Title

A Variable Age of Onset Segregation Model for Linkage Analysis, with Correction for Ascertainment, Applied to Glioma

Journal

Cancer Epidemiology Biomarkers & Prevention, 21(12)

ISSN

1055-9965

Authors

Sun, Xiangqing

Vengoechea, Jaime

Elston, Robert

et al.

Publication Date

2012-12-01

DOI

10.1158/1055-9965.epi-12-0703

Peer reviewed

Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2012 December ; 21(12): 2242–2251. doi: 10.1158/1055-9965.EPI-12-0703.

A variable age of onset segregation model for linkage analysis, with correction for ascertainment, applied to glioma

Xiangqing Sun[*,1], **Jaime Vengoechea**[*,2], **Robert Elston**[1], **Yanwen Chen**[1], **Christopher I. Amos**[3], **Georgina Armstrong**[4], **Jonine L Bernstein**[5], **Elizabeth Claus**[6], **Faith Davis**[7], **Richard S Houlston**[8], **Dora Il'yasova**[9], **Robert B Jenkins**[10], **Christoffer Johansen**[11], **Rose Lai**[12], **Ching C Lau**[4], **Yanhong Liu**[4], **Bridget J McCarthy**[7], **Sara H Olson**[5], **Siegal Sadetzki**[13], **Joellen Schildkraut**[9], **Sanjay Shete**[14], **Robert Yu**[14], **Nicholas A Vick**[15], **Ryan Merrell**[15], **Margaret Wrensch**[16], **Ping Yang**[10], **Beatrice Melin**[17], **Melissa L. Bondy**[4], **Jill S. Barnholtz-Sloan**[1,18], and **on behalf of the Gliogene Consortium**

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, Ohio

[2] Departments of Internal Medicine and Genetics, Case Western Reserve University, Cleveland, Ohio

[3] Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, Texas

[4] Department of Pediatrics, Division of Hematology-Oncology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas

[5] Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center; New York, New York

[6] Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut and Department of Neurosurgery, Brigham and Women's Hospital, Boston, Massachusetts

[7] Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, Illinois

[8] Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, United Kingdom

[9] Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina

[10] Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota

[11] Department of Neurology; Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark

[12] The Neurological Institute of Columbia University, New York, New York

[13] Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel and Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel

[14] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas

[15] Evanston Kellogg Cancer Care Center NorthShore University HealthSystem, Evanston, Illinois

[16] Department of Neurological Surgery, University of California, San Francisco, San Francisco, California

[17] Department of Radiation Sciences Oncology, Umea University, Umea, Sweden.

[18] Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio.

Abstract

Correspondance to: Jill Barnholtz-Sloan, PhD Case Comprehensive Cancer Center 11100 Euclid Ave, Wearn 152 Cleveland, Ohio 44106-5065 216-368-1506 (Phone) 216-368-2606 (Fax) jsb42@case.edu.
*These authors contributed equally to this work.

Conflict of Interest: The authors report no conflict of interest.

