**Appendix**

We note that all notation for this appendix is provided in Table 1 in the main manuscript.

*Component probability mass function for Causal Variant (CV) count data for single locus*

Before proceeding with the log-likelihoods and the test statistics, we compute the binomial probability mass function $Bin\left(x_{1,k}\middle|v_{1,k}^{t}, i_{1,k}^{t}, j_{1}^{t}\right)$ for a single locus. Here, the subscript "1" indicates that it is the "first" locus. The notation becomes important later on, when we consider multiple loci. We comment that determination of the probability mass function is necessary for computation of the log-likelihoods.

Given CV counts, coverage, and phenotype data $\left(x_{1,k}, v_{1,k}^{t}, i_{1,k}^{t}\right)$, we prove that the conditional probability of the CV counts $x_{1,k}$ for individual $k$, given coverage $v_{1,k}^{t}$, phenotype $i_{1,k}^{t}$, and genotype $j_{1}^{t}$ follows a binomial distribution, $Bin\left(x_{1,k}; v_{1,k}^{t}; p\left(i_{1,k}^{t}, j_{1}^{t}\right)\right)$, where $p\left(i_{1,k}^{t}, j_{1}^{t}\right) = \left(\frac{2-j_{1}^{t}}{2}\varepsilon_{i_{1,k}^{t}} + \frac{j_{1}^{t}}{2}\left(1-\varepsilon_{i_{1,k}^{t}}\right)\right)$ is the probability of a "success", namely the probability of observing the causal variant $A$.

To see this, consider the following: for genotype $j_{1}^{t} = 0$ and $j_{1}^{t} = 2$, it is straightforward. For instance, when $j_{1}^{t} = 0$, the individual's genotype is $a/a$. Thus every observed CV *A* is really an *a* that has been misread. The number of errors is $x_{1,k}$, so that $x_{1,k}$ follows a Binomial distribution with probability $\varepsilon_{i_{1,k}^{t}}$ given $v_{1,k}^{t}$ trials. Similarly, when $j_{1}^{t} = 2$, every read of CV *A* is now correct read from the genotype $A/A$. Thus, $x_{1,k}$ follows a Binomial distribution with probability $\left(1-\varepsilon_{i_{1,k}^{t}}\right)$ given $v_{1,k}^{t}$ trials. For the heterozygote genotype $j_{1}^{t} = 1$, the probability of CV $A$ being read on a single trial is the sum of two probabilities:

$$\Pr(a)\Pr(\text{read} = A|\text{ correct variant} = a) + \Pr(A)\Pr(\text{read} = A|\text{correct variant} = A),$$
$$= \frac{1}{2}\varepsilon_{i_{1,k}^t} + \frac{1}{2}\left(1 - \varepsilon_{i_{1,k}^t}\right),$$
$$= \frac{1}{2}.$$

Here, we specify a symmetric error model in which:

$$\Pr(\text{read} = A|\text{ correct variant} = a) = \Pr(\text{read} = a|\text{ correct variant} = A).$$

An asymmetric model, where:

$$\Pr(\text{read} = A|\text{ correct variant} = a) \neq \Pr(\text{read} = a|\text{ correct variant} = A),$$

may be considered. However, the estimation of $\Pr(\text{read} = a|\text{ correct variant} = A)$ is quite unstable due to low frequency of causal variant (even for 1000 samples).

*Log-likelihood of the observed data*

The log-likelihood function of the multiple-locus CV count data involves $3^M$ genotype configurations. If we set $\boldsymbol{x_k} = (x_{1,k}, \cdots, x_{M,k})$, $\boldsymbol{v_k^t} = (v_{1,k}^t, \cdots, v_{M,k}^t)$, $\boldsymbol{G^t} = (j_1^t, \cdots, j_M^t)$, we can write:

$$\ln(L_{\mathrm{H}_d}) = \sum_{k=1}^{N} \ln[\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t)],$$

$$= \sum_{k=1}^{N} \ln\left[\sum_{\boldsymbol{G^t}} \left(\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t, \boldsymbol{G^t})\right)\right],$$

$$= \sum_{k=1}^{N} \ln\left[\sum_{\boldsymbol{G^t}} \left(\Pr(\boldsymbol{x_k}|\boldsymbol{v_k^t}, i_k^t, \boldsymbol{G^t}) \times \Pr(i_k^t|\boldsymbol{v_k^t}, \boldsymbol{G^t}) \times \Pr(\boldsymbol{v_k^t}|\boldsymbol{G^t}) \times \Pr(\boldsymbol{G^t})\right)\right],$$

$$= \sum_{k=1}^{N} \ln\left[\sum_{\boldsymbol{G^t}} \left(\Pr(\boldsymbol{x_k}|\boldsymbol{v_k^t}, i_k^t, \boldsymbol{G^t}) \times \Pr(i_k^t|\boldsymbol{G^t}) \times \Pr(\boldsymbol{v_k^t}|\boldsymbol{G^t}) \times \boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right], \qquad (A1)$$

$$= \sum_{k=1}^{N} (1 - i_k^t) \times \ln\left[\sum_{\boldsymbol{G^t}} \left(\Pr(\boldsymbol{x_k}|\boldsymbol{v_k^t}, i_k^t = 0, \boldsymbol{G^t}) \times \Pr(i_k^t = 0|\boldsymbol{G^t}) \times \boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right]$$

$$+ \sum_{k=1}^{N} (i_k^t) \times \ln\left[\sum_{\boldsymbol{G^t}} \left(\Pr(\boldsymbol{x_k}|\boldsymbol{v_k^t}, i_k^t = 1, \boldsymbol{G^t}) \times \Pr(i_k^t = 1|\boldsymbol{G^t}) \times \boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right]$$

$$+\gamma,$$

where $\gamma = \sum_{k=1}^{N} \ln\left(\Pr(\boldsymbol{v_k^t})\right)$ and $\boldsymbol{g}_{(j_1^t,\cdots,j_M^t)} = \Pr(\boldsymbol{G^t}) = \Pr\left((j_1^t, \cdots, j_M^t)\right)$. The last equality follows from the fact that the phenotypes and genotypes are independent of the coverage.

