

# Variations on LDRecon

Thomas Mailund

April 25, 2004

We consider a new linkage disequilibrium calculation to locate markers close to a trait position. The approach is based on considering the separation of subjects into cases and controls as a pseudo-marker with two alleles, and calculating the linkage disequilibrium between all other markers and this pseudo-marker.

## INTRODUCTION

The gametic linkage disequilibrium between allele A and allele B, on markers M and N respectively, is defined as  $D_{AB} = p_{AB} - p_A p_B$  where  $p_{AB}$  is the frequency of haplotype  $p_{AB}$  and  $p_i$  is the allele frequency for allele  $i$ . The digenic disequilibrium,  $D_{A/B}$ , refers to the disequilibrium between A and B on different gametes and is defined as  $D_{A/B} = p_{A/B} - p_A p_B$  where  $p_{A/B}$  is the frequency of allele A on one gamete and allele B on the other.

When the diplotype frequencies are unknown we cannot calculate  $D_{AB}$  and  $D_{A/B}$  directly, but can instead calculate their composite disequilibrium,  $\Delta_{AB}$  defined as  $\Delta_{AB} = D_{AB} + D_{A/B} = p_{AB} + p_{A/B} - 2p_A p_B$  and can be estimated as

$$\hat{\Delta}_{AB} = \frac{n_{AB}}{n} - 2\tilde{p}_A \tilde{p}_B \quad (1)$$

where  $n_{AB}$  is the count of genotype AB

$$n_{AB} = 2n_{AABB} + n_{AABB} + n_{A\bar{A}BB} + \frac{1}{2}n_{A\bar{A}B\bar{B}} \quad (2)$$

and  $\tilde{p}_i$  is the observed frequency of allele  $i$ .

The variance of  $\hat{\Delta}_{AB}$ , under the hypothesis that  $\Delta_{AB} = 0$ , is

$$\text{Var}(\hat{\Delta}_{AB}) = \frac{1}{n} ((\pi_A + D_A)(\pi_B + D_B)) \quad (3)$$

where  $\pi_i = p_i(1-p_i)$  and  $D_i = p_{ii} - p_i^2$ . Estimating these, using the observed frequencies,  $\tilde{\pi}_i = \tilde{p}_i(1 - \tilde{p}_i)$  and  $\hat{D}_i = \tilde{p}_{ii} - \tilde{p}_i^2$ , we get a random variable  $X_{AB}^2$  define as

$$X_{AB}^2 = \frac{n\hat{\Delta}_{AB}}{(\tilde{\pi}_A + \hat{D}_A)(\tilde{\pi}_B + \hat{D}_B)} \quad (4)$$

which, again under the assumption of  $\Delta_{AB} = 0$ , is  $\chi^2$  distributed, with one degree of freedom.

To get the total linkage between marker M and N we  $X_{AB}^2$  over all alleles A on M and all alleles B on N, and get a value  $\chi^2$  distributed with  $(k-1)(l-1)$  degrees of freedom, if M has  $k$  alleles and N has  $l$  alleles.

## SEARCHING FOR A TRAIT-POSITION

---

When considering some trait that separates the population into ‘cases’ and ‘controls’, we are searching for markers highly linked with this separation, as these are more likely to be close to the mutation responsible for the trait.

To calculate the linkage between a marker and the trait, we can consider the trait a pseudo-marker with two alleles: cases and controls, and we can calculate the linkage using the method described above. Here,  $p_{\text{cases}} = \frac{|\text{cases}|}{|\text{cases}|+|\text{controls}|}$ ,  $p_{\text{controls}} = \frac{|\text{controls}|}{|\text{cases}|+|\text{controls}|}$ , and  $\hat{D}_{\text{cases}} = \hat{D}_{\text{controls}} = \frac{|\text{cases}| \cdot |\text{controls}|}{|\text{cases}|+|\text{controls}|^2}$ .

To get a candidate position for the trait gene, we can then find the marker with the maximal p-value. If there is only a single spike of p-values, that is our best guess for a position; if there are more than one spike the case is a bit more complicated since any of the spikes are candidates, but they can potentially be far apart. In that case, some more work is required – here we just assume that the highest peak is the best choice.

