

**LMM and  
meta-  
analysis**

**from the  
Health Data Science  
Sandbox**



**Samuele Soraggi, PhD**

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF  
COPENHAGEN



# GWAS with the Genomics Sandbox



## Today's topics

### Association tests

- GWAS
  - Linear mixed models
  - Tools
- Meta-analysis

# Linear Mixed Models (LLMs)

## Challenges in traditional GWAS

- Standard GWAS uses **linear regression**:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$y = X\beta + \epsilon, \quad \text{with } \epsilon \sim N(0, \sigma^2 I)$$

- **Population structure and relatedness** introduce false positives
  - The model is missing terms to describe their effect
- E.g. Height differences between populations can confound results

# Linear Mixed Models (LLMs)

## The random effect term

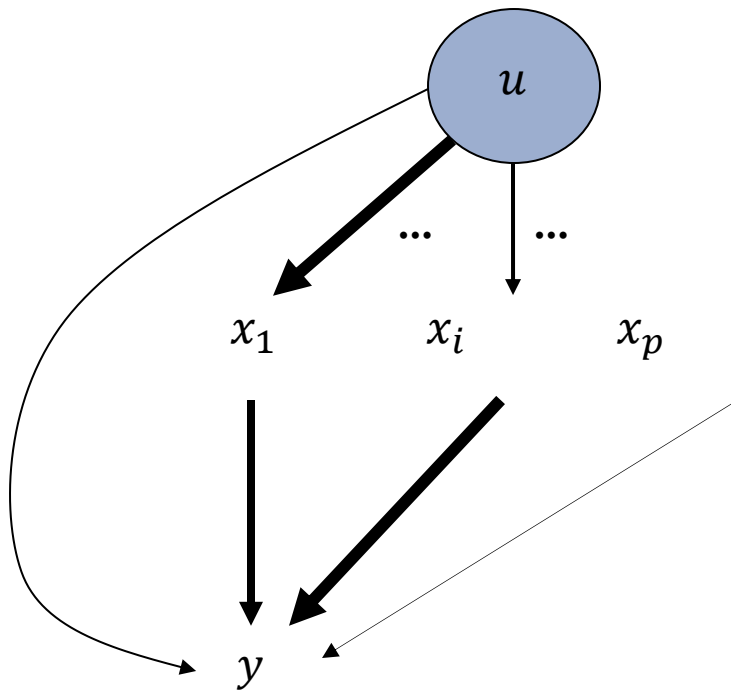
$$y = \underbrace{X\beta}_{\text{Fixed effect}} + \underbrace{u}_{\text{Polygene or Random effect}} + \underbrace{\epsilon}_{\text{Residual}}, \quad \text{with } \epsilon \sim N(0, \sigma^2 I),$$

$u \sim N(0, Z)$   
Z cov. matrix of r.e.

# Linear Mixed Models (LLMs)

## The random effect term

How does the r.e. term acts in our model?



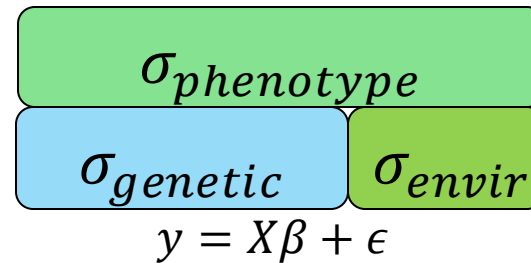
← Pop.structure, ...

