ASSORTATIVE MATING AS A STRATIFICATION PROBLEM IN GENETIC ASSOCIATION STUDIES

by

Solomon Tetteh Quaynor

B.S., Kent State University, 1996

M.A., Oklahoma State University, 2003

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PTITSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Solomon Tetteh Quaynor

It was defended on

March 30th, 2007

and approved by

Committee Chair:
Eleanor Feingold, PhD
Associate Professor
Departments of Human Genetics and Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Candace M. Kammerer, PhD
Associate Professor
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Ada O. Youk, PhD
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Eleanor Feingold, Ph.D

ASSORTATIVE MATING AS A STRATIFICATION PROBLEM IN GENETIC ASSOCIATION STUDIES

Solomon Tetteh Quaynor, M.S.

University of Pittsburgh, 2007

Genetic association studies have an important role in public health because they help us understand the biological basis of conditions (e.g. diabetes, obesity) that have important public health implications. They can help us develop and direct both treatments and prevention activities. As both Type II diabetes and obesity tend to run in families, it is reasonable to want to ascertain whether a genetic association or linkage exists between a particular allele or alleles and these conditions. Genetic association studies are, generally, the preferred method for detecting genes that are causal variants of complex diseases like diabetes because they have greater power to detect alleles that are susceptible to disease. However, the Case control genetic association studies are known to be prone to false positive associations in the presence of population stratification. We hypothesize that assortative mating in a given population can lead to a form of population stratification and subsequently false positives. We also investigate the role of gene-gene interactions in the presence of assortative mating in producing spurious results. These hypotheses are tested via studies on 10,000 simulated individuals. Our results show that assortative mating does lead to a greater than expected number of false positives as compared to a situation where there is no assortative mating. Our tests on the role of gene-gene interactions also suggest that they contribute to false positives in the presence of assortative mating.

TABLE OF CONTENTS

LIST OF TABLES

1.0 INTRODUCTION

Genetic association studies have an important role in public health because they help us understand the biological basis of conditions (e.g. diabetes, obesity) that have important public health implications. They can help us develop and direct both treatments and prevention activities. As both Type II diabetes and obesity tend to run in families, it is reasonable to want to ascertain whether a genetic association or linkage exists between a particular allele or alleles and these conditions.

Whenever one wishes to evaluate the genetic basis for a disease, it is important to know how the disease is transmitted in families and whether there is a single gene or multiple genes responsible for the disease. The classical form of disease or trait transmission in families is called Mendelian inheritance. A brief summary of Mendelian laws is as follows: traits controlled by a single gene are uniformly distributed if transmitted from the heterozygous parental generation to the offspring in a specific ratio: 1:2:1. This means that 25% of the offspring will have only the dominant or major form of the trait. The dominant trait is the trait that is more likely to be expressed from generation to generation and the minor or recessive form of the trait is likely to skip generations. The recessive trait shows up 25% of the time in the offspring while 50% of offspring have both the dominant and recessive forms of the trait. Additionally, according to Mendelian laws, unlinked traits are transmitted independent of each other. However, for complex diseases these rules do not necessarily apply. Complex diseases are typically controlled by multiple genes and the probability of transmitting the disease is often conditional on several

factors which make identifying genes responsible for such complex diseases like diabetes not as straightforward.

The task of identifying susceptible genes, responsible for complex diseases, involves the utilization of a number of statistical approaches. Among them are linkage and association studies. Genetic linkage studies attempt to determine whether a gene and a disease are co-inherited within families. The term linkage here refers to the concept that genes and other genetic markers that are close together tend to be inherited together. It has been argued, however, that linkage studies have low power to detect common alleles that confer disease susceptibility (Risch and Merikangas 1996). As a result, association approaches are currently more popularly used to detect genes which are causal variants of complex diseases.

Association studies seek to detect whether nonrandom associations exist between trait values and particular alleles in a population. Association studies can be based on any standard epidemiological study type, such as a case-control or population study, and can also be performed using family data. Association studies, however, have limitations of their own. These studies are known to be prone to spurious associations or false positives. Spurious associations erroneously suggest that certain alleles or genes are associated with some trait(s). Hence the need arises for the determination of the cause of these false positives and the subsequent development of methods to account for these spurious results in association studies.

One of the most important causes of false positives in association studies is population admixture and/or stratification. Some association tests that have been developed to address the problem of confounding in genetic association studies include the Transmission Disequilibrium Test (TDT) and the affected family based control method. (AFBAC). The TDT tests whether the ratio of alleles passed on from heterozygous parents to an affected child would differ from

expected Mendelian levels (Ziegler and Konig, 2006). This test does not consider homozygous parents because it is difficult to track which allele came from which parent in the homozygous case. The AFBAC looks at the ratio of the frequency of alleles transmitted from parent to offspring and compares it to the untransmitted alleles. These methods are meant to minimize population stratification. However, the degree to which this is reduced depends on the mating pattern and genetic model (Risch and Teng, 1998).In addition, the TDT and other family based association tests require the collecting of parental information. This process is difficult and expensive and probably impossible for a late-onset disease like diabetes.

