Genome-wide LD change as a signature of selection in experimental populations

Moi Exposito-Alonso

Genotypes and LD

Given two loci, A and B, with alleles A and a, and B and b, there must be nine possible gammetes. If the A allele and B allele are selected with coefficients s_A and s_B, and we assume no dominance and no epistasis, we can write the average fitness per genotype. For simplicity we use an haploid model, which would be similar to the diploid homozygotes case:

	B	b
A	$1 + s_B + s_A + (s_B \times s_A)\kappa$	$1 + s_A$
a	$1 + s_B$	1

Note that in this model, although it is haploid, it is general enough to represent additive as well as interaction effects in fitness. When the $\kappa > 1$, there ismultiplicative fitness, when $0 < \kappa < 1$ it is synergistic fitness, cases of $\kappa < 0$ could represent some type of antagonistic epistatic effects (although there is room for other functions).

If we know the frequency of A (p) and B (q) in the population, and also have faced genotype, so the frequency of the different gametes can be calculated, the LD between the alleles can be erxpressed as:

$$D = x_1 - pq = x_1x_4 - x_2x_3 \tag{1}$$

And the scaled measurement of LD independent on frequency can be written as:

$$r^2 = \frac{D}{\sqrt{p(1-p)q(1-q)}} \tag{2}$$

Overall, the table of frequencies would look like:

	B	b
A	$x1 = pq + D$	$x3 = p(1-q) - D$
a	$x3 = (1-p)q - D$	$x4 = (1-p)(1-q) + D$

Selection

Then selection over allele A should produce an increase of the frequency p:

$$p_{t+1} = p_t(1 + s_A)/\bar{w} \tag{3}$$

And with linkage disequilibrium:

$$p_{t+1} = \frac{x_{1\ t+1} + x_{2\ t+1}}{\bar{w}} \tag{4}$$

$$p_{t+1}\bar{w} = (pq + D)(1 + s_A + s_B + K) + (p(1-q) - D)(1 + s_A) \tag{5}$$

$$p_{t+1}\bar{w} = p(1 + s_A) + pq(s_B + K) + d(s_B + K) \tag{6}$$

$$\tag{7}$$

,where $K = s_A s_B \kappa$ for simplicity.

We got ride of the denominator term $\bar{w}$ that is the average fitness in the next generation. We can develop the equation of haplotype frequencies to express it in terms of p, q and D:

$$\bar{w} = x1(1 + s_A + s_B) + x2(1 + s_A) + x3(1 + s_B) + x4 = 1 + ps_A + q_s B + pqK + DK \tag{8}$$

Interestingly, the mean fitness does not depend in the LD term except if there is interaction of fitness effects. Note: this was solved by noting that all haplotypes need to sum 1 and that $pq + p(1 - p) = p$ and vice versa.

Change in LD due to selection

From Felsenstein we know that LD would not change with selection unless there is an epistatic term.

Relationship with Genome-Wide LD calculations

Having a X genome matrix with N individuals in the rows and M SNPs as columns, we efficiently obtain the genome-wide LD using linear algebra as:

$$V = \frac{1}{N} X^T X \tag{9}$$

The matrix V has in the off-diagonal elements the r^2 LD coefficient. Thi is because:

$$r^2 = \frac{cov(X, Y)}{sd(X) \times sdY} \tag{10}$$

$$= \frac{E[(X - \bar{x})(Y - \bar{y})]}{\sqrt{E[(X - \bar{x})^2] \times E[(X - \bar{x})^2]}} \tag{11}$$

$$\tag{12}$$

The equation (9) is only identical to (10) if the X matrix is mean centered, and also variance scaled.

