File S3

Variance of Bin Genotype Indicator Variable

1. Backcross (BC) population

Let A_1A_1 and A_2A_2 be the genotypes of two inbred lines (P_1 and P_2) used to initiate the cross of interest. The F_1 hybrid has a genotype of A_1A_2. A backcross population generated by $F_1 \times P_1$ is denoted by mating type $A_1A_2 \times A_1A_1$. The BC population contains two possible genotypes, A_1A_1 and A_1A_2. Let us denote the genotype indicator variable of individual *j* at locus τ by $Z_j(\tau)$, which is numerically coded by

$$Z_j(\tau) = \begin{cases} 1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \end{cases} \tag{21}$$

Under Mendelian segregation, the two genotypes have an equal probability. The variance of $Z_j(\tau)$ across all individuals within this BC population is

$$\text{var}\left[Z_j(\tau)\right] = \text{E}\left[Z_j^2(\tau)\right] - \text{E}^2\left[Z_j(\tau)\right] = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \tag{22}$$

The covariance of Z_j between loci τ and ω across all individuals within the BC family is

$$\text{cov}\left[Z_j(\tau), Z_j(\omega)\right] = (1-2r_{\tau\omega})\sqrt{\text{var}\left[Z_j(\tau)\right]\text{var}\left[Z_j(\omega)\right]} \tag{23}$$

where

$$r_{\tau\omega} = \frac{1}{2}(1-e^{-2|\tau-\omega|}) \tag{24}$$

is the recombination fraction between loci τ and ω (Haldane 1919) and $1-2r_{\tau\omega}$ is the correlation between the two loci. The distance between the two loci $|\tau-\omega|$ is measured in Morgan. Because the variance of $Z_j(\tau)$ is a constant (1/4), equation (23) can be reformulated as

$$\text{cov}\left[Z_j(\tau), Z_j(\omega)\right] = \frac{1}{4}(1-2r_{\tau\omega}) = \frac{1}{4}e^{-2|\tau-\omega|} \tag{25}$$

where $\tau \in \Delta_k$ and $\omega \in \Delta_k$, and Δ_k is the size of the *k*th bin in the genome. Let us define the average of $Z_j(\tau)$ for all loci within this bin (with bin size Δ_k) for individual *j* (only this individual, not across all individuals with the family) by

$$\bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k}\int_{\lambda_k-\frac{1}{2}\Delta_k}^{\lambda_k+\frac{1}{2}\Delta_k} Z_j(\tau)d\tau \tag{26}$$

This is an observed data point and determined by the breakpoints within bin *k* for this particular individual. The upper and lower limits of the above integral may be redefined by subtracting $\lambda_k - \frac{1}{2}\Delta_k$ from them, leading to a simple expression,

$$\bar{Z}_j(\Delta_k) = \frac{1}{\Delta_k}\int_0^{\Delta_k} Z_j(\tau)d\tau \tag{27}$$

It should be noted that the genome location τ here has been redefined as a number relative to this bin. The variance of $\bar{Z}_j(\Delta_k)$ across all individuals within the BC population is defined as

$$\begin{aligned}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] &= \frac{1}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\Delta_k}\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right]d\omega d\tau \\ &= \frac{1}{4\Delta_k^2}\int_0^{\Delta_k}\int_0^{\Delta_k} e^{-2|\tau-\omega|}d\omega d\tau \\ &= \frac{1}{2\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau\end{aligned} \tag{28}$$

The double integration has a closed form expression, and so does the variance. The closed form variance is

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{29}$$

which is a function of the bin size Δ_k. One can verify that

$$\begin{cases}\lim\limits_{\Delta_k\to 0}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \lim\limits_{\Delta_k\to 0}\dfrac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) = \dfrac{1}{4} \\ \lim\limits_{\Delta_k\to \infty}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \lim\limits_{\Delta_k\to \infty}\dfrac{1}{8\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) = 0\end{cases} \tag{30}$$

The situation where $\Delta_k \to 0$ is equivalent to single marker and thus the variance is $1/4$.

