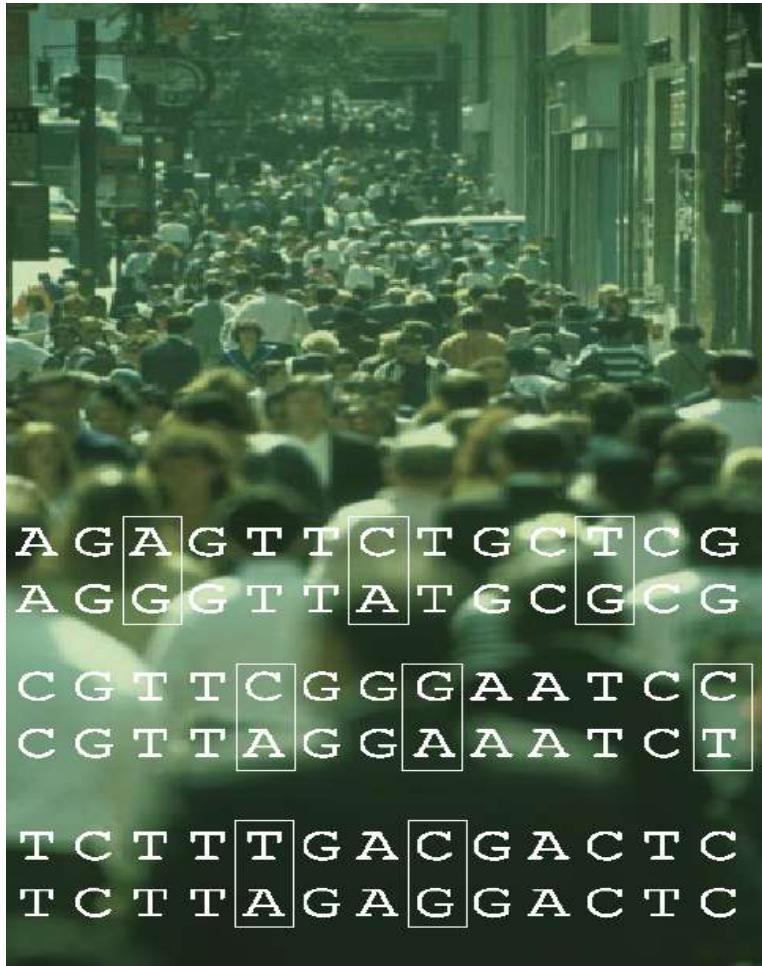


On Modern Statistical Methods for Genetic Association Study: Structured Genome-Transcriptome-Phenome Association Analysis

Eric P. Xing and Seyoung Kim
`{epxing, sssykim}@cs.cmu.edu`

Machine Learning Department
Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University

Genome Polymorphisms



The ABO Blood System				
Blood Type (genotype)	Type A (AA, AO)	Type B (BB, BO)	Type AB (AB)	Type O (OO)
Red Blood Cell Surface Proteins (phenotype)				
Plasma Antibodies (phenotype)	A agglutinogens only	B agglutinogens only	A and B agglutinogens	No agglutinogens

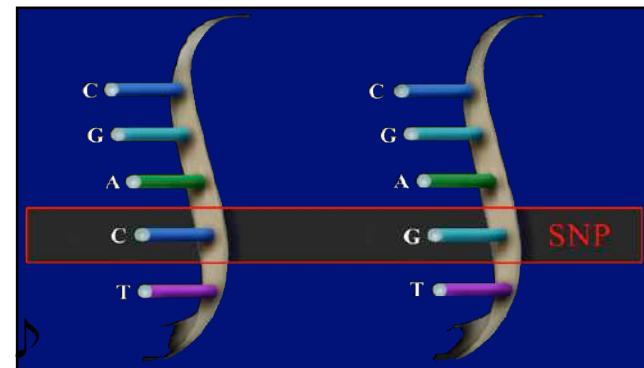
AGAGTTCTGGTGG
AGGTTTATGCGCGG
CGTTTCGGGAATCC
CGTTAGGAAATCT
TCTTTGACGACTC
TCTTAGAGGACTC

AGAGTTCTGGTGG
AGGTTTATGCGCGG
CGTTTCGGGAATCC
CGTTAGGAAATCT
TCTTTGACGACTC
TCTTAGAGGACTC

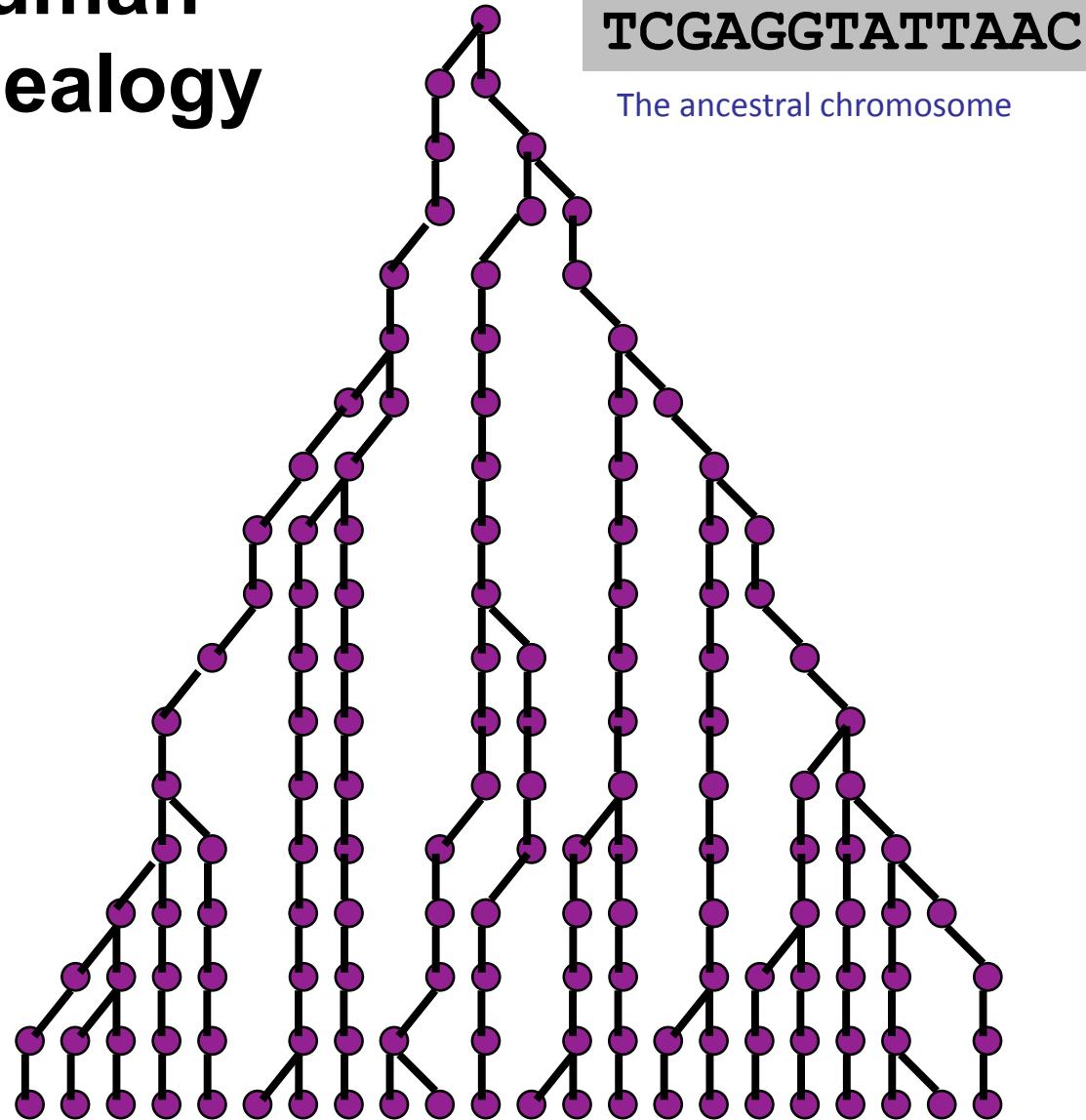


Type of Polymorphisms

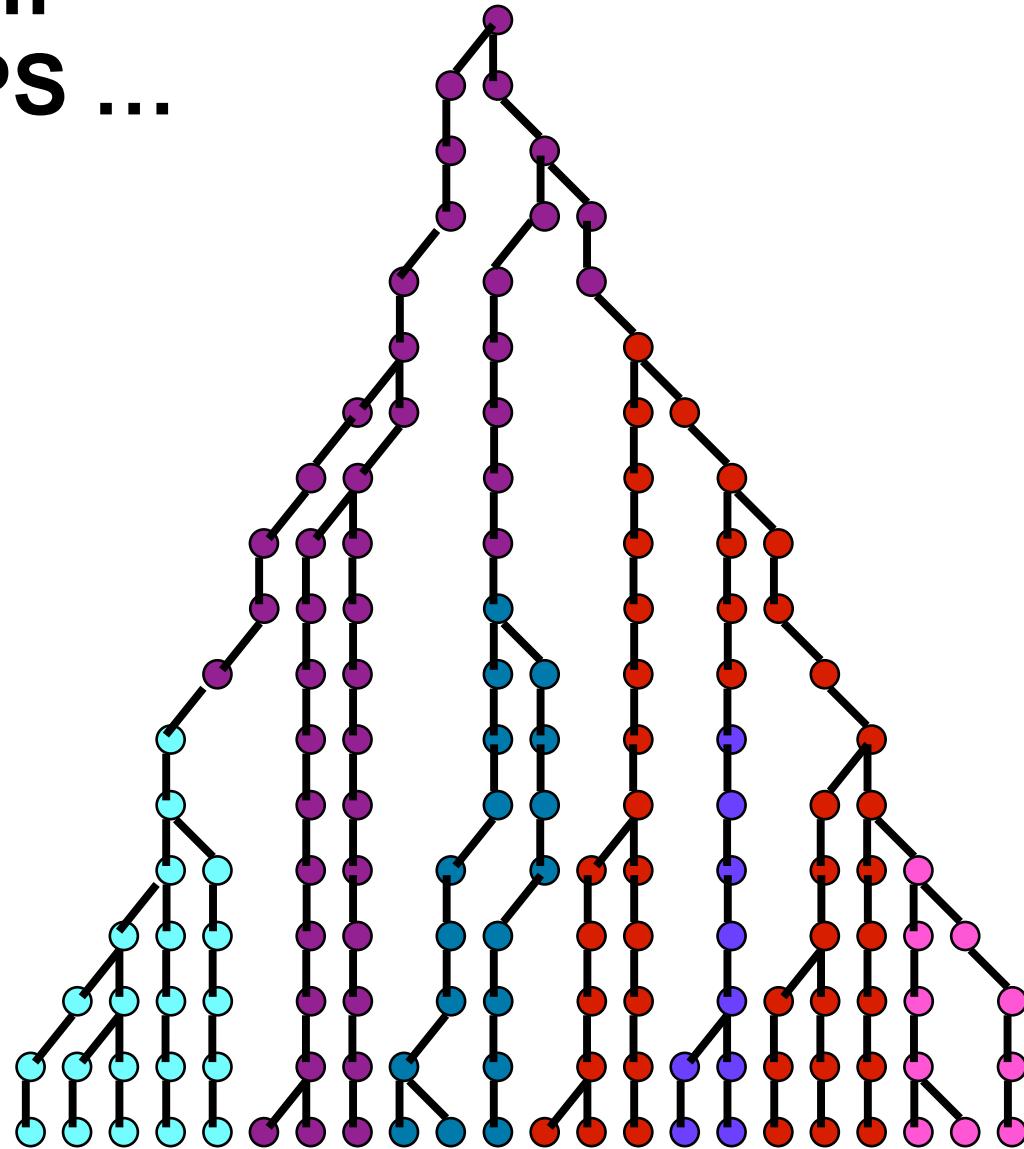
- Insertion/deletion of a section of DNA
 - Minisatellites: repeated base patterns (several hundred base pairs)
 - Microsatellites: 2-4 nucleotides repeated
 - Presence or absence of Alu segments
- **Single Nucleotide Polymorphism (SNP):**
 - DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
 - Frequency of SNPs greater than that of any other type of polymorphism
 - Each variant is called an “allele”
 - Almost always bi-allelic
 - Account for most of the genetic diversity among different (normal) individual, e.g. drug response, disease susceptibility



A Human Genealogy



From SNPs ...

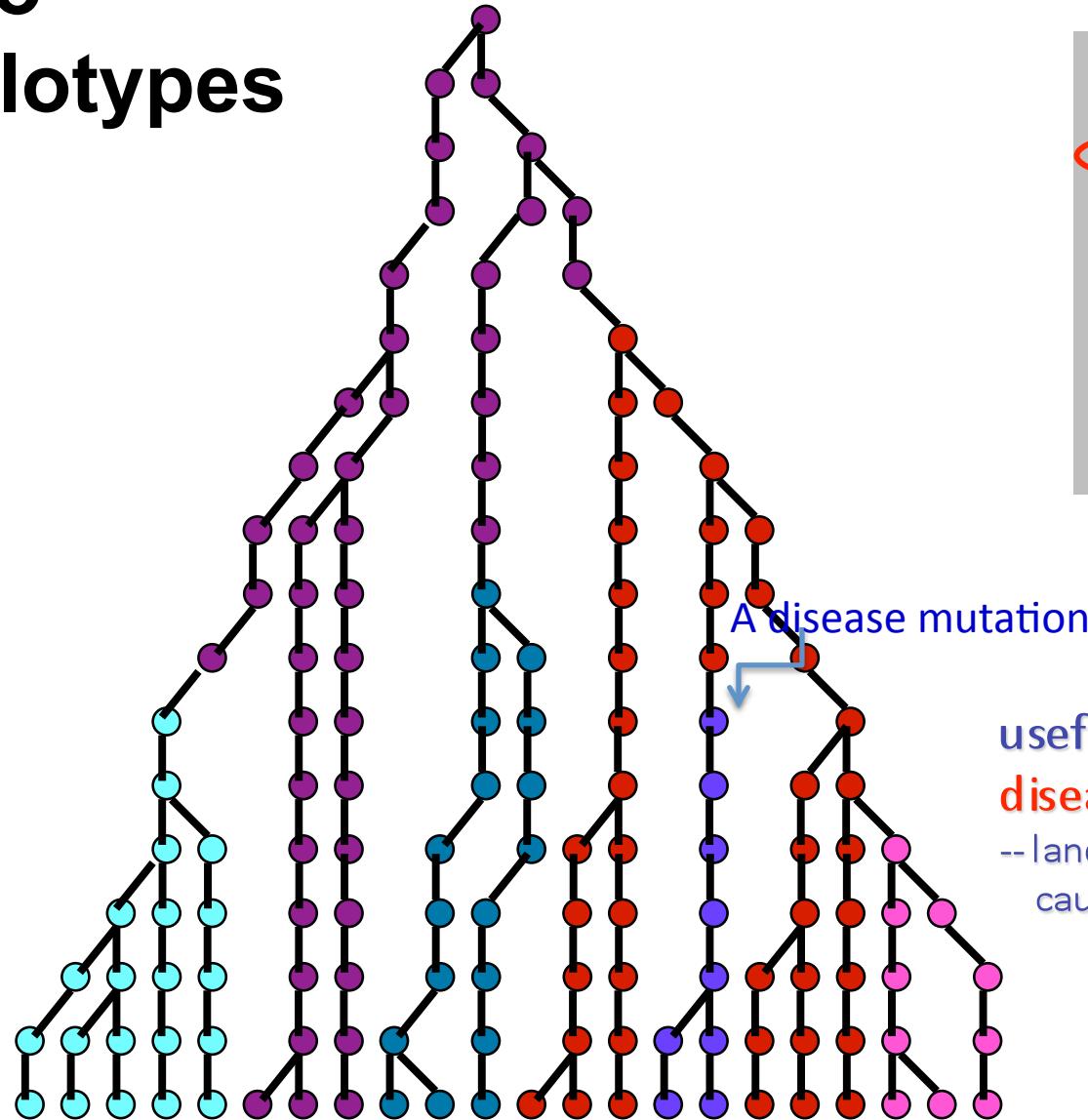


TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC

* ** * *

The SNPs

... To Haplotypes



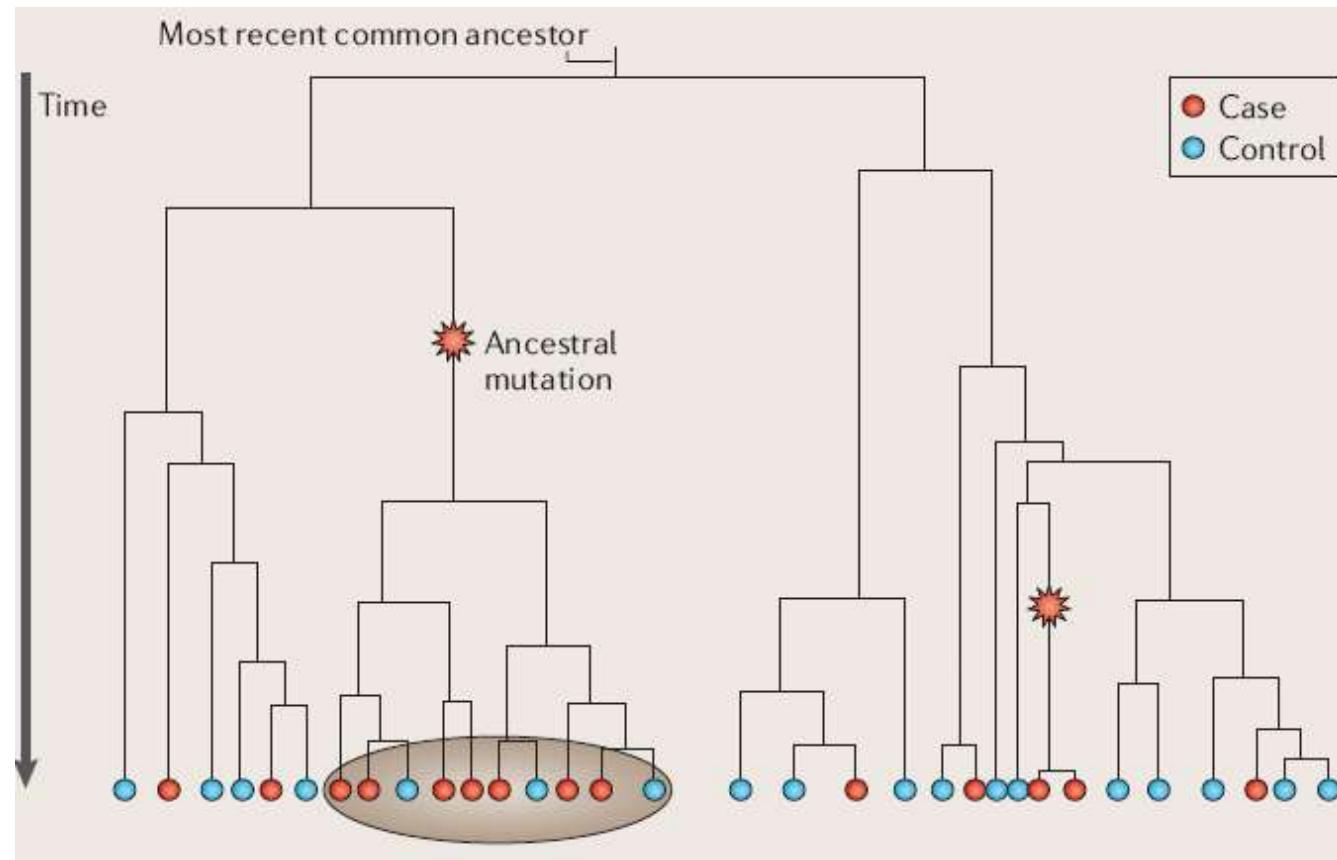
TC	AGG	T	A	C
TC	AGG	T	A	C
TC	AGG	T	A	C
TC	AGG	T	A	C
TC	AGG	T	A	C

The haplotypes

useful markers for studying
disease association
-- landmarks, indicators, co-variates,
causes ...

Population-Based Association Study

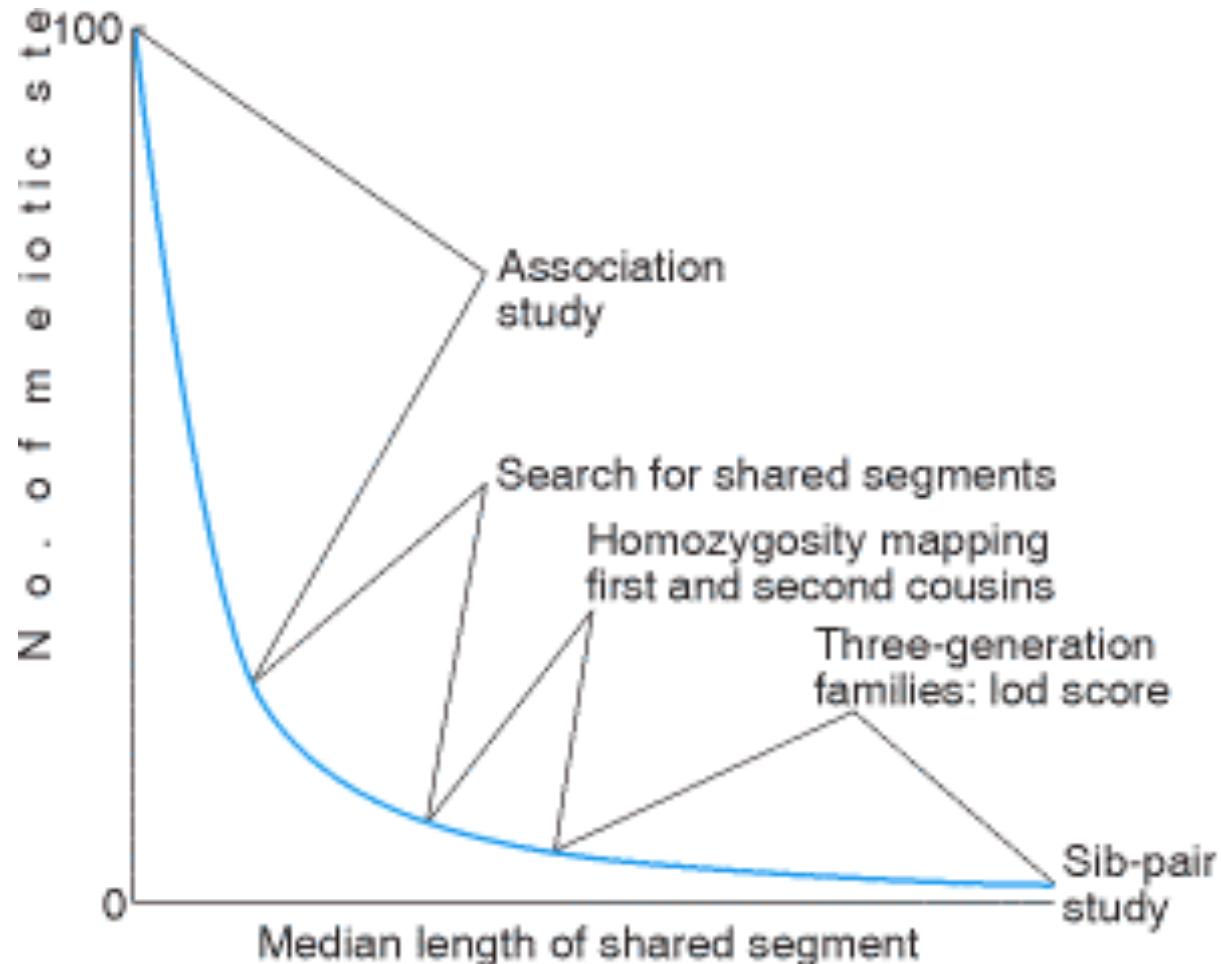
- Case/control data are collected from unrelated individuals
 - All individuals are related if we go back far enough in the ancestry



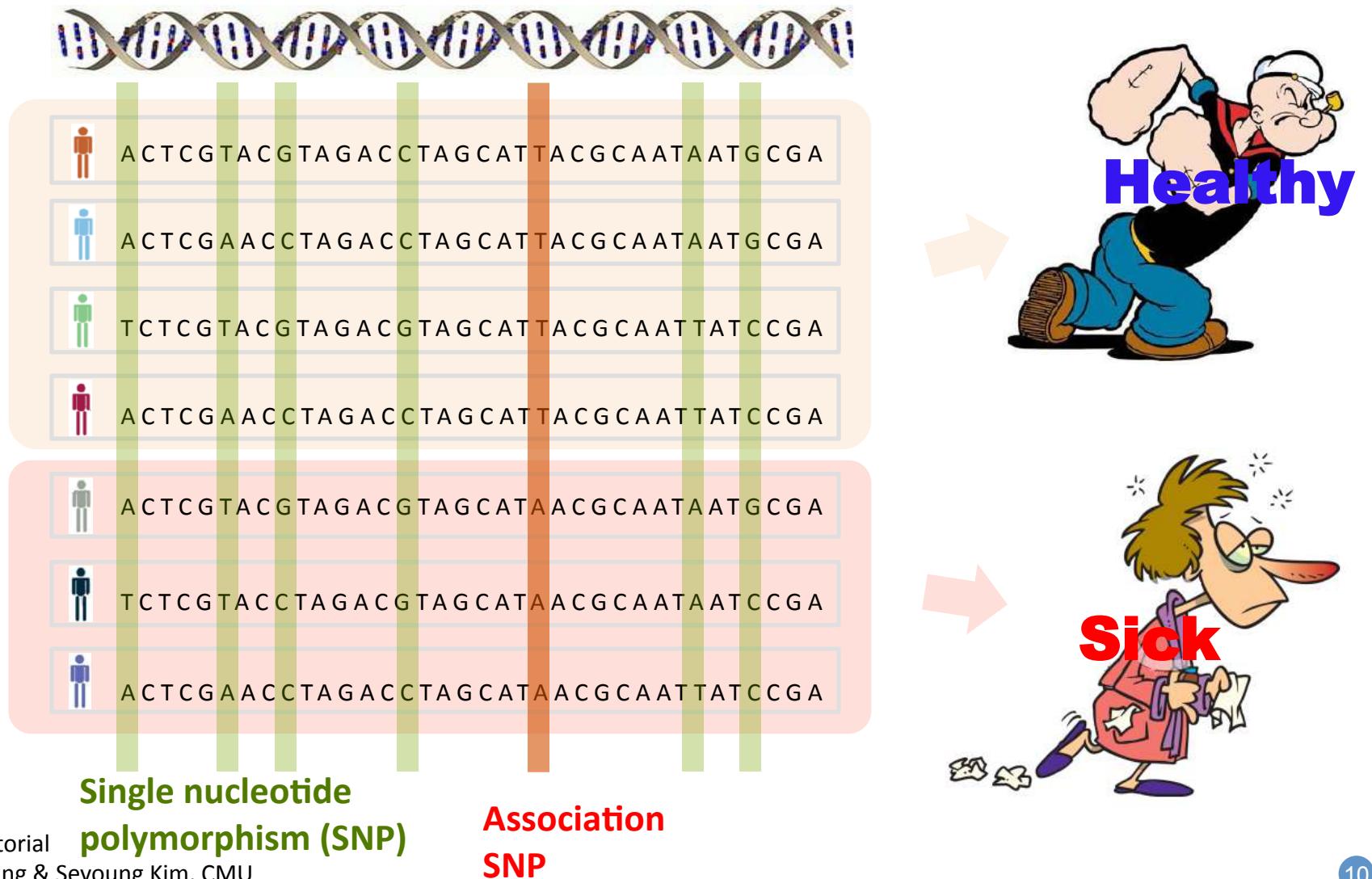
Advantages of SNPs in Genetic Analysis of Complex Traits

- Abundance: high frequency on the genome
- Position: throughout the genome
 - coding region, intron region, promoter site
- Ease of genotyping
- Less mutable than other forms of polymorphisms
- SNPs account for around 90% of human genomic variation
- About 10 million SNPs exist in human populations
- Most SNPs are outside of the protein coding regions
- 1 SNP every 600 base pairs
- More than 5 million common SNPs each with frequency 10-50% account for the bulk of human DNA sequence difference
- It is estimated that ~60,000 SNPs occur within exons; 85% of exons are within 5 kb of the nearest SNP

Linkage Analysis vs. Association Analysis



Genetic Basis of Diseases

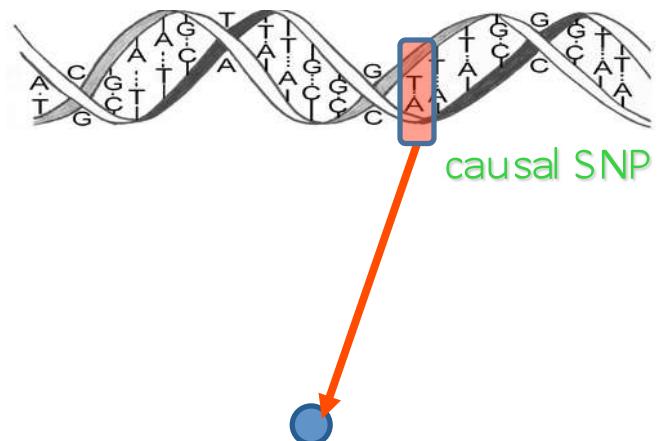


Genetic Association Mapping

Data

	<u>Genotype</u>	<u>Phenotype</u>
1	A.....T..G.....C.....T.....A..G.	
2	A.....A..C.....C.....T.....A..G.	
3	T.....T..G.....G.....T.....T..C.	
4	A.....A..C.....C.....T.....T..C.	
5	A.....T..G.....G.....A.....A..G.	
6	T.....T..C.....G.....A.....A..C.	
7	A.....A..C.....C.....A.....T..C.	

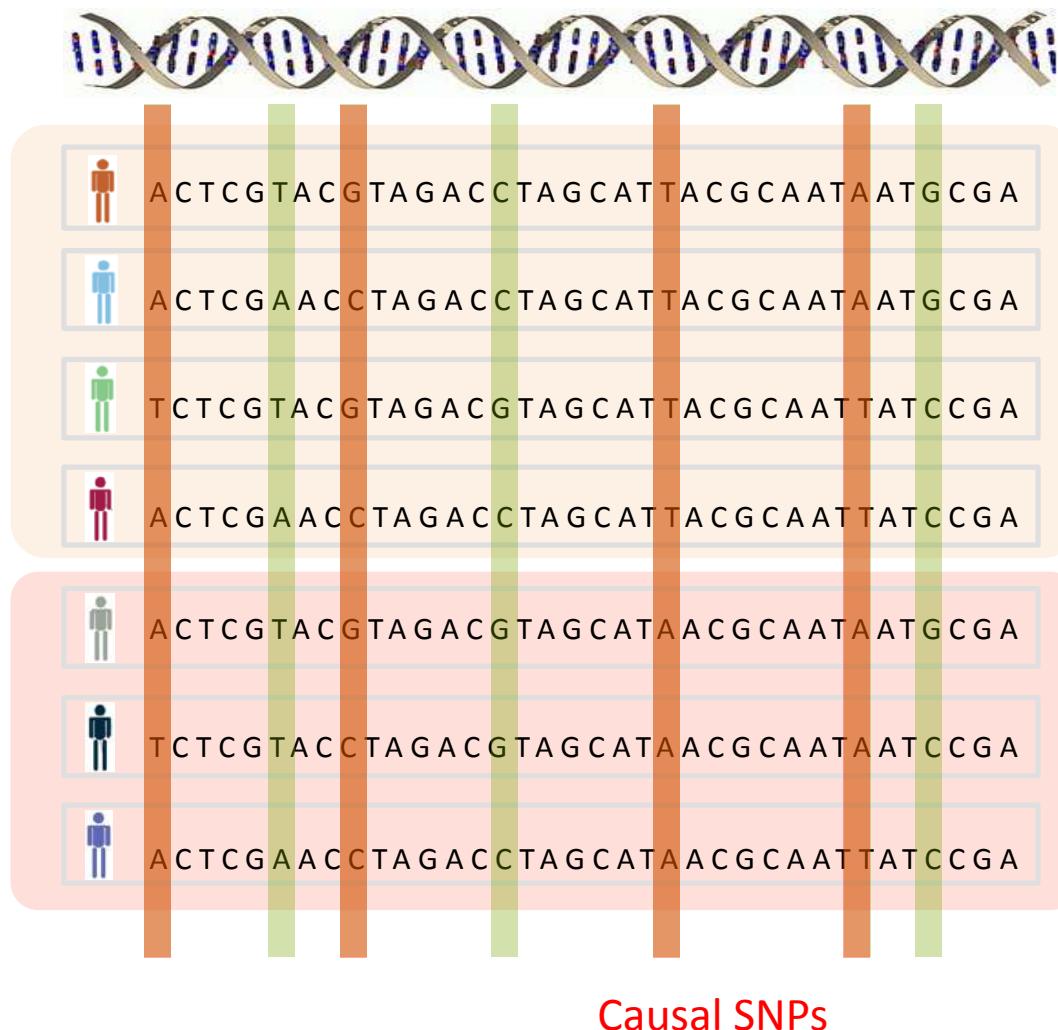
Standard Approach



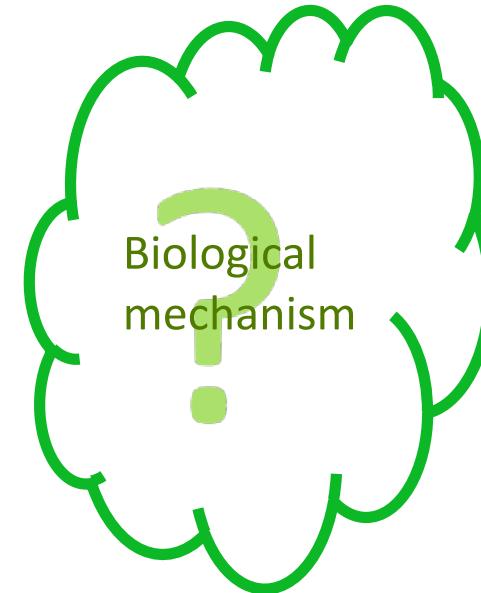
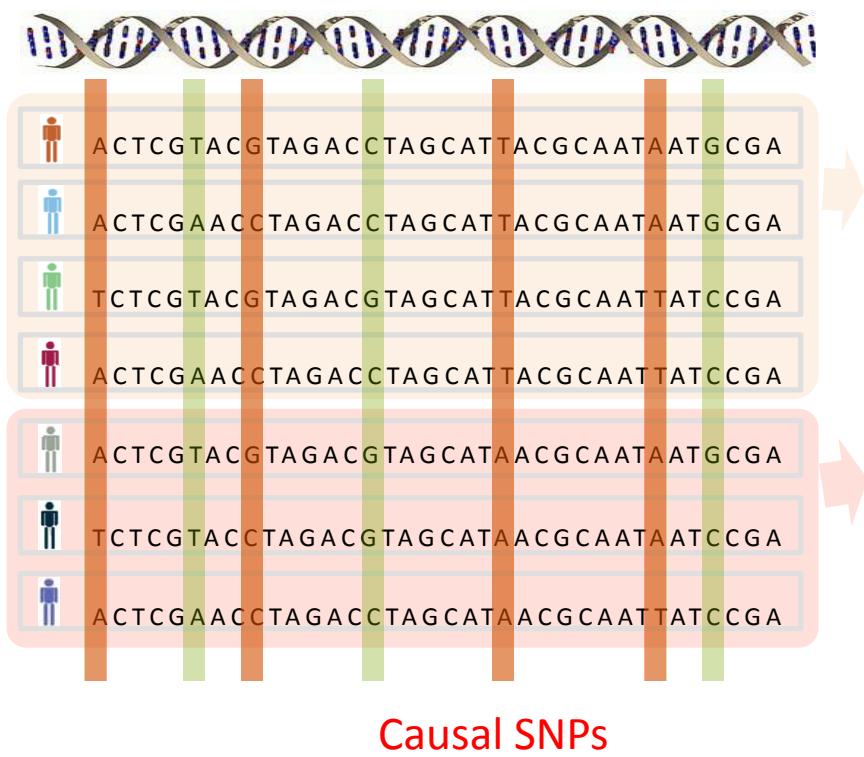
a univariate phenotype:
e.g., disease/control

- Cancer: Dunning et al. 2009.
- Diabetes: Dupuis et al. 2010.
- Atopic dermatitis: Esparza-Gordillo et al. 2009.
- Arthritis: Suzuki et al. 2008

Genetic Basis of Complex Diseases

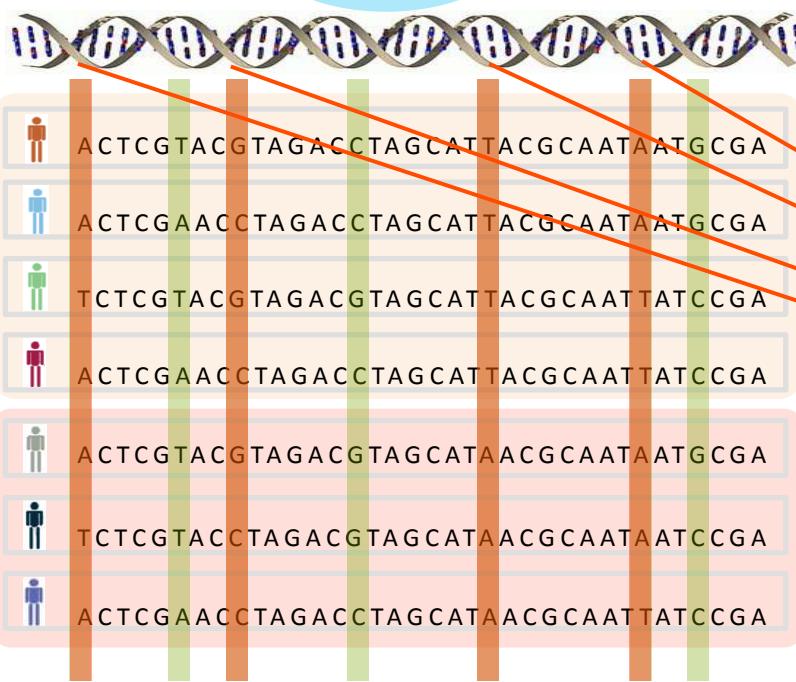


Genetic Basis of Complex Diseases



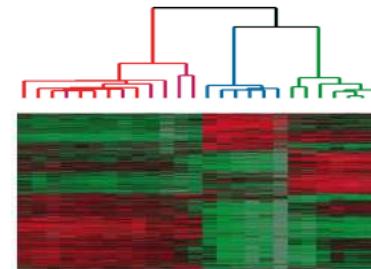
Genetic Basis of Complex Diseases

Association to intermediate phenotypes



Intermediate Phenotype

Gene expression



Clinical records



Population association analysis

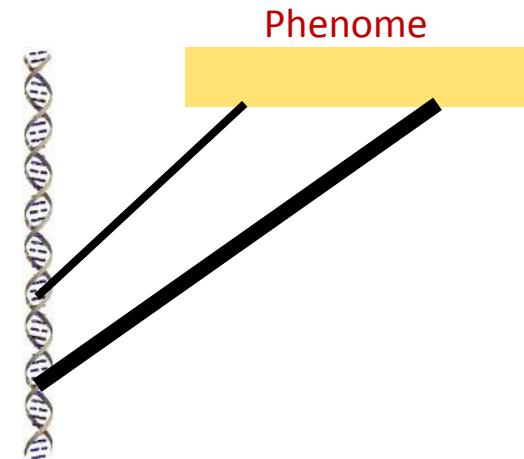
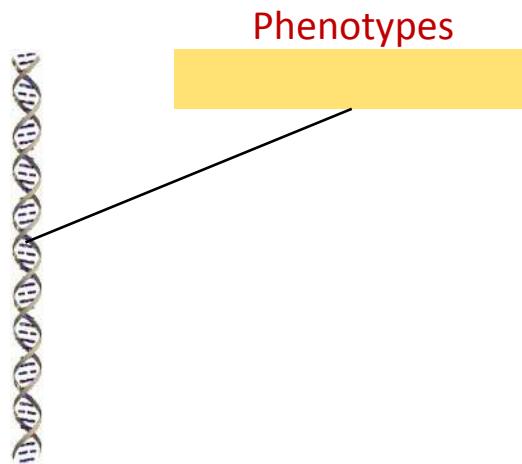
Standard Approach

Consider
one phenotype at a
time

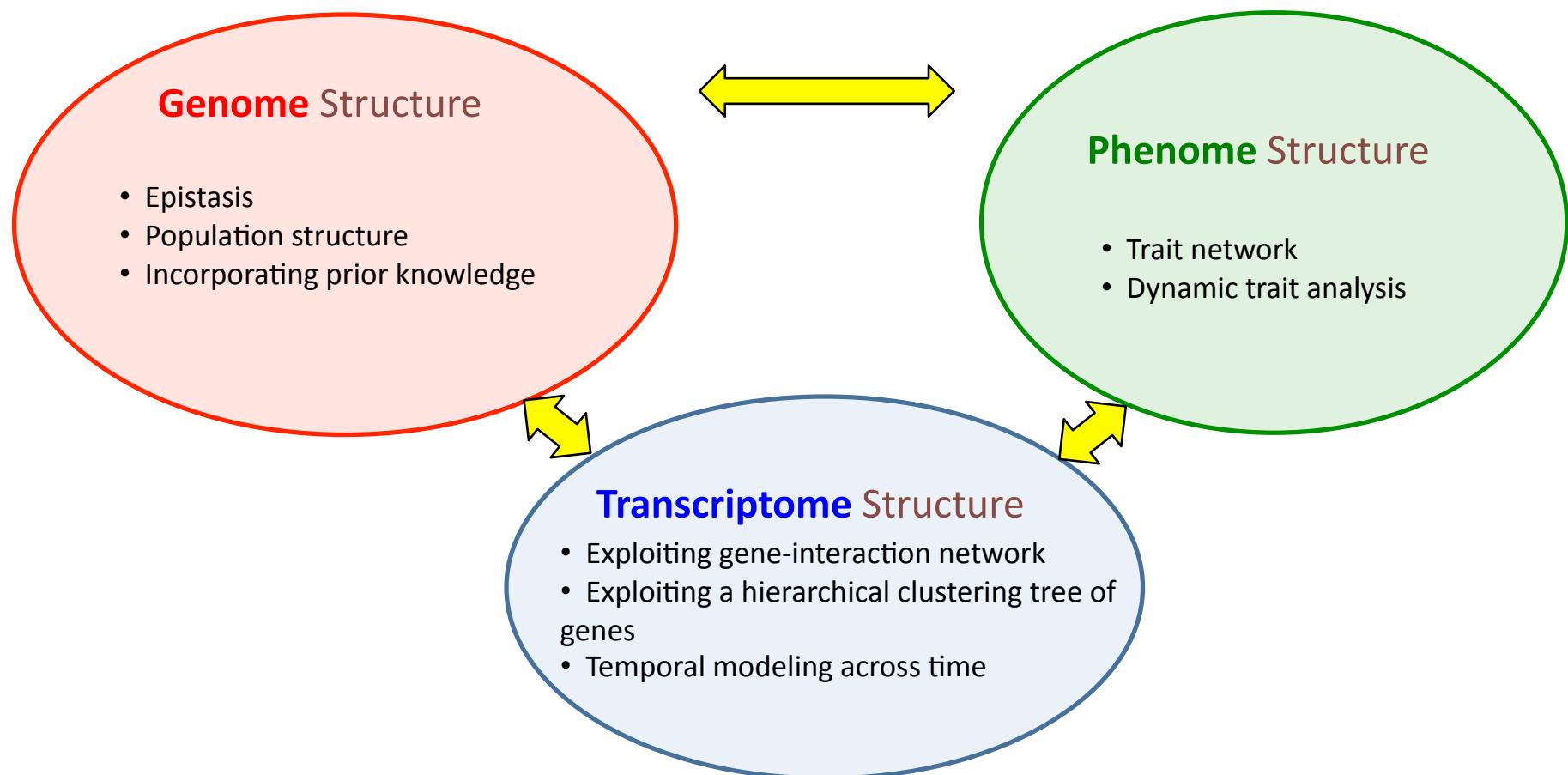
VS.

New Approach

Consider **multiple
correlated phenotypes
(phenome) and genotypes
(genome)** jointly

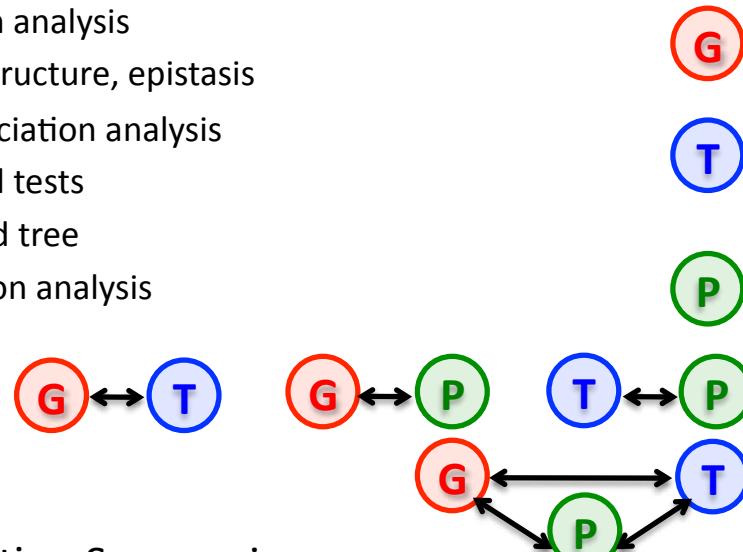


Summary: Structured Genome-Transcriptome-Phenome Association Analysis



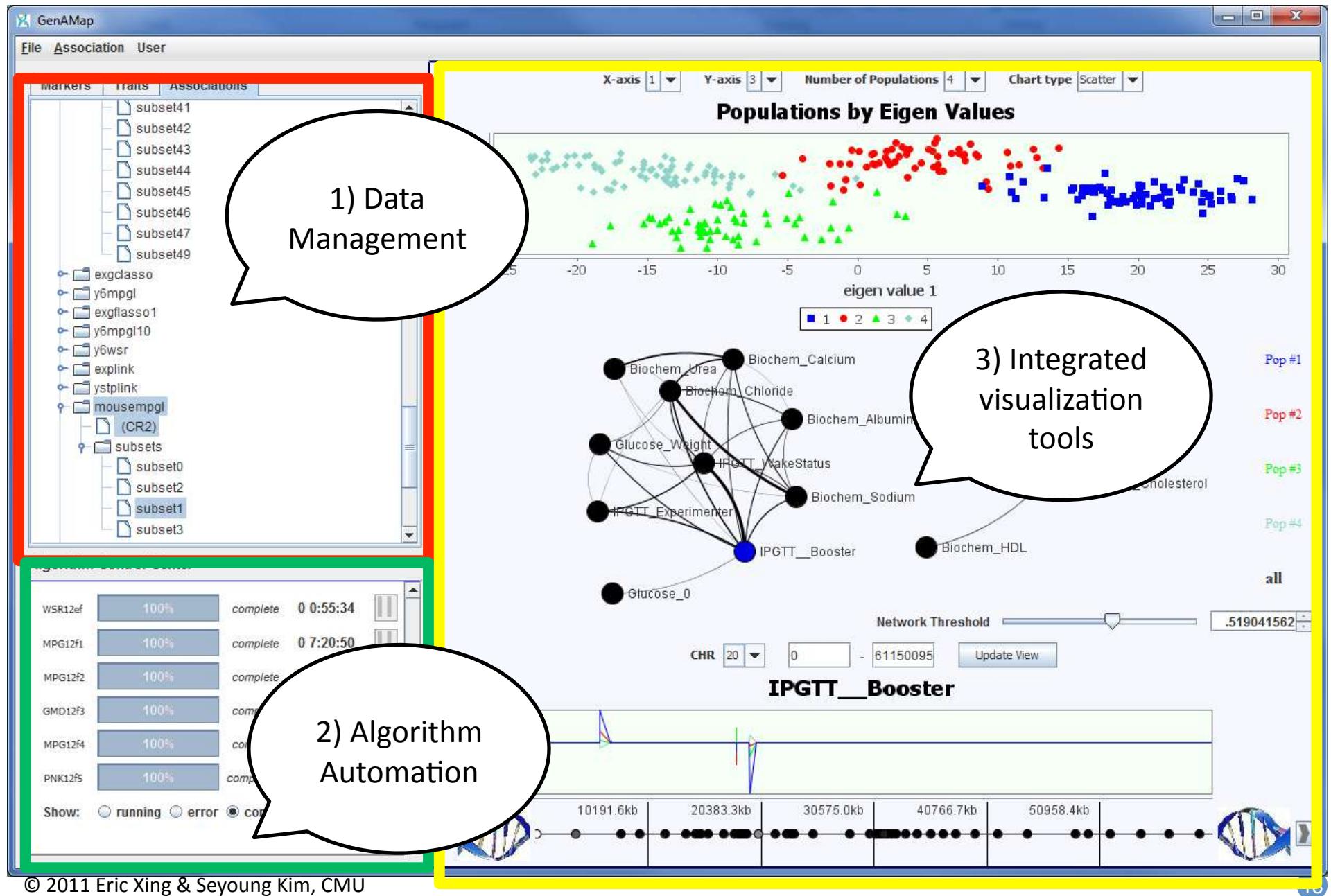
Outline

- Preparing Data for Association Analysis
 - Missing genotypes/phenotypes, tag SNP selection, haplotype inference, population structure
- Background on Association Analysis: Single Phenotype Methods
 - Single SNP approach : case/control study, quantitative trait as phenotype
 - Simultaneous analysis of all SNPs: Sparse regression method
- Structured Genome-Phenome-Transcriptome Analysis
 - Exploiting **genome structure** in association analysis
 - Linkage disequilibrium, population structure, epistasis
 - Exploiting **transcriptome structure** in association analysis
 - Pleiotropy : Pathway-based statistical tests
 - Leveraging trait structures: graph and tree
 - Exploiting **phenome structure** in association analysis
 - Pleiotropy, dynamic-trait association
 - **Two-way** structured association
 - **Three-way** structured association
- Association Analysis and Next-generation Sequencing
 - Going beyond SNPs: Structural variants
 - From common variants to rare variants



GenAMap: Visual analytics software for structured association mapping

Curtis and Xing, Carnegie Mellon University



Part I

Preparing Data for Association Analysis

Overview

- Preparing genotype data
 - Haplotype inference
 - Tag SNP selection
 - How to handle missing genotype values
 - Population structure
- Preparing phenotype data
 - How to handle missing phenotype values
 - Data normalization
 - What structure?

Haplotype and Genotype

- A collection of alleles derived from the same chromosome

Genotypes

2 13
1 6
9 15
4 17
1 9
2 6
9 17
2 12
7 12
6 14
1 7
18 18
1 4
10 10

Haplotype
Re-construction

Haplotypes

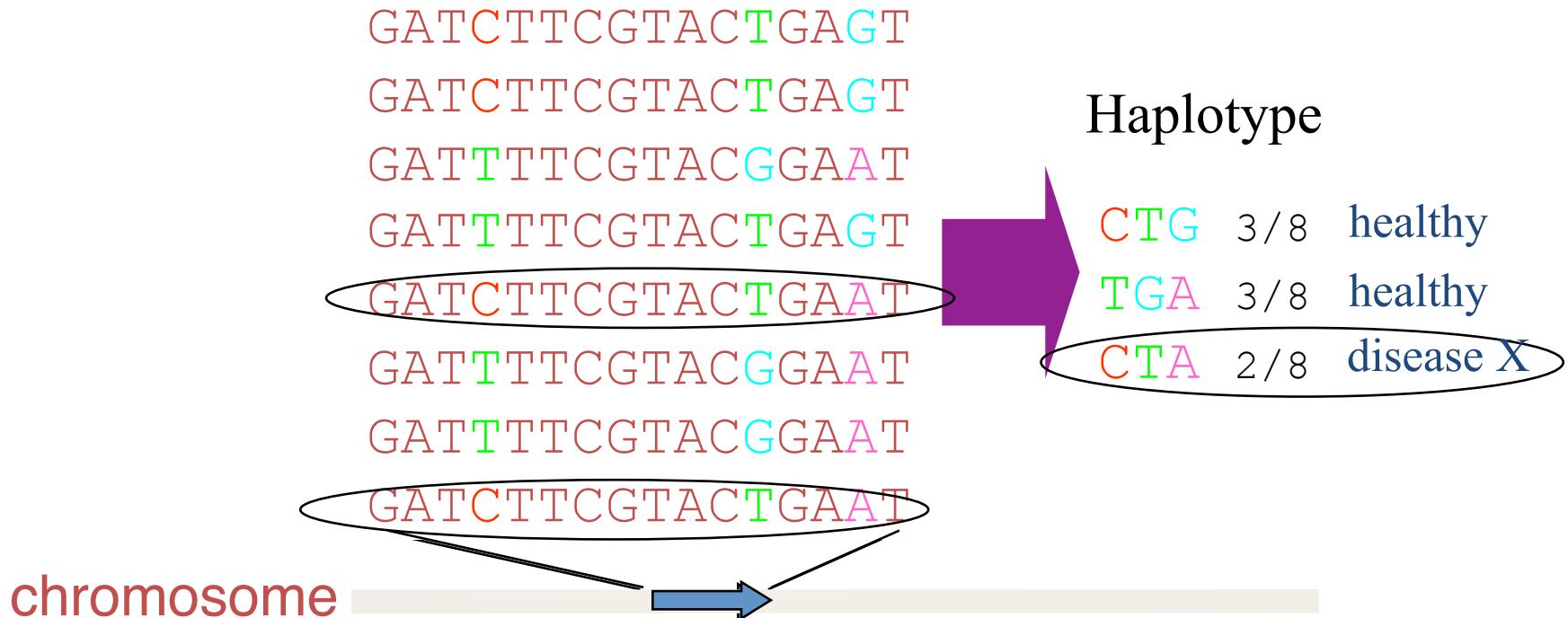
2 6 9 17 1 6 9 2 12 7 12 6 14 7 18 1 10
13 1 15 4 9 2 17 12 7 6 1 18 4 10

Chromosome phase is unknown

Chromosome phase is known

Why Haplotype in Association Mapping?

-- a more discriminative state of a chromosomal region



- Consider J binary markers in a genomic region
- There are 2^J possible haplotypes
 - but in fact, far fewer are seen in human population
- Good genetic marker for population, evolution and hereditary diseases ...

Haplotype Analyses

- Haplotype analyses
 - Linkage disequilibrium assessment
 - Disease-gene discovery
 - Genetic demography
 - Chromosomal evolution studies
- Why Haplotypes
 - Haplotypes are more powerful discriminators between cases and controls in disease association studies
 - Use of haplotypes in disease association studies reduces the number of tests to be carried out.

Linkage Disequilibrium

- LD reflects the relationship between alleles at different loci.
 - Alleles at locus A: frequencies p_1, \dots, p_m
 - Alleles at locus B: frequencies q_1, \dots, q_n
 - Haplotype frequency for A_iB_j :
 - equilibrium value: $p_i q_j$
 - Observed value: h_{ij}
 - Linkage disequilibrium: $h_{ij} - p_i q_j$
 - Linkage disequilibrium is an allelic association measure (difference between the actual haplotype frequency and the equilibrium value)
 - More precisely: **gametic association**
- Association studies.
 - If inheriting a certain allele at the disease locus increases the chance of getting the disease, and the disease and marker loci are *in LD*, then the frequencies of the marker alleles will **differ** between diseased and non-diseased individuals.

Haplotype Inference

- Given a random sample of multilocus genotypes at a set of SNPs
 - Frequency estimation of all possible haplotypes
 - Haplotype reconstruction for individuals
 - How many out of all possible haplotypes are plausible in a population
- Haplotype reconstruction algorithm
 - Clark's parsimony algorithm (Clark, Mol. Biol. Evol. 1990)
 - Haplotype (Niu et al., AJHG 2002)
 - PHASE (Stephens et al., AJHG 2001)

PHASE

(Stephens et al., AJHG 2001)

- Treat unknown haplotypes as unobserved random quantities and estimate the conditional probability of haplotypes given observed genotypes
- Gibbs sampling approach
 - Start with initial guesses on haplotypes
 - Iteratively reconstruct the haplotype of each individual assuming the haplotypes of other individuals have been correctly reconstructed.

PHASE

(Stephens et al., AJHG 2001)

- Given the genotype data G , start with some initial haplotype reconstruction $H^{(0)}$. For $t = 0, 1, 2, \dots$, obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:
 1. Choose an individual, i , uniformly and at random from all ambiguous individuals (i.e., individuals with more than one possible haplotype reconstruction).
 2. Sample $H_i^{(t+1)}$ from $P(H_i | G, H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i .
 3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j=1,\dots,n, j \neq i$

Reducing Genotyping Costs with Tag SNPs

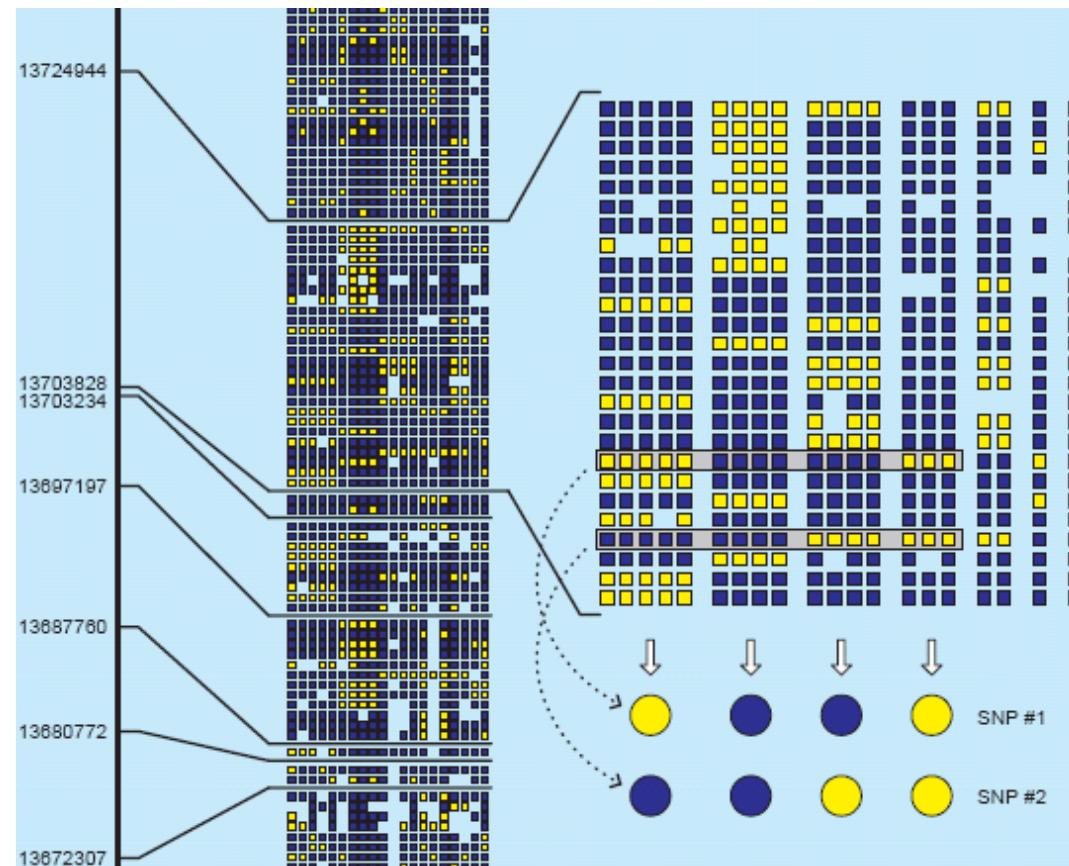
- Nearby SNPs in the genome are in linkage disequilibrium (LD), and thus contain redundant information.
- If we knew which SNPs are in LD, we can pre-select the representative SNPs for each LD block of chromosome, and genotype only for those SNPs.

Two-phase Genotyping Using Tag SNPs

- A two-phase approach
 - Phase 1: Collect a set of SNPs densely distributed across genome for a small number of individuals. Select tag SNPs based on this dataset.
 - Phase 2: Genotype only for tag SNPs for a large number of individuals

Tag SNP Selection

- Tag SNPs summarize information across multiple SNPs in linkage disequilibrium



Algorithm for Finding Tag SNP

- Problem: Find a set of tag SNPs that cover all of the non-tag SNPs in LD ($r^2 > \alpha$) with the tag SNPs
 - α : parameter that needs to be specified by the user (e.g., 0.8)
- Greedy search (Carlson et al., AJHG 2004)
 - Repeatedly select the SNP that cover the largest number of non-tag SNPs given α
- Dynamic programming approach (Zhang et al., PNAS 2002)

What is population structure?

