Genome-wide association and genomic prediction

Lecture 15, CPSC 499, Fall 2018

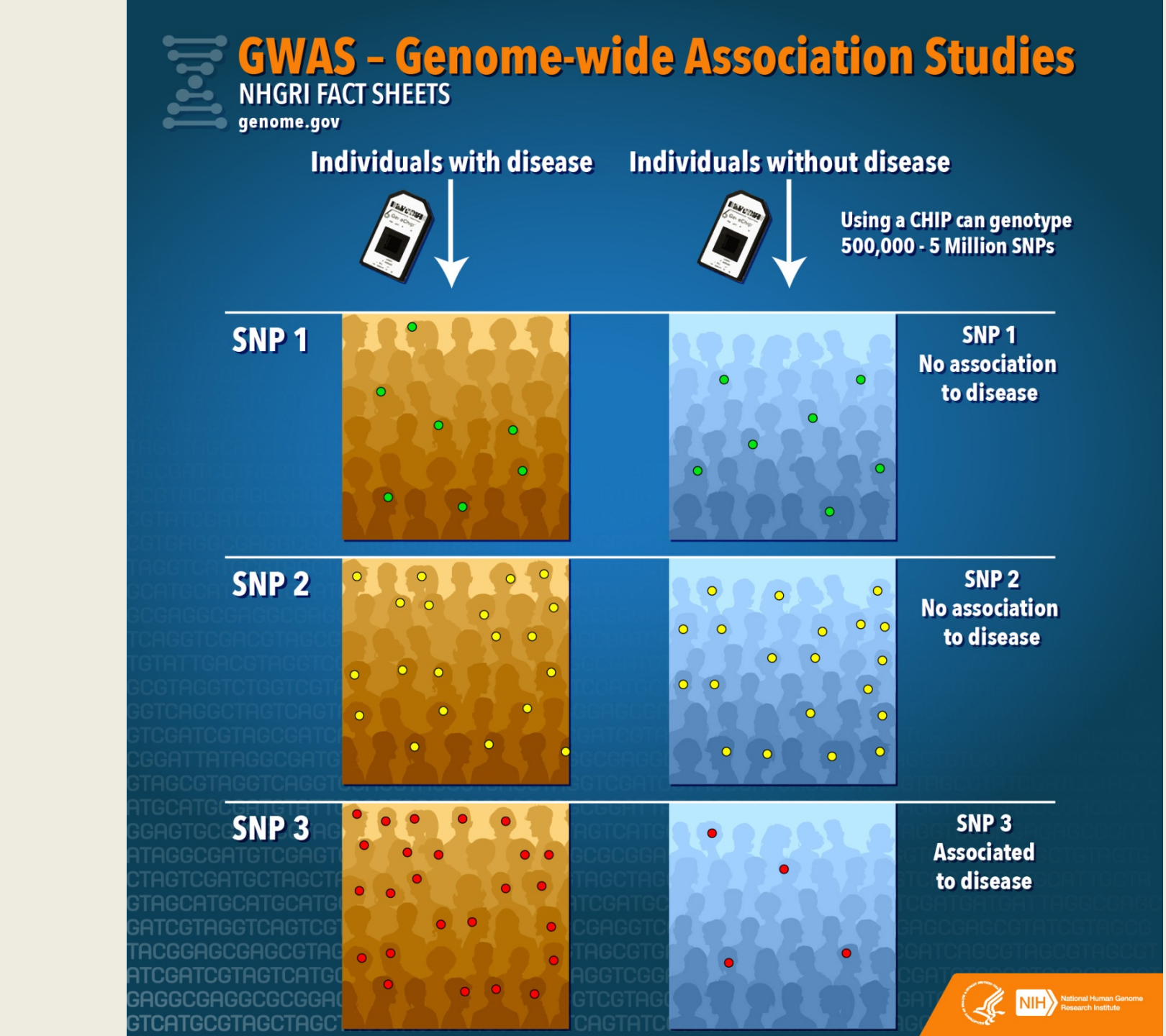
What are association studies?

