MIXED LINEAR MODEL ASSOCIATION

Aysu Okbay

VU Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

Social
Science
Genetic
Association
Consortium

WHY MIXED LINEAR MODELS?

- Effective in preventing false-positive associations due to sample structure
  - geographic population structure
  - family relatedness
  - cryptic relatedness
- Increases power by applying a correction that is specific to this structure
- Also increases power in studies without sample structure, by implicitly conditioning on associated loci other than the candidate locus.

THE BASIC APPROACH

- Build a genetic relationship matrix (GRM) modeling genome-wide sample structure
- Estimate its contribution to phenotypic variance using a random-effects model (with or without additional fixed effects)
- Compute association statistics that account for this component of phenotypic variance

MODEL

$$y = x_{test}\beta_{test} + g + e$$

y is the phenotype

x_{test} is the candidate SNP being tested

g is the genetic effect

e is the environmental effect

- Assume everything is mean-centered.
- No covariates – covariates are projected out from both genotypes and phenotypes (equivalent to including them as fixed effects).
- g and e are modeled as random effects
- x_{test} is modeled as a fixed effect with coefficient β_{test}.
- Goal is to test $H_0: \beta_{test} = 0$

MODEL

Under a standard infinitesimal model, the genetic effect is modeled as

$$g = X_{GRM}\beta_{GRM}$$

where

- X_{GRM} is a $N \times M_{GRM}$ matrix, with each column containing normalized genotypes corresponding to a SNP included in the model
  - x_{test} is excluded from X_{GRM} to avoid modelling its effect twice ("proximal contamination")
- β_{GRM} is an M_{GRM}-vector of random SNP effect sizes all drawn from the same normal distribution
- So g has multivariate normal distribution with

$$Cov(g) \propto X_{GRM}X_{GRM}'$$

MODEL

The matrix $K = X_{GRM}X'_{GRM}/M_{\mathrm{GRM}}$ is called the "genetic relationship matrix (GRM)".

$$Cov(g) = \frac{\sigma_g^2 X_{GRM}X'_{GRM}}{M_{GRM}} = \sigma_g^2 K$$

where σ_g^2 is a variance parameter.

Environmental effects are assumed i.i.d. normal, so e is multivariate normal with

$$Cov(e) = \sigma_e^2 I$$

Where I is the $N \times N$ identity matrix and σ_e^2 is another variance parameter.

MODEL

In practice, σ_g^2 and σ_e^2 are unknown.

Two-step approach:

1. Estimate the variance parameters using REML
2. Compute the chi-squared test statistic:

$$\chi^2_{LMM} = \frac{\left(x'_{test} V^{-1} y\right)^2}{x'_{test} V^{-1} x_{test}}$$

