Genetic association tests for rare variants

2025 International Statistical Genetics Workshop

Duncan Palmer

Learning objectives

Talk

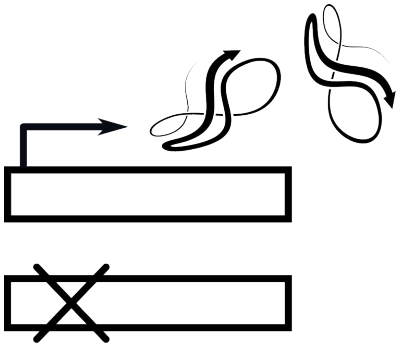
- Understand rationale and methodology behind rare variant association testing
- Be aware of approaches to increase statistical power

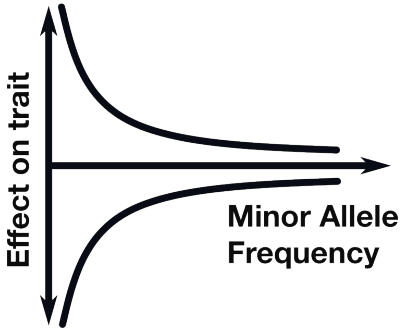
Practical

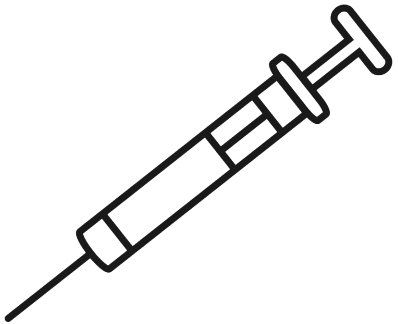
- Perform simple burden tests using R
- Get SAIGE-gene running

Why study rare genetic variation?

1. Interpretability of results
2. Large effects
3. Translational opportunities

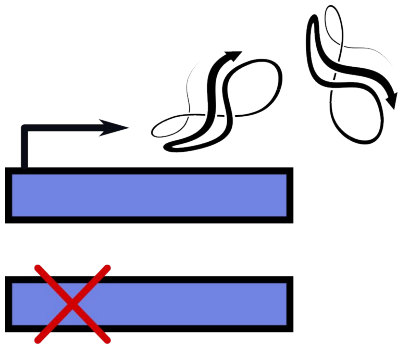


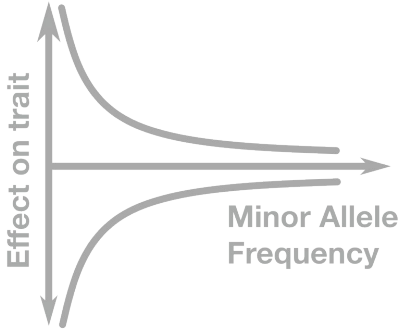


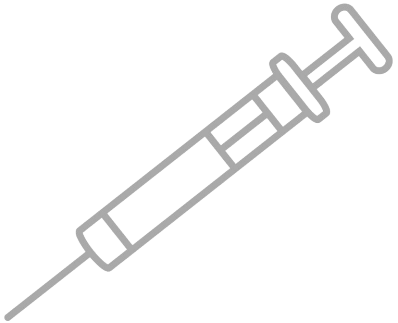


Why study rare genetic variation?

1. **Interpretability of results**
2. Large effects
3. Translational opportunities

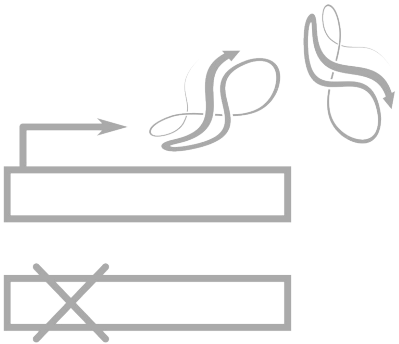


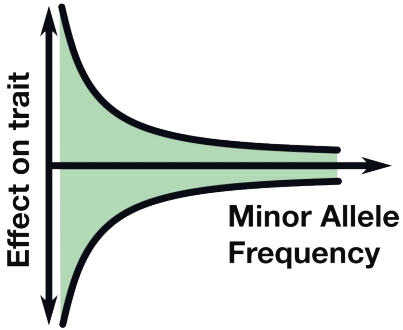


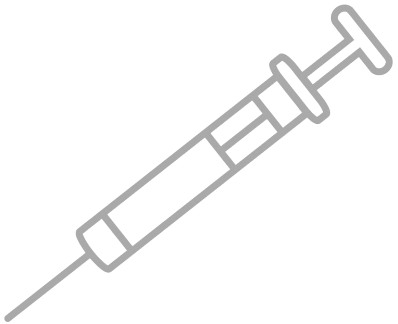


Why study rare genetic variation?

1. Interpretability of results
2. **Large effects**
3. Translational opportunities

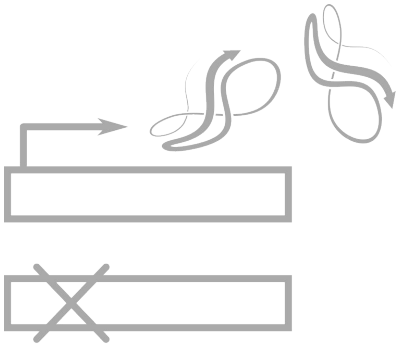


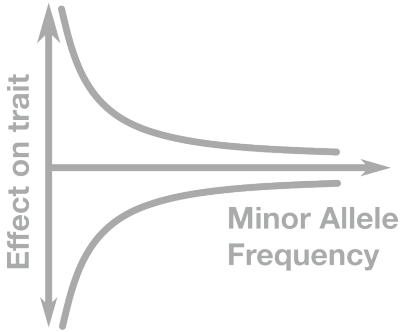


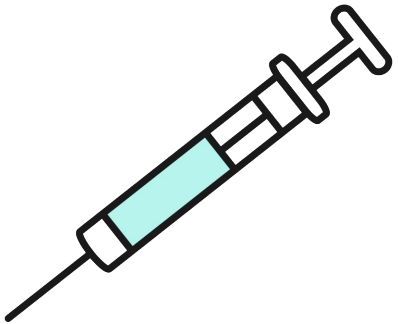


Why study rare genetic variation?

1. Interpretability of results
2. Large effects
3. **Translational opportunities**







Why study rare genetic variation?

Variants with large effects on protein function are often rare, due to **negative selection.**

Largest impact (loss-of-function)