- Population Structure
 - A set of individuals characterized by some measure of genetic distinction
 - A “population” is usually characterized by a distinct distribution over genotypes
 - Example

Genotypes  aa  aA  AA



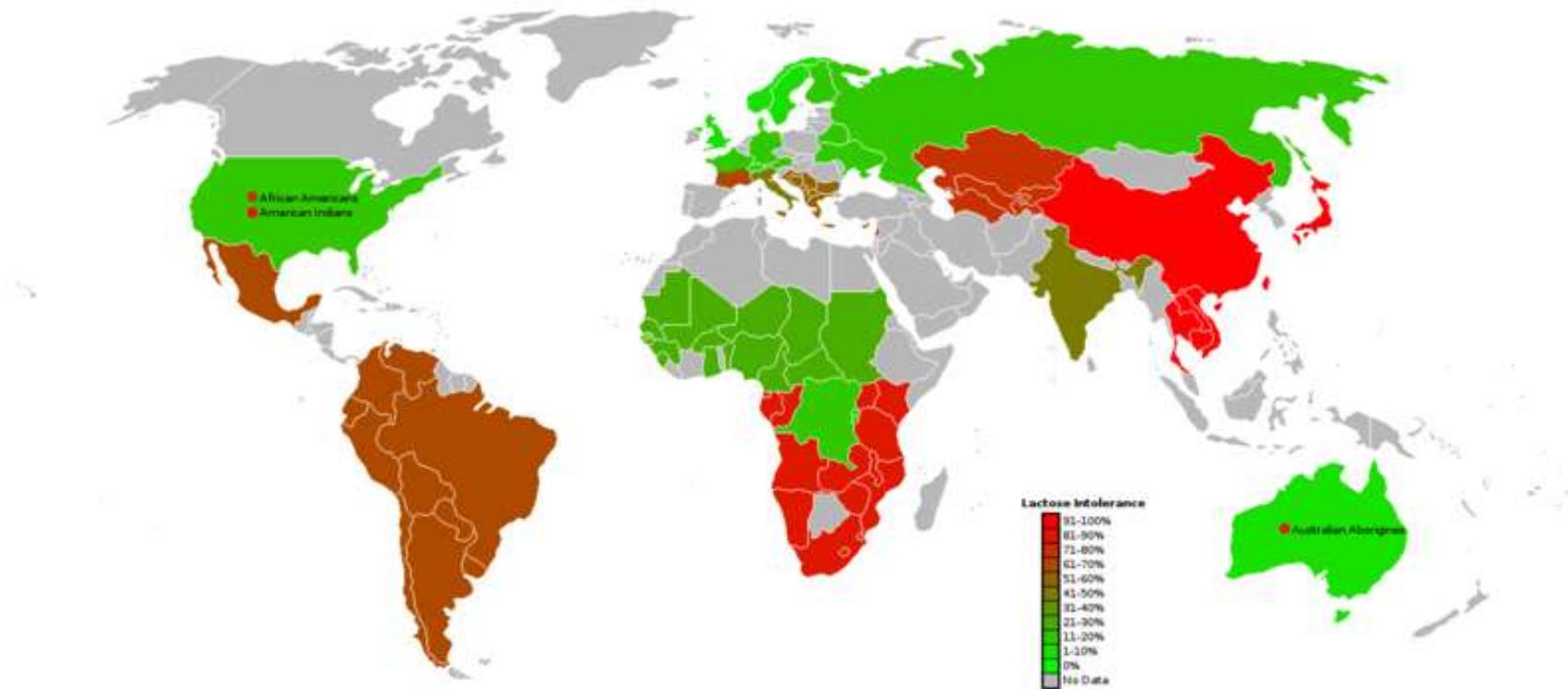
Population 1



Population 2

Motivation

- **Reconstructing individual ancestry:** The Genographic Project
 - <https://genographic.nationalgeographic.com/genographic/index.html>
- **Studying human migration**



Methods for Learning Population Structure from Genetic Markers

- Low-dimensional projection
 - PCA-based methods (Patterson et al., PLoS Genetics 2006)
- Clustering
 - Distance-based (Bowcock et al., Nature 1994)
 - Model-based
 - STRUCTURE (Pritchard et al., Genetics 2000)
 - mStruct (Shringarpure & Xing, Genetics 2008)

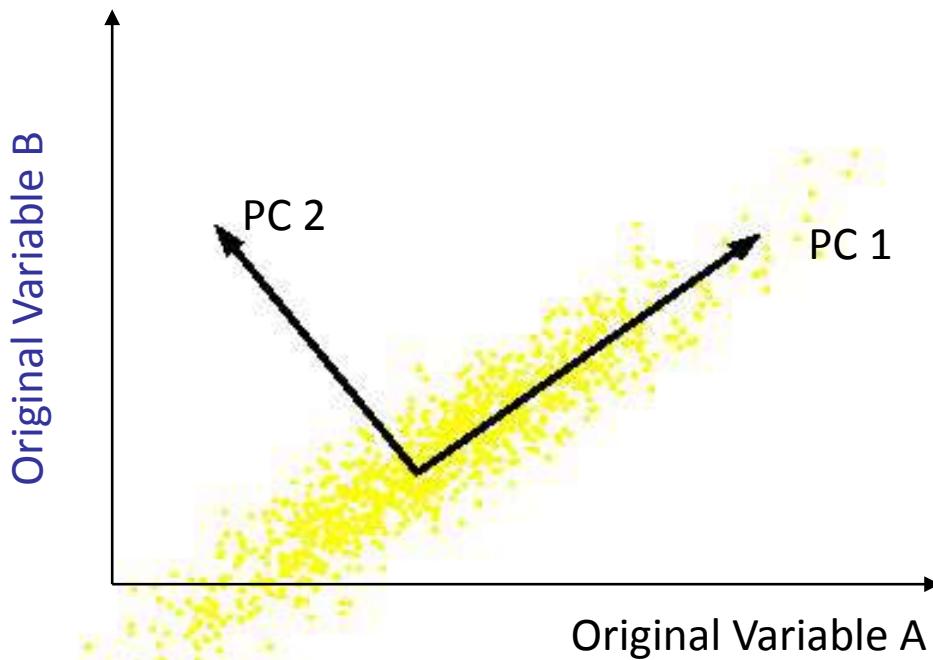
Low-dimensional Projections

- Genetic data is very large
 - Number of markers may range from a few hundreds to hundreds of thousands
 - Thus each individual is described by a high-dimensional vector of marker configurations
 - A low-dimensional projection allows easy visualization
- Technique used
 - Factor analysis
 - Many statistical methods exist – ICA, PCA, NMF etc.
 - Principal Components Analysis (next slide)
- Allows projection of individuals into a low dimensional space
- Usually projected to 2 dimensions to allow visualization

Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions ...
 - Are linear combinations of the original ones
 - Are uncorrelated with one another
 - Orthogonal in original dimension space
 - Capture as much of the original variance in the data as possible
 - Are called Principal Components
- Demo at <http://www.cs.mcgill.ca/~sqrt/dimr/dimreduction.html>

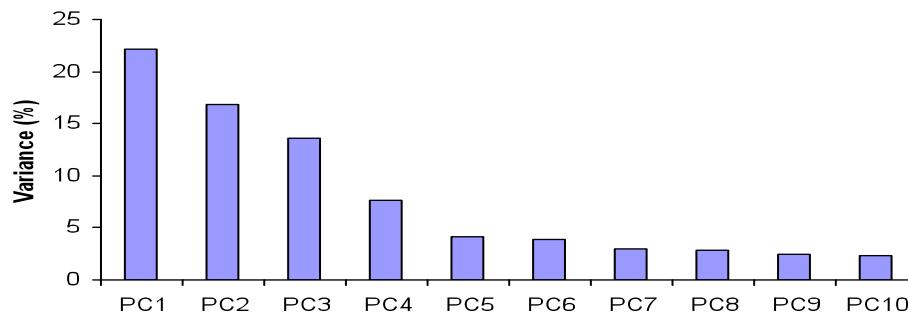
What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

Dimensionality Reduction

Can *ignore* the components of lesser significance.



You do *lose some information*, but if the eigenvalues are small, you don't lose much

- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

PCA Analysis

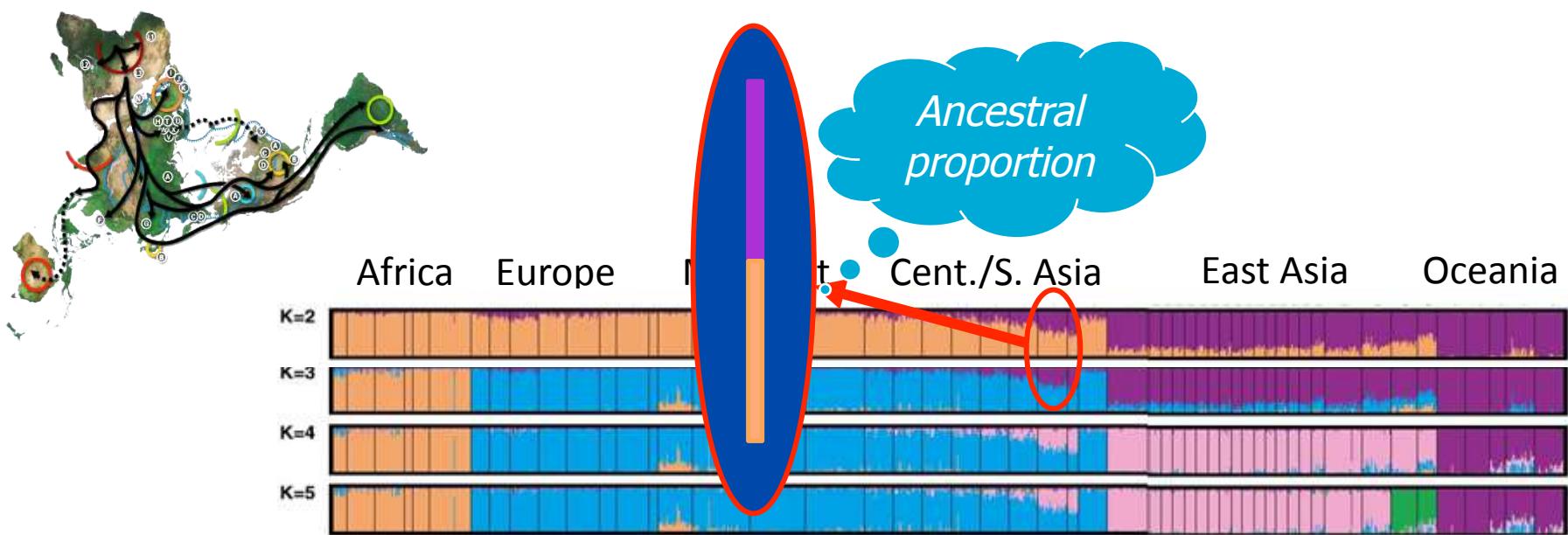
(Cavalli-sforza, 1978)

- Plot of geographical distribution of 3 PCs (Intensity proportional to value of each component)
 - First – blue
 - Second - green
 - Third - red



Model-based Clustering for Discovering Population Structure

- How to display population structure?
 - *Structure* (Pritchard et al., Genetics 2000)



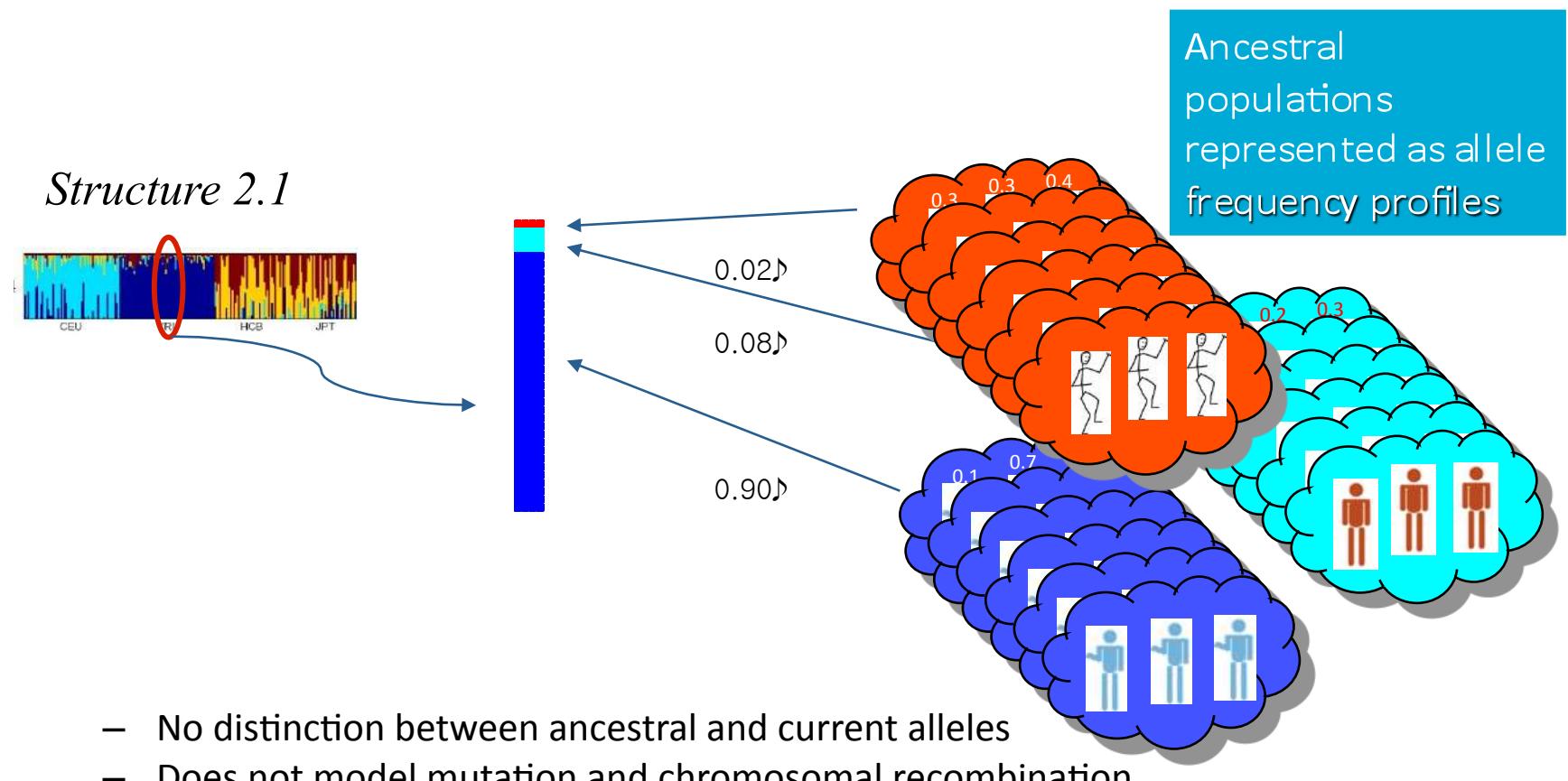
Genetic structure of Human Populations (Rosenberg et al., Science 2002)

Structure Model

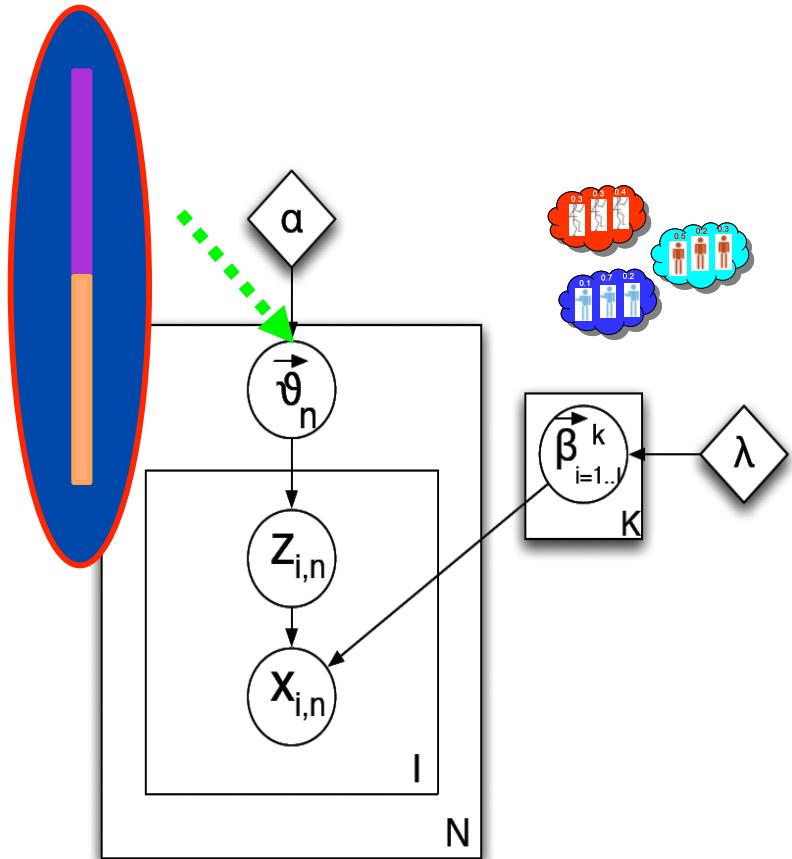
- Hypothesis: Modern populations are created by an intermixing of ancestral populations.
- An individual's genome contains contributions from one or more ancestral populations.
- The contributions of populations can be different for different individuals.
- Other assumptions
 - Hardy-weinberg equilibrium
 - No linkage disequilibrium
 - Markers are i.i.d (independent and identically distributed)

The Admixture Model

- Admixture of "ancestral frequency profiles (AP)"

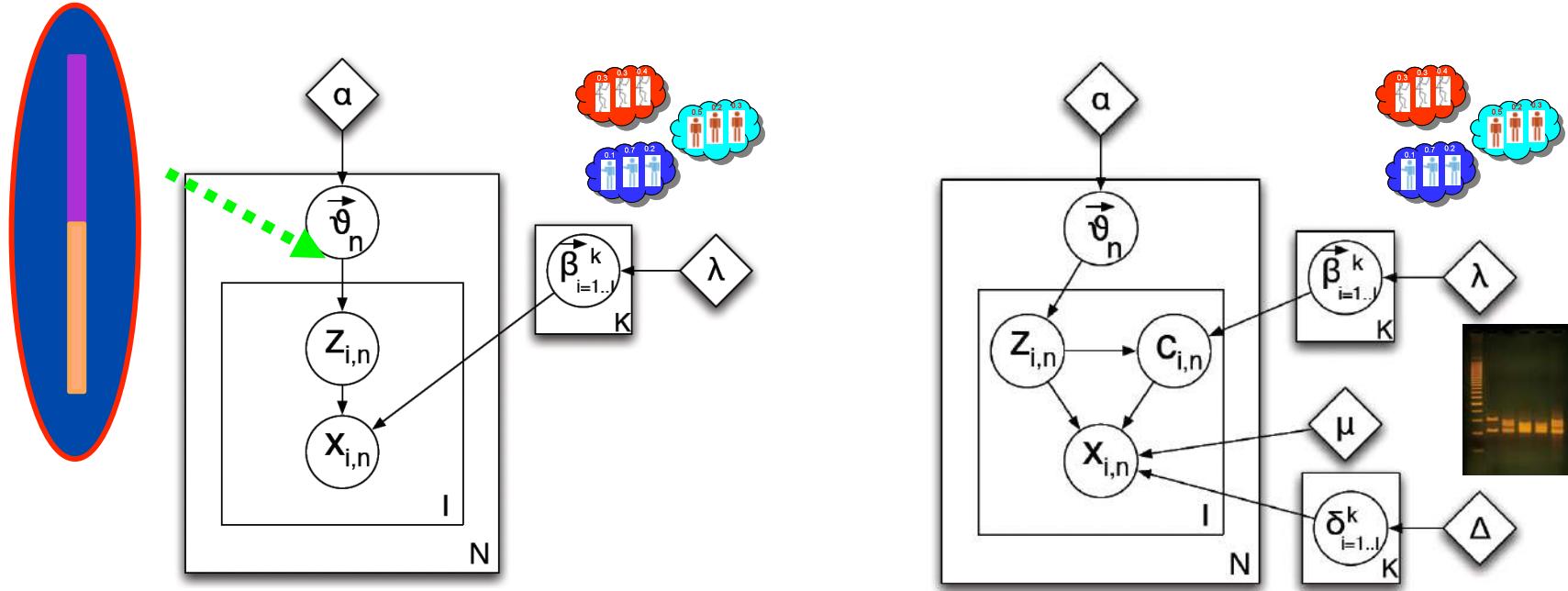


The Admixture Model



- β = Distribution over alleles
 - One per population –locus pair
- To generate an individual's genome
- Sample θ from $\text{Dirichlet}(\alpha)$
- For each locus
 - Sample z from $\text{Multinomial}(\theta)$
 - Sample x from β corresponding to the population chosen by z

From *Structure* to *mStruct* : Modelling Mutations



- From admixture of APs to admixture of MIMs
 - MiM: population-specific Mixture of Inheritance Models
- The inheritance model:

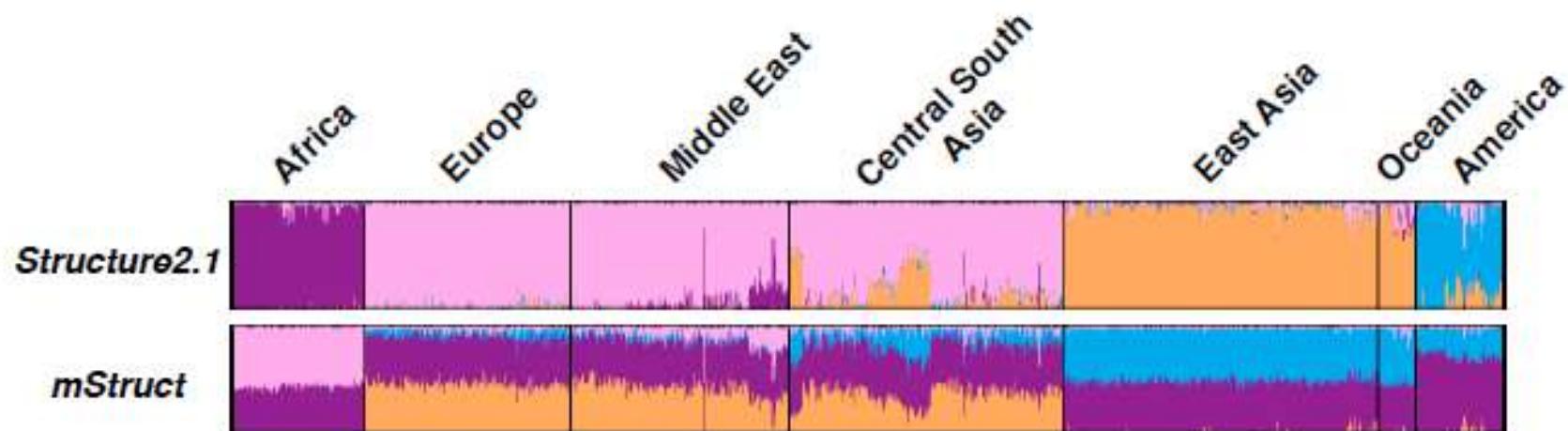
– Microsatellite

$$P(b|a) = \frac{1 - \delta}{1 - \delta^a + \delta} \delta^{|b-a|}.$$

SNPs:

$$P(b|a) = \delta^{\mathcal{I}[b=a]} \times (1 - \delta)^{\mathcal{I}[b \neq a]}; \quad a, b \in \{0, 1\}.$$

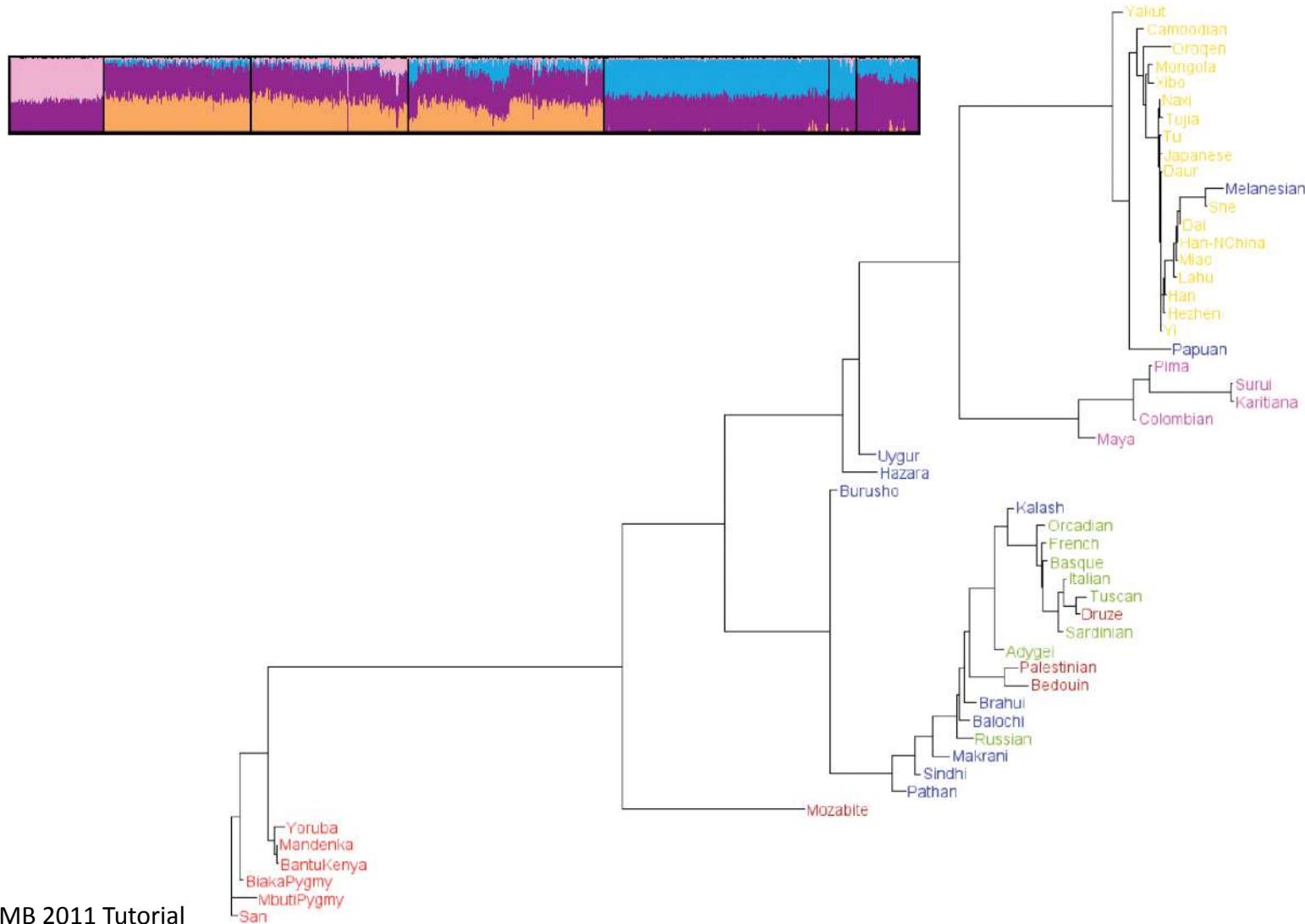
Comparing Population Structure Maps



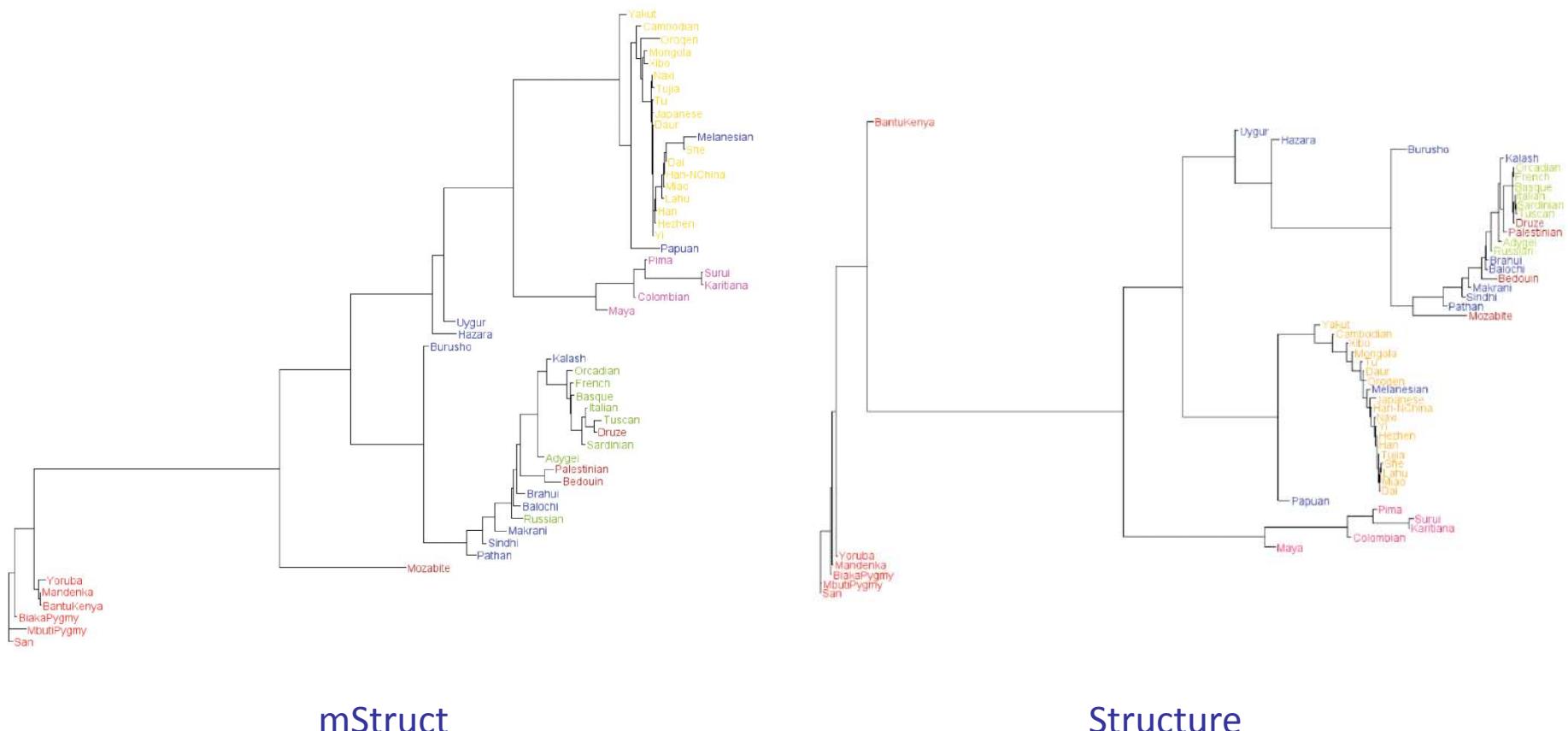
Ancestry structure maps inferred from microsatellite portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population. The colors represent different ancestral populations.

- A common ancestral population is now seen across all continents!
- Clusters remain unchanged

Phylogenetic tree from mStruct Structural map



Neighbour-joining Phylogenetic Trees from the Structural Maps



Comparison of Different Methods

	PCA	Model-based Clustering (Structure, mStruct)
Advantages	<ul style="list-style-type: none">• Statistical tests for significance of results (Patterson et al. 2006)• Easy visualization	<ul style="list-style-type: none">• Generative process that explicitly models admixture• Clustering is probabilistic: it is possible to assign confidence level of clusters
Disadvantages	<ul style="list-style-type: none">• No intuition about underlying processes	<ul style="list-style-type: none">• Computational more demanding• Based on assumptions of evolutionary models:<ul style="list-style-type: none">• Structure: No models of mutation, recombination• Mutation added in mStruct• Recombination added in extension by Falush et al.

Summary: data preparation

- Preparing genotype data
 - Haplotypes inferred
 - Tag SNP selected
 - Missing genotype values imputed
 - Population structures stratified
- Preparing phenotype data
 - Handle missing phenotype values
 - Data normalization
 - Structure discovery: network inference, hierarchical clustering

Part II

Background on Association Analysis: Single Phenotype Method

Overview

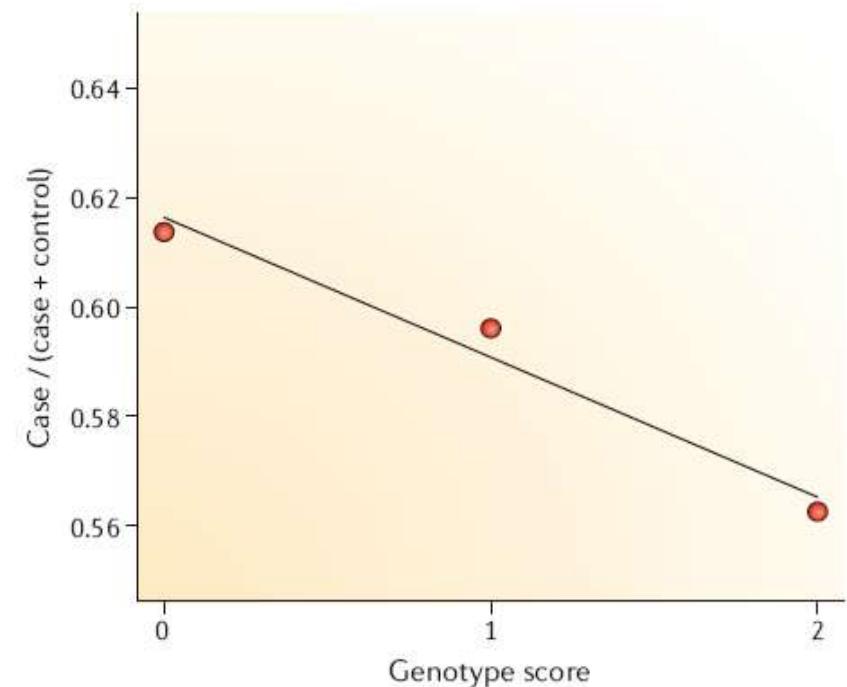
- Single SNP association test
 - Discrete-valued phenotype: case/control study
 - Continuous-valued phenotype: quantitative traits
 - Correcting for multiple testing
- Simultaneous analysis of all SNPs
 - Sparse regression method and convex optimization

Single SNP Association Analysis: Case/Control Study

- For each marker locus, find the 3x2 contingency table containing the counts of three genotypes

Genotype	Case	Control
AA	$N_{\text{case},\text{AA}}$	$N_{\text{control},\text{AA}}$
Aa	$N_{\text{case},\text{Aa}}$	$N_{\text{control},\text{Aa}}$
aa	$N_{\text{case},\text{aa}}$	$N_{\text{control},\text{aa}}$
Total	N_{case}	N_{control}

- χ^2 test with 2 df, or Fisher's exact test under the null hypothesis of no association



Genotype score = the number of minor alleles

Single SNP Association Analysis: Case/Control Study

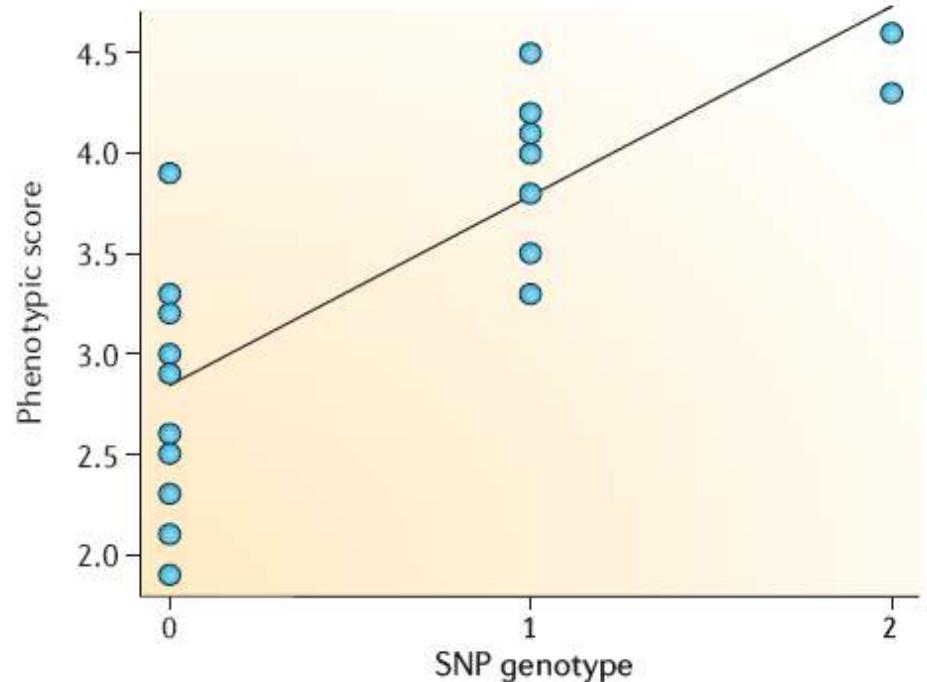
- Alternatively, assume an additive model, where the heterozygote risk is approximately between the two homozygotes
- Form a 2x2 contingency table. Each individual contributes twice from each of the two chromosomes.

Genotype	Case	Control
A	$G_{\text{case},A}$	$G_{\text{control},A}$
a	$G_{\text{case},a}$	$G_{\text{control},a}$
Total	$2 \times N_{\text{case}}$	$2 \times N_{\text{control}}$

- χ^2 test with 1df

Single SNP Association Analysis: Continuous-valued Traits

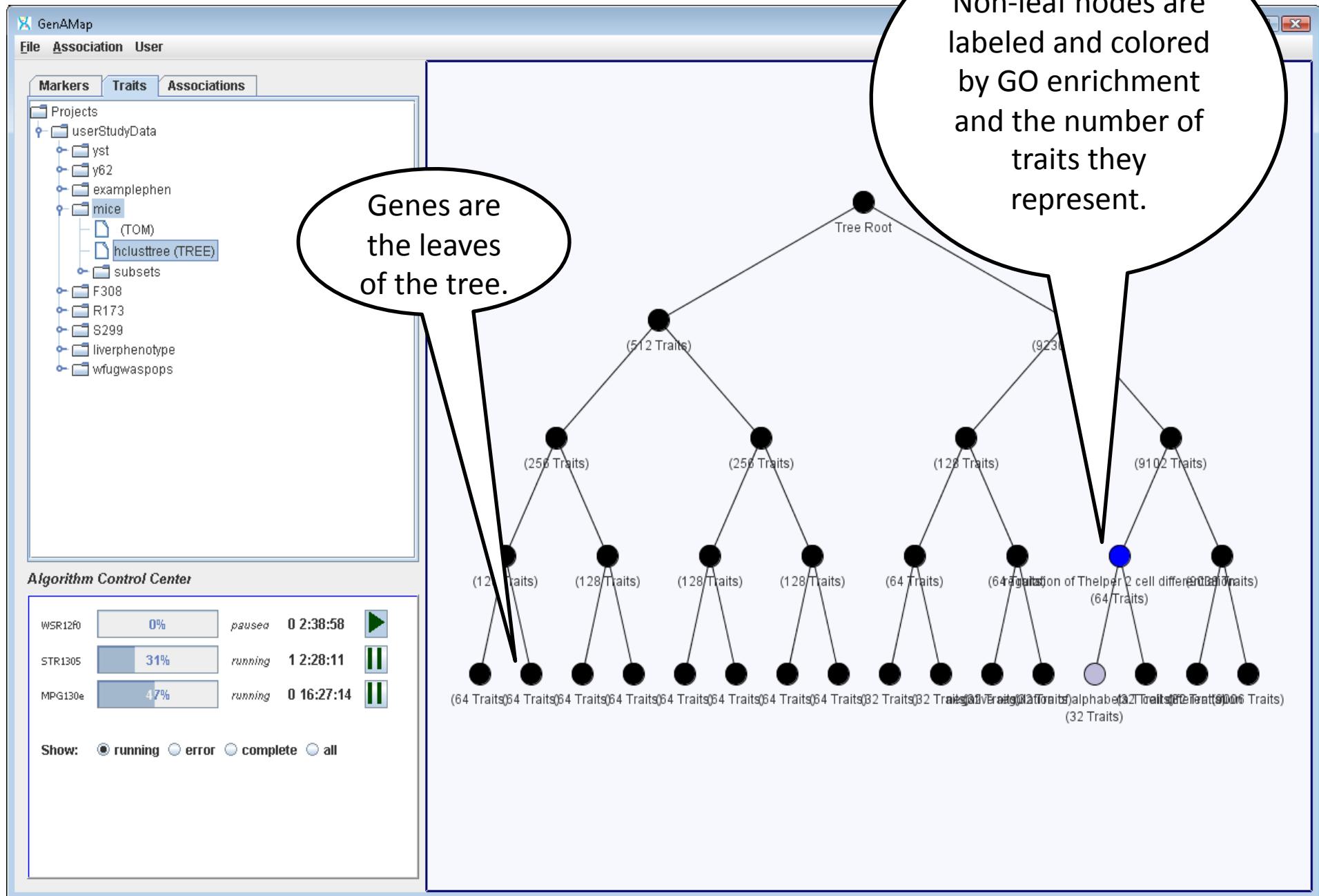
- Continuous-valued traits
 - Also called quantitative traits
 - Cholesterol level, blood pressure etc.
- For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as covariate



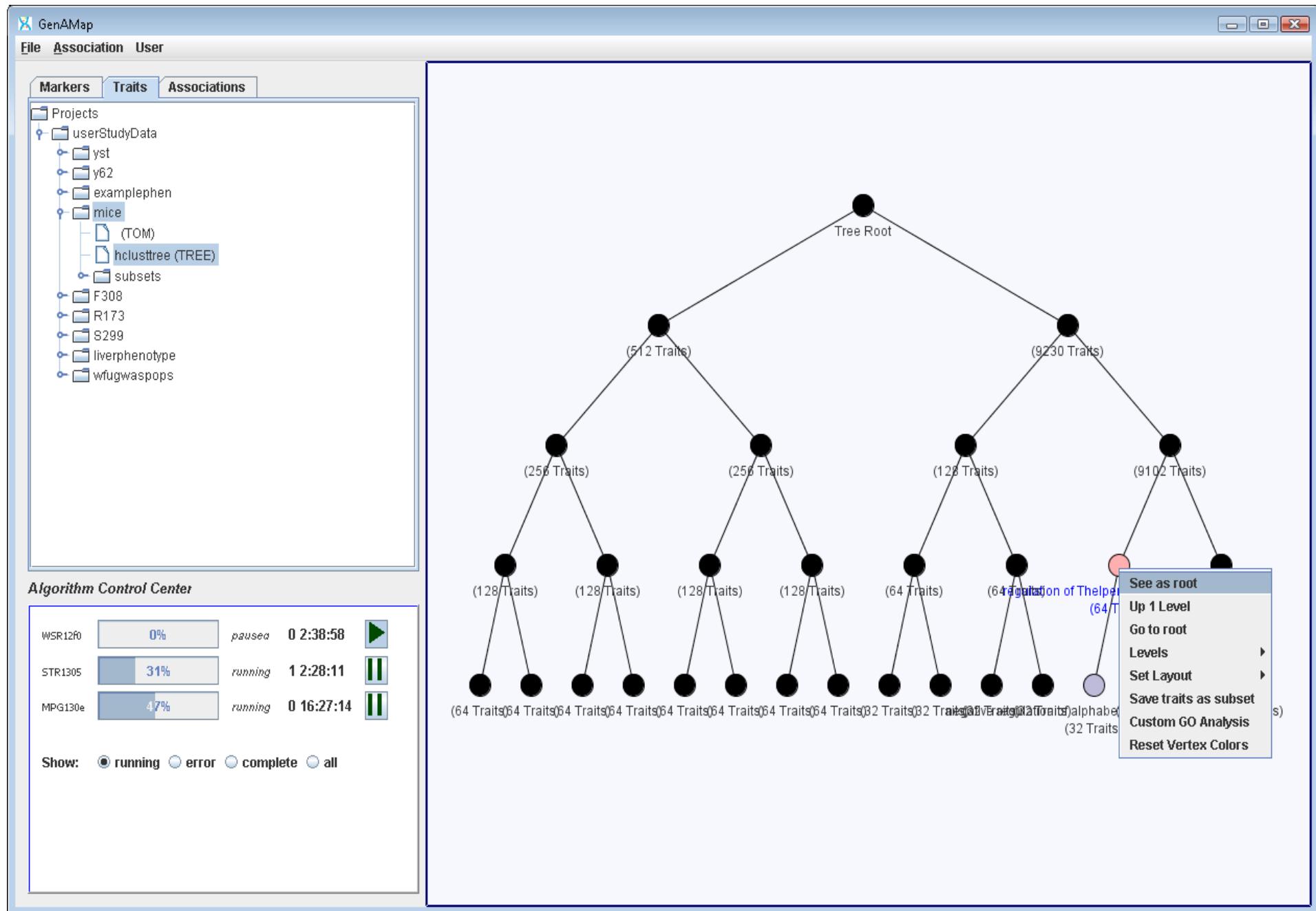
A Tree Analysis of Mouse eQTL Dataset

- NIH heterogeneous stock of mice (Johannesson et al., Genome Research 2009)
 - 259 mice from 7 inbred strains (A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL/6J, DBA/2J and LP/J)
 - 9742 gene expression
 - 12545 SNPs

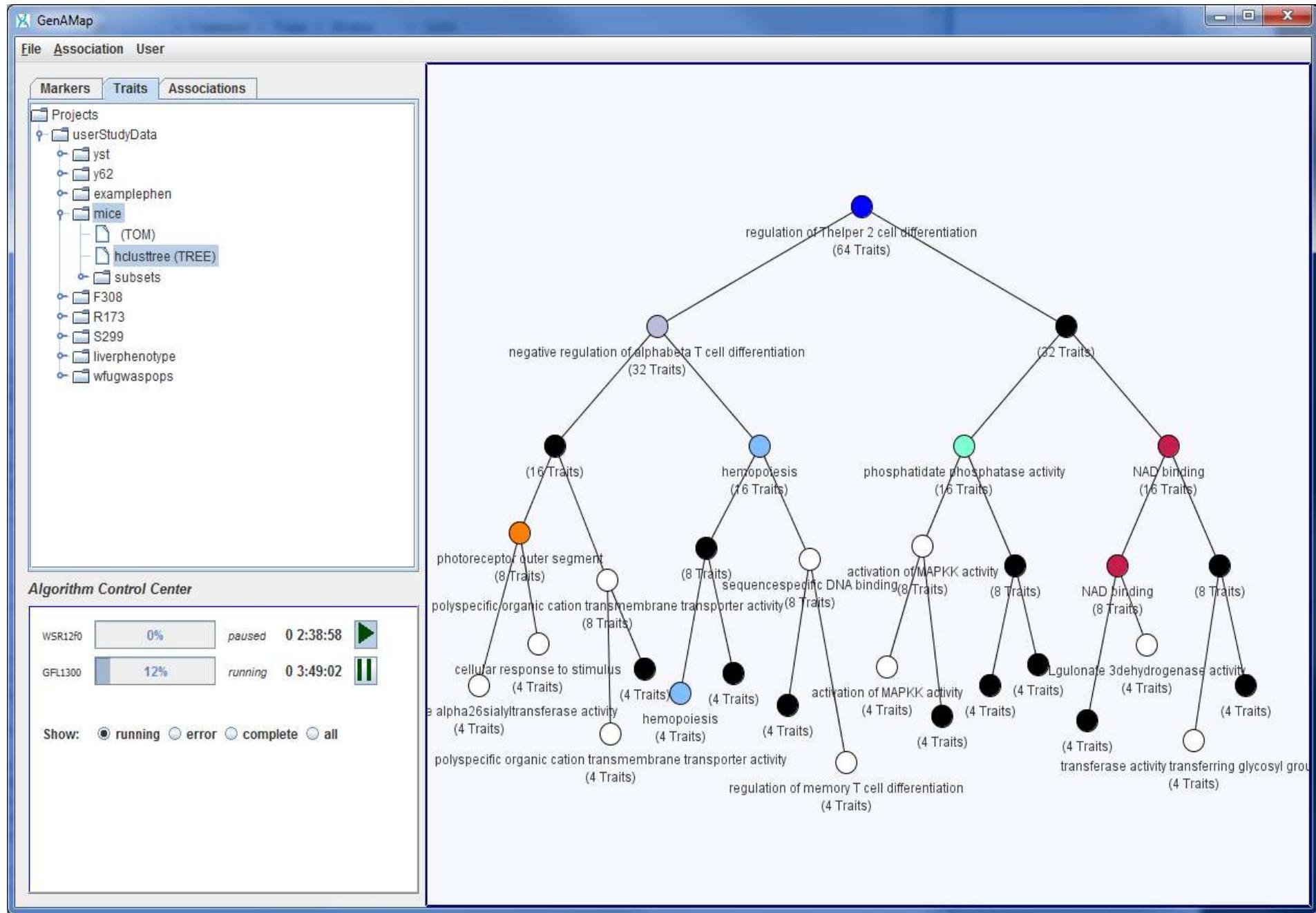
Genes displayed as a hierarchical tree



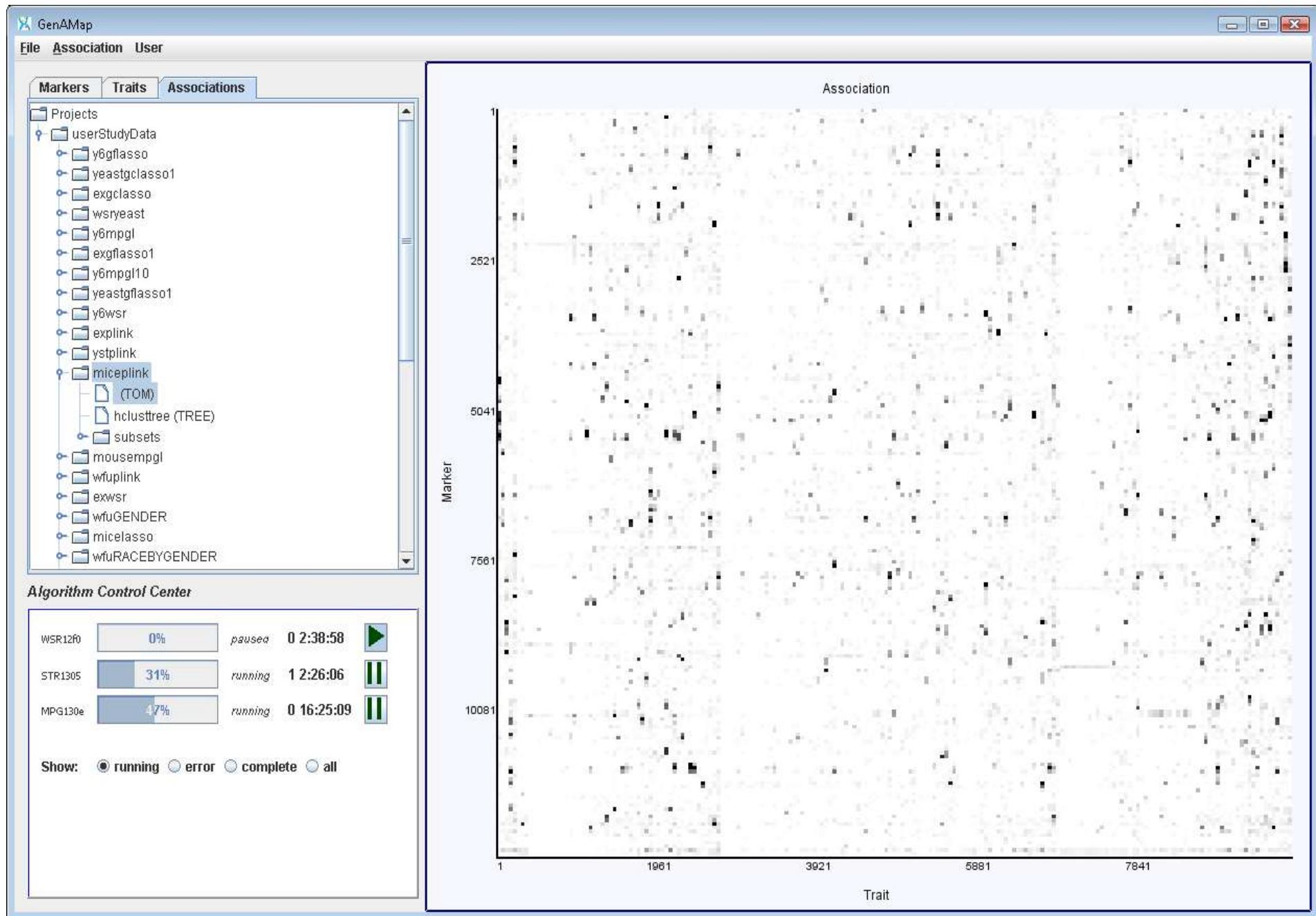
Browsing through the tree



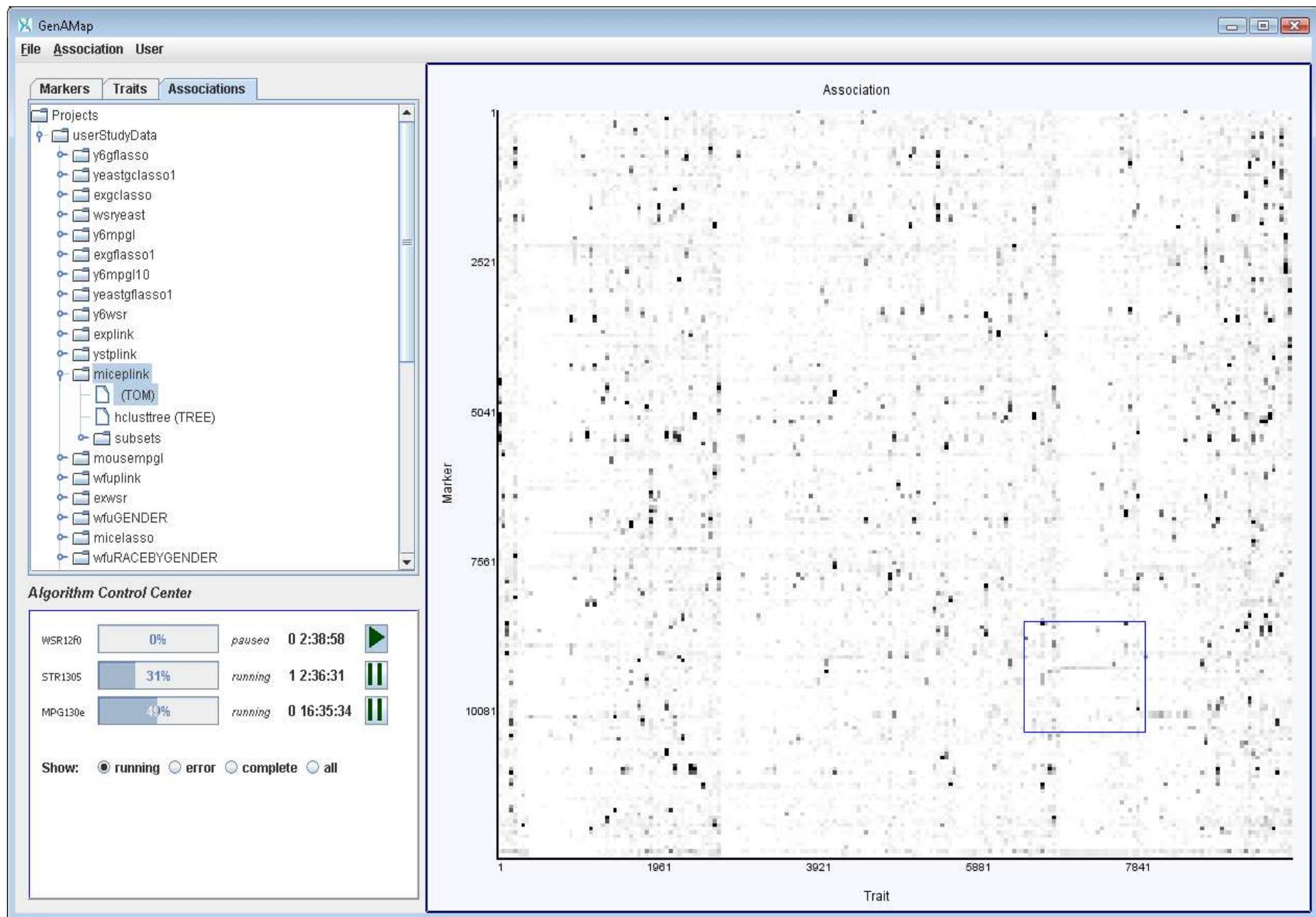
Branches of the tree are enriched for a common GO category



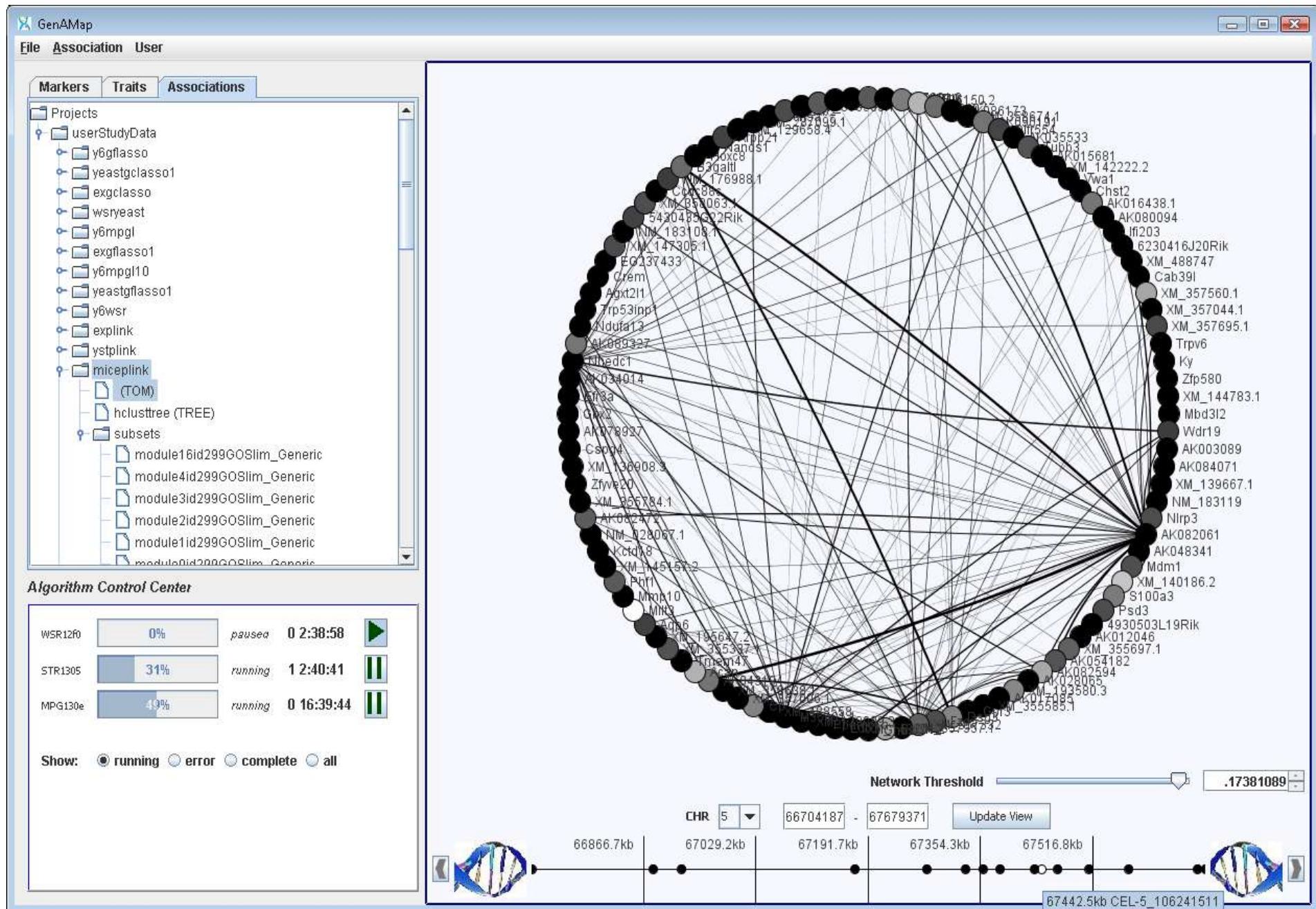
Association results (- log p-value)



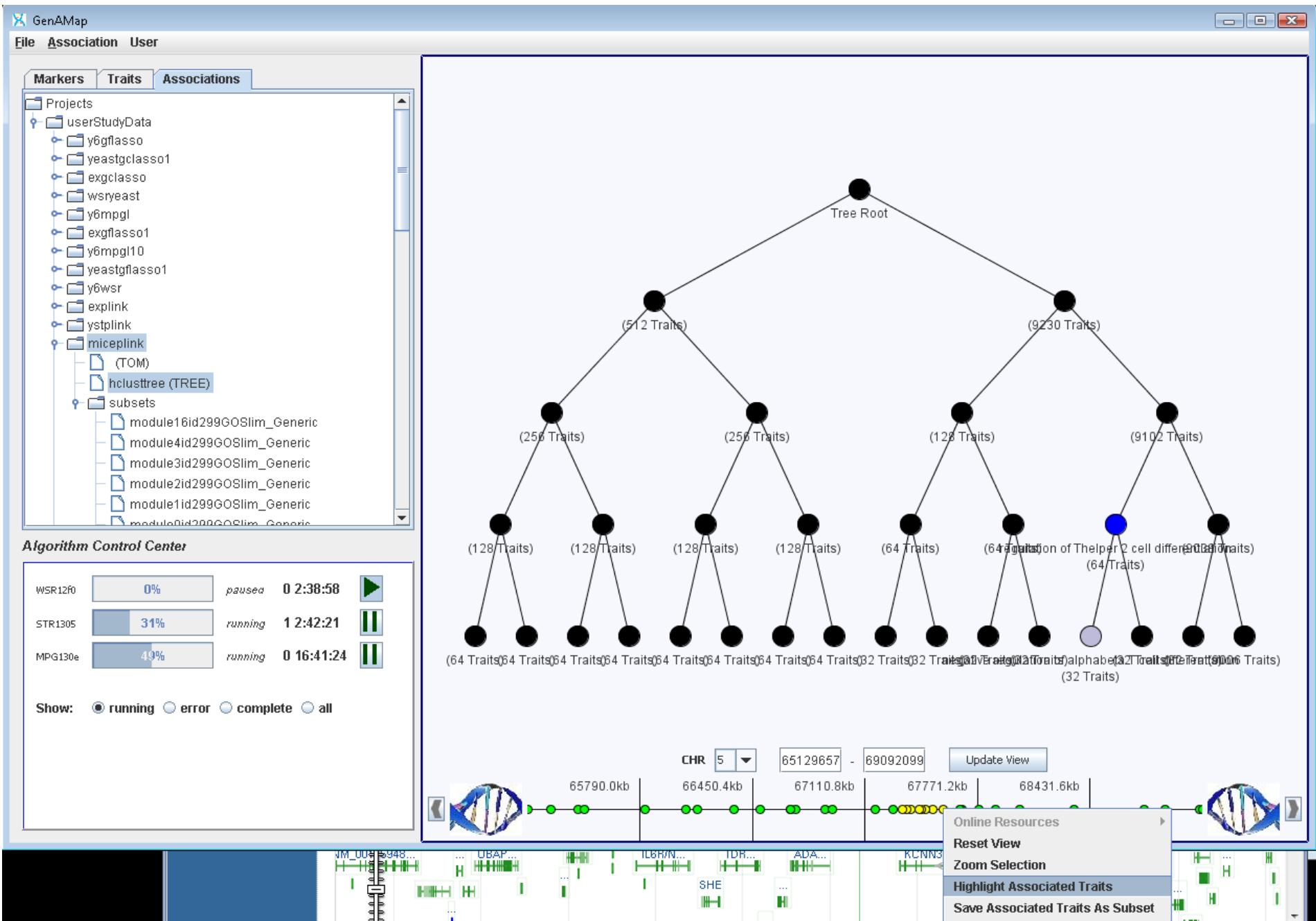
eQTL hotspot

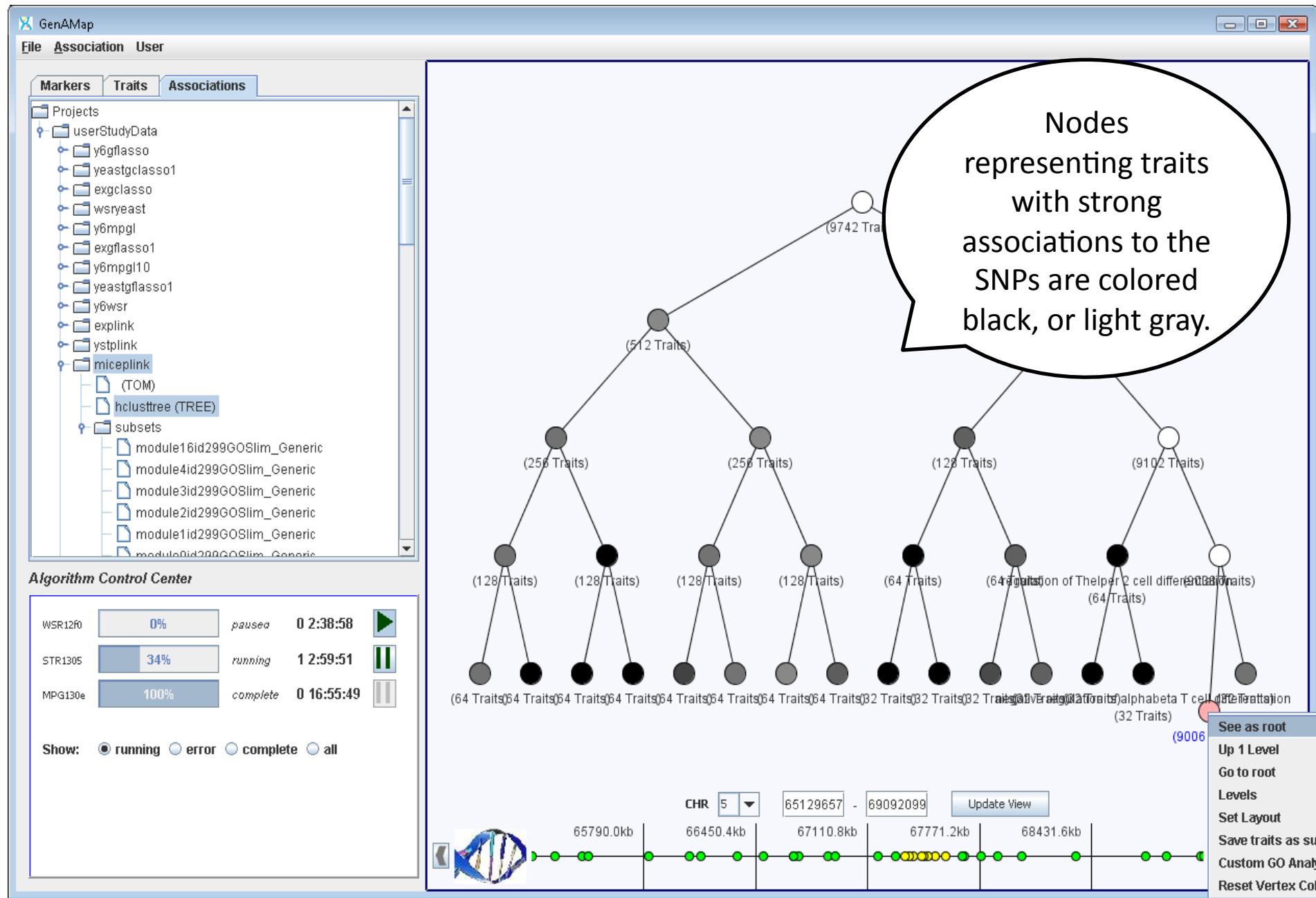


Select region to investigate on chromosome 5

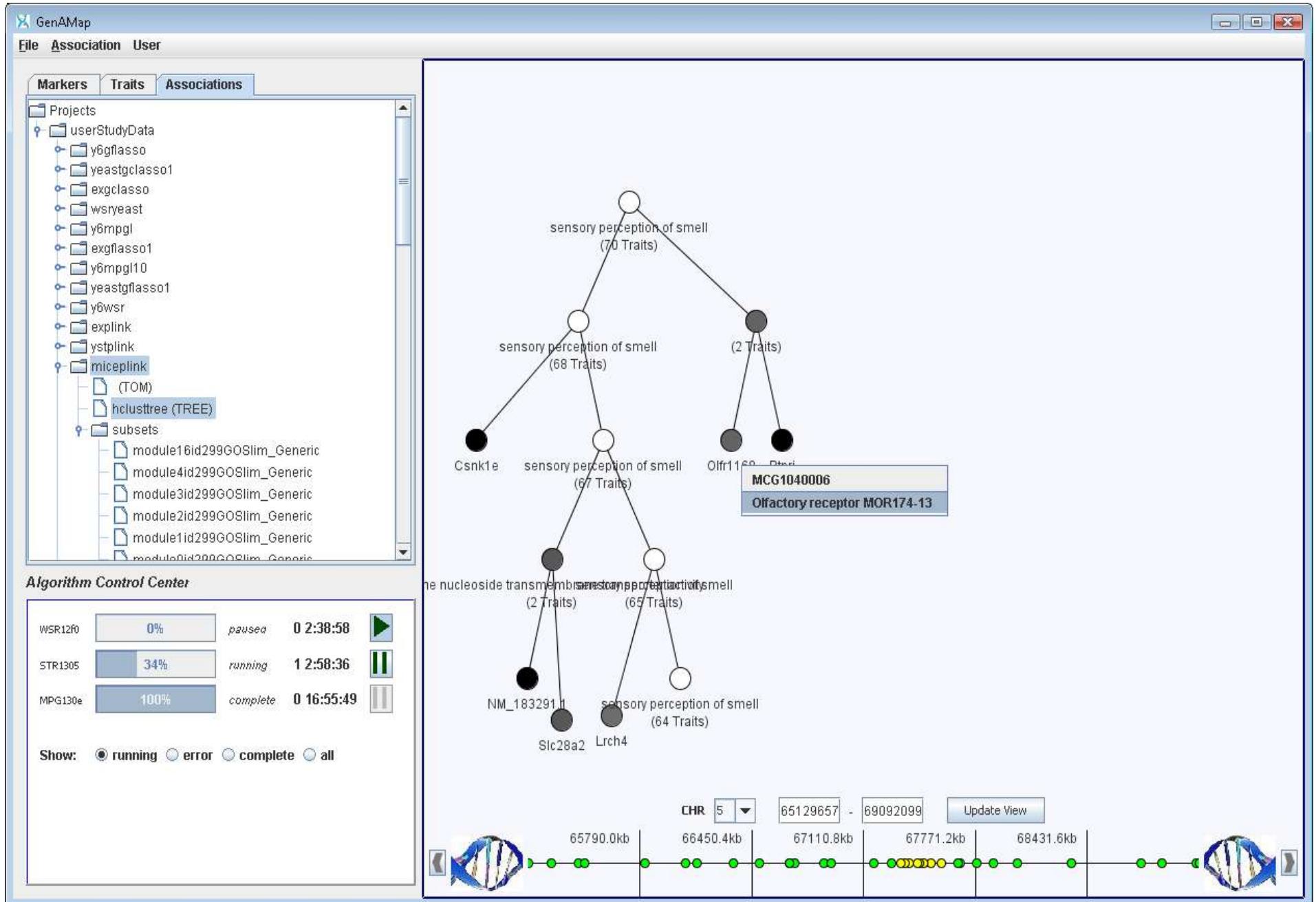


Switch to tree visualization





This region is associated with sensory perception of smell



These are our candidate genes near the SNP

More details about this RefSNP can be found in the following sections:

- [GeneView](#)
- [Map](#)
- [Submission](#)
- [Fasta](#)
- [Resource](#)
- [Diversity](#)
- [Validation](#)

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

Integrated Maps												
Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Group term	Group label	Contig label	Neighbor SNP	Map Method
37.1	5	105776341	NW_001030796.1	12617722	+	A	+	Primary Assembly	Mm_Celera	Mm_Celera	view	blast
37.1	5	109083781	NT_109320.4	32555902	+	A	+	Primary Assembly	MGSCv37	MGSCv37	view	blast

32,453,499 : 32,658,298 (204,800 bases shown, positive strand)

Open Full View

Flip Strands | ATG |

Configure

Sequence NT_109320.4: Mus musculus strain C57BL/6J chromosome 5 genomic DNA

itig, MGSCv37 C57BL/6J

SNP

Genes

Gak/NM_153569.1/NP_705797.1

Dgkq/Dgkq/NM_001057971

Tmem175/NM_001057971

Idua/Idua/NM_008...

