Statistical Challenges for Neuroimaging Data Analysis

Lecture 4. Imaging Genetics

WNAR 2018 @ Edmonton
Linglong Kong
lkong@ualberta.ca

UNIVERSITY OF ALBERTA

"uplifting the whole people"

— HENRY MARSHALL TORY, FOUNDING PRESIDENT, 1908

Reading materials (incomplete list)

1. H. Ganjgahi, A.M. Winkler, D.C. Glahn, J. Blangero, P. Kochunov, T.E. Nichol (2015) Fast and powerful heritability inference for family-based neuroimaging studies. Neuroimage, in press.
2. P. Hibar, J. L. Stein, M. Renteria, A. Arias-Vasquez, S. Desrivieres, N. Jahanshad, et al., "Common genetic variants influence human subcortical brain structures," Nature (in press), 2014.
3. Y. Zhang, Z. Xu, X. Shen, W. Pan, for the ADNI. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. NeuroImage. 96, 309-325.
4. J. L. Stein, S. E. Medland, A. A. Vasquez, D. P. Hibar, R. E. Senstad, A. M. Winkler, et al., "Identification of common variants associated with human hippocampal and intracranial volumes," Nature Genetics, vol. 44, pp. 552-561, 2012.
5. Ge, Tian, Nichols, Thomas E., Lee, Phil H., Holmes, Avram J., Roffman, Joshua L., Buckner, Randy L., Sabuncu , Mert R. and Smoller, Jordan W.. (2015) *Massively expedited genome-wide heritability analysis (MEGHA).* Proceedings of the National Academy of Sciences of the United States of America, Volume 112 (Number 8). pp. 2479-2484.
6. D. Lin, H. Cao, V. D. Calhound,e, Yu-Ping Wang, Y. P. Sparse models for correlative and integrative analysis of imaging and genetic data. Journal of Neuroscience Methods 237 (2014) 69–78.
7. H. Cao, J. Duan, D. Lin, Y. Y. Shugart, V. Calhoun, Y.-P. Wang. (2014). Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and SNPs. 102, 220–228.
8. Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N., and Nichols, T. E. (2013) Genetics of the connectome. NeuroImage.
9. Medlan, S. E., Jahanshad, N., Neale, B. M., and Thompson, P. M. (2014). Whole-geome analyses of whole-brain data: working within an expanded search space. Nature Neuroscience.
10. Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, Green RC, Hu X, Huentelman MJ, Kim S, Kauwe JSK, Li Q, Liu E, Macciardi F, Moore JH, Munsie L, Nho K, Ramanan VK, Risacher SL, Stone DJ, Swaminathan S, Toga AW, Weiner MW, Saykin AJ, for the Alzheimer's Disease Neuroimaging Initiative (2014) Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers. *Brain Imaging and Behavior*, 8:183-207.
11. Vounou, M., Nichols, T. E., Montana, G., and ADNI (2010) Discovering genetic associations with high- dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage, 53, 1147–1159.
12. Vounou, M., Janousova, E., Wolz, R., Stein, J., Thompson, P., Rueckert, D., Montana, G., and ADNI (2012) Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. Neuroimage, 60, 700–716.
13. Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012) Increasing power for voxel- wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. NeuroImage, 63, 858–873.
14. Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S., Saykin, A., Shen, L., and ADNI (2012) Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: An imaging genetics study of the adni cohort. Bioinformatics,, 28, 229–237.
15. Silver M., Janousova E., Hue X., Thompson P. and Montana G. (2012) Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage, 63(3), 1681-1694.

References

Zhao, Y.Z.,, **Zhu, H.T.,** Lu, Z., Knickmeyer, R. C. and Zou, F. Bayesian Hierarchical Variable Selection for Genome-wide Association Studies. In submission, 2015.

Zhu, H.T., Khondker, Z. S., Lu, Z.H., and Ibrahim, J. G. Bayesian Longitudinal Low Rank Regression Models for Longitudinal Imaging Genetic Data. *NeuroImage,* in revision, 2015.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., and ***Zhu, H. T.*** (2015). FVGWAS: Fast Voxelwise Genome Wide Association Analysis of Large-scale Imaging Genetic Data, *NeuroImage*, In revision.

Sun, Q., ***Zhu, H.T.,*** Liu, Y. F., and Ibrahim, J.G. SPReM: Sparse Projection Regression Model for High-dimensional Linear Regression. *Journal of American Statistical Association,* in press, 2014.

S. J. Lee, R.J. Steiner; Shikai Luo; Michael C Neale; Martin Styner; **Hongtu Zhu**; John H. Gilmore. (2015). Quantitative tract-based white matter heritability in twin neonates. *NeuroImage,* 111:123-135.

Lin, J., ***Zhu, H.T.,*** , Mihye, A., and Ibrahim, J.G. Functional Mixed Effects Models for Candidate Genetic Mapping in Imaging Genetic Studies. *Genetic Epidemiology,* 38(8):680-91, 2014.

Zhu, H.T., Khondker, Z. S., Lu, Z.H., and Ibrahim, J. G. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of American Statistical Association,* 109, 1084-1098, 2014.

Wei Gao, Amanda Elton, **H. Zhu,** Sarael Alcauter, J. Smith, John H Gilmore, and Weili Lin. (2014). Inter-subject Variability of and Genetic Effects on the Brain's Functional Connectivity during Infancy. *Journal of Neuroscience*, 34: 11288-11296.

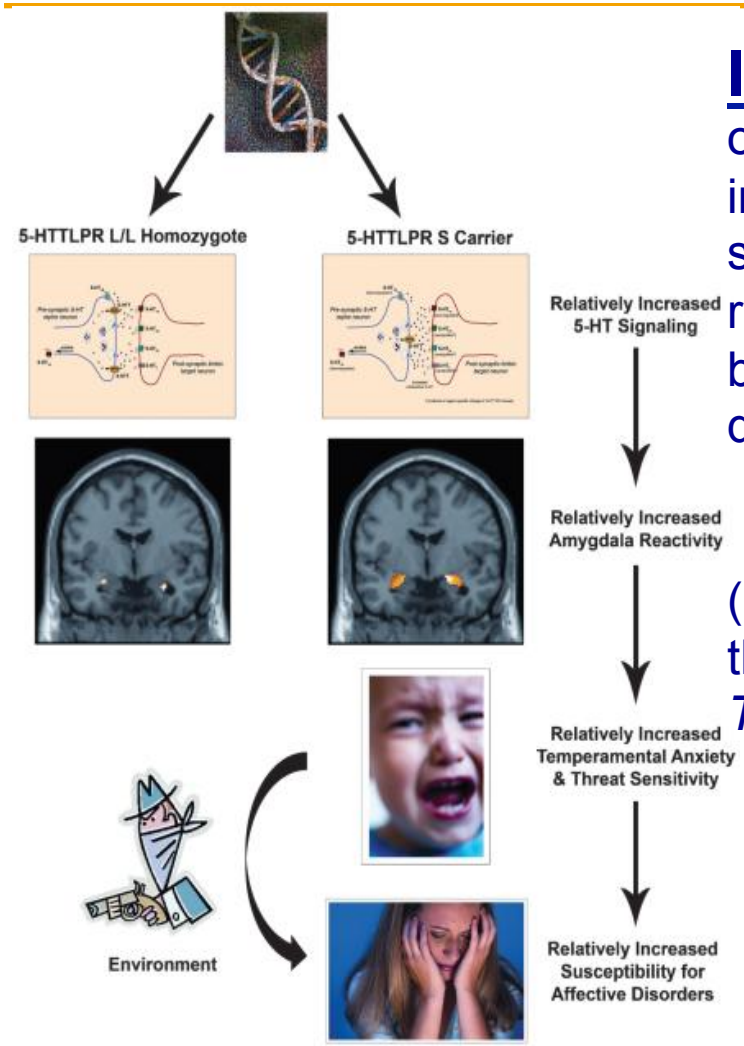
Kai Xia, Yang Yu, Mihye Ahn, **H. Zhu**, Fei Zou, John Gilmore, Rebecca Christine Knickmeyer. Environmental and genetic contributors to salivary testosterone levels in infants. *Frontiers in Endocrinology.* 2014.

Bryant, C., Giovanello, K. S., Ibrahim, J. G., Shen, D. G., Peterson, B. S., and **Zhu, H.T.** (2013) Mapping the heritability of regional brain volumes explained by all common SNPs from the ADNI study. *PLOS ONE.*

Lin, J., ***Zhu, H.T.,*** Knickmeyer, R., Styner, M., Gilmore, J. and Ibrahim, J.G. Projection Regression Models for Multivariate Imaging Phenotype. *Genetic Epidemiology,* 36, 631-641, 2012.

Motivation

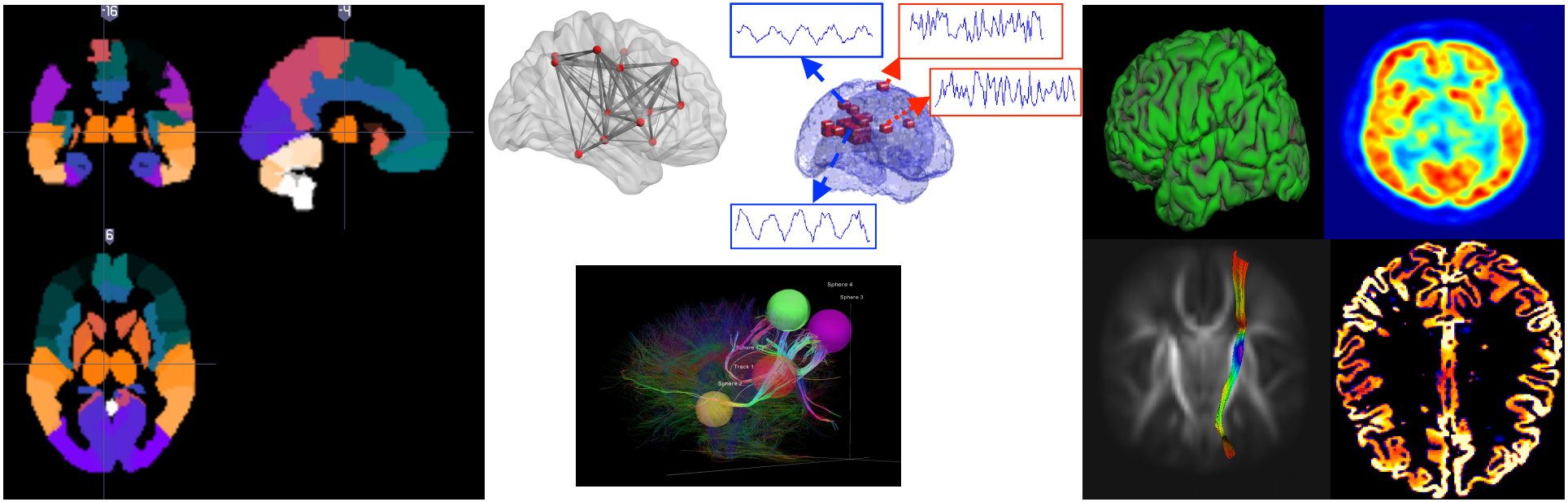
Motivation



Imaging genetics allows for the identification of how common/rare genetic polymorphisms influencing molecular processes (e.g., serotonin signaling), bias neural pathways (e.g., amygdala reactivity), mediating individual differences in complex behavioral processes (e.g., trait anxiety) related to disease risk in response to environmental adversity.

(Hariri AR, Holmes A. Genetics of emotional regulation: the role of the serotonin transporter in neural function. *Trends Cogn Sci*. [10:182–191])

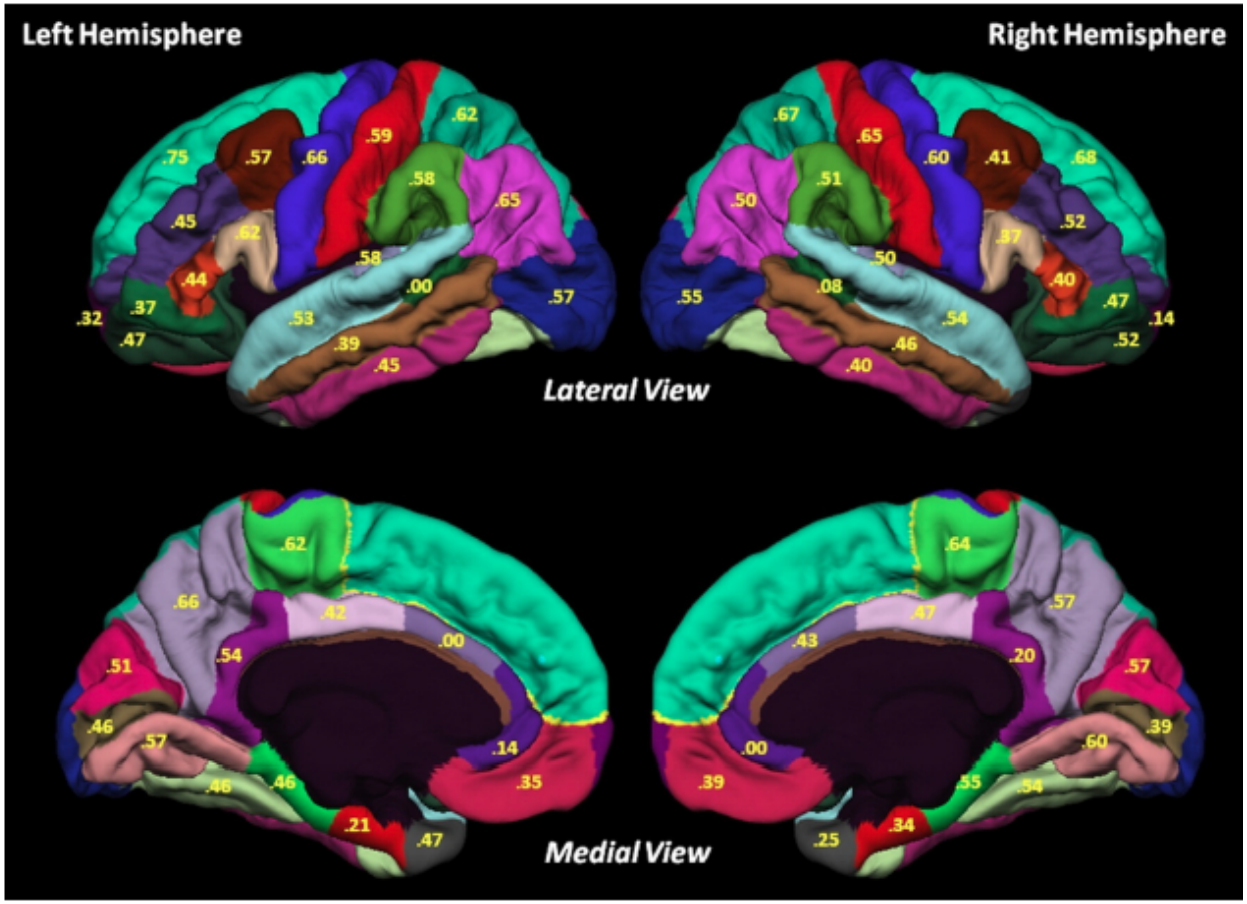
Neuroimaging Phenotype



Multivariate, smoothed functions, and piecewisely smoothed functions

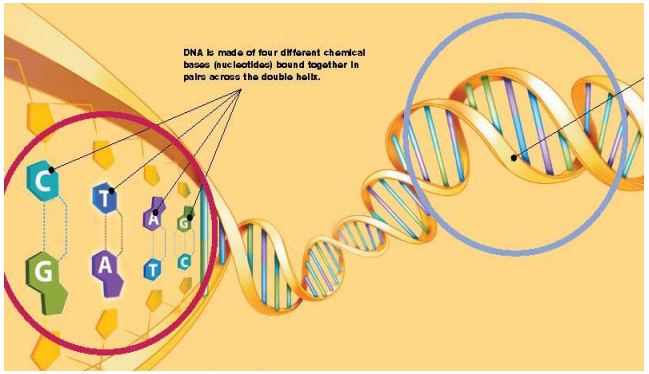
Dimension varies from 100~500,000.

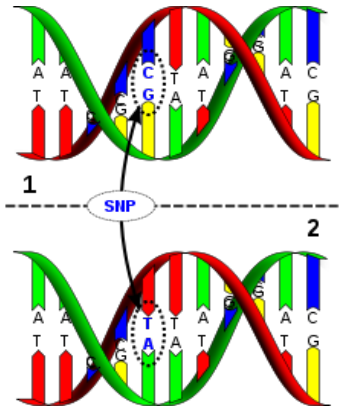
Brain Structure is highly heritable

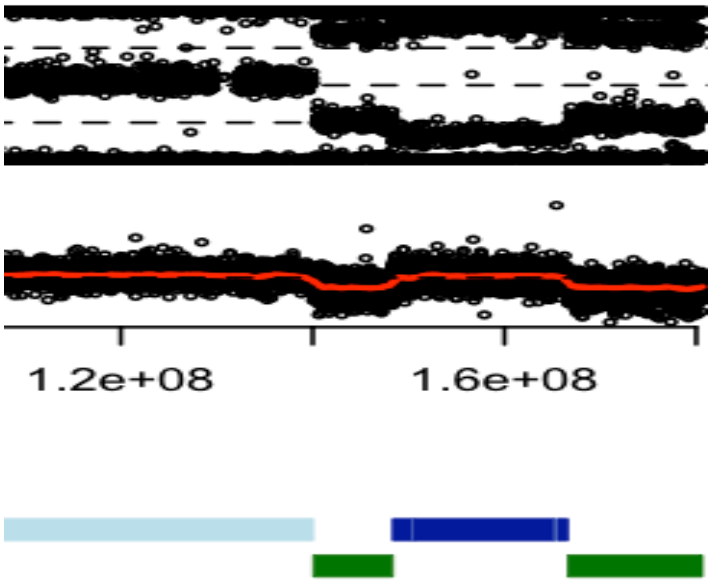


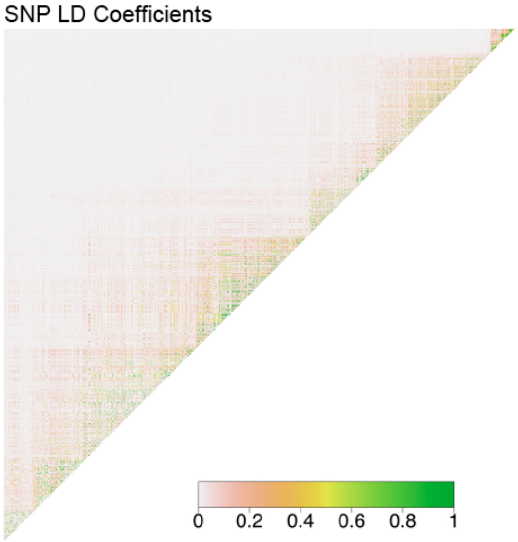
Must be specific genetic variants explaining the high heritability (Kremen et al.2010).

Genetic Data

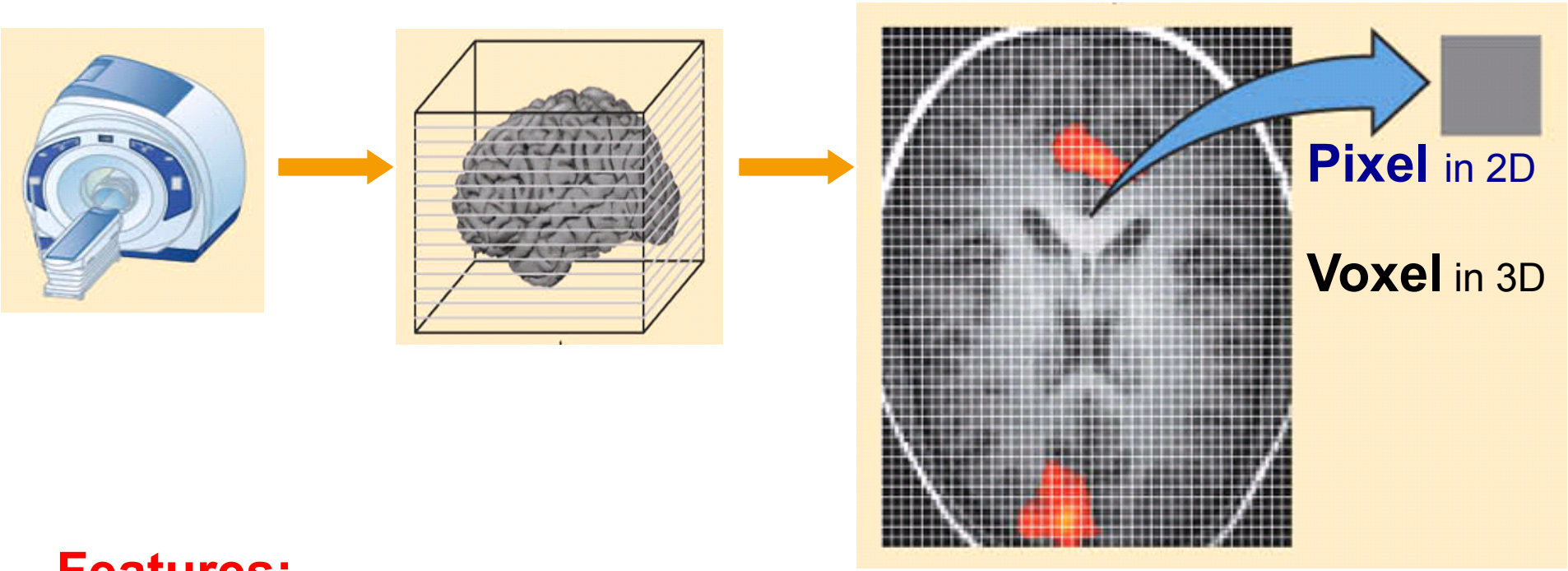








Imaging Phenotype



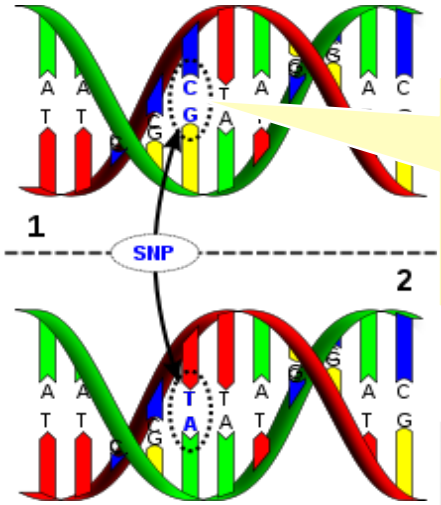
Features:

-- Spatially correlated functional data

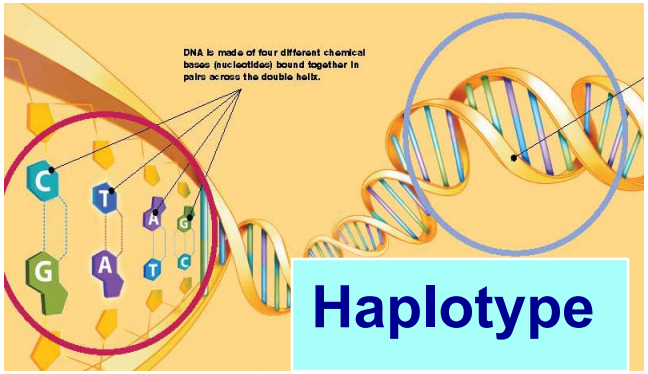
-- High dimension (~ 2^{21} = 2,097,152 voxels)

