Published in final edited form as:

Genet Epidemiol. 2014 May ; 38(4): 291–299. doi:10.1002/gepi.21806.

Detection of Mendelian Consistent Genotyping Errors in Pedigrees

Charles Y. K. Cheung^{1,2}, Elizabeth A. Thompson³, and Ellen M. Wijsman^{1,2,4}

¹Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA

²Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, 98195, USA

³Department of Statistics, University of Washington, Seattle, WA, 98195, USA

⁴Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

Abstract

Detection of genotyping errors is a necessary step to minimize false results in genetic analysis. This is especially important when the rate of genotyping errors is high, as has been reported for high-throughput sequence data. To detect genotyping errors in pedigrees, Mendelian inconsistent (MI) error checks exist, as do multi-point methods that flag Mendelian consistent (MC) errors for sparse multi-allelic markers. However, few methods exist for detecting MC genotyping errors, particularly for dense variants on large pedigrees. Here we introduce an efficient method to detect MC errors even for very dense variants (e.g. SNPs and sequencing data) on pedigrees that may be large. Our method first samples inheritance vectors (IVs) using a moderately sparse but informative set of markers using a Markov chain Monte Carlo-based sampler. Using sampled IVs, we considered two test statistics to detect MC genotyping errors: the percentage of IVs inconsistent with observed genotypes (A1) or the posterior probability of error configurations (A2). Using simulations, we show that this method, even with the simpler A1 statistic, is effective for detecting MC genotyping errors in dense variants, with sensitivity almost as high as the theoretical best sensitivity possible. We also evaluate the effectiveness of this method as a function of parameters, when including the observed pattern for genotype, density of framework markers, error rate, allele frequencies, and number of sampled inheritance vectors. Our approach provides a line of defense against false findings based on the use of dense variants in pedigrees.

Keywords

data cleaning; high-throughput; computer program; exome; genome-scan

Introduction

A major goal in human genetics is to identify DNA variants that affect risk of heritable diseases. Family studies have successfully led to the identification of over 4,500 genes

*Correspondence: Ellen M. Wijsman, Department of Medicine, Division of Medical Genetics, Box 357720, University of Washington, Seattle, WA 98195-7720. (206) 543-8987; fax: (206) 616-1973. wijsman@u.washington.edu.

associated with heritable diseases [Amberger, et al. 2011]. In these studies, the accuracy of genotype measurements ultimately affects the validity of these analyses. Despite the necessity to obtain accurate genotype measurements, technologies used to determine genotypes for Single Tandem Repeats (STRs) [Weeks, et al. 2002], Single Nucleotide Polymorphisms (SNPs) [Hinrichs and Suarez 2005], and sequence data [Zhi, et al. 2012] are all subject to genotyping errors.

Genotyping errors affect the validity of results in genetic studies. They affect the quality of genetic maps [Buetow 1991; Goldstein, et al. 1997; Hackett and Broadfoot 2003], haplotype reconstruction [Kirk and Cardon 2002], and measures of Linkage-Disequilibrium [Akey, et al. 2001; Hackett and Broadfoot 2003]. They can reduce power in association analysis [Abecasis, et al. 2001; Gordon, et al. 2002] and affect linkage analysis [Cherny, et al. 2001; LeBrec, et al. 2008; Lincoln and Lander 1992; Stringham and Boehnke 1996] at moderately low genotyping error rate [Abecasis, et al. 2001; Douglas, et al. 2000; Walters 2005], and can lead to false evidence of linkage [Douglas, et al. 2002]. Genotyping errors continue to impact results from analyses performed using current genotyping technologies. For instance, genotyping errors affect the type I error rates in the transmission disequilibrium test [Gordon, et al. 2001; Mitchell, et al. 2003] and haplotype-based tests [Knapp and Becker 2004]. Furthermore, genotyping errors can additionally propagate to imputed genotypes, thus affecting downstream analyses based on the imputed data. Hence, genotyping errors should first be corrected or removed, especially if selected analytical methods do not explicitly account for these errors. This requires first identifying possible errors.

Genotyping errors will continue to present challenges to studies based on high-throughput sequence data. At present, the rate of genotyping errors in high-throughput sequence data is still quite high [Zhi, et al. 2012]. Therefore, traditional analyses performed using these new sequence data inherit, and may even exaggerate all existing problems associated with analyses performed on genotypes derived from older technologies. One concern is that some of the identified rare variants may be artifacts of genotyping errors.

In pedigrees, genotyping errors are either Mendelian inconsistent (MI) or Mendelian consistent (MC). A MI genotyping error is an error that is detected because the observed genotypes are not consistent with the transmission pattern as specified by the Mendel's First Law. Specialized program such as PedCheck [O'Connell and Weeks 1997] and many programs to perform linkage analyses [Abecasis, et al. 2002; Lathrop, et al. 1983; Sobel, et al. 2002] can be used to detect MI errors. When a marker is flagged as Mendelian inconsistent (MI), this marker most likely has either a genotyping error or a *de-novo* mutation. The fraction of errors that escapes the Mendelian inconsistency check depends largely on a particular pedigree structure [Mukhopadhyay, et al. 2004] and can be high, especially if a considerable numbers of subjects are unobserved for genotypes [Douglas, et al. 2002; Gordon, et al. 1999]. Genotyping errors that are not MI are MC, and they cannot be detected by using MI checks.

