Asymptotic Variances of QTL Estimators With Selective DNA Pooling

C. Carleos, J. A. Baro, J. Cañón, and N. Corral

From the Departamento de Estadística, Facultad de Ciencias, C/Calvo Sotelo 33007 Oviedo, Asturias, Spain (Carleos); Departamento de Ciencias Agroforestales, Universidad de Valladolid, Palencia, Spain (Baro); Departamento de Producción Animal, Universidad Complutense, Madrid, Spain (Cañón); and Departamento de Estadística, Universidad de Oviedo, Oviedo, Spain (Corral). This work was supported by the European Regional Development Fund project 1FD97-0042.

Address correspondence to Carlos Carleos at the address above, or e-mail: carleos@pinon.ccu.uniovi.es.

Methods

We consider a QTL \( Q \) (with alleles \( q \) and \( Q \)) flanked by two markers \( (M \) and \( N \)), each with two alleles \( (M, m \) and \( N, n \), respectively). A half-sib family design is considered, where the common sire has haplotypes \( MN/MN \), while not sharing marker alleles with its mates (backcross-like). The family size is \( n \). The recombination fraction between the markers is \( \theta \). The recombination rate between \( M \) and \( Q \) is \( \theta_M \). Progeny receiving the \( Q \) allele (respectively, the \( q \) allele) from the sire has a phenotypic distribution following an \( \mathcal{N}(\mu_Q, \sigma) \) (respectively, \( \mathcal{N}(\mu_q, \sigma) \)). The parameter of interest is \( \alpha = \mu_Q - \mu_q \). Within the common framework of selective DNA pooling (Darvasi and Soller 1994), \( \alpha \) is experimentally determined by mixing DNA samples from either the top (upper tail) or worst (lower tail) performing animals for a given trait. Further, marker frequencies are determined within each tail. Information on the overall distribution is assumed sufficient to suppose the grand mean \( \mu \) and \( \sigma \) known. Algebraic approximations to the variances of the estimators of \( \theta_M \) and \( \alpha \) will be given.

Let \( p_{UM} \) denote the frequency of progeny in the upper tail that received the allele \( M \), that is, \( \text{PR}[M|U] \). The rest of the subscripts have analogous meanings. The observed proportion, corresponding to \( p_{UM} \) is \( p_{\text{tot}} \). The estimator of \( \theta_M \) based on a single tail (Dekkers 2000) is

\[
\hat{\theta}_M = \frac{\hat{p}_{UM} - \hat{p}_{UQ}}{1 - 2\hat{p}_{UQ}}
\]

where

\[
\hat{p}_{UQ} = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{(1 - 2p_{UM})(1 - 2p_{UN})}{1 - 2p_{UM}}}
\]

Let \( \mu_U \) and \( \mu_L \) be the means of phenotypes in the upper and the lower tails, respectively. Regarding \( \alpha \), if

\[
\hat{\mu}_{LL} = \frac{\hat{p}_{UL} (\hat{\mu}_L - \mu) + \hat{p}_{UL} (\hat{\mu}_U - \mu)}{\hat{p}_{UL} + \hat{p}_{UL}}
\]

The half-sib design is particularly well suited to animal genetics, for both laboratory and breeding species (Georges et al. 1995; Weller et al. 1990). DNA pooling techniques applied to selected samples allow direct estimation of marker allele frequencies within the best and worst performing animals of a class of half-sib progeny; this allows great savings in terms of genotyping and data collection compared to individual genotype determination. This method does have its drawbacks, such as loss of information about joint marker inheritance—allelic frequencies are known but genotypic frequencies are not. Additionally, imprecise values stemming from technical error constitute an additional source of inaccuracy (Lipkin et al. 1998).

Darvasi and Soller (1994) showed how DNA pooling can be combined with selection to find association between a marker and a QTL using a backcross, an \( F_2 \), or a half-sib family. Dekkers (2000) extended this method to consider two-marker (interval) mapping and to allow the estimation of the QTL position.
then an estimator for the other allele, \(q\), is
\[
\hat{\alpha} = \frac{\hat{\mu}_{U|Q} - \hat{\mu}_{U|L}}{\hat{\rho}^2}
\] (3)
where \(\hat{\rho}\) is the selection intensity associated with a tail of size \(s\) (Falconer 1989).