-- Voxels with similar color form regions

Genetic Variation Data

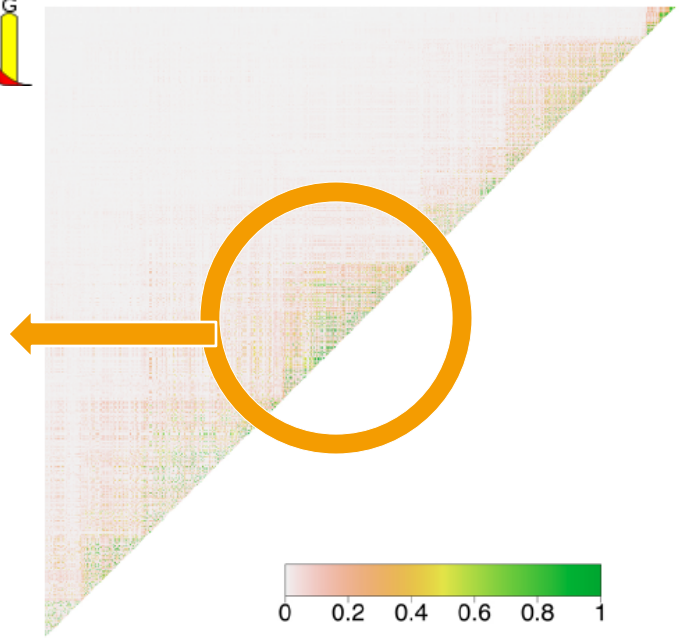


SNP
(Single Necleotide Polymorphism)



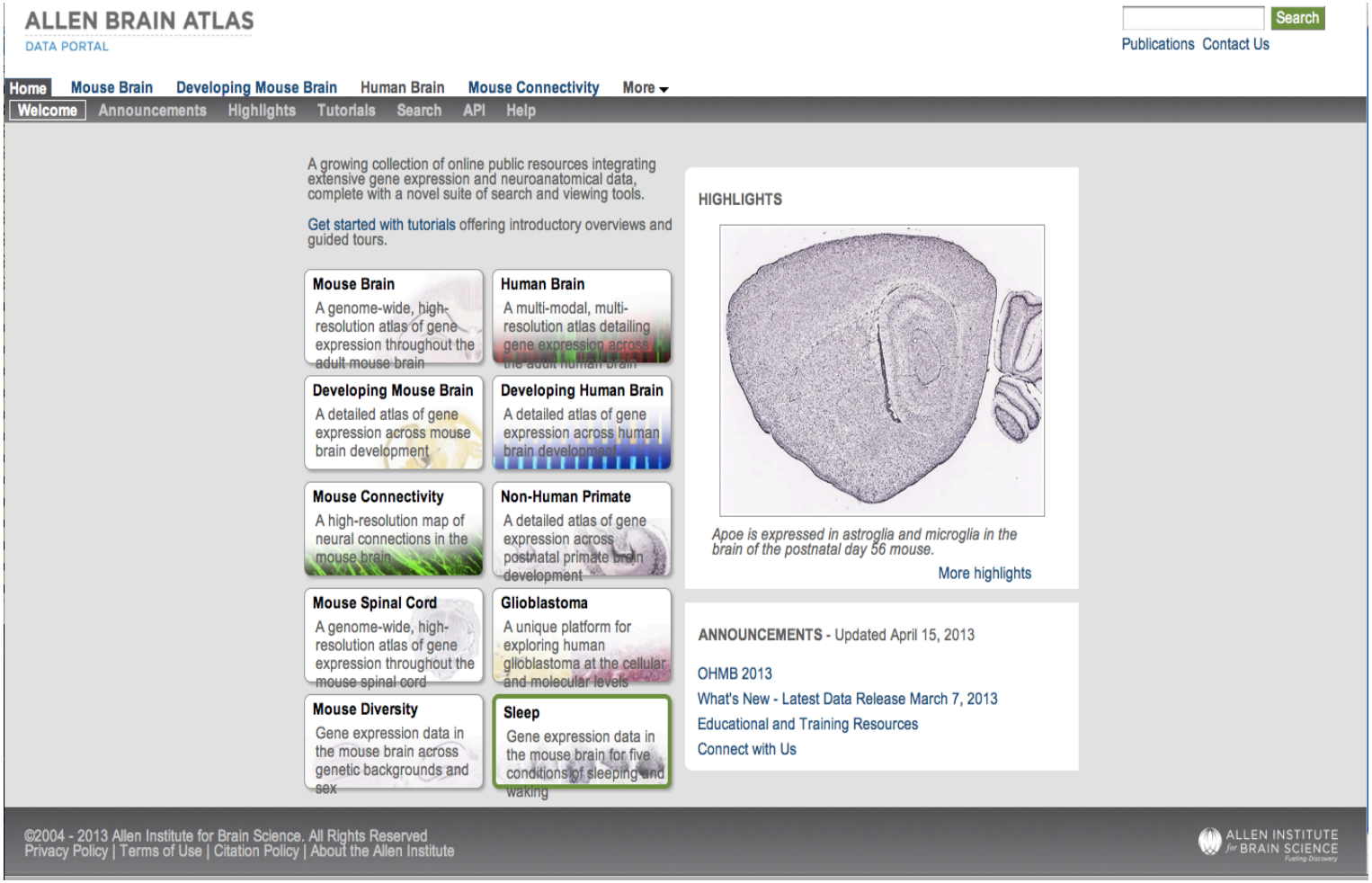
Haplotype

SNP LD Coefficients



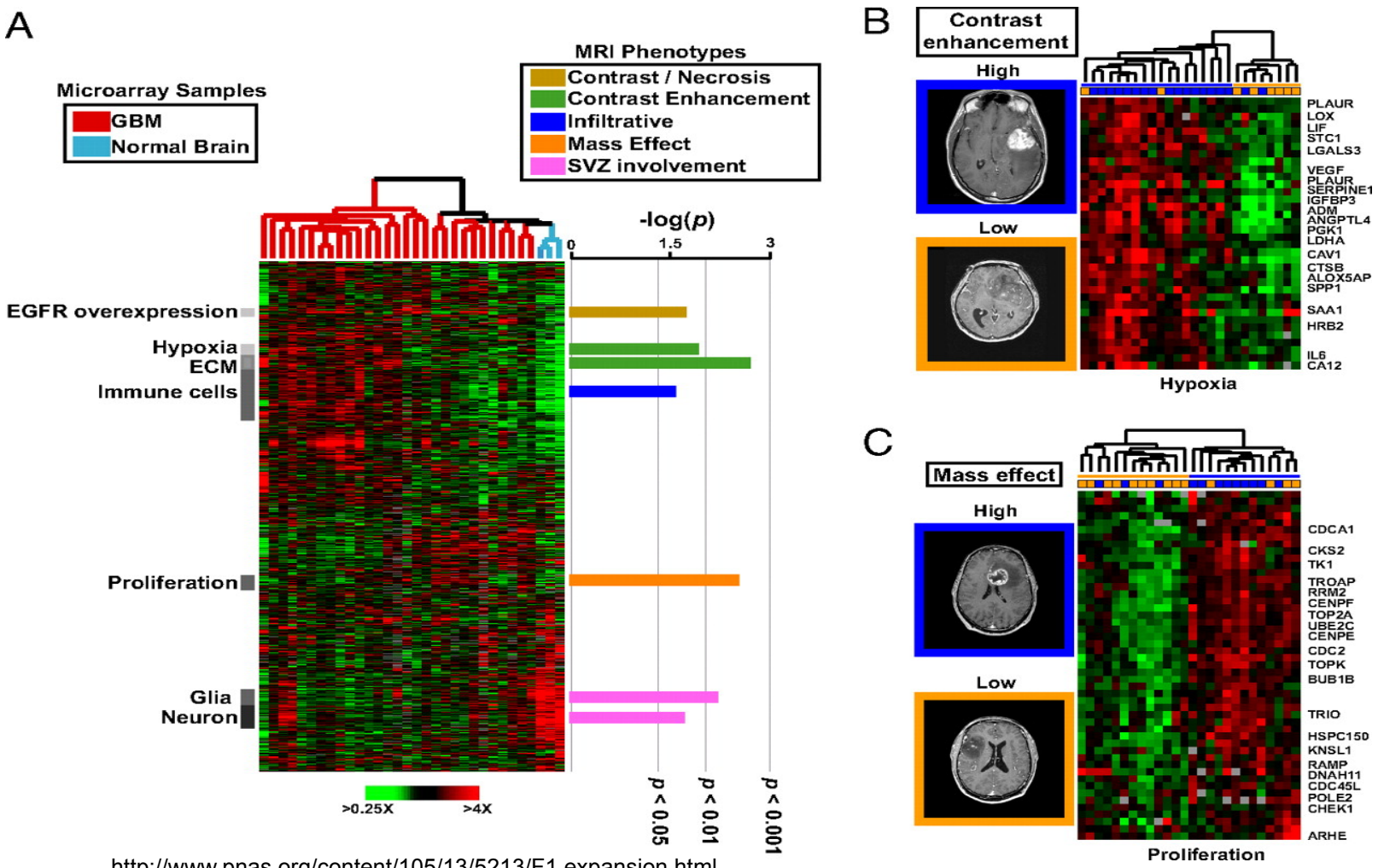
- **Correlated**
- **High Dimension**

Imaging Genetic Studies -- Brain Development



http://www.brain-map.org/

Imaging Genetic Studies -- Brain Tumor



http://www.pnas.org/content/105/13/5213/F1.expansion.html

Imaging Genetic Studies -- More Studies

ADNI
ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

NDAR National Database for Autism Research
Serving the autism research community

09/28/2012 - First Publication Using NDAR Data Now Available
09/28/2012 - NDAR Announces the Autism Genome Project (AGP) Data Release
05/24/2012 - NDAR Publication on Sharing Heterogeneous Data

Home | Query | Harmonization Tools | Contribute | Request Access | Policy | Tutorials | About | FAQ

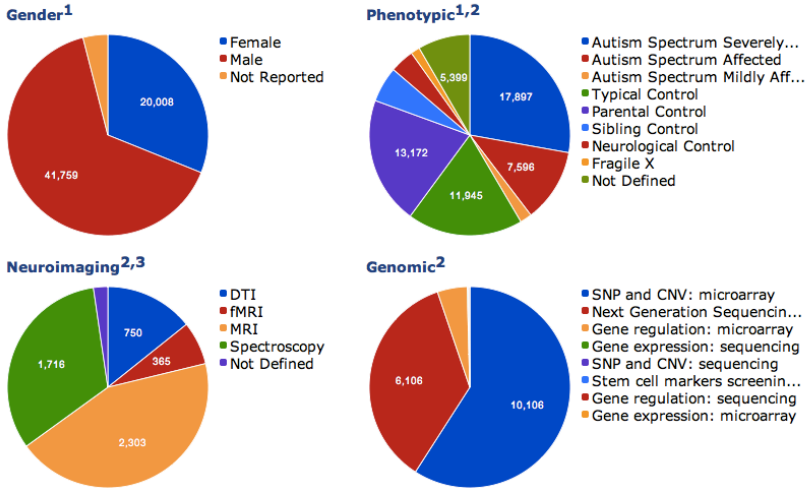
Quick Navigation

- **NDAR Data**
  - **Query**
  - **Data from Labs**
  - **Data from Papers**
- **Submission**
  - **Harmonization Standards**
  - **Prepare**
  - **Validation Tool**
  - **Submit**
- **Data Dictionary**
- **Contact Us**

Learn About NDAR

The National Database for Autism Research (NDAR) is an NIH-funded research data repository that aims to accelerate progress in autism spectrum disorders (ASD) research through data sharing, data harmonization, and the reporting of research results. NDAR also serves as a scientific community platform and portal to multiple other research repositories, allowing for aggregation and secondary analysis of data.

Data Distribution
64,335 subjects by age, 48,135 individuals



Genetic variation through collaboration

- Common genetic variation has only a **small effect** on brain structure.
- In a GWAS, many statistical tests are conducted.
- **Huge sample sizes** (>10,000 individuals) are needed in order to find and replicate association of individual genetic variants.
- As both imaging and genotyping are expensive, **a consortium is needed.**

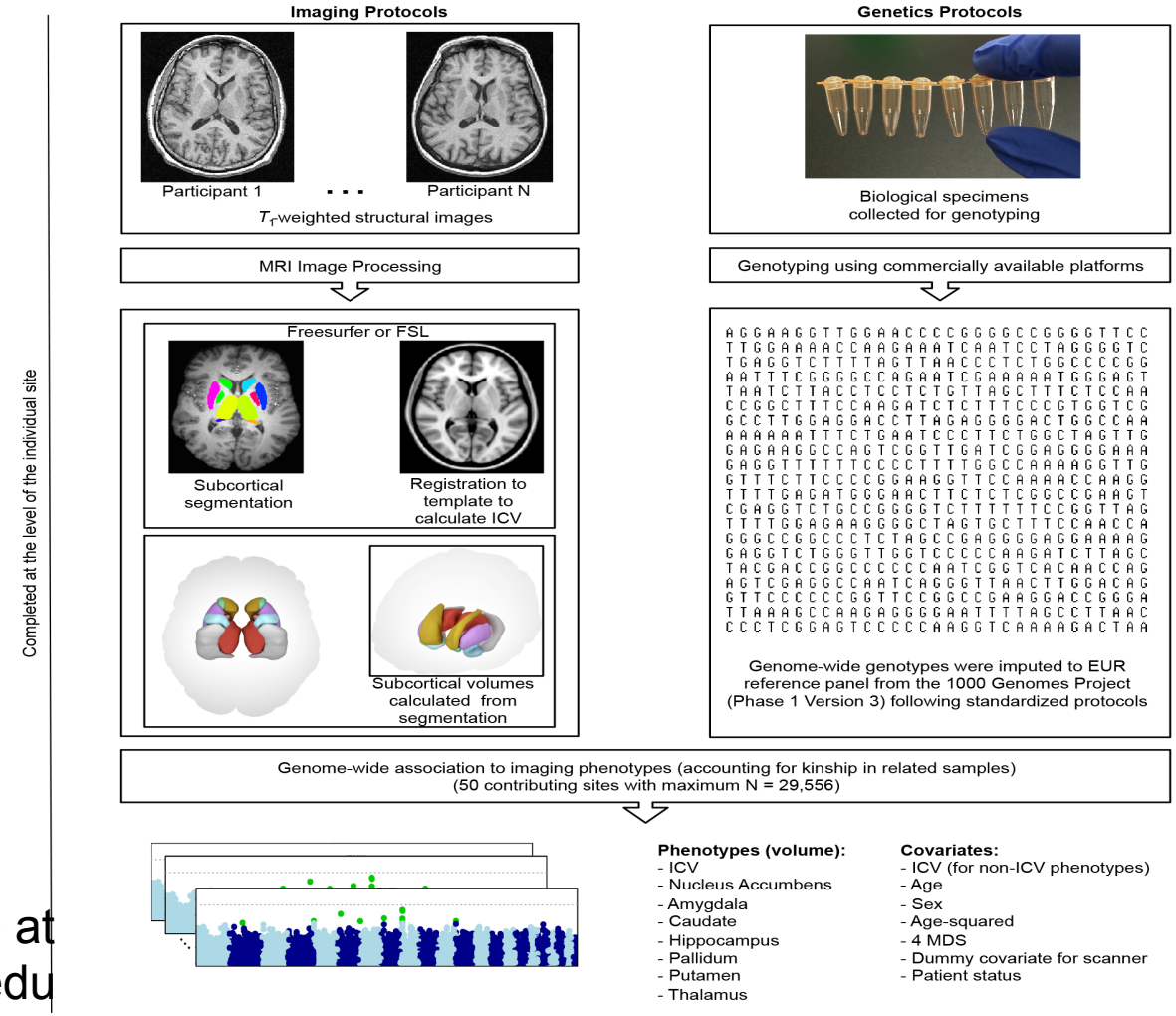
Dr. Hibar

http://enigma.ini.usc.edu

> 185 institutions, 300+ co-authors, world-wide consortium
countries resulting in **~30,000+** subjects

Now expanded to multiple multi-national meta-analytic projects:
Genetics of brain structure
Genetics of brain function
Meta-analytic effects on disease

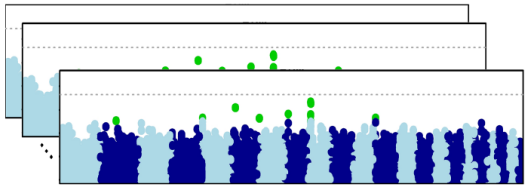
Procedure for meta-analysis (Part 1)



All protocols available at
http://enigma.ini.usc.edu

Procedure for meta-analysis (Part 2)

Genome-wide association to imaging phenotypes (accounting for kinship in related samples)
(50 contributing sites with maximum N = 29,556)



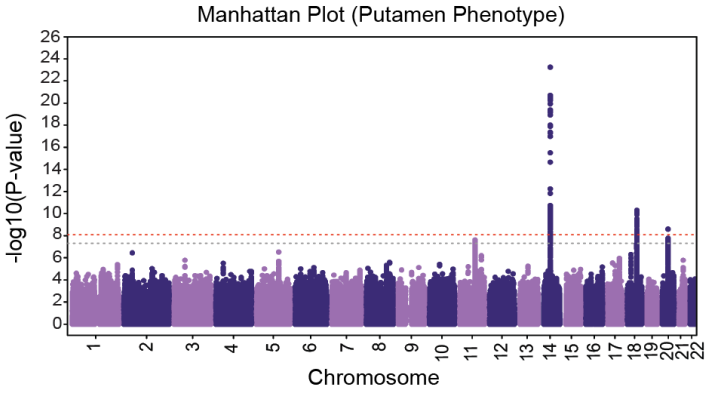
Phenotypes (volume):
- ICV
- Nucleus Accumbens
- Amygdala
- Caudate
- Hippocampus
- Pallidum
- Putamen
- Thalamus

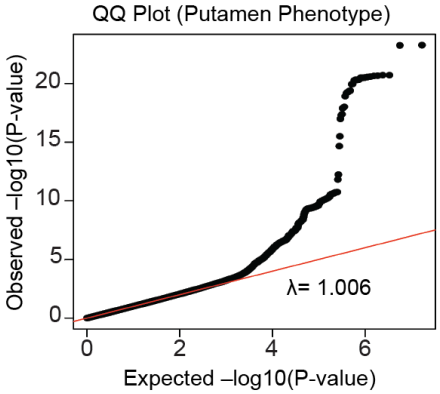
Covariates:
- ICV (for non-ICV phenotypes)
- Age
- Sex
- Age-squared
- 4 MDS
- Dummy covariate for scanner
- Patient status

Uploaded to ENIGMA server for analysis at central site

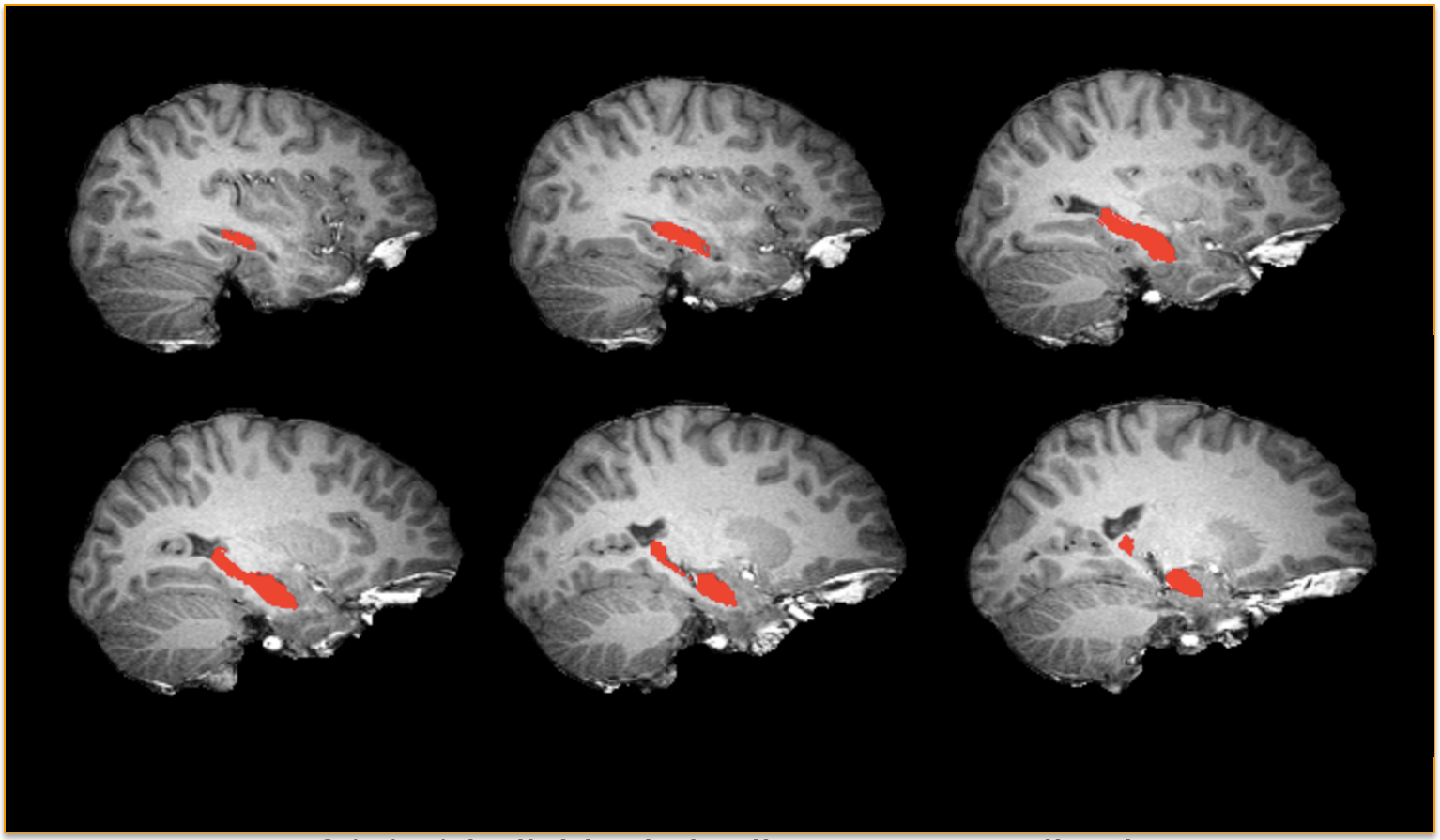
Quality Checking and Filtering (MAC < 10, R^2 < 0.5)

Meta-analyses
Fixed-effect, inverse variance-weighted model





ENIGMA1: Pilot Project Hippocampal and Intracranial Volume GWAS Meta-analysis



21,151 individuals in discovery + replication

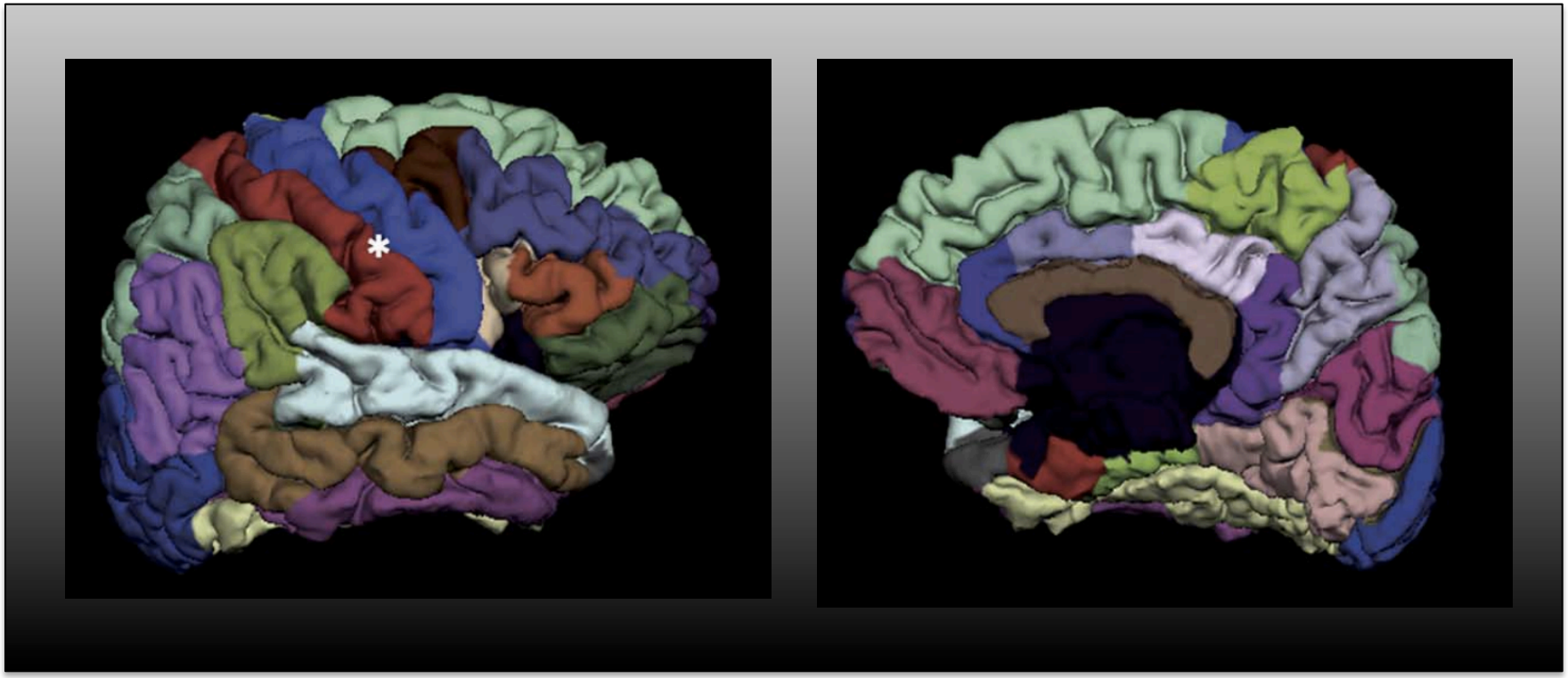
(Stein et al., *Nature Genetics*, 2012)

