Identification of and adjustment for genotyping errors in data on sibpairs when parental genotypes are unavailable

M. Bahlo[1], K.W. Broman[2] and T.P. Speed[1]

1. Cooperative Research Centre for Discovery of Genes for Common Human Diseases, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia

2. Marshfield Medical Research Foundation, Marshfield, Wisconsin

Abstract

In affected sibpair studies of late-onset diseases, such as prostate cancer, DNA samples for subjects' parents are often unavailable. In other studies, parents may not be genotyped as a cost saving measure. In such situations, genotyping errors cannot be detected by the usual methods, since all possible genotypes conform to Mendel's rules. We have studied two approaches to dealing with genotyping errors in such data, making use of multipoint marker information. First, we consider a likelihood ratio statistic to identify genotypes likely to be in error. Second, we calculate IBD probabilities given multipoint marker data using an incomplete penetrance function which allows for the presence of errors. We present the results of a simulation study comparing the power to detect a disease susceptibility gene when (a) genotyping errors are ignored, (b) genotypes likely to be in error are detected and removed, and (c) using an analytic method which allows for the presence of genotyping errors. We also investigate the bias introduced through methods (b) and (c). Genotyping errors are shown to erode the power to detect linkage. The effect may be ameliorated by use of methods to detect or allow for the presence of errors, though marker density must be quite high.

Introduction

- We studied the power to detect a disease locus in the presence of genotyping errors and the bias of two approaches
  - **Method I**: Detect errors by using log-likelihood ratio test. Remove errors by zeroing genotypes found to be in error. Carry out linkage analysis.
  - **Method II**: Allow for errors directly in the linkage analysis by making the IBD probabilities fuzzy.
- LOD scores for the various models were calculated as

$$LOD = NPL^2/(2 \ln 10)$$

where NPL denotes the statistic of Kruglyak et al. (1996)

Lander-Green algorithm

x_i = observed genotypes at marker i

v_i = IBD status 0,1 or 2 of marker i

$$\alpha_i(v) = \Pr(x_1,\ldots,x_i, v_i = v)$$

$$\alpha_1(v) = \Pr(x_1 \mid v_1 = v)\Pr(v_1 = v)$$

$$\alpha_i(v) = \Pr(x_i \mid v_i = v)\sum_{v^*}\alpha_{i-1}(v^*)\Pr(v_i = v \mid v_{i-1} = v^*)$$

$$\beta_i(v) = \Pr(x_{i+1},\ldots,x_n \mid v_i = v)$$

$$\beta_n(v) = 1$$

$$\beta_i(v) = \sum_{v^*}\beta_{i+1}(v^*)\Pr(v_{i+1} = v^* \mid v_i = v)\Pr(x_{i+1} \mid v_{i+1} = v^*)$$

$$\Pr(x) = \sum_{v}\alpha_i(v)\beta_i(v)$$

$$\Pr(v_i = v \mid x) = \alpha_i(v)\beta_i(v)/\Pr(x)$$

Method I-Identifying errors

- The Log Likelihood Ratio Statistic for a sibpair is

$$\mathrm{LOD}_{\mathrm{error}}(i)=\log_{10}[\Pr(\boldsymbol{x} \mid x_i \neq g_i) / \Pr(\boldsymbol{x} \mid x_i = g_i)]$$

Errors are introduced as penetrances:

$$\Pr(x_i \mid v_i) = (1 - \varepsilon) \Pr(x_i \mid v_i, x_i = g_i) + \varepsilon \Pr(x_i \mid v_i, x_i \neq g_i)$$

- Errors are assumed to occur independently between sibpairs, markers and genotypes

- ε (the prior error rate) and genetic map are known

Method II-Adjusting for errors

We use an error tolerant penetrance which takes the form

$$\Pr^{\varepsilon}(x_i|\, v_i = v) = (1 - \varepsilon)\, \Pr(x_i \mid v_i = v) + \varepsilon \Pr(x_i \mid v_i = 0)$$

where ε is the prior error rate and $\Pr(x_i \mid v_i = v)$ is the penetrance ignoring errors.

The $\Pr^{\varepsilon}(x_i|\, v_i = v)$ are inserted into the Lander-Green algorithm instead of $\Pr(x_i \mid v_i = v)$. The resulting $\Pr^{\varepsilon}(v_i = v|x_i)$ are then used to calculate statistics in the normal way.

