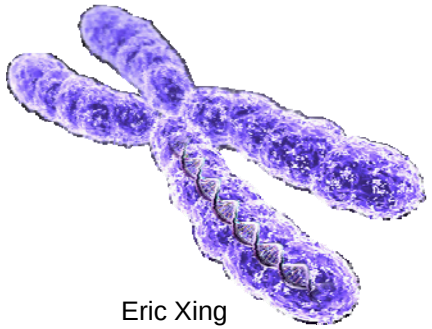
Machine Learning

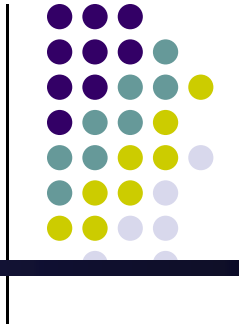
Application I:
Computational Genomics

Eric Xing



Lecture 18, August 16, 2010

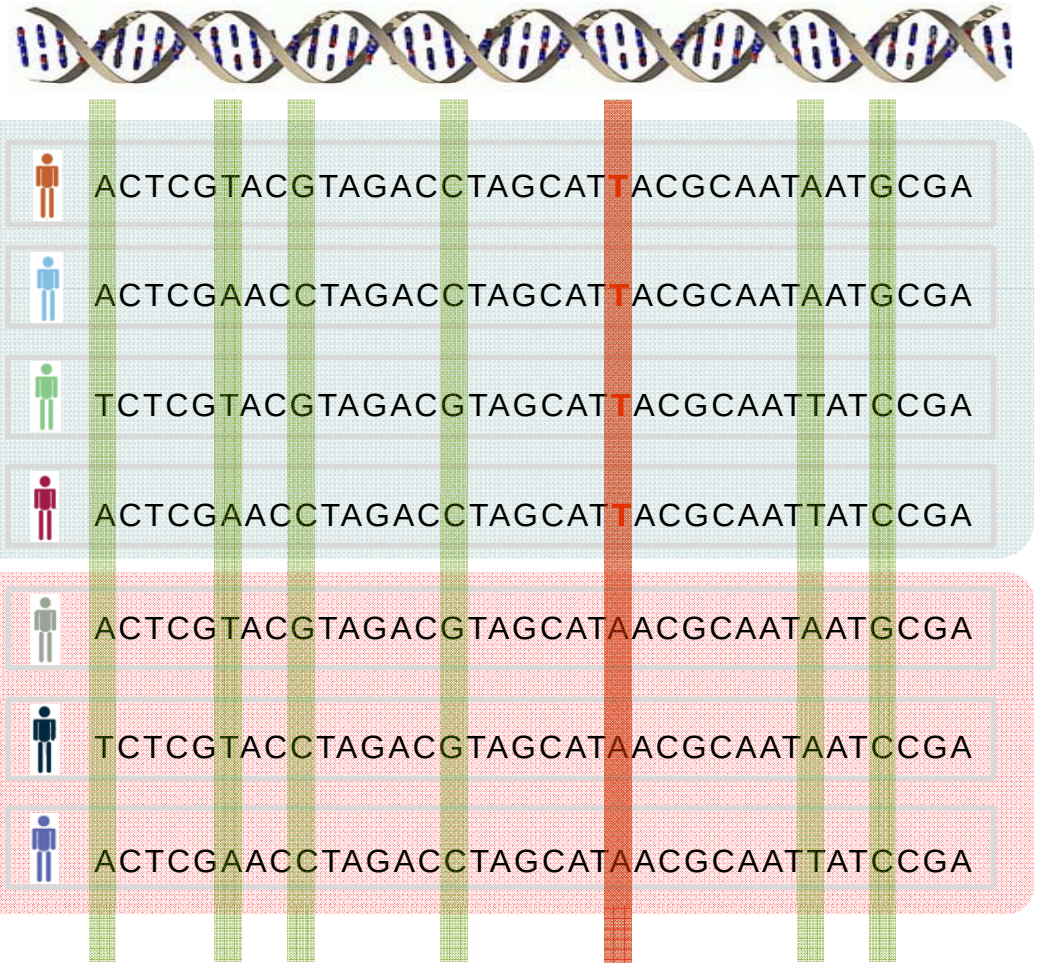
Biological Data Analysis

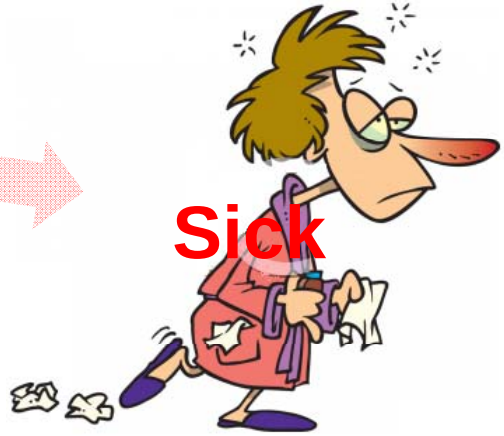


- Dynamic, noisy, heterogeneous, high-dimensional data

- "High-resolution" inference
- Parsimonious
- Scalability
- Stability
- Sample complexity
- Confidence bound

Genetic Basis of Diseases

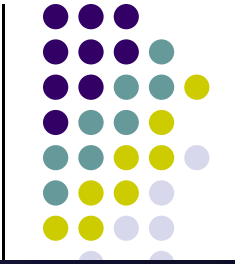




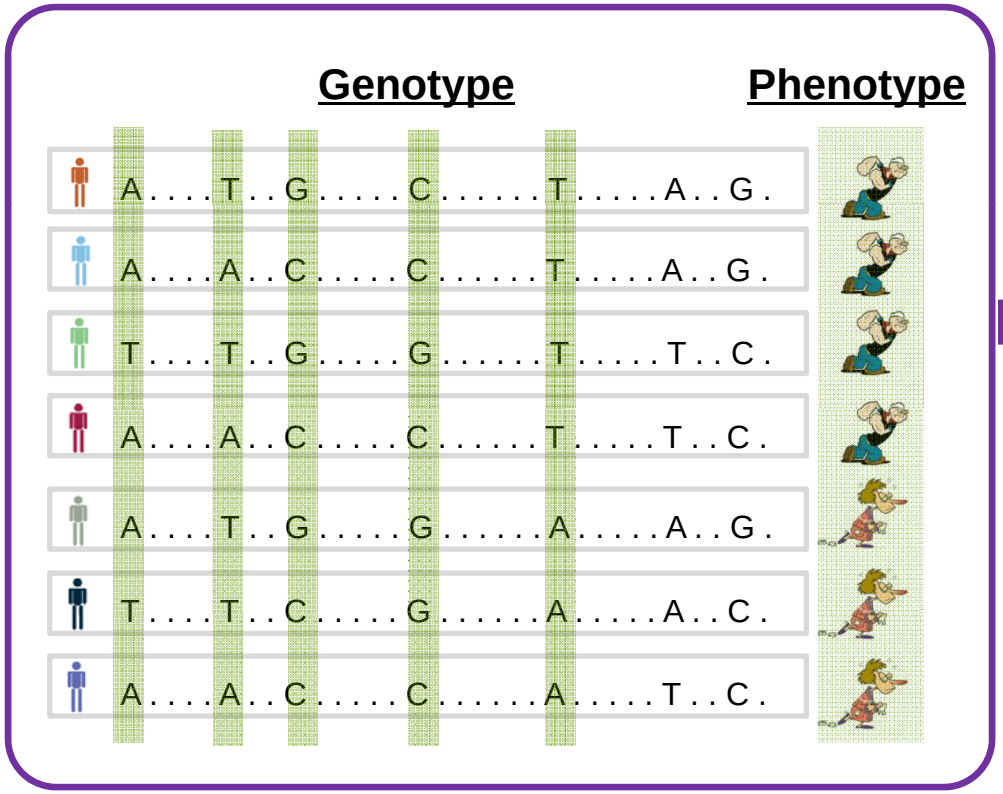
Single nucleotide polymorphism (SNP)

