POWER AND TYPE 1 ERROR FOR LARGE PEDIGREE ANALYSES OF BINARY TRAITS

By

Anna Christine Cummings

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Interdisciplinary Studies: Applied Statistics

December, 2012

Nashville, Tennessee

Approved:

Professor Jonathan L. Haines

Professor Tricia A. Thornton-Wells

Professor William S. Bush

TABLE OF CONTENTS

Page

LIST OF TABLES iii

Chapter

I. **INTRODUCTION 1**

Isolated Populations for Genetic Studies 1
Association 3
Linkage 6

II. **USING SIMULATIONS TO EVALUATE TYPE 1 ERROR AND POWER FOR ASSOCIATION AND LINKAGE ANALYSES OF AN AMISH PEDIGREE 10**

Introduction 10
Methods 11
Results 13
Discussion 19

REFERENCES 22

LIST OF TABLES

Table Page

2.1 Average percentage of times (power) per model disease SNP was under p-value thresholds when running MQLS on whole simulated pedigrees……14

2.2 Average percentage of times (power) per model disease SNP was under p-value thresholds when running MQLS on sub-pedigrees 15

2.3 Percentage of SNPs (type 1 error) above HLOD thresholds using PedCut followed by two-point parametric linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics 16

2.4 Percentage of times (power) disease SNP crossed parametric HLOD or nonparametric LOD thresholds using PedCut followed by Merlin two-point parametric and nonparametric linkage analyses 17

2.5 Percentage of SNPs (type 1 error) above parametric HLOD and nonparametric LOD thresholds using PedCut followed by multipoint linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics 18

2.6 Power to detect parametric HLOD and nonparametric LOD thresholds using PedCut followed by multipoint parametric linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics 19

CHAPTER I

INTRODUCTION

Isolated Populations for Genetic Studies

While studies involving unrelated cases and controls have become popular for studying complex genetic phenotypes, particularly with the widely used genome-wide SNP (single nucleotide polymorphism) marker sets, human genetic studies have historically examined families to map disease genes. The family structure allows the observation of the co-occurrence of genomic transmission with disease, which case-control association studies attempt to indirectly measure in a population. The traditionally used tool for studying families is linkage analysis, which studies the cosegregation of genetic markers and a phenotype within pedigrees. Studying families with multiple affected individuals using linkage analysis was extremely successful for mapping Mendelian diseases, and has also had some success with complex diseases (1-3). However, association studies, which look for differences in allele frequencies between cases and controls, have become more popular because of the lower costs of SNP genotyping. With the increasing popularity of association studies, family-based association designs have also advanced. Family-based designs-- both linkage and association-- continue to have utility for mapping complex genetic diseases, and the progress of the HapMap project to prompt genome-wide SNP genotyping has made these studies even more powerful.

Larger families provide more meioses to tease out which genetic markers are linked to disease and thereby make association analysis more powerful. Isolated founder populations provide extremely large family structures and have been successful in mapping genes for a variety of genetic diseases (4-7). Isolated founder populations can also reduce environmental noise since they typically share a common lifestyle and live in the same location. The isolated expansion of the population from a small number of founders restricts the introduction of new genetic variation**(8)**, so it can be expected that these unique groups' genomes would contain a more homogeneous set of disease risk genes. Many isolated populations have large families and often keep extensive genealogy records, making extended pedigree construction feasible.

A population isolate we have studied for many years is the Amish communities of Ohio and Indiana. This population was founded by Swiss Anabaptists fleeing religious persecution. They immigrated to the United States in two main waves starting in the early 1700's, which brought the first group to Pennsylvania. Then, in the early 1800's some of these immigrants moved to Holmes County, OH, while a second wave of immigration from Europe established more Amish communities in other areas of Ohio and Indiana (including Adams County). Later, Elkhart and LaGrange County Amish communities were started by some of the Amish from Pennsylvania and Ohio (including Holmes County) moving to these new locations (9-11).

The Amish marry almost exclusively within the community and have large families, providing pedigrees with multiple affected individuals for analyses. The Anabaptist Genealogy database (AGDB)(12;13) and the Swiss Anabaptist Genealogical Association (SAGA) keep thorough family history records, providing necessary and critical pedigree information. Because of their faith, the Amish lead a strict and

traditional lifestyle and, therefore, have more homogeneous environmental exposures than the general population.

We have been studying the Ohio and Indiana Amish communities to identify genes contributing to late-onset Alzheimer disease susceptibility (LOAD). Compared to the general population in which many genes are contributing to LOAD, the relatively homogeneous Amish population is likely to contain a smaller set of risk alleles, each with a theoretically increased population attributable risk, thereby increasing detection power. The relatively recent expansion of the population from a small number of original founders, plus isolation, results in this reduced amount of genetic variation (8).

