Effects of Common Errors In Microsatellite Data On Estimates of Population Differentiation

AND

Inferring Genotypic Structure Of Complex Disease Loci Using Genome-Wide Expression Data

by

Ellen Hepfer Breazel

(Under the Direction of Paul Schliekelman)

Abstract

In this dissertation, two different areas of statistical genetics are explored. The first is a analysis of genotyping errors in microsatellite data and their effect on population differentiation statistics. Although, genotyping errors in microsatellite data have been explored for their effects on parentage assessment, especially with exclusion and on population size estimates of mark and recapture studies. This research is the result of need to understand the effects of inevitable errors within microsatellite data on conclusions about population differentiation. Chapter 2 illustrates the statistically significant effects that three common genotyping errors (allelic dropout, binning error, and null alleles) have on the population differentiation statistic F_{ST}. These errors however, produce no change in the overall conclusions about the differences between populations.

The second is a method for improving gene mapping of complex diseases. Chapter 3 describes a process using genetical genomics methods to cluster expression level genes by their

causative locus and then inferring the genotype structure of these causative loci for each individual. General association studies of a particular locus have reduced power due to individuals present whose disease is not influenced by that locus. Our inferred genotype structure is used to eliminate individuals where this is the case to increase the power of gene mapping.

INDEX WORDS: dropout, F_{ST}, microsatellites, mutations, null alleles, population subdivision, eQTLs, genetical genomics, complex traits,

Effects of Common Errors In Microsatellite Data On Estimates of Population Differentiation

AND

Inferring Genotypic Structure Of Complex Disease Loci Using Genome-Wide Expression Data

by

Ellen Hepfer Breazel

B.S., Clemson University, 2000

M.S., The University of Georgia, 2003

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor Of Philosophy

Athens, Georgia

2008

Effects of Common Errors In Microsatellite Data On Estimates of Population Differentiation

AND

Inferring Genotypic Structure Of Complex Disease Loci Using Genome-Wide Expression Data

by

Ellen Hepfer Breazel

Major Professor:	Paul Schliekelman
Committee:	Travis Glenn Daniel Hall Jaxk Reeves Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

DEDICATION

To my husband and my parents

ACKNOWLEDGEMENTS

I would first like to thank my major professors Paul Schliekelman and Travis Glenn. Paul for all of his guidance, and patience, in the writing and research of this dissertation and Travis for his unbiased advice throughout my graduate career and guidance on writing this dissertation, both of them have made my graduate research exciting and interesting.

Thank you also to the other members of my committee: Dan Hall for his guidance on EM algorithm and its many challenges and his guidance in the classroom; Jaxk Reeves for his teaching in the classroom, his guidance on research, and for accompanying me to SREL to observe the lab; and Lynne Seymour for her help with the statistical computing challenges, all the fun, and for always being there to listen. I would also like to thank Machelle Wilson for hiring me as a research assistant at SREL and all of her introduction to statistical genetics. I have missed working with her.

I would also like to acknowledge Ross Iaci, Archan Bhattacharya, Jesse Bowling, Jimmy Cretney, Amy Vaughan, Ana Bargo, Jien Chen, Guoying Sun, Elizabeth Young, Lanier Senter, and Tremika Finney for all of their friendship and support. Thank you to Daphney for her continuous smiles and laughter.

I would like express my gratitude to my family. Thank you to my parents, Christine and Tom, who have always been supportive in everything I do. To my sister Holly, brother-in-law Abe and nephew Lance for their continued love and support. Finally, Eric, my husband, my best friend, and my rock, this is as much his accomplishment as mine.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 – MICROSATELLITE PROJECT

Microsatellites are genetic markers often used for population genetics analyses. They are short sequences of DNA usually two to nine base pairs long that are tandem repeats at more than one location. These genetic markers are found in many types of organisms from yeast to humans.

Mononucleotide – AAAAAA
Dinucleotide – CACACACACACA
Trinucleotide – ATCATCATCATCATC
Tetranucleotide - ACTCACTCACTCACTCACTC

The locus of the microsatellite is the position of the gene on its chromosome, in other words its location. For microsatellite data each location or locus in a chromosome that contains core repeats may have a different number (n) of copies of the repeat. This means that for any locus a microsatellite can have multiple allele possibilities in each population. The allele possibilities are the possible allele sizes that each individual can have at a particular locus. The allele sizes are the number of base pairs within each microsatellite. Each possible allele has a frequency associated with it known as the allele frequency. For each individual, each locus has two alleles, one paternal and one maternal, the locus is homozygote if the alleles are the same, heterozygote if they differ.

Microsatellites are used to answer questions about relationships among individuals, populations, and closely related species (BRUFORD and WAYNE 1993; QUELLER *et al.* 1993;

SELKOE and TOONEN 2006). The challenges associated with obtaining correct microsatellite-genotypes are most clear in the fields of forensics, ancient/old/museum DNA, and non-invasive sampling. It is from these fields that the most egregious errors were made, but also the most progress in establishing laboratory standards (see POMPANON *et al.* 2005 for a review). Most researchers have likewise focused on laboratory protocols and practices to reduce errors in microsatellite datasets.

When answering fundamental questions in population genetics the ideal situation is to obtain a large number of DNA samples that are able to be fully analyzed from an experimental procedure that has little to no variation (i.e., same researcher, lab, instruments, chemicals, etc.) then conclusions will be made from this data. If this ideal situation is achieved there is a high probability that the statistics obtained will be correct. However, in many to most situations geneticists have limitations on the number of DNA samples collected due to money or resources and a limited number of markers/loci are available (KALINOWSKI *et al.* 2006; SELKOE and TOONEN 2006). The amount of previous experience by researchers and the use of good laboratory practices also contribute to producing consistently correct data (PAETKAU 2003). However, even if a large number of DNA samples are able to be obtained, the large datasets make it difficult to perform experimental designs without changes in lab temperature, and/or differences in experimenter and therefore mistakes in genotyping will be made. Indeed, in large datasets mistakes are a statistical certainty (HOFFMAN and AMOS 2005).

Although genotyping errors from a variety of sources have long been recognized (see POMPANON *et al.* 2005 and references therein), few studies reported error rates until recent publications noted the value in knowing the frequency of such errors (BONIN *et al.* 2004; BROQUET and PETIT 2004; HOFFMAN and AMOS 2005). Indeed attempts to publish error rates in

the 1990's and early 2000's were nearly certain to draw criticism from reviewers. Attitudes were often expressed that allowing genotyping errors to remain in a dataset was sloppy science and a result of poor markers/methods/training/etc. Thus, huge amounts of resources were devoted to achieving an impossible goal (error-free datasets), or at least driving the error rate to some perceived level of insignificance where it would be ignored (see DAKIN and AVISE 2004; MILLER *et al.* 2002; SEFC *et al.* 2003; WAITS and PAETKAU 2005). Attitudes about genotyping errors have changed remarkably in the past four years.

Although the taboo of acknowledging the presence of genotyping errors has been broken, questions remain about the effects of these errors. Genotyping errors can clearly affect the estimated values of parameters of interest, and much work has been done recently to determine how much these errors change the conclusions drawn from the values obtained. For example, in genetic mark and recapture studies using non-invasive sampling, genotyping errors cause a systematic upward bias in population size estimates that can exceed 200% (ROON *et al.* 2005). In addition, for mark and recapture estimates a Windows based program DROPOUT has been developed to determine if a data-set is error free or if not which loci are producing the most errors (MCKELVEY and SCHWARTZ 2005). Genotyping errors also have significant effects on parentage assessment, particularly when using parental exclusion (ARAKI and BLOUIN 2005; DAKIN and AVISE 2004; VANDEPUTTE *et al.* 2006). Many methods that incorporate genotyping errors into parentage assessment have been proposed (e.g., (KALINOWSKI *et al.* 2007; MORRISSEY and WILSON 2005)) and a DOS based program FAP has been developed to aid in problematic loci for parentage assessment, when parental genotypes are known (TAGGART 2007). Alternatively, few studies have been conducted to determine the effects of genotyping errors on estimates of

population differentiation, and more work in this area is needed (BONIN *et al.* 2004; HOFFMAN and AMOS 2005; POMPANON *et al.* 2005).

We seek to understand better the effects of common genotyping errors (see below) on measures of population differentiation. Our goal for this project is both retrospective (i.e., to help interpret and apply an appropriate amount of skepticism to previous studies) and prospective (i.e., to provide tools for future researchers to better interpret their data). We are particularly interested in determining the effects of genotyping errors when small numbers of individuals and loci have been sampled.

In this study, we focus on a commonly used estimator/parameter of population differentiation F_{ST} (WRIGHT 1951; WRIGHT 1965). Herein we use the term F_{ST} to indicate the generic metric of population differentiation calculated in any of several ways (F_{ST} - (WRIGHT 1951); G_{ST} - Nei 197x (NEI 1973); $\hat{\theta}_{ST}$ - (WEIR and COCKERHAM 1984); Phi_{ST} - Rousset et al. 199x (ROUSSET 1996)), the term $\overline{\theta}_{ST}$ for the statistic calculated from a complete dataset (using the formulas for $\hat{\theta}_{ST}$), and we use $\hat{\theta}_{ST}$ to indicate the statistic calculated from a sample of the complete dataset. F_{ST} is a standard statistic based on allele frequencies used to determine the amount of genetic variation among populations by measuring the proportion of genetic variance in a subpopulation (S) to the total genetic variance (T). (GAGGIOTTI *et al.* 1999) demonstrated that F_{ST} is more consistent than R_{ST} (a similar measure based on variance of allele sizes; (SLATKIN 1995) when the numbers of individuals and loci sampled were small. (BALLOUX and GOUDET 2002) compared F_{ST} and R_{ST} for a fixed number of individuals, arranged in different population sizes, a variable number of migrants, a variable number of loci, and a variable number of mutation rates. (BALLOUX and GOUDET 2002) found that R_{ST} performed better when the populations are highly structured, but F_{ST} does better when the populations are weakly

structured. They also confirmed that F_{ST} is more accurate than R_{ST} when small numbers of individuals are sampled. Based on the results of these two studies and because we are using small numbers of loci and individuals in several of our simulations, we decided to use F_{ST}.

To test the effect of genotyping errors on F_{ST} we generated simulated data using EASYPOP (BALLOUX 2001). EASYPOP generates data according to input parameters on: ploidy (haploid, diploid, haplodiploids), number of populations, number of individuals, migration type and rate, linkage of loci, number of allelic states, and variability. This program is helpful because it simulates both migration and mutation in the data. The output from this program was then used as input for the program we developed. Our program incorporates the three of the most common types of genotyping errors (allelic dropout, binning error, and null alleles), and then calculates F_{ST} on the data with these errors.

Allelic dropout is the creation of "false homozygotes", and usually occurs from limited amounts of poor quality DNA, such as museum samples or hair samples. A false homozygote is created when one of the alleles of a particular locus is not expressed due to the software not recognizing the allele or an amplification bias in the DNA mixture and a locus which was previously heterozygous is now homozygous. Allelic dropout is typically a random effect on FST.

Binning error is one cause of the incorrect assignment of allele size and usually occurs from inconsistencies in laboratory procedures. The tools used for the analysis of microsatellite data consider allele sizes to be continuous values rather than discrete. Therefore, allele sizes for trinucleotide data could be rounded by the researcher or software to be three nucleotides more or less than it should be (if considering tri-nucleotide data). Binning errors can have a systematic or random effect on FST. In our project we are looking at the random effects.

Null Alleles occur when an allele is present in an individual, but is not assayed. For example, suppose three individuals have alleles A_1A_4 , B_1A_4, and A_4A_4 at a particular locus. If A_4 is not assayed within the entire experiment then the individuals are read in as $A_1Ø$, $B_1Ø$, and ØØ and therefore become A_1A_1, B_1B_1, and ØØ, where ØØ is missing data. Null Alleles are different from Allelic Dropout because a particular allele is not expressed throughout the entire dataset instead of randomly deleting alleles. This is more of a systematic deletion rather than the random deletion of allelic dropout.

In general, we were motivated to find error rates where the conclusions drawn from the data would be erroneous (i.e., determining how bad the data can be and still yield the correct conclusion). For this reason, we tested a variety of dataset with small sample-sizes of individuals and loci, and extremely high levels of genotyping errors.

1.2 – Gene Mapping For Complex Traits

Gene mapping is the process of identifying the location of genes on chromosomes. In particular gene mapping of diseases is the process of finding the location of the genes that cause the disease. Complex diseases are diseases that are influenced by multiple genes. It has been found that the mapping of genetically complex diseases such as schizophrenia, bipolar disorder, and diabetes, is a much more difficult task than it was once thought to be. However, single locus traits, such as cystic fibrosis, and Huntington's disease, have proven to be easier to map (Risch and Merikangas 1996). For single locus diseases, if an individual has a certain single locus disease it is highly likely that the individual has a certain genotype at a particular locus that caused that disease as well as if that certain genotype is found at that particular locus then the individual is likely to have the disease. That is to say, single locus traits have a strong correlation

between genotype at the causative locus and the trait. In multiple locus traits it is not necessarily true that a certain genotype is present when a trait is expressed nor is it true that if the certain genotype is present that the trait will be expressed. With complex diseases, the multiple genes can interact with each other or environmental factors. Because of this, there has been little success in mapping complex trait loci. Unfortunately, a majority of genetically based diseases that affect humans are complex. Therefore, there is a great need for better ways to map these complex diseases.

Some traits in observation are binary traits that do not have simple Mendelian (simple trait) inheritance patterns. These traits such as, affected versus unaffected to a complex disease have underlying quantitative attributes. Usually these binary traits are treated as threshold variables, meaning the underlying continuous variable, known as the liability, has a point which separates the two phenotypic traits. (XU and ATCHLEY 1996; YI and XU 1999) These binary traits are difficult to map because of the complexity between phenotype and the liability variable (YI and XU 2000).

In addition to binary phenotypes, complex traits may also result in continuous outcomes, these are known as quantitative traits. For example, body mass is considered a complex quantitative trait. There are many genes and environmental factors that contribute to determining body mass. Therefore, modification of only one of these genes would usually change body mass only slightly. Body mass can be considered a quantitative trait because its phenotypes are on a continuous scale. Quantitative traits can be mapped using standard methods (LYNCH and WALSH 1998) to quantitative trait loci (QTL), regions of the DNA that are associated with the phenotypic trait in question. These QTLs are genomic regions that may contain tens or hundreds of genes. If so there are fine mapping methods that are used to narrow down the region and positional

cloning (sequencing the region) is used to identify the precise genes that are associated with that phenotypic trait, if the gene sequences are known. There have been cases where mapping QTLs have been successful for diseases (GLAZIER *et al.* 2002) however, that seems to be the exception rather than the rule.

Gene mapping methods without any previous knowledge on the location of these complex traits require a genome-wide study, a method for searching the entire genome for association to the trait. There are two common genetic analysis approaches for genome scans, linkage mapping and association studies.

Linkage mapping is the detection of trait markers within families whose trait genotype is more common among individuals with the trait than would happen by chance. (HIRSCHHORN and DALY 2005) This method has proven to be very successful for locating genes associated with Mendelian traits. Linkage analysis has not been as successful with complex traits. A review study was done to look at over 100 whole-genome scan studies of complex diseases (ALTMULLER *et al.* 2001). Of these studies only a third produced significant linkages and of those few have been found to be significant if repeated. There are many factors that can attribute to the low success rate of linkage analysis including but not limited to the low heritability of complex traits, and low power due to insufficient sample sizes (RISCH and MERIKANGAS 1996). Because of the use of family data for linkage, and within families the number of recombination events is small, it is possible that the smallest DNA region of interest detected may still contain hundreds of genes (CARLSON *et al.* 2004). Linkage mapping is a good tool for identifying rare alleles that contribute to the disease however it not as good in detecting common variants that also have been shown to contribute greatly to common diseases (LOHMUELLER *et al.* 2003).

Association studies are better at detecting these common variants but are not good at detecting rare alleles (HIRSCHHORN and DALY 2005). Genome-wide association studies are the process of surveying most of the genome to determine casual genetic variants, usually by comparing affected individuals versus unaffected individuals within a genetic marker. With the completion of the human genome project and the HapMap Project (see below) it has become increasingly popular to use single nucleotide polymorphisms (SNPs) as markers to determine association with a trait. SNPs are single nucleotide variations within DNA that can be used as genetic markers. Association is usually determined using case-control studies. In order for associations to be found the most useful genetic markers are either the causal allele or ones that are highly correlated (in linkage disequilibrium) with the casual allele. This will be a smaller region of DNA than those found in linkage analysis. Association studies develop a loss of power when more than one disease causing allele is present. This is because there will be less contrast between genotypes. This is not a problem in linkage analysis because specific alleles are not observed. One popular method for association studies in humans is transmission equilibrium tests. However, it has been shown (SLAGER *et al.* 2000) that in order to achieve reasonable statistical power for mapping using this test, tens of thousands or even millions of families may be needed.

The concept of linkage disequilibrium is important to the power of genome-wide association studies (CARDON and BELL 2001). Linkage disequilibrium is the concept that makers that are near each other within the chromosome tend to be inherited together and therefore have high correlation with each other. It has been discovered that these correlation in humans occur in a block structure within the genome, the HapMap Project is a catalog of these linked locations (ALTSHULER *et al.* 2005). To perform genome-wide association studies on SNPs (or whichever

genetic marker is chosen), in order to span the entire genome, millions of makers may be tested for association. This proposes the problem of loss of power due to multiple testing. The development of the HapMap has proposed a reduction in the number of tests needed for association analysis. Instead of testing all 10 million SNPs, only a few of the SNPs that are closely related will have to be tested. Some of the recent genome-wide association studies using HapMap data have found loci associated with type 2 diabetes (SLADEK *et al.* 2007), inflammatory bowel disease (DUERR *et al.* 2006), breast cancer (HUNTER *et al.* 2007) and prostate cancer (YEAGER *et al.* 2007).

As an alternative approach to linkage or association studies there have been a number of studies recently that examine the use of microarray expression data and have found significant genetic variation in expression levels. Gene expression is a term used for the quantitative value that describes the information that is transcribed from within the DNA into messenger RNA. Microarrays are tools for being able to see the expression levels of multiple genes at a time. The outcome of a microarray experiment is an array of spots on a slide. Each spot represents a gene, and will be colored yellow, red, green, or black according to that gene's expression. For microarray expression analysis, for a study of disease versus non-diseased genes, if a gene is overexpressed in a disease state (as compared to non-disease state) those genes will appear more red than green.

Microarrays can be used for sequencing, detection of SNPS, or genetic mechanisms in living cells (DRAGHICI 2003). The most widely used method is expression analysis. Microarrays have been proven to be a reliable gene expression tool . Microarray expression data can be gathered for one sample, or multiple samples. In general for one sample the goal is to determine mean and variance of expression levels within the sample. For two sample expression analysis

the goal is to determine genes that are differentially expressed between the two samples. An example of this would examine tissue from breast cancer patients versus the same tissue from healthy individuals, to determine the genes that are differentially expressed between the two and may point to genes useful in further breast cancer studies. Usually multiple t-tests are performed on the expression levels to determine differential expression.

Over the past several years there has been an increasing interest in combining microarray expression analysis with molecular marker data. This is a strategy known as "genetical genomics" (JANSEN and NAP 2001). The idea behind genetical genomics is that one can treat microarray gene expression levels as different quantitative phenotypic traits then use linkage or association methods to determine significant quantitative trait loci. Such QTLs are known as expression QTLS (eQTLs).

This method of genetical genomics has been successful at showing high heritability (see below) of eQTLs in yeast (Brem and Kruglyak 2005; Brem et al. 2002), plants, and humans (MONKS *et al.* 2004; MORLEY *et al.* 2004; SCHADT *et al.* 2003). These studies have been reviewed (LI and BURMEISTER 2005; STAMATOYANNOPOULOS 2004) and have shown that microarray data may be extremely useful in mapping loci associated with complex traits. There have also been studies using genetical genomics methods that found causative genes related to weight in mice (GHAZALPOUR *et al.* 2006; SCHADT *et al.* 2003).

For example, Schadt *et al.* performed an F2 cross (creation of a generation where both parental phenotypes occur) between two inbred mouse strains and performed a genome-wide scan for linkage of expression levels of 23,574 genes. These mice were placed on a high-fat diet and therefore a wide array of obesity occurred among them. The mice were classified according to their fat pad mass (FPM) trait. The expression levels for the 25% of the mice with the highest

FPM were compared to those expression levels for the 25% of the mice with the lowest FPM. 280 genes were identified to be differentially expressed between these two groups. It was found that clustering the mice on this set of genes divided the mice into two high-FPM groups (high FPM group-1 and high FPM group-2) and one-low FPM group. A genome scan was then performed for those mice classified as high FPM group-1 or low FPM group and another scan on those classified as high FPM group-2 or low FPM group. They found the association log-odd scores of eQTLs for FPM were substantially increased when only one high FPM group was included. They also showed that a number of the genes found from the reduce scan had expression levels mapping to the same region as a QTL for the FPM trait prior to reduction, just with higher correlation. The idea is if there are expression levels that are highly heritable (see below), there is evidence that several of these expression levels map to the same chromosomal region and possibly the same locus. It is possible that causative loci (loci that are responsible for variation in disease or trait) affect expression levels of other genes; these expression levels would then have high power for mapping the causative loci.

The goal of gene mapping studies is to demonstrate that a phenotypic trait is influenced by inherited factors. Heritability is the percentage of the phenotypic variation among individuals that is affected by genotypic variation (GIBSON and WEIR 2005). The papers mentioned above, that have used genetical genomics methods, show high levels of heritability of expression levels with respect to mapped QTLs. Although due to the large number of genes studied it would not be surprising to see high levels of heritability by chance. For example, Monk et al (2004) found very high QTL heritability values for their 55 significant eQTLs. However, these were selected from 24,000 genes. While these heritability values were shown to be significant, because of the large number of genes they may not actually be as high as estimated. BREM and KRUGLYAK

(2005) found a way to partially avoid this problem. They mapped QTLs for the expression levels of roughly 6,000 genes in a cross between two yeast strains. They found that 3,546 expression levels had significant heritability values. They then used half of their data as a QTL detection set and the other half of the data to estimate the proportion of variance explained by the QTLs that were detected. This gave an independent estimate of the proportion of variance explained. They found a median of 27% of variance explained by QTLs with 16% of QTLS explaining more than 60% of the variance. They also estimated bounds on the number of expression levels controlled by a given number of loci. They concluded that 3% (106) of genes were consistent with 1 locus control, 17-18% (620) were consistent with 1-2 locus control, and 50% required more than 5 loci. This still proposes a multiple testing issue because they estimated variance on 3546 genes. However, due to the fact that they found heritability levels of over 69% for all 3546 of those genes, there is little doubt that there were hundreds (over 600) of expression levels with simple (1-2 loci) inheritance and high heritability.

There is a potential problem with this genetical genomics method. A hypothetical pathway for complex genetic traits is shown in Figure 1.1. Ovals represent disease causative loci – that is, loci with genotype variation that leads to variation in disease risk. Rectangles represent genes in a network, whose transcript variation between individuals leads to variation in disease risk between individuals. Arrows between loci and genes in the network represent an impact of genotypic variation at the locus on the gene's transcript level. Arrows between genes in the network represent control of transcription. Arrows between genes and the disease represent an impact of the genes transcription on disease probability. This figure is not meant to be taken literally. Realistically things are probably much more complicated than Figure 1.1 implies.

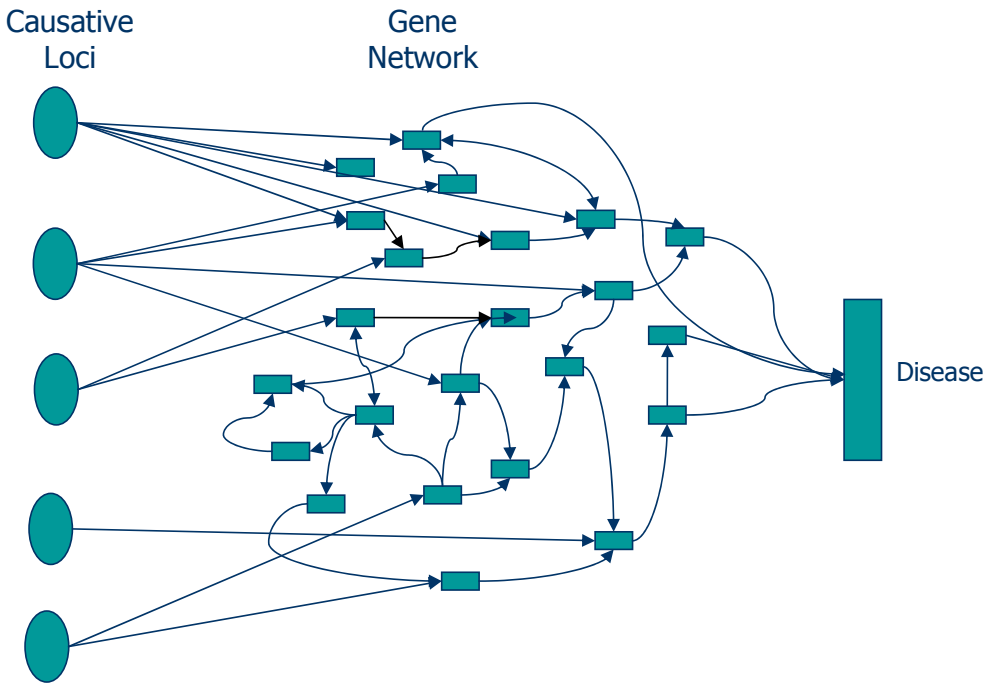


Figure 1.1 - Hypothetical Schematic of Disease Genetic Pathway

The ideal expression level for mapping a causative locus would be one whose variation is entirely determined by the genotype at a single causative locus. However, such an expression level will have poor correlation with the disease and may be difficult to detect as differentially expressed. Expression levels that are highly correlated with the disease would be easy to determine as differentially expressed, but would not be any better than the disease itself for mapping. Schadt *et al.* found a large number of eQTLs mapping to locations throughout the genome, regardless of trait status. To determine which of these eQTLs are relevant requires determining a correlation between the expression level and the trait of interests, which may be weakly correlated with causative loci. Therefore we may be exchanging the low power due to establishing a correlation between trait and genotype to a low power due to establishing a correlation between trait and expression level. It has been shown (SCHLIEKELMAN 2008) that

there is a tradeoff between the power to show an association between expression level and disease and the power to map eQTLs for those expression levels. Schliekelman also determined that power to map eQTLs under an additive penetrance model is significantly worse than with a multiplicative model.

As mentioned above, there have been several studies that use microarrays to identify causative loci, all of which have some prior knowledge of the location of the causative loci. Our goal is to develop a method that incorporates genome-wide microarray expression data as quantitative traits to map causative loci for a complex trait without any prior knowledge of locus location.

1.3 – References

ALTMULLER, J., L. J. PALMER, G. FISCHER, H. SCHERB and M. WJST, 2001 Genomewide scans of complex human diseases: True linkage is hard to find. American Journal of Human Genetics **69:** 936-950.

ALTSHULER, D., L. D. BROOKS, A. CHAKRAVARTI, F. S. COLLINS, M. J. DALY *et al.*, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

ARAKI, H., and M. S. BLOUIN, 2005 Unbiased estimation of relative reproductive success of different groups: evaluation and correction of bias caused by parentage assignment errors. Molecular Ecology **14:** 4097-4109.

BALLOUX, F., 2001 EASYPOP (version 1.7): A computer program for population genetics simulations. Journal of Heredity **92:** 301-302.

BALLOUX, F., and J. GOUDET, 2002 Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Molecular Ecology **11:** 771-783.

BONIN, A., E. BELLEMAIN, P. B. EIDESEN, F. POMPANON, C. BROCHMANN *et al.*, 2004 How to track and assess genotyping errors in population genetics studies. Molecular Ecology **13:** 3261-3273.

BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences **102:** 1572-1577.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

BROQUET, T., and E. PETIT, 2004 Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology **13:** 3601-3608.

BRUFORD, M. W., and R. K. WAYNE, 1993 Microsatellites and their application to population genetic studies. Curr Opin Genet Dev **3:** 939-943.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nature Reviews Genetics **2:** 91-99.

CARLSON, C. S., M. A. EBERLE, L. KRUGLYAK and D. A. NICKERSON, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429:** 446-452.

DAKIN, E. E., and J. C. AVISE, 2004 Microsatellite null alleles in parentage analysis. Heredity **93:** 504-509.

DRAGHICI, S., 2003 *Data Analysis Tools for DNA Microarrays*. CRC Press UK, London, UK.

DUERR, R. H., K. D. TAYLOR, S. R. BRANT, J. D. RIOUX, M. S. SILVERBERG *et al.*, 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science **314:** 1461-1463.

GAGGIOTTI, O. E., O. LANGE, K. RASSMANN and C. GLIDDON, 1999 A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Molecular Ecology **8:** 1513-1520.

GHAZALPOUR, A., S. DOSS, B. ZHANG, S. WANG, C. PLAISIER *et al.*, 2006 Integrating genetic and network analysis to characterize genes related to mouse weight. Plos Genetics **2:** 1182-1192.

GIBSON, G., and B. WEIR, 2005 The quantitative genetics of transcription. Trends in Genetics **21:** 616-623.

GLAZIER, A. M., J. H. NADEAU and T. J. AITMAN, 2002 Finding genes that underlie complex traits. Science **298:** 2345-2349.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6:** 95-108.

HOFFMAN, J. I., and W. AMOS, 2005 Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology **14:** 599-612.

HUNTER, D. J., P. KRAFT, K. B. JACOBS, D. G. COX, M. YEAGER *et al.*, 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics **39:** 870-874.

JANSEN, R. C., and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends in Genetics **17:** 388-391.

KALINOWSKI, S. T., M. L. TAPER and S. CREEL, 2006 Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. Conservation Genetics **7:** 319-329.

KALINOWSKI, S. T., M. L. TAPER and T. C. MARSHALL, 2007 Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology **16:** 1099-1106.

LI, J., and M. BURMEISTER, 2005 Genetical genomics: combining genetics with gene expression analysis. Human Molecular Genetics **14:** R163-R169.

LOHMUELLER, K. E., C. L. PEARCE, M. PIKE, E. S. LANDER and J. N. HIRSCHHORN, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics **33:** 177-182.

LYNCH, M., and B. WALSH, 1998 *Genetics and analysis of quantitative traits.* Sinauer Associates, Sunderland, MA, USA.

MCKELVEY, K. S., and M. K. SCHWARTZ, 2005 DROPOUT: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. Molecular Ecology Notes **5:** 716-718.

MILLER, C. R., P. JOYCE and L. P. WAITS, 2002 Assessing allelic dropout and genotype reliability using maximum likelihood. Genetics **160:** 357-366.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. American Journal of Human Genetics **75:** 1094-1105.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

MORRISSEY, M. B., and A. J. WILSON, 2005 The potential costs of accounting for genotypic errors in molecular parentage analyses. Molecular Ecology **14:** 4111-4121.

NEI, M., 1973 Analysis Of Gene Diversity In Subdivided Populations. Proceedings of the National Academy of Sciences of the United States of America **70:** 3321-3323.

PAETKAU, D., 2003 An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology **12:** 1375-1387.

POMPANON, F., A. BONIN, E. BELLEMAIN and P. TABERLET, 2005 Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics **6:** 847-859.

QUELLER, D. C., J. E. STRASSMANN and C. R. HUGHES, 1993 Microsatellites And Kinship. Trends in Ecology & Evolution **8:** 285-&.

RISCH, N., and K. MERIKANGAS, 1996 The future of genetic studies of complex human diseases. Science **273:** 1516-1517.

ROON, D. A., L. P. WAITS and K. C. KENDALL, 2005 A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data. Animal Conservation **8:** 203-215.

ROUSSET, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357-1362.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHLIEKELMAN, P., 2008 Statistical Power of Expression Quantitative Trait Loci for Mapping of Complex Trait Loci in Natural Populations. Genetics **178:** 2201-2216.

SEFC, K. M., R. B. PAYNE and M. D. SORENSON, 2003 Microsatellite amplification from museum feather samples: Effects of fragment size and template concentration on genotyping errors. Auk **120:** 982-989.

SELKOE, K. A., and R. J. TOONEN, 2006 Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters **9:** 615-629.

SLADEK, R., G. ROCHELEAU, J. RUNG, C. DINA, L. SHEN *et al.*, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445:** 881-885.

SLAGER, S. L., J. HUANG and V. J. VIELAND, 2000 Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genetic Epidemiology **18:** 143-156.

SLATKIN, M., 1995 A Measure Of Population Subdivision Based On Microsatellite Allele Frequencies. Genetics **139:** 457-462.

STAMATOYANNOPOULOS, J. A., 2004 The genomics of gene expression. Genomics **84:** 449-457.

TAGGART, J. B., 2007 FAP: an exclusion-based parental assignment program with enhanced predictive functions. Molecular Ecology Notes **7:** 412-415.

VANDEPUTTE, M., S. MAUGER and M. DUPONT-NIVET, 2006 An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Molecular Ecology Notes **6:** 265-267.

WAITS, L. P., and D. PAETKAU, 2005 Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. Journal of Wildlife Management **69:** 1419-1433.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-Statistics For The Analysis Of Population-Structure. Evolution **38:** 1358-1370.

WRIGHT, S., 1951 The Genetical Structure Of Populations. Annals of Eugenics **15:** 323-354.

WRIGHT, S., 1965 The Interpretation Of Population-Structure By F-Statistics With Special Regard To Systems Of Mating. Evolution **19:** 395-420.

XU, S. Z., and W. R. ATCHLEY, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417-1424.

YEAGER, M., N. ORR, R. B. HAYES, K. B. JACOBS, P. KRAFT *et al.*, 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics **39:** 645-649.

YI, N. J., and S. Z. XU, 1999 Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity **82:** 668-676.

YI, N. J., and S. Z. XU, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391-1403.

