LINKAGE DISEQUILIBRIUM IN INBRED POPULATIONS : A GEOSTATISTICAL APPROACH

J. Nsengimana and P.V. Baret

Université Catholique de Louvain, Faculté d’Ingénierie Biologique, Agronomique et Environnementale, Unité de Génétique, Croix du sud 2 bte 14, B-1348 Louvain-la-Neuve, Belgium

INTRODUCTION

Linkage disequilibrium (LD) mapping is a promising approach to finely map genes in human populations (e.g. Hästbacka et al. 1992) and in livestock (Baret and Hill, 1997 ; Farnir et al. 2000). However, this approach has two main limitations: (1) the pairwise nature of LD measurements and (2) the difficulty of reconstructing population’s history. To handle the first constraint, the theory of geostatistics (Christakos, 1992) provides interesting tools to quantify relationships between all loci of a linkage group. The basic geostatistical tool is the semi-variogram $\gamma(h)$[eq.1], which is a measure of the level of the dissimilarity between two positions.

$$\gamma (h) = \frac{1}{2} E \{ [Z(x+h) - Z(x)]^2 \}$$

where $Z(\cdot)$ is a random variable, $x$ and $x+h$ are coordinates of the two positions. Two other geostatistical functions are based on $\gamma(h)$ and measure the similarity rather than the dissimilarity. These functions are the covariance function $C(h)=S\gamma(h)$ [eq.2, also called covariogram] and the correlogram $\rho(h)=C(h)/S=1-\gamma(h)/S$ [eq.3], where $S$ (the sill) is the maximum value of $\gamma(h)$ (e.g. Christakos, 1992). The second constraint may be circumvented if we are able to model the effect of the structure of the population on LD. This study aims to handle the two limitations of LD mapping: on the basis of simulated data, (1) we introduce a chromosomewise measure of LD and (2) we extend the classical model of LD breakdown to inbred populations.

SIMULATIONS

Five populations are simulated with the same size ($N=200$) and different inbreeding levels (Table 1). In each population, a unique diploid founder is considered and his both haplotypes are known (phases 1 and 2) for 50 markers evenly spaced on a 49cM chromosome. Alleles from other individuals than the founder are not identified (phase 0). The transmission of founder alleles to his progeny is traced through 30 generations. Two mating designs are simulated in each population : half-sib from generation 1 to the second, full-sib with a fixed number of random crosses from generation 2 to 30. The sex ratio is always of 1:1. Parental chromosomes given to offspring are drawn from a binomial distribution with frequency 0.5 and the Haldane mapping function is used to model the recombination events. For each individual, the recorded data consist in phases at every locus (phase 0, 1 or 2) and the inbreeding coefficient ($F$). The simulation is run 1000 times for each population.
### Table 1. Structures of simulated populations

<table>
<thead>
<tr>
<th>Population</th>
<th>1st to 2nd generation</th>
<th>2nd to 30th generation</th>
<th>Population size</th>
<th>Inbreeding coefficient in generation 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop4</td>
<td>1</td>
<td>200</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>Pop10</td>
<td>1</td>
<td>200</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Pop25</td>
<td>1</td>
<td>200</td>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>Pop100</td>
<td>1</td>
<td>200</td>
<td>100</td>
<td>2</td>
</tr>
</tbody>
</table>

GEOSTATISTICAL TOOLS AS CHROMOSOMWISE MEASURES OF LD

We consider founder alleles carried by each chromosome in any generation as a random spatial variable (Christakos, 1992). In generation 30, founder phases 1 and 2 are transformed into an indicator variable representation, respectively 0 and 1. Indicator semi-variograms of founder phases $\gamma(h)$ are then estimated along each chromosome with eq.1 and averaged within each population. After adjustment to theoretical spatial models, the covariogram and the correlogram (eq.2 and eq.3) can be derived.

When two linked loci carry the same founder phase (haplotypes 11 and 22), the corresponding semi-variance $\gamma$ is 0 and the correlation $\rho$ is 1; while for two loci carrying opposite founder phases (haplotypes 12 and 21), $\gamma=0.5$, and $\rho=0$. Loci with similar founder phases ($\gamma=0$ and $\rho=1$) are in complete LD and loci with opposite phases ($\gamma=0.5$ and $\rho=0$) are not in LD. Thus, $\gamma$ and $\rho$ are measures of LD between a pair of loci. Extension of these pairwise LD to an entire chromosome gives chromosomewise measures of LD, i.e. $\gamma(h)$ and $\rho(h)$ [figure 1]. As the covariogram $C(h)$ is related to $\gamma(h)$, it is also a chromosomewise measure of LD. There is a relationship between the classical pairwise measure of LD, i.e. the coefficient $D$, and these geostatistical measures. In fact, $D$ is often interpreted as a covariance between loci (Lynch and Walsh, 1998). The covariance function (eq.2) along chromosomes can then be seen as an extension of $D$ to multiple loci.

