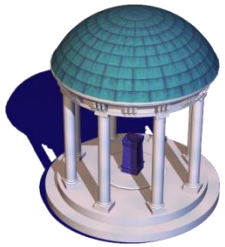
UNC SCHOOL OF PUBLIC HEALTH



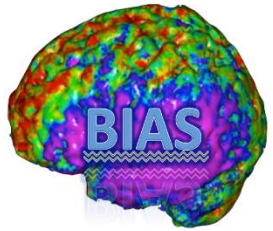
FFGWAS

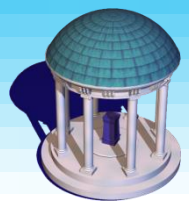
Fast Functional Genome Wide Association AnalysiS of Surface-based Imaging Genetic Data

Chao Huang

Department of Biostatistics
Biomedical Research Imaging Center
The University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA

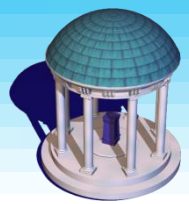
Joint work with Hongtu Zhu



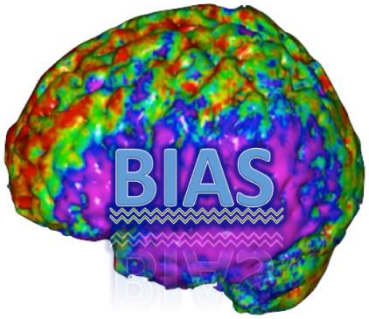


Outline

- Motivation for Imaging Genetics
- Statistical Methods for Imaging Genetics
- Fast Functional Genome-wise Association Analysis
- Conclusion

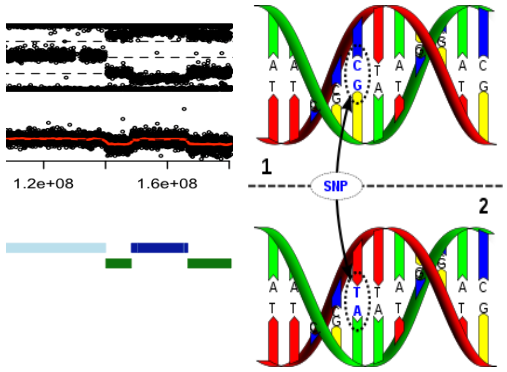


Motivation for **Imaging Genetics**

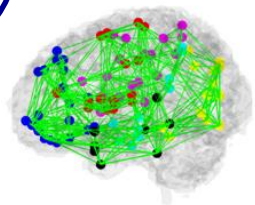


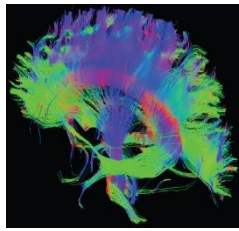
Imaging Genetics

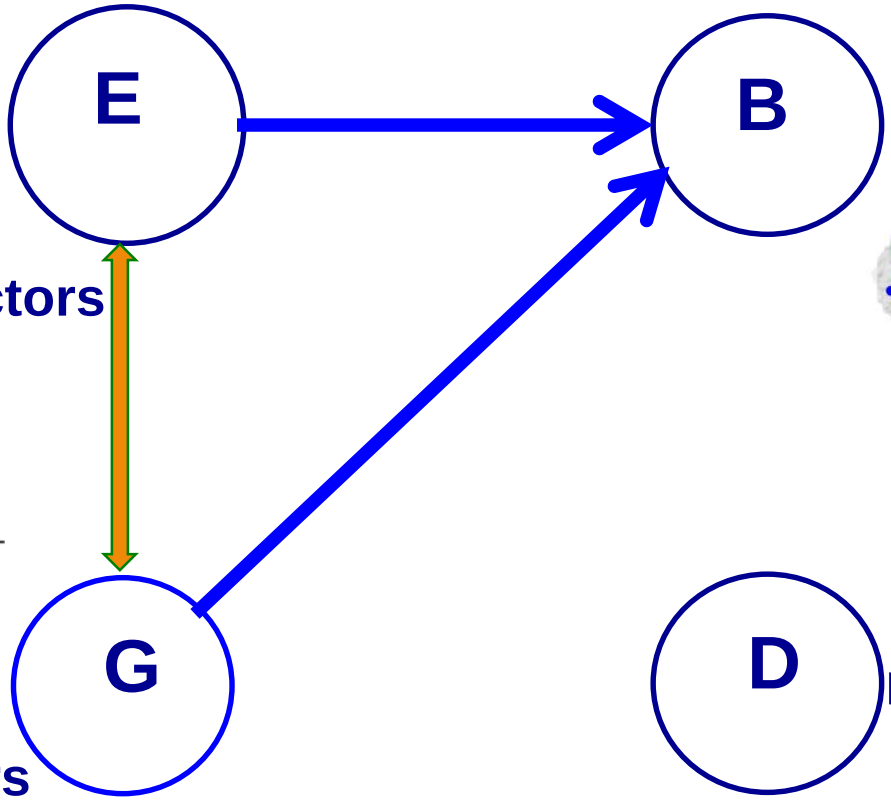
E: environmental factors

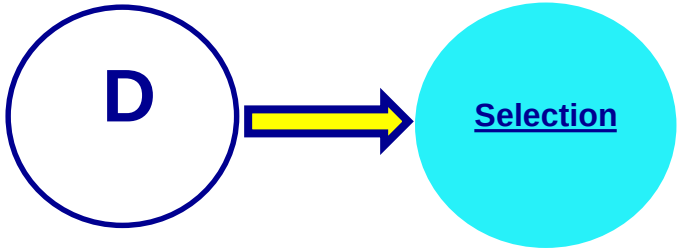


G: genetic markers



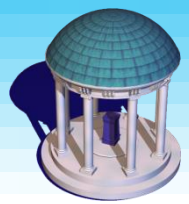






D: disease

http://en.wikipedia.org/wiki/DNA_sequence



Imaging Data

Structural MRI

Diffusion MRI

- Variety of acquisitions
- Measurement basics
- Limitations & artefacts
- Analysis principles
- Acquisition tips

Functional MRI (task)

Functional MRI (resting)

