BIS180L

Genome-Wide Association Mapping

Professor MALOOF
jnmaloof@ucdavis.edu

Association Mapping

- Goals
  - find genes associated with traits
    - humans: disease, disease susceptibility, other traits
    - crops: yield, stress resilience, quality, disease resistance
    - ecology, evolution: traits related to fitness
  - assess relative risk based on genotype (personalized medicine)

- Association mapping
  - look at associations at a few loci (you have *a priori* candidates)

- Genome Wide Association Mapping (GWAS)
  - scan the whole genome for associations

- Takes advantage of historical recombination

Lecture Outline

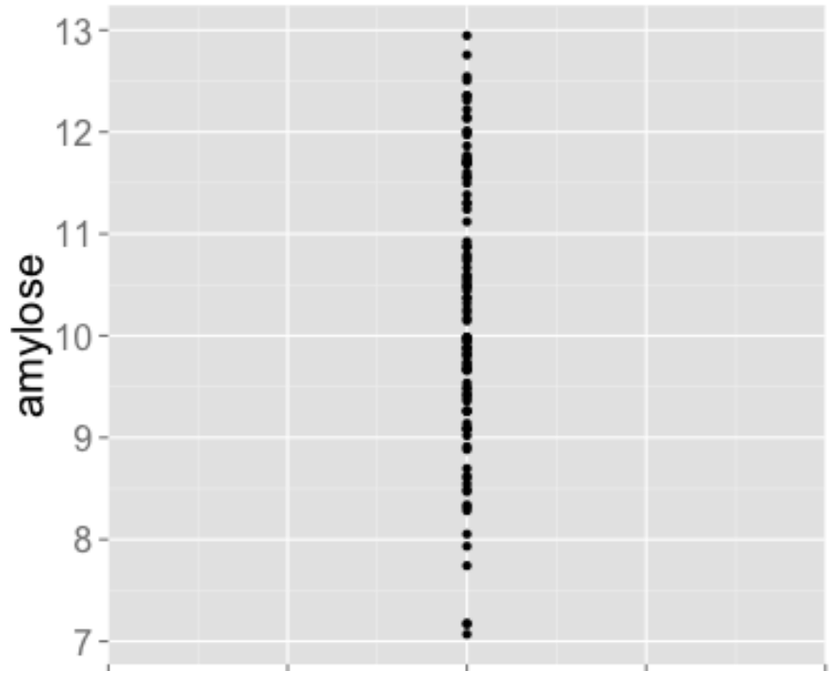
- Association Mapping
- Genome-Wide Association Mapping (GWAS)
- The problem of population structure
- QQ plot

