Rare-variant association tests in large-scale biobanks and cohorts

Wei Zhou

Post-doctoral Fellow

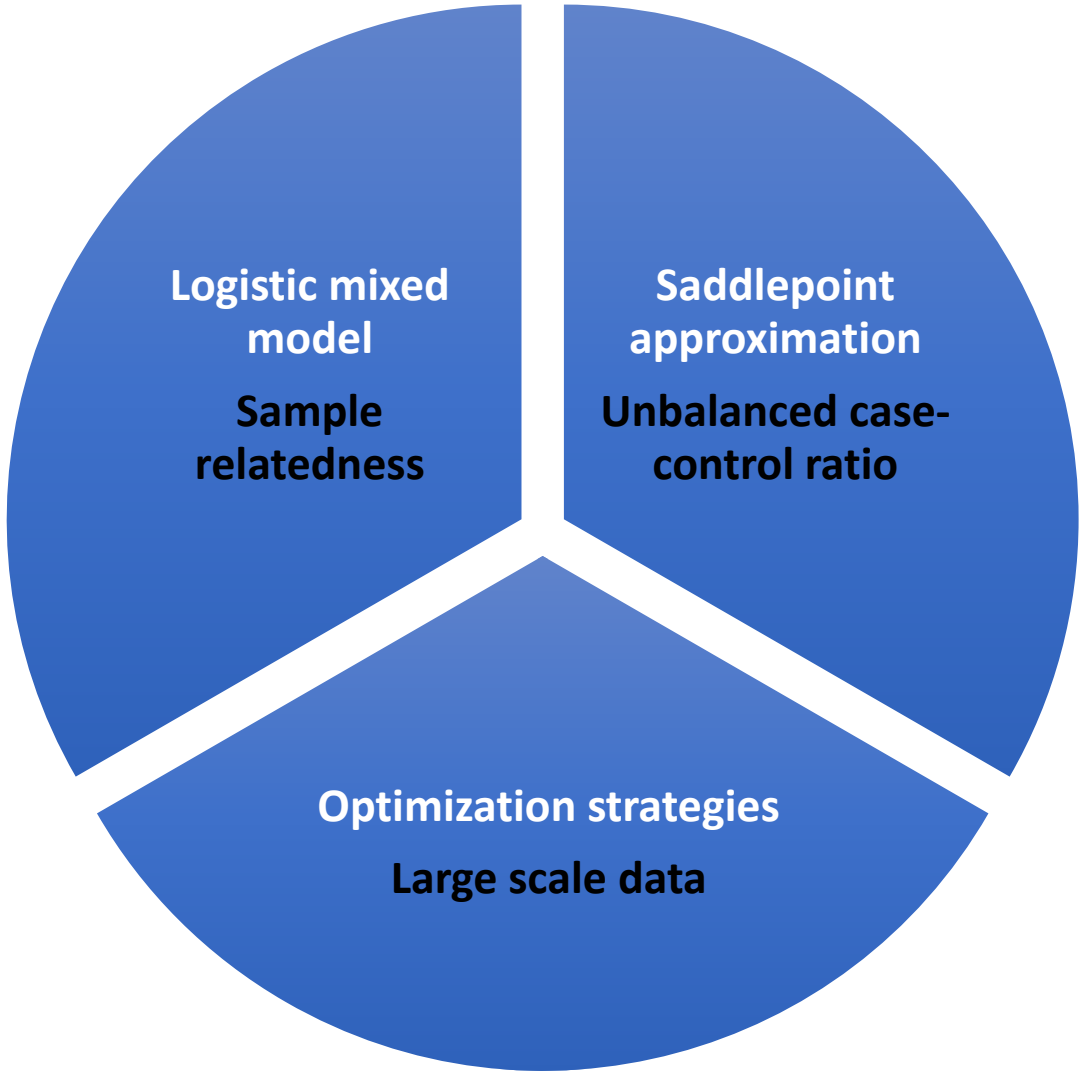
Massachusetts General Hospital, Harvard Medical School, Broad Institute

Outline

- Challenges of rare-variant association tests in large-scale cohorts/biobanks (mostly for binary phenotypes)

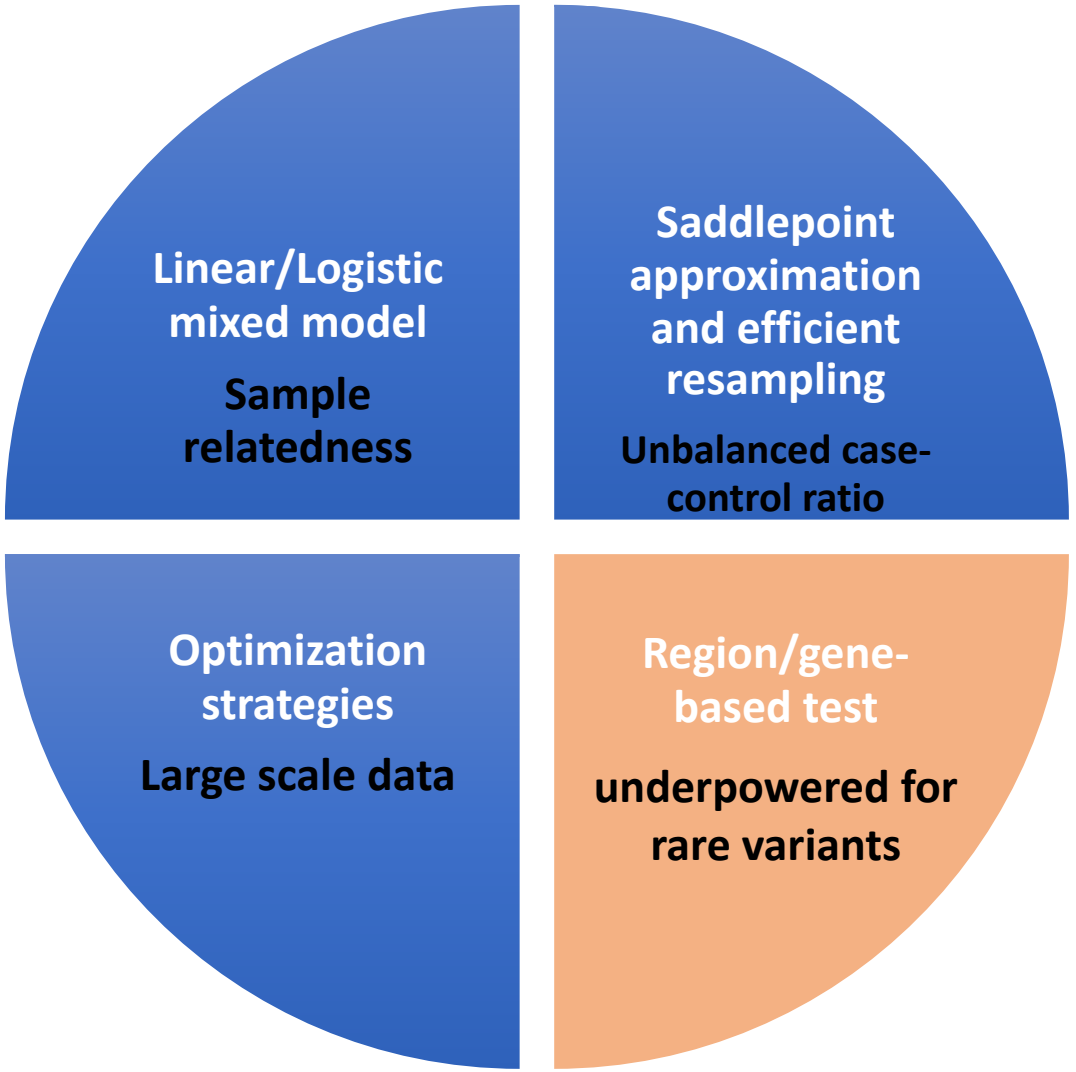
- SAIGE-GENE: Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts

Recall: GWAS in large-scale biobanks and cohorts

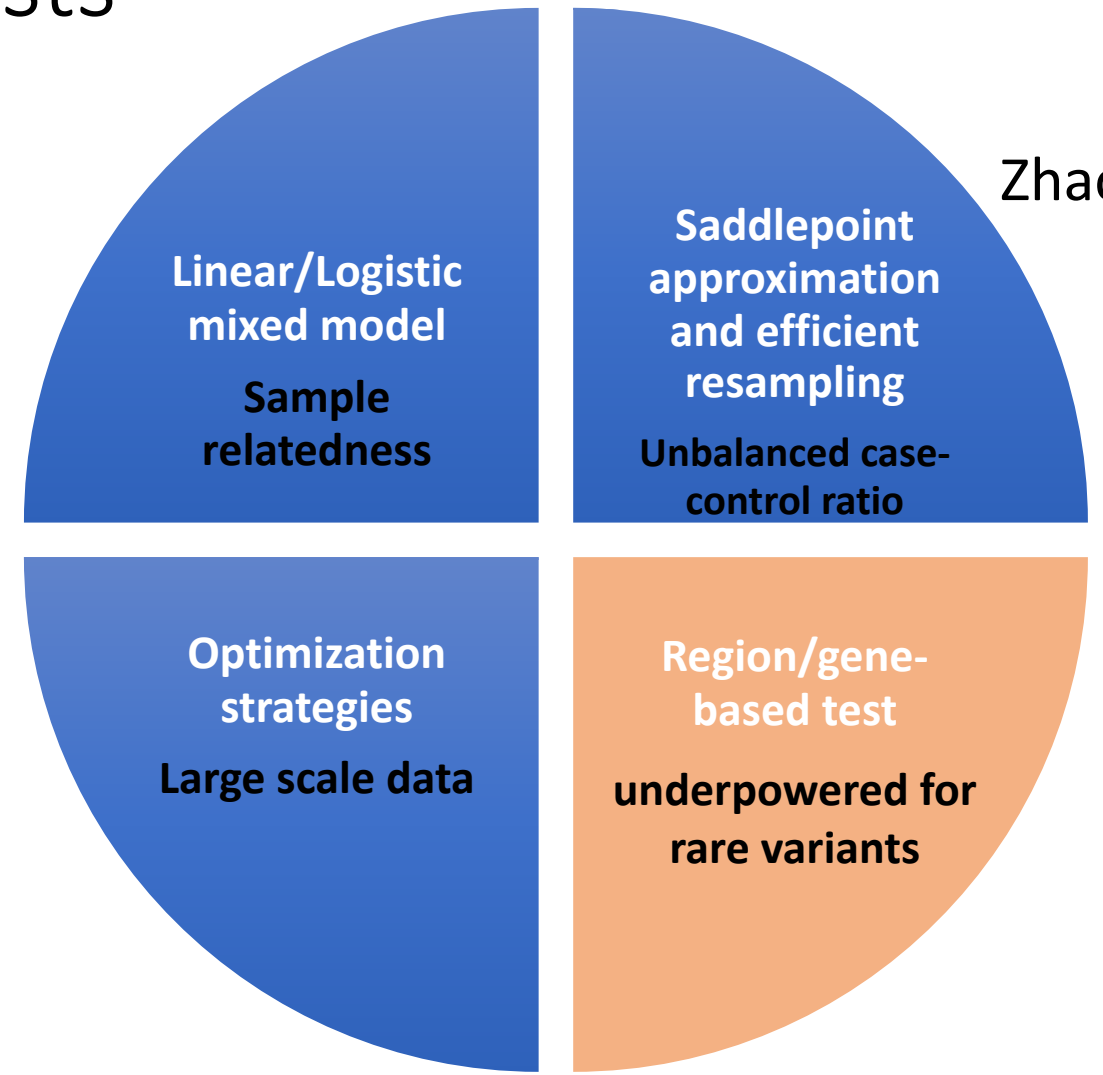


SAIGE: Zhou et al., 2018

Single-variant association tests are underpowered for very rare variants

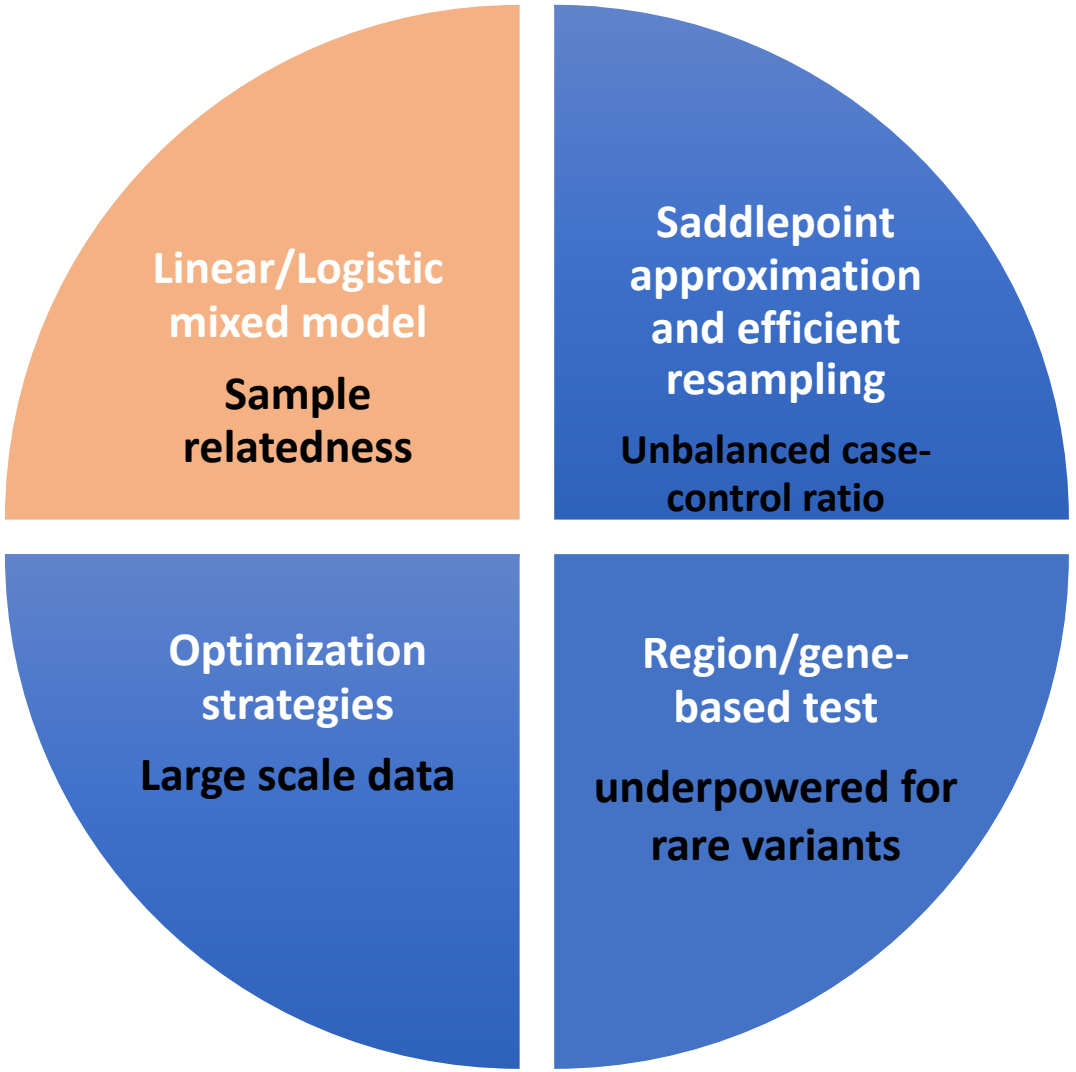


Applying saddlepoint approximation to account for unbalanced case-control ratios in set-based rare variant association tests



Zhao et al, 2020

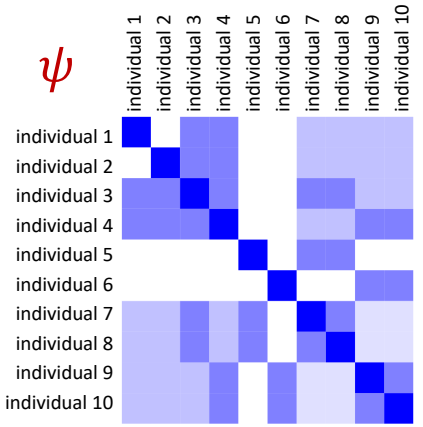
Accounting for sample relatedness using mixed models



EmmaX-SKAT: Linear mixed-model for set-based tests

$$y = X\alpha + G\beta + b + \epsilon$$

