Error Detection for Genetic Data Using Likelihood Methods

Margaret Gelder Ehm [1] Marek Kimmel [2]

Robert W. Cottingham, Jr. [3]

American Journal of Human Genetics, Vol. 58, No. 1 (1996)

[1]Department of Statistics, Rice University, P.O. Box 1892, Houston, TX 77251-1892, (713) 527-8101 ext. 2695, gelder@rice.edu.

[2]Department of Statistics, Rice University, P.O. Box 1892, Houston, TX 77251-1892, (713) 527-5255, kimmel@rice.edu.

[3]Division of Biomedical Information Sciences, Johns Hopkins University, Baltimore, MD 21205, (410) 955-9705, bc@gdb.org.

Abstract

As genetic maps become denser, the effect of laboratory typing errors becomes more serious. We review a general method for detecting errors in pedigree genotyping data which is a variant of the likelihood ratio test statistic. It pinpoints individuals and loci with relatively unlikely genotypes. Power and significance studies using Monte Carlo methods are shown using simulated data with pedigree structures similar to the CEPH pedigrees and a larger experimental pedigree used in the study of idiopathic dilated cardiomyopathy (DCM). The studies show the index detects errors for small values of θ with high power and an acceptable false positive rate. The method was also used to check for errors in DCM laboratory pedigree data and to estimate the error rate in CEPH chromosome 6 data. The errors flagged by our method in the DCM pedigree were confirmed by the laboratory. The results are consistent with estimated false positive and false negative rates obtained using simulation.

1 Introduction

There are two types of genetic linkage data errors to consider: pedigree error and typing error. Pedigree error generally involves misidentification of individuals and relationships [15]. Examples include non-paternity, unidentified adoption and sample mix-ups. Typing error includes all other types of errors such as misinterpretation of genotypes and data entry error [2].

Chromosome linkage maps are becoming denser as more marker data are collected. These high resolution linkage maps are important for locating disease genes [2]. Building high resolution maps requires large amounts of genotypic data. The typing procedures are complicated and the volume of markers is large enough for errors to occur. As genetic maps become dense, errors can obscure the data by reducing the support for the correct genetic order making gene locations harder to define.

Errors manifest themselves in the data in various forms. Pedigree error will often be detectable since the incorrect relationship will likely show genotypes not conforming to Mendelian inheritance. Typing errors and pedigree errors consistent with Mendelian inheritance will either negate a true recombination or introduce a spurious recombination. In dense maps where markers are close together and true recombination events are rare, errors are more likely to appear as spurious recombinations. Ott [16] has developed a test for identifying the persons and loci causing inconsistencies. While pedigree errors can prove troublesome in locating disease genes, the links of the pedigree need only be resolved once. Typing errors, on the other hand, are an ongoing problem in a pedigree being typed for many markers. The present paper focuses on detecting typing and pedigree errors consistent with Mendelian inheritance.

The effect of typing errors has been discussed by several authors. Ott [15] illustrates the effect of errors by letting s be the misclassification frequency, the frequency with which recombinants are misclassified as nonrecombinants and nonrecombinants are misclassified as recombinants and θ_t be the true recombination fraction. Then the apparent recombination rate is given by:

$$\hat{\theta} = \theta_t(1-s) + (1-\theta_t)s = \theta_t + (1-2\theta_t)s.$$

There is no bias when $\theta_t = \frac{1}{2}$. However as $\theta_t \rightarrow 0$, $\hat{\theta}$ approaches the error frequency and most observed recombination events will be erroneous. Since errors in the data are usually observed as erroneous crossovers, they increase the map distance and make ordering loci more difficult. Lincoln and Lander

[12] give a first order approximation of the magnitude of the problem and illustrate its effect on maps.

Several authors have estimated error frequencies for typing data. On the basis of genotyping classical markers, Lathrop *et al* [10] estimated a typing error frequency of 1% in data obtained from a South Pacific island population, Tokelau. They also noted that most typing errors did not lead to genetic inconsistencies. Dracopoli *et al* [4] estimate the error frequency to be 0.6% for CEPH chromosome 1 typings and Buetow [2] estimates the error frequency to be 1.4% for CEPH chromosome 4 data. The approach used to obtain both estimates compared data typed by several laboratories.

There are two approaches to dealing with errors in pedigree data: correction of the recombination fraction and error detection. Adjusting the recombination fraction for errors may involve using a corrected estimate of θ in the map [13] or making linkage detection procedures more robust by modeling error in the data using the penetrance function (see Lincoln and Lander [12]). Currently many laboratories use only basic error detection procedures which include confirming Mendelian transmission of alleles and identifying multiple recombination events over short intervals [1]. The latter can be made simpler using the CHROMPIC option of the CRIMAP program

[7]. Several authors [1, 16] have considered methods that pinpoint individuals and loci likely to contain errors. These more sophisticated approaches are reviewed in the discussion.

In this paper, we review the likelihood ratio method proposed by Ehm *et al* [5], systematically investigate the statistical significance and power of the method, and apply it to an experimental pedigree and CEPH chromosome 6 data. While the method was developed to detect errors compatible with Mendelian inheritance in human pedigree data, it is general enough to highlight any statistically unlikely typing including pedigree errors.

2 Materials and Methods

When pedigree data are obtained by typing individuals, the observed genotype is equal to the true genotype unless a typing error has occurred. Following [5, 6, 12, 16], error in pedigree data can be represented as incomplete penetrance of genotypes. The observed genotypes are considered "phenotypes" and may not correspond to the true genotypes due to errors. Therefore, modeling error in pedigree data can be accomplished by using a specially designed penetrance function which we will refer to as the error-penetrance function. The method, outlined in a preliminary way in Ehm *et al* [5], is designed to identify individuals and loci likely to contain errors and is equivalent to a hypothesis test for error for each individual at each locus in the pedigree. Each hypothesis test entails (1) specifying a error-penetrance function, (2) calculating a likelihood ratio statistic to test the hypothesis that the error frequency is zero, and (3) identifying test statistics with relatively large values as indicative of an unlikely genotype.

2.1 Test Statistic

Let $X_j = (x_{1j}, \ldots, x_{mj})$ be a vector of phenotypes at locus j for individuals in a pedigree of size m. Here the term phenotype means the genotype that appears in the computer file used for analysis. Let $G_j = (g_{1j}, \ldots, g_{mj})$ be the corresponding vector of actual genotypes (unobservable in general). Denote the recombination frequency between two loci by θ_{jl}. It is assumed that each genotype at each locus has a known population frequency, recombinant haplotypes are passed with probability θ_{jl}, and given a specific genotype, g_{ij}, phenotype x_{ij} is exhibited with probability $\mathcal{P}(x_{ij} \mid g_{ij})$. To quantify the fit provided by the model and calculate maximum likelihood estimates of θ_{jl}, the likelihood of the phenotypes (or pedigree likelihood), $\mathcal{P}(\mathbf{X}; \theta_{jl})$ where $\mathbf{X} = (X_j, X_l)$ can be calculated. Several loci ($k \geq 2$) can be analyzed jointly by letting $\mathbf{X} = (X_1, \ldots, X_k)$ and $\theta = (\theta_{12}, \ldots, \theta_{k-1,k})$ be a vector of recombination fractions where k is the number of loci analyzed jointly. This is called k-point linkage analysis.