- Attempts to find genetic variants that are associated with a phenotype
- Use existing populations of individuals
  - Collected from the wild
  - Diverse collection of cultivated varieties or breeds
- Perform linear regression of the phenotype on the variant, look for significant hits

GWAS – Genome-wide Association Studies

NHGRI FACT SHEETS

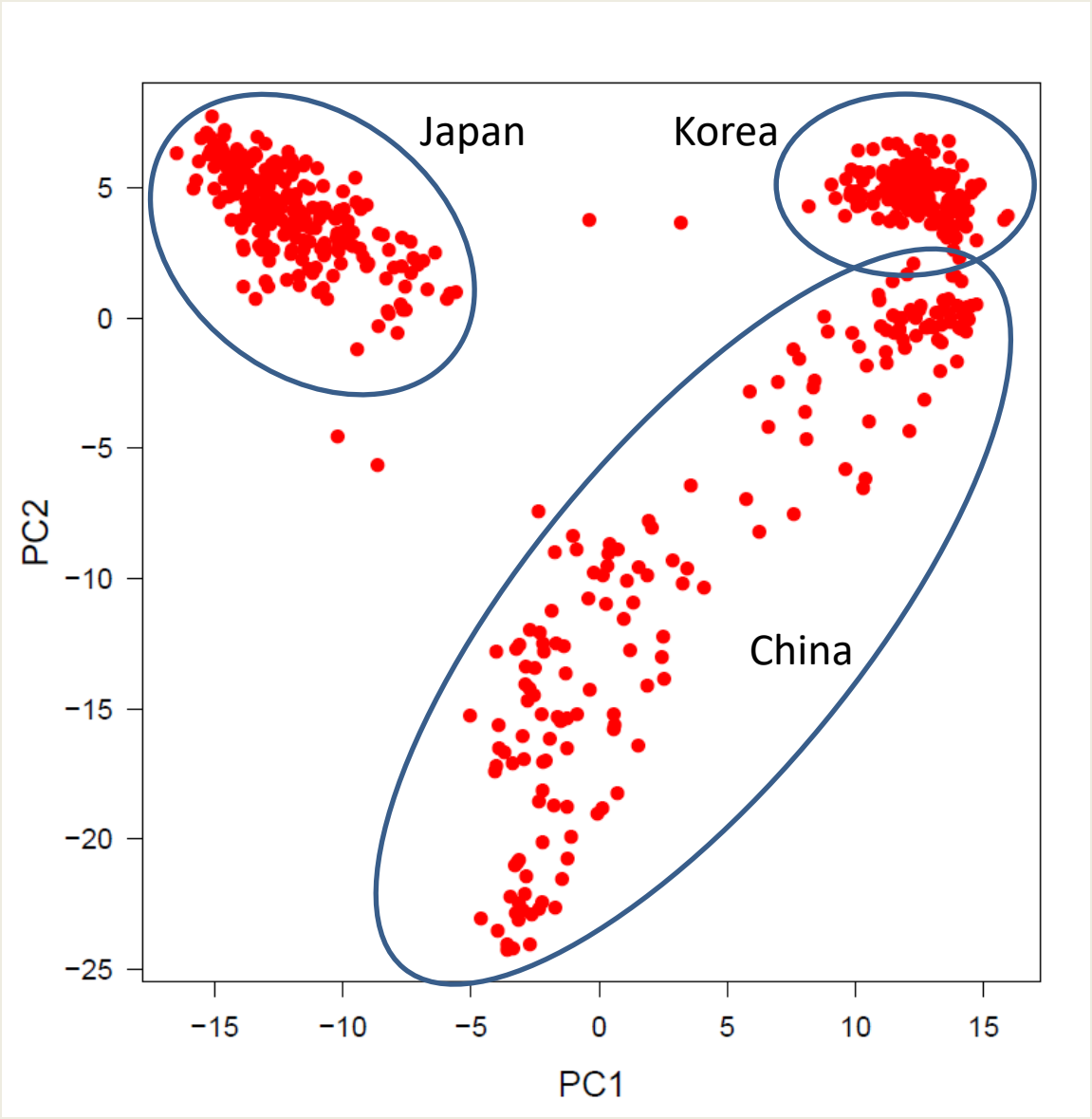
genome.gov



NIH National Human Genome Research Institute

Controlling for population structure and relatedness

- Lander and Schork (1994)
  - An allele in the HLA gene complex is associated with the ability to use chopsticks
  - That allele is simply more common in Asians
- Yu et al. (2006) proposed Q + K model
  - Q = fixed effect, several columns indicating pop. structure (PCA or Bayesian clustering)
  - K = random effect, square matrix of relatedness among all individuals
  - Phenotype = SNP + Q + K + mean + error
  - Ok to drop Q from model, but not K

