Rare variant association tests

University of Michigan

Zhangchen Zhao

Lecture Overview

- 1. Limitations of GWAS
- 2. Rationale for Rare Variant Analysis
- 3. Unadjusted SKAT, burden and SKAT-O
- 4. Robust SKAT, burden and SKAT-O

GWAS: Missing Heritability

- GWAS focus on common variants (MAF <= 5%).

- Missing heritability: Significant GWAS SNPs explain a small proportion of disease heritability.
- Possible reasons:
  - GxG and GxE interactions?
  - Many common causal variants: Each with a small effect?
  - Rare variants?

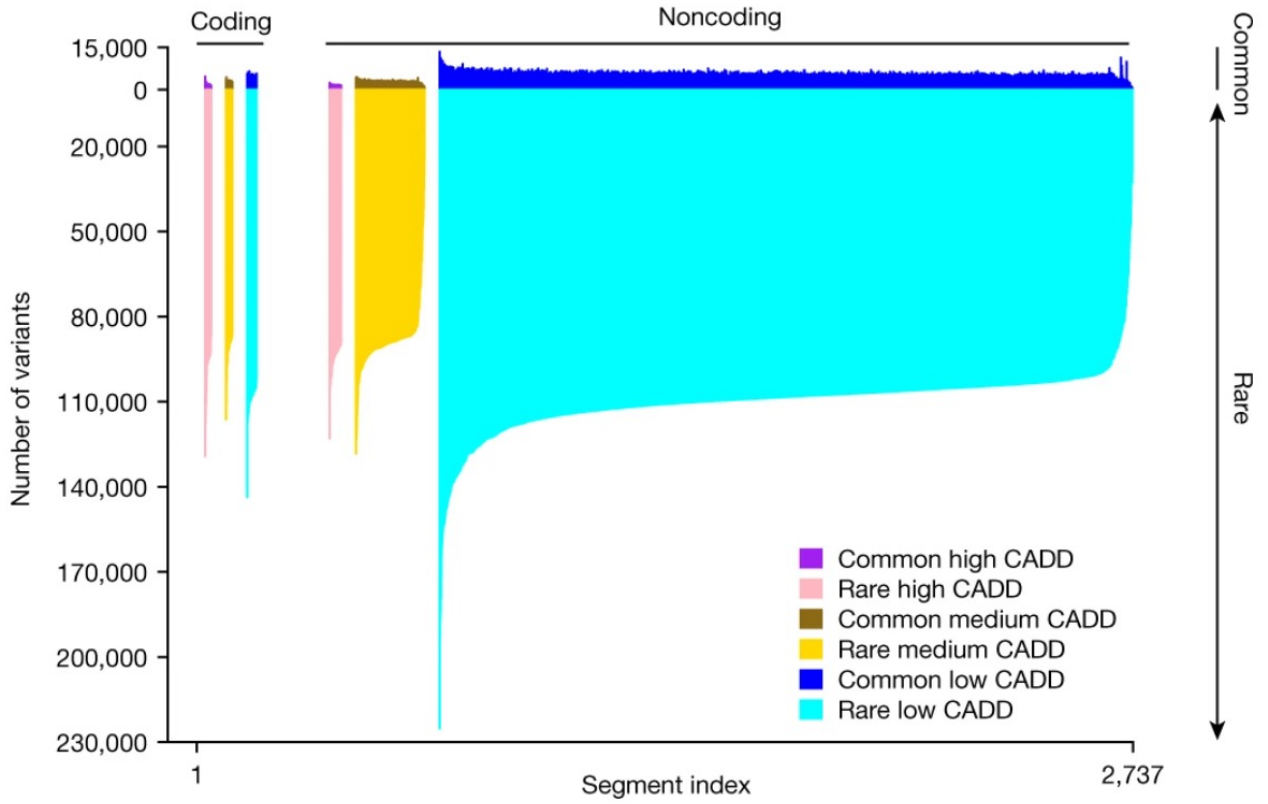
Common vs Rare variants

- Common Variants (Common SNPs):
  - MAF > 1%~5%.
  - Often high correlation with adjacent SNPs (Strong Linkage Disequilibrium(LD)).

- Rare Variants (Rare SNPs):
  - MAF<= 1%~5%.
  - Relatively new mutations.
  - Often weak correlation with other SNPs.

Why rare variants? (Taliun, 2021)

- Most of human variants are rare
- Functional variants tend to be rare.



Single variant association tests are underpowered for rare variants

(1) Large samples are required to observe rare variants.

- Sample size required to observe a variant with MAF= p with at least θ chance

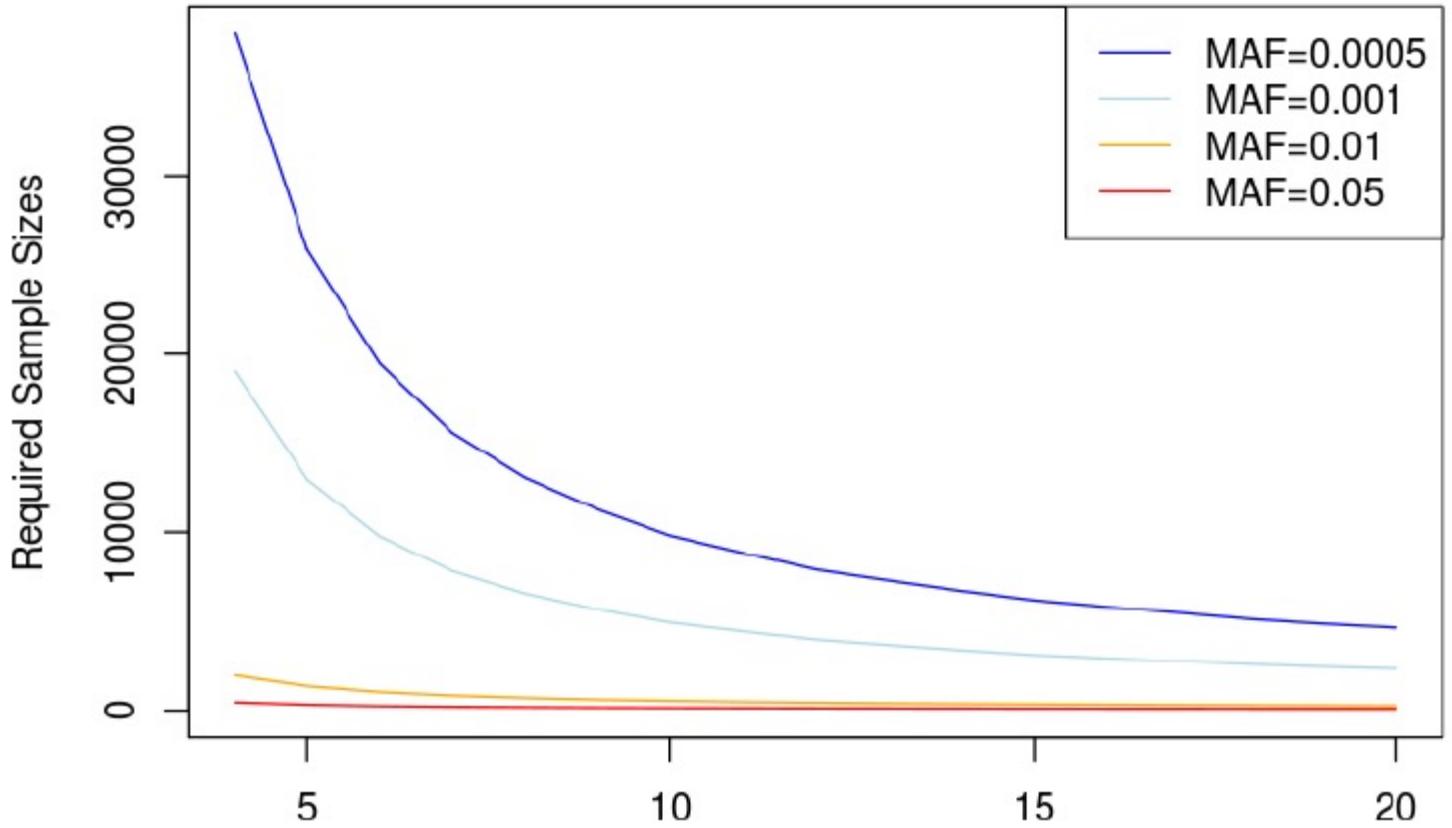
$$N > \frac{\ln(1-\theta)}{2\ln(1-p)}$$

- For θ= 99.9%, the required minimum sample size is

MAF	0.1	0.01	0.001	0.0001
Minimum N	33	344	3453	34537

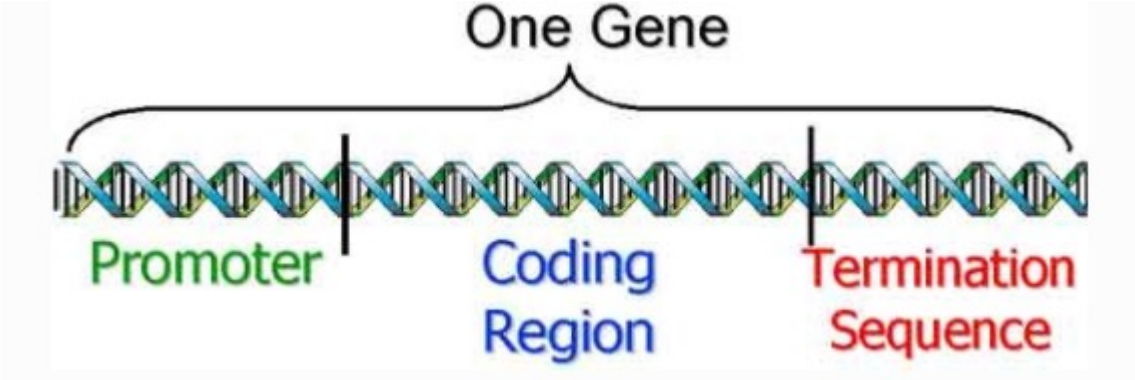
Single variant association tests are underpowered for rare variants

(2) How many subjects are needed to achieve 80% of power (α=10^(−6)) by single variant test?



To increase power in association studies for Rare Variants

- Test the joint effect of rare variants by grouping rare variants into functional units, i.e. genes.
- Region/gene-based tests



* * * * * *

* * * * * *

Methods for region/gene-based tests

- Consider the following question: Given N independent observations, we already know
  - the phenotype we are interested in
  - covariates we need to adjust
  - all genotype information of rare variants in a region

  Can we get an appropriate p-value of this rare-variant region?

- Existing methods: SKAT (Wu, 2011), burden (Wu, 2011), SKAT-O (Lee, 2012), C-alpha test (Neale, 2011), etc.

Model

- For continuous traits, we consider the following model

$$Y_i = X_i'\alpha + G_i'\beta + \varepsilon_i$$

- For binary traits, we consider the following model

$$\text{logit}[\Pr(Y_i = 1|X_i, G_i)] = X_i{}'\alpha + G_i{}'\beta$$

- For the individual i,
  - Y_i is the outcome;
  - X_i is the vector containing all the covariates, including the intercept;
  - G_i is the genotype vector of rare variants with length m;

- $H_0: \beta = 0$ vs $H_A: \beta \neq 0$

Burden, SKAT and SKAT-O (I) (Wu, 2011)

- The burden statistic:

$$Q_B = [\sum_{i=1}^{n}(y_i - \hat{\pi}_i)\,(\sum_{j=1}^{m}\omega_j g_{ij})]^2,$$

where Q_B follows scaled χ_1^2 distribution asymptotically under H_0.

- SKAT statistic:

$$Q_s = \sum_{j=1}^{m}{\omega_j}^2\left\{\sum_{i=1}^{n}g_{ij}(y_i - \hat{\pi}_i)\right\}^2,$$

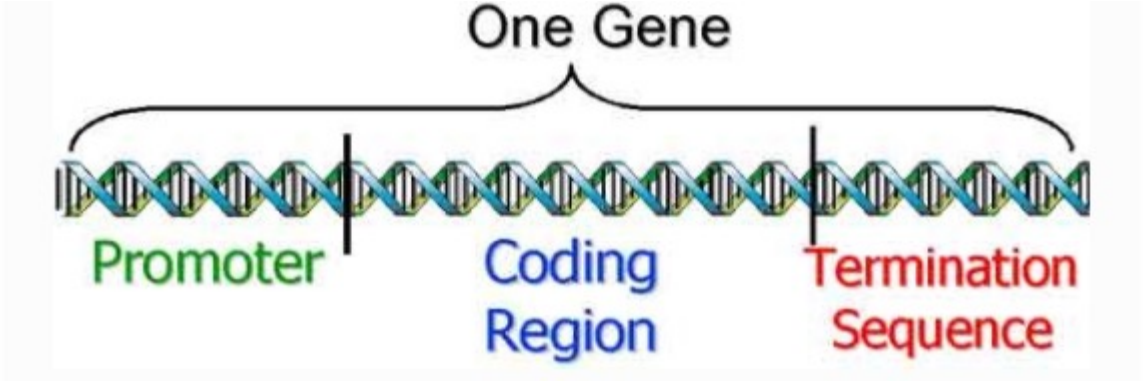
where Q_s asymptotically follows a mixture of chi-square distribution under H_0.

- SKAT-O statistic:

$$Q_\rho = (1-\rho)Q_B + \rho Q_s,$$

where ρ is a tuning parameter with range [0,1].

Burden, SKAT and SKAT-O (II)



+ + + + + + + + + + + + + → Burden tests can be used.

+ - + + - + + - + + - + + - → SKAT can be used.

? ? ? ? ? ? ? ? ? ? ? ? ? → SKAT-O can be used.

Score statistics used in Burden and SKAT

- Suppose $S_j = \sum_{i=1}^{n} g_{ij}(y_i - \hat{\pi}_i)$ is the score statistic for the variant j.
- Q_B and Q_S can be written as

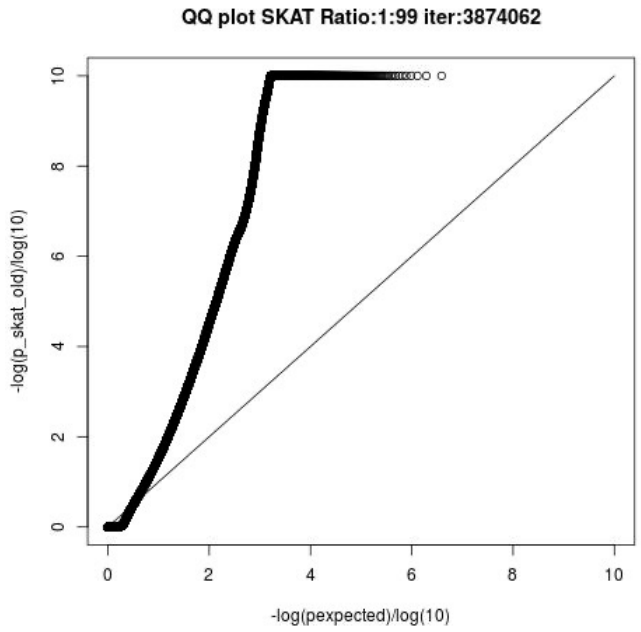
$$Q_B = \left(\sum_{j=1}^{m} \omega_j S_j \right)^2, \qquad Q_s = \sum_{j=1}^{m} {\omega_j}^2 {S_j}^2 .$$

- Under $H_0,\ S = (S_1, \ldots, S_m)^T$ asymptotically follows $MVN\left(0, V^{\frac{1}{2}} C V^{\frac{1}{2}}\right)$
  - C is the correlation matrix among m variants;
  - V is a diagonal matrix where the diagonal elements are the asymptotic variances of S.