Another method that attempts to address the stratification problem is genomic control. Genomic control methods correct the false positive rate in a case-control study by creating a test statistic that takes into account both loci that are associated with the disease and those that are not. This method incorporates a variance inflation factor that mirrors deviation from the null hypothesis of no stratification. The drawback of this method is that it assumes that the inflation factor is constant. (Ziegler and Koenig 2006;Devlin and Roeder 1999).

This study focuses on the role of assortative mating in creating population stratification and thus false positives. Almost all previous literature on the topic has assumed that stratification means ethnic stratification, but in fact assortative mating also creates population substructure that can have similar effects. This effect of assortative mating was considered by Redden and Allison (2006) at the same time that we were conducting our study, so we briefly discuss their study below and then contrast theirs with ours in the discussion.

Redden and Allison (2006) looked at the effect of assortative mating in genetic association studies in the absence of ethnic stratification. They examined the effect of non-random mating on three traits, adiposity (A), beauty (B) and intelligence (I), via simulation

studies with a large sample size of 1,000,000. Each trait was influenced by 10 separate loci. Their genotypes were randomly assigned from a multinomial distribution. The probability of each genotype was assigned based on Mendelian laws with probability of MM =0.25, Mm = 0.50, and mm = 0.25. They selected mates based on the following model:

Desirability $D = B + I – A + \varepsilon$, where ε is an error term. The rate of assortative mating was determined by the degree of desirability. Their assortative mating rates ranged from 10%-50%. They did not report results for an assortative mating level of zero. The simulations were carried out over 10 generations with the 10^{th} generation being assessed for false positives. They concluded that even in ethnically homogenous populations, spurious associations occur. Like the present study they looked at complex traits, but they did not account for the fact that, in complex traits, the interaction between the genes might contribute to the development of spurious associations between a trait and a disease.

.

1.1 POPULATION STRATIFICATION

Population stratification refers to the situation in which the population under study is actually a composite of two or more distinct subpopulations, usually thought of as different ethnic groups. Hence the differences in allele frequencies between cases and controls, instead of being an indicator of an association between allele and trait, are more than likely a reflection of different ethnic or racial origins (Redden and Allison, 2006). The fictitious example tabulated below is a simple numerical illustration of how population stratification can cause Type I errors

in results. Suppose one wants to investigate the genetic basis for procrastination. Data of cases that exhibit the allele that has the mutation for the disease and controls that do not are collected from two separate populations. The odds ratio is obtained as a measure of association between cases and controls.

Table 1. Numerical Example of Population Stratification

	Population 1		Population 2	
	mutation	no mutation	mutation	no mutation
Cases	40	10	10	20
Controls	120	30	900	1800

The odds ratio (OR) in each population is as follows:

$$OR_1 = (40x\ 30)/120x10)$$

$$= 1.0$$

$$OR_2 = (10x1800)/(20x900)$$

$$= 1.0$$

When the results are pooled,

$$OR_{12} = (40+10)\ x\ (30+1800)/\ (10+20)\ x\ (900+120)$$

$$= 2.99$$

The combined result suggests that there is in fact an association between the mutated allele and procrastination, although if fact there is not. The difference in results can be attributed to the frequency of the mutations responsible for the disease in each population. In population 1 the frequency,

F_1 = 40 + 10/40+10 +120 + 30 = 0.25. In population 2 F_2 = 10 + 20/ 10+20+900+1800 = 0.01. The discrepancy in the frequency of the procrastination in the two populations explains the difference in odds ratios.

1.2 ADMIXTURE

Admixture is similar to stratification, but more complex. It refers to the situation in which two or more ethnic groups have intermarried for a few generations. An example of a population that is a product of admixture is the contemporary Mexican population which resulted from the mating of Native Americans with Europeans (Bonilla et. al, 2005). Admixture can cause spurious associations for the same reasons as stratification.

1.3 ASSORTATIVE MATING

Assortative mating is the term used to describe the choice of mating preference based on phenotypic characteristics. There are two kinds of assortative mating; positive assortative mating (PAM) and negative assortative mating (NAM). An example of PAM would be tall people marrying tall people. This form of mating is nonrandom and does not change overall allele frequencies in a population. However, assortative mating creates semi-separate subpopulations (e.g. tall and short people), which can affect association studies in the same way as ethnic stratification. The goals of this study are to demonstrate that stratification caused by assortative mating contributes to, on average, a greater than expected frequency of Type I errors.

2.0 METHODS

We demonstrated the effect of assortative mating on association studies via a simulation study. The simulation was performed using code written in the R language. Our code is given in Appendix A. Genotypes at 3 independent biallelic loci, 3 loci (A, B, and C) with 2 (Aa, Bb, Cc) alleles each, were created for 10,000 simulated individuals. Each locus was assigned the minor alleles, A, B, and C with the probability 0.2 making the major alleles, a, b, and c, have the frequency 0.8. The frequencies of the genotypes, which were randomly drawn from a multinomial distribution, were calculated assuming Hardy-Weinberg Equilibrium (HWE): P (AA) = p^2 P (Aa) = 2pq and P (aa) = q^2. For each individual we generated two binary traits according to the penetrances given in tables 2 and 3. Trait 1 was defined as being influenced by loci A and B while Trait 2 was influenced by loci B and C.

