Two-sample testing with high-dimensional genetic and neuroimaging data

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

Nov 4, 2016

University of Georgia

Outline

- Introduction:
  1. Polygenic testing in GWAS;
  2. Functional connectivity (FC).
- Methods: 2-sample tests for high-dim data,
  Review: some existing tests;
  New: SPU/aSPU; (Pan et al 2014, *Genetics*)
  Theory: (Xu, Lin, Wei & Pan 2016, *Biometrika*)
- Applications and simulations.
- Discussion.

Introduction

- Application 1: Polygenic testing
- Example: the International Schizophrenia Consortium (ISC) (2009, *Nature*)
- $n_1 = 3322$ schizophrenia patients, $n_2 = 3587$ controls.
- $p = \sim 1$ million SNPs (single nucleotide polymorphsims) (coded 0, 1 or 2 for each).
- Any SNP associated with schizophrenia?
  univariate testing;
  high cost of multiple tests: genome-wide significance level 5×10^{-8};
  None found!
- "Dark matter" in genetics: missing heritability from genome-wide association studies (GWAS);
  Any association?

nature Vol 460|6 August 2009|doi:10.1038/nature08185

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[1,2]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. Here we show, using two analytic approaches, the extent to which common genetic variation underlies the risk of schizophrenia. First, we implicate the major histocompatibility complex. Second, we provide molecular genetic evidence for a substantial polygenic component to the risk of schizophrenia involving thousands of common alleles of very small effect. We show that this component also contributes to the risk of bipolar disorder, but not to several non-psychiatric diseases.

We genotyped the International Schizophrenia Consortium (ISC) case-control sample for up to ~1 million single nucleotide polymorphisms (SNPs), augmented by imputed common HapMap SNPs. In the genome-wide association study (GWAS; genomic control

Table 2, Supplementary Fig. 2 and section 5 and 6 in Supplementary Information).

The best imputed SNP, which reached genome-wide significance (rs3130297, $P = 4.79 \times 10^{-8}$, T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb)), was also in the MHC, 7 kilobases (kb) from *NOTCH4*, a gene with previously reported associations with schizophrenia[4]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at $P < 10^{-3}$, found on the ancestral European haplotype[5] (Table 1, Supplementary Table 3 and section 3 in Supplementary Information). However, it was not possible to ascribe the association to a specific HLA allele, haplotype or region (Supplementary Table 3 and Supplementary Fig. 4).

We exchanged GWAS summary results with the Molecular Genetics of Schizophrenia (MGS) and SGENE consortia for genotyped SNPs with $P < 10^{-3}$. There were 8,008 cases and 19,077 controls

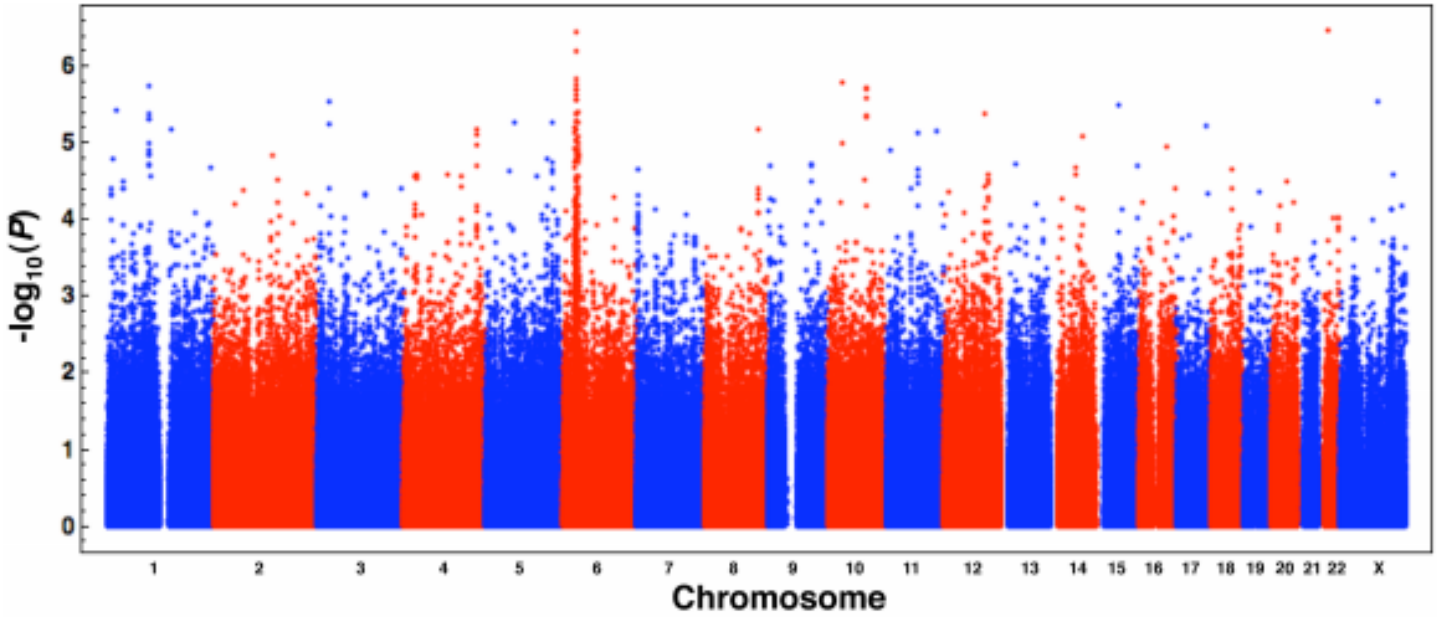


Figure S3: *Manhattan plot of single SNP Cochran-Mantel-Haenszel (CMH) test statistics, conditioning on the eight strata described above.*

Figure: ISC (2009, *Nature*), Fig S3.

ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

Schizophrenia is a highly heritable disorder. Genetic risk is conferred by a large number of alleles, including common alleles of small effect that might be detected by genome-wide association studies. Here we report a multi-stage schizophrenia genome-wide association study of up to 36,989 cases and 113,075 controls. We identify 128 independent associations spanning 108 conservatively defined loci that meet genome-wide significance, 83 of which have not been previously reported. Associations were enriched among genes expressed in brain, providing biological plausibility for the findings. Many findings have the potential to provide entirely new insights into aetiology, but associations at *DRD2* and several genes involved in glutamatergic neurotransmission highlight molecules of known and potential therapeutic relevance to schizophrenia, and are consistent with leading pathophysiological hypotheses. Independent of genes expressed in brain, associations were enriched among genes expressed in tissues that have important roles in immunity, providing support for the speculated link between the immune system and schizophrenia.

