Detecting loci under coevolution using GWAS

Miaoyan Wang

University of Wisconsin – Madison, USA

ESEB-STN 2019 workshop
Technical University of Munich

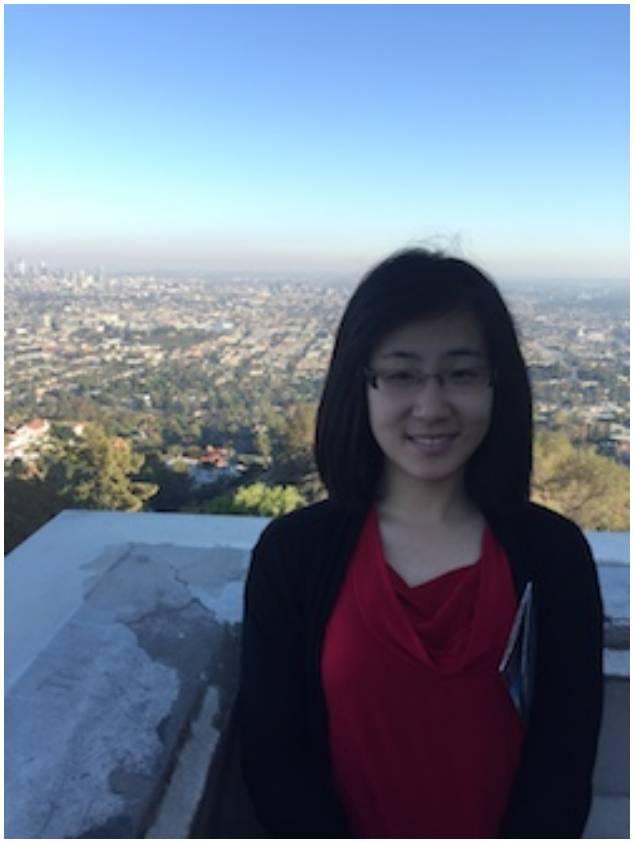
March 27, 2019

Introduction: session aim

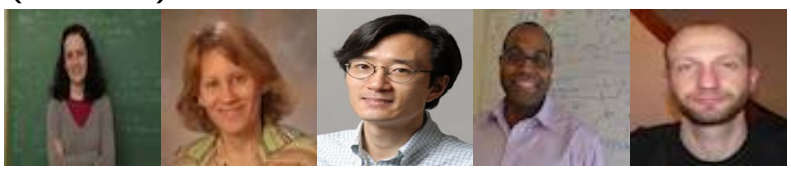
This is a session on computational methods for genetic association studies of complex traits. We aim to cover:

- Key ideas for Genetic Association Studies (GWAS)
- Population Structure/Ancestry Inference
- Joint Association Analyses Using Both Host and Pathogen Genomes.

Introduction: about me



- Assistant professor in Statistics at University of Wisconsin Madison, USA
- Past experiences:
  - Postdoc in Computer Science at UC Berkeley
  - Simons Math + Biology visitor at University of Pennsylvania
  - PhD in Statistics at UChicago, B.S in Mathematics
- Research interests: population genetics, complex traits; information theory, machine learning.
- Acknowledge: Mary Sara McPeek (UChicago), Joy Bergelson (UChicago), Yun S. Song (UC Berkeley), Tim Thornton (U Washington), Fabrice Roux (CNRS)



Introduction: resources

Importantly, the class site is
`http://www.stat.wisc.edu/~miaoyan/ESEB.html`.

- PDF copies of slides
- Datasets needed for exercises
- Exercises for you to try
- Links to software packages

Outline

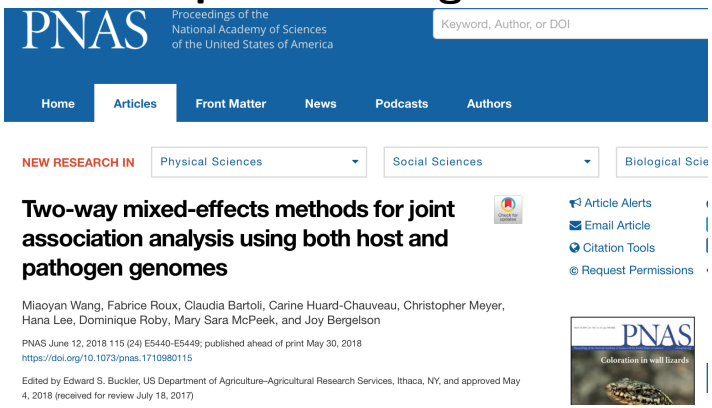
- Introduction
  - Motivation
  - Introduction to genetic association studies (GWAS)
- Topic I: Population structure inference (80 mins)
  - Principal component analysis
  - Supervised learning for ancestry admixture
- Topic II: Genetic association analysis (80 mins)
  - Linear mixed effects model
  - Interaction analysis
  - Advanced mixed method

What to expect in a typical session:

- 40 mins lecture
- 25 mins hands-on exercises
- 15 mins discussion

Suggested Literature

- D. Jiang and M. Wang. (2018) Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Studies of Complex Traits. Biostatistics and Epidemiology. Vol. 2 (1), 132-159, 2018. **A monograph on recent development of GWAS methods.** `https://www.tandfonline.com/eprint/YKvZBnbM54fkwZ5wADgk/full`
- M. Wang et al. (2018) Two-way Mixed-Effects Methods for Joint Association Analyses Using Both Host and Pathogen Genomes. PNAS. Vol. 115 (24), E5440-E5449, 2018. **A recent study on co-evolution using joint GWAS approach.**
- Nature Genetics. (2008-2013) Genome-wide association studies. **Series about best practices for doing GWAS.** `http://www.nature.com/nrg/series/gwas/index.html`
- Lynch and Walsh. (1998) Genetics and Analysis of Quantitative Traits. **A classical reference for quantitative geneticists.**



Introduction to genetic association studies (GWAS)

Motivation

- Identifying large amounts of associations efficiently is a problem that arises frequently in modern genomics data.
  - Understand the genetics of important traits, e.g. traits with medical or agricultural relevance.
  - Identifying the genomic regions that control genetic variation
  - Identifying expression QTLs
  - Cancer genetics, for identifying problematic mutations
  - Understand interaction between genotypes and the environment.
- As genomics datasets become more common and sample sizes grow, the need for efficient tests increases.
- Test association at many variants instead of some and hypothesis-free instead of hypothesis-driven.

Genomic marker

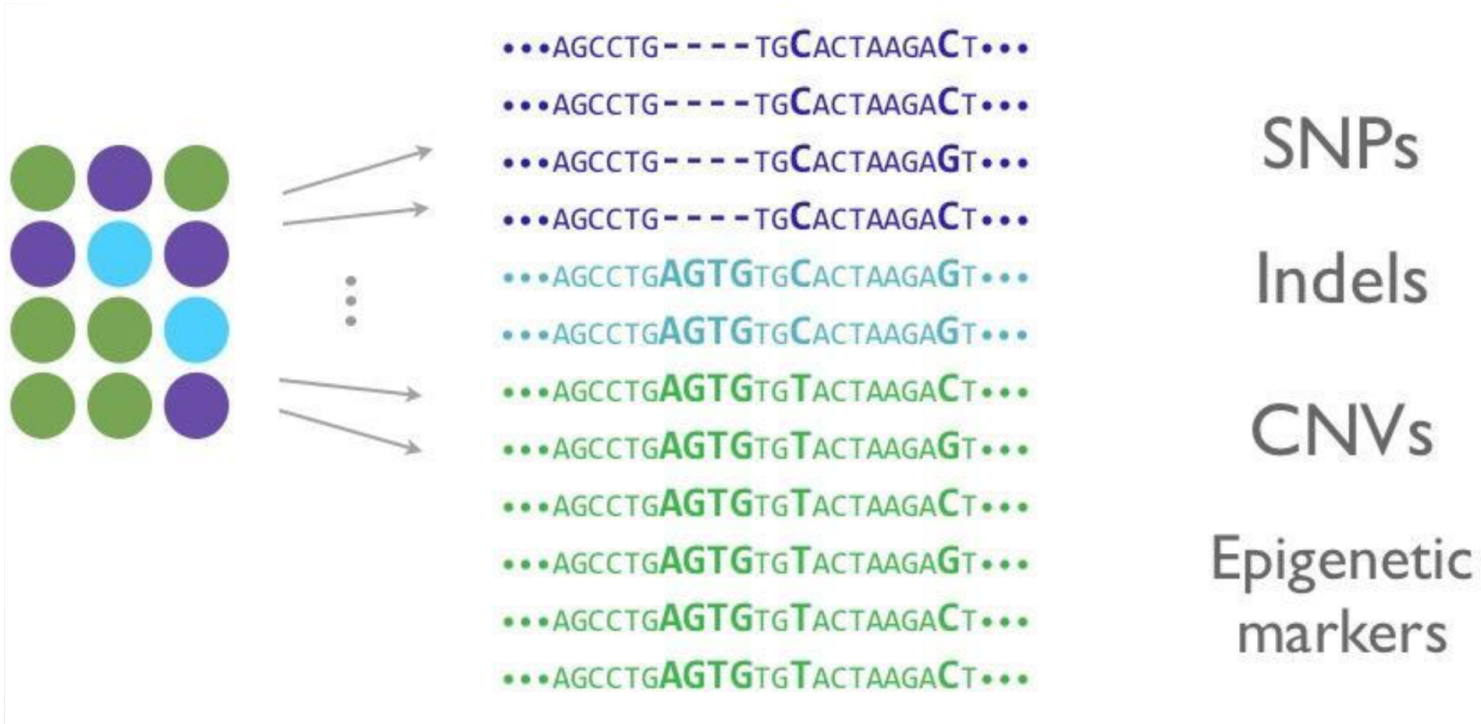
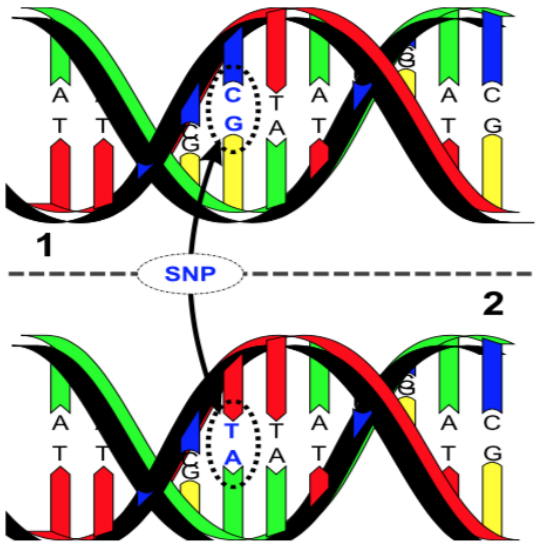


Figure source: Exploring Plant Variation Data Workshop 2015. Ümit Seren.

For this talk

- SNP (single nucleotide polymorphism): site in genome with single base-pair change that distinguishes some individuals from others.
- SNP is just one type of genetic variants. Other examples include inserts, deletions (Indels), and copy number variation (CNV).

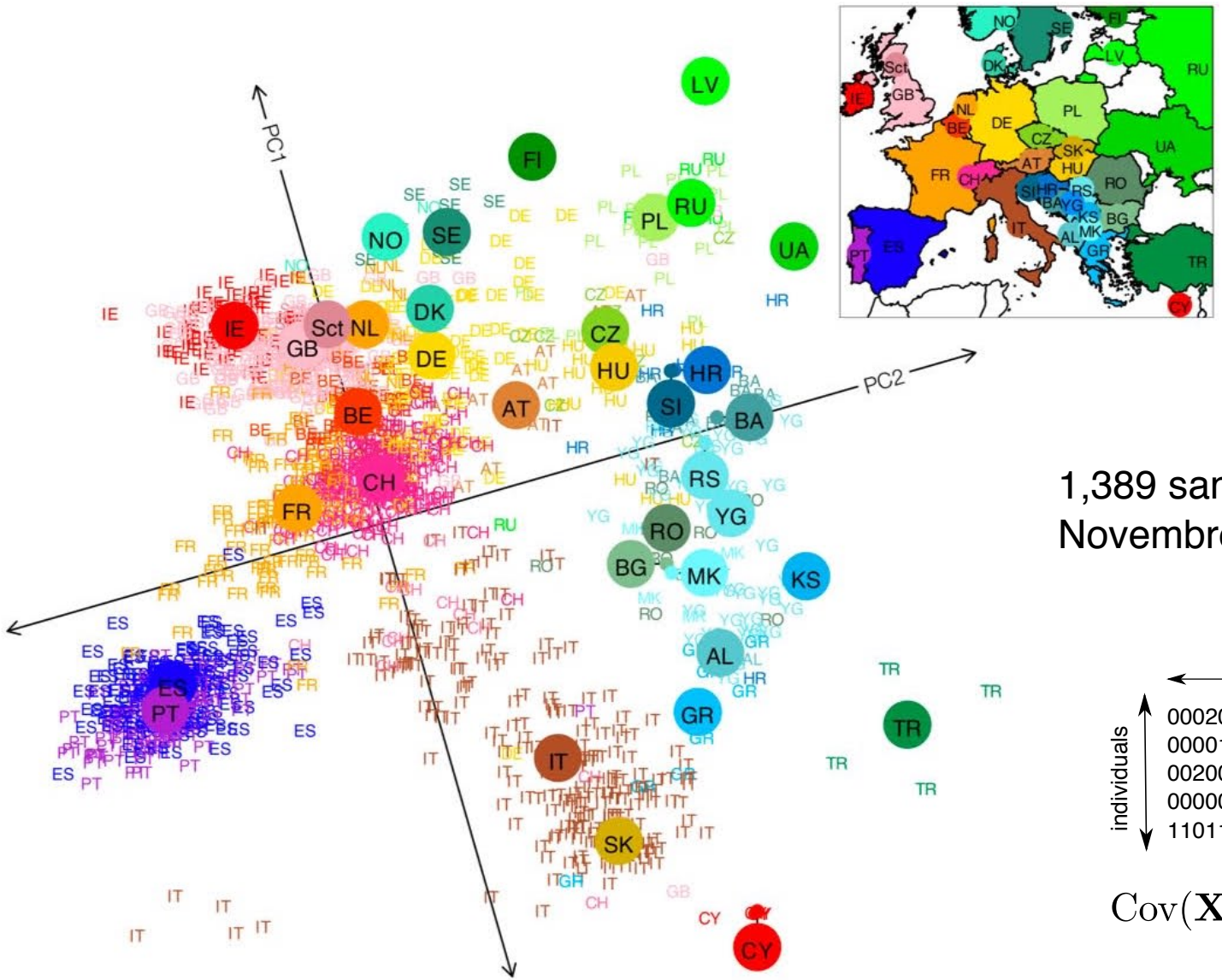


C/T polymorphism

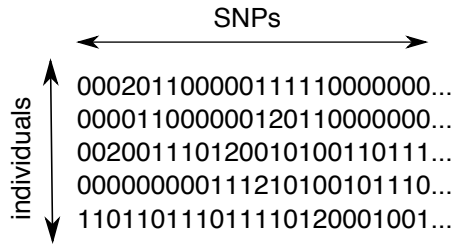
Source: David Hall, Wikipedia

- **Genotype** counts the number of copies of each allele at a SNP hold by individual, e.g. $\{0, 1, 2\}$ for a diploid organism.

