Distinct error rates for reference and non-reference genotypes estimated by pedigree analysis

Richard J. Wang^{*}, Predrag Radivojac[†] and Matthew W. Hahn^{*,‡}

^{*} Department of Biology, Indiana University, Bloomington, Indiana 47405

[†] Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, 02115

[‡] Department of Computer Science, Indiana University, Bloomington, Indiana 47405

Running title: Estimating genotyping error rates

Keywords: genotyping error rate, haplotype phase, pedigree analysis, whole-genome sequencing

Corresponding author: Richard J. Wang
1001 E. Third St.
Bloomington, IN 47405
Phone: (812) 856-7016
E-mail: rjwang@indiana.edu

Abstract

Errors in genotype calling can have perverse effects on genetic analyses, confounding association studies and obscuring rare variants. Analyses now routinely incorporate error rates to control for spurious findings. However, reliable estimates of the error rate can be difficult to obtain because of their variance between studies. Most studies also report only a single estimate of the error rate even though genotypes can be miscalled in more than one way. Here, we report a method for estimating the rates at which different types of genotyping errors occur at biallelic loci using pedigree information. Our method identifies potential genotyping errors by exploiting instances where the haplotypic phase has not been faithfully transmitted. The expected frequency of inconsistent phase depends on the combination of genotypes in a pedigree and the probability of miscalling each genotype. We develop a model that uses the differences in these frequencies to estimate rates for different types of genotype error. Simulations show that our method accurately estimates these error rates in a variety of scenarios. We apply this method to a dataset from the whole-genome sequencing of owl monkeys (*Aotus nancymae*) in three-generation pedigrees. We find significant differences between estimates for different types of genotyping error, with the most common being homozygous reference sites miscalled as heterozygous and vice versa. The approach we describe is applicable to any set of genotypes where haplotypic phase can reliably be called, and should prove useful in helping to control for false discoveries.

Introduction

When dealing with large sets of genotype data, such as those obtained from whole-genome sequencing, error rates that are low per-site can still yield millions of miscalled genotypes. Errors can be introduced anywhere along the long process, from sampling to genotyping. The frequency of error often depends on the sequencing technology employed. With next-generation sequencing (NGS), even reads that are perfectly mapped and free of base-calling errors can lead to miscalled diploid genotypes due to the random over-sampling of one allele during amplification. Genotyping errors can profoundly hinder genetic analyses, for instance by reducing power in linkage and association studies (Abecasis *et al.* 2001; Gordon *et al.* 2002; Ahn *et al.* 2007). Studies interested in identifying rare variants are especially sensitive to these errors, in the context of either disease (Powers *et al.* 2011; Yan *et al.* 2016) or the rate of *de novo* mutation (Ségurel *et al.* 2014; Carlson *et al.* 2018). The consequences of genotyping errors for biological conclusions can often be mitigated by explicitly including the possibility of errors in genetic analyses (Sobel *et al.* 2002; Cartwright *et al.* 2007; Lebrech *et al.* 2008).

Methods for attenuating the effects of error often require investigators to use an estimate of error rates (Pompanon *et al.* 2005). While error rates can broadly be classified by the type of sequencing or genotyping technology employed, each experiment will have a different error rate. Identical pipelines for generating genotype data with NGS can lead to rates of error that vary between different samples and cohorts (Dohm *et al.* 2008; Huang *et al.* 2009). For example, the population frequency of each variant in a cohort can be used to improve the accuracy of genotype calls, as in the GenotypeGVCFs workflow (Poplin *et al.* 2017). While this approach increases the confidence in each genotype call, genotype error rates can become dependent on the rarity of variants at a locus.

There are several different approaches for estimating genotyping error rates in a given experiment. The most straightforward is some form of replication, where sequencing or genotyping on one or more samples is repeated—or performed at higher read-coverage with NGS—and compared to the original results (e.g. Wall *et al.* 2014; Pfeiffer *et al.* 2018; Ma *et al.* 2019). Aside from the potential to be cost-prohibitive, this approach generates a rate of discordance between replicates rather than a true estimate of the error rate. Robust approaches to identifying genotyping errors and estimating their rates typically leverage pedigree information. These approaches identify errors by finding discordance between the observed genotypes and those expected from the laws of Mendelian inheritance (Douglas *et al.* 2002; Hao *et al.* 2004; Saunders *et al.* 2007; The 1000 Genomes Project Consortium 2015).

In this study, we develop a method to estimate the rates at which different types of genotyping errors occur. We focus on errors at biallelic sites and distinguish between errors that involve reference versus non-reference alleles. Our method is useful for establishing a distinct genotyping error rate for genetic analyses of rare variants, which are typically non-reference. We apply pedigree information to track the transmission of haplotypes and to identify errors by looking for sites that have genotypes that cannot be the result of a faithful transmission event. Loci that do not possess the expected phase within a haplotype block must be the result of either: a *de novo* mutation, a gene conversion event, or a genotyping error. Genotyping errors are by far the most common of these phenomena (Table 1). Detection of unlikely genotypes from haplotype phase has previously been used to successfully identify genotyping errors (Abecasis *et al.* 2002; Becker *et al.* 2006; O’Connell *et al.* 2014; Kothiyal *et al.* 2019). What we add here is an error model to explicitly determine the genotyping error rate, and the ability to distinguish rates for different types of errors.

We apply our new method to a dataset of genotypes collected from the whole-genome sequencing of a set of owl monkey (*Aotus nancymae*) pedigrees (Thomas *et al.* 2018). Among sites that could be phased with pedigree information, we found a significant difference in the direction in which phase errors occurred. This departure forms the signal for estimating the rate of genotyping error. Estimated error rates were significantly different among the genotypes, with the most common error being a homozygous reference site miscalled as heterozygous. The principles of our method can be applied to determine the rate of different types of genotyping error in any dataset where phase errors can be identified.

Table 1. Approximate per-site rates for phenomena leading to phase inconsistency

NGS genotyping error	10^{-3}
gene conversion	10^{-7}
recombination	10^{-8}
<i>de novo</i> mutation	10^{-8}

References

NGS error: Nielsen *et al.* (2011), Wall *et al.* (2014); gene conversion: Jeffreys and May (2004), Halldorsson *et al.* (2016); recombination: Jensen-Seaman *et al.* (2004), Coop and Przeworski (2007); *de novo* mutation: Kong *et al.* (2012), Venn *et al.* (2014).

Materials and Methods

We develop a method to estimate genotyping error rates from whole-genome data by examining autosomal sites that can be unambiguously phased from a three-generation pedigree. At phase-informative sites, the transmission of haplotypes can be determined according to the Mendelian laws of inheritance. When the genotype of the child does not match the haplotypic combination transmitted by the parents, a genotyping error is the most likely cause. The frequency of such phase violations depends on the rate of genotyping error and the relevant site frequency in the sample. The expected frequencies of different phase violations can thus be compared to the observed frequencies to estimate error rates. We focused on building an error model for three-generation pedigrees along a single line of descent (Fig. 1a), though genotypes can be phased in two-generation pedigrees when there are more than two siblings (Fledel-Alon *et al.* 2009; Roach *et al.* 2011). We also restricted ourselves to biallelic sites, the most common form of variation across the genome. The six possible miscalls at a biallelic site are labeled in Table 2.