Existing approaches for detecting errors are not suitable for use on large pedigrees with many dense markers, such as variants identified by sequence data. One class of approach aims to detect "excess recombinants". While existing methods under this class of approach

are effective for detecting MC genotyping errors, they are limited to use in sibling-pairs [Douglas, et al. 2000], trios [Sieberts, et al. 2002], or small pedigrees [Abecasis, et al. 2002], and are not computationally feasible for large pedigrees with a large number of markers [Sobel, et al. 2002]. Another class of approach detects MC genotyping errors by using information from Linkage Disequilibrium [Becker, et al. 2006; Kennedy, et al. 2008; Zou, et al. 2003]. These approaches use observed genotypes from multiple unrelated subjects or families to estimate haplotype frequencies and detect errors by computing test statistics based on likelihood ratios. However, these approaches generally have practical limitations, which include (1) the requirement of having externally phased haplotypes that may not be available, (2) a requirement that markers are tightly linked, (3) limitation to the use in trios and nuclear families, and (4) potentially inadequate performance for identifying errors for rare variants because of the lack of appropriate reference samples. As large pedigrees are important for studying complex heritable diseases [Wijsman 2012] and for implicating rare sequence variants [Marchani, et al. 2012], there is a clear need to develop methods for detection of genotyping errors in current markers with density far exceeding that used in the past.

Here we introduce a computationally efficient method to detect MC genotyping errors in dense markers. In contrast to existing methods, our method now enables the detection of such errors in large pedigrees. In a simulated large pedigree, we show that our method can detect a high percentage of MC genotyping errors. In addition, this approach can also identify which subject most likely is the source of the error. We evaluate several parameters and conditions that affect the performance of error detection. Our approach has been implemented in GIGI-Check and is available at <http://faculty.washington.edu/wijsman/software.shtml>.

Materials and Methods

A motivating example

While genotyping errors in sparse multiallelic markers are known to affect linkage results [Abecasis, et al. 2001; Buetow 1991; Douglas, et al. 2000], there has been an assumption that low error rates for SNPs would eliminate the need for error correction. However, errors will still likely be present with a large number of markers. Here, we demonstrate that MC genotyping errors can also affect linkage results from genotype data involving diallelic markers, which is consistent with other results [Markus, et al. 2011]. Using a 52-member pedigree (Figure S1) with trait and descent pattern described in previous studies [Cheung, et al. 2013; Wijsman, et al. 2006], we introduced two genotyping errors in a 20 cM neighborhood surrounding a risk locus in the middle of a 100 cM chromosome. Our results show that even one or two MC genotyping errors can affect linkage analysis on a moderately dense SNP marker panel (Figure S2). In the first example, MC genotyping error attenuated the lod score at the position of the trait locus and deflected the position with strongest evidence of linkage. This could mislead an investigator into discounting the position of the trait locus. In the second example, MC error inflated the maximum lod score.

Overview

Inheritance vectors (IVs) [Lander and Green 1987] represent the inheritance of chromosomes in related individuals and provide a framework to understand existing MI and MC genotyping error detection approaches in pedigrees. MI checks such as used by PedCheck [O'Connell and Weeks 1997] evaluate markers one-by-one. When flagging a marker, if the probability of the observed genotypes of the marker is 0, this requirement is equivalent to checking whether the observed genotypes are incompatible with all configurations of IVs. MC genotyping error detection approaches are an extension to the MI check. These approaches are referred to as multi-point because they use genotypes of neighboring markers to detect genotyping errors. Genotypes of neighboring markers are used to jointly model the transmission of chromosomes, with IVs being a common representation. A low probability of the observed data given IVs inferred by multiple markers provides evidence of MC genotyping errors.

To understand how IVs can be used to identify MC genotyping errors, we first assume that we can obtain the true IV at a locus in the pedigree. The true IV can be used to detect MC genotyping errors by checking if the true IV is consistent with the observed genotypes. We define an IV to be consistent with the observed genotypes if there is at least one legal Mendelian assignment of the observed genotypes to the IV. In the absence of genotyping error, the true IV at the position of the marker would be consistent with the observed genotypes (Figure 1). On the other hand, inconsistency between the observed marker genotypes and the true IV would be an indicator of a genotyping error. Figure 1 uses identity-by-descent graphs (IBD-graphs) [Kruglyak, et al. 1996; Sobel and Lange 1996] to illustrate the use of an IV to detect a MC genotyping error. In this figure, the observed genotypes are consistent with the pedigree structure. However, inconsistency between observed genotypes and the IV suggests the presence of a genotyping error. In reality, any biological phenomenon that leads to departure from the Mendel's First Law (eg. mutation or deletion) also results in inconsistency, but here we would ignore such possibilities.

Even though the true IV is not observable, we can infer IVs probabilistically for use in error detection. To infer IVs, we adopt a previously described approach [Cheung, et al. 2013] in which we sample IVs by using sparse markers, which we refer to as framework markers. Here, the framework markers are assumed to be free of genotyping errors. We distinguish framework markers from dense markers, which we wish to check for MC genotyping errors. Our approach consists of four steps, with the first two steps described elsewhere [Cheung, et al. 2013].

1. Sample IVs at the positions of the framework markers conditional on genotypes of the framework markers
2. Sample a set of n IVs at the positions of the dense markers conditional on the IVs sampled at the positions of the framework markers and the meiotic map
3. Flag each marker as likely to have genotyping error if its test statistic is less than a pre-defined threshold

Details

We first define some notation. Let G_v be the observed genotypes of a dense marker v , X_v be the true genotypes of the dense marker v that may be different from G_v , because G_v may contain genotyping error. Let S_v^k be the k^{th} sampled IV at the dense marker v . In addition, let be the framework markers used to infer the IVs, and let G_F be the observed genotypes of the framework markers. We assume that G_F does not contain genotyping errors.

