STATISTICAL METHODS FOR RECOVERING GWAS DATA

by

Umut Özbek

B.S. in Statistics, Middle East Technical University, Turkey, 2004

M.S. in Biostatistics, Ankara University, Turkey, 2007

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Umut Özbek

It was defended on

February 6, 2013

and approved by

Vincent C. Arena, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Yan Lin, Ph.D., Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Wei Chen, Ph.D., Assistant Professor, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC

Dissertation Advisor: Eleanor Feingold, Ph.D., Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Co-advisor: Daniel E. Weeks, Ph.D., Professor, Departments of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, PhD

STATISTICAL METHODS FOR RECOVERING GWAS DATA

Umut Özbek, Ph.D.

University of Pittsburgh, 2013

ABSTRACT

Genome-wide association studies (GWAS) are used to investigate associations between genetic variants and health and disease. GWAS often use a "chip" to genotype single nucleotide polymorphisms (SNPs) spanning the genome and test for association between genotype and phenotype at each SNP. The topic of this dissertation is methods for recovering some of the markers that are typically discarded or not analyzed in a GWAS. During data cleaning, prior to the statistical analysis, many genetic markers are discarded because they fail to meet standard quality control criteria. In addition, some analysis results are filtered out because they are considered unreliable. However, there are some discarded data that could be recovered and used in the analysis. For instance, markers that fail to meet a cutoff p-value for the Hardy-Weinberg equilibrium (HWE) test are typically discarded, as are markers with minor allele frequency below some arbitrary cutoff. In addition, markers on the X-chromosome are often not analyzed because sex chromosome analyses are not as straightforward as autosomal analyses, and the statistical methods for testing association on X-chromosome markers are not well established or well tested. In order to make use of more information from any given GWAS, the standard

quality control criteria could be modified and more flexible and data specific analysis methods could be developed. This should have the potential to increase power.

Genetic variation and environmental/behavioral factors interact to cause almost all human diseases with great public health significance. Refinements in data analysis will help improve the process of identifying the genetic factors and their interactions. These are important in order to provide better prevention and treatment for human diseases so to maintain public health. The methods in this dissertation are proposed and investigated to help with providing a better public health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

BAF ..... B allele frequency

CART ..... Classification and Regression Trees

CNVs ..... Copy number variations

dbGAP ..... Database of genotypes and phenotypes

DNA ..... Deoxyribonucleic acid

FPR ..... False positive rate

GENEVA ..... Gene and environment association studies

GWAS ..... Genome-wide association study

HWE ..... Hardy-Weinberg equilibrium

LD ..... Linkage disequilibrium

LRR ..... LogRRatio

MAF ..... Minor allele frequency

PGL ..... Posterior genotype likelihoods

SD ..... Standard deviation

SNPs ..... Single nucleotide polymorphisms

SVM ..... Support Vector Machines

TPR ..... True positive rate

PREFACE

I would like to express my deepest gratitude to my advisors, Dr. Eleanor Feingold and Dr. Daniel E. Weeks, for their guidance and support through my doctoral studies. I also would like to thank to the rest of my dissertation committee members Dr. Vincent Arena, Dr. Yan Lin, and Dr. Wei Chen for their insightful comments and suggestions.

I also owe a huge debt of gratitude to my friends Mine Bostancı, Özge Zelal Aydın, and Duygu Selçuklu, who stood by me through the good and bad times.

I am forever indebted to my wonderful parents Mehtap and Zafer Özbek and my sister Gizem Özbek for all their love and support.

This dissertation is dedicated to the memory of Mehmet Ülker.

1.0 INTRODUCTION

1.1 GENOME-WIDE ASSOCIATION STUDIES

Genome-wide association studies are used to understand the interaction among genetic, environmental and behavioral effects on health and disease by searching the genome for small variations occurring more frequently in people having the trait of interest. Deoxyribonucleic acid (DNA) is the hereditary molecule in humans and most other living organisms. DNA is usually obtained from participants' blood samples or cheek cells. Then genotypes of each participant are determined by using chips and laboratory machines. The machines assay the individuals' DNA for selected markers of genetic variation. These markers are called single nucleotide polymorphisms (SNPs). In order to test for association between a SNP and a trait phenotype, chi-squared tests or regression models can be used. If the genetic variations at a SNP occur significantly more frequently in people having a trait than in people not having the trait, then those variations are said to be associated with the trait. The associated variants are not necessarily the cause of the trait. However, they may be in the same region as the actual causal variants. Therefore, after finding a significant association, researchers determine the sequence of nucleotides in that region of DNA to identify the relevant genetic changes. After discovering the variants and studying their functions, they can use the information to diagnose, treat, and prevent the disease.

1.1.1 Genotype calling

Genotype calling is the process of deciding the genotype of an individual based on a DNA sample. In order to call genotypes, genotyping algorithms use intensity values X and Y for each SNP. To scale the data, intensities are normalized. After normalization, homozygous genotypes lie along the X and Y intensity axes [Staaf, et al. 2008]. Widely used genotype calling algorithms assume that there are two possible alleles for each SNP, which are called as biallelic. However, that may not be the case for all markers. There may be some SNPs that have more than two alleles. If this were the case for a SNP, then it would fail to meet the standard quality control criteria, which are applied to the genotype data at the beginning of a GWAS to identify and discard 'bad' SNPs. There may be some good but non-standard SNPs that fail the standard quality control criteria, such as SNPs having more than two possible alleles. In this dissertation, I focus on 'null-allele SNPs', SNPs having an extra N allele. Null alleles are non-amplified alleles and are usually caused by a mutation in the primer binding region [Callen, et al. 1993; Pemberton, et al. 1995]. Instead of discarding those null-allele SNPs, which might be quite common [Lehmann, et al. 1996], the information they carry should be analyzed to investigate possible associations and/or explore regions of copy number variation.

There are some approaches to null-allele SNPs in the literature. Franke et al. [2008] proposed a method "TriTyper" that can detect genotypes in case-control datasets for deletion copy number variations (CNVs) or SNPs with an extra, uncalled (null) allele. There is another study by Kumasaka et al [Kumasaka, et al. 2011], which mainly focuses on genotyping copy number polymorphisms.

1.1.2 GWAS data cleaning

After getting raw data based on genotype calling algorithms, the quality control steps are applied to clean the raw data [Zheng 2012]. First, contaminated DNA, low-quality SNPs, and low quality genotype calls are excluded. Then the quality measures are used for further cleaning. For SNP quality checking, the measures such as missing call rate over samples, supplicate sample discordance, Hardy-Weinberg equilibrium (HWE), minor allele frequency (MAF) can be used. For sample quality checking, the measures such as missing call rate over SNPs, B allele frequency (BAF), logRRatio (LRR), confidence score, and heterozygosity can be used [Laurie, et al. 2010].

Missing call rate is an important indicator of data quality. *Missing call rate per SNP* is the percentage of individuals uncalled for that SNP, and SNPs having 2% or higher missing rates are usually excluded from analyses. *Call rate per individual* is the percentage of SNPs uncalled for that individual. Usually an individual having call rate lower than 95% is discarded from analyses. Moreover, if there is a huge difference (magnitude would be specific to project) in call rates between genders, then those SNPs would fail.

B allele frequency (BAF) is calculated to measure the allelic imbalance. It is used to check for sample mixtures and chromosomal aberrations. BAF is the estimator of B allele frequency from a single individual. For each chromosome and each individual, the variance of BAF is calculated over all heterozygote SNPs. Also, sub chromosomal abbreviations can be detected by dividing the chromosome into sections having same number of SNPs in each and calculating the BAF variances in a window of two adjacent sections. If an individual having a variance greater than four standard deviations from the mean of all samples, then the BAF of that individual and that chromosome is investigated further. BAF plots, which is BAF versus

chromosomal position, for sample-chromosome combinations where there is a high intensity relative to other chromosomes are investigated for aneuploidy [Laurie, et al. 2010]. In cases of allelic balance, the true frequency would be 0, ½ or 1. In cases of imbalance, the frequency may vary.

LogRRatio (LRR) measures the relative intensity. LRR is used for detecting chromosomal aberrations with SNP array data [Conlin, et al. 2010; Peiffer, et al. 2006]. It is calculated as observed over expected value of R in logarithm base 2 where R is the sum of the normalized allelic probe intensities produced by SNP assays.

Confidence score measures the distance between a data point and the centroid of the closest genotype cluster in a genotype plot. This score is in (0,1) range. And if 1 indicates best, the data points having less than a certain value are set as missing.

Heterozygosity is the fraction of non-missing heterozygous genotype calls for an individual or a SNP. If there is a huge difference in heterozygosity between genders, those SNPs would fail.

Genotype concordance or *discordance rate* is calculated per individual as well as per SNP by using independently genotyped duplicate samples. For an individual, the discordance is the fraction of genotype calls that differ in duplicate pairs over all non-missing SNPs. For a SNP, the discordance rate is the number of calls that differ divided by the total number of opportunities to detect a difference.

Genomic inflation factor is calculated as the median observed over median expected statistic for a set of genome-wide tests under the null hypothesis [Devlin and Roeder 1999].

Minor allele frequency (MAF) for a SNP can be calculated by using only controls or cases and controls together. In general if a SNP has MAF less than 1%, it is removed from analyses.

Hardy-Weinberg equilibrium p-value (HWE) tests if the HWE assumption under biallelic SNP model is violated. The threshold for HWE p-value can differ depending on the dataset [Wakefield 2010].

Among the quality control measures, MAF and HWE can be considered further to prevent losing potential valuable information. There are powerful analysis tools for SNPs having low MAF, in other words rare variants [Li and Leal 2008; Morris and Zeggini 2010]. Moreover, rare variant analysis should be investigated for gonosomal SNPs, where male and female genotypes may have different properties. If a SNP fail to meet a standard biallelic HWE criterion, it does not necessarily mean that the SNP is bad or does not carry significant information for the trait. The failing SNP may fit a multi-allelic model and it can be analyzed by using appropriate techniques for multi-allelic models.

1.1.3 GWAS analysis process

GWAS analysis usually starts with single SNP testing. For binary traits, allele-based tests comparing the frequencies in cases and controls, trend tests for additive models, chi-squared tests, or robust tests can be used. For quantitative traits, regression models are preferred. The p-value of each SNP is compared to a prespecified genome-wide significance level. The p-values are ranked and the top-ranked SNPs can be considered as candidates for association. However, SNPs with the lowest p-values may not correspond to the true associations. There are some other factors affecting the ranks of p-values such as MAF, genetic models, and sample size.

After finding candidate SNPs, it is important to use independent case-control data coming from the same population as the initial study to confirm the initial phenotype-genotype association [Chanock, et al. 2007]. Multi-stage sampling or meta analysis can be used for replication study. Ideally, in the replication stage, the same phenotype, same statistic and same genetic model should be used as in the initial study [Zheng 2012].

1.2 X CHROMOSOME ANALYSIS

In a GWAS, data on the X chromosome should be analyzed carefully. If the data is from the pseudo-autosomal regions, where genes are inherited like autosomal genes, then it is treated as data from autosomal regions. Otherwise, because females carry two copies but males carry only one copy of the X chromosome, regular autosomal association methods may not be used. In order to analyze X chromosomal data, analysis methods taking into account its specific features are needed. A recent study showed that between January 2010 and March 2012, only 33% of the GWA papers reported X chromosome results [Wise, et al. 2011]. First of all, the scaling of female and male genotypes can be an issue. Under the additive model, female genotypes AA, AB, and BB are coded as 0, 1, and 2 respectively. However, the appropriate coding of the male genotypes A and B is not as obvious. Clayton [Clayton 2008] suggests that male genotypes should be scaled as homozygote females for additive models but there are other methods for X chromosome analysis using a different scaling. Some analysis programs code them as 0 and 1, whereas some other programs treat males as homozygous females, and so code them as 0 and 2. Another issue to be considered is sex-specific phenotype variances. If phenotype variances differ by sex, then regular analysis methods, which estimate an overall variance for everyone, may be

insufficient. In this case, the test statistics should estimate sex-specific variances, or a sex effect should be included in the analysis model. Moreover, allele distributions may not be same for males and females. Therefore, an overall allele frequency estimate may not be appropriate. Finally, the female/male and/or case/control ratio in the dataset may affect the analysis results. In the literature, because of all these issues for X chromosome data, either the X chromosome is not analyzed or reported, or even if it is analyzed, the analysis methods may not be appropriate.

1.3 CONTRIBUTION OF THIS DISSERTATION

In this dissertation, I aimed to recover GWAS data, which are usually discarded from analyses. First, I propose a procedure to classify SNPs genome-wide at the beginning of a study. By applying this procedure, a researcher will be able to identify non-standard, but informative 'null-allele' SNPs, which is a SNP having an extra 'null' allele besides the regular two alleles, as well as standard SNPs. Not discarding those non-standard SNPs is important not only because they can be investigated for associations by analyzing under different models, but also we can explore the regions for copy number variations. Although there are other methods to find SNPs with an extra allele, the procedure in this dissertation focuses only on null-allele SNPs. I applied three widely-used statistical classifiers in this procedure. I provide a model, which can be easily applied to genotyped data.

Secondly, I worked on X-chromosome data, which is not analyzed or reported often. Because of the special features of X chromosome, specific analysis methods are needed. There are some X-chromosome test statistics proposed. However, it is not obvious which method a researcher should use. I aimed to fill the gap in evaluating X chromosome association statistics.

Although there are some studies comparing the test statistics in the literature, I comprehensively evaluated two of the best X chromosome test statistics and compared them with regression models. Moreover, I ran real chromosome-wide data for type I error analyses. The report I provide will be helpful for determining what are the best X chromosome data analysis methods under different conditions.

