A powerful and adaptive association test for rare variants

Wei Pan[1] , Junghi Kim[1] , Yiwei Zhang[1] , Xiaotong Shen[2]
Peng Wei[3],

[1] *Division of Biostatistics, School of Public Health,* [2] *School of Statistics, University of Minnesota, Minneapolis, MN 55455*

[3] *Division of Biostatistics and Human Genetics Center, University of Texas School of Public Health, Houston, TX 77030*

UNC, Sept 25, 2014

Outline

- Introduction: problem.
- Review: some existing methods.
- New methods: SPU and aSPU tests.
  Connections with some existing tests.
- Discussion.
- Ref.: Pan et al (2014), *Genetics*.
- Application to neuroimaging: Kim et al (2014), *NeuroImage*.

Introduction

- Problem:
  - Given: a binary disease indicator Y_i for subject i; a group of of rare variants (RVs) (additively) coded as $X_i = (X_{i1}, ..., X_{ik})'$; $i = 1, ..., n >> k$.
  - Q: any association between Y_i and X_i?
  - Approaches: global testing.
- Logistic reg model:

$$\text{Logit}[Pr(Y_i = 1)] = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j.$$

or, for $j = 1, ..., k$,

$$\text{Logit}[Pr(Y_i = 1)] = \beta_{M,j0} + X_{ij}\beta_{M,j}.$$

- H_0: $\beta = (\beta_1, ..., \beta_k)' = 0$, or $\beta_M = (\beta_{M,1}, ..., \beta_{M,k})' = 0$.

- Remark: other phenotypes or covariates can be accommodated.
- The score vector $U = (U_1, ..., U_k)'$ and its covariance:

$$U = \sum_{i=1}^{n} (Y_i - \bar{Y}) X_i,$$

$$V = Cov(U|H_0) = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'.$$

Some existing tests

- Burden tests (Morgenthaler & Thilly 2007; Li & Leal 2008; Madsen & Browning 2009):
  Sum test (Chapman & Whittaker 2008): assuming $\beta_1 = \beta_2 = ... = \beta_k = \beta_c$; H_0: $\beta_c = 0$;
  $$\text{Logit}[Pr(Y_i = 1)] = \beta_{c,0} + \sum_{j=1}^{k} X_{ij}\beta_c.$$
  $T_{Sum} = 1'U = \sum_{j=1}^{k} U_j$,
- Variance components tests:
  Sum of Squared Score (SSU) test (Pan 2009): assuming β_1,...,
  $\beta_k \sim F(0, \tau^2)$, H_0: $\tau^2 = 0$,
  $T_{SSU} = U'U = \sum_{j=1}^{k} U_j^2$.
  SSU test: equivalent to KMR (Liu et al 2008) with $K = XX'$ (Pan 2011), i.e. SKAT with no weighting and a linear kernel

(Wu et al 2011); C-alpha (Neal et al 2011), an EB test (Goeman et al 2006), GDBR/MDMR (Schork et al), ...

- UminP test: $T_{UminP} = \max_{j=1}^{k} U_j^2 / V_{jj}$, close to $T_{maxU} = \max_{j=1}^{k} |U_j|$
- A challenge: no uniformly most powerful test!
- Adaptive tests: with weights $\zeta = (\zeta_1, ..., \zeta_k)'$,

$$T_G = \zeta' U = \sum_{j=1}^{k} \zeta_j U_j,$$

  - aSum (Han and Pan 2010): $\zeta_j = -1$ (or 1) if $\hat{\beta}_{M,j} < 0$ (or > 0) and p-value $p_j < 0.1$;
  - PWST (Zhang et al 2011): $\zeta_j = 2(p_j - 0.5)$;
  - EREC (Lin and Tang 2011): $\zeta_j = \hat{\beta}_{M,j} \pm d$.
  - Note: $\hat{\beta}_M = Diag(V)^{-1} U + O_p(1/n)$,

1) If $|\hat{\beta}_M|$ is large, $\zeta \approx \hat{\beta}_M \propto U \Longrightarrow \text{EREC} \approx \text{SSU}$;
2) If $|\hat{\beta}_M|$ is small, $\zeta \approx \pm d \Longrightarrow \text{EREC} \approx \text{Sum}$;

- ...

