Research Article

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

Xianging Sun1, Jaime Vengoechea2, Robert Elston1,3, Yanwen Chen1, Christopher I. Amos4, Georgina Armstrong5, Jonine L. Bernstein7, Elizabeth Clauss9,10, Faith Davis11, Richard S. Houlston12, Dora Il’yasova13, Robert B. Jenkins14, Christoffer Johansen15, Rose Lai8, Ching C. Lau6, Yanhong Liu6, Bridget J. McCarthy11, Sara H. Olson7, Siegal Sadetzki16, Joellen Schildkraut13, Sanjay Shete5, Robert Yu5, Nicholas A. Vick17, Ryan Merrell17, Margaret Wrensch18, Ping Yang14, Beatrice Melin19, Melissa L. Bondy6, and Jill S. Barnholtz-Sloan1,3, on behalf of the Gliogene Consortium

Abstract

Background: We propose a 2-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 reestimated 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 2 separate peaks on chromosome 17; the first agreed with a region identified by Shete and colleagues 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: Our approach was able to narrow the linkage peak previously published for glioma.

Impact: We provide 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. Cancer Epidemiol Biomarkers Prev; 21(12): 2242–51. ©2012 AACR.

Introduction

Successful linkage analysis of complex diseases, when carried out to obtain log of the odds (LODs), requires a number of assumptions related to both the markers and the trait of interest (See Supplementary Table S1). Both model-based and model-free linkage analysis, where

Authors’ Affiliations: Departments of 1Epidemiology and Biostatistics, School of Medicine; Internal Medicine and Genetics; 2Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio; Departments of 3Genetics, 4Biostatistics, The University of Texas MD Anderson Cancer Center; 5Department of Pediatrics, Division of Hematology-Oncology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas; 6Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center; 7The Neurological Institute of Columbia University, New York; 8Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut; 9Department of Neurosurgery, Brigham and Women’s Hospital, Boston, Massachusetts; 10Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, Illinois; 11Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, United Kingdom; 12Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina; 13Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota; 14Department of Neurology; Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark; 15Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel; 16Tel-Aviv Sackler School of Medicine, Tel-Aviv University, Tel-Aviv, Israel; 17Evanston Kellogg Cancer Care Center NorthShore University HealthSystem, Evanston, Illinois; 18Department of Neurological Surgery, University of California, San Francisco, San Francisco, California; and 19Department of Radiation Sciences Oncology, Umea University, Umea, Sweden

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

X. Sun and J. Vengoechea contributed equally to this work.

Corresponding Author: Jill Barnholtz-Sloan, Case Comprehensive Cancer Center, 11100 Euclid Ave, Wean 152, Cleveland, OH 44106. Phone: 216-368-1506; Fax: 216-368-2606; E-mail: jsb42@case.edu
doi: 10.1158/1055-9965.EPI-12-0703

©2012 American Association for Cancer Research.

Downloaded from cebp.aacrjournals.org on April 7, 2017. © 2012 American Association for Cancer Research.
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 2 types of linkage analysis, once certain model parameters are assumed to be known, is the direction of the approach: typically, model-free linkage as performed by Shete et al., (1) 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; 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 that the 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. Daw and colleagues (6) suggested the use of a general segregation/linkage Markov chain Monte Carlo Bayesian framework (7) to analyze a quantitative trait. Daw and colleagues (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 and colleagues (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 article 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 carried out (1). This data set 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 carried out using Loki with an incorrect ascertainment model (6). In this article, 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 carry out a Bayesian analysis on the same data without appropriate adjustment for ascertainment.

