UNIVERSITY OF CALIFORNIA

Los Angeles

Efficient use of Genetic Data for Mapping Complex Traits: Improved Data Management, Significance Testing for Marker Allele Sharing Statistics, and Estimation of Kinship Coefficients

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Human Genetics

by

Aaron Garth Day-Williams

2009

The dissertation of Aaron Garth Day-Williams is approved.

Robert E. Weiss

Robert E. Weiss

Aldons J. Lusis

Paivi Pajukanta

Päivi Pajukanta

Eric Sobel

Eric M. Sobel

Kenneth Lange

Kenneth L. Lange, Committee Chair

University of California, Los Angeles

2009

To Matt and Audrey.

Contents

List of Tables

List of Figures

Acknowledgments

First I would like to thank Dr. Leena Peltonen who introduced me to Dr. Kenneth Lange and Dr. Eric Sobel. This dissertation would have never materialized without Dr. Peltonen's support and suggestion that I work with Dr. Lange and Dr. Sobel. That simple introduction was the first step in the journey that has led me to this point, and I will be forever grateful.

While working on this dissertation I was supported by the UCLA National Science Foundation IGERT Bioinformatics Training Grant [DGE-9987641]. I especially want to thank Dr. Fred Fox, the principal investigator of the training grant, for all of his support and encouragement throughout my entire graduate career.

I would like to thank Dr. Paivi Pajukanta and her lab for early discussions of data management needs that led me to develop the GGSD application discussed in Chapter 2. Chris Plaisier was especially helpful in our many discussions on and his feedback of the user-interface design and functionality of the application. Tools development is always improved by close collaboration with and feedback from the intended audience and Chris and the rest of Pajukanta lab were invaluable in this role.

I would like to thank Dr. Kenneth Lange for all his support and encouragement. Without Ken's insightful suggestions and mentorship Chapter 4 would never have been as successful a project. Ken's intellectual curiosity and love of his work is inspiring and having him as a mentor has helped me grow as a scientist.

Dr. Eric Sobel has been my closest advisor and collaborator throughout my dissertation. I am deeply indebted to Eric for all of his mentorship and support. I especially appreciate Eric's support of me continuing to take classes that interested me throughout my dissertation. Those classes helped me grow as a scientist and in turn improved my research. Eric has a genuinely collaborative nature, and it has been a joy to work with him.

This dissertation would never have been finished without the love and support of my wife, Audrey. Her patience and understanding during the last month of the dissertation made it possible for me to work the weird hours of a finishing Ph.D. student.

VITA

1977	Born: Fort Walton Beach, FL, USA
1998-2000	Research Assistant Department of Microbiology University of Florida Gainesville, FL, USA
1999	Research Internship Bioscience Division Los Alamos National Laboratory Los Alamos, NM, USA
2000	B.S., Microbiology and Cell Science University of Florida, Gainesville, FL, USA
2000-2001	Bioinformatics Software Engineer Bioscience Division Los Alamos National Laboratory Los Alamos, NM, USA

PUBLICATIONS AND PRESENTATIONS

Aaron Day. Generic Genetic Studies Database (GGSD): web-based data management for large scale genetic studies, *Annual Meeting of the American Society of Human Genetics*, San Diego, CA, Oct 2007.

Day A., Sobel E., Lange K. Estimating Kinship Coefficients from High-Density SNP Genotypes for QTL Mapping and Association, *Annual Meeting of the American Society of Human Genetics*, New Orleans, LA, Oct 2006.

Lilja H.E., Suviolahti E., Soro-Paavonen A., Hiekkalinna T., Day A., Lange K., Sobel E.,Taskinen MJ, Peltonen L., Perola M., Pajukanta P. Locus for quantitative HDL-cholesterol on chromosome 10q in Finnish families with dyslipidemia, *Journal of Lipid Research*, 45(10):1876-84, 2004

Day A., Sobel E., Lange K., Lusis A.J., de Bruin T.W., Taskinen MJ, Pajukanta P. High-Density Lipoprotein Cholesterol Locus on Chromosome 16 Restricted to 3.5-Mb, *Annual Meeting of the American Society of Human Genetics*, Los Angeles, CA, Oct 2003.

Cleland C., Zheng W., Day A., Kaderali L., Desphande A., Gallagher K., White P.S., Nolan J., Brettin T. A Database-Integrated Pipeline for SNP Discovery, *Pacific Symposium on Biocomputing*, Kaua'i, Hawaii, Jan 2003

Cleland C., Zheng W., Day A., Kaderali L., Desphande A., Gallagher K., White P.S., Nolan J., Brettin T. XBase: An Open Source Gene Expression Database Project, *Fifth Annual International Conference on Computational Biology RECOMB*, Montreal, Canada , April 2001

Abstract Of The Dissertation

Efficient Use of Genetic Data for Mapping Complex Traits: Improved Data Management, Significance Testing for Marker Allele Sharing Statistics, and Estimation of Kinship Coefficients

by

Aaron Garth Day-Williams

Doctor of Philosophy in Human Genetics
University of California, Los Angeles, 2009
Professor Kenneth L. Lange, Chair

The genetic analysis of complex traits is being transformed by the generation and analysis of genome-wide genetic marker data. These new data sets open up exciting new possibilities as well as pose new challenges. But these data sets are not a panacea for elucidating the genetics of complex traits and therefore there is still a need to refine and improve traditional methods of analysis. I start by addressing one of the major challenges posed by the generation of genome-wide marker data, how to efficiently and correctly manage the data. These large datasets need to be stringently managed to minimize the introduction of error and bias, as well as integrate them with phenotype data for analysis. I present a web-based, relational database driven data management tool that allows researchers to easily manage and analyze their genetic data and integrate it with phenotype information. Due to the complex inheritance pattern of these traits, nonparametric linkage (NPL) analysis is still a good technique to study pedigrees to identify regions of the genome involved in these traits. The analysis of quantitative traits is very powerful and we present a new method that allows for researchers to perform NPL analysis using quantitative traits. Additionally, we show that using the Kong & Cox method as a general framework we can combine the exact NPL analysis of small pedigrees with estimation NPL techniques on large pedigrees to efficiently and correctly perform a single significance test. Finally, we investigate one of the exciting new possibilities that the new genome-wide genotyping technologies have opened. We develop three new algorithms that allow researchers to form and analyze pedigrees constructed only from genome-wide genotypes without any *a priori* information of genetic relatedness.

Chapter 1

Introduction

The investigation into the genetic components of human disease has been transformed over the past 7 years. From the publication of the first draft sequence of the human genome in 2001 (The International Human Genome Sequencing Consortium *et al.*, 2001; Venter *et al.*, 2001), to the development of high-throughput whole genome genotyping technologies (Matsuzaki *et al.*, 2004), to today when researchers are using new massively parallel sequencing technologies to sequence thousands of human genomes (http://www.1000genomes.org/) the study of human disease is being transformed into a data-driven, information science. These huge technological advances that produce such vast and in depth surveys of the variation in the genome makes computational genetics an exciting and rapidly changing field. The data is driving researchers to develop new computational tools that can handle and analyze the data.

The analysis of clean data that is free of error and bias is essential if the available statistical methods are to find the genetic signal. Genotyping technology has very rapidly developed over the past ten years and now allows researchers to assay hundreds of thousands of single nucleotide polymorphisms (SNPs) in thousands of samples very rapidly. Ensuring that this volume of data is stringently managed and free of error is an important task, and is daunting for some researchers. During the era of microsatellite based genome scans most researchers were managing all of their data in unlinked spreadsheets or other flat files. Even with that amount of data ensuring error free transmission of data from the lab to the analysis programs was difficult. Using flat files to manage and analyze the data generated from the new genome-wide SNP technologies is inefficient and impractical. Chapter 2 2 discusses the development of an application that harnesses the power and robustness of relational database management systems for the management of large scale genetic studies. This new data management tool allows users to easily manage and integrate their large scale genotype data with their phenotype data, and perform complex exploratory searches of the data. The developed application has a simple web-based user interface that allows easy collaboration between researchers spread across multiple labs or institutions.

The common, complex genetic traits that are currently the focus of genetics research have complex inheritance patterns, making them hard to model and therefore not good candidates for traditional parametric linkage analysis. The nonparametric linkage (NPL) analysis based on the amount of alleles shared among affected relatives is an attractive alternative. The traditional NPL method is for the analysis of qualitative traits, but quantitative traits hold more information and give more power to detect linkage. In Chapter 3 3 we develop a new test statistic that allows researchers to perform NPL analysis using quantitative traits. We develop the mathematical theory behind the method and demonstrate its ability to identify genetic signals in both simulated and real data. The most efficient and robust method of determining the statistical significance of the traditional NPL method is also investigated in this chapter. We demonstrate that there is now a framework to efficiently combine exact analysis of small pedigrees and estimation techniques used on large pedigrees for both quantitative and qualitative NPL analysis into a single significance test, thus improving computational efficiency and power.

The genome-wide SNP genotyping technologies that are posing challenges in managing data are also opening new and exciting ways to analyze data. In Chapter 4 4 we develop three new algorithms that utilize genome-wide SNP genotypes to determine the genetic relatedness between random pairs of individual without any *a prior* knowledge. The first algorithm is a methods-of-moments estimator that determines the relatedness between individuals in a global sense. The second algorithm is a penalized optimization technique that determines the relatedness between individuals on a locus by locus basis for every SNP assayed. The third algorithm uses the first algorithm with a standard algorithm from graph theory to cluster individuals into pedigrees. The chapter develops the theory and mathematics of the algorithms and illustrates their success through extensive simulation studies.

Chapter 2

The Generic Genetic Studies Database: Web-based Data Management for Large Scale Human Genetic Studies

2.1 Introduction

Recent advancements in genotyping technology are fundamentally changing the way the genetic basis of complex diseases are investigated. The new technologies allow researchers to investigate hundreds of thousands of single nucleotide polymorphisms (SNPs) across the entire genome simultaneously. These technologies enable researchers to conduct genome-wide association (GWA) and other genome-wide studies that, until these technological advances, had only been hypothesized as ways to map the genetics of complex disease (Risch and Merikangas, 1996). The National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI) catalog of published GWA studies demonstrates the tremendous growth and success of these studies (http://www.genome.gov/26525384). Despite significant progress, the impact of these discoveries has been limited because the identified variants only explain a fraction of the observed familial aggregation of the diseases studied (Altshuler and Daly, 2007).

Identification of loci with such modest effects requires sample sizes in the thousands (McCarthy *et al.*, 2008). Genotyping samples this large with the new technology generates data orders of magnitude greater than the microsatellite based whole-genome mapping studies of just a few years ago (Pajukanta *et al.*, 2003). Further complicating data management and analysis is the fact that GWA and other genome-wide SNP based studies are commonly highly collaborative with researchers spread across multiple labs and institutions. This highly collaborative setting makes storing data in flat files that need to be edited and shared impractical, inefficient, and adds an avenue to introduce error. The first wave of successful GWA studies has shown that the data needs to be accurately and precisely managed in order to eliminate error and bias, and that this has proven challenging in this collaborative environment (McCarthy *et al.*, 2008).

The establishment of a data management system is an expensive and time-consuming endeavor with a major component of the cost and time involving the development of the database schema and user interface. GWA and other large-scale genetic studies have a common set of data types and tasks that need managing, opening the possibility of a generalized solution. The Generic Model Organism Database (GMOD) project recognized that sequencing projects all share a common set of data types and tasks and have developed generalized data management tools that are provided to their research community free of charge under open source licenses (http://www.gmod.org)(Stein *et al.*, 2002; Lewis *et al.*, 2002). The genetics research community would benefit from a similar development project that produces a free data management tool that allows researchers to easily manage the massive datasets generated by the new technology and easily share and analyze the data. The Generic Genetic Studies Database (GGSD) project has developed a relational database schema and suite of web-based management tools that handles the data types and performs the tasks needed to conduct genome-wide genetic studies, and is released free under an open source license (Day, 2007).

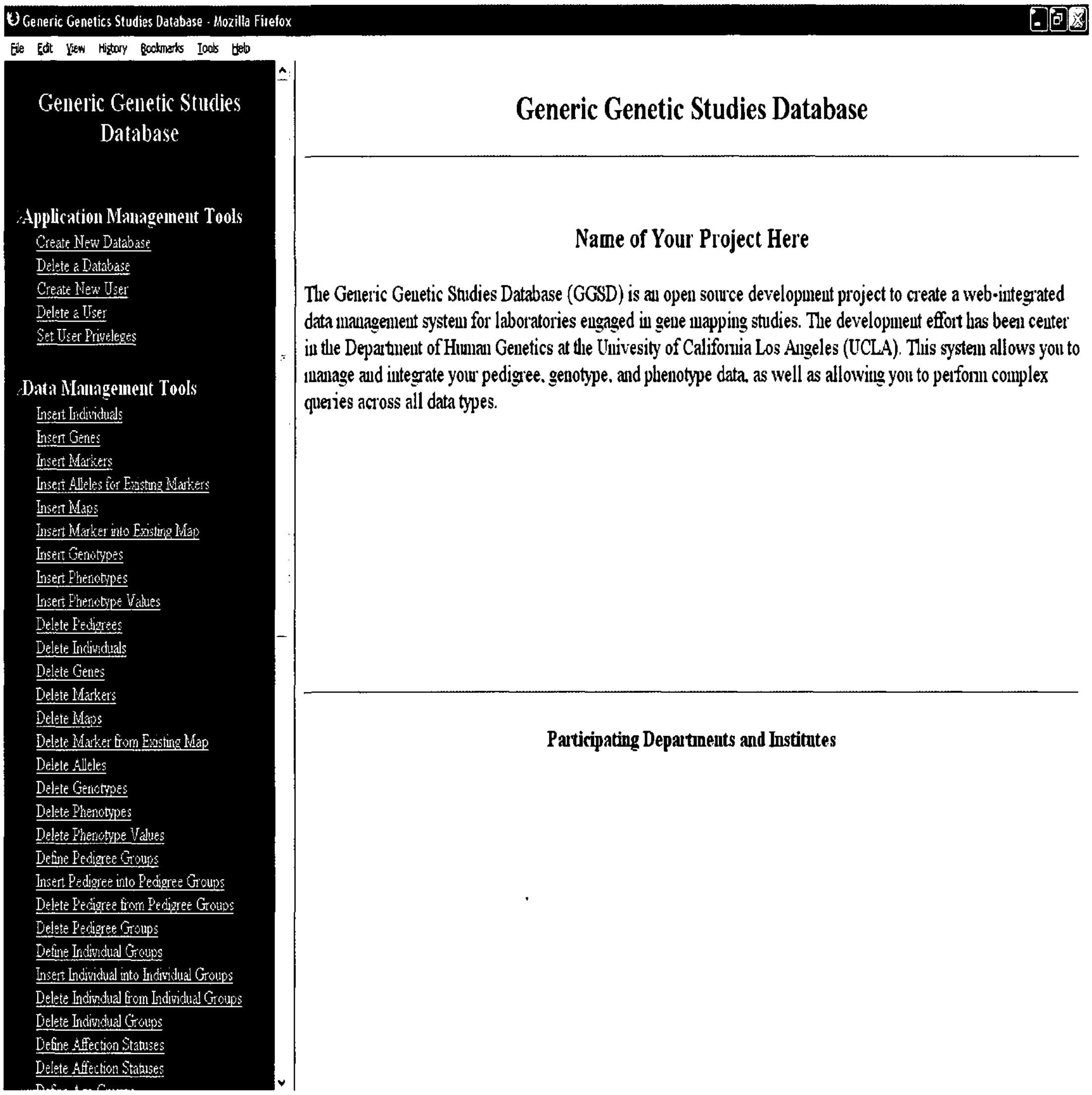
GGSD stores, organizes and links all the pedigree/individual, genomic, phenotypic and disease status information used in gene mapping studies. GGSD is not intended to be a laboratory information management system (LIMS) and therefore was not designed to manage or track information such as the plate and well location of DNA samples. LIMSs are very specific applications intended to manage laboratory work flow and improve the quality of data generated by the laboratory, and although extremely important to genetic studies was not the focus of GGSD's development. Additionally, genotyping using the new genome-wide SNP technologies is generally performed in core laboratories or as a service from the technology developer making LIMS a truly separate system. The central design principle behind GGSD is to facilitate the analysis of the data it stores. Therefore, GGSD is intended to be a central repository of data that allows researchers to easily insert, manage, search, edit, and download data. The system is suitable for any research group that must manage the above types of information and wants a single, fully integrated data management solution.

2.2 User Interface and Tools

GGSD provides a simple, web-based interface that is easy to navigate. The entry point to the system is the homepage pictured in Figure 2.1. The right frame of the homepage is configurable, allowing each group to customize the look and content of their installation. The tools of the system are all accessible from the scrollable tool bar along the left frame of the homepage. The tools monitor the size of requested data to be imported or downloaded, and if the size exceeds a configurable threshold the tool executes in the background. When the tool finishes the user receives an email with the results. Due to the sensitive nature of the data being managed all tools are password protected and all data transmissions are encrypted by requiring connection via https. Researchers can create and manage multiple projects with a single installation, as well as assign different access levels to users on a project-by-project basis.

2.2.1 Data Importing, Editing and Deleting

Importing data is performed by either uploading formatted files or entering data through simple web forms. GGSD has defined simple comma-separated file formats to import individual, gene, marker, map, phenotype and genotype data. In addition to the GGSD defined file formats for marker and genotype data, GGSD handles files generated by some of the SNP-chip manufacturers including Affymetrix (annotation and genotype call files) and Illumina (Manifest, OPA, and genotype call files). GGSD also allows users to edit data once it has been inserted. The data is edited through web forms, and only allows the user to update fields that do not alter or break the referential integrity of the data. Deleting


Generic Genetics Studies Database - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Generic Genetic Studies Database
Application Management Tools
Create New Database
Delete a Database
Create New User
Delete a User
Set User Priveleges
Data Management Tools
Insert Individuals
Insert Genes
Insert Markers
Insert Alleles for Existing Markers
Insert Maps
Insert Marker into Existing Map
Insert Genotypes
Insert Phenotypes
Insert Phenotype Values
Delete Pedigrees
Delete Individuals
Delete Genes
Delete Markers
Delete Maps
Delete Marker from Existing Map
Delete Alleles
Delete Genotypes
Delete Phenotypes
Delete Phenotype Values
Define Pedigree Groups
Insert Pedigree into Pedigree Groups
Delete Pedigree from Pedigree Groups
Delete Pedigree Groups
Define Individual Groups
Insert Individual into Individual Groups
Delete Individual from Individual Groups
Delete Individual Groups
Define Affection Statuses
Delete Affection Statuses

Generic Genetic Studies Database

Name of Your Project Here

The Generic Genetic Studies Database (GGSD) is an open source development project to create a web-integrated data management system for laboratories engaged in gene mapping studies. The development effort has been center in the Department of Human Genetics at the University of California Los Angeles (UCLA). This system allows you to manage and integrate your pedigree, genotype, and phenotype data, as well as allowing you to perform complex queries across all data types.

Participating Departments and Institutes

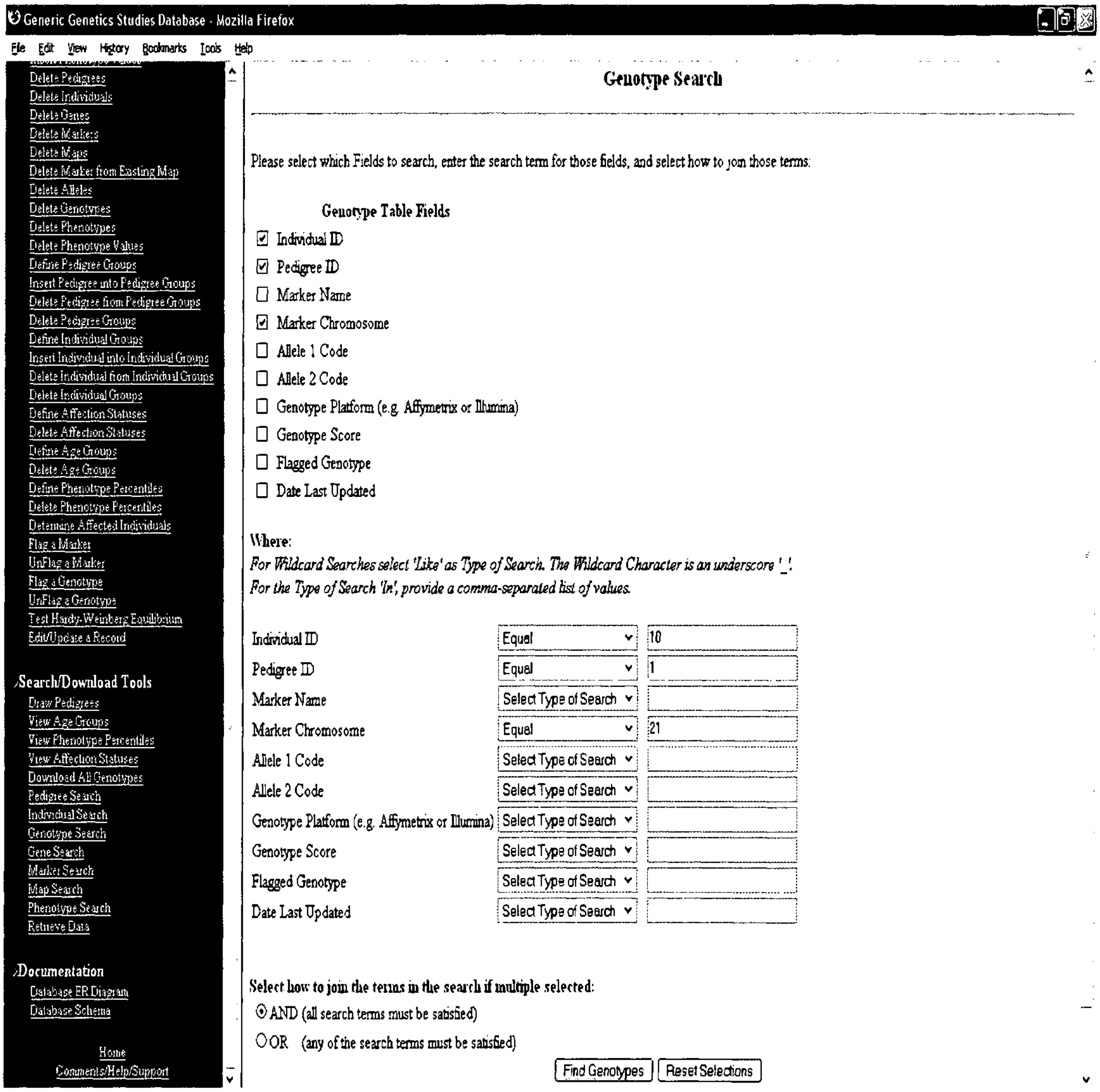
Figure 2.1: GGSD Configurable Home Page

data is just as simple as inserting data, and is performed by either uploading formatted files or entering data via web forms. Users are alerted when data is inserted, edited, or deleted through automatically generated emails that inform them of the type and amount of data that was altered. This ensures that all users are analyzing the most up to date data. Additionally, the system logs the user and time of any insertion, deletion or editing of data.

2.2.2 Searching and Downloading Data

GGSD facilitates user exploration and analysis of data by allowing complex, multi-field querying. The search tools are separated into seven data type specific search tools: pedigree search, individual search, gene search, marker search, map search, genotype search and phenotype search. The search tools allow users to select any table in the database schema and any number of fields in that table to search for the specific data type, as seen in Figure 2.2 for searching the genotype table. The tools provide the user with all the power and complexity of searching the relational database using the structured query language (SQL) without having to know any of its syntax. The user is provided a table of results that match their query and allows them to select which elements they want to download data for. All the tools allow the user to download records as they appear in the database, as well as giving the user a path to download data in analysis ready formats.

Given that the central design principle behind GGSD is to facilitate the analysis of data, the users are given the ability to generate data files for a number of analysis packages. The default download format is the pre-Makeped format which is accepted by programs such as Haploview (Barrett *et al.*, 2005). GGSD has integrated support for the Mega2 software and allows users to select, gen-



Genotype Search

Please select which Fields to search, enter the search term for those fields, and select how to join those terms:

Genotype Table Fields

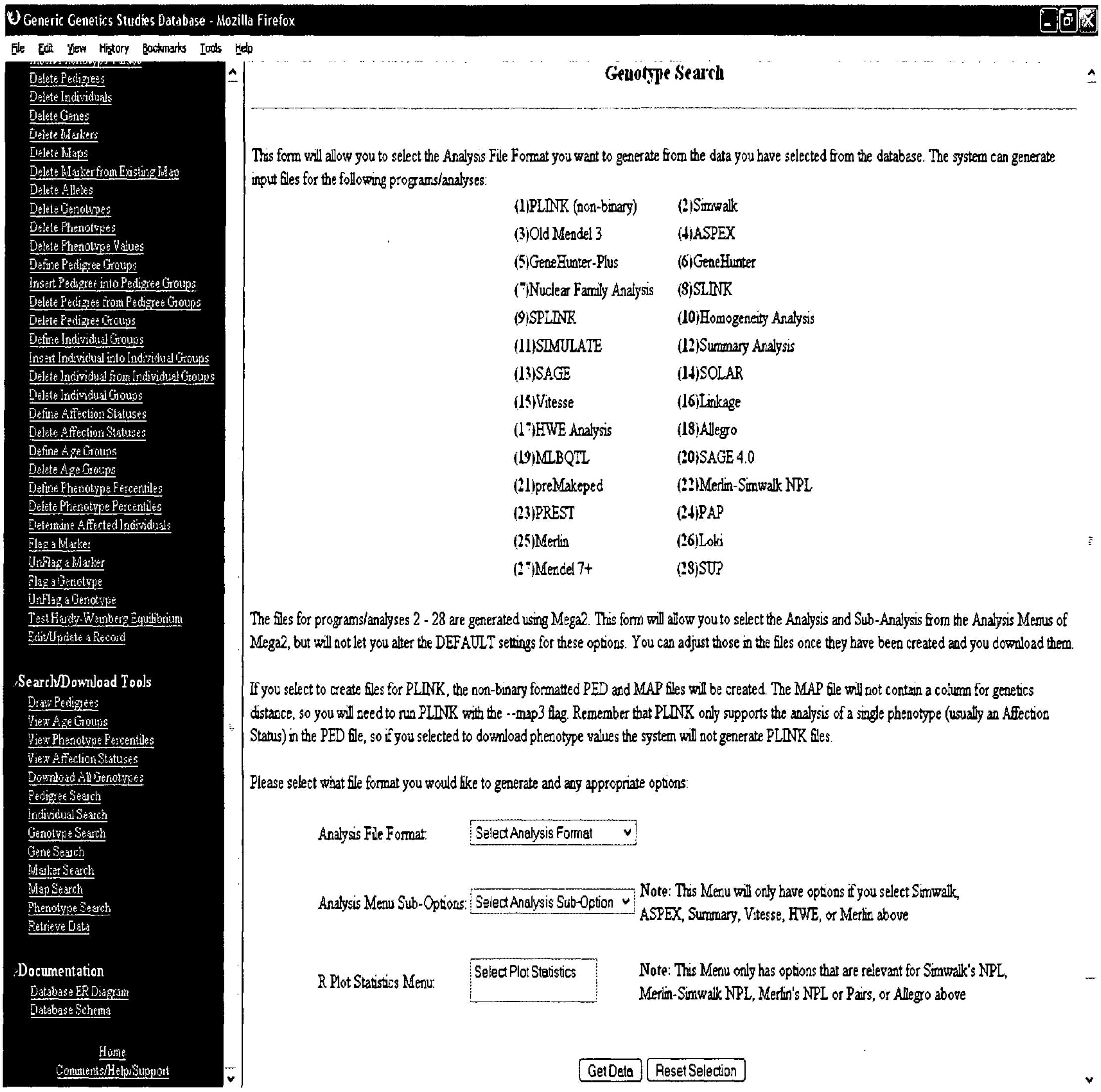
☑ Individual ID
☑ Pedigree ID
□ Marker Name
☑ Marker Chromosome
□ Allele 1 Code
□ Allele 2 Code
□ Genotype Platform (e.g. Affymetrix or Illumina)
□ Genotype Score
□ Flagged Genotype
□ Date Last Updated

Where:
For Wildcard Searches select 'Like' as Type of Search. The Wildcard Character is an underscore '_'.
For the Type of Search 'In', provide a comma-separated list of values.

Individual ID	Equal	10
Pedigree ID	Equal	1
Marker Name	Select Type of Search	
Marker Chromosome	Equal	21
Allele 1 Code	Select Type of Search	
Allele 2 Code	Select Type of Search	
Genotype Platform (e.g. Affymetrix or Illumina)	Select Type of Search	
Genotype Score	Select Type of Search	
Flagged Genotype	Select Type of Search	
Date Last Updated	Select Type of Search	

Select how to join the terms in the search if multiple selected:
● AND (all search terms must be satisfied)
○ OR (any of the search terms must be satisfied)

Figure 2.2: GGSD Example Search Definition Page



Genotype Search

This form will allow you to select the Analysis File Format you want to generate from the data you have selected from the database. The system can generate input files for the following programs/analyses:

(1)PLINK (non-binary) (2)Simwalk
(3)Old Mendel 3 (4)ASPEX
(5)GeneHunter-Plus (6)GeneHunter
(7)Nuclear Family Analysis (8)SLINK
(9)SPLINK (10)Homogeneity Analysis
(11)SIMULATE (12)Summary Analysis
(13)SAGE (14)SOLAR
(15)Vitesse (16)Linkage
(17)HWE Analysis (18)Allegro
(19)MLBQTL (20)SAGE 4.0
(21)preMakeped (22)Merlin-Simwalk NPL
(23)PREST (24)PAP
(25)Merlin (26)Loki
(27)Mendel 7+ (28)SUP

The files for programs/analyses 2 - 28 are generated using Mega2. This form will allow you to select the Analysis and Sub-Analysis from the Analysis Menus of Mega2, but will not let you alter the DEFAULT settings for these options. You can adjust those in the files once they have been created and you download them.

If you select to create files for PLINK, the non-binary formatted PED and MAP files will be created. The MAP file will not contain a column for genetics distance, so you will need to run PLINK with the --map3 flag. Remember that PLINK only supports the analysis of a single phenotype (usually an Affection Status) in the PED file, so if you selected to download phenotype values the system will not generate PLINK files.

Please select what file format you would like to generate and any appropriate options:

Analysis File Format: Select Analysis Format

Analysis Menu Sub-Options: Select Analysis Sub-Option Note: This Menu will only have options if you select Simwalk, ASPEX, Summary, Vitesse, HWE, or Merlin above

R Plot Statistics Menu: Select Plot Statistics Note: This Menu only has options that are relevant for Simwalk's NPL, Merlin-Simwalk NPL, Merlin's NPL or Pairs, or Allegro above

Figure 2.3: GGSD download file format selection screen

erate and download the files Mega2 generates for 27 different analysis options (Mukhopadhyay *et al.*, 2005). GGSD also generates non-binary formatted files for the PLINK software (Purcell *et al.*, 2007). Figure 2.3 shows the form allowing users to select which program the downloaded data should be formatted for. Additionally, the Cranefoot pedigree drawing software has been integrated allowing users to select, draw and download pedigree structures (Makinen *et al.*, 2005).

2.2.3 Incorporation of Quality Control Analysis

Researchers involved in the analysis of the first wave of GWA studies have emphasized the importance of stringent quality control measures to eliminate the introduction of error and bias (McCarthy *et al.*, 2008). GGSD allows researchers two main ways to incorporate quality control information into the system, the flagging of data that is not worthy of analysis and the creation of groups of individuals that have been cleared for analysis.

GGSD provides utilities to flag specific data types as not passing quality control procedures. The system allows markers, genotypes and phenotype values to be flagged in the database. The flagging of a marker does not remove it from the database or alter its information. What flagging a marker means is if it is selected for analysis the genotypes for all individuals will be downloaded as missing genotypes, effectively removing it from analysis. Users can either select markers to flag through the web interface or upload a file of marker names. A standard quality control procedure is testing markers for violation of Hardy-Weinberg equilibrium (HWE). GGSD has a built-in utility to perform the Chi-squared goodness-of-fit HWE test for microsatellites and the exact HWE test for SNPs as implemented in the program PEDSTATS (Wigginton and Abecasis, 2005; Wigginton *et al.*, 2005). The first large GWA studies have shown that there is not a specific p-value threshold that can be employed for identifying markers that satisfy verses violate HWE, rather the p-value cutoff should be determined on a study by study basis (McCarthy *et al.*, 2008). Therefore GGSD allows users to set the p-value threshold that separates markers that satisfy verses violate HWE. Users are given a file detailing the markers that violate HWE based upon the user define p-value threshold. Markers that violated HWE in this analysis are flagged in the database. Individual genotypes and phenotype values can also be flagged, which results in them being downloaded as missing genotypes or phenotypes for analysis purposes. The users can either select the genotypes or phenotype values through the web interface or upload a file of the values to be flagged. Important is the fact that none of the data stored in the database is altered when it is flagged, only the values printed out for analysis purposes are altered. This allows the researchers to go back and unflag the data at a later time if upon further analysis they believe the data to be suitable for analysis.

Users can define groups of individuals or pedigrees, and then link the individuals that satisfy the defined group's criteria by uploading a file of the individuals or selecting them through the web interface. The users can then filter all their searches using defined groups to restrict the individuals included in analyses. Another potential source of error is the misclassification of which individuals are affected for traits. GGSD has a utility that automates the calling of affected individuals based on the stored phenotype information for each individual, removing the potential error introduced when researchers enter data by hand. The user selects the trait they want to assign affection status for, the phenotype to base the assignment on, enters the test criteria and comparison method, and GGSD investigates all individuals in the database and assigns an affection status.

2.3 GGSD Internals

2.3.1 Data Types & Database Design

The heart of GGSD is the relational database schema that defines the entities that are tracked, the data stored for each entity, and how all the entities relate to each other. GGSD has defined 16 primary entities seen GGSD's entity-relationship model in Figure 2.4. These 16 entities are translated into the 16 primary tables of the relational database schema. The defined entities are classified into three distinct types of information: individual, genomic, and phenotypic. The database information is conceptually separated into three information spaces. The first information space being that of the project. A project is comprised of a distinct set of individual, genomic, and phenotypic data that is linked. Data is not shared between projects. The second information space is the space of users. A user is an entity that has an application managed user ID, password and email address. A user has links to projects and is assigned access rights (e.g. read access only) on a project-by-project basis. The third information space is the actual data that is stored in the database for a project.

The individual information is the central information in the schema because they are the entities that are analyzed in genetic research. The individual information is divided into 4 tables: pedigree table, pedigree group table, individual table and individual group table. A pedigree is a collection of individuals connected by genetic relationships, and the pedigree table stores the number of individuals in a pedigree and their nationality. An individual is a person that belongs to a pedigree and has a mother, father and sex. Individuals can also have genotypes for markers, phenotype values for phenotypes, and an affection status for a disease state. Pedigree and Individual groups are collections of pedigrees or individuals that can be collected together based on non-genetic information (e.g. cases and controls).

The second type of data that GGSD stores is genomic information. The top level abstraction of genomic information is a gene. The gene table stores the name, unigene accession number, chromosome, start and stop base pair positions, the number of introns and exons, and a description for a gene. The next level of information is the marker. A marker is a genome sequence element that is polymorphic in the population. GGSD is capable of storing data on both microsatellite and SNP markers. The marker table stores information on the marker name, type, the number of alleles it has, whether it belongs to a gene in the database, the chromosome it is on, its chromosome base pair position, the genomic strand, genomic position (e.g. intron, exon, 5'UTR,3'UTR, intergenic), the genome build and dbSNP version it is from, if it is a tag SNP, and whether it codes for a synonymous or nonsynonymous amino acid substitution. An allele is one of the polymorphic forms of a marker, and the database stores the marker an allele is for, its size, its sequence, its code, its frequency in the population and how the Affymetrix and Illumina technologies code the allele. Genetic markers can be combined into genetic maps. A genetic map is an ordered set of markers where the distance between markers is known. In GGSD a map has a name, the number of markers in the map and how distance is measured between markers (e.g. base pairs, Haldane centiMorgans, Kosambi centiMorgans or recombination fraction). The markers in a map have a chromosome, its position in the map (e.g. first marker) and its distance from the previous marker in the map. The final piece of genomic information stored is a genotype. A genotype is composed of the alleles (one maternal and one paternal) an individual has for a genetic marker. The information stored for a genotype is the individual it is for, the marker it is for, the two alleles that compose the genotype, the platform used for genotyping, and any associated genotype score.

The final type of data is phenotypic information. The phenotypic data is stored in 7 tables. The first table is the phenotype table. A phenotype describes a physical attribute that can be measured for an individual and the phenotype table stores the name, type (e.g. quantitative), and a description detailing what the phenotype is measuring. A phenotype value is the measurement of a phenotype on a specific individual. GGSD also stores population based phenotypic information. For the purpose of classification and measurement individuals in a population are often separated into age groups (e.g. 18-20), and GGSD allows researchers to store this information. Diseases are often defined based on population sex and age-group specific distributions of values for quantitative traits. Therefore, GGSD allows researchers to store the population based sex-specific 10^{th} and 90^{th} percentiles of a quantitative trait for stored age groups. Finally the database allows users to define traits that individuals can an affection status, and stores allele information for the traits needed in methods of analysis.

2.3.2 Software Libraries & Extending the Application

The user tools of GGSD are driven by three software libraries. The first library contains methods for querying the underlying relational database and collecting the results. This library insulates the user tools from having to know the details of the database schema and provides a well defined programming interface to process users' queries and collect the results. The second library contains methods for

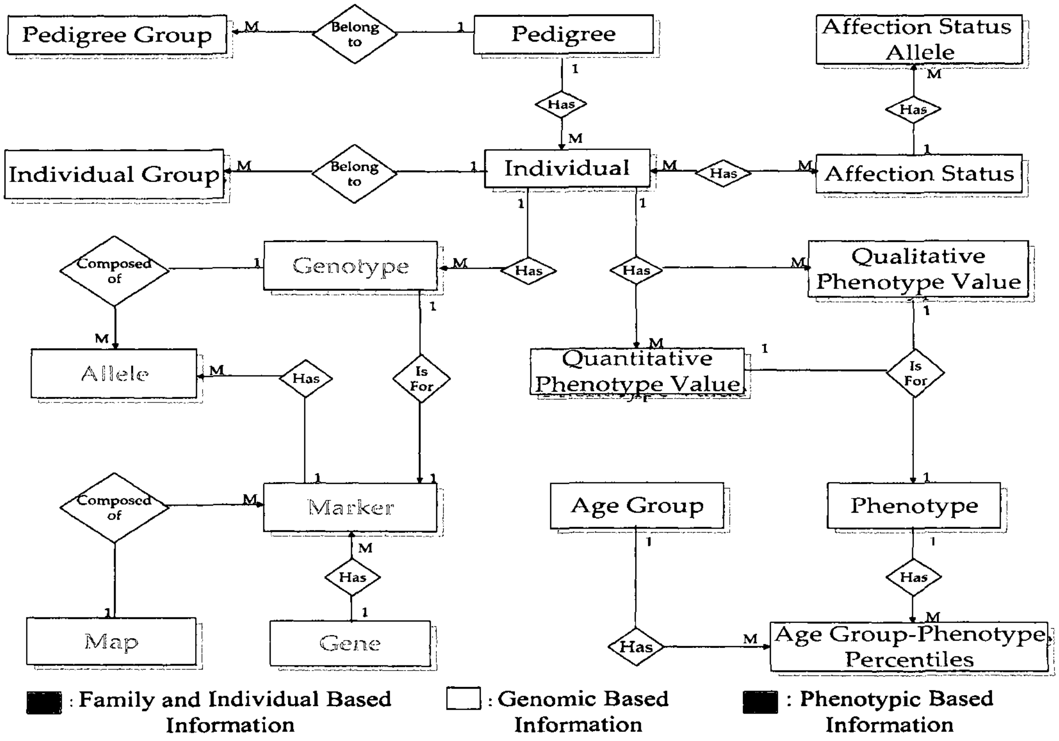


Figure 2.4: GGSD Entity-Relationship Diagram

generating html to appropriately and nicely display the results of users queries, as well as generating the JavaScript that allows the user to interact with and validate data. This library insulates the user tools from having to know how to format the different data types returned from the SQL library to display to the user by providing a well defined programming interface. The third library is a set of methods for interacting with and manipulating the file system of the web server. The methods in this library are used to monitor files stored in the individual users' data directories. When files are older than a configurable file expiration parameter the library functions are used to clean up the file system. This library ensures that the server's memory storage is not being used by old, unnecessary files.

The modularity and programming interfaces that these code libraries provide make writing new extensions and tools for GGSD straightforward. The developer is not constrained by needing an intimate knowledge of the database schema nor an understanding of how to correctly format the results into html. The process of writing extensions becomes a process of understanding what the new tool is to accomplish and calling the appropriate library functions to achieve that goal. Of course, developers are also able to add new methods to the libraries extending the underlying capabilities of what a GGSD user tool can accomplish. GGSD was designed from the bottom up to be extendable, modular and portable. That is why GGSD relies on no proprietary software, but only uses software components that are freely available under open source licenses.

2.3.3 GGSD Implementation, Requirements and Support

GGSD server side applications are written in PHP, with database connectivity using the PEAR MDB2 module. The client side form interaction and validation is written in JavaScript. The exact HWE test for SNPs is written in C (provided by the authors of PEDSTATS), and the chi-squared HWE test for microsatellites was written in Perl using the Math::CDF module. GGSD has been tested on and supports the use of both the MySQL and PostgreSQL relational database management systems. GGSD requires the server to have Mega2 installed for its Mega2 capabilities to function, and was developed against Mega2 version 4.0 R1. GGSD's Cranefoot capabilities require the user to install an executable copy of Cranefoot, and was developed against Cranefoot version 3.2.2. The GGSD application is hosted on SourceForge (http://sourceforge.net/projects/ggsd/), which provides a user interface for user support, feature requests, and reporting bugs.

2.4 Discussion

I have described a new web-based, relational database driven tool for the management of the individual, genomic and phenotypic information that researchers use in the analysis of complex traits. The need for such systems has been recognized and developed for over 20 years (Seuchter and Skolnich, 1988; Cheung *et al.*, 1996). But continuing technological advancements in genetic data generation and the types of data generated have required continual development of new data management schema and tools. There were a number of systems developed to handle the microsatellite based genome-wide scans that were prevalent just a few years ago (Li *et al.*, 2001; Gillanders *et al.*, 2004). But due to the specific design choices and the fact that they were designed to specific group's requests, they are not suitable for the genome-wide SNP based investigations of today. GGSD was designed from the beginning to manage the tremendous amount of data generated by today's high-throughput, genome-wide SNP genotyping technologies while allowing the integration of older microsatellite based data.

GGSD is not the first data management tool developed for high-throughput, genome-wide SNP data (Zhao *et al.*, 2005; Fiddy *et al.*, 2005). But it has several features and design principles that set it apart. A fundamental difference that sets GGSD apart from previously developed applications is that it was designed to be a general use application, and it was not designed around a specific groups work flow and analysis pipeline. GGSD focused on the components of gene mapping investigations that are shared and built a system that fulfills those needs. A consequence of this generalized approach, and based on today's service based genotyping model, is the dissection of LIMS related attributes and functions out of the application. This separation of LIMS functions is fundamentally different from previous data management tools (Li *et al.*, 2001; Gillanders *et al.*, 2004; Zhao *et al.*, 2005; Fiddy *et al.*, 2005). Additionally, GGSD was designed from the beginning to fully integrate phenotype data with genotype data into a single application and database schema, which sets it apart from groups that developed separate applications to manage genotypes and phenotypes (Zhao *et al.*, 2005; Li *et al.*, 2005). Central to the development of GGSD is the belief that the software should be free and released under an open source license. Therefore, GGSD does not rely on any proprietary software that requires groups to purchase licenses for software, unlike previously developed academic systems and the commercially available BC|Gene (http://www.bcplatforms.com/) and Progeny Lab(www.progenygenetics.com/) applications (Fiddy *et al.*, 2005). The open source framework that GGSD is built on makes the system economical and allows the system to grow and be modified by the genetics research community as the field continues to evolve.

Chapter 3

Unifying Nonparametric Linkage Analysis of Large and Small Pedigrees in a Kong & Cox Framework

3.1 Introduction

Linkage analysis still remains a powerful tool in mapping loci involved in complex traits (e.g. diabetes, obesity, etc). Despite the tremendous successes of genome-wide association studies to identify new loci, the majority of the heritability of the traits investigated remains unexplained (Altshuler and Daly, 2007). Due to the complex inheritance of these traits, linkage analysis based on allele sharing among affected relatives is an attractive method. Whittemore and Halpern proposed a class of these statistics that don't require the specification of an inheritance model, and the method based upon their ideas has become known as the non-parametric linkage (NPL) method (Whittemore and Halpern, 1994b,a; Kruglyak *et al.*, 1996).

Traditional NPL analysis is based upon the analysis of a dichotomous, qualitative trait. The fundamental idea is to calculate the number of alleles two affected relatives share identical by descent (IBD) and compare that to the expected proportion of IBD sharing based solely on their relationship. If two affected relatives share more alleles IBD than expected at a locus than that is suggestive of linkage. There are two major parts to this method, the scoring of IBD and the determination of statistical significance. The IBD scoring can be determined exactly on small and medium sized pedigrees. For large, complex pedigrees the problem is too computationally difficult to solve exactly so estimation techniques must be employed. The method of significance testing for both the exact and estimation implementations are the same and is called perfect data approximation (Kruglyak *et al.*, 1996). The perfect data approximation technique forms a distribution of IBD scores from analyzing a set of simulated pedigrees with 'complete' information. Complete information means that the pedigrees contain not only no missing genotypes, but also no missing or ambiguous phase information. It is known that using this method for significance testing is a conservative approach, especially when the data is missing genotype information. The perfect data approximation method was developed in the context of exact analysis, and is probably not the optimal method for significance testing for the estimation techniques that analyze large pedigrees. But we are not aware of any rigorous analysis of at what degree of missing data causes the perfect data approximation method to yield such conservative answers that true linkage signals are often missed by the exact analysis programs. Additionally, there is no analysis of how missing genotypes affect the estimation techniques of NPL analysis and their significance testing, which is the analysis paradigm in the program Simwalk2 (Sobel and Lange, 1996). Therefore we have performed an extensive analysis of the affects of missing data of the significance testing in both exact and estimation NPL analysis.

In addition to the traditional NPL method, Kong and Cox (from this point on referred to as K&C) proposed a 1-parameter extension to the method (Kong and Cox, 1997). The K&C extension uses the IBD scores from the traditional method and they stated that their extension calculates an exact log likelihood given any missing data pattern, and therefore would be a more robust method than the traditional method. A more robust method would greatly benefit the estimation technique of NPL analysis, like the analysis Simwalk2 performs. But the K&C extension was developed for the exact NPL method and the calculation uses an parameter that can be calculated exactly for small pedigrees but not for large pedigrees. We propose that we can appropriately estimate the needed parameter and implement the K&C method in Simwalk2, thereby improving its performance in the presence of missing data.

The original NPL method was based upon analyzing qualitative traits, but quantitative traits contain more information and provide more power to find linked loci. We propose a new test statistic that allows researchers to perform NPL analysis with quantitative traits, or Q-NPL analysis. Our new statistic is motivated by variance component models, but does not perform parameter estimation making the method computationally efficient. We show that our new Q-NPL method works well both in exact and estimation NPL analysis and identifies linkage to traits with a range of heritabilities.

Additionally, we see the K&C extension as a framework to unify exact NPL analysis on small pedigrees and stochastic analysis on large, complex pedigrees. We believe that this framework is suitable for both the traditional qualitative NPL and our new proposed quantitative NPL analysis. Through extensive simulation studies we will illustrate the weaknesses of perfect data approximation, the robustness of the K&C method against missing data, the favorable properties of our new quantitative NPL (Q-NPL) statistic and the ability of the K&C framework to unify exact and stochastic NPL analyses.