- Key: how to choose ζ? Is any given choice of ζ sufficiently adaptive?
  Our answers:

New Tests: SPU and aSPU

- $\zeta_j = f(U_j) = U_j^{\gamma-1}$ for $\gamma \geq 1$;
- SPU tests: for a $\gamma \geq 1$,

$$T_{SPU(\gamma)} = \sum_{j=1}^{k} U_j^{\gamma}.$$

$$T_{SPU(\infty)} \propto \lim_{\gamma \to \infty} \left(\sum_{j=1}^{k} |U_j|^{\gamma} \right)^{1/\gamma} = \max_{j=1}^{k} |U_j|.$$

- Special cases:
  SPU(1) = Sum;
  SPU(2) = SSU;
  SPU(∞) = maxU $\approx$ UminP;
- Intuition in the choice of γ:
  1) the more sparse the signals, the larger γ;

2) if (most) associations in one direction, then use an odd γ.

- Our experience: often $SPU(8) \approx SPU(16) \approx SPU(\infty)$; If $SPU(\gamma) \approx SPU(\infty)$, then no need to increase γ.
- In parctice, how to choose γ? choose the one giving the most significant p-value?
- Use an adaptive SPU (aSPU) test:

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

where $P_{SPU(\gamma)}$ is the p-value of $SPU(\gamma)$, and $\Gamma = \{1, 2, ..., 8, \infty\}$.

- Computing: one loop of permutations or parameteric bootstrap is sufficient to calculate the p-values of $SPU(\gamma)$ for $\gamma \in \Gamma$ and aSPU tests!

Simulations

- Using a multivariate Normal to simulated (possibly correlated) RVs (Wang and Eslston 2008);
- MAFs $\sim U(0.001, 0.01)$;
- $k = k_1 + 0$, 8, 16, ..., 128; $k_1 = 8$ causal ones;
- Y_i from a joint logistic reg model with various valeus of β_j's; case-control design, $n = 1000$.
- 1000 replicates for each set-up.
- Show power for one case: causal RVs with $\exp(\beta_j) \sim U(1, 2)$, $j = 1, ..., 8$.

