ASSOCIATING SNPS WITH BINARY TRAITS

Alexander E. Lipka

George P. McCabe

R. W. Doerge

ASSOCIATING SNPS WITH BINARY TRAITS

Alexander E. Lipka, George P. McCabe and R.W. Doerge
Department of Statistics, Purdue University
West Lafayette, IN 47907-2066

Abstract

Association mapping uses statistical analyses to test for relationships between genomic markers called single nucleotide polymorphisms (SNPs) and traits. This research focuses on the use of logistic regression to assess the additive, dominance, and epistatic effects when investigating associations between SNPs and binary traits, such as disease status. A very specific phenomenon that results in infinite maximum likelihood estimates (MLEs) of logistic regression parameters, called quasi-separation of points (QSP), is investigated. We provide a solution that relies on the use of Firth's MLE to estimate logistic regression parameters. Simulated and real data are utilized to investigate the use of Firth's MLE in a QSP setting.

Keywords: Single nucleotide polymorphism, quantitative trait loci (QTL), quasi-separation of points

1. Introduction

Many researchers are interested in finding associations between Deoxyribonucleic Acid (DNA) (Watson and Crick 1953) regions and biologically or economically important traits. A popular method for finding such associations is called association mapping (Balding 2006), which performs statistical tests at genomic markers called single nucleotide polymorphisms (SNPs) (Brookes 1999). The use of association mapping has led to the identification of genomic regions associated with multiple sclerosis (Haer et al. 2009), Parkinson's disease (Haugarvoll et al. 2009), and oleic acid content in maize (Belo et al. 2009).

Logistic regression is a common association mapping method used to identify associations between SNPs and binary traits (e.g., control versus cancer) (Haer et al. 2009). Although there are many benefits to using this approach, the presence of a phenomenon called quasi-separation of points (QSP) (Albert and Anderson 1984; McIntyre et al. 2001; Heinze and Schemper 2002; Allison 2008) results in infinite maximum likelihood estimates (MLEs) of logistic regression parameters. QSP arises in data when, for at least one SNP genotype (called a SNP type), all individuals have the same observed binary trait value (e.g., all individuals have the disease). The research presented here investigates the impact of QSP on binary trait association mapping and proposes the implementation of Firth's penalized likelihood function (Firth 1993) to obtain the MLEs of logistic regression parameters. Heinze and Schemper (2002) showed when QSP occurs, Firth's MLEs exist. We demonstrate the properties of Firth's MLEs, and show that the Firth's MLEs and the traditional MLEs yield similar values when QSP does not exist.

2. Single Nucleotide Polymorphisms (SNPs)

A SNP is "a single base pair position in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some populations, [with] the least frequent allele [having] an abundance of 1 % or greater." (Brookes 1999). SNPs are found by comparing DNA sequences between a subset of individuals, and then looking for base pair differences among the individuals at each base pair location (Brumfield et al. 2003) (Figure 1). Although a single base pair can have as many as four states or alleles, the vast majority of SNPs are diallelic. For diploid species, SNPs have four possible SNP types and can be classified as homozygous or heterozygous. Typically, the two heterozygote SNP types are pooled together (Balding 2006). SNP types are often referred to as having additive or dominance effects.

Ideally, SNPs used in association mapping are within a DNA region of interest, such as a gene (Balding 2006). The number of SNPs included in an association mapping study may depend upon the size of the genomic region under investigation. For example, a candidate gene polymorphism study focuses on only a few base pairs and usually includes only one SNP, while a genome-wide study investigates the entire genome and may include over 500,000 SNPs.

3. Association Mapping

The simplest statistical analyses used in association mapping assess the association between one SNP and a trait, while more complicated analyses test for the association between multiple SNPs and a trait (Balding 2006). These latter analyses are subdivided into whether or not haplotype information is used. Haplotypes are defined as an individual's alleles at SNP loci that are near each other in the genome (Hartl and Jones 2005). Although there are a wide variety of statistical analyses employed for association mapping, we focus on the use of logistic regression to estimate additive, dominance, and epistatic (the interaction of two or more SNPs) effects of SNPs that are associated with a binary trait.

4. Logistic Regression Model within the Context of Association Mapping

Consider the association between two SNPs, S_1 and S_2, within a genomic region of interest, and a disease Y (e.g., multiple sclerosis) for n individuals. Using a logistic regression model we can estimate the additive, dominance, and epistatic effects of the SNPs (Cordell 2002; Cordell and Clayton 2002; Balding 2006). Y_i are independent Bernoulli random variables with expected values $E(Y_i) = \pi(x_i)$ and variance $\mathrm{Var}(Y_i)=\pi(x_i)(1-\pi(x_i))$, where

$$\pi(x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)},$$

and:

$\boldsymbol{\beta} = (\beta_0, ..., \beta_{2212})$,
β_0 is the intercept,

β_{1j} is the additive effect of SNP S_j, $j \in \{1,2\}$,

β_{2j} is the dominance effect of SNP S_j, $j \in \{1,2\}$,

$\beta_{1112}, \beta_{1212}, \beta_{2112}, \beta_{2212}$ are the epistatic effects of SNPs S_1 and S_2,

