

GWAS model **extensions**

- **multi trait and multi locus** models (& more) -

Christian Werner

(Quantitative geneticist and biostatistician) **EiB**, **CIMMYT**, Texcoco (Mexico)

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)



HerrFalloppio

Oscar González-Recio

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)



OscarGenomics



GWAS: **multiple-trait models**



Multiple-trait GWAS: **why?**

- phenotypes may be phenotypically and (more importantly) genetically **correlated** (← **this is not collinearity! Do you know why?**)
- accounting for correlations in a multivariate model will:
 - yield **more accurate estimates of SNP coefficients**
 - as a consequence, yield **more accurate p-values** (e.g. less inflation)
 - → fewer spurious associations
- may increase power of analysis
- **beware of sample size!** (typically larger datasets are needed to fit accurately genetic correlations in a mixed linear model)

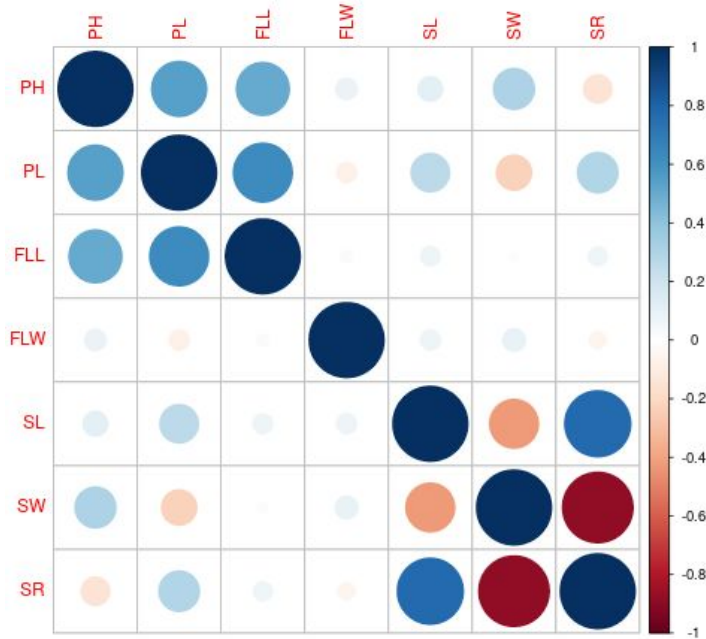


Multiple-trait GWAS: a bit of literature

1. Yoshida GM, Yáñez JM. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. *BMC genomics*. 2021 Dec;22(1):1-3.
2. Julienne H, Laville V, McCaw ZR, He Z, Guillemot V, Lasry C, Ziyatdinov A, Nerin C, Vaysse A, Lechat P, Ménager H. Multitrait GWAS to connect disease variants and biological mechanisms. *PLoS genetics*. 2021 Aug 30;17(8):e1009713.
3. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA, Wedow R, Zacher M, Furlotte NA, Magnusson P. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics*. 2018 Feb;50(2):229-37.
4. Meng X, Li L, Pascual J, Rahikainen M, Yi C, Jost R, He C, Fournier-Level A, Borevitz J, Kangasjärvi S, Whelan J. GWAS on multiple traits identifies mitochondrial ACONITASE3 as important for acclimation to submergence stress. *Plant physiology*. 2022 Apr;188(4):2039-58.
5. Malik MA, Ludl AA, Michoel T. High-dimensional multi-trait GWAS by reverse prediction of genotypes. *arXiv preprint arXiv:2111.00108*. 2021 Oct 29.
6. Bonnemaier PW, Leeuwen EM, Iglesias AI, Gharahkhani P, Vitart V, Khawaja AP, Simcoe M, Höhn R, Cree AJ, Igo RP, Gerhold-Ay A. Multi-trait genome-wide association study identifies new loci associated with optic disc parameters. *Communications biology*. 2019 Nov 27;2(1):1-2.