2.0 EFFICIENT IDENTIFICATION OF NULL-ALLELE SINGLE NUCLEOTIDE POLYMORPHISM MARKERS

2.1 ABSTRACT

At the beginning of a genome-wide association study (GWAS), many markers are discarded because they fail to meet standard quality control criteria. Some of these markers are out of Hardy-Weinberg Equilibrium (HWE) because they have 'null alleles' (which may be deletions or third alleles that do not hybridize to standard probes). It may be useful to identify null-allele markers so that they can be analyzed under different models or in order to explore regions of copy number variation. We present a model for the genotype data that are produced when a null-allele SNP is genotyped under standard (2-allele) assumptions. We show that this model can be combined with the standard HWE model to develop classification procedures based on Support Vector Machines (SVM), Classification and Regression Trees (CART) and Random Forest for identifying null-allele SNPs. We report a list of null-allele SNPs we identified on the Illumina 660W-Quad chip, and provide suggestions for applying our CART model to other SNP sets. Properly identified null-allele SNPs can be used to test for genotype-phenotype associations or to identify regions which may contain copy number variants.

2.2 INTRODUCTION

Genome-wide association studies (GWAS) are used to investigate genetic effects on health and disease. GWAS search the genome for genetic variation that is correlated with phenotype. During data cleaning, prior to statistical analysis, many genetic markers are discarded because they fail to meet standard quality control criteria. As many as 35% of single nucleotide polymorphisms (SNPs) can be discarded if different genotype calling algorithms, such as BRLMM, Chiamo++, JAPL, are applied to the same dataset and only the SNPs passing all the quality control checks are used [Vens, et al. 2009]. Usually many SNPs are discarded because they are out of Hardy-Weinberg Equilibrium (HWE). A non-trivial number of these SNPs may have "null" alleles. If the null-allele SNPs can be identified, it may be possible to recover the data and analyze these SNPs using alternative models.

Many null-allele SNPs can be readily identified by visual inspection of raw genotyping intensity plots. Figure 1 shows example plots, where Y and X are intensities of allele A and allele B respectively. Each circle in the plot represents one individual. For the genotype AA (BB), high Y (X) and low X (Y) values are expected. If the X and Y values are similar, then the individual's genotype is assumed to be AB. Each color represents a cluster. If both of the X and Y intensities fall outside the genotype clusters, then the genotype is not assigned (uncalled), which is shown as an 'x'. Figure 1A shows three distinct genotype clusters as expected for a well-behaved "standard" di-allelic marker. By contrast, figure 1B shows a typical null-allele SNP. The extra cluster near the origin consists of individuals who are homozygotes for the null allele ("NN" genotype). For a null-allele SNP we assume that the blue and red clusters in Figure 1B are each composed of two underlying genotype clusters (AA and AN, or BB and BN), with the heterozygous cluster (AN or BN) closer to the origin. These underlying genotype clusters are

rarely visually detectable. Figure 1C is an example of a plot that is visually classifiable as *neither* "standard" nor "null allele."

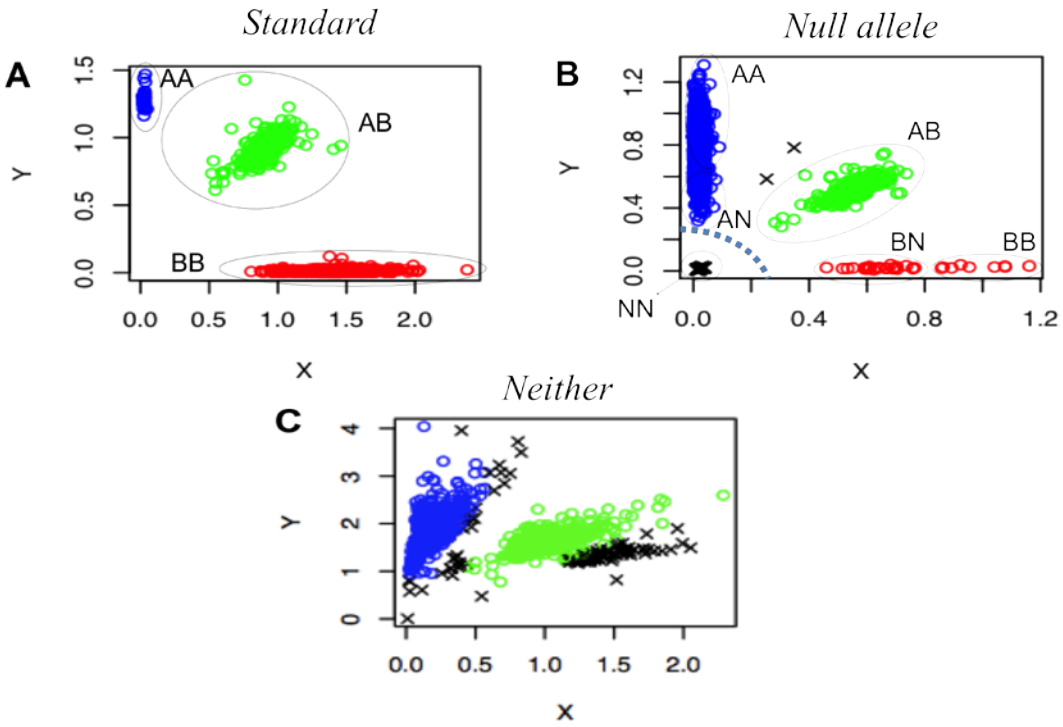

Figure 1. Example of SNP categories

Y and X are intensities of allele A and allele B respectively. Points in blue have the AA genotype, points in red correspond to the BB genotype, and points in green correspond to the heterozygous AB genotype. The black x's correspond to the uncalled genotypes. We drew a quarter circle (dashed line in Fig. 1B) centered at the origin with radius equal to the minimum distance to the nearest called genotype, and discarded the uncalled points lying out of the circle, while uncalled points within the circle are considered to correspond to the NN genotype.

Because it is not feasible to visually inspect tens of thousands of plots in order to detect null-allele SNPs, we developed methods for identifying null alleles based on the genotype data that are generated by standard genotype-calling algorithms (which assume a standard di-allelic model). To do this, we developed a likelihood model for the distribution of called genotypes that are obtained when a null-allele SNP is called by standard genotype-calling algorithms. We call

this our “null-allele model.” For each SNP, we considered three statistical tests: the goodness of fit test for this null-allele model and for the standard model (i.e., the standard HWE model), and the likelihood ratio test that compares those two models. Using these three test statistics and a hand-curated training dataset, we trained the classifiers Support Vector Machines (SVM), Classification and Regression Trees (CART), and Random Forest to identify null-allele markers.

There are also some other approaches to null-allele SNPs in the literature. Franke et al. [Franke, et al. 2008] proposed a method “TriTyper” that can detect genotypes in case-control datasets for deletion copy number variations (CNVs) or SNPs with an extra, uncalled (null) allele. They accounted for linkage disequilibrium (LD) as well as intensity data when calling the genotypes. They used data genotyped with Illumina Infinium BeadChips. They found 1,880 ‘triallelic’ SNPs - in other words, null-allele SNPs. There is another study by Kumasaka et al [Kumasaka, et al. 2011], which mainly focuses on genotyping copy number polymorphisms. They propose a Bayesian mixture model that uses intensity data, and provide software, PlatinumCNV, to detect SNP types including null-allele SNPs.

Figure 2 summarizes the main steps used to test our classifiers. We used a genome-wide association dataset from the Gene and Environment Association Studies (GENEVA) preterm birth study [Cornelis, et al. 2010]. From the set of SNPs that are out of HWE, we visually categorized 598 SNPs based on their intensity plots. We trained the SVM, CART and Random Forest models with this set of visually categorized SNPs. We then tested the trained classifiers as follows. We filtered the SNPs in our dataset based on commonly-used quality control filters in GWAS analyses: Hardy Weinberg equilibrium p-value and minor allele frequency (MAF). We included all autosomal SNPs meeting the MAF criteria but failing HWE (after first excluding the 598 training SNPs). We classified these testing SNPs by each method. In addition, we compared

our results with results from the TriTyper study [Franke, et al. 2008]. Finally, we simulated standard and null-allele SNPs and classified them by each method as an additional comparison of the methods.

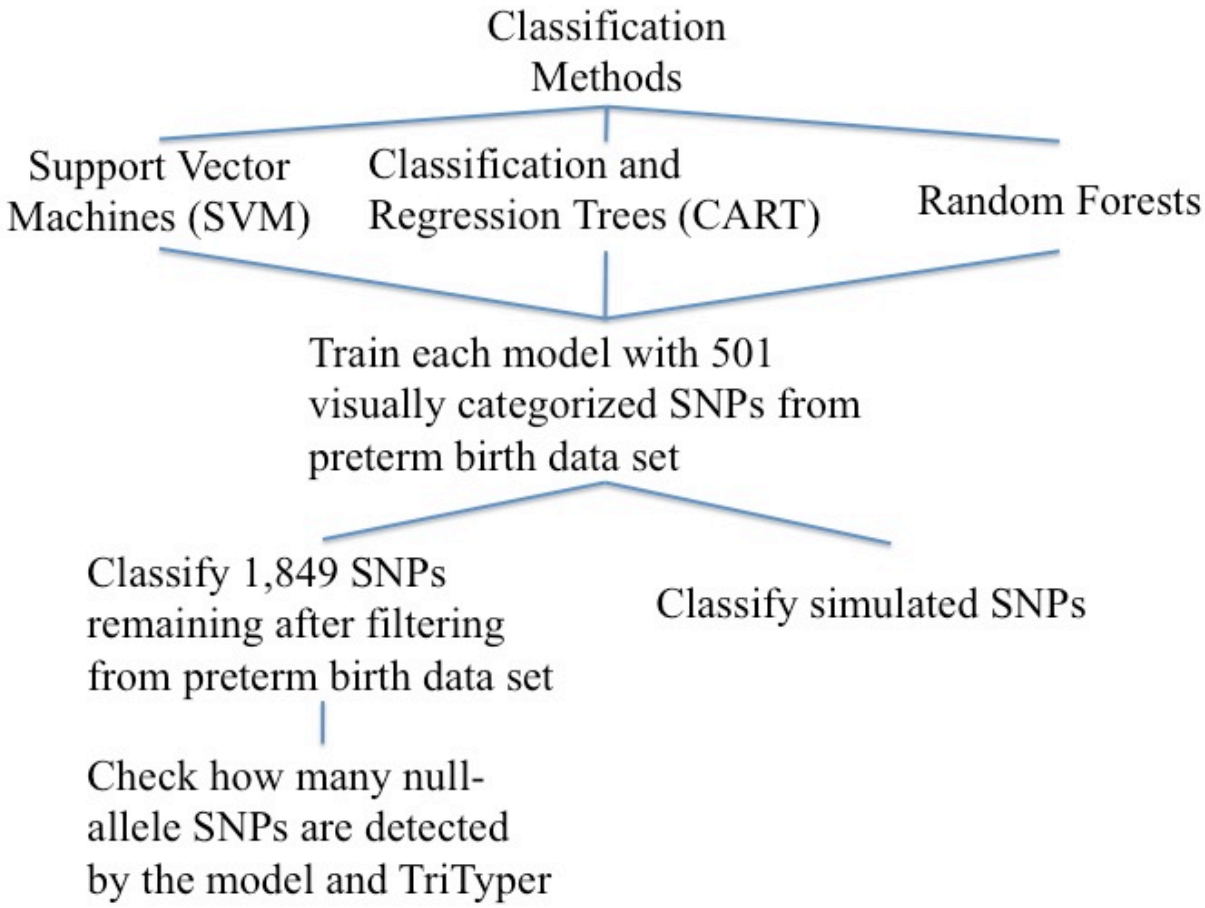

Figure 2. Classification of SNPs flowchart

2.3 LIKELIHOOD MODELS FOR GENOTYPE DATA

We fitted two statistical models to the genotype data: the standard model and the null-allele model. The standard model assumes two alleles and three genotypes, and the genotype plot of

the corresponding SNP should look like Figure 1A. Our null-allele model assumes that there is an extra allele, N, and that there are six possible genotypes, as in Figure 1B. However, since all SNPs have the genotypes called using the standard model, our data for both models is the counts of called AA, AB, and BB genotypes, plus an implicit count of NN genotypes that we infer from the number of uncalled genotypes near the origin (see below for a precise definition). For example, under our null-allele model the called AA cluster is assumed to include both AA and AN genotypes (details below). The two models are described in detail below, and the Appendix contains a detailed derivation of the relationship among the test statistics and the allele frequencies under the null-allele and standard model assumptions.

2.3.1 Standard model

The standard model has two alleles: A and B, and three genotype clusters: AA, AB and BB (Figure 1A). Each allele, A and B, has frequency p_S and q_S respectively, with $p_S + q_S = 1$. We assume Hardy-Weinberg equilibrium (HWE) holds, so that the frequencies of the three genotype clusters are p_S^2, $2p_Sq_S$, and q_S^2.