- b: random genetic effect, $b \sim N(0, \tau\psi)$
- ψ: N x N **genetic relationship matrix (GRM)**

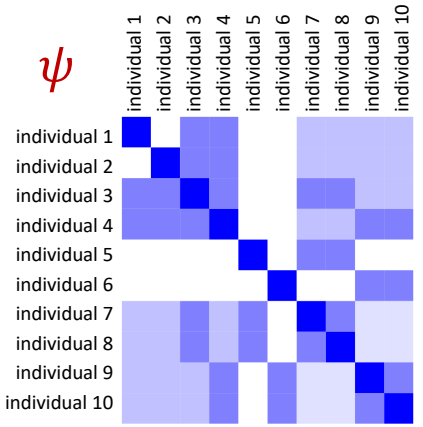


Lee *et al.* 2012

EmmaX-SKAT: Linear mixed-model for set-based tests

$$y = X\alpha + G\beta + b + \epsilon$$

- b: random genetic effect, $b \sim N(0, \tau\psi)$
- ψ: N x N **genetic relationship matrix (GRM)**



- Score statistics of marginal model for variant j is $S_j = G_j{}'\hat{P}Y$

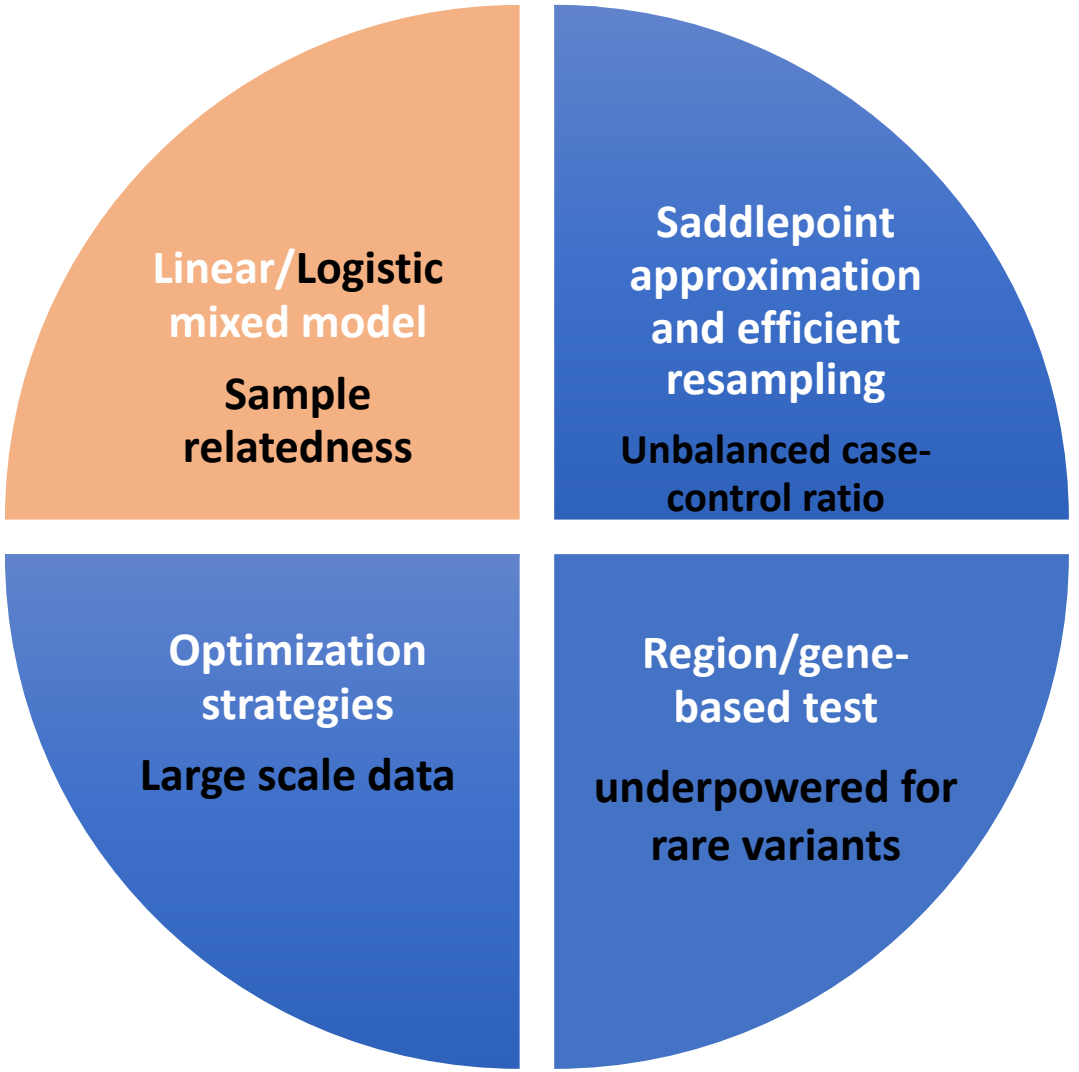
$$Q_{BURDEN} = \left(\sum_{j=1}^{q} w_j S_j\right)^2 \qquad Q_{SKAT} = \sum_{j=1}^{q} w_j^2 S_j^2$$

$$Q_{SKATO} = (1-\rho)Q_{SKAT} + \rho Q_{BURDEN}, 0 \leq \rho \leq 1$$

Lee *et al.* 2012

SMMAT: Logistic mixed model for set-based rare variant test for **binary phenotypes**

Chen *et al.* 2019



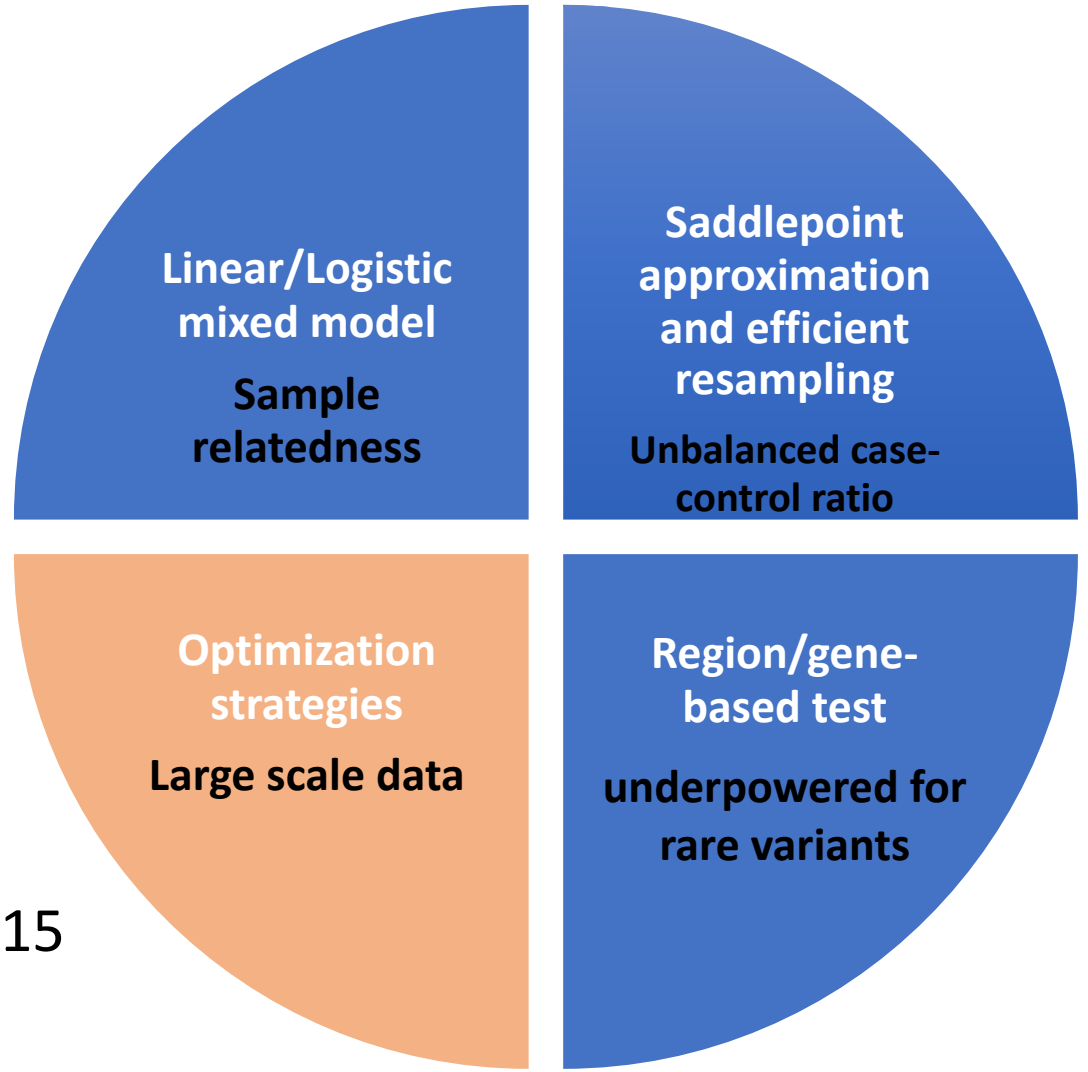
Heavy computation burden in EmmaX-SKAT and SMMAT

- **Bottleneck for computation cost:**
  - To obtain p-value for association, $G'\hat{P}G$ needs to be computed for each variant set
  - Computing $\hat{P}$ requires ψ^{-1}

EmmaX-SKAT needs
> 11 CPU years, ~ 1 Tb
for genome-wide region-based tests (16k sliding windows)
on **one phenotype in UK Biobank**

Lee *et al.* 2012

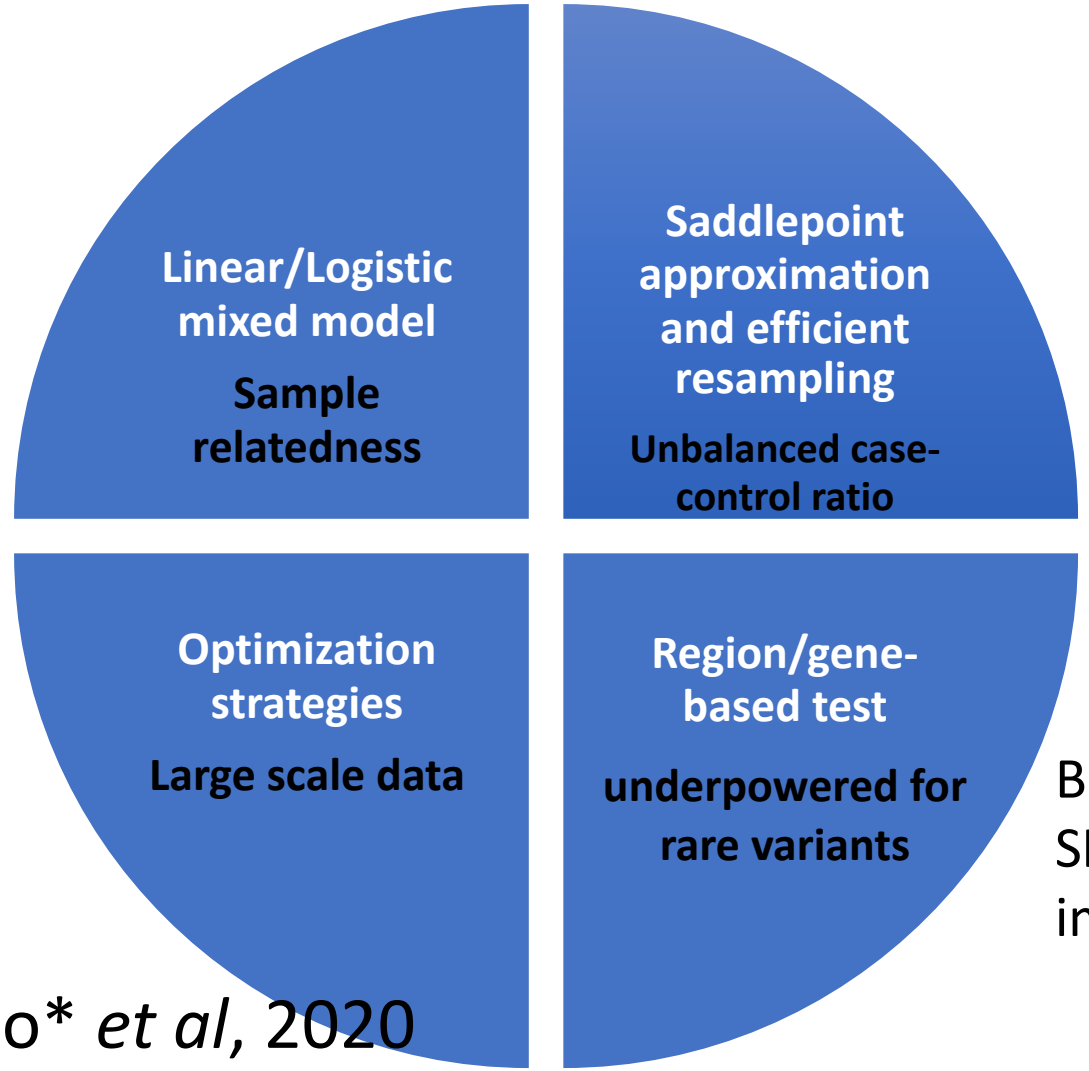
Using optimization strategies to reduce the computation cost



