Finding Genome-Transcriptome-Phenome Association with Structured Association Mapping and Visualization in GenAMap

Junming Yin, Ross Curtis, Eric P. Xing

Lane Center for Computational Biology, Carnegie Mellon University

Overview

Most of previous GWAS have considered only genome-phenome associations.

However, in the increasingly common scenario where expression, phenotype, and genomic data are available from the same cohort, it is important to integrate the expression data directly into the primary analysis. **It has the potential to**

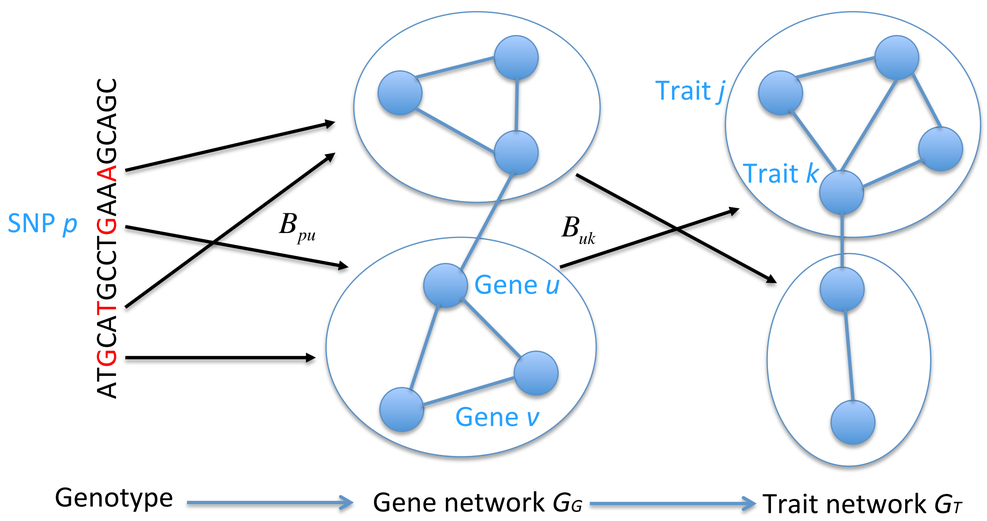
- reveal the functional relationships between associated genomic variations and physical phenotypes, via intermediate phenotypes.
- elucidate the biological mechanism behind the genome-phenome associations and uncover potential pathways to target for treatment of diseases.

We present

- a novel structured association mapping strategy for finding three-way associations.
- a visual analytics software GenAMap (http://sailing.cs.cmu.edu/genamap) that automates structured association mapping algorithms and provides visualizations to explore the complex results.

Structured Association Mapping

Association mapping by exploiting the rich structures in the transcriptome and phenome, e.g., gene-expression network, trait network.



Genome-transcriptome associations by graph-guided fused lasso (GFlasso): encourage highly correlated genes (connected by an edge in the gene network) to be associated with the same set of SNPs.

$$B_1 = \text{argmin}_B \|Y - XB\|_F^2 + \lambda \sum_u \sum_p |B_{pu}| + \gamma \sum_{(u,v)\in E_G} \sum_p |B_{pu} - \text{sign}(\rho_{uv}) B_{pu}|$$

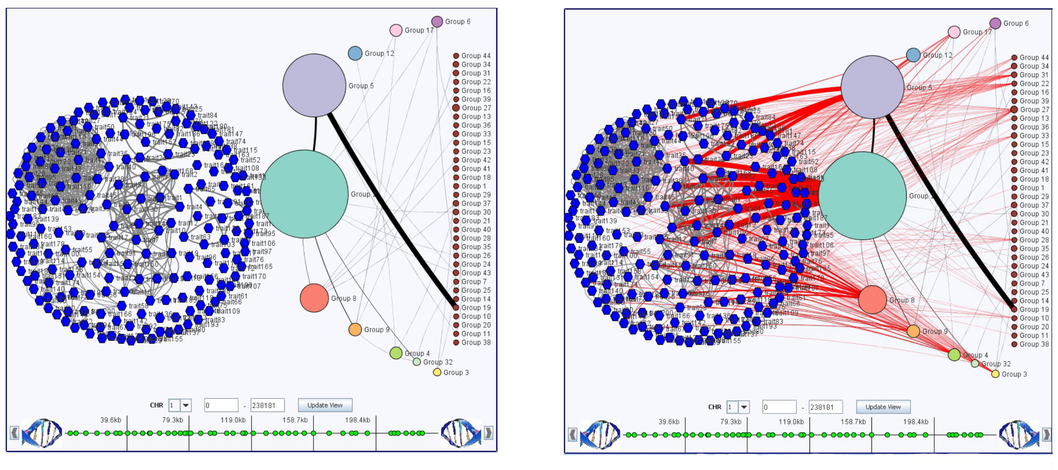
Transcriptome-phenome associations by graph-graph-guided fused lasso (gGFlasso): highly correlated genes tend to have similar influences the same subsets of traits.