Our standard model fits a log-likelihood model for this two-allele system and uses one-dimensional optimization to find the maximum likelihood estimate of the allele frequency p_S [Brent 1972]. The log-likelihood model of standard model allele frequencies given the genotype counts is derived as follows:

$$L(p_S|x) \sim P(AA)^{x_{AA}} P(AB)^{x_{AB}} P(BB)^{x_{BB}} \quad (1)$$

$$\ln L(p_S|x) \sim x_{AA} \ln(p_S^2) + x_{AB} \ln(2p_Sq_S) + x_{BB} \ln(q_S^2) \quad (2)$$

Let $n_s = x_{AA} + x_{AB} + x_{BB}$ be the total number of people. The expected frequencies of the three clusters assuming HWE are $n_{AA} = p_S^2 n_s$, $n_{AB} = 2p_S q_S n_s$, and $n_{BB} = q_S^2 n_s$. The 1 df standard model goodness of fit test statistic (χ^2_{std}) is:

$$\chi^2_{std} = \frac{(x_{AA} - n_{AA})^2}{n_{AA}} + \frac{(x_{AB} - n_{AB})^2}{n_{AB}} + \frac{(x_{BB} - n_{BB})^2}{n_{BB}} \quad (3)$$

2.3.2 Null-allele model

The null-allele model has three alleles: A, B, and N, and six genotype clusters: AA, AN, BB, BN, AB, and NN (Figure 1B). Because the AN and BN clusters are subsumed within the AA and BB clusters respectively, we refer to them as [AA, AN] and [BB, BN]. Each allele, A, B, and N, has frequency *p*, *q* and *r* respectively, with $p + q + r = 1$.

For our null-allele model, we fitted a log-likelihood model and optimized it using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm developed by Broyden [Broyden 1970], Fletcher [Fletcher 1970], Goldfarb [Goldfarb 1970] and Shanno [Shanno 1970], a quasi-Newton method, which uses function values and gradients to obtain the surface to optimize. The method is more robust than the Conjugate Gradients method, and not as sensitive to initial values.

Let $x_{[AA,AN]}, x_{BB}, x_{[BB,BN]}$ and x_{NN} be the number of people with called AA [AA or AN], AB, called BB [BB or BN] and NN genotypes respectively. For NN counts, we drew a quarter circle centered at the origin in each plot (Figure 1B). The radius of the circle is the minimum distance from the origin to the nearest called genotype. To obtain the count of NN

genotypes, we counted only genotypes that lie inside the quarter circle. The likelihood and the log-likelihood of allele frequencies given the genotype counts are as follows:

$$L(p|x) \sim P(AA,AN)^{x_{[AA,AN]}} P(AB)^{x_{AB}} P(BB,BN)^{x_{[BB,BN]}} P(NN)^{x_{NN}} \quad (4)$$

$$\ln L(p|x) \sim x_{[AA,AN]} \ln(p^2 + 2pr) + x_{AB} \ln(2pq) + x_{[BB,BN]} \ln(q^2 + 2qr) + x_{NN} \ln(r^2) \quad (5)$$

Then we estimated the allele frequencies *p*, *q* and *r*. The total number of genotypes is $n = x_{[AA,AN]} + x_{AB} + x_{[BB,BN]} + x_{NN}$. The expected counts in each cluster assuming HWE are: $n_{[AA,AN]} = (p^2 + 2pr)n$, $n_{AB} = 2pqn$, $n_{[BB,BN]} = (q^2 + 2qr)n$, and $n_{NN} = r^2 n$. The 1 df null-allele model goodness of fit test statistic (χ^2_{null}) is:

$$\chi^2_{null} = \frac{(x_{[AA,AN]} - n_{[AA,AN]})^2}{n_{[AA,AN]}} + \frac{(x_{AB} - n_{AB})^2}{n_{AB}} + \frac{(x_{[BB,BN]} - n_{[BB,BN]})^2}{n_{[BB,BN]}} + \frac{(x_{NN} - n_{NN})^2}{n_{NN}} \quad (6)$$

2.3.3 Likelihood ratio

In addition to the two goodness-of-fit test statistics described above, we also considered the likelihood ratio statistic comparing the standard and null-allele models. We computed the likelihood ratio (LR) of the two models as

$$LR = -2\big(\ln L\big(p_S|x\big) - \ln L(p|x)\big) \quad (7)$$

2.4 CLASSIFICATION OF SNPS

2.4.1 Classifiers

We applied three widely used classifiers: Support Vector Machines (SVM), Classification and Regression Trees (CART), and Random Forest. We used the null-allele and standard model goodness of fit statistics (χ^2_{null} and χ^2_{std}) and the likelihood ratio statistic as the predictors for all three methods. For SVM, we used class weights inversely proportional to the class sizes in the training set, and we tuned the model to determine cost and gamma parameters. For CART, we used 'information' parameters for the splitting function and pruned the tree to predict new SNPs. All three models input and output the three classes neither, null-allele and standard.

We used the R e1071 package for SVM by Dimitriadou et al. [Dimitriadou, et al. 2011], the rpart package for CART by Therneou et al. [Therneau and Atkinson 2012] and the randomForest package by Liaw and Wiener [Liaw and Wiener 2002] for Random Forest analysis.

2.4.2 Datasets

Our primary training and test datasets were drawn from the preterm birth data from Gene Environment Association Studies (GENEVA) [Cornelis, et al. 2010]. In this dataset, there are 657K SNPs and 3,947 individuals. Genotyping of the SNPs was performed using the Illumina Platform (www.illumina.com). Illumina's Genome Analyzer system software program converts raw image data into intensity scores and assigns genotypes based on clustering. In standard GWAS analysis, SNPs are filtered out if their MAF and HWE p-values fail to meet certain criteria. Many null-allele SNPs are filtered out because their HWE p-values are less than the

specified threshold. It is from among these discarded markers that we want to be able to detect null-allele SNPs. Thus for our training and test datasets we used only SNPs for which MAF $\geq$ 0.05 and HWE p-value < 0.001.

We also applied the classifiers to SNPs that were previously identified as 'triallelic' by the TriTyper study [Franke, et al. 2008]. Of the 1,880 TriTyper triallelic SNPs they identified, 1,757 of them were typed in our preterm birth dataset; we classified those SNPs using all methods in this paper.

Finally, we evaluated the classifiers by applying them to five simulated datasets. We simulated four null-allele model sets with the allele frequencies p, q and r, and one dataset using the standard model with the allele frequencies p, q and $r = 0$. One of the null-allele model datasets and the standard model dataset had 30K SNPs and 6K individuals. We generated the allele frequency, p, from the uniform distribution with the range [0.05, 0.4] for the standard model SNPs, where $p + q = 1$. For the first null-allele model set, p is uniformly distributed in the range [0.05, 0.70] and r varies in the range [0.01, 0.20] based on the preterm birth dataset null allele frequencies we observed, and $q = 1 - p - r$. Each of the remaining three null-allele model datasets had 10K SNPs and 4K individuals. These were generated with fixed parameters. For all three datasets, p was fixed at 0.6. In each dataset, r was fixed at 0.01, 0.02 and 0.03, and so q was 0.39, 0.38 and 0.37, respectively. All SNPs meet the MAF filtering criteria based on allele frequencies estimated assuming the standard model. We filtered out the simulated SNPs if the HWE p-value $\geq$ 0.001. We performed the analysis twice with and without filtering the simulated datasets.

PLINK was used to obtain the MAFs and the HWE p-values [Purcell, et al. 2007]. R was used for all other statistical analyses [RCoreTeam 2012].

2.4.3 Methods

We created initial training and test sets by selecting 598 SNPs among 2,350 remaining SNPs after MAF and HWE filtering on chromosomes 1-22, and visually categorized them as 'null allele', 'standard', 'neither', or 'unknown' based on their intensity plots. We visually categorized a SNP as a 'null-allele' SNP if there are four distinct clusters with two of the clusters (the [BB, BN] and [AA, AN] clusters) lying along Y and X axes, one cluster of unknown genotypes (the NN cluster) bunched right around the origin, and the last cluster (the AB cluster) around the diagonal $X = Y$ line distinctly from other clusters. Figure 1B is a canonical example of a null-allele SNP. We categorized a SNP as 'standard' if there are three distinct clusters, AA, BB and AB, but no null genotypes around the origin (e.g., Figure 1A). A SNP was categorized as 'neither' if there are not at least three distinct clusters or there are deviations in cluster shapes, such as sub-clusters lying one over another. Figure 1C illustrates a 'neither' SNP, which we classified as 'neither' because we do not see three or more distinct clusters. We categorized a SNP as 'unknown' if there is a cluster around origin composed of only called genotypes or both called and uncalled genotypes. The distinction between neither and unknown SNPs is that unknown SNPs have distinct clusters as expected in standard or null-allele SNPs, but neither SNPs do not. Of the 598 visually categorized SNPs with MAF ≥ 0.05 and HWE p-value < 0.001, there were 274 'null allele', 88 'standard', 139 'neither', and 97 'unknown' SNPs.

To evaluate our classifiers, we created 50 replicates of test and training datasets. For each replicate, we randomly sampled 200 SNPs from the 598 visually categorized SNPs as the test dataset. From the remaining 398 visually categorized SNPs, we discarded unknowns, so that null, standard and neither SNPs constituted the training dataset. Because the class 'unknown' could be either of the 'neither', 'null' and 'standard' classes, we eliminated the 'unknown' SNPs

from the training sets. We did not remove unknown SNPs from the test set because when classifying SNPs in practice, we will not be able to remove them. There were three classes, neither, null, and standard, as input in the training set, and therefore three classes as output in the prediction.

For each of the 50 replicates, we trained SVM, CART and Random Forest models with the training set using the LR, χ^2_{null}, and χ^2_{std} statistics as predictors. Then we classified 200 test SNPs using the trained models. We calculated precision, true positive rate (TPR) and false positive rate (FPR) as follows:

$$Precision = \frac{Number\ of\ visually\ categorized\ null\ SNPs\ classified\ as\ "null"}{Number\ of\ visually\ categorized\ neither,null\ and\ standard\ SNPs\ classified\ as\ "null"} \quad (8)$$

$$True\ Positive\ Rate = \frac{Number\ of\ visually\ categorized\ null\ SNPs\ classified\ as\ "null"}{Number\ of\ visually\ categorized\ null\ SNPs} \quad (9)$$

$$False\ Positive\ Rate = \frac{Number\ of\ visually\ defined\ \text{"neither"}\ or\ "standard"\ SNPs\ classified\ as\ "null"}{Number\ of\ visually\ defined\ \text{"neither"}\ or\ "standard"SNPs} \quad (10)$$

Because ‘unknown’ SNPs in the test set have an ambiguous class, we did not consider them while calculating the rates.

In each of the 50 replicates, each classifier was allowed to choose which predictors to use. We also considered all different combinations of LR, χ^2_{null}, and χ^2_{std} statistics to put into analyses manually as predictors.

As an additional method for comparing performance on the test set, we investigated the pairwise agreement between all pairs of methods in 50 replicates. We also assessed the agreement between the methods using a robust statistic, the Kappa Coefficient [Cohen 1960].

After comparing the performance of the methods on the real-data test set, we retrained the classifiers with 501 ‘neither’, ‘standard’ and ‘null’ visually categorized SNPs, and, excluding

those training SNPs, we classified the remaining genome-wide autosomal SNPs left after HWE and MAF filtering in the preterm birth dataset. Then, by using the same retrained classifiers, we classified the TriTyper triallelic SNPs and simulated sets of SNPs and evaluated the performance of all classifiers on them.

2.4.4 Results

Table 1 displays the precisions of the SVM, CART and Random Forest approaches. The means and the standard deviations are based on 50 replicates. Even though the differences are very small, the precision of CART (mean precision = 0.918) was slightly higher compared to other methods (Table 1).

Table 1. Precision for the three classifiers

	Precision	
Method	Mean	SD
SVM	0.910	0.021
CART	0.918	0.022
Random Forest	0.914	0.025

Means and standard deviations (SD) are based on 50 replicates.

Table 2 shows the true and false positive rates. Among all methods, Random Forest gave the highest mean TPR (0.947) with the lowest standard deviation, but CART had the lowest mean FPR (0.103) (Table 2).

Table 2. Mean and standard deviation (SD) of true and false positive rates based on 50 replicates

Method	Mean True Positive Rate (SD)	Mean False Positive Rate (SD)
SVM	0.932 (0.036)	0.112 (0.028)
CART	0.944 (0.031)	0.103 (0.029)
Random Forest	0.947 (0.021)	0.108 (0.032)

We also evaluated how often the methods agreed with each other. Table 3 shows the mean frequencies of visually categorized null-allele SNPs, which were classified correctly by the two corresponding methods. Table 4 summarizes the agreement between the methods using the Kappa Coefficient. Both Tables 3 and 4 represent the results of 50 replicates as in Tables 1 and 2. The highest agreement on null-allele SNPs (93%) (Table 3) and the highest overall agreement in Table 4 was between CART and Random Forest (Kappa = 0.912).

Table 3. Agreement on null-allele model SNPs of all methods

Method	SVM	CART
CART	91%	
Random Forest	92%	93%

Each cell represents the average of the percentages of visually categorized null-allele model SNPs, which were correctly classified by the two corresponding methods, in 50 replicates.

Table 4. Agreement between the methods by Kappa Coefficient

Method	SVM	CART
CART	0.871 (0.044)	
Random Forest	0.881 (0.037)	0.912 (0.038)

Each cell is the average coefficient of 50 repeated samples and standard deviations in parenthesis.

In addition to the agreement analyses, we plotted a color map showing the agreement of methods in one of the test data sets (Figure 3). We also present a graphical version of true and false positive rates summarized in Table 2, by method (Figure 4).