When genotypes are obtained by typing procedures and are assumed to be unequivocally known, the penetrance function for individual i at locus j is given by:

$$\mathcal{P}(x_{ij}|g_{ij}) = \begin{cases} 1 & \text{if } x_{ij} = g_{ij} \\ 0 & \text{if } x_{ij} \neq g_{ij}. \end{cases}$$

Typing error can be modeled by redefining the penetrance function [6, 12, 16]. Let ε_{ij} be the assumed error probability for individual i at locus j, a the number of alleles and $h = a(a+1)/2$, the number of possible combinations of two alleles. The error-penetrance function assuming a uniform distribution of error is given by:

$$\mathcal{P}(x_{ij}|g_{ij}) = \begin{cases} 1 - \varepsilon_{ij} + \frac{\varepsilon_{ij}}{h} & \text{if } x_{ij} = g_{ij} \\ \frac{\varepsilon_{ij}}{h} & \text{if } x_{ij} \neq g_{ij} \end{cases}$$

which indicates that with high probability $(1 - \varepsilon_{ij})$ the typing observed is the true genotype and with low probability (ε_{ij}) the typing observed is some other genotype. This particular error-penetrance function is defined such that all other genotypes are equally likely, but other definitions are possible.

This error-penetrance function effectively adds another parameter, ε_{ij}, to the model. In order to pinpoint individuals and loci containing errors, we conduct a series of hypothesis tests for each individual, $i = 1, \ldots, m$, at each locus, $j = 1, \ldots, k$:

$$H_0 : \varepsilon_{ij} = 0$$

versus

$$H_1 : \varepsilon_{ij} > 0.$$

To distinguish between the two hypotheses we use a test statistic which is an approximation to the likelihood ratio. Denote the likelihood function including the error-penetrance function for individual i at locus j by $L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij})$. The exact likelihood ratio statistic [18] would have the form:

$$\lambda_{ij}(\mathbf{X}) = -2\ln\left\{\frac{\sup_\theta L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij} = 0)}{\sup_{\theta,\varepsilon_{ij}} L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij})}\right\} = -2\ln\left\{\frac{L_{ij}(\mathbf{X}; \hat{\theta}_0, \varepsilon_{ij} = 0)}{L_{ij}(\mathbf{X}; \hat{\theta}_{\varepsilon^*}, \varepsilon_{ij} = \varepsilon^*)}\right\}.$$

$\hat{\theta}_\varepsilon$ denotes the value of θ maximizing $L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij})$ when $\varepsilon_{ij} = \varepsilon$. The global maximum of $L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij})$ is attained at $(\hat{\theta}_{\varepsilon^*}, \varepsilon^*)$.

Unfortunately, calculating the maximum likelihood estimate in the denominator is very time-consuming because allowing the penetrance to be incomplete for many individuals increases the number of nonzero array entries that must be multiplied and summed over. Therefore if a different denominator is used, $L_{ij}(\mathbf{X}; \hat{\theta}_0, \varepsilon_{ij} = \varepsilon_1)$ where

$$\hat{\theta}_0 = \sup_\theta L_{ij}(\mathbf{X}; \theta, \varepsilon_{ij} = 0)$$

then we obtain another test statistic:

$$\tilde{\lambda}_{ij}(\mathbf{X}) = -2\ln\left\{\frac{L_{ij}(\mathbf{X}; \hat{\theta}_0, \varepsilon_{ij} = 0)}{L_{ij}(\mathbf{X}; \hat{\theta}_0, \varepsilon_{ij} = \varepsilon_1)}\right\}$$

where ε_1 is an assumed value. As noted by Lincoln and Lander [12] and confirmed by our own simulations, the error analysis was quite insensitive to the value of ε_1 or the distribution of ε_{ij} in the error-penetrance function. We will use $\varepsilon_1 = 0.01$. Using the $\tilde{\lambda}_{ij}(\mathbf{X})$ instead of $\lambda_{ij}(\mathbf{X})$ is equivalent to using the alternative hypothesis $H_1 : \varepsilon_{ij} = \varepsilon_1, \theta = \hat{\theta}_0$ instead of $H_1 : \varepsilon_{ij} = \varepsilon^*, \theta = \hat{\theta}_{\varepsilon^*}$. From the viewpoint of statistical theory this replaces the best test, which is uncomputable, by a less powerful, but computable approximation. See Ehm *et al* [5] for more details concerning this approximation.

2.2 Test Statistic Distribution

Since the test statistic we employ [5], $\tilde{\lambda}_{ij}(\mathbf{X})$, is an approximation to the likelihood ratio test statistic and the sample sizes (number of individuals in the pedigree) are small, its distribution is difficult to derive. Therefore, we use Monte Carlo simulation to calculate p-values for the test.

In order to determine the p-values for each individual at a locus, we calculate the empirical distribution function of $\tilde{\lambda}_{ij}(\mathbf{X})$ for individuals $i = 1, \ldots, m$ and and loci $j = 1, \ldots, k$. Using the FASTSLINK or SIMULATE programs [3, 14, 20], we generate a large number of random pedigrees with a specific structure, number of loci, recombination values and specific allele frequencies under the hypothesis of no error. For each pedigree we calculate the test statistic, $\tilde{\lambda}_{ij}(\mathbf{X})$, for each individual at each locus. Then the test statistics based on all the generated pedigrees for the i^{th} individual and j^{th} locus are used to form an empirical cumulative distribution function under H_0. The p-values can be calculated by comparing observed test statistics for each individual at each locus to the empirical cumulative distribution for that individual and locus under the null hypothesis. In this way, testing $H_0 : \varepsilon_{ij} = 0$ by calculating $\tilde{\lambda}_{ij}(\mathbf{X})$ for each individual and locus allows us to pinpoint the individuals and loci that may contain errors.

