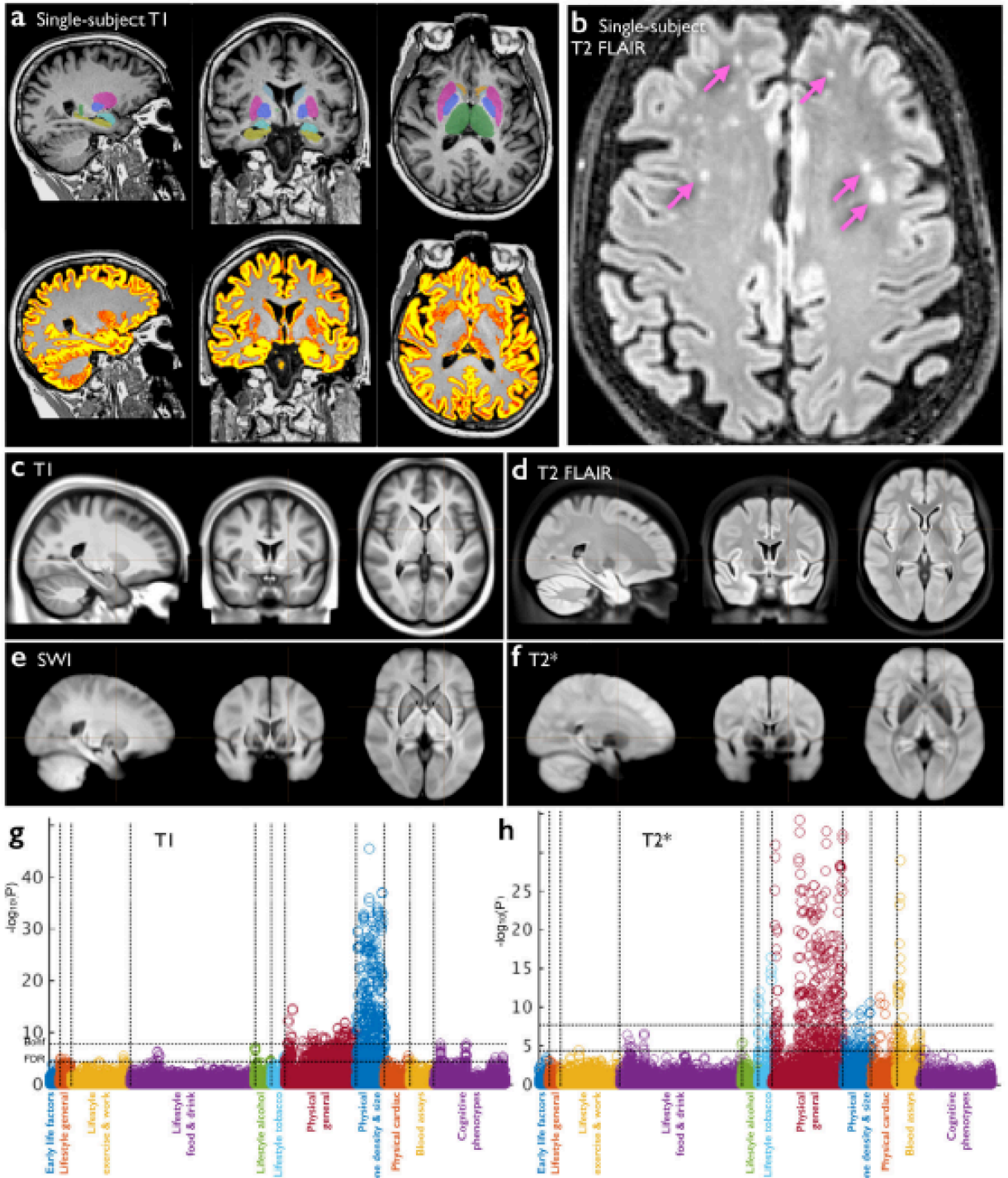
Phenotypes, Genotypes & Voxels: A Playground Next to a Nuclear Power Plant

Bryan Guillaume, Anqi Qiu
Department of Biomedical Engineering
Clinical Imaging Research Centre
National University of Singapore

Associations between Voxels and Multiple Phenotypes



UK BioBank

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J*, et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 2016; 19(11): 1523-36.

Associations between Voxels and Genotypes

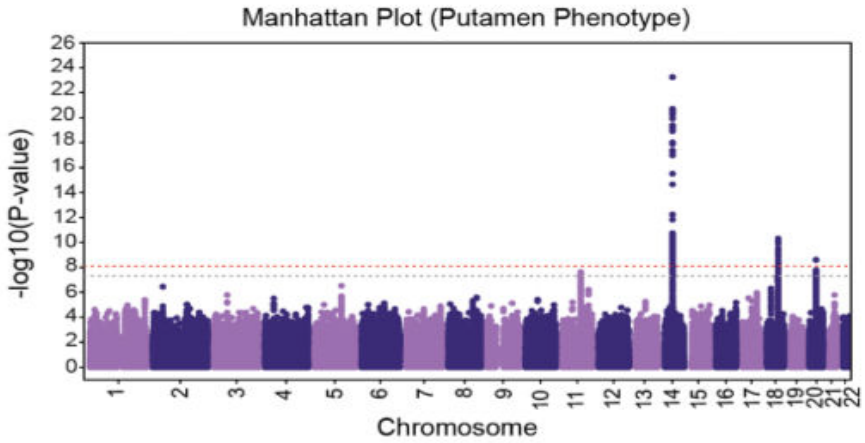
Subcortical Volumes

Individual SNPs

Whole Brain

Individual SNPs

GWAS with Structural Volumes



Hibar DP, *et al.,* "Common genetic variants influence human subcortical brain structures." *Nature* 2015.

GWAS to Brain Image Voxels

Statistical approaches to deal with correction for a large number of statistical tests.

Huang M, *et al.*, "Fast voxelwise genome wide association analysis of large-scale imaging genetic data", *Neuroimage*, 2015.

NUS National University of Singapore

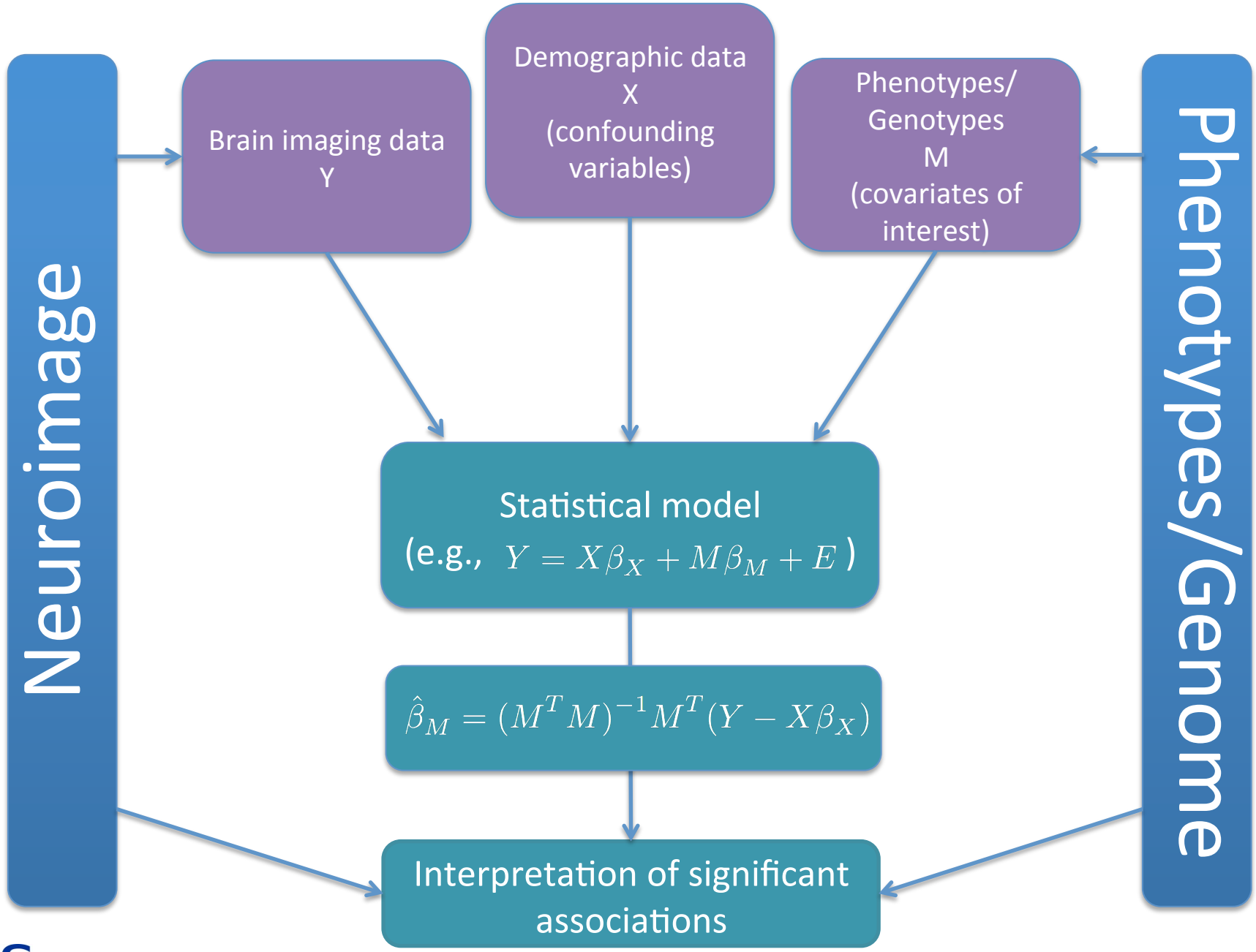
How to Examine Massive Associations?