BRIEF COMMUNICATIONS

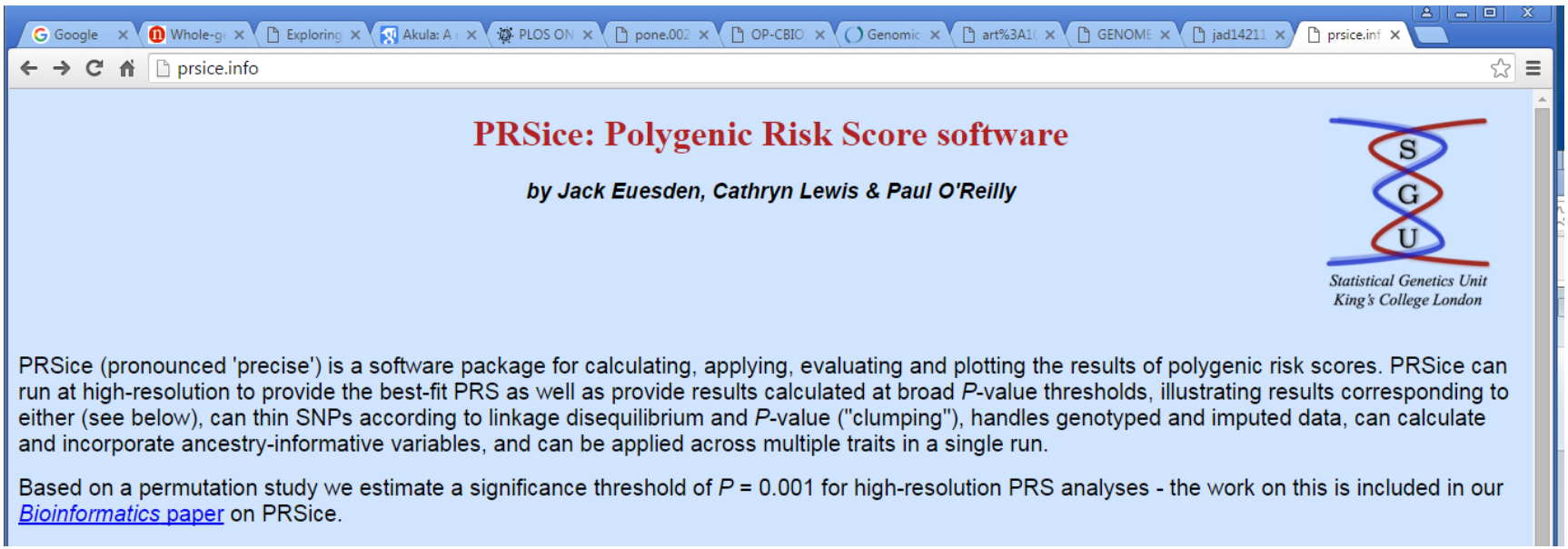
nature neuroscience

Polygenic risk scores for schizophrenia and bipolar disorder predict creativity

Robert A Power[1,2], Stacy Steinberg[1], Gyda Bjornsdottir[1], Cornelius A Rietveld[3], Abdel Abdellaoui[4], Michel M Nivard[4], Magnus Johannesson[5], Tessel E Galesloot[6], Jouke J Hottenga[4], Gonneke Willemsen[4], David Cesarini[7], Daniel J Benjamin[8], Patrik K E Magnusson[9], Fredrik Ullén[10], Henning Tiemeier[11], Albert Hofman[11], Frank J A van Rooij[11], G Bragi Walters[1], Engilbert Sigurdsson[12,13], Thorgeir E Thorgeirsson[1], Andres Ingason[1], Agnar Helgason[1,13], Augustine Kong[1], Lambertus A Kiemeney[6], Philipp Koellinger[14], Dorret I Boomsma[4], Daniel Gudbjartsson[1], Hreinn Stefansson[1] & Kari Stefansson[1,13]

with practical reasoning[8,10]. Furthermore, it has been suggested that those less restrained by practical cognitive styles may have an advantage in artistic occupations[8]. These results provide support for the notion that creativity and psychiatric disorders, particularly schizophrenia and bipolar disorder, share psychological attributes. However, whether and to what degree this is due to shared environment or genetics has not been assessed with modern genomic tools.

Creativity can be viewed in various ways[11,12], and, although it is a difficult concept to define for scientific purposes, the creative person is most often considered one who takes novel approaches requiring cognitive processes that are different from prevailing modes of thought or expression[11]. Thinking differently from others is therefore a prerequisite for creativity[11]. Schizophrenia and bipolar disorder are disorders of thoughts and emotions, which means that those affected show alterations in cognitive and affective processing. Yet it is unclear whether the cognitive deviations of psychiatric patients and of creative individuals


prsice.info
PRSice: Polygenic Risk Score software
by Jack Euesden, Cathryn Lewis & Paul O'Reilly
S
G
U
Statistical Genetics Unit
King's College London
PRSice (pronounced 'precise') is a software package for calculating, applying, evaluating and plotting the results of polygenic risk scores. PRSice can run at high-resolution to provide the best-fit PRS as well as provide results calculated at broad P-value thresholds, illustrating results corresponding to either (see below), can thin SNPs according to linkage disequilibrium and P-value ("clumping"), handles genotyped and imputed data, can calculate and incorporate ancestry-informative variables, and can be applied across multiple traits in a single run.
Based on a permutation study we estimate a significance threshold of P = 0.001 for high-resolution PRS analyses - the work on this is included in our Bioinformatics paper on PRSice.

RESEARCH ARTICLE

Genetic Epidemiolo

OFFICIAL JOURNAL
INTERNATION
EPIDEMIOLOGY S
www.geneticepi.org

Testing for Polygenic Effects in Genome-Wide Association Studies

Wei Pan,[1]* Yue-Ming Chen,[2] and Peng Wei[2]*

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota; [2]Division of Biostatistics and
Genetics Center, University of Texas School of Public Health, Houston, Texas

Received 22 December 2014; Revised 30 January 2015; accepted revised manuscript 23 February 2015.

Figure: Our approach and results.

Introduction

- Application 2: functional connectivity (FC)
- Why? How?
- Problem: based on fMRI data, estimate a functional connectivity (FC) network for each subject using Pearson's (marginal) correlations (or partial correlations or ...).
- Key Q: group comparisons
- Existinng approaches: univariate testing; network summary statistics; ...
  Powerful/flexible enough?

Disrupted Functional Brain Connectome in Individuals at Risk for Alzheimer's Disease

Jinhui Wang, Xinian Zuo, Zhengjia Dai, Mingrui Xia, Zhilian Zhao, Xiaoling Zhao, Jianping Jia, Ying Han, and Yong He

Research

Original Investigation

Disruption of Cortical Association Networks in Schizophrenia and Psychotic Bipolar Disorder

Justin T. Baker, MD, PhD; Avram J. Holmes, PhD; Grace A. Masters, MA; B. T. Thomas Yeo, PhD; Fenna Krienen, PhD; Randy L. Buckner, PhD; Dost Öngür, MD, PhD