where

$$V = Cov(y) = \sigma_g^2 K + \sigma_e^2 I$$

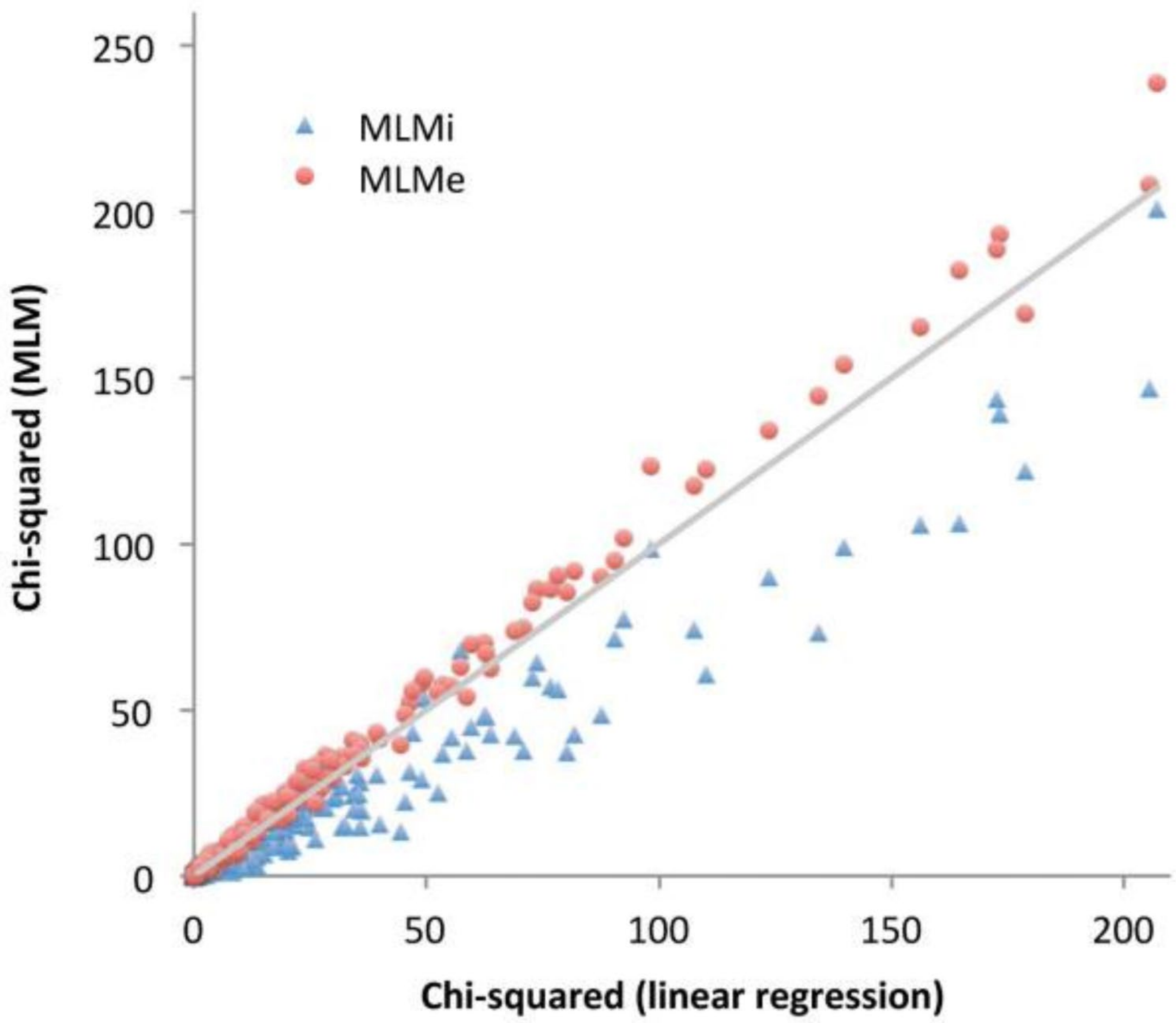
SOFTWARE

- EMMAX
- FaST-LMM
- FaST-LMM-Select
- GEMMA
- GRAMMAR-Gamma
- GCTA-LOCO
- BOLT-LMM
- SAIGE

ANY ISSUES?

PROXIMAL CONTAMINATION

- Inclusion of the candidate marker in the GRM can lead to a loss in power due to double-fitting the candidate marker in the model, both as a fixed effect tested for association and as a random effect as part of the GRM.
- MLM with candidate marker excluded is the mathematically correct approach, but
  - computation time / memory constraints!
- Usually handled by leaving out SNPs on the same chromosome as the tested SNP from the GRM (leave-one-chromosome-out, or LOCO).



Source: Yang J, Zaitlen NA, Goddard ME, Visscher PM and Price AL (2013) Mixed model association methods: advantages and pitfalls. Nat Genet. 2014 Feb;46(2):100-6.

#SNPS IN THE GRM

Two reasons to subsample the markers to be included in the GRM:

1. MLMA is computationally expensive!
   - Most algorithms require $O(MN^2)$ or $O(M^2N)$ running time, where M is the number of markers and N is the sample size.
   - Forces methods to subsample the markers so that $M < N$.
2. Efforts to more accurately model non-infinitesimal genetic architectures
   - Apply the standard infinitesimal mixed model but adapt the input data
   - Increase power by implicitly conditioning only on loci that are relatively likely to be truly associated