$\mathbf{x}_i = (x_{11i}, x_{12i}, x_{21i}, x_{22i})$,

$$x_{1ji} = \begin{cases} 2, & \text{if indivudal } i \text{ has type } SS \text{ at SNP } S_j,\ j \in \{1,2\} \\ 1, & \text{if indivudal } i \text{ has type } Ss \text{ or } sS \text{ at SNP } S_j,\ j \in \{1,2\} \\ 0, & \text{if indivudal } i \text{ has type } ss \text{ at SNP } S_j,\ j \in \{1,2\}, \end{cases}$$

$$x_{2ji} = \begin{cases} 1, & \text{if indivudal } i \text{ has type } Ss \text{ or } sS \text{ at SNP } S_j,\ j \in \{1,2\} \\ 0, & \text{otherwise} \end{cases}, \text{ and}$$

$i = 1, \ldots, n.$

The likelihood function is

$$L(\beta \mid x_1, \ldots, x_n; Y_1, \ldots Y_n) = \prod_{i=1}^{n} \pi(x_i)^{Y_i} (1 - \pi(x_i))^{1-Y_i}$$

Fisher's scoring method (Agresti 2002) is used to obtain the MLEs, which are distributed asymptotically as a multivariate normal distribution with mean β and variance-covariance matrix $I(\beta)^{-1}$, where $I(\beta)$ is the information matrix.

5. Test Statistics and Hypothesis Testing

Without loss of generality, let β_i denote a logistic regression parameter (i.e., additive, dominance, or epistatic effect). To test the effect of a logistic regression parameter (H_0: $\beta_i = 0$) a Wald statistic, a score statistic, or a likelihood ratio test (LRT) statistic can be used. (Agresti 2002). Each test statistic has an asymptotic χ^2 distribution with 1 degree of freedom under the null hypothesis. For ease of notation, let the vector $\hat{\beta}$ denote the MLE.

The Wald statistic

$$z^2 = \left(\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right)^2,$$

has been shown to reduce to zero as the distance between MLE and the parameter being tested under the null hypothesis increases (Hauck and Donner 1977). Additionally, the statistical power of the Wald statistic decreases to the type I error rate α as the distance between actual parameter value and the null parameter value increases.

The score statistic

$$S^2 = \mathbf{U}'(\hat{\boldsymbol{\beta}}_{Null})\mathbf{I}(\hat{\boldsymbol{\beta}}_{Null})\mathbf{U}(\hat{\boldsymbol{\beta}}_{Null}),$$

evaluates the slope and expected curvature of the log likelihood function when the parameters (i.e., the additive, dominance and epistatic effects) are restricted under the null hypothesis. **U** is a vector of partial derivatives of the log L and **I** is the information matrix, both are evaluated under the null hypothesis.

The LRT statistic is

$$G^2 = -2[\log L(\hat{\boldsymbol{\beta}}_{Null} \mid \mathbf{x}_1,...,\mathbf{x}_n, Y_1,...,Y_n) - \log L(\hat{\boldsymbol{\beta}} \mid \mathbf{x}_1,...,\mathbf{x}_n, Y_1,...,Y_n)].$$

Because many studies can include upwards of 500,000 SNPs, some adjustment for the multiple testing problem is needed. While many sophisticated multiple correction procedures specifically for association mapping studies are being developed (Balding 2006), it is not uncommon to find recent studies (Pillai et al. 2009) that still use the Bonferroni procedure (Neter et al. 1996). Here, we use Simes procedure (Simes 1986) to control the false discovery rate (FDR) (Benjamini and Hochberg 1995).

6. Quasi-Separation of Points (QSP) in Association Mapping

The log likelihood function is strictly concave and bounded above, and thus for most data the MLEs exist (Silvapulle 1981; Albert and Anderson 1984). However, strict concavity and an upper bound are not sufficient conditions for the existence of MLEs (Albert and Anderson 1984; Agresti 2002). Situations can arise when the log likelihood is strictly increasing towards a horizontal asymptote for some $\beta_t \to \pm$ infinity, resulting in an infinite (i.e., nonexistent) MLE $\hat{\beta}_i$ (Albert and Anderson 1984; Heinze and Schemper 2002). This research investigates one such situation in association mapping data, namely QSP.

Consider a single SNP S_1 and its association to a disease state Y (present versus absent). Furthermore, assume that a logistic regression model is being fitted, where the SNP types from the i^{th} individual are included in the vector $\mathbf{x}_i$. Then, if there exists a vector $\mathbf{b}$ such that (i) $\mathbf{b}\mathbf{x}_i \geq 0$ whenever $Y_i = 1$ and (ii) $\mathbf{b}\mathbf{x}_i \leq 0$ whenever $Y_i = 0$, with equality for at least one i in both (i) and (ii), then QSP is said to be present (Albert and Anderson 1984; Heinze and Schemper 2002). A common situation that results in QSP is when $\mathbf{b}$ subdivides the data into two groups, and all individuals in one group are either $Y_i = 1$ or $Y_i = 0$. Within the context of association mapping, QSP occurs if there is at least one SNP type (or combination of two SNP types if two SNPs are considered) where all individuals have the same disease state. To illustrate QSP in association mapping, consider Table 1, where all 144 individuals with the heterozygous (Ss / sS) S_1 SNP type have the disease. These data are simulated with a dominance effect of 3. In the presence of QSP, the Wald statistic cannot be used (because it uses the MLE directly in its calculations). However, the score statistic (which uses the MLE under H_0) can, and the LRT

statistic (which compares the approximate maximum of the log likelihood function to the maximum of the log likelihood function under H_0) can be approximated.