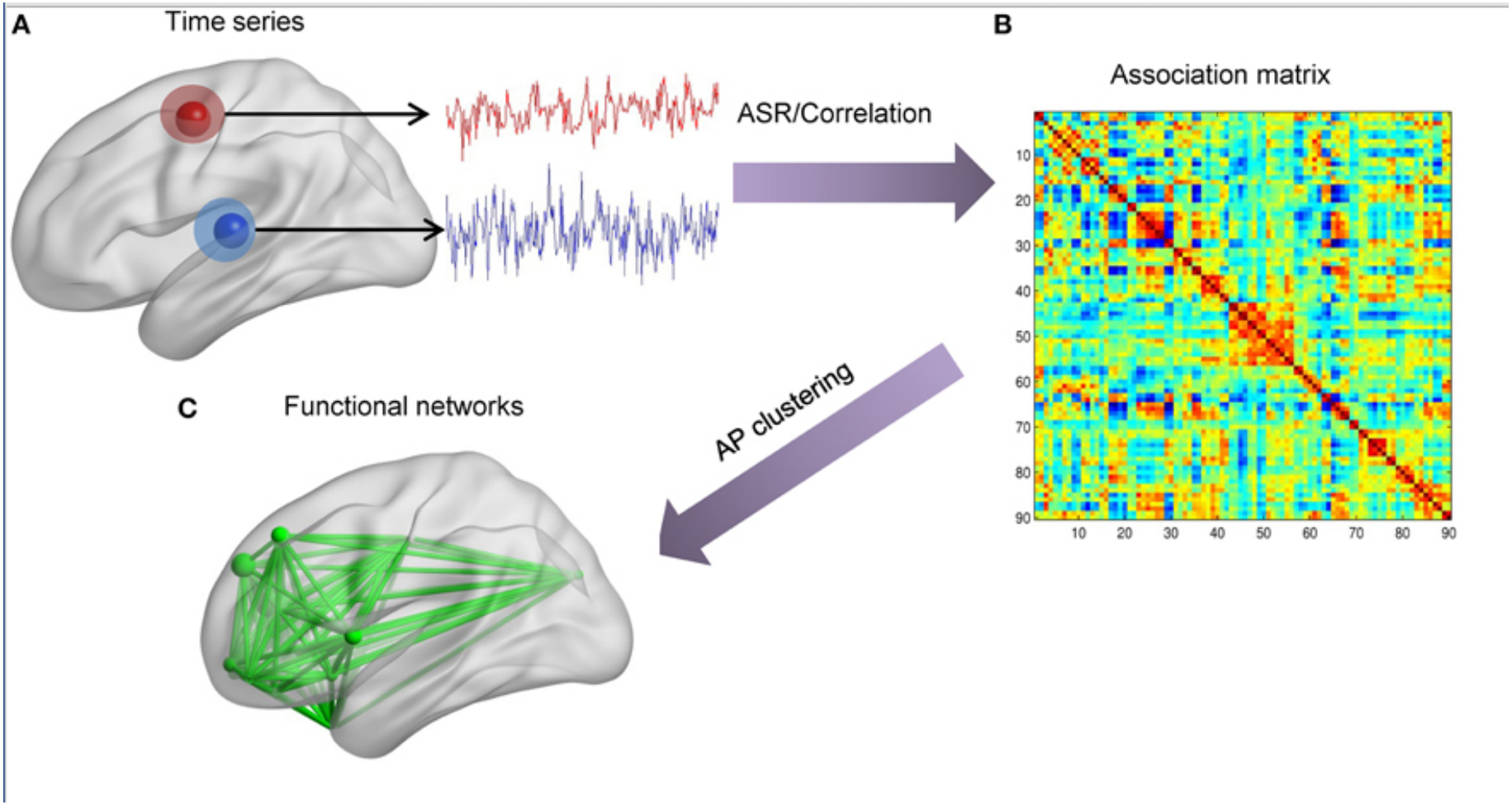
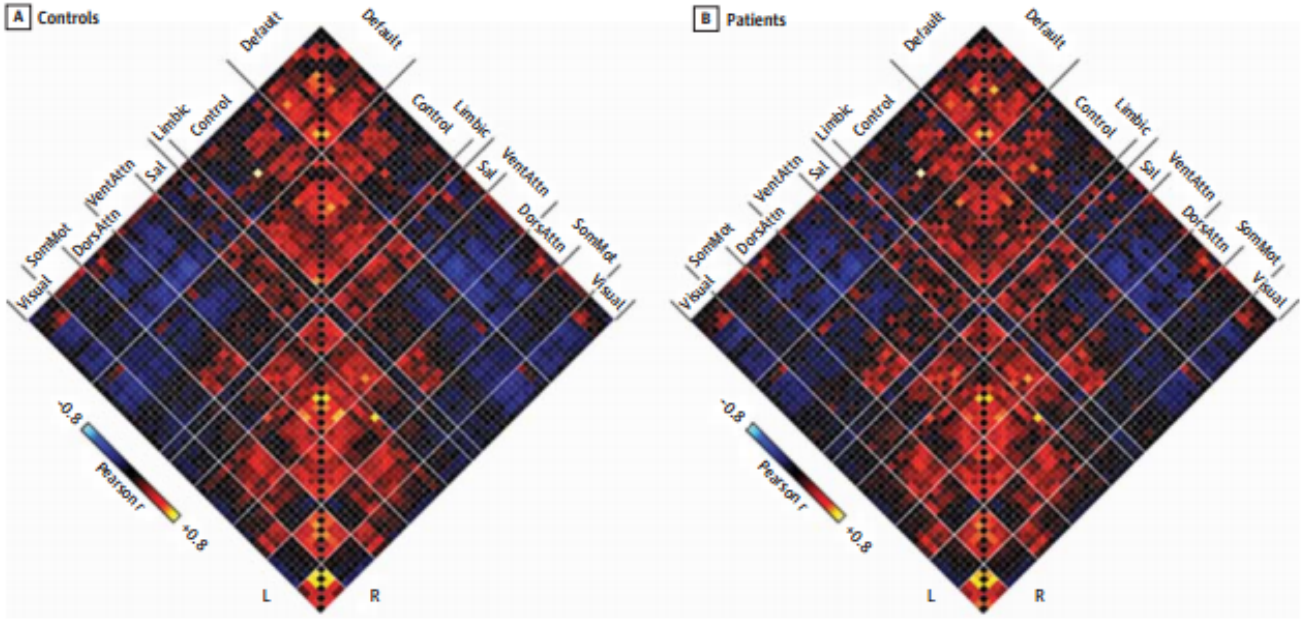


Figure: Li and Wang 2015, *Front. Neurosci.*, Fig 2.

Figure 1. Functional Connectivity Correlation Matrices in Patients and Controls



Each 61 × 61 grid shows the Pearson correlation between resting blood oxygenation level–dependent activity in intrahemispheric regional pairs for controls (A) and patients (B). Regions are ordered based on their network groupings adapted from Yeo et al.[26] Diagonal white lines represent network boundaries. DorsAttn indicates dorsal attention; L, left hemisphere; R, right hemisphere; Sal, salience; SomMot, somatomotor; and VentAttn, ventral attention.

Figure: Baker et al 2014, *JAMA Psychiatry*, Fig 1.