← SNPs affected by  $u$

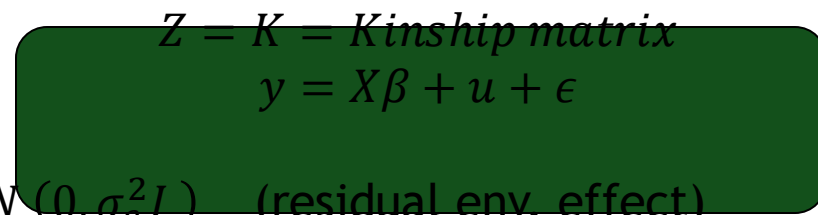
← Trait affected by  $u + \text{SNPs}$

# Linear Mixed Models (LLMs)

variances to consider

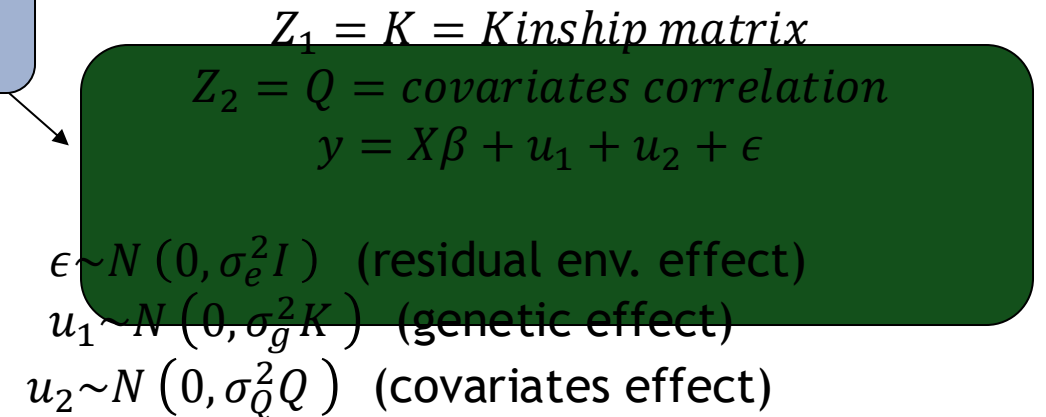
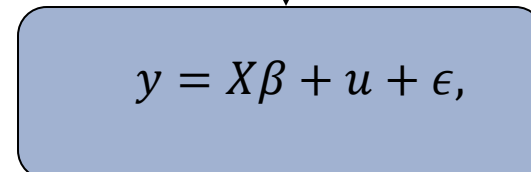


To include genetic effect



$\epsilon \sim N(0, \sigma_e^2 I)$  (residual env. effect)  
 $u \sim N(0, \sigma_g^2 K)$  (genetic effect)

To include genetic effect  
+  
other covariates



$\epsilon \sim N(0, \sigma_e^2 I)$  (residual env. effect)  
 $u_1 \sim N(0, \sigma_g^2 K)$  (genetic effect)  
 $u_2 \sim N(0, \sigma_Q^2 Q)$  (covariates effect)

# Linear Mixed Models (LLMs)

In practice - fill in the variables

$$y = X\beta + u_1 + u_2 + \epsilon \rightarrow \mathbf{y} = \mathbf{X}'\boldsymbol{\beta}' + \mathbf{u}$$

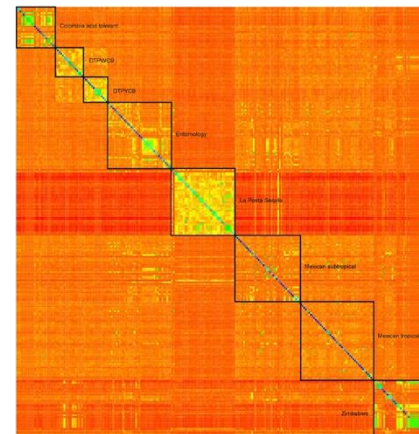
$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} & PC_{11} & PC_{12} & \dots \\ x_{21} & x_{22} & \dots & \dots & PC_{21} & PC_{22} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kn} & PC_{k1} & PC_{k2} & \dots \end{pmatrix}$$

$\underbrace{\hspace{10em}}_{\text{n SNPs (normalized) k samples}} \quad \underbrace{\hspace{10em}}_{\text{Other covariates (PCA, ...)}}$

$$\mathbf{u} \sim N(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) = N(0, \mathbf{V})$$

Total covariance matrix

Residual environmental effect and noise



Kinship matrix  
(finer relatedness structure)  
Easily calculated with plink, GCTA, ...

# Linear Mixed Models (LLMs)

In practice - parameters not directly calculated, heritability

Very innocent-looking formula

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



What about the variances  $\sigma_g^2, \sigma_e^2$

Heritability comes into play

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{\sigma_g^2}{\sigma_{phenotype}^2}$$

Heritability = variance proportion explained only by genetic variance.

The fundamental parameter for phenotype prediction



# Linear Mixed Models (LLMs)

Some approaches

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta}' + \mathbf{u}$$

$$\mathbf{u} \sim N(0, V)$$

$$h^2 = \frac{\sigma_g^2}{\sigma_y^2} \rightarrow h^2 \sigma_y^2 = \sigma_g^2$$

y usually normalized so  $\sigma_y^2 = 1$

BOLT-LMM  
(Loh *et al.* 2015)

Optimizes

$$V = \sigma_g^2 K + \sigma_e^2 I$$

Through prior on sigma's.

