On Burrows coefficients

Consider two loci, say, locus 1 and locus 2. Let A be an allele at locus 1, and B be an allele at locus 2. Let S be the set of sampled individuals, where N is assumed to be the number of individuals having *full data at both loci.* Let x_A denote frequency of allele A, h_A denote the frequency of homozygote AA. Similarly, y_B, k_B are frequencies of allele B and homozygote BB at locus 2.

For an individual $s \in S$, let $[s]_1$ and $[s]_2$ denote the genotypes of individual s at locus 1 and locus 2, respectively. Let

$$X_A(s) \;=\; \begin{cases} 0 & \text{if } [s]_1 = oo, \\ 1 & \text{if } [s]_1 = Ao, \\ 2 & \text{if } [s]_1 = AA, \end{cases} \qquad Y_B(s) \;=\; \begin{cases} 0 & \text{if } [s]_2 = oo, \\ 1 & \text{if } [s]_2 = Bo, \\ 2 & \text{if } [s]_2 = BB. \end{cases} \tag{1}$$

(Letter 'o' stands for any allele different from A at locus 1 and from B at locus 2.) Then the Burrows disequilibrium is (corresponding to allele A at locus 1, allele B at locus 2)

$$\Delta_{(A,B)} \;=\; \frac{1}{2N} \sum_{s\in S} X_A(s) Y_B(s) \;-\; 2 x_A y_B. \tag{2}$$

The expected values of X_A, Y_B are

$$E\big[X_A\big] \;=\; \frac{1}{N}\sum_{s\in S} X_A(s) \;=\; 2x_A, \qquad\qquad E\big[Y_B\big] \;=\; \frac{1}{N}\sum_{s\in S} Y_B(s) \;=\; 2y_B.$$

Thus,

$$\Delta_{(A,B)} \;=\; \tfrac{1}{2}\big(E\big[X_A Y_B\big] \;-\; E\big[X_A\big]E\big[Y_B\big]\big) \;=\; \tfrac{1}{2}\,\mathrm{cov}\,(X_A, Y_B). \tag{3}$$

Now, at locus 1, if we let

$$H_A(s) \;=\; \begin{cases} 1 & \text{if } [s]_1 = AA, \\ 0 & \text{otherwise,} \end{cases}$$

then $E\,[H_A] = h_A$ and $X_A^2 \;=\; X_A \;+\; 2H_A$. Thus,

$$\begin{aligned} \sigma_X^2 := \mathrm{Var}\,(X_A) \;&=\; E\left[X_A^2\right] \;-\; E\,[X_A]^2 \;=\; E\,[X_A] \;+\; 2\,E\,[H_A] \;-\; E\,[X_A]^2 \\ &=\; 2\,(x_A + h_A) \;-\; (2x_A)^2 \;=\; 2\left(x_A - 2x_A^2 + h_A\right). \end{aligned} \tag{4}$$

Similarly, at locus 2,

$$\sigma_Y^2 := \mathrm{Var}\,(Y_B) \;=\; 2\left(y_B - 2y_B^2 + k_B\right). \tag{5}$$

The Burrows correlation coefficient is

$$r_{(A,B)} \;=\; \frac{\Delta_{(A,B)}}{\sqrt{x_A - 2x_A^2 + h_A} \cdot \sqrt{y_B - 2y_B^2 + k_B}}. \tag{6}$$

From (3), (4), and (5), the expression for r in (6) can be written as

$$r_{(A,B)} \;=\; \frac{\frac{1}{2}\,\mathrm{cov}\,(X_A, Y_B)}{\sqrt{(1/2)\,\sigma_X^2} \cdot \sqrt{(1/2)\,\sigma_Y^2}} \;=\; \frac{\mathrm{cov}\,(X_A, Y_B)}{\sigma_X\,\sigma_Y} \;=\; \mathrm{corr}\,(X_A, Y_B). \tag{7}$$

Remark. The right side defining r in (6) is undefined when the denominator is zero, which happens when either $x_A - 2x_A^2 + h_A = \frac{1}{2}\mathrm{Var}\,(X_A) = 0$ or $y_B - 2y_B^2 + k_B = \frac{1}{2}\mathrm{Var}\,(Y_B) = 0$. Variance of a random variable is zero only when it is a constant. Thus, if the denominator is zero, then either X_A or Y_B must be a constant throughout. Since X_A, Y_B can only take values $0, 1, 2$, the following are the only possibilities that result in zero denominator:

- $X_A \equiv 1$: Locus 1 is heterozygote Ao throughout ($x_A = \frac{1}{2},\ h_A = 0$).
- $X_A \equiv 2$: Locus 1 is homozygote AA throughout ($x_A = 1,\ h_A = 1$).
- $Y_B \equiv 1$: Locus 2 is heterozygote Bo throughout ($y_B = \frac{1}{2},\ k_B = 0$).
- $Y_B \equiv 2$: Locus 2 is homozygote BB throughout ($y_B = 1,\ k_B = 1$).