Materials and Methods

Data

The segregation analysis was conducted on 6,983 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 2 siblings in 1 nuclear family married to 2 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 6,983 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 data set 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 2 types of models that represent a mixture of 2 genotypic distributions: those in which there are 2 susceptibilities and a common age of onset distribution, and those with 2 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 log of susceptibility or of the mean or variance of age of onset, so in all we fitted 6 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 ($\lambda_1$) was also simultaneously estimated, to allow for departure from this distributional form. Further details are given in the Supplementary 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 2 numbers as implemented in SEGREG—the number of affected individuals ($N$) in an independent sample of size ($N$) (see Supplementary 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 reestimated the trait locus allele frequency among founders of the 74 families chosen for linkage analysis by remaximizing the likelihoods of the 3 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 reestimating 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 reestimating 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:

### Table 1. Overall characteristics of the glioma segregation and linkage data sets

<table>
<thead>
<tr>
<th></th>
<th>Segregation data</th>
<th>Linkage data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of pedigrees</strong></td>
<td>281</td>
<td>74</td>
</tr>
<tr>
<td>2 and 3 generations</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>4 generations</td>
<td>192</td>
<td>43</td>
</tr>
<tr>
<td>5 and 6 generations</td>
<td>54</td>
<td>3</td>
</tr>
<tr>
<td><strong>Average size of pedigrees ± SD</strong></td>
<td>24.88 ± 9.93</td>
<td>15.07 ± 4.37</td>
</tr>
<tr>
<td><strong>Number of individuals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>633</td>
<td>170</td>
</tr>
<tr>
<td>Male</td>
<td>335</td>
<td>88</td>
</tr>
<tr>
<td>Female</td>
<td>298</td>
<td>82</td>
</tr>
<tr>
<td>Unknown</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Proportion of affected</strong></td>
<td>0.091</td>
<td>0.152</td>
</tr>
<tr>
<td>0.095</td>
<td>0.087</td>
<td>0.160</td>
</tr>
<tr>
<td><strong>Average age ± SD</strong></td>
<td>56.19 ± 21.41</td>
<td>54.36 ± 20.38</td>
</tr>
<tr>
<td>55.01 ± 20.99</td>
<td>53.22 ± 19.80</td>
<td></td>
</tr>
<tr>
<td>57.35 ± 21.76</td>
<td>55.39 ± 20.88</td>
<td></td>
</tr>
<tr>
<td><strong>Average age of onset ± SD</strong></td>
<td>49.39 ± 19.02</td>
<td>47.51 ± 17.98</td>
</tr>
<tr>
<td>49.28 ± 17.79</td>
<td>48.60 ± 16.31</td>
<td></td>
</tr>
<tr>
<td>49.52 ± 20.33</td>
<td>46.33 ± 19.65</td>
<td></td>
</tr>
</tbody>
</table>

*(Not fully informative, unknown for affection status or age.*
LODs were estimated at each SNP and at every 2 cM. Homogeneity across the 74 pedigrees, and multipoint models, but with the trait locus allele frequencies reestimated between consecutive markers from the genetic Kosambi map function to obtain recombination fractions. To approximate the amount of information in our pedigrees, we used 220 nuclear families each comprising 6 siblings and 2 parents.

On each data set, we simulated 2 marker SNPs with 2 different values of LD between them, and 1 unobserved trait locus that was either linked or not linked to these 2 simulated SNPs. LD between the 2 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 2 SNPs was 0.2 cM and the unobserved trait locus was in linkage equilibrium with the 2 marker SNPs, 0.2 cM away from the closest of the 2 SNPs to simulate linkage. The penetrance functions that best fitted the segregation glioma data set (model 1 in Fig. 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, that is, 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 examination; 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 data sets—of 220 nuclear families each—that satisfied the criterion of containing an offspring proband and at least 2 affected members. From these, we selected those sibships (without parents) with at least 2 affected members to form the linkage data subsets. There were 60 to 94 sibships in each linkage data set. We analyzed each of these 100 data sets 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 reestimated the allele frequencies at the trait locus in the corresponding simulated linkage data set. Type I error and power were respectively evaluated using the LOD thresholds 0.588 and 1.175, which correspond to the $P$ values 0.05 and 0.01 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.

