Published in final edited form as:

Ann Hum Genet. 2012 January ; 76(1): 63–73. doi:10.1111/j.1469-1809.2011.00683.x.

On the validity of the likelihood ratio test and consistency of resulting parameter estimates in joint linkage and linkage disequilibrium analysis under improperly specified parametric models

Tero Hiekkalinna^{1,2}, Harald HH Göring³, and Joseph D Terwilliger^{1,2,4,5,6,7}

¹Institute for Molecular Medicine Finland, University of Helsinki, Helsinki Finland ²Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, Finland ³Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA ⁴Department of Psychiatry, Columbia University, New York, NY, USA ⁵Department of Genetics and Development, Columbia University, New York, NY, USA ⁶Columbia Genome Center, Columbia University, New York, NY, USA ⁷Division of Medical Genetics, New York State Psychiatric Institute, New York, NY, USA

Summary

It has been shown that parametric analysis of linkage disequilibrium conditional on linkage using an overly deterministic model can be optimal for family-based association analysis. However if one applies this strategy carelessly there is a risk of false inference. We analyze properties of such likelihood ratio tests when the assumed disease mode-of-inheritance is inaccurate. Under some conditions problems result if one is not careful to consider what null hypothesis is being tested. We show that: (a) tests for which the null hypothesis assumes absence of both linkage and association are independent of the true mode-of-inheritance; (b) LRTs assuming either linkage or association under the null hypothesis may depend on the true mode-of-inheritance, lead to inconsistent parameter estimates, in particular under extremely deterministic models; (c) this problem cannot be eliminated by increasing sample size or adding population controls - as sample size increases, the chance of false positive inference goes to 100%; (d) this issue can lead to systematic false positive inference of association in regions of linkage. This is important because highly-deterministic models are often used intentionally in model-based analyses because they can have more power than the true model, and are implicit in many model-free analysis methods.

Keywords

Likelihood methods; Family-based association; Linkage disequilibrium; Type I error; Bias

Introduction

It is well known that the false positive rate in parametric lod score analysis does not depend asymptotically on the accuracy of the assumed etiological model (Williamson & Amos, 1990), though this may happen only with astronomical sample sizes when the assumed genotype relative risks are small. The type-I error rate of the conventional lod score is

independent of the true etiological model because in the absence of linkage, marker loci segregate randomly in pedigrees, independent of the phenotype, and therefore of the *true* mode-of-inheritance of the trait. Under the null hypothesis, the lod scores can vary stochastically as a function of the *assumed* mode-of-inheritance, though not as a function of the *true* mode of inheritance, such that the type I error rate is independent of whether or not the model used for analysis is misspecified. Furthermore, the lod score distribution is asymptotically independent of both true *and* assumed models in the absence of LD (Williamson & Amos, 1990). This property of linkage analysis has made the many successes of linkage-based gene mapping studies possible, despite the impossibility of accurately specifying the mode-of-inheritance *a priori*. While there are well-known biases and inconsistencies in the recombination fraction estimates, misspecification of parameters typically does not lead to inflated false positive rates (Clerget-Darpoux *et al.*, 1986, Ott, 1985, Ott, 1992) and often power is optimized under inaccurately specified models (Göring & Terwilliger, 2000b, Hiekkalinna *et al.*, 2011a, Hiekkalinna *et al.*, 2011b, Terwilliger, 2001).

It is equally well understood that the null hypothesis in an association test does not depend on the accuracy of the mode of inheritance assumptions, because under the null hypothesis the marker genotypes at any given locus are assumed to be i.i.d. (independent and identically distributed) for all unrelated individuals, irrespective of their trait phenotype. This is true regardless of any assumptions about the relationship between those phenotypes and the underlying trait-predisposing genotypes (unassociated to the marker being studied). For this reason, one does not need to have an accurate model of the genotype-phenotype relationship to conduct a parametric likelihood-based LD analysis, though errors in the model can induce biases in the LD parameter estimates.

If one wishes to perform joint linkage and association analysis using a parametric likelihood analysis in samples derived from multiplex pedigrees and random individuals together, model misspecification can influence the power and bias of parameter estimates, but no inflated type I error rate arises, because founder individuals have random genotypes at markers, and these would segregate randomly within pedigrees, and would thus be independent of the disease-predisposing genotypes, no matter how inaccurate the mode-of-inheritance assumptions would be.

However, when conditionally testing for LD in a genomic region where there has been prior evidence of linkage – the classical problem in Mendelian diseases where one wants to fine map the location of a disease gene – the null hypothesis is no longer one of strict independence of marker locus genotypes and trait phenotypes. Nevertheless, when the disease model is not egregiously wrong, there is generally no inflation of the type I error. However, we have identified some pathological situations under which severely inaccurate models can lead to systematic false positive evidence of LD conditional on linkage, when linkage but no LD is present. In this manuscript we will demonstrate how and why this can happen in practice, warranting caution in interpretation of family-based association results.