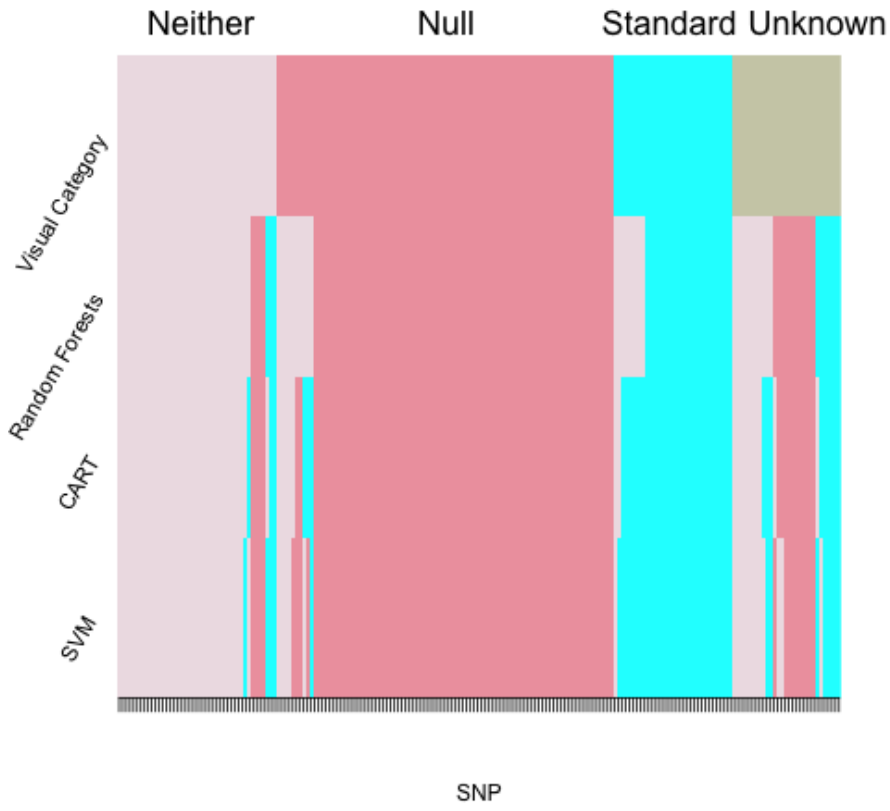

Figure 3. Color map of the agreement among methods by 200 SNPs

Neither, null, standard and unknown SNPs are represented by lavender, pink, cyan and light brown respectively.

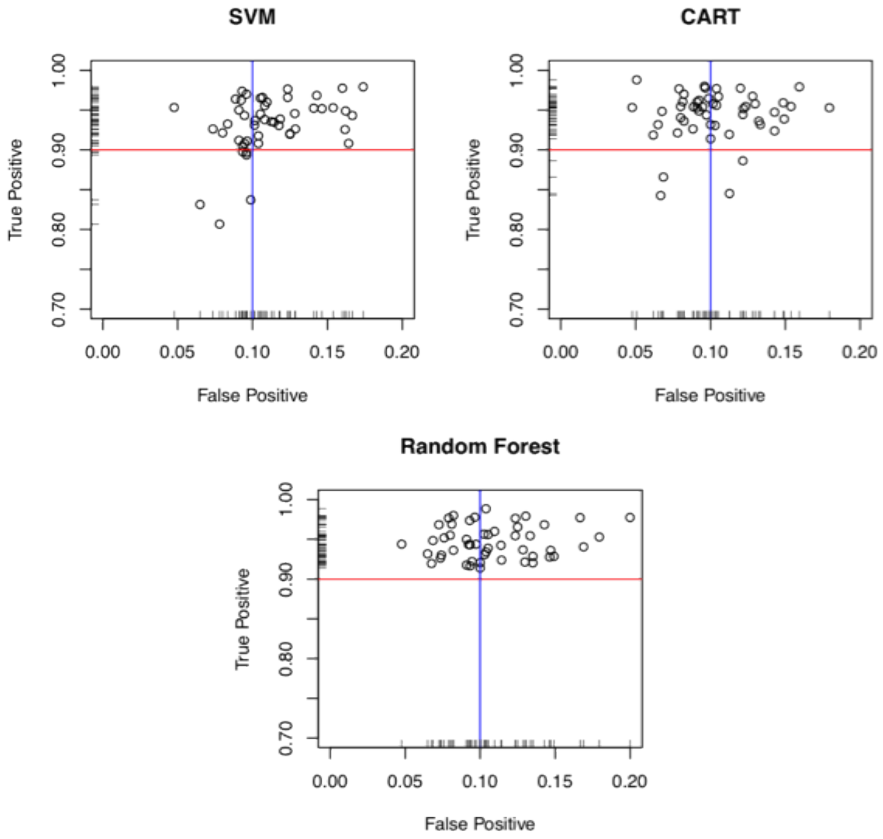

Figure 4. True positive versus false positive rate plots of all methods based on 50 replicates

All classifiers used in this study are able to select the predictors to be used in the classification process for the most effective classification. Each classifier ended up using all three statistics as predictors. In other words, all three test statistics, χ^2_{std} χ^2_{null} and LR, are needed to efficiently classify the SNPs. Figure 5 displays the CART model obtained by training with all of the 501 visually categorized null, standard and neither SNPs. The CART model classifies the SNP as 'null', 'standard' or 'neither'. A SNP is classified as 'standard' if LR $\geq$ -3.97 and χ^2_{std} <

19.67. If LR ≥ -3.97 and $\chi^2_{std} \geq 19.67$, or LR < -3.97 and $\chi^2_{null} \geq 11.86$, then the SNP is classified as 'neither'. We classify the SNP as 'null allele' if LR < -4 and $\chi^2_{null} < 11.86$.

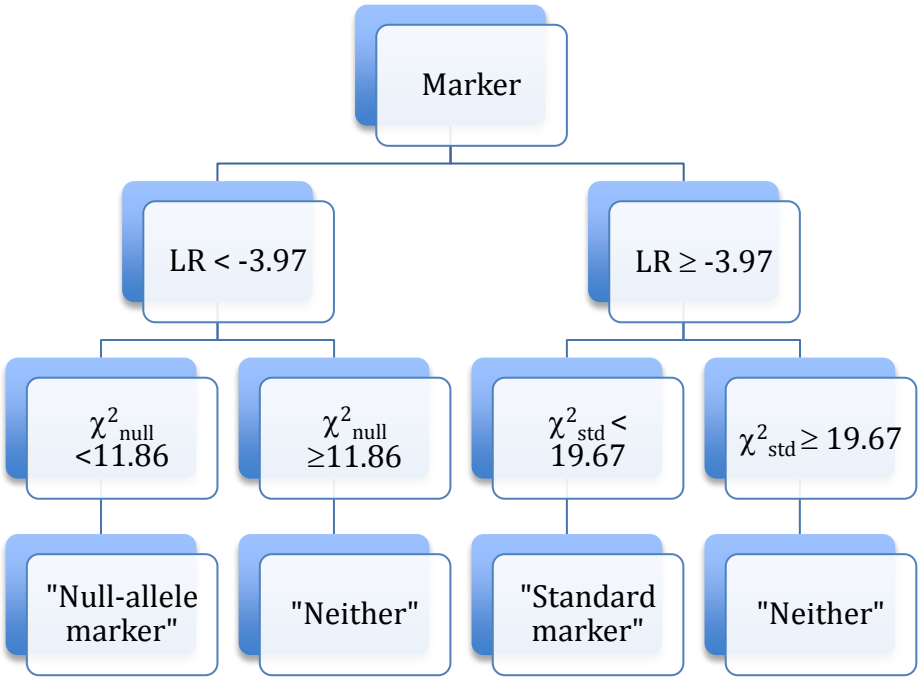

Figure 5. The CART model based on all 501 visually categorized null, standard and neither SNPs

χ^2_{std}: Standard model goodness of fit test statistic. χ^2_{null}: Null-allele model goodness of fit test statistic. LR: Likelihood ratio of null-allele and standard models.

Our next analysis was to classify all autosomal 'testing' SNPs genome-wide by using each classifier trained on 501 visually categorized null, standard and neither SNPs. There are 2,350 autosomal SNPs genome-wide meeting HWE and MAF criteria in the preterm birth dataset. We removed the 501 training SNPs and classified the remaining 1,849 testing SNPs. CART classified 998 (54%) of the 1,849 SNPs as 'null allele', Random Forest classified 938 (51%) and SVM classified 925 (50%) of the 1,849 SNPs as 'null allele'.

Our third test of the three methods was to compare their performance on the 1,880 TriTyper triallelic SNPs that Franke et al. [Franke, et al. 2008] reported. Among the 1,880 TriTyper triallelic SNPs, 1,757 SNPs are autosomal and in our preterm birth dataset. We removed SNPs from our training set and classified the remaining 1,492 TriTyper triallelic SNPs. Table 5 shows the classification results for the TriTyper triallelic SNPs, subdivided by our HWE and MAF criteria. As our classifiers were trained on all 501 visually categorized null, standard and neither SNPs satisfying our HWE and MAF criteria, they are most directly applicable to the 833 SNPs meeting those criteria (line 1 of Table 5). On these, our classifiers identified about 77% of them as null-allele SNPs.

Table 5. Classification of TriTyper triallelic SNPs

Criteria	Number of SNPs	SVM n(%)			CART n(%)			Random Forest n(%)		
		Null allele	Standard	Neither	Null allele	Standard	Neither	Null allele	Standard	Neither
HWE<0.001 & MAF≥0.05	833	640 (76.8)	3 (0.4)	190 (22.8)	637 (76.5)	2 (0.2)	194 (23.3)	646 (77.6)	4 (0.5)	183 (22.0)
HWE<0.001 & MAF<0.05	56	42 (75.0)	0 (0)	14 (25.0)	49 (87.5)	0 (0)	7 (12.5)	49 (87.5)	0 (0)	7 (12.5)
HWE≥0.001 & MAF≥0.05	573	187 (32.6)	278 (48.5)	108 (18.8)	332 (57.9)	201 (35.1)	40 (7.0)	188 (32.8)	293 (51.1)	92 (16.1)
HWE≥0.001 & MAF<0.05	30	9 (30.0)	12 (40.0)	9 (30.0)	18 (60.0)	10 (33.3)	2 (6.7)	16 (53.3)	11 (36.7)	3 (0.1%)
Overall	1,492	878 (58.8)	293 (19.6)	321 (21.5)	1036 (69.4)	213 (14.3)	243 (16.3)	899 (60.2)	308 (20.6)	285 (19.1)

Finally, we compared the performance of our classifiers on simulated data. By using simulated datasets, we are able to evaluate the performance of the methods with the simulated perfect standard and null-allele model SNPs. The results are summarized in Table 6. Almost all of the filtered or non-filtered standard SNPs, where the null allele frequency $r = 0$, were correctly classified by all methods. CART was able to correctly classify 98% of the simulated null-allele SNPs. 93% of the non-filtered null-allele SNPs were classified as null allele by SVM, and 95% by Random Forest methods (Table 6). All methods did much better when the null allele frequency $r \geq 0.03$.

Table 6. Classification of simulated dataset results

						SVM		CART		Random Forest	
Frequencies						*Classified as*		*Classified as*		*Classified as*	
p	q	r	True class	HWE p-value filtering		Null (%)	Standard (%)	Null (%)	Standard (%)	Null (%)	Standard (%)
0.6	0.39	0.01	Null	Filtered	196	0.2	85.2	35.7	62.2	1.0	77.0
				Not filtered	10K	0.2	96.6	33.2	66.7	1.2	85.0
0.6	0.38	0.02	Null	Filtered	1,964	21.8	41.5	70.2	23.1	20.0	35.2
				Not filtered	10K	13.3	57.2	77.5	21.1	26.8	35.4
0.6	0.37	0.03	Null	Filtered	6,447	70.6	5.8	92.2	3.5	65.3	4.8
				Not filtered	10K	66.3	9.2	93.8	3.4	68.5	4.4
[0.05,0.7]	1-p-r	[0.01,0.2]	Null	Filtered	27,984	92.5	3.4	98.0	1.6	94.6	2.3
				Not filtered	30K	92.1	3.5	98.0	1.6	94.1	2.5
[0.05,0.4]	1-p	0	Standard	Filtered	32	0	100	0	100	0	96.9
				Not filtered	30K	0	100	0	99.9	0	99.9

HWE p-value filtering shows if the SNPs having HWE p-value≥0.001 are filtered out or not. All MAF≥0.05.

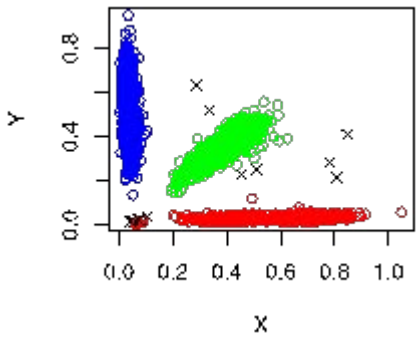

Figure 6. Genotype plot of the SNP classified as triallelic by TriTyper but 'neither' by our methods

2.5 DISCUSSION

In this study, our aim was to efficiently identify null-allele markers using only genotype data that are provided by standard calling algorithms. We applied three different classifiers, Support Vector Machines, Classification and Regression Trees and Random Forest, to the GENEVA preterm birth dataset. All of the classification methods used three predictors: the null-allele and the standard multinomial likelihoods for the genotype data, and the likelihood test comparing those two models. We also applied our methods to the TriTyper triallelic SNPs, and we compared the results with TriTyper study. Finally we classified simulated standard and null-allele model sets.

While the performances of the classifiers SVM, CART and Random Forest were similar in terms of precision and true and false positive rates (Tables 1 and 2), in certain situations CART was able to detect more simulated null-allele SNPs correctly (Table 6). In general, we were able to identify most of the standard and null-allele SNPs correctly by using the classifiers. We provide our final CART model for researchers to use for classification of markers (Figure 5).

According to the study, dataset etc., the thresholds of the CART model in Figure 5 can be adapted.

