

Genotyping Error Detection in Samples of Unrelated Individuals without Resampling

Nianjun Liu, Dabao Zhang, and Hongyu Zhao

Supplementary Material

Identifiability Conditions of the Model Parameters

1. Allelic model

The allelic model is defined as following:

$$\varepsilon_1 = \Pr(\text{observed allele is 2} \mid \text{true allele is 1})$$

$$\varepsilon_2 = \Pr(\text{observed allele is 1} \mid \text{true allele is 2})$$

Zou and Zhao showed that with fixed (known) error rates, this model is identifiable for the haplotype frequencies [1]. Otherwise, for single biallelic markers with samples of unrelated individuals, the parameters (allele frequency and error rate) are not identifiable even under the constraint that $\varepsilon_1 = \varepsilon_2$ [1]. If the allele frequencies are known, it is easy to show that the corresponding log-likelihood function attains its maximum

when the pair $(\varepsilon_1, \varepsilon_2)$ satisfies $p_1\varepsilon_1 - p_2\varepsilon_2 = \frac{n_2 + 2n_3}{2n} p_1 - \frac{2n_1 + n_2}{2n} p_2$. In other words, the

error rates $(\varepsilon_1, \varepsilon_2)$ are not identifiable, even with fixed allele frequencies. If $\varepsilon = \varepsilon_1 = \varepsilon_2$,

and allele frequencies are known, the MLE of the error rate can be derived directly:

$$\varepsilon = \frac{\frac{2n_1 + n_2}{2n} - p_1}{1 - 2p_1}.$$

Actually, if we denote $\theta = (1 - \varepsilon_1)p_1 + \varepsilon_2 p_2$, then $E[\frac{n_1}{n}] = \theta^2$, $E[\frac{n_2}{n}] = 2\theta(1 - \theta)$,

$E[\frac{n_3}{n}] = (1 - \theta)^2$, where $E[\cdot]$ indicates expectation. Therefore, this model is only

identifiable for the parameter $\theta = (1 - \varepsilon_1)p_1 + \varepsilon_2(1 - p_1) = \Pr(\text{observed allele is 1})$. In

summary, under the allelic model, the allele frequency or the error rate (assume that

$\varepsilon = \varepsilon_1 = \varepsilon_2$) is identifiable only when the other is fixed (known).

2. Simplified allelic model

The simplified allelic model is obtained from the allelic model by assuming

$\varepsilon = \varepsilon_1 = \varepsilon_2$ and ignoring the higher-order terms of error rate (i.e., keep only linear terms).

Denote $C_1 = \Pr(O = (1\ 1))$, $C_2 = \Pr(O = (1\ 2))$, $C_3 = \Pr(O = (2\ 2))$, and $\Delta_{1-3} = C_1 - C_3$. We

can have the following log-likelihood function:

$$\begin{aligned} l(\varepsilon, p_1) = & \log(n!) - \log(n_1!) - \log(n_2!) - \log(n_3!) + n_1 \cdot \log[p_1^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon] \\ & + n_2 \cdot \log 2[p_1^2\varepsilon + p_1p_2(1 - 2\varepsilon) + p_2^2\varepsilon] + n_3 \cdot \log[p_2^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon], \end{aligned} \quad (\text{A1})$$

and equations

$$\begin{aligned} C_1 = \Pr(O = (1\ 1)) &= p_1^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon, \\ C_2 = \Pr(O = (1\ 2)) &= 2[p_1^2\varepsilon + p_1p_2(1 - 2\varepsilon) + p_2^2\varepsilon], \\ C_3 = \Pr(O = (2\ 2)) &= p_2^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon. \end{aligned} \quad (\text{A2})$$

The MLE of C_1 , C_2 , and C_3 are $\hat{C}_1 = n_1/n$, $\hat{C}_2 = n_2/n$, and $\hat{C}_3 = n_3/n$, respectively. It is

known that if $\hat{\theta}$ is the MLE of θ , then for any function g the MLE of $g(\theta)$ is $g(\hat{\theta})$ [2].

Therefore, we will solve (A2) for (ε, p_1, p_2) as functions of (C_1, C_2, C_3) , which can be

used to induce the MLE of the error rate and allele frequencies. From (A2) we have

$$\Delta_{1-3} = C_1 - C_3 = (1 - 2\varepsilon)(p_1 - p_2) = (1 - 2\varepsilon)(2p_1 - 1),$$

which implies

$$p_1 = \frac{\Delta_{1-3}}{2(1 - 2\varepsilon)} + \frac{1}{2}.$$

Substituting it into the equation of C_3 in (A2), we have

$$(16C_3 + 8\Delta_{1-3} - 4)\varepsilon^2 + [4(\Delta_{1-3} - 1)^2 - 16C_3]\varepsilon + 4C_3 - (\Delta_{1-3} - 1)^2 = 0. \quad (\text{A3})$$