Association Mapping

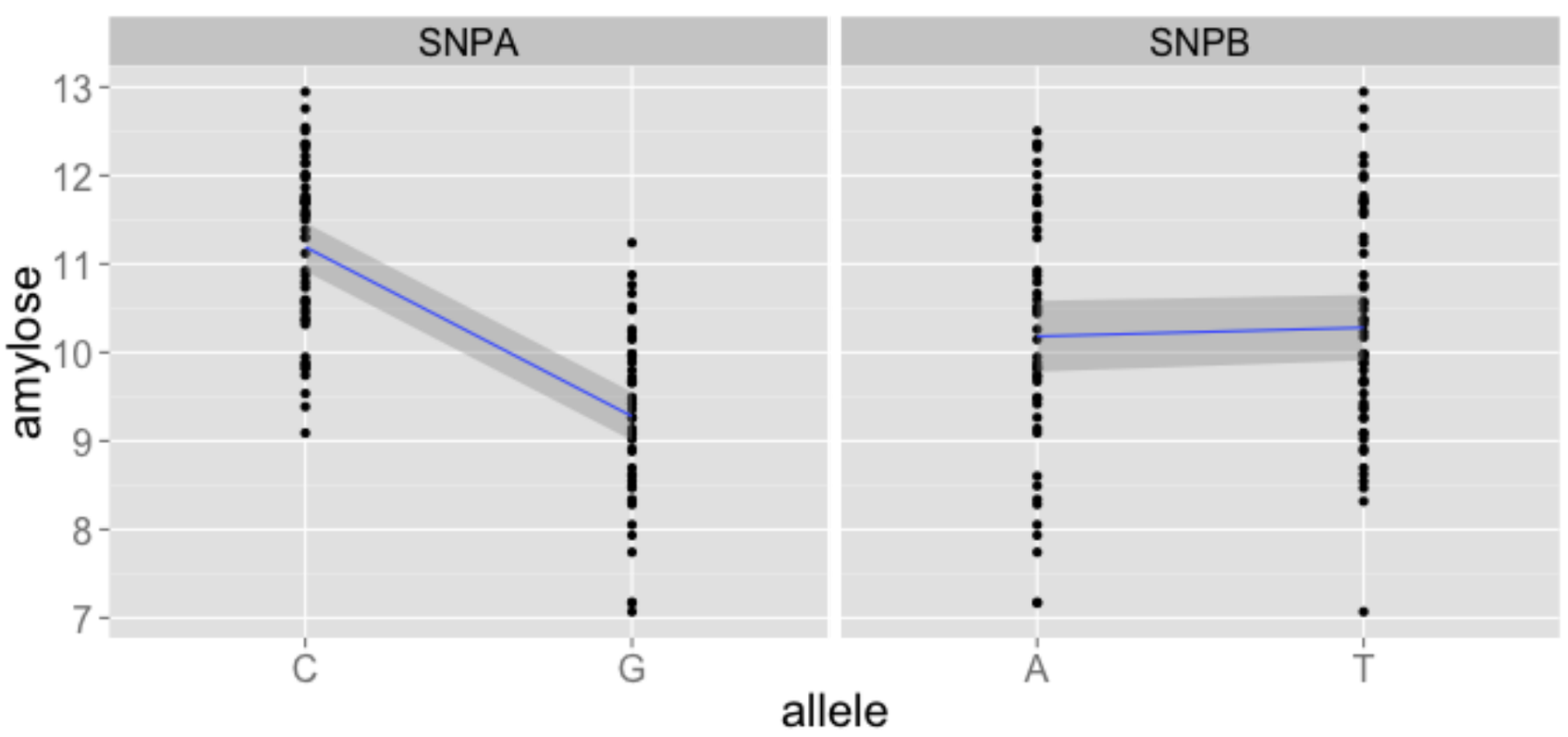
- Trying to find genetic basis for a trait or disease
- Look for statistical association between a SNP allele state and a phenotype