"Currently, the workhorse tools of neuroimaging are "mass univariate", where a regression model is fit separately at each voxel."

Smith SM, Nichols TE. (2018): Statistical Challenges in "Big Data" Human Neuroimaging. Neuron 97(2):263-268.

NUS National University of Singapore

Mass-Univariate Analysis via Regression



MASS-UNIVARIATE ANALYSIS VIA REGRESSION

$$Y = X\beta_X + M\beta_M + E$$

Y - $n_{\text{subj}} \times n_{\text{vox}}$ matrix, neuroimaging data
X - $n_{\text{subj}} \times n_X$ matrix of known covariates of nondirect interest
M - $n_{\text{subj}} \times n_M$ matrix of known covariates of interest
E - $n_{\text{subj}} \times n_{\text{vox}}$ matrix of random errors

$$\hat{\beta}_M = (M^T M)^{-1} M^T (Y - X\beta_X)$$

Assume that E is homoskedastic, that is, $Var(E) = \sigma^2 I$.

$$Var(\hat{\beta}_M) = \sigma^2 (M^T M)^{-1}$$

Statistical Non-parametric Mapping (SnPM) Toolbox --- Thomas Nichols

"Big Sample Sizes: Big Confound Trouble."

Smith SM, Nichols TE. (2018): Statistical Challenges in "Big Data" Human Neuroimaging. Neuron 97(2):263-268.

NUS National University of Singapore

Potential Issues in Linear Regression

$$Y = X\beta_X + M\beta_M + E$$

$$\hat{\beta}_M = (M^T M)^{-1} M^T (Y - X\beta_X)$$

$$Var(\hat{\beta}_M) = \sigma^2 (M^T M)^{-1}$$

→ Inflate residual
→ Inflate variance estimates
→ Deflate statistics
→ Make hypothesis tests conservative

$Cor(M, Z)$
Z are missing covariates

Positive or negative bias for the association between M and Y.

PROBLEM STATEMENT

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

- → Inflate residual
- → Inflate variance estimates
- → Deflate statistics
- → Make hypothesis tests conservative

Positive or negative bias for the association between M and Y.

1. How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$;
2. How to do statistical tests;
3. What is suitable to neuroimaging data with spatial correlation in noise.

How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

$$QR = [X\ M]$$

$$Q^\top Y = Q^\top X\beta_X + Q^\top M\beta_M + Q^\top Z\beta_Z + Q^\top E$$

$$\begin{aligned} Q_X^\top Y &= Q_X^\top X\beta_X + Q_X^\top M\beta_M + Q_X^\top Z\beta_Z + Q_X^\top E \\ Q_M^\top Y &= Q_M^\top M\beta_M + Q_M^\top Z\beta_Z + Q_M^\top E \\ Q_Z^\top Y &= Q_Z^\top Z\beta_Z + Q_Z^\top E \end{aligned}$$

How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

$$QR = [X\ M]$$

$$Q^\top Y = Q^\top X\beta_X + Q^\top M\beta_M + Q^\top Z\beta_Z + Q^\top E$$

$$Q_Z^\top Y = Q_Z^\top Z\beta_Z + Q_Z^\top E$$

Factor analysis ideally accounting for

1) Spatial heteroskedastic errors
2) Spatial correlations

Estimating n_Z, Z, β_Z and Σ

$$Q_Z^\top Y = Q_Z^\top Z\beta_Z + Q_Z^\top E$$

Factor analysis ideally accounting for

1) Spatial heteroskedastic errors
2) Spatial correlations

PCA

$$Q_Z^\top Y = USV^\top$$

Eigenvalue Ratio (ER)

$$\hat{n}_Z = \underset{0 \le k \le k_{\max}}{\operatorname{argmax}} \frac{\lambda_k}{\lambda_{k+1}}$$

$$Q_Z^\top Z = \sqrt{n_{\text{subj}} - n_X - n_M}\, U_Z$$
$$\beta_Z = S_Z V_Z^\top / \sqrt{n_{\text{subj}} - n_X - n_M}$$

How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$

$$
\begin{aligned}
Q_X^\top Y &= Q_X^\top X\beta_X + Q_X^\top M\beta_M + Q_X^\top Z\beta_Z + Q_X^\top E \\
Q_M^\top Y &= Q_M^\top M\beta_M + Q_M^\top Z\beta_Z + Q_M^\top E \\
Q_Z^\top Y &= Q_Z^\top Z\beta_Z + Q_Z^\top E
\end{aligned}
$$

$$Z = X\alpha_X + M\alpha_M + W$$

$$Q_M^\top Y = Q_M^\top M\beta_M + Q_M^\top M\alpha_M^* \hat{\beta}_Z + Q_M^\top E$$

Robust regression or

Sparse regression

How to solve $\beta_X, \beta_M, \beta_Z, Z, n_z, Var(E)$

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

$$Q^\top Y = Q^\top X\beta_X + Q^\top M\beta_M + Q^\top Z\beta_Z + Q^\top E$$

$$Q_X^\top Y = Q_X^\top X\beta_X + Q_X^\top M\beta_M + Q_X^\top Z\beta_Z + Q_X^\top E$$

$$Q_M^\top Y = Q_M^\top M\beta_M + Q_M^\top Z\beta_Z + Q_M^\top E$$

$$Q_Z^\top Y = Q_Z^\top Z\beta_Z + Q_Z^\top E$$

Not useful

Can be ignored

Factor analysis ideally accounting for

1) Spatial heteroskedastic errors
2) Spatial correlations

Robust regression or

Sparse regression

PARAMETRIC HYPOTHESIS TESTING

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

Wald score:

$$T_v = \frac{n_{\text{subj}}}{q\hat{\sigma}_v^2} C\hat{\beta}_{Mv}(C(\hat{\Omega}_M + \hat{\alpha}_M^* \hat{\alpha}_M^{*\top})C^\top)^{-1}(C\hat{\beta}_{Mv})^\top \ ,$$

where C is a contrast matrix of rank q. T_v asymptotically follows a χ^2-distribution with q degrees of freedom under the null hypothesis.