Error Models

- A simple error model was used for the simulations and error detection. Genotypes are taken to be in error, with one or both alleles changed by one tandem repeat.
- What is the best way to find errors for real data?
  - On what level should errors be modeled?
    - Alleles or genotypes
  - What kind of error model?
    - Allow all possible alternative genotypes/alleles
    - Model microsatellite marker error behavior i.e. higher probability that the true alleles/genotypes only differs by one tandem repeat

Simulations

- 100, 250 or 500 affected sibpairs
- One ~50 cM segment with 6–51 equally spaced markers
- Markers in Hardy-Weinberg and linkage equilibrium
- Markers have 8 alleles with the following frequencies

  0.4, 0.2, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05 (het ≈ 0.77)

- A single gene exactly between the two central-most markers
- Gene results in IBD probabilities (for affected sibs)
  (0.1. 0.6, 0.3)
- Genotyping error rate of 0 or 1%
- 1000 simulation replicates for each choice of parameter values
- Error model for simulations different than that for analysis

Identifying errors

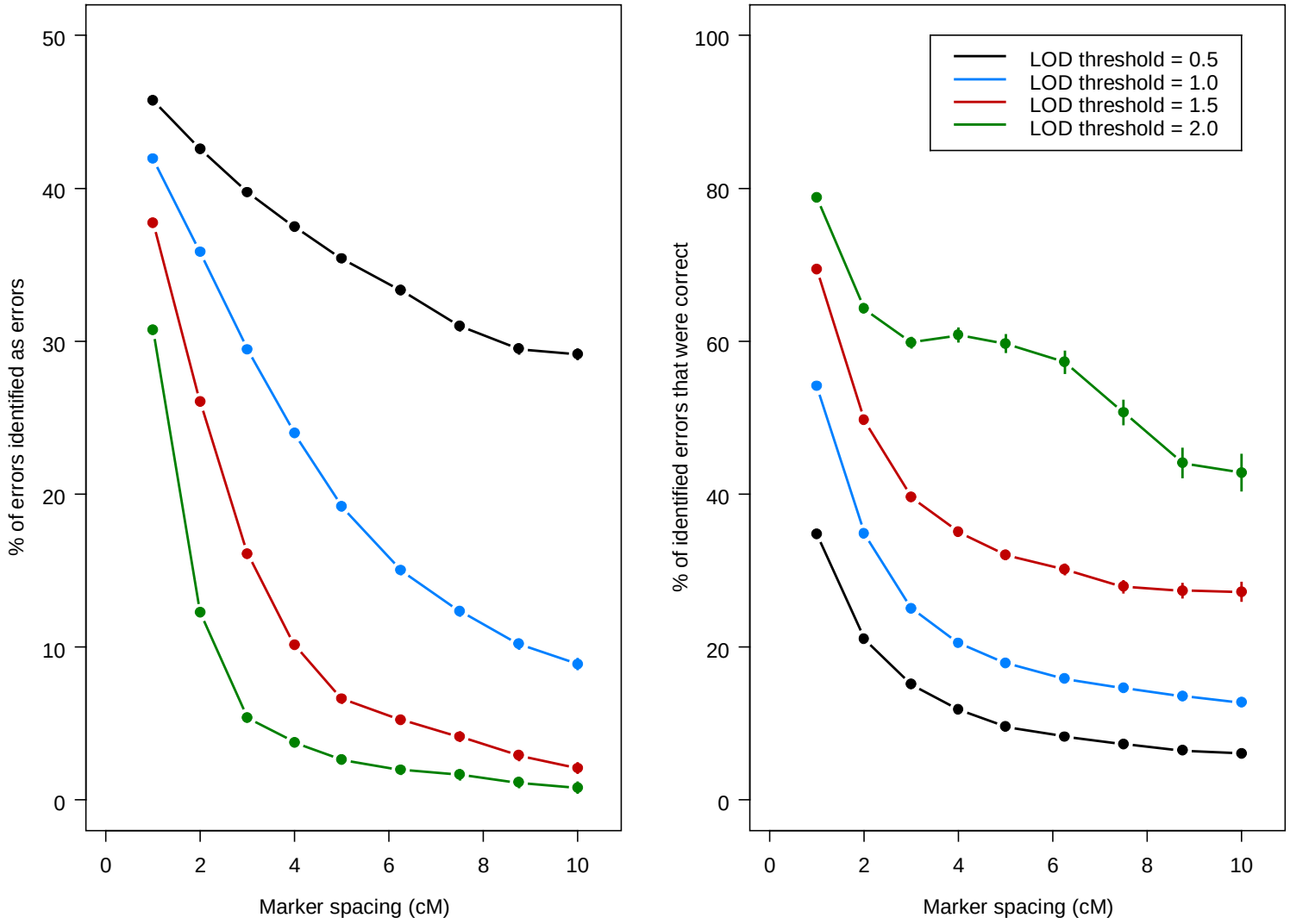

Figure 1: Ability to identify genotyping errors in data on affected sibpairs for 50 cM region with equally spaced markers in H-W and linkage equilibrium, having het ≈ 0.77, a single gene in the center of the region, and a genotyping error rate of 1%.

