Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 46

Genetic Linkage Analysis in the Presence of Germline Mosaicism

Omer Weissbrod, *Technion - Israel Institute of Technology*
Dan Geiger, *Technion - Israel Institute of Technology*

Recommended Citation:
Weissbrod, Omer and Geiger, Dan (2011) "Genetic Linkage Analysis in the Presence of Germline Mosaicism," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 46.
DOI: 10.2202/1544-6115.1709
Available at: http://www.bepress.com/sagmb/vol10/iss1/art46

Genetic Linkage Analysis in the Presence of Germline Mosaicism

Omer Weissbrod and Dan Geiger

Abstract

Germline mosaicism is a genetic condition in which some germ cells of an individual contain a mutation. This condition violates the assumptions underlying classic genetic analysis and may lead to failure of such analysis. In this work we extend the statistical model used for genetic linkage analysis in order to incorporate germline mosaicism. We develop a likelihood ratio test for detecting whether a genetic trait has been introduced into a pedigree by germline mosaicism. We analyze the statistical properties of this test and evaluate its performance via computer simulations. We demonstrate that genetic linkage analysis has high power to identify linkage in the presence of germline mosaicism when our extended model is used. We further use this extended model to provide solid statistical evidence that the MDN syndrome studied by Genzer-Nir et al. has been introduced by germline mosaicism.

KEYWORDS: linkage analysis, germline mosaicism, statistical test, statistical model

Author Notes: This work was supported by the National Institute of Health [5R01HG004175-03] and the Israeli Science Foundation. We thank Tzipora Falik-Zaccai and Mira Genzer-Nir for their assistance with the MDN data, and Mark Silberstein and Andrei Anisenia for providing us the source code of the sum of product algorithm.

1 Introduction

Genetic linkage analysis is a widely used statistical method for associating disease genes with their location on the chromosome. The principal idea behind genetic linkage is that loci which are located in close proximity on the chromosome tend to be passed together from parents to offspring. One can deduce the approximate location of a causative gene by finding a group of adjacent genetic markers that segregate healthy and affected individuals. Finding a causative gene sheds light on the biological mechanism that malfunctions in affected individuals and is sometimes the first step towards the development of a suitable remedy.

Genetic linkage analysis has been very successful in mapping genes involved in simple Mendelian diseases. Well known examples include Duchenne muscular dystrophy (DMD), cystic fibrosis (CF), and Huntington's disease (HD) (Borecki and Rice, 2010). However, it is less powerful in mapping genes involved in traits that do not follow simple Mendelian inheritance patterns. Traits that do not follow such patterns were once thought to rarely exist and received relatively little attention. With the growing availability of genomic data, it becomes increasingly clear that such human genetic traits are more frequent than previously thought (Erickson and Lewis, 1995, Gropman and Adams, 2007, Yaron and Orr-Urtreger, 2002). Such traits violate some of the assumptions underlying classic genetic linkage analysis, and thus their associated genes may elude detection. When a genetic linkage analysis study fails, it may therefore still be possible to detect linkage by changing the assumptions underlying the model used for the analysis. To date, there has been little theoretical work trying to formulate statistical tests for identifying biological phenomena that do not follow simple Mendelian inheritance in pedigrees. One common approach to dealing with such traits is employing non-parametric linkage tests, which do not assume a specific mode of inheritance. However, these tests lack statistical power in comparison to parametric tests that use an appropriate explicit model (Strauch et al., 2000).

In this paper we develop an extended statistical model for genetic linkage analysis to incorporate germline mosaicism (GM), which is a condition in which some germ cells of an individual contain a mutation. Germline mosaicism has been found in a variety of inherited traits (e.g. Barbosa et al., 2008, Choi et al., 2008, Fabrizi et al., 2001, Khan et al., 2010, Makri et al., 2009, Pauli et al., 2009). We use the extended statistical model to develop a parametric likelihood ratio test to evaluate whether a specific individual in a given pedigree has GM at a trait locus. Some theoretical aspects of GM have been studied before in the context of risk occurrence (Edwards, 1989, Grimm et al., 1990, Hartl, 1971, Jeanpierre, 1992, Murphy et al., 1974), but no statistical test has been proposed previously for identification of GM in a pedigree. We demonstrate the effectiveness of the test by providing solid statistical evidence for GM in a pedigree affected with MDN (Genzer-Nir et al., 2010), in which GM has been hypothesized. A free computer package which performs this test is available to download at `http://bioinfo.cs.technion.ac.il/superlink-GM`.

The rest of this article is organized as follows. Section 2 defines genetic linkage analysis, germline mosaicism and explains the difficulty to identify GM in pedigrees. Section 3 provides a statistical genetic model that incorporates GM and a likelihood ratio test for detecting GM. It also demonstrates the statistical properties of this test and evaluates its effectiveness via computer simulations. Section 4 uses the statistical test to provide solid statistical evidence for GM in the pedigree reported by Genzer-Nir et al. Finally, Section 5 discusses the merits and limitations of the test and proposes future extensions.

2 Background

This section provides background information regarding genetic linkage analysis and germline mosaicism, and demonstrates the difficulties of identifying GM by standard methods.

2.1 The Standard Model

The standard genetic model used in linkage analysis has been widely studied (Elston and Stewart, 1971, Friedman et al., 2000, Lander and Green, 1987). Every pedigree can be described as a Directed Acyclic Graph (DAG) under this model (Fishelson and Geiger, 2002). Figure 1 depicts a family of two parents and one child, denoted by a, b and c, respectively. The model defines two random variables for each locus of every individual in the pedigree, where the variables $G_{i,k,p}$ and $G_{i,k,m}$ denote the paternal and maternal allele of individual i at locus k, respectively. The model also defines selector variables (also called meiosis variables) denoted by $S_{i,k,p}$ and $S_{i,k,m}$, which indicate whether individual k received the paternal or the maternal allele of her respective parent at locus i. The variable $S_{i,k,p}$ is equal to 0 if the paternal allele of individual k at locus i is equal to the paternal allele of her father at locus i and is equal to 1 otherwise. The variable $S_{i,k,m}$ is defined similarly for the maternal allele.

Two loci on the same chromosome are passed together from parent to child if no recombination occurs between them during meiosis. The recombination frequency between loci i and $i+1$, denoted as $\theta_{i,i+1}$, lies in the interval $[0, 0.5]$, where higher values correspond to a higher probability of recombination between the two loci. Namely, $\theta_{i,i+1}$ is equal to 0.5 when the two loci segregate independently. The

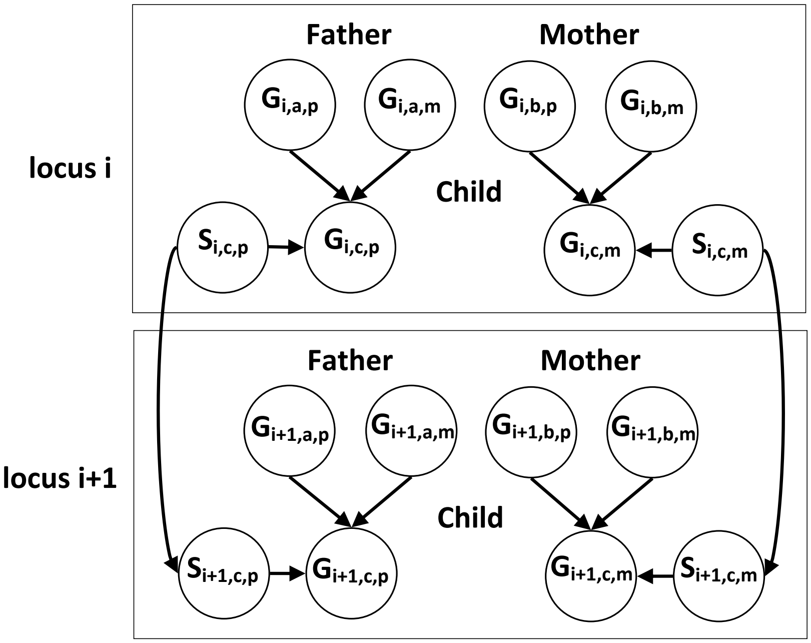

Figure 1: A DAG representing a family with two parents and one child.

value of a selector variable at locus $i+1$ is different from the selector variable at locus i when a recombination occurs between the two loci, an event whose probability is $\theta_{i,i+1}$.