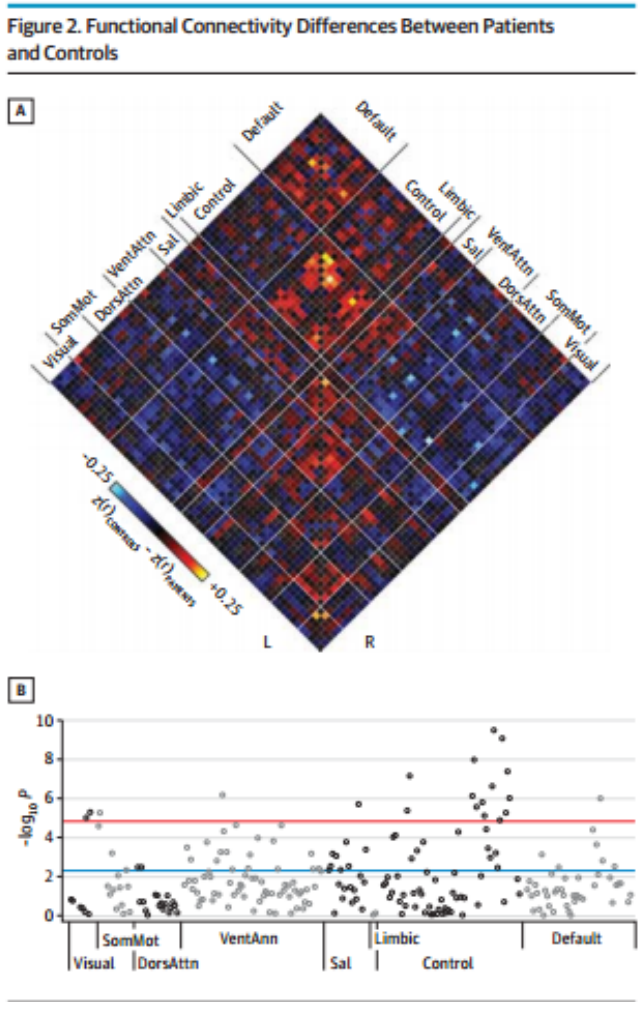


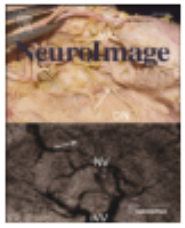
Figure: Baker et al 2014, *JAMA Psychiatry*, Fig 2.

NeuroImage 101 (2014) 681–694

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Comparison of statistical tests for group differences in brain functional networks

Junghi Kim [a], Jeffrey R. Wozniak [b], Bryon A. Mueller [b], Xiaotong Shen [c], Wei Pan [a,*]

[a] Division of Biostatistics, University of Minnesota, USA
[b] Department of Psychiatry, University of Minnesota, USA
[c] School of Statistics, University of Minnesota, USA

ARTICLE INFO

Article history:
Accepted 21 July 2014
Available online 30 July 2014

ABSTRACT

Brain functional connectivity has been studied by analyzing time series correlations in regional brain activities based on resting-state fMRI data. Brain functional connectivity can be depicted as a network or graph defined as a set of nodes linked by edges. Nodes represent brain regions and an edge measures the strength of functional

Figure: Our approach and results.

Problem formulation: two-sample testing

- Set-up: two samples, $\{\mathbf{X}_{1i}, i = 1, 2, \ldots, n_1\}$ and $\{\mathbf{X}_{2j}, j = 1, 2, \ldots, n_2\}$ with $p > \max\{n_1, n_2\}$.
  H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. (Or more generally, H_0: $F_1 = F_2$.)
- Sample means and covariance matrices: $n = n_1 + n_2$,
  $\bar{\mathbf{X}}_k = \sum_{i=1}^{n_k} \mathbf{X}_{ki}/n_k$.
  $\mathbf{S} = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(\mathbf{X}_{ki} - \bar{\mathbf{X}}_k\right) \left(\mathbf{X}_{ki} - \bar{\mathbf{X}}_k\right)^T / n$.
- Comment: here we assume $\Sigma_1 = \Sigma_2$; not necessary.
- Classic Hotelling's (1951) T^2-test,

$$T_{\mathrm{H}} = \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^T \mathbf{S}^{-1} \left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right). \tag{1}$$

  =t-test (or z-test) if $p = 1$.
  not working if $p > n$: S is singular; bad even $p \sim n$.

Review: some existing tests

- Bai and Saranadasa (1996, *Statistica Sinica*):

$$T_{\mathrm{BS}} = \frac{\frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^T\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) - \mathrm{tr}\mathbf{S}}{\sqrt{\frac{2n(n+1)}{(n-1)(n+2)}\left(\mathrm{tr}\mathbf{S}^2 - n^{-1}\left(\mathrm{tr}\mathbf{S}\right)^2\right)}},$$

  Under H_0, $T_{\mathrm{BS}} \xrightarrow{D} N(0,1)$.

- Chen et al (2010, *Ann Statist*):

$$T_{\mathrm{CQ}} = \frac{\sum_{i\neq j}^{n_1}\mathbf{X}_{1i}^T\mathbf{X}_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2}\mathbf{X}_{2i}^T\mathbf{X}_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbf{X}_{1i}^T\mathbf{X}_{2j}}{n_1 n_2}, \tag{2}$$

  which results after removing $\sum_{i=1}^{n_k}\mathbf{X}_{ki}^T\mathbf{X}_{ki}$ for $k = 1, 2$ from $\left\|\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right\|^2$. Hence

$$\frac{T_n - \left\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right\|^2}{\sqrt{\mathrm{Var}(T_n)}} \xrightarrow{D} N(0,1) \tag{3}$$

  as $n \longrightarrow \infty$ and $p \longrightarrow \infty$.

Review: some existing tests

- Cai, Liu and Xia (2014, *JRSS-B*):

$$T_{\mathrm{CLX}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le i \le p} \left(\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} \right)^2 \Big/ \sigma_{ii},$$

with σ_{ii} (always) replaced by S_{ii};
follows an asymptotic extreme value distribution under H_0.

- Chen, Li and Zhong (2014):

$$T_{\mathrm{CLZ}}(s) = \sum_{i=1}^{p} \left\{ \frac{n_1 n_2}{n_1 + n_2} \left(\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} \right)^2 \Big/ \sigma_{ii} - 1 \right\} I \left\{ \frac{n_1 n_2}{n_1 + n_2} \left(\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} \right)^2 \Big/ \sigma_{ii} > \lambda_p(s) \right\},$$

$$T_{\mathrm{CLZ}} = \max_{s \in (0, 1-\eta)} \frac{T_{\mathrm{CLZ}}(s) - \hat{\mu}_{T_{\mathrm{CLZ}}(s),0}}{\hat{\sigma}_{T_{\mathrm{CLZ}}(s),0}},$$

follows an asymptotic extreme value distribution under H_0.

- Srivastava and Du (1998, *JMA*):

$$T_{\mathrm{SD}} = \frac{\frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right)^T \mathbf{D}_S^{-1}\left(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2\right) - \frac{np}{n-2}}{\sqrt{2\left(\operatorname{tr}\mathbf{R}^2 - p^2/n\right)c_{p,n}}},$$

with $\mathbf{D}_S := \operatorname{diag}(\mathbf{S})$, $\mathbf{R} := \mathbf{D}_S^{-1/2}\mathbf{S}\mathbf{D}_S^{-1/2}$, and $c_{p,n} = 1 + \operatorname{tr}\mathbf{R}^2/p^{3/2}$.

