# permGWAS: Efficient Permutation-based Genome-wide Association Studies for Skewed Phenotypic Distributions

Maura John[1,2], Markus J. Ankenbrand[3], Carolin Artmann[3], Jan A. Freudenthal[3], Arthur Korte[3], Dominik G. Grimm[1,2,4]

1 Technical University of Munich, TUM Campus Straubing for Biotechnology and Sustainability, Bioinformatics
2 Weihenstephan-Triesdorf University of Applied Sciences, Bioinformatics
3 Center for Computational and Theoretical Biology, University of Würzburg
4 Technical University of Munich, TUM Department of Informatics

## Motivation

- Genome-wide association studies (GWAS) are a key tool to analyze relationship between genotypes and phenotypes
- Linear mixed models (LMMs) test whether a single genetic marker is associated with a given phenotype via the model

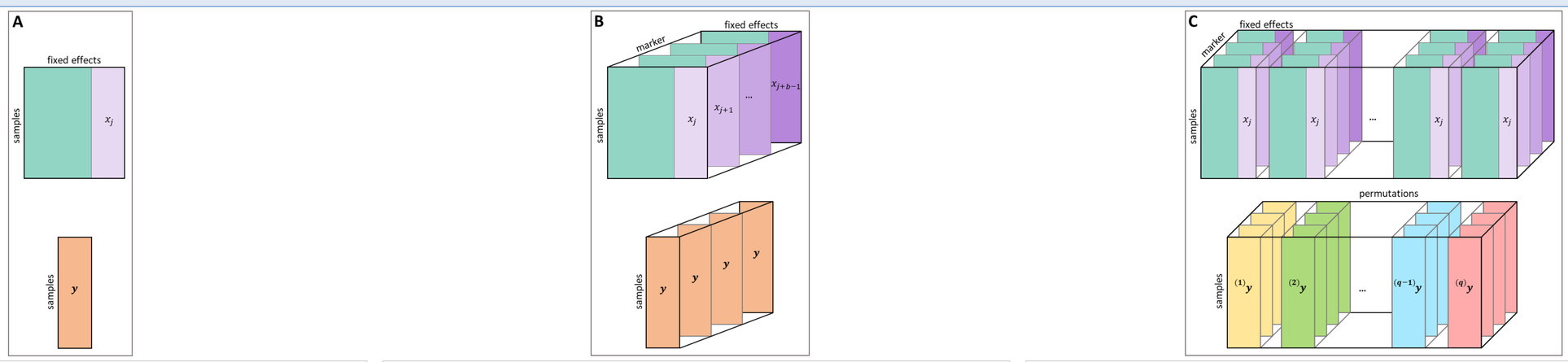
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}),\ \boldsymbol{u} \sim \mathcal{N}(0, \sigma_g^2 \boldsymbol{K})$$

phenotype | fixed effects | random effects | noise | variance components | genetic similarity matrix

- Variance components are estimated by maximizing the likelihood function $L(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2)$

- Assumptions of Gaussian distribution and independent genetic markers are often violated in real-world data
- Commonly used Bonferroni significance threshold [1] often too conservative for normally distributed phenotypes or not stringent enough for phenotypes with skewed distributions
- Idea: permutation-based threshold via *maxT* method [2]:
  Permute phenotype $q$ times and compute minimal p-value for each permutation
  The adjusted threshold is the $\alpha^{th}$ percentile
- Problem: enormous computational complexity
- Solution: permGWAS, an efficient reformulation of LMMs using 3D and 4D tensors that can provide permutation-based thresholds

## permGWAS Architecture




**Basic Linear Mixed Model**
Estimate the effect of a single genetic marker:

$$L(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X\beta}; \sigma_g^2\, \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})$$

**Batch-wise Linear Mixed Model**
Estimate the effects of several genetic markers simultaneously using 3D-tensors:

$$L(\boldsymbol{\beta}_j^b, \sigma_g^2, \sigma_e^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}_j^b \boldsymbol{\beta}_j^b; \sigma_g^2\, \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})$$

**Permutation-based Linear Mixed Model**
Estimate the effects of several genetic markers and permutations simultaneously using 4D-tensors:

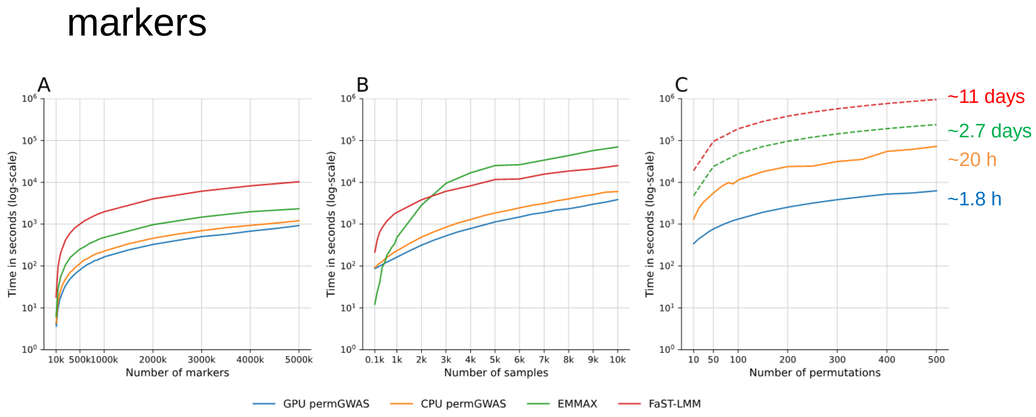
$$L({}^q\boldsymbol{\beta}_j^b, \sigma_g^2, \sigma_e^2) = \mathcal{N}(\boldsymbol{y}|{}^q\boldsymbol{X}_j^b\, {}^q\boldsymbol{\beta}_j^b; \sigma_g^2\, \boldsymbol{K} + \sigma_e^2 \boldsymbol{I})$$

## Results

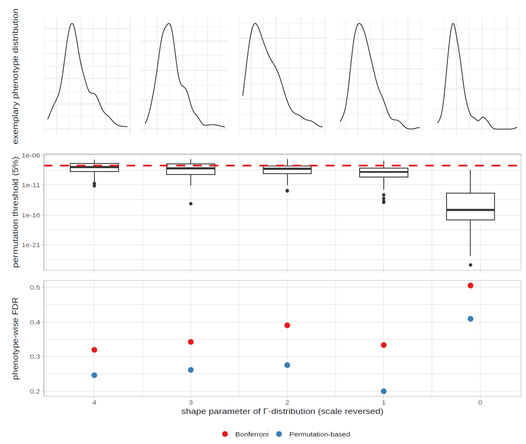
### Runtime comparison to state-of-the-art

Comparison of the runtime of permGWAS to EMMAX [3] and FaST-LMM [4] with respect to

A. Number of **markers** with fixed number of $1000$ samples
B. Number of **samples** with fixed number of $10^6$ markers
C. Number of **permutations** with $1000$ samples and $10^6$ markers



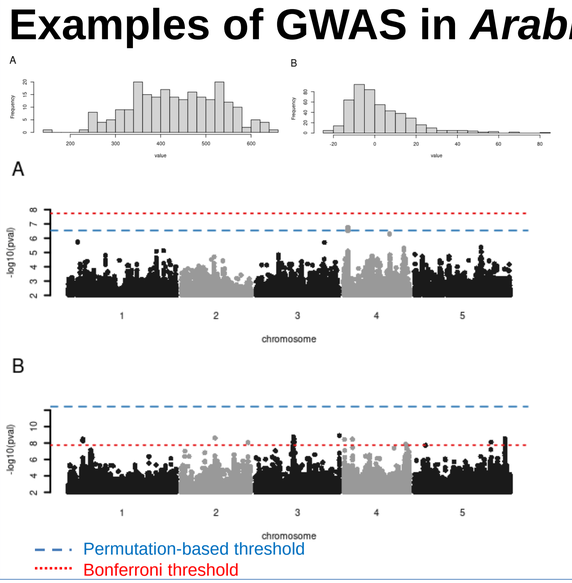

### False Discovery Rate for skewed phenotypes




Comparison of False Discovery Rate (FDR) of permutation-based thresholds and Bonferroni threshold for synthetic phenotypes with gamma-distributed noise:

- ➢ FDR increases when phenotypes become more skewed
- ➢ Permutation-based thresholds become more stringent for skewed phenotypes
- ➢ Permutation-based thresholds have lower FDR for skewed phenotypes than Bonferroni threshold

### Examples of GWAS in *Arabidopsis thaliana*




Analysis of GWAS results for real phenotypes from model plant *Arabidopsis thaliana [5,6]:*

A. For normally distributed phenotypes the permutation-based threshold is less conservative
B. For non-normally distributed phenotypes the permutation-based threshold is more stringent

## References

**Publication:**
John, M., Ankenbrand, M. J., Artmann, C., Freudenthal, J. A., Korte, A., & Grimm, D. G. (2022). Efficient Permutation-based Genome-wide Association Studies for Normal and Skewed Phenotypic Distributions. bioRxiv. **Accepted at ECCB 2022**

**References:**

1. Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8, 3–62.
2. Westfall, P. H. and Young, S. S. (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment, volume 279. John Wiley & Sons.
3. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4), 348–354.
4. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. Nature methods, 8(10), 833–835.
5. Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., et al. (2016). Arapheno: a public database for arabidopsis thaliana phenotypes. Nucleic Acids Research, page gkw986.
6. Togninalli, M., Seren, Ü., Freudenthal, J. A., Monroe, J. G., Meng, D., et al. (2020). Arapheno and the aragwas catalog 2020: a major database update including rna-seq and knockout mutation data for arabidopsis thaliana. Nucleic acids research, 48(D1), D1063–D1068.

**Website:**
http://bit.cs.tum.de

**GitHub:**
https://github.com/grimmlab/permGWAS

**Funding:**
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung