of genetic markers used in a genetic linkage study to survey the entire human genome is less than 10,000 and more often the type of marker being used is short tandem repeats (STR) or microsatellites which are more informative than bi-allelic SNP markers. Due to the advances in microarray technology, genome-wide association studies have been widely conducted in the past decade. Rather than genotyping less than 10,000 markers, in order to achieve decent genome coverage, genome-wide association studies typically genotype hundreds of thousands to millions of SNP markers on each of study individual. With a reference panel of high density SNPs from multiple populations, researchers can then perform genotype imputation across the whole

genome for GWAS or over a more targeted genomic region as part of a fine-mapping study. The imputed genotypes can then be used to boost the number of genotyped variants to be examined for association with the trait in question which increases the power of GWAS or fine-mapping study.

Compared with testing only genotyped SNPs in a GWAS, it's been shown that the use of imputation can lead to a boost in power of up to 10% [11]. Moreover, simulation study has reported that the genotype imputation has greatest benefit over inferring genotypes for rare SNPS which are harder to tag [31]. On the other hand, in fine-mapping studies the number of recombination events occurring in the region and the amount of linkage disequilibrium (LD) in the region are the limiting factor that determines mapping precision. Therefore, one can increase the chance of finding a true casual variant through increasing study sample size or increasing the mapping density in regions with lower levels of LD. Genotype imputation can impute missing genotypes and genotypes of unobserved SNPs in a possibly associated region. Thus increase the chance to directly identify the casual variant in the region. In meta-analysis, it is unlikely all the cohorts used the same type of microarray. Genotype imputation can be used to increase the number of common SNPs among studies which then be combined to boost power in meta-analysis. In addition to imputing SNP genotypes, the genotype imputation can be extended to other types of genetic variants such as STR, copy

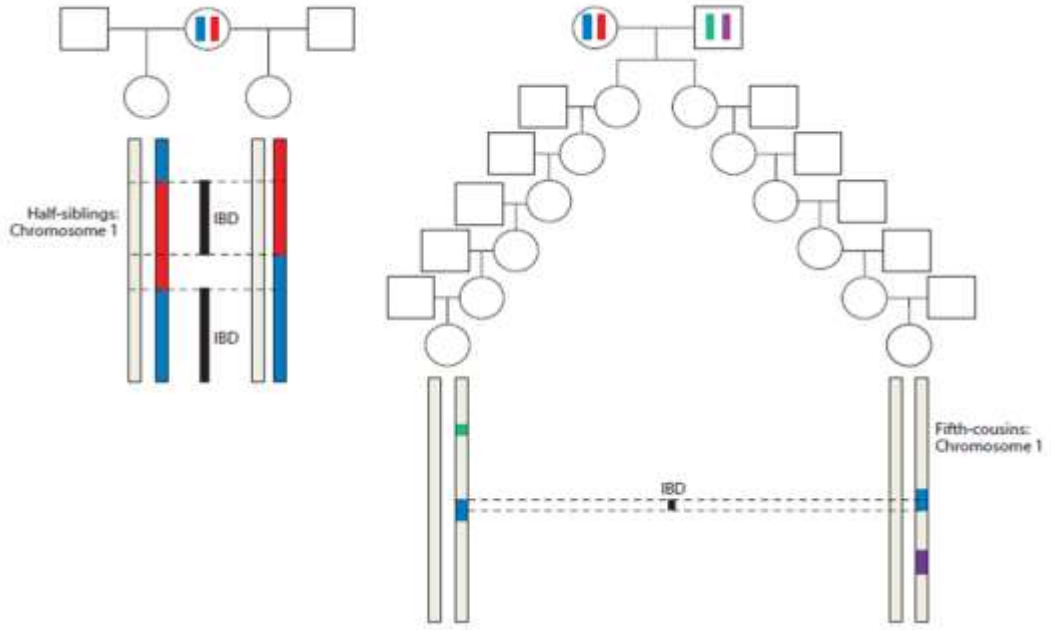
number variant and insertion/deletion variant. Finally, another use of imputation is to infer the genotype at a sporadically missing site that can happen essentially on any genotyping platforms.

In population genetics, the coalescent theory implies that all individuals have common ancestry in the distant past. When given two haplotypes share the same alleles inherited from a common ancestor, they are identical by descent (IBD). Genotype imputation is one of many applications for the IBD segments resulting from common ancestry. When detecting IBD segments, one of key considerations is the haplotype frequency. If the haplotype is found to be shared among individuals with very small frequency, the haplotype is most likely inherited from a near common ancestor. In other words, this haplotype is not likely to be seen among independently sampled individuals. Therefore, when inferring the presence of an IBD segment, one must not only consider the population genetics model but also the length of sharing, and the frequency of shared haplotype to infer probability of IBD.

Due to the smaller number of recombination happened, geneticists expect to find long stretches of shared chromosome among family members in traditional linkage and founder haplotype mapping studies. As illustrated in the family at left of Figure 1, between the two half-sib sisters, there is significant portion of chromosome 1 was inherited from a relatively recent common ancestor, the common mother in this case. On the other

hand, as illustrated by the right pedigree of Figure 1, after many meiosis and recombination happened, 2 distantly related individuals are acting like unrelated individuals who only share a very small portion of chromosome 1 from the most recent common ancestor. And it is the relatively short stretches of shared chromosome expected in GWASs among apparently unrelated individuals.

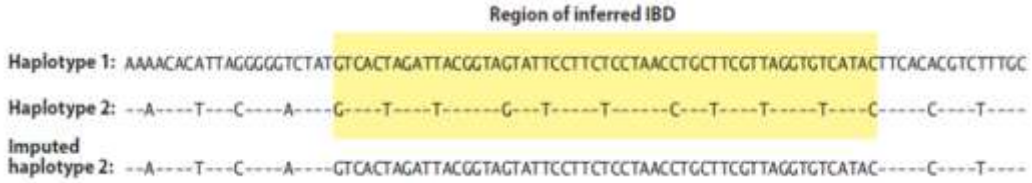
Figure 1: Closely related individuals have more and long IBD segments as compared to distantly related individuals



The principal of genotype imputation is to leverage the information carried by stretches of IBD segments to estimate the effects of variants that are not directly genotyped with great precision. The haplotype 1 in

Figure 2 is estimated from genotyping a dense genetic marker set, for example sequencing, whereas the haplotype 2 is genotyped at a limited number of selected SNPs with dashes representing unobserved alleles. Between haplotypes 1 and 2, there is an inferred IBD segment across the region highlighted by the yellow box. Since the IBD segment is thought to be identical in sequence, the unobserved alleles (dashes) of haplotype 2 can then be inferred by copying the same allele as haplotype 1 throughout the IBD segment (imputed haplotype 2). In order to predict the genotypes of unobserved variants, a reference panel with genotypes of dense variants including variants to be imputed must be available. The sample size of the reference panel could range from only a handful of family relatives owing to the high proportion of long continuous IBD or up to thousands of unrelated individuals due to low proportion of IBD which tends to be short, depending by the nature of the study design.

Figure 2: The principal behinds the genotype imputation is by leveraging the information carried by the IBD segment between Haplotype 1 and 2 (region in yellow box) for imputing unobserved alleles (dashes) of Haplotype2



The idea that related individuals share long stretches of haplotype that are IBD is the foundation of nearly all methods of linkage analysis. In 2006, Burdick and colleagues [32] first extended this idea to family-based missing genotype imputation. As shown in the Figure 1, family samples possess the most intuitive setting for genotype imputation. As mentioned earlier, an initial set of genetic markers (framework markers) are used to identify long stretches of haplotype shared among family relatives. A smaller number of family members are then selected for genotyping on a much denser set of markers (dense markers) which are used to characterize the shared IBD segment in detail. Genotypes of denser set of markers are then be inferred to family members who are typed at only the initial set of markers. This is the basis of family based genotype imputation. This method has a very high potential to increase the power of many previously conducted linkage studies by leveraging information of linkage mapping markers as the framework marker for IBD mapping and genotyping a much denser marker set, GWAS array for example, on a small number of family members. This substantially reduces the assay cost for transforming a linkage study into GWA scale family-based association study.

On the other hand, the same exact IBD sharing idea can be used for missing genotype imputation among unrelated individuals. The major difference is among unrelated individuals the shared haplotype stretches

are much shorter due to more distant common ancestors and are, therefore, harder to identify with confidence. Current tools used for genotype imputation can be classified into two categories: (i) computationally intensive tools like IMPUTE [33] and MACH [34] that consider all observed genotypes when imputing each missing genotype and (ii) computationally more efficient tools such as PLINK [14] and BEAGLE [35] that only consider genotypes from a small number of nearby markers when imputing each missing genotype. Although first category tools typically require substantially more intensive computation than the tools from second category, they actually do better at predicting missing genotype, especially for rare polymorphisms.

Although these genotype imputation tools have substantial power in imputing missing genotypes, due to the uncertainty in inferring the possibility of predicted genotype studies must have post-imputation filtering measures in place to prevent poorly imputed SNPs or genotypes from being included in downstream statistical analyses. It has been suggested to use the software reported imputation quality metrics as the post-imputation filtering measure in the community. However it's still not clear about the merit of doing post-imputation filtering solely based on this measure when different types of high density genotyping data are being used as the input for imputation, for example GWAS or Exome array. The second part of this study aims to exam the merit of using quality metric for

post-imputation filtering and identify alternative filtering measures, in the case when quality metric filtering is not adequate, through comparing the genotypes of SNP that is designed on one type of microarray but not on the other type of microarray and, yet, can be imputed by the other microarray type.

Strategies used for identifying the genetic basis of human disease have evolved considerably over the past few decades, mainly shift from family-based Linkage studies to Genome Wide Association studies (GWASs). Nevertheless, family pedigrees have been central to the discovery of genes relevant to simple Mendelian disorders, leading to the identification of nearly 4,500 such genes by the end of 2011 [36]. Despite GWASs have identified many candidate genes for common diseases, it appears now that most of common variants GWAS discovered have relative risks on the order of 1.1 to 1.2 which explains only a small fraction of heritability. Studies have suggested that most common complex diseases are likely be explained by rare variants [37-41]. Since the rare allele which is responsible for disease risk tends to aggregate and pass through family, this hypothesis once again sheds light on the use of large pedigree in a genetic study. To that extent, it also raises up a strong interest in rescuing and transforming many legacy genome-scan linkage studies into association studies. However, we are facing a number of hurdles including the expensive cost of producing dense genotypes of many subjects and

the availability of DNA sample of all family members. Family-based genotype imputation has the potential to address these hurdles. It uses the correlation of genotypes among relatives derived from sharing of genomic segments IBD within pedigrees to infer the genotypes of unobserved relatives. Although there are a number of Pedigree-based imputation methods existing in the public domain, most of them either don't handle large extended pedigree or require high-quality dense genotype data on subjects for whom we want to impute data, and do not account for recombination events. Recently a published method seems promising in performing Family-based genotype imputation on large pedigree. However, it is not clear about the merit of this method when applying on a real dataset. Most importantly, it is not clear about what is the best practical strategy in selecting limited number of family members for dense genotyping and maximizing the power of imputation. The main focus of the third part of this study is to come up with a suggestion in selecting key family members for dense genotyping through evaluating various Family-based imputation settings on a real dataset.

CHAPTER 2: GENETIC STUDY DATABASE (GSD) SYSTEM - A WEB BASED DATA MANAGEMENT SYSTEM FOR LARGE-SCALE GENETIC STUDIES

Abstract

The mapping of underlying genes for a complex disease is usually requiring a substantial number of samples and genetic markers. With current advance genotyping technology a family linkage or genome-wide association study can easily accumulate millions of genotypes. In addition, as the study data is changing dynamically from the beginning to the end of a study, the flexibility of updating and manipulating genetic data with traceability is desired. Here we describe a secure web-based genetic study database (GSD) system for high throughput population and family based genetic study. GSD is a platform independent web-based DBMS system with supporting of HTTP protocol over an encrypted Secure Sockets Layer (SSL) or Transport Layer Security (TLS) transport mechanism. Together with the comprehensive user account management and study data access control GSD is designed to accommodate the requirements of IRB proved large-scale and/or multi-sites collaborating genetic linkage or association studies. The underlying database of GSD is an Oracle relational database which offers excellent management of lager

data set, exceptional flexibility in complex data table and query designs, and effective data quality controls. The relational database primary key and foreign key relationship assures the GSD study data integrity and the table indexes and stored procedures provide efficient access to data with complex structures. The front end user interface is powered by an Apache web server with SSL/TLS encryption. Any modern web browser with encryption capability can be a client. This architecture allows efficient management and manipulation of large dataset via a user-friendly graphical interface. GSD can handle unlimited number of study existing in the system. Furthermore, each study can include unlimited number of study subject, family, marker, variable, phenotype and genotype. The data import can be readily done through copy/paste or uploading the data file through the interface. Finally, its data export functionality supports various formats for downstream analyses (e.g, LINKAGE, GENEHUNTER, MERLIN and PLINK).

2.1 Introduction

In recent years, the focus of human genetic study has shifted from Linkage analysis and/or candidate gene analysis to Genome Wide Association analysis [3,4,42]. The advent of high throughput array genotyping system has moved the study management from spreadsheet into more sophisticated relational database and user interface design. Due to the massive amount of data generated by the genotyping system, it's become unbearable to work with a spreadsheet for genetic study management when a study is going through different phases. The same burden can be seen when sharing massive amount of data between collaborators. In addition, the difficulty in handling complex and massive genetic data also jeopardizes the ability of keeping up with the data audit trail.

In contrast to single-gene disease (Mendelian trait such as Huntington's disease and Haemophilia A and B), the identification of underlying genes for a complex disease (such as hypertension, diabetes, and cancer) usually requires a substantial number of samples and genetic markers [43,44]. In addition, there are some other factors that can reduce the power of detecting the disease-causing genes including genetic heterogeneity, gene-gene interaction, gene-environment interaction, partial penetrance, phenocopies and late-onset disease. Therefore in order to obtain sufficient power in detecting underlying genes for a complex trait in a genome scan, high throughput genotyping must be done

on a larger DNA sample collection against thousands or millions of genetic markers.

One of popular applications for detecting the linkage and association between the disease and genetic marker is to use single nucleotide polymorphisms (SNPs). Unlike simple tandem repeat (STR) marker, SNP has a lower number of alleles, 2 alleles for most of time, to transmit through generations. Nevertheless, the detecting power can be maintained by increasing the number of fine-mapping SNPs in genotype process. With the advent of fast and cost-effective genotyping technology one can easily generate millions of genotypes from a microarray system, for example Affymetrix GWAS array and Illumina HumanOmni2.5 array. Therefore it has become a major challenge to handle the high-throughput laboratory genotyping data and integrate with larger volumes of clinical data for downstream genetic analyses.

Another aspect is that the genetic study data are dynamically changing over the study life cycle. For example, previous unavailable DNA samples may become available after the study began or the clinical and demographic data may change after lengthy follow-up. Moreover, a follow-up marker set may be added into the study after preliminary data are available from analyzing the initial marker set. Finally any raw genotype is subject to be invalidated after discovering a Mendelian discrepancy within family according to pedigree information or laboratory genotyping error.

Therefore it is essential to have a genetic data management system (GDMS) offering great flexibility in manipulating data changes and yet maintaining audit trail information. To complicate the situation more, considers the scale of current complex trait study it often involves multiple institutes across the globe. As a requirement of regulations imposed on any research study involving human subjects, every institute participating in the study must be approved by its Institute Review Board (IRB) prior the beginning of the study. An ideal GDMS would provide cross-platform access with sophisticated data encryption and comprehensive study access control to protect sensitive data from leaking in order to compliant with IRB regulations. In addition, simple data import mechanism and the capability of exporting complied data into various formats for downstream statistical analysis software are also crucial in facilitating study progress.

Owing to the complexity of disease etiology (for example, genetic heterogeneity, incomplete penetrance, phenocopies, age of disease onset, environmental factors, gene-environment interaction, and gene-gene interaction) the mapping of underlying genes for a complex disease is requires dissecting a substantial number of study samples and genetic markers. With current advance genotyping technology, a family linkage or genome-wide association (GWA) study can easily accumulate millions of genotypes from examining thousands samples against hundreds of thousands of genetic marker, mostly SNPs. The significant amounts of

data generated during these genome surveys pose great data management challenges in manipulating, querying, comparing, integrating and visualizing study data. These challenges come from not only the size of data generated but also the integration of data coming from different genotyping platforms, the integration between genotype and phenotype data, the integration with statistical analysis pipeline, study subject and genetic marker annotation, and family, if any, structure visualization. In addition, as the study data is changing dynamically from the beginning to the end of a study, the flexibility of updating and manipulating genetic data with traceability is desired.

Although a number of software tools have been developed and made available to assist researchers in conducting genetic association studies, they tend to focus on some specific aspects only. For example, GenoDB [45], GeneLink [46], T.I.M.S [47], SNPLims [48], SNPP [49] and OpenADAM [50] are data management systems designed to facilitate the storage and management of large volumes of genotype data generated by candidate gene study or GWAS approaches. They all lack the pedigree drawing functionality and some of them have no phenotype integration functionality or only handling limited number of phenotypes. Some of them only manage bi-allele genetic marker, mainly SNP, and cannot handle multiple-allele genetic markers. Some of them only accept association study data and do not handle linkage study data. Some of them do not handle subject

or genetic marker annotation data or only handle specific type of annotation data. On the other hand, systems such as PhD [51] and the Mouse Phenome Database [52] have been developed for managing large amount of phenotypic data but they lack the functionality to integrate with genotypic data. Therefore, the needs that are not all addressed by the existing tools include functionalities that facilitate multiple sites collaboration, simplify the data integration among various genotype platforms, integrate phenotype and genotype, annotate study subject and marker, visualize extended pedigree structure and audit data changes made through the study life cycle.

Here I propose a robust internet-based genetic study data management system, GSD, which is designed for handling multiple simultaneous studies with features mentioned above. GSD aims to address the special requirements of multiple on-going families or case-control based studies as well as the issues of multiple users. It facilitates the data management process in the post-genotyping phase of a study.

In terms of data security and user data access control, in addition to password protection this internet-based application also employs a 128-bit encrypted communication protocol between client and server. GSD imposes two types of access control mechanisms, data management control and study access control. Under the first control mechanism, a user is allowed to perform data manipulation work only with appropriate

privilege granted. Currently 4 types of privilege levels are available to assign to a user, Admin, Laboratory Manager, Laboratory User and Demonstration User. With study access control, a user must be authorized before they have access to a study.

GSD uses the concept of attribute to accommodate the database design problem arising from handling the heterogeneity among data entry. Within a study, a data entry, for example study subject, marker or phenotype, is treated as an object with unlimited number of attributes which are describing the object specific characters, for example the genetic mapping position of a marker. It does not require that all the objects in the same object type must have the same number of attributes.

During the study life cycle same type of data entry may be grouped for data review, exporting for analysis or data updating, etc. In the case of complex trait study, it is a tedious and time-consuming process of selecting certain data entries from thousands, if not millions, of data entries. GSD allows creation of a data entry cluster with an unlimited number of data entries and a study can have as many data entry clusters as needed. GSD offers 4 types of data management in performing data update: New, Update, Inactivate, and Reactivate. All can be done through a user-friendly interface in batch or interactive mode with data integrity checking.

Another measure in maintaining data integrity, GSD provides audit trail functionality which records any data update history along with comments given by the user. In terms of data exporting, GSD currently is supporting 5 export formats for a variety of analysis programs, including LINKAGE [53], SUPERLINK (post-Makeped) [54], MERLIN [13], RELPAIR [55], and PLINK [14]. Export is done in a chromosome-by-chromosome fashion and, depending on selected format, various numbers of data files is generated for each chromosome. In addition to the individual data files, GSD export interface is also providing a compressed zip file containing all data files for easy-download.

When dealing with extended pedigree, a well-illustrated pedigree drawing can facilitate the process of reviewing pedigree structure along with phenotype and genotype data. GSD is using a publicly available pedigree drawing program, CraneFoot [56], that provides extremely flexible pedigree drawing capability. It allows pedigree drawing with, technically, unlimited number of pedigree members along with unlimited number of genotypes and phenotypes.

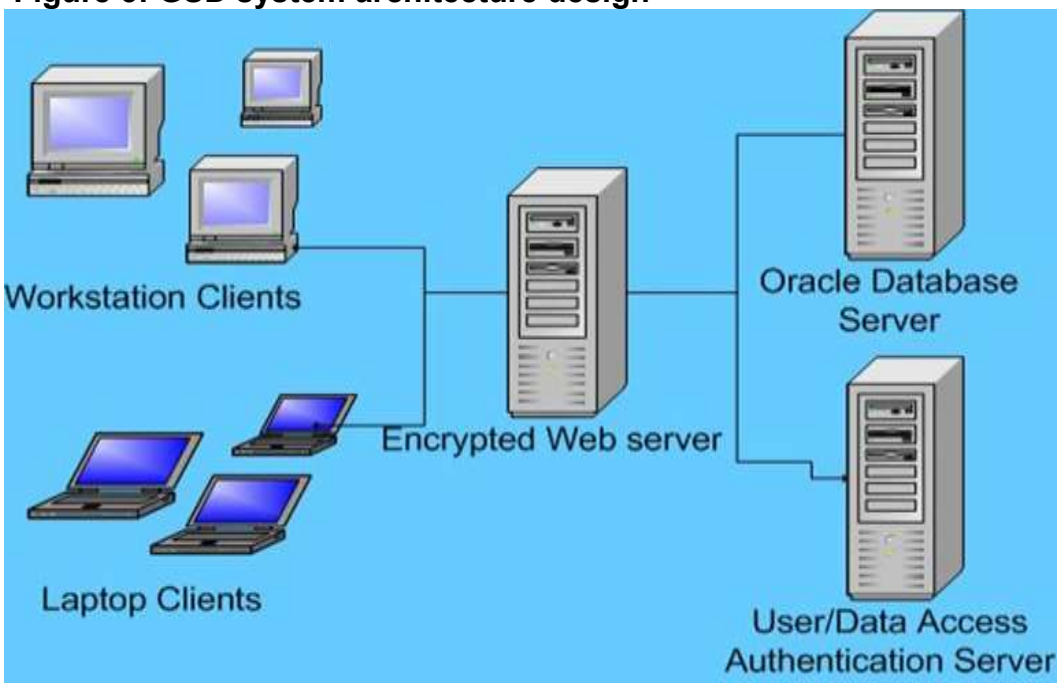
Finally, in terms of association, GSD is providing basic population basis and family adjusted association tests. With the disease model, genetic markers and study subjects selected, one can easily run the family adjusted Cochran–Armitage test, Chi-Square test or Fisher-Exact test by one click without tedious and complicated format transformation.

2.2 System design

The current software architecture design of GSD is built on top of three servers: a web server, a relational database server for genetic data management, and another relational database for user authentication (Figure 3). The basic software requirements for machine hosting the web server include CGI script support with Perl version 5.6.1 or greater and SSL/TLS 128-bit encryption. In terms of Perl modules, GSD requires CGI, Carp, DBI, DBD::Oracle and DBD::MYSQL as well as some other in-house modules. Currently an Apache web server is installed on a Linux RedHat 9.0 machine for our in-house system, although GSD is not tied to a specific type of operating system. Because of the enormous amount of data coming from a high throughput genotyping process in a complex trait study, a modern relational database is chosen to take the advantages of data storage and complex data querying capability. GSD is built on an Oracle 11g relational database as the underlying main database with sophisticated database design and optimization for managing genetic studies. In addition, a second relational database system, MySQL 3.23.54, is used to reinforce data security measures including user authentication, data manipulation function control, and study access control. Regarding pedigree drawing, GSD is employing a publicly available program, CraneFoot v3.beta [56], as the underlying drawing engine. Another public

program, MakePed [53], is required for outputting pedigree data in post-MakePed format during the data exporting process.