MASS-UNIVARIATE ANALYSIS VIA REGRESSION

$$Y = X\beta_X + M\beta_M + E$$

Y - $n_{\text{subj}} \times n_{\text{vox}}$ matrix, neuroimaging data
X - $n_{\text{subj}} \times n_X$ matrix of known covariates of nondirect interest
M - $n_{\text{subj}} \times n_M$ matrix of known covariates of interest
E - $n_{\text{subj}} \times n_{\text{vox}}$ matrix of random errors

$$\hat{\beta}_M = (M^T M)^{-1} M^T (Y - X\beta_X)$$

Assume that E is homoskedastic, that is, $Var(E) = \sigma^2 I$.

$$Var(\hat{\beta}_M) = \sigma^2 (M^T M)^{-1}$$

Statistical Non-parametric Mapping (SnPM) Toolbox --- Thomas Nichols

NON-PARAMETRIC HYPOTHESIS TESTING

$$
\begin{aligned}
Q_X^\top Y &= Q_X^\top X\beta_X + Q_X^\top M\beta_M + Q_X^\top Z\beta_Z + Q_X^\top E \\
Q_M^\top Y &= Q_M^\top M\beta_M + Q_M^\top Z\beta_Z + Q_M^\top E \\
Q_Z^\top Y &= Q_Z^\top Z\beta_Z + Q_Z^\top E
\end{aligned}
$$

$$
\begin{pmatrix} Q_M^\top Y_b \\ Q_Z^\top Y_b \end{pmatrix} = \begin{pmatrix} Q_M^\top M\hat{\alpha}_M^* \hat{\beta}_Z \\ Q_Z^\top \hat{Z}\hat{\beta}_Z \end{pmatrix} + F_b^R \begin{pmatrix} Q_M^\top \hat{E}^* \\ Q_Z^\top \hat{E}^* \end{pmatrix}
$$

b - the b^{th} resample

F_b^R - $(n_{\text{subj}} - n_X) \times (n_{\text{subj}} - n_X)$ resampling matrix under

1) Permutation resampling scheme;

2) Wild Bootstrap resampling scheme, $diag(F_b^R)$ or $diag(F_b^U)$ follows a distribution with at least mean zero and variance one.

NON-PARAMETRIC HYPOTHESIS TESTING

$$
\begin{aligned}
Q_X^\top Y &= Q_X^\top X\beta_X + Q_X^\top M\beta_M + Q_X^\top Z\beta_Z + Q_X^\top E \\
Q_M^\top Y &= Q_M^\top M\beta_M + Q_M^\top Z\beta_Z + Q_M^\top E \\
Q_Z^\top Y &= Q_Z^\top Z\beta_Z + Q_Z^\top E
\end{aligned}
$$

$$
\begin{pmatrix} Q_M^\top Y_b \\ Q_Z^\top Y_b \end{pmatrix} = \begin{pmatrix} Q_M^\top M\hat{\alpha}_M^*\hat{\beta}_Z \\ Q_Z^\top \hat{Z}\hat{\beta}_Z \end{pmatrix} + F_b^R \begin{pmatrix} Q_M^\top \hat{E}^* \\ Q_Z^\top \hat{E}^* \end{pmatrix}
$$

or

$$
\begin{pmatrix} Q_M^\top Y_b \\ Q_Z^\top Y_b \end{pmatrix} = \begin{pmatrix} Q_M^\top M\hat{\alpha}_M^*\hat{\beta}_Z \\ Q_Z^\top \hat{Z}\hat{\beta}_Z \end{pmatrix} + \begin{pmatrix} Q_M^\top \\ Q_Z^\top \end{pmatrix} F_b^U \hat{E}^*
$$

b - the b^{th} resample

F_b^R, F_b^U - $(n_{\text{subj}} - n_X) \times (n_{\text{subj}} - n_X)$ resampling matrix under

1) Permutation resampling scheme;

2) Wild Bootstrap resampling scheme, $diag(F_b^R)$ or $diag(F_b^U)$ follows a distribution with at least mean zero and variance one.

NON-PARAMETRIC HYPOTHESIS TESTING

a Family-Wise Error Rate (FWER) corrected p-value at each voxel:

$$\frac{1}{(n_{\mathrm{B}}+1)}\sum_{b=0}^{n_{\mathrm{B}}}\mathcal{I}(T_b^{max}\geq T_{v0}),$$

where T_b^{max} is the maximum score observed for the b^{th} resampled data.

Main Goals of the Simulations

1. To show the adverse effects of not modeling unknown covariates.
2. To show that non-parametric tests can improve upon parametric tests.

Simulation Setup

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

- $n_M = 1$, $n_Z = 2$
- $Z = Z^* + M\alpha_M$, where $Cov([Z^*\ M]) = I$
- $n_{subj} = 114$ and $n_{vox} = 100$ or $n_{vox} = 500$
- β_X, $\beta_Z \sim \mathcal{N}(0, I)$
- $\beta_M = 0$ to study the control of the False Positive Rate (FPR)
- 16 noise distributions

Main Criteria: observed FPR at a 5% level of significance

NOISE DISTRIBUTIONS

Label	Spatial correlation	Spatial heteroske-dasticity	Subject heteroske-dasticity	Distribution for E_t
noCorr_hom	✗	✗	✗	$\mathcal{N}(0, I)$
noCorr_het	✗	✓	✗	$\mathcal{N}(0, D^{\text{het}})$
lowCorr_hom	✓(low)	✗	✗	$\mathcal{N}(0, C_{\text{low}})$
lowCorr_het	✓(low)	✓	✗	$\mathcal{N}(0, C_{\text{low}}^{1/2} D^{\text{het}} C_{\text{low}}^{1/2})$
highCorr_hom	✓(high)	✗	✗	$\mathcal{N}(0, C_{\text{high}})$
highCorr_het	✓(high)	✓	✗	$\mathcal{N}(0, C_{\text{high}}^{1/2} D^{\text{het}} C_{\text{high}}^{1/2})$
mixCorr_hom	✓(mixture)	✗	✗	$\mathcal{N}(0, C_{\text{mix}})$
mixCorr_het	✓(mixture)	✓	✗	$\mathcal{N}(0, C_{\text{mix}}^{1/2} D^{\text{het}} C_{\text{mix}}^{1/2})$
het1_hom	✗	✗	✓	$\mathcal{N}(0, 0.4I)$ for odd i, $\mathcal{N}(0, 1.6I)$ otherwise
het1_het	✗	✓	✓	$\mathcal{N}(0, 0.4D^{\text{het}}))$ for odd i, $\mathcal{N}(0, 1.6D^{\text{het}}))$ otherwise
het2_hom	✗	✗	✓	$\mathcal{N}(0, 1.6I)$ for odd i, $\mathcal{N}(0, 0.4I)$ otherwise
het2_het	✗	✓	✓	$\mathcal{N}(0, 1.6D^{\text{het}}))$ for odd i, $\mathcal{N}(0, 0.4D^{\text{het}}))$ otherwise

Label	Spatial correlation	Spatial heteroske-dasticity	Subject heteroske-dasticity	Distribution for E_{tv}
exp_hom	✗	✗	✗	Exponential($\lambda = 1$)
exp_het	✗	✓	✗	Exponential($\lambda = 1/\sqrt{D_{vv}^{\text{het}}}$)
unif_hom	✗	✗	✗	$\mathcal{U}(-\sqrt{3}, \sqrt{3})$
unif_het	✗	✓	✗	$\mathcal{U}(-\sqrt{3D_{vv}^{\text{het}}}, \sqrt{3D_{vv}^{\text{het}}})$

noCorr, lowCorr, highCorr, mixCorr – normal distribution with no, low, high or mix of spatial correlation;