New: SPU and aSPU tests

- Sum of Powered Score (SPU) test: for a positive integer γ,

$$\text{SPU}(\gamma) = \sum_{i=1}^{p} \left(\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)} \right)^{\gamma}. \tag{4}$$

- Key: a larger γ makes "the rich get richer"!

$$\text{SPU}(\gamma) \sim ||\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2||_{\gamma} \to ||\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2||_{\infty} = \max_i |\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)}|$$

  as (an even) $\gamma \to \infty$.

- define

$$\text{SPU}(\infty) = \max_{1 \le i \le p} \left(\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)} \right)^2 \Big/ \sigma_{ii}.$$

- Weighting: $\left(\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)} \right)^{\gamma} = \left(\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)} \right)^{\gamma-1} \left(\bar{\mathbf{X}}_1^{(i)} - \bar{\mathbf{X}}_2^{(i)} \right)$.

New: SPU and aSPU tests

- Remarks:
  Chen et al (2010): $\sim$ SPU(2);
  Cai et al (2014): $\sim$ SPU(∞);
  Chen et al (2014): $\sim$ tSPU(2) $\approx$ aSPU(2)=aSSU (Pan& Shen 2011, *Genet Epi*); $\sim$ PRS (Pan et al 2015, *Genet Epi*);
  SPU(1) =Sum =Burden test in rare variant (RV) analysis ...
  SPU(2) = KMR/SKAT = MDMR/PERMANOVA if ... (Pan 2011, *Genet Epi*)
- Q: which γ to use?
- Key: no uniformly most powerful test.
- Define an adaptive SPU (aSPU) test:

$$\text{aSPU} = \min_{\gamma \in \Gamma} P_{\text{SPU}(\gamma)}$$

e.g., $\Gamma = \{1, 2, ..., 8, \infty\}$.

Theorem for SPU tests

Let Γ be a set of finite positive integers. Under H_0, we have

$$\{\sigma(\gamma)^{-1}(\mathrm{SPU}(\gamma) - \mu(\gamma)) : \gamma \in \Gamma\}' \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\xi}),$$

and for $x \in \mathbb{R}$,

$$P(n\mathrm{SPU}(\infty) - a_p \leq x) \to \exp\left\{-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right\}$$

as $n, p \to \infty$, where $a_p = 2\log p - \log\log p$ and $n = n_1 n_2/(n_1 + n_2)$.
Moreover, $\{\sigma(\gamma)^{-1}(\mathrm{SPU}(\gamma) - \mu(\gamma)) : \gamma \in \Gamma\}$ and $n\mathrm{SPU}(\infty) - a_p$ are asymptotically independent.

P-value calculations

- Asymptotics:

$$p_O = 1 - \int_{\substack{s=(s_\gamma:\ \text{odd}\ \gamma\in\Gamma)^T \\ -T_O\le s_\gamma\le T_O}} N(0, R_O)ds,$$

$$p_E = 1 - \int_{\substack{t=(t_\gamma:\ \text{even}\ \gamma\in\Gamma)^T \\ -\infty< t_\gamma\le T_E}} N(0, R_E)dt,$$

$$p_{\min} := \min\{p_O, p_E, p_\infty\},$$

$$p_{\text{aSPU}} = 1 - (1 - p_{\min})^3.$$

- Permutations: permuting group labels.

Approximation for $\mu(\gamma)$

Under the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$,

$$\mu(\gamma) = \begin{cases} 0, & \text{i} \\ \frac{\gamma!}{2^{\gamma/2}} \sum_{d=0}^{\gamma/2} \frac{1}{d!(\gamma/2-d)! n_1^d n_2^{\gamma/2-d}} \sum_{i=1}^{p} \sigma_{ii}^{\gamma/2} + o(\frac{p}{n^{\gamma/2}}), & \text{i} \\ \sum_{d=1}^{\lfloor \gamma/2 \rfloor} \frac{\gamma!}{(d-1)!(\lfloor \gamma/2 \rfloor - d)! 3! 2^{\lfloor \gamma/2 \rfloor - 1}} & \\ \times \sum_{i=1}^{p} \left(\frac{m_{1i}}{n_1^{d+1} n_2^{\lfloor \gamma/2 \rfloor - d}} - \frac{m_{2i}}{n_1^{\lfloor \gamma/2 \rfloor - d} n_2^{d+1}} \right) \sigma_{ii}^{\lfloor \gamma/2 \rfloor - 1} + o(\frac{p}{n^{\lfloor \gamma/2 \rfloor + 1}}), & \text{i} \end{cases}$$

where m_{ki} is the third central moment of the random variablein component i from group k, i.e., $m_{ki} = \mathrm{E}\left[(\mathbf{X}_k^{(i)} - \boldsymbol{\mu}_k^{(i)})^3\right]$.

Approximation for $\sigma(\gamma)$

Under some regularity conditions and $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, when $\gamma = 1$,

$$\sigma^2(1) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{1}^T \boldsymbol{\Sigma} \mathbf{1},$$

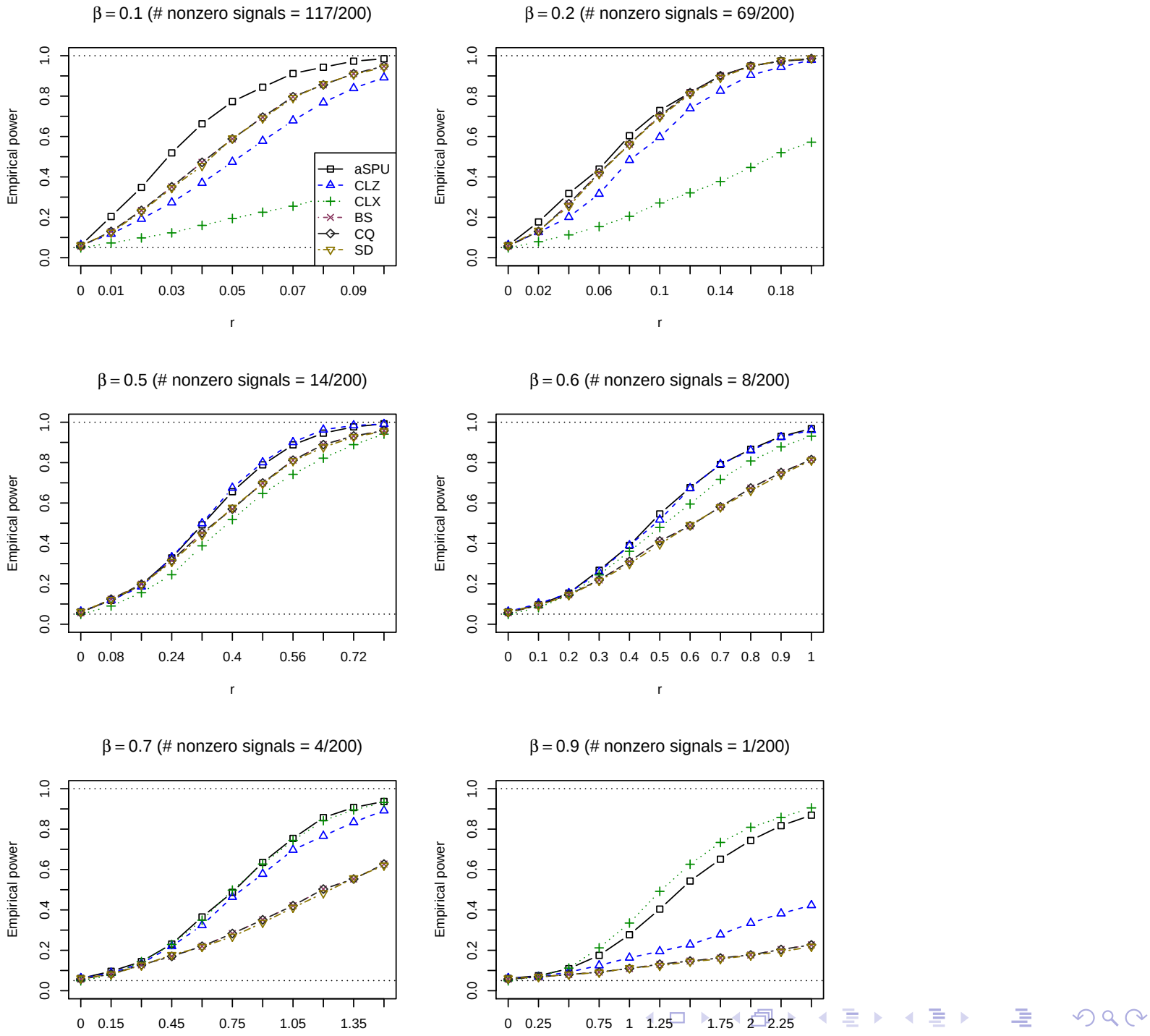
where $\mathbf{1}$ is a p-dimensional vector with all elements 1; for $\gamma \geq 2$,