We did not distinguish male and female individuals. To create the next generation, we considered our entire population of 10,000 individuals and chose each person a mate from the entire population with replacement. The mates were chosen to have positive assortative mating for trait 1. An individual with trait 1 had probability PT1 = (1/10,000 + a/R1) of choosing a mate with trait 1, where R1 is the number of people with trait 1 and ‘a’ is an arbitrary constant which we varied between 0.0 (no assortative mating) and 0.2. An individual with trait 1 had probability NPT1 = (1/10,000 - a/(10,000 -R1) of choosing a mate without trait 1.

Each couple was then given one offspring with genotypes determined according to Mendelian rules and phenotypes according to the penetrances given in tables 2 and 3. There were 10 generations simulated with a total of 10 replications per generation.

We were trying to demonstrate that after a few generations of such mating, we would be able to detect associations between both traits and all three loci. That is, we wanted to demonstrate that the assortative mating on trait 1 created non-random association among the three loci and thus false positive associations for both traits. A false positive or Type I error in our study is defined as getting a significant association between trait 1 and locus C and/or trait 2 and locus A.

Table 2: Penetrances for Trait 1

	Locus A		
Locus B	AA	Aa	aa
BB	0.8	0.8	0.3
Bb	0.8	0.8	0.3
bb	0.4	0.4	0.1

Table 3: Penetrances for Trait 2

	Locus B		
Locus C	BB	Bb	bb
CC	0.8	0.8	0.4
Cc	0.8	0.8	0.4
cc	0.3	0.3	0.1

We tested genetic associations between each trait and each locus using logistic regression. A linear trend test was performed to regress each trait on each locus. We used a very large sample size (the entire population of 10,000) because we were trying to demonstrate the existence of false positives, not to measure their frequency.

3.0 RESULTS

3.1 FREQUENCY OF PHENOTYPES

The frequencies of Trait 1 and Trait 2 in the parental generation, as well as the 9 generations that followed appear fairly constant in each generation. From the results, which are summarized in Tables 4 -7 below, it appears each trait has a 25-30% frequency in each generation. The expected frequencies for Trait 1 and Trait 2 were approximately 30%. As the degree of assortative mating was increased from 0 to 0.2, the frequency of each trait did not exhibit any appreciable differences across generations.

Table 4. Frequency of Trait 1 in every generation (a = 0.1)

	Trial									
Generation	1	2	3	4	5	6	7	8	9	10
1	0.30	0.30	0.29	0.28	0.29	0.27	0.27	0.27	0.27	0.26
2	0.30	0.29	0.29	0.28	0.28	0.28	0.26	0.28	0.26	0.26
3	0.30	0.30	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.28
4	0.31	0.30	0.29	0.29	0.28	0.27	0.27	0.27	0.26	0.26
5	0.31	0.30	0.30	0.30	0.29	0.29	0.28	0.27	0.27	0.27
6	0.29	0.29	0.28	0.28	0.27	0.26	0.26	0.25	0.25	0.25
7	0.29	0.29	0.28	0.29	0.29	0.28	0.28	0.28	0.27	0.26
8	0.29	0.29	0.29	0.29	0.28	0.28	0.27	0.27	0.27	0.25
9	0.30	0.29	0.29	0.29	0.28	0.28	0.28	0.27	0.27	0.27
10	0.29	0.29	0.27	0.26	0.25	0.25	0.23	0.22	0.21	0.21

Table 5. **Frequency of Trait 2 in every generation (a = 0.1)**

Trial

Generation	1	2	3	4	5	6	7	8	9	10
1	0.31	0.30	0.30	0.28	0.29	0.29	0.30	0.28	0.28	0.29
2	0.30	0.30	0.30	0.30	0.30	0.30	0.31	0.29	0.29	0.29
3	0.30	0.31	0.30	0.30	0.30	0.30	0.29	0.29	0.31	0.29
4	0.32	0.29	0.30	0.29	0.29	0.29	0.29	0.30	0.29	0.30
5	0.29	0.29	0.29	0.30	0.30	0.29	0.29	0.29	0.28	0.28
6	0.30	0.29	0.28	0.29	0.29	0.29	0.28	0.27	0.28	0.27
7	0.30	0.30	0.30	0.31	0.30	0.30	0.30	0.30	0.29	0.28
8	0.30	0.31	0.29	0.30	0.29	0.30	0.29	0.30	0.29	0.28
9	0.30	0.30	0.31	0.30	0.29	0.30	0.29	0.29	0.29	0.27
10	0.30	0.30	0.30	0.29	0.30	0.29	0.29	0.28	0.27	0.28