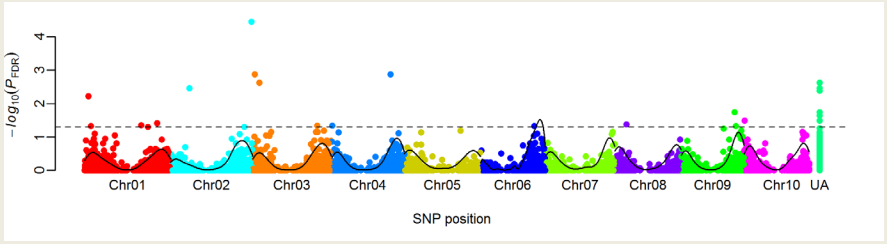


From the *Miscanthus* data we'll use in class today

For each individual, values along these axes are part of the Q matrix

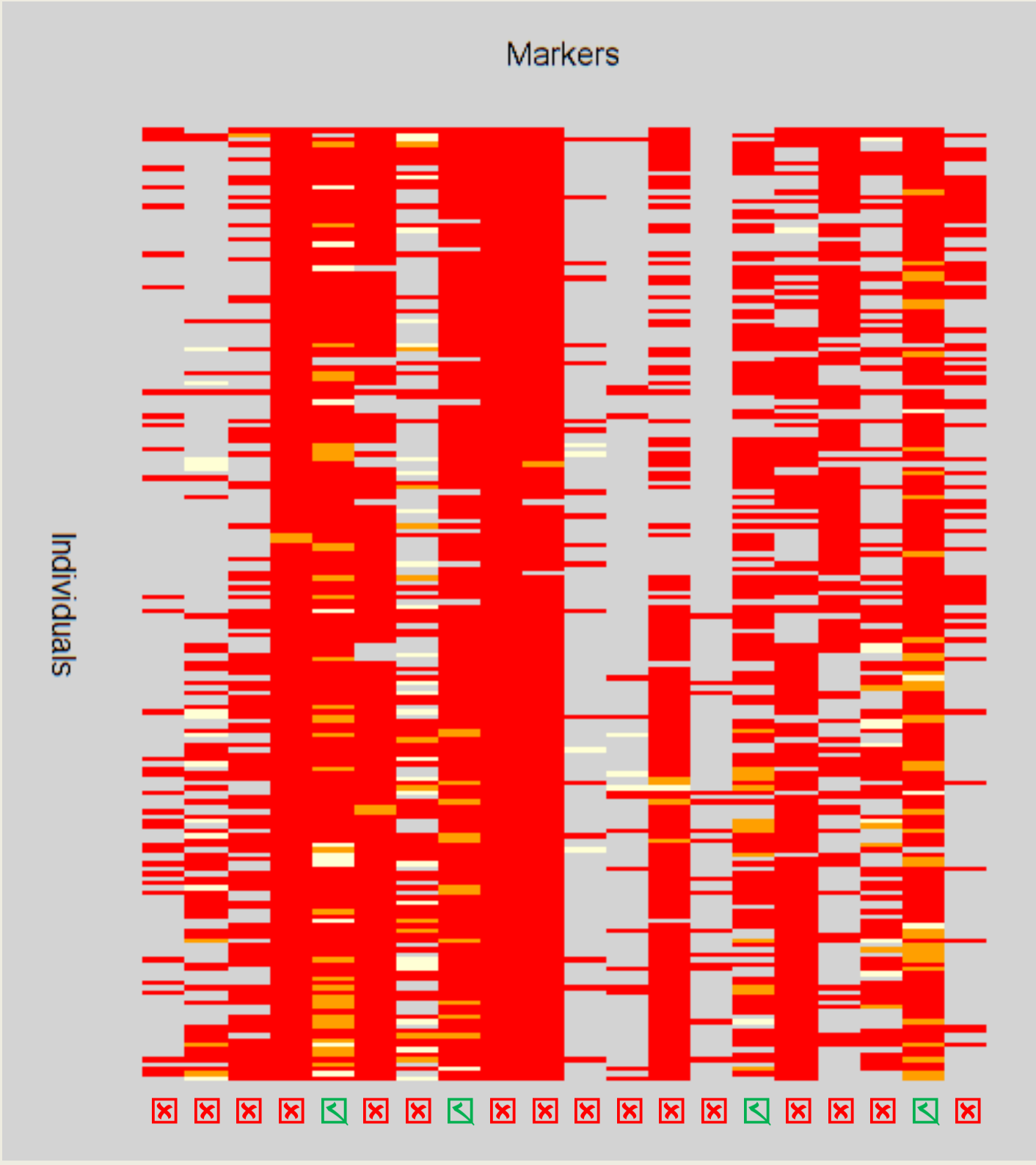
What is a genome-wide association study?