We propose two test statistics to detect MC genotyping errors at dense marker v , with both approaches flagging a marker if the test statistic is below a pre-defined threshold, c . The percent consistency (A1) is defined as:

$$\%consistency = 100 \frac{1}{n} \sum_{k=1}^n I(P(G_v | S_v^k) > 0)$$

, where the term $I(P(G_v | S_v^k) > 0)$ is an indicator variable that is 1 if $P(G_v | S_v^k) > 0$, and 0 otherwise, and is equivalent to an indicator variable of whether G_v is consistent with S_v^k .

The test statistic A1 is motivated by a simple argument. While genotypes of framework markers are consistent with all sampled IVs, dense markers genotyped at other positions are not guaranteed to be consistent with all sampled IVs [Cheung, et al. 2013]. Because a sampled IV is not necessarily the same as the true IV, inconsistency of a single sampled IV with genotypes of a dense marker does not necessarily indicate the presence of a genotyping error. On the other hand, a low percent consistency among the entire set of sampled IVs provides much stronger evidence that there is a genotyping error.

The estimated posterior probability of the unobserved genotype configuration (A2) is defined as:

$$\hat{P}(X_v = x | G_F, G_v) = \frac{1}{n^*} \sum_{k=1}^n \frac{P(X_v = x | S_v^k) P(G_v | X_v = x)}{\sum_w P(X_v = w | S_v^k) P(G_v | X_v = w)}$$

, where $n^* = \sum_{k=1}^n I(\sum_w P(X_v = w | S_v^k) P(G_v | X_v = w) > 0)$ and the term $I(\sum_w P(X_v = w | S_v^k) P(G_v | X_v = w) > 0)$ is an indicator variable of whether S_v^k is consistent with at least one configuration of w . Here, we assume that if S_v^k is not consistent with any w , then S_v^k is a sampled IV of inadequate quality that should not be used. There is no guaranteed that $n^* = n$ because the sampled IVs are not necessarily the true IV and because our error model assumes the presence of at most one genotyping error, so the set of w may not contain the true joint genotypes. Although the calculation of A2 requires more computation than that of A1 because A2 iterates over multiple w , A2 can be used to identify the subject who most likely contains the genotyping error.

A2 is an estimator of $P(X_v=x \mid G_F, G_v)$ based on Monte Carlo sampling from $P(S_v=s \mid G_F)$. To see this, $P(X_v=x \mid G_F, G_v)$ is derived as follows:

$$\begin{aligned}
 & P(X_v=x \mid G_F, G_v) \\
 &= \sum_s P(X_v=x \mid G_F, G_v, S_v=s) P(S_v=s \mid G_F, G_v) \\
 &= \sum_s \frac{P(X_v=x, G_v \mid G_F, S_v=s)}{P(G_v \mid G_F, S_v=s)} P(S_v=s \mid G_F, G_v) \\
 &= \sum_s \frac{P(X_v=x, G_v \mid G_F, S_v=s)}{P(G_v \mid G_F, S_v=s)} P(S_v=s \mid G_F) \\
 &= \sum_s \frac{P(X_v=x, G_v \mid S_v=s)}{P(G_v \mid S_v=s)} P(S_v=s \mid G_F) \\
 &= \sum_s \frac{P(X_v=x \mid S_v=s) P(G_v \mid X_v=x, S_v=s)}{\sum_w P(X_v=w \mid S_v=s) P(G_v \mid X_v=w, S_v=s)} P(S_v=s \mid G_F) \\
 &= \sum_s \frac{P(X_v=x \mid S_v=s) P(G_v \mid X_v=x)}{\sum_w P(X_v=w \mid S_v=s) P(G_v \mid X_v=w)} P(S_v=s \mid G_F)
 \end{aligned}$$

A related discussion of the inequality in the third and fourth lines of the above equation was presented earlier [Cheung, et al. 2013].

A2 estimates the posterior probability of the joint genotype configuration. When $x=G_v$, $\hat{P}(X_v=x \mid G_F, G_v)$, estimates the posterior probability of no genotyping error at the marker. For each other $x \neq G_v$, $\hat{P}(X_v=x \mid G_F, G_v)$ estimates the posterior probability of another joint genotype configuration. Thus, these estimated probabilities can be used to identify the subject who most likely has a genotyping error. The calculation of A2 is straightforward because $P(X_v=x \mid S_v^k)$ can be efficiently computed [Kruglyak, et al. 1996; Sobel and Lange 1996].

The calculation of the posterior probability in A2 requires an error model in the $P(G_v \mid X_v=g)$ term. We use a simple model. First, the model assumes that genotyping errors are independent at different markers and in different individuals with error probability ϵ . Hence, if subjects are observed at dense marker v , $P(G_v \mid X_v=G_v) = (1-\epsilon)^m$. Second, if there is a genotyping error, the model assumes that only 1 of the 2 alleles is incorrect and each allele has an equal chance to be incorrect. Third, in a marker with allelic types, the model assumes

that the true type of the incorrect allele is any of the other allelic types with an uniform probability $\frac{1}{t-1}$. For instance, if the true joint genotypes and observed joint genotypes G_v differ only at subject p , as denoted by the true genotype x_{pv} and the observed genotype G_{pv} , $p(G_v/X_v=x)=(1-\varepsilon)^{m-1}\varepsilon p(G_{pv}/x_{pv}; G_{pv} \neq x_{pv})$. For a diallelic SNP,

$P(G_{pv}=AA|x_{pv}=Aa; G_{pv} \neq x_{pv})=\frac{1}{2}$ and $P(G_{pv}=Aa|x_{pv}=AA; G_{pv} \neq x_{pv})=1$. This simple model allows efficient computation by restricting the number of possible true genotypes.