Proportion of variants that are very rare

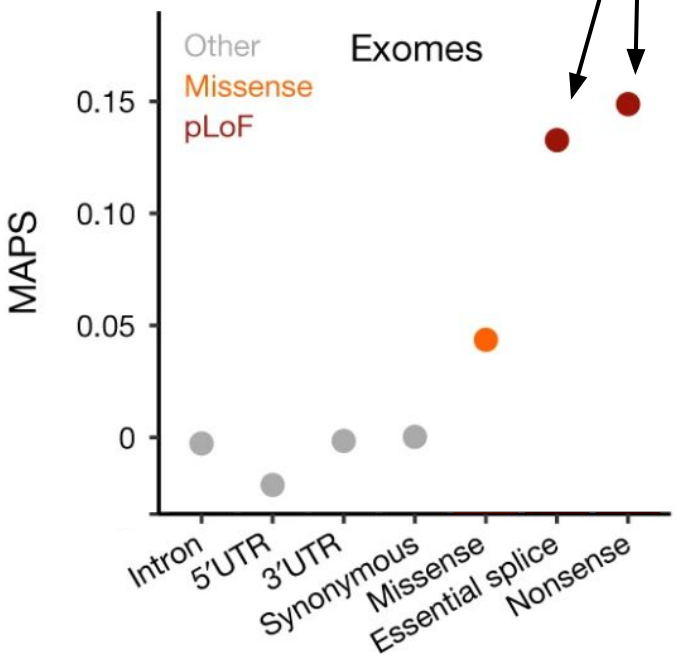


Figure adapted from Karczewski et al. *Nature* 2020

Single-variant association tests are underpowered for rare variants

Sample size required to observe a variant with MAF = *p* with > 99.9% chance is…

MAF	0.1	0.01	0.001	0.0001
N	33	344	3,453	34,537

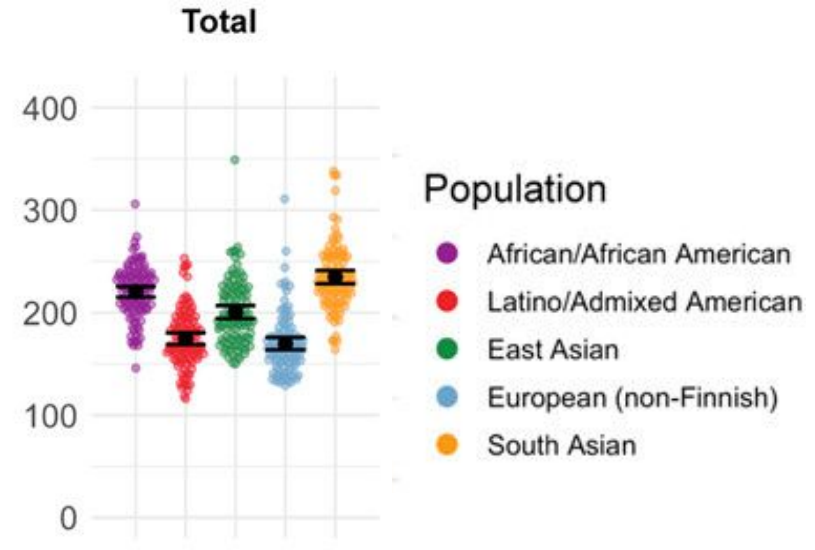
Why study rare genetic variation?

Variants with large effects on protein function are often rare, due to **negative selection.**

Collectively, rare variants are very common:

~200 very rare (MAF<0.1%) coding variants per person

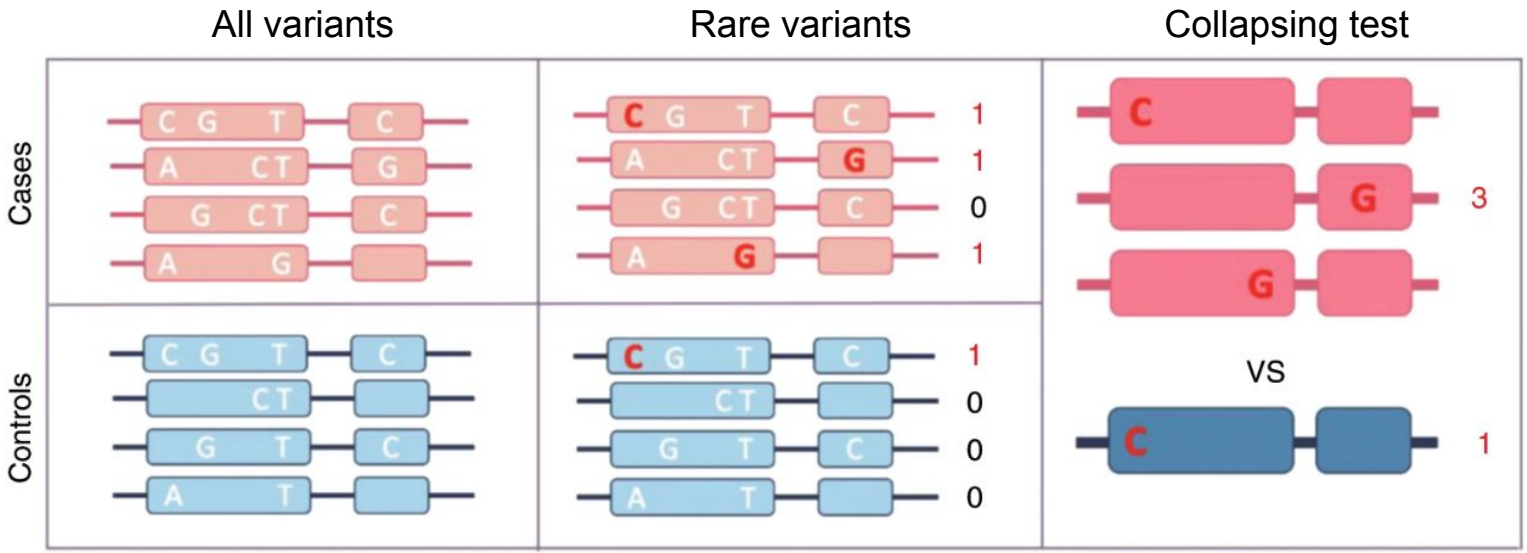
Rare variants (MAF<0.1%)



Gudmondsson et al., 2021 “Variant interpretation using population databases: Lessons from gnomAD”

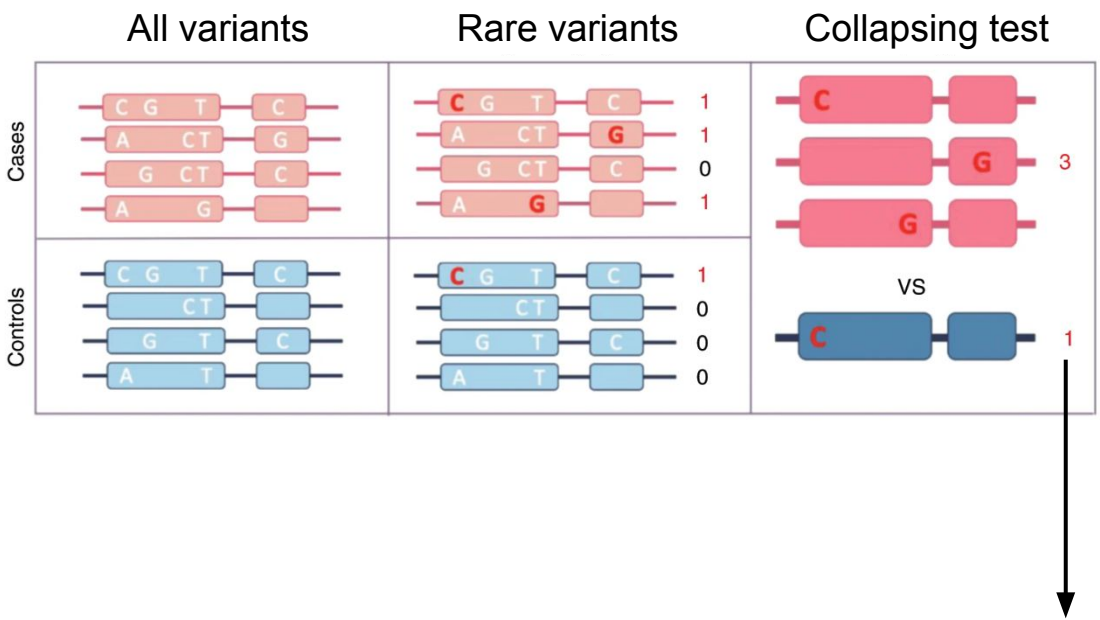
Rare variant collapsing tests

Rather than testing individual variants, we can **aggregate** across a gene or functional unit



Adapted from Cirulli et al. *Nat Comms* 2020

Simple burden test

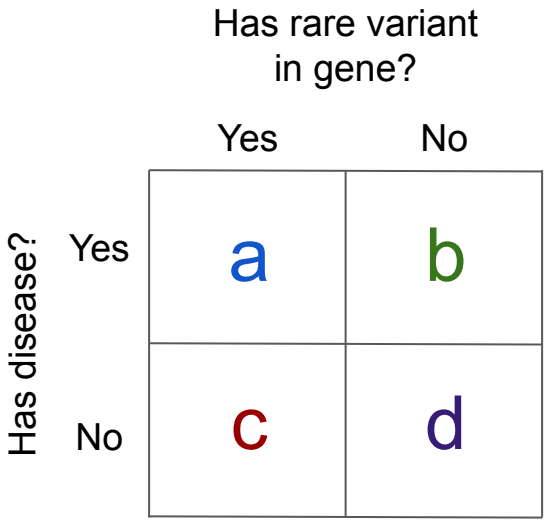


Has rare variant in gene?

