Some Capabilities for Model-Based and Model-Free Linkage Analysis using the Program Package S.A.G.E. (Statistical Analysis for Genetic Epidemiology)

A.H. Schnell X. Sun R.P. Igo, Jr. R.C. Elston

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

Key Words

Age of onset • Best linear unbiased predictor • Haseman-Elston • Individual-specific penetrance values • Multipoint • Power transform • Segregation model for a continuous trait

Abstract

For both model-free and model-based linkage analysis the S.A.G.E. (Statistical Analysis for Genetic Epidemiology) program package has some unique capabilities in analyzing both continuous traits and binary traits with variable age of onset. Here we highlight model-based linkage analysis of a quantitative trait (plasma dopamine β hydroxylase) that is known to be largely determined by monogenic inheritance, using a prior segregation analysis to produce the best fitting model for the trait. For a binary trait with variable age of onset (schizophrenia), we illustrate how using age of onset information to obtain a quantitative susceptibility trait leads to more statistically significant linkage signals, suggesting better power.

Copyright © 2011 S. Karger AG, Basel

Introduction

There are many computer programs that implement the various statistical methods of linkage analysis. Our purpose here is to indicate some of the features available

in the program package S.A.G.E. (Statistical Analysis for Genetic Epidemiology) [1], explaining how these features are best used to obtain optimal validity and power. We do this by comparing a few of the various options available for two different types of multipoint linkage analysis on one common set of pedigree data. These pedigrees were ascertained because they segregate for schizophrenia, and we chose them for illustration because they were typed for linkage analysis using the Illumina version 4 linkage panel [2] and have values of plasma dopamine β hydroxylase (pDBH) concentration on 284 of the 922 family members. We first briefly describe the main functions of some of the relevant programs available and then describe a model-based analysis of pDBH, a quantitative trait, and a model-free analysis of a binary disease, schizophrenia, with and without using information on age of onset.

S.A.G.E. comprises 17 programs for the analysis of genetic epidemiological data, designed for analyzing pedigree data but also capable of analyzing data on unrelated individuals. Here we first highlight those programs that can be used to perform linkage analysis and then illustrate the use of some of them. Programs are also available for pedigree structure statistics, and marker and relationship error detection. Furthermore the program ASSOC, though originally designed for association analysis, can also be used for linkage analysis using a TDT (transmis-

sion disequilibrium test)-type statistic [3], as well to estimate heritability.

For model-based linkage analysis, the program MLOD performs multipoint model-based LOD-score linkage analysis on small pedigrees and LODLINK is the two-point equivalent but can accommodate large pedigrees. A unique feature of S.A.G.E. is the program SEGREG, which can perform segregation analysis with the ability to output penetrance files for the model-based linkage programs LODLINK and MLOD.

For model-free linkage analysis, which we define as a linkage analysis where no particular mode of inheritance is assumed for the trait being analyzed, there are three programs. First, SIBPAL analyzes data (qualitative or quantitative) on sibships as a function of marker allele sharing identity-by-descent (IBD), and includes Hase-man-Elston (H-E) regression. Second, RELPAL is a new regression-based univariate or multivariate model-free linkage program that extends the H-E approach to multiple relative types and to a two-level analysis, as proposed by Wang and Elston [4]. Both SIBPAL and RELPAL use either the single- or the multi-marker IBD information produced by the S.A.G.E. program GENIBD, which is relatively fast, to investigate linkage at one or more trait loci, including epistatic interactions and covariate effects. Third, LODPAL was designed to perform linkage analysis based on the LOD score formulation for affected sib pairs (ASPs) [5], but has been generalized to use information from both affected and discordant relative pairs in general. The current implementation is that of the general conditional logistic model proposed by Olson [6], with the possibility of the one-parameter model of Goddard et al. [7] based on the min-max statistic of Whittemore and Tu [8]. The model allows for the inclusion of all affected relative pairs, together with covariates or discordant sibling pairs, with the possibility of pooling unaffected relative pairs together with affected relative pairs.

Methods

Model-Based Linkage Analysis of a Quantitative Trait

To perform model-based linkage analysis we need to assume a trait model, or fit the trait model by segregation analysis. The trait model includes the known (or assumed) trait locus allele frequencies and the values of the penetrance functions (which are usually probability mass functions but, in the case of continuous traits, can be probability density functions) for the genotypes at the trait locus. The SEGREG program in S.A.G.E. can fit very general segregation models for both binary and continuous traits,

and outputs a 'type' file that includes the individual-specific penetrance function values conditional on any fitted segregation model. This type file can then be used for model-based linkage analysis by LODLINK or MLOD. For the quantitative trait pD β H, a low-value dominant model under a square-root transformation was shown to produce the best genome-wide linkage signal, on chromosome 9 at the D β H structural locus [9]. Here we illustrate the fitting of the low-value dominant model with SEGREG, incorporating residual familial correlations and various power transformations and covariates; we then investigate how these factors affect the linkage result.