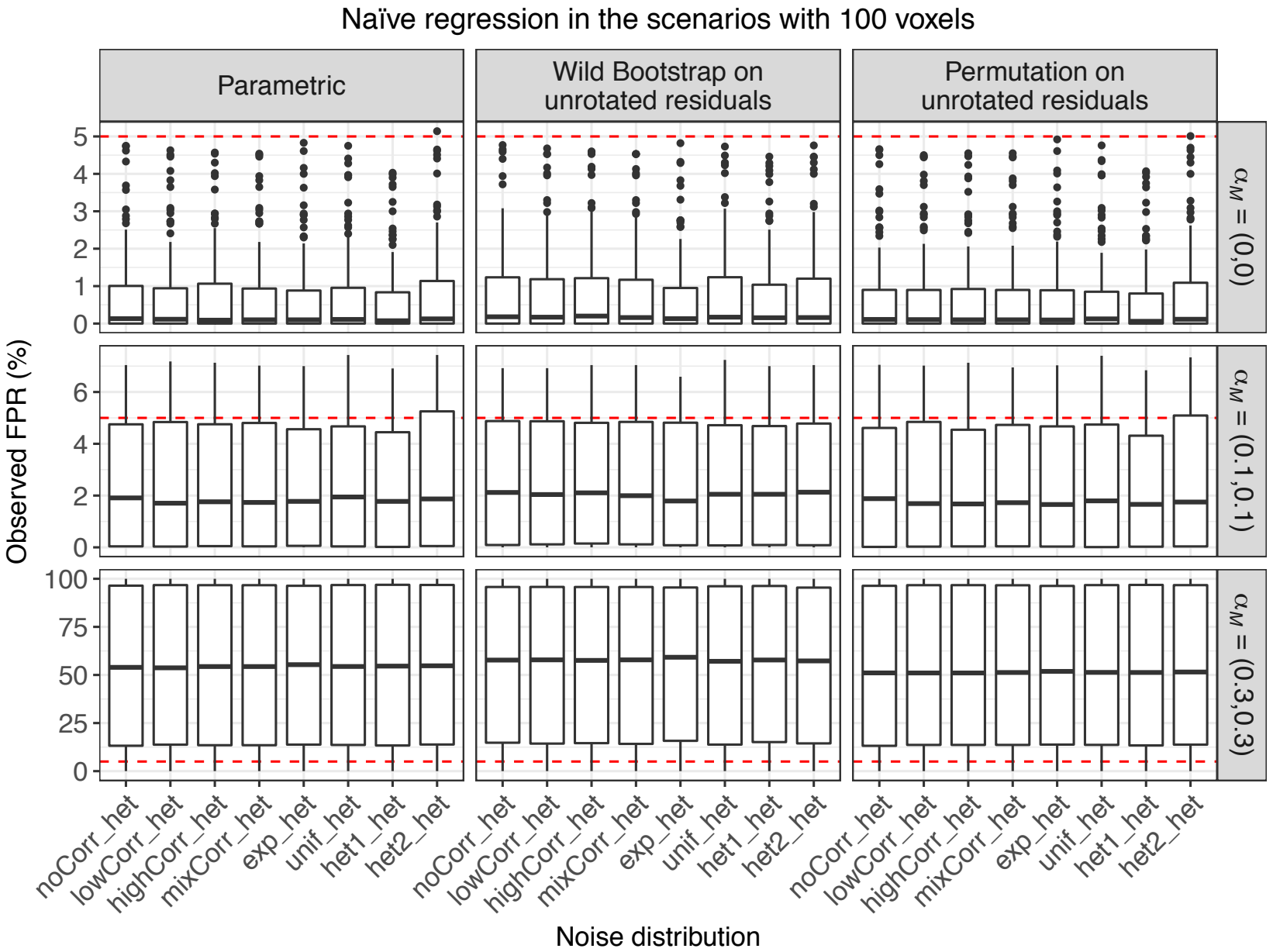
het1, het2 - two different normal distributions with no spatial correlation and subject heteroskedasticity;

exp - an exponential distribution with no spatial correlation;

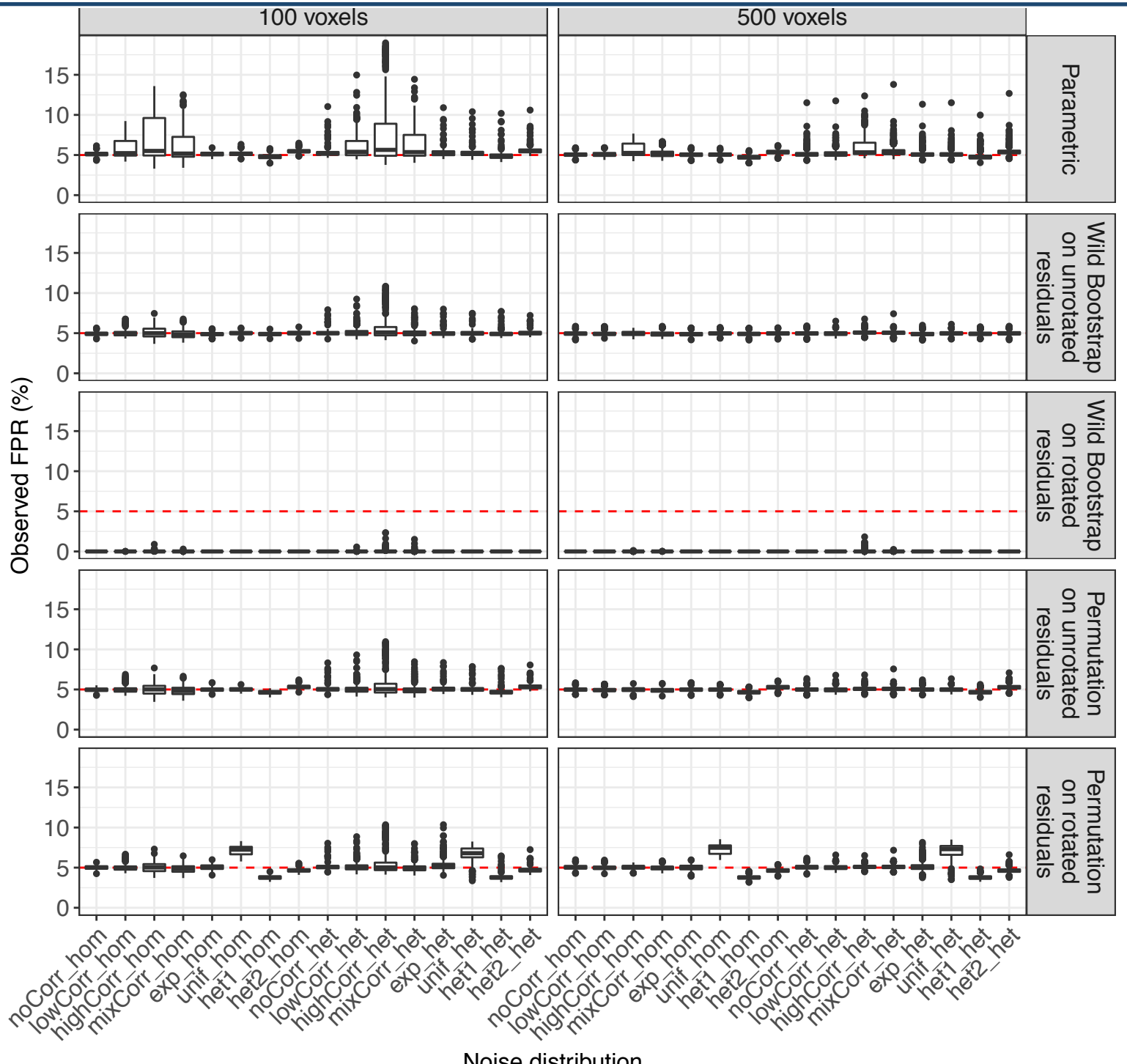
unif - a uniform distribution with no spatial correlation;

_hom, _het -spatial homoskedasticity or heteroskedasticity.