To further achieve computational efficiency, we further assume that the pedigree has at most 1 genotyping error at dense marker v . This strategy avoids the task of exhaustive enumeration of all potential true genotyping configurations, which includes unlikely configurations with multiple genotyping errors. Thus, this implementation evaluates at most $1+2m(t-1)$, which is linear in the number of observed genotypes, m , instead of up to t^{2m} configurations, which grows exponentially with m . The expectation is that the method will identify at least one error when there are multiple errors.

Evaluation in simulated data

We used a 52-member simulated pedigree (Figure S1) as previously described [Cheung, et al. 2013] to evaluate our error detection approach. We used 3 designs with genotypes observed for different numbers of subjects to measure the ability to detect MC genotyping errors under different completion rates of observed genotypes: all subjects (a : 52) as the ideal completion rate; many subjects (m : 36) and few subjects (f : 22) as realistic scenarios where only subjects in the most recent pedigree generation are available for genotyping. As will be described, we varied the completion rates in both framework (F) and dense markers (D).

We simulated both framework and dense markers on a chromosome of 100 centi-Morgans (cM) (Figure S3). To infer IVs, we simulated 200 diallelic framework markers spaced uniformly at a density of 1 marker per 0.5 cM. In addition to the framework markers, we simulated 25,000 diallelic dense markers spaced uniformly at a density of 1 marker per 0.004 cM with population allele frequencies simulated from the uniform 0–1 distribution. This large number of dense markers ensures that adequate sample sizes were available for different classes of minor allele frequencies (MAF) considered. For each run, we generated a complete genotype marker dataset by simulating founder alleles and gene-dropping through the descent patterns. Then, we introduced errors on top of the simulated dense genotypes and masked the relevant marker genotypes for analysis of the data as defined by the specific design. For each set of subjects observed for dense markers, genotyping errors were introduced independently at a pre-defined allelic error rate = 0.25%.

We briefly describe the genotyping errors created in the data. Across runs, 16.7% of the markers contained genotyping errors in the default experimental condition. For these markers with genotyping errors, 91.6% of them had only 1 error. Among the error genotypes, 33.3% were homozygous, and 11% of them were in the founders. Additional details, including distribution of errors per run, are discussed in the section below and are tabulated in the Appendix (Table S I–VI).

Evaluation metrics

We used several quality metrics for evaluation. Under each experimental condition, we computed: (1) sensitivity, the percentage of markers with MC genotyping errors that are flagged; (2) specificity, the percentage of markers without MC genotyping errors that are correctly not flagged; and (3) positive predictive value (PPV), the percentage of flagged markers that indeed contained MC genotyping errors. For A2, we also computed the accuracy of identifying subjects who had errors in markers with MC genotyping errors. To compare the performance of the test statistics A1 and A2 over a range of detection thresholds, we also plotted the Receiver Operating Characteristic (ROC) curve, which compares sensitivity vs. 1- specificity as a function of the detection threshold. In most other analyses, A1 was the only calculated test statistic.

For each experimental condition, error detection followed a sequence of three steps. First, we sampled n IVs at the dense positions of interest by using `gl_auto` to sample n IVs at framework positions followed by a previously describe method [Cheung, et al. 2013] that conditions on these framework IVs to sample an equal number of IVs at the dense positions. The program `gl_auto` can handle large pedigrees through Markov chain-Monte Carlo (MCMC) methodology or more modest pedigrees through exact calculation of likelihood followed by Monte Carlo sampling [Thompson 2011]. Second, we removed MI genotyping errors of dense markers using `gl_auto`'s Mendelian Inconsistency check option. At $\epsilon=0.25\%$, 54.5% of errors in run #1 (55.1% over all runs) were MI when all subjects were observed for dense markers (Table S V). However, only 5.2% of errors in run #1 (5.7% over all runs) were MI when only a few subjects were observed (Table S V). The percentage of MI errors increased with lower MAFs (Table S VI), but the percentage of MI errors remains low even at MAFs between 0 and 0.05 (22.7% over all runs) under the default condition when many subjects were observed. Together, these results illustrate that the use of MI check alone to detect genotyping errors is inadequate Third, we detected MC genotyping errors by using either the test statistic A1 or A2 on the remaining markers that passed the MI check.

Default experimental condition

We set up a default experimental condition in which both framework and dense genotypes were typed on the design with “many” (Figure S1) subjects (Design $D_m F_m$). The default error detection threshold was 0.05. If A2 flagged an error, the approach also attempted to identify the subject who most likely had the error. As a benchmark, we also used the simulated true IVs at the dense positions to detect MC genotyping errors. The percentage of markers inconsistent with the true IVs is the ideal percentage of markers that we hope our approach can detect. Since the quality metrics are not as meaningful when summarized as averages, results are reported per run. The default summaries reported here use results from run #1, but results from other runs are also presented whenever there is a need to illustrate inter-run variability. In the default condition, 1000 IVs were sampled.

We evaluated our error detection approach under the default experimental condition for a number of variables. First, we compared the performance of using A1 and A2, as A2 might perform better because it used an error model. Second, we varied the threshold of error

detection from the most stringent detection threshold of 0% to a relaxed threshold of 20% to evaluate the impact of choice of detection threshold. Third, we varied the number of IVs sampled to evaluate the importance of the length of the MCMC run. Fourth, we stratified the markers by bins of MAF to investigate the effect of MAF on quality metrics. Of particular interest was whether our approach can effectively detect MC genotyping errors for markers with low MAF, as the minor alleles of these markers represent rare alleles that might be of interest in a sequencing experiment.