Causal (or "associated") SNP

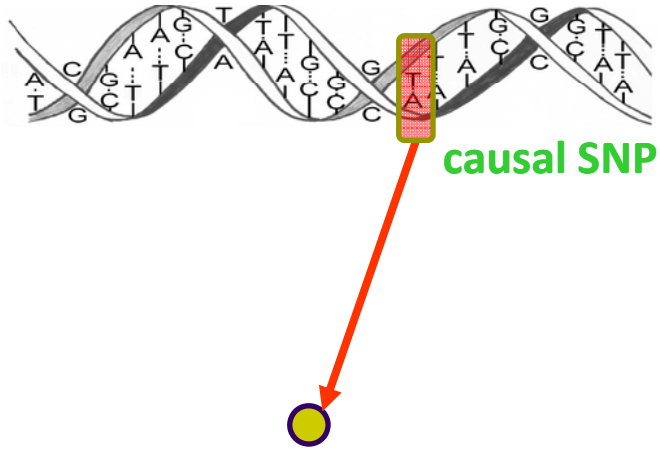
Genetic Association Mapping



Data



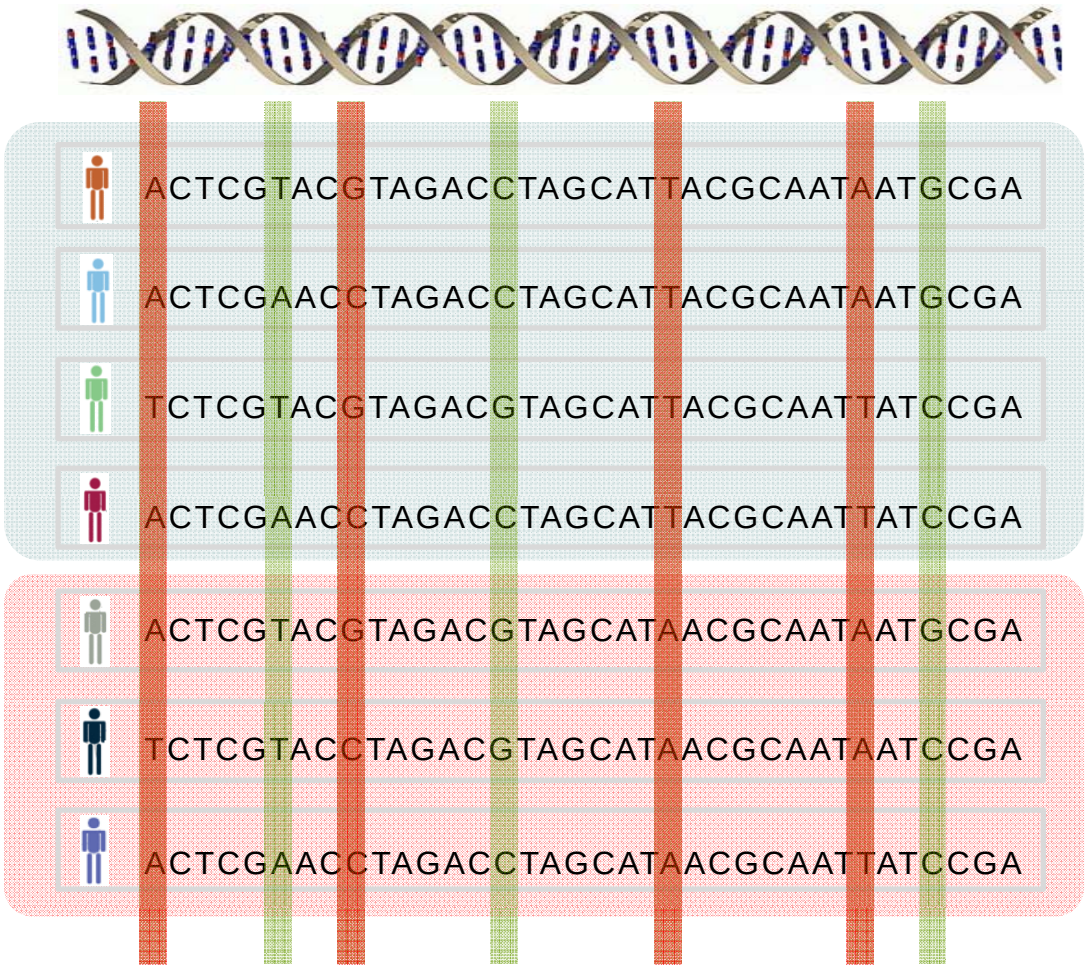
Standard Approach

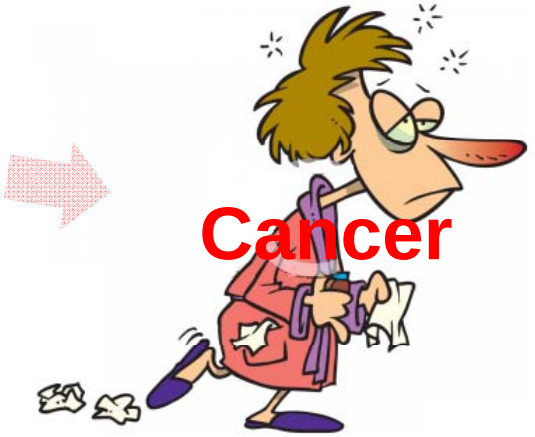


a univariate **phenotype**:
e.g., disease/control

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

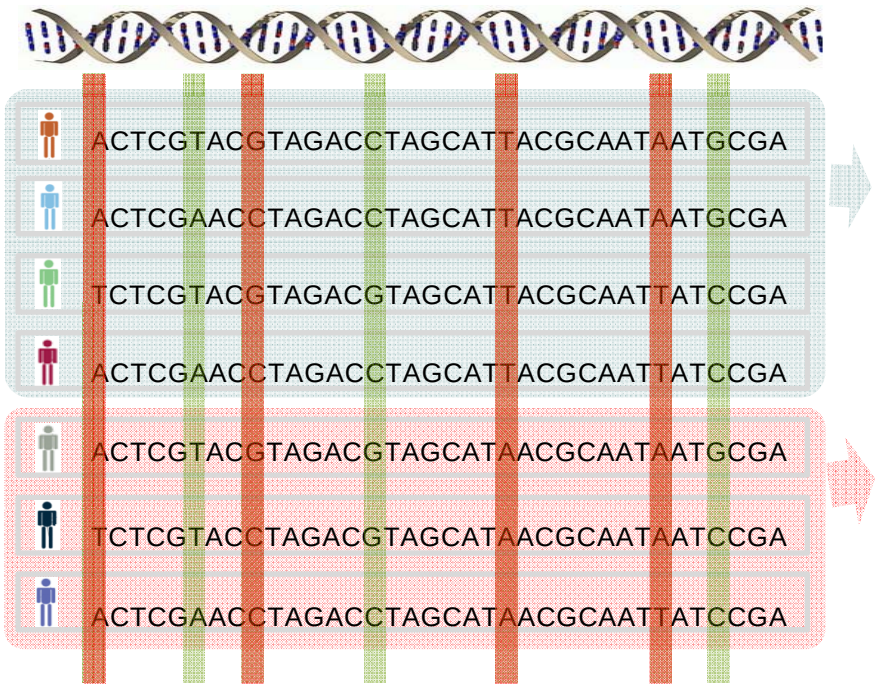
Genetic Basis of Complex Diseases



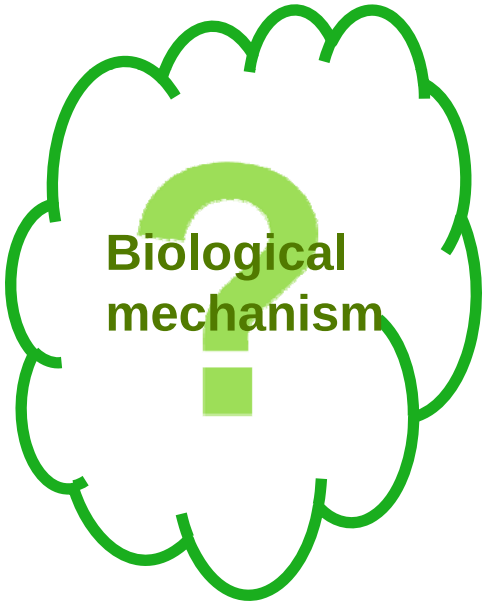


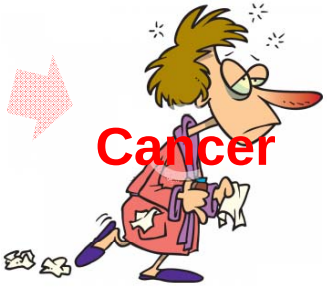
Causal SNPs

Genetic Basis of Complex Diseases



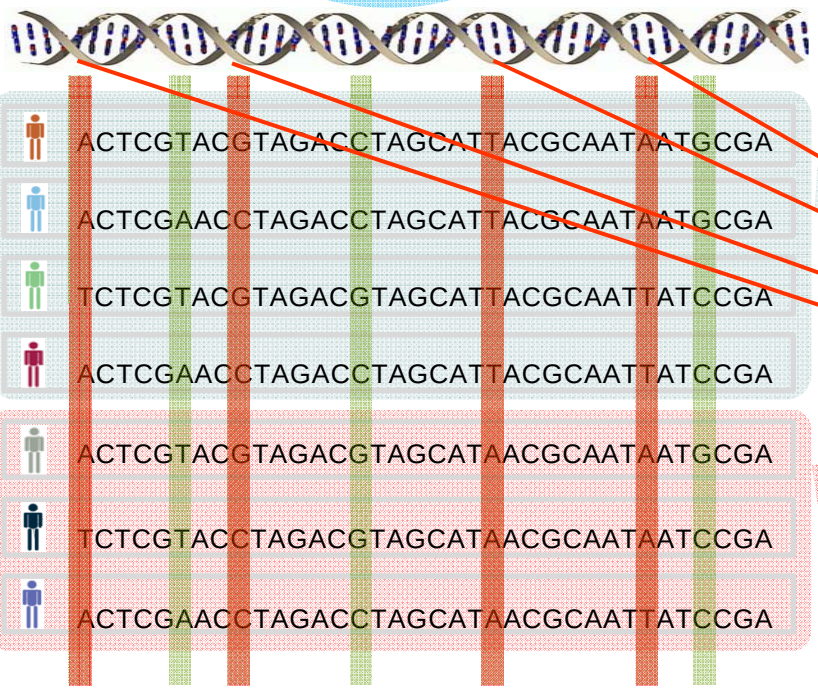
Causal SNPs





Genetic Basis of Complex Diseases

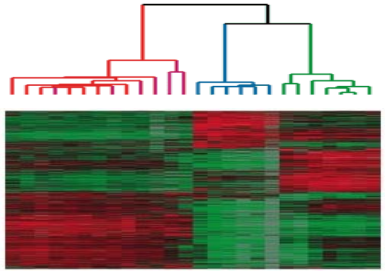
Association to intermediate phenotypes

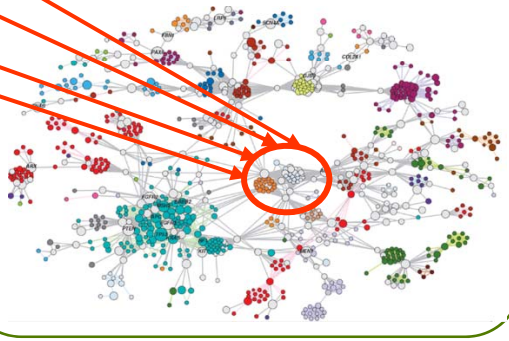


Causal SNPs

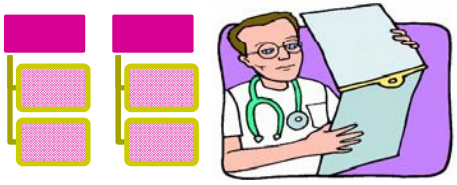
Intermediate Phenotype

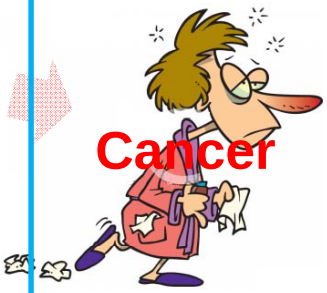
Gene expression



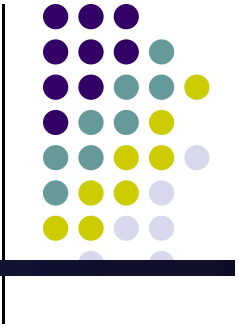


Clinical records





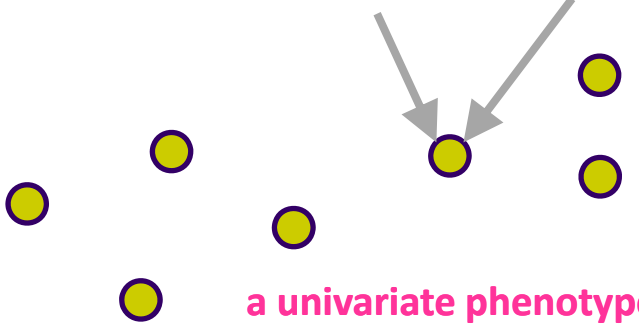
Structured Association



Traditional Approach

causal SNP

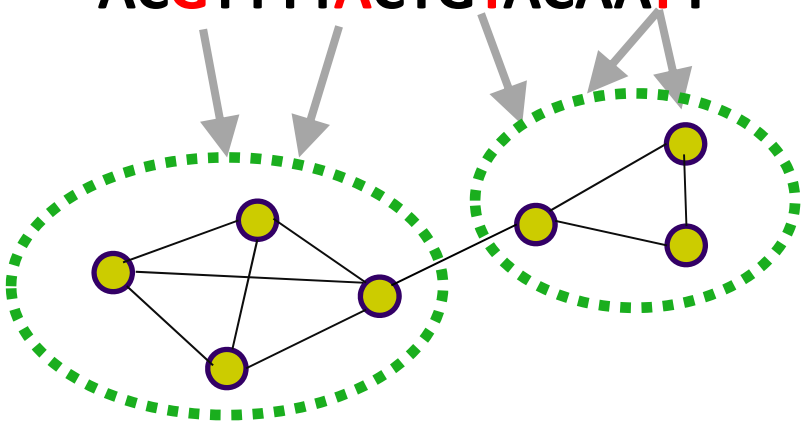
ACGTTTTACTGTACAATT



a univariate phenotype:
gene expression level

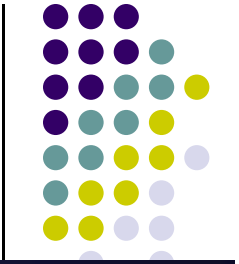
Association with Phenome

ACGTTTTACTGTACAATT



Multivariate complex syndrome (e.g., asthma)
age at onset, history of eczema
genome-wide expression profile

Structured Association : a New Paradigm



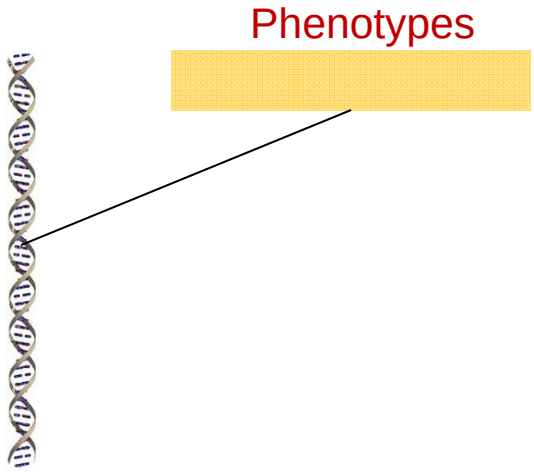
Standard Approach

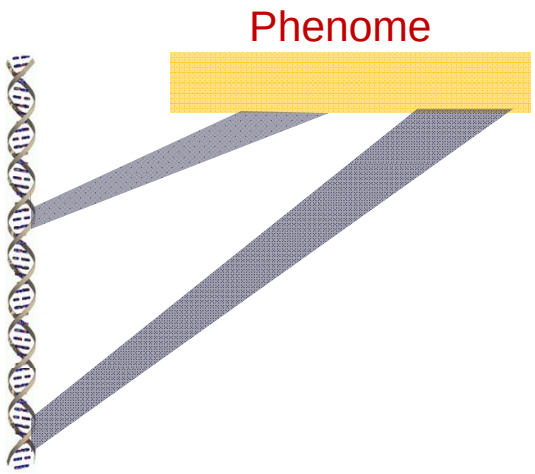
Consider **one phenotype** at a time

vs.

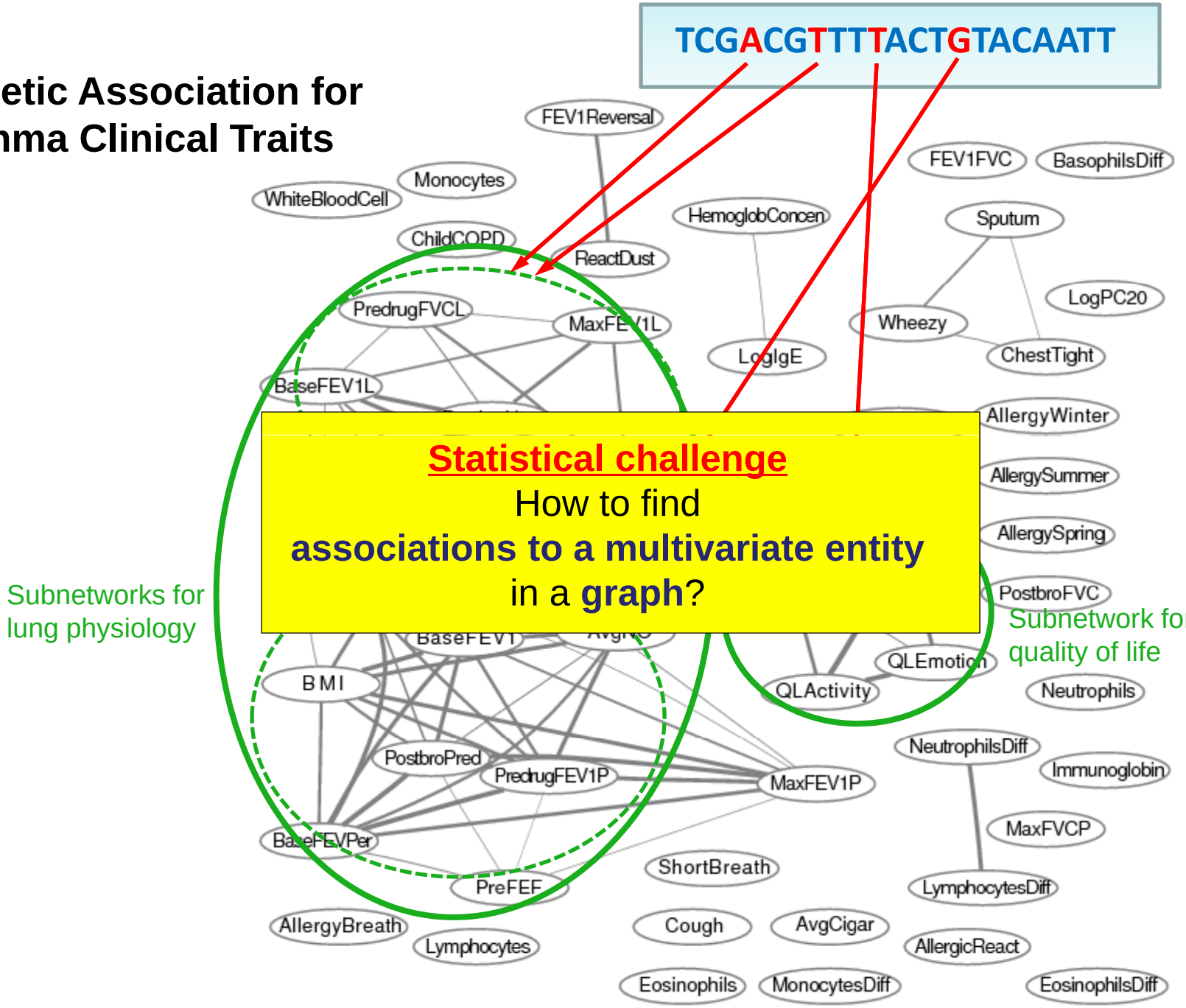
New Approach

Consider **multiple correlated phenotypes (phenome) jointly**

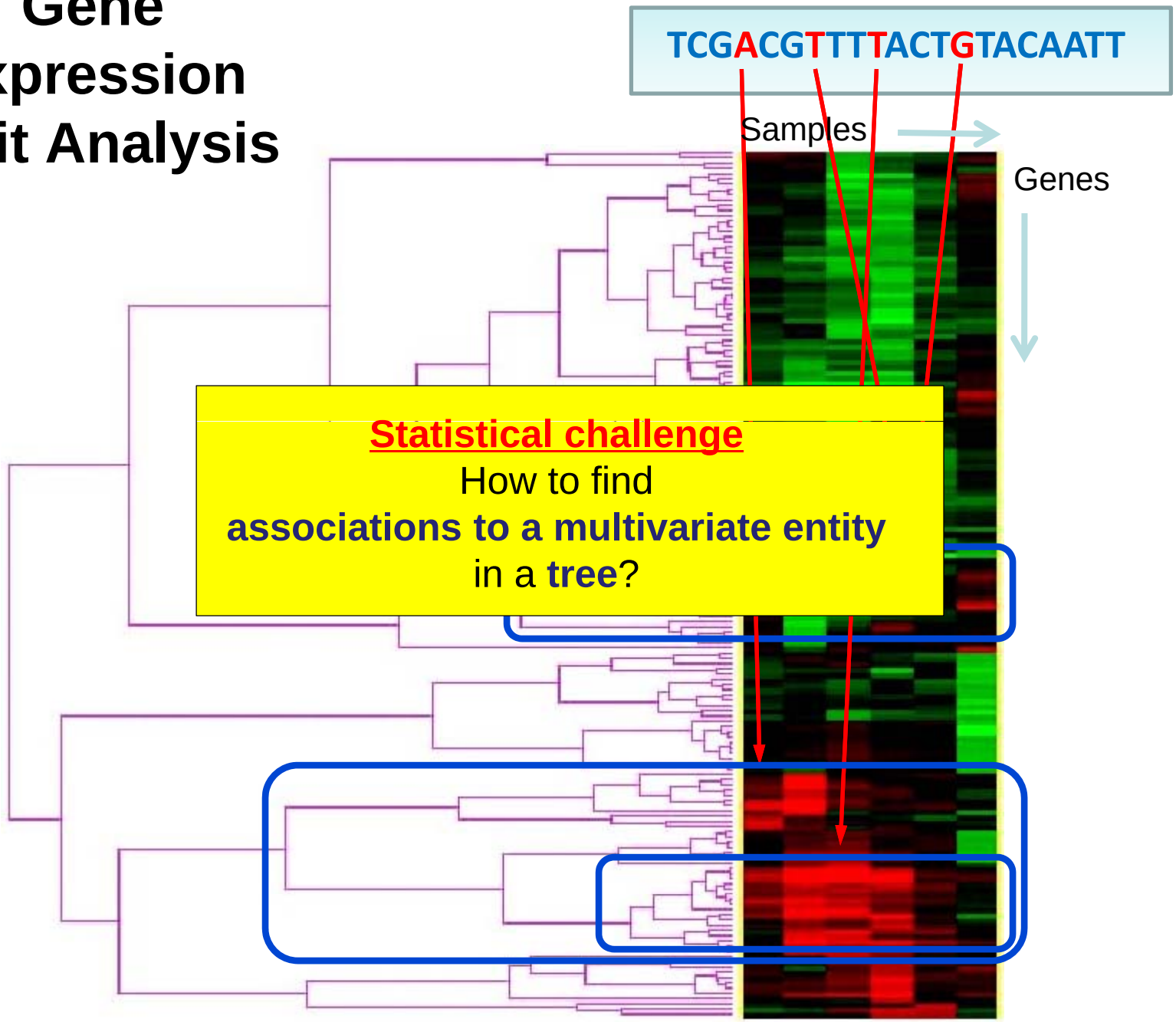




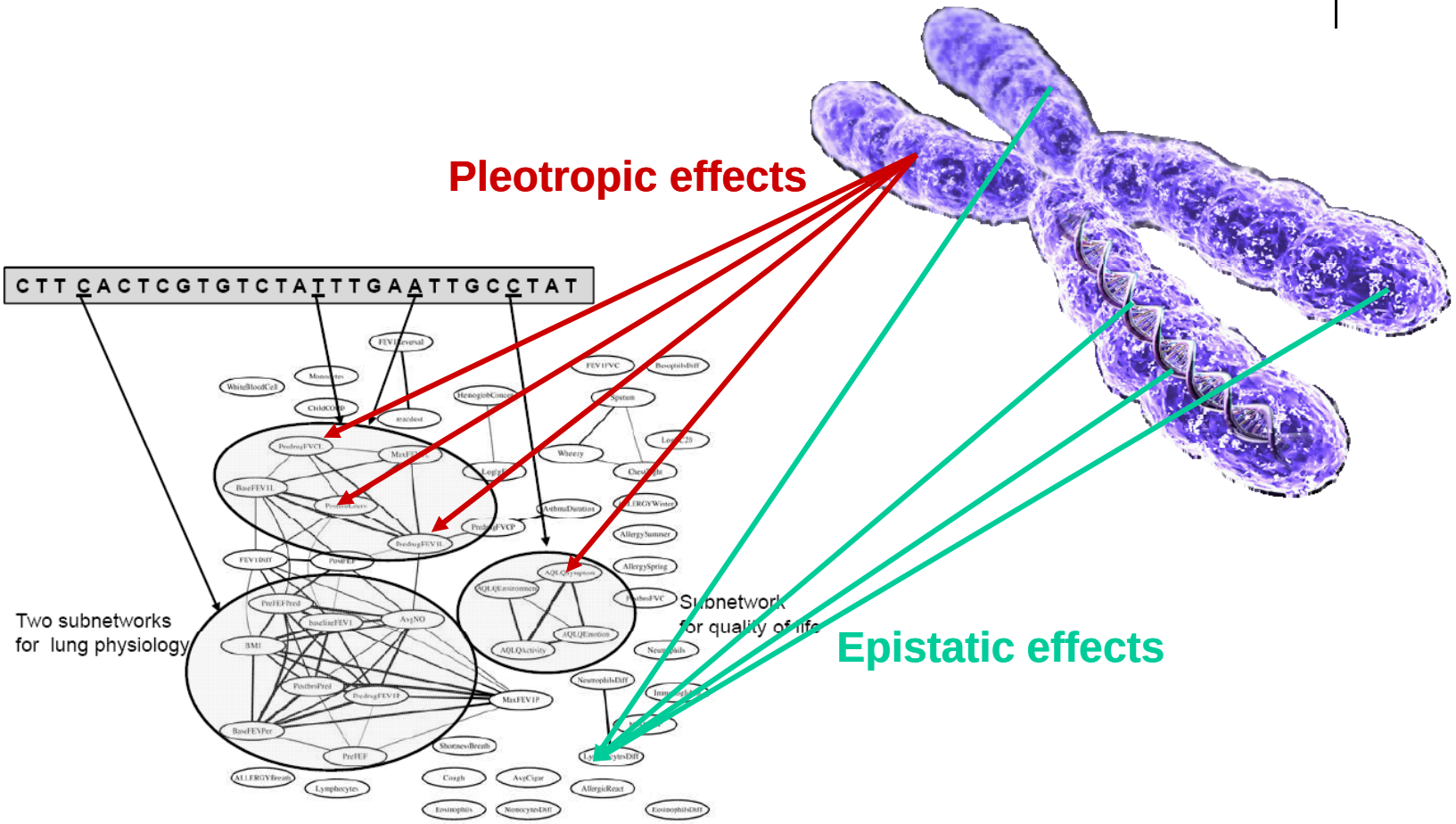
Genetic Association for Asthma Clinical Traits