(Here, we already assume allele A exists at locus 1 and allele B exists at locus 2, so X_A and Y_B are not identically zero.)

In the LD program, only loci that have at least two alleles will be considered; so the homozygosity conditions above will not occur. Therefore, zero denominator only happens in either of the following:

- Locus 1 is heterozygote Ao throughout ($x_A = \frac{1}{2},\ h_A = 0$).
- Locus 2 is heterozygote Bo throughout ($y_B = \frac{1}{2},\ k_B = 0$).

When either $\mathrm{Var}\,(X_A)$ or $\mathrm{Var}\,(Y_B)$ is zero, the covariance $\mathrm{cov}\,(X_A, Y_B)$ is also zero. Then, the Burrows correlation coefficient $r_{(A,B)}$ is set to zero.

Dependency of Burrows Coefficients

Suppose $A_1, \ldots, A_m$ are all alleles at locus 1, and $B_1, \ldots, B_n$ are all alleles at locus 2. Let $X_i = X_{A_i},\ Y_j = Y_{B_j}$ be defined as in (1) $(i = 1, \ldots, m;\ j = 1, \ldots, n)$. Then

$$X_1 \ + \ \cdots \ + X_m \ = \ 2, \qquad Y_1 \ + \ \cdots \ + Y_n \ = \ 2. \tag{8}$$

From

$$\sum_{i=1}^{m} \operatorname{cov}(X_i, Y_j) \ = \ \operatorname{cov}\left(\sum_{i=1}^{m} X_i,\ Y_j\right) \ = \ \operatorname{cov}(2, Y_j) = 0, \qquad \text{for } j = 1, \ldots, n,$$

and

$$\sum_{j=1}^{n} \operatorname{cov}(X_i, Y_j) \ = \ \operatorname{cov}\left(X_i,\ \sum_{j=1}^{n} Y_j\right) \ = \ \operatorname{cov}(X_i, 2) = 0, \qquad \text{for } i = 1, \ldots, m,$$

we have, with $\Delta_{(i,j)} \equiv \Delta_{(A_i,B_j)} = \frac{1}{2}\operatorname{cov}(X_i, Y_j)$,

$$\left.\begin{aligned} \sum_{i=1}^{m} \Delta_{(i,j)} &= 0 \qquad \text{for } j = 1, \ldots, n, \\ \sum_{j=1}^{n} \Delta_{(i,j)} &= 0 \qquad \text{for } i = 1, \ldots, m. \end{aligned}\right\} \tag{9}$$

This implies that

> **Fact 1.** *The sum of all Burrows disequilibrium coefficients taken over all allele pairs at the two loci is zero.*

Let Δ be the $(m \times n)$-matrix whose entries are $\Delta_{(i,j)}$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$. Then from (9), each entry $\Delta_{(p,q)}$ $(1 \leq p \leq m,\ 1 \leq q \leq n)$ can be deduced from other entries of the same row or of the same column:

$$- \sum_{\substack{i\,=\,1, \\ i \neq p}}^{m} \Delta_{(i,q)} \ = \ \Delta_{(p,q)} \ = \ - \sum_{\substack{j\,=\,1, \\ j \neq q}}^{n} \Delta_{(p,j)} \tag{10}$$

(The first equality comes from the first equation in (9) where j is replaced by q. The second equality comes from the second equation where i is replaced by p.) Then, starting with the second expression above,

$$\Delta_{(p,q)} \ = \ - \sum_{\substack{j\,=\,1, \\ j \neq q}}^{n} \Delta_{(p,j)} \ = \ - \sum_{\substack{j\,=\,1, \\ j \neq q}}^{n} \Big(- \sum_{\substack{i\,=\,1, \\ i \neq p}}^{m} \Delta_{(i,j)} \Big) \ = \ \sum_{\substack{j\,=\,1, \\ j \neq q}}^{n} \ \sum_{\substack{i\,=\,1, \\ i \neq p}}^{m} \Delta_{(i,j)}.$$

The rightmost sum is the sum of all entries $\Delta_{(i,j)}$ outside of row p and column q. Those are Burrows disequilibrium coefficients taken at pairs of alleles (A_i, B_j) where $A_i \neq A_p$ and $B_j \neq B_q$. Thus,