The authors acknowledge the contributions of the following individuals to the overall brain tumor research programs—MD Anderson Cancer Center: Phyllis Adatto, Fabian Morice, Sam Payen, Lacey McQuinn, Rebecca McGaha, Sandra Guerra, Leslie Paith, Katherine Roth, Dong Zeng, Hui Zhang, Dr. Alfred Yung, Dr. Howard Colman, Dr. Charles Conrad, Dr. John de Groot, Dr. Arthur Forman, Dr. Morris Groves, Dr. Victor Levin, Dr. Monica Loghin, Dr. Vinay Puduvalli, Dr. Raymond Sawaya, Dr. Amy Heimberger, Dr. Frederick Lang, Dr. Nicholas Levine, Lori Tolentino; Brigham and Women's Hospital: Kate Saunders, Donna DelloIacono; Case Western Reserve University: Dr. Stanton Gerson, Dr. Warren Selman, Dr. Robert Maciunas, Dr. Nicholas Bambakidis, Dr. David Hart, Dr. Jonathan Miller, Dr. Alan Hoffer, Dr. Mark Cohen, Dr Lisa Rogers, Dr. Charles J Nock, Wendi Barrett, Anita Merriam, Quinn Ostrom, Sarah Robbins, Perica Davitkov, Dr. Michael Vogelbaum, Dr. Robert Weil, Dr. Manmeet Ahluwalia, Dr. David Peereboom, Dr. Edward Benzel, Dr. Susan Staugaitis, Cathy Schilero, Cathy Brewer, Kathy Smolenski, Diane Fabec, Theresa Naska, Jennifer Hornacek-Guadalupe; Columbia University Medical Center: Dr. Steven Rosenfeld; Israel: Dr. Zvi Ram, Dr. Deborah T Blumenthal, Dr. Felix Bokstein (Tel-Aviv Sourasky Medical Center), Dr. Felix Umansky (Hadassah – Hebrew University Medical Center, Henry Ford Hospital), Dr. Menashe Zaaroor (Rambam – Health Care Campus) Dr. Avi Cohen (Soroka University Medical Center, Chaim Sheba Medical Center), Dr. TzeelaTzuk-Shina (Rambam Medical Center and Faculty of Medicine, Technion-Israel Institute of Technology); Denmark: Dr. Bo Voldby (Aarhus University Hospital), Dr. René Laursen M.D. (Aalborg University Hospital), Dr. Claus Andersen (Odense University Hospital), Dr. Jannick Brennum (Glostrup University Hospital), Matilde Bille Henriksen (Institute of Cancer Epidemiology, the Danish Cancer Society); Memorial Sloan-Kettering Cancer Center: Maya Marzouk, Mary Elizabeth Davis, Eamon Boland, Marcel Smith, Ogechukwu Eze, Mahalia Way; NorthShore University HealthSystem: Pat Lada, Nancy Miedzianowski, Michelle Frechette, Dr. Nina Paleologos; Sweden: Gudrun Byström, Sara Huggert, Mikael Kimdal and Monica Sandström (Umea University); University of California, San Francisco: Dr. Tarik Tihan, Dr. Shichun Zheng, Dr. Mitchel Berger, Dr. Nicholas Butowski, Dr. Susan Chang, Dr. Jennifer Clarke, Dr. Michael Prados, Terri Rice, Jeannette Sison, Valerie Kivett, Xiaoqin Duo, Helen Hansen, George Hsuang, Rosito Lamela, Christian Ramos, Joe Patoka, Katherine Wagenman, Mi Zhou, Adam Klein, Nora McGee, Jon Pfefferle, Callie Wilson, Pagan Morris, Mary Hughes, Marlin Britt-Williams, Jessica Foft, Julia Madsen, Csaba Polony; University of Illinois at Chicago: Candice Zahora, Dr. John Villano, Dr. Herbert Engelhard.
The authors acknowledge the Gliogene Consortium whose members are: Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas (Sanjay Shete, Robert K. Yu); Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, Texas (Christopher Amos); Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, Texas (Kenneth D. Aldape); Department of Neuro-Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas (Mark R. Gilbert); Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, Texas (Jeffrey Weinberg); Department of Pediatrics, Section of Hematology and Oncology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas (Ching C. Lau, Eastwood Honchiu Leung, Caleb Davis, Rita Cheng, Chris Man, Rudy Guerra, Sivashankarappa Gurusiddappa, Michael E. Scheurer, Melissa L. Bondy, Georgina N. Armstrong, Yanhong Liu); Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, United Kingdom (Richard S. Houlston, FayJ.Hosking, Lindsay Robertson, Elli Papaemmanuil); Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut (Elizabeth B. Claus); Department of Neurosurgery, Brigham and Women's Hospital, Boston, Massachusetts (Elizabeth B. Claus);Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio (Jill Barnholtz-Sloan, Andrew E. Sloan, Gene Barnett, Karen Devine, Yingli Wolinsky); The Neurological Institute of Columbia University, New York, New York (Rose Lai, Erika Florendo, Delcia Rivas, Christina Corpuz); Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina (Dora Il'yasova, Joellen Schildkraut); Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel (Siegal Sadetzki, Galit Hirsh Yechezkel, Revital Bar-Sade Bruchim, Lili Aslanov); Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel (Siega l Sadetzki); Department of Neurology; Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark (Christoffer Johansen, Hanne Bødtcher); Neurosurgery Department, Rigshospitalet, University Copenhagen (Michael Kosteljanetz), Neuropathology Department, Rigshospitalet, University Copenhagen (Helle Broholm); Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York (Jonine L. Bernstein, Sara H. Olson, Erica Schubert), Department of Neurology, Memorial Sloan-Kettering Cancer Center, New York, New York (Lisa DeAngelis); Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota (Robert B. Jenkins, Ping Yang, Amanda Rynearson); Department of Radiation Sciences Oncology, Umea University, Umea, Sweden (Beatrice S. Melin, Roger Henriksson, Ulrika Andersson), Department of Medical Biosciences, Umea University, Umea , Sweden (Thomas Brannstrom); Evanston Kellogg Cancer Care Center, North Shore University Health System, Evanston, Illinois (Nicholas A. Vick); Departments of Neurological Surgery and Epidemiology and Biostatistics (Margaret Wrensch, John Wiencke, Joe Wiemels, Lucie McCoy) Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, Illinois (Bridget J. McCarthy, Faith G. Davis).
The authors acknowledge the input of the Gliogene External Advisory Committee: Dr. Ake Borg (Department of Oncology, Lund University, Lund, Sweden), Dr. Stephen K Chanock (National Cancer Institute, U. S. National Institutes of Health), Dr. Peter Collins (University of Cambridge, United Kingdom), Dr. Robert Elston (Department of Epidemiology and Biostatistics, Case Western Reserve University), Dr. Paul Kleihues (Department of Pathology, University Hospital, Zurich, Switzerland), Carol Kruchko (Central Brain Tumor Registry of the United States), Dr. Gloria Petersen (Health Sciences Research, Mayo Clinic), Dr. Sharon Plon (Baylor Cancer Genetics Clinic, Baylor College of Medicine)
The authors would also like to acknowledge the Brain Tumor Epidemiology Consortium (BTEC) for its support of the Gliogene study. Finally, we would like to thank the patients and their families for participating in this research.

Background—We propose a two-step model-based approach, with correction for ascertainment, to linkage analysis of a binary trait with variable age of onset and apply it to a set of multiplex pedigrees segregating for adult glioma.

Methods—First, we fit segregation models by formulating the likelihood for a person to have a bivariate phenotype, affection status and age of onset, along with other covariates, and from these we estimate population trait allele frequencies and penetrance parameters as a function of age (N=281 multiplex glioma pedigrees). Second, the best fitting models are used as trait models in multipoint linkage analysis (N=74 informative multiplex glioma pedigrees). To correct for ascertainment, a prevalence constraint is used in the likelihood of the segregation models for all 281 pedigrees. Then the trait allele frequencies are re-estimated for the pedigree founders of the subset of 74 pedigrees chosen for linkage analysis.

Results—Using the best fitting segregation models in model-based multipoint linkage analysis, we identified two separate peaks on chromosome 17; the first agreed with a region identified by Shete et al. who used model-free affected-only linkage analysis, but with a narrowed peak: and the second agreed with a second region they found but had a larger maximum log of the odds (LOD).

Conclusions/Impact—Our approach has the advantage of not requiring markers to be in linkage equilibrium unless the minor allele frequency is small (markers which tend to be uninformative for linkage), and of using more of the available information for LOD-based linkage analysis.

Keywords

Glioma; model-based linkage; segregation; age of onset; prevalence constraint

Introduction

Successful linkage analysis of complex diseases, when performed to obtain log of the odds (LODs), requires a number of assumptions related to both the markers and the trait of interest (See Supplemental Table S1). Both model-based and model-free linkage analysis, where the term “model” refers to the genetic model for the trait undergoing analysis, typically assume known marker genotypic frequencies in pedigree founders, known recombination fractions between markers, and lack of interference between markers. Both models usually assume all markers are in Hardy-Weinberg equilibrium. A key difference between these two types of linkage analysis, once certain model parameters are assumed to be known, is the direction of the approach: typically, model-free linkage analyzes the markers conditional on the trait, whereas model-based linkage analyzes the trait conditional on the markers. In model-free linkage analysis the markers must be independent or their dependencies must be correctly modeled. Conversely, in model-based linkage analysis, the pedigree members’ trait values must be independent, or their dependencies must be correctly modeled. Hence the models differ in their assumptions regarding linkage equilibrium of the markers: model-free linkage requires linkage equilibrium, though this assumption may be ignored when comparing affected sib pairs to discordant sib pairs (2); model-based linkage does not require one to assume linkage equilibrium among the markers – but we do typically assume random ascertainment of markers when estimating their genotypic frequencies.

