Burden, SKAT and Optimal Unified Tests (SKAT-O) for WES Association Studies

Xihong Lin

Department of Biostatistics

Harvard School of Public Health

xlin@hsph.harvard.edu

Acknowledgement for the SKAT-O Team:

Seunggeun Lee, Mike Wu, Mary Emond, Michael Bamshad, Kathleen Barnes, Mark Rieder, Deborah Nickerson, David Christiani, Mark Wurfel

Gene-based Analysis for Rare Variant Effects

- Covariates: age, gender, PCs
- Observed SNPs in a gene: $S_1, \cdots, S_p$
- Model: continuous (linear) / binary Y (logistic):

$$E(Y) \text{ or } logit(\pi) = \alpha_0 + \boldsymbol{\alpha} Covars + \beta_1 S_1 + \cdots + \beta_p S_p$$

- Goal (Test for no genetic effect):

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

Burden Tests

- Collapse rare variants (with MAF $< c$, e.g., c=5%)
- If all β's are the same, the model becomes

$$\text{logit}(\pi) = \alpha_0 + \boldsymbol{\alpha} \mathit{Covariates} + \beta \mathit{Dose},$$

  where $\mathit{Dose} = S_1 + \cdots + S_m$=total number of rare variants in a gene.
- Various versions: CMC, MB, MZ, VT.
- Assumption: All rare variants are causal and have the same effects.

SKAT (Wu, et al, AJGH, 2011)

- Key feature (Robustness): Allow no effect of most rare variants or effects with different directions.
- SKAT: Aggregate weighted individual variant score test statistics and justified using mixed models.

$$Q_{SKAT} = w_1 U_1^2 + \cdots w_p U_p^2$$

where U's are individual SNP score statistics.

SKAT

- Easily adjust for covariates.
- P-values are quickly calculated analytically.
- Computing speed: 6 hours for the whole exome scan using a 2.33 GHz laptop.
- The C-alpha test (Neale, et al, 2011) is a special case.

Optimal Unified Test (SKAT-O)

- Burden test is more powerful when a large % of variants are causal and effects are in the same direction.
- SKAT is more powerful when a small % of variants are causal, or the effects have mixed directions.
- Both scenarios can happen when scanning the genome.

Optimal Unified Test (SKAT-O)

- Burden test is more powerful when a large % of variants are causal and effects are in the same direction.
- SKAT is more powerful when a small % of variants are causal, or the effects have mixed directions.
- Both scenarios can happen when scanning the genome.

Which test to use?

Optimal Unified Test

- Unified test:

$$Q = \rho Q_{Burden} + (1 - \rho) Q_{SKAT}, \quad 0 \leq \rho \leq 1.$$

- SKAT($\rho = 0$) and Burden ($\rho = 1$) are special cases.
- Interpretation of ρ: correlation among β's.
- Optimal Unified Test (SKAT-O):

  Use data to adaptively estimate ρ to maximize the power, and calculate p-values analytically.

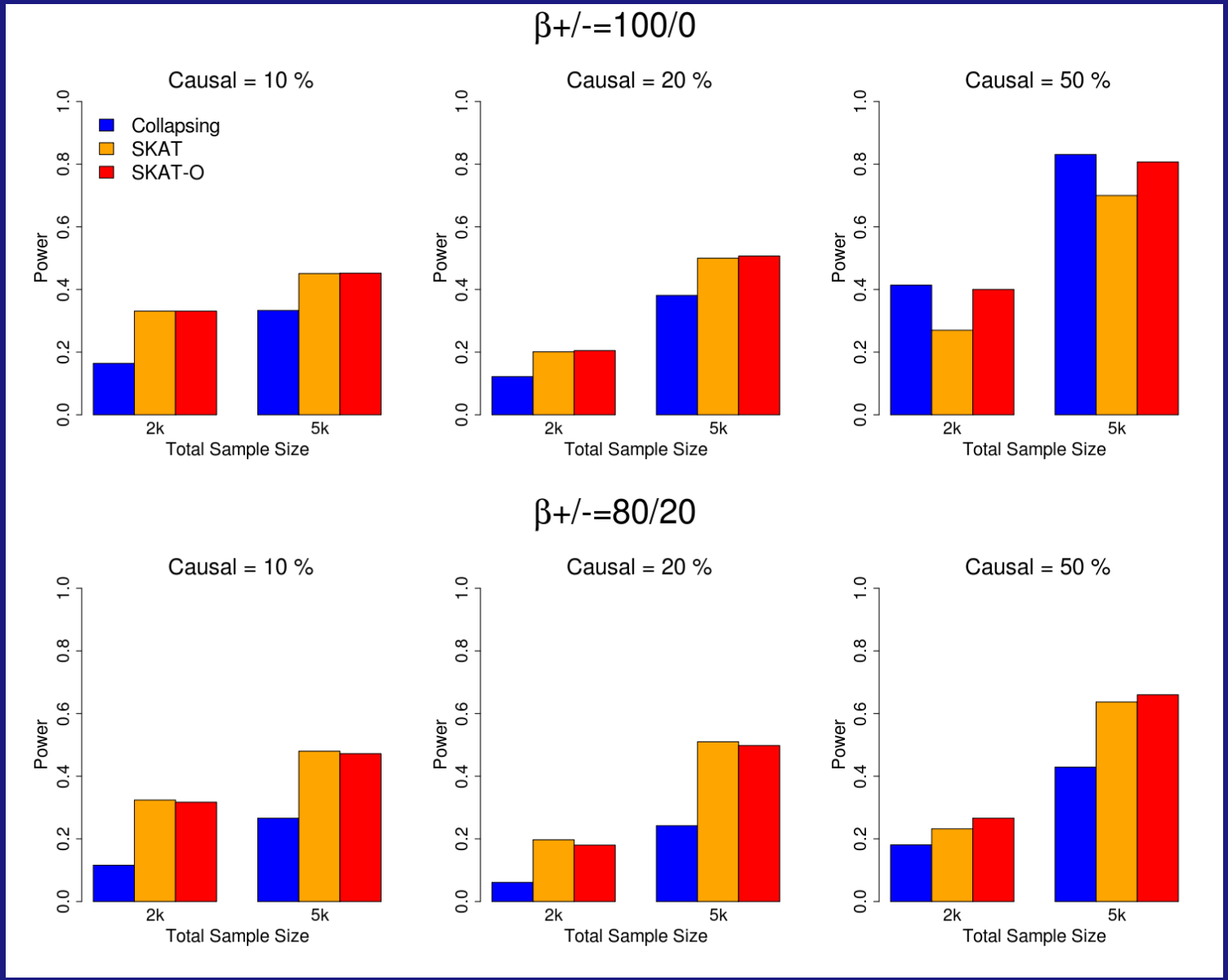
Simulations (SKAT): Genome-wide type I error