ENIGMA2: Genetics of subcortical structures

Discovery + Replication = 30,717 individuals
50 contributing cohorts around the world

(Hibar et al., 2015. *Nature*)

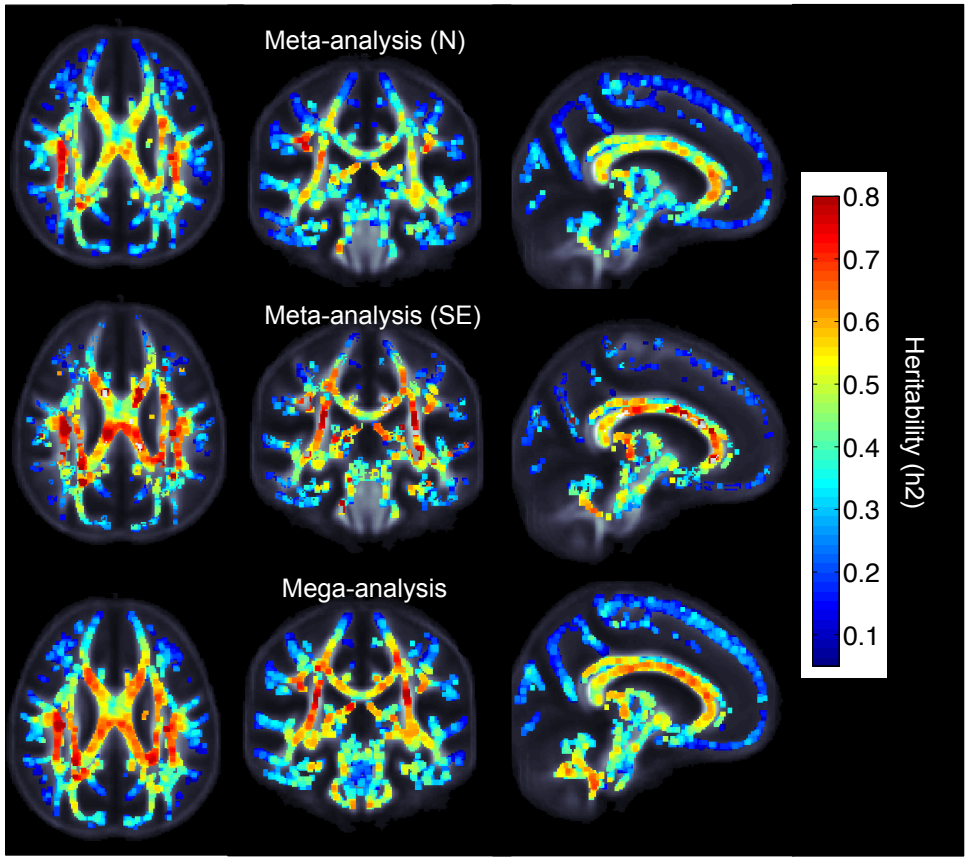
What's next for ENIGMA?

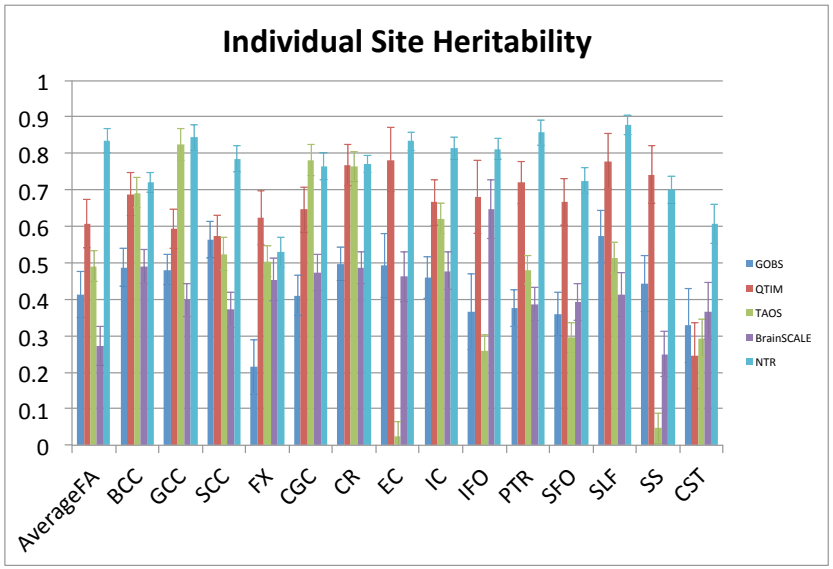


ENIGMA3: GWAS of cortical area and thickness
Projected N=30,000

Interested in joining? Sign up at http://enigma.ini.usc.edu/

ENIGMA-DTI

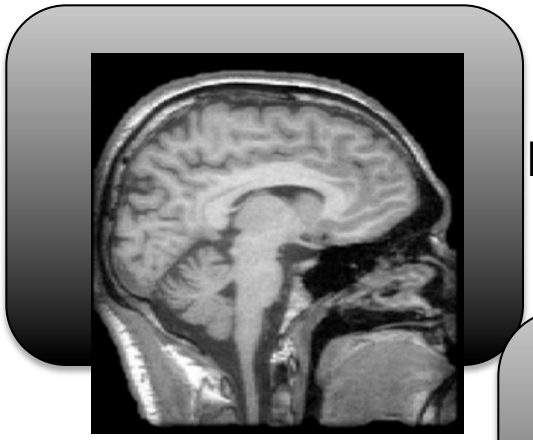




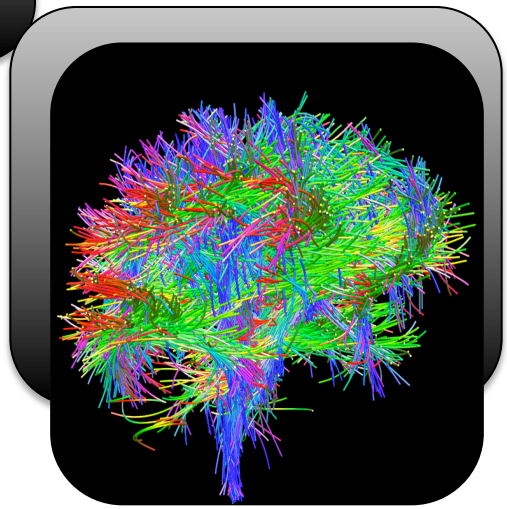
http://enigma.ini.usc.edu/ongoing/dti-working-group/

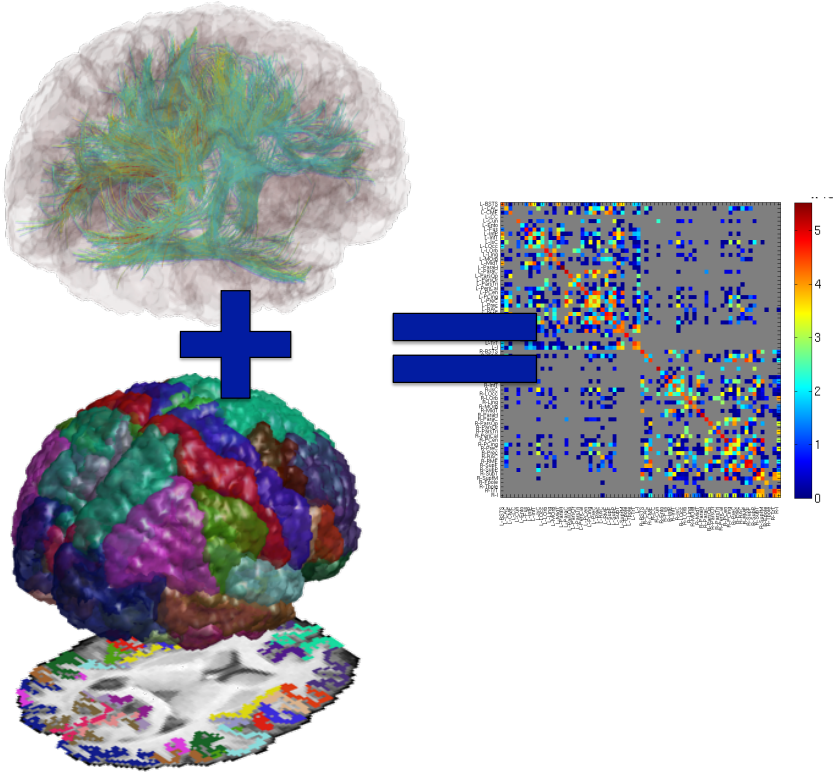
Mapping Connectivity

- Structural magnetic resonance imaging (MRI)
  - Anatomical analysis
  - Mapping Tissue Loss

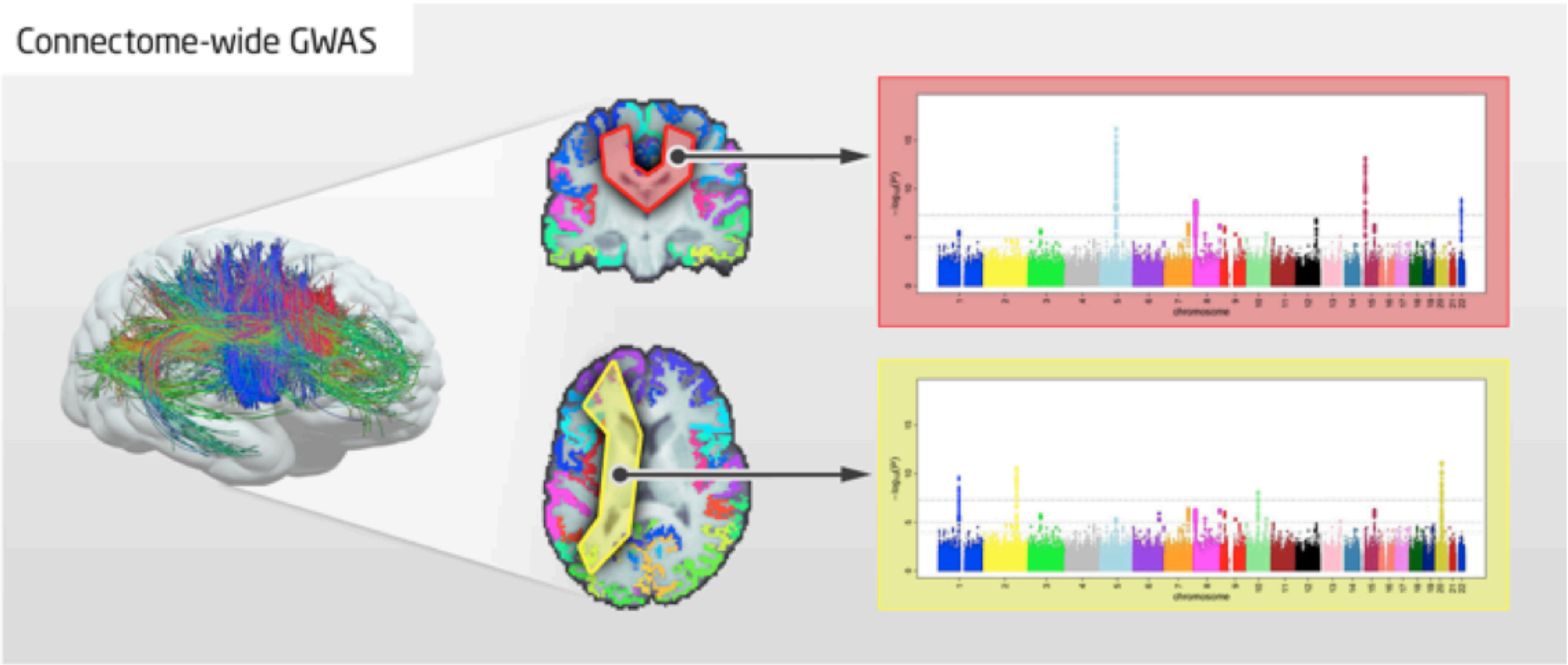


Diffusion weighted
- Microstructure
- Connectivity

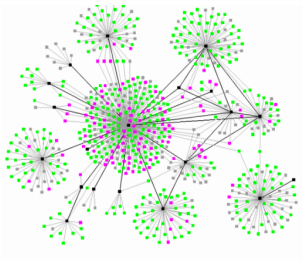




Connectome-Wide Genome-Wide Screen Alzheimer risk gene



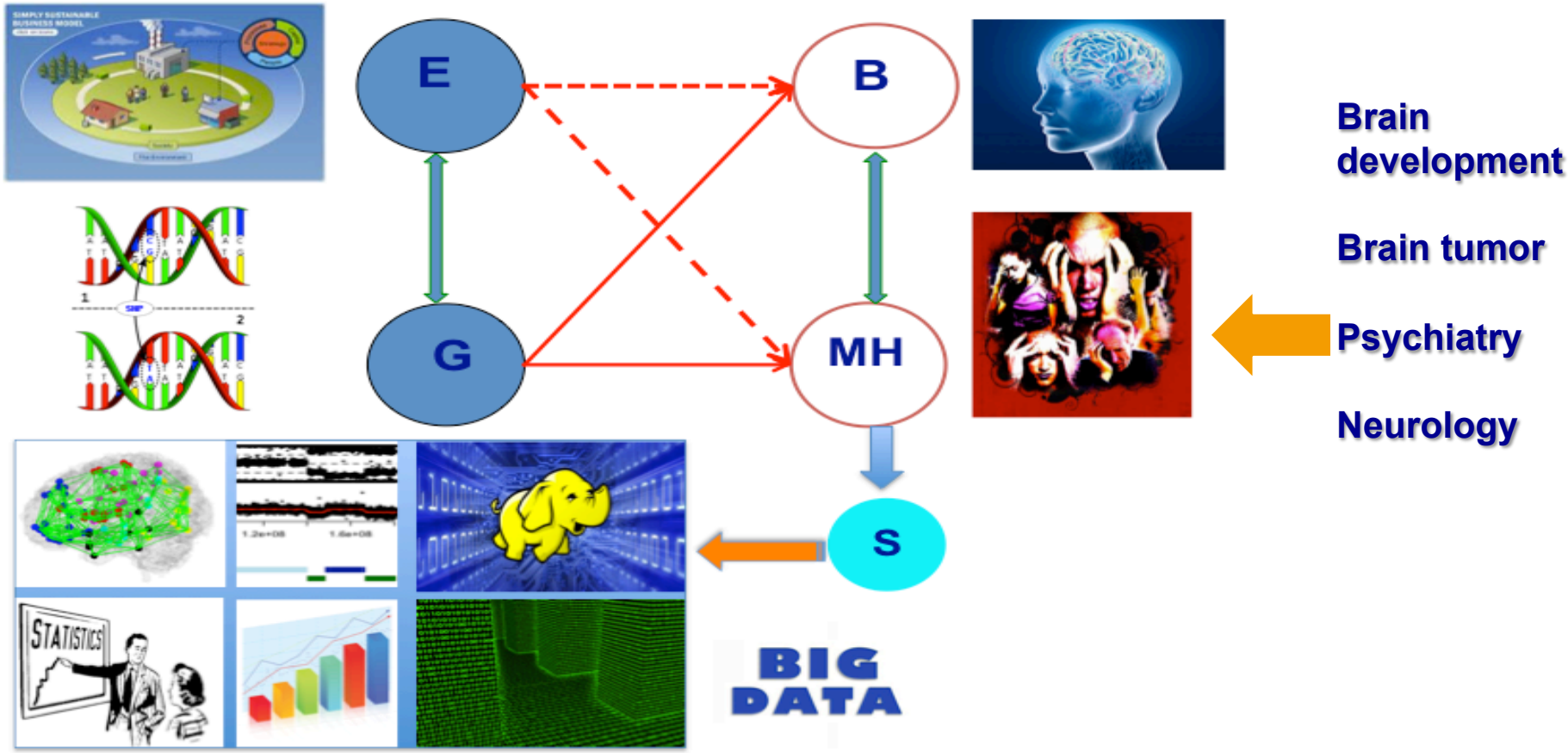
Discovery sample – Young Adults Effect in ADNI
Within 2 weeks Sherva et al. published *SPON1*
Found in a cognitive GWAS in AD



Jahanshad et al., PNAS 2013

Overview

Imaging Genetic Studies



Common Goal: Detect potential genes for inherited phenotypes

Statistical Methods

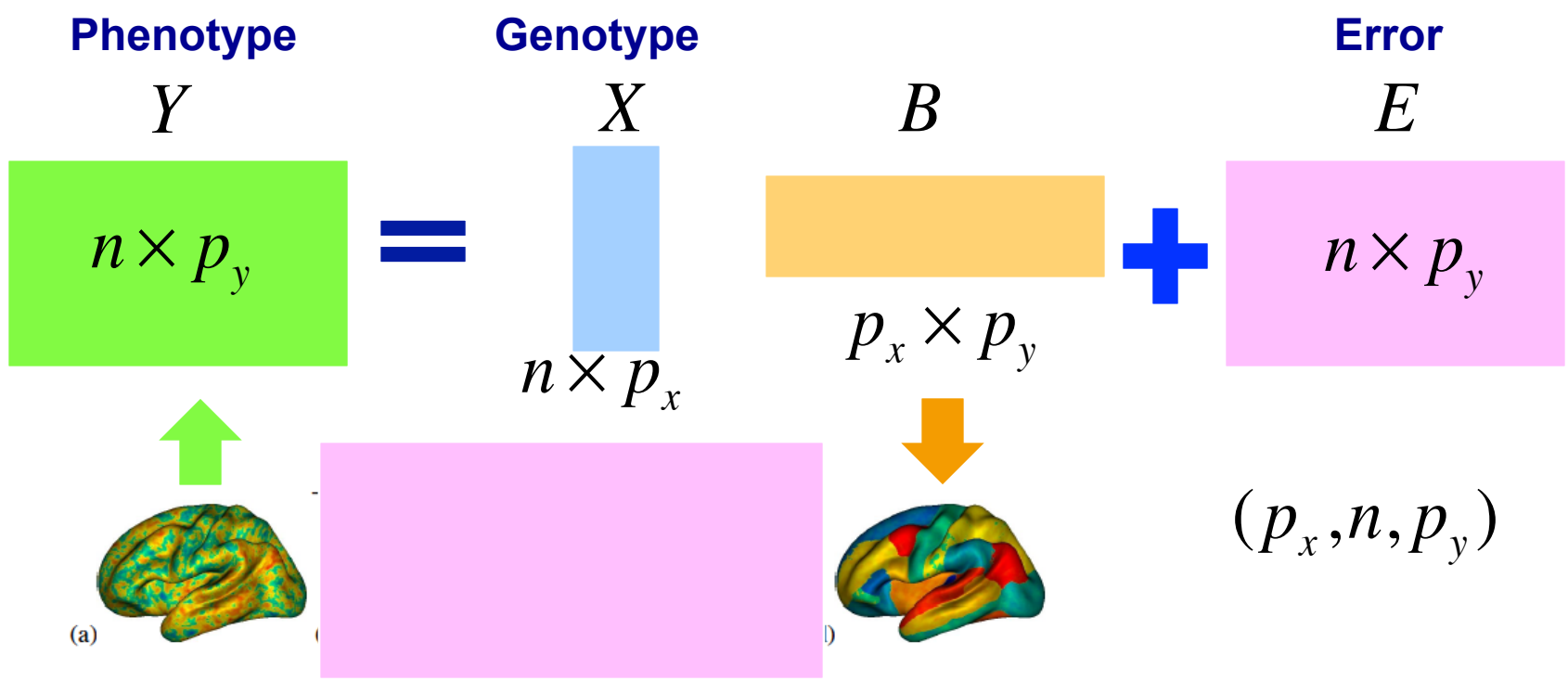
Genetics \ Imaging	Candidate ROI	Many ROI	Voxelwise
Candidate SNP	Imager	Imager	Imager
Candidate Gene	Geneticist		
Genome-wide SNP	Geneticist		
Genome-wide Gene	Geneticist		

Hibar, et al. HBM 2012

High Dimensional Regression Model

Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V\} \qquad \{X_i(g) : g \in G_0\}$$



$$(p_x, n, p_y)$$

Current Approaches and Limitations

Candidate Gene
And / Or
Candidate Imaging Region

→ Only part of the picture

http://www.tnooz.com/

Genomic-wide Search
+
Voxel-based Approach

→ ~~correlation (LD) structure~~
~~spatial correlation in voxels~~
Low statistical power

Everything in a big model
+
Penalized Method

→ Unstable variable selection
+
~~Statistical inference~~

Models for Candidate ROI Candidate Gene-set Analysis

High Dimensional Regression Model

Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V_0\} \qquad X_i = \{X_i(g) : g \in G_0\}$$

Phenotype		Genotype				Error
Y		X		B		E
$n \times p_y$	$=$	$n \times p_x$		$p_x \times p_y$	$+$	$n \times p_y$

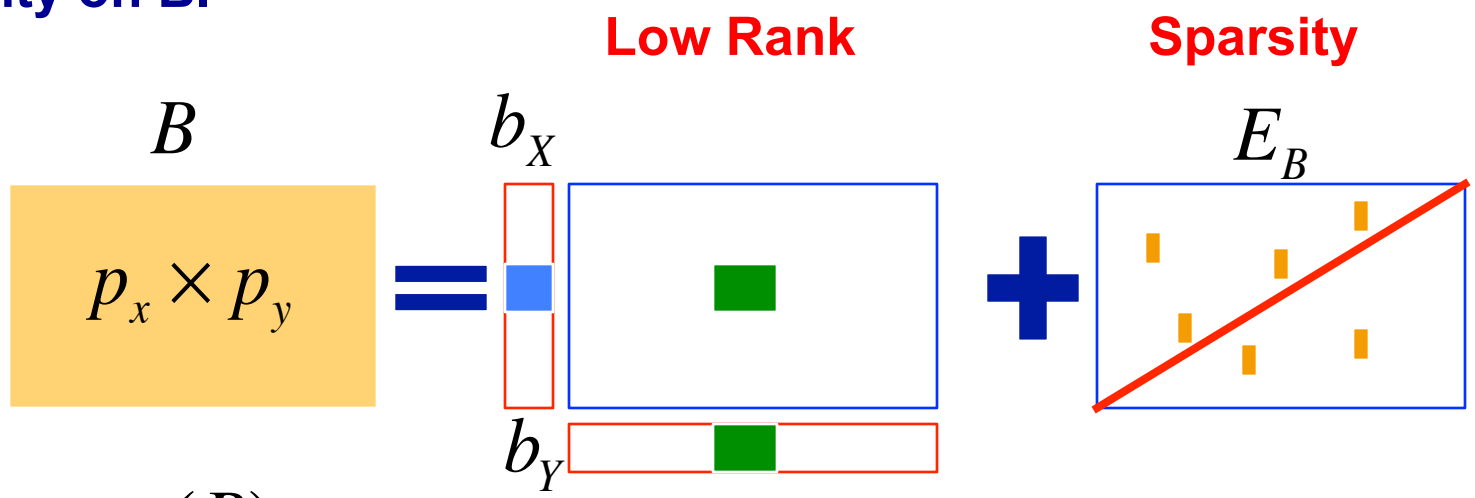
Key Conditions:

$$\max(p_x, p_y) \sim n$$

- **Sparsity of B**
- **Restricted null-space property for design matrix X**

Sparse and Low-rank Representation

Sparsity on B.



