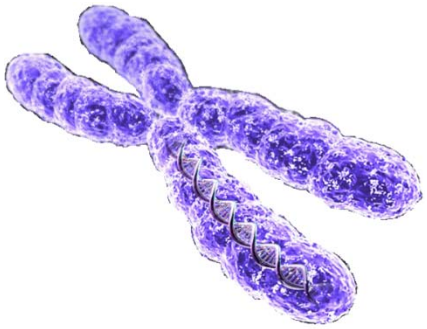
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Graph-induced structured input/output models

- Case Study: Disease Association Analysis

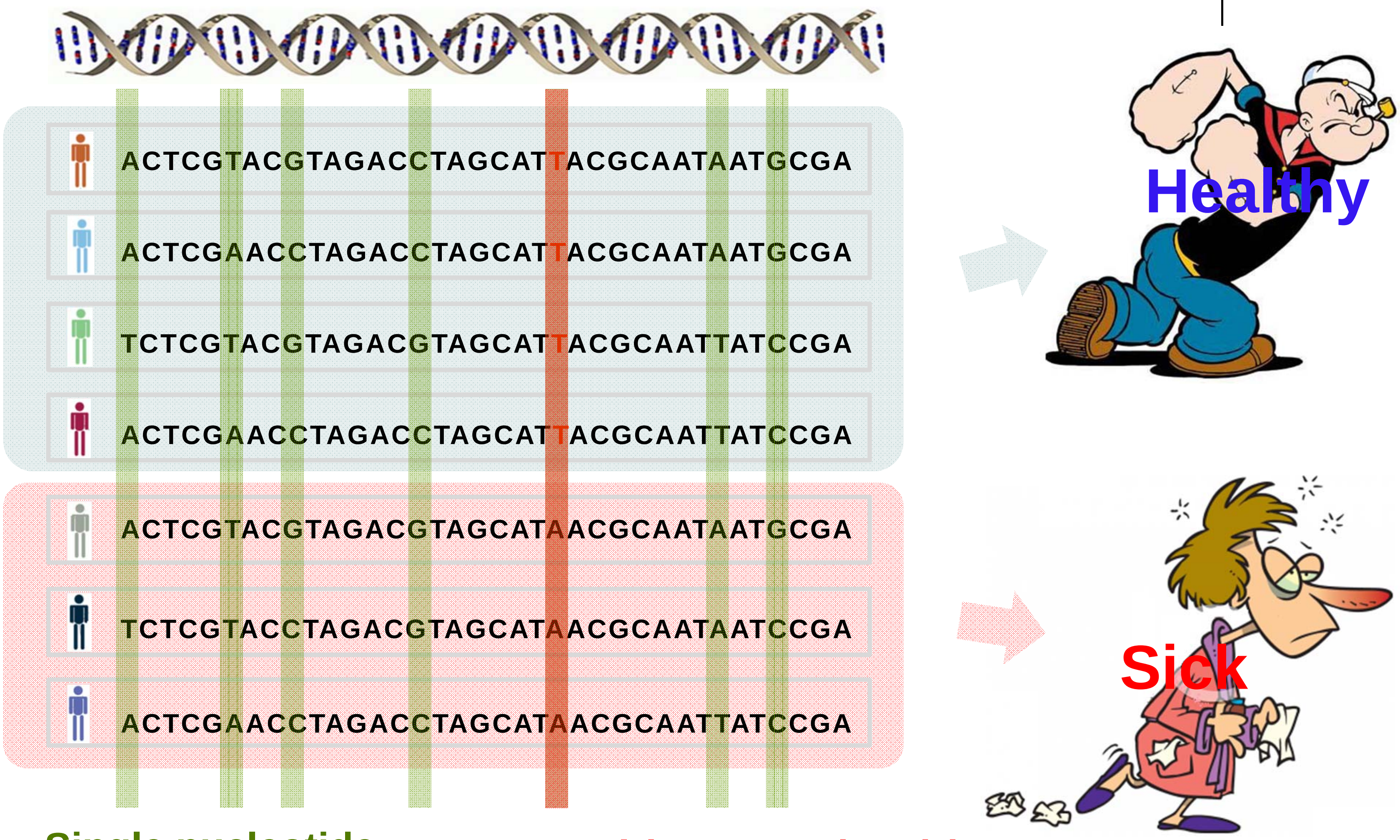


Eric Xing

Lecture 27, April 22, 2015

Reading: See class website

Genetic Basis of Diseases



ACTCGTACGTAGACCTAGCATTACGCAATAATGCGA

ACTCGAACCTAGACCTAGCATTACGCAATAATGCGA

TCTCGTACGTAGACGTAGCATTACGCAATTATCCGA

ACTCGAACCTAGACCTAGCATTACGCAATTATCCGA

ACTCGTACGTAGACGTAGCATAACGCAATAATGCGA

TCTCGTACCTAGACGTAGCATAACGCAATAATCCGA

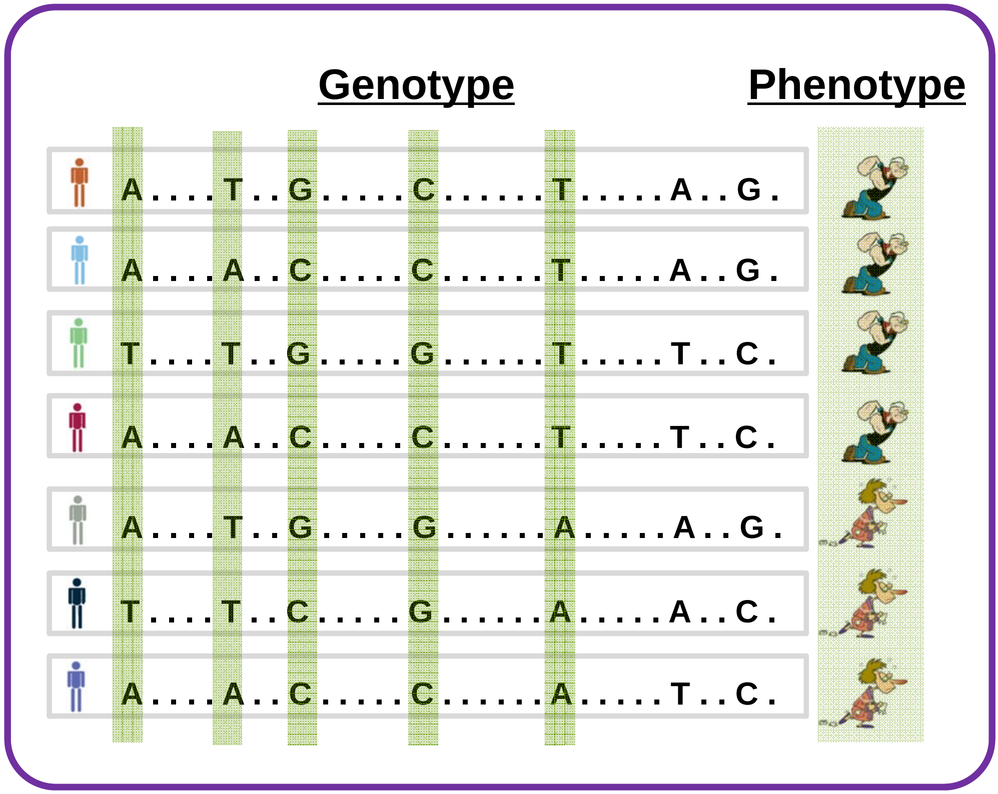
ACTCGAACCTAGACCTAGCATAACGCAATTATCCGA

Single nucleotide polymorphism (SNP)

Causal (or "associated") SNP

Genetic Association Mapping

Data



Genotype

Phenotype

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

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

T....T..G.....G......T.....T..C.

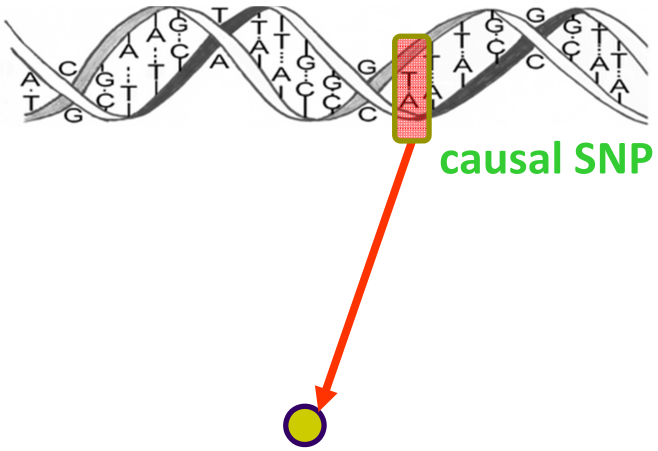
A....A..C.....C......T.....T..C.

A....T..G.....G......A.....A..G.

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

A....A..C.....C......A.....T..C.