Fgfrl1/NM_008...

Rnf212/XM_00147...

Sic...

LOC

GeneView

GeneView via analysis of contig annotation: N/A.

GeneView via direct blast against RefSeq sequences (used when no gene model is available): N/A

Submitter records for this RefSNP Cluster

The submission [ss45822725](#) has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs13478451** during BLAST analysis for the current build.

NCBI Assay ID	Handle Submitter ID	Validation Status	ss to rs Orientation	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date	Update Date	Build Added
---------------	---------------------	-------------------	----------------------	---------	-------------------	-------------------	------------	-------------	-------------

Correcting for Multiple Testing

- What happens when we scan the genome of 1 million markers for association with $\alpha = 0.05$?
 - 50,000 ($=1 \text{ million} \times 0.05$) SNPs are expected to be found significant just by chance
 - We need to be more conservative when we decide a given marker is significantly associated with the trait.
- Correction methods
 - Bonferroni correction
 - Permutation test

Bonferroni Correction

- If N markers are tested, we correct the significance level as $\alpha' = \alpha/N$
 - Assumes the N tests are independent, although this is not true because of the linkage disequilibrium.
 - Overly conservative for tightly linked markers

Permutation Procedure

- Step 1: Compute the test statistic T using the original dataset
- Step 2: Set $N_{\text{sig}} = 0$
- Step 3: Repeat 1: N_{perm}
 - Step 3a: Randomly permute the individuals in the phenotype data to generate datasets with no association (retain the original genotype)
 - Step 3b: Find the test statistics T_{perm} of SNPs using the permuted dataset
 - Step 3c: if $T > T_{\text{perm}}$, $N_{\text{sig}} = N_{\text{sig}} + 1$
- Step 4: Compute p -value as $(1 - N_{\text{sig}} / N_{\text{perm}})$

This approach is computationally demanding because often a large N_{perm} is required.

Considering All SNPs in a Single Association Model

- Sparse multivariate regression
 - Estimates the effect of each SNP on the phenotype **in the presence of all the other SNPs**
 - High-dimensional data
 - Association strengths should be estimated for a very large number of SNPs
 - Penalized regression methods can be used to set the association strengths of many irrelevant SNPs to zero
 - Estimating the parameters (association strengths)
 - Convex optimization methods

Multivariate Regression for Single-Trait Association Analysis

Trait

Genotype

Association Strength

2.1

=

T G A A C C A T G A A G T A

x

?

y

=

X

x

β

Multivariate Regression for Single-Trait Association Analysis

Trait

2.1

=

Genotype

T G A A C C A T G A A G T A

Association Strength

X



$$\operatorname{argmin}_{\beta} (y - X\beta)^T \cdot (y - X\beta)$$

Many non-zero associations:
Which SNPs are truly significant?

Sparsity

- One common assumption to make is **sparsity**.
- **Makes statistical sense:** Learning is now feasible in high dimensions with small sample size
- **Makes biological sense:** each phenotype is likely to be associated with a small number of SNPs, rather than all the SNPs.

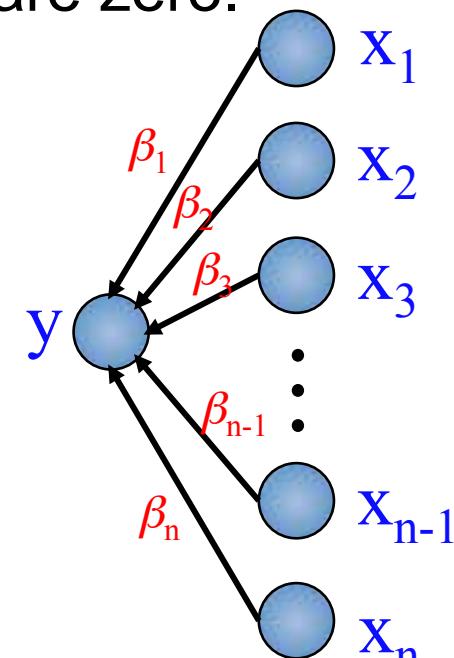
Sparsity: In a mathematical sense

- Consider least squares linear regression problem:
- Sparsity means most of the beta's are zero.

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

subject to:

$$\sum_{j=1}^p \mathbb{I}[|\beta_j| > 0] \leq C$$



- But this is not convex!!! Many local optima, computationally intractable.

L1 Regularization (LASSO)

- A convex relaxation.

Constrained Form

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

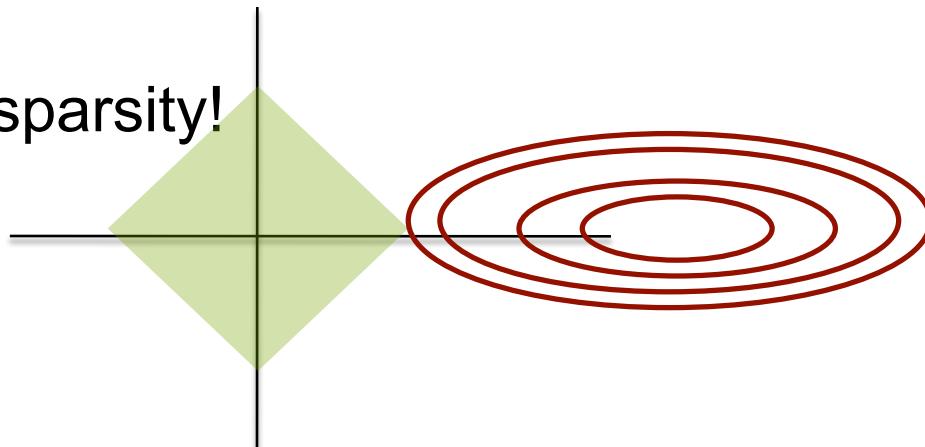
subject to:

$$\sum_{j=1}^p |\beta_j| \leq C$$

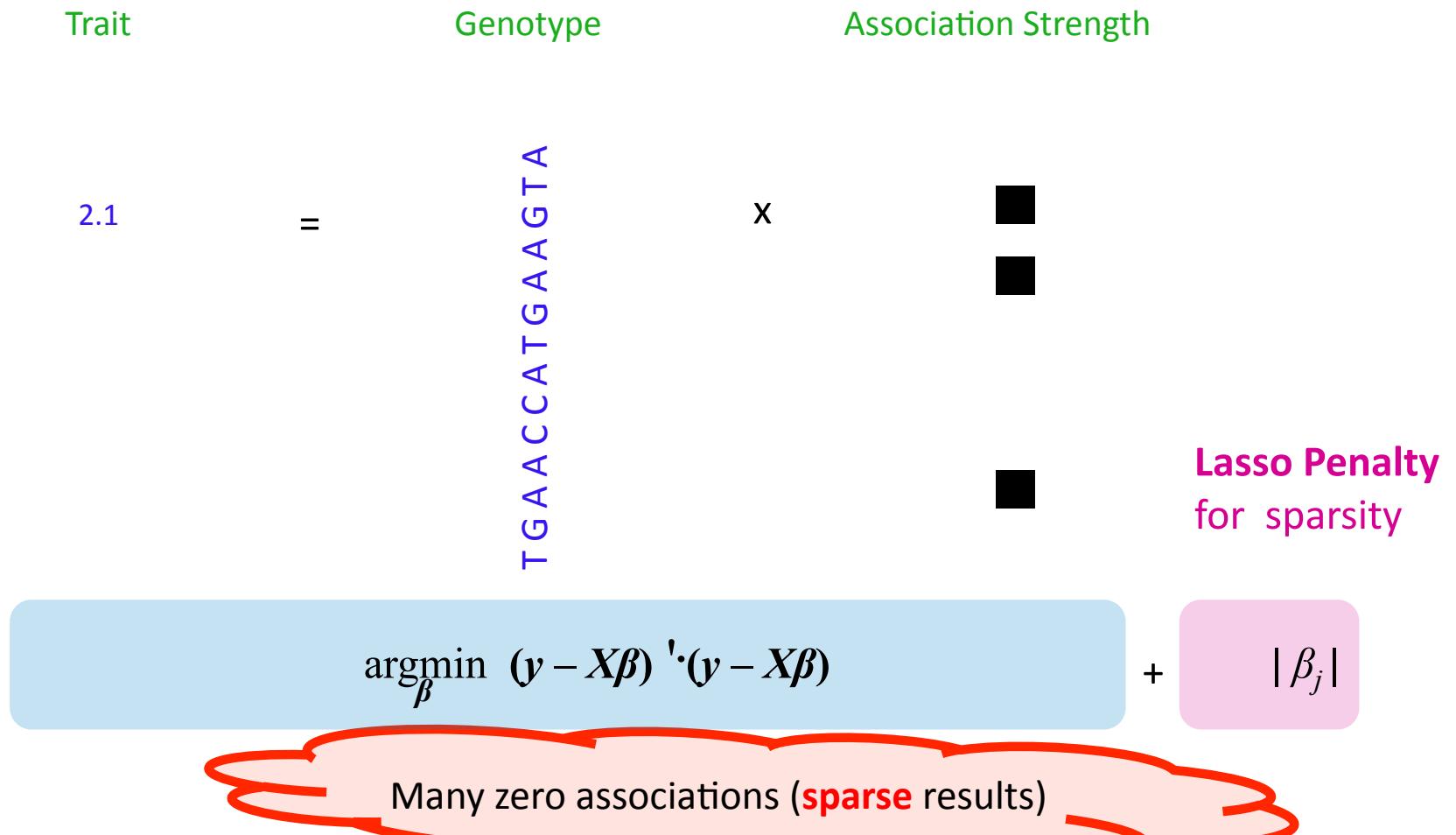
Lagrangian Form

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- Still enforces sparsity!



Lasso for Reducing False Positives



Estimating Parameters in Lasso (Association Strengths)

- Pathwise coordinate descent algorithm
(Friedman et al., Annals of Applied Statistics 2007)
 - Compute the subgradient of the objective function with respect to each β_j to find the update equation
 - In each iteration, update each β_j , and iterate until convergence
- A fast implementation available at
<http://cran.r-project.org/web/packages/glmnet/index.html>

Logistic Regression for Multiple SNPs in Case/control Association

	Case(1)/Control(0)	Genotype
Individual 1	0	...C.....T...C.....T... ...C.....A...C.....T... ...G.....A...G.....A... ...C.....T...C.....T... ...C.....A...G.....T... ...C.....T...C.....T...
Individual 2	1	
:		
Individual N	1	

$$p(y_i=\text{case}) = f\left(\sum_{k=1}^K x_{ik} \beta_k\right)$$

- f : logistic function
- β_k : weight (association strength) for the k th SNP
- x_{ik} : genotype of the k th SNP for the i th individual (0,1, or 2 depending on the number of minor alleles)

Part III

Structured

Genome-Phenome-Transcriptome Association Analysis

Structured Association: a New Paradigm

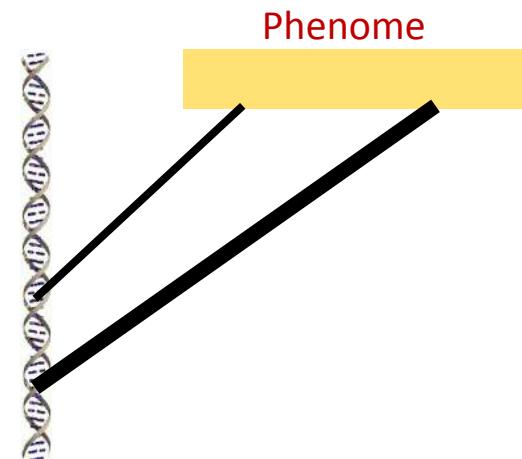
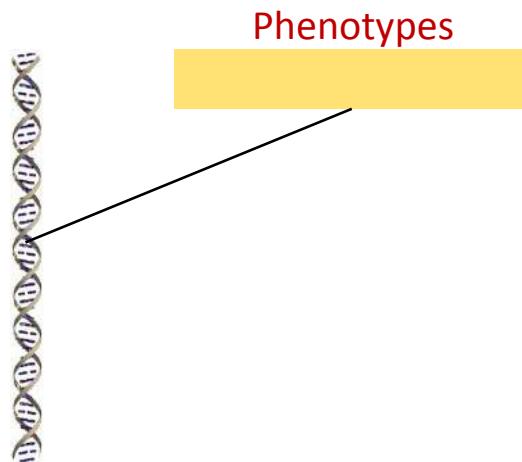
Standard Approach

Consider
one phenotype at a
time

VS.

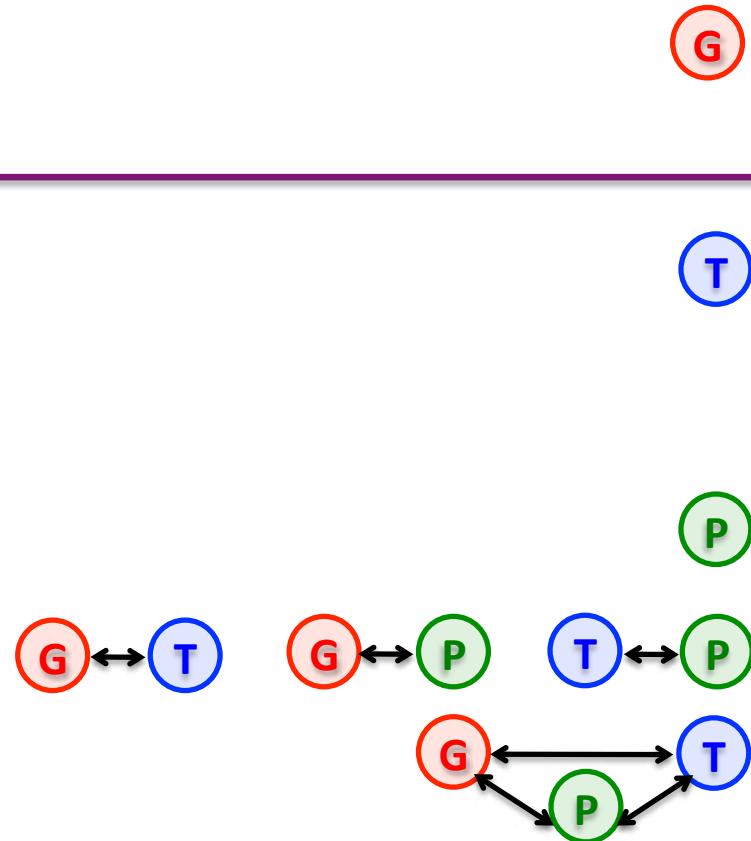
New Approach

Consider **multiple
correlated phenotypes
(phenome) and genotypes
(genome) jointly**



Overview

- **Genome structure** in association analysis
 - Linkage disequilibrium
 - Population structure
 - Epistasis
- **Transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Leveraging gene expression tree
- **Phenome structure** in association analysis
 - Pleiotropy
 - Dynamic trait
- Two-way structured association
- Three-way structured association
- Visualization software



Multi-marker Association Test

- Idea: a haplotype of multiple SNPs is a better proxy for a true causal SNP than a single SNP
 - Exploit the linkage disequilibrium structure in genome
- Form a new allele by combining multiple SNPs for a haplotype

SNP A	SNP B		Auxiliary Markers for Haplotypes			
0	0		1	0	0	0
0	1		0	1	0	0
1	0		0	0	1	0
1	1		0	0	0	1

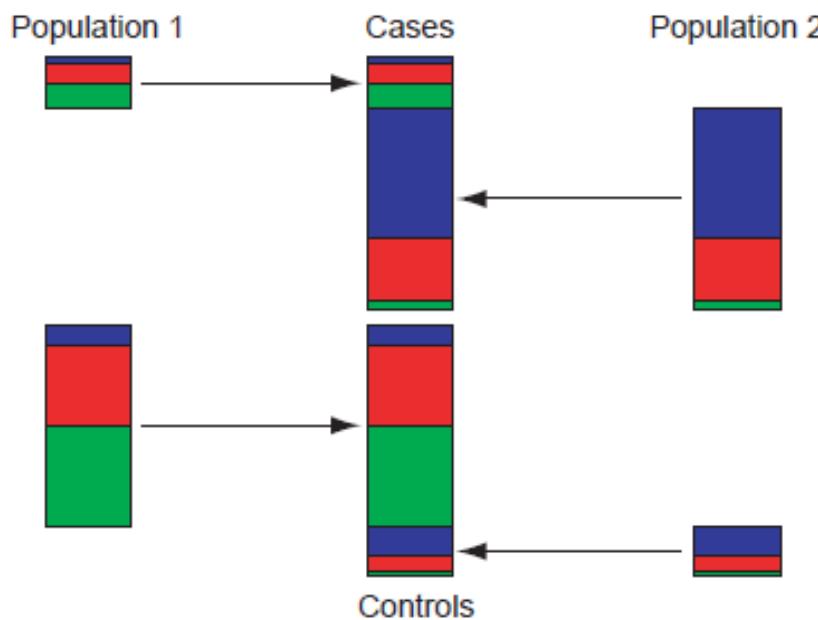
- Test the haplotype allele for association

Multi-marker Association Test

- Multi-marker approach can capture dependencies across multiple markers
 - SNPs in LD form a haplotype that can be tested as a single allele
 - Can achieve the same power with data collected for fewer samples
- Challenge as the size of haplotype increases
 - Haplotype of K SNPs results in 2^K different haplotypes, but the number of samples corresponding to each haplotype decreases quickly as we increase K
 - Large K requires a large sample size

Population Structure and Association Analysis

- Population structure in data causes false positives
 - Samples in the case population are usually more related
 - Any SNPs more prevalent in the case population will be found significantly associated with the trait.



Accounting for Population Structure in Association Analysis

- Needs to account for population structure in association mapping
 - Genomic control (Devlin & Roeder, Biometrics 1999)
 - Structured association (Pritchard et al., AJHG 2000)
 - EigenStrat: principal component analysis (Price et al., Nature Genetics 2006)
 - Multi-population lasso (Punyani et al., ISMB 2010)

Genomic Control (GC)

- **Idea:** Use the SNPs that are not associated with the trait to remove the effect of population stratification
- Genotype data consist of
 - Candidate genes to be tested
 - L supplementary loci (null loci) for estimating the inflation factor λ
- GC uses the inflation factor λ to correct the association statistic of the SNP in the candidate gene
- **Limitation:** the inflation factor λ is assumed to be the same across the genome, ignoring population admixture

Structured Association

- Idea: Within each subpopulation, an association between a genetic marker and the trait is a true association.
- Two-stage method
 - Step 1: Using Structure (Pritchard et al., Genetics 2000) and unlinked genetic markers,
 - estimate the population structure
 - assign sampled individuals to putative subpopulations
 - Step 2:
 - Test for association within the subpopulations inferred in Step 1
- Limitation
 - Running Structure is computationally demanding

EigenStrat: Structured Association with PCA

- Step 1: (Inferring Ancestry) PCA is applied to genotype data to infer continuous axes of genetic variation

Genotypes

Samples

	1	1	1	0	0
	0	1	2	1	2
	2	1	1	0	1
SNPs	0	0	1	2	2
	2	1	1	0	0
	0	0	1	1	1
	2	2	1	1	0

PCA → Axis of variation +0.7 +0.4 -0.1 -0.4 -0.5

EigenStrat: Structured Association with PCA

- Step 2: (Removing Ancestry Effects) Genotype at a candidate SNP and phenotype are continuously adjusted by amounts attributable to ancestry along each axis

Candidate SNP	2	2	1	1	0	→	1.0	1.4	1.1	1.6	0.8
Phenotype	1	1	0	0	0	→	0.3	0.6	0.1	0.4	0.5



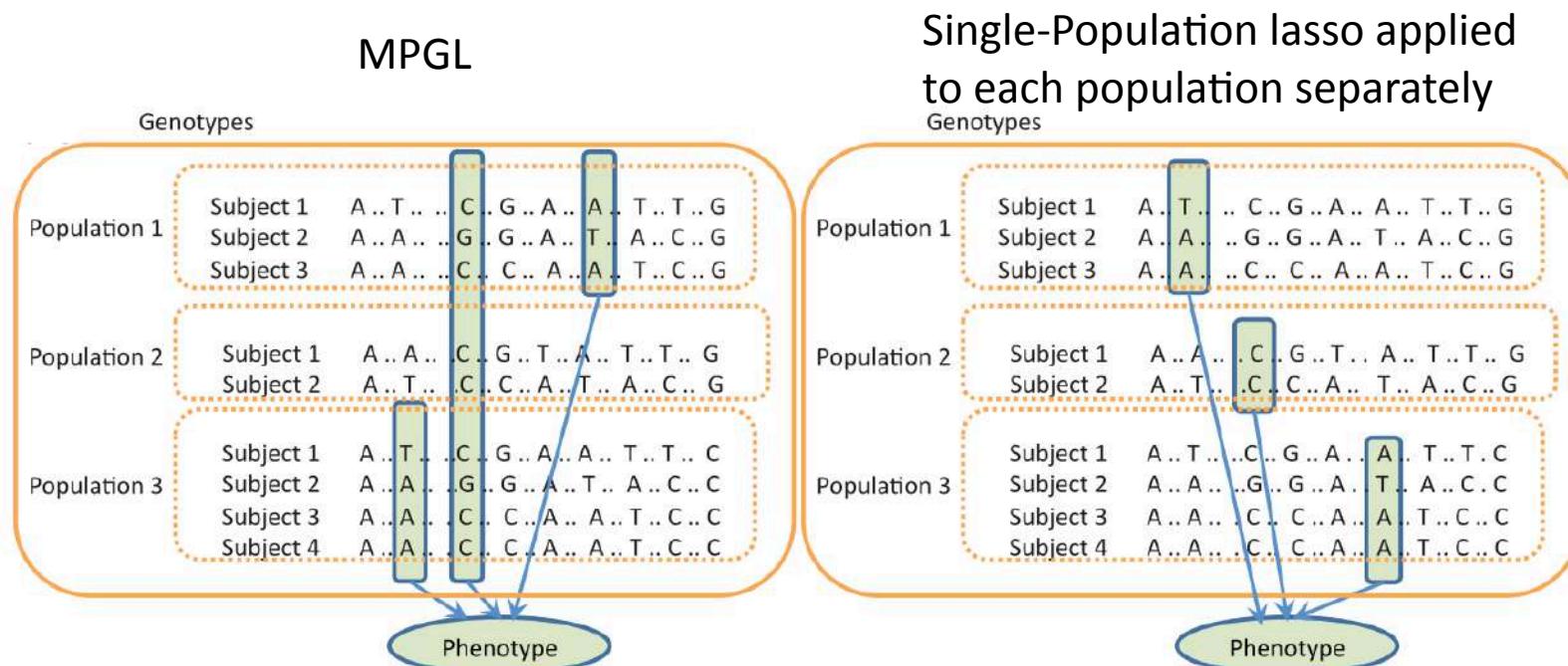
- Step 3: (Association test) $\chi^2 = 0.07 \Rightarrow$ no association

Multi-Population Group Lasso (MPGL)

- Idea: instead of correcting for the population structure (as in GC, structured association, and EigenStrat), let's exploit the population structure to detect true associations
- Assume that the population structure is known
 - If not, use existing algorithms such as Structure to infer population structure
- Sparse multivariate regression with group-lasso penalization

Multi-Population Group Lasso (MPGL)

- MPGL combines information across multiple populations to detect
 - SNPs with associations in all of the populations
 - SNPs with associations in each population



Multi-Population Group Lasso (MPGL)

- Joint estimation of shared and population-specific associations

$$\min_B \frac{1}{2} \sum_{i=1}^k (Y^{(i)} - X^{(i)}\beta^{(i)})^2 + \lambda \|B\|_{l_1/l_2}$$

For Each Population i

Least Square Error

Encourages Joint Sparsity

$$\|B\|_{l_1/l_2} = \sum_{i=1}^p \|B_i\|_2$$

Sum over SNPs

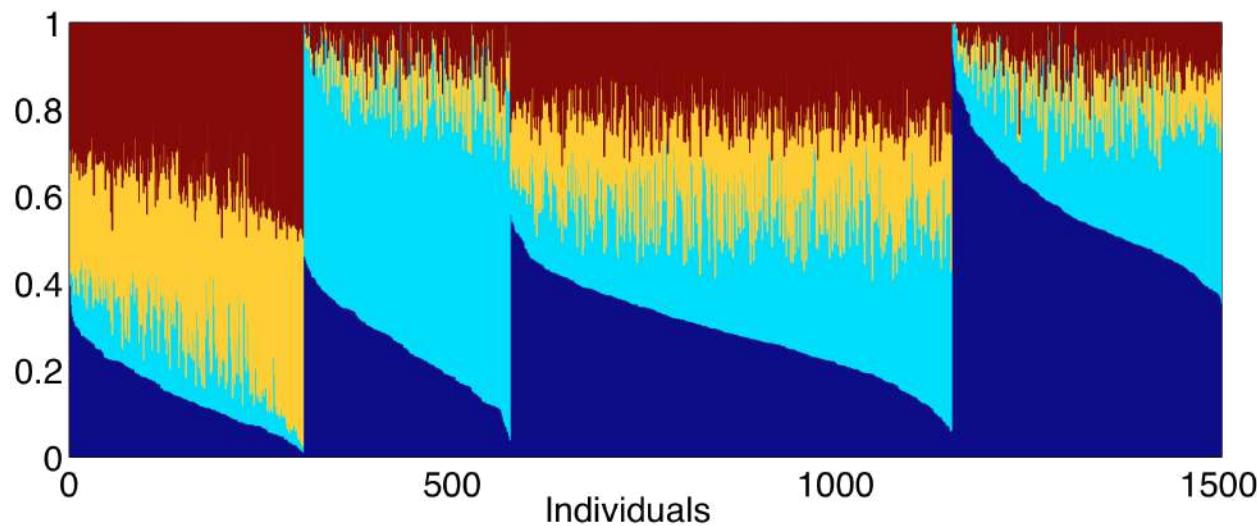
The diagram illustrates the MPGL objective function. It shows the sum of squared errors for each of the k populations, indicated by orange arrows pointing to the term $\sum_{i=1}^k$. Below this, an orange arrow points to the term $(Y^{(i)} - X^{(i)}\beta^{(i)})^2$, which is labeled 'Least Square Error'. A large orange bracket encloses the entire sum and the regularization term, with the label 'Encourages Joint Sparsity' positioned above it. This bracket also encompasses the term $\sum_{i=1}^p \|B_i\|_2$, which is labeled 'Sum over SNPs'.

Experiments: Lactose Persistence

- Data : 1400 individuals from the control group of the WTCCC dataset, all of European descent.
(The Wellcome Trust Case Control Consortium, Nature 2007)
- Genotype : 135.16-136.82Mb region on chromosome 2 (known to show geographical variation).
- Phenotype : Lactose persistence, fully determined by a particular mutation near the LCT gene (Enattah et al., 2002)
- Associated marker : SNP rs4988243 lies in a high linkage disequilibrium region ($r^2 > 0.9$) with this known genetic variant.

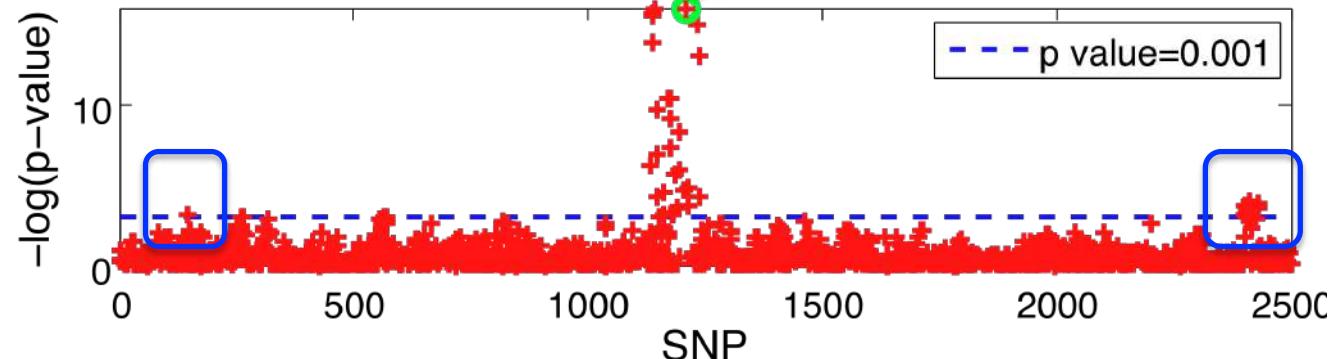
Analysis of Population Structure

- Results from *Structure* (Pritchard et al., Genetics 2000) on genotype data with four populations

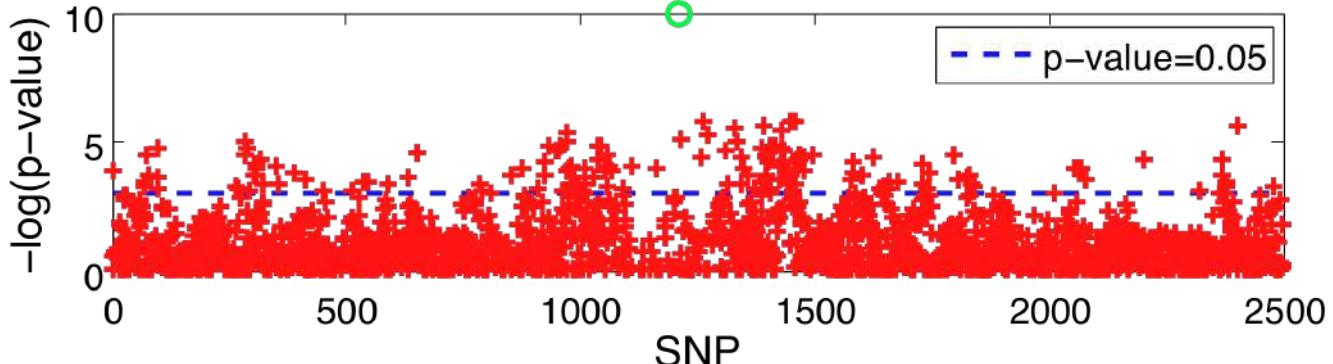


Results

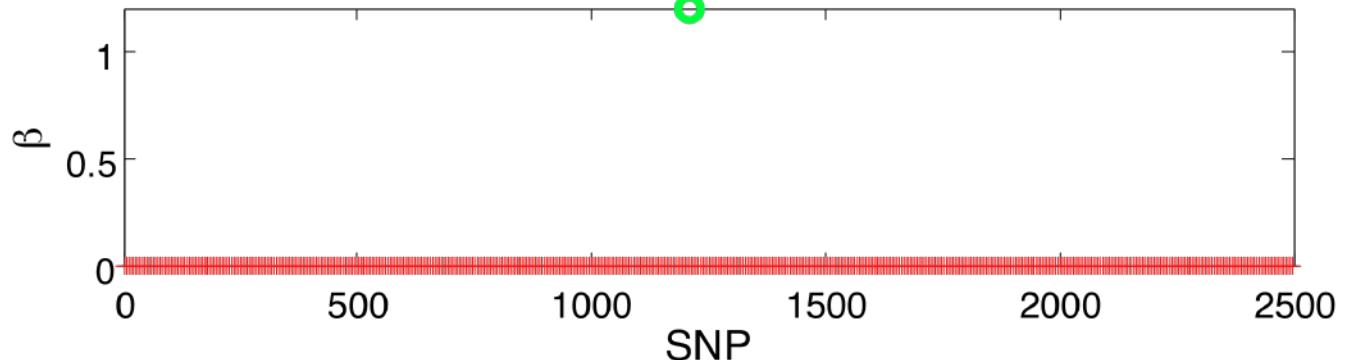
Eigenstrat



Genomic
Control

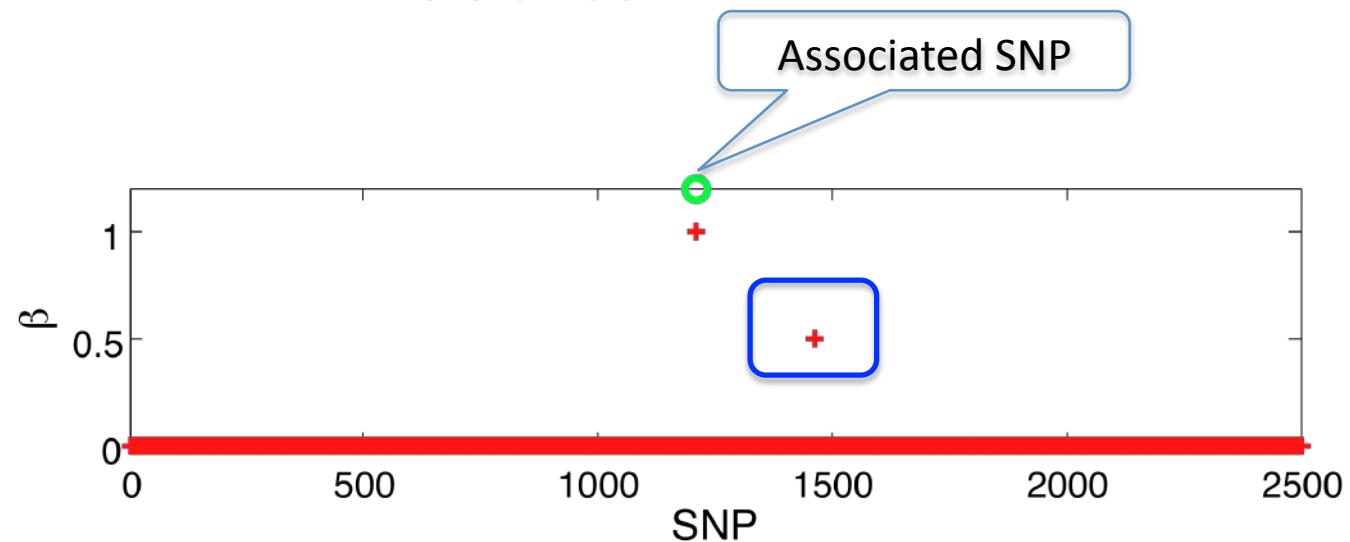


Lasso

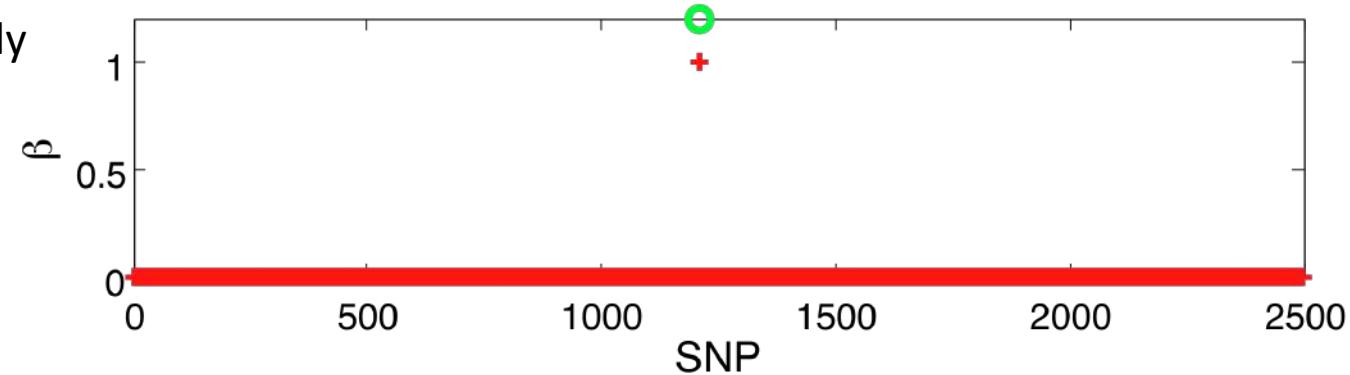


Results

Lasso for
structured
association



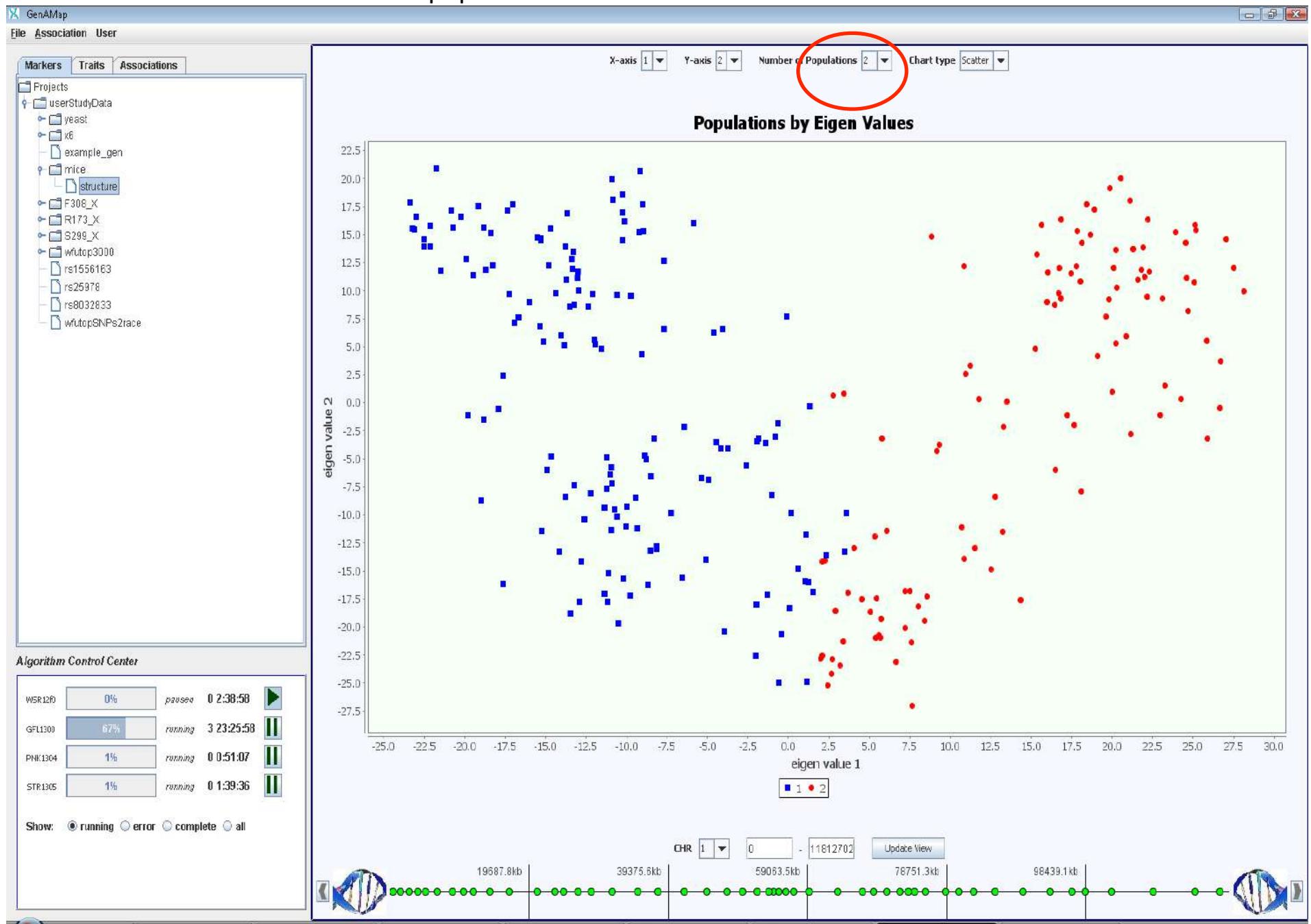
MPGL identifies only
the correct SNP



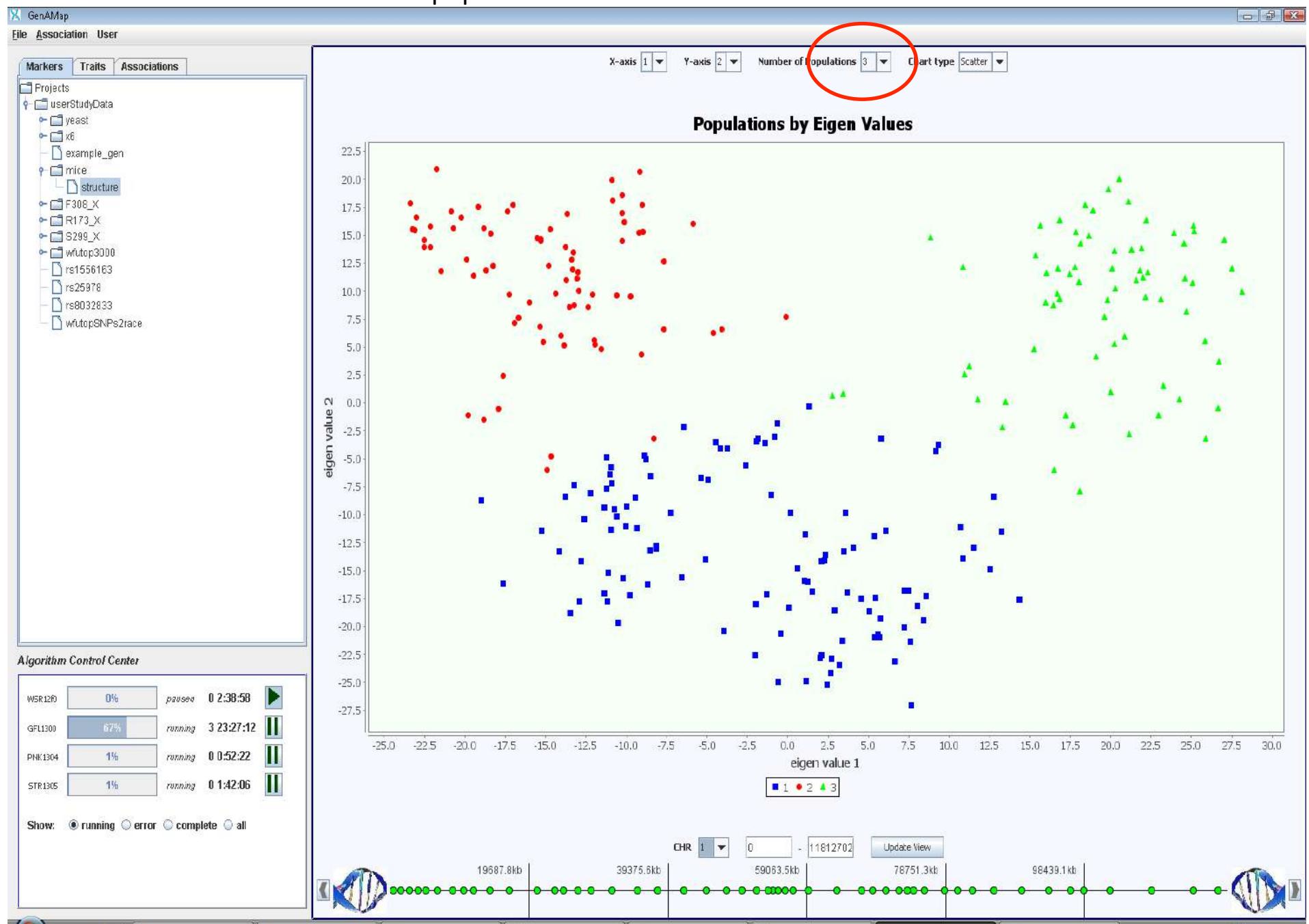
Association Analysis of Mouse Data: Considering Population Structure

- NIH heterogeneous stock of mice (Johannesson et al., Genome Research 2009)
 - 259 mice descended from 7 inbred strains (A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL/6J, DBA/2J and LP/J)
 - 176 phenotype measurements
 - 12,545 SNPs
- Specifically look at 7 asthma-related traits
 - Respiratory rate, tidal volume, minute volume, expiratory time, inspiratory time, enhanced pause

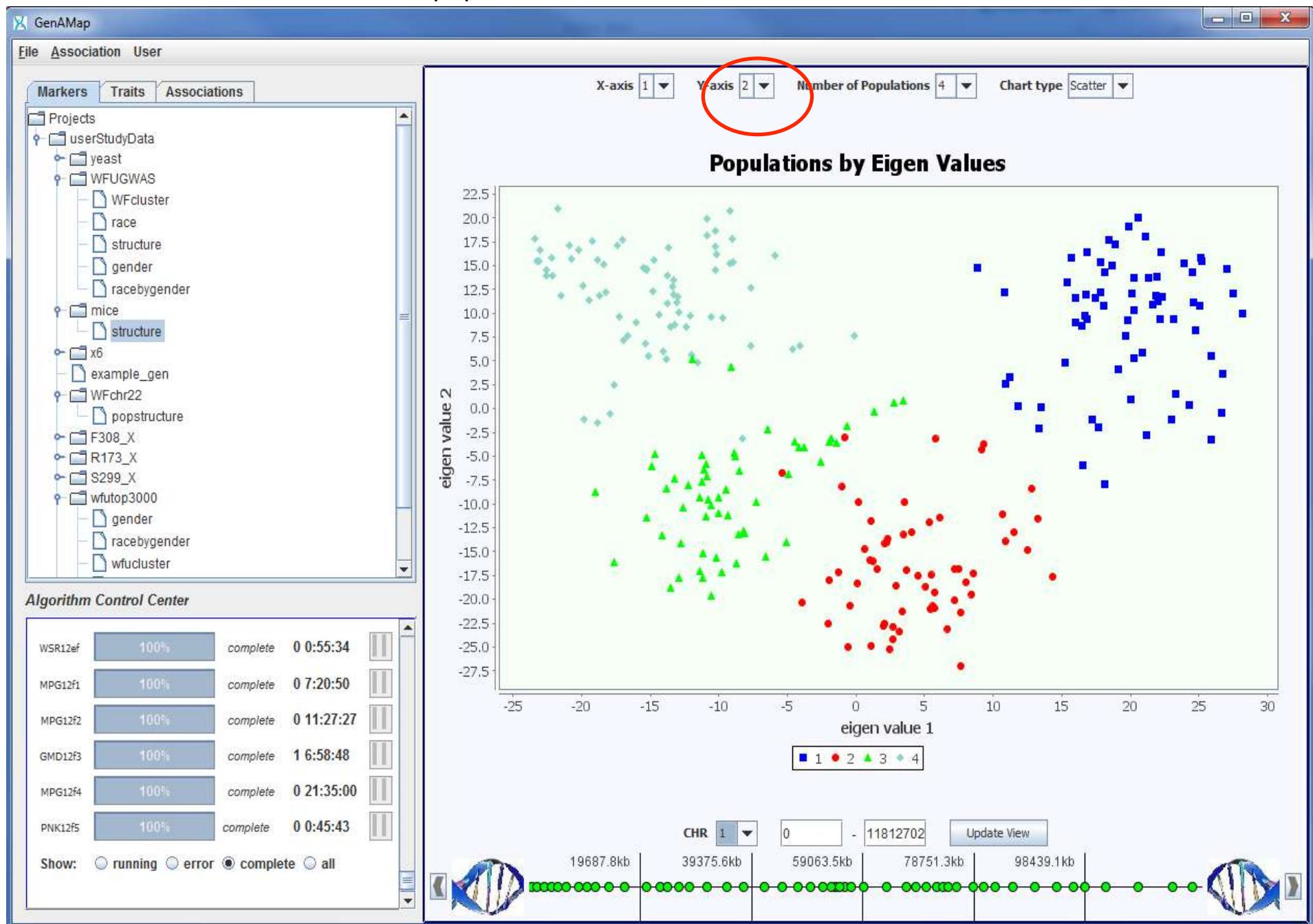
Mouse results from structure – 2 populations



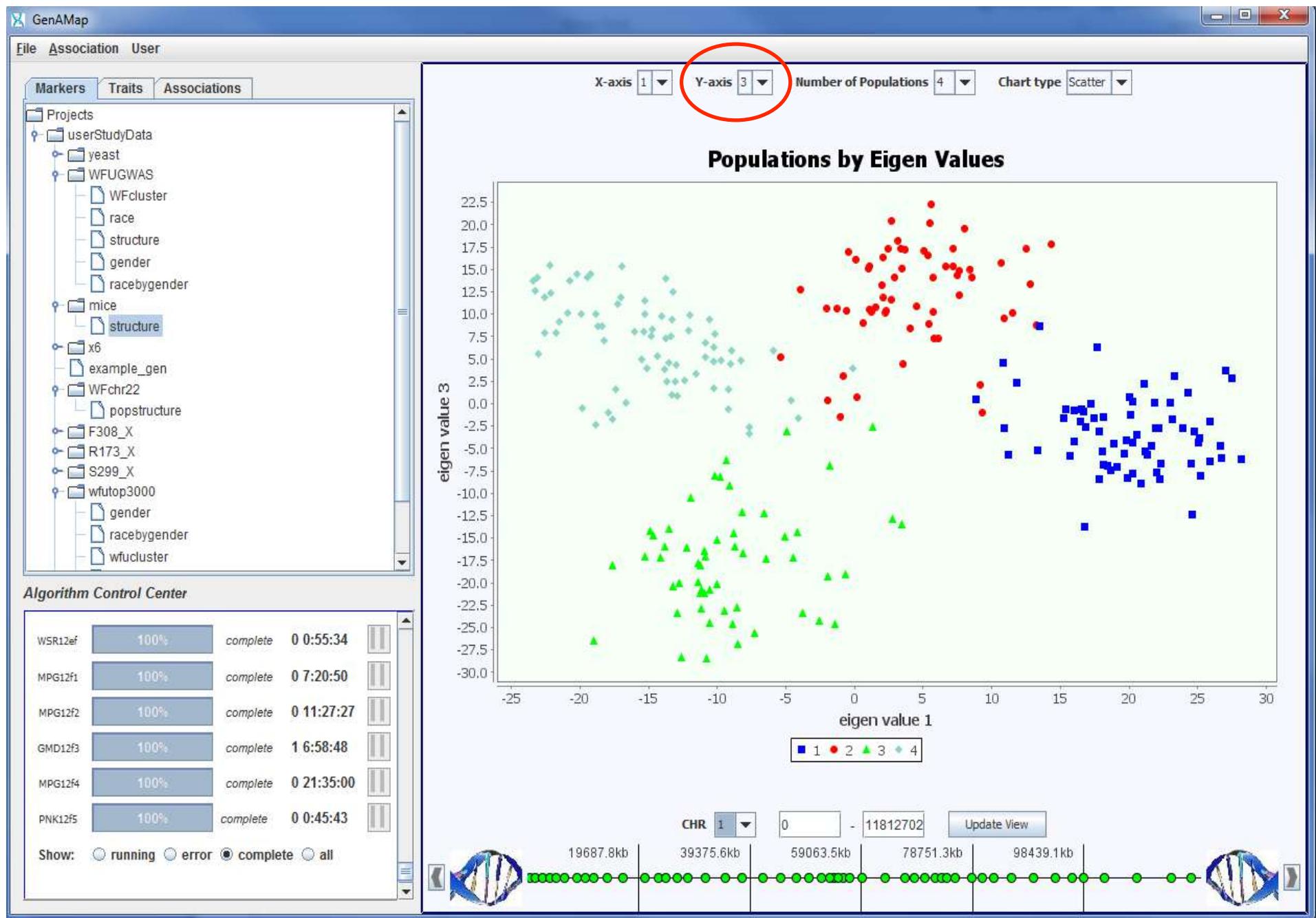
Mouse results from structure – 3 populations



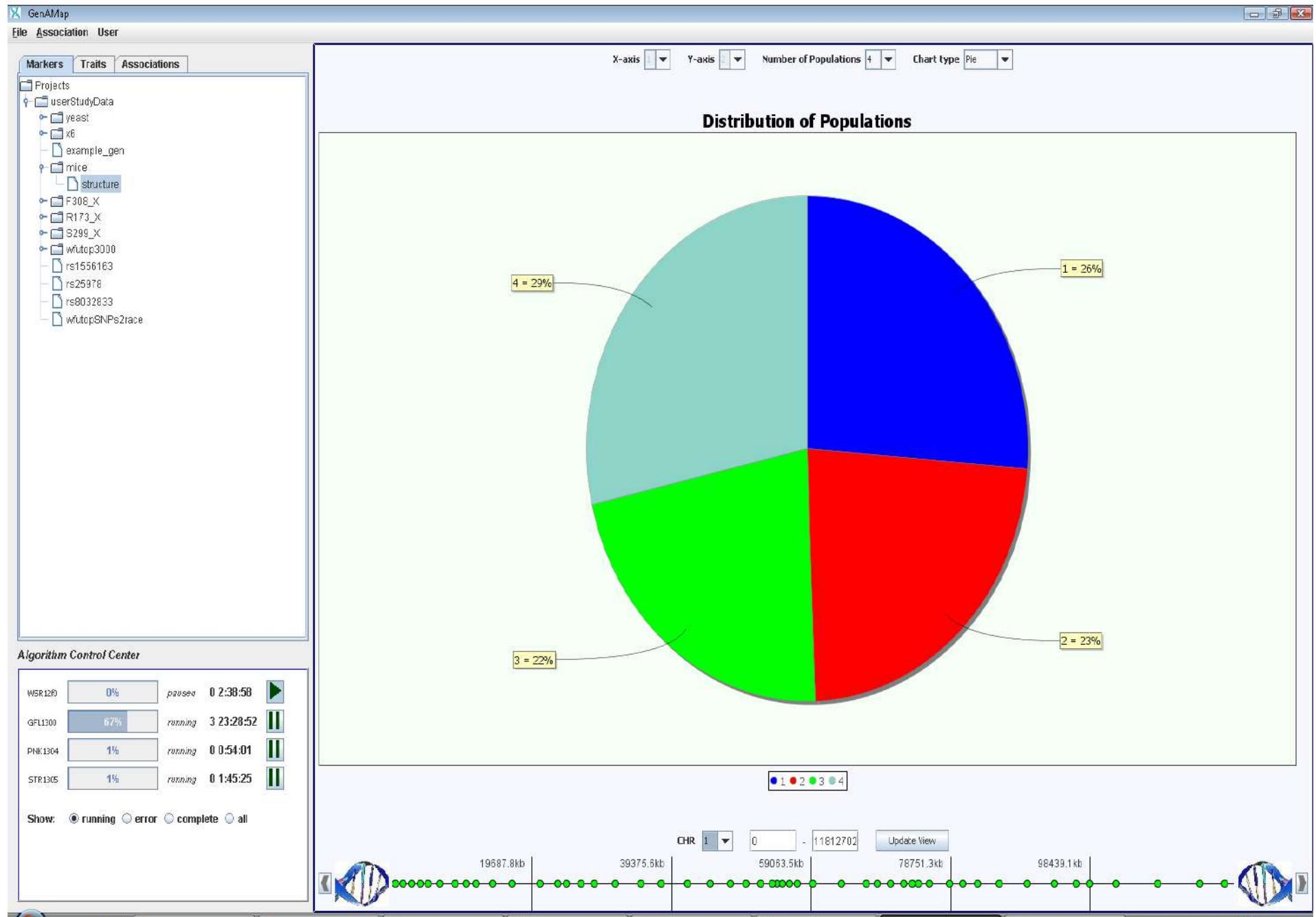
Mouse results from structure – 4 populations



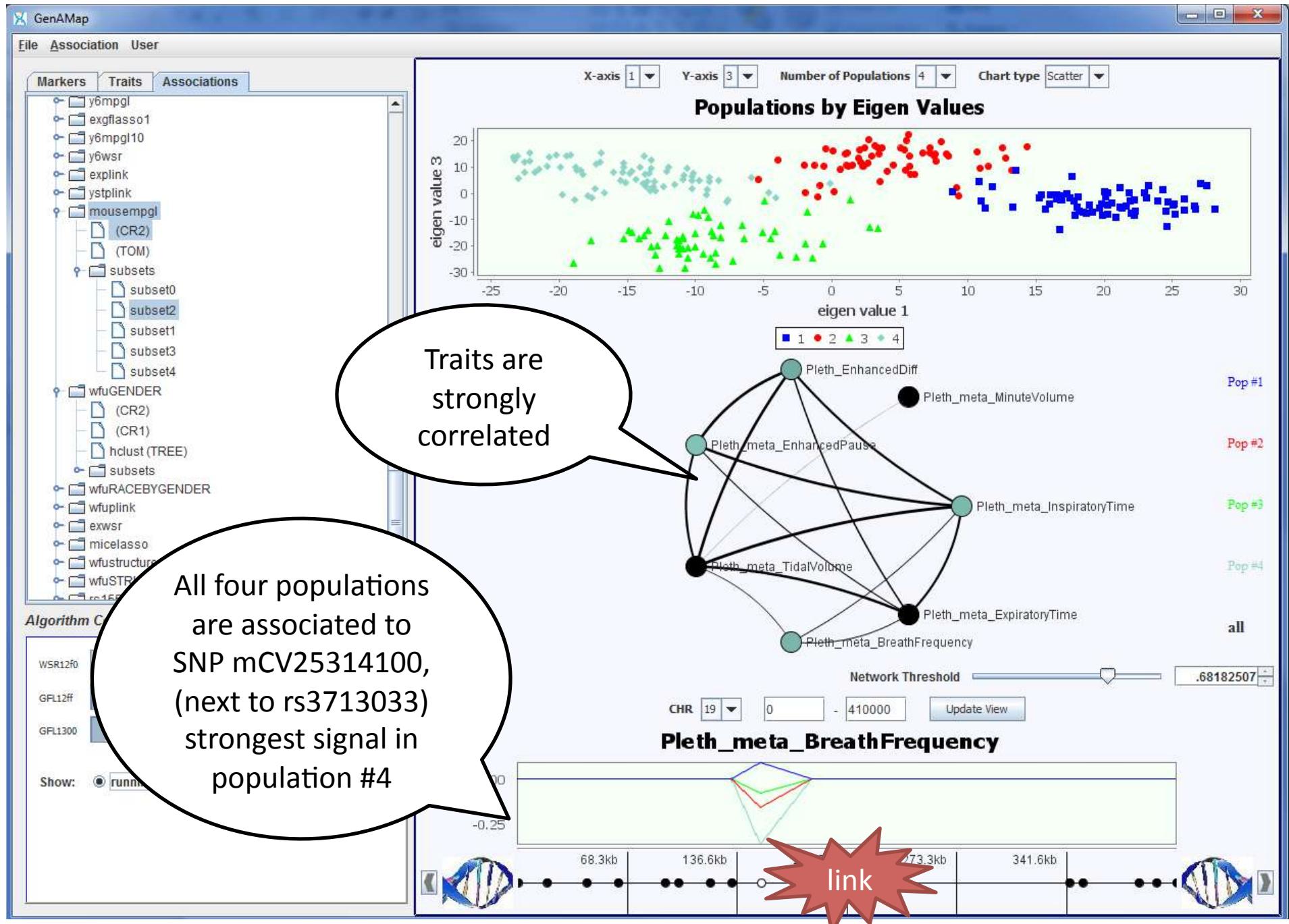
Mouse results from structure – 4 populations



Mouse results from structure – 4 populations



Mouse results from MPGL



dbSNP

www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?searchType=adhoc_search&type=rs&rs=rs3713033

Clinical Significance: NA
MAF/MinorAlleleCount: NA
MAF Source:

SNP Details are organized in the following sections:

GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Group term	Group label	Contig label	Nei S
37.1	19	4898888	NW_001030639.1	1898888	-	A	+	Primary Assembly	Mm_Celera	Mm_Celera	
37.1	19	5029791	NT_082868.6	2029791	-	G	+	Primary Assembly	MGSCv37	MGSCv37	

GeneView

GeneView via analysis of contig annotation: **Slc29a2** solute carrier family 29 (nucleoside transporters), member 2

▼ View more variation on this gene (click to hide).
 Include clinically associated: in gene region cSNP has frequency double hit

Primary Assembly Mapping

Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position
Mm_Celera	-	19	4898888	NW_001030639.1	1898888

Function class:
rs3713033 is located in the intron region of NM_007854.2.