2.2 Genetic Linkage Analysis

Genetic linkage analysis is a statistical hypothesis test which determines whether a certain genetic locus is linked to a genetic trait (Ott, 1999). The test compares the hypothesis that the tested locus and the trait locus are linked with the null hypothesis that they are unlinked, using a likelihood ratio test. The test computes:

$$\mathrm{LOD}(A,G) \triangleq \log_{10}\left[\frac{\Pr(A,G|L,\gamma,\lambda,\theta^{*})}{\Pr(A,G|U,\gamma,\lambda)}\right]. \tag{1}$$

The quantities in Equation 1 are defined as follows. The variables A and G denote the phenotypic and genotypic data of the pedigree, respectively. The events L and U denote that the trait locus is linked and unlinked to the tested locus, respectively. The event L asserts that the recombination frequency between these two loci is smaller than 0.5, while the event U asserts that it is equal to 0.5. The parameter

θ is the recombination frequency between the two loci in the event of linkage, and θ^* is the maximum likelihood estimate of θ in the range $[0, 0.5)$. Finally, the parameters γ and λ correspond to the prevalence and the penetrance parameters of the genetic trait, respectively. A value of LOD ≥ 3.3 is considered sufficient evidence for linkage (Lander and Kruglyak, 1995).

The prevalence parameter γ corresponds to the probability that a chromosome of a randomly chosen individual carries a mutated allele. The penetrance parameter λ is the probability of the phenotype given the genotype of an individual. A genetic trait is said to have full penetrance if an individual who carries the number of mutated alleles required to cause affection is affected with 100% probability. Otherwise, the trait has reduced penetrance.

2.3 Germline Mosaicism

Germline mosaicism (GM) is a condition in which part of the germ cells of an individual contain a mutation (Zlotogora, 1998). This mutation can affect the phenotype of every child who was conceived from a mutated germ cell. The earlier the mutation occurred in the development of the individual, the larger the percentage of germ cells in the body which carry this mutation. When an unaffected parent has several children affected with a genetic trait, without the parent having a family history of the trait, germline mosaicism is an optional explanation. It is not possible to directly detect GM by genotyping unless the mutation altered the allele of a marker that has been genotyped. It is therefore unlikely to directly detect GM even when very dense marker maps are used.

An example of a pedigree with GM is given in Figure 2. Assume that the first, second and fourth loci in each haplotype have been genotyped, while the third locus has not been genotyped. This locus has two alleles, denoted by *w* and *m*, where *w* is a wildtype gene and *m* is a mutated gene which causes affection. A certain percentage of the germ cells of individual I-1 carry a mutation at the third locus. The children of individual I-1 may therefore receive a mutated haplotype. Mutated haplotypes are shown in striped gray, while non-mutated haplotypes are shown in solid gray. It is impossible to distinguish mutated haplotypes from non-mutated haplotypes by genotyping since the mutation only altered the gene at the third locus, which has not been genotyped.

Figure 2 demonstrates that germline mosaicism creates inheritance patterns that are highly unlikely under the assumptions of standard linkage analysis. For example, individuals II-1 and II-3 have both received the same markers from their parents, yet individual II-1 is affected while individual II-3 is not. This is because individual II-1 received a mutated haplotype while individual II-3 received a wildtype

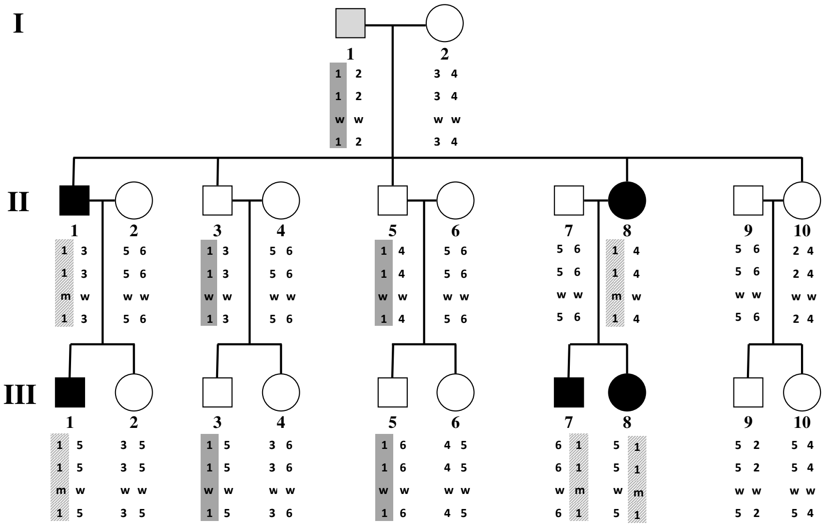

Figure 2: A three generations pedigree whose male founder (I-1) has GM. Black shading indicates affection and gray shading indicates an unknown affection status. A haplotype with four loci is shown for each individual. Affected individuals carry the mutated gene *m*.

haplotype. It is impossible to distinguish between these two haplotypes, since the mutation has not altered any of the three measured markers. Offspring of individual II-1 who received the gray-shaded haplotype are also affected, while offspring of individual II-3 are not affected, even if they carry the gray-shaded haplotype. This pattern is highly unlikely under the assumptions of standard genetic linkage analysis, since one subpedigree shows evidence of linkage between a certain haplotype and a genetic trait, while the other subpedigree does not. Double recombination before and after the trait locus in the meiosis of every affected child is a possible explanation, but it is hardly an option in dense maps. Reduced penetrance is also a possible explanation, but this assumption becomes increasingly implausible as the trait segregates more clearly in large subpedigrees. Standard genetic linkage analysis is therefore unlikely to succeed in demonstrating linkage between the trait and the haplotype. In this work we overcome this obstacle by incorporating GM into genetic linkage analysis.

2.4 Detection of GM by Model Comparison

Model comparison is typically performed in genetic linkage analysis by comparing the likelihood of the phenotypic data of a pedigree under several competing models.

For each pair of models denoted by $\mathcal{M}_1$ and $\mathcal{M}_2$, one computes the likelihood ratio $\Pr(A\,|\,\mathcal{M}_1)\ /\ \Pr(A\,|\,\mathcal{M}_2)$, where A denotes the phenotypic data of the pedigree. This method is useful for differentiating between different modes of inheritance (MOIs). For example, one may use this method to determine whether a genetic trait follows a dominant or recessive MOI. In the pedigree given in Figure 2, the likelihood ratio of these two models is $10^{5.17}$, assuming an allele prevalence of 0.1% and full penetrance. This indicates that the trait is over 100,000 more likely to follow a dominant than a recessive MOI.

Unlike differentiating recessive versus dominant models, differentiating a model which assumes a genetic trait has been introduced by GM from a model which assumes the converse via phenotypic data alone is a subtle and sometimes infeasible task. As an example, consider again the pedigree shown in Figure 2. Denote by $\mathcal{M}_1$ a model with a fully penetrant dominant MOI in which the trait has been introduced into the pedigree from individual I-1 by inheritance, and denote by $\mathcal{M}_2$ a model which assumes the same MOI and that the trait has been introduced into the pedigree from individual I-1 by GM. Assuming a mutated germ cells frequency of 50% for individual I-1, the likelihood ratio $\Pr(A\,|\,\mathcal{M}_1)\ /\ \Pr(A\,|\,\mathcal{M}_2)$ is 1.19, indicating that the trait is equally likely to be introduced into the pedigree by inheritance and by GM. More generally, the likelihood ratio of a model which assumes a standard dominant MOI versus a model which assumes GM with a 50% mutated germ cells frequency is approximately $0.5^n / \left(0.25^k \cdot 0.75^{n-k}\right)$, where n denotes the number of children of the GM suspect and k denotes the number of such children who are affected. Since human families are typically small, this ratio is not informative enough to determine whether the studied trait is caused by GM.

The suggested resolution to the difficulty of identifying GM in pedigrees is to use both the phenotypic and the genotypic data of a pedigree. One can then possibly deduce that GM has occurred due to the irregular pattern of inheritance that emerges when GM takes place; When the founder of a pedigree has GM, the pedigree can be divided into two subpedigrees, where one subpedigree shows high correlation between certain markers and the genetic trait, while the second subpedigree does not. For example, in the pedigree shown in Figure 2, individuals II-1, II-8 and their offspring show high correlation between the trait and the haplotype shaded in gray. Other individuals in the pedigree do not show such a correlation. This irregular pattern may indicate that GM has occurred. In this article we develop a statistical test that incorporates the likelihood of both the phenotypic and the genotypic data of a pedigree to identify GM.

Table 1: The probability that GMs passes a mutated allele to a child

			GMs paternal, maternal alleles			
			w,w	w,m	m,w	m,m
Standard Model	$W = 0$	$S = 0$	0	0	1	1
	$W = 0$	$S = 1$	0	1	0	1
GM Model	$W = 1$	$S = 0$	β	β	1	1
	$W = 1$	$S = 1$	0	1	0	1

3 Statistical Formulation of GM

This section incorporates GM into the statistical model used for genetic linkage analysis. It develops a likelihood ratio test for identifying occurrences of GM in a pedigree and analyzes its statistical properties. The test is evaluated empirically via simulations.