In SEGREG, we have the option to use Bonney's class D regressive model [10], which allows for the incorporation of familial correlations. The pD β H activities were assumed to have a low-value dominant mode of inheritance; i.e. we assumed a mixture of two density functions, the one with lower mean for genotypes AA and AB, and the one with higher mean for genotype BB. We investigated both sex and schizophrenia status as binary covariates of the genotype means but here only illustrate this for the one covariate sex (neither was significant). SEGREG allows for transforming the trait using the standardized Box-Cox transformation (see Appendix) [11], simultaneously estimating the power parameter together with all the other parameters of the segregation model. It is assumed that, conditional on genotype, after transformation the phenotype y is normally distributed.

Models with various transformations using the general two-parameter standardized Box-Cox transformation (Appendix) were fitted, under which we obtained maximum likelihood estimates (MLEs) of all unknown parameters. First, we set $\lambda_1 = 1$ and $\lambda_2 = 1$ to fit a model with no transformation. If $0 < \lambda_1 < 1$, in some cases $\lambda_2 = 0$ could result in erroneous transformed distributions (see Appendix) and so, for the other three power transformations used in this study, we fixed λ_2 to have a small value, 0.1. Then $\lambda_1 = 0.5$ approximates the square-root transformation and $\lambda_1 = 0$ approximates the logarithmic transformation. We also fitted a model to obtain the MLEs of all parameters other than λ_2 , but including λ_1 . SEGREG outputs likelihoods for the segregation models using the standardized Box-Cox transformation, i.e. adjusted by the Jacobian of the transformation, with the result that all the likelihoods are for the same (untransformed) data and hence comparable.

In SEGREG, the following correlations among the residuals from the predicted genotype means can be allowed in a class D regressive model: ρ_{FM} , spouse (father-mother); ρ_{MO} , mother-offspring; ρ_{FO} , father-offspring; and ρ_{SS} , sibling. Under the class D assumption, the sibling correlation is not constrained to be due to common parentage alone, as is the case for the class A regressive model; in this latter model the sibling correlation can only be larger than the parent-offspring correlation if the spouse correlation is negative. The residuals are assumed to follow a multivariate normal distribution across family members. Thus the power transformation helps to assure this normality, so that a major gene model is fitted allowing for an approximately normally distributed cumulative effect of various factors, such as polygenes and environmental factors, that are not separately distinguished. In each case we fitted familial correlations as for the usual mixed major-gene/polygenic model, i.e. equal parent-offspring and sib-sib correlations, and no spouse correlation. We also fitted models without any familial correlations, for comparison. For each fitted

model, the type file includes each individual's penetrance function values for each genotype after the class D regressive model has been fitted, thus allowing for residual familial correlations, and this was used as an input file for MLOD. We performed model-based multipoint linkage analysis for pDBH activity specifically for chromosome 9. LODs were estimated at the markers and at intervals of 2 cM.

Estimating Marker Allele Frequency and Exact Multipoint IBD Sharing

MLEs of the pedigree founder marker allele frequencies were estimated by the S.A.G.E. program *FREQ*, using the genotype data available on all the pedigree members. These were used for the model-based linkage analysis of pDBH activity, as well as for estimating multipoint IBD sharing by *GENIBD* for the model-free analysis of schizophrenia described below; in this case *GENIBD* was used to estimate exact multipoint IBD sharing distributions for five types of relative pairs, for each marker location and at intervals of 2 cM.

Model-Free Linkage Analysis of a Disease Trait with and without Age of Onset Information

For a binary trait, *SIBPAL* performs the usual mean test and proportion test of allele sharing among ASPs, as well as the analogous tests for discordant and concordantly unaffected sib pairs. For these tests, estimates are obtained of f_0 , f_1 and f_2 , the probabilities that sibs share 0, 1, or 2 alleles IBD. For quantitative traits *SIBPAL* performs various types of H-E regression [4]. For binary traits with a variable age of onset, S.A.G.E. has the capability of using age of onset information in a model-free linkage analysis. The program *AGEON* obtains MLEs of parameters that determine 'susceptibility' to disease as a function of age, producing a quantitative susceptibility trait for each individual.

Let γ be the probability a sib is susceptible, by which we mean that individual will eventually become affected. This implies a sib who is not susceptible will never become affected (even, theoretically, should the sib live infinitely long). Let the cumulative distribution function of the age of onset, a (relevant only for susceptible sibs), be $\Phi(a)$. Then the probability that a sib is affected by age a is $\gamma\Phi(a)$, and $1 - \gamma\Phi(a)$ is the probability of not being affected by age a . If unaffected by age a , the probability of a sib being susceptible is $P(\text{susceptible})P(\text{unaffected at age } a \mid \text{susceptible}) / P(\text{unaffected at age } a) = \gamma[1 - \Phi(a)] / [1 - \gamma\Phi(a)]$. Thus the probability a sib is susceptible, conditional on the binary phenotype and age a , is 1 if the sib is affected, and $\gamma[1 - \Phi(a)] / [1 - \gamma\Phi(a)]$ if the sib is unaffected at age a . We call this probability the sib's susceptibility trait.