$$p_\lambda(B) \rightarrow p_\lambda(b_X) + p_\lambda(b_Y) + p_\lambda(E_B)$$

Regularization Methods

- **Lasso 1, 2, 3, ….**
- **SCAD, MCP, …..**

$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p}|\theta_j|$$

Genetic and Imaging Networks

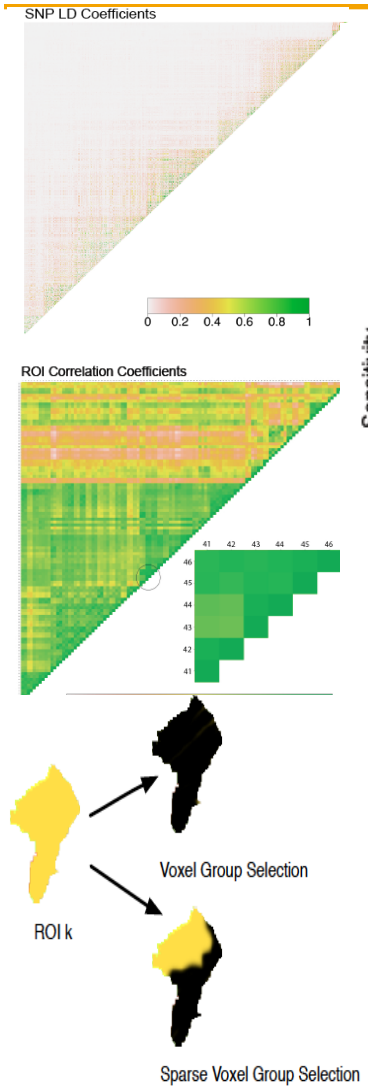
Genetic Network

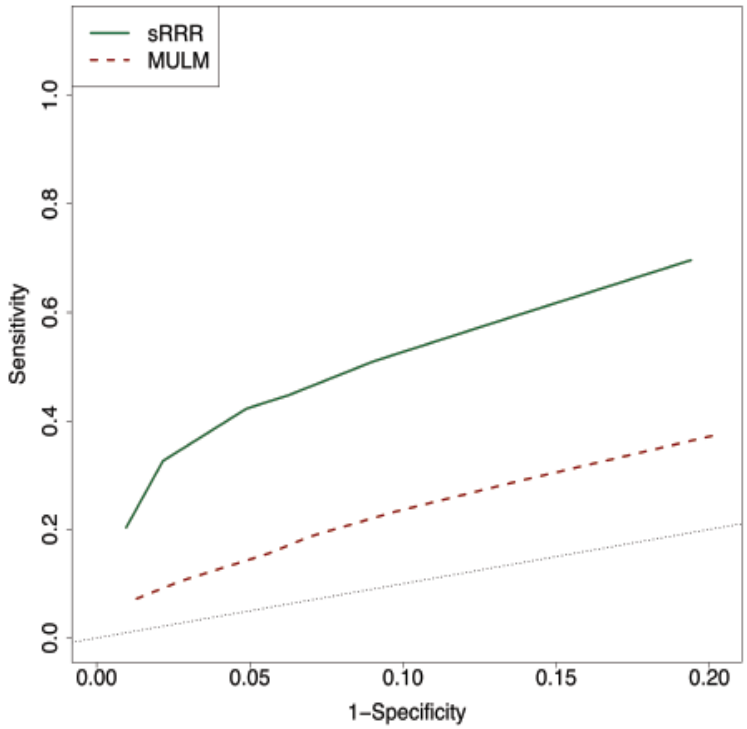
$$B \quad p_x \times p_y \rightarrow f(B) \quad f(B)^T = \underbrace{p_x \times p_x}_{f(B)f(B)^T}$$

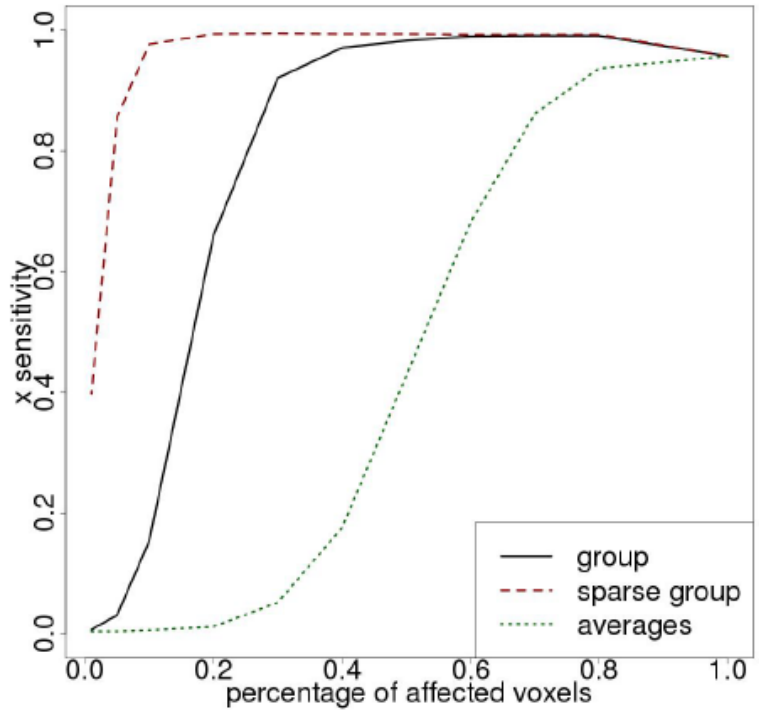
Imaging Network

$$B \quad p_x \times p_y \rightarrow f(B)^T \quad f(B) = \underbrace{p_y \times p_y}_{f(B)^T f(B)}$$

Simulation

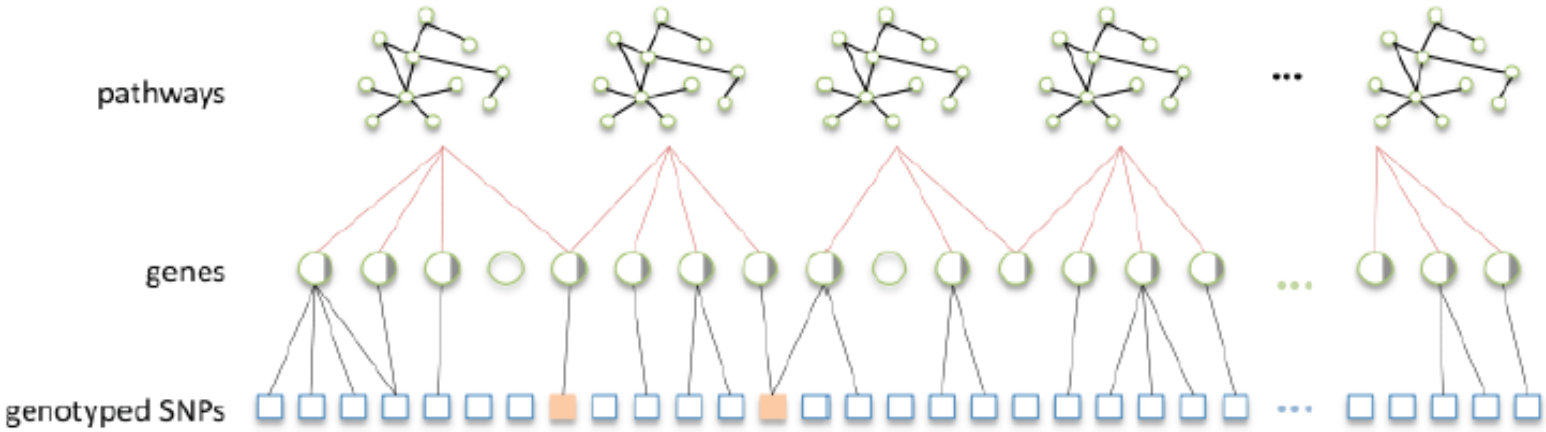






Vounou et al. (2010, 2012) NeuroImage

SNPs to Genes to Pathways



- Group all available SNPs into L pathways $\mathcal{G}_1, \ldots, \mathcal{G}_L$
- Adopt a *group lasso* penalty to force group selection,

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

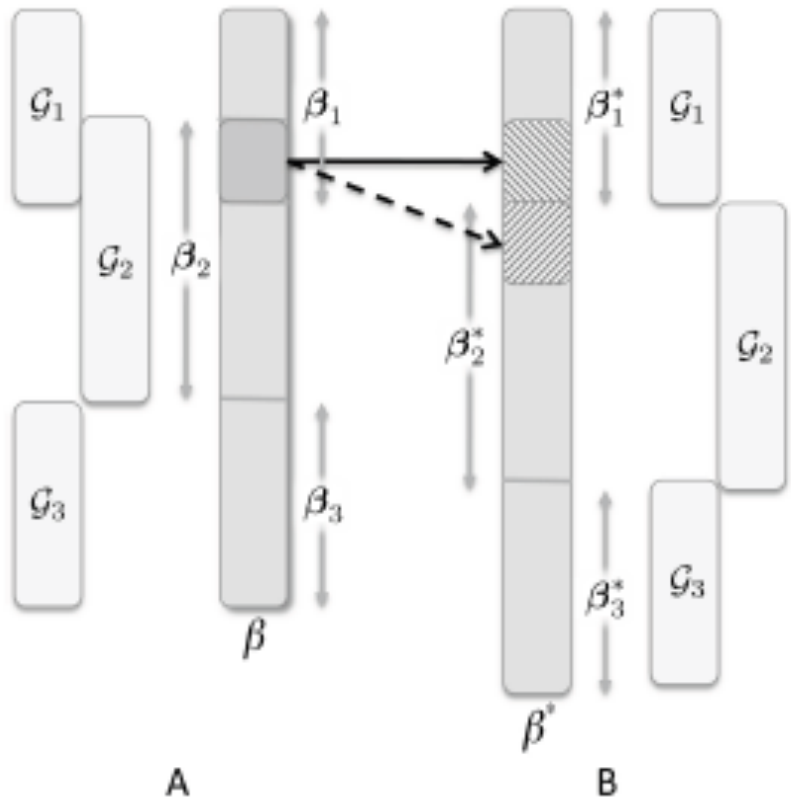
to select only most predictive pathways,

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

Overlapping Pathways

SNPs Duplication in Overlapping Pathways

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



ANDI

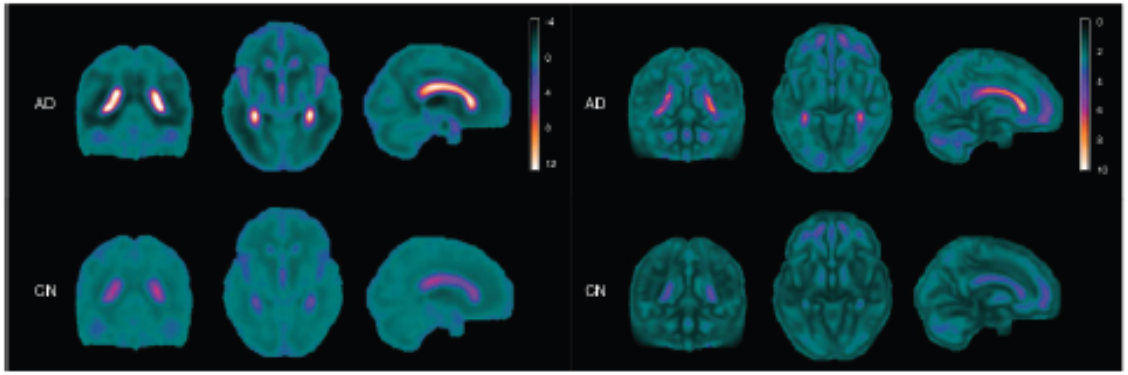
- Serial brain MRI scans were analyzed from 200 probable AD patients and 232 healthy elderly controls (CN)
- Longitudinal scans available at three time points

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

At screening:

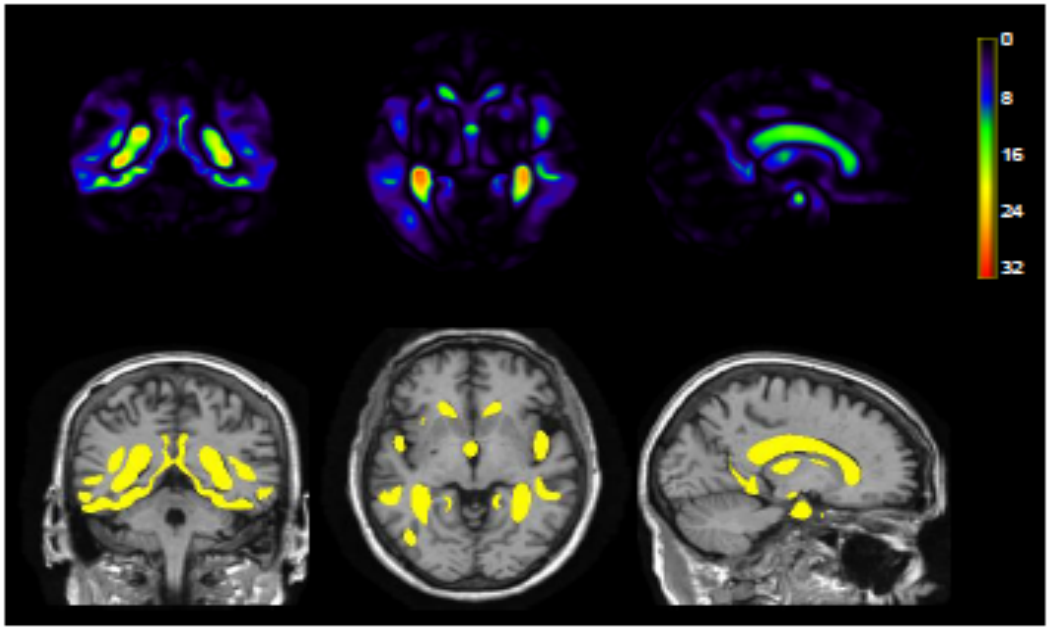
Group	age (years)	N male	N female
AD	75.7±7.7	103	97
CN	76.0±5.0	120	112

AD Imaging Signature

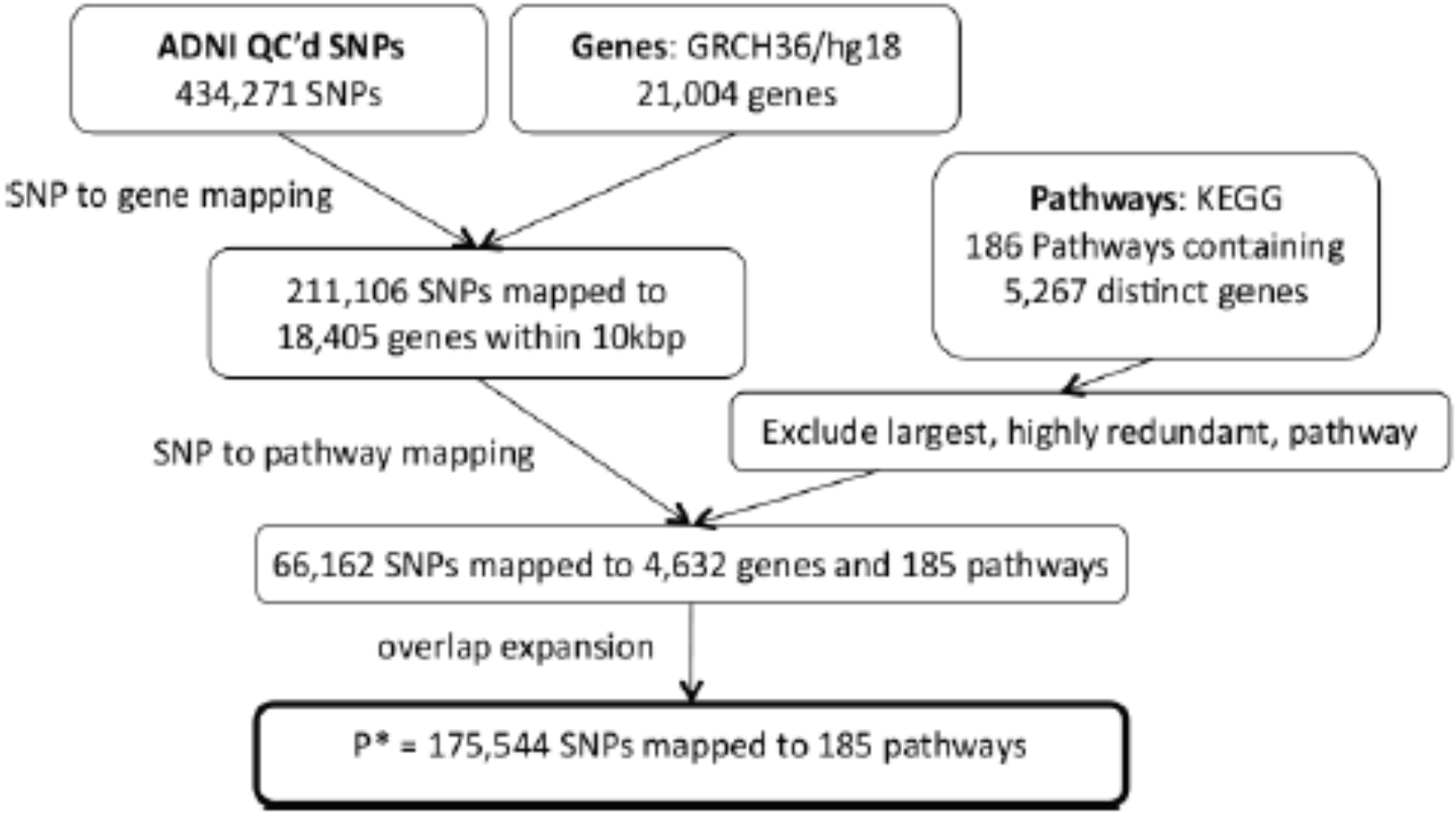


Mean and STD of slopes

-log10(p) value map



Mapping SNPs to Pathway



Top 15 SNPs and Genes Ranked by SRRR

	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

Factor Model

$$E \quad n \times p_y$$

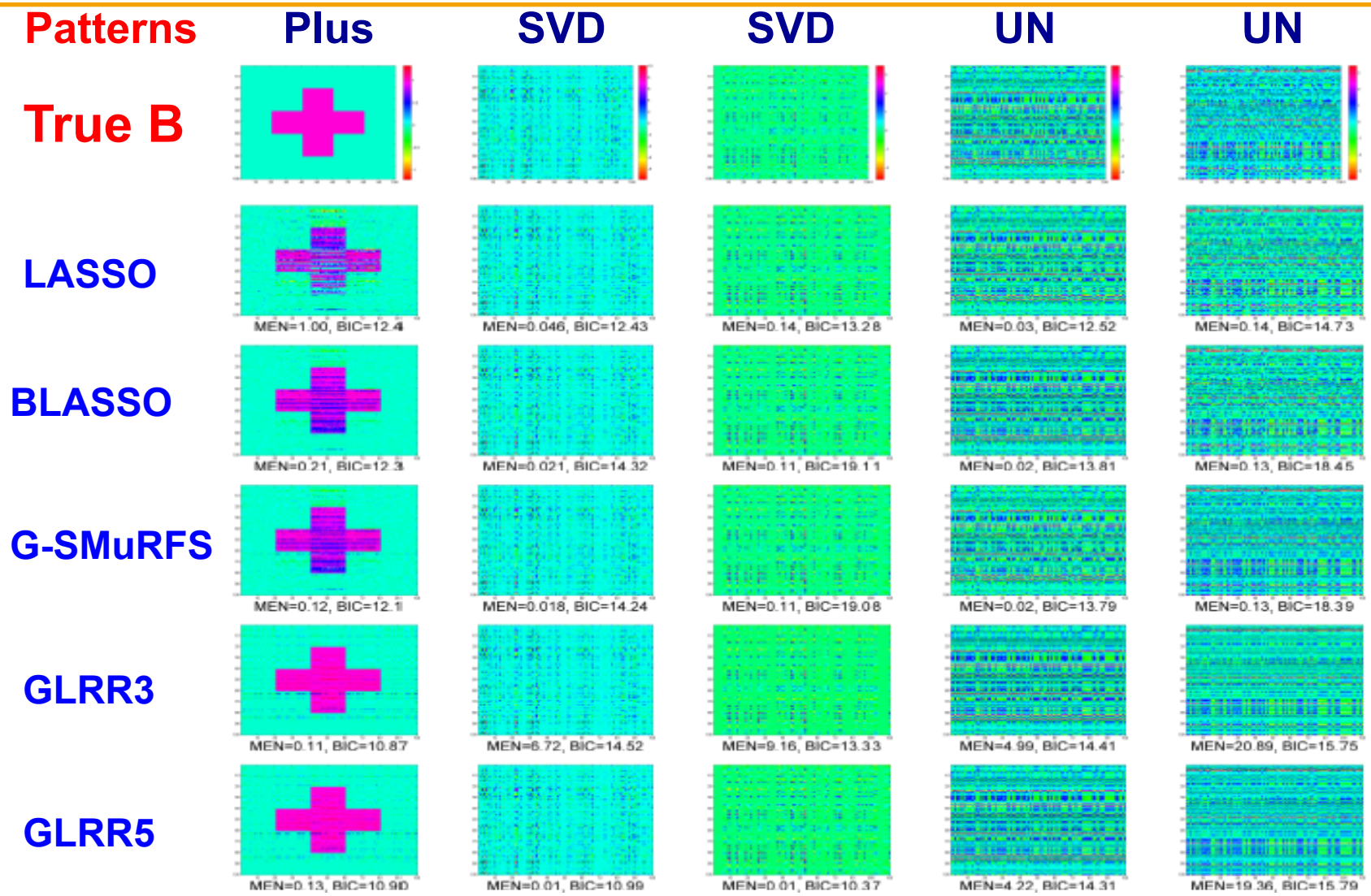
Long-range Correlation **Short-range Correlation**

$$E_i = \Lambda \xi_i + \eta_i$$

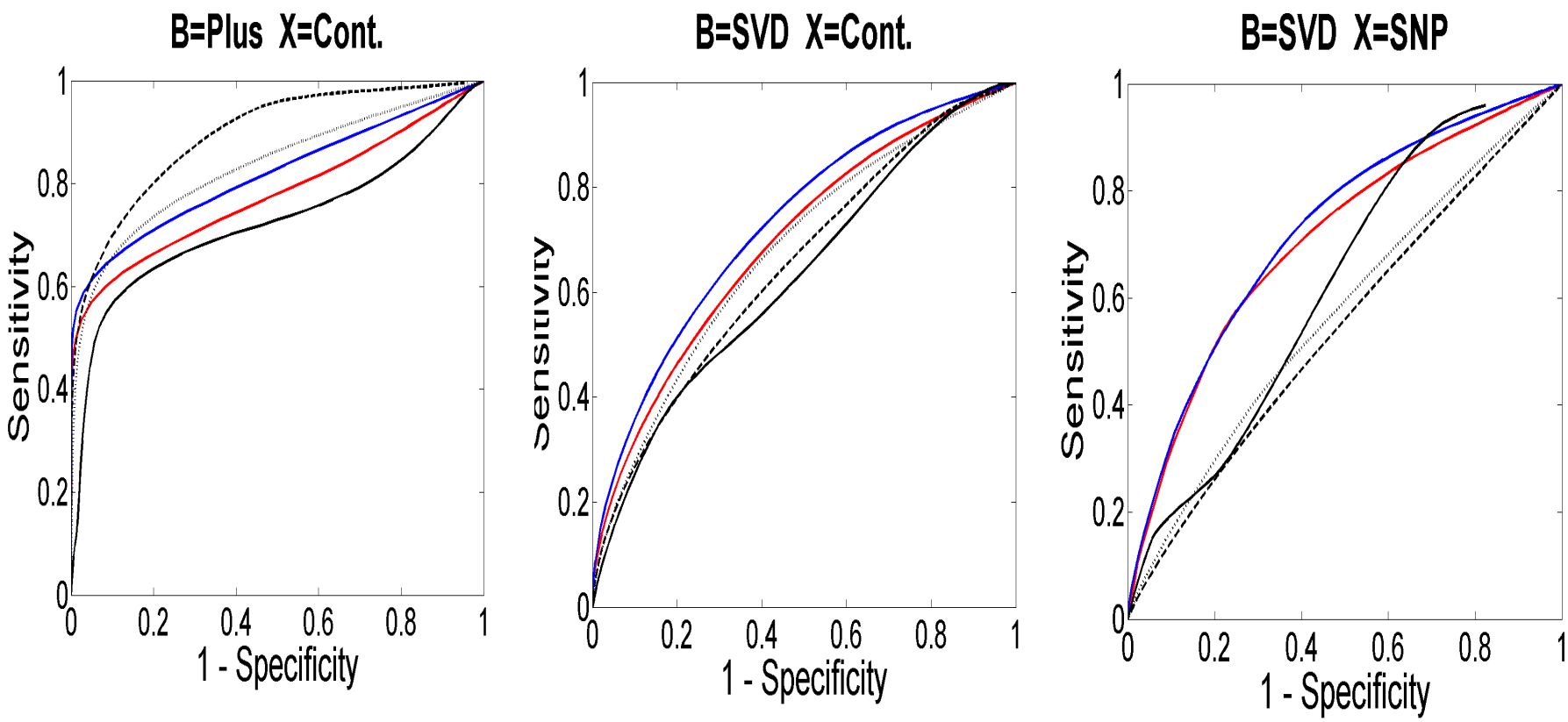
$$p_y \times 1 \qquad p_y \times q \quad q \times 1 \qquad p_y \times 1$$

$$\Sigma_E = \Lambda \Lambda^T + \Sigma_\eta$$

Simulation


Patterns
Plus
SVD
SVD
UN
UN
True B
LASSO
MEN=1.00, BIC=12.4
MEN=0.046, BIC=12.43
MEN=0.14, BIC=13.28
MEN=0.03, BIC=12.52
MEN=0.14, BIC=14.73
BLASSO
MEN=0.21, BIC=12.3
MEN=0.021, BIC=14.32
MEN=0.11, BIC=19.11
MEN=0.02, BIC=13.81
MEN=0.13, BIC=18.45
G-SMuRFS
MEN=0.12, BIC=12.1
MEN=0.018, BIC=14.24
MEN=0.11, BIC=19.08
MEN=0.02, BIC=13.79
MEN=0.13, BIC=18.39
GLRR3
MEN=0.11, BIC=10.87
MEN=6.72, BIC=14.52
MEN=9.16, BIC=13.33
MEN=4.99, BIC=14.41
MEN=20.89, BIC=15.75
GLRR5
MEN=0.13, BIC=10.90
MEN=0.01, BIC=10.99
MEN=0.01, BIC=10.37
MEN=4.22, BIC=14.31
MEN=19.36, BIC=15.79