$$\sigma^2(\gamma) \sim \mu(2\gamma) - \sum_{i=1}^{p} [\mu^{(i)}(\gamma)]^2 +$$

$$\sum_{\substack{2c_1+c_3+2d_1+d_3=\gamma \\ 2c_2+c_3+2d_2+d_3=\gamma \\ c_1,c_2,d_1,d_2 \geq 0, c_3+d_3>0}} \frac{(\gamma!)^2 \sum_{i \neq j} \sigma_{ii}^{c_1+d_1} \sigma_{jj}^{c_2+d_2} \sigma_{ij}^{c_3+d_3}}{n_1^{c_1+c_2+c_3} n_2^{d_1+d_2+d_3} c_1! c_2! c_3! d_1! d_2! d_3! 2^{c_1+c_2+d_1+d_2}}.$$

Simulations

- Simulation set-ups follow Chen et al (2014).
- $n_1 = n_2 = 50$, $p = 200$.
- Under H_0, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$; under H_1, $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2$ has $\lfloor p^{1-\beta} \rfloor$ non-zero entries of equal value, which are uniformly allocated among $\{1, 2, \ldots, p\}$. $\beta = 0, 0.1, 0.2, \ldots, 0.9$.
- The values of the non-zero entries are $\sqrt{2r \log p(1/n_1 + 1/n_2)}$. $r = 0, 0.1, 0.2, 0.3, 0.4, \ldots$.
- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = 0.6^{|i-j|}$.
- Results:
- Based on 1000 replicates; all used permutations $B = 1000$


β = 0.1 (# nonzero signals = 117/200)
β = 0.2 (# nonzero signals = 69/200)
β = 0.5 (# nonzero signals = 14/200)
β = 0.6 (# nonzero signals = 8/200)
β = 0.7 (# nonzero signals = 4/200)
β = 0.9 (# nonzero signals = 1/200)
Empirical power
r
aSPU
CLZ
CLX
BS
CQ
SD

Application 1: polygenic testing

- WTCCC (Burton et al 2007, *Nature*);
- $n_1 = 1868$ bipolar disorder (BD) patients and $n_2 = 2938$ controls;
- After QC, $p = 354,796$ SNPs; using Plink to prune to $p = 42092$ SNPs;
- There are strong polygenic effects ($P = 1 \times 10^{-12}$ for WTCCC data, ISC 2009, *Nature*), we considered chromosome-specific testing.
  Permutation (asymptotic) p-values.

Test	Chromosome (# SNPs)			
	1 (3340)	4 (2617)	13 (1592)	18 (1421)
SPU(1)	0.6431 (0.6355)	0.0024 (0.0017)	0.0372 (0.0375)	0.3229 (0.3287)
SPU(2)	<0.0001 (<0.0001)	0.0173 (0.0144)	0.0292 (0.0260)	0.2868 (0.2882)
SPU(3)	0.7454 (0.7374)	0.0314 (0.0308)	0.1264 (0.1294)	0.1740 (0.1865)
SPU(4)	<0.0001 (<0.0001)	0.0268 (0.0270)	0.0025 (0.0009)	0.3315 (0.3526)
SPU(5)	0.7323 (0.7417)	0.3606 (0.3754)	0.3713 (0.3938)	0.2344 (0.2591)
SPU(6)	0.0003 (<0.0001)	0.0407 (0.0270)	0.0040 (0.0001)	0.3864 (0.4477)
SPU(∞)	0.1183 (0.1310)	0.1194 (0.1211)	0.0800 (0.0879)	0.0038 (0.0047)
aSPU	<0.0001 (<0.0001)	0.0118 (0.0116)	0.0128 (0.0013)	0.0187 (0.0140)
CLZ	0.0004 (<0.0001)	0.1019 (0.0957)	0.0051 (0.0017)	0.0657 (0.0559)
CLX	0.1183 (0.1310)	0.1194 (0.1211)	0.0800 (0.0879)	0.0038 (0.0047)
BS	<0.0001 (<0.0001)	0.0173 (0.0146)	0.0292 (0.0263)	0.2868 (0.2885)
CQ	<0.0001 (<0.0001)	0.0173 (0.0148)	0.0292 (0.0268)	0.2868 (0.2896)
SD	<0.0001 (<0.0001)	0.0098 (<0.0001)	0.1142 (<0.0001)	0.0969 (<0.0001)

Application 1: another dataset

- Pan et al (2015, *Genet Epi*);
- SAGE GWAS on alcohol dependence (Bierut et al 2010);
  $n_1 = 1165$ cases and $n_2 = 1379$ controls;
  a total of 948,658 SNPs; 607,033 SNPs after QC;
  None reached the genome-wide significance by univariate testing!
- Previous twin/familial studies showed heritability of alcohol dependence!
- Any here?
- Use Plink to trim to $p = 62,801$ nearly uncorrelated SNPs ($r^2 \leq 0.1$ with a sliding window of 200 SNPs and a step size of 20 SNPs).
- Results: based on 10 million permutations!