7. Firth's Modified Scoring Procedure

Firth's penalized likelihood function is a function of the previous stated likelihood function multiplied by a penalty term that is known as Jeffreys prior (Jeffreys 1946).

$$L^*(\beta \mid x_1,...,x_n;Y_1,...Y_n) = \prod_{i=1}^{n} \pi(x_i)^{Y_i}(1-\pi(x_i))^{1-Y_i} \mid I(\beta) \mid^{\frac{1}{2}}$$

The influence of Jeffreys prior is asymptotically negligible (Firth 1993; Heinze and Schemper 2002) but very important in the presence of QSP since it provides MLEs, known as Firth's MLEs, that are distributed asymptotically as a multivariate normal distribution with mean β and variance-covariance matrix $I(\beta)^{-1}$ (Firth 1993; Heinze and Schemper 2002).

Firth's MLE will always be finite and unique (Firth 1993; Heinze and Schemper 2002). Recall that the traditional log likelihood function is strictly concave and bounded above. Furthermore, the log of Jeffreys prior is strictly concave and has no lower bound as the logistic regression parameters tend to tend to +/- infinity (Firth 1993). Therefore, the addition of Jeffreys prior to the traditional log likelihood function does not affect the concavity and boundedness of log L^*, and ensures the existence of Firth's MLE, even in the presence of QSP. When QSP is not present in the data the traditional MLEs exist and are comparable to the Firth MLEs.

8. Data Analysis

We investigate the impact of QSP on association mapping studies in a logistic regression setting, and compare the estimates and standard errors of Firth's MLEs to traditional MLEs. Specifically, we concentrate on SNPs where QSP occurs since, unlike the traditional MLE, Firth's MLEs exist in the presence of this phenomenon. When QSP is not present we investigate the similarity between Firth MLE and traditional MLEs.

Simulation Study 1. We investigate estimates of additive and dominance effects in the presence of QSP. Consider one binary trait and a single SNP across sample sizes n=300, 500, and 1000 individuals. The trait is simulated from a logistic regression model. The data are simulated at 16 settings of logistic regression parameters values, where the additive and dominance parameters vary from 0 to 3, and the intercept is either 0 or -2.08. Note that as the additive and dominance parameters get larger and as the sample size decreases, the chance of observing QSP increases. At each parameter value and sample size, 1000 data sets are simulated, and the proportion of data sets where QSP occurs is noted. These proportions are summarized in Figure 2. At each data set where QSP does not occur, Firth's MLE is compared to the traditional MLE. These results, summarized in Figure 3, suggest that these two estimates are similar when QSP does not occur. In conclusion, this simulation study demonstrates that QSP is more likely to occur at larger additive effects and at smaller sample sizes. Consequently, the implementation of Firth's MLE should allow researchers to estimate the effects of more SNPs with large additive effects.

Simulation Study 2. In a second simulation study we investigate QSP in the presence of epistasis (the interaction of two SNPs). Consider one binary trait and two SNPs, S_1 and S_2, respectively for sample sizes of 300, 500, or 1000 individuals. The trait is simulated from a logistic regression model that includes four epistatic effects which we allow to vary from 0 to 3. For each simulated data set, the Firth's MLE and the traditional MLE of logistic regression parameters are obtained and compared. The results of this study indicate that QSP is likely to occur in the presence of epistasis. Thus, Firth's MLEs provide estimates of epistatic effects in situations where the traditional MLEs do not exist.

Real Data. We rely on late onset Alzheimer's disease (LOAD) association mapping data to demonstrate the worth of Firth's MLE in an actual situation where QSP exists. It is well known that humans with the ε4 allele of the apoliopoprotein E (APOE) gene on Chromosome 19 have a greater risk of developing (LOAD) (Corder et al. 1993; Corder et al. 1994; Farrer et al. 1997). Reiman et al. (2007) conducted a series genome-wide association mapping studies to investigate additional genes that may be associated with LOAD. Their data consist of 1,411 LOAD cases and controls, and a Pearson chi-squared test was employed to test each of the 312,316 SNPs. Their results led to the novel discovery of the GRB-associated binding protein 2 (GAB2) gene on Chromosome 11, which contributes to an increased risk of LOAD for carriers of the APOE ε4 allele. We reanalyze data from Reiman et al. (2007) in four genome-wide studies using Firth's MLE of logistic regression parameters, with a particular focus on SNPs where QSP occurs.

The data collected by Reiman et al. (2007) consist of 861 LOAD cases and 550 controls from three cohorts and include a total of 312,316 SNPs. The data also contain genotypic information on the APOE gene for each individual. Of interest, 644 of the individuals are APOE ε4 carriers, and 767 are APOE ε4 noncarriers. Since Reiman et al. (2007) obtained similar results across all three cohorts, all individuals are pooled together in the following analyses.