2. Double haploid (DH) population

A double haploid population is generated by duplicating the gametes of the F_1 individuals derived from the cross of two inbred lines (A_1A_1 and A_2A_2). As a result, the DH population contains two possible genotypes, A_1A_1 and A_2A_2. The genotype indicator variable of individual *j* at locus τ is coded by

$$Z_j(\tau) = \begin{cases}+1 & \text{for } A_1A_1 \\ -1 & \text{for } A_2A_2\end{cases} \tag{31}$$

Under Mendelian segregation, the two genotypes have an equal probability. The variance of $Z_j(\tau)$ across all individuals within this DH population is

$$\mathrm{var}\left[Z_j(\tau)\right] = \mathrm{E}\left[Z_j^2(\tau)\right] - \mathrm{E}^2\left[Z_j(\tau)\right] = 1 \tag{32}$$

The covariance of genotype indicator variables between two different loci within the same bin is

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2r_{\tau\omega} = e^{-2|\tau-\omega|} \tag{33}$$

Using the same approach as that described in the BC population, we obtained

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{2}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau = \frac{1}{2\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{34}$$

The limits of the variance are $\lim_{\Delta_k \to 0} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1$ and $\lim_{\Delta_k \to \infty} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

3. F_2 population

A F_2 population is generated by selfing the F_1 individuals derived from the cross of two inbred lines, A_1A_1 and A_2A_2. As a result, The F_2 family contains three possible genotypes, A_1A_1, A_1A_2 and A_2A_2. Let us denote the genotype indicator variable of individual *j* at locus τ by $Z_j(\tau)$, which is coded by

$$Z_j(\tau) = \begin{cases} +1 & \text{for } A_1A_1 \quad \text{with probability } 1/4 \\ 0 & \text{for } A_1A_2 \quad \text{with probability } 1/2 \\ -1 & \text{for } A_2A_2 \quad \text{with probability } 1/4 \end{cases} \tag{35}$$

The variance of $Z_j(\tau)$ across all individuals within this F_2 population is

$$\mathrm{var}\left[Z_j(\tau)\right] = \mathrm{E}\left[Z_j^2(\tau)\right] - \mathrm{E}^2\left[Z_j(\tau)\right] = \frac{1}{2} \tag{36}$$

The covariance between the genotype indicator variables of two different loci is

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = \frac{1}{2}(1 - 2r_{\tau\omega}) = \frac{1}{2}e^{-2|\tau-\omega|} \tag{37}$$

Using the same approach as that described before, we obtained

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} e^{-2(\tau-\omega)}d\omega d\tau = \frac{1}{4\Delta_k^2}(e^{-2\Delta_k} + 2\Delta_k - 1) \tag{38}$$

The limits of the variance are $\lim_{\Delta_k \to 0} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1/2$ and $\lim_{\Delta_k \to \infty} \mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

4. Recombinant inbred lines (RIL)

There are two ways to generate recombinant inbred lines: (1) repeated selfings starting from the F_1 individual for multiple generations until all genes are fixed and (2) repeated brother-sister matings for multiple generations until all genes are fixed. The RIL generated from selfing is called RIL1 and that generated from brother-sister mating is called RIL2. The two genotypes in the RIL population are A_1A_1 and A_2A_2. The genotype indicator variable of individual *j* at locus τ is coded by

$$Z_j(\tau) = \begin{cases} +1 & \text{for } A_1A_1 \\ -1 & \text{for } A_2A_2 \end{cases} \tag{39}$$

Under Mendelian segregation, the two genotypes have an equal probability. The variance of $Z_j(\tau)$ across all individuals within the RIL population is

$$\mathrm{var}\left[Z_j(\tau)\right] = \mathrm{E}\left[Z_j^2(\tau)\right] - \mathrm{E}^2\left[Z_j(\tau)\right] = 1 \tag{40}$$

which is the same as the DH population. The covariance of the genotype indicator variables between two different loci within the same bin is

