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:

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

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 here. Should probably check to make sure this is valid.
If that equation is true then we can substitute, for example, 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:

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

Considering the entire second summation in 7, if we include the $(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})$ in the expansion we get:

$$\begin{aligned} &-\frac{1}{16}Var(A_l(p_{1l}-\epsilon_l)) - Var(D_l(1-p_{1l})(p_{l1}-\epsilon_l)) + \\ &\sum_{l\neq l'}^{L}(\frac{1}{2}A_l D_{l'}(1-p_{1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'}) + \frac{1}{2}A_{l'}D_l(1-p_{1l})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})) \end{aligned} \quad (14)$$

This is the first term in 7, multiplied out:

$$(\mathbb{E}[(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2] + \mathbb{E}[(D_l(1-2p_{1l})(p_{1l}-\epsilon_l))^2] + \sum_{l=l'}^{L} A_l D_l(1-2p_{1l})(p_{1l}-\epsilon_l)^2 \quad (15)$$

So far I see only one major problem with this approach: since in 7 the summation reads $l = l'$ and not $l = 1$ I'm not sure that the expressions for the expected values of the expanded terms (8) are valid, making 15 invalid.

Another track: forgetting about turning summations into variances and expectations for a moment, we merely expand the first and second terms of equation 7 in order to recombine them into summations where both l and l' start at 1.

$$\text{Term 1: } \sum_{l=l'}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2+\sum_{l=l'}(D_l(1-2p_{1l})(p_{1l}-\epsilon_l))^2 \tag{16}$$

$$\begin{aligned}\text{Term 2: } &\sum_{l\neq l'}\frac{1}{4}A_lA_{l'}(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+D_l(1-2p_{1l})D_{l'}(1-2p_{1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+\\ &\frac{1}{2}A_{l'}D_l(1-2p_{1l})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})\frac{1}{2}A_lD_{l'}(1-2p_{1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})\end{aligned} \tag{17}$$

If we multiply Term 1 and Term 2 above, the first term we'll get is:

$$\begin{aligned}\sum_{l=l'}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}\frac{1}{4}A_lA_{l'}(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})=\\ \sum_{l=1}^{L}\frac{1}{2}A_l(p_{1l}-\epsilon_1)\sum_{l'=1}^{L}\frac{1}{2}A_{l'}(p_{1l'}-\epsilon_{l'})\end{aligned} \tag{18}$$

Putting it all together:

$$\begin{aligned}&\sum_{l=1}^{L}\frac{1}{2}A_l(p_{1l}-\epsilon_l)\sum_{l'=1}^{L}\frac{1}{2}A_{l'}(p_{1l'}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}D_l(1-2p_{1l})D_{l'}(1-2p_{1l'})(p_{1l}-\epsilon_l)(p_{l1}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}\frac{1}{2}A_lD_{l'}(1-2p_{1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}\frac{1}{2}A_{l'}D_l(1-2p_{1l})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(D_l(1-2p_{1l})(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}\frac{1}{4}A_lA_{l'}(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(D_l(1-2p_{1l})(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}\frac{1}{2}A_{l'}D_l(1-2p_{1l})(p_{1l}-\epsilon_l)(p_{1l}-\epsilon_{l'})+\\ &\sum_{l=l'}^{L}(D_l(1-2p_{1l})(p_{1l}-\epsilon_l))^2\sum_{l\neq l'}^{L}\frac{1}{2}A_lD_{l'}(1-2p_{1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})+\\ &\sum_{l=1}^{L}D_l(1-2p_{1l})(p_{1l}-\epsilon_l)\sum_{l'=1}^{L}D_{l'}(1-2p_{1l'})(p_{1l'}-\epsilon_{l'})\end{aligned} \tag{19}$$

The equations in 18 agree more with the expressions for turning sums into expectations, but at the same time take us father from the goal of partitioning into F_s-like and LD-like terms that correspond to sums where $l=l'$ and $l\neq l'$, respectively.

to sums where $l=l'$ and $l\neq l'$, respectively.