We first estimated the additive, dominance, and epistatic (with the APOE gene) effects of the SNP data. A logistic regression model based on two markers (i.e., independent variables) is employed. One marker is a SNP (S_1) and the other marker is the APOE gene. In an effort to include as many SNPs as possible, the second, third, and fourth genome-wide analyses use a simpler logistic regression model that includes only additive and dominance parameters. Specifically, the first two analyses include all 1,411 individuals and 251,974 SNPs. In order to test for interactions between the APOE gene and SNPs the third analysis includes only the 644 APOE ε4 carriers and 234,463 SNPs, while the final analysis includes only the 767 APOE ε4 noncarriers and 242,196. Table 2 provides a summary of the number of SNPs and individuals included in each of the four studies.

In each of the four analyses, Firth's MLE and the traditional MLE of logistic regression parameters are obtained (if they exist) for each SNP. With the exception of the intercept parameter all parameters are tested using a LRT statistic. The Simes procedure (Simes 1986) is used to adjust for the genome-wide multiple testing (α=0.05)..

Recall that a SNP can have three possible SNP types (with the two heterozygote SNP types pooled together), and note that there are nine possible SNP type/APOE gene combinations. To obtain an accurate estimate of the probability of an individual having LOAD in each of these respective categories, there must be an adequate number of individuals in each category. Using the guidelines recommended in Peduzzi et al. (1996) for conducting logistic regression, a SNP is included if there are at least ten individuals observed in each category. Based on these criteria, a total of 125,613 SNPs are included in our first analysis of all the data. The previously mentioned SNP numbers for the other three studies satisfy the criterion that at least ten individuals must be observed in each SNP type category.

The number of SNPs in each of these four studies where QSP occurs is summarized in Table 3. Note that around 1% of the SNPs in the APOE ε4 carriers have QSP. The phenomenon of QSP is more prevalent in this study because, compared to APOE ε4 noncarriers, a greater proportion of APOE ε4 carriers have LOAD. Therefore, for a given SNP in APOE ε4 carriers, the chance of at least one SNP type being the same for all individuals with LOAD is greater. Similar to the results in the simulation studies, the estimates and standard errors of Firth's MLEs are similar to those of the traditional MLE at SNPs where QSP is not present, and exist in situations where QSP occurs.

The main neuropathological finding for these four genome-wide studies is that, among the APOE ε4 carriers, SNP A-2313615 in the GAB2 gene on Chromosome 11 has a statistically significant additive effect on LOAD. At this SNP, the Firth's MLE of the additive effect is 1.25 with a standard error of 0.23, and the traditional MLE of the additive effect is 1.26 with a standard error of 0.23. Note that since QSP does not occur at this SNP, the traditional MLE of the additive effect exists and is similar to Firth's MLE of the additive effect. Although none of their corresponding LRT statistics are statistically significant, the remaining nine SNPs linked to the GAB2 gene have large additive effects on LOAD (relative to the other SNPs) among the APOE ε4 carriers. Since QSP is not present at any of these SNPs, the values and standard errors of traditional MLEs and Firth's MLEs are similar. These results are suggestive of an additive biological effect of the GAB2 gene on LOAD that is not confirmed in the second (all individuals) and fourth investigation (APOE ε4 noncarriers). Therefore, there is evidence that the GAB2 gene plays a role in the risk of LOAD for APOE ε4 carriers. This reanalysis augments the main conclusion of Reiman et al. (2007) by suggesting that the association between the GAB2 gene and LOAD risk among APOE ε4 carriers is an additive effect. Additionally, an interesting novel result is that among the APOE ε4 carriers, SNP A-4202283 on Chromosome 11 has marginally significant likelihood-ratio test results when testing for its additive effect. QSP is not present at this SNP. The Firth's MLE of its additive effect is -0.98 with a standard error of 0.25, while the corresponding traditional MLE is -1.02 with a standard error of 0.24. This result for SNP A-4202283 may justify further investigation into the biological function of this SNP's surrounding genomic region. The results of the analysis of Chromosome 11 for APOE ε4 carriers are summarized in Figure 4.

9. Conclusions

It appears that the application of Firth's MLE to estimate logistic regression parameters in the context of association mapping is a viable solution to the problems that arise when QSP is present in data. When the traditional MLE exists, the Firth's MLE and the traditional MLE yield similar values and standard errors, and hence similar estimates of a SNP's additive, dominance, and epistatic effects are obtained. However, in the presence of QSP, Firth's MLE exists, guaranteeing the estimation of the additive, dominance, and epistatic effects of SNPs, and thus provides greater insight into the underlying genomic mechanisms that control binary traits.

10. Summary

This research investigates the impact of QSP on association mapping results when logistic regression is used. Two simulation investigations and four genome-wide association mapping analyses are conducted. Firth's MLEs can be successfully employed as estimates for additive, dominance, and epistatic effects of SNPs when QSP is present, and are similar to the traditional MLE when QSP does not exist.

11. Acknowledgements

The authors would like to thank the RWD research group, the Purdue University Department of Statistics, its systems administrators Doug Crabill and My Troung, and David Whittinghill for their support. The authors would also like to thank the Kansas State University Department of Statistics for the opportunity to present this work at the 2009 Applied Statistics in Agriculture Conference, as well as for their financial support.