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2R_{\tau\omega} \tag{41}$$

where

$$R_{\tau\omega} = \frac{ar_{\tau\omega}}{1 + br_{\tau\omega}} \tag{42}$$

The constant numbers, *a* and *b*, depend on the type of RIL with $a = b = 2$ for RIL1 and $a = 4$ and $b = 6$ for RIL2 (Haldane and Waddington 1931). Substituting $R_{\tau\omega}$ in equation (41) by equation (42) and replacing $r_{\tau\omega}$ by the Haldane map function (Haldane 1919) yields

$$\mathrm{cov}\left[Z_j(\tau), Z_j(\omega)\right] = 1 - 2R_{\tau\omega} = \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}} \tag{43}$$

which is a function of the distance between the two loci measured in Morgan. The variance of $\bar{Z}_j(\Delta_k)$ is defined by

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{2}{\Delta_k^2}\int_0^{\Delta_k}\int_0^{\tau} \frac{2 + b - 2a + (2a - b)e^{-2|\tau - \omega|}}{2 + b - be^{-2|\tau - \omega|}} d\omega d\tau \tag{44}$$

A closed form expression is

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{b(b+2)\Delta_k^2}\left[\Theta_3(e^{-2\Delta_k}) + \Theta_2(\Delta_k^2) + \Theta_1(\Delta_k) + \Theta_0\right] \tag{45}$$

where

$$\begin{aligned}
\Theta_0 &= 2a\xi_2\left(\frac{2}{b+2}\right) + 2\ln(2)a\ln\left(\frac{b}{b+2}\right) \\
\Theta_1(\Delta_k) &= -4\ln(2)a\Delta_k \\
\Theta_2(\Delta_k^2) &= -(2a - b - 2)b\Delta_k^2 \\
\Theta_3(e^{-2\Delta_k}) &= -2a\xi_2\left(1 - \frac{b}{b+2}e^{-2\Delta_k}\right) - 2a\ln\left(\frac{b}{b+2}e^{-2\Delta_k}\right)\ln\left(b + 2 - be^{-2\Delta_k}\right)
\end{aligned} \tag{46}$$

and $\xi_2(x)$ is the dilogarithm function defined as

$$\xi_2(x) = \sum_{i=1}^{\infty}\frac{x^i}{i^2} \tag{47}$$

The limits of the variance are $\lim_{\Delta_k \to 0}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 1$ and $\lim_{\Delta_k \to \infty}\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = 0$.

For the RIL1 population where $a = b = 2$, the variance is simplified into

$$\mathrm{var}\left[\bar{Z}_j(\Delta_k)\right] = \frac{1}{2\Delta_k^2}\left\{\xi_2\left(\frac{1}{2}e^{-2\Delta_k}\right) + 2\ln(2)\Delta_k + \frac{1}{12}\left[6\left(\ln(2)\right)^2 - \pi^2\right]\right\} \tag{48}$$

Simplification of the variance for RIL2 does not help too much except that $\Theta_2(\Delta_k^2) = 0$. Figure S3 shows the function graphically (for RIL1) in the range of $0 \le \Delta_k \le 2$. For example, when $\Delta_k = 0.01$ Morgan (1 cM), the variance is $\mathrm{var}(Z_k) = 0.9868$. When

$\Delta_k = 0.25$ Morgan (25 cM), the variance is $\text{var}(Z_k) = 0.7548$, where the new notation appearing in the figure $\text{var}(Z_k)$ is $\text{var}\left[\bar{Z}_j(\Delta_k)\right]$, which is too complicated to draw in the figure.

All derivations were conducted using Mathematica (Wolfram 1999). The final forms of the equations were then subject to manual simplification.

References

Haldane JB, Waddington CH. 1931. Inbreeding and linkage. *Genetics* **16**(4): 357-374.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**: 299-309.

Wolfram S. 1999. *The Mathematica Book*. Wlfram Media/Cambridge University Press, Oxfordshire, UK.