The above quadratic equation of ε can be solved analytically—i.e.,

$$\begin{aligned} \varepsilon &= \frac{-[(\Delta_{1-3} - 1)^2 - 4C_3] \pm \Delta_{1-3} \sqrt{(\Delta_{1-3} - 1)^2 - 4C_3}}{8C_3 + 4(\Delta_{1-3} - 1) + 2} \\ &= \frac{-[(C_1 - C_3 - 1)^2 - 4C_3] \pm (C_1 - C_3) \sqrt{(C_1 - C_3 - 1)^2 - 4C_3}}{4C_1 + 4C_3 - 2}. \end{aligned}$$

Letting τ_1 and τ_2 denote the two solutions of equation (A3), we have

$$\left(\tau_1 - \frac{1}{2}\right) \cdot \left(\tau_2 - \frac{1}{2}\right) = \frac{(C_1 - C_3)^2 (2C_1 + 2C_3 - 1)}{(4C_1 + 4C_3 - 2)^2}. \quad (\text{A4})$$

Though equation (A3) always has (real or complex) solutions, it is necessary that

$(C_1 - C_3 - 1)^2 - 4C_3 \geq 0$ in order to have real solutions. If we add $4 \cdot C_3$ on both sides and

take the square root, we have $1 - C_1 + C_3 \geq 2\sqrt{C_3}$ (since $C_1, C_3 \in [0, 1]$, therefore

$1 - C_1 + C_3 \geq 0$). Consequently $(1 - \sqrt{C_3})^2 \geq C_1$, which results in $\sqrt{C_1} + \sqrt{C_3} \leq 1$. This

means that equation (A3) does not have real solutions above the curve $\sqrt{C_1} + \sqrt{C_3} = 1$.

When $C_1 = C_3$, we have the solution $\tau_1 = \tau_2 = 0.5$. In the following discussion,

we assume that $C_1 \neq C_3$. If $\sqrt{C_1} + \sqrt{C_3} = 1$, then the discriminant of equation (A3) equals

0 and $\tau_1 = \tau_2 = 0$. We further assume that $\sqrt{C_1} + \sqrt{C_3} \neq 1$ (i.e., $\sqrt{C_1} + \sqrt{C_3} < 1$).

We can see that the sign of equation (A4) is determined by the sign of

$(2C_1 + 2C_3 - 1)$. When $C_1 + C_3 < \frac{1}{2}$, we have $(\tau_1 - \frac{1}{2}) \cdot (\tau_2 - \frac{1}{2}) < 0$, which implies that one

of two solutions of ε is larger than $1/2$, and the other is smaller than $1/2$ (and larger than

0). When $C_1 + C_3 > \frac{1}{2}$, we have $4C_1 + 4C_3 - 2 > 0$, which is the denominator of τ_1 and

τ_2 . If $C_1 + C_3 > \frac{1}{2}$ and $C_1 > C_3$ (i.e., $C_1 - C_3 > 0$), then

$$\tau_2 = \frac{-[(C_1 - C_3 - 1)^2 - 4C_3] - (C_1 - C_3)\sqrt{(C_1 - C_3 - 1)^2 - 4C_3}}{4C_1 + 4C_3 - 2} < 0,$$

because the two terms of numerator of τ_2 are all negative. Because ε lies between 0 and

1, τ_2 cannot be an estimate for ε . Therefore, ε has at most one valid solution. After some

algebra, we can also show that in this case ($C_1 + C_3 > \frac{1}{2}$, and $C_1 > C_3$) that

$$2C_1 + 2C_3 - 1 > -[(C_1 - C_3 - 1)^2 - 4C_3] + (C_1 - C_3)\sqrt{(C_1 - C_3 - 1)^2 - 4C_3},$$

and that

$$C_1 - C_3 > \sqrt{(C_1 - C_3 - 1)^2 - 4C_3}.$$

This means that τ_1 is in $(0, \frac{1}{2})$. Since C_1 and C_3 are symmetric, similarly, if $C_1 + C_3 > \frac{1}{2}$

and $C_1 < C_3$ (i.e., $C_1 - C_3 < 0$), then

$$\tau_1 = \frac{-[(C_1 - C_3 - 1)^2 - 4C_3] + (C_1 - C_3)\sqrt{(C_1 - C_3 - 1)^2 - 4C_3}}{4C_1 + 4C_3 - 2} < 0,$$

which cannot be an estimate for ε . Again, ε has at most one valid solution, and τ_2 is in

$(0, \frac{1}{2})$. In summary, the parameter pair (p_1, ε) is identifiable with the constraint that

error rate is between 0 and 0.5; i.e., $\varepsilon \in [0, \frac{1}{2}]$. Figure 2A in the main text shows the

solutions of error rate (as a function of C_1 and C_3) under this model.

