A Multi-marker Test for Analyzing Paired Transplant Genetic Data

Victoria L Arthur[1], Zhengbang Li[1,2], Rui Cao[3], Marylyn D. Ritchie[4], Weihua Guan[3], and Jinbo Chen[1]

1 Departments of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, 2 Departments of Statistics, Central China Normal University, Wuhan, 430079, China, 3 Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA, 4 Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-6116, USA

DBEI & CCEB RESEARCH DAY

Introduction

Evidence suggests that donor/recipient (D/R) matching in some genetic regions may impact transplant outcomes[1,2]. Most available matching scores account for single-nucleotide polymorphism (SNP) matching only or matching across a long range of different gene regions, making it hard to interpret association findings. In this work, we propose a multi-marker method, the Joint Score Test (JST), to jointly test for association between R genotype SNP effects and a gene-based matching score with transplant outcome. Additionally, we use a penalized testing method to test for association of a gene-based matching score with transplant outcome while adjusting for possible R genotype SNP effects.

Model and Notation

GLM for outcome Y_i $(i = 1, \dots, n)$:

$$g(\mu) = \alpha_0 + \boldsymbol{W_i}\boldsymbol{\alpha} + \boldsymbol{X_i}\boldsymbol{\beta} + Z_i\gamma$$

- $g()$: link function
- $\mu = E(Y)$
- $\boldsymbol{W_i} = (W_{i1}, \dots, W_{iK})$: vector of *K* covariates for D/R pair *i*
- $\boldsymbol{X_i} = (X_{i1}^R, \dots, X_{im}^R)$: R genotype vector of *m* SNPs for recipient *i*
- Z_i: single, gene-based genetic matching score value for D/R pair *i*

Null hypotheses of interest:

$$H_0\text{: } \boldsymbol{\beta} = 0 \text{ and } \gamma = 0$$

and

$$\gamma = 0$$

Methods

Gene Based Scores:

$$Z_i = \sum_{j=1}^{m} D(X_{ij}^D, X_{ij}^R),$$

where $D(X_{ij}^D, X_{ij}^R)$ is a measured distance between the D and R genomes

Distance Measures:

Allogenomics Mismatch Score[3]

$$D_{AMS} = \sum_{a \in X_{ij}^D} \begin{cases} 0 \text{ if } a \in X_{ij}^R \\ 1 \text{ otherwise} \end{cases}$$

Where *a* denotes alleles of a genotype

Binary Mismatch Score[4]

$$D_{MM} = \begin{cases} 1 \text{ if } \exists\, a \in X_{ij}^D \text{ such that } a \notin X_{ij}^R \\ \qquad 0 \text{ otherwise} \end{cases}$$

IBS Mismatch Score

$$D_{IBS} = |X_{ij}^D - X_{ij}^R|$$

Incompatibility Score

$$D_{Incomp} = \begin{cases} 1 \text{ if} X_{ij}^D \neq X_{ij}^R \\ 0 \text{ otherwise} \end{cases}$$

Joint Score Test (JST)

- $\hat{p}_1(\boldsymbol{W_i}) \equiv p(Y_i = 1 | \boldsymbol{W_i}; \hat{\alpha}_0; \hat{\boldsymbol{\alpha}})$: predicted probability of $Y_i = 1$ based on the null model:

$$logit\ Pr(Y_i = 1) = \alpha_0 + \sum_{k=1}^{K} W_{ik}\alpha_k \equiv \alpha_0 + \boldsymbol{W_i}\boldsymbol{\alpha}$$

- $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$: maximum likelihood estimates of α_0 and $\boldsymbol{\alpha}$
- $\hat{X}_{ij}$: fitted value from $X_{ij} = \theta_0 + \sum_{k=1}^{K} W_{ik}\theta_k$
- $\hat{Z}_i$: fitted value from $Z_i = \tau_0 + \sum_{k=1}^{K} W_{ik}\tau_k$
- Weights for above models are $\hat{p}_1(\boldsymbol{W_i})\{1 - \hat{p}_1(\boldsymbol{W_i})\}$ for R *i* or D/R pair *i*