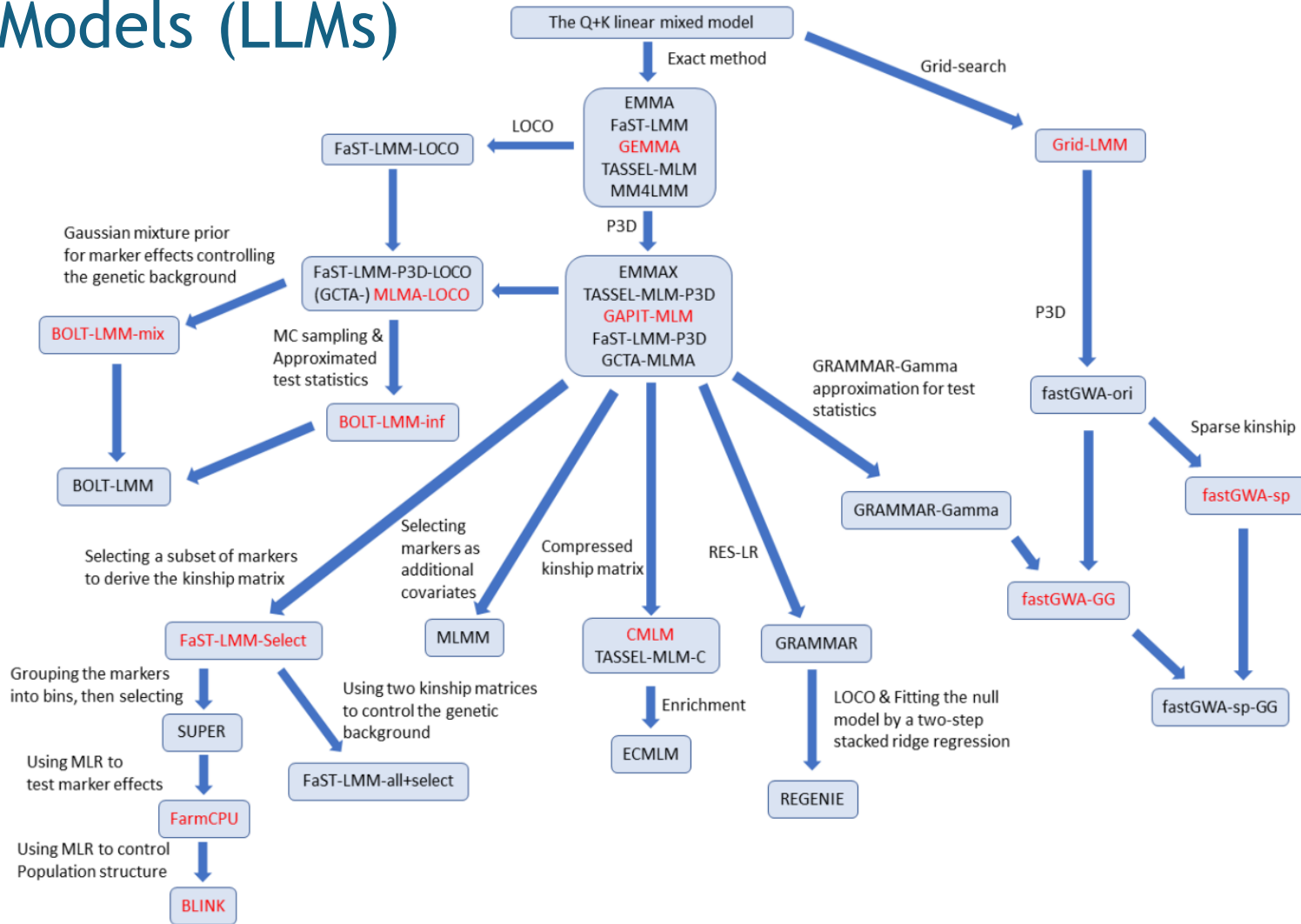
Then uses  $h^2 = \frac{\sigma_g^2}{\sigma_y^2}$  to  
define heritability.

Regenie  
(Yang *et al.* 2011)

- Does not use K, but principal components
- Shrinks effect of SNPs to 0 to avoid overfitting
- Multiple other steps to avoid overfitting such as penalties and cross-validation
- Very fast and good for large studies with > Millions of SNPs

# Linear Mixed Models (LLMs)

## Some approaches



**A phylogeny of 33 GWAS algorithms.** If two algorithms are connected by an arrow, the target is based on the source with additional techniques indicated by the text. If two algorithms target the same algorithm, the target combines the techniques implemented by the two sources. P3D, population parameters previously determined; MC, Monte-Carlo; LOCO, leave-one-chromosome-out; MLR, multi-variate linear regression; RES-LR, using the residuals from the null model as the response to test marker effects in a simple linear model. From (Liu et al, 2023, bioArxiv. DOI 10.1101/2023.12.05.570105).