Summing over all true genotype vectors $\boldsymbol{G^t}$ is equivalent to summing each locus *m*'s genotype value from 0 to 2. In this work, we specify that, conditional on the underlying data (including the genotype vectors $\boldsymbol{G^t}$), the observed CV counts are independent. Written another way, we have:

$$\Pr\left(\boldsymbol{x}_k \middle| i_k^t, \boldsymbol{v_k^t} = \left(v_{1,k}^t, \cdots, v_{M,k}^t\right), \boldsymbol{G^t} = (j_1^t, \cdots, j_M^t)\right)$$
$$= \prod_{m=1}^{M} Bin\left(x_{m,k}; v_{m,k}^t; p\left(i_{1,k}^t, j_m^t\right)\right). \qquad (A2)$$

It follows that we can rewrite the Equation $(A1)$ under the null as:

$$\ln(L_{\mathrm{H}_0}) = \sum_{k=1}^{N}(1-i_k)\ln\left[\sum_{j_1=0}^{2}\cdots\sum_{j_M=0}^{2}\left(\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p(0,j_m^t)\right)\right)\left(f_{0,*}\times\boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right] + \sum_{k=1}^{N}(i_k)\ln\left[\sum_{j_1=0}^{2}\cdots\sum_{j_M=0}^{2}\left(\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p(1,j_m^t)\right)\right)\left(f_{1,*}\times\boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right]+\gamma. \quad (A3)$$

Similarly, for the alternative hypothesis, we have:

$$\ln(L_{\mathrm{H}_1}) = \sum_{k=1}^{N}(1-i_k)\ln\left[\sum_{j_1=0}^{2}\cdots\sum_{j_M=0}^{2}\left(\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p(0,j_m^t)\right)\right)\left(f_{0,(j_1^t,\cdots,j_M^t)}\times\boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right] + \sum_{k=1}^{N}(i_k)\ln\left[\sum_{j_1=0}^{2}\cdots\sum_{j_M=0}^{2}\left(\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p(1,j_m^t)\right)\right)\left(f_{1,(j_1^t,\cdots,j_M^t)}\times\boldsymbol{g}_{(j_1^t,\cdots,j_M^t)}\right)\right]+\gamma. \quad (A4)$$

*EM Algorithm Estimates*

We provide closed formula solutions of the $(r+1)^{\mathrm{st}}$-step estimates of the parameters necessary to compute the log-likelihoods. The main difference is that estimates are determined as a function of a vector of genotypes rather than a single genotype. We determine that the posterior probability that individual *k* has genotype vector $\boldsymbol{G^t}=(j_1^t,\cdots,j_M^t)$ is given by:

$$\tau_{\boldsymbol{G^t},k}^{(r)} = \tau_{(j_1^t,\cdots,j_M^t),k}^{(r)} = \frac{f_{i_k^t,(j_1^t,\cdots,j_M^t)}\Pr\left((j_1^t,\cdots,j_M^t)\right)\left[\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p\left(i_{1,k}^t,j_m^t\right)\right)\right]}{\sum_{s_1^t=0}^{2}\cdots\sum_{s_M^t=0}^{2}\left(f_{i_k^t,(s_1^t,\cdots,s_M^t)}\Pr\left((s_1^t,\cdots,s_M^t)\right)\left[\prod_{m=1}^{M}Bin\left(x_{m,k};v_{m,k}^{t};p\left(i_{1,k}^t,s_m^t\right)\right)\right]\right)}. \quad (A5)$$

For the parameters $\alpha$ and $\beta$, we make use of the following notation:

$$\tau^{(r)}_{(j_1^t,\cdots,j_M^t)+(1)} = \sum_{k=1}^{N} (i_k)\, \tau^{(r)}_{(j_1^t,\cdots,j_M^t),k},$$

$$\tau^{(r)}_{++(1)} = \sum_{k=1}^{N} \sum_{j_1=0}^{2} \sum_{j_2=0}^{2} \dots \sum_{j_M=0}^{2} (i_k)\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k} = N_1,$$

$$\tau^{(r)}_{(j_1^t,\cdots,j_M^t)+(0)} = \sum_{k=1}^{N} (1-i_k)\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k},$$

$$\tau^{(r)}_{++(0)} = \sum_{k=1}^{N} \sum_{j_1=0}^{2} \sum_{j_2=0}^{2} \dots \sum_{j_M=0}^{2} (1-i_k)\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k} = N_0,$$

$$\tau^{(r)}_{(j_1^t,\cdots,j_M^t)+} = \sum_{k=1}^{N} \tau^{(r)}_{(j_1^t,\cdots,j_M^t),k},$$

We determine:

$$\alpha^{(r+1)} = \ln\left(\tau^{(r)}_{(0,\cdots,0)+(1)}\right) - \ln\left(\tau^{(r)}_{(0,\cdots,0)+(0)}\right), \quad (A6)$$

$$\beta^{(r+1)}$$

$$= \ln\left(N_1 - \tau^{(r)}_{(0,\cdots,0)+(1)}\right) - \ln\left(N_0 - \tau^{(r)}_{(0,\cdots,0)+(0)}\right) - \ln\left(\tau^{(r)}_{(0,\cdots,0)+(1)}\right) + \ln\left(\tau^{(r)}_{(0,\cdots,0)+(0)}\right). \quad (A7)$$

Under the null hypothesis, we have:

$$\alpha = \ln(N_1) - \ln(N_0).$$

For the $(r+1)^{\text{st}}$-step estimates of the genotype frequencies , we determine:

$$\boldsymbol{g}^{(\boldsymbol{r})}_{(j_1^t,\cdots,j_M^t)} = \frac{\sum_{k=1}^{N} \tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}}{N}. \quad (A8)$$

Finally, for the $(r+1)^{\text{st}}$-step estimates of the sequence error probabilities, we can show that:

$$\begin{aligned}
\varepsilon_0^{(r+1)} &= \frac{\sum_{k=1}^{N}\left((1-i_k)\left[\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=0}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}x_{m,k}\right)+\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=2}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}\left(v^t_{m,k}-x_{m,k}\right)\right)\right]\right)}{\sum_{k=1}^{N}\left((1-i_k)\left(\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=0}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}v^t_{m,k}\right)+\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=2}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}v^t_{m,k}\right)\right)\right)}, \\
\varepsilon_1^{(r+1)} &= \frac{\sum_{k=1}^{N}\left((i_k)\left[\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=0}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}x_{m,k}\right)+\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=2}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}\left(v^t_{m,k}-x_{m,k}\right)\right)\right]\right)}{\sum_{k=1}^{N}\left((i_k)\left(\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=0}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}v^t_{m,k}\right)+\sum_{m=1}^{M}\sum_{(j_1^t,\cdots,j_M^t),j_m^t=2}\left(\tau^{(r)}_{(j_1^t,\cdots,j_M^t),k}v^t_{m,k}\right)\right)\right)}.
\end{aligned} \quad (A9)$$

We comment that these probabilities are locus-independent, that is, the subscripts do not contain the individual locus number. Having said that, the formulas indicate that the error probabilities are computed as a composite of the individual locus data values ($v^t_{m,k}$ and $x_{m,k}$), suggesting an "average" over all loci.