2.3 Calculation of Test Statistics

The calculation of the approximate likelihood ratio statistic required for the series of tests outlined above involves several likelihood evaluations. The evaluations were performed using a modification of the ILINK program from the FASTLINK 2.2 package [3, 8, 9, 11, 17]. For each individual and locus, calculation of the test statistic requires finding the likelihood at the maximum likelihood estimate of θ with $\varepsilon_{ij} = 0$, $\hat{\theta}_0$, and then evaluating the likelihood at

$(\hat{\theta}_0, \varepsilon_1)$ assuming the error-penetrance function for that individual and locus. In effect, we are calculating the evidence for error in individual i at locus j assuming all other genotypes are correct. The evaluation time of the test statistic is reasonable since finding $\hat{\theta}_0$ (which involves several evaluations) is relatively fast. The computation time for calculating the test statistic for an entire pedigree with m individuals and k loci includes mk likelihood evaluations beyond the calculation of $\hat{\theta}_0$. An example illustrating the method is given in Ehm *et al* [5].

3 Results

3.1 Significance and Power Studies

An ideal test for genotype errors would be computationally feasible, have a low false positive rate and relatively high power (true positive rate). To evaluate our test we estimate the false positive rate and power for a three generation pedigree with structure similar to the CEPH pedigrees (three generation pedigrees with four grandparents, two parents, and many children) with two, three and four loci for various values of θ.

In order to estimate the false positive rate we generated, at random, 1500 such error-free pedigrees and used 1000 of them to construct an empirical cumulative distribution function for each individual at each locus under the null hypothesis of no error. The significance levels (false positive rates) were then studied for each individual at each locus using the remaining 500 error-free pedigrees. After setting the significance level (nominal α), the number of significant test statistics for each locus and individual were counted. The results for all individuals at all loci were pooled in order to estimate the number of individuals in a family that would be needlessly retyped. Then the true significance level (true α) was estimated by dividing the total number of significant test statistics for each individual at each locus by the total number of genotypes ($500mk$), where m is the number of individuals and k is the number of loci.

Figure 1 shows nominal versus true α, calculated using Monte Carlo simulation, for a four-point linkage analysis using several combinations of θ. Note that in some cases, α quickly increases to 1 after nominal α exceeds a certain threshold depending on θ. This is an artifact of the discreteness of pedigree data. Since individuals have discrete genotypes, the test statistic, $\tilde{\lambda}_{ij}(\mathbf{X})$, assumes a finite number of values. Not surprisingly many of the 500 values of $\tilde{\lambda}_{ij}(\mathbf{X})$ are identical. Therefore, for larger values of nominal α, the proportion of significant test statistics approaches 1. The effect will be more pronounced for small θ since the loci are by definition dependent. Apart from this effect, true α closely follows nominal α when nominal $\alpha < 0.04$. The graphs illustrating the two and three-point analyses are omitted, but the results are similar.

Our initial investigations indicated that the test would have the greatest power for individuals in the third generation of a pedigree, for whom the amount of information present is highest. In order to estimate power, we generated 1500 random pedigrees similar to CEPH pedigrees with two, three, and four loci and various values of θ. We calculated the test statistic for 1000 of these pedigrees and used the values to construct an empirical cumulative distribution function for each individual at each locus in the pedigree. Then we introduced an error into third generation individual 3 at the terminal locus 1 and the internal locus 2 in each of the remaining 500 pedigrees.

The error was introduced by randomly selecting one of the the possible genotypes (different from the original) the individual could have. If the only possible genotype for the individual was the original genotype, the pedigree was not considered (a total of four pedigrees). The test statistic and p-values for each individual and locus were calculated for the remaining 496 pedigrees containing errors. The results for each individual at each locus were pooled. The power was estimated by counting the number of times an error was detected in individual 3 at locus 1 or locus 2 and dividing by 496 for a given nominal α. The true α was again estimated using all other individuals at all loci.

Figure 2 plots the estimate of α versus the power for a four-point analysis for various values of θ when the error was introduced into individual 3 at locus 2. In general, the power is highest for small values of θ and decreases almost to 0 as θ increases to 0.5. While the graphs depicting the power for the two and three-point analyses are not shown, for true $\alpha \approx 0.10$ and $\theta = 0.02$, the power increases steadily from 79% in the two-point analysis to 81% in the three-point analysis to 85% in the four-point analysis. The power is greater when attempting to detect errors in internal loci (the middle loci in three and four-point analyses) as compared with detecting errors in the terminal loci. In summary, the power is inversely related to the recombination fraction and directly related to the number of loci jointly analyzed.

We also estimated the power to detect errors in the grandparental generation. We attempted to introduce an error into individual 10 at locus 1 in the remaining 500 pedigrees. We succeeded 86 times indicating that errors consistent with Mendelian inheritance are less likely to occur in this part of the pedigree. The 86 pedigrees were used to estimate power. The power to detect errors in the top generation was relatively low. Using two-point analyses, the method was able to detect errors in the first generation approximately 30% of the time with an expected false positive rate of 30%. Using three and four-point analyses, the index detected first generation errors approximately 30% of the time with an expected false positive rate of 18%. The poor results are a consequence of the lack of information in this part of the pedigree. These results are comparable to those obtained by Brzustowicz *et* *al* [1] (see discussion).

Our results do not include estimates of the power to detect errors in the second generation of the pedigree. We attempted to introduce an error into individual 1 at locus 1 in 500 pedigrees. We succeeded 18 times. Errors consistent with Mendelian inheritance are relatively unlikely in the second generation of a three generation pedigree. We did not perform further studies since 18 is not a large enough or representative sample on which to base estimates of power and significance.

3.2 Idiopathic Dilated Cardiomyopathy Data

In order to determine how the method may work when applied to experimental pedigrees, we used the structure of a pedigree for the study of idiopathic dilated cardiomyopathy provided by Dr. R. Roberts' laboratory at Baylor College of Medicine in Houston. The markers analyzed were on chromosome 17. The pedigree consists of 45 individuals with typing data from 36 individuals (see Figure 3). Individuals without typing data are labeled UNK (unknown). There are 15 markers of interest with an average two-point recombination fraction of 0.096. We start with a description of the simulation studies used to establish significance and power.

The simulation studies used the DCM pedigree structure with three alleles at each locus with the alleles having equal population frequencies. Simulations included two and three-point analyses with $\theta = 0.1$ and $\theta = (0.1, 0.1)$, respectively. While the estimated α does not correspond exactly to the nominal α, the inflation is small. For example, when nominal $\alpha = 0.05$, the estimated true $\alpha = 0.057$ for a two-point analysis and $\alpha = 0.063$ for a three-point analysis.

The power to detect an error in individuals 26, 32, 43, 44, and 45 at locus 2 using two and three-point analyses with $\theta = 0.1$ and $\theta = (0.1, 0.1)$ was estimated using the simulation method outlined in the results section. These particular individuals were chosen because the power obtained for them would be representative of the part of the pedigree in which they are located. For true $\alpha \approx 0.06$, the powers for the two-point analysis ranged from 39.5% to 60.2%. The powers for the three-point analysis ranged from 61.8% to 85.0%. In addition to observing higher power in the multi-point runs, the power was greater in individuals with more siblings and typed grandparents to help determine phase. Siblings seem particularly important in cases where only two grandparents were typed.