1,898,286 : 1,899,125 (1,201 bases shown, positive strand)

Flip Strands |

[1,898,300] [1,898,400] [1,898,500] [1,898,600] [1,898,700] [1,898,800] rs3713033 [1,899 K] [1,899,100] [1,899,200] [1,899,300]

Sequence NW_001030639.1: Mus musculus strain mixed chromosome 19 scaffold, alternate assembly Mm_Celera 232000009819958, whole genome shotgun sequence

SNP

Genes

Slc29a2/NM_007854.3/NP_031880.2

GeneView via direct blast against RefSeq sequences (used when no gene model is available): N/A

Submitter records for this RefSNP Cluster

The submission **ss44990541** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs3713033** during BLAST analysis.

www.ncbi.nlm.nih.gov/nuccore/NM_007854

NCBI Resources How To My NCBI Sign In

Nucleotide

Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Display Settings: GenBank Send:

Mus musculus solute carrier family 29 (nucleoside transporters), member 2 (Slc29a2), mRNA

NCBI Reference Sequence: NM_007854.3

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NM_007854 2490 bp mRNA linear ROD 12-MAR-2011

DEFINITION Mus musculus solute carrier family 29 (nucleoside transporters), member 2 (Slc29a2), mRNA.

ACCESSION NM_007854

VERSION NM_007854.3 GI:194248085

KEYWORDS .

SOURCE Mus musculus (house mouse)

ORGANISM [Mus musculus](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.

REFERENCE 1 (bases 1 to 2490)

AUTHORS Diez-Roux,G., Banfi,S., Sultan,M., Geffers,L., Anand,S., Rozado,D., Magen,A., Canidio,E., Pagani,M., Peluso,I., Lin-Marq,N., Koch,M., Bilio,M., Cantiello,I., Verde,R., De Masi,C., Bianchi,S.A., Cicchini,J., Perroud,E., Mehmeti,S., Dagand,E., Schrinner,S., Nurnberger,A., Schmidt,K., Metz,K., Zwingmann,C., Brieske,N., Springer,C., Hernandez,A.M., Herzog,S., Grabbe,F., Sieverding,C., Fischer,B., Schrader,K., Brockmeyer,M., Dettmer,S., Helbig,C., Alunni,V., Battaini,M.A., Mura,C., Henrichsen,C.N., Garcia-Torres,P., Echavarria,D., Duellon,F., Garcia-Gallego,F.

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

Articles about the Slc29a2 gene

A high-resolution anatomical atlas of the transcriptome in the mouse embryo [PLoS Biol. 2011]

Tissue distribution, ontogeny, and hormonal regulation of xenobiotic [Drug Metab Dispos. 2009]

Hypoxia-inducible factor-dependent repression of equilibrative nucleoside 1[Gastroenterology. 2008]

See all articles

Reference sequence information

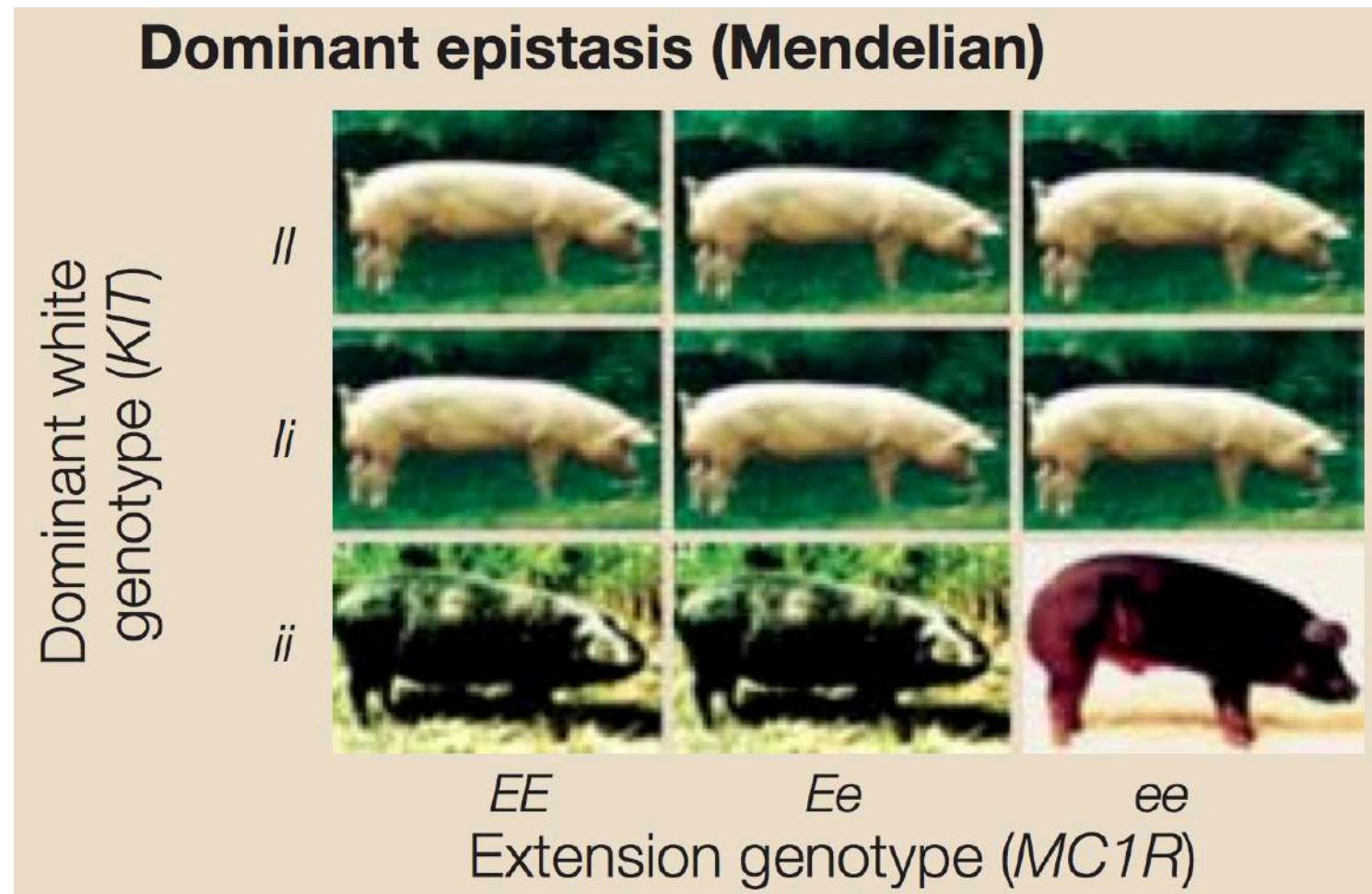
RefSeq protein product

See the reference protein sequence for equilibrative nucleoside transporter 2 (NP_031880.2).

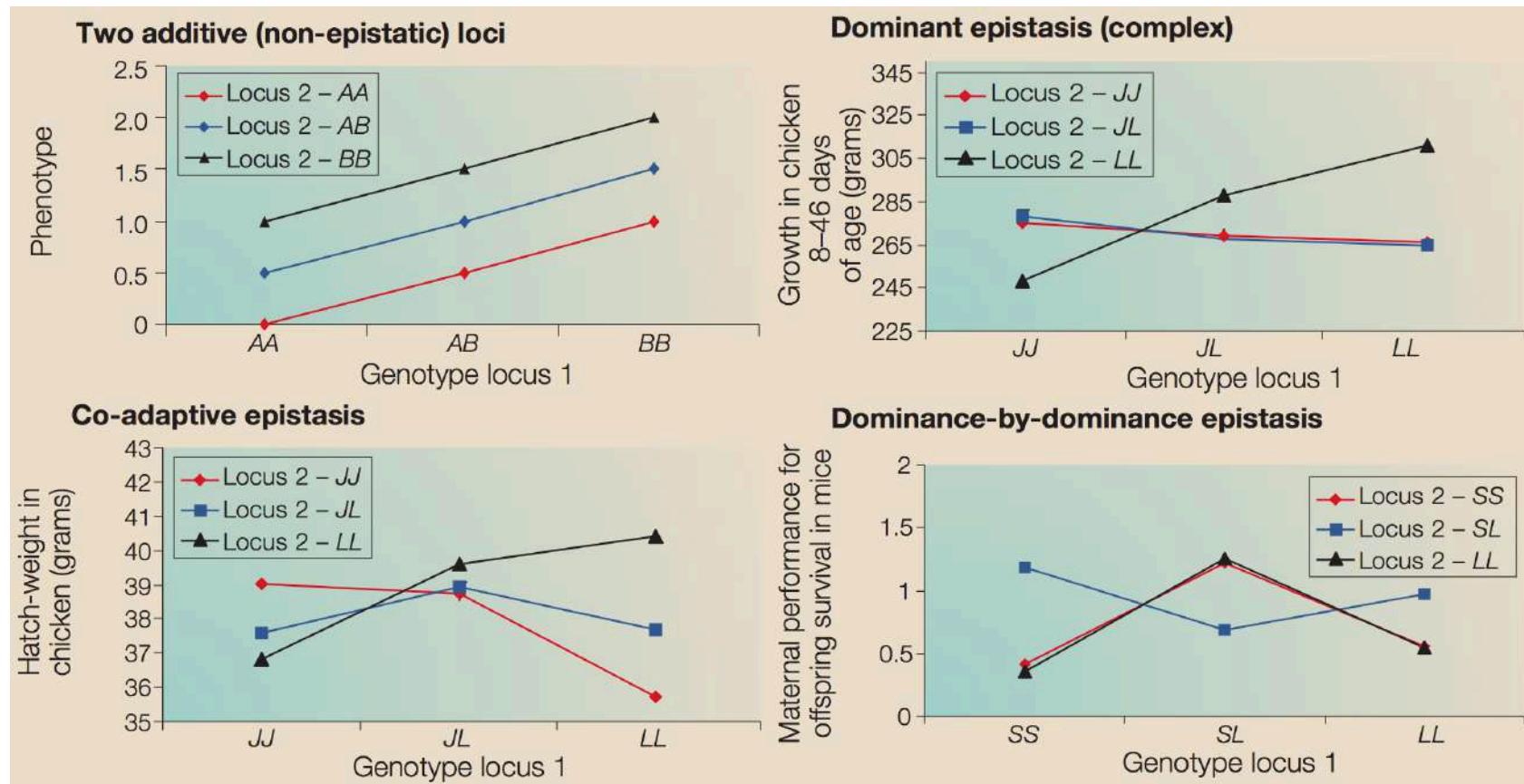
Epistasis

- Definition: The effect of one locus depends on the genotype of another locus

Epistasis for Mendelian Traits



Epistasis for Complex Traits



Epistasis

- Epistatic effects of SNPs can often be detected only if the interacting SNPs are considered jointly
 - The number of candidate SNP interactions is very large
 - For J SNPs, $J \times J$ SNP pairs need to be considered for epistasis
 - In general for J SNPs and K -way interactions, there are $O(J^K)$ candidate interactions
 - Computationally expensive to consider all possible groups of interacting SNPs
 - For a reliable detection of K -way interactions, a large sample size is required
 - Multiple testing problem

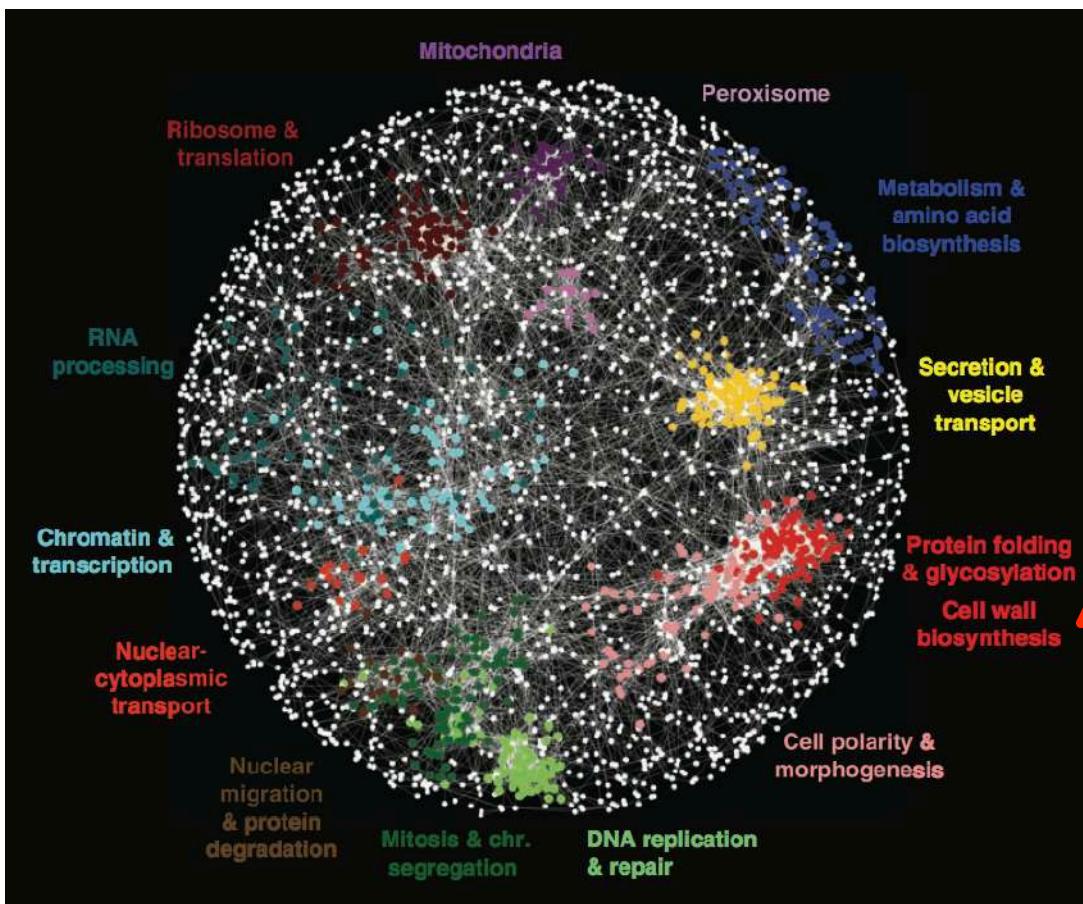
Two-stage Lasso for Detecting Epistasis

(Wu et al. Genetic Epidemiology 2010)

- Two-stage lasso
 - Stage 1: Apply lasso with no consideration of epistasis to detect SNPs with significant individual effects
 - Stage 2: Apply lasso with pairs of only those SNPs selected in Stage 1
- Reduces the computational burden
- Limitation: SNPs with epistatic effects often do not have detectable individual (marginal) effects, and many of these association signals will be missed.

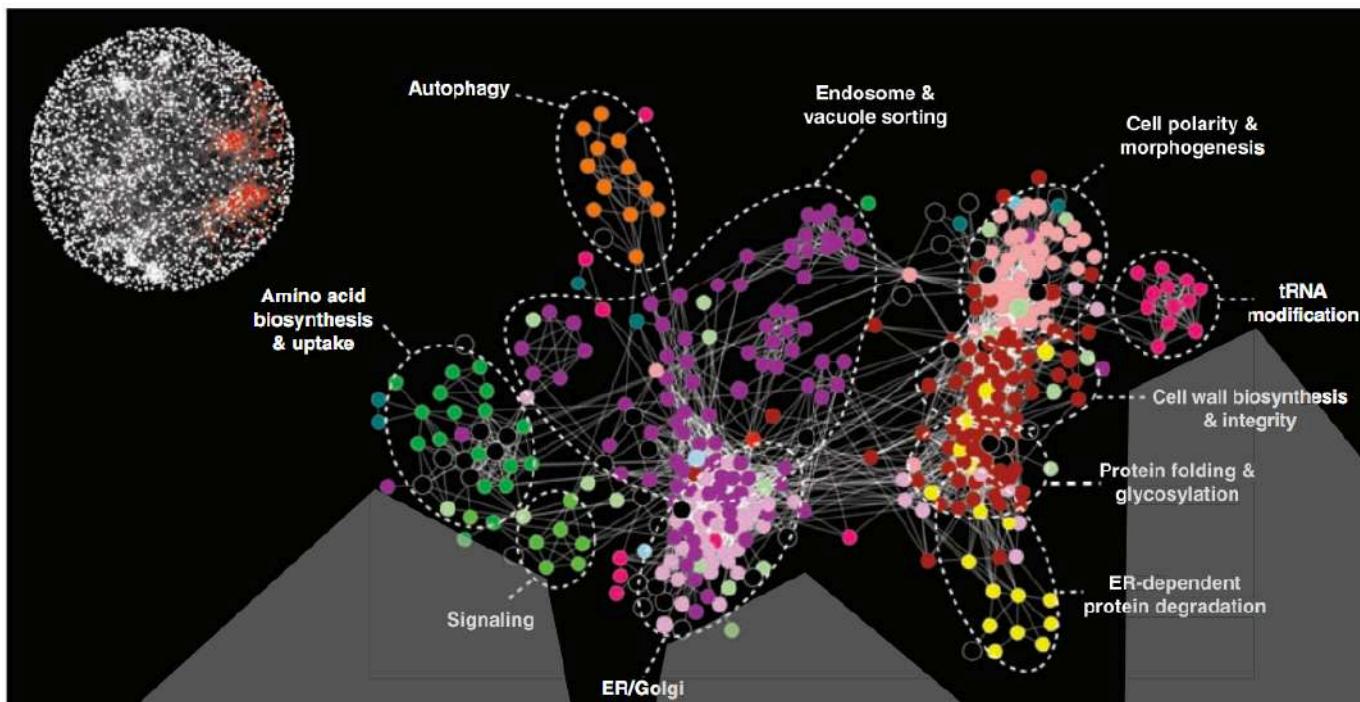
Epistasis: Exploiting Prior Knowledge

- Synthetic genetic interaction network from double-knockout experiments provide strong evidence of gene-gene interaction



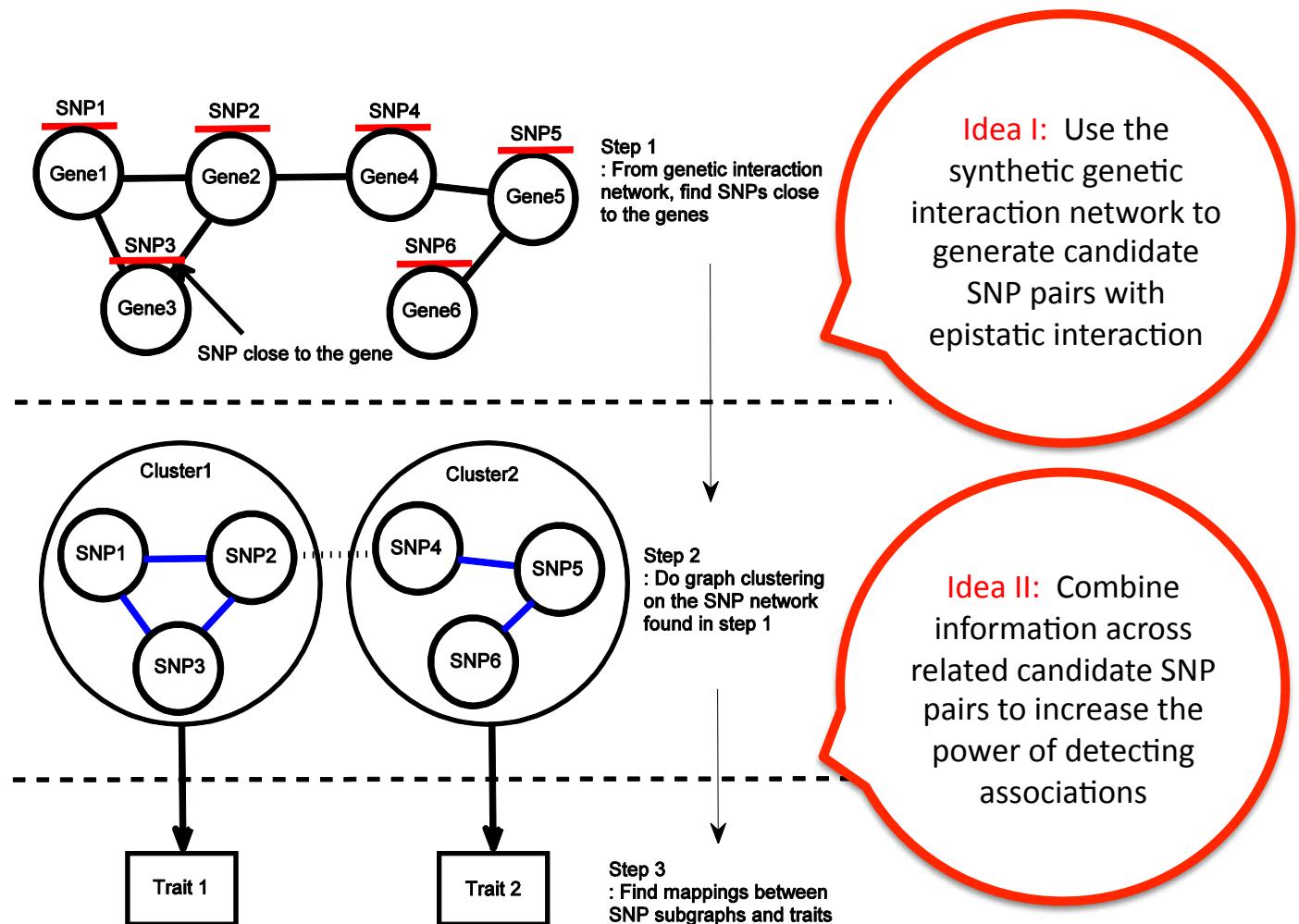
Epistasis: Exploiting Prior Knowledge

- Synthetic genetic interaction network from double-knockout experiments provide strong evidence of gene-gene interaction



Idea II: Combine information across related candidate SNP pairs to increase the power of detecting associations

Epistasis: Exploiting Prior Knowledge



Epistasis: Structured Input/Ouput (IO)-Lasso

$$\beta_{io\text{-}lasso} = \arg \min_{\beta} \sum_{k=1}^K \sum_{i=1}^N \left(Y_i^k - \sum_{j=1}^p \beta_j^k X_{ij} - \sum_{(r,s) \in U} \beta_{rs}^k Z_{i,rs} \right) + \lambda_1 \sum_{j=1} \sum_{k=1} |\beta_j^k| + \lambda_2 \sum_k \sum_m \sqrt{\sum_{(r,s) \in S_m} \beta_{rs}^{k^2}} + \lambda_3 \sum_j \sqrt{\sum_k \beta_j^{k^2}} + \lambda_4 \sum_k \sum_{(r,s) \in U} |\beta_{rs}^k|$$

Idea I: Use the synthetic genetic interaction network to generate candidate SNP pairs with epistatic interaction

Idea II: Combine information across related candidate SNP pairs to increase the power of detecting associations

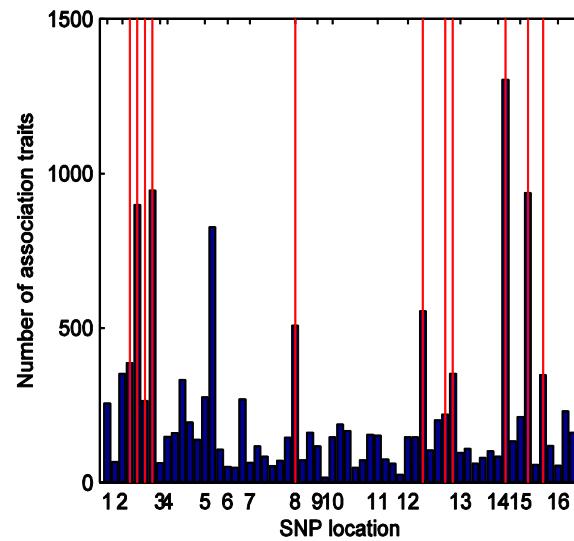
U : genetic interaction networks

S_m : m^{th} cluster in SNP network

Results on Yeast eQTL Dataset

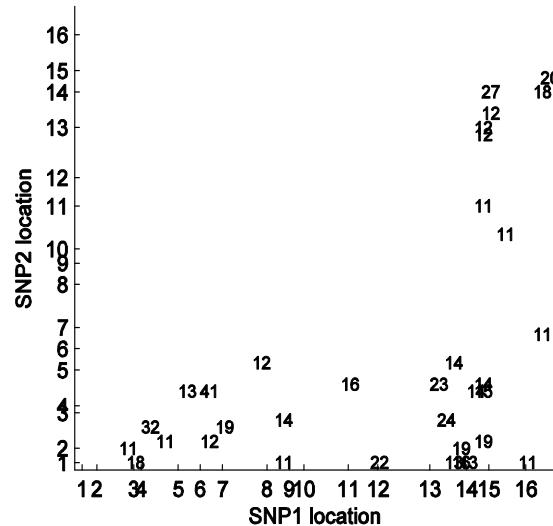
- Yeast eQTL dataset (Brem & Kruglyak, PNAS 2005)
 - Genotypes: 114 yeast strains from the cross of laboratory strains and wild strains, 1260 non-redundant SNPs
 - Phenotypes: gene-expression levels for 5,637 genes
 - 2,219 clusters of the traits
- Yeast genetic interactions (Costanzo et al., Science 2010)
 - Yeast contains ~6,000 genes
 - Synthetic genetic interactions for 5.4 million gene-gene pairs
 - We extracted 73,720 significant SNP pairs based on the results from synthetic lethal interaction experiment (0.09% of all SNP pairs)
 - We further selected 9,388 representative SNP pairs from these pairs of SNPs so that the correlation among the selected SNPs is low

eQTL hotspots of individual SNP effects

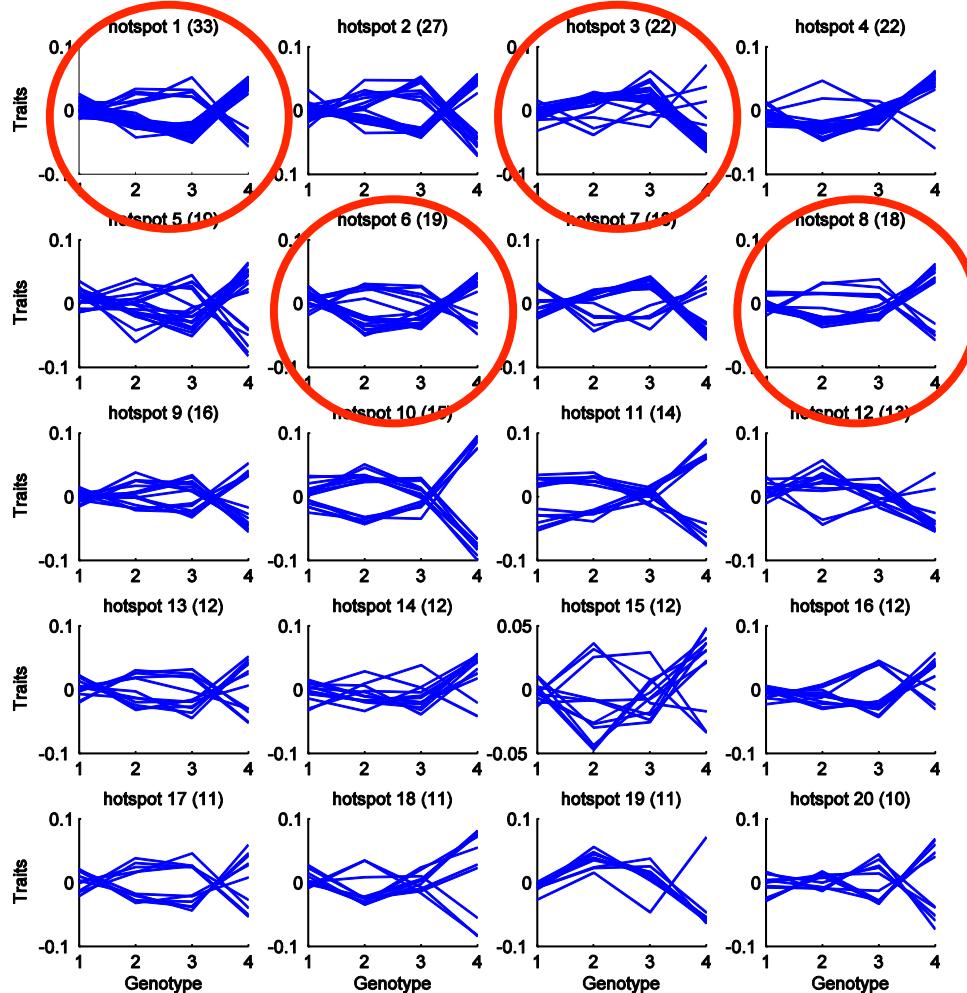


Red lines:
Previously reported eQTL hotspots
[Yvert et al. 2003]

eQTL hotspots of epistatically interacting SNPs



Top 20 Epistatic Hotspots



Overview

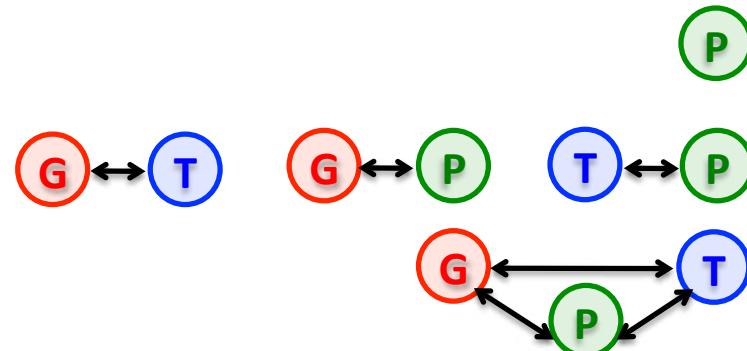
- **Genome structure** in association analysis
 - Linkage disequilibrium
 - Population structure
 - Epistasis

G

- **Transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Leveraging gene expression tree

T

- **Phenome structure** in association analysis
 - Pleiotropy
 - Dynamic trait
- **Two-way** structured association
- **Three-way** structured association
- Visualization software

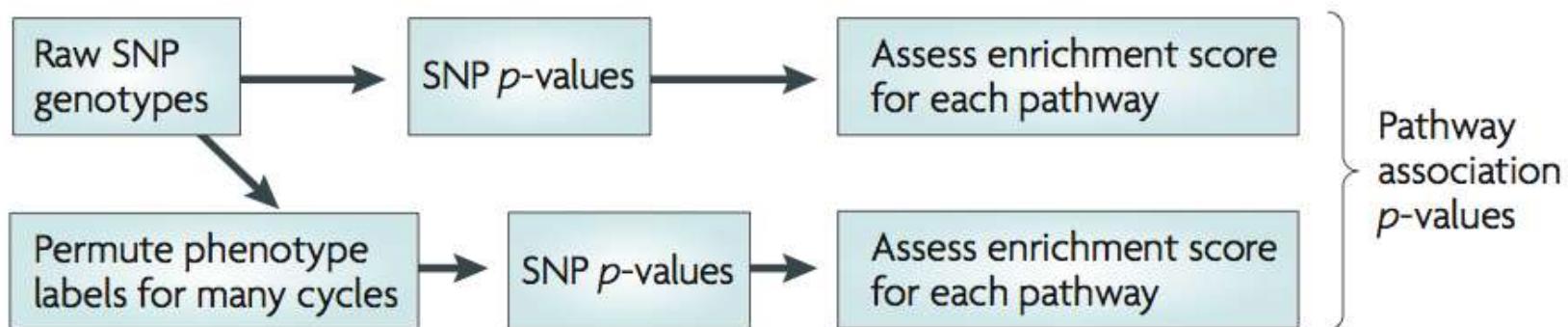


Association to the Transcriptome

- Expression quantitative trait locus (eQTL) analysis
 - Gene expression traits as phenotypes in association analysis
 - Typically involves more than 5000 genes
 - Univariate regression analysis and lasso could be applied to each gene expression trait
- Pleiotropic effects of SNPs
 - Definition: A genetic variation influences multiple phenotypes jointly
 - Instead of assessing the effect of genotype on a single phenotype, we consider multiple related phenotypes (e.g., genes in the same pathway) to detect pleiotropic effects

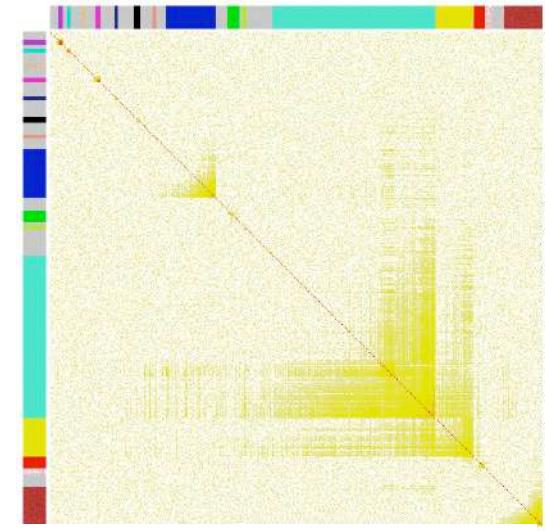
Computational Methods for Detecting Pleiotropy

- **Method 1:** Aggregating the results from a set of single-gene (trait) analyses



- **Limitation:** It does not directly search for pleiotropic effects, but looks for pleiotropy as a post-processing step after single-trait analyses

Yeast eQTL Analysis: Association to a Gene Module

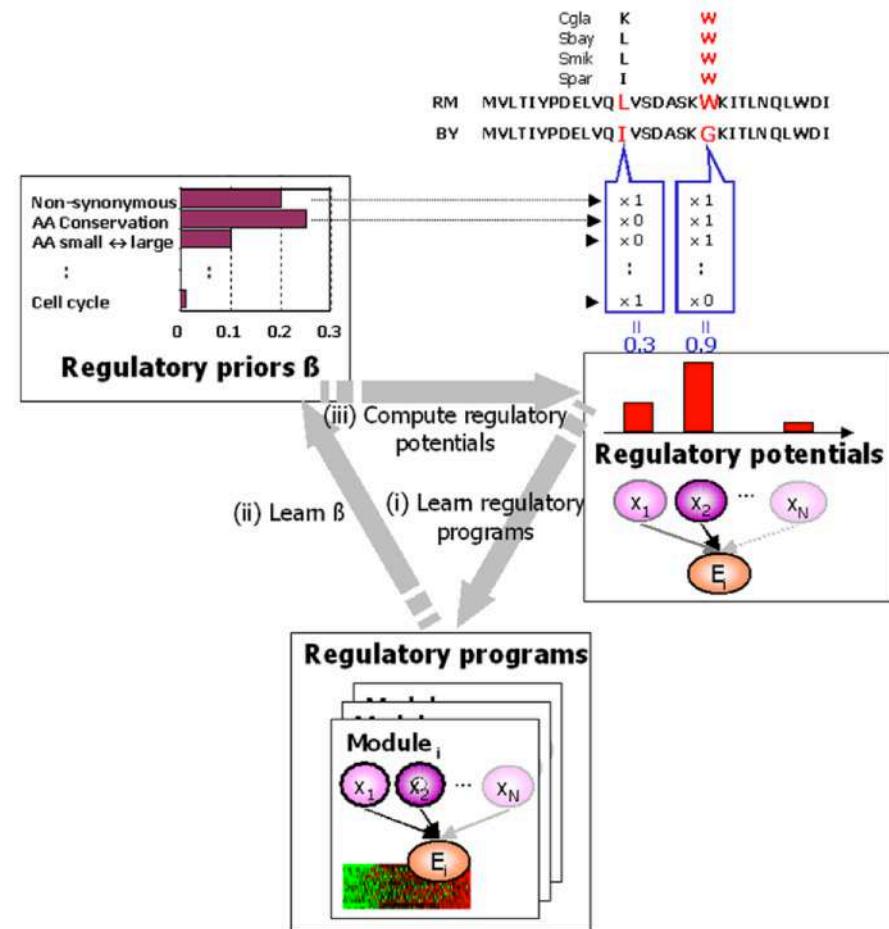


- Yeast eQTL dataset
 - Genotypes from 112 segregants from a yeast cross between BY and RM strains
 - Microarray gene-expression data

Module color ^a	Module size	GO category type ^b	GO category	GO category size (overlap)	GO enrichment nominal P value ^c	Chr.	Within-chr.		
							genome coordinate	eQTL hot spot size (overlap)	eQTL enrichment nominal P value ^c
Turquoise	1,208	BP	Cytoplasm organization and biogenesis	169 (153)	7.47×10^{-58}	2	550,000	186 (144)	2.87×10^{-39}
Blue	369	BP	Organic acid metabolism	235 (94)	4.44×10^{-34}	3	70,000	50 (46)	1.61×10^{-42}
Brown	290	BP	Protein biosynthesis	292 (98)	2.63×10^{-41}	14	450,000	206 (107)	2.92×10^{-69}
Yellow	282	BP	Generation of precursor metabolites and energy	258 (46)	2.09×10^{-8}	15	170,000	182 (134)	6.06×10^{-122}
Green	84	MF	Transferase activity	428 (20)	0.0012	2	570,000	25 (5)	0.00021
Red	83	BP	Generation of precursor metabolites and energy	168 (44)	6.54×10^{-39}	15	570,000	25 (21)	2.28×10^{-32}
Black	44	BP	Lipid metabolism	149 (20)	3.31×10^{-17}	12	650,000	52 (34)	2.37×10^{-60}
Pink	43	BP	Intracellular transport	275 (14)	1.41×10^{-6}	14	450,000	206 (19)	1.92×10^{-13}
Magenta	39	MF	RNA binding	140 (18)	3.21×10^{-16}	8	90,000	31 (4)	0.00028
Purple	37	BP	Chromosome organization and biogenesis (sensu Eukaryota)	200 (7)	0.0033	12	1,050,000	38 (31)	9.07×10^{-64}
Green-yellow	31	CC	Endoplasmic reticulum	213 (15)	2.39×10^{-11}	12	670,000	68 (5)	0.00022
Tan	29	BP	Response to chemical stimulus	153 (3)	0.12	5	110,000	24 (13)	4.84×10^{-23}
Cyan	27	BP	Response to chemical stimulus	153 (9)	7.55×10^{-7}	3	210,000	33 (23)	5.26×10^{-56}
Salmon	27	BP	Biopolymer catabolism	140 (13)	2.71×10^{-12}	12	670,000	68 (2)	0.089
Midnight blue	23	BP	Reproduction	160 (15)	7.64×10^{-16}	8	110,000	38 (23)	4.66×10^{-50}

Computational Methods for Detecting Pleiotropy

- **Method 2:** *Lirnet* finds SNPs associated with average over phenotypes in each module
- Joint estimation of
 - Regulatory modules
 - SNPs associated with each regulatory module
 - Regulatory potentials for the importance weights of prior knowledge on SNPs
- **Limitation:** *Lirnet* uses averaged phenotypes rather than the raw phenotype values
 - Loss of information
 - What about anti-correlated phenotypes?



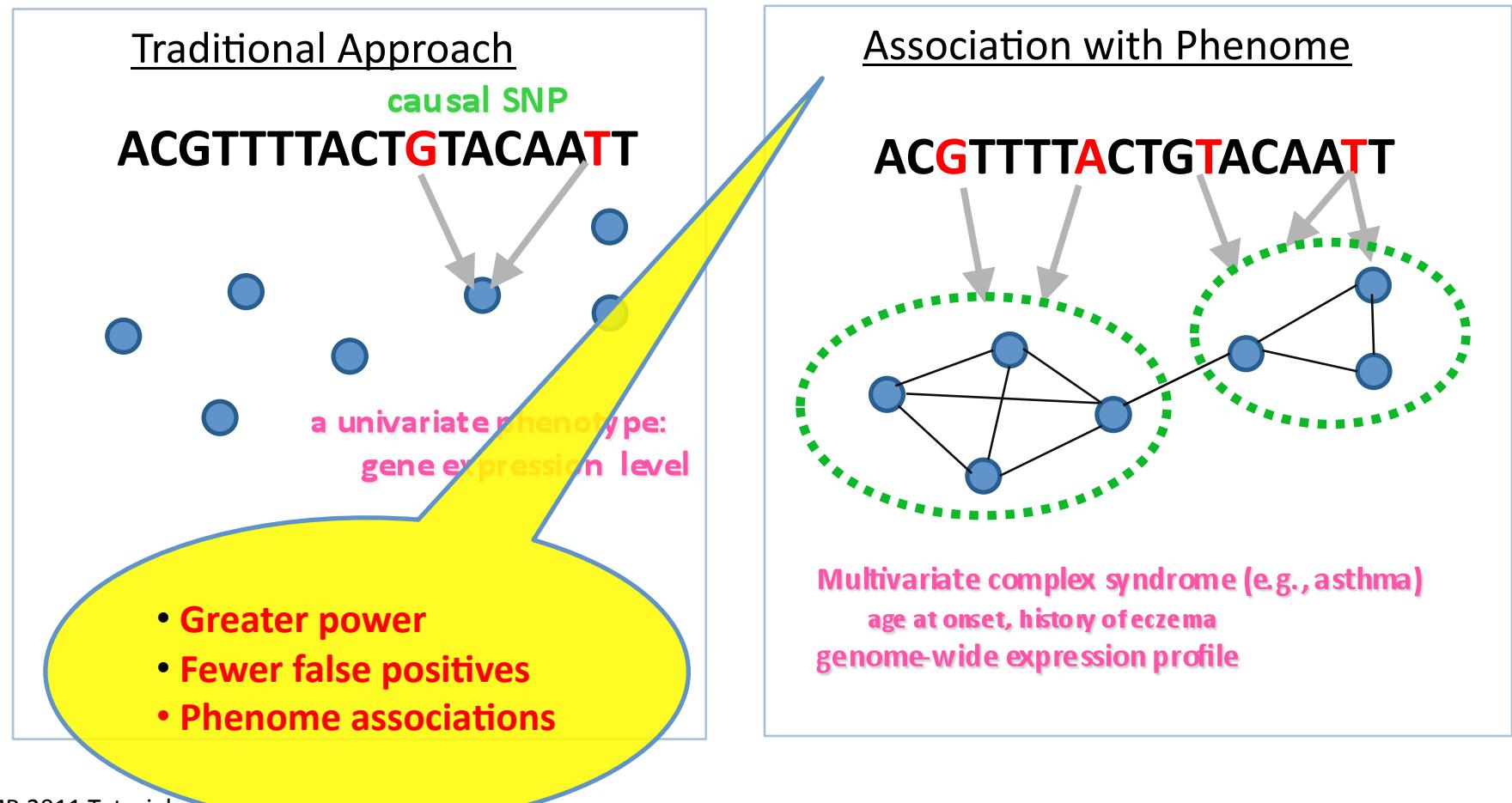
Computational Methods for Detecting Pleiotropy

- **Method 3:** Apply PCA to phenotype data and perform association analysis on the projected phenotype data
(Weller et al., Theo. Appl. Gen. 1996, Mangin et al., Biometrics 1998)
 - **Limitation:** It is not obvious how to interpret the projected phenotype values and their associations to genotypes

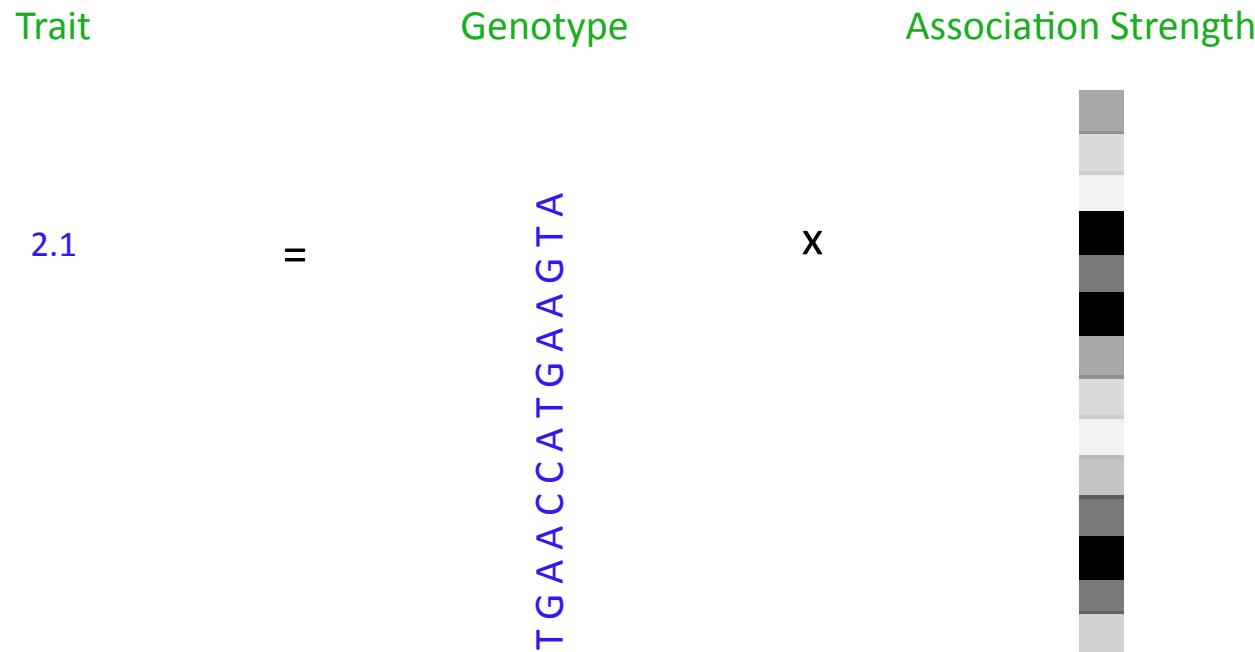
Computational Methods for Detecting Pleiotropy

- **Method 4:** Structured multi-task sparse regression that incorporate transcriptome structure as prior knowledge
 - Graph-guided fused lasso: gene-gene interaction as graph
(Kim & Xing, PLoS Genetics 2009)
 - Tree-guided group lasso: gene modules as hierarchical clustering tree
(Kim & Xing, ICML 2010)
- Extending generic toolbox in statistics for association analysis
 - Convex optimization (Boyd and Vandenberghe, Cambridge University Press 2004)
 - Fast computation for complex algorithms
 - Penalized regression method for structured sparsity
 - Increase the power of detecting associations

Structured Association



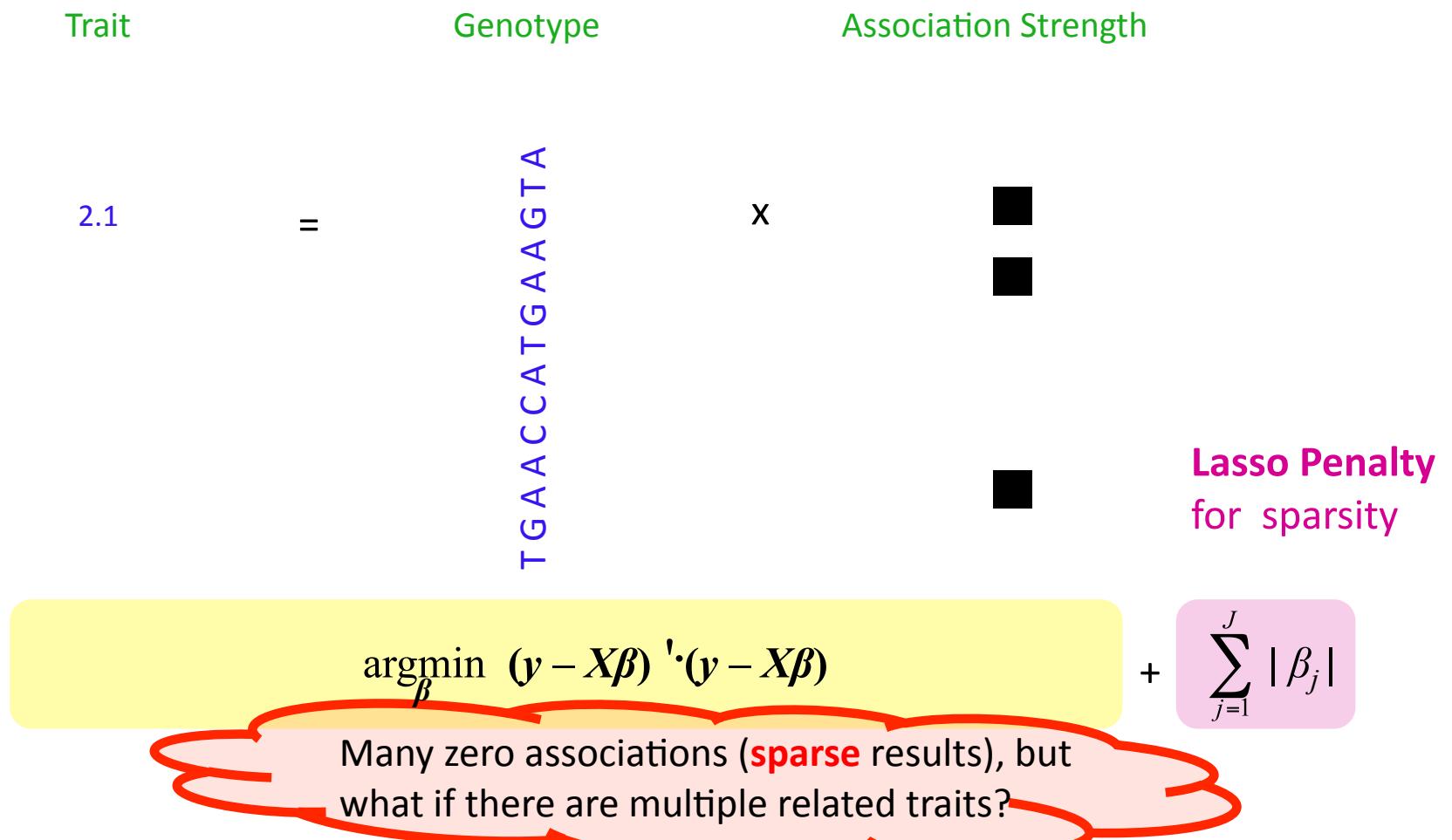
Regression for Single-Trait Association Analysis



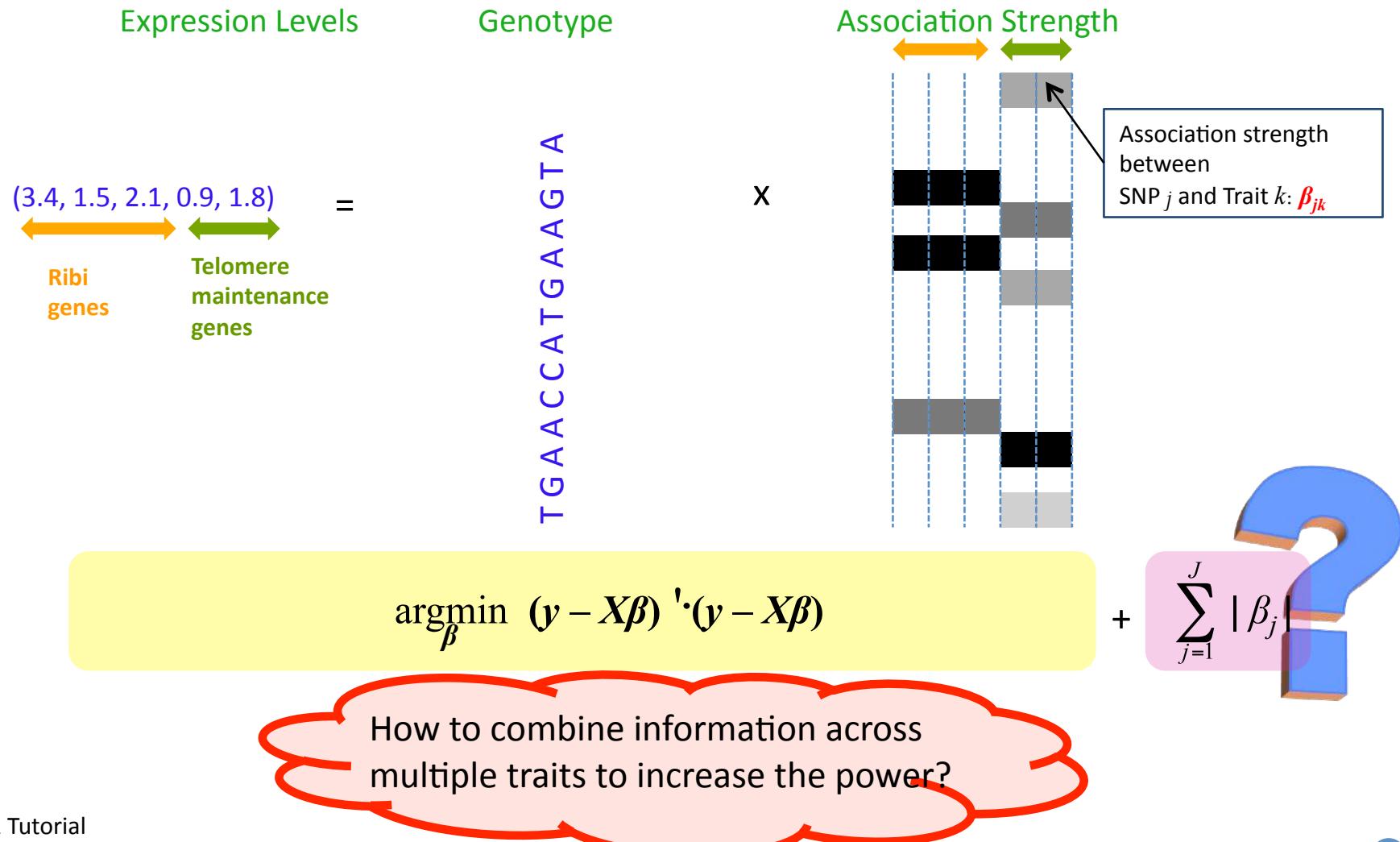
$\text{argmin}_{\beta} (y - X\beta)^T (y - X\beta)$

Many non-zero associations:
Which SNPs are truly significant?

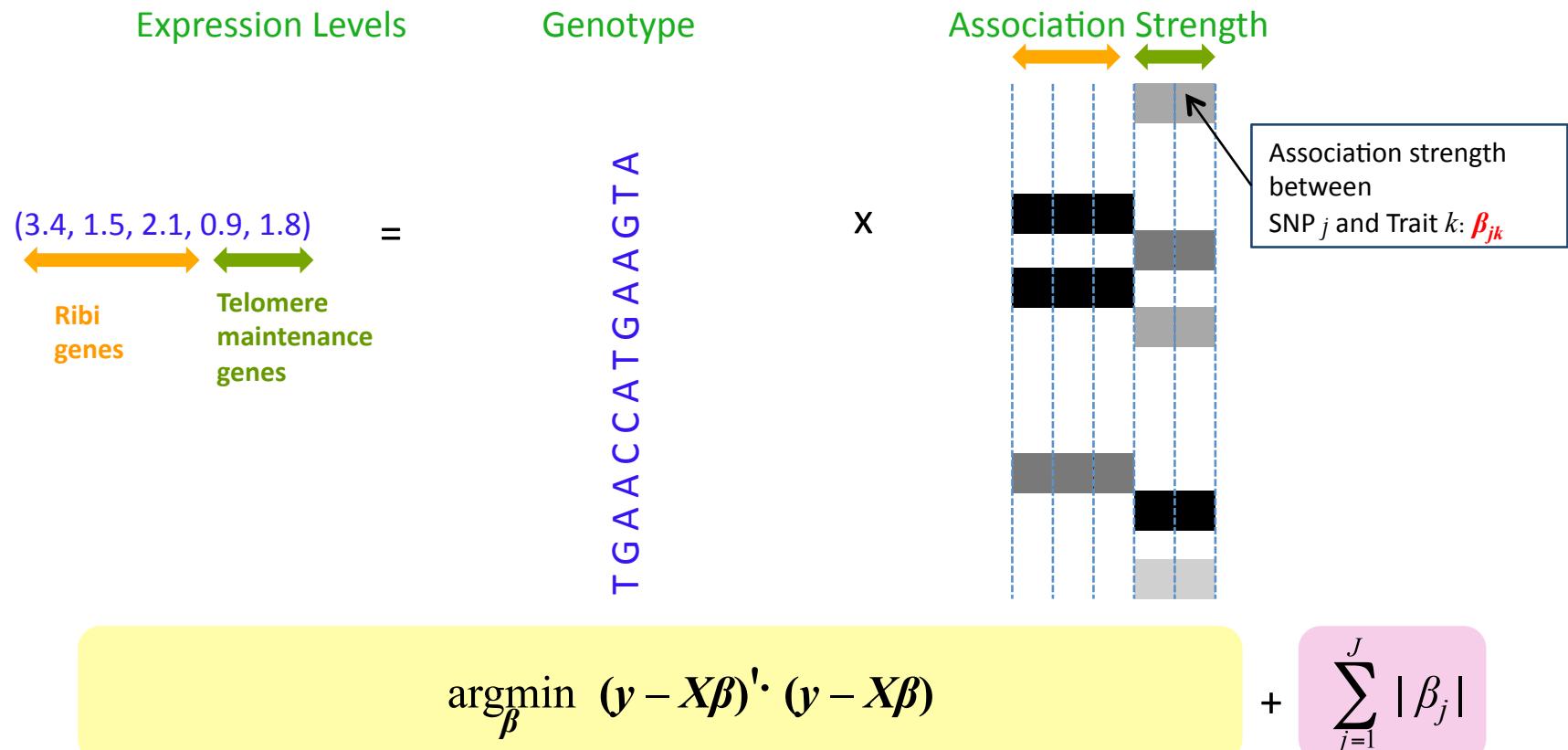
Lasso for Reducing False Positives



Multi-Task Regression for Transcriptome Association Analysis



Multi-Task Regression for Transcriptome Association Analysis

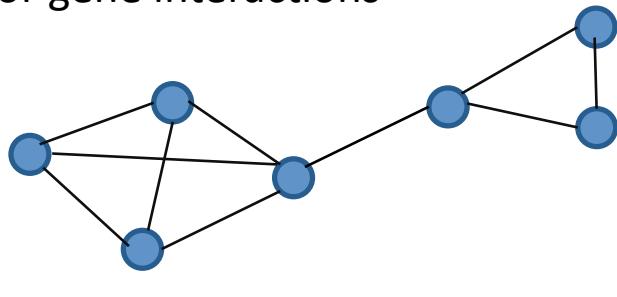


We introduce
graph-guided fusion penalty

Kim & Xing, PloS Genetics 2009

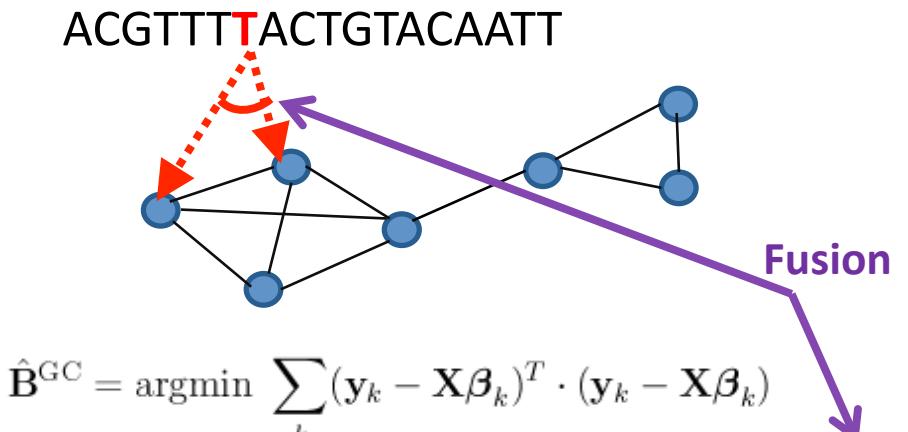
Graph-Constrained Fused Lasso

Step 1: Thresholded correlation graph
for gene interactions



Step 2: Graph-constrained fused lasso

ACGTTT**T**ACTGTACAATT

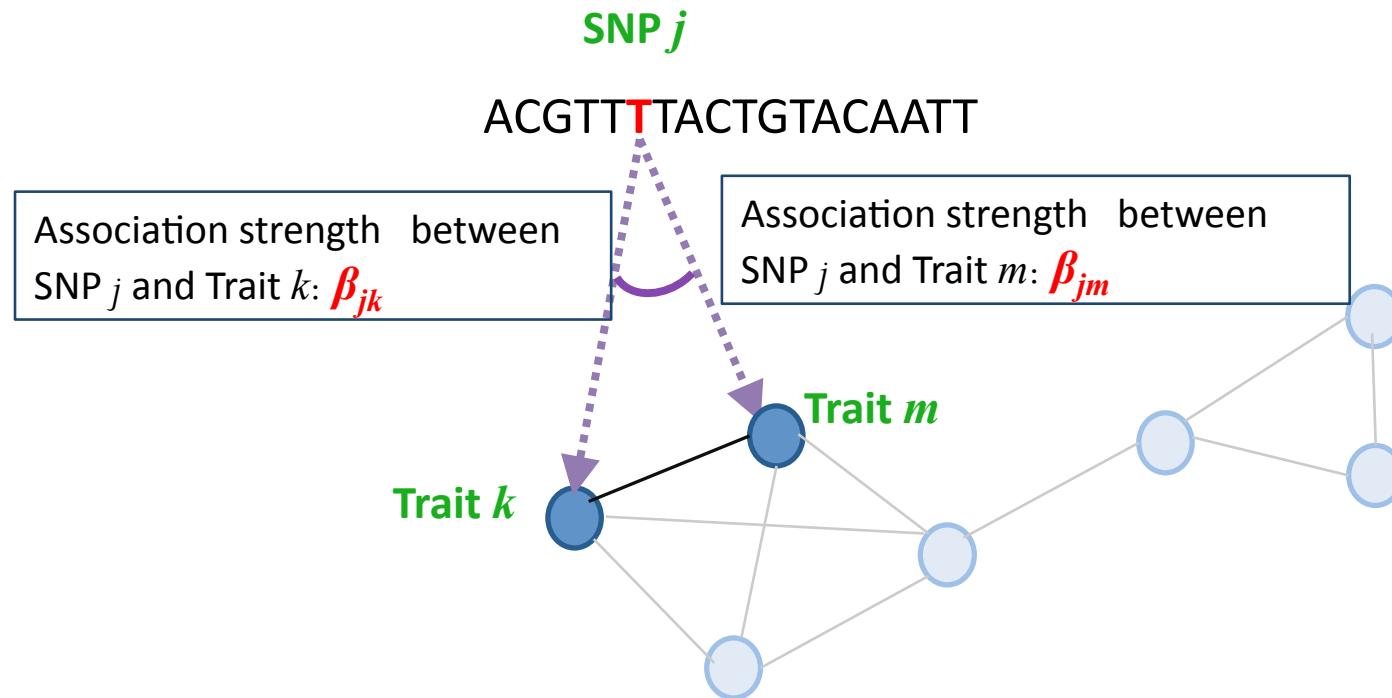


$$\hat{\mathbf{B}}^{GC} = \operatorname{argmin}_k \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l) \in E} \sum_j |\beta_{jm} - \operatorname{sign}(r_{ml})\beta_{jl}|$$

Lasso
Penalty

Graph-constrained fusion
penalty

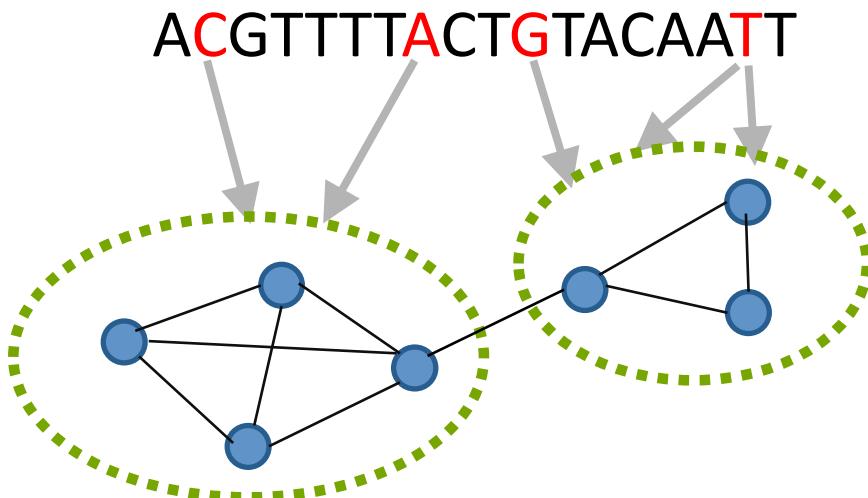
Fusion Penalty



- Fusion Penalty: $|\beta_{jk} - \beta_{jm}|$
- For two correlated traits (connected in the network), the association strengths may have similar values.