As expected, $\gamma(h)$ increases with the genetic distance ($h$) until the sill is reached in each population (figure 1A). This increasing dissimilarity between loci corresponds to a decrease of LD with $h$, as indicated by the correlogram $\rho(h)$[figure 1B]. Adjustment of $\gamma(h)$ to theoretical models showed that the exponential model is the most appropriate (the coefficient of determination $r^2$ higher than 0.99). This pattern was expected as recombination events are the only factors that separate linked loci and their frequency is a negative exponential function of the genetic distance.
JOINT EFFECT OF TIME AND INBREEDING ON LD

The classical model of LD breakdown in inbred populations is

$$\frac{D_t}{D_0} = (1 - \theta)^t$$ \hspace{1cm} (4)

where $D_t$ and $D_0$ are respectively values of LD in the current and the base generation, $\theta$ is the recombination fraction, and $t$ is time. As this model is based on a HW equilibrium assumption, it will be referred to as the restricted model on LD. In extending this model to inbred populations, the correlogram $\rho(h)$ is used as a measure of LD. Indeed, as the spatial correlation between two loci is a ratio of the present covariance and the maximum covariance (eq.3), it is equivalent to the ratio $D_t/D_0$ used as a measure of LD in eq.4.

Haplotype data from all four populations have been transformed into an indicator variable (phases 1 and 2 became respectively 0 and 1) in 10 different generations between the first and the 30th. For each time, in each population, a correlogram has been computed. Considering the inbreeding as the main characteristic of the population structure, we model the LD breakdown in general population as

$$\rho = (1 - \theta)^{b_1 t + b_2 F}$$ \hspace{1cm} (5)

where $\rho$ is the spatial correlation, $\theta$ is the recombination rate, $F$ is the mean inbreeding coefficient and $t$ is time. This model fits our data with $r^2$ always higher than 0.99 in each population and a $r^2$ of 0.98 when all populations are analysed simultaneously. We assume that $\rho$ is equivalent to $D_t/D_0$ and then eq.5 is an extension of eq.4. This model (eq.5) will be referred to as the extended model of LD. On basis of 2000 values of $\rho$, we obtained 0.80 and –10.89 as least squares squares estimates of respectively $b_1$ and $b_2$. The effect of $F$ on LD is then greater than the effect of time and these effects are antagonist. Within each population, the effect of $F$ is still more important than the effect of time ($|b_2|>|b_1|$). However, both effects are opposite ($b_1b_2<0$) in pop4 and pop10 (the most inbred) while they are additive in pop25 and pop100($b_1b_2>0$).
DISCUSSION AND CONCLUSION

This study deals with complex interactions between time, inbreeding and LD. The effects of time and the inbreeding coefficient \( F \) have been shown to be additive in pop100 and pop25 while they are opposite in pop4 and pop10. The reason is that, LD decays in time due to an accumulation of crossing-over events (COs) whereas the effect of \( F \) is to mask some COs, those in which involved chromosome segments are identical by descent (IBD). It may be expected then that the higher \( F \) (i.e. the more recombinations could remain undetected), the slower the LD decay across generations. It follows up that, an expected amount of LD in equilibrium populations will be reached generations later in non-equilibrium populations. Accordingly, a gene position can be inferred in equilibrium populations with the restricted model (eq.4.) if the level of LD and the coalescent time are known [e.g. Hästbacka et al. 1992] while the same model may underestimate \( \theta \) in non-equilibrium populations. The underestimation will be a function of the inbreeding level. For populations of livestock where the inbreeding phenomenon is common and well documented (e.g. Boichard et al. 1996) the extended model (eq.5) would be more appropriate in finely locating production QTLs.

In conclusion, we propose a solution to deal with each of the two main assumptions of LD mapping. On the one hand, the chromosomewise measure of LD will be an important tool in the context of high density marker maps as one should no more assume the independence between pairs of linked loci. On the other hand, with the extended model of LD, LD mapping may be effective both in Hardy-Weinberg populations and in more general situations.

ACKNOWLEDGEMENTS

We are grateful for the financial support of the Belgian Fund for Research in Agriculture and Industry (FRIA), to Patrick Bogaert for providing us programs to estimate and model semi-variograms and to Pierre Tilquin for Helpful comments.

REFERENCES