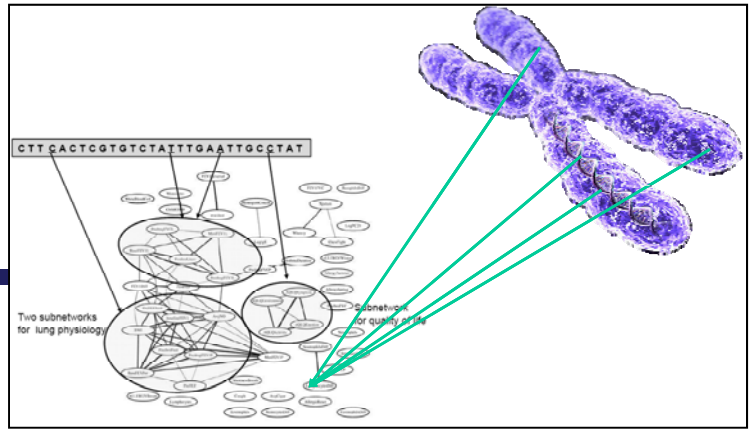
Gene Expression Trait Analysis



Sparse Associations



Sparse Learning



- Linear Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{y} \in \mathbb{R}^{N\times 1}, \quad \mathbf{X} \in \mathbb{R}^{N\times J}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{N\times N})$$

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j, \ldots, \beta_J)^T \in \mathbb{R}^J$$

- Lasso (Sparse Linear Regression) [R.Tibshirani 96]

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^J} f(\boldsymbol{\beta}) \equiv \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \Omega(\boldsymbol{\beta}) \quad \Omega(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$$

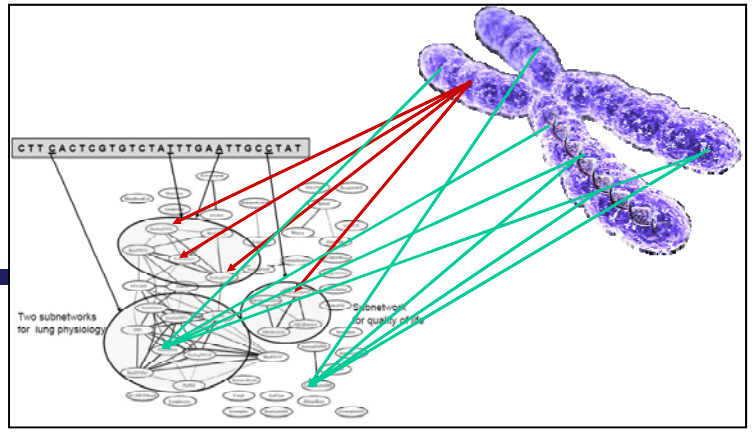
$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\beta_j|$$

- Why sparse solution?

penalizing $\lambda\|\boldsymbol{\beta}\|_1$

constraining $\|\boldsymbol{\beta}\|_1 \le \gamma$

Multi-Task Extension



- Multi-Task Linear Model:

Input: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J) \in \mathbb{R}^{N \times J}$

Output: $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_K) \in \mathbb{R}^{N \times K}$

$$\mathbf{y}_k = \mathbf{X}\boldsymbol{\beta}_k + \epsilon_k, \quad \forall k = 1, \ldots, K$$

Coefficients for *k-th* task: $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{Jk})^T \in \mathbb{R}^J$

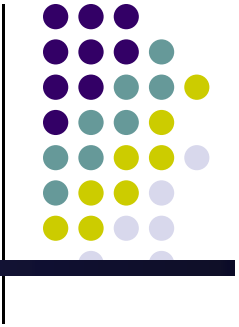
Coefficient Matrix: $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{J \times K}$

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1K} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{J1} & \beta_{J2} & \ldots & \beta_{JK} \end{pmatrix}$$

Coefficients for a variable (2nd)

Coefficients for a task (2nd)

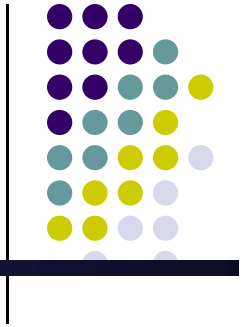
Outline



- Background: Sparse multivariate regression for disease association studies

- Structured association – a new paradigm
  - Association to a **graph**-structured phenome
    - Graph-guided fused lasso (Kim & Xing, PLoS Genetics, 2009)
  - Association to a **tree**-structured phenome
    - Tree-guided group lasso (Kim & Xing, ICML 2010)

Multivariate Regression for Single-Trait Association Analysis

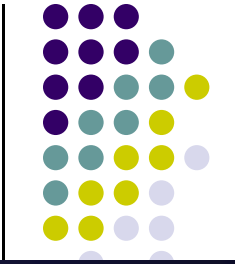


Trait | Genotype | Association Strength

2.1 = T G A A C C A T G A A G T A x ?

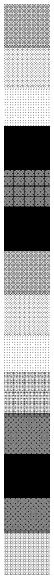
$$y = X \times \beta$$

Multivariate Regression for Single-Trait Association Analysis



Trait | Genotype | Association Strength

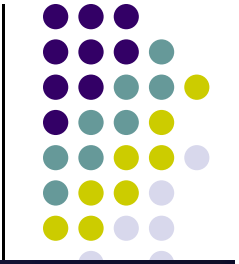
2.1 = T G A A C C A T G A A G T A x



$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

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

Lasso for Reducing False Positives (Tibshirani, 1996)



Trait | Genotype | Association Strength

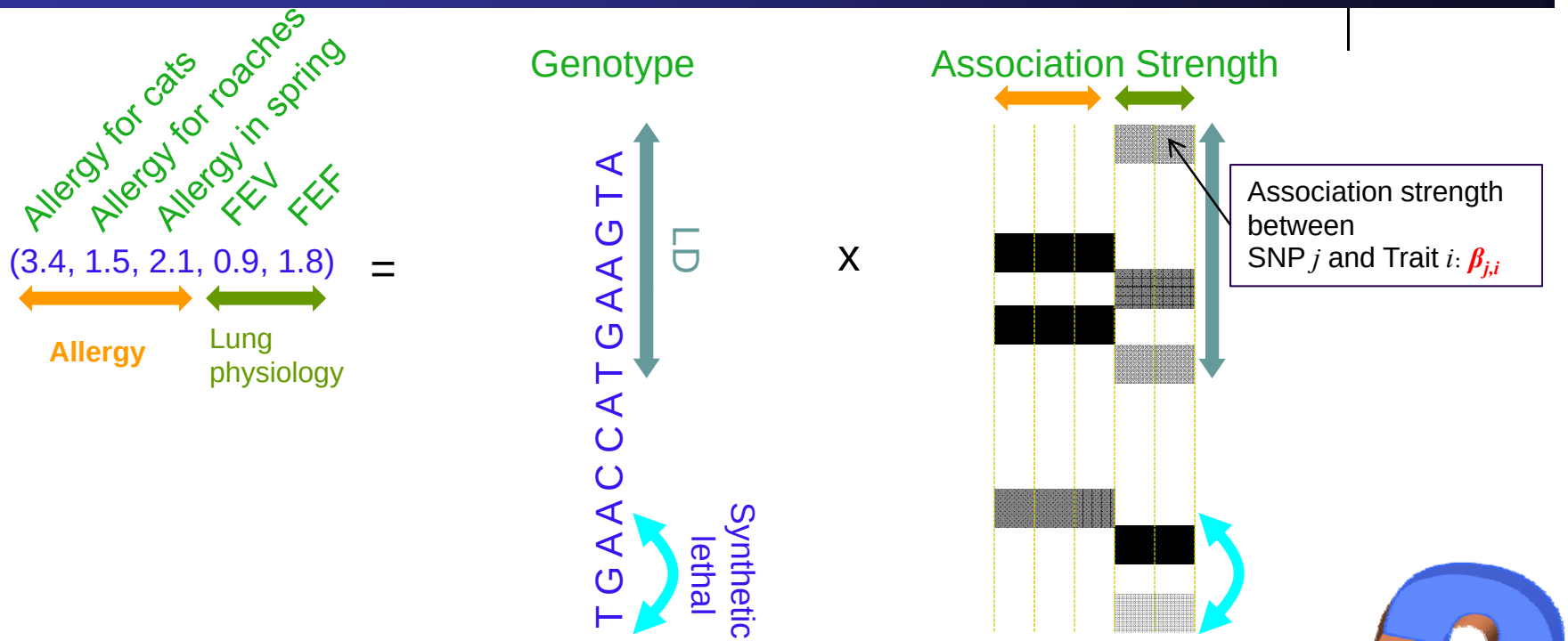
2.1 = T G A A C C A T G A A G T A x

Lasso Penalty for sparsity

$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{J} |\beta_j|$$

Many zero associations (**sparse** results), but what if there are multiple related traits?

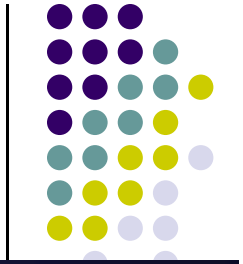
Multivariate Regression for Multiple-Trait Association Analysis



$$\beta^* = \arg\min_{\beta} \sum_i (\mathbf{y}_i - \mathbf{X}_i\beta_i)^T(\mathbf{y}_i - \mathbf{X}_i\beta_i) + \lambda \sum_{i,j} |\beta_{j,i}|$$

How to combine information across multiple traits to increase the power?

Multivariate Regression for Multiple-Trait Association Analysis



Trait

Genotype

Association Strength

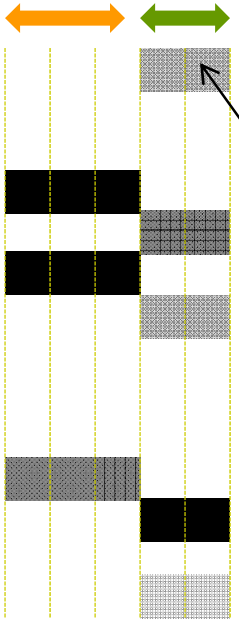
(3.4, 1.5, 2.1, 0.9, 1.8) =

Allergy

Lung physiology

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

X



Association strength between SNP j and Trait i: $\beta_{j,i}$

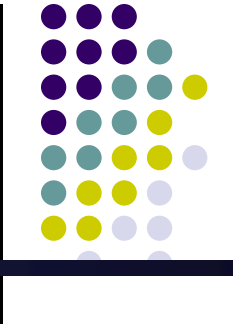
$$\beta^* = \arg\min_{\beta} \sum_i (\mathbf{y}_i - \mathbf{X}_i\beta_i)^T(\mathbf{y}_i - \mathbf{X}_i\beta_i) + \lambda \sum_{i,j} |\beta_{j,i}|$$

+

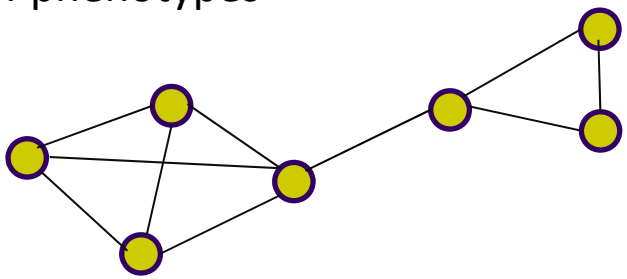
We introduce

graph-guided fusion penalty

Multiple-trait Association: Graph-Constrained Fused Lasso

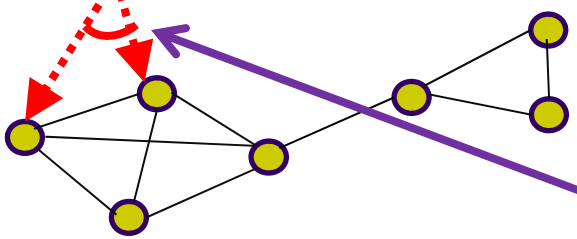


Step 1: Thresholded correlation graph of phenotypes



Step 2: Graph-constrained fused lasso

ACGTTTTACTGTACAATT



Fusion

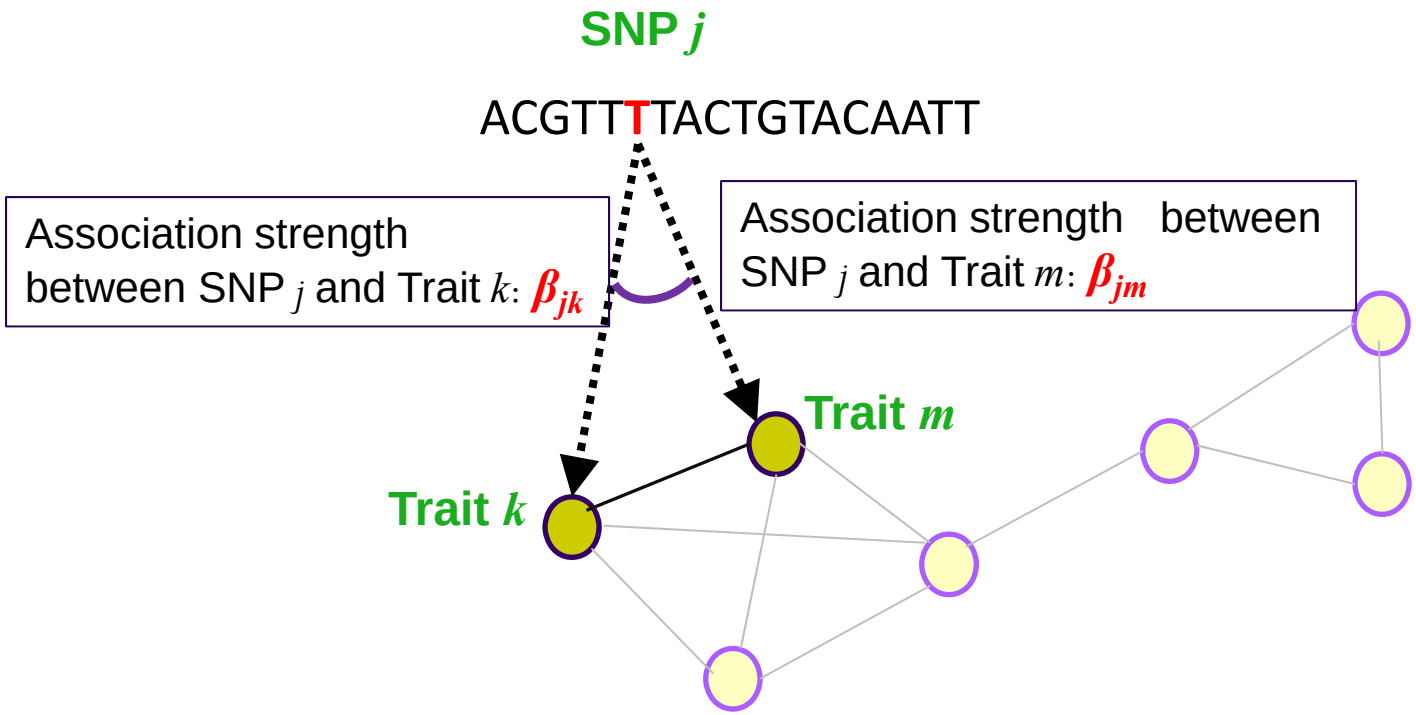
$$\hat{\mathbf{B}}^{\mathrm{GC}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)$$

$$+\lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l)\in E} \sum_j |\beta_{jm} - \operatorname{sign}(r_{ml})\beta_{jl}|$$

Lasso Penalty

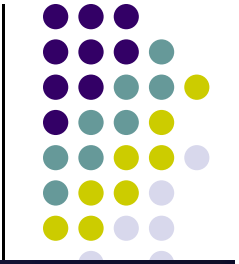
Graph-constrained fusion penalty

Fusion Penalty

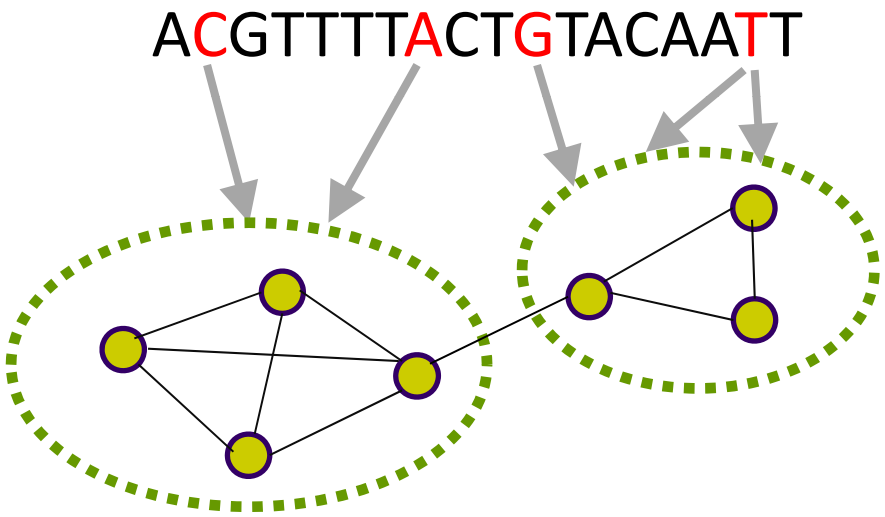


- Fusion Penalty: $|\beta_{jk} - \beta_{jm}|$
- For two correlated traits (connected in the network), the association strengths may have similar values.