The affection status of the parents (each of the two parents may be affected, unaffected or unknown) form a categorical covariate with 6 classes [12] and hence allows estimation of 6 corresponding values of γ . Using both affected and unaffected individuals, *AGEON* also Box-Cox transforms the age of onset, which is assumed to be normally distributed after transformation. Thus age of onset is assumed to follow a power-normal distribution but, for the same reasons as stated above, λ_2 is set to a small value (we used 0.5, the default value). *AGEON* obtains the MLEs of the mean and variance of the age of onset distribution, together with the 6 values of susceptibility to disease, γ – the probability of ever being affected. The likelihood is also maximized constraining these 6 sus-

ceptibilities to be equal, allowing for a likelihood ratio test of this as a null hypothesis.

We performed analyses using *SIBPAL* for (1) the binary trait affection status, and (2) two different quantitative traits. The mean test and the proportion test were used for the binary trait and H-E regression tests were performed for two quantitative traits – the binary trait as a (0, 1) quantitative trait [13] and the susceptibility trait as produced by the program *AGEON*.

SIBPAL has the option to use either the best linear unbiased predictor (BLUP) of the sibship mean or the overall sample mean when mean-correcting the squared sib-pair sum, which can be combined with the squared sib-pair difference to form the dependent variable. It has been shown that these methods that use a mean-corrected sum are asymptotically more powerful than the original H-E method, which is independent of the population mean [4, 14]. The sibship-specific mean is the mean of the trait values for a given sibship and is estimated as the average

$$\tilde{x} = \frac{1}{N_s} \sum_{i=1}^{N_s} x_i$$

where N_s is the number of sibs in a particular sibship. The BLUP of the mean for that sibship is then

$$w\tilde{x}_s + (1 - w)\tilde{x}, \text{ where } w = \sigma_b^2 / \left[\sigma_b^2 + \frac{\sigma_w^2}{N_s} \right],$$

where σ_b^2 is the variance among sibships and σ_w^2 is the variance within sibships. We used option W4 in *SIBPAL*, which is asymptotically optimal, to combine the sib-pair sum and the sib-pair difference [15], obtaining both asymptotic and empirical p values from the permutation distribution. For the latter, estimates of IBD sharing, used as the predictor in the regression, are permuted by *SIBPAL* both within sibships and across sibships of the same size. We compared these results to the original H-E method. In addition, for option W4, we compared the results obtained using the sample mean to the results using the BLUP of the sibship mean.

Results

Segregation Models

An earlier study fitted the segregation model under a square-root transformation by fixing $\lambda_1 = 0.5$ and $\lambda_2 = 0$ [9]. In this study, we fixed $\lambda_2 = 0.1$. Among the four low-value dominant models under the square-root transformation in table 1 (models 1, 2, 3, 4), the two models incorporating familial correlations (models 1 and 3, assuming equal parent-offspring and sib-sib correlations) fitted better according to Akaike's A information criterion (AIC) [16] than the two without familial correlations (models 2 and 4). Moreover, including sex as a covariate for the genotype means did not improve the fit, and in fact yielded a worse fitting model (compare the AIC values: model 1 with 3, model 2 with 4), and sex as a covariate was not significant (table 1). Models using either a log-

Table 1. Estimates of the parameters for low-value dominant inheritance under various models, defined by transformations, with (+) and without (–) sex as a covariate, and with (+) and without (–) familial correlations, using 284 individuals with pDβH activities in 70 pedigrees

Models	1	2	3	4	5	6	7 ^a
Transformation	Square-root	Square-root	Square-root	Square-root	None	Logarithm	Estimated power(λ_1)
Familial correlations	+	–	+	–	+	+	+
Covariate	+	+	–	–	–	–	–
q_A^b	0.51 ± 0.07	0.39 ± 0.05	0.52 ± 0.07	0.40 ± 0.05	0.59 ± 0.04	0.11 ± 0.03	0.52 ± 0.05
μ_{AA}^c	8.00 ± 1.07	6.41 ± 0.69	8.12 ± 1.01	6.52 ± 0.69	10.64 ± 0.72	1.91 ± 0.39	1.78 ± 0.29
μ_{AB}^c	8.00 ± 1.07	6.41 ± 0.69	8.12 ± 1.01	6.52 ± 0.69	10.64 ± 0.72	1.91 ± 0.39	15.41 ± 1.37
μ_{BB}^c	34.14 ± 4.6	26.70 ± 2.1	34.71 ± 4.40	27.11 ± 2.2	42.47 ± 1.62	13.99 ± 1.26	15.41 ± 1.37
β_{sex}^d	0.72 ± 1.03	1.40 ± 1.04	–	–	–	–	–
$\rho_{PO} = \rho_{SS}$	0.23 ± 0.07	0	0.23 ± 0.06	0	0.22 ± 0.05	0.13 ± 0.07	0.37 ± 0.07
λ_1	0.50	0.50	0.50	0.50	1	0	0.01 ± 0.06
Akaike's AIC	2,118.04	2,123.55	2,116.53	2,123.34	2,210.55	2,109.22	2,087.74

Values are given with standard errors. ^a Converged to a low-value recessive model. ^b Frequency of allele A. ^c Genotype means on the untransformed scale. ^d Coefficient for mean centered sex, coded as 0 for males and 1 for females.