With respect to the trait, model-free linkage analysis does not require known genotypic frequencies in the pedigree founders or any penetrance parameters. Model-based linkage analysis requires known penetrance parameters. The assumption that the penetrance parameters are known is a major obstacle to carrying out model-based linkage studies, and represents the main reason why most linkage studies of complex diseases are conducted using a model-free approach.

Age-of-onset data can be incorporated into segregation models to determine the penetrance parameters of the different genotypes as functions of age. Segregation models can then be used to empower subsequent linkage analysis (3, 4). Prior studies have shown use of age-of-onset data can increase the significance levels of linkage analysis, and hence the statistical power, of any joint method of analysis (5). One approach to studying age of onset has been to analyze it as a right-censored quantitative trait (6). This was done by extending the program Loki, which uses a general segregation/linkage Markov chain Monte Carlo Bayesian framework (7) to analyze a quantitative trait. Daw et al (6) suggested the location of linkage could be well estimated even though there may be appreciable bias in the estimated model parameters generated in this manner.

Adjustment for ascertainment has only been well understood for sibship studies or for cases of true single ascertainment. Elston (8) proposed a pedigree likelihood for segregation analysis that can allow for both ascertainment and age of onset. Allowance for single ascertainment has been incorporated into Loki (9). A very general likelihood approach to allow for ascertainment in general pedigrees has been formulated by Ginsburg et al. (10, 11), but this approach requires the true pedigree structures and the proband sampling frame (12) to be well defined, and full phenotypic information must be available on all members of the *sampled* pedigree who fall within the proband sampling frame. To resolve some of the assumptions of a model-based analysis, we have developed a segregation-linkage approach with correction for ascertainment by setting a prevalence constraint to determine the best fitting segregation model, and this paper illustrates its application, assuming a bivariate phenotype (affection status and age of onset) on a set of families collected to study the inheritance of glioma for which a model-free analysis has been previously performed (1). This dataset required us to allow for multiplex ascertainment (13, 14) based on the presence of a proband and an additional affected relative in the family, and then to allow for further selection of families genotyped for linkage analysis. To our knowledge, no joint segregation-linkage analysis with appropriate correction for multiplex ascertainment has been developed, though joint analyses have been successfully performed using Loki with an incorrect ascertainment model (6). In this paper, we develop an approximate method to adjust for multiplex ascertainment, in both segregation and linkage analysis, and illustrate its use for a trait, the occurrence of glioma, with variable age of onset. We justify its use with a simulation study, incidentally noting a problem that occurs when attempting to perform a Bayesian analysis on the same data without appropriate adjustment for ascertainment.

Materials and Methods

Data

The segregation analysis was performed on 6983 individuals in 281 pedigrees (Table 1), all ascertained from GLIOGENE study sites in the United States (15). Families were ascertained by the presence of a proband (i.e. an individual affected with a glioma) with a first- or second-degree relative also affected with glioma. Three pedigrees had loops which were all formed by two siblings in one nuclear family married to two siblings in another nuclear family. These loops were cut by assigning the siblings most distant to the segregating relatives as pedigree founders. Although the pedigree structure of all 6983 individuals was used, only those pedigree members whose affection status and age were known could enter the analysis: for affected persons, age was age at onset of glioma or age at examination; for unaffected persons, age was the age at which last known to be unaffected.

The linkage analysis dataset comprised a subset of 74 of these 281 pedigrees, chosen to be genotyped on the basis of their informativity for linkage (1), and both affected and unaffected persons were genotyped using the Illumina Human370 chip.

Fitting segregation models—The SEGREG program in the Statistical Analysis for Genetic Epidemiology (S.A.G.E.) version 6.1 package was used to fit diallelic monogenic models for a binary trait with variable age of onset, to define individual-specific age-dependent penetrance parameters to be used in multipoint linkage analysis. To find the best model to conduct linkage analysis, we fitted to the pedigree data two types of models that represent a mixture of two genotypic distributions: those in which there are two susceptibilities and a common age of onset distribution, and those with two age of onset distributions and a common susceptibility. Susceptibility is defined as the probability of disease if the individual lived to an infinite age and need not equal 1. Sex was included in the model as a covariate of either the logit of susceptibility or of the mean or variance of age of onset, so in all we fitted six distinct segregation models, each of which could result in dominant or recessive inheritance. The logistic density function was assumed for age of onset, but a Box-Cox (16) power transformation parameter (λ_1) was also simultaneously estimated, to allow for departure from this distributional form. Further details are given in the Supplemental Methods.

In addition to assuming single ascertainment (i.e. conditioning the likelihoods on the phenotypes of the probands) a prevalence constraint was included in the model. For this we assumed the population prevalence was on average 0.04%, with the prevalence for males being 1.5 times higher than that for females. Rather than fixing the disease prevalence, as was often typically done in early segregation analyses, we specified prevalence using two numbers as implemented in SEGREG - the number of affected individuals (R) in an independent sample of size (N) (see Supplemental Methods). For this analysis, these numbers were set to be 144 and 300,000 for males, and 96 and 300,000 for females. The lifetime prevalence R/N was taken to be the prevalence at 90 years old, and N=300,000; this was calculated from prevalence rates obtained from the Central Brain Tumor Registry of the United States registries (17).

For those relative pairs genotyped for linkage analysis, the recorded relationships were verified using genome-wide genotype data with the program RELTEST in S.A.G.E. Five MZ twin pairs were identified, and one out of each pair of the MZ twin pairs was excluded from both the segregation and the linkage analyses.