We have recently demonstrated that performing a likelihood based analysis under a highly deterministic recessive model (Table 1) can provide a robust and powerful test for association in a mixture of families and unrelated individuals (Hiekkalinna *et al.*, 2011b). Furthermore such tests are consistently more powerful than the widely-used family-based association tests incorporated in the program packages UNPHASED (Dudbridge, 2008), FBAT (Laird *et al.*, 2000, Rabinowitz & Laird, 2000), PLINK (family-based options) (Purcell *et al.*, 2007), TRANSMIT (Clayton, 1999), GENEHUNTER-TDT (Kruglyak *et al.*, 1996), MENDEL (Lange *et al.*, 2001), LAMP (Li *et al.*, 2005, Li *et al.*, 2006) and HHRR (Terwilliger & Ott, 1992), many of which are shown to be invalid tests of LD in the

presence of linkage when parental genotype data is incomplete (Hiekkalinna *et al.*, 2011a). To this end, we investigated the potential effects of model errors in the presence of strong linkage and association to determine under what conditions the conditional inference of LD given linkage would provide a valid test, as it is well known that under the alternative hypothesis of linkage, analysis under inaccurate mode of inheritance assumptions can lead to asymptotic biases in the recombination fraction estimates, and as outlined above, the marker genotypes are not independent of the disease phenotypes under the null hypothesis in the conditional test.

Likelihood model correlating marker genotypes and trait phenotypes in random samples

The observed data in either a linkage or a LD analysis are marker genotypes and trait phenotypes, and the objective of the analysis is to test for correlations between them owing to linkage and/or LD. For simplicity let us start with consideration of a randomly ascertained individual from a population, and compute the likelihood as a function of LD between the observed marker locus genotype and some (unknown) genotype of a locus which influences the observed trait according to some parametric model. In this case, $L \propto P(\mathbf{G}_M, \mathbf{Ph}) = \sum_{\mathbf{G}_D} P(\mathbf{Ph} | \mathbf{G}_D) P(\mathbf{G}_M | \mathbf{G}_D) P(\mathbf{G}_D)$, where $P(\mathbf{Ph} | \mathbf{G}_D)$ are the penetrances for each possible disease-locus genotype, $\mathbf{G}_D \in \{AA, AB, BB, \dots\}$, $P(\mathbf{G}_D)$ are the user-specified population genotype frequencies for each possible genotype $\mathbf{G}_D$, typically computed from user-specified allele frequencies under the assumption of Hardy-Weinberg equilibrium (HWE) (e.g. $P(AA) = p_A^2$, etc.). $P(\mathbf{G}_M | \mathbf{G}_D)$ is a function of the marker-locus allele frequencies in the population and the LD between the disease and marker loci, typically computed again assuming HWE. For example, $P(\mathbf{G}_M = 11 | \mathbf{G}_D = AA) = P^2(1|A)$; $P(\mathbf{G}_M = 12 | \mathbf{G}_D = AB) = P(1|A)P(2|B) + P(1|B)P(2|A)$, etc...). Under the null hypothesis of no LD, the likelihood in a dataset would be maximized under the assumption that $P(1|A) = P(1|B) = P(1)$, $P(2|A) = P(2|B) = P(2)$, etc., while under the alternative hypothesis of LD, they would be estimated freely. In this example, the statistical test of the null hypothesis of no

LD would be of the form $2 \ln \frac{\max_{P(\mathbf{G}_M | \mathbf{G}_D)} L}{\max_{P(\mathbf{G}_M)} L} \sim \chi^2_{(n-1)}$ for a marker locus with n alleles, and a putative disease locus with 2 alleles (Göing & Terwilliger, 2000b).

Likelihood model correlating marker genotypes and trait phenotypes in family samples

The likelihood in family data is computed analogously, except that one must consider all possible genotypes for all individuals in a family jointly, and sum the likelihood over all possible genotype combinations for all individuals in a family. In other words, we compute the same likelihood, $L \propto P(\mathbf{G}_M, \mathbf{Ph}) = \sum_{\mathbf{G}_D} P(\mathbf{Ph}, \mathbf{G}_M | \mathbf{G}_D) = \sum_{\mathbf{G}_D} P(\mathbf{Ph} | \mathbf{G}_D) P(\mathbf{G}_M | \mathbf{G}_D) P(\mathbf{G}_D)$ except that now $\mathbf{G}_D$ is a vector of phased trait-locus genotypes for all individuals in the pedigree jointly, and its probability is a function of disease allele frequencies for founders (individuals who do not have parents in the pedigree), and Mendelian transmission probabilities for all other individuals, conditional on the genotype of their parents. Similarly $P(\mathbf{G}_M | \mathbf{G}_D)$ is a function of the conditional marker allele frequencies (i.e. LD) for founders, and for people with parents in the pedigree, it is a function of the recombination fraction between the marker and disease-predisposing locus (i.e. linkage), conditional on parental genotypes, and the transmission of the disease-locus alleles in vector $\mathbf{G}_D$. Penetrances are independent for each individual under a simple single-locus parametric disease model.