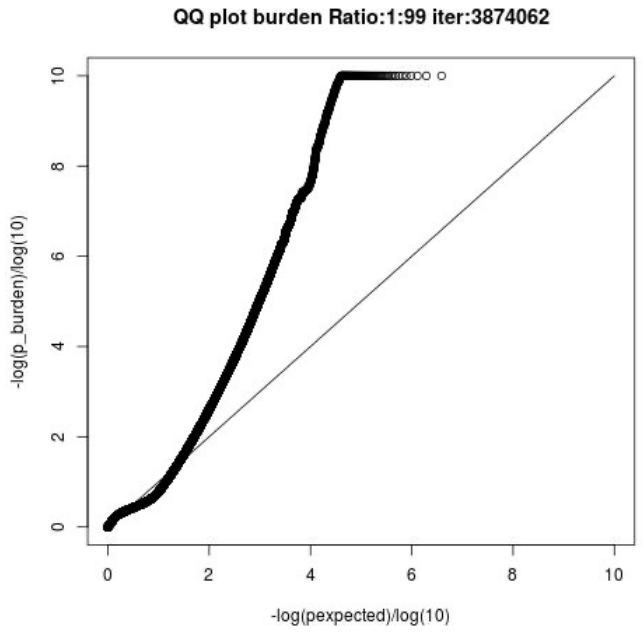
- SKAT, burden and SKAT-O can be implemented in R package SKAT.

Existing method has inflation of type I error rates

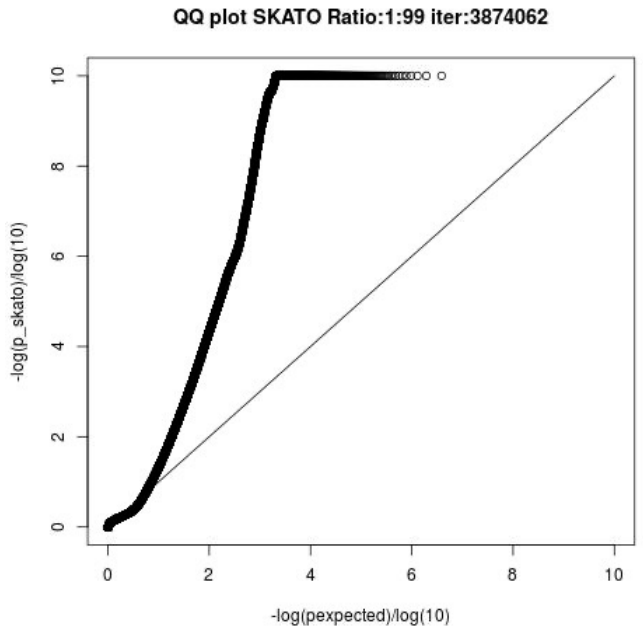
- Case: Control Ratio= 1:99
- The huge inflation can be found in QQ plots.



SKAT



Burden



SKAT-O

Robust Region-based Test (Zhao, 2021)

Main idea:

- Under H_0, $S \sim \mathrm{MVN}\left(0, V^{\frac{1}{2}} C V^{\frac{1}{2}}\right)$.
- However, in the presence of case-control imbalance, the distribution of S are not normal, causing misleading p-values.

- Solution:
  - Step1: Estimate distribution accurately to calculate single-variant p-values
    - Saddle Point Approximation (SPA)
    - Efficient Resampling (ER)
  - Step2: Re-estimate diagonal variance matrix V so that single-variant p-values are the same as p-values from SPA or ER.
- The robust methods can also be implemented in R package SKAT.

Saddle Point Approximation (SPA) (Dey, 2017)

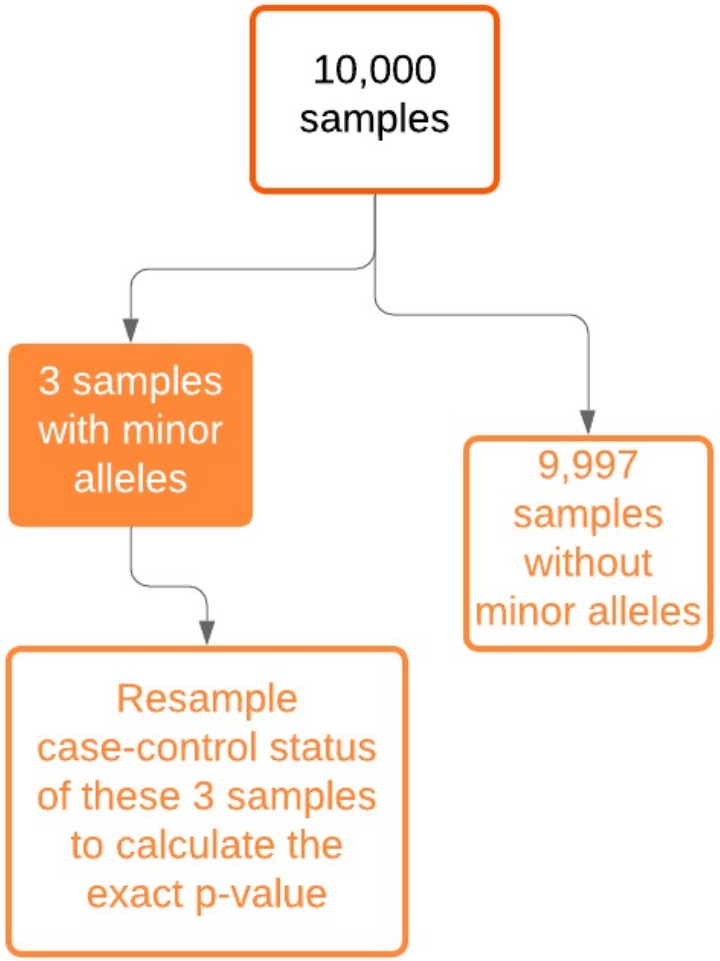
- The cumulant generating function (CGF) $K_j(t)$ of the score statistic S_j could be derived based on the fact $Y_i \sim \text{Bernoulli}(\mu_i)$ under H_0.
  - We can further calculate p-values.