Our methods also performed similar to each other on the TriTyper triallelic SNPs. 77% of the triallelic SNPs identified by TriTyper were classified as null-allele SNPs by all of our methods. However, CART was able to classify more TriTyper triallelic SNPs as null-allele than other two methods (line 5 of Table 5). Because TriTyper is a method of genotype calling by using linkage disequilibrium as well as raw intensities, and our methods rely on genotyped data based on biallelic assumption and no-called genotypes are needed around origin for null-allele SNPs, our study and TriTyper study agree only on 77% of the TriTyper triallelic SNPs. Figure 6 shows a genotype plot of the SNP which is classified as triallelic by TriTyper but as 'neither' by our methods because of the called genotypes in the cluster around origin which is supposed to be NN cluster.

From the simulation experiments, we observed that when the null allele frequency, *r*, is less than 0.02, all classifiers may misclassify more than half of the actual null-allele SNPs as standard SNPs (Table 6). When $r \leq 0.02$, SVM and Random Forest could not classify most of the simulated null-allele model SNPs correctly, but CART is able to detect more null-allele SNPs correctly. In order to compare the effect of filtering, we included the analysis results of simulated datasets with and without filtering. In general, SVM did better with filtered data. CART did better with filtered data when $r = 0.01$.

The methods used in this paper are appropriate for Illumina 660W-Quad chip output. They may be compatible with other Illumina platforms. We also examined one Affymetrix dataset in order to see how our methods might work for a different platform. Based on one dataset, clustering of intensity data instead of using genotype data may be needed to apply to

data generated by Affymetrix technology, because our method needs no-called genotypes at the origin.

Our methods provide a practical approach to identifying null-allele SNPs in a GWAS study. This can be done in one of three ways. (1) The list of null-allele SNP names that we discovered can be used directly. There are also other studies such as TriTyper, where list of null-allele SNPs can be found. (2) One could apply our CART model, which gives the best results according to the simulated SNP experiments and is the most convenient. Or (3) all of our steps can be followed to create a model and classify the SNPs.

Once null-allele SNPs are identified, the information can be used in a number of ways. One is to call genotypes at these SNPs using other software (or by hand) and analyze genotype-phenotype association rather than discarding the SNPs. Another is to look more closely at these regions for potentially functional copy number variants. Either of these approaches offers the possibility of deriving useful new information from SNPs that are now discarded in GWAS studies.

3.0 STATISTICS FOR X-CHROMOSOME ASSOCIATIONS

3.1 ABSTRACT

Association between genotype and phenotype at autosomal loci is generally tested by chi-squared tests (most often Armitage's trend test), or by regression models if there are covariates. X chromosome data are often excluded from published analyses. Failure to analyze X data at all is obviously less than ideal, and can lead to missed discoveries. Even when the X chromosome is analyzed, it is usually done with suboptimal statistics. Several mathematically sensible statistics for X-chromosome association have recently been proposed. The optimality of these statistics, however, is based on very specific simple genetic models. In addition, while the simulation studies have been very informative, they have focused on single-marker tests and have not considered the types of error that occur when an entire chromosome is scanned. In this study, we comprehensively test the most promising X-chromosome association statistics using simulation studies that include the entire chromosome. We also consider a wider range of models for sex differences and phenotypic effects of X inactivation. We found that many of the best statistics perform well even when there are variance differences between the sexes or small sex differences in allele frequency. Moreover, we find that for additive models, males should be treated as homozygous females for X chromosome loci.

3.2 INTRODUCTION

In genome wide association studies (GWAS), the first step after data cleaning is testing single nucleotide polymorphisms (SNPs) for association with a trait by chi-squared tests or regression models. Analyzing markers on autosomal chromosomes is more straightforward than sex chromosomes. Testing for association on the X chromosome, which makes up 5% of the female genome, requires specialized analysis methods – methods developed for analyzing autosomal data are not directly applicable to X data because males have only one copy of the X chromosome. Very often X chromosome data is discarded from published analyses. From January 2010 through March 2012, only 33% of the GWA papers reported X chromosome results [Wise, et al. 2011]. Moreover, even when the X chromosome is analyzed, usually suboptimal statistics are used.

Several promising statistical methods for X chromosome association testing have recently been developed. However, there is a need for complete testing of the X chromosome methods. The best of the recently proposed statistics has been shown to be most powerful only when there is no sex difference in allele frequencies. Therefore, it is still not known how the statistic will behave when we scan an entire chromosome in which there is a small amount of random variation in allele frequencies between the sexes - whether it will still find the correct loci or will just pick out the ones that have the largest sex-specific allele frequency differences by chance. We aim to fill these gaps by fully testing several commonly-used and newly-proposed X chromosome statistics. We considered potential differences in phenotypes and genotypes between genders. We also studied the behavior of statistics under X inactivation. X inactivation is randomly silencing of one allele in heterozygote females, and may cause a higher trait variance

for heterozygote females. We performed simulation studies using real chromosome-wide data in order to fully understand the practical performance of the statistics.

3.3 X CHROMOSOME TEST STATISTICS

Genotype-phenotype association is generally tested by chi-square tests, most often Armitage's trend test, or regression models. The phenotype variable can be binary or continuous, and the genotype variable can be coded as in different ways. Linear coding for genotypes is typical. At autosomal loci, if the genetic model is additive, the genotype is coded as (0, 1, 2). In order to test the association at X chromosome loci, female genotypes are coded as (0, 1, 2) and, because males have only one copy of the X chromosome, male genotypes are coded as (0, 1) or (0, 2). A good X chromosome association test statistic might need to take care of features such as male genotype coding, male/female differences in phenotype/genotype variance, and Hardy-Weinberg equilibrium (HWE) assumption.

In this study, we evaluated several regression models as well as several X chromosome statistics that will be described below. We fitted six regression models to test the effect of male genotype coding schemes, we ran each model indicated in equations 11-13 twice. First we ran with datasets where male genotypes are coded as (0,1) and we recoded the datasets so that male genotypes are (0,2) and ran again. The first two regression models have only genotype as independent variable as in equation 11. We introduced sex as a covariate into the third and forth models (equation 12). Then we added the variable genotype and sex interaction in the fifth and sixth models (equation 13). We compared the last two models with the regression models where

there is only sex as an independent variable. The results regarding the models in equation 13 in this study are the results of that comparison.

$$Phenotype \sim Genotype \quad (11)$$

$$Phenotype \sim Genotype + Sex \quad (12)$$

$$Phenotype \sim Genotype + Sex + Genotype * Sex \quad (13)$$

Even though we can control for sex in regression analysis, the analysis does not handle male/female difference in phenotype/genotype variance.

A number of different statistics have been proposed for X-chromosome association [Clayton 2008; Zheng, et al. 2007]. The statistics proposed by Clayton [Clayton 2008] are generalized linear model score tests based on genotype-phenotype covariance. They do not assume HWE and take into account X inactivation. They treat males as homozygote females. In other words, male genotypes are coded as homozygote females. They do not lose power whereas stratification does, even if the phenotype varies between sexes and if allele frequency does not [Clayton 2008]. They do assume that equal male and female allele frequencies. To compute the Clayton statistics, let subjects 1,…, F be female and F+1,…, N be male. Y_i is the phenotype and A_i is the genotype for subject i. D_i is the heterozygosity indicator. It is 0 for homozygotes and 1 for heterozygotes. P, which is assumed to be same in males and females, is the allele frequency in the population estimated from the data. The 2 degree-of-freedom test statistic for X chromosome data is:

$$T_2 = U^T \hat{V}^{-1} U \sim \chi_2^2 \quad (14)$$

where

$$U = \begin{bmatrix} U_A \\ U_D \end{bmatrix} = \begin{pmatrix} \sum_{i=1}^{N} (Y_i - \bar{Y}) A_i \\ \sum_{i=1}^{F} (Y_i - \bar{Y}_F) D_i \end{pmatrix} \quad (15)$$

$$\hat{V} = \hat{V}_F \sum_{i=1}^{F}(Y_i - \bar{Y})^2 + \hat{V}_M \sum_{i=F+1}^{N}(Y_i - \bar{Y})^2 \quad (16)$$

The female and male components of the variance are:

$$\hat{V}_F = \frac{1}{F-1}\sum_{i=1}^{F}\begin{pmatrix} (A_i - \bar{A})^2 & (A_i - \bar{A})(D_i - \bar{D}_F) \\ (A_i - \bar{A})(D_i - \bar{D}_F) & (D_i - \bar{D}_F)^2 \end{pmatrix} \quad (17)$$

$$\hat{V}_M = \begin{pmatrix} 4P(1-P) & 0 \\ 0 & 0 \end{pmatrix} \quad (18)$$

The 1 degree-of-freedom test statistic is:

$$T_1 = U_1^2/\hat{V}_{11} \sim \chi_1^2 \quad (19)$$

For autosomal loci, the variance does not include the male component, $\hat{V}_M$, and $\hat{V}_F$ is calculated over all subjects.

Despite the regression analysis, Clayton's statistic takes into account different male and female variances. However, it assumes that the allele frequency does not differ in males and females.

Zheng et al. [Zheng, et al. 2007] proposed a test statistic, which is a combination of allele-based test statistic and genotype-based trend test.

$$Z_{mfG}^2 = \left(\sqrt{\frac{n_f}{n_m+n_f}}Z_{fG} + \sqrt{\frac{n_m}{n_m+n_f}}Z_m\right)^2 \sim \chi_1^2 \quad (20)$$

where

$$Z_{fG} = \frac{n_f^{\frac{1}{2}}\left[s_f\left(\frac{1}{2}r_{f1}+r_{f2}\right)-r_f\left(\frac{1}{2}s_{f1}+s_{f2}\right)\right]}{\left[r_f s_f\left[n_f\left(\frac{1}{4}n_{f1}+n_{f2}\right)-\left(\frac{1}{2}n_{f1}+n_{f2}\right)^2\right]\right]^{\frac{1}{2}}} \quad (21)$$

$$Z_m = \frac{n_m^{\frac{1}{2}}(r_m s_{m0} - s_m r_{m0})}{(n_{m0} n_{m1} r_m s_m)^{\frac{1}{2}}} \quad (22)$$

r_{mi} (r_{fj}) and s_{mi} (s_{fj}) are number of male (female) cases and controls having i allele (j genotype). n_m (n_f) is number of males (females) and n_{mi}= r_{mi}+ s_{mi} (n_{fj}= r_{fi}+ s_{fi}).

Zheng et al.'s statistic is based on sex-specific allele frequencies. However, it assumes HWE in females and it does not take into account X inactivation. Hickey and Bahlo [Hickey and Bahlo 2011] have shown that Clayton's test statistics are generally more powerful than Zheng et al.'s [Zheng, et al. 2007].

3.4 METHODS

We comprehensively evaluated six regression models and two specialized X chromosome association test statistics mentioned above with simulation studies under a variety of statistical models for sex differences that account for errors introduced by chromosome-wide testing. Conventional statistical theory measures the optimality of statistics in terms having correct type I error and maximal power for a single test. But in genomic applications, we apply a test thousands or hundreds of thousands of times and pick out the most significant loci for further study. In that situation, it is not the expected value of the behavior of the statistic that matters, but rather the behavior of the extreme values (order statistics). For example, the best of the proposed statistics, Clayton's test statistic, has been shown to be most powerful, but only when there is no sex difference in allele frequencies. How will that statistic behave when we scan an entire chromosome in which there is a small amount of random variation in allele frequencies between

the sexes? Will it still find the correct loci, or will it just pick out the ones that have the largest sex difference by random chance? We performed our simulation studies using real chromosome-wide data in order to answer this type of question and fully understand the practical performance of the statistics.

We compared the statistics by type I error rates and statistical power. R was used for statistical analyses [RCoreTeam 2012]. There are many other tests in the literature, including Bayesian approaches, tests that attempt to gain power by assuming HWE, etc., but in practice the majority of GWAS studies use the straightforward additive test. Kuo and Feingold [Kuo and Feingold 2010] showed that this strategy has robust power both in terms of single-SNP testing and genome scanning.

3.4.1 Datasets

In this study, the datasets consist of simulated binary and continuous phenotypes, simulated SNPs and real X-chromosome SNPs. We used the X-chromosome SNPs of the preterm birth data from Gene Environment Association Studies (GENEVA) [Cornelis, et al. 2010]. In this GWAS dataset, there are 3,947 individuals genotyped using the Illumina Platform (www.illumina.com). We dropped mothers' data and used only 1,795 babies in our study. There are 863 female babies and 932 male babies in the dataset. Among all SNPs on the X chromosome in the preterm birth dataset, we filtered out the SNPs if minor allele frequency (MAF) < 0.02 or HWE p-value < 0.0001. Then 12,242 SNPs were left for the analyses. PLINK was used to obtain the MAFs and the HWE p-values [Purcell, et al. 2007].

We simulated binary phenotypes to answer the question of how type I error rate and/or power of the tests compare under various genetic models and sampling scenarios including both

unbalanced (i.e. different female/male * case/control ratios in dataset) and balanced datasets. There are 393 female cases, 470 female controls, 451 male cases, and 481 male controls in all datasets in this study. Unbalanced datasets were created from balanced datasets by randomly dropping a subset of individuals.

For binary phenotype power analyses, we simulated 200 replicates of a dataset of 393 female cases, 470 female controls, 451 male cases, and 481 male controls at a single SNP having female and male allele frequencies 0.5 and 0.46 for controls and cases under alternative hypothesis. Also, we simulated 200 replicates of the dataset at a single SNP having female allele frequencies 0.53 and 0.49 for controls and cases, and male allele frequencies 0.5 and 0.46 for controls and cases.

For both binary phenotype type I error and power analyses, in addition to our real X-chromosome loci, we spiked in 120 SNPs with extra-large male-female allele frequency differences that range in (0.07, 0.15). This allowed us to test the behavior of the statistics both for normal variation between male and female allele frequencies and for extreme situations. Figure 7 shows the density plots of real and simulated SNPs' allele frequency differences used for type I error rate analyses.