Effects of errors

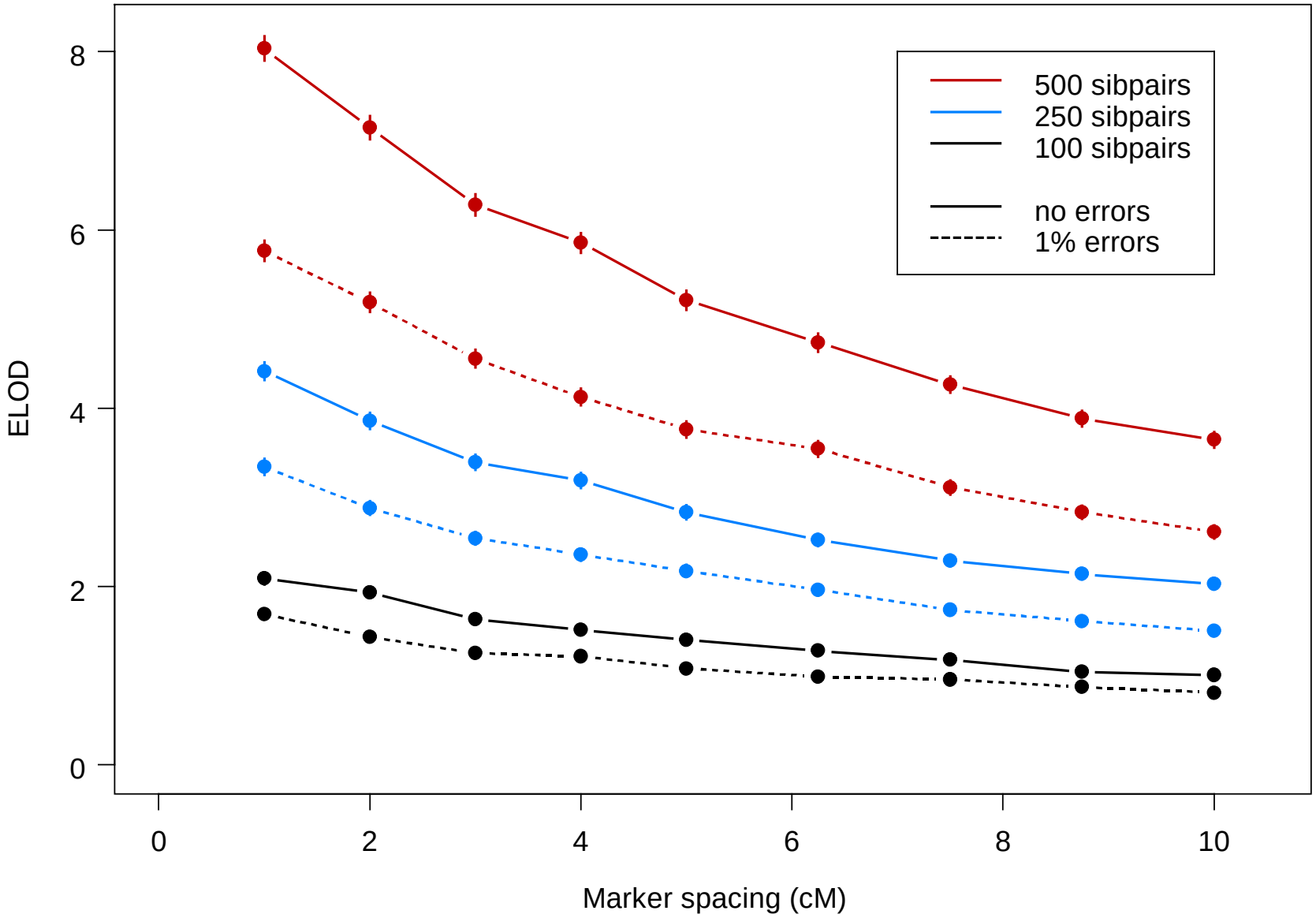

Figure 2: Average maximum LOD scores for different numbers of affected sibpairs, using data on equally spaced markers in a 50 cM region containing a single disease susceptibility locus, when the genotyping error rate is 0 or 1 %, and errors are ignored.

