Lecture 8: Emerging issues showcasing ongoing research

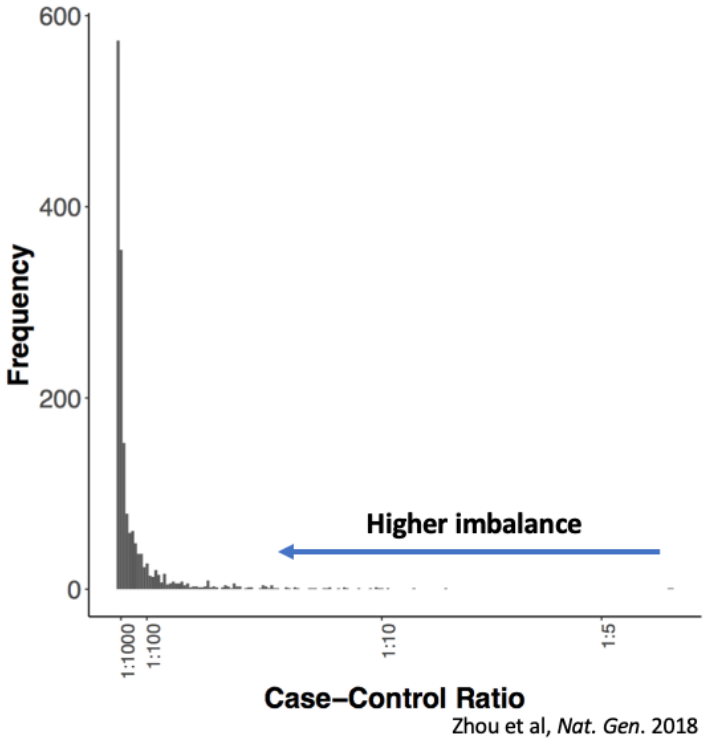
Analysis of imbalanced binary traits

Instructors: Joelle Mbatchou and Loic Yengo

Summer Institute in Statistical Genetics 2022

Binary traits with case-control imbalance

- Many conditions/diseases occur in a minor proportion of the population (i.e. low prevalence).
  - Celiac disease (CC ratio $\sim 1:180$)
  - Psoriasis (CC ratio $\sim 1:180$)
  - Rheumatoid arthritis (CC ratio $\sim 1:80$)
- There are much fewer cases than controls
  $\rightarrow$ unbalanced data