-

Association Mapping, simulated example for amylose

- Measure amylose in many rice varieties



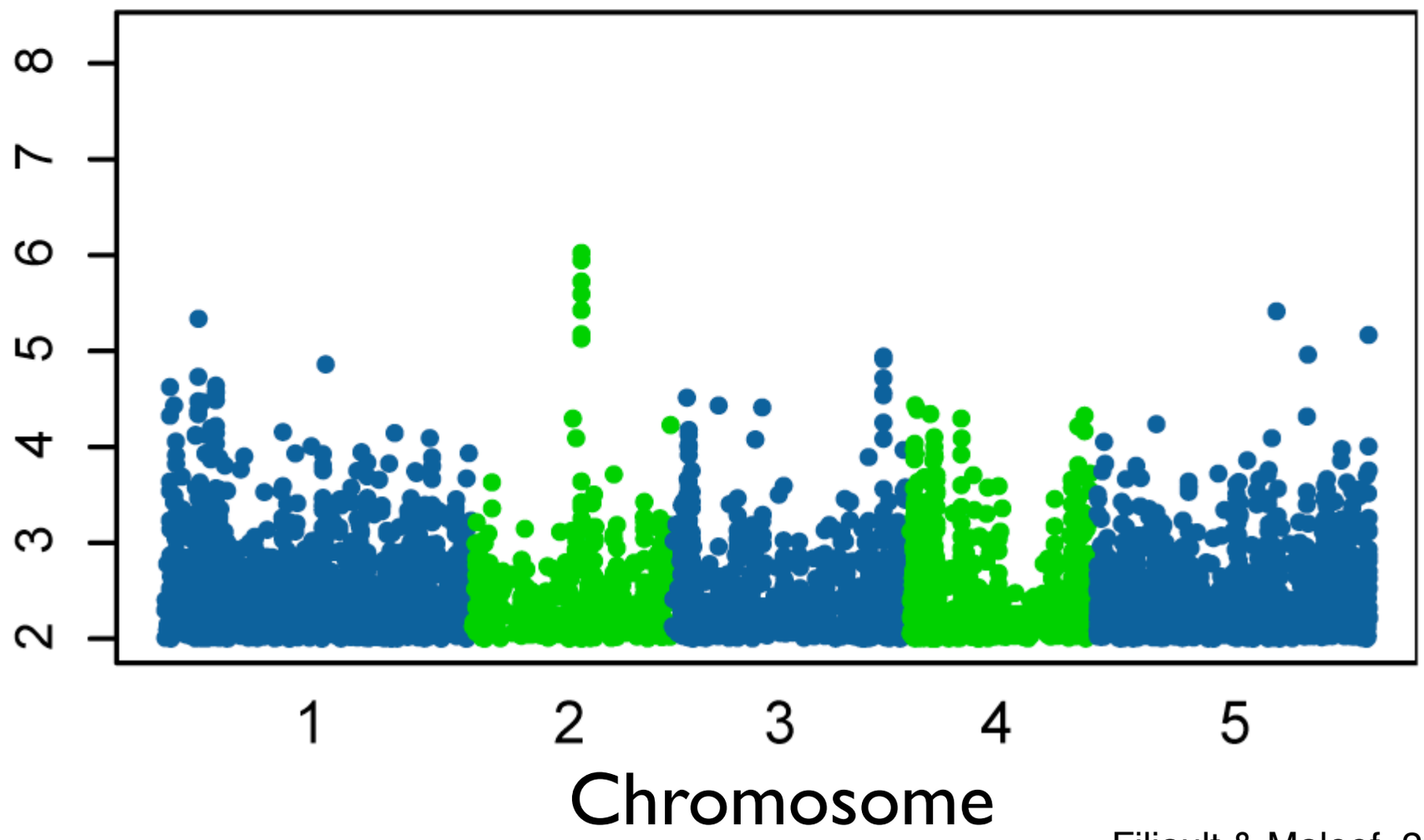
- Separate measurements according to SNP allele