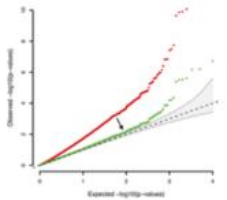
# Beyond LLMs

## New methods

New methods are

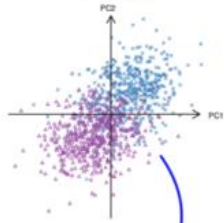
- Fast on large datasets
- Reliable in detecting association
- Use mixed models
- Have faster implementations

Genomic control



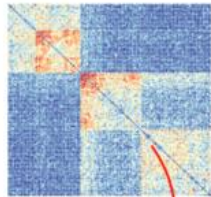
$$\hat{\beta} \rightarrow \hat{\beta}'$$

PCA



$$Y_i = \beta X_i + \gamma \bar{PC}_i + \epsilon_i$$

Mixed models



$$Y_i = \beta X_i + \eta_i + \epsilon_i$$

From basic Genomics Control (rescaling test statistics) to correcting through PCA only and to Mixed Models, of which LMMs are a special case. Credit Iain Mathieson.

Some examples

[LDAK-KVIK \(Hof and Speed, 2024\)](#)

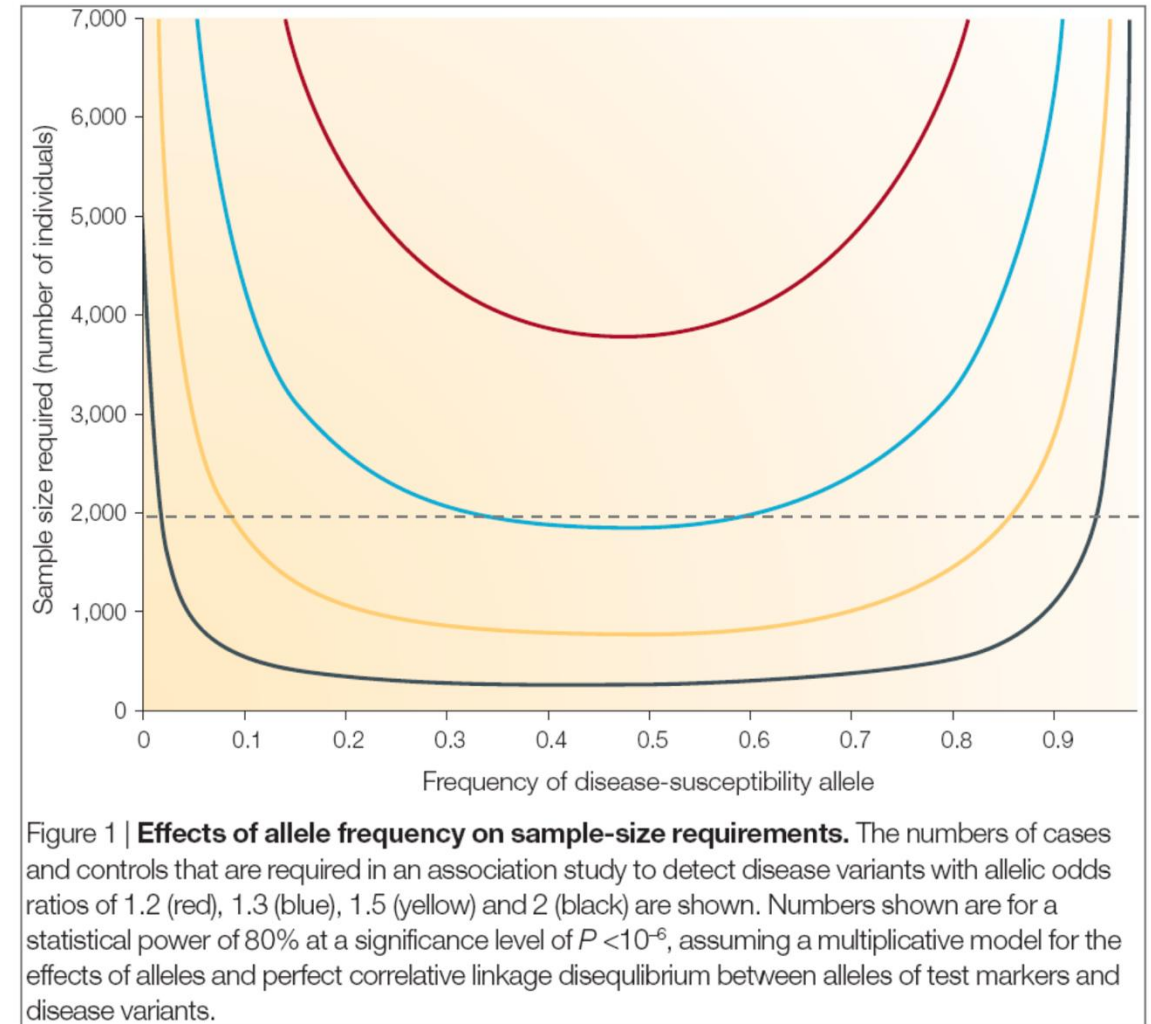
Uses mixed models: often the preferred tool, are more flexible and can be more complex than LMMs. Faster and outperforming REGENIE, BOLT-LMM

[Quickdraw \(Loya et al, 2025\)](#)

Shrinks variant effects to increase association power, computationally efficient with variational inference and GPU calculations. It also uses mixed models.