It is important to note that in the likelihood formulations we have used, the only term in which LD or linkage plays a direct role is in $P(\mathbf{G}_M | \mathbf{G}_D)$. The other terms, $P(\mathbf{G}_D)$ and $P(\mathbf{Ph} |$

$\mathbf{G_D}$) serve to weight the linkage and LD information by the probabilities of each possible phased genotype vector in each pedigree. Linkage and LD refer solely to correlations between genotypes of the marker and putative trait locus, and have nothing to do with phenotypes per se. Since the entirety of the linkage and LD information resides in the $P(\mathbf{G_M}|\mathbf{G_D})$ term, whenever the null hypothesis can be formulated as $P(\mathbf{G_M}|\mathbf{G_D}) = P(\mathbf{G_M})$, and the likelihood is maximized over all parameters in these terms (marker allele frequencies, LD, and recombination fractions), the distribution of the test statistic would be independent of the true mode of inheritance of the trait (i.e. whenever H_0 : marker and trait loci are independent). This is true whether or not biologically accurate models are used to compute $P(\mathbf{G_D})$ and $P(\mathbf{Ph}|\mathbf{G_D})$ so long as the same models are used under both null and alternative hypotheses. For this reason, tests of LD on unrelated individuals in this likelihood framework are meaningful, as are tests of linkage which do not assume LD, no matter how inaccurate the analysis model assumptions might be. Equally so, tests of the null hypothesis of “no linkage *and* no association” would be similarly meaningful, as the condition $P(\mathbf{G_M}|\mathbf{G_D}) = P(\mathbf{G_M})$ defines the null hypothesis.

Conditional tests

Since the likelihood of a pedigree is a function of both linkage and LD, it is quite natural to think about testing for them jointly, which for reasons described above would provide an appropriate test of the null hypothesis of “no linkage and no LD”. However, rejecting this simple null hypothesis does not clearly indicate which of those phenomena exists. Parameter estimates may provide a clue, but they do not themselves provide a meaningful way to make formal statistical inferences. To this end, it is of great interest to perform conditional analyses. For example, one may have prior evidence from other sources that there is linkage to a given genomic region, after which it is desired to query the region of linkage to find markers that are also in LD with the trait locus, in order to fine-map the search space. Alternatively, one may have loci known to exhibit allelic association (e.g. many HLA types and various diseases) from population studies, such that in a family-based analysis one might wish to condition on this to look for evidence of linkage. Formally speaking, neither of these conditional tests have the null hypothesis that $P(\mathbf{G_M}|\mathbf{G_D}) = P(\mathbf{G_M})$, as this only applies when there is complete independence of trait and marker locus genotypes.

If everything is modeled accurately (i.e. genotype-phenotype relationship at the putative trait locus), treating either linkage or LD parameters as nuisance parameters in the analysis should provide a valid testing framework for statistical inference. Many investigators, ourselves included, have demonstrated that this is true for conditional tests of both types, so long as the likelihood is maximized over all other parameters in the term $P(\mathbf{G_M}|\mathbf{G_D})$ separately in alternative and null hypotheses. For example, if the null hypothesis were linkage but no association, the recombination fraction and marker allele frequencies would be estimated under the null, and the recombination fraction and conditional marker allele frequencies would be estimated under the alternative hypothesis of linkage *and* LD. Or, equivalently, if association but no linkage were the null hypothesis (as in classic tests like the TDT (Spielman *et al.*, 1993, Terwilliger & Ott, 1992)), one would estimate conditional marker allele frequencies under both null and alternative, fixing the recombination fraction to 0.5 under the null, and freely estimating it under the alternative. When the “user-specified” parameters of the likelihood (like mode of inheritance) are accurately specified, this is a powerful and useful means to decompose the test into components due to linkage and association, such that a more meaningful interpretation of results can be made than “the null hypothesis of no linkage and no association was rejected”.

Expected log likelihoods and consistency of parameter estimates under wrong models

a. Demonstration of inflated false positive rate

For simplicity, we consider the expected value of the likelihood in a single sibpair under a given “true model”, $\mathbf{M}_{\text{True}}$, analyzed under some “analysis model”, $\mathbf{M}_{\text{Analysis}}$, to examine the behavior of the maximum likelihood estimations (MLE’s) when the model assumptions are inaccurately specified. We compute the expected log-likelihood as a function of θ , $p_{1|A}$, and $p_{1|B}$, the three parameters that define the term $P(\mathbf{G}_M|\mathbf{G}_D)$ in the formulation of the likelihood, with the parameters of $P(\mathbf{G}_D)$ and $P(\mathbf{Ph}|\mathbf{G}_D)$ being specified by $\mathbf{M}_{\text{Analysis}}$, and sum this expectation over all possible values of $\mathbf{G}_M$ and $\mathbf{G}_D$ for our single fixed data structure and vector $\mathbf{Ph}$ consisting of two unaffected parents having two affected children for some disease. Thus, $E[\ln L(\theta, p_{1|A}, p_{1|B})] = \sum_{\mathbf{G}_D} \sum_{\mathbf{G}_M} P(\mathbf{G}_D, \mathbf{G}_M; \mathbf{M}_{\text{True}}) \ln L(\theta, p_{1|A}, p_{1|B}; \mathbf{M}_{\text{Analysis}})$.