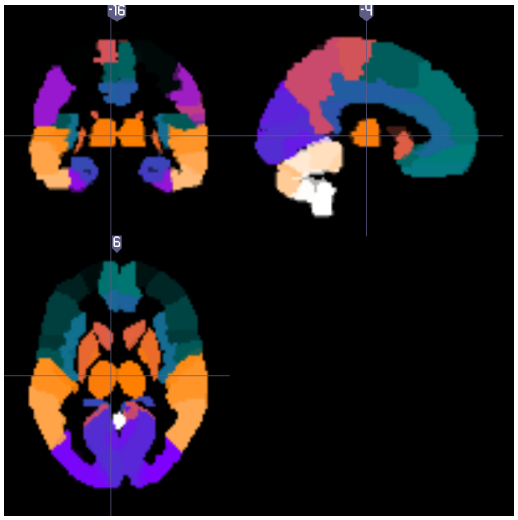
Simulation

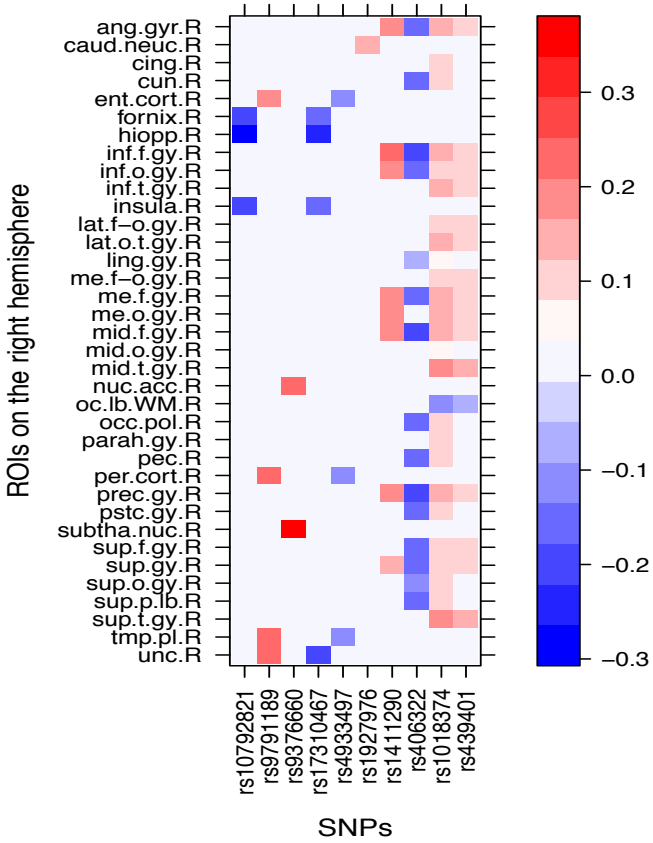

B=Plus X=Cont.
B=SVD X=Cont.
B=SVD X=SNP
Sensitivity
1 - Specificity
0
0.2
0.4
0.6
0.8
1

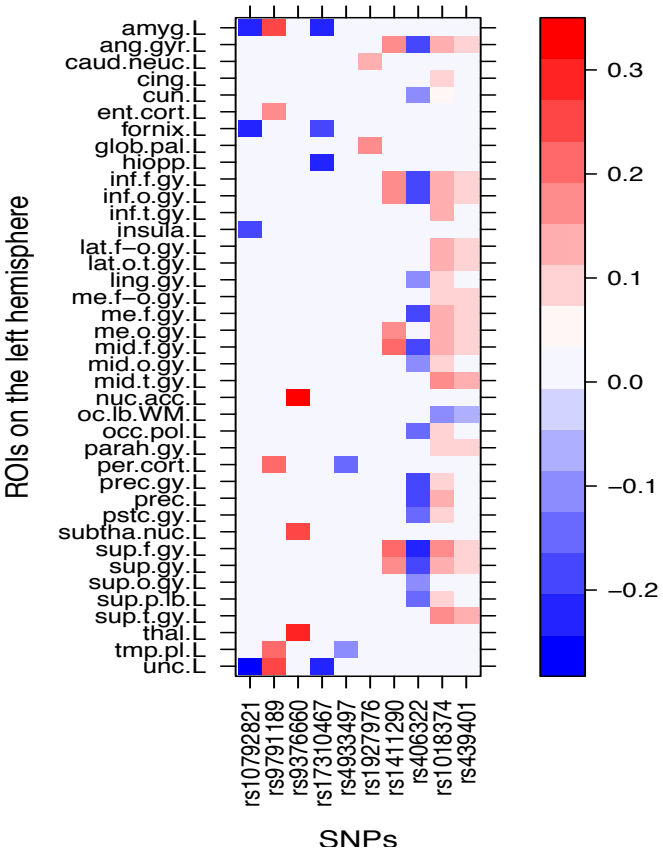
ADNI

749 AD/MCI/NC subjects, 93 ROIs

40 AD candidate genes on the AlzGene web





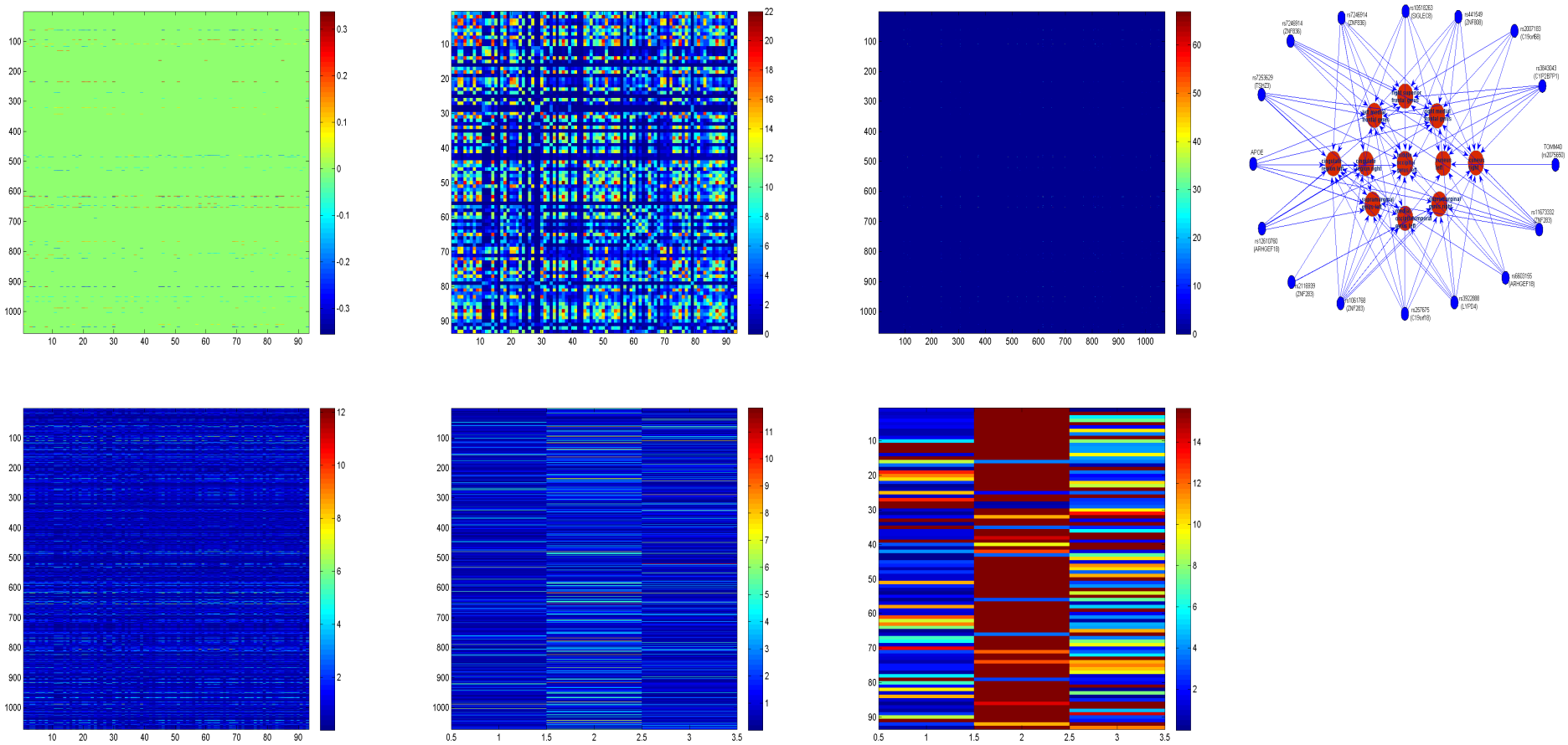


ADNI

$\widehat{B}$

ROI network

Genetic network

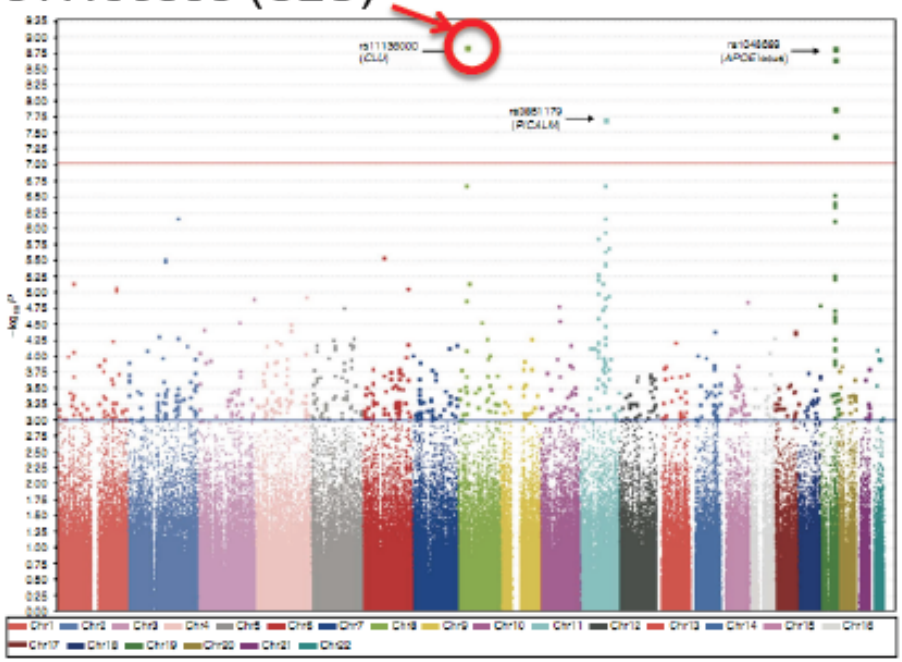


$-\log_{10}(p)$ for $\widehat{B}$

Models for Candidate SNP/Gene Voxel-wise Analysis

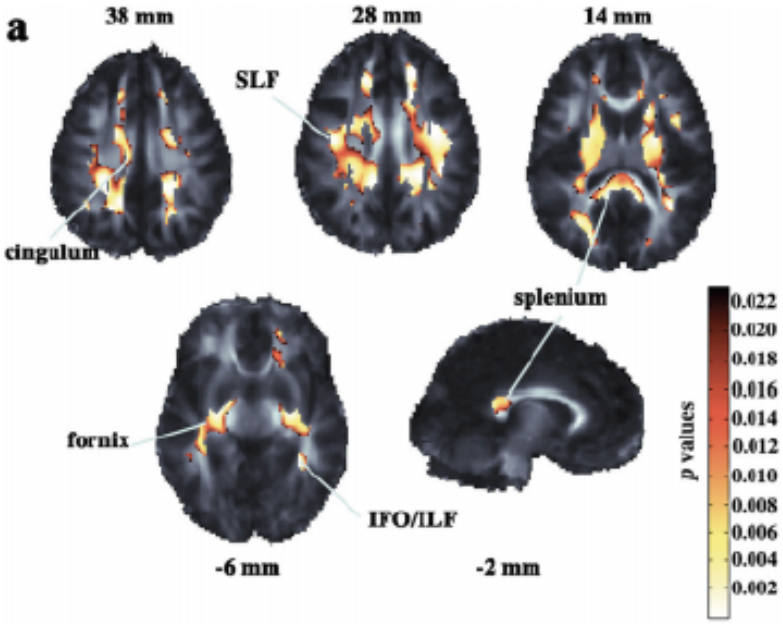
Candidate SNP

rs11136000 (*CLU*)



Genome-wide association identifies variant within the CLU gene in ~4000 Alzheimer's patients and ~8000 controls – but what does it do?

(Harold et al., 2009)



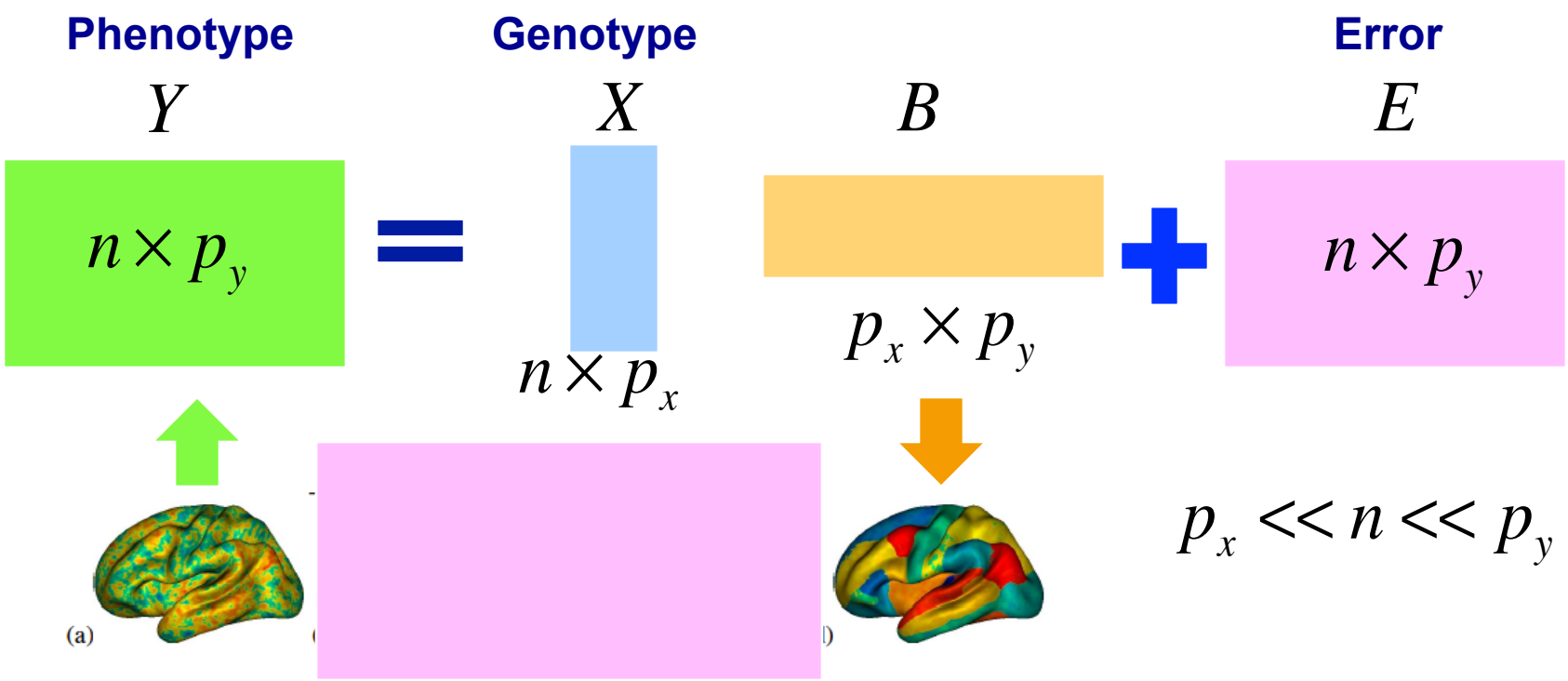
The Alzheimer's associated variant broadly affects white matter integrity in a young cohort – may create early predisposition for disease

(Braskie et al., 2009)

High Dimensional Regression Model

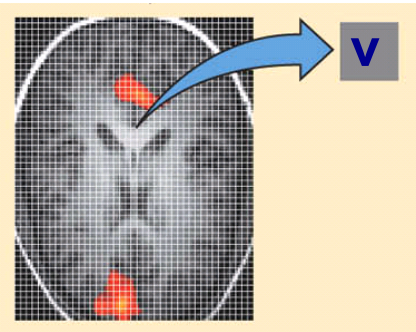
Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V\} \qquad \{X_i(g) : g \in G_0\}$$

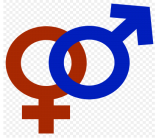


$$p_x << n << p_y$$

Functional Mixed Effects Model



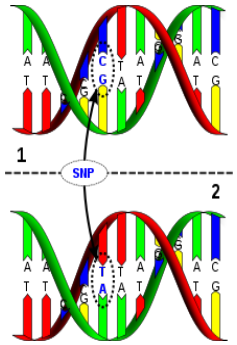
BEHAVIOR GENDER



=

Age

+



+ Error (v)

$$y_i(v) = x_i'\beta(v) + z_i'\gamma(v) + e_i(v)$$

$$\gamma(v) \sim N(\ 0\ ,\ \sigma_\gamma^2(v)\ I_L)$$

$$e_i(v) \sim N(\ 0\ ,\ \sigma_e^2(v)\ )$$

$$\dim(y_i(v)) = 1$$

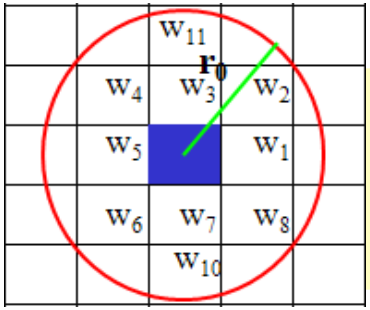
Imaging Responses

Genetic Variation

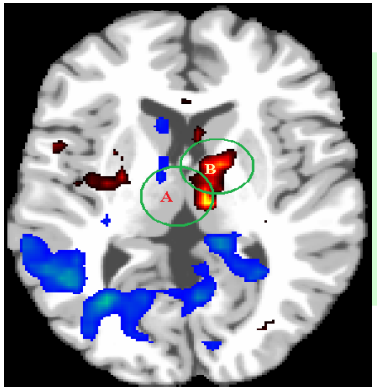
$$\sigma_\gamma^2(v) \overset{?}{=} 0$$

Multiscale Adaptive Smoothing

Estimate $\sigma_\gamma^2(v)$, $\sigma_e^2(v)$ marginally, using **regular** REML



Estimate $\sigma_\gamma^2(v)$ adaptively, using **weighted** REML



Hypothesis Testing for $H_0(v)$: $\sigma_\gamma^2(v) = 0$

Weighted REML

$$Y_{nx1}(v) = X_{nxp}\beta(v)_{px1} + Z_{nxL}\gamma(v)_{Lx1} + E(v)_{nx1}$$

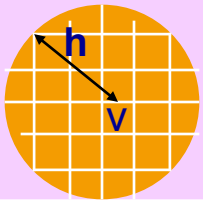
$$\Downarrow \; K_{(n-p)xn}\, X_{nxp} = 0$$

$$Y^*(v) = KY(v) = K\,Z\gamma(v) + K\,E(v) = Z^*\,\gamma(v) + E^*(v)$$

$\Downarrow$ *Regular REML*

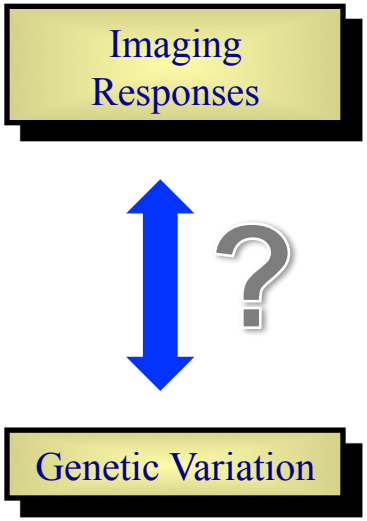
$\Downarrow$ *Weighted REML*

B(v,h) :



Hypothesis Testing

$H_0 : \sigma^2_\gamma(v) = 0$ vs $H_1 : \sigma^2_\gamma(v) \neq 0$



Test statistic T(v) :

$$2\{L_{REML}(Y^*(v)\,|\,Z,\hat{\sigma}^2_\gamma(v),\hat{\sigma}^2_e(v)) - L_{REML}(Y^*(v)\,|\,Z,0,\hat{\sigma}^2_e(v))\}$$

Traditional approach: $T \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$

Our approach : Exact Null Distribution

$$T_n(v) \sim \sup_{\sigma^2(v)\geq 0} \sum_{B(v,h)} \omega(v,v',h) \sum_{i=1}^{L} \frac{\eta_i^2(v)\ \sigma^2_\gamma(v) d_i}{\hat{\sigma}^2_e(v') + \sigma^2_\gamma(v) d_i} - \log(\hat{\sigma}^2_e(v') + \sigma^2_\gamma(v) d_i)$$

d_i : ith eigenvalue of KZ^TZK^T

$\eta_i(v) \sim N(0,1)$

Score Test

Likelihood Ratio Test:
-- Smoothing estimator then computing LRT
-- Solve it for EVERY random sample and EVERY voxels

Regular Score Test : $SC(v) = \frac{1}{2}\sigma_{ev}^{-4}\ Y_v^T \Omega Y_v$ (Tzeng2007)

-- No need to solve maximizer every time

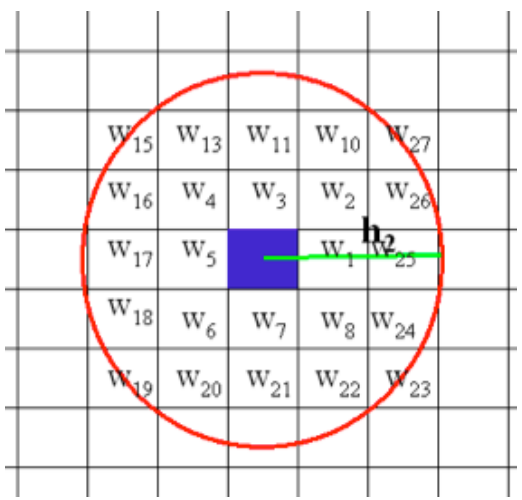
-- E(SC) and E(SC^2) feasible to calculate

But statistical power is not good enough…

Weighted Score Test

Regular Score Test : $SC_{\sigma_\gamma^2}(v) = \frac{1}{2}\sigma_{ev}^{-4}\ Y_v^T \Omega Y_v \sim Gamma(\alpha, \beta)$

(Tzeng2007)



Weighted Score Test

$$WSC_{\sigma_\gamma^2}(v) = \frac{1}{2}\sum_{v' \in B(v,h)} \omega_{v'}\sigma_{ev'}^{-4}\ Y_{v'}^T \Omega Y_{v'}$$

$$\sim Gamma(\alpha', \beta')$$

No tricks : 20 Mins!

