Lecture 7: Rare Variant Analysis: Collapsing Tests, Kernel (Variance Component) Tests and Omnibus Tests

Instructors: Joelle Mbatchou and Loic Yengo

Summer Institute in Statistical Genetics 2023

Lecture Overview

1. Limitations of GWAS
2. Rationale for Rare Variant Analysis
3. Challenges
4. Collapsing/Burden Tests
5. Variance Component Tests
6. Omnibus Tests

GWAS: Missing Heritability

- GWAS primarily focus on common variants (MAF $\geq 5\%$) whose effects are small.
- **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?
  - Epigenetics?
  - **Rare variants?**

Why rare variants?

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

Article

Table 1 | Number of variants in 40,722 unrelated individuals in TOPMed

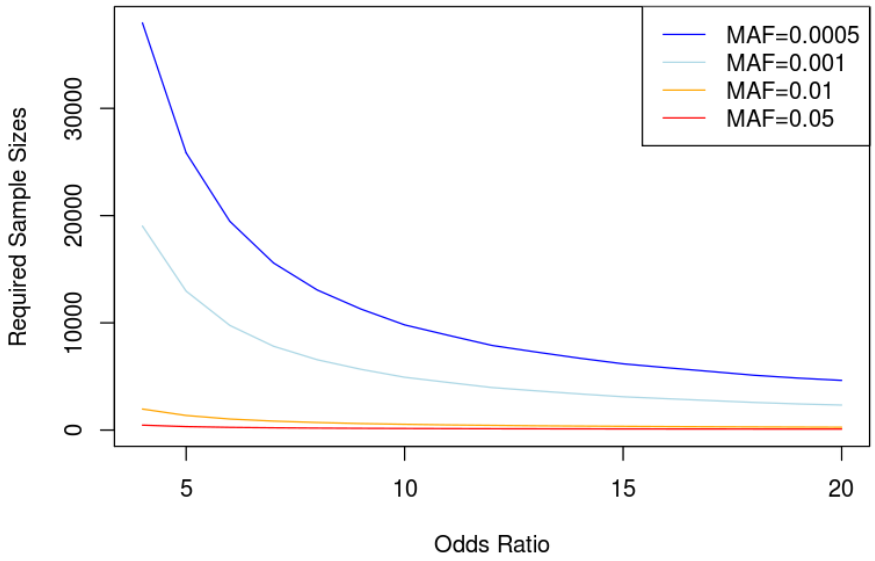
	All unrelated individuals (n = 40,722)		
	Total	Singletons (%)	A
Total variants	**384,127,954**	**203,994,740 (53)**	
SNVs	357,043,141	189,429,596 (53)	
Indels	27,084,813	14,565,144 (54)	
Novel variants	**298,373,330**	**191,557,469 (64)**	
SNVs	275,141,134	177,410,620 (64)	
Indels	23,232,196	14,146,849 (61)	
Coding variation	**4,651,453**	**2,523,257 (54)**	
Synonymous	1,435,058	715,254 (50)	
Nonsynonymous	2,965,093	1,648,672 (56)	
Stop/essential splice	97,217	60,347 (62)	
Frameshift	104,704	71,577 (68)	
In-frame	51,997	29,110 (56)	

Novel variants are taken as variants that were not present in dbSNP build 149, the most recent dbSNP version

Talium et al., *Nature* 2021

Challenges in Association Studies for Rare Variants

- Compared to common variant studies, **individual SNP analysis in rare variant studies is seriously under-powered.** $\rightarrow$ How many subjects are needed to achieve 80% of power ($\alpha = 10^{-6}$) by single variant test?



- A lot more rare variants than common variants $\rightarrow$ larger multiple testing burden

Challenges in Association Studies for Rare Variants

- Individual rare variant tests are under-powered
- Need **cost-effective study designs** to genotype a large number of individuals
- Need **powerful statistical methods and strategies** to test for associations
  - Region based analysis: genes, moving windows, networks/pathways
  - Integrate with bioinformatics: Incorporate functional information

Region Based Analysis of Rare Variants

- Gene (or Region) based tests
- Strategy:
  - Identify all observed variants within a sequenced (sub)-region.
  - Regions: gene, regulatory region, ...
  - Test the joint effect of rare variants.