Transcriptome Association: Graph-Constrained Fused Lasso

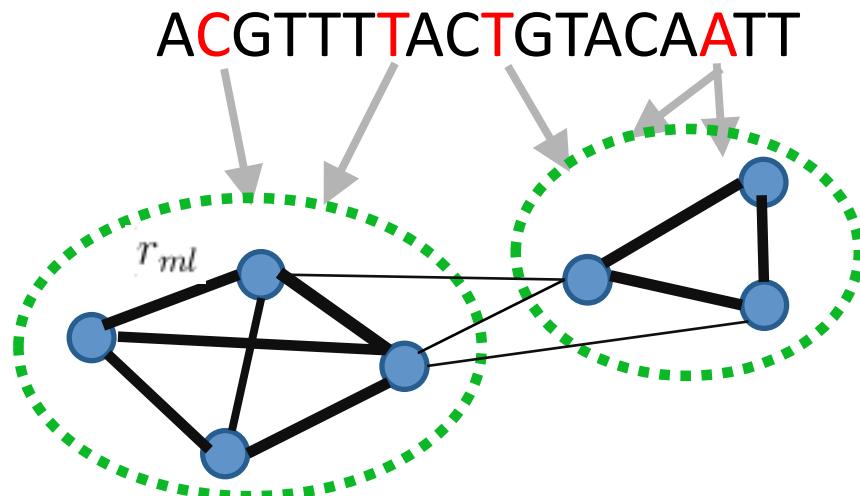
Overall effect



- Fusion effect propagates to the entire network
- Association between SNPs and subnetworks of traits

Transcriptome Association: Graph-Weighted Fused Lasso

Overall effect



- Subnetwork structure is embedded as a densely connected nodes with **large edge weights**
- Edges with small weights are effectively ignored

Estimating Parameters

- Quadratic programming formulation

$$\hat{\mathbf{B}}^{\text{GW}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)$$

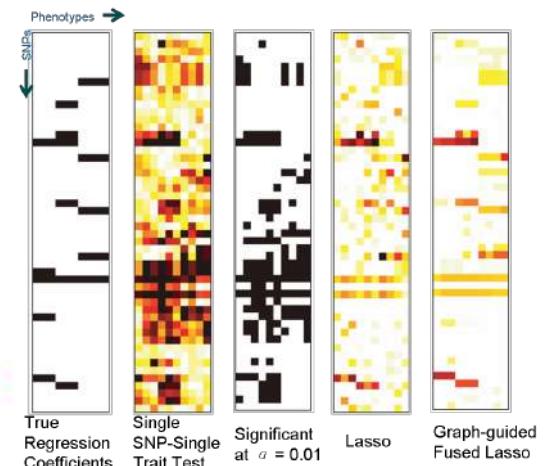
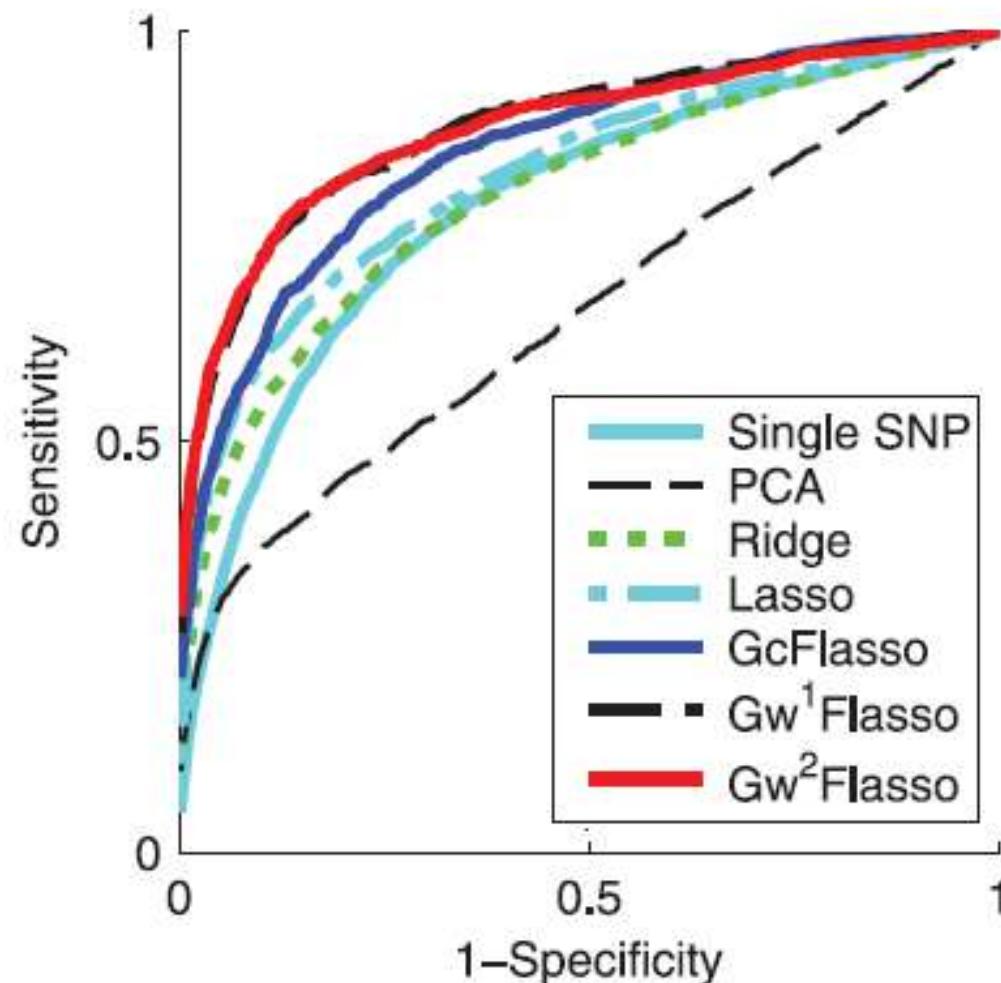
s. t. $\sum_k \sum_j |\beta_{jk}| \leq s_1$ and $\sum_{(m,l) \in E} f(r_{ml}) \sum_j |\beta_{jm} - \operatorname{sign}(r_{ml})\beta_{jl}| \leq s_2$

- Important properties:

- Convex optimization --- therefore global optimum exist
- Sparse --- therefore results are parsimonious and interpretable
- Theoretical Guarantee --- proof of consistency exist

- Many publicly available software packages for solving convex optimization problems can be used
- A new proximal gradient method can handle tens of thousands of SNPs and thousands of traits (later)

Simulation Results

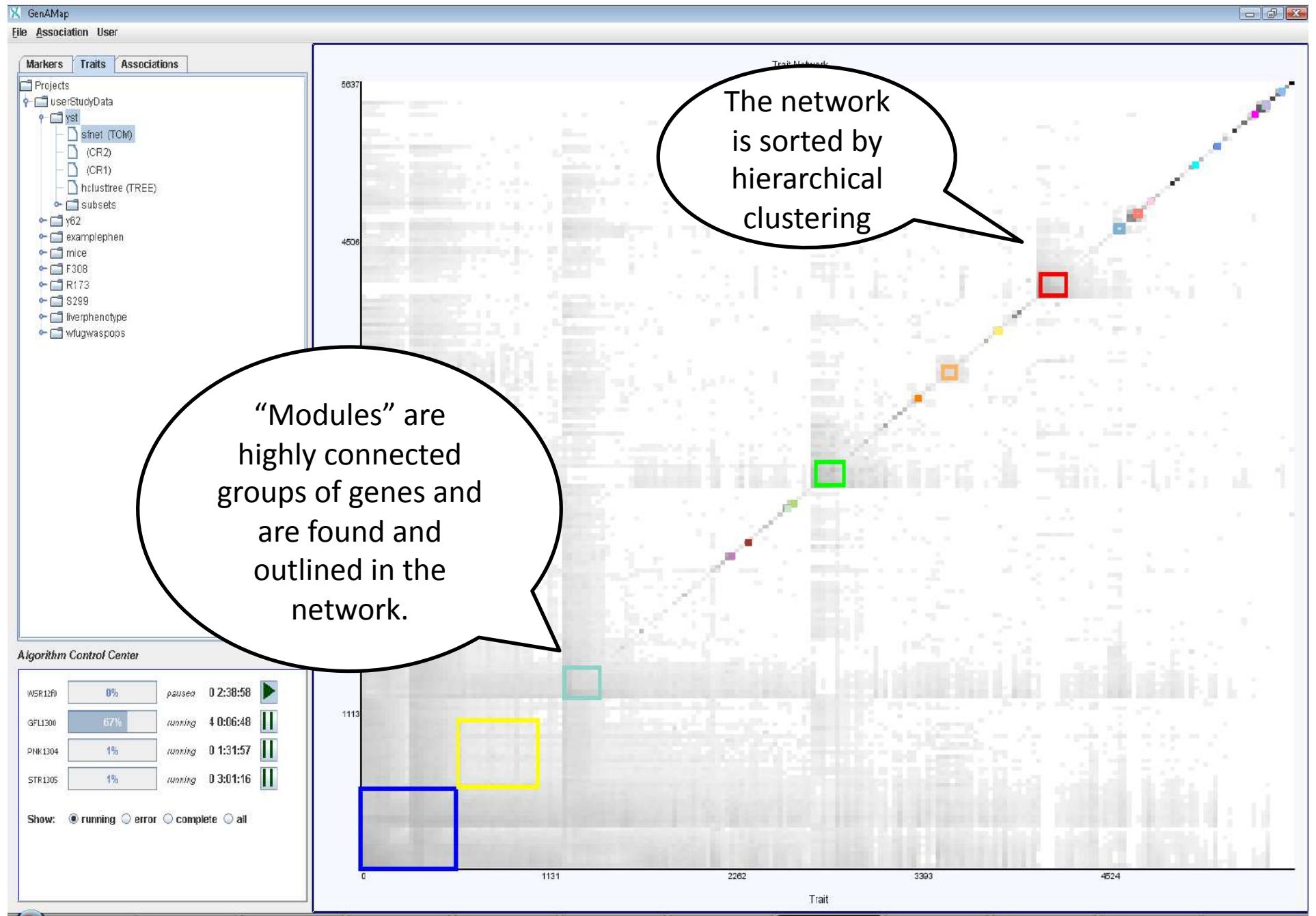


Yeast eQTL Dataset

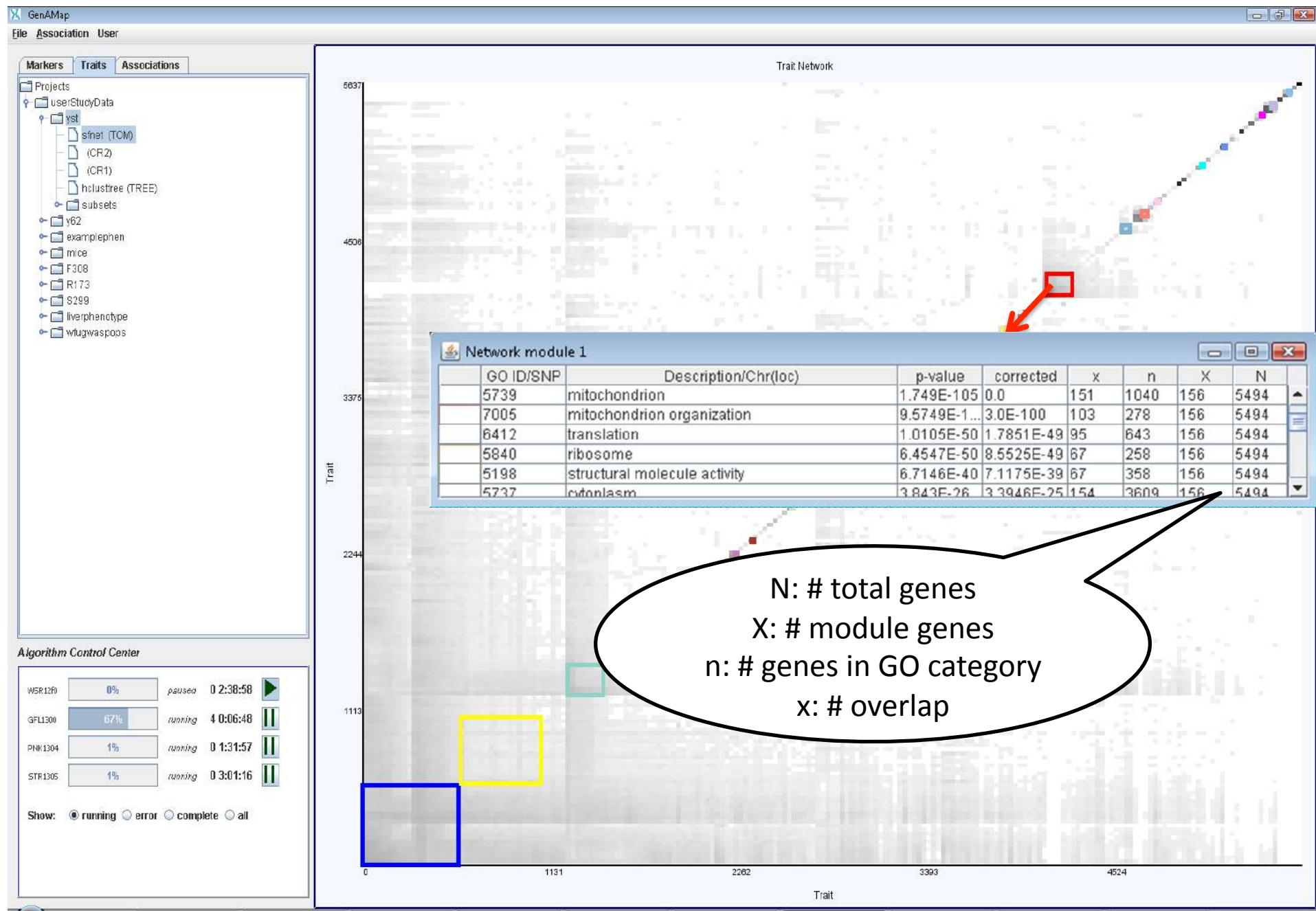
(Brem and Kruglyak, PNAS 2005)

- Genotype: 1260 SNPs
- Phenotype: expression measurements for 3684 genes
- 114 strains
- Goal: are there any SNPs that influence expression levels of genes?
 - Scale-free network construction and module finding
 - GFlasso analysis to find associations

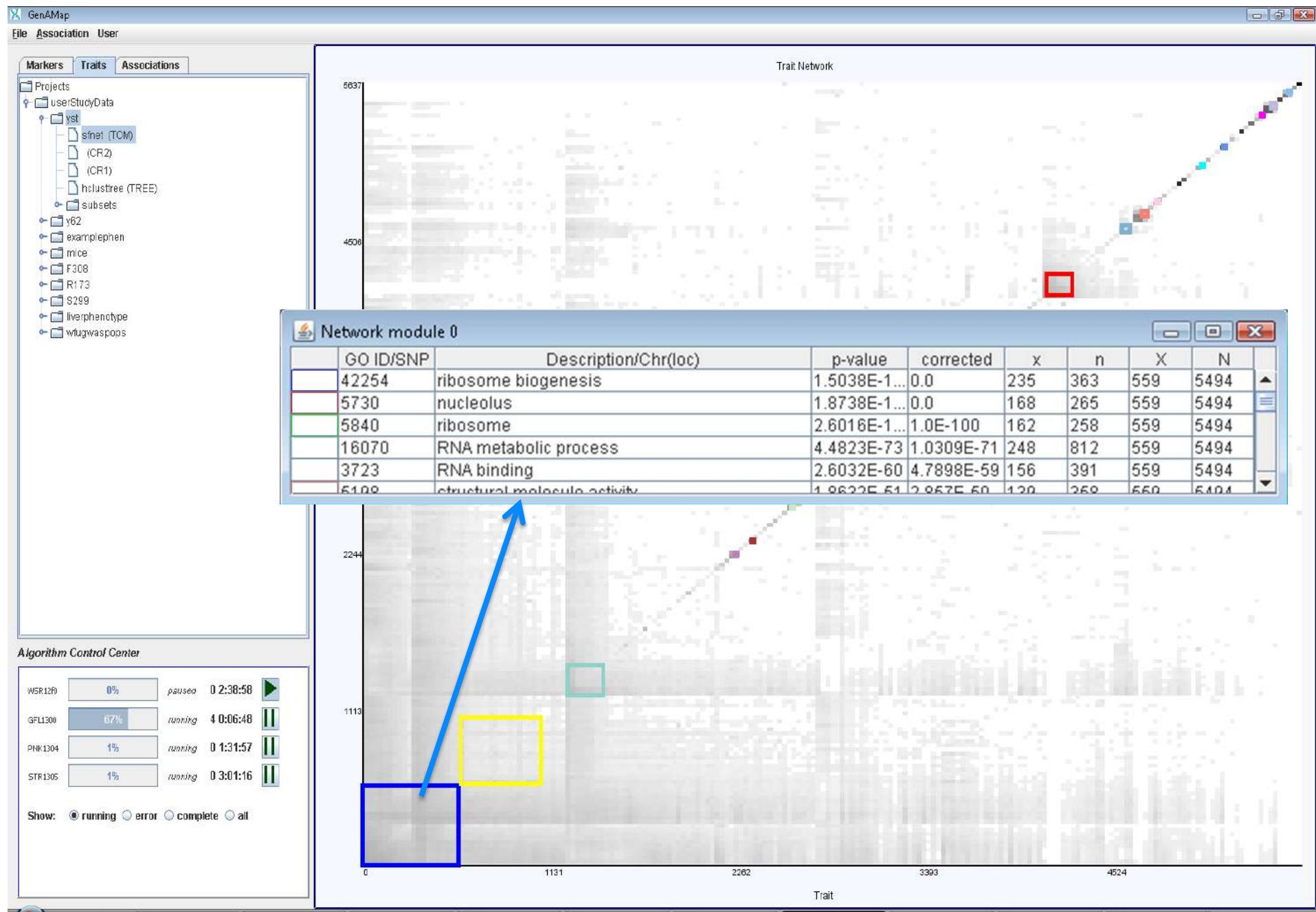
Gene-gene network exploration



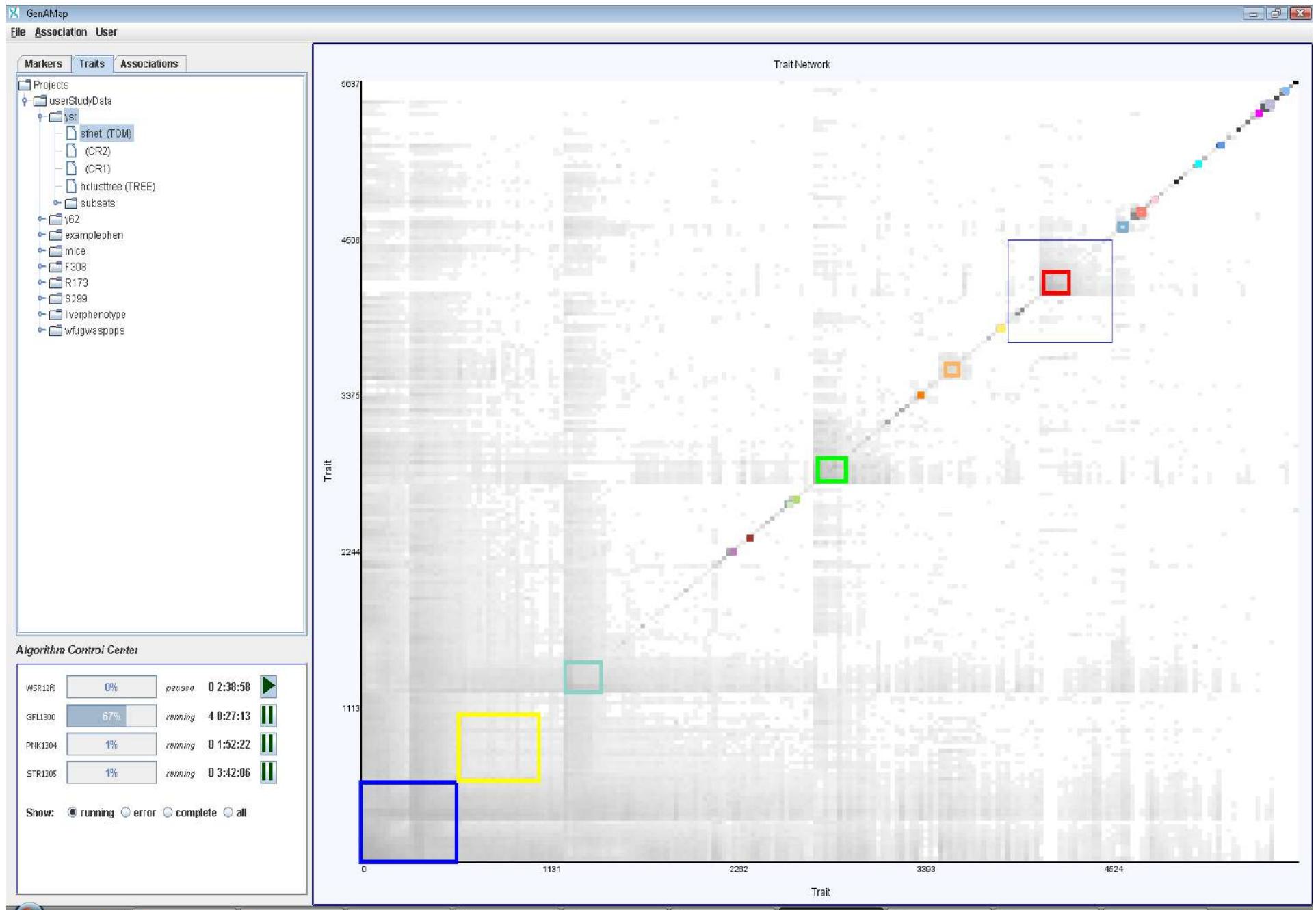
Gene-gene network exploration : Mitochondrion module

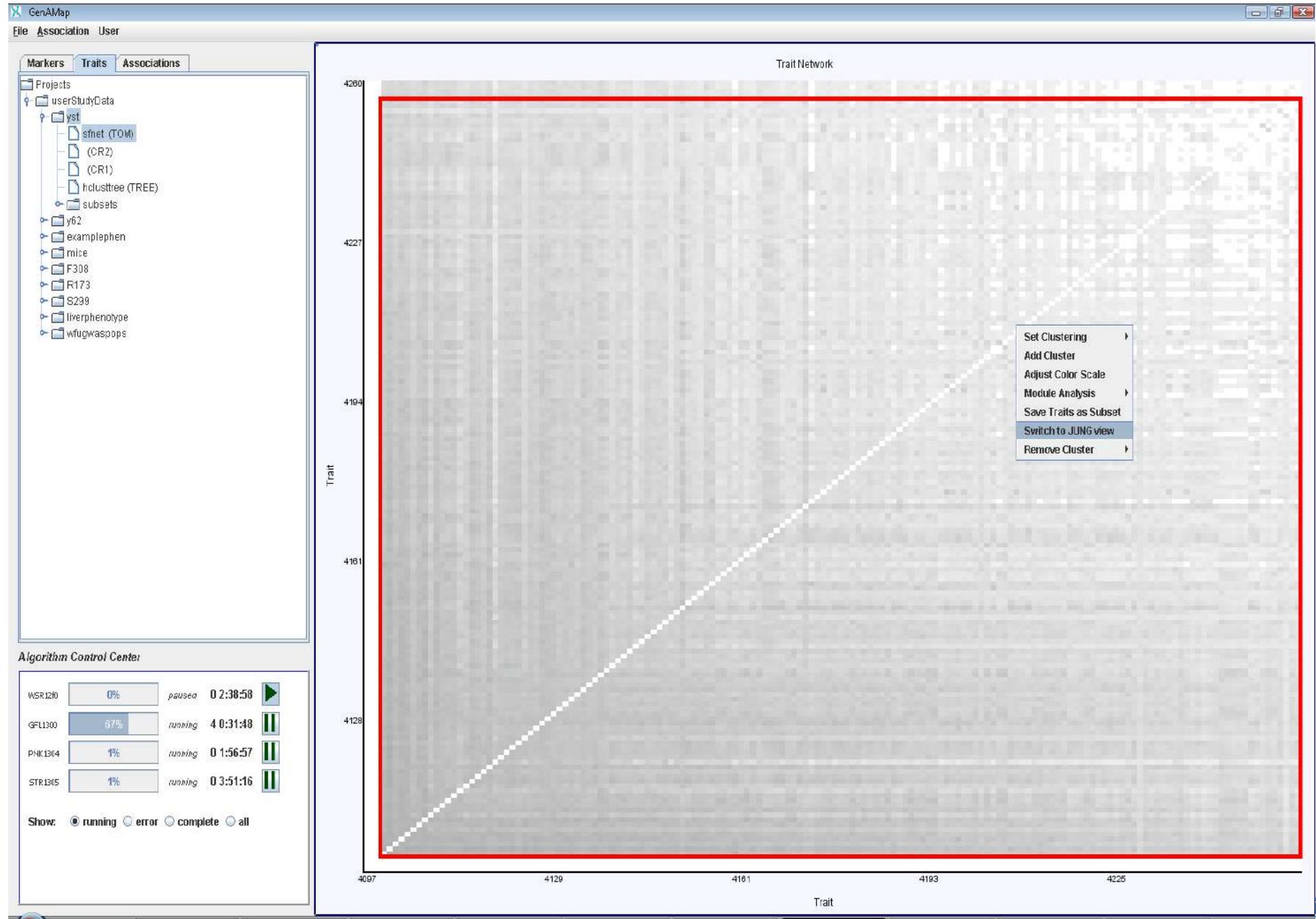


Gene-gene network exploration: ribosome biogenesis module

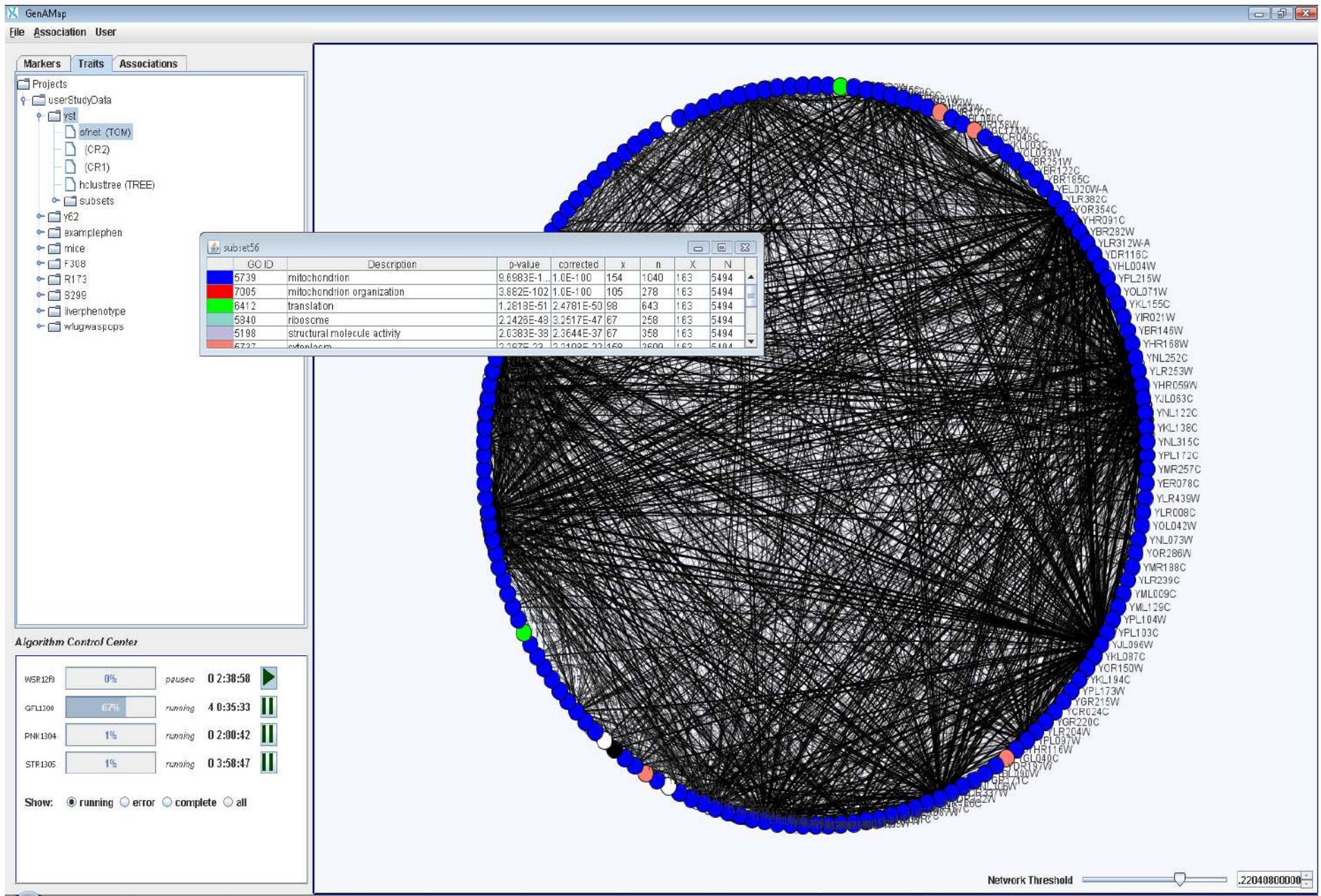


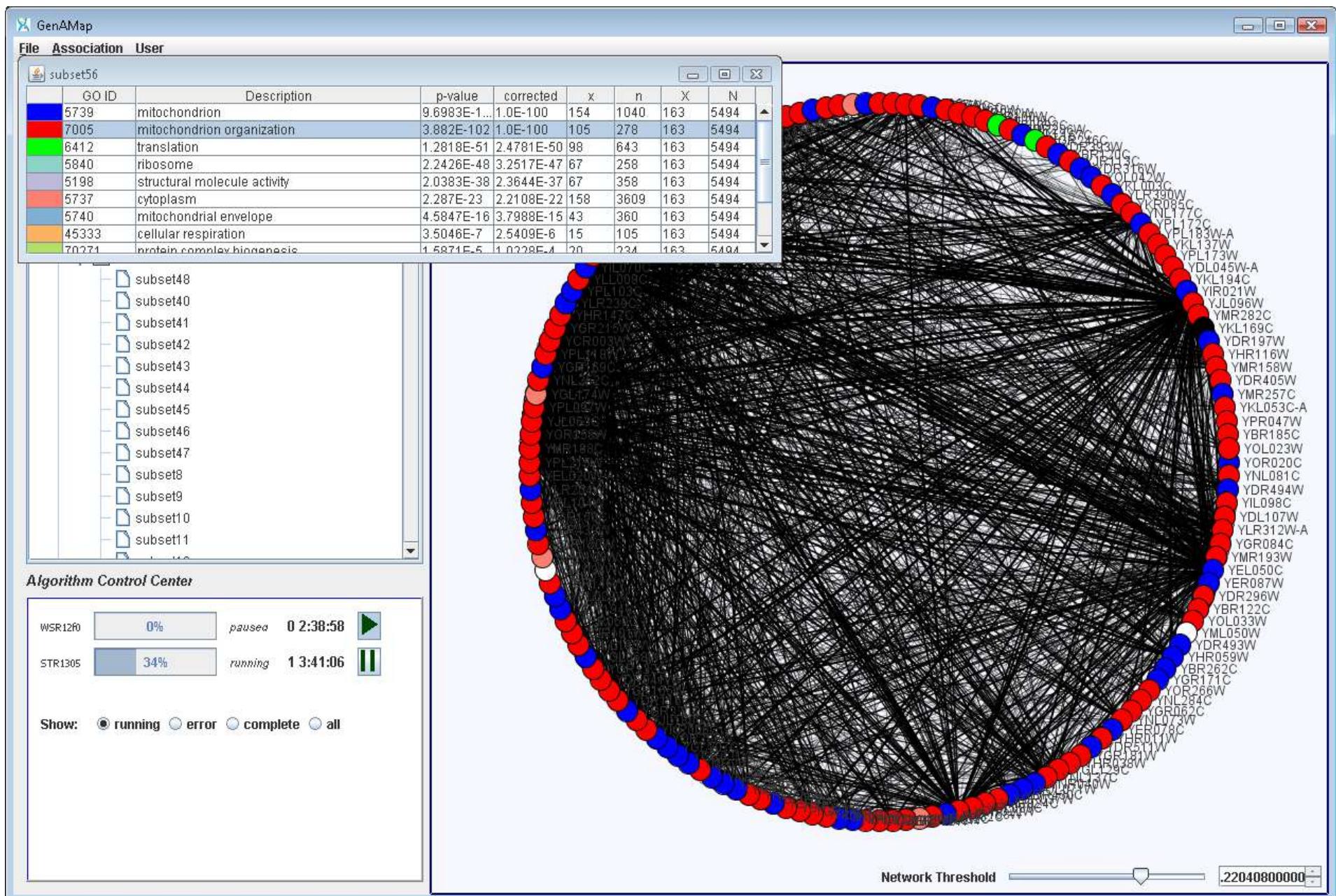
Gene-gene network exploration: zoom into mitochondrion module



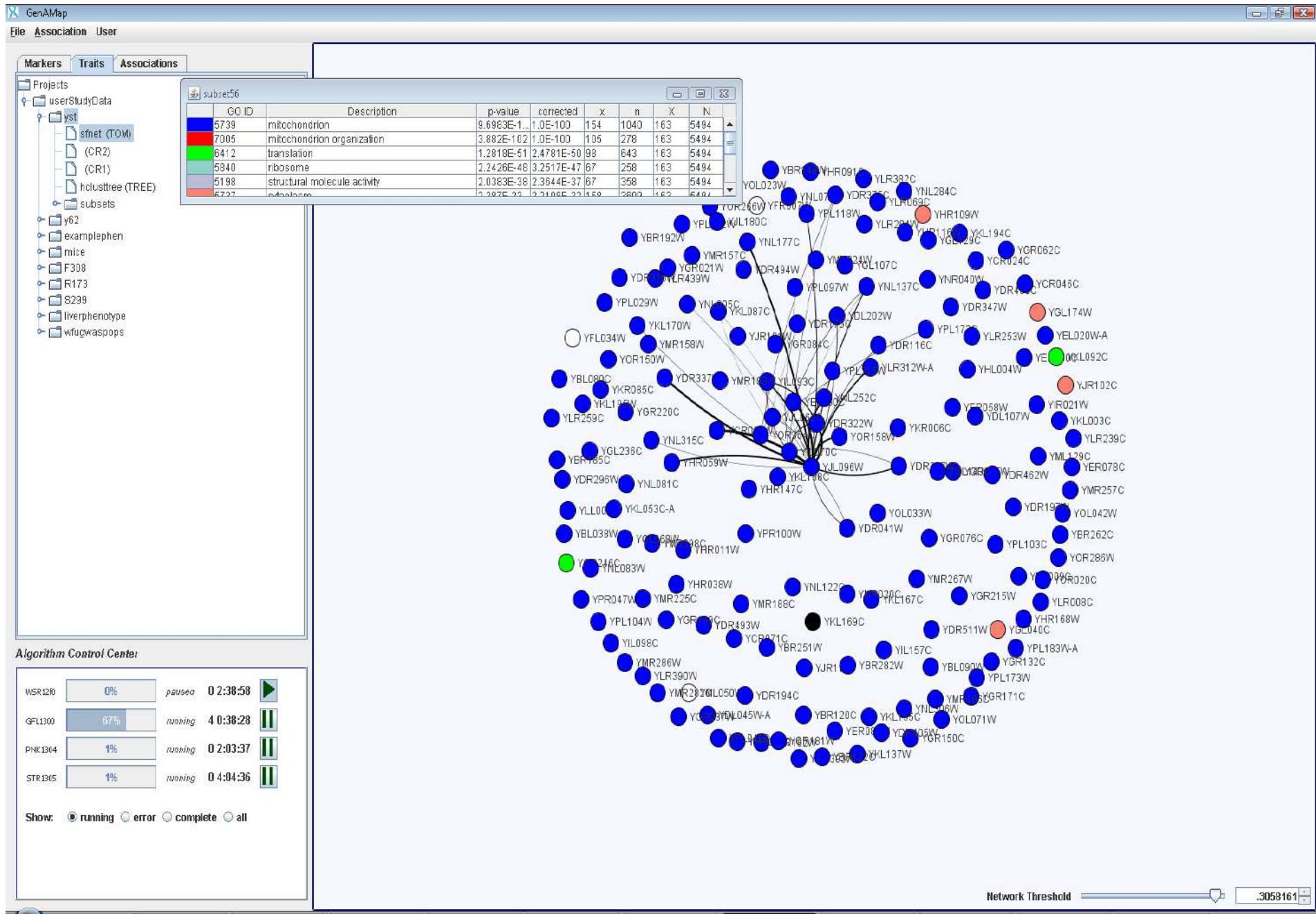


Gene-gene network exploration: GO analysis





Gene-gene network exploration: Identify hubs



Gene-gene network exploration: UniProt information

www.uniprot.org/uniprot/P40858.html

UniProt > UniProtKB Downloads · Contact · Documentation/Help

Search Blast * Align Retrieve ID Mapping *

Search in Query Protein Knowledgebase (UniProtKB) Search Clear Advanced Search »

P40858 (RN49_YEAST) ★ Reviewed, UniProtKB/Swiss-Prot

Last modified May 3, 2011. Version 87. History...

Clusters with 100%, 90%, 50% identity | Documents (4) | Third-party data

text xml rdf/xml gff fasta

Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents · Customize order

Names and origin

Protein names	<i>Recommended name:</i> 54S ribosomal protein L49, mitochondrial <i>Alternative name(s):</i> YmL49
Gene names	Name: MRPL49 Ordered Locus Names: YJL096W ORF Names: J0904
Organism	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) [Complete proteome]
Taxonomic identifier	559292 [NCBI]
Taxonomic lineage	Eukaryota > Fungi > Dikarya > Ascomycota > Saccharomycotina > Saccharomycetes > Saccharomycetales > Saccharomycetaceae > Saccharomyces

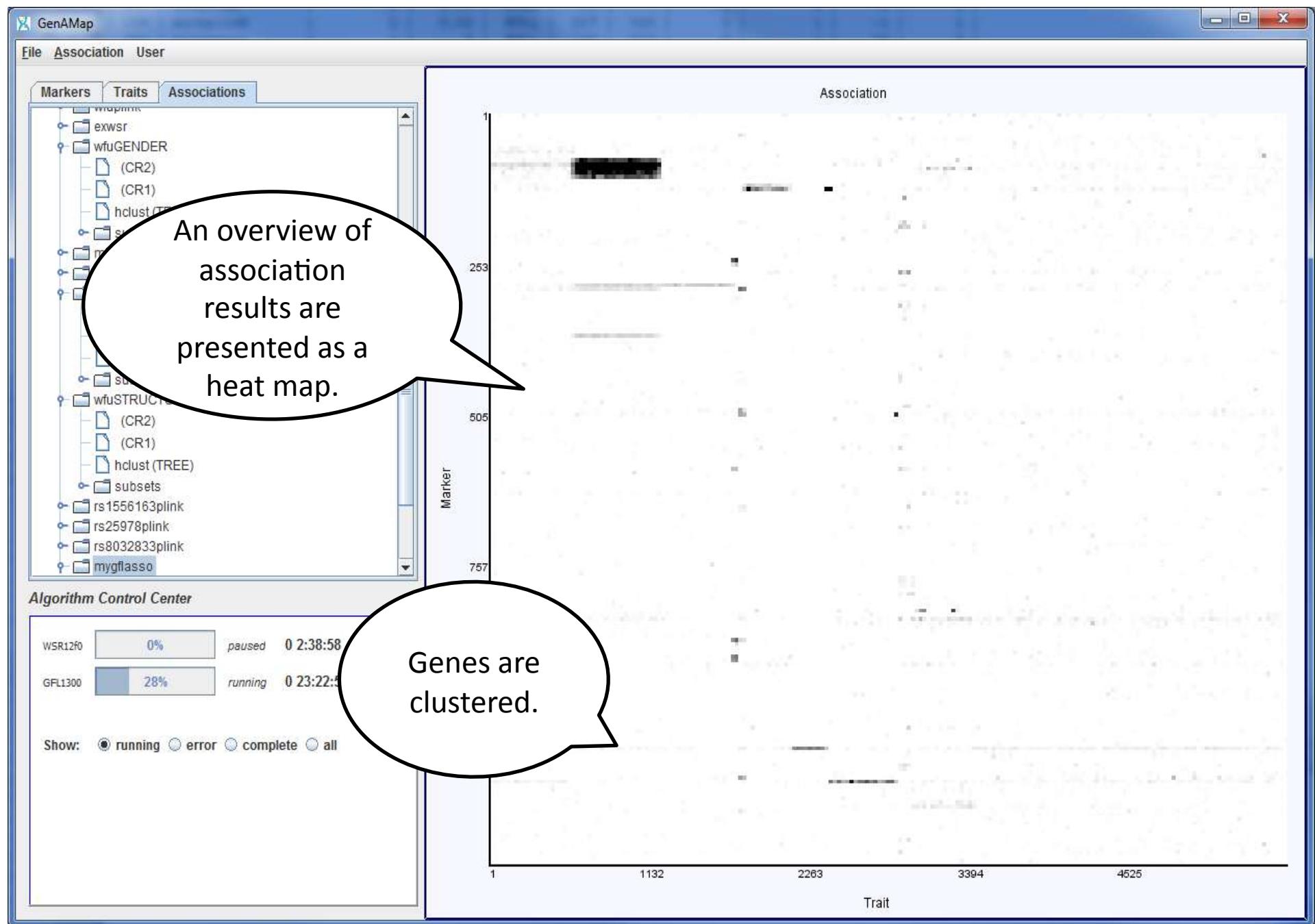
Protein attributes

Sequence length	161 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

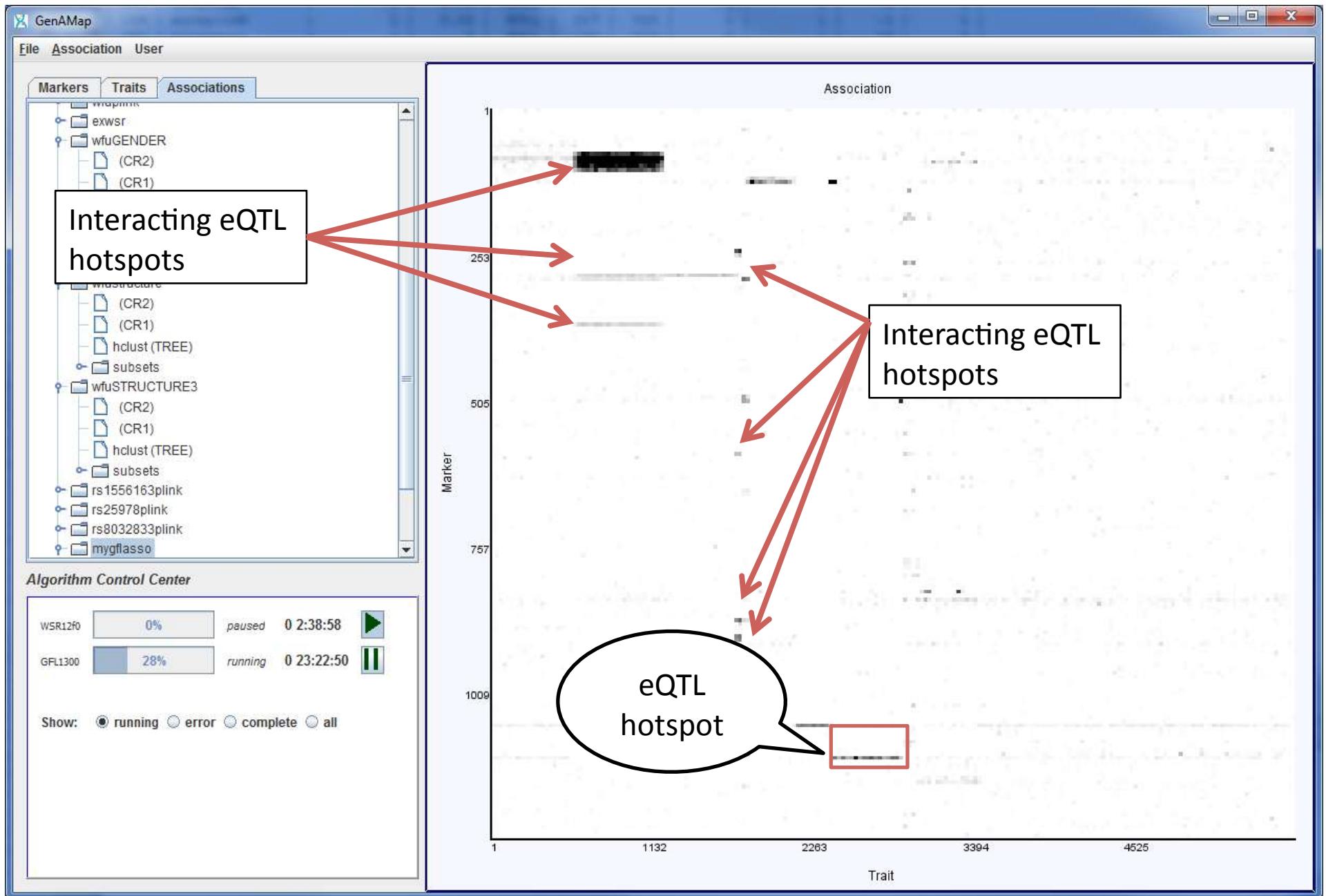
General annotation (Comments)

Subunit structure	Component of the mitochondrial large ribosomal subunit. Mature mitochondrial ribosomes consist of a small (37S) and a large (54S) subunit. The 37S subunit contains at least 33 different proteins and 1 molecule of RNA (15S). The 54S subunit contains at least 45 different proteins and 1 molecule of RNA (21S). Ref.5
Subcellular location	Mitochondrion Ref.8 Ref.10

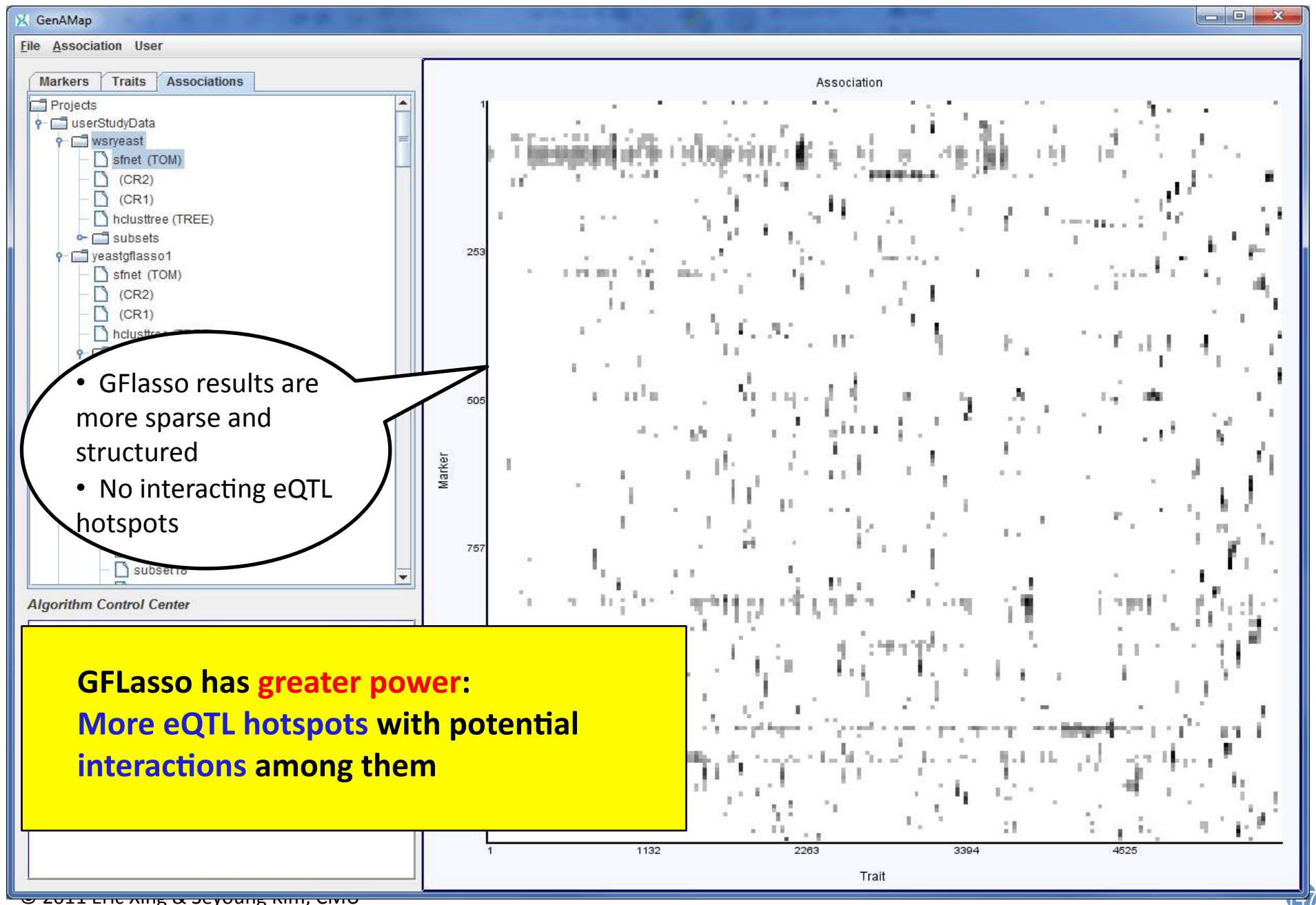
GFlasso association results



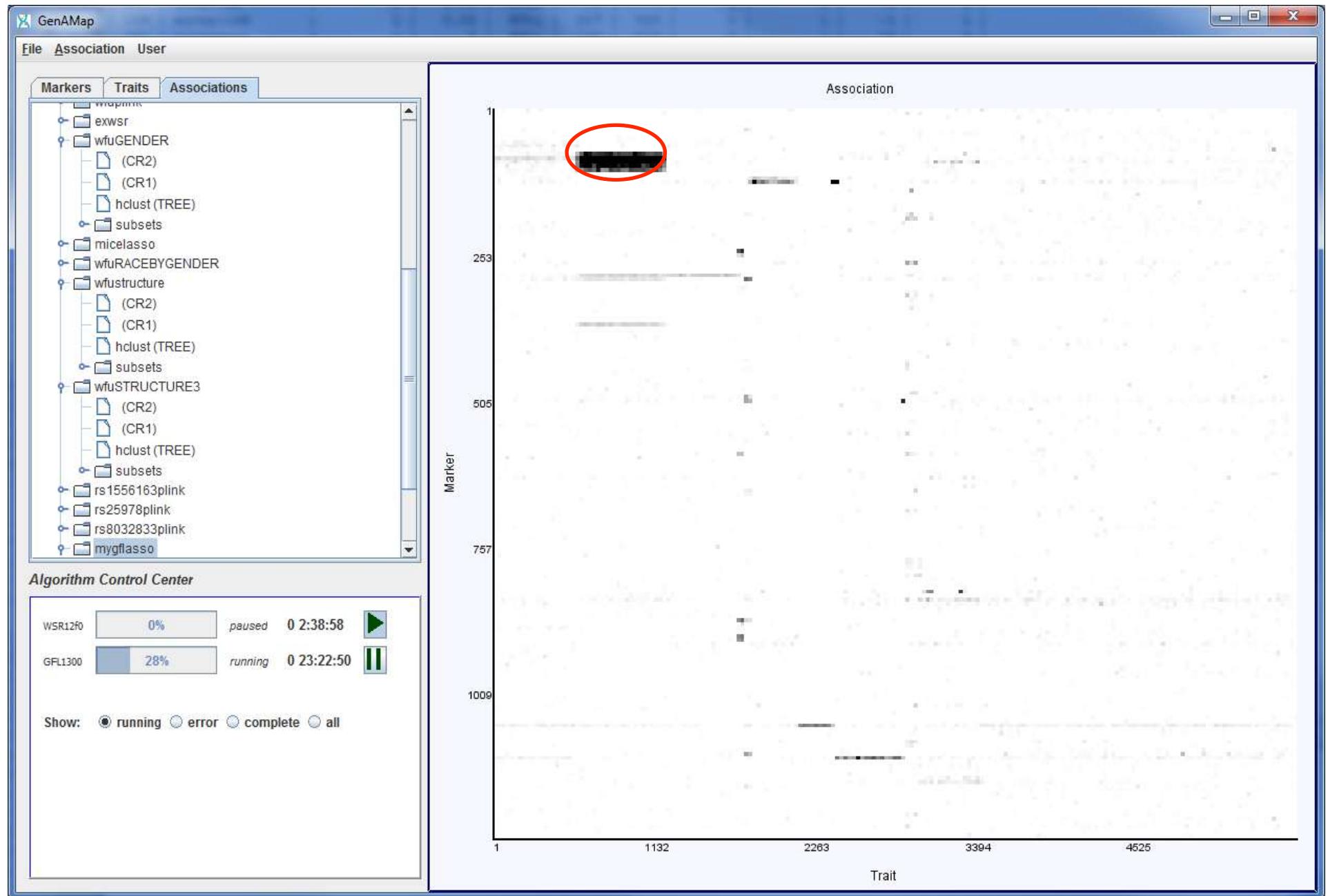
GFlasso results



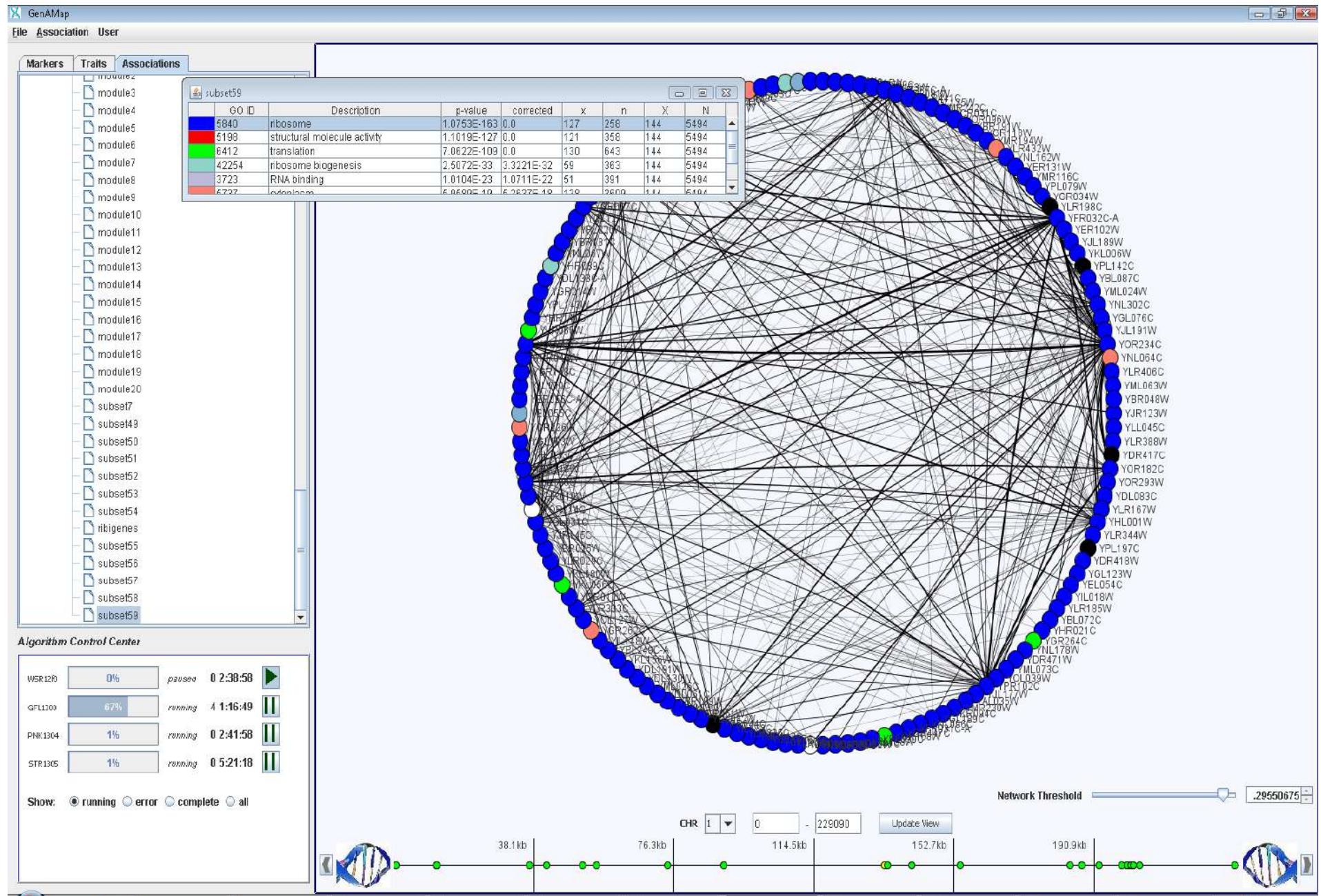
Comparing results with wilcoxon-sum-rank test (single SNP/gene analysis)



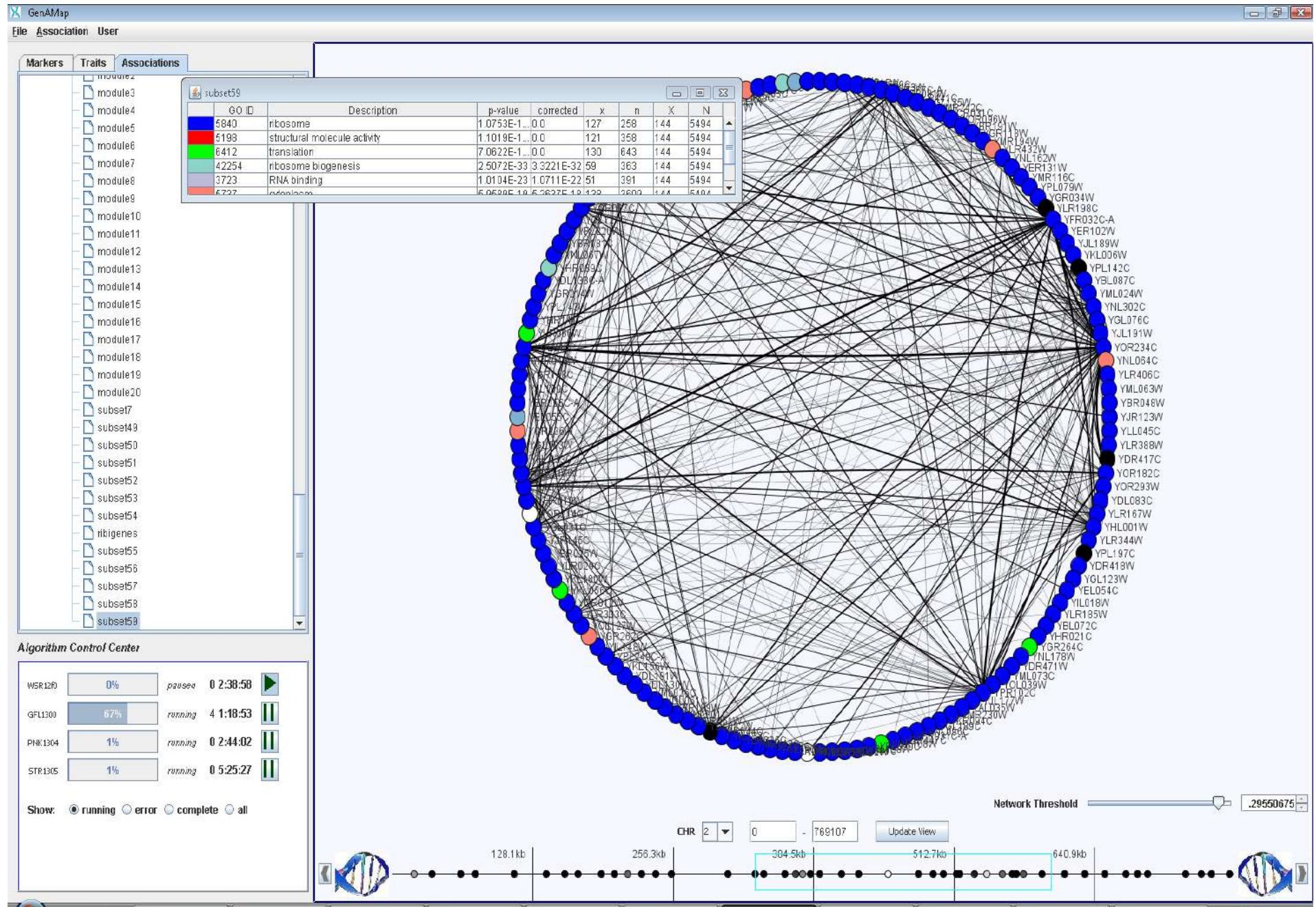
Investigate an eQTL hotspot associated with a single locus