ADNI

- **Objective:** Identify brain regions affected by CR1
- **Subject:** 335 elders (174 Controls, 161 ADs)
- **Image:** 128 * 128 * 128 RAVEN maps from baseline T1-weighted images
- **Demographic Info:** Gender, Baseline age, Baseline age square, APOE Risk, Handedness, Education Level, Baseline intracranial volume
- **Genetic Variation:** CR1 (16 SNPs, MAFs from 13% ~ 41%)
- **Method:** Functional Mixed Effects Model (FEME)

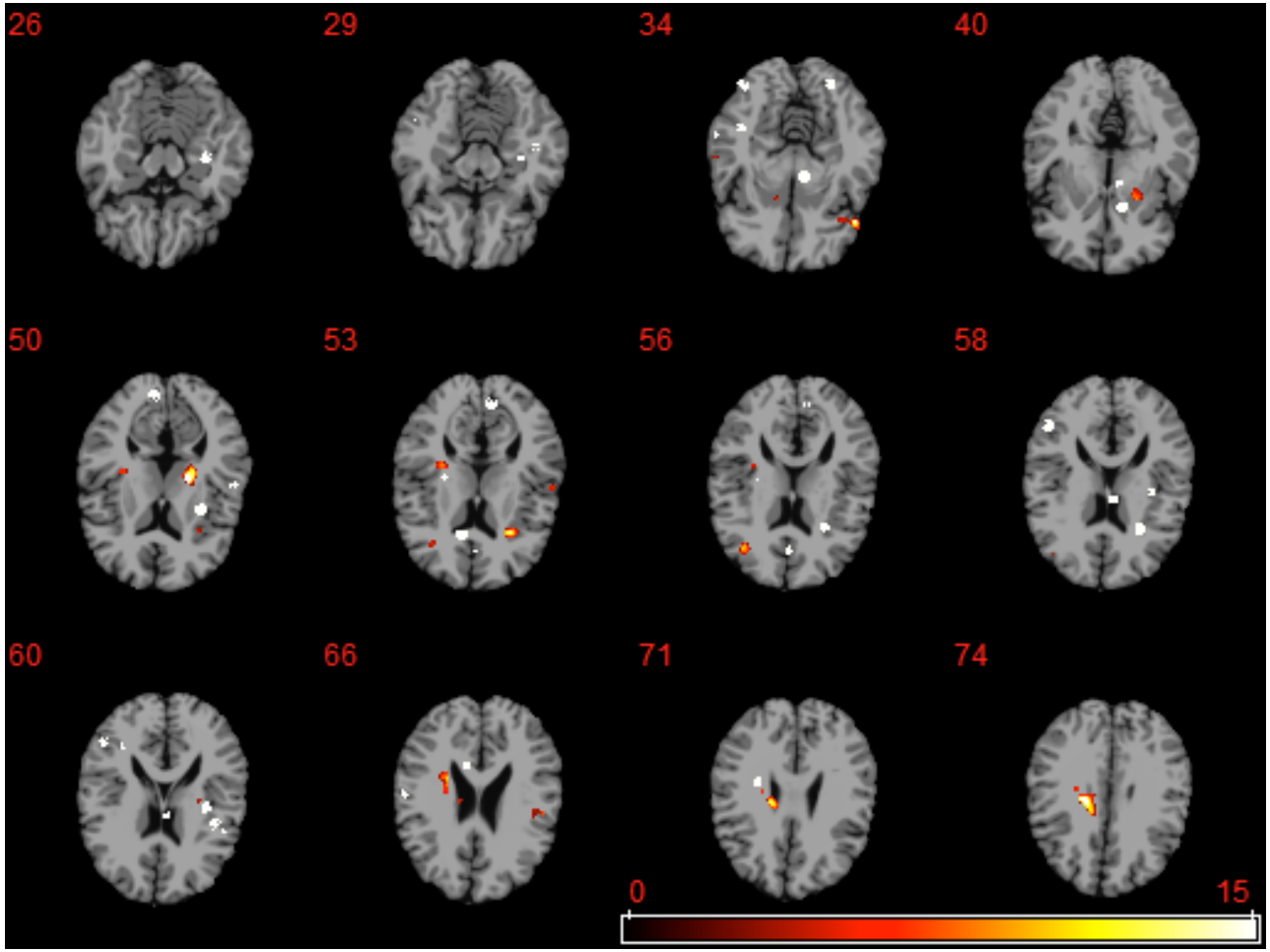
ADNI

45 ROIs by FMEM, 5 ROIs by Voxel-based

Superior temporal gryus (R167)
Inferior temporal gryus (R150, L378)
Precentral gyrus (R170, L*439)
Middle frontal gyrus (R127, L295)
Postcentral gryus (R60, **L65)
Insula (R53,L74)
Putamen (R200, L*123)
Pallidum (R58)
Fusiform (R117, L306)
Inferior temporal gryus (R219)
Inferior parietal but supramarginal and angular gyri (R87)
Angular (R433)
Inferior frontal gyrus, triangular part (R180, L*79)
Inferior occipital gyrus (R269, L133)
Superior frontal gryus (R71)
Supplementary Motor Area (R117)
Postcentral gryus (R104, L183)
Superior frontal gyrus, medial (R188,L159)
Anterior cingulate and paracingulate gyri (R428, L88)

Median cingulate and paracingulate gyri (R209, L116)
Calcarine fissure and surrounding cortex (R53, L266)
Cuneus (R151, L345)
Superior occipital gyrus (L317)
Middle occipital gyrus (R144)
Precuneus (R61)
Paracentral Lobule (R86)
Caudate (R197)
Lingual (L114)
Inferior frontal gyrus, opercular part (L78)
Middle temporal gryus (L*394)
Inferior frontal gyrus, orbital part (L61)

ADNI

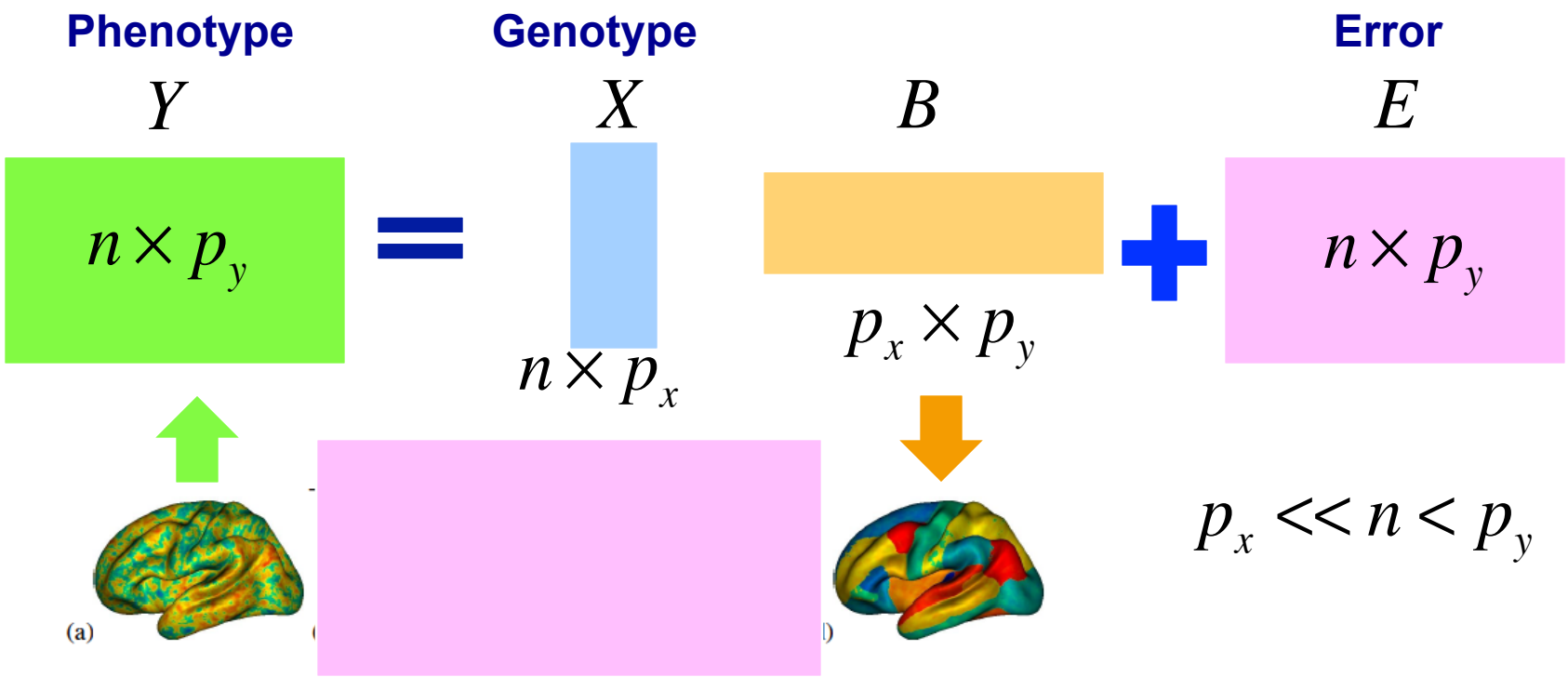


Models for Candidate ROIs Whole Genome-wise Analysis

High Dimensional Regression Model

Data $\{(Y_i, X_i) : i = 1, \cdots, n\}$

$$Y_i = \{y_i(v) : v \in V\} \qquad \{X_i(g) : g \in G_0\}$$



$$p_x << n < p_y$$

Current Approaches

Multivariate Responses

Multivariate Linear Model
+
Hotelling's T^2 Test

Principle Component Regression

Component Wise Method

(False Discovery Rate)

Partial Least Square Regression

Covariate of Interest

Limitations

Multivariate Linear Model + Hotelling's T^2 Test	→	$N > p$ ~~Hotelling T^2 Test~~
Principle Component Regression	→	Low Statistical Power ↓ Larger N
Component Wise Method	→	Inflated Type I Error Rate ↓ Wrong Conclusions
Partial Least Square Regression	→	No p-values!!

Sparse Projection Regression Model

- Multivariate regression with a high-dimensional responses and a multivariate covariate of interest
- Consider a Multivariate Linear Model (MLM):

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad or \;\; \mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i + \mathbf{e}_i$$

- We are interested in the hypothesis testing problem:

$$H_0 : \mathbf{C}\mathbf{B} = \mathbf{B}_0 \quad v.s. \quad H_1 : \mathbf{C}\mathbf{B} \neq \mathbf{B}_0$$

- Diverging q , fixed p case
  - High-dimension two sample test
  - Imaging genetics association study

Sparse Projection Regression Model

- Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$, then a projection regression model is given by:

$$\mathbf{W}^T y_i = (\mathbf{BW})^T \mathbf{x}_i + \mathbf{W}^T \mathbf{e}_i = \beta_{\mathbf{w}}^T \mathbf{x}_i + \varepsilon_i$$

- Hypothesis problem reduces to:

$$H_{0W} : \mathbf{C}\beta_{\mathbf{w}} = \mathbf{b}_0 \quad v.s. \quad H_{1W} : \mathbf{C}\beta_{\mathbf{w}} \neq \mathbf{b}_0$$

$$where\ \mathbf{C}\beta_{\mathbf{w}} = \mathbf{CBW}\ and\ \mathbf{b}_0 = \mathbf{B}_0\mathbf{W}$$

- How to determine an 'optimal' $\mathbf{W}$?

Sparse Projection Regression Model

- We show that this is achieved by optimizing the following generalized heritability ratio (GHR):

$$\mathrm{GHR}(\mathbf{w};\mathbf{C}) = \frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T S_{\tilde{X}_1}(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}} = \frac{\mathbf{w}^T\Sigma_C\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}$$

- High Dimensional Setting
- noise accumulation
  - ill-conditioned sample covariance estimator: $\hat{\Sigma}_R$
- Sparse Projection Regression Model is proposed as following:

$$\mathrm{argmax}\{\frac{\mathbf{w}^T\hat{\Sigma}_C\mathbf{w}}{\mathbf{w}^T\tilde{\Sigma}_R\mathbf{w}}\} \quad \text{s.t.} \quad ||\mathbf{w}||_1 \leq t$$

Sparse Projection Regression Model

- After estimating $\mathbf{W}$, we can calculate a $k \times k$ matrix as:

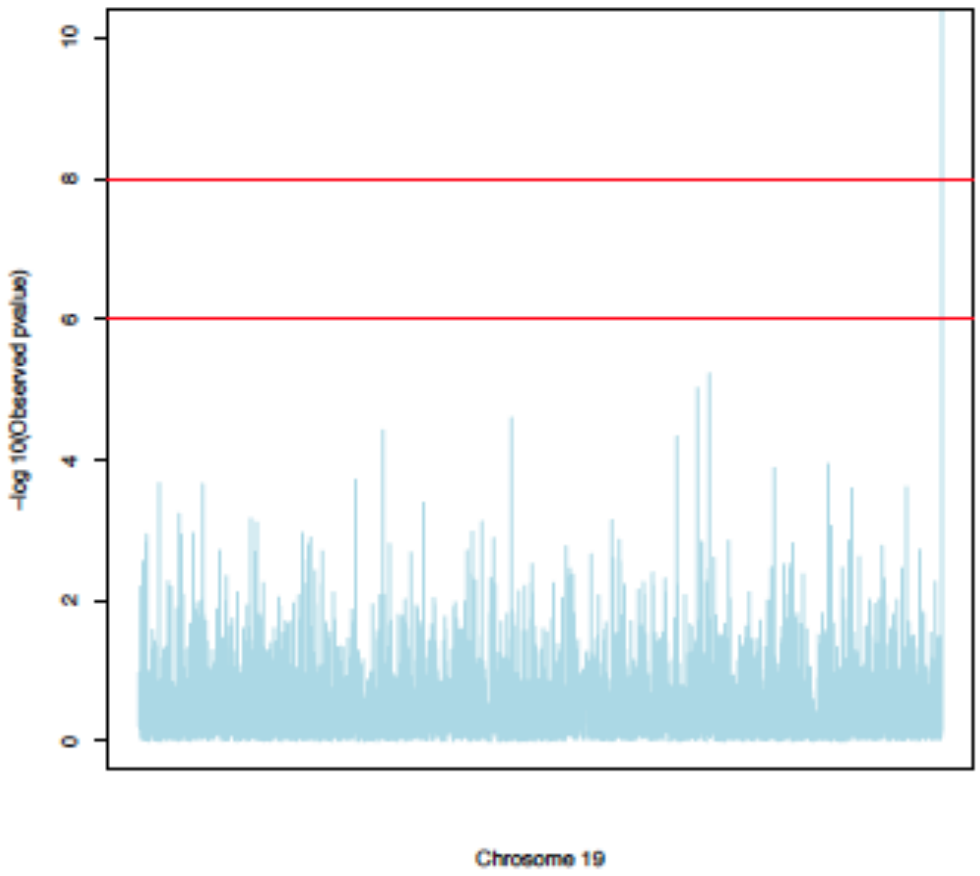
$$T_n = (\mathbf{C}\hat{\beta}_{\mathbf{w}} - \mathbf{b}_0)^T \Sigma_{\tilde{\Omega}}^{-1} (\mathbf{C}\hat{\beta}_{\mathbf{w}} - \mathbf{b}_0)$$

- Test statisitic: $\mathrm{Tr}_n = \mathrm{trace}(T_n)$
- Wild bootstrap
  - Fit MLM under the null hypothesis to calculate the estimated multivariate regression coefficient, denoted by $\widehat{\mathbf{B}}_0$, residuals $\hat{\mathbf{e}}_i = \mathbf{y}_i - \widehat{\mathbf{B}}_0^T \mathbf{x}_i$.
  - Generate G bootstrap samples $\mathbf{z}_i^{(g)} = (\widehat{\mathbf{B}}_0)^T \mathbf{x}_i + \eta_i^{(g)} \hat{\mathbf{e}}_i$.
  - Repeat the estimation procedure for estimating the optimal weights and the calculation of the test statistic $\mathrm{Tr}_n^{(g)}$.
  - p-value of Tr_n is computed as $\frac{1}{G} \sum_{g=1}^{G} \mathbf{1}(\mathrm{Tr}_n^{(g)} \geq \mathrm{Tr}_n)$.

ADNI

- Problem of interest: to perform a genome-wide search for establishing the association between the $10,479$ SNPs collected on chromosome 19 and the brain volume of 93 regions of interest (ROIs).
- 93 ROIs~intercept, a specific SNP, age, gender, whole brain volume, the top 5 principal components
- Several important findings:
  - The ApoE allele was identified as the top one significant covariate with $-\log_{10}(p) \sim 15$.
  - The rs207650 on the TOMM40 gene is found to be one of the top 3 significant SNPs with $-log_{10}(p) \sim 5$, Vounou et al. (2012).
  - Able to detect some additional SNPs, such as rs11667587 on the NOVA2 gene.

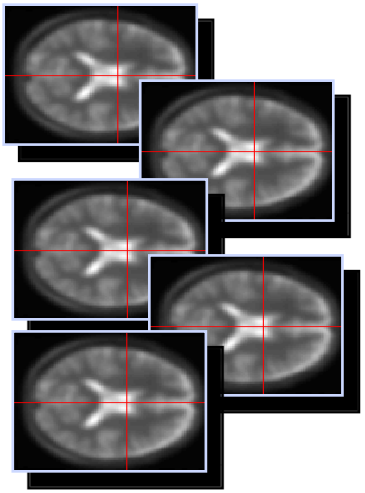
ADNI



Fast Voxel-wise Genome-wise Analysis

Data Structure

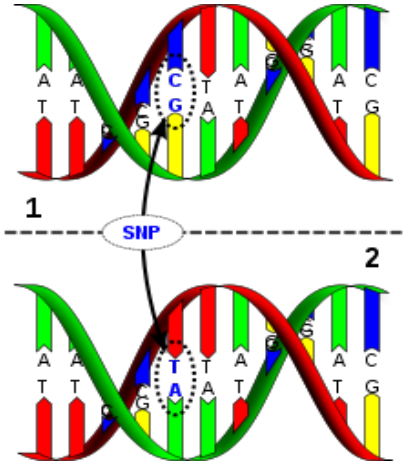
Imaging:



Person No.1 ……. Person No.100

3D Matrix ……. 3D Matrix

Genetic Variation:



SNP1 SNP2 ……. SNP2000

Person No. 1

.

.

.