Graph-Constrained Fused Lasso

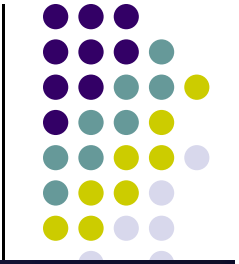


Overall effect

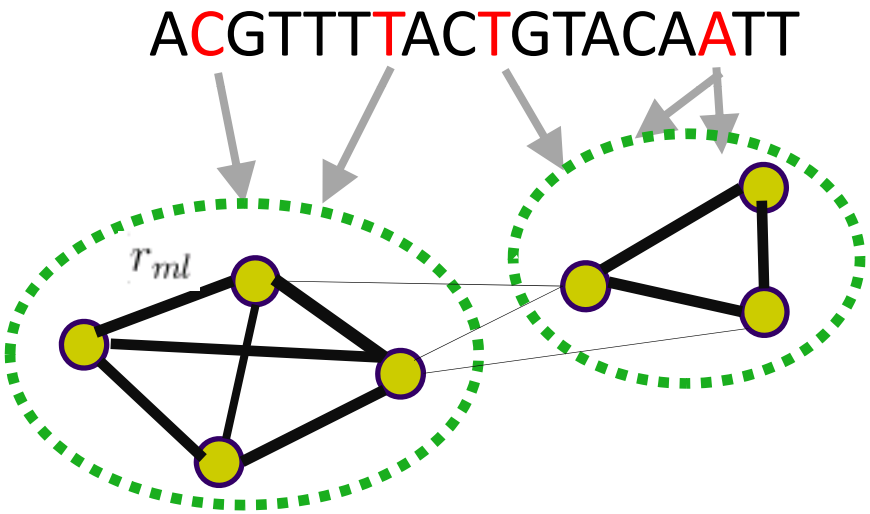


- Fusion effect propagates to the entire network
- Association between SNPs and subnetworks of traits

Multiple-trait Association: Graph-Weighted Fused Lasso

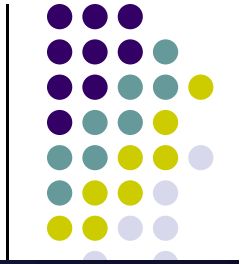


Overall effect



- 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
  - Graph-constrained fused lasso

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

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

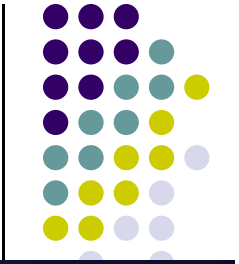
  - Graph-weighted fused lasso

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

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

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

Improving Scalability



Original problem

$$\min_{\beta_k} \sum_k (\mathbf{y}_k - \mathbf{X}\beta_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\beta_k) + \lambda \sum_{j,k} |\beta_{jk}| + \gamma \sum_{(m,l)\in E} f(r_{ml})^2 \sum_j |\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}|$$

Equivalently

$$\min_{\beta_k} \sum_k (\mathbf{y}_k - \mathbf{X}\beta_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\beta_k) + \lambda \Big(\sum_{j,k} |\beta_{jk}| \Big)^2 + \gamma \sum_{(m,l)\in E} f(r_{ml})^2 \Big(\sum_j |\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}| \Big)^2$$

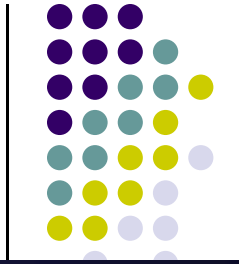
Using a variational formulation

$$\min_{\beta_k,\, d_{jk},\, d_{jml}} \sum_k (\mathbf{y}_k - \mathbf{X}\beta_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\beta_k) + \lambda \sum_{j,k} \frac{(\beta_{jk})^2}{d_{jk}} + \gamma \sum_{(m,l)\in E} f(r_{ml})^2 \sum_j \frac{(\beta_{jm} - \text{sign}(r_{ml})\beta_{jl})^2}{d_{jml}}$$

$$\text{subject to} : \sum_{j,k} d_{jk} = 1, \quad \sum_{(m,l)\in E} \sum_j d_{jml} = 1,$$
$$d_{jk} \geq 0 \text{ for all } j,\, k,$$
$$d_{jml} \geq 0 \text{ for all } j,\, (m,l) \in E,$$

Iterative optimization
- Update $\boldsymbol{\beta}_k$
- Update d_{jk}'s, d_{jml}'s

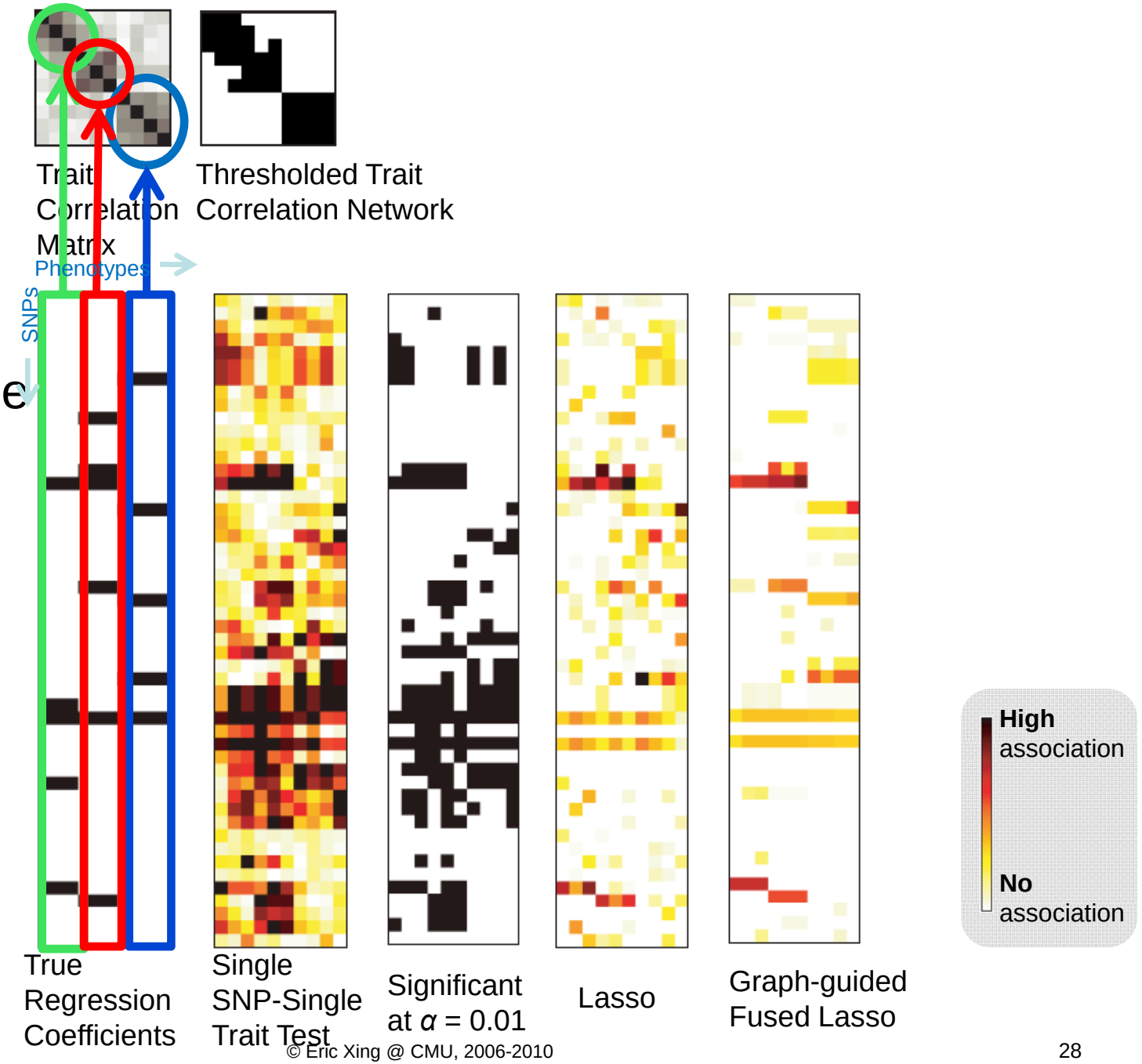
Previous Works vs. Our Approach



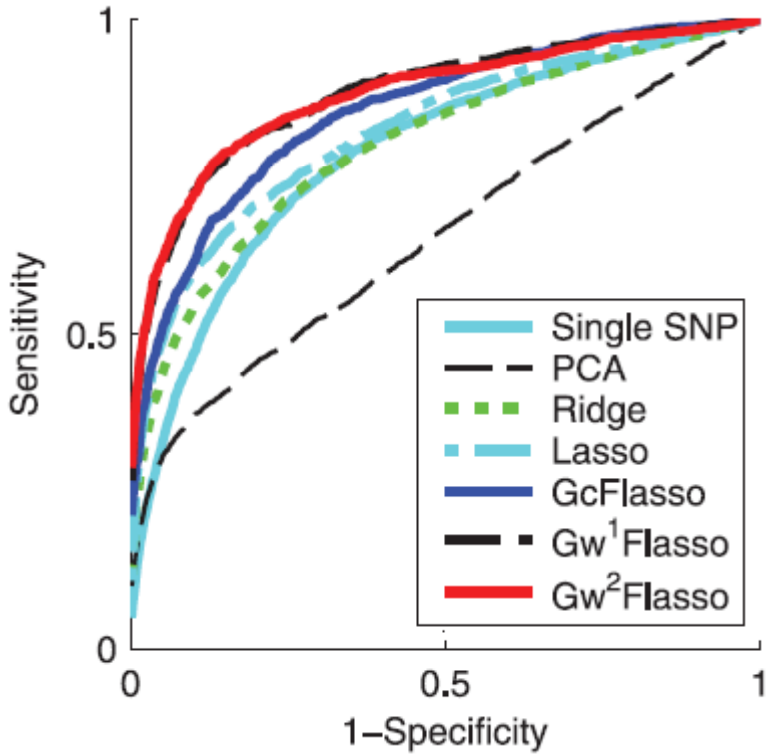
Previous approach		Our approach
PCA-based approach (Weller et al., 1996, Mangin et al., 1998)	**Implicit** representation of trait correlations Hard to interpret the derived traits	**Explicit** representation of trait correlations
Extension of module network for eQTL study (Lee et al., 2009)	**Average** traits within each trait cluster Loss of information	**Original data** for traits are used
Network-based approach (Chen et al., 2008, Emilsson et al., 2008)	**Separate** association analysis for each trait (no information sharing) Single-trait association are combined in light of trait network modules	**Joint** association analysis of multiple traits

Simulation Results

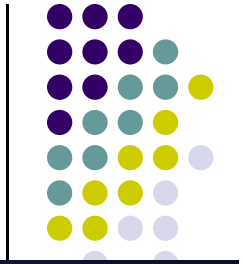
- 50 SNPs taken from HapMap chromosome 7, CEU population
- 10 traits



Simulation Results



Asthma Association Study

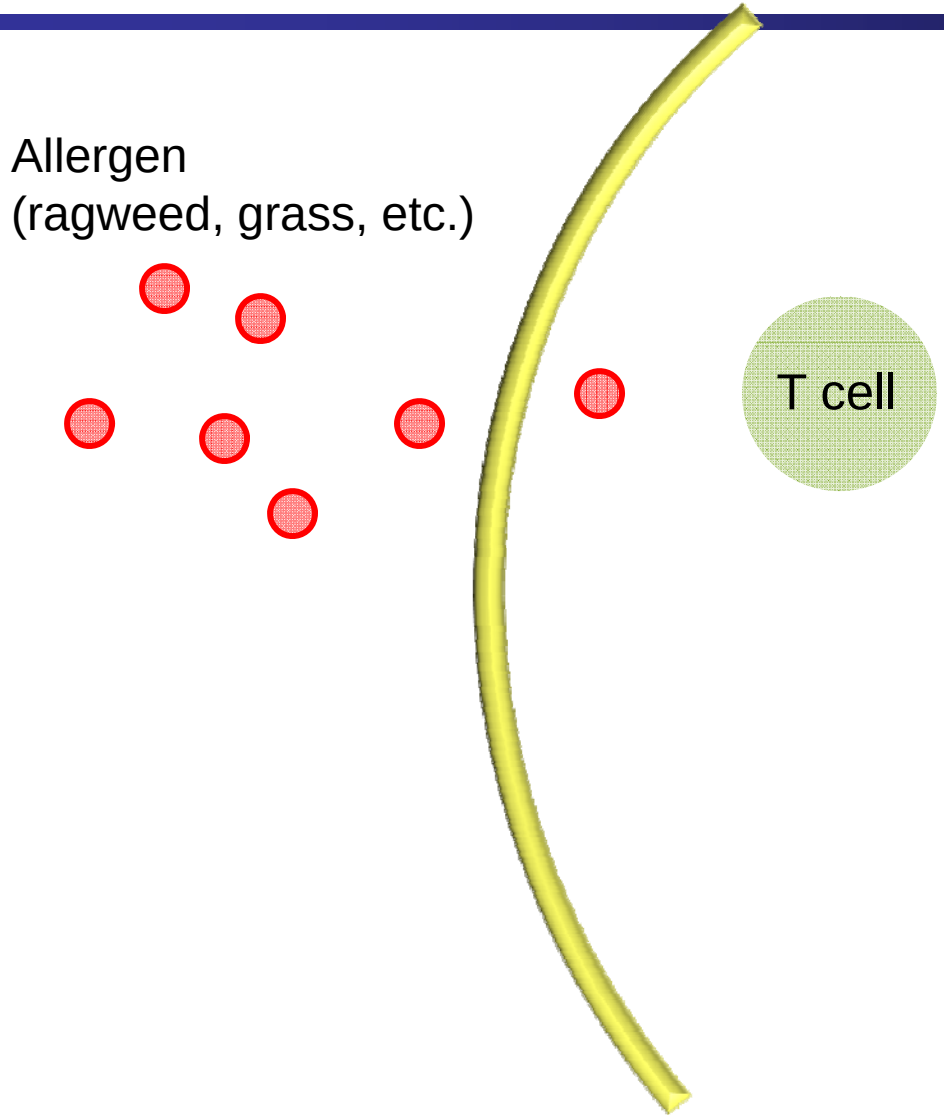


- 543 severe asthma patients from the Severe Asthma Research Program (SARP)

- Genotypes : 34 SNPs in *IL4R* gene
  - 40kb region of chromosome 16
  - Impute missing genotypes with *PHASE* (Li and Stephens, 2003)

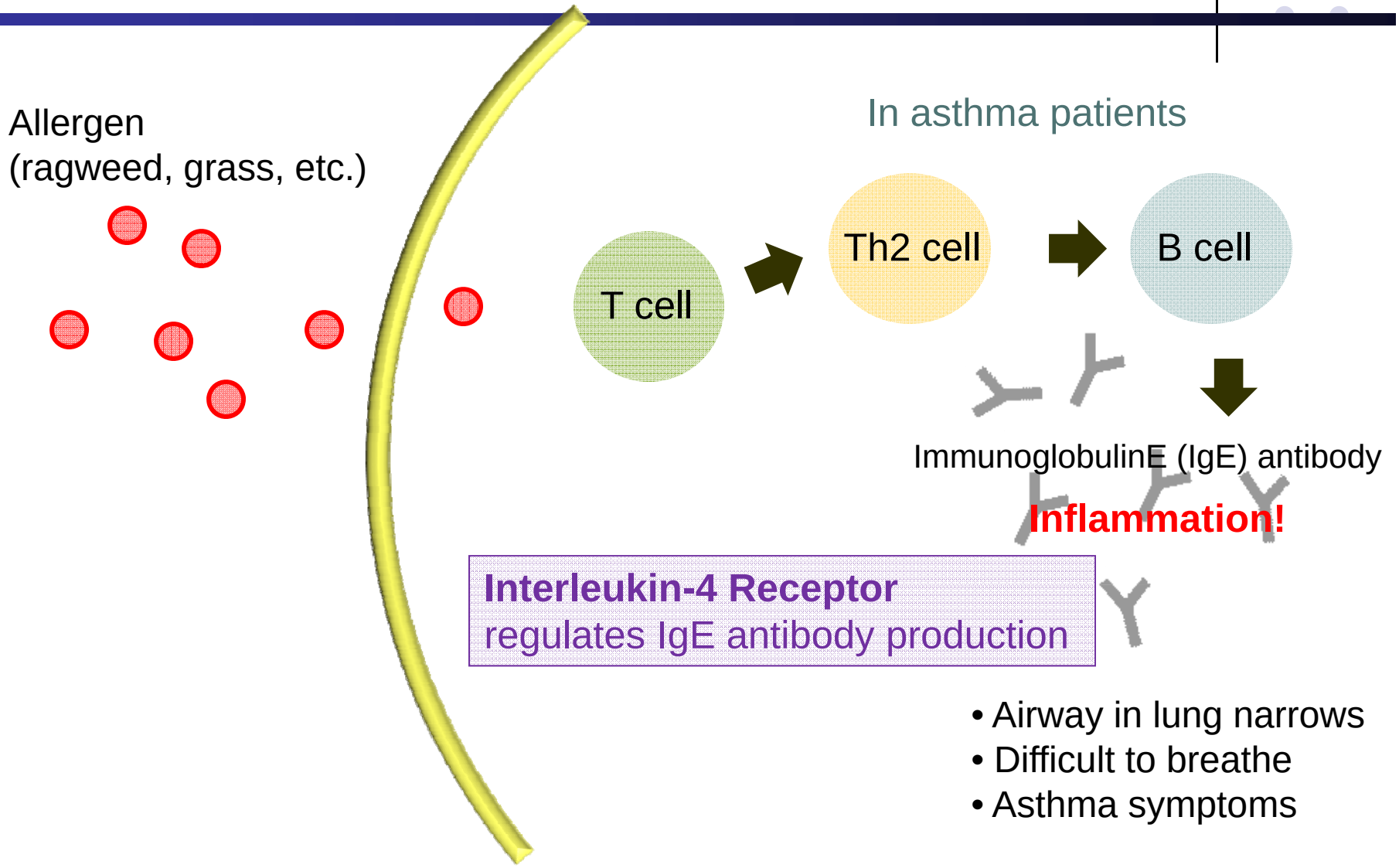
- Traits : 53 asthma-related clinical traits
  - Quality of Life: emotion, environment, activity, symptom
  - Family history: number of siblings with allergy, does the father has asthma?
  - Asthma symptoms: chest tightness, wheeziness