- Test for association. Slope not equal to 0 = association.

Association Mapping vs Genome-Wide Association Mapping

- For GWAS repeat the analysis for SNPs across the whole genome.
- Can plot the results as a manhattan plot:
  - each point is a SNP
  - X-axis is position in the genome. In this case there are 5 chromosomes
  - Y-axis is -log10(P) for association with the trait. Higher values are more significant.

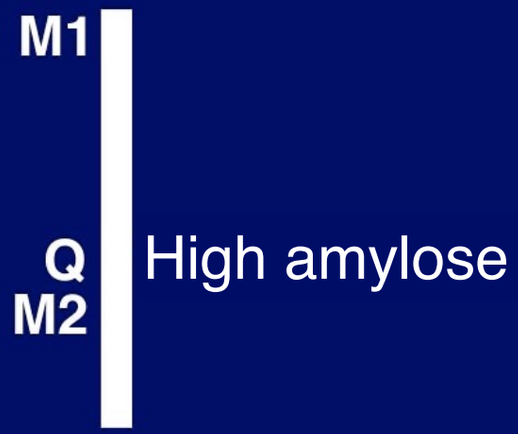


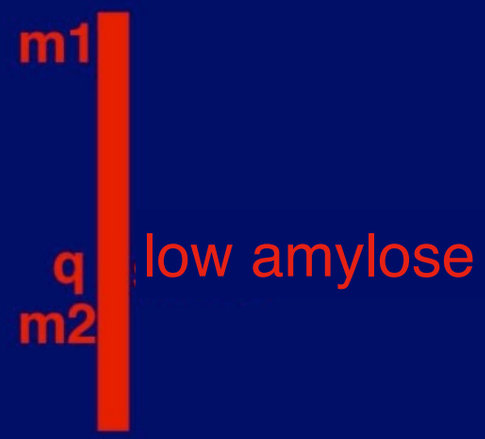
Filiault & Maloof, 2012

Association Mapping: historical recombination

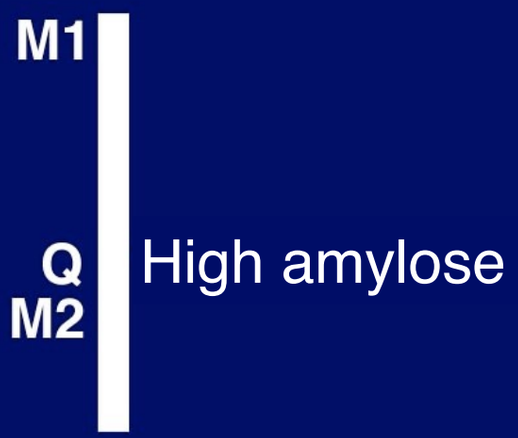
- Why might some SNPs be associated with a trait and not others?
- Historical Recombination!