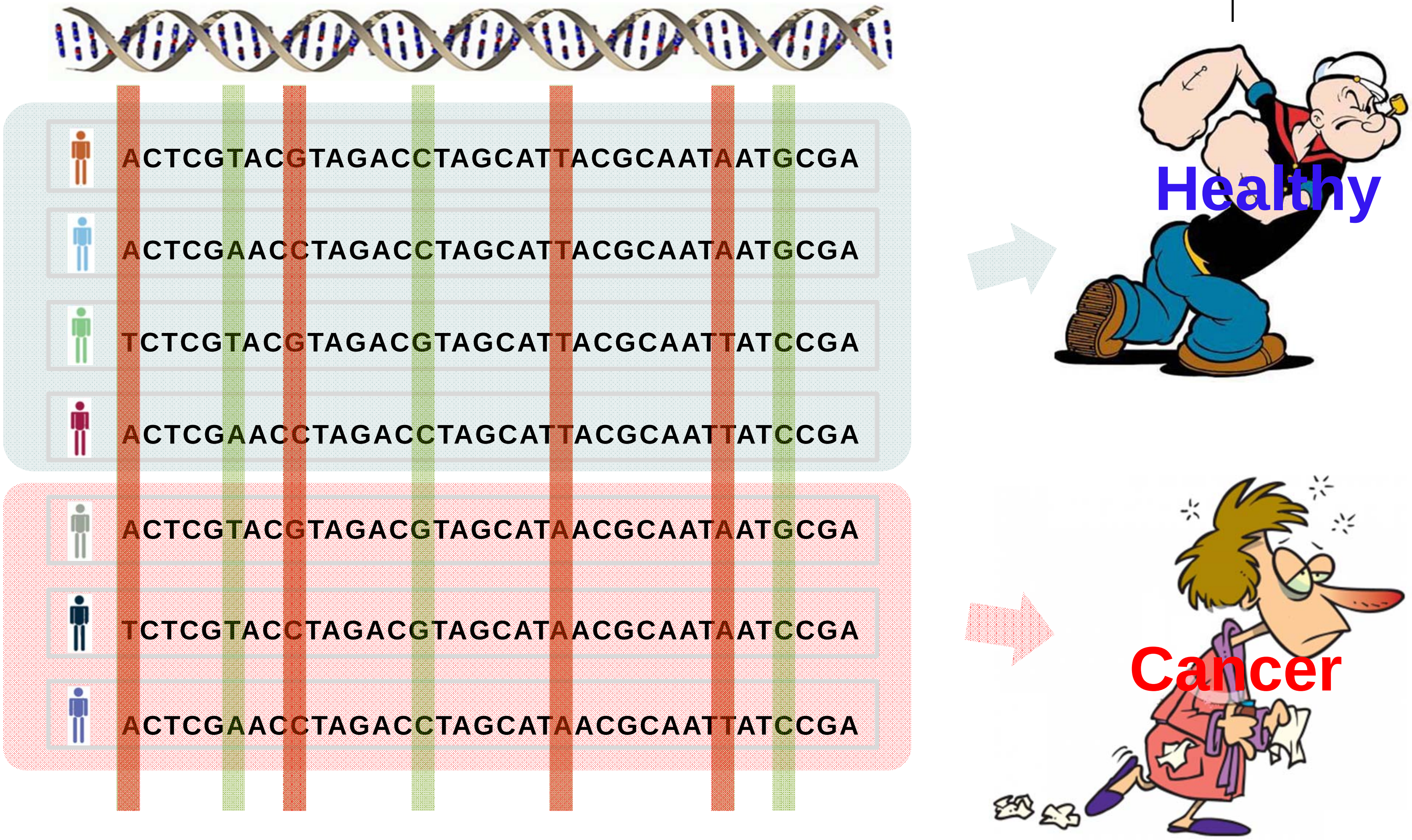
Standard Approach



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



ACTCGTACGTAGACCTAGCATTACGCAATAATGCGA

ACTCGAACCTAGACCTAGCATTACGCAATAATGCGA

TCTCGTACGTAGACGTAGCATTACGCAATTATCCGA

ACTCGAACCTAGACCTAGCATTACGCAATTATCCGA

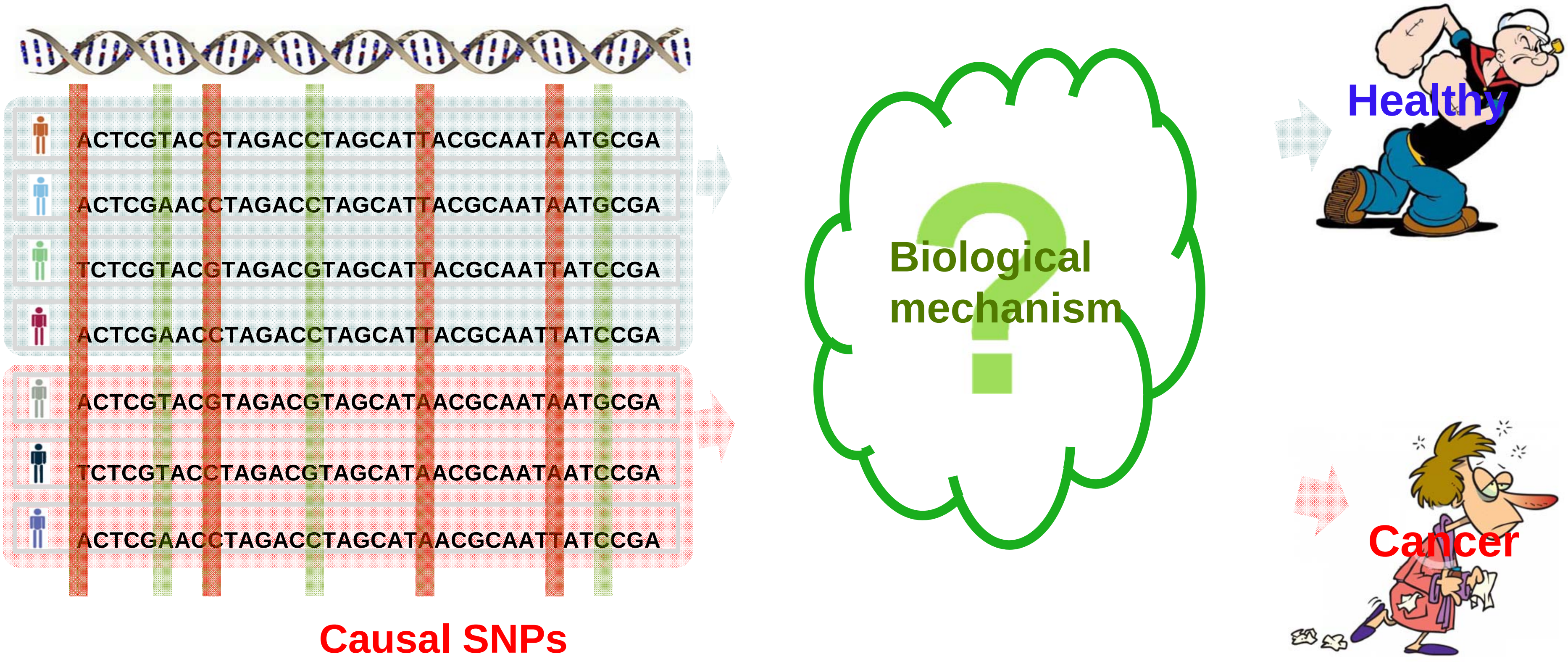
ACTCGTACGTAGACGTAGCATAACGCAATAATGCGA

TCTCGTACCTAGACGTAGCATAACGCAATAATCCGA

ACTCGAACCTAGACCTAGCATAACGCAATTATCCGA

Causal SNPs

Genetic Basis of Complex Diseases



ACTCGTACGTAGACCTAGCATTACGCAATAATGCGA

ACTCGAACCTAGACCTAGCATTACGCAATAATGCGA

TCTCGTACGTAGACGTAGCATTACGCAATTATCCGA

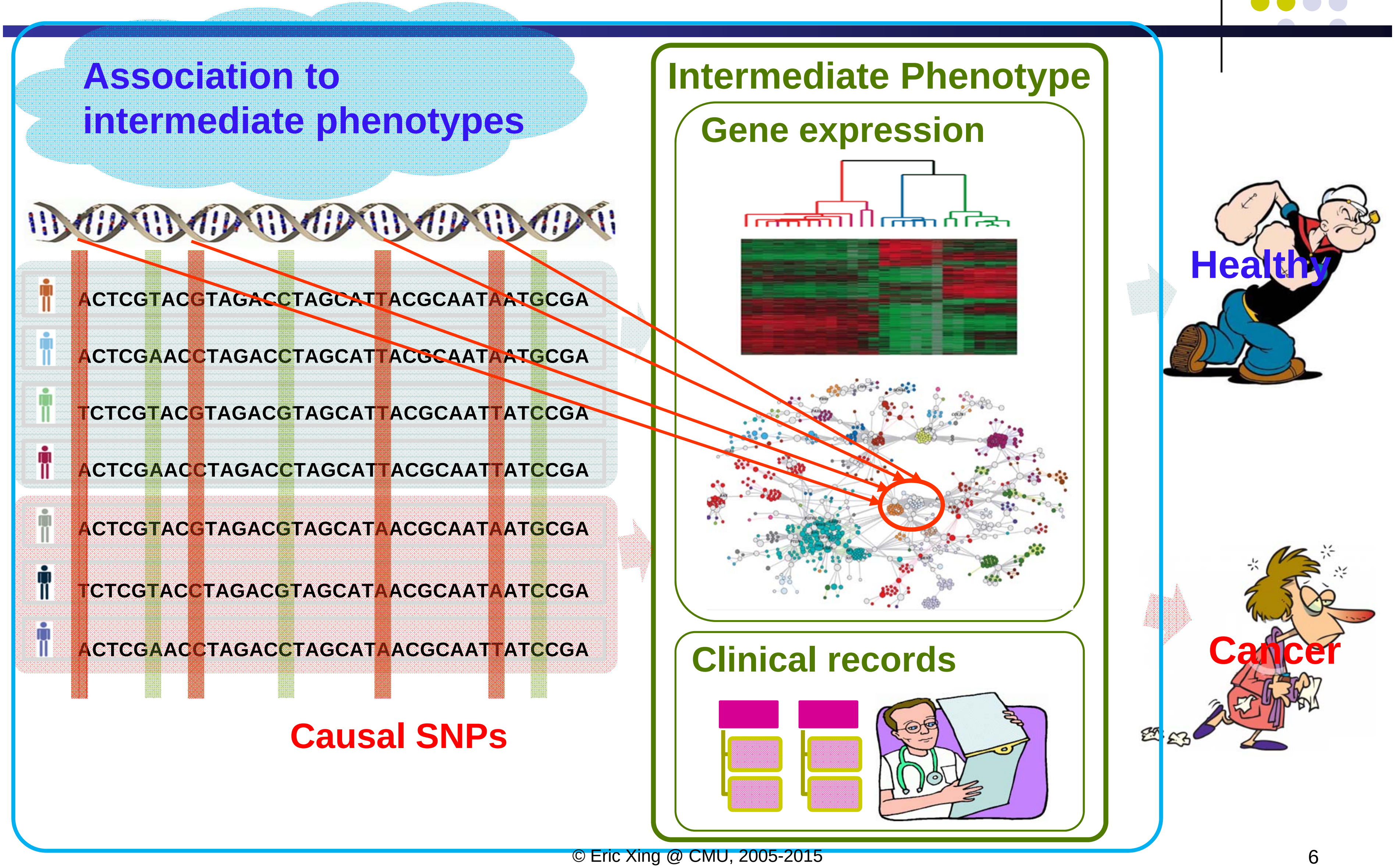
ACTCGAACCTAGACCTAGCATTACGCAATTATCCGA

ACTCGTACGTAGACGTAGCATAACGCAATAATGCGA

TCTCGTACCTAGACGTAGCATAACGCAATAATCCGA

ACTCGAACCTAGACCTAGCATAACGCAATTATCCGA

Genetic Basis of Complex Diseases



Association to intermediate phenotypes

ACTCGTACGTAGACCTAGCATTACGCAATAATGCGA

ACTCGAACCTAGACCTAGCATTACGCAATAATGCGA

TCTCGTACGTAGACGTAGCATTACGCAATTATCCGA

ACTCGAACCTAGACCTAGCATTACGCAATTATCCGA

ACTCGTACGTAGACGTAGCATAACGCAATAATGCGA

TCTCGTACCTAGACGTAGCATAACGCAATAATCCGA

ACTCGAACCTAGACCTAGCATAACGCAATTATCCGA

Causal SNPs

Intermediate Phenotype

Gene expression

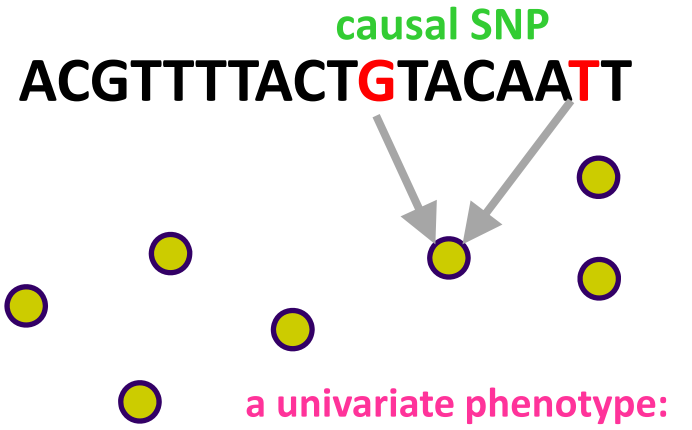
Clinical records

Healthy

Cancer

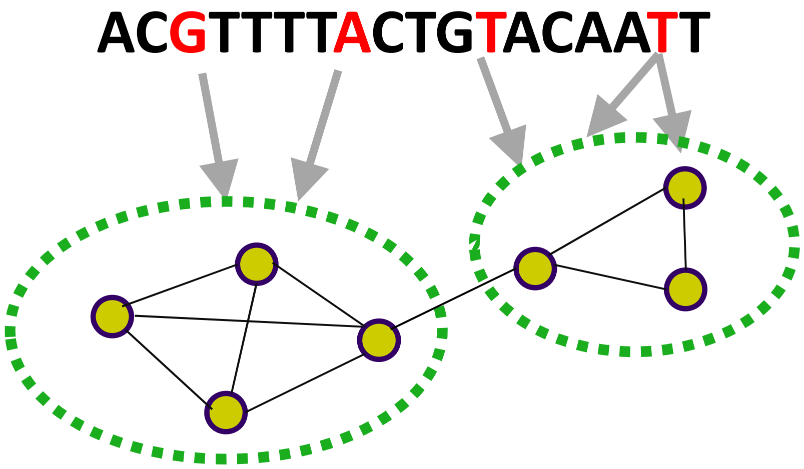
Structured Association

Traditional Approach



a univariate phenotype:
gene expression level

Association with Phenome



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

Goal:
Inferring Structured Association

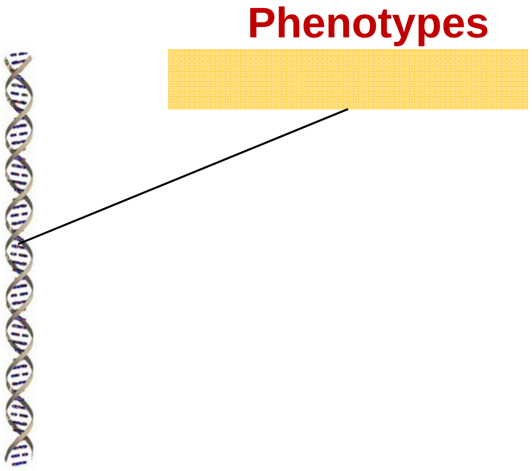
Standard Approach

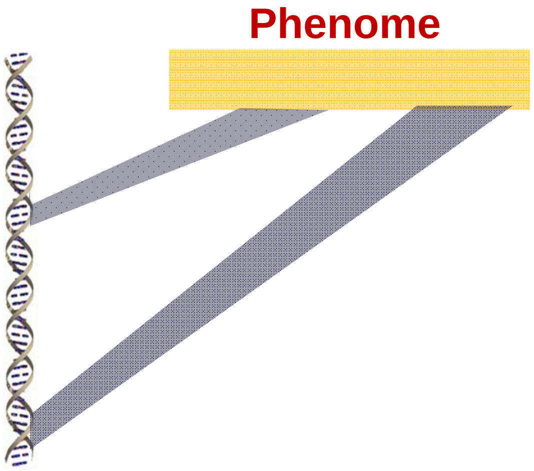
Consider
one phenotype & one genotype at a time

vs.

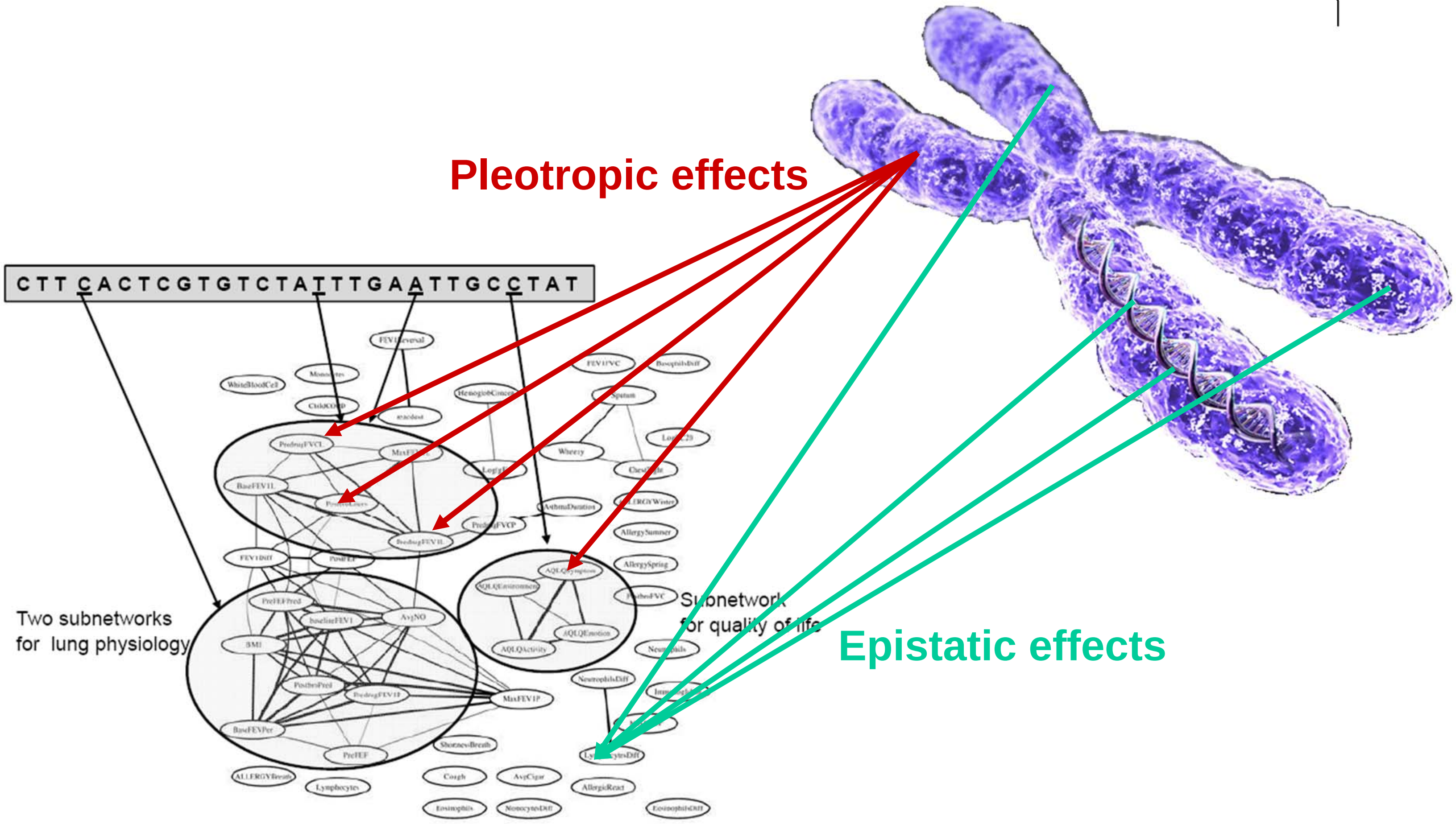
New Approach

Consider
multiple correlated phenotypes & genotypes jointly





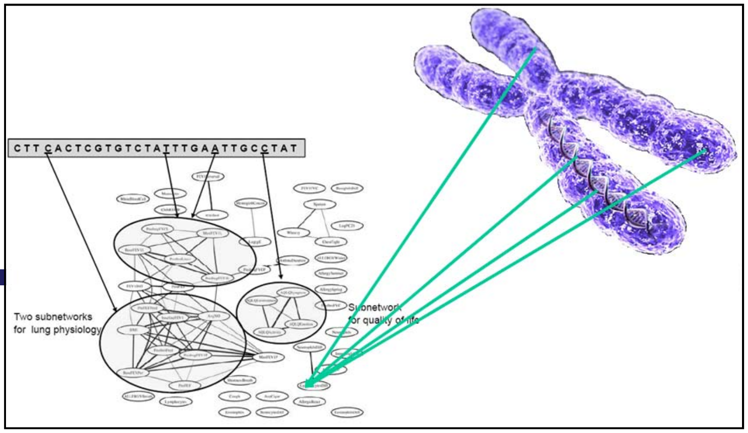
Sparse Associations



Pleotropic effects

Epistatic effects

Sparse Learning


CTT CACTCGTGTCTATTTGAATTGCCTAT
Two subnetworks for lung physiology
Subnetwork for quality of life

- 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]

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

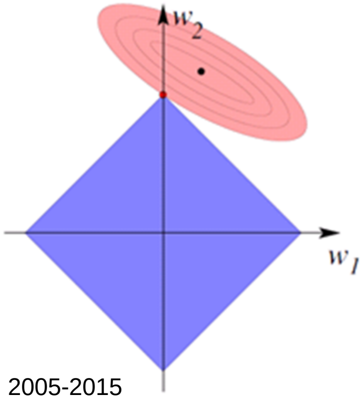
$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\beta_j|$$

- Why sparse solution?