Figure 3: GSD system architecture design



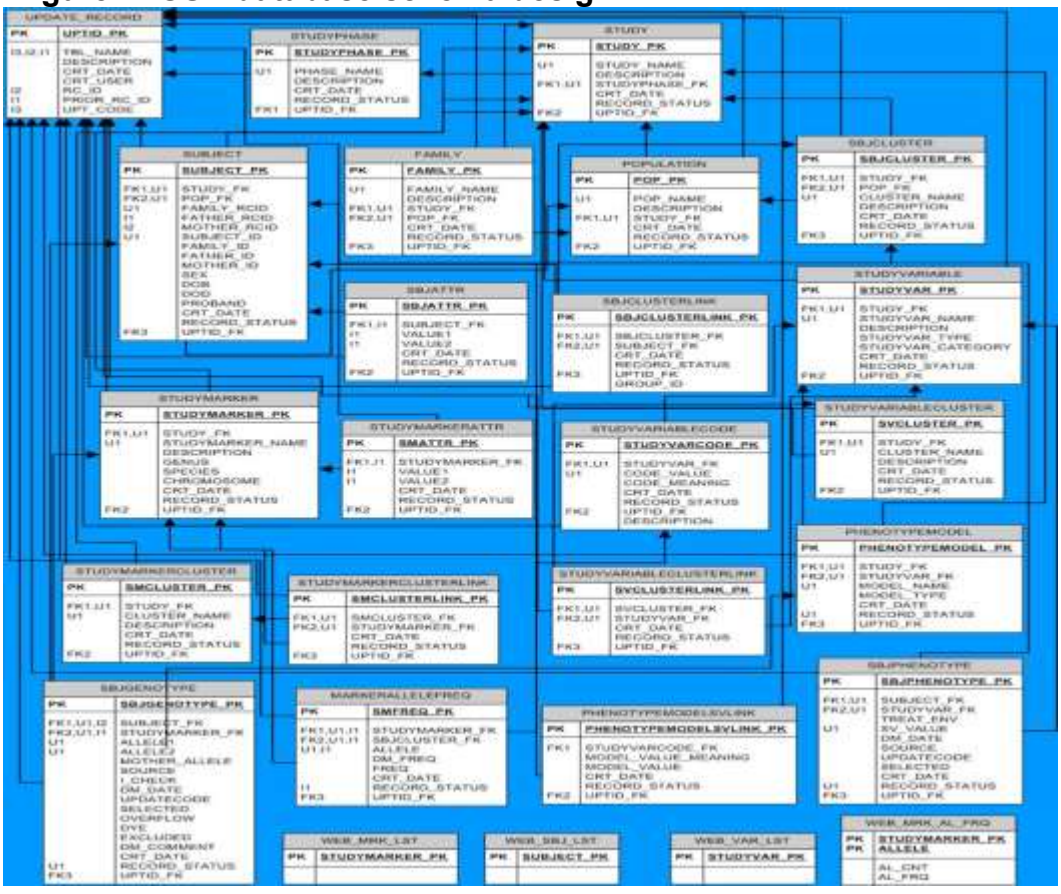
2.3 Database design

2.3.1 Overall database design and audit trail

Currently there are 26 tables in GSD database schema design (Figure 4) and all of them, except Update_Record and 4 session-specific temporary tables, have a foreign key relationship constraint to Update_Record table, which holds the key to audit trail information. This

relationship allows the audit trail information (including who, when, why/what, before-change and after-change data) to be tracked in the database. Each record of Update_Record table is pointing to two data records, representing before-change and after-change record. Therefore, one can traverse through all the changes that happened in the past from the currently active data record to the creation point of the data record and vice versa. A two-arrow line between two tables represents this data tracking functionality. And a single-arrow line is indicating a foreign key relationship from the start table to the arrow-pointed table.

Figure 4: GSD database schema design



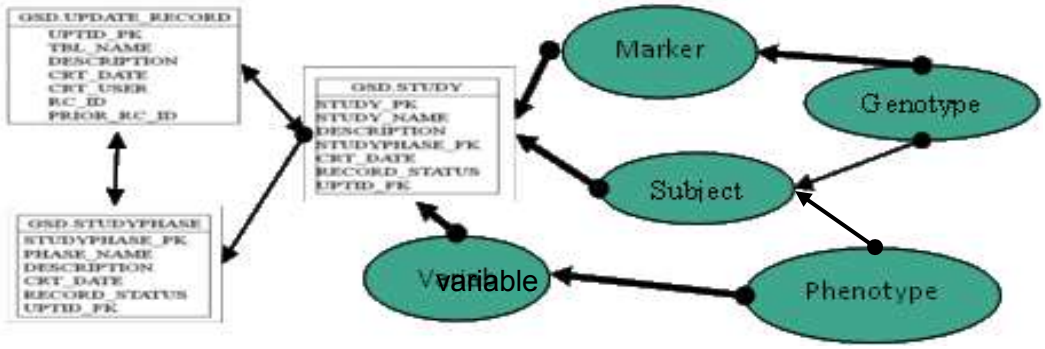
In order to insure the data integrity and speed up data records access across database tables by query joining, all the tables have primary key(s) defined and indexed. Beside the primary key(s), the vast majority of tables have foreign key(s) defined to hold the relationship between related tables. For example, the SBJCLUSTERLINK table has UPTID_FK, SBJCLUSTER_FK, SUBJECT_FK defined as its foreign keys (Table 1) which are used to maintain the relationship with UPDATE_RECORD, SBJCLUSTER, and SUBJECT tables respectively, avoid having redundant data, and insure the data integrity among these tables.

Table 1: GSD database table description and primary key and foreign key information

Table Name	Description	Primary Key(s)	Foregin Key(s)
FAMILY	Family information	FAMILY_PK	UPTID_FK, POP_FK, STUDY_FK
MARKERALLELEFREQ	Marker allele frequency data	SMFREQ_PK	UPTID_FK, STUDYMARKER_FK, SBJCLUSTER_FK
PHENOTYPEMODEL	Affection status information	PHENOTYPEMODEL_PK	UPTID_FK, STUDY_FK, STUDYVAR_FK
PHENOTYPEMODELSVLINK	Link relationship between Affection status and study variable	PHENOTYPEMODELSVLINK_PK	UPTID_FK, STUDYVARCODE_FK, PHENOTYPEMODEL_FK
POPULATION	Population information	POP_PK	UPTID_FK, STUDY_FK
SBJATTR	Subject attribute name and value	SBJATTR_PK	UPTID_FK, SUBJECT_FK
SBJCLUSTER	Subject cluster information	SBJCLUSTER_PK	UPTID_FK, POP_FK, STUDY_FK
SBJCLUSTERLINK	Link relationship between subject and subject cluster	SBJCLUSTERLINK_PK	UPTID_FK, SBJCLUSTER_FK, SUBJECT_FK
SBJGENOTYPE	Subject genotype	SBJGENOTYPE_PK	UPTID_FK, SUBJECT_FK, STUDYMARKER_FK
SBJPHENOTYPE	Subject phenotype	SBJPHENOTYPE_PK	UPTID_FK, STUDYVAR_FK, STUDYVAR_FK
STUDY	Study information	STUDY_PK	UPTID_FK, STUDYPHASE_FK
STUDYMARKER	Genetic marker information	STUDYMARKER_PK	UPTID_FK, STUDY_FK
STUDYMARKERATTR	Genetic marker attribute name and value	SMATTR_PK	UPTID_FK, STUDYMARKER_FK
STUDYMARKERCLUSTER	Genetic marker cluster information	SMCLUSTER_PK	UPTID_FK, STUDY_FK
STUDYMARKERCLUSTERLINK	Link relationship between marker and marker cluster	SMCLUSTERLINK_PK	UPTID_FK, STUDYMARKER_FK, SMCLUSTER_FK
STUDYPHASE	Study phase information	STUDYPHASE_PK	UPTID_FK
STUDYVARIABLE	Variable information	STUDYVAR_PK	UPTID_FK, STUDY_FK
STUDYVARIABLECLUSTER	Variable cluster information	SVCLUSTER_PK	UPTID_FK, STUDY_FK
STUDYVARIABLECLUSTERLINK	Link relationship between variable and variable cluster	SVCLUSTERLINK_PK	UPTID_FK, STUDYVAR_FK, SVCLUSTER_FK
STUDYVARIABLECODE	Variable data type information	STUDYVARCODE_PK	UPTID_FK, STUDYVAR_FK
SUBJECT	Subject information	SUBJECT_PK	UPTID_FK, STUDY_FK, POP_FK
UPDATE_RECORD	Audit trail information	UPTID_PK	N/A
WEB_MRK_AL_FRQ	Temporary session table for query optimization	STUDYMARKER_PK, ALLELE	N/A
WEB_MRK_LST	Temporary session table for query optimization	STUDYMARKER_PK	N/A
WEB_SBJ_LST	Temporary session table for query optimization	SUBJECT_PK	N/A
WEB_VAR_LST	Temporary session table for query optimization	STUDYVAR_PK	N/A

The overall view of GSD database design is illustrated by Figure 5. The GSD database design starts with a root entity, StudyPhase, which can be associated with as many child entities, Study, as possible by foreign key constraint. Therefore, each Study must be associated with one StudyPhase only. A Study under GSD conceptually includes five data entity groups; Subject, Study Marker, Study Variable, Phenotype and Genotype. Each entity group is composed of a number of tables which will be covered in detail later. Although StudyPhase is suggested to be used for managing a study in different data phases, for example raw data, clean data, imputed data and final data set, it is not the only way of implementing StudyPhase and Study design.

Figure 5: Over view of GSD database design

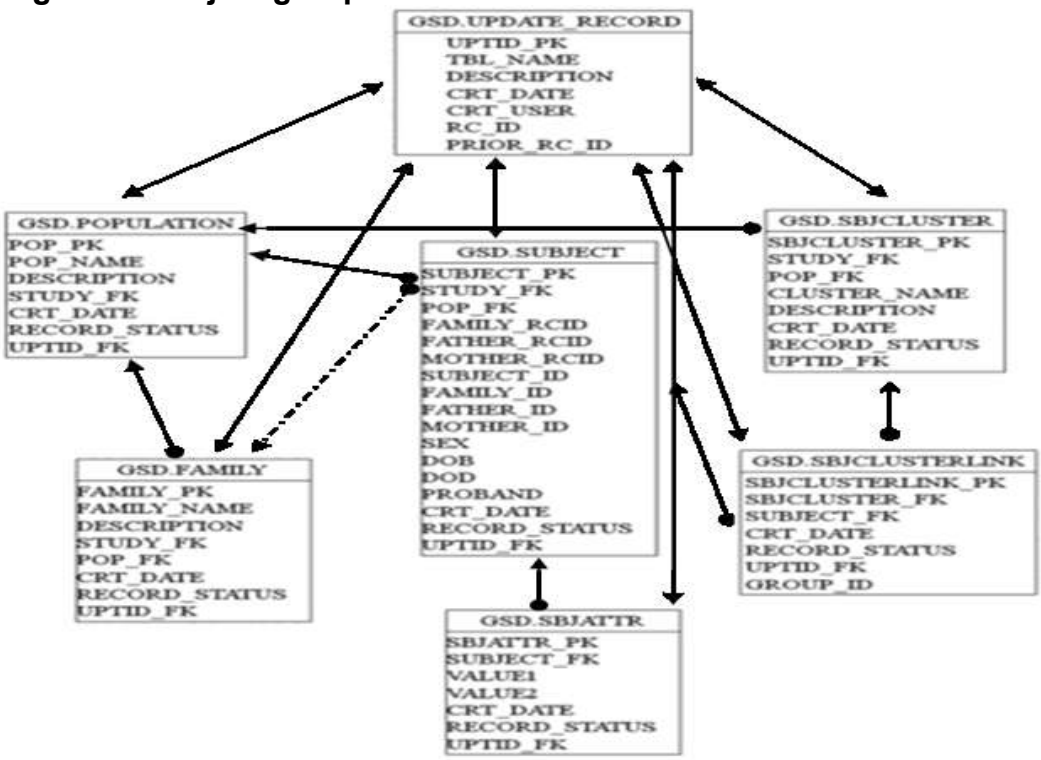


2.3.2 Subject group

The Subject Group (Figure 6) includes 6 tables with constraints among them. A Study can have as many Populations associated with it as

necessary. A Population is composed of a number of Subjects. Under a Population, Subjects can be grouped into a Family (family-based subject) or no Family (case-control subject). The arrowed dotted line between the Subject table and the Family table represents that there is a “soft” foreign key relationship between two tables allowing a Subject to be associated with no Family designed for case-control subject. In addition, any Subjects can be grouped into a SubjectCluster as a sub-group and a Subject is allowed to participate in multiple SubjectClusters.

Figure 6: Subject group database schema



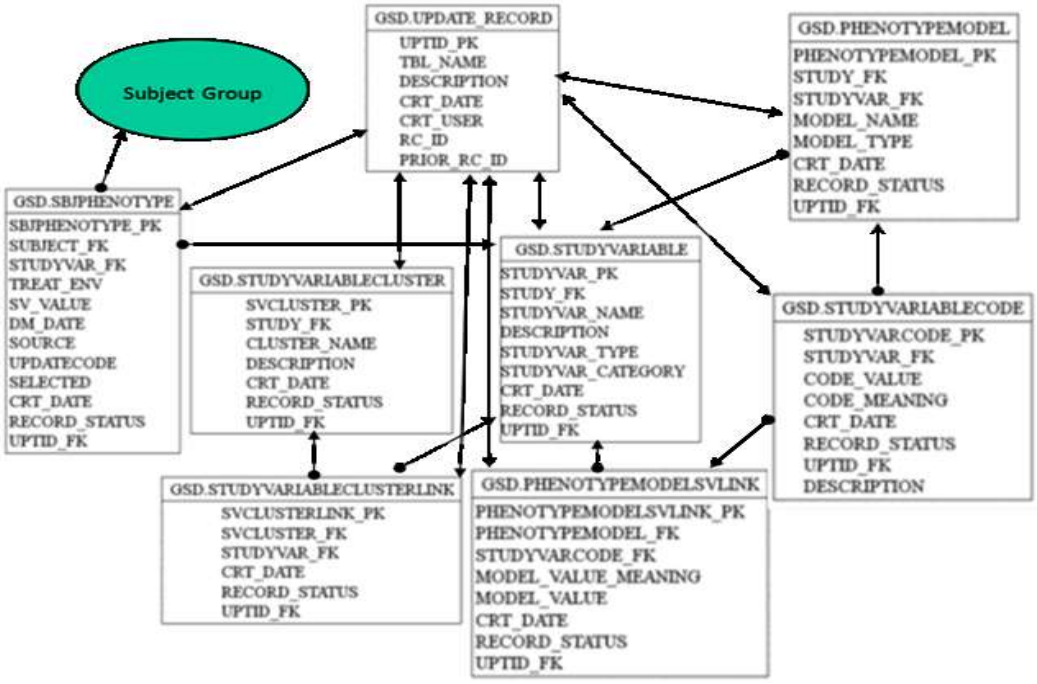
For example, a female subject with European ancestry can be defined in both Caucasian cluster and Female cluster. The SubjectClusterLink table is the key table in establishing this Subject-SubjectCluster relationship. Each record in Subject table has links pointing to the father and mother records in the table. Also each Subject can have as many attributes as possible in the SbjAttr table. As mentioned early each attribute can be used for describing an entity specific character and is composed by attribute name (value1) and attribute value (value2).

2.3.3 Phenotype group

The Phenotype group covers 7 tables (Figure 7) and is designed to handle data including subject demographic data, binary and quantitative phenotypes as well as disease affection status model. A study variable defines a study interest of Subject, which could be a physical exam measurement or disease diagnosis, and there is no limitation on the number of study variables under a Study. Like Subject, Study variables can be grouped together as a study variable cluster and a study variable can be a member of multiple study variable clusters. By design, a code-type variable can have many code values to represent different meaning through the StudyVariableCode table. The SbjPhenotype table holds the phenotype data for a Subject by keeping a foreign key relationship to the Subject and the StudyVariable tables respectively. In dealing with subject

affection status definition, GSD uses Phenotypemodel and Phenotypemodelsvlink tables to handle multiple disease definitions. Some diseases, especially those based on complex traits, may have multiple disease definitions, for example mild form and severe form, due to the complex interaction among underlying disease genes. Therefore it is crucial to be able to efficiently export data with different affection status for exploratory data analyses especially in exploring phase. Each record in Phenotypemodel table is referencing a record in Studyvariable table and each record of Studyvariablecode has a corresponding record in Phenotypemodelsvlink table.

Figure 7: Phenotype group database schema

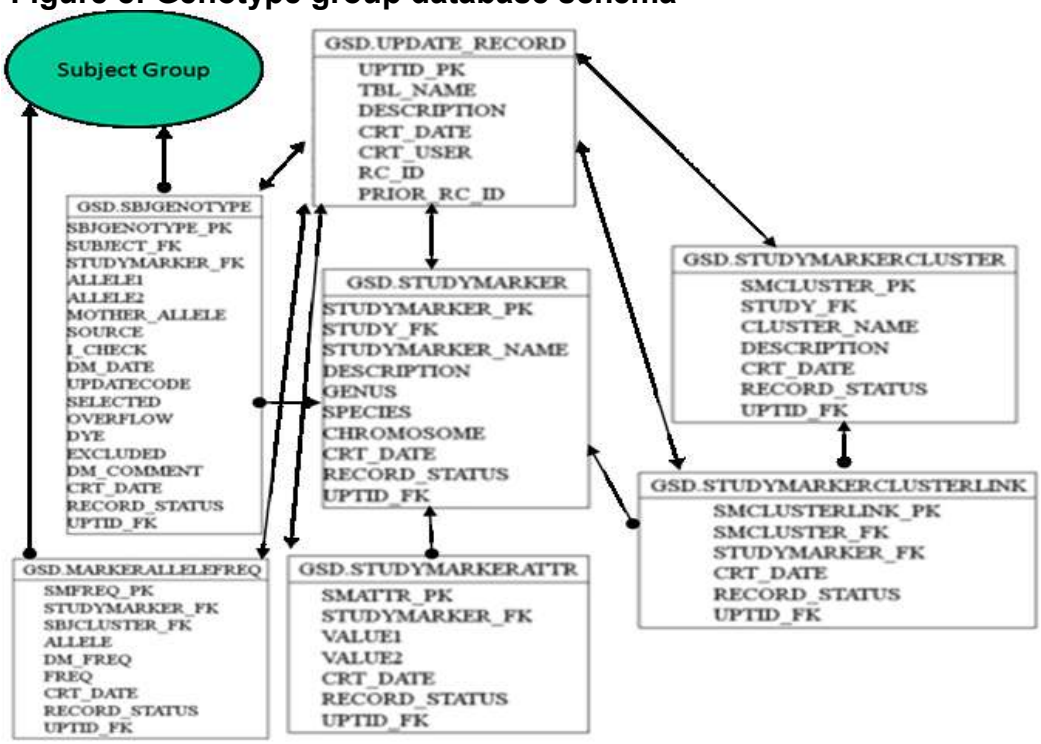


For instance, a study code-type variable may have possible 4 values, Severe, Mild, Healthy and Missing, indicating the subject disease classification. A subject with Mild form disease may be classified as Affected in a broad disease model and, on the other hand, classified as Unknown in a stringent disease model.

2.3.4 Genotype group

This group (Figure 8) covers data regarding Marker, Marker attribute, Marker cluster and Marker genotype. By design, a Study can include as many study markers as possible in the StudyMarker table. Study markers can be grouped together as a study marker cluster and a study marker is allowed to be a member of multiple study marker clusters. Records in StudyMarkerClusterLink table establish this relationship. In addition a marker can have multiple attributes, name alias or map positions for example, associated with it by having records in StudyMarkerAttr table. The allele frequency data of a marker calculated from a SubjectCluster is stored in the MarkerAlleleFreq table. Therefore, the MarkerAlleleFreq table has foreign key relationship between it and SbjCluster and StudyMarker tables. The Subject genotype data are kept in SbjGenotype table. Although a Subject-Marker pair can have as many genotype records as possible in SbjGenotype table, only Subject-Marker with one genotype record may be exported during genotype export.

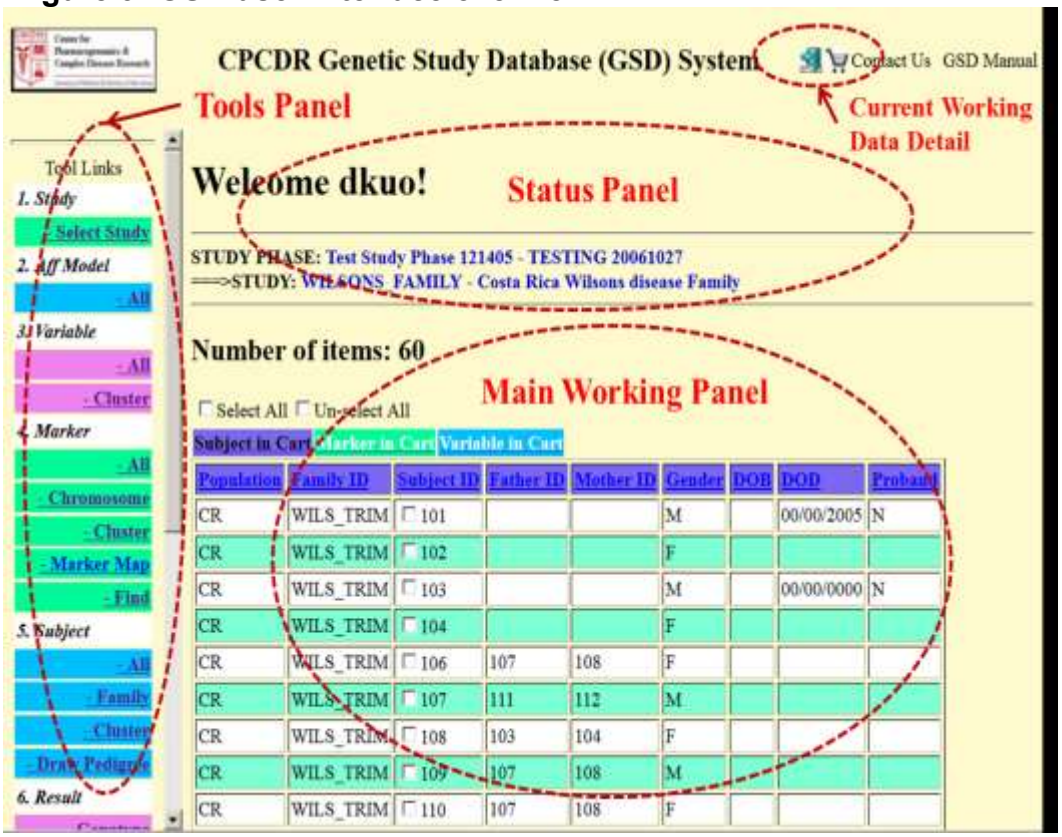
Figure 8: Genotype group database schema



2.4 User interface design

Upon successful user login, the GSD interface (Figure 9) is composed of three interface panels, including Status Panel, Tools Panel, and Main Working Panel, with main option menu showed at bottom left panel. Below sessions discuss in much detail about the design behind each panel.

