

# Genetic Association Tests for Common + Rare Variants

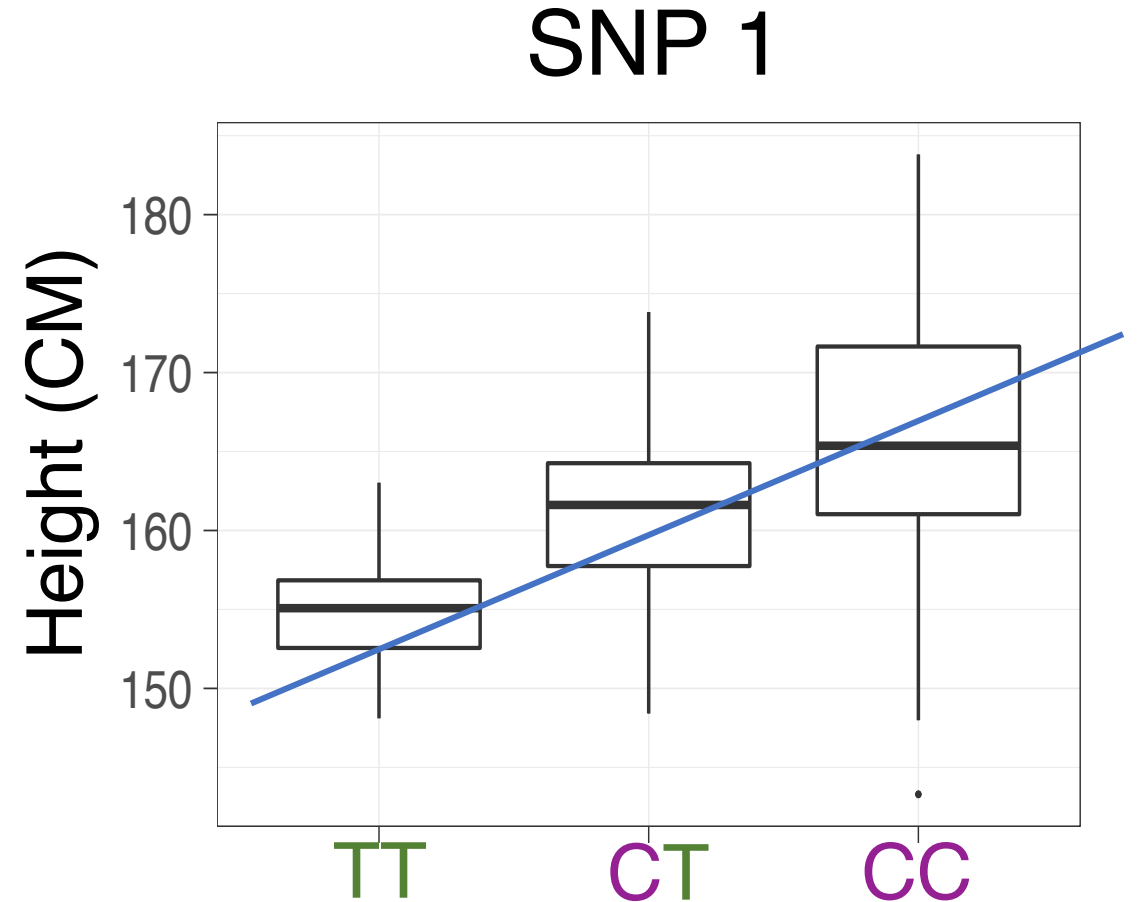
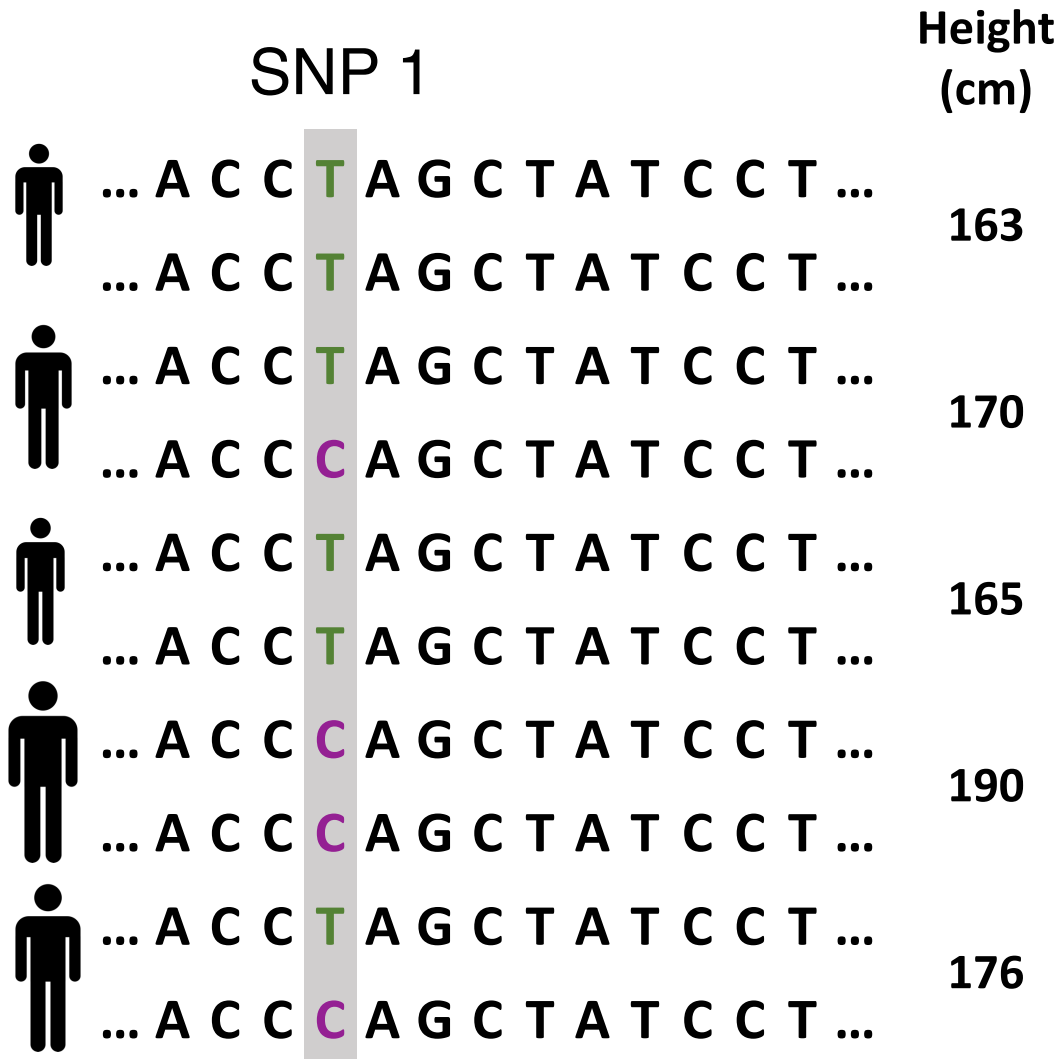
2023 International Statistical Genetics Workshop

March 6, 2023

Wei Zhou, Ph.D.

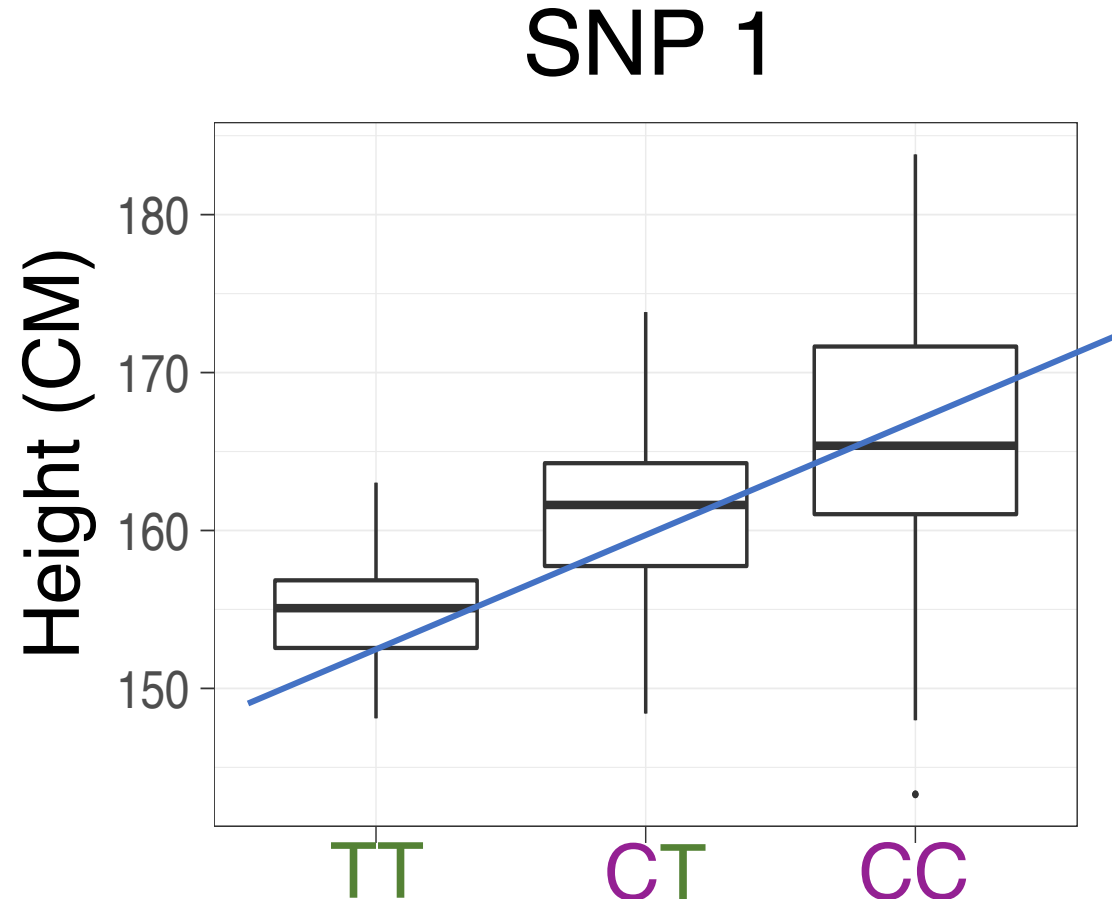
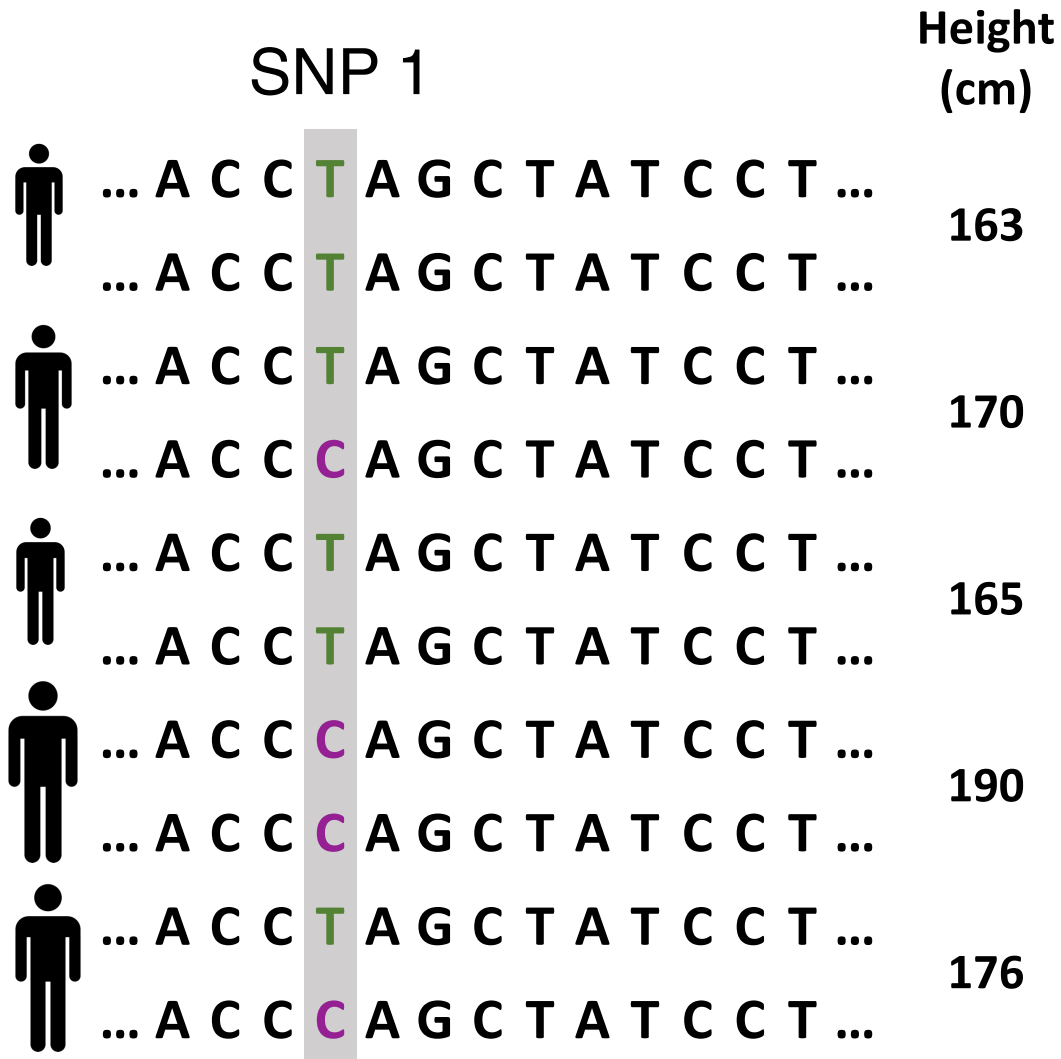
# Genetic association tests

To identify genetic variants that are associated with **a complex trait**



# Genetic association tests

To identify genetic variants that are associated with **a complex trait**



Linear regression:  $Y = X\alpha + G\beta + \epsilon$

$G = 0, 1, \text{ or } 2$

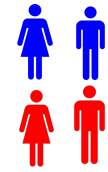
$X$ : covariates, e.g. age, sex, ancestry, batch...

$H_0: \beta = 0$

$H_1: \beta \neq 0$

# Genetic association tests











To identify genetic variants that are associated with a complex disease/disorder



No

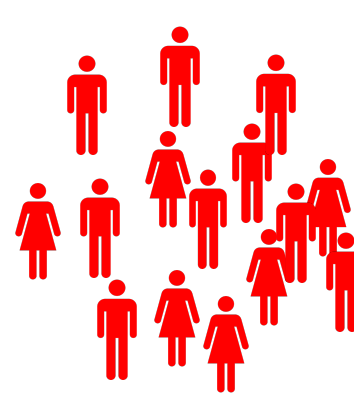
Yes

SNP 1

	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...

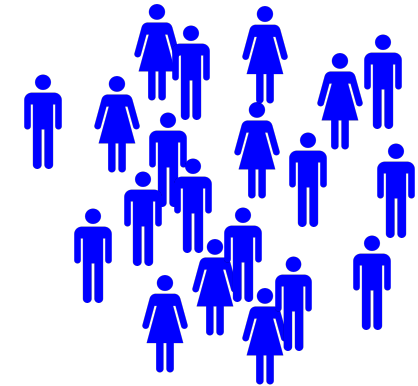
SNP 1

Frequency of C



Cases  
 $2/4 = 50\%$

VS.

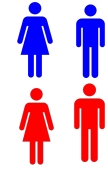


Controls:  
 $1/6 = 16.7\%$



# Genetic association tests

To identify genetic variants that are associated with **a complex disease/disorder**



No

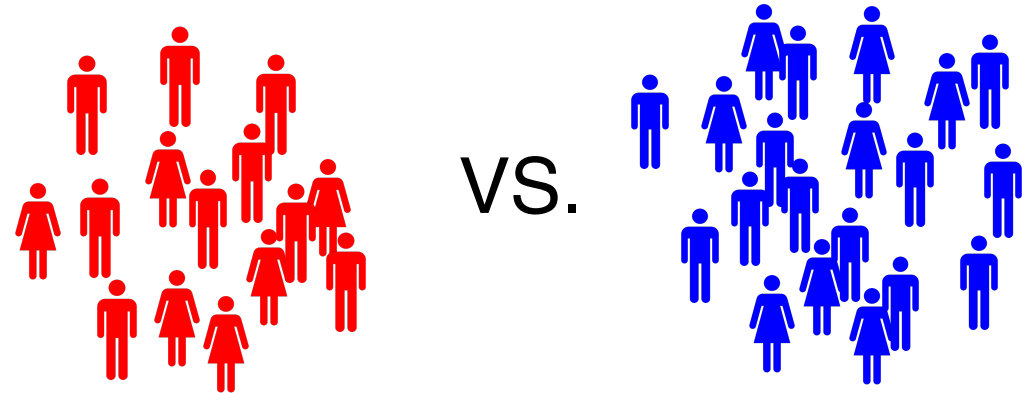
Yes

SNP 1

	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...

SNP 1

Frequency of C



VS.