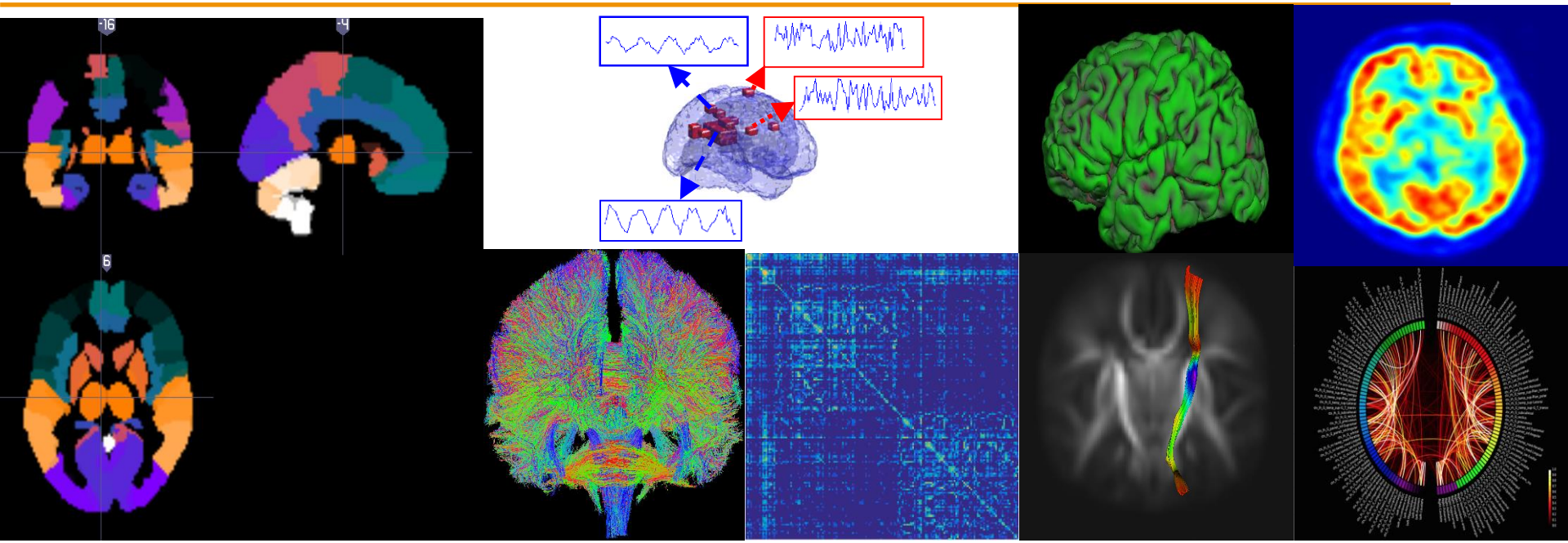
PET

EEG/MEG

CT

Calcium

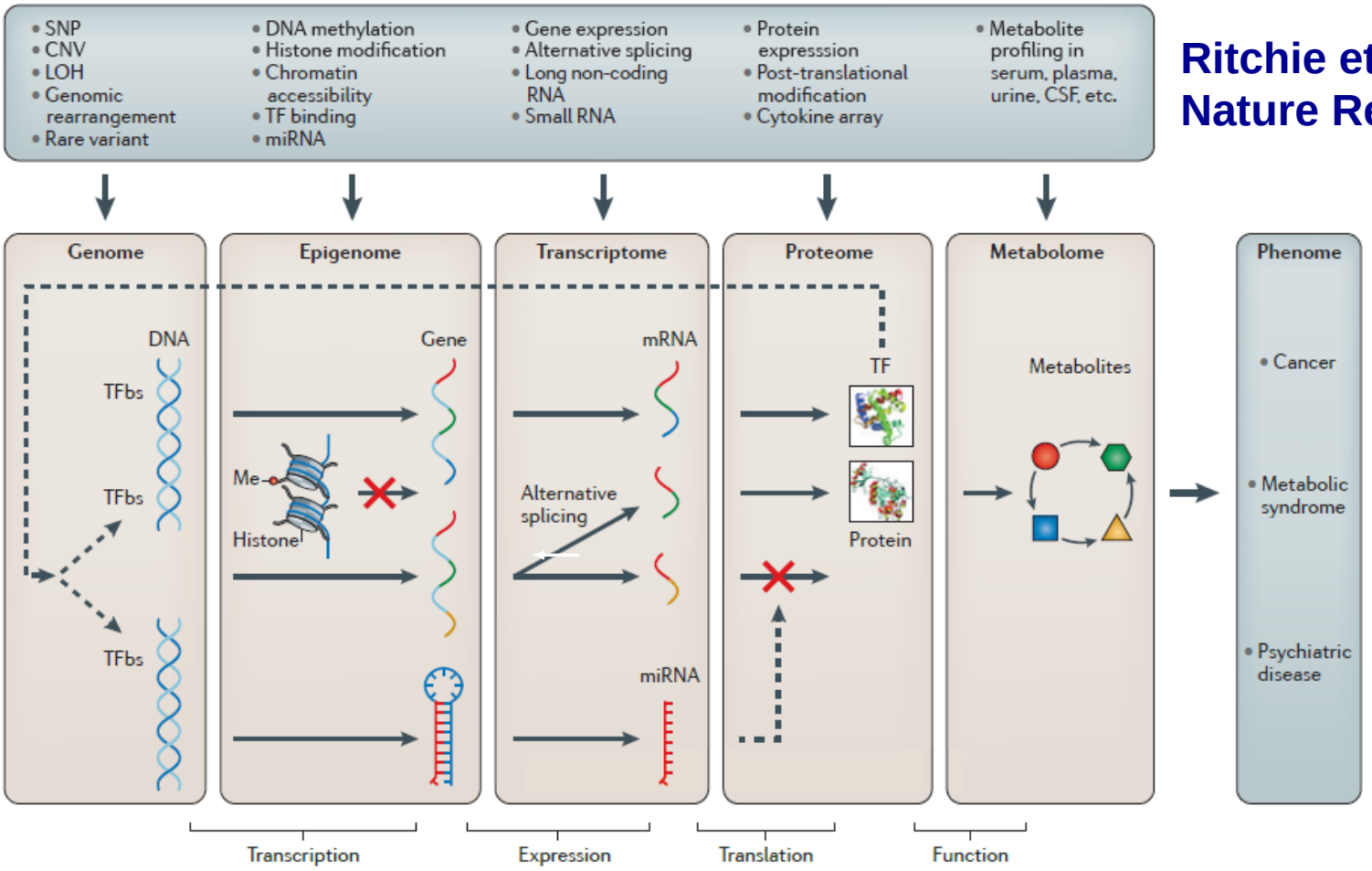
Neuroimaging Phenotype



Multivariate, smoothed functions, and piecewisely smoothed functions

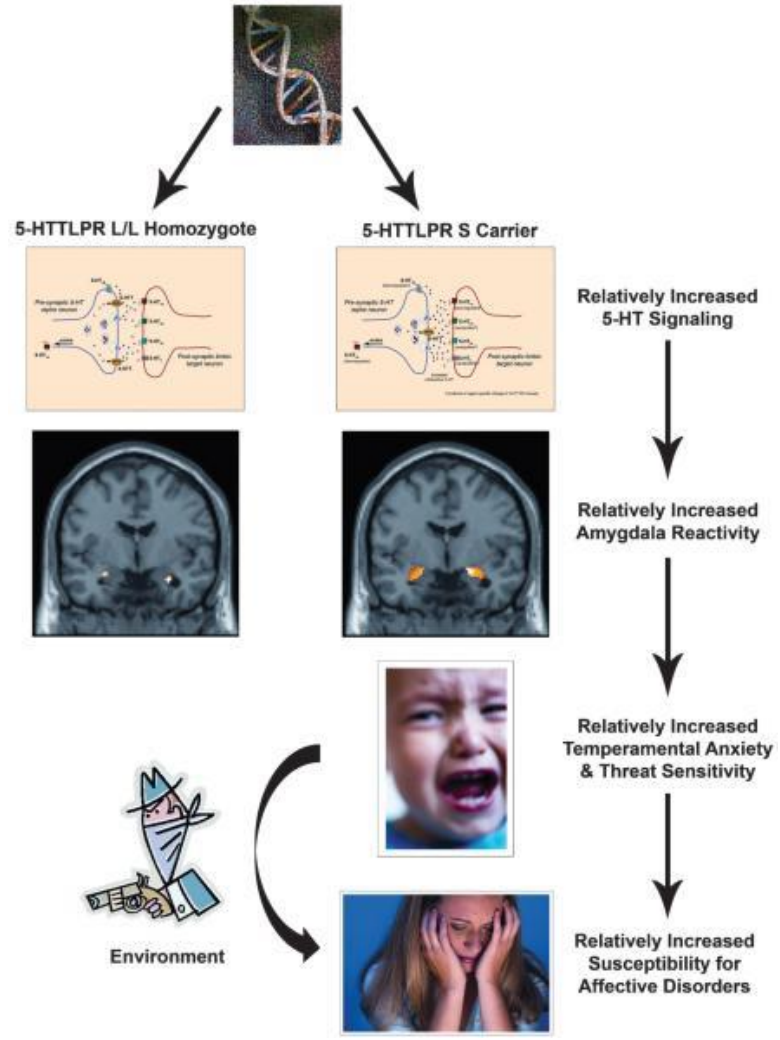
Dimension varies from 100~500,000.