To isolate disease genes in the Amish we have employed both linkage and association analyses to increase the ability to locate disease genes by tackling the problem from two different angles. Combining these approaches with large-scale SNP genotype data in the large pedigree structure increases our ability to localize disease genes. The large pedigree structure also increases the complexity of both association and linkage analyses.

Association

Isolated founder populations, such as the Amish, are unique family datasets in that their family units are interrelated and the population as a whole can be considered one family. This phenomenon reduces the validity of most family-based association methods such as the popular TDT (transmission disequilibrium test) (14) and PDT (pedigree disequilibrium test) (15;16) which assume independence between family units. In addition to the problem of non-independence, TDT only uses family triads (an

affected individual and both parents) so only part of the large pedigree dataset could be used, and the parents must have genotypes. With a late-onset phenotype like Alzheimer disease, parental genotypes are often not available. The PDT allows for more extensive pedigree structures to be used, but not as extensive as an Amish pedigree or a similar very large pedigree from a founder population.

To perform association analysis in any pedigree structure, even a large inbred pedigree, and to adopt a case-control based design which can be more powerful than family-based designs (17), Thornton and McPeek developed a quasi-likelihood score test, CC-QLS (case-control quasi-likelihood score test). The CC-QLS conditions on the pedigree structure by using kinship coefficients. A kinship coefficient is the probability that two alleles at a randomly chosen locus, one from individual i and the other from individual j, are identical by decent (i.e. came from the same common ancestor). The more related the individuals are, the more alleles they should share, and the higher the kinship coefficient will be between the two individuals. Because the Amish, like other population isolates, have reliable genealogy records, genetic sharing can be inferred using the known pedigree relationships when calculating kinship coefficients.

To improve CC-QLS, they went on to develop MQLS (modified quasi-likelihood score test) which has an even more optimal weighting scheme to increase power to detect an association. More specifically, MQLS uses unaffected controls and controls of unknown phenotype differently since it is less likely that unaffected controls will carry the risk allele. Secondly, MQLS uses phenotype data of samples without genotypes to optimize the weights of the relatives with genotypes. This optimization is based on the assumption that affected individuals with other affected relatives are more likely to carry a genetic risk factor than individuals without any affected relatives.

To evaluate type 1 error rates, Thornton and McPeek simulated a null SNP with three different allele frequencies in 60 moderately-sized pedigrees (each pedigree had three generations and sixteen individuals) and 200 unrelated controls. After performing 5,000 replicates they found that for each allele frequency setting, empirical type 1 error was ~0.05 and ~0.01 for nominal type 1 errors of 0.05 and 0.01, respectively. Therefore, they did not see any inflation of type 1 error.

To evaluate power they simulated 5,000 replicates of six different disease models in the same 200 unrelated controls and 60 pedigrees (each with 4, 5, or 6 cases). Each dataset had two or three simulated SNPs. Three of the models had two SNPs acting epistatically and dominantly. Two other models had three SNPs acting epistatically and dominantly. The sixth model had two SNPs acting epistatically with one SNP acting recessively and the other acting codominantly. They used a variety of penetrances; the highest in any model was 0.5. They calculated at least 71% power to detect a p-value ≤ 0.05 for all models. MQLS had the most difficulty detecting a significant association for the two-SNP epistatic model with one acting recessively and the other codominantly. The highest power (97%) to detect association was seen for a two-SNP epistatic model with both SNPs acting dominantly and when the disease allele frequency was high for both SNPs (0.5 and 0.4).

We have employed MQLS in many of our studies in the Amish (18;19) because the test can handle the entire 4998-member 13-generation pedigree structure; however it has been unclear to us how to estimate our type 1 error rates and power are in the large and complicated pedigree structure of our Amish dataset.

Linkage

Isolated founder populations also challenge the capabilities of available linkage analysis software available. The two main algorithms which have been developed and are implemented in software for linkage analysis are the Lander-Green (20) and Elston-Stewart (21) algorithms. The main practical difference between the two algorithms is the number of markers and the pedigree size that each algorithm can handle. The computational requirements when using the Elston-Stewart algorithm increase exponentially as the number of markers increases and increase linearly as the pedigree size increases. The Lander-Green algorithm can handle more markers because the computational complexity increases linearly as the number of markers increases but exponentially as the pedigree size increases. Therefore, the Elston-Stewart algorithm is more suitable for larger pedigrees, while the Lander-Green algorithm is more suitable for larger numbers of markers as found in SNP arrays. Neither algorithm is capable of handling genome-wide SNP (single nucleotide polymorphism) data in very large and complex pedigree structures.

