Multivariate Approaches: Joint Modeling of Imaging & Genetic Data

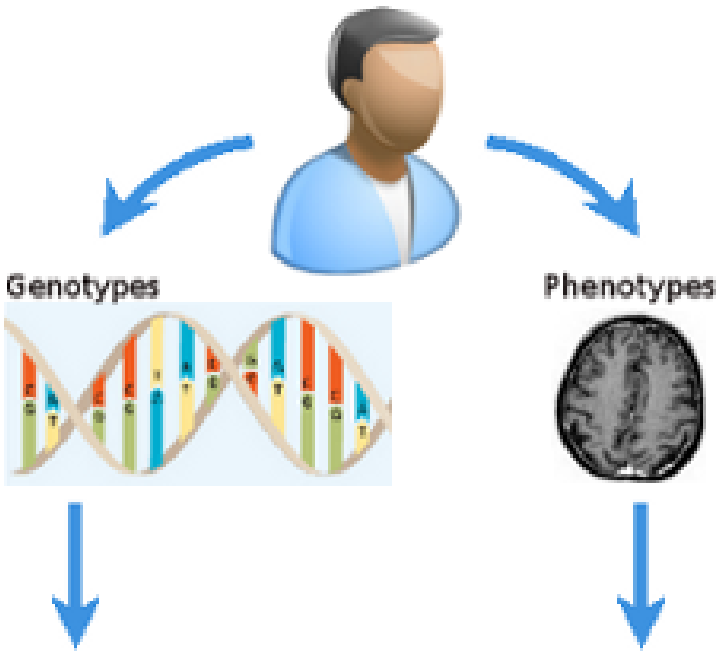
Giovanni Montana

Statistics Section
Department of Mathematics
Imperial College
London, UK

10 June, 2012

Association Mapping with Unrelated Individuals

Subject i $(i = 1, \ldots, n)$



$(x_{i1}, x_{i2}, \ldots, x_{ip})$ $(y_{i1}, y_{i2}, \ldots, y_{iq})$

- For instance, $p = 600,000$ SNPs, and $q = 2,000,000$ vocels.
- The goal is to identify markers highly predictive of *all* or *phenotypes*

Mass Univariate Linear Modelling (MULM)

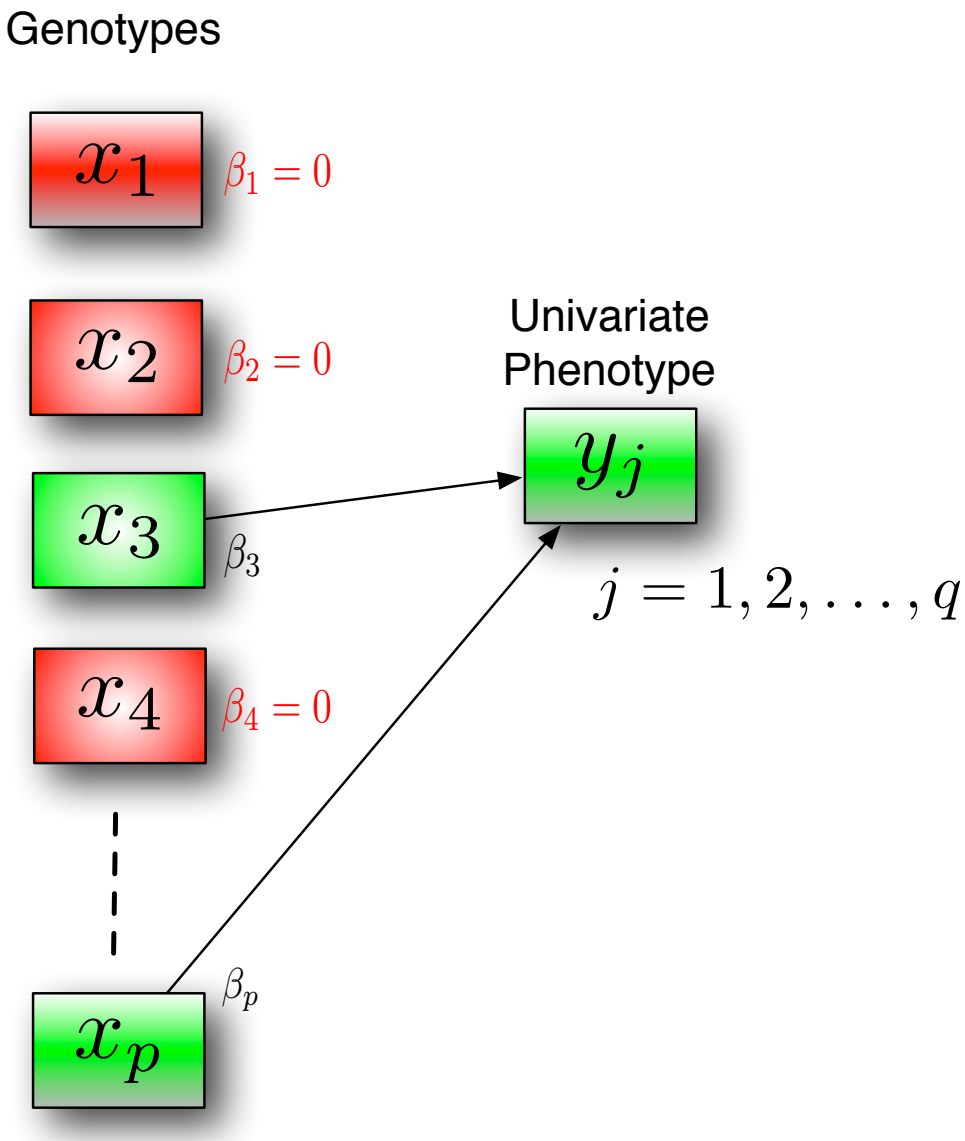
1. Fit all $(p \times q)$ linear regression models

$$y_j = \beta_{jk} x_k + \epsilon$$

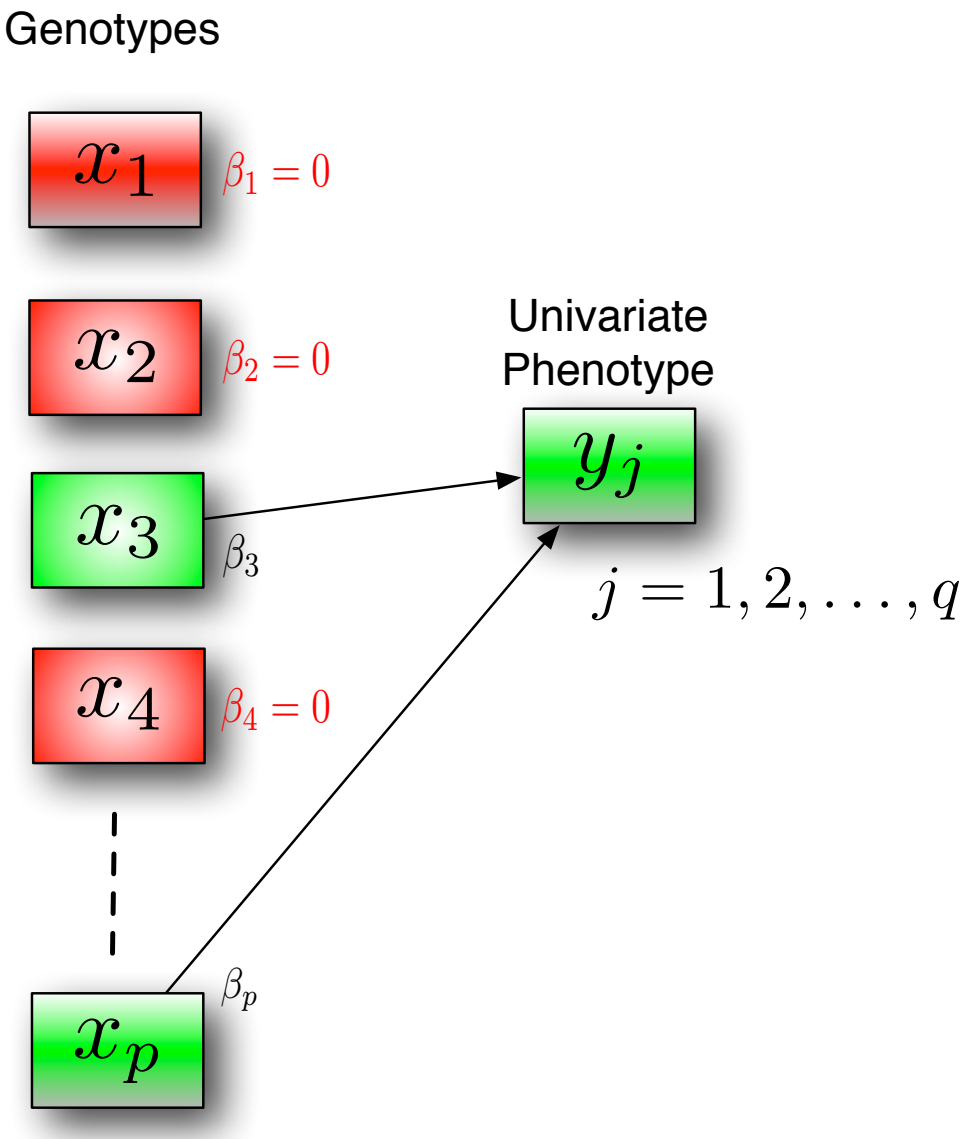
2. Test all $(p \times q)$ null hypotheses

$$H_0 : \beta_{jk} = 0$$

3. Correct for multiple testing, e.g. control FWER or FDR
4. Rank by p-values



MULM: Properties

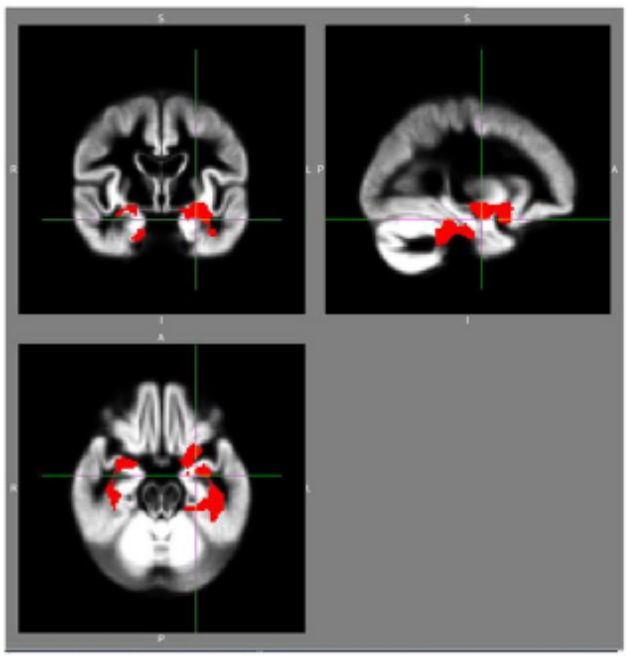


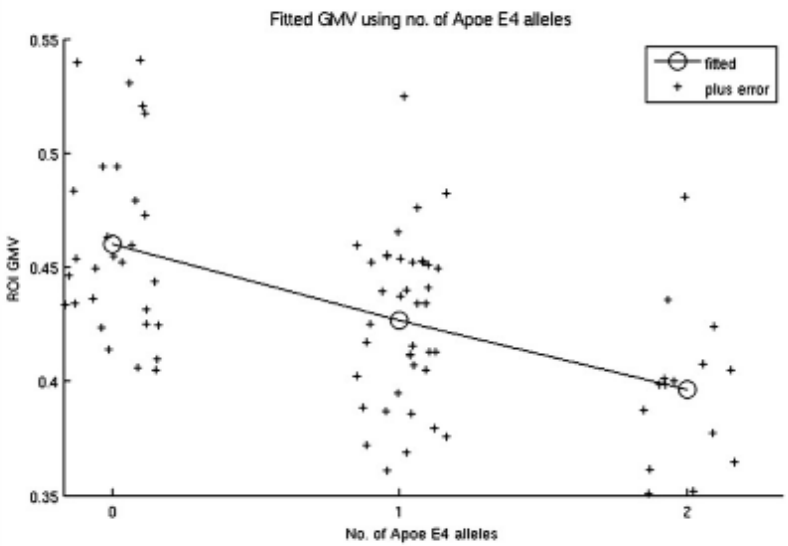
1. Linear genotype-phenotype relationship
2. Ranks all possible genotype-phenotype pairs
3. Ignores dependences among *genotypes*
4. Ignores dependences among *phenotypes*
5. Massive *multiple testing* problem which should account for dependence patterns

Example: APOE4 in Alzheimer's Disease

Filippini *et al* 2009

- APOE4 SNP: Homozygote of minor allele is 2, heterozygote is 1 and homozygote of major allele is 0
- $n = 83$ and $q = 30$k voxels (Gray Matter Volume in 15 ROIs)
- Mean GMV reduced by 14% in the red area for homozygotes compared to non-carriers





A Multivariate Regression Approach

1. For each univariate phenotype y_j, fit the multiple linear regression model

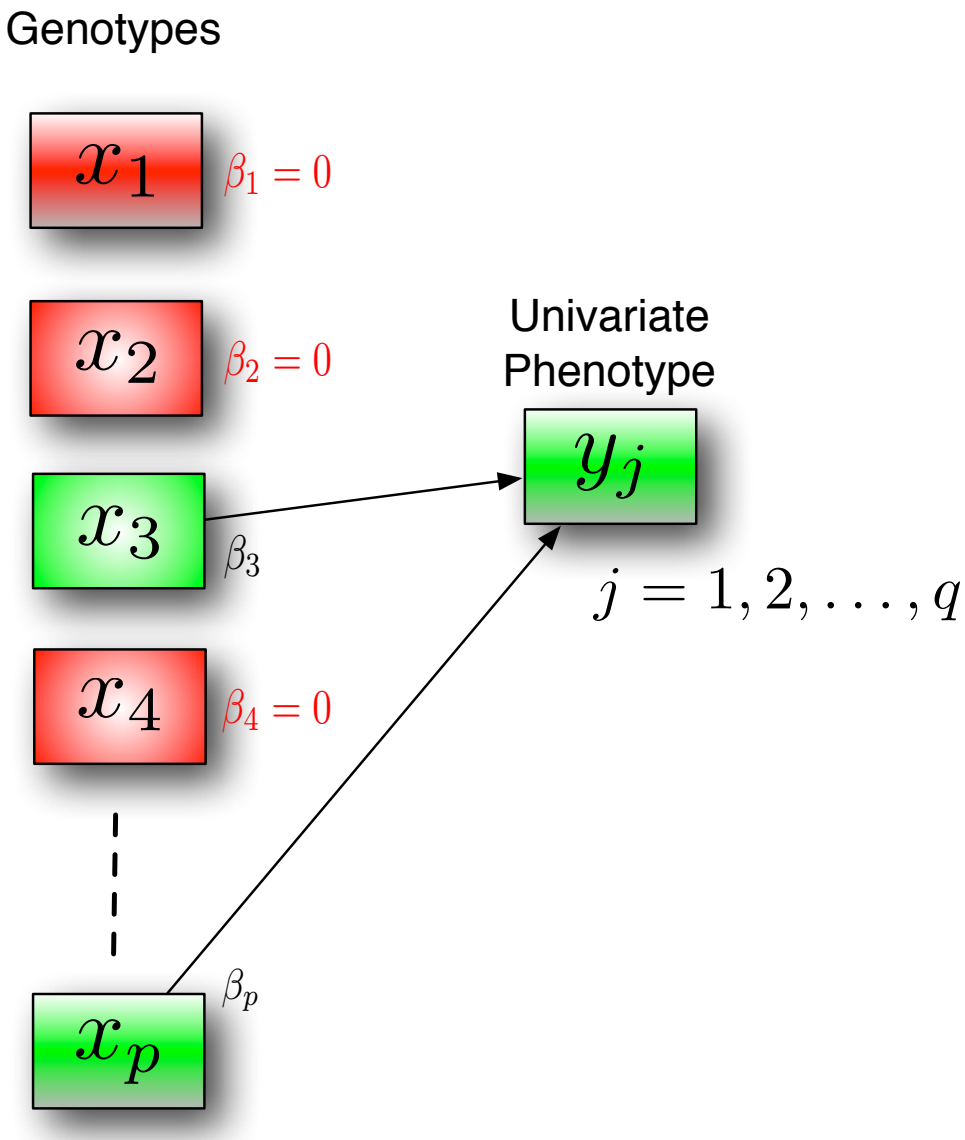
$$y_j = \sum_{k=1}^{p} \beta_k x_k + \epsilon$$

2. Solve the OLS problem after imposing a *penalty*

$$P(\boldsymbol{\beta}) < c$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$

3. Detect genetic factors influencing each one of the q univariate phenotypes



Selected Penalties

- Lasso penalty:

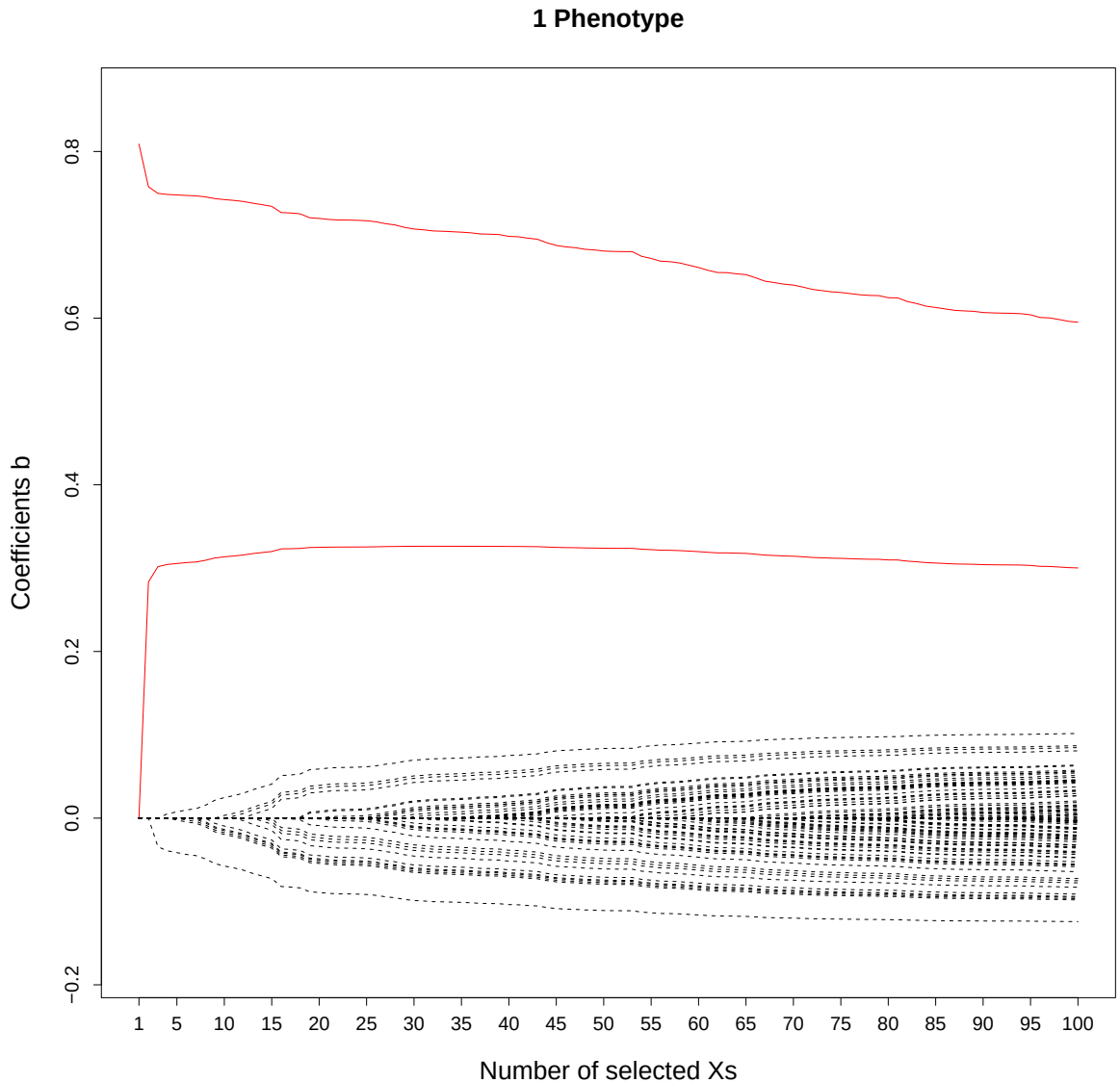
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} x_{ik}\beta_k)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right\}$$

  - The l_1 penalty ensures that some coefficients will be exactly zero
  - The λ parameter controls the number of non-zero coefficients

- Other penalties have been proposed to impose some structure while performing variable selection, e.g.
  - Elastic net (l_1 and l_2 penalties)
  - Group lasso

Lasso Regression: an Illustration

$p = 100$ genotypes, y linearly depends only on (x_1, x_2)



Fast Parameter Estimation

- Often no closed-form estimators are available
- When predictors are uncorrelated, use *soft thresholding*
  1. Find the OLS estimates $\hat{\boldsymbol{\beta}}^{\text{ols}} = \left\{\beta_1^{\text{ols}}, \beta_2^{\text{ols}}, \ldots, \beta_p^{\text{ols}}\right\}$
  2. Cycle over the coefficients and apply a thresholding update:

$$\hat{\beta}_k^{\text{lasso}} = \text{sign}(\hat{\beta}_k^{\text{ols}}) \left(|\hat{\beta}_k^{\text{ols}}| - \frac{\lambda}{2} \right)_+$$