Has disease?	Yes	No
Yes	3	1
No	1	3

Use these counts in a 2x2 table

Genetic association testing



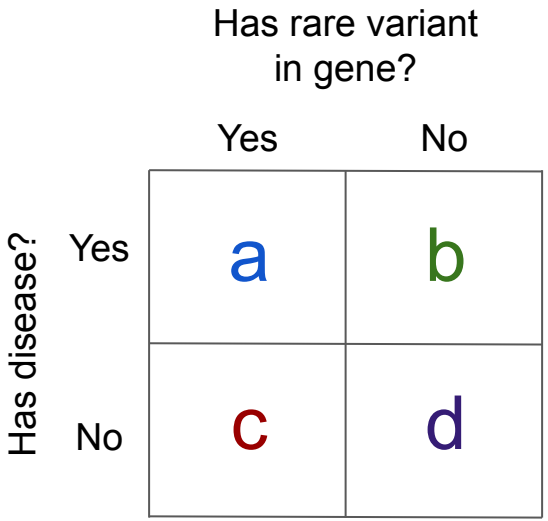
Genotype is associated if prevalence of disease is different between genotype carriers vs. non-carriers:

Odds ratio =

$$\frac{\text{Case/control ratio in carriers}}{\text{Case/control ratio in non-carriers}}$$

$$= \frac{a/c}{b/d} = \frac{a/b}{c/d}$$

Genetic association testing

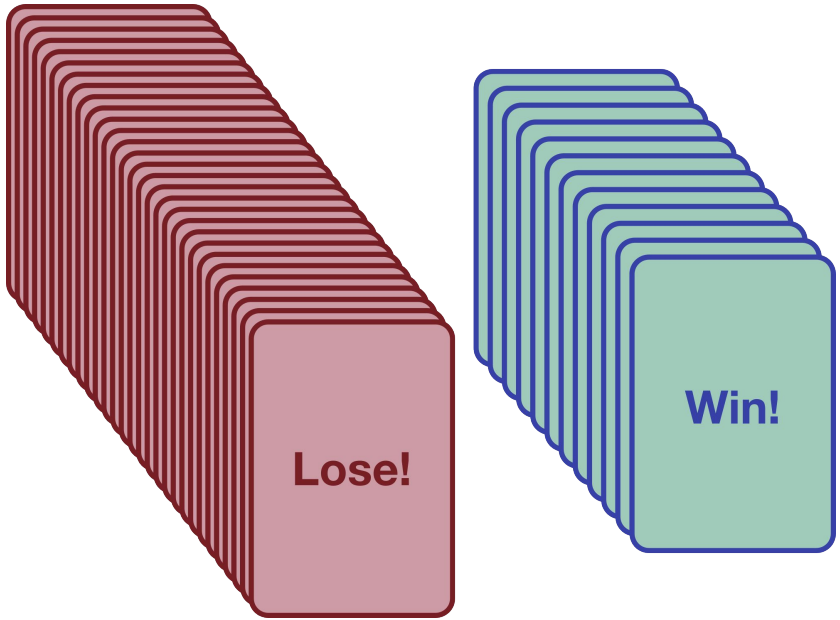


Calculating statistical significance:

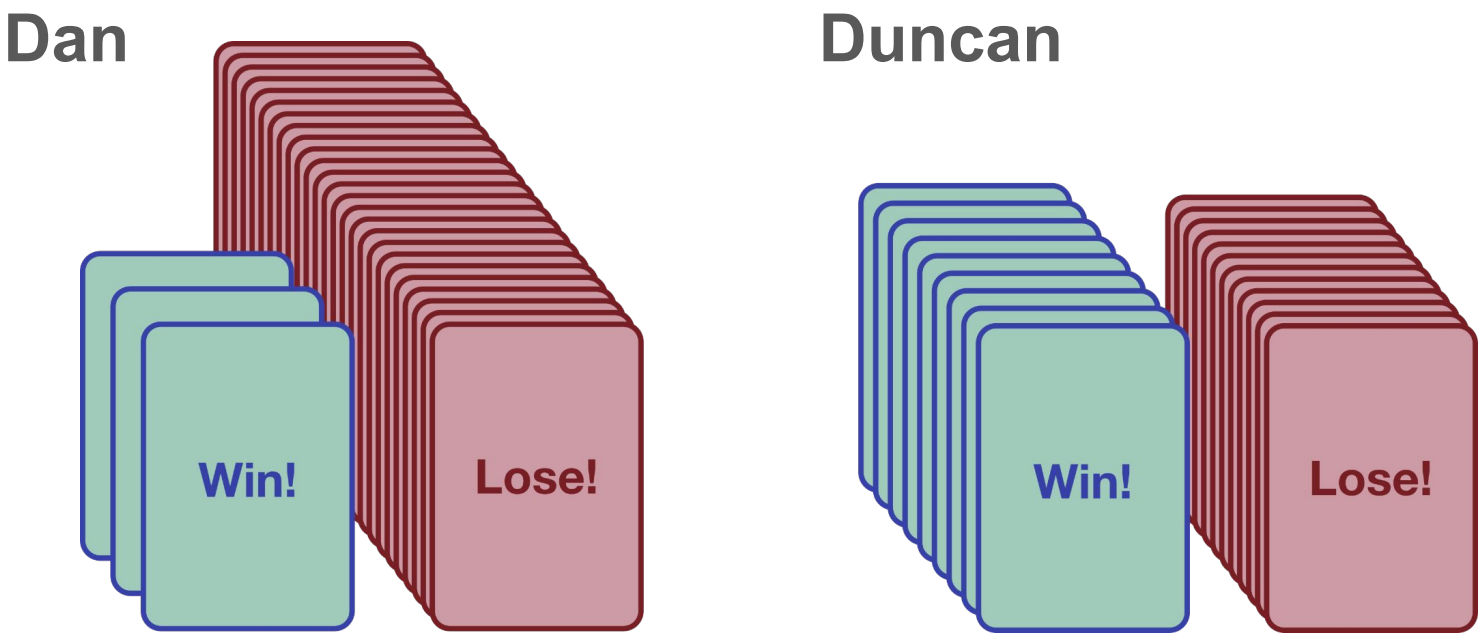
Fisher's Exact Test

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!\ b!\ c!\ d!\ (a+b+c+d)!}$$

Genetic association testing



Genetic association testing



Genetic association testing

	scratch card win? Yes	No
Dan	3	25
Duncan	9	12

$$\frac{\binom{28}{3}\binom{21}{9}}{\binom{49}{12}}$$

Genetic association testing

	scratch card win? Yes	No
Dan	3	25
Duncan	9	12

$$P = 0.017 \quad \frac{\binom{28}{3}\binom{21}{9}}{\binom{49}{12}}$$

Methods for region/gene-based tests

Consider the following question: Given *N* independent observations, and suppose we know:

- Phenotype we're interested in
- Covariates we need to adjust
- Genotype information of rare variants in a region

Can we get an appropriate *P*-value for association between rare variation in this region, and the phenotype?

