Comparison of Genetic Structure of Mouse Populations by Combinatorial Analysis of High-Order Long-Range Linkage Disequilibrium Networks Using High Density Genotypes

Yun Zhang[1], Roumyana Kirova[2], Gary A. Churchill[3], Michael A. Langston[1], Elissa J. Chesler[2]

[1] Computer Science Department, University of Tennessee, Knoxville, TN [2] Bioscience Division, Oak Ridge National Laboratory, Oak Ridge, TN
[3] The Jackson Laboratory, Bar Harbor, ME

Introduction

Genetic reference populations are panels of related mouse strains with fixed genotypes. Population structure is determined by breeding history including

number of generations of out crossing
randomization of genotype segregation
progenitor diversity.

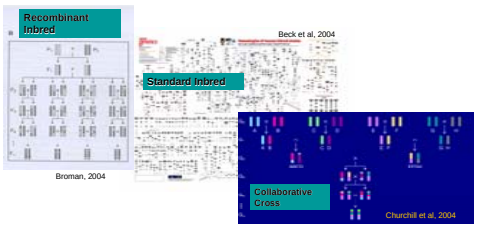
Linkage disequilibrium (LD) is a measure of statistical dependence between genetic markers. It depends on recombination frequency, genealogy of the populations, natural selection and other factors. Using large publicly available SNP sets, we applied graph analysis to compare the structure of multiple populations and sub-populations of mice.

Discrete graphs are used to represent the LD measures and graph algorithms such as clique enumeration or paraclique are applied to find completely or highly connected sets.

LD graph properties provide a quantitative comparison of populations and can be used to optimize genetic equidistance of sub-populations.

Data Sources

We studied two different mouse populations which are widely used as experimental populations. A set of 68 inbred mice consisting of several subpopulations (Beck et al, 2001), each one with a common origin and the other is the BXD RI population which has a more random mating history and three large sub-populations (Taylor, 1979, Williams et al, 1999; Peirce et al, 2004) that consists of 89 BXD strains of randomly segregating alleles. SNP genotypes were generated by Oxford-Illumina-CTC-Wellcome Trust for BXD samples provided by R. Williams, L. Lu, and SI lines from members of the Complex Trait Consorium.



BXD RI lines: 13368 SNPs
CTC-Oxford Wellcome Trust

Standard inbred lines: 11435 SNPs
CGD curated SNP set

Figure 1. Breeding of reference populations

Estimating Linkage Disequilibrium

To characterize the LD, we used three measures: *p*-values for Lewontin's *D'* (Lewontin, RC 1995), Pearson correlation coefficient (*r*) and mutual information coefficient (*MIC*). These three measures differ in the bias introduced by unbalanced allele frequencies among loci.

Lewontin's D'

$$D'_{XY} = \frac{D_{XY}}{D_{\max}} = \frac{p_{X_iY_i} - p_{X_i}p_{Y_i}}{D_{\max}},$$

$$D_{\max} = \begin{cases} \min(\, p_{X_i}p_{Y_0}, p_{X_0}p_{Y_i}), & if\ (D_{XY} > 0) \\ \min(\, p_{X_i}p_{Y_i}, p_{X_0}p_{Y_0}), & if\ (D_{XY} < 0) \end{cases}$$

Pearson's correlation

$$r_{XY} = \frac{p_{X_iY_i} - p_{X_i}p_{Y_i}}{\sqrt{p_{X_i}p_{Y_i}p_{X_0}p_{Y_0}}}$$

Mutual Information Coefficient

$$MIC_{XY} = p_{X_iY_i}\log_2\frac{p_{X_iY_i}}{p_{X_i}p_{Y_i}} + p_{X_iY_0}\log_2\frac{p_{X_iY_0}}{p_{X_i}p_{Y_0}} + p_{X_0Y_i}\log_2\frac{p_{X_0Y_i}}{p_{X_0}p_{Y_i}} + p_{X_0Y_0}\log_2\frac{p_{X_0Y_0}}{p_{X_0}p_{Y_0}}$$

Combinatorial Approaches

Once LD coefficients matrix are calculated from the SNP sets, a high-pass filter is applied incrementally to these metrics to construct unweighted graphs of LD associations. The filter consists of thresholds on each of the measures or the p-values from Fisher's Exact test. Maximal cliques (complete subgraphs, non-disjoint) or paracliques (densely-connected subgraphs, disjoint) were extracted across the range of thresholds and the resulting graphs were analysed for number, size, genome coverage, and chromosomal representation by the members. These analytic approaches provide a quantitative comparison of populations and can be used to optimize genetic equidistance of sub-populations.