Genotypes mirrors geography



1,389 samples, ~ 200k SNPs
Novembre et al. (2008)



$$\mathrm{Cov}(\mathbf{X}) = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Phenotype

- Phenotype = Genotype + Environment + Genotype × Environment

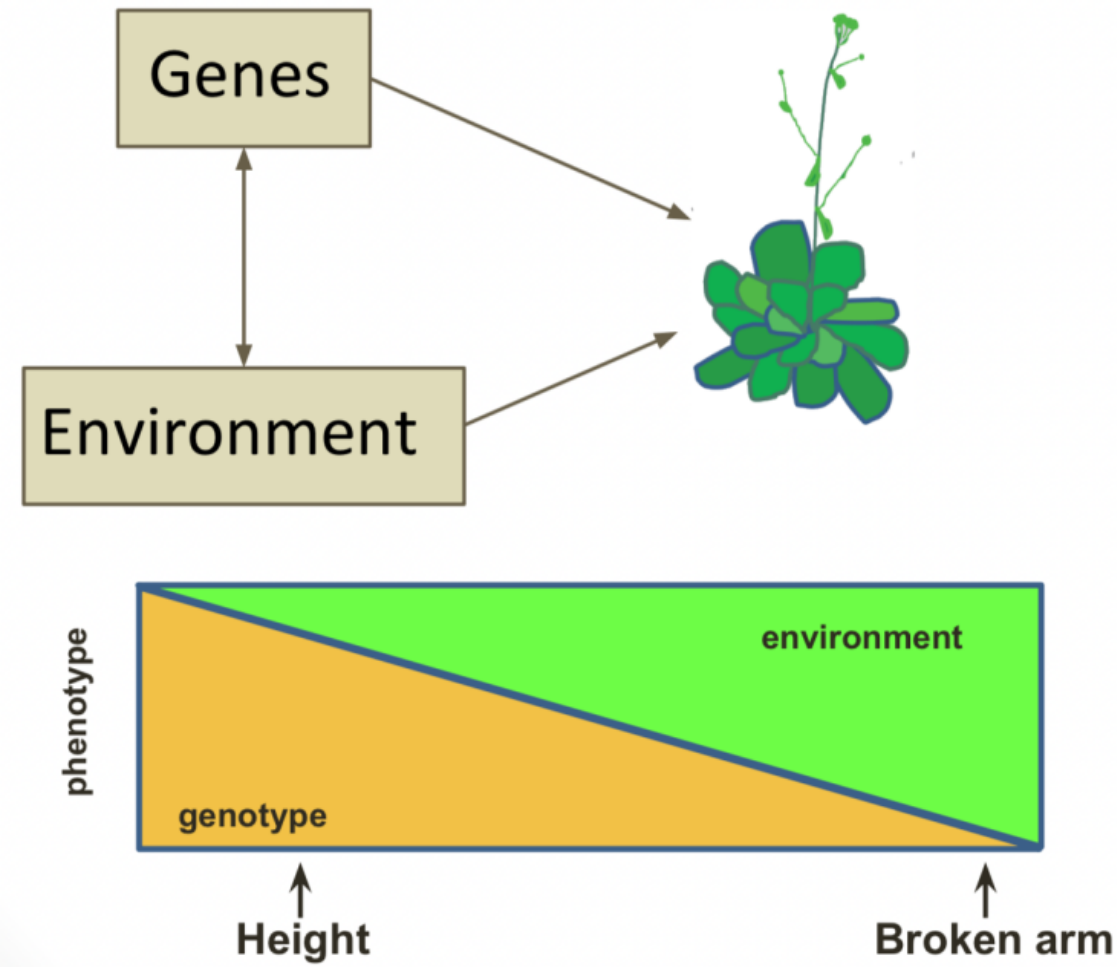


Figure source: Exploring Plant Variation Data Workshop 2015. Ümit Seren.

A typical GWAS pipeline

The primary goal of GWAS is to identify genetic variants that contribute towards the phenotypic variation of complex traits. A typical GWAS involves at least the following three broadly defined steps:

- data quality control
- association testing (will be discussed later)
- results interpretation

Data quality control

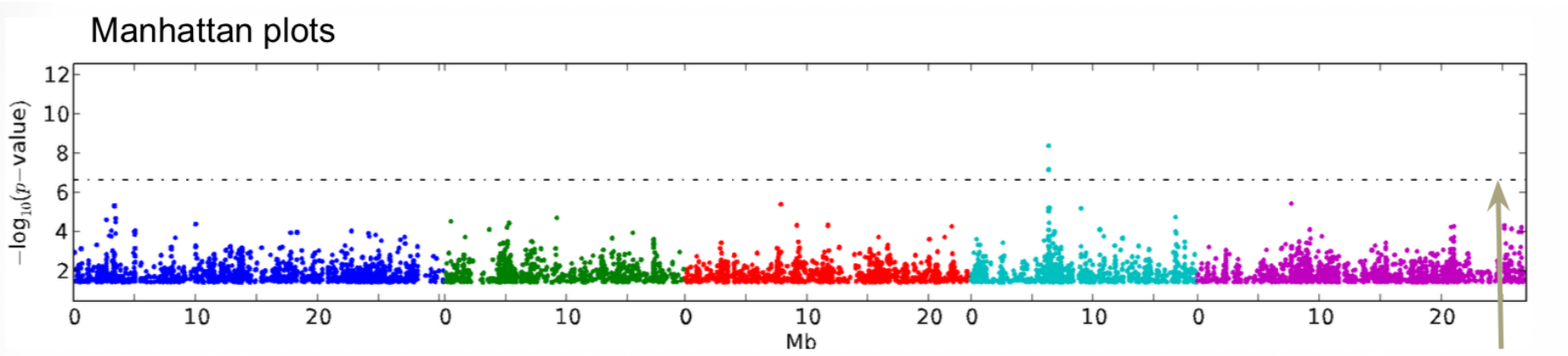
Quality control (QC) usually involves filtering out (i.e., removing) SNPs with low genotype accuracy. Common SNP filters include

- Missing call rate (MCR)
- Minor allele frequency (MAF)
- Hardy-Weinberg equilibrium (HWE)

Genotype imputation is often carried out in GWAS to allow better use of the typed SNPs.

Interpreting association results

- Statistical analysis is performed to detect the association between a SNP and a trait.
- Each SNP will produce a test statistic measuring its association with the trait of interest and a p-value measuring the statistical significance.
- Manhattan and quantile-quantile (Q-Q) plots are useful tools for visualizing GWAS results



GWAS - a successful story

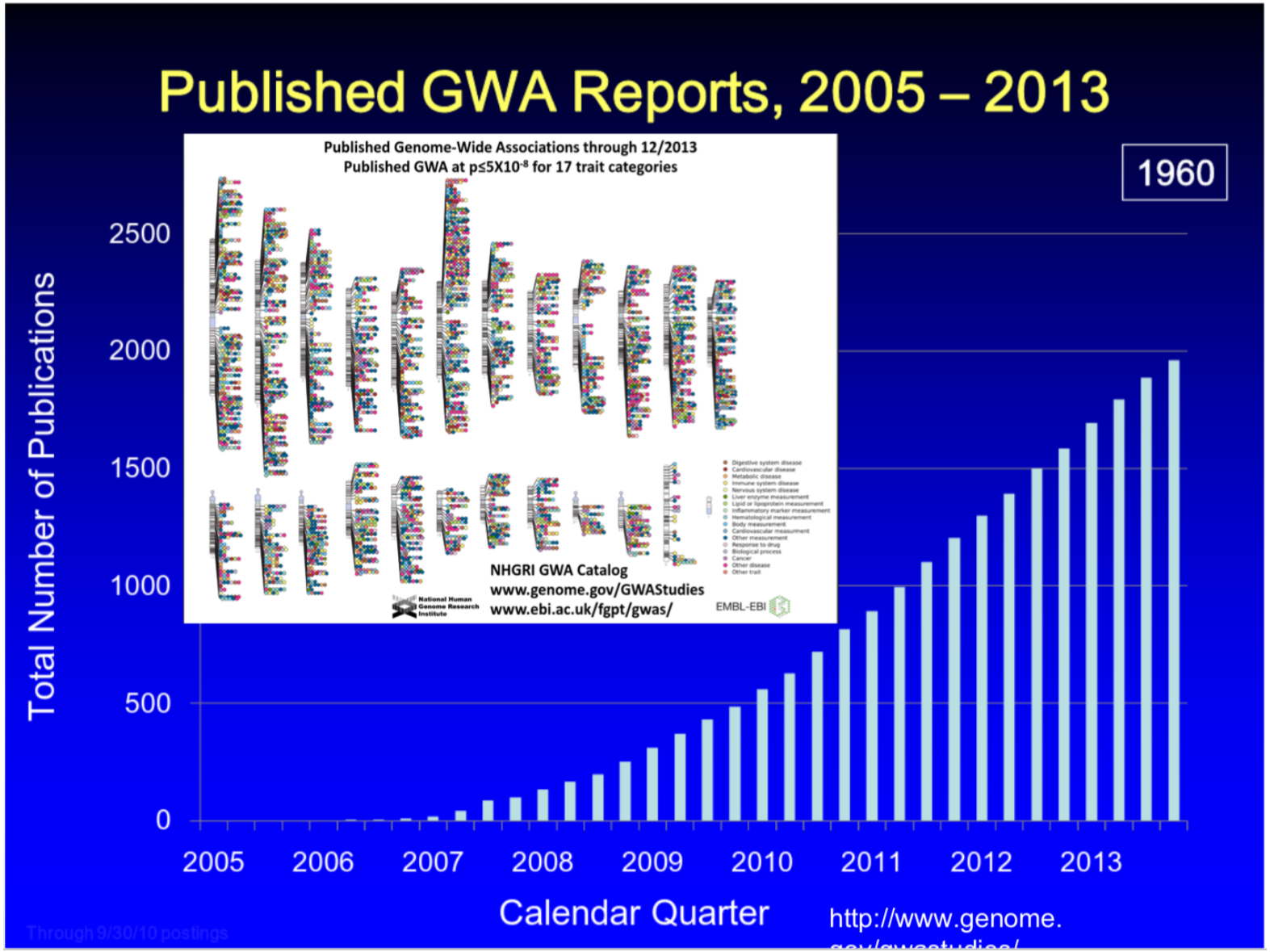
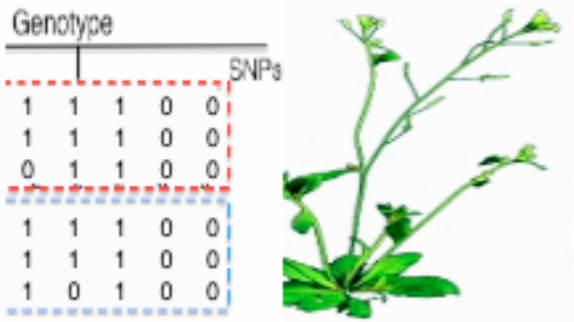


Figure source: National Human Genome Research Institute

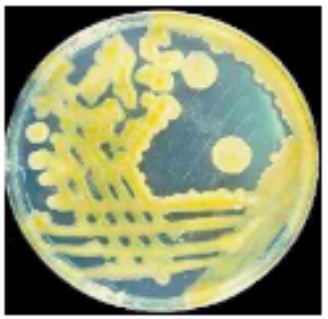
Recent advances in GWAS for co-evolution

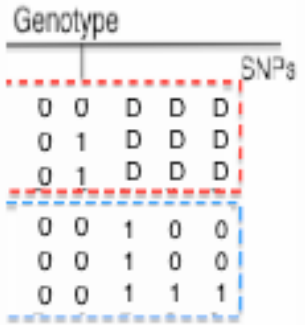
Some complex traits (e.g., infection) depend on the specific pairing of host and pathogen, and therefore on their genomes jointly.



$n_h = 130$ *Arabidopsis* accessions

Sequence data available from 1001 Genomes Project

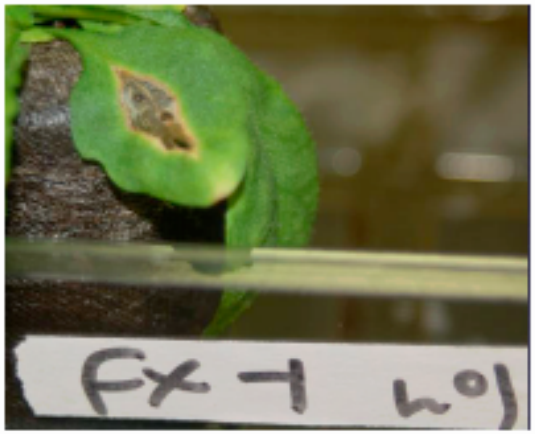




(D=Deletion)

$n_p = 22$ *Xanthomonas* strains

Whole-genome sequence data available



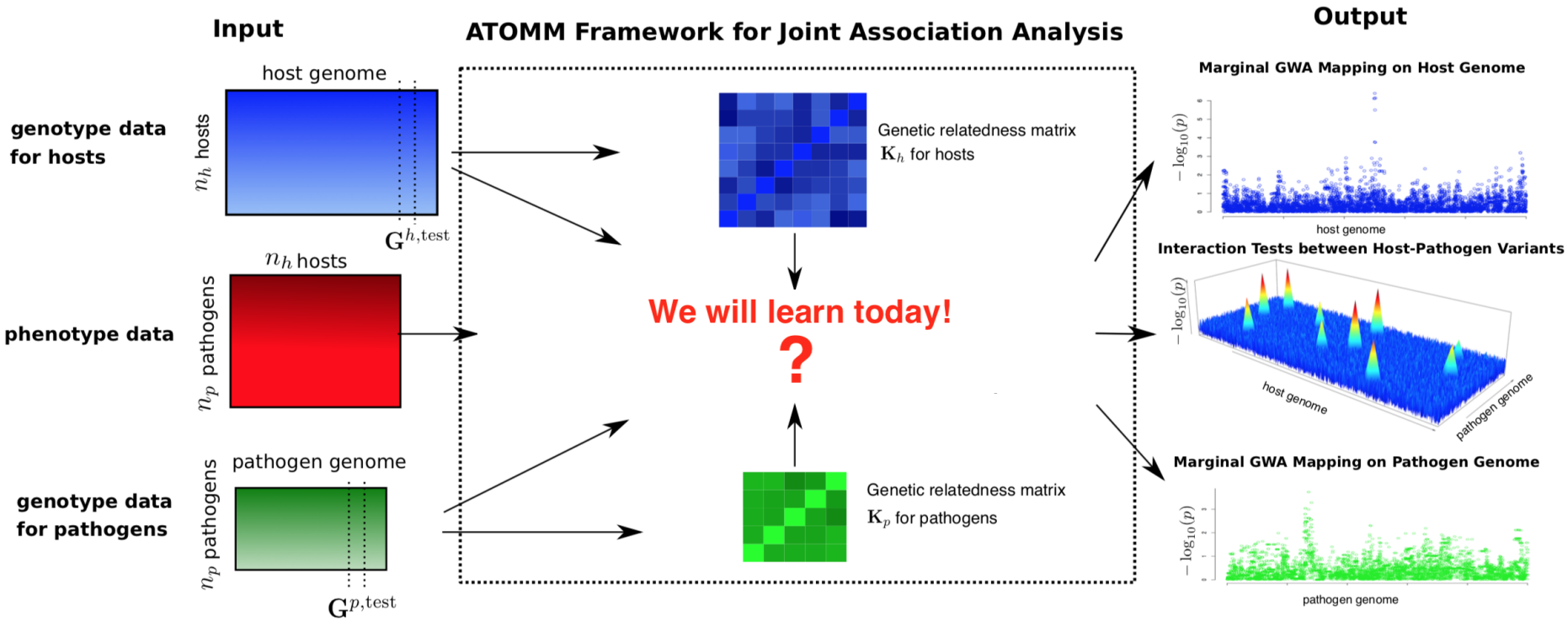
3 replicates (plants) for each accession-strain pair, 2-4 leaves on each plant

poke 3 holes on each leaf; define disease score as # {holes show infection symptoms }

$\mathbf{Y} \in \{0, 1, 2, 3, 4\}^n$ n=32,960

Joint GWAS for co-evolution

Recent research shows that GWAS can be used to test for association and gene-gene interaction in a co-evolution system that involves two interactive organisms. (M. Wang, et al. *PNAS*. Vol. 115 (24), (2018) E5440-E5449.)