#SNPS IN THE GRM

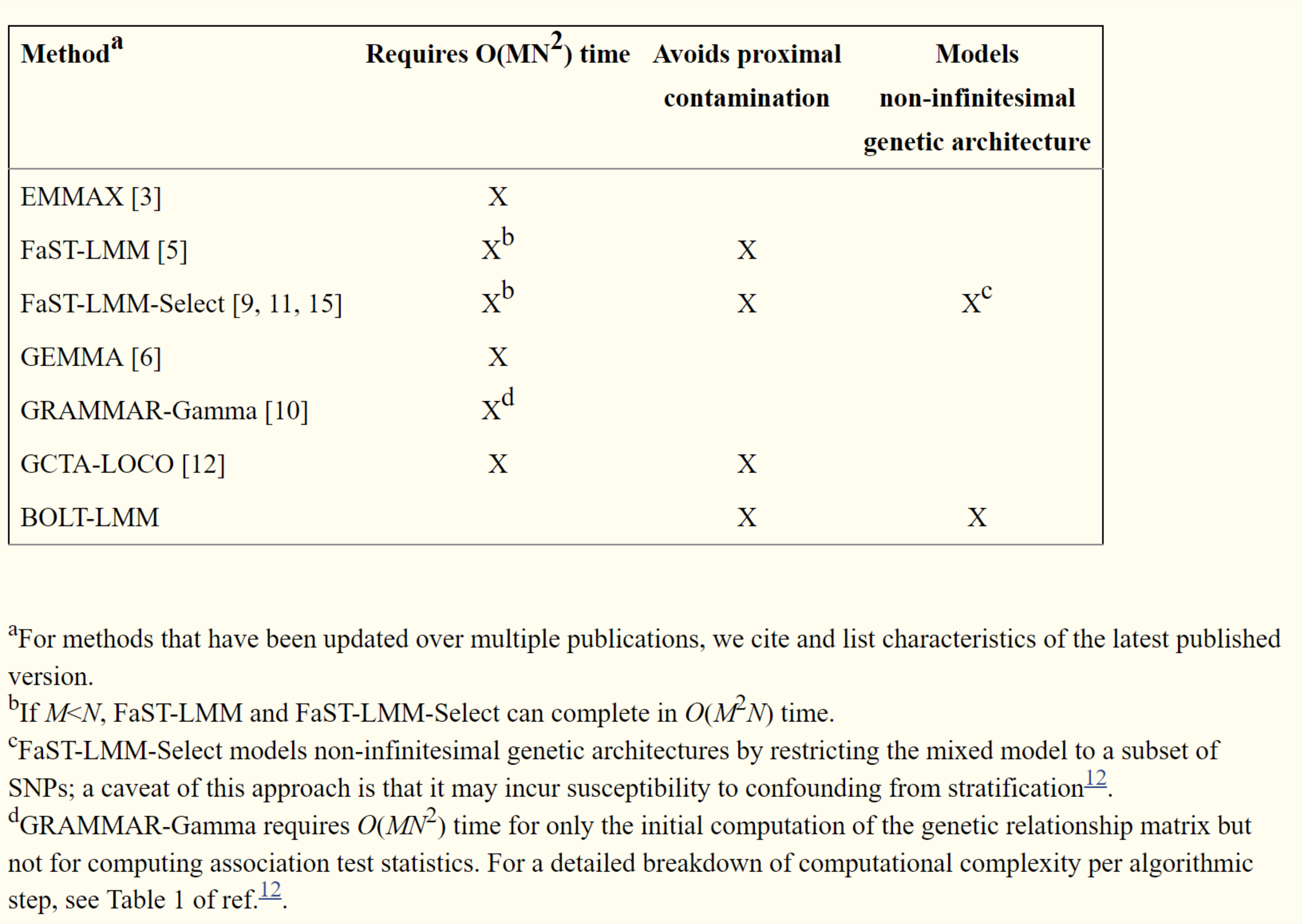
- But using a small subset of markers in GRM can compromise correction for stratification!

- If population stratification is a key concern, include all markers (except for the candidate marker and markers in LD with the candidate marker) in the GRM.

- Subsampling top associated markers is expected to perform well when maximizing power and correcting for cryptic relatedness are the primary goals.

LOSS IN POWER IN ASCERTAINED CASE-CONTROL STUDIES

- MLMA methods assume that study samples are randomly ascertained with respect to the phenotype of interest.
- True for quantitative phenotypes, not true for case-control studies, which generally oversample disease cases to increase study power.
- When disease prevalence is small, MLMA can suffer a substantial loss in power.
- SAIGE can handle this!



Method[a]	Requires O(MN^2) time	Avoids proximal contamination	Models non-infinitesimal genetic architecture
EMMAX [3]	X		
FaST-LMM [5]	X[b]	X	
FaST-LMM-Select [9, 11, 15]	X[b]	X	X[c]
GEMMA [6]	X		
GRAMMAR-Gamma [10]	X[d]		
GCTA-LOCO [12]	X	X	
BOLT-LMM		X	X

[a] For methods that have been updated over multiple publications, we cite and list characteristics of the latest published version.

