Mixed model methods for GWAS

2025 International Statistical Genetics Workshop

Wei Zhou, Ph.D.
Massachusetts General Hospital/Harvard Medical School
Broad Institute
wzhou@broadinstitute.org

International Statistical Genetics Workshop • Boulder, Colorado •
IBG

UKBB: UK Biobank

>1,000 phenotypes curated from ICD codes

Clinical data

ICD9 and ICD10 codes

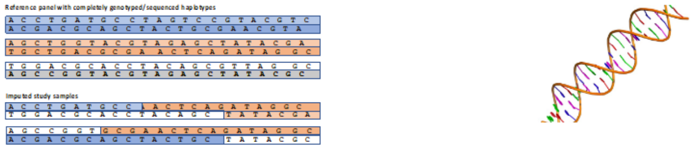
Questionnaires

Drug prescription

Imaging

Death registry
Cancer registry

Genetic Data



Reference panel with completely genotyped/sequenced haplotypes
Imputed study samples

Affymetrix UK Biobank Axiom array
Imputation from HRC+UK10K

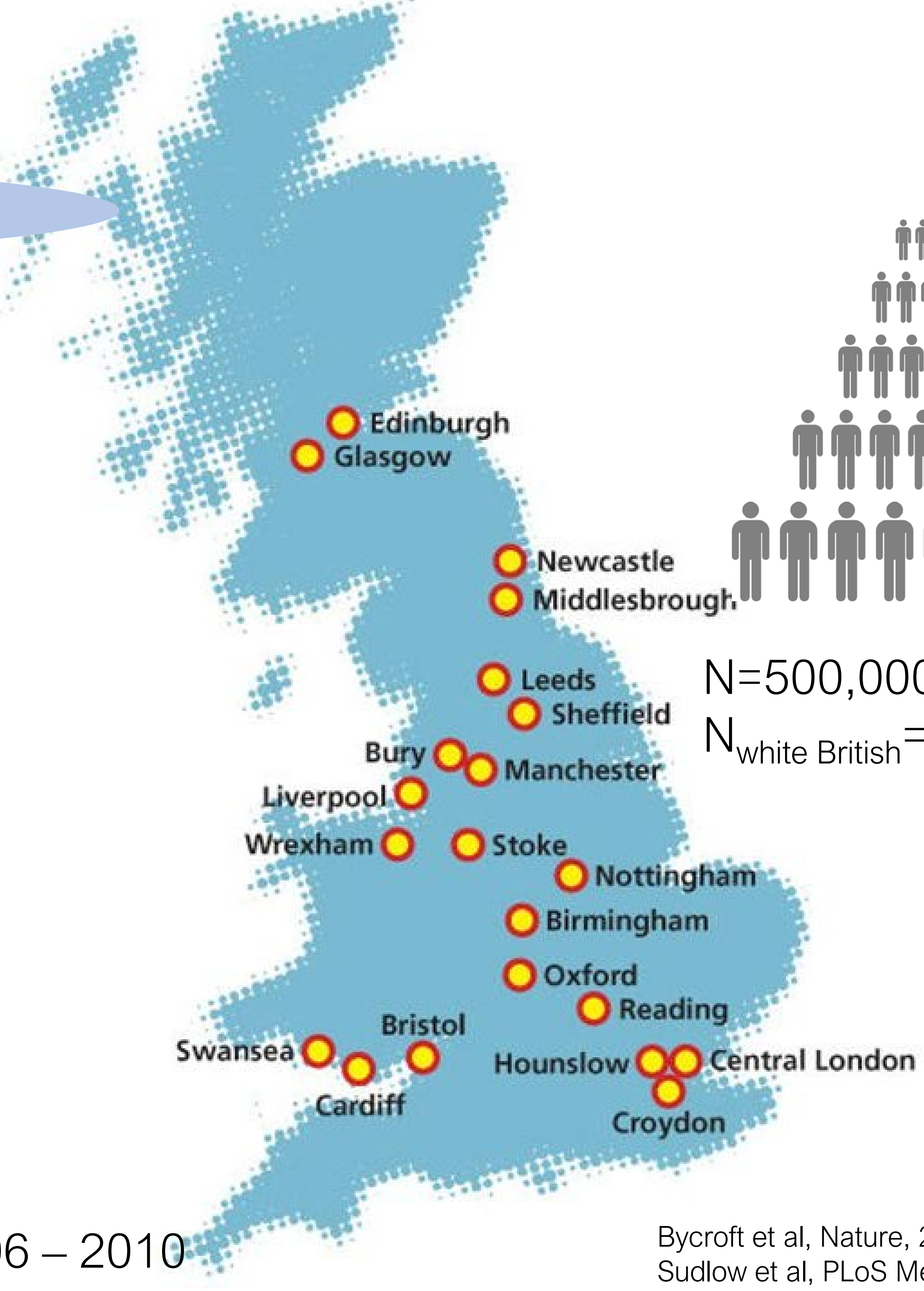
Whole exome sequencing
Whole genome sequencing

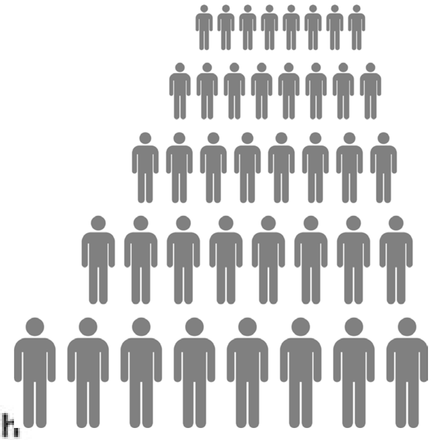
28 million genetic variants

biobank uk
Improving the health of future generations

Established 2007 in the United Kingdom

Data collected from 2006 – 2010





$N=500{,}000$

$N_{\text{white British}} = 408{,}961$

Bycroft et al, Nature, 2018
Sudlow et al, PLoS Med, 2015

Challenges in genetic association studies

Linear model:

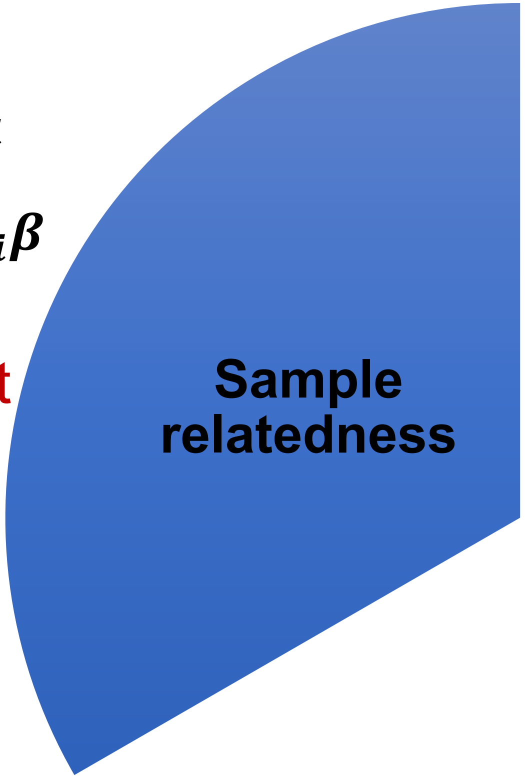
$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + \epsilon_i}$$

Logistic model:

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

$$\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I\right)$$

Assumes independent observations



Sample relatedness

1 in 3 has at least one relative up to the 3rd degree in UK Biobank

- Inflated type I errors
- Biased effect estimates

GWAS results based on linear or logistic regression can be **biased** when the **independence** assumption between samples is violated

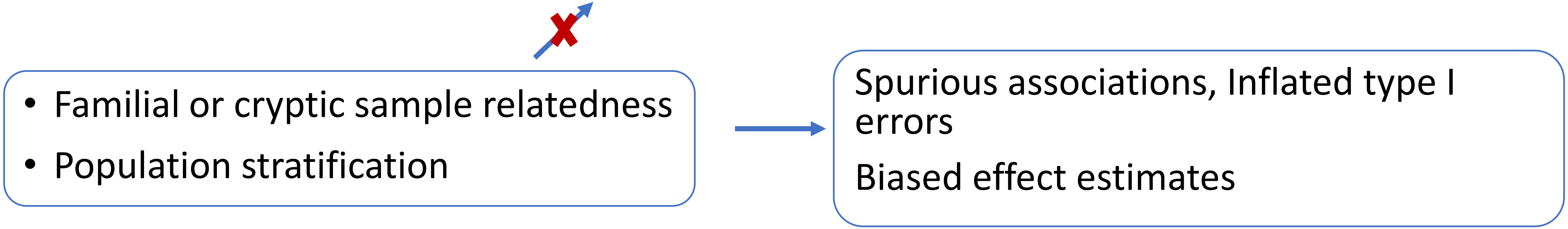
Linear model:

$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + \epsilon_i}$$

Logistic model:

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

Assumes independent observations



- Familial or cryptic sample relatedness
- Population stratification

Spurious associations, Inflated type I errors

Biased effect estimates

Challenges in genetic association studies

Linear model:

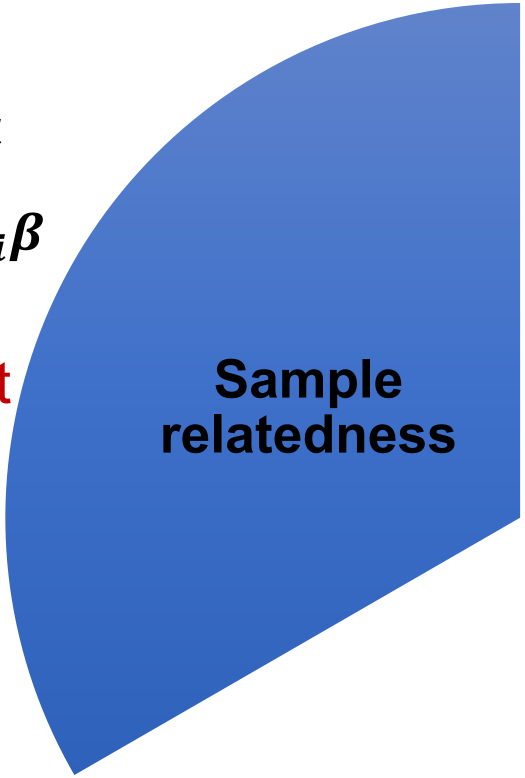
$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + \epsilon_i}$$

Logistic model:

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

$$\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I\right)$$

Assumes independent observations



Sample relatedness

1 in 3 has at least one relative up to the 3rd degree in UK Biobank

- Inflated type I errors
- Biased effect estimates

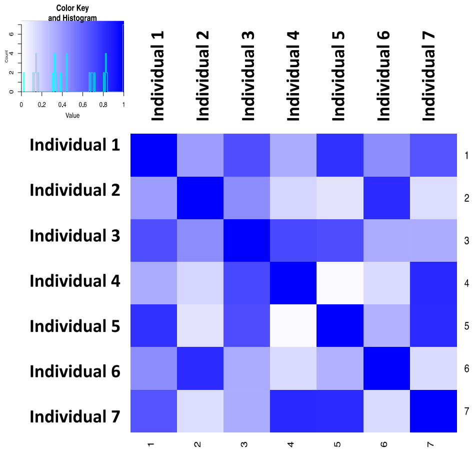
By excluding up to 3rd degree relatives in samples with EUR ancestry, we will lose ~10% of samples (out of 400k)

Linear mixed model for GWAS

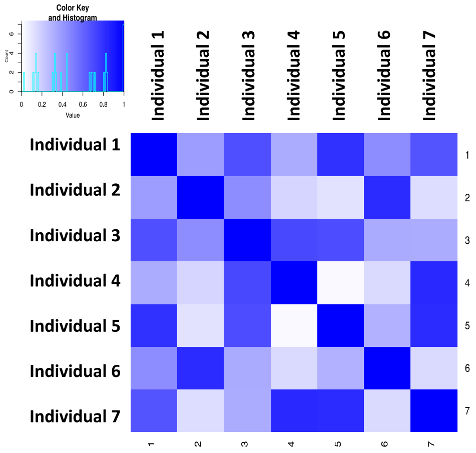
$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + b_i + \epsilon_i}$$

Accounting for sample relatedness

- b: random genetic effect, $b \sim N(0, \tau\,\psi)$, $\boldsymbol{\psi}$ **is genetic relationship matrix (GRM)**



Genetic relationship matrix



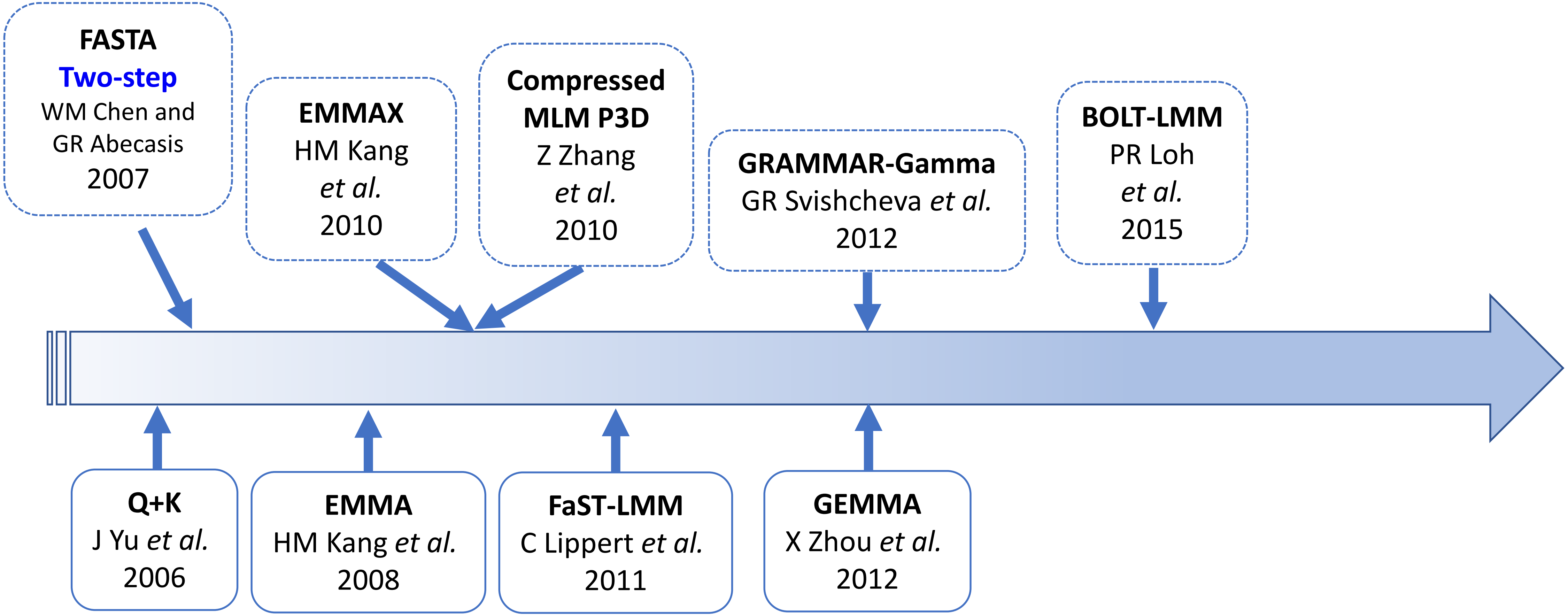
Square, symmetric matrix

- **Standardized Genotype Approach (commonly used in GWAS)**

$$\psi_{ij} = \frac{1}{M}\sum_{m=1}^{M}\frac{(g_{im}-2p_m)(g_{jm}-2p_m)}{2p_m(1-p_m)}$$