*Derivation of test statistic*

We use the log-likelihoods for each hypothesis for a given iteration value $r$ to ultimately determine the maximum log-likelihoods under each scenario (Null and Alternative), and then use these maxs to determine the value of the test statistic. We use the notation:

$\ln\left(L_{\mathrm{H}_d}\right) = \sum_{k=1}^{N} \ln[\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t)]$ = Log-likelihood Equation of the observed data.

Note:

$d = 0$ for Null Hypothesis: $(\mathrm{H}_0: \beta = 0)$,

$d = 1$ for Alternative Hypothesis: $(\mathrm{H}_1: \beta \neq 0)$.

$\ln\left(\hat{L}_{\mathrm{H}_d}\right)$ = Maximum log-likelihood of the data for each hypothesis. This maximum is achieved by applying the EM algorithm in the following way:

1. Specify a certain number of starting points (i.e., randomly generated vectors $\vec{\psi}$ of parameter settings for $\alpha, \beta$, etc.).
2. For each vector $\vec{\psi}$ in Item 1, update the log-likelihoods under each hypothesis until some stopping condition is satisfied, such as:

$$\left|(r+1)^{\text{st}}\text{- step of } \ln\left(L_{\mathrm{H}_d}\right) - (r)^{\text{th}}\text{- step of } \ln\left(L_{\mathrm{H}_d}\right)\right| < \delta, \quad (2)$$

for some tolerance $\delta$. In this work, we use $\delta = 0.00001$. The maximum log-likelihood is then the $(r)^{\text{th}}$- step of $\ln\left(L_{\mathrm{H}_d}\right)$. We denote this value by: $\ln\left(L_{\mathrm{H}_d}\right)_{r(\vec{\psi})}$.

a. NOTE: For an arbitrary vector $\vec{\psi}$ in Item 1, if the stopping condition (2) is not met after the maximum number of steps, we define the log-likelihood as: $\ln\left(L_{\mathrm{H}_d}\right)_{r_{\max}(\vec{\psi})}$, where $r_{\max}$ is the total number of steps specified for the EM algorithm. For example, in the Simulation section in the paper, $r_{\max}$= 1000 (Item (xi) in *Simulations*).

3. We define the log-likelihood of the observed data, denoted $\ln\left(\hat{L}_{\mathrm{H}_d}\right)$, as:

$$\ln\left(\hat{L}_{\mathrm{H}_d}\right) = \max_{\vec{\psi}}\left(\ln\left(L_{\mathrm{H}_d}\right)_{r(\vec{\psi})}\right).$$

$LTT_{\mathrm{ae,NGS}} = 2\left[\ln\left(\hat{L}_{\mathrm{H}_1}\right) - \ln\left(\hat{L}_{\mathrm{H}_0}\right)\right]$. As noted above in Item (3), the carat symbol $\widehat{\ }$ indicates that we have obtained the maximum log-likelihood of the data under the particular hypothesis. Note that $LTT_{\mathrm{ae,NGS}}$ is asymptotically a chi-square distribution with 1 degree of freedom. We consider two versions of the $LTT_{\mathrm{ae,NGS}}$ statistic In this work, we allow for differential misclassification in the computation of the $LTT_{\mathrm{ae,NGS}}$ statistic. Specifically, the two error model parameters are unconstrained (that is, it may be that $\varepsilon_0 \neq \varepsilon_1$).

*Example log-likelihood calculations of observed data*

In this appendix, we provide calculations of observed-data log-likelihood calculations, demonstrate how the EM algorithm is implemented, and compute the $LTT_{\mathrm{ae,NGS}}$ statistic. We perform calculations for a scenario where we have 6 cases and 4 controls sequenced at a single locus.

*Example Data Set*

Consider the following set of data:

**Table *A*1. Single locus NGS example data set.**

| **Individual ($k$)** | **Phenotype Code ($i_k^t$)** | **CV counts ($x_{1,k}$)** | **Coverage($v_{1,k}^t$)** |
|---|---|---|---|
| 1 | 1 | 0 | 4 |
| 2 | 1 | 1 | 4 |
| 3 | 1 | 2 | 4 |
| 4 | 1 | 2 | 3 |
| 5 | 1 | 0 | 3 |
| 6 | 1 | 1 | 3 |
| 7 | 0 | 0 | 4 |
| 8 | 0 | 0 | 4 |
| 9 | 0 | 3 | 3 |
| 10 | 0 | 2 | 3 |

Note that there are 6 cases and 4 controls listed in Table *A*1, and the coverage values are either 3 or 4.