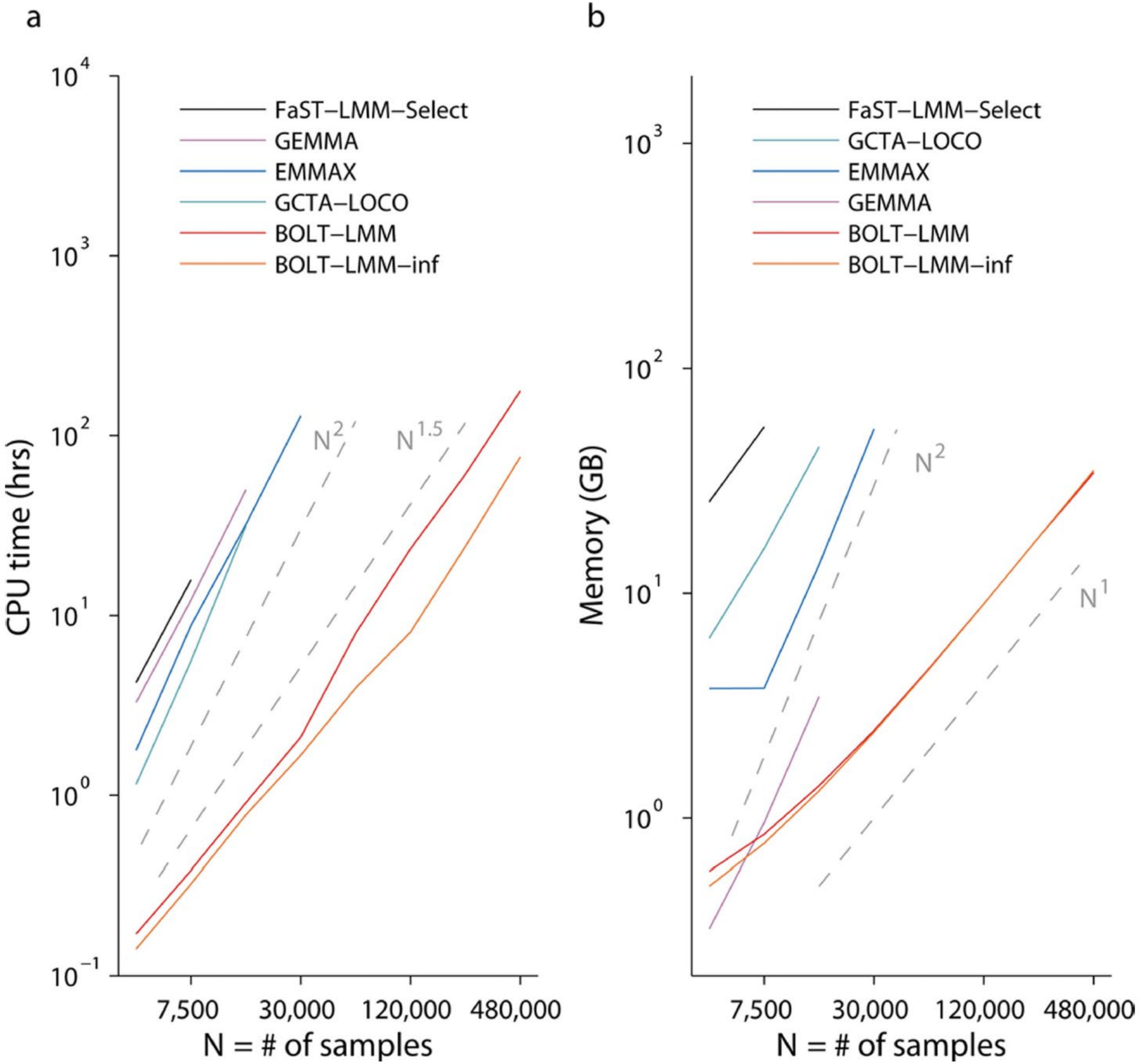
[b] If $M<N$, FaST-LMM and FaST-LMM-Select can complete in $O(M^2N)$ time.

[c] FaST-LMM-Select models non-infinitesimal genetic architectures by restricting the mixed model to a subset of SNPs; a caveat of this approach is that it may incur susceptibility to confounding from stratification[12].

[d] GRAMMAR-Gamma requires $O(MN^2)$ time for only the initial computation of the genetic relationship matrix but not for computing association test statistics. For a detailed breakdown of computational complexity per algorithmic step, see Table 1 of ref.[12].