To examine the validity of the test for LD conditional on linkage, we consider $\mathbf{M}_{\text{True}}$ in which $\theta = 0$, and $p_{1|A} = p_{1|B} = 0.1$ (complete linkage but no LD with a marker having minor allele frequency of 0.1), for a variety of mode of inheritance assumptions, and then examine whether or not the MLEs of the conditional allele frequencies lie on the line $p_{1|A} = p_{1|B}$, consistent with the null hypothesis. One would imagine that the worst behavior of these statistics would occur when $\mathbf{M}_{\text{True}}$ and $\mathbf{M}_{\text{Analysis}}$ are both deterministic and very different one from the other, and therefore we consider two highly determined models, rare recessive with very low penetrance and no phenocopies ($\mathbf{M}_{\text{Rec}}$) and rare dominant with very low penetrance and no phenocopies ($\mathbf{M}_{\text{Dom}}$) as defined in Table 1. In $\mathbf{M}_{\text{Rec}}$, both (unaffected) parents would be most likely AB at the trait locus and both (affected) children would be BB at the trait locus, while under $\mathbf{M}_{\text{Dom}}$, one parent would likely be AA and the other AB, while both children would be most likely AB at the trait locus. Since we are interested in the estimates of the conditional allele frequencies, we compute the expected profile likelihoods over the recombination fraction, which is treated as a nuisance parameter in the conditional analysis, as $E[\ln L(p_{1|A}, p_{1|B})] = \max_{\theta} E[\ln L(\theta, p_{1|A}, p_{1|B})]$.

In Figure 1, contour plots of this profile likelihood are graphed as a function of $p_{1|A}$ and $p_{1|B}$ for all combinations of the models $\mathbf{M}_{\text{Rec}}$ and $\mathbf{M}_{\text{Dom}}$. Figure 1A is for the case $\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Rec}}$ –the analysis model and the true models are both rare recessive with no phenocopies. Not surprisingly, the profile likelihood maximizes at the true parameter values, $p_{1|A} = p_{1|B} = 0.1$. The white diagonal line in the figure corresponds to the null hypothesis that $p_{1|A} = p_{1|B}$, and in this case the likelihood maximizes on this line, and thus, asymptotically, the appropriate null hypothesis is recovered. The same holds for Figure 1B in which $\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Dom}}$ – the true model and the analysis model are both rare dominant with no phenocopies. Once again, the estimates are consistent, as they should be when the analysis is done under the formally correct parametric model. In Figure 1C, we consider the case where the true model is dominant, but we incorrectly analyze the data under the assumption of a recessive model ($\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Dom}}$; $\mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Rec}}$; stochastically equivalent to many “non-parametric” approaches (Göring & Terwilliger, 2000b)). Despite the very inaccurate analysis model, consistent parameter estimates are obtained and the null hypothesis can be validly tested. However in Figure 1D, when we consider the true model to be recessive, but we analyze the data assuming a dominant model ($\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Rec}}$; $\mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Dom}}$), the maximum likelihood estimates no longer fall on the line $p_{1|A} = p_{1|B}$, but rather at $p_{1|A} = 0.084$, $p_{1|B} = 0.147$. As can be seen from the shape of the contour plot, this overall maximum would be larger in expectation than the maximum log-likelihood constrained under the null hypothesis (i.e. constrained to the line $p_{1|A} = p_{1|B}$), leading to the rejection of the null hypothesis of linkage and no LD in favor of the

hypothesis of linkage *and* LD. Expected log-likelihoods are additive across families, so that in a sample of N sibpairs, $E[\ln L(p_{1|A}, p_{1|B})]$ would be N times larger, as would the test statistic (log-likelihood-ratio), and as N goes to infinity, so does the expected LRT, meaning that false positive evidence of LD conditional on linkage is guaranteed, even when LD is absent, and thus the test is meaningless and inappropriate in those conditions. Hence, if one assumes a strong dominant risk allele and the true model is a strong recessive model, one risks making false positive inference of LD in the presence of linkage when there is none. In this situation, the properties of the LRT are dependent on both the true mode of inheritance AND the assumed mode of inheritance, a situation that can never occur in simple linkage analysis, simple association analysis, or joint linkage and association analysis. It only arises when one performs conditional analyses that try to separate the effects of linkage and LD.

b. Effects of missing parental genotypes on bias and consistency

When parents are not available for genotyping, the problem can become much worse. In Figure 2A we show the profile log-likelihood surface when $\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Rec}}$, i.e. both analysis and true models were rare recessive, and there was complete linkage. In Figure 2B we show the same surface when only one of the two parents is available for genotyping – now the log-likelihood decreased much more sharply along with changes in $p_{1|B}$ than $p_{1|A}$ because we have information from both children about marker alleles co-transmitted with the B allele, while one marker allele is not transmitted (i.e. mostly likely in phase with non-transmitted allele A in the genotyped parent, given that we assume complete linkage in $\mathbf{M}_{\text{True}}$). The estimates are still consistent, but there is much higher variance in the estimate of $p_{1|A}$ because the sample size is roughly half as large. In Figure 2C we show the likelihood surface when neither parent is available for genotyping, which of course is completely flat over the range of $p_{1|A}$ as there are no controls in the analysis, and all observed marker alleles were cotransmitted with the risk alleles to the two affected children. Still, there is no systematic false positive evidence of association as there is no information at all about the conditional allele frequencies given the A allele. For the pathological situation $\mathbf{M}_{\text{True}} = \mathbf{M}_{\text{Rec}}$; $\mathbf{M}_{\text{Analysis}} = \mathbf{M}_{\text{Dom}}$, the graphs shown in Figure 2D, 2E, and 2F correspond to the situations with both parents genotyped, one parent genotyped and no parents genotyped. In this case, the log-likelihood surface is not flat when no parents are genotyped because the analysis model assumes that each affected person has one A allele and one B allele. Note that as the number of genotyped parents decreases, the bias in the estimated conditional allele frequencies increases, as does the statistical significance of that deviation!