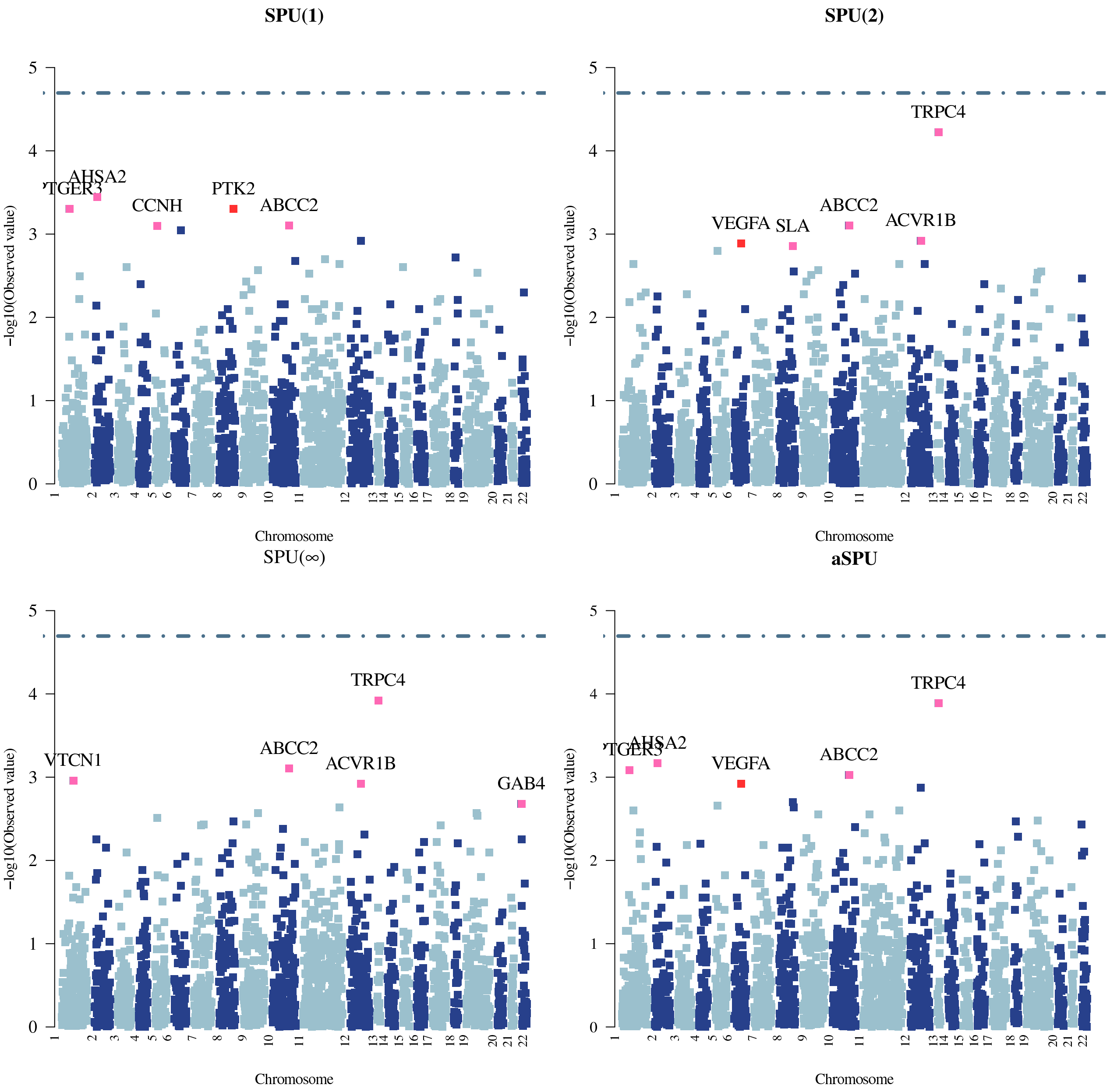
	# non-associated RVs						
Test	0	8	16	32	64	96	128
UminP	.874	.812	.768	.733	.659	.619	.586
SPU(1)	**.939**	.852	.746	.577	.411	.300	.244
SPU(2)	.926	**.908**	**.904**	.872	.832	.801	.769
SPU(3)	.917	.903	.893	.870	.829	.802	.786
SPU(4)	.909	.896	.890	**.882**	**.854**	**.840**	**.835**
SPU(5)	.902	.894	.879	.875	.843	.834	.834
SPU(6)	.901	.882	.872	.873	.843	.835	**.835**
SPU(7)	.899	.881	.868	.869	.836	.830	.828
SPU(8)	.898	.876	.863	.864	.833	.834	.826
SPU(∞)	.877	.852	.844	.846	.814	.801	.806
aSPU	.923	.898	.894	.869	**.842**	**.829**	**.811**
aSum	**.948**	.892	.855	.756	.636	.480	.407
PWST	.823	.729	.698	.613	.508	.400	.380
EREC	.943	.901	.887	.833	.738	.656	.579
SKAT	.927	**.914**	**.906**	**.870**	.823	.800	.749
SKAT-O	.940	.915	.899	.858	.799	.767	.696