$$B_2 = \text{argmin}_B \|Z - YB\|_F^2 + \lambda \sum_k \sum_u |B_{uk}| + \gamma_1 \sum_{(j,k)\in E_T} \sum_u |B_{uk} - \text{sign}(\rho_{jk}) B_{uj}|$$

$$+\gamma_2 \sum_{(u,v)\in E_G} \sum_k |B_{uk} - \text{sign}(\rho_{uv}) B_{vk}|$$

Both problems are *convex* (no local minimum) and can be solved by state-of-the-art optimization algorithms.

Visualization - GenAMap



(a) (b)

Figure: The structured (edges) of the traits (blue hexagons) and gene groups (circles) are displayed with (b) and without (a) the association edges (red edges). SNPs are represented at the bottom of GenAMap's genome browser.

Simulation Study

We simulated datasets with $N = 250$ individuals, $P = 100$ SNPs, $J = 500$ genes, and $K = 20$ traits at different noise levels. GFlasso-gGFlasso outperformed three baseline methods in terms of TPR and FPR.

	$\sigma_1^2 = \sigma_2^2 = 1/4$		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2 = 4, \sigma_2^2 = 16$	
	TPR	FPR	TPR	FPR	TPR	FPR
$\boldsymbol{B_1}$ by GFLasso	0.9454	0.0060	0.8965	0.0057	0.7884	0.0092
$\boldsymbol{B_1}$ by Lasso	0.9758	0.7632	0.9535	0.0763	0.9081	0.7528
$\boldsymbol{B_2}$ by gGFLasso	1.0000	0.0000	0.9333	0.0016	0.7067	0.0205
$\boldsymbol{B_2}$ by GFLasso	0.9800	0.0004	0.9233	0.0039	0.7000	0.0207
$\boldsymbol{B_3}$ by GFlasso-gGFlasso	0.9200	0.0266	0.8600	0.0305	0.8400	0.2450
$\boldsymbol{B_3}$ by GFlasso-GFlasso	0.9200	0.0266	0.8600	0.0615	0.8400	0.2500
$\boldsymbol{B_3}$ by Lasso-Lasso	1.0000	0.8803	1.0000	0.9030	1.0000	0.8559
$\boldsymbol{B_3}$ by PLINK	0.6300	0.0150	0.5357	0.0234	0.5000	0.0294

NIH Heterogeneous Stock Mice Data Analysis

943 SNPs were found to be associated with 746 genes, and 412 of these genes were associated with 133 traits. We found that 604 SNP-trait associations were also recovered by PLINK. This suggests that GFlasso-gGFlasso can help to explain some of these signals, in addition to the newly discovered signals.

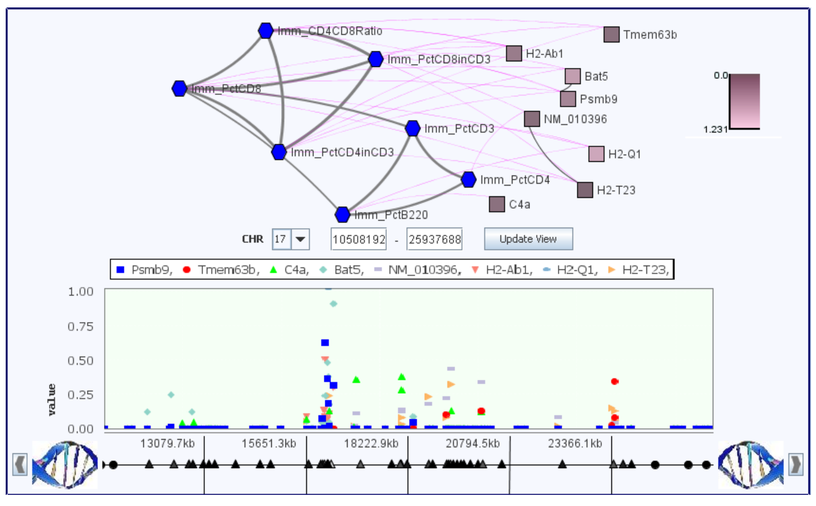


Figure: An interesting association between Chr 17, H2 genes, and immunology traits.

References

[1] R. Curtis, J. Yin, P. Kinnaird, E. P. Xing, Pac Symp Biocomput. (2012), 327-38.

[2] S. Kim and E. P. Xing, PLoS Genet 5(8), e1000587 (2009).