Some of the top linkage programs which implement one or more of these algorithms are Vitesse (22), Allegro (23;24), Superlink (25-27), and Merlin (28). Vitesse applies the Elston-Stewart algorithm but also incorporates part of the Lander-Green algorithm by using inheritance vectors. Allegro is based on the Lander-Green method. Superlink incorporates both algorithms using a Bayesian approach. Merlin applies sparse binary trees to the Lander-Green algorithm to be able to successfully handle genome-wide SNP data. Merlin has been shown to outperform the other linkage

programs in computational time and ability to handle large numbers of markers (28;29). We have also found that Merlin is best suited for SNP data in our Amish pedigrees.

Despite the advantages of using Merlin, pedigree size and complexity is still a hindrance. Therefore, the only option to analyze genome-wide SNP data in large complex pedigrees like the Amish is to divide the pedigree into smaller sub-pedigrees. Methods for subdividing the pedigree into smaller more computationally feasible pedigrees include Greffa (30) and PedCut (31). Greffa requires several user-defined parameters and does not guarantee that all resulting subpedigrees will be handled by linkage programs. When Liu et al. (2008) compared the performance of Greffa to their program PedCut, they found that Greffa did not assign as many subjects of interest to subpedigrees and that the number of subjects of interest per subpedigree was smaller. We prefer PedCut for its straightforward and automatic approach which guarantees all subpedigrees will be computationally feasible for linkage analysis. The PedCut algorithm prioritizes subjects of interest (specified by the user) and their closest relatives (measured by kinship coefficients) to be included in the subpedigrees that are all within a user-specified bit-size limit. A bit-size is defined as two times the number of nonfounders (individuals with parents represented in the pedigree structure) minus the number of founders (individuals without parents represented in the pedigree structure) (32).

While necessary to perform linkage analysis, cutting the pedigree could potentially affect power and/or type 1 error of linkage results. Type 1 error has been shown to be inflated when consanguinity is underestimated or loops are broken in the pedigree when performing homozygosity mapping of recessive traits (33;34). Power has

been shown to be reduced when splitting the pedigree prior to quantitative trait linkage analysis (35).

Liu et al. (2007) performed a type 1 error analysis using GENEHUNTER and SIMWALK assuming a dominant model and calculated a 5% type 1 error rate for a LOD score of 3.64 when no disease locus was simulated (36). In a separate publication (31), Liu et al also performed a power analysis to compare power to detect linkage using subpedigrees derived from PedCut compared to subpedigrees derived from Greffa. They simulated completely penetrant 1-locus disease models with dominant, additive (with penetrances of 0.75, 0.5, and 0.25 for the heterozygous genotype), and recessive modes of inheritance. They also set the distance between the trait and marker loci to zero, which generated a scenario of perfect linkage, and only tested each model with the corresponding correct linkage analysis model. They calculated the expected LOD scores for each model using SIMLINK.

Because these 'perfect' scenarios rarely, if ever, exist in real-life analyses of complex genetic diseases, we wanted to examine power and type 1 error with more realistic simulated binary trait models in our Amish pedigree structure using the program Merlin. Ideally, we would run linkage analysis on the whole pedigree and the divided pedigree to compare the results. Because running linkage on the whole pedigree is not possible, but running MQLS on the whole pedigree is possible, one approach is to compare MQLS results on the whole pedigree versus the divided pedigree. We also wanted to examine power when the correct model is not specified for analysis, since we most often do not know what the true underlying model is.

In this thesis work I have simulated pedigrees with the same structure as the Amish to accomplish the following goals: 1) determine power and type 1 error rates

when using MQLS to test for association; 2) determine power and type 1 error when subdividing the pedigree into subpedigrees using Merlin and subsequently performing linkage analysis using Merlin; and 3) Compare power and type 1 error rates when applying MQLS to the entire pedigree structure versus subpedigrees used for linkage analysis.

CHAPTER 2

USING SIMULATIONS TO EVALUATE TYPE 1 ERROR AND POWER FOR ASSOCIATION AND LINKAGE ANALYSES OF AN AMISH PEDIGREE

Introduction

As discussed in Chapter I, complex pedigrees from isolated populations have gained popularity for genetics studies due to their pedigree size, genetic homogeneity, and environmental homogeneity (18;19;37). Despite their advantages, pedigree size and genetic homogeneity complicate analyses and can make results difficult to interpret. Association analyses must correct for the nonindependence of samples within families. In our genetic studies of the Amish, we have employed MQLS (modified quasi-likelihood score) (38) to test for association because it can handle large complex pedigrees and uses kinship coefficients to correct for relatedness. Pedigree size and complexity also present problems when running linkage analyses because even the best available linkage programs, such as Merlin(28), can only handle pedigrees under a certain size and complexity, defined by the bit-size (two times the number of non-founders minus the number of founders (32)). Therefore, we use PedCut(31) to generate sub-pedigrees with the maximal number of subjects of interest within a specified bit-size limit conductive to two-point and multipoint linkage analyses. GenomeSIMLA(39) is a forward-time population-based simulation package for generating large-scale SNP data in both case-control and family-based designs and has been adapted to efficiently produce SNP data in any pedigree structure given a pedigree template. We have implemented this extended version of GenomeSIMLA to evaluate

the power and false-positive rates for association and linkage analyses in an Amish pedigree structure.