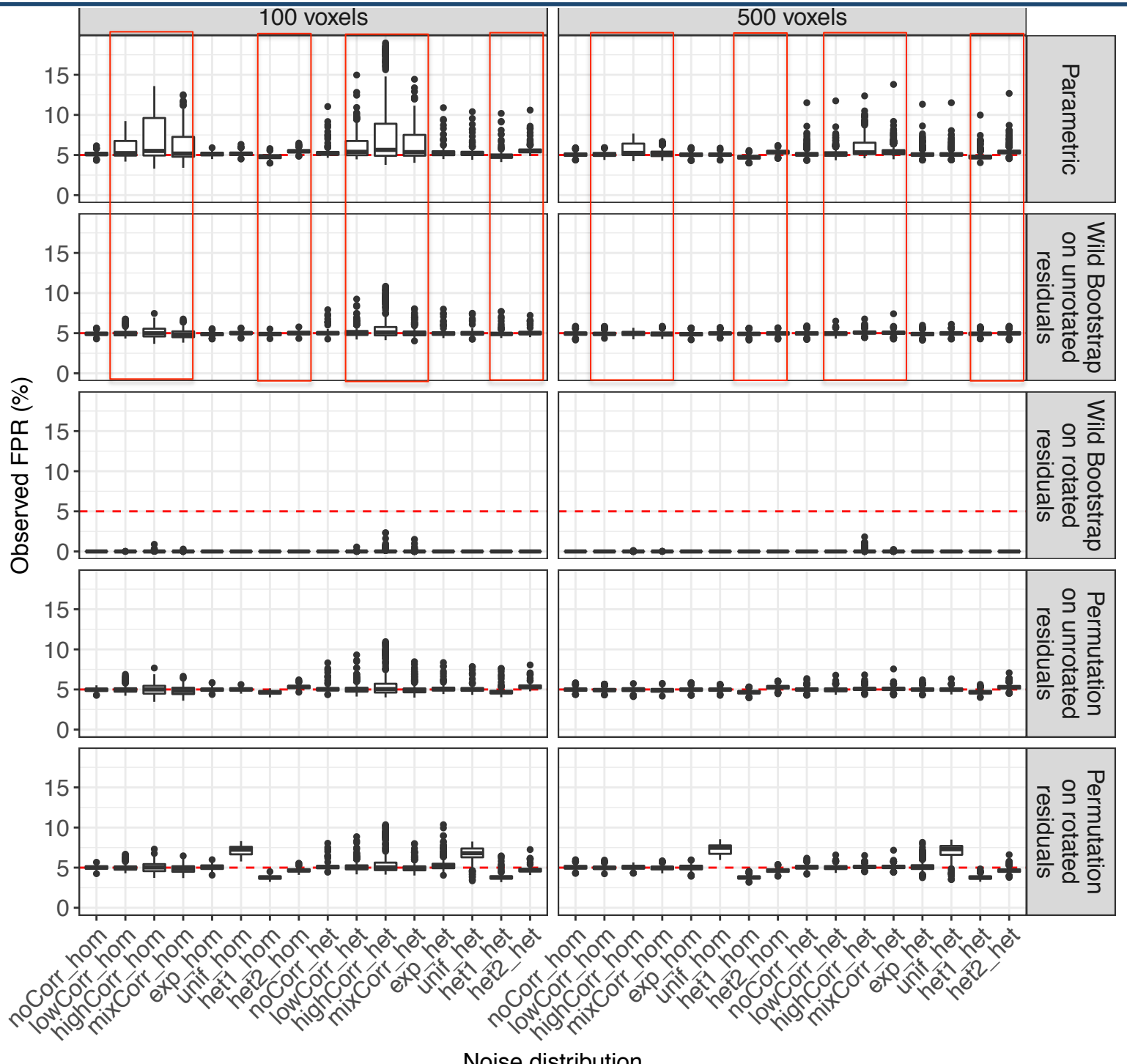
IMPORTANCE OF MODELING UNKNOWN COVARIATES



COMPARISON BETWEEN HYPOTHESIS TESTING PROCEDURES

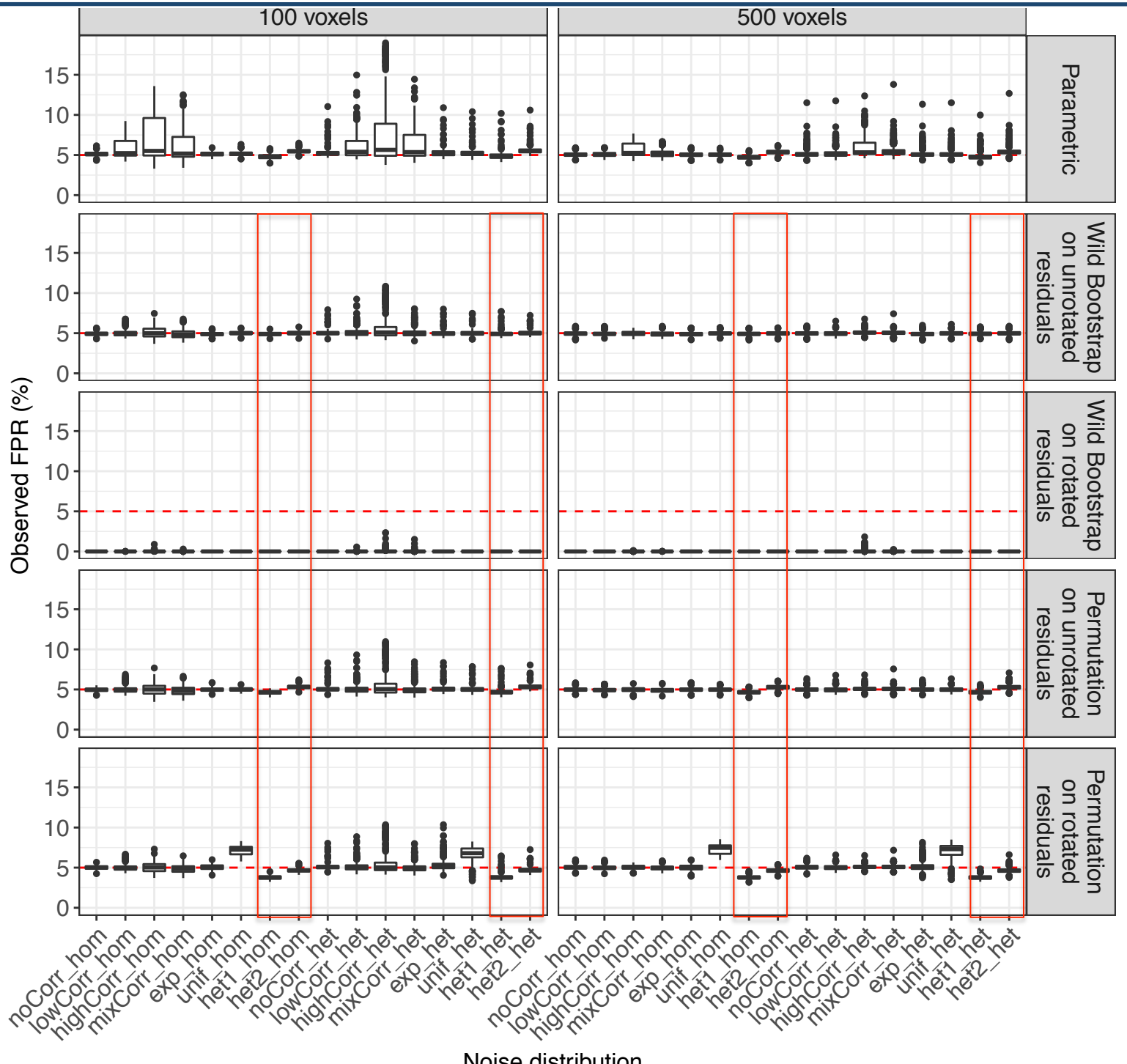


COMPARISON BETWEEN HYPOTHESIS TESTING PROCEDURES



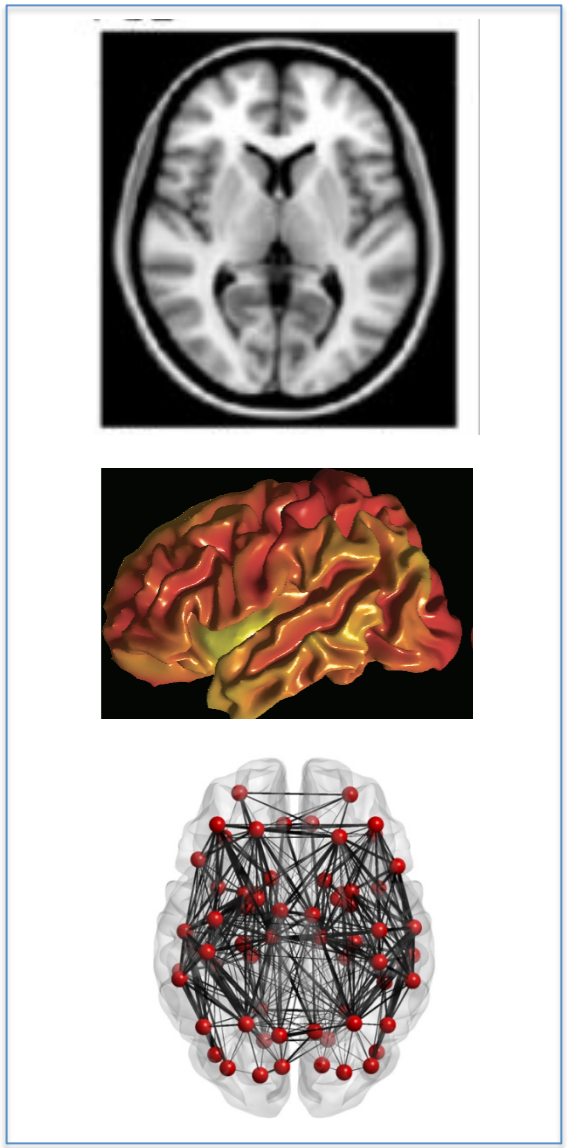
NUS National University of Singapore

COMPARISON BETWEEN HYPOTHESIS TESTING PROCEDURES



What can We do with This Model?