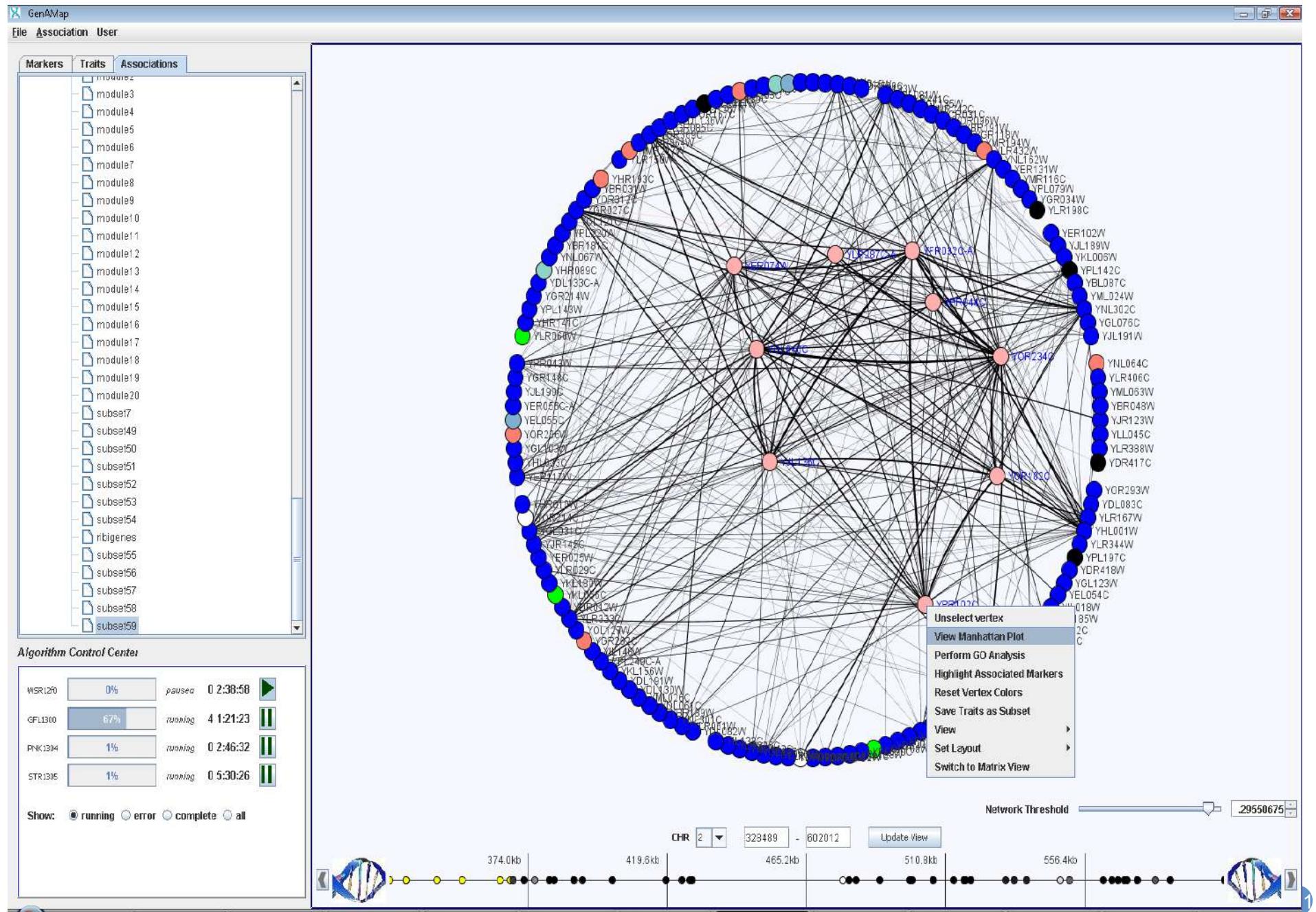
GO analysis reveals hotspot is enriched for ribosome genes



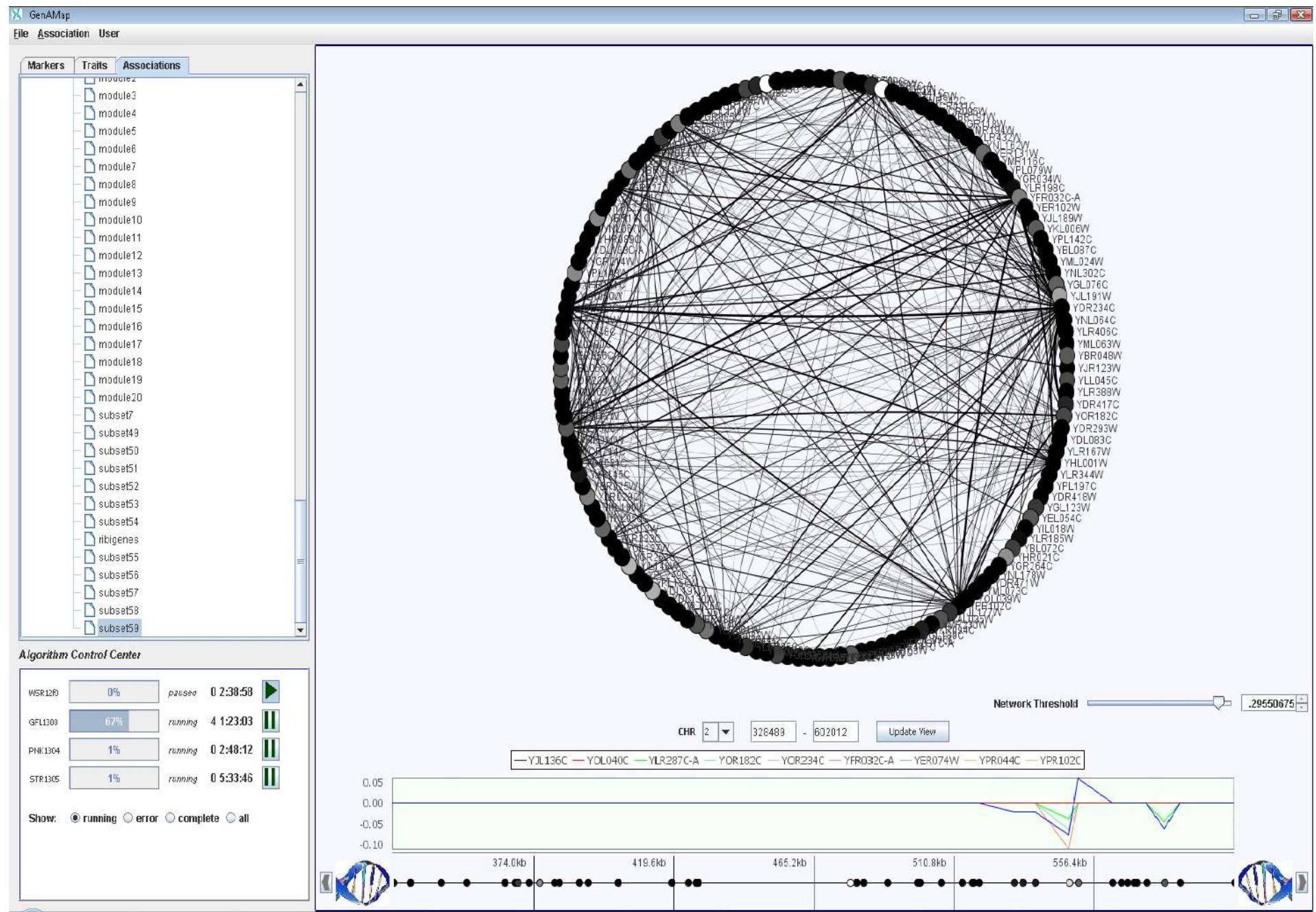
Associations on chromosome 2

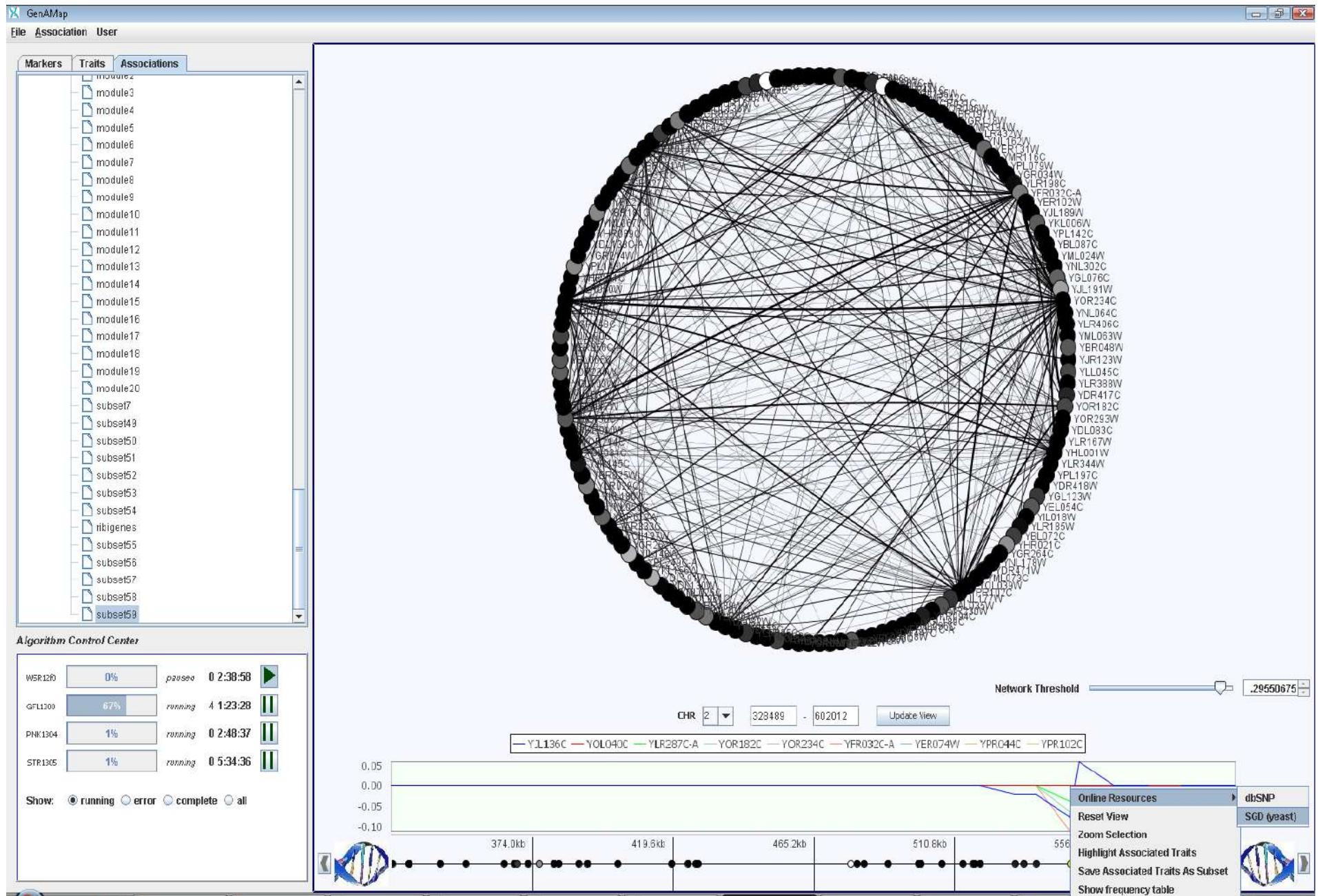


Look at association strengths to highest connected genes



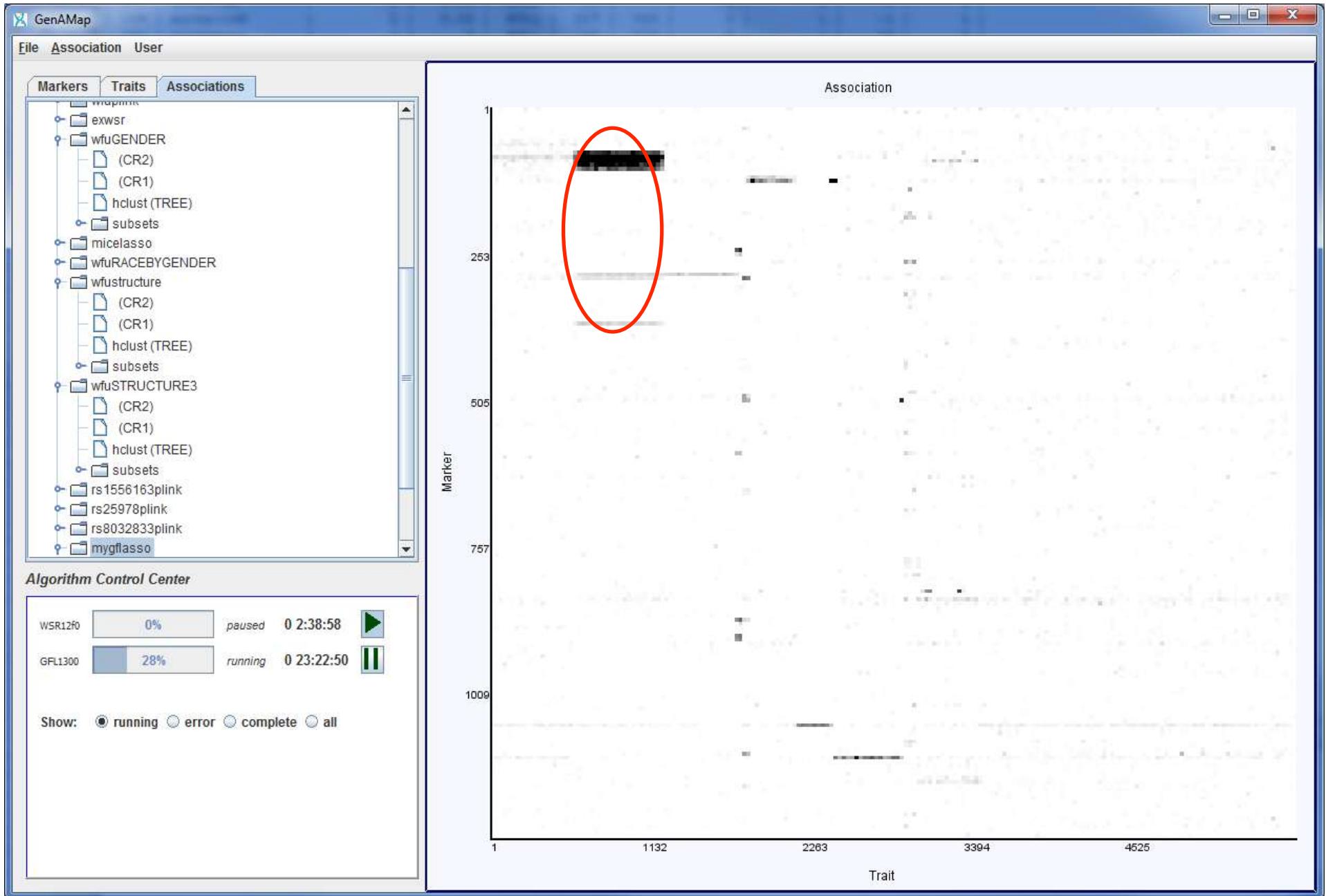
Only a few genes are associated with the hotspot around this SNP (not black), but the highest connected genes are associated with this hotspot.



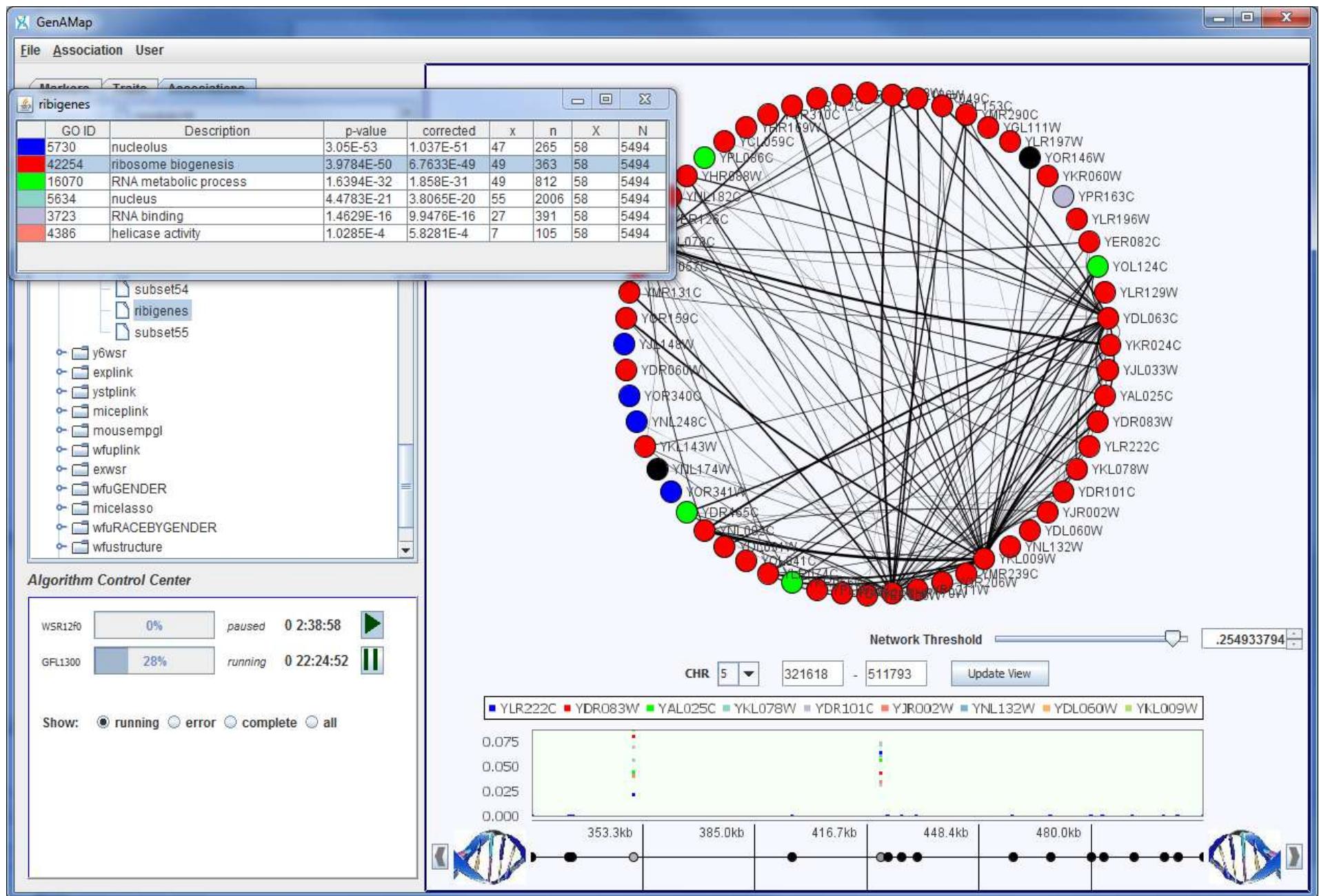


SNP in RPB5 – causal or linked?

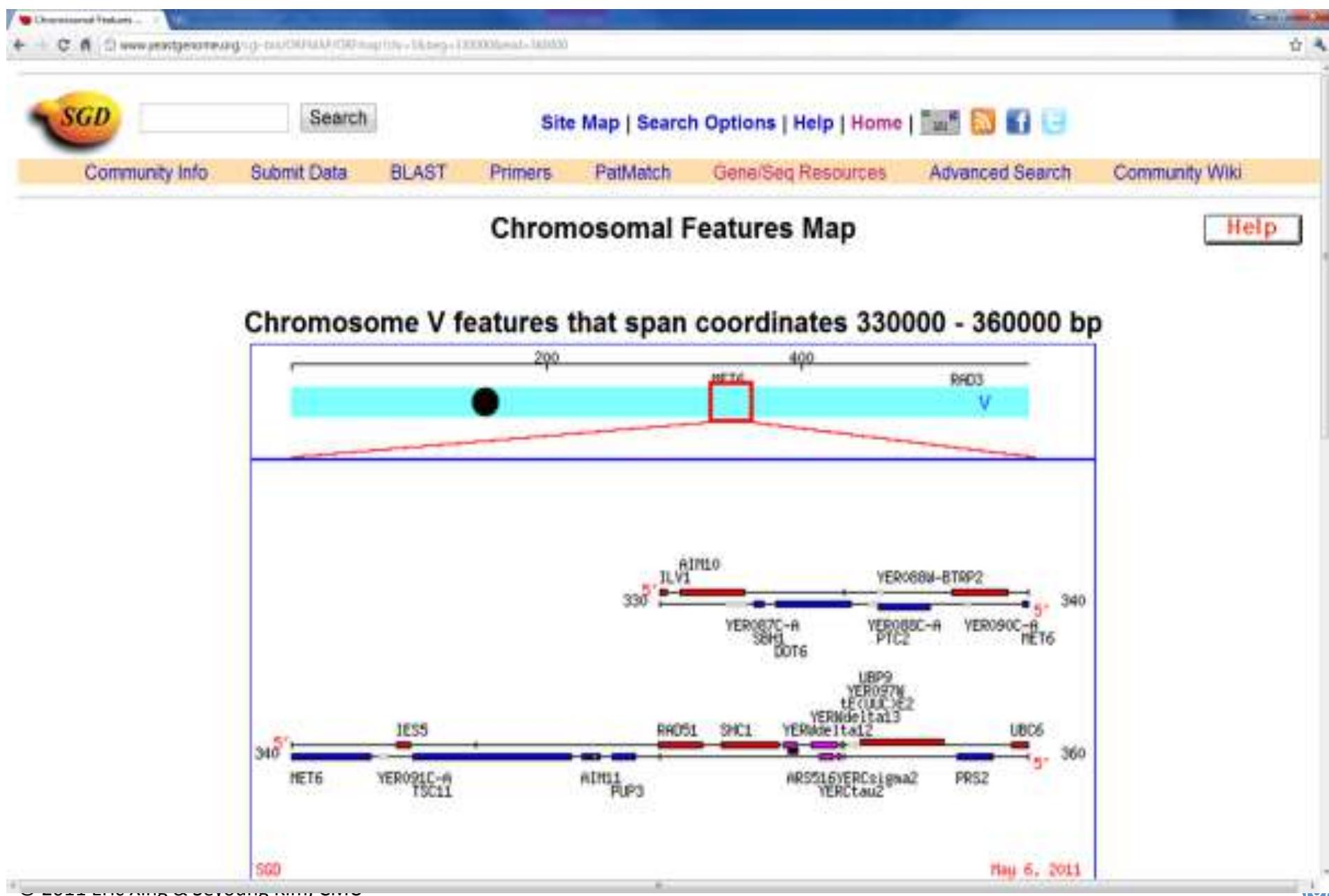
Investigate interacting eQTL hotspots



Genes are enriched for ribosome biogenesis and are associated with these SNPs on chromosome 5.



DOT6 (also known as *PBF2*) is located in *cis* with these SNPs



www.yeastgenome.org/cgi-bin/locus.fpl?locus=YER088C

SGD Search Site Map | Search Options | Help | Home | RSS | Facebook | Twitter

Community Info Submit Data BLAST Primers PatMatch Gene/Seq Resources Advanced Search

DOT6/YER088C Summary

Summary Locus History Literature Gene Ontology Phenotype Interactions Expression Protein Wiki

DOT6 BASIC INFORMATION

Standard Name **DOT6** ¹

Systematic Name **YER088C**

Alias **PBF2** ²

Feature Type ORF, Verified

Description Protein involved in rRNA and ribosome biogenesis; binds polymerase A and C motif; subunit of the RPD3L histone deacetylase complex; similar to Tod6p; has chromatin specific SANT domain; involved in telomeric gene silencing and filamentation ([1](#), [2](#), [3](#), [4](#))

Name Description Disruptor Of Telomeric silencing ¹

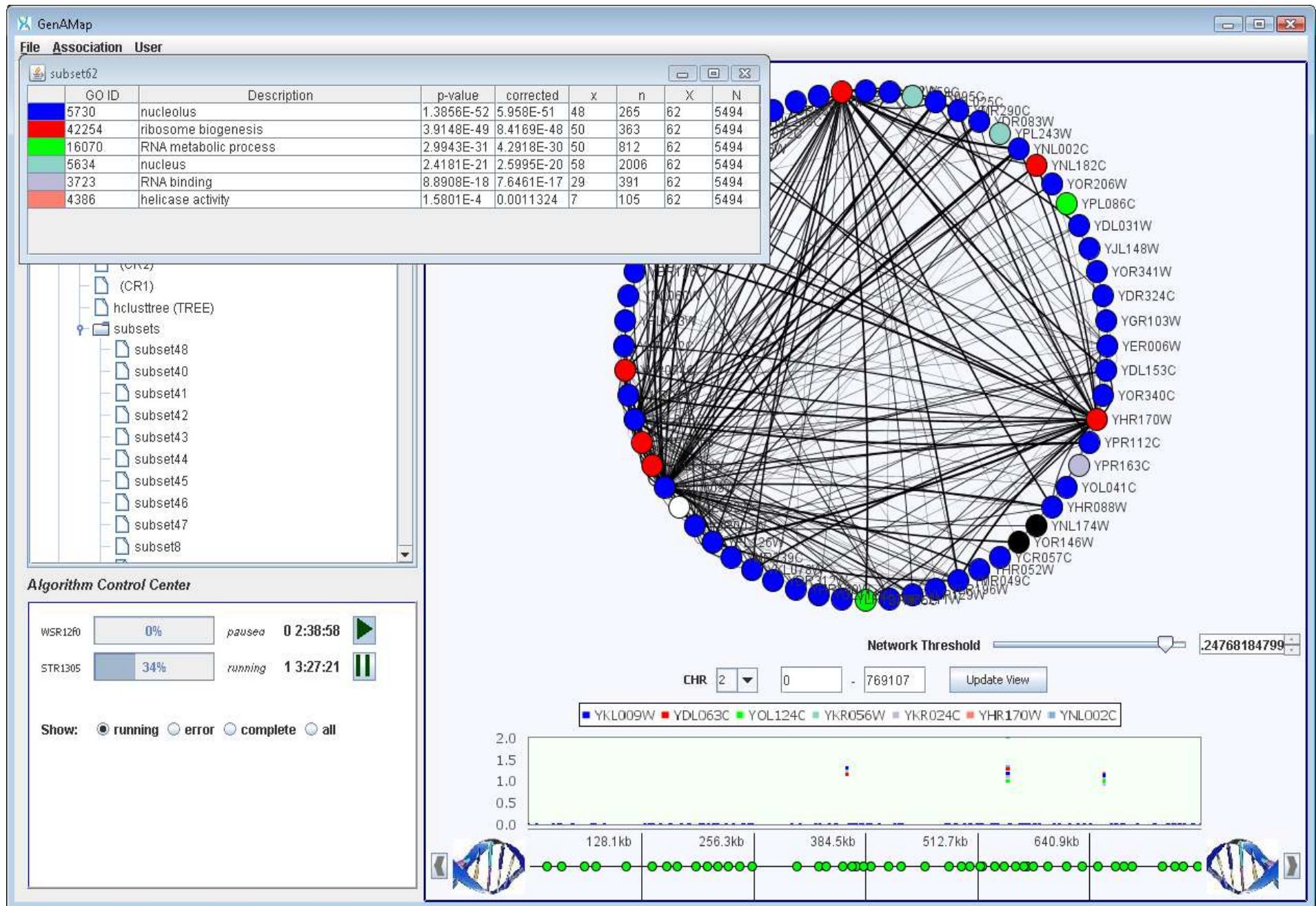
GO Annotations All DOT6 GO evidence and references

Click or SGD ORF n

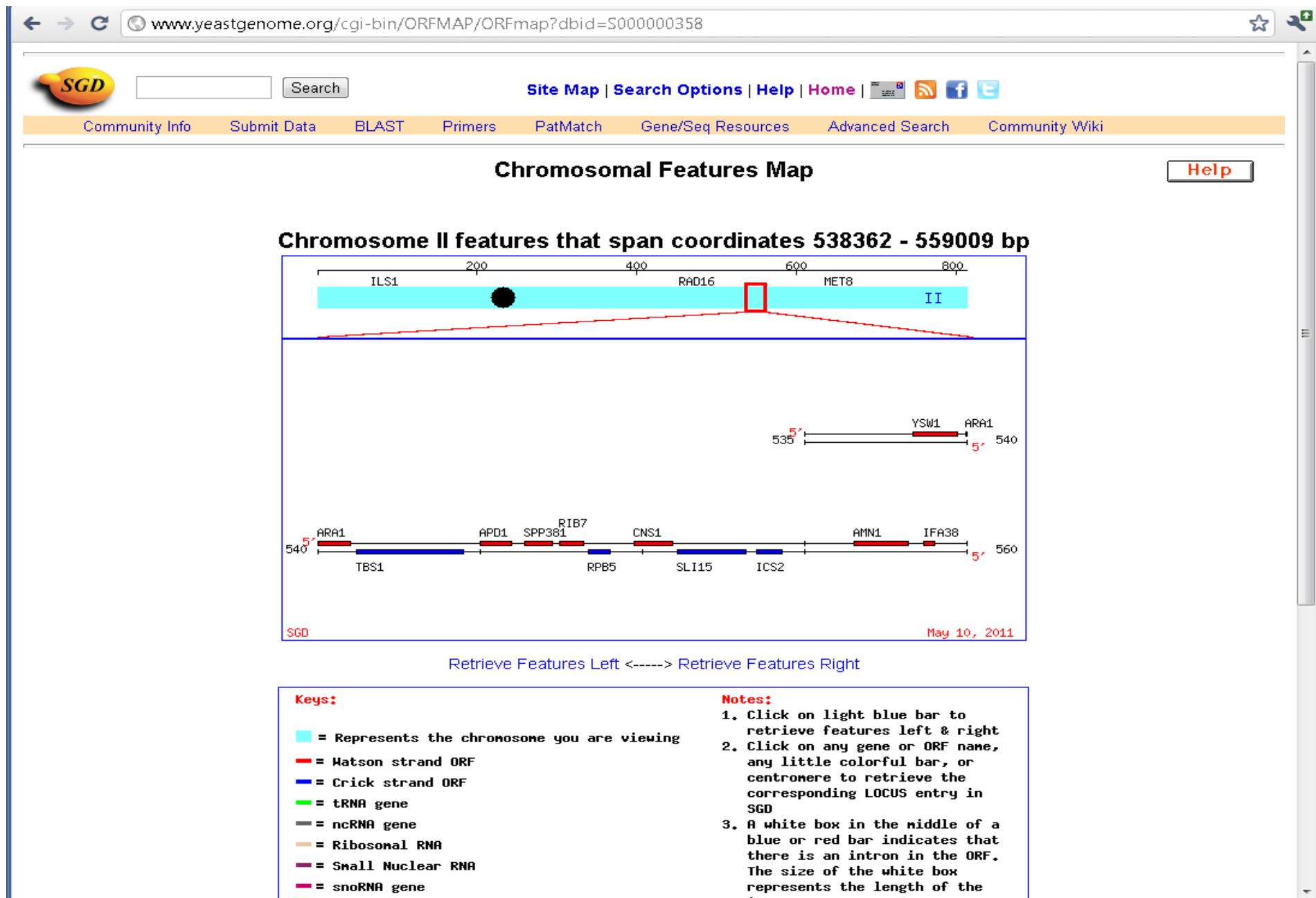
332000 to
AIM10
5' -
YER087C-A SBH1
V DOT6

- Literature Literature Guide
- Retrieve Sequence Genomic DNA

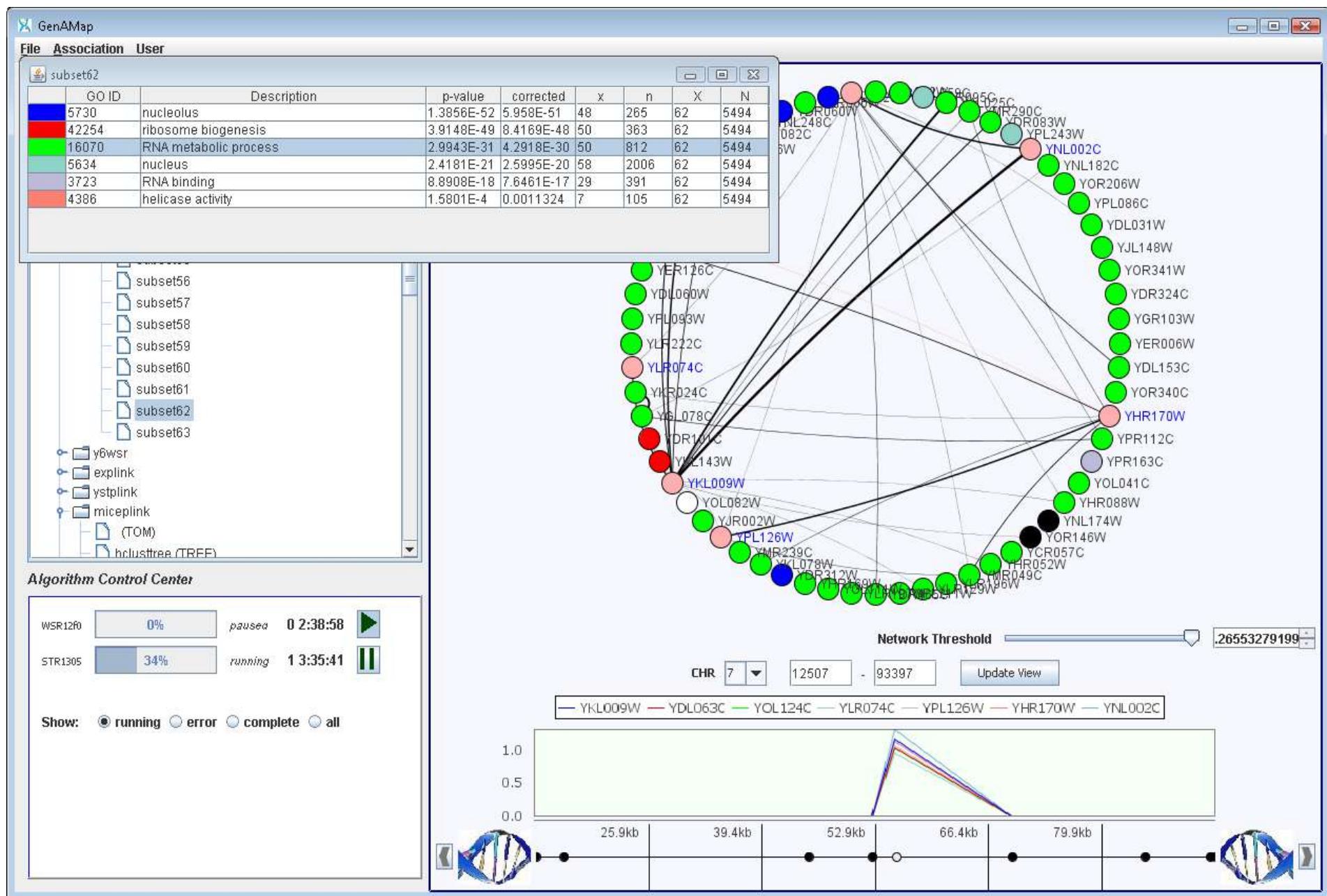
Genes are also associated with these SNPs on chromosome 2.



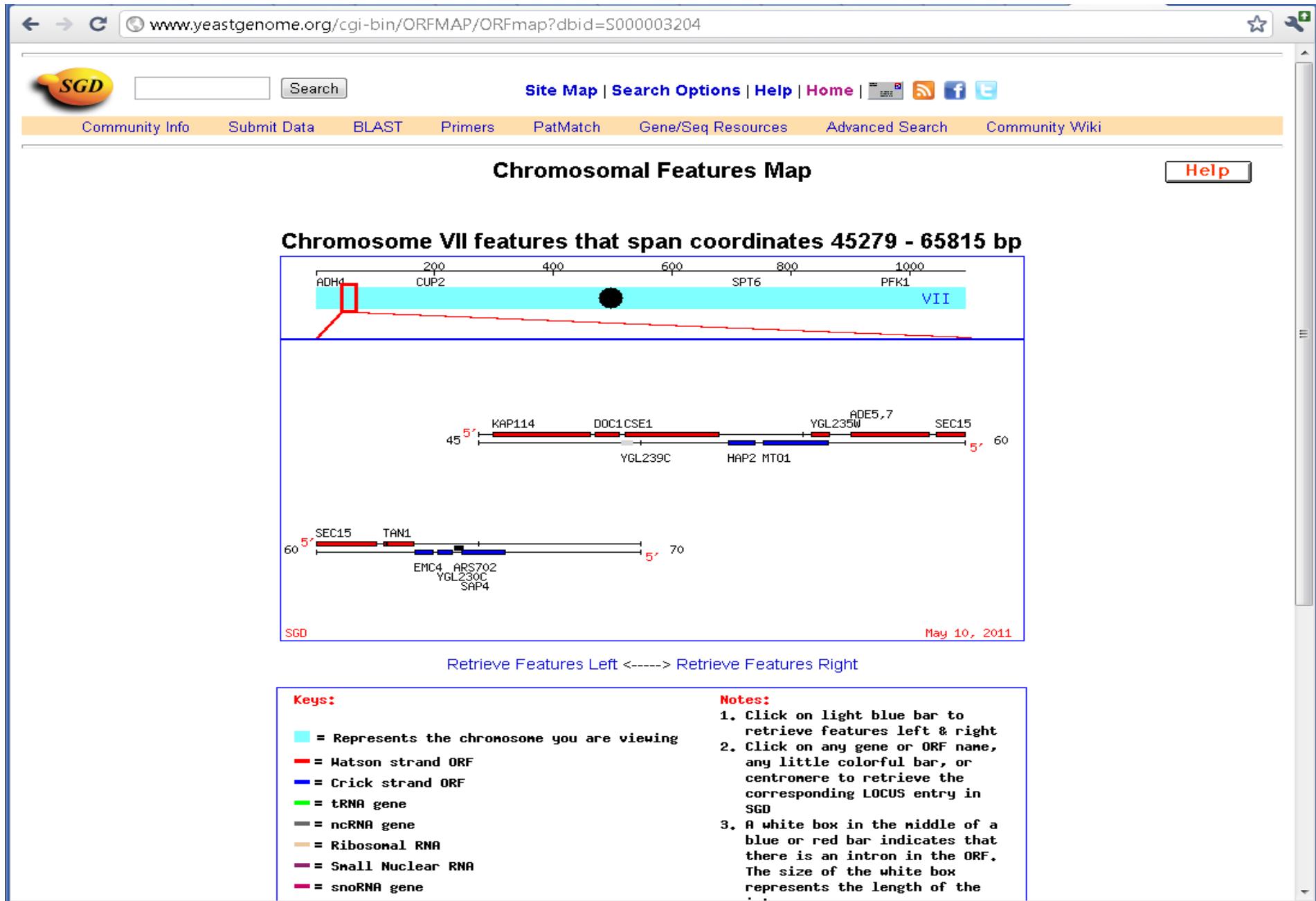
Candidate regulators on chromosome 2.



Genes are also associated with these SNPs on chromosome 7.



Candidate regulators on chromosome 7



Could KAP114 also import Ribi genes?

www.yeastgenome.org/cgi-bin/locus.fpl?locus=Kap114

SGD **Search**

Site Map | Search Options | Help | Home |    

Community Info **Submit Data** **BLAST** **Primers** **PatMatch** **Gene/Seq Resources** **Advanced Search**

KAP114/YGL241W Summary

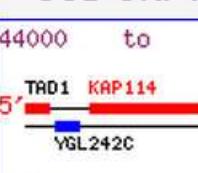
Summary **Locus History** **Literature** **Gene Ontology** **Phenotype** **Interactions** **Expression** **Protein** **Wiki**

KAP114 BASIC INFORMATION

Standard Name	KAP114 ¹
Systematic Name	YGL241W
Feature Type	ORF, Verified
Description	Karyopherin, responsible for nuclear import of Spt15p, Sua7p, histones H2A and H2B, and Nap1p; amino terminus shows similarity to those of other importins, particularly Cse1p; localization is primarily nuclear (2, 3, 4) 
Name Description	KAryoPherin ¹
GO Annotations	All KAP114 GO evidence and references View Computational GO annotations for KAP114

KAP11

Click or
SGD ORF n

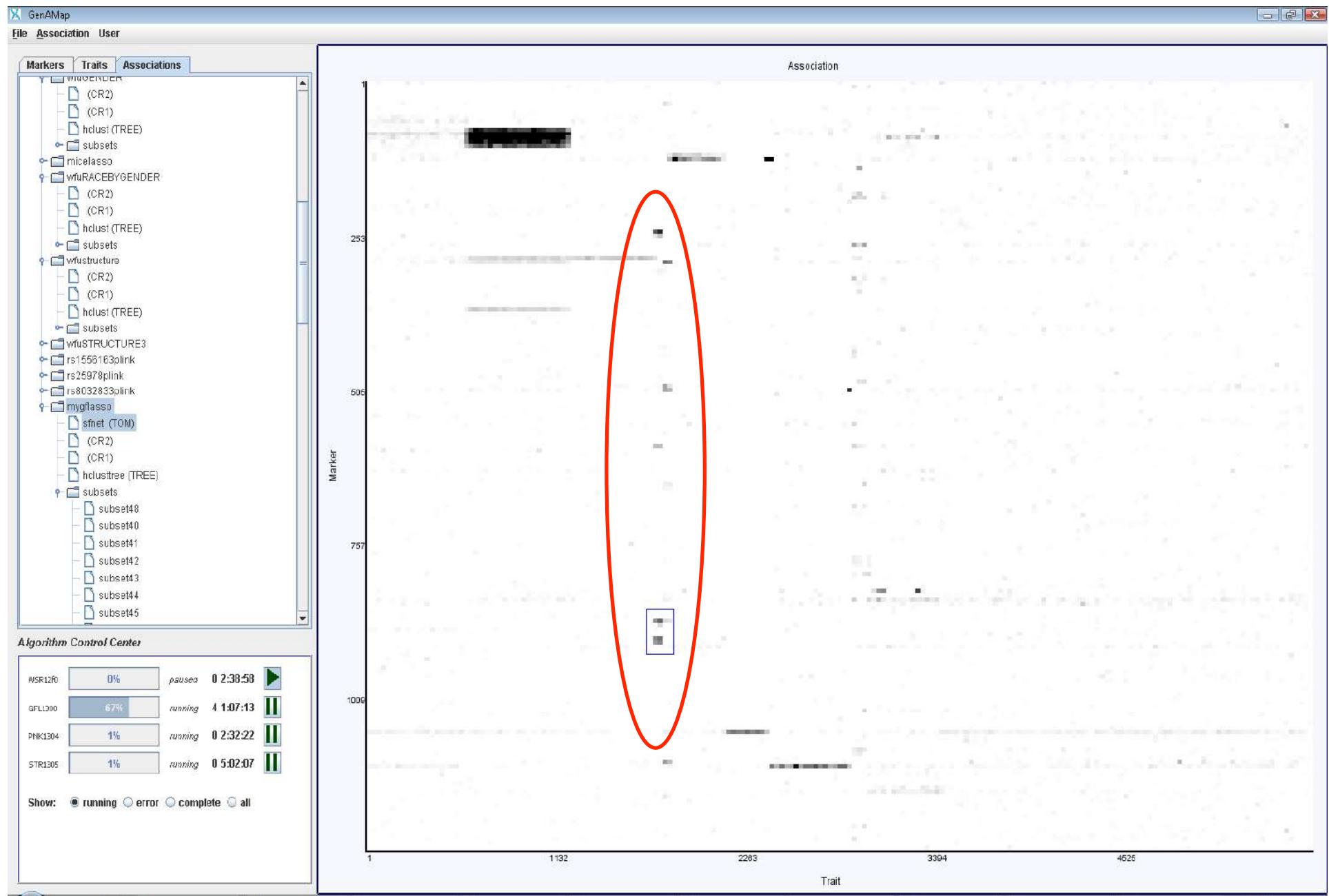


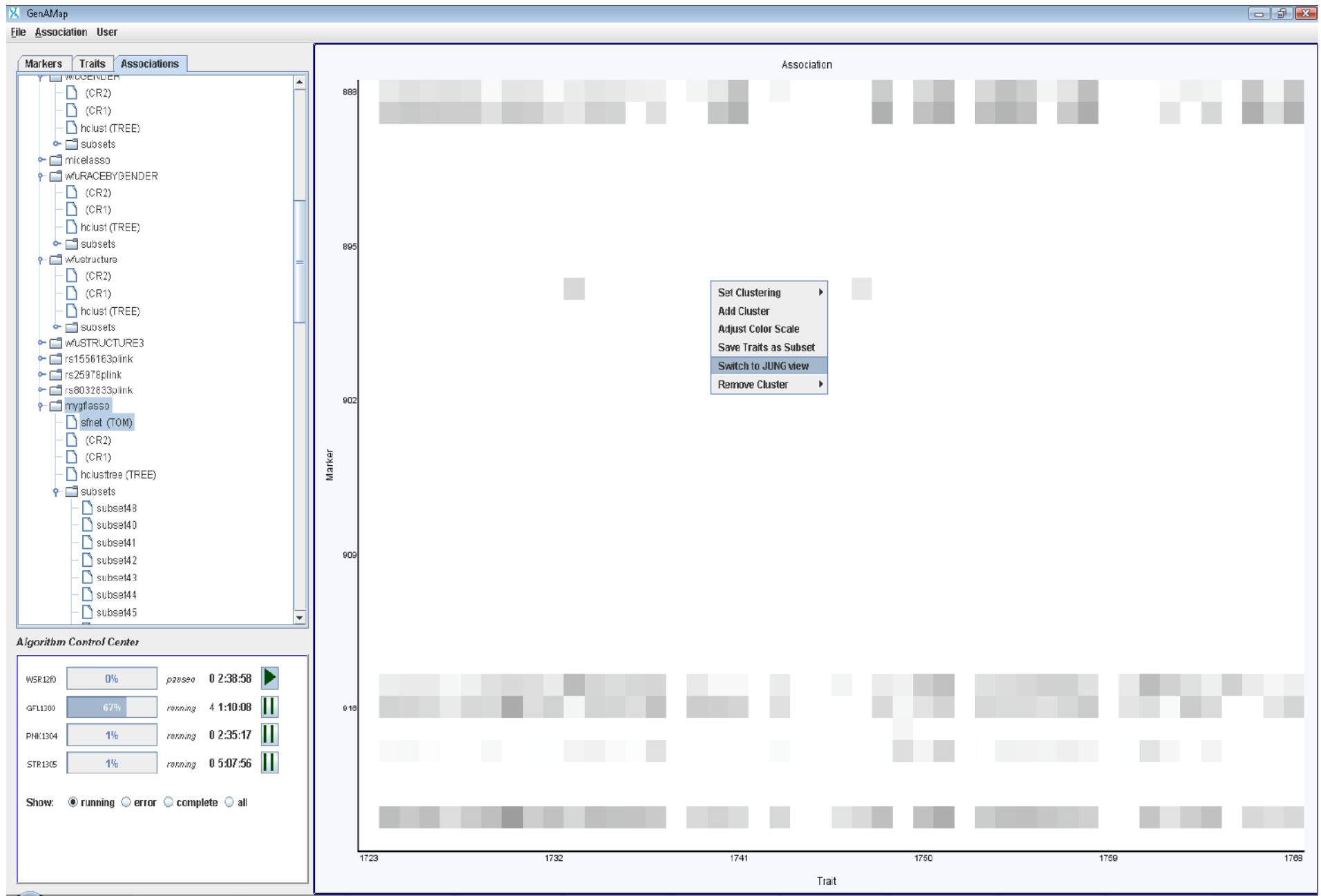
44000 to
TAD1 KAP114
5' 
YGL242C

VII

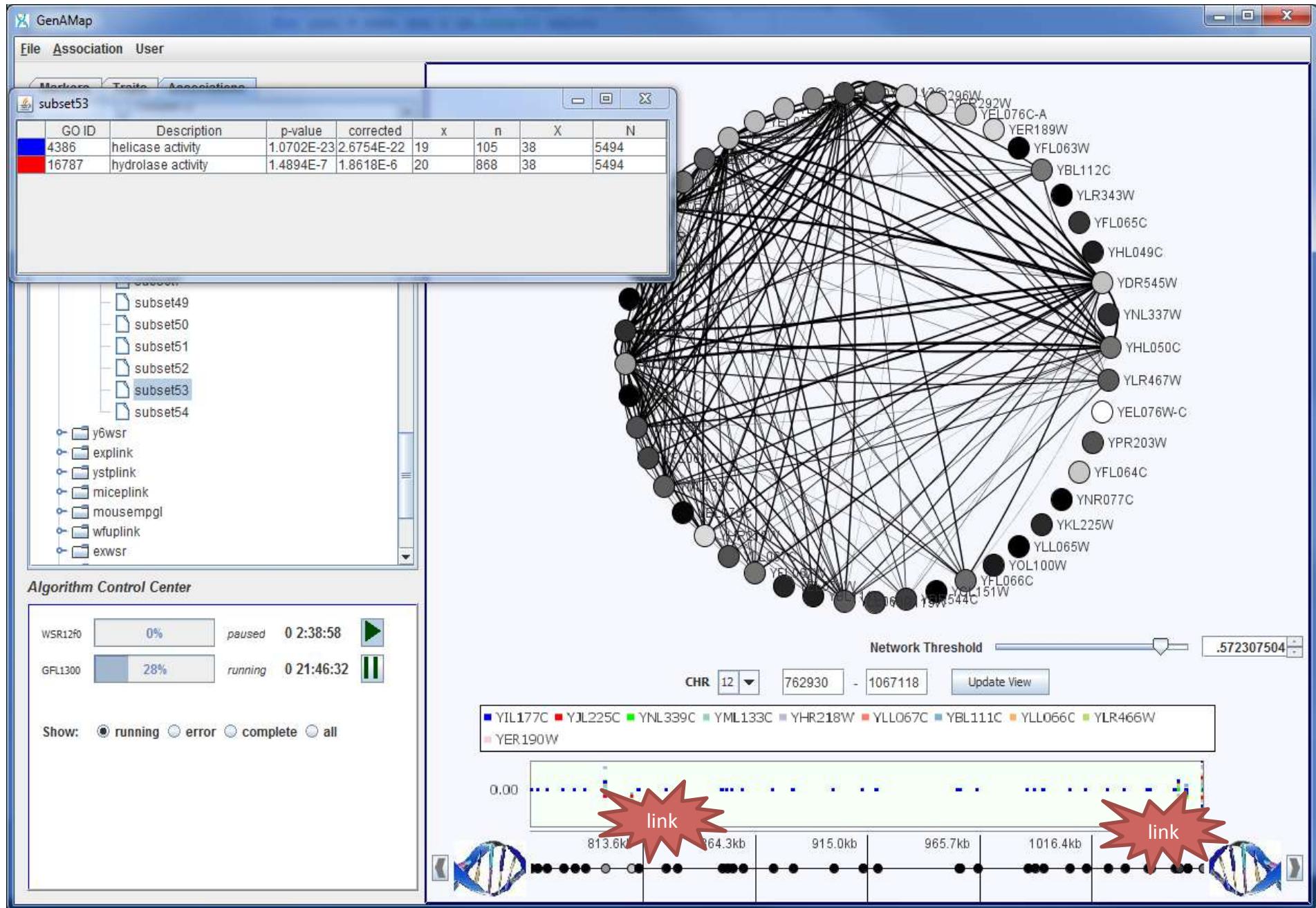
- Literature 
- Retrieve Sequ 

Investigate interacting eQTL hotspots

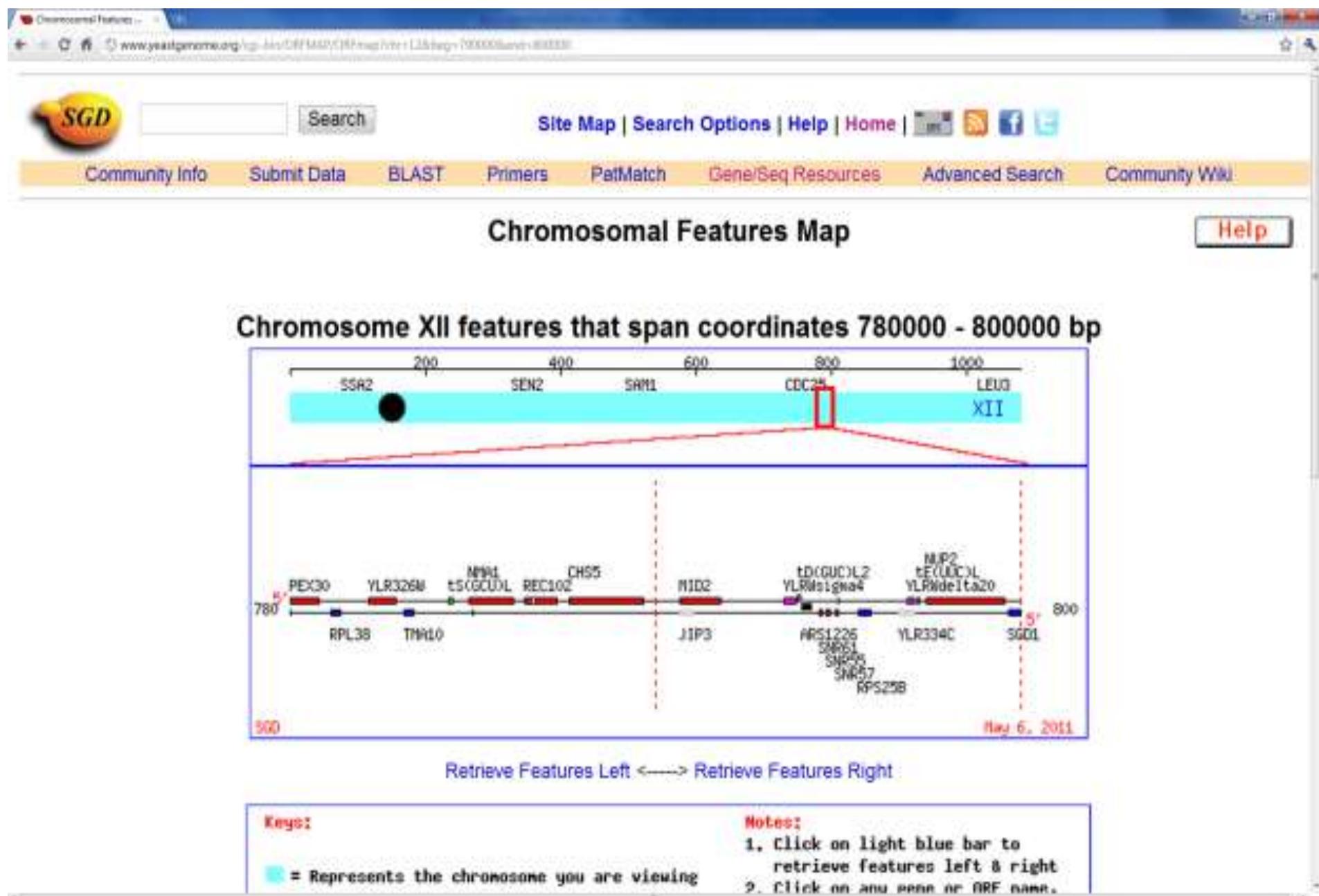




Two SNPs on chromosome 12 are associated with these genes.