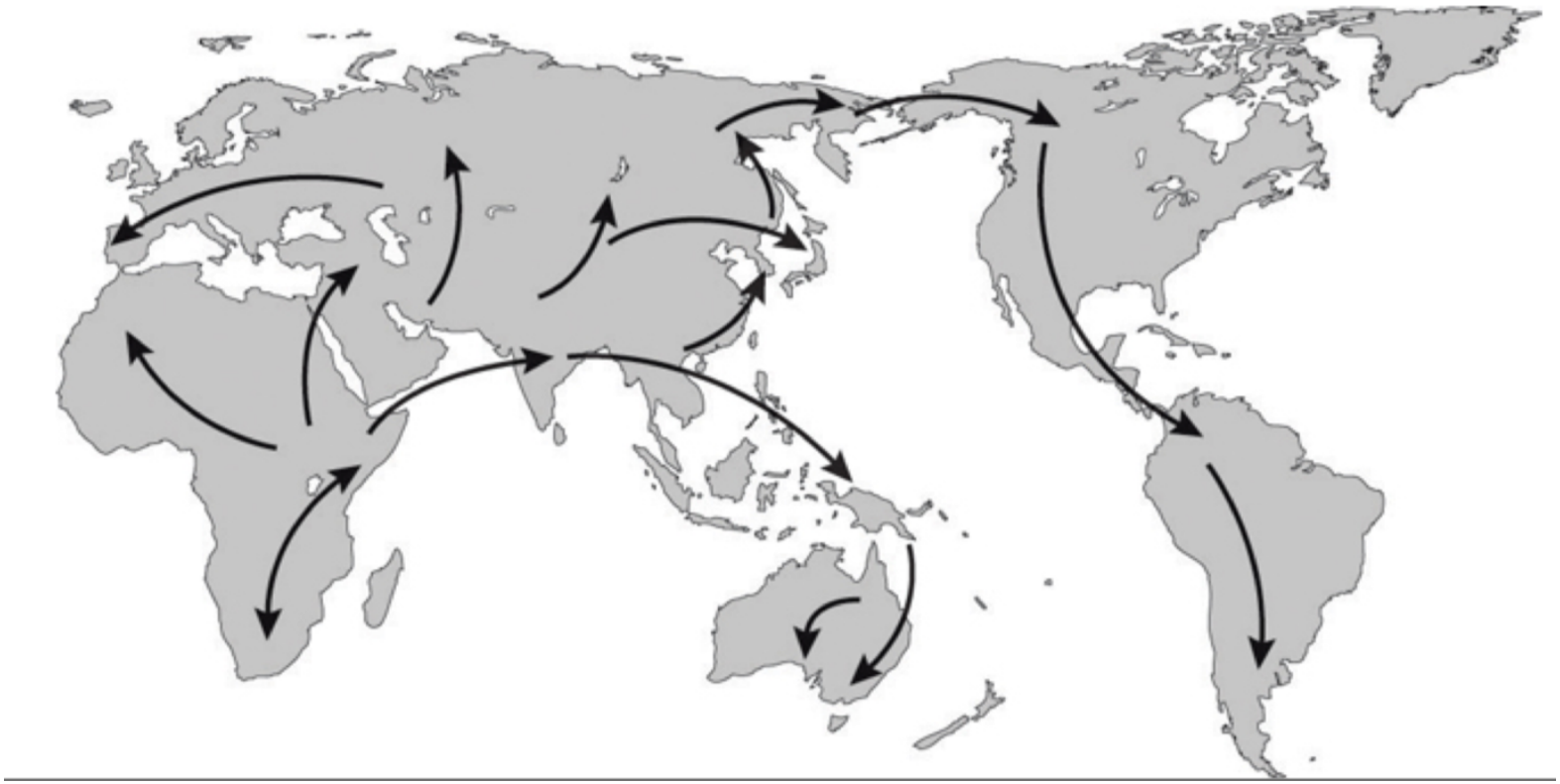


Outline

Section I: Population structure inference

Background: Population structure

- Many organisms (humans, Arabidopsis) spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between groups.



Population structure inferences

- Inference on genetic ancestry differences among individuals from different populations, or **population structure**, has been motivated by a variety of applications:
  - population genetics
  - genetic association studies
  - personalized medicine
  - forensics
- Advancements in genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail.
- A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.

Inferring Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA applied to genotype data can be used to calculate **principal components (PCs)** that explain differences among the sample individuals in the genetic data
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- PCA is an unsupervised learning tool for dimension reduction in multivariate analysis.

Data structure

- Sample of n individuals, indexed by $i = 1, 2, \ldots, n$.
- Genome screen data on m genetic autosomal markers, indexed by $\ell = 1, 2, \ldots, m$.
- At each marker, for each individual, we have a genotype value $x_{i\ell}$.
- Here we consider bi-allelic SNP data, so $x_{i\ell}$ takes values 0, 1, or 2, corresponding to the number of reference alleles.
- We center and standardize these genotype values:

$$z_{i\ell} = \frac{x_{i\ell} - 2\hat{p}_\ell}{\sqrt{2\hat{p}_\ell(1 - \hat{p}_\ell)}},$$

where $\hat{p}_\ell$ is an estimate of the reference allele frequency for marker l.

Genetic Correlation Estimation

- Create an $n \times m$ matrix, Z, of centered and standardized genotype values, and from this, a genetic correlation matrix (GRM):

$$\Phi = \frac{1}{m} Z Z^T$$

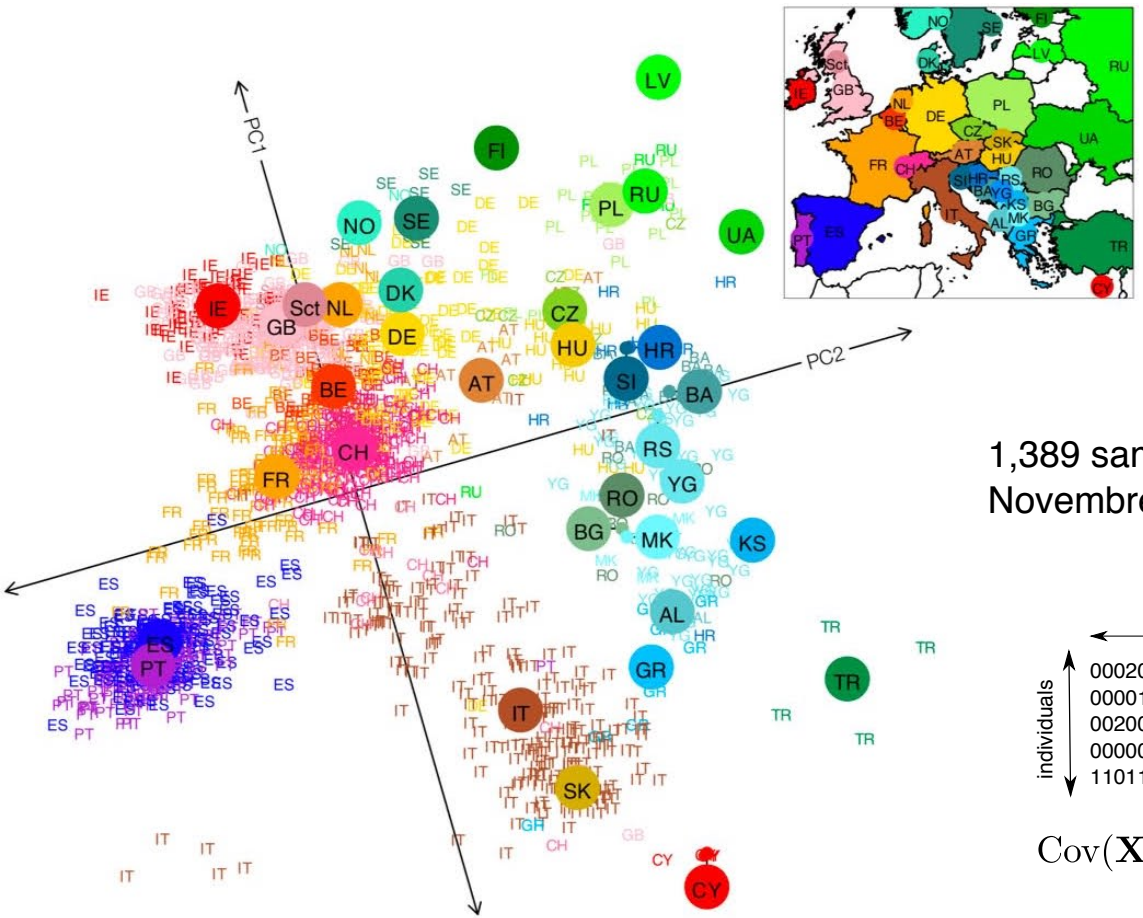
- $\hat{\Phi}_{ij}$ is an estimate of the genome-wide average genetic correlation between individuals i and j.
- PCA relies on individuals from the same ancestral population being more genetically correlated than individuals from different ancestral populations.

Standard Principal Components Analysis (PCA)

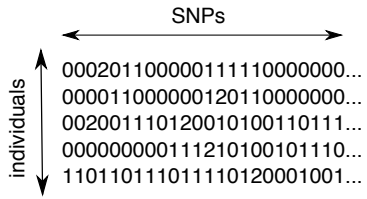
- PCA is performed by obtaining the eigen-decomposition $\hat{\Phi}$.
- Top eigenvectors (PCs) are used as surrogates for population structure.
- Orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability amongst the n sample individuals are identified.
- Individuals with "similar" values for a particular top principal component tend to have "similar" ancestry.

PCA of Europeans

An application of principal components to genetic data from European samples showed that the first two principal components computed using 200K SNPs could map their country of origin accurately.



1,389 samples, ~ 200k SNPs
Novembre et al. (2008)

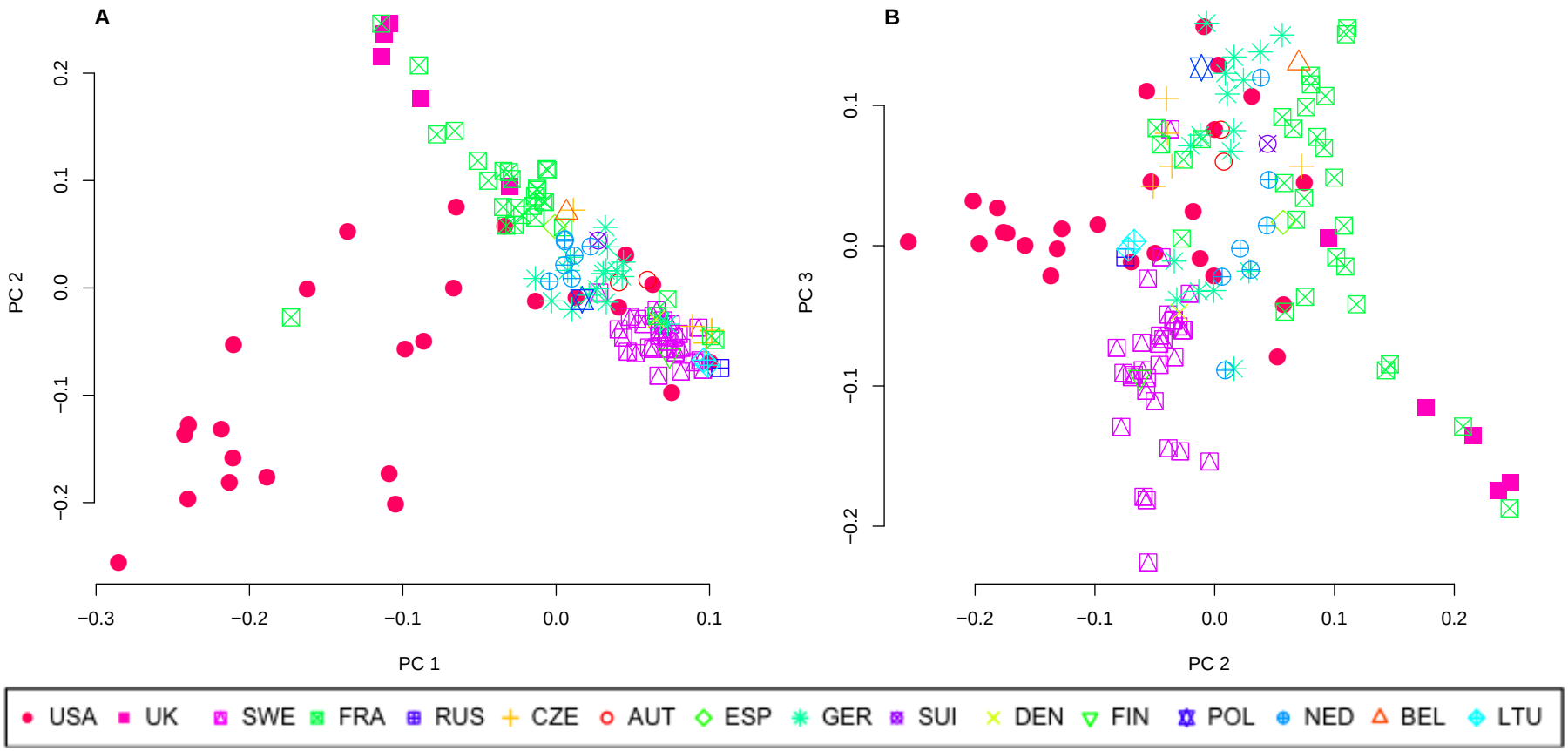


$$\mathrm{Cov}(\mathbf{X}) = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

Population structure among Arabidopsis (host) sample

An application of PCA to genetic data from 1001 Arabidopsis project largely captures the geographical origins of the *Arabidposis* accessions:

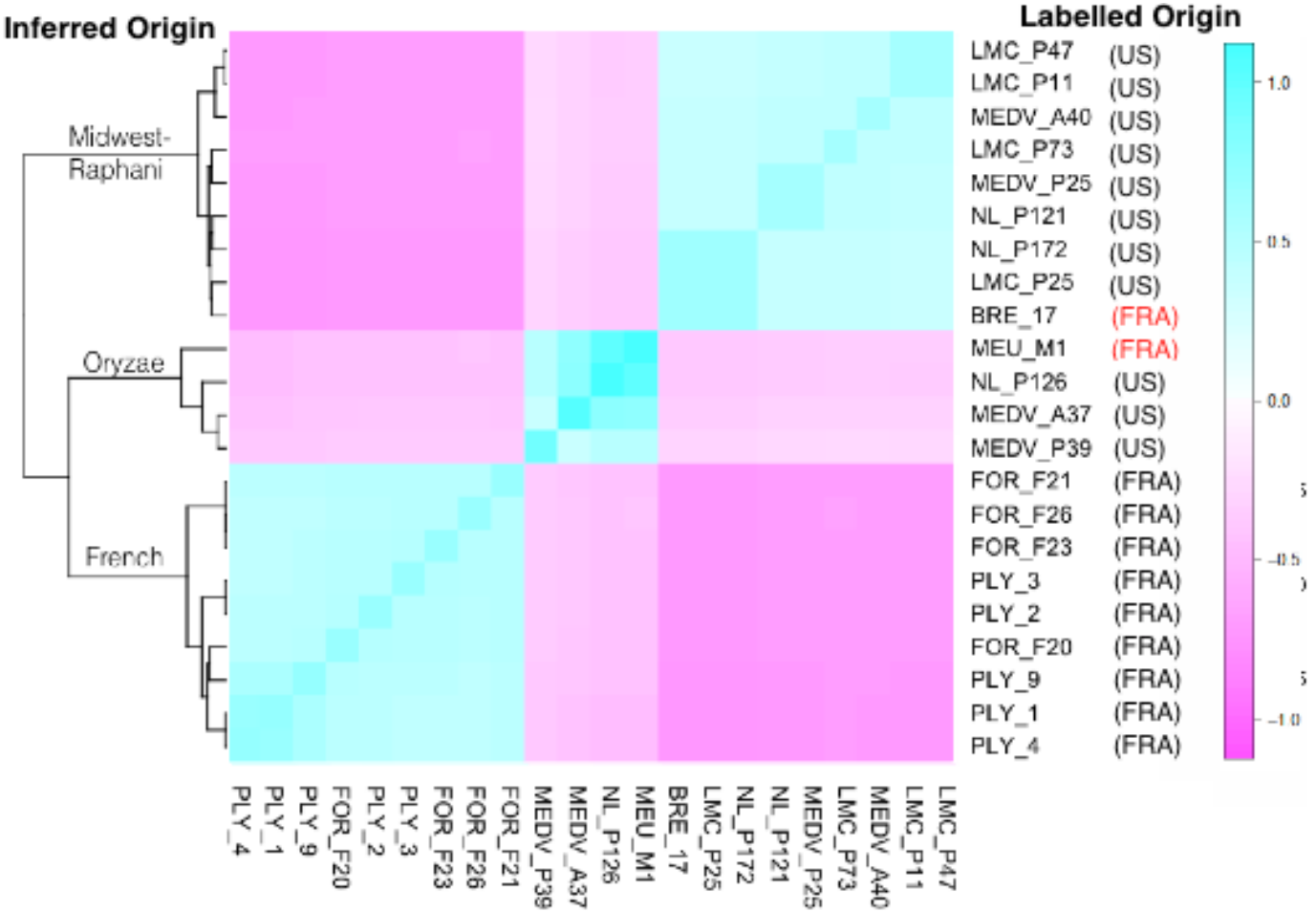
- US vs. European
- Smaller regional groups among European accessions



Population structure among pathogen sample

We develop a method for genetic correlation matrix (GRM) estimation using both mutation and deletion polymorphisms. [PNAS. Vol. 115 (24), 2018.]

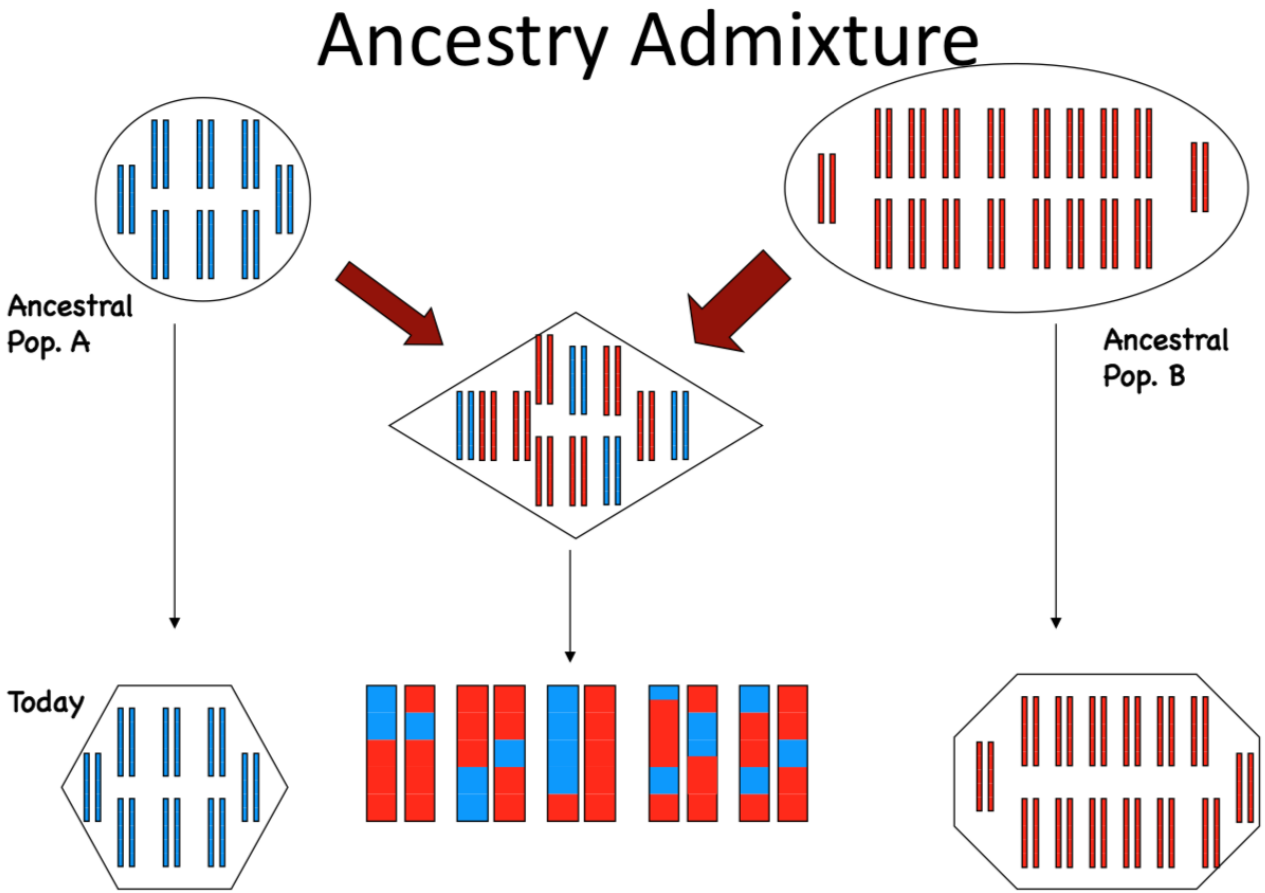
- GRM can be used for clustering analysis.
- *Xanthomonas* sample exhibits strong population stratification.



Admixed Populations

- Several recent and ongoing genetic studies have focused on **admixed populations**: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated.
- Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations.
- Examples of admixed populations include
  - African Americans and Hispanic Americans in the U.S
  - Latinos from throughout Latin America
  - Uyghur population of Central Asia
  - Cape Verdeans
  - South African "Coloured" population

Admixed Populations



- The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.

Supervised Learning for Ancestry Admixture

- Methods such as STRUCTURE (Pritchard et al, 2000) and ADMXITURE (Alexander et al,. 2009) have recently been developed for supervised learning of ancestry proportions for an admixed individuals using high-density SNP data.
- Most use either a hidden Markov model (HMM) or an Expectation-Maximization (EM) algorithm to infer ancestry
- Example: Suppose we are interested in identifying the ancestry proportions for an admixed individual

Supervised Learning for Ancestry Admixture

- Observed sequence on a chromosome for an admixed individual:

...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...

- Observed sequence on a chromosome for samples selected from a "homogenous" reference population:

...TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

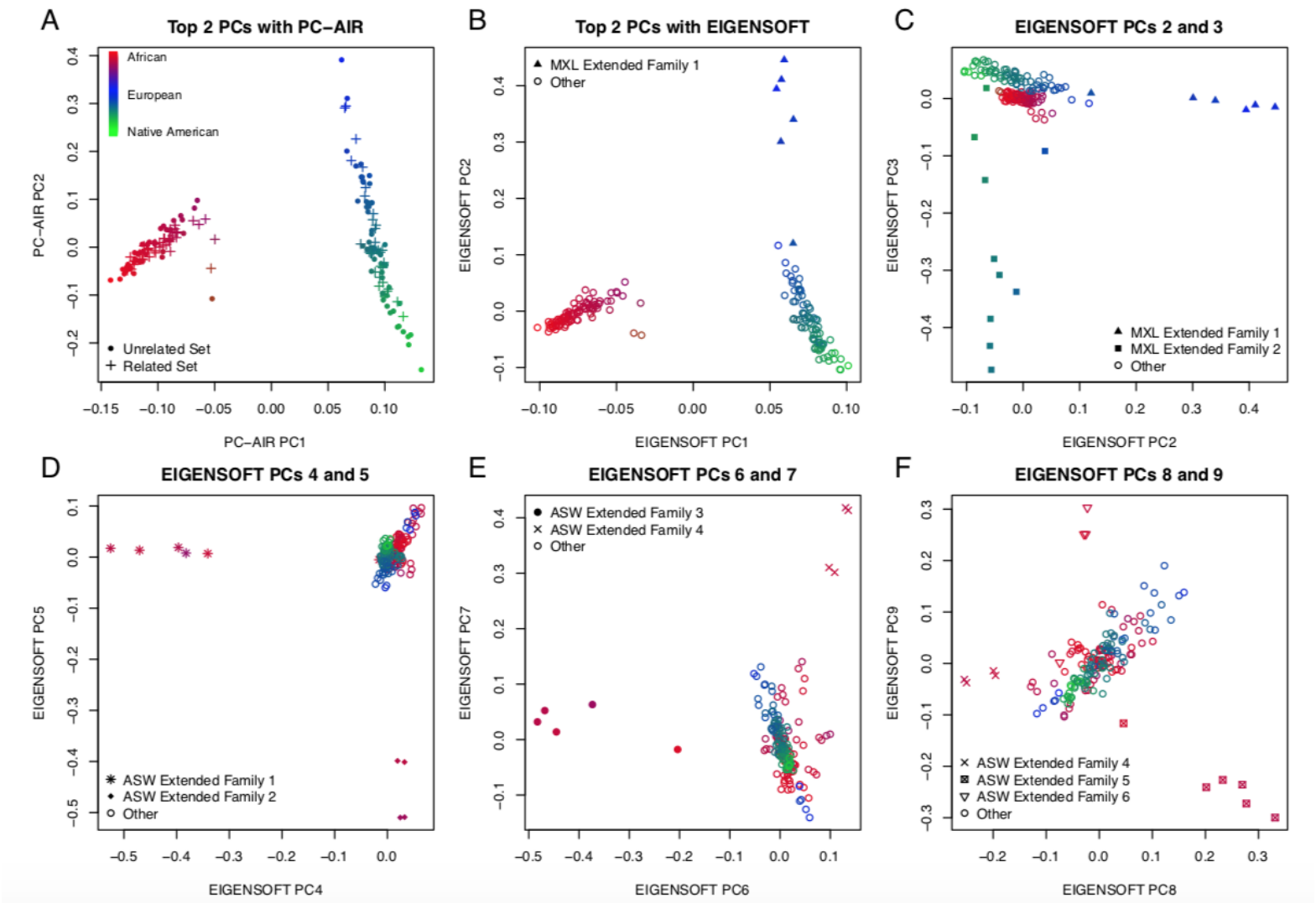
...AGATCCTGAACCTAGATTACAGATTACAGATACCAATGCTTCGATGGAC...

...CGATCCTGAACCTAGATTACAGATTTGCGTATACAATGCTTCGATGGAC...

HapMap ASW and MXL Ancestry

- Genome-screen data on 150,872 autosomal SNPs was used to estimate ancestry
- Estimated genome-wide ancestry proportions of every individual using the ADMIXTURE (Alexander et al., 2009) software
- A supervised analysis was conducted using genotype data from the following reference population samples for three "ancestral" populations
  - HapMap YRI for West African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry

Figure: **HapMap MXL + ASW Sample**



Conomos, Matthew P et al. Genetic epidemiology 39.4 (2015): 276-293.

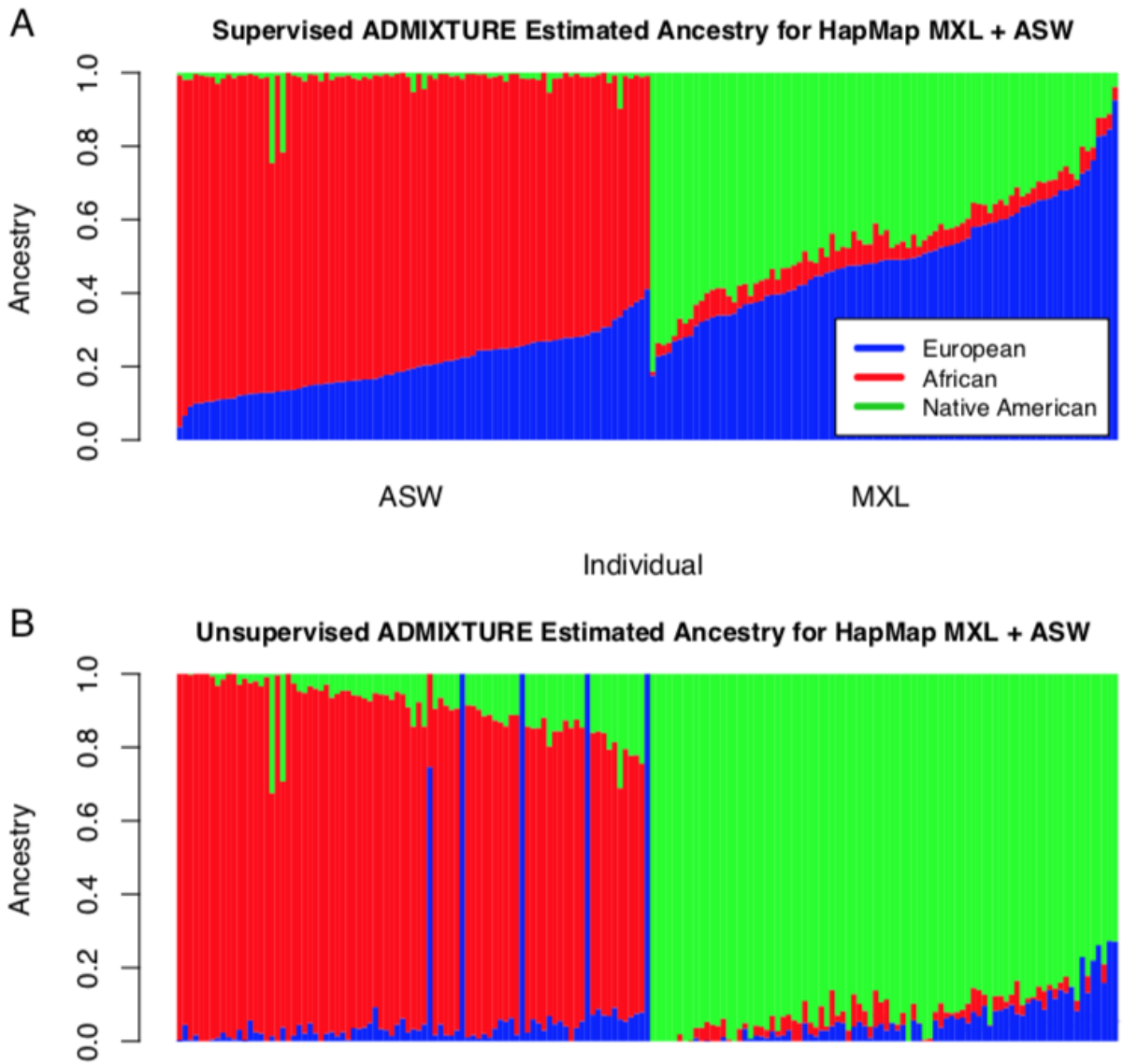


Figure source: SISG 2017. Timothy Thornton and Michael Wu.