To compute the log-likelihood under the null and alternative hypotheses, we must compute the probabilities $\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t)$ (Equation (*A*1)). Recall that:

$$\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t) = \sum_{\boldsymbol{G^t}} \Pr\,(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t, \boldsymbol{G^t}). \quad (A10)$$

Since there is only one locus, we may rewrite the summation over $\boldsymbol{G^t}$ as the sum over $j_1^t$ from 0 to 2 (the number of copies of the CV in the true genotype). Thus, equation $(A10)$ may be rewritten as:

$$\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t) = \sum_{j_1^t=0}^{2} \Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t, j_1^t). \quad (A11)$$

We documented above that:

$$\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t, j_1^t) = \Pr(\boldsymbol{x_k}|\boldsymbol{v_k^t}, i_k^t, j_1^t) \times \Pr(\boldsymbol{v_k^t}) \times \Pr(i_k^t|\, j_1^t) \times \boldsymbol{g}_{j_1^t}. \quad (A12)$$

The terms $\Pr(\boldsymbol{v_k^t})$ are fixed for each individual over all iterations of the EM algorithm and so we treat it as a constant from this point forward. We shall not include it in our calculations moving forward. From the work done above (Equation (*A*3)), including Notation from Table 1, we may rewrite Equation $(A12)$ as:

$$\Pr(\boldsymbol{x_k}, \boldsymbol{v_k^t}, i_k^t, j_1^t) = Bin\left(x_{1,k}; v_{1,k}^t p\left(i_{1,k}^t, j_m^t\right)\right) \times f_{i_k^t, j_1^t} \times \boldsymbol{g}_{j_1^t}. (A13)$$

Equation $(A13)$ is the critical equation for our log-likelihood computations. To compute this probability, we must determine the values $\varepsilon_{i_k^t}$, $f_{i_k^t, j_1^t}$, and $\boldsymbol{g}_{j_1^t}$. All values are unknown and computed via iteration by the EM algorithm. We use the superscript notation $(r)$ to indicate the $r^{\text{th}}$-iteration step estimate of the parameters. The value $r = 0$ indicates the starting value (randomly generated) for each parameter. Note that $f_{i_k^t, j_1^t} = \frac{\left(e^{\alpha + \beta w_{j_1^t}}\right)^{(i_k^t)}}{1 + e^{\alpha + \beta w_{j_1^t}}}$, so estimation of this parameter reduces to estimation of the parameters $\alpha$ and $\beta$. Under the null hypothesis, $\beta = 0$, so that $f_{i_k^t, j_1^t} = \frac{(e^{\alpha})^{(i_k^t)}}{1 + e^{\alpha}}$. We computed above that, under the null hypothesis, $\alpha^{(r)} = \ln(N_1) - \ln(N_0)$; that is, $\alpha^{(r)}$ remains constant over all EM iterations *r*. Under the null, therefore, we can reduce the penetrances $f_{i_k^t, j_1^t}$ to:

$$f_{i_k^t, j_1^t} = f_{i_{k,*}^t} = \frac{(e^{\alpha})^{(i_k^t)}}{1 + e^{\alpha}}. (A14)$$

In other words, under the null hypothesis that $\beta = 0$, the penetrance values depend only on the phenotype code of the individual.

*Log-likelihood of observed data under null hypothesis*

Given the information above, we may now perform the log-likelihood calculations of the observed data. We start by randomly generating starting values for the aforementioned parameters. Let:

$\varepsilon_{i_k^t=0}^{(0)} = 0.073;$

$\varepsilon_{i_k^t=1}^{(0)} = 0.005;$

$g_{j_1^t=0}^{(0)} = 0.316;$

$g_{j_1^t=1}^{(0)} = 0.279;$

$g_{j_1^t=2}^{(0)} = 0.405.$

$$f_{i_k^t=0,j_1^t}^{(0)} = f_{i_k^t=0,*}^{(0)} = \frac{\left(e^{\alpha^t}\right)^{(0)}}{1+e^{\alpha^t}} = \frac{1}{1+e^{\ln(N_1)-\ln(N_0)}} = \frac{1}{1+\frac{N_1}{N_0}} = \frac{N_0}{N} = \frac{4}{10} = 0.40.$$

$$f_{i_k^t=1,j_1^t}^{(0)} = f_{i_k^t=1,*}^{(0)} = \frac{\left(e^{\alpha^t}\right)^{(1)}}{1+e^{\alpha^t}} = \frac{\frac{N_1}{N_0}}{1+\frac{N_1}{N_0}} = \frac{N_1}{N} = \frac{6}{10} = 0.60.$$

(Recall that the superscripts here are the Iteration Steps (*r*). For Individual 1, using Equation $(A3)$ and the values above, we have, at EM Iteration step $r = 0$, the log-likelihood of the observed data is:

$$\ln\left[\Pr\left(x_{1,1}, v_{1,1}^t, i_1^t\right)\right] = \ln\left[\sum_{j_1^t=0}^{2} \Pr(\boldsymbol{x_1}, \boldsymbol{v_1^t}, i_1^t, j_1^t)\right]$$

$$= \ln\left[\sum_{j_1^t=0}^{2} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-j_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{j_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right) \times f_{i_{1,*}^t}^{(0)} g_{j_1^t}^{(0)}\right]. (A15)$$

From Table 1, we know $i_1^t = 1, x_{1,1} = 0,\ v_{1,1}^t = 4$. It follows that the log-likelihood $(A15)$ is equal to:

$$= \ln\left[Bin\left(0; 4, \varepsilon_1^{(0)}\right) \times f_{1,*}^{(0)} g_0^{(0)} + Bin\left(0; 4, \frac{1}{2}\right) \times f_{1,*}^{(0)} g_1^{(0)} + Bin\left(0; 4, \left(1-\varepsilon_1^{(0)}\right)\right) \times f_{1,*}^{(0)} g_2^{(0)}\right],$$

$$= \ln[Bin(0; 4, 0.005) \times (0.60)(0.316) + Bin(0; 4, 0.50) \times (0.60)(0.279) + Bin(0; 4, 0.995) \times (0.60)(0.405)],$$

$$= \ln[(0.980)(0.60)(0.316) + (0.063)(0.60)(0.279) + (0)(0.60)(0.405)],$$

$$= \ln[0.186 + 0.010],$$

$$= -1.628.$$

The log-likelihoods of every other individual are computed in the same fashion. To provide an example of someone with different data than the first individual, we choose Individual 10. This individual is unaffected, has coverage equal to 3, and has a non-zero CV count (2).

From Table 1, we have $i_{10}^t = 0, x_{10,1} = 2,\ v_{10,1}^t = 3$. It follows that the log-likelihood $(A15)$ is equal to:

$$= \ln\left[Bin\left(2; 3, \varepsilon_0^{(0)}\right) \times f_{0,*}^{(0)} g_0^{(0)} + Bin(2; 3, 0.5) \times f_{0,*}^{(0)} g_1^{(0)} + Bin\left(2; 3, \left(1-\varepsilon_0^{(0)}\right)\right) \times f_{0,*}^{(0)} g_2^{(0)}\right],$$

$$= \ln[Bin(2;3,0.073) \times (0.40)(0.316) + Bin(2;3,0.50) \times (0.40)(0.279) + Bin(2;3,0.927) \times (0.40)(0.405)],$$

$$= \ln[(0.015)(0.40)(0.316) + (0.375)(0.40)(0.279) + (0.188)(0.40)(0.405)],$$

$$= \ln[0.002 + 0.042 + 0.030],$$

$$= -2.601.$$

Log-likelihoods for the other individuals may be computed similarly. We present log-likelihood values for all individuals in Table *A*2.