Methods

Simulations

We extended GenomeSIMLA to generate complex pedigree structures based on a template pedigree. Once a population of chromosomes has been created, a collection of founders is drawn and are mated to produce all generations of the pedigree. Affection status is assigned by applying a penetrance function with the option of only assigning known phenotype and genotype data to the same individuals with known phenotype and genotype data in the template pedigree, maintaining a more realistic distribution of genotyped affected and unaffected individuals in the pedigree. We simulated a null disease model into 1000 pedigree replicates, each with 124 unlinked autosomal SNPs, using our recently published 4998-member Amish pedigree with almost identical affection status (798 genotyped) (19). Minor allele frequencies (MAFs) were randomized between 0.1 and 0.3 with a default MAF of 0.2, to approximate the mean MAF in the recent GWAS study of our Amish pedigree (19).

For studies of power, similar simulations were conducted with one of the 124 SNPs having either a dominant, recessive, or additive effect of odds ratios 1.1, 1.5, 2.0, or 5.0 on the phenotype, which generated 12 disease models. The minor allele frequency for the 'disease' SNP was held constant at 0.2. One thousand replicates were simulated for each disease model.

Analyses

We ran MQLS (software version 1.2) to test for association and used option '1' to include all individuals, cases, controls, and individuals with unknown phenotype, in the analyses. More recent versions (starting at version 1.5) of MQLS include a more robust variance estimator (40), which was not implemented in these analyses but would not likely make a significant difference in our results. We tallied the number of p-values below the relevant threshold in each of the replicates. For the type 1 error study any p-value below the threshold was included in the count, and for the power studies any p-value below the threshold at the 'disease' SNP was counted. The average number of p-values was then calculated across each set of 1000 replicates.

To generate sub-pedigrees within a bit-size limit of 24, we ran PedCut (31) in each of the simulated pedigrees using affected individuals and unaffected siblings of the affected individuals as subjects of interest. We ran two-point and multipoint parametric and nonparametric linkage analyses on the PedCut pedigrees using Merlin (28). Parametric HLOD scores were computed assuming affecteds-only autosomal dominant and recessive models of 0 penetrance for no disease allele and 0.0001 for 1 or 2 copies of the disease allele under the dominant model, and penetrances of 0 for 0 or 1 disease allele and 0.0001 for 2 disease alleles under the recessive model. A disease allele frequency of 1% was used to mimic our recently published genome-wide study. We would like to note a typographical error in that paper which misreported the disease allele frequency to be 10% (19). Nonparametric calculations (LOD*) were computed using the NP-all and NP-pairs statistics. For the two-point type 1 error results, we tallied the number of SNPs out of the 124 simulated SNPs with HLOD/LOD scores above certain thresholds. We averaged these tallies across the 1000 replicates and

divided by 124. For two-point power analyses we tallied the number of times the disease SNP crossed the HLOD/LOD threshold in each set of 1000 replicates. For type 1 error and power evaluations of multipoint linkage analysis, we tabulated the maximum parametric HLOD and nonparametric LOD of each replicate and calculated the percentage of the peak HLOD/LOD scores that crossed thresholds. We allowed the maximum HLOD/LOD to be at any of the 124 SNPs since we simulated regions similar to the regions in our previous multipoint study (3) and we do not expect the peak to always be precisely at the disease SNP every time.

We also ran MQLS on the sub-pedigrees to compare those results to running MQLS on the unmanipulated large simulated pedigrees. Prior to running MQLS, we re-calculated kinship coefficients using the sub-pedigree structures rather than the entire pedigree structure to model some of the effect of losing the entire pedigree structure that might occur during linkage analysis. We determined type 1 error rates and power as before.

All computations were performed using either the Center for Human Genetics Research (CHGR) computational cluster or the Advanced Computing Center for Research and Education (ACCRE) cluster at Vanderbilt University.

Results

MQLS

In 1000 runs of MQLS, each with the entire 4998-member pedigree and 124 null SNPs, we see average type 1 error rates of 5.06%, 1.02%, 0.56%, and 0.13% associated

with p-values less than 0.05, 0.01, 0.005, and 0.001, respectively. Therefore, we do not see an inflated type 1 error rate when running MQLS in our pedigree structure.