Regression Models

- p variants in a certain region.
- SNPs in a region $\mathbf{G_i} = (g_{i1}, g_{i2}, \ldots, g_{ip})'$, $(g_{ij} = 0, 1, 2)$
- Covariates $\mathbf{X}_i$: age, gender, PC scores (for population stratification).
- Continuous/binary traits:

$$\begin{aligned} g(\mu_i) &= \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \mathbf{G_i}'\boldsymbol{\beta} \\ &= \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \sum_j g_{ij}\beta_j \end{aligned}$$

- Joint test of no genetic effect in region:

$$H_0 : \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) = 0$$

Major Classes of Tests

- Burden/Collapsing tests
- Supervised/Adaptive Burden/Collapsing tests
- Variance component (similarity) based tests
- Omnibus tests

Collapsing/Burden Tests - Principle

- If p is large, multivariate test $\boldsymbol{\beta} = 0$ is not powerful (df=p).
- Collapsing: Suppose $\beta_1 = \cdots = \beta_p = \beta$

$$g(\mu_i) = \alpha_0 + \mathbf{X}_i'\boldsymbol{\alpha} + \sum_j g_{ij}\beta_j$$
$$= \alpha_0 + \mathbf{X}_i^T\boldsymbol{\alpha} + C_i\beta$$

- $C_i = g_{i1} + \cdots + g_{ip}$: **genetic burden/score**
- Test $H_0 : \beta = 0$ (df=1)
- **Key assumption**: all rare variants in region are causal variants with the same effect sizes and association directions.

Burden Tests

- Collapse rare variants

Y	G_1	G_2	G_3	G_4		C
1	1	0	0	0		1
1	0	1	0	0		1
1	0	0	1	1	→	2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
0	0	0	0	0		0
0	0	0	0	0		0
0	0	0	0	0		0

Burden Tests

- Many variation of burden test exist based on different aggregation rules to get C_i
  - Reflects assumptions on genetic architecture
- **Existence of any rare variants can cause loss of function of a region** (e.g. CAST)

$$C_i = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{p} g_{ij} > 0 \\ 0 & \text{if} \quad \sum_{j=1}^{p} g_{ij} = 0 \end{cases}$$

- **Dominant genetic model** (e.g.. MZ-test)

$$C_i = \sum_{j=1}^{p} I(g_{ij} > 0)$$

Weighted Burden

- Assume that **rarer variants have larger effects**
- Suppose $\beta_j = w_j \beta$, where $w_j = w(MAF_j)$.
  - Ex: $w(MAF_j) = 1/\sqrt{MAF_j(1 - MAF_j)}$ (Madsen and Browning).
- Weighted count of rare variants

$$C_i = w_1 g_{i1} + \cdots + w_p g_{ip}$$

Power of Burden Tests

- Power of burden tests depends on
  - Number of associated variants
  - Number of non-associated (null) variants
  - Direction of the effects.
- **Powerful if most variants are causal and have effects in the same direction.**

Variance component test

- Burden tests are not powerful, if
  - there exist variants with different association directions
  - many null variants
- Variance component tests have been proposed to address this limitation.

Sequence Kernal Association Test (SKAT)

- Recall the original regression models:

$$g(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + \mathbf{G}_i^T \boldsymbol{\beta}$$

- Assume $\beta_j \sim dist.(0, w_j^2 \tau)$.
- $H_0 : \beta_1 = \cdots = \beta_p = 0 \iff H_0 : \tau = 0$.
- 1df test!

Sequence Kernel Association Test (SKAT)

- Score test statistic for $\tau = 0$:

$$Q_{SKAT} = (\mathbf{y} - \hat{\boldsymbol{\mu}}_0)'\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}}_0),$$

- $\mathbf{K} = \mathbf{GWWG}'$: weighted linear kernel (where $\mathbf{W} = diag[w_1, \ldots, w_p]$).
- It is a $N \times N$ similarity matrix

SKAT

- Q_{SKAT} is a **weighted sum of single variant score statistics**

$$Q_{SKAT} = (\mathbf{y} - \hat{\boldsymbol{\mu}}_0)'\mathbf{GWWG}'(\mathbf{y} - \hat{\boldsymbol{\mu}}_0)$$
$$= \sum_{j=1}^{p} w_j^2 [\mathbf{g}_j'(\mathbf{y} - \hat{\boldsymbol{\mu}}_0)] = \sum_{j=1}^{p} w_j^2 S_j^2$$