> **Fact 2.** *The Burrows disequilibrium coefficient at a pair of alleles (A, B) is the sum of coefficients taken at all pairs distinct from (A, B).*

In the case $m = 2$, i.e., locus 1 has exactly 2 alleles A_1, A_2. Then

$$\mathrm{Var}\,(X_2) \;=\; \mathrm{Var}\,(-X_1 + 2) \;=\; (-1)^2\,\mathrm{Var}\,(X_1) \;=\; \mathrm{Var}\,(X_1).$$

(Here, the identity $\mathrm{Var}\,(aX + b) = a^2\mathrm{Var}\,(X)$ is used.) Combining with the first equation in (9) where $m = 2$, we have for any allele B_j at locus 2 $(j = 1, \ldots, n)$,

$$r_{(1,j)} + r_{(2,j)} = 0 \quad \text{or} \quad r_{(2,j)} = -\, r_{(1,j)} \qquad \text{for } \ j = 1, \ldots, n. \tag{11}$$

(Here, $r_{(i,j)} = \dfrac{2\Delta_{(i,j)}}{\sqrt{\mathrm{Var}\,(X_i)\mathrm{Var}\,(Y_{B_j})}}$ is the Burrows correlation coefficient at pair (A_i, B_j).)

Then

$$\sum_{i=1}^{2}\sum_{j=1}^{n} r_{(i,j)} \;=\; 0.$$

> **Fact 3.** *If one locus has exactly 2 alleles, then only half of Burrows correlation coefficients need to be calculated, the other half is of opposite sign. The sum of all coefficients is zero.*

In the case that locus 2 also has exactly 2 alleles $(n = 2)$, then

$$r_{(1,2)} = -\, r_{(1,1)} = r_{(2,1)} = -r_{(2,2)} \quad \Rightarrow \quad r^2_{(1,1)} \;=\; r^2_{(1,2)} \;=\; r^2_{(2,1)} \;=\; r^2_{(2,2)}.$$

> **Fact 4.** *If each locus has exactly 2 alleles, then the square of Burrows correlation coefficient is the same for all 4 allele pairs taken at the two loci.*

Remark.

(1) In the LD program, for a pair of loci, *only individuals having data at both loci will be considered. The frequencies of alleles and homozygotes are calculated against those individuals.* Therefore, all the results in the previous discussion hold true.

(2) Facts 3 and 4 are related to the comparisons of the denominators in Burrows correlations, which involve the variances. However, when those variances are zeroes, the Burrows correlation r is undefined as a ratio (the ratio is in the form $0/0$). Unless the pair of alleles is rejected, it needs to be assigned to a certain value, which will have some effect on the overall value for r^2.

- If it is assigned to be zero, just as the correlation between the two random variables when one has zero variance, then these facts still hold. This is what the LD program follows. This assignment has the effect of *decreasing* the overall r^2 for the pair of loci.
- If it is assigned to be a nonzero value, e.g., as one of the extreme values: -1 or 1, then Fact 4 still holds, but Fact 3 does not. These extreme values yield $r^2 = 1$, so the assignment has the effect of *increasing* the overall r^2 for the pair of loci.

Frequency Restriction Case

Let $0 < c < \frac{1}{2}$. In LD program, only loci where there is *at least one allele having frequency at least* c and there is *no allele with frequency surpassing* $(1 - c)$, will be considered in the pairings. For a pair of such loci, frequencies of alleles are recalculated against individuals having data at both loci, and only those alleles satisfying frequency conditions as said above will be accepted. As a result, some pair of loci may be rejected if one locus contains no allele satisfying frequency conditions (even though each of them was accepted at the first step).

Suppose there are m alleles at locus 1 and n alleles at locus 2. Among those, m^* alleles at locus 1 and n^* alleles at locus 2 meet the frequency conditions. It is possible that $m^* = 1$ or $n^* = 1$ (e.g., if $c = 0.1$ and there are 3 alleles at locus 1 whose frequencies are 0.06, 0.06, and 0.88, then $m^* = 1$). Let $A_1, \ldots, A_{m^*}$ be alleles at locus 1 and $B_1, \ldots, B_{n^*}$ be alleles at locus 2 that satisfy frequency conditions. Let Δ be the $(m^* \times n^*)$-matrix whose entries are $\Delta_{(i,j)}$ corresponding to pairs (A_i, B_j) $(i = 1, \ldots, m^*,\ j = 1, \ldots, n^*)$. The following are possibilities.

(1) No allele is dropped at either locus, i.e., $m^* = m, n^* = n$. Then all of the above conclusions still hold.

