Supplemental Material for

"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

Supplemental Figures

(a)

Parent A	Parent B	Focal	Partner
0/0	0/1	0/1	0/0
0/0	0/1	0/1	0/1
0/0	0/1	0/1	1/1
1/1	0/1	0/1	0/0
1/1	0/1	0/1	0/1
1/1	0/1	0/1	1/1
0/0	1/1	0/1	0/0
0/0	1/1	0/1	0/1
0/0	1/1	0/1	1/1
1/1	0/0	0/1	1/1
1/1	0/0	0/1	0/1
1/1	0/0	0/1	0/0
0/1	1/1	0/1	1/1
0/1	1/1	0/1	0/1
0/1	1/1	0/1	0/0
0/1	0/0	0/1	1/1
0/1	0/0	0/1	0/1
0/1	0/0	0/1	0/0

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

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

(b)

Parent A	Parent B	Focal	Partner
0/0	0/1	0/1	0/0
0/0	0/1	0/1	0/1
0/0	0/1	0/1	1/1
1/1	0/1	0/1	0/0
1/1	0/1	0/1	0/1
1/1	0/1	0/1	1/1
0/0	1/1	0/1	0/0
0/0	1/1	0/1	0/1
0/0	1/1	0/1	1/1
1/1	0/0	0/1	1/1
1/1	0/0	0/1	0/1
1/1	0/0	0/1	0/0
0/1	1/1	0/1	1/1
0/1	1/1	0/1	0/1
0/1	1/1	0/1	0/0
0/1	0/0	0/1	1/1
0/1	0/0	0/1	0/1
0/1	0/0	0/1	0/0

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

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

Figure S1. Symmetries in phase violating genotype combinations

(a) Frequency of A_{0110} and B_{1010} are expected to be the same from independent assortment. Highlighted entries indicate swap of Parent A and Parent B, flipping the expected phase.

(b) Frequency of A_{0110} and A_{2112} are expected to be the same as segregation patterns for reference and non-reference alleles are identical.

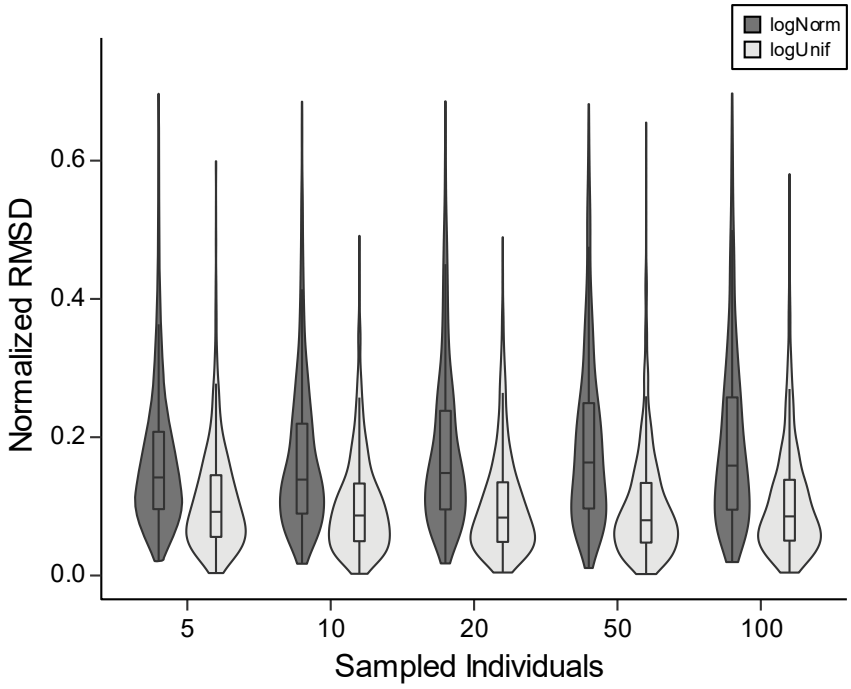

Figure S2. Effect on estimators of varying number of sampled individuals

Total deviation of rate-estimators in simulations with 40 million segregating sites. The number of sampled individuals in this range has little effect on the deviation of the estimators.

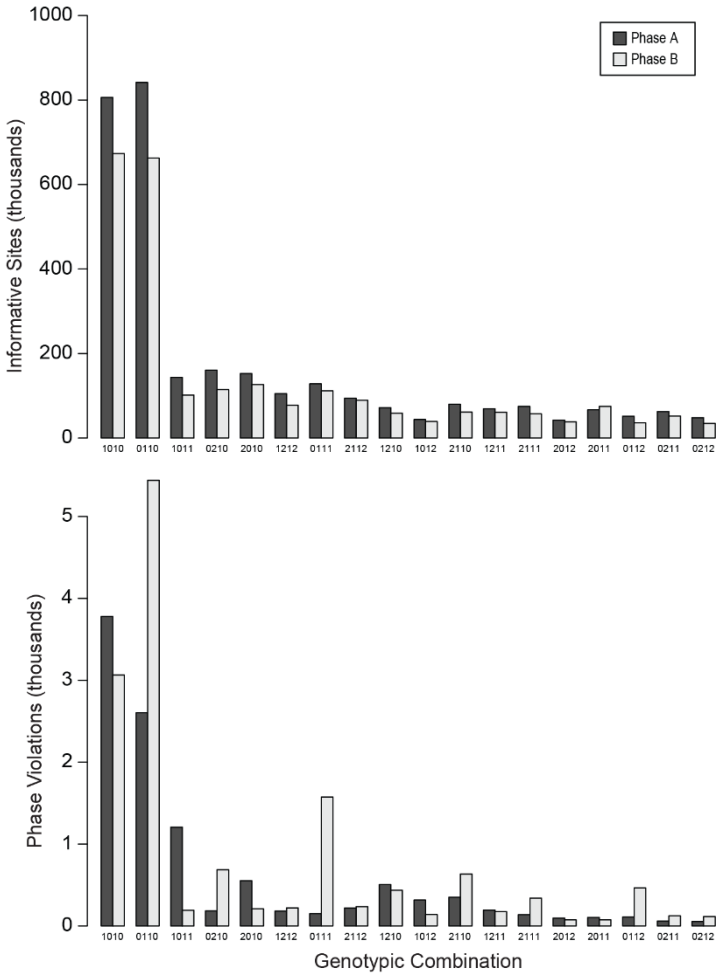

Figure S3. Mean number of informative sites and phase violations in the owl monkey dataset

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.

Supplemental Tables

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

GQ filter	$\varepsilon_{0>1}$	$\varepsilon_{1>0}$	$\varepsilon_{2>0}$	$\varepsilon_{0>2}$	$\varepsilon_{1>2}$	$\varepsilon_{2>1}$	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

Table S2. Mean genotyping error rates (×10^{-3} per bp) for sites with low minor allele frequencies

MAF filter	$\varepsilon_{0>1}$	$\varepsilon_{1>0}$	$\varepsilon_{2>0}$	$\varepsilon_{0>2}$	$\varepsilon_{1>2}$	$\varepsilon_{2>1}$	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)
< 0.5	2.5 (0.3)	3.2 (0.7)	3.2 (0.8)	0.0 (0.0)	0.7 (0.2)	3.7 (0.9)	2.8 (0.4)
< 0.2	2.3 (0.3)	3.1 (0.5)	3.9 (1.8)	0.0 (0.0)	1.3 (0.1)	2.8 (1.1)	2.8 (0.6)
< 0.1	1.6 (0.3)	3.6 (0.5)	4.1 (3.3)	0.0 (0.0)	1.1 (0.1)	3.7 (2.7)	3.0 (1.6)

Standard errors in parentheses.

[a] Genome-wide estimated calculated by weighting rates with the normalized frequencies as in Table S1.

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: $½n_{\mathbf{1}1112} \cdot \varepsilon_{1>0} + n_{\mathbf{2}1112} \cdot \varepsilon_{2>0}$

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: $n_{01110} \cdot \varepsilon_{0>2} + ½n_{01111} \cdot \varepsilon_{1>2}$

$$\mathrm{E}[|A_{0111}|] = ½n_{11112} \cdot \varepsilon_{1>0} + n_{21112} \cdot \varepsilon_{2>0} + n_{01110} \cdot \varepsilon_{0>2} + ½n_{01111} \cdot \varepsilon_{1>2}$$

$$\mathrm{E}[|A_{2111}|] = ½n_{11110} \cdot \varepsilon_{1>2} + n_{01110} \cdot \varepsilon_{0>2} + n_{21112} \cdot \varepsilon_{2>0} + ½n_{21111} \cdot \varepsilon_{1>0}$$

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: $½n_{\mathbf{1}1122} \cdot \varepsilon_{1>0} + n_{\mathbf{2}1122} \cdot \varepsilon_{2>0}$

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.

Possible miscall: 1/1
Implies more than one genotyping error across pedigree.

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

$$\mathrm{E}[|A_{0112}|] = ½n_{11122} \cdot \varepsilon_{1>0} + n_{21122} \cdot \varepsilon_{2>0} + n_{01121} \cdot \varepsilon_{1>2}$$

$$\mathrm{E}[|A_{2110}|] = ½n_{11100} \cdot \varepsilon_{1>2} + n_{01100} \cdot \varepsilon_{0>2} + n_{21101} \cdot \varepsilon_{1>0}$$

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: $½n_{02\mathbf{1}1\mathbf{1}} \cdot \varepsilon_{1>0} + n_{02\mathbf{1}2\mathbf{1}} \cdot \varepsilon_{2>0}$

Child Observed genotype: 0/1

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

Possible miscall: 1/1
Implies more than one genotyping error across pedigree.

Frequency: $n_{0210\mathbf{0}} \cdot \varepsilon_{0>1}$

$$\mathrm{E}[|A_{0210}|] = n_{02100} \cdot \varepsilon_{0>1} + ½n_{02111} \cdot \varepsilon_{1>0} + n_{02121} \cdot \varepsilon_{2>0}$$

$$\mathrm{E}[|A_{2012}|] = n_{20122} \cdot \varepsilon_{2>1} + ½n_{20111} \cdot \varepsilon_{1>2} + n_{20101} \cdot \varepsilon_{0>2}$$

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: $n_{0211\mathbf{0}} \cdot \varepsilon_{0>2} + \frac{1}{2} n_{0211\mathbf{1}} \cdot \varepsilon_{1>2}$

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

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

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: $n_{02121} \cdot \varepsilon_{1>2}$

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

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

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: $n_{1\mathbf{0}122} \cdot \varepsilon_{0>2} + \frac{1}{2} n_{1\mathbf{1}122} \cdot \varepsilon_{1>2}$

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: $\frac{1}{2} n_{12\mathbf{2}22} \cdot \varepsilon_{2>1}$

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.

Possible miscall: 1/1
Implies more than one genotyping error across pedigree.

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

$$\mathrm{E}[|A_{1212}|] = n_{10122} \cdot \varepsilon_{0>2} + (\tfrac{1}{2} n_{11122} + n_{12121}) \cdot \varepsilon_{1>2} + \tfrac{1}{2} n_{12222} \cdot \varepsilon_{2>1}$$
$$\mathrm{E}[|A_{1010}|] = n_{12100} \cdot \varepsilon_{2>0} + (\tfrac{1}{2} n_{11100} + n_{10101}) \cdot \varepsilon_{1>0} + \tfrac{1}{2} n_{10000} \cdot \varepsilon_{0>1}$$

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: $n_{\mathbf{10}112} \cdot \varepsilon_{0>2} + \frac{1}{2} n_{\mathbf{11}112} \cdot \varepsilon_{1>2}$

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: $\frac{1}{2} n_{12\mathbf{2}12} \cdot \varepsilon_{2>1}$

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: $n_{1211\mathbf{0}} \cdot \varepsilon_{0>2} + \frac{1}{2} n_{1211\mathbf{1}} \cdot \varepsilon_{1>2}$

$$\mathrm{E}[|A_{1211}|] = (n_{10112} + n_{12110}) \cdot \varepsilon_{0>2} + \tfrac{1}{2}(n_{11112} + n_{12111}) \cdot \varepsilon_{1>2} + \tfrac{1}{2} n_{12212} \cdot \varepsilon_{2>1}$$
$$\mathrm{E}[|A_{1011}|] = (n_{12111} + n_{10112}) \cdot \varepsilon_{2>0} + \tfrac{1}{2}(n_{11110} + n_{10111}) \cdot \varepsilon_{1>0} + \tfrac{1}{2} n_{10010} \cdot \varepsilon_{0>1}$$

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: $n_{\mathbf{1}0\mathbf{1}01} \cdot \varepsilon_{0>2} + \frac{1}{2} n_{\mathbf{11}101} \cdot \varepsilon_{1>2}$

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: $\frac{1}{2} n_{12\mathbf{2}01} \cdot \varepsilon_{2>1}$

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: $n_{1212\mathbf{1}} \cdot \varepsilon_{2>0} + \frac{1}{2} n_{1211\mathbf{1}} \cdot \varepsilon_{1>0}$

Child Observed genotype: 0/1

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

Possible miscall: 1/1
Implies more than one genotyping error across pedigree.

Frequency: $n_{1210\mathbf{0}} \cdot \varepsilon_{0>1}$

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

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