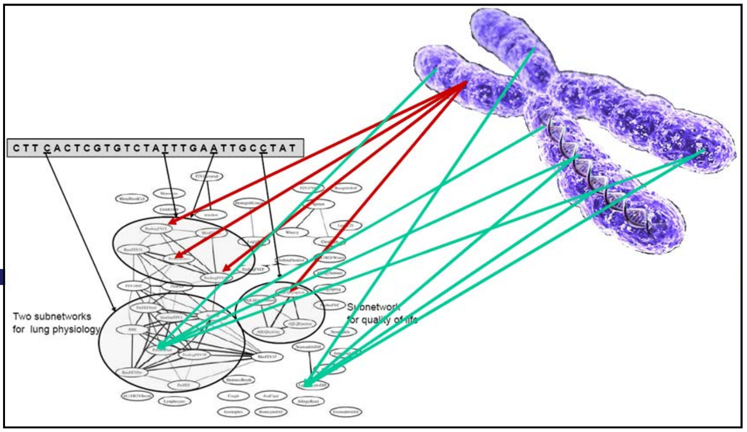
penalizing $\lambda\|\boldsymbol{\beta}\|_1$

↕

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



Multi-Task Extension


CTTCACTCGTGTCTATTTGAATTGCCTAT
Two subnetworks for lung physiology
Subnetwork for quality of life

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

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

Coefficients for a variable (2^nd^)

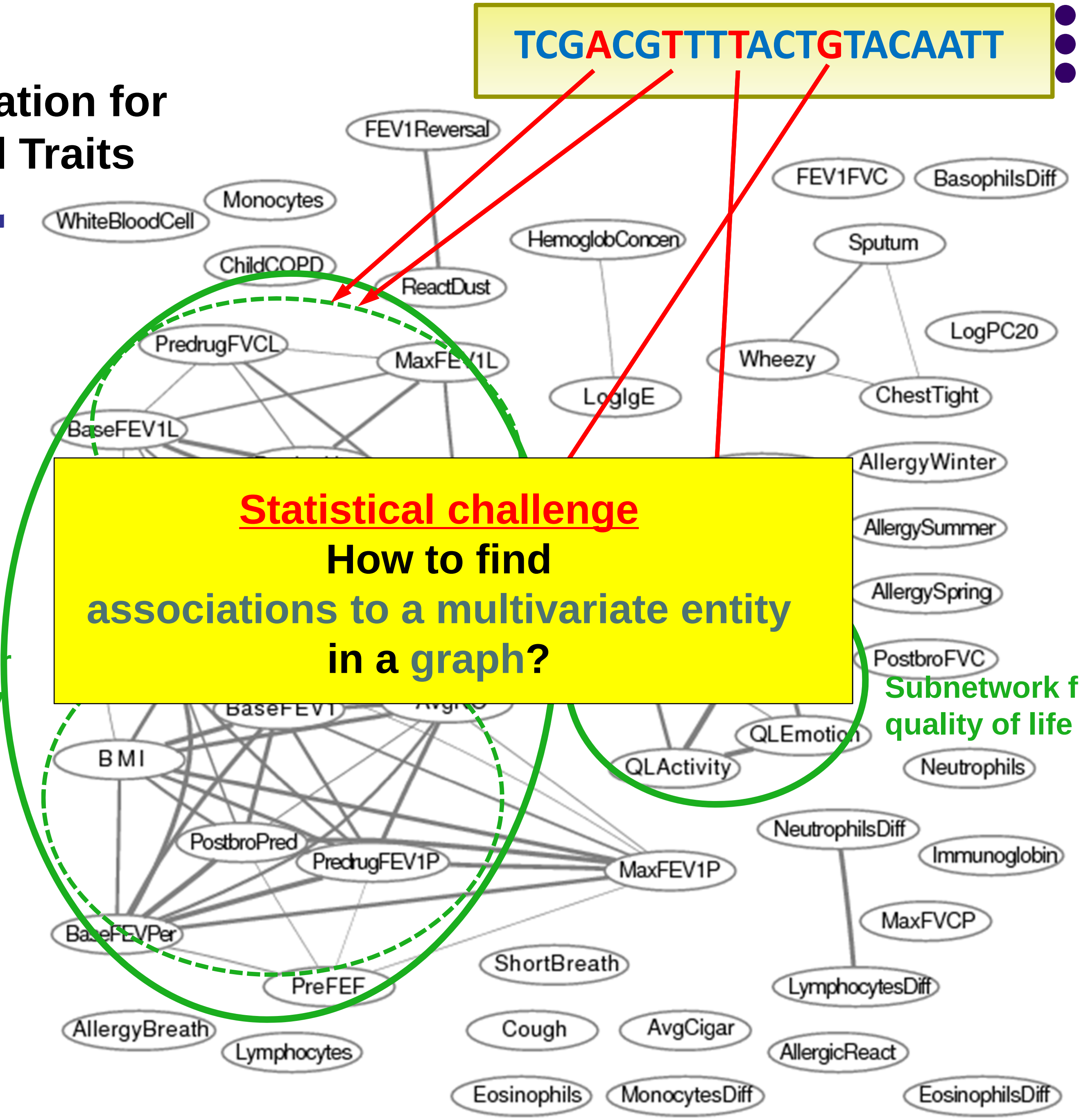
Coefficients for a task (2^nd^)

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)
  - Association between a **subnetwork** of genome and a **subnetwork** of phenme
    - Two-graph guided multi-task lasso (Chen et al., AISTATS 2012)

Genetic Association for Asthma Clinical Traits



Subnetworks for lung physiology

Subnetwork for quality of life

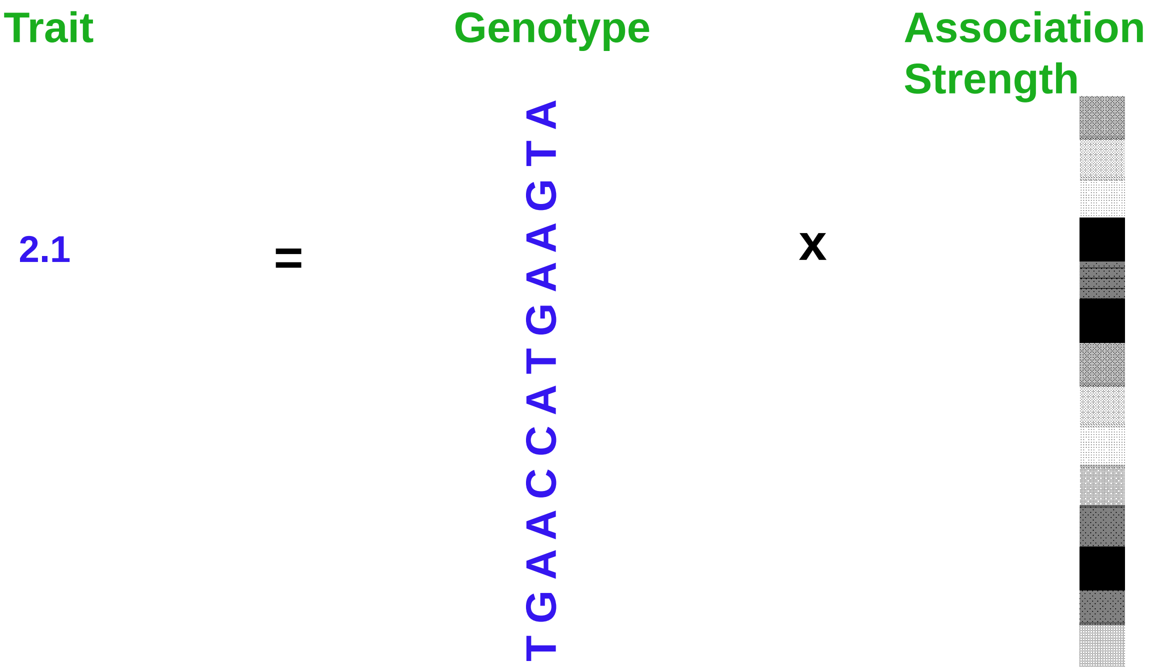
Statistical challenge

How to find associations to a multivariate entity in a graph?

Multivariate Regression for Single-Trait Association Analysis

Trait		Genotype		Association Strength
2.1	=	TGAACCATGAAGTA	x	?
y	=	X	x	β

Multivariate Regression for Single-Trait Association Analysis

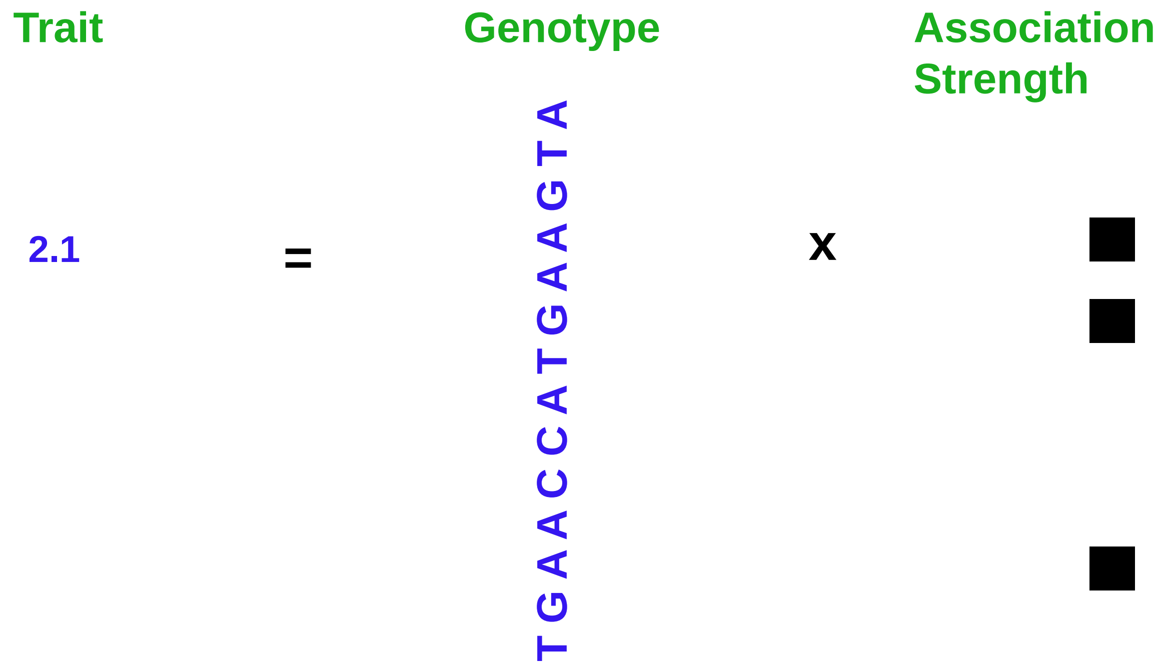


$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

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

Lasso for Reducing False Positives

(Tibshirani, 1996)

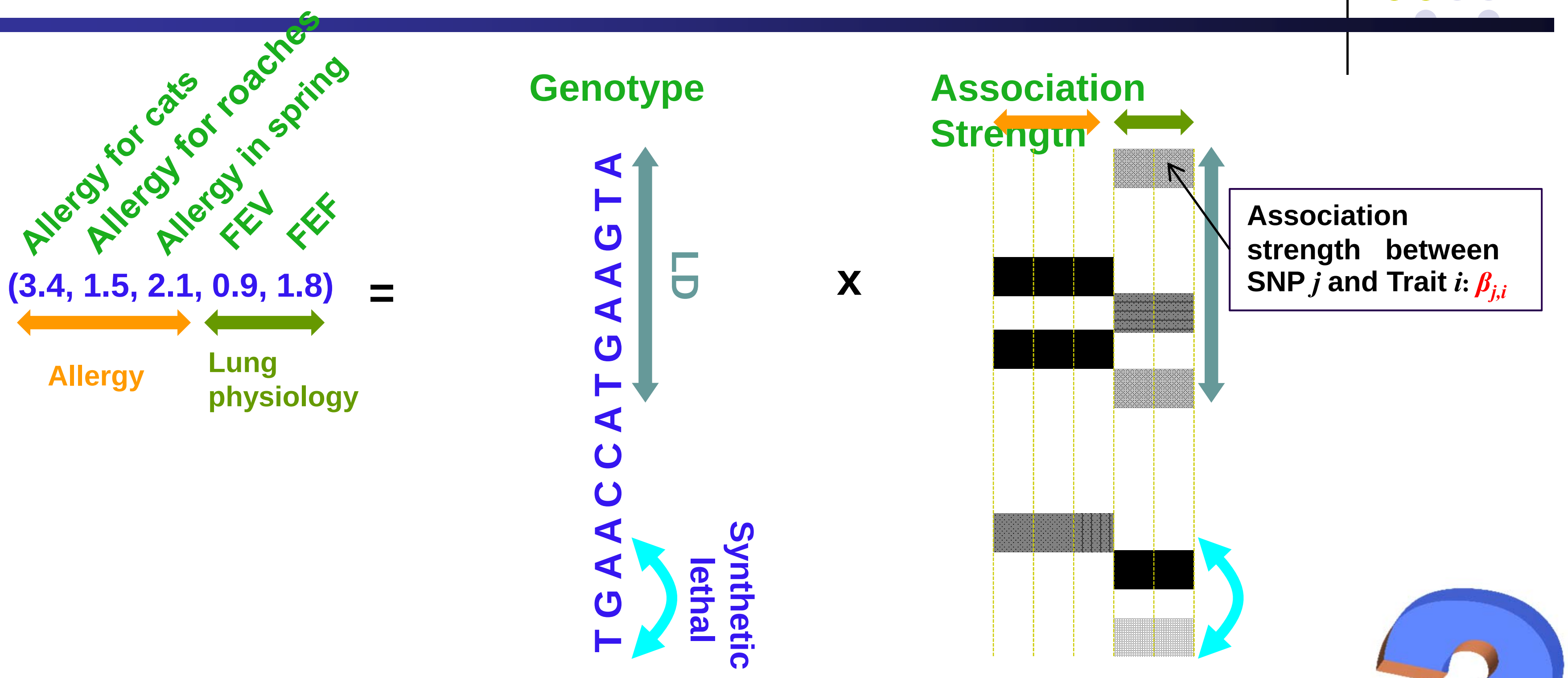


Lasso Penalty for sparsity

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

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

Multivariate Regression for Multiple-Trait Association Analysis



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

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

Multivariate Regression for Multiple-Trait Association Analysis

Trait

Genotype

Association Strength

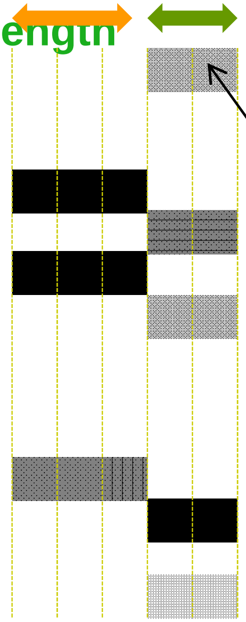
(3.4, 1.5, 2.1, 0.9, 1.8) =

Allergy

Lung physiology

TGAACCATGAAGTA

x



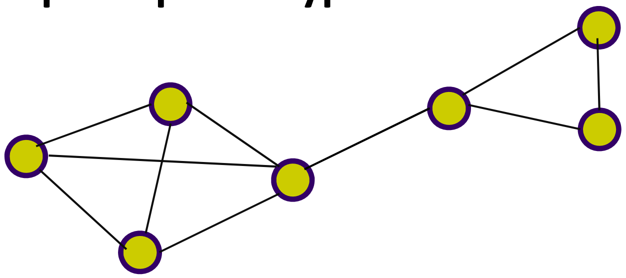
Association strength between SNP *j* and Trait *i*: $\beta_{j,i}$