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$



One Phenotype Variable, M

SOLUTION

$$Y = X\beta_X + M\beta_M + Z\beta_Z + E$$

Non-Parametric Test using Wild Bootstrap on unrotated matrix

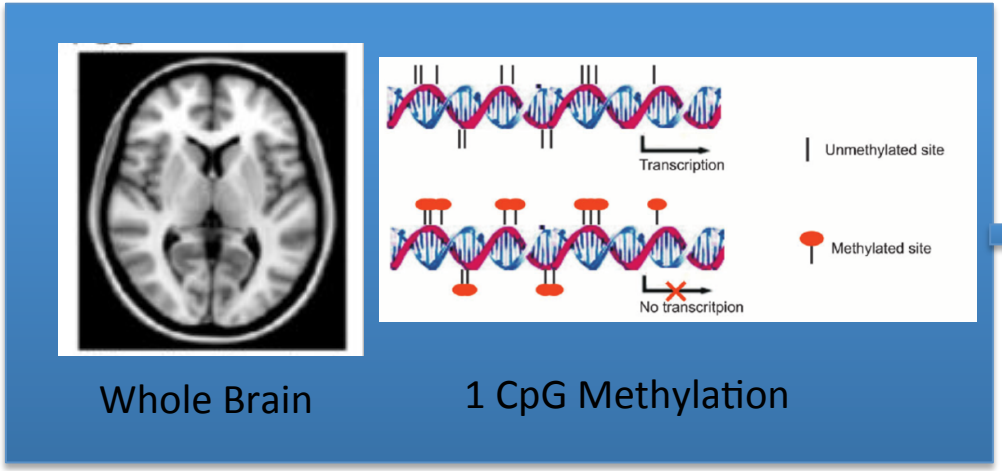
a Family-Wise Error Rate (FWER) corrected p-value at each voxel:

$$\frac{1}{(n_{\mathrm{B}}+1)} \sum_{b=0}^{n_{\mathrm{B}}} \mathcal{I}(T_b^{max} \geq T_{v0}),$$

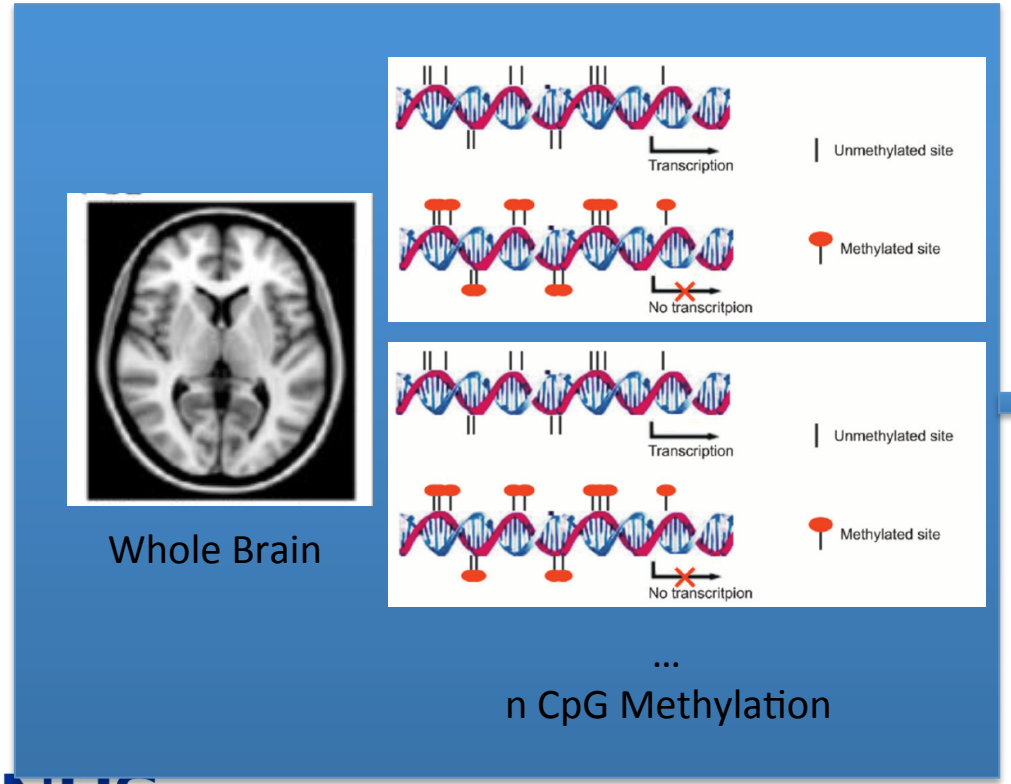
where T_b^{max} is the maximum score observed for the b^{th} resampled data.

NUS National University of Singapore

Application to Imaging Epigenetics

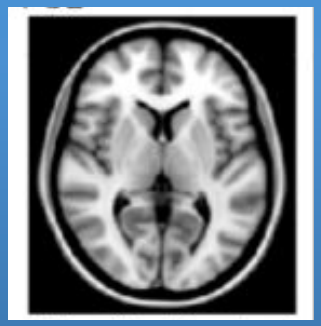


multiple comparisons across the whole brain

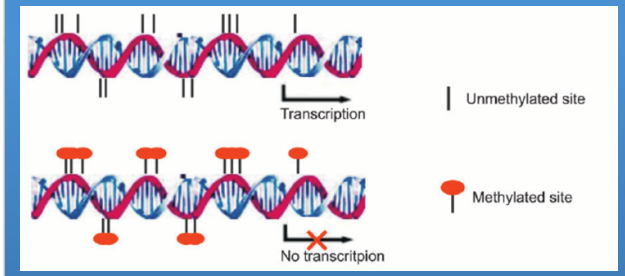


multiple comparisons across the whole brain and across n CpGs

HANDS-ON EXPERIMENTS

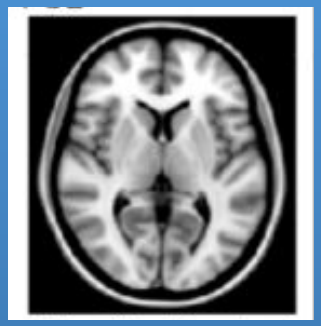


Whole Brain

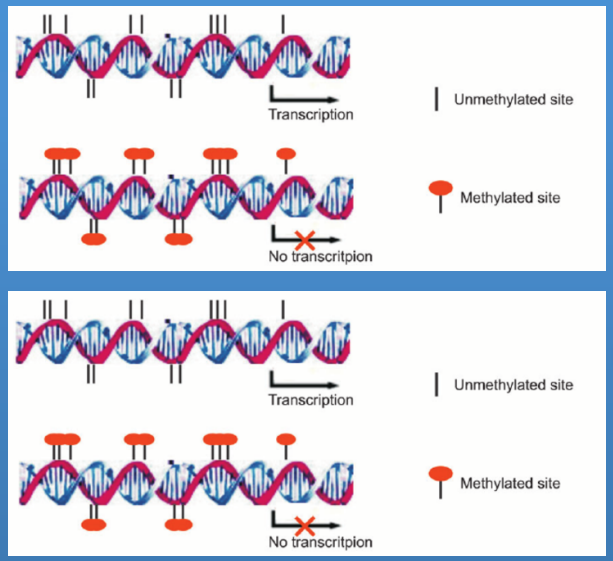


1 CpG Methylation

multiple comparisons across the whole brain



Whole Brain



...