- S_j is a score test statistic in the SNP j only model:

$$g(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + g_{ij}\beta_j$$

- Under H_0, Q_{SKAT} (asymptotically) follows a **mixture of** χ^2 **distribution** $\sum_{j=1}^{p} \lambda_j \chi_{1,j}^2$

SKAT: P-value calculation

- P-values can be computed by inverting the characteristic function using Davies' method (1973, 1980)
  - Characteristic function

$$\varphi_x(t) = E(e^{itx}).$$

  - Characteristic function of $\sum_{j=1}^{p} \lambda_j \chi^2_{1,j}$

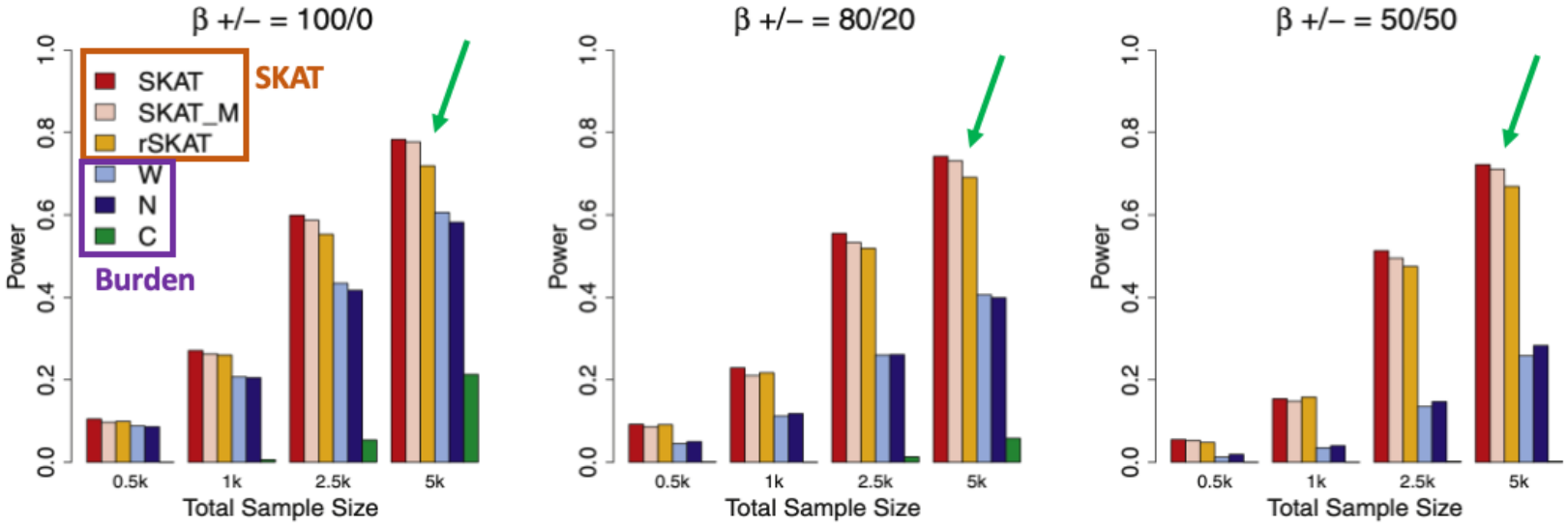
$$\varphi_x(t) = \prod_{i=j}^{p} (1 - 2\lambda_j it)^{-1/2}.$$

  - Inversion Formula

$$P(X < u) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{Im[e^{-itu}\varphi_x(t)]}{t} \, \mathrm{d}t.$$

Burden vs SKAT

- Power simulations: 5% of the variants in region are causal & vary the directions of effects
- SKAT remains powerful even if variants have different effect directions



Lee et al., *Am J Hum Genet* 2011

SKAT vs. Collapsing

- Collapsing tests are 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.
- Best test to use depends on the underlying biology.

  $\rightarrow$ Difficult to choose which test to use in practice.

SKAT vs. Collapsing

- Collapsing tests are 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.
- Best test to use depends on the underlying biology.

  $\rightarrow$ Difficult to choose which test to use in practice.