Table: Average Estimated Ancestry Proportions for HapMap African Americans and Mexican Americans

	Estimated Ancestry Proportions (SD)		
Population	European	African	Native American
MXL	49.9% (14.8%)	6%(1.8%)	44.1% (14.8%)
ASW	20.5% (7.9%)	77.5% (8.4%)	1.9% (3.5%)

Table source: SISG 2017. Timothy Thornton and Michael Wu.

Topic 2: Genetic association studies in structured population

Association analysis

- In the previous session, we gave an overview of genome-wide association studies (GWAS).
- Association analysis involves identifying genetic loci that influence the phenotypic variation of a quantitative trait.
- Association analysis is commonly conducted with GWAS using common variants, such as variants with minor allele frequencies $\geq 1\%$ - 5%
- Some quantitative traits can be largely influenced by a single gene as well as by environmental factors or gene-gene interaction.

Association analysis

- The classical quantitative genetics model introduced by Ronald Fisher (1918) is $Y = G + E$, where Y is the phenotypic value, G is the genetic value, and E is the environmental deviation.
- G is the combination of all genetic loci that influence the phenotypic value and E consists of all non-genetic factors that influence the phenotype

Heritability

- The broad-sense heritability is defined to be

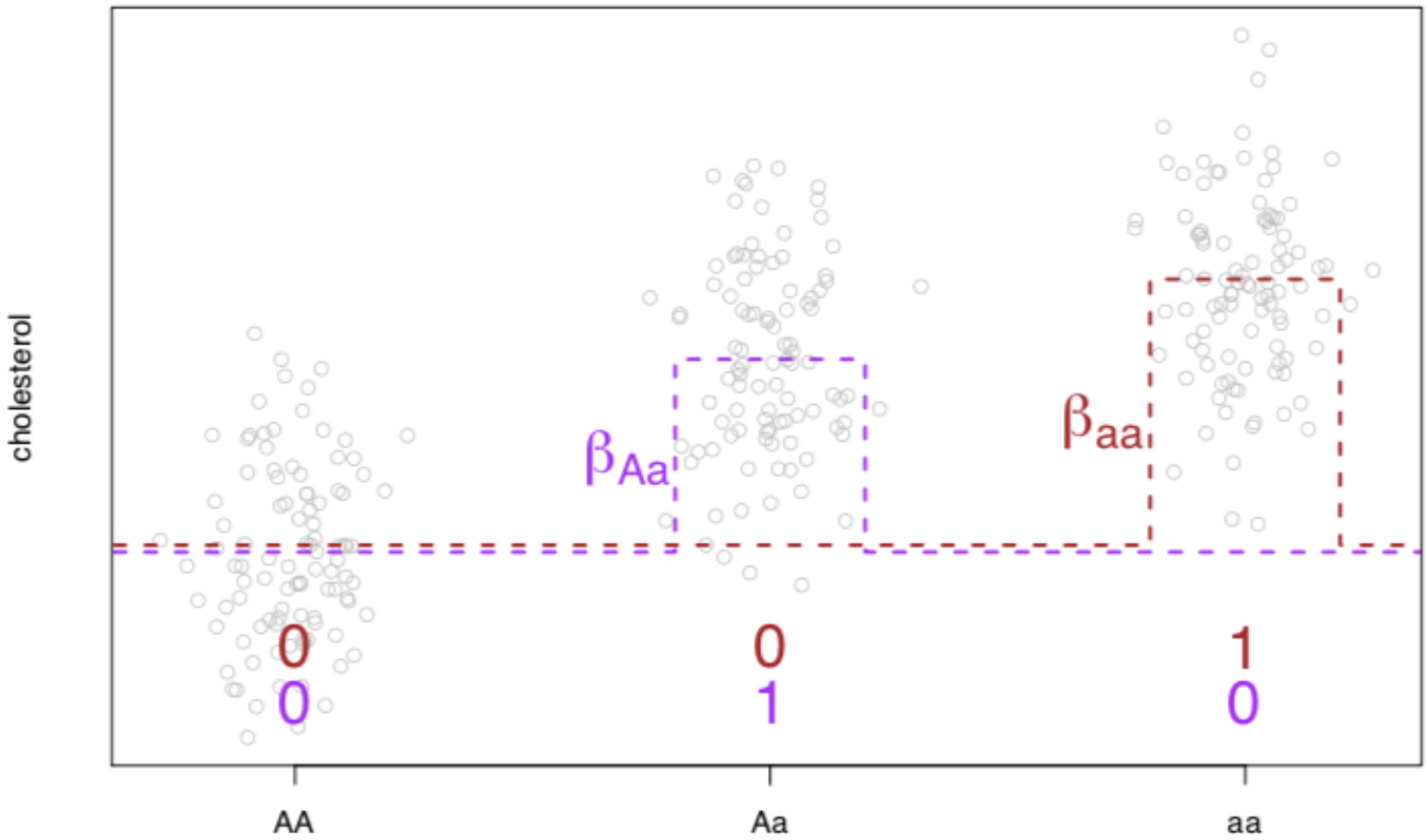
$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2}$$

- H^2 is the proportion of the total phenotypic variance that is due to all genetic effects (additive and dominance)
- There are a number of methods for heritability estimation of a trait.

Linear regression with SNPs

The "two degrees of freedom model":

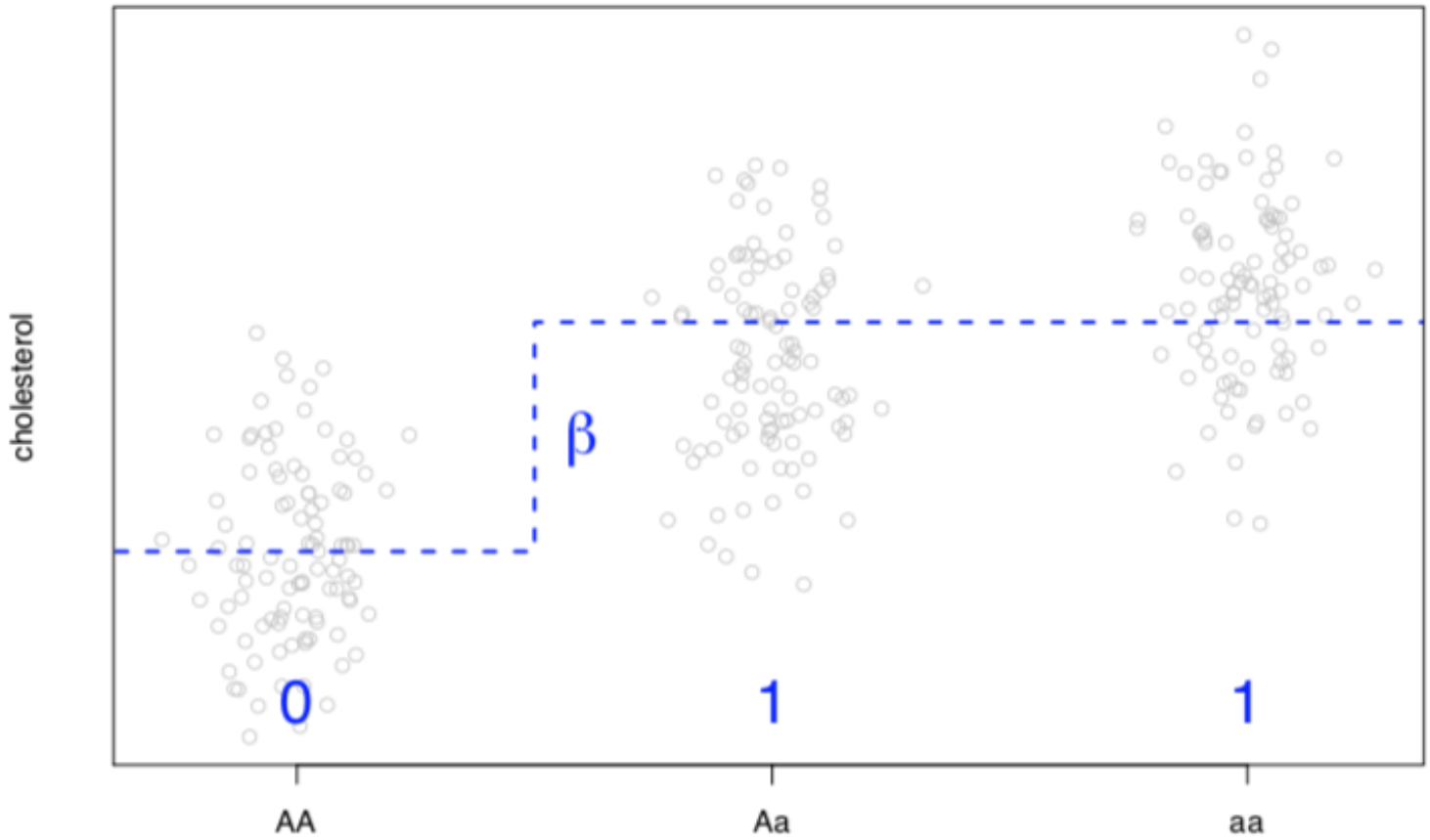
$$E(Y) = \beta_0 + \beta_{Aa} \times (G == Aa) + \beta_{aa} \times (G == aa)$$



Linear regression with SNPs

An alternative is the "dominant model":

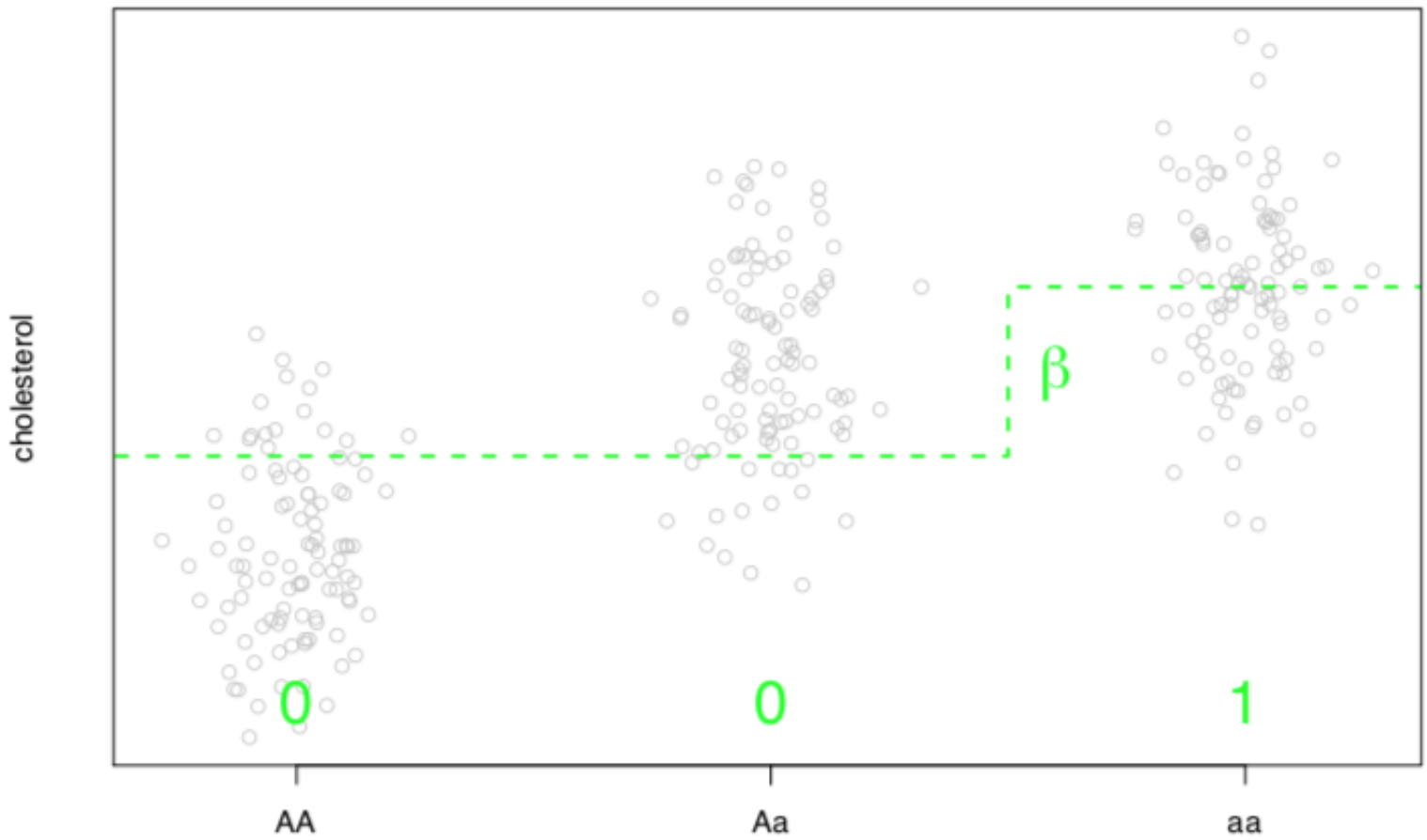
$$E(Y) = \beta_0 + \beta \times (G \neq AA)$$



Linear regression with SNPs

or the "recessive model":