Methods II

- Define $\boldsymbol{B_i} = (\boldsymbol{X_i}, Z_i)$ and $\hat{\boldsymbol{B}}_i = (\hat{\boldsymbol{X}}_{\boldsymbol{i}}, \hat{Z}_i)$
- $\boldsymbol{U} = (\boldsymbol{B} - \hat{\boldsymbol{B}})\{\boldsymbol{Y} - \hat{\boldsymbol{p}}_{\boldsymbol{1}}\}$, where $\boldsymbol{U}$ is the vector of likelihood score statistics for all R SNPs and the matching score
- $\boldsymbol{U}$ is asymptotically distributed as $\boldsymbol{\mathcal{N}_{m+1}}\ (\boldsymbol{0}, \boldsymbol{V})$
- Construct Hotelling's T^2 statistic as

$$n\boldsymbol{U}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{U} \sim \boldsymbol{\chi_{m+1}^2}$$

- Can improve power for large *m* by eliminating $\hat{\boldsymbol{V}}^{-1}$
- $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V^R} & \boldsymbol{C^{RS}} \\ \boldsymbol{C^{SR}} & V^S \end{bmatrix} = \begin{bmatrix} Var(\boldsymbol{U^R}) & Cov(\boldsymbol{U^R}, U^S) \\ Cov(U^S, \boldsymbol{U^R}) & Var(U^{\boldsymbol{S}}) \end{bmatrix}$
- JST is based on Eigen decomposition of $\boldsymbol{V^R}$
- $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \dots, \boldsymbol{a_m}]$: $m \times m$ matrix of eigenvectors of $\hat{\boldsymbol{V}}^{\boldsymbol{R}}$ with eigenvalues $(\lambda_1, \lambda_2, \dots, \lambda_m), \lambda_1 \geq \cdots \geq \lambda_m$
- Extract first $s < m$ PCs, $\boldsymbol{A_s} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \dots, \boldsymbol{a_s}]$
- Define $\boldsymbol{U^{PR}}$: vector of $\boldsymbol{U^{R\prime}} a_l / \sqrt{\lambda_l}, l = 1,2, \dots, s$
- JST is constructed as

$$\begin{pmatrix} \boldsymbol{U^{PR}} \\ U^S \end{pmatrix}^T \begin{bmatrix} I_{s\times s} & Cov(\boldsymbol{U^{PR}}, U^S) \\ Cov(U^S, \boldsymbol{U^{PR}}) & Var(U^S) \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{U^{PR}} \\ U^S \end{pmatrix}$$

- JST is asymptotically distributed as $\boldsymbol{\chi_{s+1}^2}$

Penalized Score Test[5]

- Define $\boldsymbol{X}^* = \{\boldsymbol{1}, \boldsymbol{W}, \boldsymbol{X}, Z\}, n \times p$ matrix, $(p = k + m + 2)$
- $\boldsymbol{\omega} = \{\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}. \gamma\}$, *p*-dimensional vector
- PDF of *Y* in exponential form:

$$\exp\left(\frac{Y_i\boldsymbol{X_i^*}\boldsymbol{\omega} - b(\boldsymbol{X_i^*}\boldsymbol{\omega})}{\phi_0}\right) c(Y)$$

- General null hypothesis:

$$\mathbf{C}\boldsymbol{\omega_{0,M}} = \boldsymbol{t}$$

- Our null hypothesis: $\boldsymbol{\omega_{0,M}} = 0$, where $\boldsymbol{\omega_{0,M}} = \gamma$

Methods III

- Partially penalized likelihood function:

$$L_n(\boldsymbol{\omega}, \lambda) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i\boldsymbol{X_i^*}\boldsymbol{\omega} - b(\boldsymbol{X_i^*}\boldsymbol{\omega})\} - \sum_{j \notin M} p_\lambda(|\omega_j|)$$