3. Homo-heterozygote model

With the same notation as above, we have the log-likelihood function

$$l(\varepsilon, p_1) = \log(n!) - \log(n_1!) - \log(n_2!) - \log(n_3!) + n_1 \cdot \log[p_1^2(1-\varepsilon) + 2p_1p_2\varepsilon] + n_2 \cdot \log[p_1^2\varepsilon + 2p_1p_2(1-2\varepsilon) + p_2^2\varepsilon] + n_3 \cdot \log[p_2^2(1-\varepsilon) + 2p_1p_2\varepsilon], \quad (\text{A5})$$

and equations

$$\begin{aligned} C_1 &= \Pr(O = (1 \ 1)) = p_1^2(1-\varepsilon) + 2p_1p_2\varepsilon, \\ C_2 &= \Pr(O = (1 \ 2)) = p_1^2\varepsilon + 2p_1p_2(1-2\varepsilon) + p_2^2\varepsilon, \\ C_3 &= \Pr(O = (2 \ 2)) = 2p_1p_2\varepsilon + p_2^2(1-\varepsilon). \end{aligned}$$

Following the same procedure as for the above simplified allelic model, we have

$$\varepsilon^3 - (1 + 2\Delta_{1-3})\varepsilon^2 - 4C_3(1-\varepsilon)^2 + (\Delta_{1-3} - 1)(1 - 3\Delta_{1-3})\varepsilon + (\Delta_{1-3} - 1)^2 = 0. \quad (\text{A6})$$

There exists a close form of the solutions to a polynomial equation of third degree, which is called Cardano's formula [3,4]. Although there should be three solutions to equation

(A6), some solutions may not be qualified as error rates (i.e., a real number between 0

and 1). Nickalls (1993) provided conditions under which cubic equations have real

solutions. Figure 2B shows the distribution of real solutions of equation (A6), and Figure

S1 shows some of the solutions that were chosen, as follows: if there is only one solution

for the values of (C_1, C_3) , select that solution; if there are more than one solution and only

one in $[0, 1]$, select the one in $[0, 1]$; if there are more than one solutions in $[0, 1]$, select

the smallest one in $[0, 1]$; if there is no solution in $[0, 1]$, select the one that is closest to

$[0, 1]$. Note that the three clusters correspond to the three regions in Figure 2B. Obviously,

there is an identifiability issue. For example, if $n_1 = 810$, $n_2 = 180$, and $n_3 = 10$, the two solutions to equation (A6) are 0 and $(33 - 12\sqrt{6})/25$. The corresponding values for p_1 are 0.9 and $(12 + 3\sqrt{6})/20 \approx 0.967$. The two sets of parameters give identical genotype distributions, and therefore identical maximum values, $810 \times \log(0.81) + 180 \times \log(0.18) + 10 \times \log(0.01)$, to the log-likelihood function (less a constant of $\log(1000!) - \log(810!) - \log(180!) - \log(10!)$); see Figure 1B in the main text. Hence, the parameter pair (p_1, ε) of this model is not always identifiable, and it is not straightforward how to make the model parameters identifiable by putting constraints on the parameter space.

4. Genotypic model

We have the log-likelihood function under the genotypic model as

$$l(\varepsilon, p_1) = \log(n!) - \log(n_1!) - \log(n_2!) - \log(n_3!) + n_1 \cdot \log[p_1^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon + p_2^2\varepsilon] \\ + n_2 \cdot \log[p_1^2\varepsilon + 2p_1p_2(1 - 2\varepsilon) + p_2^2\varepsilon] + n_3 \cdot \log[p_1^2\varepsilon + 2p_1p_2\varepsilon + p_2^2(1 - 2\varepsilon)]$$

and equations

$$C_1 = \Pr(O = (1 \ 1)) = p_1^2(1 - 2\varepsilon) + 2p_1p_2\varepsilon + p_2^2\varepsilon \\ C_2 = \Pr(O = (1 \ 2)) = p_1^2\varepsilon + 2p_1p_2(1 - 2\varepsilon) + p_2^2\varepsilon \\ C_3 = \Pr(O = (2 \ 2)) = p_1^2\varepsilon + 2p_1p_2\varepsilon + p_2^2(1 - 2\varepsilon).$$

Following the previous procedure, we have

$$3\varepsilon^2 + (2 - 6C_1 - 6C_3)\varepsilon + 4C_1 - (1 + C_1 - C_3)^2 = 0. \quad (\text{A7})$$

The discriminant for the above equation is $\Delta = [6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2$, which is greater than or equal to 0. This means that equation (A7) always has two real solutions,

except that, when $C_1 = C_2 = C_3 = 1/3$, equation (A7) has only a unique solution $\varepsilon = 1/3$.