Multi-Omic Data



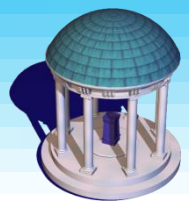
Ritchie et al. (2015). Nature Review Genetics

Motivation

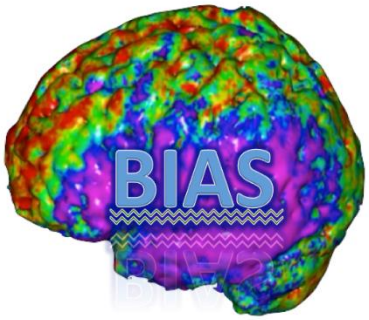


Imaging genetics allows for the identification of how common/rare genetic polymorphisms influencing molecular processes (e.g., serotonin signaling), bias neural pathways (e.g., amygdala reactivity), mediating individual differences in complex behavioral processes (e.g., trait anxiety) related to disease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



Statistical Methods for **Imaging Genetics**



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		

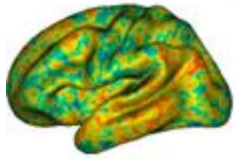
Hibar, et al. HBM 2012

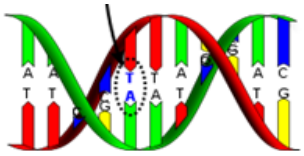
High Dimensional Regression Model

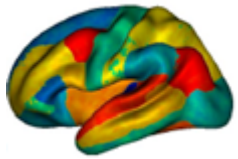
Data $\{(Y_i, X_i): i = 1, \cdots, n\}$

$$Y_i = \{y_i(v): v \hat{\mathrm{I}} V\} \quad X_i = \{X_i(g): g \hat{\mathrm{I}} G_0\}$$

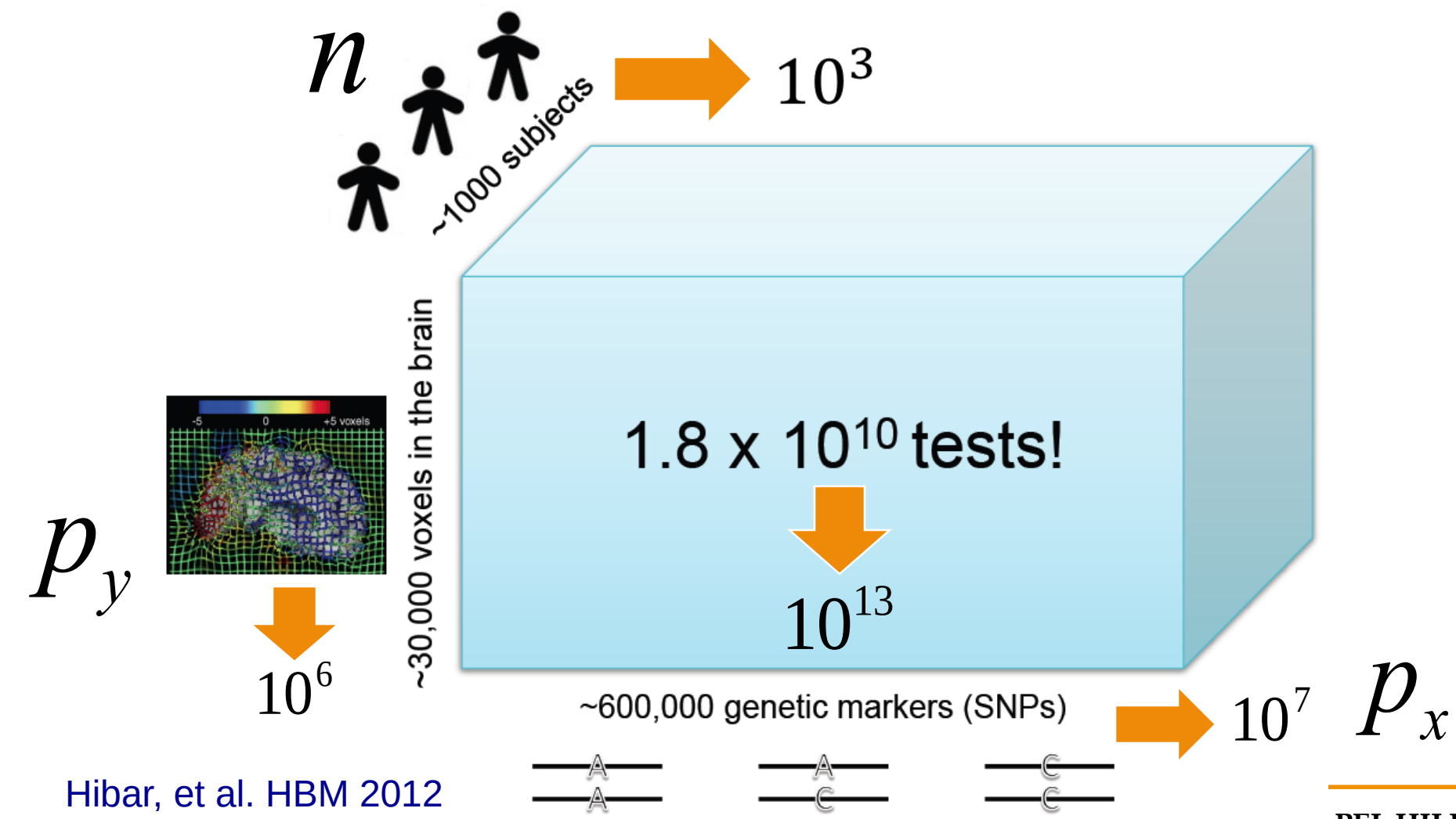
Phenotype		Genotype			Error
Y		X		B	E
$n \times p_y$	$=$	$n \times p_x$		$p_x \times p_y$	$+$ $n \times p_y$





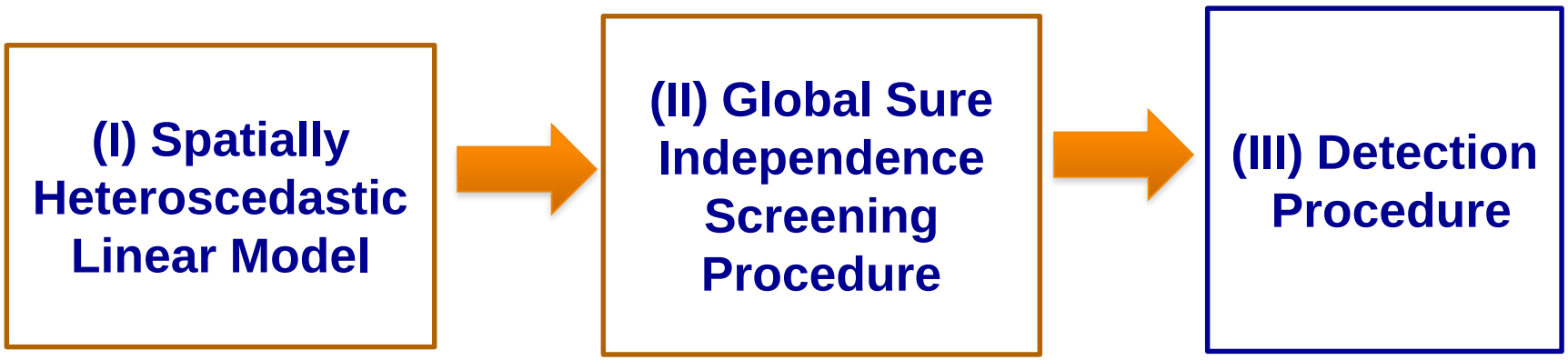


Challenges



Hibar, et al. HBM 2012

Fast Voxel-wise Genome-wise Analysis

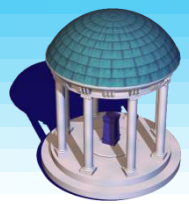


Huang, et al. Neuroimage 2015

Issues to be addressed:

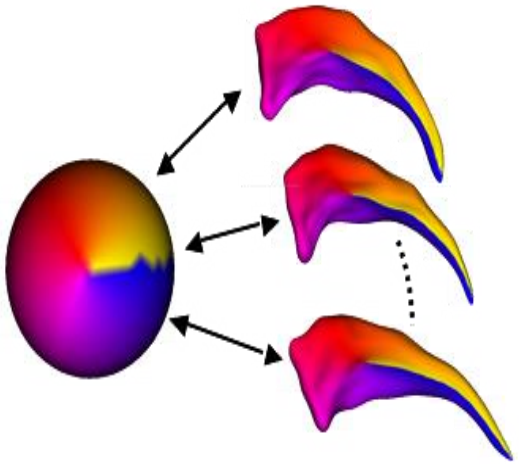
-- Spatially correlated functional data

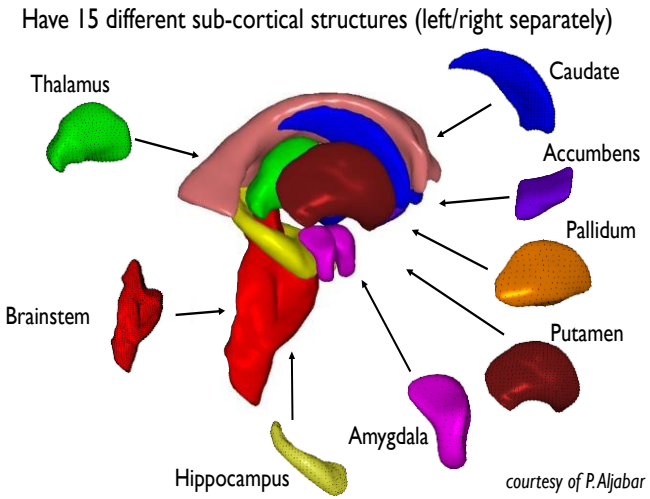
-- Multivariate imaging phenotypes

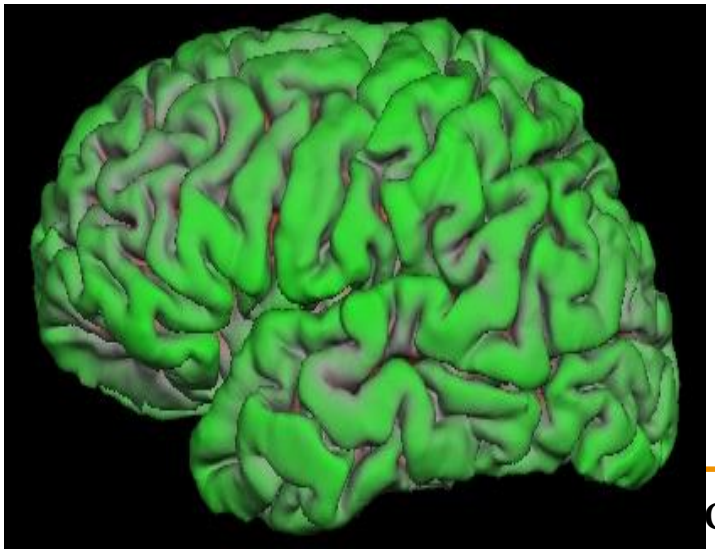


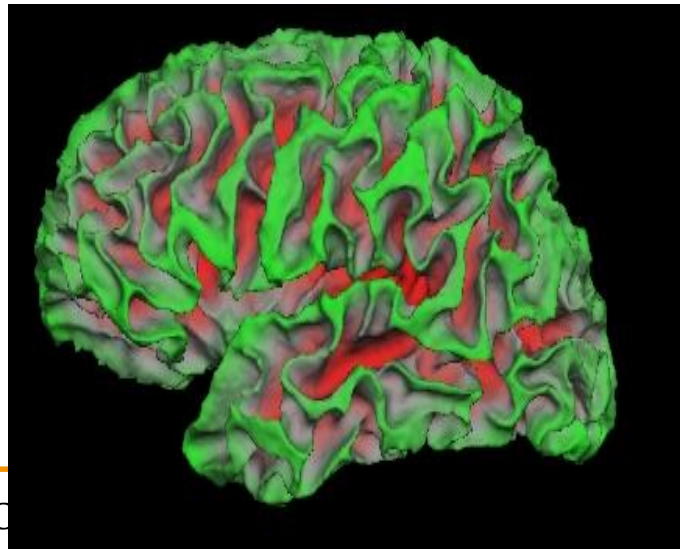
Fast Functional Genome-wise AnalysiS

Data Structure



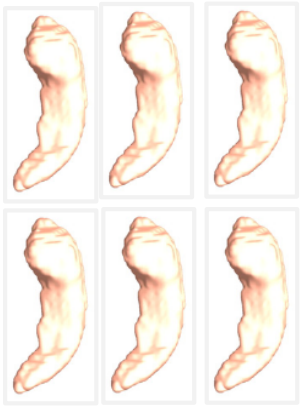




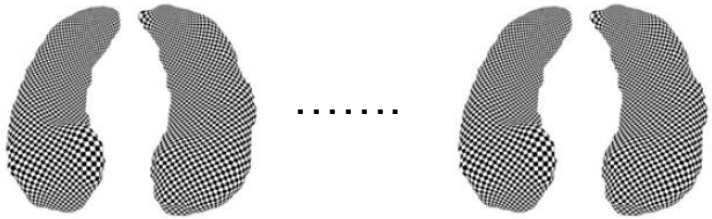


Data Structure

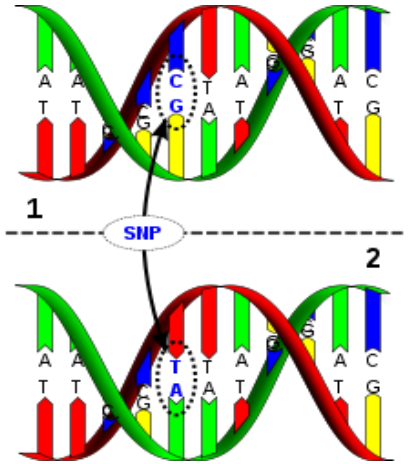
Hippocampal Surface



Person No.1 ..... Person No.100



Genetic Variation



SNP1 SNP2 ..... SNP2000

Person No. 1

Person No. 100

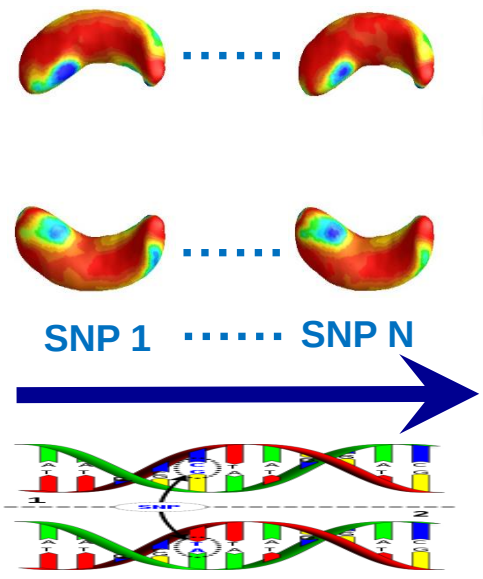
$$\begin{bmatrix} 1 & 2 & \cdots & 0 \\ 0 & \ddots & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 2 \end{bmatrix}$$