Table 6. **Frequency of Trait 1 in every generation (a = 0.2)**

Trial

Generation	1	2	3	4	5	6	7	8	9	10
1	0.29	0.28	0.29	0.26	0.25	0.24	0.24	0.24	0.23	0.21
2	0.29	0.29	0.27	0.28	0.26	0.25	0.25	0.23	0.23	0.22
3	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.24	0.23	0.21
4	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.23	0.23	0.21
5	0.29	0.28	0.28	0.27	0.25	0.25	0.25	0.24	0.23	0.22
6	0.29	0.28	0.27	0.27	0.26	0.25	0.24	0.23	0.23	0.21
7	0.28	0.28	0.27	0.26	0.25	0.23	0.23	0.21	0.20	0.20
8	0.30	0.29	0.28	0.28	0.27	0.25	0.24	0.24	0.23	0.21
9	0.29	0.29	0.28	0.27	0.26	0.25	0.24	0.24	0.23	0.22
10	0.29	0.29	0.27	0.27	0.26	0.25	0.24	0.23	0.22	0.22

Table 7. **Frequency of Trait 2 in each generation (a = 0.2)**

Trial

Generation	1	2	3	4	5	6	7	8	9	10
1	0.30	0.30	0.29	0.29	0.28	0.27	0.28	0.26	0.26	0.26
2	0.30	0.30	0.29	0.30	0.30	0.29	0.29	0.27	0.28	0.27
3	0.30	0.30	0.28	0.28	0.29	0.28	0.27	0.27	0.26	0.25
4	0.31	0.30	0.30	0.31	0.29	0.28	0.29	0.28	0.28	0.27
5	0.29	0.30	0.30	0.29	0.30	0.29	0.29	0.29	0.28	0.28
6	0.30	0.30	0.29	0.30	0.28	0.28	0.28	0.27	0.26	0.26
7	0.30	0.29	0.29	0.29	0.29	0.29	0.28	0.28	0.27	0.27
8	0.31	0.29	0.29	0.30	0.29	0.29	0.29	0.30	0.30	0.28
9	0.29	0.29	0.29	0.29	0.29	0.28	0.27	0.27	0.28	0.26
10	0.30	0.30	0.30	0.29	0.30	0.29	0.29	0.28	0.28	0.28

3.2 FREQUENCY OF TYPE 1 ERRORS

The number of false positives for each level of assortative mating, was observed in the main effects model which regressed a child's trait on a particular genotype in the main effects model. The results, in Table 8, indicate that, as the level of assortative mating was increased, there was a corresponding rise in the number of false positives after 10 generations. When the level of assortative mating was increased twofold, from 0.1 to 0.2, there was a fourfold jump in the total number of false positives. Also there were no false positives recorded when the correction factor for assortative mating was set to zero.

Investigating the possible interaction between genes was done by including an interaction term in each regression model. Since the value of a=0.2 produced the greatest number of false positives in the main effects models, the test for gene-gene interaction was done using this value. The interaction terms were the product of the two genes that are not expected to influence a particular trait. For example, the interaction between alleles A and C should have no significant effect on trait 1. When the interaction terms were placed in the regression models, there were a total of 5 false positives found. For example, there was a significant interaction found between a child expressing the trait 2 phenotype, which is influenced by alleles B and C but not allele A, and alleles A and C in generation 1. This suggests that this was a false interaction between alleles A and C in the first generation. However, in all but one case, there were no false positives in the main effects when the interaction terms were included. Results for the tests for gene-gene interactions are given in Table 12.

Table 8. False Positives for each level of Assortative Mating

Amount of Assortative Mating	Total Number of False Positives at the End of 10 Generations
a = 0	0/20
a = 0.1	1/20
a = 0.2	4/20

3.3 TEST FOR LINKAGE DISEQUILIBRIUM

As an indication of the amount of association among the genes caused by the presence of assortative mating, we also measured the correlations between genotypes in each generation. These results are given in tables 9-11. The correlations increase as the level of assortative mating is increased. Correlations are low but they are enough to create false positives because of our large sample size.

Table 9. **Correlation between Locus A and B (a = 0.2)**

Trial

Generation	1	2	3	4	5	6	7	8	9	10
1	0.01	0.04	0.02	0.03	0.02	0.04	0.03	0.03	0.02	0.00
2	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.01	0.00	0.02
3	0.01	0.01	0.02	0.03	0.03	0.04	0.03	0.04	0.02	0.02
4	0.00	0.02	0.01	0.01	-0.00	0.01	0.01	0.03	0.02	0.02
5	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.01
6	0.02	0.04	0.03	0.03	0.03	0.03	0.01	0.03	0.02	0.04
7	0.02	0.03	0.04	0.02	0.02	0.02	0.02	0.02	0.03	0.03
8	0.02	0.01	0.03	0.04	0.05	0.04	0.02	0.02	0.02	0.02
9	0.01	0.02	0.03	0.02	0.02	0.02	0.03	0.03	0.02	0.02
10	0.00	0.02	0.03	0.04	0.02	0.04	0.02	0.04	0.03	0.03