Model

For **continuous** traits

$$Y = X\alpha + G_1\beta_1 + G_2\beta_2 + \ldots G_q\beta_q + \varepsilon$$

For **binary** traits

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = X\alpha + G_1\beta_1 + G_2\beta_2 + \ldots G_q\beta_q$$

π : probability of having disease given *X* and *G*

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$$

Model for burden tests

Reduce the number of degrees of freedom

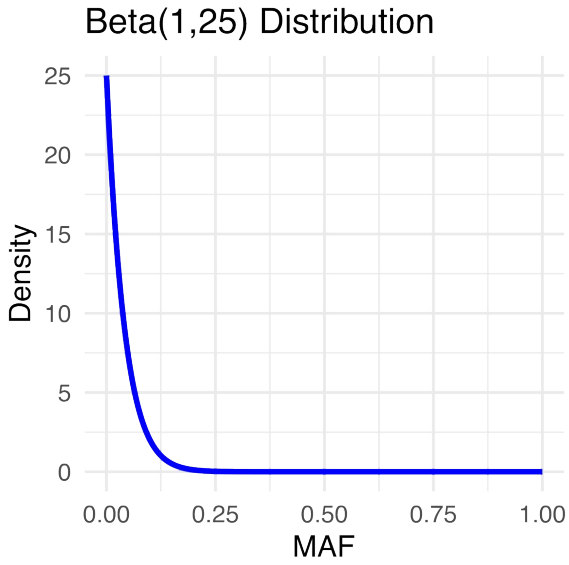
$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = X\alpha + \beta_c\left(G_1 + G_2 + \ldots G_q\right)$$

Incorporate weights for each variant

$$logit(\pi) = X\alpha + \beta_c\left(w_1G_1 + w_2G_2 + \ldots w_qG_q\right)$$

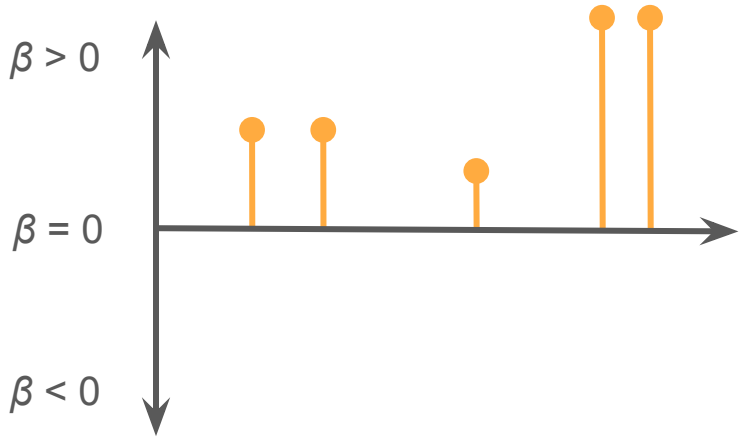
$$w_j \sim Beta(MAF, 1, 25)$$

π : probability of having disease given *X* and *G*



Score test to look for association between rare variant burden and a binary trait

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

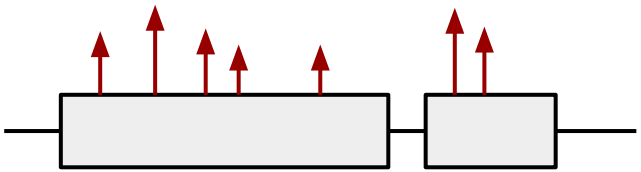


Rare variant collapsing tests

Burden tests

(unidirectional, linear statistics)

- CAST [Morgenthaler & Thilly, 2007]
- CMC [Li & Leal, 2008]
- w-Sum [Madsen & Browning, 2009]
- SST [Morris & Zeggini, 2010]
- VT [Price et al., 2010]



Best powered when:

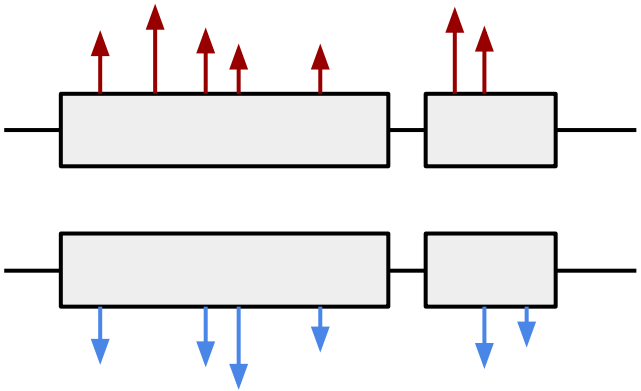
- All variants are **causal**
- All variants have the **same** direction of effect

Rare variant collapsing tests

Burden tests

(unidirectional, linear statistics)

- CAST [Morgenthaler & Thilly, 2007]
- CMC [Li & Leal, 2008]
- w-Sum [Madsen & Browning, 2009]
- SST [Morris & Zeggini, 2010]
- VT [Price et al., 2010]

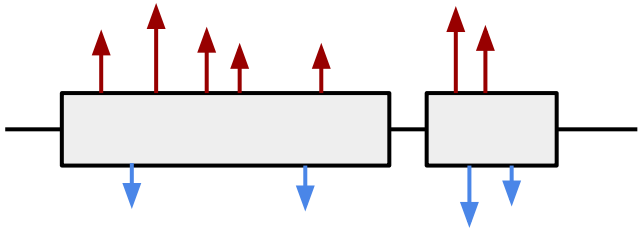


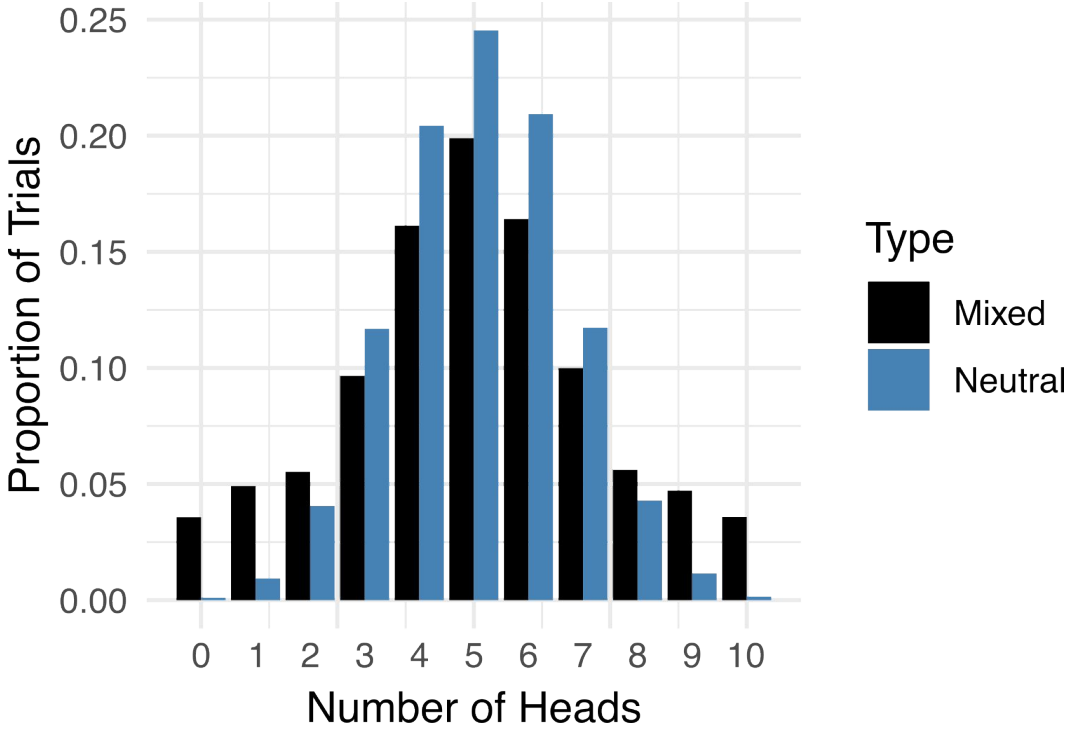
Best powered when:

- All variants are **causal**
- All variants have the **same** direction of effect

What about when the effects vary in direction?