3.1 GM Statistical Model

The standard model described in Section 2.1 can be extended to account for GM. Consider a genetic trait determined by a gene that has two alleles, where w denotes a wildtype allele and m denotes a mutated allele. Further consider a pedigree with a founder who may have GM at one of the two homologous chromosomes that carry the trait locus. Denote this founder as GM suspect (GMs). To model GM in this founder, we add a new random variable W and a new parameter β to the standard model for genetic linkage analysis. The variable W is equal to either 0 or 1, where $W = 0$ indicates that GMs does not have GM and $W = 1$ indicates that GMs has GM. The probability $\Pr(W = 1)$ corresponds to the prior probability that GMs has GM. The probability β that GMs will pass a mutated allele to a child is given by $\beta = \Pr(G_1 = m | W = 1, G_2 = w, S = 0)$, where G_1 denotes the allele that the child received from GMs, G_2 denotes the paternal allele of GMs and $S = 0$ indicates that the child received the paternal allele of GMs. In other words, β is the probability that GMs will pass a mutated allele instead of a wildtype allele to a child, given that GM occurred and given that the paternal segment of the mutated chromosome is passed. Table 1 shows the probability that GMs passes a mutated allele to a child, given the trait alleles of GMs and given W and S. Rows in Table 1 show values of W and of S, where $S = 0$ and $S = 1$ denote that the child receives the paternal and maternal allele of GMs, respectively. For example, if W=1, $S = 1$ and GMs alleles

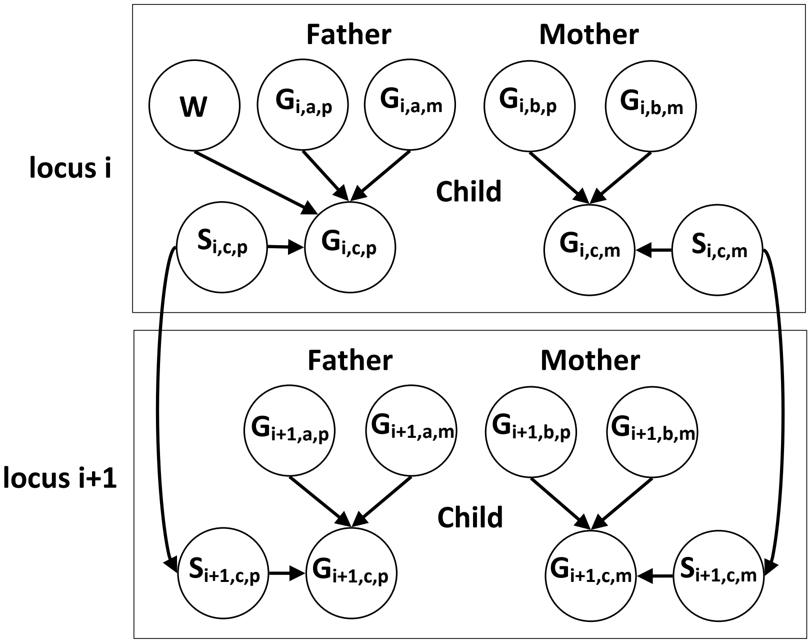

Figure 3: An extended model representation of a family whose father may have GM. The GM affects locus i but it does not affect locus $i+1$.

are w,m, then the probability that GMs passes a mutated allele to a child is 1. Note that when $\beta = 0$ the two models are identical. An illustration of the extended model is given in Figure 3. Assume that locus i is the trait locus. If $W = 1$, $S_{i,c,p} = 0$ and $G_{i,a,p} = w$, then $G_{i,c,p}$ is equal to m with probability β and to w with probability $1-\beta$. In any other case, the probability that $G_{i,c,p}$ is equal to m is the same under both the standard and the extended model. Namely, if $W = 0$ then $G_{i,c,p}$ is equal to m only if $G_{i,a,p} = m$ and $S_{i,c,p} = 0$ or if $G_{i,a,m} = m$ and $S_{i,c,p} = 1$. Consider Figure 2 for an additional illustration. Individuals II-1, II-3, II-5 and II-8 received the paternal haplotype of individual I-1 which contains the trait locus. A value of $\beta = 0.5$ reflects good agreement with the data given $W = 1$, because two of these four individuals have received a mutated allele while the other two have not. Individual II-10 cannot receive a mutated allele, since under the model assumptions GM only occurs in the paternal haplotype of individual I-1. Note that although a GM causing mutation can also affect some of the somatic cells of GMs himself and alter his own phenotype, the proposed model does not account for this rare event.

The variable W determines whether GM occurred, while the parameter β is the conditional GM rate, given that GM occurred. The parameter β can also be described as the percentage of GMs germ cells in which the paternal trait chromosome is mutated, given that GM occurred. Assuming GMs is a founder, it makes no difference if the mutation is placed at the paternal or maternal chromosome, since the choice of which chromosome is the paternal one is arbitrary. Note that the standard model for genetic linkage analysis is equivalent to the extended model when $\beta = 0$.

3.2 A Likelihood Ratio Test for Detecting GM

Consider a pedigree with a GM suspect (GMs) and denote by A the phenotypic data of individuals in the pedigree. For every locus i, we define a likelihood ratio test to determine whether GMs has GM given the phenotypic data A and the genotypic data G of markers that surround locus i. The null hypothesis states that GMs does not have GM while the alternative hypothesis states that GMs has GM. In other words, the null hypothesis states that $\beta = 0$ while the alternative hypothesis states that $\beta > 0$. We reject the null hypothesis of $\beta = 0$ if the phenotypic data A and the genotypic data G provide significant evidence for GM. Denote by θ the recombination frequency between locus i and the trait locus in the event of linkage. The likelihood ratio test is given by

$$\text{GM-LOD}(A,G) \triangleq \log\left[\frac{\Pr\left(A,G|\hat{\beta},\hat{\theta}\right)}{\Pr\left(A,G|\beta_0,\tilde{\theta}\right)}\right] \tag{2}$$

where β_0 denotes the assertion $\beta = 0$, $\left(\hat{\beta},\hat{\theta}\right) = \text{argmax}_{\beta,\theta}\Pr(A,G|\beta,\theta)$ and $\tilde{\theta} = \text{argmax}_{\theta}\Pr(A,G|\beta_0,\theta)$. The conditional recombination frequency θ is maximized under both hypotheses, since it is treated as a nuisance parameter (Ott, 1999, p. 41). The prevalence and penetrance parameters are fixed to a predetermined value under both hypotheses, as typically done in the standard LOD test. Thus, the GM-LOD test has one degree of freedom. To avoid ambiguity, we refer to the likelihood ratio test of linkage (Equation 1) as LOD and to the new test just described (Equation 2) as GM-LOD. The null hypothesis of no GM is rejected if the GM-LOD statistic exceeds a predetermined cutoff value.

Denote by W_0 and W_1 the assertions $W = 0$ and $W = 1$, respectively, and denote by ω the probability $\Pr(W_1)$. An equivalent form of the GM-LOD test is given by

$$\text{GM-LOD}(A,G) \triangleq \log\left[\frac{(1-\omega)\cdot\Pr\left(A,G|\beta_0,\hat{\theta}\right)+\omega\cdot\Pr\left(A,G|W_1,\hat{\beta},\hat{\theta}\right)}{\Pr\left(A,G|\beta_0,\tilde{\theta}\right)}\right]. \tag{3}$$

Equations 2 and 3 are equivalent because $\Pr\left(A,G|W_0,\hat{\theta}\right) = \Pr\left(A,G|\beta_0,\hat{\theta}\right)$ and $\Pr\left(A,G|\hat{\beta},\hat{\theta}\right) = (1-\omega)\cdot\Pr\left(A,G|W_0,\hat{\theta}\right) + \omega\cdot\Pr\left(A,G|W_1,\hat{\beta},\hat{\theta}\right)$. The first of these two equalities holds because the extended model is equivalent to the standard model when either $W = 0$ or $\beta = 0$, as shown in Table 1. The second equality holds because W is by definition independent of the trait parameters β and θ, and the parameter β does not affect the likelihood when $W = 0$ regardless of θ, as shown in Table 1. The likelihoods in Equation 3 are computed via the equality $\Pr(A,G|W,\beta,\theta) = \Pr(L)\cdot\Pr(A,G|L,W,\beta,\theta) + \Pr(U)\cdot\Pr(A,G|U,W,\beta)$, where L and U denote linkage and non-linkage between the tested and trait locus. This equality is derived under the assumption that the events L and U are independent of W and β. Note that the parameter θ by definition only affects the likelihood in the event of linkage. The prior probabilities $\Pr(L)$ and $\Pr(U)$ in the human genome correspond to 0.02 and 0.98, respectively (Elston and Lange, 1975, Ott, 1999, p. 35).