- $p_\lambda()$: penalty function with tuning parameter λ
- Estimates of $\boldsymbol{\omega}$ under H_0 and H_a:

$$\hat{\boldsymbol{\omega}}_{\boldsymbol{0}} = \arg\max_{\boldsymbol{\omega}} L_n(\boldsymbol{\omega}, \lambda_{n,0}) \quad \text{subject to } \boldsymbol{\omega_{0,M}} = \boldsymbol{0},$$

$$\hat{\boldsymbol{\omega}}_{\boldsymbol{a}} = \arg\max_{\boldsymbol{\omega}} L_n(\boldsymbol{\omega}, \lambda_{n,a}).$$

- Forced penalties for $\{\alpha_0, \boldsymbol{\alpha}\}$ to be 0 so only elements of $\boldsymbol{\beta}$ were penalized
- Penalized score test statistic (T_S)

$$\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\}^T \begin{pmatrix} \boldsymbol{X^*}_M \\ \boldsymbol{X^*}_{\hat{S}_0} \end{pmatrix} \hat{\boldsymbol{\Omega}}_0 \begin{pmatrix} \boldsymbol{X^*}_M \\ \boldsymbol{X^*}_{\hat{S}_0} \end{pmatrix}^T \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\}/\hat{\phi},$$

- $\hat{S}_0 = \{j \in M^C : \hat{\omega}_{0,j} \neq 0\}$
- $\hat{\boldsymbol{\Omega}}_0 =$

$$n\begin{pmatrix} \boldsymbol{X_M^{*T}}\boldsymbol{\Sigma}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\boldsymbol{X_M^*} & \boldsymbol{X_M^{*T}}\boldsymbol{\Sigma}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\boldsymbol{X^*}_{\hat{S}_0} \\ \boldsymbol{X^{*T}}_{\hat{S}_0}\boldsymbol{\Sigma}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\boldsymbol{X_M^*} & \boldsymbol{X^{*T}}_{\hat{S}_0}\boldsymbol{\Sigma}(\boldsymbol{X}^*\hat{\boldsymbol{\omega}}_{\boldsymbol{0}})\boldsymbol{X^*}_{\hat{S}_0} \end{pmatrix}^{-1}$$

- For a fixed number of constraints, *r*, and consistent estimator $\hat{\phi}$ for ϕ_0, $T_S \sim \chi_r^2$

Simulations

Study Design

- Datasets for 3 gene regions (*NAT2*, *CHI3L2*, *ASAH1*) were sampled from 1000 Genomes Phase 3 reference using HapGen2[6]
- Sample size: *n* = 500 or 1000 D/R pairs
- 5000 simulations for each gene and *n*
- *s* values account for 85, 90, 95, 99% total variance explained by PCs

A Multi-marker Test for Analyzing Paired Transplant Genetic Data

Victoria L Arthur[1], Zhengbang Li[1,2], Rui Cao[3], Marylyn D. Ritchie[4], Weihua Guan[3], and Jinbo Chen[1]

1 Departments of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, 2 Departments of Statistics, Central China Normal University, Wuhan, 430079, China, 3 Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA, 4 Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-6116, USA

DBEI & CCEB RESEARCH DAY

Simulations Continued

- Options for power analyses:
  - 5, 15, 25% R genotype SNPs associated with outcome, *Y*
  - 5, 15, 25, 50, 75, 100% D/R matching associated with outcome, *Y*
  - Associated SNPs in low or high LD
  - Small (1.25), Medium (1.50), Large (2.00) OR per SNP or matching score
  - Outcome prevalence of 5, 10, 20%
- Compared JST to:
  - Standard GLM
  - SKAT
  - Penalized score test

Simulation Results

Method	Score	Prev. 20	Prev. 10	Prev. 5	Cont.
JST (s = 85%)	IBS	0.05	0.06	0.04	0.05
	Incompatibility	0.05	0.05	0.05	0.05
	AMS	0.05	0.06	0.05	0.05
	Binary MM	0.05	0.06	0.05	0.05
SKAT*	IBS	0.05	0.05	0.05	0.05
	Incompatibility	0.05	0.05	0.05	0.05
	AMS	0.05	0.05	0.05	0.05
	Binary MM	0.05	0.05	0.05	0.05
GLM	IBS	0.22	0.17	0.16	0.05
	Incompatibility	0.22	0.15	0.15	0.05
	AMS	0.22	0.17	0.16	0.05
	Binary MM	0.23	0.19	0.15	0.06
Pen. Score	IBS	0.05	0.07	0.11	0.08
	Incompatibility	0.05	0.08	0.11	0.08
	AMS	0.05	0.09	0.11	0.09
	Binary MM	0.05	0.08	0.10	0.10

Table 1: Results of Type I Error simulations for JST using the gene *NAT2* with 500 D/R pairs. Score refers to which score was fit as Z_i. Results were similar for JST with *s* values of 90, 95, 99% variance explained. *SKAT was fit using an unweighted linear kernel.