Zhou et al, *Nat. Gen.* 2018

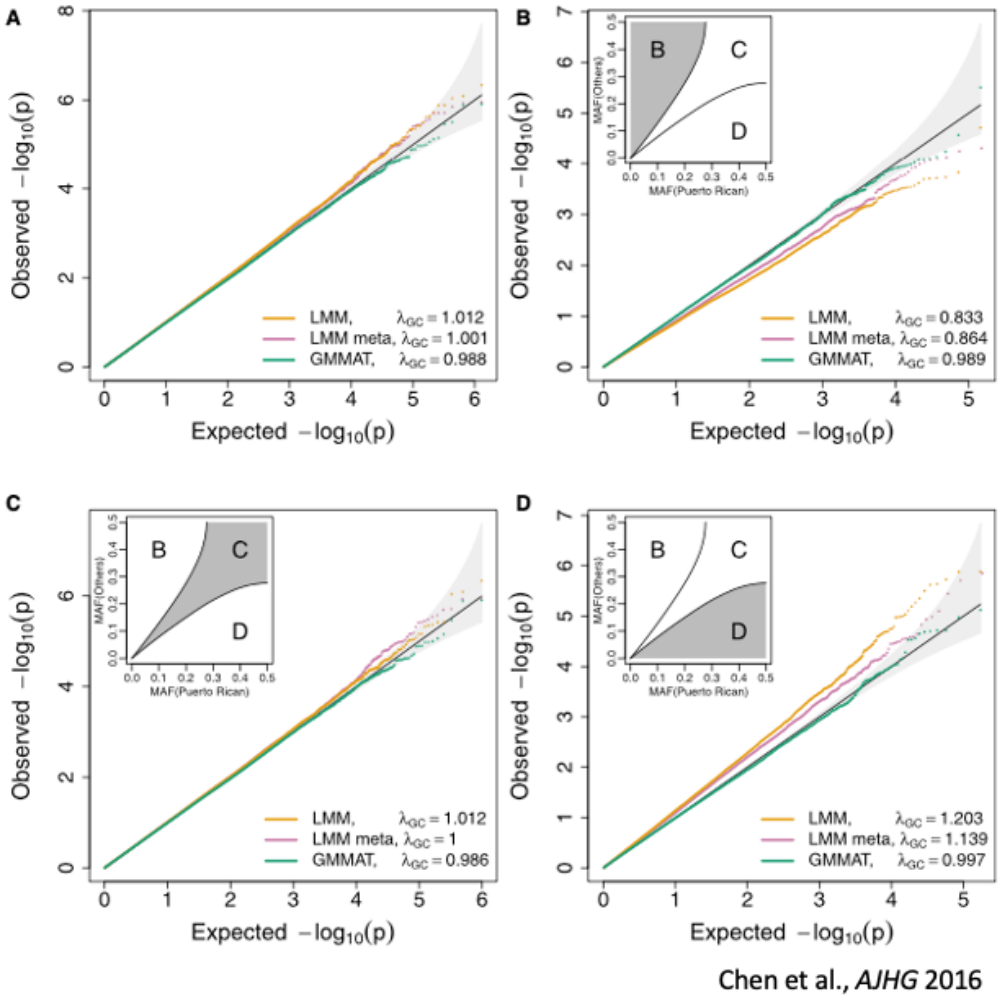
High imbalance - Limitations in GWAS

Case-control imablance can cause major challenges for GWAS:

- Analyzing them as quantitative can lead to substantial inflation
  - Linear model don't capture mean-variance relationship of binary data

$$\text{Linear} : E(Y_i) = \mu_i,$$
$$Var(Y_i) = \sigma^2$$
$$\text{vs.}$$
$$\text{Logistic} : E(Y_i) = \mu_i,$$
$$Var(Y_i) = \mu_i(1 - \mu_i)$$

Example using LMM on asthma phenotype



Chen et al., *AJHG* 2016

High imbalance - Limitations in GWAS

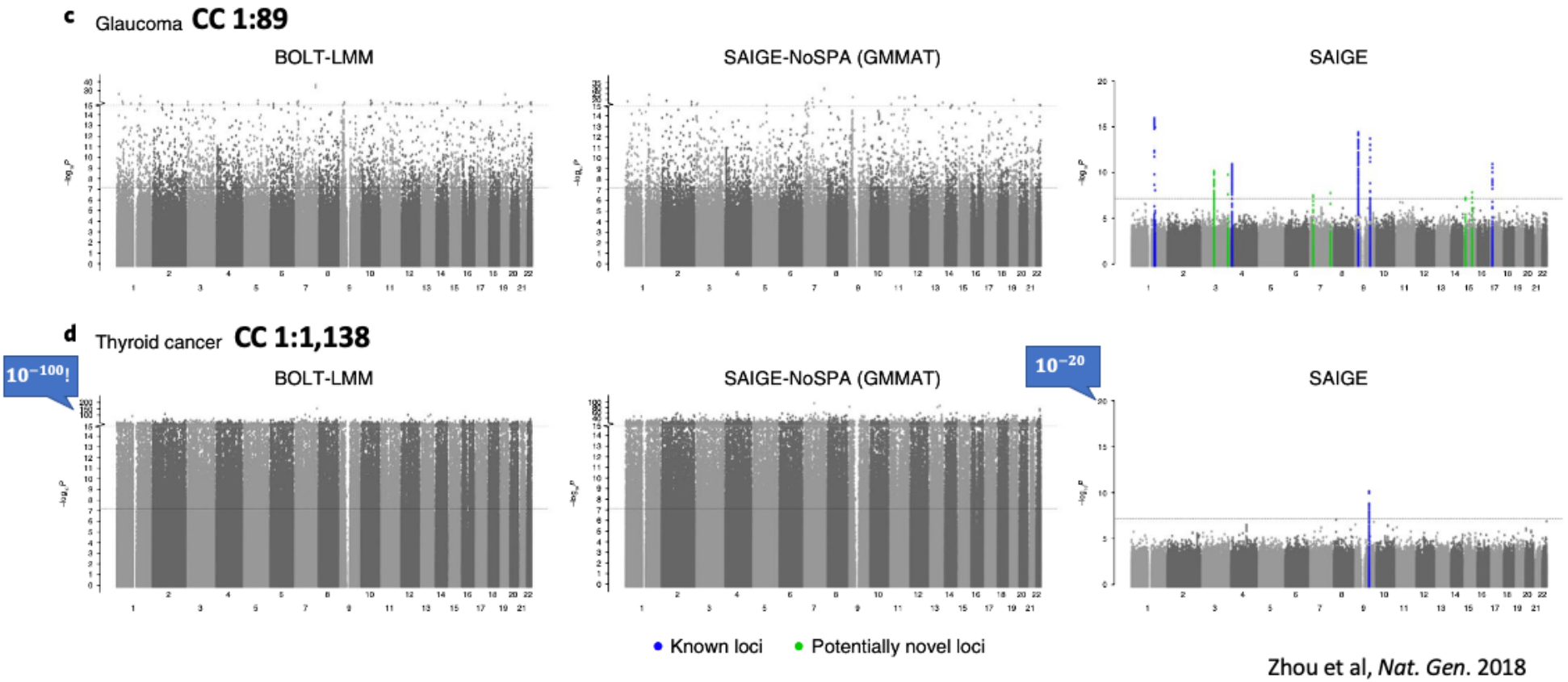
Case-control imablance can cause major challenges for GWAS:

- Using logistic models is better but can still get inflated type 1 error
  - Asymptotic assumptions can become invalid with high case-control imbalance
- With sample structure, logistic mixed models can be computationally burdensome to run