We want to develop a unified test that works well in both situations $\rightarrow$ Omnibus tests

Combine Test Statistics: Unified Test Statistics

Lee (2012). *Biostatistics*

- Combined Test of Burden tests and SKAT

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

- Q_ρ includes SKAT and burden tests.
  - $\rho = 0$: SKAT
  - $\rho = 1$: Burden

SKAT-O

- Model:

$$g(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + \mathbf{G}_i^T \boldsymbol{\beta}$$

  where β_j / w_j follows any arbitrary distribution with mean 0 and variance τ and the correlation among β_j's is ρ.
- SKAT-O considers $0 \leq \rho \leq 1$
- Special cases:
  - SKAT: $\rho = 0$
  - Burden: $\rho = 1$

SKAT-O

- Set a grid of values for ρ in $[0, 1]$ and pick ρ which maximizes power
  - Use the smallest p-value from different ρs:

$$T = \inf_{0 \leq \rho \leq 1} P_\rho.$$

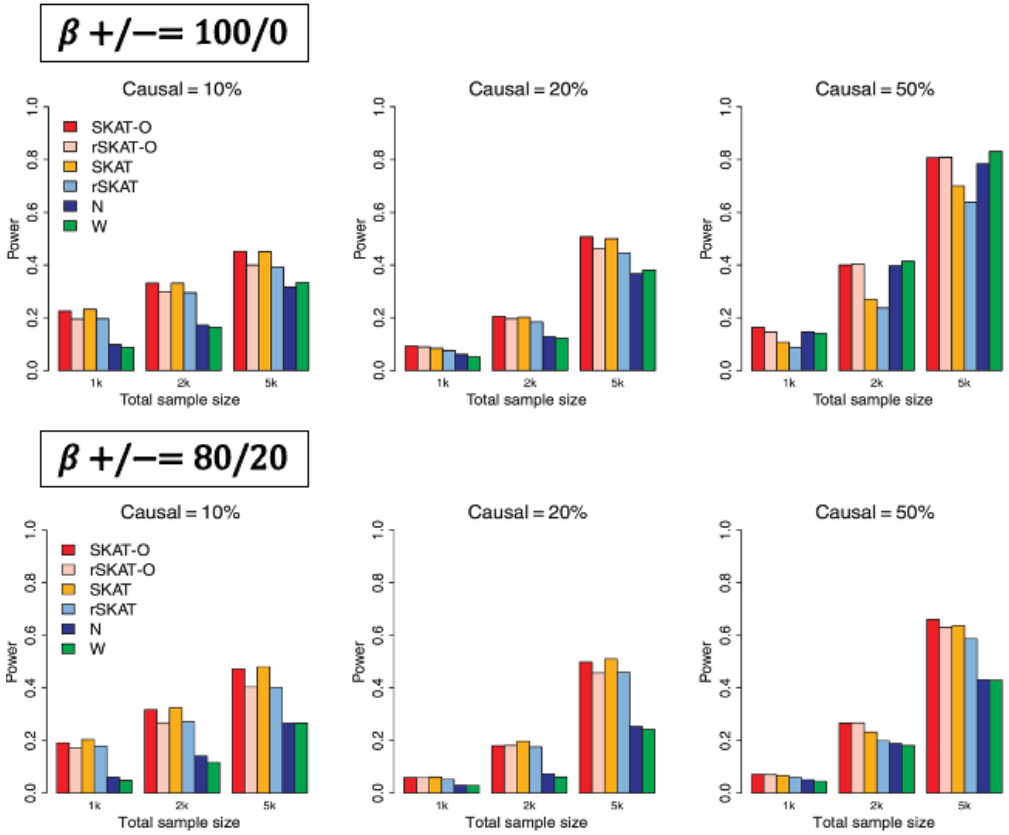
    where P_ρ is the p-value of Q_ρ for given ρ.
  - Test statistic:

$$T = min\, P_{\rho_b}, \quad 0 = \rho_1 < \ldots < \rho_B = 1.$$

- SKAT-O p-value is obtained through numerical integration

SKAT-O vs Burden/SKAT

- SKAT-O remains powerful across all scenarios



Lee et al., *Biostatistics* 2012

Aggregated Cauchy Association Test: ACAT

- Based on the Cauchy combination method to combine a set of p-values $\{p_j\}$:

$$T_{ACAT} = \sum_j w_j \tan\{\pi(0.5 - p_j)\}$$

- Computing p-value is extremely fast

$$\text{p-value} \approx 0.5 - \frac{\arctan\{T_{ACAT}/w\}}{\pi}, \quad w = \sum_j w_j$$