CHAPTER 2

EFFECTS OF COMMON ERRORS IN MICROSATELLITE DATA ON ESTIMATES OF POPULATION DIFFERENTIATION[1]

[1] Breazel, E.H. and T. C. Glenn. To be submitted to *Molecular Ecology*.

ABSTRACT

Microsatellite DNA loci are commonly used in population genetic studies. Many researchers use good lab practices to minimize genotyping errors, but it is statistically impossible to eliminate all genotyping errors in large datasets. Most researchers have thus focused on reducing errors to levels where they believe the errors won't cause erroneous conclusions to be drawn from the data. Unfortunately, answers to basic questions regarding the numbers of loci and individuals needed to estimate population differentiation and the effects of genotyping errors rely on limited numbers of simulation-based studies. Thus, the amount of effort that should be expended on reducing errors and the effects of well-known errors that occur in almost every dataset (allelic dropout, binning errors, and null alleles) remain in the realm of intuition and rules of thumb. We systematically tested the effects of these common genotyping errors on a common measure of population differentiation, F_{ST}. Allelic dropout had no effect on the average value of F_{ST}, but large amounts of dropout increase the variance of F_{ST} calculated from any particular dataset and this variance increases with decreasing sample sizes of individuals or loci. Incorrectly binning alleles and the presence of null alleles had a potentially large and statistically significant effect on the value of F_{ST}. Binning errors bias estimates of F_{ST} downward (i.e., making populations appear less different than they actually are) whereas null alleles biased estimates of F_{ST} upward (i.e., making populations appear more different than they actually are). However, despite the statistically significant bias on the estimate of F_{ST} the conclusions about population differentiation would most likely remain the same, even at unrealistically high percentages of genotyping errors.

2.1 – INTRODUCTION

Microsatellite DNA loci (known by many other names and acronyms, but simply called microsatellites hereafter) are small repetitive elements common in eukaryotic genomes (TAUTZ and RENZ 1984) that have been used as single-locus markers in genetic studies since 1989 (TAUTZ 1989; WEBER and MAY 1989). Microsatellites are used to answer questions about relationships among individuals, populations, and closely related species (BRUFORD and WAYNE 1993; QUELLER *et al.* 1993; SELKOE and TOONEN 2006). The challenges associated with obtaining correct microsatellite-genotypes are most clear in the fields of forensics, ancient/old/museum DNA, and non-invasive sampling. It is from these fields that the most egregious errors were made, but also the most progress in establishing laboratory standards (see POMPANON *et al.* 2005 for a review). Most researchers have likewise focused on laboratory protocols and practices to reduce errors in microsatellite datasets.

Ideally, a large number of DNA samples are fully analyzed at large numbers of loci using experimental procedures with little variation (i.e., same researcher, lab, instruments, chemicals, etc.), yielding consistently correct data. Unfortunately, in most studies the experimental conditions and thus, resulting data are less than ideal. Numbers of loci and DNA samples available are generally limited by money, time, or other resources (KALINOWSKI *et al.* 2006; SELKOE and TOONEN 2006). The amount of previous experience by researchers and the use of good laboratory practices also contribute to producing consistently correct data (PAETKAU 2003). However, even if a large number of DNA samples are obtained, large numbers of loci available, and all genotyping is done by experienced researchers using good laboratory practices, then the time needed to accumulate large datasets make it difficult to perform experimental designs without changes over time (e.g., lab temperature or humidity, chemicals, instruments, etc.).

Therefore, mistakes in genotyping will be made. Indeed, in large datasets mistakes are a statistical certainty (HOFFMAN and AMOS 2005).

Although genotyping errors from a variety of sources have long been recognized (see POMPANON *et al.* 2005 and references therein), few studies reported error rates until recent publications noted the value in knowing the frequency of such errors (BONIN *et al.* 2004; BROQUET and PETIT 2004; HOFFMAN and AMOS 2005). Indeed attempts to publish error rates in the 1990's and early 2000's were nearly certain to draw criticism from reviewers. Attitudes were often expressed that allowing genotyping errors to remain in a dataset was sloppy science and a result of poor markers/methods/training/etc. Thus, huge amounts of resources were devoted to achieving an impossible goal (error-free data sets), or at least driving the error rate to some perceived level of insignificance where it would be ignored (see DAKIN and AVISE 2004; MILLER *et al.* 2002; SEFC *et al.* 2003; WAITS and PAETKAU 2005). Attitudes about genotyping errors have changed remarkably in the past four years.

Although the taboo of acknowledging the presence of genotyping errors has been broken, questions remain about the effects of these errors. Genotyping errors can clearly affect the estimated values of parameters of interest, and much work has been done recently to determine how much these errors change the conclusions drawn from the values obtained. For example, in genetic mark and recapture studies using non-invasive sampling, genotyping errors cause a systematic upward bias in population size estimates that can exceed 200% (ROON *et al.* 2005). In addition, for mark and recapture estimates a Windows based program DROPOUT has been developed to determine if a dataset is "error free" or if not which loci are producing the most errors (MCKELVEY and SCHWARTZ 2005). Genotyping errors also have significant effects on parentage assessment, particularly when using parental exclusion (ARAKI and BLOUIN 2005;

DAKIN and AVISE 2004; VANDEPUTTE *et al.* 2006). Many methods that incorporate genotyping errors into parentage assessment have been proposed (e.g., KALINOWSKI *et al.* 2007; MORRISSEY and WILSON 2005) and a DOS based program FAP has been developed to aid in problematic loci for parentage assessment, when parental genotypes are known (TAGGART 2007). Alternatively, few studies have been conducted to determine the effects of genotyping errors on estimates of population differentiation, and more work in this area is needed (BONIN *et al.* 2004; HOFFMAN and AMOS 2005; POMPANON *et al.* 2005).

Study Goals and General Design

We seek to understand better the effects of common genotyping errors (Table 2.1) on measures of population differentiation. Our goal for this project is both retrospective (i.e., to help interpret and apply an appropriate amount of skepticism to previous studies) and prospective (i.e., to provide tools for future researchers to better interpret their data). We are particularly interested in determining the effects of genotyping errors when small numbers of individuals and loci have been sampled.

Table 2.1 – Summary of Genotyping Errors. We use the term pseudo-random to mean causes that are systematic in nature, but are expressed variably under typical experimental conditions so that the effect appears random. An example of this would be a single nucleotide polymorphism in the primer-binding region of a locus. Minor differences in PCR conditions (due to variance in pipetting, DNA concentrations, thermal cyclers, etc.) may cause some alleles to be null (unscorable) in some experiments, but scored in other (replicate) experiments. Pseudo-random processes were approximated in simulations with the same algorithms used for random processes.

Error	**Error Type**	**Cause**
Allelic Dropout	*Random	Bad DNA, limited DNA, PCR inhibitor
	Systematic	Large allele size differences
Binning Error	*Pseudo-random	Undetected variance in chemicals, equipment,
	Systematic	variance chemicals, equipment, researchers
Null Alleles	Pseudo-random	Minor primer mismatches, non-optimal PCR, PCR inhibitors
	*Systematic	Primer mismatches

*Modeled and tested

In this study, we focus on a commonly used estimator/parameter of population differentiation F_{ST} (WRIGHT 1951; WRIGHT 1965). Herein we use the term F_{ST} to indicate the generic metric of population differentiation calculated in any of several ways (F_{ST} - WRIGHT 1951; G_{ST} - NEI 1973; $\hat{\theta}_{ST}$ - WEIR and COCKERHAM 1984; Phi_{ST} - ROUSSET 1996), the term $\overline{\theta}_{ST}$ for the statistic calculated from a complete dataset (using the formulas for $\hat{\theta}_{ST}$), and we use $\hat{\theta}_{ST}$ to indicate the statistic calculated from a sample of the complete dataset. F_{ST} is a standard statistic based on allele frequencies used to determine the amount of genetic variation among populations by measuring the proportion of genetic variance in a subpopulation (S) to the total genetic variance (T). GAGGIOTTI *et al.* (1999) demonstrated that F_{ST} is more consistent than R_{ST} (a similar measure based on variance of allele sizes; SLATKIN (1995) when the numbers of individuals and loci sampled were small. BALLOUX and GOUDET (2002) compared F_{ST} and R_{ST} for a fixed number of individuals, arranged in different population sizes, a variable number of migrants, a variable number of loci, and a variable number of mutation rates. BALLOUX and GOUDET (2002) found that R_{ST} performed better when the populations are highly structured, but F_{ST} does better when the populations are weakly structured. They also confirmed that F_{ST} is more accurate than R_{ST} when small numbers of individuals are sampled. Based on the results of these two studies and because we are using small numbers of loci and individuals in several of our simulations, we decided to use F_{ST}.

To test the effect of genotyping errors on F_{ST} we generated simulated data using EASYPOP (BALLOUX 2001). EASYPOP generates data according to input parameters on: ploidy (haploid, diploid, haplodiploids), number of populations, number of individuals, migration type and rate, linkage of loci, number of allelic states, and variability. This program is helpful because it simulates both migration and mutation in the data. The output from this

program was then used as input for the program we developed. Our program incorporates the three types of genotyping errors (see below), and then calculates F_{ST} on the data with these errors.

We investigate the effects of three of the most common types of genotyping errors in microsatellite datasets (mathematical nature of error):

> **allelic dropout:** an allele is missed due to amplification bias in favor of another allele (systematic) or because it is not included in the aliquot of DNA amplified (random),
> **binning error:** an allele is misassigned due to a sizing error (pseudo-random or systematic), and
> **null allele:** an allele is missed because the primer binding site(s) is/are variable (pseudo-random or systematic).

Specific examples leading to the different types of errors are summarized in Table 2.1. In general, we were motivated to find error rates where the conclusions drawn from the data would be erroneous (i.e., determining how bad the data can be and still yield the correct conclusion). For this reason, we tested a variety of dataset with small sample-sizes of individuals and loci, and extremely high levels of genotyping errors.

2.2 – METHODS

To test the effect of genotyping errors in microsatellite data on F_{ST}, we first generated data that have no genotyping errors. We used EASYPOP (BALLOUX 2001) to create microsatellite data according to specific parameters. EASYPOP generates populations using a Markov Chain. A new dataset has matrix *t* which is full (the variability in the original matrix *t* can be set by the user) and matrix *t+1* which is empty. If the user specified two populations with two sexes and random mating, a female and her mate are chosen from the same population and a new individual is then created. This new offspring has a 1-*m* chance of staying in the same population or m chance of switching populations (i.e., *m* = migration rate). This process is

repeated until both populations are filled, then matrix *t+1* becomes matrix *t*. In all of our simulated files we use diploid data (containing one allele from each parent), with two sexes, random mating, at least 10,000 generations, and 99 different replicates.

These simulations yield 99 different simulated files. For now we will concentrate on only one simulated file, as the 99 different simulated files will act as replicates to our experiment. We performed our experiment on two different datasets (Dataset A and Dataset B). We considered that the effects of genotyping errors on F_{ST} may depend on the value of F_{ST} without genotyping errors. Therefore, we chose a dataset with a relatively large $\overline{\theta}_{ST}$ (0.22) and one with a relatively small $\overline{\theta}_{ST}$ (0.08). The parameters of our two datasets are listed in Table 2.2. Within the parameters a mixed mutation model is a Single-Stepwise Mutation Model (SSM) with a proportion of K-allele model (Kam) mutation events. SSM is a mutation model that increases the number of repeats by one repeat unit (either increasing or decreasing), in EASYPOP each new allele size has a 50% probability of being larger or smaller than the original allele. A user defined proportion of Kam mutation events create new alleles of random size (e.g., 20% of mutations are Kam events). Both SSM and Kam are constrained to the number of allelic states specified by the user. We treated the EASYPOP datasets as if that is the data from all individuals from 2 overall populations. Knowing that it is generally unrealistic to have a sample dataset with 1000 individuals and 20 loci, we created "sub-datasets" that are more consistent with real datasets. The sub-datasets had a number of individuals for each population (N) equal to 10, 20, 50, or 200 and the number of loci equal to 2 through 20 thus creating 76 different sub-datasets. For each of these sub-datasets we incorporated (using our program written in C#) allelic dropout and binning genotypic errors.

Table 2.2 – EASYPOP Parameters and Statistics for Experiment Datasets

	Dataset A	**Dataset B**
Ploidy level	Diploid	Diploid
# of sexes	2	2
Mating system	Random	Random
# of populations	2	2
Same # of individuals in each population	Yes	Yes
# of females in each population	500	500
# of males in each population	500	500
Same migration scheme over all simulation	Yes	Yes
Proportion of female migration	.0001	.001
Proportion of male migration	.0001	.001
# of loci	20	20
Free recombination between loci?	Yes	Yes
All loci have same mutation scheme?	Yes	Yes
Mutation rate	.001	.001
Mutation model	Mixed	Mixed
Proportion of Kam mutation events	0.1	0.1
# of possible allelic states	99	99
Variability of initial population (Minimal means all start with same allele)	Minimal	Minimal
# of generations	10000	20000
# of replicates	99	99
Average FST (Range)	0.2179 (0.1612 – 0.2715)	0.0787 (0.0577 – 0.0985)
Average # Alleles (Range)	17.4 (6-31)	16.7 (6-32)
Average Exp Heterozygosity Pop1 (Range)	0.7198 (0.0929 – 0.8963)	0.7718 (0.4030 – 0.9134)

To simulate allelic dropout a given percentage (***d)*** of alleles were randomly deleted from both populations. Randomness was determined by uniform distribution from 1 to total number

of individuals, within each locus separately (using C# random number generator). Considering that each locus has a maternal and a paternal allele, if the maternal allele was deleted it was replaced by the paternal allele (creating a false homozygous) and vice versa (Figure 2.1).

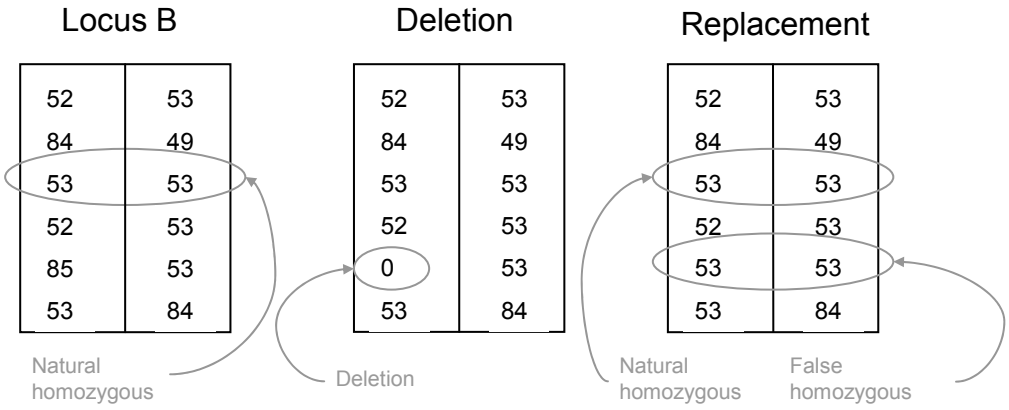


Figure 2.1 – Illustration of Allelic Dropout

If both alleles were deleted then the locus was considered missing data. Missing data is a realistic occurrence when analyzing microsatellite data. We performed these deletions using d = (0.01, 0.05, 0.1, and 0.5). In addition to observing the effect on the average over all 99 replications we also observed the effects on each individual replication as this may be more applicable for researchers with only one dataset.

To simulate binning error we assumed that the random errors of the allele sizing are normally distributed with mean 0. The standard deviation (or variance) of these random errors determines the probability of the allele being binned to another allele size. We simulated binning errors by setting a standard deviation ($\boldsymbol{\sigma}$) of the random errors of the allele sizes and calculated the percentage of alleles that needed to be binned to the next allele size, that percentage of the alleles at a locus were then binned up and down (see Figure 2.2).

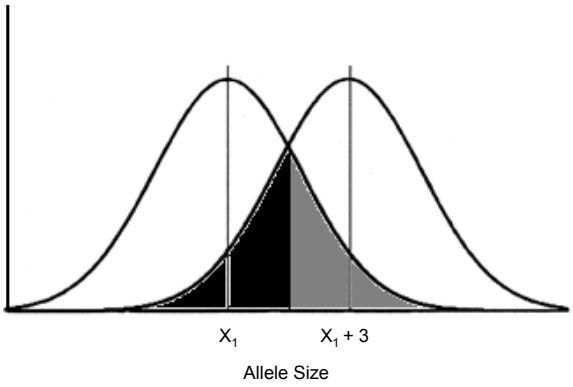


Figure 2.2 – Illustration of Binning Error

For convenience (see discussion) we assumed tri-nucleotide data and therefore binning up added 3 to the allele size, binning down subtracted 3. We performed the binning errors using $\sigma = 1$ (6.68%), 2 (22.66%), and 3 (30.85%). In other words, if the standard deviation is set to 1 then 6.68% of the allele sizes are randomly chosen and increased by 3 and 6.68% of the allele sizes are randomly chosen and decreased by 3 (i.e., a total of 13.36% of all alleles at a locus were binned/scored incorrectly). This is done likewise, for standard deviation of 2 and 3.

Allelic dropout and binning were done simultaneously using every possible combination of the values of *d* and σ as specified above. Thus, 12 different random genotyping errors were simulated for all 76 sub-datasets from each Dataset A and B, and all 99 replicates of each.

Null Alleles were simulated by deleting a particular allele size within an entire locus for both populations. For these genotypic errors, we did not use the sub-datasets used for allelic dropout and binning. Instead we created datasets from the overall populations deleting the allele or combination of alleles closest to our target null frequency (*y)*. The allele or combination of alleles (up to three) in the first population whose allele frequency(s) were closest to *y* was deleted from the entire locus. This was done for *x*% of the loci.

Each error was incorporated in the sub-datasets or datasets 100 times, to account for randomness of the simulated errors. After genotyping errors are incorporated we calculated the value of $\hat{\theta}_{ST}$ (WEIR and COCKERHAM 1984) for the data with and without incorporated errors.

$$\hat{\theta}_{ST} = \frac{a}{a+b+c}$$

$$a = \sum_{j=1}^{L}\left[\sum_{k=1}^{A_j}\left(\frac{\bar{n}_j}{Nc_j}\left\{s_{jk}^2 - \frac{1}{\bar{n}_j - 1}\left[\bar{p}_{jk}\left(1-\bar{p}_{jk}\right) - \frac{r-1}{r}s_{jk}^2 - \frac{1}{4}\bar{h}_{jk}\right]\right\}\right)\right]$$

$$b = \sum_{j=1}^{L}\left[\sum_{k=1}^{A_j}\left(\frac{\bar{n}_j}{\bar{n}_j - 1}\left[\bar{p}_{jk}\left(1-\bar{p}_{jk}\right) - \frac{r-1}{r}s_{jk}^2 - \frac{2\bar{n}_j - 1}{4\bar{n}_j}\bar{h}_{jk}\right]\right)\right]$$

$$c = \sum_{j=1}^{L}\left[\sum_{k=1}^{A_j}\left(\frac{1}{2}\bar{h}_{jk}\right)\right]$$

$\bar{n}_j = \sum_{i=1}^{r}\frac{n_{ji}}{r} =$ average number of non-zero individuals for locus j

$$Nc_j = \frac{\left(r\bar{n}_j - \sum_{i=1}^{r}\frac{n_{ji}^2}{r\bar{n}_j}\right)}{\left(r-1\right)}$$

$\bar{p}_{jk} = \sum_{i=1}^{r}\frac{n_{ji}p_{jki}}{r\bar{n}} =$ average sample freq. of allele k for locus j

$s_{jk}^2 = \sum_{i=1}^{r}\frac{n_{ji}\left(p_{jki} - \bar{p}_{jk}\right)^2}{\left(r-1\right)\bar{n}_j} =$ sample var. of allele k freqs. over pops. for locus j

$\bar{h}_{jk} = \sum_{i=1}^{r}\frac{n_{ji}h_{jki}}{r\bar{n}_j}$ = the average heterozygote freq. for allele k and locus j

$p_{jk} =$ sample freq. for allele k and locus j

$h_{jk} =$ heterozygote freq. for allele k and locus j

$r =$ # of populations

$L =$ # of loci

$A_j =$ # of alleles in locus j

We took an average of the 100 different $\hat{\theta}_{ST}$ values as the $\hat{\theta}_{ST}$ value for a particular genotypic error combination. Using the FSTAT (GOUDET 1995) program, to find bootstrap

values of F_{ST} 95% and 99% confidence intervals for the overall dataset without genotyping error, we compared the sub-dataset F_{ST} to the confidence interval of the overall dataset and compared the F_{ST} of the sub-dataset without errors to the F_{ST} of the sub-dataset with errors. This entire process after simulation was repeated 98 more times, using the additional 98 replicates from the EASYPOP program.

2.3 – RESULTS

Because different sub-datasets were created, we could determine the effect of sampling effort by comparing the F_{ST} values of the sub-datasets with the full dataset. We compared the F_{ST} values for each sub-dataset to the 95% and 99% bootstrap confidence intervals to determine if the $\hat{\theta}_{ST}$ values were significantly different from the $\overline{\theta}_{ST}$ of the original dataset.

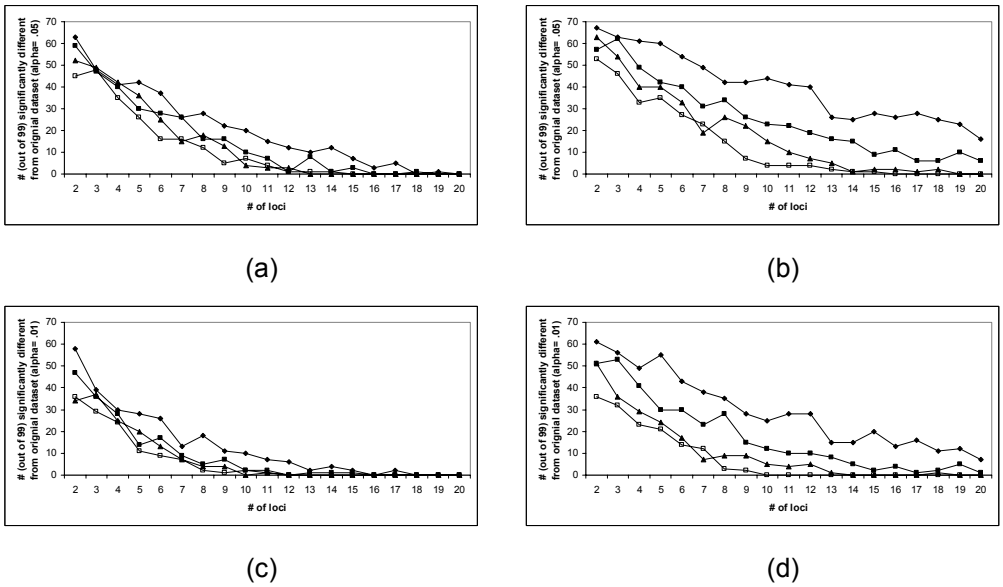


Figure 2.3 – Effect of Sample Size and Number of Loci. Graph shows number of $\hat{\theta}_{ST}$ values (estimated from sub-datasets; out of 99) that differ significantly from $\overline{\theta}_{ST}$ (calculated from the entire dataset) vs. number of loci and four different individual sampling intensities. The number of individuals, N=10 are shown with solid diamonds, N=20 solid squares, N=50 solid triangles, and N=200 open squares. Dataset A is represented in panels (a) and (c), whereas Dataset B is presented in panels (b) and (d).

The number of $\hat{\theta}_{ST}$ values out of the 99 replicates that were significantly different at the alpha = 0.05 level for Datasets A and B are pictured in Figure 2.3(a) and (b) respectively, for alpha = 0.01 level Datasets A and B are pictured in Figure 2.3 (c) and (d) respectively. In Figure 2.4 we have smoothed the lines presented in Figure 2.3.

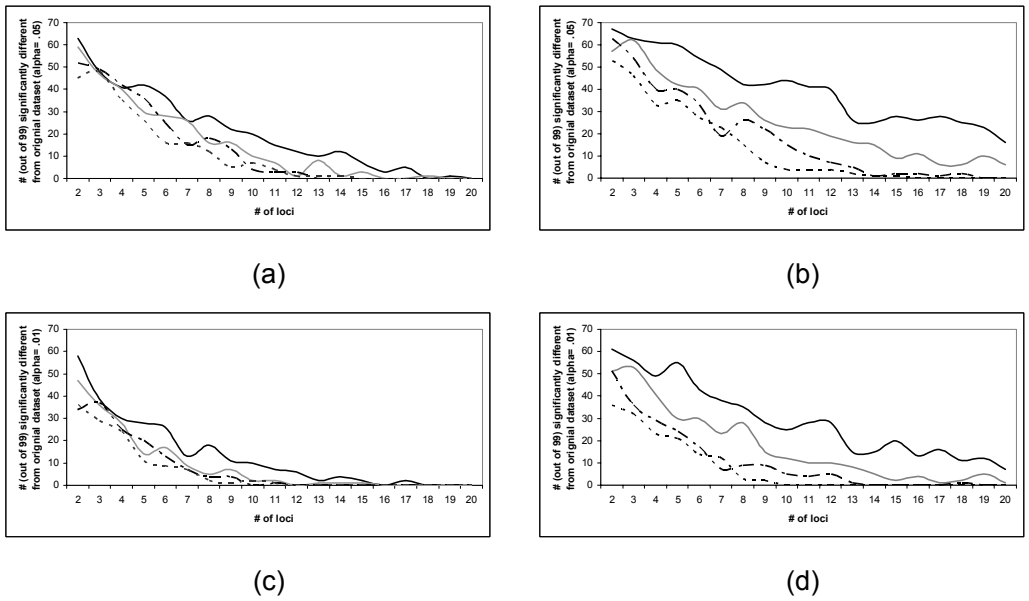


Figure 2.4 – Effect of Sample Size and Number of Loci – Smoothed. Smoothed lines representing the number of $\hat{\theta}_{ST}$ values (estimated from sub-datasets; out of 99) that differ significantly from $\overline{\theta}_{ST}$ (calculated from the entire dataset) versus number of loci and four different individual sampling intensities. N=10 individuals is represented by the solid black line, N=20 is the solid grey line, N=50 is the dashed line and N=200 is the dotted line.

From these graphs we can see that to achieve an accurate estimate of F_{ST} it is generally better to sacrifice genotyping more individuals in each population rather than the number of loci. This agrees with the findings from all previous simulation and theoretical studies, indicating that our simulations are behaving appropriately.

After incorporating allelic dropout to the sub-datasets we took the difference between the average F_{ST} with the allelic dropout errors ($\hat{\theta}_{ST\_drop}$) and the average F_{ST} without the errors

($\overline{\theta}_{ST\_sub}$). We found that no matter what the value of the randomly deleted percentage 0.01, 0.05, 0.1 or 0.5, there is no effect on the F_{ST} value due to allelic dropout; as expected, it is truly just a random error that does not influence the results. The results of the differences between the $\hat{\theta}_{ST\_drop}$ values and $\overline{\theta}_{ST\_sub}$ values recorded for Dataset A and Dataset B are displayed in Figure 2.5 (a) and (b) respectively.

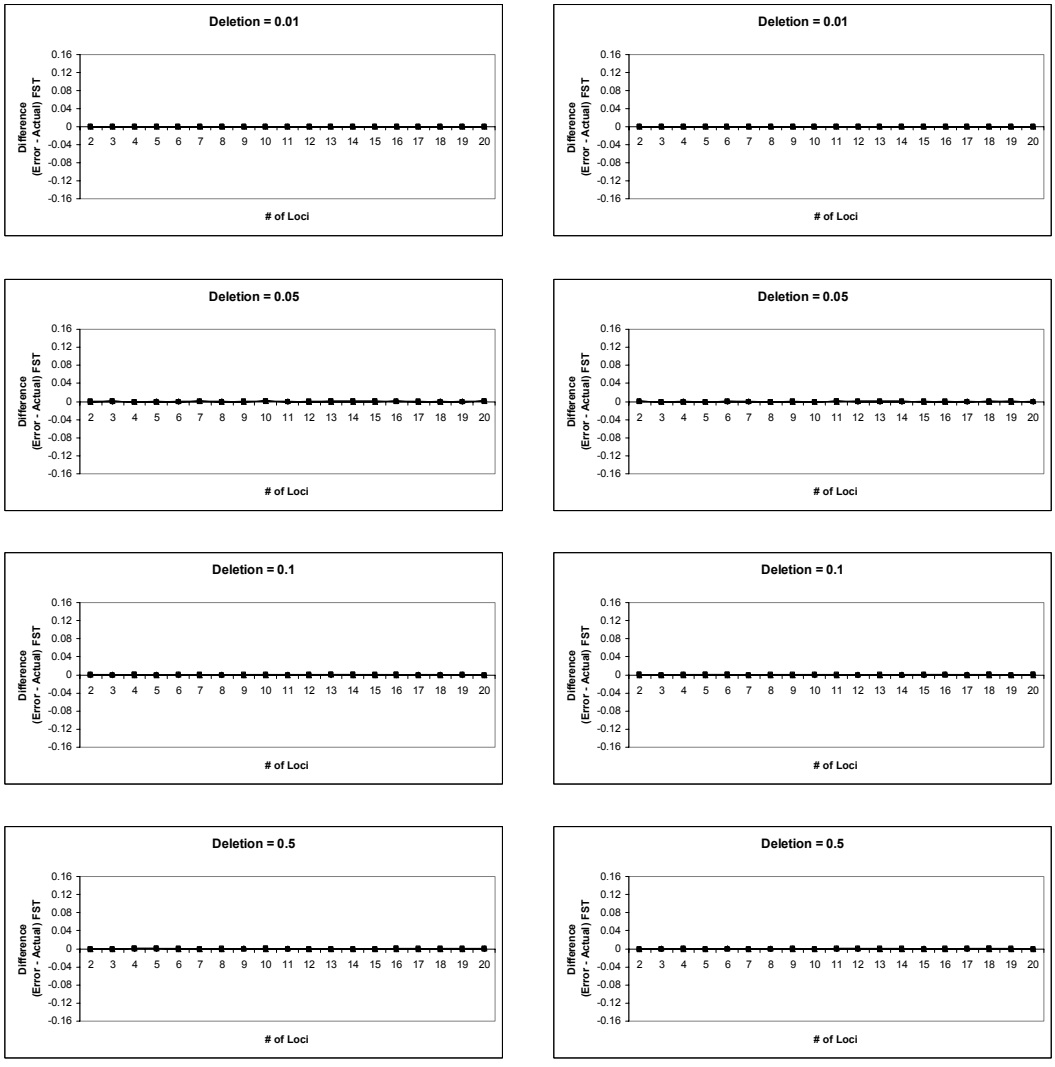
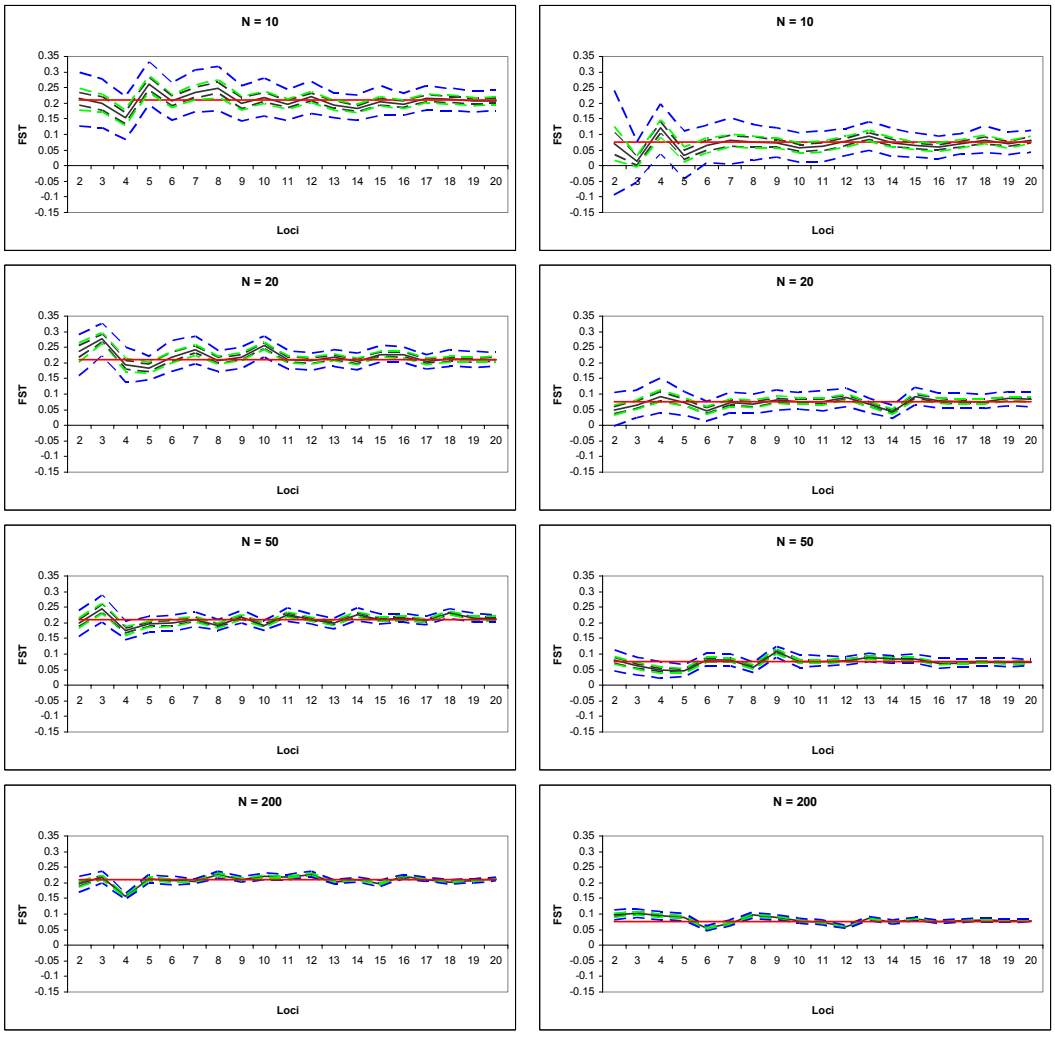


Figure 2.5 – Effect of Allelic Dropout - $\hat{\theta}_{ST\_drop}$ - $\overline{\theta}_{ST\_sub}$ versus number of loci. The number of individuals are as given in Figure 2.4. Datasets A and B are presented in panels (a) and (b) respectively.