Asthma and *IL4R* Gene



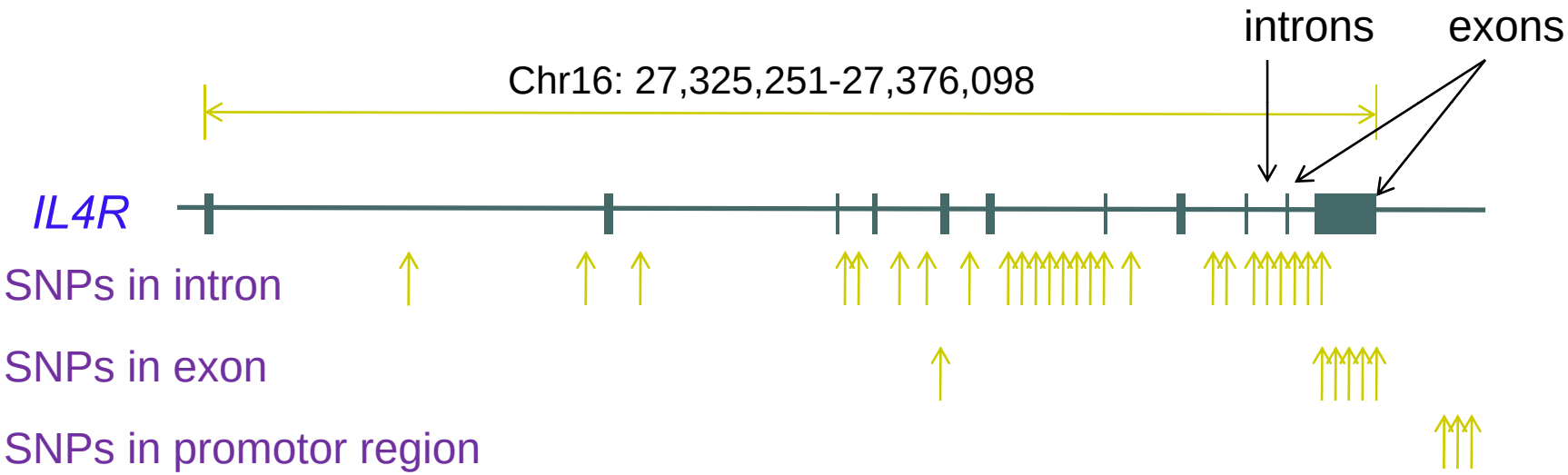
In normal individuals: allergen ignored by the immune system

Asthma and *IL4R* Gene

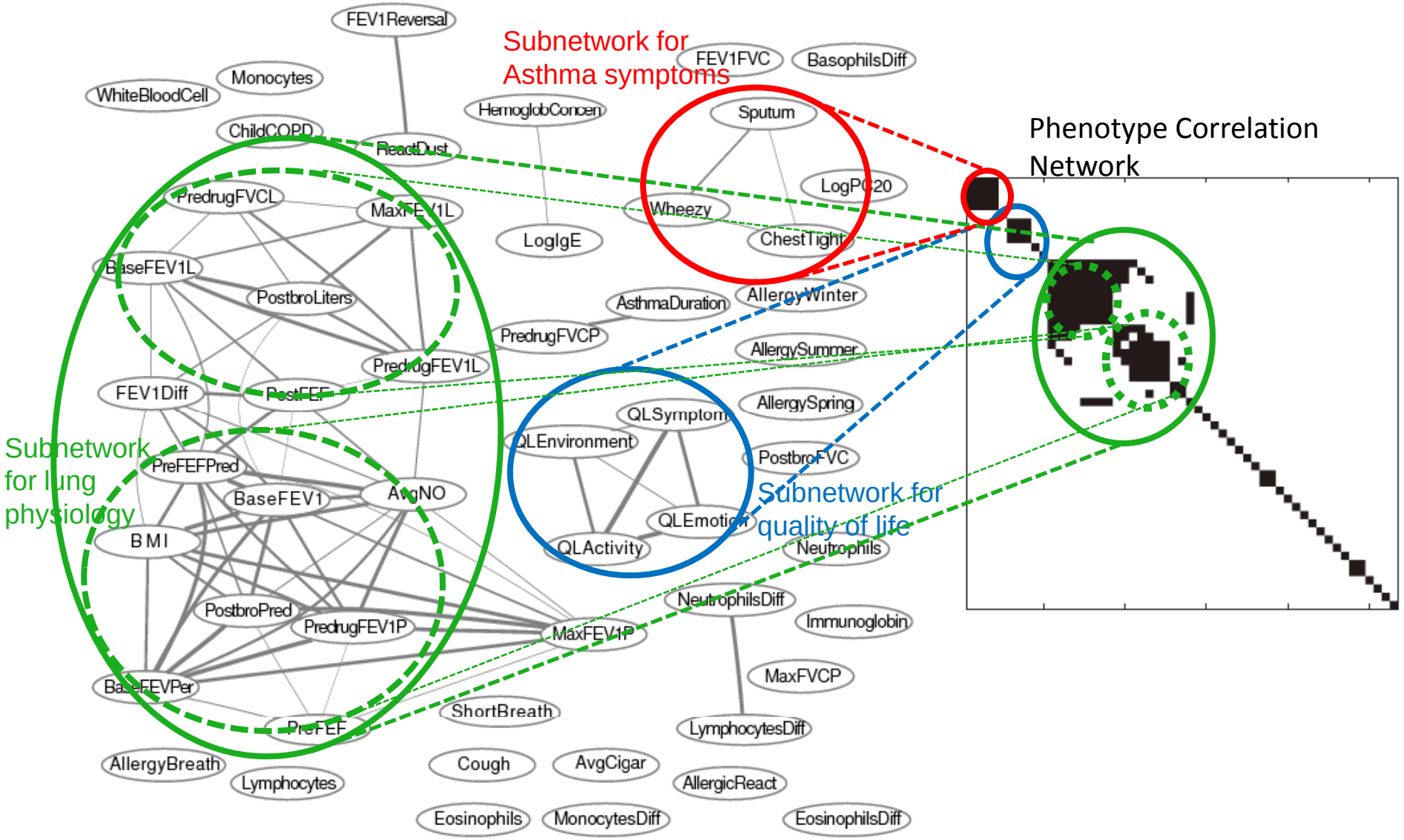


- Airway in lung narrows
- Difficult to breathe
- Asthma symptoms

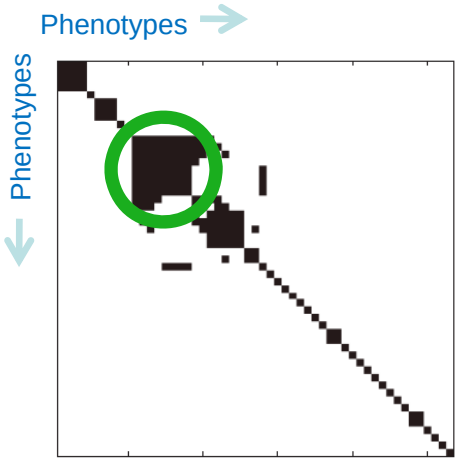
IL4R Gene



Asthma Trait Network



Results from Single-SNP/Trait Test



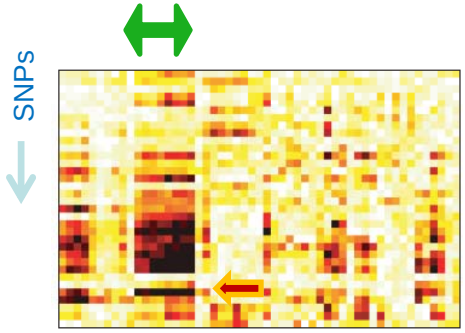
Trait Network

Lung physiology-related traits I

- Baseline FEV1 predicted value: MPVLung
- Pre FEF 25-75 predicted value
- Average nitric oxide value: online
- Body Mass Index
- Postbronchodilation FEV1, liters: Spirometry
- Baseline FEV1 % predicted: Spirometry
- Baseline predrug FEV1, % predicted
- Baseline predrug FEV1, % predicted

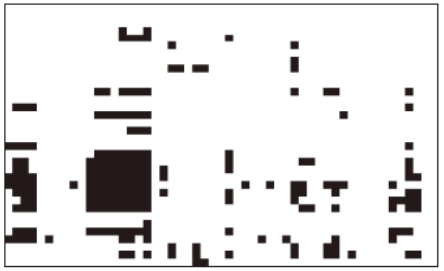
Q551R SNP

- Codes for amino-acid changes in the intracellular signaling portion of the receptor
- Exon 11

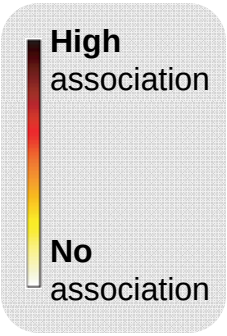


Single-Marker
Single-Trait Test

Permutation test
$\alpha = 0.05$



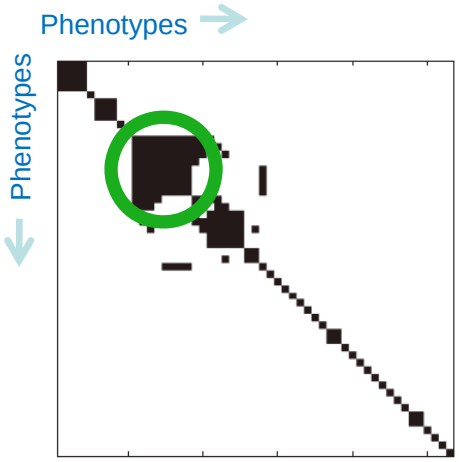
Permutation test
$\alpha = 0.01$



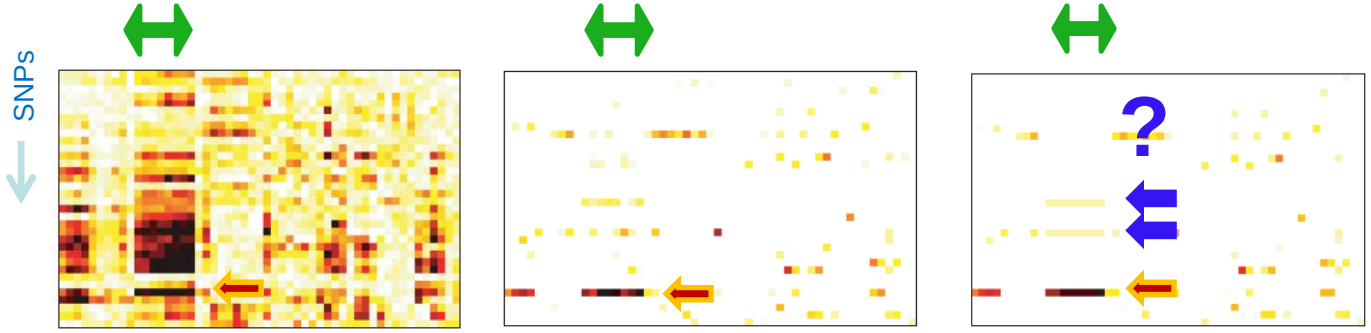
Comparison of Gflasso with Others

Lung physiology-related traits I
- Baseline FEV1 predicted value: MPVLung
- Pre FEF 25-75 predicted value
- Average nitric oxide value: online
- Body Mass Index
- Postbronchodilation FEV1, liters: Spirometry
- Baseline FEV1 % predicted: Spirometry
- Baseline predrug FEV1, % predicted
- Baseline predrug FEV1, % predicted

Q551R SNP
- Codes for amino-acid changes in the intracellular signaling portion of the receptor
- Exon 11