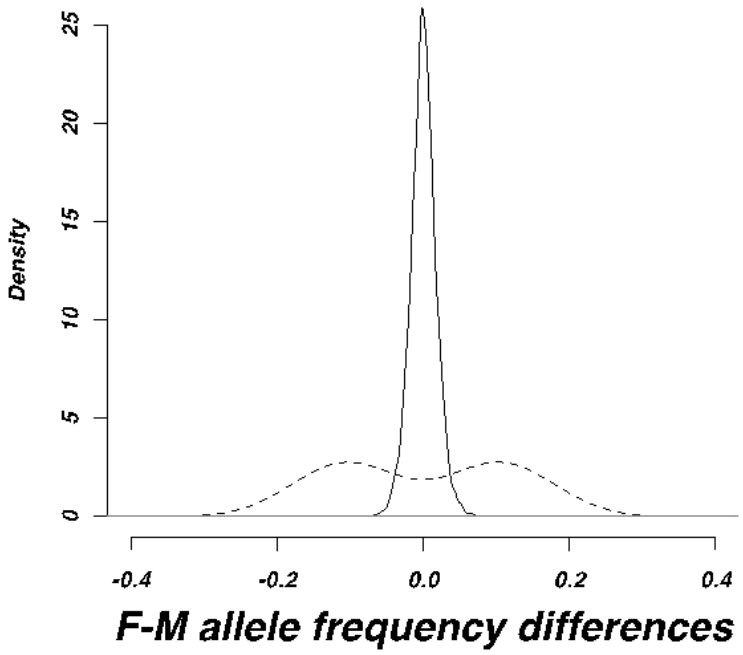

Figure 7. Allele frequency differences between males and females

Solid line refers to 12,242 real GWAS X chromosome SNPs. Dashed line refers to 120 simulated SNPs with large allele frequency differences between males and females.

For continuous phenotypes, the question of interest was whether different distributions for male and female phenotype and/or genotypes would affect the expected type I error or power of the tests. To investigate type I error rate of the tests for continuous phenotypes, we used the preterm birth dataset SNPs and the simulated 120 SNPs having extremely large allele frequency differences between males and females. We simulated two phenotype variables: first we simulated a phenotype from Normal distribution with mean 15 and variance 9, N(15, 9), for both

genders. The second phenotype variable consists of female phenotypes from N(18, 9) and male phenotypes from N(15, 9).

Table 7 shows the genetic models and sampling designs we considered for continuous phenotype power analysis. We simulated 200 replicates of a dataset of 932 males and 863 females at a single SNP and a phenotype for each setting in Table 7.

Table 7. Continuous phenotype power analysis sampling design

Phenotype distributions N(Mean, 169)					Number of males	Number of females	Male allele frequency	Female allele frequency
Mean value for males		Mean value for females						
Genotype A	Genotype B	Genotype AA	Genotype AB	Genotype BB				
15	16	15	16	17	932	863	0.30	0.30
15	17							
15	16	15	16	17	932	863	0.33	0.30
15	17							
15	16	15	16	17	932	863	0.37	0.30
15	17							
15	16	15	16	17	932	863	0.30	0.40
15	17							
15	16	15	16	17	932	863	0.45	0.30
15	17							

3.4.2 Design of Experiments

In our experiments, we ran 6 regression models, Clayton's statistic (equation 19) for binary and continuous phenotypes, and Zheng et al.'s statistic (equation 20) for binary phenotypes. We conducted experiments to find the effects of those phenotypic and genotypic differences mentioned above on the association tests under null and alternative hypotheses. We compared type I error rates and power of the analysis methods. We used all balanced and unbalanced datasets described above.

For type I error rate analyses of binary and continuous phenotypes, first, we calculated the type I error rate with only real chromosome-wide SNPs, then we only included simulated 120 SNPs, and we combined real and simulated SNPs and check the type I error rates of the tests.

For power analyses of continuous phenotypes, we ran the analyses with the datasets where male phenotypes come from the same distributions as homozygote female phenotypes. We also ran with the datasets having one male phenotype coming from the same distribution as heterozygote female phenotype. Furthermore we considered allele frequency differences between males and females.

For power analyses of binary phenotypes, we ran the analyses with the datasets where the number of female and male cases and controls are fixed and case and control allele frequencies are the same between genders. Then we used the datasets where case and control allele frequencies are different between genders.

3.5 RESULTS

First, we examined type I error rates of all models for binary phenotypes (see table 2). In a balanced design, for 12,242 real chromosome-wide SNPs or for the combination of real and 120 simulated SNPs (where the allele frequency differences between males and females are extremely high), we observed that all type I error rates fall in the Bradley's liberal criterion range of 0.025 and 0.075 [Bradley 1978]. When we only analyzed the simulated SNPs in the balanced dataset, the only test having type I error rate (0.008) (Table B.1) out of Bradley's criterion range was the logistic regression analysis where only genotype is in the model and male genotypes are coded as (0, 1). In an unbalanced design, when male genotypes were coded as (0,1), we observed that regression type I error rates may be inflated (Table B.1). If female case/control ratio is close to male case/control ratio (±0.05 in our experiments), then all tests have type I error rates in the range (0.025, 0.075) for real SNPs or combined set of real and simulated SNPs. However, if the ratios are not similar, the type I error may get as high as 1 if a sex covariate is not included. Table 8 shows the type I error rates of tests where data consists of both real and simulated SNPs. We provide a detailed summary of the binary phenotype type I error analysis results in Table B.1. When we ran the analyses only with simulated SNPs, the type I error rates can get higher than 0.05 and than the analyses where only real SNPs or combination of real and simulated SNPs are used.

Table 8. Type I error rates with both real and simulated SNPs

	~ Geno (0/1)	~ Geno (0/2)	+ Sex (0/1)	+ Sex (0/2)	Clayton	Zheng et al.
Balanced	0.038	0.045	0.052	0.044	0.055	0.046
Unbalanced	**0.867**	0.054	0.048	0.047	0.047	0.049

Unbalanced dataset include 393 female cases, 150 female controls, 150 male cases and 481 male controls whereas balanced dataset includes 393 female cases, 470 female controls, 451 male cases, and 481 male controls. Clayton results refer to Clayton's statistic in equation 19.

In the binary phenotype power analysis, we observed that when the dataset is balanced, the regression analysis having only genotype as an independent variable and male genotypes coded as (0,1) has the lowest power among all methods (Table B.2). For some of the unbalanced datasets analysis, that regression model has very high power but this is not reliable because we observed very high type I error rates in the same sampling settings (Table B.2). Under all of the sampling settings, the regression model where male genotypes are coded as (0,2) and Clayton's statistic (equation 19) have similar powers.

Secondly, we evaluated the tests with continuous phenotypes. For type I error analysis we observed type I error rate around 0.05 for two regression models (equations 12 and 13), Clayton's statistics when there is only the set of real SNPs or the combined set of real and simulated SNPs. However, when we considered only simulated 120 SNPs, then the type I error rates may be out of Bradley's liberal criterion range. A detailed summary of results can be seen in Table B.3.

In continuous phenotype power analysis (Table 9), we observed that when the male genotypes come from the same distribution as corresponding homozygote females, all other methods have higher power than the regression model having only genotype as independent variable with males coded as (0,1). However, when one of the male genotype has the same distribution as heterozygote females, then the highest power belongs to the regression model having only genotype as independent variable with males coded as (0,1). When the allele frequencies are extremely different in males and females, Clayton's statistic has slightly higher power than the other methods (Table B.4).

Table 9. Continuous phenotype power analysis results

Homozygote females and males from:	~Geno (0/1)	~Geno(0/2)	+Sex (0/1)	+Sex (0/2)	+Sex+Sex*Geno (0/1)	+Sex+Sex*Geno (0/2)	Clayton
Same distribution	0.52	0.725	0.655	0.73	0.66	0.66	0.73
Different distributions	0.58	0.35	0.39	0.34	0.33	0.33	0.36

As stated in Clayton's paper [Clayton 2008], we also conclude that Clayton's test has almost 0.05 (0.046) expected type I error rate when we combine real and the simulated SNPs with extreme allele frequency differences. Moreover, as can be seen from the Q-Q plot of the p-values from the preterm birth dataset with chromosome-wide SNPs and a real phenotype in Figure 8, there is not an extreme deviation in the distribution of observed p-values from the expected p-values. However, when we examined the results SNP by SNP, we observed that Clayton's statistic is affected if the data are unbalanced and the allele frequency differences in

males and females are extremely high (Figure 9). The smallest p values correspond to the SNPs having extreme allele frequency differences.

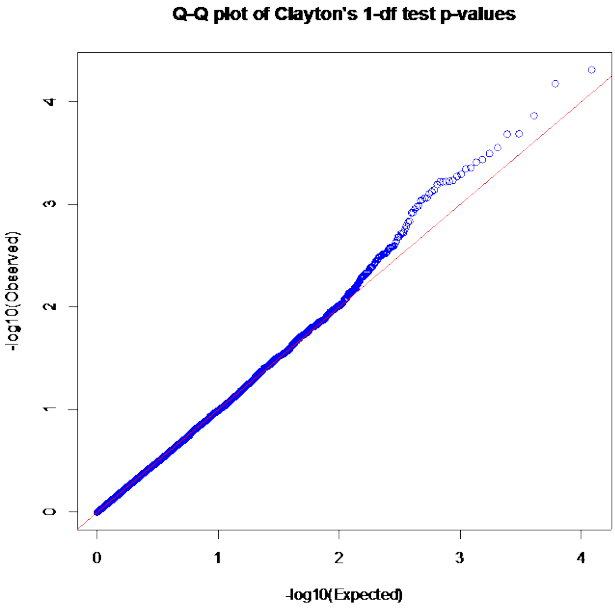

Figure 8. Q-Q plot of Clayton's statistic (equation 19) p values applied to the original preterm birth data chromosome-wide

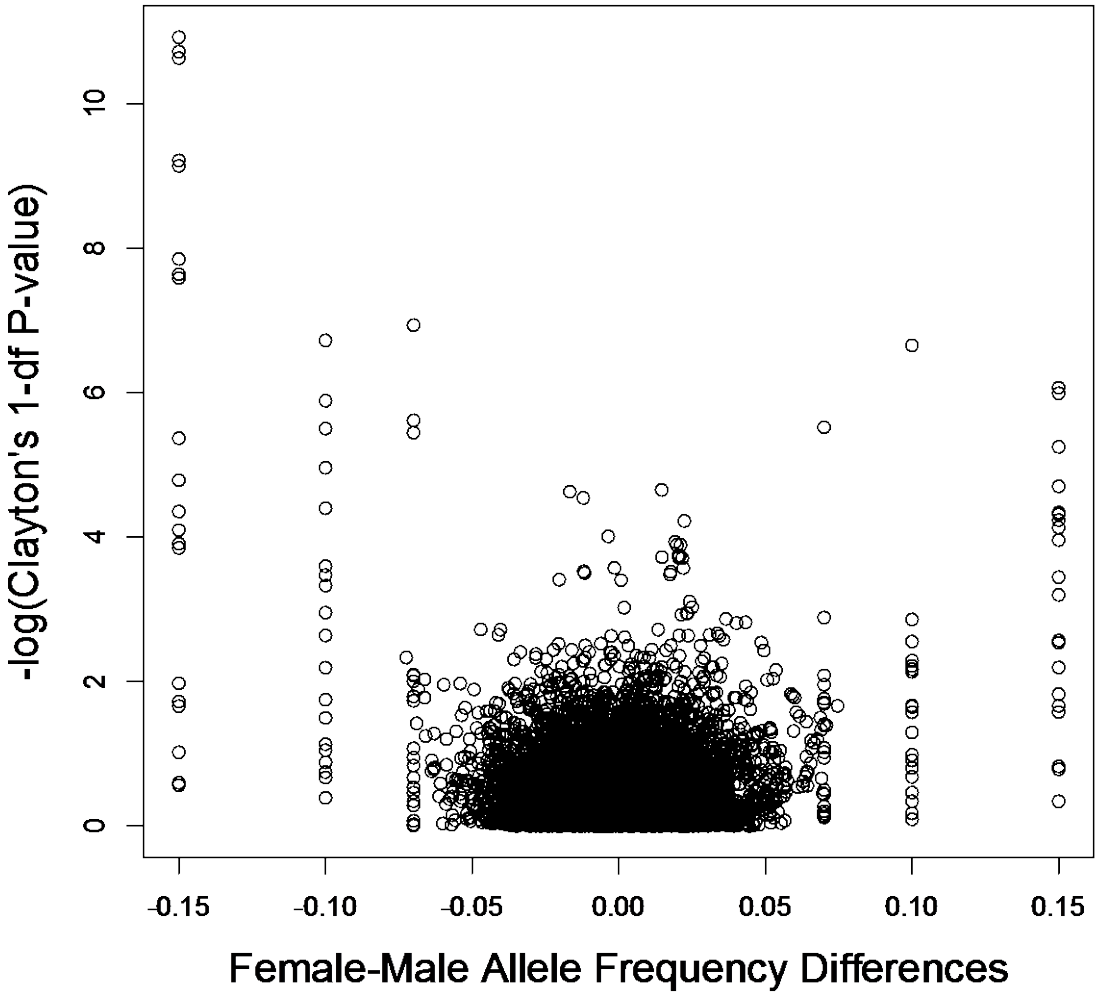

Figure 9. Clayton statistic (equation 19) p values with combined set of real and simulated SNPs vs female-male allele frequency differences under null hypothesis

3.6 X INACTIVATION

None of the models implied by the models above are quite correct because of X inactivation. X inactivation is transcriptionally silencing one of the X chromosomes in a complex manner in females [Lyon 1961]. X chromosome statistics have been modeled by assuming that one X chromosome gets dropped at random [Clayton 2009] so that AB females have a 50% chance of being A and 50% chance of being B. In this case, the variance and the mean would be estimated from the values of genotype 'A' and genotype 'B'. We propose a more realistic model where the allele silenced is not uniform throughout the organism- it may or may not even be uniform at a particular tissue. We suggest that a heterozygote female could randomly be anything between pure A and pure B. If this is the case, the variance and the mean would not be obvious as in the case above. Variance would be much higher than the one we could estimate from the sample. We studied the behavior of the statistics under different X-inactivation models. In quantitative traits, most X-inactivation models result in higher trait variance for heterozygote females than for homozygote males or females. In theory, the variance difference could increase the type I error of traditional regression-based tests. To test the potential effects of different variances in males and females, we simulated continuous phenotype variables with same and genotype-specific variances. We compared the Clayton's statistic [Clayton 2008], which is indicated as an optimal statistic for X chromosomal data, and regression methods in terms of type I error and power under X-inactivation models, where randomly one of the X chromosomes in females is transcriptionally inactivated, by using simulated datasets. From these experiments, we observed that type I error rates and powers of the robust regression analyses and Clayton's analysis are close (Table 10).