Evaluating parameters that affect detection of genotyping errors

We also evaluated several other parameters that might affect the performance of error detection. First, we varied the number of subjects typed for dense SNP markers by using 3 patterns of observed genotypes (i.e. all (52), most (36), and few (22) subjects observed) to evaluate how different levels of observed data affected error detection. Second, we varied the number of subjects typed for framework markers using the same designs to evaluate whether the use of IVs inferred with fewer subjects negatively affects error detection. Third, we thinned the framework markers to a density of one marker per 1 cM and one marker per 2 cM to test the effect of varying framework marker density.

Results

Main findings

Our approach was effective in detecting MC genotyping errors (Table I). Focusing on the default experimental condition ($D_m F_m$), our approach achieved high sensitivity (88.9%), specificity (99.0%), and PPV (93.5%). This high sensitivity was only 1.9% lower than what could ideally be obtained with using the true IV (90.8%), in which the true IV would have perfect specificity (100%) and thus perfect PPV (100%). Across all runs, the difference in sensitivity with the true IV and with our detection approach was only 1.3% (Table S VII). In addition, test statistic A2 identified the correct subjects containing error with a high accuracy of 84.2%. The test statistics A1 and A2 performed similarly (Figure 2) across a range of thresholds considered, and high ability to detect MC errors was observed across all runs, with the lowest sensitivity, specificity, and PPV still greater than 83%, 97%, and 83%, respectively (Table S VII).

The MAF of dense markers affected quality metrics (Table S VIII). Higher MAF did not tend to increase or decrease sensitivity. However, specificity gradually increased from 98.3% when $MAF=(0.40, 0.50]$ to 100% when $MAF=[0-0.05]$, so the PPV similarly increased from 89.9% to 99.7%. This result suggested that the error detection approach continues to be effective at detecting genotyping errors at very low MAF.

The use of a more stringent detection threshold increased sensitivity but decreased specificity (Figure 2). As the threshold relaxed slightly from the most stringent level of 0 to 0.01, sensitivity increased markedly (81.3% to 88.0%). As the threshold continued to relax, sensitivity only increased very slightly. In contrast, as the threshold relaxed, specificity decreased gradually. Even at the most stringent threshold of 0, specificity was not

guaranteed to be 100% (eg. 99.3% for run #2 in Figure S4). Similar qualitative findings were observed across all runs.

Increasing the number of IVs used for error detection led to an increase in specificity but not sensitivity (Table S IX). Specificity increased from 97.1% to 99.0% as the number of IVs increased from 10 to 1000 in run #1, corresponding to an increase of PPV from 83.2% to 93.5%. In contrast, sensitivity did not have a consistent trend when the number of IVs used increased from 10 to 1000. Across runs, the magnitude of improvement varied. For instance, as the number of IVs used increased from 10 to 1000, the PPV only increased minimally from 87.7% to 89.4% in run #3, but PPV markedly increased from 87.4% to 98.7% in run #4.

Results from our simulation suggest that different descent patterns of chromosomes can have differential ability for error detection using this class of approach. Sensitivity and specificity both varied across runs. For instance, the sensitivity with the true IVs and the default conditions ranged from 84.7–93.3%. In each condition, because at least 2500 markers contained genotyping errors (Table S IV), the variation between runs was unlikely to be artifacts of small sample sizes. Thus, the most likely explanation is that the method has higher ability to detect errors in some underlying descent patterns than others.

Other Parameters

Typing more subjects for dense markers improved sensitivity but decreased specificity (Table I). Sensitivity substantially increased when the number of subjects typed for dense markers increased, e.g. 66.6% for the D_fF_a design vs. 89.8% for the D_mF_a design. Nevertheless, our method continued to detect MC genotyping errors almost as well as what could be detected with the true IV. These observations were consistent across runs (Table S VII). In run #1, as well as in most other runs, sensitivity did not increase when the number of subjects typed for dense markers increased from many to all (89.8% for D_mF_a design vs. 88.7% for D_aF_a design). However, in run #10, sensitivity continued to increase when the number of subjects typed for dense markers increased from many to all subjects (88.9% for D_mF_a design vs. 95.7% for D_aF_a design). In run #1, unlike sensitivity, specificity (99.7% for D_fF_a , 99.2% for D_mF_a , and 98.9% for D_aF_a) and PPV (95.8% for D_fF_a , 94.4% for D_mF_a and 91.9% for D_aF_a) decreased as the number of subjects typed for dense markers increased (Table I). The decrease in specificity reflects an increase in false detections. An increase in false detections likely occurs because typing more subjects for dense markers imposes a tighter constraint between the inferred IVs and observed genotypes, and more clean dense markers have low numbers of consistent IVs.

In contrast, typing more subjects for framework markers did not greatly improve performance. Sensitivity, specificity, and PPV only increased slightly when the numbers of subjects typed for framework markers increased (Table I). For instance, increasing the number of subjects typed for framework markers only slightly increased sensitivity, specificity, and PPV (D_fF_f design vs. D_fF_a design: sensitivity = 65% vs. 66.6%, specificity = 99.5% vs. 99.7%, and PPV: 94% vs. 95.8%) (Table I). Even though the increases were not substantial, similar trends were observed across all runs (Table S VII). For example, the

overall difference in sensitivity between the use of true vs. estimated IVs decreased from 3.0 in the D_fF_f design to 1.4 in the D_fF_a design.