Three segregation models that best fit the data on the basis of Akaike's A Information Criterion (AIC) were selected for linkage analysis. We re-estimated the trait locus allele frequency among founders of the 74 families chosen for linkage analysis by re-maximizing the likelihoods of the three best-fitting segregation models, but fixing every parameter (other than the allele frequency at the trait locus) at the values estimated in the whole set of families. The rationale for doing this is that genotypic frequencies need to reflect those of the founders of the specific pedigrees used in the linkage analysis. No prevalence constraint was used when re-estimating allele frequencies, because that would have resulted in the population genotypic frequencies rather than the frequencies among founders of the linkage family subset. We thus assume selection of the families for linkage analysis, enriched by affected members, has only a minor effect on the penetrance parameters, but a major effect on the pedigree founder genotype frequencies. Larger likelihoods and smaller AIC values resulted after including single ascertainment in the model when re-estimating the trait locus allele frequency, so this was done.

Model-based multipoint linkage analysis

In the linkage subset, large pedigrees were trimmed to reduce the number of inheritance vector bits to 21 or less, as required for ease of computing, as follows:

A. Eliminated all linkage uninformative branches (e.g. where no DNA was available)

A.1 Trimmed off all the antecedent branches with no DNA.

A.2 Trimmed off all the descendant branches with no DNA.

A.3 Trimmed off all siblings with no DNA nor descendant branches.

B. Eliminated a few genotyped persons

B.1 In one pedigree we eliminated three genotyped unaffected siblings farthest away from the part of the pedigree segregating for the trait.

B.2 If, upon eliminating according to the above rules, the number of bits for a pedigree was still too large, the youngest unaffected genotyped offspring was eliminated. This resulted in eliminating an additional 11 unaffected genotyped offspring. In total, apart from monozygotic twins, we eliminated only 14, largely linkage-uninformative, genotyped offspring.

The SNPs used in the genome-wide multipoint linkage analysis were selected to have minor allele frequency (MAF) ≥ 0.3 to increase informativeness, and genetic distance between any two neighboring SNPs ≥ 0.4 cM to have a more accurate estimate of genetic (as opposed to physical) distance, while allowing for as many appropriate markers as possible for multipoint linkage across the genome. A total of 3404 markers were thus selected.

To analyze a linkage region discovered on chromosome 17, two sets of SNPs were used. The first set consisted of the 138 SNPs originally used by Shete et al (1), which were selected to have MAF ≥ 0.05 and pairwise linkage disequilibrium (LD) $r^2 \leq 0.004$. The second set included the SNPs in the first set after excluding those with intervals between consecutive SNPs < 0.2 cM, but adding in those with MAF > 0.3 and interval ≥ 0.2 cM. There were 173 SNPs in this set, where some SNPs were in strong LD as there was no selection of SNPs based on LD. Thus, the limitation in the second set was based on genetic distance and MAF rather than LD. SNPs were also excluded if they were more than 10 cM away from any position, because the assumption of no interference only applies up to a distance of ~10 cM; note within 10cM the Haldane and Kosambi map functions are almost identical (18).

The founder allele frequencies of SNPs were estimated by maximum likelihood with the program FREQ in S.A.G.E. Model-based multipoint linkage analysis was performed with the MLOD program, specifying the Kosambi map function to obtain recombination fractions between consecutive markers from the genetic distances in the deCODE map (19). We performed multipoint linkage analysis using the three best segregation models, but with the trait locus allele frequencies re-estimated in the linkage pedigrees. We assumed locus homogeneity across the 74 pedigrees, and multipoint LODs were estimated at each SNP and at every 2 cM.

Simulation study to investigate type I error and power

To study the performance of our approach, we conducted a small simulation study. To minimize computation time, we applied the model to nuclear family, rather than extended pedigree, structures. To approximate the amount of information in our pedigrees, we used 220 nuclear families each comprising 6 siblings and two parents.

On each dataset, we simulated two marker SNPs with two different values of LD between them, and one unobserved trait locus that was either linked or not linked to these two simulated SNPs. LD between the two SNPs was set as $r^2 = 0$, 0.4 and 0.8, and for each case we set the MAF at 0.1, 0.2, 0.3, 0.4 or 0.5 for both SNPs. The genetic distance between the

two SNPs was 0.2 cM and the unobserved trait locus was in linkage equilibrium with the two marker SNPs, 0.2 cM away from the closest of the two SNPs to simulate linkage. The penetrance functions that best fitted the segregation glioma dataset (model 1 in figure 1) were then applied to the trait locus genotypes. Because the penetrance function is age related, age was first assigned according to the age distribution in the glioma data, i.e., according to the distributions of mother's age, of the age difference between mother and her first child, between consecutive siblings, and between couples. For each affected individual, the age is age at exam, and the age of onset was assigned according to the mean difference between age of onset and age at examination in the glioma data. One affected offspring in each family was taken to be the proband, with probability assigned according to the glioma data. We simulated families until we had 100 datasets - of 220 nuclear families each - that satisfied the criterion of containing an offspring proband and at least two affected members. From these we selected those sibships (without parents) with at least two affected members to form the linkage data subsets. There were 60 to 94 sibships in each linkage dataset. We analyzed each of these 100 datasets using the same procedure used to analyze our glioma pedigrees. We assumed Hardy-Weinberg proportions for the trait locus and each marker locus.

We analyzed each of the 100 simulated segregation data sets using the same setting of the prevalence constraint as for the glioma data. Then we re-estimated the allele frequencies at the trait locus in the corresponding simulated linkage dataset. Type I error and power were respectively evaluated using the LOD thresholds 0.588 and 1.175, which correspond the p-values 0.05 and .01 for a single linkage test. The proportion of data sets with maximum LODs greater than those thresholds are reported as the type I error and power.