Figure 1. Pedigree and phase in a three-generation pedigree along a single line of descent

(a) The pedigree of five individuals and a diagram of the haplotype traced along three generations. Sites transmitted to the child can be out of phase within a haplotype block.

(b) Genotypes at an example phase-informative site. The phase of the child genotype can be traced to either Parent A or Parent B.

Table 2. Genotyping errors at a biallelic site

Error	Truth	Observed
$\epsilon_{0>1}$	0/0	0/1
$\epsilon_{1>0}$	0/1	0/0
$\epsilon_{2>0}$	1/1	0/0
$\epsilon_{0>2}$	0/0	1/1
$\epsilon_{1>2}$	0/1	1/1
$\epsilon_{2>1}$	1/1	0/1

Reference allele represented by 0, non-reference represented by 1.

Expected number of phase violations

At phase-informative sites, the focal individual in a three-generation pedigree produces gametes with traceable phase. Within a block without crossovers, the haplotype inherited from the focal individual can be identified as being derived from either the maternally or paternally inherited chromosome of the focal individual. With a three-generation pedigree (Fig 1a), phase-informative sites are those where: the focal individual is heterozygous, the parents of the focal individual are not both heterozygous, and the breeding partner of the focal individual and the child are not both heterozygous.

A genotyping error in any of the five individuals from the pedigree can cause an informative site to appear out-of-phase with its neighbors. For example, for a set of genotypes as in Figure 1b, a child called as heterozygous at a site in a parent *A* haplotype block could be an $\epsilon_{0>1}$ error, a miscalled homozygous child. Similarly, a homozygous child at a site in a parent *B* block could be an $\epsilon_{1>0}$ error. A miscalled child genotype is the most straightforward cause of a phase violation, but errors in other individuals also create apparent phase violations. Using again

the genotypic combination in Figure 1b—if parent A is miscalled as homozygous alternate instead of homozygous reference (an $\varepsilon_{2>0}$ error; Table 2) a site in a parent A block would be expected to carry the alternate allele. When parent A is miscalled in this way, the child will be heterozygous for a site in a parent A block and would appear to be a phase violation.

The expected number of violations is the sum of potential phase violations from miscalls in all individuals across the pedigree. We estimate the number of violations for each miscall as a product of the respective genotypic combination frequencies across the genome and the corresponding error rate. We demonstrate this rationale below for the genotypic combination matching the example in Figure 1b, calculating the expected number of such phase-violating sites from phase A haplotype blocks. In this calculation, we assumed that out-of-phase informative sites were the result of no more than one genotyping error among the sampled individuals for a given site.

Genotypic combinations are abbreviated as G_x , where x is composed of individual genotypes represented by a single digit: 0, 1, and 2 for homozygous reference, heterozygous, and homozygous alternate respectively. These are ordered in the subscript as: parent A , parent B , focal individual, partner of focal, and child. Let n_x be the number of such genotypic combinations and $\varepsilon_{i>j}$ be the genotyping error rate as in Table 2. The genotypic combinations and possible miscalls leading to a phase violation for a site like the one depicted in Figure 1b are then:

Expected: G_{01100} Observed: G_{01101}

Parent A Observed genotype: 0/0

Possible miscall: 1/1

Focal individual inherits alternate allele from Parent A and transmits it to the child leading to apparent phase violation.

156 Possible miscall: 0/1
 157 Focal individual inherits the alternate allele from Parent A, giving a 50% chance
 158 of transmission to child, leading to apparent phase violation.
 159 Frequency: $\frac{1}{2}n_{11101} \cdot \varepsilon_{1>0} + n_{21101} \cdot \varepsilon_{2>0}$

160 *Parent B* Observed genotype: 0/1
 161 Possible miscall: 0/0
 162 Implies more than one genotyping error across pedigree.
 163 Possible miscall: 1/1
 164 Not detectable as phase violation, as focal individual still inherits alternate allele
 165 from Parent B haplotype block.

166 *Focal* Observed genotype: 0/1
 167 Any miscall would imply more than one genotyping error.

168 *Partner* Observed genotype: 0/0
 169 Possible miscall: 1/1
 170 Child inherits alternate allele from the partner, which appears as a phase
 171 violation.
 172 Possible miscall: 0/1
 173 Child inherits alternate allele from the partner, giving a 50% chance of
 174 transmission to child, leading to apparent phase violation.
 175 Frequency: $\frac{1}{2}n_{01111} \cdot \varepsilon_{1>0} + n_{01121} \cdot \varepsilon_{2>0}$

176 *Child* Observed genotype: 0/1
 177 Possible miscall: 0/0
 178 True genotype 0 has been miscalled as 1. Each occurrence leads to this phase
 179 violation.
 180 Frequency: $n_{01100} \cdot \varepsilon_{0>1}$

181 The expected number of phase violations for this genotypic combination can be written as

$$E[|A_{0110}|] = n_{01100} \cdot \varepsilon_{0>1} + \frac{1}{2}(n_{11101} + n_{01111}) \cdot \varepsilon_{1>0} + (n_{21101} + n_{01121}) \cdot \varepsilon_{2>0} \quad (1)$$

182 where A_y is a phase violation in the parent A haplotype block, with genotypic combination y and
 183 $|A_y|$ is the number of such violations across the genome. Here, y is the abbreviated phase-
 184 informative genotypic combination, listing abbreviated genotypes for the first four individuals as
 185 ordered in x for G_x .

In pedigrees with five individuals there are 18 genotypic combinations that are phase-informative. Table 3 lists each of these combinations and the child genotypes that indicate a phase violation. Violations at sites with the haplotype from Parent *A* and Parent *B* are labeled as A_y and B_y respectively. Each genotypic combination can be evaluated to determine the total number of expected phase violations. Several symmetries between genotypic combinations reduce the number of unique calculations. For example, each violation has a matching pair where the phase and the genotypes of Parent *A* and Parent *B* are swapped. The frequency of each of these two classes is expected to be equal across the genome due to independent assortment. The matching violation to A_{0110} in the example above is B_{1010} , and the expected frequency can be calculated by swapping the two parental genotypes in each term as

$$E[|B_{1010}|] = n_{10100} \cdot \varepsilon_{0>1} + \frac{1}{2}(n_{111101} + n_{011111}) \cdot \varepsilon_{1>0} + (n_{12101} + n_{10121}) \cdot \varepsilon_{2>0} \quad (2)$$

Similarly, each violation can be paired with a case in which homozygous reference and homozygous alternate genotypes are swapped. The expected frequency of violations has the same form once these genotypes and error rates are swapped. For A_{0110} , the matching violation is A_{2112} with expected frequency

$$E[|A_{2112}|] = n_{21122} \cdot \varepsilon_{2>1} + \frac{1}{2}(n_{111121} + n_{211111}) \cdot \varepsilon_{1>2} + (n_{01121} + n_{21101}) \cdot \varepsilon_{0>2} \quad (3)$$

Derivations of the expected frequencies for the remaining set of phase violations can be found in the Supplemental Material (Appendix S1).