Even though the number of subjects typed for framework markers did not substantially affect the quality of error detection, the density of framework markers substantially affected specificity and PPV (Table S X). Specificity decreased moderately as the density of the framework panel increased from 0.5 cM to 1 cM (99% to 96.1%) and decreased substantially at 2 cM (82.8%). When specificity only dropped by 2.9% when the density decreased from 0.5 cM to 1 cM, PPV decreased substantially from 93.5% to 78.2%. At the 2 cM density, PPV was 45%. A histogram of the percentage of IVs consistent with the observed dense genotypes suggested that a higher percentage of markers had very low consistency as the framework density decreased (Figure S5). In some other runs, such as run #6, the PPV did not decrease when the density decreased from 0.5 cM to 1 cM (83.1% to 83.6%, Table S X). Nevertheless, PPV still substantially decreased as the density decreased from 1 cM to 2 cM.

Discussion

We highlight a few results. First, the results show that our error detection approach can achieve high sensitivity, specificity, and PPV for use with very dense markers on pedigrees. Second, when a true genotyping error is detected, the approach can often identify the subject who has the error. Third, the analyses suggest that the use of inferred IVs is almost as effective as the use of the true IV to identify errors, so our method works almost as well as what could optimally be accomplished under this general approach. Fourth, the computationally simpler test statistic A1 has almost identical sensitivity as test statistic A2. The incorporation of an error model and allele frequencies into A2 does not improve the ability to globally detect genotyping errors, which suggests that the constraint from sampled IVs largely determines the ability of this approach to detect errors.

Our error detection approach also effectively detects MC genotyping errors in markers with low MAF. The results from simulations show that sensitivity stays high over the entire range of MAFs. In addition, both specificity and PPV increase as the minor allele frequency of the marker decreases. This trend does not seem to be an artifact of small sample size, as different classes of MAFs are well represented. A simple explanation for this trend is that clean markers with lower MAF contain a higher percentage of homozygous alleles for the major allele in the observed genotypes, so the observed genotypes are more likely to be consistent with many sampled IVs.

In general, we expect the method to be effective at detecting a genotyping error if two or more copies of the same founder allele are observed and this allele can be unambiguously assigned to the IV. On the other hand, this method will not detect a genotyping error if only one copy of such the allele is observed or this allele cannot be assigned unambiguously to the IVs. Thus, in dense SNPs or sequence data, this class of method will be more effective when it is expected that duplicated copies of founder alleles are obtained, as would occur when multiple relatives sharing alleles are genotyped, and will not be as effective when only limited copies of founder alleles are obtained. The source of errors also may have an effect

on the effectiveness of error detection. Errors that are caused by poor study design and that lead to errors from batch effects could lead either to increased detection of such errors through excess allelic conflicts during detection, or could lead to reduced identification of errors if fewer duplicate copies of alleles are therefore observed.

The quality of sampled IVs affects the ability to detect MC genotyping errors. Since the use of the theoretical true IV always yields perfect specificity, the use of IVs that are closer to the truth should increase specificity of error detection. An important consideration is that our detection approach requires the use of a moderate number of framework markers to infer IVs to increase specificity to an acceptable level. This agrees with a general earlier remark that multipoint error detection methods are effective when the marker density is relatively high [Chang, et al. 2006]. The evaluation suggests that our approach achieves high sensitivity and specificity for detecting MC genotyping errors in diallelic markers when the density of framework markers is a marker per 1 cM. More precisely, if framework markers are not informative about the descent pattern at the dense markers, a high fraction of IVs will likely be inconsistent even with clean observed dense genotypes. Under such a scenario, the specificity of error detection would be lower because more clean markers would also be flagged as containing errors. Contrary to imputation, whether additional subjects are typed for framework markers is usually less important in error detection than in genotype imputation. This is understandable because error detection only focuses on detecting errors in the subjects observed for dense genotypes. Regardless, informative markers should still be used, and the use of other informative framework markers such as multi-allelic markers may improve the inference of IVs. As described in more detail elsewhere [Cheung, et al. 2013], a number of criteria are important for selecting framework markers. These include selection of well-typed markers that are informative for linkage analysis but that are not in strong linkage disequilibrium, as violation of the assumption may lead to biased sampling of IVs [Huang, et al. 2004; Schaid, et al. 2002].

The percentage of clean markers flagged as markers with genotyping errors depends on the threshold, and is reflected by 100%-specificity. If an error is flagged, the investigator may decide whether to drop the marker or a particular individual identified using method A2, or alternatively, to re-genotype in order to correct the error. This decision may depend on cost as well as the specific intended method of downstream analysis. For instance, in the case of linkage analysis, it may be advantageous to apply more caution to select a sub-panel of markers only from the markers without evidence of genotyping errors because it is usually unnecessary to use all markers from a dense array to carry such analysis.

Test statistics A1 and A2 offer different advantages. First, computing A1 is considerably faster than A2 because A1 does not involve calculating multiple configurations of x_v . Second, the calculation of A1 does not require the use of allele frequencies. Even though A1 contains the term $P(G_v | S_v^k)$, evaluating whether this quantity is greater than 0 does not depend on the allele frequencies. The fact that A1 is invariant to the allele frequencies is an advantage because allele frequencies may not be easily available especially when the dense markers represent sequence data. Third, an advantage of A2 is that it can be used to identify which subject most likely contains a genotyping error.