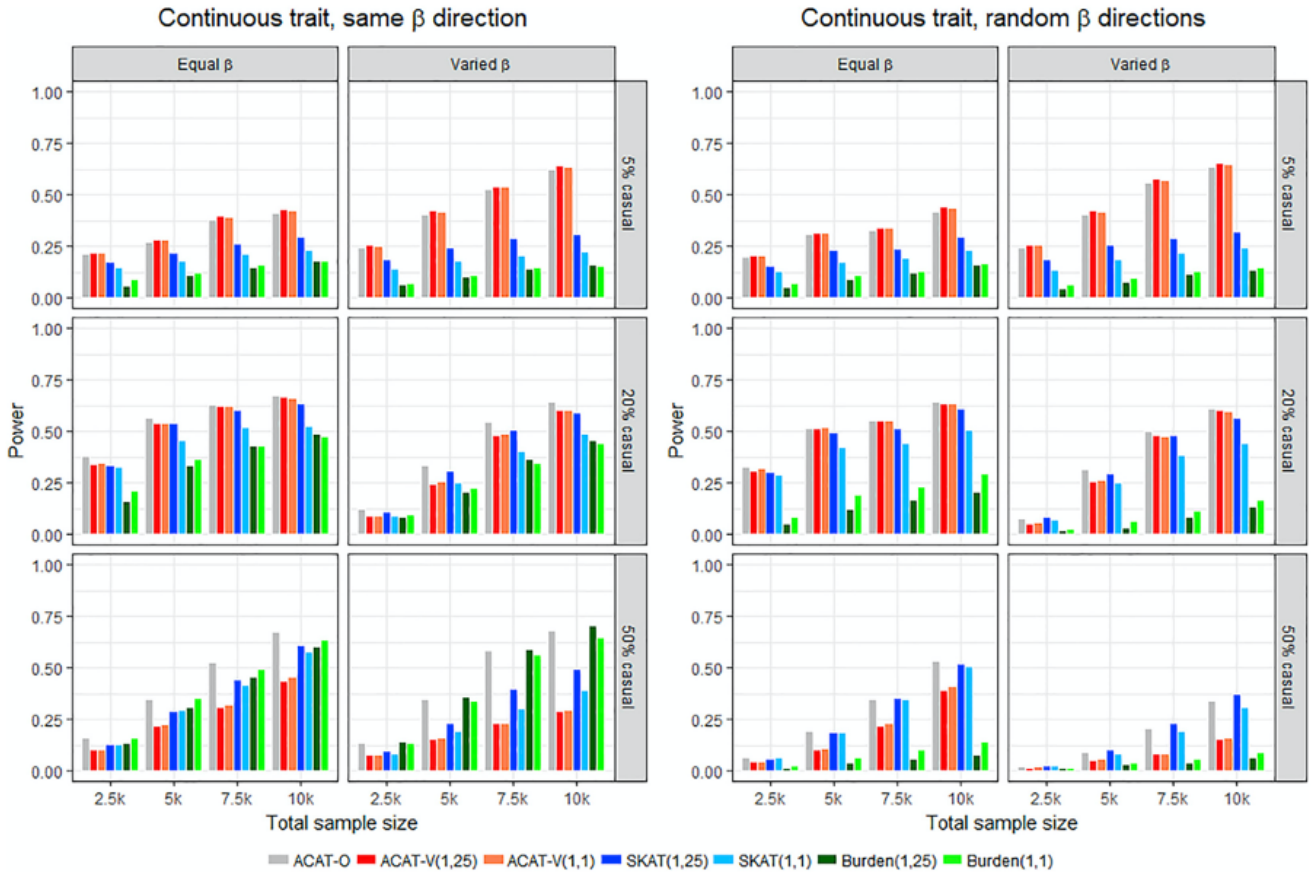
- Very accurate for small p-values
- Robust to correlation between the tests

Aggregated Cauchy Association Tests

- ACAT-V
  - Apply ACAT to single variant p-values from rare variants
  - More powerful when fewer variants are associated (i.e. sparse alternative)
  - SKAT & Burden can loose substantial power under this scenario
- ACAT-O
  - Apply ACAT to combine the p-values of SKAT, Burden and ACAT-V
  - Omnibus test which should work well whether
    - Effects are in same direction & many variants are associated
    - Effects are in different directions
    - Very few variants are causal