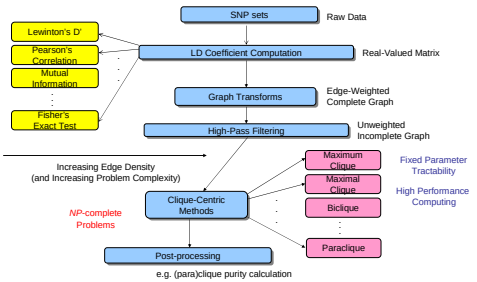


Figure 2. A overview of the graph theoretical-based model

Comparison of Properties of LD SNP Association Graphs in Populations

The number of vertices, edges and maximal cliques are larger at high LD thresholds in the BXD RI lines whether or not these measures were scaled for total graph size.

The graph size asymptotes at much larger LD (−0.75) for BXD while the graph continues expanding for inbred strains until much lower LD (−0.25).

The recombinant lines have more LD blocks observable at higher LD threshold than the inbred strains and the size is larger at each threshold.

The relatively constant genotype clique size across LD thresholds in the inbred strains indicates that lowering the threshold does not add vertices to cliques, but rather allows observation of additional LD blocks and networks of the same size.

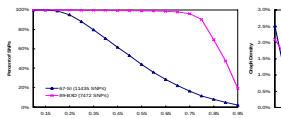


Figure 3.1

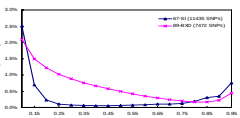


Figure 3.2

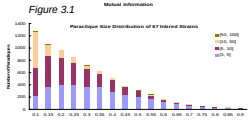


Figure 3.3

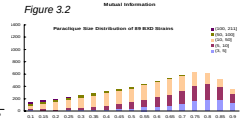
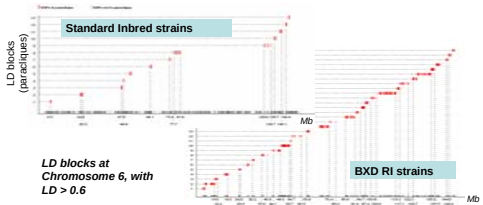


Figure 3.4

Figure 3. Properties of LD SNP association graphs for BXD and Inbred populations and LD thresholds ranging from 0 (low linkage) to 1 (high linkage) using Mutual Information Coefficient: (1) the number of vertices in the graph expressed as a percent of total number of SNPs in population (2) the number of edges in the graph expressed as a percent of total edges in the population. (3&4) number of paracliques at each LD threshold in SI and BXD RI.



Physical Distribution of LD Networks

To analyze the graph structure of LD blocks, we measured the clique purity as a number of the chromosomes in each clique. Syntenic LD blocks consist of linked loci on the same chromosome and non-syntenic LD networks contain linked loci across different chromosomes. Syntenic LDs can be further divided into long-range (several Mb pairs) or short range (<1 Mb) linkage blocks.

Figure 4.3, 4.4 show that the RI lines LD blocks have relatively constant clique purity of 1 to 3 chromosome span, while Inbred strains have a gradual decrease of clear purity which is obvious for LD < 0.3.

The existence of long-range syntenic LD or non-syntenic LD across populations or after many generations of random mating may be an indication of functional relationships of genes inside those blocks, it may also be due to co-adaptive allele selection in a common selected pathway.

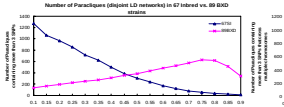


Figure 4.1

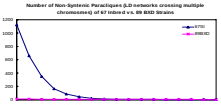


Figure 4.2

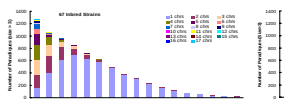


Figure 4.3

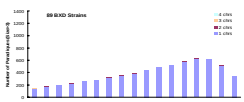


Figure 4.4

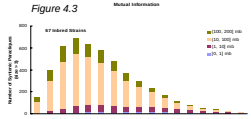


Figure 4.5

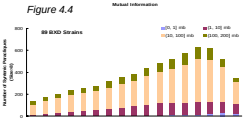
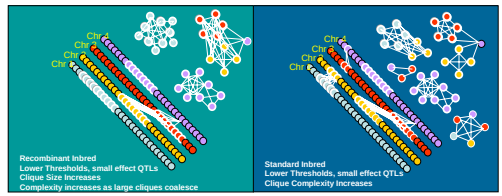
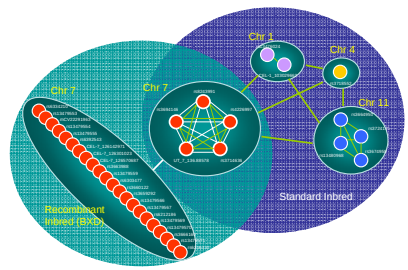


Figure 4.6

Figure 4. Distribution of LD networks of SI vs. RI: (1) the number of LD blocks, (2) the number of non-syntenic LD blocks, (3&4) LD blocks purity, (5&6) LD blocks size distribution.

Comparison of Standard and Recombinant Inbred Populations





Implications for Haplotype Associations in Reference Populations

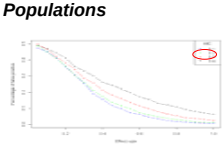


Figure 5.1

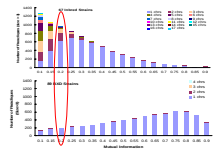


Figure 5.2

Figure 5. (1) A QTL is simulated by a mixture of two random distributions with means μ_0 and μ_1 and variance s which corresponds to the two alleles of a QTL. The phenotype distribution corresponds to a randomly drawn SNP. False positives are the percentage of non-syntenic significant SNP to phenotype associations. Note that in (2) at MIC of 0.20, for a QTL with a 0.3 standard deviation effect size, a true positive and single false positive are likely in BXD, whereas as many as 6 false positives may occur in the standard inbreds.

Conclusions

- The non-random breeding history of standard inbreds has resulted in the observation of non-syntenic linkage even at high LD thresholds while most of the LD blocks of random mated RI lines are short-range or long-range syntenic LD.
  - the genotype structure of standard inbred strains, which have had longer periods of outcrossing, consists of smaller blocks of linked loci than recombinant inbreds;
  - the non-random breeding history of standard inbreds has resulted in the infiltration of non-syntenic linkage at high LD thresholds.
- Long-range LD and the presence of LD blocks in mouse inbred strains create high rates of false positives in the SNP haplotype association analysis ("*in silico* QTL mapping"), despite the large size of the existing standard inbred strain set.
- Large syntenic LD blocks in the BXD recombinant inbred strain, though relatively uncorrelated with other genome regions, limit the power and precision of this population for genetic analysis.
- Combinatorial algorithms provide a convenient comparison of population architecture and suitability for association analysis
- Collaborative cross is predicted to have low long range LD, high precision due to many input haplotypes.
- LD observed in this population should largely reflect biological network selection.

Acknowledgements

NIGMS Center for Genome Dynamics to GAC & pilot to EJC, DOE ERKP804 to EJC, NIH BISTI to MAL

References

Beck et al 2001. Genealogies of mouse inbred strains. *Nature Genetics*.
Broman 2004. The genomes of recombinant inbred lines. *Genetics*.
Churchill et al, 2004. The collaborative cross, a community resource for the genetic analysis of complex traits, *Nature Genetics*.
Lewontin, RC 1995. The detection of linkage disequilibrium in molecular sequence data. *Genetics*.
Peirce et al, 2004. A new set of bxd recombinant inbred lines from advanced intercross populations in mice. *BMC Genetics*.