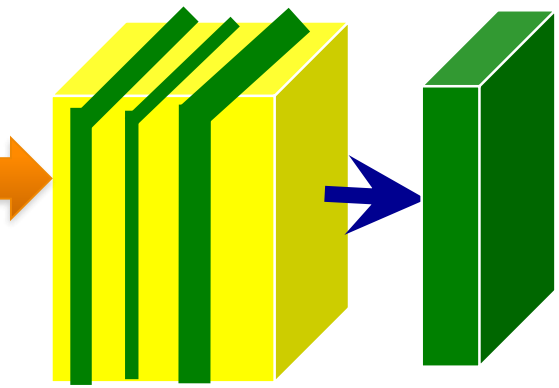
FFGWAS

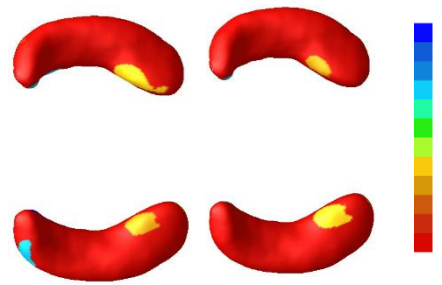
Multivariate Varying Coefficient Model

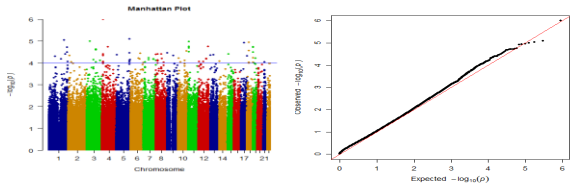
Global Sure Independence Screening Procedure

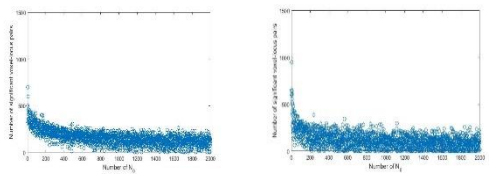
Significant Voxel-locus Cluster-locus Detection

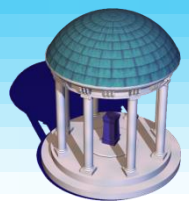








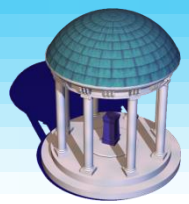




Multivariate Varying Coefficient Model

$$\boldsymbol{y_{i,j}(d) = x_i^T \beta}_{j}^{(c)}\boldsymbol{(d) + z_i(g)^T \beta}_{j}^{(g)}\boldsymbol{(d) + \eta_{i,j}(d) + \epsilon_{i,j}(d)}, i = 1, \dots, n, j = 1, \dots, J$$

where $\boldsymbol{\beta}_{j}^{(c)}\boldsymbol{(d)}$ is a $\boldsymbol{p_x \times 1}$ vector associated with non-genetic predictors (e.g., age, gender), and $\boldsymbol{\beta}_{j}^{(g)}\boldsymbol{(d)}$ is an $\boldsymbol{p_g \times 1}$ vector of genetic fixed effects (e.g., additive or dominant). Moreover, $\boldsymbol{\epsilon_i(d) = \left(\epsilon_{i,1}(d), \dots, \epsilon_{i,J}(d)\right)^T}$ are measurement errors and independent and identical copies of a stochastic process $\mathbf{SP(0,\Sigma_\epsilon)}$ and $\boldsymbol{\eta_i(d) = \left(\eta_{i,1}(d), \dots, \eta_{i,J}(d)\right)^T}$ are independent with $\boldsymbol{\epsilon_i(d)}$ and identical copies of a stochastic process $\mathbf{SP(0,\Sigma_\eta^{(g)})}$.



Multivariate Varying Coefficient Model

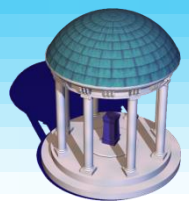
We need to test:

$$H_0 : \boldsymbol{\beta}^{(g)}(\boldsymbol{d}) = 0 \text{ for all } \boldsymbol{d} \quad \text{v.s.} \quad H_1 : \boldsymbol{\beta}^{(g)}(\boldsymbol{d}) \neq 0$$

We first consider a local Wald-type statistic as:

$$T_n(g, \boldsymbol{d}) = \boldsymbol{r}^{(g)}(\boldsymbol{d})^T \left[\hat{\boldsymbol{\Sigma}}_{\eta}^{(g)^{-1}}(\boldsymbol{d}, \boldsymbol{d}) \otimes [\sum_{i=1}^{n} \boldsymbol{z}_i(g)^{\otimes 2}]^{-1} \right] \boldsymbol{r}^{(g)}(\boldsymbol{d}),$$

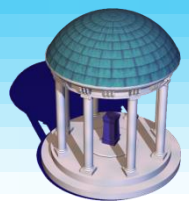
where $\boldsymbol{r}^{(g)}(\boldsymbol{d}) = \hat{\boldsymbol{\beta}}^{(g)}(\boldsymbol{d}) - \text{Bias}(\hat{\boldsymbol{\beta}}^{(g)}(\boldsymbol{d}))$



Big-data Challenges

Several big-data challenges arise from the calculation of $T_n(g, d)$ as follows.

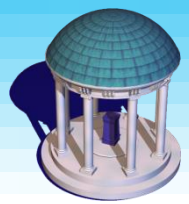
- Calculating $\hat{\Sigma}_{\eta}^{(g)}(d)$ across all loci and vertices can be computationally.
- Bandwidth selection in $T_n(g, d)$ across all loci can be also computationally.
- Holding all $T_n(g, d)$ in the computer hard drive requires substantial computer resources.
- Speeding up the calculation of $T_n(g, d)$.



FFGWAS

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

- **Calculate $\hat{\Sigma}_{\eta}^{(g)}(d)$ under the null hypothesis H_0 for all lo*ci*.**
- **Divide all loci into K groups based on their minor allele frequency (MAF), and select a common optimal bandwidth for each group.**
- **Develop a GSIS procedure to eliminate many 'noisy' loci based on a global Wald-type statistic.**



A Global Sure Independence Screening

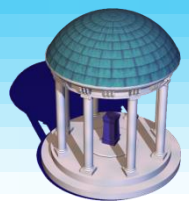
(1) The global Wald-type statistic at locus *g* is defined as

$$T_n(g) = \frac{1}{M} tr\Bigg\{ \mathrm{Vec}(\widetilde{\boldsymbol{X}})^{\otimes 2}\Big[\sum_{m=1}^{M} \boldsymbol{Y}_w(\boldsymbol{d}_m)\hat{\Sigma}_{\eta}^{-1}(\boldsymbol{d}_m)\boldsymbol{Y}_w^T(\boldsymbol{d}_m)\Big] \otimes \Big[(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)^{-1}[0_{p_g\times p_c}\ \boldsymbol{I}_{p_g}]^T[\sum_{i=1}^{n} \boldsymbol{z}_i(g)^{\otimes 2}]^{-1}[0_{p_g\times p_c}\ \boldsymbol{I}_{p_g}](\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^T)^{-1}\Big] \Bigg\}$$

(2) Calculate the p-values of $T_n(g)$ for all loci

(3) Sort the -log 10(p)-values of $T_n(g)$ and select the top N0 loci

The candidate significant locus set $\widetilde{\mathcal{G}}_0 = \{\widetilde{g}_1, \cdots, \widetilde{g}_{N_0}\}$



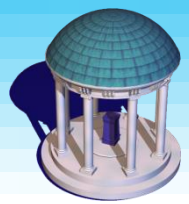
Detection Procedure

(1) The first one is to detect significant voxel-locus pairs

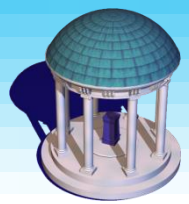
$$T_n(\widetilde{g},d) = tr\Big\{\mathrm{Vec}(\widetilde{X})^{\otimes 2}\Big[\boldsymbol{Y}_w(d)\hat{\Sigma}_\eta^{-1}(d)\boldsymbol{Y}_w^T(d)\Big]\otimes \Big[(\widetilde{X}\widetilde{X}^T)^{-1}[0_{p_g\times p_c}\ \boldsymbol{I}_{p_g}]^T[\sum_{i=1}^{n}\boldsymbol{z}_i(g)^{\otimes 2}]^{-1}[0_{p_g\times p_c}\ \boldsymbol{I}_{p_g}](\widetilde{X}\widetilde{X}^T)^{-1}\Big]\Big\}$$

(2) The second one is to detect significant cluster-locus pairs.