Table 3. Detectable phase violations in a three-generation pedigree

Parent A	Parent B	Focal	Partner	Phase Violation	Child (Expected)	Child (Observed)	Phase Violation	Child (Expected)	Child (Observed)
0/0	0/1	0/1	0/0	A_{0110}	0/0	0/1	B_{0110}	0/1	0/0
0/0	0/1	0/1	0/1	A_{0111}	0/0; 0/1	1/1	B_{0111}	1/1; 0/1	0/0
0/0	0/1	0/1	1/1	A_{0112}	0/1	1/1	B_{0112}	1/1	0/1
1/1	0/1	0/1	0/0	A_{2110}	0/1	0/0	B_{2110}	0/0	0/1
1/1	0/1	0/1	0/1	A_{2111}	1/1; 0/1	0/0	B_{2111}	0/0; 0/1	1/1
1/1	0/1	0/1	1/1	A_{2112}	1/1	0/1	B_{2112}	0/1	1/1
0/0	1/1	0/1	0/0	A_{0210}	0/0	0/1	B_{0210}	0/1	0/0
0/0	1/1	0/1	0/1	A_{0211}	0/0; 0/1	1/1	B_{0211}	1/1; 0/1	0/0
0/0	1/1	0/1	1/1	A_{0212}	0/1	1/1	B_{0212}	1/1	0/1
1/1	0/0	0/1	1/1	A_{2012}	1/1	0/1	B_{2012}	0/1	1/1
1/1	0/0	0/1	0/1	A_{2011}	1/1; 0/1	0/0	B_{2011}	0/0; 0/1	1/1
1/1	0/0	0/1	0/0	A_{2010}	0/1	0/0	B_{2010}	0/0	0/1
0/1	1/1	0/1	1/1	A_{1212}	0/1	1/1	B_{1212}	1/1	0/1
0/1	1/1	0/1	0/1	A_{1211}	0/0; 0/1	1/1	B_{1211}	1/1; 0/1	0/0
0/1	1/1	0/1	0/0	A_{1210}	0/0	0/1	B_{1210}	0/1	0/0
0/1	0/0	0/1	1/1	A_{1012}	1/1	0/1	B_{1012}	0/1	1/1
0/1	0/0	0/1	0/1	A_{1011}	1/1; 0/1	0/0	B_{1011}	0/0; 0/1	1/1
0/1	0/0	0/1	0/0	A_{1010}	0/1	0/0	B_{1010}	0/0	0/1

Reference allele represented by 0, non-reference allele represented by 1.

Violations are mirrored in the second half of the table (Parent A and Parent B genotypes swapped, A and B violation swapped).

Estimating the error rates

Taken together, the expected number of phase violations from the genotypic combinations in Table 3 form a linear system of equations that can be represented by a matrix equation. Let $\mathbf{V}_A$ be a vector for the number of observed sites that violate haplotype block A across the genome, as ordered by row in Table 3; that is, $\mathbf{V}_A = \langle |A_{0110}|, |A_{0111}|, |A_{0112}|, \dots, |A_{1011}|, |A_{1010}| \rangle$. The number of observed sites that violate block B is represented by $\mathbf{V}_B$, again as ordered in Table 3. Let $\hat{\mathbf{e}}$ be a vector of estimators for each type of error rate, $\hat{\mathbf{e}} = \langle \hat{e}_{0>1}, \hat{e}_{1>0}, \hat{e}_{2>0}, \hat{e}_{0>2}, \hat{e}_{1>2}, \hat{e}_{2>1} \rangle$. The error model can then be represented as

$$\mathbf{M}\hat{\mathbf{e}} = \begin{bmatrix} \mathbf{V}_A \\ \mathbf{V}_B \end{bmatrix} \quad (4)$$

where the matrix $\mathbf{M}$ contains the coefficients of the linear equations from the expected frequencies of genotyping errors for each violation class. We can divide $\mathbf{M}$ into two submatrices based on the coefficients for violations in phase A and phase B as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A \\ \mathbf{M}_B \end{bmatrix} \quad (5)$$

Rows in $\mathbf{M}_A$ and $\mathbf{M}_B$ are identical for violations in phase A and phase B that share the same expected frequencies (e.g. A_{0110} and B_{1010} , described in the previous section). The violations in Table 3 are ordered so that $\mathbf{M}_A$ and $\mathbf{M}_B$ can be related by an exchange matrix as

$$\mathbf{M}_B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & \ddots & 0 \\ 1 & 0 & 0 \end{bmatrix}_{n=18} \mathbf{M}_A \quad (6)$$

Coefficients from equations for each of the respective expected phase A violations (see Supplemental Information) form the $\mathbf{M}_A$ matrix:

222 $\mathbf{M}_A =$

$$\begin{array}{c|cccccc}
 n_{01100} & \frac{1}{2}(n_{11101} + n_{01111}) & n_{21101} + n_{01121} & 0 & 0 & 0 \\
 0 & \frac{1}{2}n_{11112} & n_{21112} & n_{01110} & \frac{1}{2}n_{01111} & 0 \\
 0 & \frac{1}{2}n_{11122} & n_{21122} & 0 & n_{01121} & 0 \\
 0 & n_{21101} & 0 & n_{01100} & \frac{1}{2}n_{11100} & 0 \\
 0 & \frac{1}{2}n_{21111} & n_{21112} & n_{01110} & \frac{1}{2}n_{11110} & 0 \\
 0 & 0 & 0 & n_{0112} + n_{21101} & \frac{1}{2}(n_{21111} + n_{11121}) & n_{21122} \\
 n_{02100} & \frac{1}{2}n_{02111} & n_{02121} & 0 & 0 & 0 \\
 0 & 0 & 0 & n_{02110} & \frac{1}{2}n_{02111} & 0 \\
 0 & 0 & 0 & 0 & n_{02121} & 0 \\
 0 & 0 & 0 & n_{20101} & \frac{1}{2}n_{20111} & n_{20122} \\
 0 & \frac{1}{2}n_{20111} & n_{20112} & 0 & 0 & 0 \\
 0 & n_{20101} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & n_{10122} & n_{12121} + \frac{1}{2}n_{11122} & \frac{1}{2}n_{12222} \\
 0 & 0 & 0 & n_{12110} + n_{10112} & \frac{1}{2}(n_{12111} + n_{11112}) & \frac{1}{2}n_{12212} \\
 n_{12100} & \frac{1}{2}n_{12111} & n_{12121} & n_{10101} & \frac{1}{2}n_{11101} & \frac{1}{2}n_{12201} \\
 \frac{1}{2}n_{10021} & \frac{1}{2}n_{11121} & n_{12121} & n_{10101} & \frac{1}{2}n_{10111} & n_{10122} \\
 \frac{1}{2}n_{10010} & \frac{1}{2}(n_{11110} + n_{10111}) & n_{10112} + n_{12110} & 0 & 0 & 0 \\
 \frac{1}{2}n_{10000} & n_{10101} + \frac{1}{2}n_{11100} & n_{12100} & 0 & 0 & 0
 \end{array} \quad (7)$$

223 Equation 4 is an overdetermined linear system—there are many more equations than rates
 224 to be estimated. We can fit the model using a linear least squares approach, solving for $\hat{\mathbf{\epsilon}}$ by
 225 taking

$$\hat{\mathbf{\epsilon}} = \arg \min_{\mathbf{\epsilon}} \|\mathbf{V} - \mathbf{M}\mathbf{\epsilon}\|^2 \quad (8)$$

226 Our implementation takes as initial input genotypes in the Variant Call Format (vcf) and
 227 phases for haplotype blocks in Browser Extensible Data (BED) format. We fed this data in
 228 abbreviated form into R (v. 3.5.0) and solved for the error rate estimators in Equation 8 with the
 229 linear optimization algorithm L-BFGS-B as implemented in the base stats package. The Python

(v. 3.6.1) script used to abbreviate the genotype-phase combinations, and the R script applying the algorithm, are available on GitHub.