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

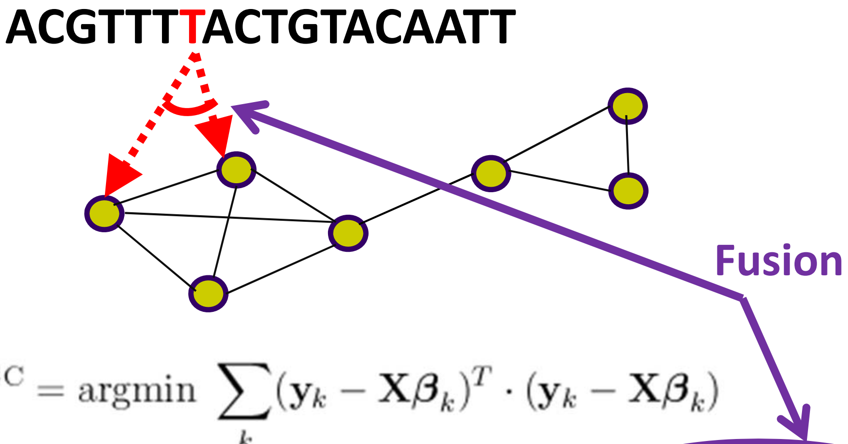
+ We introduce **graph-guided fusion penalty**

Multiple-trait Association: Graph-Constrained Fused Lasso

Step 1: Thresholded correlation graph of phenotypes



Step 2: Graph-constrained fused lasso

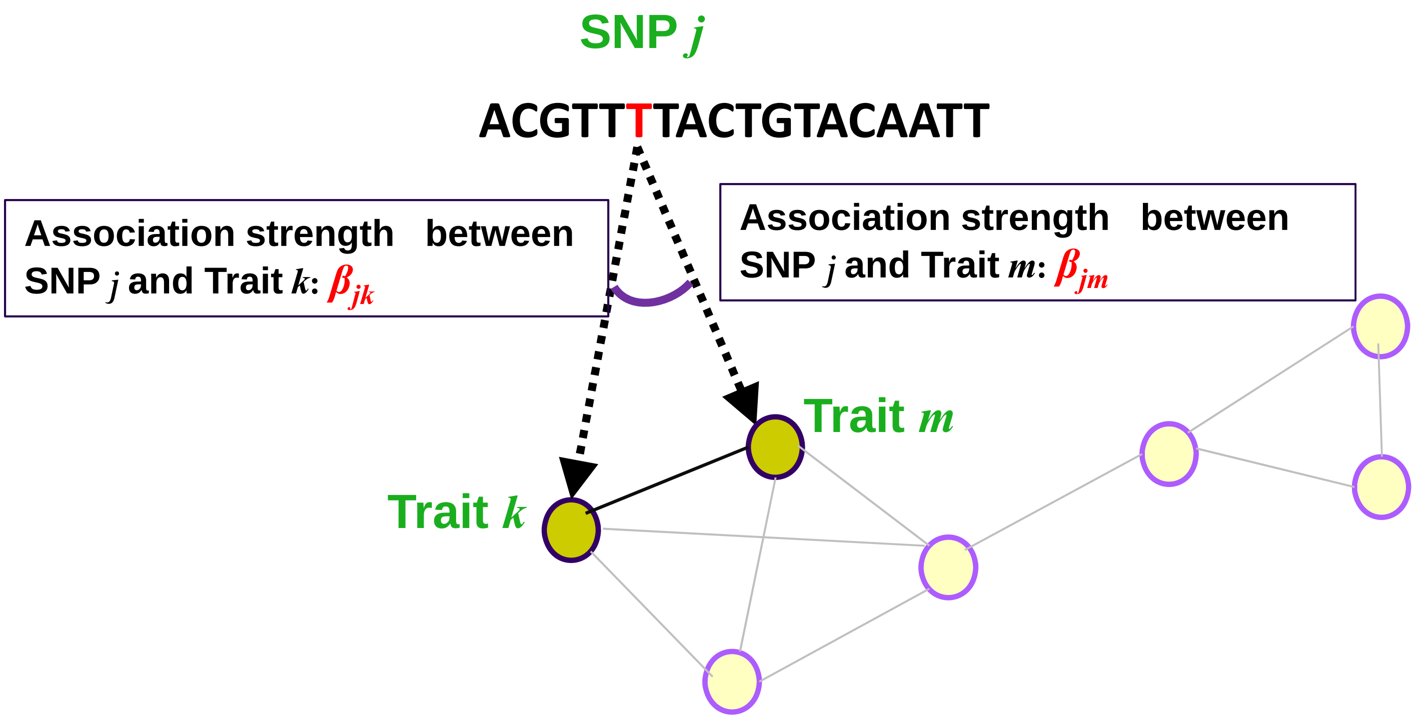


$$\hat{\mathbf{B}}^{\mathrm{GC}} = \operatorname{argmin} \sum_k (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k)^T \cdot (\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k) + \lambda \sum_k \sum_j |\beta_{jk}| + \gamma \sum_{(m,l)\in E} \sum_j |\beta_{jm} - \mathrm{sign}(r_{ml})\beta_{jl}|$$

Lasso Penalty

Graph-constrained fusion penalty

Fusion Penalty



SNP j

ACGTTTTACTGTACAATT

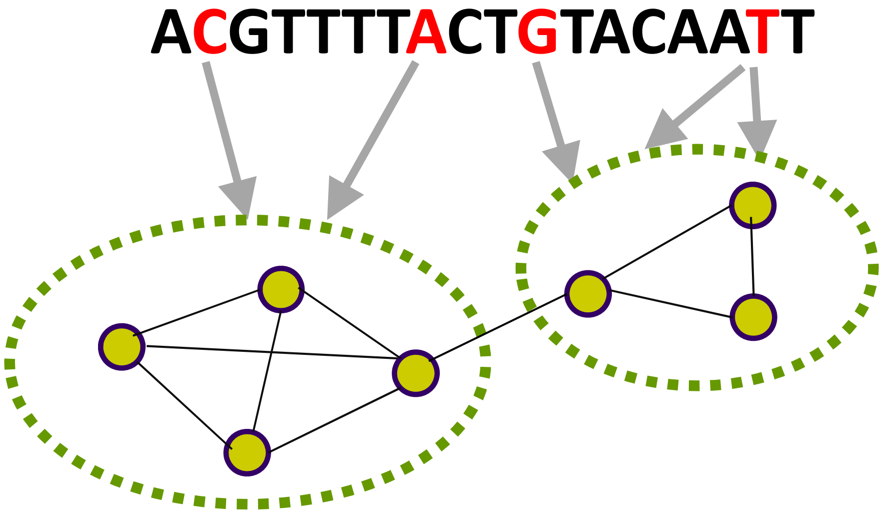
Association strength between SNP j and Trait k: β_{jk}

Association strength between SNP j and Trait m: β_{jm}

- 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

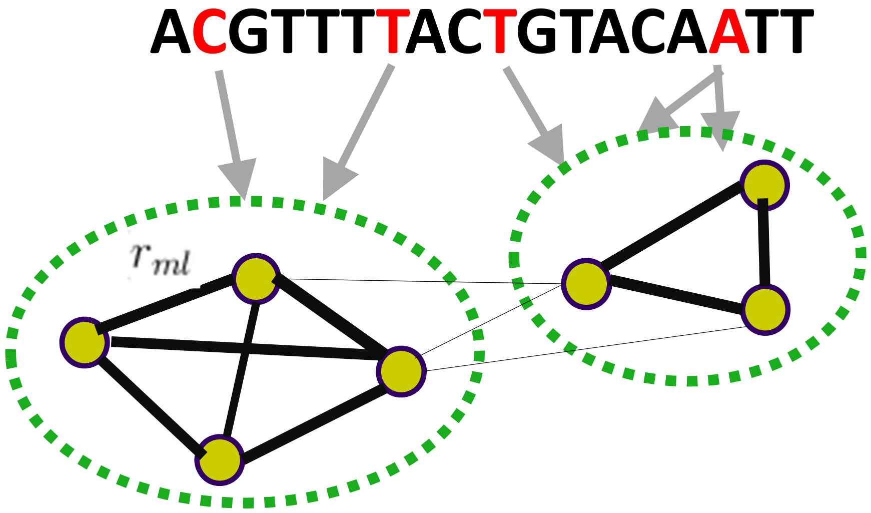


ACGTTTTACTGTACAATT

- 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

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

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

  - Graph-weighted fused lasso

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

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

- 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

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

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

$$d_{jk} \geq 0 \text{ for all } j,\ k,$$

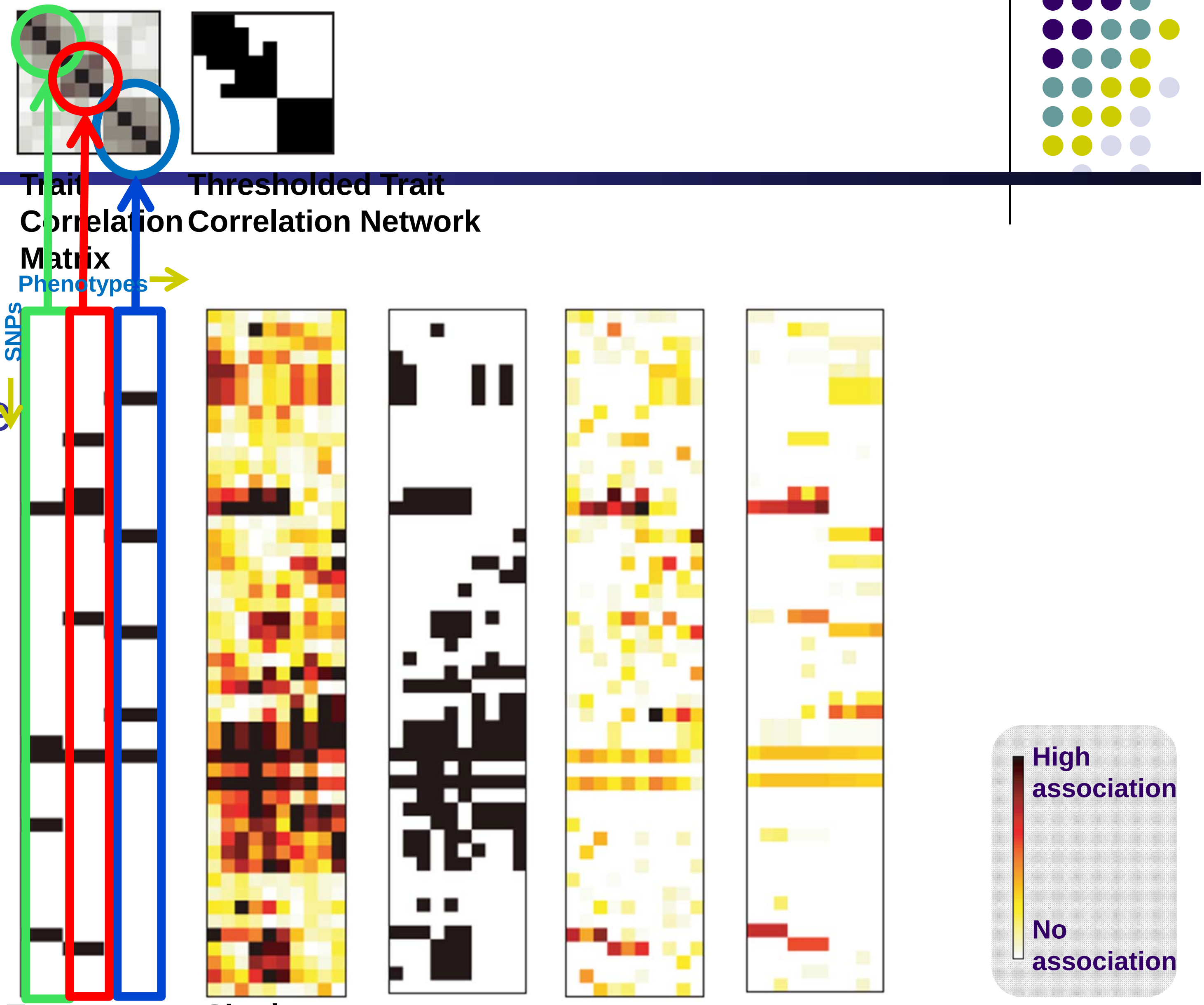
$$d_{jml} \geq 0 \text{ for all } j,\ (m,l) \in E,$$