n CpG Methylation

multiple comparisons across the whole brain and across n CpGs

HANDS-ON EXPERIMENT

Download the software package

http://www.bioeng.nus.edu.sg/cfa/imaginggenetics.html

2. Improving mass-univariate analysis of Genome-Wide and Epigenome-Wide Association Studies on neuroimage data by modelling important unknown covariates

Download Matlab Toolbox

Download Demo Data

Statistical inference on neuroimaging data is often conducted using a mass-univariate model, equivalent to fitting a univariate linear model at every voxel with a known set of covariates. Due to the large number of univariate linear models, it is challenging to check if the selection of covariates is appropriate and to modify this selection adequately. Indeed, the use of standard diagnostics for a univariate regression, such as residual plotting, is clearly not practical for neuroimaging data. However, the

HANDS-ON EXPERIMENT

Step 1. Prepare the data, including

Y --- n_subj X n_voxels (image data)

X --- n_subj X n_c (covariates)

CpG --- n_subj X n_cpg (CpG)

HANDS-ON EXPERIMENT

Step 2. Estimate the number of unknown covariates across all CpGs

```
% maximum number of unknown covariates allowed to be estimated
nZMax = 10;

for iML = 1:nML
    nZ(iML) = neuroCate_estimateNumberOfFactors(Y, [X CpGv(:,iML)],
nZMax);
end

nZ_common = mode(nZ);
```

HANDS-ON EXPERIMENT

Step 3. Compute Resampling Matrix

```
nB = 999;
seed = 0;
resamplingMatrixWB = neuroCate_resamplingMatrix(nSubj, nB, 'wild bootstrap', seed);
```

HANDS-ON EXPERIMENT

Step 4. Compute the model with Wild Bootstrap for FWER-correction

`// construction of the empirical distribution for` $T^{\text{globalMax}}$

for *each methylation locus* **do**

1. use this model on the original data at this methylation locus to compute the maximum original score at this methylation locus and assign it to T^{localMax}.
2. set the global maximum original score as $T_0^{\text{globalMax}} = \max(T_0^{\text{globalMax}}, T^{\text{localMax}})$.

for *each resampling trial* **do**

1. generate resampled data Y_b according to Permutation or Wild Bootstrap.
2. compute the maximum resampled score at this resampling trial and this methylation locus and assign it to T^{localMax}.
3. set the global maximum resampled score at resample b as $T_b^{\text{globalMax}} = \max(T_b^{\text{globalMax}}, T^{\text{localMax}})$.

end

end

$T_b^{\text{globalMax}}, b = 0, 2, \cdots, n_B$ forms the empirical distribution for $T^{\text{globalMax}}$.

`// computation of statistical maps for each methylation locus`

for *each methylation locus* **do**

1. compute T-statistic for each voxel.
2. compute p-values based on max-statistics and the aforementioned empirical distribution for $T^{\text{globalMax}}$.

end

NUS National University of Singapore

HANDS-ON EXPERIMENT

Step 4. Compute the model with Wild Bootstrap for FWER-correction

```
maxGlobalScoreF = zeros(nB + 1,1);

for iML = 1:nML
    result = neuroCate_cate(thick, X, CpGv(:,iML), 0, 'inference', 'wb', 'nB', nB, 'resamplingMatrix', resamplingMatrixWB);

    maxGlobalScoreF = max(maxGlobalScoreF, result.maxScoreF);
end
```

```
>> result

result =

       pParamUnc: [1x120117 double]
          scoreF: [1x120117 double]
           betaM: [1x120117 double]
          alphaM: []
           betaZ: []
     sigmaSquare: [1x120117 double]
               Z: []
    pNonParamUnc: [1x120117 double]
    pNonParamFWE: [1x120117 double]
       maxScoreF: [10x1 double]
```

NUS National University of Singapore

HANDS-ON EXPERIMENT

Step 5. Compute corrected p-value for each voxel

A Family-Wise Error Rate (FWER) corrected p-value is obtained at every voxel as

$$\frac{1}{(n_{\mathrm{B}}+1)}\sum_{b=0}^{n_{\mathrm{B}}}\mathcal{I}(T_b^{globalMax} \geq T_{v0}),$$

where $T_b^{globalMax}$ is the maximum score observed for the b^{th} resampled data.

TECHNICAL QUESTIONS

- Is the resampling matrix the same across all the CpGs (or all phenotypes)?

- How do we achieve correction of multiple comparisons across all the CpGs and all the voxels?

HANDS-ON EXPERIMENT

Step 4. Compute the model with Wild Bootstrap for FWER-correction

---

`// construction of the empirical distribution for` $T^{\text{globalMax}}$

for *each methylation locus* **do**

1. use this model on the original data at this methylation locus to compute the maximum original score at this methylation locus and assign it to T^{localMax}.
2. set the global maximum original score as $T_0^{\text{globalMax}} = \max(T_0^{\text{globalMax}}, T^{\text{localMax}})$.

for *each resampling trial* **do**

1. generate resampled data Y_b according to Permutation or Wild Bootstrap.
2. compute the maximum resampled score at this resampling trial and this methylation locus and assign it to T^{localMax}.
3. set the global maximum resampled score at resample b as $T_b^{\text{globalMax}} = \max(T_b^{\text{globalMax}}, T^{\text{localMax}})$.

end

end

$T_b^{\text{globalMax}}, b = 0, 2, \cdots, n_B$ forms the empirical distribution for $T^{\text{globalMax}}$.