ACAT/SKAT/Burden

- ACAT-O remains powerful across all scenarios

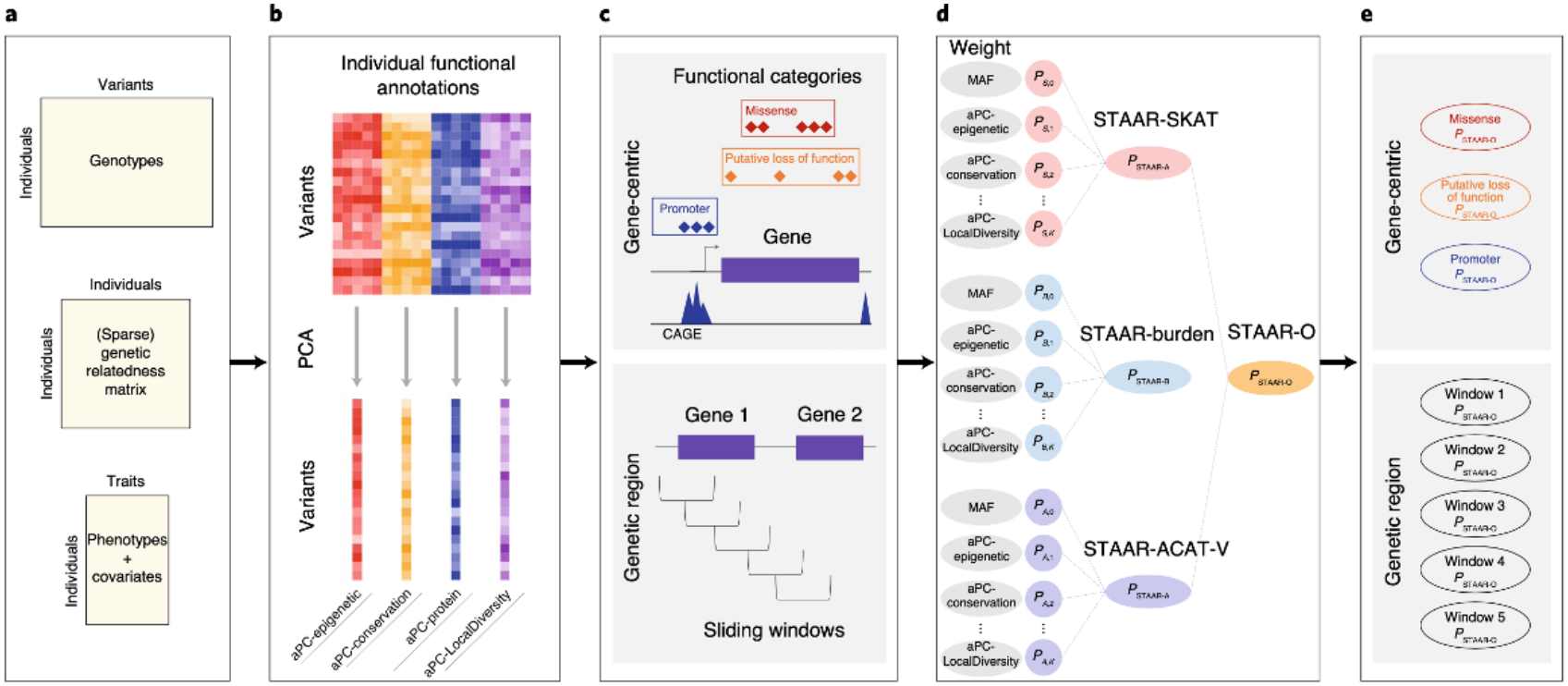


Liu et al., *AJHG* 2019

Incorporating external biological information

- What are the best variant weights to use in SKAT/Burden/ACAT-V tests?
- Using functional annotations can help improve statistical power, e.g.
  - variant effect predictor categories : loss of function, missense, ...
  - epigenetic scores (e.g. DNA methylation levels)
  - distance to coding region or transcription start/end site
- How to choose which set of variants to test jointly?
  - Within a gene
  - Sliding window

STAAR



Li et al., *Nat Gen* 2020

Summary

- Region based tests can increase the power of rare variants analysis compared to single variant tests.
- Relative performance of rare variant tests depends on underlying disease models
- Combined tests (omnibus tests), e.g, SKAT-O/ACAT-O, are more robust and powerful across different scenarios
- Can integrate functional annotation to boost statistical power

References

- Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
- Madsen, B.E. & Browning, S.R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics* **5**, e1000384 (2009).
- Wu, M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).

References

- Lee, S., Wu, M.C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-75 (2012).
- Liu, Y. et al. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410-421 (2019).
- Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, **52**, 969-983 (2020).