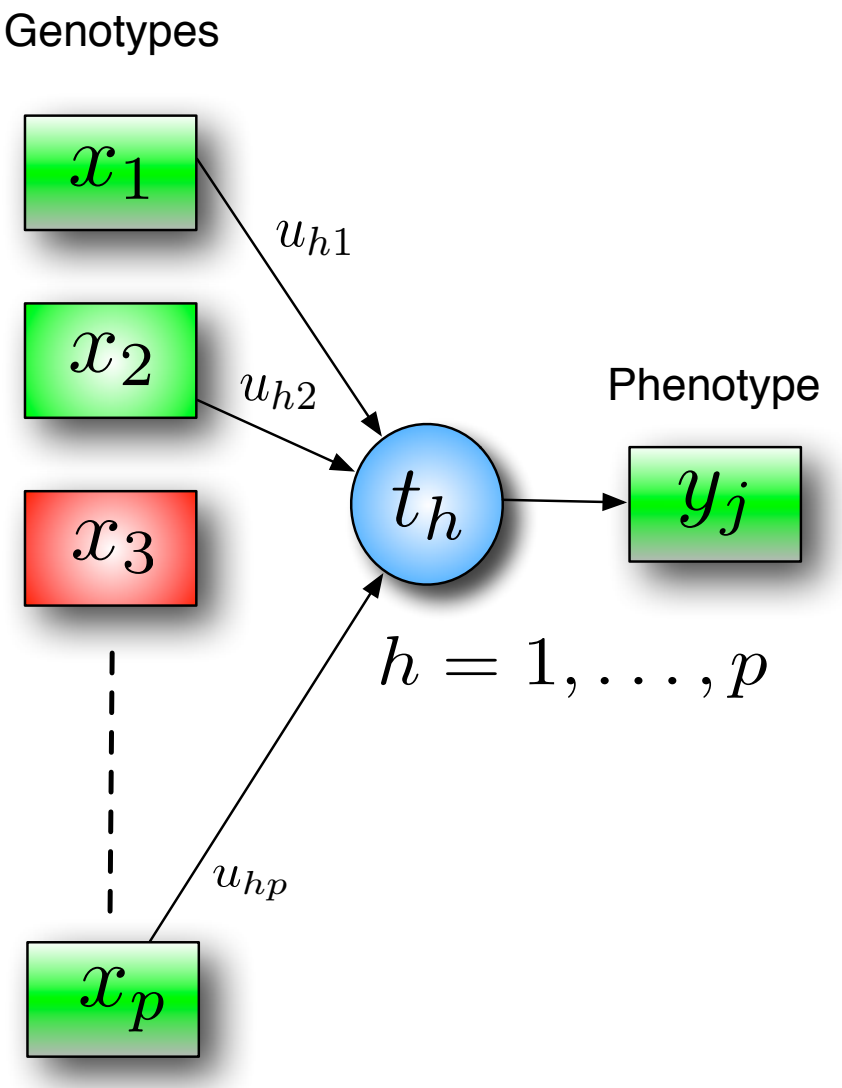
where

$$(a)_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

- With correlated predictors, use *coordinate descent*
  - Fast iterative algorithm also based on soft-thresholding
  - Can be derived for a large class of penalties

Penalised Regression on Latent Factors

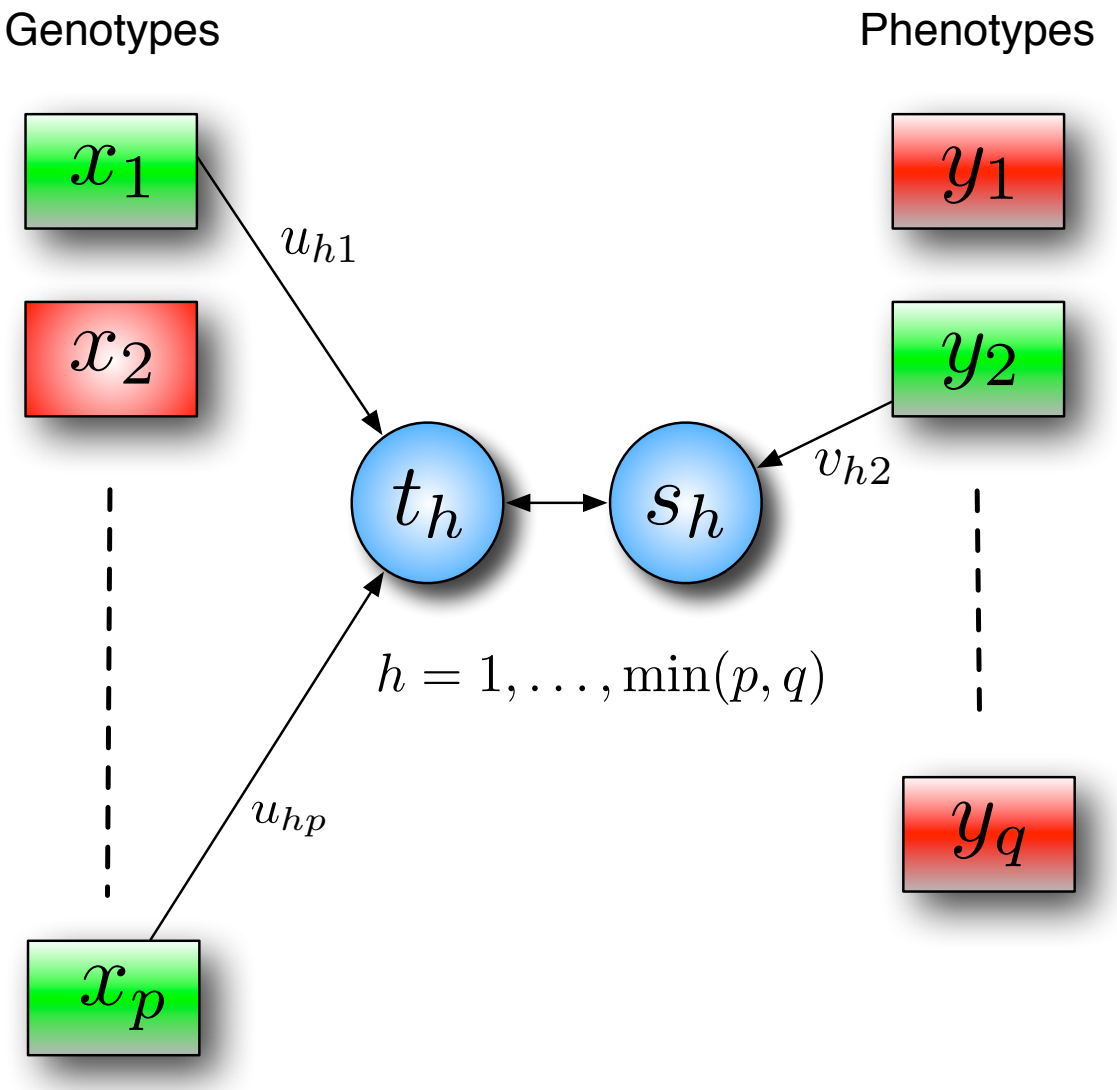
For univariate phenotypes



1. High number of correlated predictive markers
2. Assume the existence of *latent factors*, i.e. hidden factors that have high predictive power
3. A latent factor is a linear combination of SNPs
4. Estimate the contribution of each SNPs on each factor while enforcing that only a subset of SNPs have non-zero weights

Penalised Regression on Latent Factors

For multivariate phenotypes

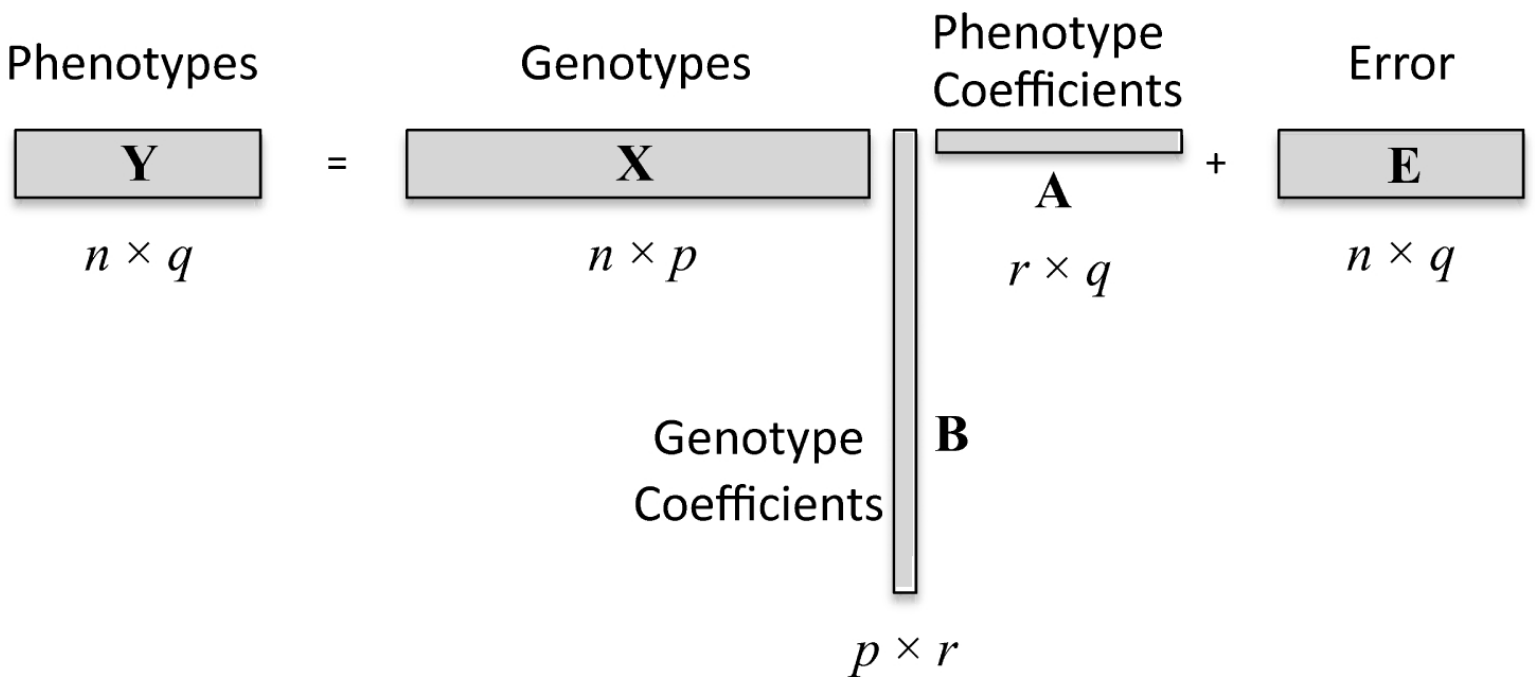


Reduced-Rank Regression

- If $\mathbf{C}$ is the $(p \times q)$ matrix of regression coefficients, then $\mathbf{Y} = \mathbf{X}\,\mathbf{C} + \mathbf{E}$
- When $\mathbf{C}$ has rank $r < \min(p, q)$, the model can be written as

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B}\mathbf{A} + \mathbf{E}$$

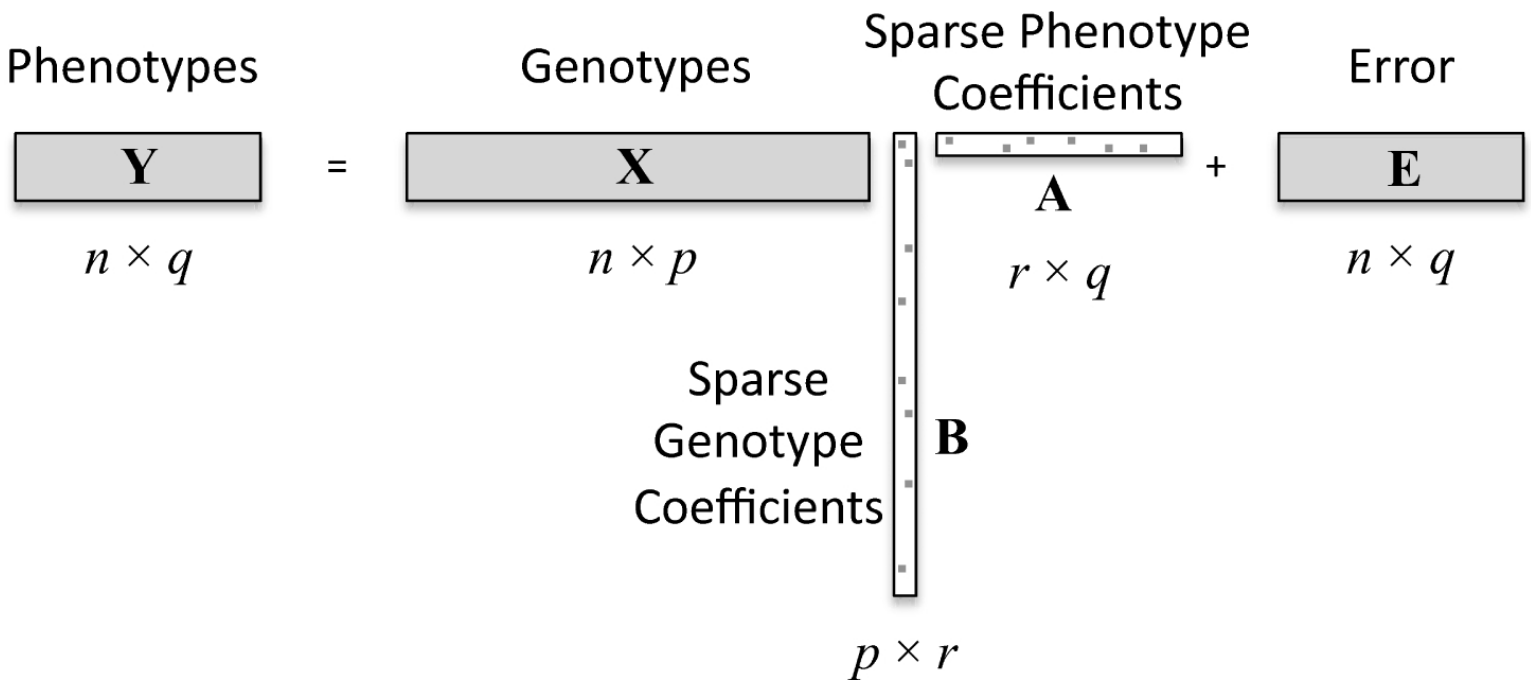
- Each one of the r ranks captures a different causal effect



Sparse Reduced-Rank Regression (sRRR)

Vounou *et al.* (2010, 2012)

- Simultaneous genotype and phenotype selection is achieved by imposing penalties on **A** and **B**



Rank-1 sRRR

- The rank-1 model is:

$$\underset{(n\times q)}{\mathbf{Y}} = \underset{(n\times p)}{\mathbf{X}}\ \underset{(p\times 1)}{\mathbf{b}}\ \underset{(1\times q)}{\mathbf{a}} + \underset{(n\times q)}{\mathbf{E}}$$

- The sparse coefficients **b** and **a** are found by solving

$$\hat{\mathbf{b}}, \hat{\mathbf{a}} = \underset{\mathbf{b},\mathbf{a}}{\operatorname{argmin}} \left\{ \operatorname{Tr}\left[(\mathbf{Y} - \mathbf{Xba})\, \mathbf{\Gamma}\, (\mathbf{Y} - \mathbf{Xba})' \right] + \lambda_a P_a(\mathbf{a}) + \lambda_b P_a(\mathbf{b}) \right\}$$

  where $\mathbf{\Gamma}$ is a given $q \times q$ positive definite matrix, e.g. $\mathbf{1}_q$

- Lasso penalties can be used for both genotypes and phenotypes

$$P_a(\mathbf{a}) = \sum_{j=1}^{q} |a_j| \qquad P_b(\mathbf{b}) = \sum_{j=1}^{p} |b_j|$$

  but many other penalties are also possible to impose more structure

Coordinate Descent Algorithm for Rank-1 sRRR

1. Initialise
   - $\mathbf{a}^0$ such that $\mathbf{a}^0\mathbf{a}'^0 = 1$
   - $\mathbf{b}^0$ such that $\mathbf{b}^{0\prime}\mathbf{b}^0 = 1$
2. Iterate until convergence
   - Update and normalise

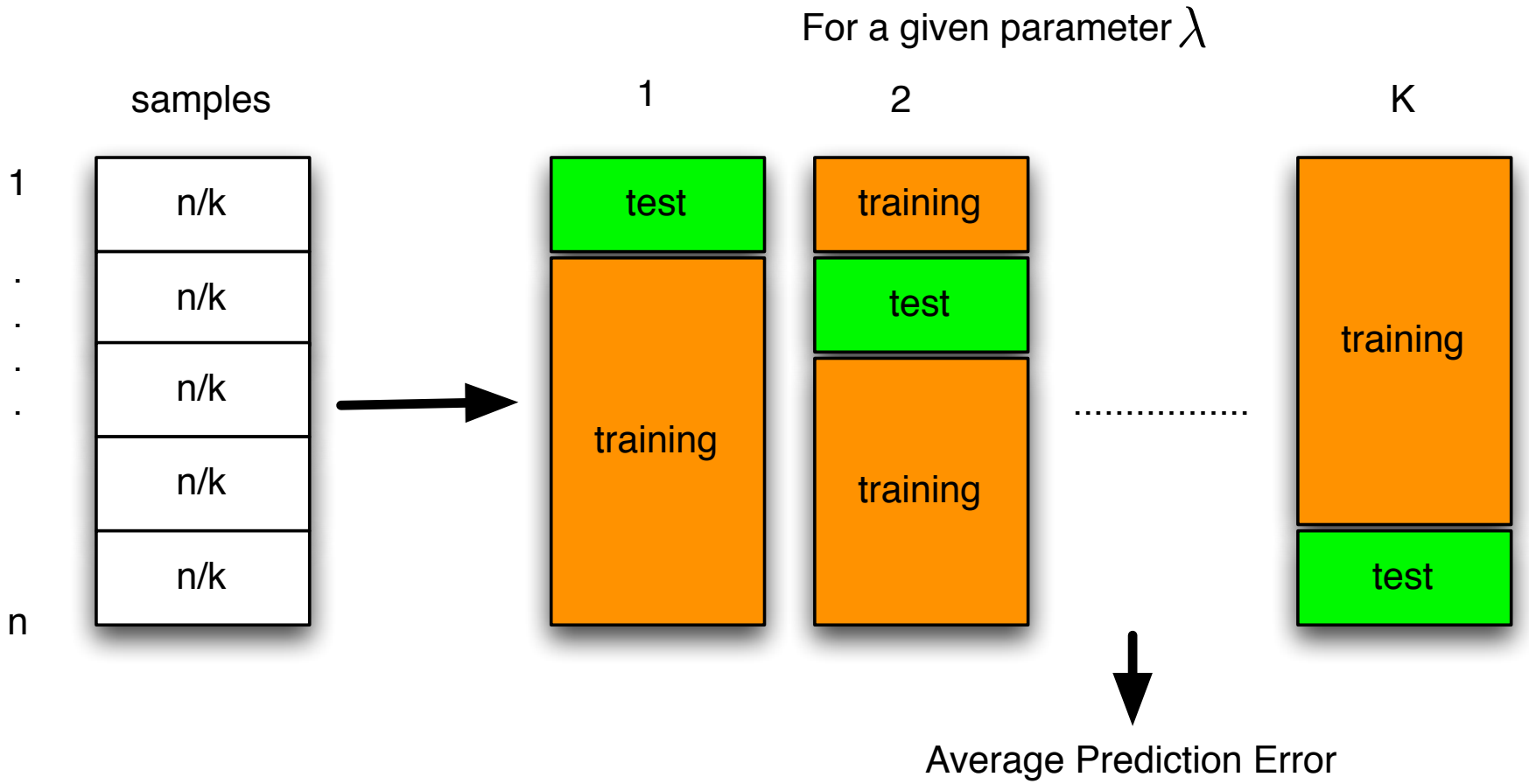
$$\hat{\mathbf{b}} = \text{sign}(\mathbf{X}'\mathbf{Y}\mathbf{a}^{0\prime}) \left(|\mathbf{X}'\mathbf{Y}\mathbf{a}^{0\prime}| - \frac{\lambda_b}{2} \right)_+$$

   - Update and normalise

$$\hat{\mathbf{a}} = \text{sign}(\mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}) \left(|\mathbf{b}^{0\prime}\mathbf{X}'\mathbf{Y}| - \frac{\lambda_a}{2} \right)_+$$

   - Set $\mathbf{a}^0 = \hat{\mathbf{a}}$ and $\mathbf{b}^0 = \hat{\mathbf{b}}$

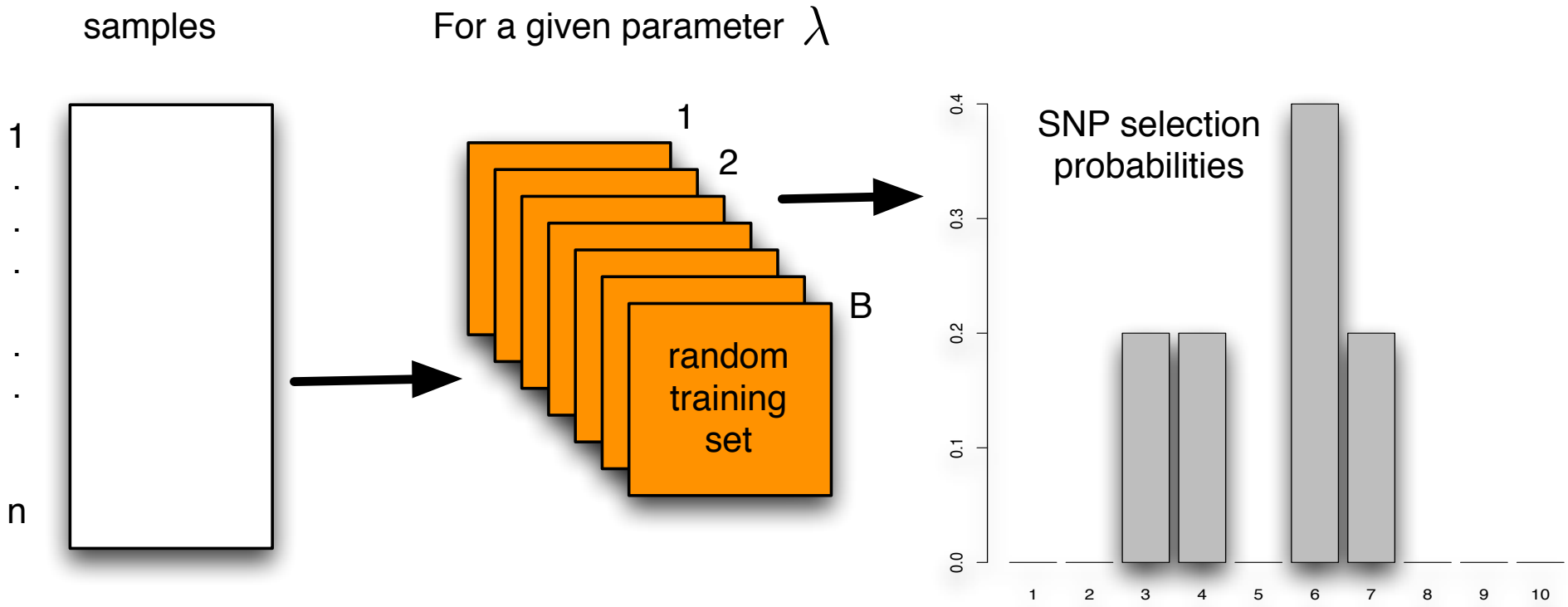
SNP Ranking using k-fold Cross Validation



1. Repeat for all λ in $[\lambda_{\text{min}}, \lambda_{\text{max}}]$
2. Select λ^* corresponding the the smaller average prediction error