Table 10. Clayton and regression results based on simulated 1000 200-sample datasets

Phenotype distribution			*Test*		
Heterozygote females	*Homozygote females and males*	*Probability*	*Clayton*	*Robust regression*	*Linear regression*
N(11, 30)	N(10+G, 8)	*Power*	0.969	0.987	0.957
N(10, 30)	N(10, 8)	*Type I error*	0.021	0.022	0.012

G takes values (0,1,2) for female genotypes (AA,AB,BB) and (0,2) for male genotypes (A,B).

3.7 DISCUSSION

There is an extreme lack of genome-wide analysis results on the X chromosome in the literature. Failure to analyze X chromosome data at all is obviously less than ideal, and can lead to missed discoveries – for example, the first step in the SNP quality control process in a GWA for diabetic nephropathy was to remove the X chromosome [Pezzolesi, et al. 2009], but when the data set was submitted to the database of genotypes and phenotypes (dbGAP), the standard pre-compute analysis by dbGAP discovered that the X-linked SNP rs16997315 was strongly associated with a p-value of 4.7×10^{-11} (according to the Phenotype-Genotype Integrator website from NCBI). Even if the X chromosome data are analyzed, there may be some suboptimal statistics used. To analyze X chromosomal data, specialized analysis methods are needed. Although there are some statistics developed for X chromosome analysis, they assume relatively simple genetic models. Moreover, these statistics are seldom used for real data analysis, at least partly because their statistical properties (strengths and weaknesses) are not well understood.

In this study we aimed to extensively evaluate two X chromosome association test statistics and compared them with regression models by using real and simulated datasets under various genetic models. In our comprehensive simulation study, we can conclude that male genotypes on X chromosome should be treated as homozygote females as Clayton suggested [Clayton 2008; Clayton 2009]. In our case, that is coding male phenotypes as (0,2) for regression analysis when additive model is assumed. Adding sex as a covariate can also solve issues related to the X chromosome. For unbalanced sets, controlling for sex may result better. However, if male phenotypes follow the same distributions as homozygote females, then having sex as a covariate may reduce power.

For binary phenotypes, if the dataset is balanced, then both of the X chromosome test statistics and regression models where male genotypes are coded as (0,2), can handle even if there are extreme differences in allele frequencies. If the dataset is unbalanced, then allele frequency differences between genders may affect the results and may lead to false positives.

For continuous phenotypes, if female and male phenotypes come from different distributions, allele frequency differences between genders in SNPs again may cause false positive results. In that situation, sex variable can be added as a covariate into analyses. Moreover, we observed that tests are more powerful when male genotypes follow the distribution of corresponding homozygote female phenotypes.

For X inactivation models, although Clayton's test performs similarly as robust regression methods, both seem to be too conservative. Further investigation of the tests under X inactivation models may be needed.

As future research, further development of test statistics appropriate for X chromosome imputed genotype data and sequence data is necessary. Imputed SNPs on X chromosome are

almost never included in the analysis. Most imputation software produces genotype results in two different forms: posterior genotype likelihoods (PGL) and best-guessed genotypes [Marchini and Howie 2010]. Most analysis methods use PGL to calculate a dosage, defined as a weighted sum of the PGL. Statistics have also been developed to use the PGL directly. For X-chromosome data best-guessed genotypes can be used in any association methods, and dosages can be used in any of the regression-based methods by substituting the continuous dosage variable for the (0,1,2) genotype variable. Direct PGL methods have not been developed. I plan to extend PGL methods to the X chromosome and test the statistics using simulated data sets.

Another gap involves methods for sequenced X-chromosome data. If an allele of an X-chromosome SNP is very rare, in females we would expect to see mostly AA individuals with a few ABs and maybe one or two BBs. The (0,1,2) model is not appropriate for this type of data because BB individuals will be influential points in the regression and can bias the analysis. Therefore, females should be coded as (0,1), where AB and BB females are 1. However, for males, it is not obvious whether (0,1) or (0,2) coding is appropriate.

4.0 DISCUSSION

At the beginning of a GWAS, during data cleaning, many SNPs are left out of the study. Some of these data are discarded because of being low-quality or contaminated. However, there is a considerable amount of discarded data which may carry information regarding association. These data are discarded because of not being standard, or the analysis methods are not directly applicable to them. In this dissertation, my aim is to find these potentially informative GWAS data and to investigate the analysis methods for these data.

4.1 IDENTIFICATION OF NULL-ALLELE SNPS

Widely used genotype calling algorithms assume biallelic SNPs and the genotypes used in GWAS are called under this assumption. However, there may be some SNPs which fail to be included in the analyses under biallelic assumptions because they may have a possible third allele. These SNPs may carry important scientific information regarding the trait of interest. In this dissertation, I developed a procedure to detect 'null-allele' SNPs genome-wide. Null-allele can be a deletion or a third allele not hybridizing to standard probes. I showed every step of my procedure and also provided a model developed using classification and regression trees (CART) approach. A researcher who wants to find null-alleles in his/her GWAS can follow the methods described in Chapter 2 or he/she can calculate the statistics and apply the CART model provided

to identify null-allele SNPs. Identifying null-allele SNPs is important to increase number of SNPs in the study and to investigate regions for potentially functional copy number variants. My procedure strictly looks for null-allele SNPs and is based on genotype data generated by the Illumina platform. A researcher should be careful about the platform used for his/her data when applying my procedure. The data from other platforms may need some processing before it can be analyzed by my procedure. For example, applying a different clustering algorithm to the intensities may be necessary.

4.2 X CHROMOSOME ASSOCIATION TESTS

Due to the special features of X-chromosome, it is usually not analyzed or even if it is analyzed, suboptimal statistics may be used. X chromosome represents approximately 5% of the DNA. Not analyzing X chromosome data is not appropriate; it may cause failure to discover truly associated variants. In this dissertation, I comprehensively evaluated two X chromosome test statistics and compared them with regression models by using simulated and real chromosome-wide data sets.

First of all, I conclude that under an additive model assumption, male genotypes should be treated as homozygote females. In other words, the coding for male genotypes should be (0,2). This is also stated by Clayton [Clayton 2008]. I also observed that having sex as a covariate in the model could also solve the problem with scaling. Especially if there are SNPs having extremely large allele frequency differences between males and females in the dataset to be analyzed, sex should be included as a covariate instead of only just coding males as (0, 2) (Tables B.1 and B.2). Zheng et al.'s statistic can handle allele frequency differences as expected because the statistic estimates allele frequencies for males and females separately. However, in

this case, Clayton's statistic with sex covariate may be preferred according to the simulation results.

The phenotype distribution of males and females may also affect the analysis results of the statistics. A researcher should be careful about the assumptions before deciding the test statistic to apply. For instance if we can not assume that a male genotype comes from the same distribution as homozygote females, then either Clayton's or Zheng et al.'s test statistic may not be the most powerful alternative to use for the association analysis.

Male/female and case/control ratios in the dataset may also be considered before analysis. Although, the main aim was not comparing data designs in this dissertation, I observed that the type I error rate may be different from statistic to statistic especially when there are huge allele frequency differences in all SNPs in the dataset (Table B.1). Moreover, the magnitude of male/female and case/control ratios may affect the results.

4.3 OVERALL CONCLUSION

Recovering potentially informative data in terms of statistical point of view should increase the power. Discarding these data is not ideal for biological reasons either. Null-allele SNPs are not contaminated or low-quality, they are only different than what we are used to seeing. Moreover, since they are 'different' than usual SNPs, we may expect to see associations with the trait, or we may want to investigate the region for copy number variations. Likewise, not analyzing any X-chromosomal data, which may contain up to 1,400 genes, is not appropriate. Dropping the whole chromosome from the analysis may cause one to miss important associations.

In this dissertation, not only I wanted to point out that recovering GWAS data is important, but also I propose a practical method applied during data cleaning to save null-allele SNPs and show analyzing X-chromosome data is not impossible. As can be seen in chapter 2, my method for null-allele SNPs works well. I also leave the options to researchers: they can use the steps that I explained in detail to find null-allele SNPs, or they can apply the model I reported. Moreover, they can make use of the list of null-allele SNPs genome-wide that I can provide. Therefore, even though finding null-allele SNPs without any guidance may require a lot of effort, applying the proposed methods, which are shown to work well in chapter 2, would make the process much easier. However, analyzing null-allele SNPs would require different analysis methods, which is not the main aim of this dissertation. On the other hand, analyzing X-chromosome data is not that difficult and such analyses should be included in every GWAS report. As shown in chapter 3, the proposed X-chromosome specific test statistics would work well under certain assumptions, also regular regression methods can be used to analyze them. Implementing the test statistics into widely-used GWAS analysis programs would help researchers to analyze and report the X-chromosome association results.

4.4 FUTURE WORK

The projects in this dissertation can be taken further for additional improvements in GWAS analysis. I plan to apply the null-allele identification procedure to other real datasets and pedigree datasets. Moreover, association tests specific to null-allele SNPs could be investigated. The analysis methods available for multi-allelic SNPs might be applicable to null-allele SNPs.

For the X chromosome association statistics, the simulations can be extended under different genetic models besides the additive model. Also, each statistic can be evaluated for different sampling designs. Further development of test statistics for imputed and/or sequence data could be studied.

APPENDIX A. DERIVATION OF GENOTYPE FREQUENCIES AND THE EXPECTED VALUE OF CHI-SQUARED TEST

We first derive the expected numbers of the called genotypes AA, AB, and BB that are seen when a null-allele SNP is called under the standard di-allelic model.

Assume a perfect null-allele system with three alleles, A, B and N, six genotypes, AA, AN, AB, BB, BN, NN, and four clusters: [AA, AN], AB, [BB, BN], NN. Let the total number of genotypes observed be $n = n_{[AA,AN]} + n_{AB} + n_{[BB,BN]} + n_{NN}$. The total number of called genotypes is then $n' = n_{[AA,AN]} + n_{AB} + n_{[BB,BN]} = n - n_{NN}$. When a null-allele SNP is called under the standard model, NN genotypes are discarded, and genotypes AN and BN are counted as AA and BB respectively, so the difference between *n* and *n'* is only n_{NN}. If we let *p*, *q* and *r* be the allele frequencies of A, B and N respectively in the perfect null-allele system, then the expected numbers of called genotypes are $(p^2 + 2pr)n, 2pqn, (q^2 + 2qr)n,$ and r^2n for the clusters [AA, AN], AB, [BB, BN] and NN respectively.

These genotype frequencies also allow us to derive the expected value of the chi-squared test statistic for the standard (HWE) model. Assume that our counts have been generated from a perfect, null-allele system. Then A allele frequency estimate we would obtain if a standard two-allele model is applied to those data is

$$p_S = \frac{2n_{[AA,AN]}+n_{AB}}{2n'} = \frac{2n(p^2+2pr)+n(2pq)}{2(n-nr^2)} = \frac{p}{(1-r)} \quad \text{(A1)}$$

Under the standard model assumption, $p_S+q_S=1$. Then,

$$q_s = 1 - p_s = q/(1-r) \quad \text{(A2)}$$

Chakraborty et al. [Chakraborty, et al. 1992] previously derived equation A1 for the more general multi-allelic case.

If we ignore the null genotypes (NN), the expected values are $p_S^2 n', 2p_S q_S n' and\ q_S^2 n'$ of the genotypes [AA, AN], AB and [BB, BN] respectively. Then the χ^2 goodness of fit test statistic for standard model SNP, where NN genotypes are ignored, is:

$$\chi^2 = \sum_{genotypes} \frac{(Observed - Expected)^2}{Expected} \quad \text{(A3)}$$

$$\chi^2 = \frac{(4r^2)(1-r)n}{1+r} \quad \text{(A4)}$$

Equation A4 is equivalent to χ^2_{std}. In the TriTyper study, Franke et al. also did the calculations above assuming HWE for three alleles. However, we ended up with a different χ^2 equation than theirs using the same n and r.