It's like flipping a bunch of biased coins

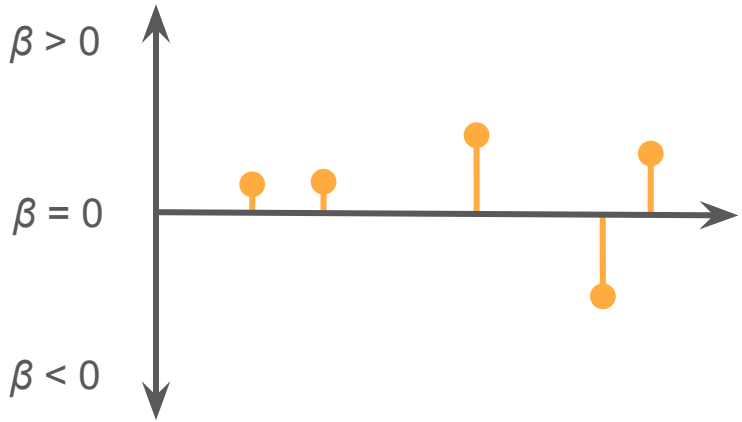




Neale et al., *PLoS Genetics*, 2011

Score test for rare variant association using variance components

$$Q_S = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} g_{i,j} \left(y_i - \hat{\pi}_i \right) \right]^2$$

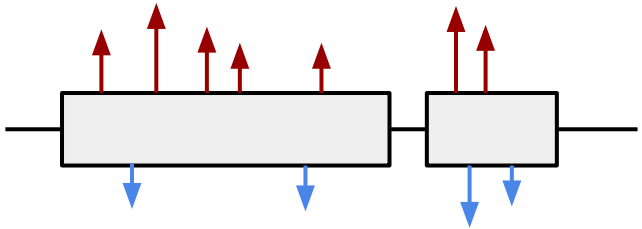


Collapsing tests

Burden tests
(unidirectional, linear statistics)
- CAST [Morgenthaler & Thilly, 2007]
- CMC [Li & Leal, 2008]
- w-Sum [Madsen & Browning, 2009]
- SST [Morris & Zeggini, 2010]
- VT [Price et al., 2010]

Better powered when:
- Not all variants are causal
- Variants have **different** directions of effect

Variance component
(bidirectional, quadratic statistics)
- C-alpha [Neale et al., 2011]
- SKAT [Wu et al., 2011]



Collapsing tests

Burden tests
(unidirectional, linear statistics)
- CAST [Morgenthaler & Thilly, 2007]
- CMC [Li & Leal, 2008]
- w-Sum [Madsen & Browning, 2009]
- SST [Morris & Zeggini, 2010]
- VT [Price et al., 2010]

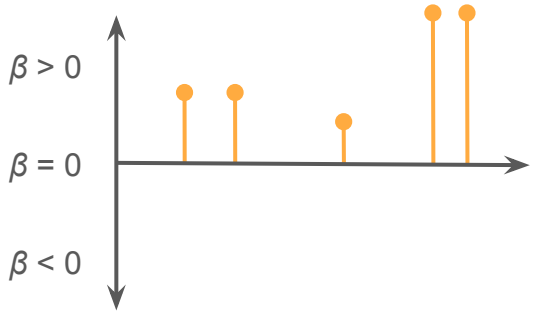
Variance component
(bidirectional, quadratic statistics)
- C-alpha [Neale et al., 2011]
- SKAT [Wu et al., 2011]

Hybrid tests
(combining both)
- **SKAT-O** [Lee et al., 2012]
- Minimum P [Derkach et al., 2013]
- Fisher's statistic [Derkach et al., 2013]

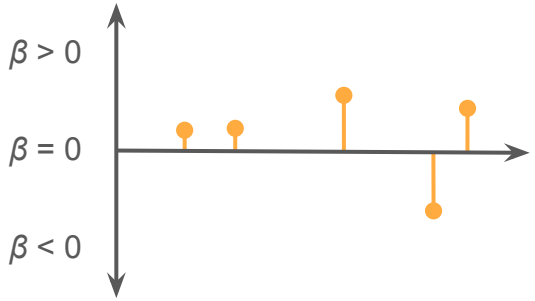
Figure adapted from Sung et al., 2014

What about combining the two?

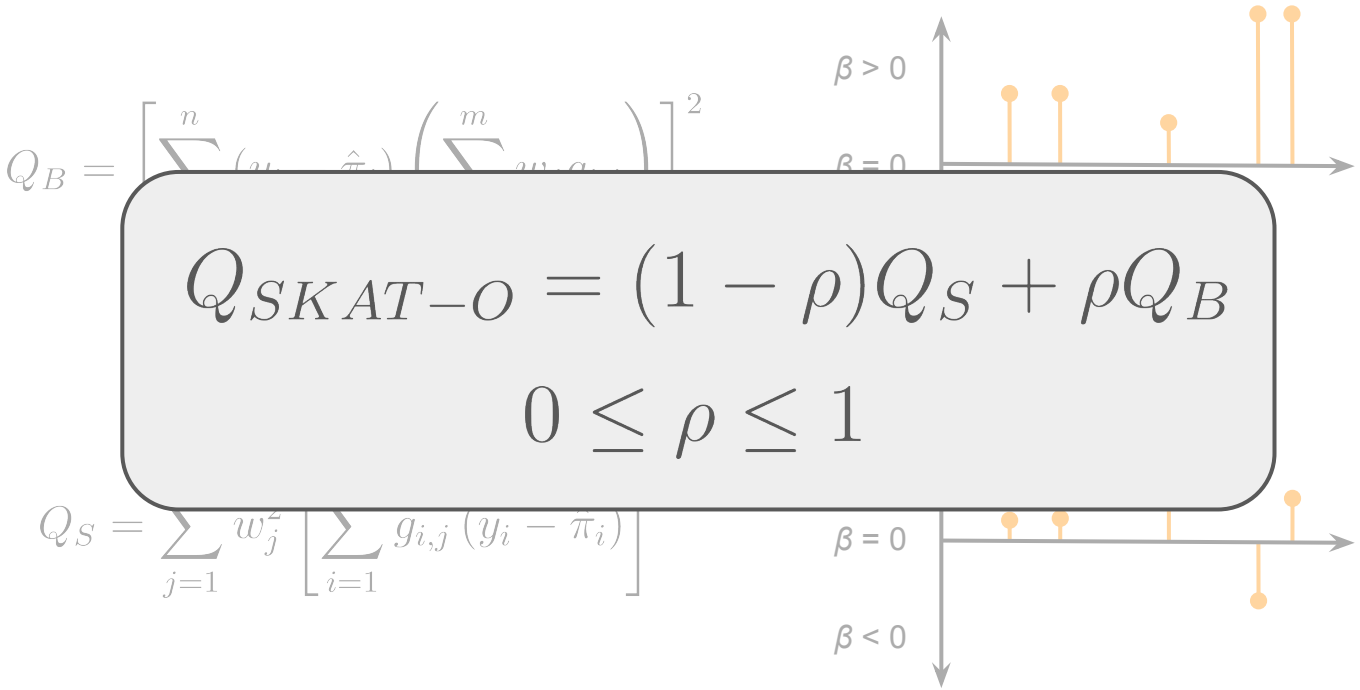
$$Q_B = \left[\sum_{i=1}^{n} (y_i - \hat{\pi}_i) \left(\sum_{j=1}^{m} w_j g_{i,j}\right)\right]^2$$



$$Q_S = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} g_{i,j} (y_i - \hat{\pi}_i)\right]^2$$