Figure 9: GSD user interface overview



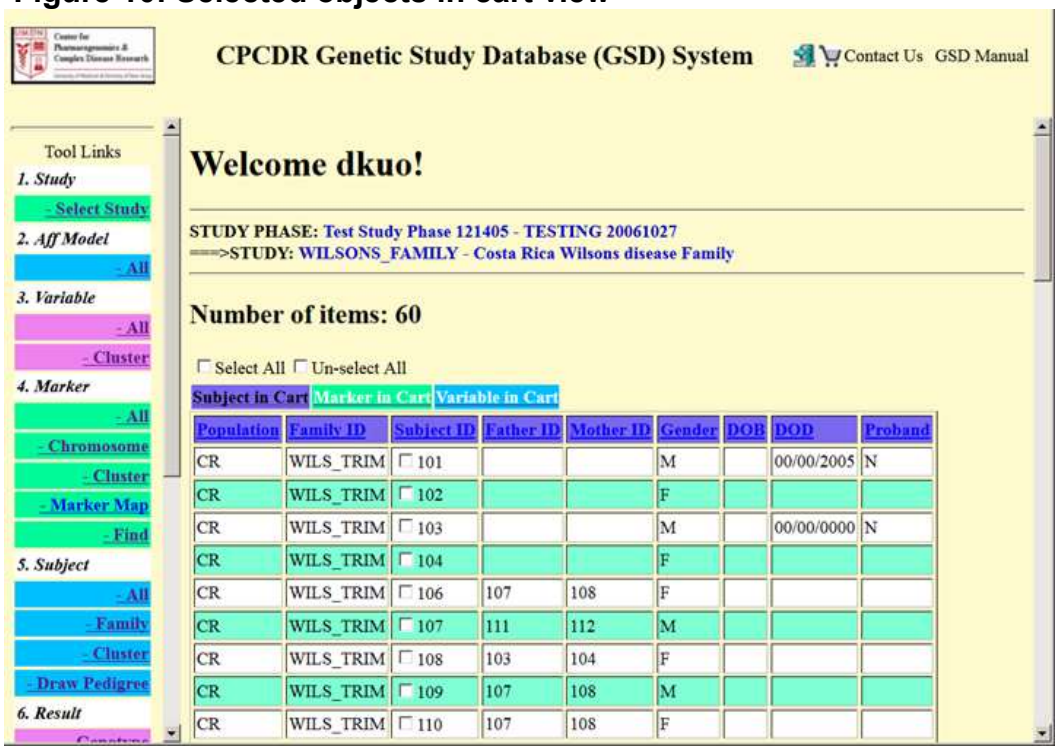
2.4.1 Main working panel

This panel is the main working panel in GSD. Through this panel GSD displays the action response or message according to user inputs. For example, the figure # shows that GSD is prompting user for selecting Study Phase through drop-down list in order to proceed for Study selection.

2.4.2 Status panel

This panel at top provides the options for user logout and shopping cart review as well as links to the center web site, online user manual, and sending message to GSD administrator. As one of security measures, GSD will invalidate the login session after the user is idle for a pre-defined amount of time; nevertheless, it is still a wise practice to logout unwanted access session through the logout button. The shopping cart button right next to the logout button gives the access to a list of current selected study subjects, markers and variables (Figure 10).

Figure 10: Selected objects in cart view



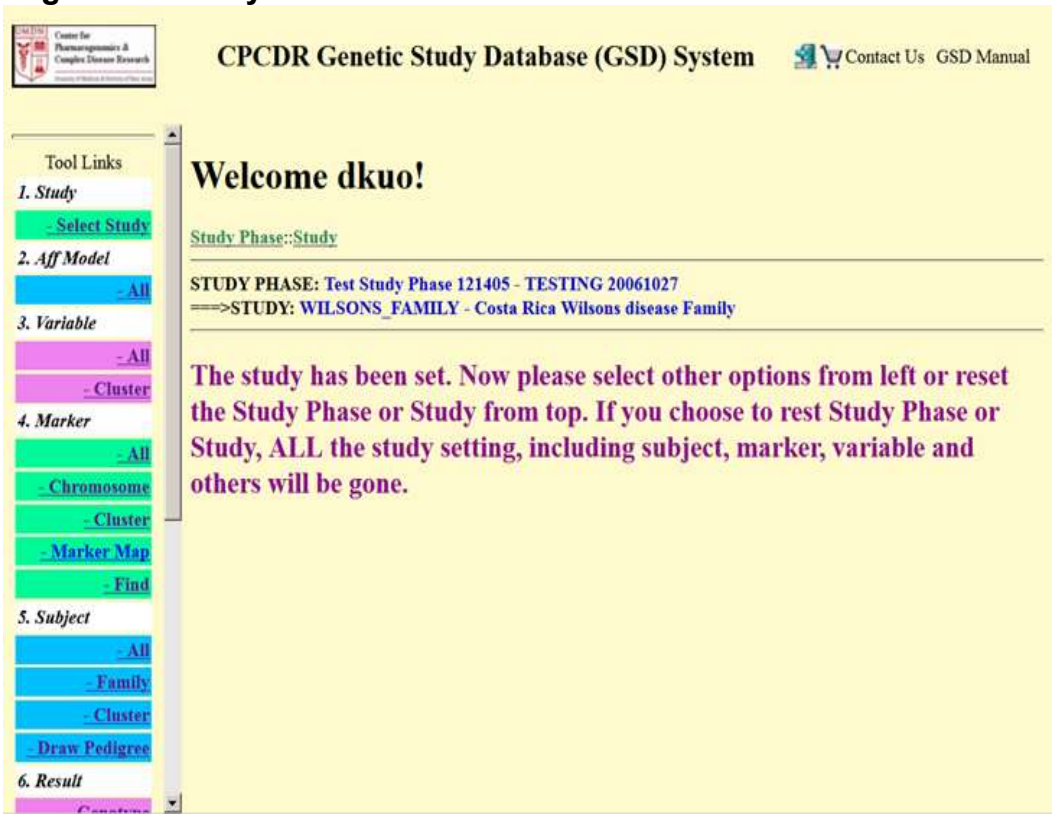
One can review the details of selected objects in the current login session or drop any undesired objects from the cart before proceeding. In addition, it also allows user to create a data record export for selected objects, which then can be used for data mining or importing the same set of objects into another study in GSD.

2.4.3 Tools panel

The panel at left provides the main option menu of all functionalities during genetic data management in nine categories, study phase/study selection, affection model selection, variable selection, subject selection, marker selection, genotype and phenotype result display, data export, system operation, and statistical analysis.

Prior to performing any data management steps in GSD a study must be selected first. Study selection is a two-step process, selecting the study phase first and then selecting the desired study under the selected study phase. Without setting desired study in the first place, GSD will not allow any functionality to be available to the user since it is not clear which study to retrieve data from. Therefore the very first step in GSD is selecting a study after a study phase is selected before performing any data management. Once a study is set, GSD will confirm and acknowledge the selected study phase and study in the working panel (Figure 11).

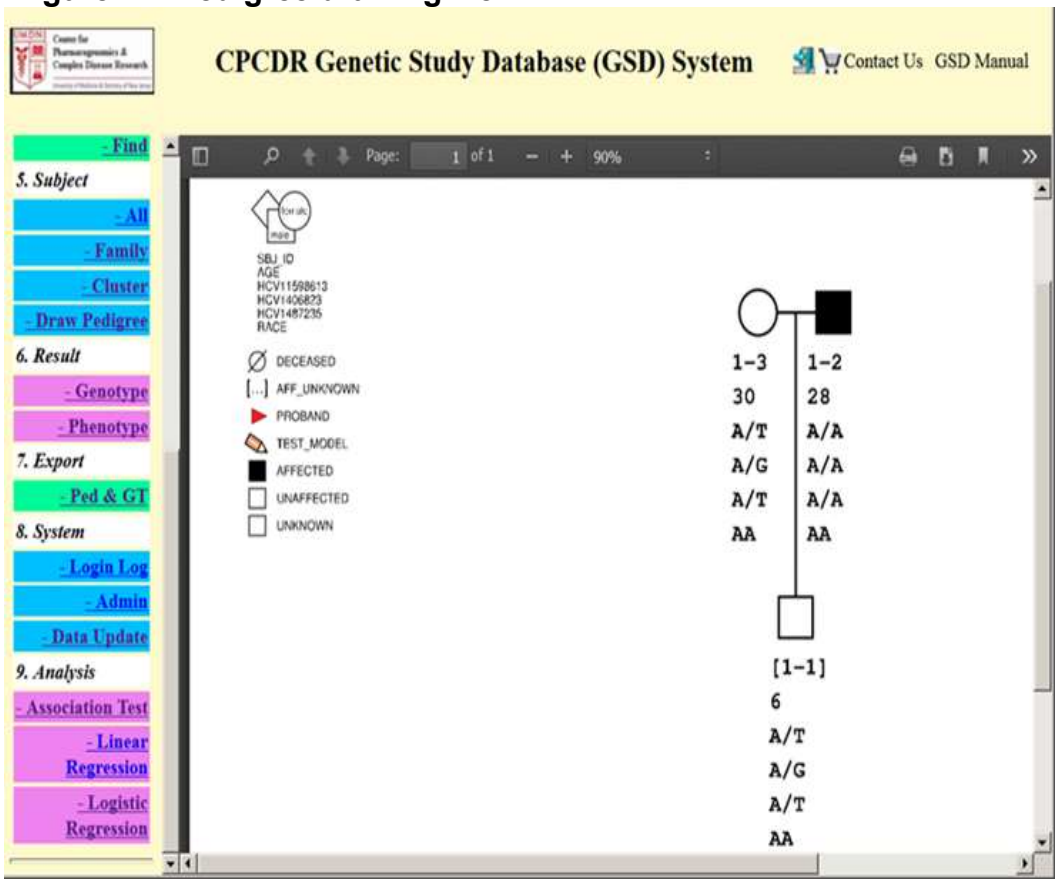
Figure 11: Study selected confirmed view



The second category, affection model selection (Aff Model), provides the option of selecting the disease affection status model. The option “All” means all the active affection status models will be listed in the working panel and only one model to be picked. Once click the “All” option under Aff Model a list of affection status models is shown alone with information including corresponding referenced study variable and the mapping between affection status values (Affected, Unaffected, Unknown) and code values of the study variable. The third category, variable selection (Variable), supports two options, "All" and "Cluster", of selecting study

variables. By clicking the “All” option, the main working panel is lists all currently active study variables from the selected study for user to select. The other way of selecting study variables is through study variable cluster. After clicking “Cluster”, a list of active study variable clusters is available to the user. User then selects the interested cluster to show all the currently active study variables under the cluster for reviewing study variable data or selecting study variables. Similar with variable selection the marker selection category (Marker) is also supporting “All” and “Cluster” options for selecting markers. In addition, it provides “Chromosome” option allowing selecting markers by chromosome number. The subject selection category (Subject) is also supporting “All” and “Cluster” options for selecting subjects as well as “Family” option for selecting subjects by family name. GSD is employing a public available program, Cranefoot v3.beta [56], as the underline pedigree-drawing engine. The option, “Draw Pedigree”, is available for drawing pedigree structure of selected family. Once click the option a list of currently active families is available to user for selecting a family to draw. After a family is selected the pedigree structure will be displayed in the working panel along with additional information including selected disease affection status model, study variables, phenotypes, study markers, and genotypes in PDF format (Figure 12).

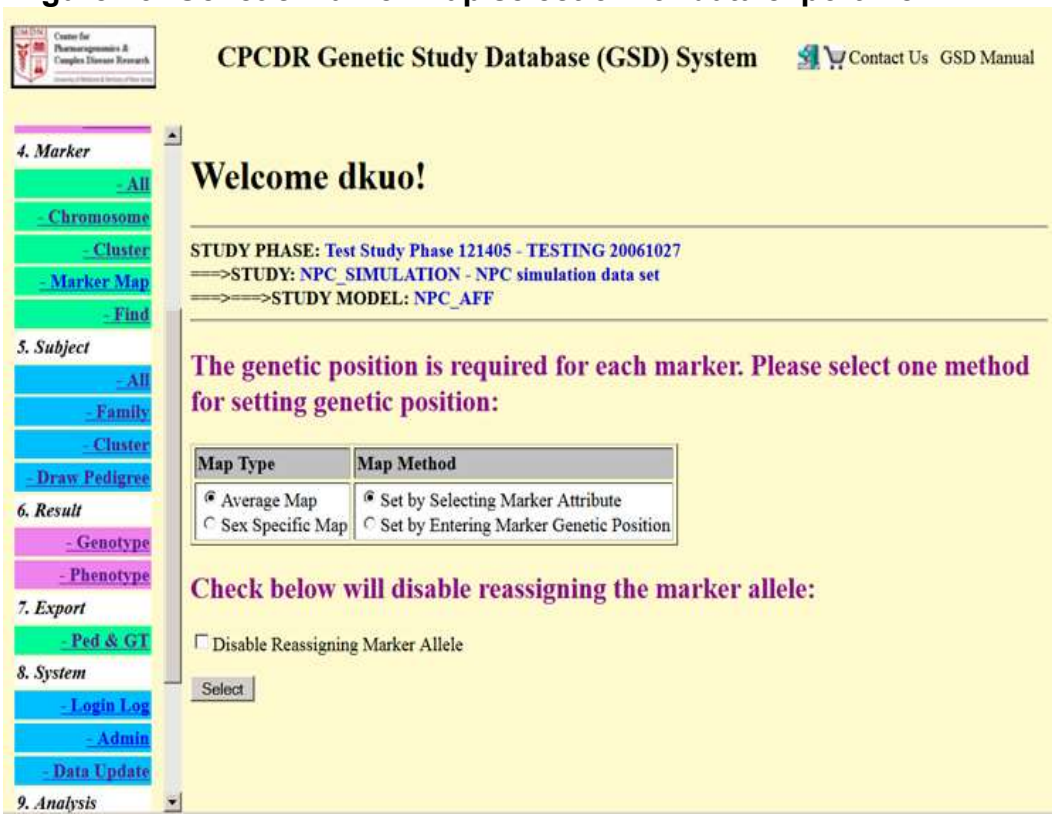
Figure 12: Pedigree drawing view



If study subjects were selected and either or both study markers and variables were selected, user can then select the options, “Genotype” or “Phenotype”, under the result selection category (Result) to display the genotype or phenotype result for review. GSD allows a subject to have different genotypes for the same marker, or to be classified as having different phenotypes for a given variable. However, in this case GSD will highlight the cell in the result table by an orange color warning the user of

the existence of these inconsistent results. It also provides a link to review the detail of all results and allow inactivation of unwanted results.

Figure 13: Genetic marker map selection for data export view

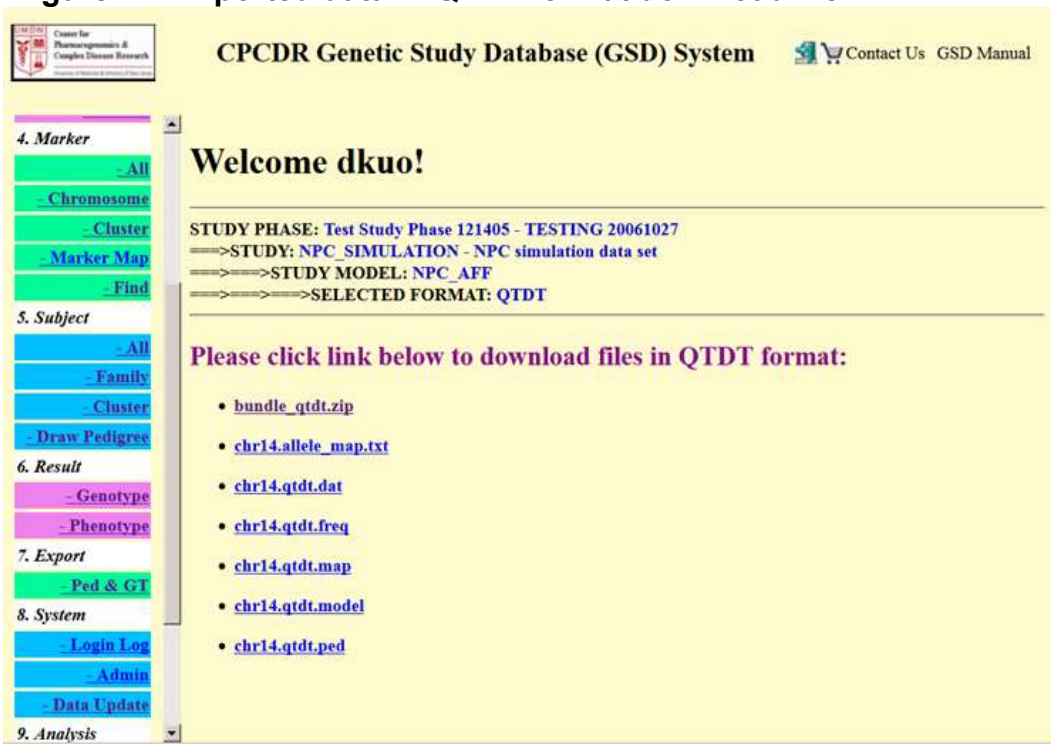


During the data export process GSD only exports single genotype/phenotype result and missing data will be used if multiple results found. One of key functions of genetic data management system is the capability of formatting data in supporting downstream linkage or association analyses. The data export category (Export) provides one option, “Ped & GT”, available for data export. After clicking the “Ped & GT”

option, GSD prompts for the source of marker genetic/physical map input and the type of genetic map to export, gender-specific map or average map (Figure 13).

Currently two kinds of marker map, physical and genetic, are required in GSD data export. Genetic map is mainly requested by the data analysis software and physical map is used to solve the problem of ordering markers with same genetic map position. There are two ways of setting the marker map information, manual entry through interface or retrieving from a marker attribute. In addition to marker map, the disease allele frequency, non-disease allele frequency, and penetrance of each of 3 genotypes at the disease locus are required for export. Currently GSD supports five data export formats for programs like Relpair, Merlin, LINKAGE, SUPERLINK, and PLINK. User has the choice of doing allele translation on the run-time or not. If allele translation is chosen an allele translation file will be included at the end of export process. Once the export process is finished all the data are exported by chromosome in the map order and data files are available for user to download individually through interface as well as a compressed file including all export files (Figure 14).

Figure 14: Exported data in QTDT format download view



2.4.4 System administration

The system operation (System) provides two key functions for administrating GSD user accounts and managing genetic data updates. The “Admin” option is only accessible to user who has Admin privilege. This option allows admin user to create new user account and assign privilege level and study access rule. Admin user also can update user information, including user info, privilege level and study access rule, or delete user account. Currently there are 4 user types, Admin, Laboratory Manager, Laboratory User and Demonstration User. Each user type has

different GSD functionality access privileges assigned. In the case of lacking access privilege, the corresponding functionality in the Tools panel will not be available to the user. In addition to functionality access control, each user account can be assigned data access privileges to only certain study phases and/or studies. The “Data Update” option is the key option in updating study data. Currently 4 types of data update are supported, including creating new records, updating existing records, inactivating and re-activating records. User have two ways of performing these data updates, either through interface to select data, subject, marker or genotype for example, for update or through uploading a comma-separate batch file with appropriate format. The “Data Update” option provides interfaces for updating various GSD data types, including study phase, study, subject population, subject, subject cluster, marker, marker cluster, genotype, variable, variable cluster, phenotype and disease affection status model. In GSD, every update generates audit trail data which include old record, type of update, reasons for update, user, and timestamp. A data record can never be updated without reason given and no data can be deleted from the system through interface. Any undesired data can be removed from user access through inactivate update and vice versa.

2.4.5 Statistical analysis functionalities

Finally, in terms of statistical analysis GSD offers three analysis options, Association test, Linear Regression, and Logistic Regression. The Association test option provides 2 test options, Chi-Squared test and Fisher's Exact test, for a selected binary trait. These tests are commonly employed in a typical case-control association study. Under the null hypothesis, for a given genetic variant there is supposed to be no significant genotype or allele frequency difference between case and control groups. These two tests examine the allele or genotype frequency deviation among case subjects as compared with control subject group. Fisher's Exact test is recommended for any cell with count below 5 in the contingency table. However, when there is suspected confounding risk factors other than genetic risk factor, for example age, gender or smoking, involved in the etiology of the disease studied, GSD provides regression analysis modeling the relationship between the dependent variable and one or more explanatory variables or covariates. For a dichotomous trait, disease status for an example, GSD offers logistic regression analysis. For a quantitative dependent variable, biomarker measurement for an example, GSD offers linear regression analysis to model the genetic effect and other possible confounding effects over the dependent variable. In addition to the test statistics and p value to be shown in the analysis result

table, user can add any meaningful marker attributes, mapping position or minor allele frequency for example, to the result table.

2.5 Results and discussion

Due to the uncertainty of underlying genes involved and gene-gene interactions in complex trait disease, genome-wide scan method is often the first option in surveying the linkage or association between the disease and markers. As a result of massive genotyping in genome-scan the management of large quantities of genotype data has created a bottleneck for high-throughput genotyping studies [57,58]. In the case of genome-wide SNP association study, one may easily generate close to one billion genotypes [43]. Therefore the development of a robust and reliable genetic study data management system helps to ease the bottleneck and thus speeds up both genotyping process and downstream data analyses. In order to deliver the capability of efficiently manipulating enormous amount of data coming from high-throughput genotyping systems and the flexibility in formatting data for various downstream analyses we have developed a genetic study data management system, GSD.

The focus of GSD design is really to facilitate the method of integrating data coming from modern high-throughput genotyping systems with subject demographic and phenotype data. The goal is to establish a platform-independent and web-accessible genetic data management

system allowing multi users to securely and efficiently manage their genetic data with features including, multiple co-existing studies, easy data import, capability of freely grouping study data into clusters, comprehensive data sorting, ability of defining multiple disease affection status, sophisticated pedigree drawing for illustrating family structure, capability of handling very large amount of data from high-throughput genotyping systems, capability of exporting data into various formats for downstream analyses, being able to individually characterize study variables, subject and marker by attribute, robust and IRB-proven user account management design and, finally, being able to preserve audit trail information. Although some data management systems have been developed and published [45,46,59-62], none of them satisfies our requirements. Some systems may lack the capability of handling enormous amounts of data coming from modern high-throughput genotyping systems. And some may have no capability in pedigree drawing, preserving audit trail information or supporting IRB-proven data management. Also some systems seem tied to the genotyping process and are limited to only one kind of genetic marker, STR, which prevents the system from incorporating SNP genotypes into studies. GSD enables the possibility of heterogeneous marker types co-existing in a study. This advantage also allows GSD to handle meta-analysis data collection from different genetic markers generated by various genotyping platforms.

GSD has several strengths in many perspectives, including networking security, user account management, data management, data annotation and data importation/exportation. First, let's discuss about networking security. GSD is a client-server application implemented over Internet using HTTP protocol. Through Internet, users across the globe can have access to the system without worrying about installing client software since GSD is a browser-based system. However, due to the sensitivity of the genetic study data exchanged between client and server, an encrypted communication protocol, HTTPS, has been implemented in GSD. With 128-bit encryption, password protection and automatic time-out functionality, GSD maintains a high degree of security in data exchange between client and server over Internet. Second, the user account management in GSD is done mainly in two categories, operation functionality and study accessibility. Before performing any data management operation a GSD user must be assigned with either one of 4 user types, Administrator, Laboratory Manager, Laboratory User and Demonstration User. Each user type is associated with different privilege levels. In a nutshell, the Administrator user has all the privileges in user account and study data management, including data creating, updating, inactivating, re-activating and exporting. A Laboratory Manager has all the privileges in study data management except user administration. A Laboratory User has no privilege involving data update and has basically

only study data querying and exporting privileges. Finally, Demonstration User has very limited data querying privilege. In study accessibility, GSD requires permission to be granted to the user for a study prior to accessing data records. Therefore, GSD has advantages in compliant IRB requirements of securing human subject data.

Third, from the data management prospective, GSD is capable of handling multiple concurrent users and studies in the system. It chooses Oracle relational database as the backend database server to take advantage of Oracle's ability in handling huge amount of genotype as well as other features. GSD is not only designed to handle both case-control and family-based studies, but also allows both case-control and family-based subjects to co-exist under the same study by assigning subjects with different population. Generally 4 types of data object, disease definition model, study variable, study subject and study marker, define a study under GSD. Under each object type GSD supports a panel of options for reviewing and searching data objects. One of advantages is through implementing the cluster option which enables sub-grouping study subjects, markers and variables. This turns out to be very convenient in selecting data records and data exporting for various kinds of downstream exploratory analyses, for example exporting and analyzing study dataset respectively in different ethnic group. Also the shopping cart option gives another advantage in reviewing and managing selected subjects, markers

and variables, with table header sorting function facilitating the process of reviewing and dropping undesired object from the shopping cart. Moreover the design of defining disease affection status by referencing a study variable can be very handy especially in exploring data analyses between various disease forms of a complex trait.