Trait Network

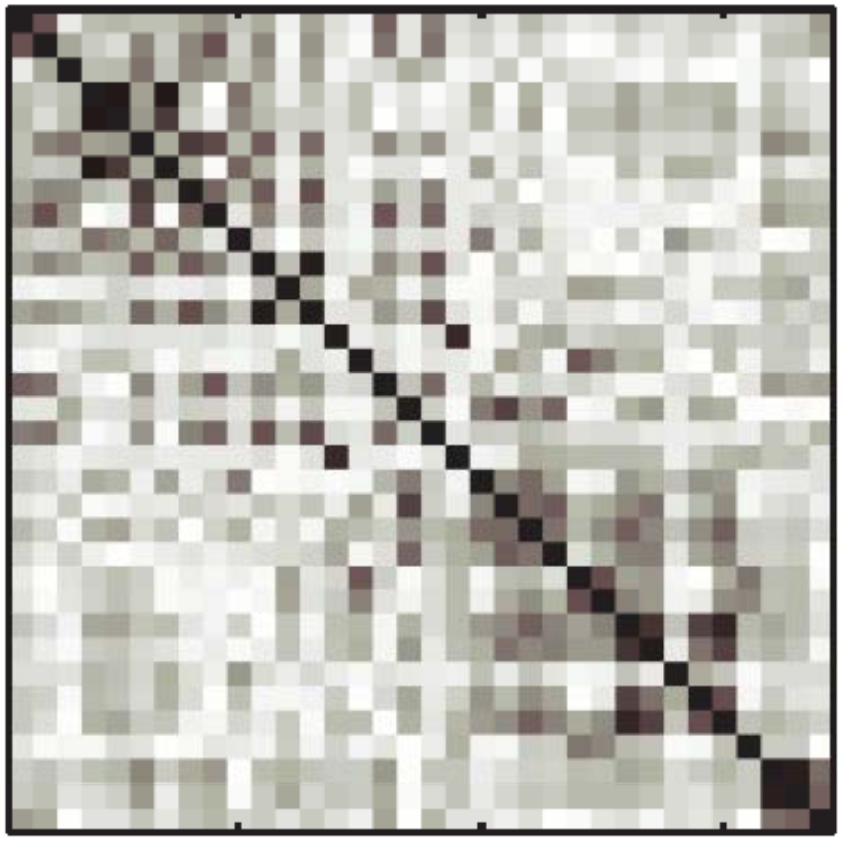


Single-Marker
Single-Trait Test

Lasso

Graph-guided
Fused Lasso

Linkage Disequilibrium Structure in *IL-4R* gene



SNP rs3024622

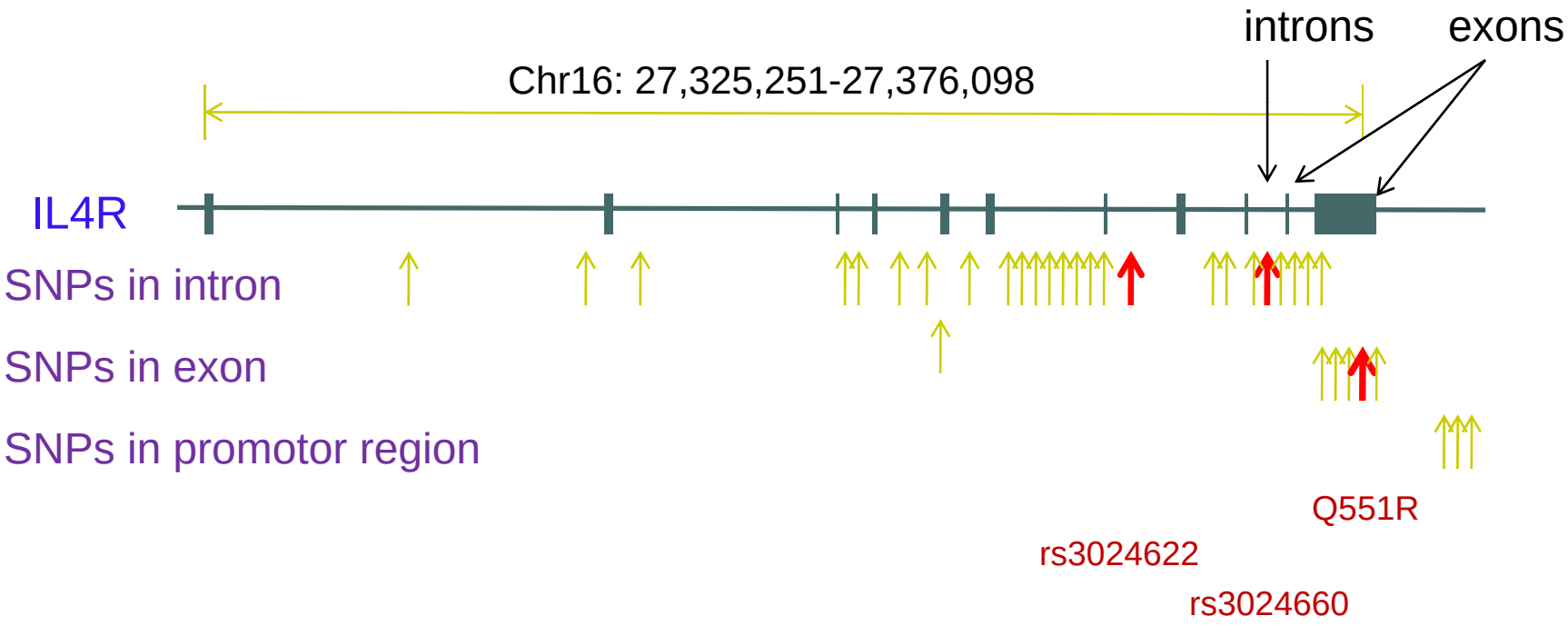
SNP rs3024660

SNP Q551R

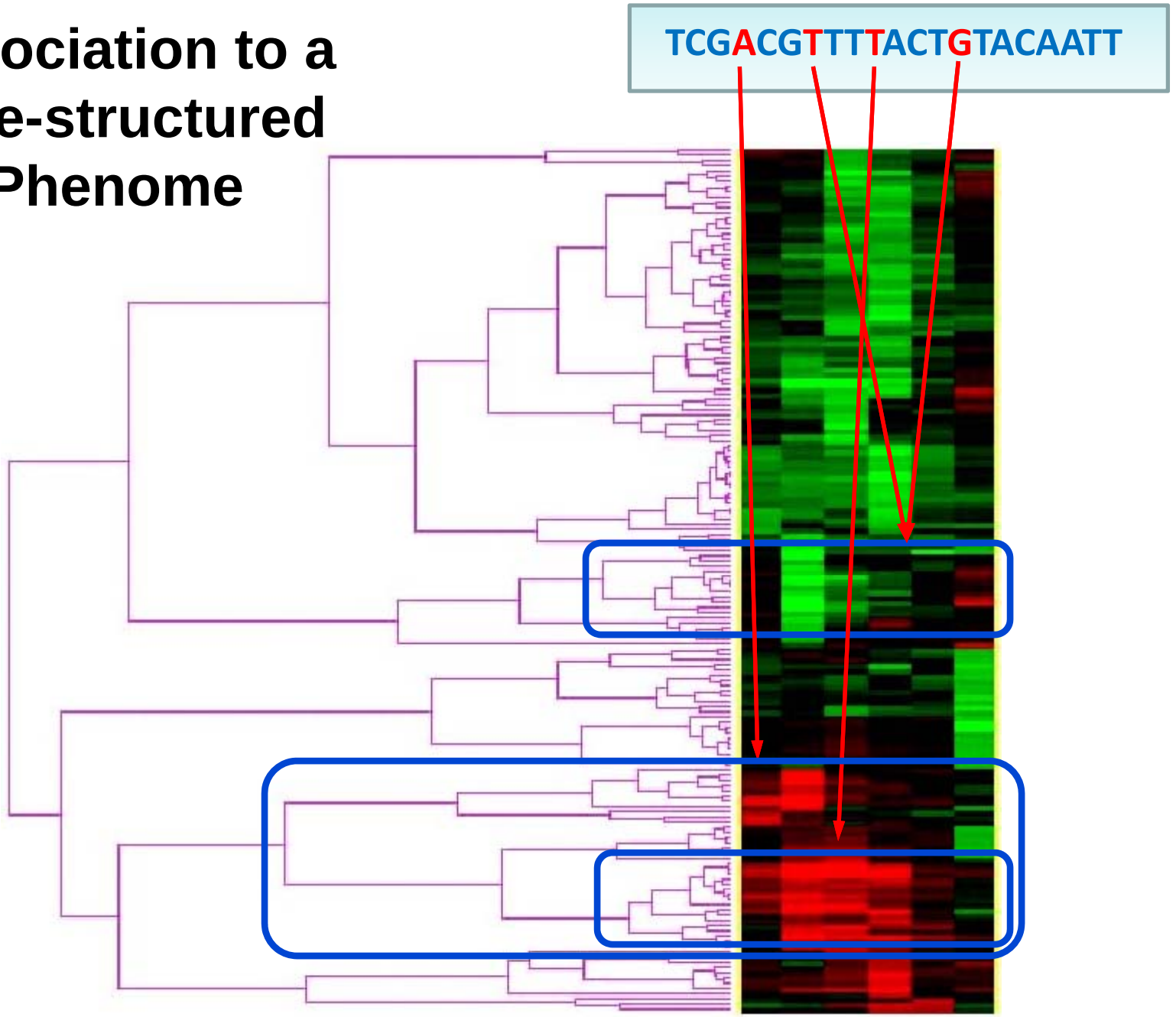
r^2 =0.64

r^2 =0.07

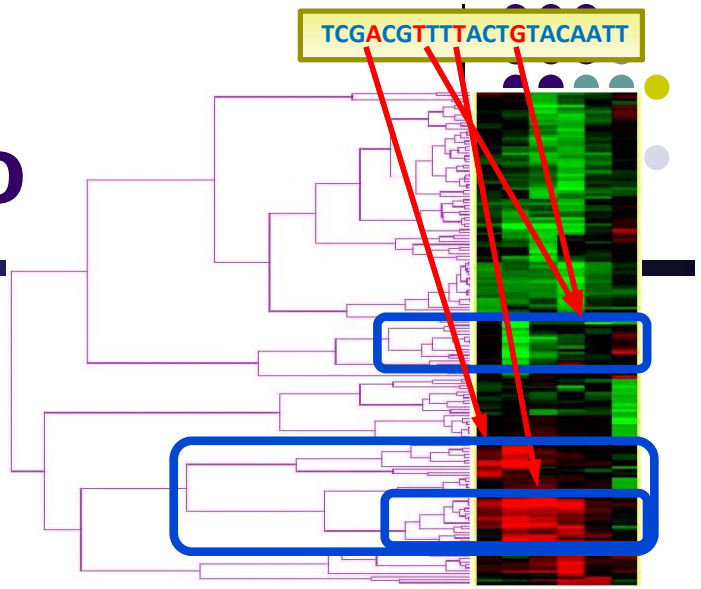
IL4R Gene



Association to a Tree-structured Phenome



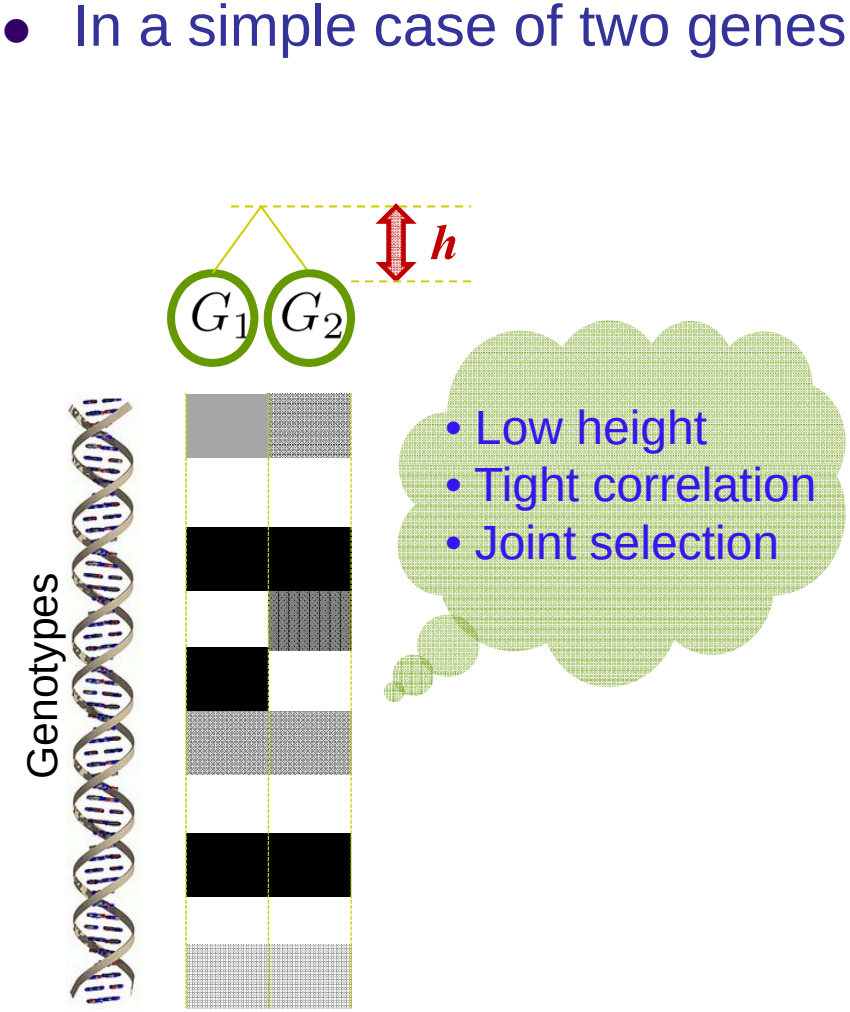
Tree-guided Group Lasso

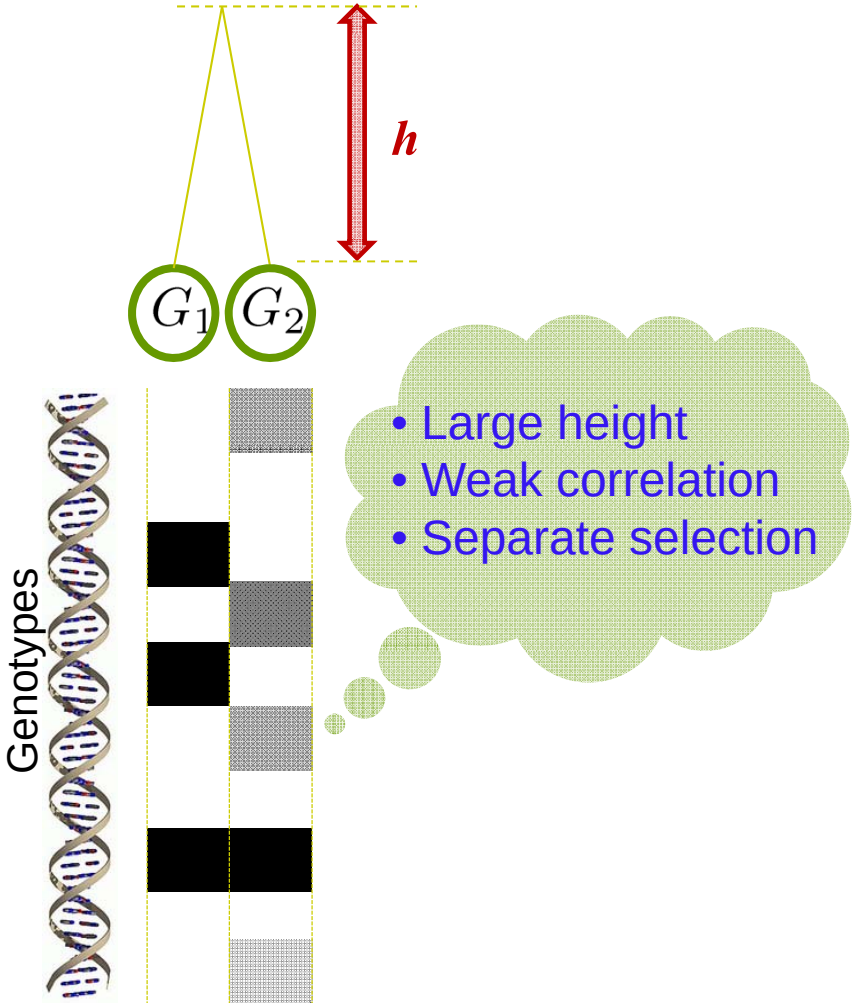


- Why tree?
  - Tree represents a clustering structure
  - Scalability to a very large number of phenotypes
    - Graph : $O(|V|^2)$ edges
    - Tree : $O(|V|)$ edges
  - Expression quantitative trait mapping (eQTL)
    - Agglomerative hierarchical clustering is a popular tool

Tree-Guided Group Lasso

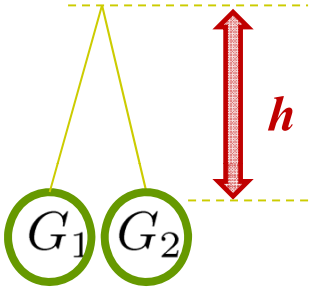
- In a simple case of two genes

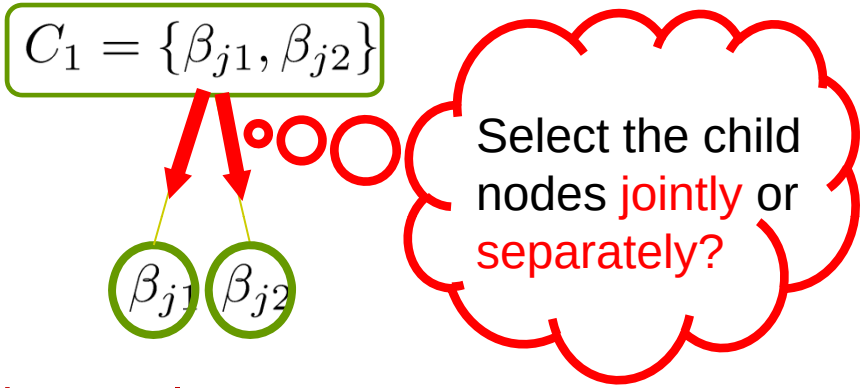


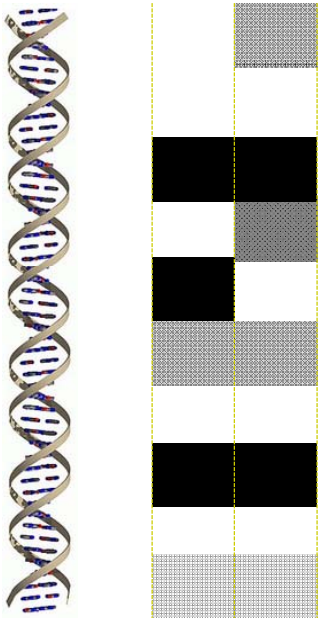


Tree-Guided Group Lasso

- In a simple case of two genes







Tree-guided group lasso

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

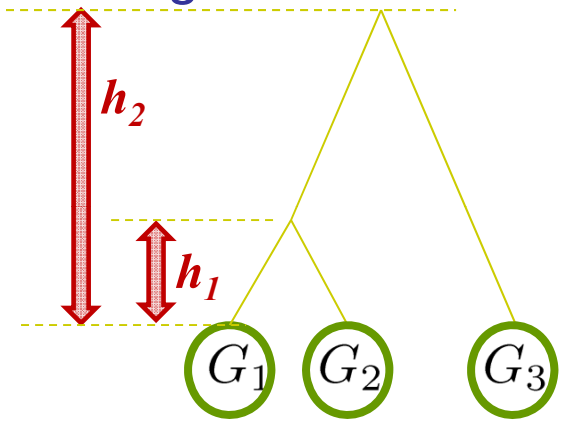
L_1 penalty
- Lasso penalty
- **Separate** selection

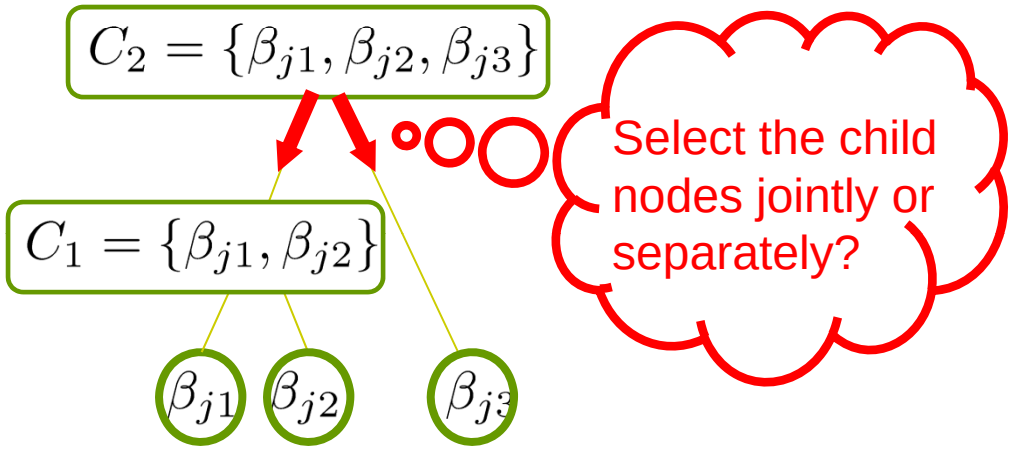
L_2 penalty
- Group lasso
- **Joint** selection