(2) One locus has no allele being dropped, and there are some alleles being dropped at the other, i.e., either $(m^* = m, n^* < n)$ or $(m^* < m, n^* = n)$, exclusively.

(i) Case $m^* = m$, $n^* < n$.
Since $A_1, \ldots, A_m$ are all alleles at locus 1, we have the first equation in (8) holds, so does the first equation in (9). However, the second equation in (9) may not hold since $Y_{B_1} + \cdots + Y_{B_{n^*}}$ is not a constant. The sum of entries in a row is still zero, but the sum of entries in a colum may not. As a result, only the first equality in (10) holds (here, we replace p in (10) by m):

$$\Delta_{(m,q)} = -\sum_{i=1}^{m-1} \Delta_{(i,q)}, \qquad q = 1, \ldots, n^*.$$

There are $(m-1) \times n^*$ coefficients $\Delta_{(i,q)}$, $(i = 1, \ldots, m-1,\ q = 1, \ldots, n^*)$ need to be calculated; and only n^* coefficients can be deduced.

Now, we evaluate Facts **1**–**4** where the coefficients are calculated on eligible alleles at the two loci, i.e., Δ is an $(m^* \times n^*)$-matrix (instead of $m \times n$). Since $m^* = m$, the row sum is zero for every row of Δ, so Fact **1** still holds; however, Fact **2** may not. Fact **3** will hold if $m^* = 2$. Fact **4** may not hold with $m^* = n^* = 2$ since $n > n^*$.

(ii) Case $m^* < m$, $n^* = n$.

Only the second equality in (10) holds (replace q in (10) by $n = n^*$):

$$\Delta_{(p,n)} \;=\; -\sum_{j=1}^{n-1} \Delta_{(p,j)}, \qquad p = 1, \ldots, m^*.$$

As in (i), Fact **1** still holds, but Fact **2** may not. Fact **3** still holds if $n^* = 2$. Fact **4** may not hold with $m^* = n^* = 2$.

(3) Both loci have some alleles being dropped. Then Burrows disequilibrium coefficients need to be calculated for all pairs.

In general, if $\overline{m}$ is the number of independent alleles at locus 1 and $\overline{n}$ is the number of independent alleles at locus 2, then the number of Burrows disequilibrium coefficients that need to be calculated is $\overline{m} \times \overline{n}$. For locus 1, $\overline{m} = m - 1$ if $m^* = m$, and $\overline{m} = m^*$ if $m^* < m$. Similarly, $\overline{n} = \min\{n^*,\ n-1\}$.

In LD program, the product of numbers of independent alleles, $\overline{m} \times \overline{n}$, *is used as a weight of the locus pair for the calculation of the weighted average of r^2 across pairs of loci.* As noted above, this is the number of Burrows disequilibrium coefficient that need to be calculated; the rest can be deduced.

A particular case in frequency restriction is made such that *singleton alleles will be rejected.* That is, c is set such that an allele is not a singleton if and only if its frequency is at least c. With N being the number of individuals having data at both loci, c can be any number satisfying

$$1/2N \;<\; c \;\leq\; 1/N.$$

When $N = 1$ (then $1/2 < c < 1$), S has only one individual having full data, this individual is either homozygote or contains 2 singleton alleles at each locus, then the locus pair will be rejected. We can take, for example, $c = 1/(2N-1)$.

Remark.

In LD code, the Burrows disequilibrium Δ is adjusted by the number of individuals, i.e., Δ in formula (2) is replaced by

$$\tilde{\Delta}_{(A,B)} \;=\; \frac{S}{S-1} \cdot \Delta_{(A,B)} \;=\; \frac{S}{S-1} \cdot \tfrac{1}{2}\,\mathrm{cov}\,(X_A, Y_B)$$

whenever $S > 1$. Then the Burrows correlation coefficient in the code is based on the adjusted value $\tilde{\Delta}_{(A,B)}$ of Δ. Thus, in terms of $r_{(A,B)}$ as given in (6) and (7), the adjusted correlation coefficient is

$$\tilde{r}_{(A,B)} \;=\; \frac{S}{S-1} \cdot r_{(A,B)} \;=\; \frac{S}{S-1} \cdot \mathrm{corr}\,(X_A, Y_B).$$

Since the difference between the original coefficient and the adjusted one is a multiplicative constant, all the results (which involve only sums of coefficients) can be applied to the adjusted coefficients. To increase efficiency when the adjusted coefficients for allele pairs are not needed for other uses (e.g. output to a file), the adjusted factor can be skipped in the Delta function, but introduced after the averages of the coefficients Δ, r being taken. In such case, the factor for the average of r^2 should be $\big[S/(S-1)\big]^2$.