Table 10. **Correlation between Locus B and C (a = 0.2)**

Trial

Generation	1	2	3	4	5	6	7	8	9	10
1	-0.01	-0.00	-0.01	-0.01	-0.00	-0.01	-0.01	-0.00	-0.00	0.00
2	-0.00	0.00	0.00	-0.01	-0.00	-0.01	-0.03	-0.01	-0.00	0.00
3	0.00	-0.01	-0.01	-0.01	0.00	-0.00	-0.01	0.00	-0.01	-0.00
4	-0.01	-0.01	-0.00	0.00	-0.00	-0.01	-0.01	-0.00	-0.00	-0.00
5	0.00	0.00	-0.01	-0.01	0.02	0.02	0.01	0.02	0.02	0.00
6	0.00	-0.00	0.01	0.00	0.01	0.00	0.00	-0.00	0.02	0.00
7	-0.00	-0.01	-0.01	-0.02	0.00	0.00	-0.01	-0.00	0.00	0.01
8	0.01	0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.01
9	0.00	-0.00	0.00	-0.01	-0.01	-0.01	-0.00	-0.02	0.00	-0.01
10	-0.01	0.00	0.01	0.00	0.01	0.01	0.01	-0.00	0.01	0.01

Table 11. **Correlation between Locus A and C.**

Generation	1	2	3	4	5	6	7	8	9	10
1	-0.00	0.00	-0.00	-0.01	-0.00	0.01	0.01	0.00	0.00	0.02
2	-0.01	-0.01	-0.00	0.01	0.00	-0.00	-0.01	-0.01	-0.01	-0.01
3	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.00
4	0.00	-0.02	-0.02	-0.01	-0.01	-0.00	0.00	0.01	0.01	0.01
5	0.01	0.01	0.01	-0.01	0.00	0.00	0.00	-0.01	-0.00	-0.01
6	0.00	-0.01	-0.00	-0.01	0.00	-0.00	0.00	0.00	0.00	0.02
7	-0.00	-0.00	0.01	0.00	0.00	0.01	-0.01	-0.01	0.01	0.00
8	-0.00	-0.00	0.01	0.00	0.00	-0.00	-0.02	-0.00	-0.01	-0.01
9	-0.01	-0.00	0.00	-0.00	-0.03	-0.03	-0.01	-0.00	-0.01	0.00
10	-0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.00	-0.00	0.00	-0.00

Table 12. Gene-Gene Interaction

Replication	Interaction	Trait
1	+	2
2	-	
3	+	1
4	+	2
5	+	1
6	+	2
7	-	
8	-	
9	-	
10	-	

4.0 DISCUSSION

This study has evaluated the effect of assortative mating on confounding due to population stratification in genetic association studies. In theory, one would expect a population that exclusively chose mates based on their mate having that certain trait, would eventually have that is equally stratified based on the number of traits. The degree of stratification was not equal in this simulated population because the probability of choosing a mate with either Trait 1 or Trait 2 was conditional on the probability of expressing the trait given a particular genotype. (P(Trait|G)).Thus the frequency of each trait was approximately between 25%-30%. When each trait was regressed onto the genotype at each locus, in a main effects generalized linear model excluding interactions, there was a total number of false positives of 4 out of 20 over the 10 generations with the level of assortative mating set at 0.2. At the 0.1 level of assortative mating, there was a 5% (1 /20) Type I error rate which, according to the literature (Redden and Allison, 2006) was the expected error rate if there was no assortative mating at all. However, these results contradict the Redden findings because when the rate of assortative mating was set to zero, there were no false positives.

When interaction terms were included in the analysis to access the effects of gene-gene interaction, the false positive rate was approximately 12%. However, the main effects no longer showed any false positives (except in one case). It appears that this is due to colinearity between main effects and interactions. There are a number of possible reasons for this: the genes

responsible for the traits in question only express the trait when working together and one allele may have a dominant effect over the other. Alternatively, or perhaps in addition to this, some genes may have an epistatic effect. An epistatic gene is one that masks the effect of another gene. Mathematically, it is represented by an interaction between different loci. Some authors represent this effect as interaction between a causative allele and a non-causative one. (Ziegler and Konig, 2006). However, as multiple genes are required for the expression of the traits under study, it is unlikely that one gene is non-causative. It is probable that as the penetrances of the alleles play a significant role in the expression of the trait, they determine the level of interaction as well. Prior literature that examines the function of gene-gene interactions in population stratification looked at populations stratified by ethnicity (Wang et. al 2006). Wang et al. found that gene-gene interactions were a significant cause of ethnic stratification. One might infer, given the results in the literature, that assortative mating is the significant contributor to the bias caused by population stratification. This conclusion is reinforced by the fact that in the absence of assortative mating, the rate of false positives found in the interaction terms is approximately 5% which is what one expects when the p-value is set at 0.05.

Our results agree with Redden and Allison's study with respect to the fact that both studies discovered that assortative mating contributes to false positives in association studies. This study also shows an increased correlation between genotypes for each generation. The prior study took a similar approach but in addition looked at correlations between mating pairs for each of their three traits: adiposity, beauty and intelligence. They also reported values for correlations between loci responsible for two of their traits: adiposity and intelligence. Their correlation results obtained here were similar to those in our study with most of their results around 0.02.