Simulation Results Continued

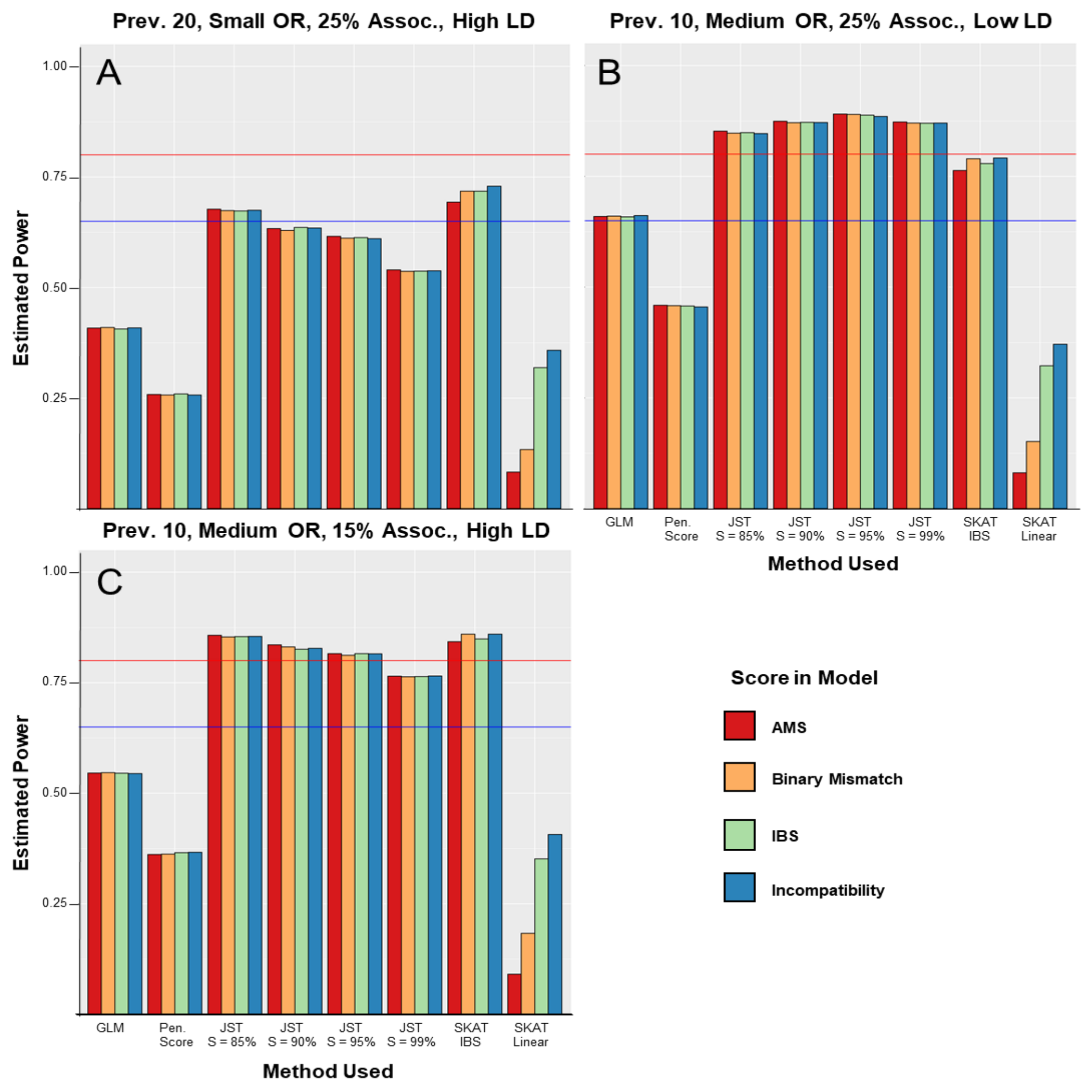


Figure 1: Power estimates from simulations using the gene NAT2 and 1000 pairs of donors and recipients under the scenario that recipient genotype SNPs were associated with outcome. The horizontal blue line corresponds to 65% power and the horizontal red line corresponds to 80% power.

