Notes on Dominance and the Q_X Test

Kyelee Ruth Fitts and Jeremy Berg

So our task is to expess the Q_X statistic in terms of the alleles frequencies, the homozygous effect, and the dominance deviation. In equation 12 of Berg and Coop (2014), we showed that, in the case where all populations are equally distant from one another (with that distance measured in terms of the parameter F_{ST}), the statistic could be written in terms of the *average* effect as

$$Q_X = \frac{1}{V_A F_{ST}} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{\ell\prime=1}^{L} \alpha_\ell \alpha_{\ell'} \left(p_{m\ell} - \overline{p}_\ell\right) \left(p_{m\ell\prime} - \overline{p}_{\ell\prime}\right) \tag{1}$$

where α_ℓ and $\alpha_{\ell\prime}$ are the average effects for site ℓ and $\ell\prime$ respectively, $p_{m\ell}$ is the allele frequency at site ℓ in population m, and $\overline{p}_\ell$ is the frequency in the ancestral population before it split into the M present day populations (and similar for $\ell\prime$). I've also rewritten in a slightly different way than I did in the Berg and Coop (2014). I've rewritten the sum over ℓ and $\ell\prime$ as two distinct sums for the sake of clarity, rather than the shorthand I used in the paper to collapse them into one. And for now I'm assuming that we know the ancestral allele frequency, rather than taking the mean, which is why the sum over m goes from 1 to M rather than $M - 1$.

Now, if we assume that the average effects are estimated in population 1, then given the allele frequency in population 1 ($p_{1\ell}$), the difference between homozygotes (A_ℓ), and the dominance deviation (D_ℓ) the average effect is given by

$$\alpha_\ell = \frac{1}{2} A_\ell + D_\ell \left(1 - 2p_{1\ell}\right). \tag{2}$$

What we want is to reexpress eqn (1) by substituting eqn (2) in for the average effects and then simplifying and partitioning in a way that provides insight. Specifically, we'll want to partition into components corresponding to the population in which the GWAS was done (i.e. population 1) and the others, as we expect the bias should come from the fact that $p_{1\ell}$ appears in the expression for the average effect (it'd probably be easiest to first consider a case with just two populations). Nested within that partition among populations, we'll also want to partition into an F_{ST}-like or variance term, which should correspond to a sum over terms where $\ell = \ell\prime$, as well as an LD-like or covariance term, which should correspond to a sum over terms where $\ell \neq \ell\prime$. Where possible, you'll want to turn sums into expectations, and expectations into variances or covariances. I supsect you're familiar with these, but just to be sure, you will probably want to make extensive use of the following identities

$$N\mathbb{E}[X] = \sum_{i=1}^{N} X_i \tag{3}$$

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{4}$$

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{5}$$

and you may find a use for some of the expressions in Bohrnstedt and Goldberger (1969) regarding the covariance of products of random variables (or not, I'm not sure). Write your algebra out below in LaTeX. You don't need to include every single piece of algebra, but make sure there's enough that I should be able to follow what you're doing.

For the case of two populations:

$$\sum_{l=1}^{L} \alpha_l (p_{1l} - \epsilon_l) \sum_{l'=1}^{L} \alpha_{l'} (p_{1l'} - \epsilon_{l'}) + \sum_{l=1}^{L} \alpha_l (p_{2l} - \epsilon_l) \sum_{l'=1}^{L} \alpha_{l'} (p_{2l'} - \epsilon_{l'}) \tag{6}$$

Taking only the first term:

$$\sum_{l=1}^{L}(\frac{1}{2}A_l + D_l(1-2p_{1l}))^2(Varp_{l1})\sum_{l\neq l'}^{L}(\frac{1}{2}A_l + D_l(1-2p_{1l}))(\frac{1}{2}A_{l'} + D_{l'}(1-2p_{1l'}))Cov(p_{1l}, p_{1l'}) \quad (7)$$

Following Jeremy's paper– I believe the first term corresponds to F_{st} while the second term is the LD-like component. Could possibly expand out first squared alpha term making use of the following identities for variance:

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \quad (8)$$

While the second alpha term with different subscripts could be another covariance term.

Also note that when you square a summation you can use this equality:

$$(\sum_{l=1}^{L} X_l)^2 = \sum_{l=1}^{L} X_l^2 + \sum_{l\neq l'}^{L} X_l X_{l'} \quad (9)$$

This is slightly modified from its source `https://math.stackexchange.com/questions/329344/what-is-the-square-of-summation`. Should probably check to make sure this is valid.

If that equation is true then looking at just the term A_L

$$(\sum_{l=1}^{L} A_l)^2 = (\mathbb{E}[A_l])^2 = \sum_{l=1}^{L} A_l^2 + \sum_{l\neq l'}^{L} A_l A_{l'} \quad (10)$$

$$Var(A_l) = \sum_{l=1}^{L} A_l^2 - (\sum_{l=1}^{L} A_l^2 + \sum_{l\neq l'}^{L} A_l A_{l'}) \quad (11)$$

$$Var(A_l) = \sum_{l\neq l'}^{L} A_l A_{l'} \quad (12)$$

Similarly for the other term of α, that is, $Var(D_l(1-2p_{1l})) = \sum_{l\neq l'}^{L} D_l(1-2p_{1l})D_{l'}(1-2p_{1l'})$.

If this is true, both of these expressions can be found by multiplying out the α terms in the second summation in 7 to yield:

$$\frac{1}{16}Var(A_l) + Var(D_l(1-p_{1l})) + \sum_{l\neq l'}^{L}(\frac{1}{2}A_l D_{l'}(1-p_{1l'}) + \frac{1}{2}A_{l'}D_l(1-p_{1l})) \quad (13)$$

References

Berg, J. J. and G. Coop, 2014, August)A population genetic signal of polygenic adaptation. PLOS Genetics *10*(8)**:** e1004412.

Bohrnstedt, G. W. and A. S. Goldberger, 1969, December)On the Exact Covariance of Products of Random Variables. Journal of the American Statistical Association *64*(328)**:** 1439–1442.