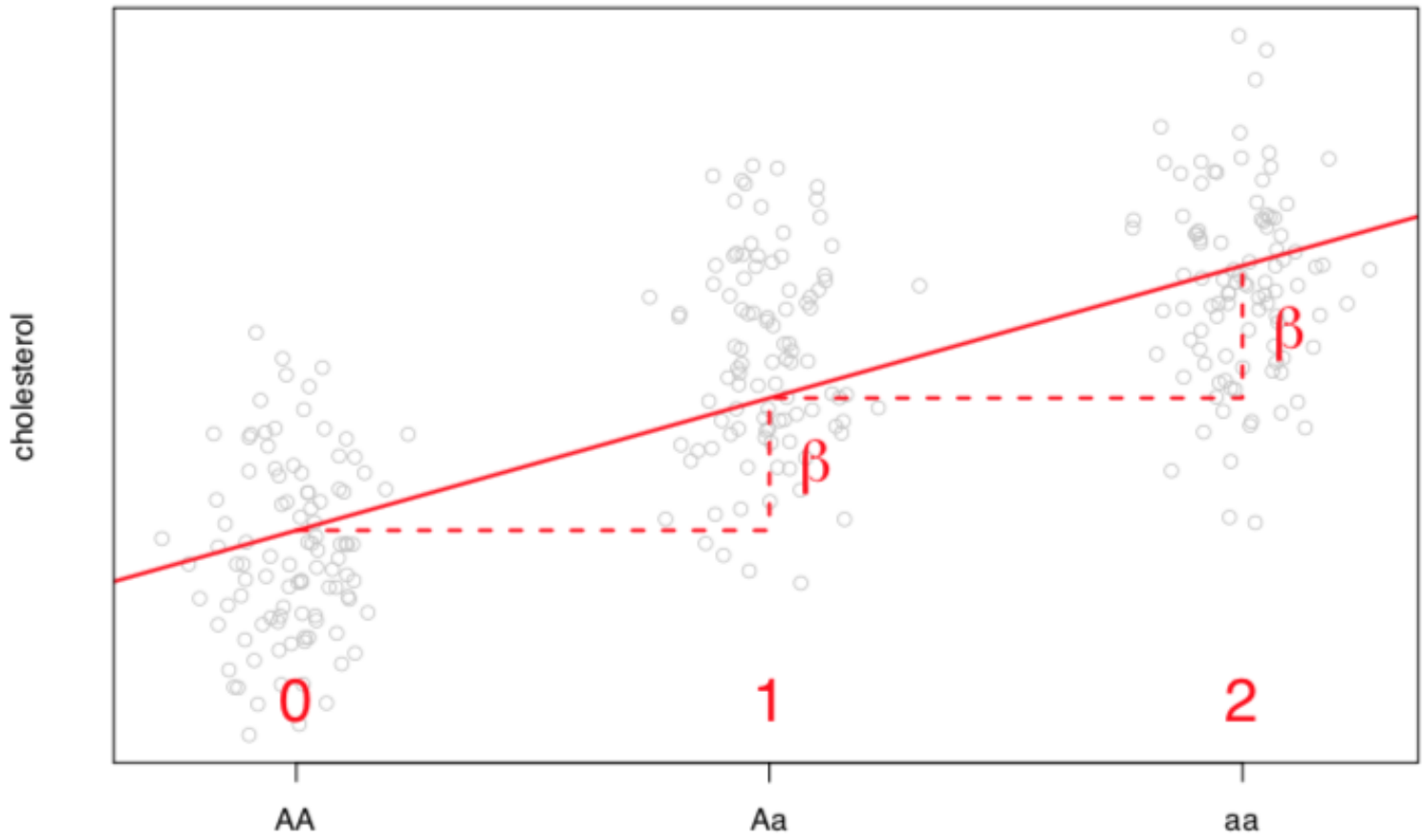
$$E(Y) = \beta_0 + \beta \times (G == aa)$$



Linear regression with SNPs

Finally many GWAS analyses fit the "additive model":

$$E(Y) = \beta_0 + \beta \times (\#\text{ minor alleles})$$

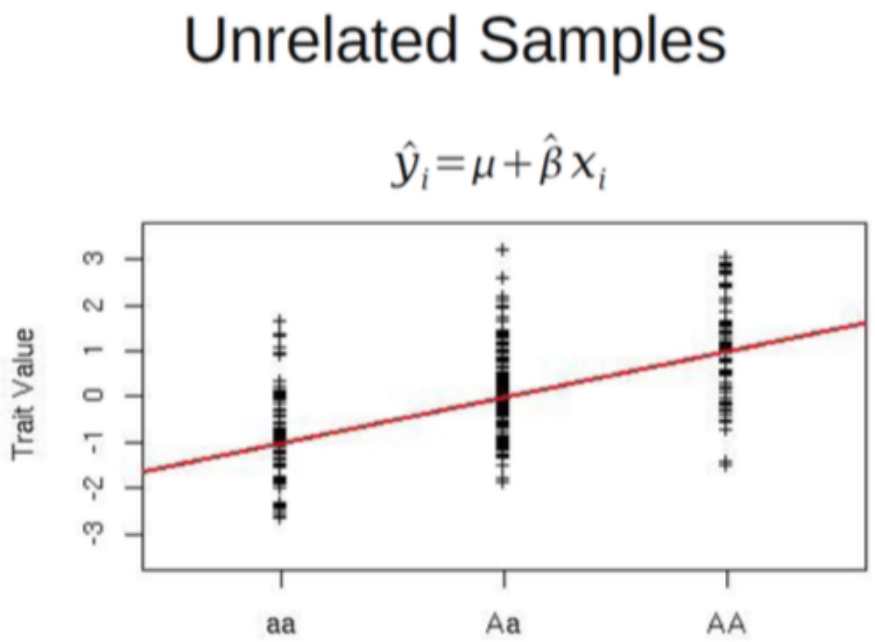


Additive Genetic Model

Most GWAS perform single SNP association testing with linear regression assuming an additive model:

$$E(Y) = \beta_0 + \beta X,$$

where X is the genotype at the SNP to be tested, e.g. $X \in \{0, 1, 2\}$ for a bi-allelic SNP.

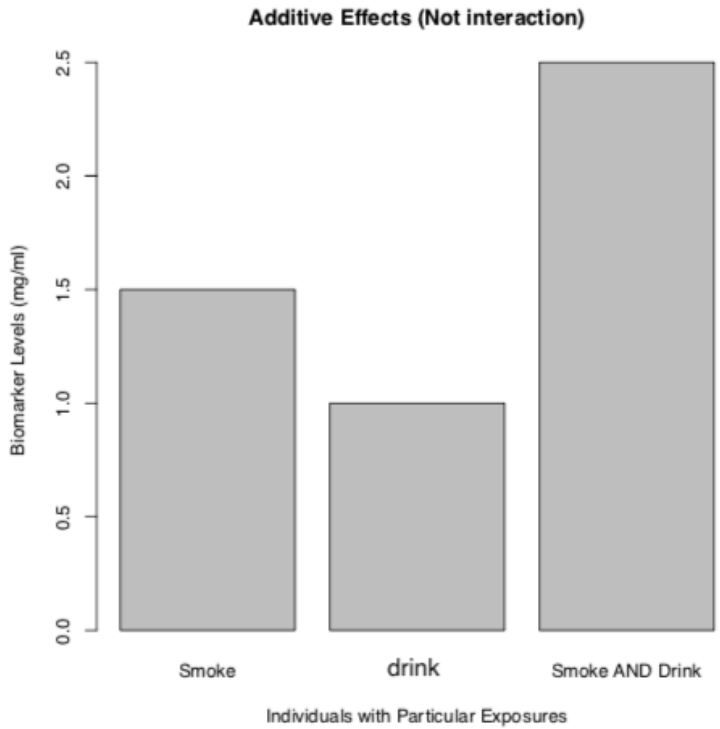


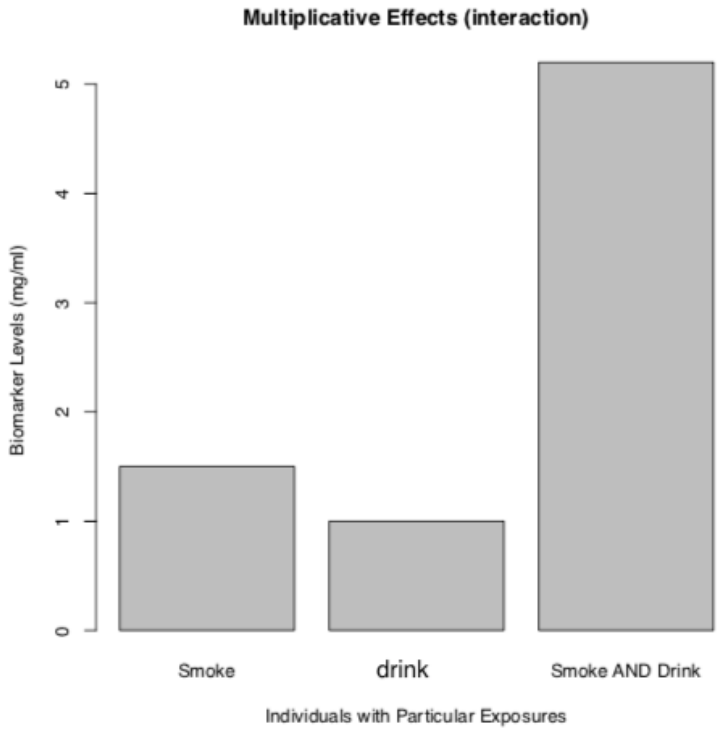
Beyond main SNP effects

- Beyond single SNP effects
  - Gene-Environment Interaction
  - Within-species gene-gene Interaction
  - Between-species gene-gene Interaction
- "Interaction" means different things in different context:
  - Communication, human-computer interaction
  - Chemistry, reaction
  - Quantitative genetics: epistasis
  - Statistics: non-additive (primarily "multiplicative")
  - Others – a lot of general vagueness
- Interaction is a three-variable concept. One of these is the response variable (Y) and the other two are predictors X_1 and X_2.
- Effect modification: one variable changes the effect of the other on outcome (deviation from additivity)

Interaction

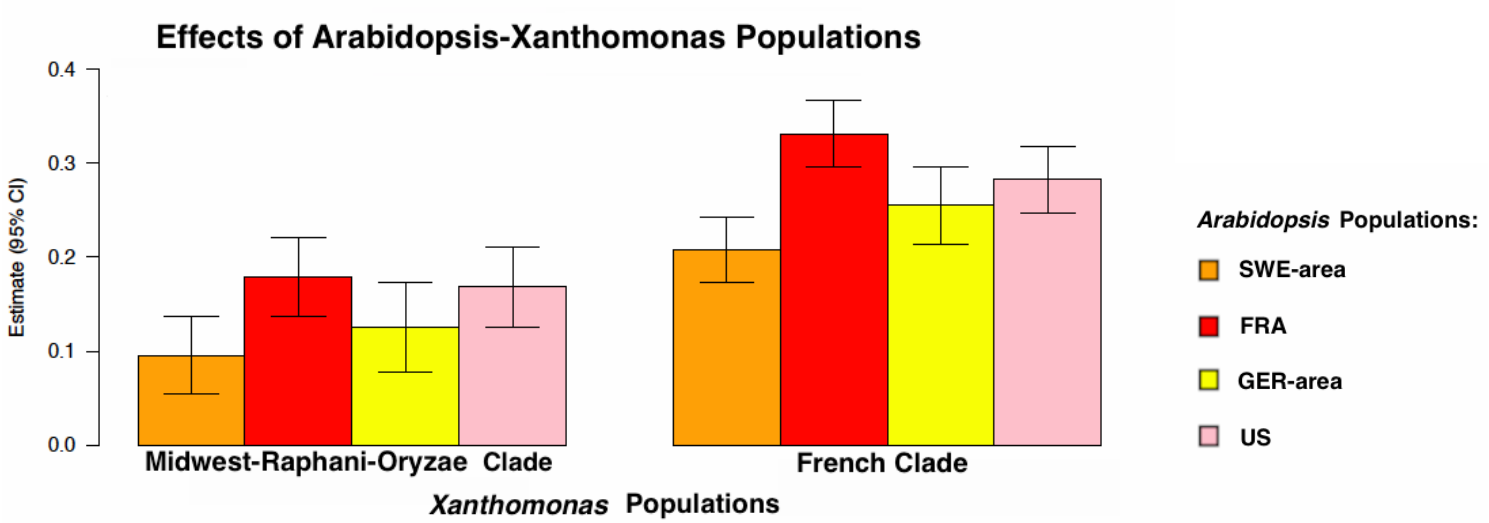
- Multiplicative interactions: combined effect exceeds the additive effects of individual variables
- Standard 2-way interaction model:
  $E(y_i) = \beta_0 + \beta_g G_i + \beta_e E_i + \beta_{ge} G_i E_i$.
- Example:



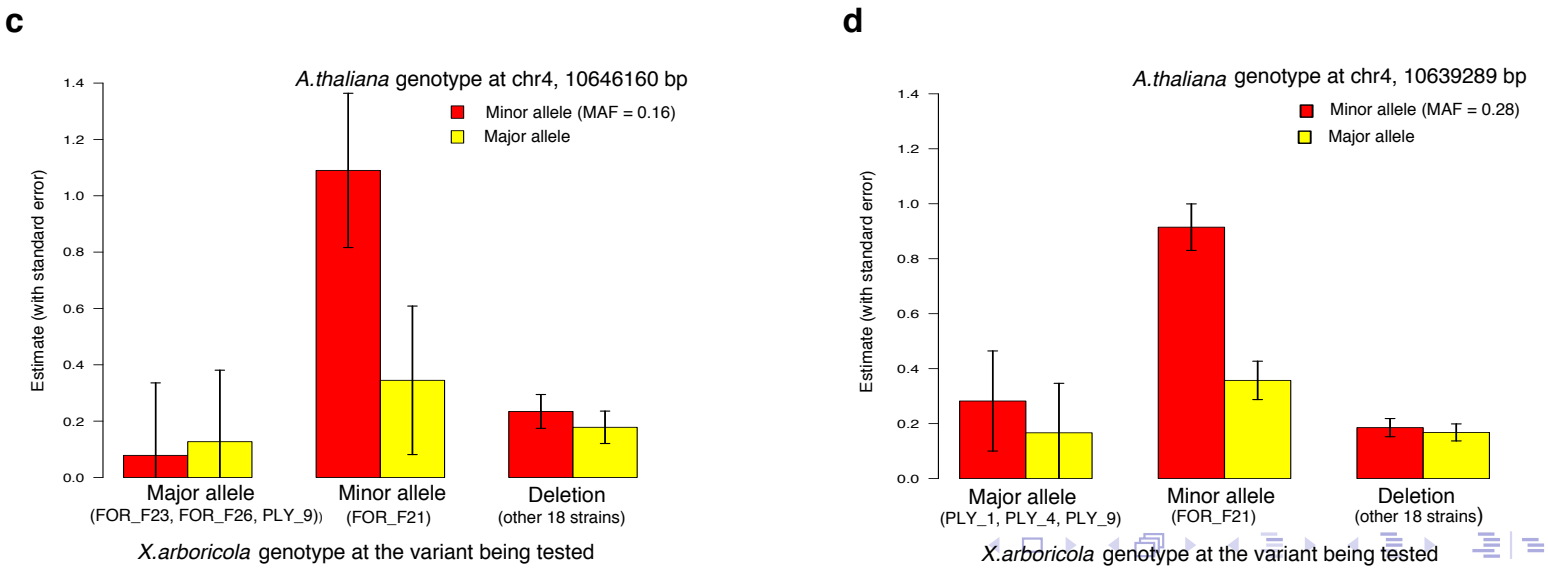


Interaction in host-pathogen system

- Population interaction:



- Gene-Gene interaction



Additive Genetic Model

A simple linear model (SLM) generally refers to the following model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

or with interaction

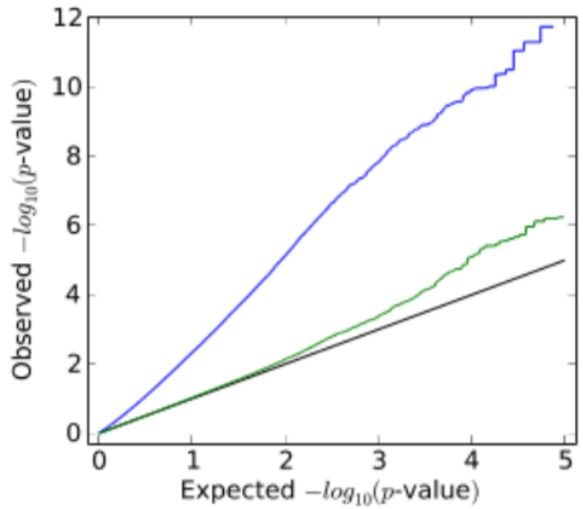
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon,$$

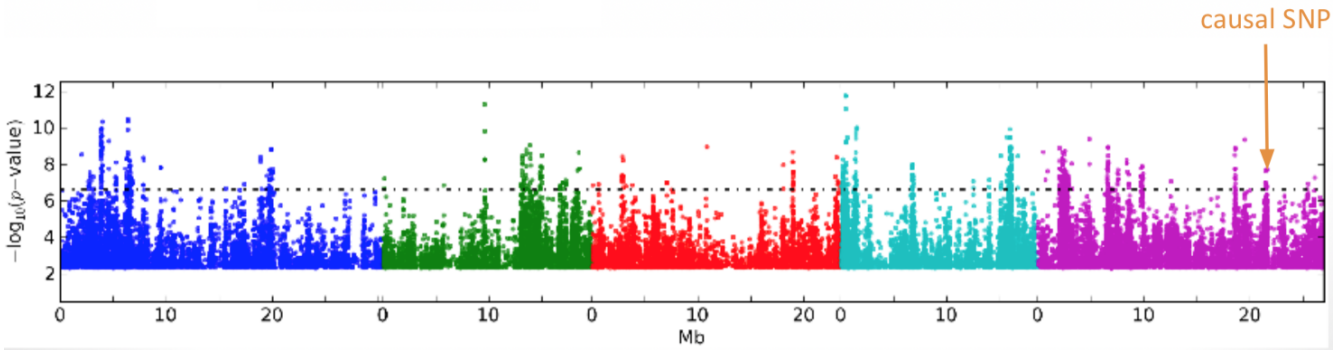
- Y consists of the phenotype values, or case-control status for N individuals.
- X_1, X_2 are the genotypes at the SNPs to be tested.