LOD bias

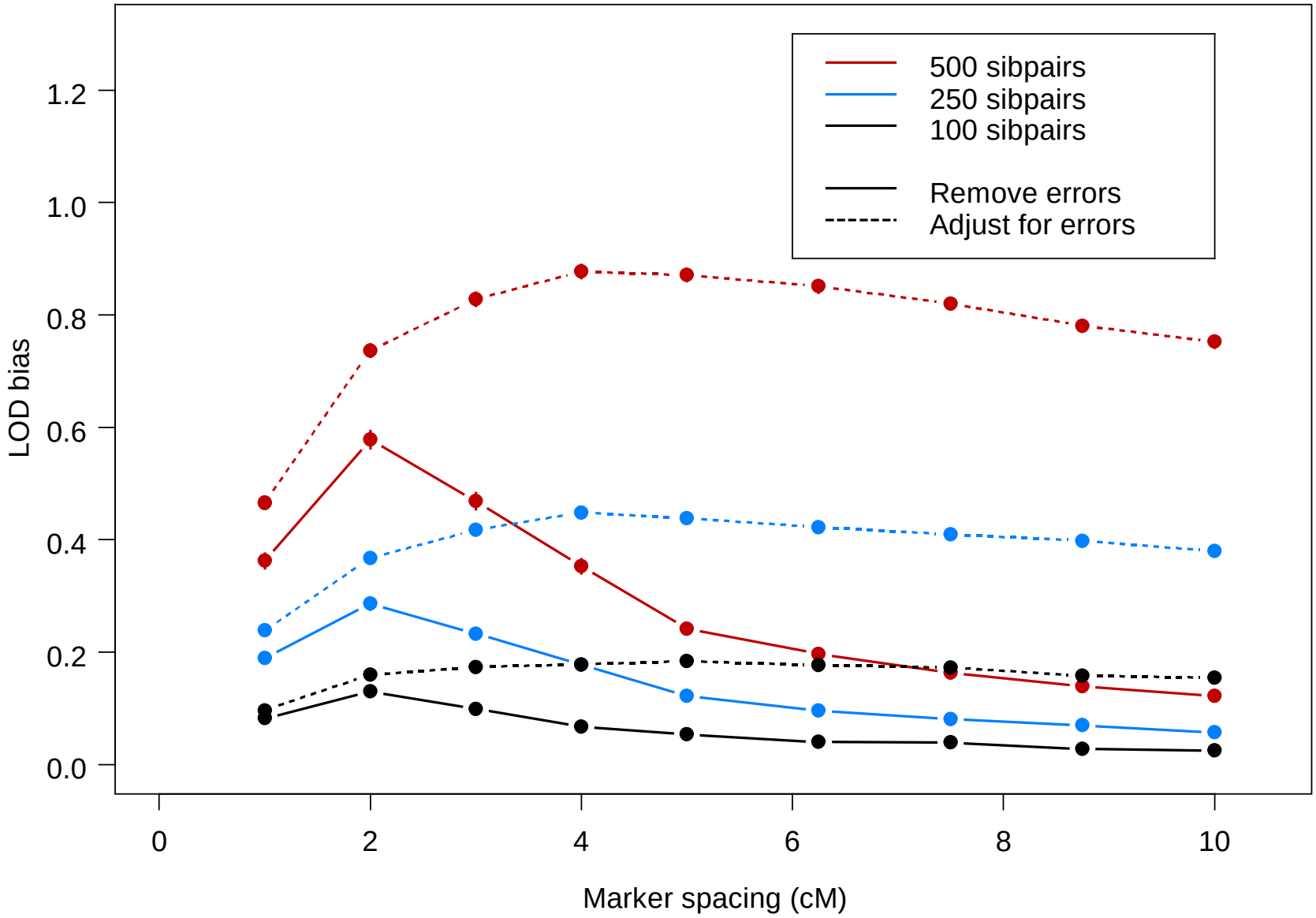

Figure 3: Difference between the average maximum LOD calculated after removing (or adjusting for) errors and versus that obtained ignoring potential errors, when the error rate is 0%. (In other words, the average inflation in the LOD score these methods give in the absence of errors.)