Simulating phase violations

We tested the performance of our method on simulated genotype combinations at biallelic sites from pedigrees as in Fig. 1a with simulated errors. The phase at each site was randomly assigned, consistent with the rules of Mendelian inheritance, and individual genotypes each had a chance of being in error. To quickly simulate a large number of such genotypes and errors, we drew from a probability distribution of genotypic combinations instead of individual genotypes. We assumed a neutral site frequency spectrum for each site and that the parents and partner of the focal individual were all unrelated to each other. Assuming that the three unrelated individuals in the pedigree are from a subset of N (minimum $N = 3$) unrelated genotyped individuals that have S segregating sites, the probability of a given unrelated genotypic combination, U_x , can be written as

$$P(U_x) = \sum_{i=1}^N \frac{i-1}{a_N} P(g_1|i) \cdot P(g_2|i) \cdot P(g_3|i) \quad (9)$$

where a_N is the Watterson correction factor

$$a_N = \sum_{i=1}^{N-1} \frac{1}{i} \quad (10)$$

g_n is the genotype of the n th-individual in the combination, and the conditional genotype frequency is given by

$$P(g_n|i) = \begin{cases} \left(1 - \frac{i}{N}\right)^2 & \text{for } g_n = \text{homozygote reference} \\ 2\left(\frac{i}{N}\right)\left(1 - \frac{i}{N}\right) & \text{for } g_n = \text{heterozygote} \\ \left(\frac{i}{N}\right)^2 & \text{for } g_n = \text{homozygote alternate} \end{cases} \quad (11)$$

The count for the 27 possible genotypic combinations, $\langle C_1, \dots, C_{27} \rangle$, can then be drawn from a multinomial with parameters $\langle P(U_1), \dots, P(U_{27}) \rangle$ for the number of segregating sites, S . For each genotypic combination, the C_x outcomes were divided among the possible genotypes of the focal individual, based on the parental genotypes in U_x . These counts were apportioned by drawing from a multinomial with an equal probability for each of the four possible phase and genotype combinations. These branching outcomes were in turn further divided by the possible child genotype from the focal and partner genotypes, drawn again from a multinomial with equal probability for each of the four possible outcomes. Counts across all branches were then summed to give a total count, $n_{x,p}$, for each unique genotype and phase combination, x and p .

We added errors to these counts by iterating over the genotypes in each combination, x , and drawing errors for the $n_{x,p}$ sites. Let the genotype be a vector, $g = \{\langle 1,0,0 \rangle, \langle 0,1,0 \rangle, \langle 0,0,1 \rangle\}$, for the homozygous reference, heterozygous, and homozygous alternate genotypes respectively. Given a set of error probabilities, $\varepsilon_{i>j}$, for each type of error as listed in Table 2, the probability of a genotypic transition can be written as $g \cdot \mathbf{E}$, where

$$\mathbf{E} = \begin{bmatrix} 1 - \varepsilon_{0>1} - \varepsilon_{0>2} & \varepsilon_{0>1} & \varepsilon_{0>2} \\ \varepsilon_{1>0} & 1 - \varepsilon_{1>0} - \varepsilon_{1>2} & \varepsilon_{1>2} \\ \varepsilon_{2>0} & \varepsilon_{2>1} & 1 - \varepsilon_{2>1} - \varepsilon_{2>2} \end{bmatrix} \quad (12)$$

For each genotype at each combination, the $n_{x,p}$ counts are divided by drawing from a multinomial using the probability for genotypic transition while retaining the phase. Unique genotype and phase combinations across all branches are then summed for the final counts.

Simulations following the above strategy were implemented in Python (v. 3.6.1) with the NumPy package (v. 1.13.3).

Data Availability

Raw sequence data for the owl monkey dataset is available from NCBI BioProject: PRJNA451475. Abbreviated counts of genotype-phase combinations from the owl monkey dataset are available at FigShare. Code used in analyses and simulations are publicly available on GitHub (<https://github.com/Wang-RJ/genotypeErrors>).

Results

Accuracy of error estimators in simulations

We simulated genotypes with varying error rates and numbers of segregating sites, maintaining the sample size of unrelated genotyped individuals at $N = 20$. Simulated error rates were drawn from two distributions, a log-uniform distribution (range: $[10^{-4}, 10^{-2}]$) and a log-normal distribution ($\mu = 10^{-3}$, $\sigma = 0.5$). We simulated 1000 genotyped pedigrees of 5 individuals for each combination of parameters tested. In the absence of genotyping errors, the frequency of parental phases at informative sites is expected to be equal due to independent assortment. However, genotyping errors cause apparent phase violations to occur at different frequencies based on the genotypic combination of individuals in the pedigree. Figure 2a shows an example of how the parental phase at detectable violations consistently varies from the balanced phases at informative sites. The frequency of simulated genotypic combinations also reflects the neutral site frequency spectrum (see Materials and Methods).

We assessed the accuracy of our error rate estimation in these simulations by calculating the normalized root mean square deviation (NRMSD). The deviation between the estimators and the simulated rate was normalized by the range of error rates as

$$NRMSD = \frac{\sqrt{\sum(\hat{\epsilon} - \epsilon)^2}}{\epsilon_{\max} - \epsilon_{\min}} \quad (13)$$

where $\epsilon_{\max}$ and $\epsilon_{\min}$ are the maximum and minimum error rates in each simulation.

There was little change to the estimators' normalized deviation with an increasing number of segregating sites for simulated error rates drawn from either the log-normal or log-uniform distribution. The mean NRMSD for simulated rates drawn from the log-normal distribution was slightly higher, ranging from 9.0% to 12.2%, than from rates drawn from the

log-uniform distribution, from 6.5% to 7.1%, for simulations with between 10 and 70 million segregating sites (Fig. 2b). These results indicate that our method of estimating error rates is robust across populations with different levels of nucleotide diversity.

Figure 2. Estimating simulated genotyping error rates from phase violations

- (a) Typical genotypic combinations from simulations, ordered by the expected frequency of informative sites. Phases are expected to occur at equal frequency for informative sites (top), but violations due to genotyping error do not occur equally (bottom). Depicted are the mean frequencies from 1,000 simulations with error rates drawn from a log-uniform distribution.
- (b) Total deviation of rate estimators decreases modestly with more segregating sites in simulated samples.
- (c) Deviation of individual estimators from actual rates used in simulations.

We also assessed our method's accuracy for each estimator, normalizing the RMSD to the median difference between maximum and minimum error rates across all simulations (Fig.

2c). We found $\hat{\epsilon}_{0>2}$ to be the most accurately estimated, with a mean NRMSD of 2.1% and 1.2%, for simulated errors drawn from the log-normal and log-uniform respectively. The estimators $\hat{\epsilon}_{1>2}$ and $\hat{\epsilon}_{2>1}$ were the least accurate, with mean NRMSDs ranging from 8.0% to 12.4%. The accuracy of the estimator for error rates at the most frequent genotype, $\hat{\epsilon}_{0>1}$, was intermediate with an NRMSD of 4.3% and 6.1% for respective draws from the log-normal and log-uniform distributions.