c. Effects of population controls

Based on these results, it is clear that whenever one is performing any type of association analysis, one should always have population controls included in the analysis. We have recently shown that substantial power gains are possible from their inclusion, and there is something intrinsically unsettling about concluding that a given genotype increases the risk of disease when no healthy controls are ascertained (Göring & Terwilliger, 2000b, Hiekkalinna *et al.*, 2011a, Hiekkalinna *et al.*, 2011b, Terwilliger & Weiss, 2003). In linkage analysis we have not traditionally relied on control samples because our null hypothesis of no linkage was quantifiable by the rules of Mendelian inheritance. However, the null hypothesis of no association implies that the frequency of the marker allele should be identical in affected and unaffected individuals (or more accurately on chromosomes with or without a risk allele). But there is no mathematical theory describing what those allele frequencies should be, and this is why it is critical to ascertain an appropriate control sample (cf [Hiekkalinna *et al.*, 2011, unpublished data]). The next obvious question is what happens to this bias when controls are added to the analysis – can one make the conditional test of LD given linkage valid by adding population controls to give better estimates of $p_{1|A}$? To examine this, we added various numbers of random population controls to the analysis and

reevaluated the MLE's of the conditional allele frequencies, obtaining the results shown in Figure 3A. In this figure we graph the asymptotic bias in the conditional allele frequency estimate of $p_{1|B}$, ($B = E[\hat{p}_{1|B}] - p_{1|B}$), and in Figure 3B the same is presented for $p_{1|A}$. Similar results were obtained from sib-trios and larger families across all combinations of missing parental genotypes, and of M_{Dom} and M_{Rec} (data not shown). As expected, the bias in $P(1|A)$ goes away with the addition of unrelated population controls, but since these individuals are all inferred to be AA at the trait locus under our analysis model, the effect on $P(1|B)$ is far less impressive. The bias is ameliorated somewhat, but certainly it is not removed even with an infinite control sample. Obviously one must be careful in sampling to avoid introducing population stratification due to unrepresentative control samples, and all data (family-based or not) should be analyzed for potential stratification using classical approaches, as discussed elsewhere (Devlin & Roeder, 1999, Hiekkalinna *et al.*, 2011b, Price *et al.*, 2006, Pritchard *et al.*, 2000).

d. Impact of specific combinations of true and assumed models

Because this bias is a real potential problem in conditional analysis, we wanted to evaluate under what combinations of models this problem materializes, since if the true model really were so deterministic as the models we considered here, one might be able to estimate it reliably from classical segregation analysis. Furthermore, for most contemporary applications, we have more common diseases characterized by relatively small effect sizes. To this end, we considered the range of true models under which analysis under this most egregiously wrong rare dominant model, M_{Dom} , would lead to inconsistent parameter estimates in small nuclear families. In Figure 4A, we graph the bias in estimation of $p_{1|B}$ as a function of the true mode of inheritance. Here we assume that the risk allele B has frequency of 0.05 in the population, the disease has a prevalence of 0.01, and we let the relative risk for individuals homozygous for the B allele range from 1 to 50, under recessive, dominant, multiplicative, and additive models. As shown, there is a bias when the mode of inheritance is recessive and the relative risk is 15 or more, while under dominant, multiplicative, and additive models, there is no notable bias unless the relative risk is enormous, which is not what one expects for complex traits, and when the relative risk is that large, segregation analysis should make the mode-of-inheritance fairly clear as well, making this issue largely moot. If there are missing data, then the results are somewhat worse as can be seen in Figure 4B, where the data structure is one affected sib-trio with parental genotypes unknown, but even then, if the actual relative risk is 10 or less, there is little to worry about. Note that the y-axis scale is ten-fold larger in Figure 4B than 4A, meaning that the biases are much more severe with missing data.

This observation begs the question of why the estimates are biased. To examine this in more detail, we examined the probabilities of each possible outcome for G_D and G_M for all persons in the sibpair to determine where the deviation was coming from. We computed the detectance of marker genotypes $P(G_M|Ph)$ for all possible genotype vectors G_M in an affected sibpair (with both parents unaffected) conditional on complete linkage and no LD under both models, M_{Rec} , and M_{Dom} . In this case, we assumed that allele 1 at the marker locus has frequency 0.1 in the population, and is completely linked to the trait locus, as above (see Hiekkalinna et al, 2011, unpublished data, (Terwilliger & Weiss, 2003, Weiss & Terwilliger, 2000)). In Table 2, we show the detectances of each possible marker genotype vector, G_M , under both models. Unshaded rows have identical genotypes for both siblings, and are consistent with both M_{Rec} and M_{Dom} . Lightly shaded rows have one of the two marker alleles shared by the sibs, and differ at the second (consistent with M_{Dom} but not with M_{Rec}), and the darkly shaded row has no alleles shared between the sibs and is not therefore consistent with either model.