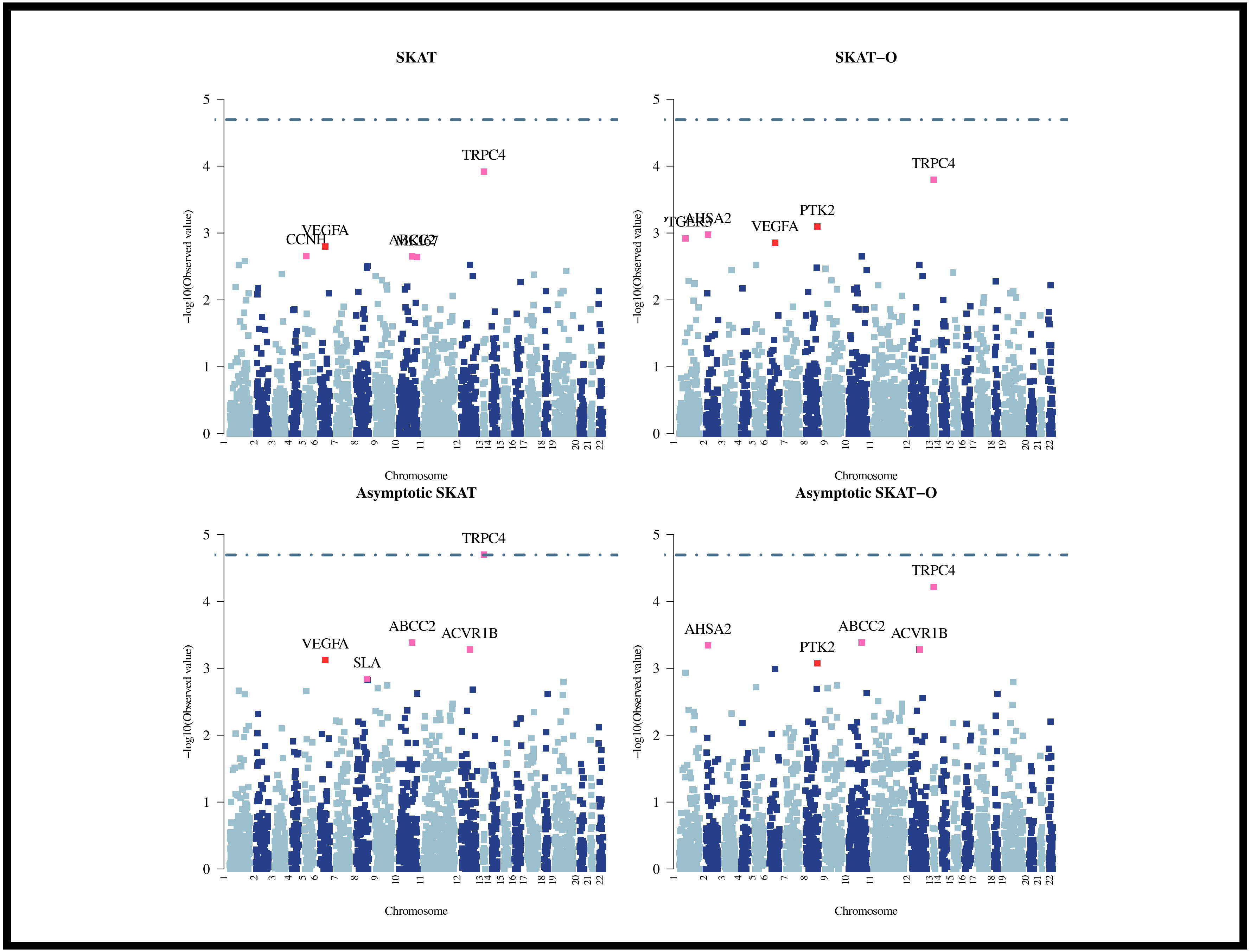
Example: GAW17 Data

- Mini-exome sequencing data with $n = 697$ unrelated subjects (Almasy et al 2011);
- After removing SNVs with MAFs $> 1\%$: 24,487 RVs in 2476 genes;
- A simulated binary phenotype; 200 sets
- Two analyses:
- 1) Applying gene-based testing on the set 1 of the binary phenotype, adjusting for age, gender and smoking status; Show p-values;
- 2) Applying causal gene-based testing on each of the 200 sets of the binary phenotype at the nominal level of 0.05; Show empirical power
- Resampling: start with $B = 10^3$; if a p-value $< 5/B$, then