Wild Bootstrap method



Simulation Studies and Real Data Analysis



Simulation Studies: Data Generation

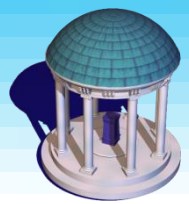
Covariate Data (non-genetic data)

Generated from either U(0,1) or the Bernoulli distribution with success probability 0.5.

Genetic Data

Linkage Disequilibrium (LD) blocks (Haploview & PLINK)

1. **Generate 2,000 blocks;**
2. **Randomly select 10 SNPs in each block;**
3. **Chose the first 100 SNPs as the causal SNPs**



Simulation Studies: Data Generation

Imaging Data

Step 1: Fitting the model without genetic predictors

$$y_{i,j}(d) = x_i^T \beta_j^{(c)}(d) + z_i(g)^T \beta_j^{(g)}(d) + \eta_{i,j}(d) + \epsilon_{i,j}(d), i = 1, \ldots, n, j = 1, \ldots, J$$

→ Estimates of $\beta_j^{(c)}(d)$ Σ_ϵ $\Sigma_\eta^{(g)}$ → True values

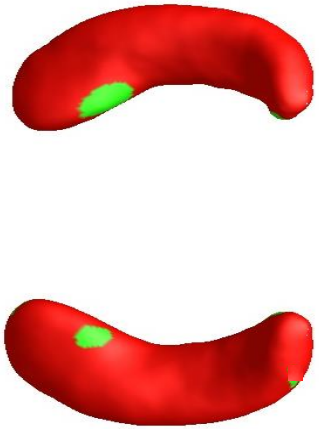
Step 2: Specifying effected Regions Of Interest associated with causal SNPs

→ $$\beta_j^{(g)}(d) = \begin{cases} r, & \forall\, d \in \text{ROIs} \\ 0, & \text{otherwise} \end{cases}$$

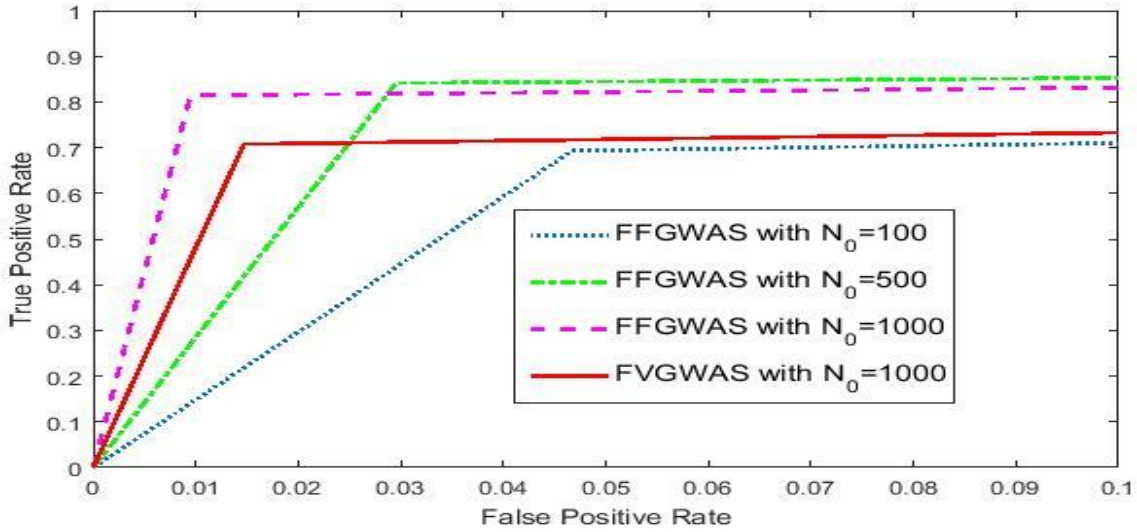
Step 3: Generating imaging data with prespecified parameters and ROIs

Simulation Studies

Simulation settings: the green and red regions in the figure, respectively, represent Hippocampal surface, and the effected ROI associated with the causal SNPs among first 20000 SNPs.



$$\boldsymbol{\beta}_j^{(g)}(\boldsymbol{d}) = 0.001$$



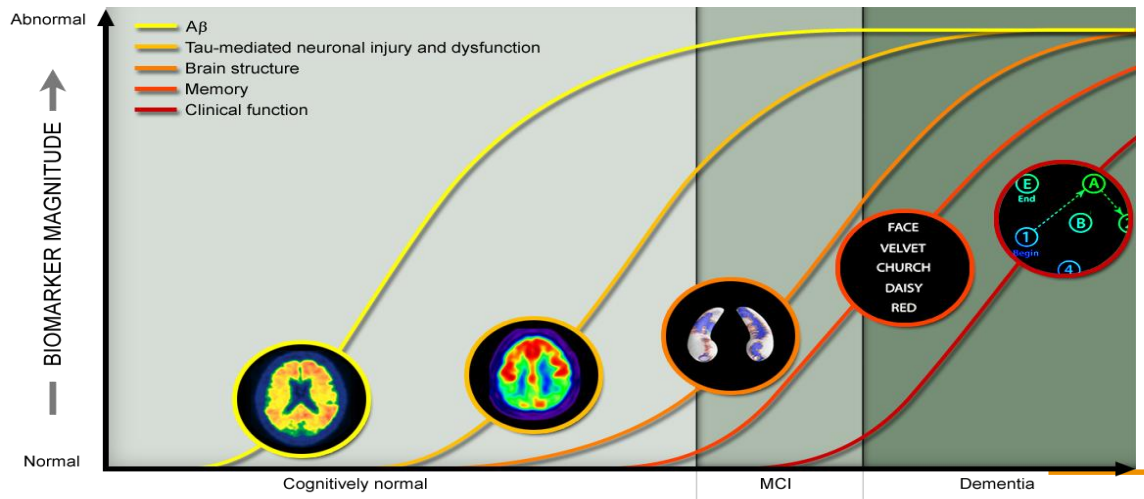
Simulation results for comparisons between FFGWAS and FVGWAS in identifying significant voxel-SNP pairs.

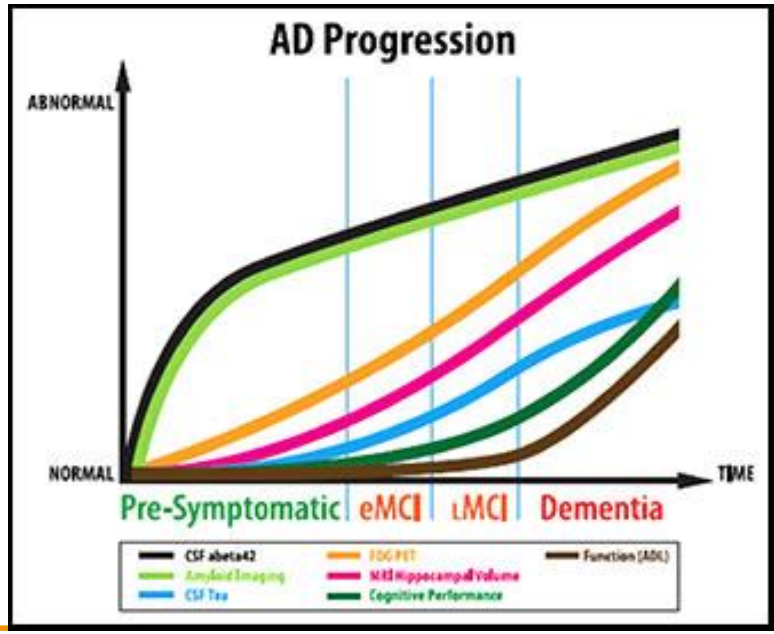
Imaging Genetics for ADNI

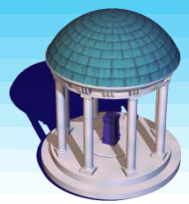
PI: Dr. Michael W. Weiner

- detecting AD at the earliest stage and marking its progress through biomarkers;
- developing new diagnostic methods for AD intervention, prevention, and treatment.

