GWAS in large-scale biobanks and cohorts

Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)

Wei Zhou

Post-doctoral Fellow

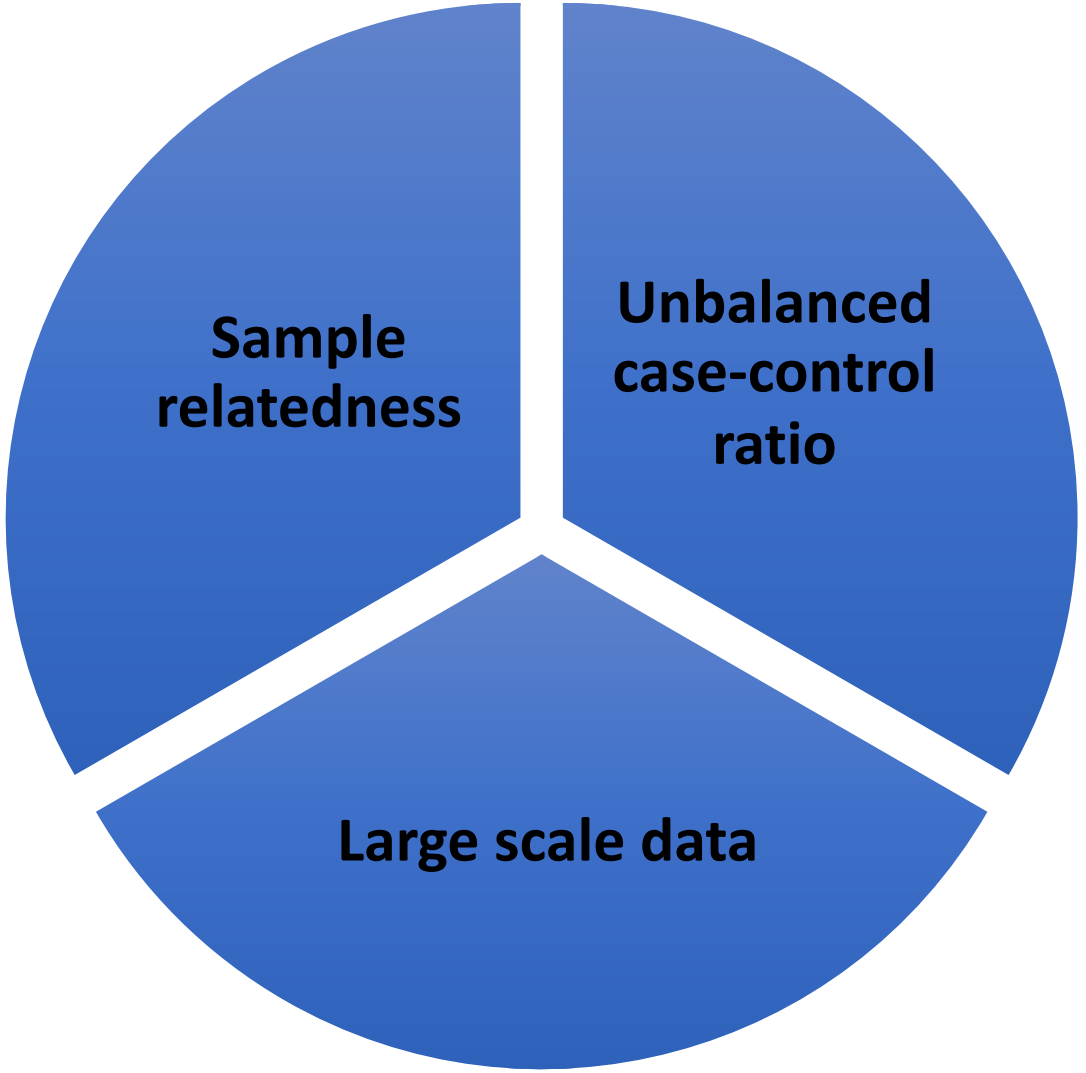
Massachusetts General Hospital, Harvard Medical School, Broad Institute

Outline

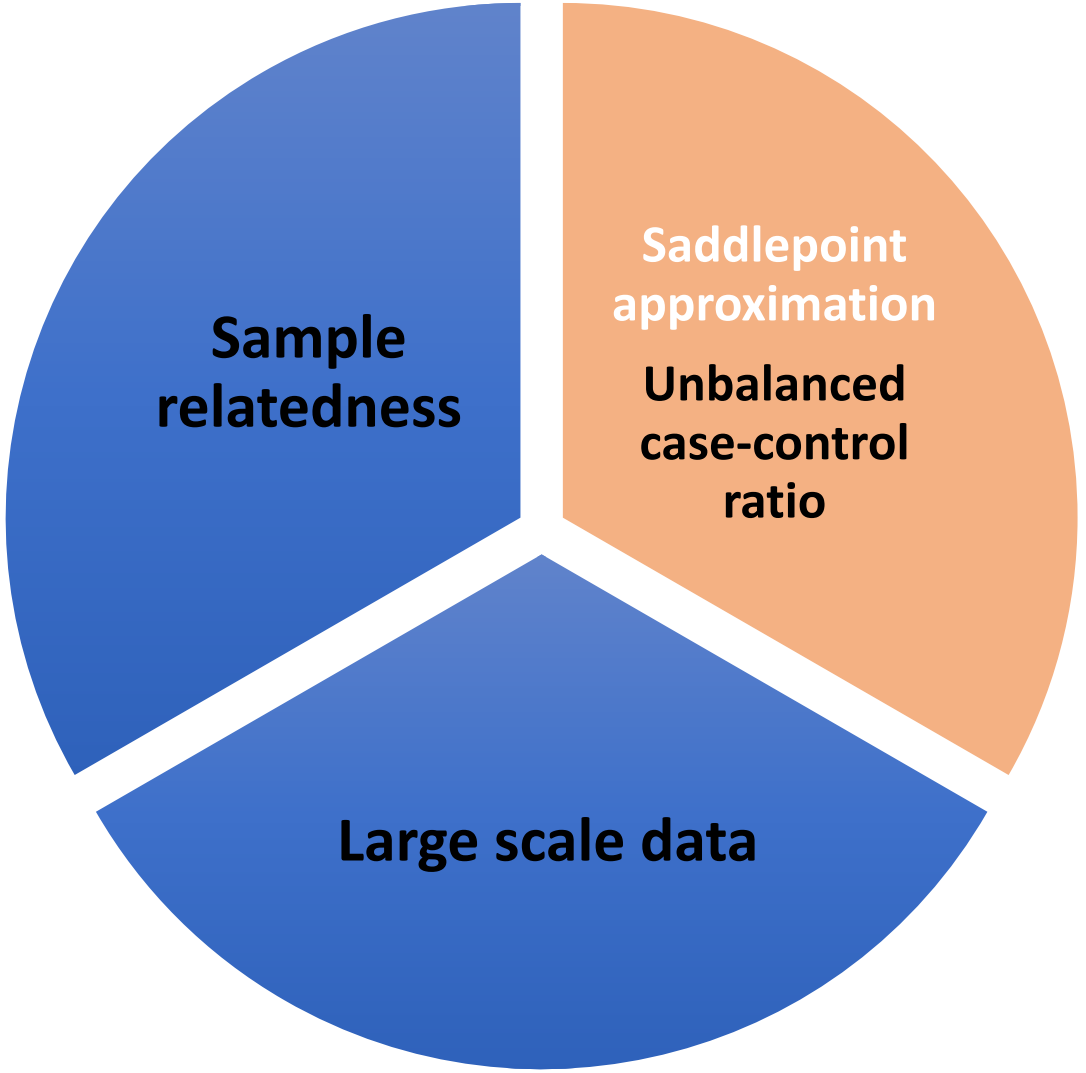
- Challenges of GWAS in large-scale cohorts/biobanks (mostly for binary phenotypes)
  - Mixed models to account for sample relatedness in GWAS

- Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)