During study life cycle the study data are dynamically changing. Some subjects whose DNA were not collected may become available for genotyping or errors were found during clean-up processes. For example, in family-based study, the clean-up processes may be using a set of high quality markers to validate reported pedigree structures. As a result of failing validation, some subjects may be dropped from the study. Once pedigree structures are confirmed, they then can be used to detect Mendelian discrepancies in genotypes among family members. Genotype discrepancy can come from genotyping error, bad DNA quality, low DNA quantity, bad assay or even by chance. In the case of finding discrepancies, the suspected genotypes should be dropped from the study. To accommodate the needs from handling the dynamic changes, GSD is providing 4 types of data update methods, New, Update, Inactivate and Re-activate. After deciding the type of data to update, changes can be made to the system through either interface-user interaction or importing a batch file. Batch update is very convenient especially in the situation of updating study with millions of genotypes since it is not feasible to make

hundreds, if not thousands, of changes to the system through interface-user interaction. Finally in order to maintain the data integrity, GSD maintains audit trail information for every update made through the system. Instead of deleting the old record from the system, GSD actually is creating a new record with links to the old record, which was inactivated after update, and audit trail record. Therefore it is achievable to review the record from the activated one to the originally created one with updates happening in between or vice versa.

In data annotation prospective, GSD is using attribute to address the issue of heterogeneity in data object annotation. It is not realistic to design a, for example, subject table with endless number of table columns to handle all kinds of possible characters which may not be seen in all subjects. Also it will be inefficient and a waste of table space to design a table with only a small portion of records having data for most of annotation columns. By implementing attribute concept GSD can add as many attributes as possible to annotate a data object at any point of a study and also make them available to object selection and data exporting processes. One of strengths in data annotation is pedigree-drawing capability. Extended pedigree is commonly seen in the complex trait study therefore the pedigree structure illustration is playing a key step in pedigree analysis. For example during the pedigree validation it sure helps to see the pedigree structure drawing with subject marked by affection

status alone with genotypes and phenotypes. Using Cranefoot as the drawing engine, GSD is not only drawing pedigree with affection status and genotypes it also shows the disease status, proband status, family member counts and marker legends. In addition, desired phenotypes can also be included in the drawing.

As one of the goals of genetic study management system is to facilitate the downstream data analysis process, the capability of easily and accurately importing and exporting data in GSD is another selling point. Mentioning previously, updates can be done in GSD through two methods, user-interface interaction or batch file importation. While reviewing the selected data objects in the shopping cart, the view export option is available for selected subjects, markers, study variables, genotype results and phenotype results. The exported data are following the batch file format; therefore this option is extremely useful in copying data objects within same study or between studies. Finally GSD currently supports 5 types of data export format including Relpair, LINKAGE, MERLIN, SUPERLINK, and PLINK format. By design GSD's exporting capabilities also provide several additional advantages. First GSD is capable of exporting only a subset of families and markers. This is a critical step in dealing with disease gene heterogeneity. For example some complex trait may have higher disease frequency in certain ethnic group. Therefore, it may not be detected in analyzing data exported with mixed ethnic groups.

Secondly, multiple versions of marker maps can be stored in GSD and it is supporting two kinds of genetic map, average genetic map or sex specific genetic map, in data export. Thirdly, GSD is doing the allele translation at run time and every export comes with an allele translation file listing mapping information between database allele and file allele. Finally, data are exported by chromosome respectively, and depending on the format chosen, the number of data files for each chromosome is varied. User can either review and save each data file individually by clicking links on the export result page or download a zip file which includes all data files exported to the local machine.

2.6 Conclusions

As the next generation of high throughput genotyping technology is rapidly coming to market and an enormous amount genotype data have been generated in complex trait linkage or association studies, the capability of managing genetic study data has become a bottleneck. The design of GSD is aimed to provide a better solution to ease the bottleneck. It has many advantages over secure data communication, comprehensive database design, easy-to-manage user interface and user-friendly import/export functions that make it a powerful and unique tool in dealing multiple-center and large-scale genome-wide association or family-base

linkage studies which have become more popular in the post human genome era.

CHAPTER 3: POPULATION-BASED GENOTYPE IMPUTATION OPTIMIZATION

Abstract

Population-based genotype imputation has been used to boost the power of Genome-Wide Association Studies (GWASs). Due to the existing uncertainty in predicting genotypes of unmeasured genetic variants, it is a standard practice to apply filtering measures after genotype imputation has completed to remove poorly imputed variants or genotypes that might introduce noise to the downstream association tests. Most of genotype imputation programs assign an imputation quality metric value to each of the imputed SNPs indicating the possible level of correlation between imputed data and perfectly observed data. Although it's often recommended to perform post-imputation filtering based the quality metric alone, here we would like to evaluate the merit of using the quality metric as the only filter in removing poorly imputed SNPs and genotypes. According to our testing results, filtering based on the quality metric alone is found to be not effective in getting rid of badly imputed SNPs. To that extent, In addition to the quality metric, more effective measures need to be included in post-imputation filtering, for example imputed genotype probability and SNP call rate. After searching for the filters which balance

between yield and accuracy of population-based genotype imputation, a combination of three measures is recommended; a quality metric cutoff value 0.4, imputed genotype probability cutoff value 0.98, and imputed SNP call rate cutoff value 0.7. This filter combination has been tested and validated in imputation runs on input datasets with different level of genome coverage and proved to be a general, robust, and effective filtering measure for obtaining high accuracy and reasonable yield in genotype imputation filtering.

3.1 Introduction

Genome-Wide Association Studies (GWASs) and meta-analyses have been successful in identifying common variants influencing many complex traits, finding candidate susceptibility variants to guide fine-mapping, and facilitate meta-analyses that combine studies genotyped on different sets of variants [18-20,33,63-69]. When compared to using only genotyped markers, genotype imputation has extended the study power through inferences of unobserved markers in a study sample by using the linkage disequilibrium among markers present in a reference panel, such as those from the HapMap project [23,24].

Although most imputations have used HapMap 2 data as the reference panel, the recently available Phase I 1000 Genomes (1KG) Project data have provided higher resolution human genome sequence variation [25,70]. The advantage of using 1KG data as the reference panel for imputation in GWASs is the ability to impute much larger number of SNPs than using HapMap data. The 1KG Project reference panel includes 1092 individuals across 14 populations which are classified further into 4 super-populations according to the ethnic background. A study evaluating the performance of genotype imputation using data from 1KG Project [71], showed that 11.4 million SNPs are found among 1KG-EUR panel as compared to 2.5 million SNPs in HapMap 2 CEU panel. The report concluded that 1) 1KG reference panel provided much higher imputation yield than the HapMap 2

panel, 2) 1KG reference panel provided high imputation accuracy which is almost identical to the accuracy from using HapMap 2 panel, 3) imputation accuracy of rare and low frequency SNPs from using 1KG reference panel is very high and almost identical to accuracy of common SNPs, and 4) 1KG-based imputation can increase the opportunity to discover significant associations for SNPs across the allele frequency spectrum.

Nevertheless, one practical problem of handling results from genotype imputation using 1KG panel is the large number of imputed SNPs. In our experience, before any filter applied we found more than 38 million SNPs could be imputed by a GWAS array when all 14 1KG populations are included in the reference panel. Although most modern imputation programs provide an imputation quality metric indicating the correlation between the true genotype and predicted genotype, for example the r-square from MaCH [34,72] or the information metric from IMPUTE2 [73,74] but they have not been shown to be mathematically equivalent to the r-square statistic for LD between a tag-SNP and a disease SNP which influences the sample-size inflation factor [75]. In practice, post-imputation filtering is still recommended to remove poorly imputed variants or genotypes that may bring in noise into downstream association tests.

Another factor which may have impact on the merit of imputation is the content of microarrays. For example the GWAS array has better genome coverage compared to the Exome array. The GWAS array tends to

include more common SNPs and is designed to have fairly good genome coverage, mainly covering as many LD blocks across the human genome as possible. On the other hand, Exome array is designed to focus mainly on functional variants in coding regions or close to gene regions and most of them are rare variants. Therefore it is predictable that imputation using GWAS array genotypes may have better result in terms of number of SNP imputed, SNP call rate, and genotype accuracy.

This study is designed to investigate the difference between GWAS and Exome arrays in terms of imputation metric distribution, imputation yield, and genotype accuracy. Most importantly, it will try to identify an optimized set of post-imputation filters which balances between genotype imputation yield and accuracy. Ninety-six individuals mostly of European origin were recruited and genotyped on both Affymetrix World GWAS (LAT) array and Exome array. Following the vender recommended genotype SNPolisher QC procedures, the array genotypes were then collected and used for genotype imputation. Although there are many popular genotype imputation programs, for example MaCH [34,72], IMPUTE2 [73,74], fastPHASE [76], PLINK [14] and BEAGLE [35], studies have found that IMPUTE2 is optimized when run with all 1KG reference panels and shows superior results among different imputation methods [18,77,78]. Therefore, IMPUTE2 was chosen for whole genome genotype imputation in this study.

Due to the uncertainty in predicting unobserved genotypes, poorly imputed SNPs must be filtered from any downstream association test. All the genotype imputation programs provide a similar imputed SNP quality metric which indicates the level of the correlation of imputed SNP data with perfectly observed genotypes, for example the R-Square from the famous BEAGLE and MaCH programs and the INFO value from IMPUTE2 program. For each SNP, IMPUTE2 reports an information metric (INFO), which has values that range between 0 and 1. The INFO values near 1 indicate that a SNP has been imputed with high certainty. As recommended by the IMPUTE2 developer group that although no universal INFO cutoff value has been established for post-imputation SNP filtering, various groups have used cutoffs of 0.3 and 0.5, for example. However they also caution the right cutoff threshold for post-imputation filtering may differ between studies. Nevertheless, it's not clear to what extent the selection of quality metric cutoff should be different between studies and if filtering on the metric alone is effective in getting rid of badly imputed SNPs and genotypes. For example, when applying an INFO cutoff of 0.4, any SNPs imputed with INFO value below or equal to 0.4 will are excluded from downstream analyses. The more stringent the INFO cutoff is used for filtering, the less number of imputed SNPs left after filtering. In addition to the SNP INFO metric, there is a posterior probability reported with each imputed genotype which could be used for filtering out

uncertain genotypes. Other statistics can be investigated for SNP filtering including, imputed SNP call rate (number of imputed study individuals / total number of study individual for impute) and allele frequency. To evaluate the performance of using quality metric as the only filter and identify the optimized set of filters, the concordance of SNPs not on the Exome array but imputed by it and present on the GWAS array are evaluated. Similarly, SNPs not on the GWAS array but imputed by it and assayed on the Exome array are evaluated, after applying different parameters to filter out potentially unreliable genotypes.

3.2 Materials and methods

3.2.1 Study population

Saliva samples of ninety-six individuals of European origin, a subset of subjects recruited in a Fluorosis study held in Newcastle and Manchester, UK [79], were obtained using Oragene self-collection kit (DNAgenoTek, http://www.dnagenotek.com/US/products/dnacollectionkits.html). DNA molecules were extracted from saliva after following manufacturer manual. All samples were selected for GWAS array and Exome array genotyping on Affymetrix Axiom platform. The study protocol was approved by the Institutional Review Board of University of Medicine and Dentistry of New

Jersey, and written informed consent was obtained from each subject prior to taking part in the study.

3.2.2 Genotyping

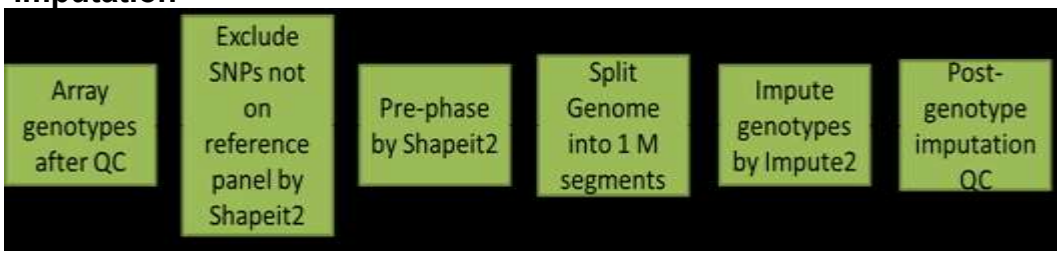
Genotyping was carried out by Affymetrix GeneTitan Multi-Channel Instrument using Affymetrix Axiom solution for GWAS array (818,154 probe sets) and Exome arrays (319,283 probe sets). Following manufacturer standard operation procedure, 20 ng of DNA molecules from each study individual were used. The Affymetrix Genotyping Console (GTC) Software and SNPolisher QC procedures were used for calling genotypes and insuring calling quality.

3.2.3 Genotype imputation and data analyses

Microarray genotypes were subject to manufacturer QC measures, including SNPolisher QC procedures, Core SNP call rate filtering and SNP metric filtering by Affymetrix GTC software. Additional QC measures were applied to SNP calls, including HWE P > 10-4 and no positive control discrepancy. A commercial software package, Golden Helix SNP & Variation Suite (Golden Helix, http://www.goldenhelix.com/SNP_Variation/index.html), was used for genotype data management. Population-based genotype imputation was

carried by first using SHAPEIT2 [80-83] for haplotype estimation (pre-phasing). Pre-phasing involves speeding up this process by first estimating haplotypes from GWA study samples, and then imputing alleles into these haplotypes from a reference haplotype panel. The phasing of the GWA studies samples needs only be done once so that when a new haplotype reference panel becomes available the imputation step is very quick. Following the pre-phasing process, the IMPUTE2 program was run with 1000 Genomes Project Phase I genotypes as the reference panel for genotype imputation. In general, the imputation procedure was following the guideline, Minimac: 1000 Genomes Imputation Cookbook (http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook).

Figure 15: The process flow of population-based genotype imputation



Genotype imputation procedures were run on a High Performance Computing Linux 62 8-cores nodes cluster. All nodes have a minimum of 12 G bytes RAM (the majority have 16 Gbytes) and 1 terabyte of on board

scratch space. Project storage is maintained by a 30 Terabyte Gluster file system with each node supplying one terabyte of storage to the ensemble. In this analysis, whole genome genotype imputation was done on chromosome segments of average length 1 million bases (Figure 15).

Raw imputed genotypes were imported into SVS and subject to various filtering measures. To reduce the complexity, insertion/deletion (INDEL) variants were excluded from analysis. Array SNPs which are not designed but imputed by the other array are identified by having the same mapping position (Human Genome Assembly Build 37) and therefore evaluated for genotype imputation accuracy test. Array genotypes and imputed genotypes are both mapped on to the forward strand of genome assembly. Statistics to be evaluated for genotype and SNP filtering include IMPUTE2 imputed posterior genotype probability, SNP INFO value, and SNP call rate. The SNP genotype concordance between array for the assayed genotypes and the genotypes imputed by the other array are calculated after applying statistical filters.

First, I compared the distribution of the Impute2 imputed SNP quality metrics INFO values between two imputation datasets, one by Exome array and the other one by GWAS array. To evaluate the effectiveness of INFO value filter, I followed the recommendation of Impute2 developer and applied an INFO cutoff value of 0.4 on both array imputation datasets and evaluated the concordance between before and after applying the

filter. To evaluate the effect of imputed genotype call rate over the genotype concordance, I inspected the concordance distribution at different call rate thresholds on both array imputation datasets. Finally, I used Impute2 Info value, imputed genotype probability and call rate for post-imputation filtering. Info has 3 proposed cutoff values including 0.1, 0.4, and 0.6. The imputed genotype probability has 4 proposed cutoff values, 0.5, 0.7, 0.9, and 0.98. In terms of call rate cutoff value, 3 values, 0.5, 0.7, and 0.9, are proposed. To find out the best combination of the cutoff values of these 3 filters which balance between genotype accuracy and yield, the concordance and yield analyses were done on all possible combination of the proposed cutoff values proposed.

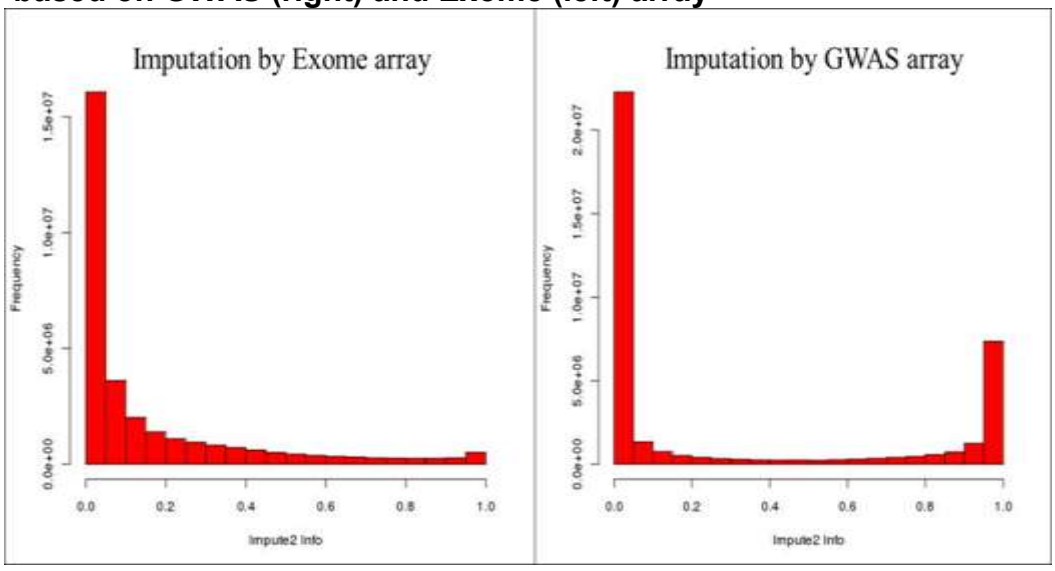
3.3 Results

3.3.1 Quality metric distribution of imputed array SNPs

The genotype imputation was done by IMPUTE2 using all 1,092 1KG Phase I samples from 14 populations as the reference panel. Due to the computation limitation, the whole genome genotype imputation is still not available. Therefore the imputation was done after splitting genome into thousands of around 1 million bases segments. After applying pre-imputation QC measures, 652,190 LAT (GWAS) array SNPs and 59,545

Exome SNPs were used for genotype imputation. Without applying any filters, 38,711,309 SNPs are found with genotypes imputed by the GWAS array and 31,146,888 SNPs are imputed by Exome array. Figure 16 shows the distribution of IMPUTE2 INFO metric of imputation done by GWAS and Exome array respectively. As expected, GWAS array has better genotype imputation yield when compared to Exome array. Most importantly, their INFO metric distribution looks different which very likely is due to the difference in array content.

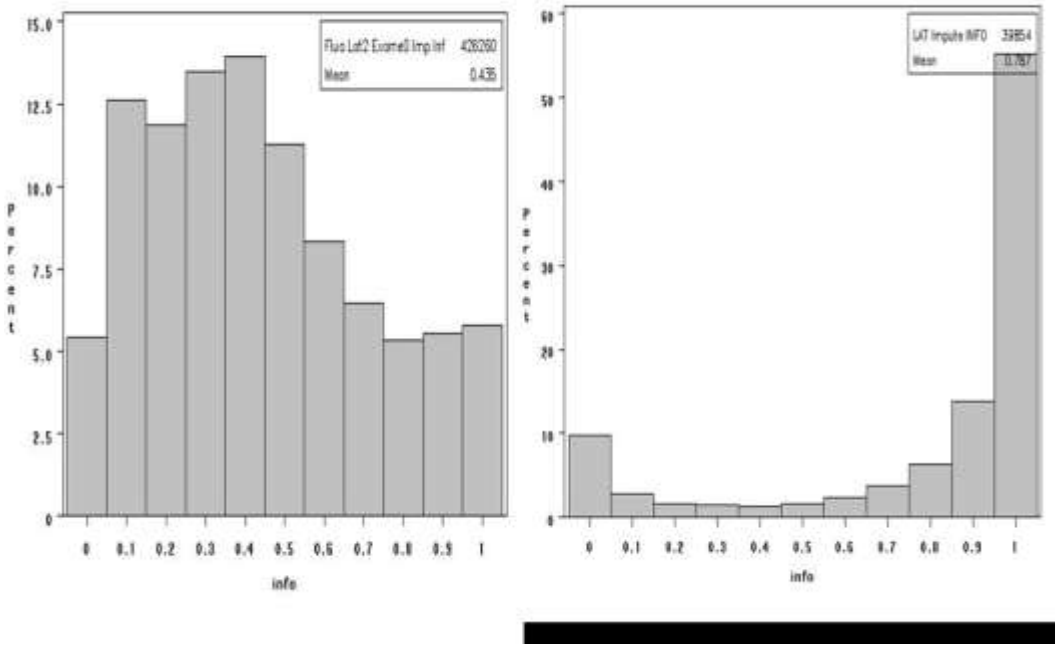
Figure 16: The IMPUTE2 INFO metric distribution of imputation runs based on GWAS (right) and Exome (left) array



The Exome array imputation has higher frequency in lower Info value SNPs and GWAS seems to have higher frequency at the higher Info end. This is most likely due to the genome coverage difference between these

2 types of array. GWAS array is designed with good genome coverage for maximizing the genotype imputation power. On the other hand, Exome array is designed with a focus on genome coding sequences and some rare variants found in NHLBI GO Exome Sequencing Project (https://esp.gs.washington.edu/drupal/). Therefore with better genome coverage, the GWAS array tends to impute more high INFO value SNP than Exome array. The INFO metric distribution looks even more different after excluding array SNPs which could not be imputed by the array content of the other array type.

Figure 17: The IMPUTE2 INFO metric distribution of imputation among array (Left - GWAS array SNPs imputed by Exome array; Right - Exome array SNPs imputed by GWAS array)



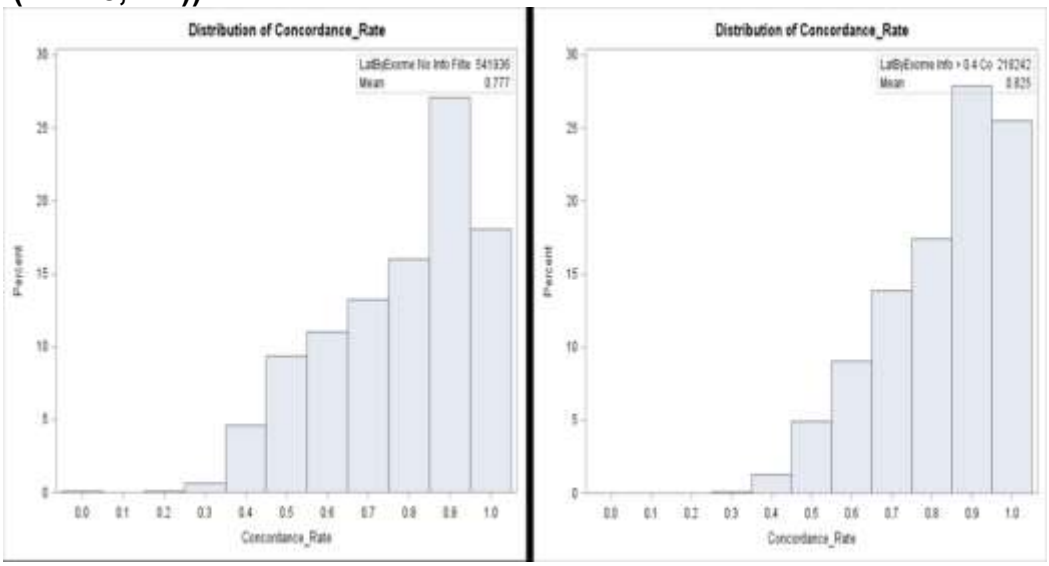
Among SNPs imputed by GWAS, 39,854 SNPs are designed on Exome array. Among SNPs imputed by Exome, 426,260 GWAS array SNPs are found. Figure 17 shows the distribution of IMPUTE2 INFO metric among GWAS array SNPs which got imputed by Exome array (left) and Exome array SNPs which got imputed by GWAS array (right). Comparing these 2 INFO distributions, it's obvious that applying INFO metric filter alone is not sufficient for post-imputation filtering. More statistics and/or combinations of statistics need to be examined to come up with a better filter set.