Evaluating power for 1-locus disease models, we find, as expected, that we have the least power to detect association when the underlying disease model is recessive and the most power to detect association when the underlying disease model is additive. For dominant and additive models we have >90% power to detect association at p ≤0.05 when the simulated odds ratio is at least 2.0, but power drops significantly at an odds ratio of 1.5. With a very strong effect of OR=5, we have very high power to detect association as low a p-value as 5.0E-10. Under the recessive models, power was >80% only when using a p-value threshold of 0.05 with an odds ratio of 5.0 (table 2.1).

Table 2.1. Average percentage of times (power) per model disease SNPs was under p-value thresholds when running MQLS on whole simulated pedigrees. Power ≥80% in bold.

Disease Model, Odds Ratio	%≤0.05	%≤ 5E-3	%≤ 5E-4	%≤ 5E-5	%≤ 5E-6	%≤ 5E-7	%≤ 5E-8	%≤ 5E-9	%≤ 5E-10
recessive, OR 1.1	6	0	0	0	0	0	0	0	0
recessive, OR 1.5	12	4	1	0	0	0	0	0	0
recessive, OR 2.0	26	9	3	1	0	0	0	0	0
recessive, OR 5.0	**87**	75	61	48	38	29	21	15	14
dominant, OR 1.1	8	2	0	0	0	0	0	0	0
dominant, OR 1.5	50	23	9	3	1	1	0	0	0
dominant, OR 2.0	**92**	72	47	28	13	7	4	1	1
dominant, OR 5.0	**100**	**100**	**100**	**100**	**100**	**99**	**98**	**94**	**92**
additive, OR 1.1	11	3	0	0	0	0	0	0	0
additive, OR 1.5	67	36	19	8	3	1	1	0	0
additive, OR 2.0	**96**	**87**	69	50	33	20	12	6	5
additive, OR 5.0	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99**

MQLS-PedCut

Using the same sets of pedigrees, but dividing them into subpedigrees using PedCut, the type 1 error rates when running MQLS hardly changed from the MQLS analysis using whole pedigrees. The type 1 error rates were 5.16%, 1.06%, 0.51%, and 0.11% for the same p-value thresholds.

On the other hand, evaluating power when subdividing the pedigree before running MQLS we do see a loss of power. Power is only >80% for dominant and additive models at an odds ratio of 5.0 (table 2.2).

Table 2.2. Average percentage of times (power) per model disease SNPs was under p-value thresholds when running MQLS on whole simulated pedigrees. All numbers are percentages. Power ≥80% in bold.

Disease Model, Odds Ratio	%≤.05	%≤ 5E-3	%≤ 5E-4	%≤ 5E-5	%≤ 5E-6	%≤ 5E-7	%≤ 5E-8	%≤ 5E-9	%≤ 5E-10
recessive, OR 1.1	6	0.5	0	0	0	0	0	0	0
recessive, OR 1.5	8	1	0.4	0.1	0	0	0	0	0
recessive, OR 2.0	15	3	0.6	0.1	0	0	0	0	0
recessive, OR 5.0	74	51	34	19	10	5	2	1	0.7
dominant, OR 1.1	8	0.3	0	0	0	0	0	0	0
dominant, OR 1.5	24	5	1	0.2	0	0	0	0	0
dominant, OR 2.0	55	21	7	2	0.6	0.1	0	0	0
dominant, OR 5.0	**99**	**90**	72	49	27	13	6	2	0.9
additive, OR 1.1	6	0.6	0	0	0	0	0	0	0
additive, OR 1.5	33	9	2	0.1	0	0	0	0	0
additive, OR 2.0	70	37	16	5	2	0.8	0	0	0
additive, OR 5.0	**100**	**98**	**92**	**80**	65	43	24	12	9

Two-point Linkage

Averaging across 1000 replicates of two-point parametric linkage analysis using sub-pedigrees with a bit-size ≤24, we see low type 1 error rates, which were nearly the same when running dominant and recessive models. The type 1 error rate for an HLOD ≥ 3 under the dominant model was only 0.01% and under the recessive model was only 0.02%. Nonparametric analyses had no type 1 error at LOD threshold of 2 and 3 (table 2.3).

Table 2.3. Percentage of SNPs (type 1 error) above HLOD thresholds using PedCut followed by two-point parametric linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics.

	HLOD/LOD >1	HLOD/LOD >2	HLOD/LOD >3
dominant	2.21%	0.18%	0.01%
recessive	2.02%	0.20%	0.02%
NPL all	0.15%	0	0
NPL pairs	0.05%	0	0

According to our simulations of 1-locus disease models, we had >80% power to detect a two-point HLOD ≥ 1.0 with a simulated additive model with OR=5.0 when a dominant model is assumed during linkage analysis. All other circumstances had <80% power; however, with the simulated dominant model with OR=5, Merlin was able to detect the disease SNP almost 80% of the time at or above an HLOD of 1 when a dominant model was assumed. Even when a recessive model was assumed two-point

linkage analysis was not powerful for any of the simulated recessive scenarios. Parametric analyses were more powerful than nonparametric analyses (table 2.4).