Iterative optimization

- **Update β_k**
- **Update d_{jk}'s, d_{jml}'s**

Simulation Results

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



True Regression Coefficients

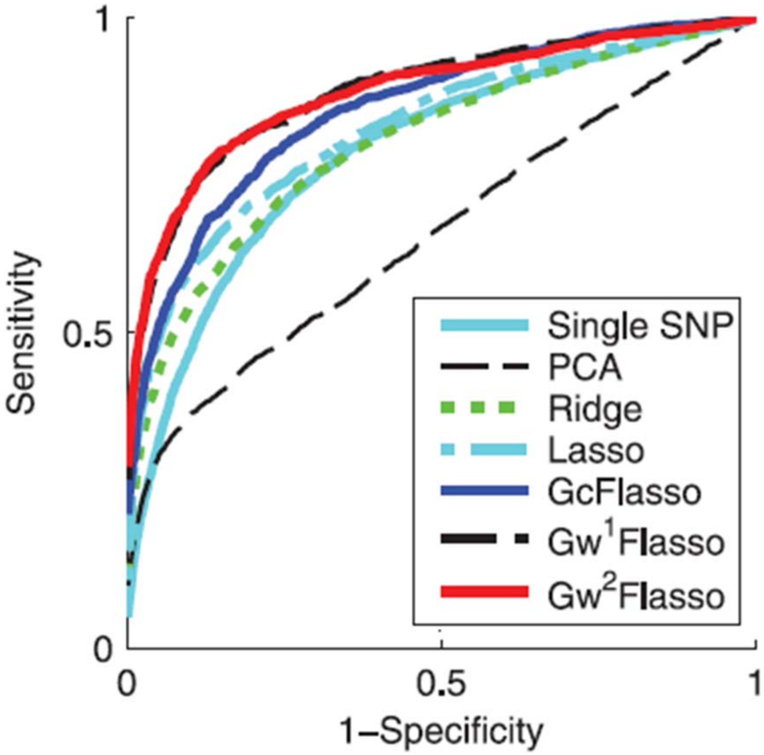
Single SNP-Single Trait Test

Significant at $\alpha = 0.01$

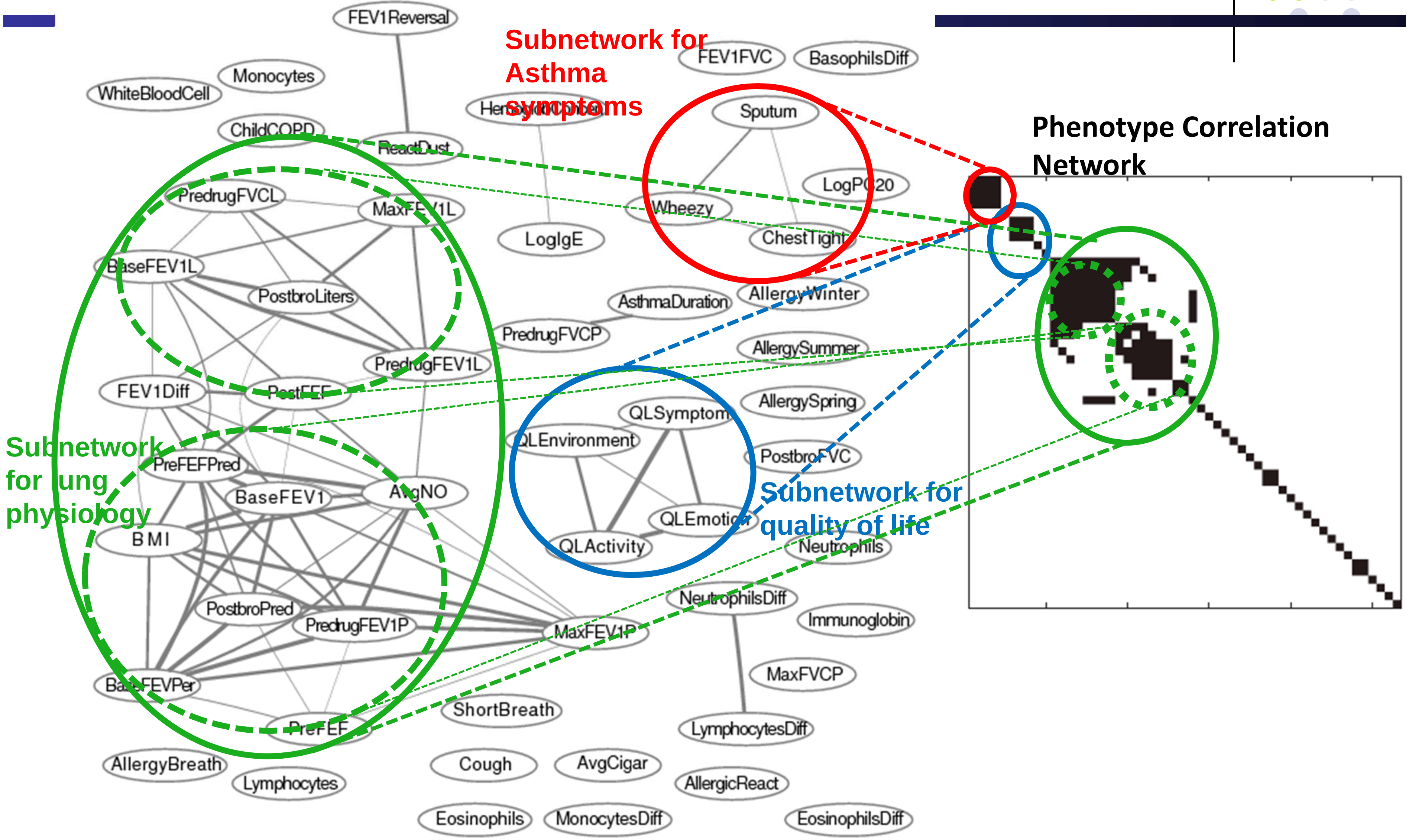
Lasso

Graph-guided Fused Lasso

Simulation Results



Asthma Trait Network



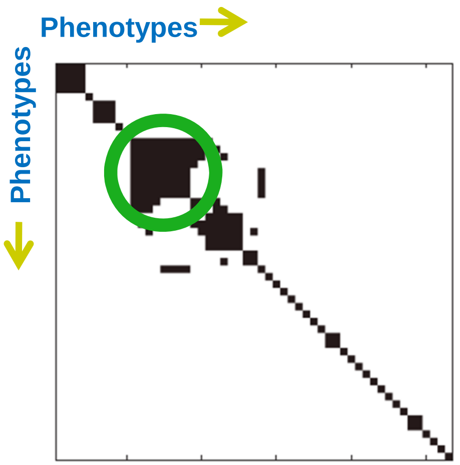
Subnetwork for Asthma symptoms

Phenotype Correlation Network

Subnetwork for lung physiology

Subnetwork for quality of life

Results from Single-SNP/Trait Test



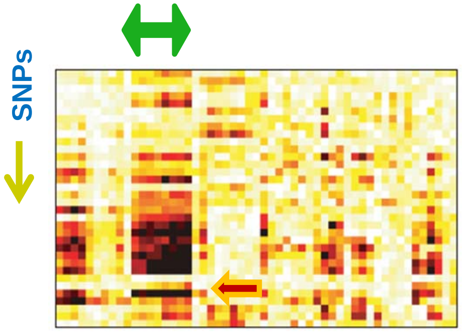
Trait Network

Lung physiology-related traits I

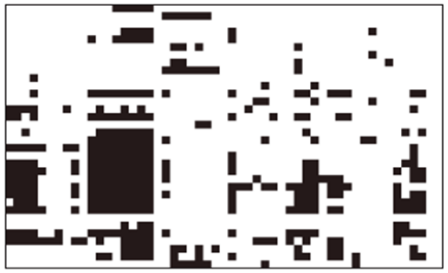
- Baseline FEV1 predicted value: MPVLung
- Pre FEF 25-75 predicted value
- Average nitric oxide value: online
- Body Mass Index
- Postbronchodilation FEV1, liters: Spirometry
- Baseline FEV1 % predicted: Spirometry
- Baseline predrug FEV1, % predicted
- Baseline predrug FEV1, % predicted

Q551R SNP

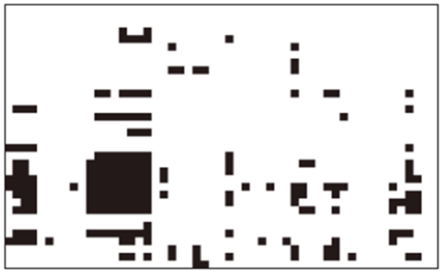
- Codes for amino-acid changes in the intracellular signaling portion of the receptor
- Exon 11