SAIGE: Zhou *et al.* 2018
BOLT-LMM: Loh *et al*, 2015

SAIGE-GENE

Scalable and **A**ccurate **I**mplementation of **GE**neralized mixed model

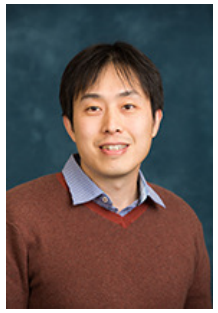


Burden, SKAT, and SKAT-O are implemented

Zhou* and Zhao* *et al*, 2020

Teamwork

Zhangchen Zhao



Seunggeun Shawn Lee

Cristen Willer

Seoul National University

Mark Daly

Benjamin Neale

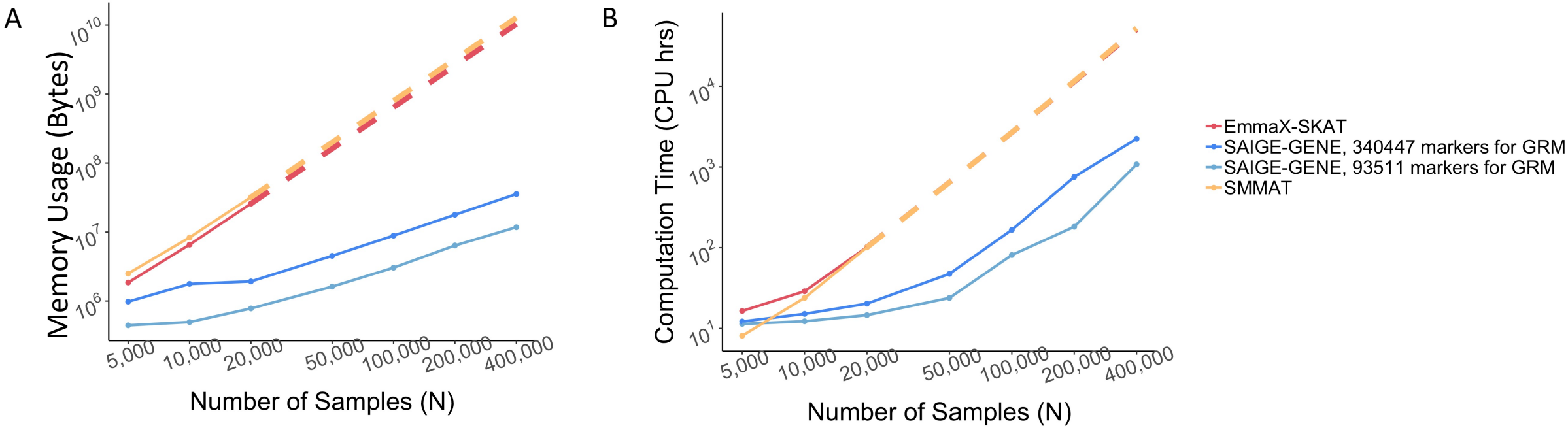
Kristian Hveem

Maiken Gabrielsen

Anne Skogholt

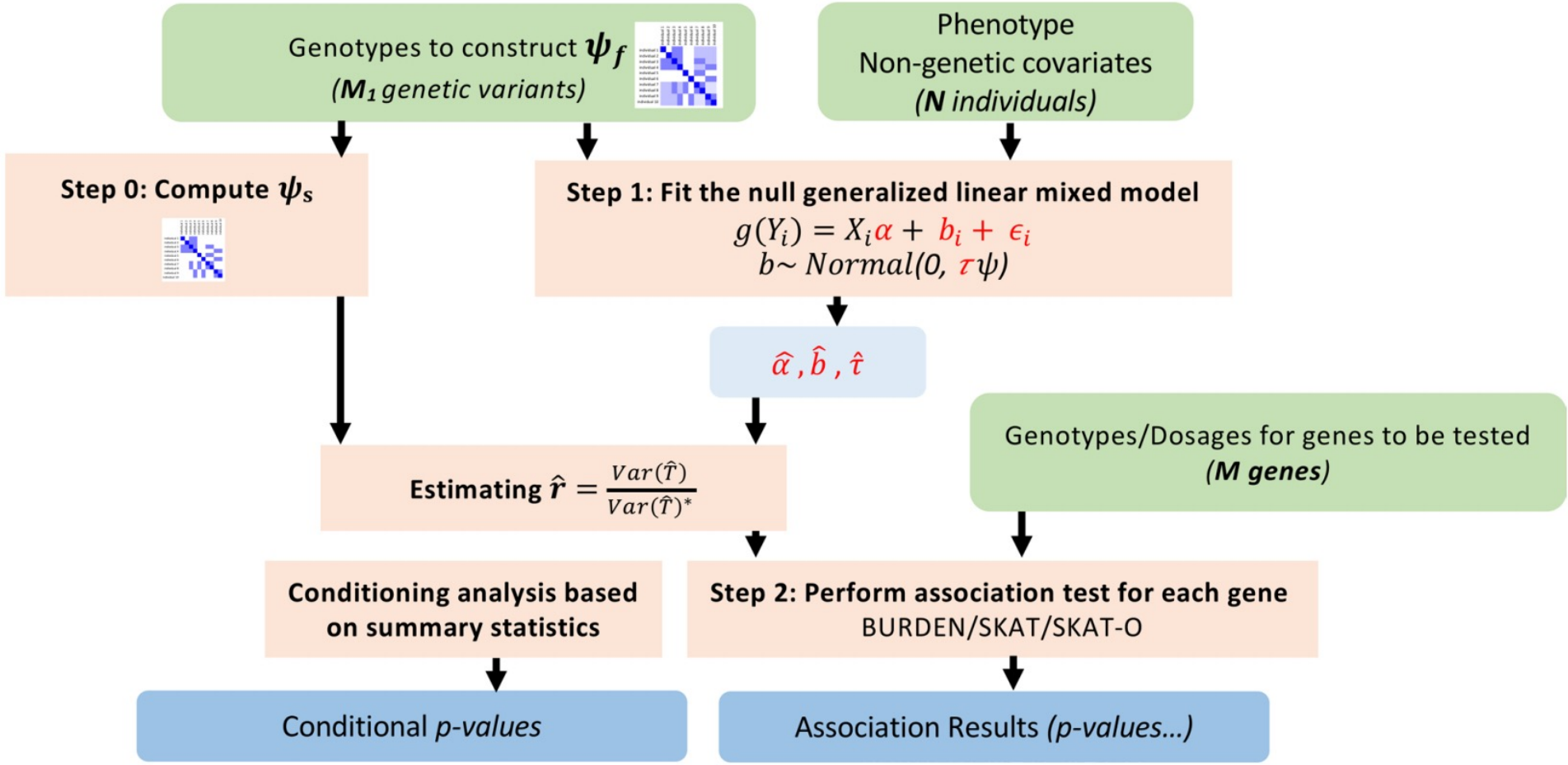
- *Jonas Nielsen*
- Lars Fritsche
- *Sarah Gagliano*
- Jonathon LeFaive
- Wenjian Bi

SAIGE-GENE is computationally efficient for large-scale biobanks



Estimated and projected computational cost by sample size (N) for gene-based tests of 15,342 genes, each containing 50 rare variants.

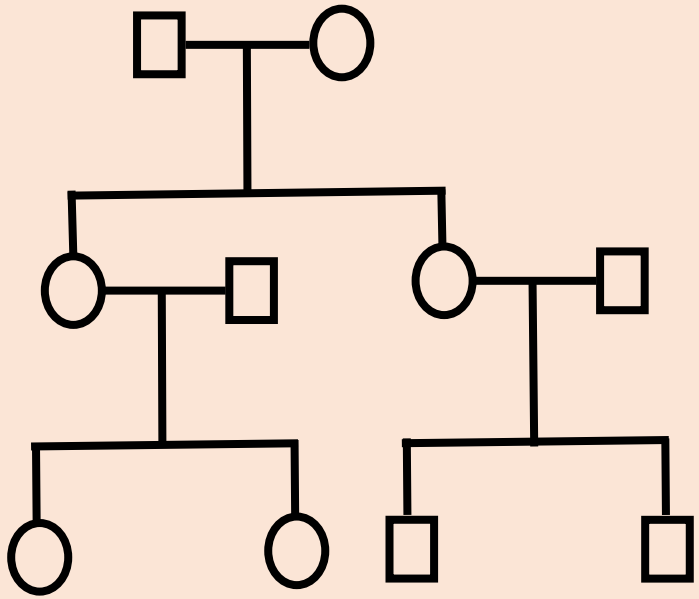
SAIGE-GENE:
Generalized Linear Mixed Model for Gene-based Association Tests



Extended Data Fig. 1 | Workflow of SAIGE-GENE. SAIGE-GENE consists of two steps: (1) Fitting the null generalized linear mixed model (GLMM) to estimate variance components and other model parameters; (2) Testing for association between each genetic variant set, such as a gene or a region, and the phenotype.

Simulation Study

500 families



5,000 independent individuals

Model:

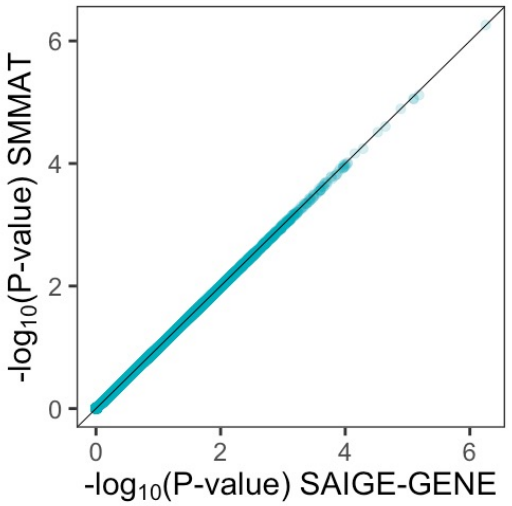
$$y_i = X_1 + G_i\beta + b_i + \epsilon_i$$

- X_1: intercept
- $b_i \sim N(0, \tau\,\psi)$, τ = 0.2 or 0.4
- $\epsilon_i \sim N(0, \sigma^2 I)$, σ^2 = 1- τ

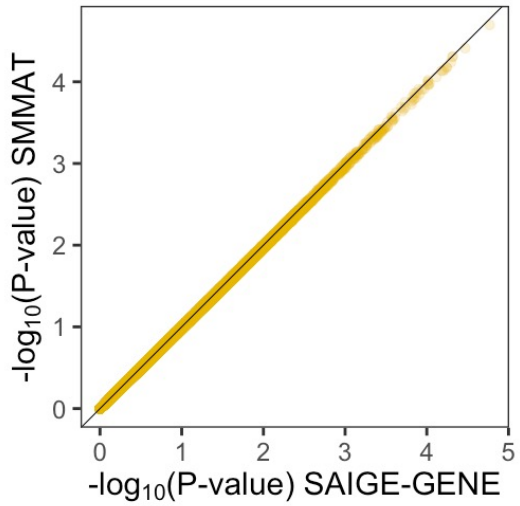
Consistent P-values from SAIGE-GENE and SMMAT for quantitative traits