3.3 Statistical Properties of the GM-LOD Statistic

As for all statistical tests, one must choose a cutoff value R such that the null hypothesis of no GM is rejected if the GM-LOD score exceeds R. A cutoff that yields a significance level of 5% is considered the right balance between statistical power and false positive rate. We analytically derived the cutoff needed to obtain a genomewide significance level of 5%. This was done by utilizing the results of (Lander and Kruglyak, 1995) who analytically determined that a cutoff of 3.3 yields a genomewide significance level of 5% for the standard LOD test. We related the probability of obtaining a false positive result in a GM-LOD test to the probability of obtaining a false positive result in a LOD test. From this relationship we derived cutoff values that depend on pedigree properties.

The cutoff required to obtain a fixed genomewide significance level depends on several pedigree properties. When GMs has a single spouse and both GMs and the spouse of GMs are founders, this cutoff depends only on the number of GMs children and the number of such children who are affected. Table 2 gives the cutoffs required for obtaining a 5% genomewide significance level for a fully penetrant trait in such a pedigree, when the allele prevalence and the prior probability of GM are equal to 0.1% and when the affection status of GMs is unknown. The rows correspond to the number of children GMs has, while the columns correspond to the number of children who are affected. For example, when GMs has 4 children and 2 of them are affected, Table 2 shows that the appropriate cutoff is 1.34. The full analysis of the GM-LOD significance level is given in the Appendix.

One can also evaluate whether a genetic trait originated by GM via Bayesian means. The posterior probability of GM, $\Pr\left(W = 1|A,G,\hat{\beta},\hat{\theta}\right)$, is the probability that

Table 2: Cutoffs required to obtain a 5% GM-LOD genomewide significance level.

#GMs	#GMs affected children									
children	1	2	3	4	5	6	7	8	9	10
2	1.34	1.34								
3	1.41	1.34	1.34							
4	1.56	1.34	1.34	1.34						
5	1.75	1.38	1.34	1.34	1.35					
6	1.96	1.48	1.34	1.34	1.34	1.35				
7	2.18	1.62	1.37	1.34	1.34	1.34	1.36			
8	2.42	1.78	1.44	1.34	1.34	1.34	1.34	1.38		
9	2.67	1.96	1.55	1.36	1.34	1.34	1.34	1.34	1.42	
10	2.92	2.16	1.69	1.42	1.34	1.34	1.34	1.34	1.34	1.46

All values computed assuming full penetrance and that GMs has a single unaffected spouse.

GMs has GM given the phenotypic data A and the genotypic data G of markers that surround the tested locus. By using Bayes formula and some algebra, this probability is bounded by

$$\Pr\left(W = 1 \,|\, A, G, \hat{\beta}, \hat{\theta}\right) \geq 1 - (1-\omega) \,/\, 10^{\text{GM-LOD}(A,G)} \tag{4}$$

where ω is the prior probability of GM, defined by $\omega = \Pr(W = 1)$. A GM-LOD score greater than $\log[20(1-\omega)]$ guarantees that the posterior probability of GM is greater than 95%. Specifically, a GM-LOD score of $\log(20) \approx 1.3$ guarantees a suitable posterior probability of GM regardless of the prior ω. This value reflects good agreement with typical scenarios given in Table 2. The derivation of Equation 4 is as follows. Denote by W_0 and W_1 the assertions $W = 0$ and $W = 1$, respectively. The following inequality holds:

$$\begin{aligned}
&\Pr\left(W_1 \,|\, A, G, \hat{\beta}, \hat{\theta}\right) \\
&= \frac{\omega \cdot \Pr\left(A, G \,|\, W_1, \hat{\beta}, \hat{\theta}\right)}{(1-\omega) \cdot \Pr\left(A, G \,|\, \beta_0, \hat{\theta}\right) + \omega \cdot \Pr\left(A, G \,|\, W_1, \hat{\beta}, \hat{\theta}\right)} \qquad \text{(5a)} \\
&= \frac{10^{\text{GM-LOD}(A,G)} - (1-\omega) \cdot \dfrac{\Pr\left(A, G \,|\, \beta_0, \hat{\theta}\right)}{\Pr\left(A, G \,|\, \beta_0, \tilde{\theta}\right)}}{10^{\text{GM-LOD}(A,G)}}
\end{aligned}$$

$$\geq \frac{10^{\text{GM-LOD}(A,G)} - (1-\omega)}{10^{\text{GM-LOD}(A,G)}} \tag{5b}$$

$$= 1 - (1-\omega) \,/\, 10^{\text{GM-LOD}(A,G)}.$$

Equation 5a follows from Bayes rule. The derivation of the denominator in Equation 5a is similar to the derivation of the enumerator in Equation 3. Equation 5b follows because $\tilde{\theta}$ by definition maximizes the likelihood $\Pr(A,G|\beta_0,\theta)$ and thus the ratio $\Pr\big(A,G|\beta_0,\hat{\theta}\big)/\Pr\big(A,G|\beta_0,\tilde{\theta}\big)$ is bounded by 1. The other equalities follow by term rearrangements and by using the definition of GM-LOD(A,G) given in Equation 3. In summary, Bayesian analysis justifies a cutoff of 1.3 for GM-LOD.

3.4 Evaluation of the Test

We evaluated the GM-LOD test empirically via simulations of random pedigrees affected with a genetic trait. In some pedigrees the trait originated by GM and in others it was inherited. We computed the GM-LOD score of each generated pedigree and estimated the significance and power of the test. We also tested the power of our model to identify linkage.

3.4.1 Tools and Methods

We generated 2000 random pedigree structures with up to five generations and up to 40 individuals in each generation, where one of the individuals in generation I has GM. For each pedigree structure, we generated two sets of random marker data consisting of 8 SNPs 5 cM apart. The first set was generated using a fully penetrant dominant MOI and $\beta = 0.5$, and the second set was generated using a reduced penetrance dominant MOI and $\beta = 0$. Thus, no pedigree displays full correlation between the genotype and the phenotype. In all simulated pedigrees the affection status of GMs was unknown. The trait prevalence and the prior probability ω of GM were both fixed at 0.1%. The penetrance parameter λ was the one that maximized the likelihood of the phenotypic data under the standard model. The computations were performed by representing the pedigree as a Bayesian Network (Jensen, 1996, Pearl, 1988) similar to the one described in (Fishelson and Geiger, 2002) and applying an efficient sum of products algorithm.

3.4.2 Significance Evaluation

We evaluated the significance of GM-LOD under linkage to provide empirical support to our analytically computed cutoffs. Recall that in the event of linkage be-

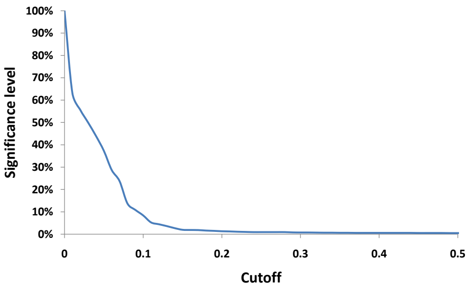

Figure 4: GM-LOD significance level given a cutoff, when the tested and trait locus are linked.

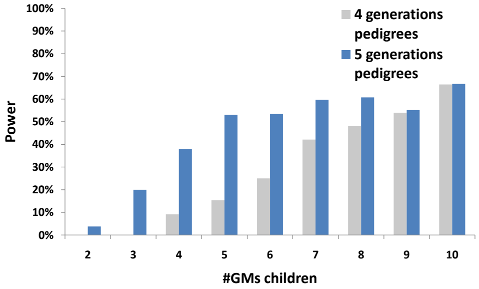

Figure 5: GM-LOD power as a function of the number of GMs children and the number of generations in the pedigree

tween the trait and tested locus, a false positive result is obtained if the GM-LOD score exceeds the cutoff but GMs does not have GM. We tested the false positive rate of the GM-LOD test under this scenario. We examined both pedigrees in which the trait originated by inheritance, and pedigrees in which the trait originated by GM from an individual other than GMs. The results are shown in Figure 4, which shows that a cutoff value as low as 0.2 yields a 5% false positive rate in the event of linkage. Thus, the cutoffs in Table 2 guarantee a 5% significance level under linkage. In the Appendix, it is shown analytically that these cutoffs also guarantee a 5% genomewide significance level under non-linkage. Taken together, the simulations and the analytical results certify that the cutoffs in Table 2 provide a 5% genomewide significance level.