**Distribution of the Position Estimator**

As seen in the Appendix, the \(\delta\)-method (Bishop et al. 1975) states that \(\theta_M\) is approximated by a normal distribution with mean \(\hat{\theta}_M\) and variance-covariance matrix
\[
\text{Var}(\hat{\theta}_M) = G' \cdot \text{Var}(\hat{\alpha}_M) \cdot G
\] (4)
where
\[
G = \frac{\partial \hat{\theta}_M}{\partial \hat{\theta}_M}
\]
and \(\hat{\alpha}_M\) is the observed allele frequencies, being its variance \(\text{Var}(\hat{\alpha}_M)\) as shown in equation (9).

**Distribution of the Effect Estimator**

The distribution of phenotypes did not affect the position estimation, except through the probability of QTL alleles in the tails. For the estimation of \(\alpha\), we assume normal distribution of phenotypes in both groups: offspring inheriting the \(Q\) allele, and offspring inheriting the \(q\) allele. Some additional notation is described:

- \(\mu_U\) and \(\sigma_U^2\) are the mean and the variance of the phenotypes above the \(u\) threshold; it is easily seen that for an overall normal distribution \(N(\mu, \sigma)\),
\[
\mu_U = \mu + \sigma \frac{\Phi(u)}{1 - \Phi(u)}
\]
and \((\text{Johnson and Kotz 1970})
\[
\sigma_U^2 = \sigma^2 \left[ 1 + \frac{\Phi(u')u'}{1 - \Phi(u')} - \left( \frac{\Phi(u')}{1 - \Phi(u')} \right)^2 \right]
\]
where \(u' = (u - \mu)/\sigma\). In our case, the phenotypic distribution is a mixture of normals, so these expressions should be altered accordingly.

- \(\mu_{U|Q}\) is the mean of offspring above the \(u\) threshold inheriting the \(q\) allele.

Analogous definitions apply for \(L\) and \(q\) subscripts.

Let the unobserved sample consist of \((X_i, Y_i)\) values, with \(i \in \{1, \ldots, n\}\), where \(X_i\) represents the parental QTL allele of individual \(i\)
\[
X_i = \begin{cases} 
0, & q \\
1, & Q 
\end{cases}, \quad X_i \sim B\left(1, \frac{1}{2}\right)
\]
and \(Y_i\) is the continuous phenotype with conditional Gaussian distribution:
\[
Y_i \mid X_i \sim N\left(\mu + \alpha \left[X_i - \frac{1}{2}\right], \sigma\right).
\]

Let us define:

- the indicator
\[
I_i = \mathcal{I}_{[3/4, \infty)}(Y_i)
\]
signaling whether the individual \(i\) is in the upper tail; analogously, \(I_i\);

- the tail size
\[
n_U = \sum_{i=1}^n I_i
\]
the number of individuals in the upper tail; analogously, \(n_L\);

- the estimator of the tail mean
\[
\hat{\mu}_U = \frac{\sum_{I_i = 1}^n Y_i}{n_U}, \quad n_U > 0
\]
and similarly for the variance (11):
\[
\text{Var}(\hat{\mu}_U) = \frac{\sum_{I_i = 1}^n (Y_i - \hat{\mu}_U)^2}{n_U}
\]
actually estimated by \(\hat{p}_{UQ}\) (2); let \(\hat{p}_{UQ} = 1 - \hat{p}_{LQ}\)
analogously for the lower tail.

The covariance matrix \(\mathbf{J} := \text{Var}(\hat{\mu}_L; \hat{\mu}_U; \hat{p}_{UQ}; \hat{p}_{LQ})\) is computed taking into account:

- \(\text{Var}(\hat{\mu}_L)\): it is seen (10):
\[
E(\hat{\mu}_L) = \mu
\]
and similarly for the variance (11):
\[
\text{Var}(\hat{\mu}_L) = \sigma_L^2 \frac{1}{n_L}
\]
where the notation \(E[1/n_L]\) implies substituting zero for the inverse of \(n_L\); in the highly unlikely case of an empty upper tail (\(n_U = 0\)) adequate bounds are computed as shown in the Appendix \((p\) stands for \(p_{UQ})\); lower bound (Equation 12)
\[
E\left[ \frac{1}{n_U} \right] \geq \frac{1}{p(n+1)} \left[ 1 - (n+1)p(n+1) - (1-p)^{n+1} \right]
\]
upper bound (Equation 13)
\[
E\left[ \frac{1}{n_U} \right] \leq \frac{1}{p(n+1)} E\left[ \frac{n_U + 1}{n_U} \right]
\]
and according to Lynch and Walsh (1998, p. 818)
\[
E\left( \frac{n_U + 1}{n_U} \right) \approx \frac{n_U + 1}{n_U} \left( 1 + \frac{n - n_U}{nn_U(n_U + 1)} \right)
\]
Note that this result is not adequate to directly approximate \(E(1/n_U)\).
It can be concluded from our results that estimation of the \( \theta_M \) variance is accurate for marker brackets wider than 20 cM. Shorter intervals lead to distribution of the position estimator not holding within the parameter space (corresponding to the intermarker gap); thus, tail values agglomerate at boundaries. The presented formulas can be used to compute the probability of erroneously locating a QTL exactly at a marker position.