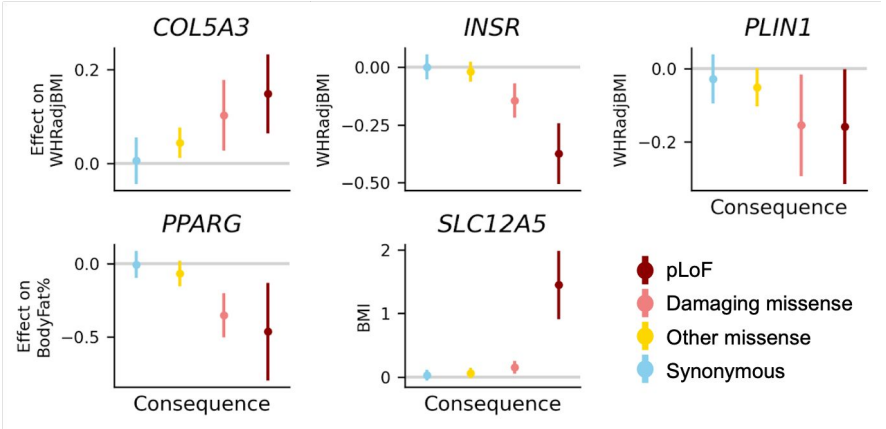
SKAT-O is more powerful that Burden and SKAT



Variants can be split into functional ‘categories’

For many genes full **loss-of-function** may be needed to cause disease

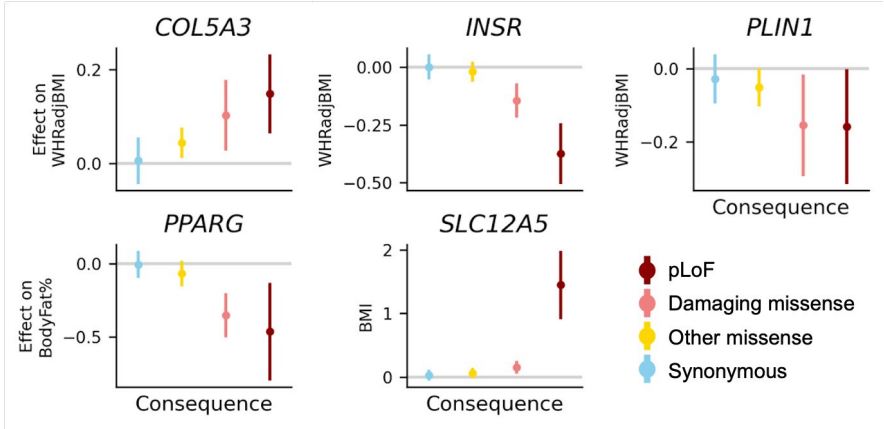
For others, an effect may be limited to **missense** variants impacting a specific protein domain

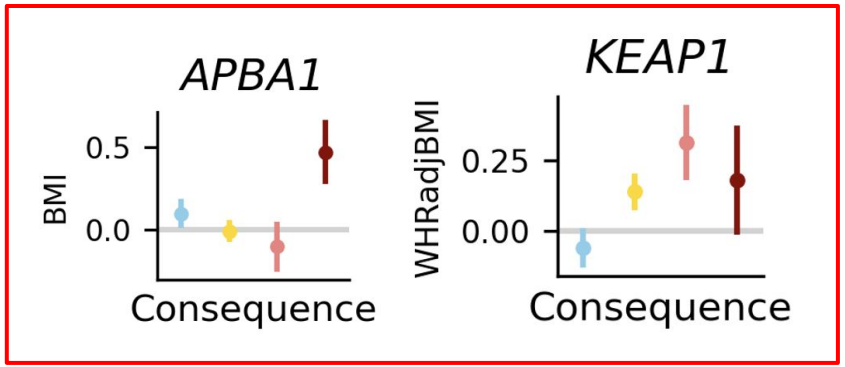


Variants can be split into functional ‘categories’

For many genes full **loss-of-function** may be needed to cause disease

For others, an effect may be limited to **missense** variants impacting a specific protein domain





Software for gene-level testing

“Perhaps the single greatest challenge facing rare-variant analyses is the issue of scalability”

- Povysil et al. *Nature Reviews Genetics* 2019

Software for gene-level testing

SAIGE-GENE

Zhou et al. *Nature Genetics* 2020

Technical Report | Published: 18 May 2020

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

Wei Zhou, Zhangchen Zhao, Jonas B. Nielsen, Lars G. Fritsche, Jonathon LeFaive, Sarah A. Gagliano Taliun, Wenjian Bi, Maiken E. Gabrielsen, Mark J. Daly, Benjamin M. Neale, Kristian Hveem, Goncalo R. Abecasis, Cristen J. Willer & Seunggeun Lee

Nature Genetics **52**, 634–639 (2020) | Cite this article

10k Accesses | **103** Citations | **16** Altmetric | Metrics

Regenie

Mbatchou et al. *Nature Genetics* 2021

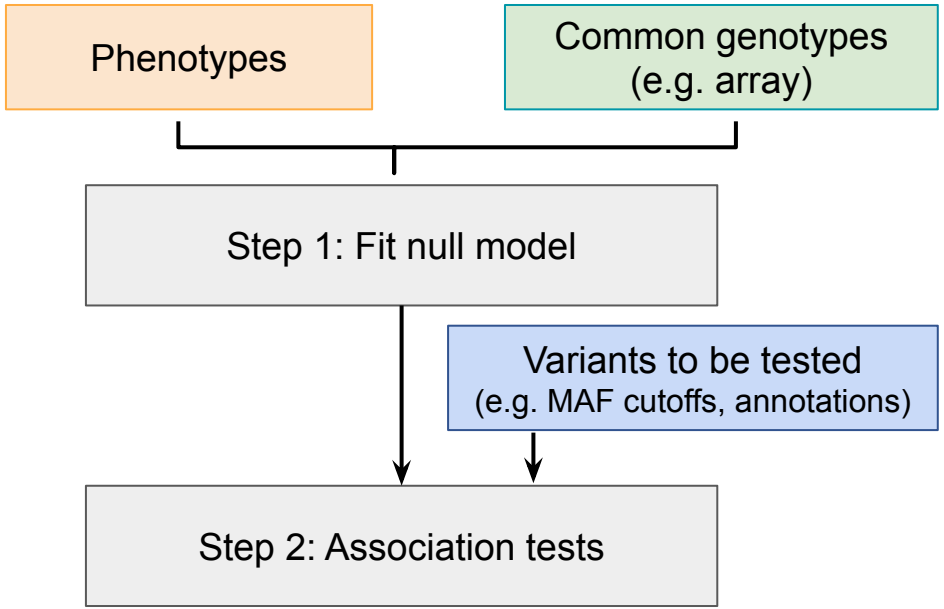
Technical Report | Published: 20 May 2021

Computationally efficient whole-genome regression for quantitative and binary traits