Results

Table 1 shows the general characteristics of the segregation and linkage pedigrees. All six segregation models using the 281 pedigrees showed an autosomal dominant model with a rare trait locus allele (allele frequency=0.00047). Three models that fit the data best (Table 2) on the basis of their AICs were subsequently used for the linkage analyses. These three models are: susceptibility dependent on genotype and mean age of onset dependent on sex (model 1); mean age of onset dependent on genotype, with that mean dependent on sex (model 2); and mean age of onset dependent on genotype, with susceptibility dependent on sex (model 3). In practical terms the three models are identical: in models 1 and 2 the susceptibilities for the AA and AB genotypes are virtually 1 (see Supplemental Methods) and Figure 1 shows the cumulative distribution of age of onset for males and females, respectively, for all three models shown in Table 2 (see Supplemental Methods). Up to age 100, the distributions of age of onset under the three models are very close, with males being more susceptible than females for AA and AB genotype carriers. Penetrance of the BB genotype is always virtually 0 up to 100 years old, for both males and females. The re-estimated trait allele frequency in the 74 linkage pedigrees was 0.13 for all three models.

All three models gave similar genome-wide multipoint linkage results (Supplemental Figure S1). The strongest evidence for linkage was identified on chromosome 17, with two peaks at the positions 72.3 cM and 87.3 cM from pter, the multipoint LODs being respectively 2.5 and 3.1 at these two positions (Figure 2). No strong linkage was found on any other chromosome region (Supplemental Figure S1). It should be noted, when a linkage analysis was performed using the segregation models shown in Table 2, i.e. without re-estimating the allele frequencies to reflect those of the families actually used for the linkage analysis, all multipoint LODs were negative, across the whole genome. When analyzing the region within 10 cM of each linkage peak on chromosome 17, the first set of SNPs yielded lower information content (20) than the second set, as expected. At the first linkage peak, where

the model-free analysis showed stronger linkage evidence, the second SNP set produced a maximum multipoint LOD 0.3 lower than the first SNP set. At the second peak, the two SNP sets resulted in similar maximum LODs.

Table 3 summarizes the findings from the simulation study. Power only considers maximum LODs within 2 cM of the trait locus. We initially used the same criterion for type I error, finding it to be inflated only when the MAF is 0.1, but the inflation increased when taking the maximum LOD at any position. However, for LOD > 1.175, the type I error is much better controlled, though perhaps increased for a small allele frequency. Note that otherwise the estimated type I error is never larger than that found for $r^2=0$.

Discussion

This study showed that by utilizing a segregation analysis procedure with a prevalence constraint, and then re-estimating the trait model allele frequencies appropriate for the actual linkage sample, a model-based multipoint linkage analysis is possible when single ascertainment was not followed. The simulation study, based on the particular model found for the glioma data, provides justification for the two-step procedure used here. The substantive findings for the dataset analyzed are similar to those of model-free linkage in the same data set (1), but yield stronger evidence for linkage at a second region in the same chromosome, 73.7 cM from pter. The observation that the LODs drop below 0 between these two regions suggests there may be two separate loci of interest on chromosome 17q. The best segregation models were consistent with autosomal dominant inheritance of a rare disease allele, consistent with a recessive cellular effect under Knudson's two-hit model, the second hit having a variable age of occurrence (21). The estimated trait locus allele frequency was 1 in 2000 in the population but 13% among the founders of the multiplex pedigrees selected for linkage, and the penetrance (i.e. probability of a heterozygous genotype becoming homozygous) by age 60 was ~20%. These segregation analysis results do not differ considerably from those used previously for homozygosity mapping in Northern Sweden (22). That study found autosomal recessive inheritance models gave better results for homozygosity mapping as compared to dominant models when assuming a higher population allele frequency (1 in 1000) and penetrance (40% in one model and 60% in the other), but without allowing for age of onset. Differences in penetrance between men and women in our analysis are, by assumption of the penetrance constraint, consistent with the known sex difference in population incidence of gliomas (17).

The cumulative age of onset distributions for all three best-fitting models were similar (up to 100 years old), and the model-based linkage results based on the three models were nearly the same, which argues for the reliability of this analysis approach and our results. In fact, a less precise prevalence constraint did not have a large effect on our segregation models: the prevalence function is nearly the same when assuming the same mean prevalence but with two quite different precisions (Figure 3).

Model-based linkage analysis including both affected and unaffected persons does not require the assumption of linkage equilibrium of the markers, unlike affected-only linkage analysis, because the likelihood function of phenotypes is conditional on markers rather than the other way around. That linkage equilibrium of the markers is an unnecessary assumption was also demonstrated by Xing et al. (2) for model-free linkage analysis when both affected and unaffected persons are included. When comparing the allele sharing with that expected under linkage equilibrium, which is the essence of affected-only model-free linkage analysis, there is a clear bias introduced by LD. However, if the bias in the allele sharing is similar for both affected and discordant pairs, the overall result is that the two biases cancel each other out when both phenotypes are included. In our study, because we included

unaffected relatives, bias would only occur as a result of mis-specifying the ascertainment of families (which led to elimination of unaffected persons, but was corrected for using the prevalence constraint) or by ignoring residual correlations among family members (which we checked by including a polygenic component in the model used for analysis and finding it to be not significant).

All current approaches to linkage analysis make the assumption of accurate specification of recombination fractions between markers, so using more SNPs in the linkage analysis could potentially provide even more linkage information (provided the genetic intervals between consecutive SNPs are accurate). Use of additional informative SNPs with intervals ≥ 0.2 cM resulted in lower multipoint linkage at the first peak, whether these markers were in LD or not. It is important to note that there would have been absolutely no evidence for linkage had we not re-estimated the trait allele frequencies in the subset of families used for the linkage analysis. Our simulation study shows the validity and efficiency of this two-step analysis.

We calculated family specific multipoint LOD scores across the region on chromosome 17 and found that 30 families contributed positive LODs to both peaks, 13 to the first peak, and 9 to the second peak. The largest family specific multipoint LOD under a peak was 0.59, under the first peak. That the family-specific LODs are small is not surprising, given the low penetrance of glioma – < 0.2 at the average age of 54 (see Figure 1). Therefore we did not calculate heterogeneity LODs, though this would be the next step if there had been higher penetrance, and hence more linkage-informative pedigrees.