3.2 Qualitative NPL Materials and Methods

3.2.1 Pedigree Configurations

Three pedigree configurations were analyzed that have previously been used in an analysis of IBD computation (Sobel *et al.*, 2001). Pedigree configuration 1 seen in Figure 3.1a is a nuclear family with 5 total individuals, pedigree configuration 2 seen in Figure 3.1b is a 3 generation pedigree with 15 individuals and pedigree configuration 3 seen in Figure 3.1c is a 45 individual extended pedigree with 4 generations. These pedigrees correspond to pedigree configurations B, D, and A respectively in the previous analysis(Sobel *et al.*, 2001). Pedigree structure C from the previous analysis (Sobel *et al.*, 2001) was also analyzed but its results were similar to pedigree configuration 2 and therefore excluded from discussion. In addition, we studied a data set composed of a mixture of the above pedigree structures. The analysis of pedigree configuration 1 used 51 pedigrees with each pedigree containing either 2 or 3 affecteds; the analysis of pedigree configuration 2 used 14 pedigrees with each pedigree containing 2, 3, or 4 affected individuals; and the analysis of pedigree configuration 3 used 17 pedigrees with each pedigree containing between 2 and 10 affecteds. The mixture data set was composed of 30 pedigrees of configuration C from (Sobel *et al.*, 2001), 10 pedigrees of configuration 1, 7 pedigrees of configuration 2, 2 inbred pedigrees of configuration

E from the previous analysis (Sobel *et al.*, 2001), and 1 pedigree of configuration 3.

3.2.2 Data Simulation

We simulated a 9-marker microsatellite map with 10 cM between markers, as would be seen in a genome-wide scan. The allele frequencies used for these 9 markers are the allele frequencies from 9 microsatellites on chromosome 16 used in a previous gene mapping study (Pajukanta *et al.*, 2003). The genetic map has Marker1 (M1) at 0 cM and Marker9 (M9) at 80 cM (Table 3.1). We simulated a disease locus located half way between M4 and M5 at 35 cM. The locus at 35 cM has two alleles, a major allele with frequency 0.8905 and a minor allele with frequency 0.1095. We modeled the disease trait as additive with P(affected | homozygous for major allele) = 0.01, P(affected | heterozygous) = 0.45, and P(affected | homozygous for minor allele) = 0.90. Individuals were assigned an affection status based upon their genotype at the putative trait locus, and the model specified. The pedigrees initially had no missing genotype information, but not necessarily complete phase information. The missing genotype patterns were developed based upon looking at the types of missing genotype patterns that were present in the 73 families investigated in the previous study (Pajukanta *et al.*, 2003). The study samples for each investigation were generated by randomly choosing from the 1000 generated pedigrees for pedigrees with 2 or more affected individuals. This sampling scheme gave us pedigrees with a variety of familial relationships among the affecteds and simulates real study sample conditions. We restricted our analyses to those pedigrees in which none of the individuals who married into the pedigree were affected.

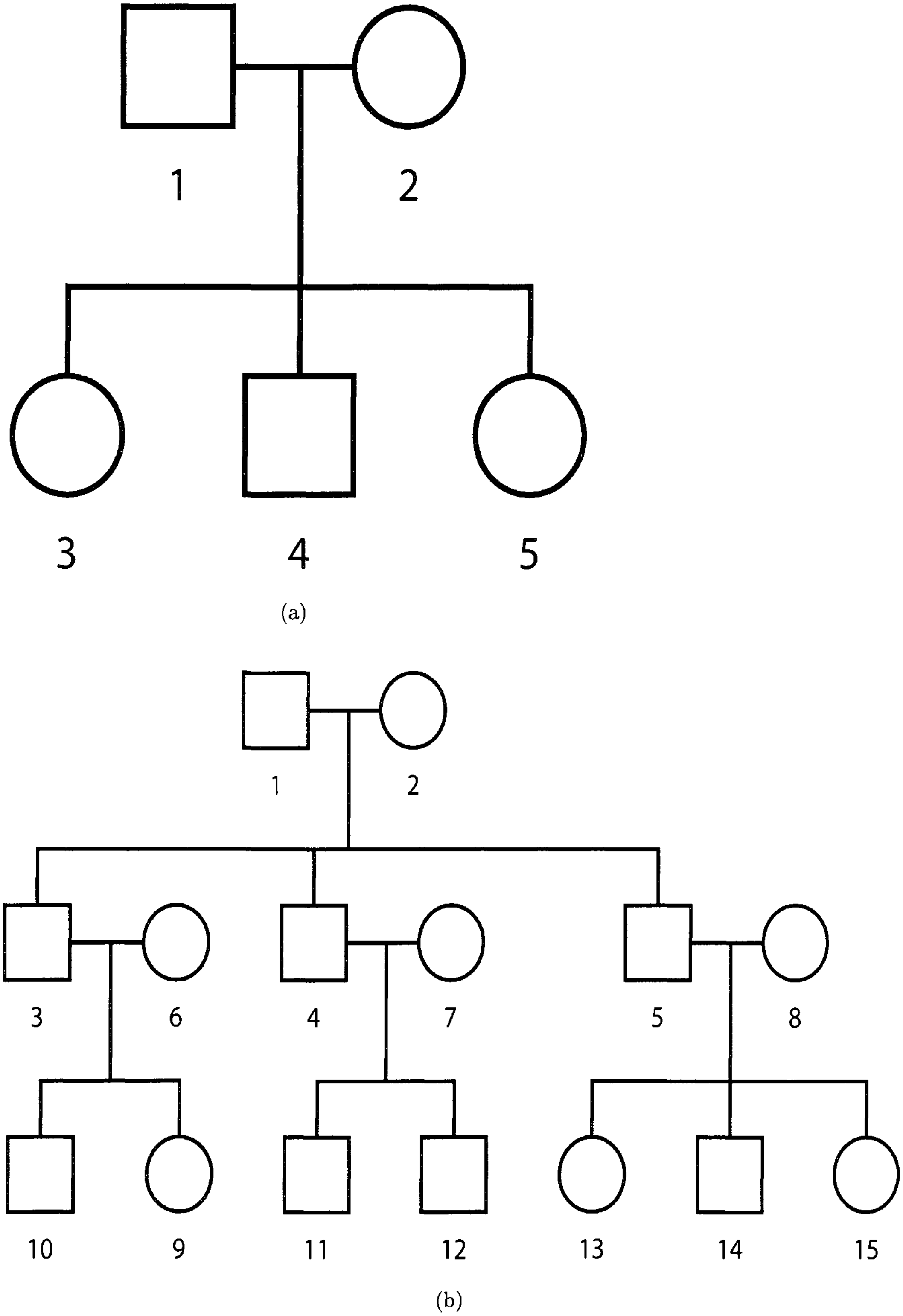
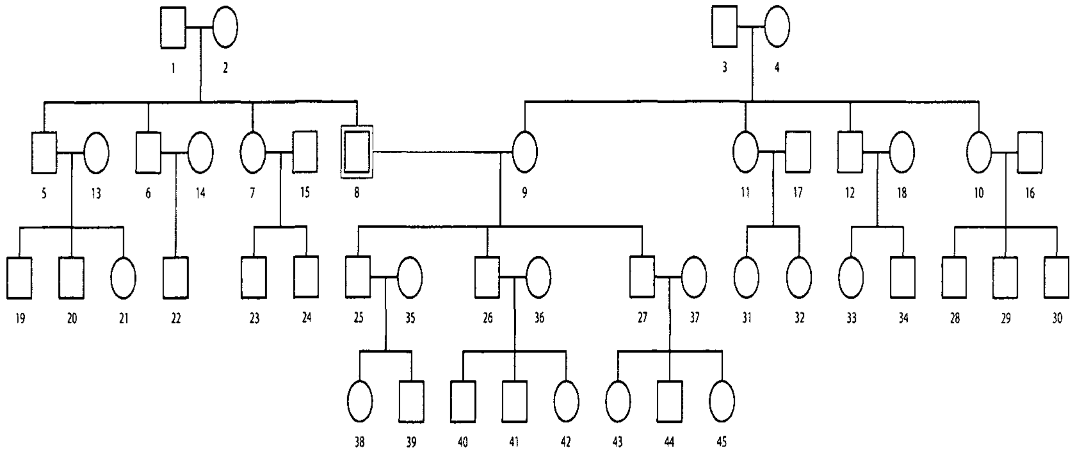


Figure 3.1: Pedigree Configurations Investigated: (a) Pedigree Configuration 1; (b) Pedigree Configuration 2



(c)

Figure 3.1: Pedigree Configurations Investigated: (c) Pedigree Configuration 3

Real Marker	Simulated Marker	Map Position (cM)	Number of Alleles
D16S518	Marker1	0	7
D16S 3096	Marker2	10	10
D16S516	Marker3	20	9
D16S3040	Marker4	30	7
	TRAIT	35	2
D16S507	Marker5	40	9
D16S505	Marker6	50	8
D16S3091	Marker7	60	10
D16S402	Marker8	70	12
D16S3061	Marker9	80	4

Table 3.1: Simulated Genetic Map

3.2.3 True Null Distribution Construction

To construct what we are calling the true null distribution of the qualitative NPL score we simulated 10,000 unlinked pedigrees for each pedigree in the study sample via gene dropping. Each of the 10,000 replicate pedigrees mimicked the corresponding pedigree in the study sample in the affection status and missing data pattern. The NPL statistic was calculated at each location along the marker map for each of the 10,000 pedigrees using Mendel (Lange *et al.*, 2001). The NPL statistics were combined at each locus to build the true null distribution for the overall study sample. The calculated total NPL score for the study sample was tested against this null distribution to determine the number of scores that were greater than or equal to the calculated score. The p-value for the significance of the allele sharing was that count divided by 10,000.

3.2.4 K&C Linear Model Implementation

K&C showed that there is a class of models where the log likelihood of a single free parameter can be written based solely upon the traditional score. They stress that this log likelihood is the exact log likelihood under any missing data pattern. The model is based upon a free parameter that is designated δ, and the test is to see if the maximum likelihood estimate of δ is different from the H_0 that $\delta = 0$. The test is conducted by calculating a score,

$$Z_{lr} = \sqrt{2\left[l(\delta) - l(0)\right]} \tag{3.1}$$

and when the number of pedigrees is large the p-value can be approximated by $1 - \Phi(Z_{lr})$, where Φ is the cumulative distribution of the standard normal distribution. According to the K&C method, the search space for the δ parameter is bounded above by what they call the b-value. The b-value for each pedigree is calculated by the equation

$$b = \frac{\sigma}{\gamma(\mu - a)} \tag{3.2}$$

where a is the smallest theoretical possible value of the score function for a particular pedigree. The b-value that bounds the search space for the δ value is the minimum b-value over all pedigrees in the study sample.

The K&C linear NPL extension was implemented in both Mendel and Simwalk2 (Lange *et al.*, 2001; Sobel and Lange, 1996). The Mendel implementation is the same as described in the original K&C paper (Kong and Cox, 1997). Simwalk2 does not carry out exact calculations; therefore it cannot determine the exact b-value as it is specified in equation (3.2). Therefore we estimate the b-value using the minimum statistic over all sampled descent graphs visited for each pedigree. This minimum b-value is used to bound the search for the K&C δ value. Using the z-scores calculated by Simwalk2 for each pedigree and our estimated b-value we carry out the K&C linear model as described in the original paper (Kong and Cox, 1997).

3.2.5 Program Comparisons

For the qualitative NPL analysis we analyzed results from Genehunter, Mendel, Merlin, and Simwalk2 (Abecasis *et al.*, 2002; Kruglyak *et al.*, 1996; Lange, 2002; Sobel and Lange, 1996). We compared how the programs' results of the NPL_{all} and the NPL_{pairs} scoring methods compared to the results determined by using the true null distributions of the two scores. We consider that a locus has significant evidence for linkage to the trait if the $-LOG_{10}$ of the p-value is $\geq$ 2. The programs Mendel and Simwalk2 weight the z-scores on a per pedigree basis by multiplying them by the square root of the number of affecteds in the pedigree, while the programs Genehunter and Merlin perform no such weighting. In order to accurately compare the results from the programs, we constructed true null distributions for both weighting schemes. Of the programs analyzed, Merlin is the only package that implements the K&C linear method as part of the standard package. Therefore, only the Merlin implementation is compared to the Simwalk2 implementation.

3.3 Qualitative NPL Simulation Results

Dichotomizing a trait and analyzing with the traditional qualitative NPL method is still a powerful method. Our analysis of the qualitative NPL showed that there is not a significant difference in results when using the weighting scheme of Mendel and Simwalk as compared to the non-weighted scheme of Genehunter and Merlin (data not shown). Therefore in the results shown and discussed we have only showed the weighted true distribution and have not shown the results from Genehunter and Merlin's traditional NPL method because their results were virtually identical to Mendel's. For the purpose of brevity we are also only showing and discussing the results from the NPL_{all} scoring function since the same trends and conclusions were made when analyzing the NPL_{pairs} scoring function.

Figure 3.2 shows the results for analyzing the nuclear family in Figure 3.1a. The missing data patterns were formed by zeroing out all the genotypes for one of the parents (20% total missing data), then additionally zeroing out all genotypes

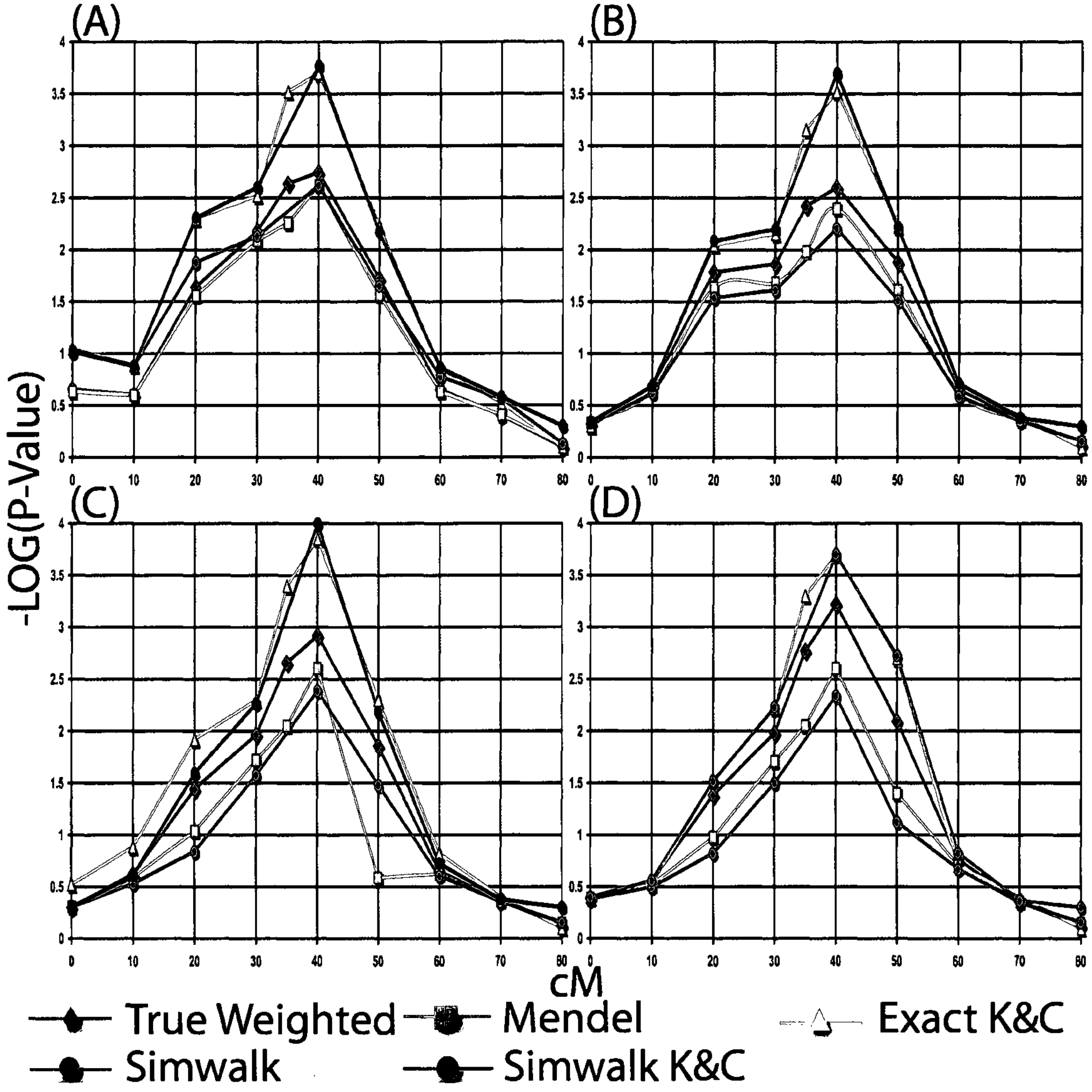


Figure 3.2: Pedigree Configuration 1 NPL Results. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

for Marker3 (30%), and finally zeroing out all genotypes for Marker6 (40%). As Figure 3.2 shows, this family structure contains a lot of information allowing for correct identification of linkage even in the presence of a large quantity of missing information. The results also show that Simwalk's estimation technique does not perform as well as the exact method, but that our implementation of the K&C method in Simwalk performs as well as the exact K&C implementation in Merlin. The results also show that the 'perfect-data approximation' method of determining significance becomes increasingly more conservative as the amount of missing data increases, as shown by the Mendel curves progressive departure from the 'true' distribution curve.

The results of our analysis for pedigree configuration 2 (Figure 3.1b) are shown in Figure 3.3. The missing data patterns were formed by zeroing out all the genotypes for both of the grandparents (20%), additionally zeroing all genotypes for marker5 except for the married-in individuals 6, 7 and 8 and zeroing 5 genotypes for marker1 spread across all generations of the pedigree (30%), and finally zeroing out genotypes for marker3 spread out over all generations (40%). The effects of missing information on the NPL method are very evident in this pedigree configuration. The traditional NPL methods (Mendel & Simwalk) fail to find any significant areas of linkage when the grandparent's genotypes are zeroed, even though the 'true' distribution and K&C implementations find significant evidence for linkage. Even with 30% missing genotypes, the 'true' distribution and the K&C implementations still find significant evidence of linkage. The analysis of this pedigree configuration also shows that Simwalk's implementation of the K&C method is performing as well as the exact implementation, and is a vast improvement over Simwalk's standard analysis.

Figure 3.4 shows the results for analyzing the data set composed of 17 replicates of pedigree configuration 3 (Figure 3.1c). This figure illustrates the tremendous amount of information in these large pedigrees. This analysis also shows that the Simwalk K&C extension is much more robust to missing information than Simwalk's standard NPL method. The missing data patterns were constructed

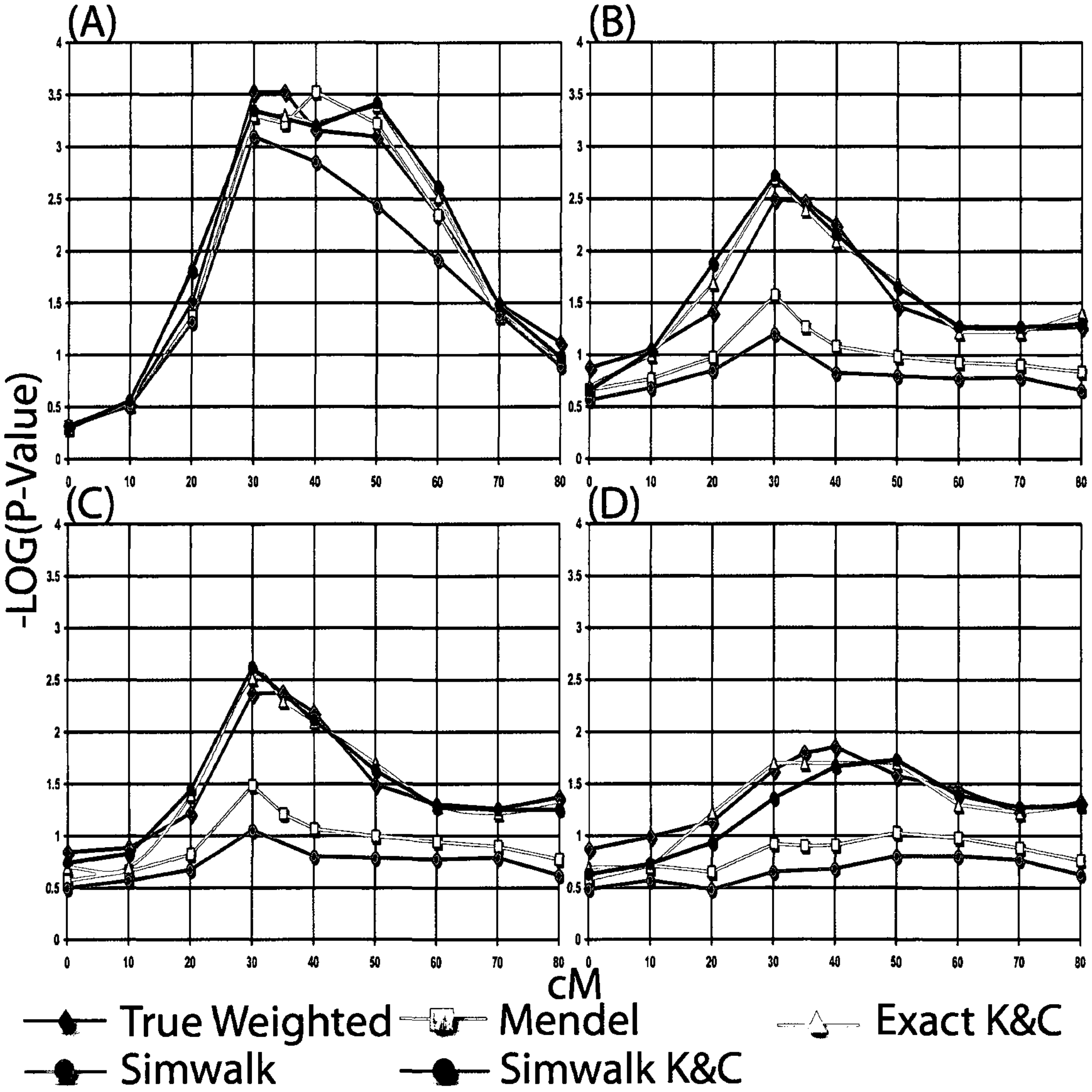


Figure 3.3: Pedigree Configuration 2 NPL Results. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

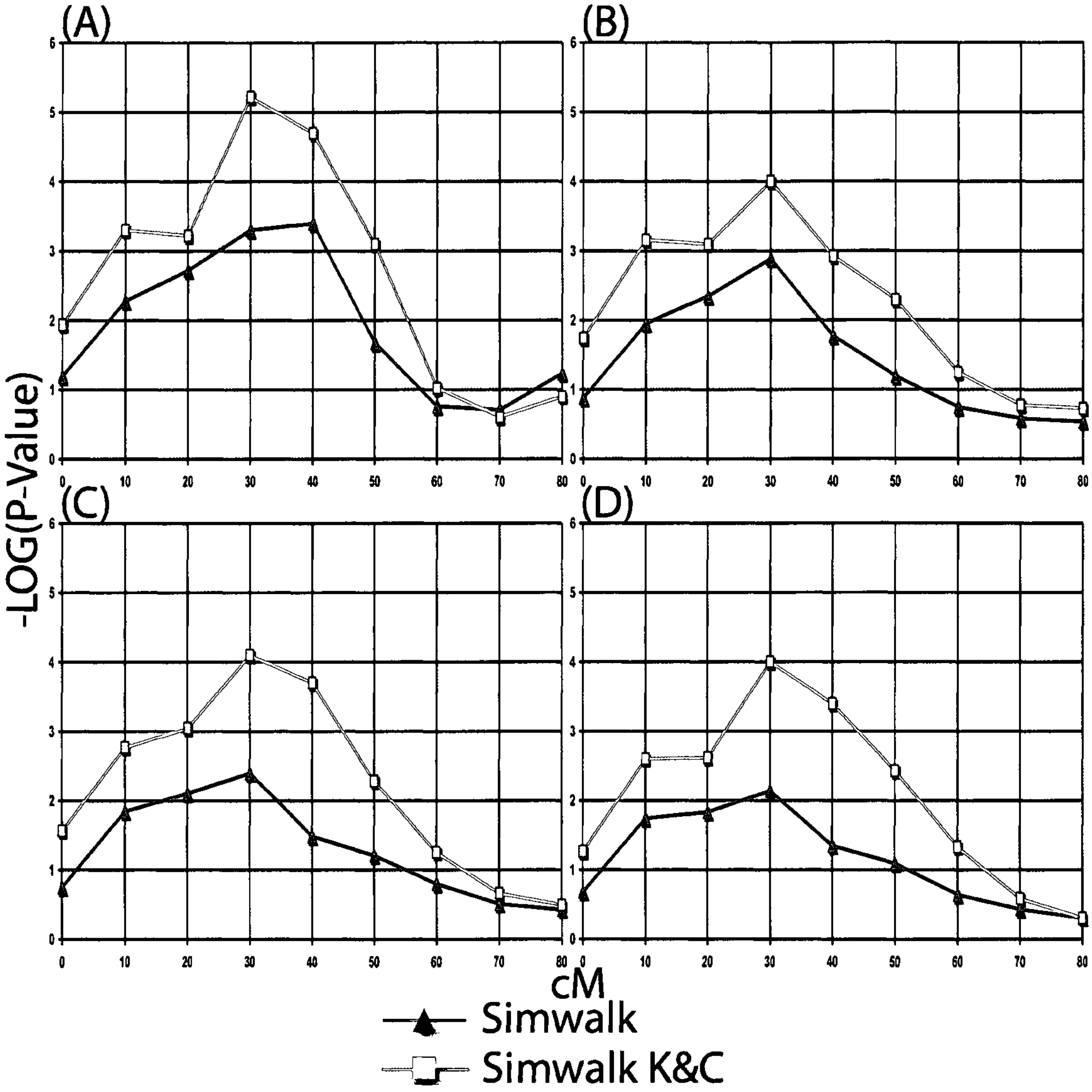


Figure 3.4: Pedigree Configuration 3 NPL Results. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

by zeroing out all the genotypes for 9 individuals at the top of the configuration (20%), additionally zeroing a large fraction of genotypes for marker5 and a small fraction of markerl genotypes (30%), and finally zeroing out more entire individuals at the top of the pedigree (40%). In fact, the Simwalk K&C method shows very little difference in results when analyzing the data set with 20%, 30%

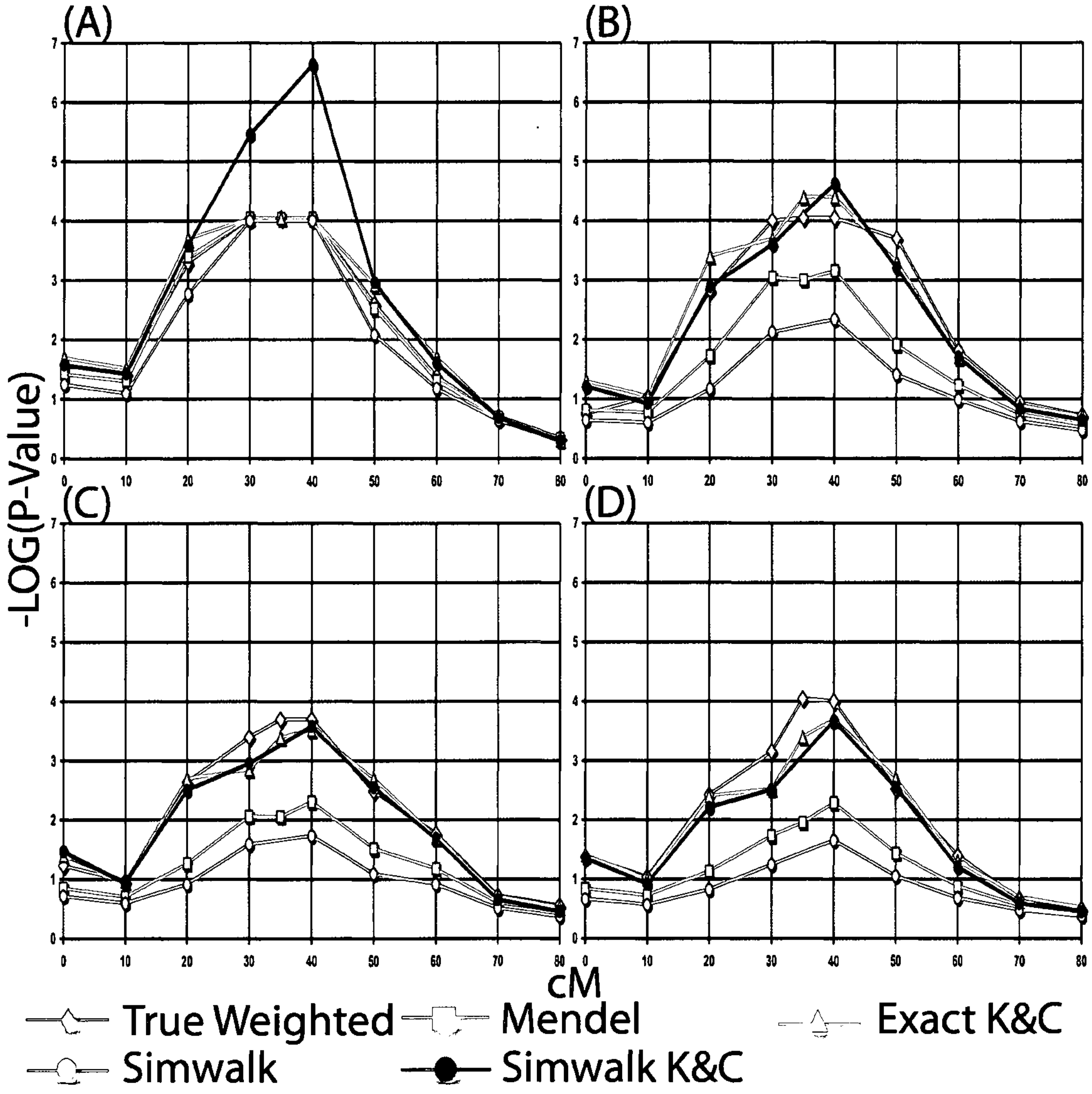


Figure 3.5: Mixed Pedigree Configurations NPL_{all} Results with: (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

and 40% missing genotypes (Figure 3.4 B, C, D). While Simwalk's standard NPL method finds it increasingly more difficult to find the linkage signal with the increasing amounts of missing data.

The final data set analyzed was the mixture configuration described above. The results of analyzing this mixed dataset can be seen in Figure 3.5. Due to the limitations in the exact analysis package Mendel, the 'true' distribution was constructed without analyzing the three large, complex pedigrees in the data set. The results from Figure 3.5 illustrate once again that the K&C method, either exact or Simwalk's, outperforms the traditional NPL method in finding evidence of linkage in the presence of missing data. This analysis also shows the same trend that Simwalk's traditional NPL implementation is the most sensitive to the effects of missing data, and that implementing K&C into Simwalk is a vast improvement and gives comparable results to the exact K&C implementation.

The results from the analysis of the traditional qualitative NPL method shows that the K&C method is a far superior method of significance testing as compared to the perfect data approximation. The K&C method is a good framework to combine the exact analysis of small pedigrees and the estimation methods on large pedigrees. Therefore, we tested to see if the K&C framework works for combining analyses for the traditional NPL. We expect it to work well considering that this is the statistic the K&C method was originally developed for. Figure 3.6 shows a comparison of the K&C results from Simwalk and a combined analysis, where the small pedigrees are analyzed by Mendel and just the 3 large pedigrees are analyzed in Simwalk and the results combined. The figure shows that the results are virtually identical, and very robust to missing data.

3.4 Quantitative NPL Materials and Methods

3.4.1 Variance Component Model Introduction

Variance component methods are very powerful tools for mapping quantitative trait loci (QTL). In the standard variance component model the value of a quan-

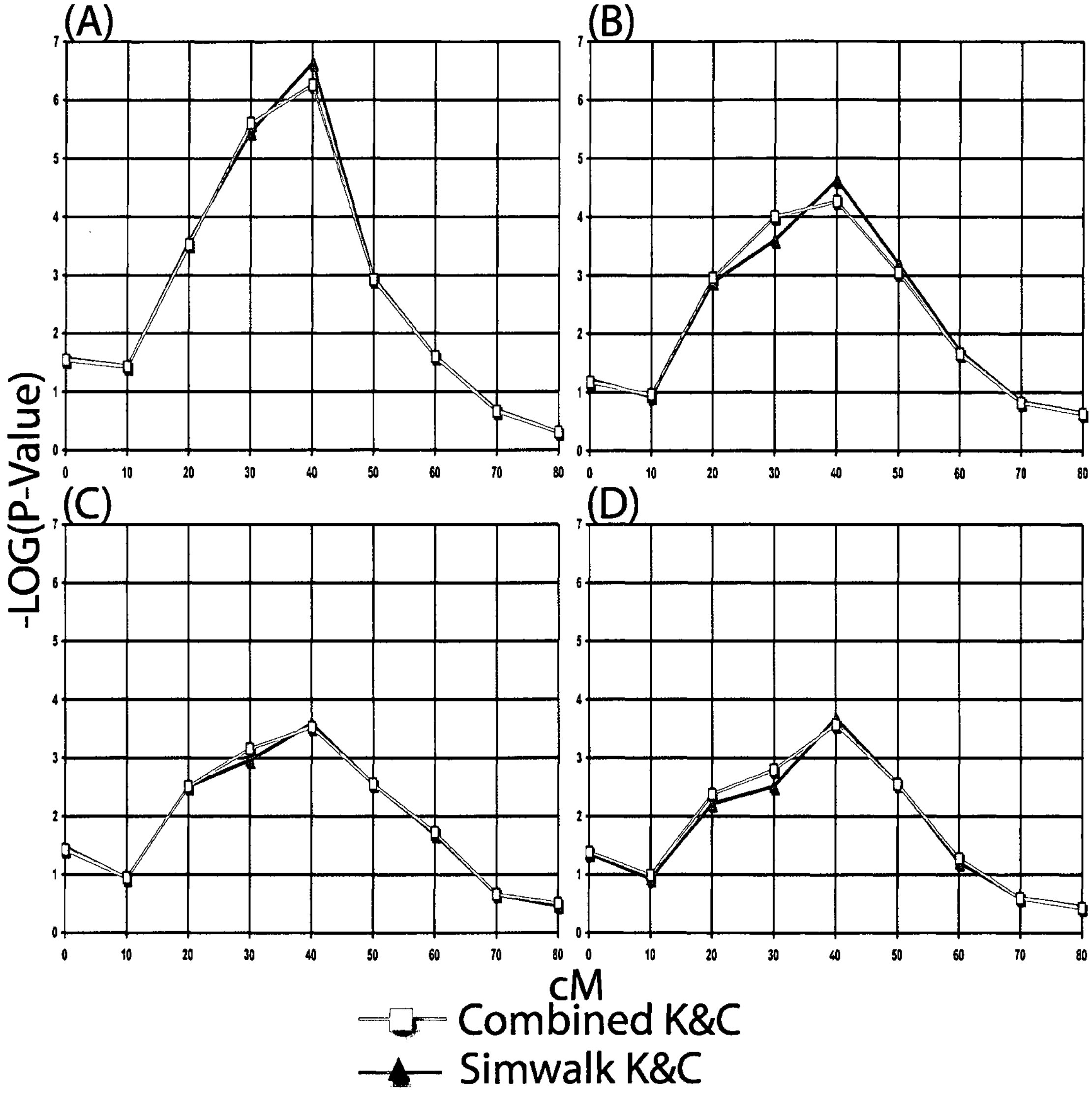


Figure 3.6: Mixed Pedigree Configuration Comparison of K&C NPL Results. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

titative trait y for individual i is modeled as

$$y_i = \mu + \beta^T z_i + a_i + q_i + e_i \tag{3.3}$$

where μ is the population mean of trait y, z are the environmental predictors, a is the polygenic effect, q is the major trait locus effect and e is the residual error. Ignoring dominance effects, the variance of Y is

$$Var(y) = V_a + V_g + V_e \tag{3.4}$$

where V_a is the additive genetic variance, V_g is the polygenic variance and V_e is the environmental variance. For non-inbred relatives i and j the trait's covariance is

$$Cov(y_i, y_j) = 2\hat{\Phi}_{ij} V_a + 2\Phi_{ij} V_g \tag{3.5}$$

where $\hat{\Phi}_{ij}$ is the conditional kinship coefficient for i and j at a map location, and Φ_{ij} is the theoretical kinship coefficient for i and j.

3.4.2 Q-NPL Statistic Definition

The variance component methods as described above rely on a multivariate normal likelihood and suggests the following statistic for capturing differences from the null

$$S = -\frac{1}{2} \ln \det \Omega - \frac{1}{2}(y - \mu)^t \Omega^{-1} (y - \mu), \tag{3.6}$$

where μ is the mean vector and Ω the variance matrix for the trait values. The rationale for this statistic will become clear as we discuss its properties. Typically μ and Ω for the underlying pedigree or pedigrees are estimated from the data by maximum likelihood. Such estimation is computationally expensive and should be avoided whenever p-values are approximated by gene dropping or other sampling procedures. This suggests that we regress the trait values y on covariates prior to analysis and employ a fixed value of Ω. Thus, our first step is to substitute regression-based standardized residuals for the entries of y and omit the trait mean μ from the model.

Using a fixed value of Ω is a less appealing step because a single Ω cannot possibly reflect gradations in marker allele sharing among relatives in different chromosome regions. The resolution to this dilemma is to employ a different Ω for each descent graph at the disease locus. A descent graph, or inheritance vector, determines the gene flow in a pedigree by specifying a grand maternal or grand paternal source for every gamete passed. The probabilities of the different descent graphs at a given genome location can be computed conditional on the marker genotypes observed in the pedigree (Sobel and Lange, 1996). These conditional probabilities can in turn be used to compute the conditional expectation of an IBD (identity by descent) scoring function such as (3.6). The conditional expectations from different pedigrees can be combined into a single NPL test statistic (Lange, 2002; Lange and Lange, 2004).

The question now becomes one of proposing a simple random effects model for the inheritance of the trait conditional on a given descent graph. Such a model has to balance computational efficiency with power to detect linkage. Computational efficiency argues for an additive model because we know only the pattern of gene flow, not what disease allele flows along each descent path. The model should also incorporate random effects. If there are f unrelated founders, then under an autosomal model there are $2f$ ancestral genes and consequently $2f$ random effects. Under Hardy-Weinberg and linkage equilibrium, these random effects are uncorrelated. These considerations prompt the model

$$\Omega = \sigma_a^2 \sum_{s=1}^{2f} u_s u_s^t + \sigma_e^2 I, \tag{3.7}$$

where I is the identity matrix, σ_a^2 is the additive variance, σ_e^2 is the environmental variance and the ith component u_{si} of the vector u_s has value 0, 1, or 2 depending on whether person i inherits 0, 1, or 2 genes from founder source s.

Several assumptions are implicit in this model. One is the incorporation of random environment via the term $\sigma_e^2 I$. A second is that a non-inbred person has total variance $2\sigma_a^2 + \sigma_e^2$. Because y is standardized, it is natural to assume $2\sigma_a^2 + \sigma_e^2 = 1$ and define the heritability $h^2 = 2\sigma_a^2$. This convention allows us to rewrite equation (3.7) as

$$\Omega = 2h^2 \sum_{s=1}^{2f} v_s v_s^t + (1 - h^2)I, \tag{3.8}$$

where the possible values of the components of v_s are 0, $\frac{1}{2}$, and 1 instead of 0, 1, and 2, respectively.

A third assumption is the neglect of additive polygenic effects. This sounds like a serious omission, but it is worth recalling that we are engaged in hypothesis testing not parameter estimation. Although it is true that additive polygenic and major gene contributions will be confounded, we are looking for excessive contributions by a major gene. The average value of Ω in the absence of marker data captures polygenic inheritance. If a major gene is present and marker coverage is good, then the conditional probabilities of one or just a few descent graphs will dominate. For these descent graphs, we want a variance decomposition that favors contributions by the major locus. The is exactly what model (3.8) achieves.

Computational speed is another reason for preferring the model. Inspection of equation (3.8) shows that Ω is a rank $2f$ perturbation of a diagonal matrix. This fact facilitates quick evaluation of $\det \Omega$ and Ω^{-1}. Ordinarily evaluation of these entities requires on the order of n^3 arithmetic operations when Ω is $n \times n$. However, the Sherman-Morrison algorithm brings these operation counts down to a multiple of n^2 (Millar, 1987). In practice, we apply each of the two identities

$$\begin{aligned} (M + ww^t)^{-1} &= M^{-1} - \tfrac{1}{1+w^t M^{-1} w} M^{-1} w w^t M^{-1} \\ \det(M + ww^t) &= \det M \det(1 + w^t M^{-1} w) \end{aligned}$$

$2f$ times, starting from $M = (1 - h^2)I$ and $\det M = (1 - h^2)^n$.

We have another reason for preferring the statistic (3.6) under the decomposition (3.8). Suppose we consider this statistic with a generic variance matrix Σ substituted for Ω. Then on average, conditional on the given descent graph, the statistic has value

$$\begin{aligned} E(S) &= -\tfrac{1}{2} \ln \det \Sigma - \tfrac{1}{2} E\left[(y-\mu)^t \Sigma^{-1} (y-\mu)\right] \\ &= -\tfrac{1}{2} \ln \det \Sigma - \tfrac{1}{2} tr\left\{\Sigma^{-1} E\left[(y-\mu)(y-\mu)^t\right]\right\} \\ &= -\tfrac{1}{2} \ln \det \Sigma - \tfrac{1}{2} tr(\Sigma^{-1}\Omega) \end{aligned}$$

Using standard results from multivariate normal theory, $E(S)$ is maximized by taking $\Sigma = \Omega$. Finally, we anticipate that the statistic from equation (3.6) will perform well because it is closely related to the Mahalanobis statistic $(y -$

$\mu)^t\Omega^{-1}(y-\mu)$ prominent in hypothesis testing and outlier detection.

3.4.3 Pedigree Configurations and Data Simulation

The same three pedigree configurations used in the analysis of the qualitative NPL were used in the analysis of the quantitative NPL (Figure 3.1). We simulated the same 9-marker microsatellite map as for the qualitative NPL as seen in Table 3.1. We simulated a quantitative trait locus located half way between M4 and M5 at 35cM. The locus at 35cM has two alleles, a major allele with frequency 0.8905 and a minor allele with frequency 0.1095. We modeled the quantitative trait as coming from 3 normal distributions based upon the individuals' genotype at the trait locus. The mean of the trait distribution for individuals homozygous for the major allele was equal to 2 (μ_{AA}), the mean for heterozygous individuals was equal to 0 ($\mu_{Aa,aA}$), and the mean of the trait for individuals homozygous for the minor allele was equal to -2 (μ_{aa}). This gave an overall trait mean (μ_{global}) of 1.56 by solving equation (3.9), where P_A is the allele frequency of the major allele and P_a is the allele frequency of the minor allele.

$$\mu_{global} = P_A^2(\mu_{AA}) + 2P_AP_a(\mu_{Aa,aA}) + P_a^2(\mu_{aa}) \tag{3.9}$$

The additive genetic variance of the trait V_A is given by equation (3.10),

$$V_A = P_A^2(\mu_{AA} - \mu_{global})^2 + 2P_AP_a(\mu_{Aa,aA} - \mu_{global})^2 + P_a^2(\mu_{aa} - \mu_{global})^2, \tag{3.10}$$

and is equal to 0.7831. The variance of the genotype specific distributions,

V_E, was determined by solving equation (3.11) for V_E, where h^2 is the heritability of the trait.

$$h^2 = \frac{V_A}{V_A + V_E} \tag{3.11}$$

We studied the trait with heritability equal to 10%, 25% and 50%. We analyzed the exact same pedigrees as were analyzed in the qualitative NPL work above. Individuals in these pedigrees were assigned a quantitative trait value based upon their genotype at the putative trait locus, and the models specified above. The pedigrees initially had no missing genotype information, but not necessarily complete phase information. The missing genotype patterns were developed based upon looking at the types of missing genotype patterns that were present in the 73 families investigated in the previous study (Pajukanta *et al.*, 2003). The study samples for each investigation were same as the study samples used in the qualitative investigations above.

The proposed Q-NPL statistic was analyzed under the null hypothesis of no linkage to determine its biases, if any, and type I error rate. For each data set and for each heritability value we generated 1000 replicates via gene dropping. For each replicate the individuals were assigned trait values as follows: P_A^2 of the assigned values were sampled from the distribution with mean 2, $2P_AP_a$ of the assigned values were sampled from the distribution with mean 0, and P_a^2 of the assigned values were sampled from the distribution with mean -2. Each replicate was analyzed using statistic (3.6) and the p-value for each marker for each replicate was stored.

3.4.4 True Null Distribution Construction

To construct the true null distribution of the quantitative NPL score we simulated 10,000 unlinked pedigrees for each pedigree in the study sample via gene dropping. Each of the 10,000 replicate pedigrees mimicked the corresponding pedigree in the study sample in the trait value and missing data pattern. The NPL statistic was calculated at each location along the marker map for each of the 10,000 pedigrees using Mendel (Lange *et al.*, 2001). The NPL statistics were combined at each locus to build the true null distribution for the overall study sample. The calculated total NPL score for the study sample was tested against this null distribution to determine the number of scores that were greater than or equal to the calculated score. The p-value for the significance of the allele sharing was that count divided by 10,000.

3.4.5 Implementation of Qualitative K&C Linear Model

The original K&C extension assumed that we are limited to utilizing the score generated by traditional NPL methods (Kong and Cox, 1997). But they state that the score used can be any function that has a higher expected value under linkage than under no linkage, which our proposed Q-NPL score satisfies and therefore allows us to use the K&C extension here as well. The K&C linear model is implemented in Mendel exactly as described in the original paper except we substitute the score from our new statistic for Mendel's calculated z-score from the traditional NPL method (Kong and Cox, 1997; Kruglyak *et al.*, 1996). The Simwalk2 implementation is the same as described above for the qualitative NPL, except we substitute the score from our new statistic for Simwalk's calculated z-score from the traditional method.

3.4.6 Program Comparisons

Mendel and Simwalk2 are the only programs that implement the new Q-NPL method, therefore only those programs and the K&C extensions are analyzed against the true null distribution of the score shown in equation (3.6).

3.5 Quantitative NPL Simulation Analysis

The first step in evaluating a new statistic is examining its behavior under the null hypothesis to assure that it is unbiased and correctly controls the type I error rate. The null hypothesis for each marker is that it is not linked to the quantitative trait under investigation. To show that the statistic is performing correctly under the null hypothesis the distribution of p-values for each marker should be uniform. Figure 3.7 shows the distribution of p-values for Marker5 from 1000 replicates of pedigree configuration 2 (Figure 3.1b) simulated with heritability equal to 50%. As the figure illustrates the distribution of p-values is approximately uniform and therefore the statistic is performing correctly under the null hypothesis. All markers under all conditions investigated showed a similar approximate uniform distribution of p-values as seen in Figure 3.7. Now that we know the statistic proposed in equation (3.6) behaves appropriately under the null hypothesis, we need to determine its power to detect linkage.