SNP Ranking using Data Resampling

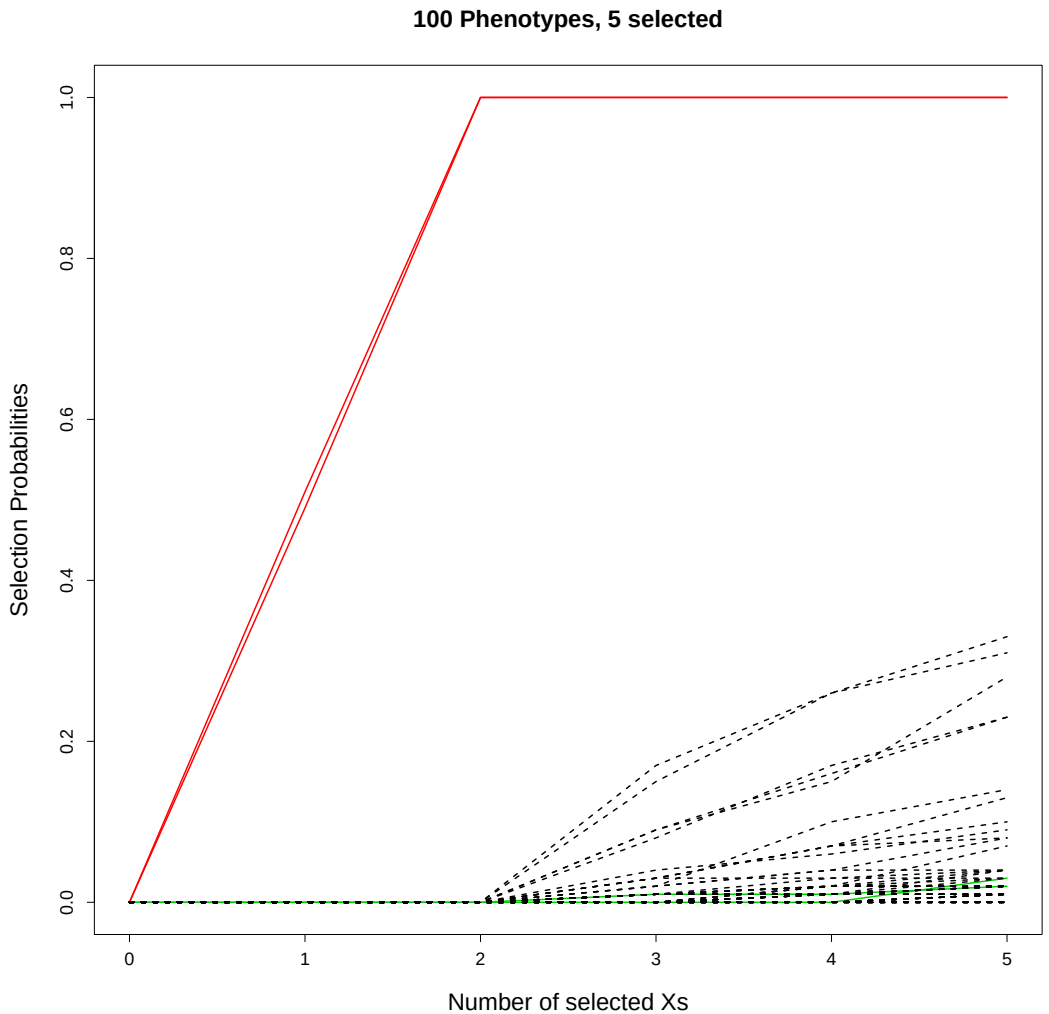
Resampling Without Replacement



1. Repeat for all λ in $[\lambda_{\min}, \lambda_{\max}]$
2. Select the SNPs having selection probability greater than threshold

Illustration: sRRR with Data Resampling

Selected SNPs from Rank 1



- $p = 100$ genotypes
- $q = 100$ phenotypes
- First signal: (y_1, y_2) on (x_1, x_2)
- Second signal: (y_3, y_4) on (x_3, x_4)

Illustration: sRRR with Data Resampling

Selected SNPs from Rank 2

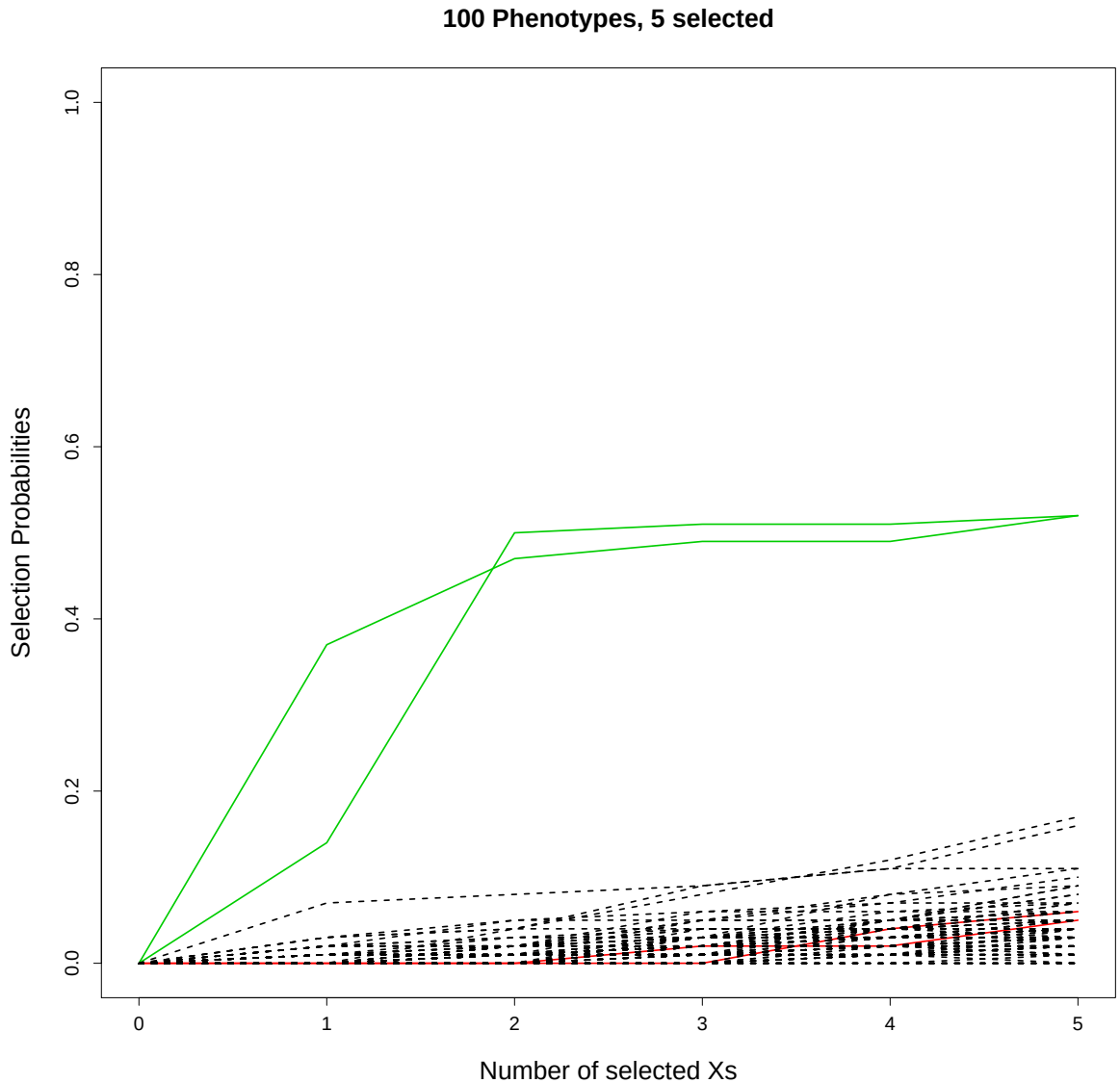
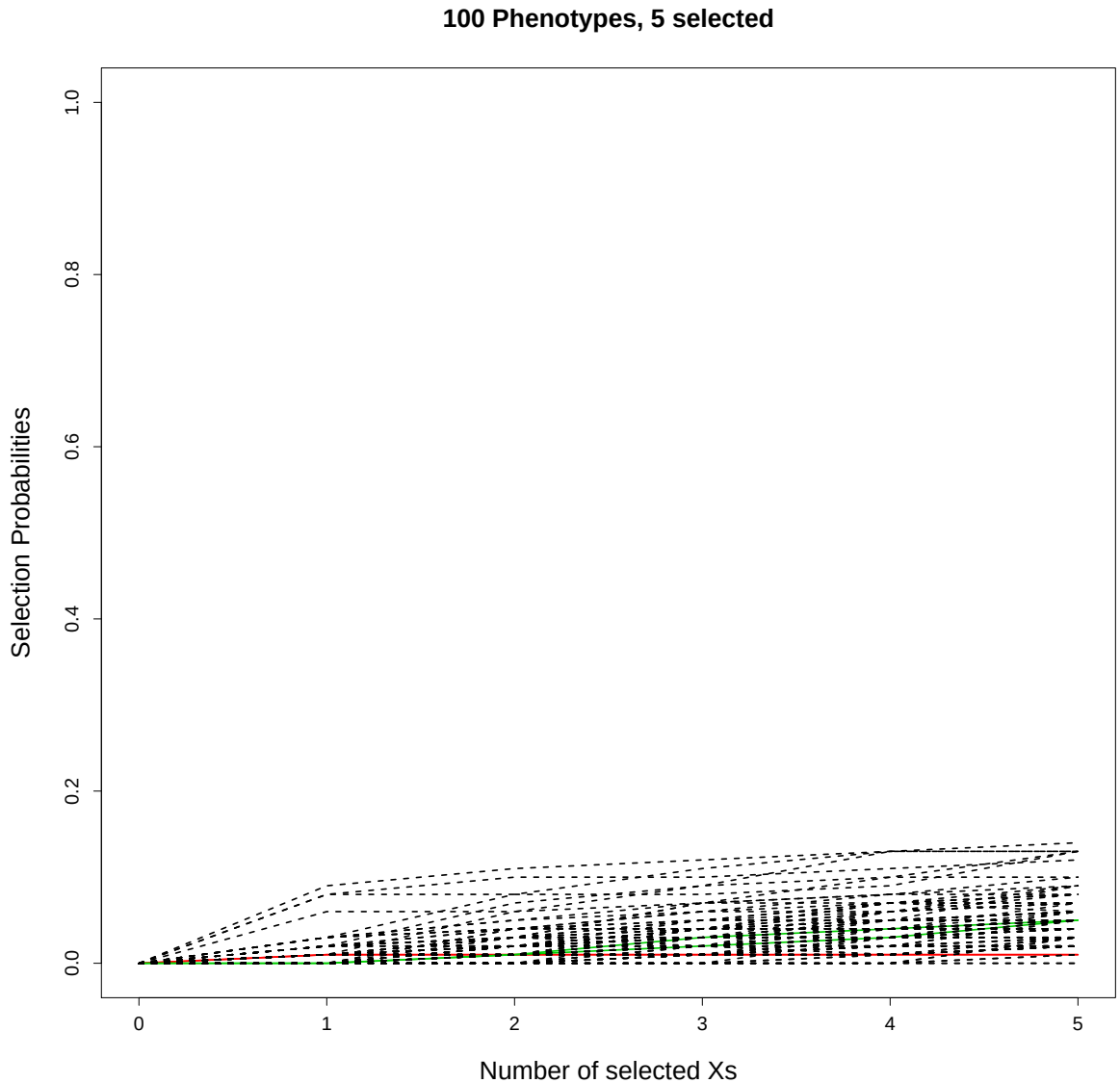


Illustration: sRRR with Data Resampling

Selected SNPs from Rank 3



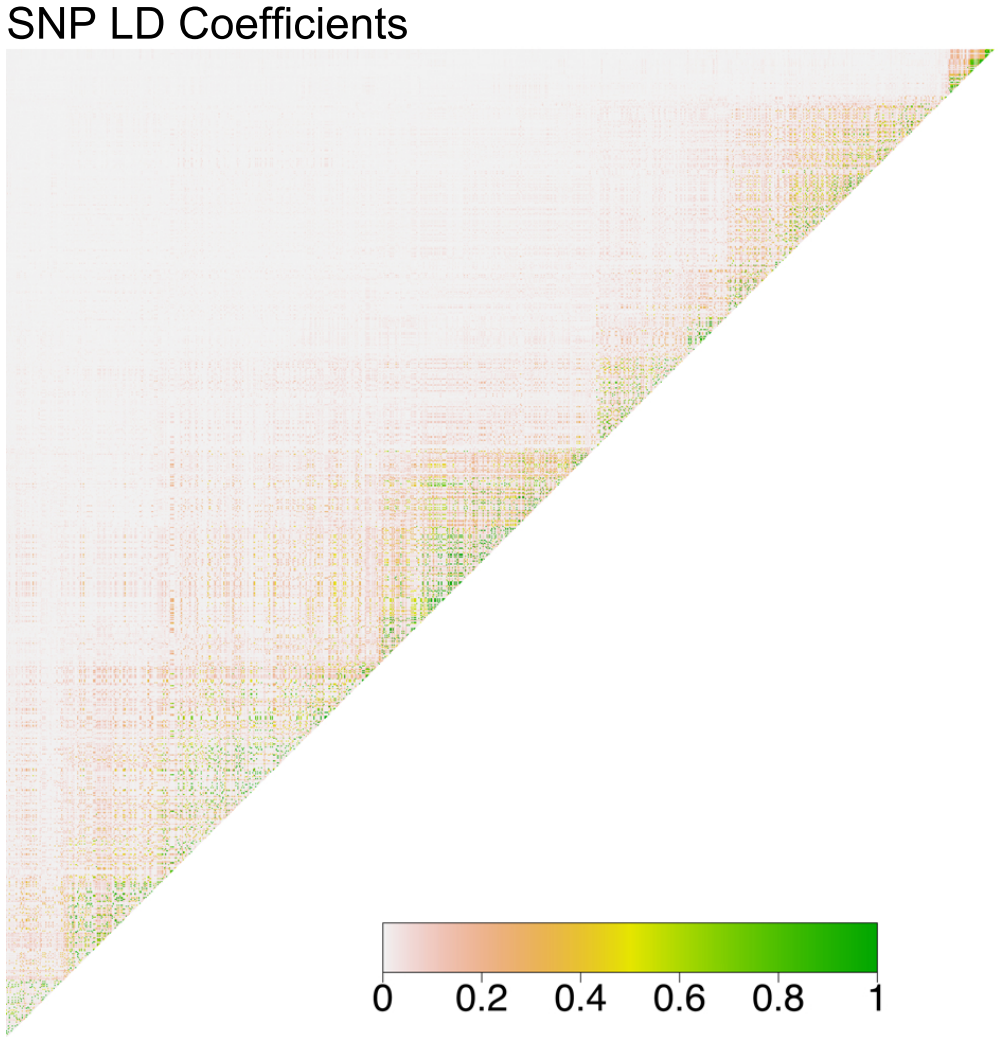
Power Studies

1. Simulation of genotypes in a human population
   - *Forwards-in-time simulation* with FREGENE (Hoggart *et al*, 2007)
   - Data is simulated realistically - evolutionary parameters are controlled
   - $N = 10$k individuals
   - $P = 40$k SNPs

2. Repeated sampling from the population
   - *Control study design*
     - ★ Sample size n
     - ★ Total number of design SNPs p
   - *Simulate phenotypes*
     - ★ $q = 111$ ROIs (GSK Brain Atlas)
     - ★ From each ROI, simulated a mean modulated GM value
     - ★ Simulations calibrated on ADNI data
   - *Induce genetic effects*
     - ★ Randomly select 10 causative SNPs
     - ★ Induce a reduction in mean GM in 6 ROIs
     - ★ Genetic effects explain 5% of phenotypic variance in the 6 ROIs