- Association studies before markers were cheap:
  - Resequence candidate genes and test variants found for association with trait
  - Use SSR markers across the genome to control for population structure
- Genome-wide associations studies
  - Genotype thousands or millions of variants across genome
  - Use the same markers for pop structure, relatedness, and association test
  - Identify candidate genes



Filtering markers for GWAS

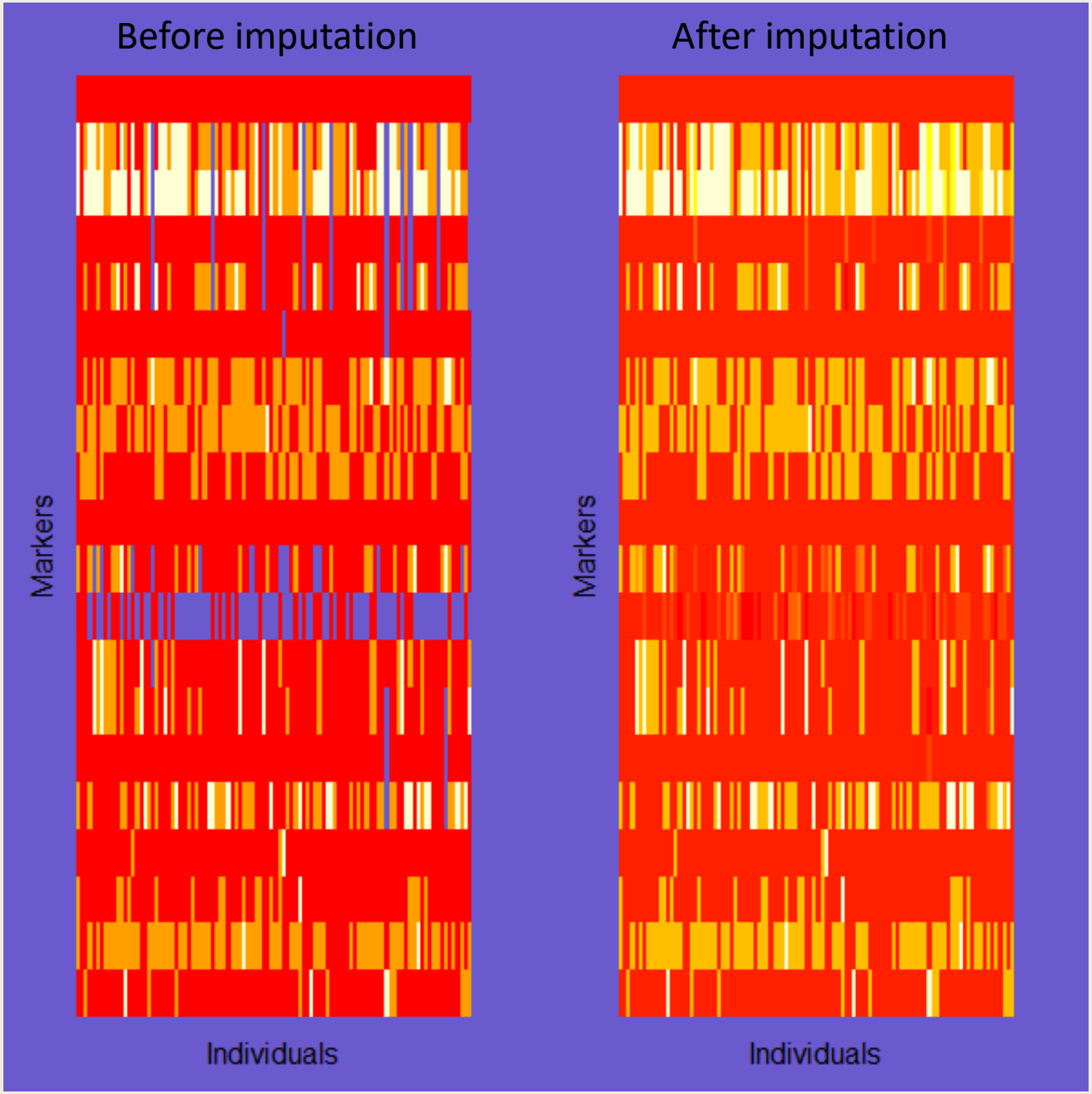
- Remove markers with high missing data
  - Lower N, less power to detect association
  - Could indicate low genotype accuracy
- Remove markers where minor allele is very rare (< 5%)
  - Alleles found only in the most extreme individuals can drive a lot of false positives
- Run these filters AFTER you have removed individuals that won't be in the analysis