Fig. 1. Multipoint model-based linkage on chromosome 9 under the low-value dominant trait models with a square-root transformation (Box-Cox transformation, $\lambda_1 = 0.5$, $\lambda_2 = 0.1$), with and without residual familial correlations and sex as a covariate of the genotype means.

arithmetic or estimated power transformation fitted these data better than models using a square-root transformation, as judged by their AIC values, but all fitted better than models with no transformation.

Model-Based Linkage Analysis

As shown in figure 1, among the four low-value dominant models using the square-root transformation, the two models incorporating familial correlations produced higher linkage signals on chromosome 9 compared with the two models ignoring familial correlations. The model including sex as a covariate for the genotype means produced a lower linkage signal compared to the model without the covariate. These results are consistent with the results from segregation analysis in that the better fitting model produced a better linkage signal. Among all the models in table 1, the best linkage signal on chromosome 9 was produced by model 1, the square-root transformation model incorporating familial correlations but without a sex covariate (fig. 1). The maximum LOD for this model was 5.94.

Comparing linkage results under the dominant models using various transformations (fig. 2), the model using a square-root transformation produced the highest linkage signal, while the model using the estimated transformation power λ_1 , which converged to a low-value recessive model, produced the lowest linkage signal. Thus, restricting the model to be dominant for low values, as found in previous studies [17, 18], was necessary to obtain a segregation model useful for linkage analysis.

Table 2. Descriptive statistics for schizophrenia: sex, age of onset, and age at exam by affection status

	Age of onset (SD), years	Age at exam (SD), years
<i>Affected (187 M, 110 F)</i>	n = 286; missing: n = 11 ^a	n = 11 ^a
Mean	21.78 (6.93)	45.82 (13.98)
Median	20	38
Range	5–48	27–69
<i>Unaffected (276 M, 270 F)</i>	not applicable	n = 478; missing: n = 68
Mean		57.27 (16.93)
Median		59
Range		11–106

^a Affected subjects missing age of onset but having age at exam.

Model-Free Linkage Analysis

Model-free linkage analysis of pDβH activity in these pedigrees was reported in [9]. Here we concentrate on schizophrenia as our trait of interest. A brief description of the data is presented in table 2. There were 286 affected individuals with (apart from 11 for whom this information was missing) average age of onset 21.78 years, and 487 unaffected individuals with (apart from 68 for whom this information was missing) average last known age at which the subject was known to be unaffected (age at exam) 57.27 years. In order to increase the amount of information to estimate the age of onset distribution, we imputed the missing ages for those unaffected to be 11, the smallest age at exam in the sample, on the reasonable assumption that this would almost certainly be an age by which these individuals would not have been affected. For the 11 affected individuals missing age of onset, age at exam was available. We now give separate results for schizophrenia considered as a binary trait and as two quantitative traits.

Binary Trait

For the binary trait affected or unaffected with schizophrenia, the smallest p value was obtained on chromosome 13. The results for the usual ASP mean and proportion tests are plotted for chromosome 13 in figure 3. This figure also shows (top three panels) the analogous tests for discordant sib pairs and concordantly unaffected sib pairs. The min-max statistic of Whittemore and Tu [8] should lie between these two lines in the top panel of figure 3, and analogous statistics should lie between the two lines in the next two panels. At position 94.8 cM p-ter, the mean allele sharing for concordantly affected and unaffected sib pairs was significantly larger than expected

Fig. 2. Multipoint model-based linkage on chromosome 9 under the low-value dominant models, with different power transformations. Residual familial correlations were included in the trait model, and no covariate was included. The Box-Cox transformations with power $\lambda_1 = 0.5, 1$ and 0 are approximately a square-root transformation, no transformation and a logarithmic transformation, respectively. The model with estimated $\lambda_1 (=0.01)$ was a low-value recessive model instead of a low-value dominant model because it had the larger likelihood.

ed under the null hypothesis ($p < 0.01$ and $p < 0.05$, respectively). The results were also significant for concordantly unaffected sib pairs sharing 0 or 2 alleles IBD, as was the result for ASPs sharing 0 alleles IBD.

Quantitative Traits

For schizophrenia analyzed as a quantitative (0, 1) trait, the analysis using the BLUP of the sibship mean in-

Fig. 3. Top three panels: IBD sharing for sib pairs on chromosome 13. Estimates of f_2 (solid lines) and mean allele sharing, $\hat{\pi} = f_2 + f_1/2$ (dashed lines) from affected (ASP), discordant (DSP) and unaffected (USP) sib pairs were normalized to a multiple of SEs from the expected value under the null hypothesis of no linkage (horizontal dotted lines). Bottom panel: the slope parameter from H-E regression using the BLUP of the sibship mean for the quantitative trait schizophrenia trait similarly normalized. Vertical lines at 77 and 94.7 cM from p-ter mark the most significant results for ASP and USP, respectively.

stead of the sample mean in H-E regression yielded a smaller p value (0.001 vs. 0.002, respectively). The bottom panel of figure 3 shows the H-E regression using the BLUP of the sibship mean; the regression parameter was equal to the allele sharing by concordant sib pairs minus that by discordant sib pairs, and is plotted for $f_2 + f_1/2$.