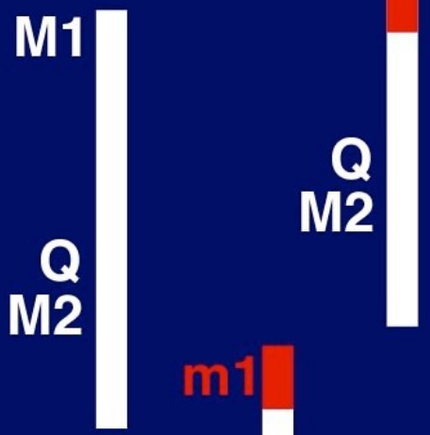
Association mapping and historical recombination

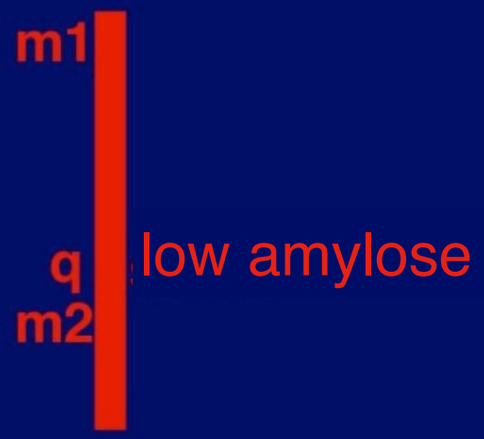


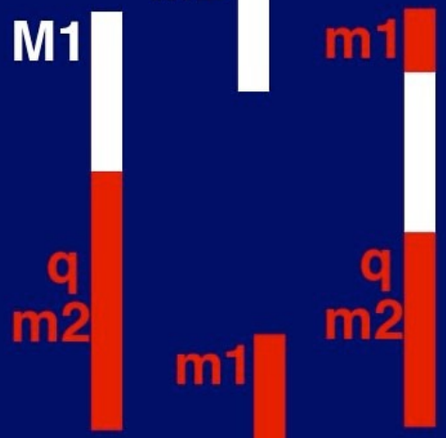


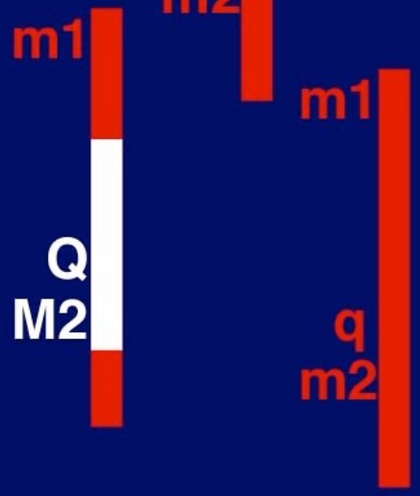
Association mapping and historical recombination