We first tested the power of the new Q-NPL statistic to detect linkage in a sample of 51 nuclear families (Figure 3.1a). We simulated the quantitative trait with heritability equal to 10%, 25% and 50%. The Q-NPL statistic was not

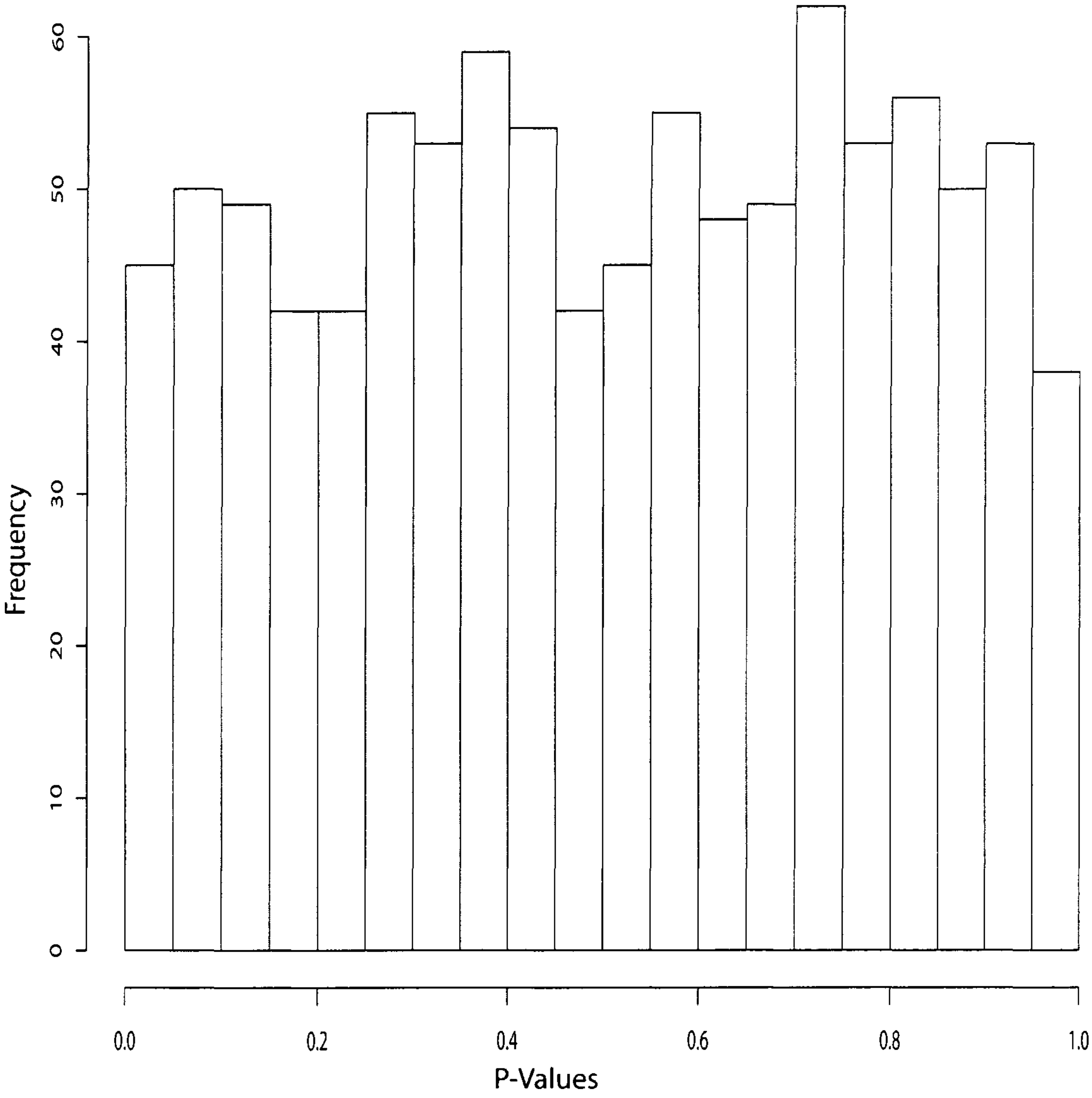


Figure 3.7: Q-NPL P-Value Distribution
Pedigree Configuration 2, Marker5, 50% Heritability.

able to detect any significant linkage signal with a trait with 10% heritability. The Q-NPL statistic does have power to detect a significant linkage signal with heritability of 25% and 50%. The results for the analysis of the trait with 25% heritability is seen in Figure 3.8. The results illustrate that the exact Q-NPL statistic as calculated by Mendel shows good power and proves to be very robust against missing genotype data. The missing data patterns were formed as described above for the qualitative trait analysis. Figure 3.8 also demonstrates that the K&C extension to the new Q-NPL statistic is virtually identical to the true null distribution we created through extensive simulation. The K&C method does not require any sampling to create the null distribution therefore providing a huge computational savings. Neither the standard Simwalk2 nor Simwalk2 K&C extension shows much power at this level of heritability, and they both are very sensitive to missing genotype information. Both Mendel and Simwalk2 show good power to detect the linkage signal when the trait is simulated with 50% heritability as seen in Figure 3.9. Again the K&C extensions are the most robust statistics against missing genotype information and closely mirror the curve from the 'true' null distribution.

We next tested the Q-NPL's performance in a sample of 14 3-generation pedigrees (Figure 3.1b). None of the implementations were able to detect a linkage signal when the trait was simulated with 10% heritability. The results for the quantitative trait simulated with 25% and 50% heritability can be seen in Figures 3.10 and 3.11, respectively. The missing data patterns were formed as described above for the qualitative analysis. As Figure 3.10(A) illustrates all the implementations have very good power to detect the linkage signal with a heritability of 25% and no missing genotypes, but as panels (B)-(D) reveal

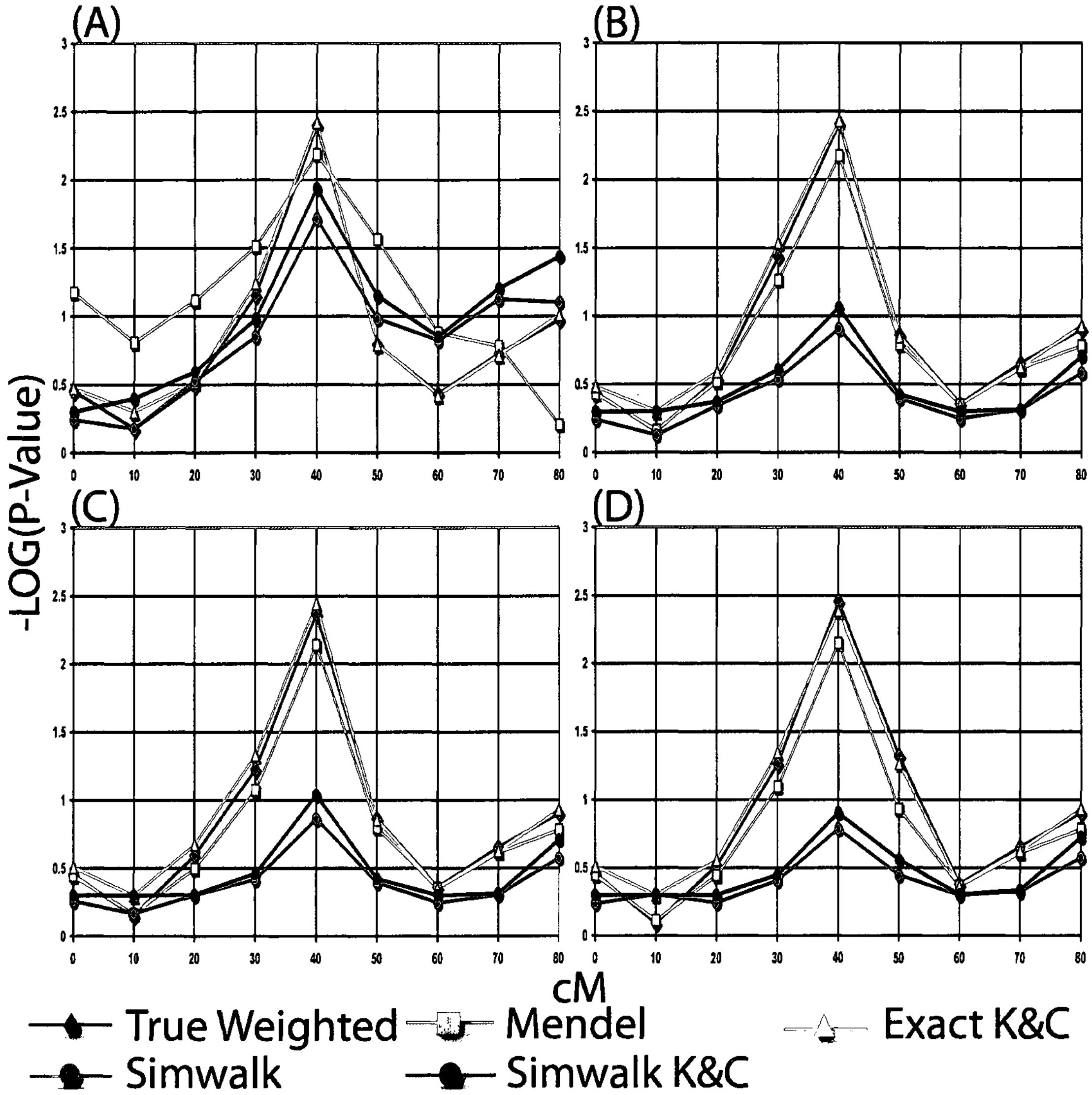


Figure 3.8: Pedigree Configuration 1 Q-NPL Results 25% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

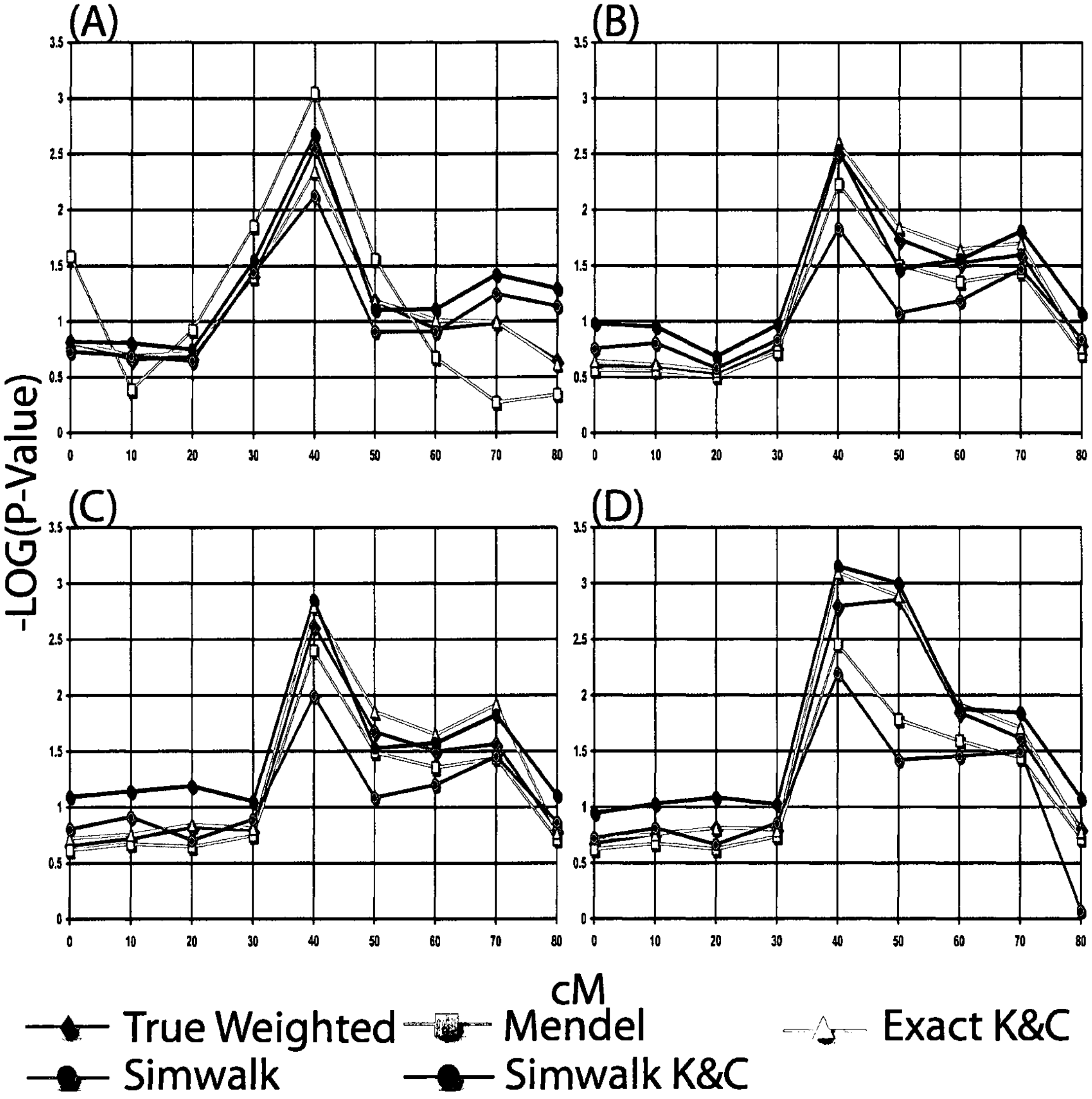


Figure 3.9: Pedigree Configuration 1 Q-NPL Results 50% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

none of the applications can detect the signal with significant amounts of missing genotypes. When the trait is simulated with 50% heritability and no missing genotypes all the methods are extremely powerful in detecting the signal as Figure 3.11(A) illustrates. In fact the plateau of significance for the true null, Mendel and Simwalk occurs at the limit of detection due to the number of replicates used in constructing the testing distribution. Panels (B)-(D) of Figure 3.11 illustrates that Simwalk2 is the most sensitive to missing genotype information, but that the K&C extension for Simwalk2 is much more robust to missing genotypes. The exact implementation in Mendel and the Mendel K&C extension are also robust to missing genotypes, achieving the significant linkage signal threshold of ≥ 2 with as much as 30% missing genotypes.

Figure 3.1c is an extended 45 individual, 5 generation pedigree structure that is too large to be handled by exact methods of analysis. We analyzed a sample of 17 such pedigrees using Simwalk. Simwalk was able to detect the linkage signal with the heritability of the quantitative trait equal to 10%, 25% and 50%. Figures 3.12 and 3.13 show the results of the analysis with heritability equal to 10% and 25% heritability respectively. The missing data patterns were constructed as described above for the qualitative analysis. As the figures demonstrate these very large extended pedigrees contain a lot of information and therefore are able to detect the linkage signal for a quantitative trait with only 10% heritability (Figure 3.12A-D). Simwalk and especially the K&C extension are very robust to the missing genotypes. The curve for the 10% heritability and 30% missing genotypes (Figure 3.12C) looks odd and may just be an artifact of the way in which we generated the missing genotype pattern. When the trait is simulated with 25% heritability, the signal is so strong that Simwalk's significance curve plateaus at

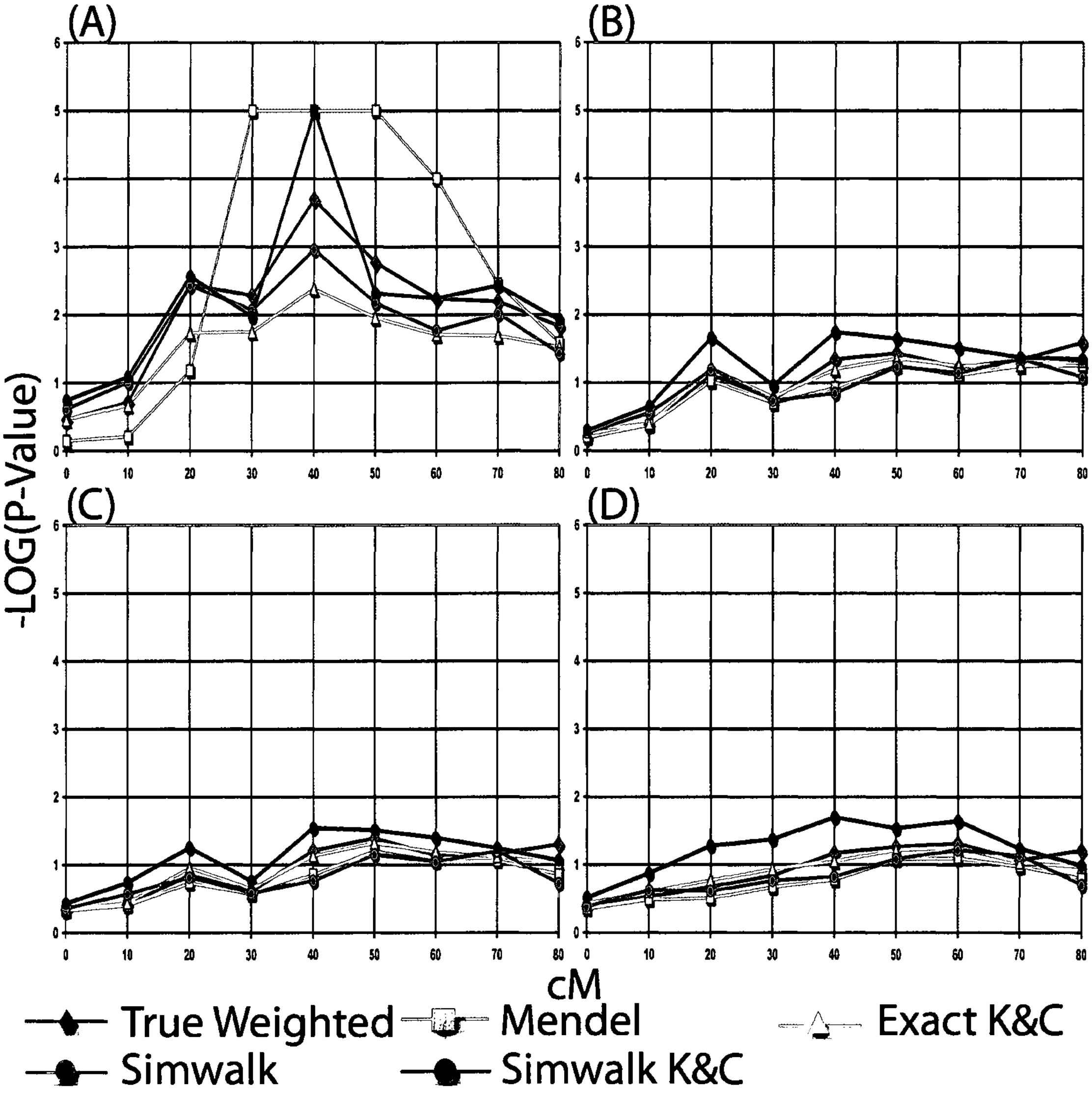


Figure 3.10: Pedigree Configuration 2 Q-NPL Results 25% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

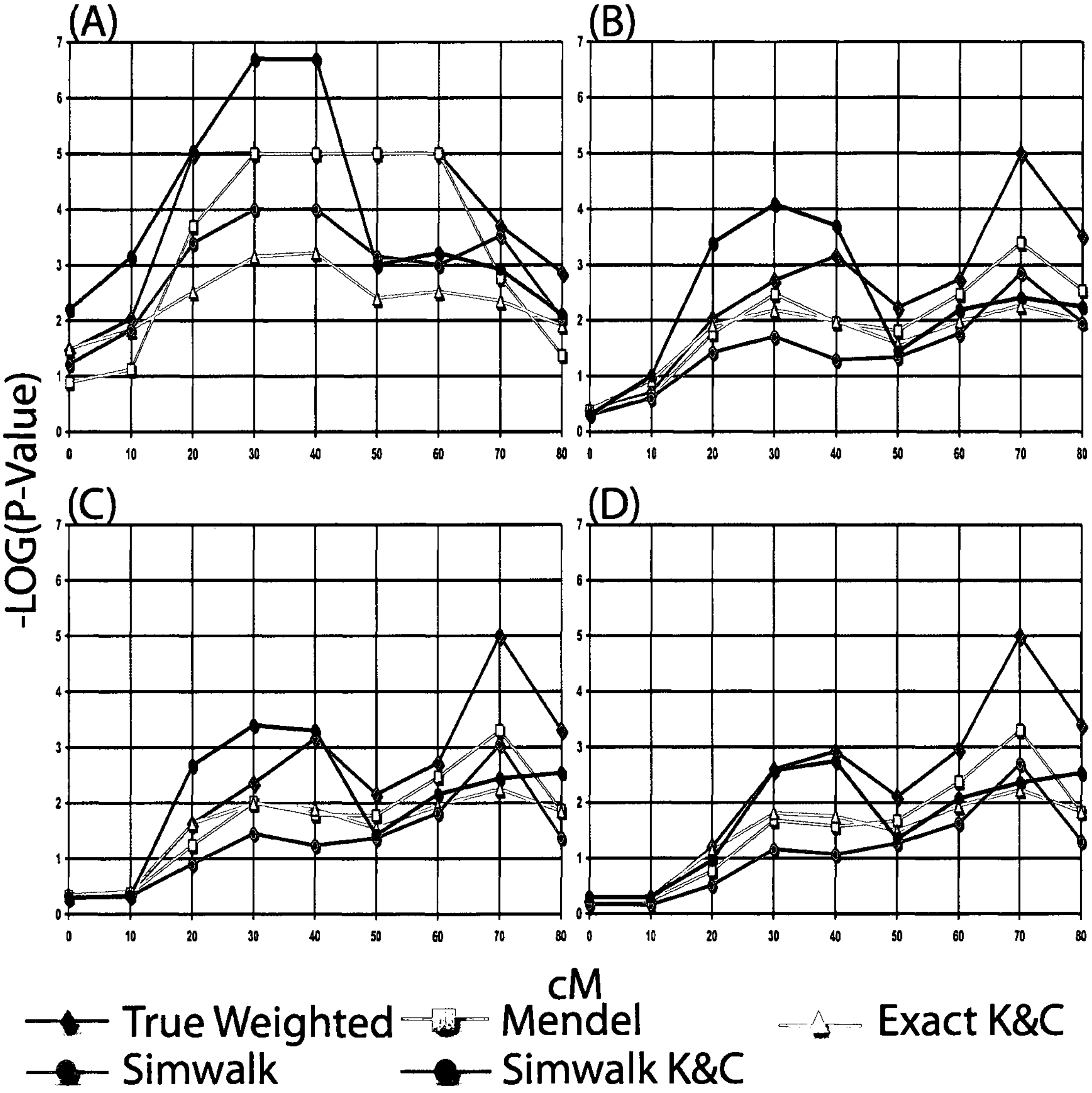


Figure 3.11: Pedigree Configuration 2 Q-NPL Results 50% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

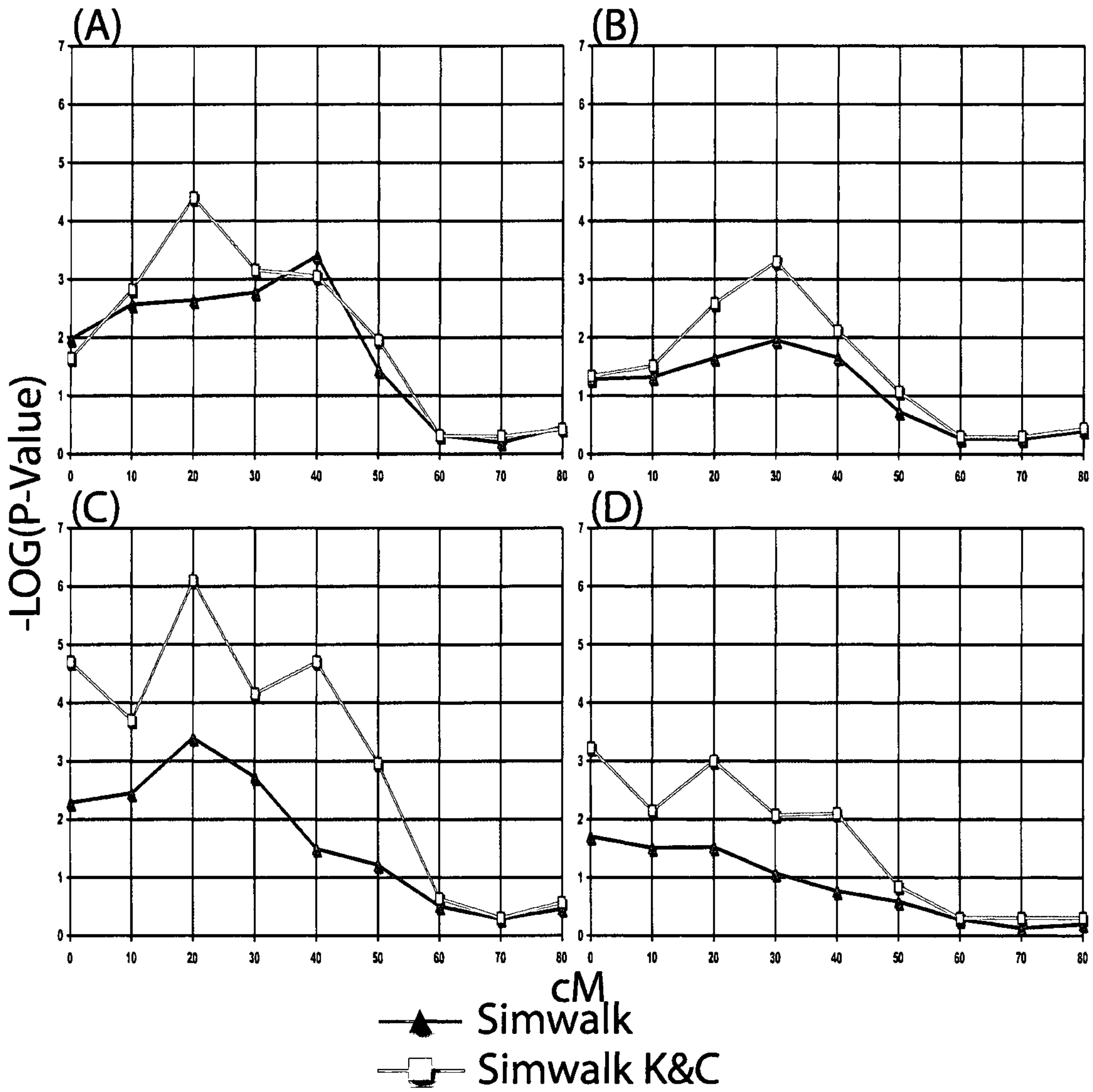


Figure 3.12: Pedigree Configuration 3 Q-NPL Results 10% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

4, its significance limit due to the number of replicates used to construct its null distribution. Again the K&C extension does not have this constraint because its significance test is based upon the standard normal distribution and does not require sampling.

Real data sets are commonly composed of a variety of pedigree structures,

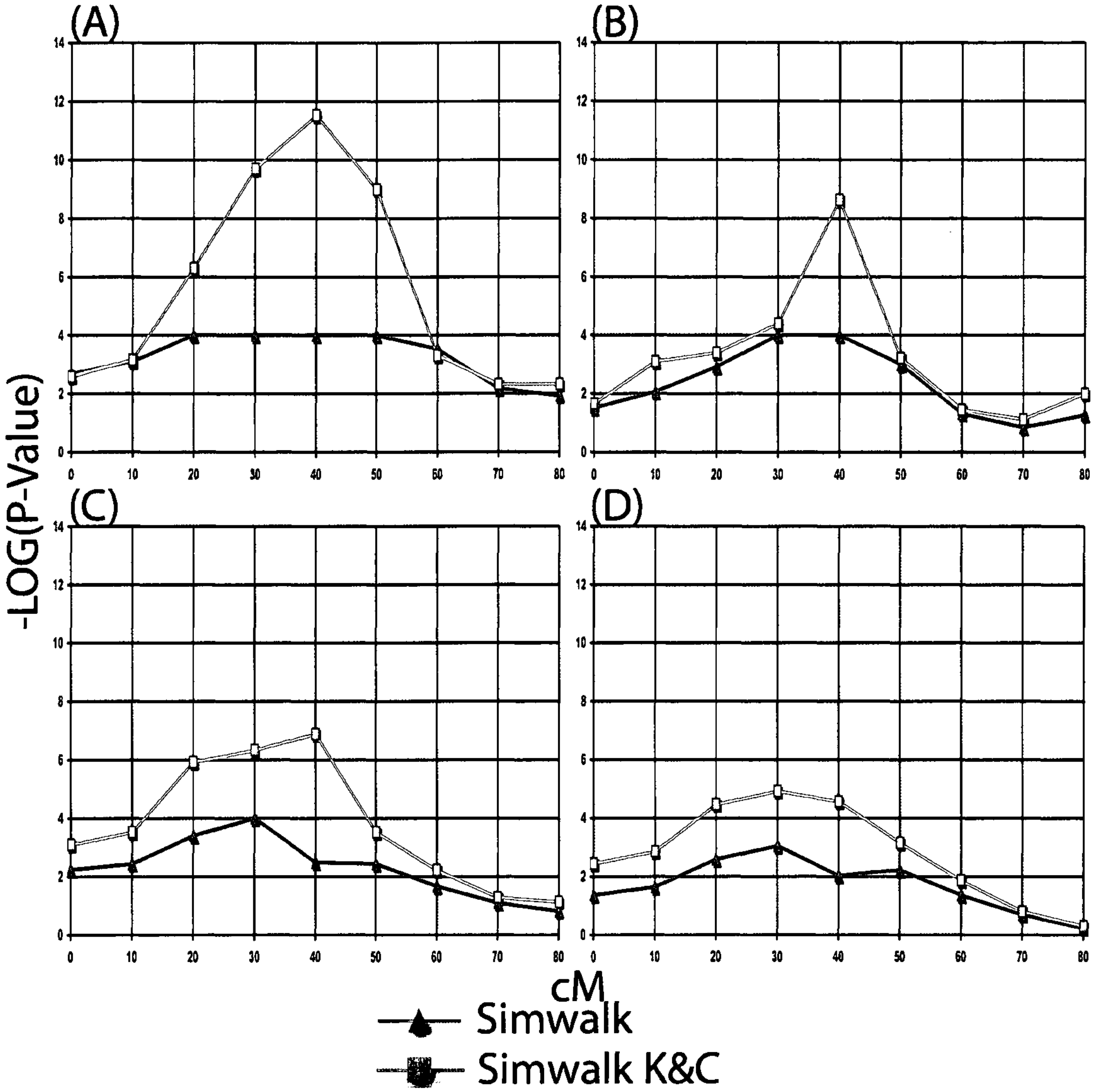


Figure 3.13: Pedigree Configuration 3 Q-NPL Results 25% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

therefore we tested the Q-NPL's power in such a sample. We constructed a data set that contains various numbers of the above data structures as described in Materials and Methods. This mixture configuration contains 3 pedigrees that are too complex to be analyzed exactly and therefore were not included in the calculations for the curves produced by Mendel or the true null distribution. The mixture configuration was analyzed with heritability of the trait equal to 10% (Figure 3.14), 25% (Figure 3.15), and 50% (Figure 3.16). The missing data patterns were the same as described above for each individual pedigree structure. As the figures illustrate, the Q-NPL statistic is able to detect a linkage signal for all three heritabilities given enough genotype information. The signal with heritability equal to 10% drops off dramatically with missing genotype information, with Simwalk being the most sensitive. Even though with missing genotypes of 20%, 30% and 40% none of the methods reach our significance threshold of 2, the Mendel K&C extension follows the true null distribution under all conditions. With a heritability of 25% the Q-NPL statistic is able to detect the linkage signal under all conditions of missing genotype information, again with Simwalk being the most sensitive to the missing genotypes. The linkage signal is so strong with heritability of 50% that all methods perform well under all missing data patterns. Figure 3.16 also demonstrates that there is significant information in the 3 large pedigrees that Mendel cannot handle, but that Simwalk is able to include.

The K&C framework allows the possibility to combine the exact calculation of Mendel on small pedigrees with the estimation method in Simwalk on large pedigrees and return a single significance score. This is accomplished by using the Z-scores and b-values from the separate programs and using them in the K&C method as if they came from a single program. This framework makes

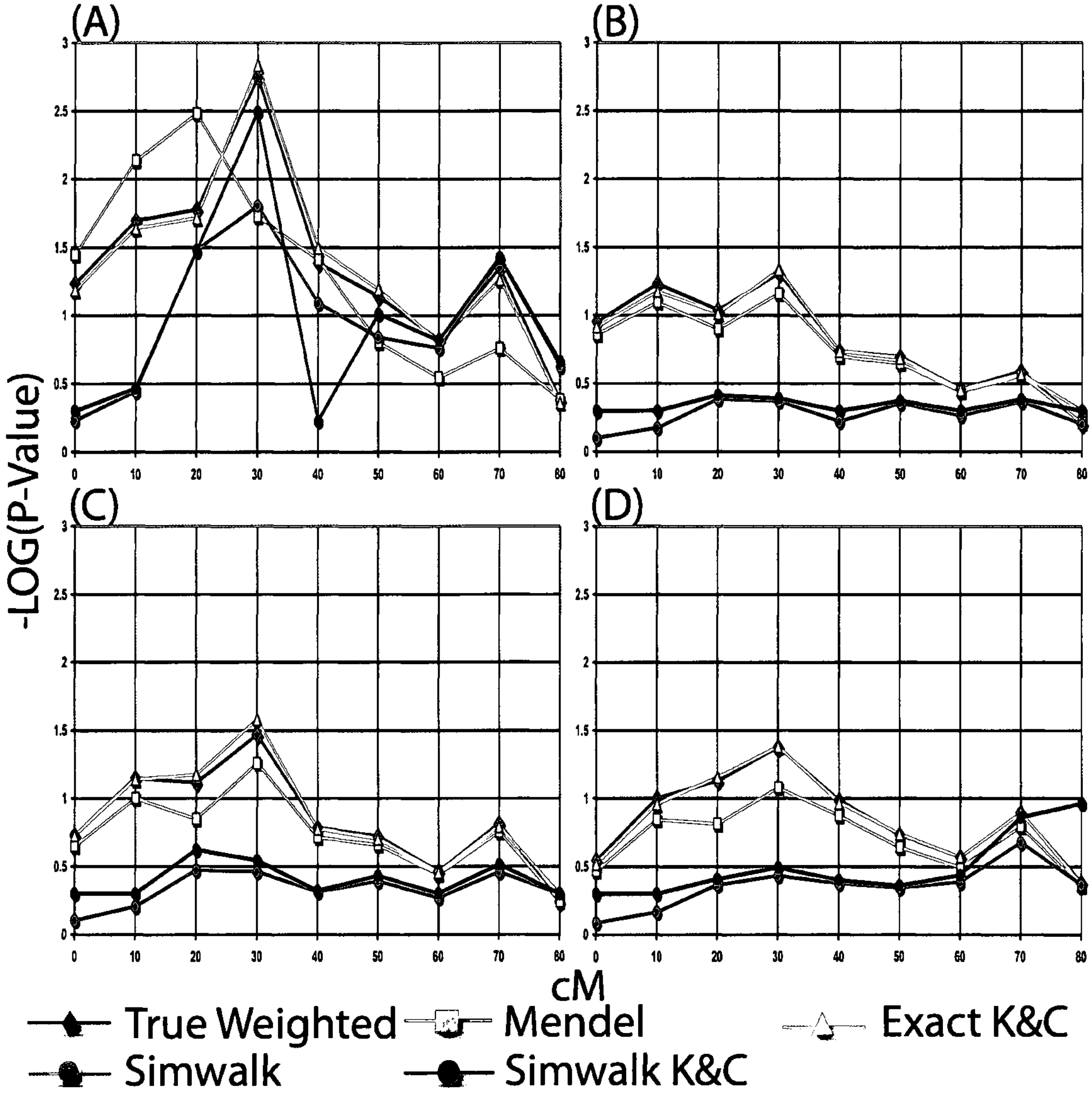


Figure 3.14: Mixed Pedigree Configuration Q-NPL Results 10% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

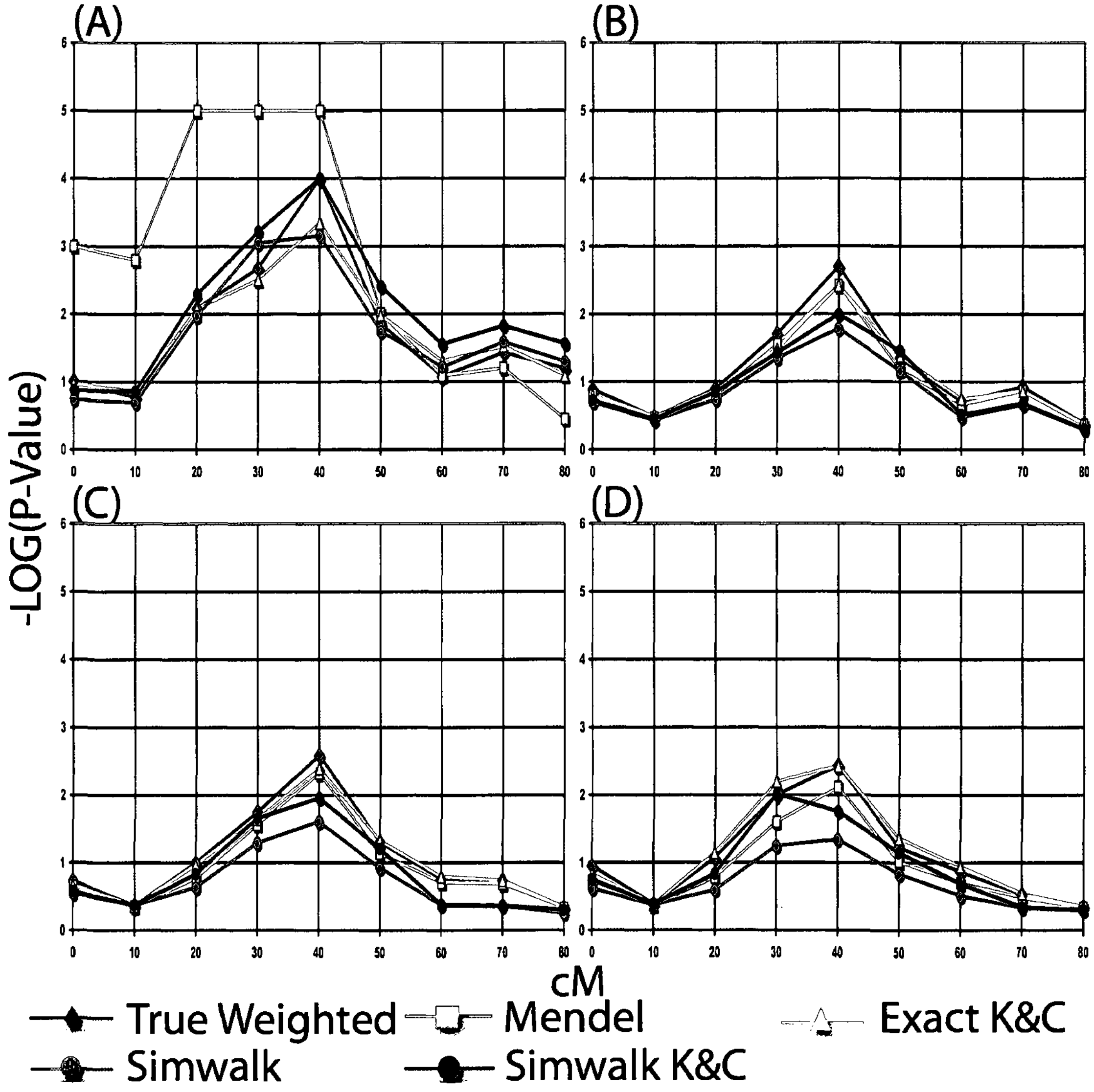


Figure 3.15: Mixed Pedigree Configuration Q-NPL Results 25% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

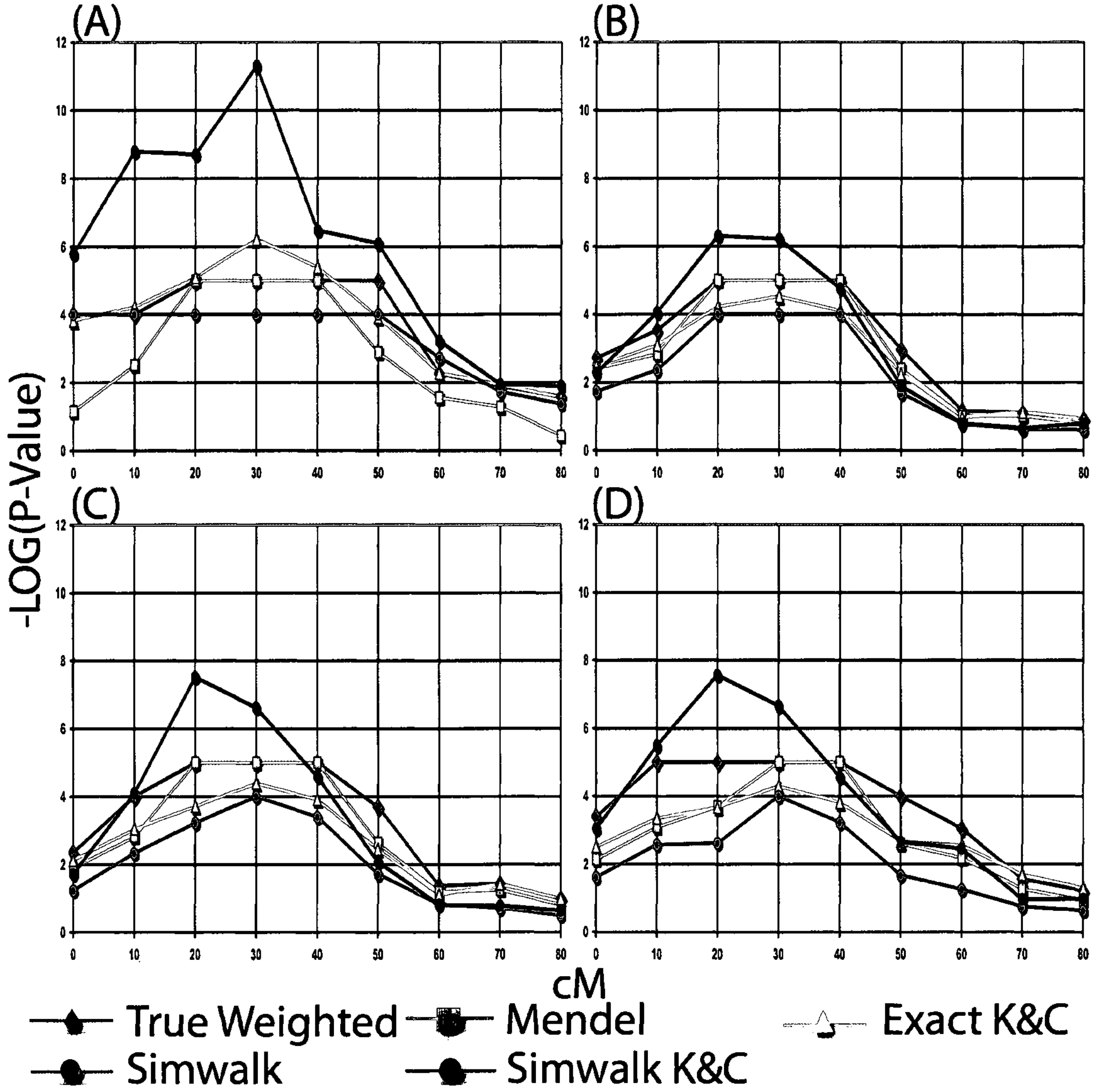


Figure 3.16: Mixed Pedigree Configuration Q-NPL Results 50% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

sense in the context of the mixture configuration. To test this idea we used the mixture configuration with 25% heritability and analyzed the small pedigrees in Mendel and just the 3 large pedigrees in Simwalk, combined the results and ran the K&C method. The results of this method are seen in Figure 3.17, where Combined K&C represents this new framework. The combining of the methods works well and is very robust to missing genotypes. This is also a very appealing framework because it alleviates the computational burden of sampling to test for significance.

3.6 Real Data NPL Analysis

The results of the simulation studies are striking, but an important question is how relevant are they to the analysis of real data. To investigate this we analyzed the data set from a previously published gene-mapping study that used traditional NPL analysis (Pajukanta *et al.*, 2003). The published results were based upon a sample of 73 pedigrees with sizes from 4 to 59 individuals. The overall percentage of missing genotypes in the study sample was approximately 56%, with individual pedigrees having percentages from 25% to 94% missing genotypes. One reason for this seemingly very high missing data rate, is that the dataset was not collected as a single data sample, but instead it was 3 separate studies combined into a single dataset. Of the nine markers investigated in our analysis one sample was genotyped at 6 markers, another at 7 markers, and the third at 8 markers. Therefore to get a better idea of missing data in each of the designed investigations we separated the dataset into its 3 independent datasets and re-analyzed the amount of missing genotypes. The results were

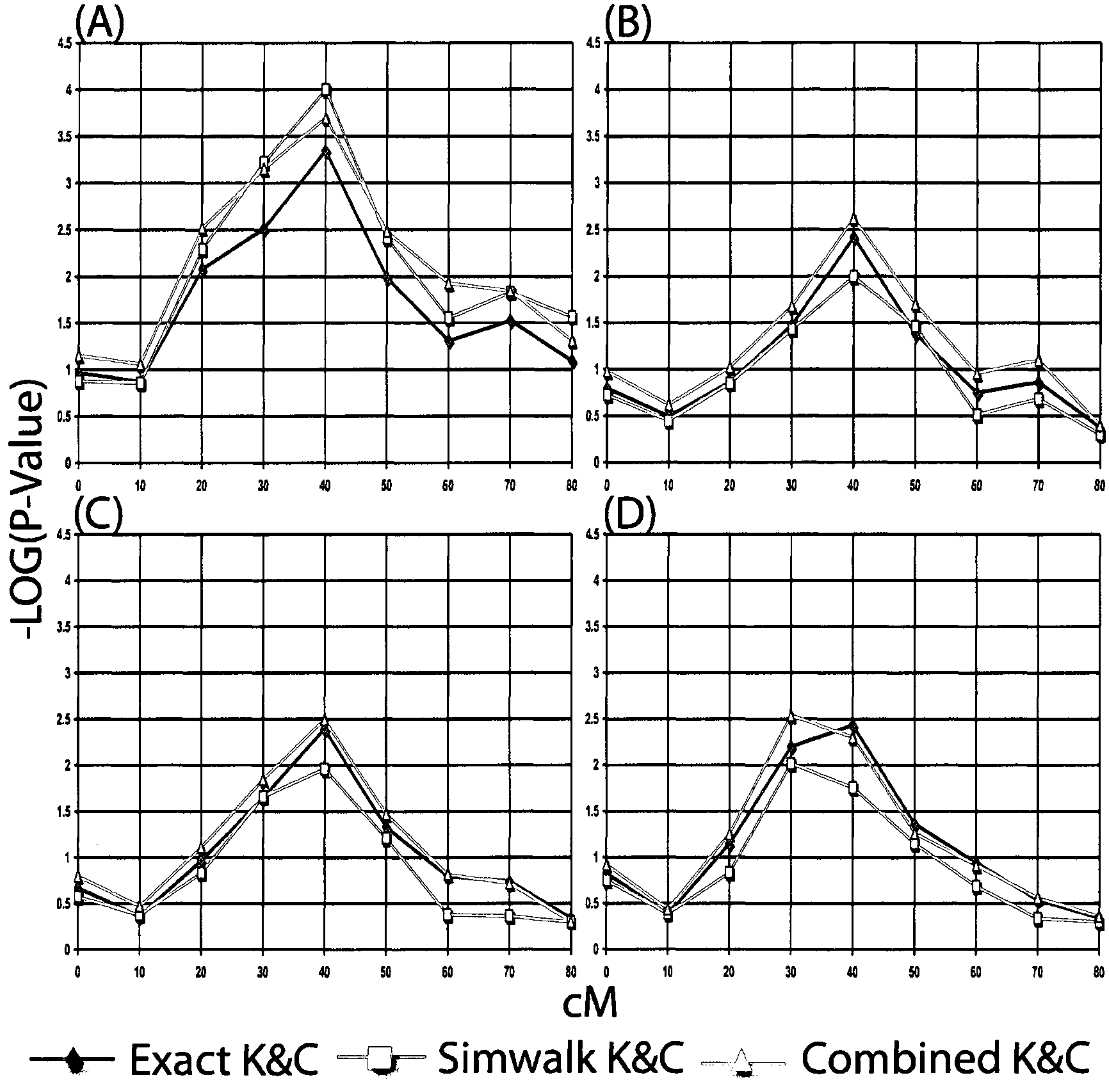


Figure 3.17: Mixed Pedigree Configuration Comparison of K&C Q-NPL Results 25% Heritability. (A) 0% (B) 20% (C) 30% (D) 40% Missing Genotypes

one data set had 38% missing genotypes, another had 55%, and the other had 36% missing genotypes. Pajukanta also indicated that only nuclear families had been genotyped for cost-benefit purposes. But missing genotypes in the pedigree structures, regardless of the reasons, reduce the amount of information in the pedigrees and affect the analysis. Based upon our analysis of the Pajukanta et al. data, the amount of missing genotype data in our analyses are consistent with actual data being used in mapping studies. Furthermore, inspecting the missing data patterns in these pedigrees showed that the patterns we investigated in the simulation studies were indeed similar to the patterns found in the real data.