Table 2.4. Percentage of times (power) disease SNP crossed parametric HLOD or nonparametric LOD thresholds using PedCut followed by Merlin two-point parametric and nonparametric linkage analyses. 1000 replicates of each disease model were performed. All numbers are percentages. Power >80% in bold.

	HLOD/LOD ≥1.0				HLOD/LOD ≥2.0				HLOD/LOD ≥3.0			
Model, Odds Ratio	Dom	Rec	All	Pairs	Dom	Rec	All	Pairs	Dom	Rec	All	Pairs
dominant, OR 1.1	2.4	2.3	0	0	0.1	0	0	0	0	0	0	0
dominant, OR 1.5	3.6	3.6	0.7	0.3	0.6	0.7	0	0	0	0	0	0
dominant, OR 2.0	8.3	9.1	2.3	0.7	1.7	1.2	0	0	0.5	0.4	0	0
dominant, OR 5.0	77.7	71	50	33.6	51.1	43.3	4.7	0.7	28.2	22.8	0	0
recessive, OR 1.1	2.6	2.6	0.4	0.1	0.4	0.4	0	0	0	0.1	0	0
recessive, OR 1.5	2.9	2.3	0.3	0.1	0.2	0.1	0	0	0	0	0	0
recessive, OR 2.0	2.4	2.3	0.2	0.1	0.2	0.2	0	0	0	0	0	0
recessive, OR 5.0	13.3	13.9	9	6.5	3.7	4.1	0.4	0.1	0.5	1.4	0	0
additive, OR 1.1	2.5	2.2	0.3	0.1	0.1	0.2	0	0	0	0	0	0
additive, OR 1.5	4.3	3.7	1	0.6	0.4	0.8	0	0	0.2	0.1	0	0
additive, OR 2.0	12.3	10.4	3	1.5	2.6	2.3	0.1	0	0.6	0.3	0	0
additive, OR 5.0	**85.5**	79.1	64.9	48.9	67.8	53.6	12.2	3.4	44	32	0.7	0

Multipoint Linkage

When running multipoint analysis on the same sets of sub-pedigrees we see both higher type 1 error and higher power for most circumstances except for a simulated dominant model with OR=5. For multipoint analyses we see higher type 1 error and power for nonparametric analyses than for parametric analyses (tables 2.5 and 2.6). For

both two-point and multipoint linkage, the highest power for detecting linkage was seen with a simulated additive model with OR=5.0 (tables 2.4 and 2.6).

Table 2.5. Percentage of SNPs (type 1 error) above parametric HLOD and nonparametric LOD thresholds using PedCut followed by multipoint parametric linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics.

	HLOD/LOD ≥1	HLOD/LOD ≥2	HLOD/LOD ≥3
dominant	23.9%	7.5%	2.5%
recessive	19.7%	6.8%	2.5%
NPL all	44.2%	16.5%	4.6%
NPL pairs	44.7%	16%	3.7%

Table 6: Power to detect parametric HLOD and nonparametric LOD thresholds using PedCut followed by multipoint parametric linkage analyses assuming dominant and recessive models and nonparametric linkage analysis using the 'all' and 'pairs' statistics. All numbers are percentages.

	HLOD/LOD ≥1.0				HLOD/LOD ≥2.0				HLOD/LOD ≥3.0			
Model, Odds Ratio	Dom	Rec	All	Pairs	Dom	Rec	All	Pairs	Dom	Rec	All	Pairs
dominant, OR 1.1	22.4	18	44.1	43.2	6.9	5.4	13.7	14	2.1	1.8	3.6	2.7
dominant, OR 1.5	23.3	21.7	44.9	44.1	7.8	6.8	15.2	15	2.4	1.6	3.5	2.6
dominant, OR 2.0	26.7	22.1	48.1	47.7	8.8	6.6	17.7	16.6	1.9	1	5.7	4.7
dominant, OR 5.0	43.8	33	72.9	72.5	20.8	13.5	41.6	41.3	7.8	5.2	19.5	16.6
recessive, OR 1.1	22.8	19.7	41.6	41.6	7.6	5.3	16	15	2.2	1.7	4	2.8
recessive, OR 1.5	24.2	20.7	43.9	44.2	6.5	5.8	16.8	16.6	1.4	1.4	4.8	4.1
recessive, OR 2.0	23.2	19.7	43.9	44.6	7.5	6.1	15.1	14.7	1.9	1.8	3.5	3.2
recessive, OR 5.0	31	26.2	54.3	56.5	10.3	8.2	23.6	23.1	3.4	3.2	7.7	6.3
additive, OR 1.1	23.5	19.2	44.1	44.2	6.9	5.7	15.4	14.7	2.9	2.6	4.4	3.6
additive, OR 1.5	26	21.5	45.5	46.2	8.6	5.8	18	17.1	1.9	1.4	5.4	4.2
additive, OR 2.0	30.7	26.5	51.4	52.7	10.6	7.3	20.8	20.2	2.5	1.5	6.4	5.7
additive, OR 5.0	50.5	39.6	77.9	77.5	26.9	18.8	52	49.9	12	8	25.9	21.7