- Normal approximation behaves well near the mean, but poorly at the tails, especially if the true distribution is skewed.
  - Normal approximation cannot incorporate higher moments such as skewness.

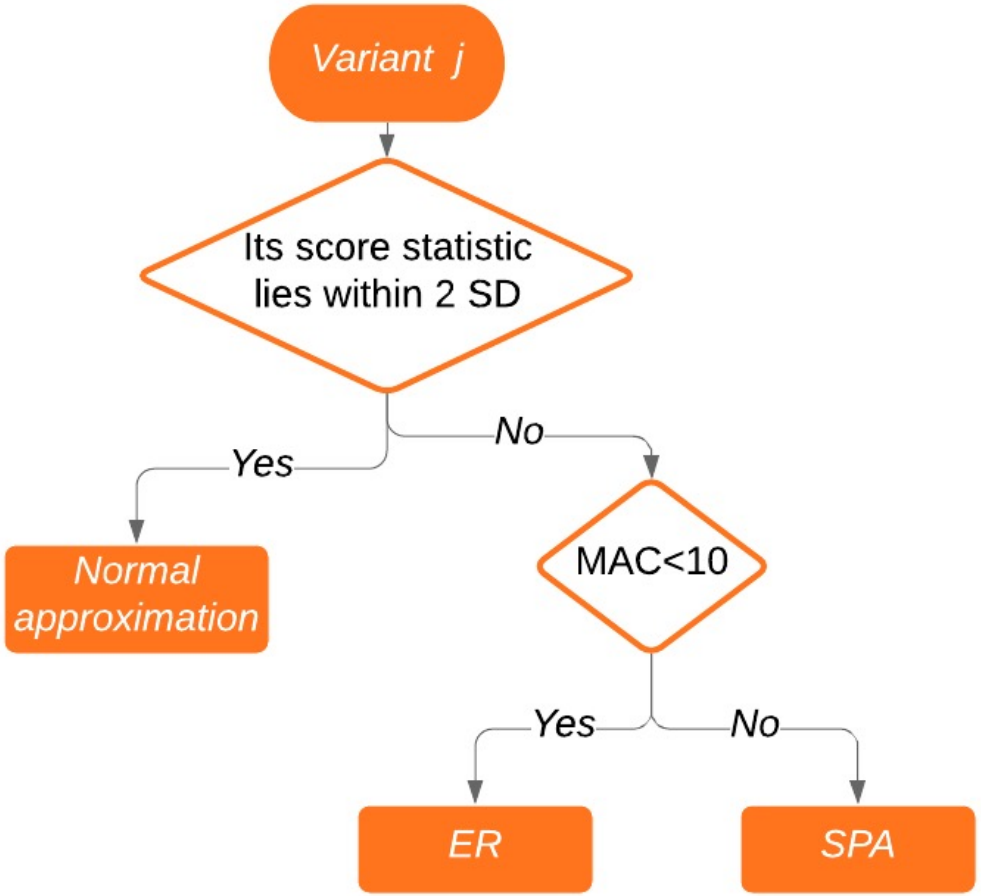
Efficient Resampling (ER) (Lee, 2016)

- Resample the case–control status of individuals with a minor allele at a given variant
  - Instead of permuting case–control status across all individuals
  - only individuals with minor alleles contribute to the score statistics S.

- When MAC is low (ex. MAC < 20), ER can calculate the exact p-value by numerating all possible configurations of case-control statuses.



Robust Region-based Test



• Adjust the variance of S_j so that the p-value is the same as $\tilde{p}_j$ from ER or SPA:

$$\tilde{V}_j = \frac{S_j^2}{\chi^2_{quantile}(1-\tilde{p}_j)}.$$

• The p-value of the region can be calculated based on the assumption that

$$S \sim MVN\left(0, \tilde{V}^{\frac{1}{2}} C \tilde{V}^{\frac{1}{2}}\right).$$

Simulation studies to evaluate type I error rates

- Generate the sequence data of mimicking European ancestry over 200 kb regions using the calibrated coalescent model

- 50,000 independent individuals with 4 case control ratios:
  - 1: 1, 1: 9, 1: 49 and 1: 99
- Two covariates:
  - $X_1 \sim Bernoulli(0.5)$
  - $X_2 \sim Normal(0, 1)$
- The binary phenotypes were simulated from