Looking at one replication at a time, however, we see that the variance for F_{ST} with allelic dropout for one replication ($\hat{\theta}_{ST\_drop\_1rep}$) is large for smaller sample sizes (see Figure 2.6). Thus, allelic dropout errors increase stochastically in estimates of F_{ST} with decreasing sample sizes of individuals or loci.



(a) (b)

Figure 2.6 – Effect of Allelic Dropout on One Replication. Graph of the values of F_{ST} versus number of loci. $\hat{\theta}_{ST\_drop\_1rep}$ values are shown in black, 95% confidence interval for d = 0.05 in black dashed lines, 95% confidence interval for d = 0.1 in green dashed lines, 95% confidence interval for d = 0.5 in blue dashed lines, and value of $\overline{\theta}_{ST}$ for replication in red line. Dataset set A (rep. 9) and Dataset B (rep. 24) are presented in panels (a) and (b) respectively.

Alternatively, Binning Error had a significant effect on the value of F_{ST}. For example, taking the standard deviation of the allele size errors to be 1 for N=20 and L=9 (for Dataset A), resulted in an average decrease in $\hat{\theta}_{ST\_bin}$ of 0.0561. This average decrease (average over all 100 pseudo-replicated sub-datasets of 99 replicated datasets) has a standard deviation of 0.0097 and a 99% confidence interval of (0.0536, 0.0586). This shows that for N=20 and L=9 that there is a significant difference between the error $\hat{\theta}_{ST\_bin}$ and the actual $\overline{\theta}_{ST\_sub}$ for allele size standard deviation equal to 1. We have similar results for N=10, N=20, N=50, and N=200 at 2-20 number of loci. Each difference is significant at the alpha = 0.01 level.

Figure 2.7 (b) demonstrates that the difference between $\hat{\theta}_{ST\_bin}$ and $\overline{\theta}_{ST\_sub}$ for Dataset B, is less than that of Dataset A. This is because the values of $\overline{\theta}_{ST}$ in Dataset B are about one third of the values of $\overline{\theta}_{ST}$ in Dataset A. Notice that for σ=1 Dataset A has a difference of about 0.05 which is approximately 25% of the original $\overline{\theta}_{ST}$ of 0.2179 and σ=1 for Dataset B has a difference of about 0.02 which is also approximately 25% of the original $\overline{\theta}_{ST}$ of 0.0787. For σ=2 the difference is approximately 55% of the original and σ=3 is approximately 60% of the original $\overline{\theta}_{ST}$ value.

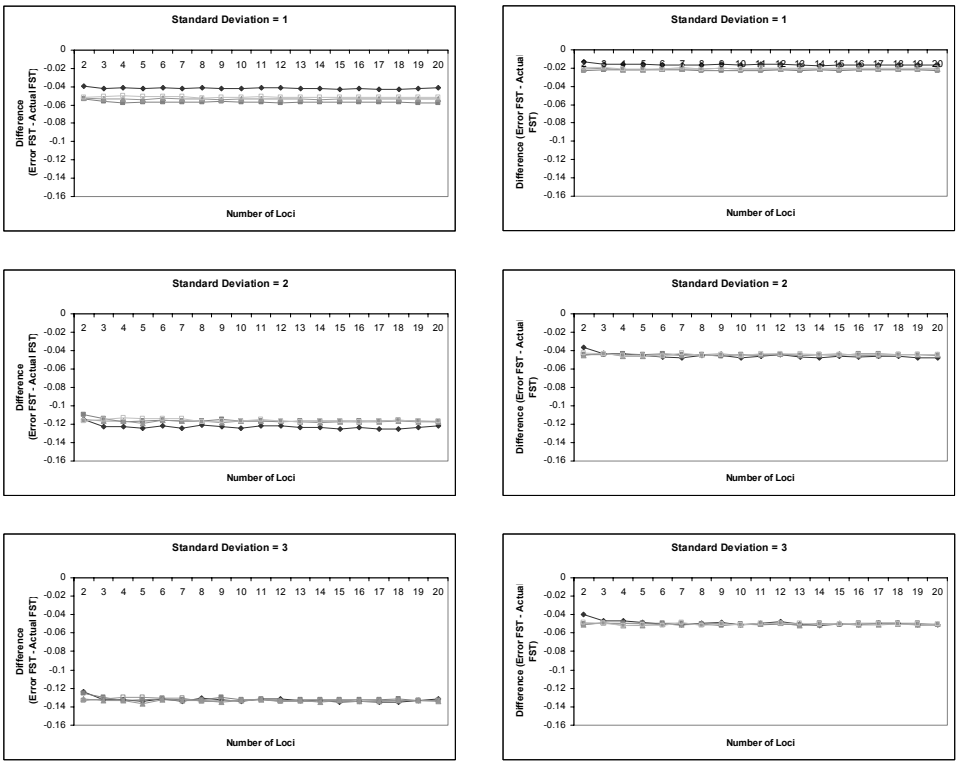


Figure 2.7 – Effect of Binning Error - $\hat{\theta}_{ST\_bin}$ - $\overline{\theta}_{ST\_sub}$ versus number of loci. The number of individuals are as given in Figure 2.4. Datasets A and B are presented in panels (a) and (b) respectively.

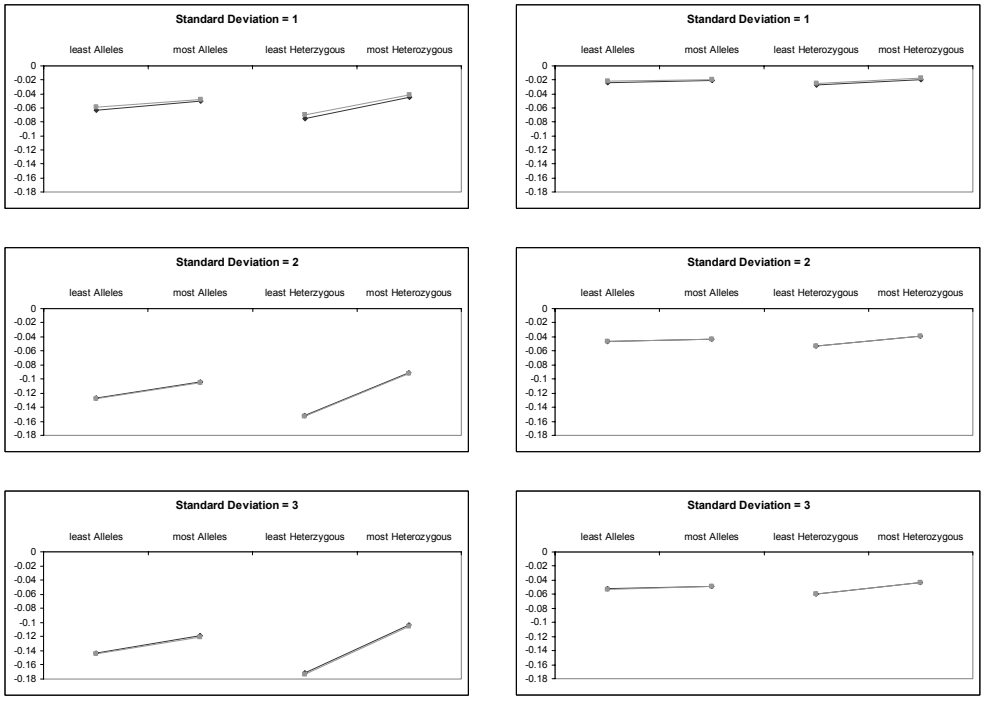
F_{ST} is a statistic that is a function of the number of alleles at each locus and the expected heterozygosity at each locus. Therefore, when applying binning errors we used the sub-datasets that have randomly chosen loci. Researchers, however, will rarely have the luxury of choosing loci at random, and indeed will likely choose loci in a biased way (e.g., choosing loci that are highly polymorphic and/or easy to score). It is therefore possible that the effect on any specific $\hat{\theta}_{ST\_bin}$ may vary significantly if the loci with the highest or lowest number of alleles were chosen or the loci with the highest or lowest expected heterozygosity were chosen. We examined the five loci within each replicate dataset with the lowest number of alleles for N=20 and N=50, then the five loci with the highest number of alleles for N=20 and N=50. Also, we looked at the five

loci with the lowest expected heterozygosity and the five loci with the highest expected heterozygosity (Table 2.3).

Table 2.3 – 5 Loci with Least/Most Alleles and Heterozygosity

	Dataset A	**Dataset B**
Least Alleles (Range)	12.6 (6 – 17)	12.0 (6 – 17)
Most Alleles (Range)	22.4 (18 – 31)	21.8 (17 – 32)
Least Heterozygous (Range)	0.5798 (0.0929 – 0.7276)	0.6651 (0.4030 – 0.8018)
Most Heterozygous (Range)	0.8236 (0.7320 – 0.8963)	0.8521 (0.7877 – 0.9134)

We can see in Dataset A that heterozygosity has a greater effect on the difference than the number of alleles (Figure 2.8).



(a) (b)

Figure 2.8 – Effect of Binning Error on Selected Loci - $\hat{\theta}_{ST\_bin}$ - $\overline{\theta}_{ST\_sub}$ sub-datasets with Most/Least Alleles and Heterozygosity. Datasets A and B are presented in panels (a) and (b) respectively.

When testing the effect of Null Alleles on F_{ST}, instead of dividing the datasets into sub-datasets we manipulated the overall dataset with N=1000 and L=20. We determined the effect of null alleles on θ_{ST} by deleting alleles from all 20 loci to achieve null allele frequencies of 0.05 to 0.50. The percentage of significantly different values from the original 99 values has a curved shape (Figure 2.9).

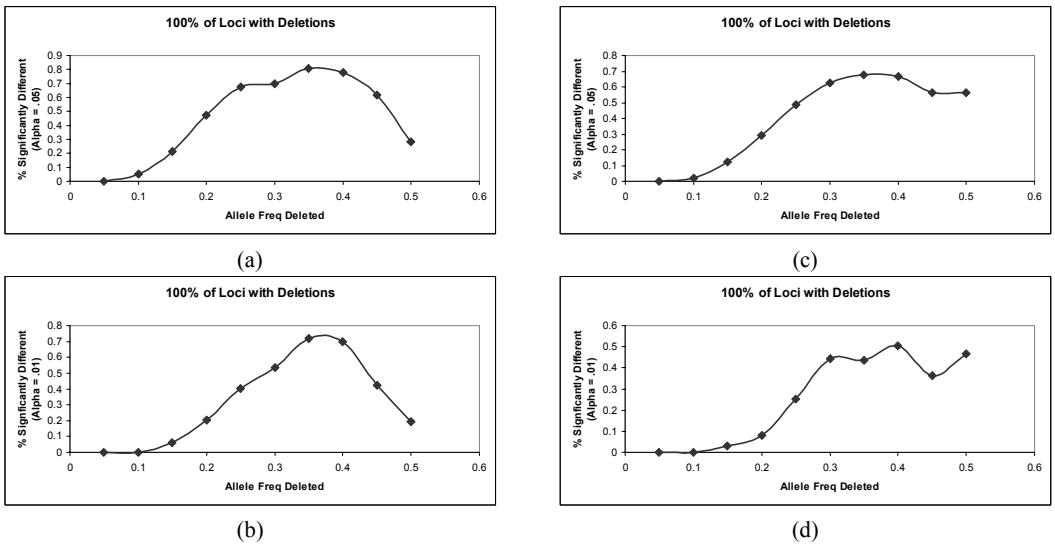
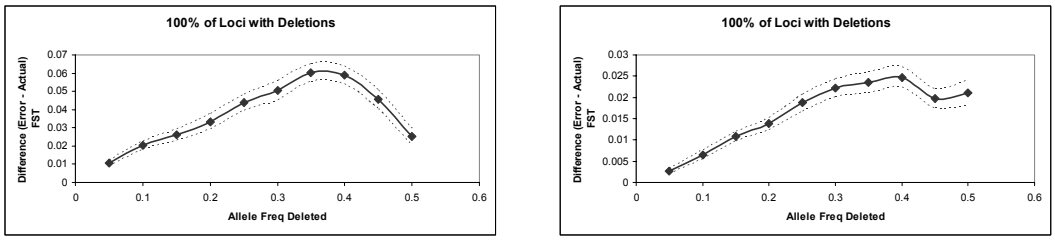


Figure 2.9 – Effect of Null Alleles (% Significantly Different) Graphs of % of Significant $\hat{\theta}_{ST\_null}$ values (out of 99) versus number of loci once Null Alleles are present. Dataset A is represented in panels (a) and (b), whereas Dataset B is presented in panels (c) and (d).

In addition, null alleles seem to increase the value of $\hat{\theta}_{ST\_null}$, which is contrary to binning error. The differences between $\overline{\theta}_{ST\_sub}$ and $\hat{\theta}_{ST\_null}$ have a similar curved shape. This suggests that there is a point when the effect becomes less as the allele frequency gets larger (Figure 2.10).



(a) (b)

Figure 2.10 – Effect of null alleles - $\hat{\theta}_{ST\_null} - \overline{\theta}_{ST}$ versus allele frequency of null allele. The dashed lines are the 95% confidence interval lines for these values. Dataset A and B are presented in panels (a) and (b) respectively.

2.4 – DISCUSSION

Overall, we found that $\hat{\theta}_{ST}$ was quite robust to even ridiculously high genotyping error rates. Allelic dropout had no effect on the average estimated F_{ST}. For binning errors and null alleles, although there were statistically significant differences in the values of $\hat{\theta}_{ST}$, the conclusions about the data (i.e., differentiation of populations) would remain the same with or without genotyping errors. We found this to be true with both large and small values of $\overline{\theta}_{ST}$, the overall population F_{ST} without genotyping errors.

Number of Individuals and Loci

If populations are strongly differentiated (i.e., $F_{ST} \geq 0.2$), then sampling ≥20 individuals at ≥10 loci provides ~90% probability of being within the 95% confidence interval of $\overline{\theta}_{ST}$. For pilot studies, however, even genotyping ≥20 individuals at as few as ≥6 loci gives ≥80% probability of being within the 99% confidence interval of $\overline{\theta}_{ST}$.

If populations are less differentiated, then more loci and individuals will be needed to achieve similar levels of confidence in the F_{ST} value obtained. For example, if populations are modestly differentiated (i.e., $F_{ST} \geq 0.08$), then sampling ≥20 individuals at ≥15 loci will provide

~90% probability of being within the 95% confidence interval of $\overline{\theta}_{ST}$. For pilot studies, however, even genotyping ≥20 individuals at as few as ≥9 loci gives a ≥80% probability of being within the 99% confidence interval of $\overline{\theta}_{ST}$. Similar to previous studies, our results indicate that to achieve an accurate estimate of F_{ST} it is generally better to sacrifice genotyping more individuals in each population rather than the number of loci. Genotyping ≥20 individuals at ≥10 loci is a reasonable initial target to determine population differentiation. If differentiation is substantial, then the estimate achieved from this effort may be sufficient. If differentiation is modest or low, then researchers will need to survey additional loci to have confidence in their estimate of F_{ST}.

Allele Dropout

We found that allelic dropout behaved as a random error and had no statistically significant effect on the average value of F_{ST}. However, researchers don't generally collect multiple replicated datasets. We found that for any particular replication there may be a significant effect on the conclusions drawn from datasets with small sample sizes of individuals and loci. That is, for dataset sets where only a small or modest number of individuals (N≤20) are genotyped at small or modest numbers of loci (N<10) and have lower values of actual F_{ST}, (such as our Dataset B) allelic dropout may cause the 95% confidence interval of sample F_{ST} to overlap with zero. This would cause a conclusion of no difference between populations when in reality there is a difference. It should be noted that we applied allelic dropout equally on both populations however, allelic dropout can occur stochastically (we have shown this in observation of one replication) and differently between populations (POMPANON *et al.* 2005). Because alleles are distributed differently among populations not having equal error rates may have a significant effect on F_{ST} and possibly the conclusions about the data as a result.

It has been found that allelic dropout may have an effect on population differentiation statistics such F_{IS} more so than F_{ST} (BJORKLUND 2005). Future work may want to explore the effect on F_{IS}.

Allele size may have an influence on amplification (SEFC *et al.* 2003). Large-sized alleles are especially vulnerable to such genotyping errors. Although this phenomena is known as large allele dropout, it would be more similar to the null allele simulations we conducted than the allelic dropout simulations.

Binning Errors (& False Alleles)

As much as 82% of the genotyping errors within previous studies are due to binning errors (AMOS *et al.* 2007). Previous studies report that estimated allele sizes are within 5% of actual allele size only 68% of the time when using automated allele calling software (AMOS *et al.* 2007). We simulated percentages of binning error that begin with realistic, but high, error percentages (i.e., d=1) and go up from there.

Binning errors seem to have the most consistently statistically significant effect on the value of F_{ST}. When the standard deviation is 1 for either dataset the reduction of F_{ST} is approximately 25% of the original F_{ST}. For standard deviation of 2 the difference is approximately 55% of the original and standard deviation of 3 is approximately 60% of the original F_{ST} value. This is true whether the alleles chosen for binning errors are randomly selected, or if we choose alleles at loci with the most/least heterozygosity or allelic diversity. Yet even with statistically significant differences in F_{ST} the conclusion about the statistical support of differentiation populations will likely stay the same. Because the 95% confidence intervals of F_{ST} with binning errors for all 99 replicates do not include 0 the conclusions for both

of the sample datasets and all 99 replications we tested will be that there is differentiation between the two species populations.

Although the conclusion of statistically supported differentiation between populations remains, it is important to note that the level of differentiation can be biased dramatically. Thus, research that depends on accurate estimates of differentiation, such as determining the conservation status of the populations, should consider the effects of binning errors. For example, datasets yielding F_{ST} values only slightly below the minimum value needed to recommend conservation actions for the populations as independent units should be viewed carefully. Resolution of binning errors in such datasets will increase the estimated F_{ST} values and thus change the management decisions about those populations. It is unlikely, however, that high quality data collected by experienced researchers will contain binning errors that will change the estimated F_{ST} values by more than ~25%.

Although tri-nucleotide data are assumed, this was done for mathematical convenience. One could easily scale the results obtained to di- or tetra-nucleotide loci. For example, if a real dataset of di-nucleotide loci had a variance of 0.67 bp, then the results of $\sigma = 1$ would apply.

Note that binning error as we implemented in our program is similar to the "false allele" genotyping error category of (HOFFMAN and AMOS 2005). False alleles observed by Hoffman & Amos (2005) were generally scoring errors due to strong stutter bands in some PCR products, and were thus, a single repeat unit above or below the real allele. False alleles in other studies often correspond to PCR contamination events and thus may contain alleles of substantially different size. Because we are using F_{ST} as our measure of genetic differentiation, the distinction between these two types of false alleles is unimportant, but the distinction would be important for R_{ST}. It should also be noted that because alleles are only binned +1 or -1 from a known

allelic state, high levels of binning error will lead to multiple alleles binned to the same incorrect state.

A possible implementation of our method would be to run the PEDANT computer program (JOHNSON and HAYDON 2007) to estimate the maximum likelihood for allelic dropout and binning errors for your datasets. Then these error probabilities can be used through our software program to determine if these errors will effect the conclusions drawn from the data.

Null Alleles

We found a curved shape for the effect of null alleles versus the allele frequency that is deleted for both of our datasets. Null allele errors increase the value of F_{ST} as much as 30% of the original F_{ST} value. That peak occurs when the allele frequency chosen for deletion is between 0.3 and 0.4. For larger allele frequencies the difference between the F_{ST} without errors and the F_{ST} with errors decreases. Dakin & Avise (2004) observed a similar curved shape for the probability of falsely excluding a parental allele when offspring should be heterozygous for that allele. This is a similar simulation to our null allele simulation. This indicates that this curve may be applicable to other statistics besides F_{ST}. We believe the curve shape in our results is because as the null frequency increases, more individuals are homozygous null (i.e., missing data) thus the less the null alleles affect the value of F_{ST}. Because F_{ST} is highly influenced by heterozygosity the alleles that have the largest allele frequency affect the F_{ST} value the least. It would be interesting to see if this curve shape is also true for other statistics used in population genetics (R_{ST}, linkage disequilibrium, distance, etc.). Another interesting approach would be to target the loci with the larger allele sizes or loci that are most heterozygous as the alleles/loci with null alleles (as we did with binning error) to see if there is any change in conclusion.

(AMOS 2006) suggested that null alleles may have significant effects on values of F_{ST} because accurate account of heterozygosity is critical. However we found even with null alleles occurring at all loci (a highly unlikely scenario, usually it would only be a few loci) although the values of F_{ST} are significantly larger the conclusion about the differentiation of our two populations will remain that there is a difference. The degree of the differentiation may increase but the conclusion is still the same.

2.5 – CONCLUSIONS

We found that although the presence of genotyping errors (specifically allelic dropout, binning error, and null alleles) may cause statistically significantly different F_{ST} values the overall conclusion about the relationship between populations will most likely not change. Therefore, although genotyping errors in microsatellite data are inevitable, the effect of these errors on population differentiation statistics seems to be limited.

This is study is only a first step toward determining the effects of genotyping errors on differentiation statistics. $\hat{\theta}_{ST}$ is only one measure of population differentiation. Many researchers use R_{ST} in addition to or instead of F_{ST}. As mentioned above estimates of F_{IS} may have varying results for smaller datasets (BJORKLUND 2005). It is important to study other statistics to see if the same conclusions apply. It will also be important to contrast effects of errors in different types of markers [e.g., microsatellite data versus single nucleotide polymorphisms, cf. WELLER *et al.* (2006)].

2.6 – REFERENCES

AMOS, W., 2006 The hidden value of missing genotypes. Molecular Biology and Evolution **23:** 1995-1996.

AMOS, W., J. I. HOFFMAN, A. FRODSHAM, L. ZHANG, S. BEST *et al.*, 2007 Automated binning of microsatellite alleles: problems and solutions. Molecular Ecology Notes **7:** 10-14.

ARAKI, H., and M. S. BLOUIN, 2005 Unbiased estimation of relative reproductive success of different groups: evaluation and correction of bias caused by parentage assignment errors. Molecular Ecology **14:** 4097-4109.

BALLOUX, F., 2001 EASYPOP (version 1.7): A computer program for population genetics simulations. Journal of Heredity **92:** 301-302.

BALLOUX, F., and J. GOUDET, 2002 Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Molecular Ecology **11:** 771-783.

BJORKLUND, M., 2005 A method for adjusting allele frequencies in the case of microsatellite allele drop-out. Molecular Ecology Notes **5:** 676-679.

BONIN, A., E. BELLEMAIN, P. B. EIDESEN, F. POMPANON, C. BROCHMANN *et al.*, 2004 How to track and assess genotyping errors in population genetics studies. Molecular Ecology **13:** 3261-3273.

BROQUET, T., and E. PETIT, 2004 Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology **13:** 3601-3608.

BRUFORD, M. W., and R. K. WAYNE, 1993 Microsatellites and their application to population genetic studies. Curr Opin Genet Dev **3:** 939-943.

DAKIN, E. E., and J. C. AVISE, 2004 Microsatellite null alleles in parentage analysis. Heredity **93:** 504-509.

GAGGIOTTI, O. E., O. LANGE, K. RASSMANN and C. GLIDDON, 1999 A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Molecular Ecology **8:** 1513-1520.

GOUDET, J., 1995 FSTAT (Version 1.2): A computer program to calculate F-statistics. Journal of Heredity **86:** 485-486.

HOFFMAN, J. I., and W. AMOS, 2005 Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology **14:** 599-612.

JOHNSON, P. C. D., and D. T. HAYDON, 2007 Maximum-likelihood estimation of allelic dropout and false allele error rates from Microsatellite genotypes in the absence of reference data. Genetics **175:** 827-842.

KALINOWSKI, S. T., M. L. TAPER and S. CREEL, 2006 Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. Conservation Genetics **7:** 319-329.

KALINOWSKI, S. T., M. L. TAPER and T. C. MARSHALL, 2007 Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology **16:** 1099-1106.

MCKELVEY, K. S., and M. K. SCHWARTZ, 2005 DROPOUT: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. Molecular Ecology Notes **5:** 716-718.

MILLER, C. R., P. JOYCE and L. P. WAITS, 2002 Assessing allelic dropout and genotype reliability using maximum likelihood. Genetics **160:** 357-366.

MORRISSEY, M. B., and A. J. WILSON, 2005 The potential costs of accounting for genotypic errors in molecular parentage analyses. Molecular Ecology **14:** 4111-4121.

NEI, M., 1973 Analysis Of Gene Diversity In Subdivided Populations. Proceedings of the National Academy of Sciences of the United States of America **70:** 3321-3323.

PAETKAU, D., 2003 An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology **12:** 1375-1387.

POMPANON, F., A. BONIN, E. BELLEMAIN and P. TABERLET, 2005 Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics **6:** 847-859.

QUELLER, D. C., J. E. STRASSMANN and C. R. HUGHES, 1993 Microsatellites And Kinship. Trends in Ecology & Evolution **8:** 285-&.

ROON, D. A., L. P. WAITS and K. C. KENDALL, 2005 A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data. Animal Conservation **8:** 203-215.

ROUSSET, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357-1362.

SEFC, K. M., R. B. PAYNE and M. D. SORENSON, 2003 Microsatellite amplification from museum feather samples: Effects of fragment size and template concentration on genotyping errors. Auk **120:** 982-989.

SELKOE, K. A., and R. J. TOONEN, 2006 Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters **9:** 615-629.

SLATKIN, M., 1995 A Measure Of Population Subdivision Based On Microsatellite Allele Frequencies. Genetics **139:** 457-462.

TAGGART, J. B., 2007 FAP: an exclusion-based parental assignment program with enhanced predictive functions. Molecular Ecology Notes **7:** 412-415.

TAUTZ, D., 1989 Hypervariability Of Simple Sequences As A General Source For Polymorphic Dna Markers. Nucleic Acids Research **17:** 6463-6471.

TAUTZ, D., and M. RENZ, 1984 Simple Sequences Are Ubiquitous Repetitive Components Of Eukaryotic Genomes. Nucleic Acids Research **12:** 4127-4138.

VANDEPUTTE, M., S. MAUGER and M. DUPONT-NIVET, 2006 An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Molecular Ecology Notes **6:** 265-267.

WAITS, L. P., and D. PAETKAU, 2005 Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. Journal of Wildlife Management **69:** 1419-1433.

WEBER, J. L., and P. E. MAY, 1989 Abundant Class Of Human Dna Polymorphisms Which Can Be Typed Using The Polymerase Chain-Reaction. American Journal of Human Genetics **44:** 388-396.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-Statistics For The Analysis Of Population-Structure. Evolution **38:** 1358-1370.

WELLER, J. I., E. SEROUSSI and M. RON, 2006 Estimation of the number of genetic markers required for individual animal identification accounting for genotyping errors. Animal Genetics **37:** 387-389.

WRIGHT, S., 1951 The Genetical Structure Of Populations. Annals of Eugenics **15:** 323-354.

WRIGHT, S., 1965 The Interpretation Of Population-Structure By F-Statistics With Special Regard To Systems Of Mating. Evolution **19:** 395-420.

CHAPTER 3

INFERRING GENOTYPIC STRUCTURE OF COMPLEX DISEASE LOCI USING GENOME-WIDE EXPRESSION DATA[2]

[2] Breazel, E.H. and P.D. Schliekelman. To be submitted to *Genetics*.

ABSTRACT

Recent studies have found that genome-wide expression data may be a useful tool in the difficult task of mapping complex traits. We have developed a method using expression level information to cluster individuals by their genotype on disease causative loci. Standard clustering methods are not well suited to identify genotypic structure because they tend to be overwhelmed by variation that is unrelated to disease genetic variation. We propose an EM algorithm-based method that targets the disease genetic variation and will thus identify disease genotypic variation via the correlation structure in differences in gene expression between disease affected and unaffected individuals. Identifying genotypic structure in a population will allow gene mapping studies to take into account heterogeneity in disease genotype and will improve mapping power.

3.1 – INTRODUCTION

Gene mapping is the process of identifying the location of genes on chromosomes. In particular gene mapping of diseases is the process of finding the location of the genes that cause the disease. Complex diseases are diseases that are influenced by multiple genes and/or environmental factors. It has been found that the mapping of genetically complex diseases such as schizophrenia, bipolar disorder, and diabetes, is a much more difficult task than it was once thought to be. However, single locus traits, such as cystic fibrosis, and Huntington's disease, have proven to be easier to map. For single locus diseases, if an individual has a certain single locus disease it is highly likely that the individual has a certain genotype at a particular locus that caused that disease as well as if that certain genotype is found at that particular locus then the individual is likely to have the disease. That is to say, single locus traits have a strong correlation

between genotype at the causative locus and the trait. In multiple locus traits it is not necessarily true that a certain genotype is present when a trait is expressed nor is it true that if the certain genotype is present that the trait will be expressed. With complex diseases, the multiple genes can interact with each other or environmental factors. Because of this, there has been little success in mapping complex trait loci. Unfortunately, a majority of genetically based diseases that affect many people are complex. Therefore, there is a great need for better ways to map these complex diseases.

Over the past several years there has been an increasing interest in combining microarray expression analysis with molecular marker data. This is a strategy known as "genetical genomics" (JANSEN and NAP 2001). The idea behind genetical genomics is that one can treat microarray gene expression levels as different quantitative phenotypic traits then use linkage or association methods to determine significant quantitative trait loci. Such QTLs are known as expression QTLS (eQTLs).

This method of genetical genomics has been successful at showing high heritability (see below) of eQTLs in yeast (Brem and Kruglyak 2005; Brem et al. 2002), plants, and humans (MONKS *et al.* 2004; MORLEY *et al.* 2004; SCHADT *et al.* 2003). These studies have been reviewed (LI and BURMEISTER 2005; STAMATOYANNOPOULOS 2004) and have shown that microarray data may be extremely useful in mapping loci associated with complex traits. There have also been studies using genetical genomics methods that found causative genes related to weight in mice (GHAZALPOUR *et al.* 2006; SCHADT *et al.* 2003).

For example, Schadt *et al.* performed an F2 cross (creation of a generation where both parental phenotypes occur) between two inbred mouse strains and performed a genome-wide scan for linkage of expression levels of 23,574 genes. These mice were placed on a high-fat diet

and therefore a wide array of obesity occurred among them. The mice were classified according to their fat pad mass (FPM) trait. The expression levels for the 25% of the mice with the highest FPM were compared to those expression levels for the 25% of the mice with the lowest FPM. 280 genes were identified to be differentially expressed between these two groups. It was found that clustering the mice on this set of genes divided the mice into two high-FPM groups (high FPM group-1 and high FPM group-2) and one-low FPM group. A genome scan was then performed for those mice classified as high FPM group-1 or low FPM group and another scan on those classified as high FPM group-2 or low FPM group. They found the association log-odd scores of eQTLs for FPM were substantially increased when only one high FPM group was included. They also showed that a number of the genes found from the reduce scan had expression levels mapping to the same region as a QTL for the FPM trait prior to reduction, just with higher correlation. The idea is if there are expression levels that are highly heritable (see below), there is evidence that several of these expression levels map to the same chromosomal region and possibly the same locus. It is possible that causative loci (loci that are responsible for variation in disease or trait) affect expression levels of other genes; these expression levels would then have high power for mapping the causative loci.

The goal of gene mapping studies is to demonstrate that a phenotypic trait is influenced by inherited factors. Heritability is the percentage of the phenotypic variation among individuals that is affected by genotypic variation (GIBSON and WEIR 2005). The papers mentioned above, that have used genetical genomics methods, show high levels of heritability of expression levels with respect to mapped QTLs. Although due to the large number of genes studied it would not be surprising to see high levels of heritability by chance. For example, Monk et al (2004) found very high QTL heritability values for their 55 significant eQTLs. However, these were selected

from 24,000 genes. While these heritability values were shown to be significant, because of the large number of genes they may not actually be as high as estimated. BREM and KRUGLYAK (2005) found a way to partially avoid this problem. They mapped QTLs for the expression levels of roughly 6,000 genes in a cross between two yeast strains. They found that 3,546 expression levels had significant heritability values. They then used half of their data as a QTL detection set and the other half of the data to estimate the proportion of variance explained by the QTLs that were detected. This gave an independent estimate of the proportion of variance explained. They found a median of 27% of variance explained by QTLs with 16% of QTLS explaining more than 60% of the variance. They also estimated bounds on the number of expression levels controlled by a given number of loci. They concluded that 3% (106) of genes were consistent with 1 locus control, 17-18% (620) were consistent with 1-2 locus control, and 50% required more than 5 loci. This still proposes a multiple testing issue because they estimated variance on 3546 genes. However, due to the fact that they found heritability levels of over 69% for all 3546 of those genes, there is little doubt that there were hundreds (over 600) of expression levels with simple (1-2 loci) inheritance and high heritability.

There is a potential problem with this genetical genomics method. A hypothetical pathway for complex genetic traits is shown in Figure 3.1. Ovals represent disease causative loci – that is, loci with genotype variation that leads to variation in disease risk. Rectangles represent genes in a network, whose transcript variation between individuals leads to variation in disease risk between individuals. Arrows between loci and genes in the network represent an impact of genotypic variation at the locus on the gene's transcript level. Arrows between genes in the network represent control of transcription. Arrows between genes and the disease represent an

impact of the genes transcription on disease probability. This figure is not meant to be taken literally. Realistically things are probably much more complicated than Figure 3.1 implies.