increase B to 10^4, ..., up to 10^6

- Computing time for a genome-scan for one phenotype: in R; with 100 cores,
  it took about 0.2 hours to test 2476 genes based on $B = 10^3$ permutations;
  0.05 hours to test 50 genes with $B = 10^4$;
  0.12 hours to test 5 with $B = 10^5$.
- Analysis I: give p-values;


SPU(1)
SPU(2)
SPU(∞)
aSPU
−log10(Observed value)
Chromosome
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
TGER3
AHSA2
CCNH
PTK2
ABCC2
VEGFA
SLA
ACVR1B
TRPC4
VTCN1
GAB4


SKAT
SKAT−O
Asymptotic SKAT
Asymptotic SKAT−O
−log10(Observed value)
0
1
2
3
4
5
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
TRPC4
VEGFA
CCNH
ABCC2
MKI67
AHSA2
PTK2
ABCC2
ACVR1B
SLA

gene	#RVs	SPU(1)	SPU(2)	SPU(∞)	aSPU	SKATa	SKAT	SKAT-Oa	SKAT-O
VEGFA*	5, 1	0.00090	0.00060	0.02020	0.00120	0.00082	0.00150	0.00091	0.00110
PTK2*	9, 2	0.00080	0.00300	0.00240	0.00230	0.00178	0.00350	0.00105	0.00120
SOS2*	7, 2	0.18600	0.03500	0.01400	0.03200	0.03838	0.04496	0.06176	0.05195
RARB*	9, 2	0.07600	0.04900	0.04900	0.07100	0.06313	0.05794	0.09546	0.09690
SIRT1*	23, 9	0.35500	0.09500	0.13600	0.07200	0.08091	0.10589	0.14543	0.17383
GAB4	5, 0	0.05240	0.00940	0.00160	0.00370	0.01358	0.01230	0.02204	0.02040

- Analysis II: empirical power based on the significance level of 0.05 and the 200 replicates
- Show for causal genes; many with power < 0.05

Chr	Gene	#RVs	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU	SKAT	SKAT-O
1	PIK3C2B	60, 23	0.565	0.445	**0.650**	0.400	0.395	0.340	**0.600**	0.435	0.560
6	VNN1	6, 1	0.185	0.230	0.315	0.235	**0.380**	0.140	**0.270**	0.215	0.185
3	BCHE	28 , 13	0.110	0.190	**0.215**	0.185	0.175	0.160	**0.210**	0.195	0.175
8	LPL	15 , 2	0.090	0.130	**0.135**	0.110	0.115	0.110	**0.135**	0.125	0.125
10	SIRT1	23 , 9	0.095	0.105	**0.110**	0.065	0.070	0.015	**0.105**	0.090	0.100
14	SOS2	7 , 2	0.100	0.270	**0.285**	0.265	0.275	0.245	0.220	**0.255**	0.200
19	RRAS	5 , 2	**0.235**	0.140	0.155	0.145	0.155	0.100	0.180	0.135	0**.200**
8	PLAT	25 , 8	**0.225**	0.135	0.145	0.110	0.105	0.070	0.155	0.130	**0.195**
9	VLDLR	23 , 8	0.080	0.120	**0.125**	0.110	0.120	0.075	0.090	**0.125**	0.090
17	SREBF1	21 , 10	0.050	0.085	0.090	**0.105**	0.100	0.100	0.085	**0.090**	0.070
4	KDR	14 , 8	**0.365**	0.350	0.160	0.105	0.105	0.020	0.280	0.365	**0.390**
13	FLT1	25 , 8	0.125	0.160	**0.170**	0.150	0.160	0.065	0.125	0.150	**0.165**
14	HSP90AA1	20,3	0.050	**0.275**	0.180	0.195	0.170	0.030	0.155	**0.335**	0.250

Discussion

- Conclusion: aSPU test is promising (and general/flexible)
- Current work:
  applied to real data;
  develop an R package;
  analytical null distribution?
- Extensions:
  Pathway analysis; ongoing ...
  Multivariate (neuroimaging) traits-single SNP (Zhang et al 2014);
  Multivariate traits-multiple SNPs; ongoing ...
  To familial and/or longitudinal data; ongoing ...