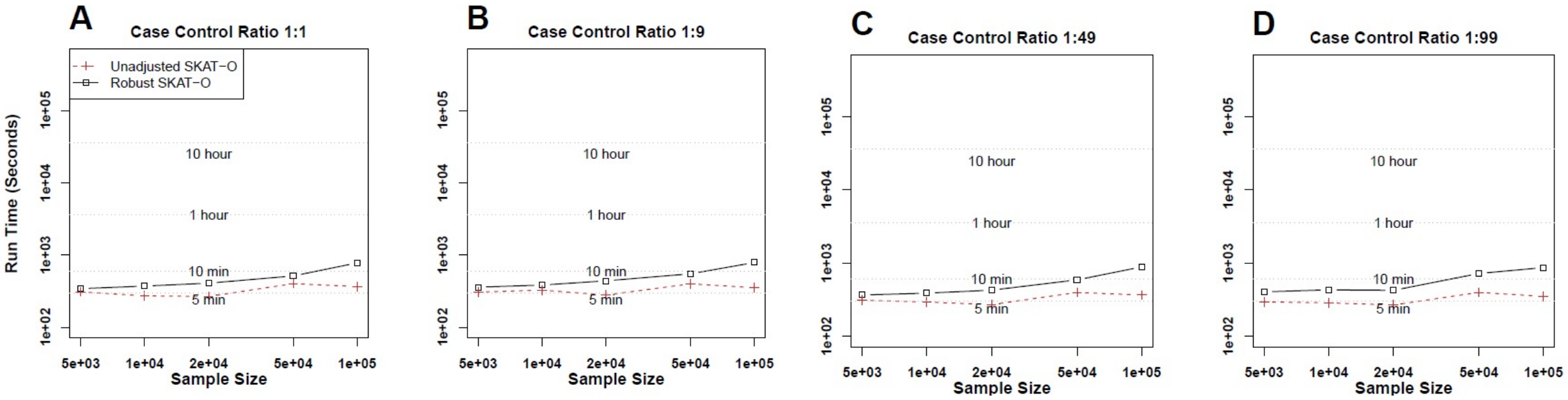
$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \beta_1 g_{1i} + \cdots + \beta_m g_{mi}.$$

Robust method has well controlled type I error rates

Alpha	Ratio	Rare variants					
		SKAT	Robust SKAT	Burden	Robust burden	SKAT-O	Robust SKAT-O
2.5×10^{-6}	1:1	1.24	1.54	1.11	1.03	1.38	1.38
	1:9	2.47	1.45	1.29	0.77	2.51	1.49
	1:49	28.27	1.91	6.88	1.06	23.70	1.98
	1:99	89.53	1.81	16.34	0.90	71.32	1.60

Note: the number in each cell represents the type I error rate divided by alpha level.

Robust method has similar computation time (1000 iterations)



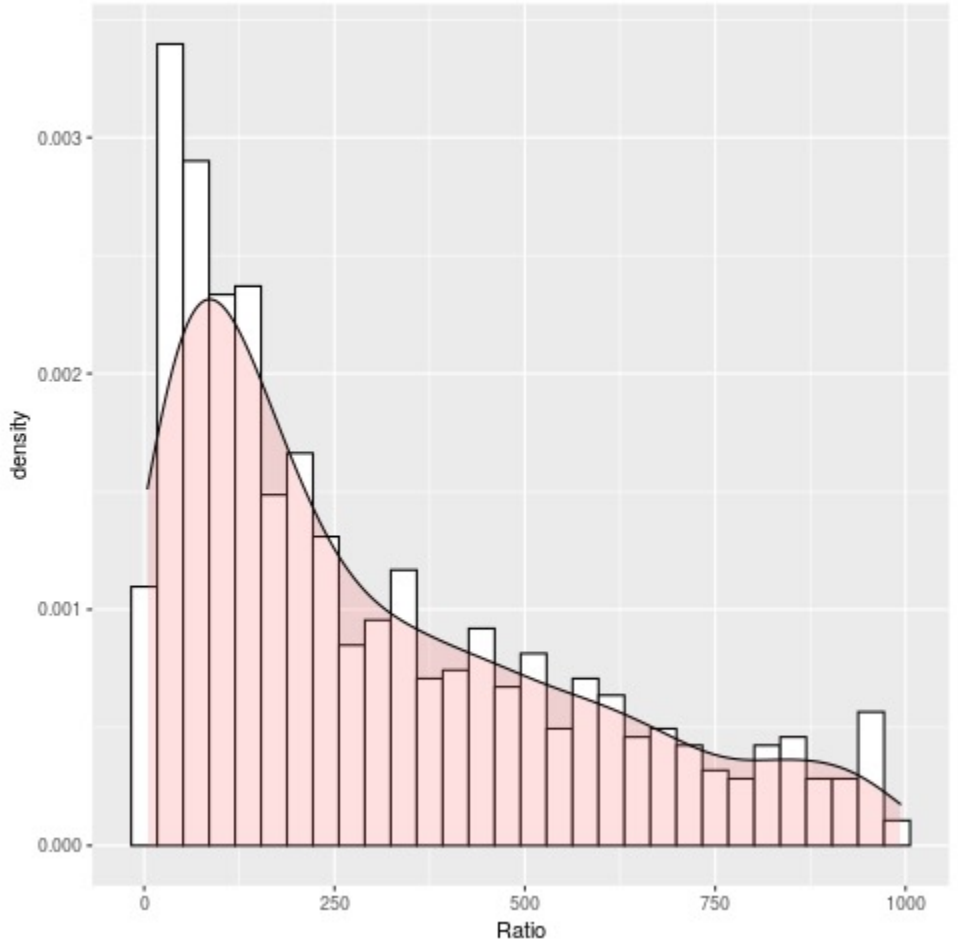
Case Control Ratio 1:1 1:9 1:49 1:99

UK Biobank WES Data

- 791 binary phenotypes with at least 50 cases based on PheCodes.
- Rare variants (MAF<=0.01) of the nonsynonymous and splicing variants in the exon and neighboring regions.