NUP2

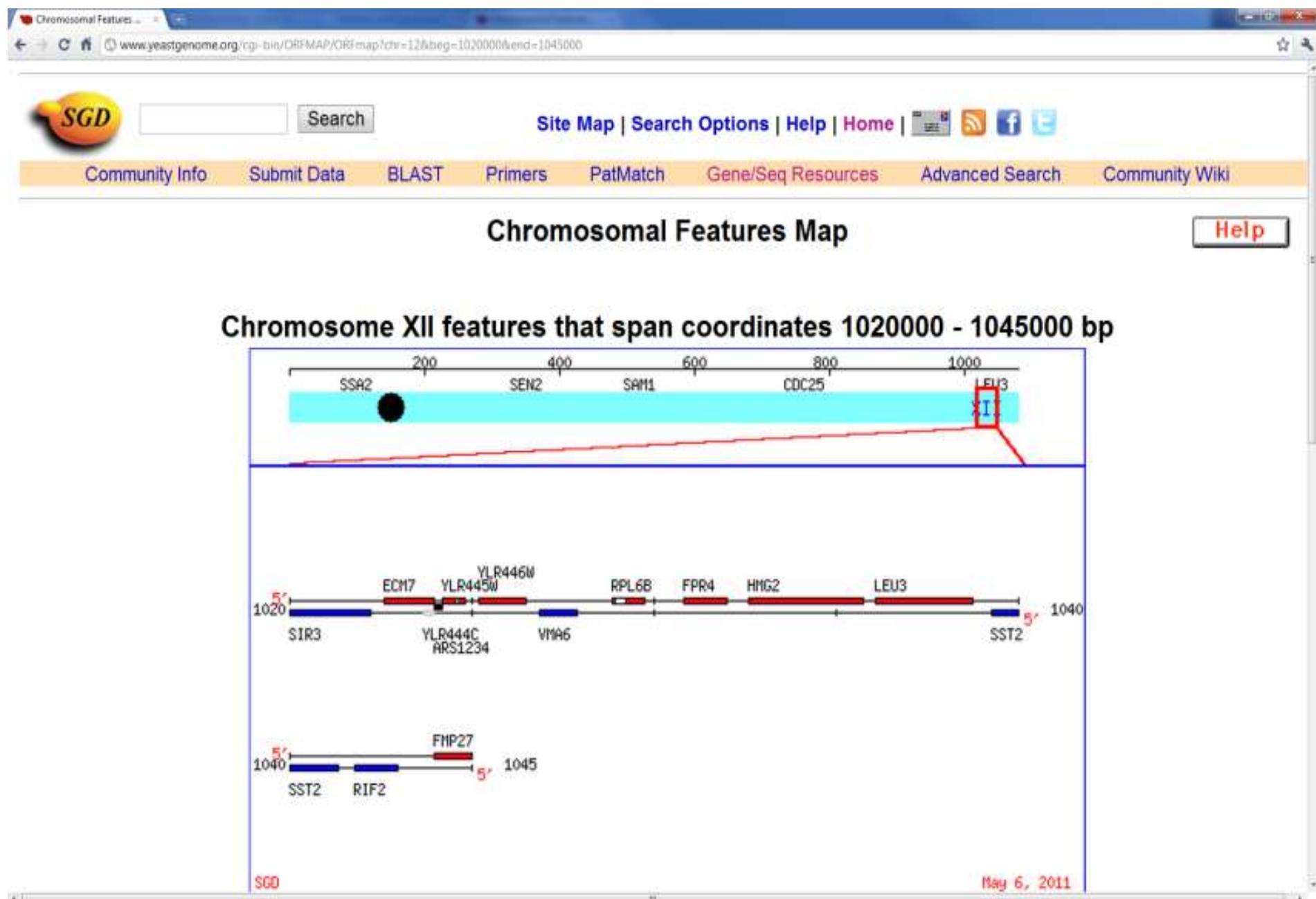


www.yeastgenome.org/cgi-bin/locus.fpl?locus=NUP2

Description	Nucleoporin involved in nucleocytoplasmic transport, bin nucleoplasmic or cytoplasmic faces of the nuclear pore complex depending on Ran-GTP levels; also has a role in chromatin organization (1 , 2 , 3 , 4 , 5 and see Summary Paragraph)
Name Description	Nuclear Pore
GO Annotations	All NUP2 GO evidence and references View Computational GO annotations for NUP2
Molecular Function	• structural molecule activity (TAS)
Biological Process	• chromatin silencing at telomere (IMP) • mRNA export from nucleus (TAS) • mRNA-binding (hnRNP) protein import into nucleus (TAS) • NLS-bearing substrate import into nucleus (TAS) • nuclear pore organization (TAS) • nuclear-transcribed mRNA catabolic process, non-stop decay (IMP) • protein export from nucleus (TAS) • protein targeting to membrane (IMP) • ribosomal protein import into nucleus (TAS) • rRNA export from nucleus (TAS) • snRNA export from nucleus (TAS) • snRNP protein import into nucleus (TAS) • tRNA export from nucleus (TAS)
Cellular Component	• nuclear chromatin (IDA) - nuclear pore (IDA)

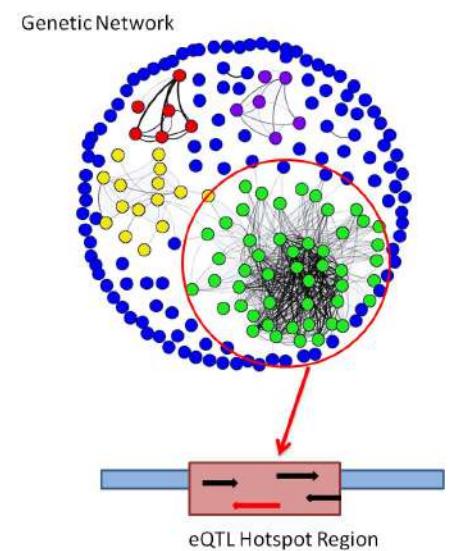
- Literature
[Literature Guide](#)
- Retrieve Sequences
[Genomic DNA](#)
- Sequence Analysis
[BLASTP](#)
- Protein Info & Tools
[Protein Info](#)
[View](#)
- Localization Resources
[YeastRC Localization](#)
- Interactions
[BioGRID \(Torch\)](#)
- Phenotype Resources
[PROPHET](#)
- Maps & Displays
[Chromosomal](#)

SIR3 / RIF2

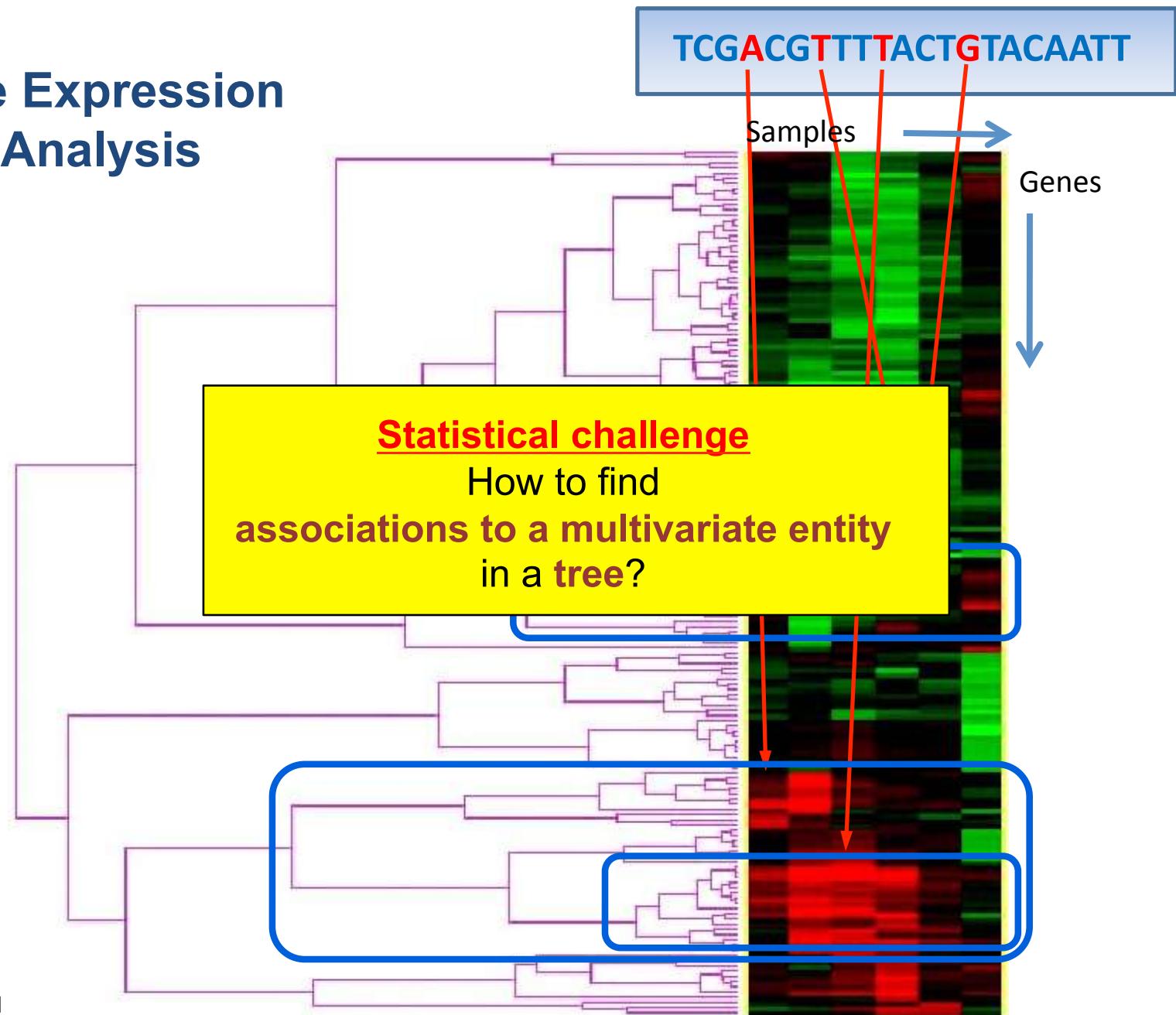


Summary

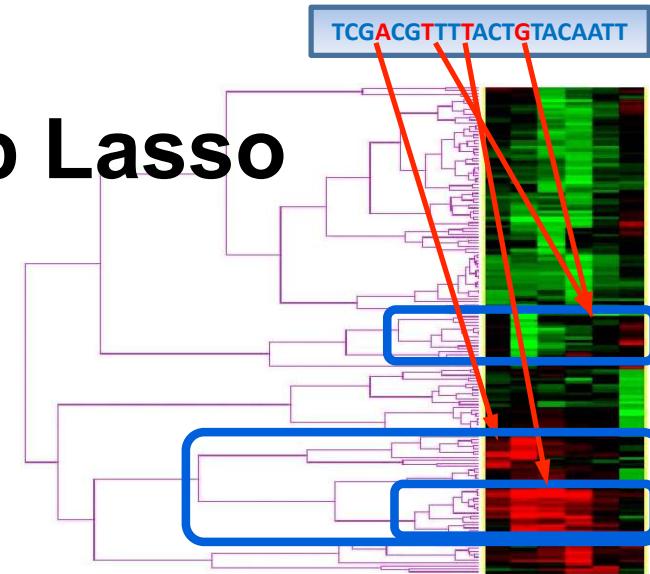
- Exploiting **transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
- Structured association – a new paradigm
 - Association to a **graph**-structured **transcriptome**
 - Graph-guided fused lasso (Kim & Xing, PLoS Genetics, 2009)
- More structures (next)
 - Leveraging gene expression hierarchical cluster
 - Association to a **tree**-structured transcriptome
 - Tree-guided group lasso (Kim & Xing, ICML 2010)



Gene Expression Trait Analysis



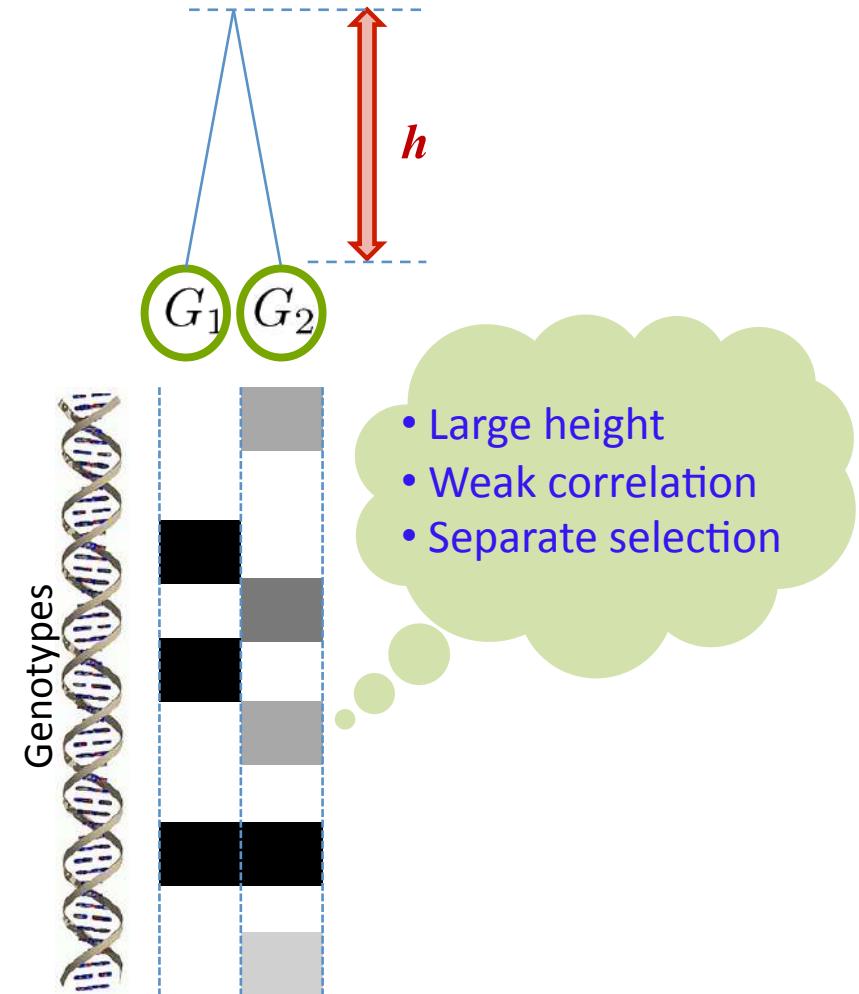
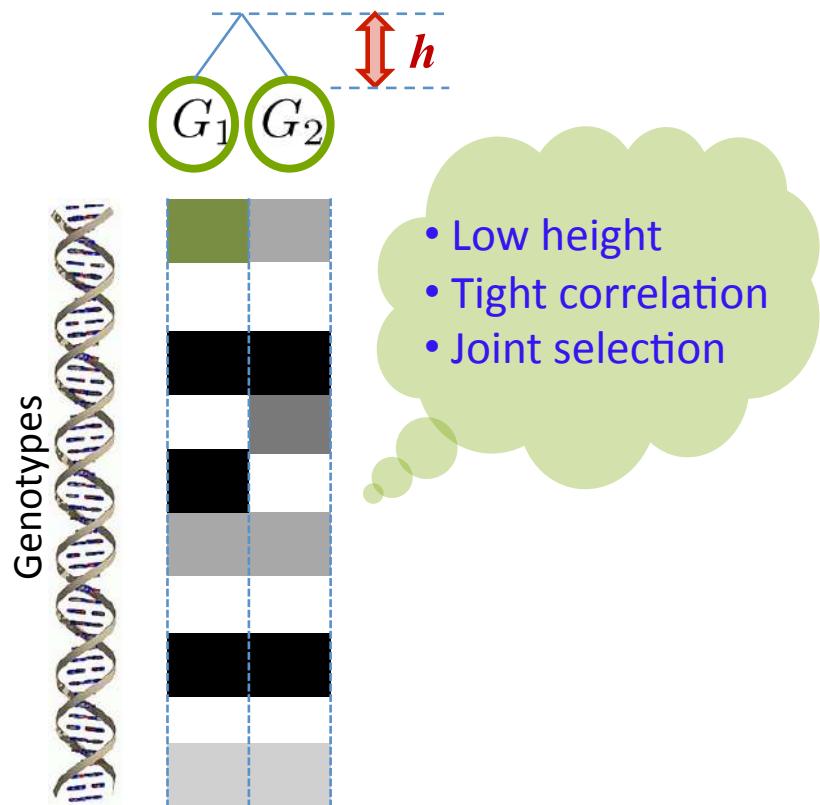
Tree-Guided Group Lasso



- Why tree?
 - Tree represents a **clustering structure**
 - **Scalability** to a very large number of phenotypes
 - Graph : $O(|\mathcal{V}|^2)$ edges
 - Tree : $O(|\mathcal{V}|)$ edges
 - Expression quantitative trait mapping (eQTL)
 - Agglomerative hierarchical clustering is a popular tool

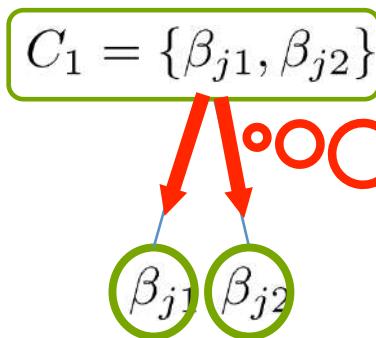
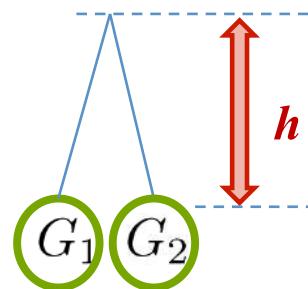
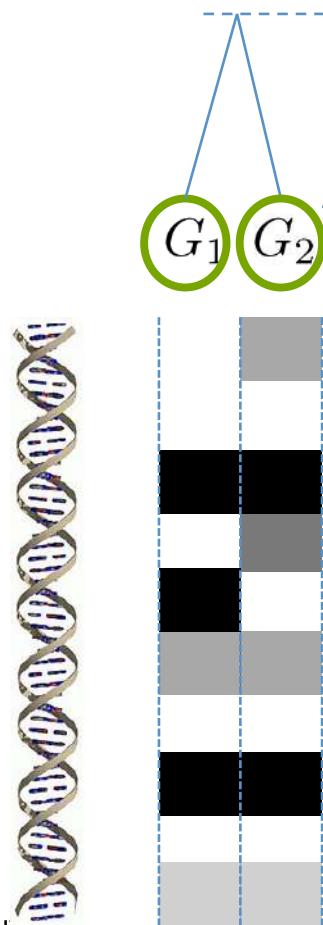
Tree-Guided Group Lasso

- In a simple case of two genes



Tree-Guided Group Lasso

- In a simple case of two genes



Select the child nodes **jointly** or **separately**?

Tree-guided group lasso

$$\operatorname{argmin} (y - X\beta)' \cdot (y - X\beta) + \lambda \sum_j \left[h(|\beta_{j1}| + |\beta_{j2}|) + (1 - h)(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}) \right]$$

L_1 penalty

- Lasso penalty
- **Separate** selection

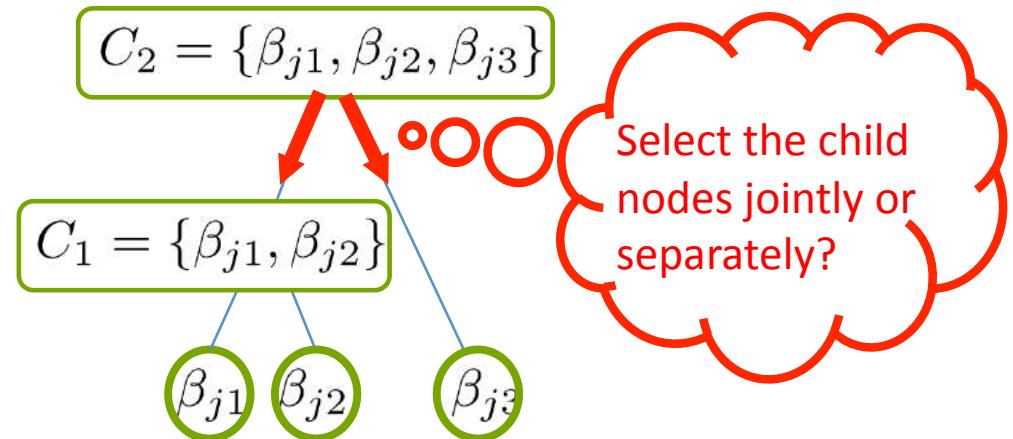
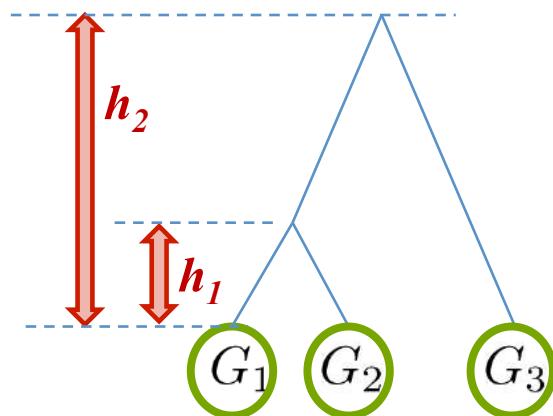
L_2 penalty

- Group lasso
- **Joint** selection

Similar to elastic net (Zou & Hastie, J. R. Statist. Soc. 2005)

Tree-Guided Group Lasso

- For a general tree



Tree-guided group lasso

$$\operatorname{argmin} (y - X\beta)' \cdot (y - X\beta)$$

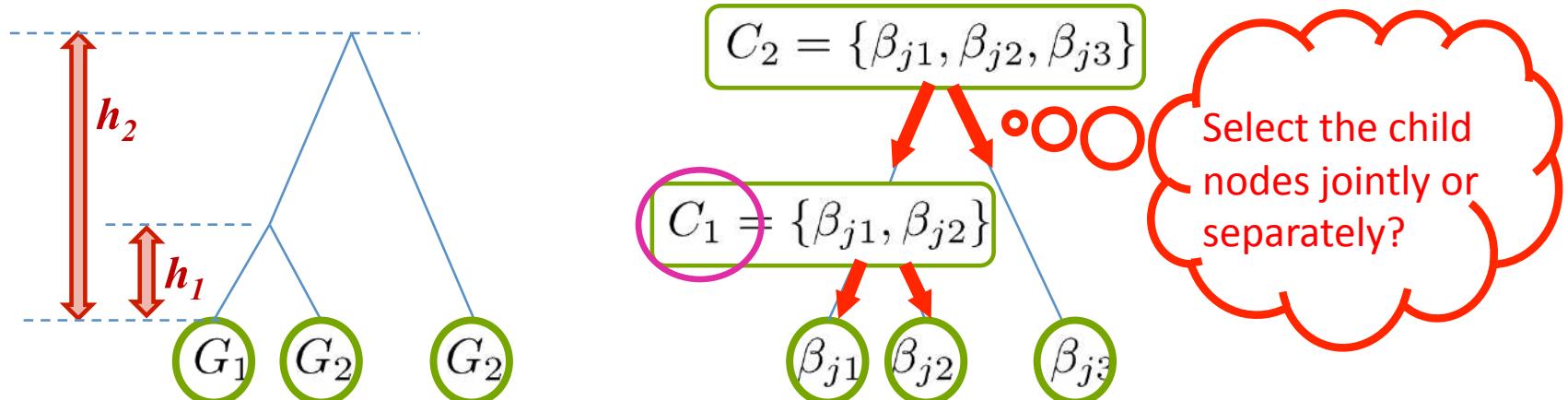
$$+ \lambda \sum_j \left[(1 - h_2) \left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2} \right) + h_2 (|C_1| + |\beta_{j3}|) \right]$$

Joint
selection

Separate
selection

Tree-Guided Group Lasso

- For a general tree



Tree-guided group lasso

$$\text{argmin } (y - X\beta)' \cdot (y - X\beta)$$

$$+ \lambda \sum_j \left[(1 - h_2) \left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2} \right) + h_2 (|C_1| + |\beta_{j3}|) \right]$$

$$(1 - h_1) \left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2} \right) + h_1 (|\beta_{j1}| + |\beta_{j2}|)$$

Joint selection

Separate selection

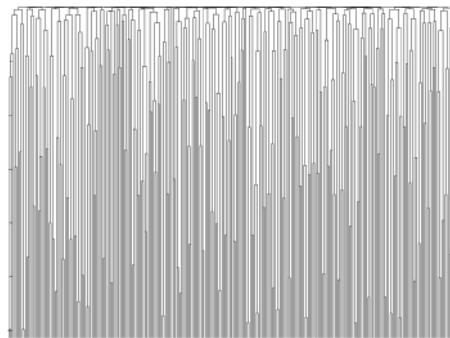
Estimating Parameters

- Second-order cone program

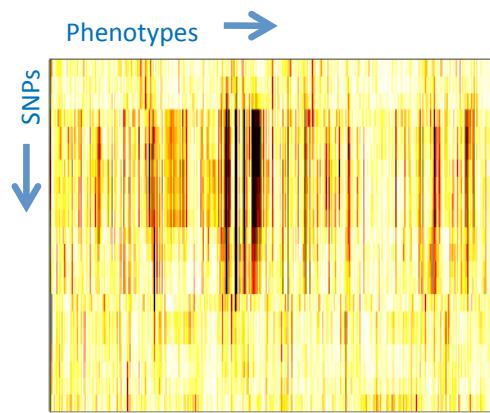
$$\hat{\mathbf{B}}^T = \operatorname{argmin}_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda \sum_j \sum_{v \in V} w_v \|\boldsymbol{\beta}_{G_v}^j\|_2$$

- Many publicly available software packages for solving convex optimization problems can be used

Yeast eQTL Analysis



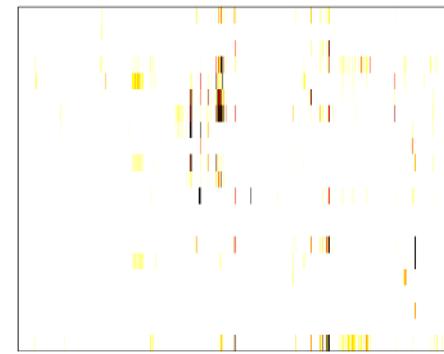
Hierarchical
clustering tree



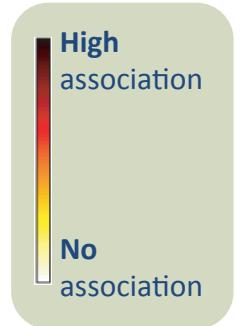
Single-Marker
Single-Trait Test



Lasso



Tree-guided
group lasso



Summary

- Exploiting transcriptome structure in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Association to a **graph**-structured transcriptome using graph-guided fused lasso
 - Leveraging gene expression hierarchical cluster
 - Association to a **tree**-structured transcriptome using tree-guided group lasso

Summary: why care about structure?

- Theoretically, it increase the power [Mladen and Xing, 2010]

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(m_{\max}^*)}] \geq 1 - C_1 \exp\left(-C_2 \frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log(p), \log(T)\}}\right).$$

Summary: why care about structure?

- Incorporating more information leads to enhanced discovery and lower false discovery (results from simulation test)
 - If our underlying model is correct!

Method	Power (multi-trait)	FDR
GFlasso	0.889	0.009
plink	0.803	0.014
Sum-rank	0.748	0.008
Screen & Clean	0.780	0.153
Lasso	0.991	0.886
Forward Selection	0.931	0.867

Improving Scalability of Structured Genome-Transcriptome-Phenome Association Methods

- Common challenges in optimization
 - Tree-guided group lasso, GFllasso, temporally-smoothed lasso
 - Non-smooth penalty
 - Non-separability
 - Coordinate descent algorithm (Friedman et al., Annals of Applied Statistics 2007) cannot be applied.
- Proximal-gradient method (xi et al., submitted)
 - Step 1: Introduce a smooth approximation to the non-smooth penalty
 - Step 2: Apply accelerated gradient method

Proximal Gradient Descent

Original Problem:

$$\arg \min_{\beta \in \mathbb{R}^J} f(\beta) \equiv \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \Omega(\beta)$$

$$\Omega(\beta) = \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta$$

Approximation Problem:

$$\arg \min_{\beta \in \mathbb{R}^J} \tilde{f}(\beta) \equiv \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + f_\mu(\beta)$$

$$f_\mu(\beta) = \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta - \mu d(\alpha)$$

Gradient of the Approximation:

$$\nabla \tilde{f}(\beta) = \mathbf{X}^T (\mathbf{X}\beta - \mathbf{y}) + C^T \alpha^*$$

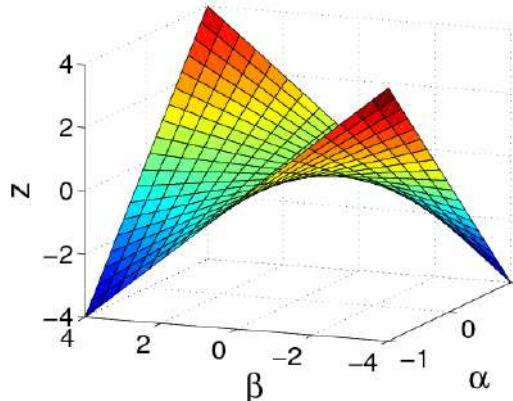
$$\alpha^* = \arg \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta - \mu d(\alpha)$$

$\nabla \tilde{f}(\beta)$ is Lipschitz continuous with the Lipschitz constant L

$$L = \lambda_{\max}(\mathbf{X}^T \mathbf{X}) + L_\mu$$

Geometric Interpretation

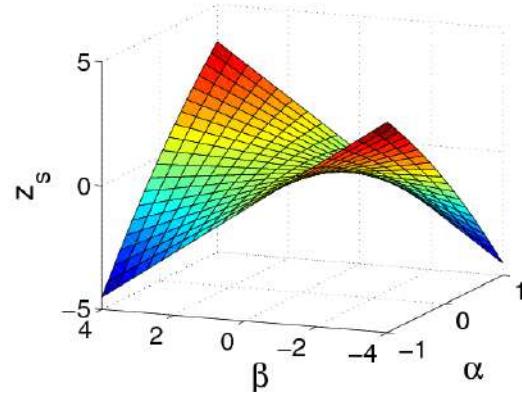
- Smooth approximation



$$z(\alpha, \beta) = \alpha\beta$$

Projection onto
 $z - \beta$ Plane

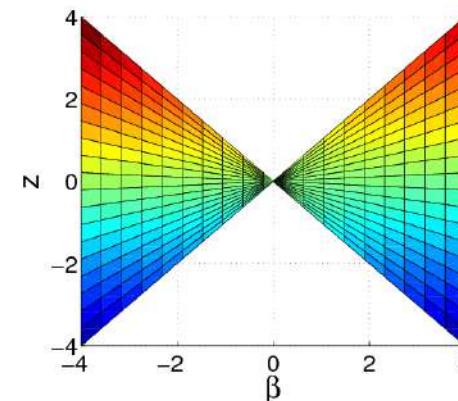
$\longrightarrow f_0(\beta) = \max_{\alpha \in [-1,1]} z(\alpha, \beta) = |\beta|$



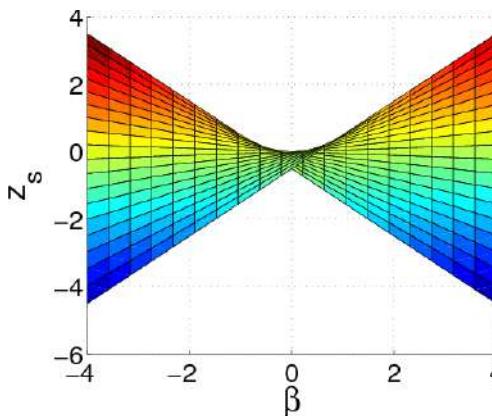
$$z_s(\alpha, \beta) = \alpha\beta - \frac{1}{2}\alpha^2$$

Projection onto
 $z_s - \beta$ Plane

$\longrightarrow f_1(\beta) = \max_{\alpha \in [-1,1]} z_s(\alpha, \beta)$



Uppermost Line
Nonsmooth



Uppermost Line
Smooth

Convergence Rate

Theorem: If we require $f(\boldsymbol{\beta}^t) - f(\boldsymbol{\beta}^*) \leq \epsilon$ and set $\mu = \frac{\epsilon}{2D}$, the number of iterations is upper bounded by:

$$t \leq \sqrt{\frac{4\|\boldsymbol{\beta}^*\|_2^2}{\epsilon} \left(\lambda_{\max}(\mathbf{X}^T \mathbf{X}) + \frac{2D\|\Gamma\|^2}{\epsilon} \right)} = O\left(\frac{1}{\epsilon}\right)$$

Remarks: state of the art IPM method for SOCP converges at a rate $O\left(\frac{1}{\epsilon^2}\right)$

Multi-Task Time Complexity

- Pre-compute:

$$\mathbf{X}^T \mathbf{X}, \mathbf{X}^T \mathbf{Y}: O(J^2N + JKN)$$

- Per-iteration Complexity (computing gradient)

Tree:

IPM for SOCP	$O\left(J^2(K + \mathcal{G})^2(KN + J(\sum_{g \in \mathcal{G}} g))\right)$
Proximal-Gradient	$O(J^2K + J \sum_{g \in \mathcal{G}} g)$

Graph:

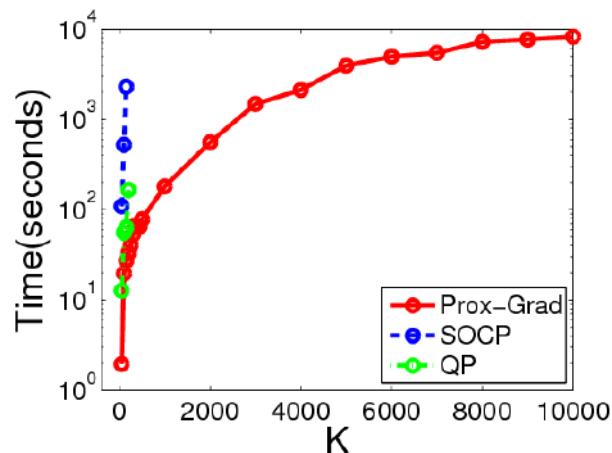
IPM for SOCP	$O\left(J^2(K + E)^2(KN + JK + J E)\right)$
Proximal-Gradient	$O(J^2K + J E)$

Proximal-Gradient: Independent of Sample Size
Linear in #.of Traits

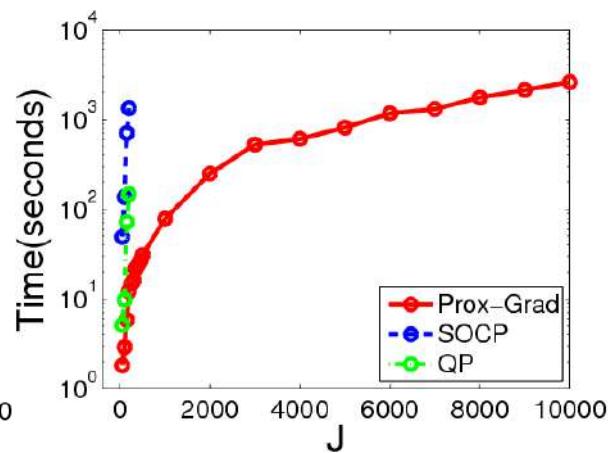


Experiments

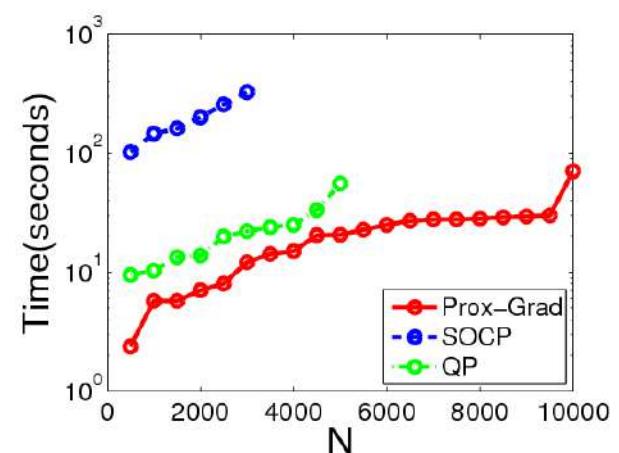
- Multi-task Graph Structured Sparse Learning (GFlasso)



$N = 500, J = 100$



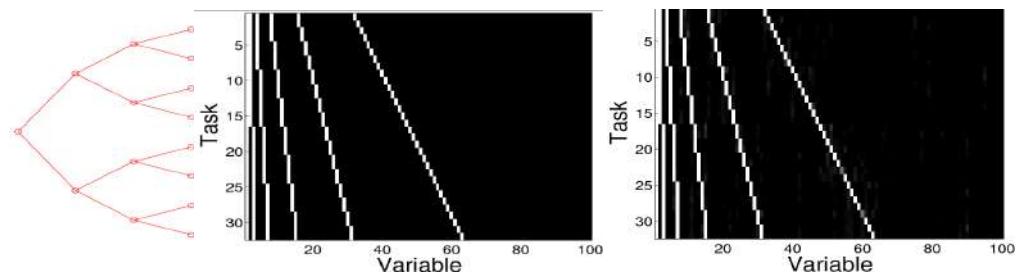
$N = 1000, K = 50$



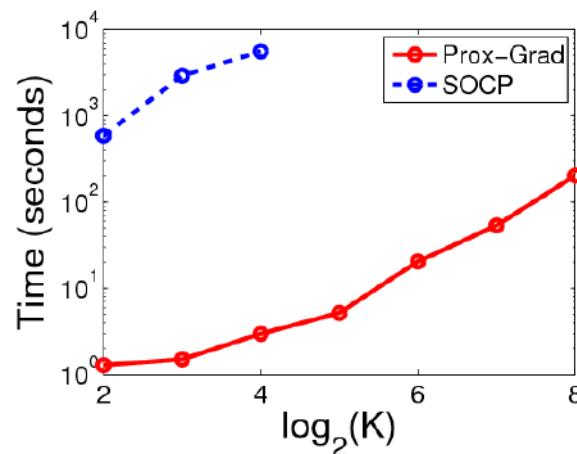
$J = 100, K = 50$

$$\mu = 10^{-4}, \rho = 0.5$$

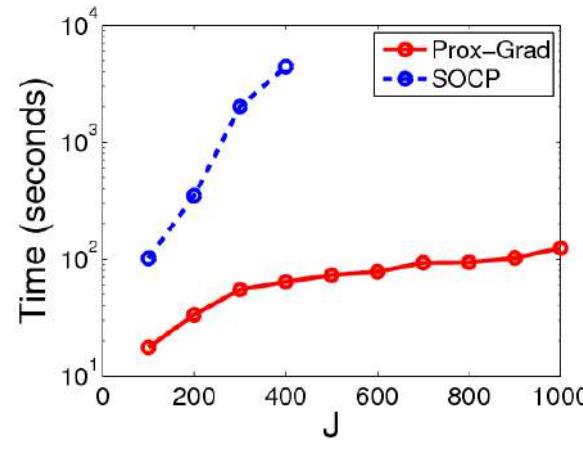
Experiments



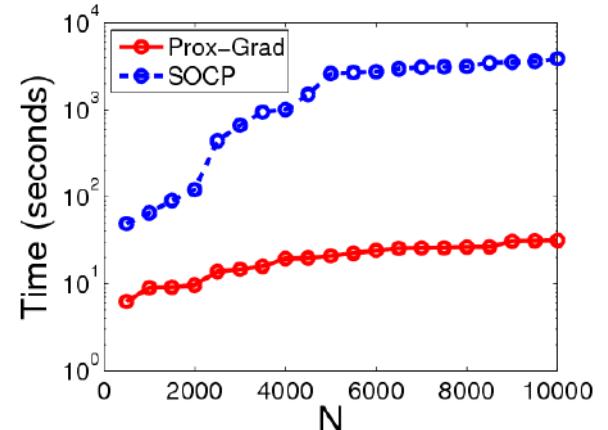
- Multi-task Tree-Structured Sparse Learning (TreeLasso)



$N = 1000, J = 600$



$N = 1000, K = 32$



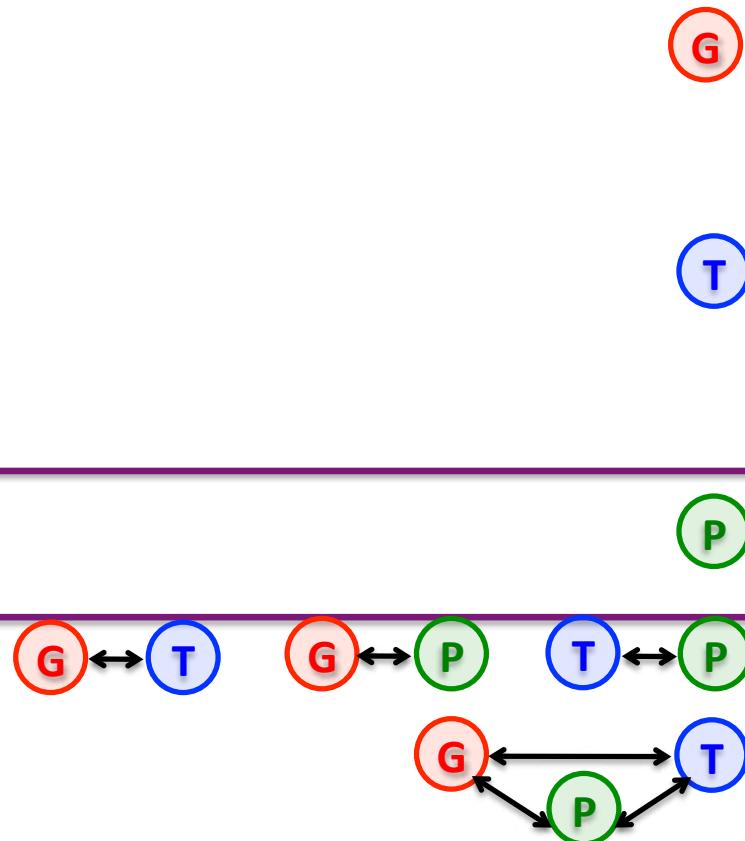
$J = 100, K = 32$

Overview

- **Genome structure** in association analysis
 - Linkage disequilibrium
 - Population structure
 - Epistasis
- **Transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Leveraging gene expression tree

- **Phenome structure** in association analysis
 - Pleiotropy
 - Dynamic trait

- **Two-way** structured association
- **Three-way** structured association
- Visualization software

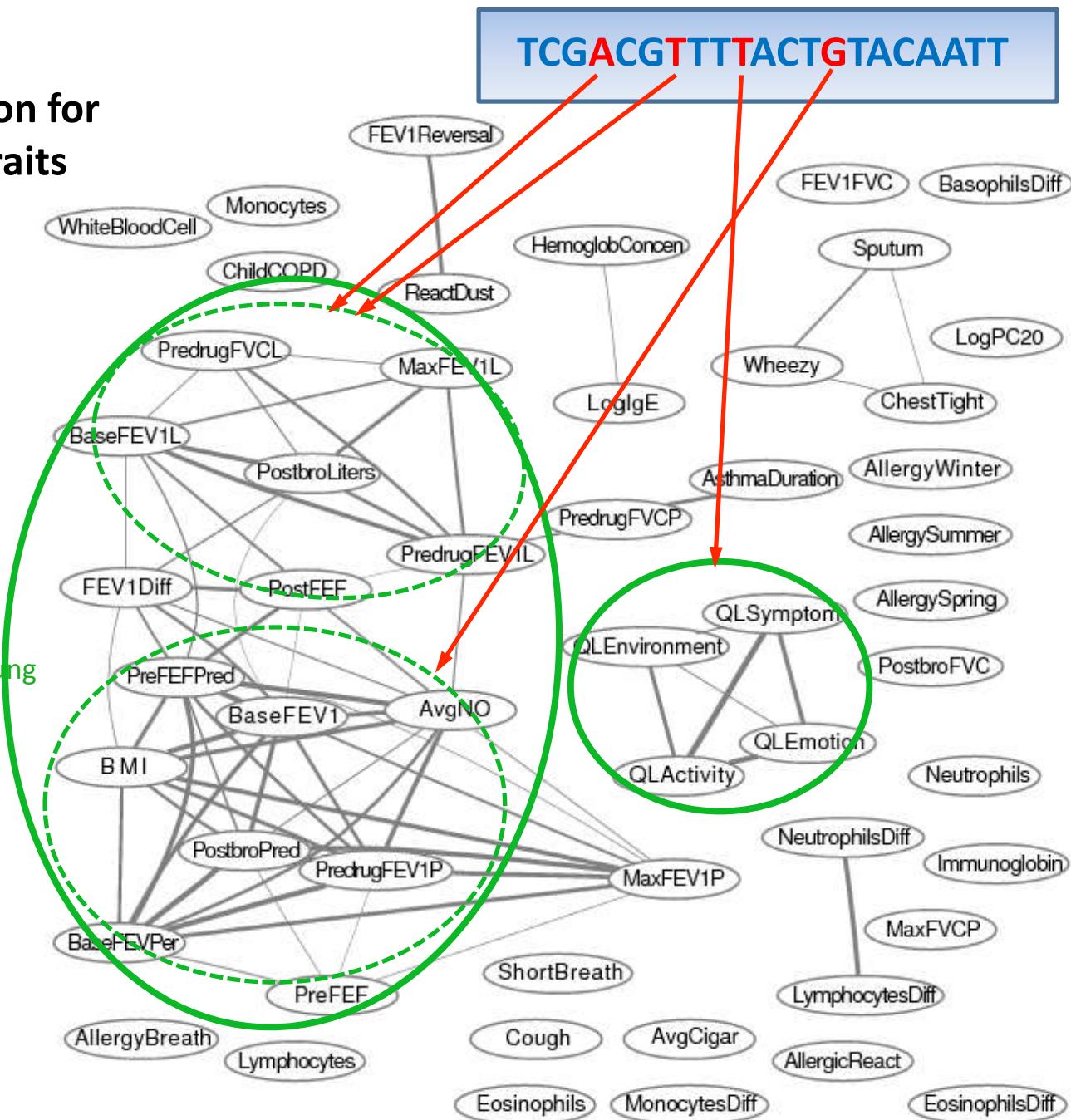


Association to Phenome

- Pleiotropy: a genetic locus influencing multiple different phenotypes
- Methods for transcriptome association can be applied
 - Graph-guided fused lasso (Kim & Xing, PLoS Genetics 2009)
 - Tree-guided fused lasso (Kim & Xing, ICML 2010)

Genetic Association for Asthma Clinical Traits

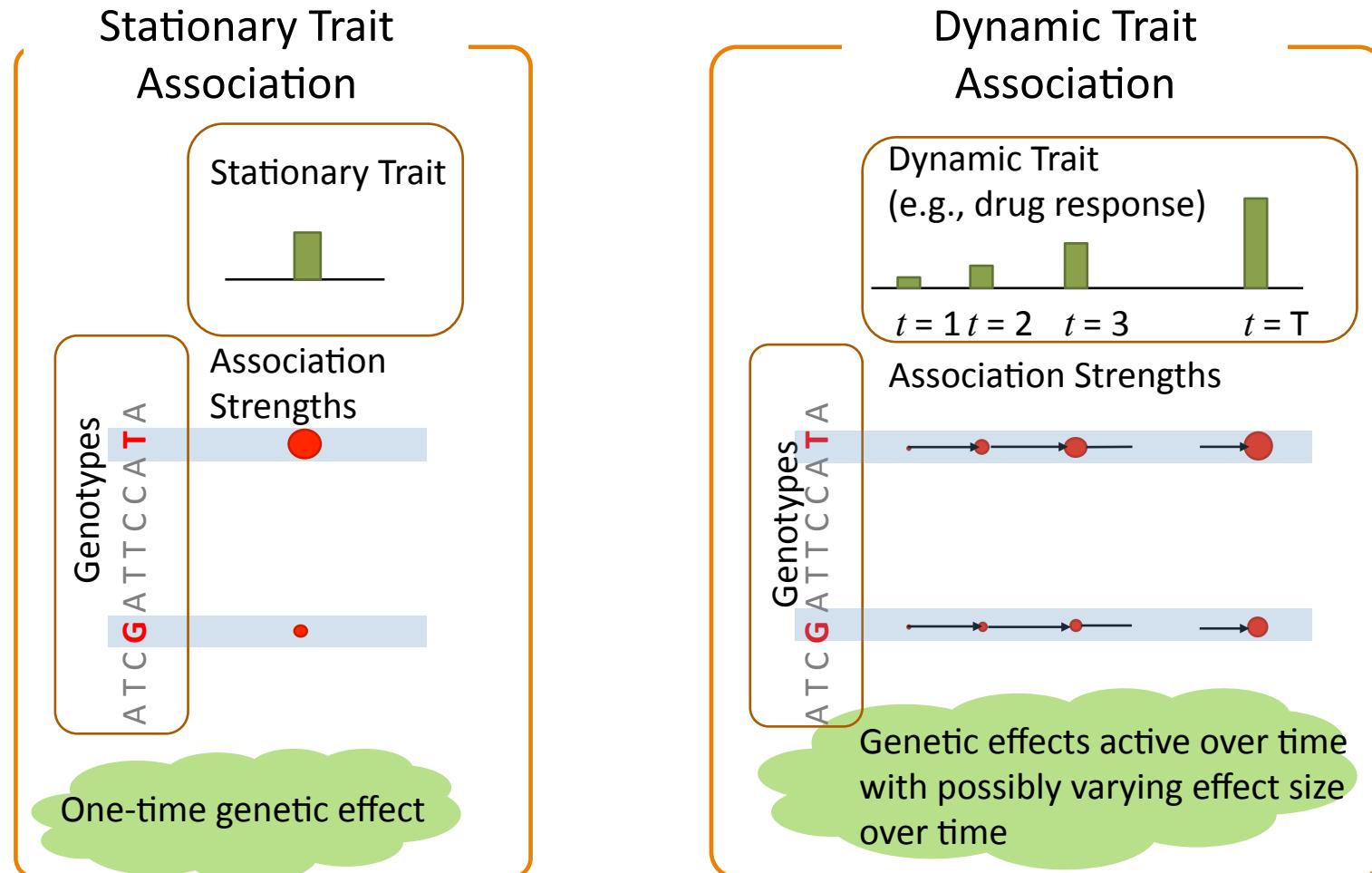
Subnetworks for lung physiology



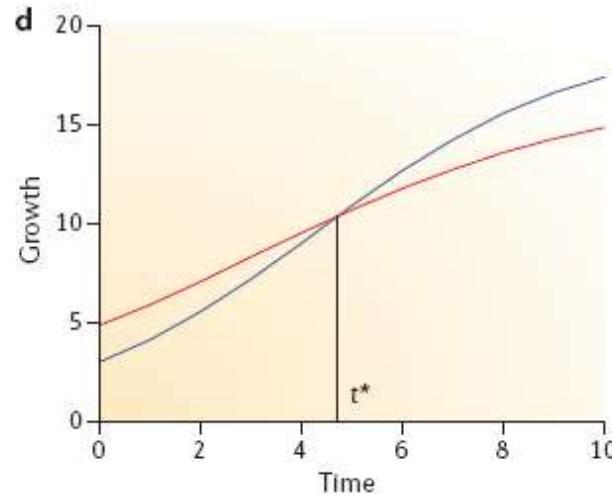
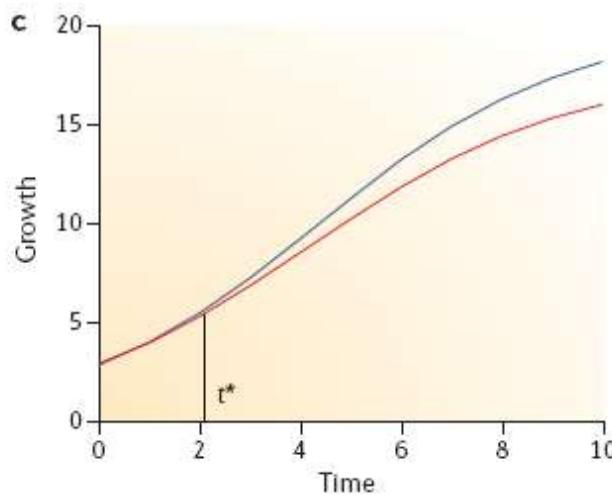
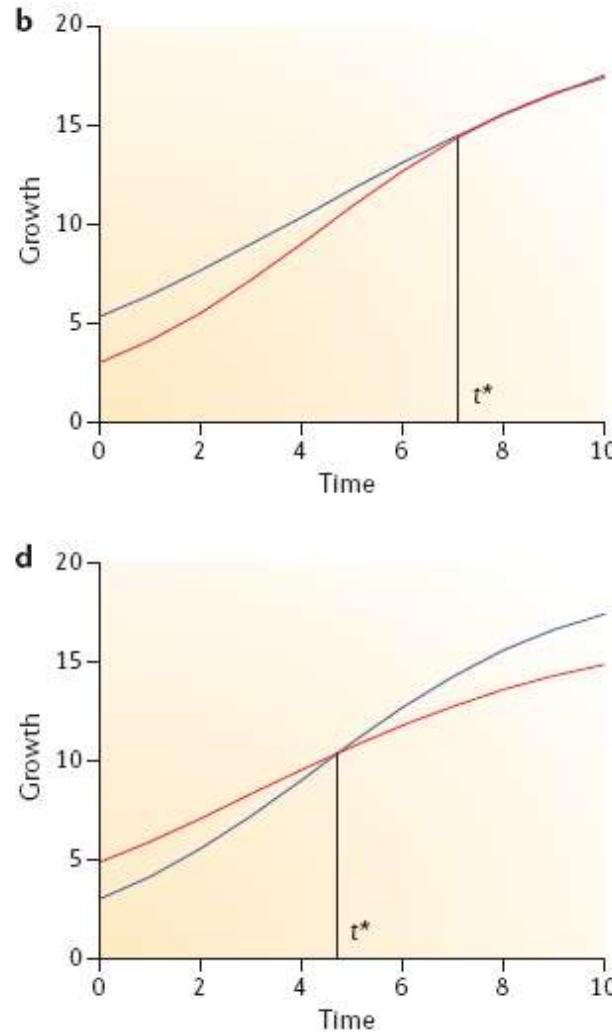
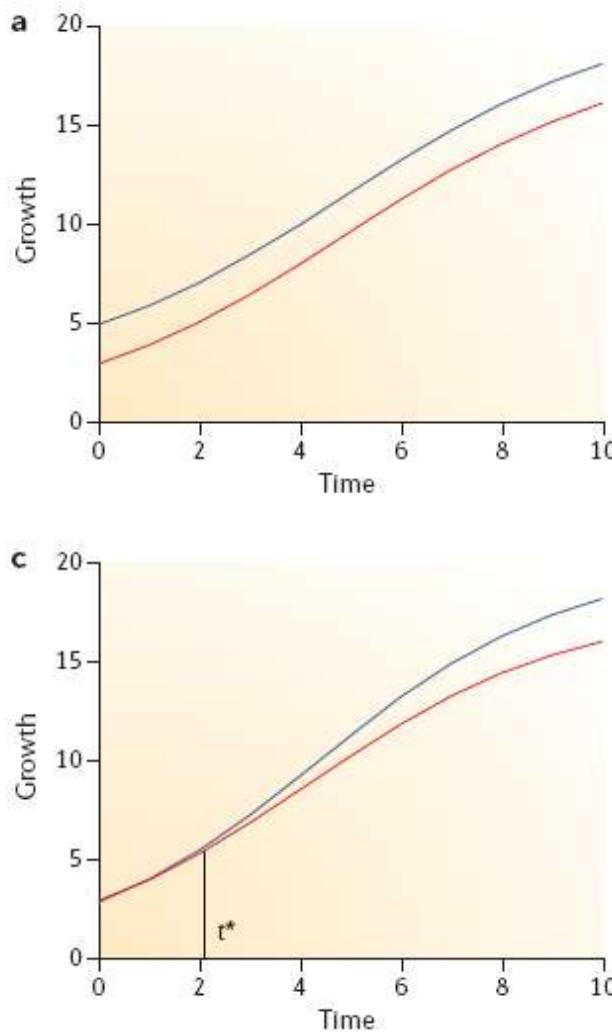
Dynamic-Trait Association Mapping

- Dynamic trait with a temporal trend
 - Growth of tumor over time
 - Height, weight over time
 - Gene expressions over time in cell cycle or embryonic development
- Are there underlying genetic variants that influence the overall trend over time?

Dynamic Trait (d-trait) Association



Genetic Control of Growth Trajectories



Temporally-Smoothed Lasso

- Step 1: Autoregressive Model
 - Captures the shape of the temporal trend in the d-trait data
 - Estimates the model parameters based on the d-trait data only
- Step 2: Temporally-Smoothed Lasso
 - Penalized regression framework
 - Incorporates the estimated d-trait shape parameters from Step 1
 - Detects time-varying genetic effects on the d-trait

Step 1: Autoregressive Model

Autoregressive Model :

$$\mathbf{y}_{k,t+1} = \alpha_{k,t} \mathbf{y}_{k,t} + \alpha_{k,t}^0 \mathbf{1} + \epsilon$$

Estimating Model Parameters:

$$\hat{\alpha}_{k,t} = \operatorname{argmin} (\mathbf{y}_{k,t+1} - \alpha_{k,t} \mathbf{y}_{k,t} - \alpha_{k,t}^0)^T \cdot (\mathbf{y}_{k,t+1} - \alpha_{k,t} \mathbf{y}_{k,t} - \alpha_{k,t}^0)$$

Estimates of the Model Parameters:

$$\hat{\alpha}_{k,t} = \frac{\mathbf{y}_{k,t}^T \cdot \mathbf{y}_{k,t+1}}{\mathbf{y}_{k,t}^T \cdot \mathbf{y}_{k,t}}$$

Step 2: Temporally-Smoothed Lasso

$$\hat{\mathbf{B}}^{\text{dyn}} = \operatorname{argmin} \sum_k \sum_t (\mathbf{y}_{k,t} - \mathbf{X}\boldsymbol{\beta}_{k,t})^T \cdot (\mathbf{y}_{k,t} - \mathbf{X}\boldsymbol{\beta}_{k,t}) + \lambda \cdot \sum_k \sum_t \sum_j \beta_{k,t}^j + \gamma \cdot \sum_j \sum_k \sum_{t=1}^{T-1} |\beta_{k,t+1}^j - \alpha_{k,t} \beta_{k,t}^j|$$

Autoregressive
parameters from
Step 1

Lasso
Penalty

Temporally-
smoothed Lasso
Penalty

Estimating Association Strengths

$$\hat{\mathbf{B}}^{\text{dyn}} = \operatorname{argmin}_k \sum_t (\mathbf{y}_{k,t} - \mathbf{X}\boldsymbol{\beta}_{k,t})^T \cdot (\mathbf{y}_{k,t} - \mathbf{X}\boldsymbol{\beta}_{k,t})$$

such that $S(\mathbf{B}) \leq s_1, T(\mathbf{B}) \leq s_2$

- Quadratic programming
- Convex optimization: many publicly available software packages can be used.
- **Proximal gradient algorithm** can be applied

Proximal Gradient Descent

Original Problem:

$$\arg \min_{\beta \in \mathbb{R}^J} f(\beta) \equiv \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \Omega(\beta)$$

$$\Omega(\beta) = \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta$$

Approximation Problem:

$$\arg \min_{\beta \in \mathbb{R}^J} \tilde{f}(\beta) \equiv \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + f_\mu(\beta)$$

$$f_\mu(\beta) = \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta - \mu d(\alpha)$$

Gradient of the Approximation:

$$\nabla \tilde{f}(\beta) = \mathbf{X}^T (\mathbf{X}\beta - \mathbf{y}) + C^T \alpha^*$$

$$\alpha^* = \arg \max_{\alpha \in \mathcal{Q}} \alpha^T C \beta - \mu d(\alpha)$$

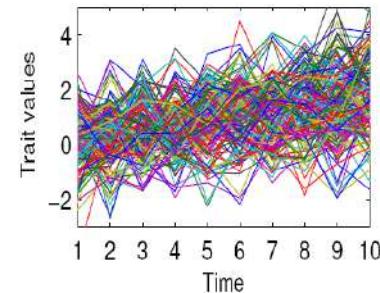
$\nabla \tilde{f}(\beta)$ is Lipschitz continuous with the Lipschitz constant L

$$L = \lambda_{\max}(\mathbf{X}^T \mathbf{X}) + L_\mu$$

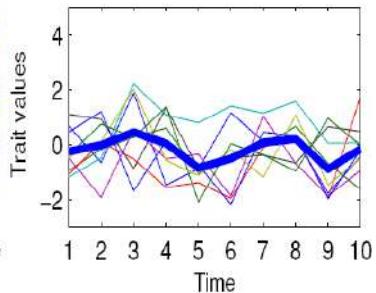
Simulation Study

- Genotype:
 - HapMap-simulated data (The international HapMap Consortium, Nature 2005)
 - 50 SNPs
- D-trait:
 - Linear dynamic
 - Cyclic dynamic
 - 10 time points
- 150 individuals

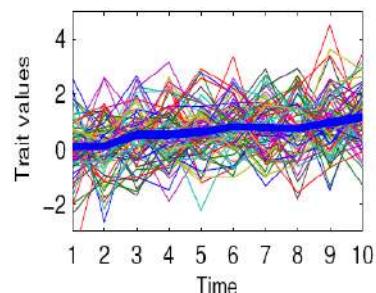
Simulation Study – Linear Dynamic



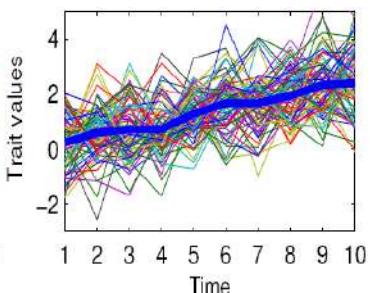
Trait data for
all individuals



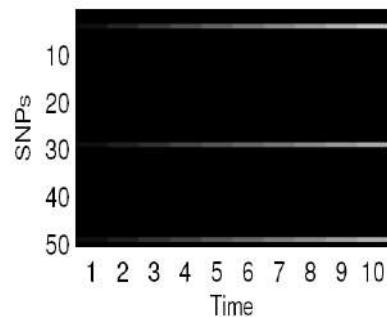
Trait data for
individuals with
no association SNPs



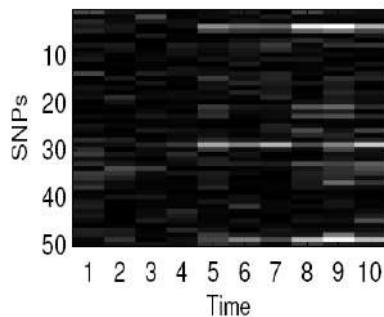
Trait data for
individuals with
1-2 association SNPs



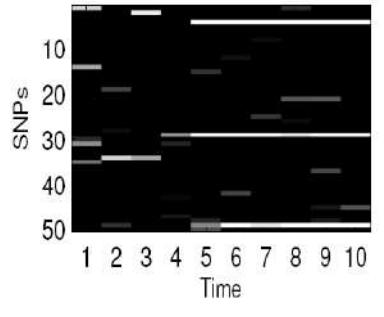
Trait data for
individuals with
>3 association SNPs



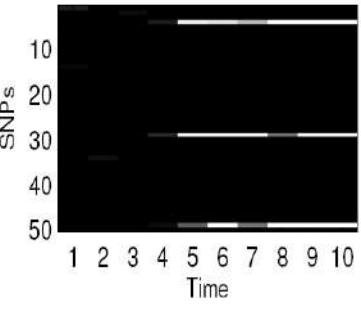
True association
strength



Estimated
association strength
(single SNP analysis)

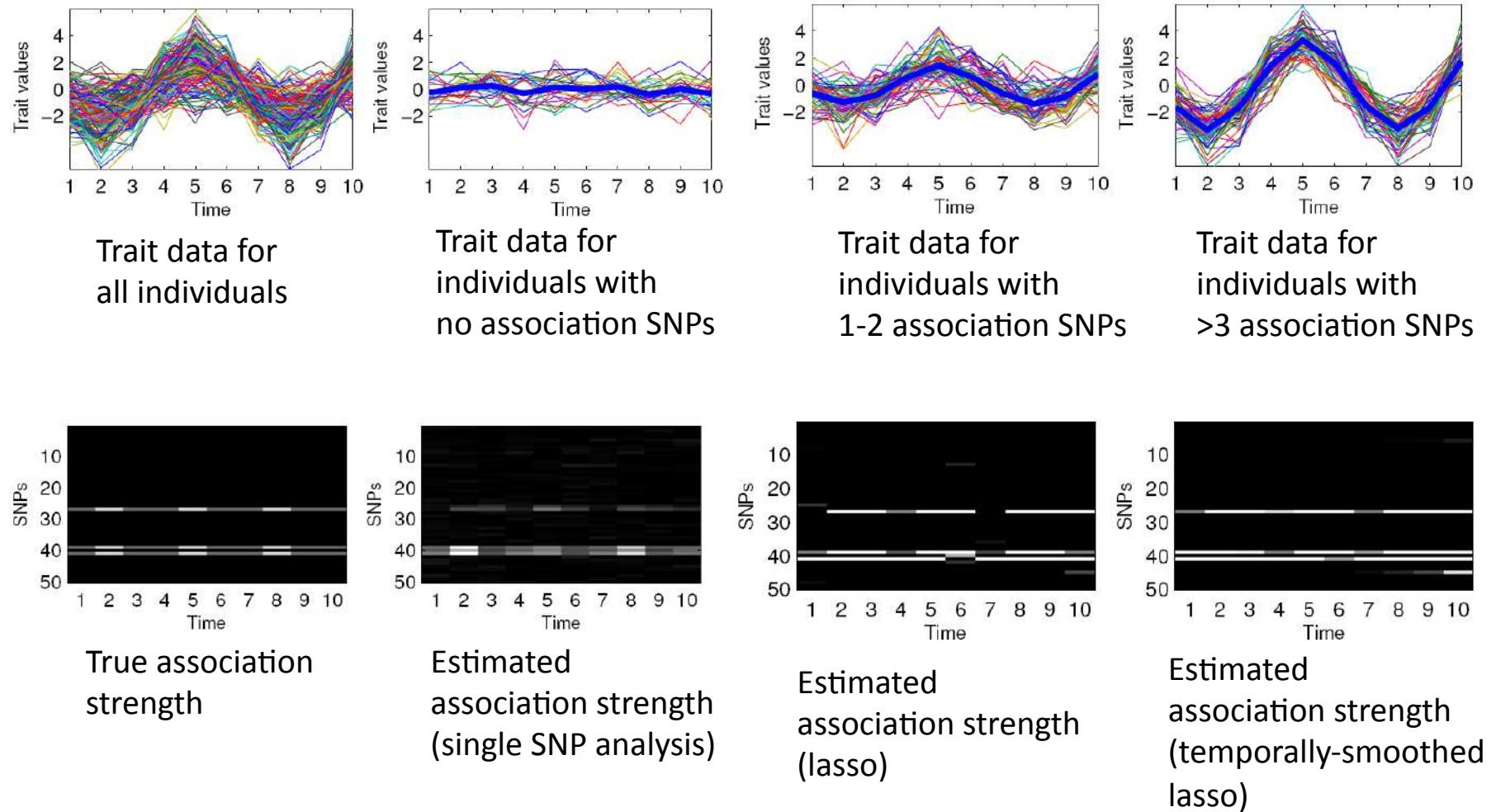


Estimated
association strength
(lasso)



Estimated
association strength
(temporally-smoothed
lasso)

Simulation Study – Cyclic Dynamic



Overview

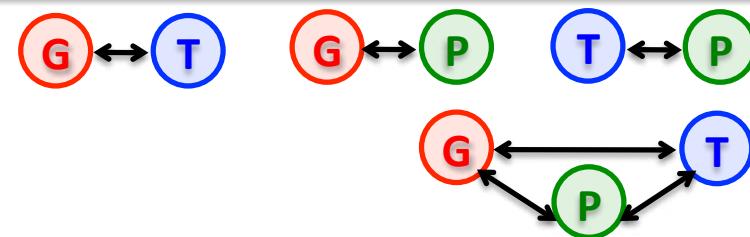
- **Genome structure** in association analysis
 - Linkage disequilibrium
 - Population structure
 - Epistasis
- **Transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Leveraging gene expression tree
- **Phenome structure** in association analysis
 - Pleiotropy
 - Dynamic trait

G

T

P

- Two-way structured association
 - Three-way structured association
- Visualization software



**Genome
Structure**

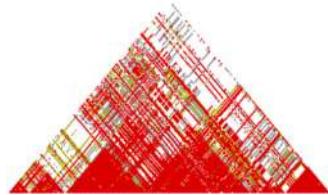
**Structured
Association**



**Phenome
Structure**

Genome Structure

Linkage Disequilibrium



Stochastic block regression
(Kim & Xing, UAI, 2008)

Population Structure



Multi-population group lasso
(Puniyani, Kim, Xing, ISMB, 2010)

Epistasis

ACGTTTACTGTACAATT



Group lasso with networks
(Lee, Kim, Xing, Submitted)

Structured Association



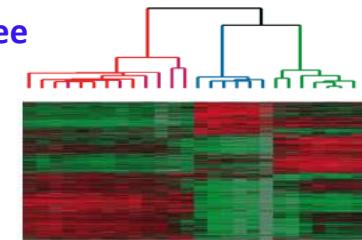
Phenome Structure

Graph



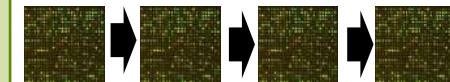
Graph-guided fused lasso
(Kim & Xing, PLoS Genetics, 2009)

Tree

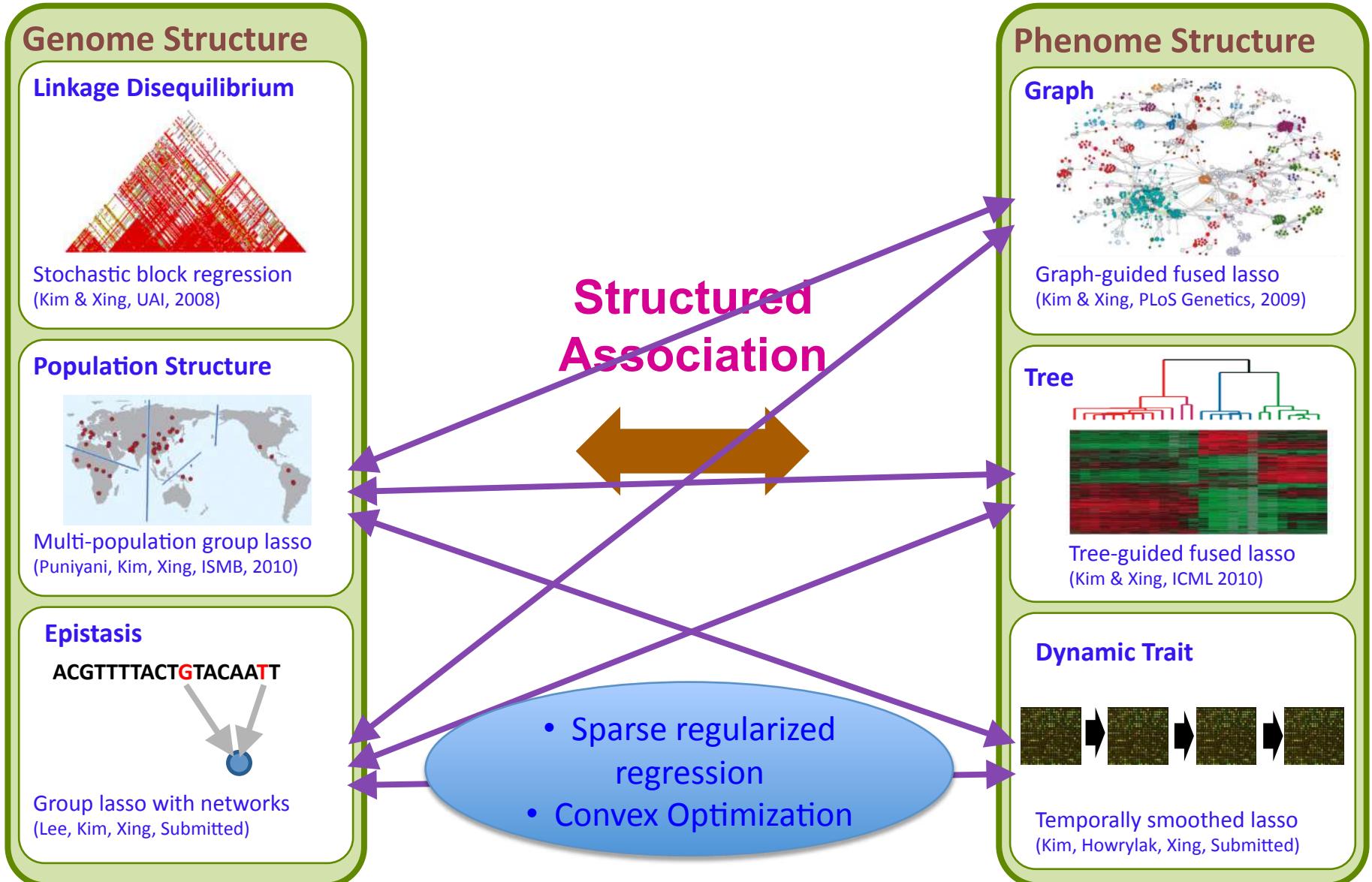


Tree-guided fused lasso
(Kim & Xing, ICML 2010)

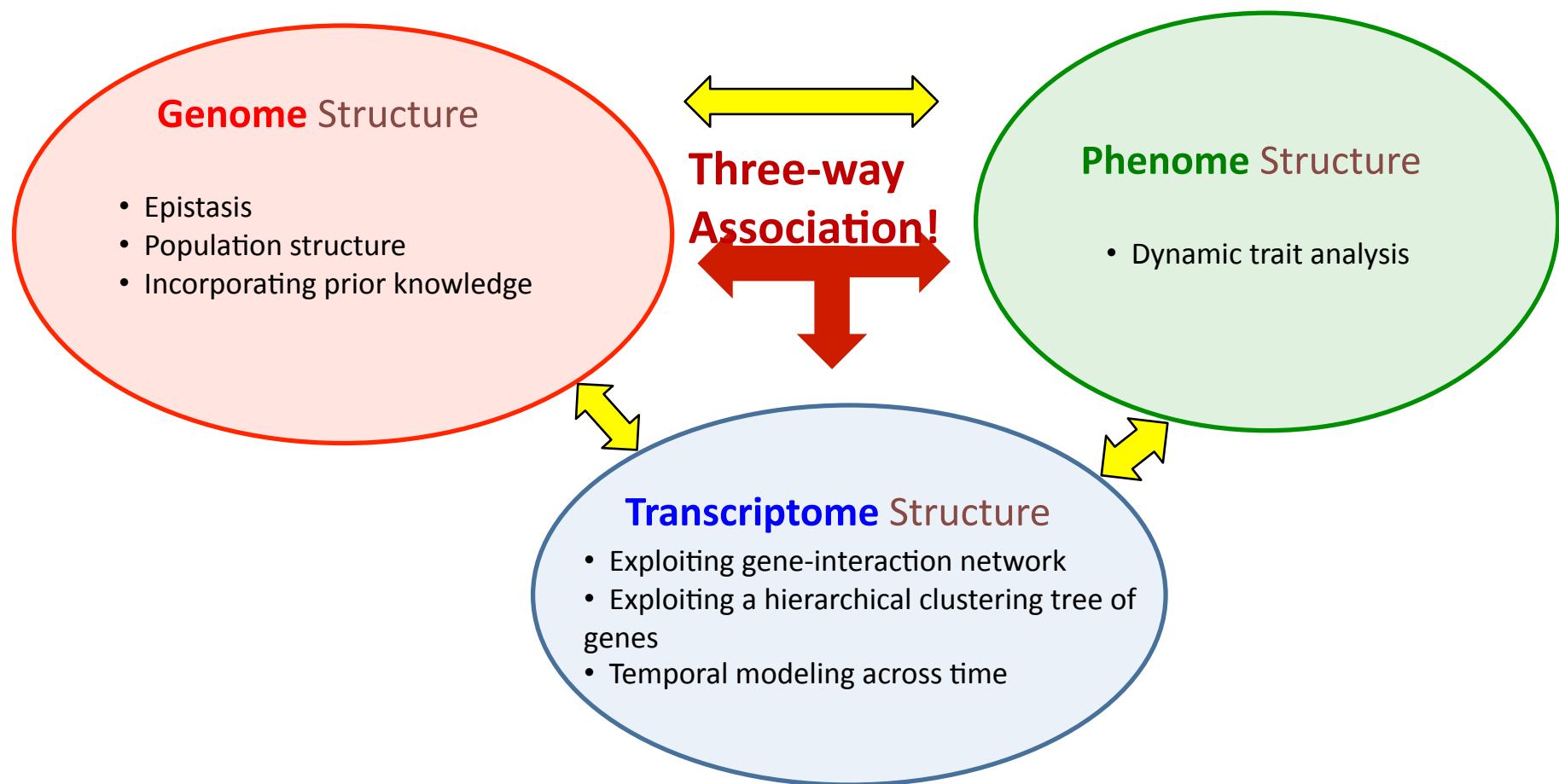
Dynamic Trait



Temporally smoothed lasso
(Kim, Howrylak, Xing, Submitted)



Structured Genome-Transcriptome-Phenome Association Analysis



Summary: why care about structure?