τ=0.2, β = 0

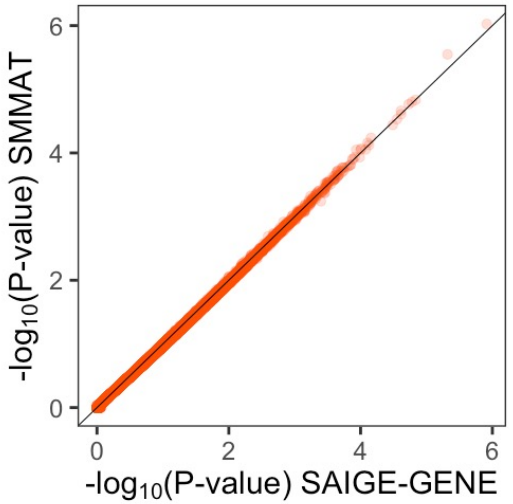
Burden



SKAT

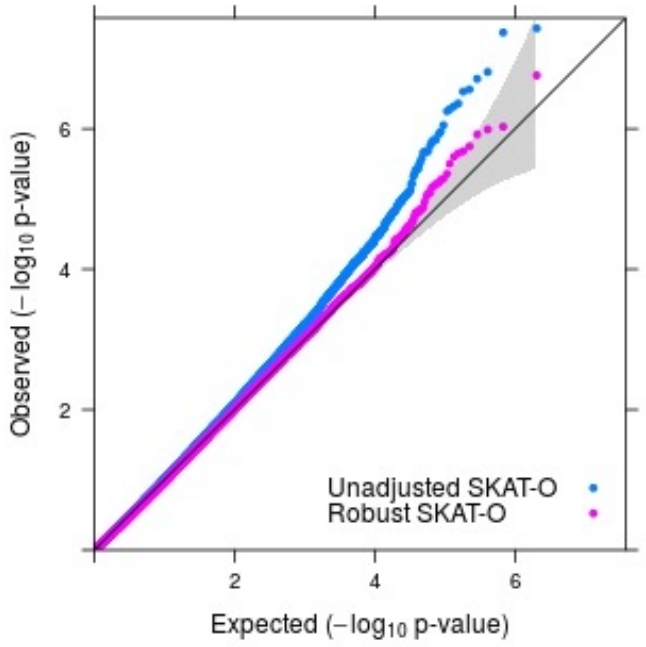


SKAT-O

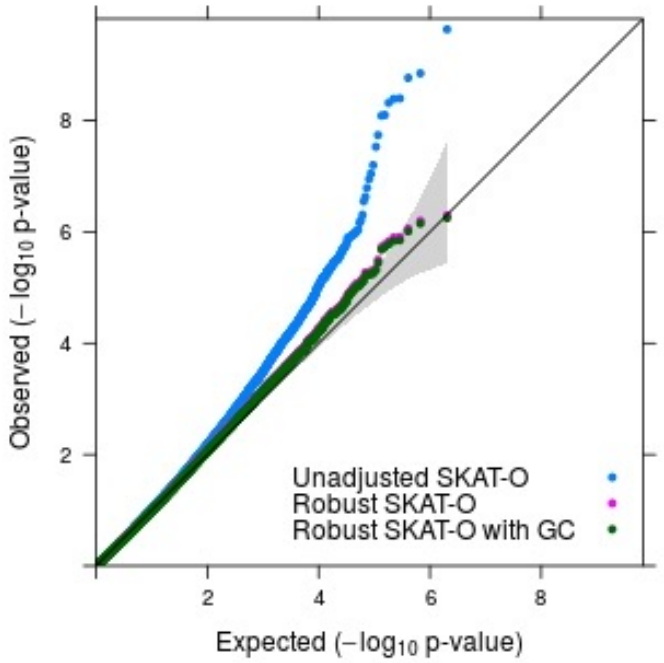


SAIGE-GENE provides greatly improved type I error control for binary traits relative to the unadjusted approach of assuming normality