$\alpha = 10^{-6}$

Total Sample Size	Continuous Trait	Binary Trait
500	5.9×10^{-7}	1.0×10^{-8}
1000	8.0×10^{-7}	2.3×10^{-7}
2500	8.4×10^{-7}	5.6×10^{-7}
5000	8.8×10^{-7}	7.0×10^{-7}

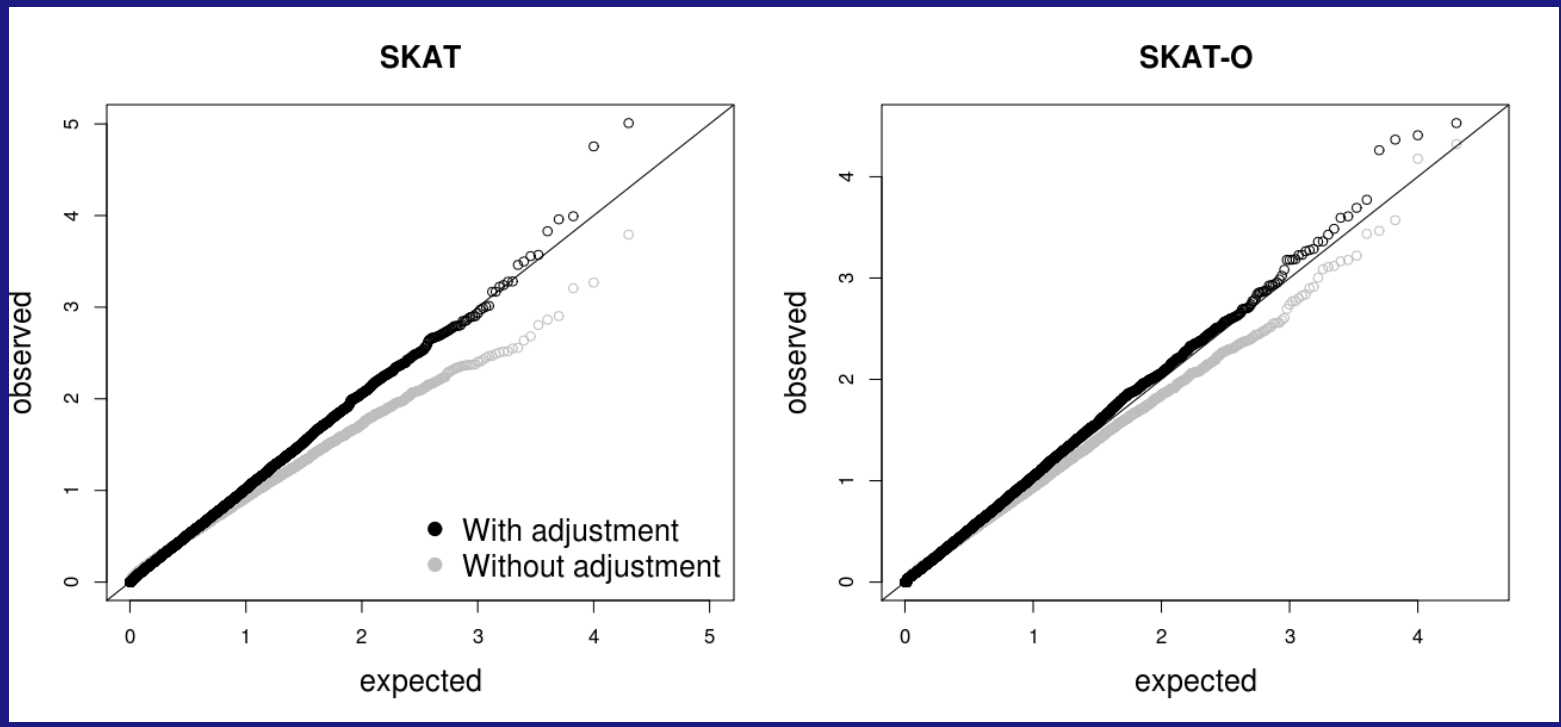
SKAT is conservative for binary traits with small n.