- A longitudinal prospective study with 1700 aged between 55 to 90 years
- Clinical Data including Clinical and Cognitive Assessments
- Genetic Data including Ilumina SNP genotyping and WGS
- MRI (fMRI, DTI, T1, T2)
- PET (PIB, Florbetapir PET and FDG-PET)
- Chemical Biomarker



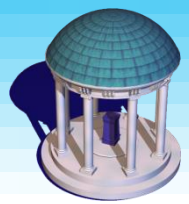




ADNI Data Analysis: Dataset Description

- 708 MRI scans of AD (186), MCI (388), and healthy controls (224) from ADNI-1.
- These scans on 462 males and 336 females are performed on a 1.5 T MRI scanners.
- The typical protocol includes the following parameters:
  (i) repetition time (TR) = 2400 ms;
  (ii) inversion time (TI) = 1000 ms;
  (iii) flip angle = 8^{o};
  (iv) field of view (FOV) = 24 cm with a 256 x 256 x 170 acquisition matrix in the x−, y−, and z−dimensions,
  (v) voxel size: 1.25 x 1.26 x 1.2 mm^{3}.

- Covariates: gender, age, APOE ε4, and the top 5 PC scores in SNPs



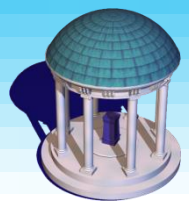
Imaging Data Preprocessing

Surface fluid registration based hippocampal sub-regional analysis package (Shi et al., Neuroimage, 2013)

- **Hippocampal surface registration**
  isothermal coordinates and fluid registration

- **Surface statistics computation**
  1. **multivariate tensor-based morphometry (mTBM) statistics**
  2. **radial distance**

Finally, we obtained left and right hippocampus shape representations as 100×150 matrices.



ADNI Data Analysis

Top 10 SNPs (Left Hippocampus)

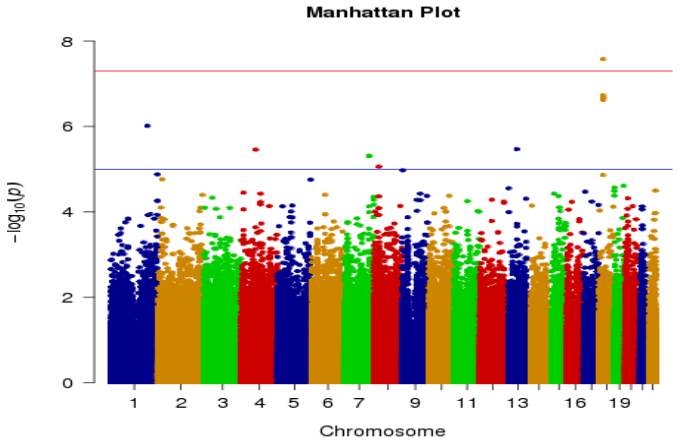
SNP	CHR	BP	-LOG 10(p)
rs657132	18	2.20533e+07	7.579767
rs604345	18	2.20033e+07	6.729377
rs582110	18	2.19954e+07	6.672876
rs546000	18	2.20031e+07	6.672876
rs489631	18	2.1989e+07	6.620395
rs16837577	1	1.94871e+08	6.016773
rs3812872	13	6.19869e+07	5.468391
rs6826085	4	7.68702e+07	5.459163
rs929714	7	1.3263e+08	5.314317
rs2042067	7	1.32651e+08	5.306583

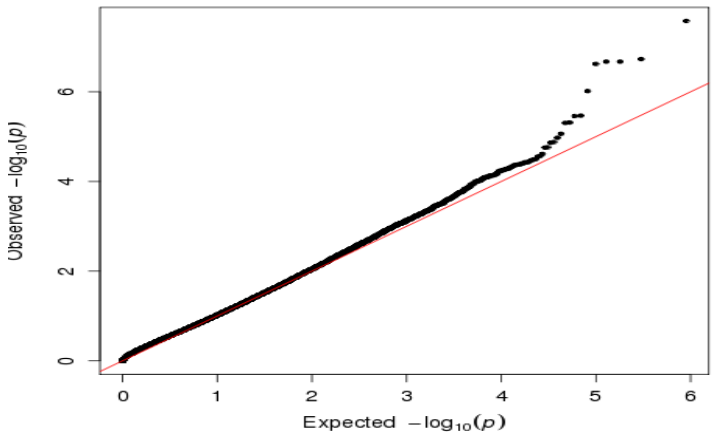
Top 10 SNPs (Right Hippocampus)

SNP	CHR	BP	-LOG 10(p)
rs4681527	3	1.44e+08	6.764886
rs3108514	2	1.51279e+08	6.274511
rs12264728	10	1.3214e+08	5.961976
rs652911	10	1.3214e+08	5.739661
rs10801705	1	8.95004e+07	5.622668
rs366346	10	1.32141e+08	5.617185
rs7312068	12	2.94352e+07	5.604041
rs7617465	3	1.43999e+08	5.522112
rs17605251	7	1.02746e+08	5.486603
rs749788	2	2.84618e+06	5.474675

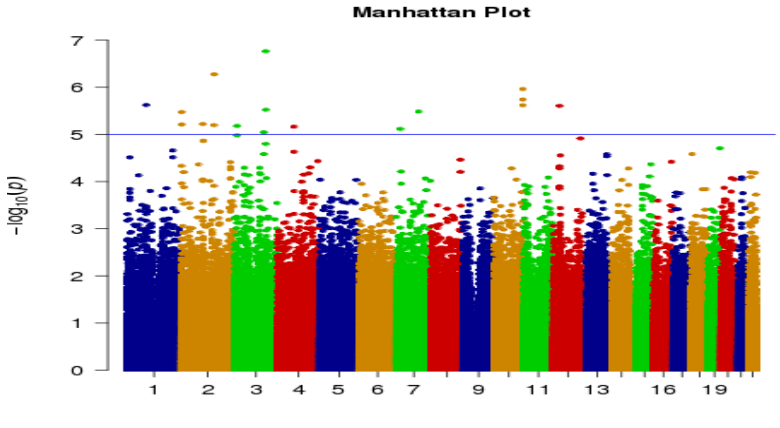
ADNI Data Analysis

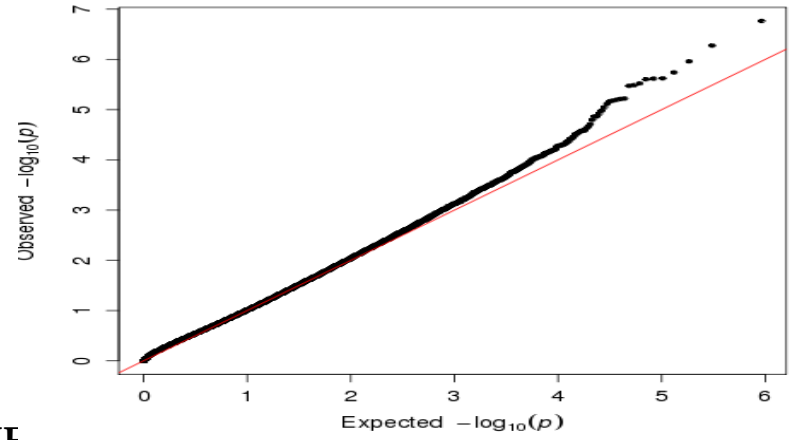
(Left Hippocampus)





(Right Hippocampus)