Application to phased owl monkey genotypes

We applied the method developed here to a dataset of genotypes from the whole-genome sequencing of owl monkey (*Aotus nancymaae*). Individuals in this dataset were part of several three-generation pedigrees, allowing us to unambiguously phase focal individuals (as in Fig. 1). We selected four unrelated pedigrees with genotypes from twenty total individuals for our analysis. These samples were sequenced to approximately 35× coverage on an Illumina HiSeq-X with 150 bp paired-end reads. Single nucleotide polymorphisms (SNPs) in this dataset were called with GATK (version 3.3.0) following best practices (Van der Auwera *et al.* 2013; Poplin *et al.* 2017), and genotypes were phased with PhaseByTransmission (Francioli *et al.* 2017) and assembled into haplotype blocks under the assumption that there would be at most one recombination crossover per Mb interval (Venn *et al.* 2014; Smeds *et al.* 2016). Complete details on the methods used to generate this dataset are available in Thomas *et al.* (2018).

We selected genotypes from biallelic SNPs on autosomes for the subsequent analyses. Based on genotypic combinations, we were able to identify approximately 5.5 million phase-informative sites in each of the pedigrees. These sites were on haplotype blocks that covered, on average, 2.46 Gb out of the 2.86 Gb genome. Among the informative sites, an average of 25,007

were out-of-phase in each family. Frequencies for the different genotypic combinations showed a similar pattern to those seen in simulations: a balanced set of phases among all informative sites and an unequal set of phases among the violations (Fig. S1). Note that the proportion of out-of-phase sites relative to the total number of informative sites does not represent an error rate because potential miscalls are spread across the genotypes of multiple individuals in the pedigree.

Using the frequencies of different phase violations, we estimated the genotyping error rates for each family with Equation 4. Estimates from each of the four pedigrees are depicted in Figure 3. The highest rates of error appear to be from distinguishing between homozygous reference and heterozygous genotypes, with mean rates $\epsilon_{0>1} = 2.9 \times 10^{-3}$ and $\epsilon_{1>0} = 2.9 \times 10^{-3}$ per bp. The rate at which homozygotes for the alternate allele were miscalled as heterozygotes was also appreciable, mean rate $\epsilon_{2>1} = 2.1 \times 10^{-3}$ per bp. In contrast, our estimate of the rate at which homozygous reference genotypes were mistaken as homozygous alternate was zero in all four pedigrees.

We repeated the analysis with genotypes that were filtered by genotype quality (GQ) as calculated in GATK. The value of GQ is based on the difference between the probability of the called genotype and the next most likely genotype. Genotypes associated with higher GQ scores are typically interpreted as being more accurate. As expected, the estimated error rates decreased as we removed sites with more stringent GQ filters (Fig. 3; Table S1). Filtering by GQ appears to be more effective at removing certain types of errors than others, with the most dramatic reduction in heterozygote false negatives, that is $\epsilon_{1>0}$.

We also calculated an average error rate by weighting each estimator with the genome-wide frequency of the corresponding genotype (e.g. the frequency of homozygous reference

genotypes for $\epsilon_{0>1}$). As with individual estimators, increasing stringency of the GQ filter reduced the overall error rate (Table S1). We estimated an overall error rate of 3.0×10^{-3} per bp, which was reduced to 1.1×10^{-3} per bp when genotypes were required to have a GQ > 60.

Figure 3. Error rates estimated from sequencing four pedigrees of owl monkeys. Each point represents an estimate from one pedigree with genotypes filtered as described in the legend. Estimated genotyping error rates consistently decline with the stringency of the GQ filter.

Discussion

We have developed a method to estimate genotyping error rates for different types of errors at biallelic loci. Leveraging pedigree information, our method directly estimates underlying error rates, rather than the discordance between experiments obtained by other approaches. Our method is more robust than those that consider only Mendelian violations in a trio of individuals because of additional transmission information, reliance on multiple biological phenomena, and the ability to distinguish different types of errors.

Our estimate of the overall genotyping error rate in the owl monkey samples is comparable to estimates calculated from discordance between replicate sequencing experiments. Wall et al. (2014) inferred a genotyping error rate of 1.18×10^{-3} per bp on the Illumina HiSeq platform at GQ ≥ 40 , remarkably similar to our overall estimate of 1.3×10^{-3} per bp at the same level of filtering stringency (Table S1). Though their resequencing approach does not distinguish between all types of miscalls, they report a false positive rate for heterozygotes ($\epsilon_{0>1}$) that is much higher than the false negative rate ($\epsilon_{1>0}$), consistent with our findings after the application of filters on GQ.

Heterozygote false positives at homozygous reference sites occur at the highest rate among all errors, after modest GQ filtering. As the most common type of site in the genome, they are also the most common genotyping error. That the lowest error rates were for erroneously called non-reference homozygotes may have been expected. The relative rarity of the non-reference allele leads to caution in calling non-reference homozygotes by most genotyping methods. More surprising is the uneven effect of filtering by GQ across error types. Heterozygote false negatives, in particular, were dramatically reduced by this filter. If the disparity between heterozygote false positive and false negative rates is common across NGS

experiments, studies that seek rare variants may not be calibrated appropriately when assuming a single error rate. While genotyping errors for low-frequency variants may have little effect for many analyses, studies looking to identify *de novo* mutations in order to estimate mutation rates are very sensitive to miscalls of these variants.

The effect of genotype filters on the false negative rate can be difficult to quantify; false positives, on the other hand, can be detected by confirming candidate sites with, for example, Sanger sequencing. We demonstrated that our method allows for an analytic estimate of the false negative rate with varying degrees of genotype filtering. At first glance, the order of magnitude difference in heterozygote false negative rates when filtering by GQ might suggest a corresponding difference in the number of false negatives in studies of *de novo* mutation. These studies, however, generally employ additional downstream filters to strictly control for the high number of false positives (Ramu *et al.* 2013; Wei *et al.* 2015). An estimate of each filter's effect on the *de novo* false positive rate may be possible by applying our method to filtered genotypes, as we have done with GQ.

Simulations demonstrated the power of our approach to estimate error rates even in samples with low levels of diversity. They also indicated differences in the accuracy of the error estimators, but these were not pronounced for the most common types of sites. One caveat to our simulations is that we did not simulate entire haplotype blocks or inaccuracies in phase calling. Low levels of diversity and high rates of recombination can make accurate phasing more difficult, while low rates of recombination could result in an imbalance in the proportion of phases across the genome. The assumption of at most one genotyping error among samples per site may slightly inflate our estimates of the error rate. The potential for two or more genotyping errors is small (on the order of 10^{-6}), and had a negligible effect in our simulations, but may be

higher at sites prone to sequencing or assembly errors. Similarly, we ignored the effects of gene conversion and *de novo* mutation, as they are expected to occur at negligible rates compared to genotyping error (Table 1). Furthermore, the signal from a genotyping error and a gene conversion event are nearly identical, though careful filtering and selection of sites have been successful in identifying gene conversion events (e.g. Williams *et al.* 2015; Miller *et al.* 2016). Finally, we note that the estimated error rates are for genotyping from a set of called variants. The heterozygote false positive rate, $\epsilon_{0>1}$, for example, does not apply to invariant reference sites.

As genomic data continues to accumulate, the consideration of genotyping errors will remain an essential part of genetic analyses. Though we have focused mainly on whole-genome sequence data, our approach is generally applicable to any collection of genotype data (e.g. SNP-chips or exome sequencing) from pedigreed samples. Interest in sequencing individuals from families, as in studies seeking to identify *de novo* mutations (Goldmann *et al.* 2016; Thomas *et al.* 2018; Sasani *et al.* 2019), provide special opportunities for this method to be useful. Studies estimating *de novo* mutation rates may be particularly interested in distinguishing the rates for different types of genotyping errors. Differences in error rates will directly affect estimates of the false positive and false negative rates, and subsequent calculations of the mutation rate (Besenbacher *et al.* 2015; Kim *et al.* 2019). We have shown here that different types of errors indeed occur at different rates, necessitating their inclusion in such studies.

Acknowledgements

We thank Gregg Thomas, Jeff Rogers, and Alan Harris for help with the owl monkey data, and Rafael Guerrero and Mark Hibbins for thoughtful discussions that led to the development of this manuscript. This work was funded by the Precision Health Initiative of Indiana University.

Literature Cited

- Abecasis G. R., S. S. Cherny, and L. R. Cardon, 2001 The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* 9: 130–134.
- Abecasis G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon, 2002 Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30: 97–101.
- Ahn K., C. Haynes, W. Kim, R. St. Fleur, D. Gordon, *et al.*, 2007 The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann. Hum. Genet.* 71: 249–261.
- Becker T., R. Valentonyte, P. J. P. Croucher, K. Strauch, S. Schreiber, *et al.*, 2006 Identification of probable genotyping errors by consideration of haplotypes. *Eur. J. Hum. Genet.* 14: 450–458.
- Besenbacher S., S. Liu, J. M. G. Izarzugaza, J. Grove, K. Belling, *et al.*, 2015 Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nat. Commun.* 6: 5969.
- Carlson J., A. E. Locke, M. Flickinger, M. Zawistowski, S. Levy, *et al.*, 2018 Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat. Commun.* 9: 3753.
- Cartwright D. A., M. Troggio, R. Velasco, and A. Gutin, 2007 Genetic mapping in the presence of genotyping errors. *Genetics* 176: 2521–2527.

Coop G., and M. Przeworski, 2007 An evolutionary view of human recombination. *Nat. Rev. Genet.* 8: 23–34.

Dohm J. C., C. Lottaz, T. Borodina, and H. Himmelbauer, 2008 Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36: e105.

Douglas J. A., A. D. Skol, and M. Boehnke, 2002 Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am. J. Hum. Genet.* 70: 487–495.

Fledel-Alon A., D. J. Wilson, K. Broman, X. Wen, C. Ober, *et al.*, 2009 Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet.* 5: e1000658.

Francioli L. C., M. Cretu-Stancu, K. V. Garimella, M. Fromer, W. P. Kloosterman, *et al.*, 2017 A framework for the detection of de novo mutations in family-based sequencing data. *Eur. J. Hum. Genet.* 25: 227–233.

Goldmann J. M., W. S. W. Wong, M. Pinelli, T. Farrah, D. Bodian, *et al.*, 2016 Parent-of-origin-specific signatures of *de novo* mutations. *Nat. Genet.* 48: 935–939.

Gordon D., S. J. Finch, M. Nothnagel, and J. Ott, 2002 Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum. Hered.* 54: 22–33.

Halldorsson B. V., M. T. Hardarson, B. Kehr, U. Styrkarsdottir, A. Gylfason, *et al.*, 2016 The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* 48: 1377–1384.

Hao K., C. Li, C. Rosenow, and W. Hung Wong, 2004 Estimation of genotype error rate using samples with pedigree information—an application on the GeneChip Mapping 10K array. *Genomics* 84: 623–630.

Huang X., Q. Feng, Q. Qian, Q. Zhao, L. Wang, *et al.*, 2009 High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19: 1068–1076.

Jeffreys A. J., and C. A. May, 2004 Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* 36: 151–156.

Jensen-Seaman M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, *et al.*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14: 528–538.

Kim Y.-H., Y. Song, J.-K. Kim, T.-M. Kim, H. W. Sim, *et al.*, 2019 False-negative errors in next-generation sequencing contribute substantially to inconsistency of mutation databases. *PLoS ONE* 14: e0222535.

Kong A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, *et al.*, 2012 Rate of *de novo* mutations and the importance of father’s age to disease risk. *Nature* 488: 471–475.

Kothiyal P., W. S. W. Wong, D. L. Bodian, and J. E. Niederhuber, 2019 Mendelian inconsistent signatures from 1314 ancestrally diverse family trios distinguish biological variation from sequencing error. *J. Comput. Biol.* 26: 405–419.

Lebrec J. J., H. Putter, J. J. Houwing-Duistermaat, and H. C. van Houwelingen, 2008 Influence of genotyping error in linkage mapping for complex traits – an analytic study. *BMC Genet.* 9: 57.

- Ma X., Y. Shao, L. Tian, D. A. Flasch, H. L. Mulder, *et al.*, 2019 Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20: 50.
- Miller D. E., C. B. Smith, N. Y. Kazemi, A. J. Cockrell, A. V. Arvanitakis, *et al.*, 2016 Whole-genome analysis of individual meiotic events in *Drosophila melanogaster* reveals that noncrossover gene conversions are insensitive to interference and the centromere effect. *Genetics* 203: 159–171.
- Nielsen R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12: 443–451.
- O’Connell J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, *et al.*, 2014 A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 10: e1004234.
- Pfeiffer F., C. Gröber, M. Blank, K. Händler, M. Beyer, *et al.*, 2018 Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* 8: 10950.
- Pompanon F., A. Bonin, E. Bellemain, and P. Taberlet, 2005 Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* 6: 847–859.
- Poplin R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, *et al.*, 2017 Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178.
- Powers S., S. Gopalakrishnan, and N. Tintle, 2011 Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Hum. Hered.* 72: 153–160.
- Ramu A., M. J. Noordam, R. S. Schwartz, A. Wuster, M. E. Hurles, *et al.*, 2013 DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* 10: 985–987.

Roach J. C., G. Glusman, R. Hubley, S. Z. Montsaroff, A. K. Holloway, *et al.*, 2011
Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* 89:
382–397.

Sasani T. A., B. S. Pedersen, Z. Gao, L. Baird, M. Przeworski, *et al.*, 2019 Large, three-
generation human families reveal post-zygotic mosaicism and variability in germline
mutation accumulation. *eLife* 8: e46922.

Saunders I. W., J. Brohede, and G. N. Hannan, 2007 Estimating genotyping error rates from
Mendelian errors in SNP array genotypes and their impact on inference. *Genomics* 90:
291–296.

Ségurel L., M. J. Wyman, and M. Przeworski, 2014 Determinants of mutation rate variation in
the human germline. *Annu. Rev. Genomics Hum. Genet.* 15: 47–70.

Smeds L., C. F. Mugal, A. Qvarnström, and H. Ellegren, 2016 High-resolution mapping of
crossover and non-crossover recombination events by whole-genome re-sequencing of an
avian pedigree. *PLoS Genet.* 12: e1006044.

Sobel E., J. C. Papp, and K. Lange, 2002 Detection and integration of genotyping errors in
statistical genetics. *Am. J. Hum. Genet.* 70: 496–508.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation.
Nature 526: 68–74.

Thomas G. W. C., R. J. Wang, A. Puri, R. A. Harris, M. Raveendran, *et al.*, 2018 Reproductive
longevity predicts mutation rates in primates. *Curr. Biol.* 28: 3193–3197.

Van der Auwera G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, *et al.*, 2013 From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43: 11.10.1-11.10.33.

Venn O., I. Turner, I. Mathieson, N. de Groot, R. Bontrop, *et al.*, 2014 Strong male bias drives germline mutation in chimpanzees. *Science* 344: 1272–1275.

Wall J. D., L. F. Tang, B. Zerbe, M. N. Kvale, P.-Y. Kwok, *et al.*, 2014 Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res.* 24: 1734–1739.

Wei Q., X. Zhan, X. Zhong, Y. Liu, Y. Han, *et al.*, 2015 A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* 31: 1375–1381.

Williams A. L., G. Genovese, T. Dyer, N. Altemose, K. Truax, *et al.*, 2015 Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* 4: e04637.

Yan Q., R. Chen, J. S. Sutcliffe, E. H. Cook, D. E. Weeks, *et al.*, 2016 The impact of genotype calling errors on family-based studies. *Sci. Rep.* 6: 28323.

Supplemental Material

Figure S1. Mean number of informative sites and phase violations in the owl monkey dataset

Phases are at roughly equal frequencies for all informative sites relative to what is observed at violations. The overall frequency of genotypic combinations depends on the site frequency spectrum across the genome in owl monkeys. Genotypic combinations ordered as in Figure 2.

Table S1. Mean genotyping error rates ($\times 10^{-3}$ per bp) estimated in the owl monkey dataset

GQ filter	$\epsilon_{0>1}$	$\epsilon_{1>0}$	$\epsilon_{2>0}$	$\epsilon_{0>2}$	$\epsilon_{1>2}$	$\epsilon_{2>1}$	Genome-wide ^a
None	2.9 (0.2)	2.9 (0.6)	0.8 (0.5)	0.0 (0.0)	0.4 (0.1)	2.1 (0.3)	3.0 (0.3)
> 20	2.0 (0.1)	0.4 (0.3)	0.5 (0.3)	0.0 (0.0)	0.1 (0.1)	1.2 (0.2)	1.6 (0.2)
> 40	1.7 (0.1)	0.3 (0.2)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.8 (0.2)	1.3 (0.1)
> 60	1.6 (0.1)	0.2 (0.2)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.5 (0.2)	1.1 (0.1)

Standard errors in parentheses.

^a Genome-wide estimated calculated by weighting rates with the normalized frequency of homozygote reference, heterozygote, and homozygote alternate genotypes in the owl monkey genome, respectively: 0.64, 0.25, and 0.11

Appendix S1. Derivation of phase violation expected frequencies

Expected number of phase violations A_{0111} , A_{2111}

Expected: G_{01110} ; G_{01111}

Observed: G_{01112}

Parent A Observed genotype: 0/0

Possible miscall: 1/1

Focal individual inherits alternate allele from Parent A and transmits it to the child leading to apparent phase violation.

Possible miscall: 0/1

Focal individual inherits the alternate allele from Parent A, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{11112} \cdot \varepsilon_{1>0} + n_{21112} \cdot \varepsilon_{2>0}$

Parent B Observed genotype: 0/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 1/1

Not detectable as phase violation, as focal individual still inherits alternate allele from Parent B haplotype block.

Focal Observed genotype: 0/1

Any miscall would imply more than one genotyping error.

Partner Observed genotype: 0/1

Possible miscall: 1/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/0

True genotype 0 has been miscalled as 2. Each occurrence leads to this phase violation.

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. There is a 50% chance the alternate allele was inherited from the focal individual, leading to apparent phase violation.

Frequency: $n_{01110} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{01111} \cdot \varepsilon_{1>2}$

$$E[A_{0111}] = \frac{1}{2}n_{11112} \cdot \varepsilon_{1>0} + n_{21112} \cdot \varepsilon_{2>0} + n_{01110} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{01111} \cdot \varepsilon_{1>2}$$

$$E[A_{2111}] = \frac{1}{2}n_{11110} \cdot \varepsilon_{1>2} + n_{01110} \cdot \varepsilon_{0>2} + n_{21112} \cdot \varepsilon_{2>0} + \frac{1}{2}n_{21111} \cdot \varepsilon_{1>0}$$

Expected number of phase violations A_{0112} , A_{2110}

Expected: G_{01121}

Observed: G_{01122}

Parent A Observed genotype: 0/0

Possible miscall: 1/1

Focal individual inherits alternate allele from Parent A and transmits it to the child leading to apparent phase violation.

Possible miscall: 0/1

Focal individual inherits the alternate allele from Parent A, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{11122} \cdot \varepsilon_{1>0} + n_{21122} \cdot \varepsilon_{2>0}$

Parent B Observed genotype: 0/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 1/1

Not detectable as phase violation, as focal individual still inherits alternate allele from Parent B haplotype block.

Focal Observed genotype: 0/1

Any miscall would imply more than one genotyping error.

Partner Observed genotype: 1/1

Possible miscall: 0/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. Each occurrence leads to this phase violation.

Frequency: $n_{01121} \cdot \varepsilon_{1>2}$

$$E[A_{0112}] = \frac{1}{2}n_{11122} \cdot \varepsilon_{1>0} + n_{21122} \cdot \varepsilon_{2>0} + n_{01121} \cdot \varepsilon_{1>2}$$

$$E[A_{2110}] = \frac{1}{2}n_{11100} \cdot \varepsilon_{1>2} + n_{01100} \cdot \varepsilon_{0>2} + n_{21101} \cdot \varepsilon_{1>0}$$

Expected number of phase violations A_{0210} , A_{2012}

Expected: G_{02100}

Observed: G_{02101}

Parent A Observed genotype: 0/0

Any miscall would imply more than one genotyping error.

Parent B Observed genotype: 1/1

Any miscall would imply more than one genotyping error.

Focal Observed genotype: 0/1

Any miscall would imply more than one genotyping error.

Partner Observed genotype: 0/0

Possible miscall: 1/1

Child inherits alternate allele from the partner, which appears as a phase violation.

Possible miscall: 0/1

Child inherits alternate allele from the partner, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{02111} \cdot \varepsilon_{1>0} + n_{02121} \cdot \varepsilon_{2>0}$

Child Observed genotype: 0/1

Possible miscall: 0/0

True genotype 0 has been miscalled as 1. Each occurrence leads to this phase violation.

Frequency: $n_{02100} \cdot \varepsilon_{0>1}$

$$E[A_{0210}] = n_{02100} \cdot \varepsilon_{0>1} + \frac{1}{2}n_{02111} \cdot \varepsilon_{1>0} + n_{02121} \cdot \varepsilon_{2>0}$$

$$E[A_{2012}] = n_{20122} \cdot \varepsilon_{2>1} + \frac{1}{2}n_{20111} \cdot \varepsilon_{1>2} + n_{20101} \cdot \varepsilon_{0>2}$$

Expected number of phase violations A_{0211} , A_{2011}

Expected: G_{02110} ; G_{02111}

Observed: G_{02112}

Parent A Observed genotype: 0/0

Any miscall would imply more than one genotyping error.

Parent B Observed genotype: 1/1

Any miscall would imply more than one genotyping error.

Focal Observed genotype: 0/1

Any miscall would imply more than one genotyping error.

Partner Observed genotype: 0/1

Possible miscall: 1/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/0

True genotype 0 has been miscalled as 2. Each occurrence leads to this phase violation.

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. There is a 50% chance the alternate allele was inherited from the focal individual, leading to apparent phase violation.

Frequency: $n_{02110} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{02111} \cdot \varepsilon_{1>2}$

$$E[|A_{0211}|] = n_{02110} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{02111} \cdot \varepsilon_{1>2}$$

$$E[|A_{2011}|] = n_{20112} \cdot \varepsilon_{2>0} + \frac{1}{2}n_{20111} \cdot \varepsilon_{1>0}$$

Expected number of phase violations A_{0212} , A_{2010}

Expected: G_{02121}

Observed: G_{02122}

Parent A Observed genotype: 0/0

Any miscall would imply more than one genotyping error.

Parent B Observed genotype: 1/1

Any miscall would imply more than one genotyping error.

Focal Observed genotype: 0/1

Any miscall would imply more than one genotyping error.

Partner Observed genotype: 1/1

Possible miscall: 0/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. Each occurrence leads to this phase violation.

Frequency: $n_{02121} \cdot \epsilon_{1>2}$

$$E[A_{0212}] = n_{02121} \cdot \epsilon_{1>2}$$

$$E[A_{2010}] = n_{20101} \cdot \epsilon_{2>1}$$

Expected number of phase violations A_{1212} , A_{1010}

Expected: G_{12121}

Observed: G_{12122}

Parent A Observed genotype: 0/1

Possible miscall: 0/0

Not detectable as phase violation, as focal individual still inherits reference allele from Parent A haplotype block.

Possible miscall: 1/1

Implies more than one genotyping error across pedigree.

Parent B Observed genotype: 1/1

Possible miscall: 0/0

Focal individual inherits reference allele from Parent B and transmits it to the child leading to apparent phase violation.

Possible miscall: 0/1

Focal individual inherits the reference allele from Parent B, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $n_{10122} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{11122} \cdot \varepsilon_{1>2}$

Focal Observed genotype: 0/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 1/1

True genotype 2 has been miscalled as 1. There is a 50% chance child inherits alternate allele from Parent A, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{12222} \cdot \varepsilon_{2>1}$

Partner Observed genotype: 1/1

Possible miscall: 0/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. Each occurrence leads to this phase violation.

Frequency: $n_{12121} \cdot \varepsilon_{1>2}$

$$E[|A_{1212}|] = n_{10122} \cdot \varepsilon_{0>2} + (\frac{1}{2}n_{11122} + n_{12121}) \cdot \varepsilon_{1>2} + \frac{1}{2}n_{12222} \cdot \varepsilon_{2>1}$$

$$E[|A_{1010}|] = n_{12100} \cdot \varepsilon_{2>0} + (\frac{1}{2}n_{11100} + n_{10101}) \cdot \varepsilon_{1>0} + \frac{1}{2}n_{10000} \cdot \varepsilon_{0>1}$$

Expected number of phase violations A_{1211} , A_{1011}

Expected: G_{12110} ; G_{12111}

Observed: G_{12112}

Parent A Observed genotype: 0/1

Possible miscall: 0/0

Not detectable as phase violation.

Possible miscall: 1/1

Implies more than one genotyping error across pedigree.

Parent B Observed genotype: 1/1

Possible miscall: 0/0

Focal individual inherits reference allele from Parent *B* and transmits it to the child leading to apparent phase violation.

Possible miscall: 0/1

Focal individual inherits the reference allele from Parent *B*, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $n_{10112} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{11112} \cdot \varepsilon_{1>2}$

Focal Observed genotype: 0/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 1/1

True genotype 2 has been miscalled as 1. There is a 50% chance child inherits alternate allele from Parent *A*, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{12212} \cdot \varepsilon_{2>1}$

Partner Observed genotype: 0/1

Possible miscall: 1/1

Not detectable as phase violation, child still inherits alternate allele from Partner.

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Child Observed genotype: 1/1

Possible miscall: 0/0

Each occurrence leads to this phase violation.

Possible miscall: 0/1

True genotype 1 has been miscalled as 2. There is a 50% chance the alternate allele was inherited from the focal individual, leading to apparent phase violation.

Frequency: $n_{12110} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{12111} \cdot \varepsilon_{1>2}$

$$E[|A_{1211}|] = (n_{10112} + n_{12110}) \cdot \varepsilon_{0>2} + \frac{1}{2}(n_{11112} + n_{12111}) \cdot \varepsilon_{1>2} + \frac{1}{2}n_{12212} \cdot \varepsilon_{2>1}$$

$$E[|A_{1011}|] = (n_{12111} + n_{10112}) \cdot \varepsilon_{2>0} + \frac{1}{2}(n_{11110} + n_{10111}) \cdot \varepsilon_{1>0} + \frac{1}{2}n_{10010} \cdot \varepsilon_{0>1}$$

Expected number of phase violations A_{1210} , A_{1012}

Expected: G_{12100}

Observed: G_{12101}

Parent A Observed genotype: 0/1

Possible miscall: 0/0

Not detectable as phase violation.

Possible miscall: 1/1

Implies more than one genotyping error across pedigree.

Parent B Observed genotype: 1/1

Possible miscall: 0/0

Focal individual inherits reference allele from Parent *B* and transmits it to the child leading to apparent phase violation.

Possible miscall: 0/1

Focal individual inherits the reference allele from Parent *B*, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $n_{10101} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{11101} \cdot \varepsilon_{1>2}$

Focal Observed genotype: 0/1

Possible miscall: 0/0

Implies more than one genotyping error across pedigree.

Possible miscall: 1/1

True genotype 2 has been miscalled as 1. There is a 50% chance child inherits alternate allele from Parent *A*, leading to apparent phase violation.

Frequency: $\frac{1}{2}n_{12201} \cdot \varepsilon_{2>1}$

Partner Observed genotype: 0/0

Possible miscall: 1/1

Each occurrence leads to this phase violation.

Possible miscall: 0/1

Child inherits alternate allele from the partner, giving a 50% chance of transmission to child, leading to apparent phase violation.

Frequency: $n_{12121} \cdot \varepsilon_{2>0} + \frac{1}{2}n_{12111} \cdot \varepsilon_{1>0}$

Child Observed genotype: 0/1

Possible miscall: 0/0

True genotype 0 has been miscalled as 1. Each occurrence leads to this phase violation.

Frequency: $n_{12100} \cdot \varepsilon_{0>1}$

$$E[|A_{1210}|] = n_{12100} \cdot \varepsilon_{0>1} + \frac{1}{2}n_{12111} \cdot \varepsilon_{1>0} + n_{12121} \cdot \varepsilon_{2>0} + n_{10101} \cdot \varepsilon_{0>2} + \frac{1}{2}n_{11101} \cdot \varepsilon_{1>2} + \frac{1}{2}n_{12201} \cdot \varepsilon_{2>1}$$

$$E[|A_{1012}|] = n_{10122} \cdot \varepsilon_{2>1} + \frac{1}{2}n_{10111} \cdot \varepsilon_{1>2} + n_{10101} \cdot \varepsilon_{0>2} + n_{12121} \cdot \varepsilon_{2>0} + \frac{1}{2}n_{11121} \cdot \varepsilon_{1>0} + \frac{1}{2}n_{10021} \cdot \varepsilon_{0>1}$$