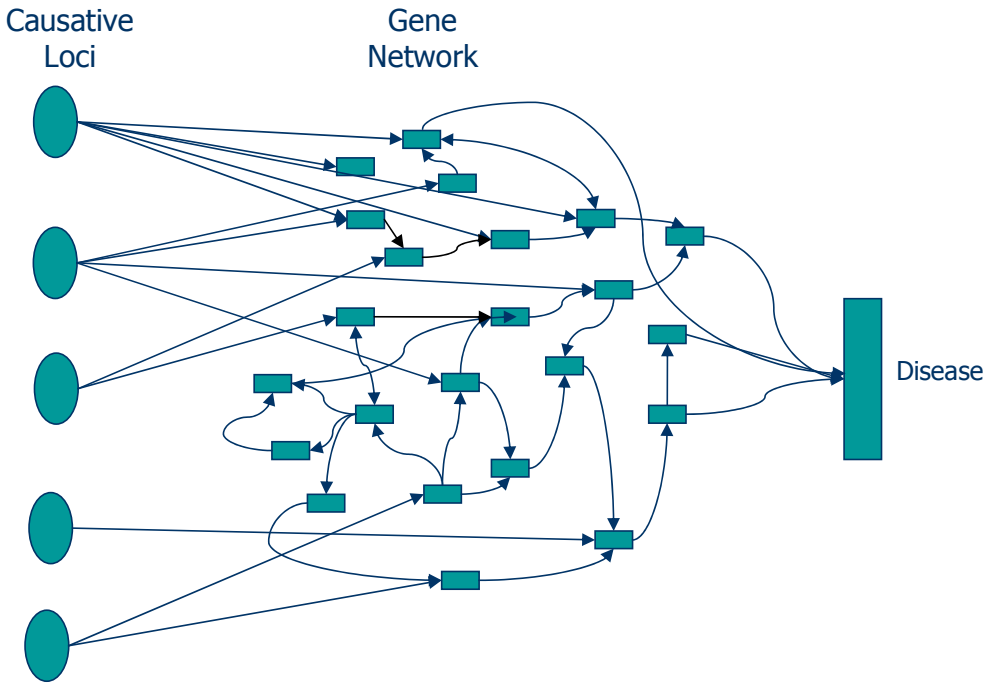


Figure 3.1 - Hypothetical Schematic of Disease Genetic Pathway

The ideal expression level for mapping a causative locus would be one whose variation is entirely determined by the genotype at a single causative locus. However, such an expression level will have poor correlation with the disease and may be difficult to detect as differentially expressed. Expression levels that are highly correlated with the disease would be easy to determine as differentially expressed, but would not be any better than the disease itself for mapping. Schadt *et al.* found a large number of eQTLs mapping to locations throughout the genome, regardless of trait status. To determine which of these eQTLs are relevant requires determining a correlation between the expression level and the trait of interests, which may be weakly correlated with causative loci. Therefore we may be exchanging the low power due to

establishing a correlation between trait and genotype to a low power due to establishing a correlation between trait and expression level. It has been shown (SCHLIEKELMAN 2008) that there is a tradeoff between the power to show an association between expression level and disease and the power to map eQTLs for those expression levels. Schliekelman also determined that power to map eQTLs under an additive penetrance model is significantly worse than with a multiplicative model.

As mentioned above, there have been several studies that use microarrays to identify causative loci, all of which have some prior knowledge of the location of the causative loci. Our goal is to develop a method that incorporates genome-wide microarray expression data as quantitative traits to map causative loci for a complex trait without any prior knowledge of locus location.

The purpose of this project is to develop methodology for integrating genome-wide expression data with gene mapping for complex traits. Our approach involves a two step process. First, a method is implemented to cluster individuals according to genotype on L causative loci. Assuming there is a finite set of loci that could affect the complex trait; the affected or unaffected individuals are clustered according to the genotype on these loci. These clusters are then used to allow traditional gene mapping approaches to take account of this new data structure. The thought is the power will be greater if when mapping the data the clusters are considered individually.

3.2 – METHODS AND RESULTS

Genetic Heterogeneity and Optimal Samples

Consider a disease caused by a single locus. If we were to conduct a case-control study testing alleles at this locus, we might get a contingency table like Table 3.1:

Table 3.1 – Sample Contingency Table: Fully Penetrant, Single Locus

	Case	Control	
dd	0	49	49
Dd	60	26	26
DD	40	0	100
	100	75	175

Only individuals with at least one D allele get the disease. All DD individuals get it, while only some Dd ones do. In this case there is a very strong contrast between case and control because all disease affected individuals have a D allele. The p-value for association for this table is $2.2x10^{-16}$.

Unfortunately, with most real traits of interest, things are not this easy. If the disease has multiple causative loci, then it is possible to exhibit the disease without having a disease genotype on a particular disease causative locus or to have a disease genotype on a particular disease locus without exhibiting the disease. This dramatically reduces the power to detect disease loci because there is not a consistent genotypic difference between cases and controls. Consider, for example, a situation where mutations at any one of multiple loci can cause the disease. In this case we might get a contingency table like Table 3.2:

Table 3.2: Sample Contingency Table: Complex Disease

	Case	Control	
dd	84	56	49
Dd	14	19	26
DD	2	0	100
	100	75	175

Out of 100 cases, only 16 are caused by this locus. The remaining cases are caused by mutations at other loci. The p-value for association in this case is 0.087 (which would be far from significant in the case of a genome scan where there would be a large multiple testing correction). There is a major loss in power because now most affected individuals are not

affected due to this locus and thus there is no longer a consistent genotypic difference at this locus between affected and unaffected. However, if we could limit our sample of cases to only those individuals with a disease genotype at the D locus, then we could greatly improve the power to detect an association between D and the disease.

The goal of this project is to use genome-wide expression information to do exactly this. We use the expression information to infer the existence of disease loci and to infer individual genotype at those loci, without using any marker genotype information. We then use these inferred genotypes to construct an optimal sample for detecting association between the inferred disease locus and the disease. If we are conducting a genome scan with many markers, we will not know in advance which (if any) markers correspond to the inferred locus. However, if there is a marker locus that is closely linked to the inferred disease locus, then the optimal sample will give improved power to detect it. Suppose that we correctly inferred the existence of the locus D in table 2 and genotypes at this locus are correctly inferred for each sample individual. Then, an improved sample for this locus is created by dropping all *dd* individuals from the cases. Thus, the table becomes

Table 3.3 – Sample Contingency Table: Improved Sample for Complex Disease

	Case	**Control**	
dd	0	56	49
Dd	14	19	26
DD	2	0	100
	16	75	175

The p-value for association from this table is 2.0×10^{-8}. The power from this "optimal" sample will be lower for markers linked to other disease loci, because we have reduced the sample size. However, for markers linked to the inferred locus, the power is greatly improved.

Determining Effect of Incorrectly Inferred Genotype on Power

When inferring the genotypic structure of individuals for each diseased locus, it is likely there will be some errors. The question is how many individuals can be incorrectly genotyped on a particular locus and still increase the power? We will focus on case-control tests because of their simplicity. However, similar principles will apply for any gene mapping method. We can find the power of a case-control study using a non-centrality parameter of the chi-squared test. According to Lachin (LACHIN 1977), the non-centrality parameter λ for an *r*x*c* chi-square test can be calculated as λ=Nτ, where N is the sample size and

$$\tau = \sum_{j}^{c} \frac{\left(\pi_j^1 - \pi_j^0\right)^2}{\pi_j^0}$$

where π_j^0 and π_j^1 are the sets of multinomial parameters specified under the null and alternative hypotheses, respectively. The null hypothesis for a case control study is that there is no association between the genotypes at the given locus and the disease status. The alternative is that there is association. For the contingency table with 2 treatment groups (disease and non-diseased) and 3 classes (genotypes *aa*, *Aa*, and *AA*), the non-centrality parameter becomes

$$\tau = \sum_{i}^{\text{disease status}} \sum_{j}^{genotype} \frac{\left(Q_i p_{ij}^1 - Q_i \alpha_j\right)^2}{Q_i \alpha_j}$$

where Q_0, Q_1 are the relative frequencies of non diseased and diseased individuals in the sample respectively, α_j is the relative frequency of genotype *j* in the sample and p_{ij}^1 is the conditional probability that the genotype is *j* given treatment *i* (disease/non-disease) under the alternative hypothesis. That is,

$$Q_0 = P\left(D_{sample} = 0\right) = \frac{f(0,0) + f(1,0) + f(2,0)}{N}$$
$$Q_1 = P\left(D_{sample} = 1\right) = \frac{f(0,1) + f(1,1) + f(2,1)}{N}$$
$f(j,i)$ = frequency of genotype j case i
$p_{ij}^1 = P\left(\text{g = j | affected status = i}\right)$ under alternative hypothesis
$\alpha_j = P\left(\text{genotype=j}\right)$

where D_{sample} is the disease status in the sample. Now the non-centrality parameter becomes,

$$\tau = \sum_{j}^{genotype} \left[\frac{\left(Q_{case}\left(P(g = j \mid case) - \alpha_j\right)^2\right)}{\alpha_j} + \frac{\left(Q_{control}\left(P(g = j \mid control) - \alpha_j\right)^2\right)}{\alpha_j} \right]$$

if we let,

$$\alpha_j = \sum_i Q_i p_{ij}^1 = \sum_i Q_i P\left(g = j \mid i\right)$$

after some manipulation the non-centrality parameter becomes

$$\tau = \sum_{j}^{genotype} \left[\begin{array}{l} \dfrac{\left(Q_{case}\left((1 - Q_{case})P(g = j \mid case) - Q_{control}P\left(g = j \mid control\right)\right)^2\right)}{\alpha_j} \\ + \dfrac{\left(Q_{control}\left((1 - Q_{control})P(g = j \mid control) - Q_{case}P\left(g = j \mid case\right)\right)^2\right)}{\alpha_j} \end{array} \right]$$

Take $P(G_o = j, G = g, D = 1)$ to be the probability that a randomly selected individual has an inferred genotype G_o of *j*, a true genotype G of *g*, and disease status *D* of affected (case).

Next, we define variables *A*, *B*, *C*, *E*, *F*, and *H* to be the probabilities of the possible errors in the inferred genotype that may occur within the contingency table and express the non-centrality parameter in terms of these variables. We can then find the values of these parameters that will still increase the power.

For simplicity, we assume that if the inferred genotype is incorrect, then it has an equal probability of being the other two genotypes. For example, if an individual has true genotype *aa* then if the inferred genotype is incorrect there is an equal chance that it is *AA* or *Aa*. Therefore, let

$$A = P\left(G = aa \text{ incorrectly identified in case}\right)$$
$$1 - A = P\left(G_O = aa \mid G = aa, D = 1\right)$$
$$\frac{A}{2} = P\left(G_O = Aa \mid G = aa, D = 1\right)$$
$$\frac{A}{2} = P\left(G_O = AA \mid G = aa, D = 1\right)$$

Similar relationships hold for other genotypes.

Table 3.4 - Probabilities of The Inferred Genotype Is Incorrect

	$G =$ '*aa*' incorrectly identified	$G =$ '*Aa*' incorrectly identified	G = '*AA*' incorrectly identified
Case	*A*	*B*	*C*
Control	*E*	*F*	*H*

Where, $P\left(G_0 = j \mid G = g, D\right)$ is the probability of the inferred genotype is equal to *j* given the true genotype is *g* and disease status.

In order to construct optimal samples for mapping a given inferred disease locus, we will eliminate all individuals with inferred genotype *aa* from among the diseased individuals and all individuals with inferred genotype *AA* from among the non-diseased individuals. Note that this is not necessarily the best possible sample. That is, depending on the form of the relationship between genotype and disease, it might be beneficial to delete other genotypes (e.g. if the disease allele is recessive then it might be beneficial to delete heterozygotes from among disease individuals). This issue will be addressed in future work.

Let

$$P\left(G_{del}=j \mid D_{del}=status\right)=\frac{P\left(G_d=j, D_{del}=status\right)}{P(D_{del}=status)}=\frac{P\left(G=j, D=1\right)-P\left(G_O=aa, G=j, D=1\right)}{P\left(D=1\right)-P\left(G_O=aa, D=1\right)}$$

be the probability that the genotype of an individual in the sample after deletion is *j* given their disease status. A similar relationship applies for unaffected individuals. We want to know what happens to the power after these deletions therefore, our non-centrality parameter becomes

$$\tau=\sum_{j}^{genotype}\left[\begin{array}{l}\dfrac{\left(Q_{case}\left(\left(1-Q_{case}\right)P(G_{del}=j \mid case)-Q_{control}P\left(G_{del}=j \mid control\right)\right)^2\right)}{\alpha_j} \\ +\dfrac{\left(Q_{control}\left(\left(1-Q_{control}\right)P(G_{del}=j \mid control)-Q_{case}P\left(G_{del}=j \mid case\right)\right)^2\right)}{\alpha_j}\end{array}\right]$$

The probability of the deleted genotypes are a function of the true genotype given the status *P(G=j|status)* and the error rate.

$$\begin{aligned}P\left(G_d=j \mid D_{del}=1\right)&=\frac{P\left(G_d=j, D_{del}=1\right)}{P\left(D_{del}=1\right)} \\ &=\frac{P\left(G=j \mid D=1\right)\left[1-P\left(G_O=aa \mid G=j, D=1\right)\right]}{1-\left[\begin{array}{l}P\left(G_O=aa \mid G=aa, D=1\right)P\left(G=aa \mid D=1\right) \\ +P\left(G_O=aa \mid G=Aa, D=1\right)P\left(G=Aa \mid D=1\right) \\ +P\left(G_O=aa \mid G=AA, D=1\right)P\left(G=AA \mid D=1\right)\end{array}\right]} \\ &=\frac{P\left(G=j \mid D=1\right)\left[1-P\left(G_O=aa \mid G=j, D=1\right)\right]}{1-\left[\left(1-A\right)P\left(G=aa \mid D=1\right)+\left(\frac{B}{2}\right)P\left(G=Aa \mid D=1\right)+\left(\frac{C}{2}\right)P\left(G=AA \mid D=1\right)\right]}\end{aligned}$$

Similarly,

$$P\left(G_d=j \mid D_{del}=0\right)=\frac{P\left(G=j \mid D=0\right)\left[1-P\left(G_O=AA \mid G=j, D=0\right)\right]}{1-\left[\left(\frac{E}{2}\right)P\left(G=aa \mid D=0\right)+\left(\frac{F}{2}\right)P\left(G=Aa \mid D=0\right)+\left(1-H\right)P\left(G=AA \mid D=0\right)\right]}$$

Note that,

$$1-P\left(G_O=AA\mid D=0\right)$$
$$=1-\left[\frac{(E)}{2}P\left(G=aa\mid D=0\right)+\left(\frac{F}{2}\right)P\left(G=Aa\mid D=0\right)+\left(1-H\right)P\left(G=AA\mid D=0\right)\right]$$

and

$$1-P\left(G_O=aa\mid D=1\right)$$
$$=1-\left[\left(1-A\right)P\left(G=aa\mid D=1\right)+\left(\frac{B}{2}\right)P\left(G=aa\mid D=1\right)+\left(\frac{C}{2}\right)P\left(G=aa\mid D=1\right)\right]$$

The probability of the true genotype given the disease status $P\left(G=g\mid status\right)$ can be set as constants using disease penetrance model and genotype probabilities. We are assuming a multiplicative model where the disease probability is $u(G)=u_1(g_1)u_2(g_2)\ldots u_L(g_L)$ for multi-locus genotype *G* and $u_j(g_j)$ is the contribution to the penetrance from the one-locus genotype *g* on locus *j*. Take *K* as the probability of having the disease within the overall population. It can be shown that $K=K_1K_2...K_L$, where $K_i=\sum_{g=1}^{3}u_i\left(g\right)P_i\left(g\right)$ would be the population disease prevalence if locus *i* were the only disease locus. $P_i(g)$ is the probability of genotype *g* at locus *i*. We can now find the genotype probabilities given the disease status from the overall population (see SCHLIEKELMAN (2008) for details):

$$P\left(G=aa\mid D=1\right)=\frac{P\left(D=1\mid G=aa\right)P\left(G=aa\right)}{P\left(D=1\right)}=\frac{u_1(g_1)P\left(g_1\right)}{K_1}$$

$$P\left(G=aa\mid D=0\right)=\frac{P\left(D=0\mid G=aa\right)P\left(G=aa\right)}{P\left(D=0\right)}=\frac{1-\left(u_1(g_1)\frac{K}{K_1}\right)P\left(g_1\right)}{1-K}$$

Default parameter values will be as in Table 3.5.

Table 3.5 - Default Parameter Values of Overall Population

Parameter	Value
K	0.01
K_1	0.01^1/L
L	9
$u_1(aa)= \pi$	0
$u_1(AA)= \delta$	1
h	0.5
$u_1(Aa)= \psi= \pi+ h(\delta- \pi)$	0.5
p	0.5995323

where p is the frequency of the disease allele A at the target locus, solved from the relationship that disease prevalence $K=K_1K_2\ldots K_L$, assuming that the locus contributes disease risk $K^{1/L}$. L is the number of causative loci, π is the contribution to disease risk for genotype aa, ψ is the contribution to disease risk for genotype Aa, and δ is the contribution to disease risk for genotype AA.

Recall that Q_i is the relative frequency of disease status i in the sample. Prior to deleting any individuals, we take Q_0 and Q_1 as equal to 0.5. However, once individuals are deleted from the sample, the values of Q_0 and Q_1 become functions of the deletion probabilities. If we delete all affected individuals with inferred genotype aa and delete all unaffected individuals with inferred genotype AA, then we have

$$Q_{case} = \frac{Freq\left(D_{del}=1\right)}{N_{del}} = \frac{N_{aff} - Freq\left(G_O = aa \mid D=1\right)}{N_{del}}$$

where $Freq\left(D_{del}=1\right)$ is the number of affected individuals in the sample after deletion. If we assume that $Freq\left(G_O = aa \mid D=1\right)$ takes its expected value, then we have

$$Q_{case} = \frac{N_{aff}\left[1 - P\left(G_O = aa \mid D=1\right)\right]}{N_{del}}$$

where N_{aff} is the number of individuals affected by the disease, N_{unaff} is the number of individuals not affected by the disease, and N_{del} is the number of individuals after deletion. Similarly,

$$Q_{control} = \frac{N_{unaff}\left[1 - P\left(G_O = AA \mid D = 0\right)\right]}{N_{del}}$$

All that is left to calculate is the value of N after the deletions.

$$\begin{aligned} N_{del} &= N - Freq\left(G_O = aa, D = 1\right) - Freq\left(G_O = AA, D = 0\right) \\ &= N_{aff} + N_{unaff} - Freq\left(G_O = aa, D = 1\right) - Freq\left(G_O = AA, D = 0\right) \end{aligned}$$

If we again assume that the counts take their expected values, we have

$$N_{del} = N_{aff}\left[1 - P\left(G_O = aa \mid D = 1\right)\right] + N_{unaff}\left[1 - P\left(G_O = AA \mid D = 0\right)\right]$$

Now we can see how changing the values of the error rates *A*, *B*, *C*, *E*, *F*, and *H* affects the power. Prior to any deletions the genotype probabilities are:

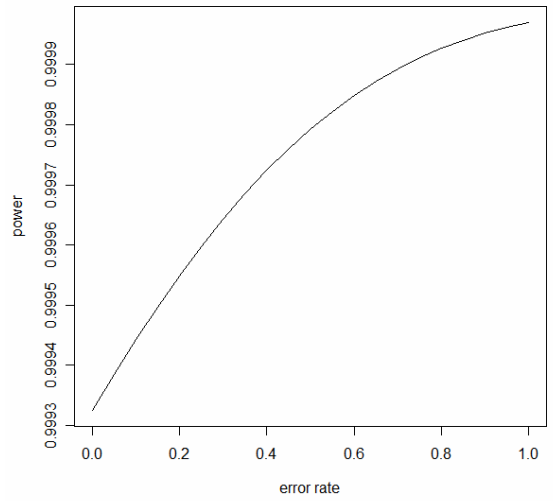
Table 3.6 – Genotype Probabilities Without Incorrectly Inferring Genotype and Prior To Deletions

	P(G=0|D)	P(G=1|D)	P(G=2|D)
Disease	0	.40	.60
Non-Disease	.16	.48	.36

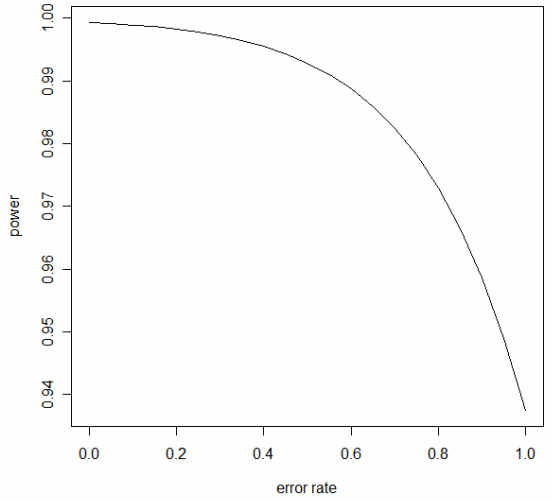
for L=9, N=200, and other parameter values as in Table 3.5. If we assume that a case-control study is being conducted with 100,000 markers such that we compare to alpha = .05/100,000, the power is 0.3014583.

Assuming that an error in the inferred genotype has a 50% chance of becoming either of the other two genotypes, the following graphs show the power with increasing values of *B*, *C*, *E*, *F*, and *H* respectively. Note that, because the $P\left(G = aa \mid D = 1\right) = 0$ for the parameters that we have chosen, then the error *A* does not occur,

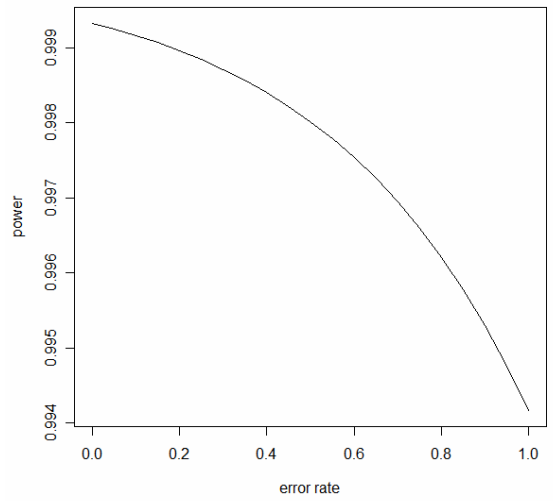
B = P(G=Aa incorrectly identified at D=1)
All other errors = 0



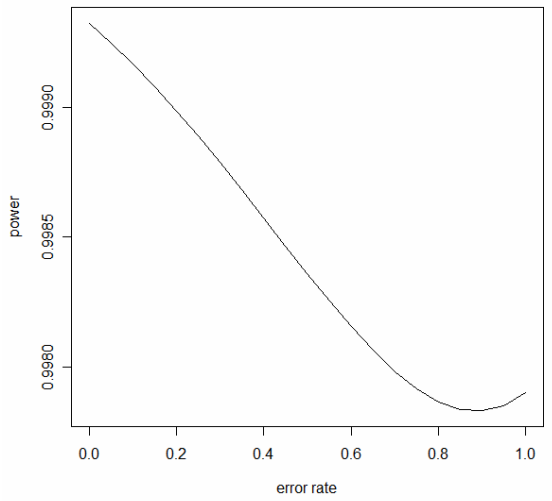
C = P(G=AA incorrectly identified at D=1)
All other errors = 0



E = P(G=aa incorrectly identified at D=0)
All other errors = 0



F = P(G=Aa incorrectly identified at D=0)
All other errors = 0



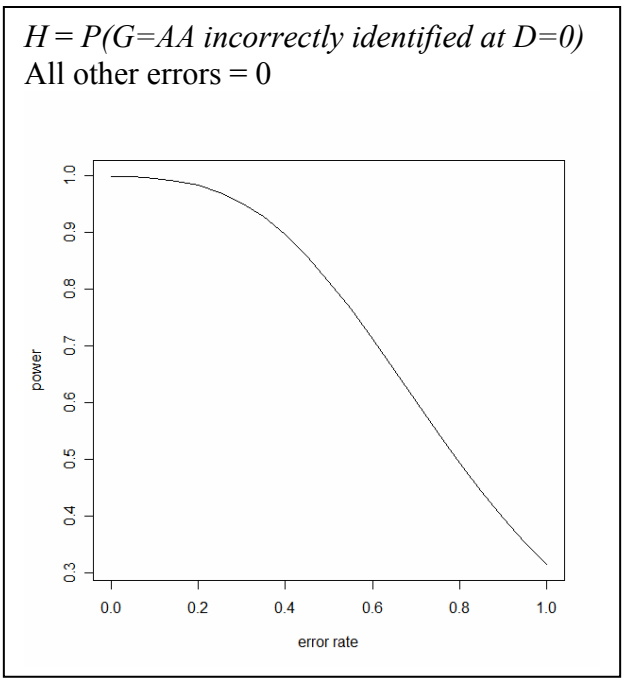


Figure 3.2 - Graphs of Error Rates of Incorrectly Inferring Genotype vs. Power

First of all, we see that power is near one when the error rate is low. That is, if we can correctly infer the genotypes, then removing genotype *AA* from among unaffected individuals increases the power from about 30% to near 100%. This shows that the idea of constructing optimal samples for mapping a locus can be very powerful if it is possible to infer genotypes well. The question then is how well this works when the genotypes are not inferred perfectly. The error rate that has the most effect on power is *H*, the probability that a genotype of *AA* is identified incorrectly for unaffected individuals. This seems logical; as *H* goes to 1 the probability of having a genotype with two disease alleles for an unaffected individual becomes larger. When there is no genotyping error, *AA* individuals will be deleted. As *H* gets larger, more *AA* individuals will remain in the dataset as incorrect genotypes. The values of *B*, *C*, *E*, and *F* do not affect the power that much. These errors each involve genotypes that should remain in the data set. If these errors occur, then half of the time there is no effect because the error is to

other genotypes (*AA* and *Aa* in affecteds and *Aa* and *aa* in unaffecteds) that remain in the dataset. The other half of the time, the error is to a genotype that is removed. In this case, the effect is to reduce the sample size. For example, if error *C* occurs then *AA* case individuals will either be incorrectly classified as *Aa* genotype and remain in the study (which has no effect) or be classified as *aa* individuals and mistakenly be removed. Errors *C*, *E*, and *F* all decrease the power as the error rates get larger, while *B* increases the power. This is an indication that it may be best to also delete disease individuals with genotype *Aa* from the case-control study.

The assumption of equal error to other genotypes will often be untrue. This is shown with simulated data in Appendix C and in the "Overall Example". Furthermore, it will not be the case that only a single type of error will occur at a time. Thus, these results are only intended as a general indication of the affect of error.

EM Algorithm for Inferring Genotypic Structure

The above results show that if we can infer genotype with good accuracy in a manner independent from the marker data, then we can use this to construct a sample with greatly improved power for detecting a specific disease locus. We next discuss how we will go about inferring the genotype. We do this using genome-wide expression data.

(SCHADT *et al.* 2003) used simple hierarchical clustering to identify the genetic structure in their data. Although this method was successful, it may not be so with a more complicated genetic structure. As an example, we generated simulated data from a model with eight genetic loci interacting multiplicatively to determine disease risk. The sample sizes were 100 disease affected and 100 unaffected individuals. The genotype at each disease locus determined the mean for four genes (for a total of 32 genes controlled by disease loci). In addition, there were 100 genes with means determined directly by disease status. Finally, there were 1000 non-disease

loci whose genotypes each determine the mean for four genes (for a total of 4000 genes). Other parameter values and assumptions are as in the example in the previous section. We applied hierarchical clustering to this simulated data set. Although it did correctly cluster the genes with shared controlling loci, these clusters were indistinguishable from the many clusters formed by non-disease controlling loci. There is little question that hierarchical clustering was successful at revealing diseased genetic structure with the data of SCHADT et al (2003). However, this structure was likely very simple, with expression level changes in many genes resulting from genotypic change at a single obesity locus. Now, it may turn out this simple structure exists for most complex traits and hierarchical clustering will be adequate. However, if the structure is often more complicated, our results show that hierarchical clustering is not adequate because the expression variation due to disease loci is overwhelmed by the variation unrelated to disease.

The goal of this project is to infer the genotypic structure of affected and unaffected individuals. Suppose we have N_{aff} affected individuals and $N_{unaff} = N\text{-}N_{aff}$ unaffected individuals. X_{ij} is the expression level for the i^{th} individual for gene j. The disease status of individual i will be denoted as Y_i, where $Y_i = 1$ for affected individuals and $Y_i = 0$ for unaffected individuals. We assume that there are a finite number, L, of disease-causative loci. We assume that each expression level is affected by only one locus. The locus could be the gene itself or another gene and could be one of the L causative loci or one of many "null loci", which we define as loci that do not affect the disease status. The variable, Z_j, is the controlling locus for gene j.

We cluster the genes by controlling locus using an approach similar to the K-means algorithm. Each cluster corresponds to a genetic locus (which may be disease causative or not). Initially, we randomly assign genes to clusters. We then use an EM algorithm (described later) to estimate the parameters for the cluster: μ_{jgk} is the mean of gene j genotype g and controlling

locus k; σ^2_{jgk} is the variance of gene j genotype g and controlling locus k, and λ_{kg} is the probability that controlling locus k has genotype g. We estimate these parameters for each locus cluster. Then, using these parameter estimates we calculate a likelihood-like quantity for each gene for membership in each cluster. Each gene is then assigned to the cluster which maximizes this likelihood. The process is continued iteratively until convergence.

Assume that the controlling locus (correct cluster) is known for every gene except the target gene r. Then, take $f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \theta_0\right)$ as the probability density for the expression data given that the controlling locus for gene r is q and given the current set of parameter estimates θ_0. The expression data are separated into $\vec{X}_r$, the vector of expression values for gene r, and $\vec{X}$ the set of expression values for all other genes.

Then we have

$$f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right) = \prod_{i=1}^{N} f\left(\vec{X}_i, X_{ir} \mid Z_r = q, \vec{\theta}_0\right)$$

, where X_{ir} is the expression value for gene r individual i and $\vec{X}_i$ is the set of other expression values for this individual. Now, take g_k to be the genotype for locus k, X_{ijk} as the expression value for the i^{th} individual of the j^{th} gene that is in cluster k, and $X_{i.k}$ as the set of expression values in individual i for all genes controlled by locus k. Then, expanding on genotype on the controlling loci indexed 1 to L, we have

$$f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right) = \prod_{i=1}^{N}\left[\sum_{g_1}\sum_{g_2}\cdots\sum_{g_L} f\left(\vec{X}_i, X_{ir} \mid Z_r = q, g_1, g_2, \ldots, g_L, \vec{\theta}_0\right) P\left(g_1, g_2, \ldots, g_L\right)\right],$$

$$= \prod_{i=1}^{N}\left[\sum_{g_1}\sum_{g_2}\cdots\sum_{g_L} f\left(X_{ir} \mid Z_r = q, \vec{g}, \vec{\theta}_0\right) f\left(X_{i.1} \mid Z_r = q, \vec{g}, \vec{\theta}_0\right) \ldots f\left(\vec{X}_{iiL} \mid Z_r = q, \vec{g}, \vec{\theta}_0\right) P\left(\vec{g}\right)\right]$$

It is worth noting that $f\left(X_{i.k} \mid Z_r = q, \vec{g}, \vec{\theta}_0\right) = f\left(X_{i.k} \mid \vec{g}, \vec{\theta}_0\right)$. That is, knowing the genotype gives full knowledge of the expression distribution and knowing the controlling locus for another gene gives no additional information. Next we take $P\left(\vec{g}\right) \approx P\left(g_1\right)P\left(g_2\right)...P\left(g_L\right)$. Note that even if this is true in the natural population, it will not be true in the sample. That is, the process of creating the sample will introduce correlations because affected individuals have more genotypic similarity than do randomly chosen ones. However, our calculations (SCHLIEKELMAN, unpublished) show these correlations will not usually be large.. Then we have,

$$f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right) = \prod_{i=1}^{N}\left[\sum_{g_1}\sum_{g_2}\cdots\sum_{g_L}\begin{pmatrix} f\left(X_{i.1} \mid g_1, \vec{\theta}_0\right)P\left(g_1\right)\cdots \\ f\left(X_{i.q} \mid Z_r = q, g_q, \vec{\theta}_0\right)P\left(g_q\right)f\left(X_{ir} \mid Z_r = q, g_q, \vec{\theta}_0\right) \\ \cdots f\left(\vec{X}_{i.L} \mid g_L, \vec{\theta}_0\right)P\left(g_L\right)\end{pmatrix}\right]$$

$$= \prod_{i=1}^{N}\left[\begin{array}{l} \sum_{g_1} f\left(X_{i.1} \mid g_1, \vec{\theta}_0\right)P\left(g_1\right)\sum_{g_2} f\left(X_{i.2} \mid g_2, \vec{\theta}_0\right)P\left(g_2\right)\cdots \\ \sum_{g_q} f\left(X_{i.q} \mid Z_r = q, g_q, \vec{\theta}_0\right)P\left(g_q\right)f\left(X_{ir} \mid Z_r = q, g_q, \vec{\theta}_0\right) \\ \cdots\sum_{g_L} f\left(\vec{X}_{i.L} \mid g_L, \vec{\theta}_0\right)P\left(g_L\right)\end{array}\right]$$

After some manipulation we get,

$$f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right)$$

$$= \prod_{i=1}^{N}\left[\frac{\prod_{k=1}^{L}\left(\sum_{g_k} f\left(X_{i.k} \mid g_k, \vec{\theta}_0\right)P\left(g_k\right)\right)}{\sum_{g_q} f\left(X_{i.q} \mid g_q, \vec{\theta}_0\right)P\left(g_q\right)}\sum_{g_q} f\left(X_{i.q} \mid Z_r = q, g_q, \vec{\theta}_0\right)f\left(X_{ir} \mid Z_r = q, g_q, \vec{\theta}_0\right)P\left(g_q\right)\right]$$

Take

$$A = \sum_{i=0}^{N}\log\left(\prod_{k=0}^{L}\left(\sum_{g=0}^{genotype}\left[\prod_{j=0}^{M_k}\left(f\left(X_{ij} \mid \mu_{jgk}, \sigma_{jgk}^{\ 2}\right)\right)\lambda_{kg}\right]\right)\right)$$

$$B = \sum_{i=0}^{N}\log\left(\sum_{g=0}^{genotype}\left[\prod_{j=0}^{M_q}\left(f\left(X_{ij} \mid \mu_{jgq}, \sigma_{jgq}^{\ 2}\right)\right)f\left(X_{ir} \mid \mu_{rgq}, \sigma_{rgq}^{\ 2}\right)\lambda_{qg}\right]\right)$$

$$C = \sum_{i=0}^{N}\log\left(\sum_{g=0}^{genotype}\left[\prod_{j=0}^{M_q}\left(f\left(X_{ij} \mid \mu_{jgq}, \sigma_{jgq}^{\ 2}\right)\right)\lambda_{qg}\right]\right)$$

where μ_{jgk} is the mean of gene *j* genotype *g* and controlling locus *k*, σ^2_{jgk} is the variance of gene *j* genotype *g* and controlling locus *k*, and λ_{kg} is the probability that controlling locus *k* has genotype *g*. Then we have $\log f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right) = A + B - C$.