Markers
Individuals

Marker imputation

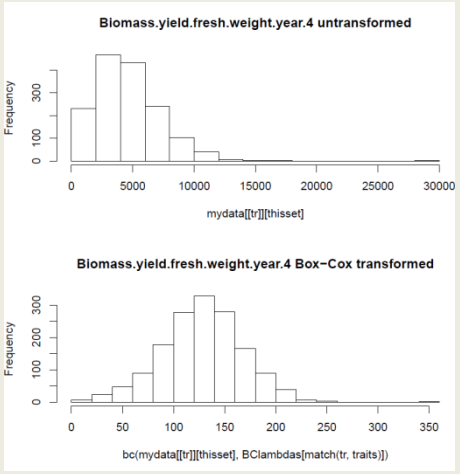
- For genotypes that are missing, fill in a best guess about the real value
- We'll do this today because GAPIT requires it
- Generally optional
  - How much missing data do you have?
  - How accurately can you impute genotypes?
- Species with low LD (like *Miscanthus*): do based on relatedness/pop structure (RR-BLUP)
- Species with high LD (many crops): based on nearby markers (Beagle, LinkImpute, etc.)

LD = linkage disequilibrium



Note that, depending on the imputation method, imputed genotypes might be intermediate reflecting genotype uncertainty.

Preparing phenotypic data for GWAS



- Aim for a normal distribution
  - Log transform, Box-Cox transform, etc. if necessary
  - High skew or kurtosis leads to false positives
- If genotypes are replicated across blocks or environments
  - build a random effects model (`lmer` inpackage lme4)
  - get Best Linear Unbiased Predictors (BLUPs) from model with `ranef`

Now we'll run through a GWAS in R

- Follow along
- R code is in Rmd document on Compass
- Go to http://www.zzlab.net/FarmCPU/FarmCPU_help_document.pdf to get FarmCPU instructions

Mini-exercise

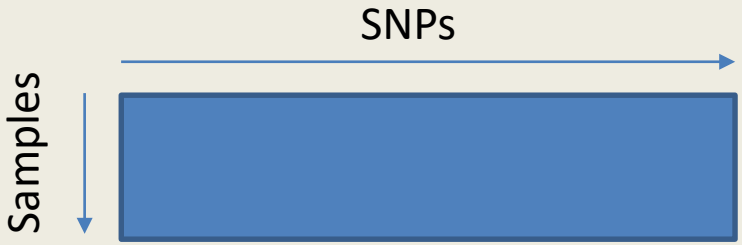
- Effect estimates can be positive or negative. In practical terms, what does that mean?
- In `results$GWAS`, look at the effect estimates for the significant hits. Make genotype x phenotype scatter plots for some SNPs with positive effects and some with negative effects.

GWAS doesn’t detect everything

- For quantitative traits, there are probably very many loci with effects too small to be detectable with significance

- A variant might have a real effect on a trait, but if strongly associated with population structure, we can’t detect it

Challenge of genomic prediction



- How do we deal with more variables than observations?
  - Variable selection – only retain the SNPs with the most predictive power
  - Shrinkage of effect estimates – assume effects of SNPs are overestimated, and shrink back towards zero
  - Assume variance attributable to SNPs is known

Genomic prediction with RR-BLUP

- Assume every LD block along the entire genome has an equal, small effect on the phenotype
- Can then reduce your genotype data to a square matrix of kinship among individuals
- Use this matrix to predict breeding values of all individuals
- Can make predictions for non-phenotyped individuals, save phenotyping effort