Subsequently, we carried out an experiment aimed at testing our method under conditions close to laboratory practice. We identified individuals and loci likely to contain errors in the DCM pedigree. The laboratory re-read the gels and rechecked the computer files for the genotypes in question. Using the results, we estimate the true and false positive rates for the method and the error rate for the pedigree.

The loci were ordered according to published maps. For computational reasons the 15 loci were partitioned into five two-point runs and two three-point runs. We assumed $\varepsilon = 0.01$. In order to determine significance, 100 pedigrees with the same familial structure, allele frequencies and recombination fractions were generated and used to determine an empirical cumulative distribution function for the test statistic for each individual and locus in the pedigree. Using a nominal $\alpha = 0.05$, the method identified 79 genotypes as likely to contain error.

The laboratory re-read the gels and rechecked the computer files for these particular individual/locus combinations and confirmed 12 errors. The laboratory also found 11 additional errors which were not flagged by the program. Assuming that the laboratory can detect true errors 100% of the time, we performed a hypothesis test to determine if the error checking procedure was beneficial. Let X be the event that the method identifies an individual, locus combination as likely to contain error. Let Y be the event that the laboratory detects a true error in an individual at a locus. Table 1 gives the two by two contingency table summarizing the results. The χ^2_1 test statistic for testing the null hypothesis that X and Y are independent is 19.04 with a p-value of 10^{-5}. Therefore we accept the alternative hypothesis and conclude that our error detection method is beneficial.

The best way to visualize the meaning of the data in Table 1 is to provide the theoretical counterpart of this table given false positive rate α, power $1-\beta$ and true error rate ε (see Table 2). It is assumed, for the moment, that the laboratory's false positive rate is 0 and their power is 1: the laboratory can detect true errors 100% of the time. Comparing Tables 1 and 2 the estimated false positive rate is $\hat{\alpha} = 0.15$, the estimated power (true positive rate) is $1 - \hat{\beta} = 0.52$ and the estimated overall error rate is $\hat{\varepsilon} = 0.053$. The power, $1 - \hat{\beta} = 0.52$, is close to that expected based on simulations ($1 - \beta \approx 0.6$). The estimated value $\hat{\alpha} = 0.15$ is higher than the value expected from simulations ($\alpha = 0.063$ when using nominal $\alpha = 0.05$). There are two possible explanations for this inflation: propagation of errors and rechecking of the genotypes. Since the genotypes for a single locus from different individuals within a pedigree are dependent, their corresponding test statistics will be dependent. Therefore a typing error in one individual may influence the test statistics of other individuals, thereby increasing the number of false positives. Another explanation is that re-reading of the gels and checking the computer files by the laboratory are not error-free themselves. If we assume that these procedures have joint power less than 1 (such as $1 - \beta' = 0.7$), then $\hat{\alpha}$ will be lower and $\hat{\varepsilon}$ higher (e.g. $\hat{\alpha} = 0.14$ and $\hat{\varepsilon} = 0.075$). The large number of individual/locus pairs flagged by the method which correspond to no errors is related to the generally high value of the specificity $(1 - \varepsilon)\alpha$ caused by the fact that error rate, ε, is usually less than 0.1. This disadvantage will be common to all statistical error detection methods.

The errors identified were removed from the pedigree file and the analysis was run again. There were 29 genotypes flagged as likely to contain error. Again, if we assume that the laboratory detection rate is 1, then the false positive rate would be 0.066. None of the genotypes with errors that were removed were flagged again and several of the genotypes that were erroneously flagged were not flagged again. This suggests that error propagation is indeed present. Figure 4 shows the published map for this region and the maps with distances estimated using two-point analyses before and after error checking. The Haldane map function was used to transform the recombination fractions. Removal of errors resulted in considerable shrinking of the length of the map (see [12]) although it is still longer than the published map.

3.3 CEPH Pedigrees - Chromosome 6 Data

Using our error detection algorithm, we tested CEPH chromosome 6 data [19] for errors and derived estimates of the error frequency, ε_j, for third generation individuals for each locus, $j = 1, \ldots, 41$. This dataset consists of 65 families and 41 markers located on chromosome 6. Testing for errors entailed determining which loci were typed in each family, ordering the loci according to the current CEPH chromosome 6 consortium map and partitioning the list of loci into sets of two, three and four loci which became two, three and four-point analyses. For each set the vector θ was estimated and test statistics for each individual at each locus in the run were calculated. Then using the structure of the pedigree, allele frequencies and estimates of θ, 200 random pedigrees were generated for each run and used to construct an empirical cumulative distribution function for each individual at each locus in the run.

The empirical cumulative distribution function was used to calculate p-values for each individual at each locus in all two, three and four-point runs.

In order to estimate ε, a maximum likelihood approach was used. Let

$$Y_{ijk} = \begin{cases} 1 & \text{if error detected in pedigree } k, \text{ individual } i, \text{ locus } j \\ 0 & \text{otherwise.} \end{cases}$$

The probability of a likely error in pedigree k in individual i at locus j can be written:

$$\begin{aligned} \mathcal{P}(Y_{ijk} = 1) &= \mathcal{P}(Y_{ijk} = 1 \mid H_0)\mathcal{P}(H_0) + \mathcal{P}(Y_{ijk} = 1 \mid H_1)\mathcal{P}(H_1) \\ &= \alpha(1-\varepsilon) + (1-\beta)\varepsilon \end{aligned}$$

where ε is the probability of an error in pedigree k in individual i. The null and alternative hypotheses are H_0: no error in pedigree k in individual i versus H_1: error in pedigree k in individual i. Let N_{jk} be the number of individuals tested for errors at locus j in pedigree k, $N'_{jk} = \sum_{i=1}^{N_{jk}} Y_{ijk}$ be the number of individuals with apparent errors detected at locus j in pedigree k, and n be the number of families tested. Let $N''_{jk} = N_{jk} - N'_{jk}$. The log-likelihood can be written:

$$l(\mathbf{Y}; \varepsilon_j, \alpha) = \sum_{k=1}^{n} \left\{ N'_{jk} \left[\alpha(1-\varepsilon_j) + (1-\beta)\varepsilon_j\right] + N''_{jk} \left[1 - \alpha(1-\varepsilon_j) - (1-\beta)\varepsilon_j\right] \right\}.$$

$l(\mathbf{Y}; \varepsilon_j, \alpha)$ can be approximated by letting $\alpha = 0.02$ be the nominal Type I error probability and $1 - \beta$ be the power estimate given nominal $\alpha = 0.02$ and the vector θ for pedigree k.