Estimating the parameter

We need to estimate the parameter values for each locus cluster. We use an EM algorithm as developed in SUN and SCHLIEKELMAN (2008), applied to each cluster separately. In this study the likelihood equation is

$$L(\vec{\theta}_0 \mid \vec{X}, \vec{Y}, \vec{G}) = f(\vec{X}, \vec{Y}, \vec{G} \mid \vec{\theta}_0)$$
$$= \prod_{i=1}^{N} \left[f\left(\vec{X}_i, Y_i \mid G_i, \vec{\theta}_0\right) f\left(G_i \mid \vec{\theta}_0\right) \right]$$

For each cluster (i.e., *k*) we will consider $\vec{X}$ as the expression information for the genes within cluster *k*, and $\vec{G}$ to be the vector of genotypes of each individual for locus *k*. This likelihood equation is maximized by find estimates for μ_{ab}, the mean of gene *a* for genotype *b*, σ^2_{ab}, the variance for gene *a* genotype *b*, $P\left(G_i = g \mid \vec{\theta}_0\right)$, the probability individual *i* has genotype *g*, and $P(Y_i \mid G_i)$, the probability of disease status of individual *i* given genotype for individual *i*. The genotype $\vec{G}$ is unobserved and thus the EM algorithm is used to maximize the likelihood (see SUN and SCHLIEKELMAN (2008) for more details). The EM algorithm is performed separately on each cluster (*k*) of genes and the resulting parameters are used to calculate the value of $f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right)$. The values of μ_{ab} are used directly as the mean of gene *a* for genotype *b* for locus *k*, and σ^2_{ab} the variance for gene *a* genotype *b* for locus *k*. $P\left(G_i = g \mid \vec{\theta}_0\right)$ for each individual cluster is used to determine the value of λ_{kg}, the probability that controlling locus *k* has genotype *g*

$$\lambda_{kg} = \frac{\sum_{i=0}^{N} P\left(G_i = g \mid X_i, Y_i, \vec{\theta}_0\right)}{N}.$$

$P(Y_i \mid G_i)$ is not directly used in the calculation of $f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right)$,but will be used when inferring genotypic structure (see below). Once $f\left(\vec{X}, \vec{X}_r \mid Z_r = q, \vec{\theta}_0\right)$ for every gene in every locus cluster is calculated the cluster (*q*) with the highest likelihood for gene *r* is the new cluster for gene *r*. The EM algorithm is then performed on these new clusters separately. Again, this process is repeated until convergence (i.e., there is no difference between the previous clusters and current clusters for several iterations.)

Implementation of Algorithm

This algorithm was implemented as a C++ program. To begin the EM algorithm each parameter had initial values. For the initial locus clusters each gene was randomly assigned to one of *W* loci clusters. The initial values for μ_{abc} were chosen by sorting the expression levels for gene '*a*' and finding the quartiles. Then the initial value for $\mu_{a1.}$ was a chosen value from the zero to the first quartile, the initial value for $\mu_{a2.}$ was a chosen value between the first and third quartile, and the initial value for $\mu_{a3.}$ was a chosen value from the third quartile to *N*. The initial values for σ^2_{ac} were set to the sample variance of the expression levels for gene *a*. The initial values of $P\left(Y \mid V\right)$ were set to $P\left(Y=1 \mid V=0\right)$=0.1, $P\left(Y=1 \mid V=1\right)$=0.3, and $P\left(Y=1 \mid V=2\right)$=0.6. These are chosen as plausible values, based on a multiplicative disease model. The initial values of λ_{cb} were set to λ_{c0}=0.25, λ_{c1}=0.5, and λ_{c2}=0.25.

Avoiding local maxima

Like many clustering methods, this procedure is not guaranteed to find a globally best set of clusters and is quite dependent on the initial clusters. In order to avoid local maximums the entire process (i.e., calculating μ_{abc}, σ^2_{ac}, *P(Y|V)*, λ_{cb}, and $f\left(\bar{X}, \bar{X}_r \mid Z_r = q, \bar{\theta}_0\right)$) is performed with 30 different sets of initial cluster assignments. A log likelihood value was calculated for each using

$$\text{log likelihood} = \sum_{i=0}^{N} \log\left(\prod_{k=0}^{L}\left(\sum_{g=0}^{genotype}\left[\prod_{j=0}^{M_k}\left(f\left(X_{ij} \mid \mu_{jgk}, \sigma_{jk}^{\ 2}\right)\right)\lambda_{kg}\right]\right)\right)$$

This likelihood is the probability density for the expression data assuming that the clusters and their corresponding parameter values are correct. The cluster producing the highest likelihood value was chosen as the optimal one out of the 30. To illustrate how this method works we will do a simulated example. For our simulations we take *L*, the number of loci associated with the disease (causative loci), equal to 9; *M*, the number of genes per causative locus, equal to 6; *LND*, the number of loci not associated with disease, equal to 9; *MND*, the number of genes per non-causative locus, equal to 6; and sample size as 400 (200 affected and 200 unaffected individuals). We take *W*, the number of initial clusters, to be the correct value of 18 (see the next section). The method runs 30 times in parallel and the run with the highest likelihood is chosen as the best clusters. This iteration produces the maximum number of clusters with all genes that are controlled by the same locus clustered together. This is shown for our example in Figure 3.3.

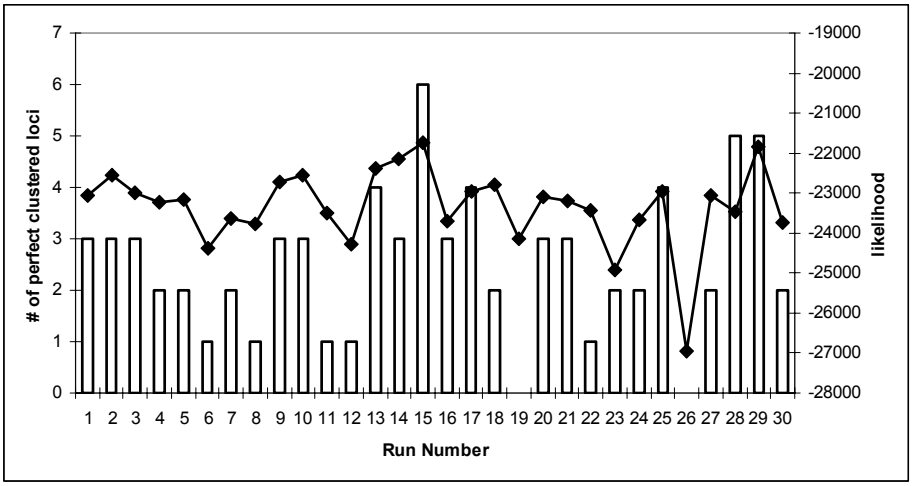


Figure 3.3 - Graph of Likelihood for Each of 30 Runs and # of Perfect Clusters

A perfectly clustered locus is defined as one with all genes controlled by the locus assigned to the cluster. The bars in the figure represent the number of perfectly clustered loci and the lines are the values of the likelihood. As seen from the figure, run 15 was chosen as the best cluster set. The clusters for this example result as follows:

Table 3.7 – Gene Clusters for L=9, M=6, LND=9, MND=6, N=400

[0]	55	64	69	73	82	91	100			
[1]	**6**	**15**	**24**	**33**						
[2]	42	66	75	84	87	93	102			
[3]	**7**	**16**	**25**	**34**	**43**	**52**	**57**			
[4]	**3**	**12**	**21**	**30**	**39**	**48**				
[5]	54	63	72	76	81	86	90	99	103	106
[6]	**0**	**9**	**18**	**27**	**36**	**45**	**78**			
[7]	**4**	**13**	**22**	**31**	**49**	**60**				
[8]	**51**									
[9]	56	59	74	83	92	101	105			
[10]	62	71	80	89	94	98	107			
[11]	**8**	**17**	**26**	**35**	**44**	**53**	**65**			
[12]	**1**	**10**	**19**	**28**	**37**	**46**	**96**			
[13]	68	77	95	104						
[14]	**2**	**11**	**20**	**29**	**38**	**47**	**85**	**88**		
[15]	**5**	**14**	**23**							
[16]	61	67	70	79	97					
[17]	**32**	**40**	**41**	**50**	**58**					

Each number in Table 3.7 represents a gene and each row is a locus cluster. The sample data in Table 3.6 and 3.7 is simulated (using an R program) so that the first *L*x*M*=54 genes are associated with the disease, with every M^{th} gene associated with the same locus. Thus, for example, genes 1, 10, 19, 28, 37, and 46 are controlled by locus 1. The next *LND*x*MND* genes

are all not associated with the disease. The bolded clusters are the clusters chosen as causative loci (See below). Notice that clusters 3, 4, 6, 11, 12, and 14 are the six clusters that have all genes controlled by the same locus clustered together. Six of the nine loci are clustered perfectly and the other three have 3-4 correctly clustered genes.

How many clusters?

Ideally, we should have one cluster for each locus that is producing significant variation in the data. Of course, we have no priori information about the number of segregating loci in the data. We can again use a log likelihood function to choose the best cluster number, *W*. We vary the number of initial clusters over a plausible range, and calculate a log-likelihood for the resulting locus clusters. . The value of *W* producing the highest log likelihood is designated as the optimal value. This log likelihood equation is the same as the one used previously:

$$\text{log likelihood} = \sum_{i=0}^{N} \log\left(\prod_{k=0}^{L} \left(\sum_{g=0}^{genotype} \left[\prod_{j=0}^{M_k} \left(P\left(X_{ij} \mid \mu^*_{jgk}, \sigma^{*2}_{jk} \right) \right) \lambda^*_{kg} \right] \right) \right)$$

When applied to the previous example the log likelihood had its maximum at *W*=18, the true value. See Figure 3.4.

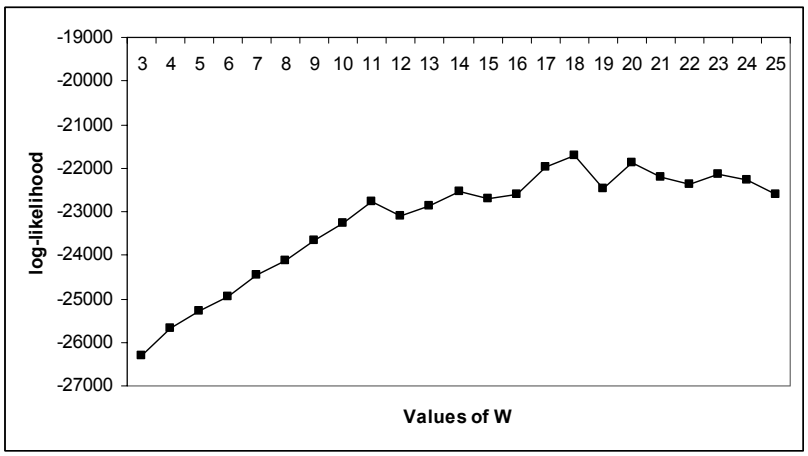


Figure 3.4 - Graph of Log-Likelihoods of Varying Values of W

From these clusters and the values of μ_{abc}, σ^2_{ac}, $P(Y|V), \lambda_{cb}$ we can infer the genotypic structure for each cluster.

Our example shows a typical outcome for our clustering method. When the percentage of disease genes is 50% or more the clustering methods seems to work very well. At 25% disease loci it works well but not on a consistent basis. This method also works well when N is smaller (i.e., 200) or larger (i.e., 600). See Figure 3.5 and Appendix B for results from other simulations.

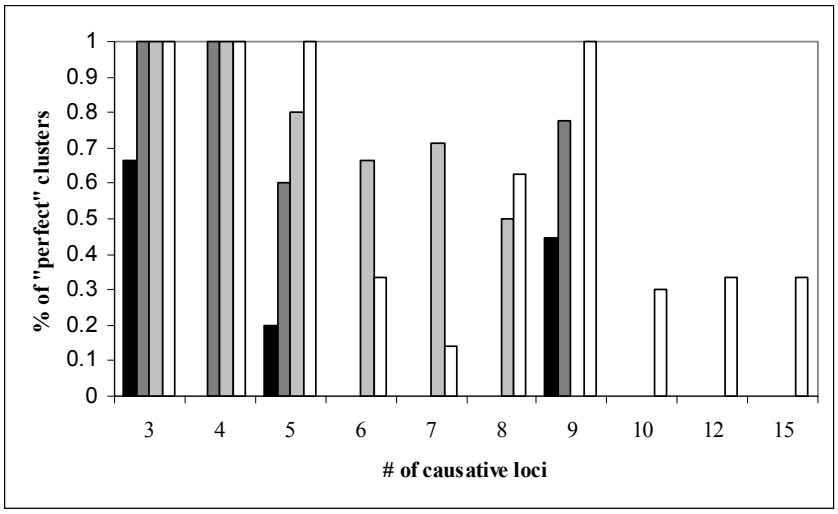


Figure 3.5 - Performance of Clustering Method.

Figure 3.5 is a graph of the # of simulated causative loci versus the percentage of "perfect" clusters produced by the clustering method. Black bars indicate datasets with 25% disease loci, dark grey is 50% disease loci, light grey is 75% disease loci and white bars are datasets with 100% disease loci present. This graph shows that the clustering method perfectly clusters at least 50% of causative loci for datasets with 50% or more disease genes present for values of *L* less than 10. These simulations are performed on sample sizes of 200, 400, or 600. It is worth noting that the clusters for L=10, 12, and 15 were calculated only for *W* = 10, 12, and 15 total gene

clusters respectively. The outcomes may have been different had the value of *W* been increased (to determine the maximum value of likelihood – see above).

It is worth noting that the purpose of this "EM clustering method" is to infer genotype for individuals on a particular locus. It may be true that other clustering methods do as well or better at clustering genes into locus clusters. However, this method allows for not only clustering genes into locus clusters but also finding probability of disease given genotype and probability of genotypes that are necessary to infer genotypic structure.

Assigning Genotypes to Individuals.

If the algorithm performs well, the gene clusters produced correspond to genetic loci. The next step is to assign individual genotypes at those loci. Define

$$\psi_{ijk} = f\left(V_{ij} = k \mid Y_i, R_{ij}\right)$$

where R_{ij} is a vector containing the values of expression levels in individual *i* that are controlled by locus *j* and V_{ij} is the genotype of individual *i* at locus *j*.. Assuming again that expression levels are independent of disease status if genotype is known we can see that

$$\psi_{ijk} = f\left(Y_i \mid V_{ij} = k\right) f\left(R_{ij} \mid V_{ij} = k\right) f\left(V_{ij} = k\right)$$

Each individual *i* is assigned the genotype *k* at locus *j* that maximizes ψ_{ijk} .The three component probabilities are calculated using the parameters estimated for the locus cluster *j*, That is,

$$f\left(Y_i \mid V_{ij} = k\right) = P\left(Y_i \mid V_{ij} = k\right)$$
$$f\left(R_{ij} \mid V_{ij} = k\right) = \prod_{r=0}^{M_j}\left[f\left(X_{ir} \mid \mu_{rkj}, \sigma_{rj}^2\right)\right]$$
$$f\left(V_{ij} = k\right) = \lambda_{jk}$$

Determining which clusters correspond to genetic loci.

Not all calculated gene locus clusters are clusters of disease genes and therefore do not all correspond to disease causative loci. In order to determine the best clusters we calculate $P\left(\vec{Y} \mid \vec{G}\right)$. That is, the probability of the disease status vector given the vector of genotypes calculated from above at each individual.

$$P\left(\vec{Y} \mid \vec{G}\right) = \prod_{i=1}^{N} \left[P\left(Y_i = 1 \mid G = g_i\right)^{Y_i} \left(1 - P\left(Y_i = 1 \mid G = g_i\right)\right)^{1-Y_i} \right]$$

The locus clusters with the highest $P\left(\vec{Y} \mid \vec{G}\right)$ are the ones that are well clustered and have the strongest association with the disease. For our example we calculate the values of $P\left(\vec{Y} \mid \vec{G}\right)$.

Table 3.8 – Values of $P\left(\vec{Y} \mid \vec{G}\right)$ For Each Locus Cluster in Descending Order

Gene Cluster	$P\left(\vec{Y} \mid \vec{G}\right)$
8	**1.38E-38**
6	**3.54E-106**
3	**6.68E-107**
12	**2.52E-107**
17	**5.51E-108**
15	**4.93E-108**
7	**5.18E-110**
1	**1.88E-110**
11	**1.93E-111**
14	**8.75E-112**
4	**4.87E-112**
5	4.39E-120
2	3.94E-120
9	1.06E-120
10	8.01E-121
13	6.54E-121
16	5.54E-121
0	4.03E-121

The locus with the highest value of $P\left(\vec{Y} \mid \vec{G}\right)$ is a cluster with only one gene that is not controlled by a disease locus. The next ten highest values were produced by disease loci. We note a substantial drop-off in the probability to the next locus (5), which is not a disease locus. We have

not devised a formal rule for selecting the disease loci. However, the following ad hoc approach works reliably: 1) eliminate any loci at the top of the list with probability much higher than the rest and 2) cut off the list at the point where the probability drops by more than 2-3 orders of magnitude. Under this approach we could eliminate locus 8 but we will continue with it in our example.

The loci chosen are bolded in Table 3.7 and Table 3.8. To show how well the method performs at genotyping we compare the genotypes of these calculated locus clusters to the true (simulated) genotypes:

Table 3.9 – Number of Individuals With Correct Genotype Out of *N*=400 for Each Locus Cluster

# of Individuals Correct	212	385	387	380	347	349	366	373	376	379	381
Gene Cluster/Locus	[8]	**[**6]	[3]	[12]	[17]	[15]	[7]	[1]	[11]	[14]	[4]

Table 3.9 shows the number of individuals out of the sample size (*N*=400) for that are correctly genotyped, for each of the chosen loci. None of the loci are genotyped perfectly, but the method does very well. We showed earlier in this paper that there is a major increase in power even with substantial error in inferred genotypes. Table 3.9 shows a typical outcome for our genotyping method. Figure 3.6 shows the performance of this genotyping method for various datasets and values of L.

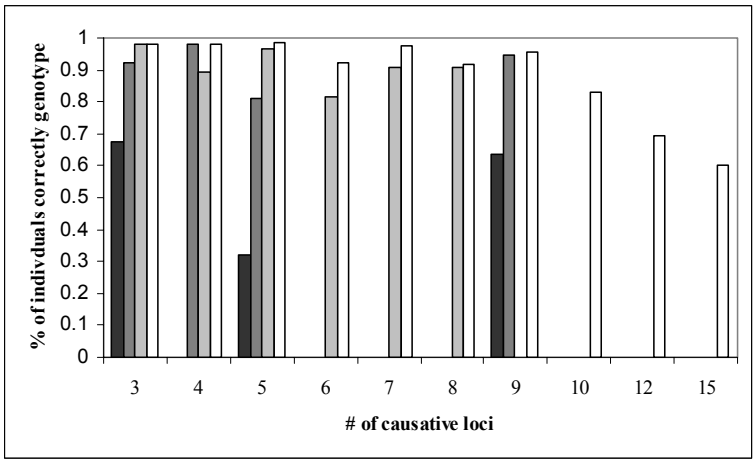


Figure 3.6 – Performance of Inferring Genotype

Figure 3.6 is a graph of the number of simulated causative loci versus the percentage of average over the correct loci of individuals correctly genotyped by inferred genotyping method. These bars represent the same simulations as those in Figure 3.5. Black bars indicate datasets with 25% disease loci, dark grey is 50% disease loci, light grey is 75% disease loci and white bars are datasets with 100% disease loci present. When comparing Figure 3.5 to Figure 3.6 it is easy to see that even though the clustering method may have resulted in only a few "perfect" clusters those that are clustered well perform well in the genotyping method. Notice, even though only a few gene clusters were "perfect" for L =10, 12, and 15, these few have at least 60% of the genotypes inferred correctly. Note that some of the individuals that are incorrectly genotyped may be deleted when performing the case-control study for association (see below).

Using the inferred genotypes

Now that we have a group of causative loci and the genotypic structure has been inferred for each causative locus, we can use this to construct optimal samples for gene mapping. We will demonstrate this with mapping via case-control test, but it will work similarly with any gene mapping method.

In a typical gene mapping study, we will have genotype data for hundreds or thousands of markers for each sample individual. We will not have any a priori information about the genomic location of our inferred loci. Therefore, we will not have information about which markers may correspond to these loci. If the method has performed well, then each inferred locus will correspond to some real locus. We will construct separate samples for each inferred disease locus and the sample will give improved power for detecting the corresponding true locus and lesser power for all other loci because of reduced sample size. If the method has performed poorly, then the inferred genotypes will not correspond to any true locus. This could occur either because the inferred locus is not a true locus (meaning the clustered genes do not share a controlling locus) or the inferred genotypes at a true locus are not accurate.

If an inferred locus corresponds to some true non-disease locus and if the error rate is substantially different between cases and controls for some genotype, then false positives could result. However, our simulation results do not show any evidence of such differences in error rate. Excepting this possibility, our approach should not produce excess false positives. If the method has performed poorly, then inferred genotypes will not correspond to any real genotypes and we will be in effect reducing sample size by removing randomly chosen individuals.

It should be noted that there will be a multiple testing correction and subsequent loss of power because we will be repeating the full genome scan for the reduced data set corresponding to each inferred disease locus. However, this loss of power will be greatly exceeded by the gain in power from the optimal data sets if the method performs well.

For our example, we performed an association test individually on our 10 chosen gene clusters. In each case the p-value of the chi-squared association test was calculated for what the case-control study would be in a true association test. For us that is the contingency table of

genotypes by disease status for the simulated genotypes. Note that in a true association study these genotypes are not known. For our example we will look at gene cluster 14. The genotypes at the corresponding simulated loci are as follows:

Table 3.10 - Simulated Genotypes for Gene Cluster 14 in Example

	G=*aa*	G=*Aa*	G=*AA*
Disease	0	83	117
Non-Disease	31	91	78

The p-value for the chi-squared test for this contingency-table is 3.124761e-09. That p-value is compared to the p-value of the chi-squared test for our genotyping method. That is, the contingency table of the inferred genotypes with the observed diseased individuals with genotype *a* and observed non-diseased individuals with genotype *AA* deleted. The non-diseased individuals that were incorrectly inferred as *aa* or *Aa* but were actually *AA* genotype were moved in the contingency table as non-diseased individuals with *AA* genotype. This is because the true power is affected by these *AA* individuals remaining in the contingency table. (See discussion about incorrect inferred genotype effect on power). To understand which individuals are moved to Non-Diseased Individuals with genotype *AA* the following table (Table 3.11) shows the errors in inferred genotypes for this gene cluster.

Table 3.11 - Incorrectly Inferred Genotypes for Example

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	4	0

Non-Disease	0	1	2
0->	0	6	0
1->	0	0	4
2->	0	5	0

In Table 3.11 only the individuals that are incorrectly inferred are present, the rows represent the true genotype, the columns represent the inferred genotype. Notice that for non-diseased individuals 5 individuals were inferred to be genotype *Aa* (or 1) that should have been genotype *AA* (or 2). These individuals will be removed from the Aa genotype and moved to the *AA*

genotype for non-diseased individuals. Our contingency table under these conditions is as follows:

Table 3.12: Contingency Table for Inferred Genotypes for Example

	G=*aa*	G=*Aa*	G=*AA*
Disease	0	85	115
Non-Disease	25	93	5

The p-value for this contingency table (Table 3.12) is 1.487951e-25. This is a big improvement on the association for this locus.

What happens if genes directly influenced by disease are included?

Most of the discussion above use examples of how the clustering method and inferring genotypic structure works for genes whose expression level is directly influence by disease loci or non-disease loci. What happens if genes whose expression levels are directly associated with the disease are included in the analysis? The clustering method and method of inferring genotypic structure is the same. As a result, the genes influenced directly with the disease are clustered with genes influenced by disease-loci (See Appendix C). This does seem to influence the inferred genotype of those causative loci. However, those clusters where genes influenced by causative loci are not clustered with genes directly associated with disease the inferred genotypes perform as before. Because some of the genes in the clusters are genes directly associated with the disease the value of $P\left(\bar{Y} \mid \bar{G}\right)$ for these clusters is 1. For example, we repeated the same example as above but added in 10 disease associated directly with the disease. We varied W from 3 to 22 and using the likelihood equation found $W = 13$ to be the largest value. We found 7 gene clusters that grouped all the genes influenced by the same causative loci together. These are clusters 0, 1, 3,6, 8, 9, and 11 in Table 3.13.

Table 3.13 - Gene Clusters for *L*=9, *M*=6, *LND*=9, *MND*=6, *N*=400 and 10 Genes Influenced Directly by Disease

[0]	**6**	**15**	**24**	**33**	**42**	**51**				
[1]	**0**	**9**	**18**	**27**	**36**	**45**	**67**	**70**	**114**	
[2]	**7**	**43**	61	62	**79**	**88**	**97**	**106**	**115**	
[3]	**5**	**14**	**23**	**32**	**41**	**50**	**72**	**103**	**117**	
[4]	37	71	80	81	89	90	98	102	107	108	112
	116									
[5]	**1**	**19**	**28**	**46**	**73**	**91**	**94**	**100**	**109**	
[6]	**4**	**13**	**22**	**31**	**40**	**49**	**69**	**105**		
[7]	65	74	76	78	83	92	101	110		
[8]	**2**	**11**	**20**	**29**	**38**	**47**	**96**			
[9]	**3**	**12**	**21**	**30**	**39**	**48**	**111**			
[10]	10	52	68	77	84	86	87	95	104	113	
[11]	**8**	**17**	**25**	**26**	**35**	**44**	**53**	**64**	**82**	**99**	
[12]	**16**	**34**	54	55	56	57	58	59	60	63	**66**
	75	**85**	**93**							

Again, each row represents a gene cluster and each number is a gene. The genes were simulated as before where the first *LxM* (0-53) genes are associated with the disease loci, with every M^{th} gene influenced by the same loci. The next 10 genes (54-63) are influenced directly by disease. The next *LNDxMND* (64-107) genes are influenced by non-disease loci with every *MND* gene influenced by the same loci. The clusters in bold are the clusters that are taken to be influenced by causative loci, by the values of $P\left(\vec{Y} \mid \vec{G}\right)$. Those values for this example see Table 3.14.

Table 3.14 - Values of $P\left(\vec{Y} \mid \vec{G}\right)$ for *L*=9, *M*=6, *LND*=9, *MND*=6, *N*=400 and 10 Genes Influenced Directly by Disease

Gene Cluster	$P\left(\vec{Y} \mid \vec{G}\right)$
Locus 2	**1**
Locus 12	**1**
Locus 11	**1.23E-103**
Locus 1	**5.22E-106**
Locus 8	**2.13E-107**
Locus 0	**3.23E-108**
Locus 6	**1.46E-108**
Locus 3	**8.40E-110**
Locus 9	**8.13E-112**
Locus 5	**3.32E-113**
Locus 4	3.47E-120
Locus 10	4.62E-121
Locus 7	4.35E-121

The clusters with the value of $P\left(\vec{Y} \mid \vec{G}\right)$ equal to 1 are chosen as causative loci. We chose the gene clusters that have the largest values of $P\left(\vec{Y} \mid \vec{G}\right)$ until what appears to be a reasonable cut off. The number of individuals, out of 400, that are genotyped correctly in these 10 clusters are shown in the table below.

Table 3.15 - Number of Individuals Genotyped Correctly When Genes Directly Influenced By Disease Are Included

# of Individuals Correct	**126**	**118**	**372**	**383**	**379**	**387**	**380**	**385**	**384**	**203**
Gene Cluster/Locus	**[2]**	**[12]**	**[11]**	**[1]**	**[8]**	**[0]**	**[6]**	**[3]**	**[9]**	**[5]**

The genotyping method performs not as well when genes influenced directly by disease are included but exactly as good, as previous example, for clusters where these genes are not included.

3.3 – Overall Example

To illustrate the entire process from microarray data to contingency table we will show an overall example. This example would be a typical experiment using our methods. The data we use is a combination of simulation data (created from an R program exactly as above examples) and real prostate cancer data. The cancer data was obtained from the National Center for Biotechnology Information (NCBI) database (Yu *et al.* 2004). Series GSE6919 is expression data from both individuals with and without prostate cancer. This data has 171 individuals with expression levels at 12625 genes. There were 25 individuals that had some missing data so those individuals were deleted from the dataset leaving 146 individuals. This dataset is used to create a background set of genes with realistic correlation structure.

Using our R program we simulated 400 individuals (200 disease affected and 200 unaffected) each with 54 genes from 9 loci (L) with 6 genes associated with each locus (M). These genes are treated as genes influenced by disease loci. For each one of these 400 individuals we randomly chose (with replacement) an individual from the prostate cancer data and the 12625 expression levels were added to the 54 for genes for that individual. The individuals from the real data set were chosen without regard to their prostate cancer status and disease status was determined randomly from the simulated genotype. The genes from the prostate cancer data set were used only to create a realistic background set of genes. A small amount of random noise was added to the genes from the prostate cancer dataset so that there would not be multiple individuals with exactly the same expression values, which was found to cause problems in some cases.

From these 12679 genes we must reduce the number of genes a much smaller set of those genes most likely controlled by disease loci. We perform a t-test for each gene between the 200 affected and 200 unaffected individuals. Using the Benjamini method (BENJAMINI and HOCHBERG 1995) for false discovery rate we created a set of genes with 50% FDR. We chose 50% because we know from simulation that our method works well with 50% genes influenced by disease loci and 50% genes influenced by non-disease loci. As a result the following genes were selected to be used in our dataset. These genes are listed in order of selection.

Table 3.16 - Genes Selected for Overall Example, False Discovery Rate = 50%

37, 32, 27, 38, 19, 2, 4, 26, 9, 36, 18, 16, 21, 30, 3, 47, 46, 28, 12, 39, 43, 51, 10, 14, 29, 49, 23, 48, 8, 20, 41, 15, 0, 35, 44, 7, 13, 50, 52, 22, 5, 17, 34, 11, 1, 11816, 45, 31, 24, 33, 4754, 3887, 5973, 653, 3383, 42, 2375, 9798, 1295, 6, 2755, 7030, 53, 9012, 9950, 1065, 2029, 3419, 2281, 10034, 7883, 1697, 5573, 8709, 4156, 7639

Recall that the first *L*x*M* (0-53) genes are from the simulated program. These are the genes that are controlled by disease loci. The other 12625 were from the prostate cancer dataset. All but two (25 and 40) of the 54 simulated genes are selected and the prostate cancer genes comprised only 31.6% of the data.

Because the prostate cancer data is rather variable we excluded any outliers from those genes and replaced those values with the lower or upper bound (depending on direction of outlier). Gene 2375 was deleted from the dataset because over 10% of the data was outliers. We also took the natural log of the remaining prostate cancer data.

As stated before our clustering method may not be the best possible clustering method. Its main purpose is to calculate the values needed to infer genotype. Because this is true prior to implementing our clustering method we clustered the genes using k-means. This will help eliminate any extremely wrong clusters that may result in an underflow problem with the program (clusters with extremely low probabilities of being clustered together can cause all values to go to zero). The clusters from the k-means are used as initial clusters in our clustering method rather than randomly assigning genes to a cluster. Now we can use these clusters and the reduced dataset as the input into our clustering method and inference of genotypic structure.