Challenges of GWAS in large-scale cohorts/biobanks

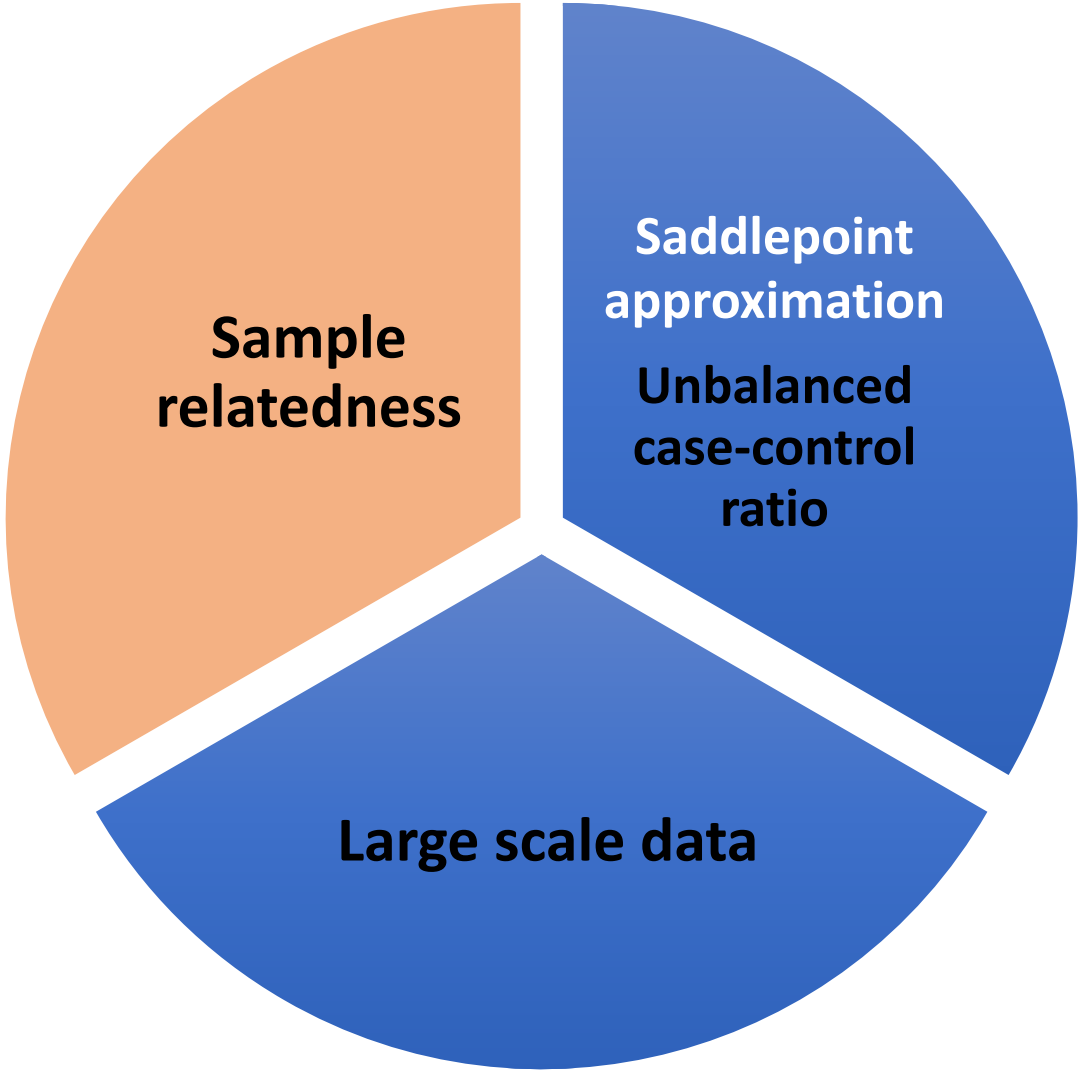


Challenges of GWAS in large-scale cohorts/biobanks



Dey *et al.* 2017

Challenges of GWAS in large-scale cohorts/biobanks



What if individuals are inter-related?

- Linear and Logistic regression models assume individuals are unrelated.
- Known and unknown family relatives can be included in the GWAS studies

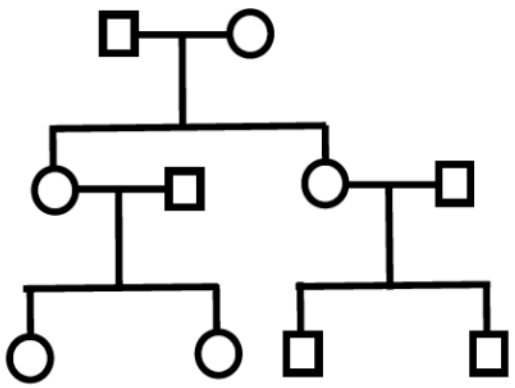
In the UK Biobank data, almost one-third of the individuals have a third degree (e.g., first cousin) or closer relative in the cohort (Bycroft et al, 2017)

What if individuals are inter-related?

- To accommodate this in GWASs,
  - First, we need to quantify unknown relatedness.
  - Second, we need to account for the relatedness in the association tests

Genetic Relatedness Matrix (GRM): Quantifying relatedness

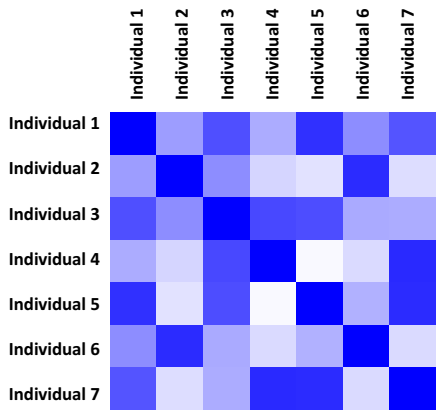
- First, let's look at the case when the pedigree is known.



$$\begin{bmatrix} 1 & 0 & 0.5 & 0 & 0.5 & 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 1 & 0.5 & 0 & 0.5 & 0 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.5 & 1 & 0 & 0.5 & 0 & 0.5 & 0.5 & 0.25 & 0.25 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0.5 & 0 & 1 & 0 & 0.25 & 0.25 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.5 & 0.5 & 0.25 & 0 & 1 & 0.5 & 0.125 & 0.125 \\ 0.25 & 0.25 & 0.5 & 0.5 & 0.25 & 0 & 0.5 & 1 & 0.125 & 0.125 \\ 0.25 & 0.25 & 0.25 & 0 & 0.5 & 0.5 & 0.125 & 0.125 & 1 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0 & 0.5 & 0.5 & 0.125 & 0.125 & 0.5 & 1 \end{bmatrix}$$

Genetic Relatedness Matrix (GRM): Quantifying relatedness

- When the pedigrees are unknown, we approximate the relatedness.
- Let G be the $n \times p$ genotype matrix (centered and appropriately scaled)
- Then, the empirical GRM is,
  - $\widehat{\Psi} = \frac{1}{p} G G^{\top}$



Recall the linear regression model:

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

Assume Y_is are independent given X_i, G_i.

Y_i: phenotype vector for the *ith* individual
X_i: covariates matrix for the *ith* individual
G_i: genotype vector for the *ith* individual

Recall the linear regression model:

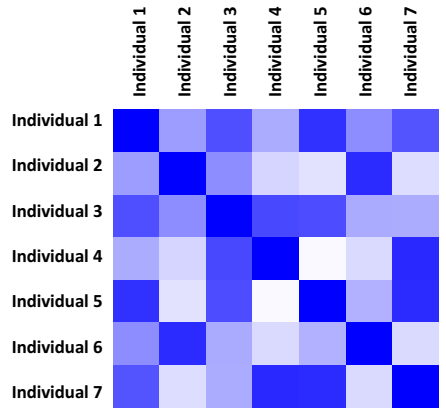
$$\boldsymbol{Y_i = X_i\alpha + G_i\beta + \epsilon_i}$$

Assume Y_is are independent given X_i, G_i.

Linear mixed model:

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

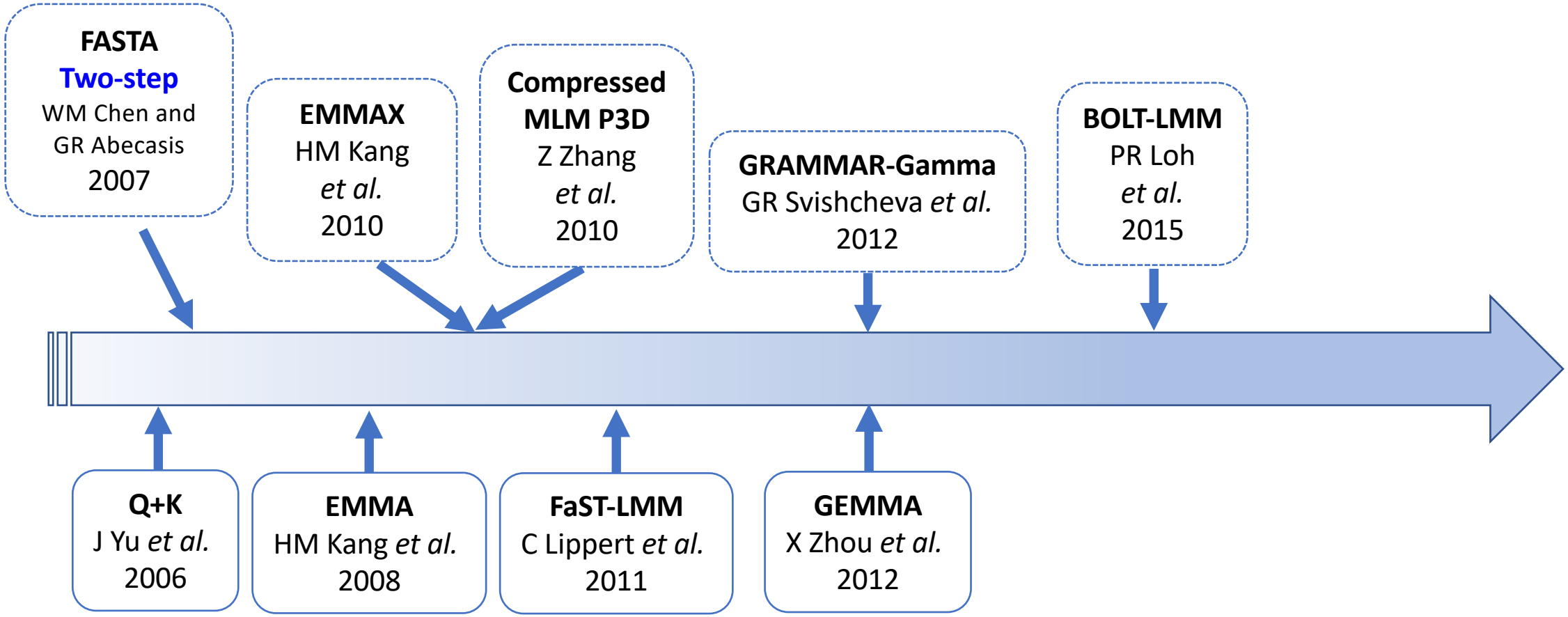
Assume Y_is are independent given X_i, G_i, and b_i