(comment from Jeremy) So I think you can rearrange equation 18 as

$$\sum_{l=1}^{L} \frac{1}{2} A_l (p_{1l} - \epsilon_1) \sum_{l'=1}^{L} \frac{1}{2} A_{l'} (p_{1l'} - \epsilon_{l'}) = \frac{1}{4} \sum_{l=1}^{L} \sum_{l'=1}^{L} A_l (p_{1l} - \epsilon_1) A_{l'} (p_{1l'} - \epsilon_{l'}) \tag{20}$$

$$= \frac{1}{4} \left(\sum_{l=1}^{L} (A_l (p_{1l} - \epsilon_1))^2 + \sum_{l=1}^{L} \sum_{l \neq l'}^{L} (A_l (p_{1l} - \epsilon_1))(A_{l'} (p_{1l'} - \epsilon_1)) \right) \tag{21}$$

$$= \frac{1}{4} \left(\sum_{l=1}^{L} Var\left(A_l (p_l - \epsilon_l)\right) + \sum_{l \neq l'} Cov\left(A_l (p_l - \epsilon_l), A_{l'} (p_{l'} - \epsilon_{l'})\right) \right) \tag{22}$$

$$= \frac{1}{4} \left(\sum_{l=1}^{L} A_l^2 Var\left(p_l - \epsilon_l\right) + \sum_{l \neq l'} A_l A_{l'} Cov\left(p_l - \epsilon_l, p_{l'} - \epsilon_{l'}\right) \right) \tag{23}$$

This expression lines is essentially the one from my paper (in the case where $D = 0$ for all loci, the additive effects (α) just reduce to $\frac{A}{2}$. The trick to get from line (22) to line (23) is that the A_l are constants with respect to the evolutionary process (i.e. genetic drift), and therefore can be pulled outside of the variance and covariance terms.

Ultimately, in order to work out what the bias in the statistic is with dominance we need to work out its expected value in the presence of dominance. Comparing this to the expected value in the case of no dominance will give us a sense of how strong the bias is. For example, in the case of no dominance, the statistic is just given by line (23). When we take the expectation, the second term in line (23) is zero, because alleles that are not tightly linked to one another drift independently (i.e. $\mathbb{E}[Cov(p_l - \epsilon_l, p_{l'} - \epsilon_{l'})] = 0$ for all l and l'). This leaves the expectation of the numerator of our statistic as

$$\frac{1}{4} \mathbb{E} \left[\sum_{l=1}^{L} A_l^2 Var(p_l - \epsilon_l) \right] \tag{24}$$

and we can use the linearity of expectation to push the expectation inside the sum

$$\frac{1}{4} \sum_{l=1}^{L} \mathbb{E} \left[A_l^2 Var(p_l - \epsilon_l) \right] \tag{25}$$

which is a familiar quanitity in population/quantitative genetics, such that we can reexpress it in terms of parameters like F_{ST} and V_A, but we'll leave that for later.

My intution says that the decomposition of the test statistic that we're looking for includes terms that look like

$$\sum_{l=1}^{L} \left(\frac{A_l}{2} D_l (1 - 2p_l)(p_{1l} - \epsilon_1)^2 \right) + \sum_{l=1}^{L} \sum_{l \neq l'}^{L} \left(\left(\frac{A_l}{2} (p_{1l} - \epsilon_1) \right) \left(D_{l'} \left(1 - 2p_{l'}\right) \left(p_{1l'} - \epsilon_1\right) \right) \right) \tag{26}$$

and

$$\sum_{l=1}^{L} \left(\left(D_l (1 - 2p_l)(p_{1l} - \epsilon_1) \right)^2 \right) + \sum_{l=1}^{L} \sum_{l \neq l'}^{L} \left(\left(D_l \left(1 - 2p_l\right) \left(p_{1l} - \epsilon_1\right) \right) \left(D_{l'} \left(1 - 2p_{l'}\right) \left(p_{1l'} - \epsilon_1\right) \right) \right) \tag{27}$$

and in fact I suspect that the ultimate expression we're looking for is basically a sum of line (21), 2 times line (26), and line (27).

Let's step back from worrying about variances and covariances for a minute (I think I was trying to have you solve too many steps at once by suggesting to focus on that interpretation). See if you can

see your way to verifying my intuition about the above expressions. The stuff you have written in lines (**??**) seems very close to what we're looking for, but 1) I think you may be missing some addition signs in between the two different sums on each line (but I'm not certain, as I had a little trouble following exactly how you arrived at these expressions), and 2) I don't totally follow your indexing, as for each case you say that l goes from l' to L but at no point is it specified what l' is, so that seems like it can't be quite right.

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.