**Table *A*2. Log-likelihood values for all individuals for Iteration Step 0.**

| **Individual ($k$)** | $\ln[\Pr(x_{1,k}, v_{1,k}^t, i_k^t)]$ |
|---|---|
| 1 | -1.628 |
| 2 | -3.088 |
| 3 | -2.767 |
| 4 | -2.712 |
| 5 | -1.572 |
| 6 | -2.724 |
| 7 | -2.299 |
| 8 | -2.299 |
| 9 | -1.945 |
| 10 | -2.601 |
| Total | -23.635 |

As noted in the table, we compute a total log-likelihood of -23.635.

Critically important parameters for the EM-parameter-estimate updates (and consequently, the updated log-likelihoods) are the posterior probabilities $\tau_{j_1^t,k}^{(r)}$. These probabilities are used to determine the updates. To compute the terms, $\tau_{j_1^t,k}^{(r)}$, we use the formula provided above:

$$\tau_{j_1^t,k}^{(r)} = \frac{f_{i_{k^*}^t,*}^{(0)} g_{j_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^t; \left(\frac{2-j_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{j_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)}{\sum_{s_1^t=0}^{2} f_{i_{k^*}^t,*}^{(0)} g_{s_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^t; \left(\frac{2-s_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{s_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)} = \frac{g_{j_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^t; \left(\frac{2-j_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{j_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)}{\sum_{s_1^t=0}^{2} g_{s_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^t; \left(\frac{2-s_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{s_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)}.$$

Let us perform this calculation for Individuals 1 and 10. We compute the singleton terms $f_{i_{k^*}^t,*}^{(0)} g_{s_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^t; \left(\frac{2-s_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{s_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)$ for $s_1^t = 0,1,2$ first. Once we have these values, it is straightforward to compute each of the $\tau_{j_1^t,k}^{(r)}$. For Individual 1, we have:

$$f_{1,*}^{(0)} g_0^{(0)} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-0}{2}\varepsilon_1^{(0)} + \frac{0}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \quad (s_1^t = 0)$$
$$= (0.60)(0.316) \times Bin(0; 4, 0.005)$$
$$= 0.186.$$

$$f_{1,*}^{(0)} g_1^{(0)} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-1}{2}\varepsilon_1^{(0)} + \frac{1}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \quad (s_1^t = 1)$$
$$= (0.60)(0.279) \times Bin(0; 4, 0.5)$$
$$= 0.010.$$

$$f_{1,*}^{(0)} g_2^{(0)} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-2}{2}\varepsilon_1^{(0)} + \frac{2}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \quad (s_1^t = 2)$$
$$= (0.60)(0.405) \times Bin(0; 4, 0.995)$$
$$= 1.52\text{E}-10 \approx 0.$$

The sum of these terms is 0.186 + 0.010 = 0.196.

Thus,

$$\tau_{0,1}^{(0)} = \frac{0.186}{0.196} = 0.947,$$

$$\tau_{1,1}^{(0)} = \frac{0.010}{0.196} = 0.053,$$

$$\tau_{2,1}^{(0)} = \frac{0.000}{0.196} = 0.000.$$

Based on these calculations, we conclude that the most likely genotype code for individual 1 is 0 (from just one iteration!).

For Individual 10, we have:

$$f_{0,*}^{(0)} g_0^{(0)} Bin\left(x_{1,10}; v_{1,10}^t, \left(\frac{2-0}{2}\varepsilon_0^{(0)} + \frac{0}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \qquad (s_1^t = 0)$$
$$= (0.40)(0.316) \times Bin(2; 3, 0.073),$$
$$= 0.002.$$

$$f_{0,*}^{(0)} g_1^{(0)} Bin\left(x_{1,10}; v_{1,10}^t, \left(\frac{2-1}{2}\varepsilon_0^{(0)} + \frac{1}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \qquad (s_1^t = 1)$$
$$= (0.40)(0.279) \times Bin(2; 3, 0.50),$$
$$= 0.042.$$

$$f_{0,*}^{(0)} g_2^{(0)} Bin\left(x_{1,10}; v_{1,10}^t, \left(\frac{2-2}{2}\varepsilon_0^{(0)} + \frac{2}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \qquad (s_1^t = 2)$$
$$= (0.40)(0.405) \times Bin(2; 3, 0.927),$$
$$= 0.030.$$

The sum of these terms is 0.074.

Thus,

$$\tau_{0,10}^{(0)} = \frac{0.002}{0.074} = 0.025,$$

$$\tau_{1,10}^{(0)} = \frac{0.042}{0.074} = 0.564,$$

$\tau_{2,10}^{(0)} = \frac{0.030}{0.074} = 0.411.$

Based on these calculations, we can only reasonably rule out genotype code 0 for Individual 10. The other genotype probabilities are relatively large. The posterior probabilities may be determined in a similar fashion for the other individuals. We report all posterior probabilities in Table *A*3.

**Table *A*3. Posterior probabilities of a given genotype for all individuals in Table *A*1 for Iteration 0.**

| Individual | $\boldsymbol{\tau_{0,k}^{(0)}}$ | $\boldsymbol{\tau_{1,k}^{(0)}}$ | $\boldsymbol{\tau_{2,k}^{(0)}}$ |
|---|---|---|---|
| 1 | 0.947 | 0.053 | 8E-10 |
| 2 | 0.082 | 0.918 | 3E-06 |
| 3 | 0.000 | 0.999 | 0.001 |
| 4 | 0.000 | 0.946 | 0.054 |
| 5 | 0.899 | 0.101 | 1E-07 |
| 6 | 0.043 | 0.957 | 0.000 |
| 7 | 0.930 | 0.070 | 5E-05 |
| 8 | 0.930 | 0.070 | 5E-05 |
| 9 | 0.000 | 0.098 | 0.902 |
| 10 | 0.025 | 0.564 | 0.411 |

Using the formula above, we compute the 1st Iteration Step estimates for the genotype frequencies as: $g_{j_1^t}^{(1)} = \frac{\sum_{k=1}^{N} \tau_{j_1^t,k}^{(0)}}{N}$. To determine, $g_0^{(1)}$, we sum up the first column of Table *A*3 and divide by $N = 10$. We get: $g_0^{(1)} = \frac{3.858}{10} = 0.386$. Similarly, $g_1^{(1)} = \frac{4.774}{10} = 0.477$, $g_2^{(1)} = \frac{1.368}{10} = 0.137$. As a check, we note that the frequencies sum to 1.0.