The maximum likelihood estimate of ε_j was obtained using numerical maximization of $l(\mathbf{Y}; \varepsilon_j, \alpha = 0.02)$. In order to calculate a confidence interval for $\hat{\varepsilon_j}$, we used the logit transformation, $x(\varepsilon_j) = \ln[\varepsilon_j/(1-\varepsilon_j)]$, with inverse $\varepsilon_j(x) = \varepsilon^x/(1+\varepsilon^x)$, to calculate $\hat{\sigma}_x^2$ using Fisher's information, $\hat{\sigma}_x^2 = -\left[\partial^2 l_1(\mathbf{Y}; x, \alpha) / \partial x^2\right]^{-1}$ where

$$l_1(\mathbf{Y}; x, \alpha) = l(\mathbf{Y}; x(\varepsilon_j), \alpha).$$

The inverse transformation was applied to yield the 95% confidence interval for $\hat{\varepsilon_j}$ given by

$$\varepsilon_j(\hat{x} \pm 1.96\hat{\sigma}_x).$$

Figure 5 gives the 95% confidence intervals for $\hat{\varepsilon}_j$, the maximum likelihood estimate of an error at locus j, for $j = 1, \ldots, 41$. When $\hat{\varepsilon}_j = 0$, the confidence bounds could not be estimated and are omitted on the graph. $\hat{\varepsilon}_j$ is plotted versus the distance from the terminal end of the short arm of chromosome 6 (in Morgans). The solid line at $\hat{\varepsilon} = 0.068$ is the global estimate of ε.

Note that $\hat{\varepsilon}_j$ is fairly constant and close to 0.068 for many of the markers. Three markers show exceptionally high error estimates ($\varepsilon_j > 0.15$): PGK1P2, D6S44, and D6S39. Three other markers show relatively high error estimates: TCTE1, D6S30 and MYB.

4 Discussion

While many laboratories use only basic error detection schemes, several authors have considered sophisticated methods of detecting pedigree and typing errors. Ott [16] proposes locating individuals likely to contain errors by comparing the predicted genotype conditional on individual phenotype with the predicted genotype conditional on all the pedigree data under the model of incomplete penetrance. The first predictor is given by:

$$\mathcal{P}(g_{ij} \mid x_{ij}) = \frac{\mathcal{P}(x_{ij} \mid g_{ij})\mathcal{P}(g_{ij})}{\sum_{h_{ij}} \mathcal{P}(x_{ij} \mid h_{ij})\mathcal{P}(h_{ij})}$$

where the sum is over all possible genotypes for individual i at locus j. The second predictor is given by:

$$\mathcal{P}(g_{ij} \mid x_{1j}, \ldots, x_{mj}) = \frac{\mathcal{P}(x_{1j}, \ldots, x_{mj}, g_{ij}, \varepsilon)}{\sum_{h_{ij}} \mathcal{P}(x_{1j}, \ldots, x_{mj}, h_{ij}, \varepsilon).}$$

The comparison is made using:

$$SS_{ij} = \sum_{g_{ij}} \{\mathcal{P}(g_{ij} \mid x_{ij}) - \mathcal{P}(g_{ij} \mid x_{1j}, \ldots, x_{mj})\}^2.$$

Large values of SS_{ij} indicate individuals at loci likely to be errors. While this method was first proposed to identify individuals leading to inconsistencies, Brzustowicz *et al* [1] (including Ott) apply it in conjunction with the "one family out" (OFO) method. The OFO method involves identifying families possibly containing errors by comparing the lod score obtained using all families to that obtained leaving the i^{th} family out. If leaving out the i^{th} family increases the lod score, then it is likely to contain errors. After families likely to contain errors are identified, individuals likely to contain errors are pinpointed using the conditional genotypes. They showed that retyping approximately one-third of the individuals with the highest SS_{ij} caught roughly half of the errors. Our method seems to be more sensitive.

Lincoln and Lander propose a similar approach. They also assume the structure of incomplete penetrance and then calculate the index:

$$LOD_{error}(i,j) = \log_{10}\left\{ \frac{1-\mathcal{P}(g_{ij} \mid x_{mj},\ldots,x_{mj},\hat{\theta},\varepsilon)}{\mathcal{P}(g_{ij} \mid x_{mj},\ldots,x_{mj},\hat{\theta},\varepsilon)} \bigg/ \frac{\varepsilon}{1-\varepsilon} \right\}.$$

Again individuals at all loci exhibiting high values for LOD_{error} are identified as likely to contain errors. The method proved effective for detecting errors in internal loci for real and simulated data corresponding to informative markers in experimental animals. Whether or not it would be effective for human pedigrees is an open question.

The present paper describes a comprehensive study of the test for pedigree and typing errors given in Ehm*et al* [5] including power and significance studies and examples. The simulation studies investigating the performance of $\tilde{\lambda}_{ij}(\mathbf{X})$ using two, three and four-point analyses for various values of θ are the most comprehensive and complete studies performed for the evaluation of an error detection scheme to date. The results obtained using the CEPH pedigree structure show that the test has a slightly increased false positive rate given a nominal α. The test has good power to detect errors in individuals with typed parents and grandparents in internal and terminal loci using two, three and four-point analyses when $\theta < 0.05$ and nominal α in the range $0.02 < \alpha < 0.04$. Higher power is achieved for internal loci as compared to terminal loci. Power increases for multi-point analyses when compared to two-point analyses. The magnitude of this latter effect is smaller than was expected. As θ increases, the power to detect errors drops steadily. The test had low power to detect errors in the first generation and we did not perform power studies for detecting errors in the second generation. The conclusions from these studies are applicable to three generation pedigrees of the CEPH type, and more complex pedigrees extending the regular CEPH structure over a complete range of θ.

The simulation results for the DCM pedigree using two and three-point analyses with $\theta = 0.1$ are slightly different than the results for the CEPH pedigrees. The DCM pedigree showed an inflated false positive rate given a nominal α. The power of the method was highly variable depending on the nature of the pedigree near the individual of interest. In contrast to the results obtained for the CEPH pedigrees, the power showed a substantial increase for the three-point analyses when compared to two-point analyses. This effect may be present since in experimental pedigrees, many untyped individuals make individuals uninformative in two-point analyses. This disadvantage can be overcome by using three and four-point analyses since phase can often be determined by analyzing more loci simultaneously.

As a practical application of our method, we tested the DCM pedigree collected by Dr. R. Roberts' laboratory for typing errors. A total of 79 genotypes were flagged as likely to contain errors and 12 errors were confirmed. Reanalyzing the pedigree decreased the length of the genetic map toward the length of the published map. The lower bound of the error rate in the DCM pedigree was estimated to be $\hat{\varepsilon} = 0.053$.

In another experiment, we analyzed publicly available CEPH data from chromosome 6 and found an overall error estimate of $\hat{\varepsilon} = 0.068$. We have no way to double check the validity of these estimates to provide independent corroboration of the results. However, we will be happy to provide more details to those interested in this chromosomal region.