Discussion

Pedigrees from population isolates provide rich datasets for genetic analyses; however, the size and complexity of the pedigrees contribute to ambiguity when running analyses and interpreting results. We have used this approach to discover novel susceptibility loci for complex diseases, such as Alzheimer disease and Parkinson's disease, by studying the Amish communities of Ohio and Indiana. In a recent genome-wide study using this population (19), 798 successfully genotyped

individuals connected into one 13-generation, 4998-member pedigree with consanguineous loops. Using this same pedigree structure, we simulated 1000 pedigree replicates.

Simulations of pedigrees as large and as complex as an Amish pedigree to assess the type 1 error rate and power of MQLS have not been previously published, so we sought to fill this void. We did not see an inflated type 1 error rate in our simulated pedigrees. Therefore, MQLS is an appropriate method for analyzing pedigrees as large and as complex as the Amish. MQLS is very powerful for detecting a strong effect of OR=5 when the mode of inheritance is recessive, dominant, and additive and OR=2 when the mode of inheritance model is dominant or additive. While these are large effect sizes compared to those typical of complex diseases, in a homogeneous founder population a larger effect size is more likely.

Linkage analyses for a pedigree of this size and complexity require pedigree splitting, but the effect on the type 1 error and power was not known for our pedigree structure using PedCut to subdivide the pedigrees followed by linkage analysis using Merlin. Using a bit-size limit of 24, we saw a low type 1 error rate associated with an HLOD of 3.0 for both two-point and multipoint linkage (lower for two-point). An HLOD of ~3 has traditionally been a 'significant' HLOD score, and the low type 1 error rate in this instance all allows us to confidently use this threshold when evaluating results from the Amish sub-pedigrees. These approaches, however, were not powerful when we analyzed simulated 1-locus disease models.

Unfortunately, we cannot analyze the entire 4998 member pedigree for linkage to compare the type 1 error and power to analyses of sub-pedigrees for linkage. We can, however, compare the type 1 error of association analysis using MQLS on the entire

pedigree versus using MQLS on the sub-pedigrees. Splitting the simulated pedigrees did not affect the type 1 error when running MQLS. This result does not guarantee that splitting a pedigree will not lead to any spurious positive results, since other studies suggest otherwise (14). We do see a loss of power due to splitting the pedigrees because many pedigree connections are disrupted.

Through these simulations we see that MQLS has acceptable type 1 error rates even when using an extremely complex pedigree structure. Type 1 error rates are also acceptable when splitting pedigrees prior to linkage analysis, consistent with a related study (13). Unfortunately, but not surprisingly, significant power is lost when pedigrees are divided. Development of new methods or extensions of current methods to use more pedigree information to perform multipoint linkage analyses would greatly improve our ability to query the rich genetic information of founder populations.

Acknowledgements

We would like to thank the Anabaptist Genealogy Database for providing the template pedigree. We thank the family participants and community members for graciously agreeing to participate, making research in these communities possible. This study is supported by the National Institutes of Health grants AG019085 (to JLH and MAP-V) and AG019726 (to WKS). Some of the samples used in this study were collected while WKS, JRG, and MAP-V were faculty members at Duke University.

REFERENCES

(1) Edwards AO, Ritter R, III, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science 2005 Apr 15;308(5720):421-4.

(2) Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science 2005 Apr 15;308(5720):419-21.

(3) Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, et al. Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. Am J Hum Genet 1991;48:1034-50.

(4) Lander ES, Botstein D. Homozygostiy mapping: a way to map human recessive traits with the DNA of inbred children. Science 1987;236:1567-70.

(5) Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, et al. Mutation in the à-synuclein gene identifed in families with Parkinson's disease. Science 1997;276(5321):2045-7.

(6) Sheffield V, Carmi R, Kwitek-Black A, Rokhlina T, Nishimura D, Duyk GM, et al. Identification of a Bardet-Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. Human Molecular Genetics 1994;3(8):1331-5.

(7) Vance JM, Jonasson F, Lennon F, Sarrica J, Damji KF, Stauffer J, et al. Linkage of a gene for macular corneal dystrophy to chromosome 16. Am J Hum Genet 1996 Apr;58(4):757-62.