Association mapping and historical recombination

M1
Q
M2
High amylose

m1
Q
M2

M1
q
m2

M1
Q
M2

m1
Q
M2

m1
Q
M2

M1
q
m2

M1
Q
M2

m1
Q
M2

m1
q
m2

M1
q
m2

m1
Q
M2

M1
Q
M2

m1
q
m2

m1
q
m2

M1
q
m2

m1
Q
M2

m1
q
m2

M1
Q
M2

m1
q
m2

m1
Q
M2

M1
q
m2

m1
Q
M2

m1
q
m2

m1
q
m2

M1
Q
M2

m1
q
m2

m1
Q
M2

m1
q
m2

m1
Q
M2

m1
q
m2

m1
Q
M2

m1
q
m2

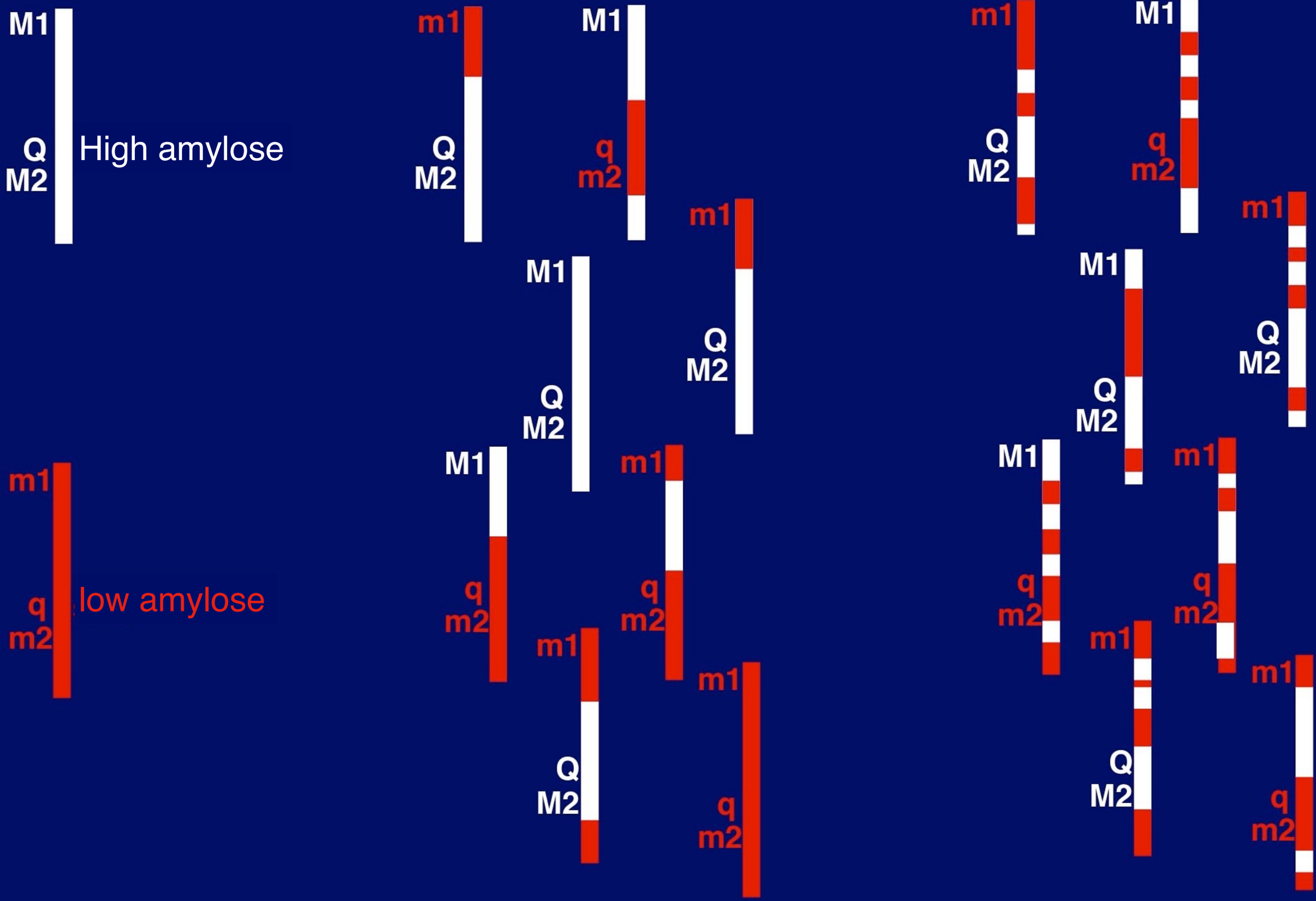
m1
Q
M2

m1
q
m2
low amylose

The importance of tagSNPs

- Our rice SNP data set has ~44,000 SNPs.
- There are ~500,000 SNPs segregating among the rice varieties.
- Is it hopeless?
  - Do we have less than a 1 in 10 chance of finding an association because we are assaying less than 10% of the SNPs?
- Not hopeless
  - Because of linkage disequilibrium there is a strong correlation among closely linked SNPs

Not hopeless: SNPs near to one another are correlated...



Not hopeless: SNPs near to one another are correlated...