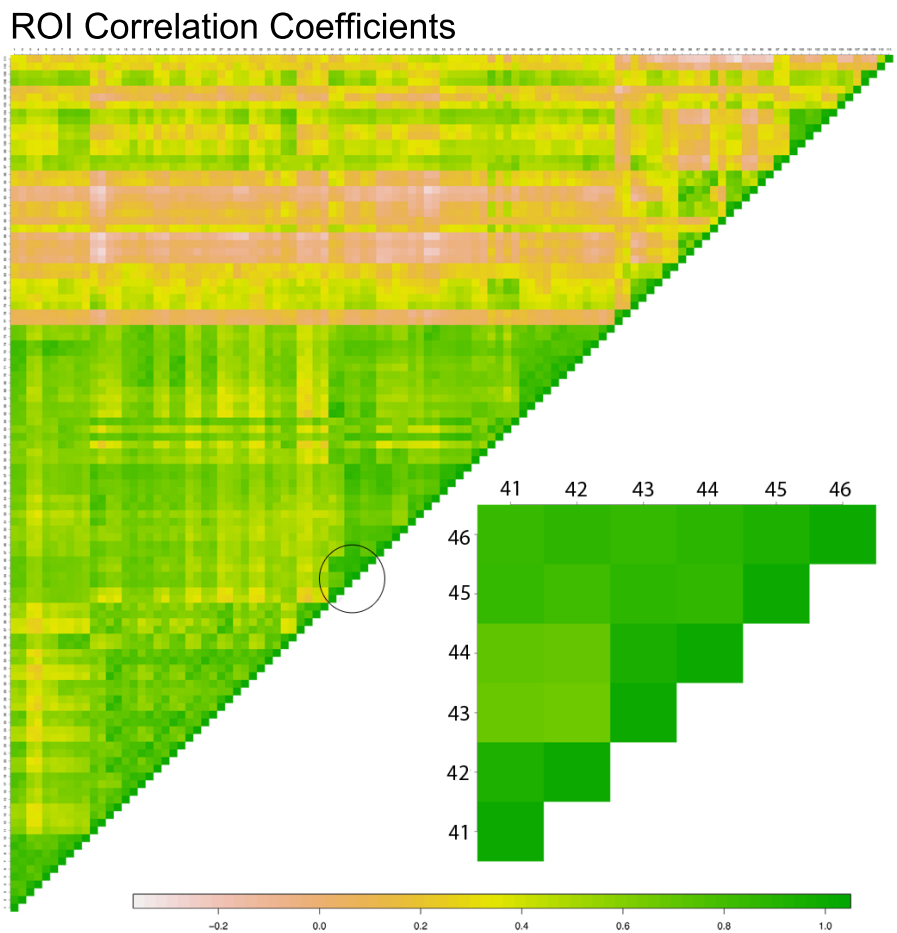
Simulated Genotypes

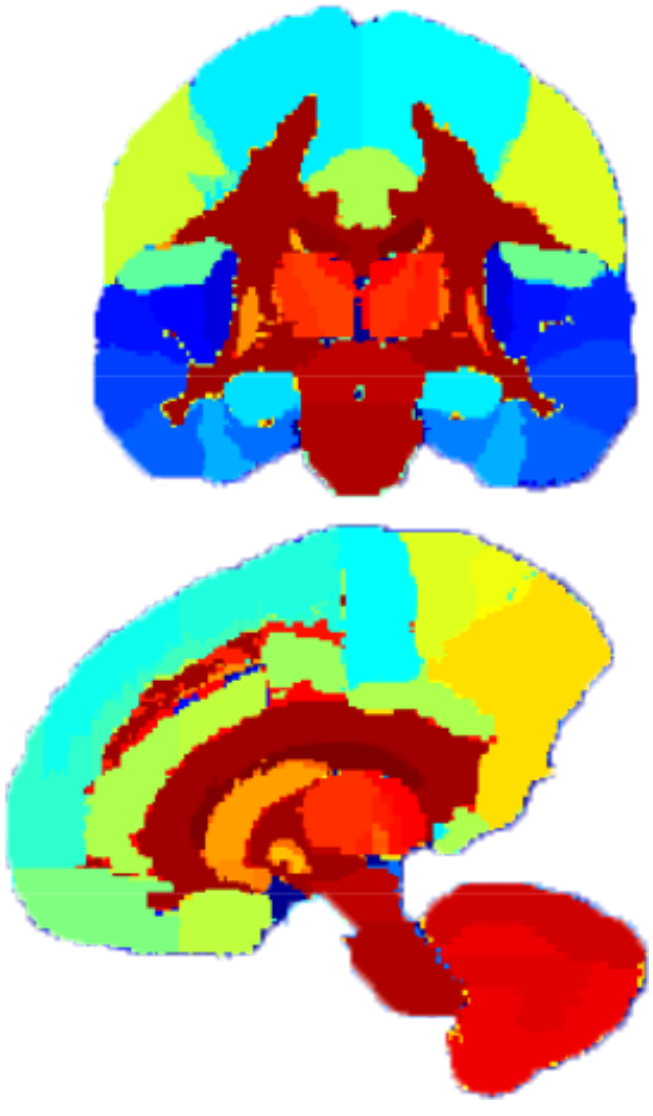
Linkage disequilibrium patterns



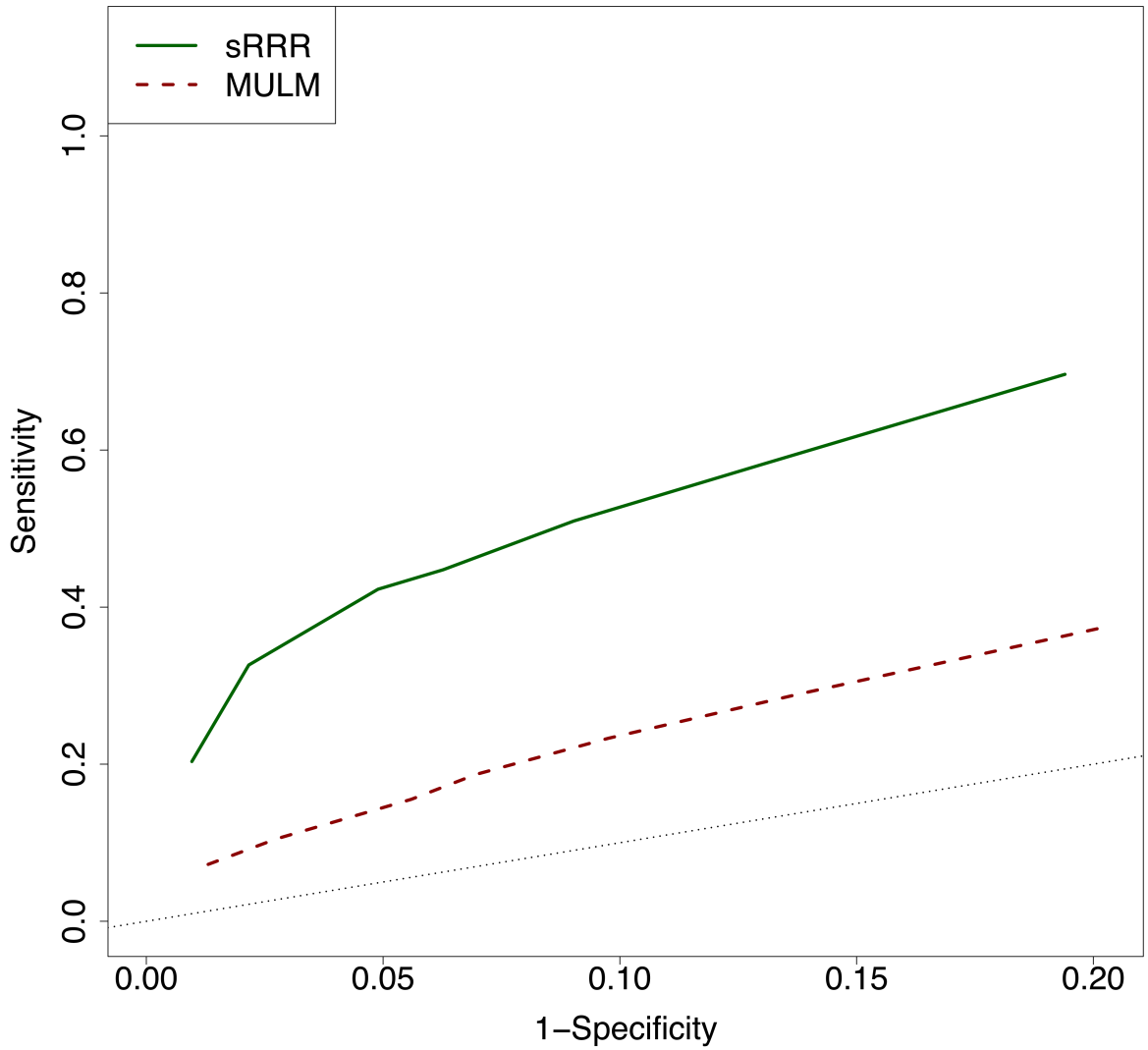
Simulated Phenotypes

ROI correlation matrix

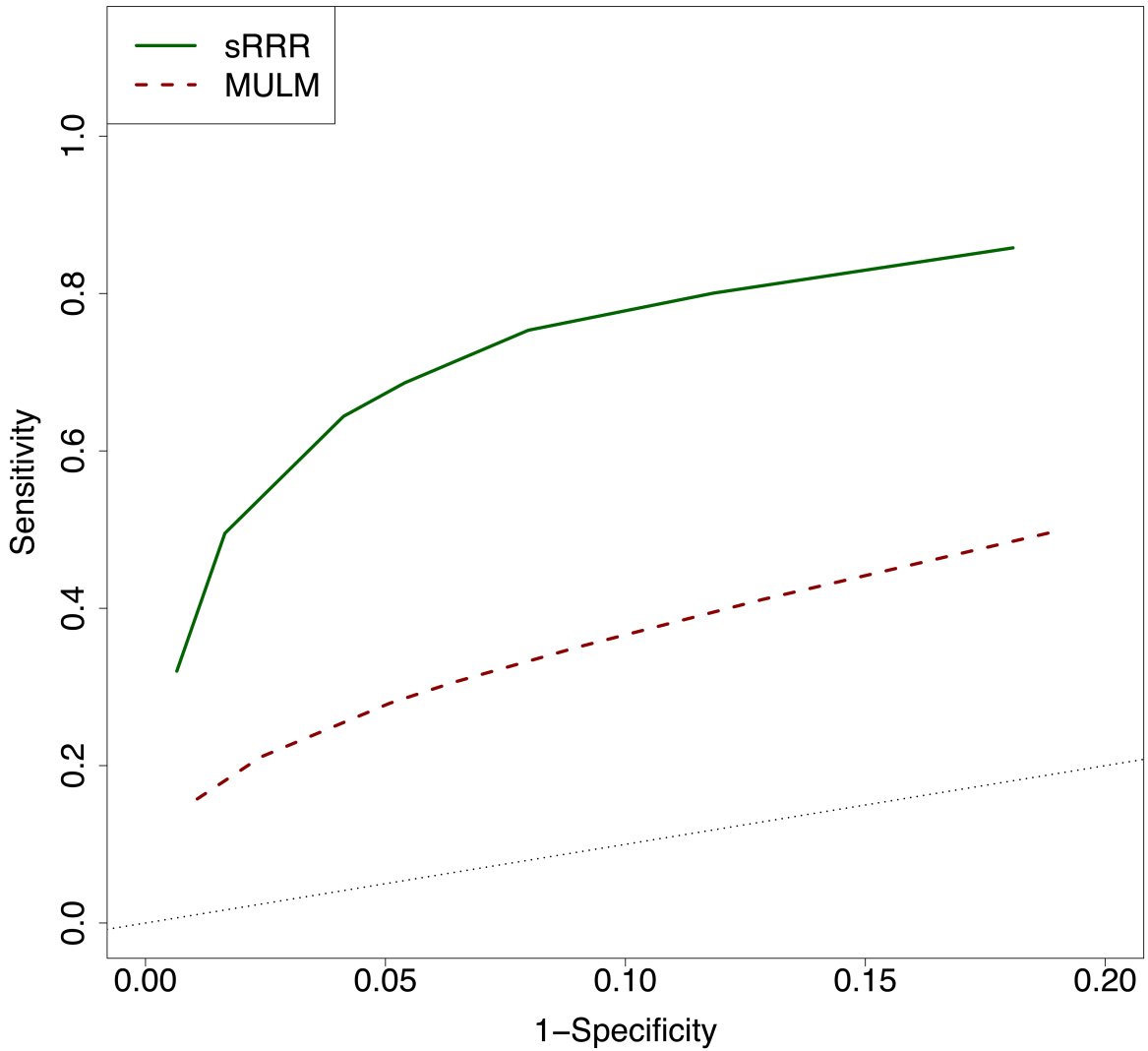




Power to Detect Causative SNPs ($n = 500$)

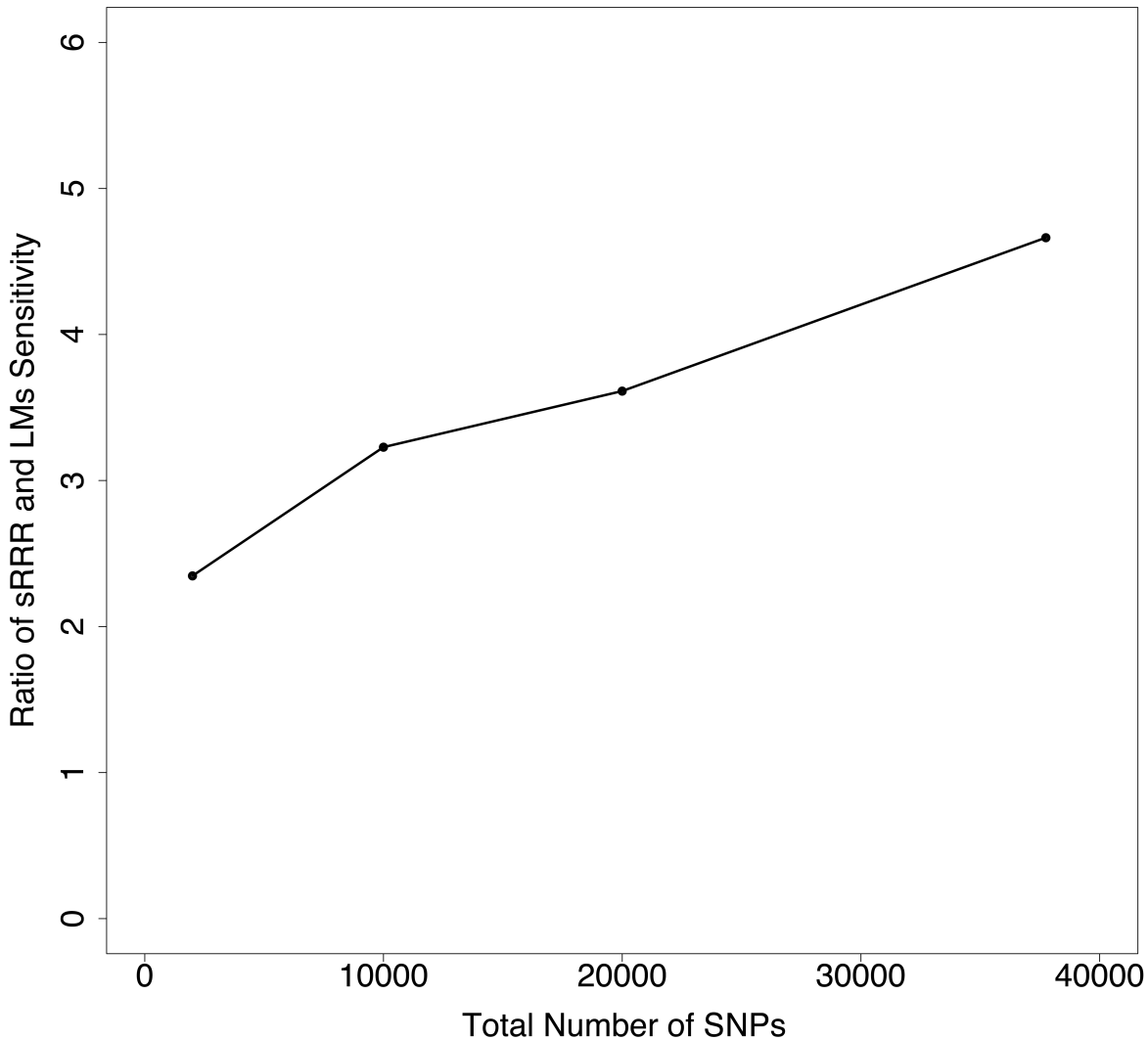


Power to Detect Causative SNPs ($n = 1000$)



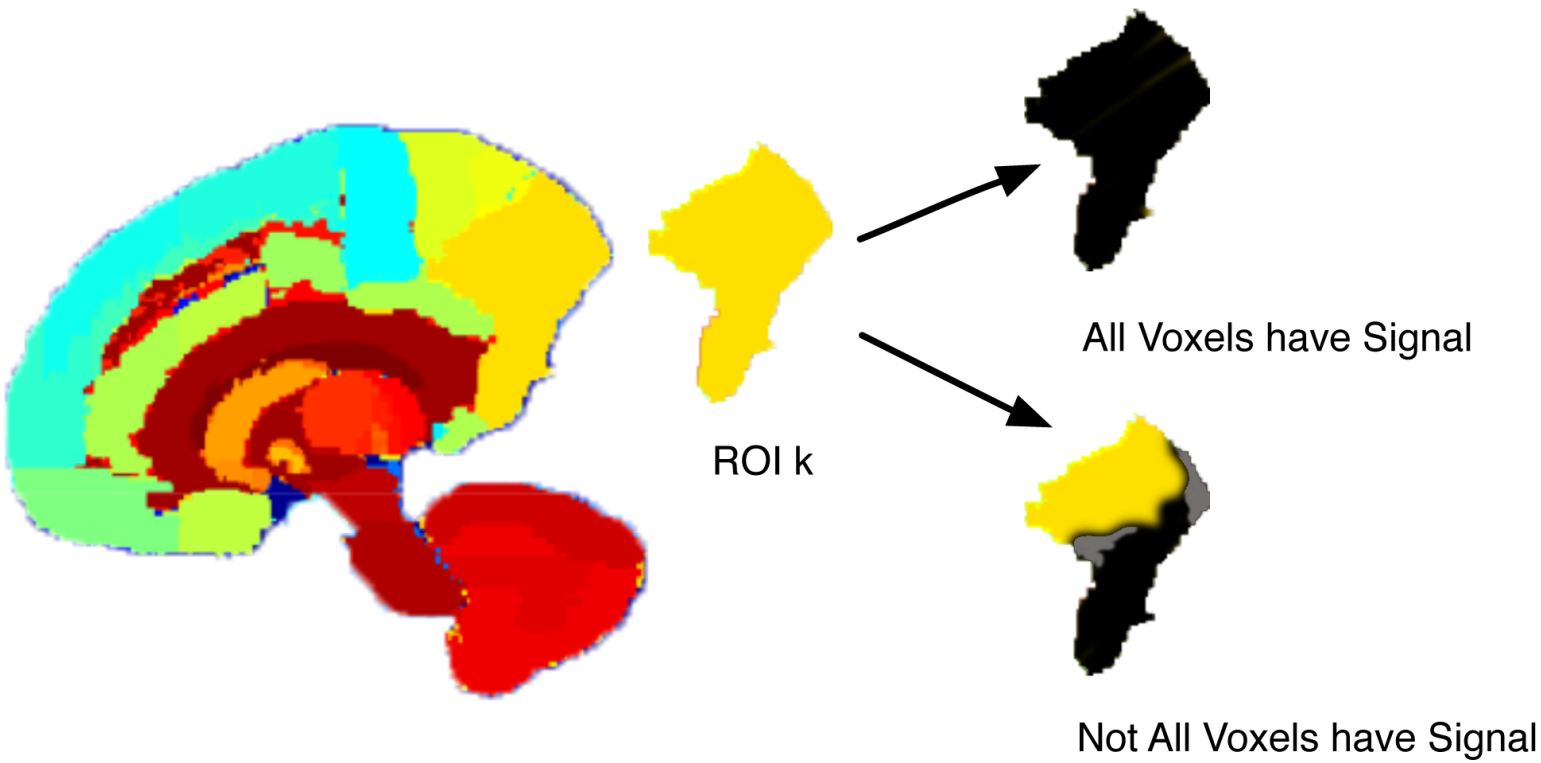
Relative Power in Large-scale GWA Studies

Ratio of SNP sensitivities (sRRR/MULM) as a function of the total number of SNPs



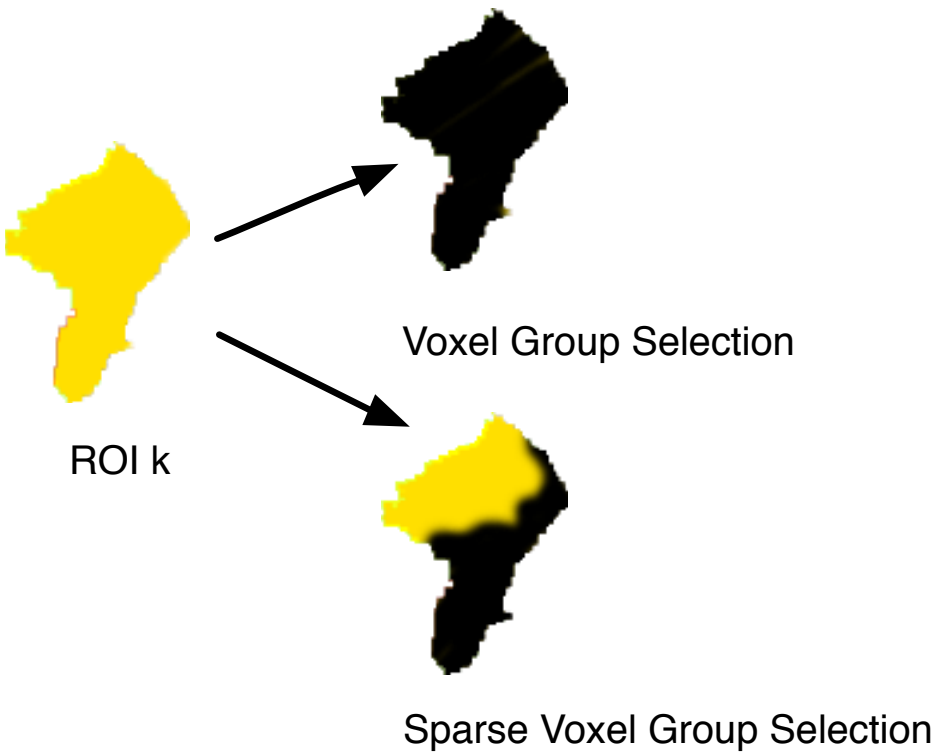
Atlas-Guided sRRR

- K non-overlapping voxel groups (ROIs) $g_1, \ldots, g_K$
- Group k contain q_k voxels
- It is common to take ROI averages $z_1, \ldots, z_K$ as phenotypes
- However *averaging* may reduce the power to detect genetic effects



Two Atlas-guided sRRR Models

Group and sparse group selection penalties



1. A model to select *all voxels* within one or more ROIs (*Group Selection*)
2. A model to select only *subset of voxels* within one or more ROI (*Sparse Group Selection*)

Penalties

$$\hat{\mathbf{b}}, \hat{\mathbf{a}} = \underset{\mathbf{b},\mathbf{a}}{\operatorname{argmin}} \left\{ \operatorname{Tr}\left[(\mathbf{Y} - \mathbf{Xba})\, \mathbf{\Gamma}\, (\mathbf{Y} - \mathbf{Xba})' \right] + \lambda_a P_a(\mathbf{a}) + \lambda_b P_b(\mathbf{b}) \right\}$$

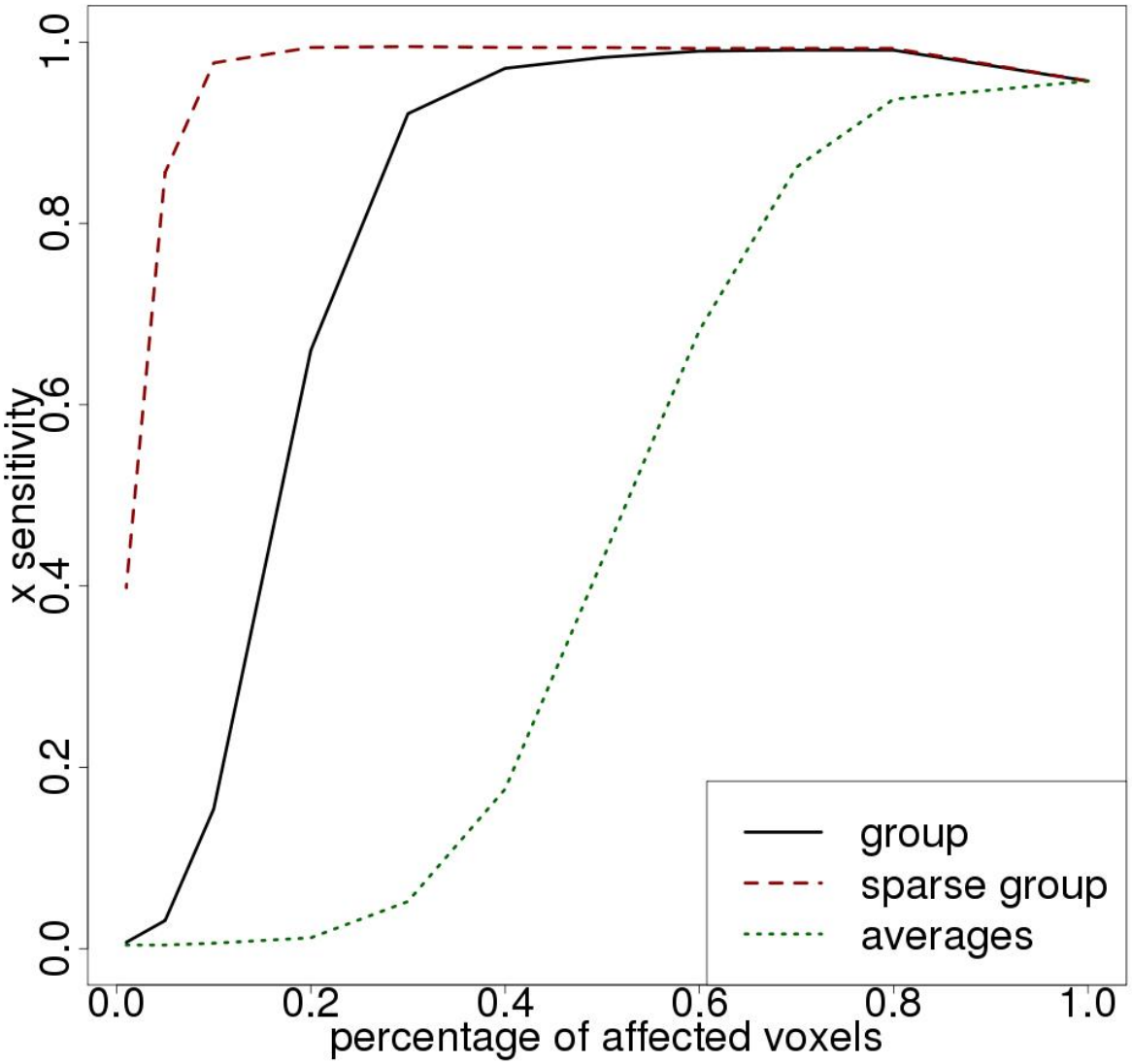
1. **Group lasso**: select *all* voxels in a ROI

$$P_a(\mathbf{a}) = \sum_{k=1}^{K} \|\mathbf{a}_{g_k}\|_2$$

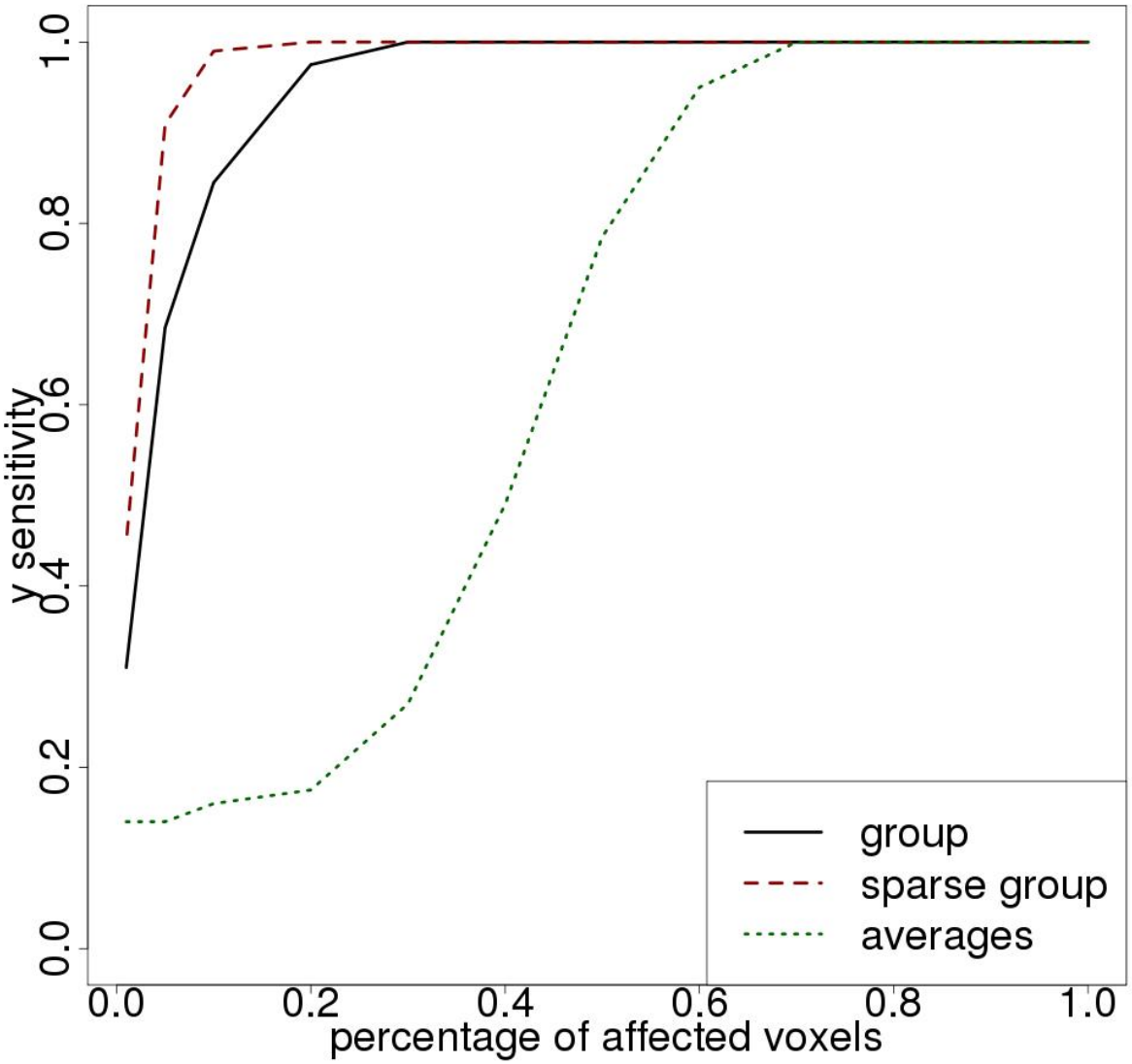
2. **Sparse group lasso**: select a *subset* of voxels within a ROI

$$P_a(\mathbf{a}) = m \sum_{j=1}^{q} |a_j| + (1-m) \sum_{k=1}^{K} \|\mathbf{a}_{g_k}\|_2$$