Increase in ELOD

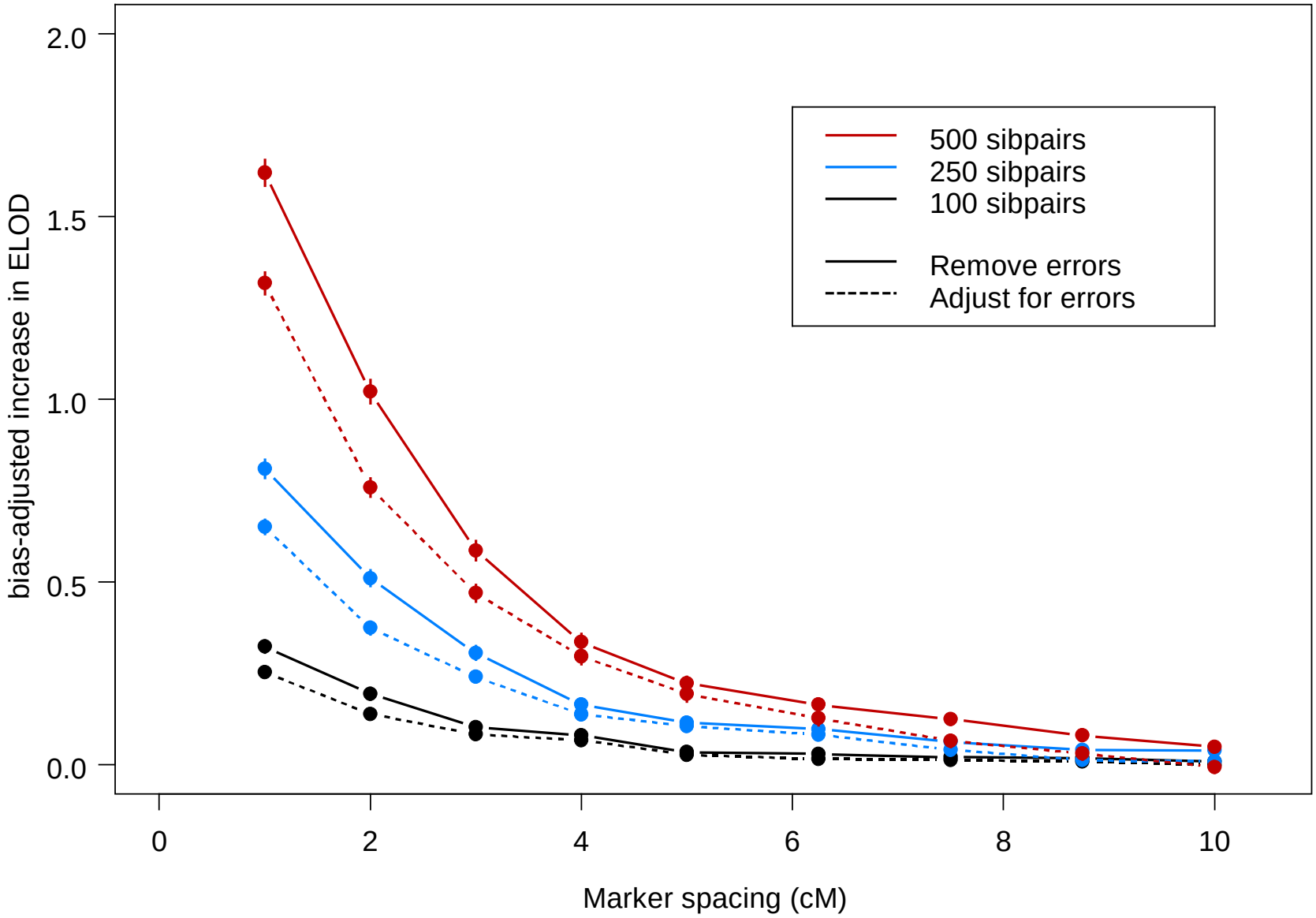

Figure 4: Average increase in maximum LOD score obtained by removing or adjusting for errors, in the presence of genotyping error rate of 1%. (Bias adjustment was performed by subtracting off the estimated LOD bias, given the method, sample size and marker density.)