We also analyzed our age of onset data on chromosome 17 using the multipoint linkage package Loki, where age of onset for the unaffected is assumed right censored and a posterior distribution is obtained for all unknown parameters. The form of the model assumed is similar to our model 2, but with 6 sex-dependent normal age-of-onset distributions, two for each genotype, rather than 4 sex-dependent logistic age-of-onset distributions after power transformation (assuming dominance) (Supplemental Methods). Loki identified four possible linkage locations on chromosome 17 (Supplementary Figure S2), including the two found by our method but shifted slightly, with more evidence for linkage at 87-89 cM from pter (further details are given in the Supplemental Methods and Supplementary Figures S2, S3 and S4). But by far the highest peak, expressed as a Bayes' factor - the posterior probability divided by the prior probability, was found at 2.5 cM from pter on chromosome 17. The estimated model at all three peaks was one of over-dominance, which simulation studies have suggested could be due to not allowing for ascertainment, though estimation of the linkage location does not seem to be affected (23). Because neither our model-based analysis nor the previous model-free analysis found any peak at this location, and because there is evidence that the Monte Carlo Markov Chain sampler was not mixing well at that location (Supplementary Figure S3), this new linkage peak could well be a type I error. With hindsight we repeated all the Loki analyses disallowing overdominance, but then no linkage signals were found on chromosome 17 (Supplementary Figure S4).

Our study provides an approach for linkage analysis for a bivariate trait (comprising a binary disease affection status and a censored quantitative age of onset) when there is multiplex ascertainment. Recent advances in genome-wide sequencing often reveal thousands of low penetrance, low frequency sequence variants. Hence, it can be challenging to filter out true deleterious variants from those that are benign. Linkage methods can both help decide true genomic areas of interest and screen families that will be most informative for sequencing. In our two-step analysis, we fitted segregation models for both disease affection status and age of onset using the whole sample, while we adjusted the likelihood for ascertainment (together with a correction for single ascertainment) by incorporating a prevalence constraint to obtain estimates of the penetrance parameters, and then we re-estimated the trait allele

frequencies that correspond to those of the founders of the linkage pedigrees. Therefore, this method provides a practical solution to model-based linkage analysis for disease affection status with variable age of onset for the kinds of pedigree data often collected for linkage analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant Support

This work was supported by grants from the NIH, Bethesda, Maryland (5R01 CA119215, 5R01 CA070917, R01CA52689, P50097257, R01CA126831, 5P30CA16672, P30 CA043703) and by a National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004) This publication was also made possible by the Case Western Reserve University/Cleveland Clinic CTSA Grant Number UL1 RR024989 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Additional support was provided by the American Brain Tumor Association, The National Brain Tumor Society, and the Tug McGraw Foundation. For more information about the Gliogene Consortium, refer to the following Web site: http://www.gliogene.org.

Appendix

Author Contributions:

Accrual of Families: Amos, Armstrong, Barnholtz-Sloan, Bernstein, Bondy Claus, Davis, Houslton, Il'yasova, Jenkins, Johansen, Lai, Liu, McCarthy, Melin Olson, Sadetzki, Schildkraut, Vick, Wrensch, Yang.

Data procurement and management: Armstrong, Amos, Bondy, Barnholtz-Sloan, Chen, Elston, Liu, Shete, Sun, Vengoechea.

Data Analysis: Barnholtz-Sloan, Chen, Elston, Sun, Vengoechea.

Writing of first draft of manuscript: Armstrong, Bondy, Barnholtz-Sloan, Chen, Elston, Shete, Sun, Vengoechea.

Final editing, review and approval of manuscript: All authors

References

1. Shete S, Lau CC, Houlston RS, Claus EB, Barnholtz-Sloan J, Lai R, et al. Genome-wide high-density SNP linkage search for glioma susceptibility loci: results from the Gliogene Consortium. Cancer Res. 2011; 71:7568–7575. [PubMed: 22037877]
2. Xing C, Sinha R, Xing G, Lu Q, Elston RC. The affected/discordant-sib-pair design can guarantee validity of multipoint model-free linkage analysis of incomplete pedigrees when there is marker-marker disequilibrium. Am J Hum Genet. 2006; 79:396–401. [PubMed: 16826532]
3. Jenkins MA, Baglietto L, Dite GS, Jolley DJ, Southey MC, Whitty J, et al. After hMSH2 and hMLH1– what next? Analysis of three-generational, population based, early-onset colorectal cancer families. Int J Cancer. 2002; 102:166–171. [PubMed: 12385013]
4. Cicek MS, Cunningham JM, Fridley BL, Serie DJ, Bamlet WR, Diergaarde B, et al. Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. PloS one. 2012; 7:e38175. [PubMed: 22675446]

