Linear Model Genome Scans for Expression QTL Analysis

Shirng-Wern Tsaih, Hyuna Yang, Renhua Li, Ioannis M. Stylianou, Keith DiPetrillo,Beverly Paigen, Gary Churchill

The Jackson Laboratory , Bar Harbor, Maine 04609 USA

Abstract

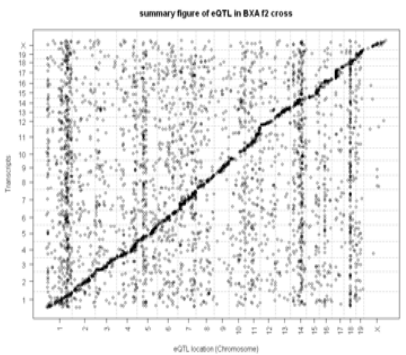
Large-scale data integration and data analysis requires systematic approach. We illustrate genome scanning strategies using high density lipoprotein (HDL) as trait of interest to explore associations in expression data from a mouse BXA intercross population to identify polymorphic genetic loci and transcript abundance associated with variation in traits. A defining feature of transcript abundance as a quantitative trait is that the phenotype is identified with a specific genetic locus, usually the gene that gives rise to the transcript. This distinctive property of transcript abundance, in some cases, allows us to identify the gene responsible for the effect of QTL. Using transcript abundance together with classical phenotype and genotype data, one can establish causal models/hypotheses to test the interaction of small or moderate numbers of transcripts, phenotypic traits and genetic loci. Liner models developed in this paper can be used to interpret these complex relationships. Especially, we demonstrate how to use these linear models in combination with the concept of conditional independence to distinguish a spurious association created by linkage from a true association.

Mice and Expression data

The BXA data set. **An F2 population consisting of 383 mice was constructed from two inbred strains of mice, C56BL/6J and A/J. Only male mice were maintained in this population. This cross was designed to study kidney disease (Doorenobos et al submitted) and the mice were on a chow diet (4% fat). Plasma HDL was measured as previously described and QTL results of HDL has been reported in Stylianou et al (2006). Mice were sacrified at 11 weeks of age. Expression profiling of liver tissue were available for a subset of 120 mice.**

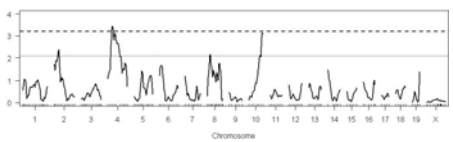
We re-mapped Affy probe using the custom CDF file (Dai 2005) from Brain Array (http://brainarray.mbni.med.umich.edu/Brainarray). Based on custom CDF file, perferct match intensities were normalized and summarized by RMA (Irizarray et al, 2003). Normalized gene expression intensities of 16,578 transcripts were used to illustrate our genome scan strategy.

Figure 1: Summary results of eQTL in BXA F2 cross local vs distant QTL

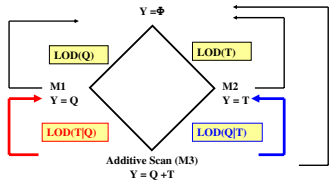


We used FWER-FDR procedure to obtain the significance level of genome scans. This approach can correctly control the number of false discovered transcripts having at least one linked QTL.

Figure 2 : Genome scan figure of HDL



Search loci (Q) and transcripts (T) that affect a phenotype of interest (Y) with three genome scans and 4 essential LOD scores:



Standard Genome Scan (M1). **Standard genomic model at locus Q can be represented as a linear model. We simplify the notation and denote it as Y = Q. The location of the putative QTL is scanned across the genome in increments (2cM) to identify associations (Lander and Botstein 1989). We compute a LOD score, LOT(Q), by contrasting M1 to the null model (Y =Φ).**

Trait Correlation Scan (M2). **The problem of identifying transcripts that are associated with variation in Y can be recast as a genome scan. We denote it as Y = T. We compute a LOD score, LOT(T), by contrasting M2 to the null model (Y =Φ).**

Additive Genome Scan (M3). **We denote additive scan as Y = T + Q. The first conditional LOD socre LOD(Q|T) accounts for the effect of the transcript. The second conditional LOD score, LOD(T|Q) accounts for the effect of the genetic effect.**

Based on the pattern of these 4 LOD scores, we conclude:

- **T ← Q → HDL : LOD(T) > 2.1, LOD(T|Q) = 0**
- **Q → T → HDL : LOD(T) > 2.1, LOD(T|Q) ≠ 0, LOD(Q) for T > 2.1, LOD(Q|T) << LOD(Q) for HDL**
- **Q → HDL ← T : LOD(T) > 2.1, LOD(T|Q) ≠ 0, LOD(Q) for T < 2.1, LOD(Q|T) >> LOD(Q) for HDL**

Figure 3 : HDL vs intermediate transcripts

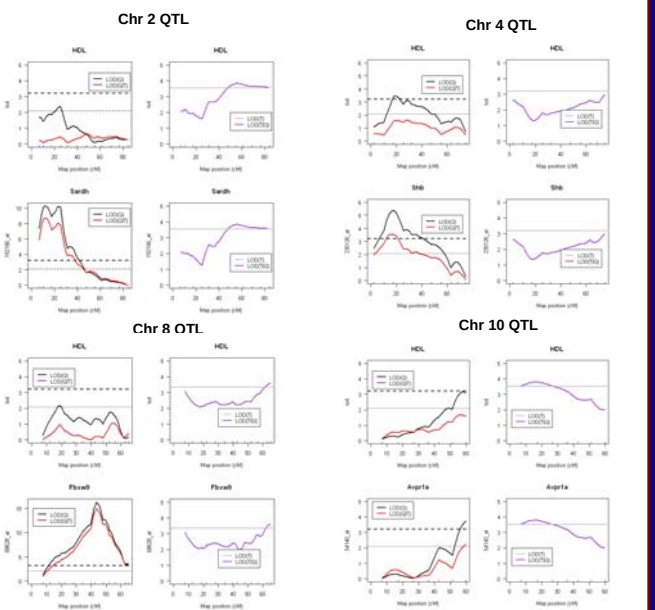
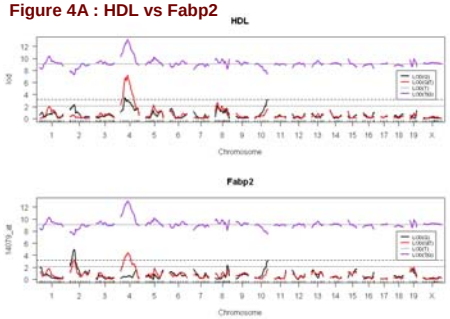


Figure 4A : HDL vs Fabp2



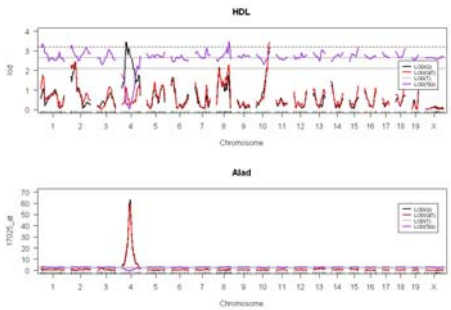
Conditioning on Fabp2 affects chr 2, 4 and 8 peaks for QTL. Based on the pattern of these 4 LOD score, we conclude that

Chr 2:
Q →Fabp2 →HDL

Chr 4:
Q → HDL ← Fabp2

Chr 10: Either
Q →Fabp2 →HDL, or
Q → HDL → Fabp2 is possible. We need additional information to conclude the direction.

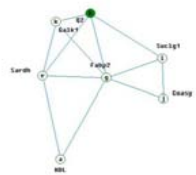
Figure 4B : HDL vs Alad



Conditioning on Alad only affects chr 4 peak for QTL. Based on the pattern of these 4 LOD score, we conclude that

Chr 4:
Alad ← Q →HDL

Figure 5: Graphical model of HDL. chr 2 OTL and intermediate transcripts



Conclusions

- **The obvious limitation of our model is including one QTL and one transcript at a time. It may miss important additive or interactive effect which can be only revealed by including all related pieces. One can extend the models, however, model space complexity and computational time should be considered.**
- **Our procedure also depends on experimental, measurement and model errors which may lead to incorrect conclusion of the direction of the relationship between HDL and transcripts.**
- **We build graphical model including all transcripts that we identified using genome scan strategy. This integrated network shows how the genetic effect, intermediate transcripts and interesting trait can be connected to each other. This two-step approach, i.e. select potential transcripts first and build the network, has a big advantage in a model searching procedure.**