The discussion above implies that, without any further information being provided, that the observed trait from one generation to the next in the absence of assortative mating is a more reliable indicator of the association of the trait with a given genotype. However, as the extent of assortative mating increases, so too does the unreliability of the trait as an indicator of a particular genotype. Therefore caution must be taken in extending the traditional case control to all cases of genetic association studies. In other words, going back to the diabetes and obesity example, if one sees a patient who is clinically considered obese, one cannot assume that the patient will necessarily develop diabetes if they mate with another individual who is obese on the basis of the results of a case control study that predicts that they are likely to develop diabetes.

5.0 CONCLUSION

One might be tempted to conclude from the analysis above that only in the presence of assortative mating is one likely to have false positives in stratified populations. However, the results obtained here indicate that even in populations with no assortative mating there is some amount of spurious associations. It appears that the false associations are due to both the main effects as well as the interactions. It is however not clear how the interactions play a role in the number of false positives attained.

One of the limitations of this study is the inability to model the effect of the interaction: it is clear from the results that there is a marked effect of the interactions on spurious associations but if one were to model the type of interaction, i.e. epistatic, codominant, etc., in a future study, it might shed more light on the effect on the frequency of false positives. It might also be important to replicate the study to determine whether spurious interactions show up in almost every generation as they do here. A future study could also look at solutions for the assortative mating issue and assess whether alternatives to the TDT and genomic control tests, which presently are quite expensive and time consuming, could be found.