Elastic net

Tree-Guided Group Lasso

- For a general tree





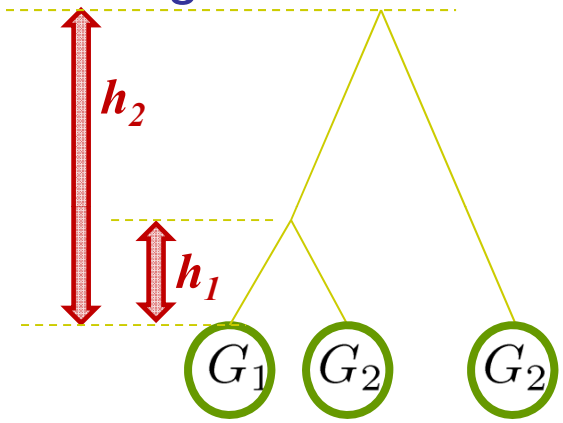
Tree-guided group lasso

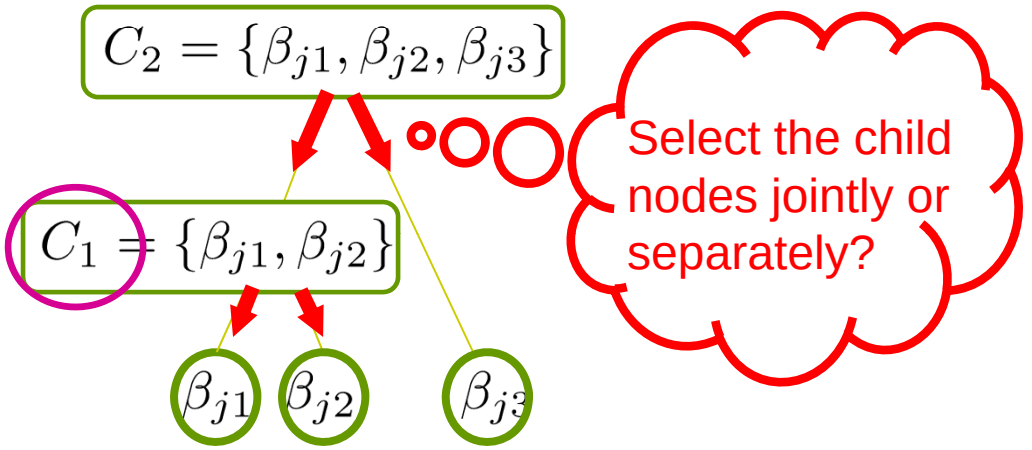
$$\text{argmin} \ (y - X\beta)' \cdot (y - X\beta)$$
$$+\lambda \sum_j \left[(1 - h_2) \left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2} \right) + h_2 \left(|C_1| + |\beta_{j3}| \right) \right]$$

Joint selection — **Separate selection**

Tree-Guided Group Lasso

- For a general tree





Tree-guided group lasso

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

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

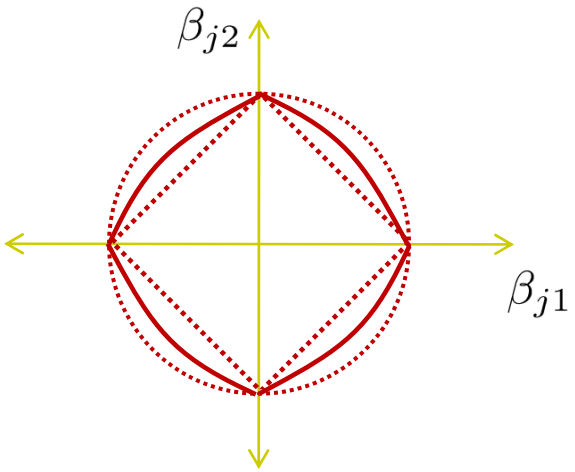
$$(1-h_1)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}\right) + h_1\left(|\beta_{j1}| + |\beta_{j2}|\right)$$

Joint selection **Separate selection**

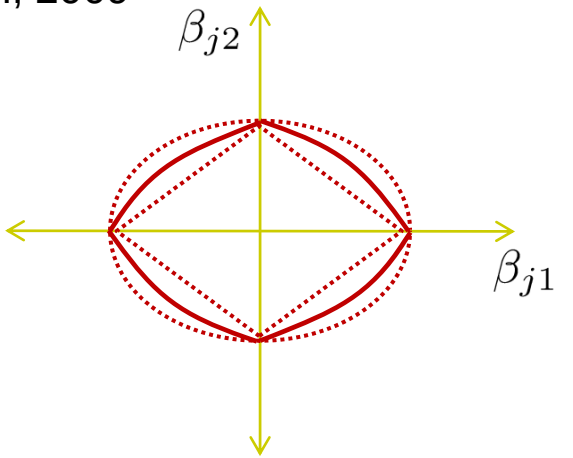
Balanced Shrinkage

Proposition 1 *For each of the k-th output (gene), the sum of the weights w_v for all nodes $v \in V$ in T whose group G_v contains the k-th output (gene) as a member equals one. In other words, the following holds:*

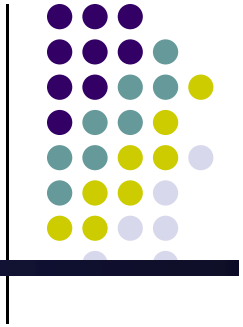
$$\sum_{v:k\in G_v} w_v = \prod_{m\in Ancestors(v_k)} h_m + \sum_{l\in Ancestors(v_k)} (1-h_l) \prod_{m\in Ancestors(v_l)} h_m = 1.$$



Previously, in Jenatton, Audibert & Bach, 2009



Estimating Parameters



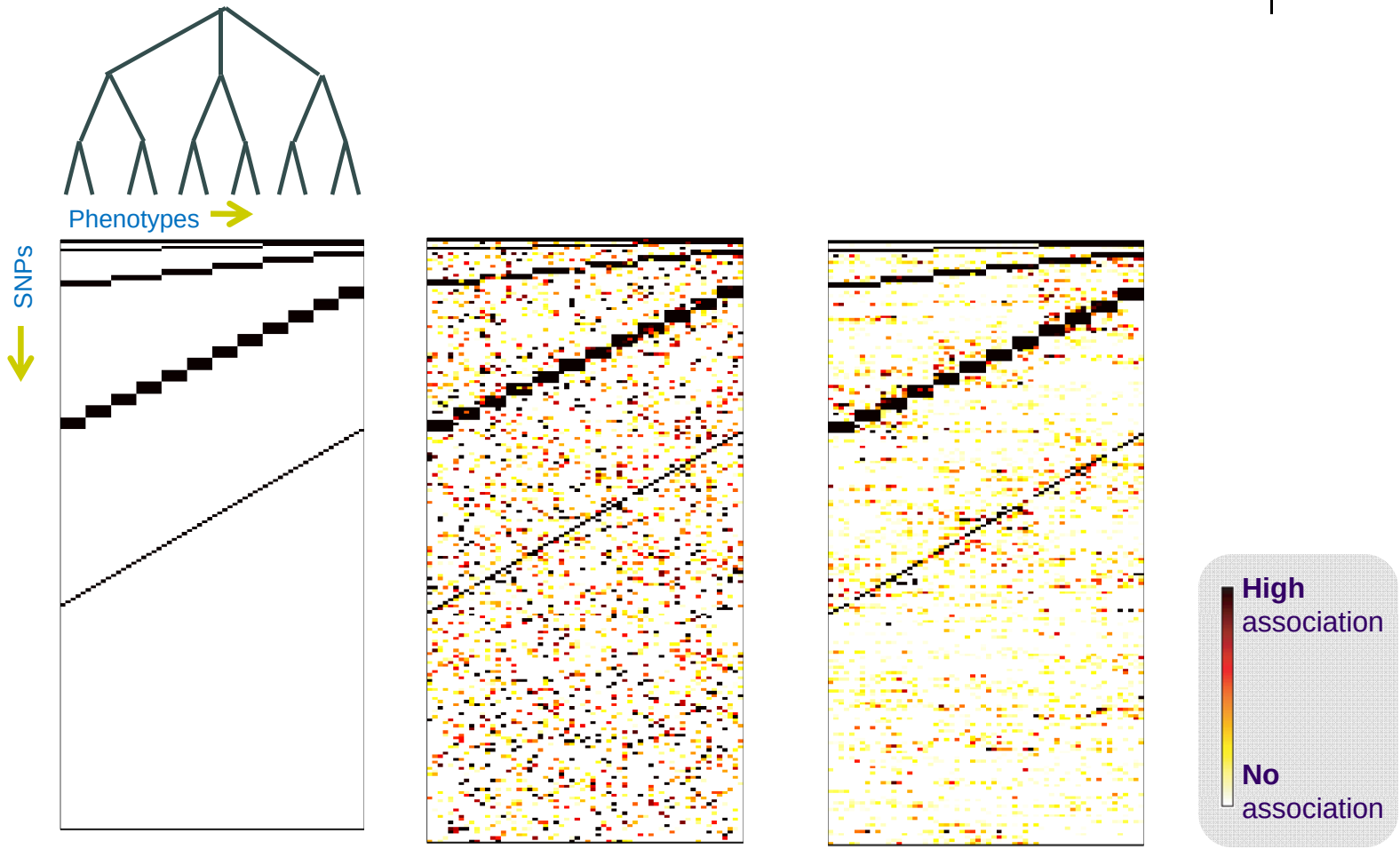
- Second-order cone program

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

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

- Also, variational formulation

Illustration with Simulated Data

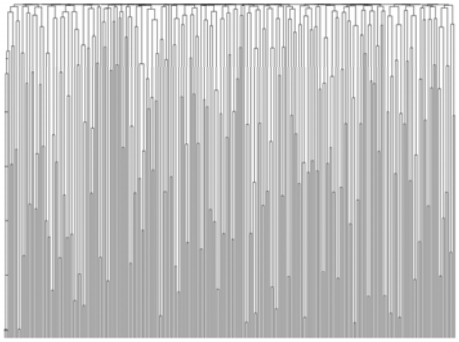


True association strengths

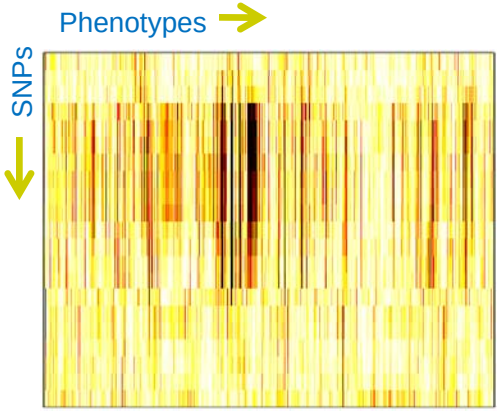
Lasso

Tree-guided group lasso

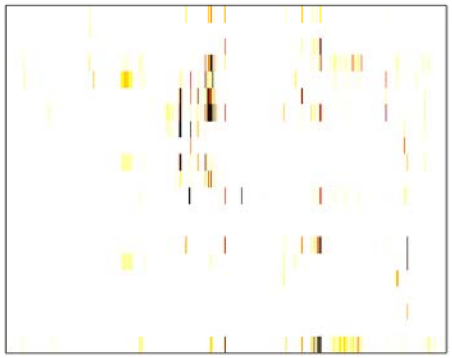
Yeast eQTL Analysis



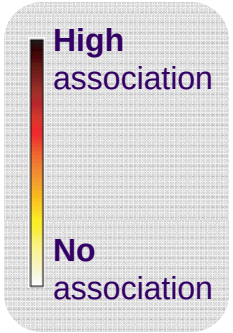
Hierarchical clustering tree



Single-Marker Single-Trait Test

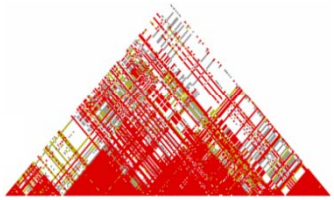


Tree-guided group lasso



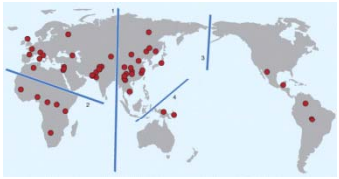
Genome Structure

Linkage Disequilibrium



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

Population Structure



Multi-population group lasso
(Puniyani, Kim, Xing, Submitted)

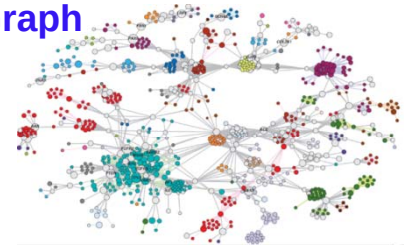
Epistasis

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

Structured Association

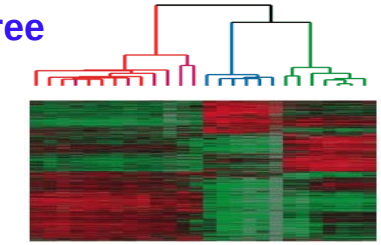
Phenome Structure

Graph



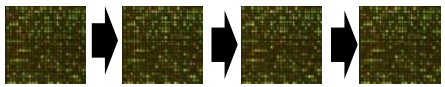
Graph-guided fused lasso
(Kim & Xing, PLoS Genetics, 2009)

Tree

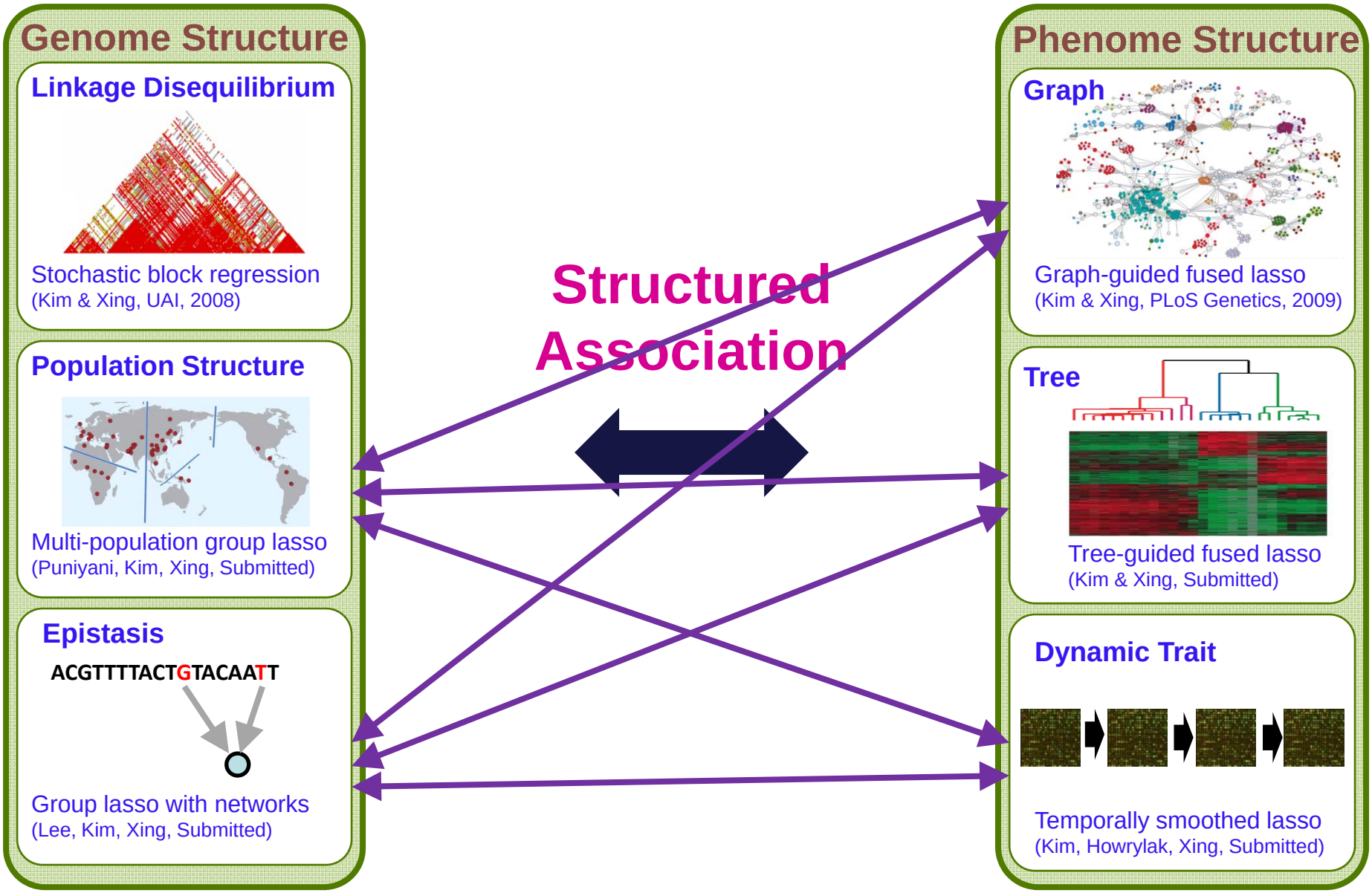


Tree-guided fused lasso
(Kim & Xing, Submitted)