Test	P_T	p-value
PRS	0.01	0.0042
	0.05	7.29×10^{-5}
	0.10	5.04×10^{-5}
	0.20	1.61×10^{-5}
	0.30	5.85×10^{-6}
	0.40	1.37×10^{-6}
	0.50	1.23×10^{-6}
Bonferroni-adjusted p-value		8.64×10^{-6}
SPU(1)		5.12×10^{-4}
SPU(2)		$< 1 \times 10^{-7}$
SPU(3)		0.0433
SPU(4)		$< 1 \times 10^{-7}$
SPU(5)		0.1925
SPU(6)		6.54×10^{-5}
SPU(7)		0.3111
SPU(8)		0.0235
SPU(∞)		0.3383
aSPU		9.00×10^{-7}

Simulations: SNP data

Empirical Type I error rate (for $OR = 1$) and power (for $a > 1$) for polygenic tests (with sample splitting) and SPU/aSPU tests (without sample splitting) for 1000 independent SNPs, including k_1 causal SNPs (among $p = 1000$ SNPs) with OR_j's $\sim U(1, a)$.

Test	P_T	Null	$k_1 = 20$			$k_1 = 50$			$k_1 = 100$		
		$a = 1$	$a = 1.2$	1.3	1.4	1.1	1.2	1.3	1.1	1.15	1.2
PRS	0.05	.044	.109	.344	.728	.056	.298	.769	.093	.240	.674
	0.1	.053	.115	.299	.676	.057	.311	.767	.106	.284	.738
	0.5	.041	.101	.258	.488	.078	.298	.731	.121	.377	.769
SPU(1)		.053	.139	.182	.296	.162	.439	.733	.490	.781	.946
SPU(2)		.062	.234	.565	.819	.158	.657	.966	.327	.756	.981
SPU(4)		.058	.364	.817	.984	.159	.763	.994	.292	.782	.986
SPU(8)		.049	.348	.830	.982	.122	.630	.978	.166	.495	.918
SPU(16)		.056	.308	.769	.961	.105	.465	.924	.114	.339	.744
SPU(32)		.056	.293	.741	.950	.103	.413	.903	.110	.307	.682
SPU(∞)		.058	.297	.737	.949	.109	.408	.887	.115	.307	.674
aSPU		.055	.348	.806	.971	.203	.747	.992	.464	.877	.995

Review: PRS test

- The Polygenic Risk Score (PRS) test:
  1) Divide data $D = D_1 \cup D_2$;
  2) $w_j = w_j(D_1) = \hat{\beta}_{M,j} I(p_j < P_T)$ from marginal models;
  3) $s_i = \sum_j w_j(D_1) X_{ij}(D_2)$;
  4) t-test on s_i's with $i \in D_2$;
- The ISC-PRS is the same as the Sum (Poly-Sum) test on H_0': $\alpha_1 = 0$ in

$$\text{Logit}[Pr(Y_i = 1)] = \alpha_0 + \alpha_1 \sum_{j=1} w_j X_{ij},$$

  with the new genotype score $w_j X_{ij}$ and $i \in D_2$.
- Can construct Poly-SSU, Poly-UminP, ...
- Key: use a half of the sample to construct weights w_j's; use the other half for hypothesis testing.
  sample splitting is **not** efficient!

- Some algebra (and asymptotics) shows

$$T_{PRS(P_T)} \propto \frac{\sum_j U_j(D_1)U_j(D_2)I(p_j(D_1) < P_T)}{\mathrm{Var}(U_j(D_1))},$$

- Better to use

$$T_{tSSUw(P_T)} = \frac{\sum_j U_j(D)U_j(D)I(p_j(D) < P_T)}{\mathrm{Var}(U_j(D))},$$

- Thresholding and inverse-variance weighting are not really effective $\Longrightarrow$

$$T_{SSU} = \sum_j U_j(D)U_j(D),$$

  or even better, SPU(γ), and aSPU!
- aSSU (Pan and Shen 2011, *Genetic Epi*; Fan 1997, *JASA*) vs aSPU (Pan et al 2014, *Genetics*)...

Application 2: functional connectivity (FC)

- Kim, Wozniak, Mueller, Shen & Pan (2014, *NeuroImage*);
- A rs-fMRI dataset (Wozniak et al 2013);
  Group 1: patients, fatal alcohol spectrum disorder, $n_1 = 24$;
  Group 2: controls, $n_2 = 31$;
  74 (sub)cortical ROIs; $p = 2701$ possible edges;
  Each subject measured at 180 time points;
- For each subject i, calculate a $N \times N$ sample correlation matrix $\hat{R}_i$, then $X_i = \text{vech}(h(R_i))$ with $h()$ as Fisher's z-transformation.
- Then compare two groups of X_i's.
- Remarks: testing H_0: $\Sigma_1 = \Sigma_2$,
  Li & Chen (2012, *Ann Statist*): $\sim$ SPU(2);
  Cai, Liu & Xia (2013, *JASA*): SPU(∞).

Table: P-values after adjusting for age and gender for the FASD data.

Test	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
P-value	0.009	0.312	0.085	0.348	0.236	0.391	0.366	0.437	0.759	0.031
Test	MDMR		nbs(0.1)	nbs(0.25)	nbs(0.5)	nbs(0.75)	CharPath	Eclust	Eglob	Eloc
P-value	0.468		0.009	0.017	0.064	0.081	0.673	0.862	0.919	0.925

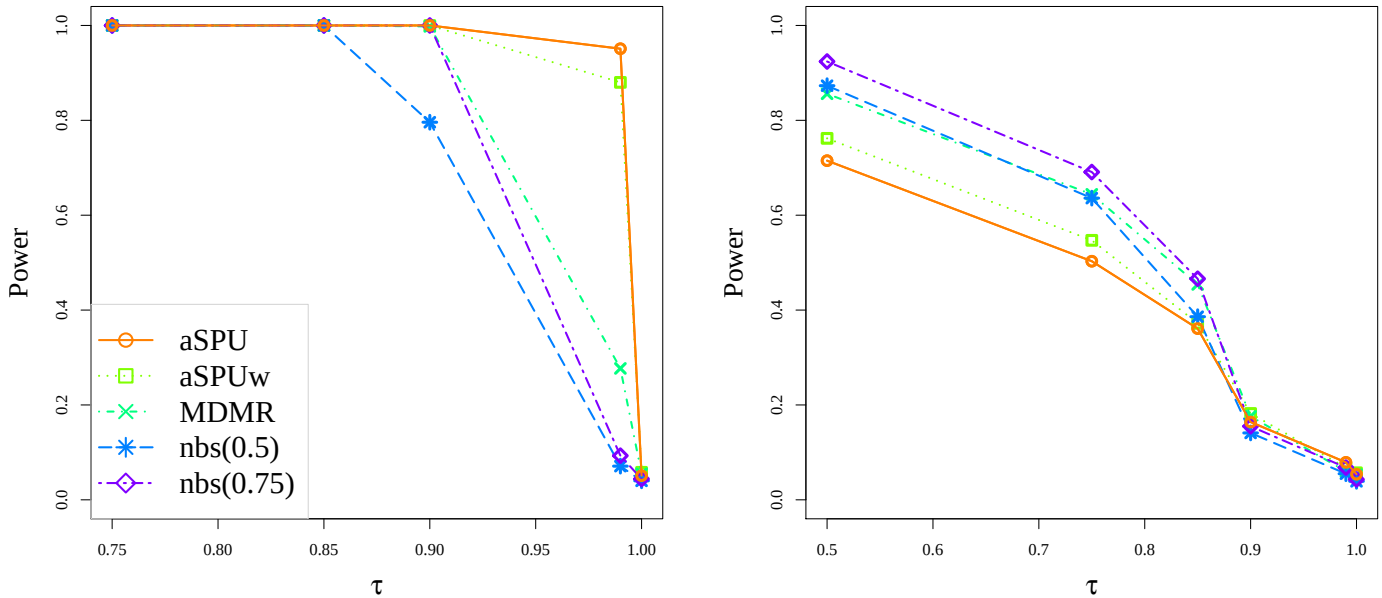


Figure: Sparse networks: empirical Type I error (for $\tau = 1$) and power (for $\tau < 1$) based on 1000 simulations.