**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 2 parents.

On each data set, we simulated 2 marker SNPs with 2 different values of LD between them, and 1 unobserved trait locus that was either linked or not linked to these 2 simulated SNPs. LD between the 2 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 2 SNPs was 0.2 cM and the unobserved trait locus was in linkage equilibrium with the 2 marker SNPs, 0.2 cM away from the closest of the 2 SNPs to simulate linkage. The penetrance functions that best fitted the segregation glioma data set (model 1 in Fig. 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, that is, 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 examination; 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 data sets—of 220 nuclear families each—that satisfied the criterion of containing an offspring proband and at least 2 affected members. From these, we selected those sibships (without parents) with at least 2 affected members to form the linkage data subsets. There were 60 to 94 sibships in each linkage data set. We analyzed each of these 100 data sets 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 reestimated the allele frequencies at the trait locus in the corresponding simulated linkage data set. Type I error and power were respectively evaluated using the LOD thresholds 0.588 and 1.175, which correspond to the $P$ values 0.05 and 0.01 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 6 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 3 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 3 models are identical: in models 1 and 2 the susceptibilities for the AA and AB genotypes are virtually 1 (see Supplementary Methods) and Fig. 1 shows the cumulative distribution of age of onset for males and females, respectively, for all 3 models shown in Table 2 (see Supplementary Methods). Up to age 100, the distributions of age of onset under the 3 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 reestimated trait allele frequency in the 74 linkage pedigrees was 0.13 for all 3 models.

All 3 models gave similar genome-wide multipoint linkage results (Supplementary Fig. S1). The strongest evidence for linkage was identified on chromosome 17, with 2 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 2 positions (Fig. 2). No strong linkage was found on any other chromosome region (Supplementary Fig. S1). It should be noted, when a linkage analysis was conducted using the segregation models shown in Table 2, that is, without reestimating 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 2 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 $r^2 = 0$. 

![Figure 1. Cumulative age of onset distribution for genotypes AA and AB as estimated from the segregation models. The top plot is for males, the lower one is for females.](image-url)
Discussion

This study showed that by using a segregation analysis procedure with a prevalence constraint, and then reestimating 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 2-step procedure used here. The substantive findings for the data set 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, 87.3 cM from pter. The observation that the LODs drop below 0 between these 2 regions suggests there may be 2 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 2-hit model, the second hit having a variable age of occurrence (21). The estimated trait locus allele frequencies were consistent with those found in model-free linkage, but yield stronger evidence for linkage at a second region in the same chromosome.

Table 2. Segregation model parameter estimates ± SEs for the 281 glioma pedigrees using SEGREG

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{AA} = \mu_{AB} = \mu_{BB}$</td>
<td>$\mu_{AA}^a$</td>
<td>$\mu_{AA}^a$</td>
</tr>
<tr>
<td>$\sigma^2 = 13.67 ± 3.01$</td>
<td>$\mu_{AB}^a$</td>
<td>$\mu_{AB}^a$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\mu_{BB}^a$</td>
<td>$\mu_{BB}^a$</td>
</tr>
<tr>
<td>$\mu_{AA}^a$</td>
<td>$\mu_{BB}^a$</td>
<td>$\mu_{BB}^a$</td>
</tr>
<tr>
<td>$\mu_{AB}^a$</td>
<td>$\mu_{AB}^a$</td>
<td>$\mu_{AB}^a$</td>
</tr>
<tr>
<td>$\mu_{BB}^a$</td>
<td>$\mu_{BB}^a$</td>
<td>$\mu_{BB}^a$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
<tr>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
<td>$\sigma^2 = 26.12 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
<td>$\sigma^2 = 8.83 ± INF$</td>
</tr>
<tr>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
<td>$\sigma^2 = 895.81 ± 79.22$</td>
</tr>
</tbody>
</table>

NOTE: ± INF indicates that the likelihood is flat and it is not possible to estimate a SE

- $\mu_{AA}$, $\mu_{AB}$, $\mu_{BB}$ are median unbiased estimates of the mean ages of onset for genotypes AA, AB, and BB, respectively
- $\sigma^2$ is the variance of age of onset on the transformed scale
- $\beta_{AA}$, $\beta_{AB}$, $\beta_{BB}$ are the logits of susceptibility for genotypes AA, AB, and BB, respectively
- $\lambda^a$ 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
- $\lambda^b$ is the power parameter in the Box–Cox transformation, the shift parameter $\lambda^c$ is fixed at 0
- $\lambda^d$ is the allele parameter for allele A at the trait locus

Figure 2. Model-based multipoint linkage LODs on chromosome 17 using the 3 trait locus segregation models. Multipoint LOD scores have been truncated at −3, some scores are below −3.
frequency was 1 in 2,000 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 approximately 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 with dominant models when assuming a higher population allele frequency (1 in 1,000) 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 3 best-fitting models were similar (up to 100 years old), and the model-based linkage results based on the 3 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, 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, which argues for the reliability of this analysis approach and our results.

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 shown by Xing and colleagues (2) for model-free linkage analysis when both affected and discordant pairs, the overall result is that the 2 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 misspecifying 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 reestimated the trait allele frequencies in the subset of families used for the linkage analysis. Our simulation study shows the validity and efficiency of this 2-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 Fig. 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.

### Table 3. Type I error and power of multipoint 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

<table>
<thead>
<tr>
<th>MAF of the 2 SNPs</th>
<th>LOD &gt; 0.588&lt;sup&gt;a&lt;/sup&gt;</th>
<th>LOD &gt; 0.588&lt;sup&gt;b&lt;/sup&gt;</th>
<th>LOD &gt; 1.175&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r² = 0</td>
<td>r² = 0.4</td>
<td>r² = 0.8</td>
</tr>
<tr>
<td>Type I error</td>
<td>0.01</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>0.2</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>0.4</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>0.5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Power</td>
<td>0.96</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td>0.2</td>
<td>0.99</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1</td>
<td>0.99</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Evaluated within 2 cM of the trait locus.

<sup>b</sup>Type I error evaluated anywhere over the genome, power evaluated within 2 cM of the trait locus.
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, 2 for each genotype, rather than 4 sex-dependent logistic age-of-onset distributions after power transformation (assuming dominance) (Supplementary Methods). Loki identified 4 possible linkage locations on chromosome 17 (Supplementary Fig. S2), including the 2 found by our method but shifted slightly, with more evidence for linkage at 87 to 89 cM from pter (further details are given in the Supplementary 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 (Supplementary Figures S2 and S4). The estimated model at all 3 peaks was one of overdominance, 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 Fig. 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.

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 2-step analysis, we fitted segregation models for both disease affection status and age of onset using the whole sample, whereas 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 reestimated the trait allele

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 2 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.
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.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors' Contributions
Conception and design: J. Vengeoechea, J.L. Bernstein, R.B. Jenkins, C. Johansen, P. Yang, B. Melin, M.L. Bondy, J.S. Barnholtz-Sloan
Development of methodology: X. Sun, J. Vengeoechea, R. Elston, B. Melin, M.L. Bondy, J.S. Barnholtz-Sloan
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C.I. Amos, J.L. Bernstein, E.B. Claus, F. Davis
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): X. Sun, R. Elston, Y. Chen, C.I. Amos, J.S. Barnholtz-Sloan, R.B. Claus, R.B. Jenkins, Y. Liu, S.S. Shete, P. Yang, B. Melin, M.L. Bondy, J.S. Barnholtz-Sloan
Review, writing, and/or revision of the manuscript (i.e., reporting or organizing data, constructing databases): J. Vengeoechea, C.I. Amos, G. Armstrong, C. Johansen, L. Lai, B. Melin, Y. Liu, M.L. Bondy, J.S. Barnholtz-Sloan
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J. Vengeoechea, C.I. Amos, G. Armstrong, C. Johansen, L. Lai, B. Melin, Y. Liu, M.L. Bondy, J.S. Barnholtz-Sloan