Another Application

- To brain connectivity data: $k >> n$; Kim et al (2014).
- Problem: based on fMRI data, estimate a functional connectivity (FC) network for each subject using marginal correlations (i.e. sample covariance) or partial correlations (i.e. precision matrix).
- Key Q: group comparisons; not many studies ...
- Example: a rs-fMRI dataset (Wozniak et al 2013);
  Group 1: patients with fatal alcohol spectrum disorder (FASD), $n_1 = 24$;
  Group 2: controls, $n_2 = 31$;
  $N = 62 + 12 = 74$ cortical and sub-cortical ROIs; $k = 2701$ possible edges;
  Each subject measured at 180 time points;

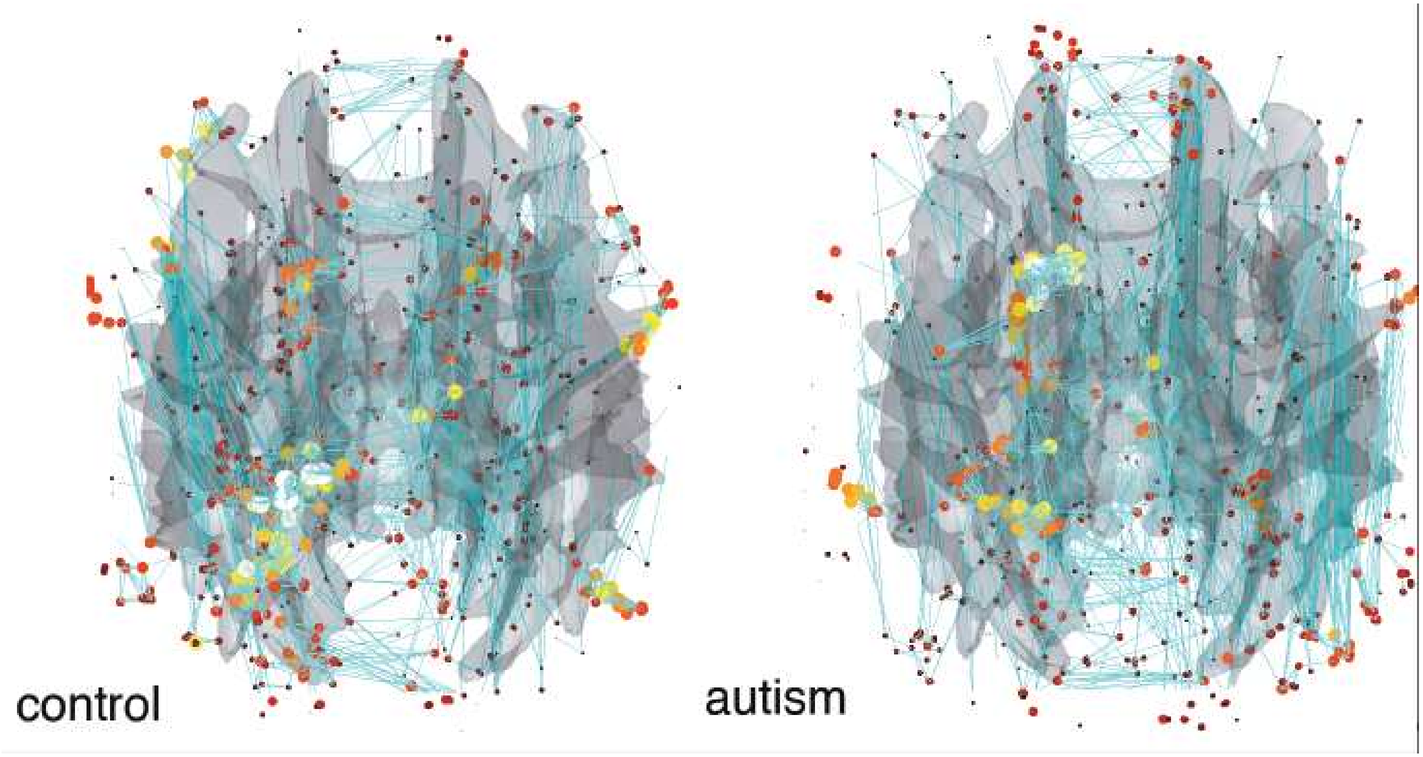


Figure 1: Structural networks (from DTI); taken from Moo Chung's website at UW-Madison.

Table 1: 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	DiProPerm	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

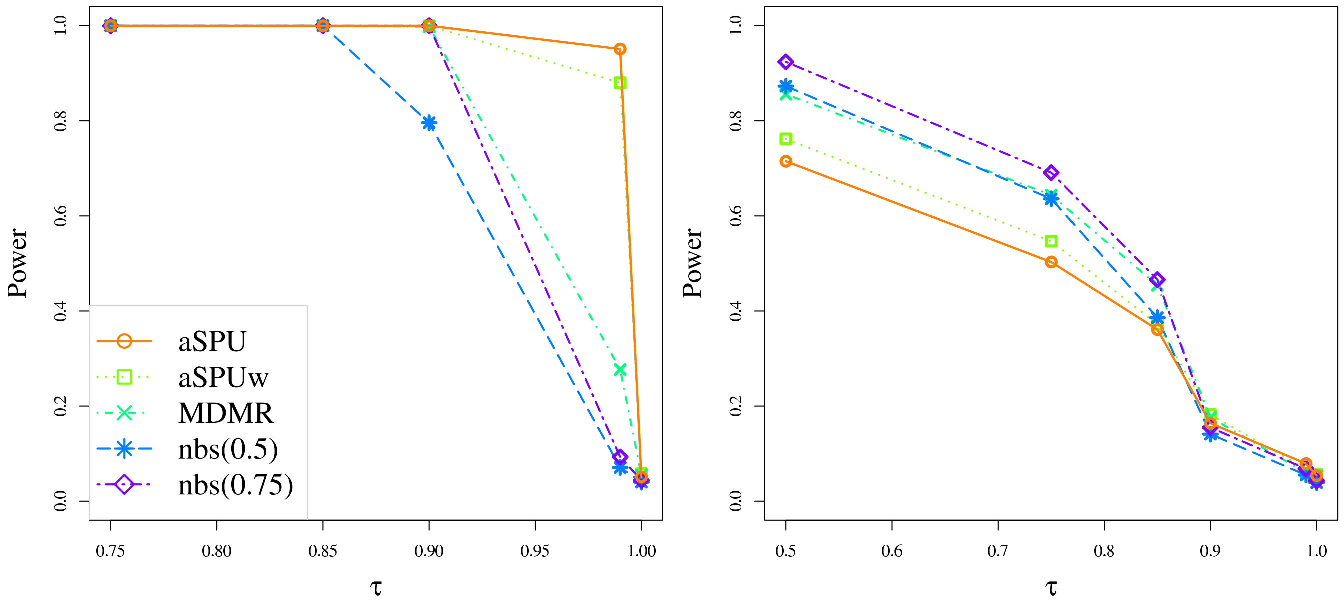


Figure 2: Sparse networks: empirical Type I error (for $\tau = 1$) and power (for $\tau < 1$) based on 1000 simulations.

Acknowledgement: This research was supported by NIH.

You can download our papers from
http://www.biostat.umn.edu/rrs.php

Thank you!