## EXPERIMENTS

---

We have conducted 500 CoaSim simulations (with growth=0, 2000 leaf-nodes, and 50 as recombination rate), with 10 evenly placed SNP markers and one randomly placed trait marker, separated the resulting data into 100 cases and 100 controls.

Since we are only working with SNP markers, the degrees of freedom is the same for all markers and we do not need to calculate the p-values; we can just take the maximal linkage value.

We found the marker with highest linkage, and we then calculated the distance from that marker to the true position, and plotted the results as shown in Figure 1.

As mentioned above, blindly selecting the marker with highest linkage is not always a good idea, since we can have two or more spikes where only one is at the marker, so reducing the data to a single point is throwing away important information. Consider Figure 2 (a) and (b), here there are peaks closer to the true position than the one selected by taking the maximum value. Of course, the linkage value can also just be completely off at guessing the true position, as in (c) and (d).

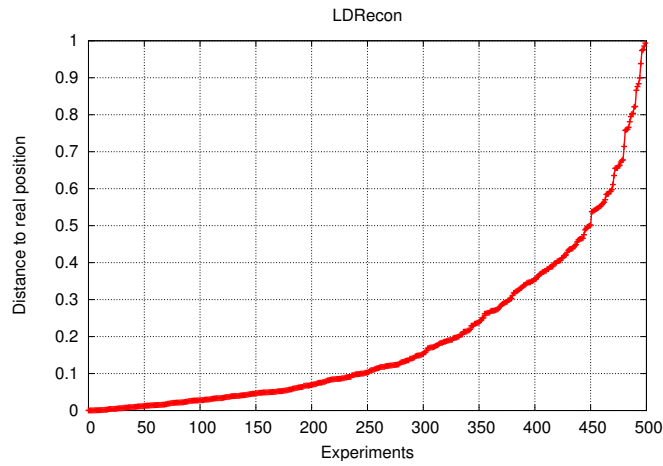
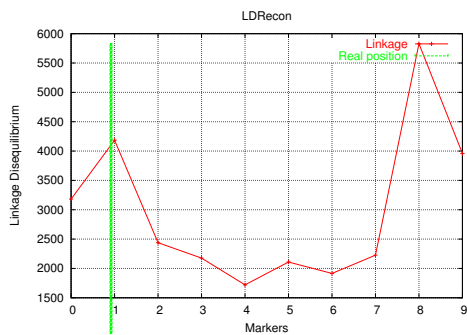
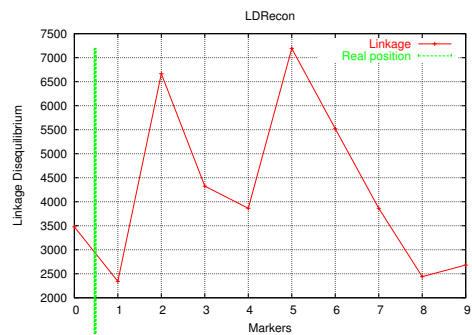


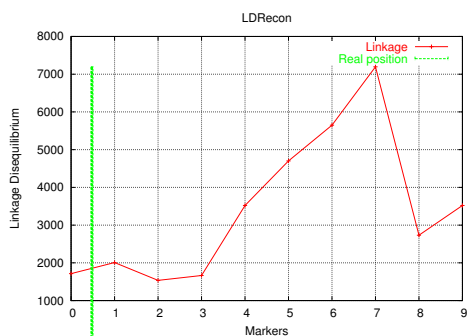
Figure 1: Distance from best guess to true position. The 10 SNP markers are placed at an even distance of 0.1, the trait is placed at random in the interval (0.0, 1.0).



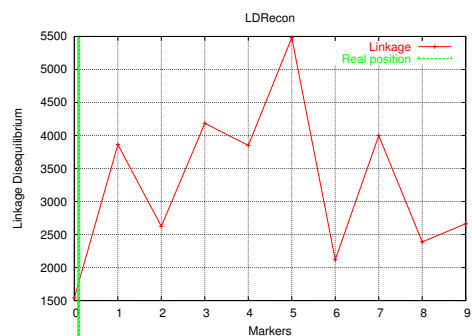
(a)



(b)



(c)



(d)

Figure 2: Examples of linkage of the different markers.