12. References

A. Agresti. Categorical Data Analysis, Second Edition. Hoboken, NJ: Wiley, 2002.

A. Albert and J.A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71:1-10, 1984

P.D. Allison. Convergence failures in logistic regression. SAS Global Forum, Statistics and Data Analysis, 360-2008, 2008.

D.J. Balding. A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7:781-791, 2006.

A. Belo, P. Zheng, S. Luck, B. Shen, D. J. Meyer, B. Li, S. Tingey, and A. Rafalski. Whole genome scan detects an allelic variant of Fad2 associated with increased oleic acid levels in maize. Molecular Genetics and Genomics, 279:1-10, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57:289-300, 1995.

A. J. Brookes. The essence of SNPs. Gene, 234:177-186, 1999.

R.T. Brumfield, P. Beerli, D.A. Nickerson, and S.V. Edwards. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution, 18(5):249-256, 2003.

H.J. Cordell and D.G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in Type 1 Diabetes. American Journal of Human Genetics, 70:124-141, 2002.

H.J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics, 11 (20):2463-2468, 2002.

E.H. Corder, A.M. Saunders, N.J. Risch, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, J.B. Rimmler, P.A. Locke, P.M. Conneally, and K.E. Schmader. Protective effect of Apoliopoprotein E Type 2 allele for late onset Alzheimer's disease. Nature Genetics, 7:180-184, 1994.

E.H. Corder, A.M. Saunders, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, G.W. Small, A.D. Roses, J.L. Haines, and M.A. Pericak-Vance. Gene dose of Apoliopoprotein E Type 4 allele and the risk of Alzheimer's disease in late onset families. Science, 261:921-923, 1993.

L.A. Farrer, L.A. Cupples, J.L. Haines, B. Hyman, W.A. Kukull, R. Mayeux, R.H. Myers, M.A. Paricak-Vance, N. Risch, and C.M. van Duijn. E_ects of age, sex and ethnicity on the association between Apolipoprotein E genotype and Alzheimer's disease. A meta-analysis. APOE and Alzheimer's Disease Meta Analysis Consortium. JAMA, 278:1349-1356, 1997.

A. Farwick, B. Dasch, B. H. F. Weber, D. Pauleikhoff, M. Stoll, and H.W. Hense. Variations in five genes and the severity of age-related macular degeneration: results from the muenster aging and retina study. Eye, doi: 10.1038/eye.2008.426, 2009.

D. Firth. Bias reduction of maximum likelihood estimates. Biometrika, 80:27-38, 1993.

A. Franke, T. Balschun, T. H. Karlsen, J. Hedderich, S. May, T. Lu, D. Schuldt, S. Nikolaus, P. Rosenstiel, M. Krawczak, and S. Schreiber. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nature Genetics, 40:713-715, 2008.

J.P. Haer, L. M. Maier, J. D. Cooper, V. Plagnol, A. Hinks, M. J. Simmonds, H. E. Stevens, N. M. Walker, B. Healy, J. M. M. Howson, M. Maisuria, S. Duley, G. Coleman, S. C. L. Gough, The International Multiple Sclerosis Genetics Consortium (IMSGC), J. Worthington, V. K. Kuchroo, L. S. Wicker, and J. A. Todd. Cd226 Gly307Ser association with multiple autoimmune diseases. Genes and Immunity, 10:5-10, 2009.

D.L . Hartl and E.W. Jones. Genetics: Analysis of Genes, and Genomes, Sixth Edition. Sudbury, MA: Jones and Bartlett, 2005.

W.W. Hauck and A. Donner. Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association, 72:851-853, 1977.

K. Haugarvoll, M. Toft, L. Skipper, M. G. Heckman, J. E. Crook, A. Soto, O. A. Ross, M. M. Hulihan, J. M. Kachergus, S. B. Sando, L. R. White, T. Lynch, M.Gibson, R. J. Uitti, Z.K. Wszolek, J. O. Aasly, and M. J. Farrer. Fine-mapping and candidate gene investigation within the PARK10 locus. European Journal of Human Genetics, 17:336-343, 2009.

G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. Statistics in Medicine, 21:2409-2419, 2002.

H. Jeffreys. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London, Series A, 186:453-461, 1946.

L.M. McIntyre, C.J. Coffman, and R.W. Doerge. Detection and localization of a single binary trait locus in experimental populations. Genetics Research, 78:79-92, 2001.

J. Neter, M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. Applied Linear Statistical Models, Fourth Edition. Boston: McGraw-Hill, 1996.

A. Papassotiropoulos, M. Fountoulakis, T. Dunckley, D.A. Stephan, and E.M. Reiman. Genetics, transcriptomics and protemoics of Alzheimer's disease. Journal of Clinical Psychiatry, 67:652-670, 2006.

P. J. Peduzzi, E. Concato, T.R. Kemper, and Holford A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49:1373-1379, 1996.

S. G. Pillai, D. Ge, G. Zhu, X. Kong, K. V. Shianna, A. C. Need, S. Feng, C. P. Hersh, P. Bakke, A. Gulsvik A. Ruppert, K. C. Ldrup Carlsen, A. Roses, W. Anderson, ICGN Investigators, S. I. Rennard, D. A. Lomas, E. K. Silverman, and D. B. Goldstein. A genome-wide association study in Chronic Obstructive Pulmonary Disease (COPD): Identification of two major susceptibility loci. PLoS Genetics, 5(3):e1000421, 2009.