Power to Detect Causative SNPs



Power to Detect Signal Voxels



ADNI: Genetic Data

- Samples available for AD's and MCI (Mild Cognitive Impairment) patients and healthy controls (CN)
  - Progressive MCI (P-MCI): those who converted to AD
  - Stable MCI (S-MCI): those who did not convert to AD
- n is the sample size broken down by class (n_H and n_D)
- p total SNPs that survived quality control
- Baseline and 24-month follow-up scans were available
- Three studies were performed:

Comparison	n	n_H	n_D	p
AD vs CN	254	101	153	322875
P-MCI vs CN	260	107	153	309730
P-MCI vs S-MCI	221	107	114	304209

ADNI: Imaging Data

1. Data preprocessing:
   - Baseline and 24 month follow-up MR images available (Oct 2010)
   - Follow-up scans aligned with baseline scans using non-rigid registration
   - Jacobian determinants were extracted from the resulting deformation fields and represent the expansion/contraction on a voxel basis
   - After extracting Jacobian maps for all subjects, they were transformed to a template (using non-rigid registration) estimated for baseline scans
   - $q = 1,650,857$ voxel intensities (Jacobian determinants) representing longitudinal changes corrected for age and sex

2. Disease imaging signature extraction (voxel selection):
   - For each comparison, we identified regions of *highly discriminative* voxels
   - Penalised linear discriminant analysis (pLDA) with data resampling
   - Predictive power of disease signature was assessed using SVMs

Linear Discriminant Analysis (LDA)

- LDA finds a linear combination of the voxels intensities that best discriminates between classes - it finds a projection vector $\mathbf{w}$ such that $\mathbf{Yw}$ gives best linear discrimination
- We have two classes: H (Healthy) and D (Diseased) - the mean vectors for H and D, and the overall mean are, respectively

$$\mathbf{m}_H = \frac{1}{n_H}\sum_{i\in H}\mathbf{y}_{i\cdot} \quad \mathbf{m}_D = \frac{1}{n_D}\sum_{i\in D}\mathbf{y}_{i\cdot} \quad \mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{i\cdot}$$

- The *between-class* scatter matrix is

$$\Sigma_B = (\mathbf{m}_H - \mathbf{m}_D)'(\mathbf{m}_H - \mathbf{m}_D)$$

- The *within-class* scatter matrix is

$$\Sigma_W = \sum_{i\in H}(\mathbf{y}_{i\cdot} - \mathbf{m}_H)'(\mathbf{y}_{i\cdot} - \mathbf{m}_H) + \sum_{i\in D}(\mathbf{y}_{i\cdot} - \mathbf{m}_D)'(\mathbf{y}_{i\cdot} - \mathbf{m}_D)$$

LDA Solution

- The optimal direction vector $\mathbf{w}$ solves

$$\max_{\mathbf{w}} \frac{\mathbf{w}'\Sigma_B\mathbf{w}}{\mathbf{w}'\Sigma_W\mathbf{w}}$$

- This is equivalent to

$$\max_{\mathbf{w}} \mathbf{w}'\Sigma_B\mathbf{w} \quad \text{subject to} \quad \mathbf{w}'\Sigma_W\mathbf{w} = 1$$

- This solution involves *all* the q voxels, but we want to filter non-informative voxels out

Penalised LDA

- We assume a diagonal within-class scatter matrix, $\mathbf{S}_W$, where $\text{diag}(\mathbf{S}_W) = (s_1^2, \ldots, s_q^2)$ and call $\mathbf{S}_B$ the estimated between-group scatter matrix
- Impose $\mathbf{w}$ to be *sparse* by adding a penalty term, and carry out constrained maximisation

$$\max_{\mathbf{w}} \left\{ \mathbf{w}'\mathbf{S}_B\mathbf{w} - \lambda \sum_{j=1}^{q} s_j |w_j| \right\}$$

subject to

$$\mathbf{w}'\mathbf{S}_W\mathbf{w} = 1$$

- Since the objective function is non-convace, standard convex optimization methods cannot be used, but we use a *minorization-maximization* algorithm
  - find a concave function that minorizes the objective function
  - then use convex maximisation

Parameter Tuning and Image Classification Results

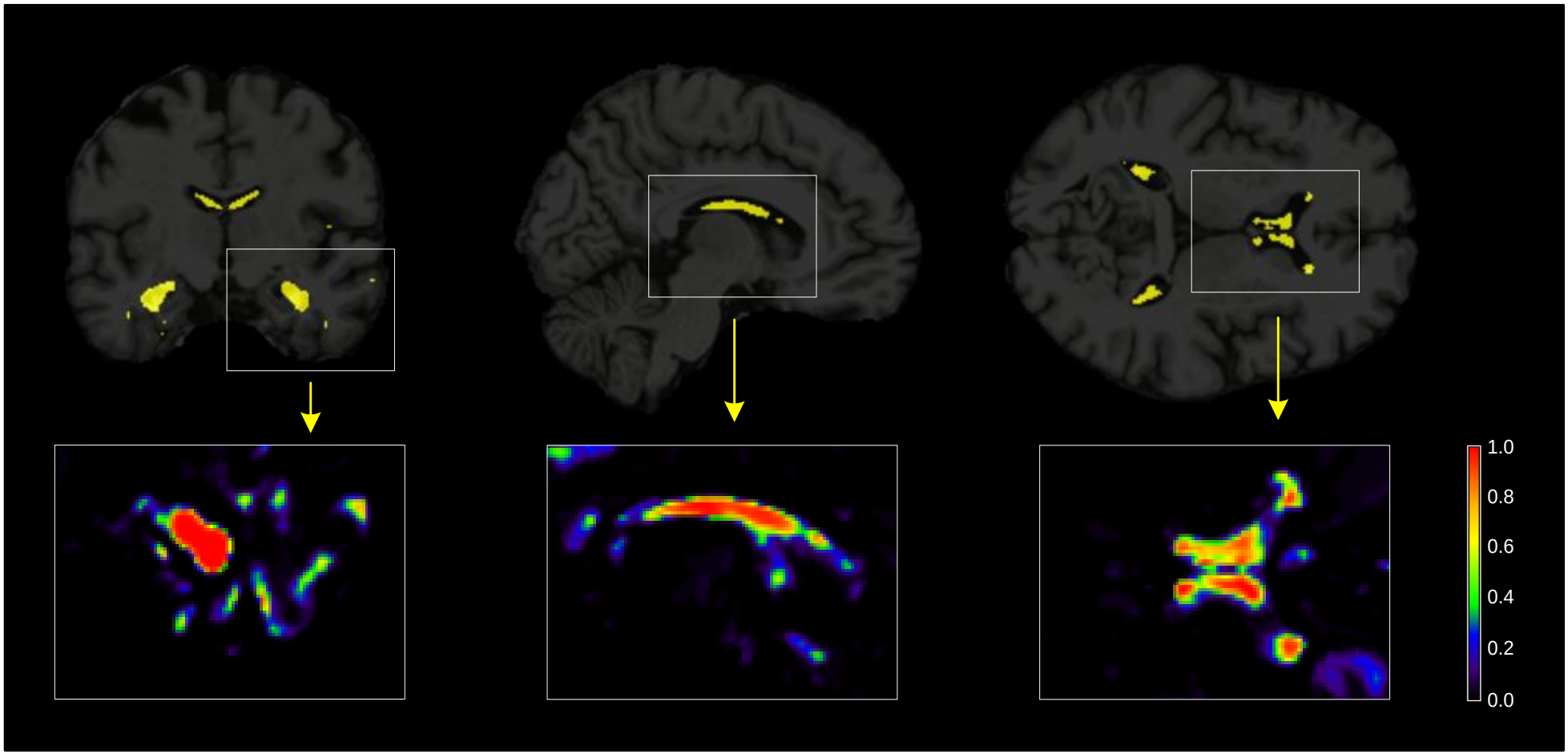
Jasounova *at al* (2011)

- We use a data resampling strategy combined with a non-linear Support Vector Machine classifier to (a) select the optimal number of voxels, (b) obtain a sparse image classification
- `vox` is the number of selected voxels using sparse LDA
- 10-fold cross validated performance measures: accuracy (`acc`), sensitivity (`sen`) and specificity (`spe`)

Experiment	`vox`	`acc`	`sen`	`spe`
AD vs CN	11394	90.3	87.5	92.1
P-MCI vs CN	12664	86.9	81.2	90.9
P-MCI vs S-MCI	10593	82.1	81.5	82.9

AD Imaging Signature

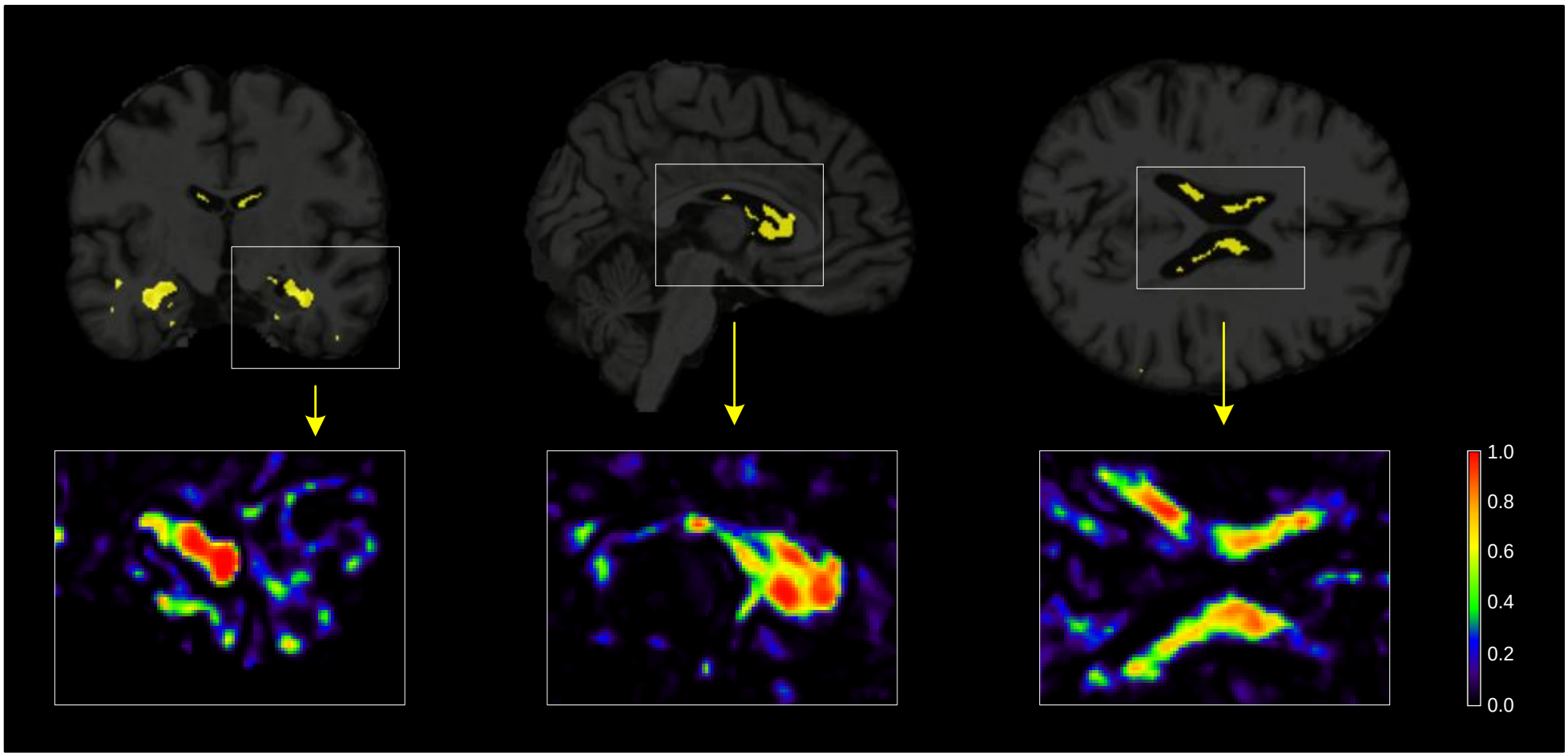
The selected voxels are in yellow (coronal, sagittal and axial view from left to right).



- Many informative voxels cluster in the hippocampus and lateral ventricles
- Also involved are the temporal lobe, amygdala and caudate nucleus

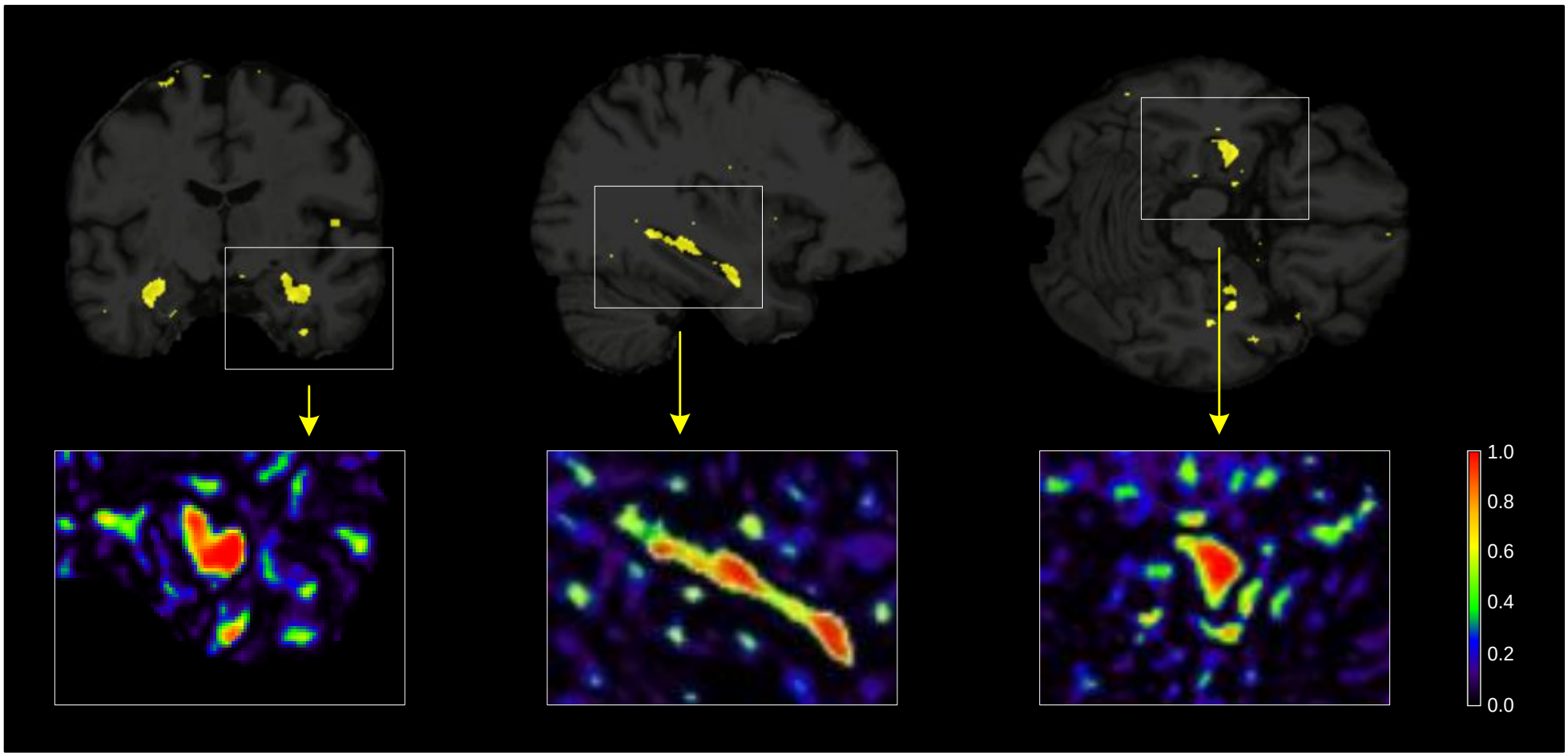
P-MCI Imaging Signature

The selected voxels are in yellow (coronal, sagittal and axial views from left to right).



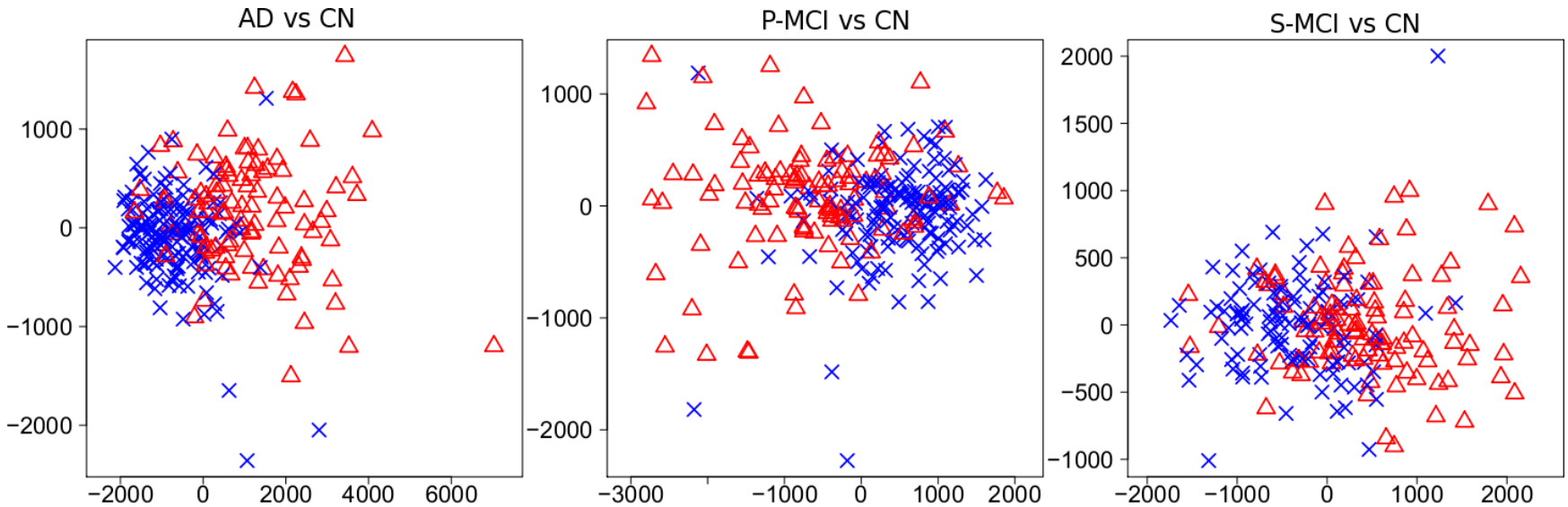
P-MCI/S-MCI Imaging Signature

The selected voxels are in yellow (coronal, sagittal and axial views from left to right).



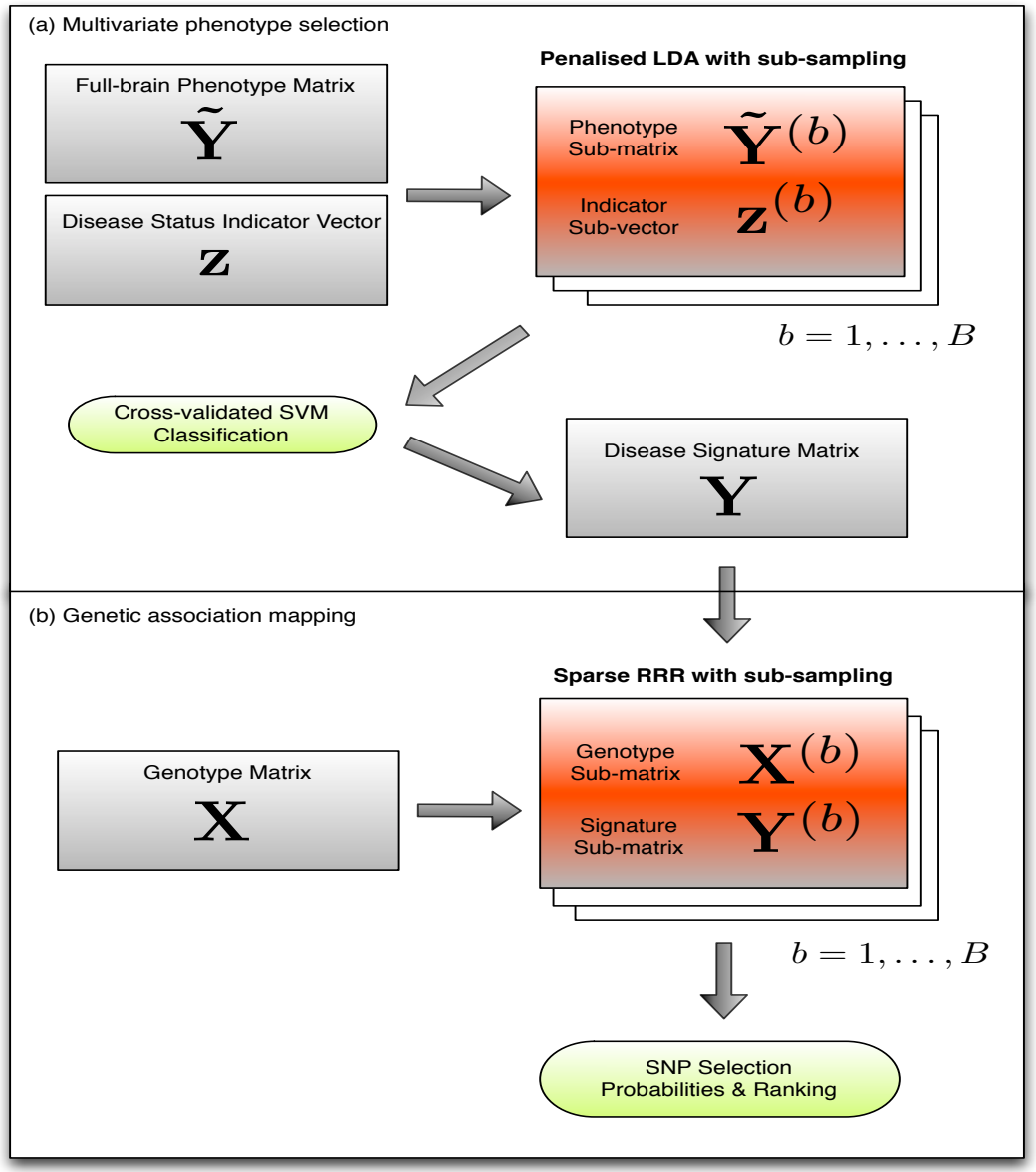
Sample Proximity using 2D Projections

MDS plots using the extracted imaging signatures



Voxel-wide GWA Analysis Flowchart

Vunou *et al* (2012)