Multiple-trait GWAS: **rice data**



some phenotypes are correlated: e.g. seed length (SL) and seed width (SW)

Practical session

AWS/locally:

model_extensions/multi_trait



GWAS: **multiple-locus models**



why just **one SNP** at a time?

Improvements on basic GWAS models mainly in these directions:

- computing speed
- statistical power
- reduce false positive signals



multiple-SNP GWAS

Multiple associated markers can be genetically linked → remove redundancy:

- only **one associated marker is selected from each bin/cluster** of associated markers
- **variable kinship** to eliminate the confounding between kinship and testing markers (trait-associated markers are excluded from the kinship calculation if they are also associated with the testing markers)
- fits two models (iteratively):
 - one tests one marker at a time with multiple associated markers as fixed effect in the MLM (mixed linear model) → forward stepwise regression
 - the second selects covariate markers to test the remaining markers with a fixed effect model (FEM)
- and many more details ...



BLINK and FarmCPU: **multilocus models**

- **BLINK** and **FarmCPU** are software implementations of the multi-locus model for testing markers across genome
- BLINK is an enhanced version of FarmCPU
- BLINK conducts two fixed effect models iteratively:
 - one tests marker one at time with multiple associated markers fitted as covariates to account for population stratification;
 - the other selects the covariate markers to directly control spurious association instead of kinship, unmasking the confounding between testing marker and kinship.
- BLINK eliminates the requirement that genes underlying a trait are distributed equally across genome to further improve statistical power.
- BLINK also replaces the REstricted Maximum Likelihood (REML) in a mixed linear model (FarmCPU) with Bayesian Information Content (BIC) in a fixed effect model to boost computing speed



material for further reading

1. BLINK

- i. [Paper](#)
- ii. [Software](#)

2. FARMCpu

- i. [paper](#)
- ii. [Software](#)

3. Examples:

- i. [Linge et al.](#)
- ii. [Kaler et al.](#)

BLINK: multilocus models

Practical session

aws:

→ `model_extensions/multi_locus/`

GRAMMAR: Genome-wide Association using Mixed Model and Regression



GRAMMAR

To be done ...



GWAS: more software



GEMMA: Genome-wide Efficient Mixed Model Association

- software for the application of **linear mixed models** (LMMs) to genome-wide association studies (GWAS) [and more ...]
- <https://github.com/genetics-statistics/GEMMA>
- C++ scripts
- Install:
 - precompiled binaries
 - Docker images
 - **Conda (bioconda)** [chosen here]
 - Compile from source



GEMMA: **G**enome-wide **E**fficient **M**ixed **M**odel **A**ssociation

- imputed data are needed
- Plink binary input files
- phenotypes to be provided in the .fam file (Plink, 6th column)
- relatedness matrix (same order as Plink files)
- covariates file (optional)

Practical session

AWS:

`model_extensions/gemma`



GEMMA: Genome-wide Efficient Mixed Model Association

Dare try it with a binary trait?

- dog data → data preparation (no covariates)
- fit a binary trait using a **linear model**
 - `gemma -bfile <file> -k <file> -lmm -o <file>`
- fit a binary trait using a **probit model** instead
 - `gemma -bfile <file> -k <file> -bslmm 3 -o <file>`



REGENIE: Genome-wide Efficient Mixed Model Association

Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, Habegger L. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics*. 2021 Jul;53(7):1097-103.

- C++ computer package for large-scale GWAS
- developed at the Regeneron Genetics Center
- works on quantitative and binary traits
- can handle population structure and relatedness
- can process multiple phenotypes at once efficiently
- **fast and memory efficient** 🔥
- accepts BGEN, PLINK bed/bim/fam and PLINK2 pgen/pvar/psam genetic data formats



REGENIE: **G**enome-wide **E**fficient **M**ixed **M**odel **A**ssociation

Practical session

AWS:

`model_extensions/regenie`

NEXT LECTURE

GWASpoly example for polyploid species