Overcoming limitations: SAIGE

- Zhou et al. (2018) proposed SAIGE which used **efficient computational strategies** (e.g. pre-conditioned conjugate gradient, compact storage of genotypes) to fit logistic mixed models
- In addition, they introduce **Saddlepoint Approximation** (SPA) as a way to obtain more accurate p-values under high case-control imablance
  - SPA approximates the null distribution by using all of the higher order moments, i.e. $E(X^k)$, instead of traditional approaches which only look at first two moments

Overcoming limitations: SAIGE



Zhou et al, *Nat. Gen.* 2018

Challenges with SAIGE

- Still highly intensive when 1,000's of phenotypes have to be analyzed such as in large scale biobanks
- Effect sizes from SPA can be quite inflated for rare variants

REGENIE extension for binary traits

- Switch linear ridge regression model to logistic ridge model
  - Remains computationally efficient due to the two-level approach in Step 1 to reduce number of predictors (e.g. 500K to 2,500)
- To accommodate for case-control imbalance when performing association tests, REGENIE uses Firth penalization:

$$\tilde{\ell}(\beta) = \ell(\beta) + 0.5 \log |I(\beta)|$$

- This has been shown to be effective at addressing issues with quasi/complete separation (e.g. no minor alleles in cases)
  - An approximate implementation has been derived in REGENIE to make it faster (60x) while still remaining accurate

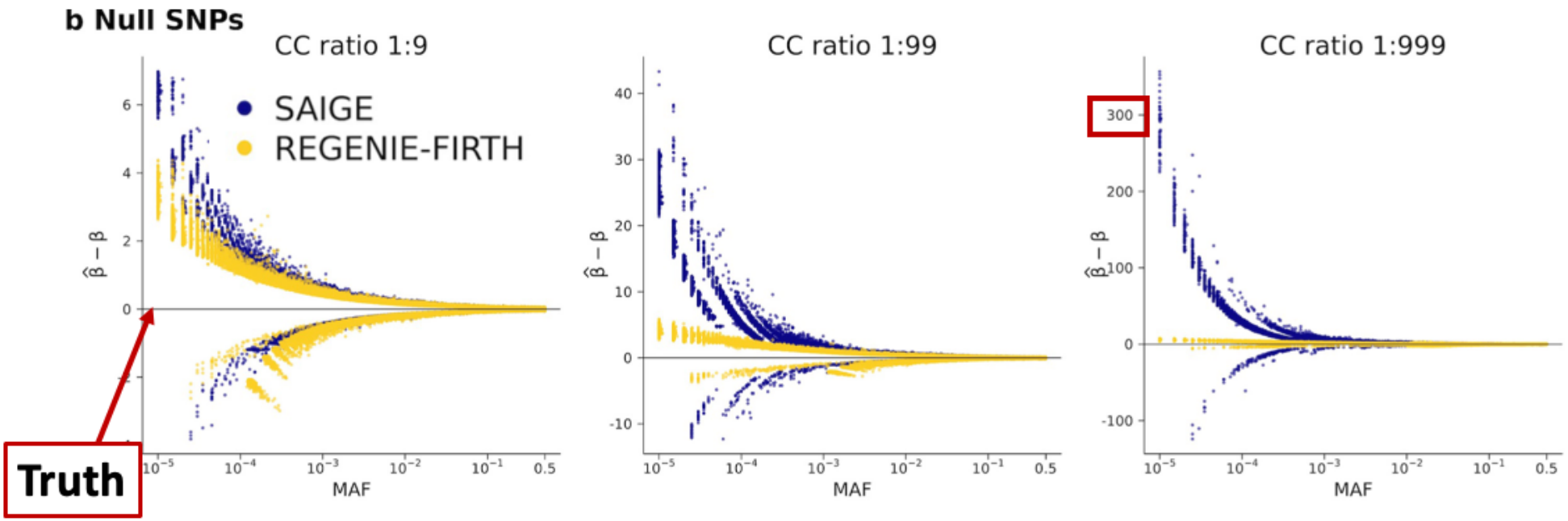
Computational Timing

Table 2 | Computational performance of REGENIE-Firth, REGENIE-SPA and SAIGE when analyzing 50 binary traits with UK Biobank data

Method	Step	Benchmark		
		CPU time (h)	Elapsed time (h)	Memory usage (GB)
REGENIE	1	1,590	117	11.8
REGENIE-LOOCV	1	777	108	19.5
REGENIE-Firth	2	115,492	8,237	7.7
REGENIE-SPA	2	79,363	5,090	9.1
SAIGE	1	275,070	21,428	48.7
SAIGE	2	239,865	173,992	2.1

>4x faster (>350x for Step 1!)

Effect Sizes Estimation



Future directions

- Using more flexible models which don't assume same prior for variant effects to gain power (e.g. spike and slab prior)
- Generalize to other phenotypic measurements
  - Count data (e.g. number of occurences of a specific outcome)
  - Time-to-Event data (e.g. time until disease occurrence)