Code Distribution

A variant of the error detection program described above is available from the anonymous FTP server softlib.cs.rice.edu. Please contact the first author by electronic mail (gelder@rice.edu) for specific instructions.

Calculating the p-values associated with the test statistic given is computationally intensive and may not be practical for many laboratories. The publically available code calculates the values of the test statistic and generates a file containing a list of the locus/individual combinations and their associated test statistics in order of decreasing test statistic.

The test statistics with relatively large values are indicative of an unlikely genotype for that individual at that locus. The ordered list of locus/individual combinations given should be thought of as a priority list for retyping. Interpreting an error checking run includes the following two steps. First, reread gels and check computer file entries for those in the top 20% of the list. If there are any errors, correct them and run the analysis again. Second, retype each individual/locus combination in the top 10% of the list. If there are no errors, then stop error checking. If errors are present, correct them and run the analysis again.

5 Acknowledgments

The authors appreciate Dr. Alejandro Schäffer for his helpful comments and for providing assistance with FASTLINK. We are also grateful to Dr. Linda Bachinski and Dr. Robert Robert's laboratory at Baylor College of Medicine for providing and maintaining the idiopathic dilated cardiomyopathy data. M. Ehm's research is supported by grants from the U. S. National Library of Medicine Medical Informatics Research Training Program and the W.M. Keck Foundation. M. Kimmel's research has been supported by NSF Grants 9203436 and 9409909 and the W.M. Keck Foundation. R. W. Cottingham's research is supported by a grant from the National Center for Human Genome Research.

References

[1] L.M. Brzustowicz, C. Merette, X. Xie, L. Townsend, T.C. Gilliam, and J. Ott. Molecular and statistical approaches to the detection and correction of errors in genotype databases. *American Journal of Human Genetics*, 53:1137–1145, 1991.

[2] K.H. Buetow. Influence of aberrant observations on high-resolution linkage analysis outcomes. *American Journal of Human Genetics*, 49:985–994, 1991.

[3] R.W. Cottingham Jr., R.M. Idury, and A.A. Schäffer. Faster sequential genetic linkage computations. *American Journal of Human Genetics*, 53:252–263, 1993.

[4] N.C. Dracopoli, P. O'Connel, T.I. Elsner, J-M. Lalouel, R.L. White, K.H. Buetow, D.Y. Nishimura, and *et al.* The CEPH consortium linkage map of human chromosome 1. *Genomics*, 9:686–700, 1991.

[5] M. Gelder Ehm, R.W. Cottingham Jr., and M. Kimmel. Error detection in genetic linkage data for human pedigrees using likelihood ratio methods. *Journal of Biological Systems*, 3(1), 1995.

[6] B.J.B. Keats, S.L. Sherman, and J. Ott. Report of the committee on linkage and gene order. *Cytogenetics and Cell Genetics*, 55:387–394, 1990.

[7] E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences USA*, 84:2363–2367, 1987.

[8] G. M. Lathrop and J-M. Lalouel. Easy calculations of lod scores and genetic risks on small computers. *American Journal of Human Genetics*, 36:460–465, 1984.

[9] G. M. Lathrop, J-M. Lalouel, and R. L. White. Construction of human genetic linkage maps: Likelihood calculations for multilocus analysis. *Genetic Epidemiology*, 3:39–52, 1986.

[10] G.M. Lathrop, A.B. Hooper, J.W. Huntsman, and R.H. Ward. Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker mistyping. *American Journal of Human Genetics*, 25:241–262, 1983.

[11] G.M. Lathrop, J.M. Lalouel, C. Julier, and J. Ott. Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences*, 8:3443–3446, 1984.

[12] S.E. Lincoln and E.S. Lander. Systematic detection of errors in genetic linkage data. *Genomics*, 14:604–610, 1992.

[13] N.E. Morton and A. Collins. Standard maps of chromosome 10. *Annals of Human Genetics*, 54:235–251, 1990.

[14] J. Ott. Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences*, 86:4175–4178, 1989.

[15] J. Ott. *Analysis of human genetic linkage*. Johns Hopkins University, Baltimore, revised edition, 1991.

[16] J. Ott. Detecting marker inconsistencies in human gene mapping. *Human Heredity*, 43:25–30, 1993.

[17] A.A. Schäffer, S.K. Gupta, K. Shriram, and R.W. Cottingham Jr. Avoiding recomputation in genetic linkage analysis. *Human Heredity*, 44:225–237, 1994.

[18] A. Stuart and K. Ord. *Kendall's advanced theory of statistics*, volume 2. Oxford University Press, New York, fifth edition, 1991.

[19] A. Volz, J.M. Boyle, H.M. Cann, R. W. Cottingham Jr., H. T. Orr, and Andreas Ziegler. Report of the second international workshop on human chromosome 6. *Genomics*, 21:464–472, 1994.

[20] D.E. Weeks, J. Ott, and G.M. Lathrop. SLINK: A general simulation program for linkage analysis. *American Journal of Human Genetics*, 47:A204, 1990. Abstract.

6 Table and Figure Legends

Table 1: Contingency table summarizing results for DCM pedigree error analysis. The χ_1^2 test statistic for testing the null hypothesis that X and Y are independent is 19.04 with a p-value of 10^{-5}.

Table 2: Model used to describe results for DCM pedigree error analysis given false positive rate α, power $1-\beta$ and true error rate ε.

Figure 1: Nominal versus true α simulated using four-point analysis for a pedigree typical of the CEPH pedigrees. Using Monte Carlo simulation, 1500 random pedigrees having the same structure as pedigree 13292, but with various values of θ were generated. 1000 pedigrees were used to construct empirical cumulative distribution functions for each individual at each locus. For a given significance level (nominal α), the number of significant test statistics in the remaining 500 pedigrees were counted. The results for all individuals at all loci were pooled and divided by the number of genotypes ($500mk$) in order to estimate the true α.

Figure 2: True α versus power for four loci. Using Monte Carlo simulation, 1500 random pedigrees having the same structure as pedigree 13292, but with various values of θ were generated. 1000 pedigrees were used to construct empirical cumulative distribution functions for each individual at each locus. In the remaining 500 pedigrees we attempted to introduce an error in a third generation individual (individual 3 at locus 2). The test statistic and p-values for each individual at each locus were calculated for the 496 pedigrees containing errors. The power was estimated by counting the number of times an error was detected in individual 3 at locus 2 divided by 496. True α was estimated by counting the number of times the test statistic was found significant in all other individuals at all loci.

Figure 3: The DCM pedigree investigated for typing errors.

Figure 4: Published map for the investigated region of chromosome 17 along with the map generated by using two-point analyses before and after error analysis. The Haldane map function was used to transform the recombination fractions. Note that the map distances are generally shorter after the errors were removed, but still longer than the published map.