- A total of 18,360 genes remained for analysis
  - gene size ranged from 2 to 7,439 with a highly skewed distribution
- Covariates: age, gender and the first four principal components.

Control: Case of 791 phenotypes

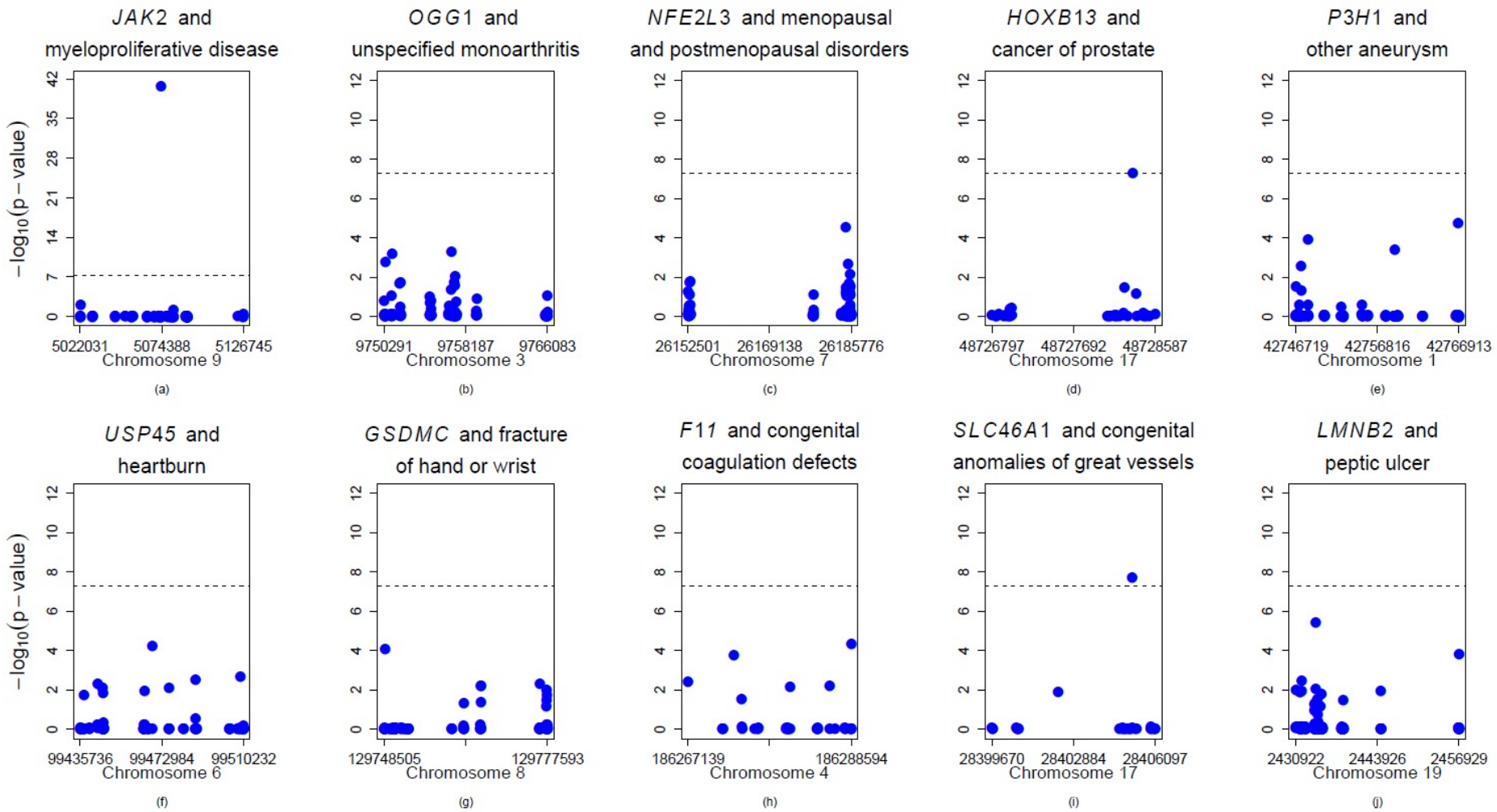


Significant Gene-Phenotype Associations (p-value< 10^{-7})