It was also noted that the proposed approximations degrade when the effect \( \alpha \) exceeds one standard deviation. A likely explanation is the departure from normality of the overall phenotypic distribution when the \( Q, g \) mixture components are too separated.

All of the limitations mentioned so far are related to the single fact that variances of estimators are computed making use of asymptotic theory, which notably relies on regularity conditions.

This study did not address issues such as influence of phase for small QTL effects, and narrow marker intervals, on normality of asymptotic distribution of the effect and position estimators. These deserve further study.

Large samples, above 5,000 half-sibs, are required for the proposed formulas to achieve fair results. Study of small sample distributions must account for lack of normality, and inferences should no longer rely on asymptotic theory.

Exploration of the behavior of the estimators under small-sample scenarios requires additional research. A novel approach based on resampling methods is being developed by the authors (Carleos et al. 2002).

### Appendix

#### Distribution of the Position Estimator

Assuming fixed selection thresholds, the unobserved absolute genotypic frequencies follow a multinomial distribution:

\[
\text{Var}(\hat{\theta}_M) \approx \frac{p_{\mu}(1 - p_{\mu})}{n_{\mu}}
\]

\[
\text{Cov}(\hat{\theta}_M, \hat{\theta}_S) = \frac{p_{\mu}(\mu_{\mu} - \mu_S)}{n_{\mu}}
\]

\[
\text{Cov}(\mu_M, \hat{\mu}_S) = \frac{p_{\mu}(\mu_{\mu} - \mu_S)}{n_{\mu}}
\]

Eventually:

\[
J = \begin{bmatrix}
\frac{\sigma_\mu^2}{n_\mu} & 0 & \frac{\mu_{\mu}(\mu_{\mu} - \mu_S)}{n_\mu} & 0 \\
0 & \frac{\sigma_\mu^2}{n_S} & 0 & \frac{\mu_S(\mu_{\mu} - \mu_S)}{n_S} \\
\frac{\mu_{\mu}(\mu_{\mu} - \mu_S)}{n_\mu} & 0 & \frac{\mu_{\mu}(1 - \mu_{\mu})}{n_\mu} & 0 \\
0 & \frac{\mu_S(1 - \mu_S)}{n_S} & 0 & \frac{\mu_S(1 - \mu_{\mu})}{n_S}
\end{bmatrix}
\]

The matrix of partial derivatives of the estimator (3) of \( \alpha \) with respect to (\( \mu_M, \mu_S, \mu_{\mu}, p_{\mu} \)) is:

\[
G := \frac{\partial \alpha}{\partial (\mu_M, \mu_S, \mu, p_{\mu}, p_{\mu})}
\]

\[
= \begin{bmatrix}
\frac{\mu_{\mu}(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_\mu} & \frac{\mu_{\mu}(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_\mu} \\
\frac{\mu_{\mu}(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_\mu} & \frac{\mu_S(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_S} \\
\frac{\mu_S(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_S} & \frac{\mu_S(\mu_{\mu} - \mu_S)(1 - \mu_{\mu})}{n_S}
\end{bmatrix}
\]

therefore, after application of the \( \delta \)-method,

\[
\text{Var}(\hat{\alpha}) \approx G' \cdot J \cdot G.
\]

#### Table 1. Variances of position and effect estimators.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Position estimator</th>
<th>Effect estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>Predicted</td>
<td>Simulation</td>
</tr>
<tr>
<td>0.25</td>
<td>0.064</td>
<td>0.064</td>
</tr>
<tr>
<td>0.50</td>
<td>0.032</td>
<td>0.033</td>
</tr>
<tr>
<td>1.00</td>
<td>0.017</td>
<td>0.018</td>
</tr>
<tr>
<td>2.00</td>
<td>0.012</td>
<td>0.012</td>
</tr>
</tbody>
</table>

\( n \) is the sample size.
\[ \hat{\theta}_M = \left( \hat{n}_{LMN}, \hat{n}_{LMM}, \hat{n}_{LMN}, \hat{n}_{LMM}, \hat{n}_{LMN}, \hat{n}_{UMN}, \hat{n}_{UMN}, \hat{n}_{UMN} \right) \]

\[ \rightarrow \mathcal{E}(\theta, \left[ p_{LMN} p_{LMM} p_{LMN} p_{UMN} p_{UMN} p_{UMN} p_{UMN} p_{UMN} \right]) \]

\[ \begin{pmatrix} \hat{n}_{LMN} + \hat{n}_{LMM} + \hat{n}_{LMN} + \hat{n}_{LMM} + \hat{n}_{LMN} + \hat{n}_{UMN} + \hat{n}_{UMN} + \hat{n}_{UMN} \end{pmatrix} = A \cdot \hat{\theta}_M \]  

where counts \( \hat{\theta}_M \) carry subscripts indicating tail and allele, and with \( A \), the matrix that relates observed and unobserved absolute frequencies, being

\[ A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \]

so

\[ \text{Var}(\hat{\theta}_M) = A \cdot \text{Var}(\hat{\theta}_M) \cdot A' \]

The estimator (1) of \( \theta_M \) averaged over the two tails, is rewritten as

\[ \hat{\theta}_M = \frac{1}{2} - \frac{1}{4} \sqrt{1 - 20 \left[ \frac{2\hat{n}_{LMN} - \hat{n}_U}{2\hat{n}_{UMN} - \hat{n}_U} + \frac{2\hat{n}_{LMN} - \hat{n}_L}{2\hat{n}_{UMN} - \hat{n}_L} \right] } \]

The multinomial frequencies (7) can be approximated by a normal distribution. The observed frequencies (8) are a linear transformation of those, so they are asymptotically normal. The estimator \( \hat{\theta}_M \) is a nonlinear function of the observed frequencies, defined on an open subset and differentiable at their expected values, \( E[\theta_M] = AE[\theta_M] \). Should these conditions hold, the \( \delta \)-method states that \( \hat{\theta}_M \) is approximated by a normal distribution with mean \( \theta_M \) and variance-covariance matrix

\[ \text{Var}(\hat{\theta}_M) \approx G' \cdot \text{Var}(\hat{\theta}_M) \cdot G \]

where

\[ G = \frac{\partial \theta_M}{\partial \hat{\theta}_M} \]

and \( \text{Var}(\hat{\theta}_M) \) is (9). \( G \) can be estimated by:

\[ E(\hat{\mu}_U) = \frac{1}{2} \left[ \left( \sum_{I_l} Y_{l} \right) - \mu_U \right] \approx \left( \sum_{I_l} Y_{l} \right) \]

\[ \text{Var}(\hat{\mu}_U) = \left[ \left( \sum_{I_l} Y_{l} \right) - \mu_U \right] \approx \left( \sum_{I_l} Y_{l} \right) \]

Appropriate bounding values for the expectation constituting the last factor, \( E[1/\hat{\theta}_M] \) in (11), can be determined as follows (\( \hat{p} \) denotes \( p_{UL} \)):
lower bound

\[
\begin{align*}
E\left[ \frac{1}{n_U} \right] &= \sum_{k=1}^{n} \frac{1}{k} \frac{1}{k!(n-k)!} \rho^k (1 - \rho)^{n-k} \\
&\geq \sum_{k=1}^{n} \frac{1}{1} \frac{1}{k!(n-k)!} \rho^k (1 - \rho)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(n-k)!k^n(1-\rho)^k} \\
&= \frac{1}{\rho(n+1)} \sum_{k=1}^{n+1} \frac{(n+1)!}{(n+1-k)!}\rho^k (1 - \rho)^{n-k} \\
&= \frac{1}{\rho(n+1)} [1 - (n+1)\rho^1 (1 - \rho) - (1 - \rho)^{n+1}] \\
&= \frac{1}{\rho(n+1)} \left[ \frac{n+1}{n} \right] \\
\end{align*}
\]

upper bound

\[
\begin{align*}
E\left[ \frac{1}{n_U} \right] &= \sum_{k=1}^{n} \frac{1}{k} \frac{1}{k!(n-k)!} \rho^k (1 - \rho)^{n-k} \\
&\leq \sum_{k=1}^{n} \frac{1}{k+1} \frac{1}{k!(n-k)!} \rho^k (1 - \rho)^{n-k} \\
&= \frac{1}{\rho(n+1)} \sum_{k=1}^{n+1} \frac{k+1}{k} \frac{(n+1)!}{(n+1-k)!} \rho^k (1 - \rho)^{n-k} \\
&\times (1 - \rho)^{n-k} \\
&= \frac{1}{\rho(n+1)} \left[ \frac{n+1}{n} \right] \\
\end{align*}
\]