3.4.3 Power Evaluation

We evaluated the power of the GM-LOD test by computing the percentage of positive results obtained when GMs has GM and the trait and tested locus are linked, using a cutoff of 1.3. Figure 5 demonstrates that the power to detect GM grows with the number of GMs children and with the number of tested generations. For 5 generations pedigrees, the power is greater than 50% when GMs has 5 children or more.

3.4.4 Parameters Sensitivity

We tested the sensitivity of GM-LOD power to the ratio ω/γ, where ω is the prior probability of GM and γ is the trait prevalence. Higher values of this ratio indicate

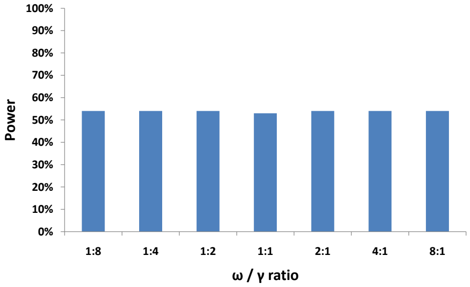

Figure 6: GM-LOD power using various values of the ratio ω/γ for pedigrees with 5 generations and 6 GMs children.

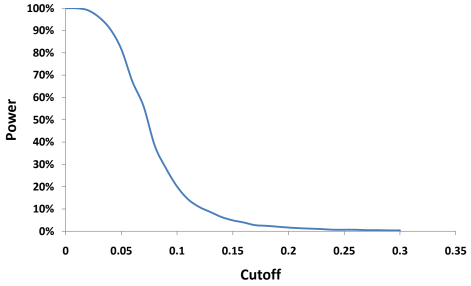

Figure 7: GM-LOD power given a cutoff when using phenotypic data only.

that a priori the trait is more likely to have originated by GM, while lower values indicate it is more likely to be inherited. For each value of the ratio, an appropriate cutoff can be computed. The power of the test is then evaluated via the percentage of GM-LOD scores that exceed the cutoff when using this ratio. Figure 6 shows that the power to detect GM is constant wrt to the ratio ω/γ. The GM-LOD test is therefore robust with regards to this ratio.

3.4.5 Testing for GM Using Phenotypic Data Only

We tested whether GM can be identified using phenotypic data only. We used a likelihood ratio test which is similar to the GM-LOD test, but uses phenotypic data only. The test is given by $\log\left[(1-\omega)+\omega\cdot\Pr\left(A\,|\,W_1,\hat{\beta}\right)/\Pr(A\,|\,\beta_0)\right]$, where $\hat{\beta} = \text{argmax}_{\beta}\Pr(A\,|\,W_1,\beta)$. Figure 7 shows that this test has no power to detect GM with a cutoff greater than 0.3. This confirms the argument in Section 2.4 that it is rarely possible to detect GM by using phenotypic data alone.

3.4.6 Testing for Linkage

When one has rejected the null hypothesis of no GM, one should test for linkage using a LOD test. A LOD test (Equation 1) should be taken after evidence for GM is found, to directly test the hypothesis of linkage in the suspect region, with the variable W set to 1 and the parameter β set to the maximum likelihood estimate. We examined the power of our statistical model to identify linkage in the presence of GM versus other methods.

DOI: 10.2202/1544-6115.1709

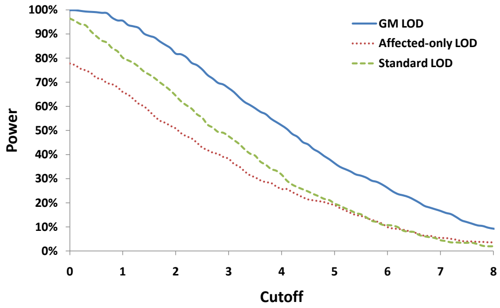

Figure 8: LOD power in the presence of GM, using different genetic models

We compare three different methods of testing for linkage when GM is present. The first method computes the standard LOD test, using reduced penetrance. The second method performs an affected-only analysis by treating the phenotypic data of all unaffected individuals in the pedigree as unknown and computing the standard LOD test. The third method computes a LOD score with the parameter β which maximizes the likelihood and the variable W set to 1. All three methods assume a mutated allele prevalence of 0.1% and a dominant MOI. The power of the three methods to detect linkage is given in Figure 8, which shows that the power to detect linkage is higher when our model is utilized. For a cutoff of LOD = 3.3, the power is 20% higher than for the other methods.

4 A Case Study

We used the GM-LOD statistic to test for the occurrence of GM in a pedigree in which GM has been hypothesized (Genzer-Nir et al., 2010). The pedigree, shown in Figure 9, is affected with the MDN syndrome. Genzer-Nir et al. have found evidence of linkage between a certain locus and this syndrome. Furthermore, they observed that the pedigree can be divided into two sub-pedigrees, one of which exhibits evidence of linkage while the second does not. This pattern is similar to the one shown in Figure 2. Consequently, Genzer-Nir et al. hypothesized that a mutated allele has been introduced into the pedigree by GM. We provide solid statistical evidence for this hypothesis using the GM-LOD test.

4.1 Tools and Methods

In order to perform the GM-LOD test, we removed individuals who are not descendants of the GM suspect and have no descendants in common with him from the

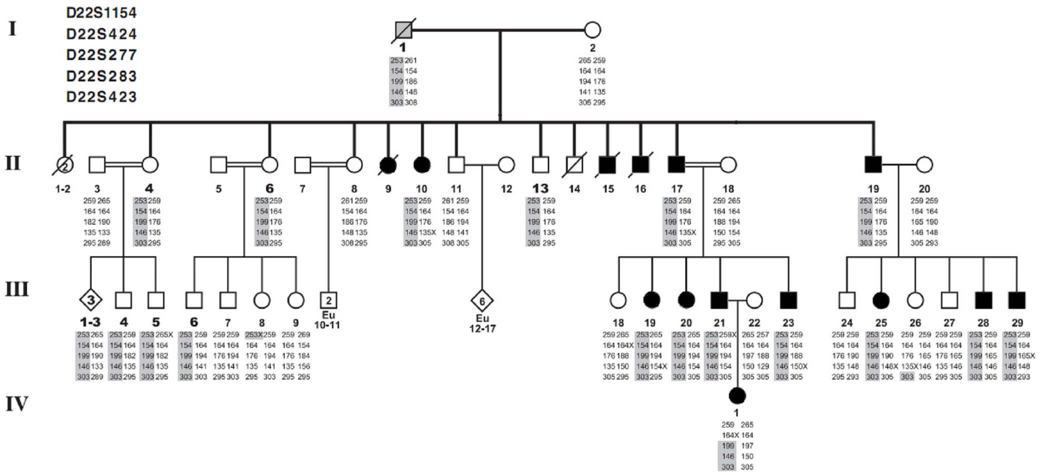

Figure 9: The pedigree studied in Genzer-Nir et al. with affected individuals shaded in black. A haplotype with five markers is shows for each individual, with the marker names shows on the top left. Individuals with irrelevant haplotypes are denoted with E_u. The haplotype suspected of being linked to the trait is shaded in gray. The ten unaffected individuals who carry the suspected haplotype are emphasized as bold numbers. The image is adapted from Genzer-Nir et al.

original pedigree described in Genzer-Nir et al. We also removed the parents of the GM suspect and his spouse. All these individuals are untyped and unaffected, and thus their removal does not increases the likelihood of GM. The resulting pedigree contains 52 individuals only, versus 84 individuals in the original pedigree, rendering the GM-LOD computation less computationally demanding. The affection status of the GM suspect was marked as unknown instead of unaffected to avoid biasing the test in favor of GM.

We conducted a genomewide screen for GM by evaluating the GM-LOD score at loci 10 cM apart, with a window of five adjacent markers for each multipoint computation. We used a value of 0.1% mutated allele prevalence as in Genzer-Nir et al. and a value of 0.1% for the prior probability ω of GM. Thus, the prior probabilities that the mutated allele originated by GM and did not originate by GM are equal. We used a fully penetrant dominant MOI, since this is the maximum likelihood estimate of the penetrance given the phenotypic data of the pedigree. The stringent cutoff required to establish GM in the MDN pedigree is 1.40.