Because the 1 allele is rare, if it is seen at all in a dataset, most often it will have occurred only once, in a mating type $1,2 \times 2,2$. Under model $\mathbf{M}_{\text{Rec}}$, 50% of the time, this would lead to children who have both received one copy of the 1 allele (genotype 1,2), and 50% of the time, this would lead to both children being 2,2. However, under $\mathbf{M}_{\text{Dom}}$, if the 1,2 parent carried the risk allele, then 50% of the time, both children would be 1,2, and 50% of the time both children would be 2,2. While if the 2,2 parent carried the risk allele, then 25% of the time both children would be 1,2, 25% of the time both children would be 2,2, and 50% of the time there would be a 1,2 child and a 2,2 child. This last 50% of possible outcomes is therefore missing. The data are consistent with the 1,2 parent being more often the carrier of the disease allele than the 2,2 parent, creating a bias in favor of LD between the risk allele and the minor allele at the marker locus.

Why use wrong over-deterministic models at all?

Because family data are ascertained in a biased manner from populations, in that generally only multiplex pedigrees are informative for linkage, it has been demonstrated that linkage analysis using overly deterministic models has greater power to detect linkage than the “true” biological model in most situations (e.g. (Göring & Terwilliger, 2000b, Terwilliger, 2001, Terwilliger & Göring, 2000)), even though an upward bias in the estimated recombination fraction would result. If one models this bias explicitly through the use of complex-valued recombination fractions (Göring & Terwilliger, 2000a), power is typically maximized when using models that assume the maximal number of meioses are informative for linkage (Terwilliger, 2001) – that is to say models which assume rare disease alleles and do not allow for any phenocopies – even though that is clearly not a reasonable model of the underlying biology. The reason is that this model implies that affected individuals in pedigrees share some risk allele (the assumption that motivated the collection of multiplex families in the first place). When this assumption is untrue, the estimated recombination fraction will be inflated. However, as shown in Terwilliger & Göring (2000), it also maximizes the information used in the analysis, such that no “actually informative” meioses are falsely assumed to be uninformative for linkage. As outlined in Terwilliger (2001), using models that imply a stronger genetic component of the trait than is truly there does not negatively impact the power of a study (at least when using complex-valued recombination fractions), while models that imply a weaker genetic component than is true systematically lead to reduced power and accuracy in mapping studies because essentially they integrate out a lot of the actual signal in the data.

To demonstrate this principle in general, we simulated 1000 replicates of a dataset consisting of 800 affected sibpairs under each of a range of models in which a disease allele with 10% frequency increases the risk of some disease with 5% prevalence in a recessive manner. RR_{BB} , the relative risk of disease given two copies of allele B at the trait locus, is

varied from 1 to 6, where $RR_{BB} = \frac{P(\text{Disease}|BB)}{P(\text{Disease})|AA}$. We also simulated a marker locus with minor allele frequency $P(1) = 0.1$, at recombination fraction 0 with the risk locus. Expected lod scores were computed under the true model ($\text{ELOD}_{\text{true}}$) and under the overly deterministic recessive model (ELOD_R) - ($\mathbf{M}_{\text{Rec}}$ defined in Table 1). Replicates were simulated with the (Fast)SLINK program (Ott, 1989, Weeks *et al.*, 1990) and lod scores (maximized over recombination fraction) were computed with the MLINK program from FASTLINK 4.1P package (Cottingham *et al.*, 1993, Lathrop & Lalouel, 1984, Lathrop *et al.*, 1984, Lathrop *et al.*, 1986, Schäffer *et al.*, 1994). In Figure 5, we graph the ratio of $\text{ELOD}_{\text{Rec}}/\text{ELOD}_{\text{true}}$ to demonstrate how much signal can be gained from use of a highly inaccurate, overly deterministic analysis model. Note that the weaker the true effect, the greater the gain in power from using an overly determined model (see also (Terwilliger & Göring, 2000)). However, as shown in this paper, one needs to use caution in interpreting

the results of conditional tests of LD given linkage under circumstances in which certain types of model misspecification are employed.

Discussion

We have demonstrated a potential problem with conditional analysis of LD given linkage, in which spurious LD can be systematically inferred because of errors in the mode of inheritance assumed in the analysis, relative to the true mode of inheritance. These errors are seen when a dominant model is assumed for purposes of analysis, in pedigrees with unaffected parents, in which the true mode of inheritance is recessive, and in which there is known to be complete linkage between marker and trait loci. We demonstrate that while this problem can be alleviated somewhat through addition of population controls to the analysis, there remains an asymptotic bias. When M_{Rec} is used, no bias results in any situation, so we recommend this as a preferred strategy, though when one has clear intergenerational transmission of phenotypes, a dominant analysis is likely warranted and more appropriate. In the case of multigenerational pedigrees in which the lineage through which the disease alleles are transmitted can be inferred, the bias does not appear to exist, and in those situations the possibility that the true mode of inheritance would be recessive is likewise minimal, and of little concern. It should be noted that we have considered only fairly extreme cases in this paper – it is possible that similar biases may result under other conditions where more common disease alleles are assumed to segregate. We advise anyone applying conditional testing of LD given linkage under analysis models outside the range we have considered to perform a similar analysis to check for potential problems.