3.3.2 Inadequate power of quality metric filtering

To further show that INFO value filtering alone is not sufficient for filtering out poorly imputed SNPs and genotypes, genotype concordance was also compared before and after applying the INFO cutoff value 0.4 in both array imputation datasets. In order to assign predicted genotype based on predicted genotype probability, a probability cutoff value must be applied. Therefore a loose genotype probability, 0.5, is chosen which means one of 3 possible genotypes (minor allele homozygous, heterozygous, and major allele homozygous) is assigned to a subject only when its predicted probability is greater than 0.5 for a given imputed SNP. Figure 18 shows the concordance distribution of GWAS designed array

SNPs which were imputed by Exome array before (left) and after (right) applying INFO filtering. It seems that although more than 50% of imputed SNPs are filtered by applying INFO cutoff value 0.4 as recommended by the IMPUTE2 developer, the genotype concordance actually has not been improved by the filtering.

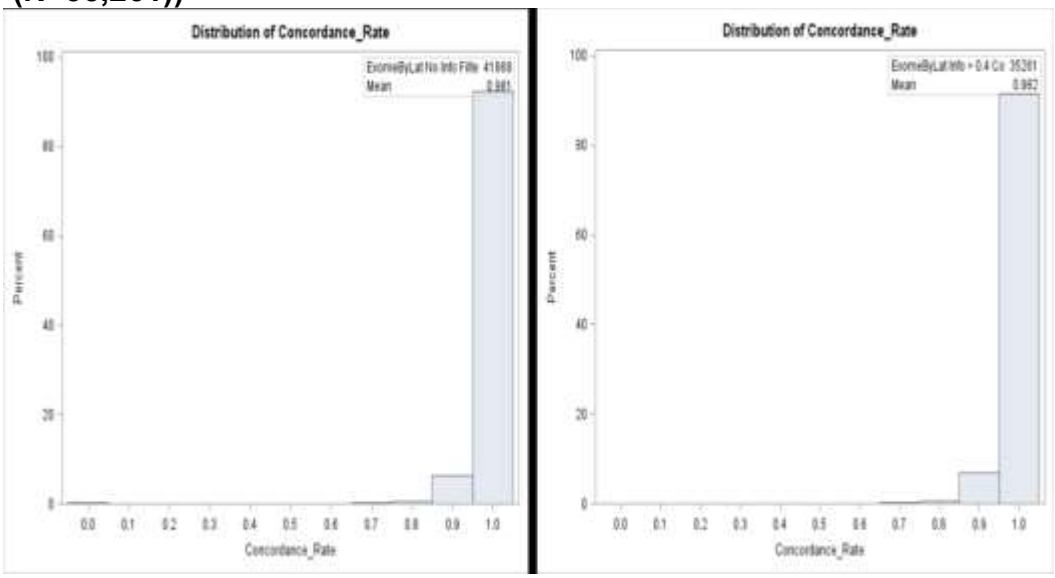
Figure 18: The effect of INFO filtering over Exome array imputation (Left - No Info filtering (N=541,936); Right - Info value > 0.4 Filtering (N=218,242))



Similar phenomenon is seen over applying the INFO cutoff value 0.4 on the Exome array designed SNPs which were imputed by GWAS array (Figure 19). As expected due to the higher confidence in imputation, same INFO filtering measure actually filters out less percentage of SNPs imputed by GWAS array. Nevertheless, the genotype concordance

actually has not been improved by the filtering. These results further confirm the doubt of treating the quality metrics as the analogous to the r-square statistic for LD between a tag-SNP and a disease SNP used in the estimation of sample-size inflation factor (Huang et al., 2009) and demonstrate the lack of power of using INFO metric as the only post-genotype imputation filter.

Figure 19: The effect of INFO filtering over GWAS array imputation (Left - No Info filtering (N=41,868); Right - Info value > 0.4 Filtering (N=35,261))



3.3.3 Other candidate measures for post-imputation filtering

In searching for the other measure which can be effectively used in post-genotype imputation filtering, the imputed SNP call rate is a potential candidate. First, the call rate distribution of GWAS array SNPs which got

imputed by Exome array and Exome array SNPs which got imputed by GWAS array is investigated. To assign predicted genotype and insure the call rate is based on imputed genotypes with high confidence, a predicted genotype probability 0.98 is applied prior to call rate calculation.

Figure 20: The genotype call rate distribution of array imputations (Left – GWAS array SNPs imputed by Exome array (426,260); Right – Exome array SNPs imputed by GWAS array (39,854))

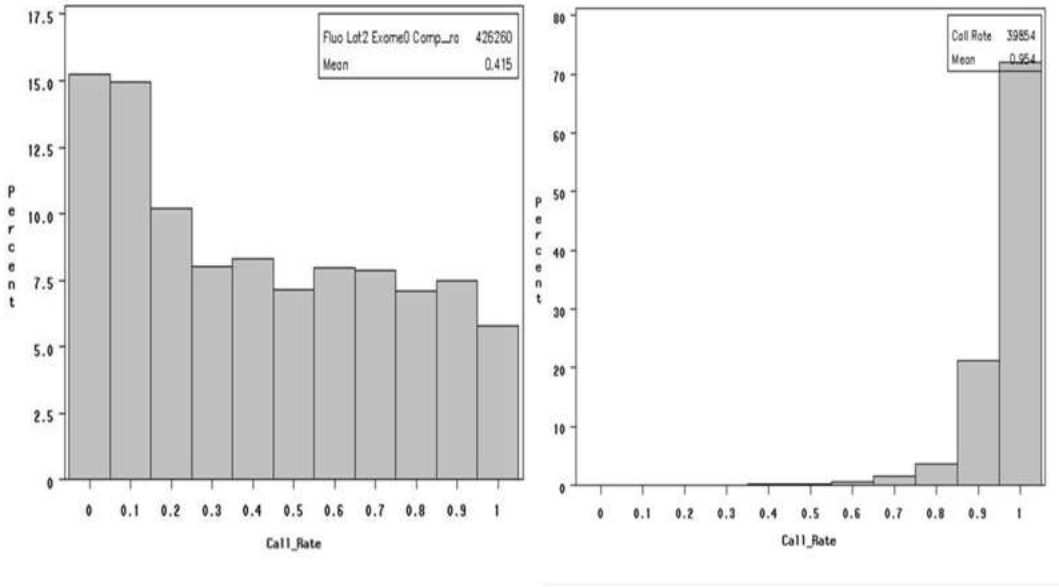
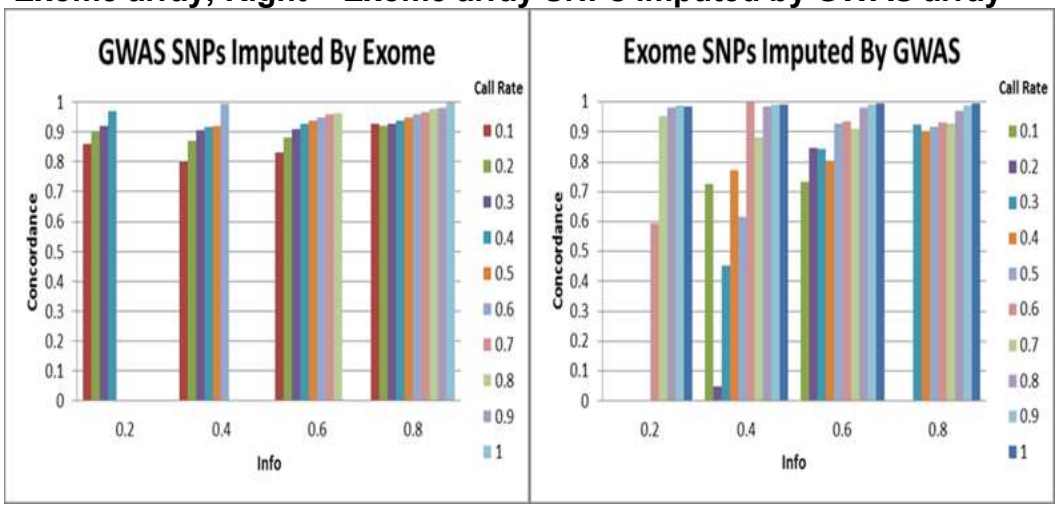


Figure 20 shows very different call rate distributions between these 2 datasets. GWAS array imputation (Right) has more SNPs with high call rate than Exome array imputation (Left). More than 90% of Exome array SNPs imputed by GWAS array has genotype call rate greater than 90%. On the other hand, less than 15% of GWAS array SNPs imputed by

Exome array is found with genotype call rate greater than 90%. Again, this is most likely due to the genome coverage difference between GWAS and Exome arrays.

Figure 21: The effect of genotype call rate over genotype concordance distribution (Left – GWAS array SNPs imputed by Exome array; Right – Exome array SNPs imputed by GWAS array



Next, to further check the effect of genotype call rate versus the genotype concordance under different INFO value groups, the genotype concordance analysis is done between different call rate and INFO value group combinations. The left hand side plot of the Figure 21 shows the concordance of GWAS array SNPs which are imputed by Exome at different combinations of call rates, including 0.1, 0.2, 0.3, and so on, and Info values, including 0.2, 0.4, 0.6 & 0.8. The Y axis is the genotype concordance and X axis is the 4 INFO value groups. Within each INFO

value group, genotype concordance is calculated for each call rate bin. As shown in the Figure 21, the concordance seems correlated with the call rate and this is also true across all 4 INFO groups. Similar phenomenon is seen among the Exome array SNPs which are imputed by GWAS. This analysis demonstrates clearly that call rate can serve as a good candidate for post-imputation filtering.

Therefore, three measures, INFO value, imputed genotype probability, and genotype call rate, are chosen for post-imputation filtering optimization. Three cutoff values of INFO, including 0.1, 0.4, and 0.6, and 4 cutoff values of imputed genotype probability, including 0.5, 0.7, 0.9, and 0.98, and 3 call rate cutoff values, including 0.5, 0.7, and 0.9, are proposed. To find out the best combination of the cutoff values of these 3 filters, the concordance and yield analyses were done on all possible combination of these proposed cutoff values. For each genotype probability group, genotypes are filtered by the predicted genotype probability cutoff followed by applying 3 INFO cutoff filters, 0.1, 0.4, and 0.6, respectively which excludes imputed SNP with IMPUTE2 INFO value below the cutoff. Finally the genotype call rate cutoff, 0.5, 0.7, and 0.9, are applied to each INFO cutoff group and the genotype concordance is calculated for the GWAS array SNPs imputed by Exome array as well as the Exome array SNPs imputed by GWAS array. In addition, the yield after

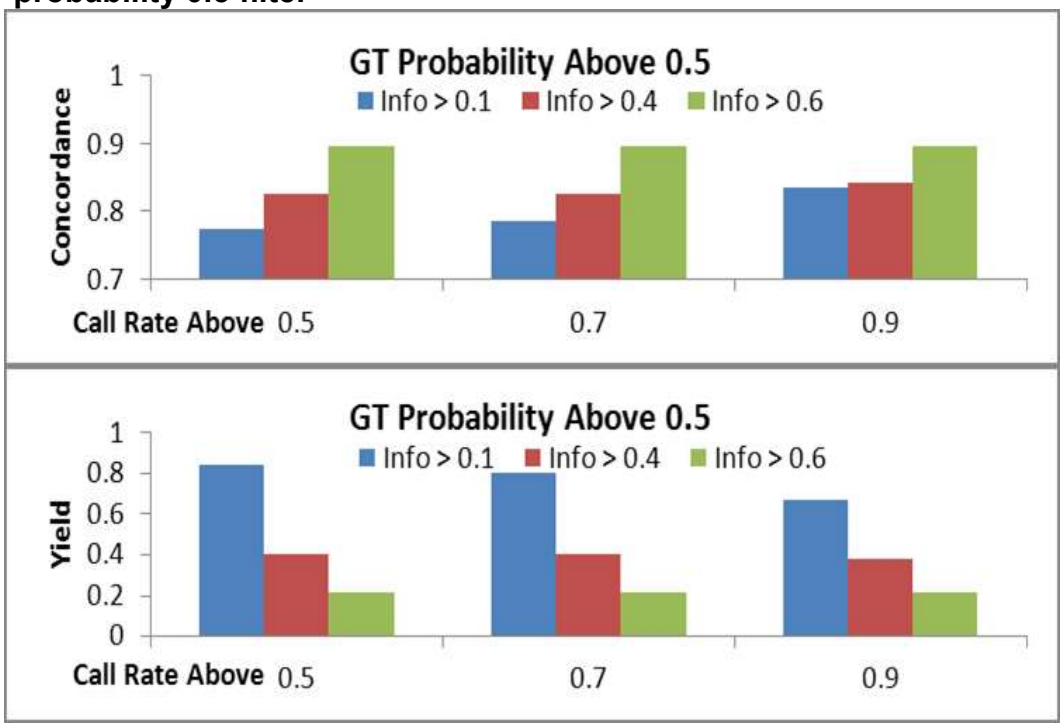
applying each combination of genotype probability, INFO value, and genotype call rate is also calculated.

3.3.4 Exome array imputation result optimization

First, 542,210 GWAS array SNPs imputed by Exome array dataset is analyzed. Figure 22 shows the concordance (top) and yield (bottom) analysis results after applying the genotype probability cutoff 0.5. It clearly shows within each call rate cutoff group, the concordance increased when the more stringent INFO cutoff was applied. For example, after applying probability cutoff 0.5, call rate cutoff 0.5 and Info cutoff 0.1, the concordance is 0.77 which increases to 0.90 after applying probability cutoff 0.5, call rate cutoff 0.5 and Info cutoff 0.6. Similarly, within each INFO cutoff group the concordance increases after applying a more stringent call rate filter. For example, within the INFO cutoff 0.1 group the concordance increases from 0.77 to 0.84 after call rate cutoff changing from 0.5 to 0.9.

On the other hand, as expected applying more stringent filtering measures reduces the yield of imputed SNPs. Nevertheless, after passing through these filtering measures these SNPs presumably are imputed with better accuracy.

Figure 22: Genotype concordance and yield after applying genotype probability 0.5 filter

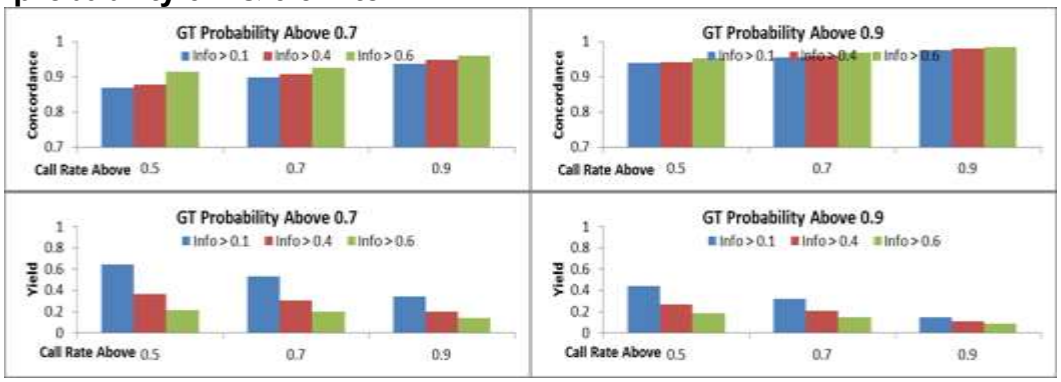


As shown in the bottom yield plot of Figure 22, in the call rate 0.5 group although the concordance increases from 0.77 to 0.90 the yield dramatically reduces from 0.84 to 0.21 after changing the INFO cutoff from 0.1 to 0.6. When compared within INFO cutoff 0.1 group, concordance increases from 0.77 to 0.84 after call rate cutoff changing from 0.5 to 0.9, the yield drops from 0.84 to 0.67.

The concordance gets improved further after more stringent genotype probability cutoffs are applied. Figure 23 shows the concordance and yield plots after applying 0.7 (left) and 0.9 (right) as the genotype probability

cutoff. Similarly the concordance increases after more stringent INFO value or genotype call rate cutoff applied for filtering.

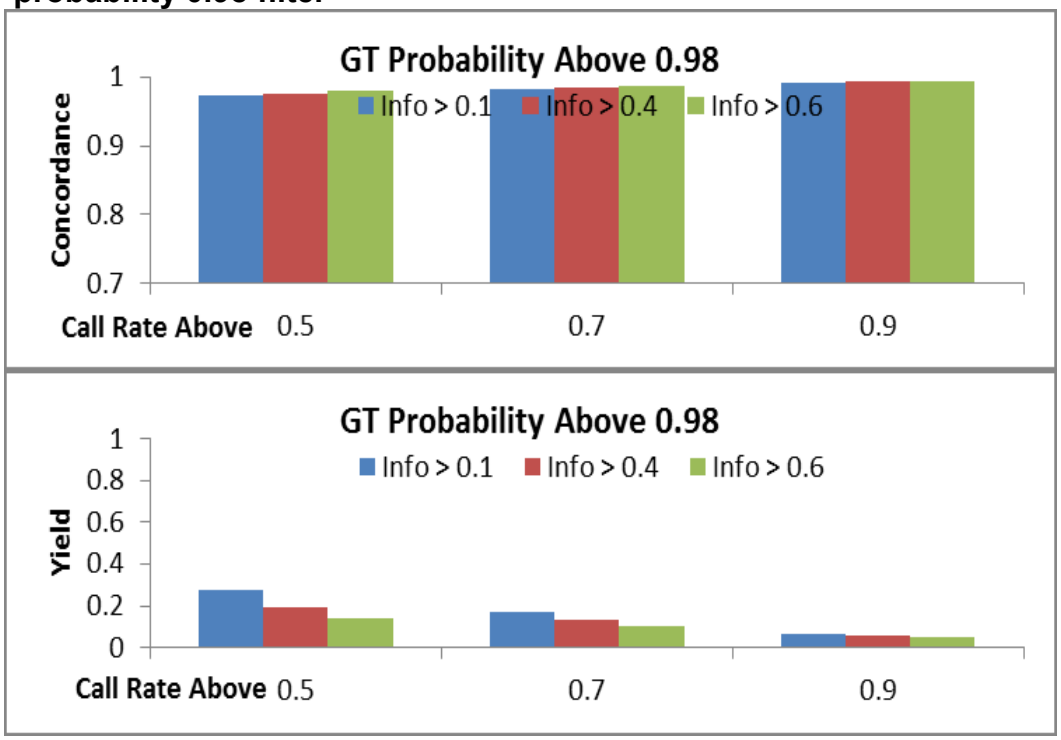
Figure 23: Genotype concordance and yield after applying genotype probability 0.7 & 0.9 filter



For example, under genotype probability cutoff 0.7 and call rate cutoff 0.5 filters, the concordance increases from 0.87 to 0.91 after INFO value cutoff changes from 0.1 to 0.6 and, as expected, the yield changes from 0.65 to 0.21. Instead of applying a more stringent INFO cutoff, if more stringent call rate cutoff is applied, 0.9 for example, the concordance increases from 0.87 to 0.94 and the yield changes from 0.65 to 0.34. Apparently in this setting, applying higher call rate cutoff filter gives better concordance and yield as compared to applying higher INFO cutoff filter. Another example, when more stringent genotype probability cutoff is applied, 0.9 for example, with call rate cutoff 0.5 and INFO value cutoff 0.6, it gives a concordance of 0.953 and yield of 0.18 which is probably

slightly better than applying genotype probability cutoff 0.7, call rate cutoff 0.9 and INFO value cutoff 0.6 which gives a concordance of 0.959 and yield of 0.14. As expected, there is always a tradeoff between impute accuracy and between these two filter sets giving close concordance level and higher call yield one would probably choose the former one for post-imputation filtering.

Figure 24: Genotype concordance and yield after applying genotype probability 0.98 filter



To further investigate the impact of more stringent genotype probability filter over the concordance and yield of post-imputation filtering, the

genotype probability cutoff 0.98 is applied and the concordance and yield plots are shown in the Figure 24. Under the same setting of call rate cutoff 0.5 and INFO cutoff 0.1 after applying genotype probability cutoff 0.98 the concordance increases to 0.97 and yield drops to 0.28. Again, applying more stringent cutoff of call rate and/or INFO further improves the concordance and reduces the yield.

It's been reported in studies that even 2% genotype imputation error can have profound influence over the association analyses [75,84,85]. Therefore it would be better to optimize the post-imputation filtering to have concordance equal to or over 0.98 and, yet, balance the yield. After excluding filter combinations which generate concordance below 0.98, Table 2 presents filter combinations with concordance greater than 0.98. Among these filter combinations, the one (shadow) combining genotype probability cutoff 0.98, INFO cutoff 0.4, and genotype call rate cutoff 0.7 generates genotype concordance 0.98 and the best yield, 0.13, among the other combinations. The yield 0.13 is translated as 13% of 542,210 GWAS array SNPs imputed by 59,545 Exome array with mean concordance 0.98. This means doing genotype imputation with Exome array SNPs and applying post-imputation filters (genotype probability > 0.98, INFO > 0.4, and genotype call rate > 0.7) can predict genotypes of 70,533 GWAS array SNPs with mean concordance 0.98 which is 118% of the input SNPs for imputation.

Table 2: Exome array imputation candidate filter combinations with corresponding concordance and yield

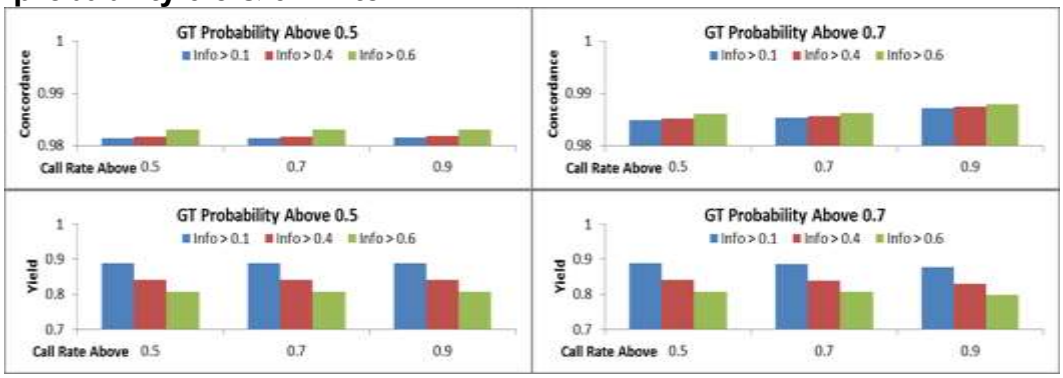
Genotype Probability	INFO	Call Rate	Mean Concordance	Yield
0.90	0.6	0.9	0.98	0.09
0.90	0.4	0.9	0.98	0.11
0.98	0.4	0.7	0.98	0.13
0.98	0.6	0.9	0.99	0.05
0.98	0.4	0.9	0.99	0.06
0.98	0.1	0.9	0.99	0.07
0.98	0.6	0.7	0.99	0.10

3.3.5 GWAS array imputation result optimization

Next, 41,909 Exome array SNPs imputed by GWAS array dataset is analyzed by the same filter combinations. Similarly, among Exome array SNPs imputed by GWAS array the trend of up raising concordance and diminishing yield is seen after more stringent genotype probability cutoff, INFO value cutoff, and genotype call rate cutoff are applied to the dataset. Figure 25 presents the concordance and yield plots after applying genotype probability cutoff 0.5 (left) and 0.7 (right). When applied genotype probability cutoff 0.5, INFO value cutoff 0.4, and genotype call rate cutoff 0.5 the concordance is 0.982 and yield is 0.841. After applying more stringent genotype probability cutoff 0.7, the genotype concordance increases to 0.985 and the yield stays at 0.841. Compared with GWAS

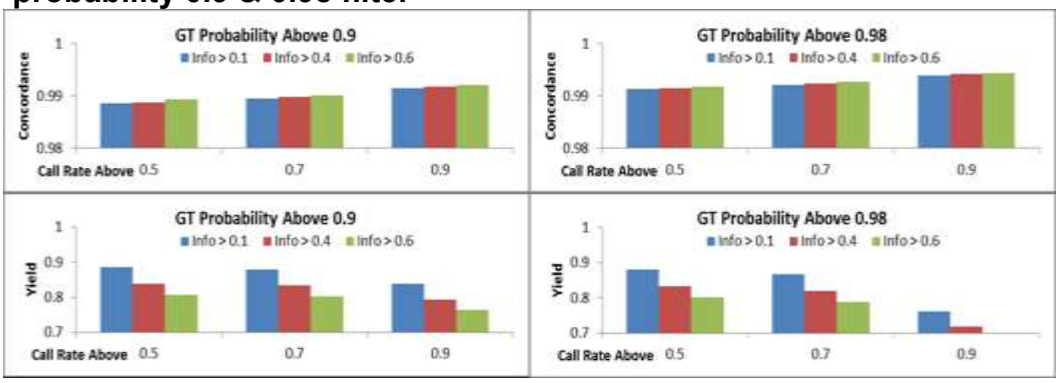
array SNPs imputed by Exome array dataset, these Exome array SNPs imputed by GWAS array have better concordance and yield to start with and even after applying same filter sets.