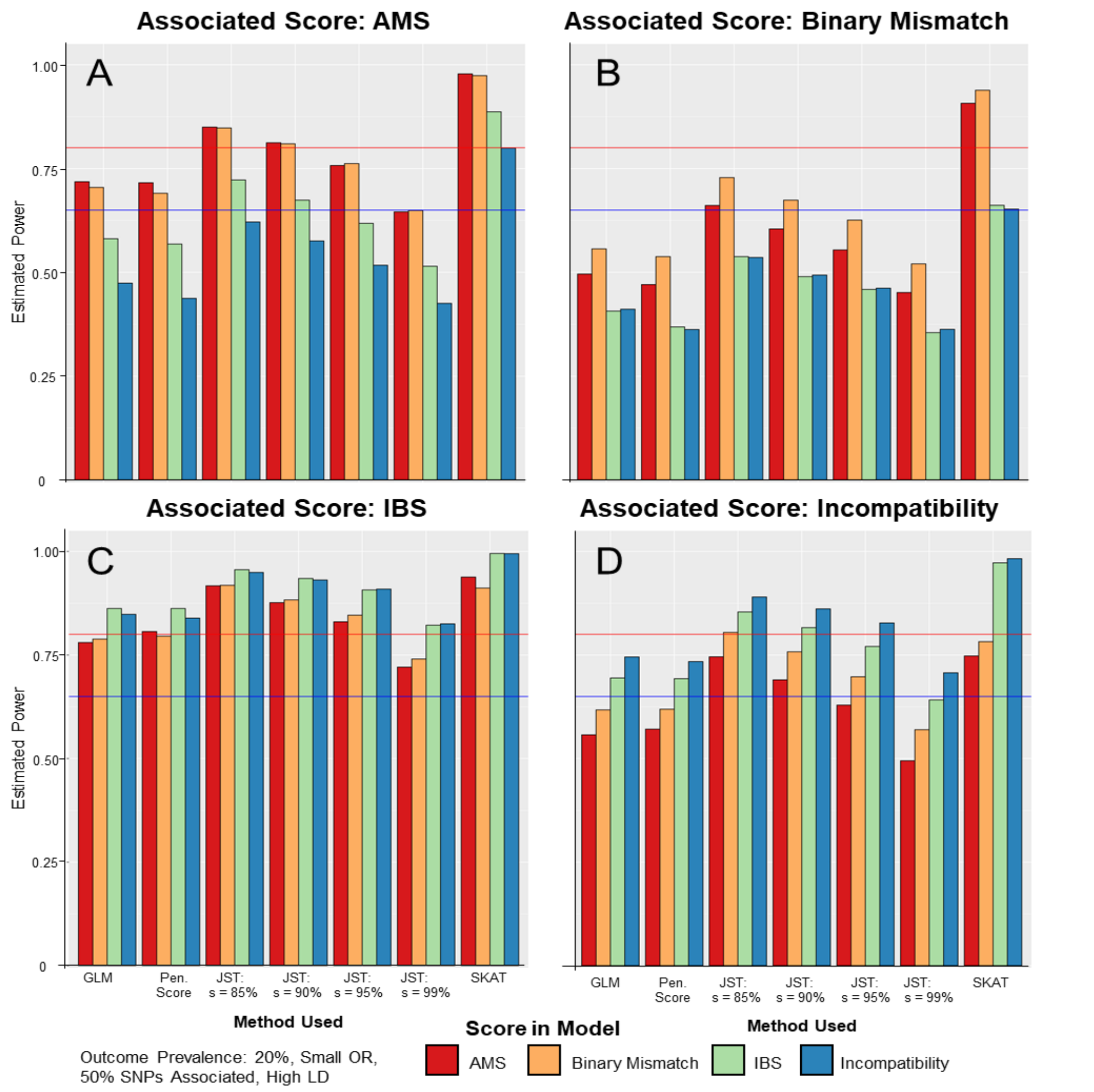


Figure 2: Power estimates from simulations using the gene *NAT2* and 1000 pairs of donors and recipients under the scenario that the gene-based score was associated with outcome. The horizontal blue line corresponds to 65% power and the horizontal red line corresponds to 80% power.