To compute the error parameter updates, we use the formula (*A*9). In our special case of a single locus,

$$\varepsilon_0^{(r+1)} = \frac{\sum_{k=1}^{N}\left\{(1-i_k)\left(\tau_{0,k}^{(r)}x_{1,k} + \tau_{2,k}^{(r)}\left(v_{1,k}^{t} - x_{1,k}\right)\right)\right\}}{\sum_{k=1}^{N}\left\{(1-i_k)\left(\tau_{0,k}^{(r)} + \tau_{2,k}^{(r)}\right)v_{1,k}^{t}\right\}},$$
$$\varepsilon_1^{(r+1)} = \frac{\sum_{k=1}^{N}\left\{(i_k)\left(\tau_{0,k}^{(r)}x_{1,k} + \tau_{2,k}^{(r)}\left(v_{1,k}^{t} - x_{1,k}\right)\right)\right\}}{\sum_{k=1}^{N}\left\{(i_k)\left(\tau_{0,k}^{(r)} + \tau_{2,k}^{(r)}\right)v_{1,k}^{t}\right\}}. \qquad (A16)$$

In Table *A*4, we compute the numerator and denominators for each of the error-parameter estimates. The interested reader can verify these values. Using the columns total, we compute that the updated error-parameter estimates for Iteration Step $r = 1$ are:

$$\varepsilon_0^{(1)} = \frac{0.4627}{11.4594} = 0.040,$$

$$\varepsilon_1^{(1)} = \frac{0.1822}{7.1097} = 0.026.$$

**Table *A*4. Numerator and denominator terms from formula (*A*16).**

| Individual | Numerator-Term $\left(\varepsilon_0^{(1)}\right)$ | Denominator-Term $\left(\varepsilon_0^{(1)}\right)$ | Numerator-Term $\left(\varepsilon_1^{(1)}\right)$ | Denominator-Term $\left(\varepsilon_1^{(1)}\right)$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 3.1E-09 | 3.787 |
| 2 | 0 | 0 | 0.0820 | 0.3278 |
| 3 | 0 | 0 | 0.0020 | 0.0041 |
| 4 | 0 | 0 | 0.0548 | 0.1637 |
| 5 | 0 | 0 | 4.4E-07 | 2.6978 |
| 6 | 0 | 0 | 0.0435 | 0.1296 |
| 7 | 0.0002 | 3.7219 | 0 | 0 |
| 8 | 0.0002 | 3.7219 | 0 | 0 |
| 9 | 0.0010 | 2.7074 | 0 | 0 |
| 10 | 0.4613 | 1.3082 | 0 | 0 |
| **Total** | **0.4627** | **11.4594** | **0.1822** | **7.1097** |

Using these values and following the same steps listed above (with the updated parameter estimates), we obtain the following log-likelihoods (including the total log-likelihood) for Iteration Step 1.

**Table *A*5. Log-likelihood values for all individuals for Iteration Step 1 (null hypothesis).**

| **Individual** $(\boldsymbol{k})$ | $\mathbf{ln}\left[\mathbf{Pr}\left(\boldsymbol{x}_{\mathbf{1},\boldsymbol{k}}, \boldsymbol{v}_{\mathbf{1},\boldsymbol{k}}^{\boldsymbol{t}}, \boldsymbol{i}_{\boldsymbol{k}}^{\boldsymbol{t}}\right)\right]$ |
|---|---|
| 1 | -1.485 |
| 2 | -2.369 |
| 3 | -2.220 |
| 4 | -2.173 |
| 5 | -1.387 |
| 6 | -2.084 |
| 7 | -1.946 |
| 8 | -1.946 |
| 9 | -2.628 |
| 10 | -2.546 |
| Total | -20.783 |

As noted in the table, we compute a total log-likelihood of -20.783.

Using the inequality **(2)** listed above and the information from the two tables, we have:

$$\left|(1)^{\text{st}}\text{- step of } \ln\left(L_{\text{H}_0}\right) - (0)^{\text{th}}\text{- step of } \ln\left(L_{\text{H}_0}\right)\right| = |-20.783 - (-23.635)| = 2.852.$$

The interested reader can check that it takes up to the $r = 25^{\text{th}}$ Iteration Step before the tolerance inequality is achieve. The $(25)^{\text{st}}$- step of $\ln\left(L_{\text{H}_0}\right)$ is -19.48327932, and $(24)^{\text{th}}$- step of $\ln\left(L_{\text{H}_0}\right)$ is -19.48328661, so the difference is approximately 7.30E-06, less than the tolerance of 1E-05.

It is interesting to note that the final error-parameters for this set of starting points are: $\varepsilon_0^{(25)} = 0.0, \varepsilon_1^{(25)} = 0.224$. This result suggests differential misclassification. Of course, these estimates must be taken "with a grain of salt" since the sample size is very small. Also, the genotype frequency estimates at this step are: $g_0^{(25)} = 0.587, g_1^{(25)} = 0.309, g_2^{(25)} = 0.104$ . We

use these parameter estimates as starting values for the log-likelihood calculations under the alternative hypothesis.

*Log-likelihood of observed data under alternative hypothesis*

The calculations for log-likelihoods under the alternative hypothesis are nearly the same as under the null hypothesis, with the exception that the $\alpha$ and $\beta$ parameters (and hence the penetrances) are updated with each iteration step.