Figure 25: Genotype concordance and yield after applying genotype probability 0.5 & 0.7 filter



Similarly like the trend seen in the other dataset, when more stringent genotype probability cutoff 0.9 and 0.98 are applied to the filtering the concordance gets improved and, as expected, the yield gets smaller. Figure 26 presents the concordance and yield plots after applying genotype probability cutoffs of 0.9 (left) and 0.98 (right). When applied genotype probability cutoff 0.9, INFO value cutoff 0.4, and genotype call rate cutoff 0.5 the concordance is 0.989 and yield is 0.839. After applying more stringent genotype probability cutoff of 0.98, the genotype concordance increases to 0.991 and the yield stays at 0.834.

Figure 26: Genotype concordance and yield after applying genotype probability 0.9 & 0.98 filter



As pointed out earlier, SNPs imputed by GWAS array seem to have better concordance above 0.98 when applying the less stringent filter set (genotype probability cutoff 0.5, INFO value cutoff 0.1, and genotype call rate cutoff 0.5), it would be inappropriate to assume this lower level of filtering is adequate. Since Exome array aims to cover coding and rare variants instead of maximizing genome coverage as GWAS array design. With that and the phenomena seen in Exome array imputation result filtering, to be on the safe side, more stringent filter is recommended. Also considering the recommendation of IMPUTE2 developer and the impact of low genotype prediction probability and the low call rate implication in potential genotyping error and losing test power, any filter set including lower INFO value (0.1), lower low genotype probability (0.5 and 0.7), or lower call rate (0.5) should not be considered for optimized post-imputation filtering. Table 3 presents filter combinations and results after

excluding filter set with any of lower cutoff. Among them the one in shadow which includes genotype probability cutoff 0.98, INFO cutoff 0.4, and genotype call rate cutoff 0.7 generates the better concordance and yield combination, which translates into 34,345 imputed SNPs passing through the filtering.

Table 3: GWAS array imputation candidate filter combinations with corresponding concordance and yield

Genotype Probability	INFO	Call Rate	Mean Concordance	Yield
0.90	0.4	0.7	0.990	0.83
0.90	0.6	0.7	0.990	0.80
0.90	0.4	0.9	0.992	0.79
0.90	0.6	0.9	0.992	0.76
0.98	0.4	0.7	0.992	0.82
0.98	0.6	0.7	0.993	0.79
0.98	0.4	0.9	0.994	0.72
0.98	0.6	0.9	0.994	0.69

3.4 Discussion

The results in Figure 16 and 17 clearly show that the distribution of INFO metric is very different between GWAS imputation and Exome imputation datasets. As shown in Figure 16, Exome array imputation has higher frequency in lower INFO value SNPs and GWAS seems have higher frequency at the higher INFO end. In addition, the difference is

even more dramatic when only comparing results from array SNPs which are imputed by the other array (Figure 17). By comparing the genotypes of GWAS array laboratory assay and genotypes of GWAS array imputed by Exome array and the genotypes of Exome array laboratory assay and genotypes of Exome array imputed by GWAS array, one can identify other potential candidates for post-imputation filtering and come up with an optimized post-imputation filtering strategy. This concordance evaluation between laboratory genotype and computer genotype serves as a powerful tool for optimizing post-imputation filtering. After applying an INFO cutoff value of 0.4 and comparing the genotypes between laboratory assay and imputation, the results of Figure 18 and 19 demonstrate the inadequate power of applying INFO cutoff as the only post-imputation filter as recommended by the IMPUTE2 developer in removing poorly imputed SNPs and genotypes.

Other than the INFO quality metric, IMPUTE2 reports the imputed genotype probability of each of 3 possible genotypes, DD, Dd, and dd (D=Minor allele, d=Major allele). Therefore, it's an appropriate candidate to be included in post-imputation filtering. Another candidate is the in silico-genotyping call rate of imputed SNPs. It's common in laboratory genotyping process that a low call rate of genetic marker is indicating possible genotyping error. To that extent, the in silico-genotyping call rate may serve as a good candidate in post-imputation filtering as well. Figure

20 presents the call rate distribution difference between using GWAS array and Exome array for genotype imputation. This difference is most likely due to the genome coverage difference between these 2 types of array. The results in Figure 21 further demonstrate that applying a more stringent call rate filter improves the genotype concordance. This improvement is also seen among different INFO filtering groups. Therefore along with INFO and imputed genotype probability, the imputed SNP call rate is chosen as the measures for optimizing the post-imputation filtering.

Different cutoff values of each of 3 measures are tested for the mean genotype concordance and imputation yield. Three cutoff values of INFO (0.1, 0.4, 0.6), and 4 cutoff values of imputed genotype probability (0.5, 0.7, 0.9, 0.98), and 3 call rate cutoff values (0.5, 0.7, 0.9) are tested in 36 combinations of INFO, genotype probability, and call rate. After applying each filter combination, the mean genotype concordance and imputation yield are calculated and the one balancing between genotype concordance and yield is identified. The results from Exome array imputation and GWAS array imputation datasets, shown in Figure 22-24 and Figure 25-26 respectively, demonstrate the effectiveness of INFO, genotype probability, and call rate filtering. The mean genotype concordance is improved when more stringent cutoffs are applied. On the other hand, the yield of imputation drops when more poorly imputed SNPs and genotypes are filtered out.

Without question, the genotyping error plays a significant role over the Type I and Type II error rate in the downstream statistical analysis. It's been reported [75,84,85] that even as low as 2% genotyping error can have profound effect over the association analyses. To that extent, it is better to find the optimized post-imputation filtering with concordance greater or equal to 0.98 and a good yield. Among the GWAS array SNPs imputed by Exome and Exome array SNPs imputed by GWAS, the filter combination with INFO cutoff 0.4, genotype probability cutoff 0.98, and call rate cutoff 0.7 seems generating a balanced set of genotype concordance and SNP yield, 0.984 and 0.13, 0.992 and 0.82 respectively (Table 2 and Table 3). The difference over yield between these 2 datasets is actually a reflection of the content difference between these 2 array types which has been reported in INFO distribution (Figure 16 and Figure 17), imputed genotype concordance (Figure 18 and Figure 19), and the call rate (Figure 20).

Although the GWAS imputation has better averaged genome coverage and generates higher concordance and yield among imputed Exome array SNPs, the number of Exome array SNPs that got imputed (41,909) is actually much less than the other dataset, GWAS array SNPs imputed by Exome (542,210). Again, this is due to the fundamental difference between these 2 types of array. GWAS array is designed to survey the human genome with good genome coverage. On the other hand, Exome array is designed with a main focus on the coding sequences which is just

about 3% of the genome. In that sense, GWAS array has much more SNPs (~ 900K) designed as compared to Exome array (~ 320K), and therefore, disregarding the quality of imputed SNPs, Exome array imputation has more target SNPs to be imputed as compared to GWAS array imputation.

Indeed, the results shown here have indicated that SNPs imputed by Exome array have poorer quality in many aspects, INFO metric, genotype concordance, as well as call rate, when compared to SNPs imputed by GWAS array. The Exome SNPs imputed by GWAS array tend to have better quality due to the genome coverage of GWAS array and, probably, the dense coverage over the coding sequence of Exome array which, to a certain extent, may indicate some level of linkage disequilibrium among gene coding SNPs. Nevertheless it will be wrong to extend this genotype concordance level to the SNPs mapped on to the other 97% of genome. Although Exome array does not have the same power, in terms of genome coverage, as GWAS array in whole genome SNP imputation, it can actually serve as the “worst case” scenario for post-imputation filtering optimization. After all, not every whole genome SNP imputation study has funding for a validation SNP set to be used for filter optimization. Results here propose the use of filter combination with INFO cutoff 0.4, genotype probability cutoff 0.98, and call rate cutoff 0.7 for post-imputation filtering. This filter set increases the concordance of GWAS SNPs imputed by

Exome array from 0.775 (after applying the less stringent filter) to 0.984 and drops the yield from 0.84 to 0.13. However, the same filter set actually only increases the concordance of Exome SNPs imputed by GWAS array from 0.981 (after applying the less stringent filter) to 0.992 and drops the yield from 0.89 to 0.82. Therefore it is recommended to err on the safe side of having a more stringent filter set in place for various imputation designs.

Finally to summarize the conclusions in this analysis, it's shown that depending on the input content for whole genome SNP imputation, the quality of output imputed SNPs could be very different. Secondly, doing post-imputation filtering based on the SNP quality metric cutoff alone, INFO in this case, risks of leaving poorly imputed SNPs or genotypes for downstream statistical analyses. Finally, the combination of INFO cutoff 0.4, genotype probability cutoff 0.98, and call rate cutoff 0.7 is the optimized filter set for post-imputation filtering.

CHAPTER 4: COST EFFECTIVE DESIGN OF FAMILY-BASED IMPUTATION

Abstract

The idea of family-based genotyped imputation leveraging the rich information possessed in a large pedigree and shared genome segments among relatives has been proven feasible through recently published GIGI method. Rare variant alleles tend to aggregate and pass in families; therefore, the use of extended and complex pedigrees in searching the rare disease-causing variants is an effective study design. Owing to the power of family-based imputation, a cost-effective family study using high density SNP microarray or whole exome or genome sequencing now can be achieved. Given the high cost in high throughput sequencing and limited study budget, it is very important to identify a cost-effective approach and maximizes the study power through careful prioritization of the family members to be sequenced or genotyped. Another prospect is to evaluate an economically feasible way to transform legacy genome scan linkage studies into family genome-wide association studies through family-based imputation. This study is aimed to optimize the selection of family members to be included for high density genotyping whose genotypes can then be used for predicting un-assayed genotypes of other

family members. After evaluating various family-based imputation designs, at least 3 family members are required for high density genotyping in order to gain more power in inferring genotypes of unobserved members. The results also indicate 2 parents and one offspring for high density genotyping design has the greatest power over genotype imputation when compared with other designs like one parent and 2 offspring or 3 offspring only designs for example. In addition, the common genome-scan linkage marker, Short Tandem Repeat (STR), is found to have compatible imputation power like the SNP in predicting genotypes of other family members.

4.1 Introduction

Although pedigrees have been central to the discovery of genes underlying Mendelian traits, in recent years it's the GWA studies of large population-based samples that have been used to search for variants of complex traits based on the common disease common variant hypothesis. However, even though GWA studies have identified many candidate loci [63,69], common variants now seem to only explain a small percentage of heritability [86]. The missing heritability is suggested to be found among the rare but high disease risk variants [87], the hypothesis of common disease rare variant. This hypothesis has brought the use of large pedigrees back to focus once again, owing to the power of combining the high density genotyping and the extended pedigree information in identifying rare variants passing through generations. Compared with population-based GWA study, family studies with extended pedigrees have potential to identify longer IBD segments by examining a small subset of relatives. Therefore the family-based genotype imputation, undoubtedly, can further boost the power of family study in identifying rare variants.

Another aspect of family-based genotype imputation is to bring legacy genome scan linkage studies back to life. By leveraging the information carried by the genome scan markers, family-based genotype imputation can provide a new life to many legacy linkage studies through offering a cost effective study design of genotyping a subset of family members.

Instead of high density genotyping every family member, family-based imputation can impute the genotypes of unobserved family members by analyzing the genome scan genotypes from most family members and combining the high density marker genotypes from selected family members. This strategy is taking advantage of the power of previous genome scan studies and the power of family-based imputation. Unlike the population-based imputation, family imputation only needs a small number of high density genotyped family individuals and has more power and accuracy in predicting rare variant genotypes [88,89]. Therefore it is obvious that providing guidance in selecting which subjects to genotype for dense markers in a cost effective study design holds the key to the success in finding rare disease-causing variants.

Although a number of pedigree-based genotype imputation methods exist [32,90-92]. They tend to have limitation in either handling large pedigrees with many markers because of computational constraints or require high quality dense genotype data on subjects for whom we want to impute data and do not account for recombination events. Therefore, a computationally efficient method implemented in the Genotype Imputation Given Inheritance program (GIGI) for imputing dense genotypes in large pedigrees [93] is chosen for performing family-based genotype imputation. This Markov Chain Monte Carlo (MCMC)-based approach uses a sparse set of markers (Framework Markers) typed on most subjects plus dense

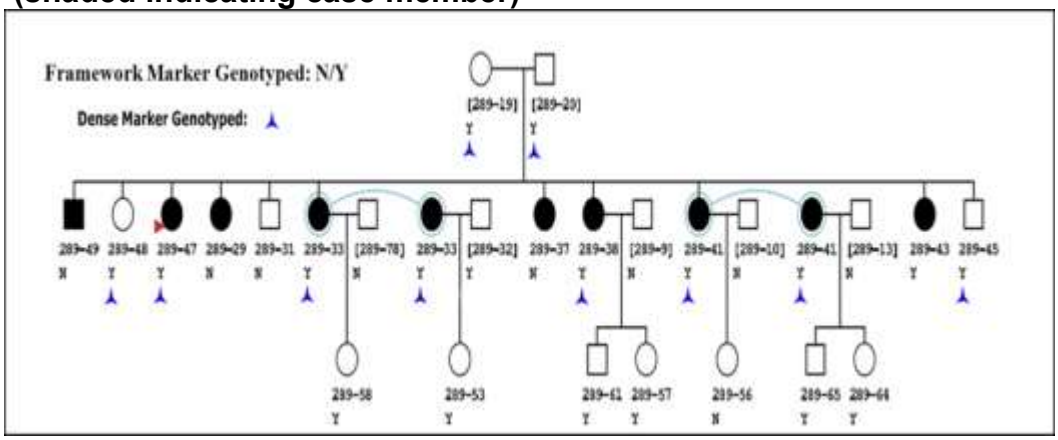
markers (Dense Markers) typed on a few subjects to impute genotypes of dense markers for unobserved family members. Although it's been reported that compared with other publically available software, GIGI seems to have better imputed genotype accuracy and performs very well when imputing genotypes for especially rare SNPs [93], it's not clear to what extent this excellent performance is affected by the number of members missing dense marker genotypes. For example, given an extended pedigree of 3 generations which is commonly seen is typical family study, among many family members, who must have framework marker genotypes and who must be dense marker genotyped in order to obtain the greatest power of the family-based imputation with smallest cost? Do all the members possess equal power in predicting genotypes of other members? A guideline for selecting effective family members for dense marker genotyping for imputation is critical to researchers in designing a cost effective family study with genotype imputation. With limited budget, it is not feasible to include all study subjects with DNA available for high density microarray genotyping or whole exome or genome sequencing. The main advantage of doing genotype imputation is precisely about cost saving. Unlike the population-based imputation, family-based imputation which leverages the pedigree information, does not need all study subjects to be genotyped in order to predict genotype of unobserved genetic markers. However, among family members, who is the most

effective member to be genotyped for imputation? In addition, many genome-scan linkage studies have used microsatellite markers or Short Tandem Repeats (STR) for dissecting the linkage between genomic regions and diseases. In terms of testing the association with the diseases or traits, STR is not the appropriate genetic marker for that purpose. However, it is a huge waste to discard the rich data collected in those genome-scan linkage studies. Therefore, if STR markers can be used as the framework markers in doing family-based imputation, many legacy genome-scan linkage studies can then be transformed into association study by a highly cost-effective method.

4.2 Materials and methods

In this study an African American family with extended pedigree structure (Figure 27) from an Early Onset Periodontitis (EOP) linkage study (Diehl et al., 1999) was selected for analyses. This family had been genotyped in 3 different sets of genetic markers including genome scan microsatellite markers, linkage mapping SNP set, and Illumina 2.5M SNP array. Thirteen family members were genotyped for genome scan markers and 16 family members were genotyped for linkage mapping SNP markers, and 8 family members were genotyped for Illumina HumanOmni2.5 SNP array.

Figure 27: A three generation African American EOP study family (shaded indicating case member)



The baseline benchmark of family-based imputation result came from imputation with all data available including dense marker genotypes of two 1st generation members (parents) and 6 members among 2nd generation siblings and framework marker genotypes of 16 members. Various genotyping scenarios are evaluated by the imputed genotype call rate and imputed genotype accuracy. Genotype accuracy is measured by comparing the imputed genotypes with array genotypes of members who was masked as either no dense marker genotype or no DNA available.

4.2.1 Study population

EOP linkage study population included 2,151 subjects from 300 families recruited in Virginia. DNA samples are available from 1,149 individuals, including 349 Caucasian (183 female, 166 male) and 800

African American (456 female, 344 male). DNA was extracted from whole blood using a standard protocol. Subjects were categorized according to the criteria described in the paper [94]. Among African American families, the family 289 (Figure 27) was selected for this family imputation analysis. As shown by the figure 27, 16 family members were assayed for the framework marker genotyping on Life Technologies SNPLEX platform [95], including founder parents 19 and 20, offspring 33, 47, 48, 38, 41, 43, and 45, grandchildren 58, 53, 61, 57, 65, and 64, and one of daughter-in-laws 32. In terms of dense marker genotyping, 8 members were selected for Illumina HumanOmni2.5 array genotyping, including founder parents 19 and 20, and their offspring 33, 47, 48, 38, 41, and 45. In addition, genotypes of 24 STR markers from the legacy EOP linkage study are also available for 12 family members (19, 20, 33, 38, 48, 45, 41, 47, 43, 37, 49, and 53) and used to evaluate the power of family-based imputation between different framework marker types. The study protocol was approved by the Institutional Review Board of University of Medicine and Dentistry of New Jersey, and written informed consent was obtained from each subject prior to taking part in the study.

4.2.2 Genotyping

Framework marker genotyping was carried out by genotyping Life Technologies SNP-based Linkage Mapping Set panel with approximately

3,500 SNPs by Life Technologies SNPLEX multiplex genotyping system [95] using capillary electrophoresis. Among the family members of Family 289, 16 members were selected for framework marker genotyping as shown in Figure 27. Following manufacturer standard operation procedure, 20 ng of DNA molecules from each study subject was used for framework marker genotyping. Dense marker genotyping was outsourced to Beckman Coulter and done on Illumina HumanOmni2.5 GWAS array platform. Among 16 members whom were genotyped for framework markers, 8 members were chosen for dense marker genotyping (Figure 27).

4.2.3 Genotype imputation and data analyses

Assuming no significant difference among chromosomes and to reduce the complexity of analysis, chromosome 1 genetic markers were selected for analysis. Various QC measures were applied to framework marker genotypes including Mendelian discrepancy. LD SNPs pruning was done on software package Golden Helix SVS (Golden Helix, http://www.goldenhelix.com/SNP_Variation/index.html) with R-square threshold of 0.8. Similarly dense marker genotype with Mendelian discrepancy was filtered out by SVS. Only dense markers with call rate equal to or greater than 95% are included for imputation. Following the user manual instruction of the family-based genotype imputation program, GIGI, framework marker genotypes were formatted into MORGAN

(http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml) acceptable format. The program gl_auto from MORGAN package was run on framework marker genotypes to compute the inheritance vectors. The GIGI program then took in the dense marker genotypes of observed family members to impute dense marker genotypes for unobserved members based on the inheritance vectors calculated by gl_auto. Genotype imputation procedures were run on a High Performance Computing 4 core Linux SMP machine provisioned with 128 G bytes of memory and 20 terabytes of user storage. Post-imputation genotype QC measures were applied to imputed genotypes. Various genotype imputation runs were conducted with different settings of family members with framework marker genotypes and/or dense marker genotypes. Genotype accuracy is measured by comparing the imputed genotypes with array genotypes of members who were masked as either no dense marker genotype or no DNA available.

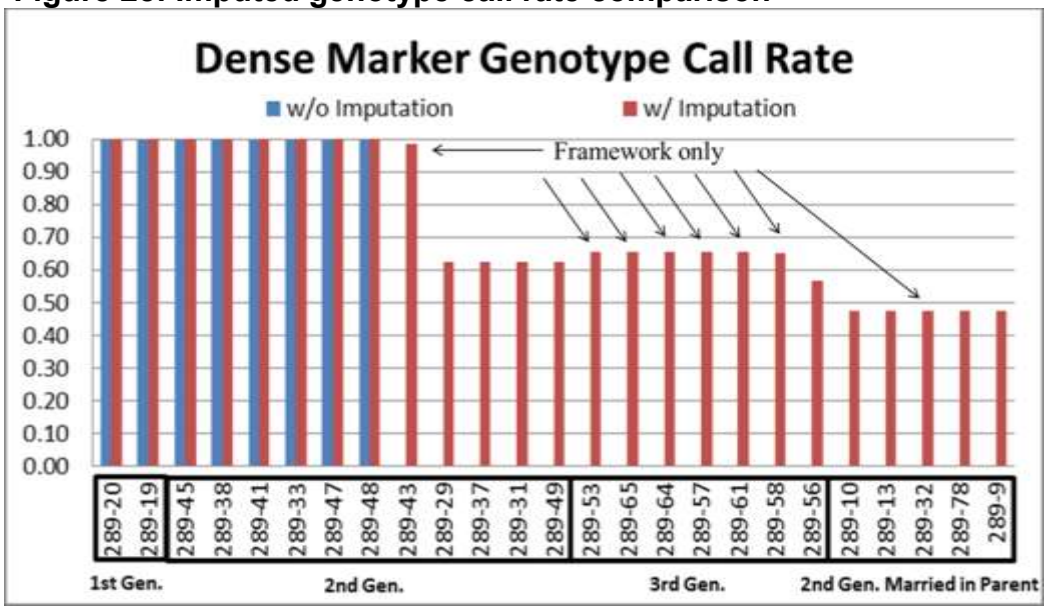
4.3 Results

To reduce the design complexity, only chromosome 1 markers are selected in the family-based imputation, including 299 framework (linkage mapping) markers and 161,991 dense (array) markers. Figure 28 shows the dense marker call rate before and after running family-based imputation.

4.3.1 The power of family-based imputation

This plot on Figure 28 shows the call rate of each family member for chr1 dense SNPs. The blue bar represents the run with no genotype imputation and the red bar represents the run with genotype imputation implemented using entire data set 16 family members genotyped for framework SNPs and 8 members genotyped for dense SNPs. The Y axis is the call rate which indicates the proportion of dense SNPs got called. The X axis indicates each family member.