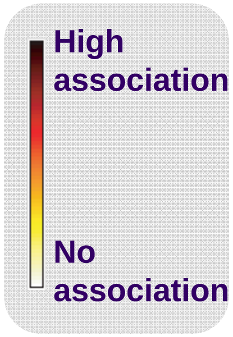
Single-Marker Single-Trait Test



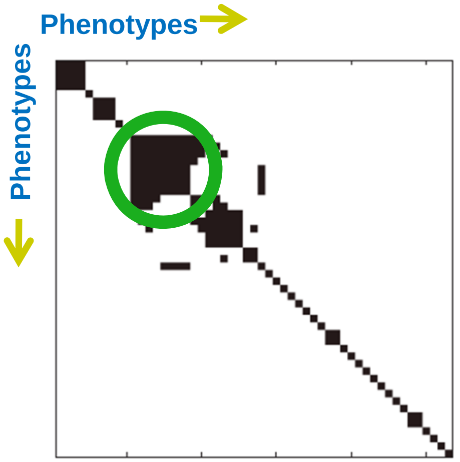
Permutation test α = 0.05



Permutation test α = 0.01



Comparison of Gflasso with Others



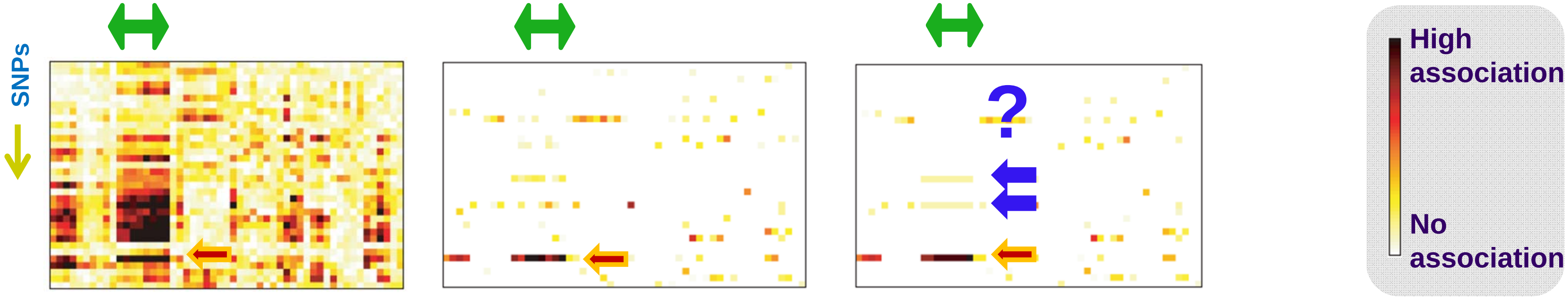
Trait Network

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

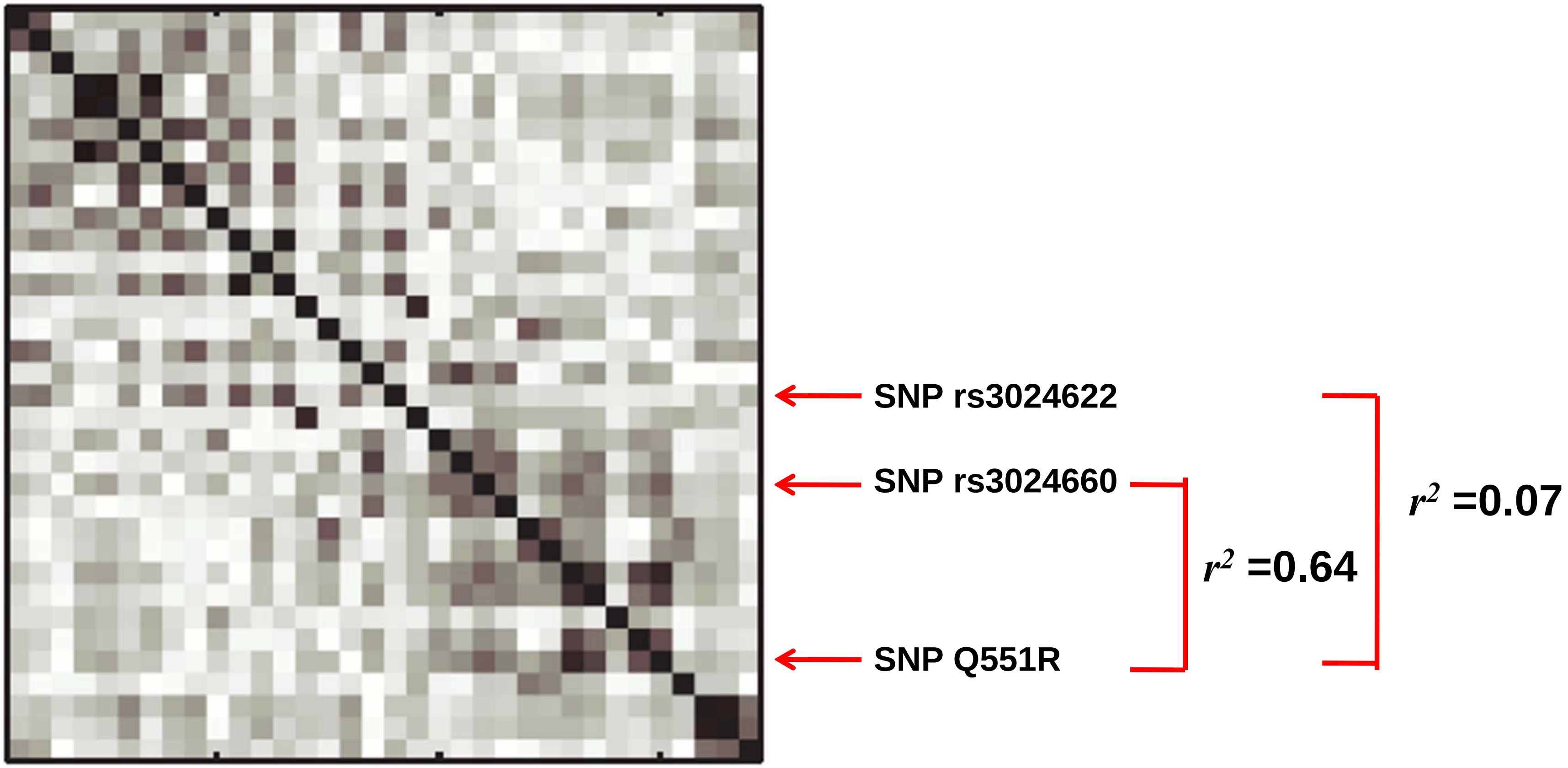


Single-Marker Single-Trait Test

Lasso

Graph-guided Fused Lasso

Linkage Disequilibrium Structure in *IL-4R* gene



SNP rs3024622

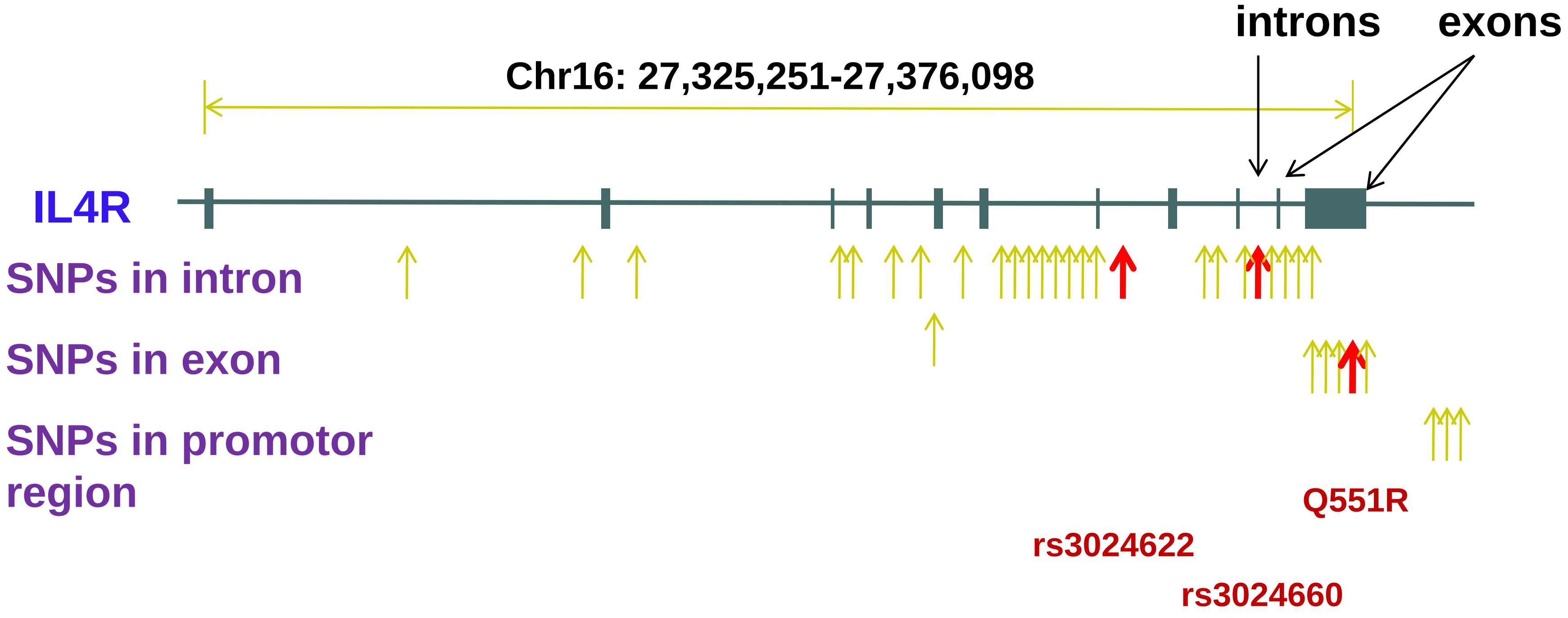
SNP rs3024660

SNP Q551R

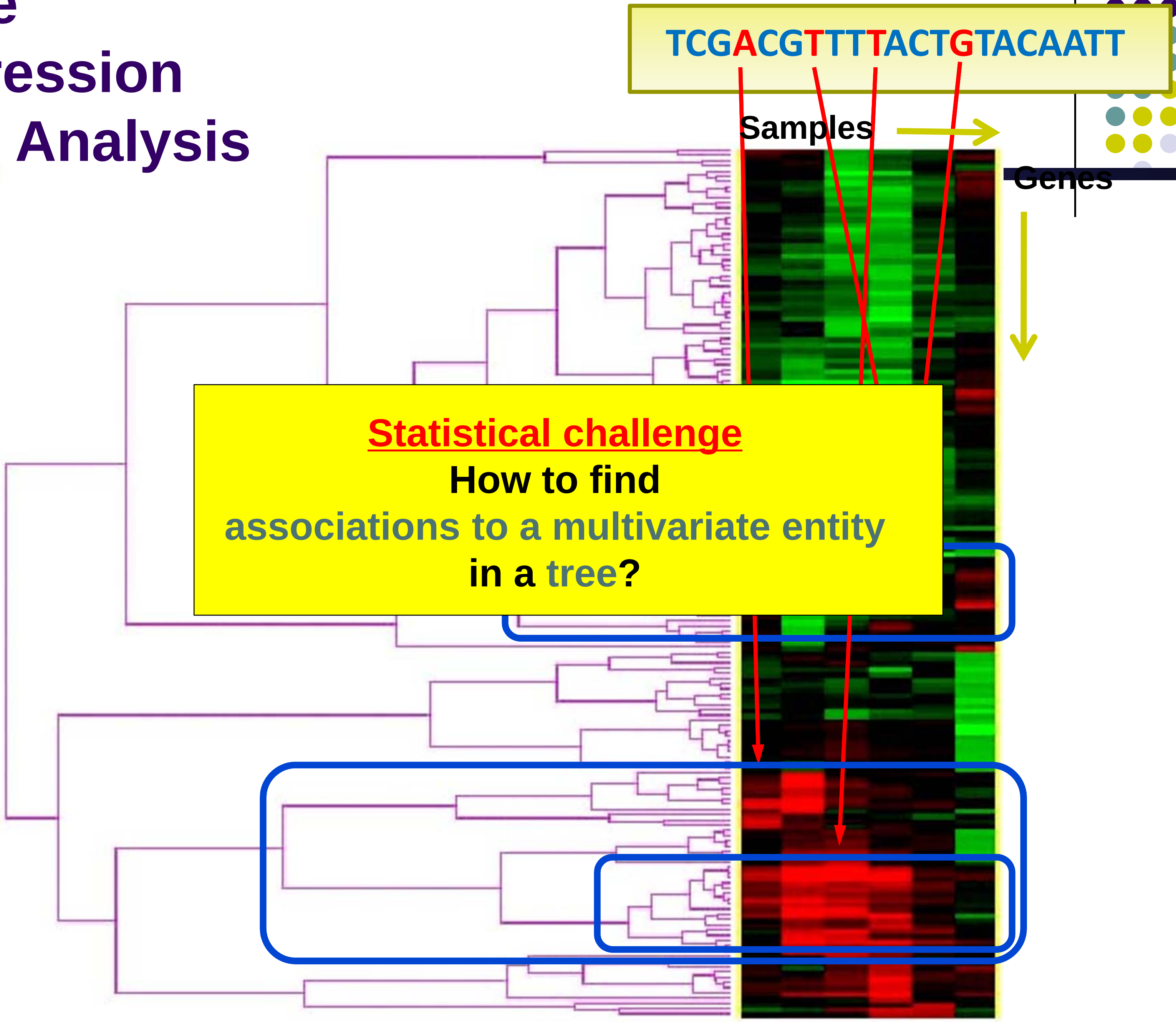
r^2 =0.64

r^2 =0.07

IL4R Gene


introns
exons
Chr16: 27,325,251-27,376,098
IL4R
SNPs in intron
SNPs in exon
SNPs in promotor region
Q551R
rs3024622
rs3024660

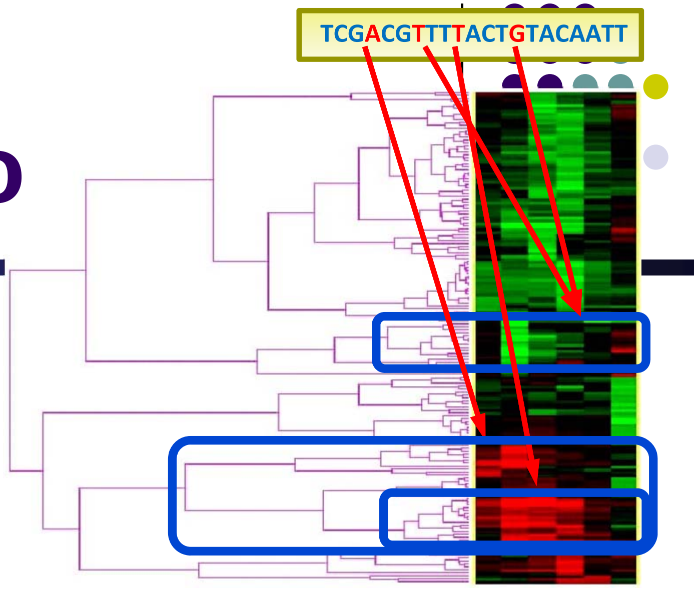
Gene Expression Trait Analysis



Statistical challenge

How to find associations to a multivariate entity in a tree?

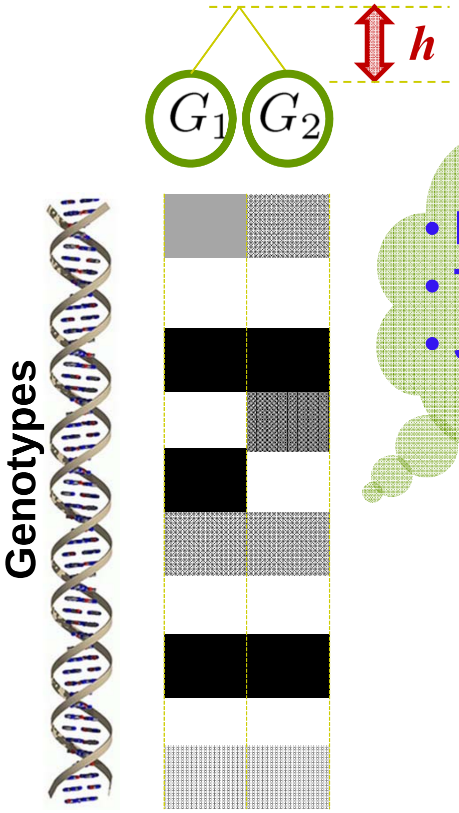
Tree-guided Group Lasso



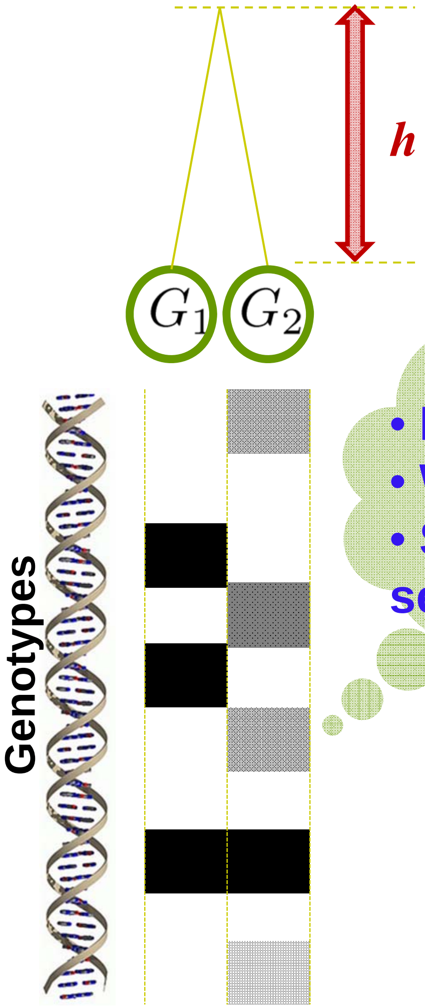
- Why tree?
  - Tree represents a clustering structure
  - Scalability to a very large number of phenotypes
    - Graph : $O(|V|^2)$ edges
    - Tree : $O(|V|)$ edges
  - Expression quantitative trait mapping (eQTL)
    - Agglomerative hierarchical clustering is a popular tool

Tree-Guided Group Lasso

- In a simple case of two genes



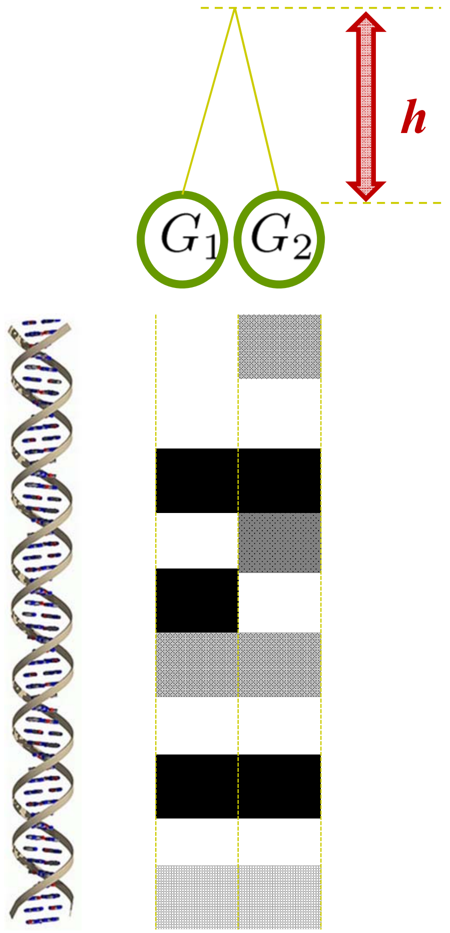
- Low height
- Tight correlation
- Joint selection

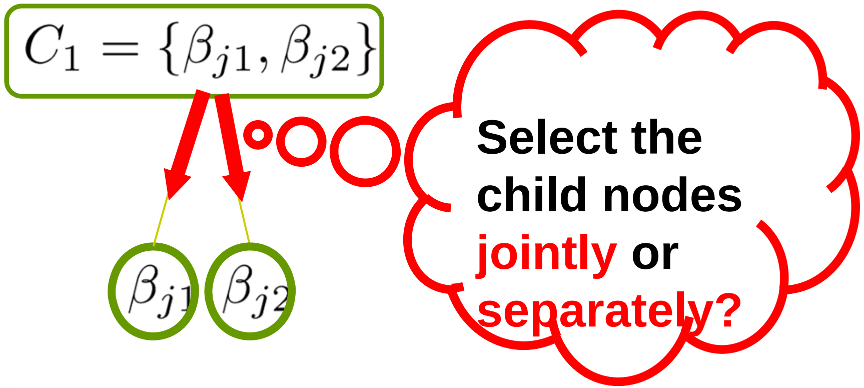


- Large height
- Weak correlation
- Separate selection

Tree-Guided Group Lasso

- In a simple case of two genes





Select the child nodes jointly or separately?

Tree-guided group lasso

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

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

L_1 **penalty**

- Lasso penalty
- Separate selection

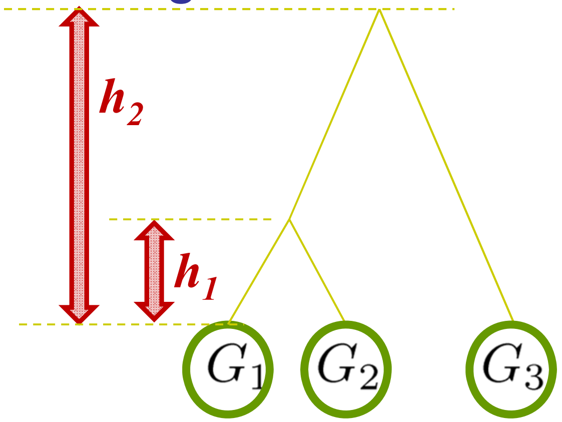
L_2 **penalty**

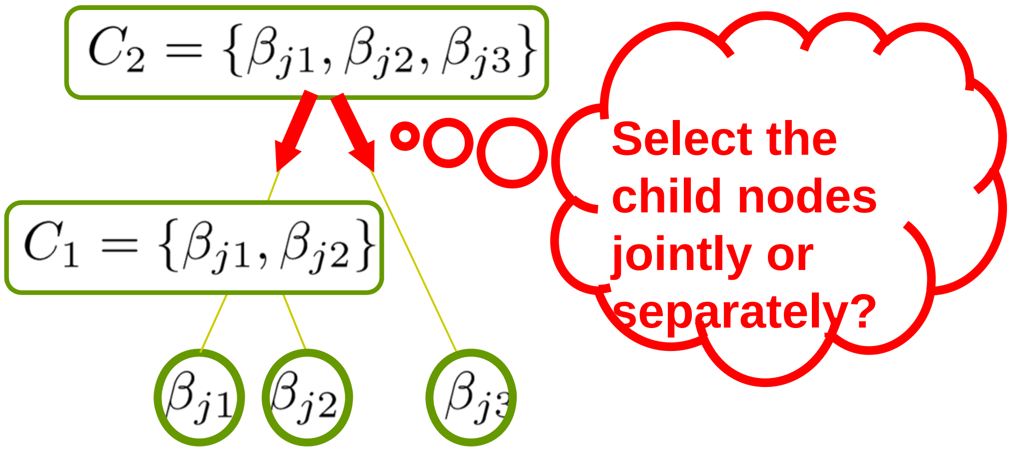
- Group lasso
- Joint selection

Elastic net

Tree-Guided Group Lasso

- For a general tree




$C_2 = \{\beta_{j1}, \beta_{j2}, \beta_{j3}\}$
$C_1 = \{\beta_{j1}, \beta_{j2}\}$
β_{j1}
β_{j2}
β_{j3}

Select the child nodes jointly or separately?