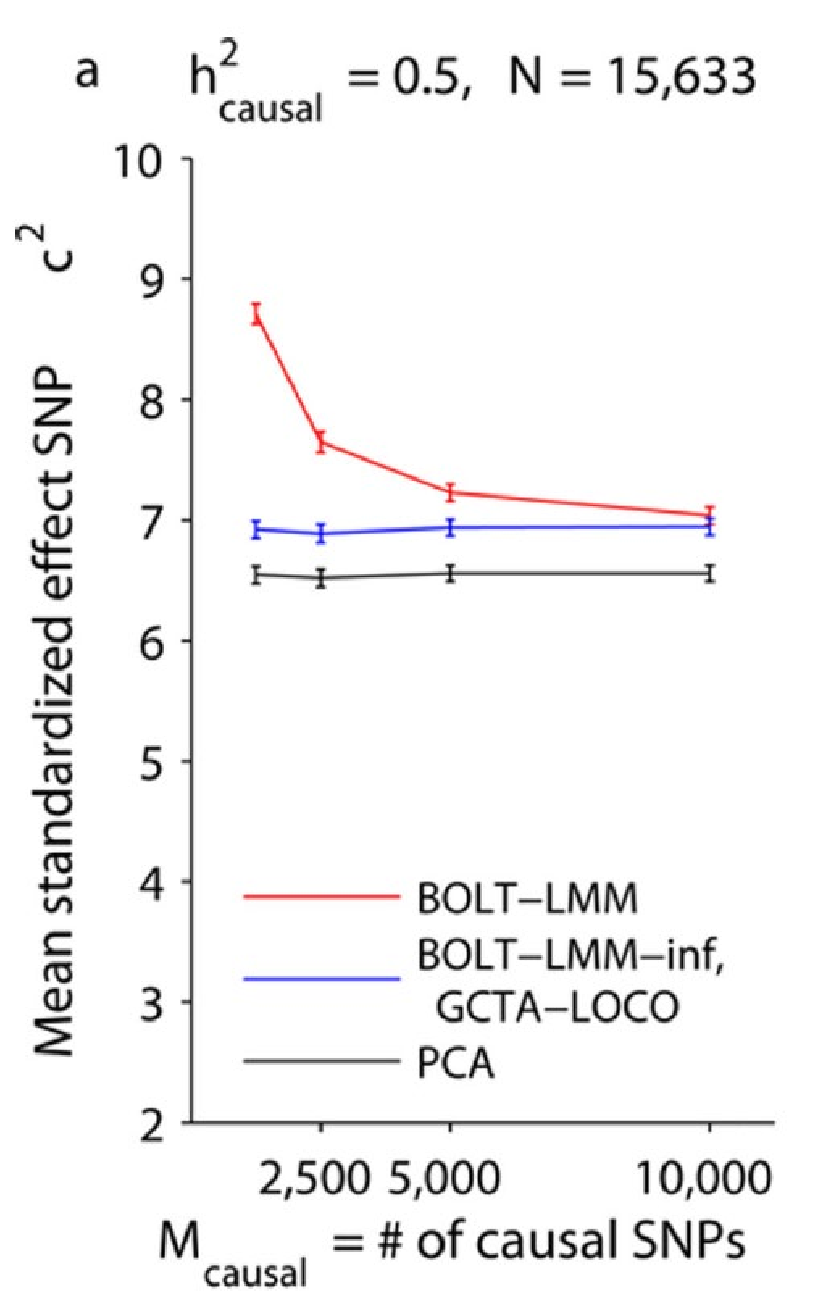
MORE ON BOLT-LMM

- Uses some approximation algorithms that reduce the time and memory cost
- Runs in a small number of $O(MN)$-time iterations
- Increases power by modeling non-infinitesimal genetic architectures
  - Generalizes the standard model by imposing a Bayesian mixture prior on marker effect sizes
- Gives the model greater flexibility to accommodate large-effect SNPs while maintaining effective modelling of genome-wide effects such as ancestry.



Source: Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M.,…, Price, A.L.(2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. (2015) 47:284–90. 10.1038/ng.3190

Power gain decreases with increasing number of causal SNPs (BOLT-LMM-inf ≈ GCTA-LOCO)



Source: Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M.,…, Price, A.L.(2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. (2015) 47:284–90. 10.1038/ng.3190