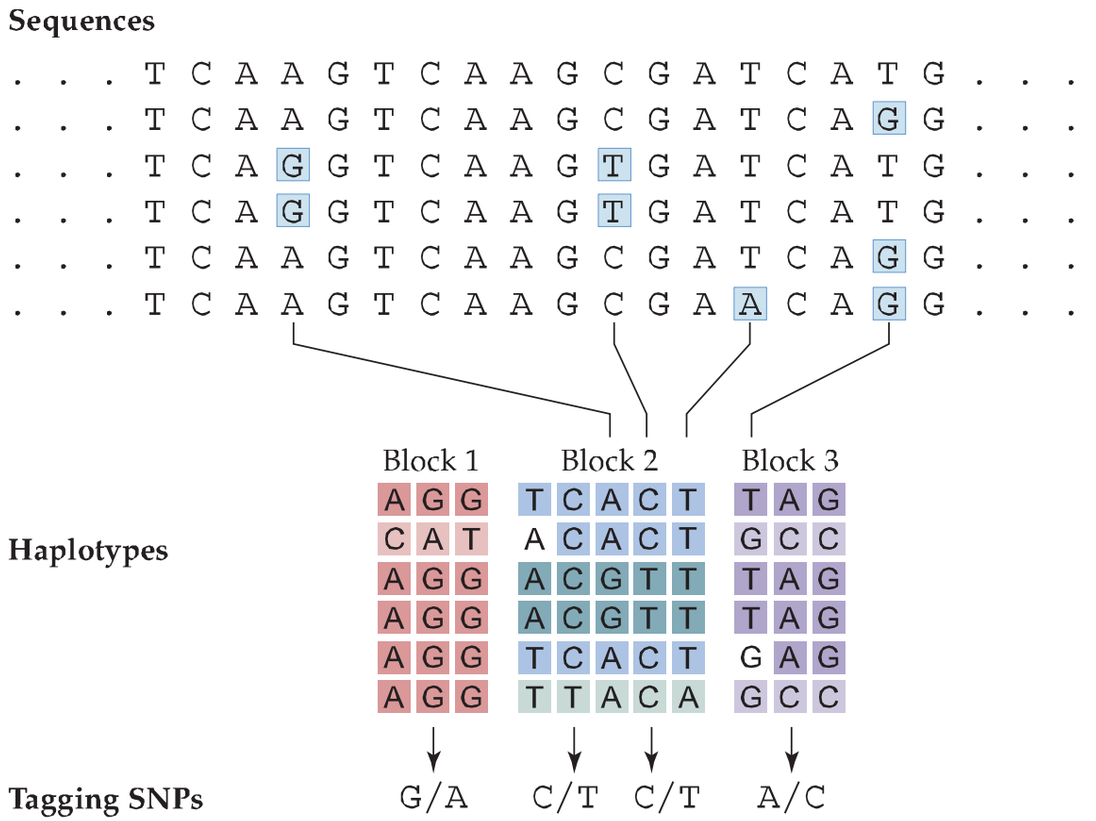
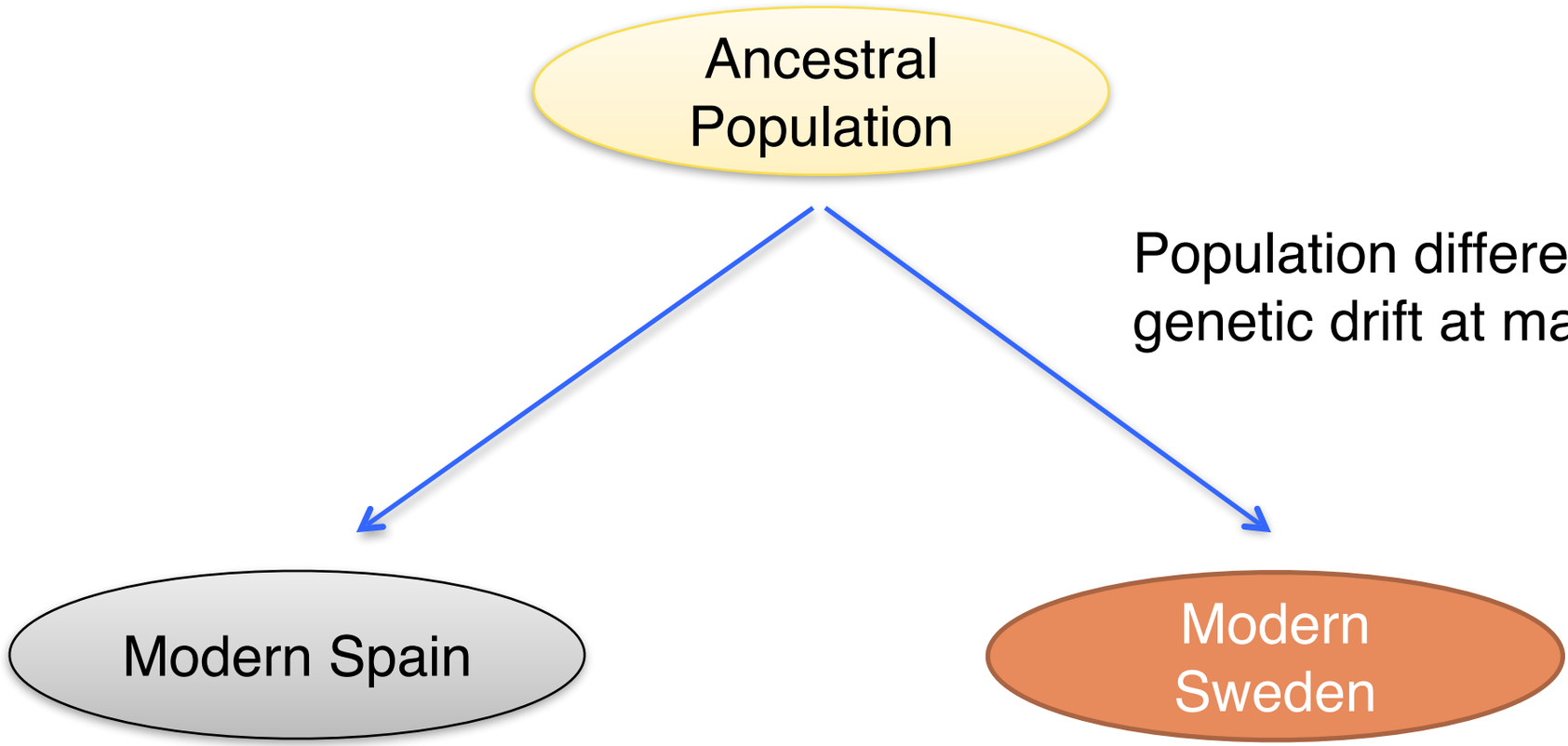