We ran the clustering method for values of W= 3 to 25. The maximum value of the log equation determines the best value for W as 22. (See Figure 3.7)

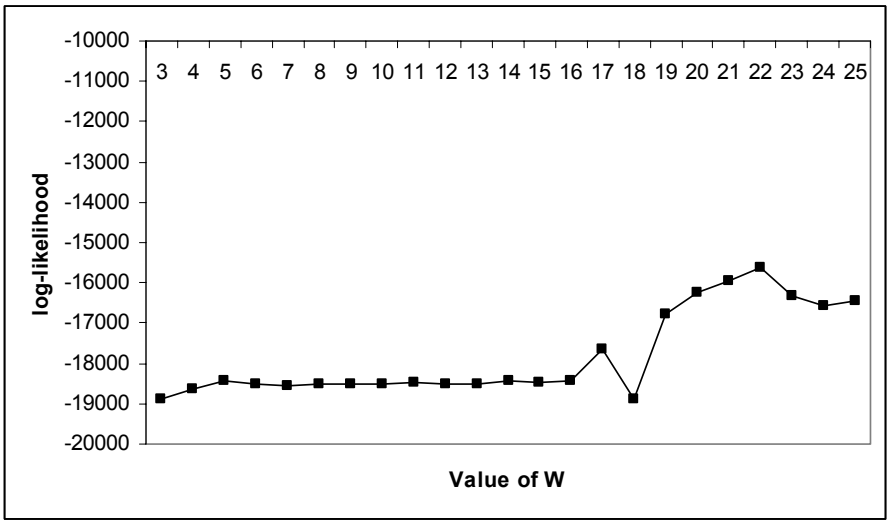


Figure 3.7 – Graph of Log-Likelihood of Varying Values of *W* for Overall Example

The genes are clusters as follows:

Table 3.17 – Gene Clusters for Overall Example

[0]	5573					
[1]	**13**					
[2]	**1**	**10**	**19**	**28**	**37**	**46**
[3]	**6**	**15**	**33**	**42**	**51**	
[4]	9798					
[5]	1697	3419				
[6]	**4754**	**7030**				
[7]	**0**	**9**	**18**	**27**	**36**	**45**
[8]	**5**	**14**	**23**	**32**	**41**	**50**
[9]	3383					
[10]	**2**	**11**	**20**	**29**	**38**	**47**
[11]	**8**	**35**	**44**			
[12]	24					
[13]	5973	9012	8709			
[14]	**7**	**16**	**34**	**43**	**52**	
[15]	2029					
[16]	**1295**	**2281**	**3887**	**7639**		
[17]	653					
[18]	**3**	**12**	**21**	**30**	**39**	**48**
[19]	**4**	**22**	**31**	**49**		
[20]	1065					
[21]	**17**	**26**	**53**			

From simulation we know that gene clusters 2, 7, 8, 10, 14, and 18 are all clustered perfectly. Cluster 14 only has 5 genes because gene 25 was not included in the dataset (due to FDR

selection). However, if this information is not known we look at the values of $P\left(\vec{Y} \mid \vec{G}\right)$ (See Table 3.18). Since we do not have a concrete way of selecting these gene clusters we will select the first several until a reasonable cut-off. The gene clusters selected are bolded in Table 3.17 and Table 3.18. Notice that gene clusters 16 and 6 are not clusters of genes associated with causative loci. Those gene clusters that are influenced by causative loci are compared to the genotypes of the simulated data to determine how many individuals (out of 400) our method correctly inferred. (See Table 3.19)

Table 3.18 – Values of $P\left(\vec{Y} \mid \vec{G}\right)$ for Overall Example

Gene Cluster	$P\left(\vec{Y} \mid \vec{G}\right)$
1	**3.54E-35**
16	**6.00E-101**
2	**9.79E-105**
11	**7.29E-107**
18	**4.18E-107**
10	**2.45E-107**
7	**1.55E-107**
3	**1.54E-107**
8	**1.03E-108**
19	**4.39E-109**
21	**3.02E-110**
14	**6.64E-111**
6	**2.11E-111**
12	3.21E-115
5	4.48E-116
17	7.06E-117
13	5.13E-117
20	4.81E-118
9	3.70E-118
0	2.69E-119
4	6.25E-120
15	8.54E-121

Table 3.19 – Number of Individuals Genotyped Correctly for Overall Example

# of Individuals Correct	**222**	**384**	**343**	**383**	**389**	**384**	**380**	**382**	**366**	**350**	**362**
Gene Cluster/Locus	**[1]**	**[2]**	**[11]**	**[18]**	**[10]**	**[7]**	**[3]**	**[8]**	**[19]**	**[21]**	**[14]**

Now we can compare the contingency tables of the simulated data (what the data would be in regular association study without our method) and the inferred data with deletions (see Methods). All 11 gene clusters associated with causative loci showed drastically smaller p-values for the chi-squared tests.

Table 3.20 – P-values for Chi-Squared Test for Association (without method vs. with method)

Locus	**without method**	**with method**
1	1.593644×10^{-9}	1.542096×10^{-54}
2	3.222184×10^{-13}	1.507401×10^{-34}
11	1.412036×10^{-9}	4.512904×10^{-21}
18	2.214003×10^{-11}	2.33019×10^{-30}
10	5.113686×10^{-21}	$6.4445291 \times 10^{-31}$
7	9.114662×10^{-13}	4.686652×10^{-30}
3	2.722611×10^{-10}	4.763017×10^{-21}
8	1.871605×10^{-10}	9.130038×10^{-27}
19	1.593644×10^{-9}	7.772012×10^{-27}
21	1.412036×10^{-9}	4.454229×10^{-26}
14	9.956697×10^{-11}	2.49172×10^{-30}

3.4 – DISCUSSION

The purpose of this method is to use gene expression information to identify heterogeneity in disease-causative genotypes and allow genome scans to take account of that heterogeneity. We do this by clustering genome-wide expression data into possible genetic causative loci clusters whose inferred genotype will determine which individuals should be deleted from the association study in order to increase the power of the association. Our procedure is a multi-step process: 1) Use a t-test with a liberal FDR to identify genes with expression patterns most strongly associated with the disease; 2) Cluster those genes to identify

groups that share a putative controlling factor assumed to be a genetic locus (either by method presented here or some other clustering method); 3) Use the method of SUN and SCHLIEKLEMAN (2008) to estimate the parameters of each locus cluster, including genotype probabilities, genotype-specific disease probabilities, and genotype-specific expression means and variances; 4) Using these parameter fits and associated models, assign genotype at each inferred locus for each sample individual; 5) Assuming that these genotype assignments are right, rank the inferred loci by their ability to predict disease status and select a set of putative disease loci from this list; 6) For each such inferred locus, construct an optimal sample by removing all disease-affected individuals whose inferred genotype has no disease alleles and removing all unaffected individuals whose inferred genotype has two disease alleles. Conduct a full genome scan separately for each such data set.

If the method has performed well and the genotypes at an inferred disease locus match well with genotype at some true disease locus, then the corresponding optimal sample will give greatly improved power for detecting that disease locus. Our simulations show that this procedure gives excellent results under many circumstances.

We performed our method on a variety of datasets. If only genes associated with the disease locus are considered in the dataset the method perfectly clusters all genes almost 100% of the time, this true for values of N from 200 to 600 for values of L ranging from 3 to 9. These perfect clusters also genotype well, having >90% of the individuals have correctly inferred genotype. When datasets contain 50% and 75% genes influenced by disease loci (50% and 25% genes influenced by non-disease genes) the clustering method and the genotyping method continue to result in 65% of causative loci perfectly clustered. These clusters result in around 80% of their individuals having correctly inferred genotypes. The methods do not work as

consistently or as well when there are only 25% of dataset consisting of genes influenced by causative loci. Both the clustering and the genotyping methods perform well for lower values of *L*. It would be interesting to explore these methods for values beyond 9.

If genes that are directly associated with the disease are included in the dataset the methods still produce useful information for a handful of loci but are not as successful at inferring genotype. This is mostly due to genes directly associated with disease being clustered with genes influenced by disease-loci.

Our method produces information for a large number of loci for various values of *L* causative loci, with 50%, 75%, and 100% of the data influenced by disease-loci, and even if genes directly influenced by the disease are included.

Traditionally genome-wide studies have used linkage analysis or association studies. Linkage analysis requires large samples of family data and has been shown to have low power in identifying common variants that may contribute greatly to common complex diseases (LOHMUELLER *ET AL.* 2003). The results in this paper from our method were obtained with sample sizes of 200, 400 or 600. Although these sample sizes are rather large relative to previous genetical genomics studies, they are reasonable sample sizes for a genome scan, and quite low relative to sample sizes required for linkage studies to give good power for a disease with the level of genetic complexity that we have assumed here.

It has been shown that association studies have much better power than linkage studies to detect common variants in a genome-wide scan when the number of alleles is small (HIRSCHHORN and DALY 2005). We use expression data to assign genotype at inferred loci and then rank these inferred loci by their ability to predict disease risk. Why should this procedure have better power to detect association to disease than does the association study? That is, in

order for our method to work we must be able to tell which inferred loci are disease loci. Why does a method that infers genotype indirectly via expression data do better than an association study that looks at genotype directly? There are several reasons. First of all, in the stage at which we do the ranking, we are not actually doing a hypothesis test. Thus, while it is required that some of the disease loci fall near the top of the list, it is not required they meet any threshold for significance. Second, our method infers genotype of the actual locus - not a marker linked to that locus. In many cases, this will greatly increase the power because there is not recombination between marker and locus. Of course, we are trading uncertainty resulting from recombination for uncertainty resulting from variance in gene expression. However, results of SUN and SCHLIEKELMAN (2008) show that the effect of this variation is greatly reduced when there are multiple transcripts controlled by a locus. Of course, the exact balance between the effects of expression variation and the effect of recombination depends on the specific genes and the specific markers. The third major benefit results from the fact that power in association studies can be greatly reduced when more than one disease causing allele is present (SLAGER *et al.* 2000). Our method effectively reduces all alleles into two types; those that increase the expression and those that decrease the expression. Therefore, regardless of the actual number of alleles present the power to detect association between inferred loci and disease will be similar to that of an association study with two alleles.

Our method will work with any method for doing genome scans, whether association or linkage based. Recent successes in genome-wide association studies in humans (DUERR *et al.* 2006; HUNTER *et al.* 2007; SLADEK *et al.* 2007; YEAGER *et al.* 2007) are dependent on the haplotype block structure found in humans and on the existence of the Human Haplotype Map (ALTSHULER *et al.* 2005). The approach developed here can be adapted for genome scans in any

species, and may be particularly useful in species where a haplotype block structure does not exist and multiple alleles will be a major problem for association mapping.

This method depends fundamentally on our assumption that the basic structure of the relationship between genetic loci, gene expression levels, and disease is similar to that depicted in Figure 3.1. Most importantly, genotype at some disease loci must affect the expression of multiple other genes. Furthermore, these genes must be reasonably close to the loci in the genetic pathway and not far downstream where the statistical link will be small. Our simulations showed that genes that respond directly to disease (as opposed to disease loci) tend to cause problems in the assignment of genotypes. This is because these genes will tend to be clustered with some disease locus, but have little correlation with that locus. It may be beneficial to simply eliminate genes with very high correlation to the trait because it is not likely that they will give much information about any disease locus.

Our simulations have assumed a multiplicative disease model. Although such a model is consistent with family history data for many human diseases (RISCH 1990; SCHLIEKELMAN and SLATKIN 2002), it is completely unknown how well it represents the true relationship between disease genotype for any disease. Furthermore, the simulations depend on many assumptions about parameter values and specific details of the relationship between genotype, expression, and disease. Thus, it is difficult to say how the performance of our method in simulations will compare with the performance with real data. See SCHLIEKELMAN (2008) for further discussion of these issues.

Currently, we do not have well-developed method for choosing which inferred loci should be considered as disease causative. We have found that an experienced person can pick the correct disease loci with near perfect accuracy by looking at the ranked list of disease

prediction probabilities (e.g. Table 3.8). This suggests that an objective rule is possible, but we have not determined what it is. Our current method is an ad hoc one of simply choosing by sight. We then conduct a separate genome scan for each inferred disease locus and apply a Bonferonni correction to the case-control p-values to account for the multiple testing. There will be a loss of power resulting from this multiple testing. If the genotypes at even one inferred disease locus match a true disease locus well, this loss in power will be far outweighed by the gain in power from the optimal sample. However, if no true disease loci are matched well, then our procedure will result in an overall loss of power.

There will clearly be a large amount of correlation between genome scans using different optimal data sets and hence a Bonferonni correction will likely be highly conservative. A randomization based procedure would produce better p-values, but may be infeasible because of the large computational time each run requires. An adaptive procedure for determining the number of loci to test could also offer some benefits.

It will be important to explore the outcomes of this method on data where expression levels are linked to more than one causative locus (i.e., c>1). It is possible that c is greater than one in real datasets. It would also be beneficial to develop a better method for determining which gene clusters are causative loci. Currently our approach is very ad hoc.

We make no claim that our method for clustering is superior to any other. Our method is very similar to the k-means algorithm, which is well known to have poor performance relative to other methods. The main advance here is a method for using clusters to infer genotype at unobserved disease loci. It may well be optimal to use a procedure similar to the example: First, filter the data for genes most likely to be associated with disease loci; second, use some other clustering method to produce the gene clusters; third, use our method to estimate parameters for

those clusters and assign genotypes at the their controlling loci. Future work will explore this issue.

3.5 - REFERENCES

ALTSHULER, D., L. D. BROOKS, A. CHAKRAVARTI, F. S. COLLINS, M. J. DALY *et al.*, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological **57:** 289-300.

BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences **102:** 1572-1577.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

DUERR, R. H., K. D. TAYLOR, S. R. BRANT, J. D. RIOUX, M. S. SILVERBERG *et al.*, 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science **314:** 1461-1463.

GHAZALPOUR, A., S. DOSS, B. ZHANG, S. WANG, C. PLAISIER *et al.*, 2006 Integrating genetic and network analysis to characterize genes related to mouse weight. Plos Genetics **2:** 1182-1192.

GIBSON, G., and B. WEIR, 2005 The quantitative genetics of transcription. Trends in Genetics **21:** 616-623.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6:** 95-108.

HUNTER, D. J., P. KRAFT, K. B. JACOBS, D. G. COX, M. YEAGER *et al.*, 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics **39:** 870-874.

JANSEN, R. C., and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends in Genetics **17:** 388-391.

LACHIN, J. M., 1977 Sample-Size Determinations For Rxc Comparative Trials. Biometrics **33:** 315-324.

LI, J., and M. BURMEISTER, 2005 Genetical genomics: combining genetics with gene expression analysis. Human Molecular Genetics **14:** R163-R169.

LOHMUELLER, K. E., C. L. PEARCE, M. PIKE, E. S. LANDER and J. N. HIRSCHHORN, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics **33:** 177-182.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. American Journal of Human Genetics **75:** 1094-1105.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

RISCH, N., 1990 Linkage Strategies For Genetically Complex Traits .2. The Power Of Affected Relative Pairs. American Journal of Human Genetics **46:** 229-241.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHLIEKELMAN, P., and M. SLATKIN, 2002 Multiplex relative risk and estimation of the number of loci underlying an inherited disease. American Journal of Human Genetics **71:** 1369-1385.

SLADEK, R., G. ROCHELEAU, J. RUNG, C. DINA, L. SHEN *et al.*, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445:** 881-885.

SLAGER, S. L., J. HUANG and V. J. VIELAND, 2000 Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genetic Epidemiology **18:** 143-156.

STAMATOYANNOPOULOS, J. A., 2004 The genomics of gene expression. Genomics **84:** 449-457.

SUN, G. and SCHLIEKELMAN, P., 2008 A genetical genomics approach to genome scans for complex traits. (to be submitted)

YEAGER, M., N. ORR, R. B. HAYES, K. B. JACOBS, P. KRAFT *et al.*, 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics **39:** 645-649.

YU, Y. P., D. LANDSITTEL, L. JING, J. NELSON, B. G. REN *et al.*, 2004 Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. Journal of Clinical Oncology **22:** 2790-2799.

CHAPTER 4

CONCLUSIONS

In this chapter we give a brief overview of both statistical genetic methods developed in this dissertation. In Chapter 2 we explored the effects of three types of genotyping errors on the population differentiation statistic F_{ST} and the conclusions made from these statistics. This research is motivated by the recent understanding that although eliminating genotyping errors from microsatellite data would be ideal this is not statistically possible. The effects of genotyping errors have been developed for parentage assessment and population estimation but few studies have been done to determine the effects on population differentiation. Our research explored the effects of allelic dropout, binning errors, and null alleles on F_{ST}. We performed our analysis over multiple datasets of varying sizes at large percentages of genotyping error. We found that allelic dropout has no statistically significant effect on the values of FST. Binning error has a statistically significant effect on the FST values however; the overall conclusion that the populations are different remains the same. Null Alleles also have a significant effect on F_{ST} but again the conclusions remain the same. As a result of our multiple simulations we were also able to illustrate the effects of sample size and number of loci on the population differentiation statistic. Future methods should see if these same results occur with different population differentiation statistics.

In Chapter 3 we develop a method for improving gene mapping of complex diseases. Our goal was to perform a genome-wide study, with no prior knowledge of location of disease

genes, to identify disease causative loci. In general, genome-wide studies are analyzed using linkage or association studies. Linkage analysis, although useful for single locus disease and identifying rare variants in complex diseases, requires extremely large datasets of family data and is has been proven to be ineffective at finding common variants of complex diseases. Association studies, although good at identifying common variants, can have a significant loss of power due the fact that all individuals' disease status is not determined by the same causative loci. That is, there may be some individuals who disease status is not determined by the particular locus in question. Our method uses genome-wide expression data to identify these individuals for a particular causative locus and eliminate them, creating increase power of association. We use microarray expression data and disease status of individuals to cluster the expression levels into locus clusters. These gene clusters ideally correspond to genetic loci. Using our clustered information and likelihood parameters we infer genotype structure of the causative locus clusters for each individual. This information is used to eliminate the individuals whose disease status in not determined by that particular causative locus. Through examples of case-control studies we show that our method does increase the power of association to the disease.

In conclusion, useful statistical genetic results were produced from our research efforts.

REFERENCES

ADAMS, J. R., and L. P. WAITS, 2007 An efficient method for screening faecal DNA genotypes and detecting new individuals and hybrids in the red wolf (Canis rufus) experimental population area. Conservation Genetics **8:** 123-131.

AGAPITO, J., J. RODRIGUEZ, P. HERRERA-VELIT, O. TIMOTEO, P. ROJAS *et al.*, 2008 Parentage testing in alpacas (Vicugna pacos) using semi-automated fluorescent multiplex PCRs with 10 microsatellite markers. Animal Genetics **39:** 201-203.

AKEY, J. M., K. ZHANG, M. M. XIONG, P. DORIS and L. JIN, 2001 The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. American Journal of Human Genetics **68:** 1447-1456.

ALTMULLER, J., L. J. PALMER, G. FISCHER, H. SCHERB and M. WJST, 2001 Genomewide scans of complex human diseases: True linkage is hard to find. American Journal of Human Genetics **69:** 936-950.

ALTSHULER, D., L. D. BROOKS, A. CHAKRAVARTI, F. S. COLLINS, M. J. DALY *et al.*, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.

AMOS, W., 2006 The hidden value of missing genotypes. Molecular Biology and Evolution **23:** 1995-1996.

AMOS, W., J. I. HOFFMAN, A. FRODSHAM, L. ZHANG, S. BEST *et al.*, 2007 Automated binning of microsatellite alleles: problems and solutions. Molecular Ecology Notes **7:** 10-14.

ARAKI, H., and M. S. BLOUIN, 2005 Unbiased estimation of relative reproductive success of different groups: evaluation and correction of bias caused by parentage assignment errors. Molecular Ecology **14:** 4097-4109.

BADZIOCH, M. D., H. B. DEFRANCE and G. P. JARVIK, 2003 An examination of the genotyping error detection function of SIMWALK2. Bmc Genetics **4**.

BALLOUX, F., 2001 EASYPOP (version 1.7): A computer program for population genetics simulations. Journal of Heredity **92:** 301-302.

BALLOUX, F., and J. GOUDET, 2002 Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Molecular Ecology **11:** 771-783.

BARIC, S., S. MONSCHEIN, M. HOFER, D. GRILL and J. D. VIA, 2008 Comparability of genotyping data obtained by different procedures an inter-laboratory survey. Journal of Horticultural

Science & Biotechnology **83:** 183-190.

BELLEMAIN, E., J. E. SWENSON, O. TALLMON, S. BRUNBERG and P. TABERLET, 2005 Estimating population size of elusive animals with DNA from hunter-collected feces: Four methods for brown bears. Conservation Biology **19:** 150-161.

BENJAMINI, Y., and Y. HOCHBERG, 1995 CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING. Journal of the Royal Statistical Society Series B-Methodological **57:** 289-300.

BJORKLUND, M., 2005 A method for adjusting allele frequencies in the case of microsatellite allele drop-out. Molecular Ecology Notes **5:** 676-679.

BONIN, A., E. BELLEMAIN, P. B. EIDESEN, F. POMPANON, C. BROCHMANN *et al.*, 2004 How to track and assess genotyping errors in population genetics studies. Molecular Ecology **13:** 3261-3273.

BONIN, A., D. EHRICH and S. MANEL, 2007 Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Molecular Ecology **16:** 3737-3758.

BONIN, A., F. NICOLE, F. POMPANON, C. MIAUD and P. TABERLET, 2007 Population adaptive index: a new method to help measure intraspecific genetic diversity and prioritize populations for conservation. Conservation Biology **21:** 697-708.

BRADLEY, B. J., and L. VIGILANT, 2002 False alleles derived from microbial DNA pose a potential source of error in microsatellite genotyping of DNA from faeces. Molecular Ecology Notes **2:** 602-605.

BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences **102:** 1572-1577.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

BROQUET, T., N. MENARD and E. PETIT, 2007 Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. Conservation Genetics **8:** 249-260.

BROQUET, T., and E. PETIT, 2004 Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology **13:** 3601-3608.

BRUFORD, M. W., and R. K. WAYNE, 1993 Microsatellites and their application to population genetic studies. Curr Opin Genet Dev **3:** 939-943.

BUCHAN, J. C., E. A. ARCHIE, R. C. VAN HORN, C. J. MOSS and S. C. ALBERTS, 2005 Locus effects and sources of error in noninvasive genotyping. Molecular Ecology Notes **5:** 680-683.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nature Reviews Genetics **2:** 91-99.

CARLSON, C. S., M. A. EBERLE, L. KRUGLYAK and D. A. NICKERSON, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429:** 446-452.

CARVAJAL-RODRIGUEZ, A., 2007 FAMSPHERE: a computer program for parental allocation from known genotypic pools. Molecular Ecology Notes **7:** 213-216.

CASTRO, J., A. PINO, M. HERMIDA, C. BOUZA, A. RIAZA *et al.*, 2006 A microsatellite marker tool for parentage analysis in Senegal sole (Solea senegalensis): Genotyping errors, null alleles and conformance to theoretical assumptions. Aquaculture **261:** 1194-1203.

CAUDRON, A. K., S. S. NEGRO, C. G. MULLER, L. J. BOREN and N. J. GEMMELL, 2007 Hair sampling and genotyping from hair follicles: A minimally-invasive alternative for genetics studies in small, mobile pinnipeds and other mammals. Marine Mammal Science **23:** 184-192.

CHAPUSOT, C., L. MARTIN, P. L. PUIG, T. PONNELLE, N. CHEYNEL *et al.*, 2004 What is the best way to assess microsatellite instability status in colorectal cancer? Study on a population base of 462 colorectal cancers. American Journal of Surgical Pathology **28:** 1553-1559.

CHEN, M., and C. KENDZIORSKI, 2007 A statistical framework for expression quantitative trait loci mapping. Genetics **177:** 761-771.

CHENG, K. F., and W. J. LIN, 2007 Simultaneously correcting for population stratification and for genotyping error in case-control association studies. American Journal of Human Genetics **81:** 726-743.

CREEL, S., G. SPONG, J. L. SANDS, J. ROTELLA, J. ZEIGLE *et al.*, 2003 Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology **12:** 2003-2009.

DAKIN, E. E., and J. C. AVISE, 2004 Microsatellite null alleles in parentage analysis. Heredity **93:** 504-509.

DATTA, S., 2003 Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics **19:** 459-466.

DAVID, P., B. PUJOL, F. VIARD, V. CASTELLA and J. GOUDET, 2007 Reliable selfing rate estimates from imperfect population genetic data. Molecular Ecology **16:** 2474-2487.

DAVISON, A., and S. CHIBA, 2003 Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. Molecular Ecology Notes **3:**

321-323.

DEWOODY, J., J. D. NASON and V. D. HIPKINS, 2006 Mitigating scoring errors in microsatellite data from wild populations. Molecular Ecology Notes **6:** 951-957.

DRAGHICI, S., 2003 *Data Analysis Tools for DNA Microarrays*. CRC Press UK, London, UK.

DREHER, B. P., S. R. WINTERSTEIN, K. T. SCRIBNER, P. M. LUKACS, D. R. ETTER *et al.*, 2007 Noninvasive estimation of black bear abundance incorporating genotyping errors and harvested bear. Journal of Wildlife Management **71:** 2684-2693.

DUCHESNE, P., T. MELDGAARD and P. BERREBI, 2008 Parentage analysis with few contributing breeders: Validation and improvement. Journal of Heredity **99:** 323-334.

DUERR, R. H., K. D. TAYLOR, S. R. BRANT, J. D. RIOUX, M. S. SILVERBERG *et al.*, 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science **314:** 1461-1463.

EKSTROM, C. T., 2003 Detecting low-quality markers using map expanders. Genetic Epidemiology **25:** 214-224.

EWEN, K. R., M. BAHLO, S. A. TRELOAR, D. F. LEVINSON, B. MOWRY *et al.*, 2000 Identification and analysis of error types in high-throughput genotyping. American Journal of Human Genetics **67:** 727-736.

FERNANDO, P., B. J. EVANS, J. C. MORALES and D. J. MELNICK, 2001 Electrophoresis artefacts - a previously unrecognized cause of error in microsatellite analysis. Molecular Ecology Notes **1:** 325-328.

FERNANDO, P., T. N. C. VIDYA, C. RAJAPAKSE, A. DANGOLLA and D. J. MELNICK, 2003 Reliable noninvasive genotyping: Fantasy or reality? Journal of Heredity **94:** 115-123.

GAGGIOTTI, O. E., O. LANGE, K. RASSMANN and C. GLIDDON, 1999 A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Molecular Ecology **8:** 1513-1520.

GAGNEUX, P., C. BOESCH and D. S. WOODRUFF, 1997 Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology **6:** 861-868.

GHOSH, S., Z. E. KARANJAWALA, E. R. HAUSER, D. ALLY, J. I. KNAPP *et al.*, 1997 Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. Genome Research **7:** 165-178.

GIBSON, G., and B. WEIR, 2005 The quantitative genetics of transcription. Trends in Genetics **21:**

616-623.

GLAZIER, A. M., J. H. NADEAU and T. J. AITMAN, 2002 Finding genes that underlie complex traits. Science **298:** 2345-2349.

GOMES, I., A. COLLINS, C. LONJOU, N. S. THOMAS, J. WILKINSON *et al.*, 1999 Hardy-Weinberg quality control. Annals of Human Genetics **63:** 535-538.

GOOSSENS, B., L. P. WAITS and P. TABERLET, 1998 Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. Molecular Ecology **7:** 1237-1241.

GOUDET, J., 1995 FSTAT (Version 1.2): A computer program to calculate F-statistics. Journal of Heredity **86:** 485-486.

GUNN, M. R., K. HARTNUP, S. BOUTIN, J. SLATE and D. W. COLTMAN, 2007 A test of the efficacy of whole-genome amplification on DNA obtained from low-yield samples. Molecular Ecology Notes **7:** 393-399.

HAJKOVA, P., B. ZEMANOVA, J. BRYJA, B. HAJEK, K. ROCHE *et al.*, 2006 Factors affecting success of PCR amplification of microsatellite loci from otter faeces. Molecular Ecology Notes **6:** 559-562.

HARTL, D. L., 2000 *A Primer of Population Genetics*. Sinauer Associates, Inc, Suderland, Mass.

HAUSKNECHT, R., R. GULA, B. PIRGA and R. KUEHN, 2007 Urine - a source for noninvasive genetic monitoring in wildlife. Molecular Ecology Notes **7:** 208-212.

HEDMARK, E., and H. ELLEGREN, 2006 A test of the multiplex pre-amplification approach in microsatellite genotyping of wolverine faecal DNA. Conservation Genetics **7:** 289-293.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6:** 95-108.

HOFFMAN, J. I., and W. AMOS, 2005 Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology **14:** 599-612.

HUNTER, D. J., P. KRAFT, K. B. JACOBS, D. G. COX, M. YEAGER *et al.*, 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics **39:** 870-874.

ISLER, J. A., O. E. VESTERQVIST and M. E. BURCZYNSKI, 2007 Analytical validation of genotyping assays in the biomarker laboratory. Pharmacogenomics **8:** 353-368.

JANECKA, J. E., L. I. GRASSMAN, R. L. HONEYCUTT and M. E. TEWES, 2007 Whole genome amplification for sequencing and applications in conservation genetics. Journal of

Wildlife Management **71:** 1357-1360.

JANSEN, R. C., and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends in Genetics **17:** 388-391.

JI, F., Y. N. YANG, C. HAYNES, S. J. FINCH and D. GORDON, 2005 Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. Statistical Applications in Genetics and Molecular Biology **4**.

JOHNSON, P. C. D., and D. T. HAYDON, 2007 Maximum-likelihood estimation of allelic dropout and false allele error rates from Microsatellite genotypes in the absence of reference data. Genetics **175:** 827-842.

KALINOWSKI, S. T., 2006 HW-QUICKCHECK: an easy-to-use computer program for checking genotypes for agreement with Hardy-Weinberg expectations. Molecular Ecology Notes **6:** 974-979.

KALINOWSKI, S. T., M. L. TAPER and S. CREEL, 2006 Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. Conservation Genetics **7:** 319-329.

KALINOWSKI, S. T., M. L. TAPER and T. C. MARSHALL, 2007 Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology **16:** 1099-1106.

KENDZIORSKI, C. M., M. CHEN, M. YUAN, H. LAN and A. D. ATTIE, 2006 Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics **62:** 19-27.

KIROV, G., N. WILLIAMS, P. SHAM, N. CRADDOCK and M. J. OWEN, 2000 Pooled genotyping of microsatellite markers in parent-offspring trios. Genome Research **10:** 105-115.

KNAPP, L. A., 2005 Facts, faeces and setting standards for the study of MHC genes using noninvasive samples. Molecular Ecology **14:** 1597-1599.

LACHIN, J. M., 1977 SAMPLE-SIZE DETERMINATIONS FOR RXC COMPARATIVE TRIALS. Biometrics **33:** 315-324.

LAI, Y. L., and F. Z. SUN, 2004 Sampling distribution for microsatellites amplified by PCR: mean field approximation and its applications to genotyping. Journal of Theoretical Biology **228:** 185-194.

LEVINSON, D. F., K. R. EWEN, M. BAHLO, S. A. TRELOAR, B. MOWRY *et al.*, 2000 Estimating genotyping error rate in a fine-mapping project. American Journal of Medical Genetics **96:** 570-571.

LI, J., and M. BURMEISTER, 2005 Genetical genomics: combining genetics with gene expression analysis. Human Molecular Genetics **14:** R163-R169.

LIVIA, L., P. ANTONELLA, L. HOVIRAG, N. MAURO and F. PANARA, 2006 A nondestructive, rapid, reliable and inexpensive method to sample, store and extract high-quality DNA from fish body mucus and buccal cells. Molecular Ecology Notes **6:** 257-260.

LOHMUELLER, K. E., C. L. PEARCE, M. PIKE, E. S. LANDER and J. N. HIRSCHHORN, 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics **33:** 177-182.

LUIKART, G., S. ZUNDEL, D. RIOUX, C. MIQUEL, K. A. KEATING *et al.*, 2008 Low genotyping error rates and noninvasive sampling in bighorn sheep. Journal of Wildlife Management **72:** 299-304.

LUKACS, P. M., and K. P. BURNHAM, 2005 Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. Journal of Wildlife Management **69:** 396-403.

LYNCH, M., and B. WALSH, 1998 *Genetics and analysis of quantitative traits.* Sinauer Associates, Sunderland, MA, USA.

MANEL, S., F. BERTHOUD, E. BELLEMAIN, M. GAUDEUL, G. LUIKART *et al.*, 2007 A new individual-based spatial approach for identifying genetic discontinuities in natural populations. Molecular Ecology **16:** 2031-2043.

MARTINEZ, J. G., and T. BURKE, 2003 Microsatellite typing of sperm trapped in the perivitelline layers of avian eggs: a cautionary note. Journal of Avian Biology **34:** 20-24.

MAUDET, C., G. LUIKART, D. DUBRAY, A. VON HARDENBERG and P. TABERLET, 2004 Low genotyping error rates in wild ungulate faeces sampled in winter. Molecular Ecology Notes **4:** 772-775.

MCKELVEY, K. S., and M. K. SCHWARTZ, 2004 Genetic errors associated with population estimation using non-invasive molecular tagging: Problems and new solutions. Journal of Wildlife Management **68:** 439-448.

MCKELVEY, K. S., and M. K. SCHWARTZ, 2004 Providing reliable and accurate genetic capture-mark-recapture estimates in a cost-effective way. Journal of Wildlife Management **68:** 453-456.

MCKELVEY, K. S., and M. K. SCHWARTZ, 2005 DROPOUT: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. Molecular Ecology Notes **5:** 716-718.

MILLER, C. R., P. JOYCE and L. P. WAITS, 2002 Assessing allelic dropout and genotype reliability

using maximum likelihood. Genetics **160:** 357-366.

MIQUEL, C., E. BELLEMAIN, C. POILLOT, J. BESSIERE, A. DURAND *et al.*, 2006 Quality indexes to assess the reliability of genotypes in studies using noninvasive sampling and multiple-tube approach. Molecular Ecology Notes **6:** 985-988.

MITCHELL, A. A., D. J. CUTLER and A. CHAKRAVARTI, 2002 Genotyping error introduces bias to the transmission disequilibrium test (TDT). American Journal of Human Genetics **71:** 204-204.

MITCHELL, A. A., D. J. CUTLER and A. CHAKRAVARTI, 2003 Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. American Journal of Human Genetics **72:** 598-610.

MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. American Journal of Human Genetics **75:** 1094-1105.

MONTGOMERY, G. W., M. J. CAMPBELL, P. DICKSON, S. HERBERT, K. SIEMERING *et al.*, 2005 Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. Twin Research and Human Genetics **8:** 346-352.

MOONESINGHE, R., M. J. KHOURY, T. LIU and J. P. A. IOANNIDIS, 2008 Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences of the United States of America **105:** 617-622.

MORIN, P. A., and M. MCCARTHY, 2007 Highly accurate SNP genotyping from historical and low-quality samples. Molecular Ecology Notes **7:** 937-946.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature **430:** 743-747.

MORRISSEY, M. B., and A. J. WILSON, 2005 The potential costs of accounting for genotypic errors in molecular parentage analyses. Molecular Ecology **14:** 4111-4121.

MUKHOPADHYAY, N., S. G. BUXBAUM and D. E. WEEKS, 2004 Comparative study of multipoint methods for genotype error detection. Human Heredity **58:** 175-189.

MURPHY, M. A., K. C. KENDALL, A. ROBINSON and L. P. WAITS, 2007 The impact of time and field conditions on brown bear (Ursus arctos) faecal DNA amplification. Conservation Genetics **8:** 1219-1224.