Tree-guided group lasso

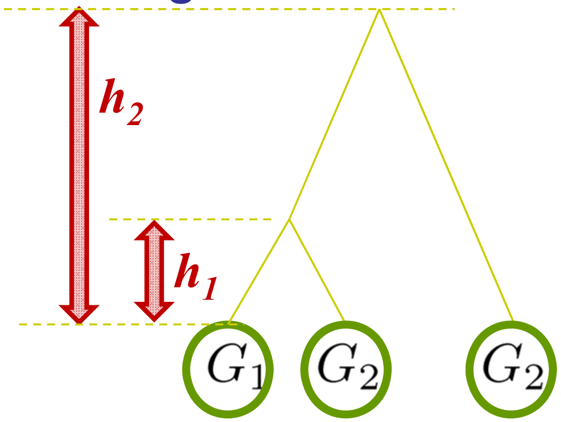
$$\operatorname{argmin} \ (y - X\beta)' \cdot (y - X\beta)$$
$$+\lambda \sum_j \left[(1 - h_2)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2}\right) + h_2\left(|C_1| + |\beta_{j3}|\right)\right]$$

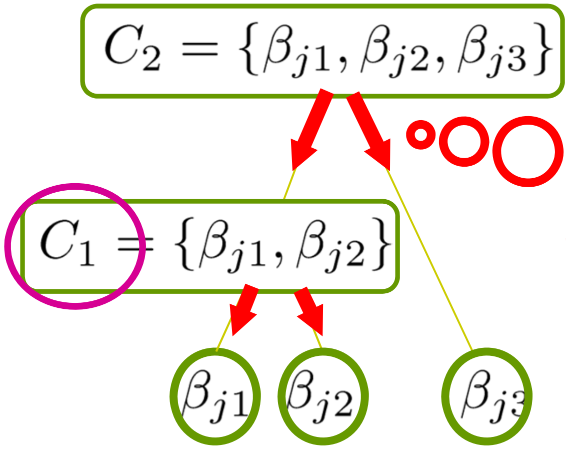
Joint selection

Separate selection

Tree-Guided Group Lasso

- For a general tree





$C_2 = \{\beta_{j1}, \beta_{j2}, \beta_{j3}\}$
$C_1 = \{\beta_{j1}, \beta_{j2}\}$
β_{j1}
β_{j2}
β_{j3}

Select the child nodes jointly or separately?

Tree-guided group lasso

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

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

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

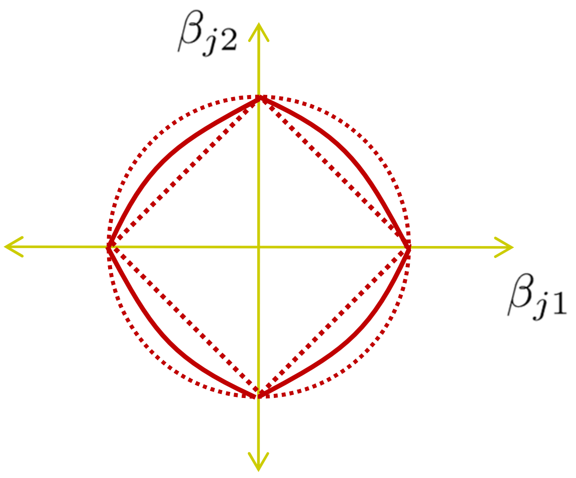
Joint selection

Separate selection

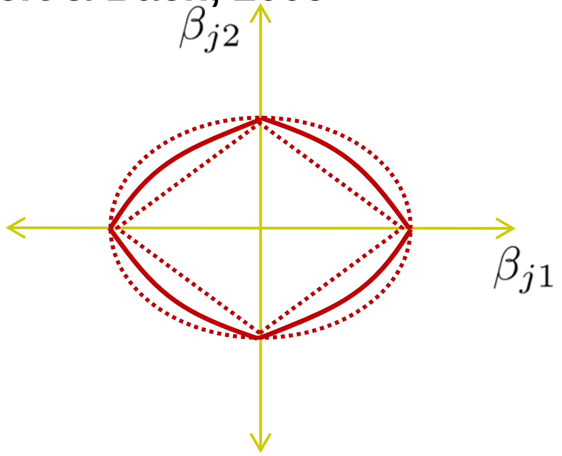
Balanced Shrinkage

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

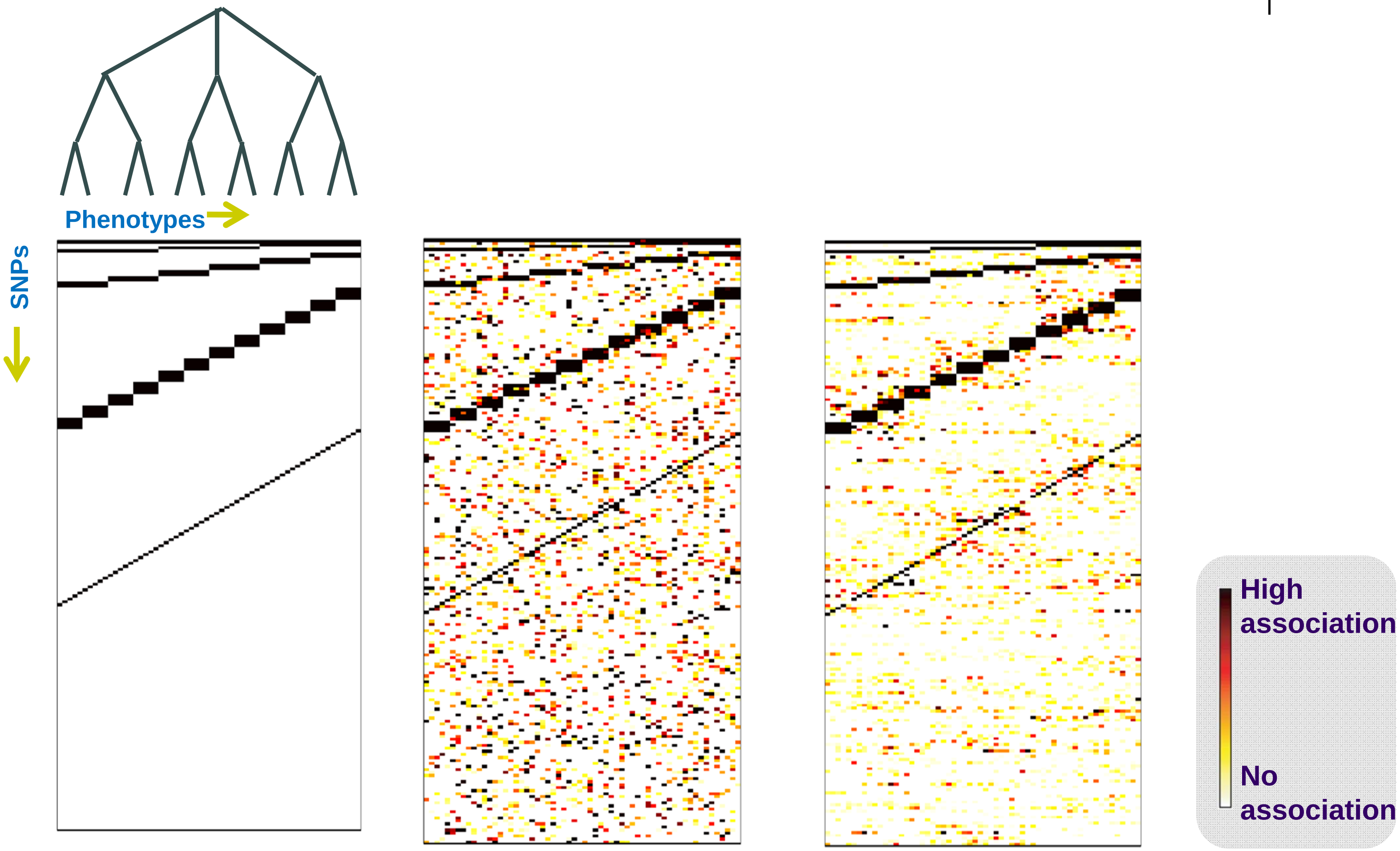
- Second-order cone program

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

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

- Also, variational formulation

Illustration with Simulated Data



True association strengths

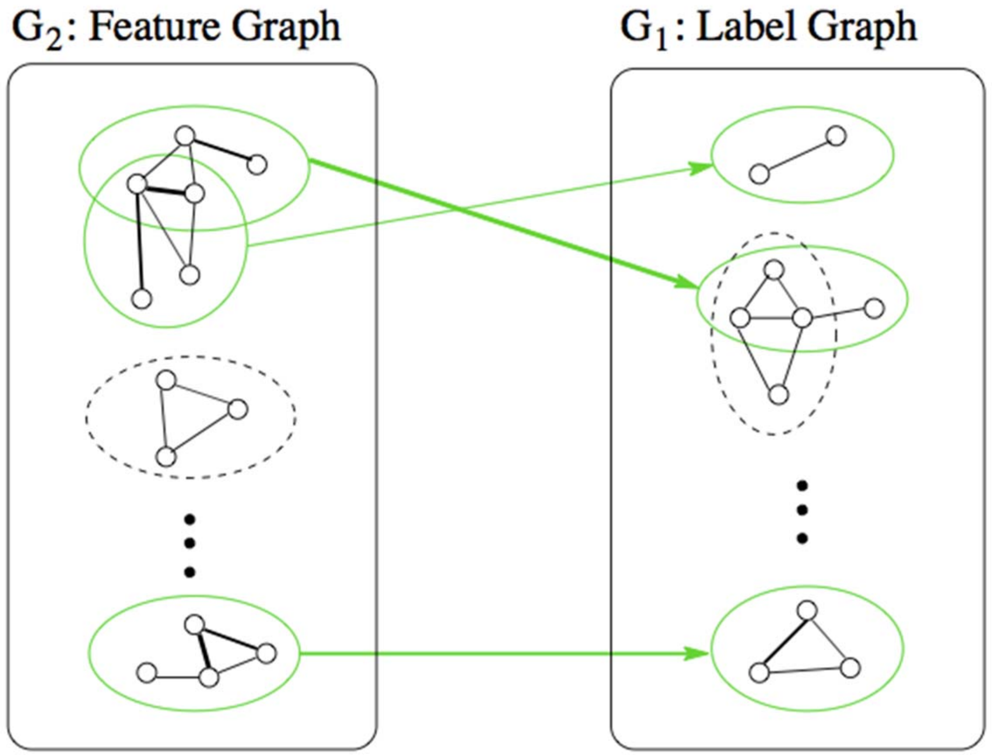
Lasso

Tree-guided group lasso

Incorporating Both Genome and Phenome Strucrues

- Find associations between subnetworks of genome and subnetworks of phenome

Genome structure



Phenome structure

Two-graph Guided Multi-task Lasso

- Motivated by graph structures in both genome and phenome
  - Gnome structure: pathway, linkage disequilibrium blocks
  - Phenome structure: trait networks

- How to take advantage of the two side information simultaneously?
  - Extend the graph-guided fused lasso to incorporate genome structures embedded in a graph
  - Use fusion penalty to use genome structures

Two-graph Guided Multi-task Lasso

$$\hat{B}^{\text{TCML}} = \text{argmin} \sum_k (y_k - \mathbf{X}\beta_k)^T (y_k - \mathbf{X}\beta_k) + \lambda \|\mathbf{B}\|_1 + \underbrace{\gamma_1 pen_1(E_1, \mathbf{B})}_{\textbf{Trait network}} + \underbrace{\gamma_2 pen_2(E_2, \mathbf{B})}_{\textbf{Genome network}}$$

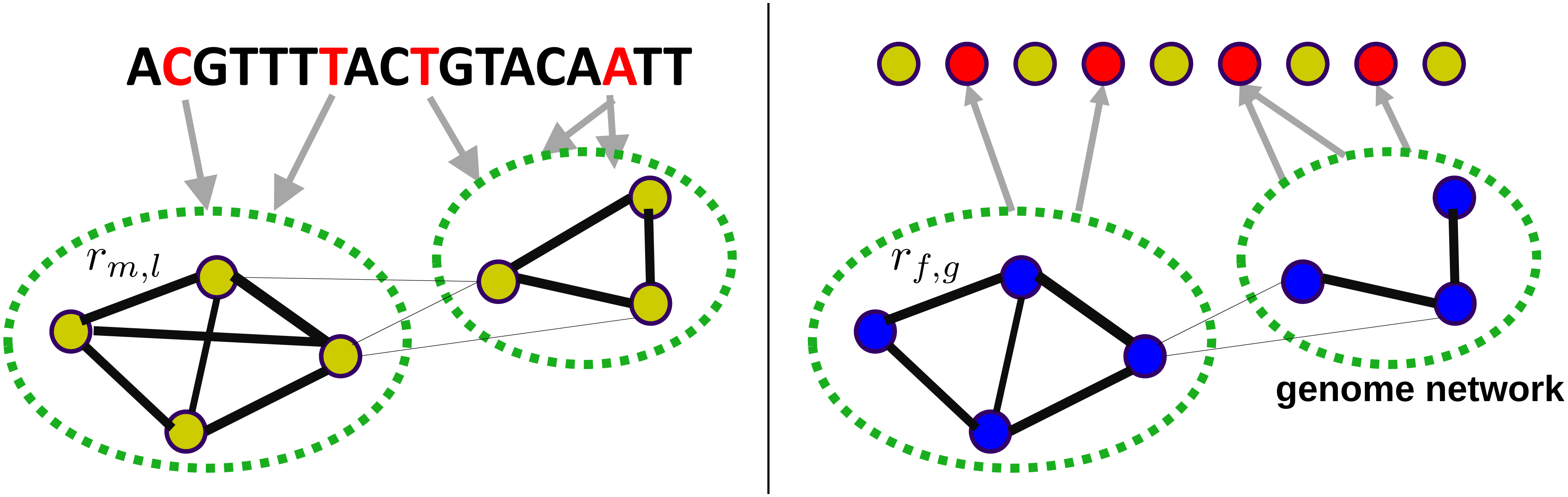
$$pen_1(E_1, B) = \sum_{e_{m,l} \in E_1} w(e_{m,l}) \sum_{j=1}^{J} |b_{jm} - \text{sign}(r_{m,l}) b_{jl}|$$

$$pen_2(E_2, B) = \sum_{e_{f,g} \in E_2} w(e_{f,g}) \sum_{k=1}^{K} |b_{fk} - \text{sign}(r_{f,g}) b_{gk}|,$$

- This model can be solved by using coordinate descent
- Similar to graph-guided fused Lasso, transform this objective into a differentiable function; then apply coordinate descent

Two-graph Guided Multi-task Lasso

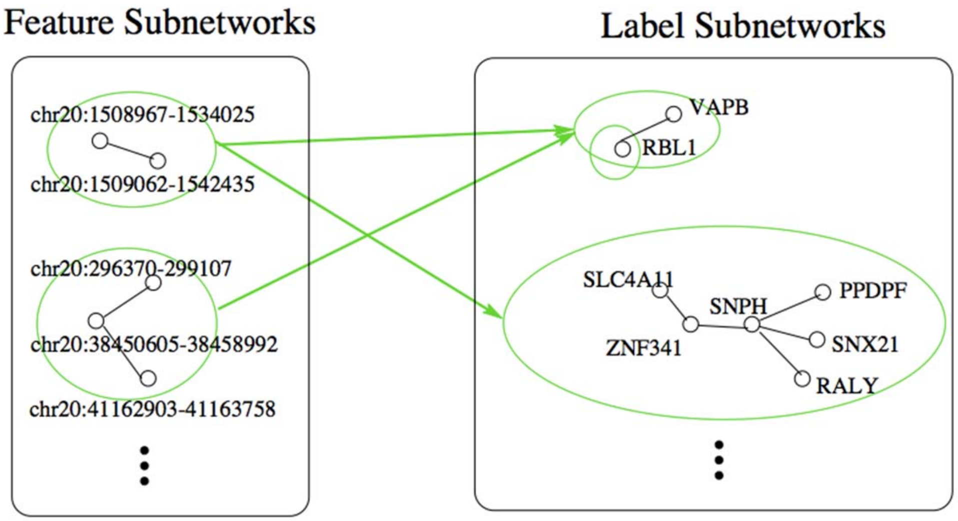
Overall effect



- Use both trait network and genome network simultaneously
  - Two traits connected in trait network are coupled though paths between the two nodes
  - Two SNPs connected in genome network are coupled through paths between the two nodes