Pajukanta et al. used Simwalk to perform their analyses since they had several large pedigrees that the exact programs couldn't handle and did not want to discard any of their data. Given the amount of missing genotypes in the study sample and the results from the simulation studies showing Simwalk's limitations with that degree of missing information we decided to re-analyze the Pajukanta et al. data with the new K&C implementation in Simwalk and the Combined K&C implementation. The results of our re-analysis of the data can be seen in Figure 3.18. As a comparison we also analyzed the data with the exact traditional method from Mendel and the exact K&C implementation in Merlin. Due to the algorithms used in these programs which limit the size of the pedigrees that can be analyzed, Mendel's curve is based upon excluding 8 pedigrees that were too big and Merlin's K&C curve is based upon excluding 4 pedigrees that were too big. Figure 3.18 shows that excluding these large, complex pedigrees reduces the amount of relevant information and therefore misses a significant linkage signal around 13 cM in the map. The figure also reveals again that Simwalk's K&C implementation far outperforms its traditional method. The figure also

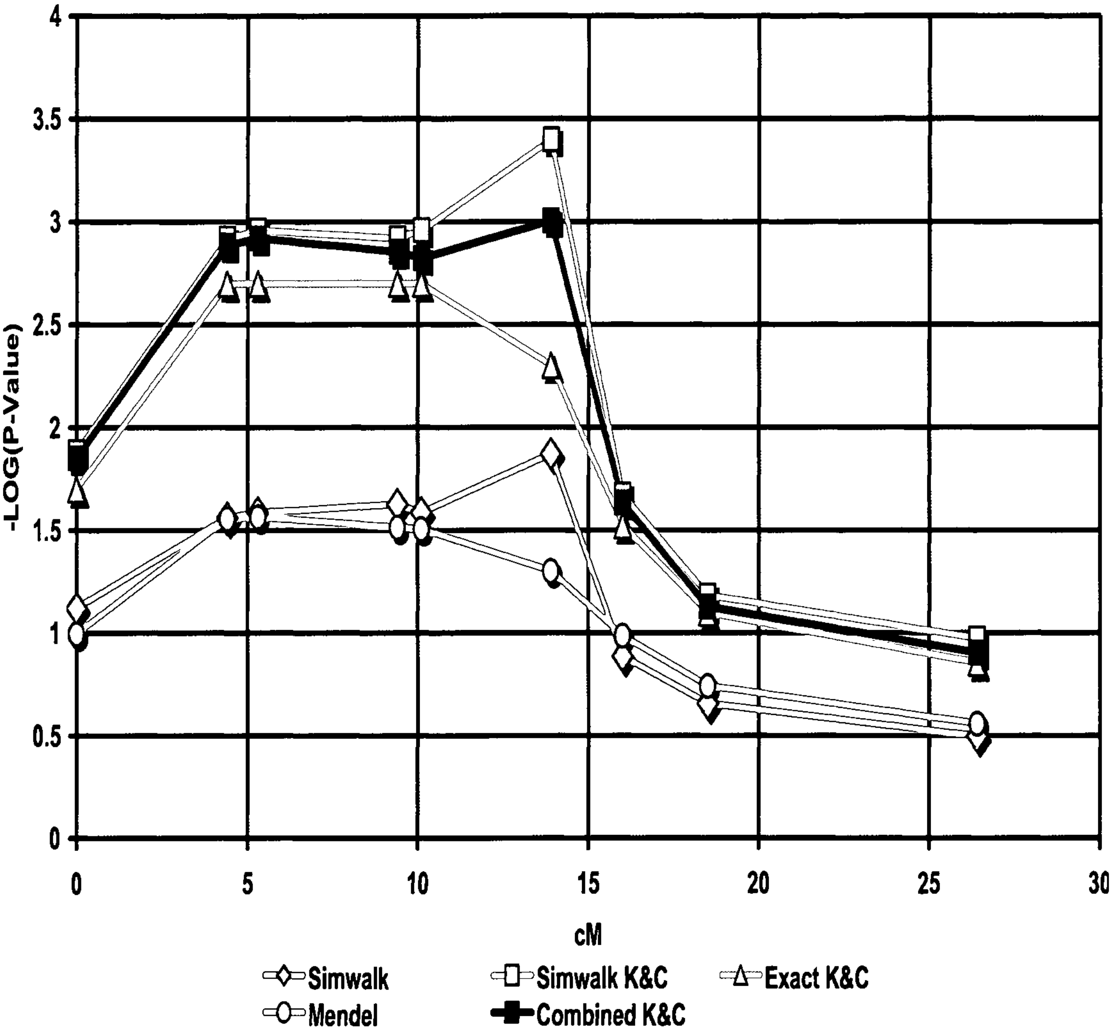


Figure 3.18: Reanalysis of Pajukanta et al. data.

shows that the traditional NPL method, as represented by Mendel, performs much worse than the K&C method when this level of missing data is present. This analysis also reveals that the Combined K&C analysis closely mirrors the Simwalk K&C curve, and at a significant computational savings. In addition, we obtained HDL cholesterol measurements for the individuals in the study sample and performed Q-NPL analysis. In our analysis we set the heritability equal to 50%, used the raw HDL values with the outliers removed and standardized the values. The results of our analysis can be seen in Figure 3.19. The results for the Q-NPL analysis are not as significant as the traditional qualitative NPL, which is counter intuitive. Although the qualitative NPL analysis gave a strong signal for the dichotomized HDL trait, this data set is not ideal for a quantitative analysis. As Pajukanta et al. stated in the original paper, the genotyping strategy they used which resulted in the large amount of missing genotypes discussed above also resulted in limited phenotypic variation in the measured quantitative traits (Pajukanta *et al.*, 2003). Additionally, of the 73 pedigrees analyzed 48 were ascertained for familial combined hyperlipidemia and only 25 were ascertained based on their HDL-C values thus creating heterogeneity in the sample. Given these considerations, it reflects favorably on our new method that Mendel's Q-NPL is able to reach the significance threshold of 2 for a linkage peak around 5 cM. The fact that Mendel was unable to reach that threshold when analyzing the data with the traditional qualitative NPL method as seen in Figure 3.18 also reflects favorably on the new method. The results also show Simwalk's sensitivity to missing genotype information in the Q-NPL analysis.

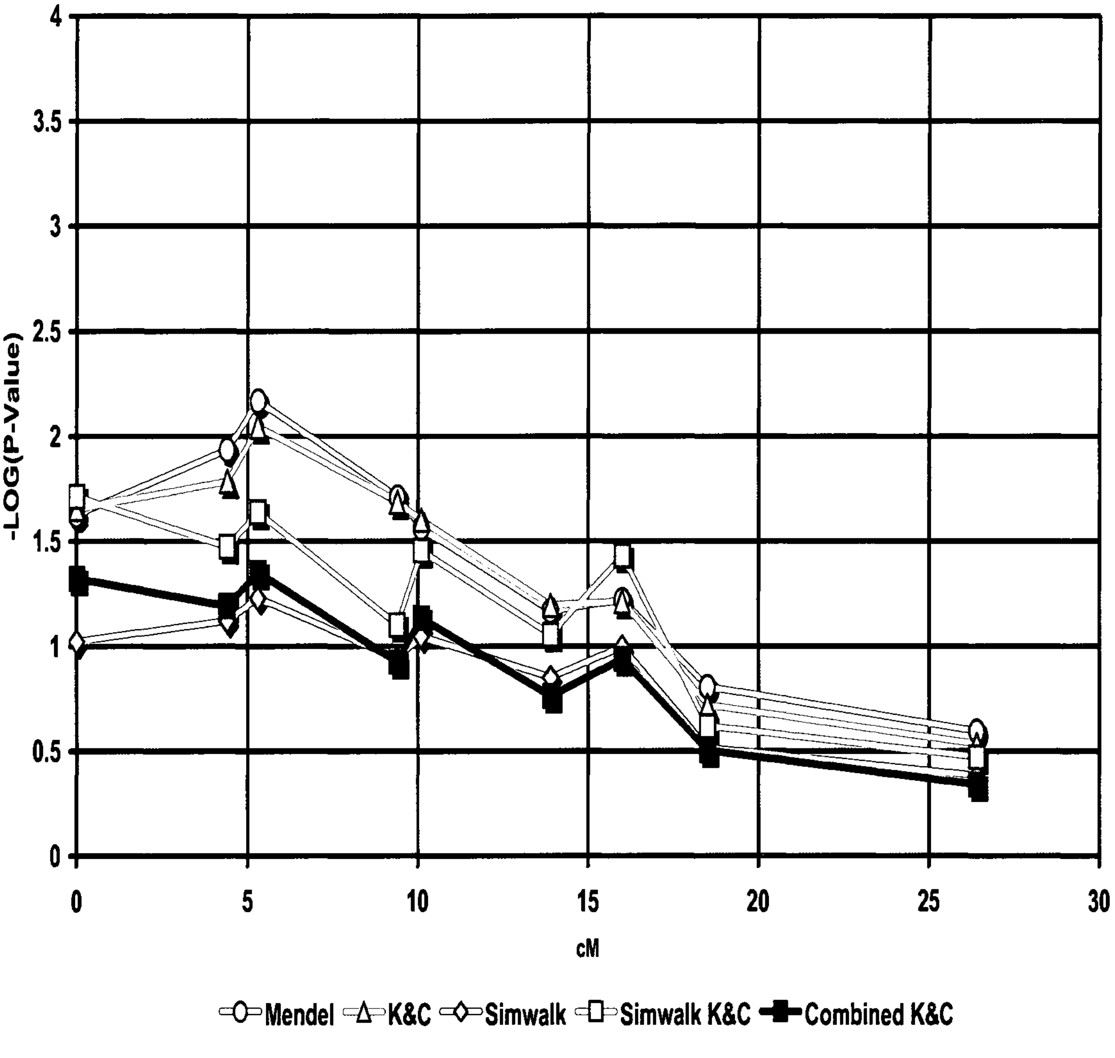


Figure 3.19: Q-NPL analysis of Pajukanta et al. data.

3.7 Discussion

We have proposed and shown results from a new test statistic for performing quantitative NPL (Q-NPL) analysis based upon a simple random effects model motivated from variance component models. The test statistic looks for excessive contributions by a major gene, and is involved in hypothesis testing not parameter estimation which makes the statistic computationally efficient. We have illustrated that the method is applicable to general pedigrees, works well in the context of exact analysis on small pedigrees and estimation techniques utilized in the analysis of large, complex pedigrees and fits nicely into previous work for approximating p-values for NPL statistics (Lange and Lange, 2004). Additionally, we have shown that the statistic works within the framework of the 1-parameter K&C extension to the traditional qualitative NPL method. The method correctly controls the type I error, displays a good ability to detect linkage and is robust to missing genotype information.

The research has revealed that Simwalk's NPL method, both qualitative and quantitative, is extremely sensitive to missing data. In fact, with missing genotypes at levels in real data sets it is potentially missing significant signals of linkage. We have also shown that the K&C NPL extension, which was originally developed for the exact methods, can be extended to work with the estimation techniques that Simwalk employs to perform NPL analysis on large, complex pedigrees. Even more significantly, the qualitative Simwalk K&C implementation has been shown to perform virtually identically to an exact qualitative analysis method. Thus researchers will be able to analyze large, complex pedigrees that the exact methods can't analyze, yet still get results as robust as if they were analyzed by an exact method.

This research has also given an insight into the workings of the perfect data approximation technique of determining significance in the traditional NPL method. Kruglyak et al. in their original paper stated that the perfect data approximation was a conservative method of determining significance and would be less conservative the closer to complete information the data was, but to our knowledge no one has ever published an analysis of when perfect data approximation gives too conservative an estimate and misses a significant linkage signal. Our analysis of Genehunter's (data not shown), Merlin's (data not shown), and Mendel's traditional NPL methods have shown that in datasets with 30% missing genotypes or greater, which based upon our experience is not uncommon, the use of perfect data approximation to determine significance actually means that these programs will often miss linkage signals. This makes us concerned that previous analyses might have given researchers false negatives. Therefore we believe that researchers should consider re-analyzing previous NPL investigations if upon review the data sets have $\geq$ 30% missing genotypes. We recommend using the K&C method in the re-analysis. Another good alternative is the replicate-pool method (Lange and Lange, 2004; Lange, 2002; Song *et al.*, 2004). Our analysis shows that using the replicate pool method gives similar results as our true distribution with significant computational savings, and Wigginton et al. have come to similar conclusions (Wigginton and Abecasis, 2006).

An important and exciting development from this work is that the Kong&Cox method provides a framework to create a state of the art NPL method. The K&C framework allows the results from exact analysis to be combined with results from estimation techniques for either the quantitative or qualitative NPL, as can be accomplished today for parametric linkage analysis. This combined analysis option significantly speeds up the calculation, alleviates the conservativeness of perfect data approximation, and most importantly allows each pedigree to be analyzed by the best available method without losing information from the total sample.

Chapter 4

Efficient use of Dense SNP Maps for Relationship Construction and Gene Mapping

4.1 Introduction

Modern genetic analysis methods can use information about the degree to which the subjects are related to increase the power to identify the genetic etiology of the trait under investigation. There are a number of ways to measure relatedness, but perhaps the most widely utilized measure is the number of alleles that a pair of individuals share identical-by-descent (IBD). Identical by descent means that the two alleles are copies of a gene from a shared ancestral relative. This differs from individuals sharing alleles identical by state (IBS) which means that the alleles are of the same form, but are not copies of a gene from a shared relative. Measurements of IBD are used in linkage studies, association tests, and in quantitative-trait locus (QTL) mapping. Misspecification of the degree of relatedness in a sample can have dramatic affects on the results of these analyses, in some cases reducing the power to detect a signal and in some cases giving false positive results. Therefore it is very important to accurately measure, specify and account for the IBD among individuals in genetic studies.

Due to its importance in genetic analyses, the development of methods to determine pairwise relatedness or account for it in the tests conducted has been an active area of research. With the large number of molecular markers now typed in genetic studies, researchers have focused on using these markers to both determine IBD among subjects as well as adjust the test statistics used by accounting for the amount of IBD in the sample. The methods developed to identify and quantify pairwise relationships stems from Thompson's early work showing that markers can be used to estimate relationships (Thompson, 1974, 1975). Since Thompson's work there have been a number of different methods developed to either test specified IBD relationships or estimate IBD relationships without prior specification. The development of methods to estimate measures of IBD from molecular markers have, up to this point, been focused on the analysis of natural populations where pedigree construction is difficult; and the developed methods have been based upon the 'method of moments' concept (Lynch and Ritland, 1999; Mousseau *et al.*, 1998; Queller and Goodnight, 1989; Wang, 2002). Methods developed to test the accuracy of specified IBD relationships have been developed for human genetics pedigree analysis, and have primarily taken the form of statistical tests to identify incorrectly specified relationships (Boehnke and Cox, 1997; Ehm and Wagner, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000; Sun *et al.*, 2002).

In recent years, the use of genome-wide case-control association tests has gained favor as the method of choice for mapping variants of complex disease (Risch and Merikangas, 1996). Unspecified or unaccounted relatedness, including population structure, in the study sample can increase the false positive rate.

This fact has lead to the development of association tests that account for relatedness in the test statistic itself, by investigating the markers in the data. The methods test and account for population stratification/substructure (Pritchard and Rosenberg, 1999; Reich and Goldstein, 2001; Satten *et al.*, 2001), account for so-called cryptic relatedness among individuals (Devlin and Roeder, 1999; Bacanu *et al.*, 2000; Voight and Pritchard, 2005; Slager and Schaid, 2001), or both (Yu *et al.*, 2006). A recent study showed that in certain populations or studies with "poor" design/ascertainment that there can indeed be high levels of cryptic relatedness that can drastically inflate the false positive rate in association tests (Voight and Pritchard, 2005). But we believe that instead of just accounting for relatedness in the test statistic, if there was a method to estimate a useful measure of relatedness (e.g. theoretical and conditional kinship coefficients) it would allow for more powerful quantitative association analysis to be performed. The theoretical kinship coefficient is determined solely from a pedigree structure, and for a pair of relatives i and j at a locus k is the probability that an allele selected randomly at k in individual i is IBD to an allele chosen randomly at k in individual j. The conditional kinship coefficient also gives the probability for randomly chosen alleles at k for individuals i and j matching IBD, but this probability uses the pedigree structure and is conditioned on all known genotypes of i and j.

In this paper we describe three new, related algorithms using whole-genome SNP data to estimate pairwise IBD coefficients without prior information of relatedness. The first algorithm is used to estimate theoretical kinship coefficients. The second estimates conditional kinship coefficients. The third uses the first procedure to cluster individuals into pedigrees. We then show how these procedures can be used to improve association and linkage-based gene mapping studies.

The first algorithm, to estimate theoretical kinship coefficients, is a simple closed-form, likelihood-based method of moments algorithm that estimates the kinship coefficient between pairs of individuals in a homogeneous population. Our method only requires allele frequencies for the markers utilized in the method, and investigates IBS matching in order to model the amount of IBD sharing between pairs of individuals. Our method is based upon the availability of high-density whole-genome SNP panels, and confirms that with a large number of markers a method of moments approach can accurately estimate relatedness coefficients (Milligan, 2003). This method allows researchers to quickly identify misspecified relationships similar to the statistical tests mentioned above (Boehnke and Cox, 1997; Ehm and Wagner, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000; Sun *et al.*, 2002), but also gives a parameter that would allow the inclusion of the misspecified data with corrected values. This method will also allow researchers with case-control data to discover cryptic relatedness in their sample and take appropriate steps to maximize the power of their analysis.

The second algorithm uses the pairwise theoretical kinship estimate in a discrete, penalized optimization technique to estimate the pairwise conditional kinship coefficient at every SNP. This procedure uses the method-of-moments point estimates at every SNP and the theoretical kinship coefficient estimate to assign every SNP to one of the four possible conditional kinship coefficients ($\Phi = 0, \frac{1}{4}, \frac{1}{2}, 1$) for two individuals at a biallelic SNP. Each kinship coefficient corresponds one of the four possible IBD configurations that two individuals can display (IBD = 0, 1, 2, 4 alleles shared identical-by-descent).

The third algorithm combines the first algorithm with an algorithm from graph theory to cluster individuals into pedigrees. This hybrid algorithm considers all individuals with genotype data as nodes in an undirected graph. The first algorithm described above defines the edges between the individuals, and the standard algorithm to find the connected components of a graph clusters individuals into pedigrees. The pedigree clusters defined by this third method can then be analyzed by the first two algorithms to estimate the coefficients for use in downstream analysis, such as QTL mapping.

In this chapter we will demonstrate our methods' ability to correctly estimate the theoretical and conditional kinship coefficients of known genetic relationships. We also discuss our method's sensitivity to allele frequency misspecification. The methods' utility and performance are illustrated using both real and simulated data.

4.2 Materials and Methods

4.2.1 Materials

We used the Affymetrix Mapping 10K, 100K, 200K, and 500K SNP sets as representative whole-genome high-density SNP panels. In the simulation studies, we used the Caucasian allele frequencies specified in the chip annotation files provided by Affymetrix. We then generated genotypes via gene-dropping using the software package Mendel (Lange *et al.*, 2001). For analysis of the methods' applicability in QTL analysis we analyzed a data set provided by John Blangero at the Southwest Foundation for Biomedical Research consisting of 1942 Mexican-American individuals from the San Antonio, TX area with 107 singletons and the rest of the individuals spread over 46 pedigrees. Of the 1942 individuals in the sample 858 of them were genotyped on the Illumina 550K genotyping array and phenotyped for a quantitative trait. In our analysis we used the maximum likelihood based allele frequencies provided by the Southwest Foundation.

4.2.2 Methods

Assumptions

The methods are based upon the assumption that every homogeneous population sample is part of a large extended pedigree in which the familial relations are not observed. The methods also assume that the allele frequencies for every marker are known with low error rates. The conditional kinship estimation method also assumes that haplotypes occur in blocks and therefore neighboring loci should have the same kinship coefficient (Patil *et al.*, 2001; Daly *et al.*, 2001; The International HapMap Consortium *et al.*, 2003).

Algorithm 1: Estimating Theoretical Kinship Coefficient

To estimate the theoretical kinship coefficient between two individuals we have constructed a simple, closed-form, likelihood-based algorithm. The algorithm is based on using the number of identity-by-state (IBS) matches observed in an extended segment of SNPs as an estimate for the expected number, which is a function of the theoretical kinship coefficient. The observed fraction of matches IBS between two individuals, designated e_{uv} for individuals u and v, for an autosomal chromosome is calculated as

$$e_{uv} = \sum_{i=1}^{m} \left[\frac{1}{4} 1_{\{I_i \,=\, K_i\}} + \frac{1}{4} 1_{\{I_i \,=\, L_i\}} + \frac{1}{4} 1_{\{J_i \,=\, K_i\}} + \frac{1}{4} 1_{\{J_i \,=\, L_i\}} \right] , \qquad (4.1)$$

where the sum is taken over all m marker loci, I and J are the alleles of individual u, K and L are the alleles of individual v, and the 1 represents an indicator function that equals one when the IBS condition is met between the alleles and zero otherwise. For the analysis of the X chromosome, if individuals u and v are both females then no alteration of equation (4.1) is needed. If individual u is a male and individual v is a female then equation (4.1) is altered to

$$e_{uv} = \sum_{i=1}^{m} \left[\frac{1}{2} 1_{\{I_i \;=\; K_i\}} + \frac{1}{2} 1_{\{I_i \;=\; Li\}} \right] \ , \tag{4.2}$$

and if both individual u and individual v are male then equation 4.1 is altered to

$$e_{uv} = \sum_{i=1}^{m} \left[1_{\{I_i \;=\; K_i\}} \right] \ , \tag{4.3}$$

The total expected number of matches, either IBS or IBD, on either an autosome or the X chromosome is given by:

$$\sum_{i=1}^{m} \left[\Phi + (1 - \Phi) \sum_{j=1}^{n} (p_{i,j})^2 \right] \ , \tag{4.4}$$

where the sum is taken over all m marker loci, Φ is the kinship coefficient, and the sum of the second term is taken over all n alleles at each loci with $p_{i,j}$ representing the allele frequency for allele j at marker i. The first term of the sum accounts for the probability of matching IBD, while the second term accounts for the probability of matching IBS but not IBD. Setting either equation (4.1) or (4.2) or (4.3) equal to equation (4.4) and solving for the kinship coefficient Φ gives,

$$\Phi = \frac{e_{uv} - \sum_{i=1}^{m} \sum_{j=1}^{n} (p_{i,j})^2}{m - \sum_{i=1}^{m} \sum_{j=1}^{n} (p_{i,j})^2} \ , \tag{4.5}$$

which provides an unbiased estimator of Φ. When considering a single marker, equation (4.5) yields a point estimate of Φ, which is related to the traditional conditional kinship coefficient. In the next section we describe an algorithm that uses these point estimates to obtain a much better estimate for the conditional kinship coefficient. Although we provide a framework for performing the analysis on the X chromosome, we only discuss and show results for our analysis of the autosomes.

Algorithm 2: Estimating Conditional Kinship Coefficient

To estimate the conditional kinship coefficient between two individuals at a specific locus we have developed a discrete, penalized estimation algorithm that uses a chromosome-wide estimate of the theoretical kinship coefficient from equation (4.5), and the point estimates that equation (4.5) gives for a specific locus. Given that all the markers under consideration are biallelic SNPs, the only possible values for the conditional kinship coefficient at a locus are $0, \frac{1}{4}, \frac{1}{2}$, or 1. Therefore the fundamental idea of the algorithm is to assign the continuous point estimate of the kinship coefficient for a locus to one of these four discrete sets. The set of loci where the kinship coefficient equals 0 is designated S_0; similarly, the set with kinship coefficient $\frac{1}{4}$ is designated S_1; the set with kinship coefficient $\frac{1}{2}$ is designated S_2; and the set with kinship coefficient 1 is designated S_4. The algorithm proceeds through a four step process for each chromosome to assign each locus to one of the four sets:

1. Calculate chromosome-specific estimate of theoretical kinship coefficient

2. Block the stretches of IBS = 0

3. Refine locus-specific kinship coefficient point estimates

4. Find optimal set assignment for each locus by minimizing via dynamic programming an objective function based on penalized optimization .

The first three steps in the algorithm can be considered pre-processing steps to initialize the data for the penalized estimation.

The first step is to calculate a chromosome-specific theoretical kinship coefficient as described in the previous algorithm. During this process, save each locus' point estimate, and keep track of whether the locus was IBS $= 0$ (i.e., no alleles in common between the two individuals) or IBS $= 1$ (i.e., all four alleles identical among the two individuals).

In step two we block stretches of loci that were flagged IBS $= 0$. The reason for flagging loci that are IBS $= 0$ is that these loci actually are the most informative for correctly assigning the locus to the correct set. If a locus is IBS $= 0$ we know with certainty that the locus is IBD $= 0$ and therefore can anchor our estimation at these loci. The assumption of the block nature of haplotypes means that neighboring loci should be similar; therefore we search for blocks of loci that have been flagged as IBS/IBD $= 0$. We scan each chromosome from the start to the end and look for flagged loci that are separated by less than one megabase. For each such interval we set the kinship coefficient point estimate for each locus in the interval equal to 0. Tracking loci that have IBS $= 1$ allows a computational savings in the estimation algorithm because we do not need to consider loci as belonging to S_4 unless they are IBS $= 1$.

In the third step, we refine the point estimates for the kinship coefficient obtained at each SNP. Since the initial point estimates are poor, we use our assumption that neighboring loci should have similar values to refine these estimates for each locus by considering the point estimates of neighboring loci in a small window centered on the current locus. We refine the estimates by calculating a modified version of equation (4.5) using all the markers in the window. Equation (4.5) is modified by weighting the point estimate at each locus in the window by its heterozygosity, where the heterozygosity of locus i is calculated by $w_i = 1 - \sum_{j=1}^{n}(p_{i,j})^2$. The estimate of the kinship coefficient for the SNP at the center of the window becomes, after substitution and simplification,

$$\Phi = \frac{\sum_{i=1}^{m}\left(e_{uv}^{i} - \sum_{j=1}^{n}(p_{i,j})^2\right)}{\sum_{i=1}^{m}\left(1 - \sum_{j=1}^{n}(p_{i,j})^2\right)}, \qquad (4.6)$$

where m is the number of loci in the window, e_{uv}^{i} is the number of IBS matches between individuals u and v at locus i and $p_{i,j}$ is the allele frequency of allele j at locus i. The number of loci m in the window is based on the average spacing between loci on a chip, but are selected so that the window is approximately 50KB on either side of the centered SNP. This is the last preprocessing step and the refined point estimates are the input into the penalized estimation technique.

In step 4, the penalized estimation finds the $z_1, \ldots, z_m$ which minimize the objective function

$$\begin{aligned} f(z) &= \sum_{i \in S_0}(y_i - 0)^2 + \sum_{i \in S_1}(y_i - \frac{1}{4})^2 + \sum_{i \in S_2}(y_i - \frac{1}{2})^2 \\ &\quad + \sum_{i \in S_4}(y_i - 1)^2 + \lambda_1 \sum_{i=1}^{m}(z_i - \Phi_{chr})^2 + \lambda_2 \sum_{i=1}^{m-1}(z_{i+1} - z_i)^2 \qquad (4.7) \end{aligned}$$

where z_i is the conditional kinship coefficient at SNP i, and has the possible values 0, $\frac{1}{4}$, $\frac{1}{2}$, and 1, and y_i is the point estimate for the kinship coefficient at locus i obtained in the previous step. The λ_1 is the penalty for the conditional kinship coefficient being different from the chromosome-specific theoretical kinship coefficient estimate Φ_{chr}. The coefficient λ_2 is the penalty for neighboring loci belonging to different kinship coefficient sets.

This problem can be solved using dynamic programming in a single pass of the data. The dynamic programming solution begins by formulating the objective function (4.7) as

$$f(z) = \sum_{i=1}^{m} f_i(z_i) + \sum_{i=1}^{m-1} g_i(z_i, z_{i+1}). \tag{4.8}$$

The function $f_i(z_i)$ has four possible formulations, one for each possible state at a locus:

$$\begin{aligned} f_i(z_i = 0) &= (y_i - 0)^2 &+& \lambda_1(0 - \phi_{chr})^2 \\ f_i(z_i = \tfrac{1}{4}) &= (y_i - \tfrac{1}{4})^2 &+& \lambda_1(\tfrac{1}{4} - \phi_{chr})^2 \\ f_i(z_i = \tfrac{1}{2}) &= (y_i - \tfrac{1}{2})^2 &+& \lambda_1(\tfrac{1}{2} - \phi_{chr})^2 \\ f_i(z_i = 1) &= (y_i - 1)^2 &+& \lambda_1(1 - \phi_{chr})^2, \end{aligned} \tag{4.9}$$

and the penalty function $g_i(z_i, z_{i+1})$ is

$$g_i(z_i, z_{i+1}) = \lambda_2(z_{i+1} - z_i)^2. \tag{4.10}$$

To solve by dynamic programming define $h_1(z_1) = f_1(z_1)$ and for $k > 1$,

$$h_k(z_k) = \min_{z_1,\ldots,z_{k-1}} \sum_{i=1}^{k} f_i(z_i) + \sum_{i=1}^{k-1} g_i(z_i, z_{i+1}). \tag{4.11}$$

We record the values $h_k(0), h_k(\frac{1}{4}), h_k(\frac{1}{2})$, and $h_k(1)$. Then

$$h_k(z_k) = \min_{z_{k-1}} \{h_{k-1}(z_{k-1}) + g_{k-1}(z_{k-1}, z_k)\} + f_k(z_k) \tag{4.12}$$

is the needed recursion. Finally, we set z_m to the value that minimizes $h_m(z_m)$ and obtain the $z_1, \ldots, z_{m-1}$ through the standard dynamic programming traceback procedure.

Algorithm 3: Pedigree Construction

There has been a substantial body of work in utilizing genotype data to test for pedigree misspecification (Boehnke and Cox, 1997; Ehm and Wagner, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000; Sun *et al.*, 2002). All of these methods perform statistical inference on specified pedigrees and return relationships with a high probability of misspecification, and some propose the most probable true relationship for the identified errors. But none of these methods perform inference between pedigrees to test for cryptic relatedness in the study sample. We propose a new method to construct 'pedigrees' from the genotype data and therefore avoid the computational cost of statistical inference testing while allowing for the discovery of cryptic relationships between pedigrees.

The method we have developed is based on graph theory. We consider every individual with genotype data as a node in an undirected graph. The pedigree discovery problem is equivalent to finding all the connected components of this individual graph. We use our method for estimation of the theoretical kinship coefficient to find the edges between individuals that link them in a component or pedigree. The inputs into the method are the individuals' genotypes and a cutoff value that the theoretical kinship estimation method must reach to designate an edge between individuals. Construction of these pedigrees is an efficient process to run because of the use of the standard algorithm for finding the connected components of a graph, and in the worst case scenario of no individuals in the data set being linked for a specified cutoff our method performs $O(n^2)$ theoretical kinship estimation calculations where n is the number of individuals.

Assessing Performance and Properties in Simulations

The method for estimating the theoretical kinship coefficient was tested using simulated data. The simulations were undertaken by using the pedigree structure in Figure 4.1, and generating 500 different pedigrees of that structure by producing genotypes for the individuals via gene dropping in the program Mendel (Lange *et al.*, 2001). The true theoretical kinship coefficient was calculated exactly for the pedigree using the algorithm described by Lange (Lange, 2002). The performance of our method was then assessed by analyzing the distribution of the estimated coefficients of all the pairwise relationships from the 500 generated pedigrees and comparing to the true theoretical kinship coefficient. Of course the estimated values were found without any knowledge of the pedigree structures. We also tested the constructed distributions for normality by using the Kolmogorov-Smirnov test. We conducted a one sample Kolmogorov-Smirnov test using our estimates and testing against a normal distribution with mean equal to the sample mean and standard deviation equal to the sample standard deviation. This process was conducted on the 10K, 100K, 200K, and 500K Affymetrix chip information to assess the relationship between the number of SNPs and the accuracy of the estimate. Since our method assumes correct allele frequencies, we also tested our method's sensitivity to allele frequency misspecification. This was accomplished by gene-dropping under one set of allele frequencies and analyzing the data with a different set of allele frequencies. Each chip's performance was assessed under the following conditions:

1. All loci gene-dropped with major allele frequency increased and decreased by 1% and 5% from allele frequency used in analysis

2. 25% of loci gene-dropped with major allele frequency increased and decreased by 1% and 5% from allele frequency used in analysis

3. 50% of loci gene-dropped with major allele frequency increased and decreased by 1% and 5% from allele frequency used in analysis

4. All loci gene-dropped with major allele frequency drawn from a normal distribution with mean equal to major allele frequency and standard deviation 0.025 × major allele frequency

5. All loci gene-dropped with major allele frequency drawn from a normal distribution with mean equal to major allele frequency and standard deviation 0.05 × major allele frequency

6. All loci gene-dropped with major allele frequency drawn from a normal distribution with mean equal to major allele frequency and standard deviation 0.075 × major allele frequency

7. All loci gene-dropped with major allele frequency drawn from a normal distribution with mean equal to major allele frequency and standard deviation 0.10 × major allele frequency

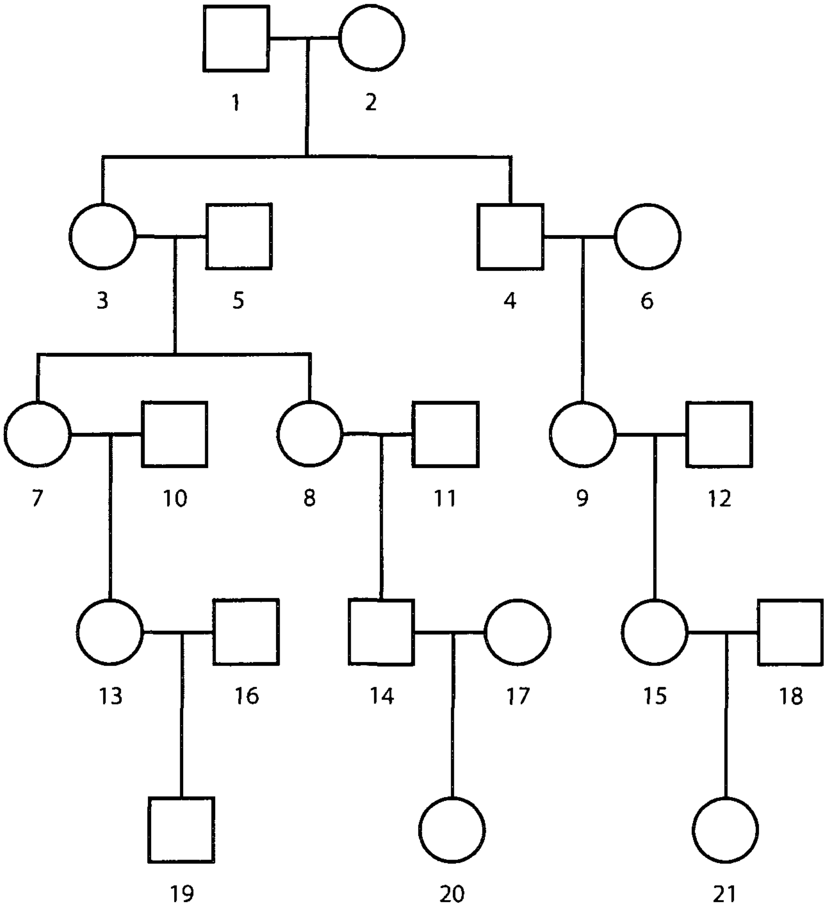


Figure 4.1: Simulated Pedigree Structure

For each of the above conditions we performed the same analysis as with the true allele frequencies.

The estimation procedure for conditional kinship coefficients was also assessed via simulation studies. The same 500 pedigrees analyzed for the theoretical kinship estimation were used for testing the conditional kinship coefficient estimation. The accuracy of the method was determined by comparing the calculated estimate for each locus z_i, described above, to the true value t_i obtained from using equation 4.1 on a single locus at a time from a gene-dropping where each founder allele is uniquely labeled. The result is reported as the average absolute difference (aad) between the true value and the calculated estimate over all loci analyzed

$$aad = \frac{\sum_{i=1}^{m} |t_i - z_i|}{m} \tag{4.13}$$

where m is the number of loci analyzed. The conditional kinship coefficient estimation technique was not analyzed under allele frequency misspecification.

Implications for Gene Mapping of Quantitative Traits

Variance component methods are very powerful tools for mapping quantitative trait loci (QTL). In the standard variance component model the value of a quantitative trait Y for individual i is modeled as

$$Y_i = \mu + \beta^T Z_i + A_i + q_i + e_i \tag{4.14}$$

where μ is the population mean of the trait, Z are the environmental predictors, A is the polygenic effect, q is the major trait locus effect and e is the residual error. Ignoring dominance effects, the variance of Y is

$$Var(Y) = V_A + V_G + V_E \tag{4.15}$$

where V_A is the additive genetic variance, V_G is the polygenic variance and V_E is the environmental variance. For non-inbred relatives i and j the trait's covariance is

$$Cov(Y_i, Y_j) = 2\hat{\Phi}_{ij} V_A + 2\Phi_{ij} V_G \tag{4.16}$$

where $\hat{\Phi}_{ij}$ is the conditional kinship coefficient for i and j at a map location, and Φ_{ij} is the theoretical kinship coefficient for i and j. Algorithms 1 and 2 described above give you estimates for Φ_{ij} and $\hat{\Phi}_{ij}$ at every SNP on a chip.

In order for our estimates of Φ_{ij} and $\hat{\Phi}_{ij}$ to have relevance they must be able to be substituted for the values calculated traditionally in the analysis of quantitative traits. To assess the quality and utility of the estimates produced we used our estimates to perform QTL mapping in the Southwest Foundation data set described above, which has a known QTL signal. We compared the results of their QTL analysis to results of analyses using our estimates for the coefficients. We performed two different analyses. For the first, we used the stated pedigree structures provided by the Southwest Foundation but used our methods to construct the theoretical kinship matrix for the pedigrees and the conditional kinship coefficient estimates at each locus on the chip for each pair in the pedigrees. For the second investigation, we first ran the pedigree construction algorithm described above with a kinship coefficient cutoff of 0.20 to construct 'pedigrees', then calculated all the theoretical and conditional coefficients in the constructed pedigrees and ran the QTL analysis on this new data set. Both of our investigations only analyzed individuals with both genotype and phenotype values.

4.3 Results

4.3.1 Theoretical Kinship Coefficient Estimation

To test the characteristics and accuracy of our theoretical kinship coefficient estimator we performed extensive gene-dropping simulations using the Pedigree structure seen in Figure 4.1. For each SNP-chip analyzed (Affymetrix 10K, 100K, 200K, 500K) we generated 500 replicates of the pedigree structure and estimated the theoretical kinship coefficient for all pairwise relationships in the pedigree. We analyzed the distribution of the estimated coefficients to determine its accuracy and whether the estimator is indeed an unbiased estimator. Here we are showing and discussing the results for eight pairs of relationships that span the spectrum of genetic relationships in the pedigree and give a clear answer as to the validity of our method.

The first relationship analyzed is a pair of unrelated individuals. This analysis will allow us to determine if our estimator is unbiased and gives us the lower bound of the kinship coefficients we are able to accurately estimate. For the analysis of unrelated individuals we compared the two founders of the pedigree, individuals 1 and 2 (Figure 4.1). Table 4.1 gives the minimum, mean and maximum calculated values for this pair of unrelated individuals over 500 replicates for each of the 4 chips analyzed. As the data in the table shows, the mean value on all the chips is very close to zero which is the true genetic relatedness of these two individuals. Additionally it appears that our method is unbiased because of its ability to obtain both positive and negative values. To more formally test the results for bias we conducted the Kolmogorov-Smirnov (KS) test by comparing the empiric distribution of the estimated coefficients against a normal distribution

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0	-0.02464	-0.015208	-0.0001802	0.014848	0.02109	0.00751
100K	0	-0.006876	-0.00540	-0.0001257	0.00515	0.008267	0.00264
200K	0	-0.004975	-0.00388	9.059e-05	0.00406	0.006254	0.001987
500K	0	-0.004236	-0.00250	4.285e-05	0.00258	0.004536	0.00127

Table 4.1: Individual 1 vs Individual 2 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

Chip	D	P-Value
10K	0.0220464	0.9682626
100K	0.02459163	0.9229316
200K	0.02233089	0.9643697
500K	0.02090449	0.981261

Table 4.2: Individual 1 vs Individual 2 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

with mean equal to the mean of a chip's estimates and the standard deviation equal to the standard deviation of the estimates. The results of the KS test for this pair of unrelated individuals is shown in Table 4.2, and the test was unable to reject the null hypothesis that the values are normally distributed. Figure 4.2 plots the histograms for this pair of individuals for the 10K (4.2A), 100K(4.2B), 200K (4.2C) and 500K (4.2D) chips. This figure clearly illustrates the shrinking standard deviation and improved accuracy of our estimator as the number of SNPs in the calculation increases. The red vertical lines on the plots show the ±2 standard deviation area of the distribution. When using the 500K SNP-chip it appears that the lower bound for our estimator for distinguishing between related and unrelated individuals is a kinship coefficient of approximately 0.003, which is about the relatedness of 3rd cousins.

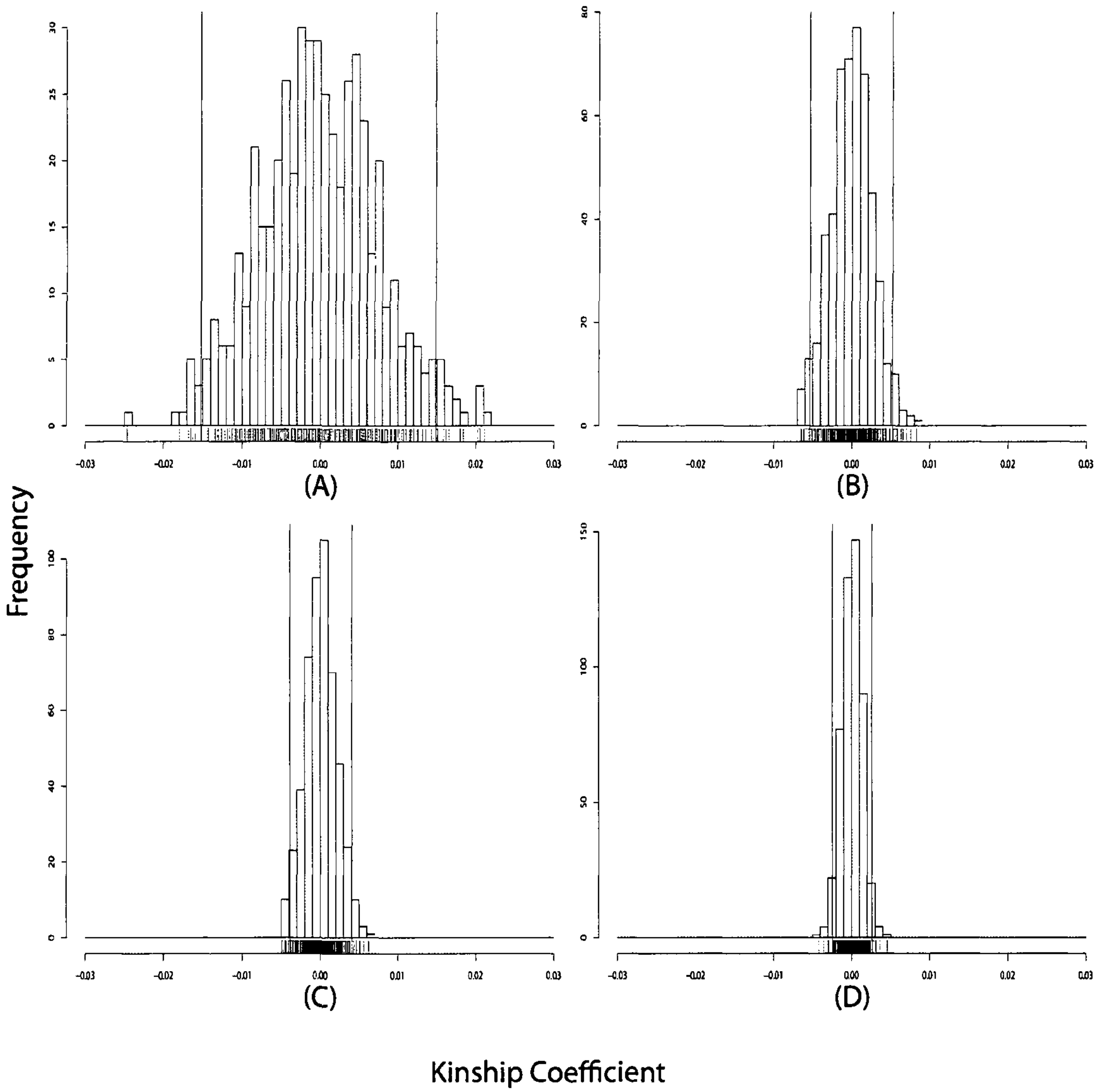


Figure 4.2: Individual 1 vs Individual 2 Kinship Coefficient Estimation, True Value = 0.0: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.25	0.1793	0.2016	0.2478	0.2940	0.3135	0.02310
100K	0.25	0.1949	0.20920	0.2485	0.28789	0.3109	0.01967
200K	0.25	0.1976	0.21280	0.2503	0.28772	0.2973	0.01873
500K	0.25	0.2139	0.22050	0.2502	0.27995	0.3007	0.01486

Table 4.3: Individual 3 vs Individual 4 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

The next relationship analyzed has a theoretical kinship coefficient of 0.25, which is the genetic relatedness of siblings and parent-offspring pairs. To represent this level of relatedness we analyzed the sibling pair of individuals 3 and 4 of Figure 4.1. As Table 4.3 shows the mean of the distribution of calculated coefficients correctly estimates the kinship coefficient for all 4 chips. As was seen with the unrelated pair, with increasing number of SNPs the standard deviation of the distribution shrinks. Figure 4.3 illustrates how with increasing number of SNPs analyzed the method produces a very tight distribution of coefficients centered on the true coefficient value. As Table 4.4 shows the distributions for this relatedness level cannot be rejected by the KS test as coming from a normal distribution with mean equal to the sample mean and standard deviation equal to the sample standard deviation. Further supporting our assertion that our estimator is unbiased. The results also support that this indeed is a powerful and robust estimator. The results for the 500K chip are very impressive with the minimum calculated value only 0.04 from the true value.

Chip	D	P-Value
10K	0.03096577	0.7238213
100K	0.02664091	0.8699132
200K	0.02375163	0.9405131
500K	0.02586740	0.8922024

Table 4.4: Individual 3 vs Individual 4 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

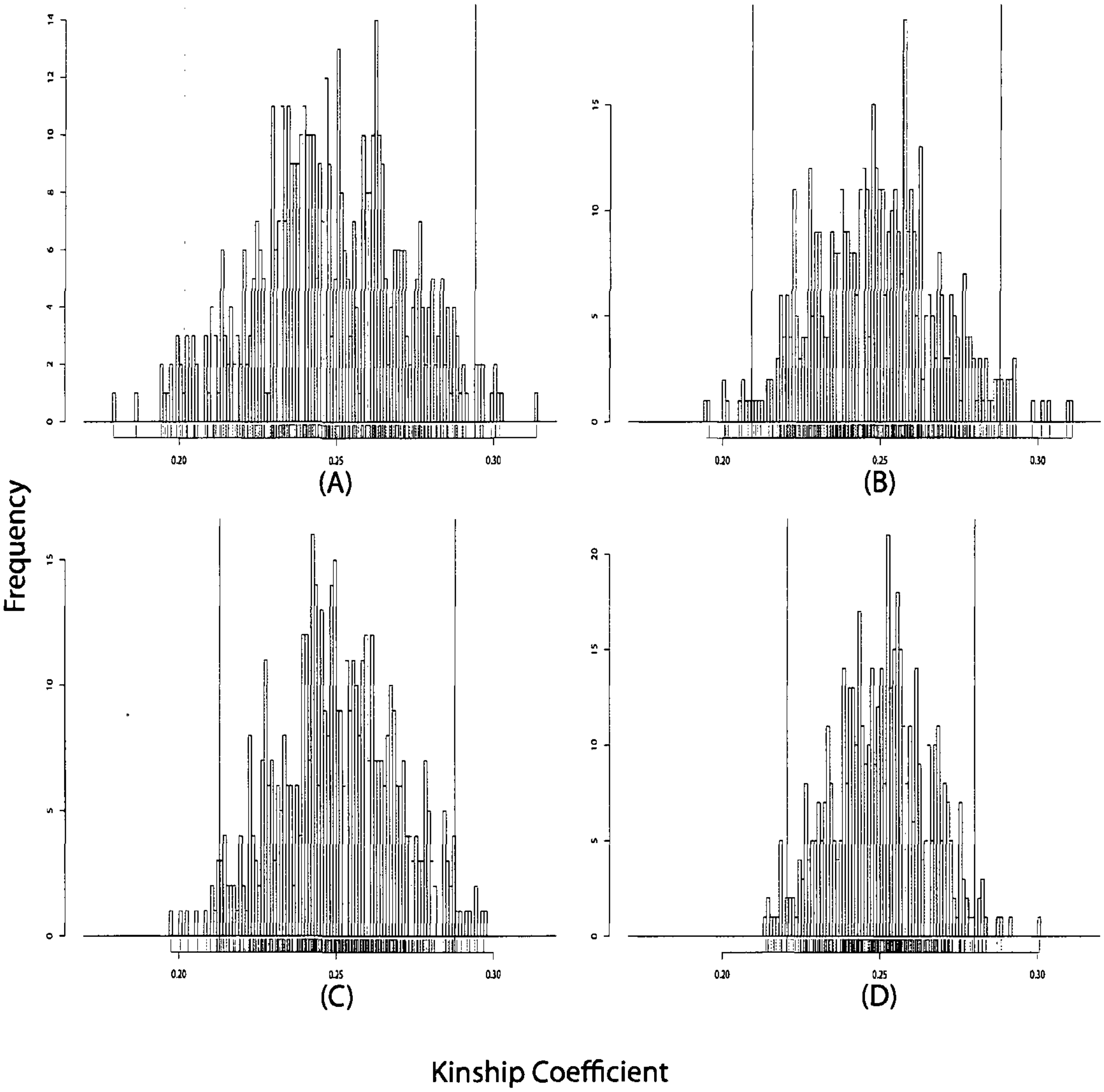


Figure 4.3: Individual 3 vs Individual 4 Kinship Coefficient Estimation, True Value = 0.25: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.08004	0.09047	0.1236	0.15664	0.1739	0.01654
100K	0.125	0.08556	0.09705	0.1238	0.15047	0.1594	0.01335
200K	0.125	0.09049	0.09948	0.1249	0.15033	0.1645	0.01271
500K	0.125	0.09986	0.10635	0.1254	0.14442	0.1519	0.00952

Table 4.5: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

Chip	D	P-Value
10K	0.01885266	0.994255
100K	0.02859544	0.8082093
200K	0.03205186	0.6832826
500K	0.04013574	0.3974943

Table 4.6: Individual 4 vs Individual 7 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

The third relatedness level analyzed has a theoretical kinship coefficient of 0.125, and designates individuals related at the level of uncle-niece/nephew or grandparent-grandchild pairs. To represent this type of relatedness we analyzed the individual pair 4 and 7 from Figure 4.1. The mean value of our theoretical kinship estimate distribution correctly estimates this level of relatedness for all 4 chips as seen in Table 4.5. The same trend of better mean estimates and smaller standard deviations with increasing number of SNPs is continued here. Table 4.6 also shows that the distributions cannot be rejected as coming from a normal distribution as determined by the KS test. The reduction in the standard deviation of the distribution with increasing numbers of SNPs is clearly illustrated in Figure 4.4. The results for the 500K chip also show that there is a clear separation between the distributions for individuals 3 & 4 and that for individuals 4 & 7, shown in Figure 4.5.