- Theoretically, it increase the power [Mladen and Xing, 2010]

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(m_{\max}^*)}] \geq 1 - C_1 \exp\left(-C_2 \frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log(p), \log(T)\}}\right).$$

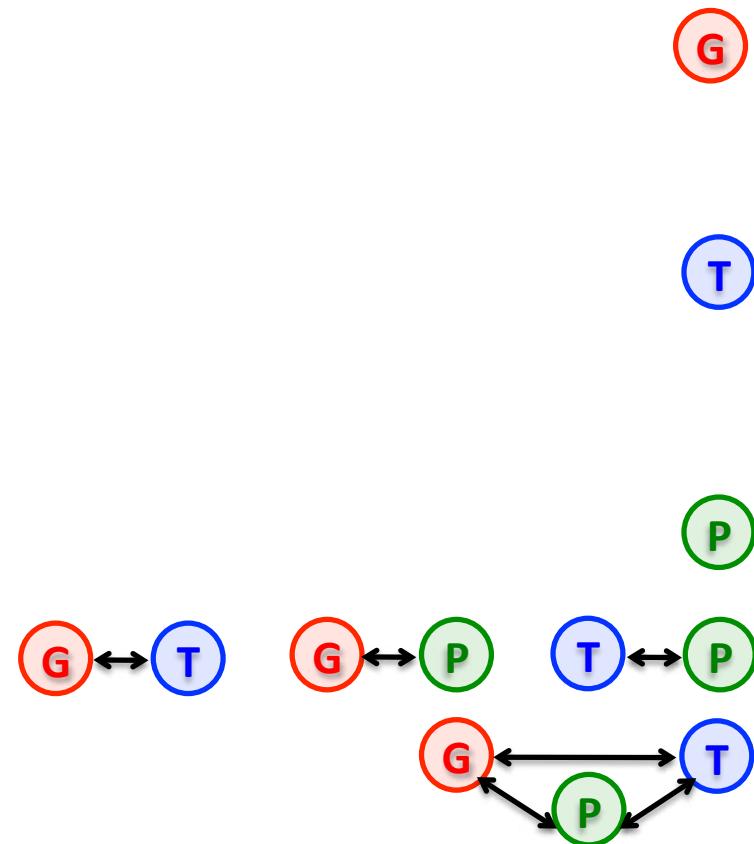
Summary: why care about structure?

- Incorporating more information leads to enhanced discovery and lower false discovery (results from simulation test)
 - If our underlying model is correct!

Method	Power (multi-trait)	FDR
GFlasso	0.889	0.009
plink	0.803	0.014
Sum-rank	0.748	0.008
Screen & Clean	0.780	0.153
Lasso	0.991	0.886
Forward Selection	0.931	0.867

Overview

- **Genome structure** in association analysis
 - Linkage disequilibrium
 - Population structure
 - Epistasis
- **Transcriptome structure** in association analysis
 - Pleiotropy
 - Pathway-based statistical tests
 - Leveraging gene expression network
 - Leveraging gene expression tree
- **Phenome structure** in association analysis
 - Pleiotropy
 - Dynamic trait
- **Two-way** structured association
- **Three-way** structured association
- Visualization software



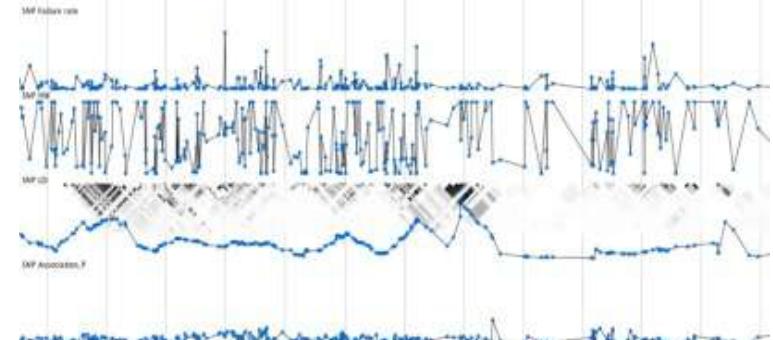
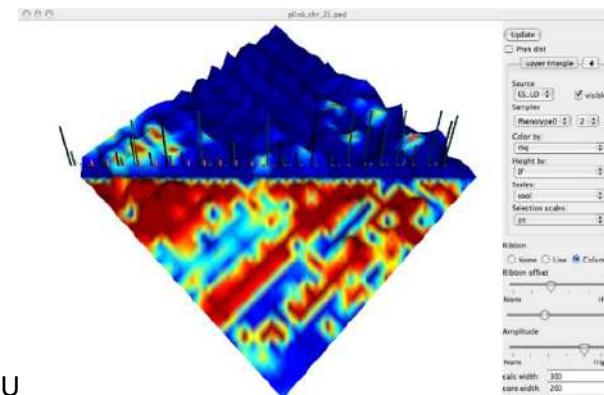
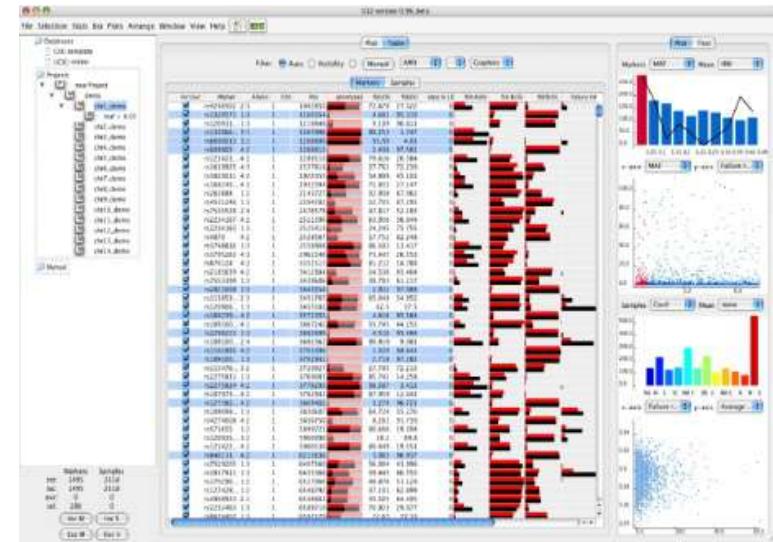
Visualization Softwares for Association Mapping

- **GoldSurfer2** <http://www.well.ox.ac.uk/gs2>
- **GoldenHelix** <http://www.goldenhelix.com/>
- **GWAS GUI** <http://www.sph.umich.edu/csg/weich/browser/>
- **MAVEN** <http://cbc.case.edu/maven/home.do>
- **Association Viewer** http://www.improvedoutcomes.com/docs/WebSiteDocs/Plots/Classification_and_Prediction/SLAM_Assocation_Viewer.htm
- **eQTLExplorer** <http://web.bioinformatics.ic.ac.uk/eqtlexplorer/>
- **eQTL Viewer** <http://statgen.ncsu.edu/eQTLViewer/svgHome.html>



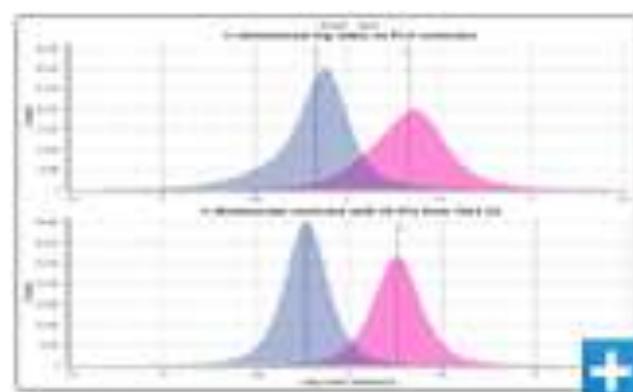
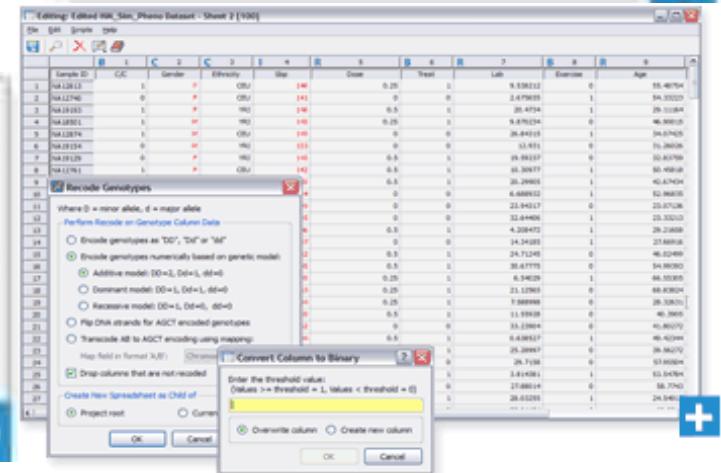
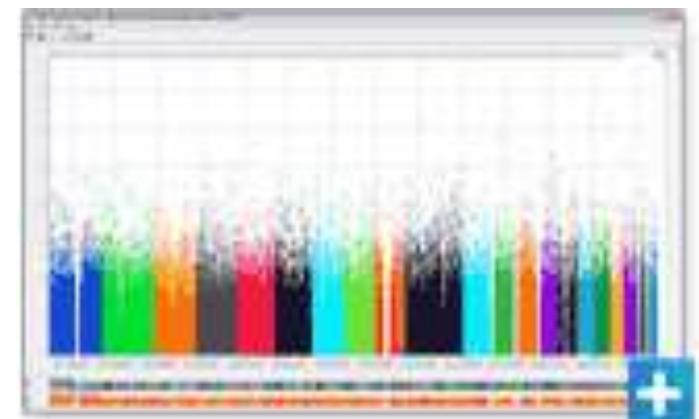
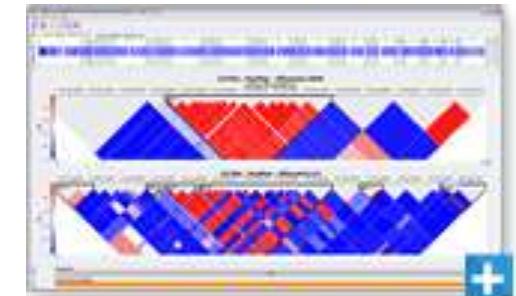
GoldSurfer2

- Developed by Glaxo SmithKline and Oxford University
- Features
 - Explore genome LD structure
 - Standard statistical tests for association
- Disadvantages
 - Runs locally



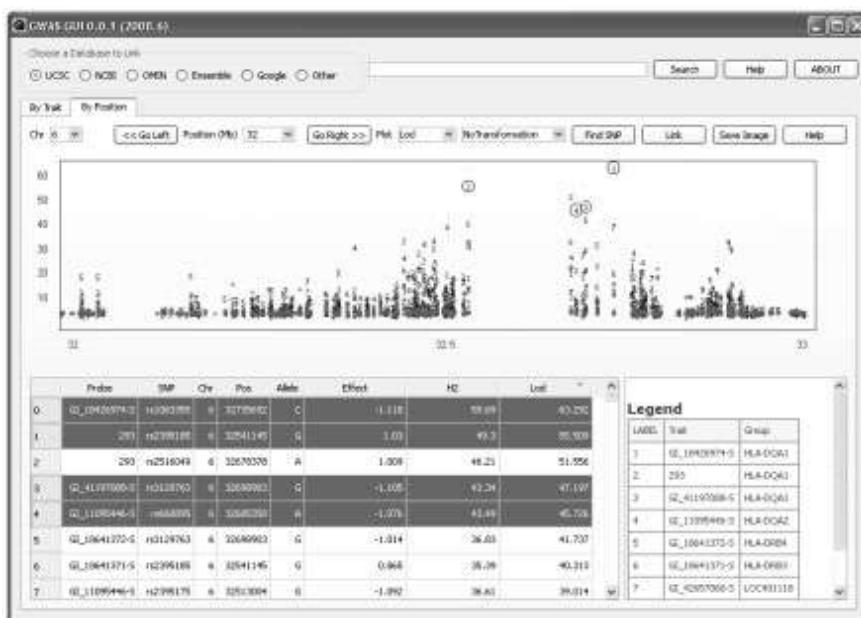
Golden Helix

- Features
 - Standard conventional association tests
 - Links to external data sources
 - Fast computation and storage
 - Data management
 - Quality assurance
- Disadvantages
 - Case/control methodology



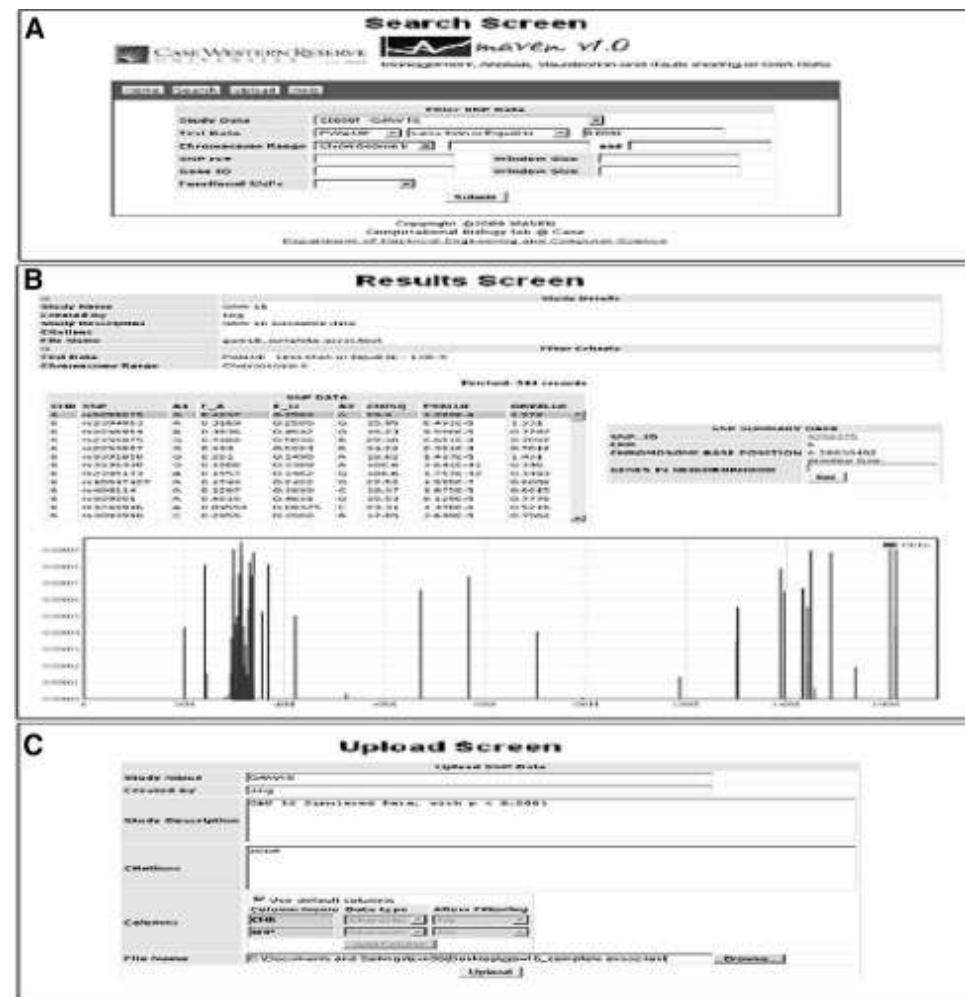
GWAS GUI

- Developed by Center for Statistical Genetics, U of Michigan
- Features
 - Manages huge datasets – including large gene expression datasets
 - Results in a graph or tabular form
 - Searching based on trait or locus
- Disadvantages
 - Must import your own association results
 - Cannot view structure in the gene expression
 - Limited visualizations



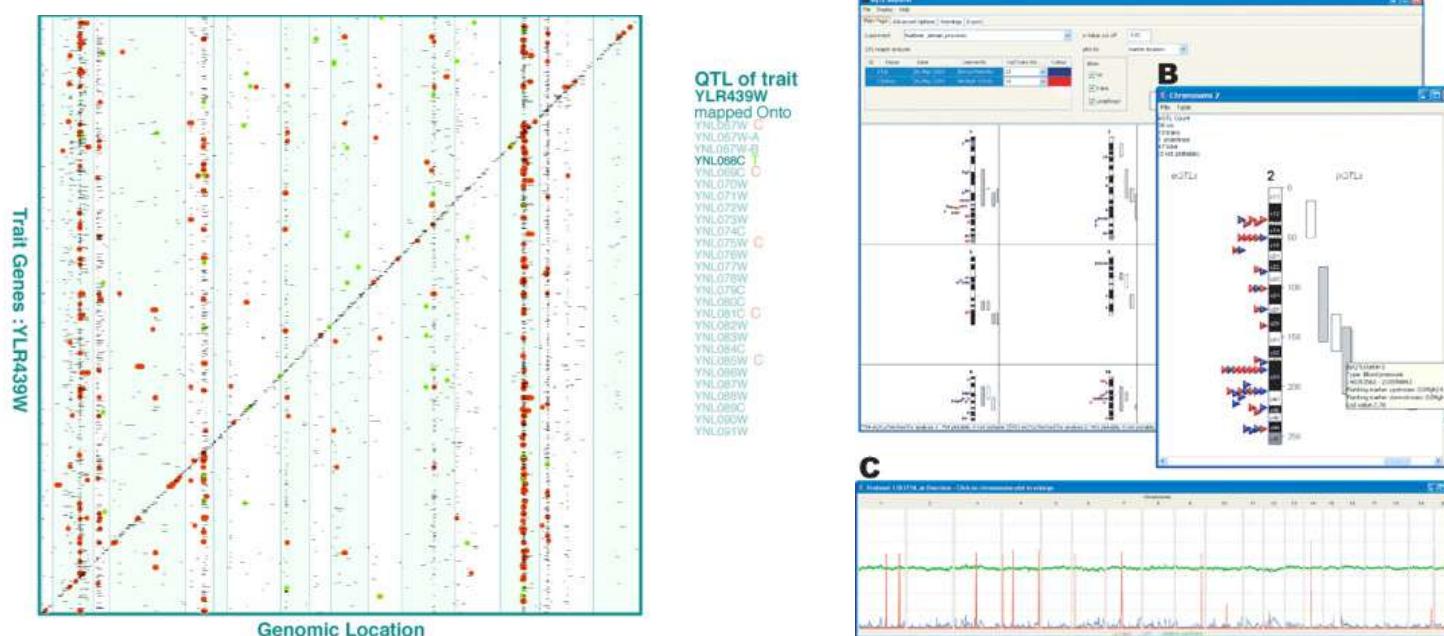
MAVEN and AssociationViewer

- Similar to GWAS GUI, but only for case/control studies
- Maven links to external databases
- AssociationViewer has advanced memory management capability.

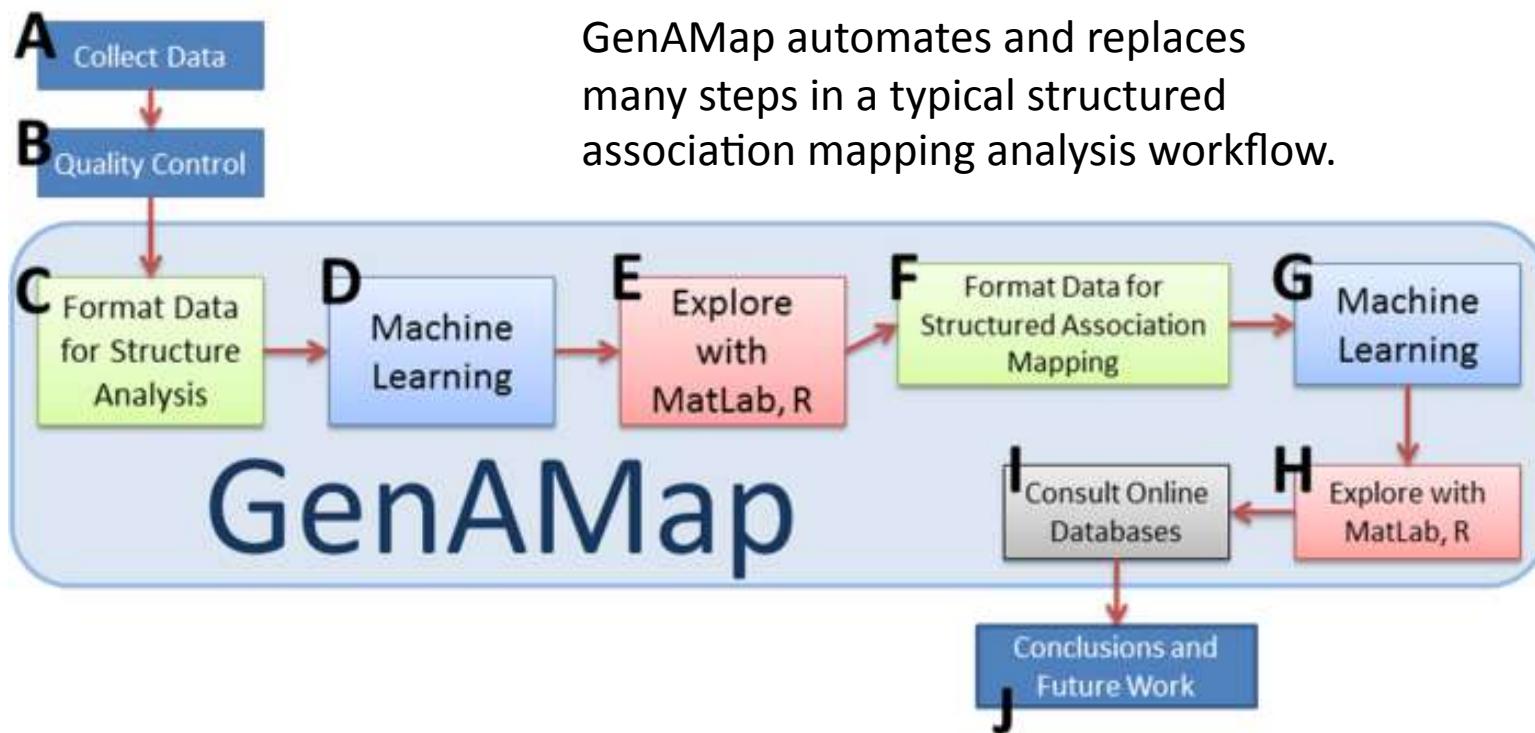


eQTL Explorer and eQTL Viewer

- Features – can explore eQTL results, link out to external information

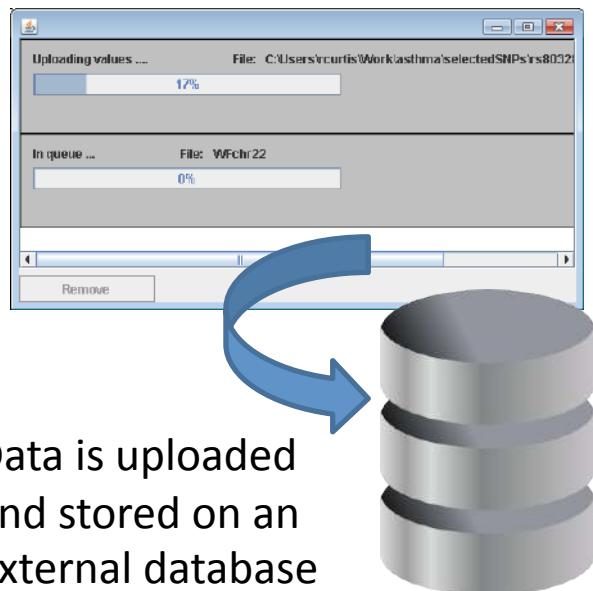


GenAMap: Structured Association Mapping Workflow



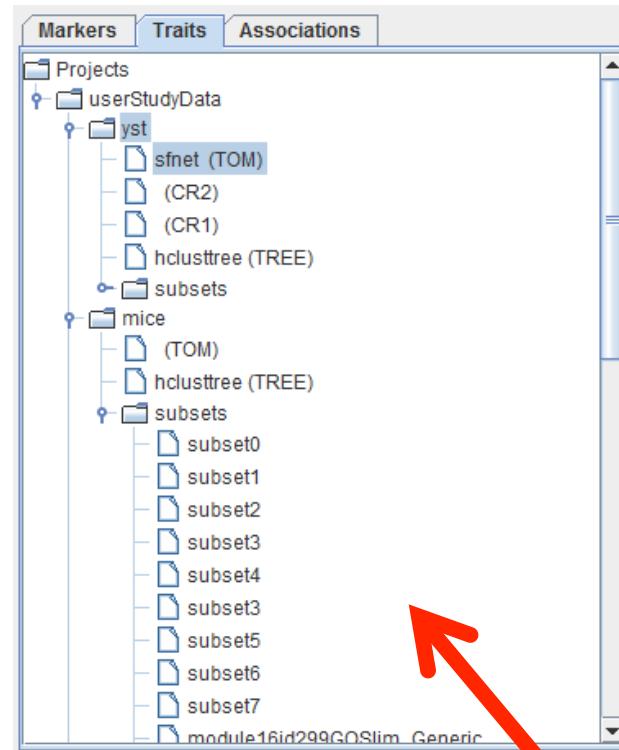
GenAMap: Data Management

GenAMap organizes genome, phenome, and association data with the corresponding structures.

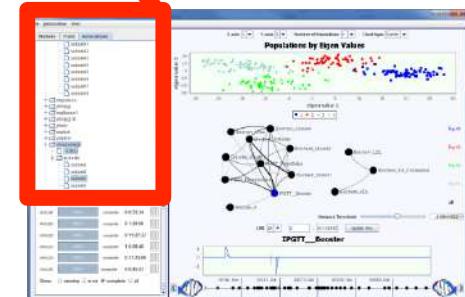


Data is uploaded and stored on an external database for fast processing and collaborative work.

ISMB 2011 Tutorial
© 2011 Eric Xing & Seyoung Kim, CMU



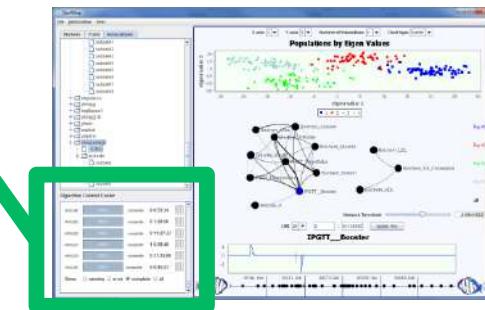
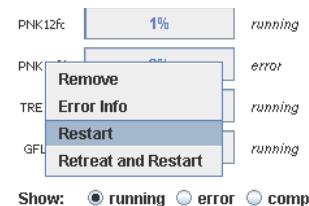
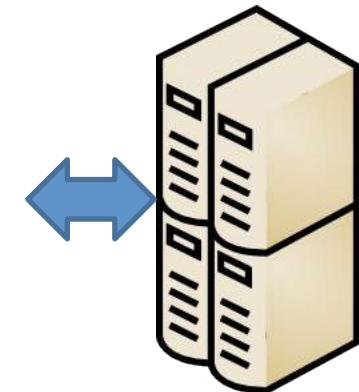
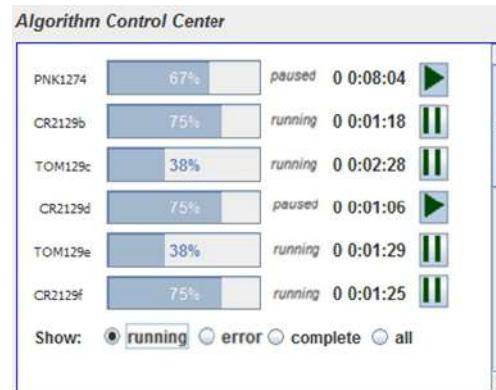
GenAMap supports multiple file formats: delimited text, .PED, and .BED files



GenAMap: Algorithm Automation

- Association Algorithms
 - PLINK
 - Lasso
 - Wilcoxon-Sum Rank
- Algorithms to create structure
 - Population structure (structure)
 - Network structure (glasso, correlation, scale-free network learning)
 - Tree structure (hierarchical clustering)
- Structured Association Algorithms
 - TreeLasso
 - GFlasso
 - MPGL

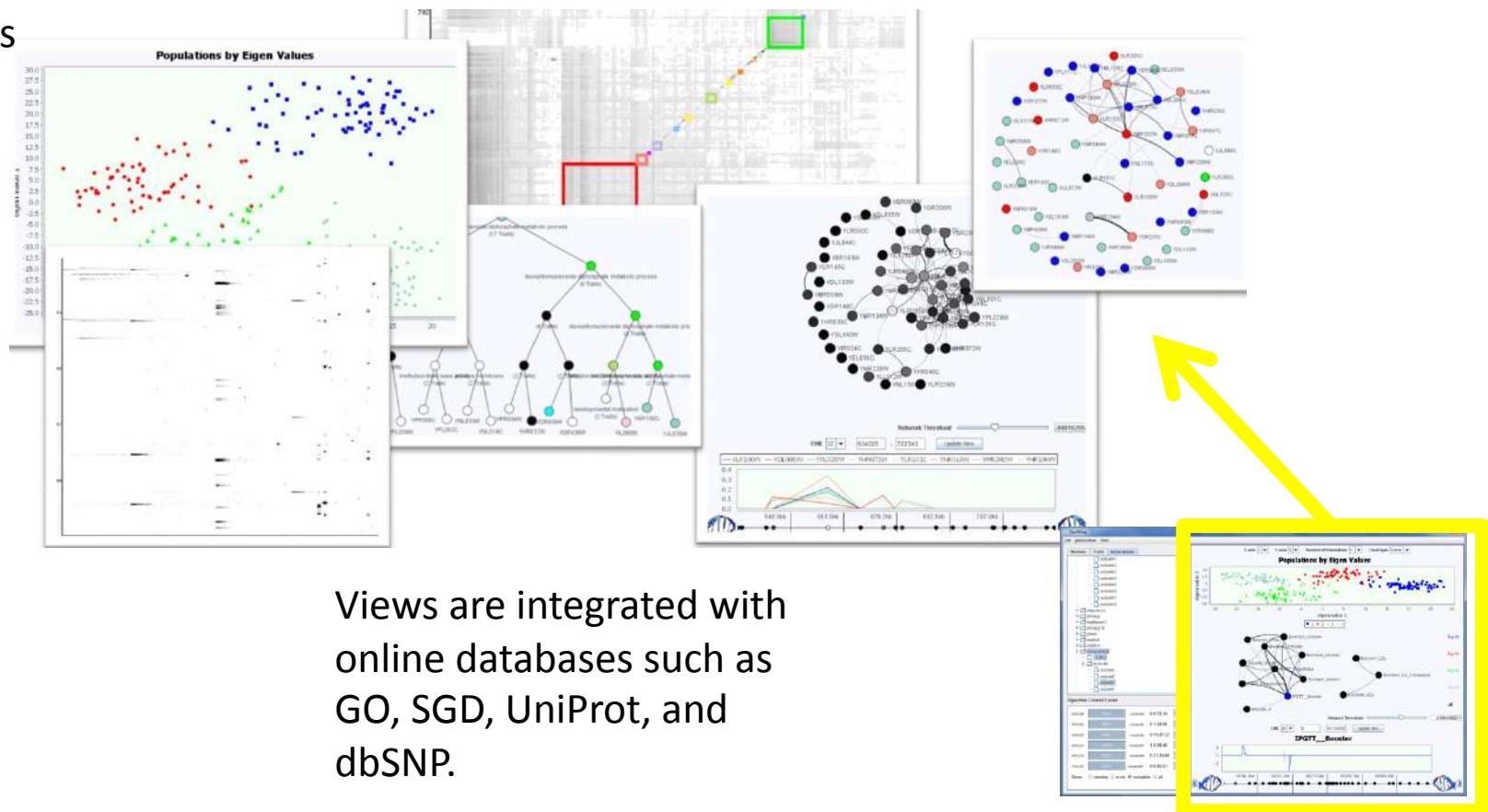
Algorithms are automatically run on a cluster in parallel.



User monitors and interacts with algorithms through GenAMap.

GenAMap: Visualization

Multiple coordinated views allow for exploration of data structure and association signals



GenAMap: User Study

- Preliminary qualitative user study
 - 8 participants, experts in genetics
 - Semi-structured tasks to explore GenAMap
- Survey results
 - GenAMap is better than other tools that explore association results: average score 5.0/5
 - GenAMap led to insight not available using other tools: average score 4.71/5
 - Would recommend GenAMap to other researchers: average score 4.75/5
- Comments
 - 6 users commented that GenAMap was more convenient than using MATLAB
 - 5 users specifically mentioned that using GenAMap was easier, more convenient, and saved time over how they normally do association analysis.
 - *“By myself, I would have to go back and forth between tools. It is really nice that this is integrated into this software.”*
 - *“The ability to interact with the network and the genome is excellent.”*
 - *“GenAMap’s visualizations really put the associations into perspective.”*

Part IV

Association Analysis and Next Generation Sequencing



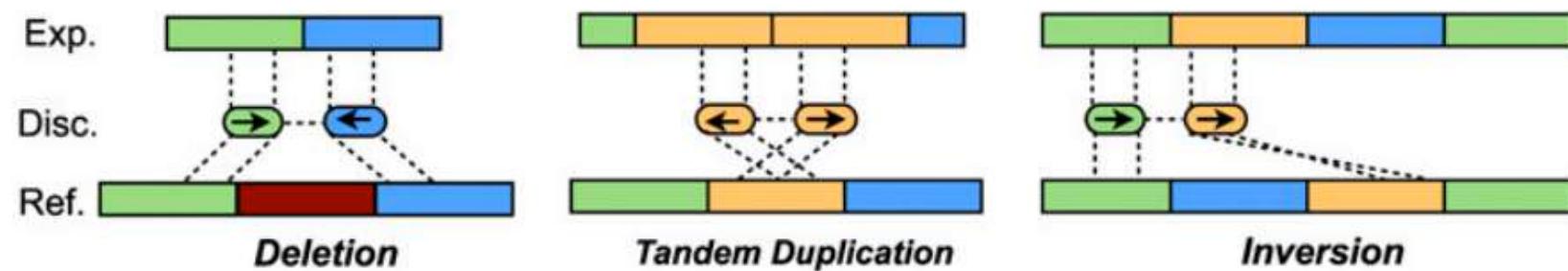
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Explaining Missing Heritability

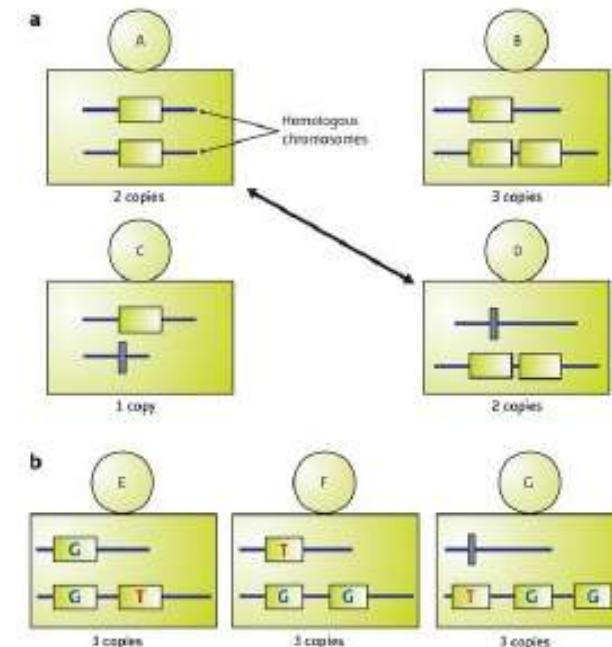
- Undetected epistasis
- Going beyond SNPs
 - Structural variants: insertions, deletions, duplication, copy number variants
- From common variants to rare variants

Other Genetic Variations



Other Genetic Variations

- Copy Number Variation
 - DNA segment whose numbers differ in different genomes
 - Kilobases to megabases in size
 - Usually two copies of all autosomal regions, one per chromosome
 - Variation due to deletion or duplication



Copy-number variation (CNV) can occur in ambiguous patterns. (a) Individuals in a population may have different copy numbers on homologous chromosomes at CNV loci. For example, here individual A and D have two copies, although the patterns are different: A has one copy on each chromosome, whereas D has two on one chromosome and zero on the other. (b) Individuals may also have CNVs that contain SNPs. For example, individuals E, F, and G each have three copies, but the patterns can be distinguished by the numbers of copies on each chromosome and variations defined by SNPs.

Common Variants vs. Rare Variants

- First-generation genome-wide association study (GWAS): common variant common disease hypothesis
- Common variants with minor allele frequency (MAF)>5%
 - dbGap: ~11 million SNPs
 - HapMap: 3.5 million SNPs
 - A successful GWAS requires a more complete catalogue of genetic variations
- Rare variants (MAF<0.5%), low-frequency variants (MAF:0.5%~5%)
 - Captured by sequencing with next-generation sequencing technology
 - Possibly significant contributors to the genetic architecture of disease
 - Causal variants are subject to negative selection

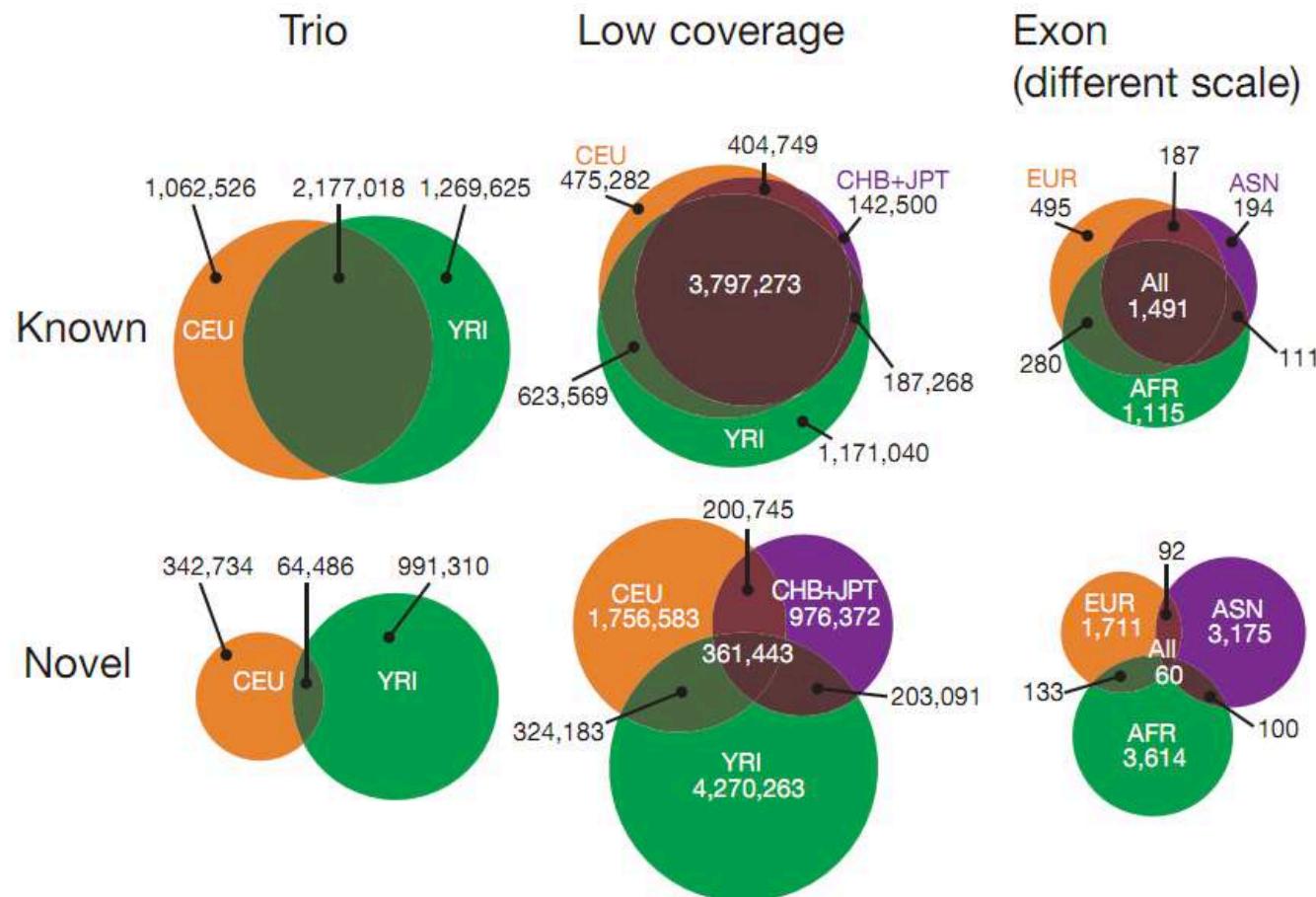
1000 Genome Project

(The 1000 Genome Project Consortium, Nature 2010)

The **goal** is to characterize over **95% of variants** that are in genomic regions accessible to current high-throughput sequencing technologies and that have **allele frequency of 1% or higher** (the classical definition of polymorphism) in each of **five major population groups** (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas)

- 179 individuals (low coverage)
- 6 individuals in two trios (deep sequencing)
- 697 individuals (exon sequencing of 8,140 exons)

1000 Genome Projects: Known vs. Novel Variants



Associations to Rare Variants

- Often GWA studies are underpowered for functional rare variants

Common Variant Association

	Case	Control
Allele a	60	20
Allele A	40	80

Rare Variant Association

	Case	Control
Allele a	7	2
Allele A	93	98

- Common variant GWA approaches are appropriate only for common variants

Recent Methods for Detecting Rare Variant Associations

- Test **combined effect** of multiple rare variants
- Fixed-Threshold Approach (Li & Leal, AJHG 2008)
 - Include only the SNPs with allele frequency below a fixed threshold
 - For SNP $i=1,\dots,m$ in a genomic region or a gene,

$$Score = \sum_{i=1}^m \epsilon_i C_i$$

» C_i : The allele frequency of SNP i in cases

$$\epsilon_i = \begin{cases} 1 & \text{If the allele frequency of SNP } i \text{ is below a specified threshold} \\ 0 & \text{Otherwise} \end{cases}$$

- Evaluate the significance of $Score$ by permutation test

Recent Methods for Detecting Rare Variant Associations

- Weighted Approach (Madson & Browning, PLoS Genetics 2009)
 - Instead of fixed threshold, weight each SNP with the inverse square root of the variance at each SNP locus
 - For SNP $i=1,\dots,m$ in a genomic region or a gene,

$$Score = \sum_{i=1}^m \mathcal{E}_i C_i$$

» C_i : The allele frequency of SNP i in cases

$$\mathcal{E}_i = 1/\sqrt{p_i(1 - p_i)} \quad p_i : \text{allele frequency}$$

- Evaluate the significance of $Score$ by permutation test

Recent Methods for Detecting Rare Variant Associations

- Incorporating computational predictions of functional effects
(Price et al., AJHG 2010)
 - PolyPhen-2 : a probabilistic scoring system for neutral and functional amino acids
 - For SNP $i=1,\dots,m$ in a genomic region or a gene,

$$Score = \sum_{i=1}^m \mathcal{E}_i C_i$$

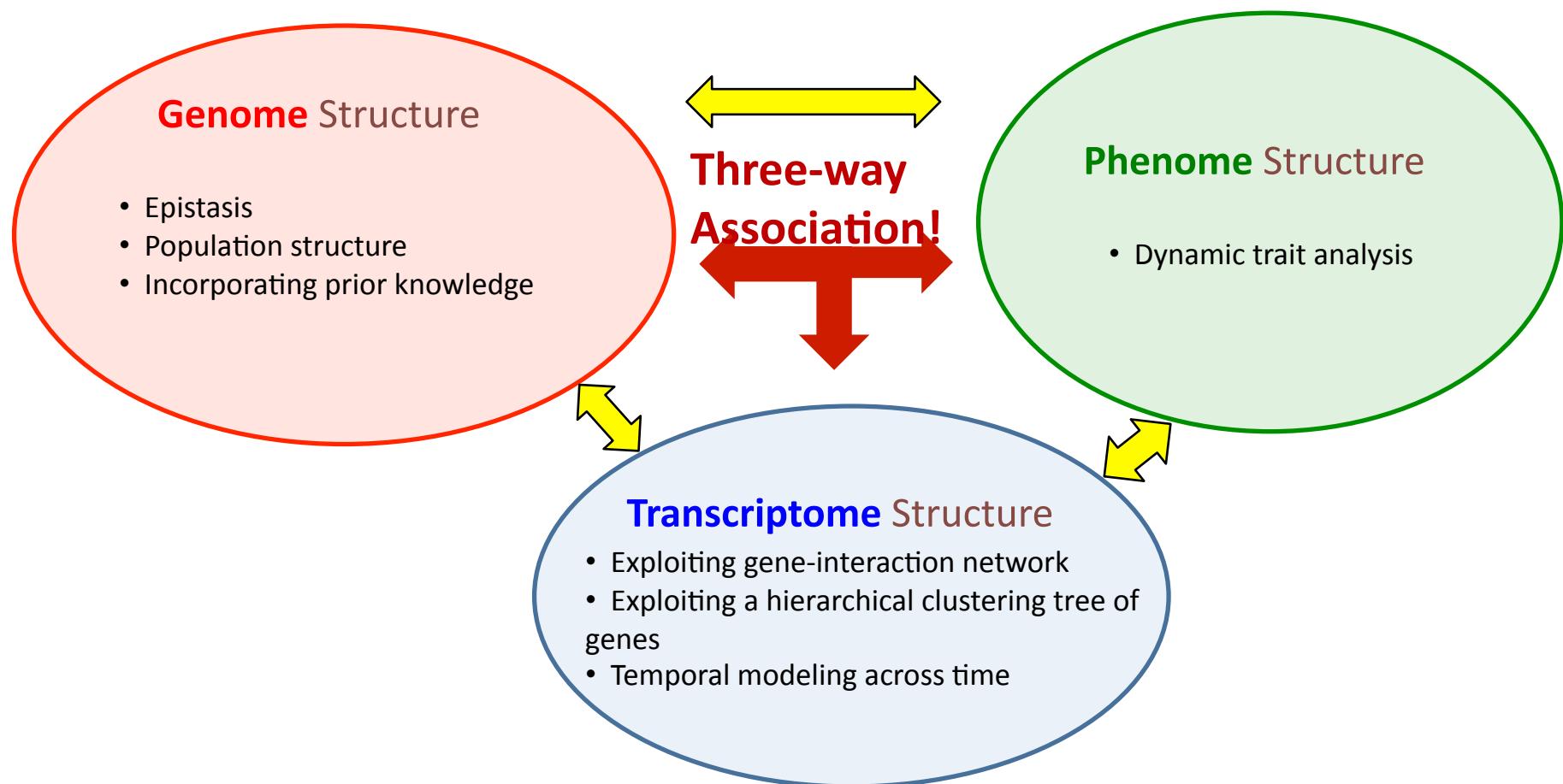
» C_i : The allele frequency of SNP i in cases

» $\mathcal{E}_i = P(S_i)$

$P(S_i)$: score for SNP i being functional, given PolyPhen-2 score S_i

- Evaluate the significance of $Score$ by permutation test

Summary: Structured Genome-Transcriptome-Phenome Association Analysis



Publicly Available Software

- Haplotyper: <http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm>
- PHASE, FastPHASE: <http://stephenslab.uchicago.edu/software.html>
- Structure: <http://pritch.bsd.uchicago.edu/structure.html>
- EigenStrat: <http://genepath.med.harvard.edu/~reich/Software.htm>
- mStruct: <http://cogito-b.ml.cmu.edu/mstruct/>
- Plink: <http://pngu.mgh.harvard.edu/~purcell/plink/>
- Lasso: <http://cran.r-project.org/web/packages/glmnet/index.html>
- Strat: <http://pritch.bsd.uchicago.edu/software/STRAT.html>
- Genomic Control: <http://www.wpic.pitt.edu/wpiccompgen/software.htm>
- Gflasso: <http://cogito-b.ml.cmu.edu/gflasso/>
- Structured IO Lasso: <http://cogito-b.ml.cmu.edu/epistaticQTL/>
- Lirnet: <http://www.cs.washington.edu/homes/suinlee/lirnet/>
- GenAMap: <http://cogito-b.ml.cmu.edu/genamap/>

Acknowledgements

- Sailing Lab
 - Ross Curtis
 - Seunghak Lee
 - Suyash Shringapure
 - Judie Howrylak
 - Kyung-Ah Sohn
 - Kriti Puniyani
- University of Pittsburgh Medical Center
 - Sally Wenzel (MD)
- Harvard Medical School
 - Scott Weiss (MD)
 - Benjamin Raby (MD)

Acknowledgement



SAILING LAB



Laboratory for Statistical Artificial Intelligence & INtegrative Genomics

Postdocs



Jacob Eisenstein Seyoung Kim Le Song (Andrew) Xu Wei
PhD, MIT PhD, UC Irvine PhD, USydney PhD, U Ottawa

PhD students



Jun Zhu Amr Ahmed Ross Curtis Judie Howrylik Hetunandan K. Mladen Kolar
PhD, Tsinghua LTI Comp Bio Comp Bio CSD MLD

www.sailing.cs.cmu.edu

\$\$\$:



Jing Xiang

MLD

André Martins

LTI / UT Lisboa

Kriti Puniyani

LTI

Gunhee Kim

CSD

Suyash Shringarpure

MLD

Kyung-Ah Sohn

CSD

Anuj Goyal

MS, LTI

Qirong Ho

MLD

Seunghak Lee

CSD

Ankur Pankh

MLD

Bin Zhao

MLD

Matt Wytock

MLD

References

- [1] D.J. Balding, "A tutorial on statistical methods for population association studies." *Nature reviews. Genetics*, vol. 7, Oct. 2006, pp. 781-91.
- [2] R.B. Brem and L. Kruglyak, "The landscape of genetic complexity across 5,700 gene expression traits in yeast." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, Feb. 2005, pp. 1572-7.
- [3] O. Carlborg and C.S. Haley, "Epistasis: too often neglected in complex trait studies?" *Nature Reviews Genetics*, vol. 5, Aug. 2004, pp. 618-25.
- [4] C.S. Carlson, M. a Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D. a Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." *American Journal of Human Genetics*, vol. 74, Jan. 2004, pp. 106-20.
- [5] L.L. Cavalli-Sforza, "The Human Genome Diversity Project: past, present and future." *Nature Reviews Genetics*, vol. 6, Apr. 2005, pp. 333-40.
- [6] G. Clark, "Inference of haplotypes from PCR-amplified samples of diploid populations." *Molecular Biology and Evolution*, vol. 7, Mar. 1990, pp. 111-22.
- [7] G. Control, "Genomic Control to the extreme To the editor :" *Nature Genetics*, vol. 36, 2004, pp. 1129-1131.
- [8] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, C.S. Sevier, H. Ding, J.L.Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R.P. St Onge, B. VanderSluis, T. Makhnevych, F.J. Vizeacoumar, S. Alizadeh, S. Bahr, R.L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. San Luis, E. Shuteriqi, A.H.Y. Tong, N. van Dyk, I.M. Wallace, J. a Whitney, M.T. Weirauch, G. Zhong, H. Zhu, W. a Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pál, F.P. Roth, G. Giaever, C. Nislow, O.G. Troyanskaya, H. Bussey, G.D. Bader, A.-C. Gingras, Q.D. Morris, P.M. Kim, C. a Kaiser, C.L. Myers, B.J. Andrews, and C. Boone, "The genetic landscape of a cell." *Science (New York, N.Y.)*, vol. 327, Jan. 2010, pp. 425-31.
- [9] B. Devlin, K. Roeder, and L. Wasserman, "Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing." *Biostatistics (Oxford, England)*, vol. 1, Dec. 2000, pp. 369-87.
- [10] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Applied Statistics*, vol. 1, Dec. 2007, pp. 302-332.

- [11] S.-I. Lee, A.M. Dudley, D. Drubin, P. a Silver, N.J. Krogan, D. Pe'er, and D. Koller, "Learning a prior on regulatory potential from eQTL data.", *PLoS Genetics*, vol. 5, Jan. 2009, e1000358.
- [12] B. Li and S.M. Leal, "Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data," *Journal of Human Genetics*, 2008, pp. 311-321.
- [13] T. Niu, Z.S. Qin, X. Xu, and J.S. Liu, "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.", *American Journal of Human Genetics*, vol. 70, Jan. 2002, pp. 157-69.
- [14] N. Patil, a J. Berno, D. a Hinds, W. a Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K. a Frazer, S.P. Fodor, and D.R. Cox, "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.", *Science (New York, N.Y.)*, vol. 294, Nov. 2001, pp. 1719-23.
- [15] N. Patterson, A.L. Price, and D. Reich, "Population structure and eigenanalysis." *PLoS Genetics*, vol. 2, Dec. 2006, e190.
- [16] A.L. Price, G.V. Kryukov, P.I.W. de Bakker, S.M. Purcell, J. Staples, L.-J. Wei, and S.R. Sunyaev, "Pooled Association Tests for Rare Variants in Exon-Resequencing Studies.", *American Journal of Human Genetics*, vol. 86, May. 2010, pp. 832-838.
- [17] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N. a Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies.", *Nature Genetics*, vol. 38, Aug. 2006, pp. 904-9.
- [18] J.K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data.", *Genetics*, vol. 155, Jun. 2000, pp. 945-59.
- [19] J.K. Pritchard, M. Stephens, N. a Rosenberg, and P. Donnelly, "Association mapping in structured populations.", *American Journal of Human Genetics*, vol. 67, Jul. 2000, pp. 170-81.
- [20] A.R. Quinlan, R. a Clark, S. Sokolova, M.L. Leibowitz, Y. Zhang, M.E. Hurles, J.C. Mell, and I.M. Hall, "Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome.", *Genome research*, vol. 20, May. 2010, pp. 623-35.
- [21] N. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L. a Zhivotovsky, and M.W. Feldman, "Genetic structure of human populations.", *Science (New York, N.Y.)*, vol. 298, Dec. 2002, pp. 2381-5.

- [22] S. Shringarpure and E.P. Xing, "mStruct: Inference of Population Structure in Light of Both Genetic Admixing and Allele Mutations," *Genetics*, vol. 593, 2009, pp. 575-593.
- [23] M. Stephens, N.J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, vol. 68, Apr. 2001, pp. 978-89.
- [24] C.J. Vaske, S.C. Benz, J.Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J.M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, Jun. 2010, pp. 237-245.
- [25] K. Wang, M. Li, and H. Hakonarson, "Analysing biological pathways in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, Dec. 2010, pp. 843-854.
- [26] R. Wu and M. Lin, "Functional mapping - how to map and study the genetic architecture of dynamic complex traits," *Nature Reviews Genetics*, vol. 7, Mar. 2006, pp. 229-37.
- [27] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics (Oxford, England)*, vol. 25, Mar. 2009, pp. 714-21.
- [28] N. Zaitlen, H.M. Kang, E. Eskin, and E. Halperin, "Leveraging the HapMap correlation structure in association studies," *American Journal of Human Genetics*, vol. 80, Apr. 2007, pp. 683-91.
- [29] K. Zhang, M. Deng, T. Chen, M.S. Waterman, and F. Sun, "A dynamic programming algorithm for haplotype block partitioning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, May. 2002, pp. 7335-9.
- [30] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner, and E.E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature Genetics*, vol. 40, Jul. 2008, pp. 854-61.
- [31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, Apr. 2005, pp. 301-320.
- [32] The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, Oct. 2005, pp. 1299-320.
- [33] B. Devlin and K. Roeder, "Genomic Control for Association Studies," *Biometrics*, vol. 55, Dec. 1999, pp. 997-1004.

- [34] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, Jun. 2007, pp. 661-78.
- [35] A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza, "High resolution of human evolutionary trees with polymorphic microsatellites," *Nature*, vol. 368, Mar. 1994, pp. 455-457.
- [36] T. Strachan and A. Read, "Human Molecular Genetics", Garland Science, 2nd Edition, 2001.
- [37] S. Boyd and L. Vandenberghe, "Convex Optimization", Cambridge University Press, 2004.
- [38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Royal Statistical Society, Series B*, vol. 58, 1996, pp. 267–288.
- [39] J. Weller, G. Wiggins, P. Vanraden, and M. Ron, "Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment," *Theoretical and Applied Genetics*, vol. 92, 1996, pp. 998–1002.
- [40] B. Mangin, B. Thoquet, and N. Grimsley, "Pleiotropic QTL analysis," *Biometrics*, vol. 54, 1998, pp. 89–99.
- [41] Y. Chen, J. Zhu, P. Lum, X.Yang, S. Pinto, et al., "Variations in DNA elucidate molecular networks that cause disease," *Nature*, vol. 452, 2008, pp. 429–35.
- [42] V. Emilsson, G. Thorleifsson, B. Zhang, A. Leonardson, F. Zink, et al., "Genetics of gene expression and its effect on disease," *Nature*, vol. 452, 2008, pp. 423–28.
- [43] S. Kim and E.P. Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLoS Genetics*, vol. 5, 2009, e1000587.
- [44] S. Kim and E.P. Xing, "Sparse feature learning in high-dimensional space via block regularized regression," In *Proceedings of the 24th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008, pp. 325-332. AUAI Press.
- [45] S. Kim and E.P. Xing, "Exploiting a hierarchical clustering tree of gene-expression traits in eQTL analysis," In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.
- [46] S. Kim, J. Howrylak, B. Raby, S. Weiss, and E.P. Xing, "Dynamic-trait association analysis via temporally-smoothed lasso," Submitted.
- [47] K. Punyani, S. Kim, and E.P. Xing, "Multi-population GWA mapping via multi-task regularized regression," *Bioinformatics*, vol. 26, 2010, pp. 208-216.

- [48] S. Lee, S. Kim, and E.P. Xing, "Leveraging genetic interaction networks and regulatory pathways for joint mapping of epistatic and marginal eQTLs," Submitted.
- [49] A.M. Dunning et al., "Association of ESR1 gene tagging SNPs with breast cancer risk," *Hum Mol Genet.* vol. 18, 2008, pp. 1131-9.
- [50] J. Esparza-Gordillo et al., "A common variant on chromosome 11q13 is associated with atopic dermatitis," *Nature Genetics*, vol. 41, 2009, pp. 596-601.
- [51] A. Suzuki et al., "Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population," *Nature Genetics*, vol. 40, 2008, pp. 1224-1229.
- [52] J. Dupuis et al., "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk," *Nature Genetics*, vol. 42, 2010, pp. 105-116.
- [53] M. Johannesson et al., "A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock." *Genome Res.* Vol. 19, Jan. 2009, pp. 150-8.
- [54] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E.P. Xing, "Smoothing Proximal Gradient Method for General Structured Sparse Regression," Submitted.