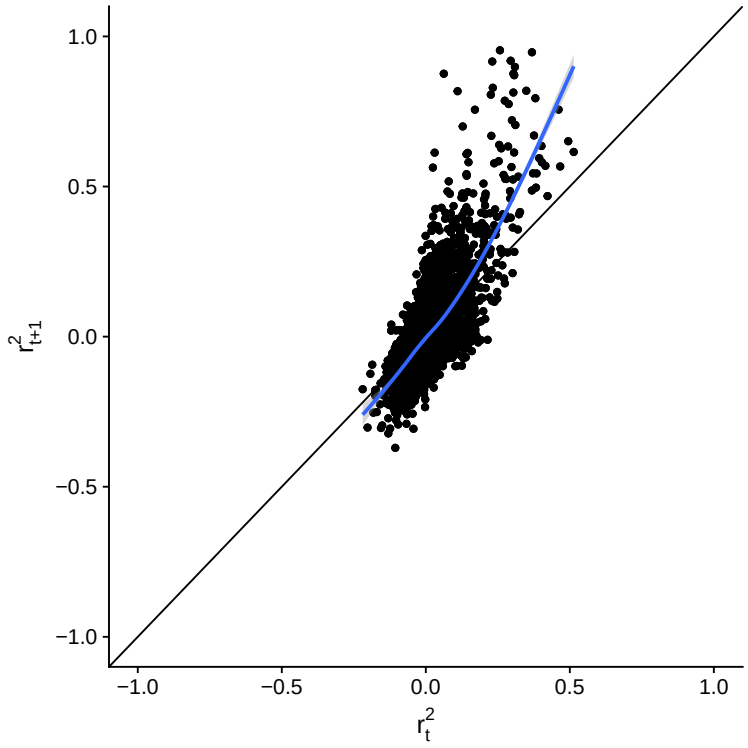
If X would not be variance centered, V would correspond to a covariance matrix, and if it was not mean centered, it would be other non-standard statistical entity.

One more relationship about the covariance:

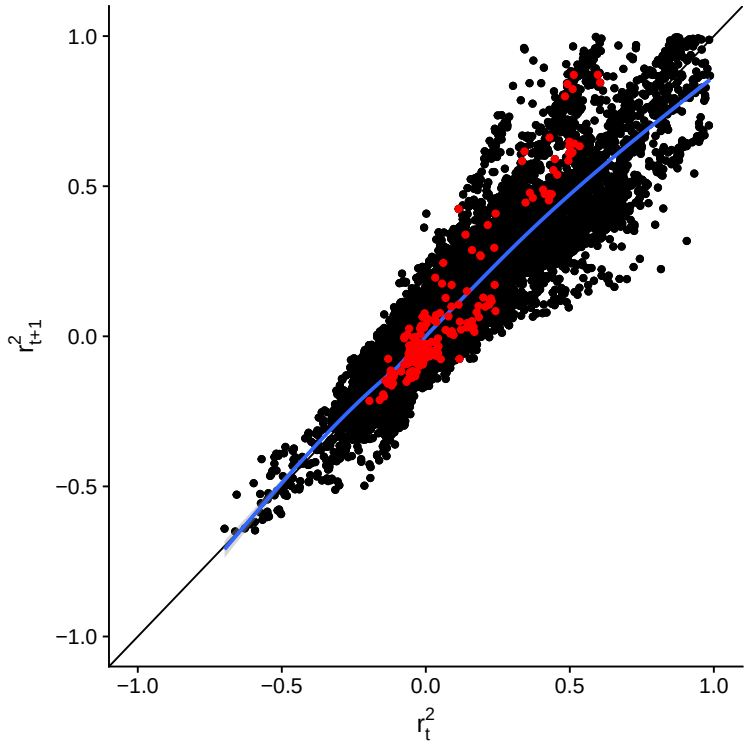
$$cov(XY) = E[(X - \bar{x})(Y - \bar{y})] = E[XY] - E[X]E[Y] = \frac{XY^T}{N} - \frac{E[X]E[Y]^T}{N} = D. \tag{13}$$

Data on LD change

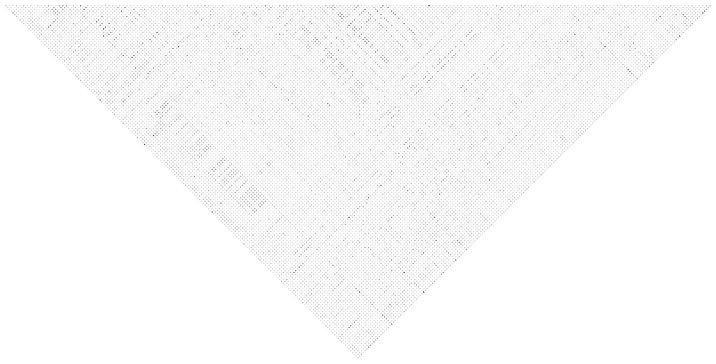
Using the absolute fitness from the field experiment under drought conditions and a randomly selected set of SNPs, the change in LD can be calculated.



The same can be done in a window of 200 SNPs around a region that had a peak in the GWA.



And this is the raw change.



The use of absolute fitness

Having absolute fitness, how can I calculate the new covariance without actually building a new genome matrix?

Quantification of the total potential interference of LD

Summatory of all LD values relative to the total genetic distances. It needs to be expressed as per bp or cM in order to make it comparable with other systems that might not have the same density of markers.

Inference of selection from LD

It would be of interst to genearte an expectation of V given 2 loci (or more) under selection.