AD: SNP Selection Probabilities

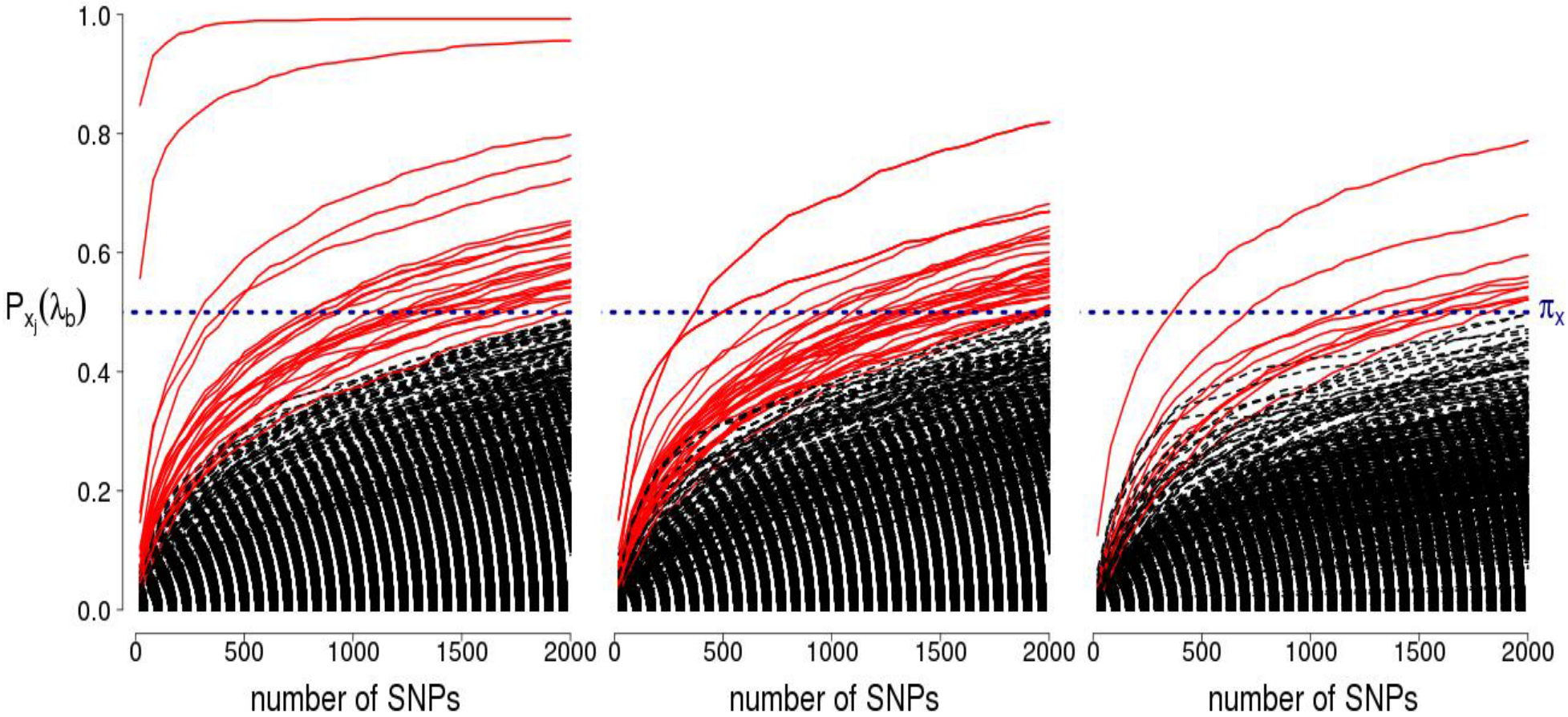


Figure: Ranks 1, 2 and 3 (from left to right).

AD: Top Ranked Genes

- APOE-ϵ4 (~ 1) - well know and replicated risk factor
- TOMM40 (0.96) - located in close proximity to the APOE gene, it has also been linked to AD in recent studies
- BZW1 (0.8) - no prior implication, but expressed in brain, differentially expressed in a microarray analysis on a mouse model related to a neurodegenerative disease (amyotrophic lateral sclerosis)
- PDZD2 (0.65) - interact with CST3, which is suspected to be implicated
- YES1 (0.5) - Three SNPs in the genes have high selection probability, a possible link between this gene and AD suggested in the literature

P-MCI: SNP Selection Probabilities

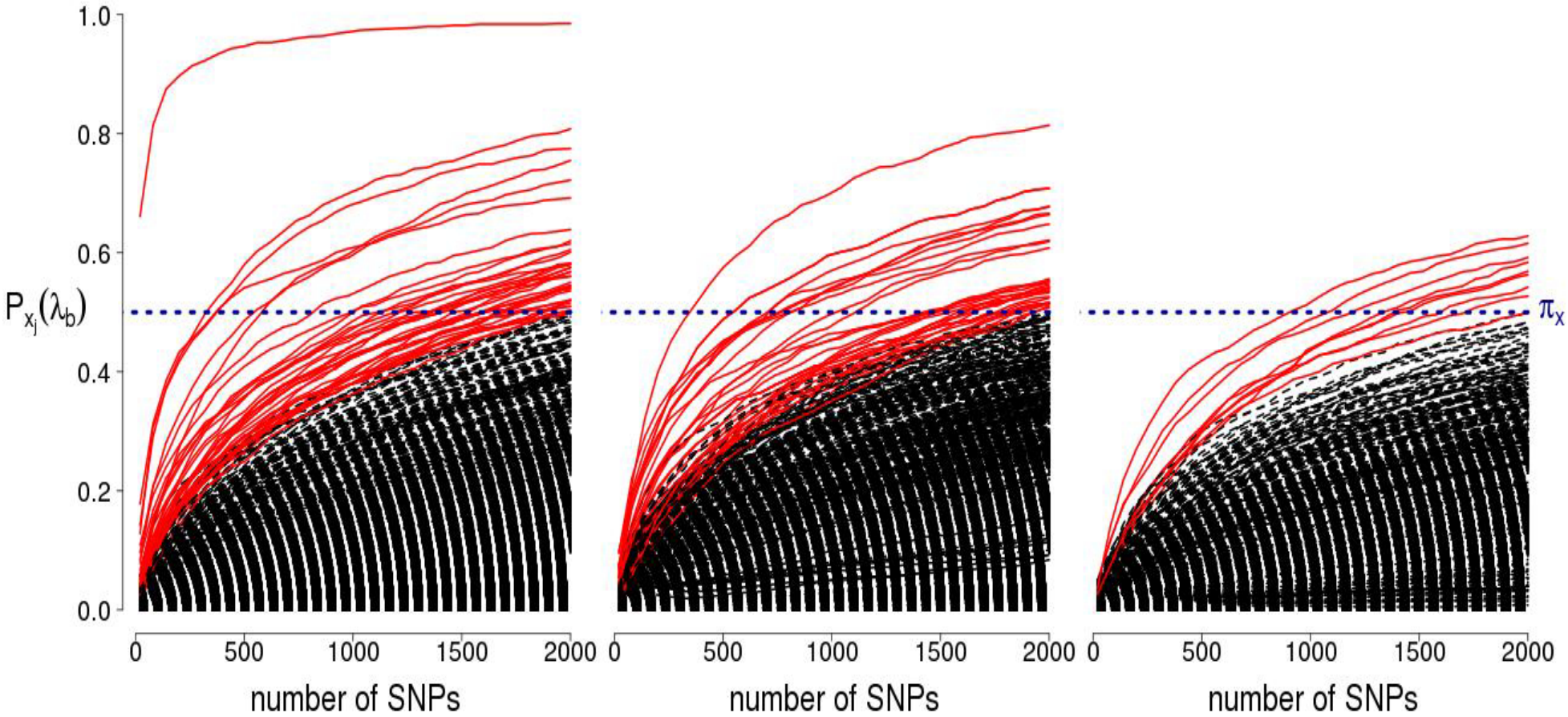


Figure: Ranks 1, 2 and 3 (from left to right).

P-MCI: Top Ranked Genes

- APOE-ϵ4 (~ 1)
- TOMM40 (0.59)
- RBFOX1 (0.57) - associated to autism, bipolar disorder, mental retardation and epilepsy
- COX7A2L (0.53) - belongs in the AD KEGG pathway and physical interactions between the key AD risk factor TOMM40 and COX7A2L have been previously reported

P-MCI/S-MCI: SNP Selection Probabilities

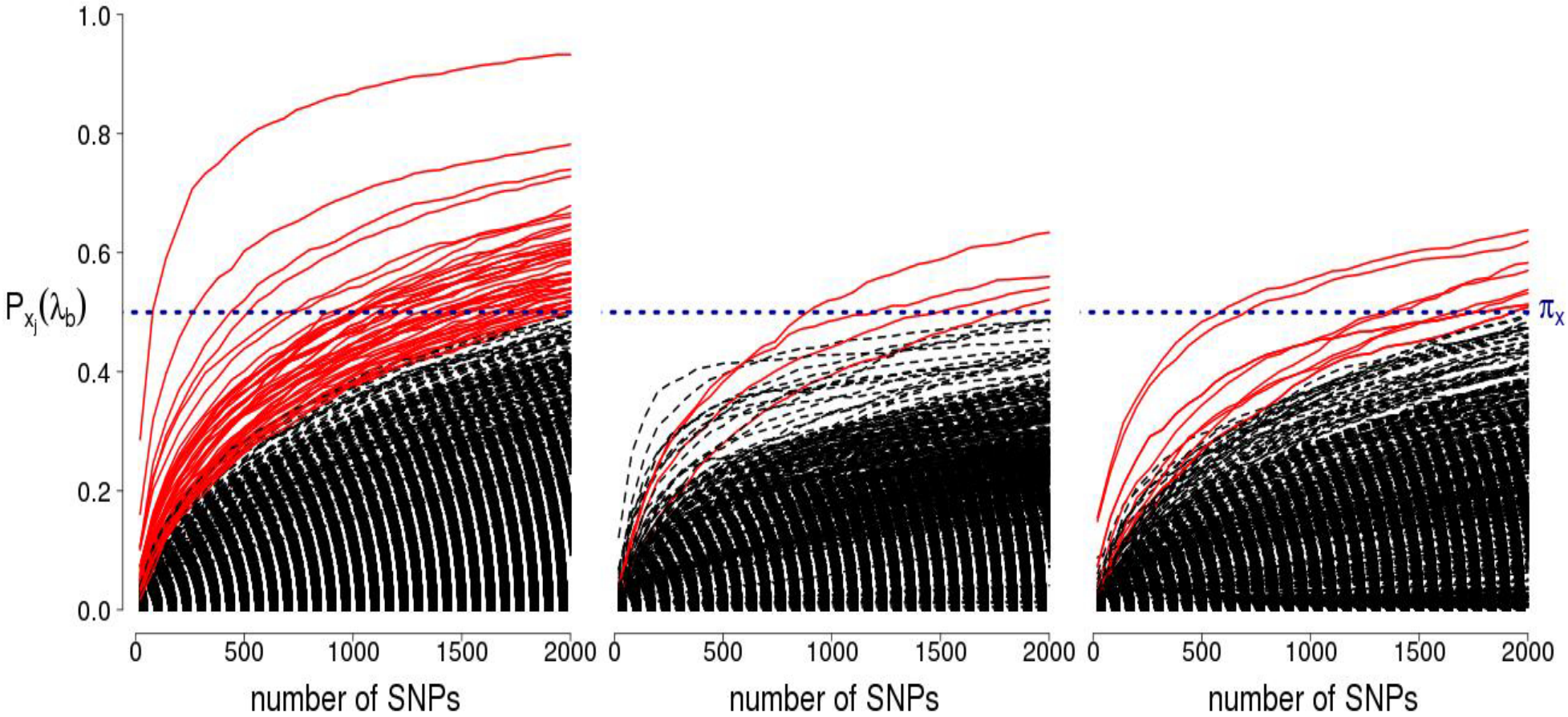


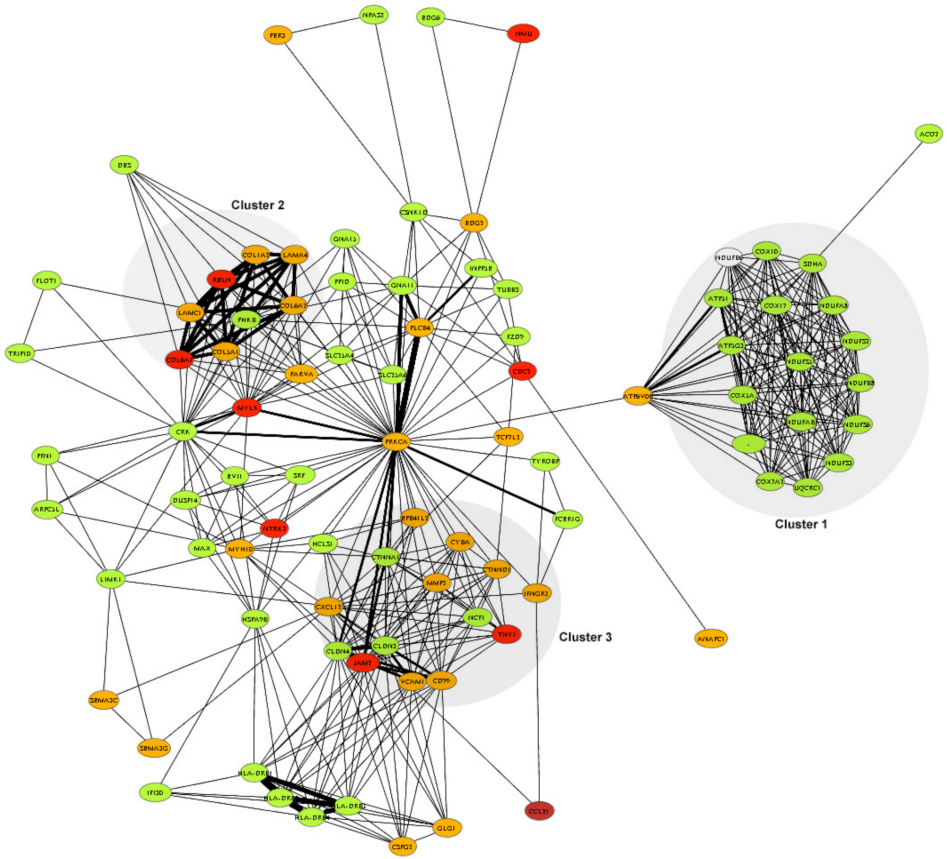
Figure: Ranks 1, 2 and 3 (from left to right).

P-MCI/S-MCI: Top Ranked Genes

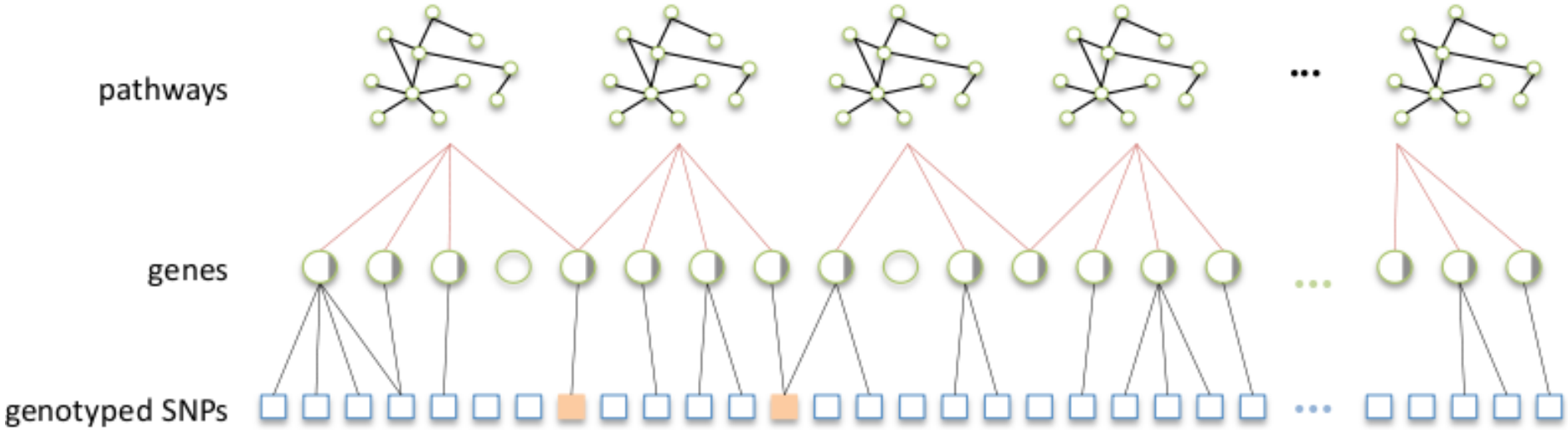
- APOE-ϵ4
- MGMT - using the Allen Brain Atlas, we confirmed that this gene is expressed in the brain regions where the selected voxels mostly lie
- Other previously unreported associations

From SNPs to Biological Pathways

- Genes act together in functionally related pathways
- Pathways GWAS can reveal aspects of a disease's genetic architecture that would otherwise be missed when considering variants individually
- Increase power due to the detection of coordinated small signals within pathways
- Easier biological interpretation and comparisons across studies



The SNPs to Genes to Pathway Mapping Process



- Known genes are mapped to known pathways, e.g. KEGG
- Many genes do not map to any known pathway (unfilled circles), some genes may map to more than one pathway.
- Genes that map to a pathway are in turn mapped to genotyped SNPs within a specified distance. Many SNPs cannot be mapped to a pathway since they do not map to a mapped gene (unfilled squares).
- SNPs may map to more than one gene and some SNPs (orange squares) may map to more than one pathway.

Existing Methods for Pathways Selection

- Existing methods are fundamentally *univariate*
  - SNPs are independently scored
  - The individual genetic effects are then combined over pathways
- For instance, GenGen (Wang *et al.*, 2007):
  - Rank all genes by assigning the value of the highest-scoring SNP within 500kb of each gene
  - Assess pathway significance by determining the degree to which high-ranking genes are over-represented or enriched in a given gene set, in comparison with genomic background
- No methods exist for *multivariate quantitative traits*

Pathways-based Sparse Regression Modelling

- We approach the problem differently: we include all the known pathways in a multivariate regression model so they can compete against each other
- The *assumption* is that, where causal SNPs are enriched in a pathway, a regression model that selects all SNPs *grouped into pathways* will have increased power
  - Group all available SNPs into L pathways $\mathcal{G}_1, \ldots, \mathcal{G}_L$
  - Adopt a *group lasso* penalty to force group selection,

$$P_b(\mathbf{b}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{b}||_2^2 + \lambda \sum_{l=1}^{L} w_l ||\mathbf{b}_l||_2$$

to select only most predictive pathways,

$$\mathbf{b} = \{\underbrace{(0, \ldots, 0)}_{\mathcal{G}_1}, \ldots, \underbrace{(0, \ldots, \mathbf{b}_{l_a}, 0 \ldots, \mathbf{b}_{l_b}, 0, \ldots, 0)}_{\mathcal{G}_l}, \ldots, \underbrace{(0, \ldots, 0)}_{\mathcal{G}_L}\}$$

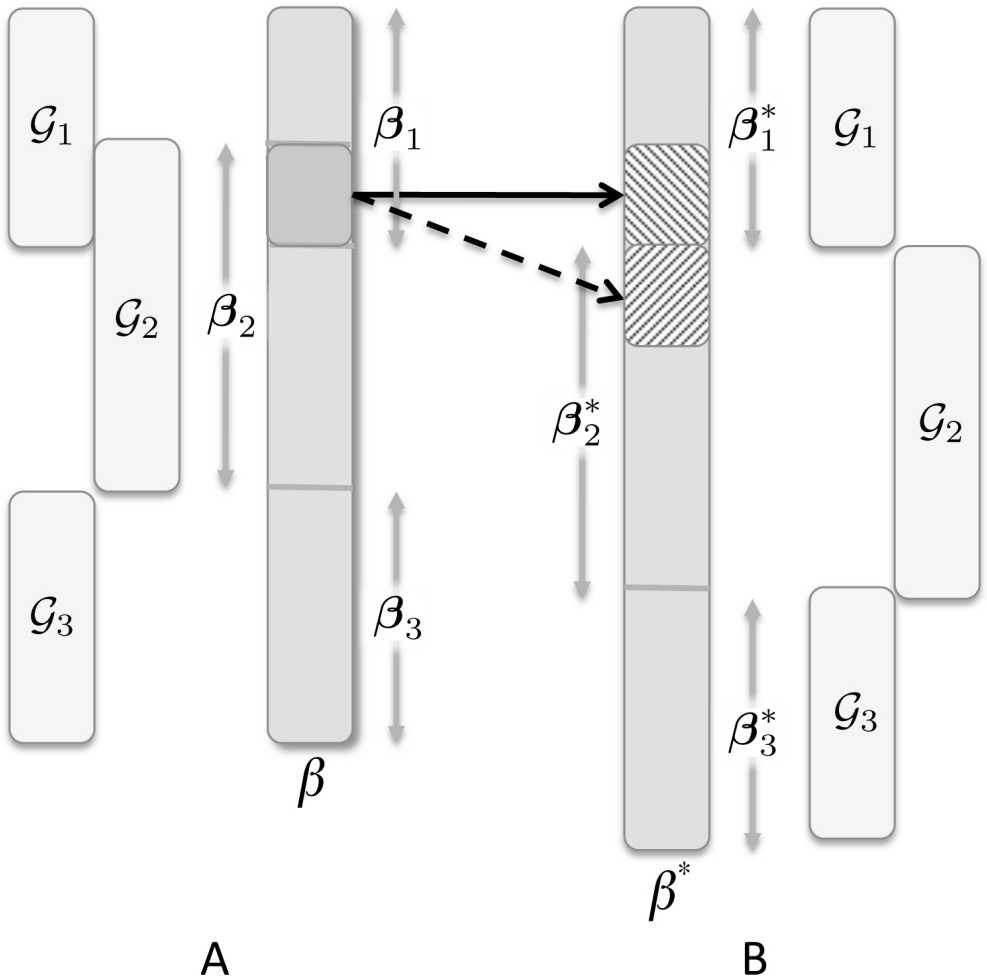
Challenges and Solutions

Silver and Montana (2011)

- Challenges:
  1. Pathways often *overlap*, since many SNPs map to multiple pathways
  2. *Selection bias* due to pathways heterogeneity in size, LD distribution, etc.
  3. Sheer scale of datasets, *efficient estimation* is a necessity
- Solutions:
  1. Expanded design matrix by SNP duplication (non-orthogonal groups)
  2. Adaptive pathway weights $\{w_1, \ldots, w_L\}$ for bias correction
  3. A fast iterative estimation procedure: block-coordinate descent (BCD+) algorithm for non-orthogonal groups