Computation can proceed with high efficiency. On a computer with one Intel L5420 Xeon 2.50GHz CPU with 8GB of memory, GIGI-check with both Method A1 and A2 takes 18 minutes to detect errors with 25,000 markers on the 52 member pedigree with 100 sampled IVs (and only 1 minute 47 seconds with 10 sampled IVs, which as the results suggest, is often enough for initial scan of errors). Computation scaled linearly with the number of markers and number of sampled IVs. Thus, on an Illumina core array with ~250,000 SNPs, a run with similar condition can be completed in ~3 hours with 100 IVs and 18 minutes with 10 IVs. Further computation efficiency may also be gained with future improvements. Our results from simulations suggest that the use of a smaller set of IVs decreases specificity but does not substantially decrease sensitivity. Because of these properties, one may first run the error detection program on a small number of IVs (e.g. 50 IVs or less) as an initial scan with a more lenient threshold to screen out most markers that are free of errors. Then, on a much reduced set of markers that are enriched for potential errors, one can re-run the error detection program using a full set of IVs. In addition, because computing A1 is much faster than computing A2, and because the two scores give comparable performance, A1 should be used in the initial scan with the number of sampled IVs that one can afford in terms of computational time. Then, on a subsequent scan, A2 can be used to additionally identify the subjects that most likely have the genotyping errors. Thus, along with parallel computation using multiple CPUs that is often available, GIGI-check would be able to process millions of markers in a realistic amount of time. If we are not confident that our framework markers are free of errors, it is beneficial to run the error detection procedure using a few sets of framework markers, because framework markers that contain genotyping errors may lead to biased sampling of IVs [Lebrec, et al. 2008; Markus, et al. 2011]. To reduce the number of false error calls, it may be useful to use a few sets of framework markers to detect genotyping errors.

It is important to minimize false findings in genetic studies. Because of the time, money, and effort invested in these studies, one should not neglect proper checks to improve the quality of genotype data before carrying out major analyses, such as linkage or family-based association studies with dense variants. Our practical approach described here coupled with implementation in the program GIGI-Pick enables efficient performance of this crucial quality check of dense genotypes in small to large pedigrees. Once errors have been identified then depending on the goals of subsequent data use, re-genotyping to correct errors, removal of potential errors, or avoidance of use of markers with errors are all options. Our approach therefore provides a line of defense against false findings driven by genotyping error in studies that incorporate dense genotyping, as may be obtained by sequencing, into family studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Supported by funding from the National Institutes of Health, grants R37GM046255, R01MH094293, P50AG005136, R01MH092367 and R01AG039700.

References

- Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*. 2001; 9(2):130–134. [PubMed: 11313746]
- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*. 2002; 30(1):97–101. [PubMed: 11731797]
- Akey JM, Zhang K, Xiong MM, Doris P, Jin L. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *American Journal of Human Genetics*. 2001; 68(6):1447–1456. [PubMed: 11359212]
- Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM (R)). *Human Mutation*. 2011; 32(5):564–567. [PubMed: 21472891]
- Becker T, Valentonyte R, Croucher PJP, Strauch K, Schreiber S, Hampe J, Knapp M. Identification of probable genotyping errors by consideration of haplotypes. *European Journal of Human Genetics*. 2006; 14(4):450–458. [PubMed: 16435001]
- Buetow KH. Influence of aberrant observations on high-resolution linkage analysis outcomes. *American Journal of Human Genetics*. 1991; 49(5):985–994. [PubMed: 1928104]
- Chang YPC, Kim JDO, Schwander K, Rao DC, Miller MB, Weder AB, Cooper RS, Schork NJ, Province MA, Morrison AC. and others The impact of data quality on the identification of complex disease genes: experience from the Family Blood Pressure Program. *European Journal of Human Genetics*. 2006; 14(4):469–477. [PubMed: 16493446]
- Cherny SS, Abecasis GR, Cookson WOC, Sham PC, Cardon LR. The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. *Genetic Epidemiology*. 2001; 21:S117–S122. [PubMed: 11793653]
- Cheung CYK, Thompson EA, Wijsman EM. GIGI: An Approach to Effective Imputation of Dense Genotypes on Large Pedigrees. *American Journal of Human Genetics*. 2013; 92(4):504–516. [PubMed: 23561844]
- Douglas JA, Boehnke M, Lange K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *American Journal of Human Genetics*. 2000; 66(4):1287–1297. [PubMed: 10739757]
- Douglas JA, Skol AD, Boehnke M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics*. 2002; 70(2):487–495. [PubMed: 11791214]
- Goldstein DR, Zhao HY, Speed TP. The effects of genotyping errors and interference on estimation of genetic distance. *Human Heredity*. 1997; 47(2):86–100. [PubMed: 9097090]
- Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Human Heredity*. 2002; 54(1):22–33. [PubMed: 12446984]
- Gordon D, Heath SC, Liu X, Ott J. A transmission disequilibrium test that allows for genotyping errors in the analysis of single nucleotide polymorphism data. *American Journal of Human Genetics*. 2001; 69(2):371–380. [PubMed: 11443542]
- Gordon D, Heath SC, Ott J. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Human Heredity*. 1999; 49(2):65–70. [PubMed: 10077724]
- Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity*. 2003; 90(1):33–38. [PubMed: 12522423]
- Hinrichs AL, Suarez BK. Genotyping errors, pedigree errors, and missing data. *Genetic Epidemiology*. 2005; 29:S120–S124. [PubMed: 16342180]
- Huang QQ, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *American Journal of Human Genetics*. 2004; 75(6):1106–1112. [PubMed: 15492927]
- Kennedy J, Mandoiu I, Pasaniuc B. Genotype error detection using Hidden Markov Models of haplotype diversity. *Journal of Computational Biology*. 2008; 15(9):1155–1171. [PubMed: 18973433]