Cases

2/4 = 50%

Controls:

1/6 = 16.7%

Logistic regression:  $logit(\pi) = X\alpha + G\beta$   
 $\pi$ : probability of being a case given  $X$  and  $G$

$G = 0, 1, \text{ or } 2$

$H_0: \beta = 0$

$H_1: \beta \neq 0$

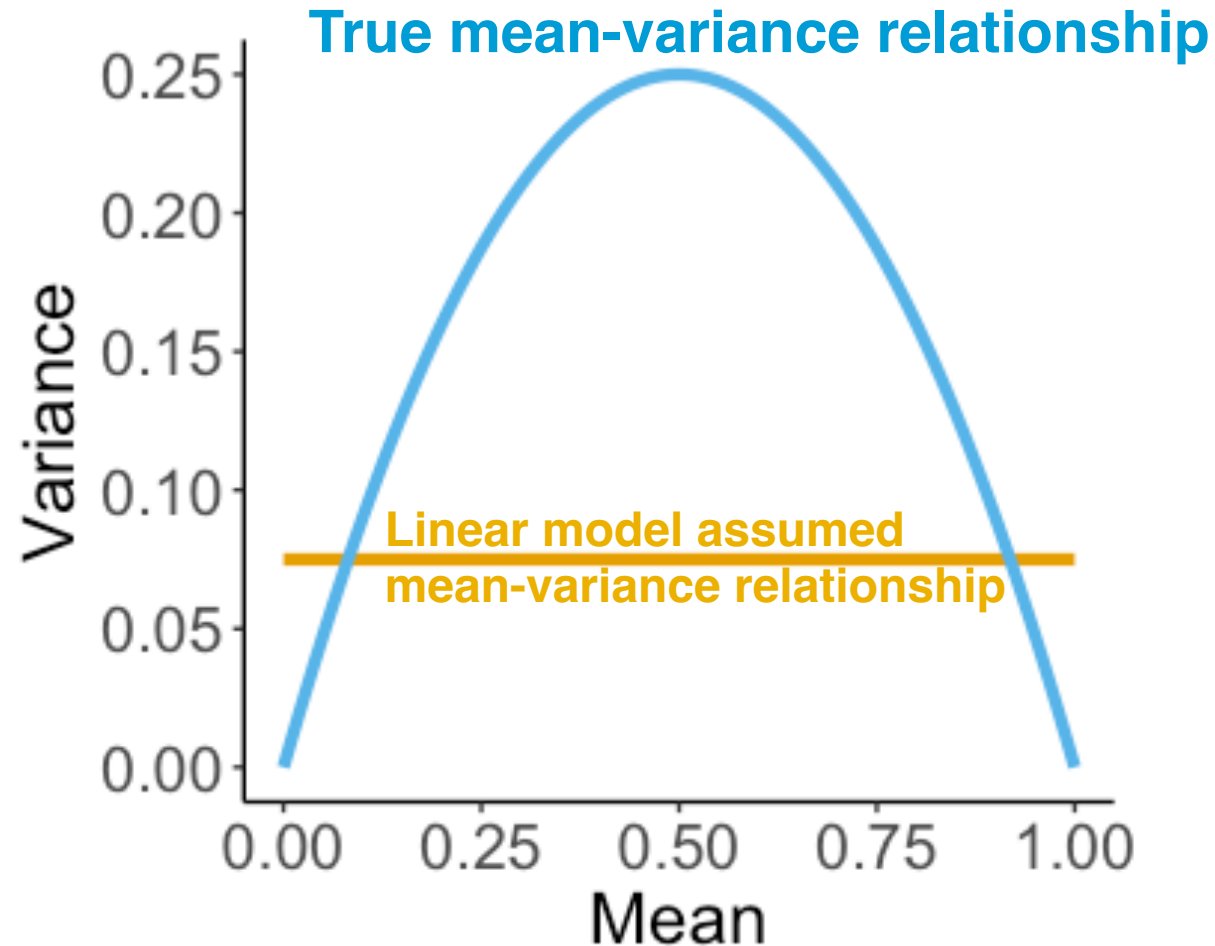
# Different types of phenotypes require different statistical models for association tests

- Quantitative
  - eg. LDL cholesterol level, height
- Binary
  - eg. Schizophrenia, Type 2 Diabetes
- Ordinal/categorical
  - eg. On a scale of 1-10 how much do you like smoking
- Time-to-event (TTE)
  - eg. Age at skin cancer onset, Time of death after diagnosis of lung cancer

# Different types of phenotypes require different statistical models for association tests

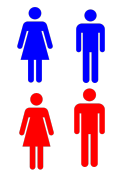
- Quantitative
  - eg. LDL cholesterol level, height
  - Linear regression
- Binary
  - eg. Schizophrenia, Type 2 Diabetes
  - Logistic regression
- Ordinal/categorical
  - eg. On a scale of 1-10 how much do you like smoking
  - Proportional odds logistic regression, Multinomial regression
- Time-to-event (TTE)
  - eg. Age at skin cancer onset, Time of death after diagnosis of lung cancer
  - Survival analysis model

Using linear regression for binary phenotypes (coded as 0 and 1) can lead to inflated type I errors



# Genetic association tests

To identify genetic variants that are associated with **a complex disease/disorder**



No

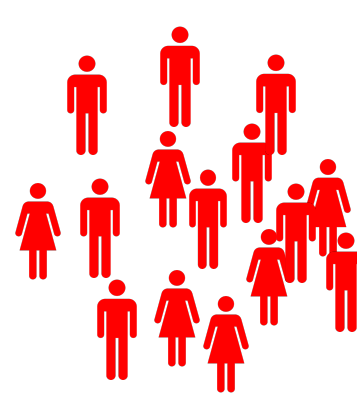
Yes

SNP 1

	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...
	... A C C T A G C T A T C C T ...
	... A C C C A G C T A T C C T ...

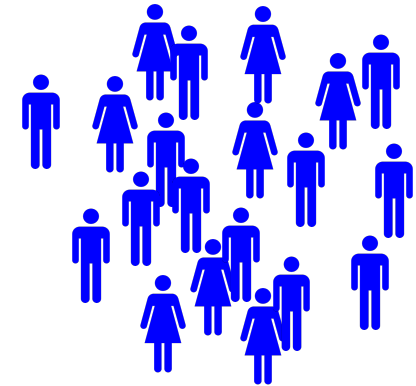
SNP 1

Frequency of C



Cases  
2/4 = 50%

VS.



Controls:  
1/6 = 16.7%

Logistic regression:  $logit(\pi) = X\alpha + G\beta$   
 $\pi$ : probability of being a case given  $X$  and  $G$

$G = 0, 1, \text{ or } 2$

$H_0: \beta = 0$

$H_1: \beta \neq 0$

# Standard asymptotic tests

In the example of testing 10 million genetic markers, one at a time:

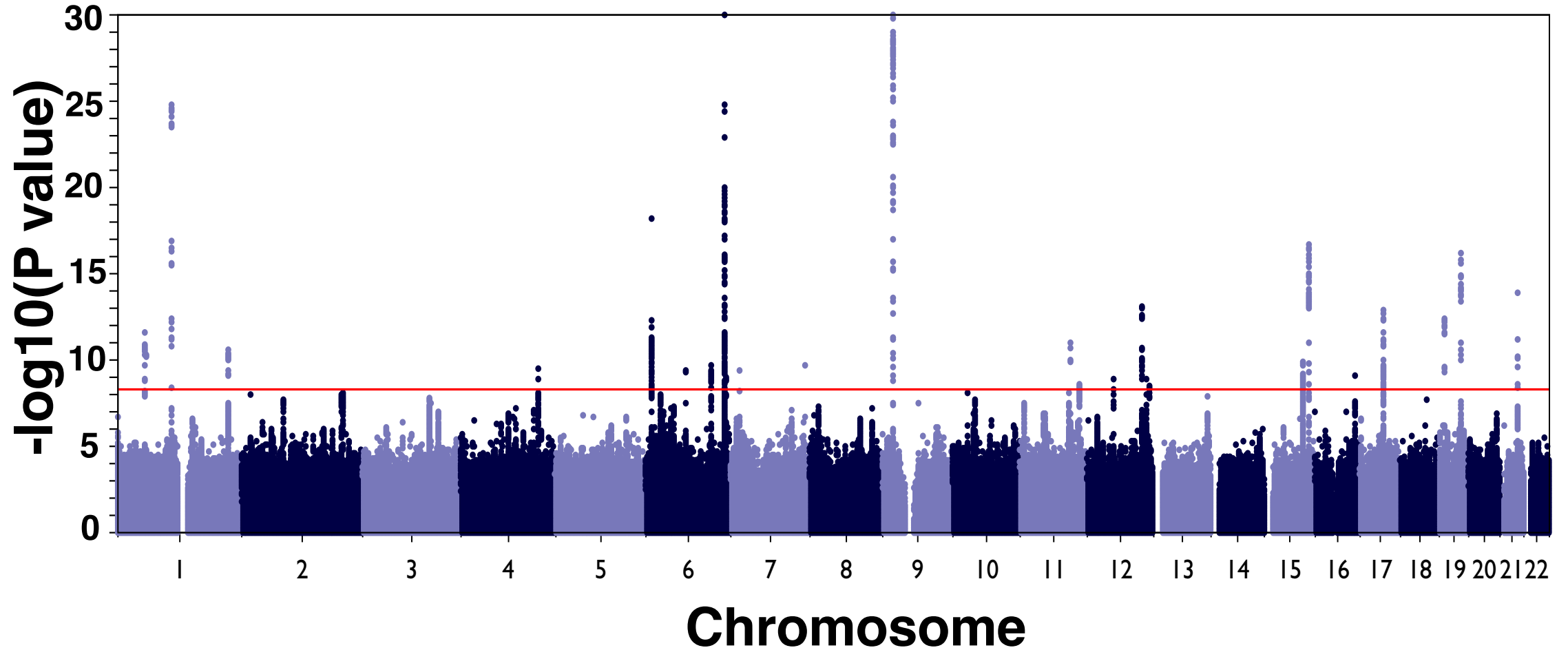
$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Test	Fit null model ( $H_0$ )	Fit full model ( $H_1$ )
Likelihood Ratio	1	10M
Wald	0	10M
<b>Score</b>	<b>1</b>	<b>0</b>

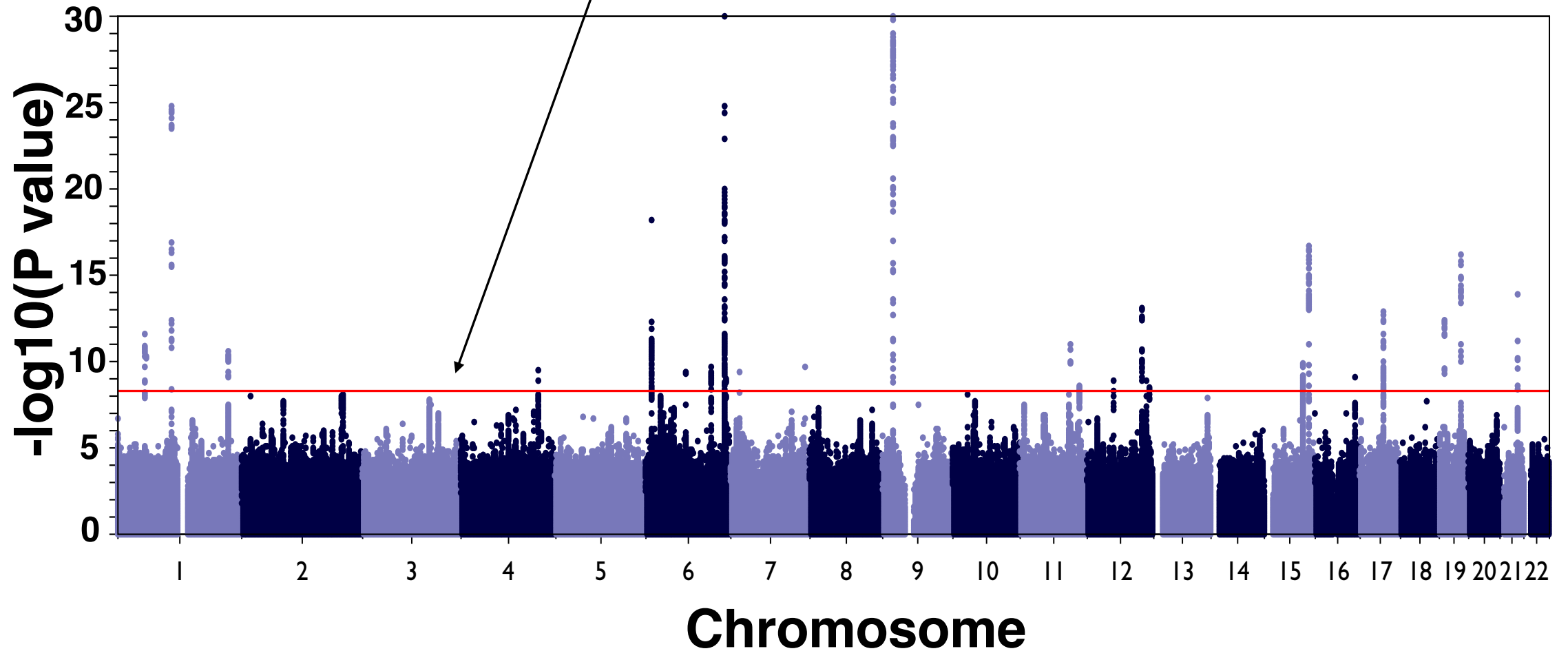
- Clear computational advantage of running score test for GWAS
- Tradeoff of score test: Cannot provide accurate effect-size estimate for  $\beta$
- Solution: Run the GWAS using score test, then only calculate MLE of the effect-sizes for the significant SNPs using Wald for Likelihood Ratio test

# Genome-Wide Association Study (GWAS)



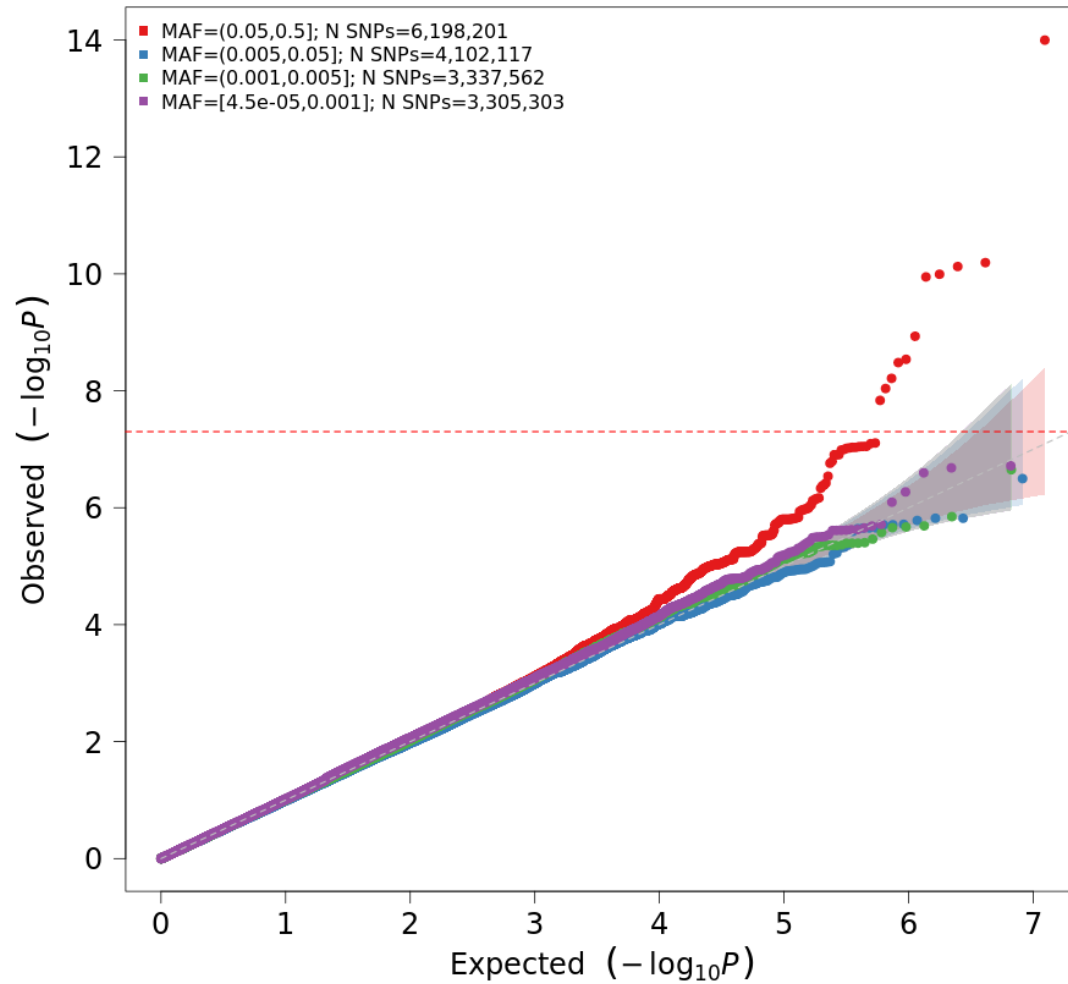
# Genome-Wide Association Study (GWAS)

Genome-wide significant threshold for p-value:  $5 \times 10^{-8}$





# Visualize GWAS: quantile-quantile (QQ) plot



# Different types of phenotypes require different statistical models for association tests

- **Quantitative**

- eg. LDL cholesterol level, height
- Linear regression

- **Binary**

- eg. Schizophrenia, Type 2 Diabetes
- Logistic regression

- **Ordinal/categorical**

- eg. On a scale of 1-10 how much do you like smoking
- Proportional odds logistic regression, Multinomial regression

- **Time-to-event (TTE)**

- eg. Age at skin cancer onset, Time of death after diagnosis of lung cancer
- Survival analysis model

# Challenges in genetic association studies

**Linear model:**

$$Y_i = X_i\alpha + G_i\beta + \epsilon_i$$

**Logistic model:**

$$\text{logit}(\pi_i) = X_i\alpha + G_i\beta$$

$$\epsilon \sim N(0, \sigma^2 I)$$

Assumes independent observations



**Sample relatedness**

**1 in 3 has at least one relative up to the 3<sup>rd</sup> degree in UK Biobank**  
- inflated type I errors

# Mixed models are used for genetic association tests with related samples

**Linear mixed model:**

$$Y_i = X_i\alpha + G_i\beta + b_i + \epsilon_i$$

**Logistic mixed model:**

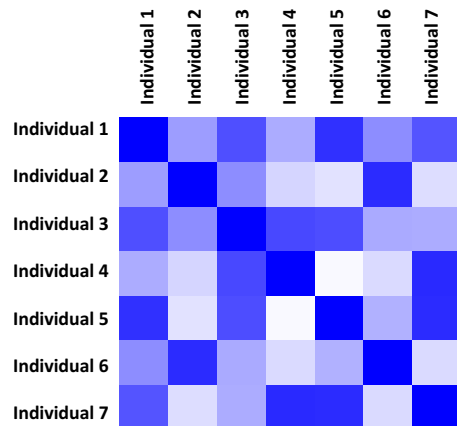
$$\text{logit}(\pi_i) = X_i\alpha + G_i\beta + b_i$$