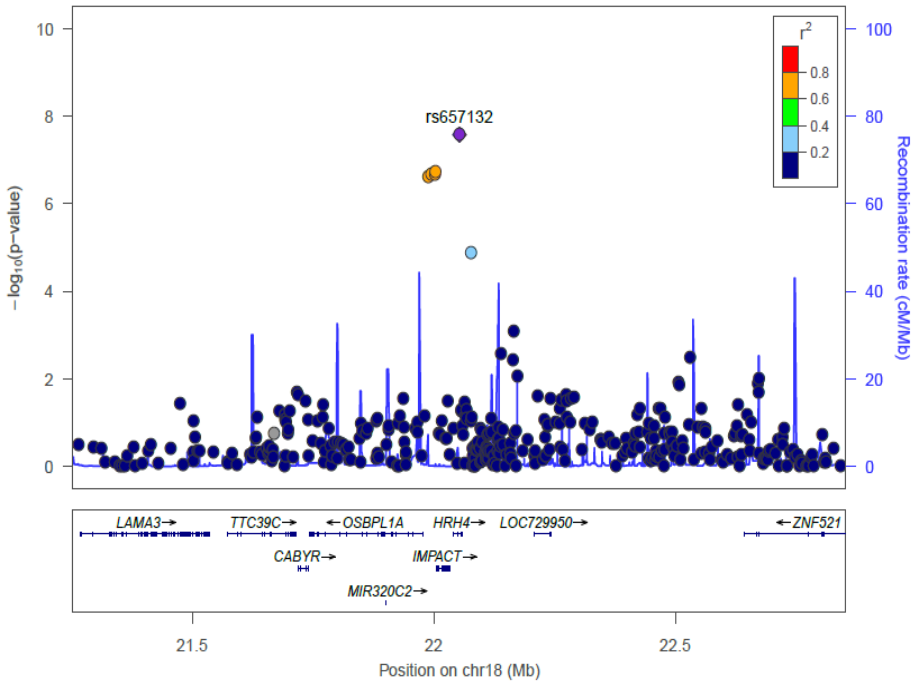




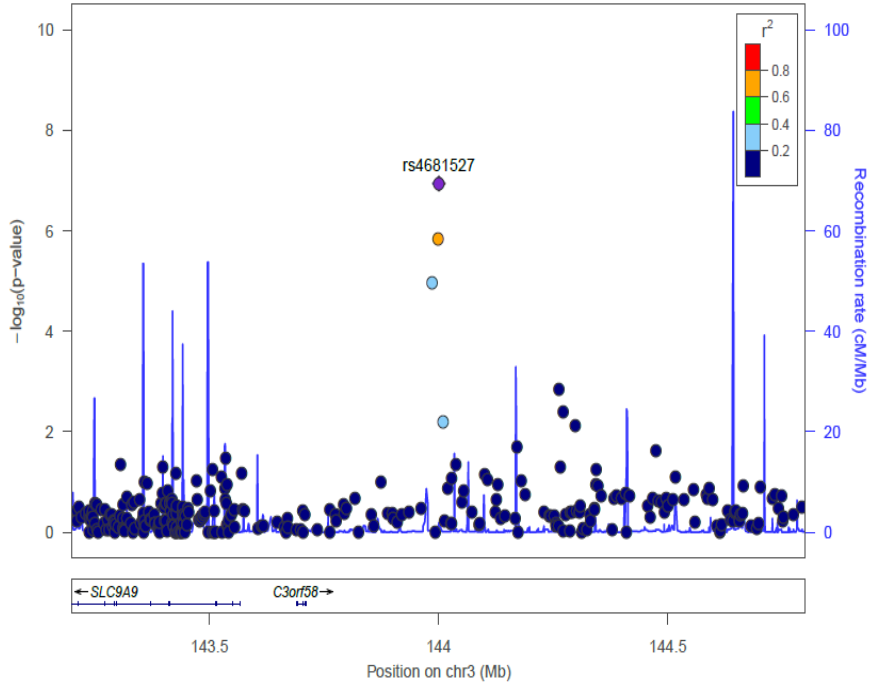
ADNI Data Analysis: Left Hippocampus

Significant Loci Zoom

(Left Hippocampus)



(Right Hippocampus)



ADNI Data Analysis: Left Hippocampus

(Left Hippocampus)

Top 1 SNP: rs657132
Closed Gene: HRH4

HRH4 (Histamine Receptor H4) is a Protein Coding gene.
Diseases associated with HRH4: cerebellar degeneration
An important paralog of this gene: CHRM4

Mirshafiey & Naddafi, Am J Alzheimers Dis Other Demen. 2013

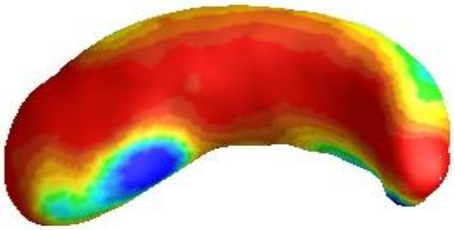
(Right Hippocampus)

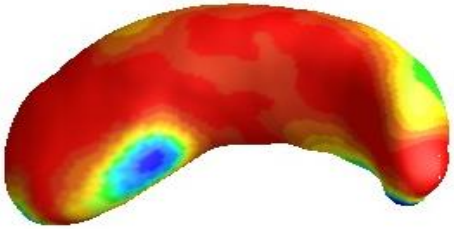
Top 1 SNP: rs4681527
Closed Gene: C3orf58

C3orf58 (Chromosome 3 Open Reading Frame 58) is a Protein Coding gene.
Diseases associated with C3orf58: hypoxia

ADNI Data Analysis

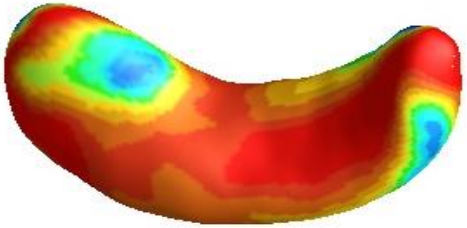
Left

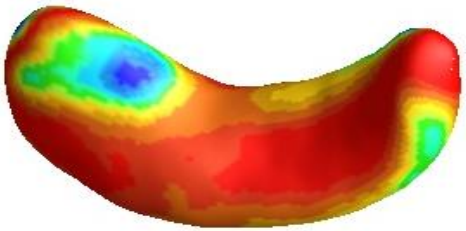




rs657132 rs604345

Right





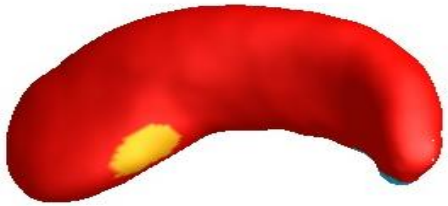
rs4681527 rs3108514

$-\log_{10}(p)$ values on Hippocampus (L & R) corresponding to Top 2 SNPs

ADNI Data Analysis

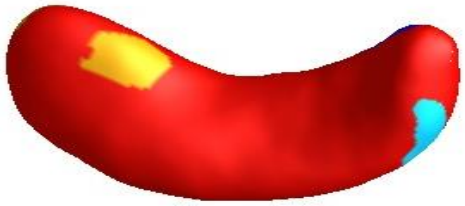
Left

rs657132

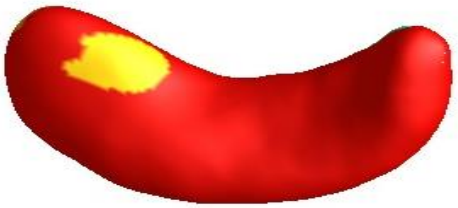


rs604345

Right

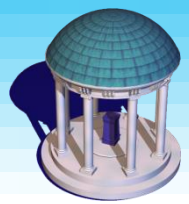


rs4681527



rs3108514

$-\log_{10}(p)$ values of significant clusters on Hippocampus (L & R) corresponding to Top 2 SNPs



Conclusion

- **We have developed a FFGWAS pipeline for efficiently carrying out whole-genome analyses of multimodal imaging data.**
- **Our FFGWAS consists of a multivariate varying coefficient model, a global sure independence screening (GSIS) procedure, and a detection procedure based on wild bootstrap methods.**
- **Two key advantages of using FFGWAS include**

  (i) Much smaller computational complexity;

  (ii) GSIS for screening many noisy SNPs.
- **We have successfully applied FFGWAS to hippocampal surface data & genetic data of ADNI study.**

A Software for FFGWAS