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell & Jonathan Marchini

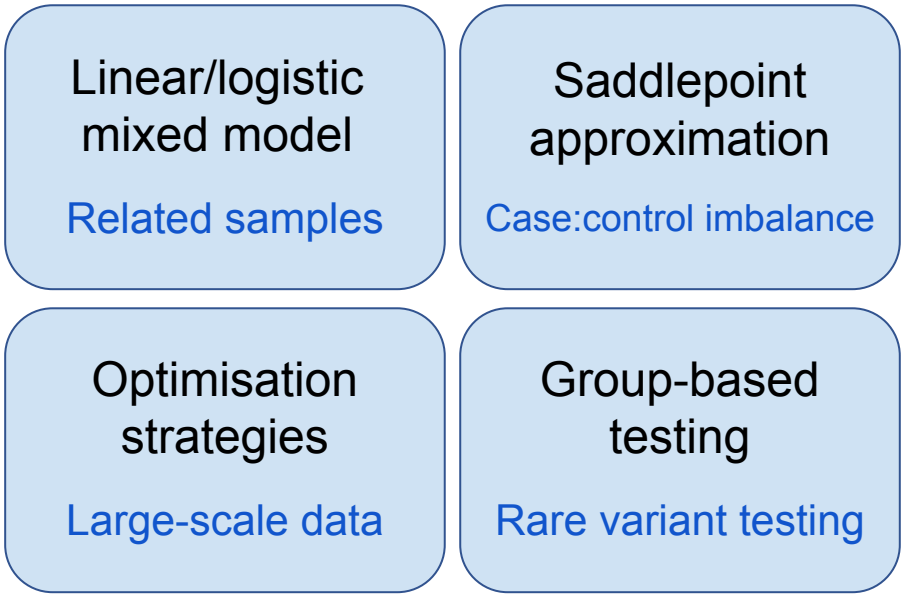
Nature Genetics **53**, 1097–1103 (2021) | Cite this article

67k Accesses | **395** Citations | **41** Altmetric | Metrics



Note: Possible (but now outdated) to code gene burden as variant-level genotypes and use variant-level testing frameworks (see Cirulli et al., 2020 using BOLT-LMM)

SAIGE-GENE+



Improved computational efficiency

Improved type 1 error

Improved power

Multiple functional annotations, e.g.

- LoF
- LoF+non-synonymous

Multiple max-MAF cutoffs, e.g.

- 0.01%
- 0.1%
- 1%

Zhou*, Bi*, Zhao* et al., *Nature Genetics*, 2022

Recent applications

Regeneron UKB 450k exomes

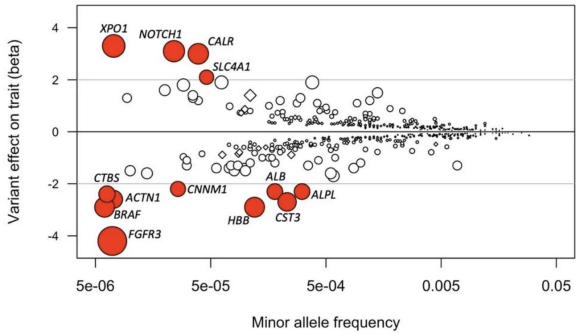
Backman et al. *Nature* 2021

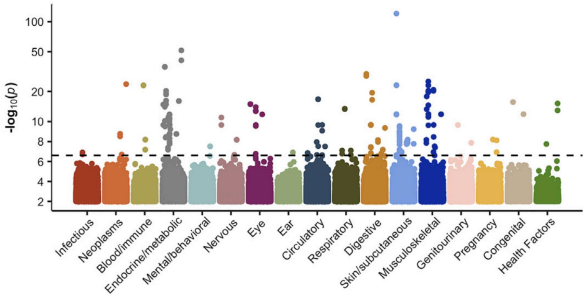
Genebass UKB 450k exomes

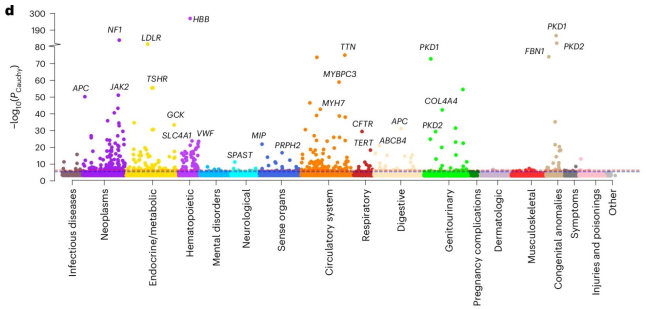
Karczewski et al. *Cell Genomics* 2022

Pan-ancestry ~750k (UKB, AoU, MGB)