E.M. Reiman, J.A. Webster, A.J. Myers, J. Hardy, T. Dunckley, V.L. Zismann, K.D. Joshipura, J.V. Pearson, D. Hu-Lince, M. J. Huentelman, D. W. Craig, K.D. Coon, W.S. Liang, R.H. Herbert, T. Beach, K. C. Rohrer, A.S. Zhao, D. leung, L. Bryden, L. Marlowe, M. Kaleem, D. Mastroeni, A. Grover, C. B. Heward, R. Ravid, J. Rogers, M.L. Hutton S. Melquist, R. C. Petersen, G.E. Alexander, R. J. Caselli, W. Kukull, A. Papassotiropoulos, and D. A. Stephan. GAB2 alleles modify Alzheimer's risk in APOE Epsilon-4 carriers. Neuron, 54:713-720, 2007.

SAS Institute Inc. SAS/STAT User's Guide, Version 8. SAS Institute Inc.: Cary, NC, 1999.

M.J. Silvapulle. On the existence of maximum likelihood estimates for the binomial reponse models. Journal of the Royal Statistical Society B, 43:310-313, 1981.

R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73:751-754, 1986.

C. I. G. Vogel, B. Greene, A. Scherag, T. D. Mller, S. Friedel, H. Grallert I. M. Heid, T. Illig, H. E. Wichmann, H. Schfer, J. Hebebrand, and A. Hinney. Non-replication of an association of CTNNBL1 polymorphisms and obesity in a population of central European ancestry. BMC Medical Genetics, 10:14, 2009.

J.D. Watson and F.H.C. Crick. Molecular strucutre of nucleic acids. Nature, 171(4356):737-738, 1953.

Figures and Tables

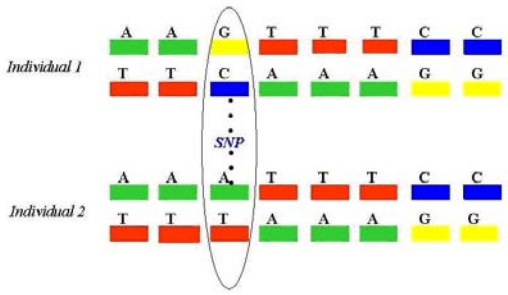

Figure 1. Comparison of DNA sequences between two individuals. Alleles at the circled site differ, and are considered a SNP.