When fitting the susceptibility model with AGEON, the MLE of the power parameter was 0 (natural log transformation). When estimating five susceptibility parameters (γ) (one category had no observations, i.e. there were no sibships where both parents were affected), the category where both parents were unaffected had the lowest estimate of susceptibility (0.53), but the estimate for the case where one parent was unknown and the other parent unaffected was slightly higher than that for one unknown and one affected parent (0.633 vs. 0.630) – i.e. the values of γ were not in a logical order.

Twice the difference in the log likelihoods between the model with five susceptibilities compared to equal susceptibilities was not significant ($\chi^2 = 3.91$, d.f. = 4), and therefore estimates with the susceptibilities equal were used in the analysis with SIBPAL. The susceptibility for this model was 0.56 and the estimated age of onset distribution had mean 20.9 years and standard deviation 1.92 on the original scale. Using the quantitative susceptibility trait in H-E regression yielded a smaller p value than did the binary unaffected/affected (0, 1) trait (0.00002 vs. 0.001, table 3). The susceptibility trait yielded empirical p values <0.001 on chromosomes 2, 11 and 13, with the smallest p value on chromosome 13. We chose to report p values instead of LOD scores as the LOD score can be biased [19]. Table 4 summarizes these quantitative trait results using both the BLUP of the sibship mean and the sample mean to correct the sib-pair trait sum. Empirical p values were calculated using the permutation option in SIBPAL. In all cases, using the original H-E regression resulted in larger p values than the W4 option.

Discussion

SEGREG and MLOD are two programs in S.A.G.E. that respectively fit segregation models and perform model-based multipoint linkage analysis. SEGREG can fit models for both quantitative and binary traits; it allows for including covariates for genotype means and genotype variances in the case of a quantitative trait, or covariates of susceptibilities for a binary trait. An example of the flexibility of SEGREG for the segregation analysis of a binary trait is given in Sun et al. [20]. SEGREG incorporates a transformation in the likelihood function for a quantitative trait, and can incorporate a multifactorial effect, and hence residual familial correlations, in two distinct ways – via a regressive model as illustrated here and via a special form of the usual mixed major gene/polygenic model to incorporate a finite number of polygenic loci. It also allows for ascertainment by conditioning the likelihood on the phenotypes of a subset of individuals. MLOD can perform model-based linkage based on a trait model fitted by SEGREG. Our study illustrated fitting genetic models for pD β H activity allowing for familial correlations by using a regressive model, as well as including covariates and incorporating various transformations. The program REGRESS [21, 22] and the updated faster program FINESSE [Florence Deme-nais, personal communication] incorporate regressive

Table 3. Most significant p values for tests of f_2 and H-E regression (original and the W4 option), chromosome 13

Analysis	Location cM	Asymptotic p	Empirical p
ASP	77.2	0.019	–
DSP	90.9	0.052	–
USP	94.7	0.0022	–
H-E ^a (0, 1 quantitative)	94.7	0.002	0.001
H-E (Susc. ^b)	94.0	0.00007	0.00002
H-E (Susc. ^c)	94.8	0.0001	0.00004

DSP = Discordant sib pairs; USP = unaffected sib pairs. ^a H-E regression ($\hat{\pi} = f^2 + f^1/2$). ^b Susceptibility trait with option W4. ^c Susceptibility trait with original H-E.

Table 4. Summary of SIBPAL results for quantitative trait analysis (H-E regression)

Trait	Chromosome	BLUP of sibship mean		Sample mean	
		position from p-ter	empirical p value	position from p-ter	empirical p value
(0, 1) trait	13	94.8	0.001	94.8	0.002
Susceptibility trait	2	133.3–135.5	0.0007	128.9	0.006
	11	90	0.0006	90.3	0.007
	13	94	0.00002	94	0.0006

models into the LINKAGE program, and they can simultaneously fit a regressive model while performing linkage analysis ('combined segregation and linkage analysis'), but there are advantages to a two-stage analysis using SEGREG and MLOD separately: incorporating covariates, adjusting for ascertainment, and simultaneously estimating the best transformation. In addition, performing segregation analysis separately enables use of all the data available to obtain the penetrance values and then trim off those individuals with no DNA data for the linkage analysis. If genotype means or variances are dependent on some covariates, SEGREG can incorporate such covariates in the model to overcome the difficulty of directly incorporating them in model-based linkage analysis.

Our results indicated pDBH activity, using the appropriate square-root transformation and fitting the appropriate low-value dominant penetrance model, still required a model with familial correlations, but not a sex covariate, for this model to show higher power to detect linkage. However, among all the models with various

transformations, the best fitting model (model 7 in table 1) did not produce a higher linkage signal (the lowest linkage signal in fig. 2). This is consistent with the earlier study showing that the best fitting model may not produce a better linkage result [9]. However, note that in the best fitting model, when λ_1 was estimated, its standard error was relatively large, i.e. the model is suspect because it has a poorly estimated parameter. This emphasizes the assumption, always made in a model-based linkage analysis, that the genetic model is true. Thus we should be cautious in choosing a model for a model-based linkage analysis.