$b$ : random genetic effect

$b \sim N(0, \tau \psi)$ ,  $\psi$  is **genetic relationship matrix (GRM)**

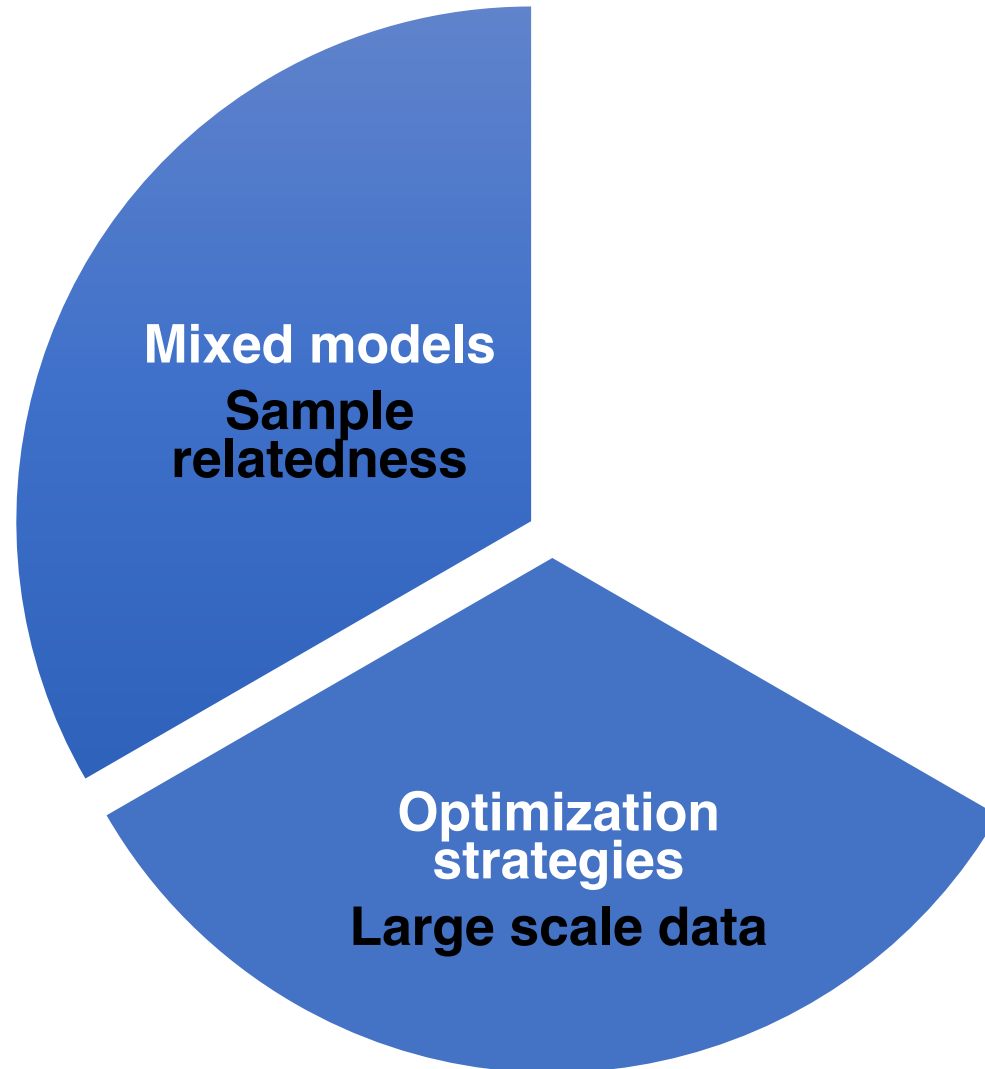
Mixed models  
Sample  
relatedness

**1 in 3 has at least one relative up to the 3<sup>rd</sup> degree in UK Biobank**

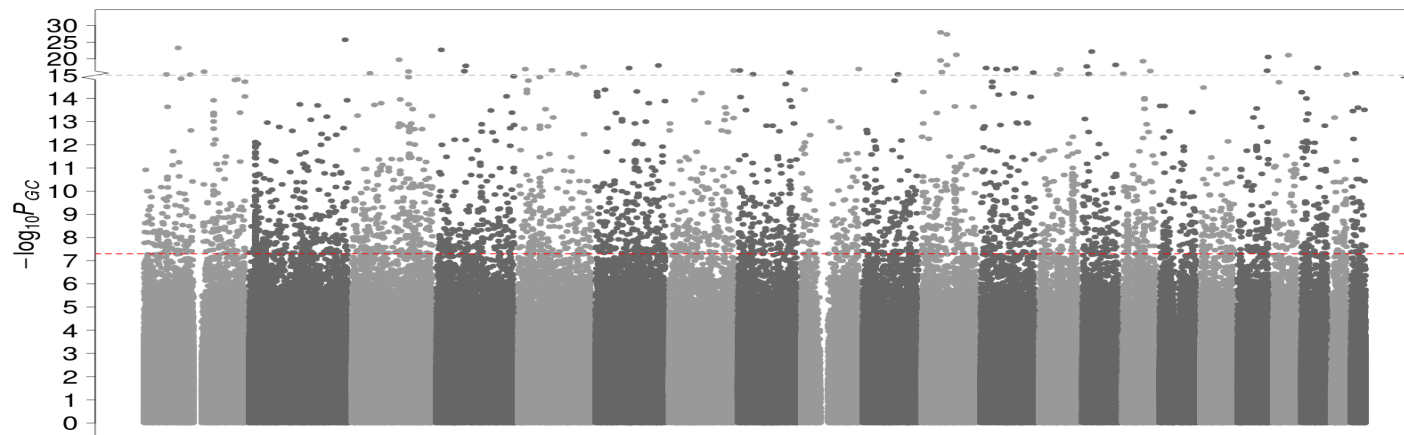




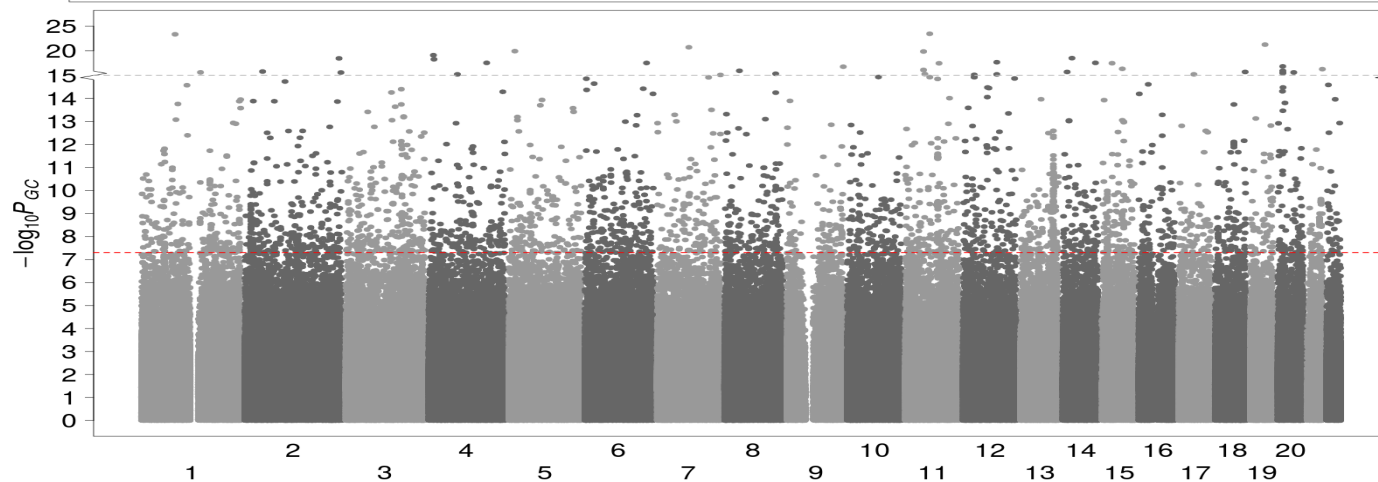
Optimizations were applied for large-scale data



**Linear mixed model**



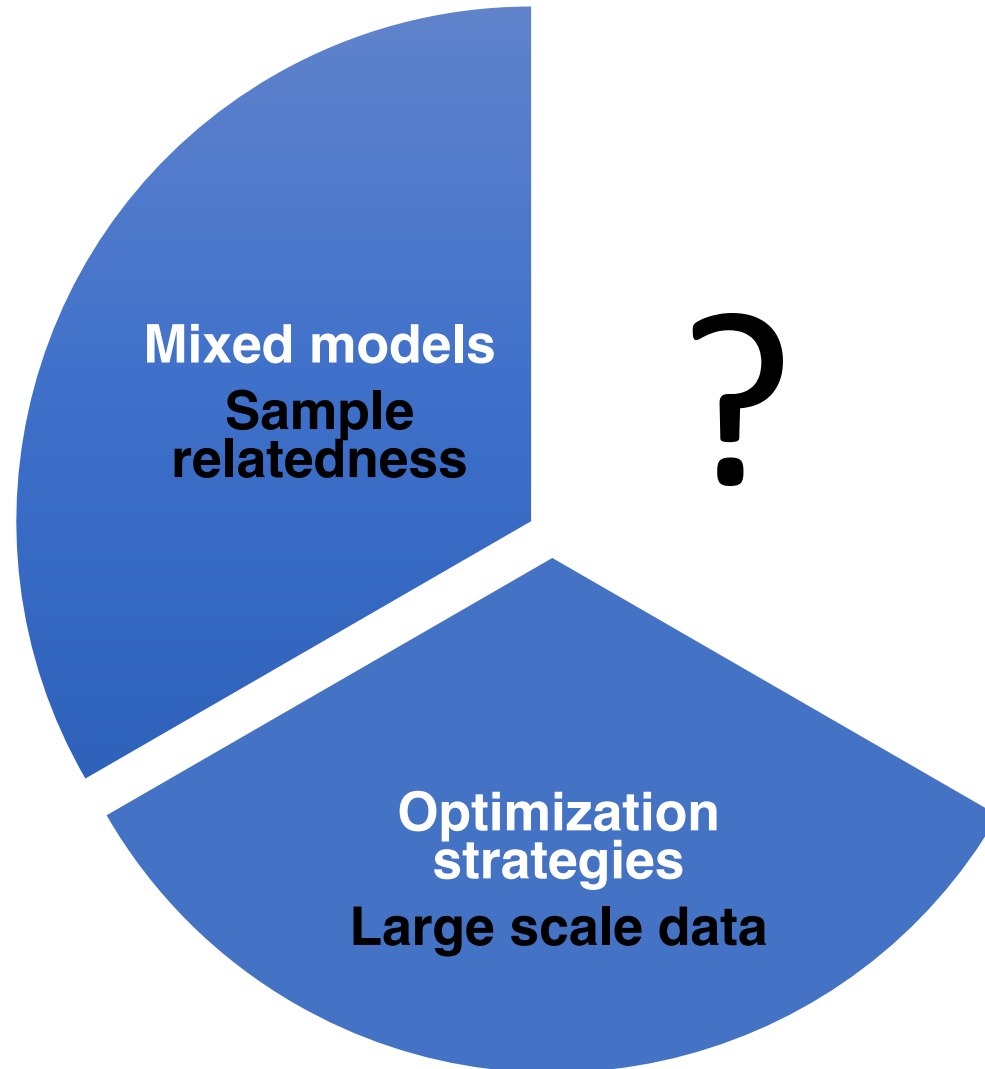
**Logistic mixed model**



**Colorectal cancer**

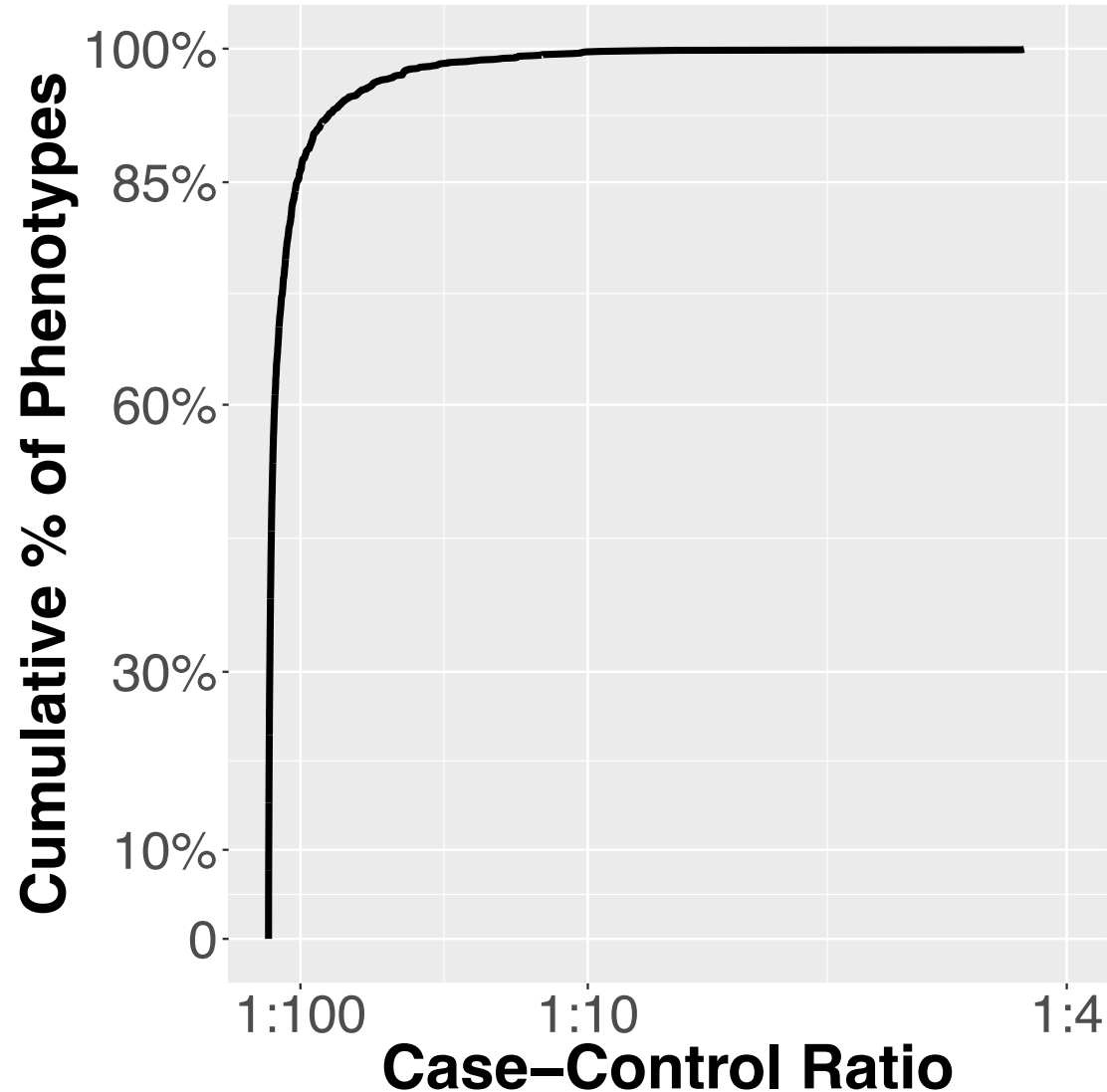
- **4,562 Cases**
- **382,756 Controls**
- **Case: Control = 1:84**

Optimizations were applied for large-scale data



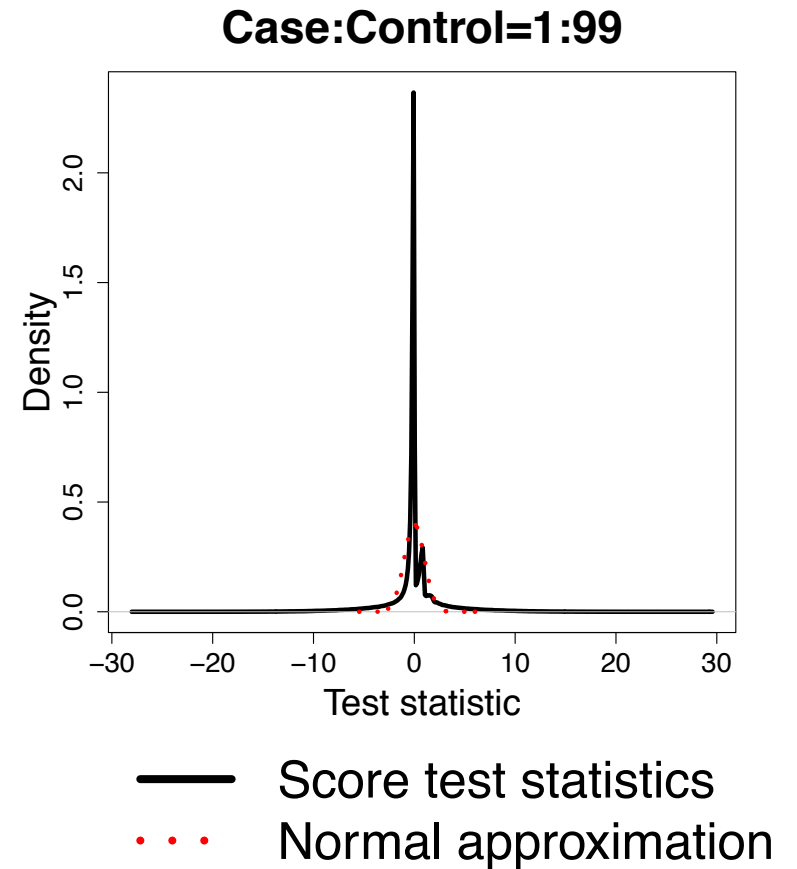
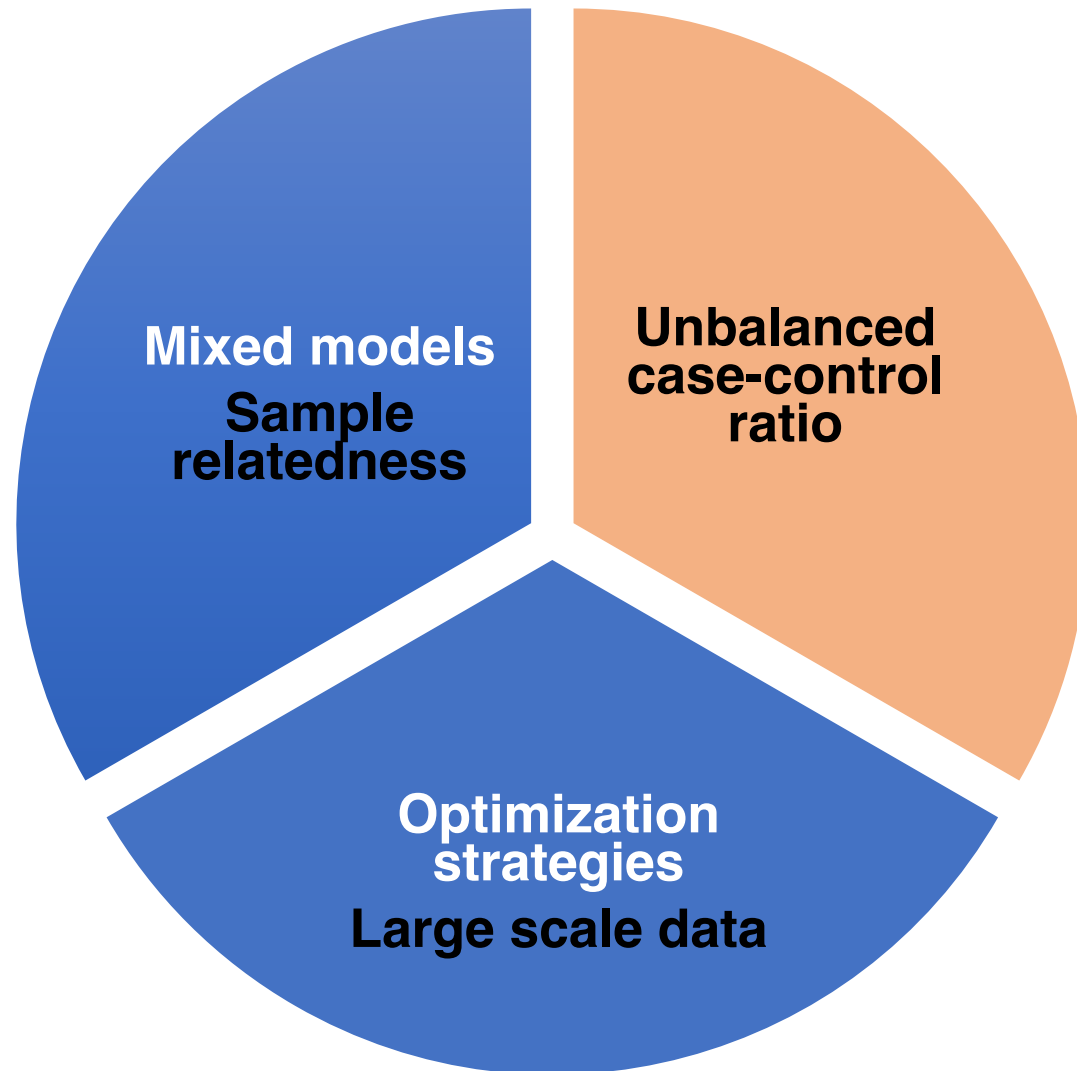