For the 0th Iteration Step under the alternative, we use starting parameters determined by the last iteration step of the null ($r$ = 25). As noted above, we have:

$\varepsilon_0^{(0)} = 0.000;$

$\varepsilon_1^{(0)} = 0.224;$

$g_0^{(0)} = 0.587;$

$g_1^{(0)} = 0.309;$

$g_2^{(0)} = 0.104.$

$f_{0,j_1^t}^{(0)} = 0.40, 0 \leq j_1^t \leq 2.$

$f_{1,j_1^t}^{(0)} = 0.60, 0 \leq j_1^t \leq 2.$

For Individual 1, since the 0th Step penetrances are the same as under the null, we may use Equation $(A3)$ to compute the log-likelihood. We get:

$$\ln\left[Bin\left(0; 4, \varepsilon_1^{(0)}\right) \times f_{1,*}^{(0)} g_0^{(0)} + Bin\left(0; 4, \frac{1}{2}\right) \times f_{1,*}^{(0)} g_1^{(0)} + Bin\left(0; 4, \left(1 - \varepsilon_1^{(0)}\right)\right) \times f_{1,*}^{(0)} g_2^{(0)}\right],$$

$$= \ln[Bin(0; 4, 0.224) \times (0.60)(0.587) + Bin(0; 4,0.50) \times (0.60)(0.309) + Bin(0; 4, 0.776) \times (0.60)(0.104)],$$

$$= \ln[(0.363) \times (0.60)(0.587) + (0.0625) \times (0.60)(0.309) + (0.003) \times (0.60)(0.104)],$$

$$= \ln[0.140],$$

$$= -1.969.$$

As above, we also consider Individual 10.

The log-likelihood is equal to:

$$= \ln\left[Bin\left(2; 3, \varepsilon_0^{(0)}\right) \times f_{0,*}^{(0)} g_0^{(0)} + Bin(2; 3,0.5) \times f_{0,*}^{(0)} g_1^{(0)} + Bin\left(2; 3, \left(1 - \varepsilon_0^{(0)}\right)\right) \times f_{0,*}^{(0)} g_2^{(0)}\right],$$

$$= \ln[Bin(2; 3, 0.000) \times (0.40)(0.587) + Bin(2; 3,0.50) \times (0.40)(0.309) + Bin(2; 3, 1.00) \times (0.40)(0.104)],$$

$$= \ln[(0.375) \times (0.40)(0.309)],$$

$$= \ln[0.046],$$

$$= -3.073.$$

Log-likelihoods for the other individuals may be computed similarly. We present log-likelihood values for all individuals in Table *A*6.

**Table *A*6. Log-likelihood values for all individuals for Iteration Step 0 under alternative hypothesis.**

| Individual ($k$) | $\ln\left[\Pr\left(x_{1,k}, v_{1,k}^{t}, i_{k}^{t}\right)\right]$ |
|---|---|
| 1 | -1.970 |
| 2 | -1.630 |
| 3 | -1.934 |
| 4 | -1.996 |
| 5 | -1.669 |
| 6 | -1.518 |
| 7 | -1.417 |
| 8 | -1.417 |
| 9 | -2.862 |
| 10 | -3.073 |
| **Total** | **-19.483** |

As noted in the table, we compute a total log-likelihood of -19.483.

To compute the terms, $\tau_{j_1^t,k}^{(r)}$, as with the null, we use the formula:

$$\tau_{j_1^t,k}^{(r)} = \frac{f_{i_k^t,j_1^t}^{(0)} g_{j_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^{t}, \left(\frac{2-j_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{j_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)}{\Sigma_{s_1^t=0}^{2}\left(f_{i_k^t,s_1^t}^{(0)} g_{s_1^t}^{(0)} Bin\left(x_{1,1}; v_{1,1}^{t}, \left(\frac{2-s_1^t}{2}\varepsilon_{i_1^t}^{(0)} + \frac{s_1^t}{2}\left(1-\varepsilon_{i_1^t}^{(0)}\right)\right)\right)\right)}.$$

The main difference between this calculation and the one for the null is that the penetrance value $f_{i_k^t,j_1^t}^{(0)}$ now depends upon the genotype.

Let us perform this calculation for Individuals 1 and 10. For Individual 1, we have:

$$f_{1,0}^{(0)} g_0^{(0)} Bin\left(x_{1,1}; v_{1,1}^{t}, \left(\frac{2-0}{2}\varepsilon_1^{(0)} + \frac{0}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \qquad (s_1^t = 0)$$
$$= (0.60)(0.587) \times Bin(0; 4, 0.224)$$
$$= 0.128.$$

$$f_{1,1}^{(0)} g_1^{(0)} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-1}{2}\varepsilon_1^{(0)} + \frac{1}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \quad (s_1^t = 1)$$
$$= (0.60)(0.309) \times Bin(0; 4, 0.5)$$
$$= 0.012.$$

$$f_{1,2}^{(0)} g_2^{(0)} Bin\left(x_{1,1}; v_{1,1}^t, \left(\frac{2-2}{2}\varepsilon_1^{(0)} + \frac{2}{2}\left(1-\varepsilon_1^{(0)}\right)\right)\right) \quad (s_1^t = 2)$$
$$= (0.60)(0.104) \times Bin(0; 4, 0.776)$$
$$= 0.2\text{E}-05.$$

The sum of these terms is 0.128 + 0.012 + 0.00002 = 0.140.

Thus,

$$\tau_{0,1}^{(0)} = \frac{0.128}{0.140} = 0.914,$$

$$\tau_{1,1}^{(0)} = \frac{0.012}{0.140} = 0.086,$$

$$\tau_{2,1}^{(0)} = \frac{0.000}{0.140} = 0.000.$$

Based on these calculations, we conclude that the most likely genotype code for individual 1 is 0.

For Individual 10, we have:

$$f_{0,0}^{(0)} g_0^{(0)} Bin\left(x_{1,10}; v_{1,10}^t; \left(\tfrac{2-0}{2}\varepsilon_0^{(0)} + \tfrac{0}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \quad (s_1^t = 0)$$
$$= (0.40)(0.587) \times Bin(2; 3; 0.000)$$
$$= 0.000.$$

$$f_{0,1}^{(0)} g_1^{(0)} Bin\left(x_{1,10}; v_{1,10}^t; \left(\frac{2-1}{2}\varepsilon_0^{(0)} + \frac{1}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \quad (s_1^t = 1)$$
$$= (0.40)(0.309) \times Bin(2; 3; 0.500)$$
$$= 0.012.$$

$$f_{0,2}^{(0)} g_2^{(0)} Bin\left(x_{1,10}; v_{1,10}^t; \left(\frac{2-2}{2}\varepsilon_0^{(0)} + \frac{2}{2}\left(1-\varepsilon_0^{(0)}\right)\right)\right) \quad (s_1^t = 2)$$
$$= (0.40)(0.104) \times Bin(2; 3; 1.000)$$
$$= 0.000.$$

The sum of these terms is 0.012.