A joint test of linkage and LD does not suffer from this problem, as when there is no linkage between trait and marker loci, no spurious LD is inferred under any incorrectly specified model. Therefore, unless one has prior evidence of linkage, joint testing strategies may be preferred. And if there is significant evidence of linkage in a dataset, presumably the mode of inheritance would be less ambiguous and thus less likely to cause the sort of problem described in this paper.

The problem described in this manuscript is a general feature of family-based association tests that are based on some parametric mode-of-inheritance assumptions, and is not unique to any of the various implementations of such tests. It is rather an intrinsic property of the likelihoods, and investigators should be careful of any situation in which their test looks at association in a dominant or quasi-dominant manner in the absence of clear evidence of which parent would be transmitting the risk allele. The software used for the likelihood calculations in this paper is available from the authors.

Acknowledgments

Funding from the FiDiPro program of the Academy of Finland, grants MH84995, MH059490, and RR017515 from the National Institutes of Health, the Helsingin Sanomat Centennial Foundation, Biomedicum Helsinki Foundation, Emil Aaltonen Foundation, Otto A. Malm Foundation, Jenny and Antti Wihuri Foundation and Finnish Cultural Society are gratefully acknowledged. We gratefully acknowledge the assistance, guidance, and support provided by Markus Perola and the late Leena Peltonen-Palotie over many years. Finnish IT Center for Science (CSC) Linux-based supercomputers Murska and Vuori were used for doing computations and support from CSC is greatly acknowledged.

References

- Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet.* 1999; 65:1170–7. [PubMed: 10486336]
- Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics.* 1986; 42:393–9. [PubMed: 3741977]

- Cottingham RW Jr, Idury RM, Schäffer AA. Faster sequential genetic linkage computations. *Am J Hum Genet.* 1993; 53:252–63. [PubMed: 8317490]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered.* 2008; 66:87–98. [PubMed: 18382088]
- Görling HH, Terwilliger JD. Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet.* 2000a; 66:1095–106.
- Görling HH, Terwilliger JD. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet.* 2000b; 66:1310–27.
- Hiekkalinna T, Görling HHH, Lambert BW, Weiss KM, Norrgrann P, Schäffer AA, Terwilliger JD. On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples. Under review. 2011a
- Hiekkalinna T, Schäffer AA, Lambert BW, Norrgrann P, Görling HHH, Terwilliger JD. PSEUDOMARKER: A powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals. *Hum Hered.* 2011b; 71:256–266. [PubMed: 21811076]
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet.* 1996; 58:1347–63. [PubMed: 8651312]
- Kuokkanen S, Sundvall M, Terwilliger JD, Tienari PJ, Wikström J, Holmdahl R, Pettersson U, Peltonen L. A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2. *Nat Genet.* 1996; 13:477–80. [PubMed: 8696346]
- Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. *Genetic Epidemiol.* 2000; 19(Suppl 1):S36–42.
- Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet.* 2001; 69(Suppl):504. [PubMed: 11462172]
- Lathrop GM, Lalouel JM. Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet.* 1984; 36:460–5. [PubMed: 6585139]
- Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci U S A.* 1984; 81:3443–6. [PubMed: 6587361]
- Lathrop GM, Lalouel JM, White RL. Construction of human linkage maps: likelihood calculations for multilocus linkage analysis. *Genetic Epidemiol.* 1986; 3:39–52.
- Li M, Boehnke M, Abecasis GR. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am J Hum Genet.* 2005; 76:934–49. [PubMed: 15877278]
- Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet.* 2006; 78:778–92. [PubMed: 16642434]
- Ott J. *Analysis of Human Genetic Linkage.* Johns Hopkins University Press; Baltimore: 1985.
- Ott J. Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci U S A.* 1989; 86:4175–8. [PubMed: 2726769]
- Ott J. Strategies for characterizing highly polymorphic markers in human gene mapping. *American journal of human genetics.* 1992; 51:283–90. [PubMed: 1642229]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics.* 2006; 38:904–9. [PubMed: 16862161]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *American journal of human genetics.* 2000; 67:170–81. [PubMed: 10827107]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]

- Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered.* 2000; 50:211–23. [PubMed: 10782012]
- Satsangi J, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K, Terwilliger JD, Lathrop GM, Bell JI, Jewell DP. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet.* 1996; 14:199–202. [PubMed: 8841195]
- Schäffer AA, Gupta SK, Shriram K, Cottingham RW Jr. Avoiding recomputation in linkage analysis. *Hum Hered.* 1994; 44:225–37. [PubMed: 8056435]
- Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993; 52:506–16. [PubMed: 8447318]
- Terwilliger JD. On the resolution and feasibility of genome scanning approaches. *Advances in Genetics.* 2001; 42:351–91. [PubMed: 11037331]
- Terwilliger JD, Göring HH. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol.* 2000; 72:63–132. [PubMed: 10721614]
- Terwilliger JD, Ott J. A haplotype-based ‘haplotype relative risk’ approach to detecting allelic associations. *Hum Hered.* 1992; 42:337–46. [PubMed: 1493912]
- Terwilliger JD, Weiss KM. Confounding, ascertainment bias, and the blind quest for a genetic ‘fountain of youth’. *Ann Med.* 2003; 35:532–44. [PubMed: 14649335]
- Weeks DE, Ott J, Lathrop GM. SLINK: a general simulation program for linkage analysis. *Am J Hum Genet.* 1990:A204.
- Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nat Genet.* 2000; 26:151–7. [PubMed: 11017069]
- Williamson JA, Amos CI. On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genetic Epidemiology.* 1990; 7:309–18.

Figure 1.

The expected profile log-likelihood surface for a single sibpair. (A) True and analysis models M_{Rec} , (B) true model and analysis models M_{Dom} , (C) true model M_{Dom} and analysis model M_{Rec} , and (D) true model M_{Rec} and analysis model M_{Dom} .

Figure 2. The expected profile log-likelihood surface for a single sibpair. In A, B, and C true model and analysis model are M_{Rec} . In D, E, and F true model is M_{Rec} and analysis model is M_{Dom} . A and D both parent genotypes are known, B and E one parent genotype is unknown, and C and F, both parental genotypes are unknown.

Figure 3.

The bias of parameters estimates and effect on estimates when adding controls. (A) The bias of $P(1|B)$, and (B) $P(1|A)$. A complete linkage between disease locus and marker was assumed ($\theta=0$) and no LD [$P(1|D)=P(1|+)=0.1$, i.e. $\delta=0$], and true model is $\mathbf{M}_{\text{Rec}}$ and analysis model is $\mathbf{M}_{\text{Dom}}$.

Figure 4.

The bias of estimated parameter $P(1|B)$ as a function of the true model parameters RR_{BB} and RR_{AB} when the analysis model is M_{Dom} : $P(\text{Disease}|AA) = 0$; $P(\text{Disease}|AB \text{ or } BB) = 0.00001$; $P(B) = 0.00001$. The data structure is one affected sib-pair with parental genotypes known. The true disease allele frequency was $P(B) = 0.05$, $P(\text{Disease}) = 0.01$, and $\theta=0$. Note that the line for $RR_{AB} = 1$ is equivalent to a purely recessive mode of inheritance with relative risk RR_{BB} , and the other lines are for intermediate models with intermediate amounts of dominance. (B) The bias of estimated parameter $P(1|B)$ over true model parameters RR_{BB} and RR_{AB} when analysis was model M_{Dom} . The data structure is one sib-trio and parental genotypes are unknown. True model disease allele frequency was 5%, disease prevalence 1%, and $\theta=0$.

Figure 5.

The ratio of M_{Rec} and true model-based linkage expected lod scores. The data was 800 affected sib-pairs with the disease allele frequency of 10%, disease prevalence of 5%, and $\theta=0$.

Table 1

Model parameters for deterministic recessive ($\mathbf{M}_{\text{Rec}}$) and dominant ($\mathbf{M}_{\text{Dom}}$) models (Göring & Terwilliger, 2000b, Kuokkanen *et al.*, 1996, Satsangi *et al.*, 1996):

	$\mathbf{M}_{\text{Rec}}$	$\mathbf{M}_{\text{Dom}}$
P(B)	0.00001	0.00001
P(Disease AA)	0	0
P(Disease AB)	0	0.00001
P(Disease BB)	0.00001	0.00001

Table 2

Detectances for marker genotype vectors $\mathbf{G}_M$, assuming deterministic recessive ($\mathbf{M}_{\text{Rec}}$) and dominant ($\mathbf{M}_{\text{Dom}}$) models. Shaded rows indicate situations where the genotypes of the two children are different. The one situation where both alleles are different (11, 22) for the two children is darkly shaded.

$\mathbf{G}_M$ (Parents)	$P(\mathbf{G}_M\text{-Parents})$	$\mathbf{G}_M$ (Sibs)	$P(\mathbf{G}_M \mathbf{M}_{\text{Rec}})$	$P(\mathbf{G}_M \mathbf{M}_{\text{Dom}})$
11×11	0.0001	11-11	0.0001	0.0001
11×12	0.0036	11-11	0.0018	0.00135
		11-12	0.00000036	0.0009
		12-12	0.0018	0.00135
11×22	0.0162	12-12	0.0162	0.0162
12×12	0.0324	11-11	0.0081	0.00405
		11-12	0.00000324	0.0081
		11-22	0.000000000000162	0.000000162
		12-12	0.0162	0.0081
		12-22	0.000000324	0.00809998
		22-22	0.0081	0.00405
12×22	0.2916	12-12	0.146	0.10935
		12-22	0.00000292	0.0729027
		22-22	0.146	0.10935
22×22	0.6561	22-22	0.6561	0.6561