`// computation of statistical maps for each methylation locus`

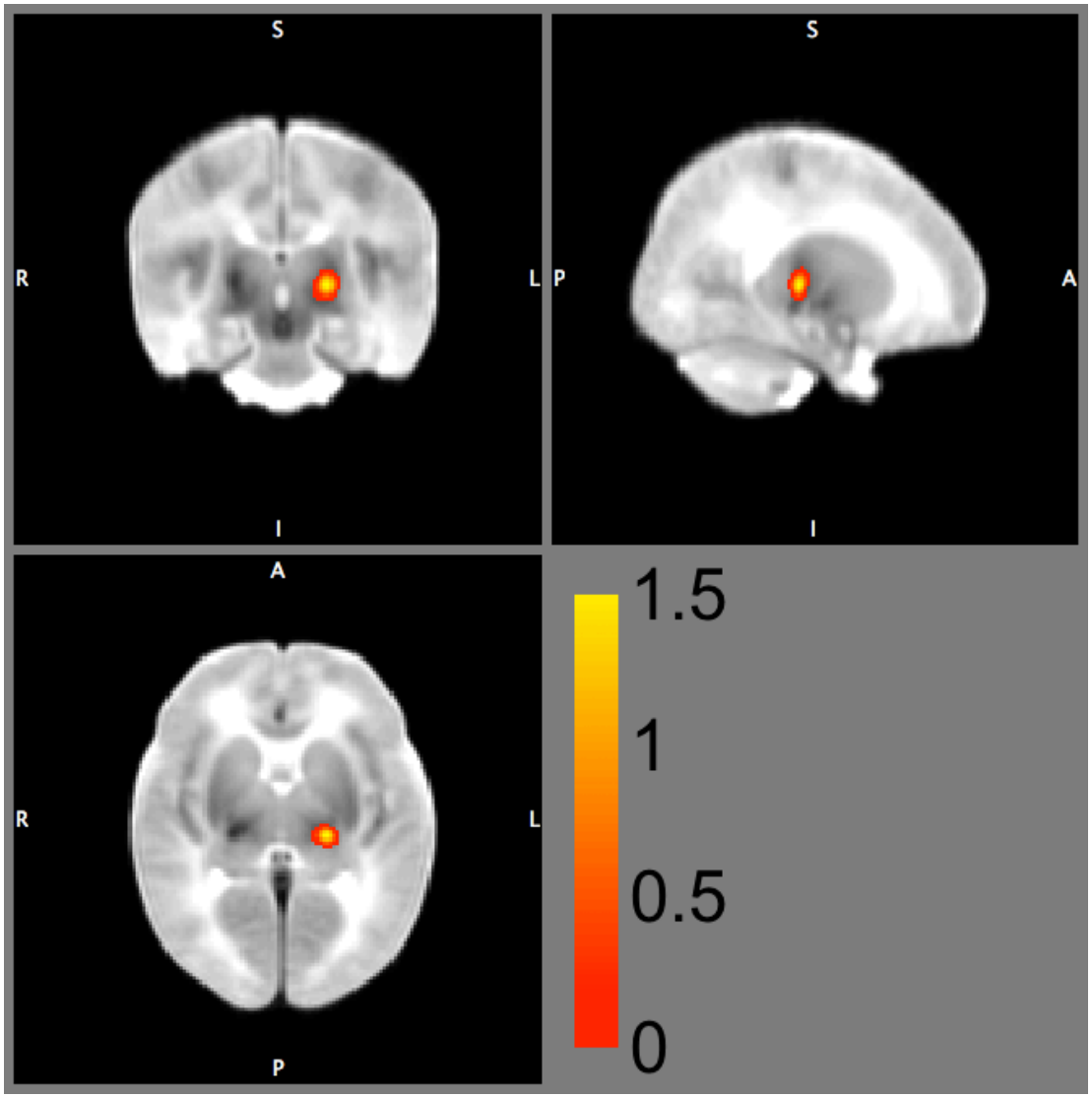
for *each methylation locus* **do**

1. compute T-statistic for each voxel.
2. compute p-values based on max-statistics and the aforementioned empirical distribution for $T^{\text{globalMax}}$.

end

NUS National University of Singapore

SIGNIFICANT ASSOCIATIONS BETWEEN CG19641625 AND VOXELS IN THE LEFT VENTROLATERAL THALAMUS



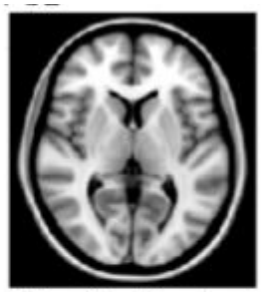
-log10(FWER-corrected p-value) image obtained for the methylation locus cg19641625

SMOC2 - SPARC-related modular calcium-binding protein 2.

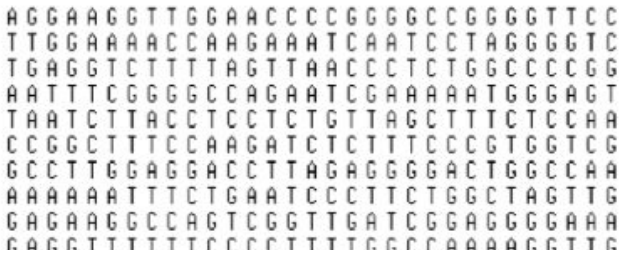
NUS National University of Singapore

SUMMARY

- A Statistical Model Incorporates Missing Covariates;
- Non-parametric Testing Improves Statistical Power and Controls FPR well.



Whole Brain

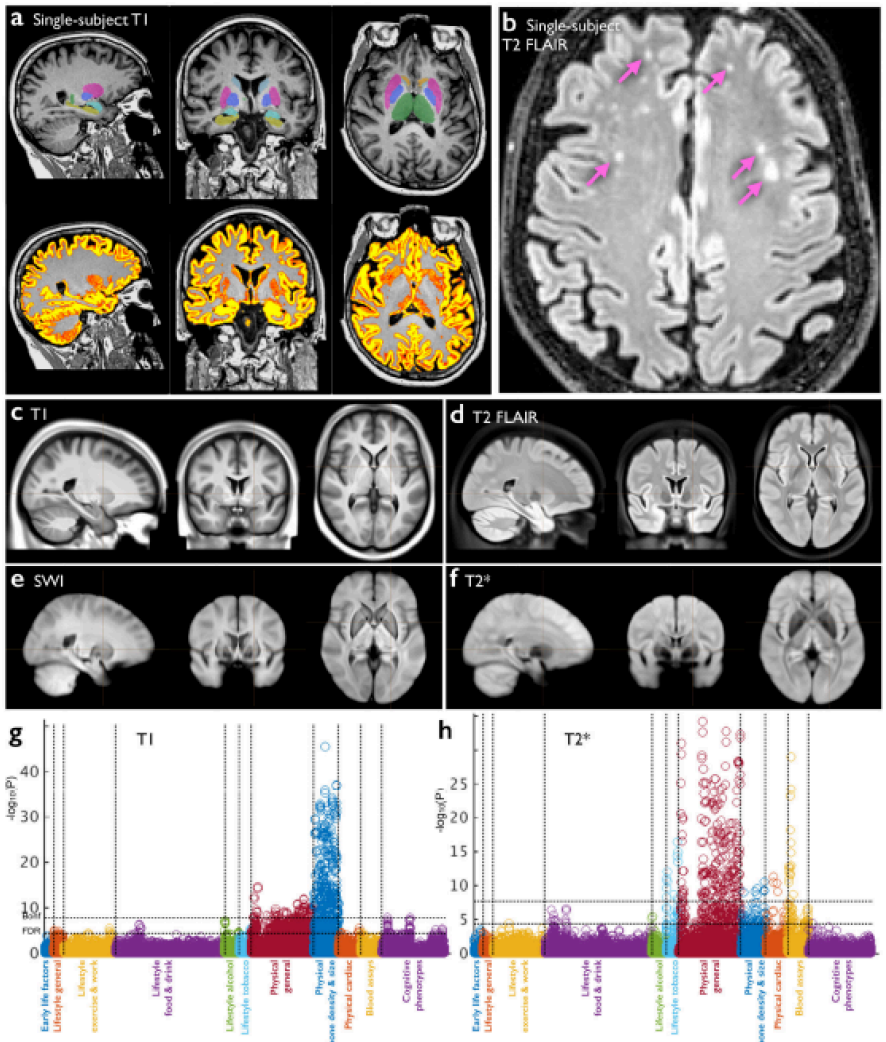


Individual SNPs/CpGs

GWAS/EWAS to Brain Image Voxels

⇩

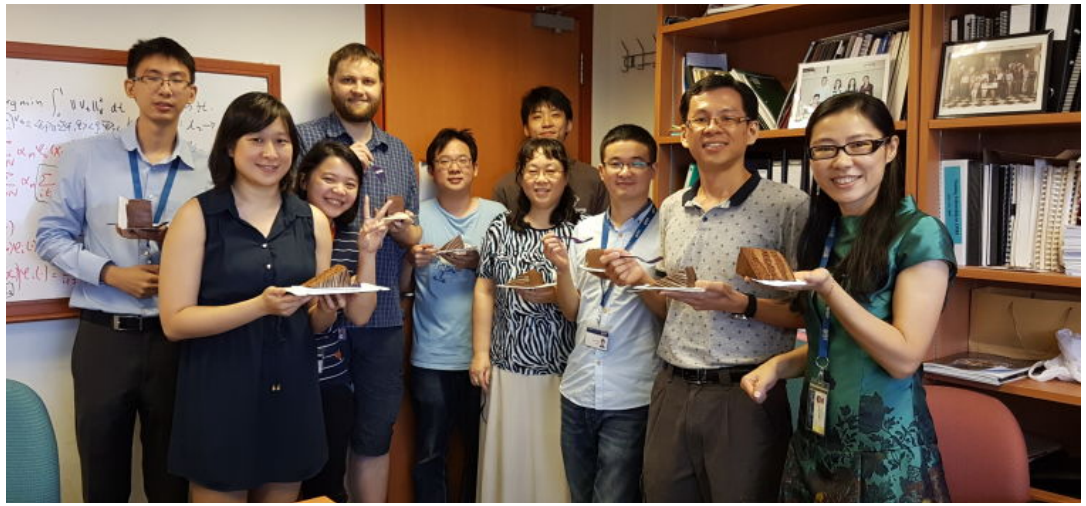
Statistical approaches to deal with correction for a large number of statistical tests.



NUS National University of Singapore

ACKNOWLEDGEMENTS

Laboratory for Medical Image Data Sciences at the National University of Singapore



NUS National University of Singapore

KK Women's and Children's Hospital
SingHealth

NUH National University Hospital

A*STAR Singapore Institute for Clinical Sciences