# Meta studies

- Individual genetic variants often have **small effects**.
- **Large sample sizes** are required to detect novel associations.
- **Low minor allele frequency (MAF)** reduces statistical power.
- **Combining individual level data** is technically and administratively challenging (large dataset sizes, variations in study designs, and data protection constraints)



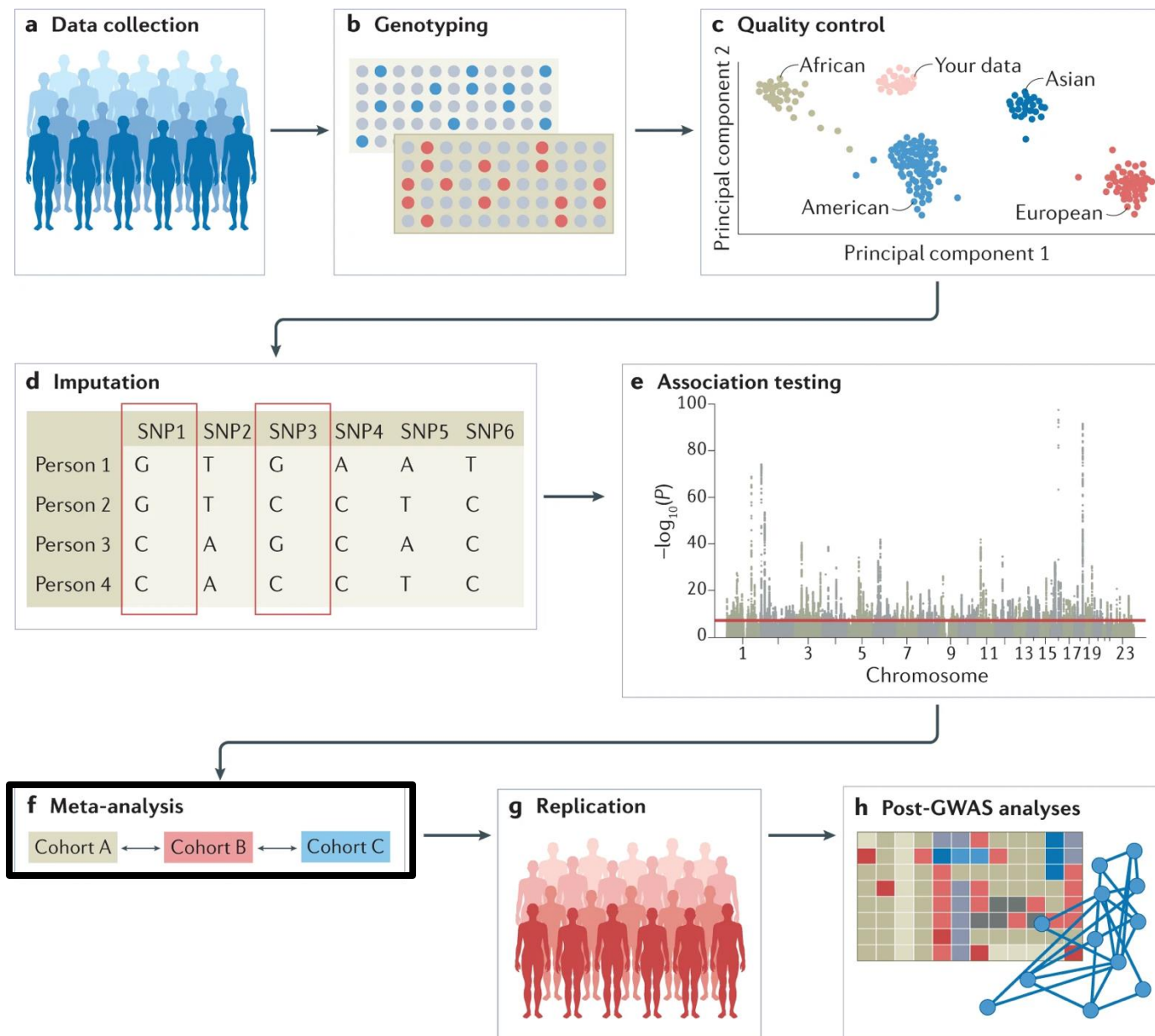
# Meta studies

GWAS summary statistics are publicly available:

- Meta-studies integrate those summary statistics
- Increased statistical power as sample size increases