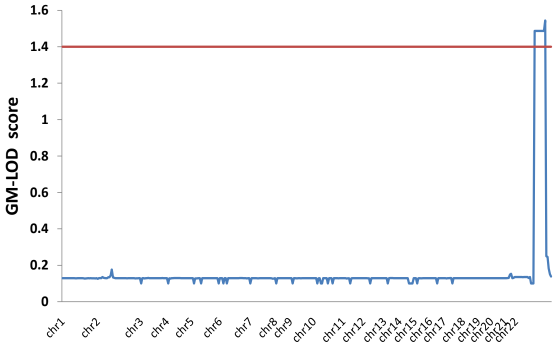

Figure 10: GM-LOD scores for the MDN pedigree. The horizontal line is the stringent cutoff of 1.40.

4.2 Results

The GM-LOD values across all loci tested are given in Figure 10. The GM-LOD values across chromosome 22q12.3-13.1 show significant evidence of GM. The GM-LOD score across this chromosomal segment is 1.49. This value exceeds the required cutoff of 1.40 and corresponds to a p-value of 0.034. In all other chromosomal segments, the GM-LOD value is close to zero. This supports the hypothesis raised by Genzer-Nir et al. that the trait locus has been introduced into the pedigree by GM and is located in chromosome 22q12.3-13.1.

After determining that the suspect region indeed shows evidence of GM, we tested for linkage in this region using a LOD test (Equation 1) carried with the extended statistical model. The variable W was set to 1, indicating that individual I-1 has GM, and the parameter β was fixed at 0.61, which is the computed MLE of β for this pedigree (this value of β reflects good agreement with the fact that six out of the nine GMs children carrying the gray-shaded haplotype are affected, assuming all affected children carry this haplotype). The cutoff required to establish linkage with this LOD test is 3.3, the same as in the standard LOD test. The LOD score across the suspect region is 3.76, indicating linkage. For comparison, we also computed the LOD score obtained using the standard model ($\beta = 0$) and obtained a maximal LOD score of -0.02 in the suspect region. Thus, the standard LOD test fails to detect linkage in this region. No other region yields a significant LOD score, as reported by Genzer-Nir et al. Note that since the GM-LOD score indicates that there is GM only in one specific region, testing for linkage in other regions should still be carried with the standard LOD test.

5 Discussion

We developed a statistical test to determine whether a genetic trait has been introduced into a pedigree by GM. We derived analytical cutoffs that guarantee a 5% genomewide significance level for this test under non-linkage, and demonstrated via simulations that these cutoffs guarantee a 5% significance level under linkage as well. Thus, these cutoffs guarantee a 5% genomewide significance level. We focused on traits that follow a dominant mode of inheritance (MOI), since all known genetic traits caused by GM that we are aware of follow this MOI (Zlotogora, 1998). Nevertheless, our extended model can be used with any MOI. Existing software for LOD score computation can easily be modified to compute GM-LOD scores, as both tests compute similar likelihoods, with only a slight modification to the genetic model.

The GM-LOD test is meant to be used when standard genetic linkage analysis fails to detect linkage or provides inconclusive results. Parametric affected-only (AF) analysis, where the affection status of unaffected individuals is regarded as unknown, is an alternative method to detect linkage in such scenarios. This method circumvents the difficulty of sib-pairs with different phenotypes who share the same markers around the trait locus. Genzer-Nir et al. have shown that there is evidence of linkage in the MDN pedigree by using this method. However, an AF analysis disregards much of the pedigree data, and is thus less powerful than our method, as shown in Section 3.4.6. One may also employ non-parametric linkage analysis methods, which compute allele-sharing statistics between affected pairs in the pedigree. (e.g. Kruglyak et al., 1996). We have not been able to test the power of these methods to detect linkage in the presence of GM, since the pedigrees we analyzed are too large for computer packages that perform non-parametric linkage analysis, such as GENEHUNTER (Kruglyak et al., 1996), Merlin (Abecasis et al., 2002) and Allegro (Gudbjartsson et al., 2005). Nevertheless, these methods are inevitably less powerful than parametric methods that use an appropriate explicit model (Strauch et al., 2000). Other non-parametric methods, which take non-affected individuals into consideration (e.g. Blackwelder et al., 1985, Commenges, 1994, Green and Montasser, 1988) are bound to run into the same difficulty of siblings with different phenotypes who share the same markers around the trait locus.

Model selection tests may also be developed to detect other types of non-standard inheritance patterns. For example, allelic and locus heterogeneity (Benayoun et al., 2009), modifier genes (Dipple and McCabe, 2000) and uniparental disomy (Kotzot, 2001) are biological phenomena that do not follow the assumptions of classic genetic linkage analysis. As genotyping becomes widely affordable more variates of non-standard genetic diseases will be discovered. Our work demonstrates that adapting the genetic model is beneficial to deal with such unusual cases.

Appendix

Statistical Properties of the GM-LOD Statistic

Here we derive the cutoff which guarantees a 5% significance level for the GM-LOD statistic, as described in Section 3.3. Recall that the GM-LOD test is defined by

$$\text{GM-LOD}(A,G) \triangleq \log\left[\frac{(1-\omega)\cdot \Pr\left(A,G\,|\,\beta_0,\hat{\theta}\right) + \omega\cdot \Pr\left(A,G\,|\,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr\left(A,G\,|\,\beta_0,\tilde{\theta}\right)}\right]. \qquad \text{(A1)}$$

The terms in Equation A1 have all been defined previously. In short, recall that A and G are the phenotypic and the marker data of the pedigree, respectively, W is an indicator for whether GMs has GM, β is the GM rate given that $W = 1$, ω is the prior probability $\Pr(W = 1)$, θ is the recombination frequency between the trait and the tested locus in the event of linkage, β_0 denotes the assertion $\beta = 0$, $\tilde{\theta} = \text{argmax}_{\theta}\Pr(A,G\,|\,\beta_0,\theta)$ and the parameters $\hat{\beta}$ and $\hat{\theta}$ are defined as $\left(\hat{\beta},\hat{\theta}\right) = \text{argmax}_{\beta,\theta}\Pr(A,G\,|\,\beta,\theta)$.

In the GM-LOD test, the null hypothesis H_0 holds when $\beta = 0$ and its negation H_1 holds when $\beta > 0$. We analyze the GM-LOD statistic by defining two alternative hypotheses denoted by H_0^* and H_1^*. The hypothesis H_0^* holds when either $\beta = 0$ or the tested and trait locus are unlinked. The hypothesis H_1^* holds when $\beta > 0$ and the two loci are linked. This definition encodes the fact that there is no power to detect GM when the tested and trait locus and unlinked, since the genotypic data used in the test is then independent of the trait and thus does not affect the likelihood ratio. Because human families are typically small, there is no power to detect GM by using phenotypic data alone. When conducting a genome-wide screen for GM, and assuming that the trait originated by GM, the GM-LOD score is expected to exceed the cutoff only in the region that is linked to the trait. The subsequent analysis determines the cutoff needed to ensure that the probability of falsely rejecting H_0^* is smaller than 5%. This cutoff ensures a 5% probability of falsely rejecting H_0 as well, since when one rejects H_0^* one also rejects H_0.

The analysis of the significance level for the GM-LOD statistic is carried as follows. In a genomewide screen, a GM-LOD test is performed for each locus. The tested locus in each test can be either unlinked or linked to the trait locus. We first derive a cutoff value which guarantees a false positive rate smaller than 5% for tests performed on unlinked loci. In other words, we derive an upper bound on the cutoff needed to obtain a 5% significance level in a screen performed on all regions that are unlinked to the trait locus. Since the prior probability of non-linkage in the human genome is 98%, this significance level accounts for 98% of all tests performed in a genomewide screen.