Our results confirmed accounting for residual familial correlations in the regressive models can increase the power to detect linkage [21]. It is difficult to incorporate a regressive model when parents have missing trait values – typically it is then assumed the grandparent-grandchild residuals are zero. However, the usual mixed model with a latent polygenic component does not require that all parents have trait values. SEGREG could be used to fit a mixed model and hence estimate a residual familial cor-

relation (the same for sibs and parent-offspring) even when parental values are missing (in these data, 38 and 25% of the children have one parent and both parents missing pDβH values, respectively). Using this correlation estimated from the mixed model, we could impute from the siblings' trait values all missing parental values and then fit a regressive model. Including this information in a linkage analysis via the penetrance function, as we illustrated, would be expected to further improve the linkage signal.

For schizophrenia, analysis of the binary trait yielded nominally significant results on chromosome 13. Using the BLUP of the sibship mean produced a smaller *p* value than the sample mean. However, the BLUP of the sibship mean may not always lead to most power [14]; it is expected to do so in large-enough samples (i.e. when the estimate *w* is good). The susceptibility trait, however, produced a more significant result at the same position (table 4). In addition, the susceptibility trait also produced nominally significant findings on chromosomes 2 and 11. Previous studies have found linkage to chromosome 8 for these data [23]. Other studies have previously reported findings suggesting the importance of chromosome 11 [24–29] and chromosome 13 [30–33], and a recent meta-analysis found evidence for linkage to 1, 2q, 3q, 4q, 5q, 8p and 10q [34]. We therefore believe our analysis of susceptibility has probably increased the power of linkage analysis in these data.

Acknowledgements

This work was supported in part by a US Public Health Service Resource grant (RR03655). We wish to thank Dr. Ann E. Pulver of The Johns Hopkins School of Medicine Epidemiology-Genetics Program in Psychiatry as the source of the schizophrenia data and Dr. Joseph F. Cubells of Emory University School of Medicine for the plasma DβH data, obtained in part with support from a US Public Health Service research grant (MH 077233).

Appendix

The general two-parameter standardized Box-Cox transformation of a continuous variate *y* is given by

$$h(y) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1 y_G^{\lambda_1 - 1}} & \text{if } \lambda_1 \neq 0 \\ y_G \ln(y + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

where

$$y_G = \left[\prod_{i=1}^N (y_i + \lambda_2) \right]^{\frac{1}{N}}$$

and *N* is the number of data values *y* in the dataset.

For the original trait *x*, assume that after transformation *y* is normally distributed with mean μ and variance σ^2 . The relationship between *y* and *x* (Box-Cox transformation without inclusion of the Jacobian) is

$$y = \begin{cases} \frac{(x + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \ln(x + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}, \quad x + \lambda_2 \geq 0.$$

The probability density function of *y* is

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

and the probability density function of *x* is

$$f_x(x) = f_y(y) \left| \frac{dy}{dx} \right|,$$

$$\text{where } \frac{dy}{dx} = \begin{cases} (x + \lambda_2)^{\lambda_1 - 1} & \text{if } \lambda_1 \neq 0 \\ 1 & \text{if } \lambda_1 = 0 \end{cases}, \quad x + \lambda_2 \geq 0.$$

So, for $x + \lambda_2 \geq 0$,

$$f_x(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left[\frac{(x + \lambda_2)^{\lambda_1}}{\lambda_1} - \left(\frac{1}{\lambda_1} + \mu \right) \right]^2}{2\sigma^2}} (x + \lambda_2)^{\lambda_1 - 1} & \text{if } \lambda_1 \neq 0 \\ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\{\ln(x + \lambda_2) - \mu\}^2}{2\sigma^2}} \frac{1}{x + \lambda_2} & \text{if } \lambda_1 = 0. \end{cases}$$

When performing a segregation (or commingling) analysis, we wish to avoid any transformation that introduces any spurious modes or antimodes. In order to check if the transformation might cause this in $f_x(x)$, we deduce the location of the possible modes or antimodes of $f_x(x)$ by setting

$$\frac{df_x(x)}{dx} = 0,$$

i.e. $z^2 - (1 + \lambda_1 \mu)z - \sigma^2 \lambda_1 (1 - \lambda_1) = 0$ ($\lambda_1 \neq 0$), where $z = (x + \lambda_2)^{\lambda_1} > 0$. The solution for *z* is

$$z = \frac{1 + \lambda_1 \mu \pm \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2}, \lambda_1 \neq 0,$$

and, in order for *z* to have at least one solution, we must have $\sigma^2 \leq 1 + \mu$.

(1) If $\lambda_1 \geq 1$, *z* has one positive solution, which means the transformation does not change the number of modes (both $f_x(x)$ and $f_y(y)$ have one mode, no antimode).