Softwares:

- METAL
- GWAMA
- MANTRA



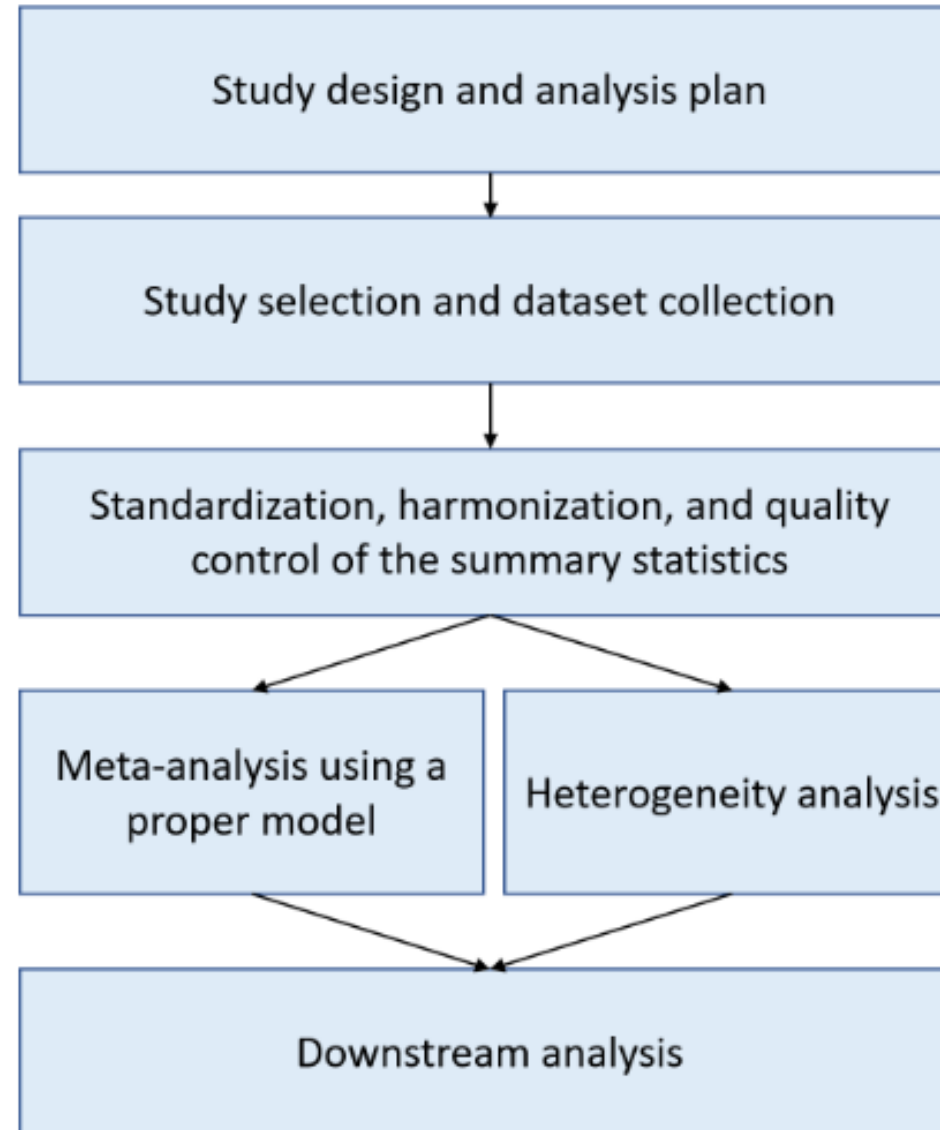
# Meta studies

GWAS summary statistics are publicly available:

- Meta-studies integrate those summary statistics
- Increased statistical power as sample size increases

Softwares:

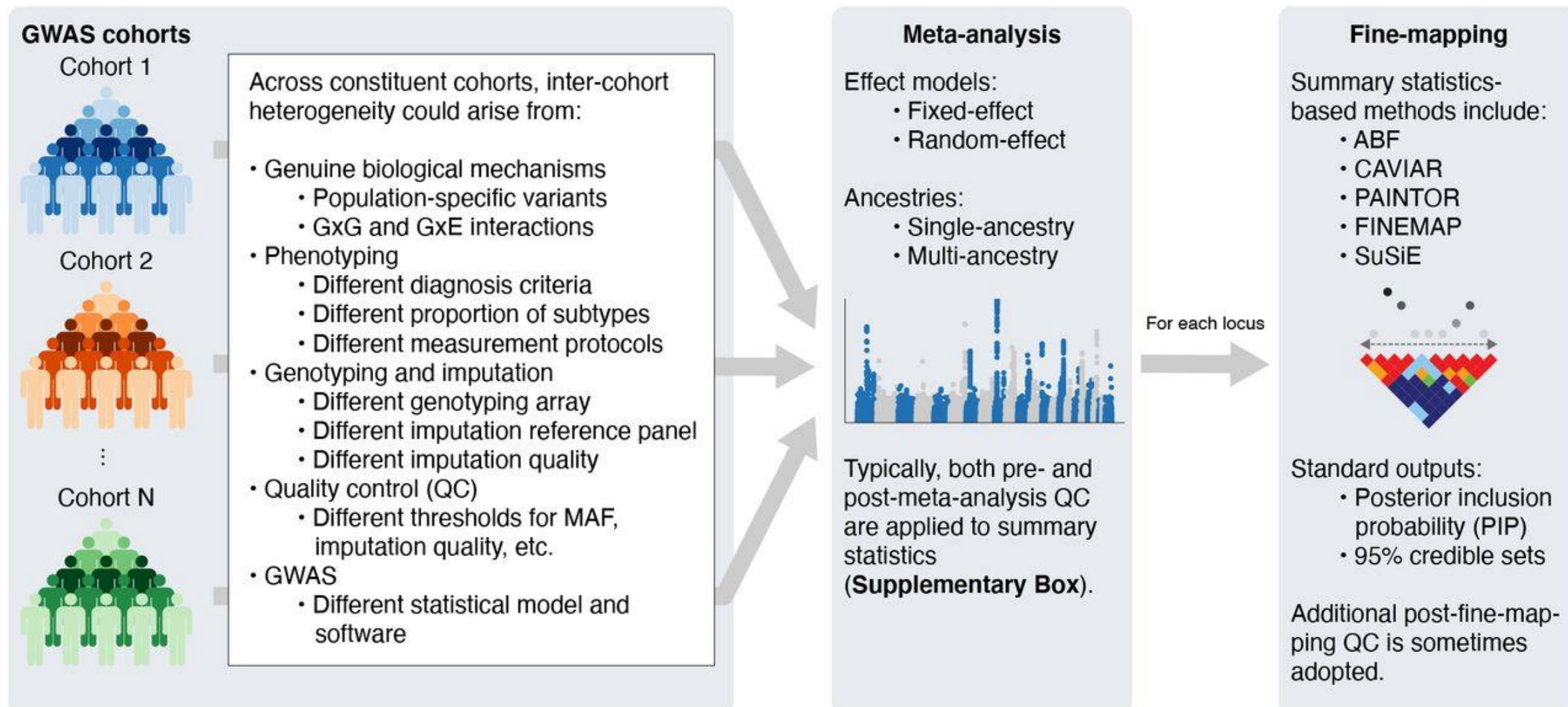
- METAL
- GWAMA
- MANTRA



Credit: Yunye He



# Meta studies



# Meta-analysis

## Approaches

- **Fixed Effects**
  - Most commonly used and most powerful for discovery when assuming a consistent effect of each risk allele across datasets.
    - **Inverse variance weighting** is the most common method.
    - **Sample size weighting** (z-score based) is also widely used.
- **Random Effects**
  - Less common but useful for assessing the generalizability of associations.
  - Estimates the average effect size and its uncertainty across different populations.
- **Bayesian Approaches** (rarely used)



# Meta studies

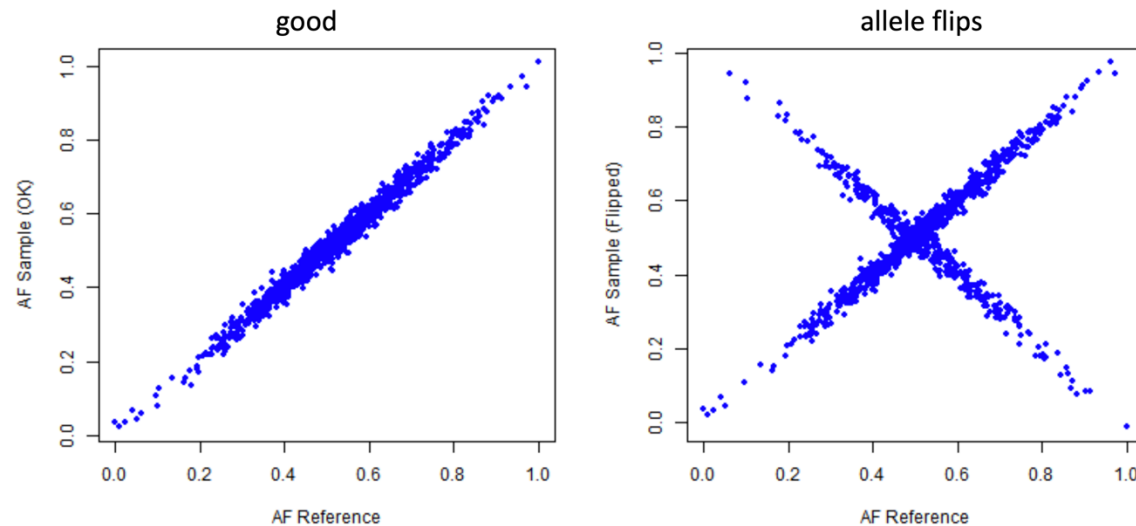
Quality control is crucial!