APPENDIX A

ASSORTATIVE MATING SIMULATION

```
###Allele Frequencies###
p<-0.8 #p=P (a)
q<-0.2 #q=P (A)
n.reps<-10 #Number of replications
n.ind<-10000 ###Number of individuals###
n.gen <- 10 ###Number of generations###
#######Arrays####
num.a<-array()###Number of A alleles in parents
num.b<-array()###Number of B alleles in parents
num.c<-array()###Number of C alleles in parents
child.a<-array() #Child's Genotype at locus A
child.b <-array() #child's genotype at locus B
child.c<-array() #child's genotype at locus C
child.r <-array() # temporary variable
child.1<-array() #Child's Phenotype for trait 1
child.2 <-array() #child's phenotype for trait 2
trait1<-array() #parent phenotype Influenced by allele A and B
trait2<-array() #parent phenotype Influenced by allele B and C
freq1<-array() # frequency of trait 1 in each generation
freq2 <-array() #frequency of trait2 in each generation
corAB <- array() #correlation between A and B genotypes in each generation
corBC <- array() #correlation between B and C genotypes in each generation
corAC <- array() #correlation between A and C genotypes in each generation
s<-1:10000 ###Vector to sample from
#####Loop over replicates#####
for (jj in 1:n.reps) {
#####Simulation of parental genotypes#####
for (j in 1:n.ind){
ind.r<-rmultinom(1,size=1, prob=c(p^2,2*p*q,q^2))
   if (ind.r[1,1]==1){num.a[j]<-0}
```

```
else if (ind.r[2,1]==1){num.a[j]<-1}
else if (ind.r[3,1]==1) {num.a[j]<-2}
ind.r2<-rmultinom(1,size=1, prob=c(p^2,2*p*q,q^2))
   if (ind.r2[1,1]==1){num.b[j]<-0}
else if (ind.r2[2,1]==1){num.b[j]<-1}
else if (ind.r2[3,1]==1) {num.b[j]<-2}
ind.r3<-rmultinom(1,size=1, prob=c(p^2,2*p*q,q^2))
   if (ind.r3[1,1]==1){num.c[j]<-0}
else if (ind.r3[2,1]==1){num.c[j]<-1}
else if (ind.r3[3,1]==1){num.c[j]<-2}
###simulation of parental Traits###
##For Trait 1##
if(num.a[j]==0 & num.b[j]==0){trait1[j]<-rbinom(1,size=1, prob=c(0.1))}
else if(num.a[j]==0 & num.b[j]==1){trait1[j]<-rbinom(1,size=1, prob=c(0.3))}
else if(num.a[j]==0 & num.b[j]==2){trait1[j]<-rbinom(1,size=1, prob=c(0.3))}
else if(num.a[j]==1 & num.b[j]==0){trait1[j]<-rbinom(1,size=1, prob=c(0.4))}
else if(num.a[j]==1 & num.b[j]==1){trait1[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.a[j]==1 & num.b[j]==2){trait1[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.a[j]==2 & num.b[j]==0){trait1[j]<-rbinom(1,size=1, prob=c(0.4))}
else if(num.a[j]==2 & num.b[j]==1){trait1[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.a[j]==2 & num.b[j]==2){trait1[j]<-rbinom(1,size=1, prob=c(0.8))}
##For Trait2##
if(num.b[j]==0 & num.c[j]==0){trait2[j]<-rbinom(1,size=1, prob=c(0.1))}
else if(num.b[j]==0 & num.c[j]==1){trait2[j]<-rbinom(1,size=1, prob=c(0.4))}
else if(num.b[j]==0 & num.c[j]==2){trait2[j]<-rbinom(1,size=1, prob=c(0.4))}
else if(num.b[j]==1 & num.c[j]==0){trait2[j]<-rbinom(1,size=1, prob=c(0.3))}
else if(num.b[j]==1 & num.c[j]==1){trait2[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.b[j]==1 & num.c[j]==2){trait2[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.b[j]==2 & num.c[j]==0){trait2[j]<-rbinom(1,size=1, prob=c(0.3))}
else if(num.b[j]==2 & num.c[j]==1){trait2[j]<-rbinom(1,size=1, prob=c(0.8))}
else if(num.b[j]==2 & num.c[j]==2){trait2[j]<-rbinom(1,size=1, prob=c(0.8))}
}
for (i in 1:n.gen) {
#####Create Vectors of Probabilities for choosing mates
R1<-sum(trait1) #number of people with trait1
a<-0.2 ##Correction factor for formula
PT1<-(1/n.ind + a/R1)# Probability for those with trait 1
NPT1<-(1/n.ind - a/(n.ind -R1))# Probability for those without trait 1
PT1B<-(1/n.ind + a/(n.ind -R1)) # Probability for those without trait 1
NPT1B<-(1/n.ind - a/R1)# Probability for those having trait 1
prob1<-array() #Vector of probabilities for PT1, NPT1
probN1<-array() #Vector of probabilities from PT1B, NPT1B
for (j in 1:n.ind){
        if (trait1[j]==1) {prob1[j]=PT1}
  else if (trait1[j]==0) {prob1[j]=NPT1}
}
```

```
for (j in 1:n.ind){
        if (trait1[j]==0) {probN1[j]=PT1B}
  else if (trait1[j]==1) {probN1[j]=NPT1B}
}

#####Loop through people, choose a mate for each, and give each couple a child

s <- c(1:n.ind)
for (j in 1:n.ind) {

        if (trait1[j]==1) {mate<- sample(s, 1, replace = FALSE, prob = prob1)}
        else if (trait1[j]==0) {mate<-sample(s, 1, replace = FALSE, prob = probN1)}


##### child's genotype at locus A by Mendelian Rules  #####
     if (num.a[j]==0 & num.a[mate]==0) {child.a[j] <- 0}
                else if ((num.a[j]==0 & num.a[mate]==1)|( num.a[j]==1 & num.a[mate]==0))
                    {child.a[j] <- rbinom(1, size=1, prob=c(0.5))}
                else if ((num.a[j]==0 & num.a[mate]==2)|( num.a[j]==2 & num.a[mate]==0))
                    {child.a[j] <- 1}
                else if (num.a[j]==2 & num.a[mate]==2) {child.a[j] <- 2}
                else if ((num.a[j]==2 & num.a[mate]==1)|( num.a[j]==1 & num.a[mate]==2))
                    {child.a[j] <- rbinom(1, size=1, prob=c(0.5))+1}
                else if (num.a[j]==1 & num.a[mate]==1)
                        {child.r <- rmultinom(1, size=1, prob=c(.25, .5, .25))
            if (child.r[1,1]==1) {child.a[j]<-0}
            else if (child.r[2,1]==1) {child.a[j]<-1}
            else if (child.r[3,1]==1) {child.a[j]<-2}
                        }

##### child's genotype at locus B by Mendelian Rules  #####
     if (num.b[j]==0 & num.b[mate]==0) {child.b[j] <- 0}
                else if ((num.b[j]==0 & num.b[mate]==1)|( num.b[j]==1 & num.b[mate]==0))
                    {child.b[j] <- rbinom(1, size=1, prob=c(0.5))}
                else if ((num.b[j]==0 & num.b[mate]==2)|( num.b[j]==2 & num.b[mate]==0))
                    {child.b[j] <- 1}
                else if (num.b[j]==2 & num.b[mate]==2) {child.b[j] <- 2}