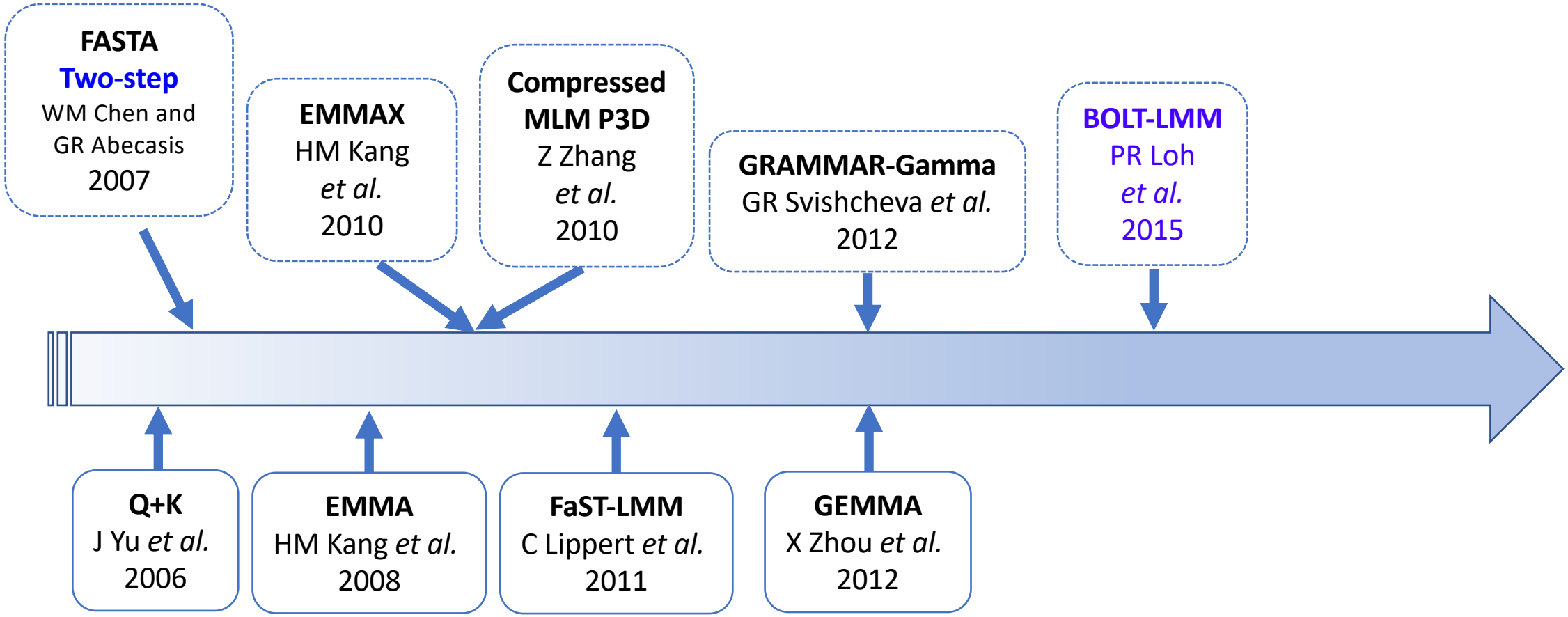
Accounting for sample relatedness

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

Linear mixed model methods for GWAS

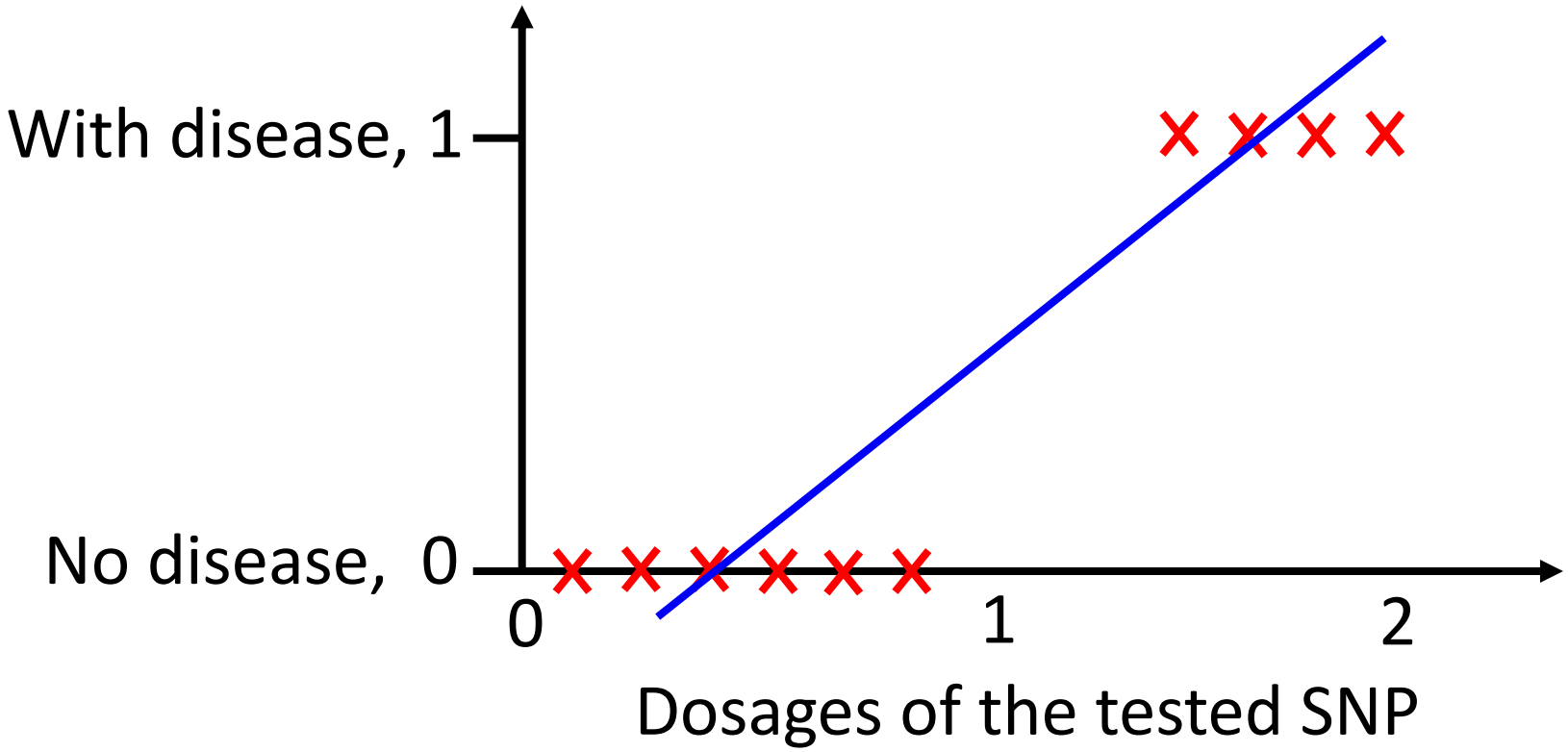


BOLT-LMM: first linear mixed model method for GWAS in biobank-scale data

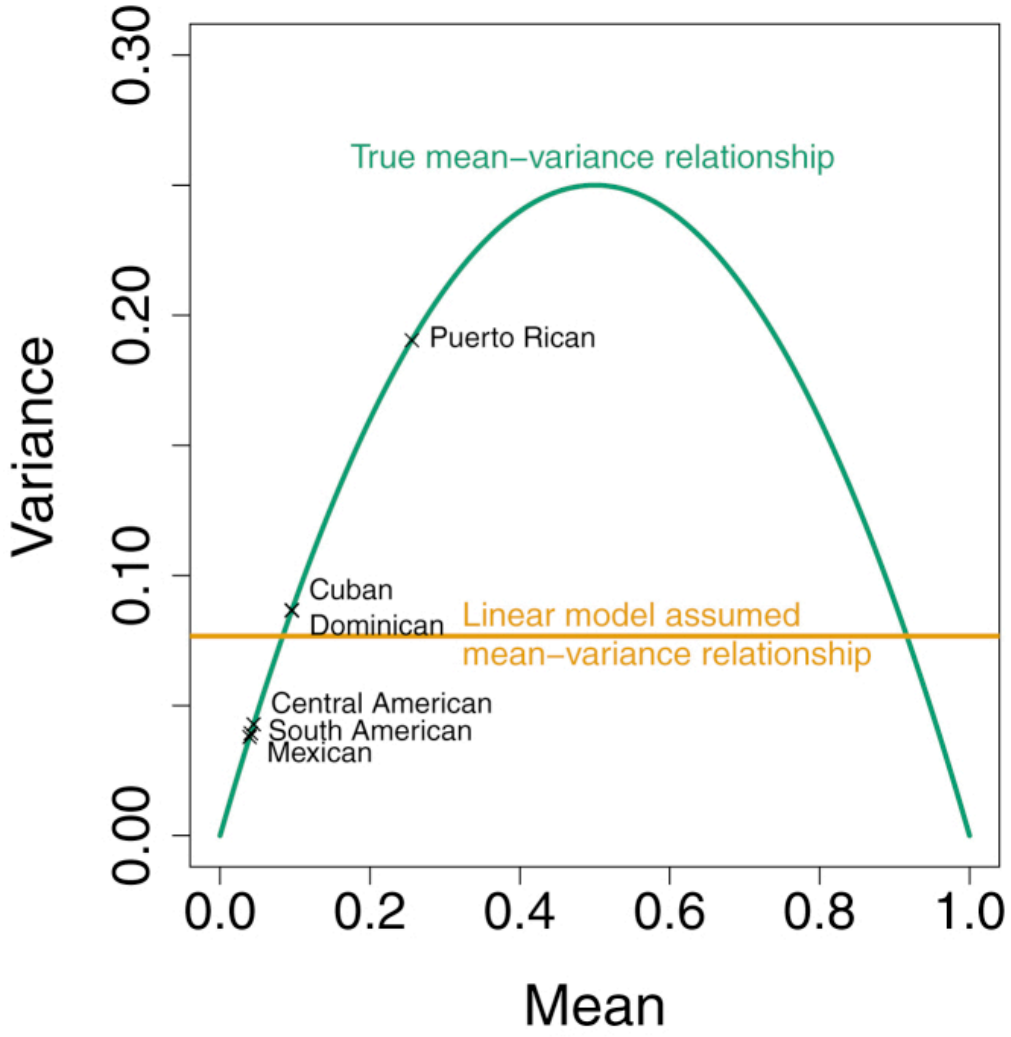


Linear Mixed Model for Binary Phenotypes?

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



Linear Mixed Model for Binary Phenotypes?



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

Logistic mixed model:

$$\mu_i = \Pr(Y_i = 1 | X_i, G_i, \boldsymbol{b_i})$$

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

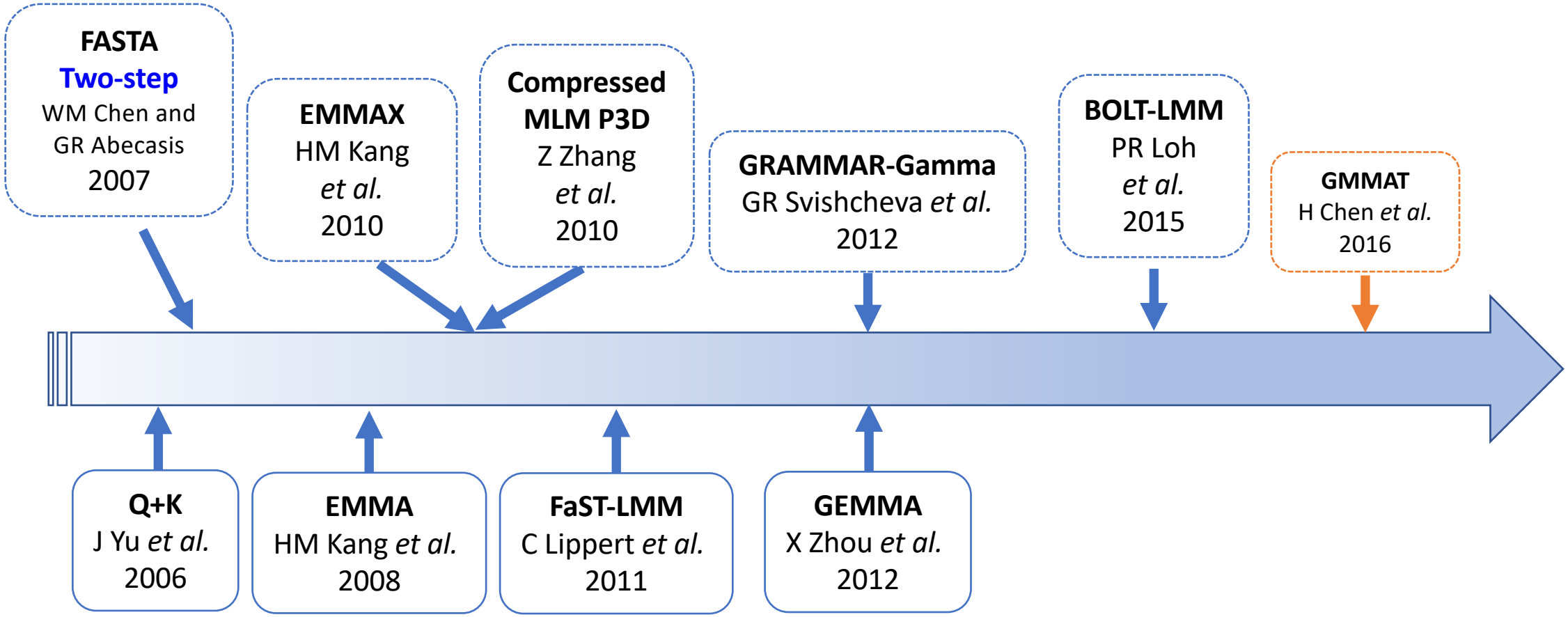
Linear mixed model:

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

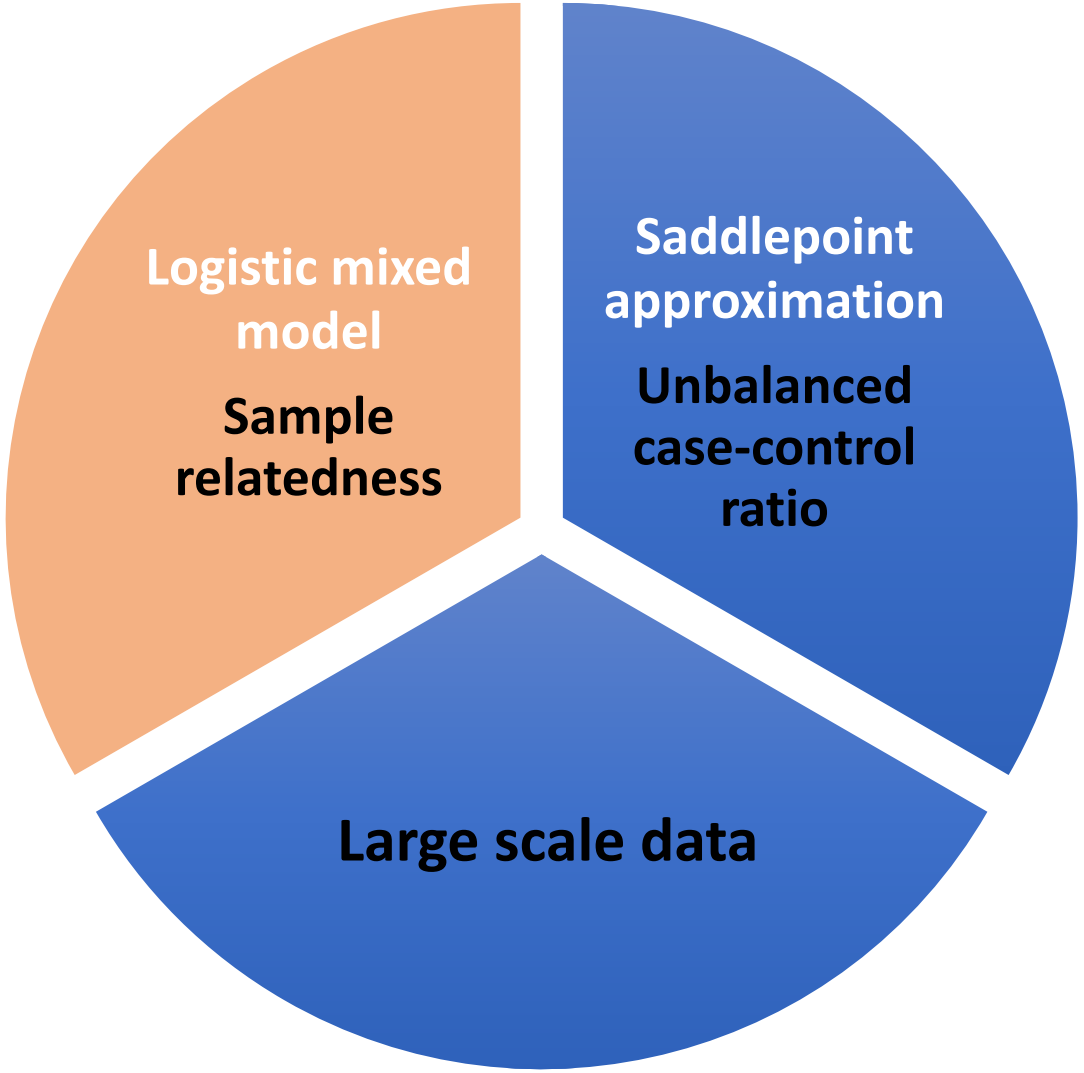
Accounting for sample relatedness

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

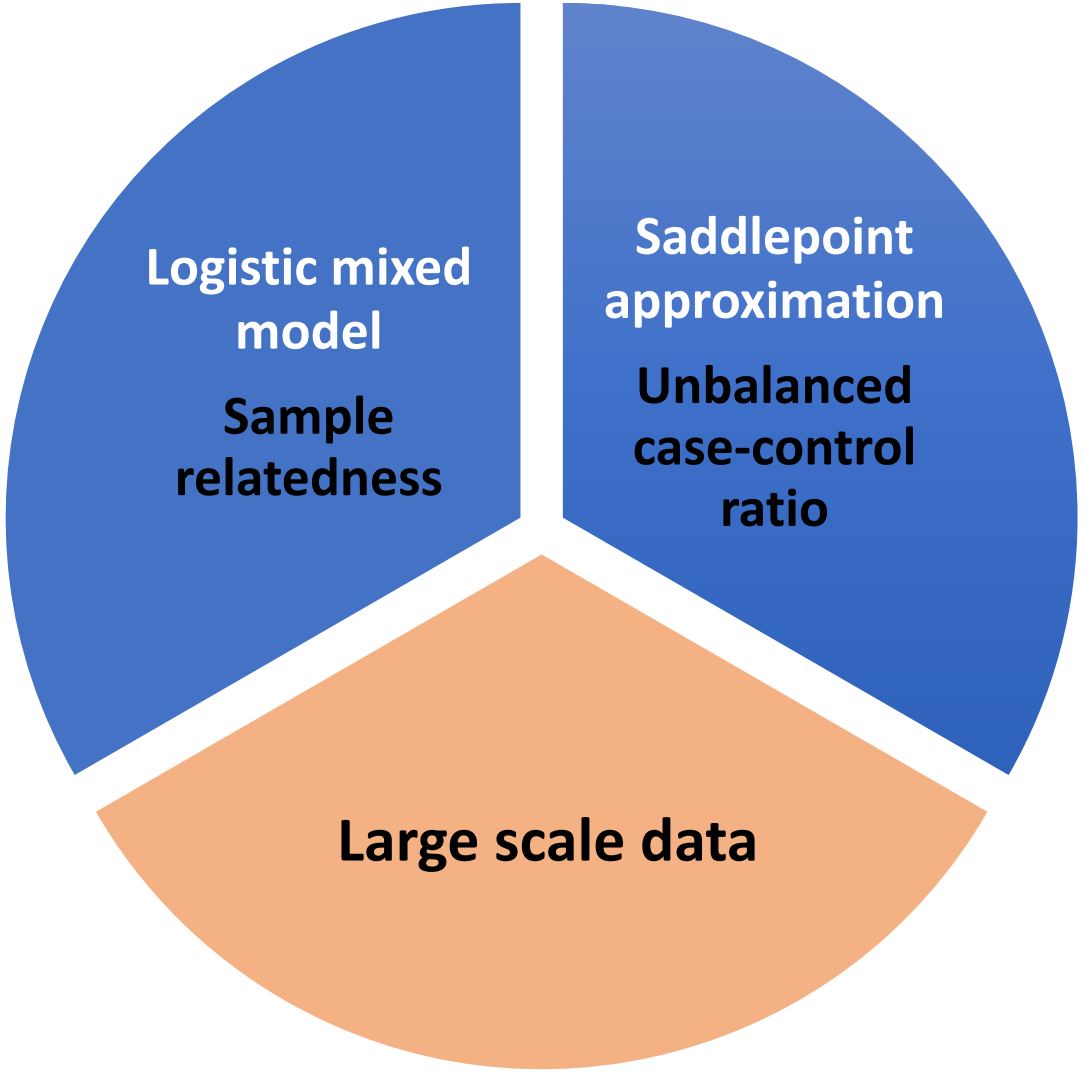
GMMAT: Generalized linear Mixed Model Association Test



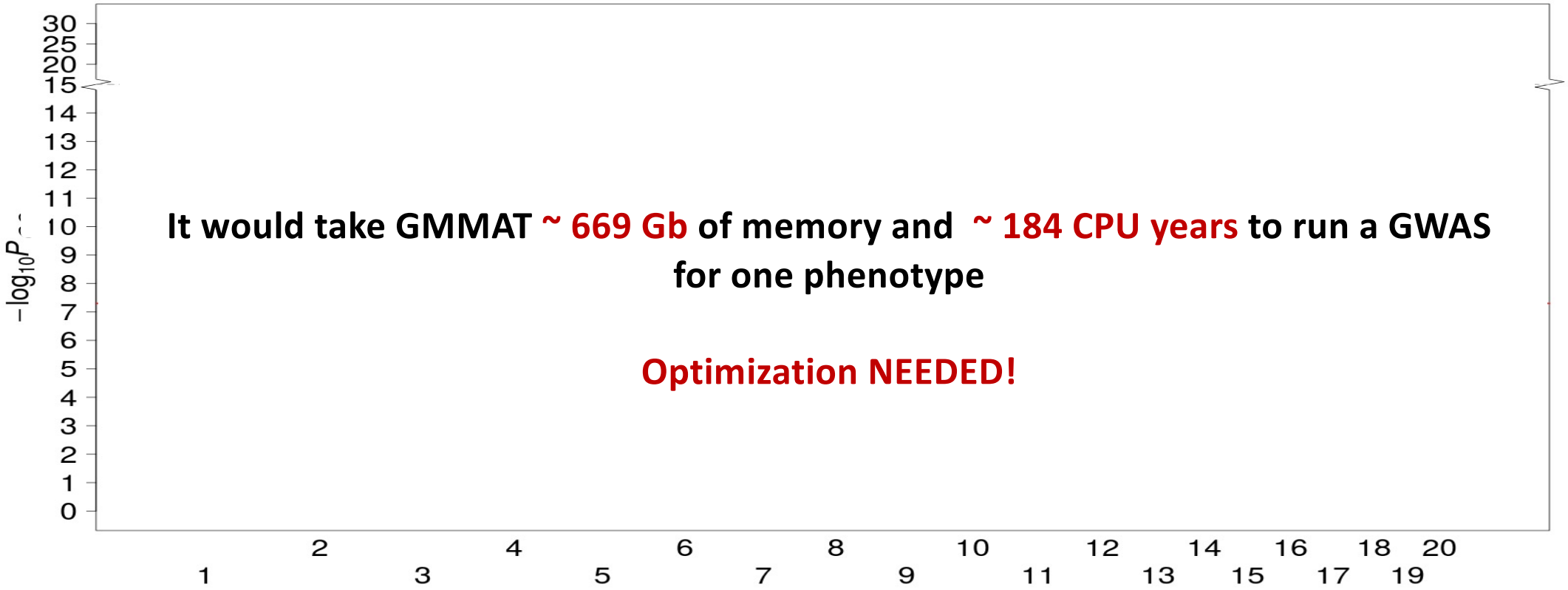
Challenges of GWAS in large-scale cohorts/biobanks



Challenges of GWAS in large-scale cohorts/biobanks



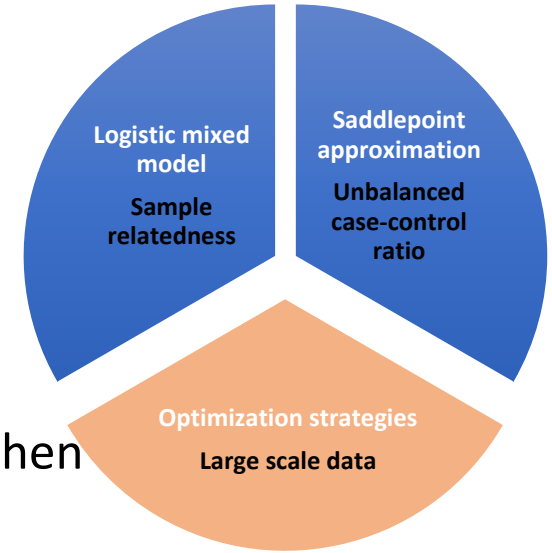
Binary Traits	N_{Case}	$N_{Control}$
Colorectal cancer	4,562	382,756



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

Optimization NEEDED!

Strategies to make the algorithm computationally practical for large data sets



Reduce memory usage

- Store raw genotypes in a binary vector to compute GRM ($\boldsymbol{\psi}$) elements when needed

❖ $\mathbf{N\times(N+1)\times 4}$ **to** $\frac{\mathbf{N}M_1}{\mathbf{4}}$

❖ In the example of UK Biobank: N = 408,961 and M_1 = 93,511, memory usage drops from **669Gb to 9.56Gb**

Reduce Computation time

- Using pre-conditioned conjugate gradient to calculate the product of $\boldsymbol{\Sigma^{-1}\mathbf{b}}$ by iteratively solving the linear system $\boldsymbol{\Sigma x = b}$
- Hutchinson's randomized trace estimator is used to estimate the traces of matrix $P\psi$ (M. F. Hutchinson, 1989)