The upper bound on the cutoff is derived via the following inequality:

$$\text{GM-LOD}(A,G) \leq \log\left[(1-\omega)+\omega\cdot\frac{\Pr(L)}{\Pr(U)}\cdot K\cdot 10^{\text{LOD}\left(A,G|\hat{\beta}\right)}+\omega\cdot K\right]. \quad \text{(A2)}$$

The quantities in Equation A2 are the following. The ratio $\Pr(L)/\Pr(U)$ is the prior probability of linkage versus no linkage. The term $\text{LOD}\left(A,G|\hat{\beta}\right)$ corresponds to the standard LOD test carried using the extended model, with the variable $W = 1$ and the parameter β which maximizes the likelihood $\Pr(A,G\,|\,W_1,\beta)$. It is defined by $\text{LOD}\left(A,G|\hat{\beta}\right) \triangleq \log\left[\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\theta^*\right)/\Pr\left(A,G\,|\,U,W_1,\hat{\beta}\right)\right]$, where θ^* is the maximum in the range $[0,0.5)$. The term K denotes the likelihood ratio $\Pr\left(A\,|\,W_1,\hat{\beta}\right)/\Pr\left(A\,|\beta_0\right)$. It is bounded by

$$K \leq \max_T\left\{\frac{\Pr\left(A_C,T\,|\,W_1,\hat{\beta}\right)}{\Pr\left(A_C,T\,|\beta_0\right)}\;\middle|\;\Pr(A_S\,|\,T) > 0\right\}. \quad \text{(A3)}$$

In Equation A3, C is the set of individuals that consists of GMs and every other parent of a child of GMs, and S is the set of individuals not in C who have a parent in C or a child in C or a child with a parent in C. The variables A_C and A_S are the phenotypic data of individuals in sets C and S, respectively. The variable T is a consistent assignment of trait alleles to individuals in S, where each individual receives zero, one or two mutated alleles. For example, in the pedigree given in Figure 2, A_C corresponds to the affection status of individuals I-1 and I-2, and A_S is the affection status of individuals in generation II. A consistent assignment T for example is one mutated allele to individuals II-1 and II-8 and no mutated alleles to individuals II-3, II-5 and II-10. Note that Equation A3 depends only on the phenotypic data of individuals in sets C and S and on the set of consistent assignments T. Therefore, Equation A3 provides a bound which holds for every pedigree with the same values of A_C and A_S, regardless of the phenotypic data of other individuals. The derivation of Equations A2 and A3 is given in the next section.

Equation A2 shows that $\text{GM-LOD}(A,G)$ is bounded by a monotone function of $\text{LOD}\left(A,G|\hat{\beta}\right)$. The test $\text{LOD}\left(A,G|\hat{\beta}\right)$ tests for linkage, and thus the probability of it exceeding a given cutoff in the absence of linkage is bounded. Therefore, the false positive rate of a GM-LOD test on an unlinked locus is related to the false positive rate of a LOD test. The pointwise null distribution of the test $\text{LOD}\left(A,G|\hat{\beta}\right)$ is the same as the pointwise null distribution of the standard LOD test, $\text{LOD}(A,G|\beta_0)$, since maximizing the null and the alternative hypotheses of a likelihood ratio test over an additional parameter β does not affect the null distribution of the test. Thus, the probability of exceeding a cutoff R in the standard LOD test given in (Lander and Botstein, 1989, Lander and Kruglyak, 1995) also holds for the test $\mathrm{LOD}\left(A, G|\hat{\beta}\right)$. We conclude that Equation A2 provides a bound on the probability of obtaining a false positive result when the tested locus is unlinked to the trait locus, which covers approximately 98% of all tests.

Equation A2 does not provide a sufficiently tight bound on loci linked to the trait locus, because in the event of linkage, $\mathrm{LOD}\left(A, G|\hat{\beta}\right)$ is not expected to be low. The cutoffs computed with Equations A2 and A3 will continue to guarantee a 5% genomewide significance level, also in the case of linkage, if the false positive rate is shown to be smaller under linkage than under non-linkage when using the cutoff determined by Equation A2. This claim can be proved in the case where all phases are known and the trait is fully penetrant, but a proof of the general case remains an open mathematical problem. The justification for this claim is as follows. Under non-linkage, the genotypic and the phenotypic data are independent, and the genotypic data is thus not affected by the value of β. Consequently, the distribution of the phenotypic and genotypic data is less sensitive to the value of β under non-linkage, and thus the probability of obtaining a false positive result is higher under non-linkage. We determined via simulations that when using the cutoffs determined by Equations A2 and A3, the false positive rate in the event of no GM and linkage is smaller than 5% as required.

Proofs of Inequalities

Propositions 1 and 2 prove Equations A2 and A3, respectively.

Proposition 1. *Equation* A2 *holds.*

Proof. Equation A2 is obtained via the following two inequalities:

$$\frac{\Pr\left(A, G|\beta_0, \hat{\theta}\right)}{\Pr\left(A, G|\beta_0, \tilde{\theta}\right)} \leq 1. \tag{A4}$$

$$\frac{\Pr\left(A, G|W_1, \hat{\beta}, \hat{\theta}\right)}{\Pr\left(A, G|\beta_0, \tilde{\theta}\right)} \leq \frac{\Pr(L)}{\Pr(U)} \cdot K \cdot 10^{\mathrm{LOD}\left(A, G|\hat{\beta}\right)} + K. \tag{A5}$$

Equation A2 immediately follows by using Equations A4 and A5 on the definition of GM-LOD(A, G) given in Equation A1. Equation A4 holds because $\tilde{\theta}$ by definition maximizes the likelihood $\Pr(A,G|\beta_0,\theta)$. Equation A5 is derived as follows.

$$\frac{\Pr\left(A,G\,|\,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr\left(A,G|\beta_0,\tilde{\theta}\right)}$$

$$= \frac{\Pr(L)\cdot\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)+\Pr(U)\cdot\Pr\left(A,G\,|\,U,W_1,\hat{\beta}\right)}{\Pr(L)\cdot\Pr\left(A,G\,|\,L,\beta_0,\tilde{\theta}\right)+\Pr(U)\cdot\Pr(A,G\,|\,U,\beta_0)}$$

$$\leq \frac{\Pr(L)\cdot\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr(U)\cdot\Pr(A,G\,|\,U,\beta_0)}+\frac{\Pr(U)\cdot\Pr\left(A,G\,|\,U,W_1,\hat{\beta}\right)}{\Pr(U)\cdot\Pr(A,G\,|\,U,\beta_0)} \tag{A6a}$$

$$= \frac{\Pr(L)\cdot\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr(U)\cdot\Pr(G\,|\,U)\cdot\Pr(A\,|\,U,\beta_0)}+\frac{\Pr(U)\cdot\Pr(G\,|\,U)\cdot\Pr\left(A\,|\,U,W_1,\hat{\beta}\right)}{\Pr(U)\cdot\Pr(G\,|\,U)\cdot\Pr(A\,|\,U,\beta_0)} \tag{A6b}$$

$$= \frac{\Pr(L)\cdot\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr(U)\cdot\Pr(G)\cdot\Pr(A\,|\beta_0)}+\frac{\Pr(U)\cdot\Pr(G)\cdot\Pr\left(A\,|\,W_1,\hat{\beta}\right)}{\Pr(U)\cdot\Pr(G)\cdot\Pr(A\,|\beta_0)} \tag{A6c}$$

$$= \frac{\Pr(L)}{\Pr(U)}\cdot K\cdot\frac{\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr(G)\cdot\Pr\left(A\,|\,W_1,\hat{\beta}\right)}+K$$

$$= \frac{\Pr(L)}{\Pr(U)}\cdot K\cdot\frac{\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\hat{\theta}\right)}{\Pr\left(A,G\,|\,U,W_1,\hat{\beta}\right)}+K \tag{A6d}$$

$$\leq \frac{\Pr(L)}{\Pr(U)}\cdot K\cdot\frac{\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\theta^*\right)}{\Pr\left(A,G\,|\,U,W_1,\hat{\beta}\right)}+K \tag{A6e}$$

$$= \frac{\Pr(L)}{\Pr(U)}\cdot K\cdot 10^{\mathrm{LOD}\left(A,G|\hat{\beta}\right)}+K. \tag{A6f}$$

Equation A6a is obtained by removing the first term from the denominator. Equation A6b follows because A and G are independent given the assertion U of non-linkage regardless of other parameters, and G is independent of β when A is not given. Equation A6c follows because the likelihood of A is unaffected by U when G is not given and vice versa. Equation A6d follows because A and G are independent given U, regardless of W and β. Equation A6e follows because θ^* by definition maximizes the likelihood $\Pr\left(A,G\,|\,L,W_1,\hat{\beta},\theta\right)$. Finally, Equation A6f is obtained by using the definition of $\mathrm{LOD}\big(A,G\,|\,\hat{\beta}\big)$. The other equalities are term rearrangements. □

Proposition 2. *Equation* A3 *holds.*

Proof. We split the individuals in the pedigree into three sets denoted by C, S and R. Recall that C contains GMs and every other parent of a child of GMs and that S contains individuals not in C who have a parent in C or a child in C or a child with a parent in C, and define R as the set containing the rest of the individuals. Further recall that A_C and A_S denote the phenotypic data of individuals in sets C and S, respectively, and denote A_R as the phenotypic data of individuals in set R. Finally, recall that T corresponds to an assignment of trait alleles to individuals in S and define M by $M = \max_T \Big\{ \Pr\big(A_C, T\,|\,W_1, \hat{\beta}\big) / \Pr\big(A_C, T\,|\,\beta_0\big) \;\Big|\; \Pr(A_S\,|\,T) > 0 \Big\}$.