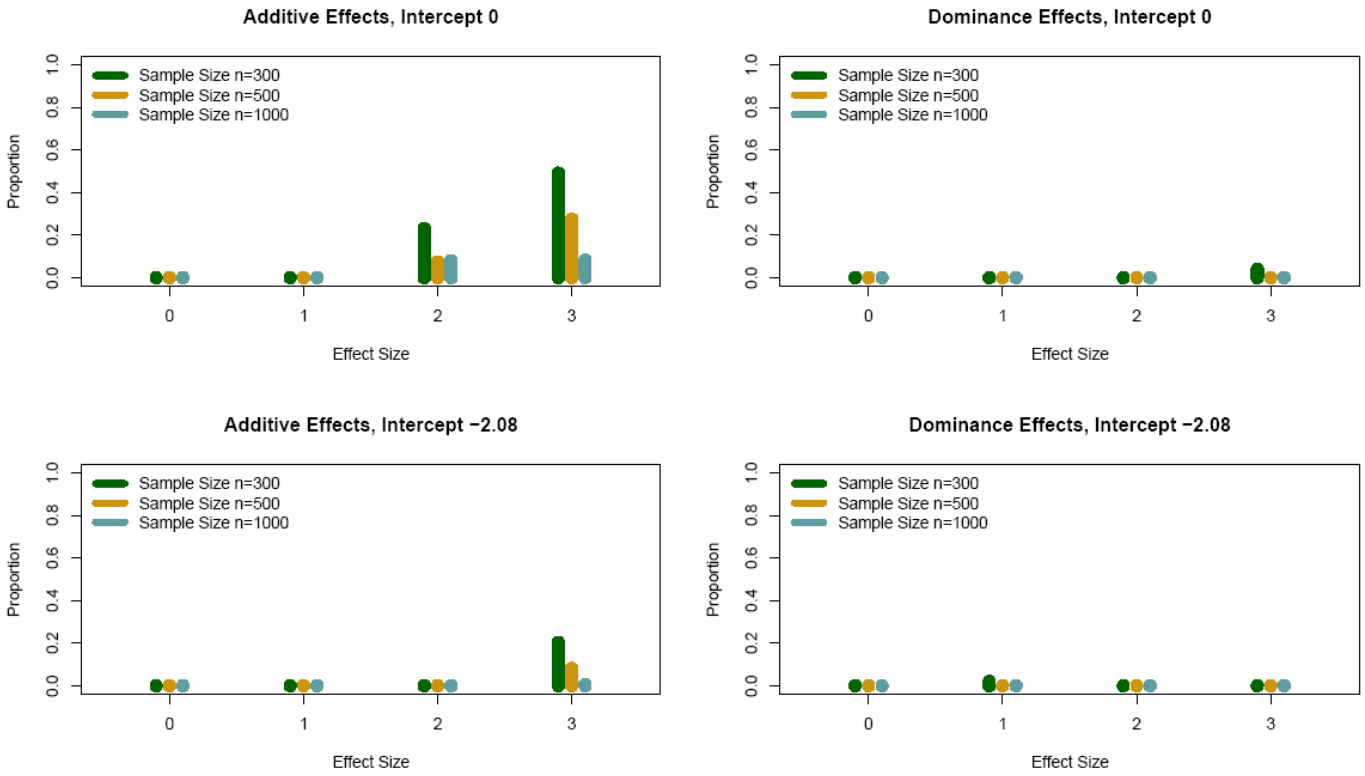

Figure 2. The proportion of data sets (out of 1000) that have QSP in each of the 48 simulation settings.

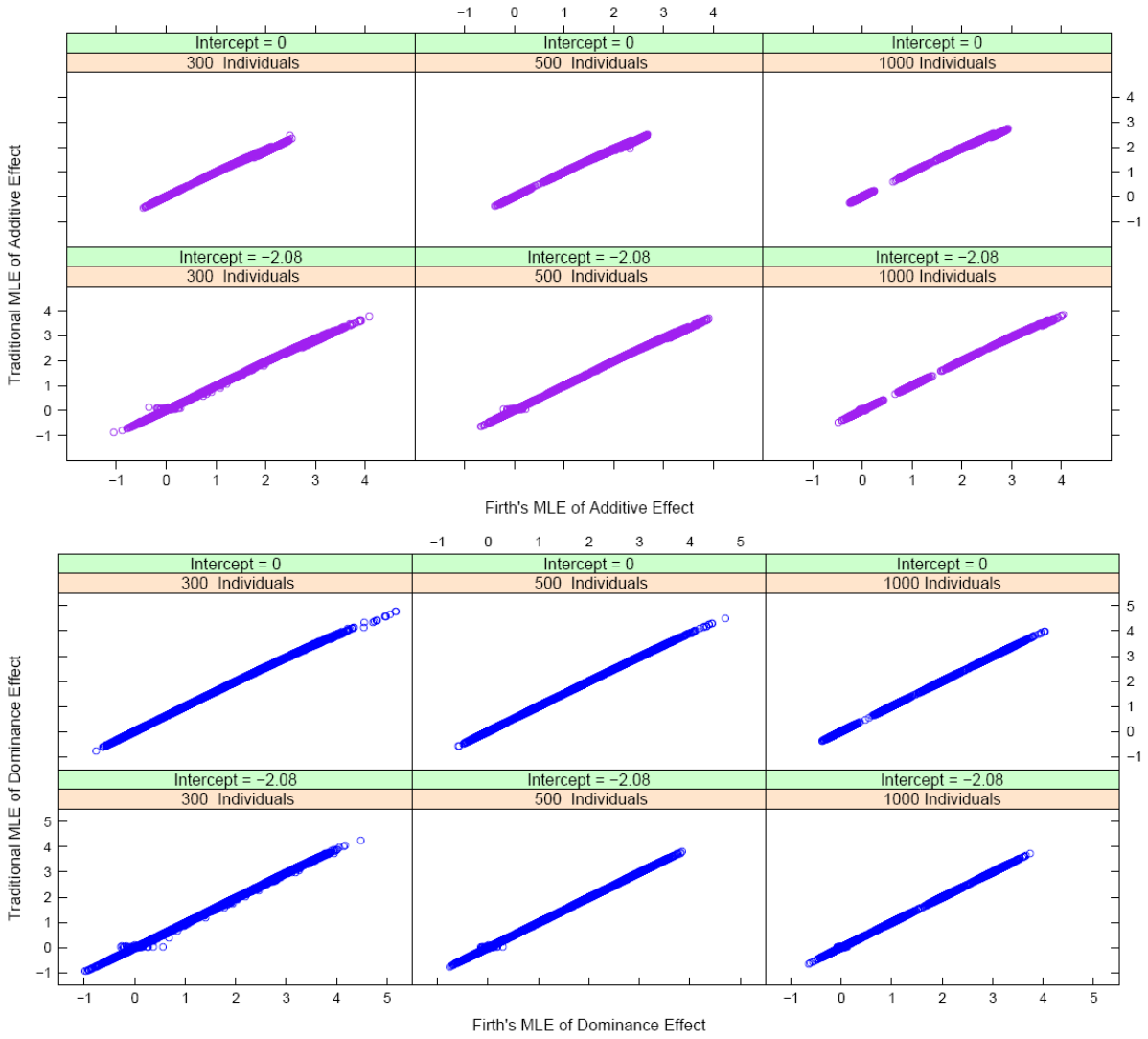

Figure 3. For every simulated data set in Simulation Study 1 where QSP does not occur, the traditional MLE of the indicated effect is plotted against its corresponding Firth's MLE. The x-axis on each graph is the Firth's MLE and the y-axis is the traditional MLE. The majority of the points in these plots lie approximately on the identity line. This indicates that the Firth's MLE and the traditional MLE of the additive and dominance effects are similar when the traditional MLE exists.