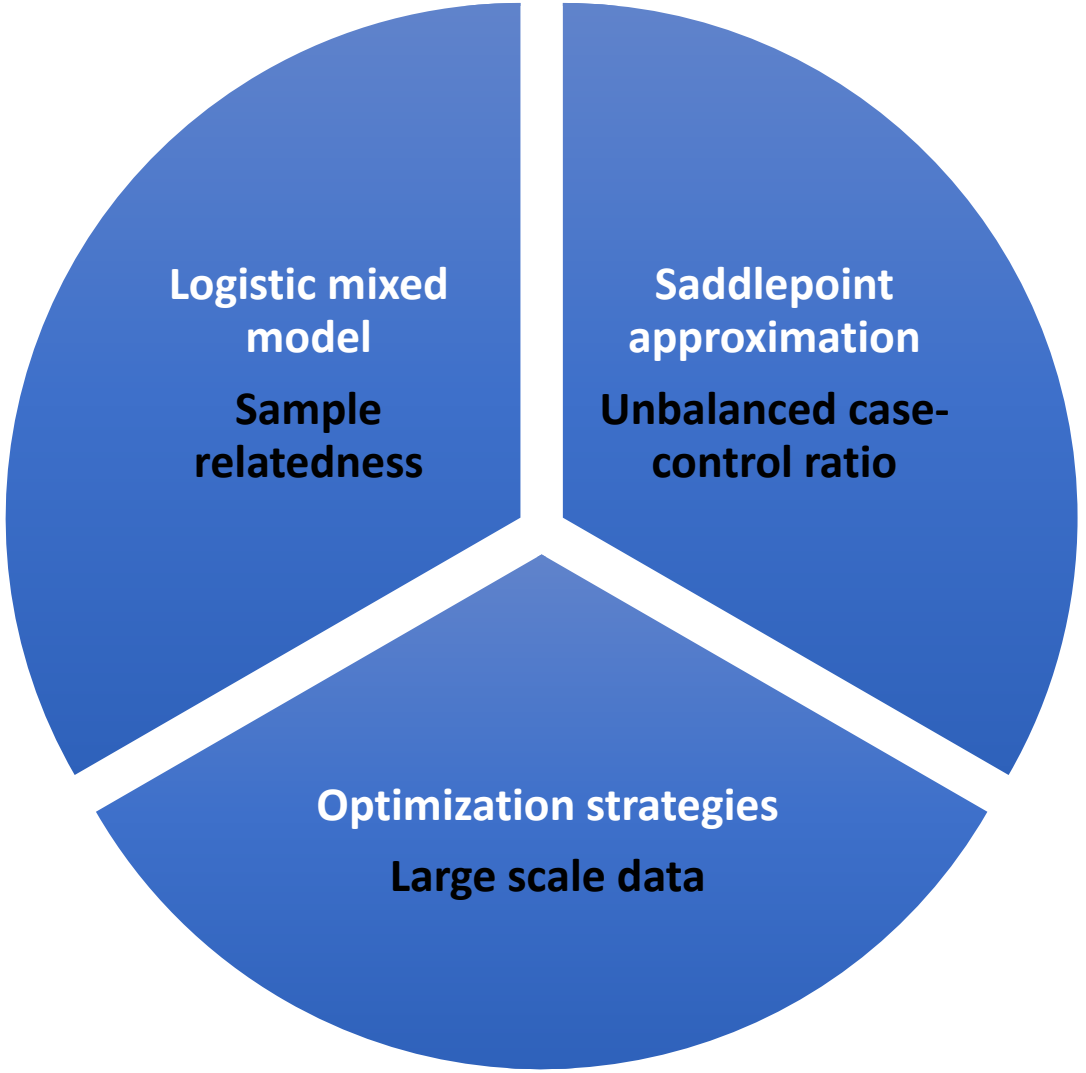
❖ **$O(N^3)$ to $O(M_1N^{1.5})$**

$$S(\tau) = \frac{\partial q l_R(\hat{\alpha}(\phi,\tau),\beta=0,\phi,\tau)}{\partial \tau} = \frac{1}{2}(\tilde{Y}^T P\psi P\tilde{Y} - tr(P\psi))$$

N: number of samples

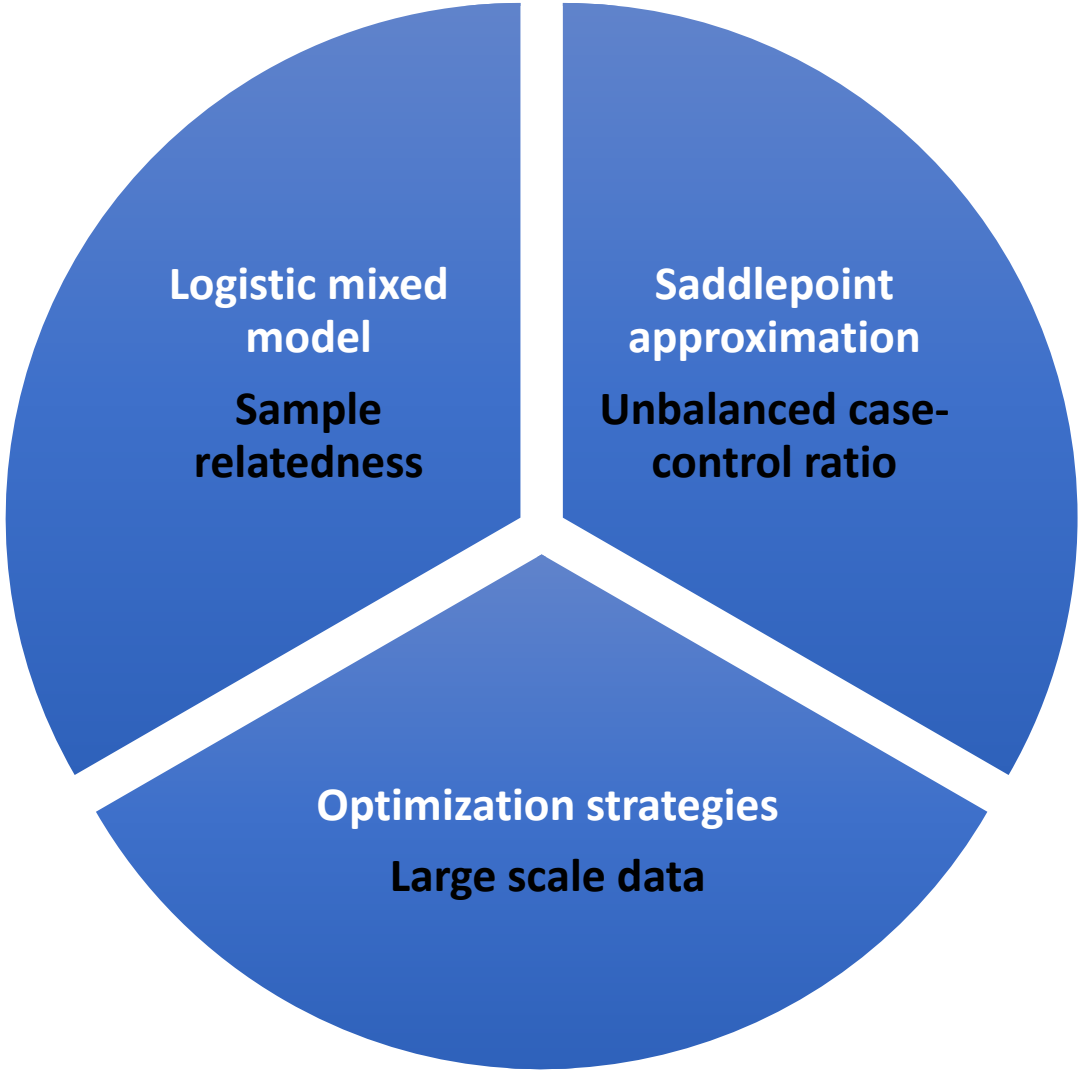
M_1: number of genetic markers used to construct the genetic relationship matrix

Challenges and Solutions of GWAS in large-scale cohorts/biobanks



SAIGE

Scalable and Accurate Implementation of GEneralized mixed model



Zhou et al., 2018

31,355 cases
377,103 controls
1:12

4,562 cases
382,756 controls
1:84

4,462 cases
397,761 controls
1:89

358 cases
407,399 controls
1:1138

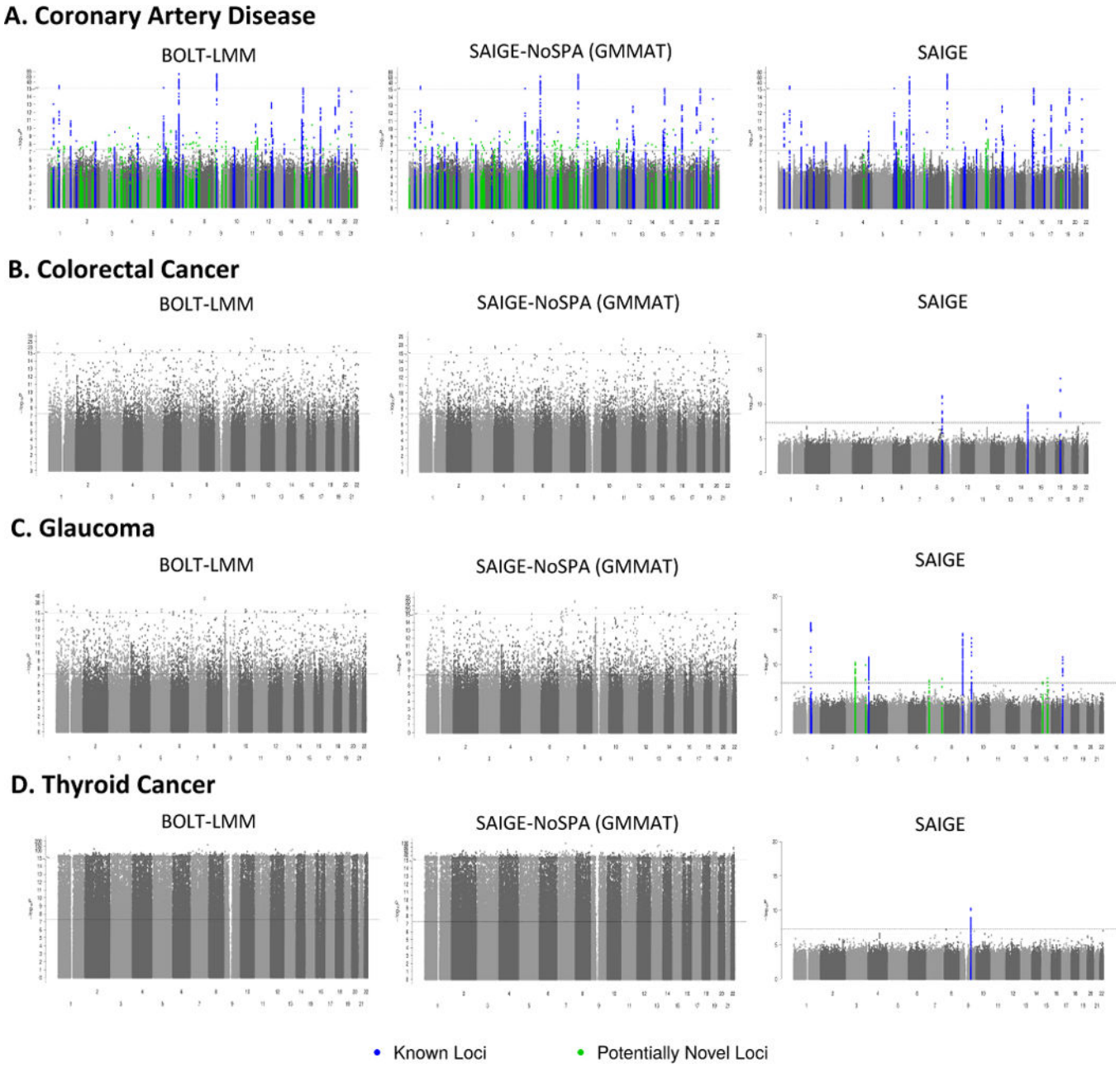


Figure 1.
Manhattan plots of GWAS results for four binary phenotypes with various case-control ratios in the UK Biobank.

Zhou et al., 2018

31,355 cases
377,103 controls
1:12

4,562 cases
382,756 controls
1:84

4,462 cases
397,761 controls
1:89

358 cases
407,399 controls
1:1138

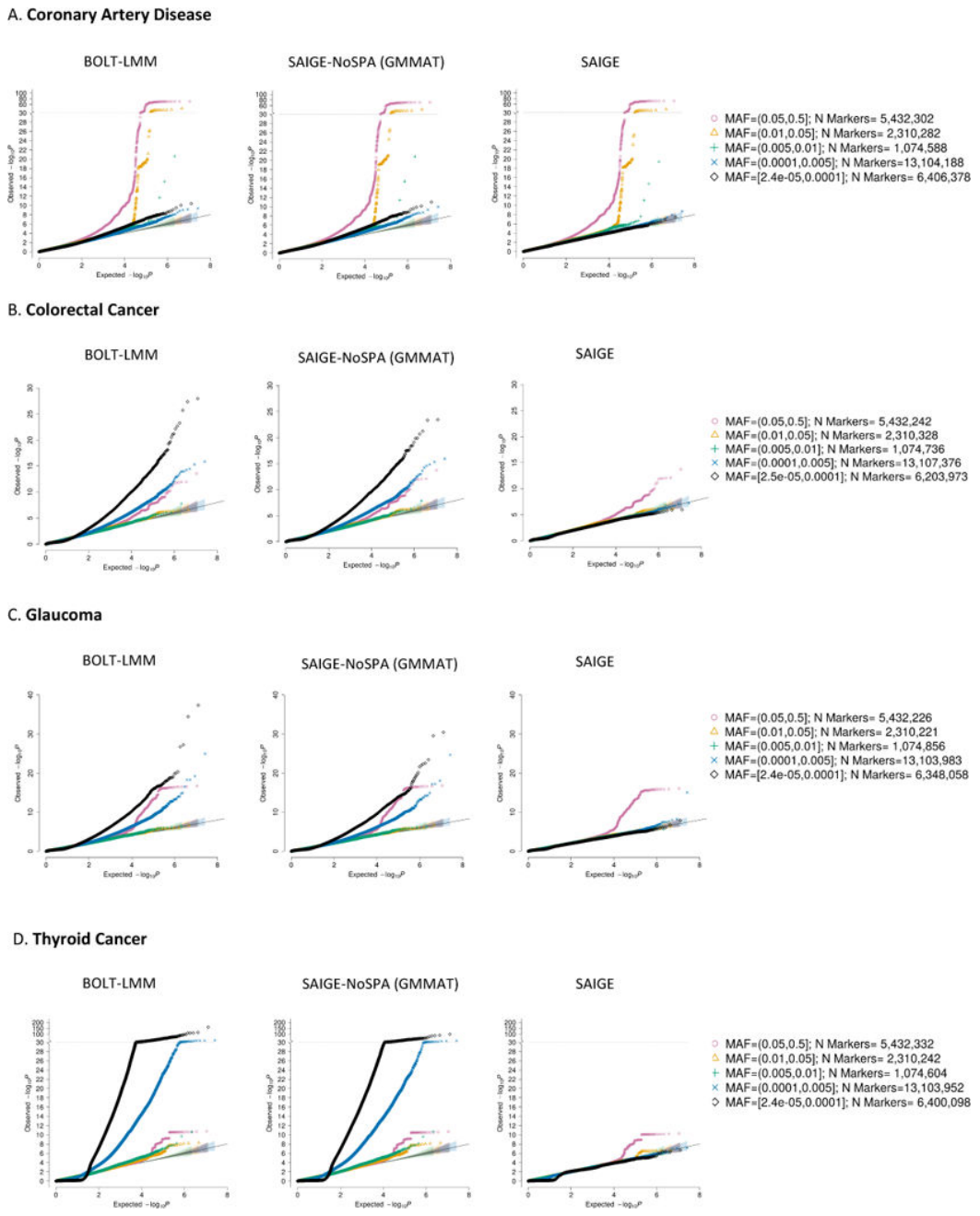
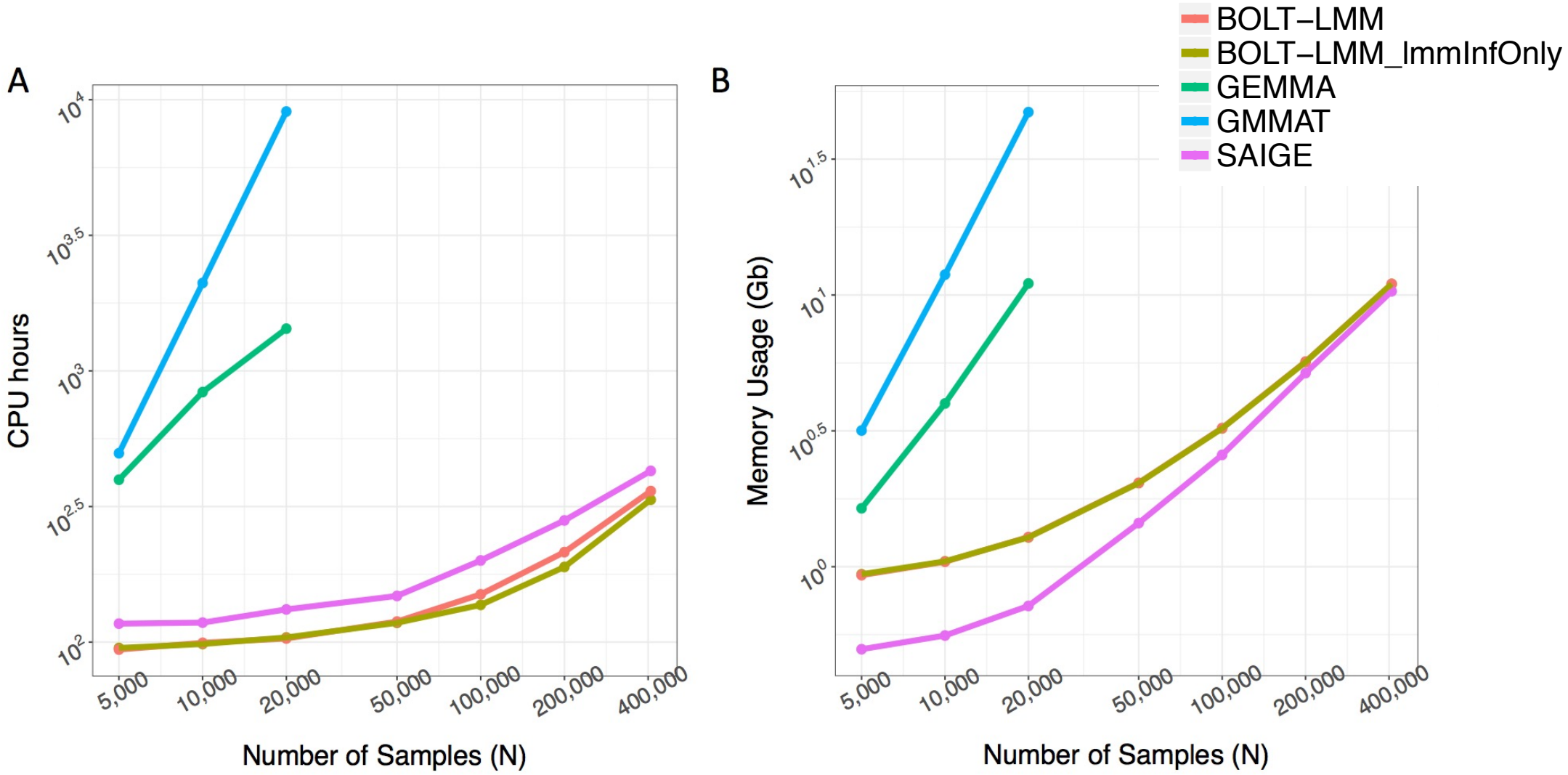


Figure 2.
Quantile-quantile plots of GWAS results for four binary phenotypes with various case-control ratios in the UK Biobank.

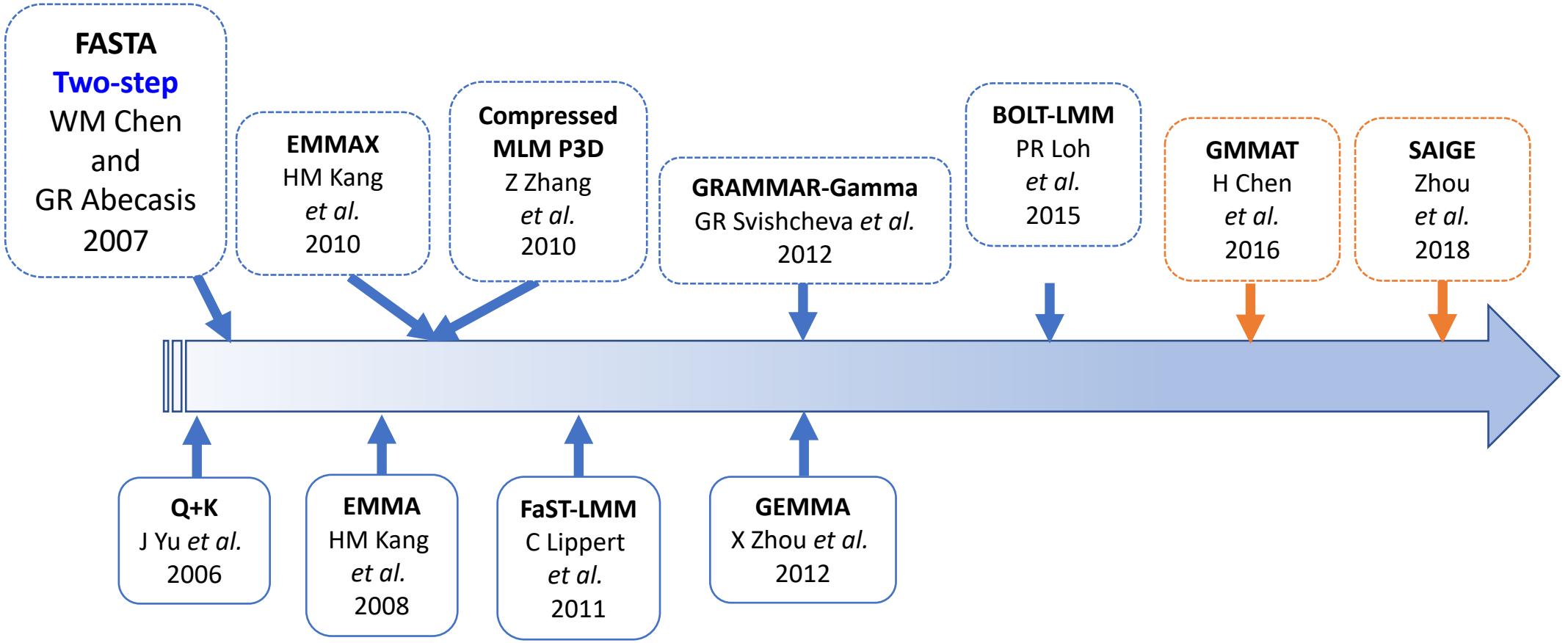
Zhou et al., 2018

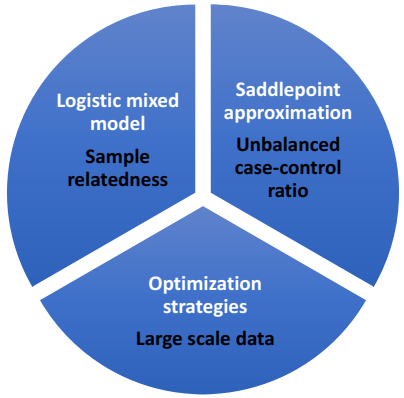
Run Time and Memory Usage



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.

Mixed model methods for GWAS





SAIGE

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

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

Step 1: Fit the null logistic mixed model

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

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

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

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…)

Code and Data Availability

- SAIGE is implemented as an open-source R package available at
  - https://github.com/weizhouUMICH/SAIGE/

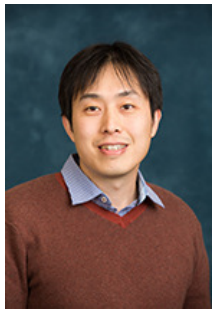
- The GWAS results for 1,403 binary phenotypes with the PheCodes constructed based on ICD codes in UK Biobank using SAIGE are currently available for public download at
  - https://www.dropbox.com/sh/wuj4y8wsqjz78om/AAACfAJK54KtvnzSTAoaZTLma?dl=0

- Michigan PheWeb
  - HRC-imputed UKBB https://pheweb.org/UKB-SAIGE/
  - TOPmed-imputed UKBB https://pheweb.org/UKB-TOPMed/

- Pan-UKBB has conducted a multi-ancestry analysis of 7,221 phenotypes, across 6 continental ancestry groups, for a total of 16,119 genome-wide association studies. https://pan.ukbb.broadinstitute.org/

Teamwork

Cristen Willer



Seunggeun Shawn Lee

- *Goncalo Abecasis*
- *Jonas Nielsen*
- Lars Fritsche
- *Rounak Dey*
- *Sayantan Das*
- *Sarah Gagliano*
- Jonathon LeFaive
- Peter VandeHaar

Seoul National University

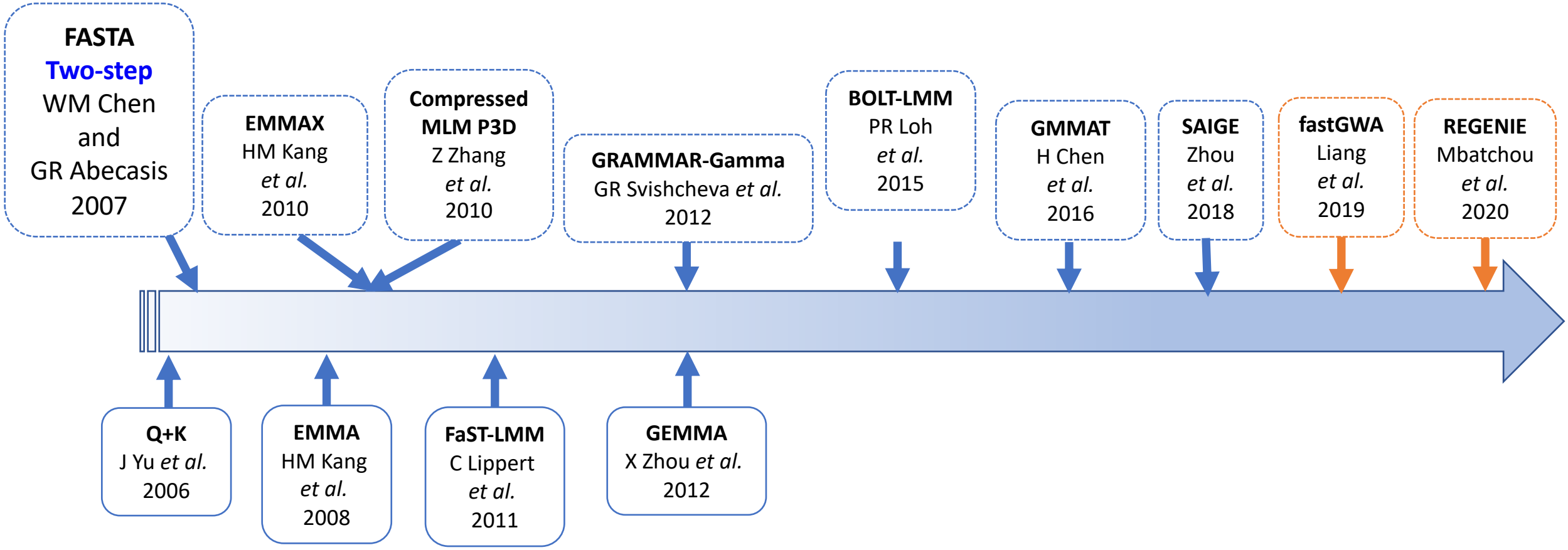
NTNU
K.G. Jebsen Center for Genetic Epidemiology

Kristian Hveem

Maiken Gabrielsen

Anne Skogholt

Efficient mixed model methods for biobank-scale GWAS



Mixed model method for other trait types in large-scale biobanks

- Time-to-event phenotypes
  - GATE: **G**enetic **A**nalysis of **T**ime-to-**E**vent phenotypes
  - R library: https://github.com/weizhou0/GATE
  - Pre-print: https://www.biorxiv.org/content/10.1101/2020.10.31.358234v1.full

- Categorical phenotypes
  - POLMM: **P**roportional **O**dds **L**ogistic **M**ixed **M**odel
  - **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.

Limitations

- Asymptotic approaches were used to achieve scalability for large data sizes, whose performance may be poor when sample sizes are too small.

- Score tests cannot provide accurate effect sizes.

In summary

- Challenges of GWAS exist in large-scale cohorts/biobanks
  - Mixed models can be used to account for sample relatedness in GWAS

- Methods have been developed for biobank-scale GWAS
  - Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)

- SAIGE has been extended for set-based tests to gain more power for rare variant associations, called SAIGE-GENE (Zhou* and Zhao* et al, 2020)

References

- 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.
- 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.
- 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.
- 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." *bioRxiv* (2020).