Person No. 100

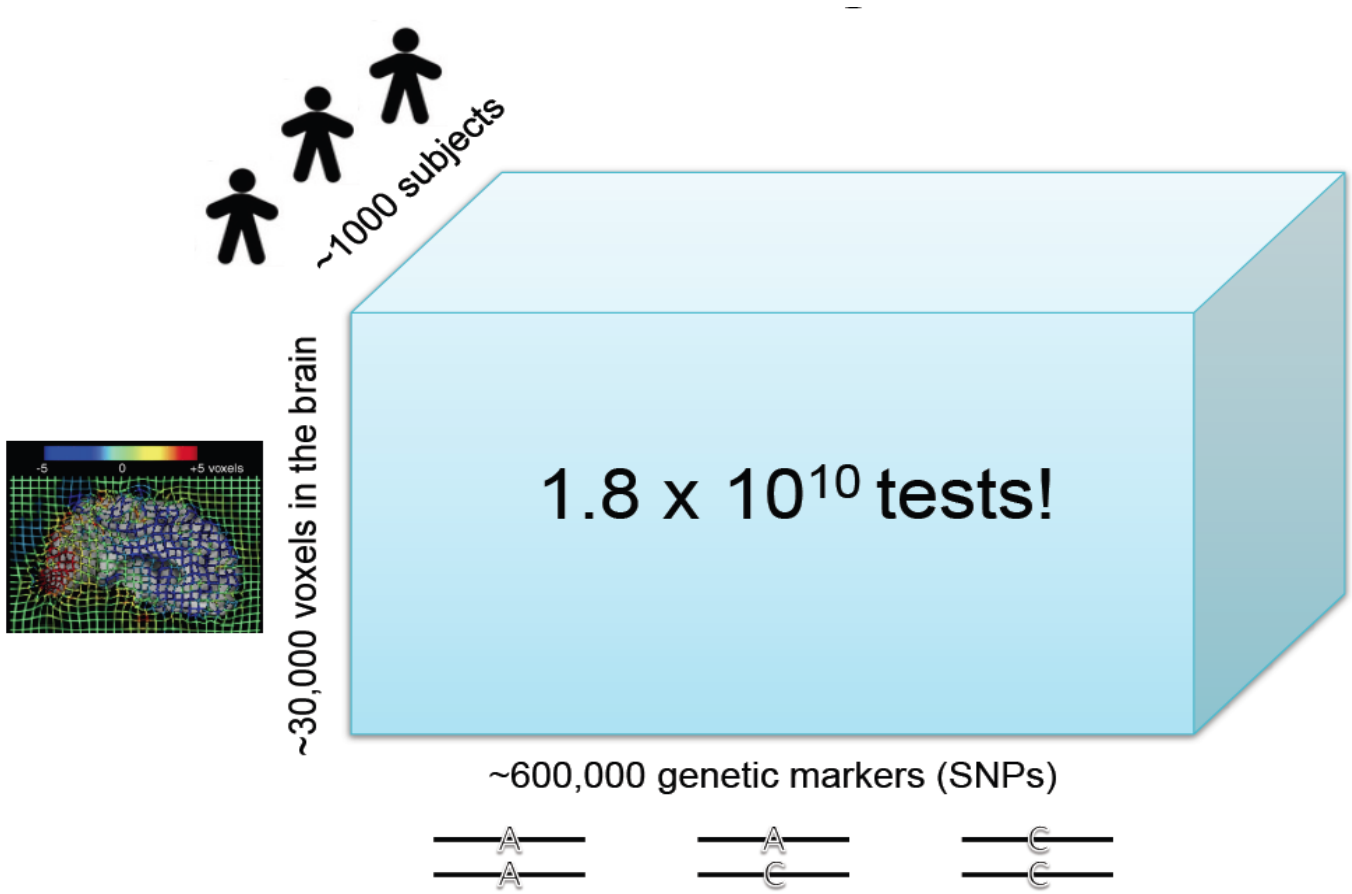
$$\begin{bmatrix} 1 & 2 & \cdots & 0 \\ 0 & \ddots & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 2 \end{bmatrix}$$

vGWAS

Advantages	Disadvantages
Candidate SNPs allow you to test a specific biological hypothesis	It is highly likely that we don't know the genetic underpinnings of a trait like brain structure so in general don't know the right SNP to pick
Strong hypothesis drives clearly interpretable results	In order to be widely accepted, the variant needs to have strong prior evidence (genome-wide significant in a meta-analysis or have clear function)
Multiple comparisons burden is reduced (one SNP – many voxels)	Unable to search the genome, only characterize the effect of a known variant
Quick way to provide functional relevance to unbiased genome-wide search results	

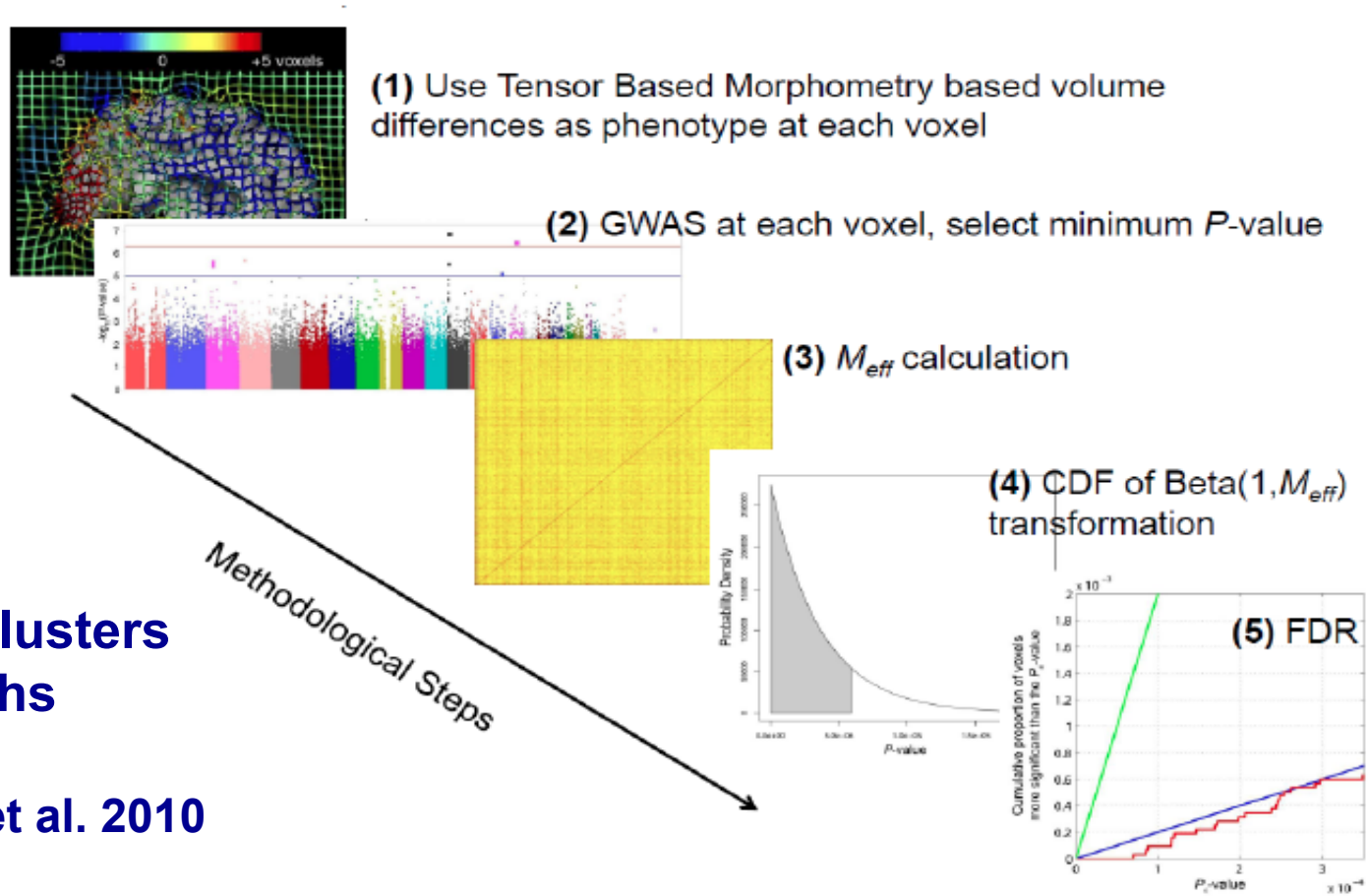
Dr. Hibar

vGWAS



Hibar, et al. HBM 2012

vGWAS

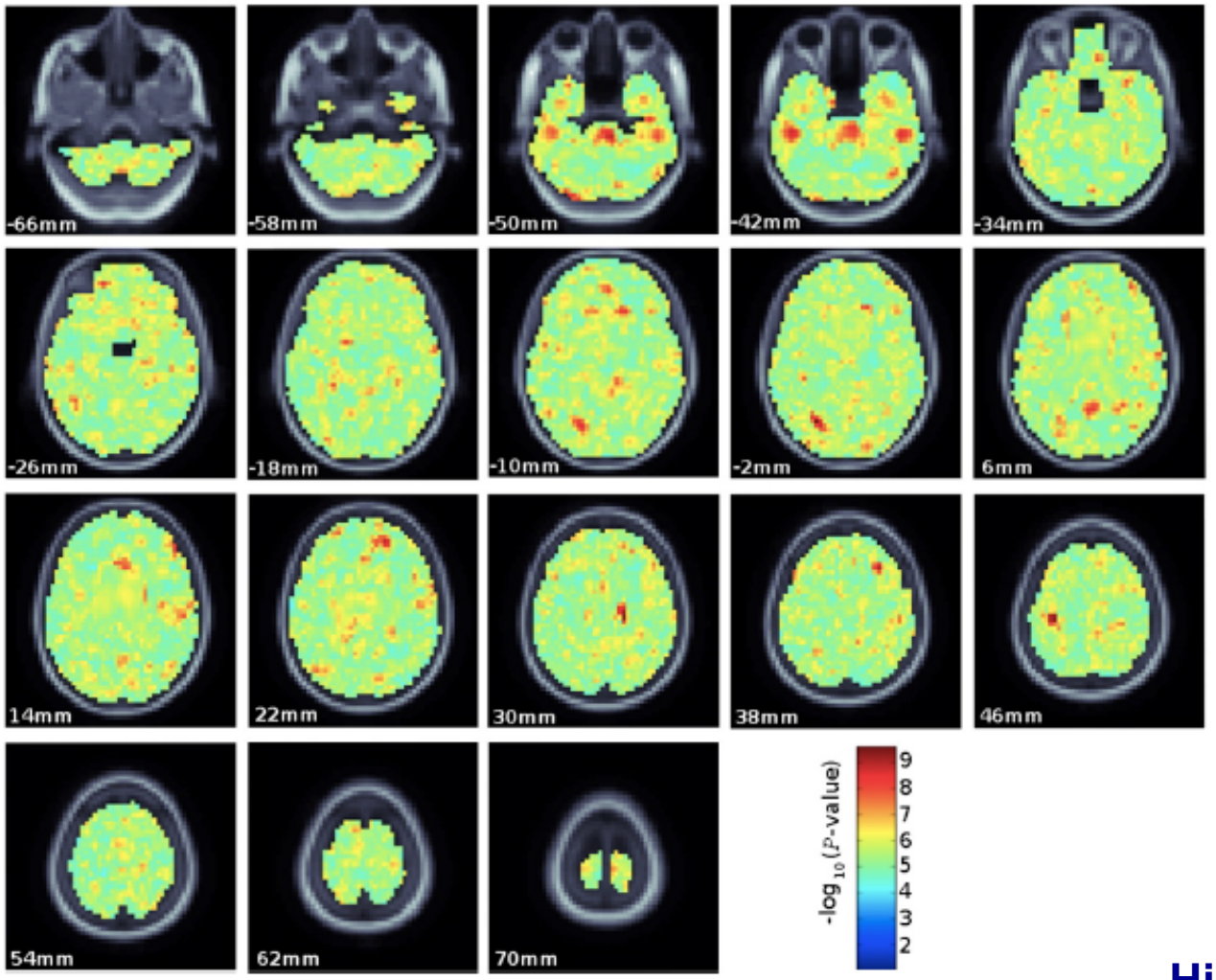


500 clusters months

Stein et al. 2010

Figure 1: Summary of vGWAS method for conducting association and assessing statistical significance.

Raw Minimum p-value at each voxel



Hibar et al. 2011

Most associated genes

Chr	Gene	# of SNPs in gene	# of			Volume (mm3)	Proportion of brain volume	Clustermax (mm3)	# of clusters
11	GAB2	20				6336	0.0049	2688	9
11	LRDD	2				8128	0.0063	7872	4
12	PTPRB	17				3200	0.0024	3008	5
9	ZNF462	9			10^{-5}	2688	0.0021	2688	1
21	IGSF5	27	14	[illegible]	1.62×10^{-5}	16,384	0.013	9344	3
2	SLC25A12	10	5	9.48×10^{-9}	2.66×10^{-5}	1856	0.0014	1792	2
11	MRE11A	9	6	9.86×10^{-9}	8.80×10^{-6}	9344	0.0072	9216	3
19	SLC8A2	11	7	1.06×10^{-8}	3.18×10^{-5}	5632	0.0043	5376	3
15	CHRM5	3	3	1.71×10^{-8}	1.77×10^{-5}	1280	0.00099	1216	2
18	SPIRE1	19	14	2.94×10^{-8}	2.88×10^{-5}	6016	0.0046	3072	12
3	C3orf64	9	8	3.71×10^{-8}	2.43×10^{-5}	4352	0.0034	2112	4
21	S100B	1	1	4.75×10^{-8}	2.81×10^{-5}	9344	0.0072	6656	7
1	CRCT1	1	1	5.54×10^{-8}	2.90×10^{-5}	4096	0.0032	3456	4
19	ZNF626	6	5	5.85×10^{-8}	2.12×10^{-5}	2560	0.002	2112	3
1	ELK4	1	1	6.05×10^{-8}	3.27×10^{-5}	4032	0.0031	2688	4
11	RSF1	8	6	9.30×10^{-8}	1.31×10^{-5}	768	0.00059	768	1
20	WFDC11	2	2	1.06×10^{-7}	2.49×10^{-5}	1280	0.00099	512	5
6	SCML4	27	18	1.07×10^{-7}	1.67×10^{-5}	3328	0.0026	3328	1
12	ERP27	8	14	1.08×10^{-7}	2.61×10^{-5}	2624	0.002	2176	2

AD Risk Gene (Reiman et al., 2007)

A Heteroscedastic Linear Model

$$y_i(v) = \boldsymbol{x}_i^T \boldsymbol{\beta}(v) + \boldsymbol{z}_i(c)^T \boldsymbol{\gamma}(c,v) + e_i(v) \quad \text{for} \quad i = 1,\ldots,n$$

where $\boldsymbol{\beta}(v) = \left(\beta_1(v),\ldots,\beta_K(v)\right)^T$ is a $K \times 1$ vector associated with non-genetic predictors, and $\boldsymbol{\gamma}(c,v) = \left(\gamma_1(c,v),\ldots,\gamma_L(c,v)\right)^T$ is an $L \times 1$ vector of genetic fixed effects (e.g., additive or dominant). Moreover, $e_i(v)$ are measurement errors with zero mean and $\boldsymbol{e}_i = \left\{e_i(v) : v \in V\right\}$ are independent across i.

A Heteroscedastic Linear Model

We need to test:

$$H_0(c,v):\gamma(c,v)=0 \quad \text{versus} \quad H_1(c,v):\gamma(c,v)\neq 0 \quad \text{for} \quad \text{each} \quad (c,v)$$

We calculate a Wald-type statistic as:

$$W(c,v)=\tilde{\gamma}(c,v)^T\left\{\mathrm{Cov}\left(\tilde{\gamma}(c,v)\right)\right\}^{-1}\tilde{\gamma}(c,v)$$

$$=\mathrm{tr}\left\{\left\{\boldsymbol{Z}_c^T\left(I_n-P_X\right)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T\left(\boldsymbol{I}_n-P_X\right)\sigma_e^{-2}(c,v)\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\left(\boldsymbol{I}_n-P_X\right)\boldsymbol{Z}_c\right\}$$

Big-Data Challenges

10^4

$X : p_x \times n$

10^6

10^6

$B : p_x \times p_y$

10^7

10^4

$Y : n \times p_y$

10^7

Memory:

$$O((p_x + p_y)n + p_x p_y)$$

Computational time: $O(p_x p_y n) = O(10^{17})$

Big-data Challenges

Several big-data challenges arise from the calculation of $W(c,v)$ as follows.

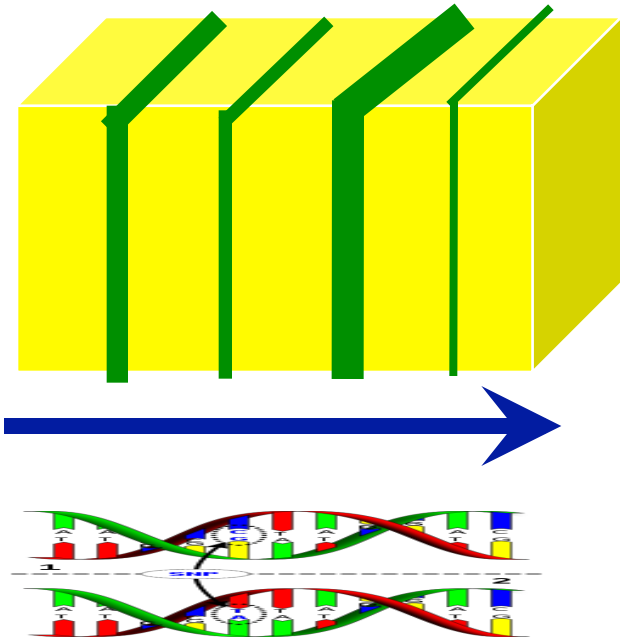
- **Calculating $\sigma_e^2(c,v)$ across all (c,v)s' can be computationally intensive.**
- **Holding all $W(c,v)$ in the computer hard drive requires substantial computer resources.**
- **Speeding up the calculation of $W(c,v)$.**

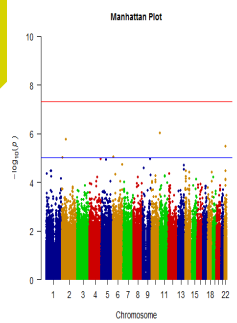
FVGWAS

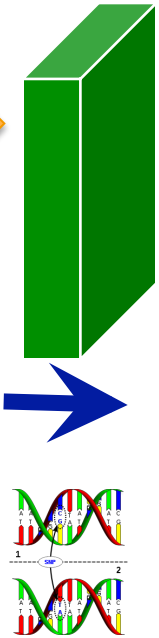
(I) Spatially Heteroscedastic Linear Model

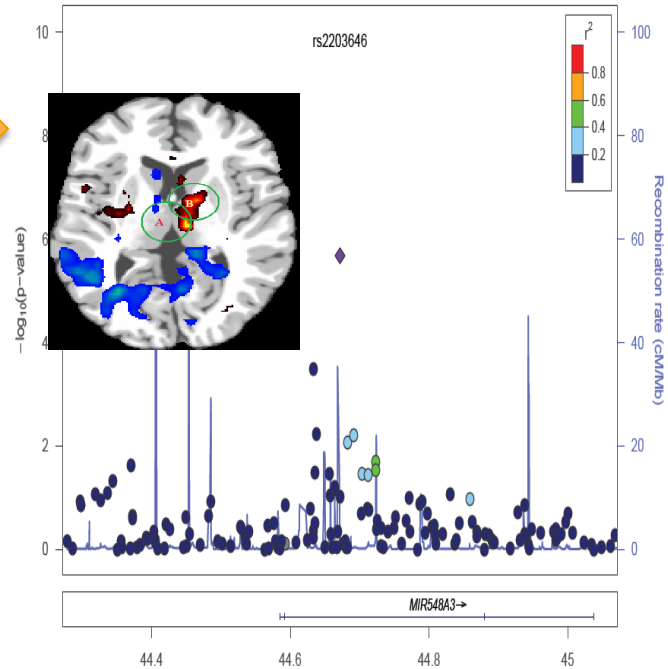
(II) Global Sure Independence Screening Procedure

(III) Detection Procedure







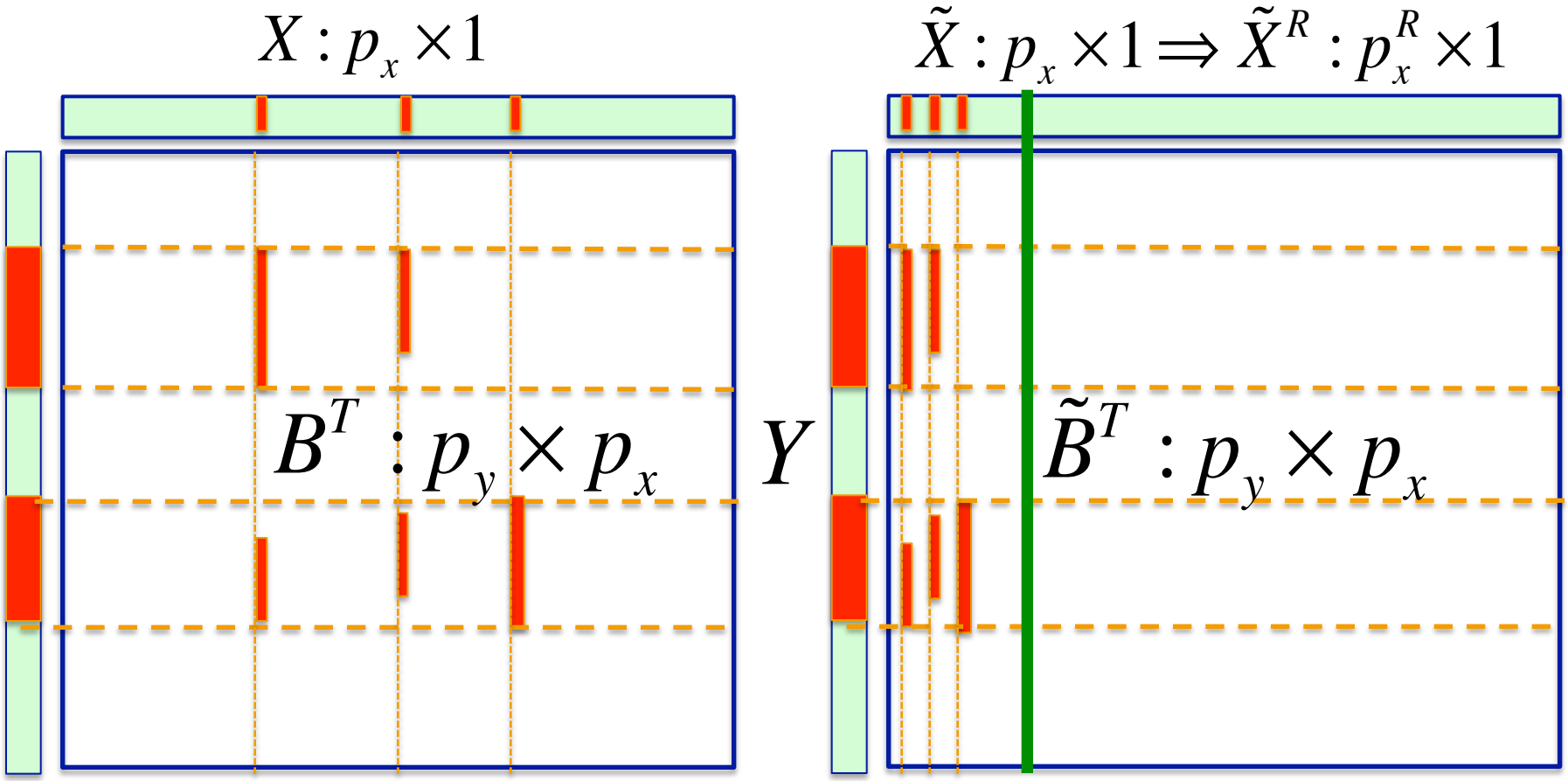


UNIVERSITY OF ALBERTA

"uplifting the whole people"

— HENRY MARSHALL TORY, FOUNDING PRESIDENT, 1908

Key Features



X: Sparsity; Y|X: Clustered ROIs

FVGAWS

To solve these computational bottlenecks, we propose two solutions as follows.

•Calculate $\sigma_e^2(c,v)$ under the null hypothesis $H_0(c,v)$ for each *v* and *c*.

•Develop a GSIS procedure to eliminate many 'noisy' loci based on a global Wald-type statistic.

❖By using these two solutions, we are able to reduce the computational complexity from $O(p_x p_y n)$ to $O((p_x + p_y)n^2)$.

A Global Sure Independence Screening

The global Wald-type statistic at locus *c* is defined as

$$W(c) = N_V^{-1}\mathrm{tr}\left\{\left\{\boldsymbol{Z}_c^T\left(\boldsymbol{I}_n - P_X\right)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T\left(\boldsymbol{I}_n - P_X\right)\left\{\sum_{v\in V}\hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\right\}\left(\boldsymbol{I}_n - P_X\right)\boldsymbol{Z}_c\right\}$$

$\left\{\sum_{v\in V}\hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\right\}$ **is independent of *c*,**

the complexity of computing $\{W(c)\}$ **is** $O\left(\left(p_x + p_y\right)n^2\right)$

the complexity of computing $\{W(c,v)\}$ **is** $O\left(p_x p_y n\right)$

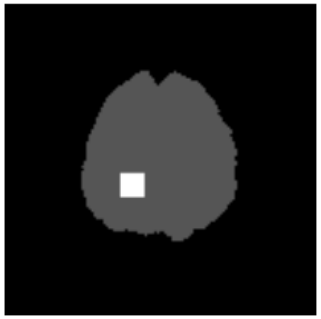
$$p_x p_y / \left\{\left(p_x + p_y\right)n\right\} = 10^{6+7} / \left\{\left(10^7 + 10^6\right)10^4\right\} = 10^2$$

$$= 10^{6+7} / \left\{\left(10^7 + 10^6\right)10^3\right\} = 10^3$$

Simulation Studies

Simulation settings: the dark, gray, and white regions in the figure, respectively, represent background, brain region, and the effected ROI associated with the causal SNPs.