Thus,

$$\tau_{0,10}^{(0)} = \frac{0.000}{0.012} = 0.000,$$

$$\tau_{1,10}^{(0)} = \frac{0.012}{0.012} = 1.000,$$

$$\tau_{2,10}^{(0)} = \frac{0.000}{0.012} = 0.000.$$

These results are consistent with the parameter estimates. Note that the error estimate for controls is $\varepsilon_0^{(0)} = 0.000$. Because the observed number of variants is 2, we know with certainty that the genotype must be a heterozygote. When there is no sequence error, it is impossible for either homozygote to have any number of counts other than 0 or $v_{m,k}^t$.

**Table *A7*. Posterior probabilities of a given genotype for all individuals in Table 1 for Iteration 0 (alternative hypothesis).**

| Individual | $\boldsymbol{\tau_{0,k}^{(0)}}$ | $\boldsymbol{\tau_{1,k}^{(0)}}$ | $\boldsymbol{\tau_{2,k}^{(0)}}$ |
|---|---|---|---|
| 1 | 0.916 | 0.083 | 0.001 |
| 2 | 0.753 | 0.236 | 0.011 |
| 3 | 0.441 | 0.480 | 0.078 |
| 4 | 0.303 | 0.511 | 0.186 |
| 5 | 0.873 | 0.122 | 0.004 |
| 6 | 0.650 | 0.317 | 0.033 |
| 7 | 0.968 | 0.032 | 0.000 |
| 8 | 0.968 | 0.032 | 0.000 |
| 9 | 0.000 | 0.270 | 0.730 |
| 10 | 0.000 | 1.000 | 0.000 |

We compute the 1st Iteration Step estimates for the genotype frequencies as: $g_{j_1^t}^{t(1)} = \frac{\sum_{k=1}^{N} \tau_{j_1^t,k}^{(0)}}{N}$. As in the null situation, we sum up the first column of Table *A*7 and divide by *N* = 10. We get:

$g_0^{(1)} = \frac{5.872}{10} = 0.587$. Similarly, $g_1^{(1)} = \frac{3.084}{10} = 0.308$, $g_2^{(1)} = \frac{1.0441}{10} = 0.104$.

To compute the error parameter updates, we use the formula (*A*16). The interested reader can verify that the component values necessary to compute the updated error parameter estimates are the ones provided in Table *A*8 . Using the columns total, we compute that the updated error-parameter estimates for Iteration Step $r = 1$ are:

$$\varepsilon_0^{(1)} = \frac{0.000}{9.936} = 0.000,$$

$$\varepsilon_1^{(1)} = \frac{3.350}{14.951} = 0.224.$$

One can check that the log-likelihood tolerance condition is met at Iteration Step *r* = 53. At that step, the maximum log-likelihood under the alternative is -18.8506. Thus, the value of the test statistic is $LTT_{\text{ae,NGS}} = 2\ln[-18.851 - (-19.483)] = 1.265$. The corresponding p-value using a central chi-square distribution as the null distribution is 0.261.

**Table *A*8. Numerator and denominator terms from formula (*A*16).**

| Individual | Numerator-Term $\left(\varepsilon_0^{(1)}\right)$ | Denominator-Term $\left(\varepsilon_0^{(1)}\right)$ | Numerator-Term $\left(\varepsilon_1^{(1)}\right)$ | Denominator-Term $\left(\varepsilon_1^{(1)}\right)$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 0.005 | 3.668 |
| 2 | 0 | 0 | 0.786 | 3.055 |
| 3 | 0 | 0 | 1.040 | 2.079 |
| 4 | 0 | 0 | 0.792 | 1.467 |
| 5 | 0 | 0 | 0.011 | 2.632 |
| 6 | 0 | 0 | 0.717 | 2.050 |
| 7 | 0 | 3.873 | 0 | 0 |
| 8 | 0 | 3.873 | 0 | 0 |
| 9 | 0 | 2.190 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 |
| **Total** | **0** | **9.936** | **3.350** | **14.951** |

*Genotype probability calculations for SKAT and CMAT*

The genotype probabilities are calculated from the posterior probabilities using Bayes' Rule.

Given CV count data $x_k = (x_{1,k}, \dots, x_{M,k})$, the genotype probabilities of are obtained as follows.

At each indexed site $m$ $(m = 1,2, \dots, M)$ , we have:

$$\Pr\left(G_{m,k} = j \middle| x_{m,k}, i_k^t, v_{m,k}^t\right) = \frac{\Pr\left(G_{m,k} = j\right) \Pr\left(x_{m,k} \middle| G_{m,k} = j, i_k^t, v_{m,k}^t\right)}{\sum_{l=0}^{2} \Pr\left(G_{m,k} = l\right) \Pr\left(x_{m,k} \middle| G_{m,k} = l, i_k^t, v_{m,k}^t\right)}. (A17)$$

We may include the sequence error parameter $\varepsilon$ in equation ($A$17) by rewriting it as:

$$\Pr\left(G_{m,k} = j \middle| x_{m,k}, i_k^t, v_{m,k}^t\right) = \frac{\Pr\left(G_{m,k} = j\right) \Pr\left(x_{m,k} \middle| G_{m,k} = j, \varepsilon_{i_k^t}, v_{m,k}^t\right)}{\sum_{l=0}^{2} \Pr\left(G_{m,k} = l\right) \Pr\left(x_{m,k} \middle| G_{m,k} = l, \varepsilon_{i_k^t}, v_{m,k}^t\right)}. (A18)$$

Since the equation $(A18)$ requires the probability $\Pr\left(G_{m,k} = j\right)$ and the sequence error parameters $\varepsilon_{i_k^t}$, we estimate these values internally using the $LTT_{\text{ae,NGS}}$ statistic.