Jurgens et al. *Nature Genetics* 2024







Recent applications

Regeneron UKB 450k exomes

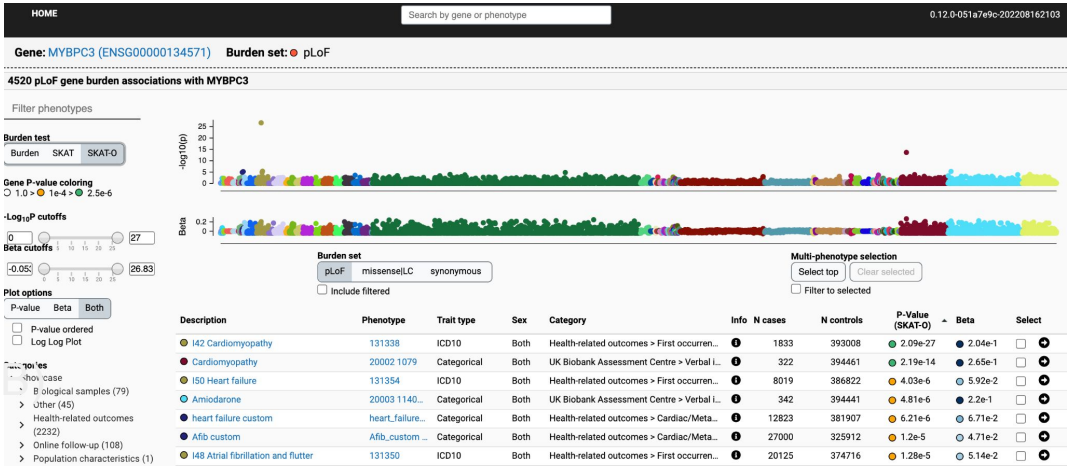
Backman et al. *Nature* 2021

Genebass

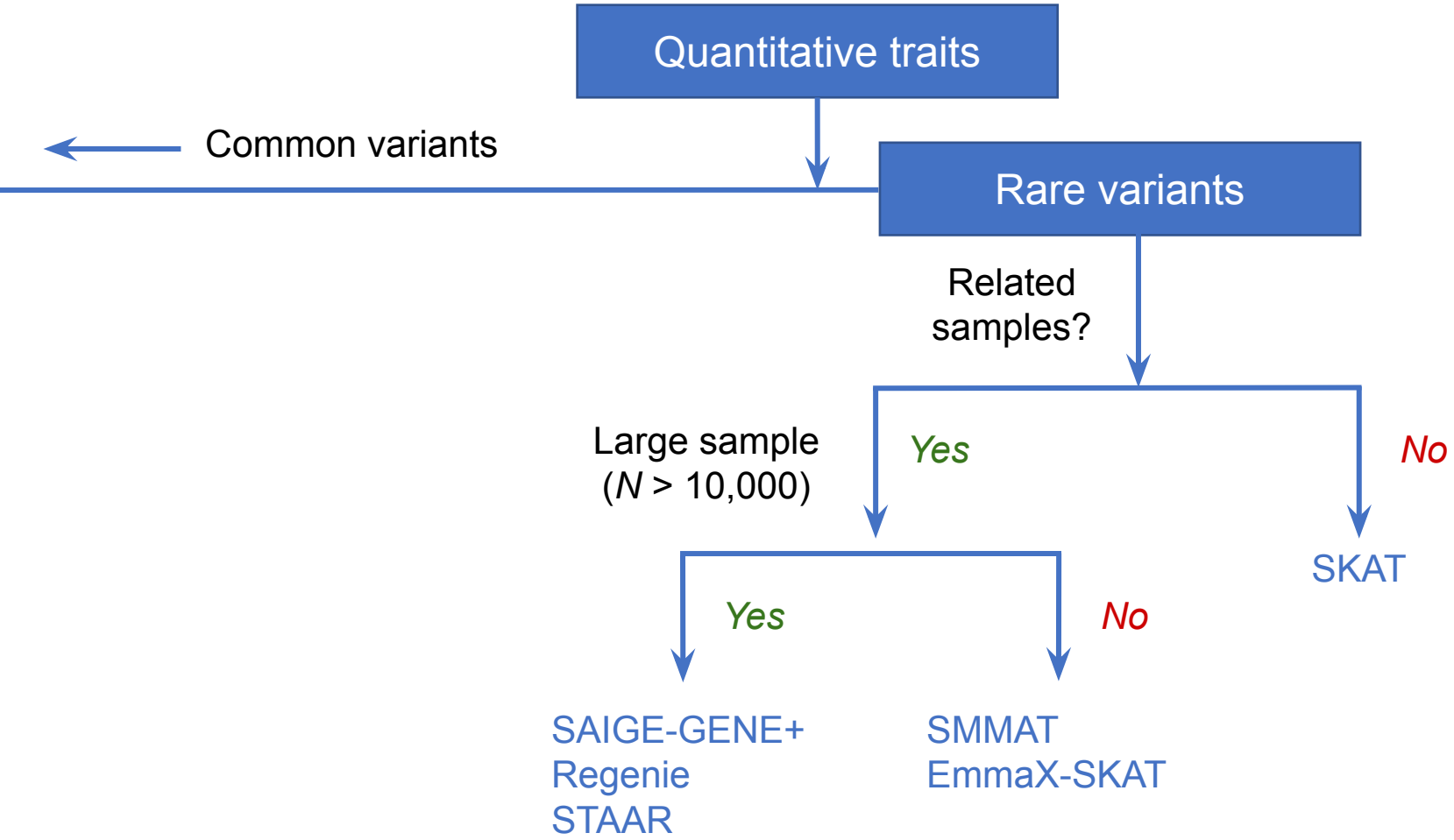
Karczewski et al. *Cell Genomics* 2022

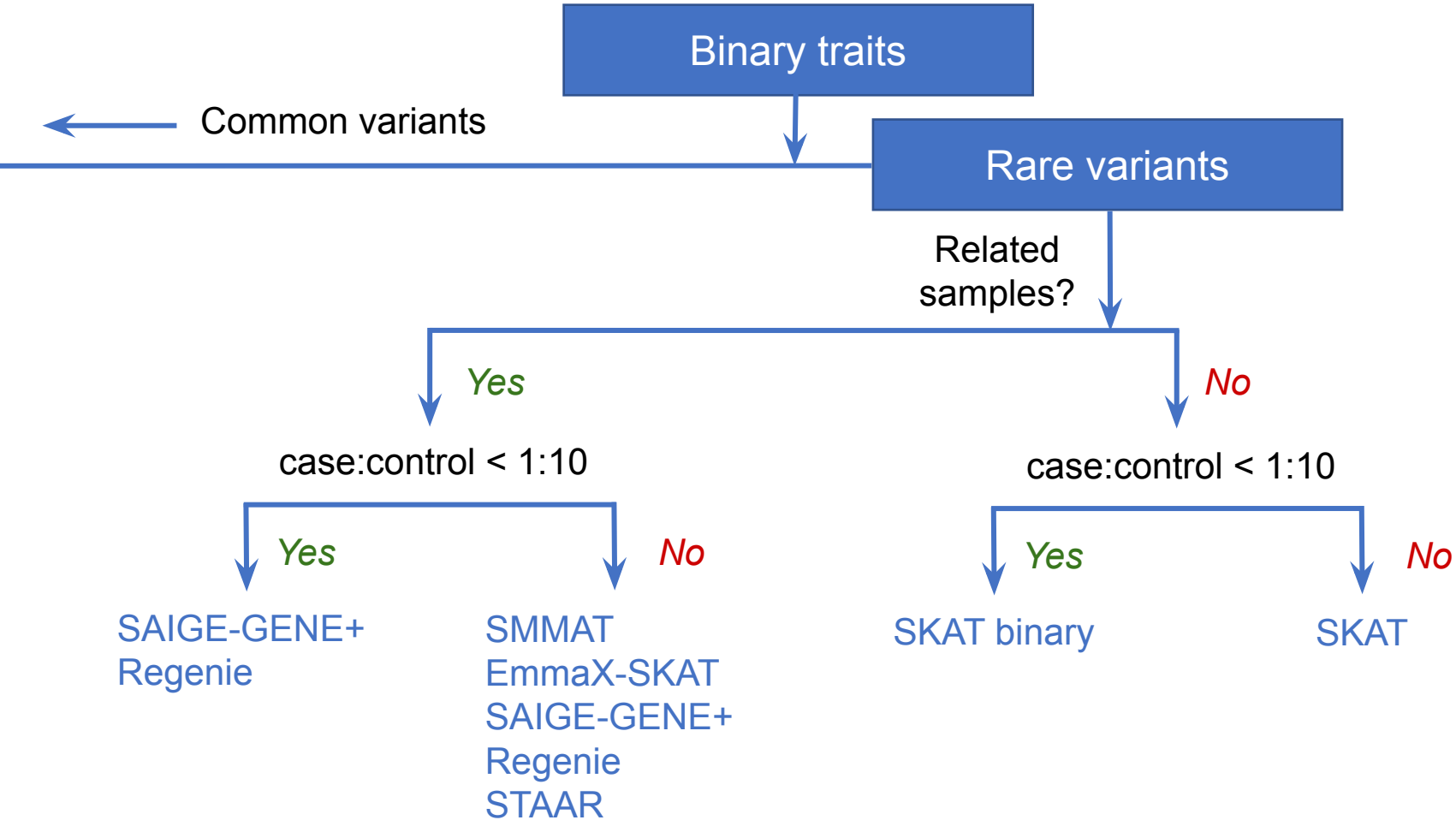
Pan-ancestry ~750k (UKB, AoU, MGB)

Jurgens et al. *Nature Genetics* 2024



app.genebass.org


Quantitative traits
Common variants
Rare variants
Related samples?
Large sample (N > 10,000)
Yes
No
SKAT
Yes
No
SAIGE-GENE+
Regenie
STAAR
SMMAT
EmmaX-SKAT


Binary traits
Common variants
Rare variants
Related samples?
Yes
No
case:control < 1:10
case:control < 1:10
Yes
No
Yes
No
SAIGE-GENE+
Regenie
SMMAT
EmmaX-SKAT
SAIGE-GENE+
Regenie
STAAR
SKAT binary
SKAT

Lecture summary

1. Low power to detect single-variant associations with low frequency
2. Grouping rare variants increases power
3. Various methods to test grouped variants (e.g. burden, SKAT, SKAT-O)
4. Various software to make this scalable (e.g. SAIGE, Regenie)

Next up: Put this knowledge into practice and do some association testing!

Questions and instructions are here:

https://qimr.az1.qualtrics.com/jfe/form/SV_5vZEC9z5y2RXckS

Slides adapted from earlier versions from Nicky Whiffin and Nik Baya

Reading list

Review articles:

- Lee et al., 2014 *
- Povysil et al., 2019 ***

Applications:

- Questions:
  - What methods are used to annotate variants with consequence? What variant masks are used?
  - What association tests are used? (e.g. burden, SKAT, SKAT-O)
  - What multiple testing correction do they use?
- Cirulli et al., Nature Comms 2020 (first phenome-wide burden testing: 4.2k phenos, 50k + 22k individuals)
- Backman et al. Nature 2021 (Regeneron 450k WES flagship)
- Karczewski et al. Cell Genomics 2022 (Genebass, 4.5k phenos, 400k individuals)
- Jurgens et al. Nature Genetics 2024 (Pan-ancestry, 601 diseases, ~750k individuals across 3 biobanks)

Methods:

- CMC (early burden testing) – Li & Leal 2008
- SKAT-O: Lee et al., 2012