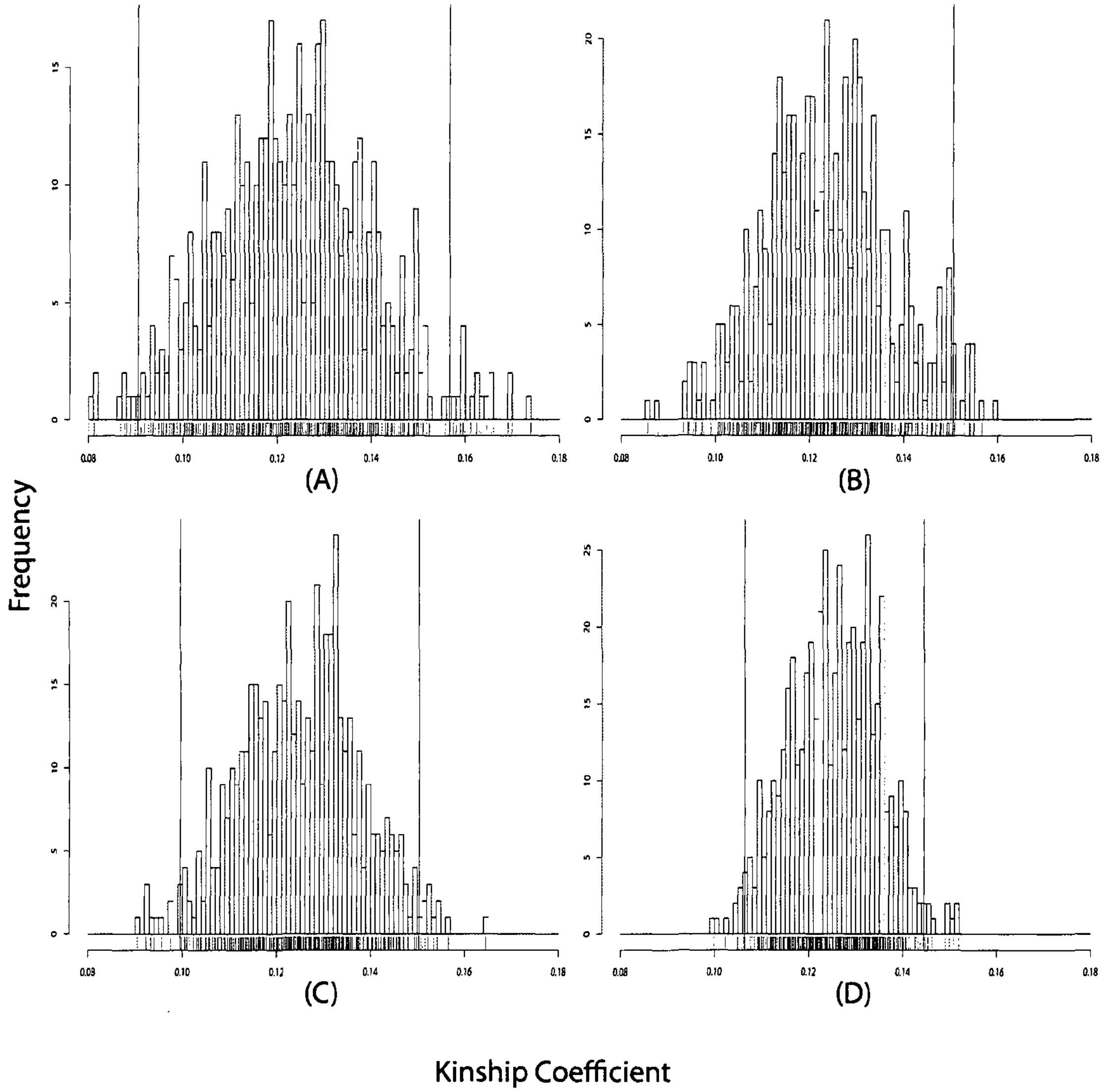


Figure 4.4: Individual 4 vs Individual 7 Kinship Coefficient Estimation, True Value = 0.125: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

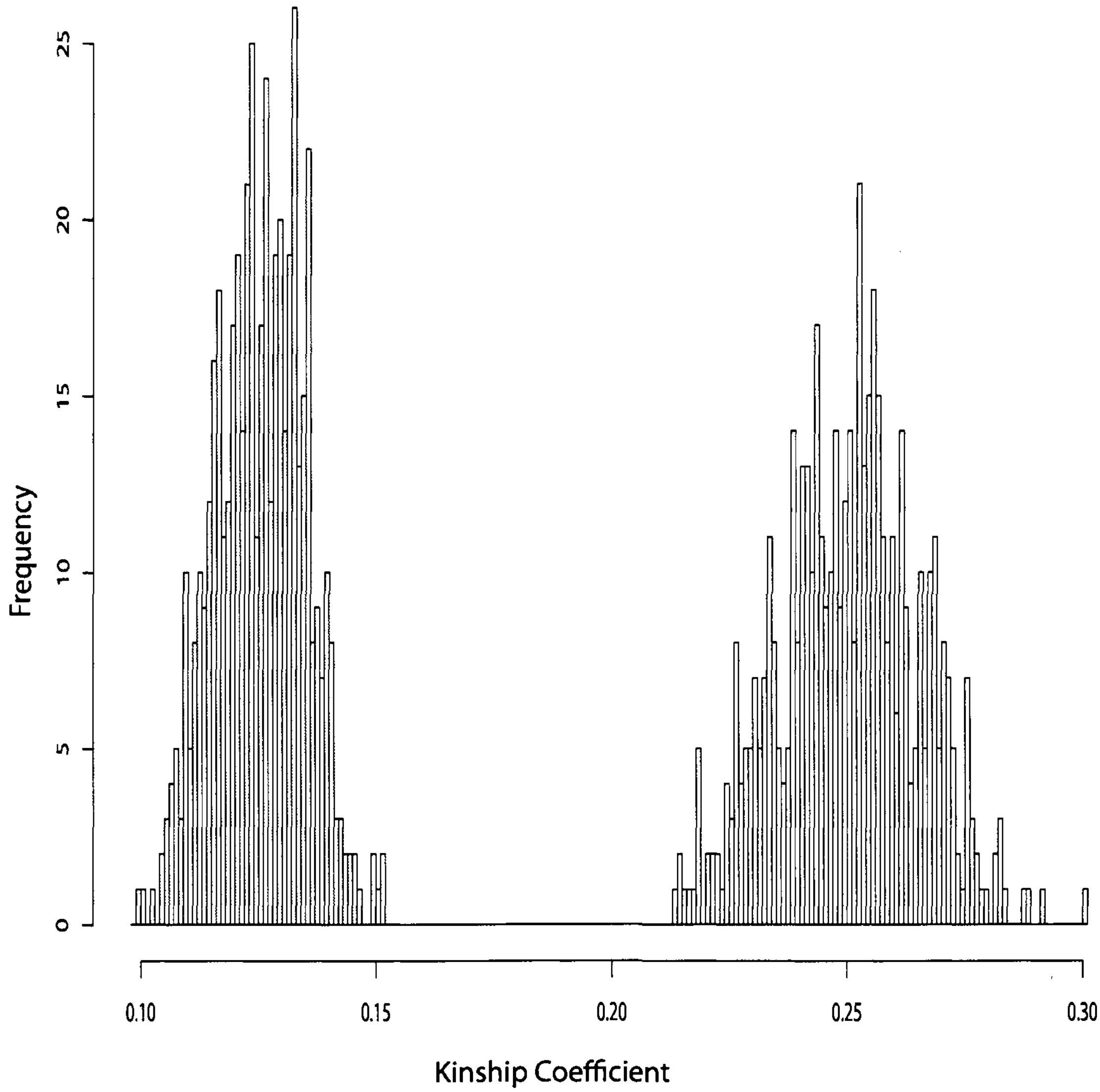


Figure 4.5: Comparison of Distributions of Kinship Estimates for 0.25 true coefficient (red) and 0.125 true coefficient (blue) for the 500K Chip

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.0625	0.006607	0.02701	0.06214	0.09728	0.1153	0.01757
100K	0.0625	0.02548	0.03093	0.0619	0.09288	0.1109	0.01549
200K	0.0625	0.01616	0.03239	0.0623	0.09221	0.1082	0.01495
500K	0.0625	0.0316	0.03869	0.06268	0.08667	0.09739	0.01200

Table 4.7: Individual 4 vs Individual 21 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

Chip	D	P-Value
10K	0.03375575	0.6190896
100K	0.03792196	0.4684229
200K	0.03472519	0.5828407
500K	0.04334377	0.3045118

Table 4.8: Individual 4 vs Individual 21 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

The fourth relationship analyzed has a theoretical kinship coefficient of 0.0625, and reveals relatedness of the order of a great grandparent-great grandchild pair. To represent this level of relatedness we analyzed the great grandparent-great grandchild pair of individuals 4 and 21 from Figure 4.1. The mean estimate from our method for all four chips is accurate to the third decimal place as seen in Table 4.7. The standard deviation shrinks as the number of SNPs analyzed increase which Figure 4.6 illustrates nicely. The estimator continues to be unbiased as evidenced in Table 4.8 where the KS test is unable to reject the null hypothesis that the estimates belong to a normal distribution. For the 500K chip there is still a clear separation between the distribution for this relationship and the next highest relationship represented above for individuals 4 and 7, as seen in Figure 4.7.

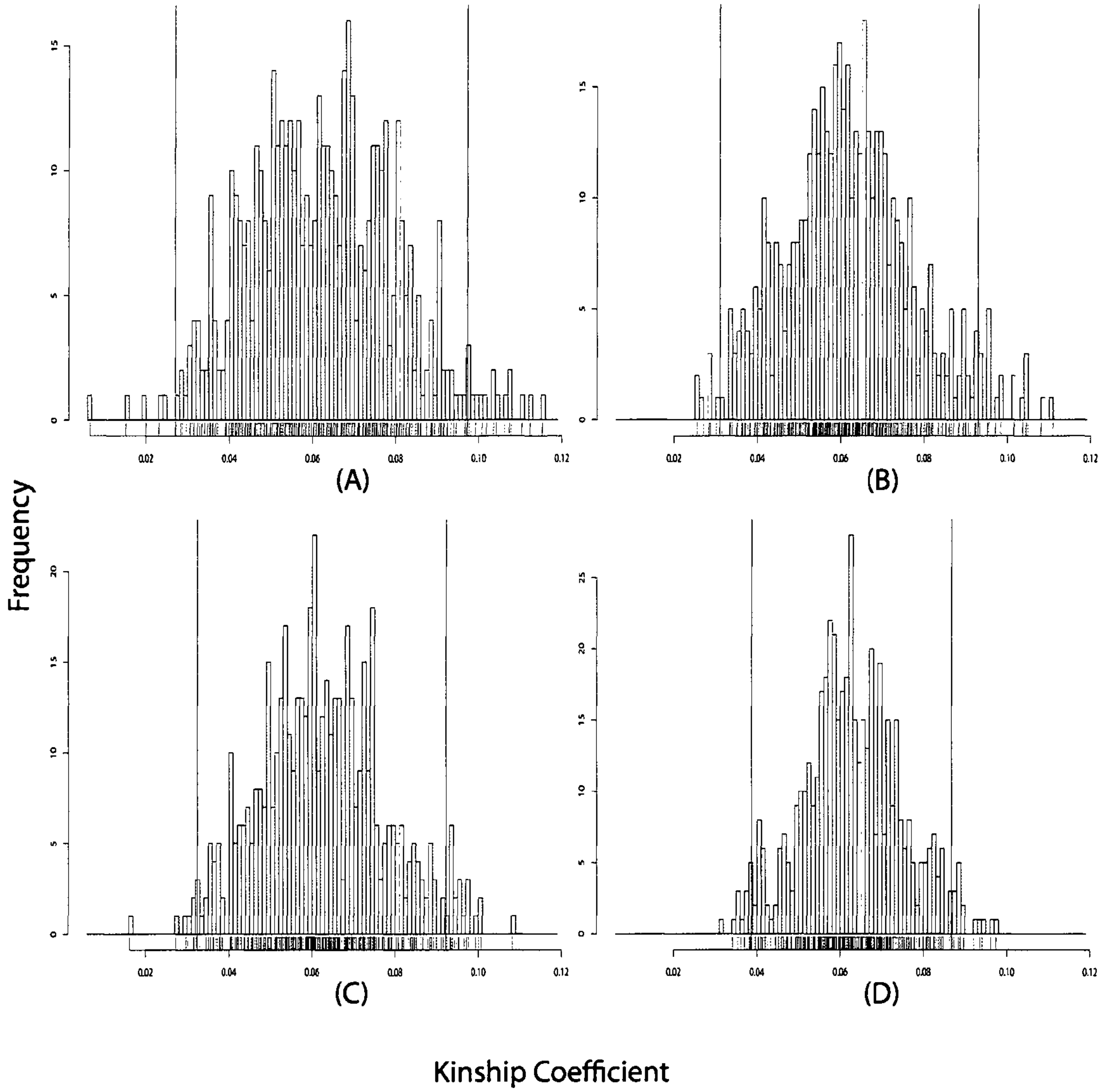


Figure 4.6: Individual 4 vs Individual 21 Kinship Coefficient Estimation, True Value = 0.0625: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

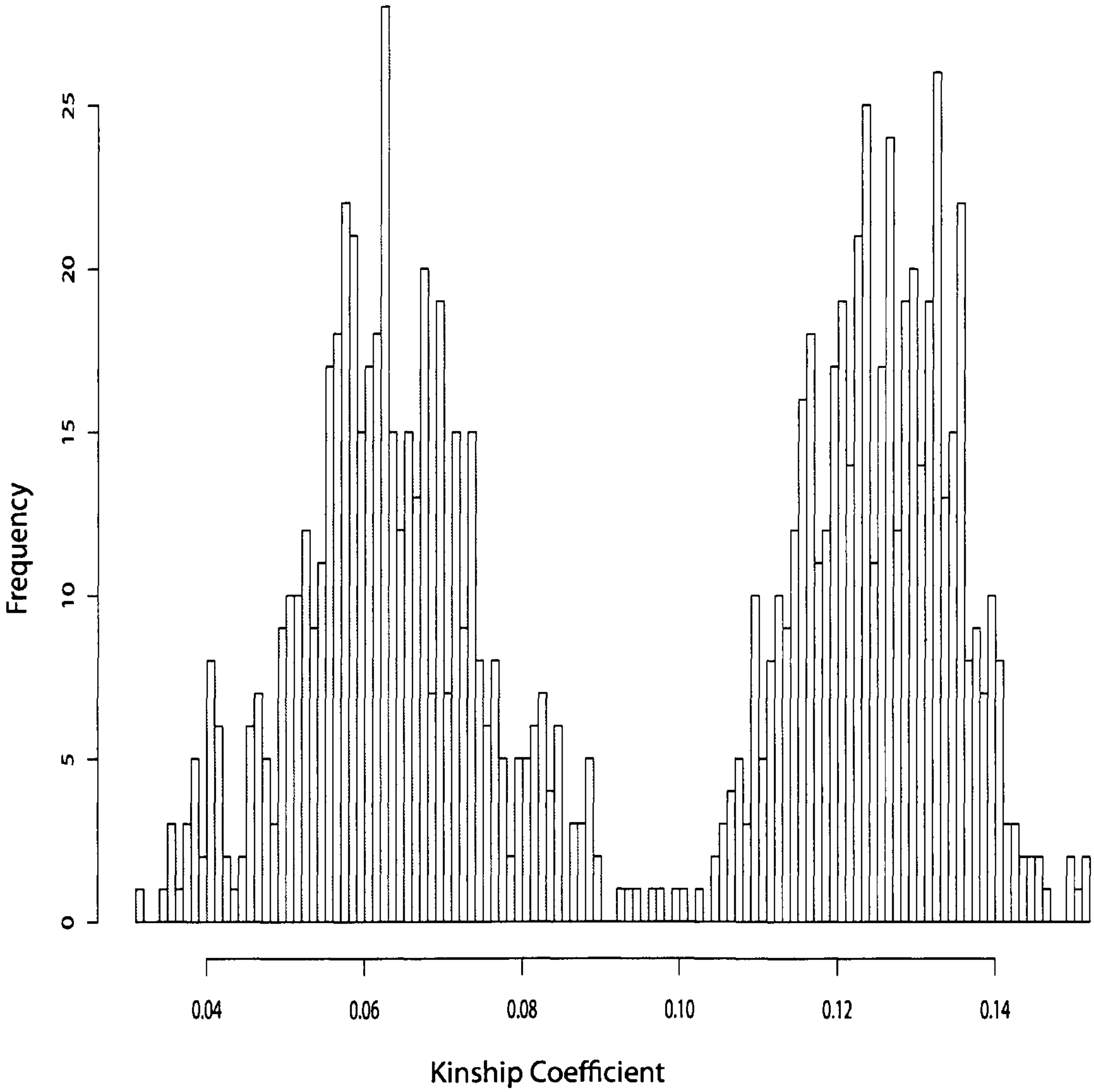


Figure 4.7: Comparison of Distributions of Kinship Estimates for 0.125 true coefficient (red) and 0.0625 true coefficient (blue) for the 500K Chip

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.03125	-0.0003526	0.00660	0.03154	0.05648	0.07819	0.01247
100K	0.03125	0.00413	0.01153	0.03113	0.05073	0.06088	0.00980
200K	0.03125	0.009827	0.01293	0.03073	0.04854	0.06285	0.00890
500K	0.03125	0.01424	0.01698	0.03127	0.04556	0.05073	0.00715

Table 4.9: Individual 4 vs Individual 19 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

Chip	D	P-Value
10K	0.03656664	0.5157035
100K	0.04312246	0.3103269
200K	0.04747099	0.2098196
500K	0.03630494	0.5263683

Table 4.10: Individual 4 vs Individual 19 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

The fifth relationship analyzed has a theoretical kinship coefficient of 0.03125, and represents relatedness on the order of one's self to one's great-grandparent's sibling. We analyzed the pair of individuals 4 and 19 from the pedigree in Figure 4.1 to represent this level of relatedness (highlighted in Figure 4.1). The mean estimate from the distribution generated by the method is still accurate out to the third decimal place for this distant level of relatedness, as seen in Table 4.9. The mean value for the 500K chip is almost exactly at the true value of 0.03125, and has a very small standard deviation. The KS test results seen in Table 4.10 still fails to reject that the estimates from our method are drawn from a normal distribution. Figure 4.8 illustrates the need for a large number of SNPs when estimating coefficients this small. Figure 4.9 shows that the tails of the distribution for this level of relatedness and the previous example (individuals 4 & 21) start to overlap, but data in Tables 4.7 and 4.9 show that the ±2 standard deviation areas of the two distributions are non-overlapping.

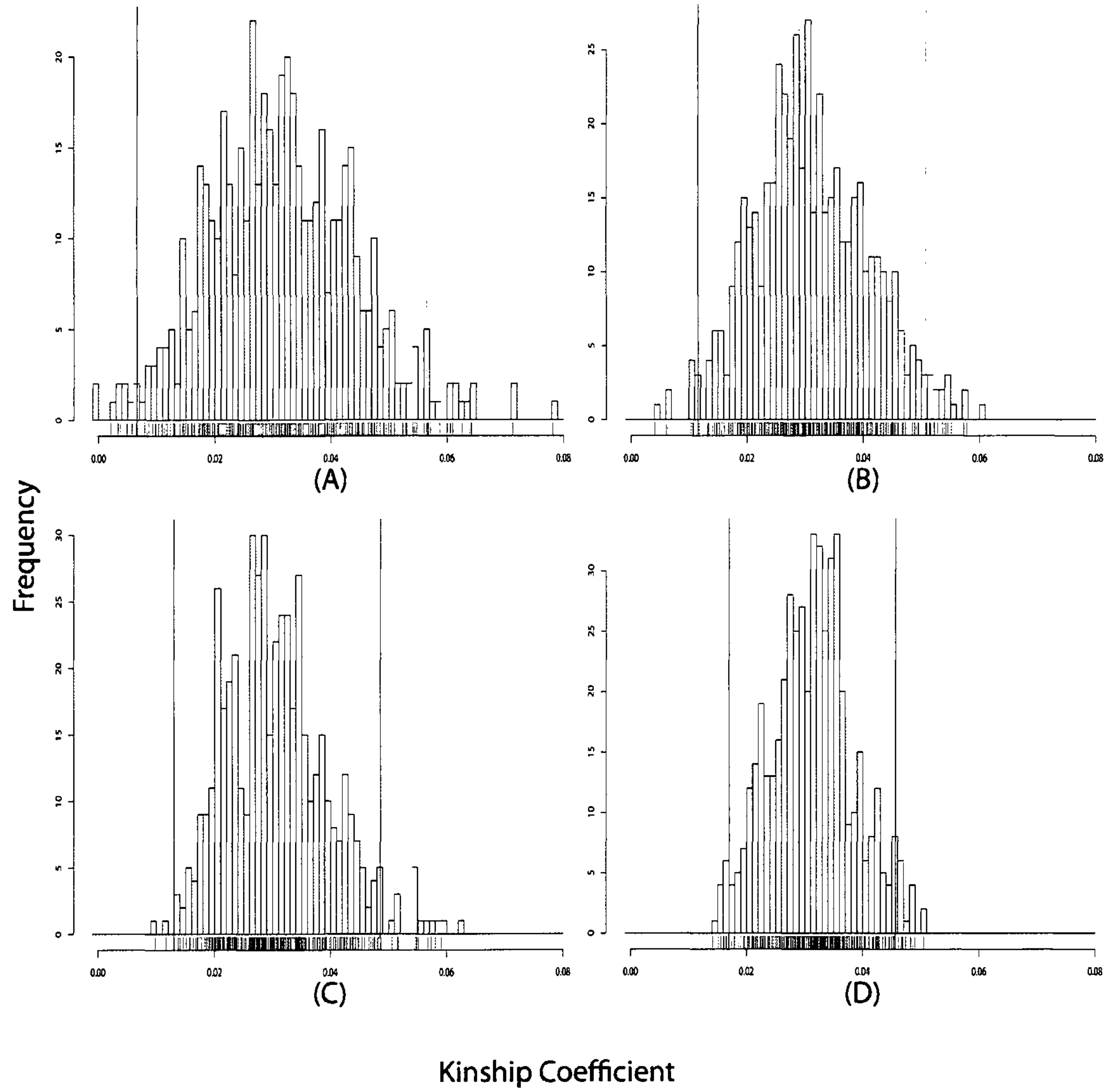


Figure 4.8: Individual 4 vs Individual 19 Kinship Coefficient Estimation, True Value = 0.03125: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

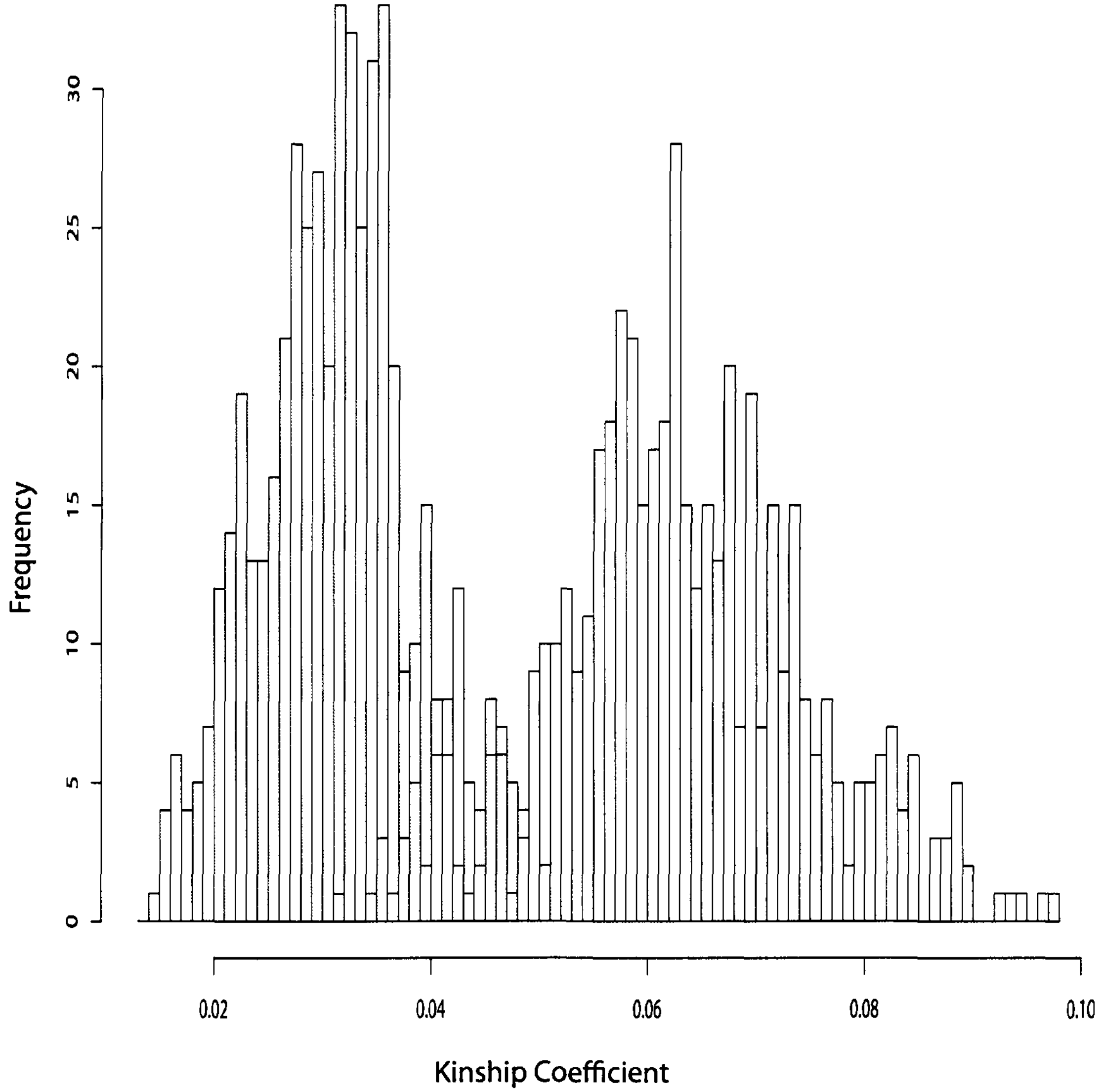


Figure 4.9: Comparison of Distributions of Kinship Estimates for 0.0625 true coefficient (red) and 0.03125 true coefficient (blue) for the 500K Chip

Chip	True	Min Value	-2 SD	Mean	+2 SD	Max	SD
10K	0.015625	-0.01541	-0.00505	0.01503	0.03512	0.04567	0.01004
100K	0.015625	0.00005471	0.00263	0.01559	0.02855	0.04016	0.00648
200K	0.015625	0.00165	0.00349	0.01515	0.02681	0.03488	0.00583
500K	0.015625	0.002539	0.00631	0.01579	0.02527	0.03324	0.00474

Table 4.11: Individual 7 vs Individual 21 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

The sixth level of relatedness analyzed has a theoretical kinship coefficient of 0.015625, and represents the relationship from one's self to one's grandparent's cousin. The individual pair of 7 and 21 represent this level of relatedness in the pedigree in Figure 4.1. As the relationships become more and more distant, the more important a large number of SNPs becomes for accurate estimation, as can be seen in Table 4.11 and Figure 4.10. Although the mean estimate for all 4 chips is accurate to the third decimal place, the standard deviation for the 10K chip is quite large with the -2 standard deviation area covering negative coefficients, which overlaps with the estimate for unrelated pairs. But our method continues to be unbiased with the KS test unable to reject that the coefficients are drawn from a normal distribution for all 4 chips seen in Table 4.12. The 500K chip clearly performs the best as Figure 4.10 illustrates. For the 500K chip the distribution for this relationship and the previous one (individual pair 4 & 19) show clear overlap, with the +2 SD area of this distribution overlapping with the -2 SD area of the previous distribution as seen in Figure 4.11. However, when considering the 500K chip, the distribution for this relationship shows a separation between it and the distribution for the unrelated pair analyzed above (individuals 1 & 2) as seen in Figure 4.12.

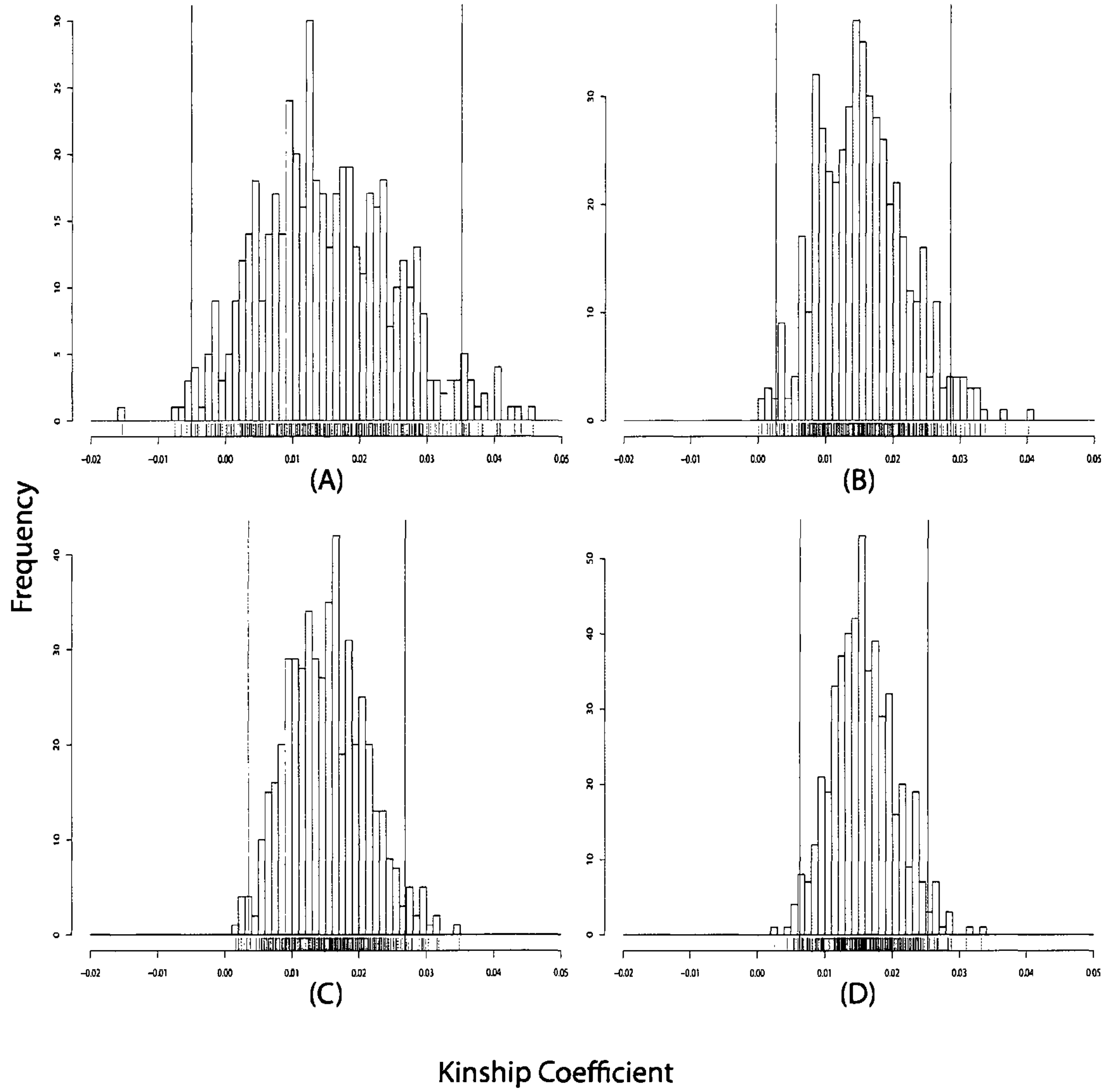


Figure 4.10: Individual 7 vs Individual 21 Kinship Coefficient Estimation, True Value = 0.015625: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

Chip	D	P-Value
10K	0.04249858	0.3271328
100K	0.04110611	0.3668298
200K	0.03798183	0.4663846
500K	0.04268127	0.3221485

Table 4.12: Individual 7 vs Individual 21 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

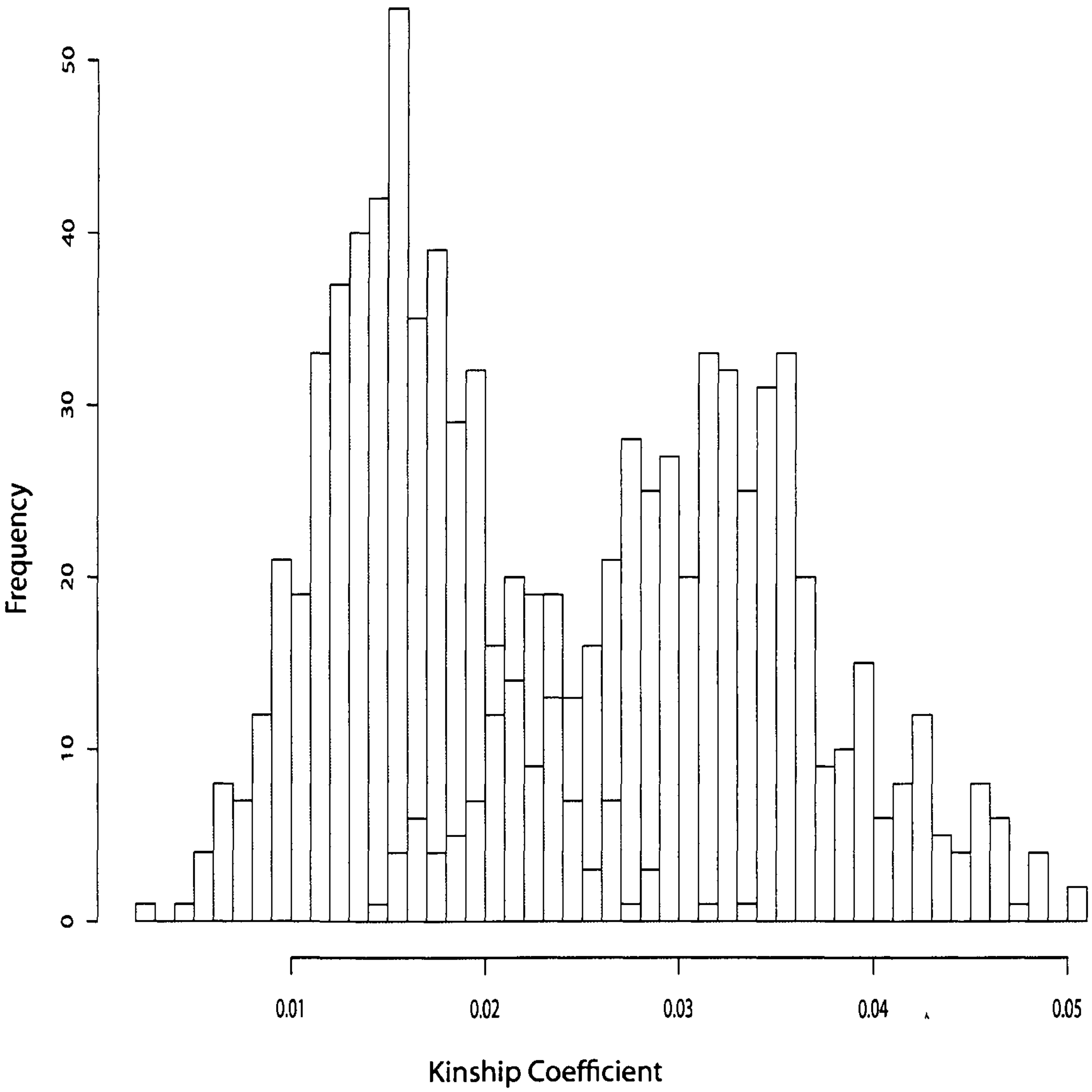


Figure 4.11: Comparison of Distributions of Kinship Estimates for 0.03125 true coefficient (red) and 0.015625 true coefficient (blue) for the 500K Chip

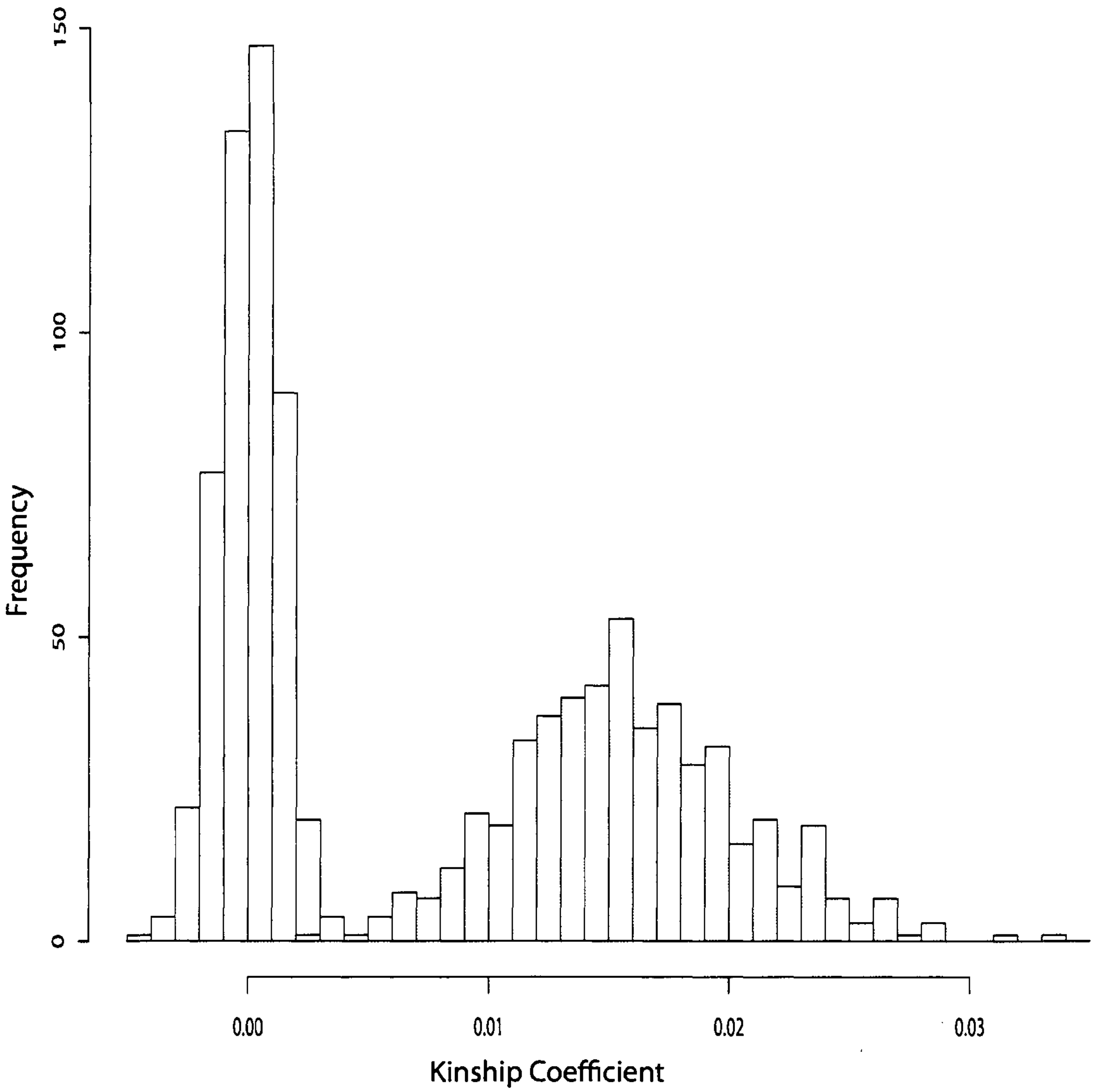


Figure 4.12: Comparison of Distributions of Kinship Estimates for 0 true coefficient (red) and 0.015625 true coefficient (blue) for the 500K Chip

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.0078125	-0.01725	-0.01032	0.007738	0.02580	0.03445	0.00903
100K	0.0078125	-0.003156	-0.00218	0.007441	0.01706	0.0229	0.00481
200K	0.0078125	-0.003564	-0.00119	0.007413	0.01602	0.02254	0.00430
500K	0.0078125	0.0004696	0.00109	0.007839	0.01458	0.02185	0.00337

Table 4.13: Individual 14 vs Individual 21 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

The seventh level of relatedness analyzed has a theoretical kinship coefficient of 0.0078125, which is a very distant relationship illustrated by the individual pair 14 & 21 in the pedigree in Figure 4.1. Even at this very distant genetic relatedness, the mean estimate from our method is accurate to the third decimal place for all 4 chips as seen in Table 4.13. In fact the 500K chip is still performing so well that its mean estimate is accurate to the 4th decimal place. The tight distribution for the 500K chip and its clear outperformance of the other chips is illustrated in Figure 4.13. The method remains normally distributed according to the KS test results shown in Table 4.14. As the data in Tables 4.11 and 4.13 show there is clear overlap in the distributions between this level of relatedness and the distribution for individuals 7 & 21. This level of relatedness is also starting to approach the lower bound of our method's ability to distinguish related from unrelated individuals even in the 500K chip. Figure 4.14 shows that there is overlap between this distribution and the distribution of estimates for unrelated individuals 1 & 2 when using the 500K chip, and in fact Tables 4.13 and 4.1 show that the -2 SD area of this distribution overlaps with the +2 SD area of the unrelated distribution.

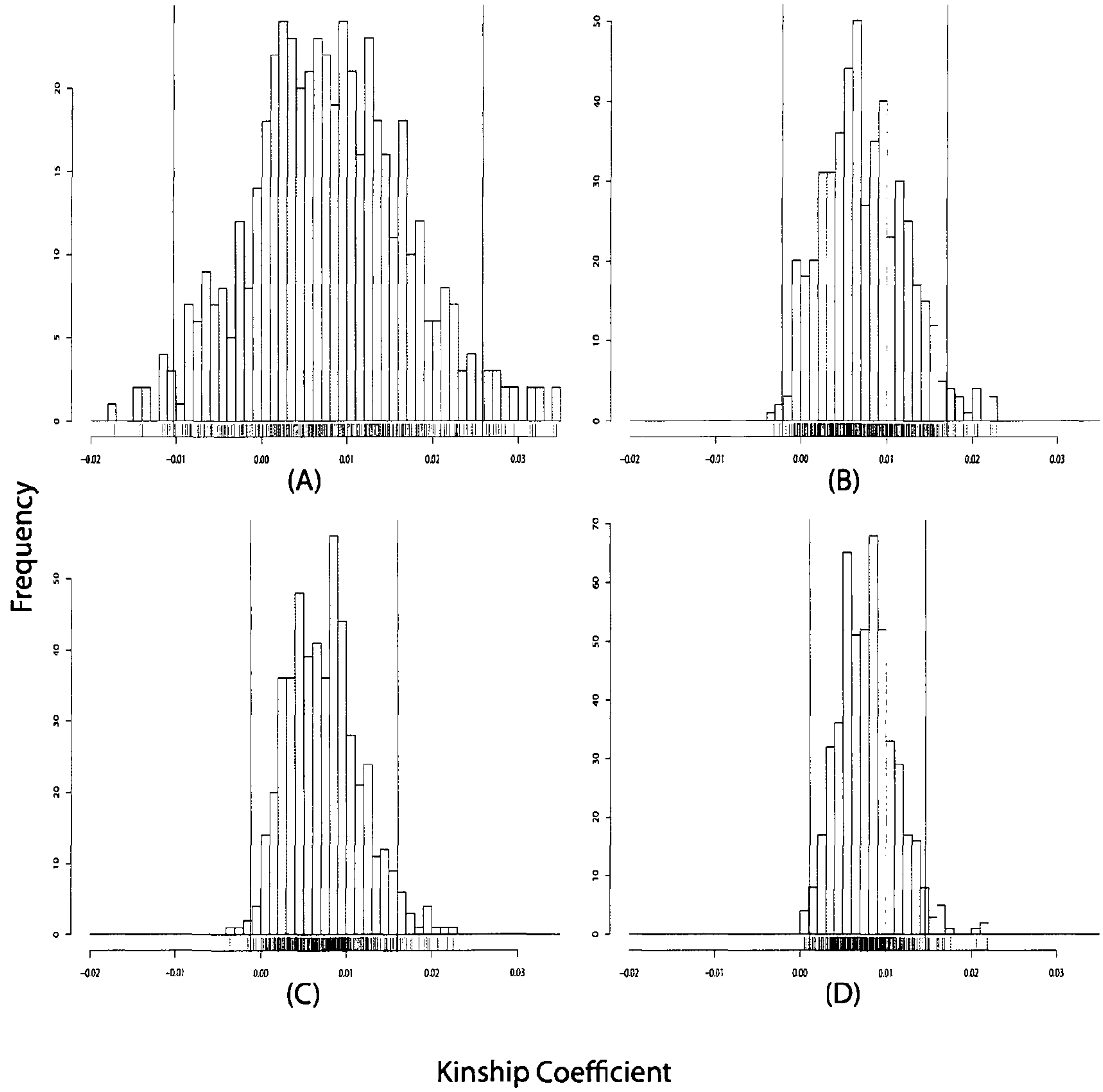


Figure 4.13: Individual 14 vs Individual 21 Kinship Coefficient Estimation, True Value = 0.0078125: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

Chip	D	P-Value
10K	0.02704439	0.8579572
100K	0.05384713	0.1100817
200K	0.03973323	0.4088592
500K	0.03955475	0.4145268

Table 4.14: Individual 14 vs Individual 21 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

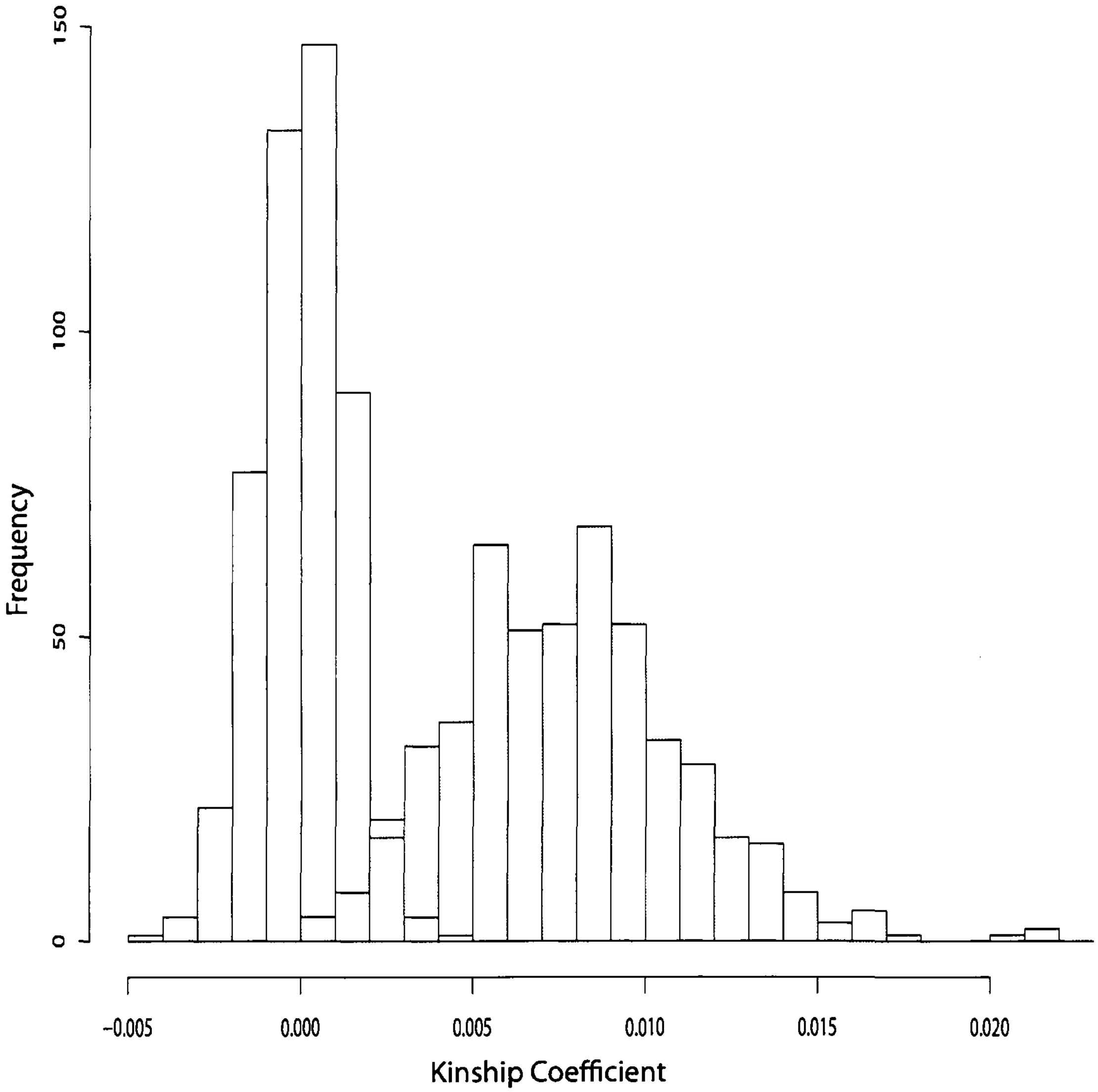


Figure 4.14: Comparison of Distributions of Kinship Estimates for 0 true coefficient (red) and 0.0078125 true coefficient (blue) for the 500K Chip

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.003906	-0.0174	-0.01219	0.003557	0.01930	0.03047	0.00787
100K	0.003906	-0.00649	-0.00414	0.00377	0.01168	0.02326	0.00396
200K	0.003906	-0.004102	-0.00301	0.003731	0.01047	0.01785	0.00337
500K	0.003906	-0.002393	-0.00113	0.004023	0.00917	0.01542	0.00257

Table 4.15: Individual 19 vs Individual 21 Kinship Coefficient Estimation Results by Chip Type. (SD = Standard Deviation)

Chip	D	P-Value
10K	0.03491145	0.5759316
100K	0.03793606	0.4679426
200K	0.06698928	0.02249599
500K	0.05490507	0.0987158

Table 4.16: Individual 19 vs Individual 21 Kolmogorov-Smirnov Test of Normality for Kinship Coefficient Estimation Results by Chip Type.