Phenotype (PheCode)	Gene Name	Case-control Ratio	NSNP	Robust SKAT-O	Lowest P SNP	Conditional P-value (SKAT-O)
Myeloproliferative disease (200)	*JAK2*	94:9306	73	1.36E-33	1.81E-41	1.06E-35
Unspecified monoarthritis (716.2)	*OGG1*	1728:41060	117	7.73E-09	4.67E-04	7.79E-09
Menopausal and postmenopausal disorders (627)	*NFE2L3*	1345:21226	171	2.54E-08	2.72E-05	3.94E-08
Cancer of prostate (185)	*HOXB13*	741:18940	37	3.00E-08	5.24E-08	2.50E-08
Other aneurysm (442)	*P3H1*	164:16236	110	5.76E-08	1.71E-05	4.03E-07
Heartburn (530.9)	*USP45*	189:18711	103	6.34E-08	5.39E-05	1.46E-09
Fracture of hand or wrist (804)	*GSDMC*	382:37818	109	7.12E-08	8.17E-05	1.49E-07
Congenital coagulation defects (286.1)	*F11*	76:7524	38	7.40E-08	4.52E-05	4.09E-08
Congenital anomalies of great vessels (747.13)	*SLC46A1*	134:13266	28	9.38E-08	1.86E-08	3.87E-08
Peptic ulcer (excl. esophageal) (531)	*LMNB2*	773:44818	171	9.89E-08	3.83E-06	9.54E-08

Note: Associations with red color are previously reported.

Scatterplots of single variants in 10 significant genes

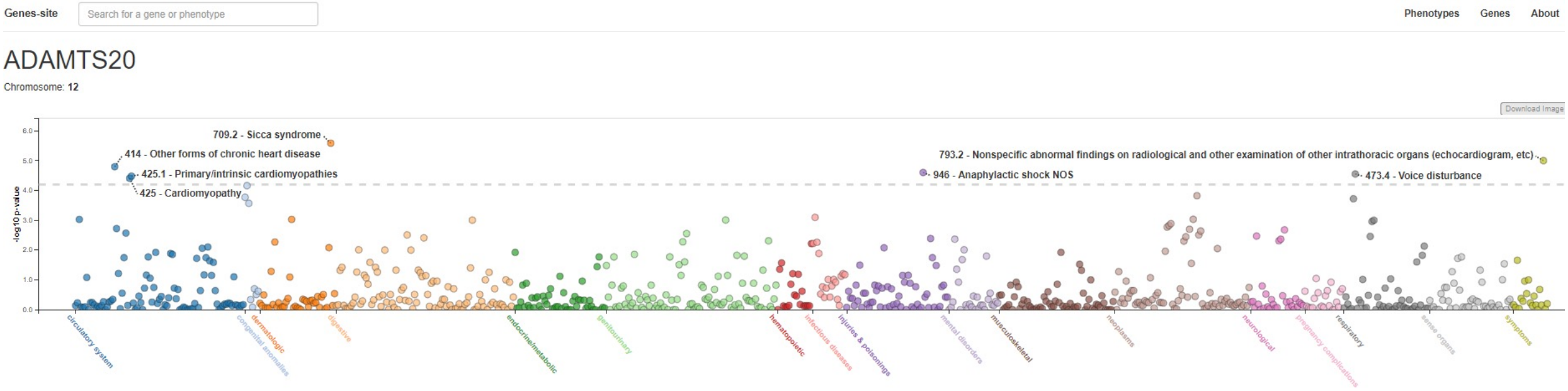


GWAS figures in PheWeb (Denny, 2010) (I)

Gene	P-value	#Rare Variants	Chromosome	Start-End	Case MAC (Minor Allele ...	Control MAC (Minor Allele...
JAK2	1.4e-33	73	9	5,022,031 - 5,081,724	27	442.01
CTAGE4;CTAGE8	2.2e-7	90	7	144,184,113 - 144,268,388	25.25	1666.99
ERVV-2	2.7e-6	33	19	53,049,361 - 53,050,859	7.05	149.77

Link: http://ukb-50kexome.leelabsg.org/

PheWAS figures in PheWeb (Denny, 2010)(II)



Category	Code	Name	#Cases	#Controls	P-value	#Rare Variants	Start-End	Case MAC (Minor Alle...	Control MAC (Minor Al...
filter column...	filter column...	filter column...							
dermatologic	709.2	Sicca syndrome	58	5,742	2.6e-6	110	43,354,237 - 43,462,894	10	185
symptoms	793.2	Nonspecific abnormal findings on radiological and other exa...	53	5,247	1.0e-5	114	43,354,237 - 43,550,908	7	167
circulatory system	414	Other forms of chronic heart disease	145	14,355	1.6e-5	186	43,354,237 - 43,550,908	13	468.01

Link: http://ukb-50kexome.leelabsg.org/

Acknowledgement and references

- The slides are modified based on “Introduction to Rare Variant Analysis and Collapsing Tests” by Timothy Thornton and Michael Wu at Summer Institute in Statistical Genetics 2015

- Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in medicine*, *4*(2), 45-61.
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82-93.
- Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., ... & NHLBI GO Exome Sequencing Project. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics, 91(2), 224-237.
- Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., ... & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet, 7(3), e1001322.
- Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L. G., & Lee, S. (2020). Uk biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *The American Journal of Human Genetics*, *106*(1), 3-12.
- Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics*, *101*(1), 37-49.
- Lee, S., Fuchsberger, C., Kim, S., & Scott, L. (2016). An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. *Biostatistics*, *17*(1), 1-15.
- Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., ... & Crawford, D. C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics, 26(9), 1205-1210.