ROI: 10×10



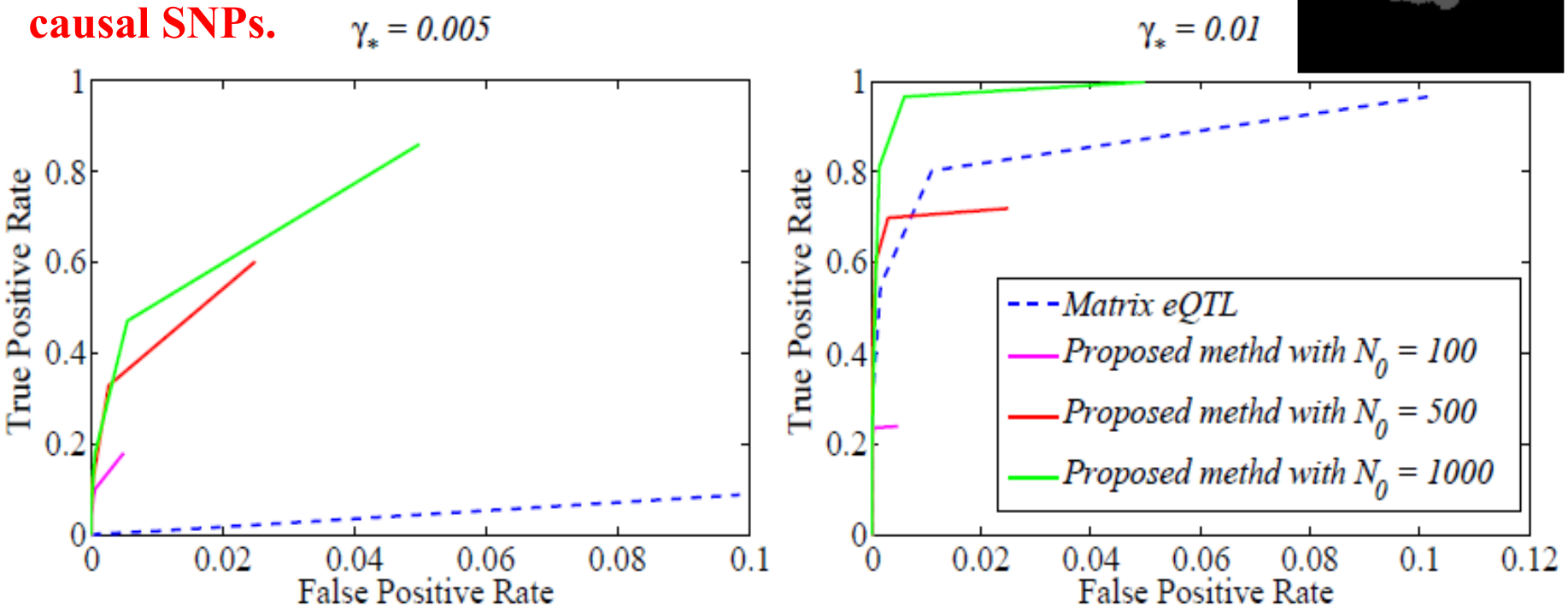
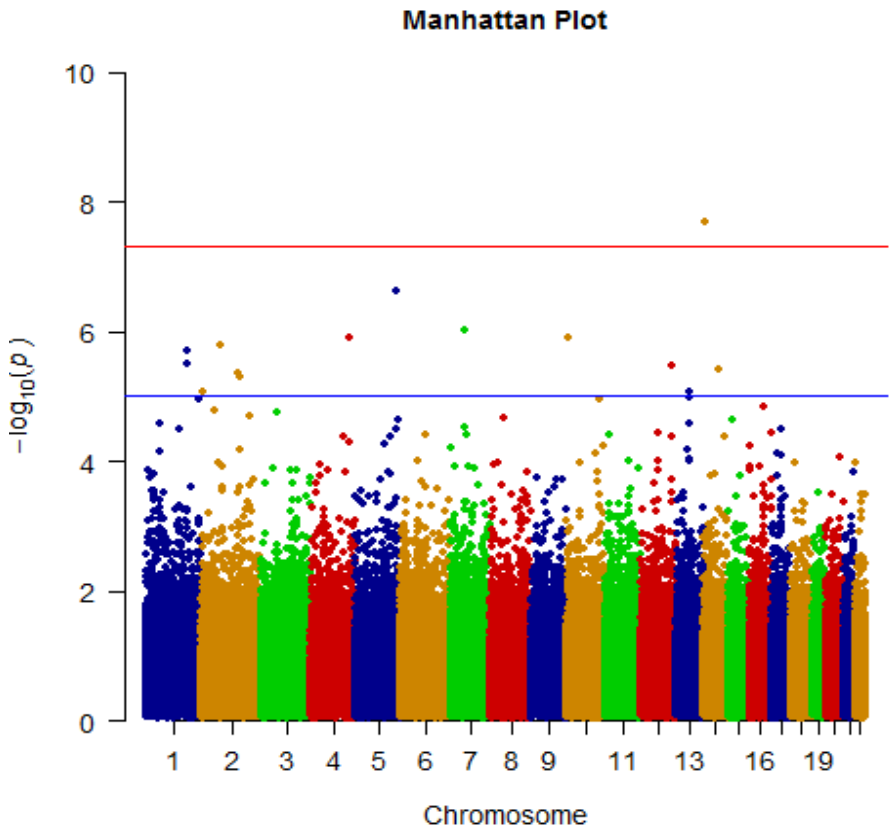
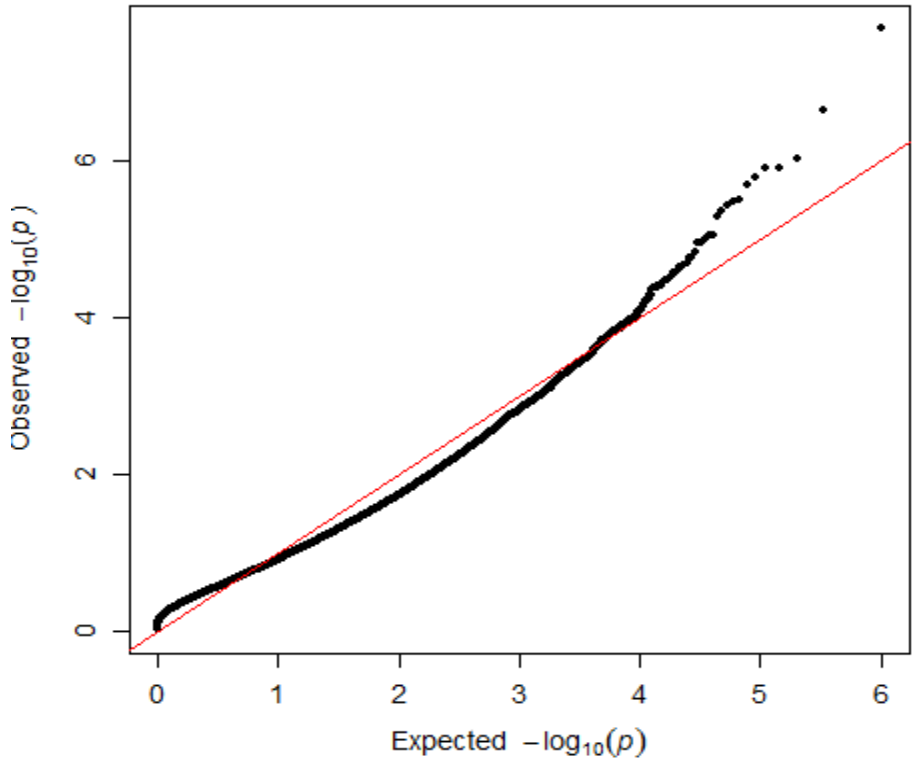


Fig. Simulation results for comparisons between FVGWAS and the Matrix eQTL in identifying significant voxel-SNP pairs.

Results

Our computational time
About 33,800 s





ADNI Data Analysis

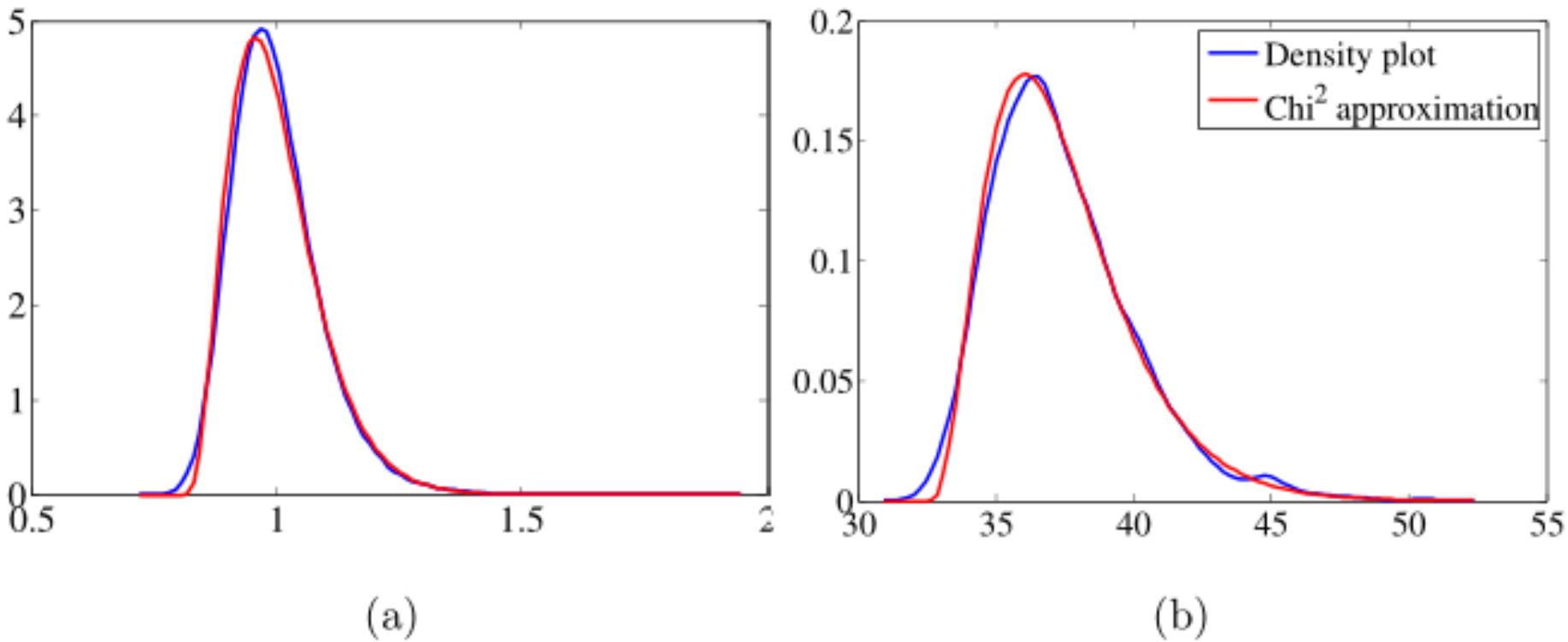


Fig. ADNI whole-brain GWAS: (a) the density plot of and its approximation; (b) the density plot of and its approximation

ADNI Data Analysis

Table. RAVEN map GWAS: significant voxel-SNP pairs at the 0.5 significance level (left) and significant cluster-SNP pairs at the 0.5 significance level (right)

SNP	Number of voxel-SNP pairs	SNP	Number of cluster-SNP pairs	Max cluster	*p*-value of the max cluster
rs2075650 (TOMM40)	23	rs11815438	1	7906	0.11
rs9490103	4	rs2480271	1	7365	0.23
rs2244634	2	rs7001339	1	6864	0.45

ADNI Data Analysis

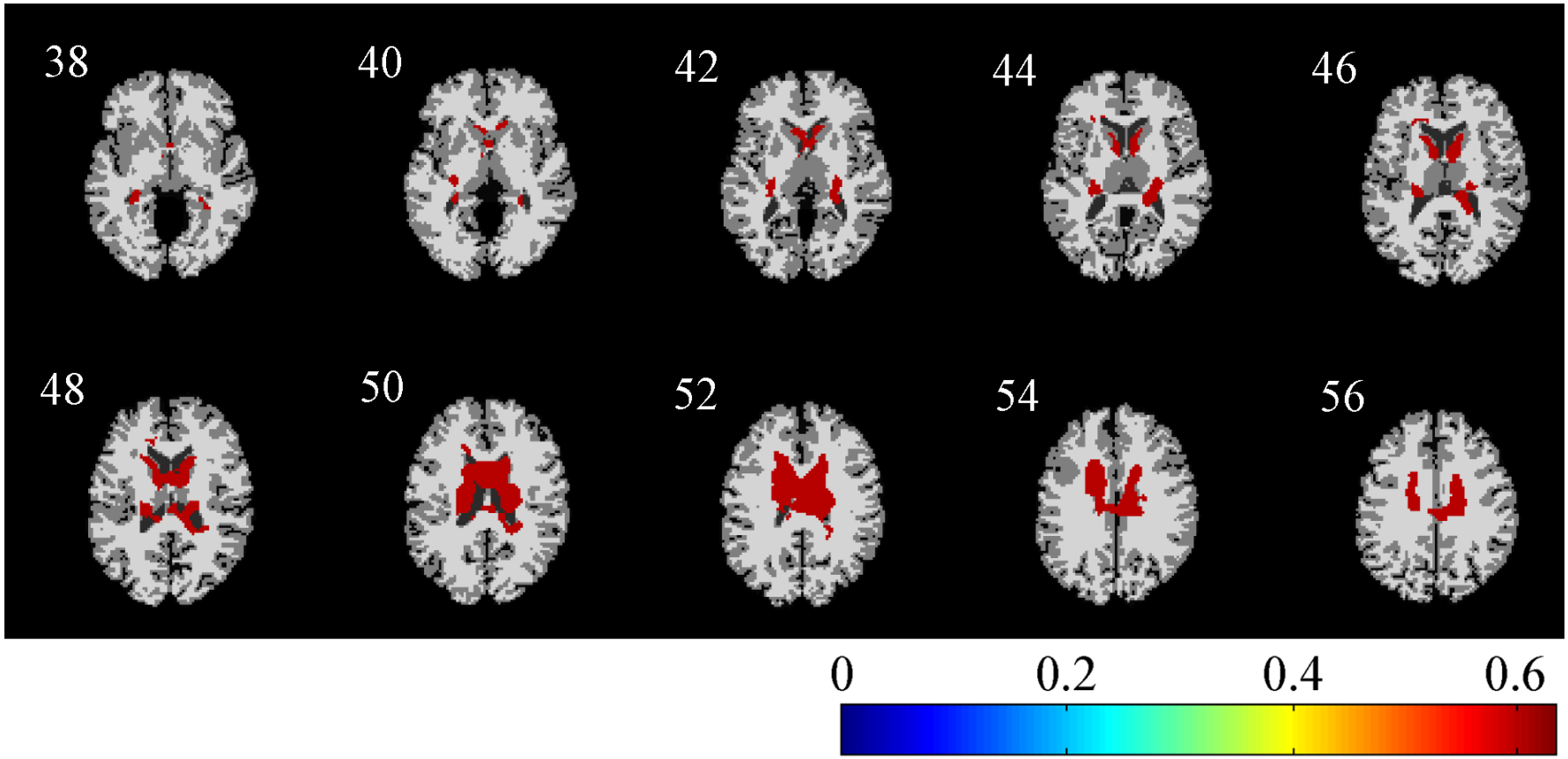
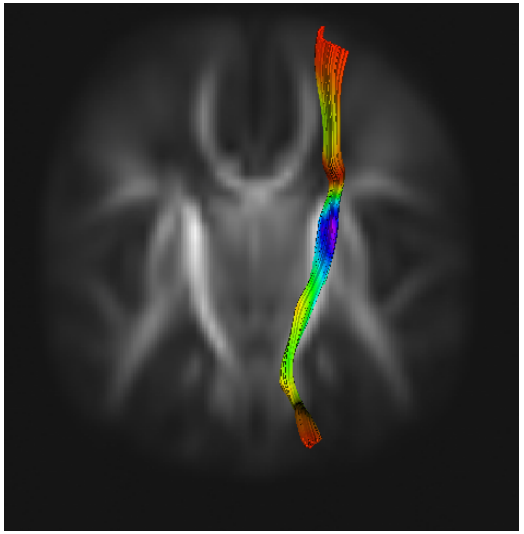


Fig. ADNI whole-brain GWAS: selected slices of $-\log_{10}(p)$ for significant clusters corresponding to a SNP (rs2480271).

Motivating Ex: DTI Fiber Tract Data



Data

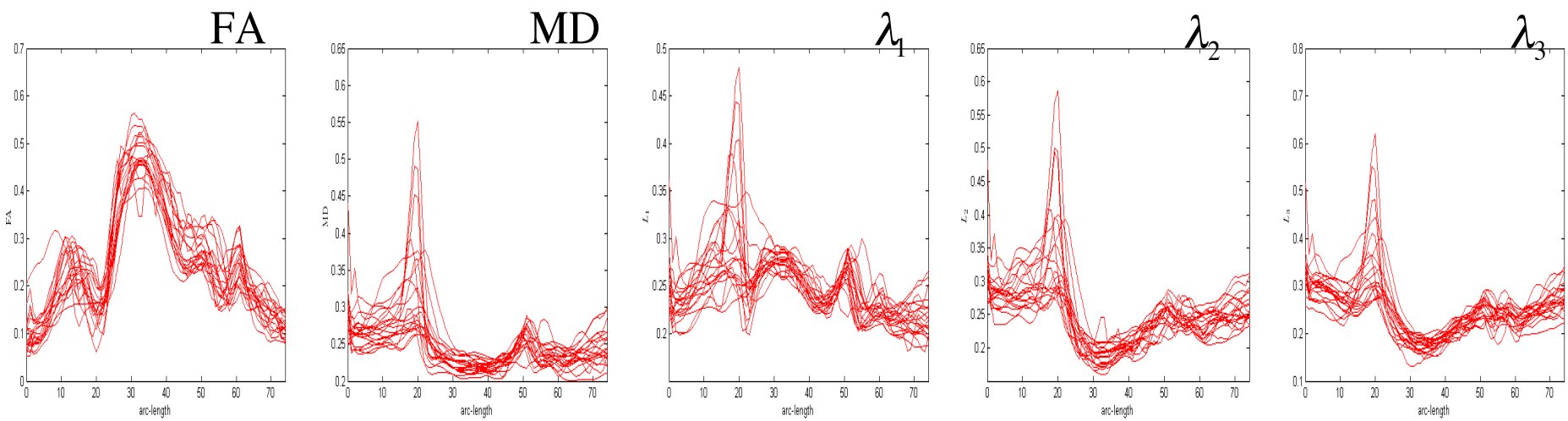
- **Diffusion properties (e.g., FA, RA)**

$$Y_i(s_j) = (y_{i,1}(s_j), \cdots, y_{i,m}(s_j))^T$$

- **Grids** $\{s_1, \cdots, s_{n_G}\}$

- **Covariates (e.g., age, gender, diagnostic)**

$$x_1, \cdots, x_n$$



MVCM

Decomposition:

$$y_{i,k}(s) = x_i^T B_k(s) + z_i(g)^T \beta_k(s,g) + \eta_{i,k}(s) + \varepsilon_{i,k}(s)$$

Coefficients

$x_1, \cdots, x_n$

Long-range Correlation

$\eta_{i,k}(\bullet) \sim SP(0, \Sigma_\eta)$

Short-range Correlation

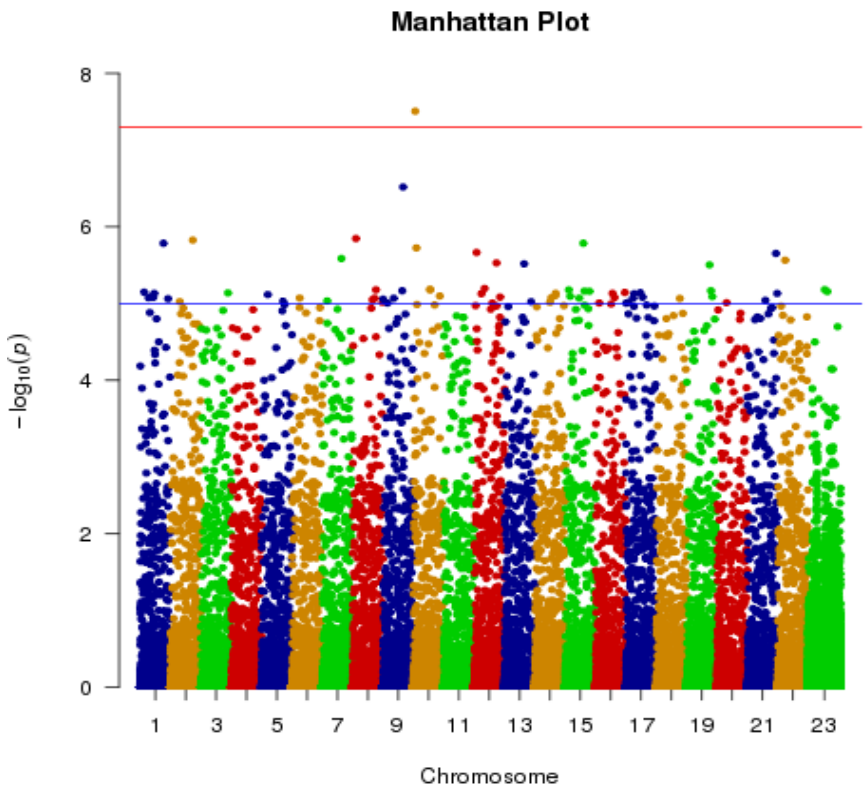
$\varepsilon_{i,k}(\bullet) \sim SP(0, \Sigma_\varepsilon),$

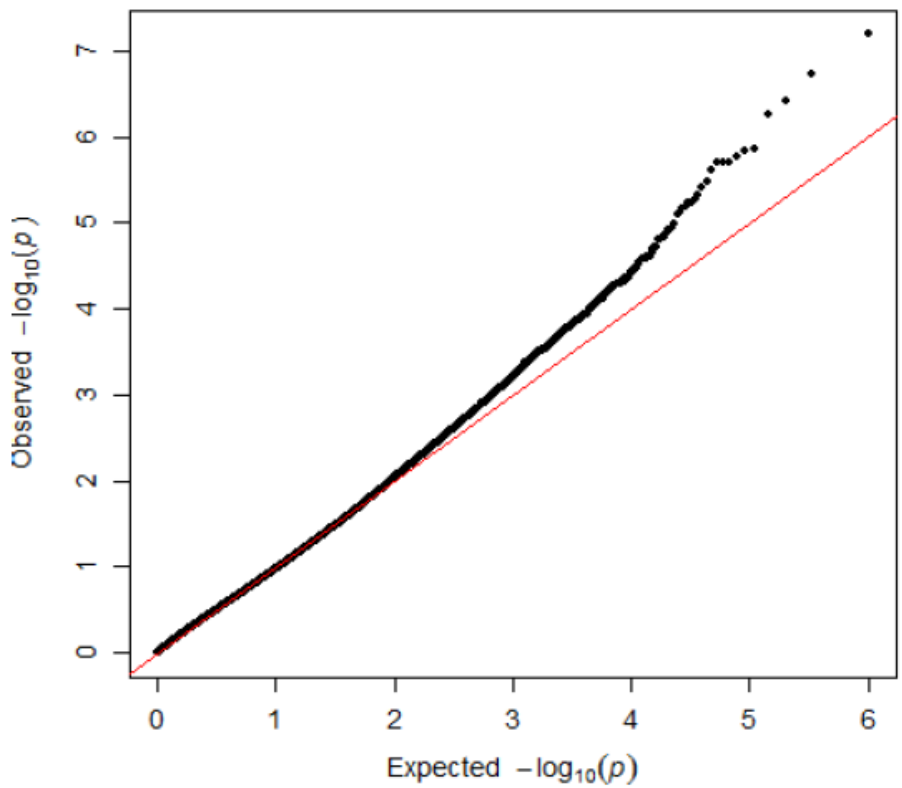
Covariance operator: $\Sigma_y(s,s') = \Sigma_\eta(s,s') + \Sigma_\varepsilon(s,s')$

$$\sqrt{n}\{\mathrm{vec}(\hat{B}(d) - B(d) - 0.5O(H^2)) : d \in D\} \xrightarrow{L} G(0, \Sigma_B(d,d'))$$

Zhu, Li, and Kong (2012). AOS

Tract-GWAS


Manhattan Plot
8
6
4
2
0
$-\log_{10}(p)$
1
3
5
7
9
11
13
15
17
19
21
23
Chromosome


0
1
2
3
4
5
6
7
Observed $-\log_{10}(p)$
0
1
2
3
4
5
6
Expected $-\log_{10}(p)$

Top SNPs

SNP ID	CHR ID	−log10(p-value)
'rs1384333'	10	7.3057
'rs2805816'	9	6.6177
'rs34608777'	8	6.2466
'2:170144362:T_TTA'	2	6.1259
'rs11263882'	1	5.8854
’ rs78723271’	15	5.7848
'rs11580558'	10	5.7236
'rs10841375'	12	5.6652
'rs11910115'	21	5.6525
'rs28839626'	7	5.5871

A Software for FVGWAS