In general, the solutions of equation (A7) can be expressed as

$$\varepsilon = \frac{2(3C_1 + 3C_3 - 1) \pm \sqrt{\Delta}}{6} = \frac{2(3C_1 + 3C_3 - 1) \pm \sqrt{[6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2}}{6}.$$

It is obvious that there may be more than one solution of ε in $[0, 1]$. For example, if $C_1 = 1/2$, and $C_3 = 1/6$, there are two sets of solutions for parameters (p_1, ε) —i.e.,

$$((3 - \sqrt{3})/6, (1 + 1/\sqrt{3})/3) \text{ and } ((3 + \sqrt{3})/6, (1 - 1/\sqrt{3})/3),$$

which provide identical genotype distributions, and hence identical maximum values of the likelihood function.

Thus, the parameter pair (p_1, ε) is not always identifiable under this error model. Figure

2C in the main text shows the solutions of error rate under the genotypic model.

Letting η_1 and η_2 denote the two solutions of equation (A7), we have

$$\begin{aligned} & (\eta_1 - 1/3) \times (\eta_2 - 1/3) \\ &= \left(\frac{2(3C_1 + 3C_3 - 1) + \sqrt{[6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2}}{6} - 1/3 \right) \\ & \quad \times \left(\frac{2(3C_1 + 3C_3 - 1) - \sqrt{[6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2}}{6} - 1/3 \right) \\ &= -\frac{1}{3}(C_1 - C_3)^2 \leq 0. \end{aligned}$$

The above equality holds if and only if $C_1 = C_3$. It implies that one of the solutions to equation (A7) is greater than or equal to $1/3$, and that the other is smaller than or equal to $1/3$. We can show that

$$\begin{aligned} & [6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2 - [2(3C_1 + 3C_3 - 1)]^2 \\ &= (1 + \sqrt{C_1} + \sqrt{C_3})(1 + \sqrt{C_1} - \sqrt{C_3})(1 - \sqrt{C_1} + \sqrt{C_3})(1 - \sqrt{C_1} - \sqrt{C_3}). \end{aligned}$$

The sign of the above equation is determined by the sign of the last term $(1 - \sqrt{C_1} - \sqrt{C_3})$ because the first three terms are non-negative. Therefore, when $\sqrt{C_1} + \sqrt{C_3} \geq 1$, the above equation is smaller than or equal to 0. This means that

$$2(3C_1 + 3C_3 - 1) \geq \sqrt{[6(C_1 + C_3) - 4]^2 + 12(C_1 - C_3)^2},$$

which results in that $\varepsilon \geq 0$ (i.e., $\eta_1 \geq \frac{1}{3}, 0 \leq \eta_2 \leq \frac{1}{3}$). When $\sqrt{C_1} + \sqrt{C_3} < 1$, $\eta_1 \geq \frac{1}{3}$

(equality holds when $C_1 = C_3$) but $\eta_2 \leq 0$ (equality holds only when C_1 or C_3 is 1).

Therefore, the model parameter pair (p_1, ε) is identifiable when restricting $\varepsilon \in [0, \frac{1}{3}]$ on

the region $\sqrt{C_1} + \sqrt{C_3} \geq 1$. When $\sqrt{C_1} + \sqrt{C_3} < 1$, ε has only one solution in $[0, 1]$ and is

in $[\frac{1}{3}, 1]$. Note that for (C_1, C_3) in the region $(C_1 - C_3)^2 + 4(C_1 + C_3) > 4$, $C_1 + C_3 \leq 1$,

equation (A7) has two real solutions, with one being larger than 1 and the other being in

$[0, \frac{1}{3}]$. Therefore, restricting $\varepsilon \in [\frac{1}{3}, 1]$ does not work on the entire parameter space. In

addition, it is not plausible to have an error rate larger than 1/3 in practice. An interesting

observation is that in the region $\sqrt{C_1} + \sqrt{C_3} < 1$ the likelihood of the simplified allelic

model is always larger than that of the genotypic model.

References:

- 1 Zou G, Zhao H: Haplotype frequency estimation in the presence of genotyping errors. *Hum Hered* 2003;56:131-138.
- 2 Casella G, Berger LR: *Statistical inference* ed 2. Duxbury Press, 2001.
- 3 Nickalls RWD: A new approach to solving the cubic: Cardan's solution revealed. *Mathematical Gazette* 1993;77:354-359.
- 4 Weisstein EW: Cubic formula, *MathWorld-A Wolfram Web Resource*.

Figure Legends

Figure S1

Shown are some of the solutions of equation (A6) as a function of (C_1, C_3) . The solutions were selected as follows: select the solutions that are in $[0, 1]$; if there are more than one solution in $[0, 1]$, select the smallest one in $[0, 1]$; if there are no solutions in $[0, 1]$, select the one that is closest to $[0, 1]$.