Acknowledgments
The authors thank 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 Rynearson; Department of Neuro-Oncology, 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 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 (Jeffrey Weinberg); Department of Pediatrics, Section of Hematology and Oncology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas (Ching C. Lai, Ena Goud Horchita Leng); Department of Surgery, Rigshospitalet, University Copenhagen (Helle Broholm); Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York (Joline L. Bernstein, Sara H. Olson, Erica Schubert); Department of Neurology, Memorial Sloan-Kettering Cancer Center, New York (Lisa DeAngelis); Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota (Robert B. Jenkins, Ping Yang, Amanda Rynesort); Department of Radiation Sciences Oncology, Umeå University School of Medicine, Umeå, Sweden (Beatrice S. Melin, Roger Henriksson, Ulrika Andersson); Department of Medical Biociences, Umeå University, Umeå, Sweden (Thomas Brannstrom); Yale School of Medicine, Yale University School Health System, Evanston, Illinois (Nicholas A. Vick); Departments of Neurological Surgery and Epidemiology and Biostatistics (Margaret Wrensch, John Wienecke, Joe Wiemels, Lucie McCoy) Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, Illinois (Bridge J. McCarthy, Faith G. Davis).

The authors also thank the input of the Gliogene External Advisory Committee: Dr. Ake Borg (Department of Oncology, Lund University, Sweden), Dr. Stephen K Chanock (National Cancer Institute, U. S. National Institutes of Health), Dr. Peter Collins (University of Cambridge, U. K.), Dr. Richard Houlston (University of Manchester, U. K.), Dr. Robert Elston (Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio), Dr. Kenneth E. Offit (Children's Hospital of Philadelphia, Philadelphia, Pennsylvania), Dr. Cary N Engl (Boston, Massachusetts), Dr. Michael Silverman (Memorial Sloan-Kettering Cancer Center, New York, New York), Dr. Fred Saad (University of Toronto, Canada), Dr. Clemens Schulte (German Cancer Research Center, Heidelberg, Germany), Dr. Keisuke Ohgaki (National Cancer Center Research Institute, Tokyo, Japan). Finally, the authors thank the patients and their families for participating in this research.

Grant Support
This work was supported by grants from the NIH (5R01 CA119215, 5R01 CA170971, R01CA52689, P50CA097257, R01CA126831, 5P30CA16672, P01CA543703) and by a National Research Foundation of Korea Grant (2009-0065011). The funding for the Korean component (NRF-2011–220-C00004) of 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 (NIH). 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

Sun et al.

2250
Cancer Epidemiol Biomarkers Prev; 21(12) December 2012 Cancer Epidemiology, Biomarkers & Prevention

Copyright © 2012 American Association for Cancer Research. cebp.aacrjournals.org Downloaded from on April 7, 2017. © 2012 American Association for Cancer Research.
National Brain Tumor Society, and the Tug McGraw Foundation. For more information about the Gliogene Consortium, refer to the following Website: http://www.gliogene.org.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 15, 2012; revised July 31, 2012; accepted August 23, 2012; published OnlineFirst September 7, 2012.

References

Cancer Epidemiology, Biomarkers & Prevention

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

Xiangqing Sun, Jaime Vengoechea, Robert Elston, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-12-0703

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2012/09/10/1055-9965.EPI-12-0703.DC1

Cited articles
This article cites 19 articles, 3 of which you can access for free at:
http://cebp.aacrjournals.org/content/21/12/2242.full.html#ref-list-1

Citing articles
This article has been cited by 3 HighWire-hosted articles. Access the articles at:
/content/21/12/2242.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.