Power with 1% errors

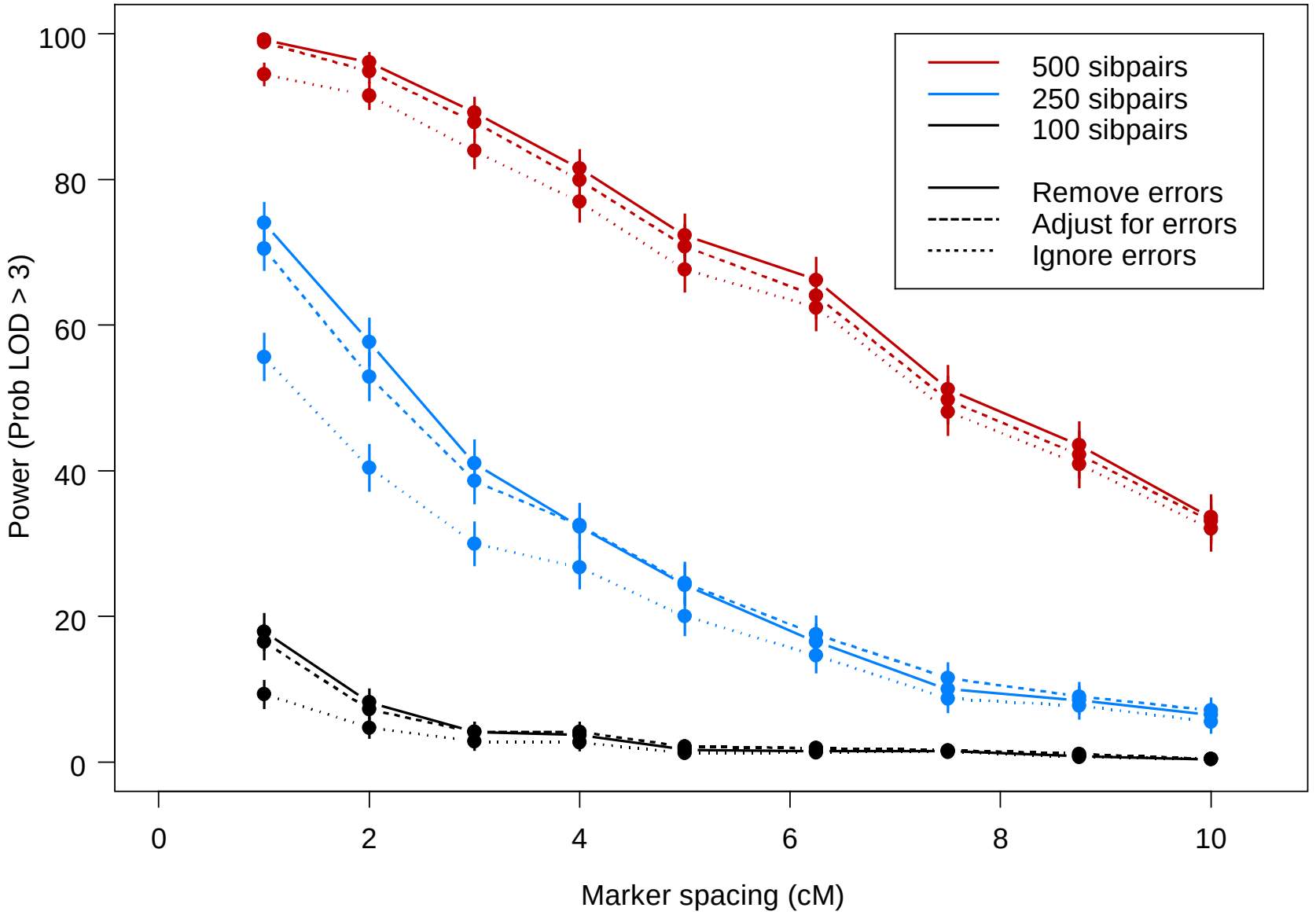

Figure 5: Power (using LOD threshold of 3) when ignoring, removing or adjusting for errors, in the presence of genotyping error rate of 1%. (LOD scores were adjusted for the estimated LOD bias.)

Discussion

- Genotyping errors reduce power to detect genes
- Methods to identify errors introduce bias: errors are easier to detect when IBD=2 than when IBD=0
- The bias leads to increased LOD scores, even when there are no genotyping errors
- Method I (removing genotypes found to be in error) leads to somewhat less bias than Method II (adjusting for errors)
- Identifying and removing errors leads to considerable increases in power if proper adjustment is made for the bias
- Power is approximately equal for both methods
- Need to find a general way to adjust for the bias, which is a function of the sample size, marker density, marker heterozygosity and size of gene effect