SNPs Duplication in Overlapping Pathways

Three pathways $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_2$ and grouped regression coefficients $\beta_1, \beta_2, \beta_3$



Pathways Group Lasso with Adaptive Weights

1. In order to control for group size the common choice is to use a weight

$$w_l = \sqrt{S_l}$$

   but other factors may bias the group selection process
2. In case of no association and no selection bias, a pathway $\mathcal{G}_l$ should be selected according to a *uniform distribution*, that is with probability Π_l
3. The empirical selection probability is called Π_l^*
4. We propose an adaptive strategy whereby the weights $\mathbf{w} = \{w_1, \ldots, w_L\}$ are tuned so that the distance D between Π and Π^* is minimised, where

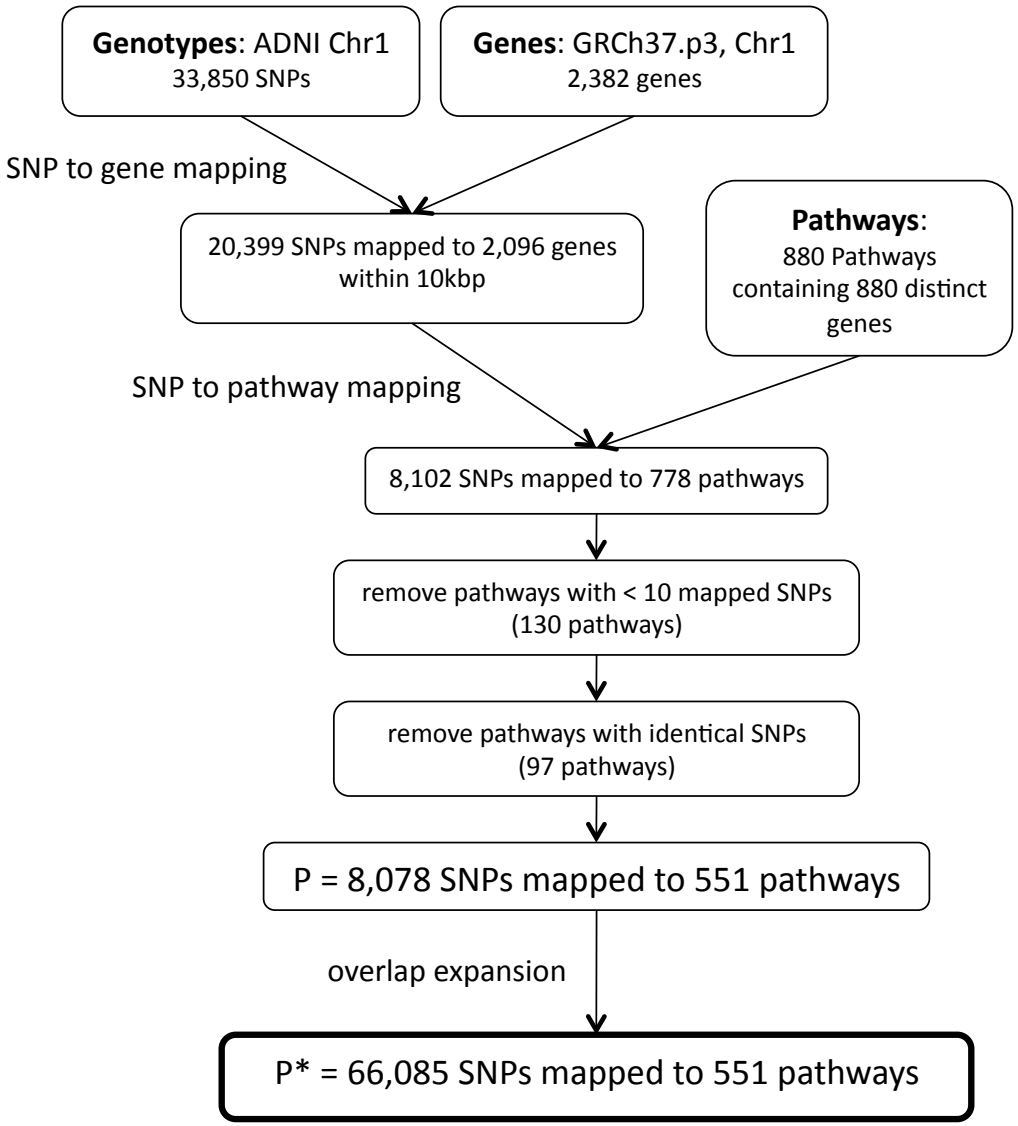
$$D = \sum_l \Pi_l^*(\mathbf{w}) \log \frac{\Pi_l^*(\mathbf{w})}{\Pi_l}$$

   is taken to be the Kullback-Leibler (KL) divergence

Fast Parameter Estimation Algorithms

1. Block coordinate descent (BCD) has generally used for group lasso with orthogonal groups
2. We propose a BCD+ algorithm for non-orthogonal groups due to the expanded design matrix
3. The proposed BCD+ algorithm is particularly fast as it relies on a number of techniques:
   - Taylor approximation of the group lasso penalty
   - *Active sets* strategy
   - Efficient computation of block regression residuals

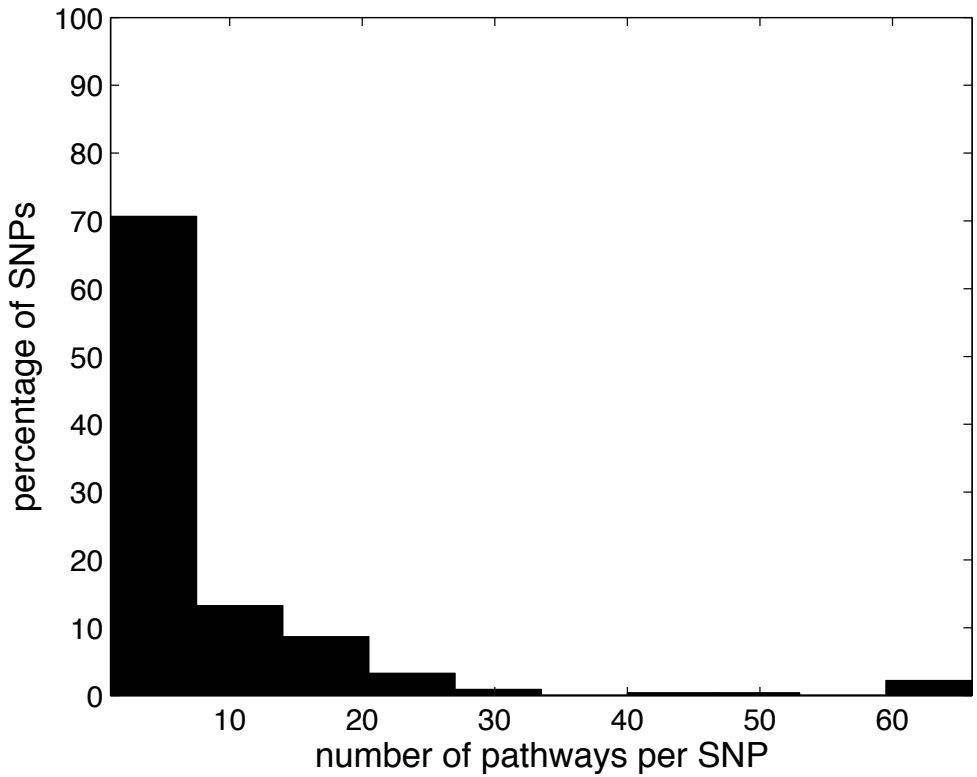
Simulation Studies: ADNI Data, Chromosome 1



Number of Pathways per SNP

Frequency distribution of ADNI SNPs by number of pathways they map to

- SNPs are mapped to genes within 10kbp
- $8,078$ SNPs and 551 pathways



Genetic Effects

- Randomly chose a causative pathway $\mathcal{G}_{\mathcal{C}}$
- Generate a set $\mathcal{S}$ of causal SNPs within $\mathcal{G}_{\mathcal{C}}$
- Form the set $\mathcal{C}$, of causal pathways that contain all the SNPs in $\mathcal{S}$
- Simulate a univariate quantitative phenotype,

$$y = \sum_{k \in \mathcal{S}} \zeta_k x_k + \epsilon$$

  - ζ_k is the allelic effect per minor allele due to causal SNP k
  - $\epsilon \sim \mathcal{N}(1, \sigma^2_\epsilon)$
- The *effect size* of each SNP k in S, that is

$$\delta_k = \mathrm{E}(\zeta_k x_k)/\mathrm{E}(y)$$

Simulation Scenarios

- Given the real ADNI genotype data, phenotypes are simulated
- We control three design factors:
  - The number of causative SNPs, $|\mathcal{S}|$
  - Where the causative SNPs are in the pathway
  - The effect size, δ_k (same for all causative SNPs)

scenario	$\|\mathcal{S}\|$	δ_k	distribution	description
(a)	10	0.005	random from $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ large; δ_k large; random distribn
(b)	3	0.005	random from $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ small; δ_k large; random distribn
(c)	3	0.005	random from single gene in $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ small; δ_k large; single gene
(d)	10	0.001	random from $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ large; δ_k small; random distribn
(e)	3	0.001	random from $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ small; δ_k small; random distribn
(f)	3	0.001	random from single gene in $\mathcal{G}_{\mathcal{C}_1}$	$\|\mathcal{S}\|$ small; δ_k small; single gene

BCD+ vs BCD: Computational Speed-ups

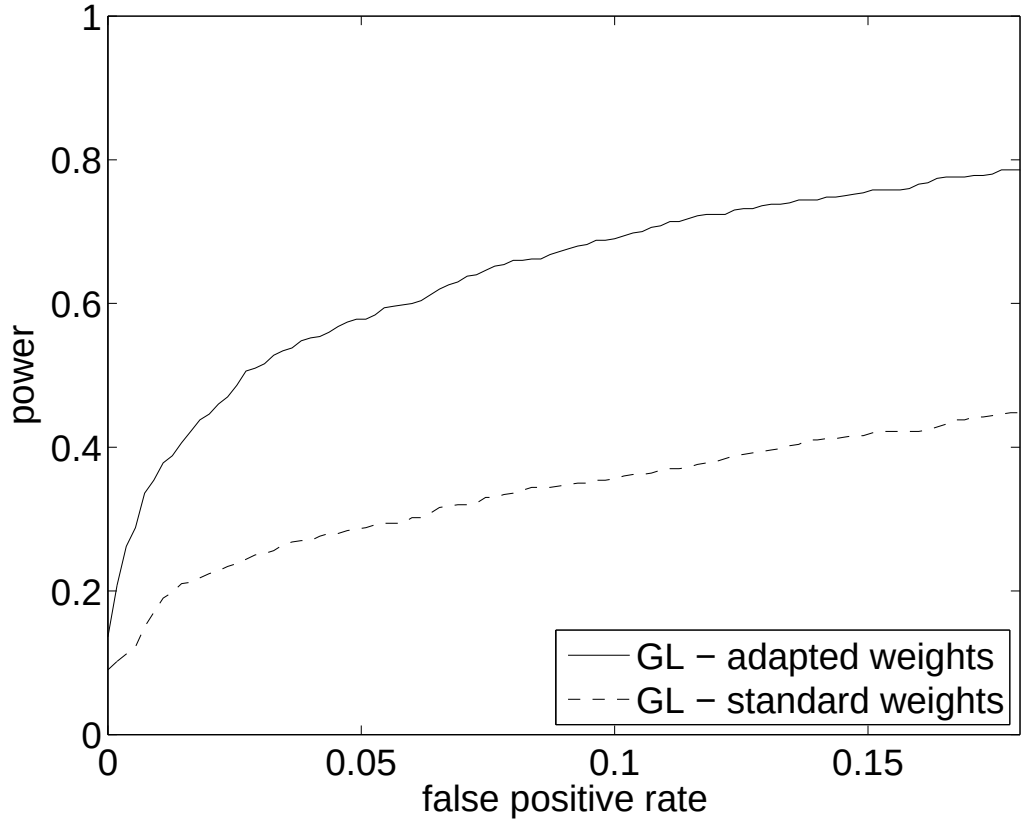
- BCD: block coordinate descent for non-orthogonal groups
- BCD+: improved BCD version with speed-ups
- We report on estimation times (seconds)
- Computations performed using multi-threading on a single machine with 8 3.2 GHz processors and 64GB RAM.

	$P^* = 4$k		$P^* = 66$k		$P^* = 647$k	
sample size	BCD	BCD+	BCD	BCD+	BCD	BCD+
371 ($N/2$)	7.93	0.17	421	1.35	5490	16
743 (N)	16.9	0.27	511	2.5	6430	30.0

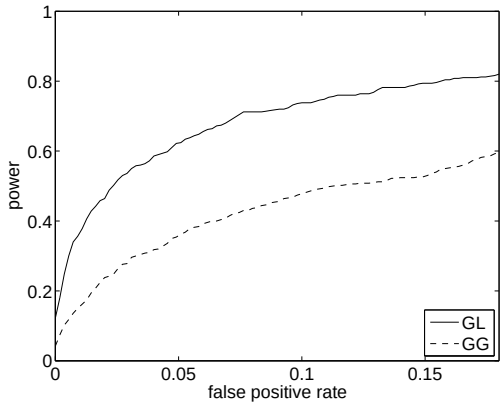
Adaptive Weights: Power Advantages

Adaptive weighting scheme vs. standard pathway size weighting

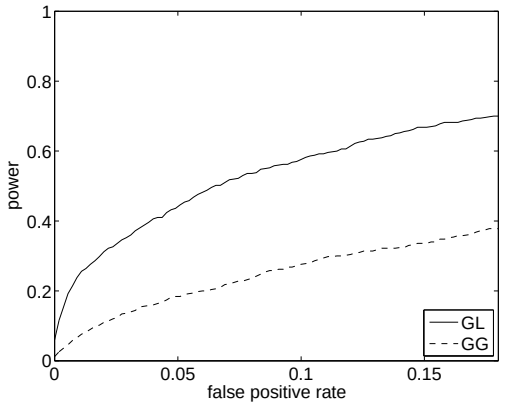
- $|\mathcal{S}| = 10; \delta_k = 0.005$
- SNPs randomly distributed across causative pathway $\mathcal{G}_\mathcal{C}$.
- ROC curves illustrating power to identify at least one causal pathway in the top 100. Power is average across 500 simulations.



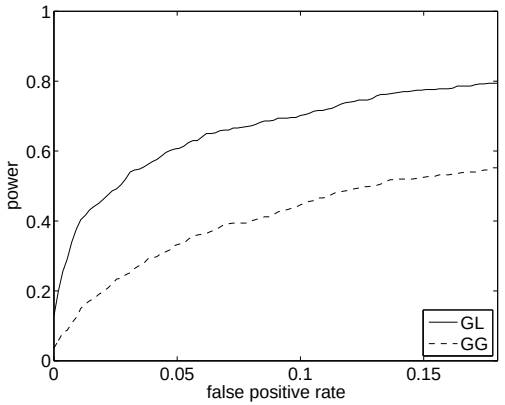
Power Comparisons to GenGen



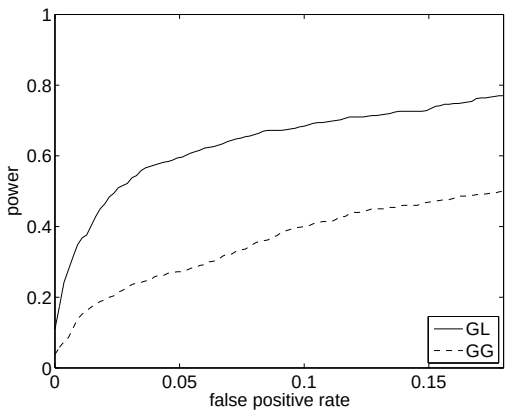
(a) $|\mathcal{S}| = 10; \delta_k = 0.005$; random



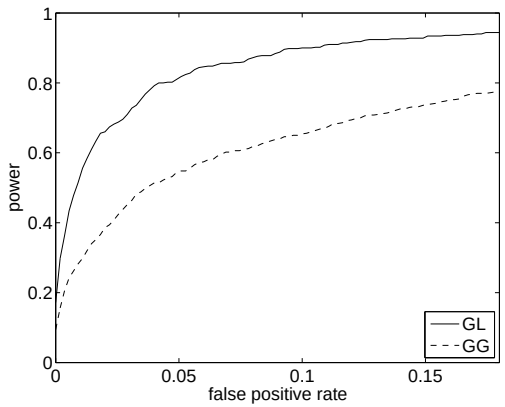
(b) $|\mathcal{S}| = 10; \delta_k = 0.001$; random



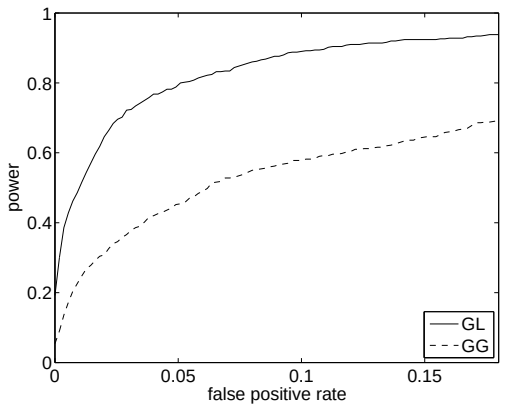
(c) $|\mathcal{S}| = 3; \delta_k = 0.005$; random



(d) $|\mathcal{S}| = 3; \delta_k = 0.001$; random



(e) $|\mathcal{S}| = 3; \delta_k = 0.005$; single gene



(f) $|\mathcal{S}| = 3; \delta_k = 0.001$; single gene

ADNI-1: Available Samples

- Serial brain MRI scans were analyzed from 200 probable AD patients and 232 healthy elderly controls (CN)
- Longitudinal scans available at three time points

	Screening	6Mo	12Mo	24Mo
AD	200	165	144	111
CN	232	214	202	178
Total	432	379	346	289

At screening:

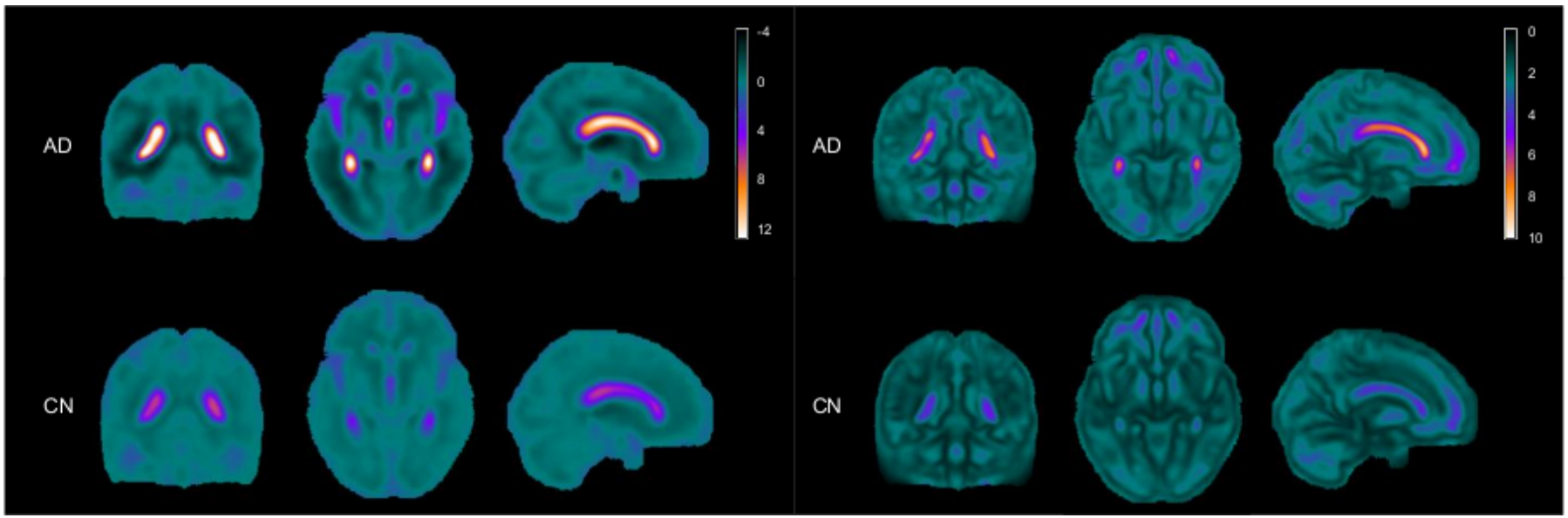
Group	age (years)	N male	N female
AD	75.7±7.7	103	97
CN	76.0±5.0	120	112

ADNI-1: Imaging Data Preprocessing and Voxel Filtering

1. Individual Jacobian maps were created to estimate 3D patterns of structural brain change over time - longitudinal maps of tissue change were spatially normalized across subjects by nonlinearly aligning all individual Jacobian maps to an average group template minimal deformation target (MDT)
2. For each one of the $Q^* = 2,153,231$ voxels, we obtain a single real-value measurement that capture the temporal changes at that voxel by fitting a linear regression model with *time* as covariate and use estimated *slope* as the associated phenotype
3. All voxels where the difference in the slopes in AD vs CN is not significantly difference from zero are removed, while also controlling for sex and age as covariates - the family-wise error rate is controlled by using a Bonferroni correction, and $Q = 148,023$ voxels are retained