Figure 28: Imputed genotype call rate comparison



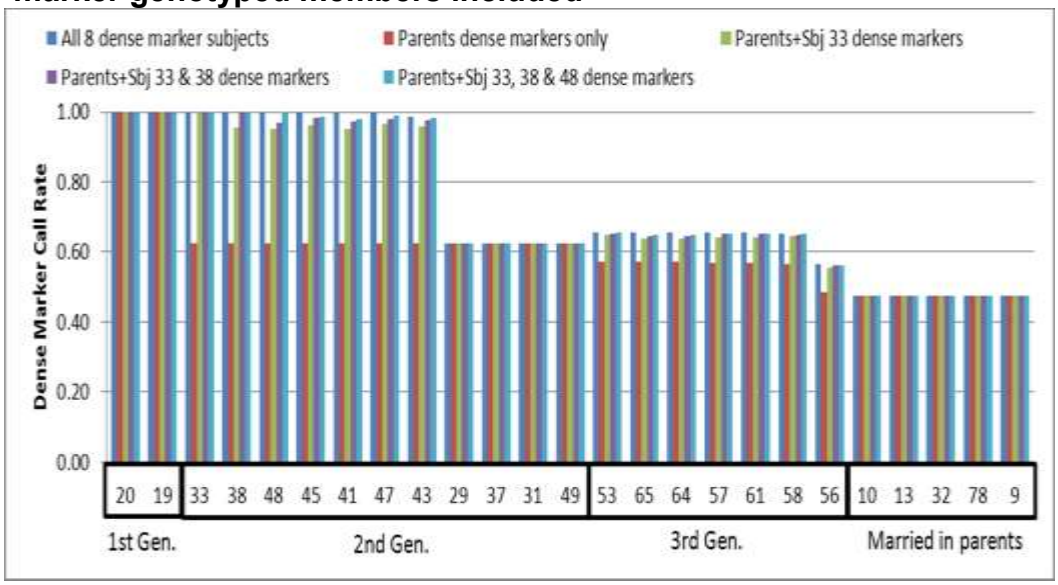
The first rectangle box includes the parents. The 2nd rectangle box includes offspring and among them 6 has both SNP sets genotyped and one (289-43) has only framework SNP set genotyped and 4 has no genotype. The 3rd rectangle box includes the 3rd generation grandchildren and first 6 has framework SNP set genotyped and one has no genotype. The 4th rectangle box includes married-in members and one (289-32) has framework SNP set genotyped. The results show that with the framework genotypes available, the Member 289-43 who has only framework marker genotypes actually has about 99% of dense markers imputed by the imputation. When considering the siblings who have no DNA available for genotyping, the call rate is around 0.63 after imputation which is still a significant gain in number of genotypes available for further analysis.

4.3.2 Number of offspring needed for imputation

Next, to test the performance of genotype imputation when only parents' dense SNP genotypes are available, the dense marker genotypes from 6 offspring are excluded from genotype imputation. All the framework SNP genotypes from 16 family members are included for imputation. This imputation design is aiming to answer the question of imputation merit when only both parents are chosen for dense marker genotyping. Following the parents dense marker genotypes only imputation, 3

additional imputation runs with different numbers of family members with dense marker genotypes included are implemented. The first one, “Parents+Sbj 33 dense markers”, is conducted with dense marker genotypes from both parents and one of the offspring, the Subject 33. Then another offspring, Subject 38, with dense marker genotypes is added for the second imputation run, “Parents+Sbj 33 &38 dense markers”. Finally the third imputation is conducted after adding dense marker genotypes from Subject 48 for imputation. The Figure 29 shows the imputed genotype call rate of dense marker for each family member.

Figure 29: Imputed genotype call rate at different numbers of dense marker genotyped members included



Shown in the Figure 29, the blue bar represents the imputation result of including dense marker genotypes of all 8 genotyped members. The

Subject 43 is the offspring who has only framework marker genotypes and 99% of dense markers got genotype imputed. Subject 56 is the only 3rd generation member who has no genotype for imputation and Subject 32 is the only married in parent who has the framework marker genotypes for imputation. The red bar represents the imputation result of including only dense marker genotypes of both parents for imputation. The call rate change is so dramatic that even with all framework marker genotypes included for imputation when only 2 parents are genotyped for dense markers, the imputed genotype call rate drops from 0.99 to 0.62 and 0.63 to 0.62 for no DNA subjects among 2nd generation members. Among 3rd generation members, the call rate drops from 0.66 to 0.57 for framework genotyped subjects and 0.57 to 0.49 for no DNA subjects. No call rate change is seen among married-in parents irrespective of the availability of framework marker genotype. However, this reduction can actually be rescued by adding dense marker genotypes of one 2nd generation members, in this example Subject 33. The green bar represents the imputation result of including dense marker genotypes of Subject 33 in addition to genotypes of both parents for imputation. Surprisingly, by adding the dense marker genotypes of just one offspring, the call rate is increased from 0.62 to 0.95, 0.95, 0.96, 0.95, 0.97, and 0.96 for 2nd generation framework genotyped subject 38, 48, 45, 41, 47, and 43 respectively. The call rate of 2nd generation no DNA members stays the

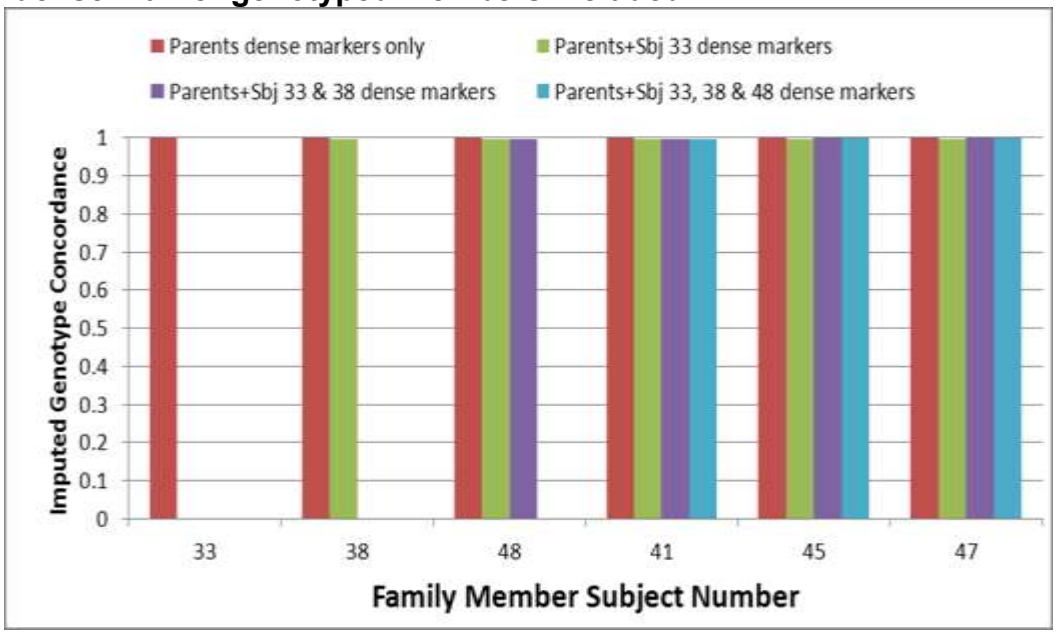
same at 0.62. Among 3rd generation members, the call rate drops from 0.57 to an average of 0.64 for framework genotyped subjects and 0.49 to 0.56 for no DNA subjects. Again, no call rate change is seen among married-in parents irrespective of the availability of framework marker genotype. The call rate improves after more 2nd generation members with dense marker genotypes are added for imputation although in a much smaller scale (purple bar: “Parents+Sbj 33 & 38 dense markers” & cyan bar: “Parents+Sbj 33, 38 & 48 dense markers”).

4.3.3 Imputed genotype concordance

Other than subject call rate of imputed dense markers, another measure to be used for validating the family-based imputation is the concordance of imputed SNPs with genotype calls based on lab assays. In this series of dense marker genotypes masked analysis, the concordance of imputed SNPs can be obtained by comparing the imputed genotype with lab assayed genotype which is masked out for imputation. Figure 30 displays the concordance between imputed and lab assayed genotypes of 2nd generation members whose dense marker genotypes are masked out during the imputation runs of “Parents dense markers only”, “Parents+Sbj 33 dense markers”, ”Parents+Sbj 33 & 38 dense markers”, and “Parents+Sbj 33, 38 & 48 dense markers” (red, green, purple, and cyan respectively). In “Parents dense markers only”

imputation, since only both parent dense marker genotypes are included all six masked out 2nd generation members have genotypes imputed. In “Parents+Sbj 33 dense markers” imputation, since only both parent and Subject 33 dense marker genotypes are included which leaves 5 masked out, 2nd generation members have imputed genotypes for concordance estimation. Therefore the ”Parents+Sbj 33 & 38 dense markers” imputation has 4 masked-out 2nd generation members, and the “Parents+Sbj 33, 38 & 48 dense markers” imputation has 3 masked out 2nd generation members who have imputed genotypes available to evaluate concordance.

Figure 30: Imputed genotype concordance at different numbers of dense marker genotyped members included



Although when only parents' dense marker genotypes are available for imputation, the imputed genotype call rate of framework-marker-genotyped-only subjects is only about 0.62 (Figure 29), the quality of imputed genotypes actually is very impressive with genotype concordance above 0.99. Even after adding more subjects with dense marker genotypes to the imputation which increases the imputed genotype call rate to above 0.95, the genotype concordance stays at the same level (> 0.99). Considering Figures 2, 3 & 4 results, it is clear that family-based imputation is a very powerful and cost-effective genotype imputation method which gives high yield and, yet, high accuracy.

4.3.4 Missing parent(s) in imputation

Next, to evaluate the importance of parents' genotypes in family-based genotype imputation, an imputation is conducted with all the genotypes of both parents excluded from imputation. In other words, it is mimicking the situation when parents are not available for genetic test which is commonly seen in late-onset diseases. In this run, framework marker genotypes of 14 members, including 7 2nd generation members, 1 married-in 2nd generation member, and 6 3rd generation members, and dense marker genotypes of 6 2nd generation members are used for genotype imputation. In addition, another imputation run mimicking the situation where only one parent, Subject 20 in this case, is available for

genotyping is conducted. Finally another imputation run is implemented with one parent available for genotyping and framework marker genotypes of 14 members, as mentioned before, but dense marker genotypes of only 4 2nd generation members.

Figure 31: Imputed genotype call rate when parents are not available

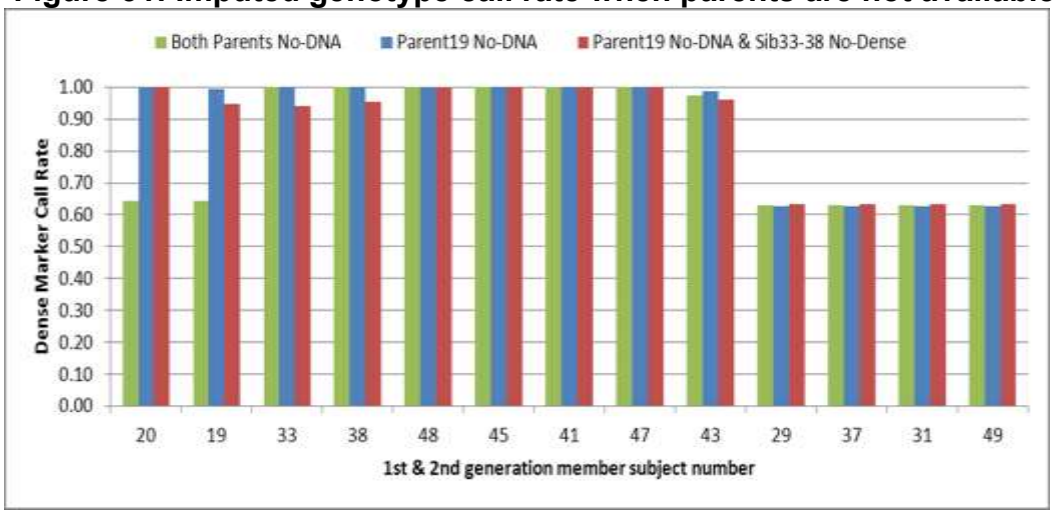
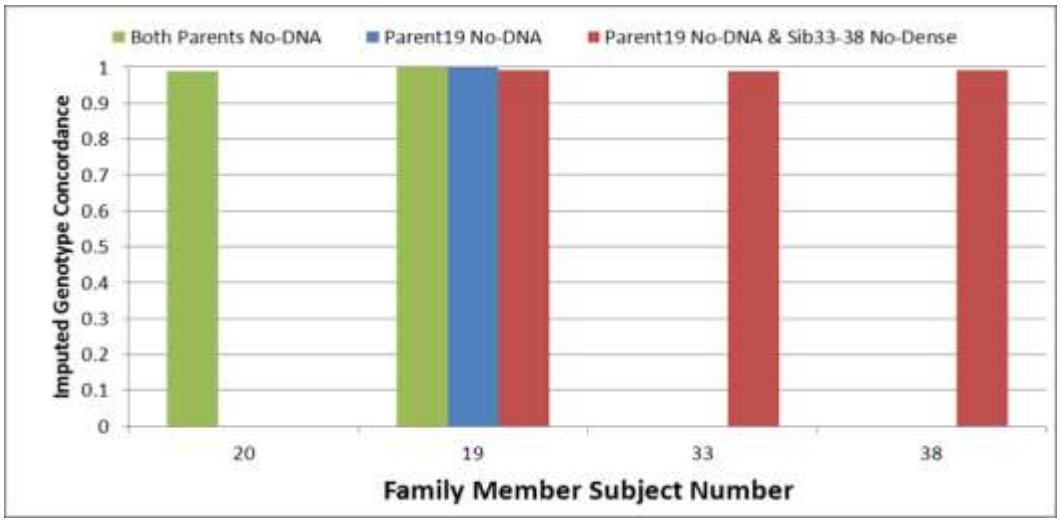


Figure 31 presents the results from these 3 imputation runs. The green bar group represents the results of runs with both parents not available for genotyping. The blue bar group represents the results of runs with only one parent available for genotyping, Subject 20 in this case. The red bar group represents the results of runs with one parent not available for genotyping, meaning no framework and dense marker genotypes, and 2 more offspring missing dense marker genotypes. Only 1st and 2nd generation parent-offspring members, including no DNA members, are

shown with results in this figure. The results indicate that even when parents are not available for genetic marker testing, imputation power can still be obtained by the rich genetic information provided by the other siblings. This is demonstrated by the imputed dense marker call rate, 0.98, of Subject 43 who has framework marker genotypes but no dense marker genotypes. Nevertheless, in this setting there seems to be no power in predicting genotypes of parents. The imputed genotype call rate of both parents is 0.64 which is close to the call rate of no DNA siblings (0.63). However, when one of parents is available for dense marker genotyping (blue bar group) the imputed genotype call rate of the no DNA parent is increased from 0.64 to 0.99. And the imputed genotype call rate of Subject 43 increased from 0.98 to 0.99. When 2 siblings (Subjects 33 and 38) are dropped for dense marker genotyping, the imputed genotype call rate of the no DNA parent drops to 0.95 and 0.96, 0.94, 0.95 for Subjects 43, 33, and 38 respectively. In all 3 runs, there is no call rate change among 2nd generation members who have no DNA for genotyping (average call rate: 0.63).

In these 3 runs, the genotype concordance can be estimated for members masked as no dense marker genotypes are available for imputation. Figure 32 presents the genotype concordance report of these 3 runs.

Figure 32: Imputed genotype concordance when parents are not available



When both parents are not available for genotyping, “Both Parents No-DNA” run (green), with framework marker genotypes from 7 offspring and dense marker genotypes from 6 offspring, 64% of dense markers are imputed with concordance of 0.99 for both parents. When only one parent is not available for genotyping, “Parent19 No-DNA” run (blue), with framework and dense marker genotypes from one parent and the framework marker genotypes from 7 offspring and dense marker genotypes from 6 offspring the missing genotype parent actually has 99% of dense markers imputed with concordance of 0.99. Finally, when one parent is not available for genotyping and 2 offspring are missing dense marker genotypes, “Parent19 No-DNA & Sib33-38 No-Dense” run (red), about 95% of dense markers can be imputed with concordance of 0.99.

4.3.5 Members needed for dense marker genotyping

So far the results seem to indicate that when implementing a family-based imputation, at least one offspring and one parent for dense marker genotyping are sufficient for better yield in imputing parent and offspring genotypes. However, in this experimental setting, there are 16 framework marker genotyped family members and 8 dense marker genotyped members. Although results show that adding one densely genotyped offspring along with both parents dense genotypes can achieve great power in imputing genotypes for other offspring, it's under the condition that all the framework marker genotypes are included. When testing the missing-parent design, similarly, all the framework genotypes are included and at least 5 densely genotyped members are included. Actually the results show that the imputation yield starts diminishing when less number of densely genotyped offspring are available for imputation. Also the framework marker genotype availability status among grandchildren and married-in parents seems to have no significant effect over the imputation yield and accuracy. In the real world situation, most study families do not have similar amount of data for imputation. Therefore, more imputation analyses focusing on smaller realistic family designs are carried out. Table 4 presents the summary results of 13 imputation designs (A ~ M).

Table 4: Imputed genotype call rate and concordance of various smaller family designs

	Total Dense	Parent		Offspring		Mother		Father		Framework Offspring		No-DNA Offspring
Design	Subject Count	Framework	Dense	Framework	Dense	Call Rate	Concordance	Call Rate	Concordance	Call Rate	Concordance	Call Rate
A	4	2	2	2	2							0.63
B	3	2	2	2	1					0.96	1.00	0.62
C	3	2	1	3	2	0.92	0.98			0.91	0.99	0.63
D	3	2	1	2	2	0.92	0.98					0.63
E	3	1	1	2	2	0.91	0.98					0.63
F	3	0	0	4	3	0.63	0.99	0.63	0.97	0.85	0.98	0.61
G	2	2	0	2	2	0.72	0.98	0.61	0.95			0.59
H	2	2	0	3	2	0.72	0.98	0.61	0.95	0.67	0.96	0.59
I	2	2	1	2	1	0.64	0.93			0.70	0.94	0.64
J	2	1	1	2	1	0.64	0.93			0.70	0.94	0.64
K	2	0	0	2	2	0.61	0.99	0.61	0.94			0.59
L	2	0	0	3	2	0.61	0.99	0.62	0.94	0.67	0.96	0.59
M	1	2	0	2	1	0.57	0.93	0.57	0.92	0.61	0.93	0.57

Design A is a 2 generation family including 2 parents, 2 offspring who are both framework and dense marker genotyped and one no-DNA offspring. Since the dense marker genotyped subject count is 4 in this 5 member family, only the no-DNA offspring is imputed with call rate 0.63 which is similar with previous analysis result. Design B is a 5 member family which includes both parents and one offspring genotyped for framework and dense marker, one offspring genotyped for framework marker only, and one no-DNA offspring. In this design, the framework marker genotyped only offspring has imputation call rate 0.96 and concordance 1. The no-DNA offspring has call rate 0.62. Design C represents a 6 member family with both parents and 3 offspring genotyped for framework markers and one no-DNA offspring. Only the father and 2 of framework genotyped offspring are genotyped for dense markers. Therefore, the mother, one framework offspring, and the no-DNA offspring

are imputed for dense marker genotypes with call rate 0.92, 0.91, 0.63 respectively. The imputed genotype concordance is 0.98, 0.99 for the mother and the framework offspring respectively. Design D is similar with Design C, except it has only 5 members and no framework marker genotyped only offspring. Therefore, only the mother and the no-DNA offspring have imputed genotypes. Similarly, the mother is found with call rate 0.92 and genotype concordance 0.98 and call rate 0.63 for the no-DNA offspring. Between Design C and D, it seems adding framework marker genotypes from one more offspring does not make a difference in imputing dense marker genotypes for the parent. Design E basically is the same as Design D, except, instead of having framework marker genotyped only for the mother, Design E has the mother masked as no-DNA. Nevertheless, under this design the mother is still found with call rate 0.91 and genotype concordance 0.98 and call rate 0.63 for the no-DNA offspring. Between Design D and E results, it indicates that having mother framework marker genotypes does not gain more power in imputing her dense marker genotypes. Design F is a 7 member family with both parents having no-DNA, one no-DNA offspring, and 4 framework genotyped offspring and 3 of them are dense marker genotyped. Therefore, the parents, the no-DNA offspring, and the framework marker genotyped only offspring are imputed for dense marker genotypes. Compared with Design E, this design has same number (3) of dense

marker genotyped members and 2 more framework genotyped offspring for imputation. Nevertheless, it has no power in imputing parent genotypes, call rate 0.63, 0.63 and concordance 0.99, 0.97 for mother and father respectively although it has some power in imputing framework genotyped only offspring with call rate 0.85, concordance 0.98 and call rate 0.61 for no-DNA offspring.

Designs from G to L are the designs which in total have only 2 dense marker genotyped members for imputation. Design G is a 5 member family with parents available for only framework marker genotypes, 2 offspring available for both framework and dense marker genotypes, and one no-DNA offspring. The mother has 72% of dense markers imputed with concordance 0.98. The father is imputed with 61% dense markers and concordance 0.95. The no-DNA offspring has 59% dense markers imputed. Similarly, Design H is like Design G, except, in addition to the 5 members mentioned it has one more framework marker genotyped only offspring for imputation. However, this addition seems to add no power in imputing dense marker genotyped for parents, no-DNA offspring. It generates the same call rate and concordance for them and call rate 0.67 and concordance 0.96 for the framework marker genotyped only offspring. Both Design I and J are 5 member family design with the father and one offspring available for framework and dense marker genotypes, and one framework marker genotyped only offspring. The only difference is, in

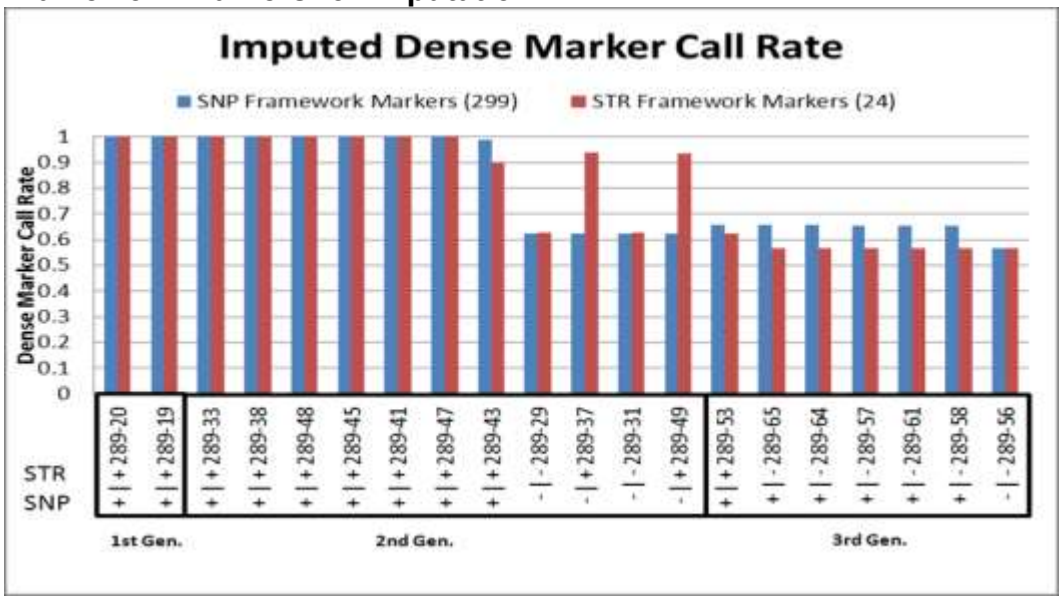
Design I, the mother is available for framework marker genotypes, and in Design J, the mother is not available for both framework and dense marker genotypes. Nevertheless, the addition of the mother's framework marker genotypes does not seem to add power to the imputation. Both designs are found with call rate 0.64 and concordance 0.93 for the mother, and call rate 0.70 and concordance 0.94 for the framework marker genotyped only offspring and call rate 0.64 for no-DNA offspring. Designs K and L are almost the same with Designs G and H, respectively, except the parents are no-DNA member in Designs K and L. When none of parents are available for dense marker genotypes, having no framework marker genotypes from both parents seems to affect the prediction of mother genotypes but not for the father, framework marker genotyped only offspring, and no-DNA offspring. In Designs K and L, the mother call rate and genotype concordance are 0.61 and 0.99 as compared to 0.72 and 0.98 from Designs G and H respectively. All 4 designs have the father found with call rate about 0.61 and concordance 0.94 for the father and call rate 0.67 and concordance 0.96 for the framework marker genotyped only offspring, and call rate 0.59 for the no-DNA offspring. Finally, the Design M is the 5 member family design with only one dense marker genotyped member for imputation as well as framework marker genotyped only parents, one framework marker genotyped only offspring, and one no-DNA offspring. Although 4 of 5 members are available for framework

marker genotypes, with only one dense marker genotyped offspring, this design barely has any power in genotype imputation. Under this design, the mother has call rate 0.57 and concordance 0.93, similarly the father has call rate 0.57 and concordance 0.92, the framework marker genotyped only offspring has call rate 0.61 and concordance 0.93, and the no-DNA offspring has call rate 0.57.