5. Lin PI, McInnis MG, Potash JB, Willour VL, Mackinnon DF, Miao K, et al. Assessment of the effect of age at onset on linkage to bipolar disorder: evidence on chromosomes 18p and 21q. Am J Hum Genet. 2005; 77:545–555. [PubMed: 16175501]
6. Daw EW, Heath SC, Wijsman EM. Multipoint oligogenic analysis of age-of-onset data with applications to Alzheimer Disease pedigrees. Am J Hum Genet. 1999; 64:839–851. [PubMed: 10053019]
7. Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet. 1997; 61:748–760. [PubMed: 9326339]
8. Elston RC. Ascertainment and age of onset in pedigree analysis. Hum Hered. 1973; 23:105–12. [PubMed: 4756850]
9. Ma J, Amos CI, Daw EW. Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait. Genet Epidemiol. 2007; 31:594–604. [PubMed: 17487893]
10. Ginsburg E, Malkin I, Elston RC. Sampling correction in pedigree analysis. Stat Appl Genet Mol Biol. 2003; 2:1–22. Article2.
11. Ginsburg, E.; Malkin, I.; Elston, RC. Theoretical aspects of pedigree analysis. Ramot Publishing House; Tel-Aviv: 2006. p. 31-55.
12. Elston RC, Sobel E. Sample consideration in the gathering and analysis of pedigree data. Am J Hum Genet. 1979; 31:62–69. [PubMed: 373427]
13. Tiwari JL, Betuel H, Gebuhrer L, Morton NE. Genetic epidemiology of coeliac disease. Genet Epidemiol. 1984; 1:37–42. [PubMed: 6336186]
14. Bonney GE, Elston RC, Correa P, Haenszel W, Zavala DE, Zarama G, et al. Genetic etiology of gastric carcinoma: I. Chronic atrophic gastritis. Genet Epidemiol. 1986; 3:213–24. [PubMed: 3744019]
15. Malmer B, Adatto P, Armstrong G, et al. GLIOGENE—an international consortium to understand familialgGlioma. Cancer Epidemiol Biomarkers Prev. 2007; 16:1730–1734. [PubMed: 17855690]
16. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Series B Stat Methodol. 1964; 26:211–252.
17. CBTRUS. Statistical Report: Primary brain tumors in the United States, 2004-2009. Central Brain Tumor Registry of the United States; Chicago: 2012.
18. Ott, J. Analysis of Human Genetic Linkage. third edition. Johns Hopkins Univ Press; Baltimore: 1999. p. 20
19. Kong A, Gudbjartsson DF, Sainz J, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31:241–247. [PubMed: 12053178]
20. Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet. 1995; 57:439–454. [PubMed: 7668271]
21. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971; 68:820–823. [PubMed: 5279523]
22. Malmer B, Haraldsson S, Einarsdottir E, Lindgren P, Holmberg D. Homozygosity mapping of familial glioma in Northern Sweden. Acta Oncol. 2005; 44:114–119. [PubMed: 15788289]
23. Ma J, Amos CI. Ascertainment Bias for Markov Chain Monte Carlo Segregation and Linkage Analysis of Age-at-onset data. Genet Epidemiol. 2009; 33:801.

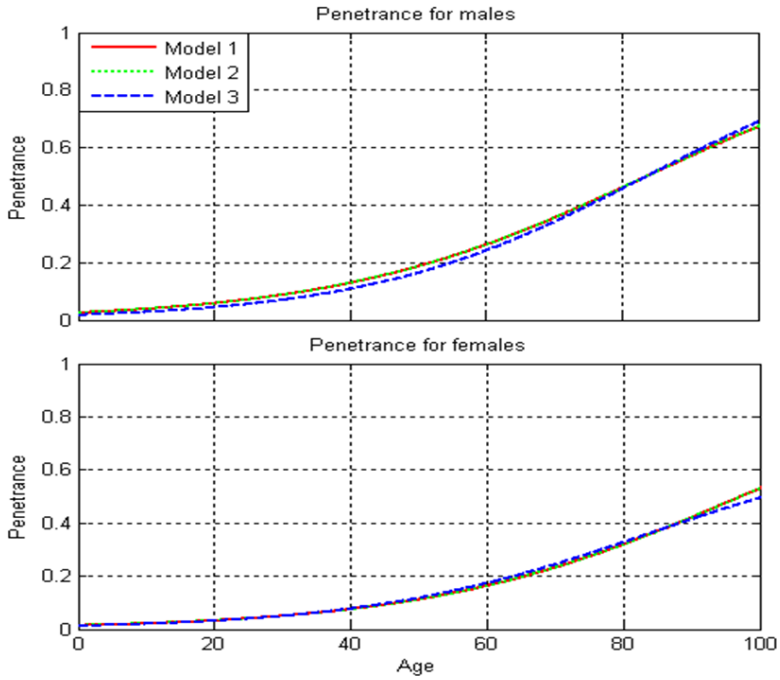

Figure 1.
Cumulative age of onset distribution for genotypes AA and AB as estimated from the segregation models. The upper plot is for males, the lower one is for females.

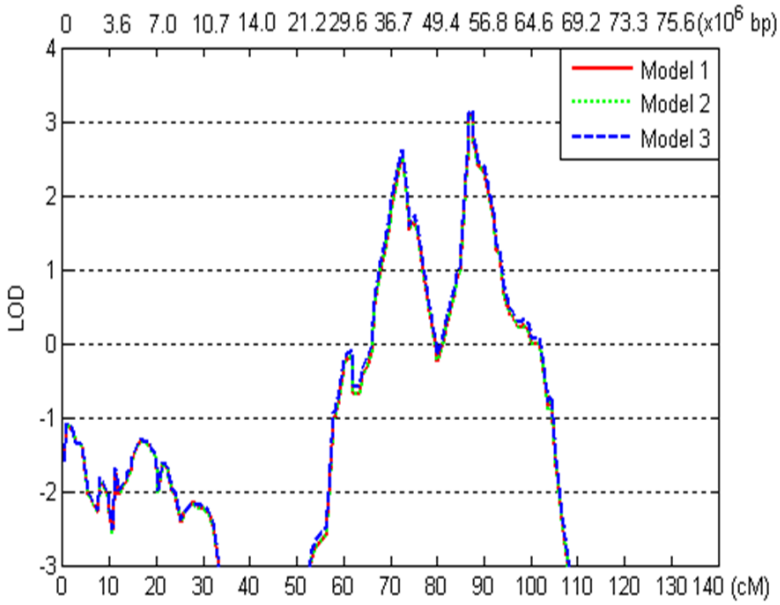

Figure 2.

Model-based multipoint linkage LODs on chromosome 17 using the three trait locus segregation models. Multipoint LOD scores have been truncated at -3, some scores are below -3.