APPENDIX B. SUPPLEMENTARY TABLES

Table B.1. Type I error rates of the methods with binary phenotypes

Sample size			Regression models								
Females (Case/Cont)	Males (Case/Cont)	Data*	I: P~G(0,1)	II: P~G(0,2)	III: P~G(0,1)+S	IV: P~G(0,2)+S	V[#]: P~G(0,1)+S+I	VI[#]: P~G(0,2)+S+I	Zheng	Clayton (1-df)	Clayton + Sex
393/470	451/481	R	0.048	0.043	0.054	0.043	0.044	0.044	0.046	0.055	0.055
		S	0.008	0.058	0.067	0.058	0.05	0.05	0.067	0.067	0.067
		R+S	0.038	0.045	0.052	0.044	0.044	0.044	0.046	0.055	0.055
150/470	451/481	R	**0.707**	0.047	0.043	0.044	0.043	0.043	0.044	0.053	0.053
		S	**0.925**	0.342	0.067	0.05	0.05	0.05	0.05	0.092	0.058
		R+S	**0.710**	0.050	0.043	0.044	0.043	0.043	0.044	0.053	0.053
393/470	150/481	R	**0.645**	0.043	0.046	0.048	0.048	0.048	0.047	0.056	0.057
		S	**0.883**	0.358	0.075	0.025	0.067	0.067	0.058	0.075	0.033
		R+S	**0.647**	0.047	0.046	0.047	0.048	0.048	0.047	0.057	0.056
393/150	451/481	R	**0.677**	0.048	0.047	0.045	0.045	0.045	0.046	0.055	0.054
		S	**0.908**	0.333	0.008	0.017	0.017	0.017	0.008	0.05	0.058
		R+S	**0.680**	0.051	0.047	0.045	0.045	0.044	0.046	0.055	0.054
393/470	451/150	R	**0.740**	0.038	0.048	0.045	0.048	0.048	0.046	0.053	0.053
		S	**0.975**	0.45	0.083	0.033	0.058	0.058	0.075	0.05	0.042
		R+S	**0.742**	0.042	0.049	0.045	0.048	0.048	0.046	0.053	0.053
393/150	451/150	R	0.051	0.048	0.048	0.049	0.045	0.045	0.051	0.052	0.051
		S	0.042	0.017	0.05	0.033	0.075	0.075	0.033	0.067	0.033

Table B.1 continued

		R+S	0.051	0.047	0.048	0.049	0.045	0.046	0.050	0.052	0.051
150/470	150/481	R	0.034	0.048	0.050	0.049	0.052	0.052	0.051	0.052	0.051
		S	0.033	0.05	0.050	0.05	0.067	0.067	0.058	0.033	0.033
		R+S	0.034	0.048	0.050	0.049	0.052	0.052	0.051	0.051	0.051
393/150	150/481	R	**0.866**	0.047	0.048	0.048	0.045	0.045	0.049	0.047	0.047
		S	**1**	0.733	0.033	0.042	0.05	0.05	0.025	0.058	0.033
		R+S	**0.867**	0.054	0.048	0.047	0.045	0.045	0.049	0.047	0.048
150/470	451/150	R	**0.890**	0.042	0.046	0.041	0.044	0.044	0.043	0.056	0.056
		S	**1**	**0.7**	0.058	0.05	0.042	0.042	0.092	0.1	0.075
		R+S	**0.891**	0.048	0.046	0.041	0.044	0.044	0.043	0.057	0.056

Datasets include 12,242 real SNPs and/or 120 simulated SNPs with the allele frequency differences 0.07, 0.10 and 0.15.

*R: Real SNPs, *S: Simulated SNPs
I: Phenotype ~ Genotype (males coded as (0,1))
II: Phenotype ~ Genotype (males coded as (0,2))
III: Phenotype ~ Genotype + Sex (males coded as (0,1))
IV: Phenotype ~ Genotype + Sex (males coded as (0,2))
V: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,1))
VI: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,2))
#: Results of 2 df test statistic which compares the models with the model Phenotype ~ Sex
Bold rates are unacceptably high.

Table B.2. Power of the methods with continuous phenotypes based on 200 simulated phenotype and genotype pairs

Phenotype distributions N(Mean, Variance=169)					Sample size		Allele frequency		Regression models							
Mean for males		Mean for females			Males	Females	Male	Female	I: P~ G(0,1)	II: P~ G(0,2)	III: P~ G(0,1)+S	IV: P~ G(0,2)+S	V[#]: P~ G(0,1)+S+I	VI[#]: P~ G(0,2)+S+I	Clayton (1-df)	Clayton + Sex
Genotype A	Genotype B	Genotype AA	Genotype AB	Genotype BB												
15	16	15	16	17	932	863	0.3	0.3	0.580	0.350	0.390	0.340	0.330	0.330	0.360	0.340
15	17	15	16	17					0.520	0.725	0.655	0.730	0.660	0.660	0.730	0.730
15	16	15	16	17	932	863	0.33	0.3	0.585	0.430	0.495	0.425	0.360	0.360	0.460	0.425
15	17	15	16	17					0.565	0.755	0.705	0.745	0.645	0.645	0.755	0.745
15	16	15	16	17	932	863	0.37	0.3	0.640	0.485	0.450	0.435	0.395	0.395	0.490	0.435
15	17	15	16	17					0.615	0.780	0.720	0.780	0.685	0.685	0.795	0.780
15	16	15	16	17	932	863	0.3	0.4	0.575	0.400	0.505	0.435	0.380	0.380	0.385	0.435
15	17	15	16	17					0.620	0.790	0.740	0.795	0.700	0.700	0.785	0.795
15	16	15	16	17	932	863	0.45	0.3	0.625	0.490	0.470	0.430	0.375	0.375	0.520	0.430
15	17	15	16	17					0.620	0.755	0.660	0.725	0.600	0.600	0.780	0.725

I: Phenotype ~ Genotype (males coded as (0,1))
II: Phenotype ~ Genotype (males coded as (0,2))
III: Phenotype ~ Genotype + Sex (males coded as (0,1))
IV: Phenotype ~ Genotype + Sex (males coded as (0,2))
V: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,1))
VI: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,2))
#: Results of 2 df test statistic which compares the models with the model Phenotype ~ Sex

Table B.3. Type I error rates of the methods with continuous phenotypes

Sample size				Regression models							
Females	Males	Data*	F-M pheno means (sd)	I: P~ G(0,1)	II: P~ G(0,2)	III: P~ G(0,1) +S	IV: P~ G(0,2)+ S	V[#]: P~ G(0,1)+S+I	VI[#]: P~ G(0,2)+S+I	Clayton (1-df)	Clayton + Sex
863	932	R	15-15 (3)	0.250	0.050	0.049	0.050	0.052	0.052	0.051	0.05
		S		0.50	0.083	0.033	0.033	0.033	0.033	0.117	0.033
		R+S		0.253	0.050	0.049	0.050	0.052	0.052	0.052	0.05
863	932	R	18-15 (3)	0.884	0.053	0.054	0.053	0.053	0.053	0.055	0.053
		S		1	0.742	0.025	0.025	0.025	0.025	0.767	0.025
		R+S		0.885	0.059	0.053	0.053	0.053	0.053	0.061	0.052

Datasets include 12,242 real SNPs and/or 120 simulated SNPs with the allele frequency differences 0.07, 0.10 and 0.15.

*R: Real SNPs, *S: Simulated SNPs
I: Phenotype ~ Genotype (males coded as (0,1))
II: Phenotype ~ Genotype (males coded as (0,2))
III: Phenotype ~ Genotype + Sex (males coded as (0,1))
IV: Phenotype ~ Genotype + Sex (males coded as (0,2))
V: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,1))
VI: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,2))
#: Results of 2 df test statistic which compares the models with the model Phenotype ~ Sex

Table B.4. Power rates of the methods with binary phenotypes based on 200 simulated phenotype and genotype pairs

Sample size		Allele frequency		Regression models								
Females (Case/Cont)	Males (Case/Cont)	M/F controls	M/F cases	I: P~G(0,1)	II: P~G(0,2)	III: P~G(0,1)+S	IV: P~G(0,2)+S	V#: P~G(0,1)+S+I	VI#: P~G(0,2)+S+I	Zheng	Clayton (1-df)	Clayton + Sex
393/470	451/481	0.5/0.5	0.46/0.46	0.245	0.455	0.5	0.475	0.445	0.445	0.455	0.475	0.505
393/470	150/481			1	0.37	0.405	0.35	0.315	0.315	0.38	0.35	0.36
150/470	451/481			0.615	0.355	0.38	0.335	0.325	0.325	0.35	0.335	0.35
393/470	451/150			0.715	0.425	0.435	0.385	0.32	0.32	0.435	0.385	0.405
393/150	451/481			1	0.365	0.375	0.315	0.27	0.27	0.36	0.315	0.36
393/150	451/150			0.11	0.305	0.325	0.31	0.245	0.245	0.305	0.31	0.325
393/150	150/481			1	0.315	0.24	0.245	0.195	0.195	0.325	0.245	0.26
150/470	150/481			0.23	0.22	0.255	0.225	0.23	0.23	0.225	0.23	0.245
150/470	451/150			1	0.355	0.26	0.25	0.21	0.21	0.355	0.25	0.275
393/470	451/481	0.5/0.53	0.46/0.49	0.375	0.53	0.505	0.515	0.435	0.435	0.53	0.515	0.52
393/470	150/481			1	0.335	0.45	0.43	0.37	0.37	0.345	0.43	0.42
150/470	451/481			0.4	0.53	0.395	0.385	0.34	0.34	0.52	0.385	0.405
393/470	451/150			0.515	0.585	0.46	0.425	0.355	0.355	0.595	0.43	0.445
393/150	451/481			1	0.325	0.425	0.385	0.34	0.34	0.325	0.39	0.41
393/150	451/150			0.16	0.3	0.285	0.29	0.25	0.25	0.305	0.29	0.3
393/150	150/481			1	0.185	0.325	0.31	0.235	0.235	0.19	0.31	0.33
150/470	150/481			0.33	0.34	0.35	0.345	0.25	0.25	0.345	0.345	0.35

Table B.4 continued

150/470	451/150			0.995	0.66	0.3	0.285	0.235	0.235	0.66	0.285	0.295

I: Penotype ~ Genotype (males coded as (0,1))
II: Phenotype ~ Genotype (males coded as (0,2))
III: Phenotype ~ Genotype + Sex (males coded as (0,1))
IV: Phenotype ~ Genotype + Sex (males coded as (0,2))
V: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,1))
VI: Phenotype ~ Genotype + Sex + Genotype*Sex (males coded as (0,2))
#: Results of 2 df test statistic which compares the models with the model Phenotype ~ Sex

BIBLIOGRAPHY

Bradley JV. 1978. Robustness? British Journal of Mathematicala and Statistical Psychology 31(2):144-152.

Brent RP. 1972. Algorithms for minimization without derivatives. Englewood Cliffs, N.J., Prentice-Hall.

Broyden CG. 1970. The convergence of a class of double-rank minimization algorithms. Journal of the Institute of Mathematics and Its Applications 6:76-90.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR. 1993. Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am J Hum Genet 52(5):922-7.

Chakraborty R, De Andrade M, Daiger SP, Budowle B. 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56(Pt 1):45-57.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE and others. 2007. Replicating genotype-phenotype associations. Nature 447(7145):655-60.

Clayton D. 2008. Testing for association on the X chromosome. Biostatistics 9(4):593-600.

Clayton DG. 2009. Sex chromosomes and genetic association studies. Genome medicine 1(11):110.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37-46.

Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, Zackai EH, Deardorff MA, Krantz ID, Hakonarson H, Spinner NB. 2010. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum Mol Genet 19(7):1263-75.

Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF and others. 2010. The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 34(4):364-72.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55(4):997-1004.

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. 2011. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.

Fletcher R. 1970. A New Approach to Variable Metric Algorithms. Computer Journal 13(3):317-322.

Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zhernakova A, Hunt KA, Blauw HM, van den Berg LH, Ophoff R, Deloukas P and others. 2008. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. American Journal of Human Genetics 82(6):1316-33.

Goldfarb D. 1970. A Family of Variable-Metric Methods Derived by Variational Means. Mathematics of Computation 24(109):23-26.

Hickey PF, Bahlo M. 2011. X chromosome association testing in genome wide association studies. Genetic epidemiology 35(7):664-70.

Kumasaka N, Fujisawa H, Hosono N, Okada Y, Takahashi A, Nakamura Y, Kubo M, Kamatani N. 2011. PlatinumCNV: a Bayesian Gaussian mixture model for genotyping copy number polymorphisms using SNP array signal intensity data. Genet Epidemiol 35(8):831-44.

Kuo CL, Feingold E. 2010. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol 34(3):246-53.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ and others. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34(6):591-602.

Lehmann T, Hawley WA, Collins FH. 1996. An evaluation of evolutionary constraints on microsatellite loci using null alleles. Genetics 144(3):1155-63.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311-21.

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News 2(3):18-22.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature 190:372-3.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11(7):499-511.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188-93.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J and others. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16(9):1136-48.

Pemberton JM, Slate J, Bancroft DR, Barrett JA. 1995. Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Mol Ecol 4(2):249-52.

Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB, Ng DP, Placha G, Canani LH, Bochenski J and others. 2009. Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. Diabetes 58(6):1403-10.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81(3):559-75.

RCoreTeam. 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Shanno DF. 1970. Conditioning of Quasi-Newton Methods for Function Minimization. Mathematics of Computation 24(111):647-656.

Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M. 2008. Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC Bioinformatics 9:409.

Therneau TM, Atkinson B. 2012. rpart: Recursive Partitioning.

Vens M, Schillert A, Konig IR, Ziegler A. 2009. Look who is calling: a comparison of genotype calling algorithms. BMC Proc 3 Suppl 7:S59.

Wakefield J. 2010. Bayesian methods for examining Hardy-Weinberg equilibrium. Biometrics 66(1):257-65.

Wise AL, Gyi L, Manolio TA. 2011. Analyze X: A Comparison of X Chromosome and Autosomal Results in GWAS. 61st Annual Meeting of the American Society of Human Genetics. Montréal, Canada.

Zheng G. 2012. Analysis of genetic association studies. Statistics for biology and health,. New York: Springer.

Zheng G, Joo J, Zhang C, Geller NL. 2007. Testing association for markers on the X chromosome. Genet Epidemiol 31(8):834-43.