The term K is bounded by M as follows.

$$K \triangleq \frac{\Pr\big(A\,|\,W_1, \hat{\beta}\big)}{\Pr(A\,|\,\beta_0)} = \frac{\Pr\big(A_C, A_S, A_R\,|\,W_1, \hat{\beta}\big)}{\Pr(A_C, A_S, A_R\,|\,\beta_0)} \tag{A7a}$$

$$= \frac{\sum_T \Pr\big(A_C, T\,|\,W_1, \hat{\beta}\big) \cdot \Pr(A_S\,|\,T) \cdot \Pr(A_R\,|\,T, A_S)}{\sum_T \Pr(A_C, T\,|\,\beta_0) \cdot \Pr(A_S\,|\,T) \cdot \Pr(A_R\,|\,T, A_S)} \tag{A7b}$$

$$\leq \frac{\sum_T \left[M \cdot \Pr(A_C, T\,|\,\beta_0)\right] \cdot \Pr(A_S\,|\,T) \cdot \Pr(A_R\,|\,T, A_S)}{\sum_T \Pr(A_C, T\,|\,\beta_0) \cdot \Pr(A_S\,|\,T) \cdot \Pr(A_R\,|\,T, A_S)} = M. \tag{A7c}$$

Equation A7a follows because the phenotypic data A is composed of A_C, A_S and A_R. Equation A7c follows according to the definition of M. Equation A7b, which is the essence of the proof, follows because $\Pr(A_S\,|\,T) = \Pr\big(A_S\,|\,T, A_C, W_1, \hat{\beta}\big)$ and $\Pr(A_R\,|\,T, A_S) = \Pr\big(A_R\,|\,T, A_S, A_C, W_1, \hat{\beta}\big)$. In other words, given the genotypic trait data T, the probability of the phenotypic data A_S and A_R is unaffected by the quantities A_C, W and β. The direct proof shows that the probability of A_S and A_R given a specific value of T is constant for every given combination of A_C, W and β. This claim can also be proved in a Bayesian networks terminology via the definition of d-separation (Pearl, 1988). □

References

Abecasis, G., S. Cherny, W. Cookson, and L. Cardon (2002): "Merlin-rapid analysis of dense genetic maps using sparse gene flow trees," *Nature genetics*, 30, 97–101.

Barbosa, R., F. Vargas, F. Aguiar, S. Ferman, E. Lucena, C. Bonvicino, and H. Seuánez (2008): "Hereditary retinoblastoma transmitted by maternal germline mosaicism," *Pediatric Blood & Cancer*, 51, 598–602.

Benayoun, L., R. Spiegel, N. Auslender, A. Abbasi, L. Rizel, Y. Hujeirat, I. Salama, H. Garzozi, S. Allon-Shalev, and T. Ben-Yosef (2009): "Genetic heterogeneity in two consanguineous families segregating early onset retinal degeneration: The pitfalls of homozygosity mapping," *American Journal of Medical Genetics Part A*, 149, 650–656.

Blackwelder, W., R. Elston, and D. Rao (1985): "A comparison of sib-pair linkage tests for disease susceptibility loci," *Genetic Epidemiology*, 2, 85–97.

Borecki, I. and J. Rice (2010): "Linkage analysis of discrete traits." *CSH protocols*, 2010.

Choi, H., B. Lee, H. Cho, K. Moon, I. Ha, M. Nagata, Y. Choi, and H. Cheong (2008): "Familial focal segmental glomerulosclerosis associated with an ACTN4 mutation and paternal germline mosaicism," *American Journal of Kidney Diseases*, 51, 834–838.

Commenges, D. (1994): "Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic," *Genetic epidemiology*, 11, 189–200.

Dipple, K. and E. McCabe (2000): "Modifier genes convert "simple" Mendelian disorders to complex traits." *Molecular genetics and metabolism*, 71, 43.

Edwards, J. (1989): "Familiarity, recessivity and germline mosaicism," *Annals of human genetics*, 53, 33–47.

Elston, R. and K. Lange (1975): "The prior probability of autosomal linkage," *Annals of Human Genetics*, 38, 341–350.

Elston, R. and J. Stewart (1971): "A general model for the genetic analysis of pedigree data," *Human Heredity*, 21, 523–542.

Erickson, R. and S. Lewis (1995): "The new human genetics," *Environmental and molecular mutagenesis*, 25, 7–12.

Fabrizi, G., M. Ferrarini, T. Cavallaro, L. Jarre, A. Polo, and N. Rizzuto (2001): "A somatic and germline mosaic mutation in MPZ/P0 mimics recessive inheritance of CMT1B," *Neurology*, 57, 101.

Fishelson, M. and D. Geiger (2002): "Exact genetic linkage computations for general pedigrees," *Bioinformatics*, 18, S189.

DOI: 10.2202/1544-6115.1709

Friedman, N., D. Geiger, and N. Lotner (2000): "Likelihood computations using value abstraction," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, Citeseer, 192–200.

Genzer-Nir, M., M. Khayat, L. Kogan, H. Cohen, M. Hershkowitz, D. Geiger, and T. Falik-Zaccai (2010): "Mammary-digital-nail (MDN) syndrome: a novel phenotype maps to human chromosome 22q12. 3–13.1," *European Journal of Human Genetics*.

Green, J. and M. Montasser (1988): "HLA haplotype discordance," *Biometrics*, 44, 941–950.

Grimm, T., B. Müller, C. Müller, and M. Janka (1990): "Theoretical considerations on germline mosaicism in Duchenne muscular dystrophy." *Journal of medical genetics*, 27, 683.

Gropman, A. and D. Adams (2007): "Atypical patterns of inheritance," in *Seminars in Pediatric Neurology*, volume 14, Elsevier, volume 14, 34–45.

Gudbjartsson, D., T. Thorvaldsson, A. Kong, G. Gunnarsson, and A. Ingolfsdottir (2005): "Allegro version 2," *Nature genetics*, 37, 1015–1016.

Hartl, D. (1971): "Recurrence risks for germinal mosaics." *American Journal of Human Genetics*, 23, 124.

Jeanpierre, M. (1992): "Germinal mosaicism and risk calculation in X-linked diseases." *American journal of human genetics*, 50, 960.

Jensen, F. (1996): *An introduction to Bayesian networks*, volume 210, UCL press London.

Khan, A., D. Khalil, L. Al Sharif, F. Al-Ghadhfan, and N. Al Tassan (2010): "Germline mosaicism for KIF21A mutation (p. R954L) mimicking recessive inheritance for congenital fibrosis of the extraocular muscles," *Ophthalmology*, 117, 154–158.

Kotzot, D. (2001): "Complex and segmental uniparental disomy (UPD): review and lessons from rare chromosomal complements," *Journal of medical genetics*, 38, 497.

Kruglyak, L., M. Daly, M. Reeve-Daly, and E. Lander (1996): "Parametric and non-parametric linkage analysis: a unified multipoint approach." *American Journal of Human Genetics*, 58, 1347.

Lander, E. and D. Botstein (1989): "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps," *Genetics*, 121, 185.

Lander, E. and P. Green (1987): "Construction of multilocus genetic linkage maps in humans," *Proceedings of the National Academy of Sciences of the United States of America*, 84, 2363.

Lander, E. and L. Kruglyak (1995): "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results," *Nat. Genet.*, 11, 241–247.

Makri, S., N. Clarke, P. Richard, S. Maugenre, L. Demay, G. Bonne, and P. Guicheney (2009): "Germinal mosaicism for LMNA mimics autosomal recessive congenital muscular dystrophy," *Neuromuscular Disorders*, 19, 26–28.

Murphy, E., D. Cramer, R. Kryscio, C. Brown, and E. Pierce (1974): "Gonadal mosaicism and genetic counseling for X-linked recessive lethals." *American Journal of Human Genetics*, 26, 207.

Ott, J. (1999): *Analysis of human genetic linkage*, Johns Hopkins Univ Pr, third edition.

Pauli, S., L. Pieper, J. Häberle, P. Grzmil, P. Burfeind, M. Steckel, U. Lenz, and H. Michelmann (2009): "Proven germline mosaicism in a father of two children with CHARGE syndrome," *Clinical genetics*, 75, 473–479.

Pearl, J. (1988): *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann.

Strauch, K., R. Fimmers, M. Baur, and T. Wienker (2000): "How to model a complex trait," *Human heredity*, 55, 202–210.

Yaron, Y. and A. Orr-Urtreger (2002): "New genetic principles," *Clinical Obstetrics and Gynecology*, 45, 593.

Zlotogora, J. (1998): "Germ line mosaicism," *Human genetics*, 102, 381–386.

DOI: 10.2202/1544-6115.1709