Power: Continuous traits, $\alpha = 2.5 \times 10^{-6}$

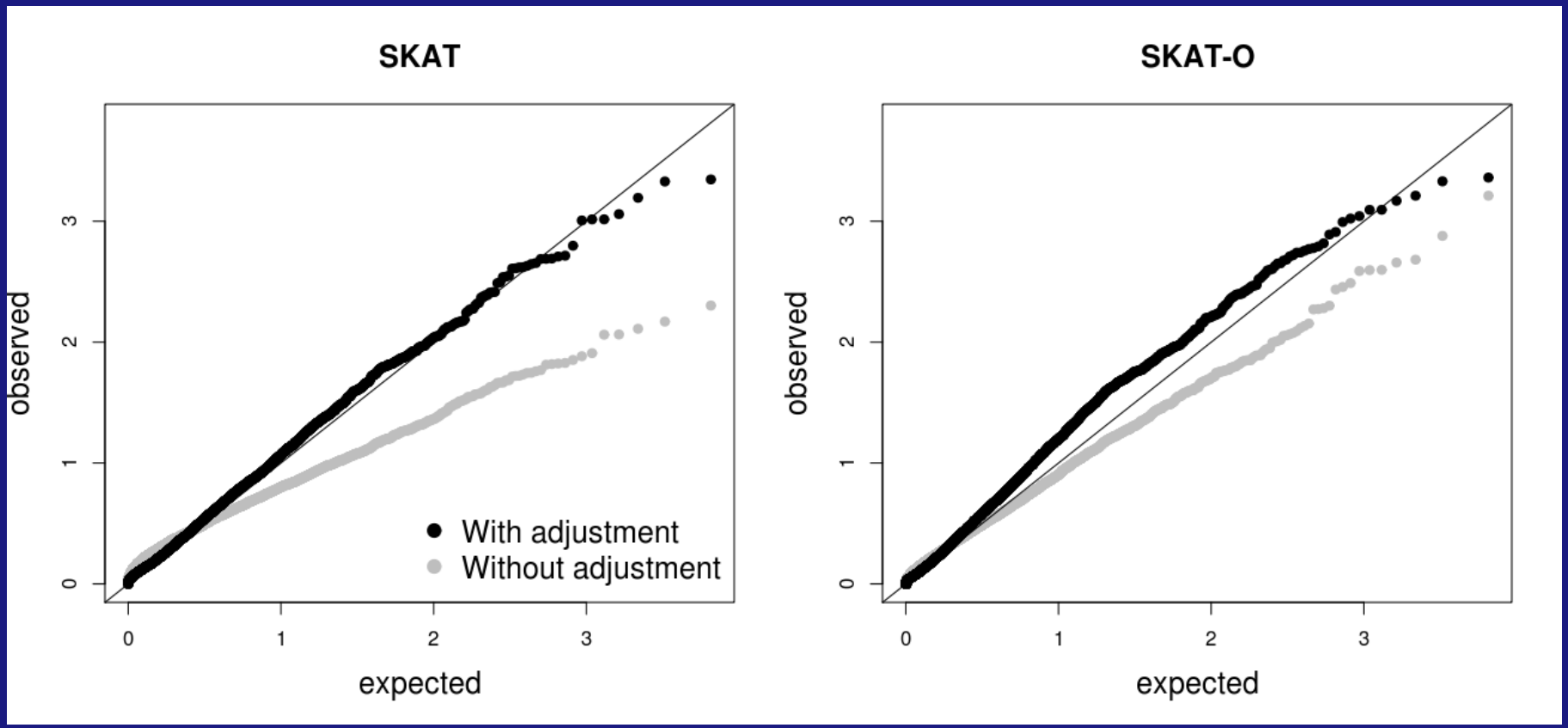


Analytic small sample correction for SKAT/SKAT-O

- Current WES, e.g. in ESP, have small n.
- Large sample based p-values are conservative.
- n =100 cases/100 controls with no genetic effect.



QQ Plot of the NHLBI ALI Exome Seq Study (43/45 cases/controls)



SKAT Package

- SKAT and SKAT-O Software:
  http://www.hsph.harvard.edu/∼xlin/software
- Phenotypes: Binary, continuous and extreme phenotypes.
- Perform analytic power and sample size calculations for designing WES and WGS.

Extreme Phenotype Sampling Enriches Rare Variants and Powers

- RS: Random Sampling
- EDP: Extreme Dichotomized Phenotype
- ECP: Extreme Continuous Phenotype