Illustration with Simulated Data

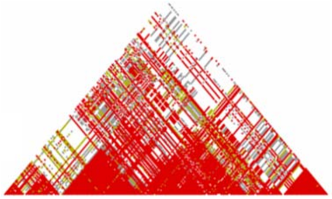
- eQTL Analysis using two-graph guided multi-task Lasso



- Genome: copy number variants from the latest release of the 1000 genome project
- Phenome: gene expression profiles from the RNA sequencing data

Genome Structure

Linkage Disequilibrium



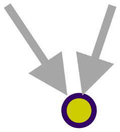
Stochastic block regression
(Kim & Xing, UAI, 2008)

Population Structure

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

Epistasis

ACGTTTTACTGTACAATT



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

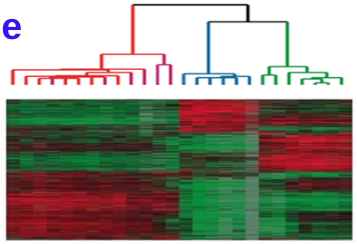
Structured Association

Phenome Structure

Graph

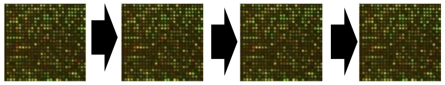
Graph-guided fused lasso
(Kim & Xing, PLoS Genetics, 2009)

Tree



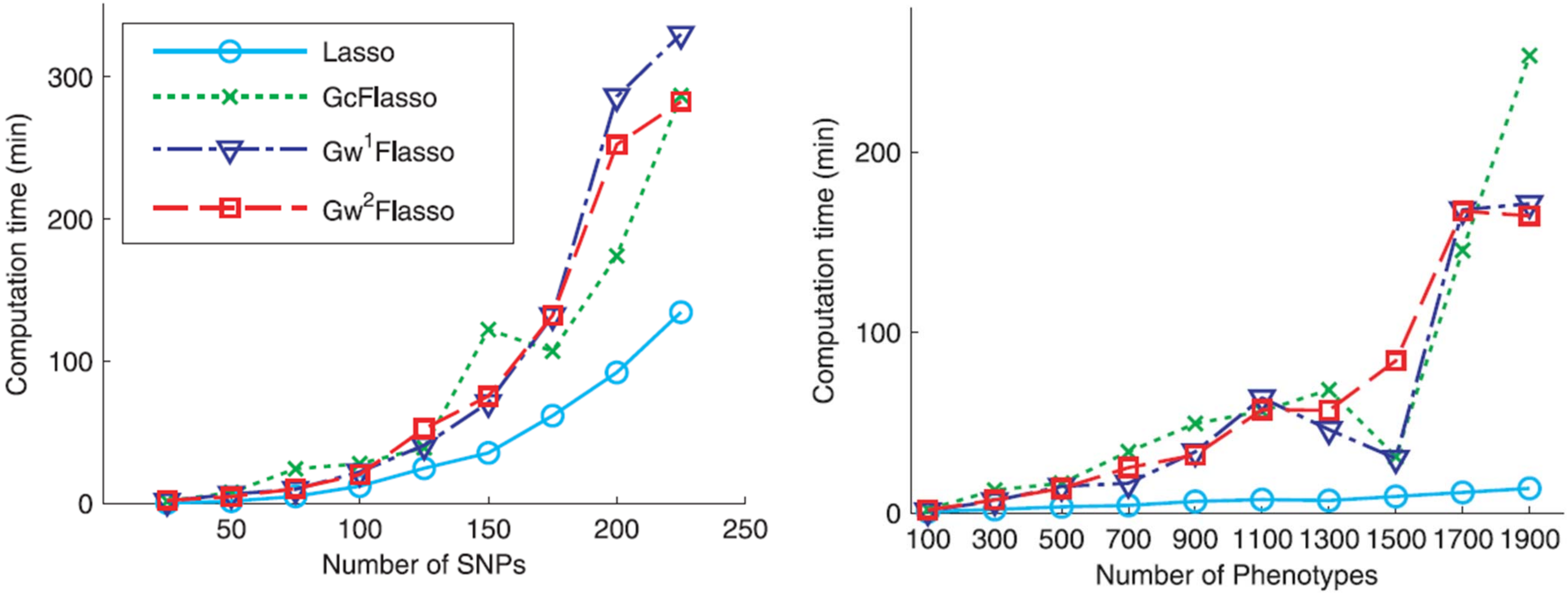
Tree-guided fused lasso
(Kim & Xing, Submitted)

Dynamic Trait

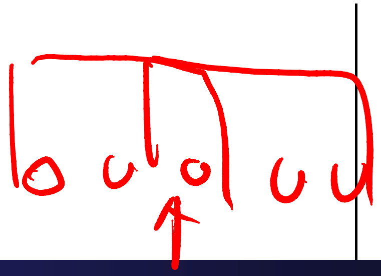


Temporally smoothed lasso
(Kim, Howrylak, Xing, Submitted)

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:

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

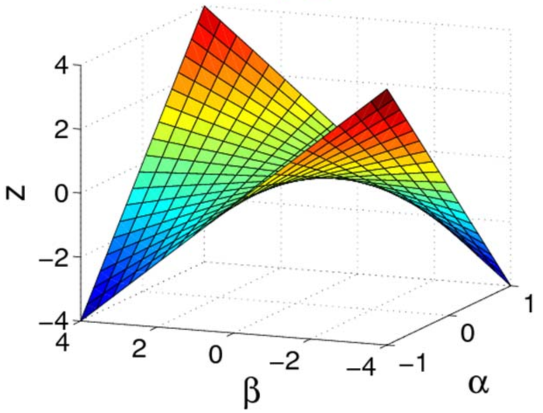
$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}\in\mathcal{Q}} \boldsymbol{\alpha}^T C\boldsymbol{\beta} - \mu d(\boldsymbol{\alpha})$$

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

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

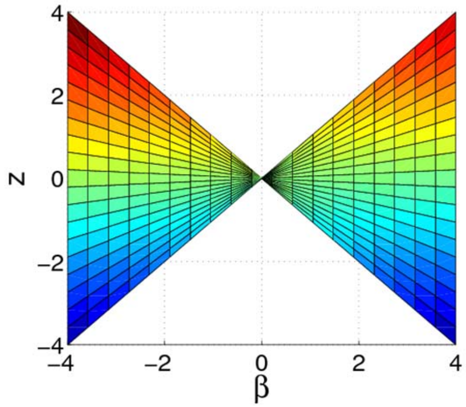
Geometric Interpretation

- Smooth approximation



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

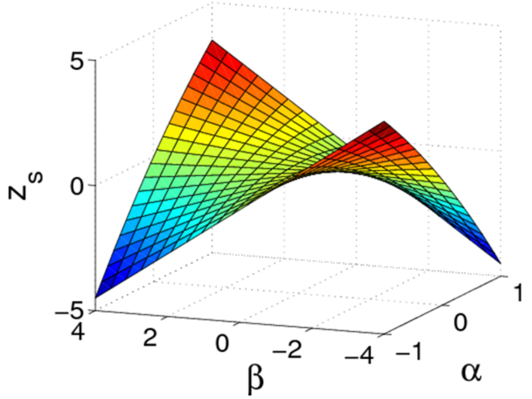
Projection onto
$z - \beta$ Plane



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

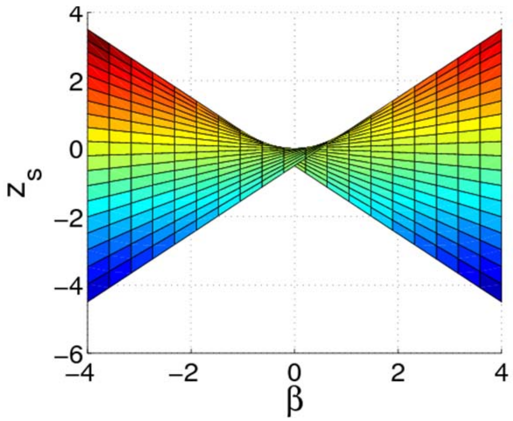
Uppermost Line Nonsmooth

☹



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

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

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

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: $\mathbf{X}^T\mathbf{X},\ \mathbf{X}^T\mathbf{Y}:\quad O(J^2N+JKN)$
- 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	$O(J^2K+J\sum_{g\in\mathcal{G}}\lvert g\rvert)$

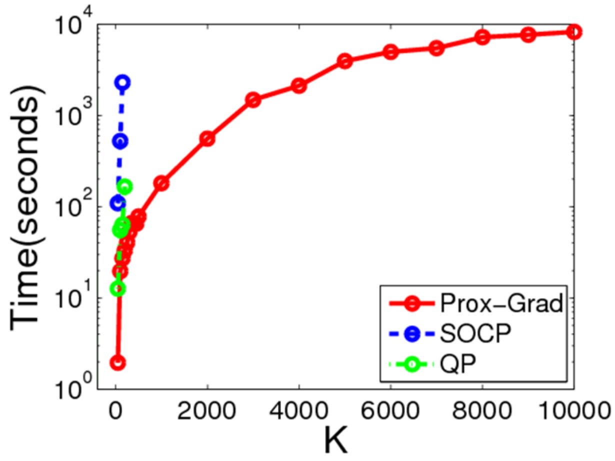
Graph:

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

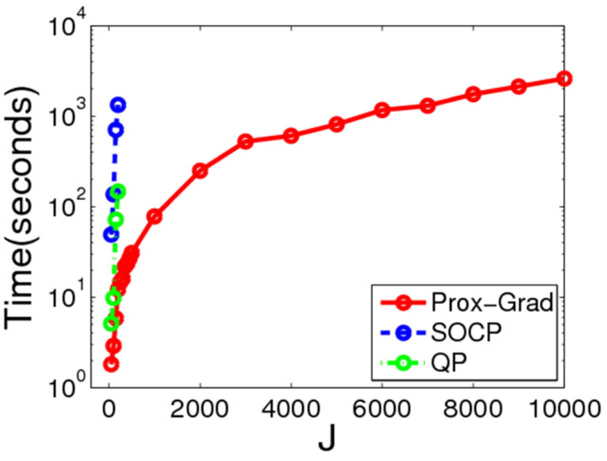
Proximal-Gradient: **Independent of Sample Size**
Linear in #.of Tasks

Experiments

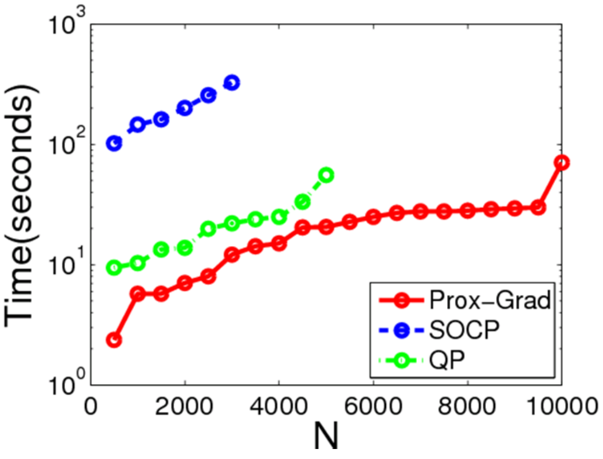
- Multi-task Graph Structured Sparse Learning (GFlasso)



$N = 500, J = 100$



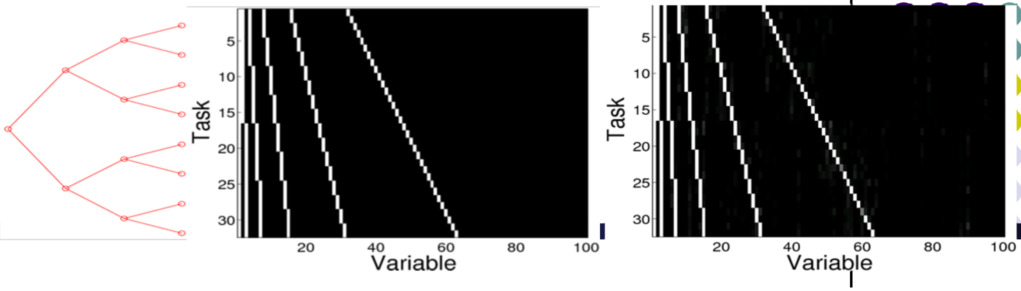
$N = 1000, K = 50$



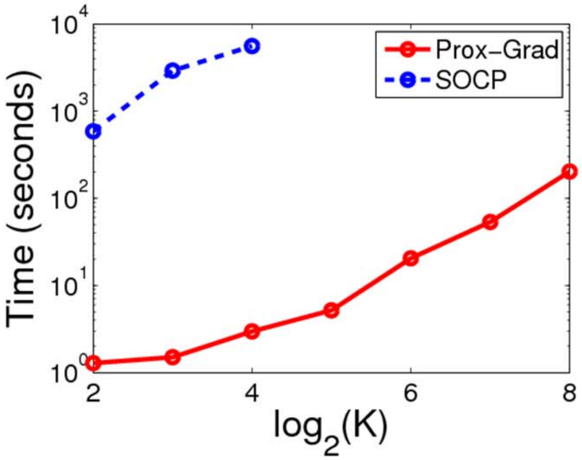
$J = 100, K = 50$

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

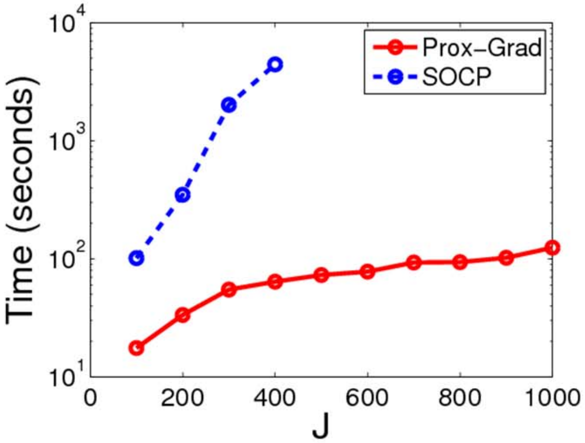
Experiments



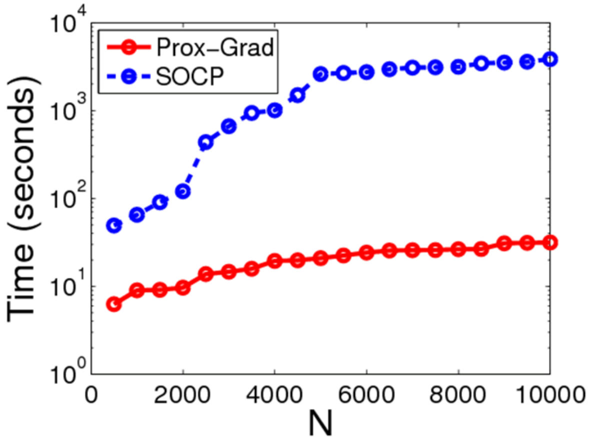
- Multi-task Tree-Structured Sparse Learning (TreeLasso)



$$N = 1000, J = 600$$



$$N = 1000, K = 32$$



$$J = 100, K = 32$$

$$\epsilon = 0.1$$

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.