What would your interpretation of ε be for these models?

Risk

- Neglecting or not accounting for ancestry differences among sample individuals can lead to false positive or spurious associations!
- This is a serious concern for all genetic association studies.

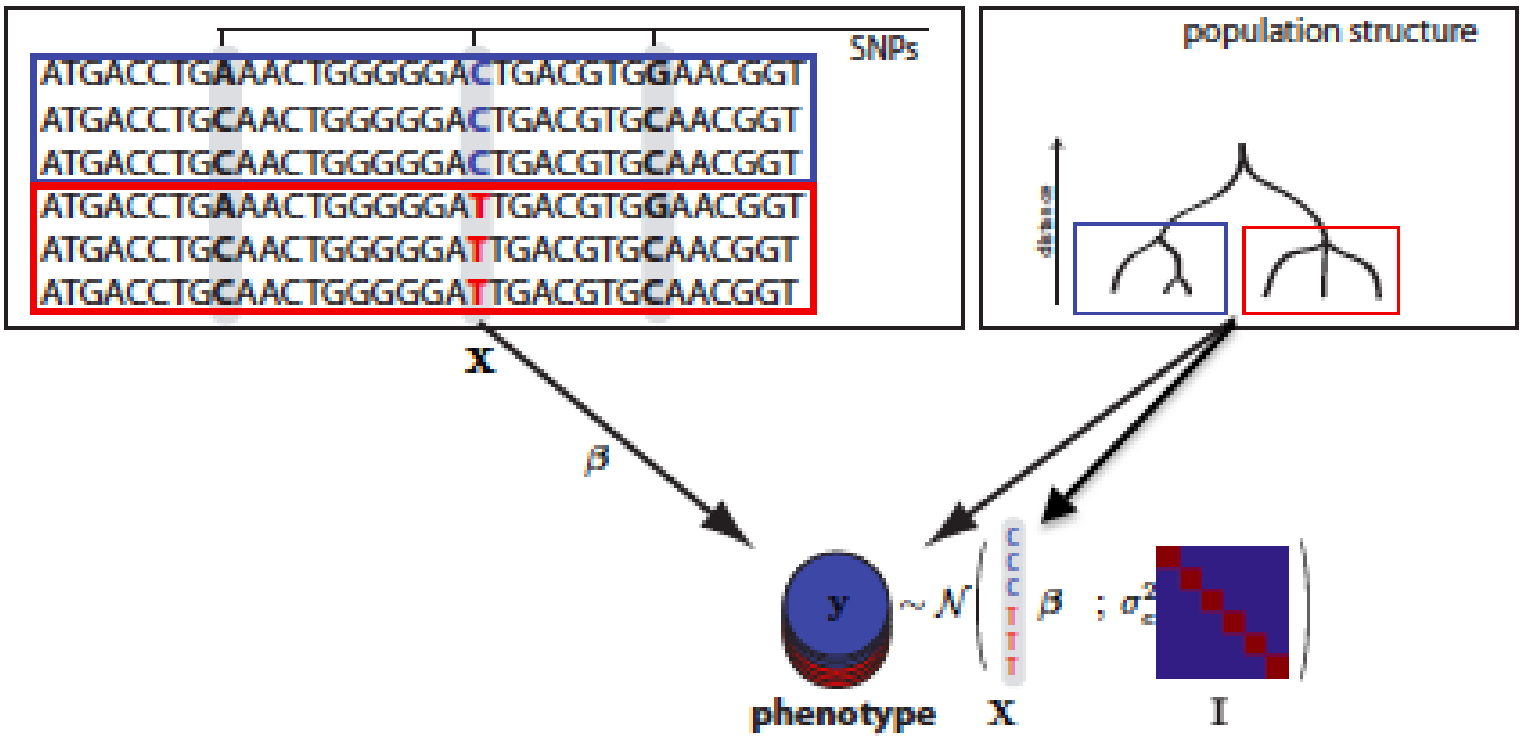




Confounding due to Hidden sample Structure

Spurious association due to confounding factors:

- Population Stratification

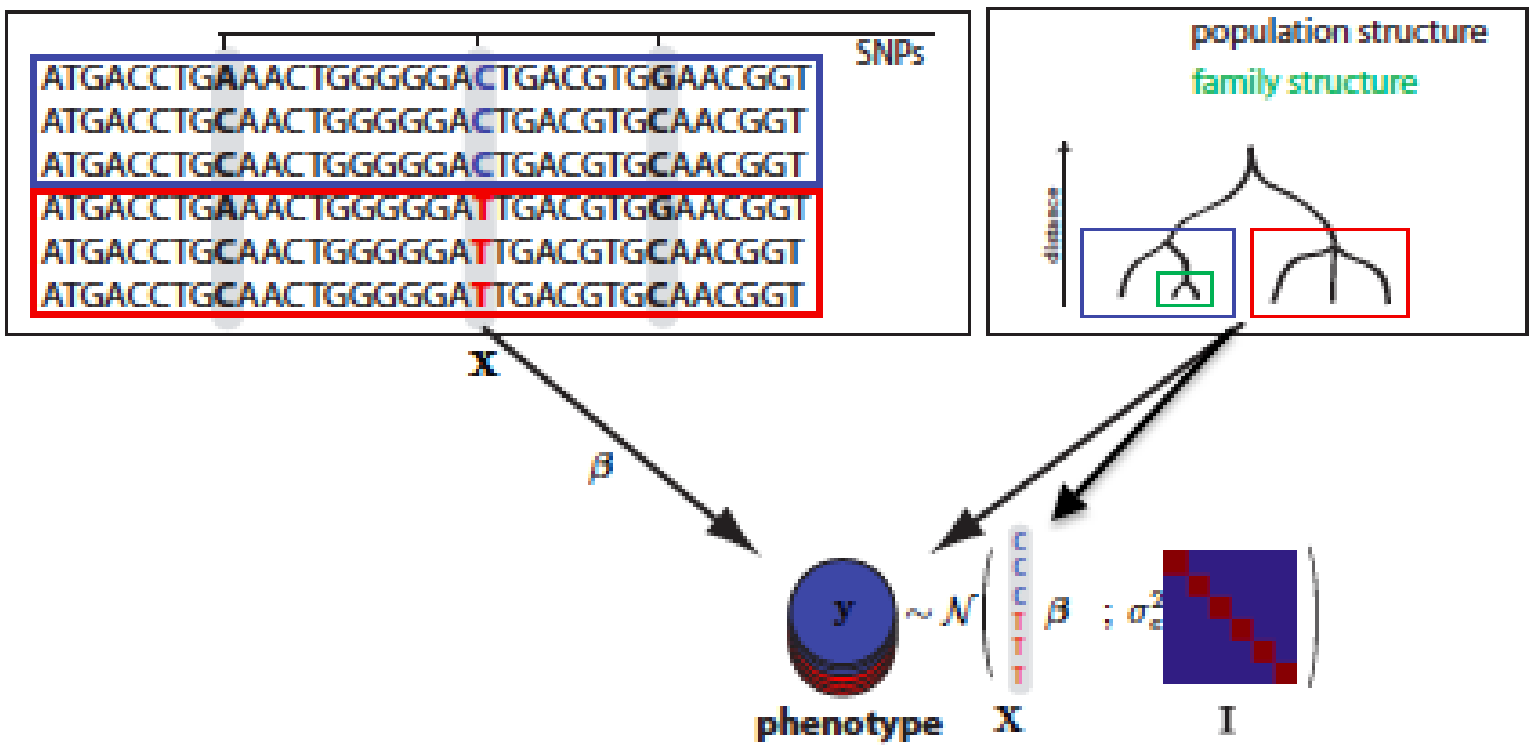


(Modified) Figure Source: Machine Learning and Statistics in Genetics and Genomics. Christoph Lippert. Lectures in UCLA. Winter 2014.

Confounding due to Hidden sample Structure

Spurious association due to confounding factors:

- Population Stratification
- Family Relationship

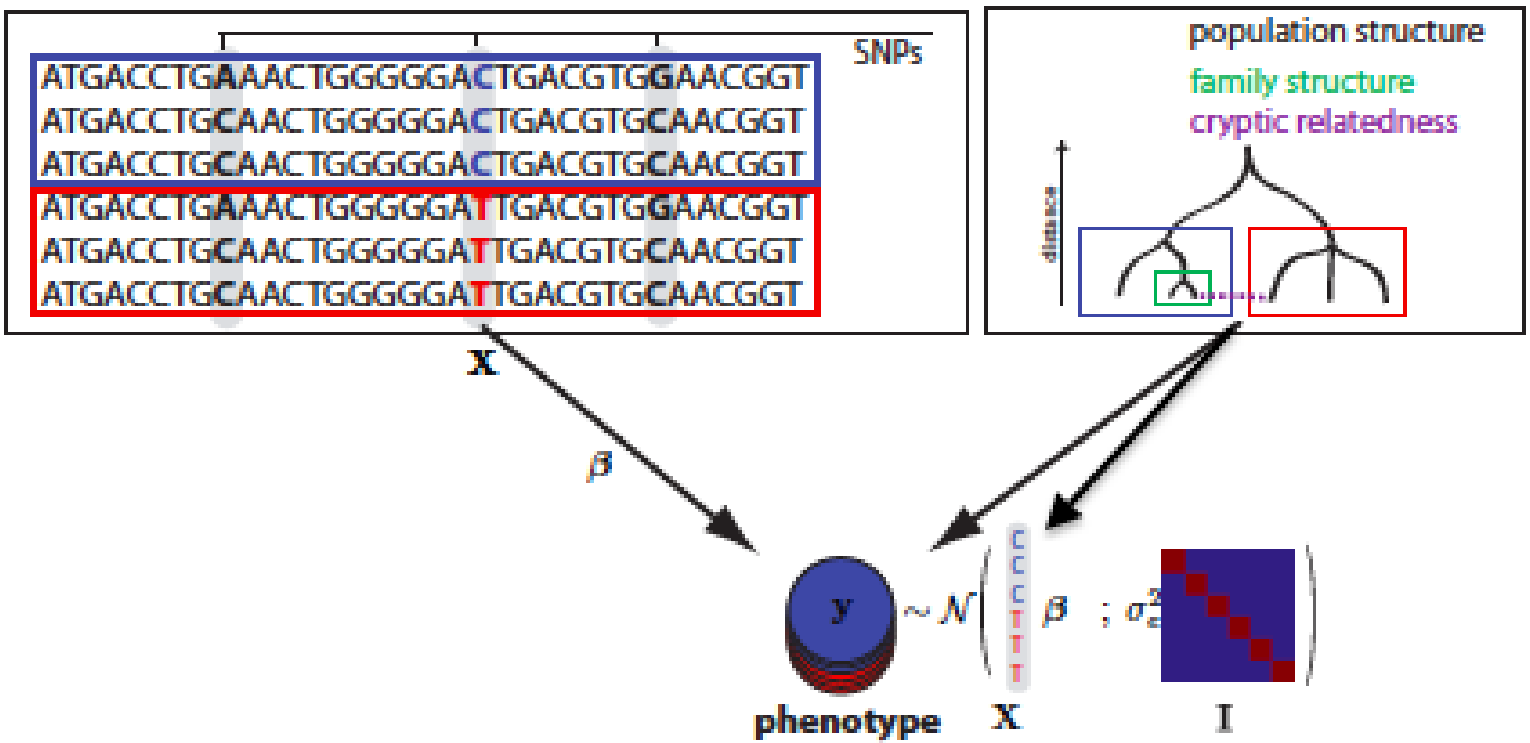


(Modified) Figure Source: Machine Learning and Statistics in Genetics and Genomics. Christoph Lippert. Lectures in UCLA. Winter 2014.

Confounding due to Hidden sample Structure

Spurious association due to confounding factors:

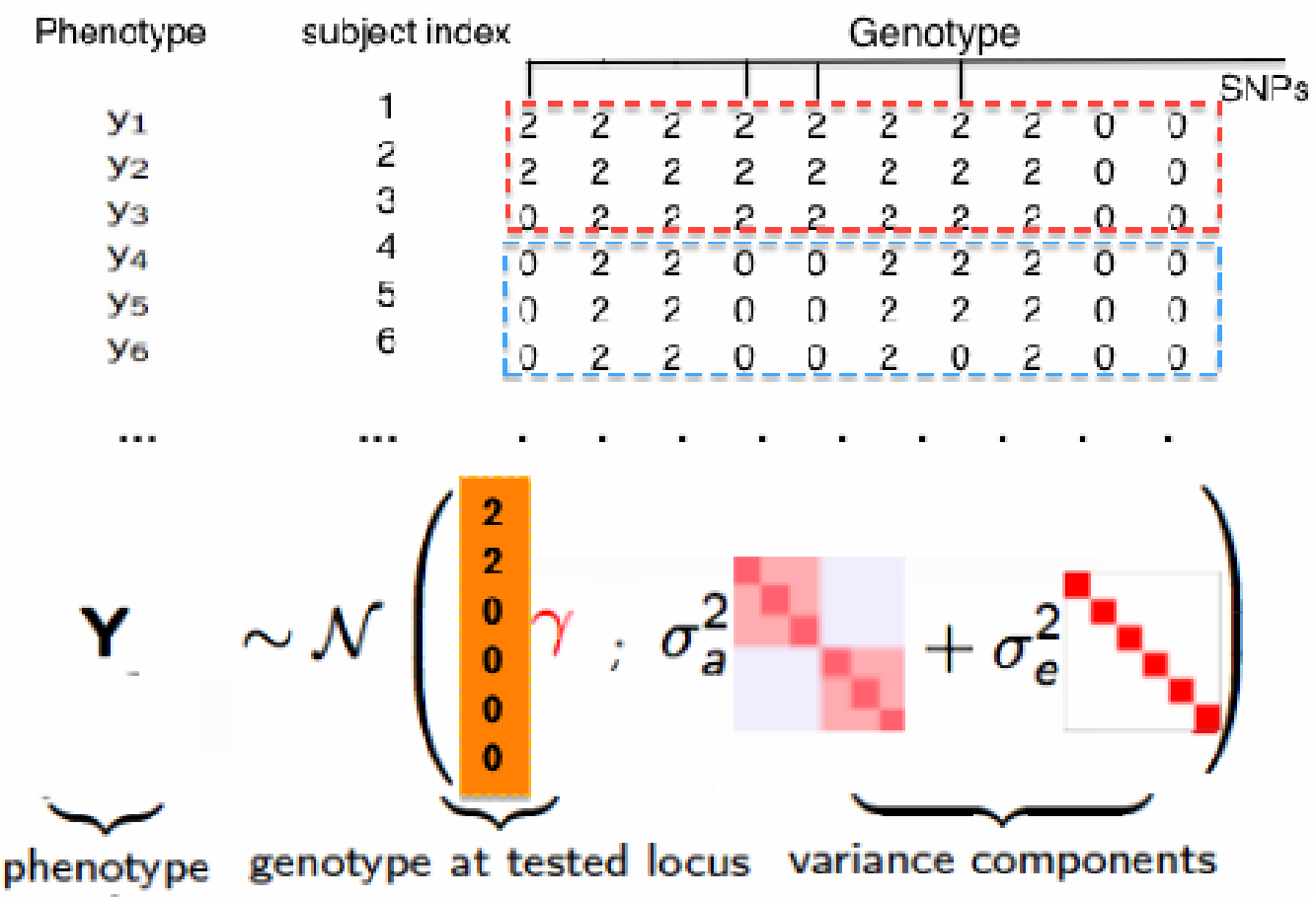
- Population Stratification
- Family Relationship
- Cryptic Relatedness



(Modified) Figure Source: Machine Learning and Statistics in Genetics and Genomics. Christoph Lippert. Lectures in UCLA. Winter 2014.