Figure 3.5 Tagging SNPs and haplotype blocks. Extraction of the polymorphisms from a set of sequences typically reveals a blocklike pattern of haplotypes. In this hypothetical example, Block 1 has two classes of haplotypes, one rare and one common; Block 2 has three classes of haplotypes; and Block 3 has two classes of haplotypes. Note that the boundaries between blocks are relatively sharp. The tagging SNPs can be used to define most of the variation in the sample.

the problem of population structure

- If a trait is correlated with population structure then many spurious associations will result.
- What if GWAS for hair color in Spain + Sweden?

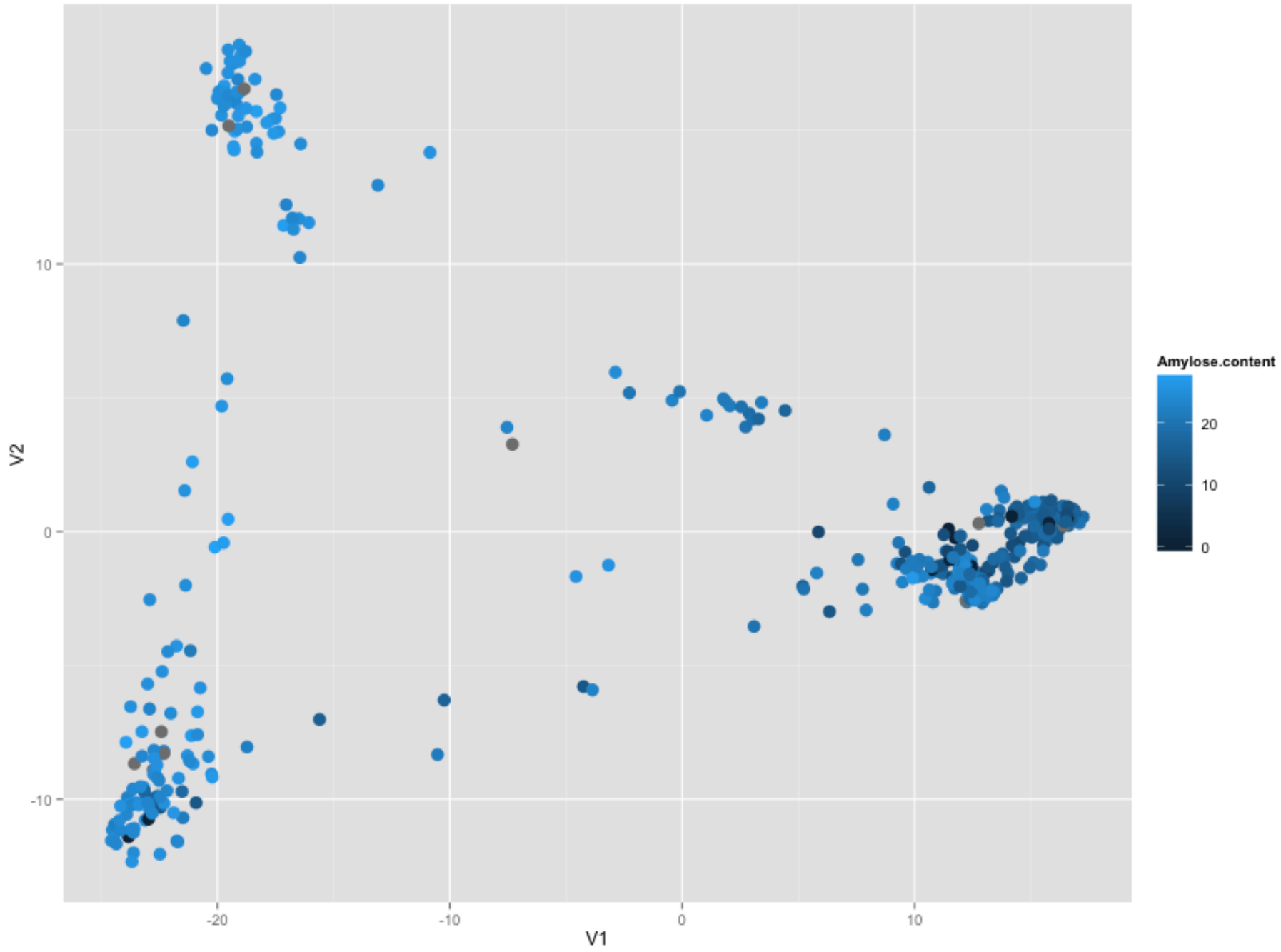


Because the Spanish and Swedish populations have been separated for a long time they will differ at many, many loci due to drift.

Almost none of these will have anything to do with hair color, but many of them will be associated with hair color in a GWAS due to population structure.

Population Structure can present a problem for GWAS

- What is the potential problem with a GWAS for amylose content?



Population structure corrections

- Do GWAS separately for each sub-population

  OR

- Include population information in the statistical GWAS model.

Include population structure information in GWAS model

- Three ways to include structure information:

- "Q" matrix
  - population assignment, e.g. from fastStructure

- Genotype principal components
  - Like the PCs we generated in the last lab

- Kinship matrix ("K")
  - Matrix of pairwise genetic relatedness (in our case 413 X 413 matrix)
  - Values represent genetic similarity and range from 0 to 1
    - Identical twins have kinship = 1
    - Siblings or parent/children have kinship = 0.5
    - Half-siblings have kinship = 0.25
  - If you don't know pedigree (we don't), then the Kinship matrix can be estimated from the SNPs

Include population structure information in GWAS model

- In formula notation:
  - amylose ~ SNPgenotype (no correction)
  - amylose ~ SNPgenotype + Q (population membership)
  - amylose ~ SNPgenotype + K (kinship matrix)
  - amylose ~ SNPgenotype + Q + K
- The statistical test is whether SNP is associated with amylose BEYOND what is expected given Q or K
- Often it is best to include BOTH Q and K

Q-Q Plot

- Q-Q plots can be used to determine if population structure correction is working
- Plots observed -log10 p-value VS -log10 p-value expected if no associations
- Because the VAST majority of the SNPs should not influence the trait, we expect the points to be on the 1:1 diagonal except at the very right hand side of the plot.
- Example of a good Q-Q plot