(8) Hastbacka J, de la Chappelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations:diastrophic dysplasia in Finland. Nature Genetics 1992;2(november):204-11.

(9) Amish Heritage Committee. Amish and Mennonites in Eastern Elkhart & LaGrange Counties, Indiana 1841-1991. 2nd printing ed. Goshen, Indiana: Amish Heritage Committee; 2009.

(10) Beachy L. Unser Leit: The Story of the Amish. Millersburg, OH: Goodly Heritage Books; 2011.

(11) Hostetler J. Amish Society, 4th ed. Baltimore, MD: Johns hopkins University Press; 1993.

(12) Agarwala R, Biesecker LG, Schaffer AA. Anabaptist genealogy database. Am J Med Genet C Semin Med Genet 2003 Aug 15;121(1):32-7.

(13) Agarwala R, Biesecker LG, Tomlin JF, Schaffer AA. Towards a complete North American Anabaptist genealogy: A systematic approach to merging partially overlapping genealogy resources. Am J Med Genet 1999 Sep 10;86(2):156-61.

(14) Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.

(15) Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 2000 Jul;67(1):146-54.

(16) Martin ER, Bass MP, Kaplan NL. Correcting for a potential bias in the pedigree disequilibrium test. Am J Hum Genet 2001 Apr;68(4):1065-7.

(17) Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 1998 Dec;8(12):1273-88.

(18) Cummings AC, Lee SL, McCauley JL, Jiang L, Crunk A, McFarland LL, et al. A genome-wide linkage screen in the amish with Parkinson disease points to chromosome 6. Ann Hum Genet 2011 May;75(3):351-8.

(19) Cummings AC, Jiang L, Velez Edwards DR, McCauley JL, Laux R, McFarland LL, et al. Genome-wide association and linkage study in the amish detects a novel candidate late-onset Alzheimer disease gene. Ann Hum Genet 2012 Sep;76(5):342-51.

(20) Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 1987 Apr;84(8):2363-7.

(21) Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Human Heredity 1971;21:523-42.

(22) O'Connell JR, Weeks DE. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recording and fuzzy inheritance. Nature Genetics 1995;11:402-8.

(23) Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. Allegro, a new computer program for multipoint linkage analysis. Nat Genet 2000 May;25(1):12-3.

(24) Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A. Allegro version 2. Nat Genet 2005 Oct;37(10):1015-6.

(25) Fishelson M, Geiger D. Exact genetic linkage computations for general pedigrees. Bioinformatics 2002;18 Suppl 1:S189-S198.

(26) Fishelson M, Geiger D. Optimizing exact genetic linkage computations. J Comput Biol 2004;11(2-3):263-75.

(27) Fishelson M, Dovgolevsky N, Geiger D. Maximum likelihood haplotyping for general pedigrees. Hum Hered 2005;59(1):41-60.

(28) Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002 Jan;30(1):97-101.

(29) Williams AL, Housman DE, Rinard MC, Gifford DK. Rapid haplotype inference for nuclear families. Genome Biol 2010;11(10):R108.

(30) Falchi M, Forabosco P, Mocci E, Borlino CC, Picciau A, Virdis E, et al. A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. Am J Hum Genet 2004 Dec;75(6):1015-31.

(31) Liu F, Kirichenko A, Axenovich TI, van Duijn CM, Aulchenko YS. An approach for cutting large and complex pedigrees for linkage analysis. Eur J Hum Genet 2008 Jul;16(7):854-60.

(32) Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 1996 Jun;58(6):1347-63.

(33) Liu F, Elefante S, van Duijn CM, Aulchenko YS. Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Ann Hum Genet 2006 Nov;70(Pt 6):965-70.

(34) Miano MG, Jacobson SG, Carothers A, Hanson I, Teague P, Lovell J, et al. Pitfalls in homozygosity mapping. Am J Hum Genet 2000 Nov;67(5):1348-51.

(35) Dyer TD, Blangero J, Williams JT, Goring HH, Mahaney MC. The effect of pedigree complexity on quantitative trait linkage analysis. Genet Epidemiol 2001;21 Suppl 1:S236-S243.

(36) Liu F, Arias-Vasquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P, et al. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. Am J Hum Genet 2007 Jul;81(1):17-31.

(37) Wang Y, O'Connell JR, McArdle PF, Wade JB, Dorff SE, Shah SJ, et al. From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. Proc Natl Acad Sci U S A 2009 Jan 6;106(1):226-31.

(38) Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 2007 Aug;81(2):321-37.

(39) Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, et al. Generating Linkage Disequilibrium Patterns in Data Simulations using genomeSIMLA. Lect Notes Comput Sci 2008 Jan 1;4973(2008):24-35.

(40) Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet 2010 Feb 12;86(2):172-84.