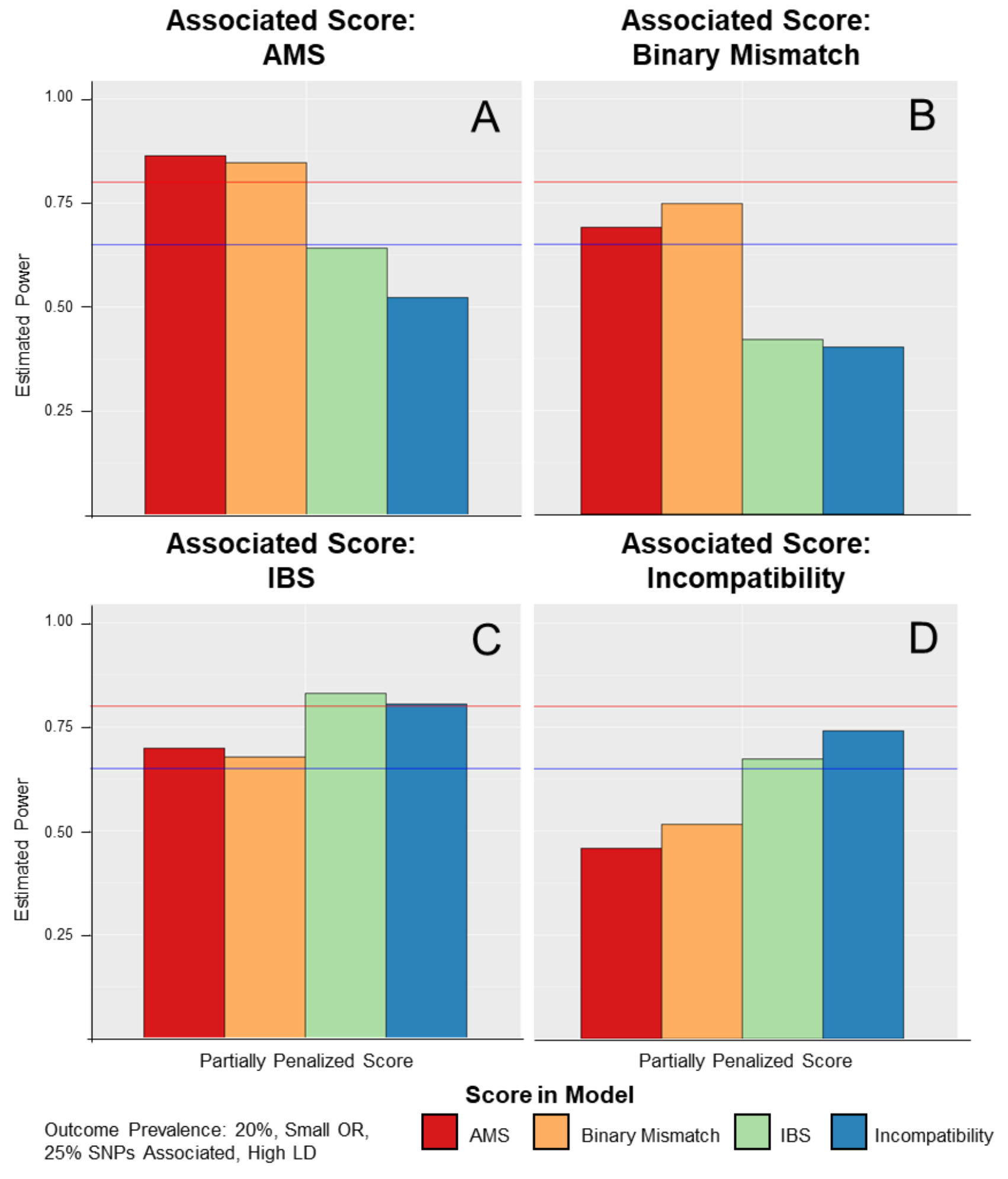
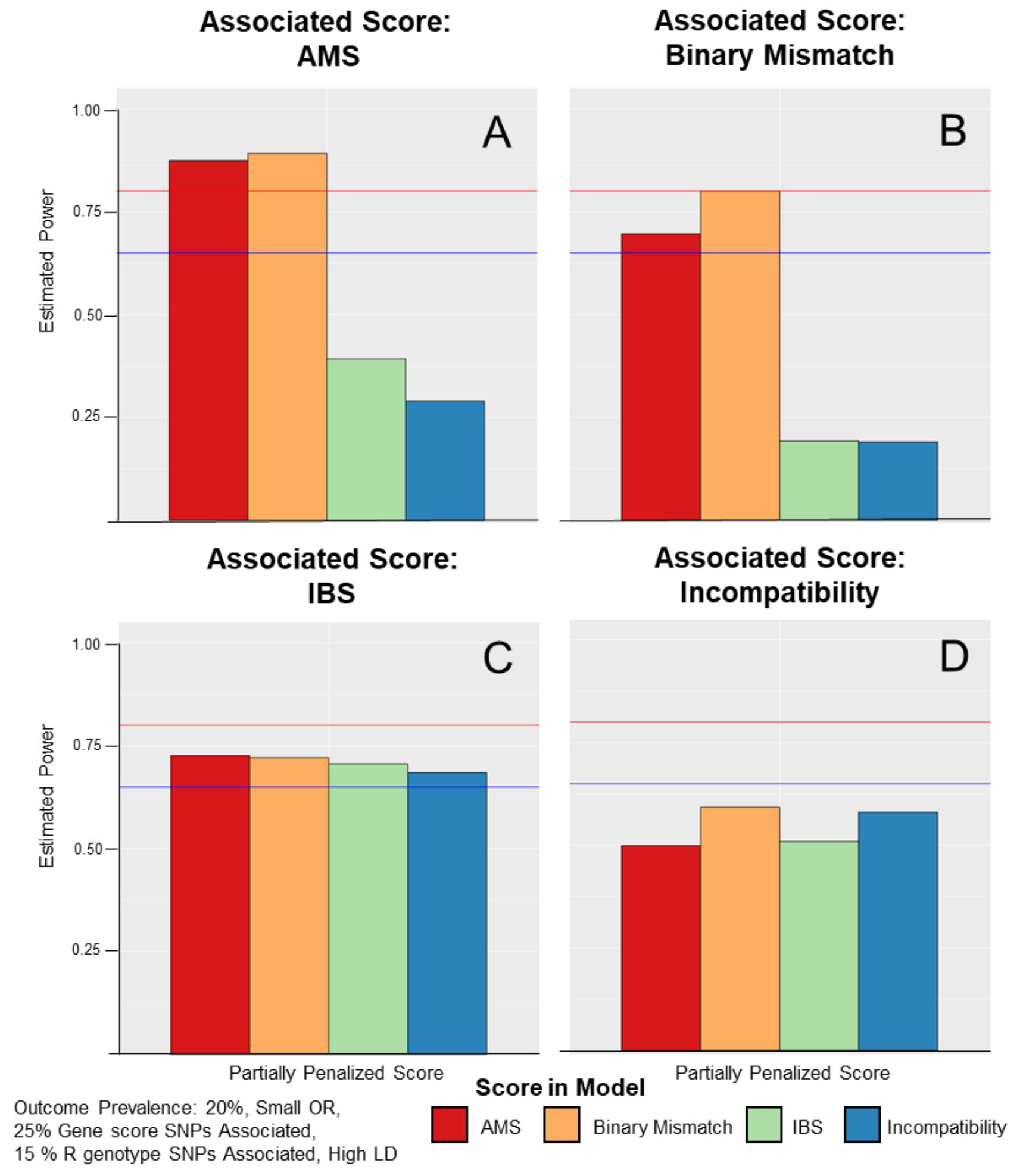


Figure 3 (Left): Estimated power plots for simulations testing whether gene-based score was associated with outcome. All models were fit using the partially penalized score test, with 1000 donor/recipient pairs. The blue line corresponds to 65% power and the horizontal red line corresponds to 80% power.