Appendix **Fig. 1.** Illustration of a situation where the Box-Cox transformation changes a distribution $f_x(x)$ with an antimode to a unimodal $f_y(y)$. The distribution of y is $N(0.5, 1)$, and the transformation is $y = (x^{0.5} - 1) / (0.5)$ ($\lambda_1 = 0.5$, $\lambda_2 = 0$).

(2) If $\lambda_1 < 1$, z has two positive solutions, which means that $f_x(x) = 0$ at

$$x_0 = \left\{ \frac{1 + \lambda_1 \mu \pm \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}} - \lambda_2 (\lambda_1 < 1, x + \lambda_2 \geq 0).$$

Furthermore, when

$$-x_{\min} \leq \lambda_2 < \left\{ \frac{1 + \lambda_1 \mu - \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}} - x_{\min} (\lambda_1 < 1),$$

$f_x(x)$ ($x \geq 0$) has an antimode, which means that the distribution of a trait x with two suprema was transformed to have one (Appendix fig. 1). Therefore, if we want both $f_x(x)$ and $f_y(y)$ to have just one supremum, λ_2 should satisfy, when $\lambda_1 < 1$,

$$\left\{ \frac{1 + \lambda_1 \mu - \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}} - x_{\min} \leq \lambda_2 < \left\{ \frac{1 + \lambda_1 \mu + \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}} - x_{\min}.$$

If the trait values are all positive ($x \geq 0$), i.e. $x_{\min} = 0$, then the restriction on λ_2 when $\lambda_1 < 1$ becomes

$$\left\{ \frac{1 + \lambda_1 \mu - \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}} \leq \lambda_2 < \left\{ \frac{1 + \lambda_1 \mu + \sqrt{(1 + \lambda_1 \mu)^2 - 4\sigma^2 \lambda_1 (1 - \lambda_1)}}{2} \right\}^{\frac{1}{\lambda_1}}.$$

References

- 1 S.A.G.E. program. Available at <http://darwin.cwru.edu/sage> 2012.
- 2 Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan JB, Chee MS, Hansen MST: A highly informative SNP linkage panel for human genetic studies. *Nat Methods* 2004;1:113–117.
- 3 George V, Tiwari HK, Zhu X, Elston RC: A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* 1999;65:236–245.
- 4 Wang T, Elston RC: Two-level Haseman-Elston regression for general pedigree data analysis. *Genet Epidemiol* 2005;29:12–22.
- 5 Risch N: Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet* 1990;46:229–241.
- 6 Olson JM: A general conditional-logistic model for affected-relative-pair linkage studies. *Am J Hum Genet* 1999;65:1760–1769.
- 7 Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM: Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* 2001;68:1197–1206.
- 8 Whittemore AS, Tu IP: Simple, robust linkage tests for affected sibs. *Am J Hum Genet* 1998;62:1228–1242.
- 9 Cubells JF, Sun X, Li W, Bonsall RW, McGrath JA, Avramopoulos D, Lasseter VK, Wolyniec PS, Tang YL, Mercer K, Pulver AE, Elston RC: Linkage analysis of plasma dopamine beta-hydroxylase activity in families of patients with schizophrenia. *Hum Genet* 2011;130:635–645.
- 10 Bonney GE: On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *Am J Med Genet* 1984;18:731–749.
- 11 Box GEP, Cox DR: An analysis of transformations. *J Roy Stat Soc B* 1964;26:211–252.
- 12 Elston RC: The use of polymorphic markers to detect genetic variability; in Woodhead A, (ed): *Phenotypic Variation in Populations: Relevance to Risk Assessment*. New York, Plenum Press, 1988, pp 105–112.
- 13 Elston RC, Song D, Iyengar SK: Mathematical assumptions versus biological reality: myths in affected sib pair linkage analysis. *Am J Hum Genet* 2005;76:152–156.
- 14 Sinha R, Gray-McGuire C: Haseman-Elston regression in ascertained samples: importance of dependent variable and mean correction factor selection. *Hum Hered* 2008; 65:66–76.
- 15 Shete S, Jacobs KB, Elston RC: Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum Hered* 2003;55:79–85.
- 16 Akaike H: A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;AC-19:716–713.
- 17 Asamoah A, Wilson AF, Elston RC, Dalferes E Jr, Berenson GS: Segregation and linkage analyses of dopamine-beta-hydroxylase activity in a six-generation pedigree. *Am J Med Genet* 1987;27:613–621.