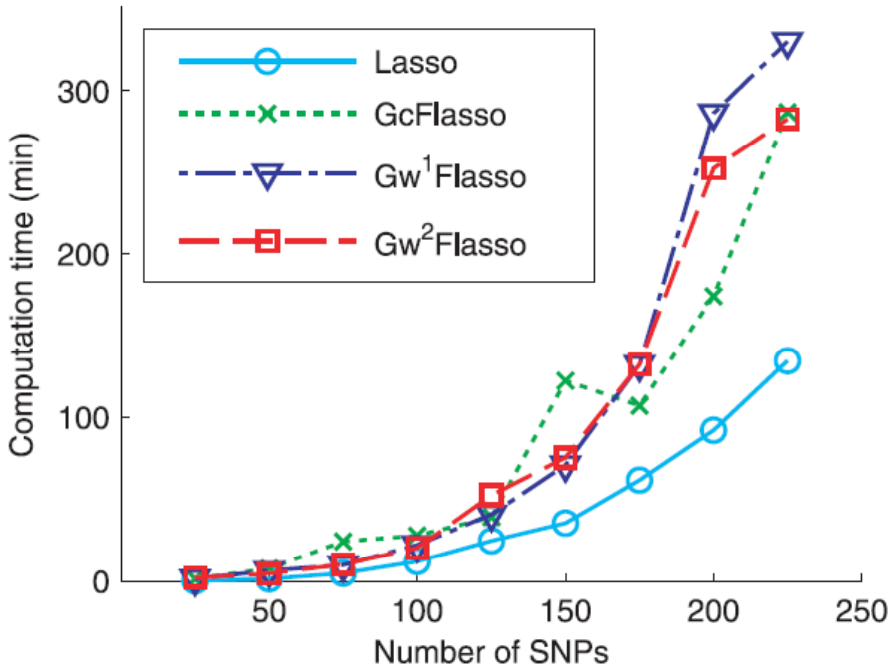
Dynamic Trait

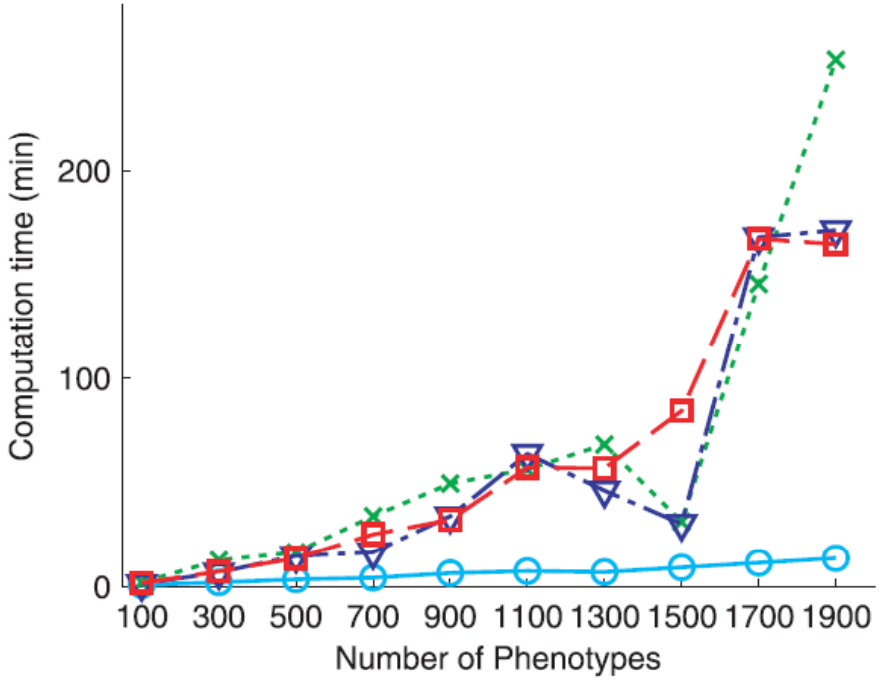


Temporally smoothed lasso
(Kim, Howrylak, Xing, Submitted)

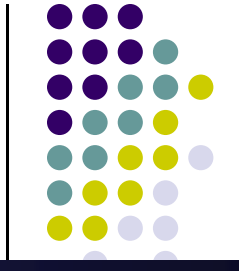


Computation Time





Proximal Gradient Descent



Original Problem:

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

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

Approximation Problem:

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

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

Gradient of the Approximation:

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

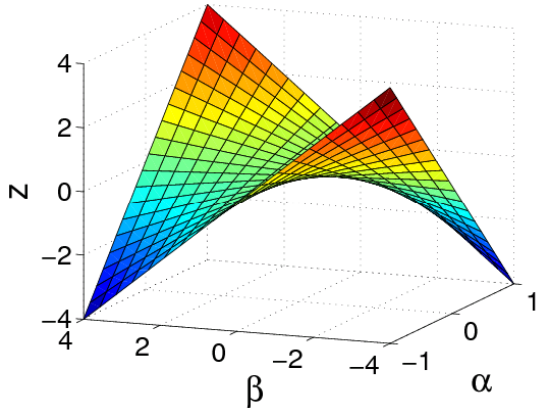
$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}\in\mathcal{Q}} \boldsymbol{\alpha}^T C\boldsymbol{\beta} - \mu d(\boldsymbol{\alpha})$$

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

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

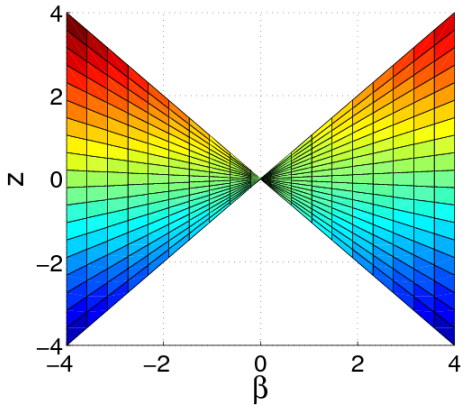
Geometric Interpretation

- Smooth approximation



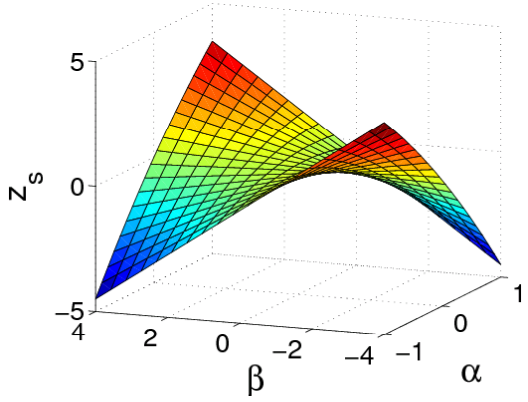
$z(\alpha, \beta) = \alpha\beta$

Projection onto $z - \beta$ Plane



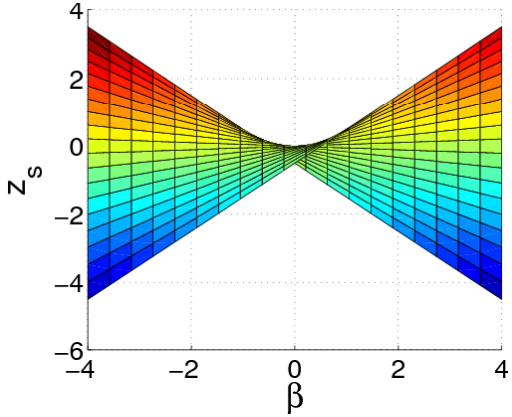
$f_0(\beta) = \max_{\alpha \in [-1,1]} z(\alpha, \beta) = |\beta|$

Uppermost Line Nonsmooth



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

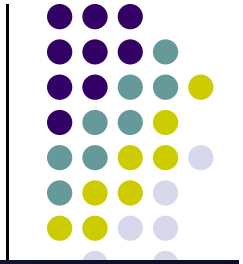
Projection onto $z_s - \beta$ Plane



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

Uppermost Line Smooth

Convergence Rate

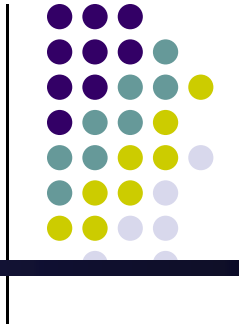


Theorem: If we require $f(\beta^t) - f(\beta^*) \leq \epsilon$ and set $\mu = \frac{\epsilon}{2D}$, the number of iterations is upper bounded by:

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

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

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

Graph:

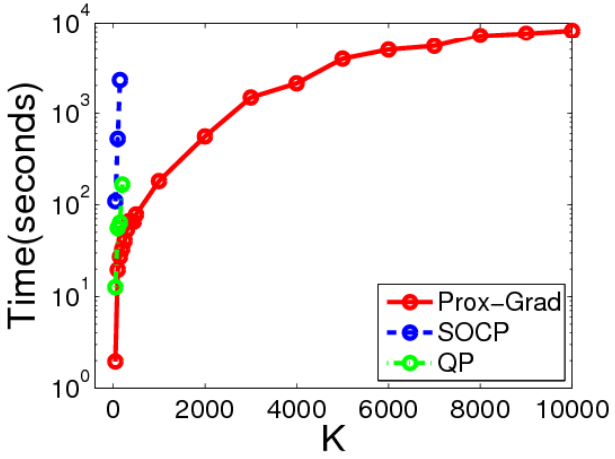
IPM for SOCP	$O\left(J^2(K+\lvert E\rvert)^2(KN + JK + J\lvert E\rvert)\right)$
Proximal-Gradient	$O(J^2K + J\lvert E\rvert)$

Proximal-Gradient: Independent of Sample Size

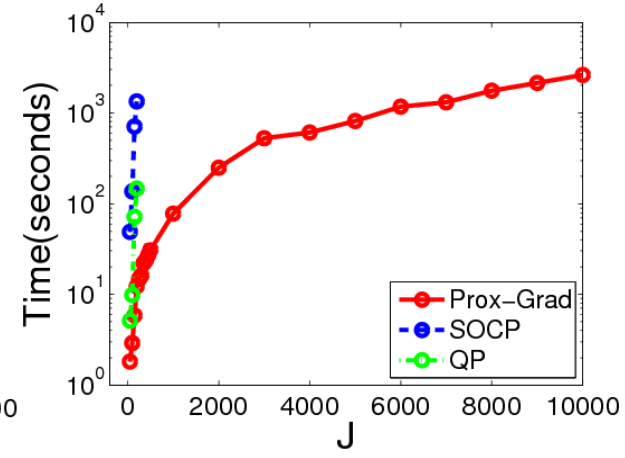
Linear in #.of Tasks

Experiments

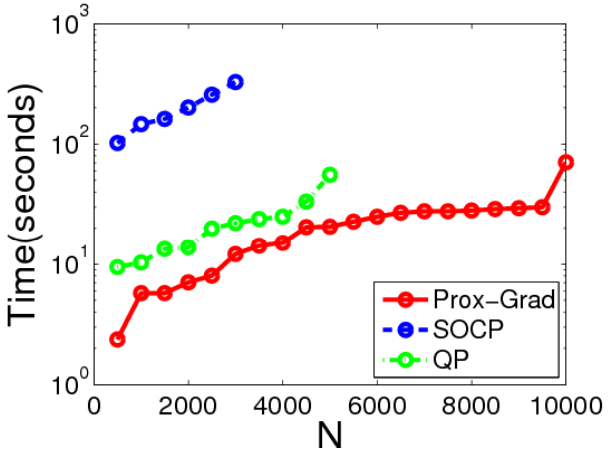
- Multi-task Graph Structured Sparse Learning (GFlasso)



$N = 500, J = 100$



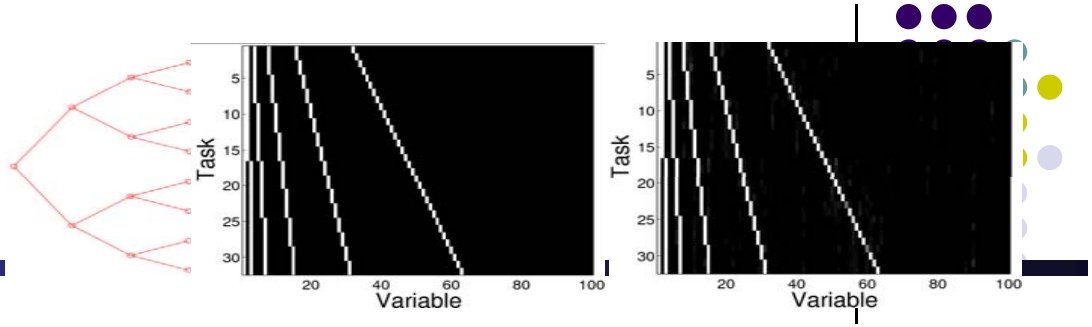
$N = 1000, K = 50$



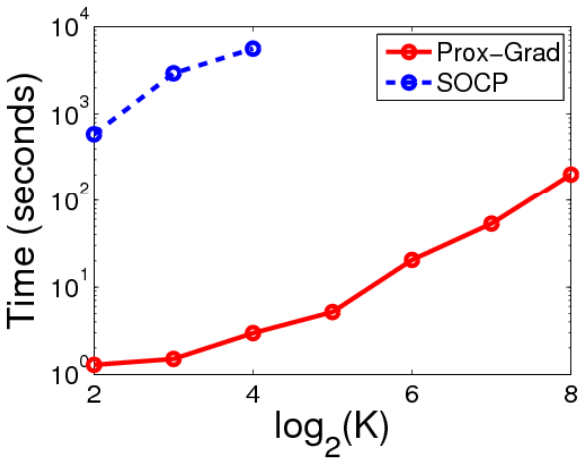
$J = 100, K = 50$

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

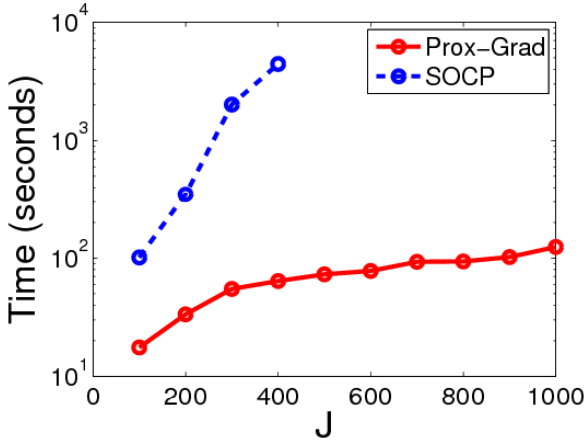
Experiments



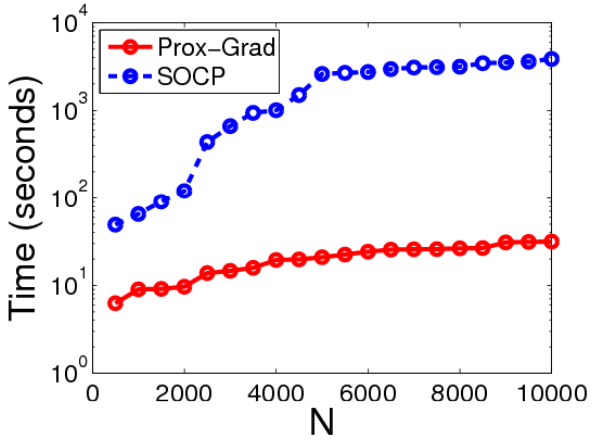
- Multi-task Tree-Structured Sparse Learning (TreeLasso)



$N = 1000, J = 600$



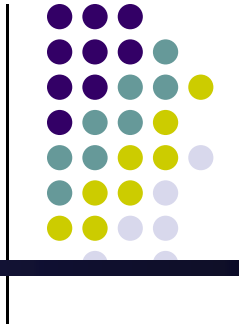
$N = 1000, K = 32$



$J = 100, K = 32$

$\epsilon = 0.1$

Conclusions



- Novel statistical methods for joint association analysis to correlated phenotypes
  - Graph-structured phenome : graph-guided fused lasso
  - Tree-structured phenome : tree-guided group lasso
- Advantages
  - Greater power to detect weak association signals
  - Fewer false positives
  - Joint association to multiple correlated phenotypes
- Other structures
  - In phenotypes: dynamic trait
  - In genotypes: linkage disequilibrium, population structure, epistasis

Reference

- Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society, Series B 58:267–288.
- Weller J, Wiggans G, Vanraden P, Ron M (1996) 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 92:998–1002.
- Mangin B, Thoquet B, Grimsley N (1998) Pleiotropic QTL analysis. Biometrics 54:89–99.
- Chen Y, Zhu J, Lum P, Yang X, Pinto S, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–35.
- Lee SI, Dudley A, Drubin D, Silver P, Krogan N, et al. (2009) Learning a prior on regulatory potential from eQTL data. PLoS Genetics 5:e1000358.
- Emilsson V, Thorleifsson G, Zhang B, Leonardson A, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423–28.
- Kim S, Xing EP (2009) Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genetics 5(8): e1000587.
- Kim S, Xing EP (2008) Sparse feature learning in high-dimensional space via block regularized regression. In Proceedings of the 24th International Conference on Uncertainty in Artificial Intelligence (UAI), pages 325-332. AUAI Press.
- Kim S, Xing EP (2010) Exploiting a hierarchical clustering tree of gene-expression traits in eQTL analysis. Submitted.
- Kim S, Howrylak J, Xing EP (2010) Dynamic-trait association analysis via temporally-smoothed lasso. Submitted.
- Puniyani K, Kim S, Xing EP (2010) Multi-population GWA mapping via multi-task regularized regression. Submitted.
- Lee S, Kim S, Xing EP (2010) Leveraging genetic interaction networks and regulatory pathways for joint mapping of epistatic and marginal eQTLs. Submitted.
- Dunning AM et al. (2009) Association of ESR1 gene tagging SNPs with breast cancer risk. Hum Mol Genet. 18(6):1131-9.
- Esparza-Gordillo J et al. (2009) A common variant on chromosome 11q13 is associated with atopic dermatitis. Nature Genetics 41:596-601.
- Suzuki A et al. (2008) Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population. Nature Genetics 40:1224-1229.
- Dupuis J et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature Genetics 42:105-116.