Discussion

- Genetics
  - can be generalized to GLMs with covariates, RVs, $p < n$ (Pan et al 2014, *Genetics*);
  - extended to gene- and pathway-based association analysis (Pan et al 2015, *AJHG*);
  - extended to multiple traits (Zhang et al 2015, *NeuroImage*; Kim et al 2016, *Genetics*),
  - to that with only summary statistics (meta-analysis) (Kim et al, 2015, *Genet Epi*; Kwak and Pan 2016a, 2016b, *Bioinformatics*).
- Neuroimaging:
  - generalized to using regularized cov and precision matrices (Kim et al, 2015, *NeuroImage: Clinical*);
  - neuroimaging genetics: WGCNA/module detection (Gao, Kim & Pan 2017, *Pacific Biocomputing Symposium*; Kim & Pan (to appear), *Genet Epi*).

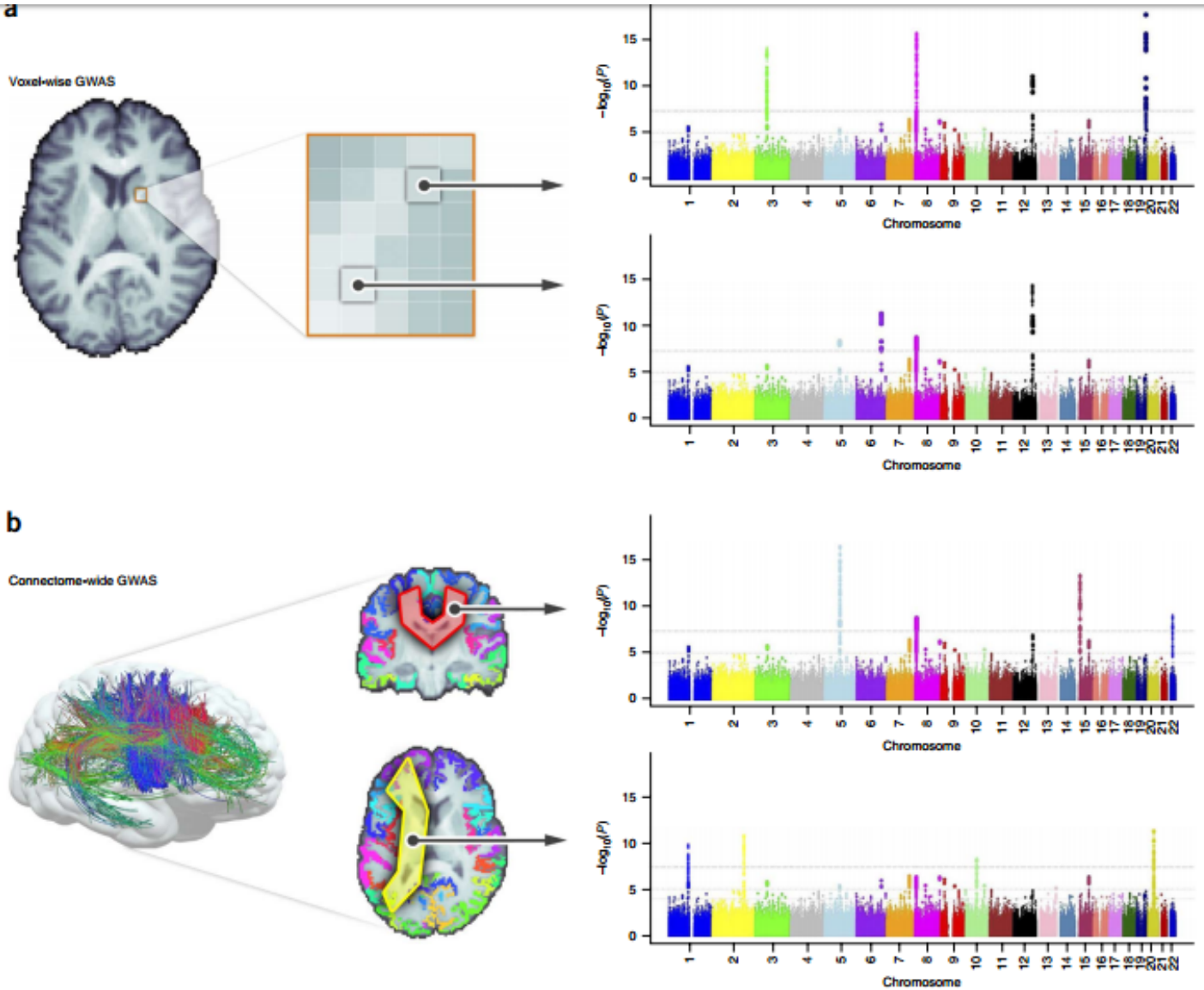


Figure 1 Whole-brain GWAS. **(a)** Voxel-wise genetic association analysis. This kind of analysis involves a genome-wide search at each voxel in the brain, after aligning all subjects' images to a common template. **(b)** Extending this method to study brain connections, Jahanshad *et al.*[30] described connectome-wide searches. They combined diffusion-based MRI tractography and cortical parcellations to perform GWAS at all connections between cortical regions of interest. Artificial Manhattan plots are illustrated here, with thresholds shown based on a single GWAS. Despite the vast number of tests, promising findings emerged, even after correction, from these whole-connectome genetic screens.

Acknowledgement

- This research was supported by NIH: R01 GM113250 (PI: Pan), R01 HL105397 (MPI: Pan/Shen), R01 HL116720 (MPI: Pan/Wei) and R01 GM081536 (MPI: Shen/Pan). .
- Polygenic testing: Peng Wei (UT-Houston);
- SPU/aSPU for RVs: Peng Wei (UT-Houston), Junghi Kim, Yiwei Zhang, Xiaotong Shen (UofM Statistics);
- 2-sample high-dim tests: Lifeng Lin, Gongjun Xu (UofM Statistics).
- Neuroimaging data: JR Wozniak, BA Mueller (UofM CMRR), ADNI

- `http://www.biostat.umn.edu/~weip`
  Code: `http://www.biostat.umn.edu/~weip/prog.html`
  R packages aSPU, highmean, GEEaSPU, POMaSPU, MiSPU; prclust, pGMGM, all on CRAN.

- **Thank you!**