- 18 Goldin LR: Segregation analysis of dopamine-beta-hydroxylase (dbh) and catechol-o-methyltransferase (comt): identification of major locus and polygenic components. *Genet Epidemiol* 1985;2:317–325.
- 19 Province MA: Sequential methods of analysis for genome scans. *Adv Genet* 2001;42:499–514.
- 20 Sun X, Elston R, Barnholtz-Sloan J, Falk G, Grady WM, Kinnard M, Mittal SK, Willis JE, Markowitz S, Brock W, Chak A: A segregation analysis of Barrett's esophagus and associated adenocarcinomas. *Cancer Epidemiol Biomarkers Prev* 2010;19:666–674.
- 21 Demenais F, Lathrop M: Use of the regressive models in linkage analysis of quantitative traits. *Genet Epidemiol* 1993;10:587–592.
- 22 Demenais F, Lathrop M: Regress: a computer program including the regression approach into the LINKAGE package. *Genet Epidemiol* 1994;11:291.
- 23 Fallin MD, Lasseter VK, Liu Y, Avramopoulos D, McGrath J, Wolyniec PS, Nestadt G, Liang KY, Chen PL, Valle D, Pulver AE: Linkage and association on 8p21.2-p21.1 in schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 2010.
- 24 Klar AJ: The chromosome 1;11 translocation provides the best evidence supporting genetic etiology for schizophrenia and bipolar affective disorders. *Genetics* 2002;160:1745–1747.
- 25 Klar AJ: A genetic mechanism implicates chromosome 11 in schizophrenia and bipolar diseases. *Genetics* 2004;167:1833–1840.
- 26 Choudhury K, McQuillin A, Puri V, Pimm J, Datta S, Thirumalai S, Krasucki R, Lawrence J, Bass NJ, Queded D, Crombie C, Fraser G, Walker N, Nadeem H, Johnson S, Curtis D, St Clair D, Gurling HM: A genetic association study of chromosome 11q22–24 in two different samples implicates the FXYD6 gene, encoding phosphohippolin, in susceptibility to schizophrenia. *Am J Hum Genet* 2007;80:664–672.
- 27 Fletcher JM, Evans K, Baillie D, Byrd P, Hanratty D, Leach S, Julier C, Gosden JR, Muir W, Porteous DJ, et al: Schizophrenia-associated chromosome 11q21 translocation: identification of flanking markers and development of chromosome 11q fragment hybrids as cloning and mapping resources. *Am J Hum Genet* 1993;52:478–490.
- 28 Gill M, McGuffin P, Parfitt E, Mant R, Asherson P, Collier D, Vallada H, Powell J, Shaikh S, Taylor C, et al: A linkage study of schizophrenia with DNA markers from the long arm of chromosome 11. *Psychol Med* 1993;23:27–44.
- 29 Wang ZW, Black D, Andreasen NC, Crowe RR: A linkage study of chromosome 11q in schizophrenia. *Arch Gen Psychiatry* 1993;50:212–216.
- 30 Maziade M, Chagnon YC, Roy MA, Bureau A, Fournier A, Merette C: Chromosome 13q13-q14 locus overlaps mood and psychotic disorders: the relevance for redefining phenotype. *Eur J Hum Genet* 2009;17:1034–1042.
- 31 Brzustowicz LM, Honer WG, Chow EW, Little D, Hogan J, Hodgkinson K, Bassett AS: Linkage of familial schizophrenia to chromosome 13q32. *Am J Hum Genet* 1999;65:1096–1103.
- 32 Mulle JG, McDonough JA, Chowdari KV, Nimgaonkar V, Chakravarti A: Evidence for linkage to chromosome 13q32 in an independent sample of schizophrenia families. *Mol Psychiatry* 2005;10:429–431.
- 33 Blouin JL, Dombroski BA, Nath SK, Lasseter VK, Wolyniec PS, Nestadt G, Thornquist M, Ullrich G, McGrath J, Kasch L, Lamacz M, Thomas MG, Gehrig C, Radhakrishna U, Snyder SE, Balk KG, Neufeld K, Swartz KL, DeMarchi N, Papadimitriou GN, Dikeos DG, Stefanis CN, Chakravarti A, Childs B, Housman DE, Kazazian HH, Antonarakis S, Pulver AE: Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nat Genet* 1998;20:70–73.
- 34 Ng MY, Levinson DF, Faraone SV, Suarez BK, DeLisi LE, Arinami T, Riley B, Paunio T, Pulver AE, Irmansyah, Holmans PA, Escamilla M, Wildenauer DB, Williams NM, Laurent C, Mowry BJ, Brzustowicz LM, Maziade M, Sklar P, Garver DL, Abecasis GR, Lerer B, Fallin MD, Gurling HM, Gejman PV, Lindholm E, Moises HW, Byerley W, Wijsman EM, Forabosco P, Tsuang MT, Hwu HG, Okazaki Y, Kendler KS, Wormley B, Fanous A, Walsh D, O'Neill FA, Peltonen L, Nestadt G, Lasseter VK, Liang KY, Papadimitriou GM, Dikeos DG, Schwab SG, Owen MJ, O'Donovan MC, Norton N, Hare E, Raventos H, Nicolini H, Albus M, Maier W, Nimgaonkar VL, Terenius L, Mallet J, Jay M, Godard S, Nertney D, Alexander M, Crowe RR, Silverman JM, Bassett AS, Roy MA, Merette C, Pato CN, Pato MT, Roos JL, Kohn Y, Amann-Zalcenstein D, Kalsi G, McQuillin A, Curtis D, Brynjolfsson J, Sigmundsson T, Petursson H, Sanders AR, Duan J, Jazin E, Myles-Worsley M, Karayiorgou M, Lewis CM: Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry* 2009;14:774–785.