Linear Mixed Model (LMM)

Linear mixed model (LMM) corrects for confounding and increases power for association:



Linear Mixed Model (LMM)

- Standard linear mixed-effect model (LMM):

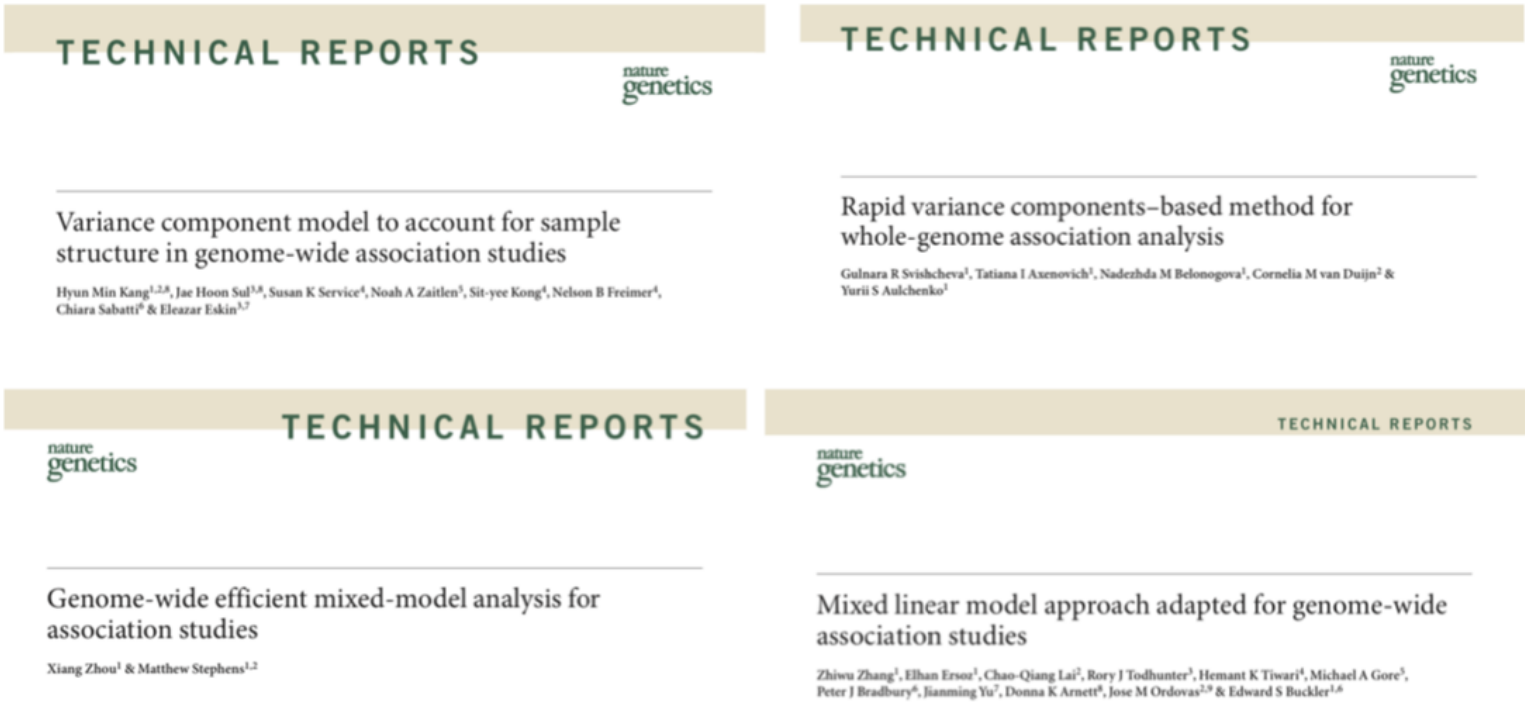
$$\underbrace{\mathbf{Y}}_{\text{phenotype}} = \underbrace{\mathbf{G}\gamma}_{\text{genotype at tested locus}} + \underbrace{\mathbf{W}\beta}_{\text{covariates}} + \varepsilon,$$

$$\varepsilon \sim N(0, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \underbrace{\sigma_a^2 \mathbf{\Phi} + \sigma_e^2 \mathbf{I}}_{\text{variance components}}.$$

  where $\mathbf{\Phi}$ is the structure matrix designated to reflect the dependence among sampled subjects, and could be chosen to be
  - function of the genealogies among sampled subjects (e.g, kinship matrix)
  - or, genetic relatedness matrix (also called empirical kinship matrix) estimated from genome-wide SNP data
- Mixed-effect model is widely used in genetic association studies.

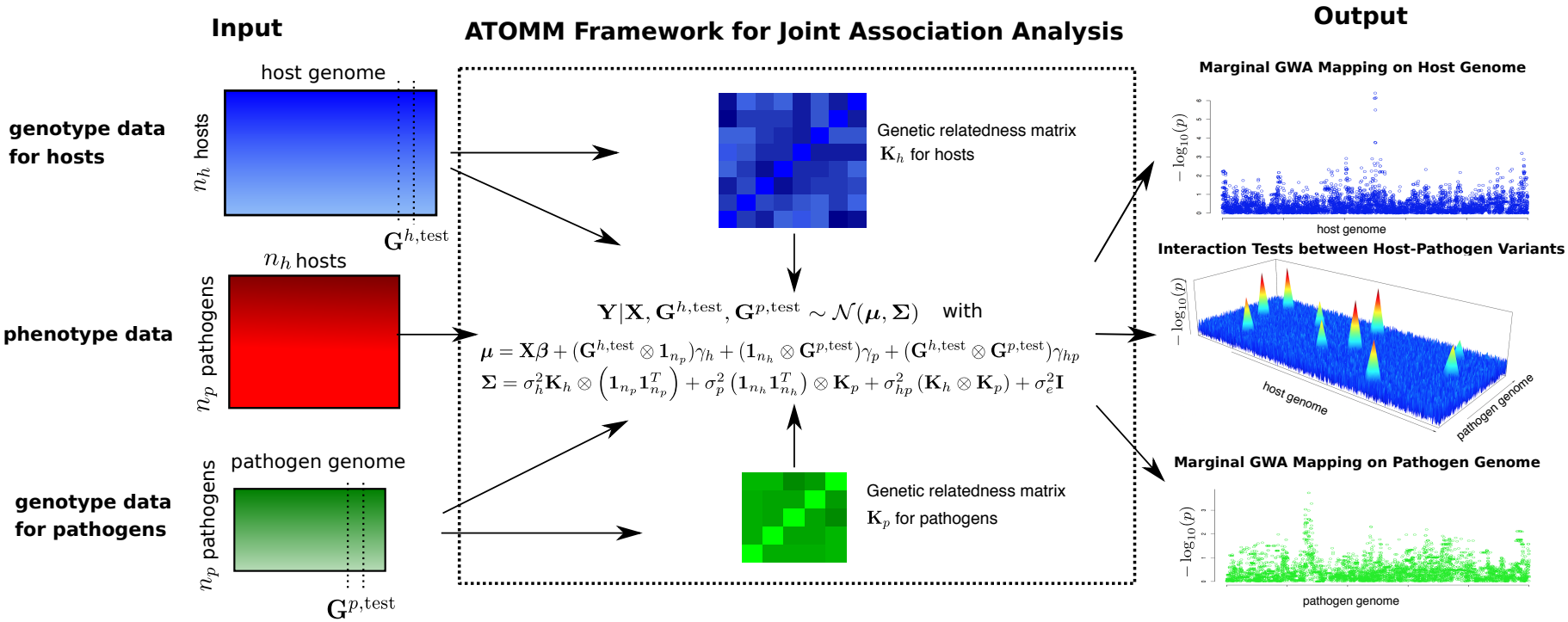
LMM approaches for quantitative traits

- A number of similar linear mixed-effects methods have recently been proposed for association testing when there is cryptic structure: Kang HM et al [2010, Nat Genet, EMMAX], Lippert et al [2011, Nat Methods], Zhou & Stephen [2012, Nat Genet], and others.



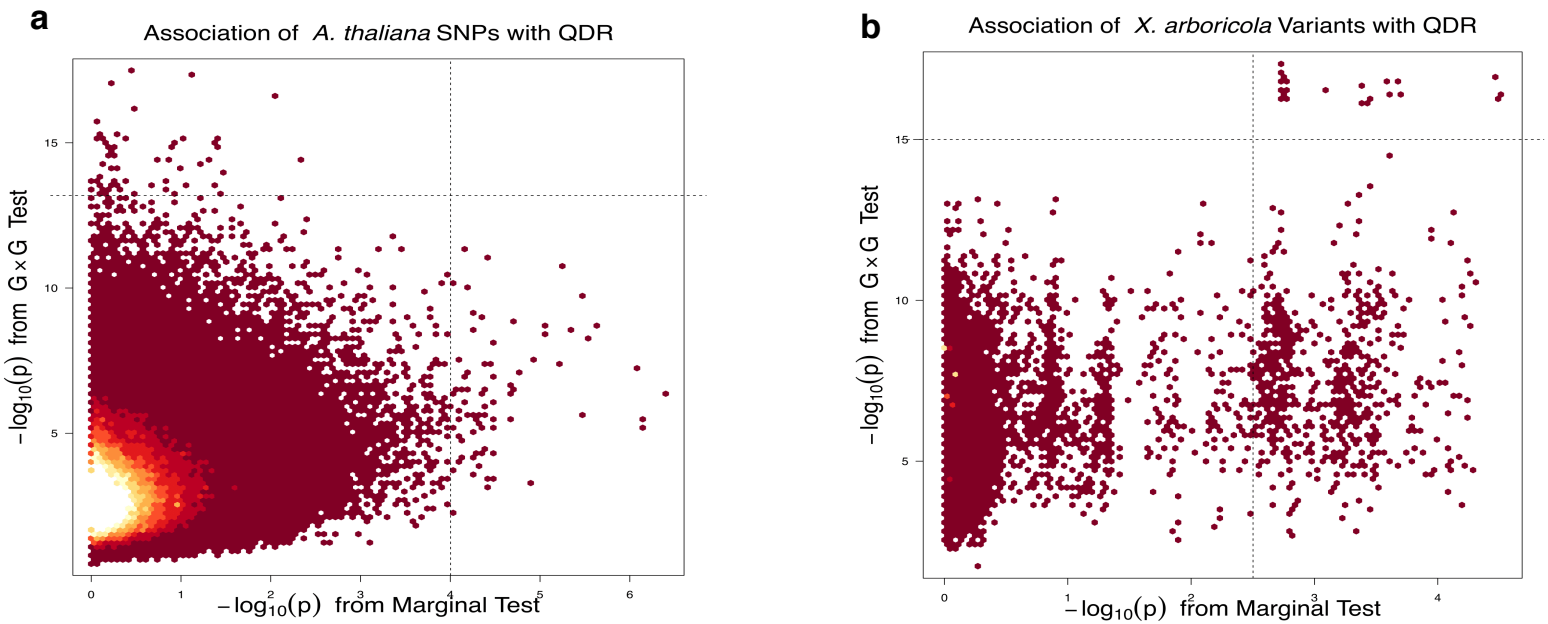
Joint GWAS for co-evolution

We have developed ATOMM (for **Analysis** with a **T**wo-**O**rganism **M**ixed **Model**) method for simultaneously detects association signals on a pair of genomes, while controlling for population structure in both species.



M. Wang, et al. *PNAS*. Vol. 115 (24), (2018) E5440-E5449.

Top interactive SNPs vs. top marginal SNPs.



Estimate Under the Null:	Gaussian trait	Binomial-like trait
Parameter	Estimate	Estimate
Intercept (β_0)	.19 (se .010)	-3.01 (se .046)
Other covariates (omitted)	...	...
Total Variance (σ_t^2)	1.54	5.14
Proportion of Residual Variance due to:		
Arabidopsis (ξ_1)	.027	.028
Xanthomonas (ξ_2)	.567	.545
Arab. - Xan. Interaction (ξ_3)	.020	.020
Batch effect (ξ_4)	.075	.081

References:

- Two-way mixed-effects methods for joint association analyses using both host and pathogen genomes. M. Wang et al. *PNAS*. Vol. 115 (24), (2018) E5440-E5449.
- Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Studies of Complex Traits. D. Jiang and M. Wang. Biostatistics and Epidemiology. Vol. 2 (1), 132-159, 2018.
- Summer Institute in Statistical Genetics (SISG) 2017. Timothy Thornton and Michael Wu.
- Exploring Plant Variation Data Workshop 2015. Ümit Seren.
- ATOMM software: https://github.com/Miaoyanwang/ATOMM_matlab
- Plink software: http://zzz.bwh.harvard.edu/plink/

Appendix: Plink Software

Plink Overview

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner:
- PLINK has numerous useful features for managing and analyzing genetic data:
  - Gene-based tests of association
  - Screen for epistasis
  - Gene-environment interaction with continuous and dichotomous environments

Input Files

- Genotype data is a text file
  - Pedigree file (.ped)
  - Map file (.map)
- Genotype data is a compressed binary file
  - Fam File (.fam)
  - Bim file (.bim)
  - Bed file (.bed)

Input Files

- Pedigree File - the first six columns are mandatory:
  - Family ID
  - Individual ID
  - Paternal ID
  - Maternal ID
  - Sex (1=male; 2=female; other=unknown)
  - Phenotype

Input Files

- MAP File has 4 columns:
  - chromosome (1-22, X, Y or 0 if unplaced)
  - rs# or snp identifier
  - Genetic distance (morgans)
  - Base-pair position (bp units)

Quality Control (QC)

- Summary statistics options:
  - minor allele frequency (MAF): –freq
  - SNP missing rate: –missing
  - Individual missing rate: –missing
  - Hardy-Weinberg: –hardy
- MAF: –maf
- SNP missing rate: –geno
- Individual missing rate: –mind
- Hardy-Weinberg: –hwe