# Unbalanced case-control ratios are commonly observed for binary phenotypes in biobanks

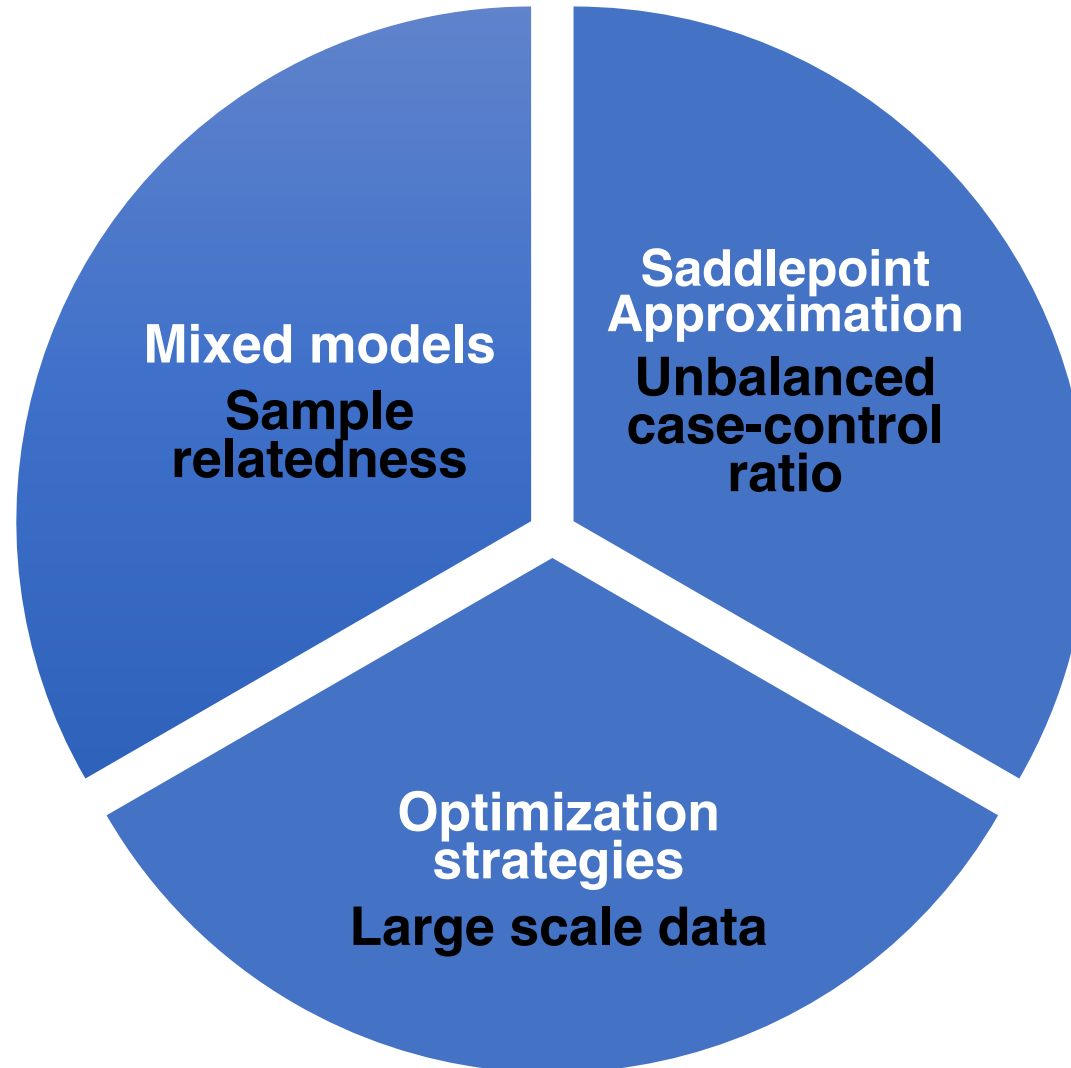


1,663 Binary Phenotypes in the UK Biobank

Test statistics do not converge to Normal distribution,  
leading to inflated type I error rates



# Saddlepoint approximation (SPA) is used to account for unbalanced case-control ratio



**SPA** uses the **entire moment generating function** -> **more accurate p-values**

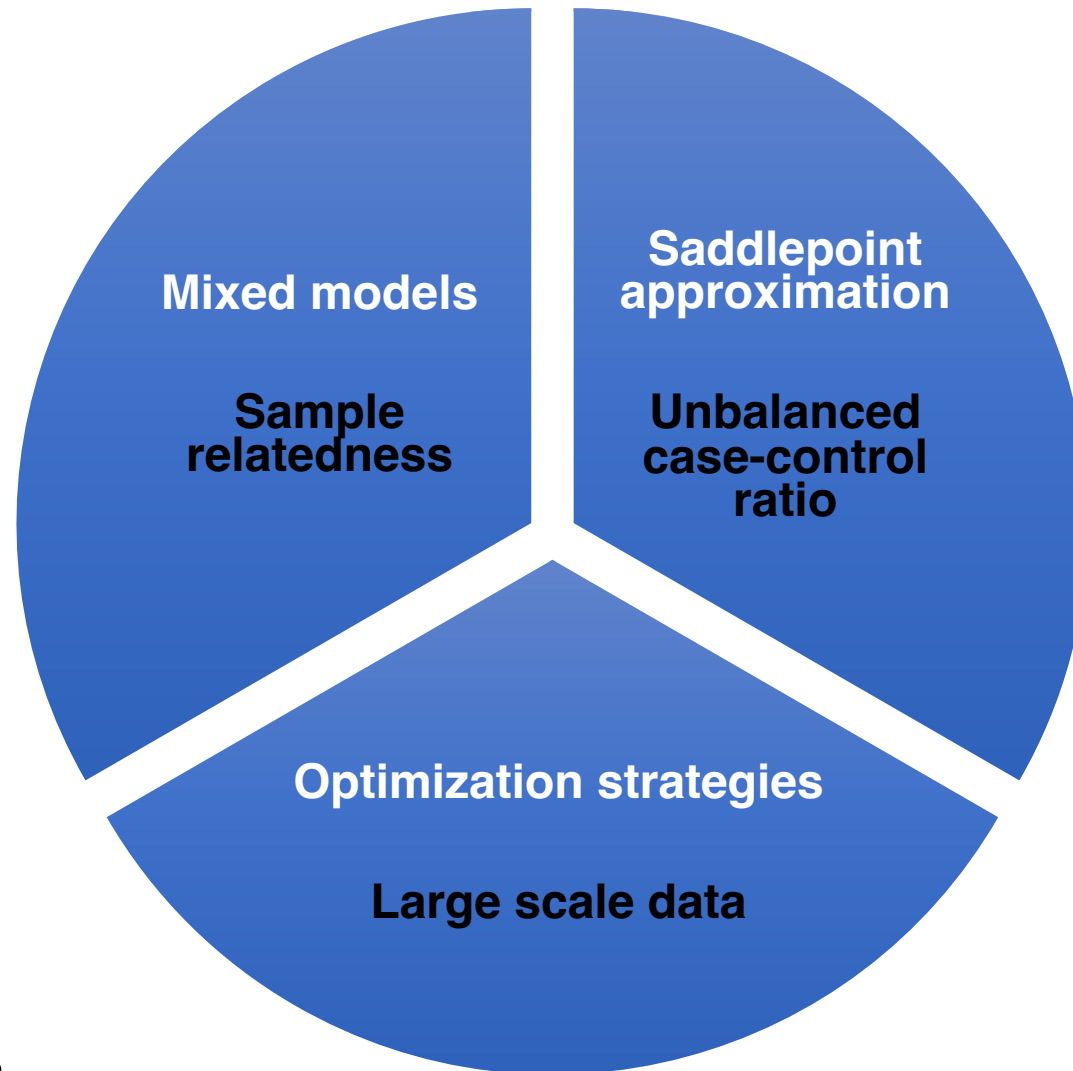
vs.

**Normal distribution** only uses the first two moments (**mean and variance**)

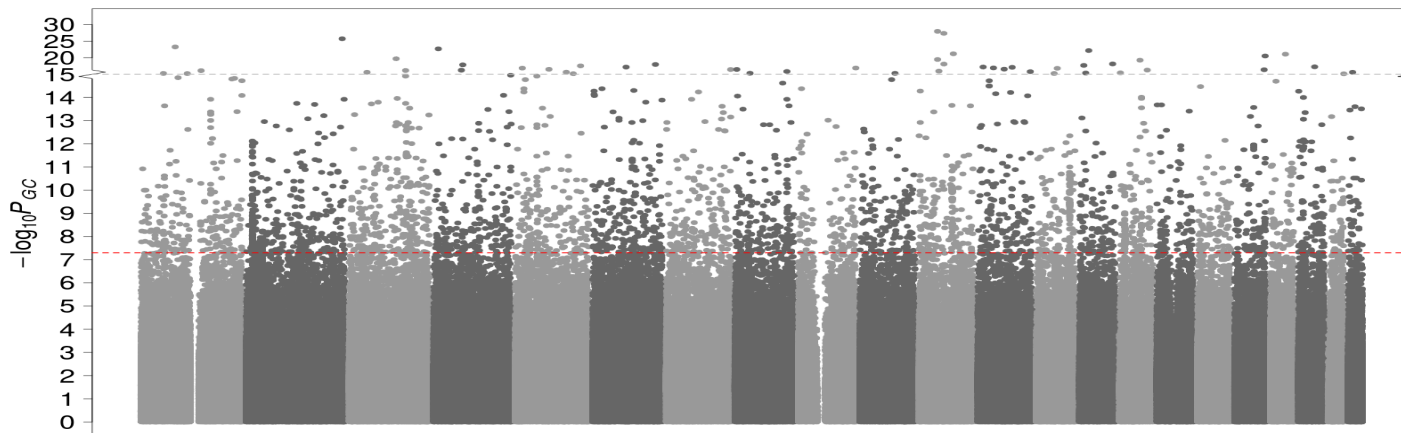
Daniels, 1954  
Dey et al., 2017

# SAIGE

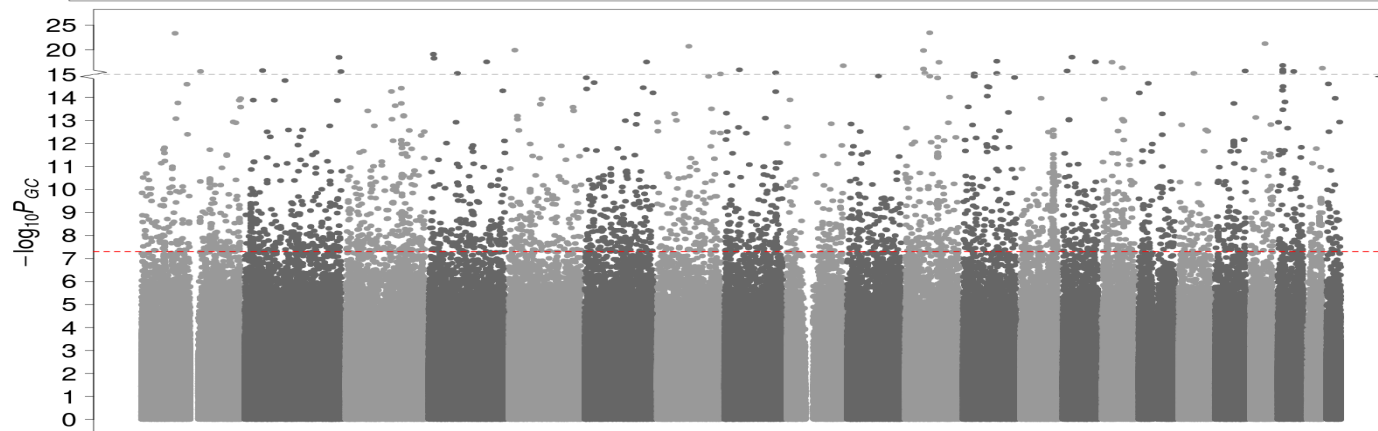
(Scalable and Accurate Implementation of GEneralized mixed model)  
was developed to conduct GWAS in large-scale biobanks



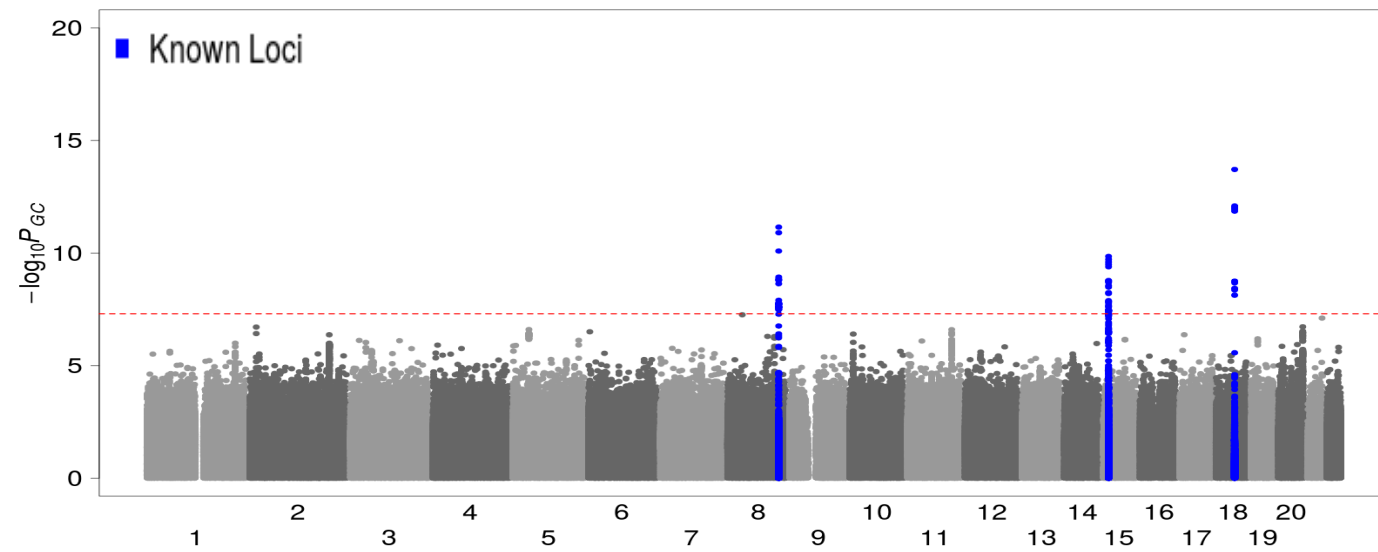
**Linear mixed model**



**Logistic mixed model**



**Logistic mixed model +SPA (SAIGE)**



**Colorectal cancer**

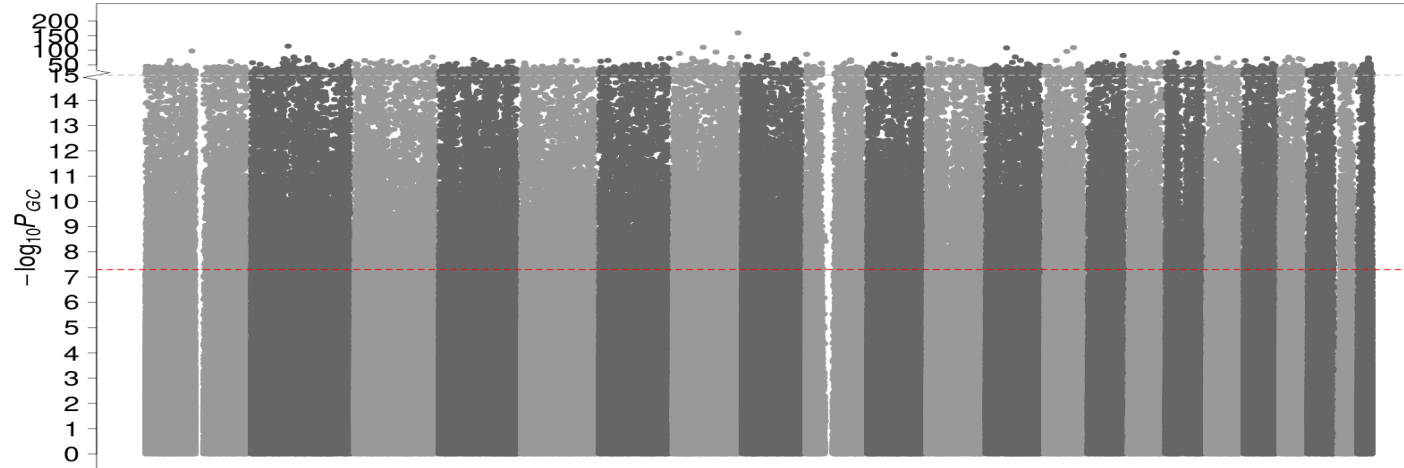
- **4,562 Cases**
- **382,756 Controls**
- **Case: Control = 1:84**

■ Known Loci

## **Thyroid cancer**

- **358 Cases**
- **407,399 Controls**
- **Case: Control = 1:1138**

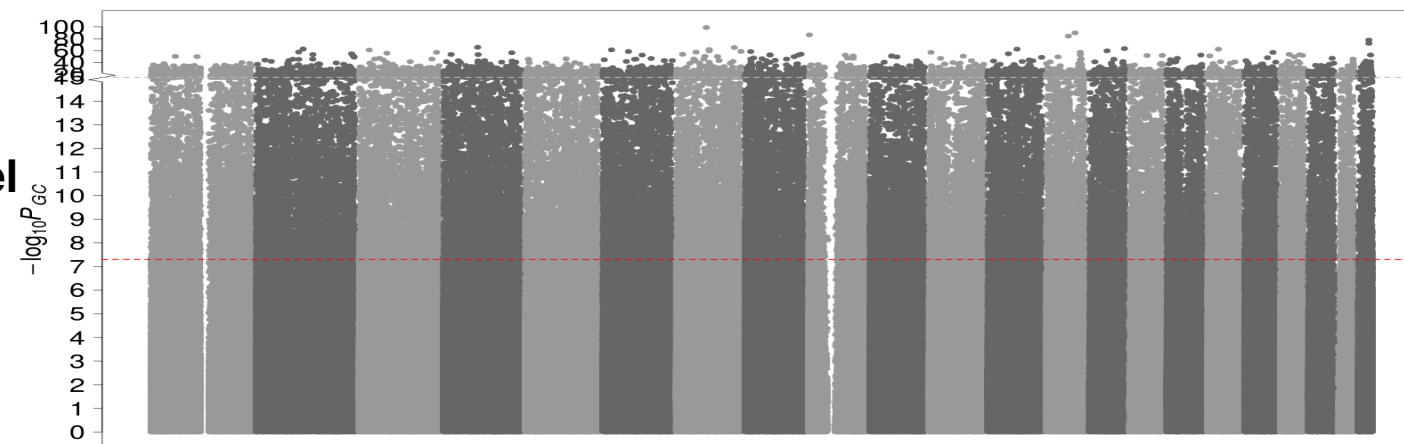
**Linear mixed model**



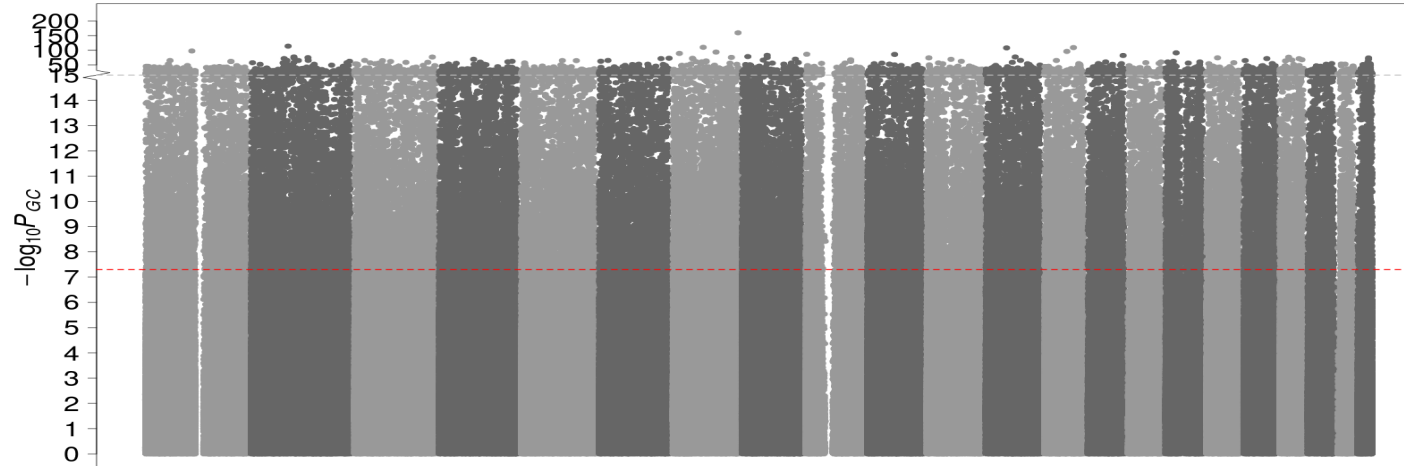
**Thyroid cancer**

- **358 Cases**
- **407,399 Controls**
- **Case: Control = 1:1138**

**Logistic mixed model**



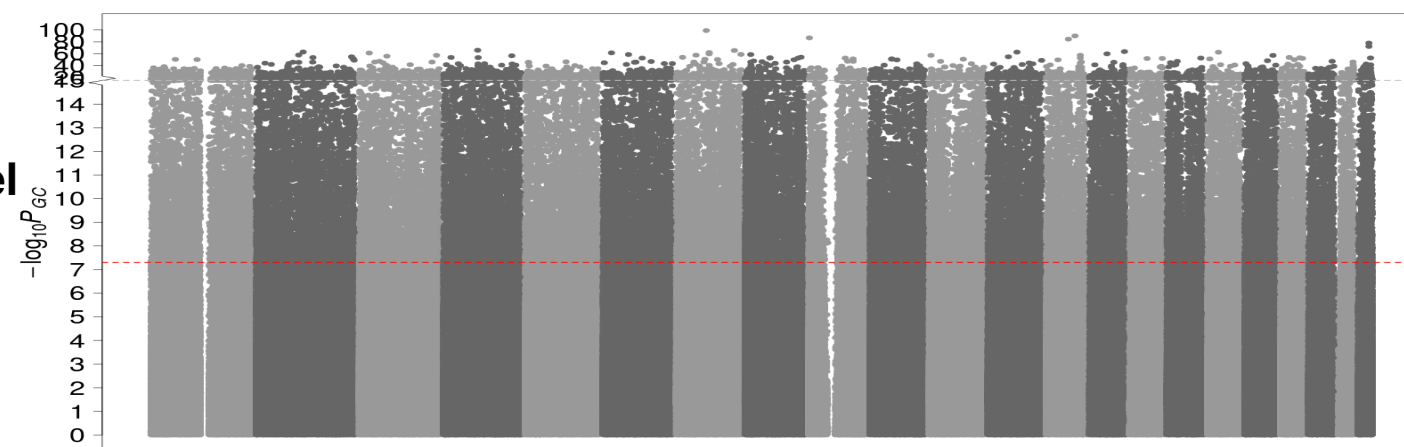
**Linear mixed model**



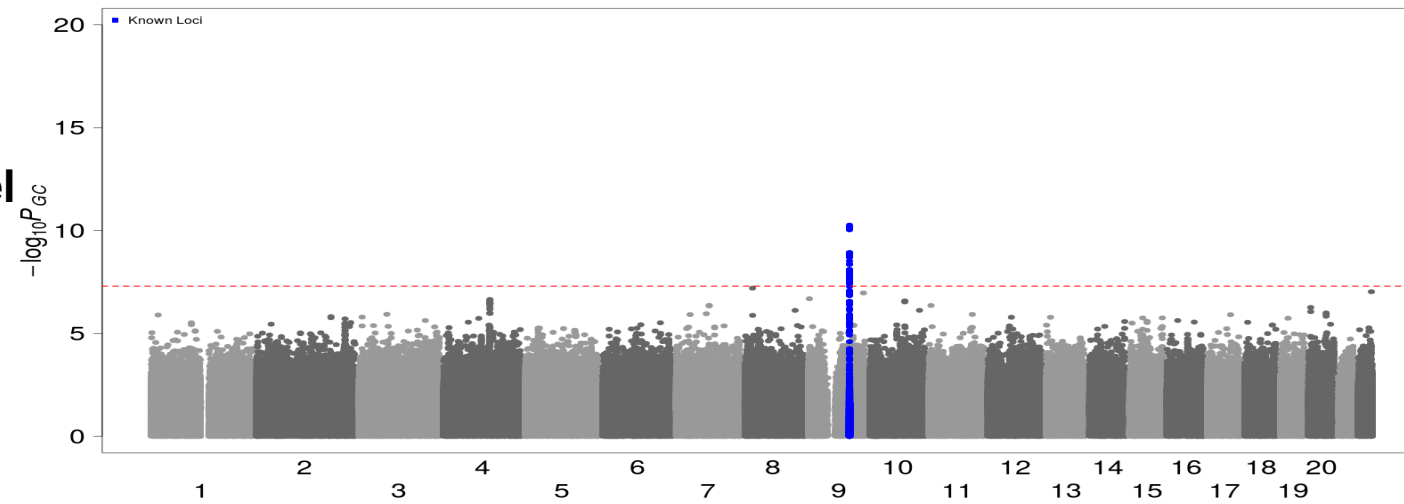
**Thyroid cancer**

- **358 Cases**
- **407,399 Controls**
- **Case: Control = 1:1138**

**Logistic mixed model**

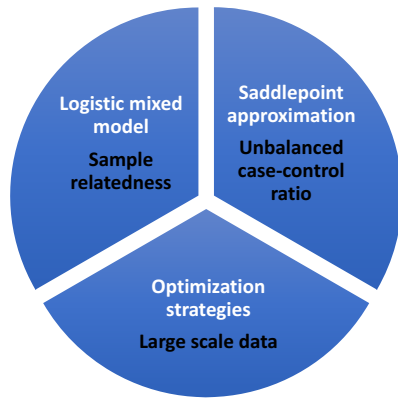


**Logistic mixed model +SPA**



■ Known Loci





# SAIGE

**Step 1: Fit the logistic mixed model under the null  $H_0: \beta = 0$**

$$\text{logit}(\pi_i) = X_i \alpha + b_i$$
$$b \sim \text{Normal}(0, \tau \psi)$$

$\hat{\alpha}, \hat{b}, \hat{\tau}$

**Step 2: Perform association test for each genetic marker**  
*Apply SPA to score tests*

*Association Results (p-values...)*

**FASTA (Two-step)**

Chen and Abecasis, 2007

# Pan-UKBB: run SAIGE for 7,228 phenotypes, across 6 continental ancestry groups, for a total of 16,131 GWAS



## Pan-UK Biobank

Pan-ancestry genetic analysis of the UK Biobank

The UK Biobank is a collection of a half million individuals with paired genetic and phenotype information that has been enormously valuable in studies of genetic etiology for common diseases and traits. However, most genome-wide analyses of this dataset use only the European ancestry individuals. Analyzing a more inclusive and diverse dataset increases power and improves the potential for discovery. Here, we present a multi-ancestry analysis of 7,228 phenotypes, across 6 continental ancestry groups, for a total of 16,131 genome-wide association studies. We release these summary statistics freely to the community ahead of publication.

Konrad Karczewski  
Alicia Martin  
Hilary Finucane  
Benjamin Neale  
Mark Daly  
Hail Team

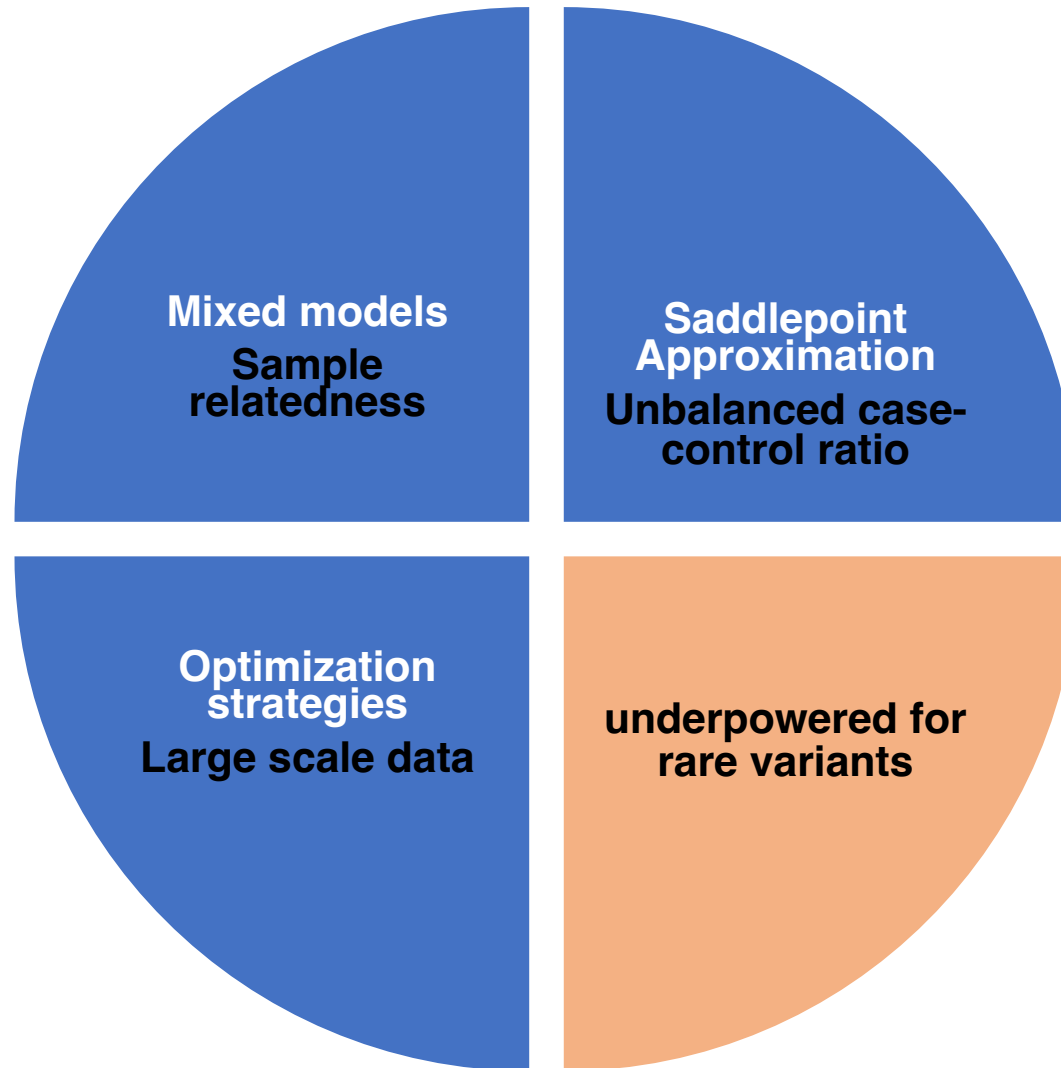
# Sequencing data are being generated by biobanks allow for studying rare variant associations for complex diseases

The screenshot shows the UK Biobank website header with navigation links for 'Researcher log in' and 'Participant log in'. Below the header is a dark grey banner with the text 'Final data release from the world's largest whole exome sequencing project'. A secondary navigation bar includes links for 'About us', 'Governance', 'Our impact', 'News', 'Winter Scientific Conference 2022', and 'COVID-19 hub'. The main content area features a news article with the same title and a date of 'July 26<sup>th</sup> 2022'. The article includes a large graphic with the text 'World's LARGEST Whole Exome Sequencing project COMPLETED!' and 'Sequenced data on 470,000 UK Biobank participants now available'. The graphic also features the UK Biobank logo and an illustration of a scientist in a white coat interacting with a server rack. At the bottom of the graphic, logos for partner companies are displayed: REGENERON, abbvie, Alnylam, AstraZeneca, Biogen, Bristol Myers Squibb, Pfizer, and Takeda.

## Why studying rare variations?

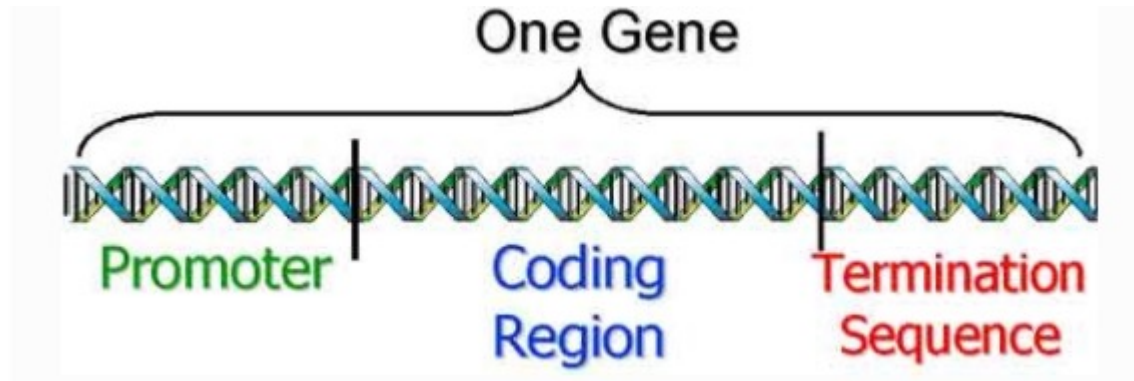
- Unexplained heritability
- Precision medicine
- Rare coding variants
  - function
  - therapeutic targets

# Single-variant association tests are underpowered for rare variants



# Solution: Test the joint effects of rare variants

- Grouping rare variants into functional units, i.e. genes, epigenetic features..
- Set-based tests
  - Gene-based tests, group-based tests



**Logistic regression:**  $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = X\alpha + G_1\beta_1 + G_1\beta_2 + \dots + G_q\beta_q$

$\pi$ : *probability of having disease given X and G*

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

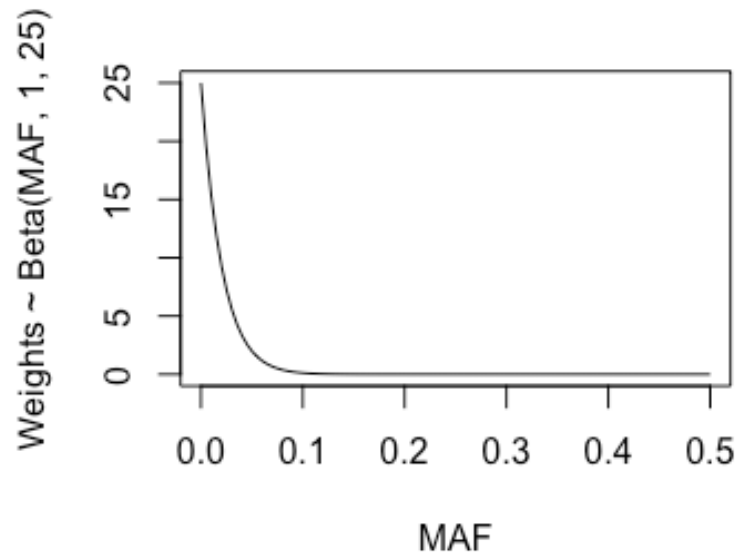
# Incorporating weight for each variant

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = X\alpha + G_1 \mathbf{w}_1 \beta_1 + G_2 \mathbf{w}_2 \beta_2 + \dots + G_q \mathbf{w}_q \beta_q$$

$\pi$ : probability of having disease given  $X$  and  $G$

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

$$\mathbf{w} \sim \text{Beta}(\text{MAF}, 1, 25)$$

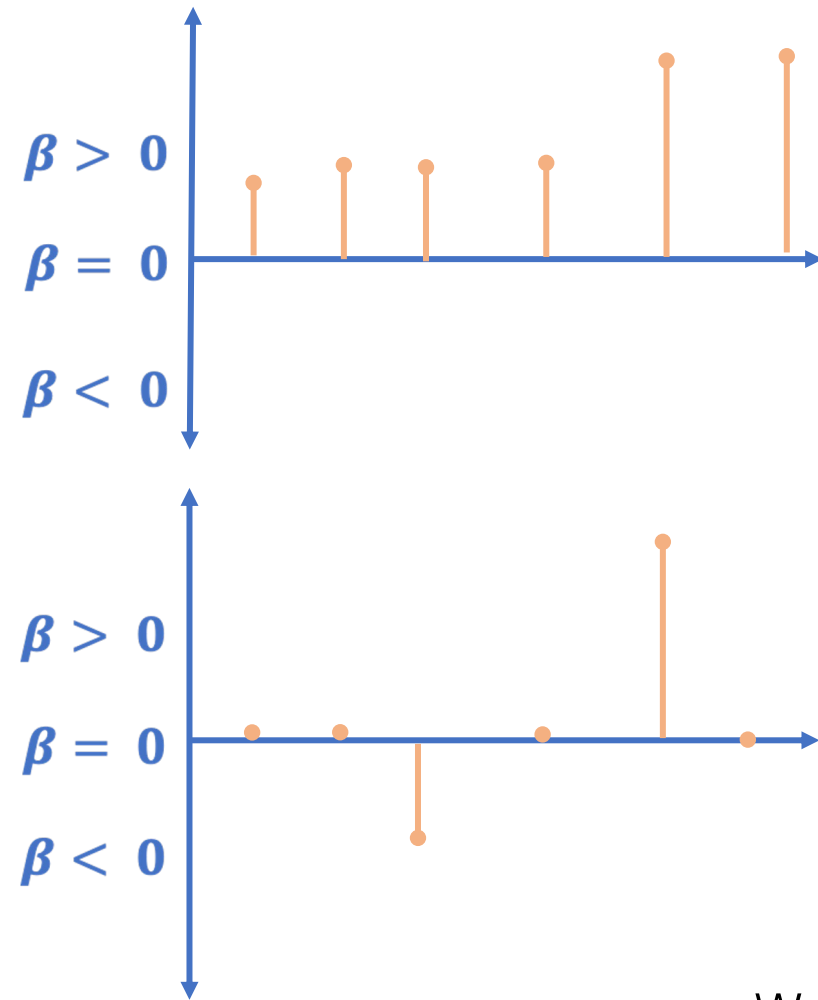


# Set-based association tests: grouping rare variants to test (by genes, epigenetic features...)

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

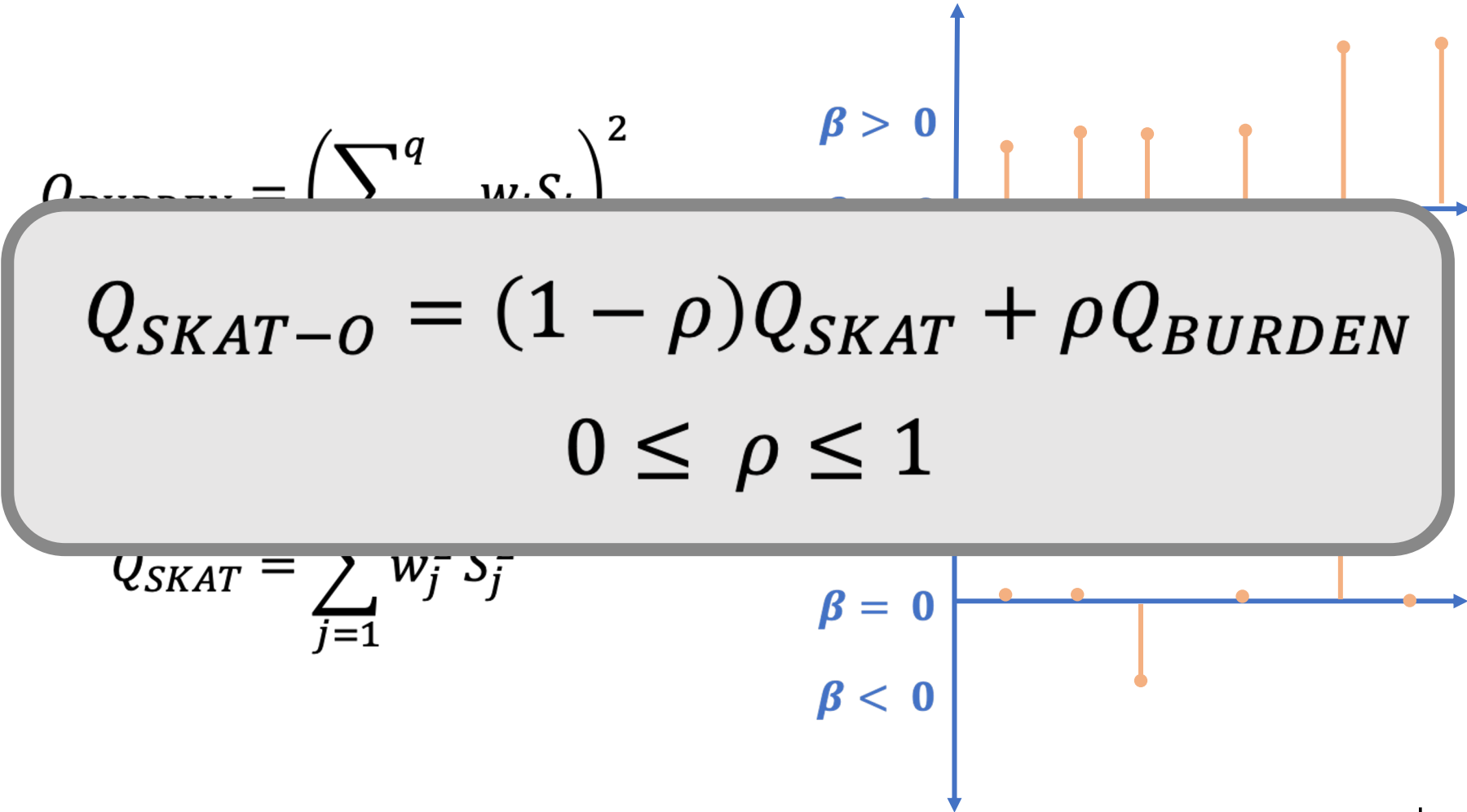
$$Q_{SKAT} = \sum_{j=1}^q w_j^2 S_j^2$$

$S_j = \sum_{i=1}^n g_{ij} (y_i - \hat{\pi}_i)$  is the score statistic for the variant  $j$   
Under the null  $H_0: \beta_j = 0$





# SKAT-O is more powerful than Burden and SKAT



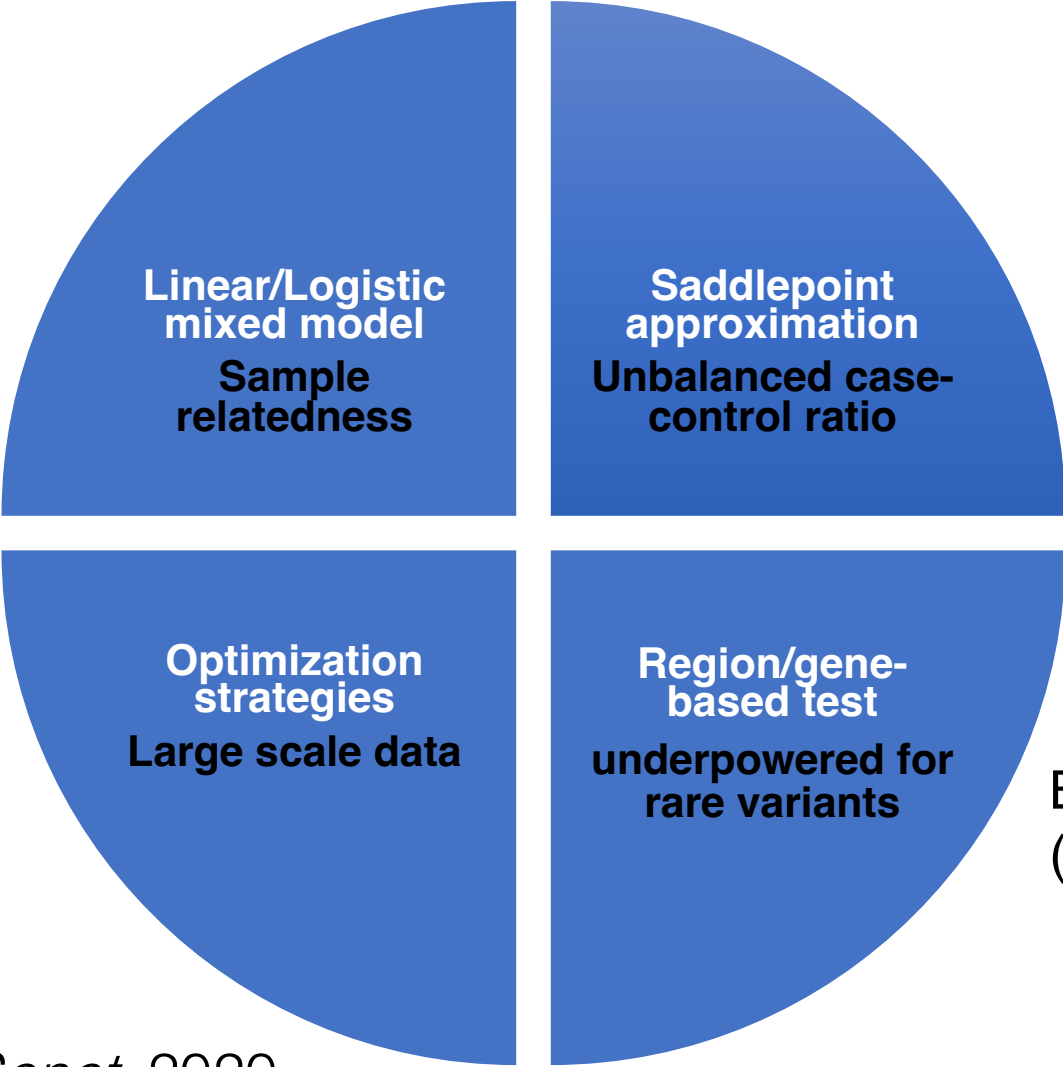
**Table 2. Summary of Statistical Methods for Rare-Variant Association Testing**

	<b>Description</b>	<b>Methods</b>	<b>Advantage</b>	<b>Disadvantage</b>	<b>Software Packages<sup>a</sup></b>
Burden tests	collapse rare variants into genetic scores	ARIEL test, <sup>50</sup> CAST, <sup>51</sup> CMC method, <sup>52</sup> MZ test, <sup>53</sup> WSS <sup>54</sup>	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum, <sup>55</sup> Step-up, <sup>56</sup> EREC test, <sup>57</sup> VT, <sup>58</sup> KBAC method, <sup>59</sup> RBT <sup>60</sup>	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT, <sup>61</sup> SSU test, <sup>62</sup> C-alpha test <sup>63</sup>	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O, <sup>64</sup> Fisher method, <sup>65</sup> MiST <sup>66</sup>	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test <sup>67</sup>	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

Abbreviations are as follows: ARIEL, accumulation of rare variants integrated and extended locus-specific; aSum, data-adaptive sum test; CAST, cohort allelic sums test; CMC, combined multivariate and collapsing; EC, exponential combination; EPACTS, efficient and parallelizable association container toolbox; EREC, estimated regression coefficient; GRANVIL, gene- or region-based analysis of variants of intermediate and low frequency; KBAC, kernel-based adaptive cluster; MiST, mixed-effects score test for continuous outcomes; MZ, Morris and Zeggini; RBT, replication-based test; Rvtests, rare-variant tests; SKAT, sequence kernel association test; SSU, sum of squared score; VAT, variant association tools; VT, variable threshold; and WSS, weighted-sum statistic.

<sup>a</sup>More information is given in [Table 3](#).

SAIGE-GENE was the first method for rare variant associations tests of binary phenotypes in large-scale data



Burden, SKAT, and SKAT-O (Lee et al., 2012)

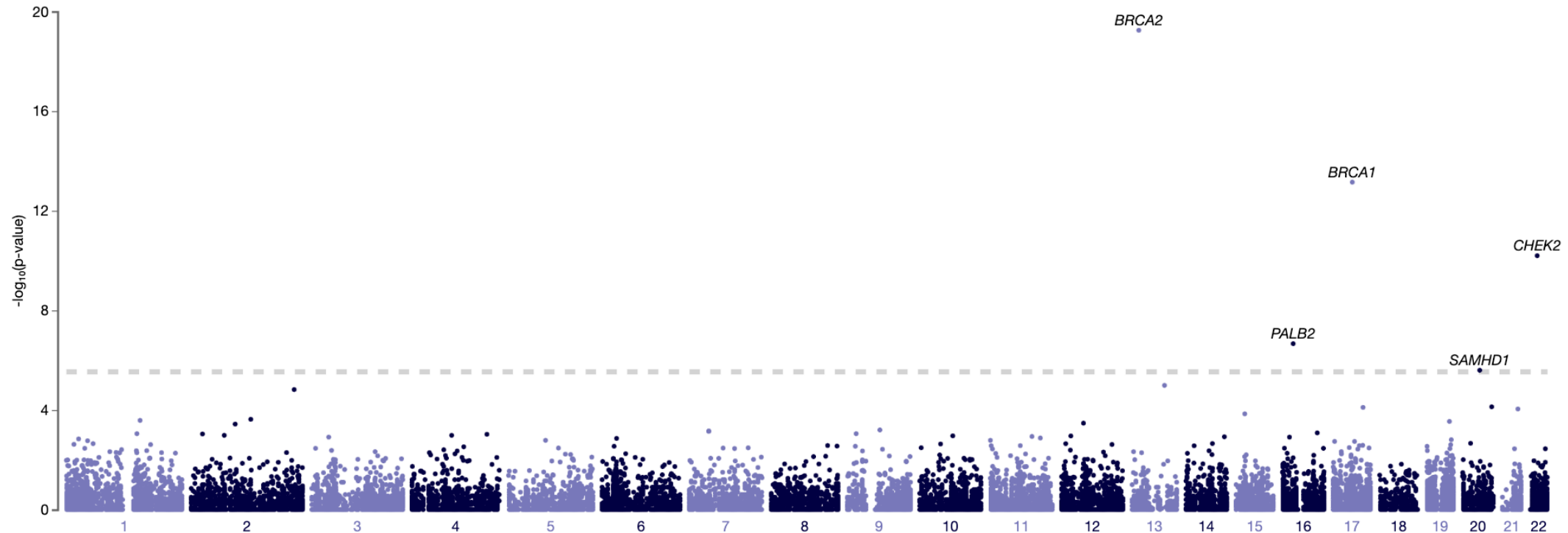
# Set-based tests help identify genetic associations that are missed by single-variant tests

## 20001\_1002: breast cancer

Category: **self-report**

Sample Size: **3905:87740**

Showing the top 1000 genes



[https://ukb-200kexome.leelabsg.org/pheno/20001\\_1002](https://ukb-200kexome.leelabsg.org/pheno/20001_1002)

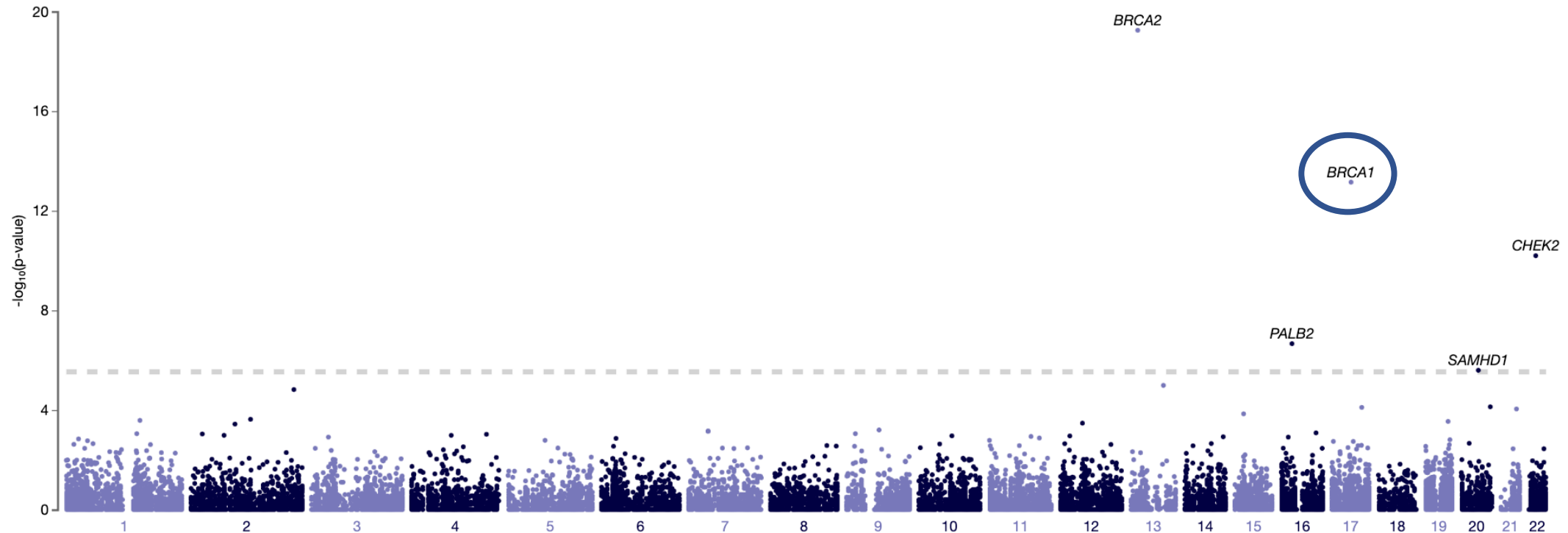
# Set-based tests help identify genetic associations that are missed by single-variant tests

## 20001\_1002: breast cancer

Category: **self-report**

Sample Size: **3905:87740**

Showing the top 1000 genes



[https://ukb-200kexome.leelabsg.org/pheno/20001\\_1002](https://ukb-200kexome.leelabsg.org/pheno/20001_1002)

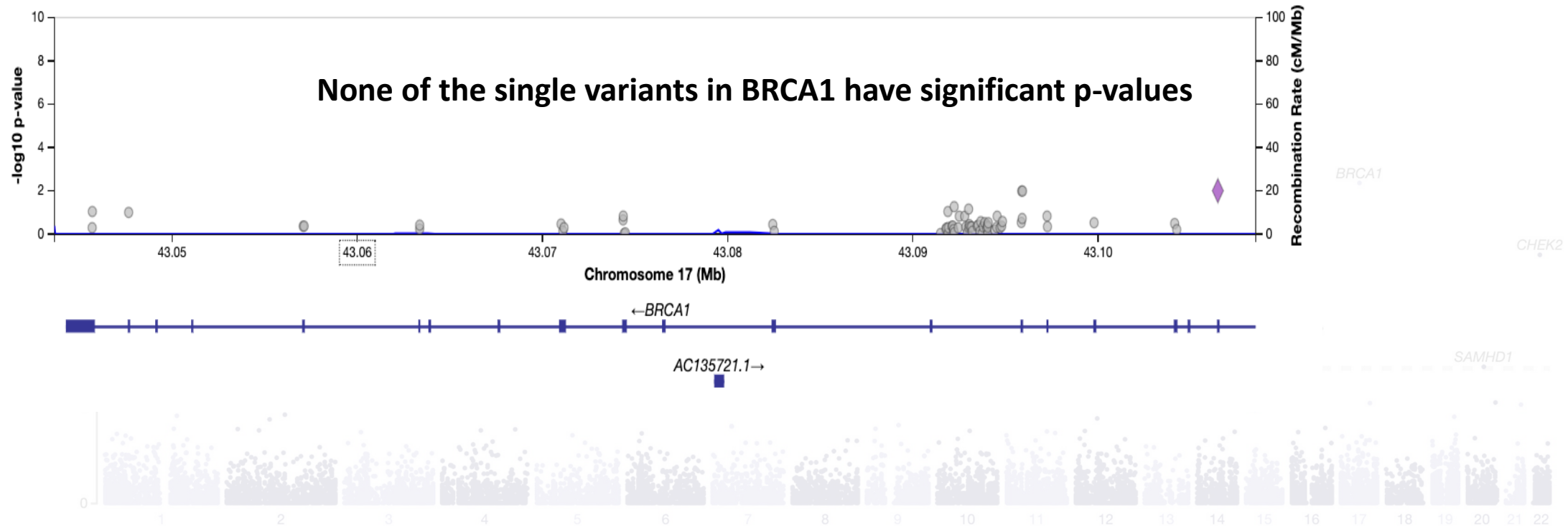
# Set-based tests help identify genetic associations that are missed by single-variant tests

20001\_1002: breast cancer

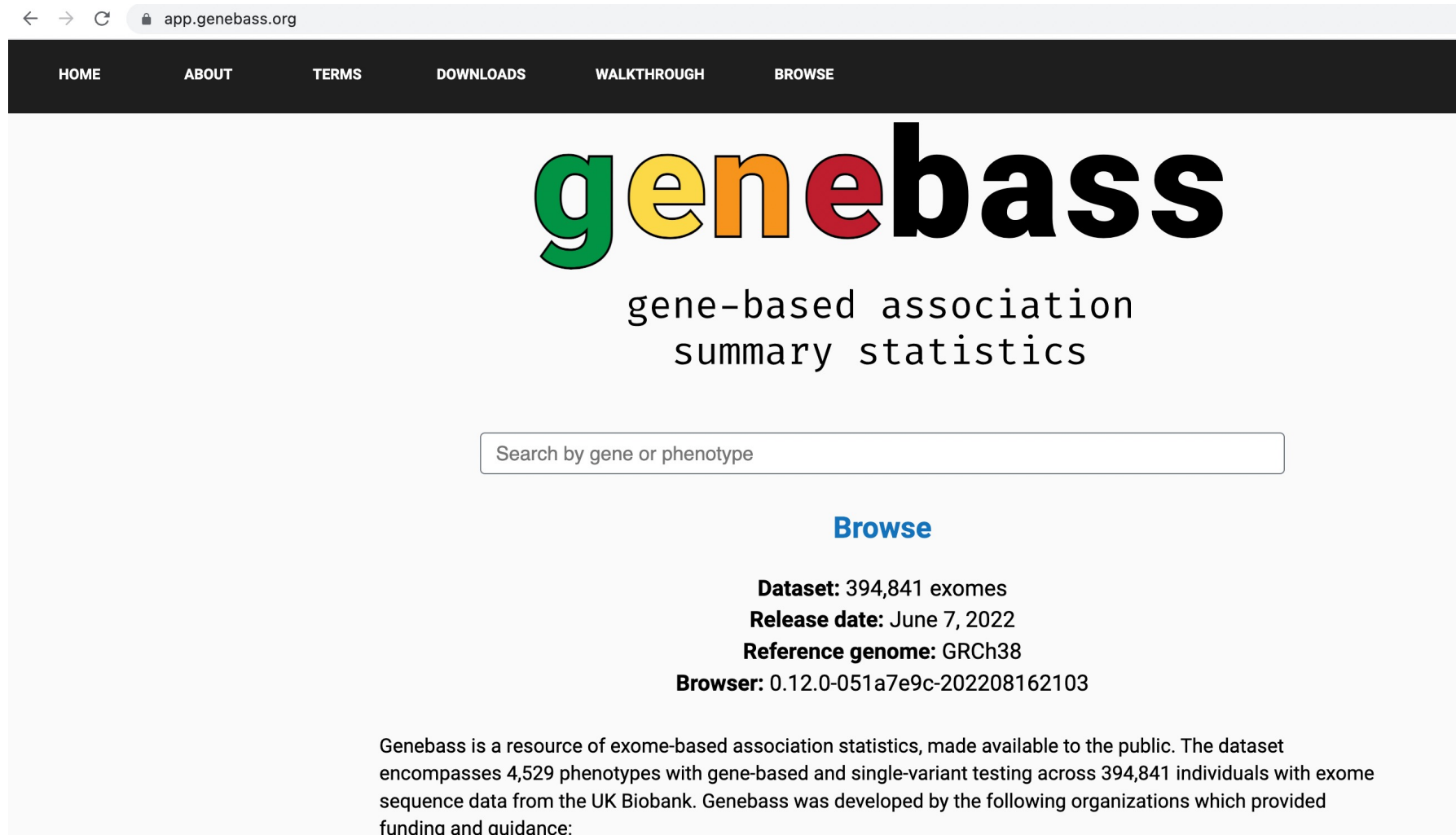
Category: self-report

Sample Size: 3905:87740

Showing the top 1000 genes



SAIGE-GENE was used to analyze 4,529 phenotypes on 394,841 exomes in UKBB



The screenshot shows the Genebass website interface. At the top, there is a navigation bar with links for HOME, ABOUT, TERMS, DOWNLOADS, WALKTHROUGH, and BROWSE. Below the navigation bar is the Genebass logo, which consists of the word "genebass" in a stylized font where "gene" is multi-colored (green, yellow, orange, red) and "bass" is black. Underneath the logo is the text "gene-based association summary statistics". A search bar is present with the placeholder text "Search by gene or phenotype". Below the search bar is a "Browse" button. Further down, there are several key statistics: "Dataset: 394,841 exomes", "Release date: June 7, 2022", "Reference genome: GRCh38", and "Browser: 0.12.0-051a7e9c-202208162103". At the bottom, there is a paragraph of text describing the resource and its development.

app.genebass.org

HOME ABOUT TERMS DOWNLOADS WALKTHROUGH BROWSE

# genebass

gene-based association  
summary statistics

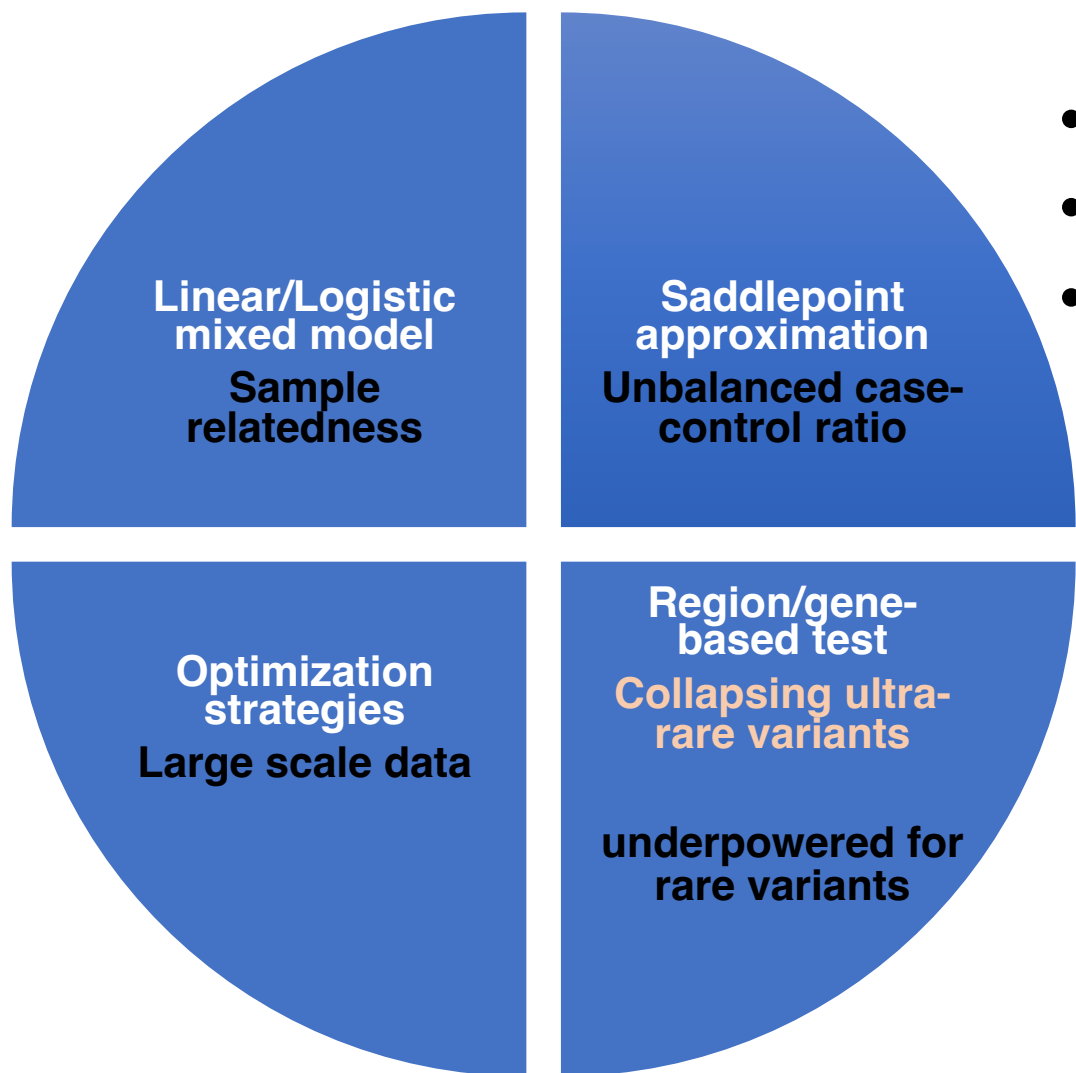
Search by gene or phenotype

[Browse](#)

**Dataset:** 394,841 exomes  
**Release date:** June 7, 2022  
**Reference genome:** GRCh38  
**Browser:** 0.12.0-051a7e9c-202208162103

Genebass is a resource of exome-based association statistics, made available to the public. The dataset encompasses 4,529 phenotypes with gene-based and single-variant testing across 394,841 individuals with exome sequence data from the UK Biobank. Genebass was developed by the following organizations which provided funding and guidance:

# SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests



- **Improved computational efficiency**
- **Improved type I error**
- **Improved power**

- multiple functional annotations
  - LoF
  - LoF+nonsynonymous
  - LoF+nonsynonymous+synonymous
- multiple max MAF cutoffs
  - 0.01%
  - 0.1%
  - 1%

**Combined p-values using Cauchy combination method**



# Rare variant associations can aggregate in different annotation and MAF groups

## BRCA1 + 20001\_1002 (breast cancer)

[Go to gene BRCA1](#)

[Go to pheno 20001\\_1002](#)

pval: **6.78e-14**

Chrom : Start - End: 17 : **43,044,294 - 43,125,483**

Sample Size: **3905:87740**

Group	P-value	MAC(case:control)	#Rare Variants
Lof_0.0001	9.7e-15	25:72	51
Lof_0.001	6.7e-14	29:121	53
Lof_0.01	6.7e-14	29:121	53
MissenseLof_0.0001	8.9e-3	57:1002	375
MissenseLof_0.001	0.011	125:2276	403
MissenseLof_0.01	0.033	255:4971	407
MissenseLofSynonym...	0.011	72:1349	503
MissenseLofSynonym...	0.014	145:3037	539
MissenseLofSynonym...	0.078	284:5952	544

[https://ukb-200kexome.leelabsg.org/assoc/BRCA1/20001\\_1002](https://ukb-200kexome.leelabsg.org/assoc/BRCA1/20001_1002)

## GCK + 250.2 (Type 2 diabetes)

[Go to gene GCK](#)

[Go to pheno 250.2](#)

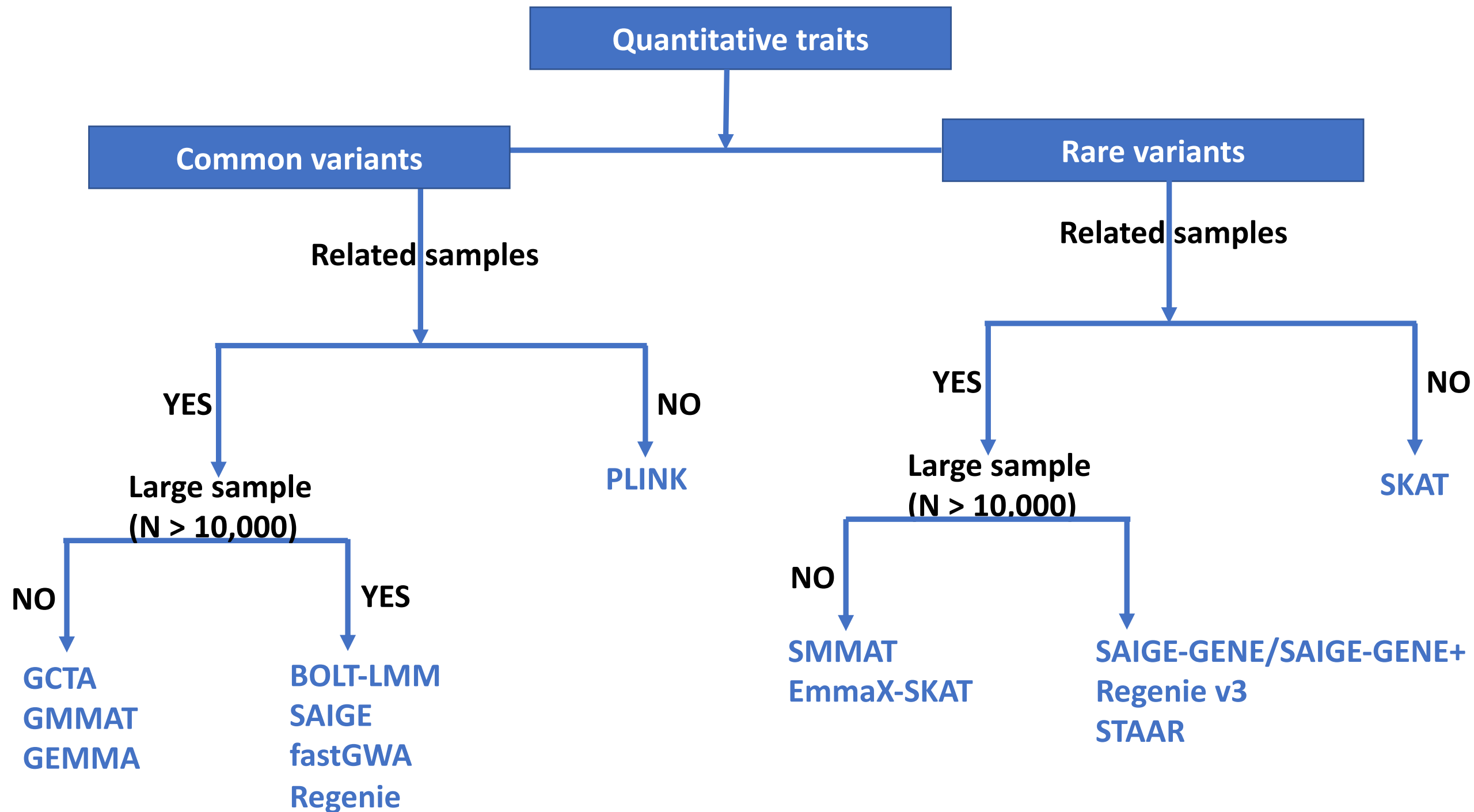
pval: **2.29e-12**

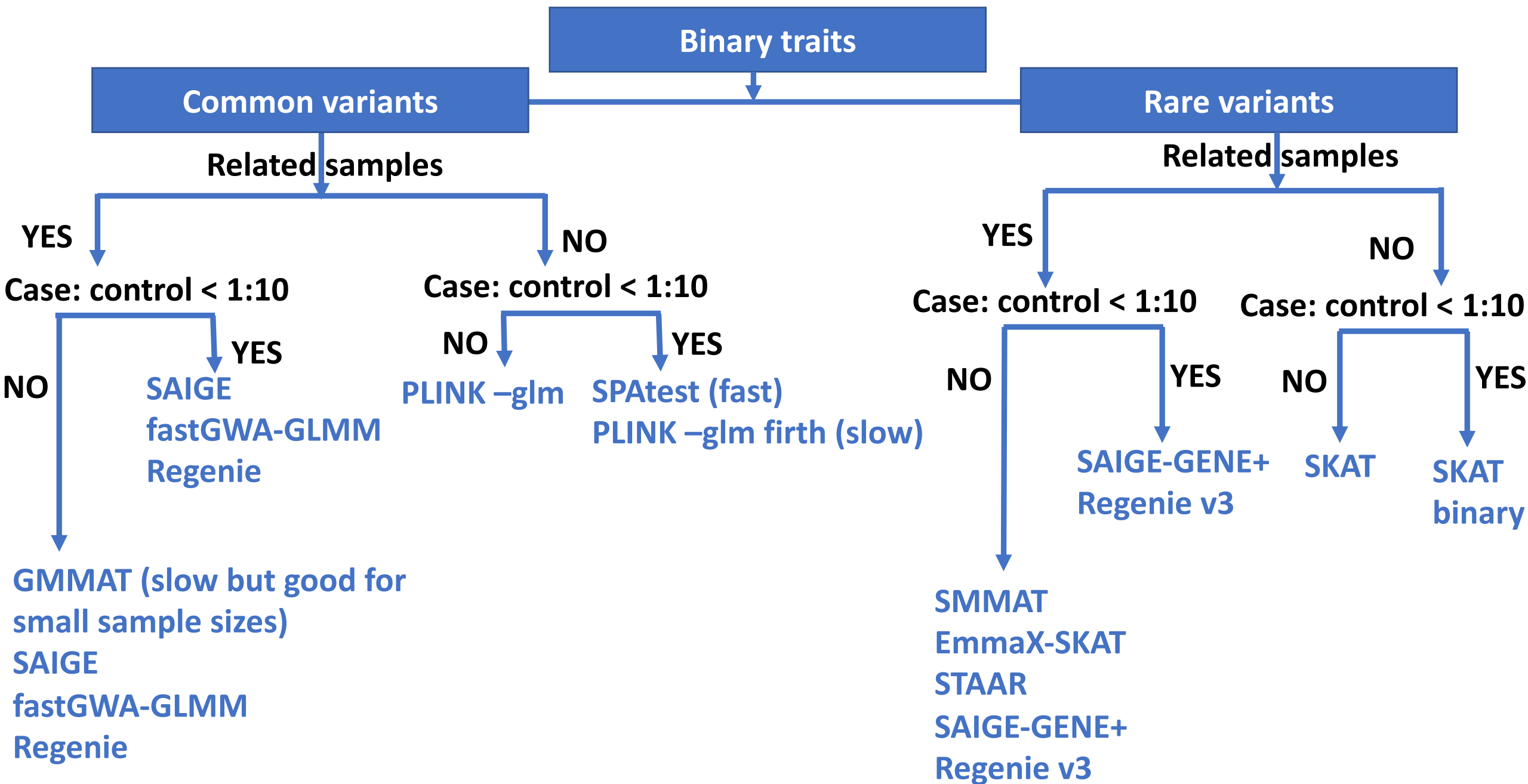
Chrom : Start - End: 7 : **44,144,270 - 44,189,423**

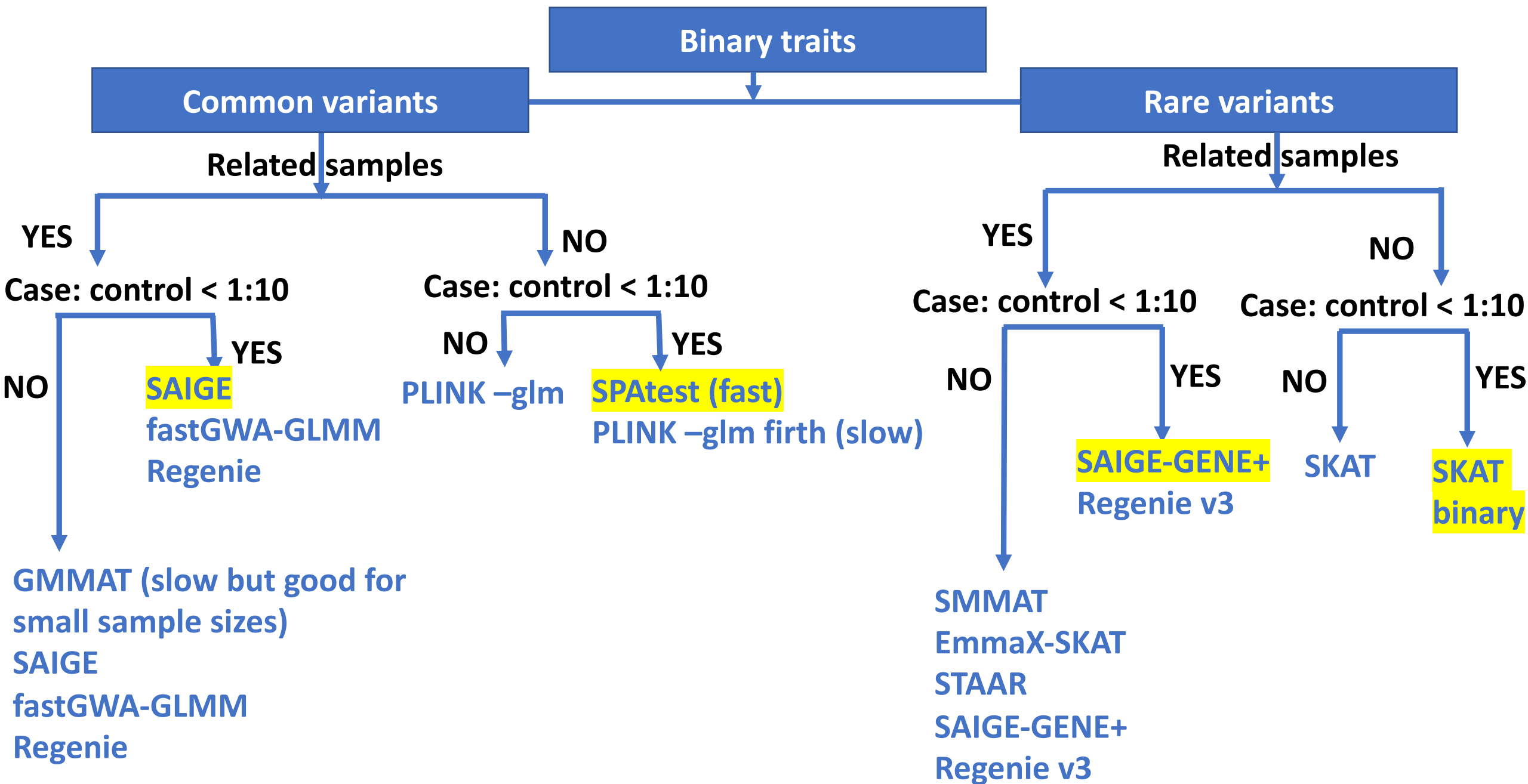
Sample Size: **7342:159103**

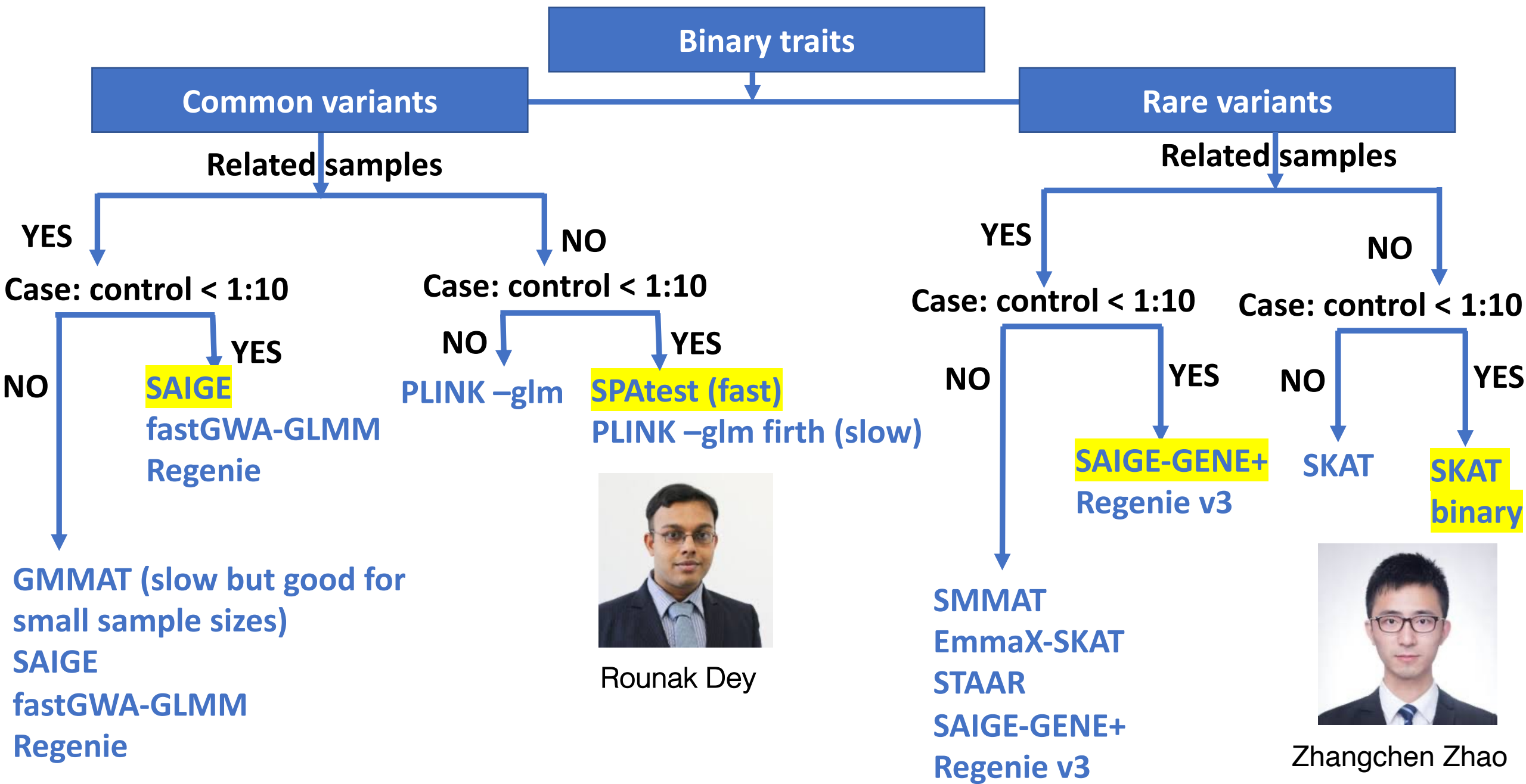
Group	P-value	MAC(case:control)	#Rare Variants
Lof_0.0001	3.2e-8	10:13	13
Lof_0.001	3.2e-8	10:13	13
Lof_0.01	3.2e-8	10:13	13
MissenseLof_0.0001	1.5e-12	47:387	116
MissenseLof_0.001	6.1e-13	51:528	119
MissenseLof_0.01	6.1e-13	51:528	119
MissenseLofSynonym...	3.2e-8	57:617	184
MissenseLofSynonym...	3.7e-9	81:1170	192
MissenseLofSynonym...	8.0e-3	210:4150	193

<https://ukb-200kexome.leelabsg.org/assoc/GCK/250.2>









# Reference

- SKAT-O: Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani et al. "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, no. 2 (2012): 224-237.
- SKAT: 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.
- SKAT binary: Zhao, Zhangchen, Wenjian Bi, Wei Zhou, Peter VandeHaar, Lars G. Fritsche, and Seunggeun Lee. "UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test." *The American Journal of Human Genetics* 106, no. 1 (2020): 3-12.
- SPAtest: Dey, Rounak, Ellen M. Schmidt, Goncalo R. Abecasis, and Seunggeun Lee. "A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS." *The American Journal of Human Genetics* 101, no. 1 (2017): 37-49.
- BOLT-LMM: Loh, Po-Ru, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjalmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts." *Nature genetics* 47, no. 3 (2015): 284.
- SAIGE: Zhou, Wei, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies." *Nature genetics* 50, no. 9 (2018): 1335-1341.
- SAIGE-GENE: 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.
- SAIGE-GENE+: Zhou, Wei\*, Wenjian Bi\*, Zhangchen Zhao\*, Kushal K. Dey, Karthik A. Jagadeesh, Konrad J. Karczewski, Mark J. Daly, Benjamin M. Neale, and Seunggeun Lee. "SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests." *Nature genetics* 54, no. 10 (2022): 1466-1469.

# Reference

- GEMMA: Zhou, Xiang, and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nature genetics* 44, no. 7 (2012): 821-824.
- GMMAT: Chen, Han, Chaolong Wang, Matthew P. Conomos, Adrienne M. Stilp, Zilin Li, Tamar Sofer, Adam A. Szpiro et al. "Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models." *The American Journal of Human Genetics* 98, no. 4 (2016): 653-666.
- Regenie: Mbatchou, Joelle, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner et al. "Computationally efficient whole-genome regression for quantitative and binary traits." *Nature genetics* 53, no. 7 (2021): 1097-110
- SMMAT: 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.
- fastGWA-GLMM: Jiang, Longda, Zhili Zheng, Hailing Fang, and Jian Yang. "A generalized linear mixed model association tool for biobank-scale data." *Nature genetics* 53, no. 11 (2021): 1616-1621.
- fastGWA: Jiang, Longda, Zhili Zheng, Ting Qi, Kathryn E. Kemper, Naomi R. Wray, Peter M. Visscher, and Jian Yang. "A resource-efficient tool for mixed model association analysis of large-scale data." *Nature genetics* 51, no. 12 (2019): 1749-1755.
- STAAR: Li, Xihao, Zilin Li, Hufeng Zhou, Sheila M. Gaynor, Yaowu Liu, Han Chen, Ryan Sun 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, no. 9 (2020): 969-983.
- PLINK: Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American journal of human genetics* 81, no. 3 (2007): 559-575.

# Different types of phenotypes require different statistical models for association tests

- Quantitative
  - eg. LDL cholesterol level, height
  - Linear regression
- Binary
  - eg. Schizophrenia, Type 2 Diabetes
  - Logistic regression
- **Ordinal/categorical**
  - **eg. On a scale of 1-10 how much do you like smoking**
  - **Proportional odds logistic regression, Multinomial regression**
- **Time-to-event (TTE)**
  - **eg. Age at skin cancer onset, Time of death after diagnosis of lung cancer**
  - **Survival analysis model**



# Mixed model method for other phenotype types

- Ordinal phenotypes
  - Common variants:
    - **POLMM: Proportional Odds Logistic Mixed Model**
    - Bi, Wenjian, Wei Zhou, Rounak Dey, Bhramar Mukherjee, Joshua N. Sampson, and Seunggeun Lee. "Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes." *The American Journal of Human Genetics* 108, no. 5 (2021): 825-839.
  - Rare variants:
    - **POLMM-GENE** (under development)
- Time-to-event phenotypes
  - Common variants:
    - **GATE: Genetic Analysis of Time-to-Event phenotypes**
    - R library: <https://github.com/weizhou0/GATE>
    - Common variants: Dey, Rounak\*, Wei Zhou\*, Tuomo Kiiskinen, Aki Havulinna, Amanda Elliott, Juha Karjalainen, Mitja Kurki et al. "Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks." *Nature Communications* 13, no. 1 (2022): 5437.

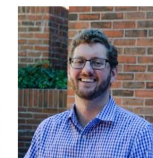
Rounak Dey



Mark Daly



Benjamin Neale



Xihong Lin



# Analyzing X Chromosome

JOURNAL ARTICLE

## A systematic review of analytical methods used in genetic association analysis of the X-chromosome

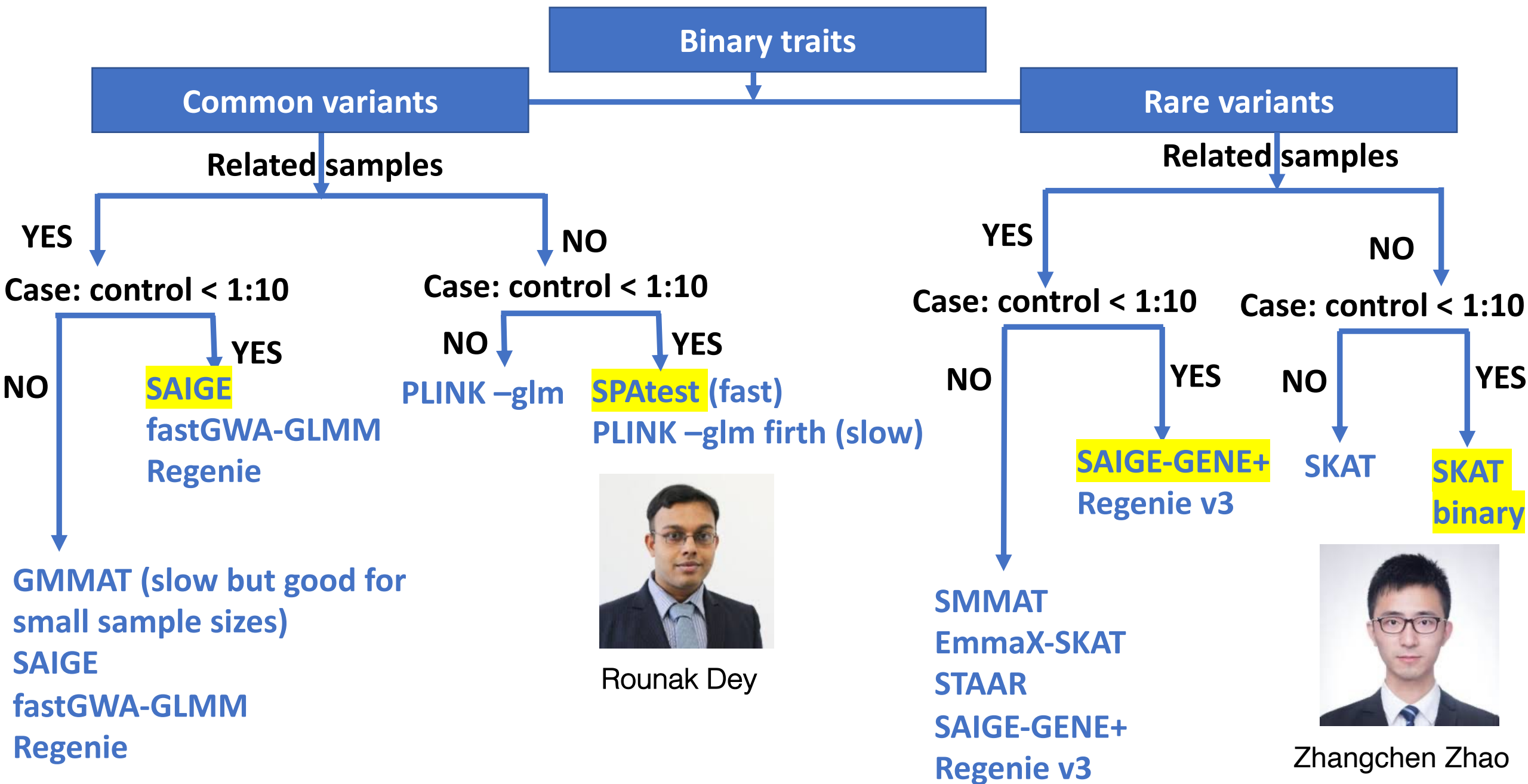
[Nick Keur](#), [Isis Ricaño-Ponce](#), [Vinod Kumar](#), [Vasiliki Matzaraki](#) 

- Complex diseases/traits present sexual dimorphic prevalence which points toward a potential contribution of the X-chromosome.
- Quality control and imputation of X-chromosome genetic data require special attention to account for its unique properties.
- Selection of statistical tests to identify associations with X-chromosome loci depends on the underlying X-chromosomal inactivation (XCI) model, HWE, sex-specific alleles and confounding variables.

**Table 3**

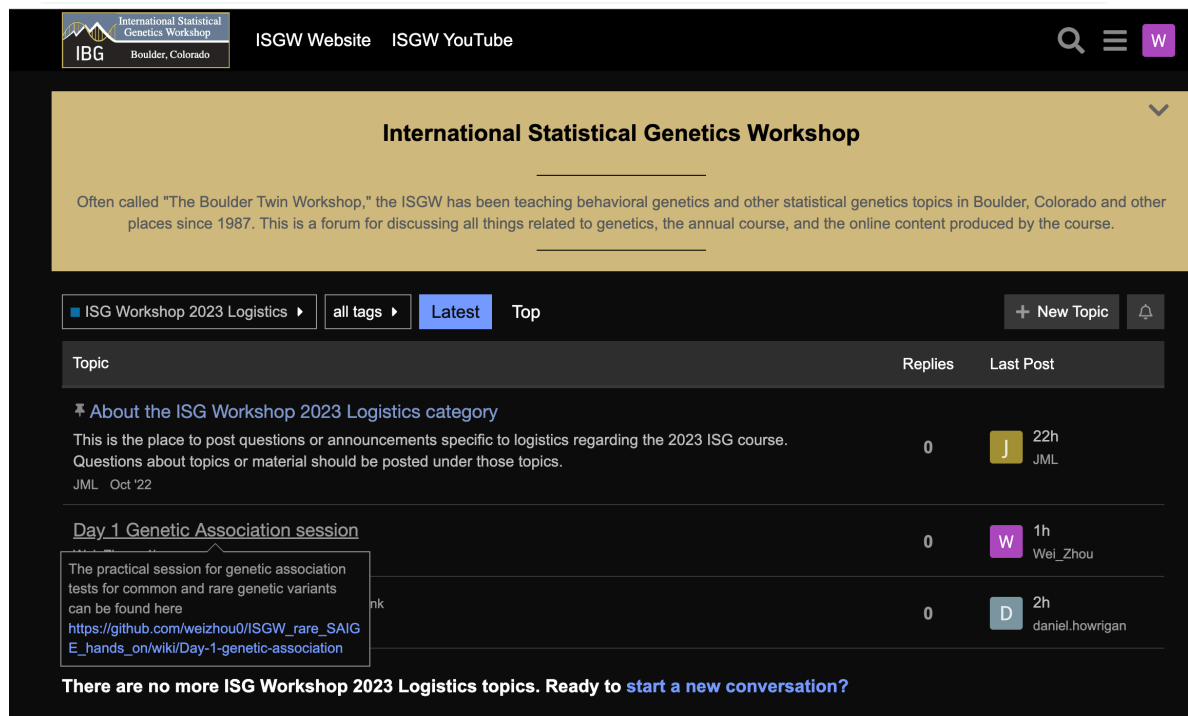
Overview of available tools used for the analysis of X-chromosome

<b>Name</b>	<b>Method</b>	<b>Platform</b>	<b>Weblink</b>
Impute2	Imputation	Unix/Win	<a href="#">Link</a>
PLINK	Quality control	Unix/Win	<a href="#">Link</a>
GWASTools	Quality control	R package	<a href="#">Link</a>
SNPTEST	Association testing	Unix/Win	<a href="#">Link</a>
snpStats	Association testing	Unix	<a href="#">Link</a>
XCMAX4	Association testing	R package	<a href="#">Link</a>
XCIR	XCI Inference	R Package	<a href="#">Link</a>
SkewXCI	XCI Inference	R Package	<a href="#">Link</a>
XWAS	Pipeline/Workflow	Unix/Win	<a href="#">Link</a>
GCTA	Association testing	Unix/Win	<a href="#">Link</a>
	XCI Inference		
Matrix eQTL	Association testing	R package	<a href="#">Link</a>



# Hands-on

- [https://github.com/weizhou0/ISGW\\_rare\\_SAIGE\\_hands\\_on/wiki/Day-1-genetic-association](https://github.com/weizhou0/ISGW_rare_SAIGE_hands_on/wiki/Day-1-genetic-association)
- Questions: <https://isgw-forum.colorado.edu/t/about-the-common-rare-variant-association-category/29/1>



The screenshot shows the forum interface for the International Statistical Genetics Workshop. At the top, there is a navigation bar with the workshop logo, 'ISGW Website', and 'ISGW YouTube'. Below this is a header section with the title 'International Statistical Genetics Workshop' and a brief description: 'Often called "The Boulder Twin Workshop," the ISGW has been teaching behavioral genetics and other statistical genetics topics in Boulder, Colorado and other places since 1987. This is a forum for discussing all things related to genetics, the annual course, and the online content produced by the course.'

The main content area features a filter bar with 'ISG Workshop 2023 Logistics' selected, 'all tags', 'Latest', and 'Top' buttons, along with '+ New Topic' and a notification bell. Below the filter bar is a table of forum topics:

Topic	Replies	Last Post
<a href="#">About the ISG Workshop 2023 Logistics category</a> This is the place to post questions or announcements specific to logistics regarding the 2023 ISG course. Questions about topics or material should be posted under those topics. JML Oct '22	0	JML 22h
<a href="#">Day 1 Genetic Association session</a> The practical session for genetic association tests for common and rare genetic variants can be found here <a href="https://github.com/weizhou0/ISGW_rare_SAIGE_hands_on/wiki/Day-1-genetic-association">https://github.com/weizhou0/ISGW_rare_SAIGE_hands_on/wiki/Day-1-genetic-association</a>	0	Wei_Zhou 1h
	0	daniel.howrigan 2h

At the bottom of the page, a message states: 'There are no more ISG Workshop 2023 Logistics topics. Ready to [start a new conversation?](#)'