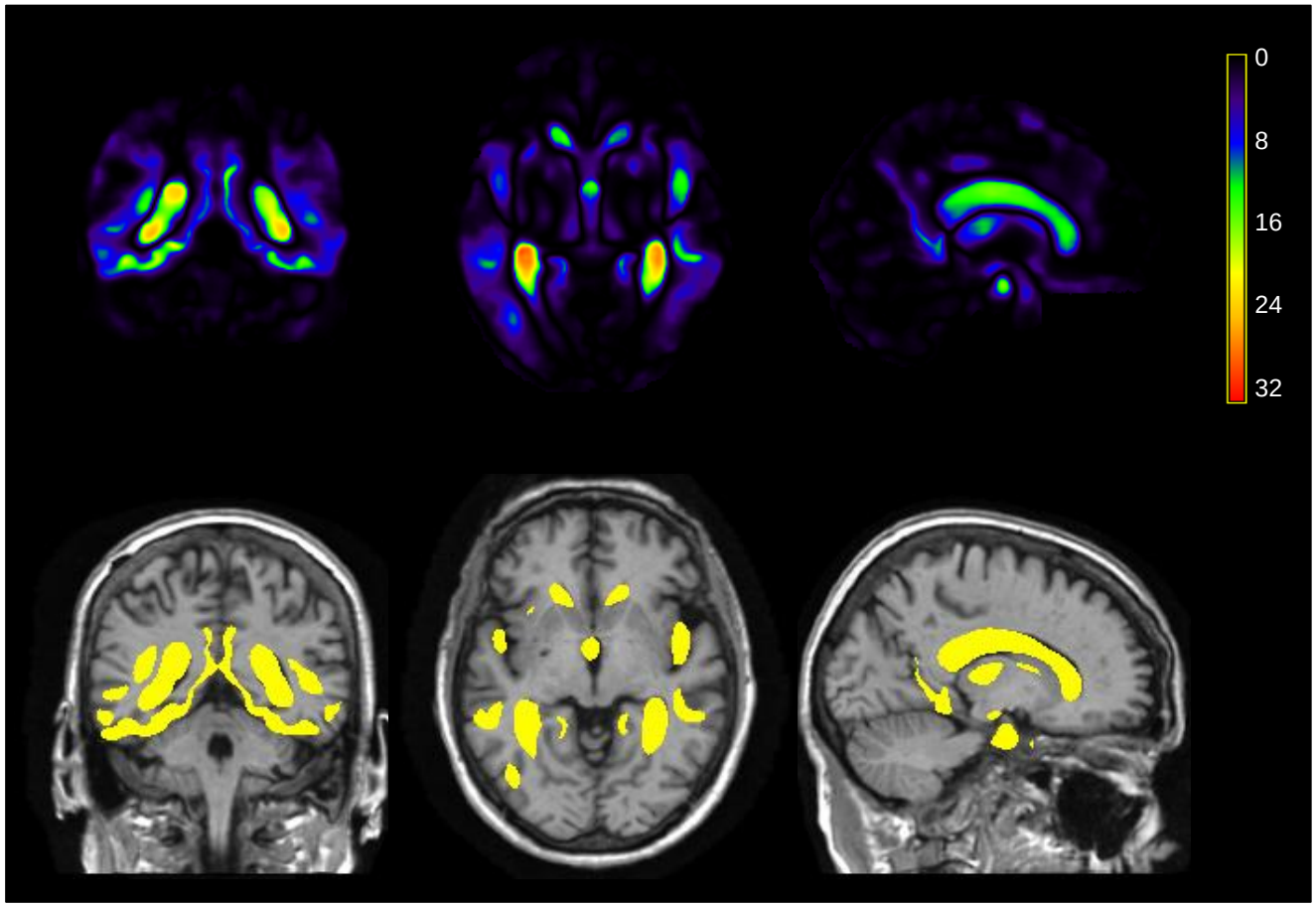
AD Imaging Signature

Mean and Std Dev Maps of Slope Coefficients



- Increased expansion of ventricular volumes is clear in all subjects, but is most marked in AD patients, where ventricular volumes expand by an average 1.2% per year (whiter regions in left hand subplot).
- AD patients show the most variation in structural change over time

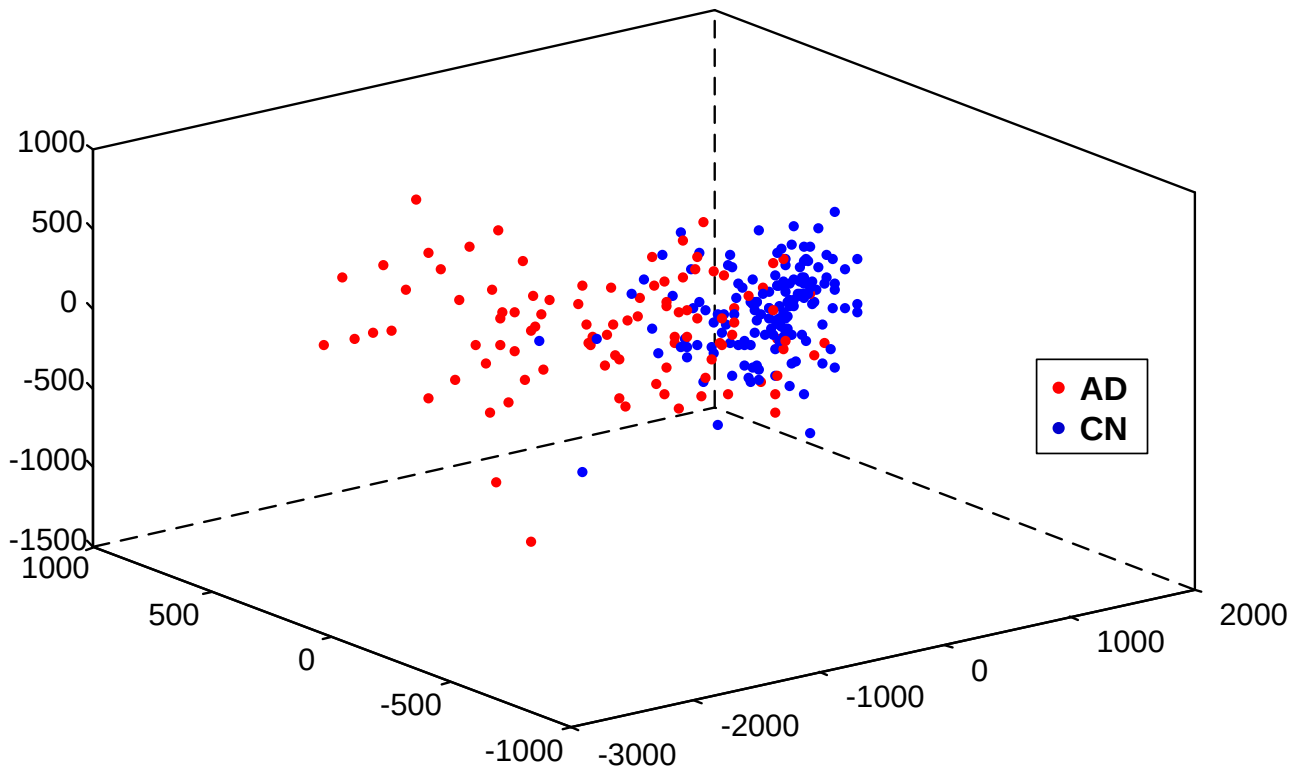
AD Imaging Signature: P-values Map



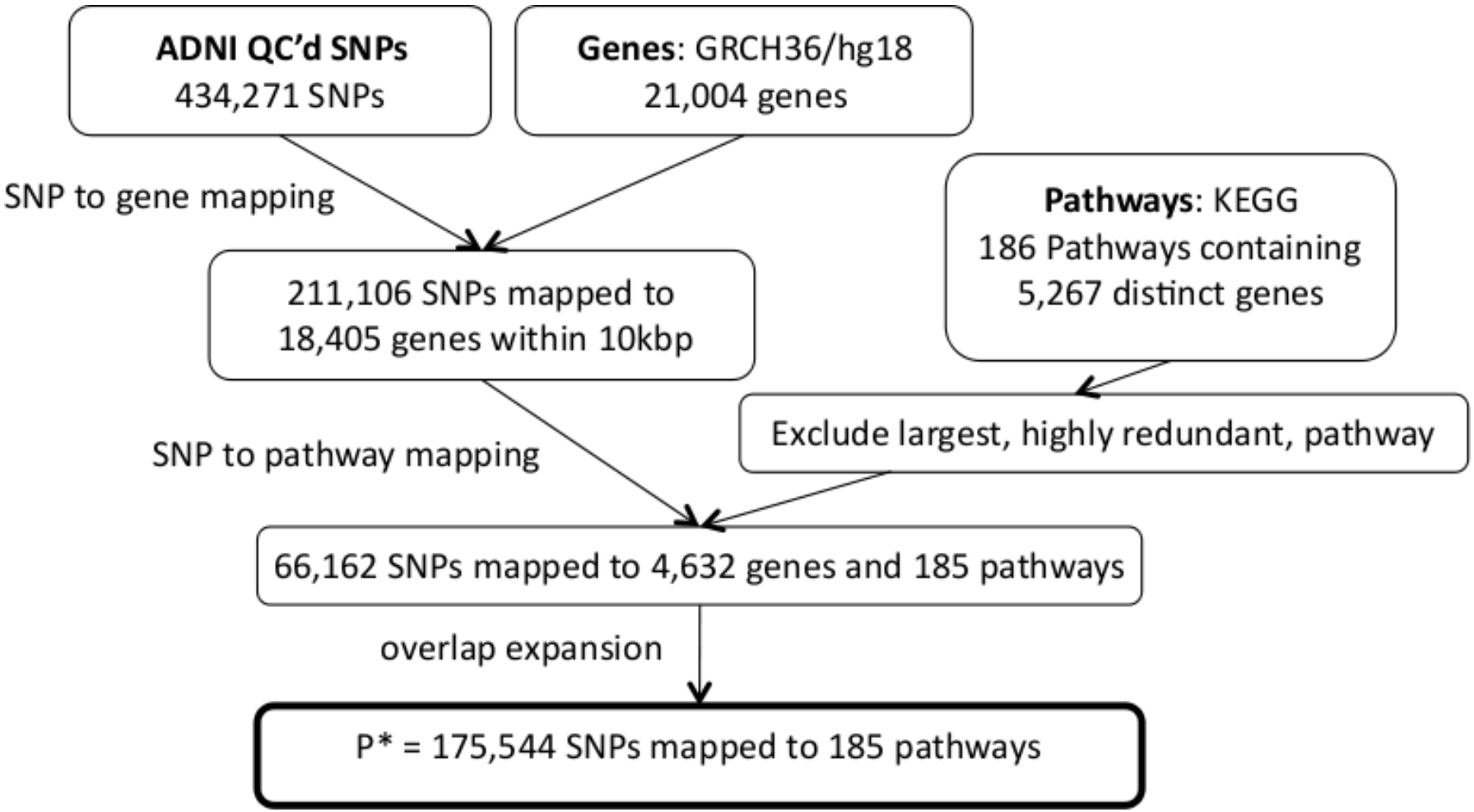
- P-values ($-\log_{10}$ scale) obtained from voxelwise ANOVA models
- The final set of $Q = 148,023$ selected voxels with p-values exceeding a Bonferroni-corrected threshold $\alpha_B = 0.05/2153231, (-\log_{10} \alpha_B = 7.6)$ are highlighted in yellow.

Sample Proximities using 3D Projections

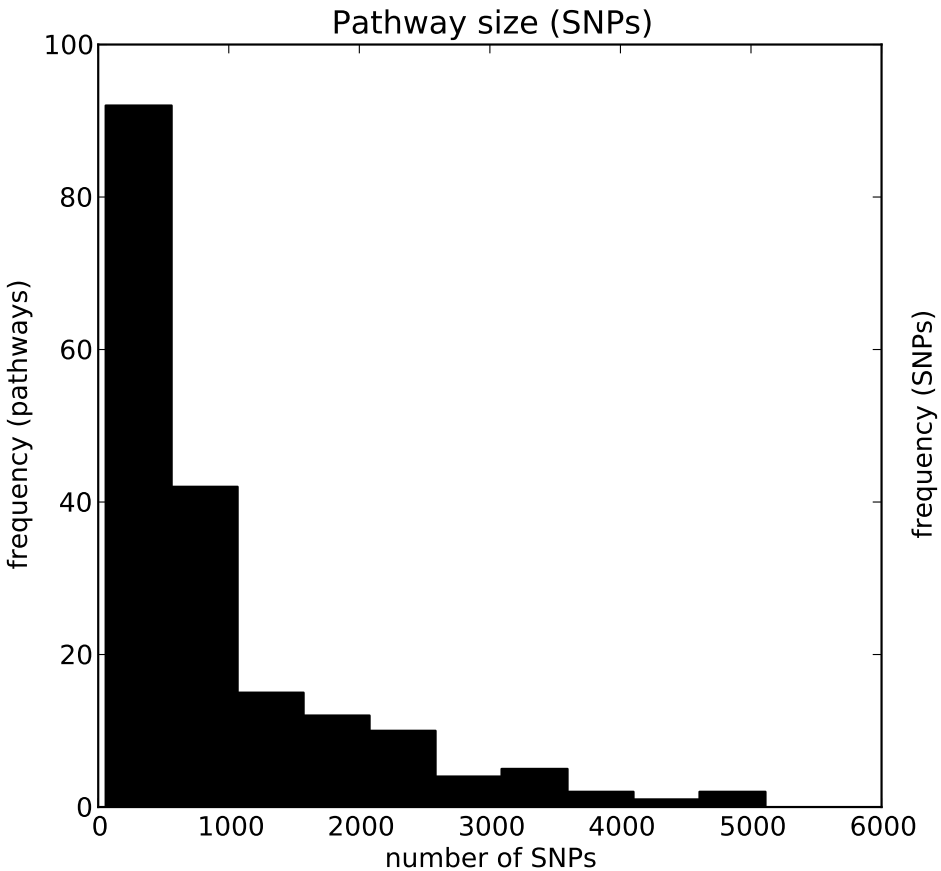
MDS plot illustrating the spread of imaging signatures across samples

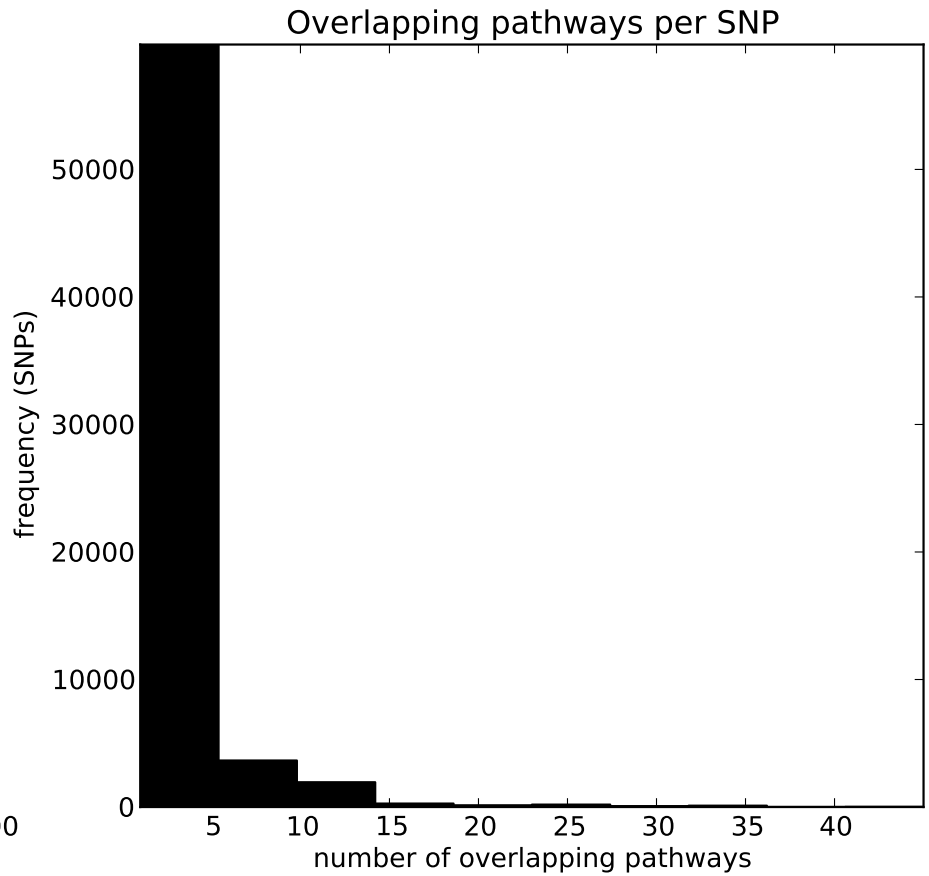


Mapping SNPs to Pathways



Pathway Sizes and SNP Overlaps





Top 15 Pathways Ranked by the PsRRR Algorithm

Selection Frequencies obtained from 1000 subsamples

Rank	KEGG pathway name	π^{path}	Size
1.	Insulin signaling pathway	0.524	1517
2.	Vascular smooth muscle contraction	0.456	3236
3.	Melanogenesis	0.331	1638
4.	Focal adhesion	0.232	4009
5.	Gap junction	0.180	2350
6.	Huntingtons disease	0.155	1980
7.	Purine metabolism	0.154	2896
8.	Pyruvate metabolism	0.153	456
9.	Propanoate metabolism	0.152	471
10.	Amyotrophic lateral sclerosis als	0.151	865
11.	Chemokine signaling pathway	0.145	2769
12.	Phosphatidylinositol signaling system	0.138	2067
13.	Citrate cycle tca cycle	0.137	210
14.	Glycosphingolipid biosynthesis globo series	0.135	227
15.	Alzheimers disease	0.127	2500

Top 15 SNPs and Genes Ranked by sRRR

SNP and gene ranking performed on the highly-ranked pathways

	SNP RANKING			GENE RANKING		
Rank	SNP	π^{SNP}	Mapped gene(s)	Gene	π^{gene}	# SNPs
1	rs4788426	0.451	*PRKCB*	PRKCB	0.451	73
2	rs11074601	0.429	*PRKCB*	ADCY8	0.411	69
3	rs263264	0.411	*ADCY8*	ADCY2	0.392	106
4	rs13189711	0.392	*ADCY2*	HK2	0.302	28
5	rs680545	0.302	*HK2*	PRKCA	0.290	99
6	rs4622543	0.290	*PRKCA*	PIK3R3	0.267	9
7	rs9896483	0.274	*PRKCA*	MYLK	0.234	24
8	rs1052610	0.267	*PIK3R3*	PIK3CG	0.207	9
9	APOϵ4	0.251	*TOMM40 APOE*	COL5A3	0.174	14
10	rs1254403	0.234	*MYLK*	GNAI1	0.167	22
11	rs4730205	0.207	*PIK3CG*	ACACA	0.164	23
12	rs889130	0.174	*COL5A3*	G6PC	0.163	6
13	rs6973616	0.167	*GNAI1*	DGKA	0.160	3
14	rs9906543	0.164	*ACACA*	CR1	0.154	21
15	rs2229611	0.163	*G6PC*	TOMM40	0.152	6

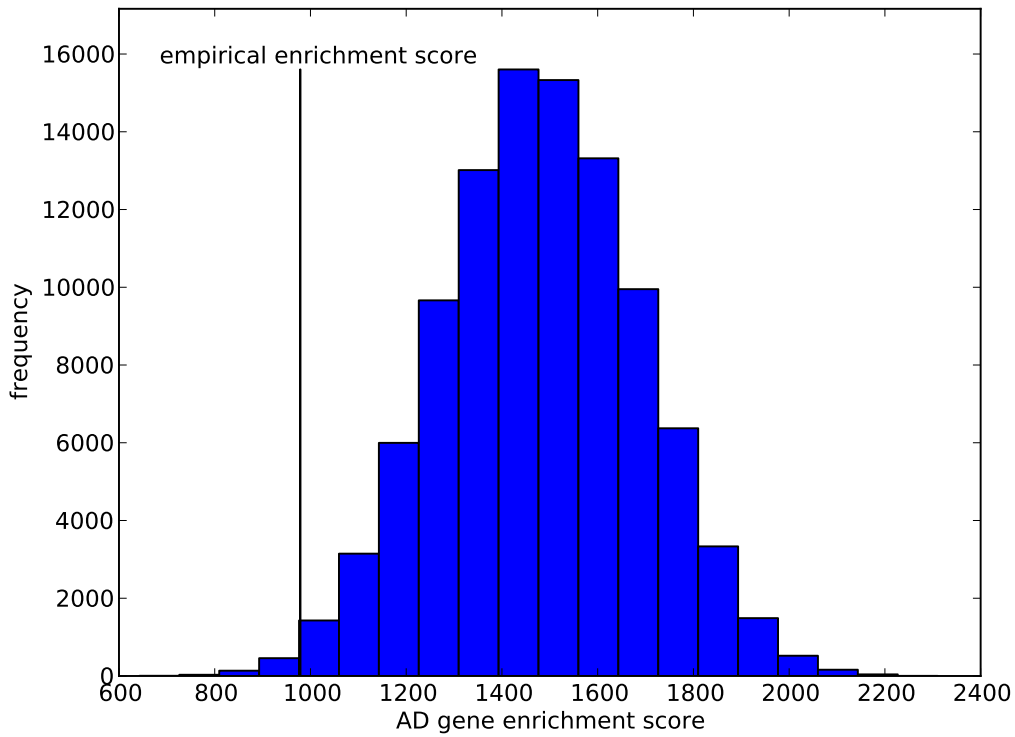
Known AD Genes Included in the Study

- 12 out of 30 known genes in Braskie, Ringman, and Thompson, 2011
- These genes: (a) map to a KEGG pathway and (b) have a genotyped SNP within 10kbp.

Implicated gene	Mapped genes in study
TOMM40	*TOMM40 APOE PVRL2*
ACE	*ACE*
EPHA4	*EPHA4*
CCR2	*CCR2 CCR5*
APOE	*TOMM40 APOE PVRL2*
FAS	*FAS*
CHRNB2	*ADAR CHRNB2*
EFNA5	*EFNA5*
LDLR	*LDLR*
CR1	*CR1 CR2*
GRIN2B	*GRIN2B*
IL8	*IL8*

AD Genes Enrichment Score in Top Pathways

- Distribution of AD gene enrichment scores obtained when permuting pathway rankings $100,000$ times.
- The vertical black line indicates the observed AD gene enrichment score using the true pathway rankings obtained in the study.
- The AD gene enrichment score has p-value $p = 0.0051$.



Biological Relevance

- High-ranking, AD endophenotype-associated pathways include those describing *insulin signalling*, *vascular smooth muscle contraction* and *focal adhesion*- all known to be implicated in AD biology
- Other functions previously associated with AD biology among high-ranking pathways include those related to *focal adhesion, gap junctions, chemokine signalling and phosphatidylinositol signalling*
- High ranking genes include a number previously linked in gene expression studies to β-amyloid plaque formation in the AD brain (PIK3R3; PIK3CG; PRKCA and PRKCB), and to AD related changes in hippocampal gene expression (ADCY2, ACTN1, ACACA, GNAI1).
- Other high ranking previously validated AD endophenotype-related genes include CR1, TOMM40 and APOE.

Conclusions

- Sparse reduced-rank regression model combined with a data resampling scheme provides a strategy for SNP and phenotype prioritisation and ranking
- Different penalties induce different sparsity patterns and allow prior knowledge (e.g. about gene-gene interactions or phenotypic structures) to be easily included in the model
- Extensive realistic simulation results show that sRRR is more powerful than mass univariate linear modelling and other models for pathways selection
- Real studies on Alzheimer's disease have confirmed known causal variants and pathways implicated with the AD biology
- Other studies on Multiple Sclerosis (not presented here) have also confirmed that sRRR is a valid approach for neuroimaging genetics

References

1. Vounou M., Nichols T. and Montana G. (2010) *Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach*. NeuroImage.
2. Vounou, M. *al* (2012) *Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimers disease*. NeuroImage
3. Jasounova, E. *et al* (2012) *Biomarker discovery for sparse classification of brain images in Alzheimers disease*. Annals of Computer Vision Association. *To appear*
4. Silver, M. and Montana, G. (2012) *Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps*. Statistical Applications in Genetics and Molecular Biology.
5. Silver *et al* (2012) *Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression*. Preprint

Software Availability

Coming up soon ...

- Open-source software will be released later this year
- R and Python libraries for sRRR model fitting and data resampling using CUDA for GPU computing
- Python library for P-sRRR fitting and resampling using Parallel Python
- Python Scripts for SNP-to-Pathways mapping
- Email me: `g.montana@ic.ac.uk`