4.3.6 STRs used as framework marker for imputation

So far, the family-based genotype imputation using SNP as the framework marker has proven to be a very effective and cost-saving method to increase family study power in both number of imputed genotypes and imputed genotype accuracy.

Figure 33: Comparison between using SNP and STR as the framework markers for imputation



Before SNP was popular to be used in genetic study, mainly association but some linkage studies, STR Linkage Mapping set had been employed in many linkage studies. Ever since the focus of genetic study moved from linkage to association analysis, the STR genotypes of those legacy linkage studies had been kept in the closet. If STR markers can serve as the framework marker in family-based genotype imputation, one can bring those legacy linkage studies back to life with minimum cost. To that extent, a run of imputation with genotypes of 24 chromosome 1 STR markers from 12 family members as the framework marker genotypes was done. Figure 33 shows the call rate comparison between using 299 SNPs (blue) and using 24 STRs (red) as the framework markers for imputation.

Subjects who have framework genotypes for imputation are marked by a “+” under the subject ID for each SNP and/or STR. A “+” on the SNP row, indicates the subject has genotypes of SNP framework markers for imputation. Thus a “+” on the STR row, indicates the subject has genotypes of STR framework markers for imputation. As mentioned earlier, only 8 members are genotyped for dense markers, including 2 1st generation parents and 6 2nd generation offspring. Notice that the 2nd generation married-in members are shown in Figure 7 and the 2nd generation members, Subjects 37 and 49, are actually available for STR framework marker genotyping but not for SNP framework marker

genotyping. Thus these 2 siblings along with Subject 43 serve as the indicator of imputation power when using STR as the framework markers. When using 299 LMS SNPs as the framework markers for imputation, Subject 43 has about 99% of dense markers imputed. On the other hand, when 24 STR markers being used as the framework marker for imputation, Subject 43 has around 90% dense markers imputed and 94% for Subjects 37 and 49. Among the 3rd generation members, Subject 53 has genotypes from both SNP and STR framework markers for imputation. Under the imputation with SNP framework markers, her imputed genotype call rate is 0.66. When STR framework markers are used, instead, the call rate changes to 0.62. Overall, as compared with SNP framework marker run, with 10 times less number of markers the STR framework markers can actually yield comparable result in genotype imputation.

4.4 Discussion

In recent years, genotype imputation has been routinely employed in Genome-Wide Association Studies (GWASs) leveraging the shared genome segments among individuals in the hope of boosting the power of assayed genetic markers at the expense of computational power. Nevertheless, most of the genotype imputations are performed in the population-based approach as compared to family-based approach.

Although the family-based approach has many advantages over the population-based approach, due to the lack of effective method in handling large complex extended pedigree its power had not been fully unearthed despite the rich pedigree information. Because of a smaller number of meiosis recombination events happening between family relatives, the IBD segments are much longer as compared with unrelated individuals which in return provides more power in predicting unobserved genotypes with excellent yield and accuracy from the long stretch of shared genome segments. Many GWASs of large population-based samples have been carried out to search for variants responsible for complex traits according to common-disease-common-variant hypothesis and reported in the NHGRI GWAS Catalog [63,69]. Nevertheless most of common variants appear to explain only a small fraction of heritability [86,87]. In recent years, the hypothesis of rare variants explaining the missing heritability starts emerging which is bringing the use of large pedigrees back to life. Rare variant alleles tend to aggregate and pass within families as compared to general population. The large pedigree next-generation sequencing study is a particularly efficient design for identifying rare variants that affect disease risk. To that extent, family-based genotype imputation is really an ideal method in boosting family study power and, yet, minimizing the cost. On the other hand, the idea of using “framework” markers to impute genotypes for “dense” markers has shed light on those

legacy genome scan linkage studies which mostly use STRs in finding the linkage between disease trait and genome regions. If STRs can be used as the framework marker and work as good as SNP markers in genotype imputation then it's a big cost saving and resurgence in bringing back those legacy linkage studies to life with a cost-effective study design.

The combination of MORGAN and GIGI programs shed light on the possibility of doing family-based imputation on large and complex pedigrees. These Markov Chain Monte Carlo (MCMC)-based methods enable feasible and accurate analyses of large pedigrees with many markers on large pedigrees. With family-based imputation based on just a number of family members to be assayed on high density array genotyping platforms or next-generation sequencing technologies, almost every member can be imputed for dense marker genotypes. This power of imputation has been demonstrated by the results of Figure 28. With genotypes of 299 LMS SNPs from 16 family members and 8 GWAS array chromosome 1 SNPs genotyped members, the family-based imputation method can actually predict genotypes of 99% of 161,991 GWAS array chromosome 1 SNPs. However in a real world situation, no study will have the luxury in genotyping so many family members by high density GWAS array or whole genome or exome sequencing. To that extent, a guide in identifying the key family members which maximize the power of imputation and, yet, minimize the cost of dense marker genotyping is

needed. Results shown in Figure 29 clearly indicate the need of dense marker genotypes from at least one offspring along with parents' dense marker genotypes for imputation with decent power. The imputed genotype call rate dramatically increases from 0.62 to 0.96 after adding just one offspring with dense marker genotypes for imputation. Although the imputed genotype call rate gets improved after adding more dense marker genotyped offspring for imputation, the gain is trivial when compared to the gain from adding one dense marker genotyped offspring. This gain of imputation power is not only seen in the number of dense markers got imputed but also in the quality of imputed genotypes. As shown in the Figure 30, all the imputation designs have imputed genotype concordance above 0.99 which, indeed, proves the power of this family-based imputation method.

The analyses have shown that at least one dense marker genotyped offspring is needed with parents' dense marker genotypes for imputation. However, one of common problem happened, especially in late-onset disease studies, is the lack of one or both of parents DNA for genotyping. Through different imputation designs, the results from Figure 31 seem to indicate the need of dense marker genotypes of one parent in maintaining power in imputing genotypes for the missing parent and missing dense marker genotypes of both of parents seems not having significant impact in imputing genotypes of framework marker genotyped offspring. In

addition, the imputed genotype concordance results from Figure 32, again, show the robustness of this family-based imputation method. Even when both parents are not available for dense marker genotyping, the imputed genotype concordance is about 0.99 for both parents. Nevertheless, it would be wrong to conclude that only one dense marker genotyped offspring is needed for imputation or parent dense marker genotypes are not needed for family-based imputation. Because the imputation power demonstrated so far, most likely, is due to large number of dense marker genotypes offspring, since the married-in parents' and 3rd generation members' framework marker genotypes seem not having significant impact on the imputation results. Therefore imputation designs which focus on smaller family setting are carried out and the results are present in Table 4. The results from 13 small size family imputation designs indicate that 3 dense marker genotyped family members are required for having an efficient and cost-effective family-based imputation with high yield and accuracy. Moreover, when both of parents having dense marker genotypes for imputation, it generates the best imputation results with imputed genotypes call rate 0.96 and concordance 1 for framework genotyped only offspring. If only one of parents is available with dense marker genotypes, the imputation power is still impressive given imputed genotypes call rate 0.91 and concordance 0.99 for framework genotyped only offspring and call rate 0.91 and concordance 0.98 for the genotype

missing parent. However, the power starts diminishing when none of parents is available for dense marker genotyping. With only 2 dense marker genotyped offspring, the imputed genotypes call rate and concordance drop to 0.67 and 0.96 respectively for framework genotyped only offspring as well as call rate 0.61 and concordance 0.96 for genotype missing parents. Nevertheless, in this case with one more dense marker genotyped offspring the imputed genotypes call rate and concordance can actually be increased to 0.85 and 0.98 respectively for framework genotyped only offspring and 0.63 and 0.98 as imputed genotypes call rate and concordance respectively for genotype missing parents. In a short summary, when choosing family members for GWAS array genotyping or whole genome or exome sequencing with the consideration of family-based imputation in mind, it is recommended to have at least 3 members for dense marker genotyping. It's better to include both parents and one offspring for dense marker genotyping in order to obtain better imputation power. Otherwise, at least one parent better to be available along with 2 offspring for dense marker genotyping. If none of parents are available then adding more offspring can only improve the power in imputing genotypes for other offspring. This is great news for late-onset disease family studies. With only one parent available for laboratory assay, it is still possible to archive impressive imputation power through family-based imputation.

Finally, most legacy linkage studies used multi-allele informative STR markers as the tool in surveying the linkage between trait and genome regions. It will be interested to see if the STR markers can be a good candidate for framework marker in family-based imputation. If that is the case then it will a big cost-saving study design to give power to those legacy linkage studies in dissecting the association between the traits and genetic variants. The imputation result comparison between using 299 SNPs as the framework markers and 24 STRs as the framework markers in Figure 33 shows that, indeed, STR marker can be a good candidate for the framework marker in this family-based imputation method. With 10 times less than SNPs in number of markers, under the similar setting, 24 multi-allele STR markers can still impute genotypes for 94% of dense markers for offspring with framework marker genotypes available for imputation.

CHAPTER 5: STUDY CONCLUSIONS

In recent years, the field of genetics has been evolving at a speed that has never been seen in its history. Owing to the advent of new technologies that have brought about enormous changes in terms of both the size and the complexity of genetic research, the focus has shifted from rare Mendelian disorders to more complex common diseases. The type of studies has also been moving away from mostly single-center linkage studies to more collaborative multi-center association studies. Although the cost of the whole genome sequencing is still too high for conducting genetic studies of thousands of subjects, it is expected to go down substantially in the near future. As one can imagine, the enormous size and the complexity of data in the collaborative environment have also brought about great challenges for genetic study management. Although the practical benefits of employing informatics tools with database implementation have been long been recognized in genomic research for uses such as microarray genotyping data management, high throughput sequencing data management and gene repositories, the need for comprehensive tools that integrate genotype and phenotype data with public annotation information for genome-wide genetic studies has not been addressed.

In this dissertation, a Genetic Study Database (GSD) system is designed and developed to address this need and ease the data management burden of mapping genetic variants, especially for complex disorders. With sophisticated relational database design, a comprehensive user interface, and a wide range of data management utilities it provides a powerful tool for genetic linkage and/or association research. GSD simplifies merging genotype data from various assay systems with subject and genetic marker annotation information, pedigree, phenotype, and risk factor data. Specifically, GSD's design makes it suitable for large-scale, multiple-center projects that have become more common recently for studies of association between common and/or rare variants and complex disease etiology. GSD implementation significantly eases the burden of managing the large volumes of data generated by such studies.

The ultimate goal in a genetic association study is to thoroughly survey the complete human genome to dissect association between genetic variants and diseases. However, before the cost of whole-genome sequencing falls to a range where this becomes feasible, the use of genotype imputation will be an essential tool for boosting study power. Even if sequencing cost only pennies, we would still need imputation to infer data for family members who do not have DNA available either because they were already deceased at the start of the study or were not available to participate. Genotype imputation has been routinely

performed across the whole genome in genome-wide association studies or in more targeted genomic regions for fine-mapping studies. In contrast to determining genotypes by assaying DNA in the laboratory, genotype imputation involves “in silico” genotyping by predicting the genotypes for SNPs which are not directly assayed in the study sample. The predicted genotypes can then be included in testing for association along with laboratory assayed genotypes and boost up study power by increasing the number of genotypes as well as the number of SNPs. This increases the ability to discover or fine-map causal variants and facilitates meta-analysis in merging data from various genotyping platforms. Nevertheless, owing to the uncertainty in predicting genotypes for SNPs located in regions that lack high enough levels of linkage disequilibrium, it’s important to filter out poorly imputed SNP genotypes after imputation (i.e., genotype imputation error). If the imputation quality is low at a SNP, it is better to remove such SNPs before association testing is performed. Studies have found genotyping error as low as 2% can reduce power of association tests, especially when dealing with rare genetic variants that are especially sensitive to such error. Most imputation software programs recommend that once the imputation has been carried out, in the absence of true genotypes to compare with imputed genotypes, it is important to assess the quality metrics provided by the programs for the imputed SNPs. However, it is not clear to what extent the quality metrics are effective in

filtering poorly imputed SNPs. Moreover, it is also not clear that if filtering based on the imputed SNP quality metrics alone is adequate and robust when using different types of microarrays such as those that cover common variants for the whole genome versus exome arrays that cover only the protein coding regions of the genome.

In this study, by comparing genotypes from laboratory assays with genotypes obtained from imputation for the same subjects, it's shown that the imputed SNP quality metrics do not adequately filter out data with low accuracy. Furthermore, different types of SNP microarrays perform quite differently in terms of the accuracy of their imputation. Therefore, additional quality measures have been tested with the goal of obtaining higher imputation accuracy and while maximizing imputation yield. In most genotype imputation studies, one does not have the true genotypes of imputed SNPs to validate the genotypes imputed. Since the imputation quality and accuracy may vary depending on the contents of the imputation input dataset, it is better to have a stringent post-imputation filtering strategy in place to reduce the error for downstream statistical analyses. After evaluating various combinations of measures for filtering, this study suggests a single set of filters combining the imputed genotype probability (0.98), imputed SNP quality metric (0.4), and imputed SNP call rate (0.7) to be used for general post-imputation filtering in population-based imputation studies.

Although in recent years genetic studies have shifted from family-based linkage approaches to population-based association methods, the more recent focus on the role of rare variants in common diseases has caused a resurgence of interest in using large pedigrees for detecting the association of rare variants with common diseases. Since rare variants tend to aggregate and pass through a pedigree, family-based genotype imputation is especially effective in predicting missing genotypes of rare variants for close relatives. Utilizing the pedigree information to identify large segments of shared genome among relatives, family-based genotype imputation can be a very cost-effective method compared to population-based approaches. Family-based imputation also has the unique advantage which population-based method does not possess in its ability to impute genotypes for individuals who have no DNA available for genotyping. Additionally, the family-based method has the merit in transforming existing legacy genome-scan linkage studies into low-cost high throughput sequencing association studies. Because of the high cost of high density genotyping, whether GWAS array or whole-genome sequencing, it is important to carefully choose which family subjects to assay by genotyping or sequencing in the laboratory. Through testing various family-based imputation designs based on real data, this study has demonstrated the need for dense genotyping of at least, 3 family members in order to obtain high power for imputation. When both parents

and at least one offspring are available for dense genotyping, genotypes of 96% of dense markers can be imputed with high accuracy for offspring who have only framework marker genotypes. In the situation where only one parent is available for dense genotyping and imputation, 91% of dense markers can be imputed with high accuracy for both other offspring who have only framework markers and the missing parent. Even when both parents are unavailable, 85% of dense markers can be imputed with accurate genotypes for siblings having only framework markers as well as 63% of the dense marker genotypes for both unobserved parents. Finally, the results also demonstrate that STR markers from existing genome-scan studies can be leveraged to allow genotype imputation of dense markers on many individuals when these existing STR marker genotypes are coupled with dense markers typed on parents or siblings. This finding demonstrates the potential for transforming many legacy genome-scan linkage studies into powerful family-based association studies with high density sequencing or genotyping data obtained at a small fraction of the cost if every subject had to be assayed in the laboratory for the high density genetic variants.

REFERENCES

1. Riordan JR, Rommens JM, Kerem B, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 1989;245:1066-73.

2. MacDonald ME, Novelletto A, Lin C, et al. The Huntington's disease candidate region exhibits many different haplotypes. Nature genetics 1992;1:99-103.

3. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics 2005;6:95-108.

4. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nature reviews Genetics 2005;6:109-18.

5. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881-5.

6. Diabetes Genetics Initiative of Broad Institute of H, Mit LU, Novartis Institutes of BioMedical R, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331-6.

7. Hara K, Fujita H, Johnson TA, et al. Genome-wide association study identifies three novel loci for type 2 diabetes. Human molecular genetics 2014;23:239-46.

8. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nature genetics 2007;39:207-11.

9. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature genetics 2007;39:596-604.

10. Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 1997;278:1580-1.

11. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS genetics 2009;5:e1000477.

12. Ball RD. Designing a GWAS: power, sample size, and data structure. Methods in molecular biology 2013;1019:37-98.

13. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 2002;30:97-101.

14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 2007;81:559-75.

15. de la Chapelle A. Disease gene mapping in isolated human populations: the example of Finland. Journal of medical genetics 1993;30:857-65.

16. de la Chapelle A, Wright FA. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proceedings of the National Academy of Sciences of the United States of America 1998;95:12416-23.

17. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037-48.

18. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews Genetics 2010;11:499-511.

19. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annual review of genomics and human genetics 2009;10:387-406.

20. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Human genetics 2008;124:439-50.

21. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics 2010;26:445-55.

22. Neale BM. Introduction to linkage disequilibrium, the HapMap, and imputation. Cold Spring Harbor protocols 2010;2010:pdb top74.

23. International HapMap C, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.

24. International HapMap C. A haplotype map of the human genome. Nature 2005;437:1299-320.

25. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.

26. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics 2007;81:1084-97.

27. Greenspan G, Geiger D. Model-based inference of haplotype block variation. Journal of computational biology : a journal of computational molecular cell biology 2004;11:493-504.

28. Kimmel G, Shamir R. GERBIL: Genotype resolution and block identification using likelihood. Proceedings of the National Academy of Sciences of the United States of America 2005;102:158-62.

29. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. American journal of human genetics 2001;68:978-89.

30. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. American journal of human genetics 2003;73:1162-9.

31. Latz E, Verma A, Visintin A, et al. Ligand-induced conformational changes allosterically activate Toll-like receptor 9. Nature immunology 2007;8:772-9.

32. Burdick JT, Chen WM, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. Nature genetics 2006;38:1002-4.

33. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 2007;39:906-13.

34. Li Y DJ, Abecasis GR. MACH 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. American journal of human genetics 2006;79:S 2290.

35. Browning SR. Multilocus association mapping using variable-length Markov chains. American journal of human genetics 2006;78:903-13.

36. Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). Human mutation 2011;32:564-7.

37. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 2004;305:869-72.

38. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics 2008;40:695-701.

39. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. Clinical genetics 2011;79:199-206.

40. Sanna S, Li B, Mulas A, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS genetics 2011;7:e1002198.

41. Leigh SE, Foster AH, Whittall RA, Hubbart CS, Humphries SE. Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database. Annals of human genetics 2008;72:485-98.

42. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics 2008;9:356-69.

43. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics 2003;33 Suppl:228-37.

44. Mayeux R. Mapping the new frontier: complex genetic disorders. The Journal of clinical investigation 2005;115:1404-7.

45. Li JL, Deng H, Lai DB, et al. Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. Genome research 2001;11:1304-14.

46. Gillanders EM, Masiello A, Gildea D, et al. GeneLink: a database to facilitate genetic studies of complex traits. BMC genomics 2004;5:81.

47. Monnier S, Cox DG, Albion T, Canzian F. T.I.M.S: TaqMan Information Management System, tools to organize data flow in a genotyping laboratory. BMC bioinformatics 2005;6:246.

48. Orro A, Guffanti G, Salvi E, Macciardi F, Milanesi L. SNPLims: a data management system for genome wide association studies. BMC bioinformatics 2008;9 Suppl 2:S13.

49. Zhao LJ, Li MX, Guo YF, Xu FH, Li JL, Deng HW. SNPP: automating large-scale SNP genotype data management. Bioinformatics 2005;21:266-8.

50. Yeung JM, Sham PC, Chan AS, Cherny SS. OpenADAM: an open source genome-wide association data management system for Affymetrix SNP arrays. BMC genomics 2008;9:636.

51. Li JL, Li MX, Deng HY, Duffy PE, Deng HW. PhD: a web database application for phenotype data management. Bioinformatics 2005;21:3443-4.

52. Grubb SC, Maddatu TP, Bult CJ, Bogue MA. Mouse phenome database. Nucleic acids research 2009;37:D720-30.

53. Ott J. Analysis of human genetic linkage. Baltimore: The Johns Hopkins University Press; 1991.

54. Fishelson M, Dovgolevsky N, Geiger D. Maximum likelihood haplotyping for general pedigrees. Human heredity 2005;59:41-60.

55. Epstein MP, Duren WL, Boehnke M. Improved inference of relationship for pairs of individuals. American journal of human genetics 2000;67:1219-31.

56. Makinen VP, Parkkonen M, Wessman M, Groop PH, Kanninen T, Kaski K. High-throughput pedigree drawing. European journal of human genetics : EJHG 2005;13:987-9.

57. Perlin MW, Burks MB, Hoop RC, Hoffman EP. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. American journal of human genetics 1994;55:777-87.

58. Hall JM, LeDuc CA, Watson AR, Roter AH. An approach to high-throughput genotyping. Genome research 1996;6:781-90.

59. Adams P. LABMAN and LINKMAN: a data management system specifically designed for genome searches of complex diseases. Genetic epidemiology 1994;11:87-98.

60. Cheung KH, Nadkarni P, Silverstein S, et al. PhenoDB: an integrated client/server database for linkage and population genetics. Computers and biomedical research, an international journal 1996;29:327-37.

61. McMahon FJ, Thomas CJ, Koskela RJ, et al. Integrating clinical and laboratory data in genetic studies of complex phenotypes: a network-based data management system. American journal of medical genetics 1998;81:248-56.

62. Seuchter SA, Skolnick MH. HGDBMS: a human genetics database management system. Computers and biomedical research, an international journal 1988;21:478-87.

63. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human

diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 2009;106:9362-7.

64. Guan Y, Stephens M. Practical issues in imputation-based association mapping. PLoS genetics 2008;4:e1000279.

65. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS genetics 2007;3:e114.

66. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nature genetics 2010;42:436-40.

67. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human molecular genetics 2008;17:R122-8.

68. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics 2009;10:191-201.

69. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research 2014;42:D1001-6.

70. Genomes Project C, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.

71. Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC. Performance of genotype imputations using data from the 1000 Genomes Project. Human heredity 2012;73:18-25.

72. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology 2010;34:816-34.

73. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 2009;5:e1000529.

74. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 2011;1:457-70.

75. Huang L, Wang C, Rosenberg NA. The relationship between imputation error and statistical power in genetic association studies in diverse populations. American journal of human genetics 2009;85:692-8.

76. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 2006;78:629-44.

77. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. Analyses and comparison of accuracy of different genotype imputation methods. PloS one 2008;3:e3551.

78. Hancock DB, Levy JL, Gaddis NC, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PloS one 2012;7:e50610.

79. McGrady MG, Ellwood RP, Maguire A, Goodwin M, Boothman N, Pretty IA. The association between social deprivation and the prevalence and severity of dental caries and fluorosis in populations with and without water fluoridation. BMC public health 2012;12:1122.

80. Delaneau O, Coulonges C, Zagury JF. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC bioinformatics 2008;9:540.

81. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods 2012;9:179-81.

82. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. American journal of human genetics 2013;93:687-96.

83. O'Connell J, Gurdasani D, Delaneau O, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS genetics 2014;10:e1004234.

84. Powers S, Gopalakrishnan S, Tintle N. Assessing the impact of non-differential genotyping errors on rare variant tests of association. Human heredity 2011;72:153-60.

85. Mayer-Jochimsen M, Fast S, Tintle NL. Assessing the impact of differential genotyping errors on rare variant tests of association. PloS one 2013;8:e56626.

86. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.

87. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 2010;11:446-50.

88. Krithika S, Valladares-Salgado A, Peralta J, et al. Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs. BMC medical genomics 2012;5:12.

89. Li L, Li Y, Browning SR, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PloS one 2011;6:e24945.

90. Chen WM, Abecasis GR. Estimating the power of variance component linkage analysis in large pedigrees. Genetic epidemiology 2006;30:471-84.

91. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature genetics 2008;40:1068-75.

92. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 2011;189:317-27.

93. Cheung CY, Thompson EA, Wijsman EM. GIGI: an approach to effective imputation of dense genotypes on large pedigrees. American journal of human genetics 2013;92:504-16.

94. Diehl SR, Wang Y, Brooks CN, et al. Linkage disequilibrium of interleukin-1 genetic polymorphisms with early-onset periodontitis. Journal of periodontology 1999;70:418-30.

95. Tobler AR, Short S, Andersen MR, et al. The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. Journal of biomolecular techniques : JBT 2005;16:398-406.