The eighth and final level of relatedness analyzed has a kinship coefficient of 0.003906, a very distant relationship illustrated by individuals 19 & 21 in the pedigree in Figure 4.1. Even for this very distantly related pair, the method still performs quite well as shown in Table 4.15. Figure 4.15 illustrates that to detect this distant of a relationship with our method that the 500K chip is necessary. But even with the 500K chip Figure 4.14 and Tables 4.15 and 4.1 demonstrates that there is significant overlap between the distribution for this pair and the distribution of coefficients for the unrelated pair 1 & 2. The unbiased nature of our statistic is upheld with the KS test being unable to reject that the coefficients are drawn from a normal distribution, except for the 200K chip if you use a p-value cutoff of 0.05. Figure 4.15 shows that the reason for this might be the extended right tail of the distribution, but this could just be an artifact of sampling.

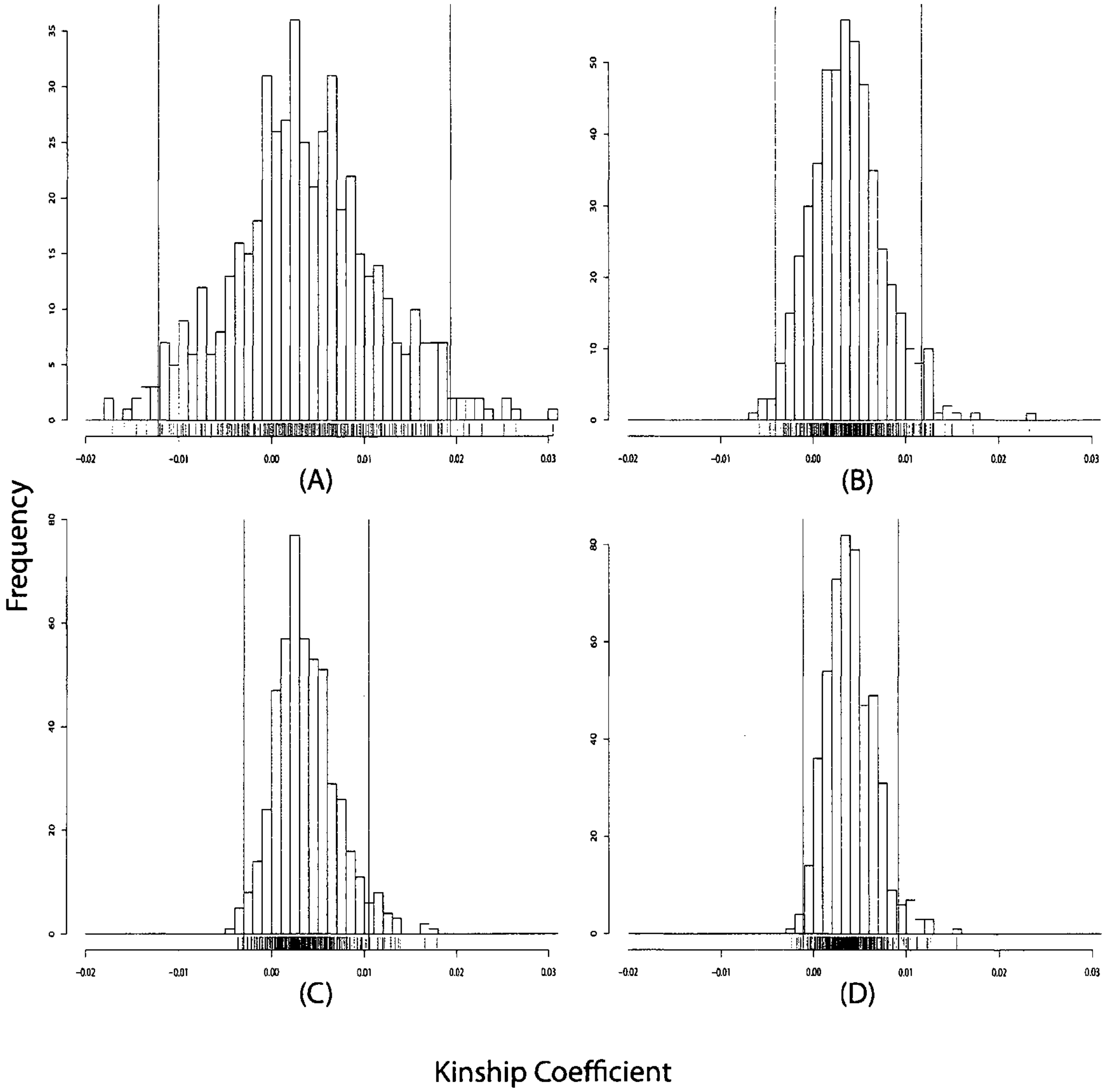


Figure 4.15: Individual 19 vs Individual 21 Kinship Coefficient Estimation, True Value = 0.003906: (A) 10K Chip; (B) 100K Chip; (C) 200K Chip; (D) 500K Chip. (The left and right red vertical lines represent the ±2 Standard Deviation area)

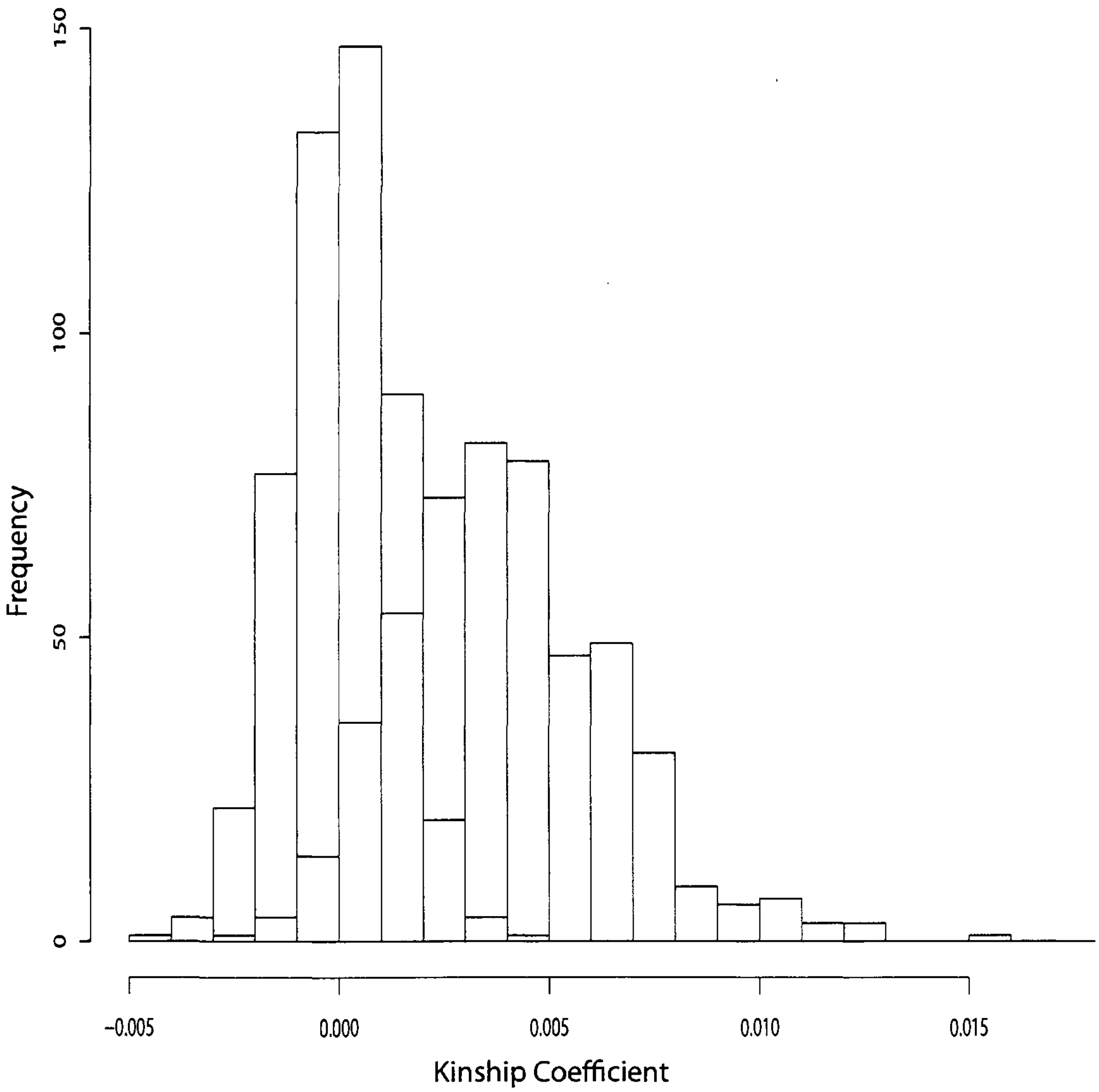


Figure 4.16: Comparison of Distributions of Kinship Estimates for 0 true coefficient (red) and 0.003906 true coefficient (blue) for the 500K Chip

The above results show that with correctly specified allele frequencies our estimator performs very well. But the formulation of our method is heavily dependent on the specified allele frequencies. To assess how allele frequency misspecification affects our methods performance we conducted extensive simulation studies, as detailed in the materials and methods section. We are only going to show and discuss the results for a single relationship pair because the affect on the methods performance is the same no matter the relationship investigated. For the purpose of discussing allele frequency misspecification we will be discussing individual pair 4 & 7, who have a theoretical kinship coefficient of 0.125.

The first analysis was to determine the behavior of our kinship estimates when the analysis major allele frequencies are greater than the actual major allele frequencies. Table 4.17 shows the results when the major allele frequencies used in the analysis of the data are 1% greater than the major allele frequencies used to generate the data at 25% of the loci on a chip. This allele frequency misspecification shifts the mean of the distribution to underestimate the level of relatedness, but does not affect the size of the standard deviation of the distribution as can be seen by comparison with Table 4.5. The next analysis was to have the analysis major allele frequency greater than the actual major allele frequency by 1% at 50% of the loci on the chip. The results of this analysis are shown in Table 4.18 and show a further downward shift in the mean of the distribution without affecting the standard deviation. Finally, we set the analysis major allele frequencies greater than the actual major allele frequencies by 1% at 100% of the loci. The results of this analysis are shown in Table 4.19 and reveal a very dramatic downward shift in the mean kinship coefficient estimate without affecting the structure of the distribution. These results are exactly what would be expected from an

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.07307	0.08492	0.11819	0.15145	0.16682	0.0166
100K	0.125	0.07577	0.08838	0.11538	0.14238	0.15171	0.0135
200K	0.125	0.08326	0.09202	0.11769	0.14335	0.15699	0.0128
500K	0.125	0.09157	0.09660	0.11676	0.13690	0.14299	0.0101

Table 4.17: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency > Simulated by 1 percent at 25 percent of the loci

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.06583	0.07941	0.11291	0.14639	0.16256	0.0167
100K	0.125	0.06730	0.07975	0.10705	0.13433	0.14345	0.0136
200K	0.125	0.07336	0.08132	0.10860	0.13588	0.14757	0.0136
500K	0.125	0.08276	0.08951	0.10896	0.12840	0.13583	0.0097

Table 4.18: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency > Simulated by 1 percent at 50 percent of the loci

analysis of equation 4.5 used in the calculation. As the major allele frequency increases so does the second term in the numerator and denominator which is the sum of the squares of all the allele frequencies. This term down weights the observed number of IBS matches because more are expected and therefore reduces the estimate of the relatedness.

The second analysis was to determine the behavior of our kinship estimates when the analysis major allele frequencies are less than the actual major allele

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.05546	0.06875	0.10234	0.13592	0.15205	0.0168
100K	0.125	0.05016	0.06269	0.09038	0.11806	0.12668	0.0138
200K	0.125	0.05290	0.06217	0.08871	0.11525	0.12973	0.0133
500K	0.125	0.06579	0.07273	0.09246	0.11218	0.11968	0.0099

Table 4.19: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency > Simulated by 1 percent at All of the loci

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.08557	0.09628	0.12910	0.16192	0.17917	0.0164
100K	0.125	0.09296	0.10562	0.13221	0.15879	0.16888	0.0133
200K	0.125	0.10158	0.11003	0.13517	0.16031	0.17376	0.0126
500K	0.125	0.10885	0.11561	0.13398	0.15234	0.16069	0.0092

Table 4.20: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency < Simulated by 1 percent at 25 percent of the loci

frequencies. Based upon the results of the above analyses and the analysis of the equation used to estimate the kinship coefficient, we expect that this analysis will show an upward shift in the mean of the distribution and inflation of the estimated relatedness. The logic is that the decreased major allele frequency will decrease the sum of the squares of the allele frequencies and up weights the observed number of IBS matches and inflate the estimate of relatedness. Table 4.20 shows the results when 25% of the loci on a chip have the analysis major allele frequency 1% less than the actual major allele frequency. The table does in fact show an upward shift in the mean of the distribution with no affect on the standard deviation. The same upward shift in the mean of the distribution is observed when 50% (Table 4.21) and 100% (Table 4.22) of the loci on the chip have the analysis major allele frequency less than the actual major allele frequency by 1%. The method behaves in a predictable manner in terms of the affects on the mean of the distribution of kinship estimates when the allele frequencies are uniformly misspecified. But we do not believe that these examples simulate the types of allele frequency misspecification that would occur in real data.

To analyze the performance of the method in a more realistic setting we set a normally distributed error distribution on the major allele for each locus. This situation will result in having a mixture of loci allele misspecification. Tables

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.09111	0.10207	0.13455	0.16703	0.18386	0.0162
100K	0.125	0.10183	0.11429	0.14062	0.16694	0.17743	0.0134
200K	0.125	0.10987	0.11811	0.14304	0.16796	0.18073	0.0125
500K	0.125	0.11749	0.12370	0.14239	0.16107	0.16882	0.0093

Table 4.21: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency < Simulated by 1 percent at 50 percent of the loci

Chip	True	Min Value	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.10035	0.11351	0.14562	0.17772	0.19480	0.0161
100K	0.125	0.11842	0.13170	0.15752	0.18334	0.19294	0.0129
200K	0.125	0.12660	0.13430	0.15878	0.18326	0.19586	0.0122
500K	0.125	0.13515	0.14116	0.15946	0.17776	0.18546	0.0092

Table 4.22: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; Analysis major allele frequency < Simulated by 1 percent at All of the loci

4.23, 4.24, 4.25 and 4.26 shows the affects on the kinship estimates from setting the standard deviation of the error distribution for each locus to (0.025×major allele frequency), (0.05×major allele frequency), (0.075×major allele frequency) and (0.10×major allele frequency) respectively. The expectation is that for each analysis approximately half the loci will have the analysis major allele frequency greater than the actual allele frequency and that approximately half the loci will have the analysis major allele frequency less than the actual allele frequency. This was not actually the case. Although the skewness was not dramatic, in all the analyses the were more loci where the analysis major allele frequency was greater than the actual allele frequency, and this fact accounts for the mean of the coefficient distribution being shift towards underestimating the relatedness of the individuals in all the analyses. The reason for this bias is that the SNPs on the chips are biased towards loci with large major allele frequencies and in our

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.07805	0.09096	0.12367	0.15637	0.17435	0.0163
100K	0.125	0.07925	0.09238	0.11934	0.14628	0.15470	0.0135
200K	0.125	0.06907	0.07950	0.10539	0.13127	0.14538	0.0129
500K	0.125	0.09810	0.10513	0.12412	0.14313	0.15066	0.0095

Table 4.23: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; 2.5 percent standard deviation in distribution of major allele frequency misspecification

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.07151	0.08962	0.12245	0.15527	0.17236	0.0164
100K	0.125	0.06598	0.08119	0.10846	0.13572	0.14514	0.0136
200K	0.125	0.04417	0.05395	0.08013	0.10631	0.11409	0.0131
500K	0.125	0.08877	0.09485	0.11412	0.13284	0.14232	0.0096

Table 4.24: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; 5 percent standard deviation in distribution of major allele frequency misspecification

sampling we had to reject allele frequencies > 1, therefore for the loci with very large major allele frequencies the probability of sampling frequency larger than the actual allele frequency and < 1 was very small. An interesting observation is that the 10K and 500K chips appear to be more robust against allele misspecifications than the 100K and 200K chip. We are not sure why, but might have something to due with the criteria Affymetrix employed in selecting the SNPs to include on their chips.

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.07009	0.08464	0.11745	0.15025	0.16682	0.0164
100K	0.125	0.05182	0.06487	0.09264	0..1204	0.12901	0.0139
200K	0.125	0.01709	0.02645	0.05385	0.08123	0.09608	0.01379
500K	0.125	0.07242	0.07937	0.09881	0.11825	0.12637	0.0097

Table 4.25: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; 7.5 percent standard deviation in distribution of major allele frequency misspecification

Chip	True	Min	-2 SD	Mean	+2 SD	Max	SD
10K	0.125	0.06697	0.08416	0.11711	0.15005	0.16554	0.0165
100K	0.125	0.03205	0.04576	0.07400	0.10222	0.11135	0.0141
200K	0.125	-0.01301	-0.00294	0.02523	0.05339	0.06977	0.0141
500K	0.125	0.05347	0.06081	0.08077	0.10073	0.10837	0.001

Table 4.26: Individual 4 vs Individual 7 Kinship Coefficient Estimation Results by Chip Type; 10 percent standard deviation in distribution of major allele frequency misspecification

These analyses of our methods performance under allele frequency misspecification are important to note and understand. They suggest that much effort should be employed to ensure the correct estimation of the allele frequencies for any population being analyzed by this method. The results are also not surprising given that the allele frequencies play an important role in our estimation technique.

4.3.2 Conditional Kinship Coefficient Estimation

The penalized optimization technique we propose for estimating the conditional kinship coefficient for each SNP on a chip has two penalty terms, λ_1 and λ_2. The λ_1 parameter is to penalize the point estimate at a SNP for being different than the chromosome specific theoretical kinship coefficient. The λ_2 parameter is to penalize neighboring SNPs for belonging to different conditional kinship coefficient sets. The first step for our method is to find the optimal λ_1, λ_2 combination that minimizes the error between the true conditional kinship coefficient and our estimated conditional kinship coefficient in a variety of IBD sharing configurations and relationships. We chose four relationships from the pedigree structure (Figure 4.1) that represented a wide range of relationships and would exhibit complex patterns of IBD sharing along the chromosomes. For each of these relationships we analyzed the first 100 replicates of the 500 replicates analyzed for the theoretical kinship estimation above, and performed a grid search of all combinations of λ_1 and λ_2 for λ_1 from 0-2 in increments of 0.2 and for λ_2 from 0 - 110 in increments of 5. For each replicate, for each SNP, for each pair of λ_1 and λ_2, we compared our estimated conditional coefficient to the true conditional coefficient. The true conditional coefficient was determined by performing gene-dropping in the program Mendel with its option to uniquely label all founder alleles so IBD sharing can be calculated exactly. For each replicate we calculated the average absolute difference (*aad*) for a chromosome as defined by equation 4.13. We wanted to find the λ_1, λ_2 combination that minimized the *aad* over the 100 replicates. We chose to use chromosome 21 of the 200K chip to perform the grid search to find the optimal λ_1, λ_2 combination. The 4 relationships we tested were individual pairs 3 & 4, 4 & 7, 4 & 19 and 19 & 21. The hope was to find a single λ_1, λ_2 combination that works well across all relationships and all chromosomes.

The results of the grid search for the sibling pair 3 & 4 can be seen in Figure 4.17. The λ_1, λ_2 combination that gave the minimum *aad* for this pair was λ_1 equal to 0 and λ_2 equal to 90, and designated by a red diamond in the figure. The optimal λ_1, λ_2 combination that gave the minimum *aad* for the individual pair 4 & 7 was λ_1 equal to 0.2 and λ_2 equal to 110 designated by a red diamond in Figure 4.18. The minimum *aad* for individual pair 4 & 19 occurred at λ_1 equal to 0 and λ_2 equal to 100 designated by a red diamond in Figure 4.19. The final pair analyzed, individuals 19 & 21, found the minimum *aad* with λ_1 equal to 0.6 and λ_2 equal to 90, designated by a red diamond in Figure 4.20. The optimal λ_1, λ_2 combination for all pairs are all very close. In order to simplify the calculation and increase the generality of the technique we would like a single λ_1, λ_2 combination. Therefore we decided to test to see if we used the optimal λ_1, λ_2 combination for individual pair 4 & 19 ($\lambda_1 = 0, \lambda_2 = 100$) for all pairs would the results be significantly different than if we used the true optimal λ_1, λ_2 for each pair. The average *aad* across the 100 replicates for the pairs 3 & 4, 4 & 7, and 19 & 21 using the combination $\lambda_1 = 0, \lambda_2 = 100$ is shown by a blue open circle on Figures 4.17, 4.18 and 4.20 respectively.

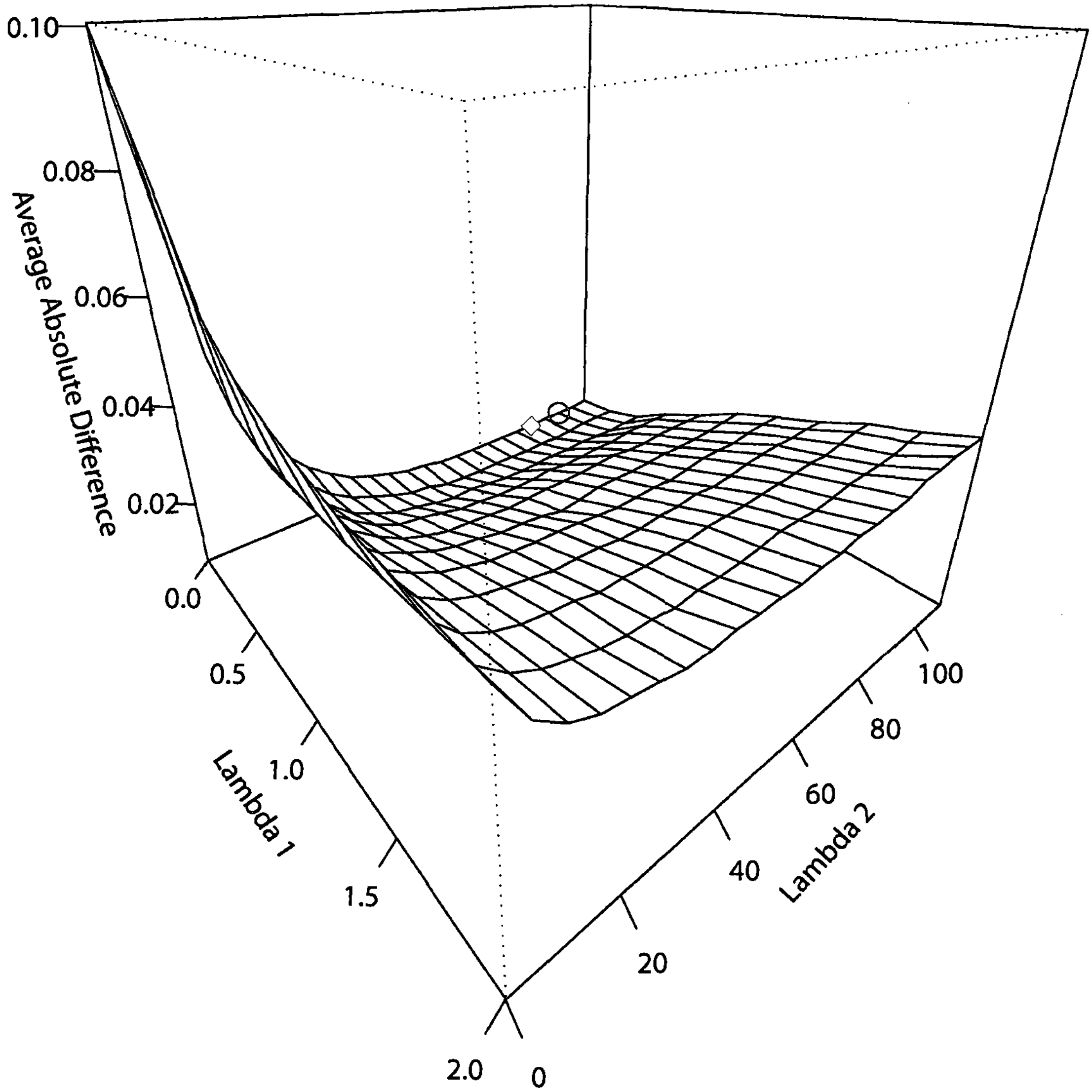


Figure 4.17: Optimization of λ_1 and λ_2 penalties for conditional kinship coefficient estimation for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) for Chr 21 on the 200K chip, 100 replicates. (Red diamond true minimum, Blue circle generalized penalties)

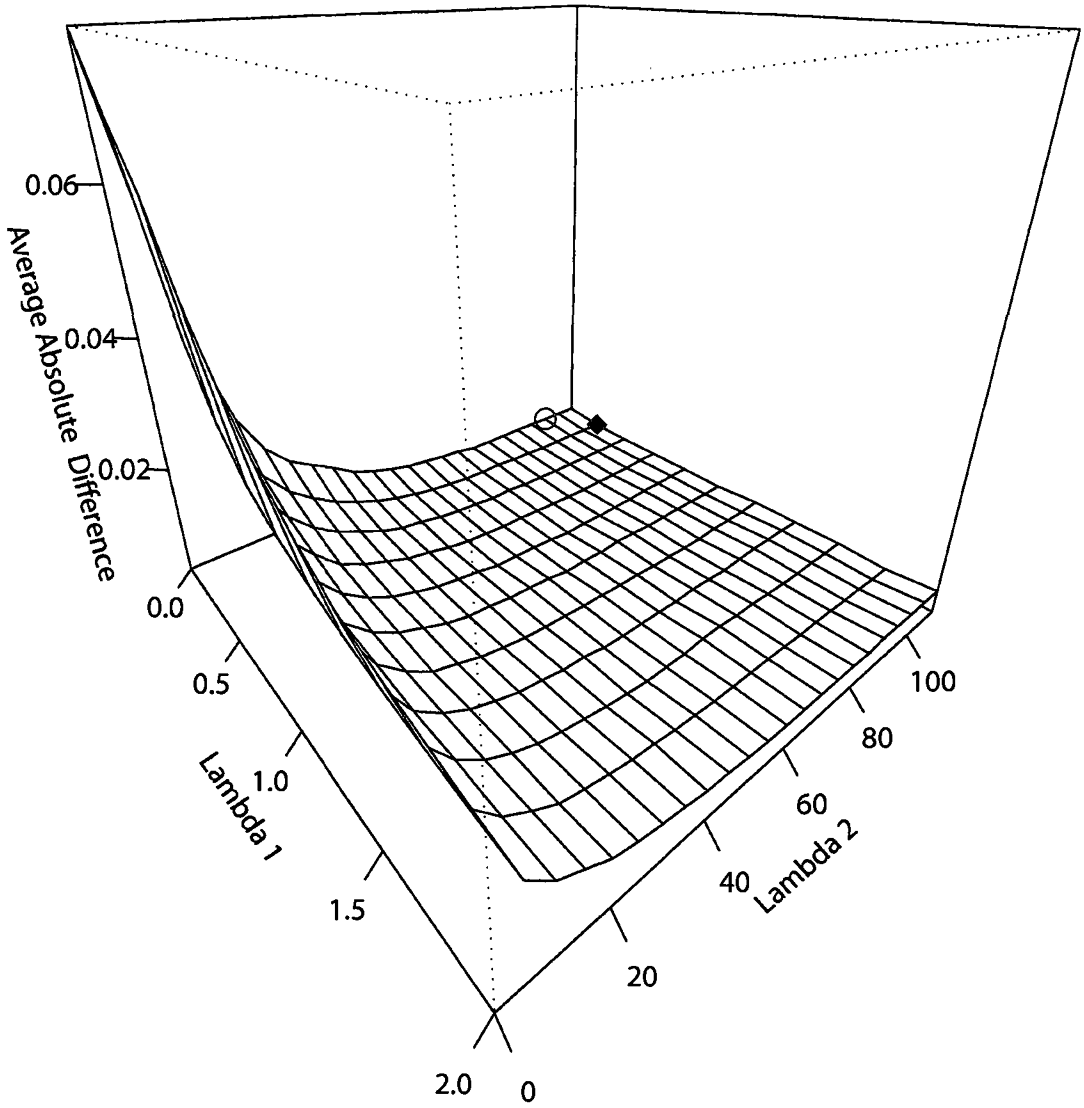


Figure 4.18: Optimization of λ_1 and λ_2 penalties for conditional kinship coefficient estimation for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) for 200K chip, 100 replicates. (Red diamond true minimum, Blue circle generalized penalties)

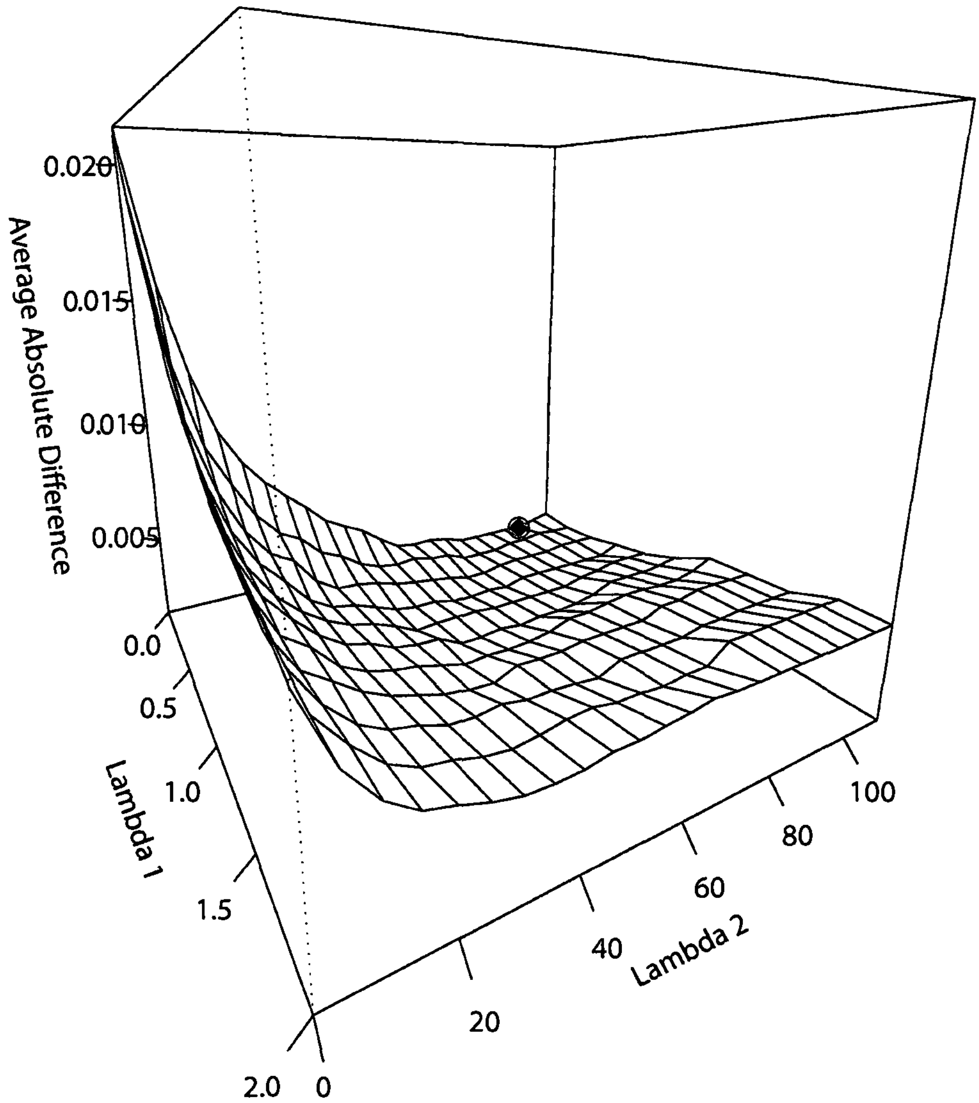


Figure 4.19: Optimization of λ_1 and λ_2 penalties for conditional kinship coefficient estimation for individual pair 4 & 19 (theoretical kinship coefficient = 0.03125) for 200K chip, 100 replicates. (Red diamond true minimum, Blue circle generalized penalties)

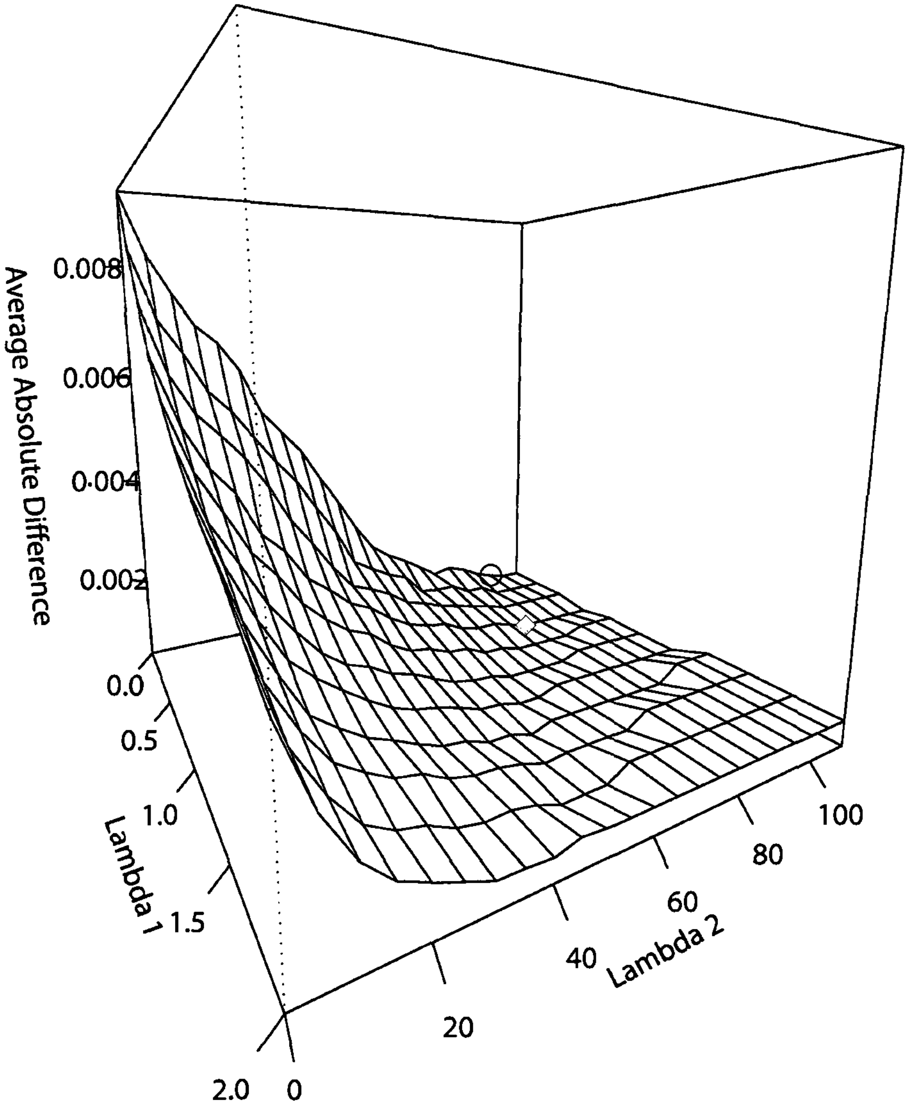


Figure 4.20: Optimization of λ_1 and λ_2 penalties for conditional kinship coefficient estimation for individual pair 19 & 21 (theoretical kinship coefficient = 0.003906) for 200K chip, 100 replicates. (Red diamond true minimum, Blue circle generalized penalties)

Figure 4.21 compares the distribution of the *aad* over 500 replicates for the pair 3 & 4 using the true optimal lambdas (Figure 4.21A) and using the generalize lambdas (Figure 4.21B). As the figure shows, there is not a significant difference in the distribution of error between the two sets of lambdas. The mean of the distribution using the optimal lambda combination is 0.0076, while the mean with the generalized lambdas is 0.0080. Additionally, the two lambda sets both achieve zero error in 19 replicates and both have the same maximum error of 0.04034. Therefore we conclude that there is not a significant difference between the results using the optimal vs the generalized lambdas, and will use the generalized lambdas for all our analyses. Next we analyzed the individual pair 4 & 7 to determine if the proposed generalized lambda combination yields significantly different results from the optimal lambda combination. Figure 4.22 compares the distribution of *aad* for 500 replicates using the optimal lambda combination (4.22A) and the generalized combination (4.22B). The mean of the distribution using the optimal lambda combination is 0.003388 while the mean of using the generalized lambdas is 0.003397. The minimum and maximums of the distributions are also very similar leading us to conclude that the generalized lambdas do not give significantly different results from the optimal lambda combination. Finally we needed to determine if the generalized lambda combination will work for the most distantly related pair, individuals 19 & 21. The comparison of the distribution of *aad* for 500 replicates on chromosome 21 of the 200K chip is seen in Figure 4.23, and it reveals that the two distributions are virtually identical. Therefore we are going to use the generalized lambda combination for all future analysis.

The results from these analyses show that we can use a single λ_1, λ_2 combina-

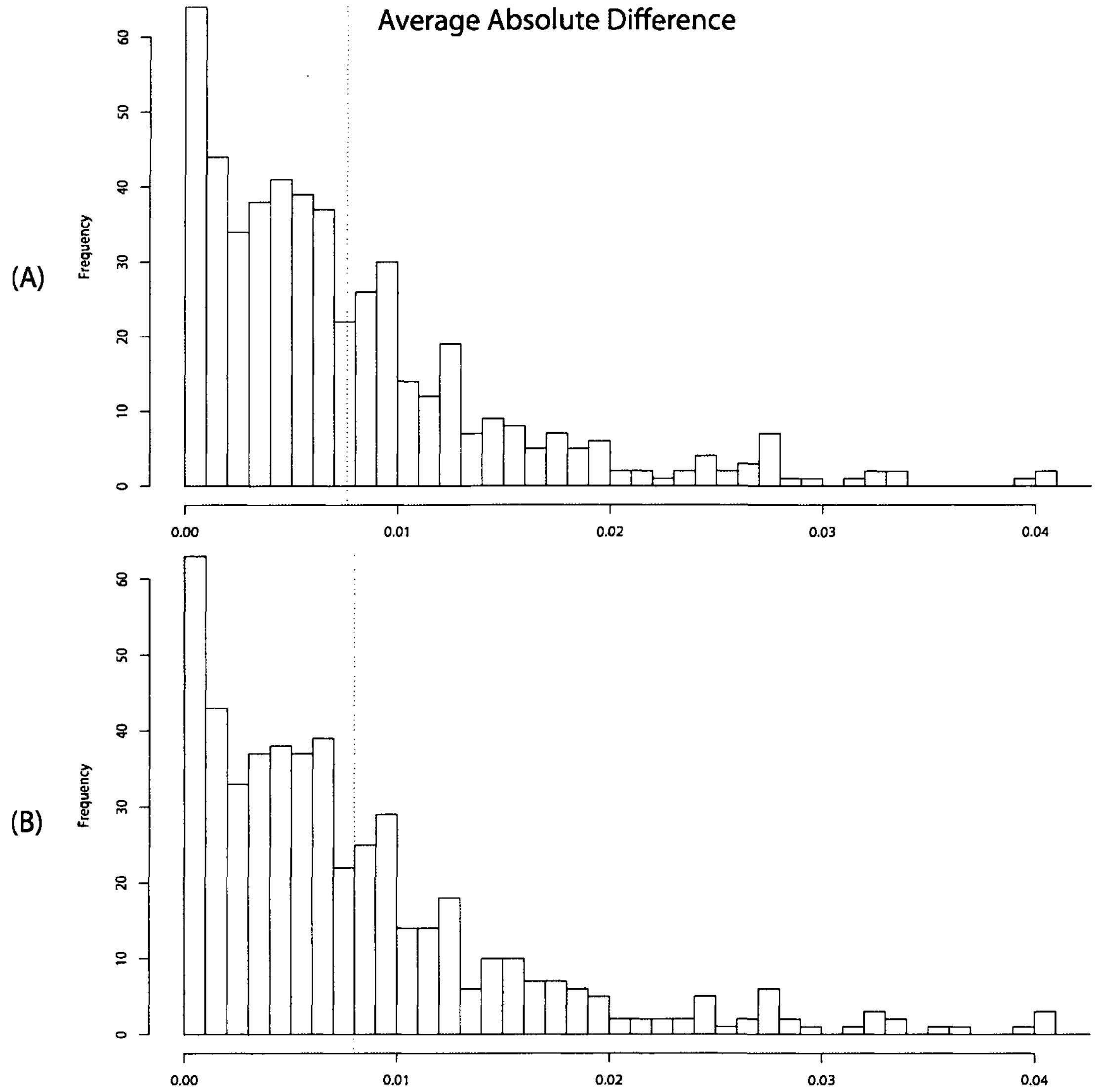


Figure 4.21: Comparison of Distribution of Average Absolute Difference (*aad*) of the conditional kinship coefficient estimation for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) for Chr 21 on the 200K chip, 500 replicates. (A) True Minimum λ_1, λ_2 (B) Generalized λ_1, λ_2. Red dotted line is at mean of distribution.

tion for estimating the conditional kinship coefficient for the entire spectrum of theoretical kinship coefficients that we can confidently estimate. The next step is to test the accuracy of our estimator on the 10K, 100K and 500K chips while employing the coefficients we optimized for the 200K chip.

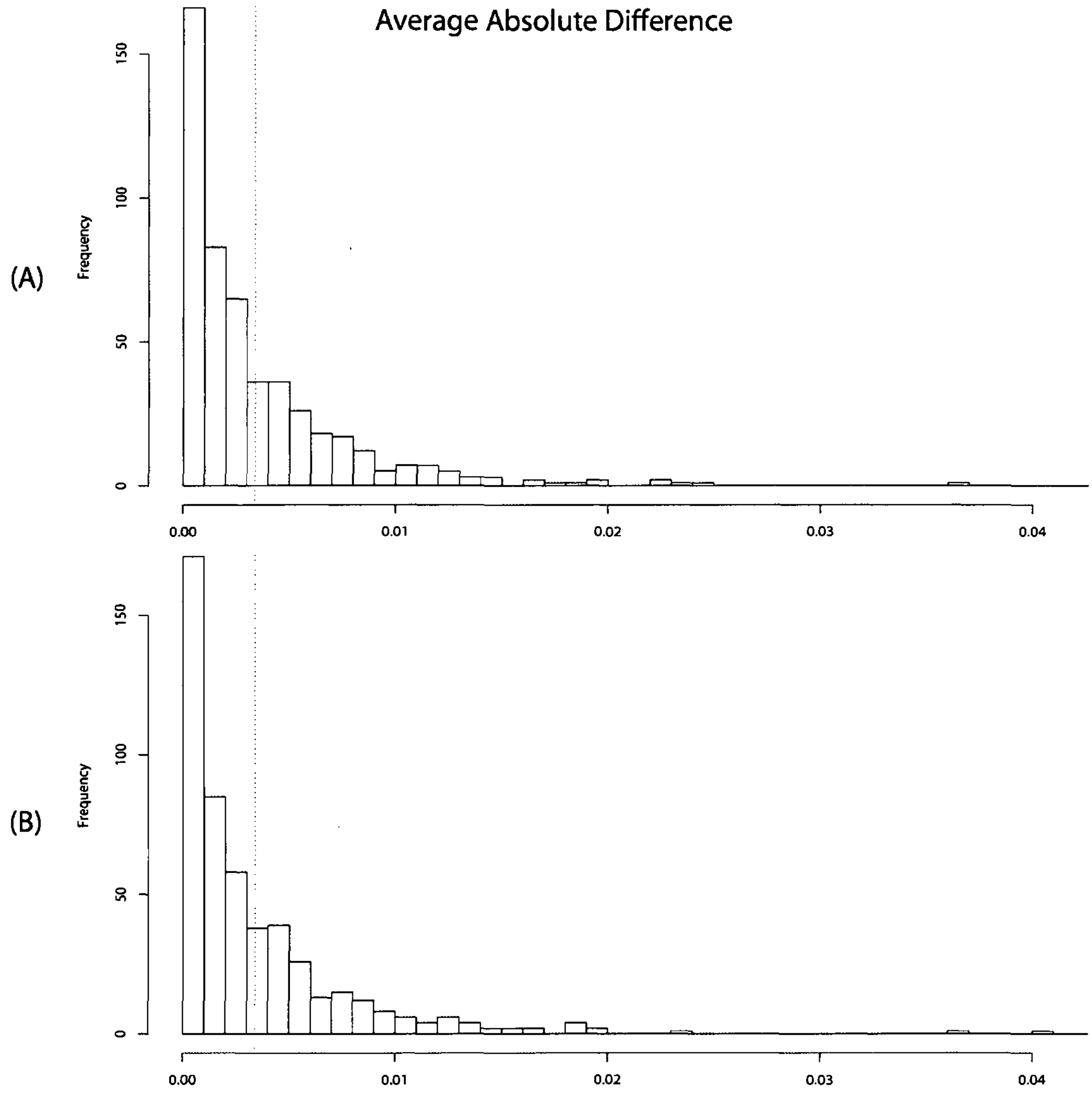


Figure 4.22: Comparison of Distribution of Average Absolute Difference (*aad*) of the conditional kinship coefficient estimation for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) for Chr 21 on the 200K chip, 500 replicates. (A) True Minimum λ_1, λ_2 (B) Generalized λ_1, λ_2. Red dotted line is at mean of distribution.

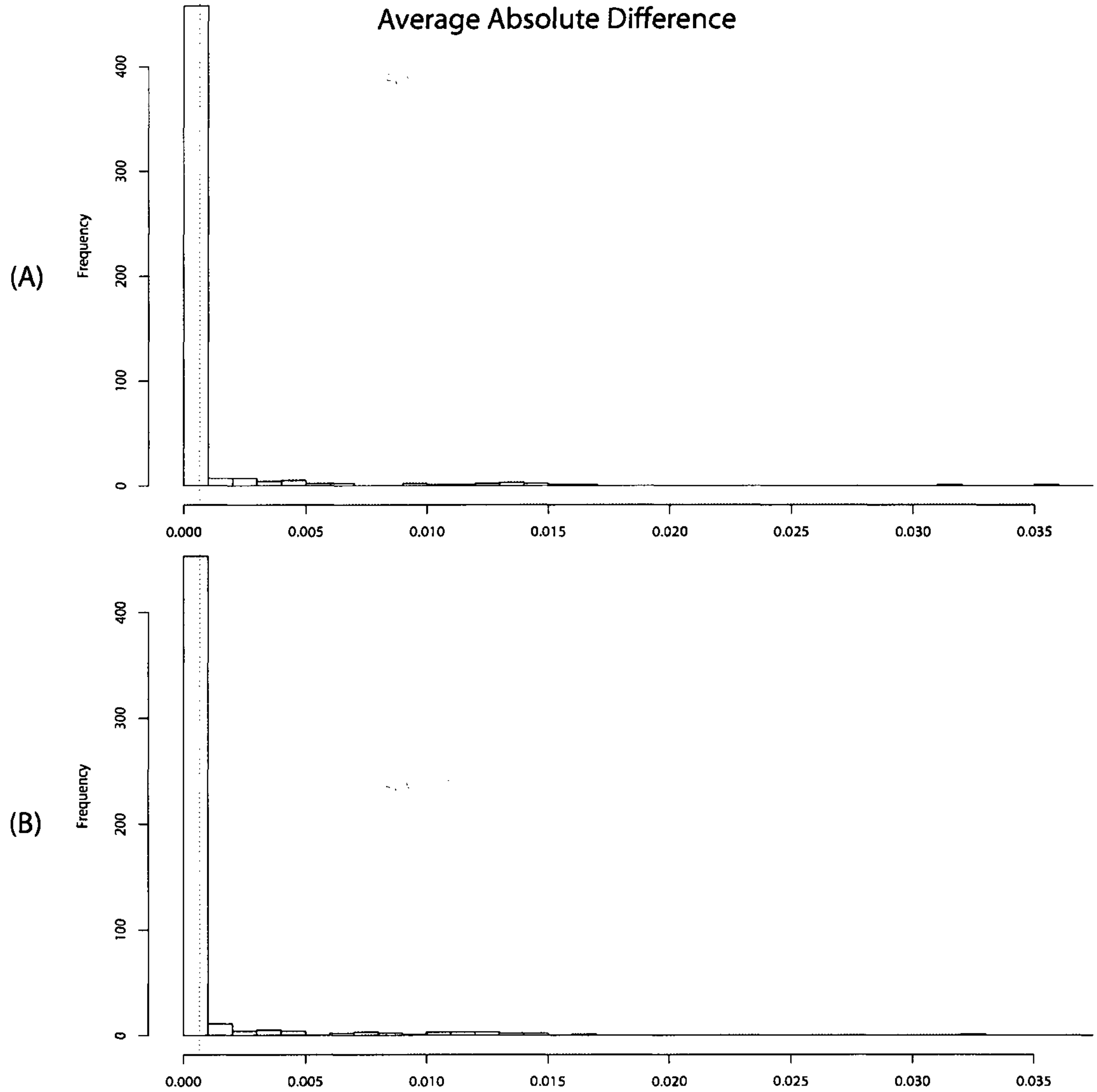


Figure 4.23: Comparison of Distribution of Average Absolute Difference (*aad*) of the conditional kinship coefficient estimation for individual pair 4 & 19 (theoretical kinship coefficient = 0.003906) for Chr 21 on the 200K chip, 500 replicates. (A) True Minimum λ_1, λ_2 (B) Generalized λ_1, λ_2. Red dotted line is at mean of distribution.

Figure 4.24 shows the distribution of the *aad* over 500 replicates for individual pair 3 & 4 using the generalized lambdas for chromosome 21 for the 10K (4.24A), 100K (4.24B), 200K (4.24C) and 500K (4.24D). As the distributions reveal, the estimation of the conditional kinship coefficient is best using the 500K chip. Figure 4.25 shows a conditional kinship coefficient plot for all 7,143 SNPs on chromosome 21 on the 500K chip for the replicate with approximately the mean *aad* of the distribution shown in Figure 4.24D. As the figure reveals we do an extremely good job of estimation, and only assign 12 of the 7,143 SNPs incorrect conditional kinship coefficients. Figure 4.26 shows the distribution of *aad* on chromosome 21 for individual pair 4 & 7 for the 10K (4.26A), 100K (4.26B), 200K (4.26C) and 500K (4.26D) chips. The 500K chip performs the best and Figure 4.27 shows the conditional kinship coefficient plot for the 7,143 SNPs on chromosome 21 on the chip for a replicate with approximately the mean *aad.* Our estimation method works well and only incorrectly assigns 58 SNPs to the wrong IBD set. The distribution of *aad* for individual pair 4 & 19 for 500 replicates on chromosome 21 for the 4 chips is found in Figure 4.28. The figure shows that all 4 chips are performing well, but that the 500K chip still outperforms the rest of the chips. The comparison of the true and estimated conditional kinship coefficients for the replicate with approximately the mean *aad* is seen in Figure 4.28, and shows that the method only misassigns 29 of the 7,143 SNPs. Figure 4.30 shows how this lambda combination performs at estimating the conditional kinship coefficient for individual pair 19 & 21 on all 4 chips. The 500K chip performs the best and Figure 4.31 show the coefficient plot for a replicate with approximately the mean *aad* and shows that only 7 of the 7,143 SNPs were misassigned IBD status.

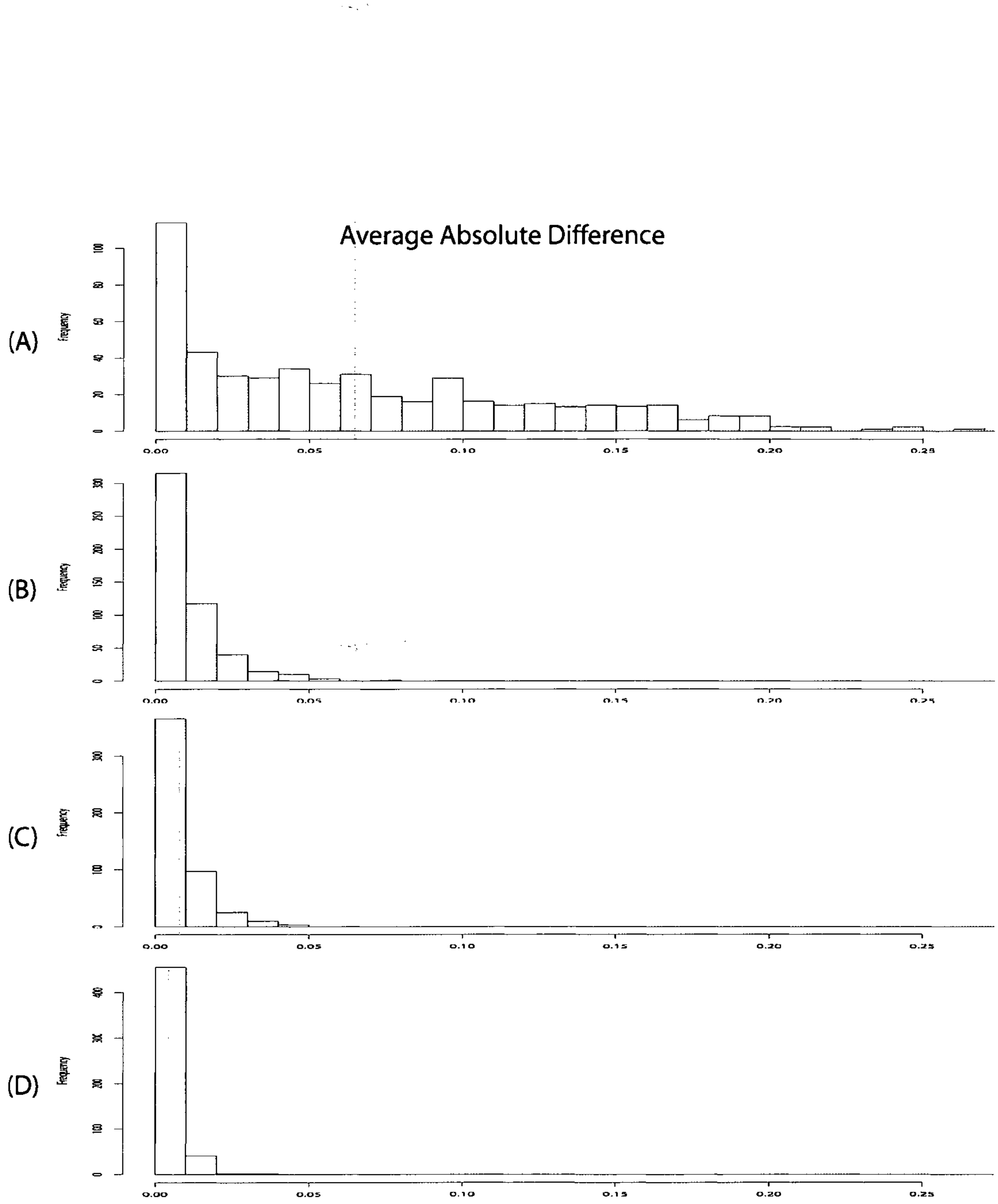


Figure 4.24: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) for Chr 21, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution

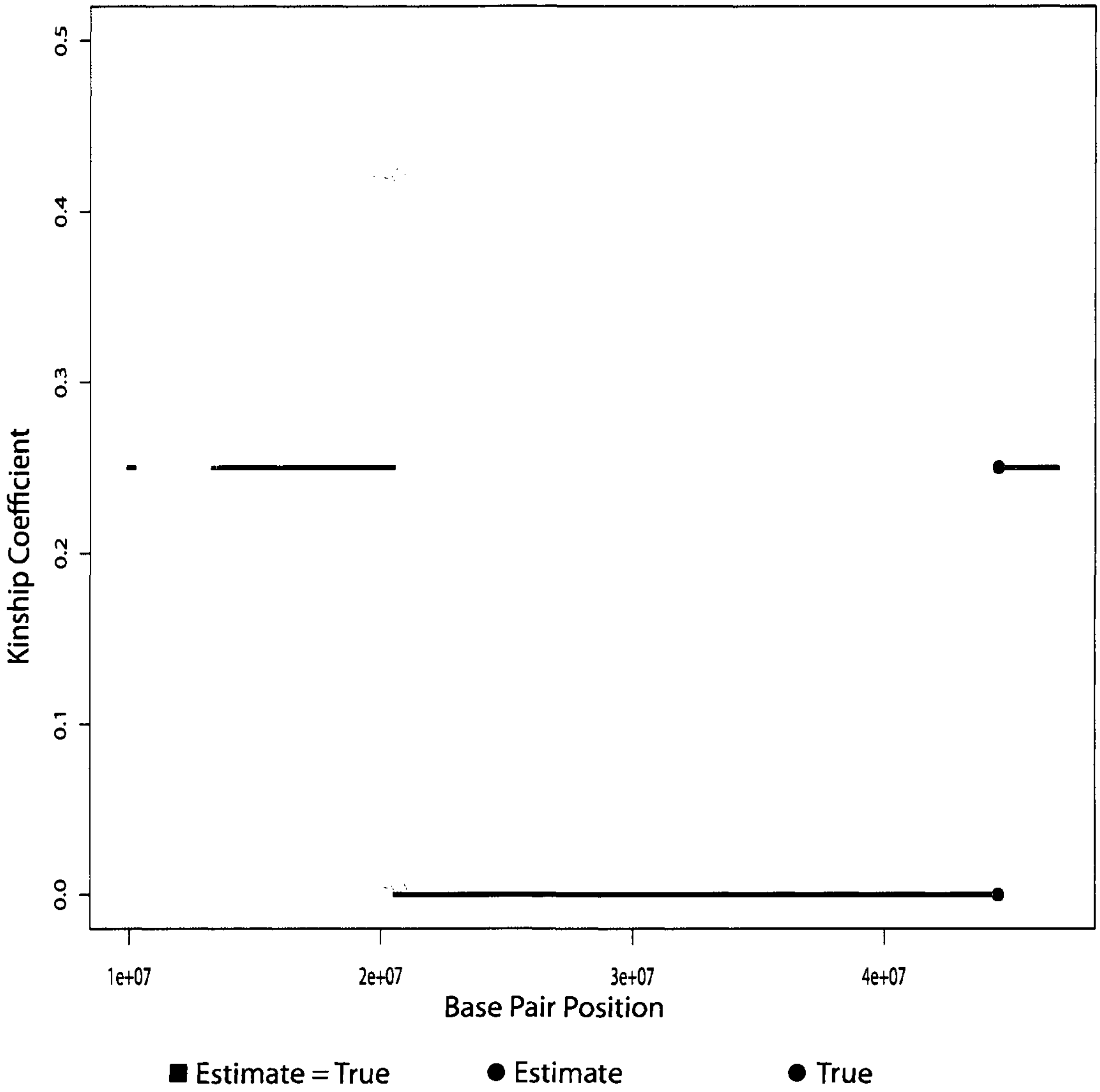


Figure 4.25: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) 500K chip, Chr 21. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

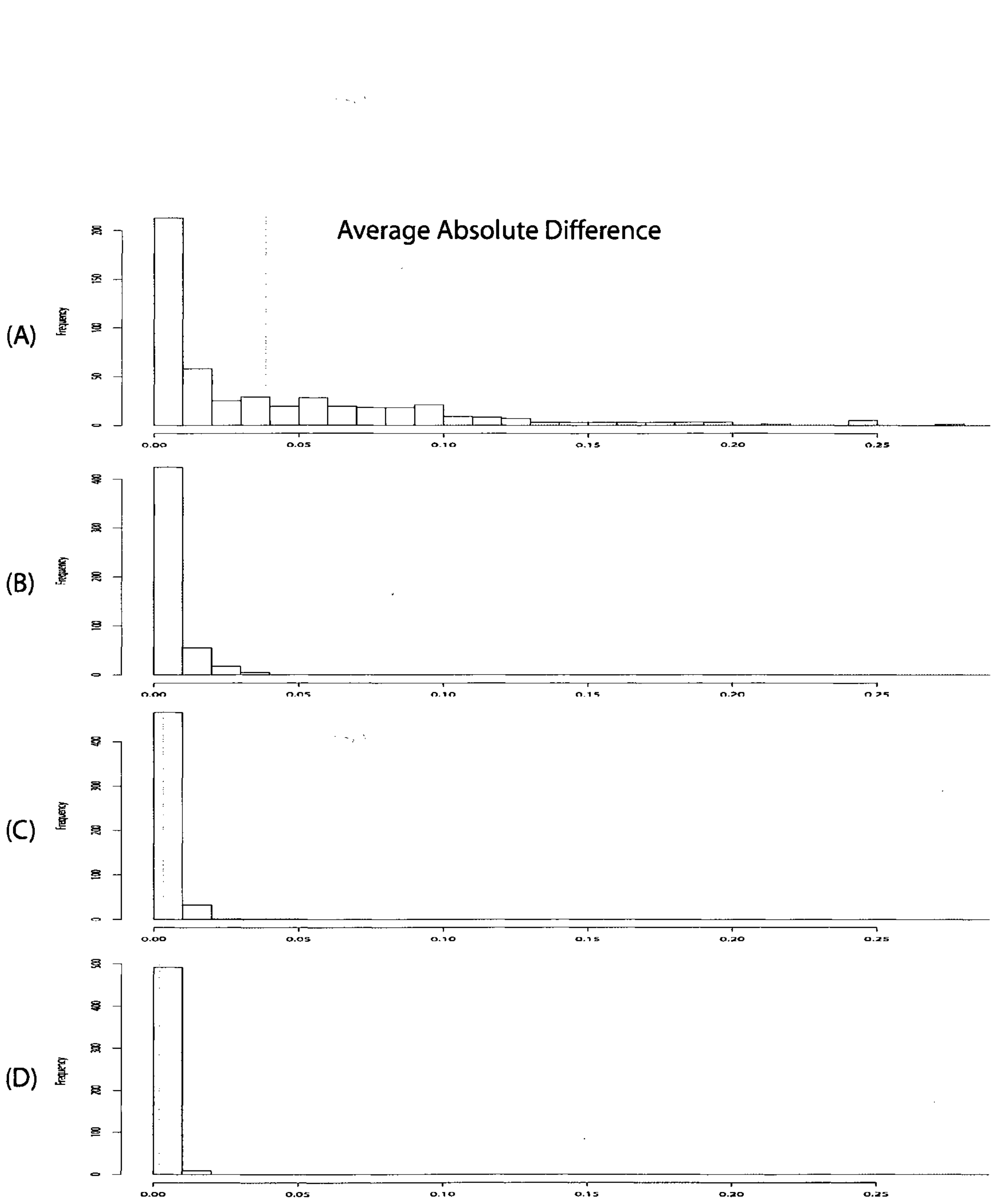


Figure 4.26: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) for Chr 21, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution

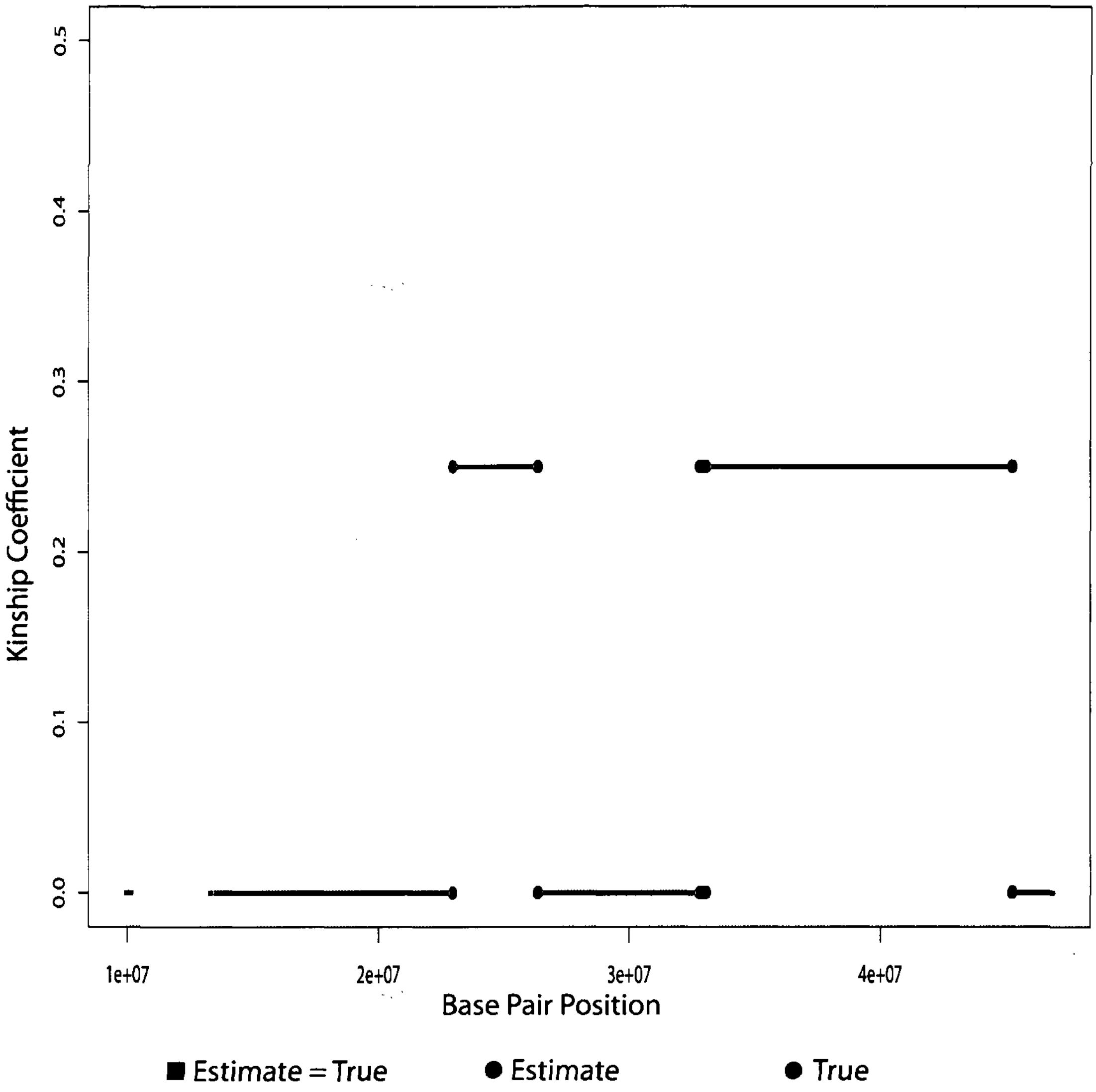


Figure 4.27: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) 500K chip, Chr 21. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

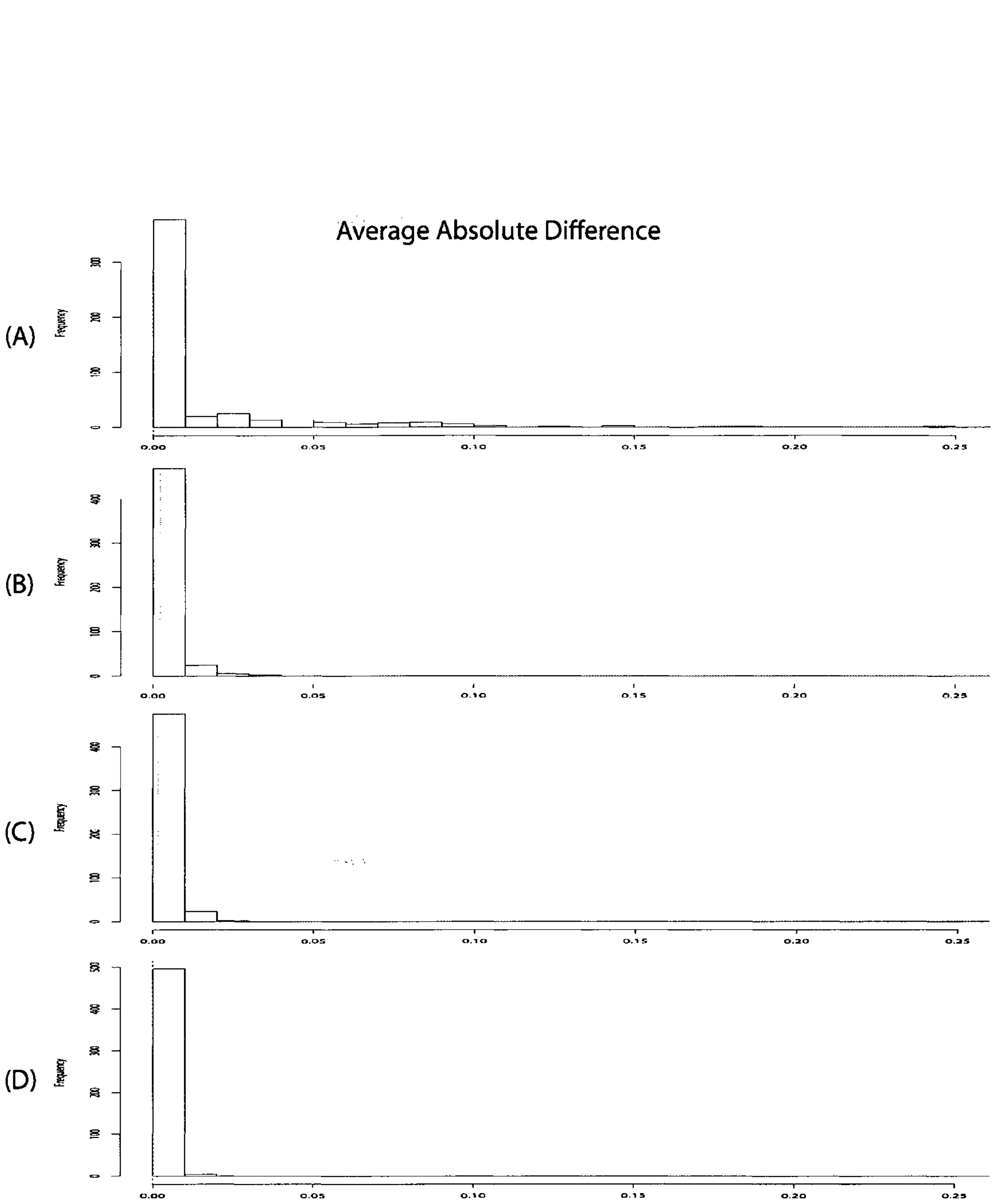


Figure 4.28: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 4 & 19 (theoretical kinship coefficient = 0.03125) for Chr 21, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution

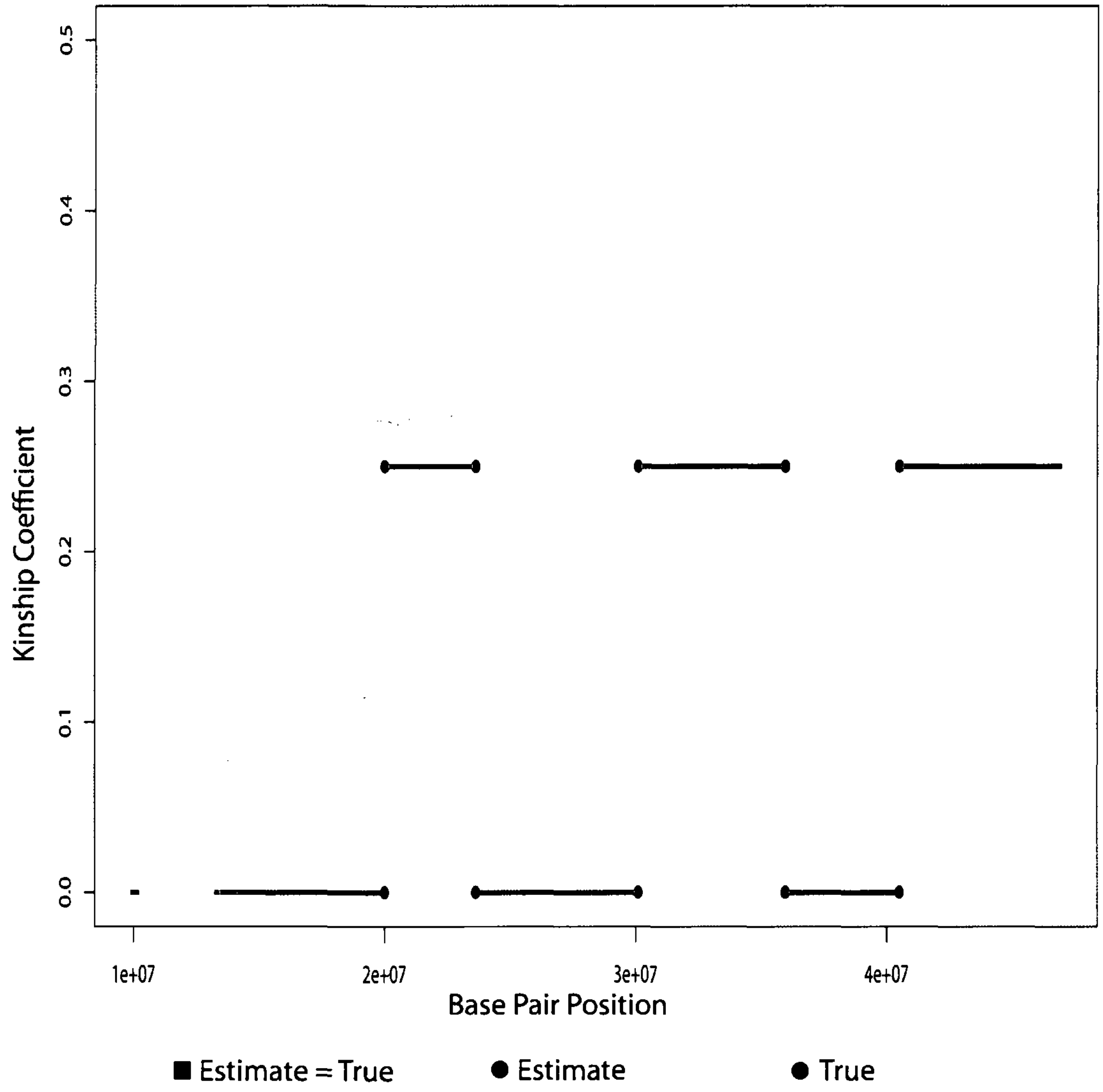


Figure 4.29: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 4 & 19 (theoretical kinship coefficient = 0.03125) 500K chip, Chr 21. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

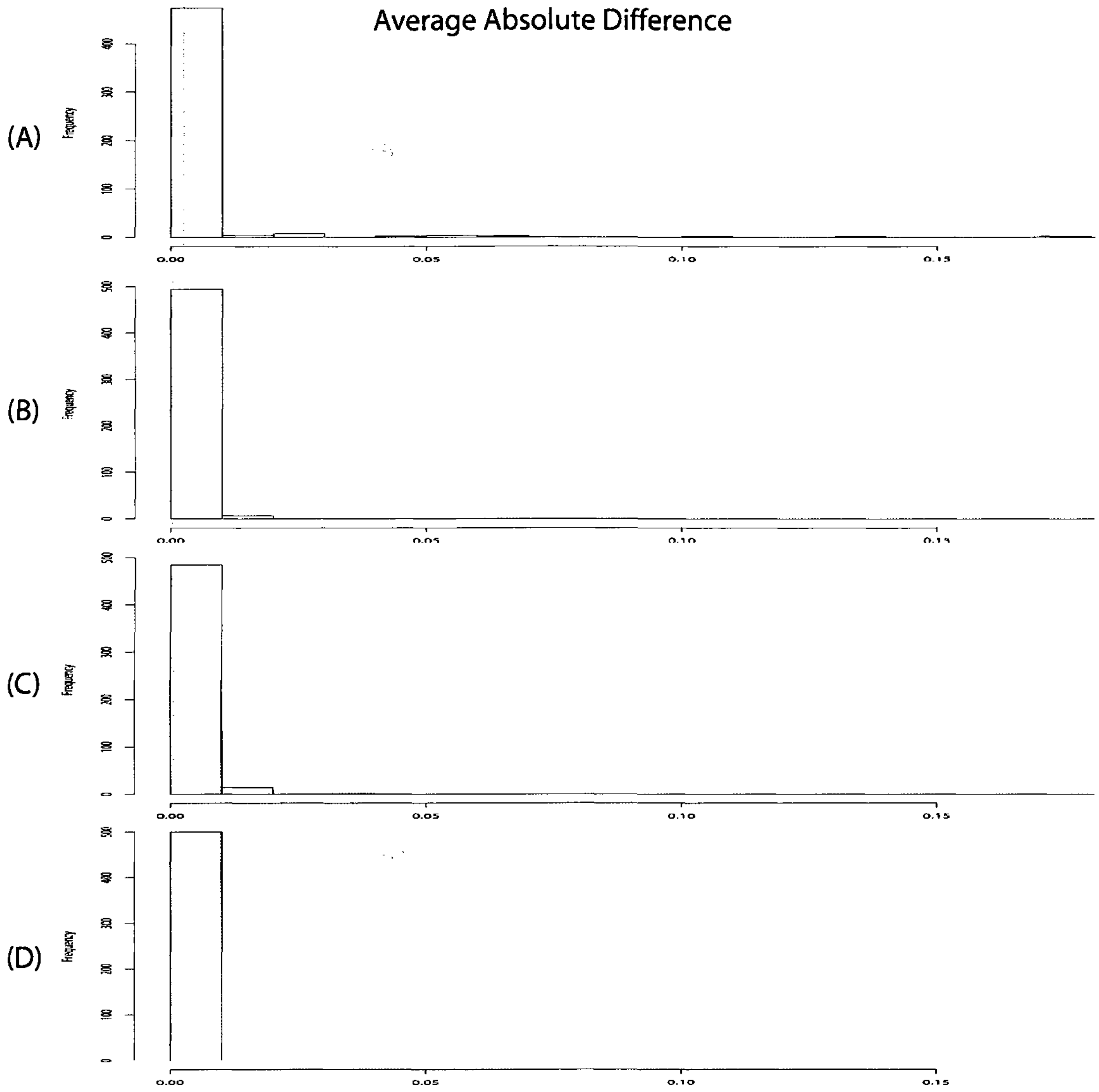


Figure 4.30: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 19 & 21 (theoretical kinship coefficient = 0.003906) for Chr 21, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution.

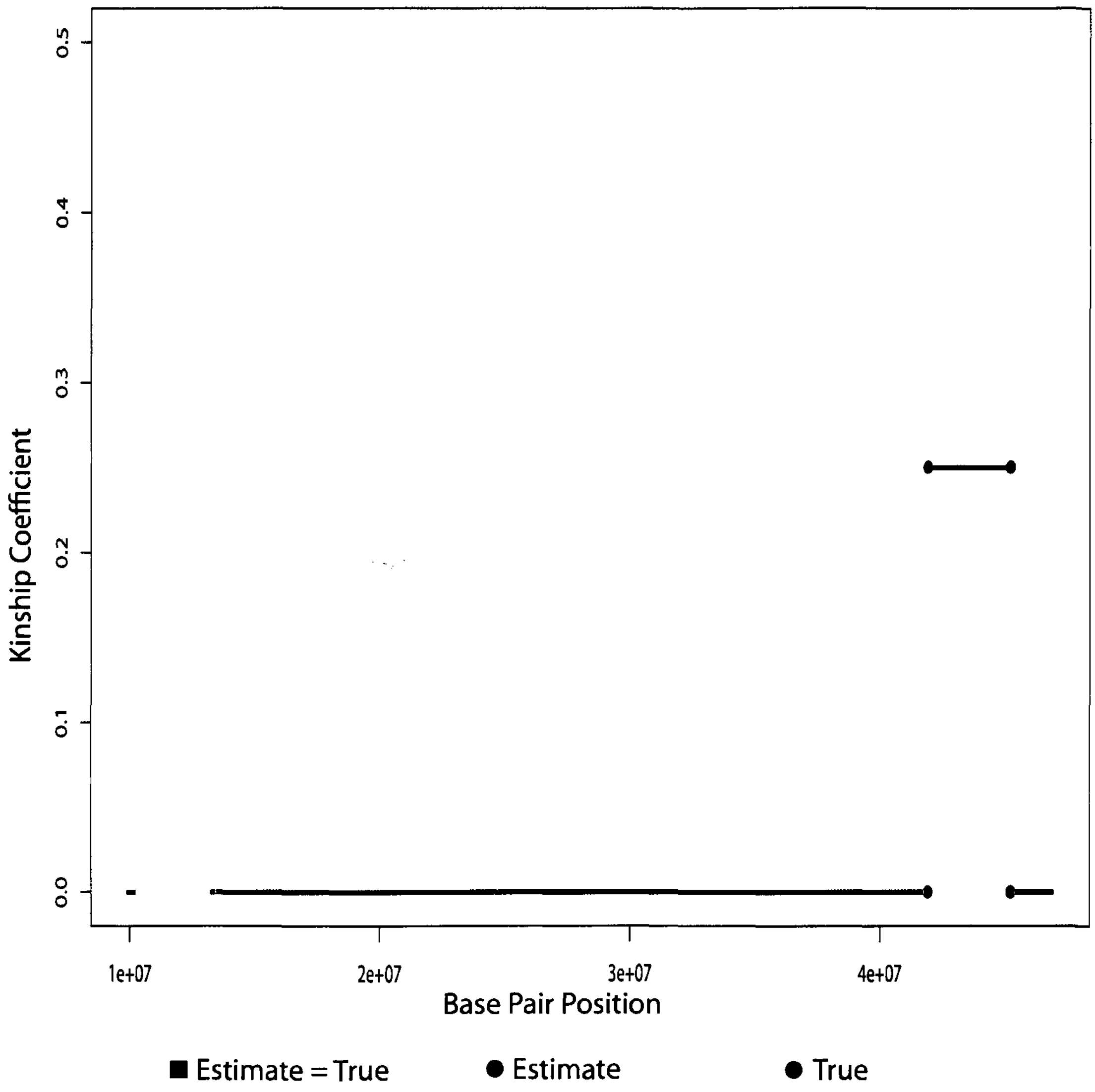


Figure 4.31: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 19 & 21 (theoretical kinship coefficient = 0.003906) 500K chip, Chr 21. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

We have demonstrated that our conditional kinship coefficient estimator using the generalized λ_1, λ_2 combination optimized on the 200K chip for chromosome 21 works well for estimating the conditional coefficients on chromosome 21 on the 100K, 200K and 500K chips. We now need to demonstrate that the generalized λ_1, λ_2 combination optimized for chromosome 21 is applicable genome wide. Additionally, chromosome 21 is one of the smallest chromosomes so we need to illustrate how our method performs on the large chromosomes. We answered both of these questions by using the generalized λ_1, λ_2 combination and analyzing chromosome 1 (one of the largest chromosomes) for the same individual pairs as above.

Figure 4.32 shows the distribution of *aad* for 500 replicates for individual pair 3 & 4 on chromosome 1 using the generalized lambdas for the 10K (4.32A), 100K (4.32B), 200K (4.32C) and 500K (4.32D) chips. The 500K chip again performs the best, and Figure 4.33 shows a conditional kinship coefficient plot for all 40,326 SNPs on chromosome 1 on the 500K chip for a replicate with approximately the mean *aad*. This figure shows the ability of our method to capture very complicated IBD patterns, and in fact we only misassigned 506 of the 40,326 SNPs to the incorrect IBD set. The performance of our method on chromosome 1 for individual pair 4 & 7 for the four chip types can be seen in Figure 4.34. The conditional kinship plot for a replicate with approximately the mean error for the 500K chip is in Figure 4.35, and shows that only 249 of the 40,326 SNPs were assigned to the incorrect IBD set. The distribution of *aad* for individual pair 4 & 19 for chromosome 1 for the four chips is found in Figure 4.36, which shows that the 500K chip is outperforming the other chips for this relationship as well. The conditional kinship coefficient plot for the replicate with approximately the mean error for the 500K chip is seen in Figure 4.37, and for this replicate the method only misassigned 127 of the 40,326 SNPs. Figure 4.38 shows the distribution of *aad* for individual pair 19 & 21 for chromosome 1 from the four chips and reveals that the 500K chip performs the best. Figure 4.39 is the coefficient plot for the replicate with approximately the mean *aad* and shows that only 36 of the 40,326 SNPs were misassigned.

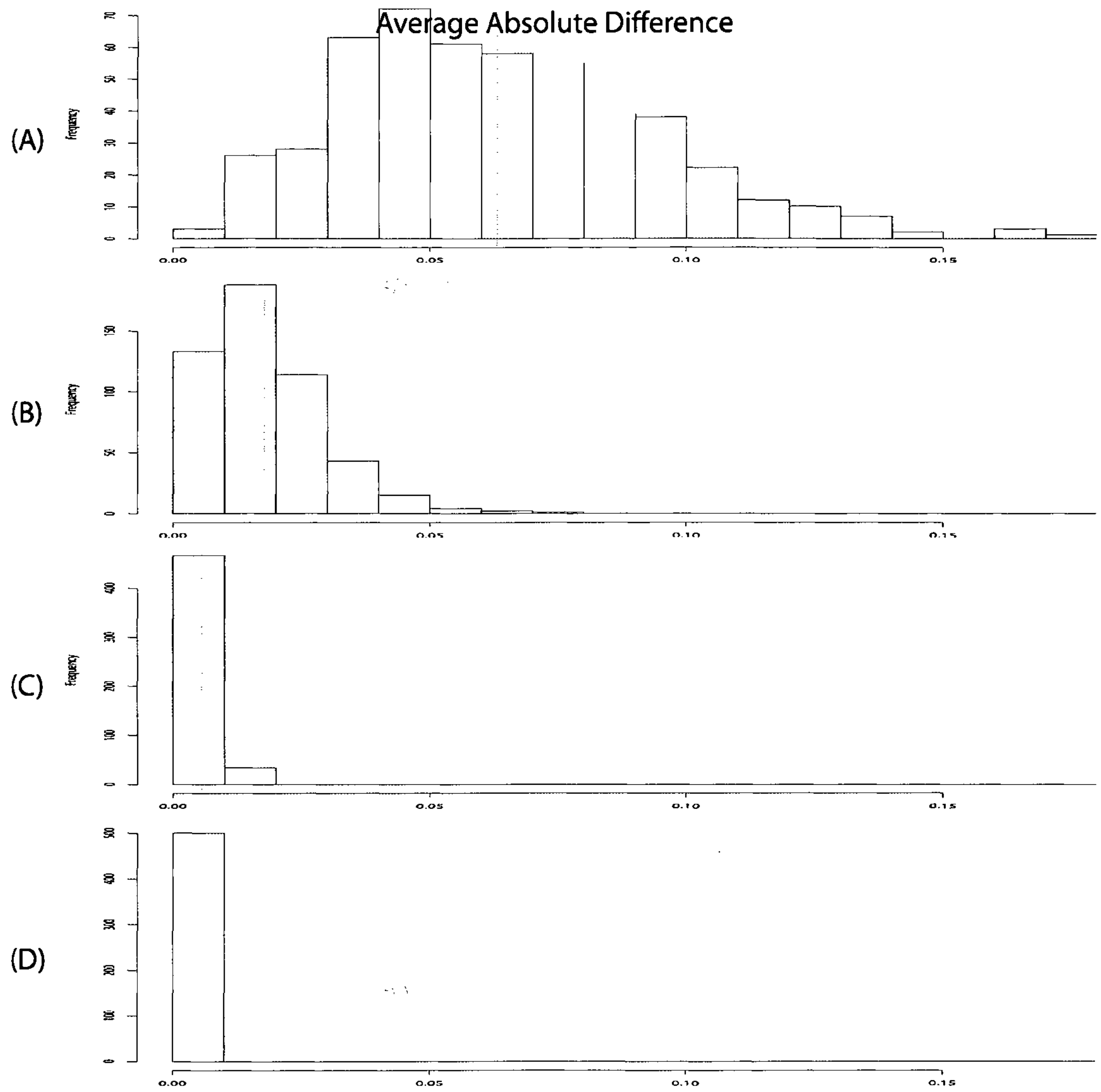


Figure 4.32: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) for Chr 1, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution.

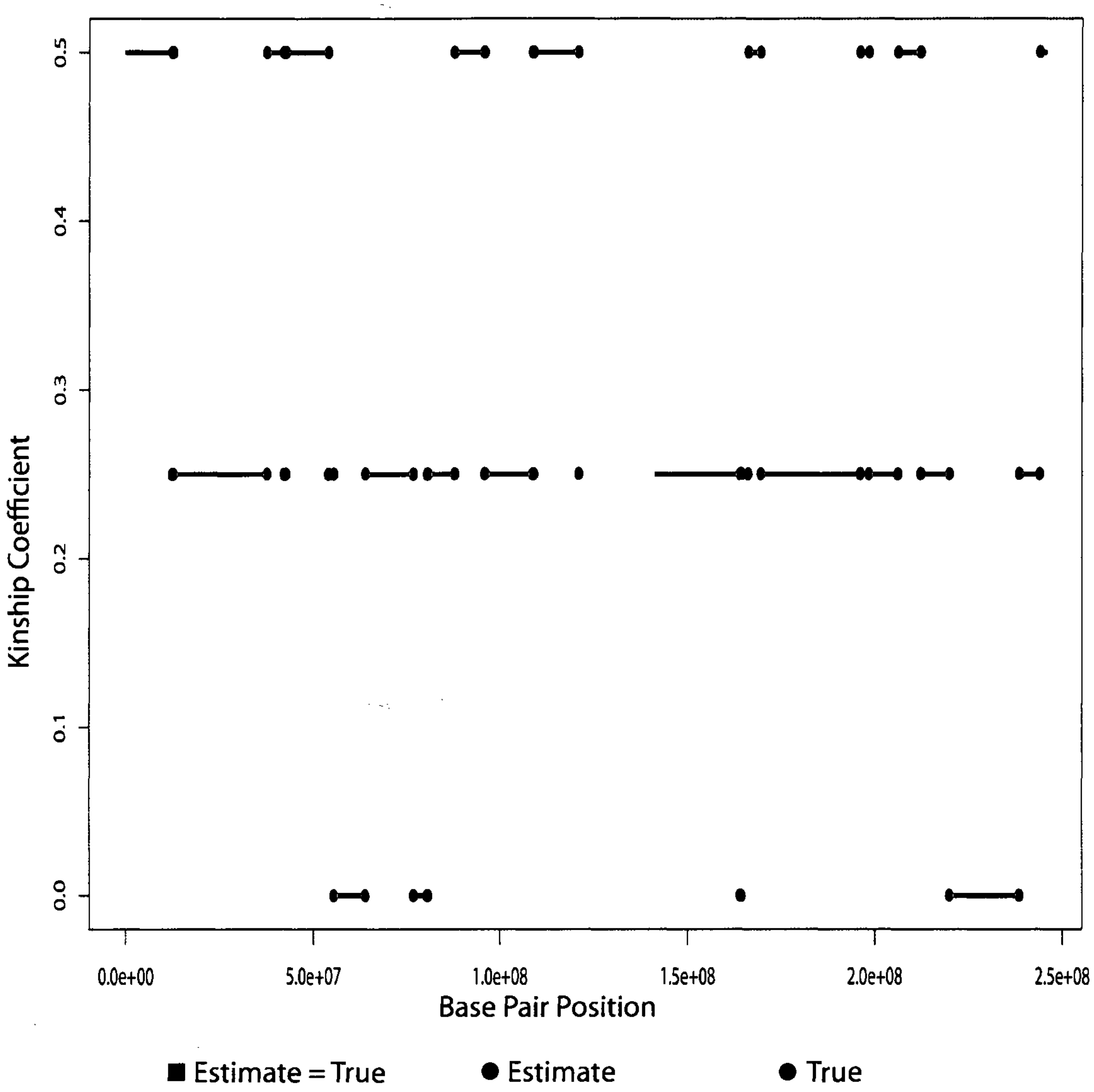


Figure 4.33: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 3 & 4 (theoretical kinship coefficient = 0.25) 500K chip, Chr 1. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

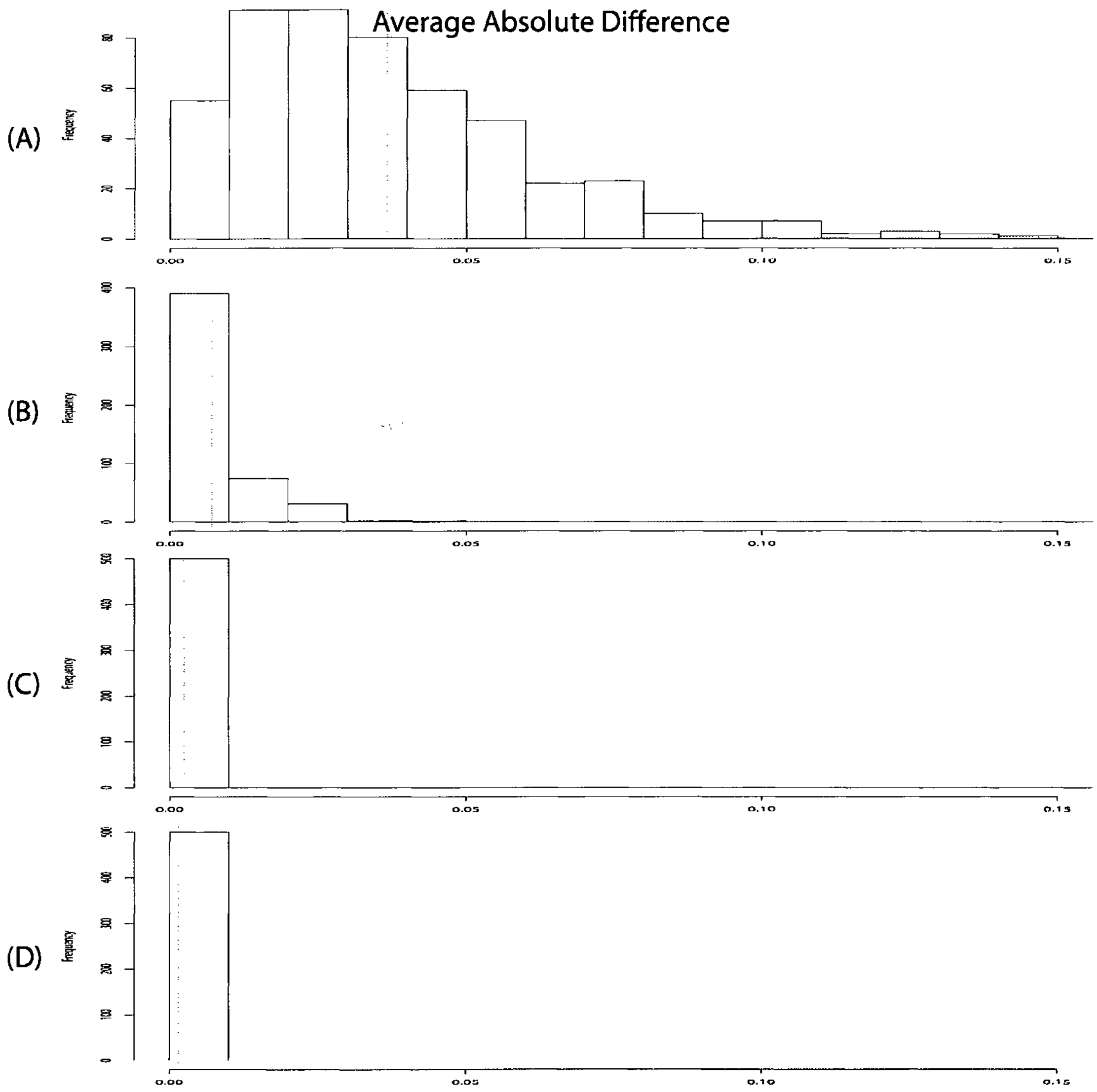


Figure 4.34: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) for Chr 1, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution.

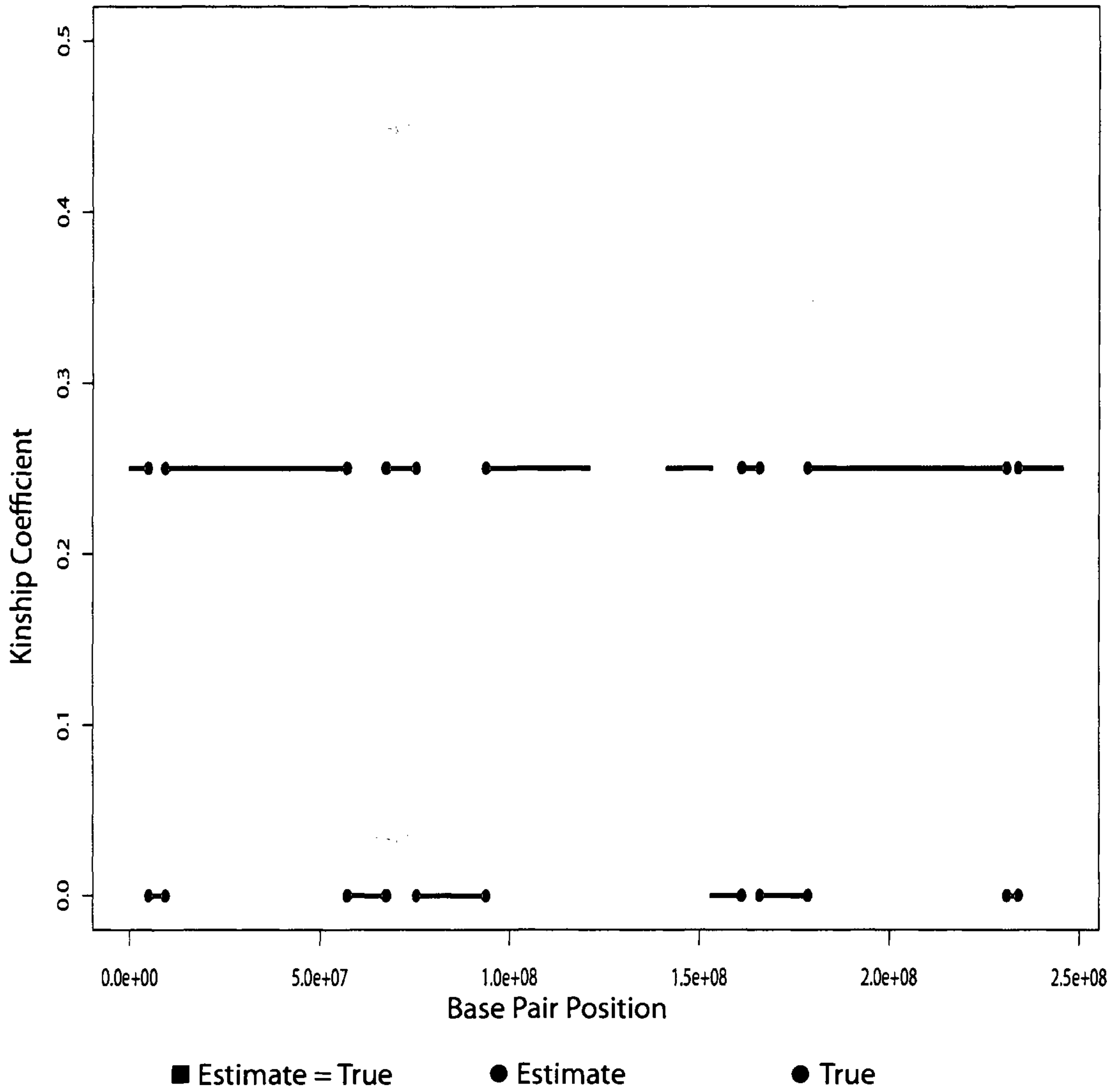


Figure 4.35: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 4 & 7 (theoretical kinship coefficient = 0.125) 500K chip, Chr 1. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

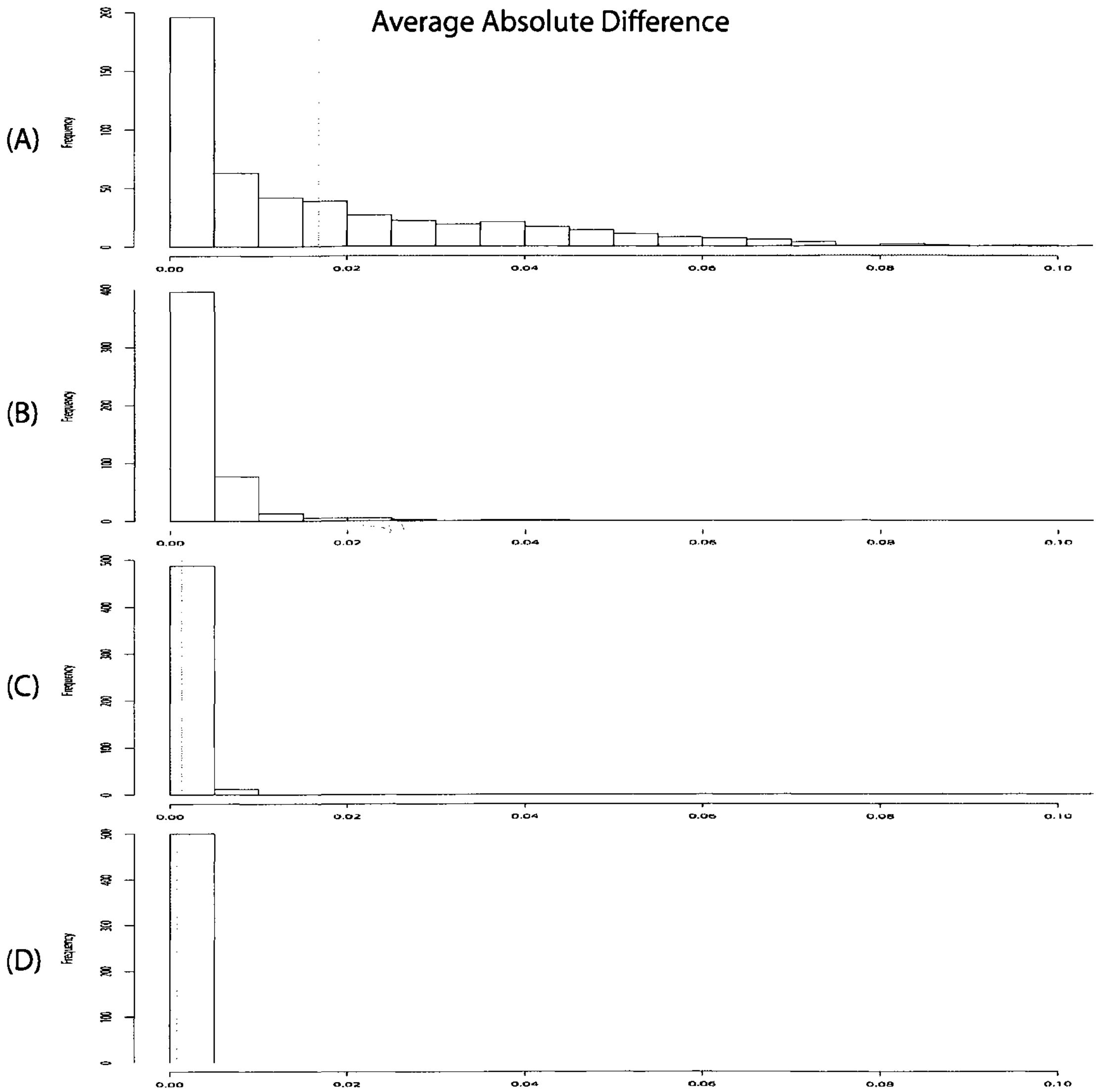


Figure 4.36: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 4 & 19 (theoretical kinship coefficient = 0.03125) for Chr 1, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution.

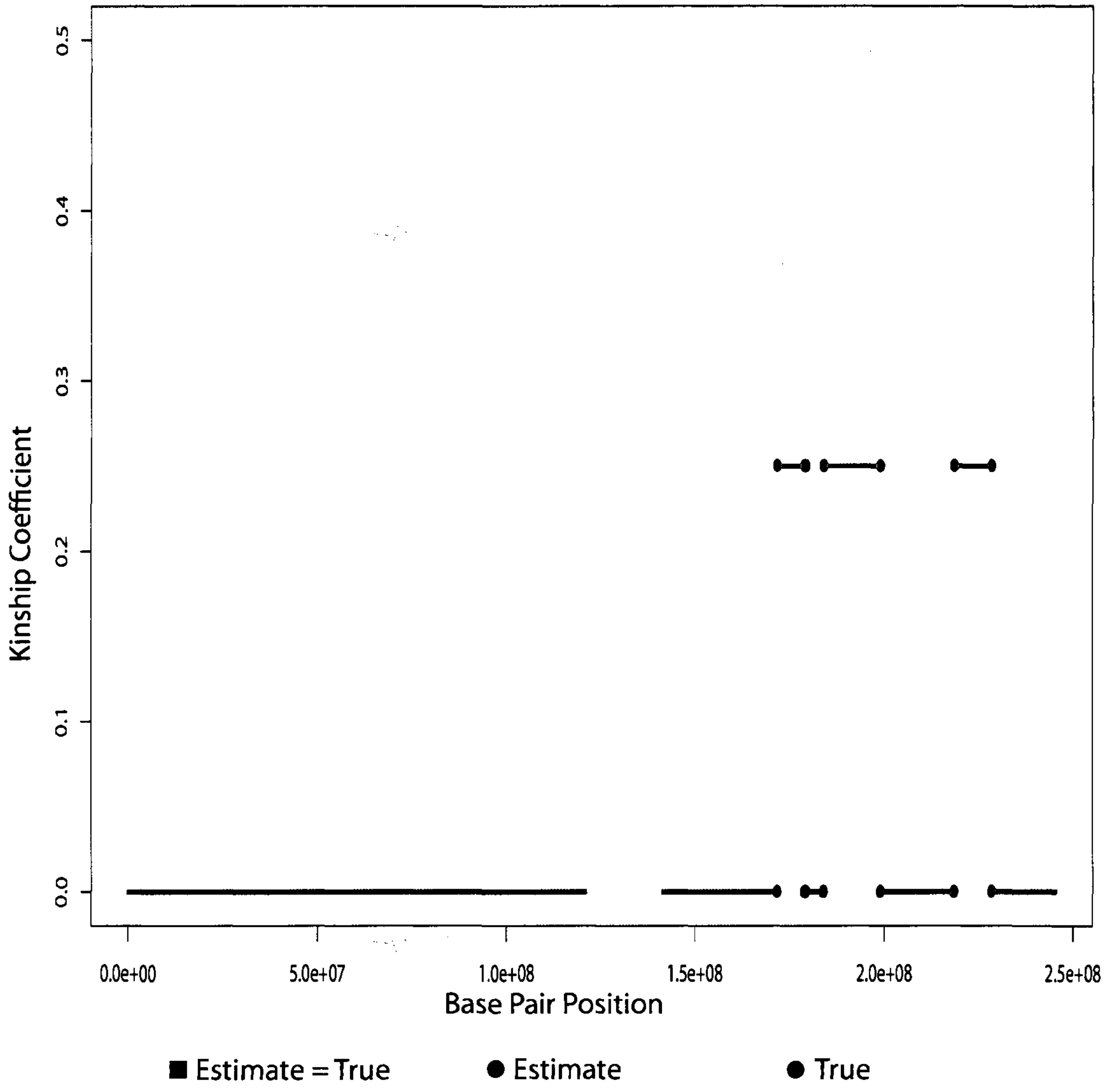


Figure 4.37: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 4 & 19 (theoretical kinship coefficient = 0.03125) 500K chip, Chr 1. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

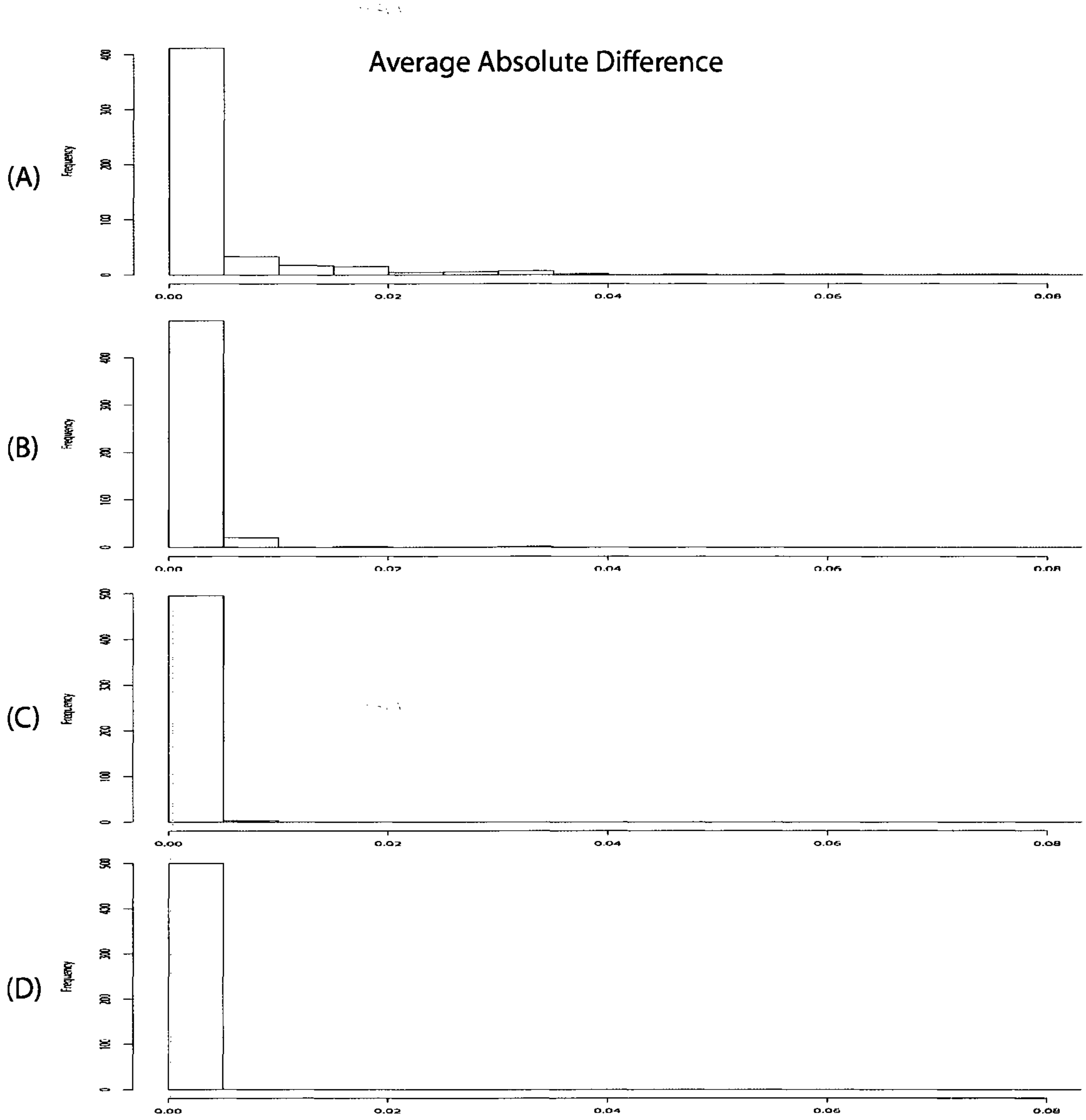


Figure 4.38: Distribution of Average Absolute Difference of the conditional kinship coefficient estimation for individual pair 19 & 21 (theoretical kinship coefficient = 0.003906) for Chr 1, 500 replicates: (A) 10K chip, (B) 100K chip, (C) 200K chip, (D) 500K chip. Red dotted line is at mean of distribution.

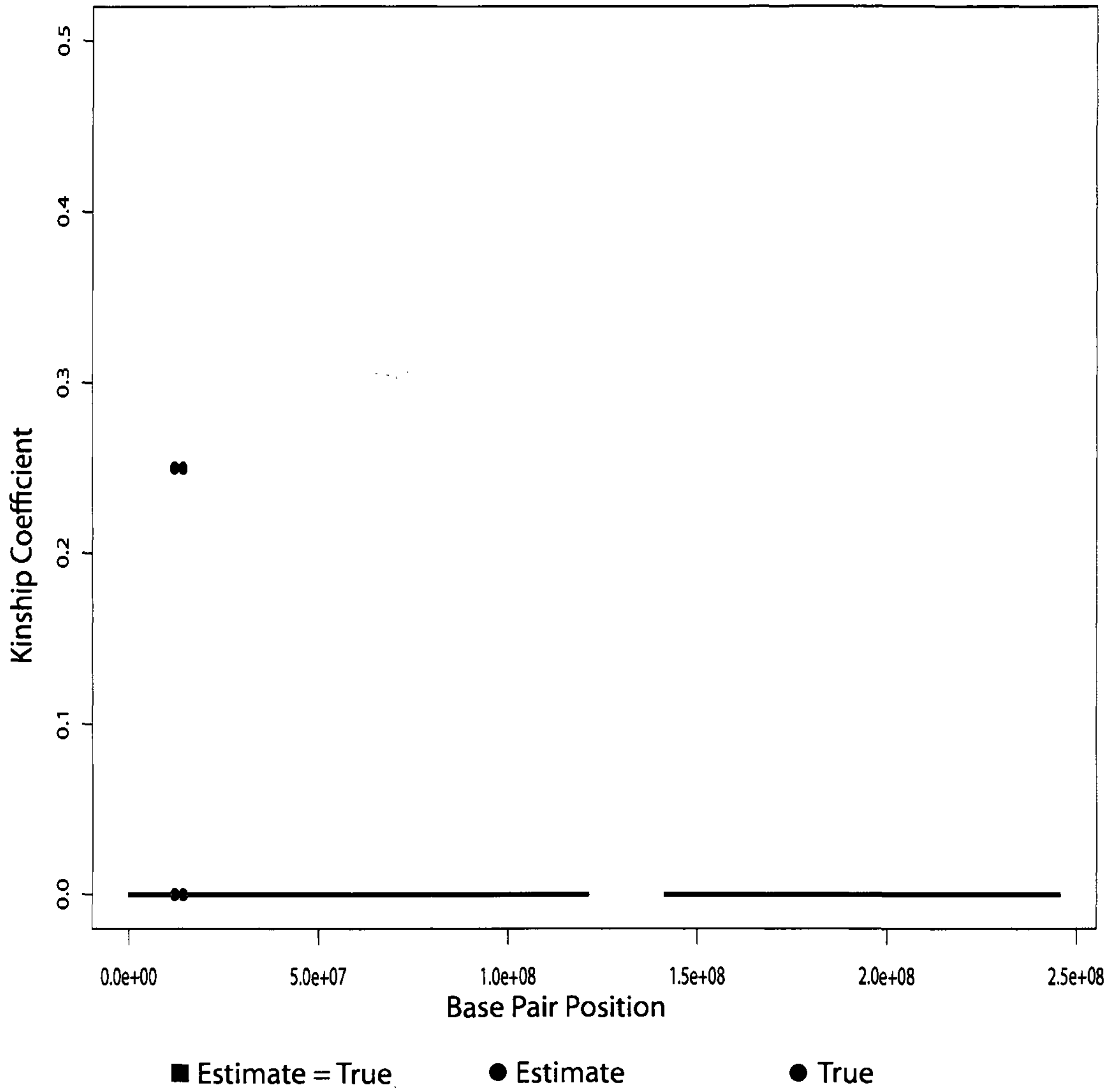


Figure 4.39: Comparison between True Conditional Kinship Coefficient and Estimated Conditional Kinship Coefficient for replicate with mean Average Absolute Difference for individual pair 19 & 21 (theoretical kinship coefficient = 0.003906) 500K chip, Chr 1. A locus has a black box when the estimated value and the true value are identical. For loci where the estimate is different than the true value, the estimated value is a blue dot and the true value is a red dot.

We have demonstrated that our penalized optimization method of estimating conditional kinship coefficients is a good general estimator and is able to discover the complex patterns of IBD sharing for both large and small chromosomes and across the entire spectrum of genetic relatedness that our theoretical kinship coefficient estimator can confidently estimate. The results from the theoretical kinship estimation and conditional kinship coefficient estimation leads us to belief that our estimates can be used in the traditional QTL mapping framework without lose of information or power to uncover linkage signals.

4.3.3 Pedigree Construction

We have demonstrated the ability to correctly estimate both the global and per locus genetic relatedness between individuals using only their genome wide SNP genotypes and no other information. Traditionally the only way to calculate these measures of relatedness was through the use of specified pedigrees. Therefore we hypothesized that we can perform inference in the opposite direction, given our estimates for the theoretical kinship coefficient we can infer the pedigrees contained within the data. Our method combines our theoretical kinship coefficient estimation method with the standard algorithm to find connected components of a graph to cluster individuals into pedigrees. These constructed pedigrees can then be analyzed in the same manner as traditionally collected and annotated pedigrees because we can estimate all the needed coefficients. This would allow the data collector to not spend the large amount of time and money to determine all pedigree relationships among the individuals genotyped.

We tested our hypothesis through simulation studies. We simulated genotypes via gene dropping for two replicates of the pedigree structure used in the theoretical and conditional kinship coefficient simulations (Figure 4.1) using the Illumina 550K SNP chip as the genotyping platform. We performed two different analyses: (1) everyone in both pedigrees had genotype data, and (2) individuals 7-12 from both pedigrees were excluded from the analysis. For these two setups we used different kinship coefficient cutoffs to see how the constructed pedigrees compared to the true pedigrees. We analyzed the two datasets with cutoff values of 0.25 (data not shown), 0.20, 0.125, and 0.10. These cutoffs were chosen to see the affects of using the mean value versus the -2SD value of the distributions for individual pairs 3 & 4 and 4 & 7. The results from using the 0.25 cutoff (mean of the distribution for pair 3 & 4) were not good. The expected number of pedigrees is 2 for both data set 1 and 2, but the analysis returned 23 and 24 pedigrees respectively. The next analysis was using a cutoff of 0.20 which is the -2SD value for individual pair 3 & 4. For data set 1 the correct configuration is two pedigrees each with 21 individuals, that is exactly what we found (Table 4.27). Next we analyzed data set 2 with the 0.20 cutoff. The expected number of pedigrees formed from this set up is 6 per input pedigree for a total of 12, and again we correctly recover all pedigrees. Next we used a cutoff of 0.125 which is the mean value of the distribution for individual pair 4 & 7. The expected number of pedigrees for both data set 1 and 2 is two pedigrees. The analysis of data set 1 correctly constructed the two pedigrees, but the analysis of data set 2 returned 7 pedigrees instead (Table 4.27). The last analysis used a cutoff of 0.10 which is the -2SD value of the distribution for individual pair 4 & 7. For both the analysis of data set 1 and 2 the expected number of pedigrees is 2, and we correctly reconstruct the pedigrees. We believe these results clearly demonstrate the ability of our new method to correctly cluster individuals into pedigrees. The

Cutoff	Data Set	Expected	Actual	Formed Pedigrees
0.20	1	2	2	[1-1 to 1-21],[2-1 to 2-21]
0.20	2	12	12	[1-1,1-2,1-3,1-4],[1-5],[1-6], [1-13,1-16,1-19],[1-14,1-17,1-20], [1-15,1-18,1-21],[2-1,2-2,2-3,2-4] [2-5],[2-6],[2-13,2-16,2-19] [2-14,2-17,2-20],[2-15,2-18,2-21]
0.125	1	2	2	[1-1 to 1-21],[2-1 to 2-21]
0.125	2	2	7	[1-1,1-2,1-3,1-4,1-14,1-15, 1-17,1-18,1-20,1-21],[1-6], [1-5,1-13,1-16,1-19],[2-1,2-2,2-3,2-4], [2-5,2-13,2-14,2-16,2-17,2-19,2-20], [2-15,2-18,2-21],[2-6]
0.10	1	2	2	[1-1 to 1-21], [2-1 to 2-21]
0.10	2	2	2	[1-1 to 1-21], [2-1 to 2-21]

Table 4.27: Pedigree Construction Results from Simulated Pedigrees. Data Set 1 refers to data set with everyone included, and Data Set 2 refers to the data set with individuals 7-12 from both pedigrees excluded. Individuals within [] were clustered into a pedigree. The 1- means the individual was from the first simulated pedigree and 2- means the individual was from the second simulated pedigree.

next test is to see how this and the methods tested above perform in real data.

4.3.4 Analysis of Southwest Foundation Data

We have demonstrated that our methods work well on simulated data and now will test the performance on the real data set from the Southwest Foundation described in the Materials and Methods section. The analysis undertaken was to use our pedigree construction algorithm, estimate the theoretical and conditional kinship coefficients in those pedigrees, and then perform QTL mapping using our data and compare the results to the known QTL in the data set. Based upon the simulation results we constructed pedigrees with three different cutoffs: 0.20, 0.10 and 0.0625. Using the cutoff of 0.20 we constructed 122 pedigrees of 2 or more individuals and included 740 of 858 individuals with genotype data, the remaining genotyped individuals forming singleton pedigrees. With the cutoff of 0.10 we constructed 31 pedigrees of 2 or more individuals and included 793 of the 858 typed individuals, and this included a pedigree with 568 individuals. Using a cutoff of 0.0625 we constructed a single pedigree that included all 858 individuals. The remainder of the discussion of our analysis will focus on the analysis using a cutoff of 0.20.

The first analysis we performed was to investigate how our theoretical kinship coefficient estimator performed in this real data set. We found all pairs of individuals with theoretical kinship coefficients of 0.25, 0.125, 0.0625 and 0.03125 and determined the distribution of our estimated coefficients for these pairs. There were 1,218 individual pairs with theoretical kinship coefficients of 0.25 and the distribution of our estimates for those pairs are seen in Figure 4.40A. The mean of the distribution was 0.2533 with a standard deviation of 0.03. There were 1,521 individual pairs with theoretical kinship coefficient of 0.125 and the distribution of our estimates had a mean of 0.1291 and standard deviation of 0.021 and is seen in Figure 4.40B. The analysis of pairs with theoretical kinship coefficient of 0.0625 was based on 1,950 pairs and the distribution of estimates had mean 0.0667 and standard deviation of 0.021 and is in Figure 4.40C. The final theoretical kinship coefficient pairs we looked at have a coefficient of 0.03125 and our analysis of the 1,454 pairs had a mean of 0.0349 with a standard deviation of 0.017 as seen in Figure 4.40D. All these analyses show that our method is performing well in real data and we believe revealing relationship misspecifications in the outliers of the distributions.

With these results showing that our theoretical kinship estimator is working well, we ran the pedigree construction algorithm on the data with a cutoff of 0.20. As stated above, this resulted in the formation of 122 pedigrees with 2 or more individuals for a total of 740 individuals being included in these pedigrees. Just as a reminder, the data as specified by the Southwest Foundation had 1942 individuals, with 107 singletons and the other 1,835 individuals in 46 pedigrees of 2 or more individuals. Of the 1942 individuals only 858 individuals had genotype information. So with a cutoff that is very good at recognizing sibling/parent-offspring relationships we were able to assign 86.25% of the individuals with genotypes to pedigrees. Figure 4.41 is an example of how our pedigree construction compares to the specified pedigrees. This figure shows the structures of two pedigrees as specified by the Southwest Foundation. Individuals 1a and 2a were clustered with Pedigree B because individual 1a has estimated kinship coefficients of 0.329, 0.201, 0.215 and 0.326 with individuals 1b, 2b, 3b and 4b respectively. Whereas individual 1a had estimated coefficients of 0.193, 0.190 and 0.175 with individuals 3a, 4a and 5a respectively. We believe this example shows that our method performs well because individuals 1b, 2b and 4b would not have been included in their specified pedigree if it was not for including individual 1a in the analysis.

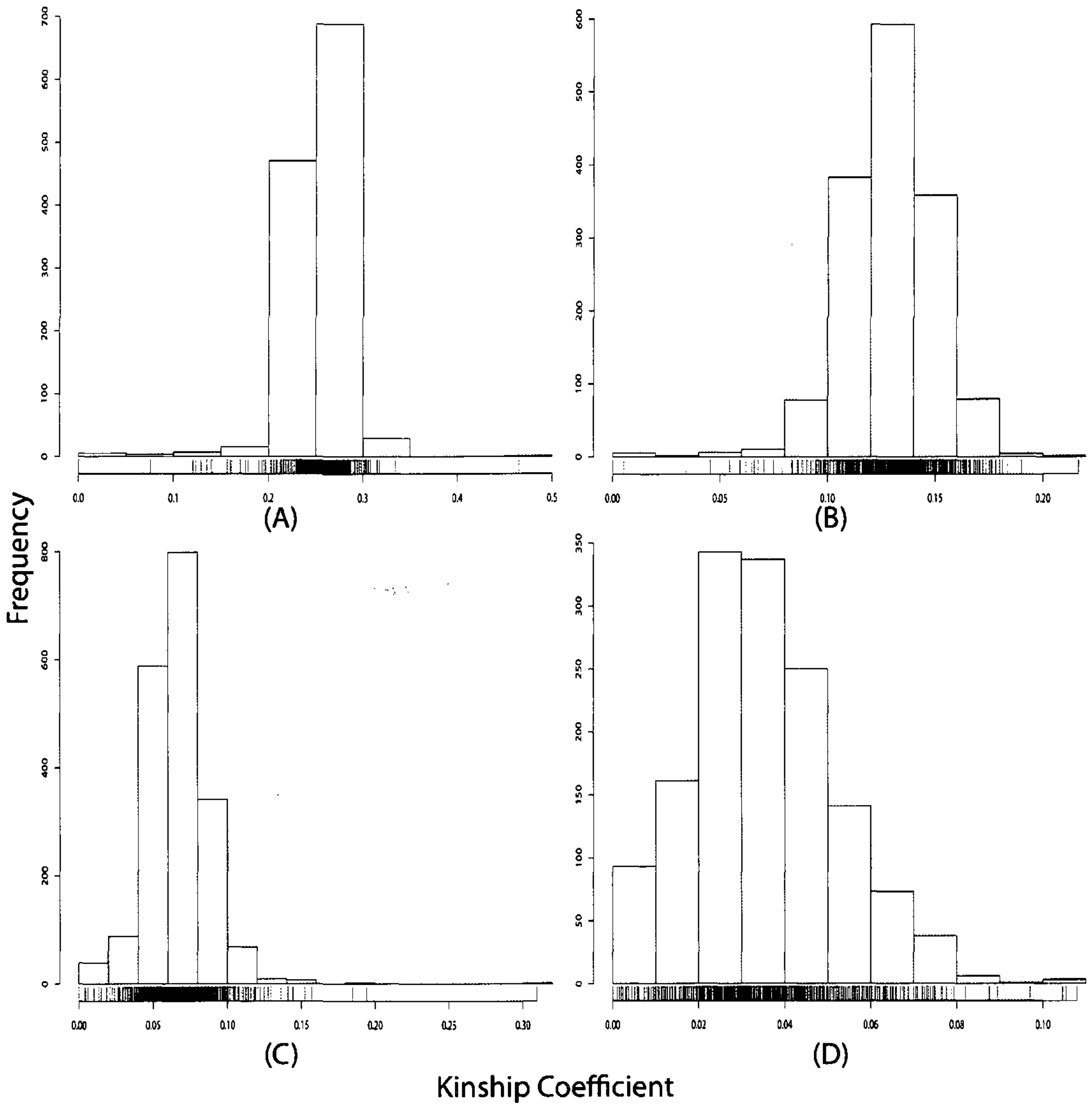


Figure 4.40: Theoretical Kinship Coefficient Estimation for Southwest Foundation data: (A) Pairs designated with 0.25 (B) Pairs with 0.125 (C) Pairs with 0.0625 (D) Pairs with 0.03125 Theoretical Kinship Coefficient based upon specified pedigree structures

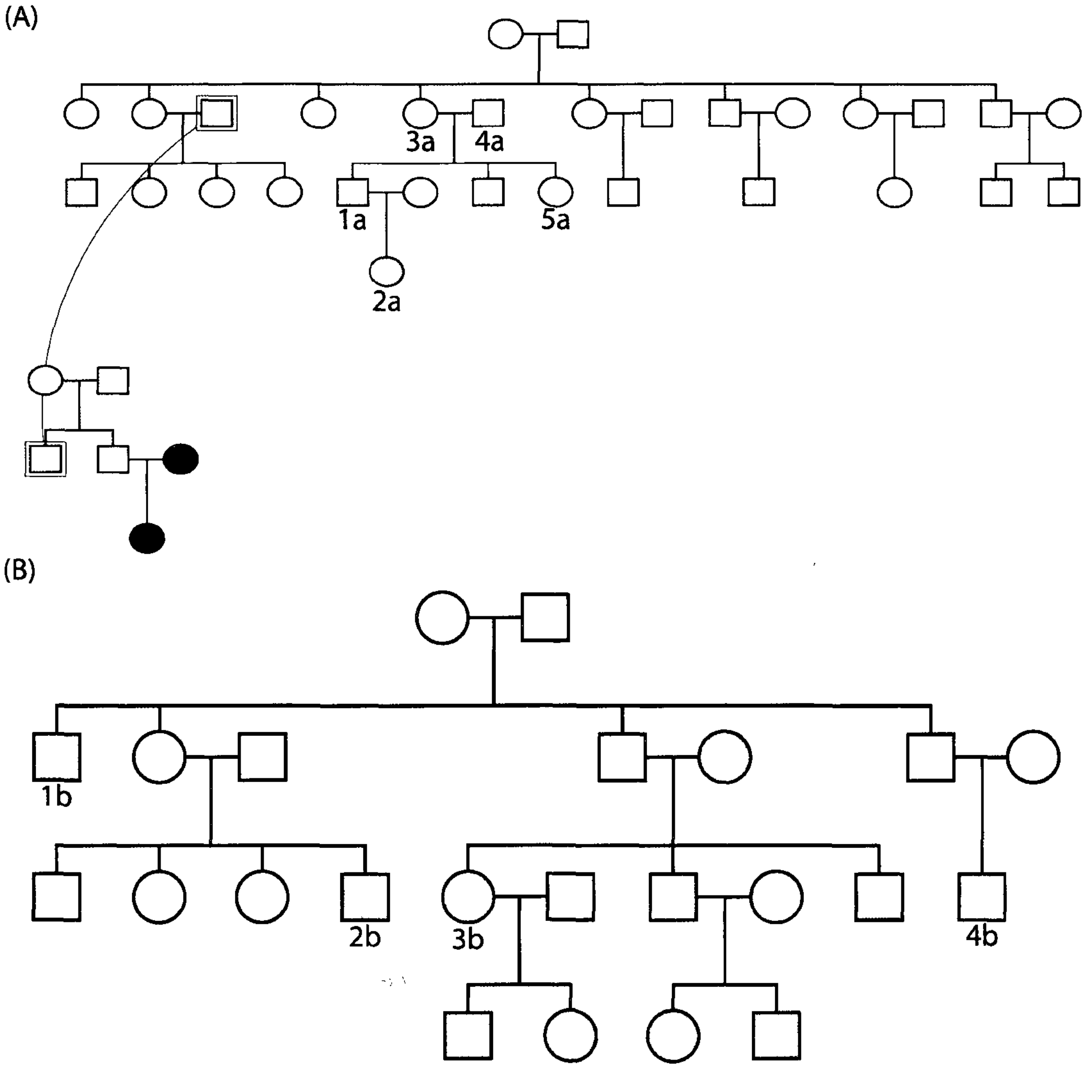


Figure 4.41: Example of our pedigree construction using 0.20 cutoff versus specified pedigree structure for Southwest Foundation data. Only colored individuals had genotype data and were analyzed. (A) & (B) shows pedigree structures supplied by Southwest Foundation. Individuals of the same color were assigned to the same pedigree by our algorithm.

With confidence in our pedigree construction and theoretical kinship estimation methods we are currently performing QTL analysis in our constructed pedigrees with our estimated coefficients.

4.4 Discussion

We have proposed three new methods utilizing genome-wide SNP genotypes to estimate the level of pairwise genetic relatedness both globally and locally between individuals with no prior knowledge of their relationship. The first method we developed was a method-of-moments technique to utilize SNP genotypes to estimate the theoretical kinship coefficient between pairs of individuals. We have shown that our method is an unbiased and good estimator of the theoretical kinship coefficient for very distantly related individuals. Our results also illustrate that our method increases its accuracy and reduces its variance with increasing numbers of markers included in the analysis. This method can be used to discover cryptic relationships in study samples or test specified relationships for accuracy.

Building upon our theoretical kinship coefficient estimator we developed a penalized optimization algorithm to estimate the conditional kinship coefficient for every SNP on a genome-wide basis for pairs of individuals. We found a single penalty value that works for all levels of relatedness and works well on both large and small chromosomes. The results show that our method is capable of uncovering complex patterns of IBD sharing between individuals, and does so with a small degree of error. The results from the simulation studies give us confidence that the estimates that our method produces can be used in conjunction with our theoretical kinship coefficient estimates in QTL analysis using variance components with little or no lose of power.

Finally we showed that we have created a self-contained analysis paradigm: by combining our theoretical kinship estimator with an algorithm to find connected components of a graph we construct pedigrees from genome-wide SNP data; we can also estimate the conditional kinship among pairs within the pedigrees unites. These algorithms together supply all the material (pedigree structures, theoretical, and conditional kinship coefficients) necessary for standard gene mapping algorithms. The results of our simulation investigations reveal that a soft threshold cutoff should be applied to pedigree construction, and when it is applied the algorithm performs well at correctly identifying individuals that belong together. In addition, our analysis of the Southwest Foundation data illustrates the ability of our procedures to discover cryptic relationships in data. Taken together we believe that these three new analysis techniques open up new possibilities for study design. The methods allow researchers to use the power of pedigree analysis without the cost and time of correctly collecting and annotating pedigrees.

Chapter 5

Outlook

In the span of my graduate career I've gone from analyzing a 9 marker microsatellite map on chromosome 16 in approximately 750 individuals, to analyzing the Southwest Foundation data set in Chapter 4 that was composed of 858 individuals that had genotype data at 547,458 autosomal SNPs from across the entire genome. That is an astonishing expansion in scale in a short period of time and gives an example of the tremendous technological advancements that are driving the field of human genetics. We hope this dissertation has given the reader some of the challenges posed by this huge expansion in the scale of data generation as well as ways to overcome those challenges and use the data in novel ways to help map the genetic components involved in the complex human diseases. The research in this dissertation has revealed further avenues of investigation that can build on its success.

The GGSD data management application discussed in Chapter 2 has many areas for further development and enhancement. An important area of development is the addition of the capability to store and search analysis results. This is a large undertaking considering the vast number of different analysis techniques, especially since that list is continually growing as new methods are developed to analyze these large datasets. But such a capability is necessary and if the results are searchable an extremely powerful tool. Additionally, the analysis of copy number variation (CNV) is becoming a very interesting and widely-used analysis. The database schema needs to be able to handle this type of data and potentially develop new tools to search and analyze this important type of variation. Other small enhancements include supporting the production of the binary formatted files for the PLINK software (Purcell *et al.*, 2007). As the human genetics research community evolves the types of data generated and analyses performed, GGSD will grow and evolve with it.

The algorithms for the estimation of identity-by-descent (IBD) sharing between general individuals developed in Chapter 4 has many further applications. Many genetic analysis methods are based upon utilizing information on IBD sharing, including the NPL analysis methods discussed in Chapter 3. The estimation techniques employed by the program Simwalk2 to estimate IBD sharing can take a long time to converge. An investigation into whether the algorithms developed in Chapter 4 are good enough and allow for a substantial savings in computational efficiency to replace Simwalk2's method of IBD estimation should be undertaken. Further development of the algorithms is possible as well. All three algorithms of Chapter 4 lend themselves to straight forward parallelization and therefore a tremendous increase in its computational efficiency. For example, the estimation of the conditional kinship coefficients along each chromosome is independent of all other chromosomes allowing the per chromosome calculation to be split among available processors. These are but two of the possible extensions to these powerful algorithms.

With the continued development of new ways to assay genetic variation and the cost of sequencing plummeting due to new massively parallel technologies, there will be no shortage of data or new possibilities for the development of analysis methods any time soon. This makes the computational genetics field an exciting field with a strong future.

Bibliography

Abecasis, G. *et al.* (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, **30**, 97–101.

Altshuler, D. and Daly, M. (2007). Guilt beyond a reasonable doubt. *Nature Genet*, **39**(7), 813–815.

Bacanu, S. *et al.* (2000). The power of genomic control. *Am J Hum Genet*, **66**, 1933–1944.

Barrett, J. *et al.* (2005). Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, **21**(2), 263–5.

Boehnke, M. and Cox, N. (1997). Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet*, **61**, 423–429.

Cheung, K. *et al.* (1996). Phenodb: an integrated client/server database for linkage and population genetics. *Comput Biomed Res*, **29**, 327–337.

Daly, M. *et al.* (2001). High-resolution haplotype structure in the human genome. *Nat Genet*, **29**(2), 229–32.

Day, A. G. (2007). The generic genetic studies database: A data management system for large scale genetic studies. *Presented at the annual meeting of The American Society of Human Genetics, October 23-27*. Available from http://www.ashg.org/cgi-bin/ashg07s/ashg07.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometics*, **55**, 997–1004.

Ehm, M. and Wagner, M. (1998). A test statistic to detect error in sib-pair relationships. *Am J Hum Genet*, **62**, 181–188.

Epstein, M. *et al.* (2000). Improved inference of relationship for pairs of individuals. *Am J Hum Genet*, **67**, 1219–1231.

Fiddy, A. *et al.* (2005). An integrated system for genetic analysis. *BMC Bioinformatics*, **7**, 210.

Gillanders, E. *et al.* (2004). Genelink: a database to facilitate genetic studies of complex traits. *BMC Genomics*, **5**, 81.

Kong, A. and Cox, N. (1997). Allele-sharing models: Lod scores and accurate linkage tests. *Am J Hum Genet*, **61**, 1179–1188.

Kruglyak, L. *et al.* (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, **58**, 1347–1363.

Lange, E. and Lange, K. (2004). Powerful allele sharing statistics for nonparametric linkage analysis. *Hum Hered*, **57**, 49–58.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York, 2nd edition.

Lange, K. *et al.* (2001). Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet*, **69(Suppl)**, A1886.

Lewis, S. *et al.* (2002). Apollo: a sequence annotation editor. *Genome Biol*, **3**, RESEARCH0082.

Li, J.-L. *et al.* (2001). Toward high-throughput genotyping: dynamic and automatic software for manipulated large-scale genotype data using fluorescently labeled dinucleotide markers. *Genome Res*, **11**, 1304–1314.

Li, J.-L. *et al.* (2005). Phd: a web database application for phenotype data management. *Bioinformatics*, **21**(16), 3443–3444.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, **152**, 1753–1766.

Makinen, V. *et al.* (2005). High-throughput pedigree drawing. *Eur J Hum Genet*, **13**, 987–989.

Matsuzaki, H. *et al.* (2004). Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nat Methods*, **1**, 109–111.

McCarthy, M. I. *et al.* (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, **9**(5), 356–369.

McPeek, M. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet*, **66**, 1076–1094.

Millar, K. (1987). *Some Eclectic Matrix Theory*. Robert E Krieger Publishing, Malabar, FL.

Milligan, B. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Mousseau, T. *et al.* (1998). A novel method for estimating heritability using molecular markers. *Heredity*, **80**, 218–224.

Mukhopadhyay, N. *et al.* (2005). Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics*, **21**, 2556–7.

Pajukanta, P. *et al.* (2003). Combined analysis of genome scans of dutch and finnish families reveals a susceptibility locus for high-density lipoprotein cholesterol on chromosome 16q. *Am J Hum Genet*, **72**, 903–917.

Patil, N. *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–23.

Pritchard, J. and Rosenberg, N. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, **65**, 220–228.

Purcell, S. *et al.* (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, **81**(3), 559–75.

Queller, D. and Goodnight, K. (1989). Estimating relatedness using genetic markers. *Evolution*, **43**, 258–275.

Reich, D. and Goldstein, D. (2001). Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol*, **20**, 4–16.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Satten, G. *et al.* (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet*, **68**, 466–477.

Seuchter, S. and Skolnich, M. (1988). Hgdbms: A human genetics database management system. *Comput Biomed Res*, **21**, 478–487.

Slager, S. and Schaid, D. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *Am J Hum Genet*, **68**, 1457–1462.

Sobel, E. *et al.* (2001). Multipoint estimation of indentity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered*, **52**, 121–131.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet*, **58**, 1323–1337.

Song, K. *et al.* (2004). Efficient simulation of p values for linkage analysis. *Genet Epidemiol*, **24**, 1–9.

Stein, L. *et al.* (2002). The generic genome browser: a building block for a generic model organism database. *Genome Res*, **12**, 1599–610.

Sun, L. *et al.* (2002). Enhanced pedigree error detection. *Hum Hered*, **54**, 99–110.

The International HapMap Consortium *et al.* (2003). The international hapmap project. *Nature*, **426**, 789–96.

The International Human Genome Sequencing Consortium *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Thompson, E. (1974). Gene identities and multiple relationships. *Biometrics*, **30**, 667–680.

Thompson, E. (1975). The estimation of pairwise relationships. *Ann Hum Genet*, **39**, 173–188.

Venter, J. *et al.* (2001). The sequence of the human genome. *Science*, **291**, 1304–51.

Voight, B. and Pritchard, J. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, **1**, e32.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, **160**, 1203–1215.

Whittemore, A. and Halpern, J. (1994a). A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

Whittemore, A. and Halpern, J. (1994b). Probability of gene identity by descent: computation and applications. *Biometrics*, **50**, 109–117.

Wigginton, J. and Abecasis, G. (2005). Pedstats: descriptive statistics, graphics, and quality assessment for gene mapping data. *Bioinformatics*, **21**, 3445–3447.

Wigginton, J. and Abecasis, G. (2006). Evaluation of the replicate pool method: quick estimation of genome-wide linkage peak p-values. *Genet Epidemiol*, **30**, 320–332.

Wigginton, J. *et al.* (2005). A note on exact tests of hardy-weinberg equilibrium. *Am J Hum Genet*, **76**, 887–893.

Yu, J. *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, **38**, 203–8.

Zhao, L.-J. *et al.* (2005). Snpp: automating large-scle snp genotype data management. *Bioinformatics*, **21**(2), 266–268.