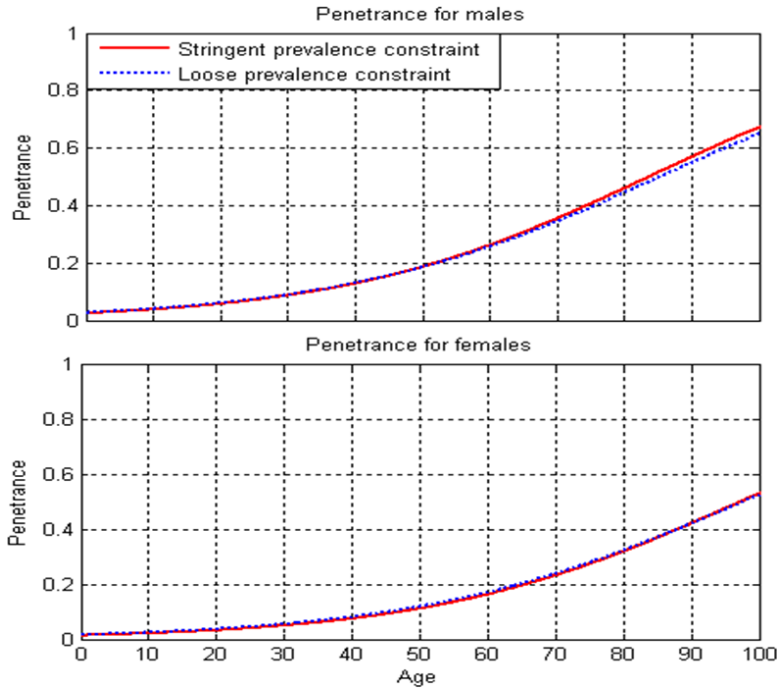

Figure 3.

Cumulative age of onset distribution taking into account the prevalence constraint (for genotypes AA and AB) as estimated from the segregation models. Averaging over the two sex groups, R= 240 and N= 600,000 for the precise prevalence constraint and R=2.4 and N=6,000 for the less precise prevalence constraint. The upper plot is for males, the lower one is for females. Note that R and N do not need to be integers.

Table 1

Overall characteristics of the glioma segregation and linkage data sets.

	Segregation data				Linkage data		
Number of pedigrees				281			74
2 and 3 generations				35			28
4 generations				192			43
5 and 6 generations				54			3
Average size of pedigrees±s.d.				24.88±9.93			15.07±4.37
Number of individuals	All	Male	Female	Unknown	All	Male	Female
Affected	633	335	298	0	170	88	82
Unaffected	3561	1743	1818	0	727	338	389
Unknown*	2789	1445	1334	10	218	123	95
Total	6983	3523	3450	10	1115	549	566
Proportion of affected	0.091	0.095	0.087	0	0.152	0.160	0.145
Average age±s.d.	56.19±21.41	55.01±20.99	57.35±21.76	-	54.36±20.38	53.22±19.80	55.39±20.88
Average age of onset±s.d.	49.39±19.02	49.28±17.79	49.52±20.33	-	47.51±17.98	48.60±16.31	46.33±19.65

*not fully informative, unknown for affection status or age

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2013 December 01.

Table 2

Segregation model parameter estimates ± standard errors for the 281 glioma pedigrees using SEGREG

Model 1		Model 2		Model 3	
$\mu_{AA}=\mu_{AB}=\mu_{BB}$	90.38±2.38	μ_{AA}	90.36±1.41	μ_{AA}	83.61±2.29
β_{sex}	13.67±3.01	μ_{AB}	90.36±1.41	μ_{AB}	83.61±2.29
σ^2	895.81±79.22	μ_{BB}	205614170.7±INF	μ_{BB}	10603.79± INF
θ_{AA}	26.12±INF	μ_{sex}	13.671490±0.000004	σ^2	838.71±65.08
θ_{AB}	26.12±INF	σ^2	895.2753±0.0003	$\theta_{AA}=\theta_{AB}=\theta_{BB}$	8.83±0.89
θ_{BB}	-60.36± INF	$\theta_{AA}=\theta_{AB}=\theta_{BB}$	424.29±INF	β_{sex}	-15.59±1.76
λ_1	0.47±0.08	λ_1	0.4711195±0.0000002	λ_1	0.52±0.08
q_A	0.00047±0.00004	q_A	0.00047±0.0000	q_A	0.00047±0.00004
-2ln(L)	10077.9	-2ln(L)	10077.9	-2ln(L)	10080.8
Akaike's AIC	10091.9	Akaike's AIC	10091.9	Akaike's AIC	10094.8

μ_{AA}, μ_{AB}, μ_{BB} are median unbiased estimates of the mean ages of onset for genotypes AA, AB and BB, respectively

σ^2 is the variance of age of onset on the transformed scale

θ_{AA}, θ_{AB}, θ_{BB} are the logits of susceptibility for genotypes AA, AB and BB, respectively

β_{sex} is the effect of sex on mean age of onset for model 1 and model 2, the effect of sex on the logit of susceptibility for model 3

λ_1 is the power parameter in the Box-Cox transformation, the shift parameter λ_2 is fixed at 0

q_A is the allele frequency for allele A at the trait locus

±INF indicates that the likelihood is flat and it is not possible to estimate a standard error.

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC 2013 December 01.

Table 3

Type I error and power of multi-point linkage analysis using simulation data, evaluated by the LOD thresholds 0.588 and 1.175, which correspond to the p values 0.05 and 0.01 for a single linkage test

	MAF of the two SNPs	LOD > 0.588[a]			LOD > 0.588[b]			LOD > 1.175[b]		
		$r^2 = 0$	$r^2 = 0.4$	$r^2 = 0.8$	$r^2 = 0$	$r^2 = 0.4$	$r^2 = 0.8$	$r^2 = 0$	$r^2 = 0.4$	$r^2 = 0.8$
Type I error	0.1	0.01	0.06	0.08	0.07	0.09	0.08	0.01	0	0.03
	0.2	0.03	0.01	0.01	0.06	0.04	0.03	0	0	0
	0.3	0	0.02	0.01	0.10	0.06	0.07	0	0	0
	0.4	0.01	0.02	0.03	0.10	0.11	0.09	0	0.02	0
	0.5	0.01	0.01	0.03	0.10	0.06	0.07	0.02	0.01	0.01
Power	0.1	0.96	0.91	0.86	0.96	0.91	0.86	0.90	0.83	0.75
	0.2	0.99	1	0.99	0.99	1	0.99	0.99	0.97	0.92
	0.3	1	1	0.97	1	1	0.97	1	0.99	0.95
	0.4	1	1	0.99	1	1	0.99	1	1	0.97
	0.5	1	1	1	1	1	1	1	1	0.99

[a] evaluated within 2 cM of the trait locus

[b] type I error evaluated anywhere over the genome, power evaluated within 2 cM of the trait locus