Figure 5: $\hat{\varepsilon}_j$ versus the distance from the terminal end of the short arm of chromosome 6. CEPH data for chromosome 6 was tested for errors using our method. Using a maximum likelihood approach, ε and $\hat{\sigma}(\varepsilon)$ were estimated globally and for each locus, j. The 95% confidence intervals are represented by error bars.

7 Tables

	Flagged	Not Flagged	Total
Error	12	11	23
No Error	67	347	414
Total	79	358	437

Table 1: Contingency table summarizing results for DCM pedigree error analysis.

	Flagged	Not Flagged	Total
Error	$N\varepsilon(1-\beta)$	$N\varepsilon\beta$	$N\varepsilon$
No Error	$N(1-\varepsilon)\alpha$	$N(1-\varepsilon)(1-\alpha)$	$N(1-\varepsilon)$
Total	$N[\varepsilon\beta+(1-\varepsilon)(1-\alpha)]$	$N[\varepsilon\beta+(1-\varepsilon)(1-\alpha)]$	N

Table 2: Model used to describe results for DCM pedigree error analysis.

8 Figures

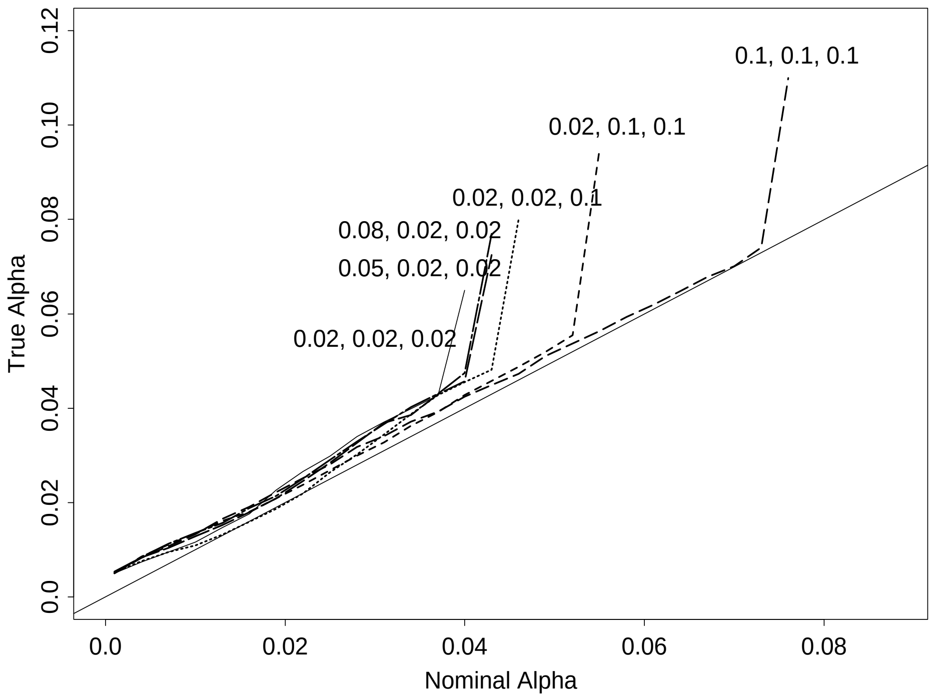

Figure 1:

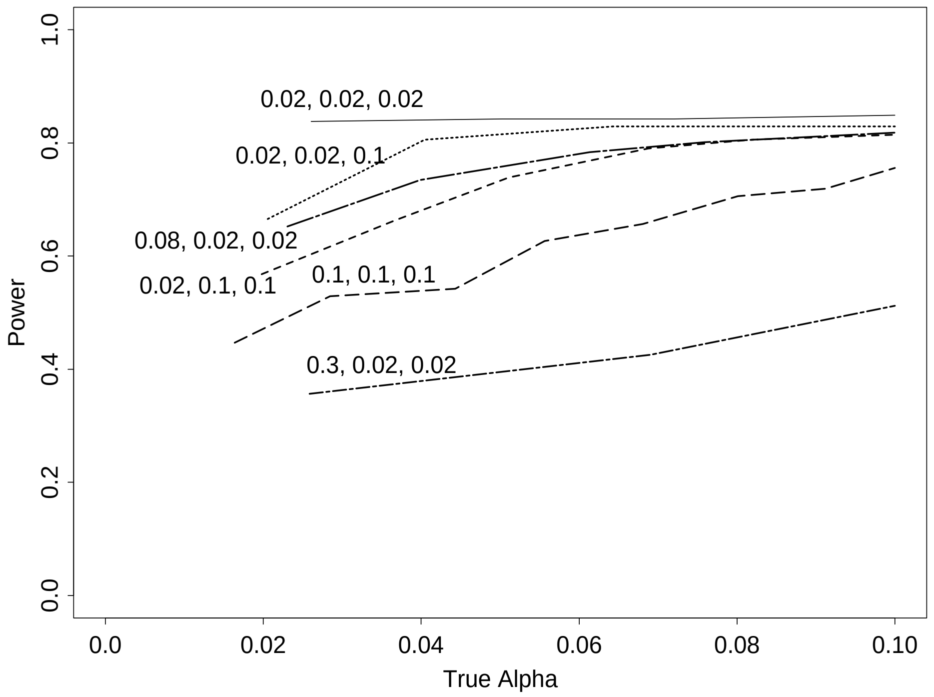
0.02, 0.02, 0.02
0.02, 0.02, 0.1
0.08, 0.02, 0.02
0.02, 0.1, 0.1
0.1, 0.1, 0.1
0.3, 0.02, 0.02
Power
True Alpha
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.02
0.04
0.06
0.08
0.10

Figure 2:

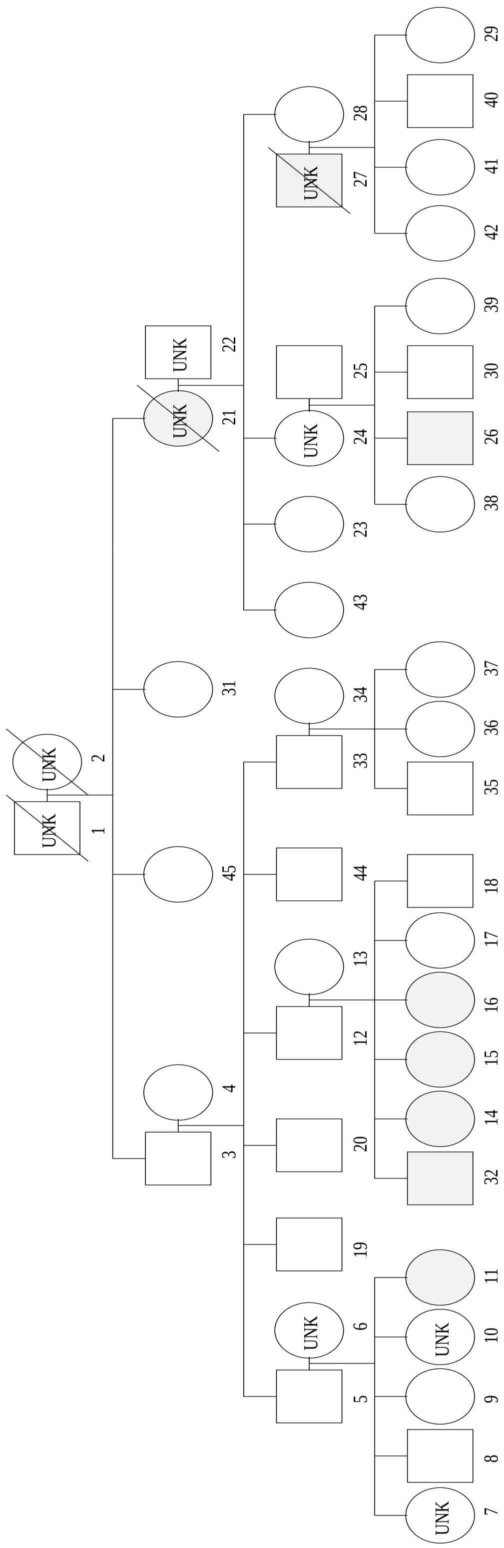
UNK 1
UNK 2
3
4
45
31
UNK 21
UNK 22
5
UNK 6
19
20
12
13
44
33
34
43
23
UNK 24
25
UNK 27
28
UNK 7
8
9
UNK 10
11
32
14
15
16
17
18
35
36
37
38
26
30
39
42
41
40
29

Figure 3:

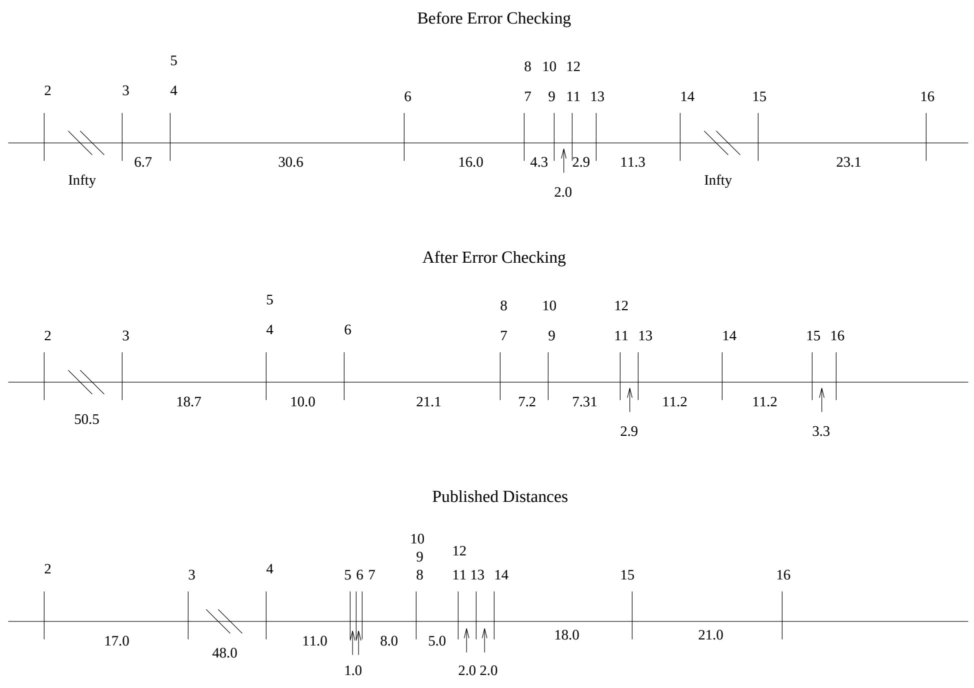
Before Error Checking
2
3
5
4
6
8 10 12
7 9 11 13
14
15
16
Infty
6.7
30.6
16.0
4.3
2.9
11.3
Infty
23.1
2.0
After Error Checking
2
3
5
4
6
8 10 12
7 9 11 13
14
15 16
50.5
18.7
10.0
21.1
7.2
7.31
11.2
11.2
2.9
3.3
Published Distances
2
3
4
5 6 7
10
9
8
12
11 13 14
15
16
17.0
48.0
11.0
8.0
5.0
18.0
21.0
1.0
2.0 2.0

Figure 4:

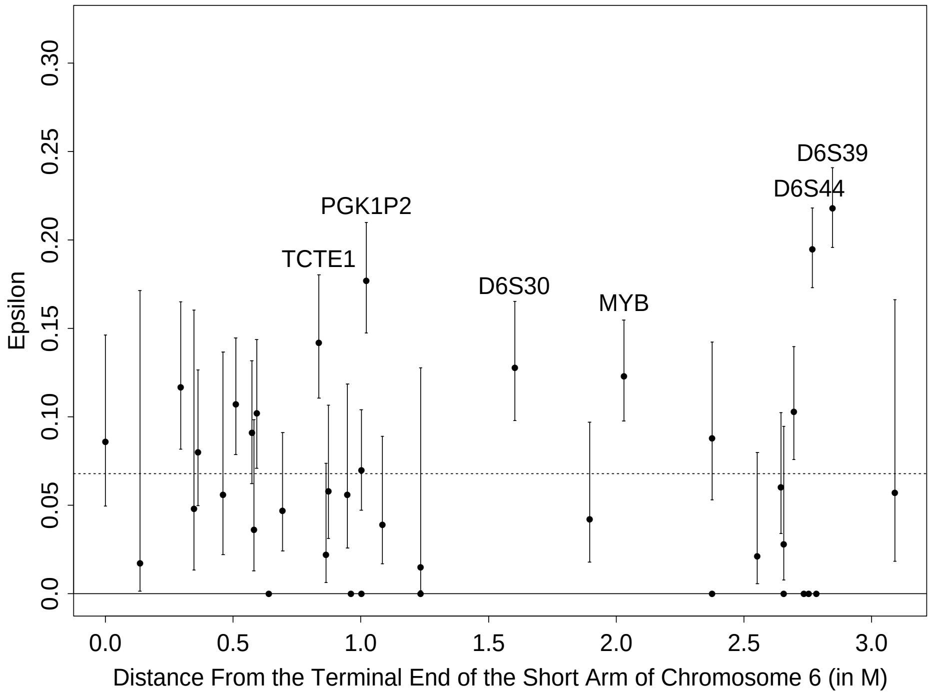
Epsilon
0.0
0.05
0.10
0.15
0.20
0.25
0.30
PGK1P2
TCTE1
D6S30
MYB
D6S39
D6S44
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Distance From the Terminal End of the Short Arm of Chromosome 6 (in M)

Figure 5: