Estimating and accounting for genotyping errors in RAD-seq experiments

Supporting Information

Luisa Bresadola\textsuperscript{1,*}, Vivian Link\textsuperscript{1,2,*}, C. Alex Buerkle\textsuperscript{3}, Christian Lexer\textsuperscript{3}, and Daniel Wegmann\textsuperscript{1,2}

\textsuperscript{1}Department of Biology, University of Fribourg, Fribourg, Switzerland
\textsuperscript{2}Swiss Institute of Bioinformatics, Fribourg, Switzerland
\textsuperscript{3}Department of Botany, University of Wyoming, Laramie, WY, USA
\textsuperscript{4}Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

\* These authors contributed equally

Contents

Inferring Genotyping Error Rates 2

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate Model</td>
<td>2</td>
</tr>
<tr>
<td>Inferring Genotyping Error Rates by Comparing to a Truth Set</td>
<td>2</td>
</tr>
<tr>
<td>Inferring Genotyping Error Rates from Sample Replicates</td>
<td>3</td>
</tr>
<tr>
<td>Inferring Genotyping Error Rates from Population Samples</td>
<td>4</td>
</tr>
</tbody>
</table>

Supplementary Tables 6

Supplementary Figures 7
Inferring Genotyping Error Rates

Error rate Model

Here we describe algorithms to infer per-allele genotyping error rates from hard genotype calls \( g_{il} \) obtained for individuals \( i = 1, \ldots, I \) at loci \( l = 1, \ldots, L \). Let \( g_{il} = 0, 1, 2 \) reflects the number of reference alleles and \( \gamma_{il} = 0, 1, 2 \) the corresponding true genotypes. Further, \( g = \{g_{11}, \ldots, g_{1L}, \ldots, g_{IL}\} \) and \( \gamma = \{\gamma_{11}, \ldots, \gamma_{1L}, \ldots, \gamma_{IL}\} \).

We assume a simple error model with the per allele genotyping error rates \( \epsilon = \{\epsilon_0, \epsilon_1\} \) such that

\[
\mathbb{P}(g|\gamma, \epsilon_0, \epsilon_1) = \begin{cases} 
(1 - \epsilon_0)^2 & \text{if } g = \gamma; \gamma = 0, 2 \\
\epsilon_1^2 + (1 - \epsilon_1)^2 & \text{if } g = \gamma = 1 \\
\epsilon_0^2 & \text{if } g = 0; \gamma = 2 \text{ or } g = 2; \gamma = 0 \\
2\epsilon_0(1 - \epsilon_0) & \text{if } g = 1; \gamma = 0, 2 \\
\epsilon(1 - \epsilon_1) & \text{if } g = 0, 2; \gamma = 1 
\end{cases} \tag{1}
\]

Often, not all genotype calls are affected by the same genotyping error rates. For instance, genotyping error rates might vary between calls obtained with different sequencing depth or between samples of different batches such as from different libraries or sequencing experiments. We account for this by stratifying the error rates by different sets \( s = 1, \ldots, S \) of genotypes and denote the corresponding error rates by \( \epsilon_0(s) \) and \( \epsilon_1(s) \), respectively. We further denote by \( s_{il} \) the set to which genotype call \( g_{il} \) belongs. Finally, \( \epsilon_0 = \{\epsilon_0(1), \ldots, \epsilon_0(S)\} \), \( \epsilon_1 = \{\epsilon_1(1), \ldots, \epsilon_1(S)\} \) and \( \epsilon = \{\epsilon_0, \epsilon_1\} \).

Inferring Genotyping Error Rates by Comparing to a Truth Set

We first describe an algorithm to infer per-allele genotyping error rates by comparing calls to a truth set, i.e. the case in which all \( \gamma_{il} \) are known. Under the assumption that genotyping errors are independent between loci and individuals, the likelihood of the full data is given by

\[
\mathbb{P}(g|\gamma, \epsilon) = \prod_{i=1}^{I} \prod_{l=1}^{L} \mathbb{P}(g_{il}|\gamma_{il}, \epsilon)
\]

where \( \mathbb{P}(g_{il}|\gamma_{il}, \epsilon) = \mathbb{P}(g_{il}|\gamma_{il}, \epsilon_0(s_{il}), \epsilon_1(s_{il})) \) according to (1).

In order to obtain a maximum likelihood (ML) estimate of \( \epsilon \), we numerically maximize the likelihood function by finding \( \epsilon \) such that

\[
\frac{\partial}{\partial \epsilon} \log \mathbb{P}(g|\gamma, \epsilon) = 0.
\]

Consider a table with entries \( n_{g, \gamma}(s) \) reflecting the counts of observing genotype \( g \) at a site with true genotype \( \gamma \) for all calls of set \( s \). Given this table, the log-likelihood function can be calculated as

\[
\log \mathbb{P}(n|\epsilon) = \sum_{g=0}^{2} \sum_{\gamma=0}^{2} n_{g, \gamma}(s) \log \mathbb{P}(g|\gamma, \epsilon) = \sum_{s} a_0(s) \log \epsilon_0(s) + b_0(s) \log(1 - \epsilon_0(s)) + \ldots
\]

\[
\ldots + a_1(s) \log \epsilon_1(s) + b_1(s) \log(1 - \epsilon_1(s)) + c(s) \log(\epsilon_1^2(s) + (1 - \epsilon_1(s))^2) + C
\]
We next develop an algorithm to infer genotyping error rates from individuals for which genotypes were estimated.

\[ a_0(s) = 2n_{0,2}(s) + n_{1,0}(s) + n_{1,2}(s) + 2n_{2,0}(s) \]
\[ a_1(s) = n_{0,1}(s) + n_{2,1}(s) \]
\[ b_0(s) = 2n_{0,0}(s) + n_{1,0}(s) + n_{1,2}(s) + 2n_{2,0}(s) \]
\[ b_1 = n_{0,1} + n_{2,1} \]
\[ c_0 = n_{1,1} \]

From this we get for \( e_0(s) \)

\[ \frac{\partial}{\partial e_0(s)} \log \mathbb{P}(\mathbf{n}|\mathbf{e}) = \frac{a}{e_0(s)} - \frac{b}{1 - e_0(s)} = 0, \]

and the MLE estimate

\[ e_0(s) = \frac{a_0(s)}{a_0(s) + b_0(s)}. \]

For \( e_1(s) \) we get

\[ \frac{\partial}{\partial e_1(s)} \log \mathbb{P}(\mathbf{n}|\mathbf{e}) = -2[a_1(s) + b_1(s) + 2c(s)]e_1(s)^3 + 2[2a_1(s) + b_1(s) + 3c(s)]e_1(s)^2 - \ldots \]
\[ \ldots - [3a_1(s) + b_1(s) + 2c(s)]e_1(s) + a_1(s), \]

which we set to zero using a numerical search algorithm to obtain the MLE estimate \( e_1(s) \).

**Inferring Genotyping Error Rates from Sample Replicates**

We next develop an algorithm to infer genotyping error rates from individuals for which genotypes were estimated from multiple, independent replicate data sets. Consider inferred genotypes \( \mathbf{g}_i = \{g_i^{(1)}, \ldots, g_i^{(r_i)}\} \) for multiple individuals \( i = 1, \ldots, I \), where \( g_i^{(j)} = \{g_i^{(j)}(1), \ldots, g_i^{(j)}(L)\} \) indicates the inferred genotypes of individual \( i \) at sites \( l = 1 \ldots, L \) from the \( j \)th replicate. The likelihood of the full data is then given by

\[ \mathbb{P}(\mathbf{g}|\mathbf{e}) = \prod_{i=1}^{I} \prod_{l=1}^{L} \sum_{g=0}^{2} \mathbb{P}(\gamma_{il} = g) \prod_{j=1}^{r_i} \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = g, e_0(s_{il}^{(j)}), e_1(s_{il}^{(j)})), \tag{2} \]

where \( \mathbb{P}(\gamma_{il} = g) = f_{ig} \) is the frequency of genotype \( g \) among all sites typed in individual \( i \), \( s_{il}^{(j)} \) the set to which call \( g_{il}^{(j)} \) belongs, and \( \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = g, e_0(s_{il}^{(j)}), e_1(s_{il}^{(j)})) \) is given by (1).

In order to find the MLE estimates of \( e \) and \( \mathbf{f} = \{f_1, \ldots, f_L\}, f_i = \{f_{i0}, f_{i1}, f_{i2}\} \) we will employ an EM algorithm.

**E-step.** The expected complete data likelihood is given by

\[ Q(\mathbf{e}, \mathbf{f}; \mathbf{e}', \mathbf{f}') = \mathbb{E}[\ell_c(\mathbf{e}, \mathbf{f})|\mathbf{g}, \mathbf{e}', \mathbf{f}'] \]
\[ = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{g=0}^{2} \mathbb{P}(\gamma_{il} = g|\mathbf{g}_i, \mathbf{e}_0, \mathbf{e}_1, \mathbf{f}') \left[ \log f_{ig} + \sum_{j=1}^{r_i} \log \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = g, e_0(s_{il}^{(j)}), e_1(s_{il}^{(j)})) \right], \]

where we used \( \mathbf{g}_i = \{g_{i1}^{(1)}, \ldots, g_{i1}^{(r_i)}\} \) and we have, according to Bayes formula,

\[ \mathbb{P}(\gamma_{il} = g|\mathbf{g}_i, \mathbf{e}_0(s_{il}^{(j)}), \mathbf{e}_1(s_{il}^{(j)}), \mathbf{f}') = \frac{f_{ig} \prod_{j=1}^{r_i} \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = g, e_0(s_{il}^{(j)}), e_1(s_{il}^{(j)}))}{\sum_{h=0}^{2} f_{ih} \prod_{j=1}^{r_i} \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = h, e_0(s_{il}^{(j)}), e_1(s_{il}^{(j)}))}. \]

**M-step.** We have to maximize \( Q \) subject to the constraints

\[ \sum_{g=0}^{2} f_{ig} = f_{i0} + f_{i1} + f_{i2} = 1 \tag{3} \]
for all individuals $i = 1, \ldots, I$, we form the Lagrangian

$$
\mathcal{L}(\epsilon, f, \mu) = Q - \sum_{i=0}^{I} \mu_i \left( \sum_{g=0}^{2} f_{ig} - 1 \right),
$$

where $\mu = \{\mu_1, \ldots, \mu_I\}$ is the vector of Lagrangian multipliers. All $f$ can be maximized analytically. We get the following derivatives of the Lagrangian:

$$
\frac{\partial}{\partial \mu_i} \mathcal{L} = \sum_{g=0}^{2} f_{ig} - 1 = 0
$$

$$
\frac{\partial}{\partial f_{ig}} \mathcal{L} = \frac{1}{f_{ig}} \sum_{l=1}^{L} \mathbb{P}\left( \gamma_{il} = g|g_{il}, \epsilon_0(s_{il}^{(j)}), \epsilon_1(s_{il}^{(j)}), f' \right) - \mu_i = 0.
$$

From these we get

$$
f_{ig} = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}\left( \gamma_{il} = g|g_{il}, \epsilon_0(s_{il}^{(j)}), \epsilon_1(s_{il}^{(j)}), f' \right).
$$

In contrast, we will need to maximize $Q$ with respect to each $\epsilon_0(s)$ and $\epsilon_1(s)$ numerically. For this, we need to maximize

$$
Q_{\epsilon_0(s)} = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{g \in \{0, 2\}} I(s_{il}^{(j)} = s) \mathbb{P}(\gamma_{il} = g|g_{il}, \epsilon_0(s), f') \sum_{j=1}^{r_i} \log \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = g, \epsilon_0(s)),
$$

$$
Q_{\epsilon_1(s)} = \sum_{i=1}^{I} \sum_{l=1}^{L} I(s_{il}^{(j)} = s) \mathbb{P}(\gamma_{il} = 1|g_{il}, \epsilon_1(s), f') \sum_{j=1}^{r_i} \log \mathbb{P}(g_{il}^{(j)}|\gamma_{il} = 1, \epsilon_1(s)),
$$

where the sums run only over the genotypes of set $s$, as given by the indicator function

$$
\mathcal{I}(s_{il}^{(j)} = s) = \begin{cases} 
1 & \text{if } s_{il}^{(j)} = s, \\
0 & \text{otherwise}.
\end{cases}
$$

Inferring Genotyping Error Rates from Population Samples

We finally develop an algorithm to infer genotyping error rates from multiple samples from each of multiple populations. Consider observed genotypes $g_{pi}$ from individual $i = 1, \ldots, n_p$ from population $p = 1, \ldots, P$ at locus $l = 1, \ldots, L$. Let us further denote by $g_{pi} = \{g_{pi1}, \ldots, g_{pil}\}$ and by $g_{p} = \{g_{p1}, \ldots, g_{pL}\}$. The likelihood of the full data given per allele genotyping error rates $\epsilon = \{\epsilon_1, \ldots, \epsilon_S\}$ for different sets $s = 1, \ldots, S$ of genotype calls is then given by

$$
\mathbb{P}(g|\epsilon) = \prod_{p=1}^{P} \prod_{i=1}^{n_p} \prod_{l=1}^{L} \sum_{g=0}^{2} \mathbb{P}(g_{pil}|\epsilon_0(s_{pil}), \epsilon_1(s_{pil})),
$$

where $g = \{g_1, \ldots, g_P\}$, $f_{pl}$ is the frequency in population $p$ of one of the two alleles at locus $l$, and under the assumption of Hardy-Weinberg equilibrium in each population,

$$
\mathbb{P}(g_{pil}|f_{pil}) = \begin{cases} 
2 \left( \frac{1}{f_{pil}} \right)^{g_{pil}} & \text{if } g_{pil} = 0, \\
2 f_{pil}(1 - f_{pil}) & \text{if } g_{pil} = 1, \\
\left( 1 - f_{pil} \right)^{2 - g_{pil}} & \text{if } g_{pil} = 2.
\end{cases}
$$

Finally, $s_{pil}$ is the set of call $g_{pil}$ and $\mathbb{P}\left( g_{il}^{(j)}|\epsilon_0(s), \epsilon_1(s) \right)$ is given by (1).
Maximum Likelihood Inference

In order to find the MLE estimates of $\epsilon$ and $f = \{f_1, \ldots, f_p\}$, $f_p = \{f_{p1}, \ldots, f_{pL}\}$ we will employ an EM algorithm.

**E-step.** The expected complete data likelihood is given by

$$Q(\epsilon, f; \epsilon', f') = E[\ell(\epsilon, f)|g, \epsilon', f']$$

$$= \sum_{p=1}^{P} \sum_{n_p} \sum_{l=1}^{L} \sum_{g=0}^{2} \mathbb{P}(\gamma_{pil} = g|g_{pil}, \epsilon_0(s_{pil}), \epsilon_1(s_{pil}), f'_{pil}) \left\{ \log f_{p0g} + \log \mathbb{P}(g_{pil}|\gamma_{pil} = g, \epsilon_0(s_{pil}), \epsilon_1(s_{pil})) \right\},$$

According to Bayes formula,

$$\mathbb{P}(\gamma_{pil} = g|g_{pil}, \epsilon_0(s_{pil}), \epsilon_1(s_{pil}), f'_{pil}) = \frac{\mathbb{P}(g_{pil}|\gamma_{pil} = g, \epsilon_0(s_{pil}), \epsilon_1(s_{pil})) f'_{pilg}}{\sum_{h=0}^{2} \mathbb{P}(g_{pil}|\gamma_{pil} = h, \epsilon_0(s_{pil}), \epsilon_1(s_{pil})) f'_{plhg}}.$$

**M-step.** We have to maximize $Q$ subject to the constraints

$$\sum_{g=0}^{2} f_{pilg} = f_{p0} + f_{p1} + f_{p2} = 1$$

for all populations $p = 1, \ldots, P$, we form the Lagrangian

$$\mathcal{L}(\epsilon, f, \mu) = Q - \sum_{p=1}^{P} \mu_p \left( \sum_{g=0}^{2} f_{pilg} - 1 \right),$$

where $\mu = \{\mu_1, \ldots, \mu_P\}$ is the vector of Lagrangian multipliers. All $f$ can be maximized analytically. We get the following derivatives of the Lagrangian:

$$\frac{\partial}{\partial \mu_p} \mathcal{L} = \sum_{g=0}^{2} f_{pilg} - 1 = 0$$

$$\frac{\partial}{\partial f_{pilg}} \mathcal{L} = \frac{1}{f_{pilg}} \sum_{i=1}^{n_p} \mathbb{P}(\gamma_{pil} = g|g_{pil}, \epsilon_0(s_{pil}), \epsilon_1(s_{pil}), f'_{pilg}) - \mu_p = 0.$$

From these we get

$$f_{pilg} = \frac{1}{n_p} \sum_{i=1}^{n_p} \mathbb{P}(\gamma_{pil} = g|g_{pil}, \epsilon_0(s_{pil}), \epsilon_1(s_{pil}), f'_{pilg}).$$

In contrast, we will need to maximize $Q$ with respect to each $\epsilon_s$ numerically. For this, we need to maximize

$$Q_{\epsilon_s} = \sum_{p=1}^{P} \sum_{i=1}^{n_p} \sum_{l=1}^{L} \mathcal{I}(s_{pil} = s) \sum_{g=0}^{2} \mathbb{P}(\gamma_{pil} = g|g_{pil}, \epsilon_0(s_{pil}), \epsilon_1(s_{pil}), f'_{pilg}) \left\{ \log f_{p0g} + \log \mathbb{P}(g_{pil}|\gamma_{pil} = g, \epsilon_0(s_{pil}), \epsilon_1(s_{pil})) \right\},$$

where the sums run only over the genotypes of set $s$, as given by the indicator function

$$\mathcal{I}(s_{pil} = s) = \begin{cases} 1 & \text{if } s_{pil} = s, \\ 0 & \text{otherwise.} \end{cases}$$

Bayesian Inference

The MLE estimator introduced above results in an underestimation of the error rates because the allele frequencies are estimated too close to the data, i.e. the MLE of the allele frequencies are biased. Importantly, this problem only disappears when using excessively many samples, but not when using more loci since an allele frequency must be estimated for every locus. To obtain meaningful estimates also for reasonable sample sizes, we develop a Bayesian inference method for the posterior distribution that integrates over the uncertainty of allele frequencies. This results in an unbiased estimation if sufficiently many loci are used (see Figure 1 in the main text).
\[
\mathbb{P}(g|\epsilon, f) = \frac{\mathbb{P}(g|\epsilon, f)\mathbb{P}(\epsilon)\mathbb{P}(f)}{\mathbb{P}(\epsilon, f)},
\]
assuming an exponential prior on \(\epsilon\)

\[
\mathbb{P}(\epsilon) = \prod_{s=1}^{C} \mathbb{P}(\epsilon_0(s))\mathbb{P}(\epsilon_1(s)); \quad \epsilon_0(s), \epsilon_1(s) \sim \text{Exponential}(\lambda)
\]

and the uninformative Jeffrey’s prior on \(\mathbb{P}(f)\)

\[
\mathbb{P}(f) = \prod_{p=1}^{P} \prod_{l=1}^{L} \mathbb{P}(f_{pl}); \quad \mathbb{P}(f_{pl}) \sim \text{Beta}(0.5, 0.5)
\]

Since \(\mathbb{P}(\epsilon, f)\) can not be evaluated analytically, we resort to an MCMC algorithm to generate samples from the posterior distribution \(\mathbb{P}(g|\epsilon, f)\).

**Supplementary Tables**

**Table S1:** Samples included in the RAD vs GBS comparison. The RAD data is available through the SRA Biproject PRJNA528699. Compared genotypes are available at Zenodo through DOI 10.5281/zenodo.2604109.

<table>
<thead>
<tr>
<th>F008_01</th>
<th>F008_15</th>
<th>F009_12</th>
<th>F021_06</th>
<th>F026_06</th>
<th>F030_03</th>
<th>F032_03</th>
<th>F033_05</th>
<th>F039_05</th>
<th>I373_05</th>
</tr>
</thead>
<tbody>
<tr>
<td>F008_02</td>
<td>F008_16</td>
<td>F009_13</td>
<td>F021_07</td>
<td>F026_07</td>
<td>F030_04</td>
<td>F032_04</td>
<td>F033_06</td>
<td>F039_06</td>
<td>I373_06</td>
</tr>
<tr>
<td>F008_03</td>
<td>F008_17</td>
<td>F020_01</td>
<td>F022_01</td>
<td>F026_08</td>
<td>F030_05</td>
<td>F032_05</td>
<td>F033_07</td>
<td>F039_07</td>
<td>I373_07</td>
</tr>
<tr>
<td>F008_04</td>
<td>F009_01</td>
<td>F020_02</td>
<td>F022_02</td>
<td>F026_09</td>
<td>F030_06</td>
<td>F032_06</td>
<td>F036_01</td>
<td>I345_01</td>
<td>I396_01</td>
</tr>
<tr>
<td>F008_05</td>
<td>F009_02</td>
<td>F020_03</td>
<td>F022_03</td>
<td>F026_10</td>
<td>F030_07</td>
<td>F032_08</td>
<td>F036_02</td>
<td>I345_02</td>
<td>I396_02</td>
</tr>
<tr>
<td>F008_06</td>
<td>F009_03</td>
<td>F020_04</td>
<td>F022_04</td>
<td>F026_11</td>
<td>F031_01</td>
<td>F032_09</td>
<td>F036_03</td>
<td>I345_03</td>
<td>I396_03</td>
</tr>
<tr>
<td>F008_07</td>
<td>F009_04</td>
<td>F020_05</td>
<td>F022_05</td>
<td>F026_12</td>
<td>F031_02</td>
<td>F032_10</td>
<td>F036_04</td>
<td>I345_04</td>
<td>I396_04</td>
</tr>
<tr>
<td>F008_08</td>
<td>F009_05</td>
<td>F020_06</td>
<td>F022_06</td>
<td>F026_13</td>
<td>F031_03</td>
<td>F032_11</td>
<td>F036_05</td>
<td>I345_05</td>
<td>I396_05</td>
</tr>
<tr>
<td>F008_09</td>
<td>F009_06</td>
<td>F020_07</td>
<td>F022_07</td>
<td>F026_14</td>
<td>F031_04</td>
<td>F032_12</td>
<td>F036_06</td>
<td>I345_06</td>
<td>I396_06</td>
</tr>
<tr>
<td>F008_10</td>
<td>F009_07</td>
<td>F021_01</td>
<td>F026_01</td>
<td>F026_15</td>
<td>F031_05</td>
<td>F032_13</td>
<td>F036_07</td>
<td>I345_07</td>
<td>I396_07</td>
</tr>
<tr>
<td>F008_11</td>
<td>F009_08</td>
<td>F021_02</td>
<td>F026_02</td>
<td>F026_16</td>
<td>F031_06</td>
<td>F033_01</td>
<td>F039_01</td>
<td>I373_01</td>
<td></td>
</tr>
<tr>
<td>F008_12</td>
<td>F009_09</td>
<td>F021_03</td>
<td>F026_03</td>
<td>F026_17</td>
<td>F031_07</td>
<td>F033_02</td>
<td>F039_02</td>
<td>I373_02</td>
<td></td>
</tr>
<tr>
<td>F008_13</td>
<td>F009_10</td>
<td>F021_04</td>
<td>F026_04</td>
<td>F030_01</td>
<td>F032_01</td>
<td>F033_03</td>
<td>F039_03</td>
<td>I373_03</td>
<td></td>
</tr>
<tr>
<td>F008_14</td>
<td>F009_11</td>
<td>F021_05</td>
<td>F026_05</td>
<td>F030_02</td>
<td>F032_02</td>
<td>F033_04</td>
<td>F039_04</td>
<td>I373_04</td>
<td></td>
</tr>
</tbody>
</table>

6
Supplementary Figures

Figure S1: Accuracy of error rate estimates from two (bottom row), three (middle row) and five (top row) replicates per samples. Shown are the estimated error rates relative to the true error rates (red line) of 100 replicates for different samples sizes $n$ (shown on top), the two error rates 0.1 and 0.01 (first and last two columns, respectively) and different numbers of unlinked loci. The horizontal dashed lines indicate Q2, the interval within which an estimate is less than a factor of two away from the true value.

Figure S2: Distribution of the relative bias when inferring error rates from a truth set generated with next-generation sequencing data with a specific mean depth. Each line indicates a quantile $q$ such that a smaller bias was observed in a fraction $q$ of $10^6$ simulations of $10^4$ homozygous and $10^4$ heterozygous loci. The left two panels are for the case $\epsilon_0 = \epsilon_1 = 0.01$ and the right two panels for $\epsilon_0 = \epsilon_1 = 0.1$. 
Figure S3: Distribution of the allelic depth observed in the RAD-seq data for genotypes found to be homozygous reference (left, 8,701 genotypes), heterozygous (middle, 6,022 genotypes) and homozygous alternative (right, 1,914 genotypes) genotypes in the GBS truth set exploiting familial relationships. Allelic depth distributions are normalized per total depth for plotting (i.e. they sum to 1 per top-left to lower-right diagonal).

Figure S4: Accuracy of error rate estimates under the population sample model if data deviates from Hardy-Weinberg proportion as quantified by the inbreeding coefficient $F$. Shown are the estimated error rates relative to the true error rates (red line) of 100 replicates for different samples sizes $n$ (shown on top), the two error rates 0.1 and 0.01 (first and last two columns, respectively) and different numbers of unlinked loci. The horizontal dashed lines indicate Q2, the interval within which an estimate is less than a factor of two away from the true value.