  - g_{im} is the genotype (0, 1, or 2) for individual i at SNP m
  - p_m is the allele frequency of SNP m
  - M is the total number of SNPs

Linear mixed model methods for GWAS



FASTA
Two-step
WM Chen and
GR Abecasis
2007

EMMAX
HM Kang
et al.
2010

Compressed MLM P3D
Z Zhang
et al.
2010

GRAMMAR-Gamma
GR Svishcheva *et al.*
2012

BOLT-LMM
PR Loh
et al.
2015

Q+K
J Yu *et al.*
2006

EMMA
HM Kang *et al.*
2008

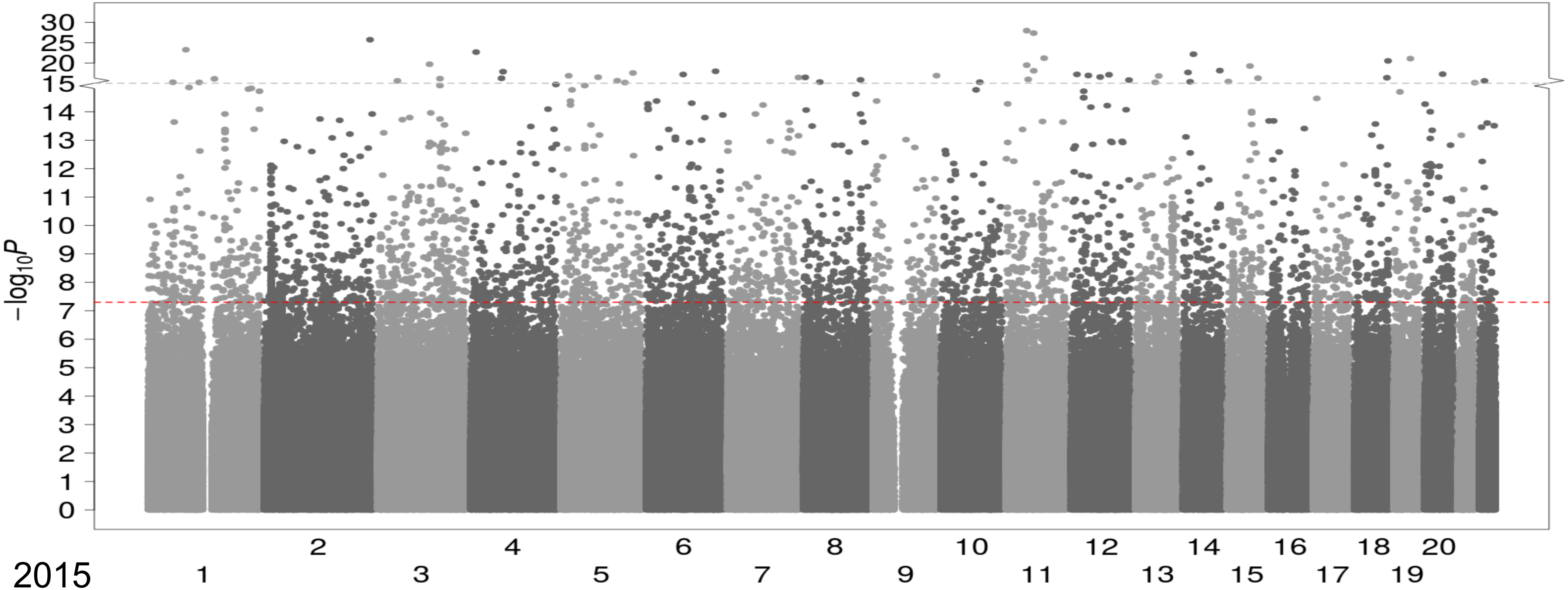
FaST-LMM
C Lippert *et al.*
2011

GEMMA
X Zhou *et al.*
2012

Binary trait in UKBB	N_{Case}	$N_{Control}$
Colorectal cancer	4,562	382,756

Inflated type I error rates were observed after using BOLT-LMM for binary phenotypes

Binary Traits	N $_{Case}$	N $_{Control}$
Colorectal cancer	4,562	382,756

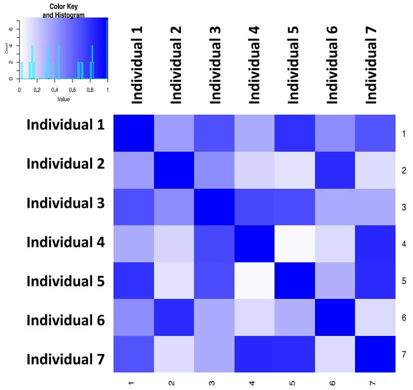


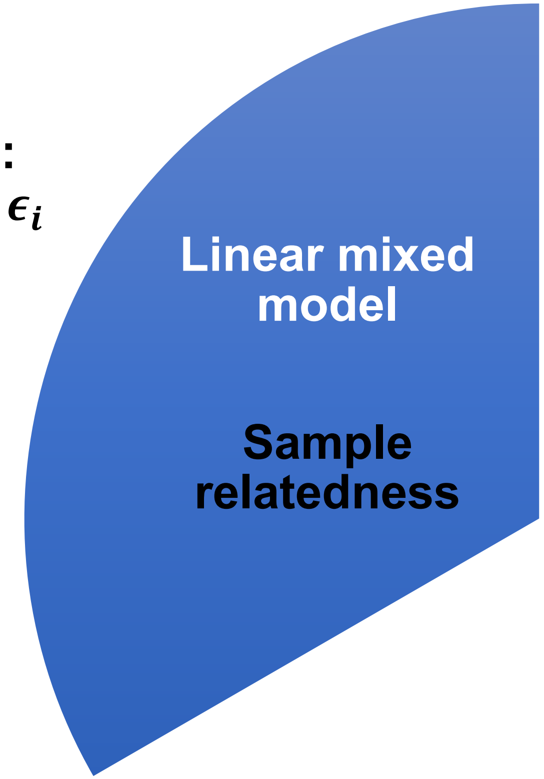
Loh *et al.*, 2015

Challenges in biobank-based GWASs

Linear mixed model:

$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + b_i + \epsilon_i}$$





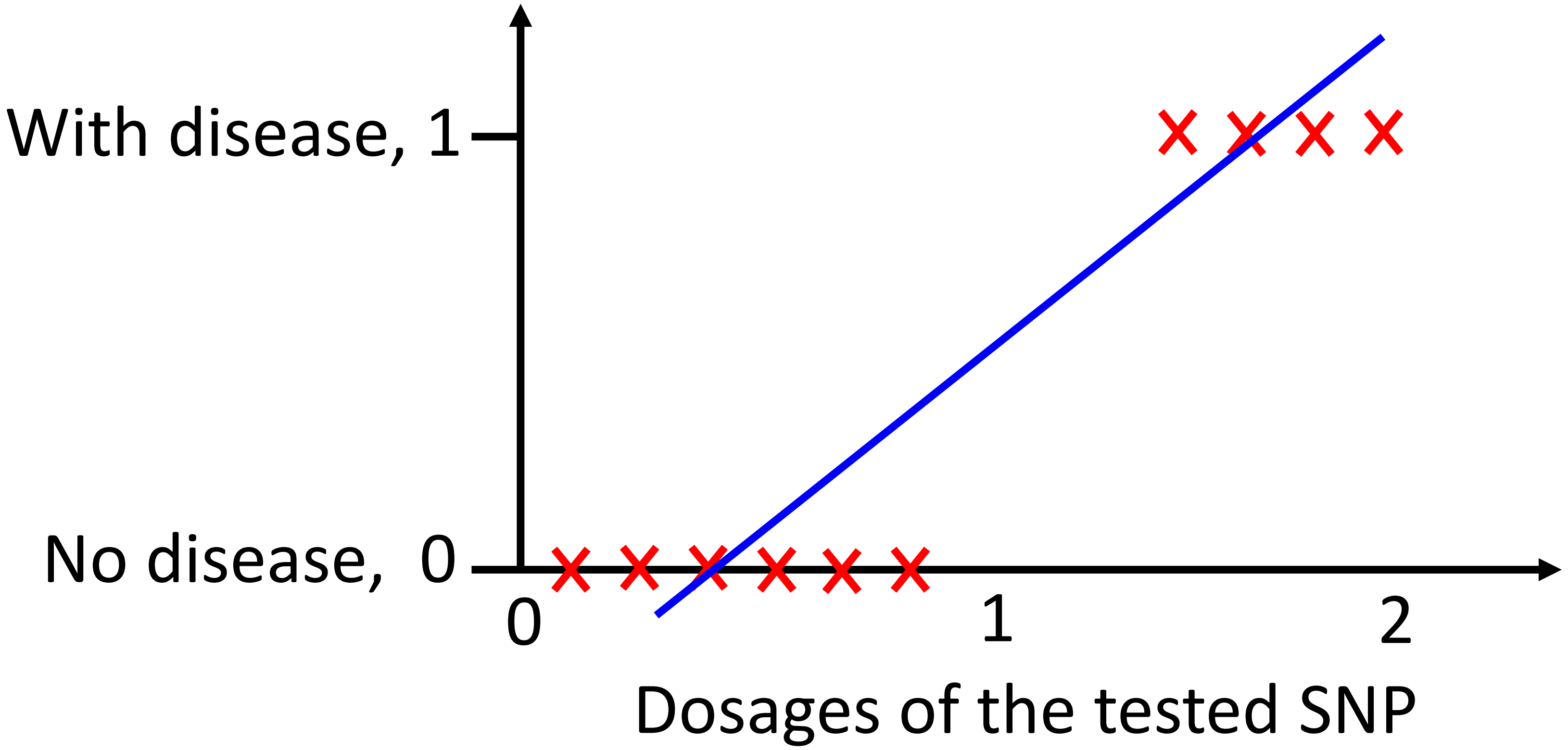
Linear mixed model

Sample relatedness

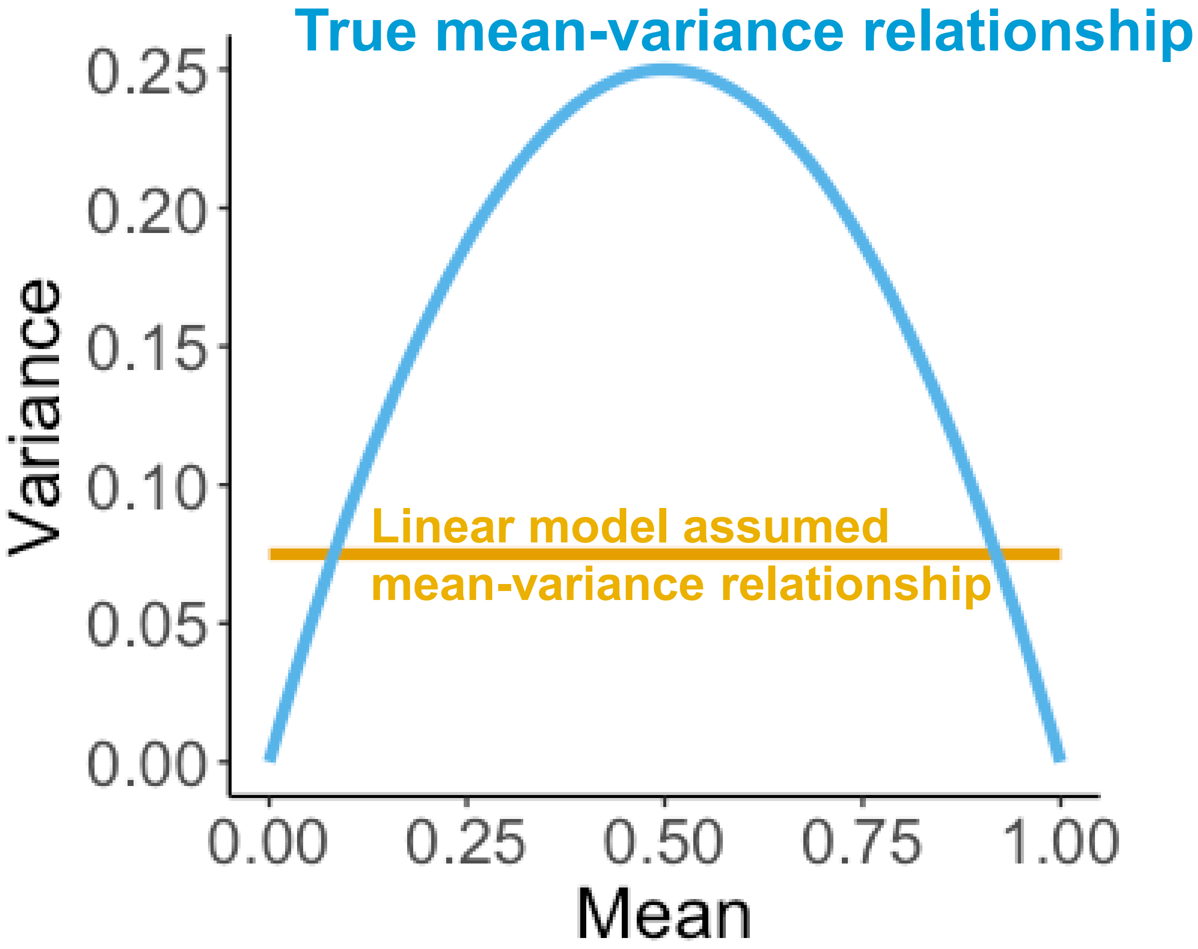
?

Linear Mixed Model for Binary Phenotypes?

- Assumes homoscedasticity (constant residual variance)
  - Violated by binary traits ⟶ **Inflated type I error rates**



Linear mixed model for **binary phenotypes (0 and 1)**?



adapted from Chen, H., Wang, C., *et. al.* (2016)

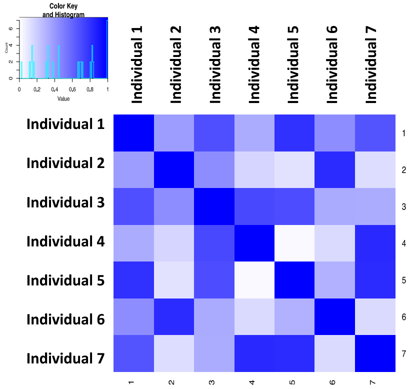
Use logistic mixed model for binary phenotypes

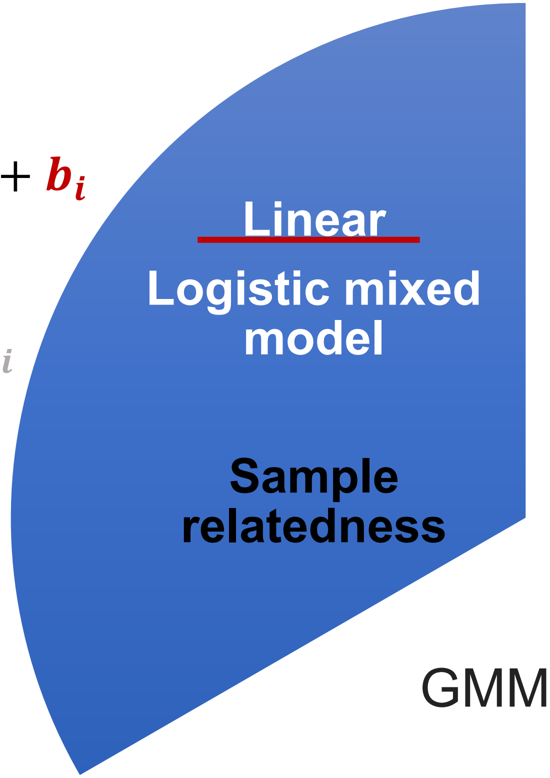
Logistic mixed model:

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

Linear mixed model:

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





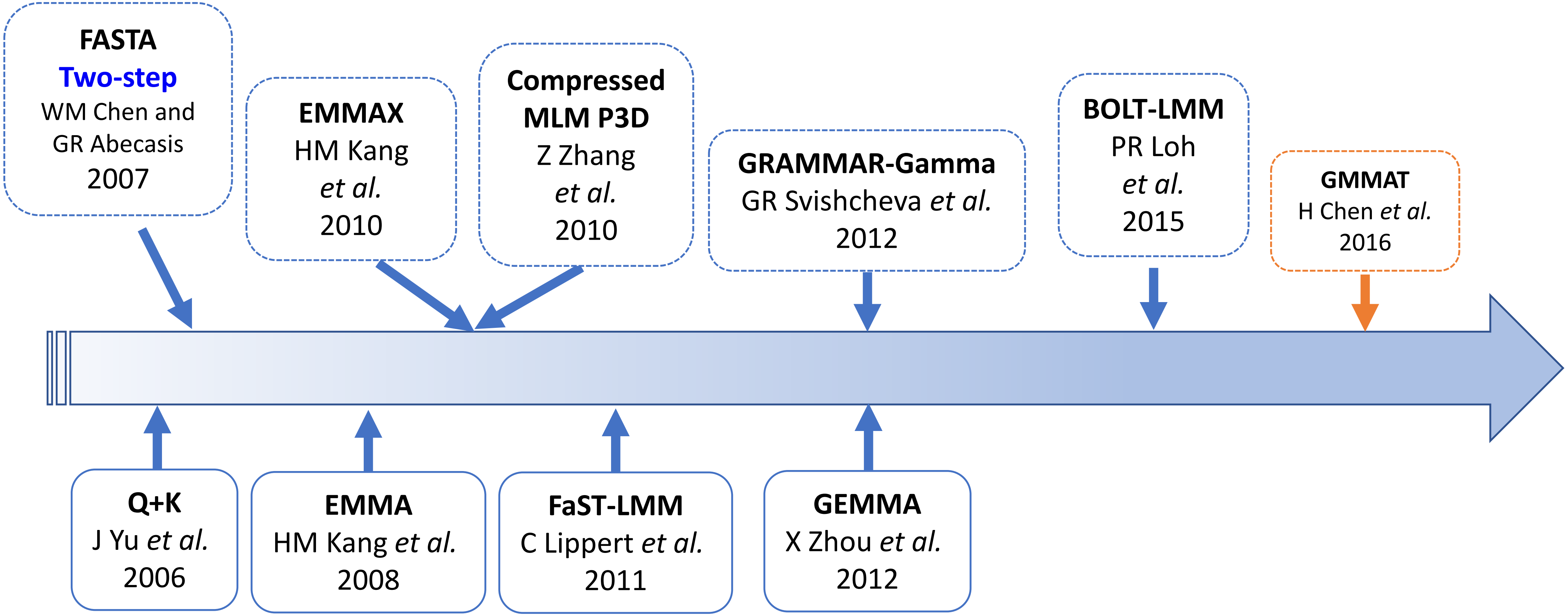
~~Linear~~

Logistic mixed model

Sample relatedness

GMMAT: Chen, H., Wang, C., *et. al.* (2016)

GMMAT: Logistic Mixed Model Association Test



FASTA
Two-step
WM Chen and
GR Abecasis
2007

EMMAX
HM Kang
et al.
2010

Compressed MLM P3D
Z Zhang
et al.
2010

GRAMMAR-Gamma
GR Svishcheva *et al.*
2012

BOLT-LMM
PR Loh
et al.
2015

GMMAT
H Chen *et al.*
2016

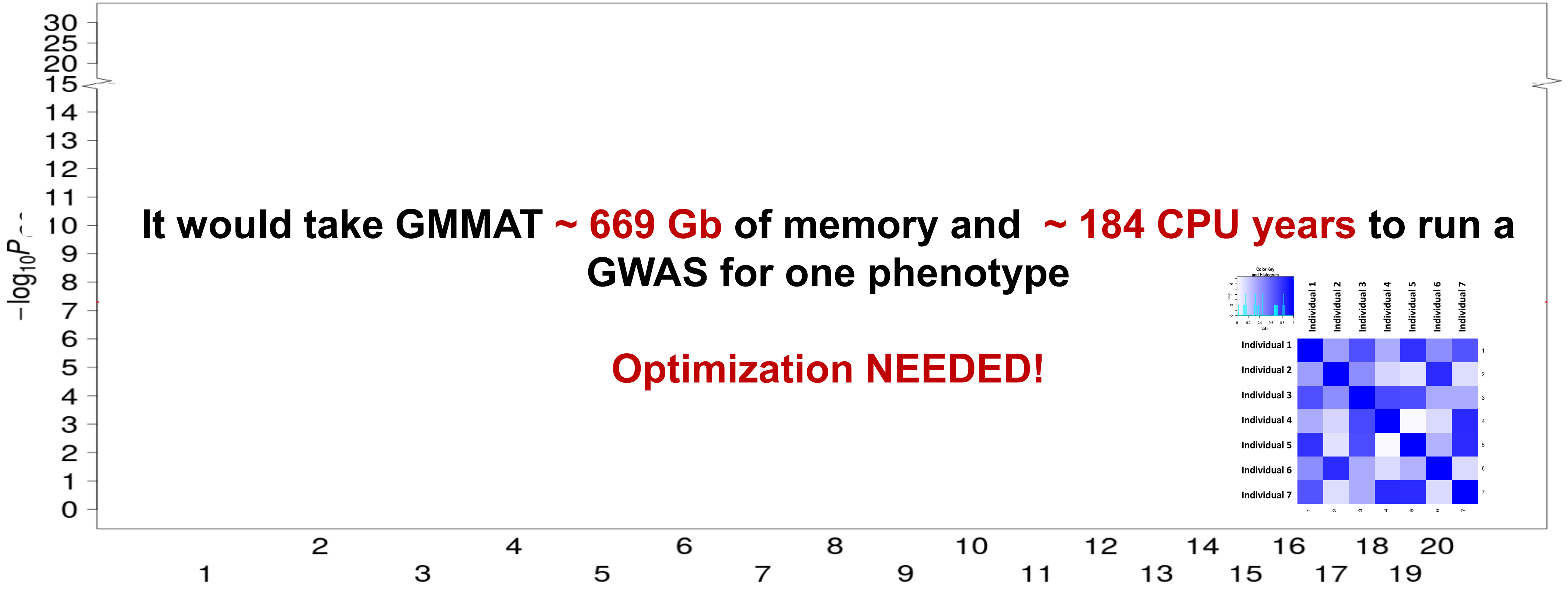
Q+K
J Yu *et al.*
2006

EMMA
HM Kang *et al.*
2008

FaST-LMM
C Lippert *et al.*
2011

GEMMA
X Zhou *et al.*
2012

Binary Traits	N_{Case}	$N_{Control}$
Colorectal cancer	4,562	382,756

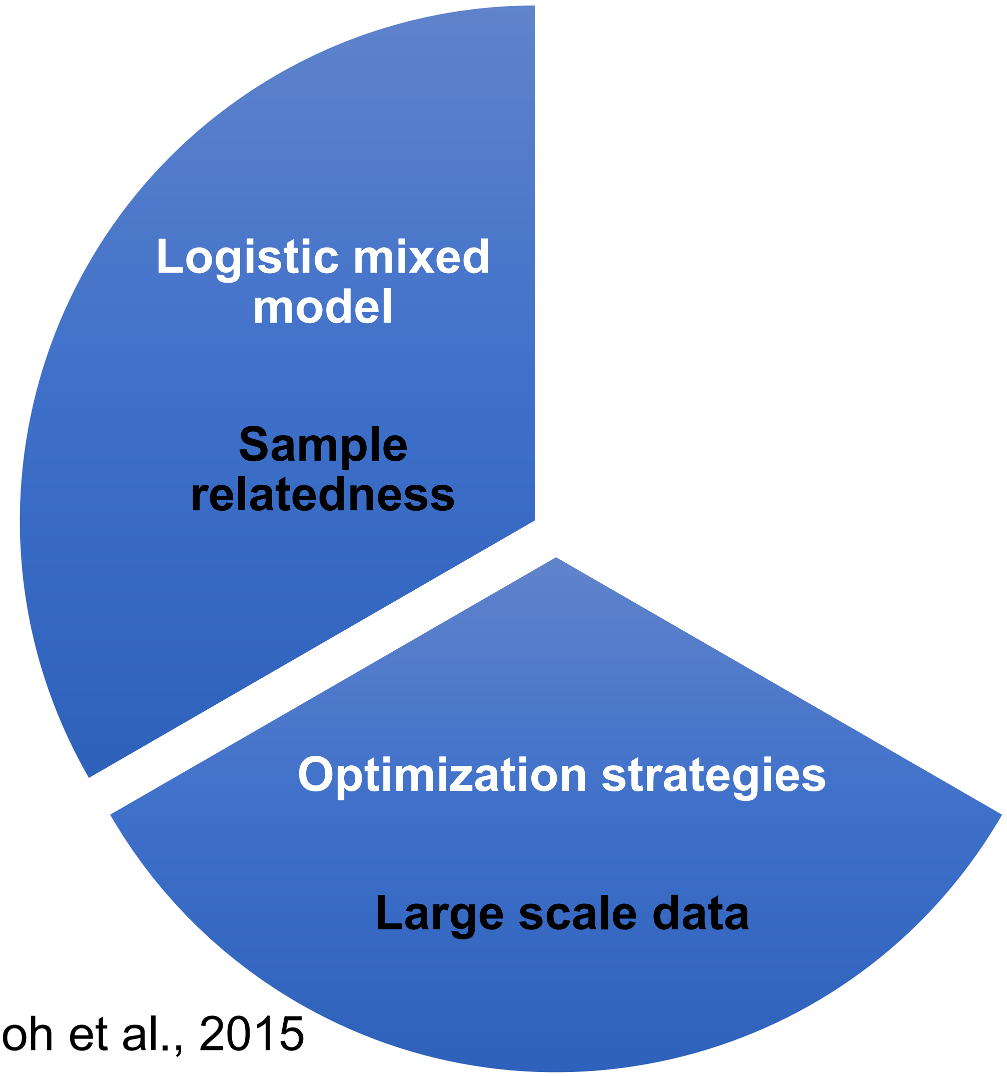


It would take GMMAT ~ 669 Gb of memory and ~ 184 CPU years to run a GWAS for one phenotype

Optimization NEEDED!

GMMAT: Chen, H., Wang, C., *et. al.* (2016)

Optimize logistic mixed model method for biobank-scale data



Logistic mixed model

Sample relatedness

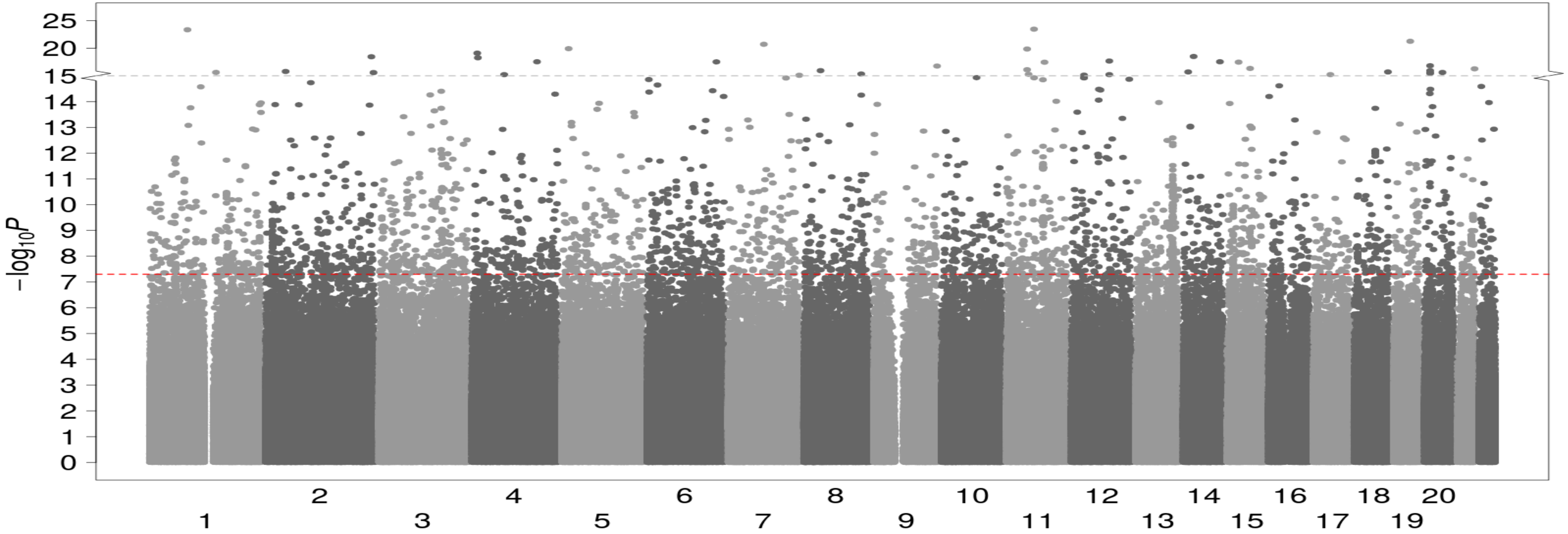
Optimization strategies

Large scale data

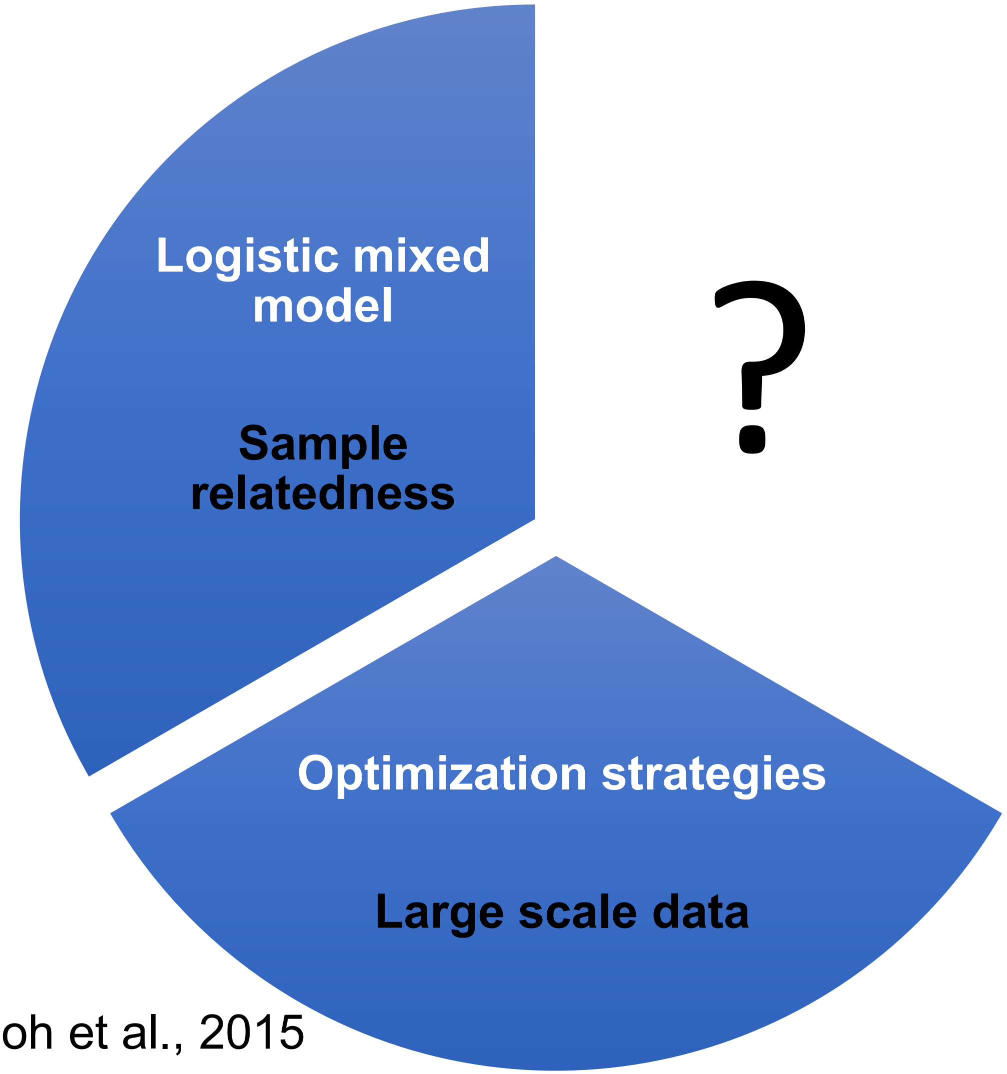
Loh et al., 2015

Inflated type I error rates were still observed after using logistic mixed model for binary phenotypes

Binary Traits	N $_{\text{Case}}$	N $_{\text{Control}}$
Colorectal cancer	4,562	382,756



Challenges in biobank-based GWASs



Logistic mixed model

Sample relatedness

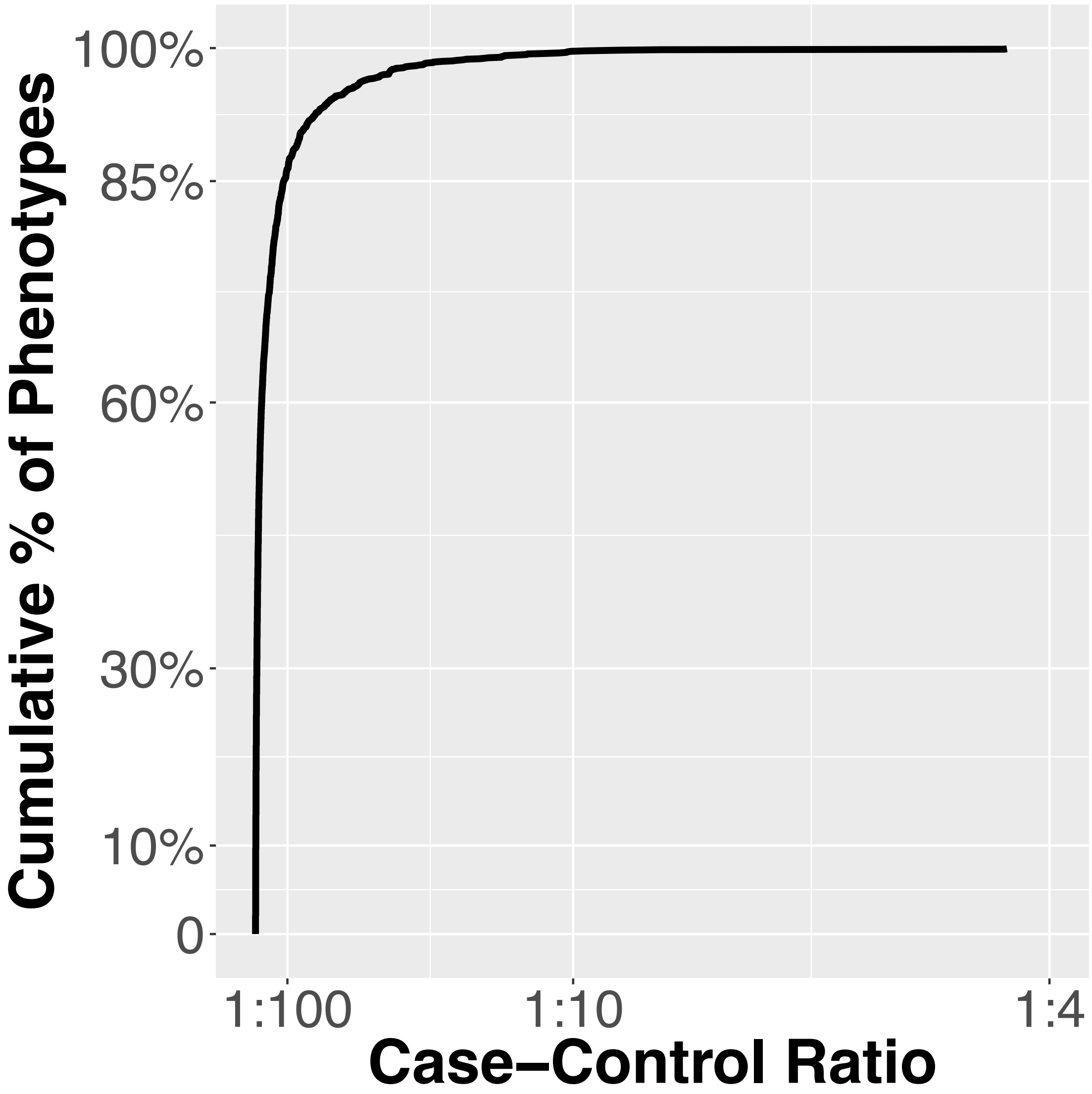
?

Optimization strategies

Large scale data

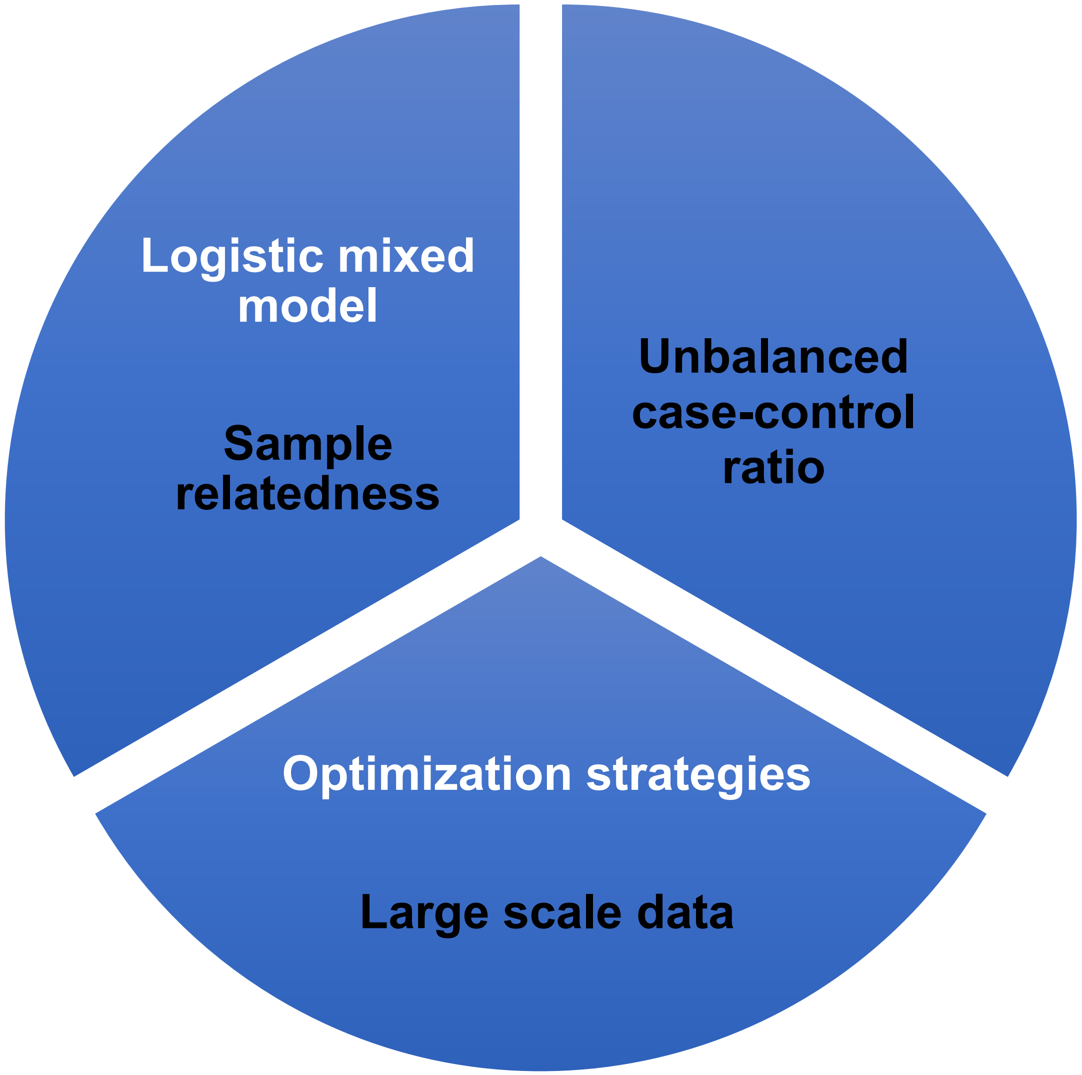
Loh et al., 2015

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



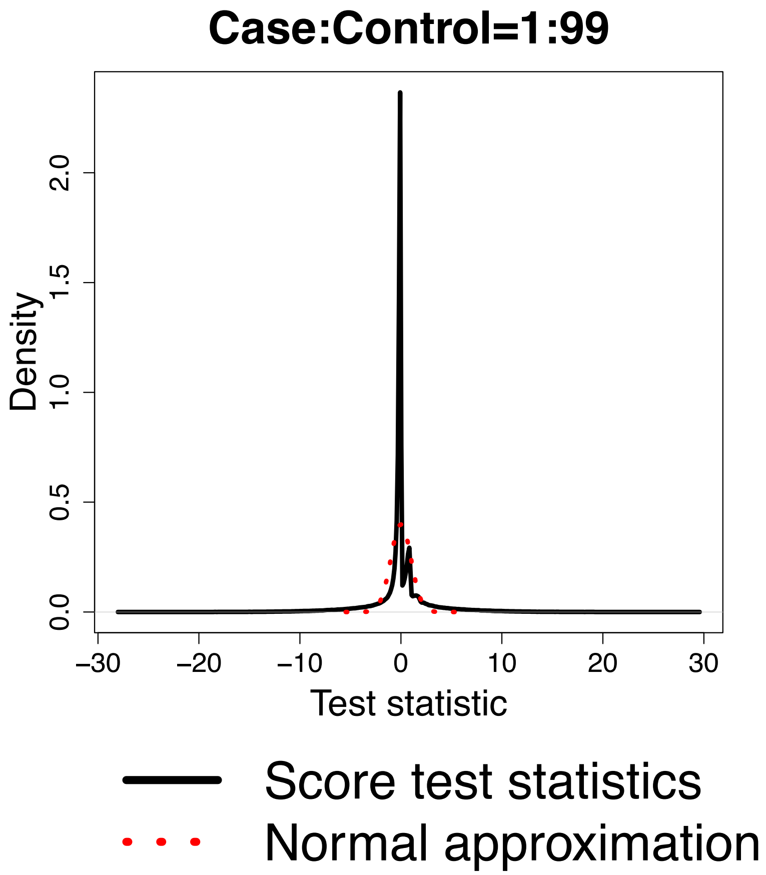
Logistic mixed model

Sample relatedness

Unbalanced case-control ratio

Optimization strategies

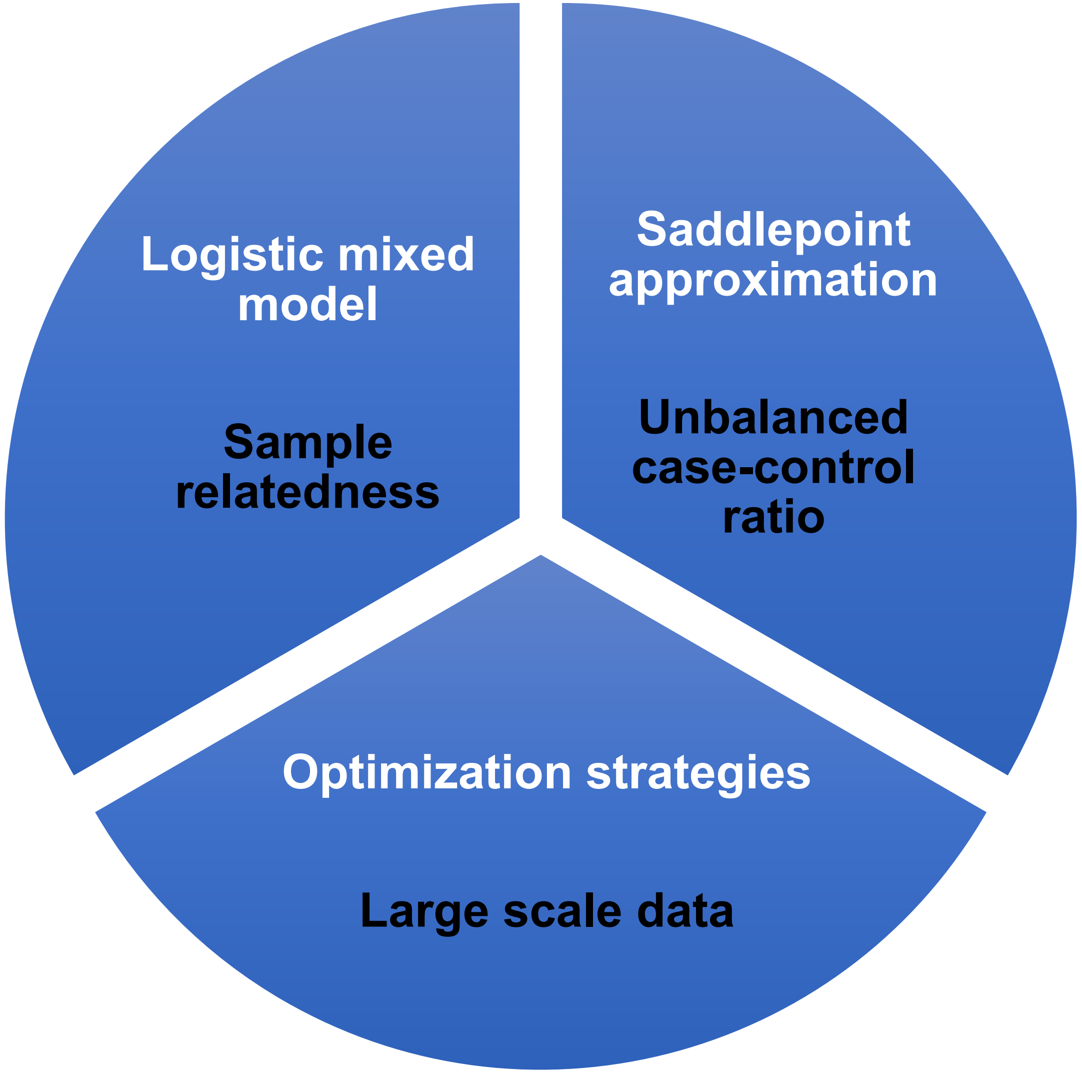
Large scale data



Case:Control=1:99

Ma, *et al.* (2013)

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



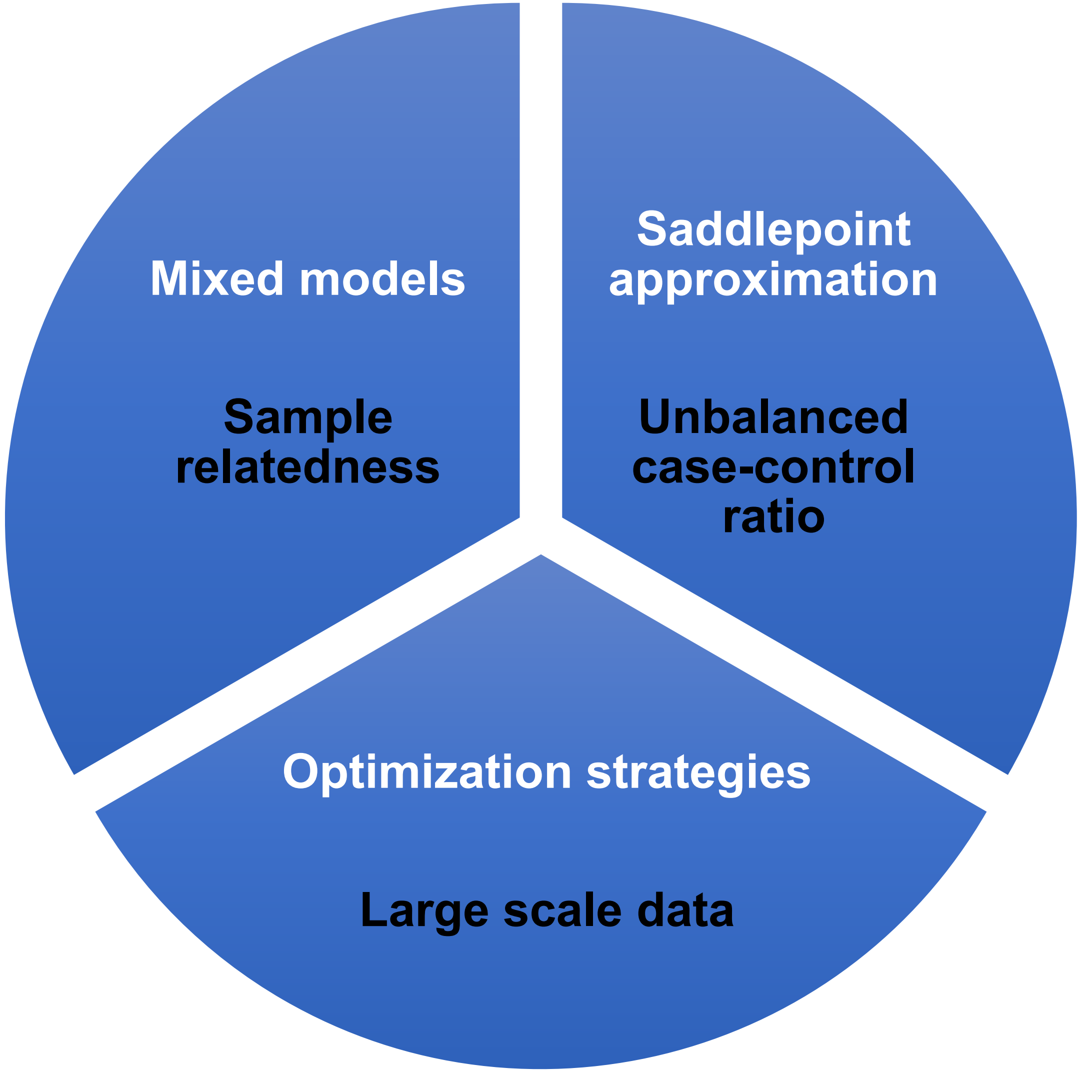
SPA uses the **entire moment generating function -> more accurate p-values**

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



Mixed models

Sample relatedness

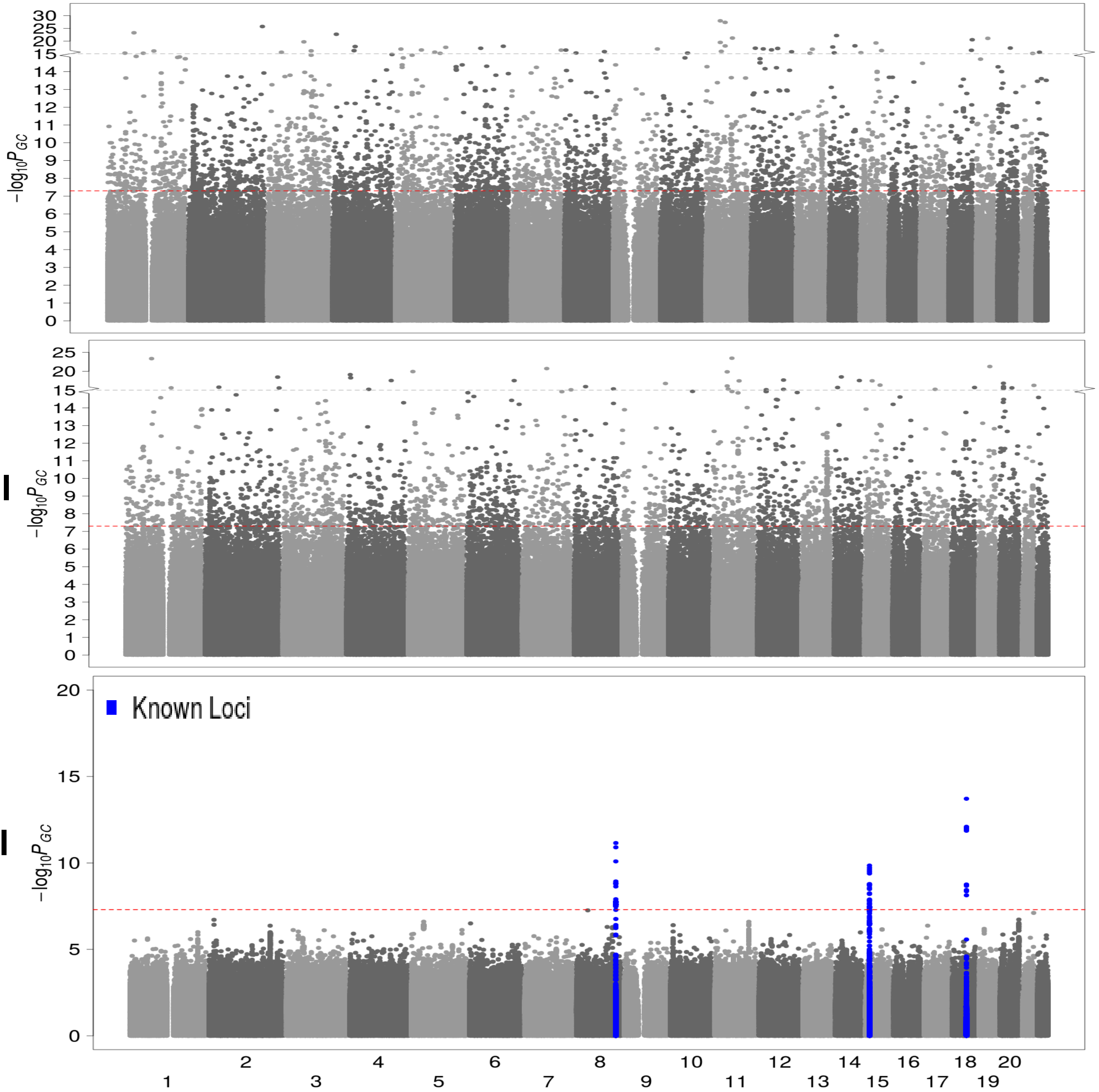
Saddlepoint approximation

Unbalanced case-control ratio

Optimization strategies

Large scale data

Zhou et al. *Nat. Genet*. 2018



Linear mixed model

Logistic mixed model

Logistic mixed model
+SPA (SAIGE)

Colorectal cancer

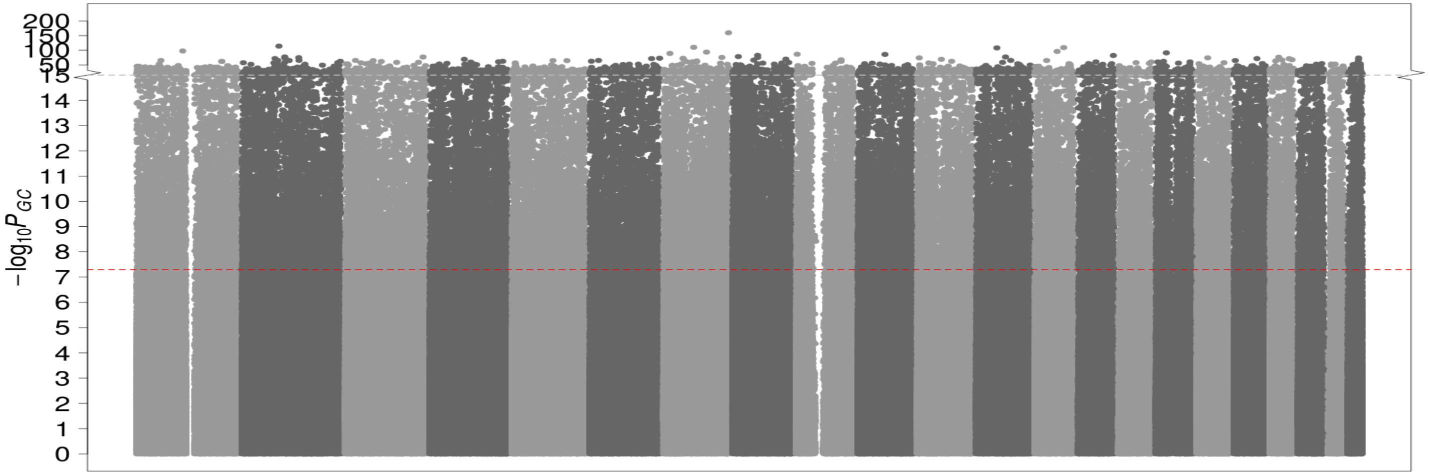
- **4,562 Cases**
- **382,756 Controls**
- **Case: Control = 1:84**

■ Known Loci

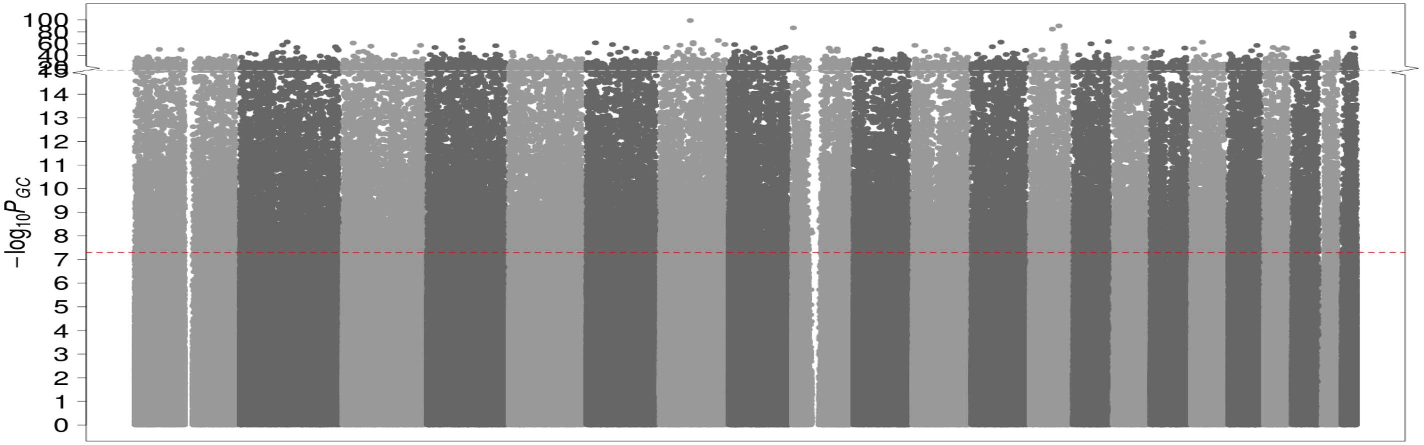
Thyroid cancer

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

Linear mixed model



Logistic mixed model



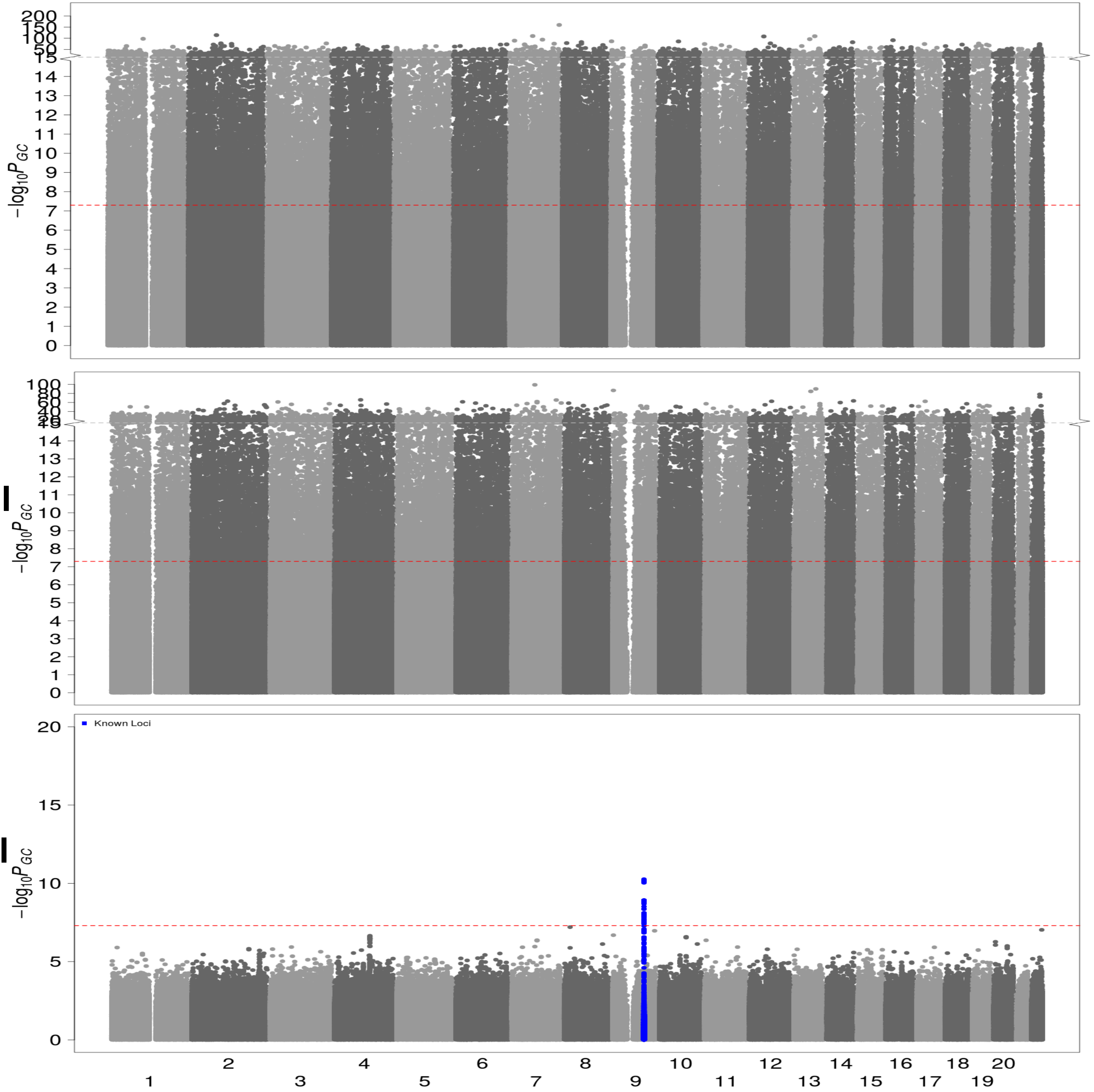
Thyroid cancer

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

Linear mixed model

Logistic mixed model

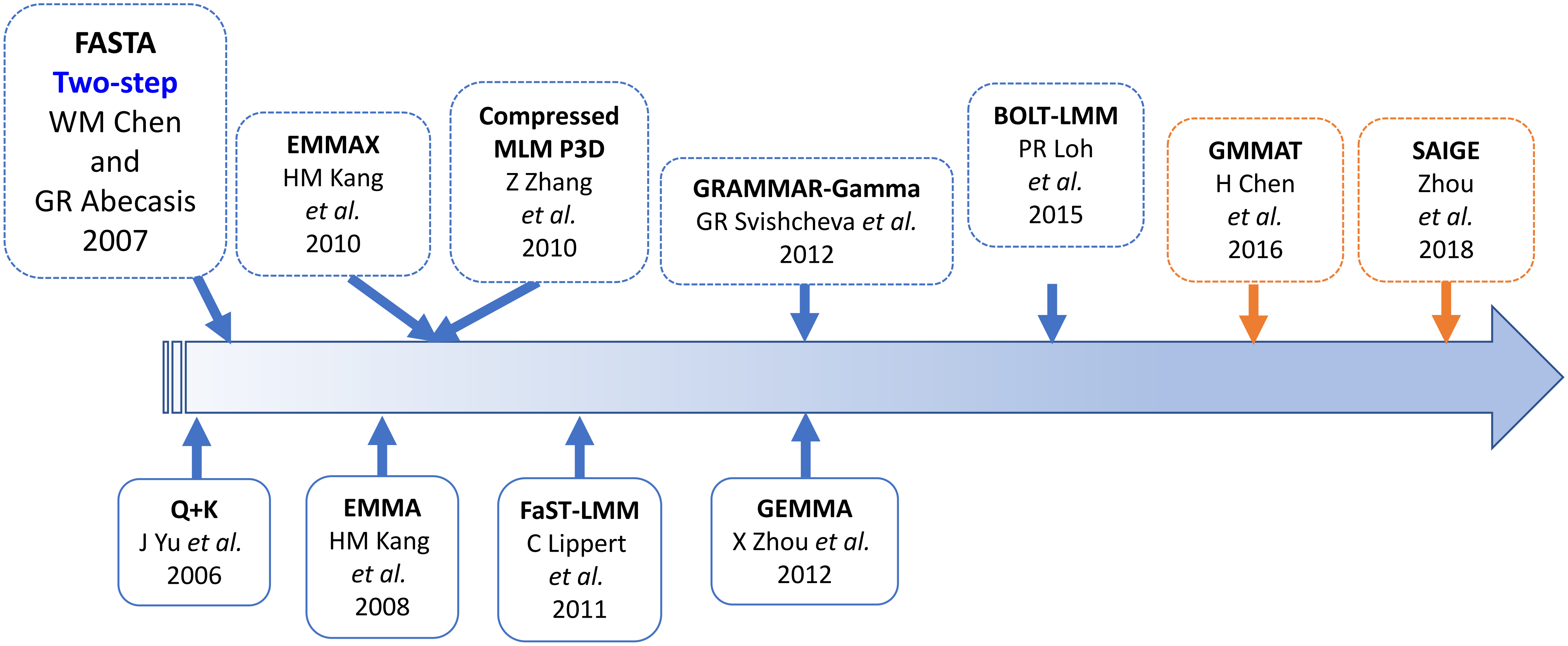
Logistic mixed model
+SPA



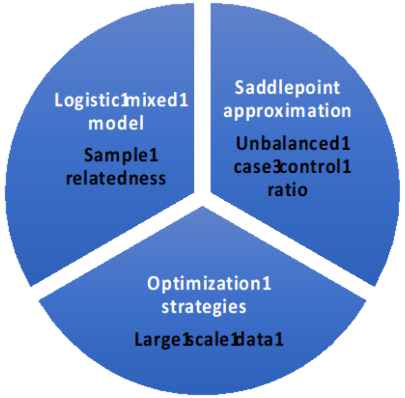
Thyroid cancer

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

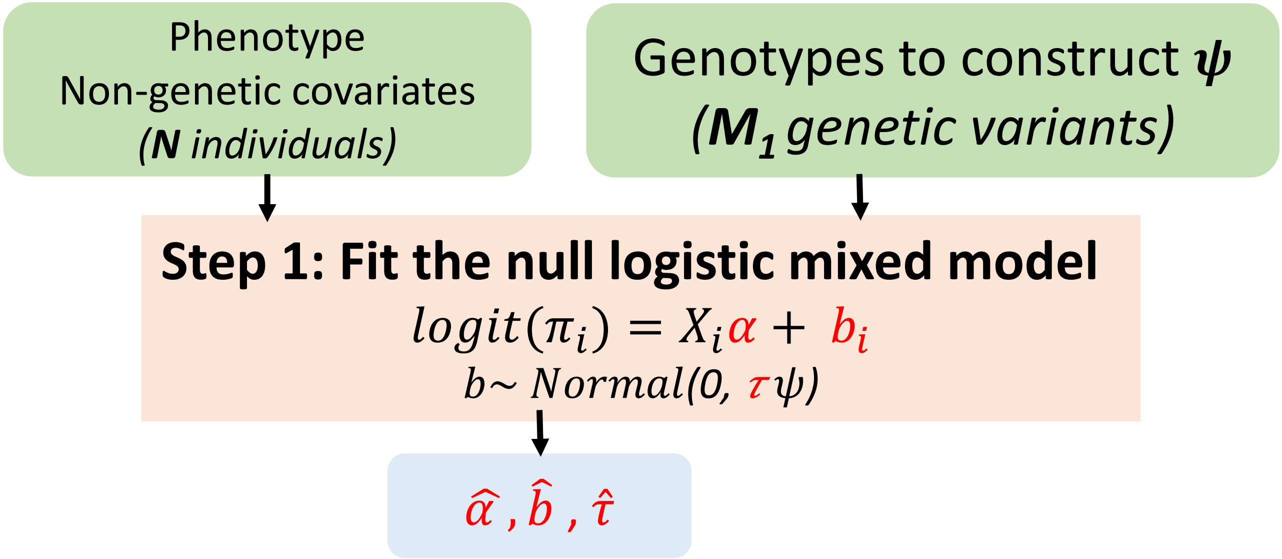
■ Known Loci



SAIGE documentation: https://saigegit.github.io/SAIGE-doc/



SAIGE



Phenotype
Non-genetic covariates
*(**N** individuals)*

Genotypes to construct $\boldsymbol{\psi}$
*(**M_1** genetic variants)*

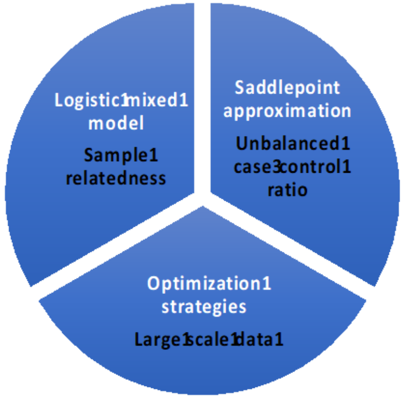
Step 1: Fit the null logistic mixed model

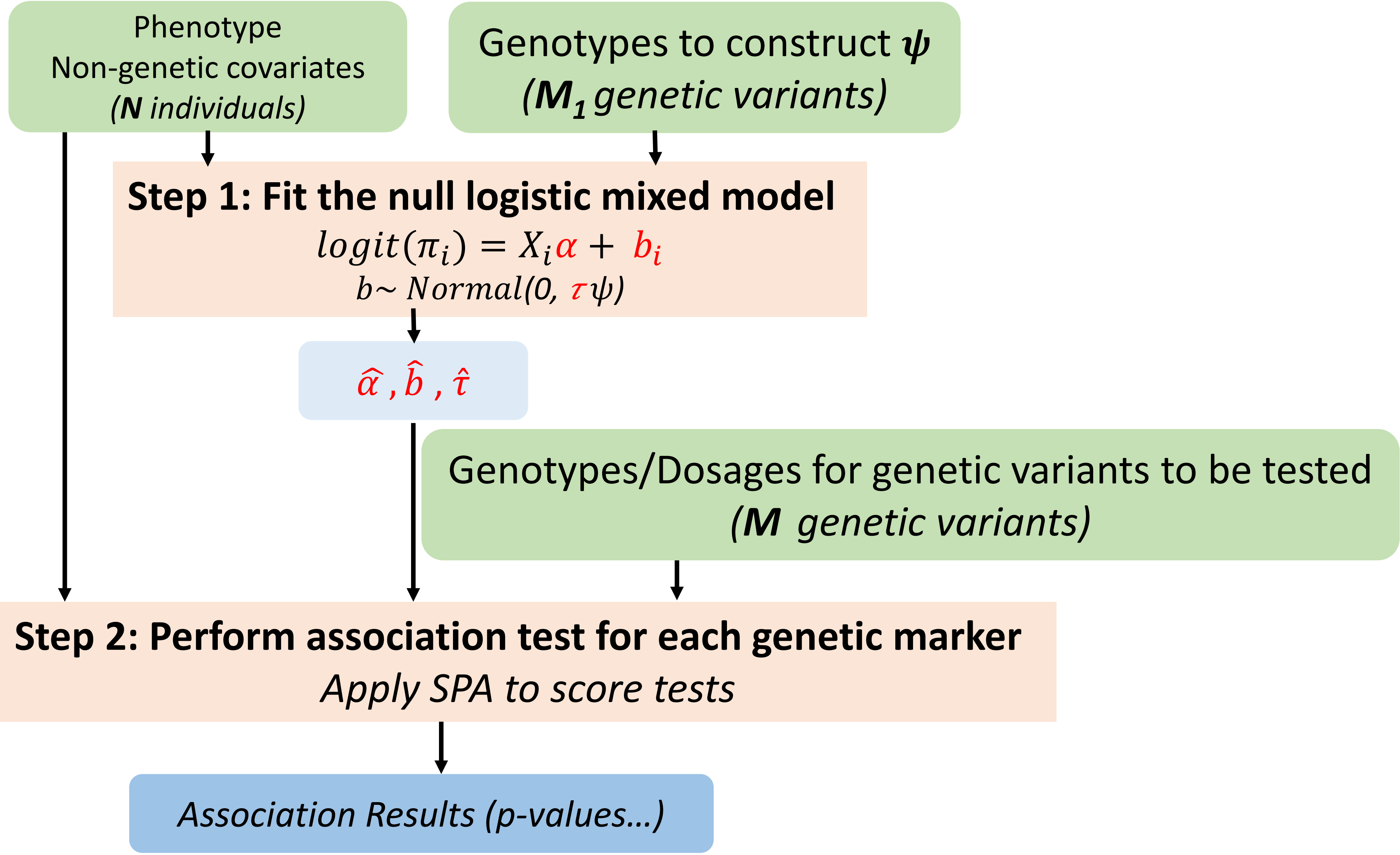
$$logit(\pi_i) = X_i\alpha + b_i$$

$$b \sim Normal(0, \tau\psi)$$

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

SAIGE





Phenotype
Non-genetic covariates
*(**N** individuals)*

Genotypes to construct $\boldsymbol{\psi}$
*(**M₁** genetic variants)*

Step 1: Fit the null logistic mixed model

$$logit(\pi_i) = X_i\alpha + b_i$$

$$b \sim Normal(0, \tau\psi)$$

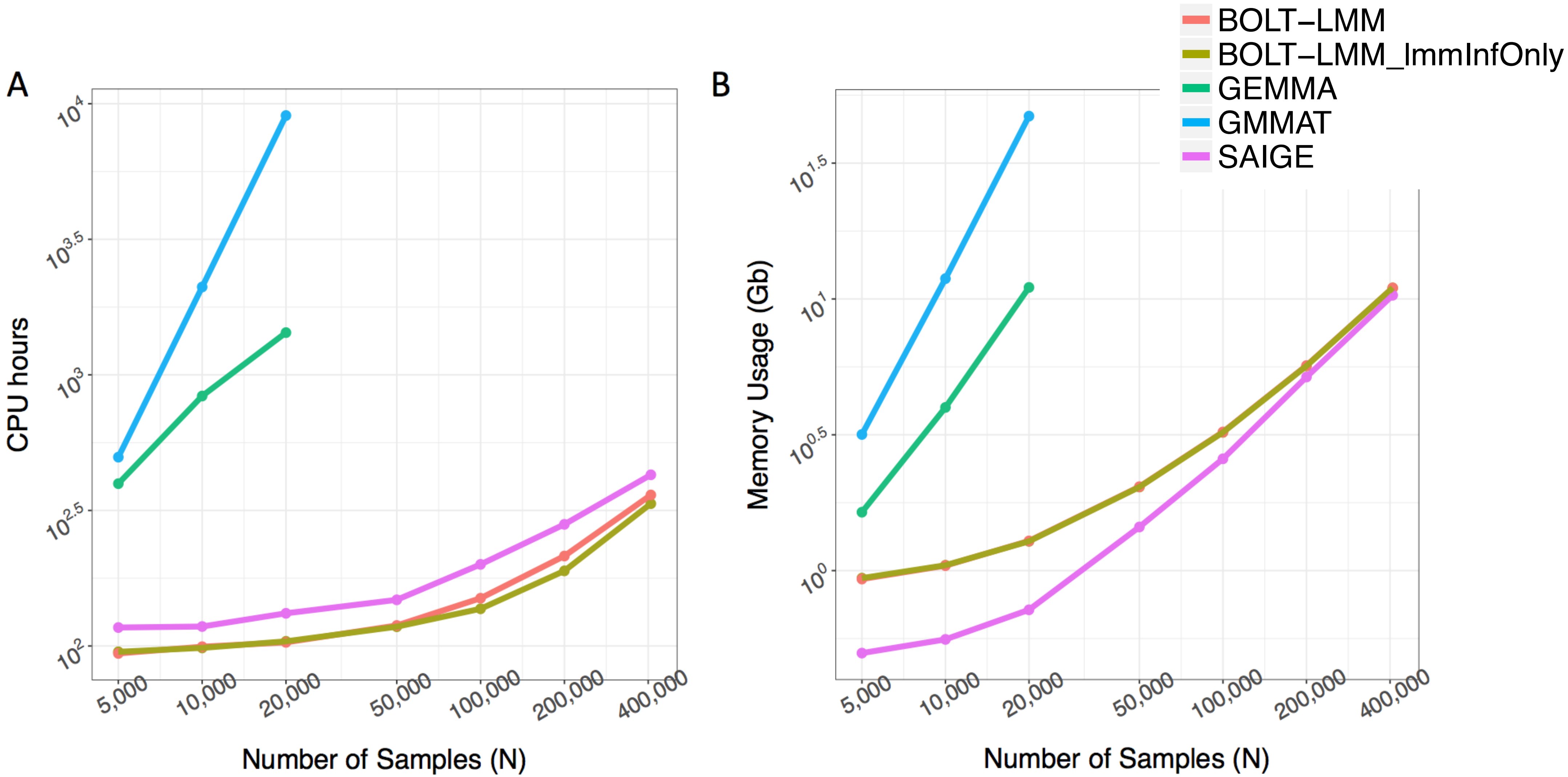
$$\widehat{\alpha}, \widehat{b}, \hat{\tau}$$

Genotypes/Dosages for genetic variants to be tested
*(**M** genetic variants)*

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

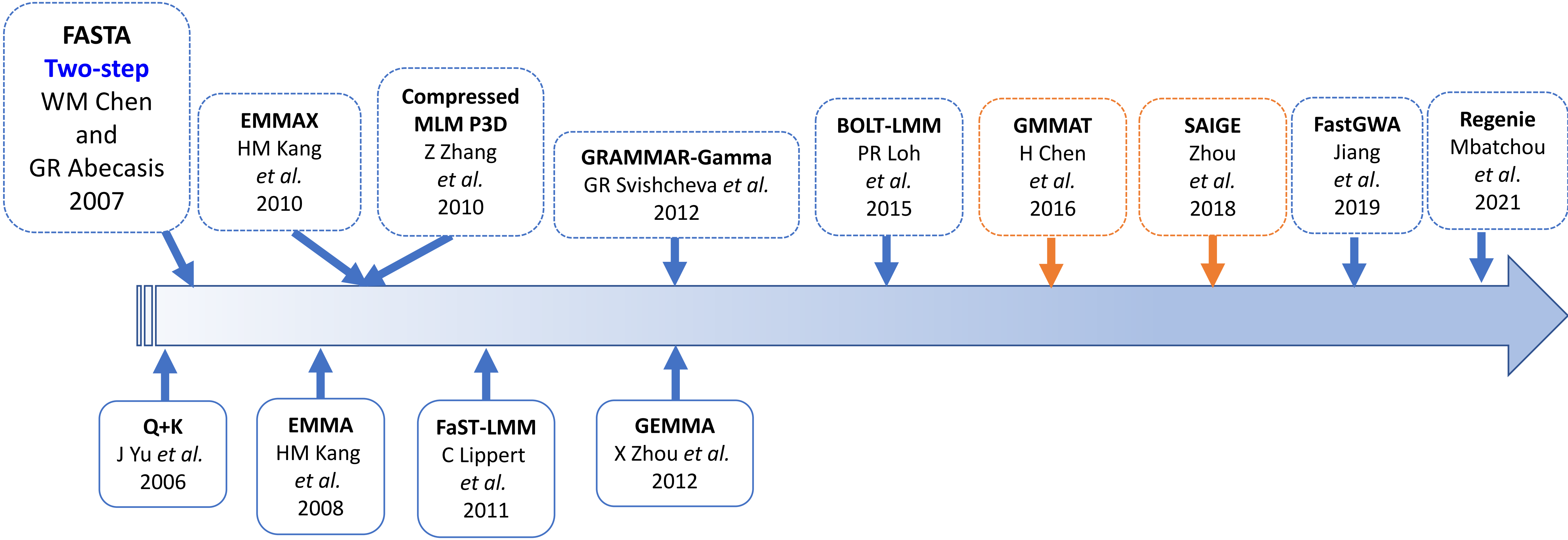
Association Results (p-values…)

Run Time and Memory Usage



Log-log plots of the **estimated** run time (A) and memory use (B) as a function of sample size (N) for testing for testing 71 million markers with info ≥ 0.3 as in UK Biobank.

More popular methods developed to improve the computational efficiency for running mixed model-based GWAS in large biobanks



FASTA
Two-step
WM Chen
and
GR Abecasis
2007

EMMAX
HM Kang
et al.
2010

Compressed MLM P3D
Z Zhang
et al.
2010

GRAMMAR-Gamma
GR Svishcheva *et al.*
2012

BOLT-LMM
PR Loh
et al.
2015

GMMAT
H Chen
et al.
2016

SAIGE
Zhou
et al.
2018

FastGWA
Jiang
et al.
2019

Regenie
Mbatchou
et al.
2021

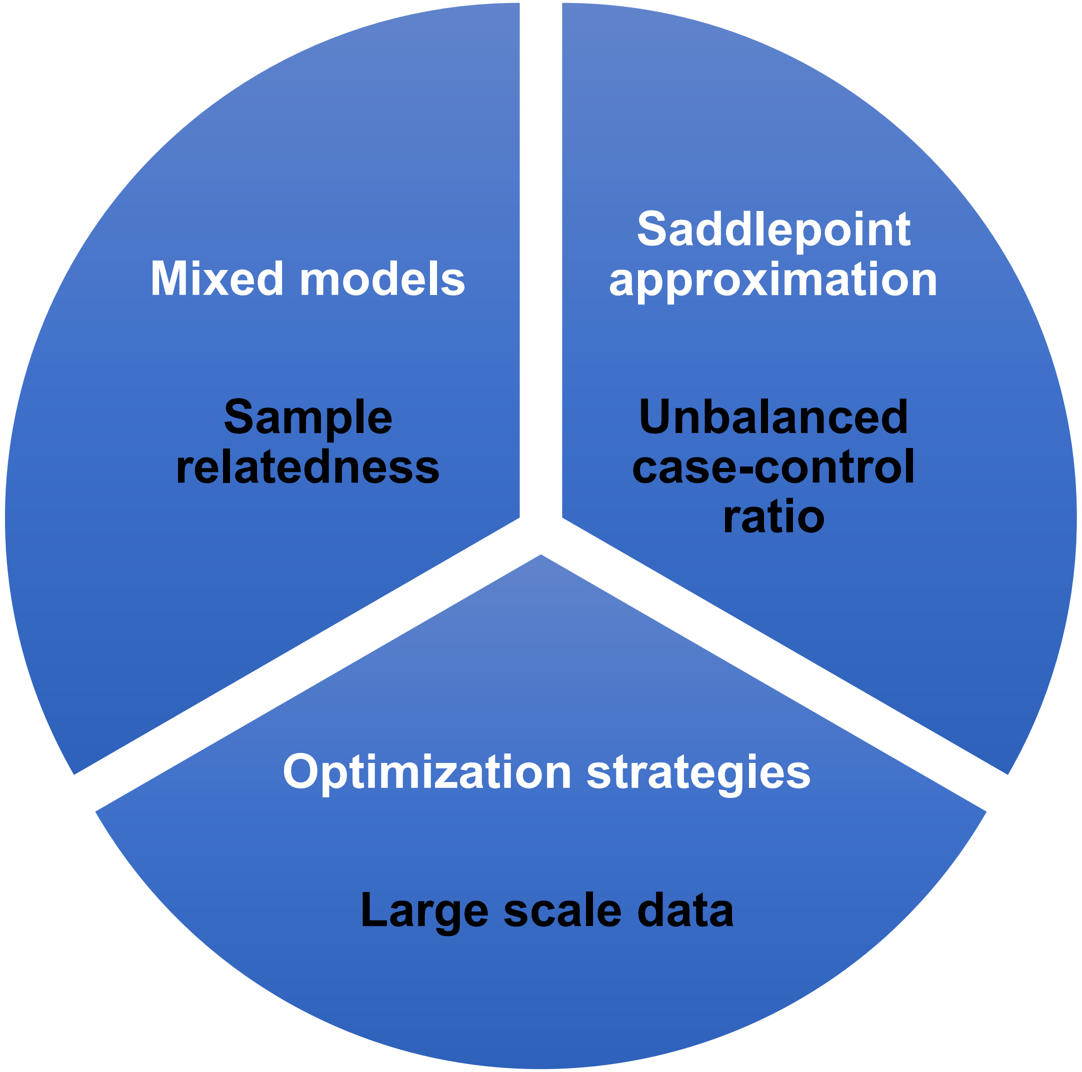
Q+K
J Yu *et al.*
2006

EMMA
HM Kang
et al.
2008

FaST-LMM
C Lippert
et al.
2011

GEMMA
X Zhou *et al.*
2012

Challenges of GWAS in large-scale cohorts/biobanks



Mixed models

Sample relatedness

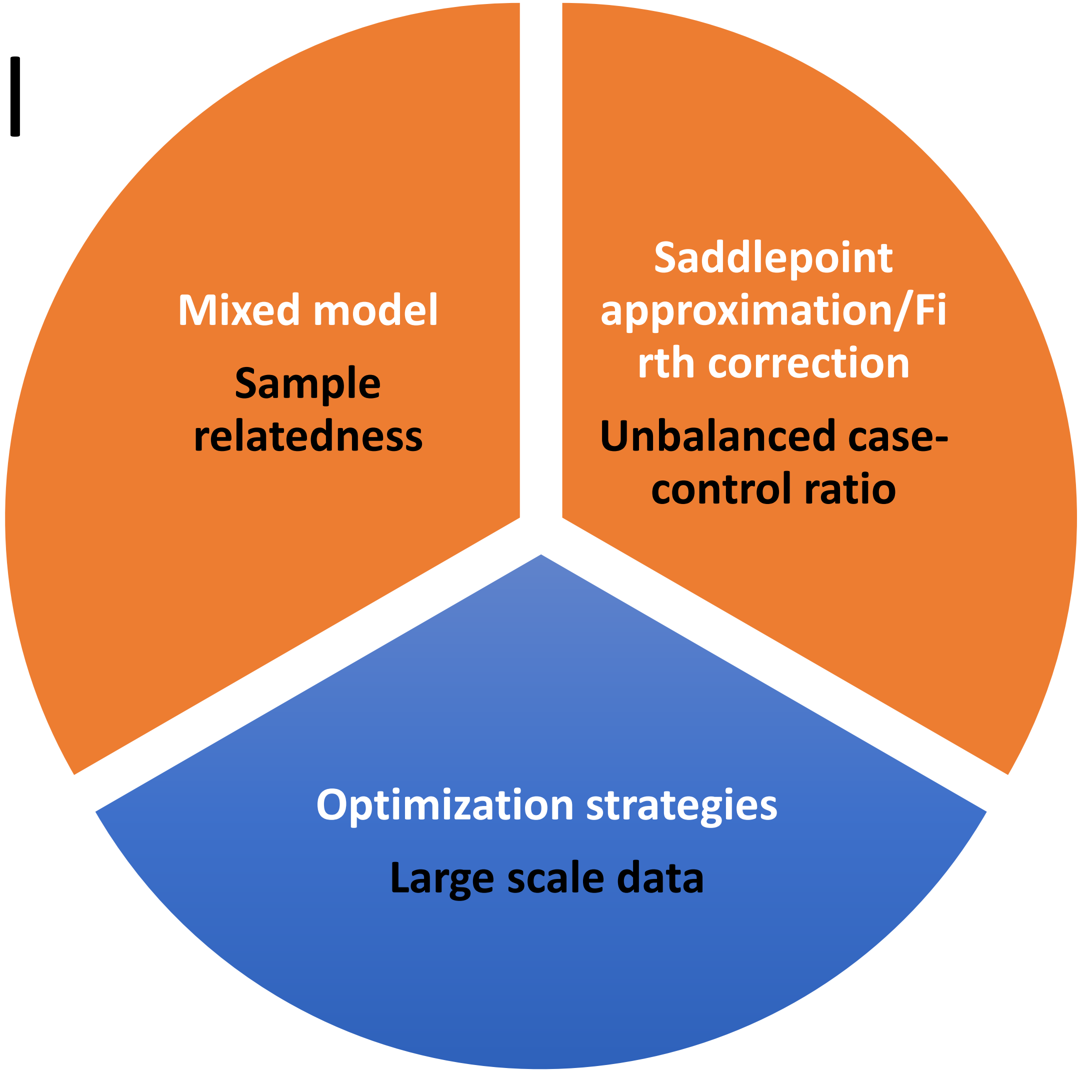
Saddlepoint approximation

Unbalanced case-control ratio

Optimization strategies

Large scale data

In today’s practical



Mixed model

Sample relatedness

Saddlepoint approximation/Firth correction

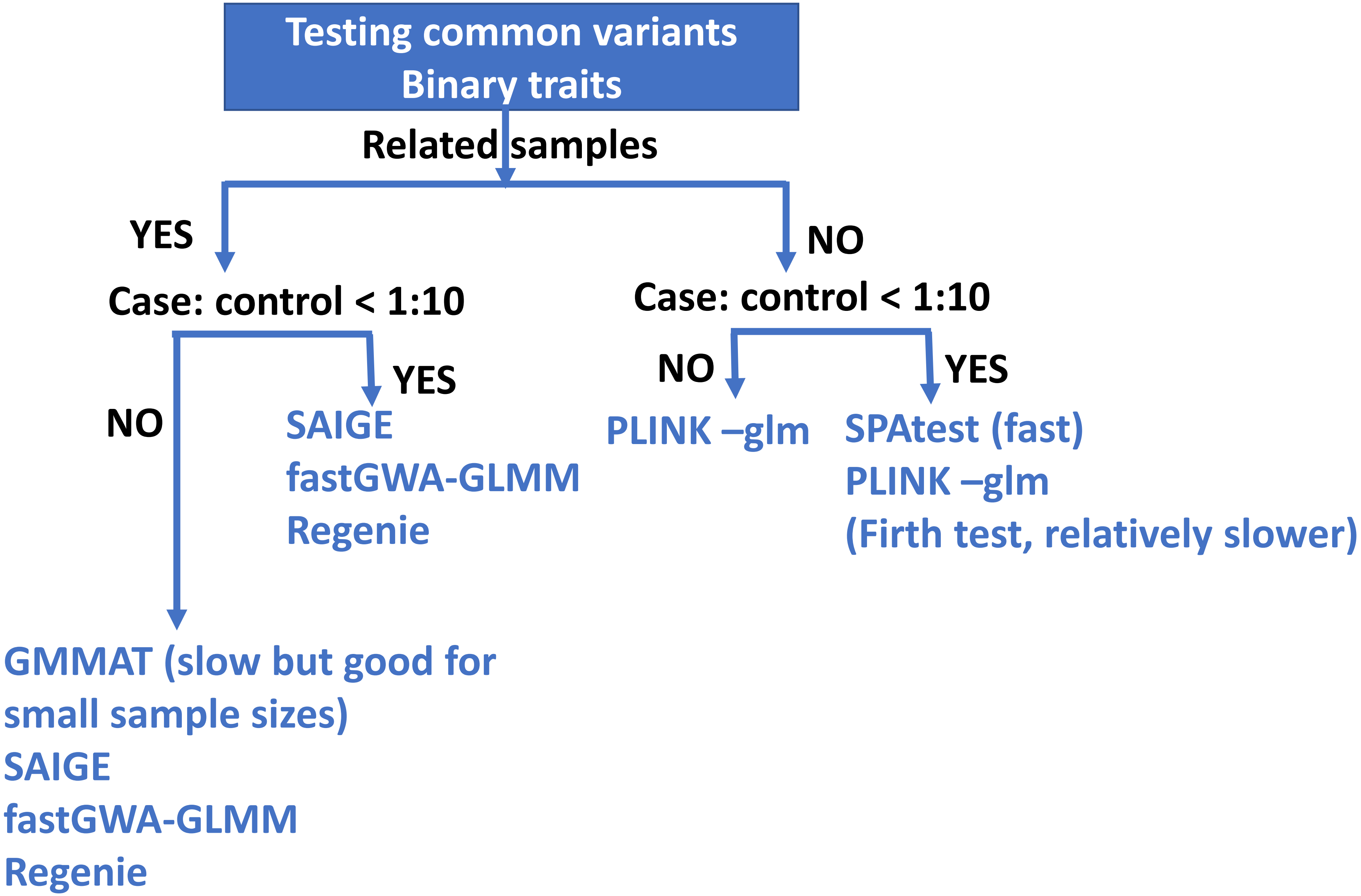
Unbalanced case-control ratio

Optimization strategies

Large scale data

Today’s practical is on Qualtrics:
https://qimr.az1.qualtrics.com/jfe/form/SV_036Ckv3AMwjqewu

Link is in Qualtrics.txt



Testing common variants
Binary traits

Related samples

YES

Case: control < 1:10

NO

GMMAT (slow but good for small sample sizes)
SAIGE
fastGWA-GLMM
Regenie

YES

SAIGE
fastGWA-GLMM
Regenie

NO

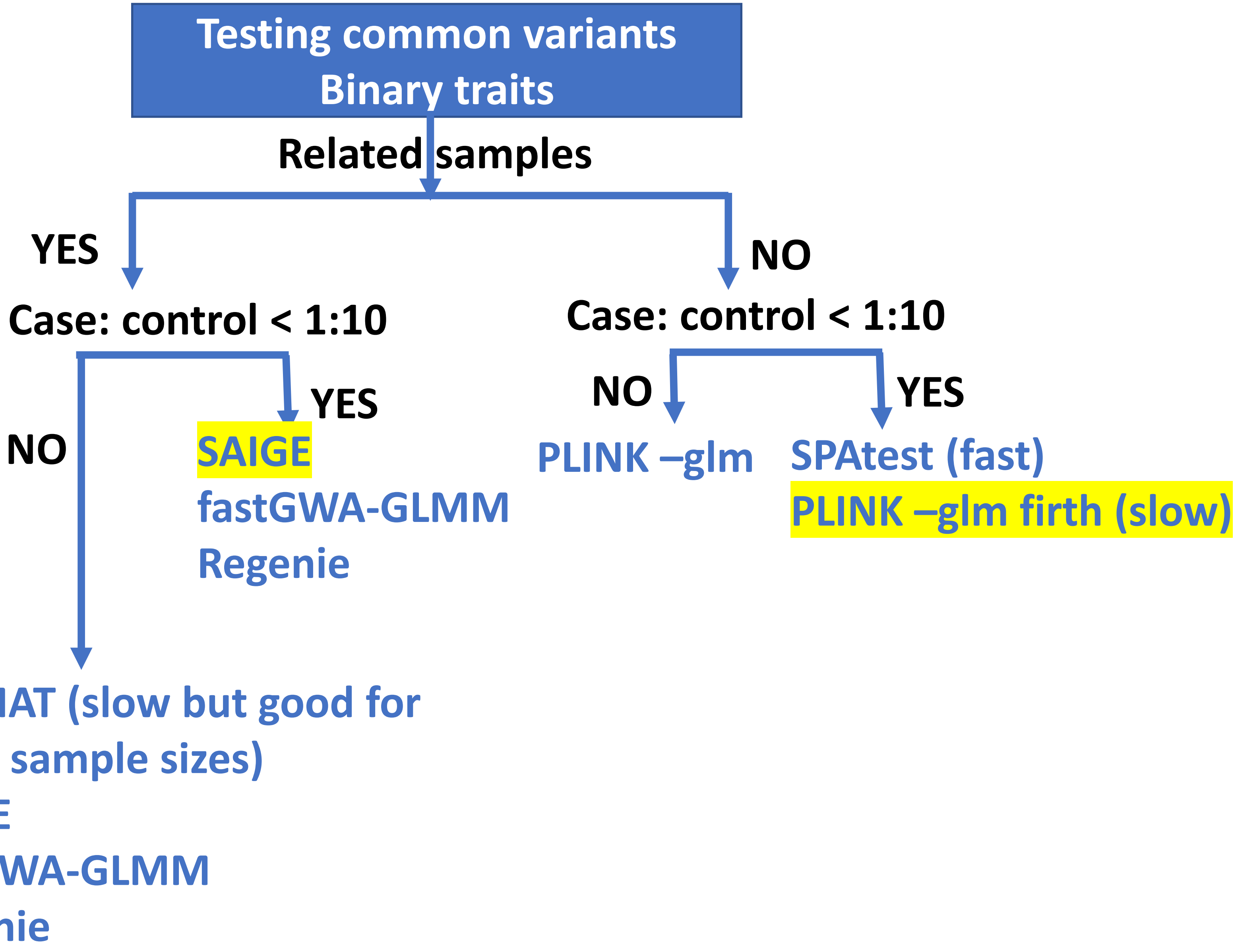
Case: control < 1:10

NO

PLINK –glm

YES

SPAtest (fast)
PLINK –glm
(Firth test, relatively slower)



Testing common variants
Binary traits

Related samples

YES

Case: control < 1:10

YES

SAIGE

fastGWA-GLMM

Regenie

NO

GMMAT (slow but good for small sample sizes)

SAIGE

fastGWA-GLMM

Regenie

NO

Case: control < 1:10

NO

PLINK –glm

YES

SPAtest (fast)

PLINK –glm firth (slow)

Reference

- 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.
- 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.

Backup slides

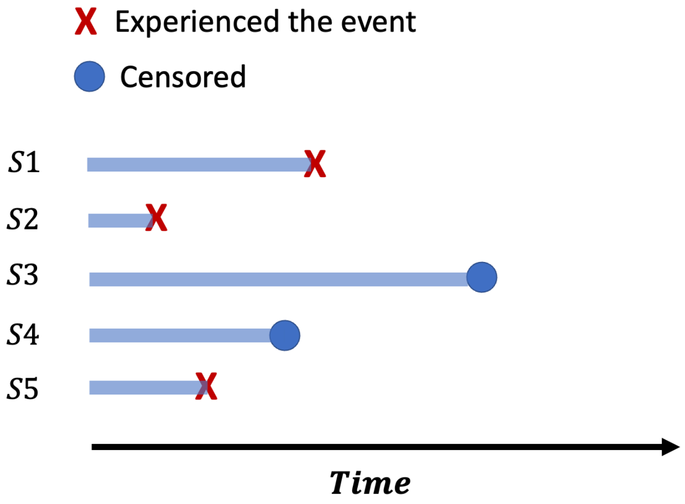
Different types of phenotypes require different statistical models for association tests

- 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 types of phenotypes

- Ordinal phenotypes
  - Common variants:
    - **POLMM**: **P**roportional **O**dds **L**ogistic **M**ixed **M**odel
    - Bi et al., *AJHG* 2021
  - Rare variants:
    - **POLMM-GENE**
    - Bi* and Zhou* *et al., AJHG* 2023

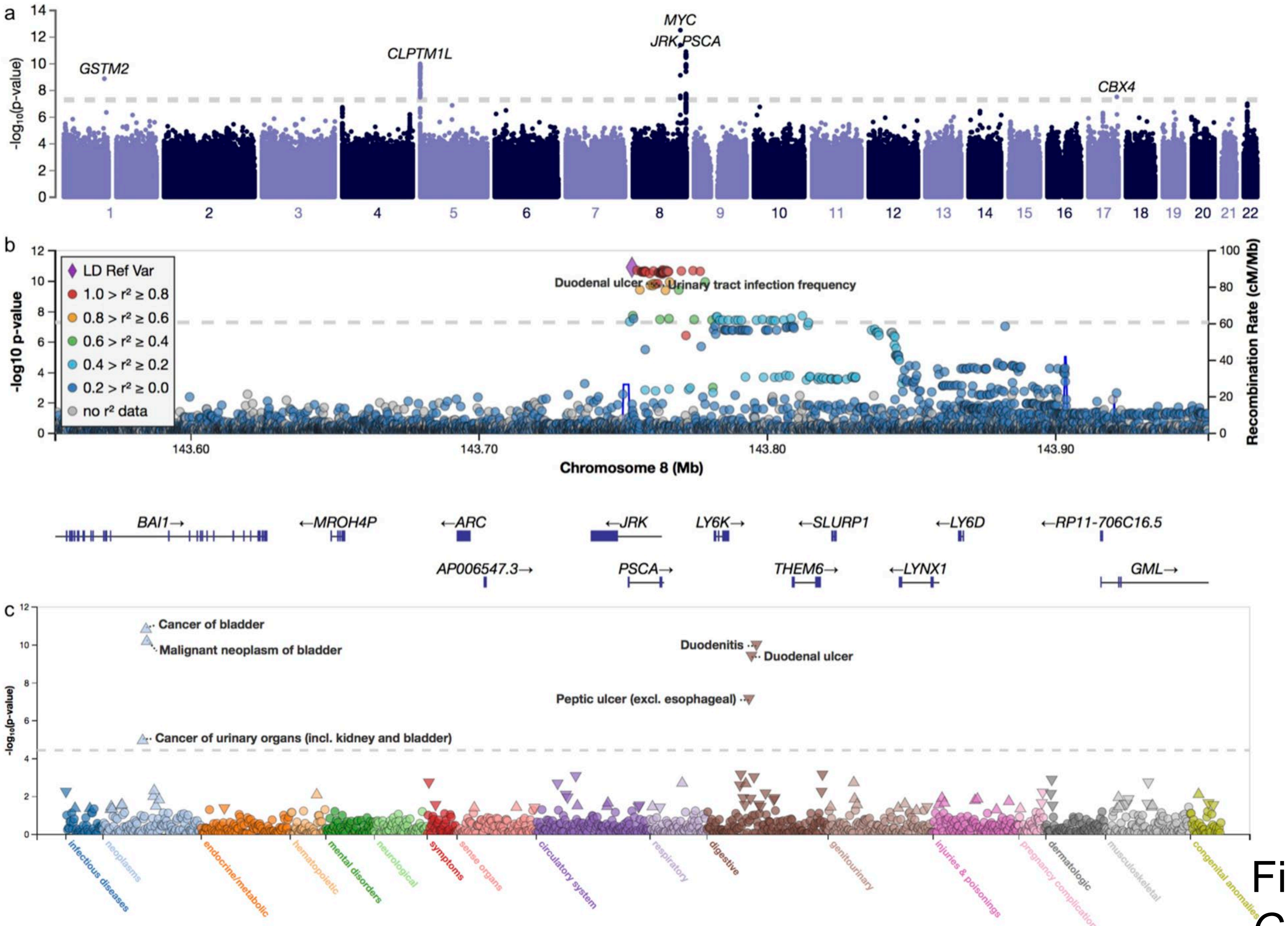
- Time-to-event phenotypes
  - Common variants:
    - **GATE**: **G**enetic **A**nalysis of **T**ime-to-**E**vent phenotypes
    - R library: https://github.com/weizhou0/GATE
    - Recently merged to SAIGE v1.4.4
    - Dey* and Zhou* *et al., Nature Comm* 2022: 5437.



Phenome-wide GWAS resources for large biobanks

- UK Biobank
    - HRC imputed and TopMed imputed
        - https://pheweb.sph.umich.edu/
        - 1403 binary phenotypes based on Phecodes in White British samples
        - PheWeb: Taliun *et al., Nat Genet*. 2020. https://github.com/statgen/pheweb

Interactive views of genetic associations in the UK Biobank instance of PheWeb



Manhattan Plot - GWAS

LocusZoom Plot
(Pruim et al., 2010)

Figure 1 from Taliun *et al.*, *Nat Genet*. 2020

Phenome-wide GWAS resources for large biobanks

- UK Biobank
  - HRC imputed and TopMed imputed
    - https://pheweb.sph.umich.edu/
    - 1403 binary phenotypes based on Phecodes in White British samples
    - PheWeb: Taliun *et al., Nat Genet*. 2020. https://github.com/statgen/pheweb
  - Neale lab – round 1 and 2
    - https://www.nealelab.is/uk-biobank
    - Sex stratified PheGWAS
  - Pan-UKBB:
    - https://pan.ukbb.broadinstitute.org/
    - 7,228 quantitative and binary phenotypes, across 6 continental ancestry groups, for a total of 16,131 GWAS
    - All variants with INFO > 0.8, and MAC > 20 in that population (up to 28 million variants)
  - Genebass:
    - https://app.genebass.org/
    - 4,529 quantitative and binary phenotypes on 394,841 exomes
    - 57,650 group tests per phenotype (pLoF, missense, synonymous for each gene)

Phenome-wide GWAS resources for large biobanks

- FinnGen project (412,181 individuals, 2,408 endpoints): https://r10.finngen.fi/
- Michigan Genomic Initiative (80,381 individuals, 1,728 endpoints): https://pheweb.org/MGI/
- Biobank Japan: https://pheweb.jp/
- Taiwan Biobank: https://taiwanview.twbiobank.org.tw/pheweb.php
- Million Veteran Program (MVP):
    - Verma et al.,*Science*, 2024

Biobank meta-analysis results

- Global Biobank Meta-analysis Flagship project (24 biobanks with up to 2.2 million individuals for 14 exemplary disease endpoints)
    - http://results.globalbiobankmeta.org/

- FinnGen + UKBB + MVP (1.5 million individuals for ~300 binary phenotypes)
    - https://mvp-ukbb.finngen.fi/

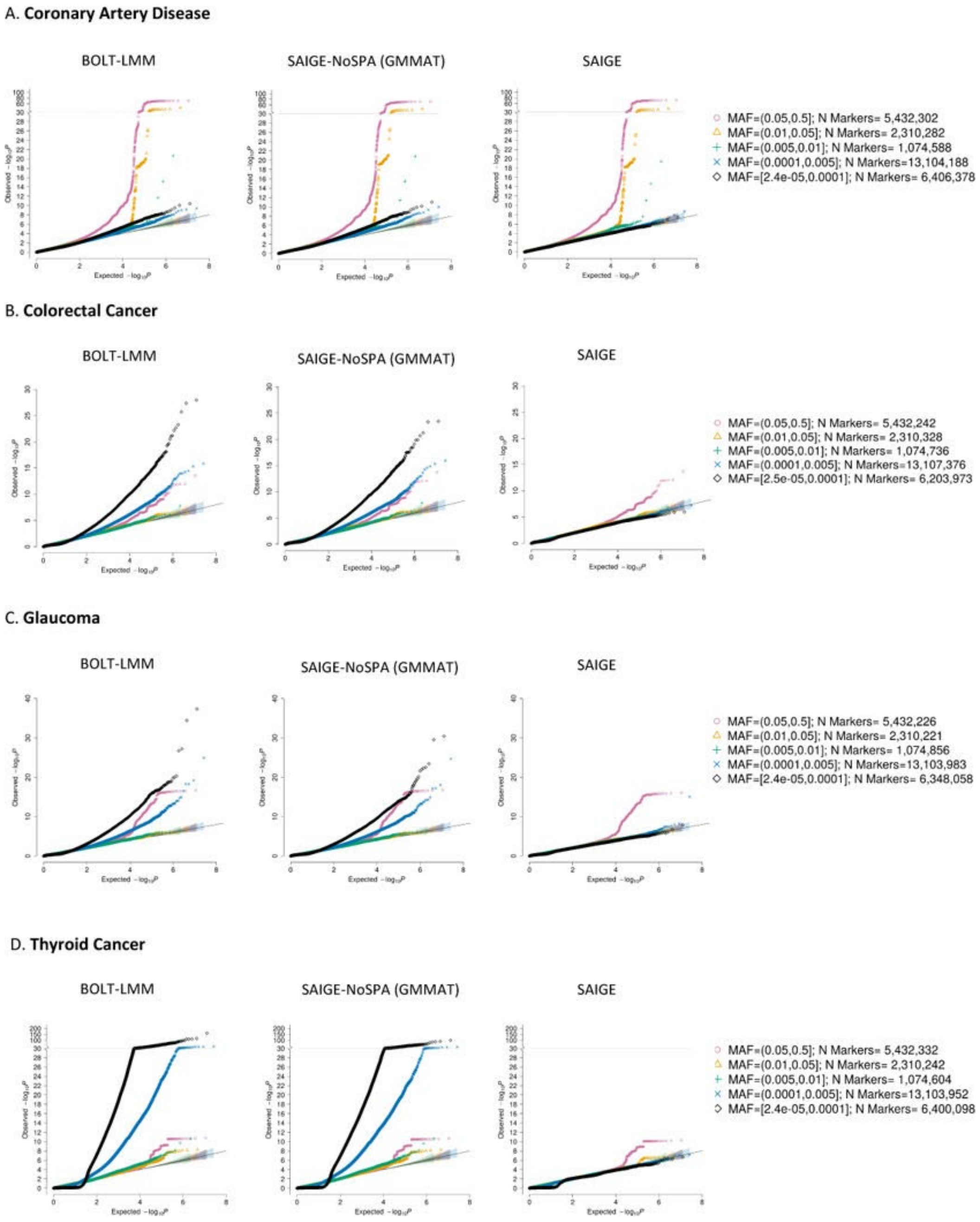


Figure 2.

Quantile-quantile plots of GWAS results for four binary phenotypes with various case-control ratios in the UK Biobank.

Zhou et al., 2018