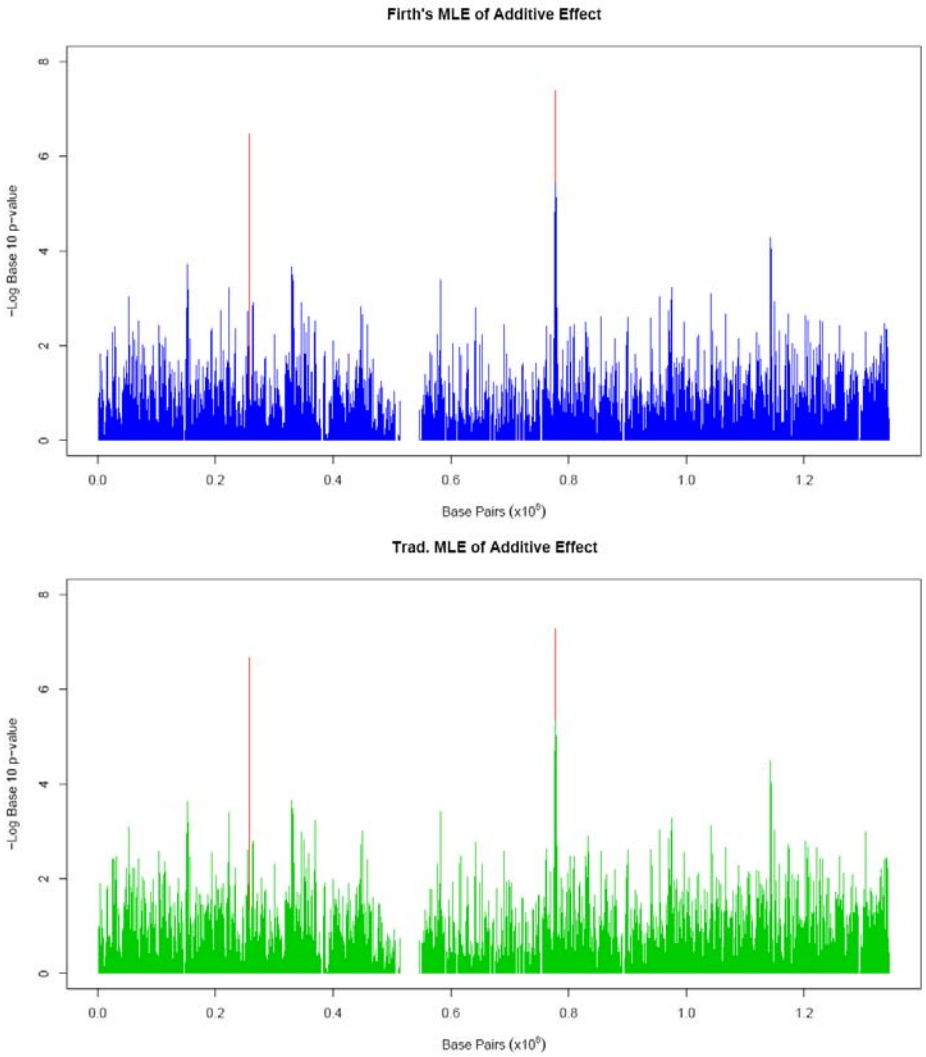

Figure 4. Chromosome 11 LRT statistic results for all 644 APOE ε4 carriers. The top graph shows results from Firth's penalized likelihood function, and the bottom graph shows the results from the traditional likelihood function. The statistically significant additive effect of the SNP (A-2313615) in the GAB2 gene is located at base pair position 77722719 and is highlighted in red. The GAB2 gene is located approximately between base pairs 77608147 and 77768798. Additionally, SNP A-4202283, located at base pair position 25747721 and highlighted in red, has a significant additive effect.

Table 1. Illustrative example where with 300 disease cases and controls, and one SNP S_1. These data are simulated with a dominance effect of 3. QSP occurs since all 144 individuals with SNP type Ss/sS are disease cases.

S1 SNP Type	Disease Case	Disease Control
Ss	36	38
Ss/sS	144	0
SS	42	40

.

Table 2. Summary of the effects assessed, the subjects included, and the number of SNPs included in four studies conducted for the reanalysis of the LOAD data (Reiman et al., 2007).

Study	Effects Assessed	Subjects Included	Number of SNPs
1	Additive, Dominance, Epistatic	1,411	125,613
2	Additive, Dominance	1,411	251,974
3	Additive, Dominance	644 APOE ε4 Carriers	234,463
4	Additive, Dominance	767 APOE ε4 Noncarriers	242,196

Table 3. Numerical summary of the number of SNPs in each of the four LOAD studies where QSP occurs. Study 1 has epistatic parameters (with the APOE gene) in the logistic regression model and includes all 1,411 individuals; Study 2 includes all 1,411 individuals, Study 3 includes all 644 APOE ε4 carriers; and Study 4 includes all 767 APOE ε4 noncarriers.

	Study 1	Study 2	Study 3	Study 4
Number of SNPs	125,613	251,974	234,463	242,196
Number of SNPs with QSP	8	88	3,226	37