Thursday, 19 June 2014

Test statistic for epistatic interaction in the presence of an imperfectly tagged additive QTL

(Peter Visscher & Mike Goddard)

Notation and assumptions

Consider 3 loci. Locus 1 is causal but not observed, locus 2 is correlated by LD and locus 3 is uncorrelated. We are considering a haploid system to keep the calculation as simple as possible.

$y_i = 0$ or 1 ($i = 1, 2, 3$), with $E(y_i) = p_i$ ($i = 1, 2, 3$).

The haplotype frequencies for loci 1 and 2 are

$p_{00} = (1-p_1)(1-p_2) + D$
$p_{01} = (1-p_1)p_2 - D$
$p_{10} = p_1(1-p_2) - D$
$p_{11}= p_1p_2 + D$

The LD correlation between loci 1 and 2 is $r_{12} = D / \text{sqrt}[p_1(1-p_1)\ p_2(1-p_2)]$

Also,

$E(y_1y_2) = p_{11}= p_1p_2 + D$, $\text{cov}(y_1,y_2) = D$
$E(y_1y_3) = p_1p_3$, $\text{cov}(y_1,y_3) = 0$
$E(y_2y_3) = p_2p_3$, $\text{cov}(y_2,y_3) = 0$
$E(y_1y_2y_3) = p_{11}p_3 = (p_1p_2 + D)p_3$, and
$\text{cov}(y_1,y_2*y_3) = (p_1p_2 + D)p_3 - p_1p_2p_3 = Dp_3$.

Therefore, y_1 and the interaction term y_2*y_3 are not independent.

Model for analysis

Real phenotypic data (Y), e.g. expression data, would have the form $Y = b*y_1 + e$, but initially we will assume the phenotype is simply y_1.

Fitting linear models to y1 we compare

$y_1 \sim y_2 + y_3 + e_1$ [M1]
$y_1 \sim y_2 + y_3 + y_2*y_3 + e_2$ [M2]

The test statistic for the interaction term is $F_{12} = \{(SSE_1 - SSE_2)/1\} / \{(SSE_2 / (n-4)\}$ where SSE is the error sum of squares from model 1 or 2, and {1} and {n-4} are the df for the terms in the numerator and denominator, respectively. If the errors e_1 and e_2 were to be normally distributed then under the null hypothesis of no interaction, F_{12} would follow a central F-distribution with 1 and (n-4) degrees of freedom. For large n, this distribution is approximately χ^2 with 1 df.

For large n, we can assume that

SSE_2 / (n-4)) ≈ $(1 – r_{12}^2)var(y_1) = (1 – r_{12}^2)p_1(1-p_1)$. This is the residual variance from M1 and M2, i.e. under the null hypothesis of no interaction variance. Hence, under the null of no interaction,

$var(e_1) = var(e_2) = (1 – r_{12}^2)p_1(1-p_1)$

We are assuming that the estimate of the residual variance for both models is unbiased.

The test statistic $F_{12} = (SSE_2 – SSE_1) / (SSE_2 / (n-4)) = Q / var(e_2)$

The observation from simulations (and real data) is that Q / $var(e_2)$ is not (always) distributed as a central F or χ^2 with 1 df under the null distribution of no interaction. **The reason (as explained below) is that the variance of the test statistic is not as assumed, that is, it is an issue of the distribution of the data.**

Simulations

2 locus haplotypes on y_1 and y_2 were generated from the population haplotype frequencies, which follow from the input parameters. The haplotypes were fixed according to the population haplotype frequencies and kept constant across replicates (hence one element of the X'X matrix and one element of X'y are constant). This was to mimic a case where a single tag-SNP is compared to many uncorrelated SNPs for pairwise interactions. Not fixing the haplotype frequencies was also implemented and didn't make much difference to the results. For each replicate, a third unlinked locus was generated using binomial sampling. Data were analysed using the models M1 and M2 as described above.

Example: n = 1000, $p_1 = p_2 = 0.1$, $p_3 = 0.5$, $r_{12} = 0.5$. 10,000 replicates.
Haplotype frequencies (out of a 1000) for y_1 and y_2 are 855, 45, 45 and 55.

$var(y_1) = 0.1*0.9 = 0.09$. $var(e_1) = var(e_2) = (1 – 0.5^2)*0.09 = 0.0675$.

The results from 10,000 simulations give an average F-statistic for the interaction term of ~3.4 with a variance across replicates of ~23. The expected values are 1.0 and 2.0, respectively. The simulations suggest a false-positive rate of 29% (using an F-value threshold corresponding to 5%).

Theory

Let $y_{1.ij} = (y_1 = 1|y_2=i, y_3=j)$, with i,j = {0,1}.

Let $n_{ij} = n * (1-p_2+i*(2p_2-1))(1-p_3+i*(2p_3-1))$. These give the numbers in the 2x2 cells of combinations of y_2 and y_3:

y_2	y_3	n_{ij}/n
0	0	$(1 – p_2)(1-p_3)$
1	0	$p_2(1-p_3)$

0	1	$(1 - p_2)p_3$
1	1	p_2p_3

A test for interaction is:

$\delta = \text{mean}(y_{1.11}) + \text{mean}(y_{1.00}) - \text{mean}(y_{1.10}) - \text{mean}(y_{1.01})$,

with $\text{mean}(y_{1.ij}) = \Sigma(y_{1.ij}) / n_{ij}$.