                else if ((num.b[j]==2 & num.b[mate]==1)|( num.b[j]==1 & num.b[mate]==2))
                    {child.b[j] <- rbinom(1, size=1, prob=c(0.5))+1}
                else if (num.b[j]==1 & num.b[mate]==1)
                        {child.r <- rmultinom(1, size=1, prob=c(.25, .5, .25))
            if (child.r[1,1]==1) {child.b[j]<-0}
            else if (child.r[2,1]==1) {child.b[j]<-1}
            else if (child.r[3,1]==1) {child.b[j]<-2}
                        }
```

```
        ##### child's genotype at locus C by Mendelian Rules  #####
            if (num.c[j]==0 & num.c[mate]==0) {child.c[j] <- 0}
                        else    if    ((num.c[j]==0    &    num.c[mate]==1)|(    num.c[j]==1    &
num.c[mate]==0))                        {child.c[j] <- rbinom(1, size=1, prob=c(0.5))}
                        else    if    ((num.c[j]==0    &    num.c[mate]==2)|(    num.c[j]==2    &
num.c[mate]==0))                        {child.c[j] <- 1}
                        else if (num.c[j]==2 & num.c[mate]==2) {child.c[j] <- 2}
                        else    if    ((num.c[j]==2    &    num.c[mate]==1)|(    num.c[j]==1    &
num.c[mate]==2))                        {child.c[j] <- rbinom(1, size=1, prob=c(0.5))+1}
                        else if (num.c[j]==1 & num.c[mate]==1)
                                {child.r <- rmultinom(1, size=1, prob=c(.25, .5, .25))
                    if (child.r[1,1]==1) {child.c[j]<-0}
                    else if (child.r[2,1]==1) {child.c[j]<-1}
                    else if (child.r[3,1]==1) {child.c[j]<-2}
                                }
        ###simulation of child's Traits###
        ##For Trait 1##

        if(child.a[j]==0 & child.b[j]==0){child.1[j]<-rbinom(1,size=1, prob=c(0.1))}
        else if(child.a[j]==0 & child.b[j]==1){child.1[j]<-rbinom(1,size=1, prob=c(0.3))}
        else if(child.a[j]==0 & child.b[j]==2){child.1[j]<-rbinom(1,size=1, prob=c(0.3))}
        else if(child.a[j]==1 & child.b[j]==0){child.1[j]<-rbinom(1,size=1, prob=c(0.4))}
        else if(child.a[j]==1 & child.b[j]==1){ child.1[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.a[j]==1 & child.b[j]==2){ child.1[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.a[j]==2 & child.b[j]==0){ child.1[j]<-rbinom(1,size=1, prob=c(0.4))}
        else if(child.a[j]==2 & child.b[j]==1){ child.1[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.a[j]==2 & child.b[j]==2){ child.1[j]<-rbinom(1,size=1, prob=c(0.8))}

        ##For Trait2##
        if(child.b[j]==0 & child.c[j]==0){child.2[j]<-rbinom(1,size=1, prob=c(0.1))}
        else if(child.b[j]==0 & child.c[j]==1){child.2[j]<-rbinom(1,size=1, prob=c(0.4))}
        else if(child.b[j]==0 & child.c[j]==2){ child.2[j]<-rbinom(1,size=1, prob=c(0.4))}
        else if(child.b[j]==1 & child.c[j]==0){ child.2[j]<-rbinom(1,size=1, prob=c(0.3))}
        else if(child.b[j]==1 & child.c[j]==1){ child.2[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.b[j]==1 & child.c[j]==2){ child.2[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.b[j]==2 & child.c[j]==0){ child.2[j]<-rbinom(1,size=1, prob=c(0.3))}
        else if(child.b[j]==2 & child.c[j]==1){ child.2[j]<-rbinom(1,size=1, prob=c(0.8))}
        else if(child.b[j]==2 & child.c[j]==2){ child.2[j]<-rbinom(1,size=1, prob=c(0.8))}

                } ## end of loop through individuals



        ### Make children into new parental generation###
```

```
num.a <- child.a
num.b <- child.b
num.c <- child.c
trait1 <- child.1
trait2 <- child.2
### store results for this generation ###

freq1[i] <- sum(trait1)
freq2[i] <- sum(trait2)
corAB[i] <- cor(num.a, num.b)
corBC[i] <- cor(num.b, num.c)
corAC[i] <- cor(num.a, num.c)

}  ### end of loop through generations


### logistic regression to do trend test of each trait on each locus in kids##

#####print(summary (glm(child.1~child.a,family=binomial)))
#####print(summary (glm(child.1~child.b,family=binomial)))
print(summary (glm(child.1~child.c,family=binomial)))
print(summary (glm(child.2~child.a,family=binomial)))
#####print(summary (glm(child.2~child.b,family=binomial)))
#####print(summary (glm(child.2~child.c,family=binomial)))
print(summary (glm(child.1~child.a*child.c,family=binomial)))
print(summary (glm(child.1~child.b*child.c,family=binomial)))
print(summary (glm(child.2~child.a*child.c,family=binomial)))
print(summary (glm(child.2~child.a*child.b,family=binomial)))

print(freq1/n.ind)  # freqency of trait 1 in each generation
print(freq2/n.ind)  # frequency of trait 2 in each generation
print(corAB) # genetic locus correlations in each generation
print(corBC)
print(corAC)




## End loop over replicates ##
}
```

BIBLIOGRAPHY

Bonilla, C, Gutierrez,G, Parra, EJ, Kline, C, Shriver, MD (2005) Admixture analysis of a rural population of the state of Guerrero, Mexico. Am. J. Phys. Anthropol. 128: 861-869.

Cordell HJ, Barratt, BJ, Clayton DG. (2004) Case/Pseudocontrol Analysis in Genetic Association Studies: A Unified Framework for Detection of Genotype and Haplotype Associations, Gene-Gene and Gene-Environment Interactions, and Parent of Origin Effects. Genetic Epidemiology 26 : 167-185.

Devlin B, Roeder K. (1999) Genomic Control for Association Studies. Biometrics 55(4):997-1004.

Redden DT, Allison DB. (2006) The Effect of Assortative Mating upon Genetic Association Studies: Spurious Associations and Population Substructure in the Absence of Admixture. Behavior Genetics 36 (5):678-686.

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

Risch N, Teng J (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273 –1288.

Wang Y, Localio R, Rebbeck TR.(2006) Evaluating Bias due to Population Stratification in Epidemiologic Studies of Gene-Gene or Gene-Environment Interactions. Cancer Epidemiol Biomarkers 15 (1): 124-132.