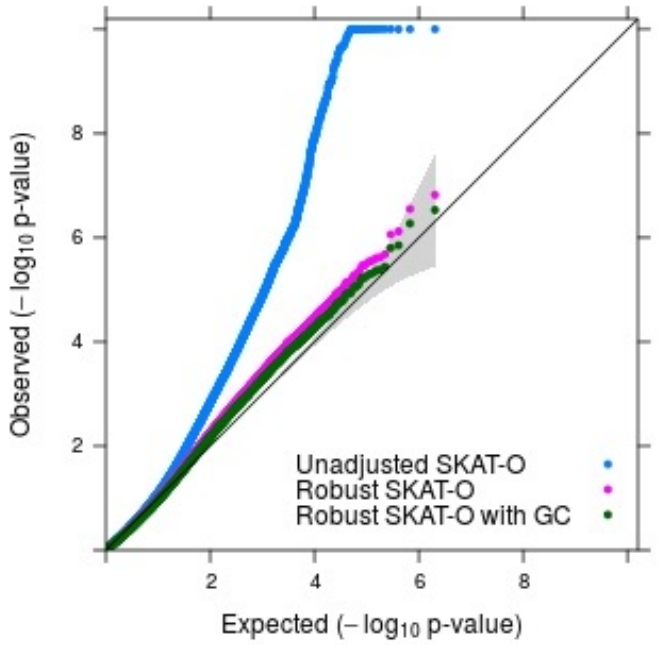
A. Case: Control = 1:9



B. Case: Control = 1:19



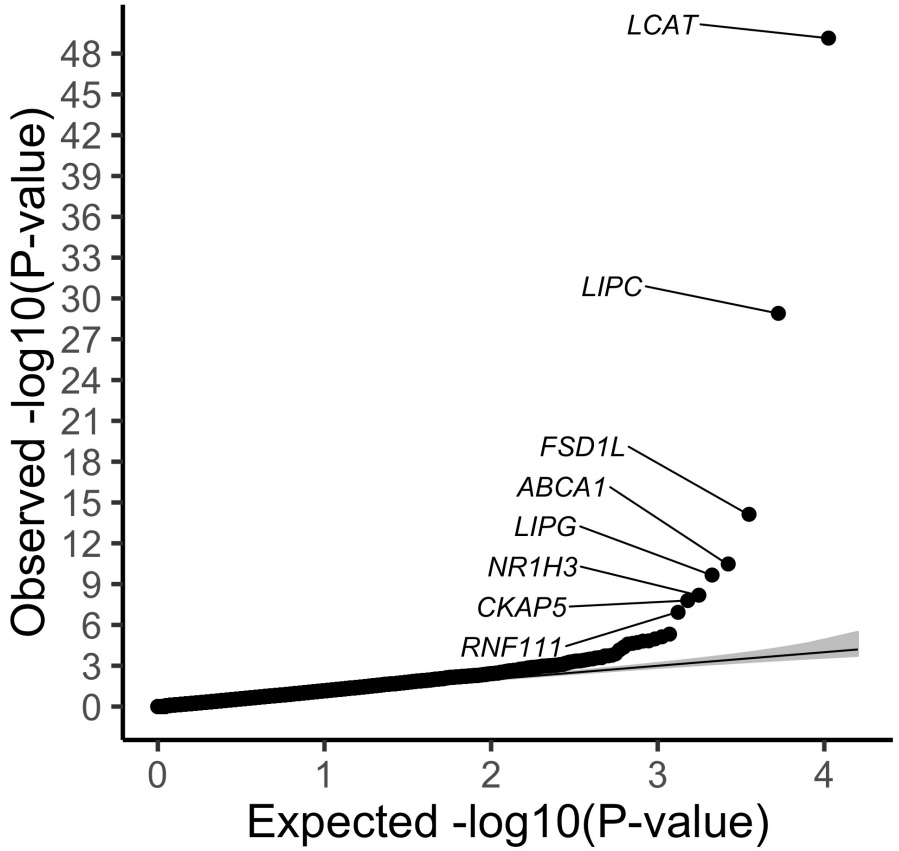
C. Case: Control = 1:99



Apply SAIGE-GENE to quantitative traits in the HUNT study and UK Biobank

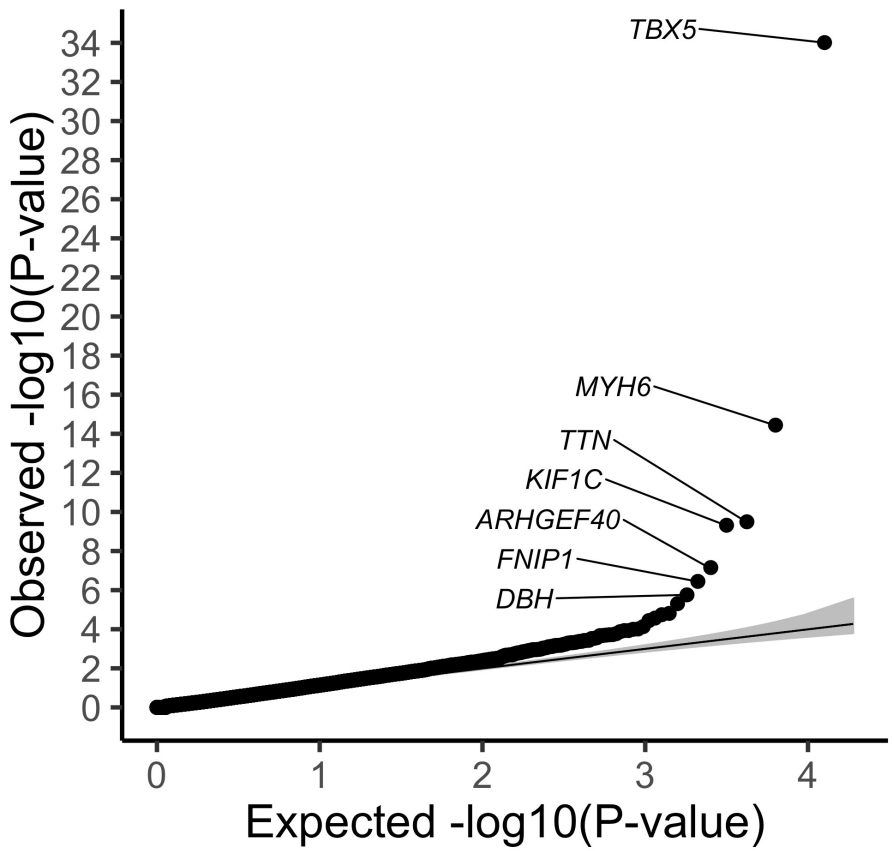
HDL in the HUNT study (N= 69,214)

A



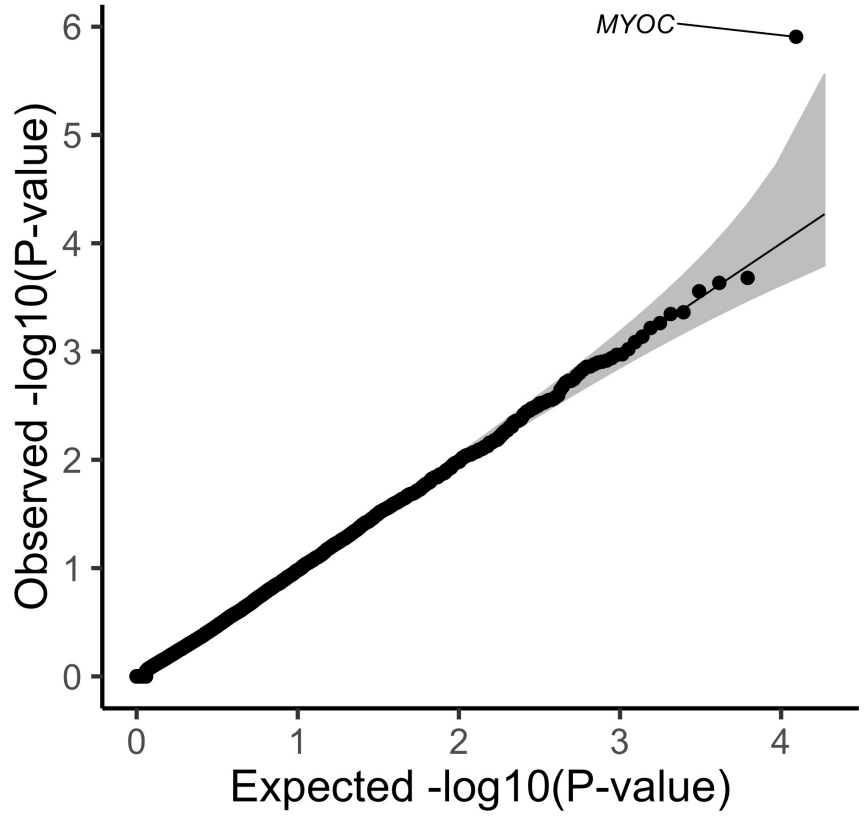
Automated readings of pulse rate in the UK Biobank (N= 385,365).

B



Apply SAIGE-GENE to the binary phenotype in UK Biobank

Glaucoma in the UK Biobank (N cases = 4,462; N controls = 397,761)



Code and Data Availability

- SAIGE-GENE is implemented as an open-source R package available at
  - https://github.com/weizhouUMICH/SAIGE/

- The summary statistics and quantile–quantile plots for 53 quantitative phenotypes and 10 binary phenotypes in the UK Biobank by SAIGE-GENE are available for public download at
  - https://www.leelabsg.org/resources

References

- Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. "Rare-variant association testing for sequencing data with the sequence kernel association test." *The American Journal of Human Genetics* 89, no. 1 (2011): 82-93.
- Lee, Seunggeun, Michael C. Wu, and Xihong Lin. "Optimal tests for rare variant effects in sequencing association studies." *Biostatistics* 13, no. 4 (2012): 762-775.
- Chen, Han, Jennifer E. Huffman, Jennifer A. Brody, Chaolong Wang, Seunggeun Lee, Zilin Li, Stephanie M. Gogarten et al. "Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies." *The American Journal of Human Genetics* 104, no. 2 (2019): 260-274.
- Zhou, Wei*, Zhangchen Zhao*, Jonas B. Nielsen, Lars G. Fritsche, Jonathon LeFaive, Sarah A. Gagliano Taliun, Wenjian Bi et al. "Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts." *Nature genetics* 52, no. 6 (2020): 634-639.