The exact variance of δ can be calculated from the allele and haplotype frequencies, and compared to its variance under a linear model (LM). From the given haplotype frequencies:

$E(y_{1.11}) = E(y_{1.10}) = P(y_1=1|y_2=1) = p_{11}/p_2 = p_1 + D/p_2$. Similarly,
$E(y_{1.00}) = E(y_{1.01}) = p_1 - D/(1-p_2)$

Each of the terms has a binomial variance:

$\text{var}(\text{mean}(y_{1.ij})) = E(y_{1.ij})(1 - E(y_{1.ij})) / n_{ij}$.

Putting all the terms together gives the exact variance of the test statistic as,

$\text{var}(\delta)$

$$= \{p_1(1-p_1) + (1-2p_1)(1-2p_2)D/[p_2(1-p_2)] - D^2(1 - 3p_2(1-p_2))/[p_2^2(1-p_2)^2]\}/c \qquad \text{[Eq 1]}$$

with $c = n * p_2(1-p_2)p_3(1-p_3)$

In the linear model, the error variance is assumed to be the same in each cell of the 2x2 table and a pooled estimate is used. As a result

$$c * \text{var}(\delta_{LM}) = p_1(1-p_1)(1 - r_{12}^2) = \text{var}(e_{LM}) \qquad \text{[Eq 2]}$$

Thus, when using a linear model, an incorrect error variance of the interaction test is assumed, and this can lead to inflated (or deflated) type-I error rates. The ratio of the exact and linear model variances is the expected value of the linear model F-test.

Re-arranging gives,

$$c * \text{var}(\delta) = \text{var}(e_{LM}) + (1-2p_1)(1-2p_2)D/[p_2(1-p_2)] - D^2(1-2p_2)^2/[p_2^2(1-p_2)^2] \qquad \text{[Eq 3]}$$

Unless $D = 0$ or $p_2 = ½$, the exact variance is different from that under the linear model. If loci 1 and 2 are the same ($p_1=p_2$, $r_{12} = 1$) then the variance is also the same as that from the linear model. The middle term can result in a substantial inflation of the test statistic when using a linear model. The ratio of the variances can be expressed as:

$\mathrm{var}(\delta)/\mathrm{var}(\delta_{LM}) = E(F)$,

$E(F) = 1 + (1\text{-}2p_1)(1\text{-}2p_2)D/[\mathrm{var}(e_{LM})p_2(1\text{-}p_2)] – D^2(1\text{-}2p_2)^2/[\mathrm{var}(e_{LM})p_2^2(1\text{-}p_2)^2]$

[Eq 4]

Note that his expression (Eq 4) does not depend on the allele frequency of the unlinked locus.

For the parameters of the example (Box above), the expected F-statistic using Eq 4 gives E(F) = 3.47, as observed from simulations.

Quantitative traits

Let $Y = y_1 + e$,

with $e \sim N(0, \mathrm{var}(e))$, with

$\mathrm{var}(e) = p_1(1\text{-}p_1)(1\text{-}R^2)/R^2$, and $R^2 = \mathrm{var}(y_1)/\mathrm{var}(Y)$, i.e. the proportion of variance in the quantitative trait explained by locus 1. The variance of the trait is now a mixture of a binomial and normal.

For the test statistics, the same principles apply as before, with the difference that $\mathrm{var}(e_{LM}) = p_1(1\text{-}p_1)(1/R^2 – r_{12}^2)$. Hence the inflation (or deflation) of the test statistic is,

$E(F) = 1 + (1\text{-}2p_1)(1\text{-}2p_2)D/[\mathrm{var}(e_{LM})p_2(1\text{-}p_2)] – D^2(1\text{-}2p_2)^2/[\mathrm{var}(e_{LM})p_2^2(1\text{-}p_2)^2]$

For R^2 values of 0.25, e.g. the size of a large cis-eQTL, the inflation can still be considerable. For example, for $p_1 = p_2 = 0.1$ and $r_{12} = 0.25$, the mean F-test is 1.47.

In conclusion: the presence of a large unobserved QTL can lead to a substantial inflation of the test statistic for interaction between a linked marker and an unlinked marker, as in cis-trans interactions. The same principle applies for cis-cis interactions, if one of the cis-markers is not in LD with the causal variant. Although the theory and simulations were on haploids, the same problem of the distribution of the data is an issue for the diploid F-test on 4 degrees of freedom. It might be worse for a 4-df test because the test is across multiple genotype combinations, all with a mixture of binomial and normal errors. The exact variance of the 4-df test under the assumption of a linked major QTL could be derived using the same principle as described above, but is likely to result in an even more complicated equation and it might not give more insight. The exact variance could be calculated without giving an explicit equation, but this would be akin to performing a simulation experiment.

If an F-test is used on real data then one could use a permutation test to get the empirical distribution of the test statistic. This could be done in validation data if only a few interaction pairs are to be tested for significance. It is not clear how one would implement a permutation test in a discovery sample because an empirical p-value threshold might be needed for every pair that is tested.