- Kirk KM, Cardon LR. The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*. 2002; 10(10):616–622. [PubMed: 12357332]
- Knapp M, Becker T. Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *American Journal of Human Genetics*. 2004; 74(3): 589–591. [PubMed: 14973788]
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*. 1996; 58(6):1347–1363. [PubMed: 8651312]
- Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*. 1987; 84(8):2363–2367. [PubMed: 3470801]
- Lathrop GM, Hooper AB, Huntsman JW, Ward RH. Evaluating pedigree data.1. The estimation of pedigree error in the presence of marker mistyping. *American Journal of Human Genetics*. 1983; 35(2):241–262. [PubMed: 6573130]
- Lebrech JJP, Putter H, Houwing-Duistermaat JJ, Van Houwelingen HC. Influence of genotyping error in linkage mapping for complex traits - an analytic study. *BMC Genetics*. 2008;9. [PubMed: 18221501]
- Lincoln SE, Lander ES. Systematic detection of errors in genetic-linkage data. *Genomics*. 1992; 14(3): 604–610. [PubMed: 1427888]
- Marchani EE, Chapman NH, Cheung CYK, Ankenman K, Stanaway IB, Coon HH, Nickerson D, Bernier R, Brkanac Z, Wijsman EM. Identification of Rare Variants from Exome Sequence in a Large Pedigree with Autism. *Human Heredity*. 2012; 74(3–4):153–164. [PubMed: 23594493]
- Markus B, Birk OS, Geiger D. Integration of SNP genotyping confidence scores in IBD inference. *Bioinformatics*. 2011; 27(20):2880–2887. [PubMed: 21862568]
- Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics*. 2003; 72(3):598–610. [PubMed: 12587097]
- Mukhopadhyay N, Buxbaum SG, Weeks DE. Comparative study of multipoint methods for genotype error detection. *Human Heredity*. 2004; 58(3–4):175–189. [PubMed: 15812175]
- O'Connell JR, Weeks DE. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*. 1998; 63(1):259–266. [PubMed: 9634505]
- Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *American Journal of Human Genetics*. 2002; 71(4):992–995. [PubMed: 12387273]
- Sieberts SK, Wijsman EM, Thompson EA. Relationship inference from trios of individuals, in the presence of typing error. *American Journal of Human Genetics*. 2002; 70(1):170–180. [PubMed: 11727198]
- Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*. 1996; 58(6):1323–1337. [PubMed: 8651310]
- Sobel E, Papp JC, Lange K. Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*. 2002; 70(2):496–508. [PubMed: 11791215]
- Stringham HM, Boehnke M. Identifying marker typing incompatibilities in linkage analysis. *American Journal of Human Genetics*. 1996; 59(4):946–950. [PubMed: 8808612]
- Thompson E. The structure of genetic linkage data: From LIPED to 1M SNPs. *Human Heredity*. 2011; 71(2):86–96. [PubMed: 21734399]
- Walters K. The effect of genotyping error in sib-pair genomewide linkage scans depends crucially upon the method of analysis. *Journal of Human Genetics*. 2005; 50(7):329–337. [PubMed: 16059746]
- Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB. A tale of two genotypes: Consistency between two high-throughput genotyping centers. *Genome Research*. 2002; 12(3):430–435. [PubMed: 11875031]
- Wijsman EM. The role of large pedigrees in an era of high-throughput sequencing. *Human Genetics*. 2012; 131(10):1555–1563. [PubMed: 22714655]

- Wijman EM, Rothstein JH, Thompson EA. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics*. 2006; 79(5):846–858. [PubMed: 17033961]
- Zhi DG, Wu JH, Liu NJ, Zhang K. Genotype calling from next-generation sequencing data using haplotype information of reads. *Bioinformatics*. 2012; 28(7):938–946. [PubMed: 22285565]
- Zou GH, Pan DY, Zhao HG. Genotyping error detection through tightly linked markers. *Genetics*. 2003; 164(3):1161–1173. [PubMed: 12871922]

Figure 1. An example illustrating that MC genotyping error can be detected by IVs

A) In this pedigree, subjects are labeled by lower-case roman numerals, distinct founder chromosomes are labeled by numbers, and observed genotypes are labeled by letters. B) Incompatibility between the observed genotypes and the identity-by-descent graph (IBD-graph) occurs when subjects with observed genotypes are added in any order to the IBD-graph. The IBD-graph is the graph of founder chromosomes connected by edges indicating the pairs of founder chromosomes that observed individuals share. The order of subjects added on the left is: iii, iv, then ii, leading to an inconsistency between the observed genotype (Aa) and that predicted by the IBD-graph (AA) for subject ii. The order of subjects added on the right is: iii, ii, and then iv, leading to an inconsistency between observed (AA) and predicted (Aa) genotype for subject iv.

Figure 2. Receiver Operating Characteristic curves comparing the performance of error detection using test statistics A1 and A2 as a function of detection thresholds in run #1
 Different detection thresholds are indicated by points at an increment of 0.01 from the left-most threshold of 0. The vertical dash line indicates the default detection threshold. The horizontal line represents the maximum sensitivity that can be obtained with the true IV.

Table 1

Effect of the number of subjects typed for dense and framework markers in run#1

	Design <i>I</i>					
	Default	Completion rate of dense and framework markers				
	D _m F _m	D _a F _a	D _m F _a	D _f F _a	D _f F _m	D _f F _f
Sensitivity	88.9	88.7	89.8	66.6	65.4	65.0
Specificity	99.0	98.9	99.2	99.7	99.6	99.5
PPV	93.5	91.9	94.4	95.8	95.3	94.0
Sensitivity with true IV	90.8	89.7	90.8	68.0	68.0	68.0

I, Design with subjects typed for dense (D) and framework (F) markers on all (a), many (m), or few (f) subjects