NEI, M., 1973 ANALYSIS OF GENE DIVERSITY IN SUBDIVIDED POPULATIONS. Proceedings of the National Academy of Sciences of the United States of America **70:** 3321-3323.

OKELLO, J. B. A., G. WITTEMYER, H. B. RASMUSSEN, I. DOUGLAS-HAMILTON, S. NYAKAANA *et al.*, 2005 Noninvasive genotyping and mendelian analysis of microsatellites in African savannah elephants. Journal of Heredity **96:** 679-687.

PAETKAU, D., 2003 An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology **12:** 1375-1387.

PARSONS, K. M., 2001 Reliable microsatellite genotyping of dolphin DNA from faeces. Molecular Ecology Notes **1:** 341-344.

PASQUALOTTO, A. C., D. W. DENNING and M. J. ANDERSON, 2007 A cautionary tale: Lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. Journal of Clinical Microbiology **45:** 522-528.

PEAKALL, R., and P. E. SMOUSE, 2006 GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes **6:** 288-295.

PIGGOTT, M. P., 2004 Effect of sample age and season of collection on the reliability of microsatellite genotyping of faecal DNA. Wildlife Research **31:** 485-493.

PIGGOTT, M. P., E. BELLEMAIN, P. TABERLET and A. C. TAYLOR, 2004 A multiplex pre-amplification method that significantly improves microsatellite amplification and error rates for faecal DNA in limiting conditions. Conservation Genetics **5:** 417-420.

POMILLA, C., and H. C. ROSENBAUM, 2006 Estimates of relatedness in groups of humpback whales (Megaptera novaeangliae) on two wintering grounds of the Southern Hemisphere. Molecular Ecology **15:** 2541-2555.

POMPANON, F., A. BONIN, E. BELLEMAIN and P. TABERLET, 2005 Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics **6:** 847-859.

PRESSON, A. P., E. SOBEL, K. LANGE and J. C. PAPP, 2006 Merging microsatellite data. Journal of Computational Biology **13:** 1131-1147.

PUECHMAILLE, S. J., G. MATHY and E. J. PETIT, 2007 Good DNA from bat droppings. Acta Chiropterologica **9:** 269-276.

QUELLER, D. C., J. E. STRASSMANN and C. R. HUGHES, 1993 Microsatellites And Kinship. Trends in Ecology & Evolution **8:** 285-&.

RISCH, N., 1990 Linkage Strategies For Genetically Complex Traits .2. The Power Of Affected Relative Pairs. American Journal of Human Genetics **46:** 229-241.

RISCH, N., and K. MERIKANGAS, 1996 The future of genetic studies of complex human diseases. Science **273:** 1516-1517.

RODRIGUEZ, S., G. VISEDO and C. ZAPATA, 2001 Detection of errors in dinucleotide repeat typing by nondenaturing electrophoresis. Electrophoresis **22:** 2656-2664.

ROON, D. A., M. E. THOMAS, K. C. KENDALL and L. P. WAITS, 2005 Evaluating mixed samples as a source of error in non-invasive genetic studies using microsatellites. Molecular Ecology **14:** 195-201.

ROON, D. A., L. P. WAITS and K. C. KENDALL, 2005 A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data. Animal Conservation **8:** 203-215.

ROUSSET, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357-1362.

SCANDURA, M., C. CAPITANI, L. IACOLINA and A. MARCO, 2006 An empirical approach for reliable microsatellite genotyping of wolf DNA from multiple noninvasive sources. Conservation Genetics **7:** 813-823.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHLIEKELMAN, P., 2008 Statistical Power of Expression Quantitative Trait Loci for Mapping of Complex Trait Loci in Natural Populations. Genetics **178:** 2201-2216.

SCHLIEKELMAN, P., and M. SLATKIN, 2002 Multiplex relative risk and estimation of the number of loci underlying an inherited disease. American Journal of Human Genetics **71:** 1369-1385.

SCHNEIDER, P. M., 2007 Scientific standards for studies in forensic genetics. Forensic Science International **165:** 238-243.

SCHWARTZ, M. K., S. A. CUSHMAN, K. S. MCKELVEY, J. HAYDEN and C. ENGKJER, 2006 Detecting genotyping errors and describing American black bear movement in northern Idaho. Ursus **17:** 138-148.

SCHWARTZ, M. K., G. LUIKART and R. S. WAPLES, 2007 Genetic monitoring as a promising tool for conservation and management. Trends in Ecology & Evolution **22:** 25-33.

SEFC, K. M., R. B. PAYNE and M. D. SORENSON, 2003 Microsatellite amplification from museum feather samples: Effects of fragment size and template concentration on genotyping errors. Auk **120:** 982-989.

SEFC, K. M., R. B. PAYNE and M. D. SORENSON, 2007 Single base errors in PCR products from avian museum specimens and their effect on estimates of historical genetic diversity. Conservation Genetics **8:** 879-884.

SELKOE, K. A., and R. J. TOONEN, 2006 Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters **9:** 615-629.

SLADEK, R., G. ROCHELEAU, J. RUNG, C. DINA, L. SHEN *et al.*, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445:** 881-885.

SLAGER, S. L., J. HUANG and V. J. VIELAND, 2000 Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genetic Epidemiology **18:** 143-156.

SLATKIN, M., 1995 A MEASURE OF POPULATION SUBDIVISION BASED ON MICROSATELLITE ALLELE FREQUENCIES. Genetics **139:** 457-462.

SLAVOV, G. T., G. T. HOWE, A. V. GYAOUROVA, D. S. BIRKES and W. T. ADAMS, 2005 Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. Molecular Ecology **14:** 3109-3121.

SMITH, D. A., K. RALLS, A. HURT, B. ADAMS, M. PARKER *et al.*, 2006 Assessing reliability of microsatellite genotypes from kit fox faecal samples using genetic and GIS analyses. Molecular Ecology **15:** 387-406.

SMITH, J. R., J. D. CARPTEN, M. J. BROWNSTEIN, S. GHOSH, V. L. MAGNUSON *et al.*, 1995 Approach to Genotyping Errors Caused by Nontemplated Nucleotide Addition by Taq DNA-Polymerase. Genome Research **5:** 312-317.

SOULSBURY, C. D., G. IOSSA, K. J. EDWARDS, P. J. BAKER and S. HARRIS, 2007 Allelic dropout from a high-quality DNA source. Conservation Genetics **8:** 733-738.

STAMATOYANNOPOULOS, J. A., 2004 The genomics of gene expression. Genomics **84:** 449-457.

SUN, G. and SCHLIEKELMAN, P., 2008 A genetical genomics approach to genome scans for complex traits. (to be submitted)

TAGGART, J. B., 2007 FAP: an exclusion-based parental assignment program with enhanced predictive functions. Molecular Ecology Notes **7:** 412-415.

TAUTZ, D., 1989 Hypervariability Of Simple Sequences As A General Source For Polymorphic Dna Markers. Nucleic Acids Research **17:** 6463-6471.

TAUTZ, D., and M. RENZ, 1984 Simple Sequences Are Ubiquitous Repetitive Components Of Eukaryotic Genomes. Nucleic Acids Research **12:** 4127-4138.

THOMPSON, C. L., D. BAECHLE, Q. LU, G. MATHEW, Y. J. SONG *et al.*, 2005 Effect of genotyping error in model-free linkage analysis using microsatellite or single-nucleotide polymorphism marker maps. Bmc Genetics **6**.

TINTLE, N. L., D. GORDON, F. J. MCMAHON and S. J. FINCH, 2007 Using duplicate genotyped data

in genetic analyses: Testing association and estimating error rates. Statistical Applications in Genetics and Molecular Biology **6**.

VALI, U., M. BRANDSTROM, M. JOHANSSON and H. ELLEGREN, 2008 Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. Bmc Genetics **9**.

VALIERE, N., 2002 GIMLET: a computer program for analysing genetic individual identification data. Molecular Ecology Notes **2:** 377-379.

VALIERE, N., C. BONENFANT, C. TOIGO, G. LUIKART, J. M. GAILLARD *et al.*, 2007 Importance of a pilot study for non-invasive genetic sampling: genotyping errors and population size estimation in red deer. Conservation Genetics **8:** 69-78.

VANDEPUTTE, M., S. MAUGER and M. DUPONT-NIVET, 2006 An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. Molecular Ecology Notes **6:** 265-267.

WAITS, J. L., and P. L. LEBERG, 2000 Biases associated with population estimation using molecular tagging. Animal Conservation **3:** 191-199.

WAITS, L. P., and D. PAETKAU, 2005 Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. Journal of Wildlife Management **69:** 1419-1433.

WALTERS, K., 2005 The effect of genotyping error in sib-pair genomewide linkage scans depends crucially upon the method of analysis. Journal of Human Genetics **50:** 329-337.

WEBER, J. L., and P. E. MAY, 1989 ABUNDANT CLASS OF HUMAN DNA POLYMORPHISMS WHICH CAN BE TYPED USING THE POLYMERASE CHAIN-REACTION. American Journal of Human Genetics **44:** 388-396.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-Statistics For The Analysis Of Population-Structure. Evolution **38:** 1358-1370.

WELLER, J. I., E. SEROUSSI and M. RON, 2006 Estimation of the number of genetic markers required for individual animal identification accounting for genotyping errors. Animal Genetics **37:** 387-389.

WRIGHT, S., 1951 The Genetical Structure Of Populations. Annals of Eugenics **15:** 323-354.

WRIGHT, S., 1965 The Interpretation Of Population-Structure By F-Statistics With Special Regard To Systems Of Mating. Evolution **19:** 395-420.

XU, S. Z., and W. R. ATCHLEY, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417-1424.

YEAGER, M., N. ORR, R. B. HAYES, K. B. JACOBS, P. KRAFT *et al.*, 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature Genetics **39:** 645-649.

YI, N. J., and S. Z. XU, 1999 Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity **82:** 668-676.

YI, N. J., and S. Z. XU, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391-1403.

YU, Y. P., D. LANDSITTEL, L. JING, J. NELSON, B. G. REN *et al.*, 2004 Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. Journal of Clinical Oncology **22:** 2790-2799.

ZEDROSSER, A., E. BELLEMAIN, P. TABERLET and J. E. SWENSON, 2007 Genetic estimates of annual reproductive success in male brown bears: the effects of body size, age, internal relatedness and population density. Journal of Animal Ecology **76:** 368-375.

ZHANG, H. M., and H. STERN, 2006 Assessment of ancestry probabilities in the presence of genotyping errors. Theoretical and Applied Genetics **112:** 472-482.

APPENDIX A

RESULTS OF ALL 74 SUB-DATASETS FOR ALL 99 REPLICATIONS – DATASET A

Because of its length this appendix is in electronic form attached to this dissertation.

APPENDIX B

RESULTS OF ALL 74 SUB-DATASETS FOR ALL 99 REPLICATIONS – DATASET B.

Because of its length this appendix is in electronic form attached to this dissertation.

APPENDIX C

SUMMARY OF SIMULATIONS FOR CLUSTERING METHOD AND INFERRING GENOTYPE

1. L=9 M=6 LND=9 MND=18 N=200 25% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	**Likelihood**
3	-24592.2
4	-23862.6
5	-23349.5
6	-22412.9
7	-21995.6
8	-20225.7
9	-20208.1
10	-21064.5
11	-20889.2
12	-20184.7
13	-20285.7
14	-20127.5
15	**-20007.2**
16	-20089.7
17	-20188.7
18	-20167.4
19	-20058.1
20	-20227.8

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   13   16   27   34   40   43   80   89   107  116  125  134
      152  161  170  173  175  188  189  206
[1]   7    21   59   64   73   82   91   100  127  136  145  163
      190  199  207  208
[2]   56   65   74   83   98   110  119  146  155  182  200  215
[3]   29   31   52   77   86   95   104  113  122  131  140  149
      158  167  176  194  203  212
[4]   2    4    25   60   71   87   96   195
[5]   69   114  132  141  150  159  168  177  213
[6]   1    3    10   19   28   30   37   39   46   47   48   55
      109  154  172
[7]   18   61   70   88   106  115  118  124  133  142  151  160
      169  178  181  187  196  205  214
[8]   5    6    14   15   23   24   32   33   36   41   42   50
      51   180
```

```
[9]   8    9    17   20   26   35   38   44   45   53   68   79
      111  156  162  185  210
[10]  12   57   66   75   84   93   102  120  129  138  147  165
      174  183  192  201
[11]  0    22   92   101  128  137  164  191  209
[12]  78   105  123  179  186  197  204
[13]  11   49   54   63   72   81   90   97   99   117  126  135
      144  153  171  198
[14]  58   62   67   76   85   94   103  108  112  121  130  139
      143  148  157  166  184  193  202  211
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	78	136
Locus	[6]	[9]

- values of $P(\vec{Y} \mid \vec{G})$ to determine which locus clusters are associated with disease and genes are well clustered

LOCUS	P(Y|G)
6	**6.61E-53**
9	**1.27E-55**
11	8.55E-59
14	2.91E-59
8	2.43E-59
0	8.75E-60
2	8.31E-60
13	5.91E-60
7	2.39E-60
5	1.72E-60
4	1.57E-60
1	1.24E-60
10	1.20E-60
12	8.91E-61
3	7.75E-61

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 6

Simulated

	0	1	2
Disease	0	39	61
Non-Disease	22	45	33

Calculated

	0	1	2
Disease	18	48	34
Non-Disease	59	26	15

Disease	0	1	2
0->	0	0	0
1->	18	0	9
2->	0	36	0

Non-Disease	0	1	2
0->	0	0	0
1->	37	0	2
2->	0	20	0

Locus 9

	0	1	2
Disease	0	38	62
Non-Disease	11	50	39

	0	1	2
Disease	29	9	62
Non-Disease	33	34	33

Disease	0	1	2
0->	0	0	0
1->	28	0	2
2->	1	1	0

Non-Disease	0	1	2
0->	0	2	0
1->	22	0	1
2->	2	5	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 6	0	0	0.18	0.09	0	0.36
Locus 9	0	0	0.28	0.02	0.01	0.01

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 6	0	0	0.37	0.02	0	0.2
Locus 9	0.02	0	0.22	0.01	0.02	0.05

2. L=9 M=6 LND=9 MND=18 N=400 25% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	**Likelihood**
3	-50456.9
4	-49044.7
5	-47874.5
6	-45944
7	-43461.8
8	-42019.4
9	-42678.6
10	-42184.8
11	-41356.8
12	-42029.2
13	-40395.1
14	-40255.5
15	-40910.2
16	-41570.5
17	-41197.7
18	-40918.6
19	**-40222.9**
20	-41247.8

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   24   35   62   71   80   89   98   107  116  125  143  159
      161  170  179  188  197  215
[1]   33   54   61   79   88   97   106  108  115  124  133  160
      169  178  187  196  214
[2]   56   74   84   134  137  151  182  191  200
[3]   2    11   29   38   42
[4]   16   77   122  131  158  167  177  185  194  203  212
[5]   4    5    13   14   22   23   31   32   40   41   49   50
      157
[6]   7    25   34   43   52   94   130  184  193  206  211
[7]   15   65   83   92   101  110  119  146  155  164  209
[8]   75   93   129  138  147  152  174
[9]   0    9    18   20   36   45   47   128
[10]  57   59   66   111  120  156  165  183  192  201  210
[11]  8    58   67   76   85   103  112  121  139  148  166  175
      202
[12]  63   72   81   90   99   117  126  135  144  145  153  162
      171  180  189  207
[13]  26   68   86   95   104  113  149  176
[14]  3    12   21   30   39   48   53   205
[15]  6    27   44   55   64   70   73   82   91   100  109  118
      127  154  163  172  190  198  199  208
[16]  17   69   102  114  140  142  173  213
[17]  60   78   87   96   105  123  132  141  150  168  186  195
      204
[18]  1    10   19   28   37   46   51   136  181
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	191	388	367	378	191
Locus	[5]	[18]	[3]	[14]	[5]

- **-** values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	**P(Y|G)**
5	**1.76E-107**
18	**1.12E-108**
3	**2.68E-109**
14	**1.87E-109**
9	**2.78E-112**
6	3.11E-119
15	3.49E-120
11	3.27E-120
13	1.78E-120
2	1.40E-120
4	1.27E-120
8	9.48E-121
10	9.07E-121
17	5.01E-121
12	4.95E-121
16	4.81E-121
0	4.58E-121
1	4.23E-121
7	4.18E-121

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 5

Simulated

	0	1	2
Disease	0	85	115
Non-Disease	34	98	68

Calculated

	0	1	2
Disease	36	116	48
Non-Disease	106	50	44

Disease	0	1	2
0->	0	0	0
1->	33	0	0
2->	3	64	0

Non-Disease	0	1	2
0->	0	8	0
1->	75	0	1
2->	5	20	0

Locus 18

	0	1	2
Disease	0	79	121
Non-Disease	32	97	71

	0	1	2
Disease	0	79	121
Non-Disease	31	97	72

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	3	0

Non-Disease	0	1	2
0->	0	1	0
1->	0	0	3
2->	0	2	0

Locus 3

	0	1	2
Disease	0	83	117
Non-Disease	39	90	71

	0	1	2
Disease	3	74	123
Non-Disease	43	86	71

Disease	0	1	2
0->	0	0	0
1->	3	0	11
2->	0	5	0

Non-Disease	0	1	2
0->	0	0	0
1->	4	0	5
2->	0	5	0

Locus 14

	0	1	2
Disease	0	88	112
Non-Disease	24	104	72

	0	1	2
Disease	0	86	114
Non-Disease	31	98	71

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	2	0

Non-Disease	0	1	2
0->	0	2	0
1->	9	0	2
2->	0	3	0

Locus 9

	0	1	2
Disease	0	69	131
Non-Disease	24	83	93

	0	1	2
Disease	60	47	93
Non-Disease	103	62	35

Disease	0	1	2
0->	0	0	0
1->	59	0	8
2->	1	45	0

Non-Disease	0	1	2
0->	0	0	0
1->	77	0	3
2->	2	59	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 5	0	0	0.165	0	0.015	0.32
Locus 18	0	0	0	0.015	0	0.015

Locus 3	0	0	0.015	0.055	0	0.025
Locus 14	0	0	0	0.02	0	0.01
Locus 9	0	0	0.295	0.04	0.005	0.225

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 5	0.04	0	0.375	0.005	0.025	0.1
Locus 18	0.005	0	0	0.015	0	0.01
Locus 3	0	0	0.02	0.025	0	0.025
Locus 14	0.01	0	0.045	0.01	0	0.015
Locus 9	0	0	0.385	0.015	0.01	0.295

3. L=9 M=6 LND=9 MND=18 N=600 25% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-76979.4
4	-74169.8
5	
6	-69689.1
7	-69030.5
8	-65883.7
9	-64405.7
10	-64147.3
11	-64433.2
12	-65466.8
13	-62764.5
14	-63470.6
15	-64975.9
16	-62688.7
17	-62953.8
18	**-62634.3**
19	-63476.3
20	-63577.4

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   77    131   149   158   167   185   194   212
[1]   55    73    91    100   109   127   136   145   163   172   192   208
[2]   3     7     12    16    21    30    32    39    48
[3]   41    56    74    83    92    101   110   119   146   155   164   173
      182   191   200   209   213
[4]   8     17    26    35    43    44    53    65    76    85    94    139
      148   156   157
[5]   14    20    22    51    59    68    86    95    113   122   140   203
[6]   60    69    96    128   137   150   159   186   204
[7]   0     9     18    27    31    36    40    45
[8]   67    112   165   166   175   202
[9]   2     11    38    47    49    63    64    70    82    99    106   117
      118   135   154   181   190
[10]  6     13    15    23    24    33    42    108
[11]  25    34    54    72    81    90    126   143   144   153   162   176
      180   189   197   198   207
[12]  50    78    87    105   114   123   132   141   168   177   195
[13]  1     4     5     10    19    28    29    37    46    52    66    84
      93    104   134   138   147   174   201
[14]  61    79    88    97    124   133   142   151   160   169   178   187
      196   205   214
[15]  58    103   116   121   171   184   193   199   211
```

```
[16]  62    71    80    89    98    107   115   125   130   152   161   170
      179   188   206   215
[17]  57    75    102   111   120   129   183   210
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	579	566	576	199	178
Locus	[7]	[10]	[2]	[13]	[4]

- - values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	**P(Y|G)**
7	**3.64E-156**
10	**2.74E-160**
2	**5.35E-164**
13	**1.58E-165**
4	**2.06E-176**
8	5.39E-180
17	5.04E-180
15	4.92E-180
14	1.27E-180
16	1.08E-180
0	6.52E-181
12	4.72E-181
5	4.45E-181
6	4.26E-181
9	3.72E-181
3	3.59E-181
11	3.02E-181
1	2.95E-181

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 7

Simulated

	0	1	2
Disease	0	128	172
Non-Disease	66	139	95

Calculated

	0	1	2
Disease	0	132	168
Non-Disease	66	144	90

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	7	0

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	2
2->	0	7	0

Locus 10

	0	1	2
Disease	0	128	172
Non-Disease	46	156	98

	0	1	2
Disease	0	129	171
Non-Disease	49	161	90

Disease	0	1	2
0->	0	0	0
1->	0	0	8
2->	0	9	0

Non-Disease	0	1	2
0->	0	2	0
1->	5	0	1
2->	0	9	0

Locus 2

	0	1	2
Disease	0	125	175
Non-Disease	49	152	99

	0	1	2
Disease	1	129	170
Non-Disease	48	154	98

Disease	0	1	2
0->	0	0	0
1->	1	0	2
2->	0	7	0

Non-Disease	0	1	2
0->	0	4	0
1->	3	0	3
2->	0	4	0

Locus 13

	0	1	2
Disease	0	111	189
Non-Disease	54	150	96

	0	1	2
Disease	49	121	130
Non-Disease	145	69	86

Disease	0	1	2
0->	0	0	0
1->	49	0	46
2->	0	105	0

Non-Disease	0	1	2
0->	0	0	2
1->	91	0	48
2->	2	58	0

Locus 4

	0	1	2
Disease	0	124	176
Non-Disease	42	138	120

	0	1	2
Disease	71	110	119
Non-Disease	120	68	112

Disease	0	1	2
0->	0	0	0
1->	71	0	45
2->	0	102	0

Non-Disease	0	1	2
0->	0	0	4
1->	82	0	53
2->	0	65	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	A(1)	A(2)	B(0)	B(2)	C(0)	C(1)
Locus 7	0	0	0	0.01	0	0.023333
Locus 10	0	0	0	0.026667	0	0.03
Locus 2	0	0	0.003333	0.006667	0	0.023333
Locus 13	0	0	0.163333	0.153333	0	0.35
Locus 4	0	0	0.236667	0.15	0	0.34

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	E(1)	E(2)	F(0)	F(2)	H(0)	H(1)
Locus 7	0.003333	0	0.003333	0.006667	0	0.023333
Locus 10	0.006667	0	0.016667	0.003333	0	0.03
Locus 2	0.013333	0	0.01	0.01	0	0.013333
Locus 13	0	0.006667	0.303333	0.16	0.006667	0.193333
Locus 4	0	0.013333	0.273333	0.176667	0	0.216667

4. L=9 M=6 LND=9 MND=6 N=200 50% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-12906.3
4	-12594.7
5	-12451.9
6	-12238.6
7	-12075.9
8	-11989.6
9	-11902
10	-11656.9
11	-11488.4
12	-11465.5
13	-11411.5
14	-11322.3
15	-11303.8
16	**-10851.9**
17	-11063.4
18	-11436.7
19	-11066.6
20	-10861.9

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]    0     9     27    36    55    64    91    100
[1]    26    57    71    75    84
[2]    2     20    29    38    47
[3]    3     12    18    21    30    39    48
[4]    6     15    33    51    79
[5]    54    63    80    81    90    99
[6]    59    62    68    77    86    89    95    98    104
[7]    16    25    34    42    43    52    102
[8]    7     32    61    70    88    97    106
[9]    58    67    72    85    94    103
[10]   11    24    60    69    78    87    96    105
[11]   1     10    19    28    37    46    73    76
[12]   4     13    22    31    40    45    49    83
[13]   8     17    35    44    53    107
[14]   56    65    74    92    93    101
[15]   5     14    23    41    50    66    82
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	185	188	195	179	196	186
Locus	[13]	[7]	[12]	[15]	[11]	[2]

- - values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	P(Y|G)
13	**2.76E-52**
7	**3.14E-53**
12	**8.29E-54**
15	**1.40E-54**
11	**2.71E-55**
2	**2.10E-55**
3	9.11E-57
4	1.42E-57
0	8.66E-58
10	1.30E-58
8	9.76E-60
9	4.59E-60
14	2.44E-60
1	1.92E-60
5	1.64E-60
6	5.86E-61

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 13

Simulated

	0	1	2
Disease	0	43	57
Non-Disease	23	45	32

Disease	0	1	2
0->	0	0	0
1->	0	0	1
2->	0	7	0

Calculated

	0	1	2
Disease	0	49	51
Non-Disease	25	44	31

Non-Disease	0	1	2
0->	0	1	0
1->	3	0	1
2->	0	2	0

Locus 7

	0	1	2
Disease	0	32	68
Non-Disease	19	45	36

Disease	0	1	2
0->	0	0	0
1->	0	0	1
2->	0	4	0

	0	1	2
Disease	0	35	65
Non-Disease	20	42	38

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	3
2->	0	1	0

Locus 12

	0	1	2
Disease	0	41	59
Non-Disease	21	45	34

	0	1	2
Disease	0	41	59
Non-Disease	19	46	35

Disease	0	1	2
0->	0	0	0
1->	0	0	0
2->	0	0	0

Non-Disease	0	1	2
0->	0	3	0
1->	1	0	1
2->	0	0	0

Locus 15

	0	1	2
Disease	0	37	63
Non-Disease	10	59	31

	0	1	2
Disease	1	38	61
Non-Disease	11	63	26

Disease	0	1	2
0->	0	0	0
1->	1	0	3
2->	0	5	0

Non-Disease	0	1	2
0->	0	2	0
1->	3	0	1
2->	0	6	0

Locus 11

	0	1	2
Disease	0	43	57
Non-Disease	14	49	37

	0	1	2
Disease	0	43	57
Non-Disease	15	49	36

Disease	0	1	2
0->	0	0	0
1->	0	0	0
2->	0	0	0

Non-Disease	0	1	2
0->	0	0	0
1->	1	0	1
2->	0	2	0

Locus 2

	0	1	2
Disease	0	37	63
Non-Disease	15	46	39

	0	1	2
Disease	0	36	64
Non-Disease	16	39	45

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	2	0

Non-Disease	0	1	2
0->	0	0	0
1->	1	0	7
2->	0	1	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	A(1)	A(2)	B(0)	B(2)	C(0)	C(1)
Locus 13	0	0	0	0.01	0	0.07
Locus 7	0	0	0	0.01	0	0.04
Locus 12	0	0	0	0	0	0
Locus 15	0	0	0.01	0.03	0	0.05
Locus 11	0	0	0	0	0	0
Locus 2	0	0	0	0.03	0	0.02

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	E(1)	E(2)	F(0)	F(2)	H(0)	H(1)
Locus 13	0.01	0	0.03	0.01	0	0.02
Locus 7	0.01	0	0.02	0.03	0	0.01
Locus 12	0.03	0	0.01	0.01	0	0
Locus 15	0.02	0	0.03	0.01	0	0.06
Locus 11	0	0	0.01	0.01	0	0.02
Locus 2	0	0	0.01	0.07	0	0.01

5. L=9 M=6 LND=9 MND=6 N=400 50% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	**Likelihood**
3	-26300.9
4	-25691.9
5	-25278.4
6	-24964.7
7	-24453.3
8	-24132.2
9	-23656.3
10	-23278.7
11	-22783.2
12	-23110.8
13	-22881.1
14	-22543.1
15	-22694.6
16	-22602.1
17	-21973.7
18	**-21725.8**
19	-22467.3
20	-21870.5
21	-22196.8
22	-22387.8
23	-22130.3
24	-22264.1
25	-22619.5

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   55    64    69    73    82    91    100
[1]   6     15    24    33
[2]   42    66    75    84    87    93    102
[3]   7     16    25    34    43    52    57
[4]   3     12    21    30    39    48
[5]   54    63    72    76    81    86    90    99    103   106
[6]   0     9     18    27    36    45    78
[7]   4     13    22    31    49    60
[8]   51
[9]   56    59    74    83    92    101   105
[10]  62    71    80    89    94    98    107
[11]  8     17    26    35    44    53    65
[12]  1     10    19    28    37    46    96
[13]  68    77    95    104
[14]  2     11    20    29    38    47    85    88
[15]  5     14    23
[16]  61    67    70    79    97
[17]  32    40    41    50    58
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	**385**	**387**	**380**	**347**	**349**	**366**	**373**	**376**	**379**	**381**
Locus	**[6]**	**[3]**	**[12]**	**[17]**	**[15]**	**[7]**	**[1]**	**[11]**	**[14]**	**[4]**

- values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	P(Y|G)
8	1.38E-38
6	**3.54E-106**
3	**6.68E-107**
12	**2.52E-107**
17	**5.51E-108**
15	**4.93E-108**
7	**5.18E-110**
1	**1.88E-110**
11	**1.93E-111**
14	**8.75E-112**
4	**4.87E-112**
5	4.39E-120
2	3.94E-120
9	1.06E-120
10	8.01E-121
13	6.54E-121
16	5.54E-121
0	4.03E-121

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 6

Simulated

	0	1	2
Disease	0	75	125
Non-Disease	32	108	60

Calculated

	0	1	2
Disease	0	80	120
Non-Disease	32	108	60

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	7	0

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	2
2->	0	2	0

Locus 3

	0	1	2
Disease	0	77	123
Non-Disease	37	87	76

	0	1	2
Disease	0	77	123
Non-Disease	38	88	74

Disease	0	1	2
0->	0	0	0
1->	0	0	1
2->	0	1	0

Non-Disease	0	1	2
0->	0	2	0
1->	3	0	2
2->	0	4	0

Locus 12

	0	1	2
Disease	0	82	118
Non-Disease	34	101	65

	0	1	2
Disease	0	87	113
Non-Disease	37	96	67

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	7	0

Non-Disease	0	1	2
0->	0	0	0
1->	3	0	5
2->	0	3	0

Locus 17

	0	1	2
Disease	0	75	125
Non-Disease	30	101	69

	0	1	2
Disease	0	85	115
Non-Disease	34	100	66

Disease	0	1	2
0->	0	0	0
1->	0	0	7
2->	0	17	0

Non-Disease	0	1	2
0->	0	3	0
1->	7	0	8
2->	0	11	0

Locus 15

	0	1	2
Disease	0	75	125
Non-Disease	30	101	69

	0	1	2
Disease	1	58	141
Non-Disease	22	107	71

Disease	0	1	2
0->	0	0	0
1->	1	0	20
2->	0	4	0

Non-Disease	0	1	2
0->	0	10	0
1->	2	0	8
2->	0	6	0

Locus 7

	0	1	2
Disease	0	83	117
Non-Disease	29	102	69

	0	1	2
Disease	1	85	114
Non-Disease	30	110	60

Disease	0	1	2
0->	0	0	0
1->	1	0	6
2->	0	9	0

Non-Disease	0	1	2
0->	0	3	0
1->	4	0	1
2->	0	10	0

Locus 1

	0	1	2
Disease	0	82	118
Non-Disease	28	98	74

Disease	0	1	2
0->	0	0	0
1->	0	0	9
2->	0	6	0

	0	1	2
Disease	0	79	121
Non-Disease	24	104	72

Non-Disease	0	1	2
0->	0	4	0
1->	0	0	3
2->	0	5	0

Locus 11

	0	1	2
Disease	0	74	126
Non-Disease	27	88	85

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	4	0

	0	1	2
Disease	0	74	126
Non-Disease	29	82	89

Non-Disease	0	1	2
0->	0	1	0
1->	3	0	8
2->	0	4	0

Locus 14

	0	1	2
Disease	0	83	117
Non-Disease	31	91	78

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	4	0

	0	1	2
Disease	0	85	115
Non-Disease	25	98	77

Non-Disease	0	1	2
0->	0	6	0
1->	0	0	4
2->	0	5	0

Locus 4

	0	1	2
Disease	0	83	117
Non-Disease	29	89	82

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	6	0

	0	1	2
Disease	0	86	114
Non-Disease	26	93	81

Non-Disease	0	1	2
0->	0	3	0
1->	0	0	3
2->	0	4	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	A(1)	A(2)	B(0)	B(2)	C(0)	C(1)
Locus 6	0	0	0	0.01	0	0.035
Locus 3	0	0	0	0.005	0	0.005
Locus 12	0	0	0	0.01	0	0.035
Locus 17	0	0	0	0.035	0	0.085
Locus 15	0	0	0.005	0.1	0	0.02
Locus 7	0	0	0.005	0.03	0	0.045
Locus 1	0	0	0	0.045	0	0.03
Locus 11	0	0	0	0.02	0	0.02
Locus 14	0	0	0	0.01	0	0.02
Locus 4	0	0	0	0.015	0	0.03

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	E(1)	E(2)	F(0)	F(2)	H(0)	H(1)
Locus 6	0.005	0	0.005	0.01	0	0.01
Locus 3	0.01	0	0.015	0.01	0	0.02
Locus 12	0	0	0.015	0.025	0	0.015
Locus 17	0.015	0	0.035	0.04	0	0.055
Locus 15	0.05	0	0.01	0.04	0	0.03
Locus 15	0.05	0	0.01	0.04	0	0.03
Locus 7	0.015	0	0.02	0.005	0	0.05
Locus 1	0.02	0	0	0.015	0	0.025
Locus 11	0.005	0	0.015	0.04	0	0.02
Locus 14	0.03	0	0	0.02	0	0.025
Locus 4	0.015	0	0	0.015	0	0.02

6. L=9 M=6 LND=6 MND=3 N=200 75% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-8171.41
4	-7967.31
5	-7900.37
6	-7711.14
7	-7606.42
8	-7480.71
9	-7362.81
10	-7415.68
11	-7513.18
12	-7368.7
13	-7463.97
14	**-7301.46**
15	-7500.35
16	-7365.33
17	-7307.82
18	-7369.06
19	-7544.85
20	-7645.54

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   17    55    61    62    67    68
[1]   57    63    69
[2]   2     11    20    29    38    47
[3]   0     18    27    36    45    58    64    70
[4]   19    37
[5]   6     8     26    35    44    53
[6]   1     10    28    46    59    71
[7]   15    24    33    42
[8]   4     13    22    31    40    49    56
[9]   30
[10]  3     12    21    39    48    65
[11]  7     16    25    34    43    52
[12]  9     51    54    60    66
[13]  5     14    23    32    41    50
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	188	190	193	79	189	189	149	94	195	188
Locus	[8]	[13]	[10]	[6]	[5]	[2]	[4]	[3]	[11]	[8]

- values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	P(Y|G)
9	1.46E-17
12	1.33E-38
8	**1.07E-53**
13	**6.06E-54**
10	**2.49E-55**
6	**2.47E-55**
5	**1.97E-55**
2	**1.22E-55**
4	**7.24E-56**
3	**4.58E-56**
11	**1.72E-56**
7	3.65E-58
0	6.84E-60
1	1.01E-60

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 8

Simulated

	0	1	2
Disease	0	41	59
Non-Disease	18	47	35

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	6	0

Calculated

	0	1	2
Disease	0	45	55
Non-Disease	20	47	33

Non-Disease	0	1	2
0->	0	0	0
1->	2	0	0
2->	0	2	0

Locus 13

	0	1	2
Disease	0	43	57
Non-Disease	18	49	33

Disease	0	1	2
0->	0	0	0
1->	0	0	0
2->	0	2	0

	0	1	2
Disease	0	45	55
Non-Disease	20	45	35

Non-Disease	0	1	2
0->	0	1	0
1->	3	0	3
2->	0	1	0

Locus 10

	0	1	2
Disease	0	40	60
Non-Disease	19	50	31

	0	1	2
Disease	1	40	59
Non-Disease	19	49	32

Disease	0	1	2
0->	0	0	0
1->	1	0	0
2->	0	1	0

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	2
2->	0	1	0

Locus 6

	0	1	2
Disease	0	38	62
Non-Disease	14	48	38

	0	1	2
Disease	17	57	26
Non-Disease	49	41	10

Disease	0	1	2
0->	0	0	0
1->	17	0	2
2->	0	38	0

Non-Disease	0	1	2
0->	0	0	0
1->	34	0	1
2->	1	28	0

Locus 5

	0	1	2
Disease	0	40	60
Non-Disease	14	52	34

	0	1	2
Disease	0	37	63
Non-Disease	11	54	35

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	1	0

Non-Disease	0	1	2
0->	0	3	0
1->	0	0	2
2->	0	1	0

Locus 2

	0	1	2
Disease	0	45	55
Non-Disease	17	48	35

	0	1	2
Disease	1	46	53
Non-Disease	20	44	36

Disease	0	1	2
0->	0	0	0
1->	1	0	1
2->	0	3	0

Non-Disease	0	1	2
0->	0	0	0
1->	3	0	2
2->	0	1	0

Locus 4

	0	1	2
Disease	0	38	62
Non-Disease	14	48	38

	0	1	2
Disease	0	49	51
Non-Disease	17	54	29

Disease	0	1	2
0->	0	0	0
1->	0	0	8
2->	0	19	0

Non-Disease	0	1	2
0->	0	3	0
1->	6	0	3
2->	0	12	0

Locus 3

	0	1	2
Disease	0	34	66
Non-Disease	19	40	41

	0	1	2
Disease	20	42	38
Non-Disease	51	30	19

Disease	0	1	2
0->	0	0	0
1->	20	0	1
2->	0	29	0

Non-Disease	0	1	2
0->	0	0	0
1->	32	0	1
2->	0	23	0

Locus 11

	0	1	2
Disease	0	43	57
Non-Disease	11	55	34

	0	1	2
Disease	0	44	56
Non-Disease	11	53	36

Disease	0	1	2
0->	0	0	0
1->	0	0	0
2->	0	1	0

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	2
2->	0	0	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 8	0	0	0	0.02	0	0.06
Locus 13	0	0	0	0	0	0.02
Locus 10	0	0	0.01	0	0	0.01
Locus 6	0	0	0.17	0.02	0	0.38
Locus 5	0	0	0	0.04	0	0.01
Locus 2	0	0	0.01	0.01	0	0.03
Locus 4	0	0	0	0.08	0	0.19
Locus 3	0	0	0.2	0.01	0	0.29
Locus 11	0	0	0	0	0	0.01

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(0)**	**E(1)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**

Locus 8	0	0	0.02	0	0	0.02
Locus 13	0	0.01	0.03	0.03	0	0.01
Locus 10	0	0.01	0.01	0.02	0	0.01
Locus 6	0	0	0.34	0.01	0.01	0.28
Locus 5	0	0.03	0	0.02	0	0.01
Locus 2	0	0	0.03	0.02	0	0.01
Locus 1	0	0.03	0.06	0.03	0	0.12
Locus 3	0	0	0.32	0.01	0	0.23
Locus 11	0	0.01	0.01	0.02	0	0

7. L=9 M=6 LND=0 MND=0 N=200 100% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-5821.78
4	-5567.75
5	-5399.38
6	-5318.98
7	-5209.67
8	-5092.39
9	**-5040.07**
10	-5290.21
11	-5113.74
12	-5194.67
13	-5333.27
14	-5336.43
15	-5383.7
16	-5399.93
17	-5436.46
18	-5524.4
19	-5349.71
20	-5389.94

- Genes clustered into loci. Each number is a gene and each row is a locus.

```
[0]   0    9    18   27   36   45
[1]   19   28   37
[2]   2    11   20   29   38   47
[3]   1    3    10   12   21   30   39   46   48
[4]   4    13   22   31   40   49
[5]   5    14   23   50
[6]   6    15   24   32   33   41   42   51
[7]   7    16   25   34   43   52
[8]   8    17   26   35   44   53
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	**192**	**153**	**185**	**73**	**192**	**139**	**191**	**189**	**193**
Locus	**[0]**	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**	**[8]**

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 0

Simulated

	0	1	2
Disease	0	46	54
Non-Disease	19	41	40

Calculated

	0	1	2
Disease	0	45	55
Non-Disease	23	36	41

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	1	0

Non-Disease	0	1	2
0->	0	0	0
1->	4	0	1
2->	0	0	0

Locus 1

	0	1	2
Disease	0	35	65
Non-Disease	17	52	31

	0	1	2
Disease	6	26	68
Non-Disease	32	40	28

Disease	0	1	2
0->	0	0	0
1->	6	0	8
2->	0	5	0

Non-Disease	0	1	2
0->	0	0	0
1->	15	0	5
2->	0	8	0

Locus 2

	0	1	2
Disease	0	41	59
Non-Disease	22	39	39

	0	1	2
Disease	0	38	62
Non-Disease	28	35	37

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	1	0

Non-Disease	0	1	2
0->	0	0	0
1->	6	0	1
2->	0	3	0

Locus 3

	0	1	2
Disease	0	38	62
Non-Disease	11	43	46

	0	1	2
Disease	35	17	48
Non-Disease	46	36	18

Disease	0	1	2
0->	0	0	0
1->	34	0	4
2->	1	17	0

Non-Disease	0	1	2
0->	0	0	0
1->	35	0	4
2->	0	32	0

Locus 4

	0	1	2
Disease	0	45	55
Non-Disease	16	41	43

	0	1	2
Disease	0	45	55
Non-Disease	16	43	41

Disease	0	1	2
0->	0	0	0
1->	0	0	1
2->	0	1	0

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	1
2->	0	3	0

Locus 5

	0	1	2
Disease	0	42	58
Non-Disease	9	49	42

Disease	0	1	2
0->	0	0	0
1->	25	0	4
2->	0	3	0

	0	1	2
Disease	25	16	59
Non-Disease	29	31	40

Non-Disease	0	1	2
0->	0	0	0
1->	19	0	4
2->	1	5	0

Locus 6

	0	1	2
Disease	0	43	57
Non-Disease	14	57	29

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	1	0

	0	1	2
Disease	0	41	59
Non-Disease	17	56	27

Non-Disease	0	1	2
0->	0	0	0
1->	3	0	0
2->	0	2	0

Locus 7

	0	1	2
Disease	0	47	53
Non-Disease	15	49	36

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	2	0

	0	1	2
Disease	0	46	54
Non-Disease	16	51	33

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	0
2->	0	3	0

Locus 8

	0	1	2
Disease	0	44	56
Non-Disease	18	50	32

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	2	0

	0	1	2
Disease	0	44	56
Non-Disease	19	47	34

Non-Disease	0	1	2
0->	0	0	0
1->	1	0	2
2->	0	0	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	A(1)	A(2)	B(0)	B(2)	C(0)	C(1)
Locus 0	0	0	0	0.02	0	0.01
Locus 1	0	0	0.06	0.08	0	0.05
Locus 2	0	0	0	0.04	0	0.01
Locus 3	0	0	0.34	0.04	0.01	0.17
Locus 4	0	0	0	0.01	0	0.01
Locus 5	0	0	0.25	0.04	0	0.03
Locus 6	0	0	0	0.03	0	0.01
Locus 7	0	0	0	0.03	0	0.02
Locus 8	0	0	0	0.02	0	0.02

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	E(1)	E(2)	F(0)	F(2)	H(0)	H(1)
Locus 0	0	0	0.04	0.01	0	0
Locus 1	0	0	0.15	0.05	0	0.08
Locus 2	0	0	0.06	0.01	0	0.03
Locus 3	0	0	0.35	0.04	0	0.32
Locus 4	0.01	0	0.01	0.01	0	0.03
Locus 5	0	0	0.19	0.04	0.01	0.05
Locus 6	0	0	0.03	0	0	0.02
Locus 7	0.01	0	0.02	0	0	0.03
Locus 8	0	0	0.01	0.02	0	0

8. L=9 M=6 LND=0 MND=0 N=400 100% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-11692.5
4	-11151.6
5	-10910
6	-10335.1
7	-10061.5
8	-9789.86
9	-9774.42
10	**-9695.73**
11	-9786.47
12	-10016.6
13	-10180.4
14	-9993
15	-10374.6
16	-10410.6
17	-10439.1
18	-10597.1
19	-10758.7
20	-10839.3

- Genes clustered into loci. Each number is a gene and each row is a locus.

[0]	0	9	18	27	36	45
[1]	1	10	19	28	37	46
[2]	2	11	20	29	38	47
[3]	3	12	21	30	39	48
[4]	4	13	22	31	40	49
[5]	5	14	23	32	41	50
[6]	6	15	24	33	42	51
[7]	7	16	25	34	43	52
[8]	8	26	53			
[9]	17	35	44			

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	**383**	**385**	**381**	**385**	**381**	**389**	**383**	**388**	**363**
Locus	**[0]**	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**	**[8]**

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 0

Simulated

	0	1	2
Disease	0	86	114
Non-Disease	32	91	77

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	1	0

Calculated

	0	1	2
Disease	0	84	116
Non-Disease	25	96	79

Non-Disease	0	1	2
0->	0	7	0
1->	0	0	4
2->	0	2	0

Locus 1

	0	1	2
Disease	0	82	118
Non-Disease	34	101	65

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	6	0

	0	1	2
Disease	0	86	114
Non-Disease	35	104	61

Non-Disease	0	1	2
0->	0	0	0
1->	1	0	1
2->	0	5	0

Locus 2

	0	1	2
Disease	0	71	129
Non-Disease	23	102	75

Disease	0	1	2
0->	0	0	0
1->	1	0	3
2->	0	5	0

	0	1	2
Disease	1	72	127
Non-Disease	27	98	75

Non-Disease	0	1	2
0->	0	0	0
1->	4	0	3
2->	0	3	0

Locus 3

	0	1	2
Disease	0	85	115
Non-Disease	37	95	68

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	3	0

	0	1	2
Disease	0	86	114
Non-Disease	39	93	68

Non-Disease	0	1	2
0->	0	2	0
1->	4	0	2
2->	0	2	0

Locus 4

	0	1	2
Disease	0	82	118
Non-Disease	29	93	78

	0	1	2
Disease	0	82	118
Non-Disease	29	92	79

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	4	0

Non-Disease	0	1	2
0->	0	3	0
1->	3	0	3
2->	0	2	0

Locus 5

	0	1	2
Disease	0	82	118
Non-Disease	39	102	59

	0	1	2
Disease	0	81	119
Non-Disease	40	98	62

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	1	0

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	4
2->	0	1	0

Locus 6

	0	1	2
Disease	0	93	107
Non-Disease	31	99	70

	0	1	2
Disease	0	93	107
Non-Disease	30	102	68

Disease	0	1	2
0->	0	0	0
1->	0	0	5
2->	0	5	0

Non-Disease	0	1	2
0->	0	2	0
1->	1	0	1
2->	0	3	0

Locus 7

	0	1	2
Disease	0	87	113
Non-Disease	30	90	80

	0	1	2
Disease	0	87	113
Non-Disease	31	88	81

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	2	0

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	3
2->	0	2	0

Locus 8

	0	1	2
Disease	0	74	126
Non-Disease	39	101	60

	0	1	2
Disease	0	65	135
Non-Disease	36	105	59

Disease	0	1	2
0->	0	0	0
1->	0	0	14
2->	0	5	0

Non-Disease	0	1	2
0->	0	5	0
1->	2	0	5
2->	0	6	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 0	0	0	0	0.026667	0	0.02
Locus 1	0	0	0	0.01	0	0.01
Locus 2	0	0	0	0.016667	0	0.01
Locus 3	0	0	0.003333	0.026667	0	0.02
Locus 4	0	0	0.003333	0.003333	0	0.016667
Locus 5	0	0	0	0.026667	0	0.023333
Locus 6	0	0	0	0.016667	0	0.01
Locus 7	0	0	0	0.02	0	0.023333
Locus 8	0	0	0	0.023333	0	0.02

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 0	0.046667	0	0.003333	0.01	0	0.006667
Locus 1	0.013333	0	0.013333	0.016667	0	0.01
Locus 2	0.013333	0	0.013333	0.01	0	0.003333
Locus 3	0.006667	0	0.01	0.02	0	0.006667
Locus 4	0.01	0	0.01	0.01	0	0.003333
Locus 5	0.006667	0	0.013333	0.016667	0	0.016667
Locus 6	0.01	0	0.016667	0.02	0	0.01
Locus 7	0.01	0	0	0.016667	0	0.013333
Locus 8	0.013333	0	0	0.023333	0	0.016667

9. L=9 M=6 LND=0 MND=0 N=600 100% disease genes

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	**Likelihood**
3	-17776.7
4	-17051.8
5	-16475.2
6	-16141.3
7	-15585.2
8	-14990.9
9	**-14605.4**
10	-14869.5
11	-15009.8
12	-14830.9
13	-15318.1
14	-15448.8
15	-15725.4
16	-15851
17	-15710.6
18	-16158.8
19	-15883.9
20	-16024.6

- Genes clustered into loci. Each number is a gene and each row is a locus

```
[0]   0     9     18    27    36    45
[1]   1     10    19    28    37    46
[2]   2     11    20    29    38    47
[3]   3     12    21    30    39    48
[4]   4     13    22    31    40    49
[5]   5     14    23    32    41    50
[6]   6     15    24    33    42    51
[7]   7     16    25    34    43    52
[8]   8     17    26    35    44    53
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	566	578	580	572	583	569	575	575	571
Locus	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 0

Simulated

	0	1	2
Disease	0	124	176
Non-Disease	59	136	105

Calculated

	0	1	2
Disease	0	122	178
Non-Disease	46	148	106

Disease	0	1	2
0->	0	0	0
1->	0	0	8
2->	0	6	0

Non-Disease	0	1	2
0->	0	14	0
1->	1	0	3
2->	0	2	0

Locus 1

	0	1	2
Disease	0	130	170
Non-Disease	54	144	102

	0	1	2
Disease	0	130	170
Non-Disease	54	142	104

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	3	0

Non-Disease	0	1	2
0->	0	4	0
1->	4	0	5
2->	0	3	0

Locus 2

	0	1	2
Disease	0	131	169
Non-Disease	43	138	119

	0	1	2
Disease	0	129	171
Non-Disease	43	136	121

Disease	0	1	2
0->	0	0	0
1->	0	0	5
2->	0	3	0

Non-Disease	0	1	2
0->	0	4	0
1->	4	0	3
2->	0	1	0

Locus 3

	0	1	2
Disease	0	108	192
Non-Disease	47	145	108

	0	1	2
Disease	1	105	194
Non-Disease	48	140	112

Disease	0	1	2
0->	0	0	0
1->	1	0	8
2->	0	6	0

Non-Disease	0	1	2
0->	0	2	0
1->	3	0	6
2->	0	2	0

Locus 4

	0	1	2
Disease	0	128	172
Non-Disease	53	136	111

	0	1	2
Disease	1	131	168
Non-Disease	53	134	113

Disease	0	1	2
0->	0	0	0
1->	1	0	1
2->	0	5	0

Non-Disease	0	1	2
0->	0	3	0
1->	3	0	3
2->	0	1	0

Locus 5

	0	1	2
Disease	0	119	181
Non-Disease	58	140	102

	0	1	2
Disease	0	118	182
Non-Disease	60	138	102

Disease	0	1	2
0->	0	0	0
1->	0	0	8
2->	0	7	0

Non-Disease	0	1	2
0->	0	2	0
1->	4	0	5
2->	0	5	0

Locus 6

	0	1	2
Disease	0	107	193
Non-Disease	60	136	104

	0	1	2
Disease	0	105	195
Non-Disease	62	131	107

Disease	0	1	2
0->	0	0	0
1->	0	0	5
2->	0	3	0

Non-Disease	0	1	2
0->	0	3	0
1->	5	0	6
2->	0	3	0

Locus 7

	0	1	2
Disease	0	109	191
Non-Disease	43	146	111

	0	1	2
Disease	0	110	190
Non-Disease	40	148	112

Disease	0	1	2
0->	0	0	0
1->	0	0	6
2->	0	7	0

Non-Disease	0	1	2
0->	0	3	0
1->	0	0	5
2->	0	4	0

Locus 8

	0	1	2
Disease	0	129	171
Non-Disease	39	142	119

	0	1	2
Disease	0	128	172
Non-Disease	35	144	121

Disease	0	1	2
0->	0	0	0
1->	0	0	7
2->	0	6	0

Non-Disease	0	1	2
0->	0	4	0
1->	0	0	7
2->	0	5	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**

Locus 0	0	0	0	0.026667	0	0.02
Locus 1	0	0	0	0.01	0	0.01
Locus 2	0	0	0	0.016667	0	0.01
Locus 3	0	0	0.003333	0.026667	0	0.02
Locus 4	0	0	0.003333	0.003333	0	0.016667
Locus 5	0	0	0	0.026667	0	0.023333
Locus 6	0	0	0	0.016667	0	0.01
Locus 7	0	0	0	0.02	0	0.023333
Locus 8	0	0	0	0.023333	0	0.02

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 0	0.046667	0	0.003333	0.01	0	0.006667
Locus 1	0.013333	0	0.013333	0.016667	0	0.01
Locus 2	0.013333	0	0.013333	0.01	0	0.003333
Locus 3	0.006667	0	0.01	0.02	0	0.006667
Locus 4	0.01	0	0.01	0.01	0	0.003333
Locus 5	0.006667	0	0.013333	0.016667	0	0.016667
Locus 6	0.01	0	0.016667	0.02	0	0.01
Locus 7	0.01	0	0	0.016667	0	0.013333
Locus 8	0.013333	0	0	0.023333	0	0.016667

L=9 M=6 LND=9 MND=6 NDE=10 N=400

10. Genes directly associated with disease added.

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-28017.5
4	-27892.7
5	-27477.6
6	-27122.4
7	-26958.1
8	-26423.7
9	-26652.7
10	-25884.6
11	-26057.8
12	-25815.6
13	**-24748.9**
14	-24762
15	-24970.1
16	-24911.7
17	-24983.7
18	-25322.4
19	-24848.4
20	-24773.7
21	-25249.2
22	-25504.1

- Genes clustered into loci. Each number is a gene and each row is a locus

```
[0]   6    15   24   33   42   51
[1]   0    9    18   27   36   45   67   70   114
[2]   7    43   61   62   79   88   97   106  115
[3]   5    14   23   32   41   50   72   103  117
[4]   37   71   80   81   89   90   98   102  107  108  112  116
[5]   1    19   28   46   73   91   94   100  109
[6]   4    13   22   31   40   49   69   105
[7]   65   74   76   78   83   92   101  110
[8]   2    11   20   29   38   47   96
[9]   3    12   21   30   39   48   111
[10]  10   52   68   77   84   86   87   95   104  113
[11]  8    17   25   26   35   44   53   64   82   99
[12]  16   34   54   55   56   57   58   59   60   63   66   75
      85   93
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	126	118	372	383	379	387	380	385	384	203
Locus	[2]	[12]	[11]	[1]	[8]	[0]	[6]	[3]	[9]	[5]

- values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	P(Y|G)
2	**1**
12	**1**
11	**1.23E-103**
1	**5.22E-106**
8	**2.13E-107**
0	**3.23E-108**
6	**1.46E-108**
3	**8.40E-110**
9	**8.13E-112**
5	**3.32E-113**
4	3.47E-120
10	4.62E-121
7	4.35E-121

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 2

Simulated

	0	1	2
Disease	0	77	123
Non-Disease	24	86	90

Disease	0	1	2
0->	0	0	0
1->	0	0	27
2->	0	71	0

Calculated

	0	1	2
Disease	0	121	79
Non-Disease	200	0	0

Non-Disease	0	1	2
0->	0	0	0
1->	86	0	0
2->	90	0	0

Locus 12

	0	1	2
Disease	0	77	123
Non-Disease	24	86	90

Disease	0	1	2
0->	0	0	0
1->	0	0	33
2->	0	73	0

	0	1	2
Disease	0	117	83
Non-Disease	200	0	0

Non-Disease	0	1	2
0->	0	0	0
1->	86	0	0
2->	90	0	0

Locus 11

	0	1	2
Disease	0	84	116
Non-Disease	37	111	52

Disease	0	1	2
0->	0	0	0
1->	0	0	10
2->	0	4	0

	0	1	2
Disease	0	78	122
Non-Disease	36	109	55

Non-Disease	0	1	2
0->	0	3	0
1->	2	0	6
2->	0	3	0

Locus 1

	0	1	2
Disease	0	81	119
Non-Disease	39	89	72

Disease	0	1	2
0->	0	0	0
1->	0	0	7
2->	0	3	0

	0	1	2
Disease	0	77	123
Non-Disease	40	88	72

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	2
2->	0	2	0

Locus 8

	0	1	2
Disease	0	78	122
Non-Disease	28	96	76

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	2	0

	0	1	2
Disease	0	78	122
Non-Disease	36	91	73

Non-Disease	0	1	2
0->	0	0	0
1->	8	0	3
2->	0	6	0

Locus 0

	0	1	2
Disease	0	69	131
Non-Disease	29	94	77

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	1	0

	0	1	2
Disease	0	68	132
Non-Disease	31	90	79

Non-Disease	0	1	2
0->	0	1	0
1->	3	0	4
2->	0	2	0

Locus 6

	0	1	2
Disease	0	85	115
Non-Disease	31	97	72

	0	1	2
Disease	0	83	117
Non-Disease	34	93	73

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	2	0

Non-Disease	0	1	2
0->	0	3	0
1->	6	0	3
2->	0	2	0

Locus 3

	0	1	2
Disease	0	85	115
Non-Disease	37	92	71

	0	1	2
Disease	1	84	115
Non-Disease	36	92	72

Disease	0	1	2
0->	0	0	0
1->	1	0	3
2->	0	3	0

Non-Disease	0	1	2
0->	0	2	0
1->	1	0	3
2->	0	2	0

Locus 9

	0	1	2
Disease	0	87	113
Non-Disease	25	107	68

	0	1	2
Disease	1	85	114
Non-Disease	24	111	65

Disease	0	1	2
0->	0	0	0
1->	1	0	2
2->	0	1	0

Non-Disease	0	1	2
0->	0	3	0
1->	2	0	2
2->	0	5	0

Locus 5

	0	1	2
Disease	0	71	129
Non-Disease	36	83	81

	0	1	2
Disease	27	89	84
Non-Disease	78	52	70

Disease	0	1	2
0->	0	0	0
1->	27	0	20
2->	0	65	0

Non-Disease	0	1	2
0->	0	0	0
1->	42	0	16
2->	0	27	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 2	0	0	0	0.135	0	0.355
Locus 12	0	0	0	0.165	0	0.365
Locus 11	0	0	0	0.05	0	0.02
Locus 1	0	0	0	0.035	0	0.015

Locus 8	0	0	0	0.01	0	0.01
Locus 0	0	0	0	0.01	0	0.005
Locus 6	0	0	0	0.02	0	0.01
Locus 3	0	0	0.005	0.015	0	0.015
Locus 9	0	0	0.005	0.01	0	0.005
Locus 5	0	0	0.135	0.1	0	0.325

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 2	0	0	0.43	0	0.45	0
Locus 12	0	0	0.43	0	0.45	0
Locus 11	0.015	0	0.01	0.03	0	0.015
Locus 1	0.005	0	0.01	0.01	0	0.01
Locus 8	0	0	0.04	0.015	0	0.03
Locus 0	0.005	0	0.015	0.02	0	0.01
Locus 6	0.015	0	0.03	0.015	0	0.01
Locus 3	0.01	0	0.005	0.015	0	0.01
Locus 9	0.015	0	0.01	0.01	0	0.025
Locus 5	0	0	0.21	0.08	0	0.135

11. Overall Example

L=9 M=6 genes from cancer dataset MND=0 N=400 <50% disease genes = 12624

- Vary number of initial clusters (W) to determine best number of locus clusters (bold is W chosen)

W	Likelihood
3	-18888.5
4	-18626.4
5	-18424.2
6	-18506.1
7	-18534.1
8	-18506.8
9	-18497.4
10	-18508.8
11	-18486.1
12	-18505.4
13	-18523.2
14	-18444.7
15	-18454.1
16	-18427.4
17	-17625.2
18	-18888.5
19	-16777.7
20	-16254.4
21	-15965.7
22	**-15613.7**
23	-16341.9
24	-16552.1
25	-16440.4

- Genes clustered into loci. Each number is a gene and each row is a locus

```
[0]     5573
[1]     13
[2]     1     10    19    28    37    46
[3]     6     15    33    42    51
[4]     9798
[5]     1697  3419
[6]     4754  7030
[7]     0     9     18    27    36    45
[8]     5     14    23    32    41    50
[9]     3383
[10]    2     11    20    29    38    47
[11]    8     35    44
[12]    24
[13]    5973  9012  8709
[14]    7     16    34    43    52
```

```
[15]    2029
[16]    1295  2281  3887  7639
[17]    653
[18]    3     12    21    30    39   48
[19]    4     22    31    49
[20]    1065
[21]    17    26    53
```

- Number of individuals out of N that are correctly genotyped

# of Correct Individuals	222	384	343	383	389	384	380	382	366	350	362
Locus	[1]	[2]	[11]	[18]	[10]	[7]	[3]	[8]	[19]	[21]	[14]

- values of $P\left(\vec{Y} \mid \vec{G}\right)$ to determine which locus clusters are associated with disease and genes are well clustered.

LOCUS	P(Y|G)
1	**3.54E-35**
16	**6.00E-101**
2	**9.79E-105**
11	**7.29E-107**
18	**4.18E-107**
10	**2.45E-107**
7	**1.55E-107**
3	**1.54E-107**
8	**1.03E-108**
19	**4.39E-109**
21	**3.02E-110**
14	**6.64E-111**
6	**2.11E-111**
12	3.21E-115
5	4.48E-116
17	7.06E-117
13	5.13E-117
20	4.81E-118
9	3.70E-118
0	2.69E-119
4	6.25E-120
15	8.54E-121

- Contingency Tables of Simulated data and Calculated data
- Tables of Genotyping Errors (i.e., 1->2 means 1 in simulated moves to 2 in calculated)

Locus 1

Simulated

	0	1	2
Disease	0	96	104
Non-Disease	35	92	73

Calculated

	0	1	2
Disease	0	122	78
Non-Disease	157	0	43

Disease	0	1	2
0->	0	0	0
1->	0	0	12
2->	0	38	0

Non-Disease	0	1	2
0->	0	0	0
1->	86	0	6
2->	36	0	0

Locus 2

	0	1	2
Disease	0	66	134
Non-Disease	29	102	69

	0	1	2
Disease	0	68	132
Non-Disease	36	96	68

Disease	0	1	2
0->	0	0	0
1->	0	0	1
2->	0	3	0

Non-Disease	0	1	2
0->	0	0	0
1->	7	0	2
2->	0	3	0

Locus 11

	0	1	2
Disease	0	91	109
Non-Disease	32	98	70

	0	1	2
Disease	0	93	107
Non-Disease	35	109	56

Disease	0	1	2
0->	0	0	0
1->	0	0	11
2->	0	13	0

Non-Disease	0	1	2
0->	0	5	0
1->	8	0	3
2->	0	17	0

Locus 18

	0	1	2
Disease	0	83	117
Non-Disease	35	97	68

	0	1	2
Disease	0	84	116
Non-Disease	37	95	68

Disease	0	1	2
0->	0	0	0
1->	0	0	3
2->	0	4	0

Non-Disease	0	1	2
0->	0	1	0
1->	3	0	3
2->	0	3	0

Locus 10

	0	1	2
Disease	0	81	119
Non-Disease	35	93	72

	0	1	2
Disease	0	80	120
Non-Disease	36	93	71

Disease	0	1	2
0->	0	0	0
1->	0	0	2
2->	0	1	0

Non-Disease	0	1	2
0->	0	1	0
1->	2	0	2
2->	0	3	0

Locus 7

	0	1	2
Disease	0	79	121
Non-Disease	34	103	63

Disease	0	1	2
0->	0	0	0
1->	1	0	2
2->	0	4	0

	0	1	2
Disease	1	80	119
Non-Disease	34	106	60

Non-Disease	0	1	2
0->	0	1	0
1->	1	0	2
2->	0	5	0

Locus 3

	0	1	2
Disease	0	82	118
Non-Disease	34	91	75

Disease	0	1	2
0->	0	0	0
1->	0	0	6
2->	0	3	0

	0	1	2
Disease	0	79	121
Non-Disease	37	88	75

Non-Disease	0	1	2
0->	0	1	0
1->	4	0	3
2->	0	3	0

Locus 8

	0	1	2
Disease	0	87	113
Non-Disease	32	101	67

Disease	0	1	2
0->	0	0	0
1->	0	0	4
2->	0	7	0

	0	1	2
Disease	0	90	110
Non-Disease	31	106	63

Non-Disease	0	1	2
0->	0	1	0
1->	0	0	1
2->	0	5	0

Locus 19

	0	1	2
Disease	0	96	104
Non-Disease	35	92	73

Disease	0	1	2
0->	0	0	0
1->	0	0	7
2->	0	3	0

	0	1	2
Disease	0	92	108
Non-Disease	37	90	73

Non-Disease	0	1	2
0->	0	4	0
1->	6	0	7
2->	0	7	0

Locus 21

	0	1	2
Disease	0	91	109
Non-Disease	32	98	70

	0	1	2
Disease	2	88	110
Non-Disease	39	95	66

Disease	0	1	2
0->	0	0	0
1->	2	0	11
2->	0	10	0

Non-Disease	0	1	2
0->	0	3	0
1->	10	0	5
2->	0	9	0

Locus 14

	0	1	2
Disease	0	79	121
Non-Disease	31	98	71

	0	1	2
Disease	4	74	122
Non-Disease	40	93	67

Disease	0	1	2
0->	0	0	0
1->	4	0	5
2->	0	4	0

Non-Disease	0	1	2
0->	0	3	0
1->	12	0	3
2->	0	7	0

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Diseased Individuals]
- A = genotyping error at genotype 0 diseased for diseased individuals
- B = genotyping error at genotype 1 diseased for diseased individuals (B(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- C = genotyping error at genotype 2 diseased for diseased individuals

Disease	**A(1)**	**A(2)**	**B(0)**	**B(2)**	**C(0)**	**C(1)**
Locus 1	0	0	0	0.06	0	0.19
Locus 2	0	0	0	0.005	0	0.015
Locus 11	0	0	0	0.055	0	0.065
Locus 18	0	0	0	0.015	0	0.02
Locus 10	0	0	0	0.01	0	0.005
Locus 7	0	0	0.005	0.01	0	0.02
Locus 3	0	0	0	0.03	0	0.015
Locus 8	0	0	0	0.02	0	0.035
Locus 19	0	0	0	0.035	0	0.015
Locus 21	0	0	0.01	0.055	0	0.05
Locus 14	0	0	0.02	0.025	0	0.02

- Table of Percentage of Genotyping Errors [# of Genotyping Errors/# of Non-Diseased Individuals]
- E = genotyping error at genotype 0 non-diseased for diseased individuals
- F = genotyping error at genotype 1 non-diseased for diseased individuals (F(2) is % of genotyping errors that move to genotype 2 from genotype 1, etc.
- H = genotyping error at genotype 2 non-diseased for diseased individuals

Non-Disease	**E(1)**	**E(2)**	**F(0)**	**F(2)**	**H(0)**	**H(1)**
Locus 1	0	0	0.43	0.03	0.18	0
Locus 2	0	0	0.035	0.01	0	0.015
Locus 11	0.025	0	0.04	0.015	0	0.085
Locus 18	0.005	0	0.015	0.015	0	0.015

Locus 10	0.005	0	0.01	0.01	0	0.015
Locus 7	0.005	0	0.005	0.01	0	0.025
Locus 3	0.005	0	0.02	0.015	0	0.015
Locus 8	0.005	0	0	0.005	0	0.025
Locus 19	0.02	0	0.03	0.035	0	0.035
Locus 21	0.015	0	0.05	0.025	0	0.045