Figure 4 (Right): Estimated power plots for simulations testing whether gene-based score was associated with outcome. All models were fit using the partially penalized score test, with 1000 donor/recipient pairs. The blue line corresponds to 65% power and the horizontal red line corresponds to 80% power.



Real Data Analysis

- Samples: 404 D/R kidney transplant pairs (56 cases of Acute Rejection)
- Genome-wide SNPs (785,458 Bi-alleles),
- Grouped by 25,265 genes (physical position)

JST Results

Gene ID	IBS Score	P-value	AMS Score	P-value
AC119677.1	29.25	4.46E-07	13.08	1.44E-03
OVCH2	33.14	1.12E-06	28.71	8.95E-06

SKAT Results

Gene ID	IBS Score	P-value	AMS Score	P-value
OVCH2	454.93	7.916E-06	230.06	3.18E-04
AC119677.1	107.41	5.143E-04	32.94	2.87E-02

Matching Score Test Results

Gene ID	IBS Score	P-value	AMS Score	P-value
OVCH2	19.31	1.11E-05	12.01	5.28E-04
AC119677.1	16.19	5.74E-05	4.96	2.60E-02

Table 2: After Bonferroni correction, two genes were found to be associated in joint testing. Of these, *OVCH2* was also found to be significant using SKAT testing and the matching score only test. Results for incompatibility score and binary mismatch score match those for the nonbinary scores.

References

1. Reindl-Schwaighofer, R., Heinzel, A., Gualdoni, G. A., Mesnard, L., Claas, F. H. J., & Oberbauer, R. (2020). Novel insights into non-HLA alloimmunity in kidney transplantation. *Transplant International, 33*(1), 5–17. https://doi.org/10.1111/tri.13546
2. Marin, E. P., Cohen, E., & Dahl, N. (2020). Clinical Applications of Genetic Discoveries in Kidney Transplantation: A Review. *Kidney360, 1*(4), 300–305. https://doi.org/10.34067/KID.0000312019
3. Mesnard, L., Muthukumar, T., Burbach, M., Li, C., Shang, H., Dadhania, D., Lee, J. R., Sharma, V. K., Xiang, J., Suberbielle, C., Carmagnat, M., Ouali, N., Rondeau, E., Friedewald, J. J., Abecassis, M. M., Suthanthiran, M., & Campagne, F. (2016). Exome Sequencing and Prediction of Long-Term Kidney Allograft Function. *PLOS Computational Biology, 12*(9), e1005088. https://doi.org/10.1371/journal.pcbi.1005088
4. Reindl-Schwaighofer, R., Heinzel, A., Kainz, A., van Setten, J., Jelencsics, K., Hu, K., Loza, B.-L., Kammer, M., Heinze, G., Hruba, P., Koňaříková, A., Viklicky, O., Boehmig, G. A., Eskandary, F., Fischer, G., Claas, F., Tan, J. C., Albert, T. J., Patel, J., … Oberbauer, R. (2019). Contribution of non-HLA incompatibility between donor and recipient to kidney allograft survival: Genome-wide analysis in a prospective cohort. *The Lancet, 393*(10174), 910–917. https://doi.org/10.1016/S0140-6736(18)32473-5
5. Shi, C., Song, R., Chen, Z., & Li, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics*, 47(5), 2671–2703. https://doi.org/10.1214/18-AOS1761
6. Su, Z., Marchini, J. and Donnelly, P. (2011) HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011 Aug 15;27(16):2304-5. doi: 10.1093/bioinformatics/btr341. Epub 2011 Jun 8. PMID: 21653516; PMCID: PMC3150040.