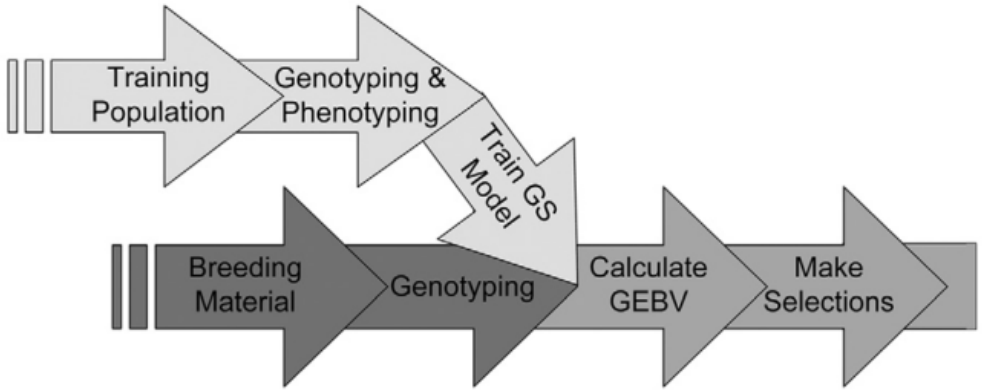


Figure 1. Diagram of genomic selection (GS) processes starting from the training population and selection candidates continuing through to genomic estimated breeding value (GEBV)–based selection. Note that while we show here a single occurrence of model training, training can be performed iteratively as new phenotype and marker data accumulate.

Figs from Heffner et al. (2009)
Doi:10.2135/cropsci2008.08.0512

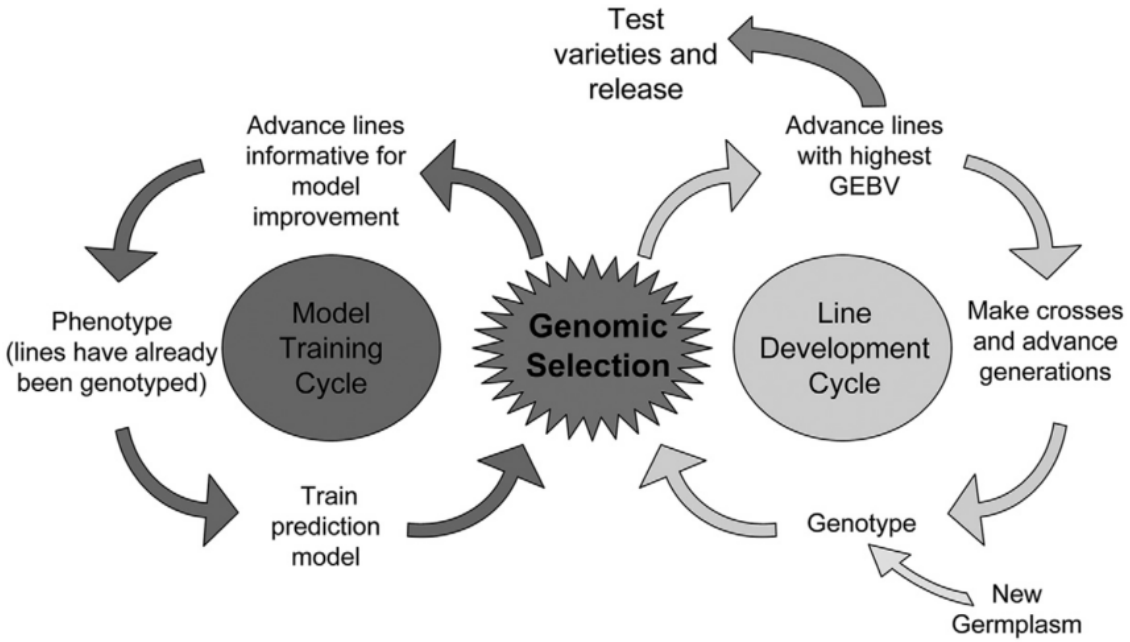


Figure 2. Flow diagram of a genomic selection breeding program. Breeding cycle time is shortened by removing phenotypic evaluation of lines before selection as parents for the next cycle. Model training and line development cycle length will be crop and breeding program specific. (GEBV = genomic estimated breeding value.)

Estimating genomic prediction accuracy

- K-fold cross-validation (generally 5-fold or 10-fold)
- For a random 20% of individuals, predict breeding values using the other 80%.
- Do many iterations of this, get average predicted breeding values
- Regress predicted breeding values on known phenotype values

Now we'll run through genomic prediction

How is prediction accuracy so high?

- We are only looking at one chromosome out of 19 in *Miscanthus sinensis*
- Yet we have explained 32% of phenotypic variation with our genomic prediction model
- This is due to high population structure, correlated with phenotype
- In subsequent generations pop. struct. would break down and we would re-make model

Mini-exercise

- Look at the GEBV matrix. How much do the predicted values vary from run to run?
- Use `apply` and `sd` to get the standard deviation for each GEBV estimate.

In summary

- GWAS and genomic prediction/GS are two ways to relate genomic variants to phenotype
- GWAS: for finding genes that may impact phenotype
- Genomic prediction: for predicting genotype using whole genome
- Some combination of the two is probably most useful for breeding or disease prediction