- Rigorous QC on the individual GWAS results
- Exclude rare variants and poorly imputed variants
- Control for population stratification and ancestry differences
- Verify input data and identify differences (tools: GWAtoolbox, EasyQC, GWASinspector)
- Harmonization of the data (effect allele polarization)
- Perform both fixed effects approaches and compare the results
- As in GWAS, QQ and Manhattan plots are important.

# Quality control

## Allele flipping

- Effect allele must be the same across GWAS studies.
- How does it look if the effect direction is not the same?



# Meta-analysis software

- Most commonly used software for common variant analysis: METAL
  - Automatic strand flipping of non-ambiguous SNPs
  - Calculation of max/min/mean allele frequency
  - Inverse variance & sample size weightings
  - Automatic genomic control correction
  - Heterogeneity tests
- 
- Link: [www.sph.umich.edu/csg/abecasis/metal/](http://www.sph.umich.edu/csg/abecasis/metal/)
  - Documentation: [genome.sph.umich.edu/wiki/Metal\\_Documentation](http://genome.sph.umich.edu/wiki/Metal_Documentation)

# Meta-analysis example!

- Setup

- Modify files to include:

- all information
- consistent **marker name**

- Tools: WAtoolbox, EasyQC, GWASinspector

## Input: Script file

```
# Execute analysis on 2 studies
# GENOMICCONTROL ON
# SCHEME STDERR

#-- DESCRIBE AND PROCESS 1st FILE --
MARKER SNP
ALLELE REF_ALLELE OTHER_ALLELE
EFFECT BETA
PVALUE PVALUE
WEIGHT N
STDERR SE
PROCESS gwas1.txt.gz

#-- DESCRIBE AND PROCESS 2nd FILE --
MARKER SNP
ALLELE A1 A2
EFFECT EFFECT1
PVALUE pvalue
WEIGHT N
STDERR SE
PROCESS gwas2.txt.gz

OUTFILE META_GWAS1-2
MINWEIGHT 10000
ANALYZE HETEROGENEITY
```

## Running METAL

### META\_GWAS1-2.TBL.INFO


# This file contains a short description of the columns  
# meta-analysis summary file, named 'META\_GWAS1-2.TBL'

# Marker - this is the marker name  
# Allele1 - the first allele for this marker in the first file where it occurs  
.  
.  
# Input for this meta-analysis was stored in the files: # --> Input File 1 :  
gwas1.txt.gz  
# --> Input File 2 : gwas2.txt.gz

### META\_GWAS1-2.TBL

MarkerName	Allele1	Allele2	Weight	Zscore	P-value	Direction
rs560887	t	c	6806	-7.075	1.491*10 <sup>-12</sup>	---
rs853787	t	g	6806	6.691	2.221*10 <sup>-11</sup>	+++
rs853789	a	g	5339	-6.597	4.189*10 <sup>-11</sup>	?--
rs853773	a	g	6806	-6.132	8.662*10 <sup>-10</sup>	---
rs537183	t	c	6806	6.007	1.887*10 <sup>-9</sup>	+++
rs557462	t	c	6806	6.005	1.917*10 <sup>-9</sup>	+++
rs502570	a	g	6806	-6.001	1.955*10 <sup>-9</sup>	---
rs563694	a	c	6806	5.975	2.300*10 <sup>-9</sup>	+++
rs475612	t	c	6806	-5.867	4.423*10 <sup>-9</sup>	---
rs853781	a	g	6806	-5.844	5.092*10 <sup>-9</sup>	---

## **GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies**

Stephanie M. Gogarten , Tushar Bhangale, Matthew P. Conomos, Cecelia A. Laurie, Caitlin P. McHugh, Ian Painter, Xiuwen Zheng, David R. Crosslin, David Levine, Thomas Lumley ... [Show more](#)

[Author Notes](#)

*Bioinformatics*, Volume 28, Issue 24, December 2012, Pages 3329–3331, <https://doi.org/10.1093/bioinformatics/bts610>

- Ensures consistency of input file columns
- Compares effect size distributions across cohorts
- Harmonized header and separator across input files
- Calculated effective N and corrects for genomic control

