HOS 6236 Molecular Marker Assisted Plant Breeding

Fall 2017

Last Class:

Linkage Disequilibrium

Today's Class:

Population Structure

Genome Wide Association Studies (GWAS)

Linkage Disequilibrium

- **Two loci are in linkage equilibrium, if they are completely independently in each generation.**
- **If two genes are in linkage disequilibrium, it means that certain alleles of each gene are inherited together more often that would be expected by chance.**

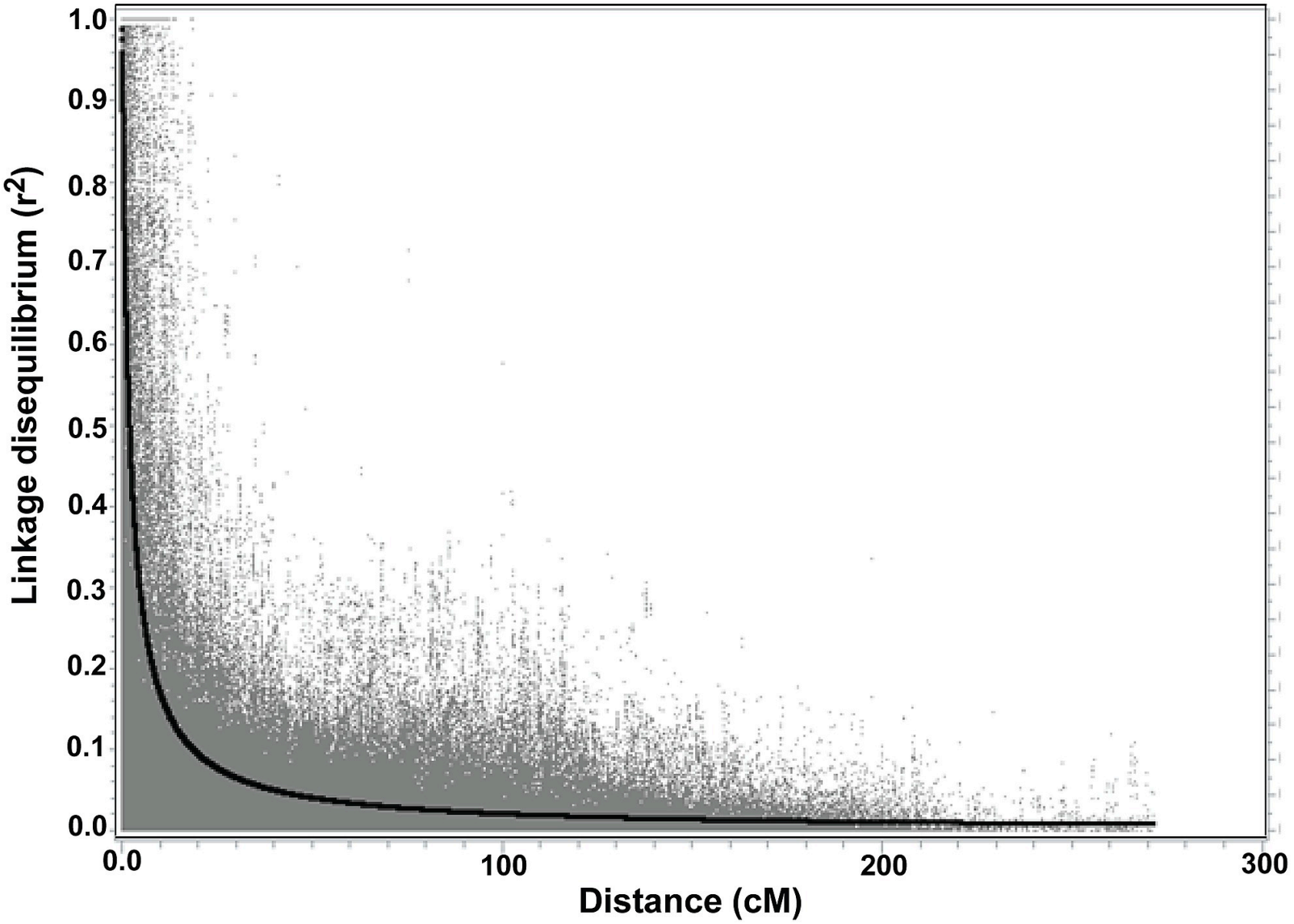
Independently segregating SNPs:

Haplotype Frequency p(ab) = p(a) x p(b)

LINKAGE DISEQUILIBRIUM

Haplotype Frequency p(ab)≠ p(a) x p(b)

Visualizing LD



Gurung *et al.* 2014

Creation of LD

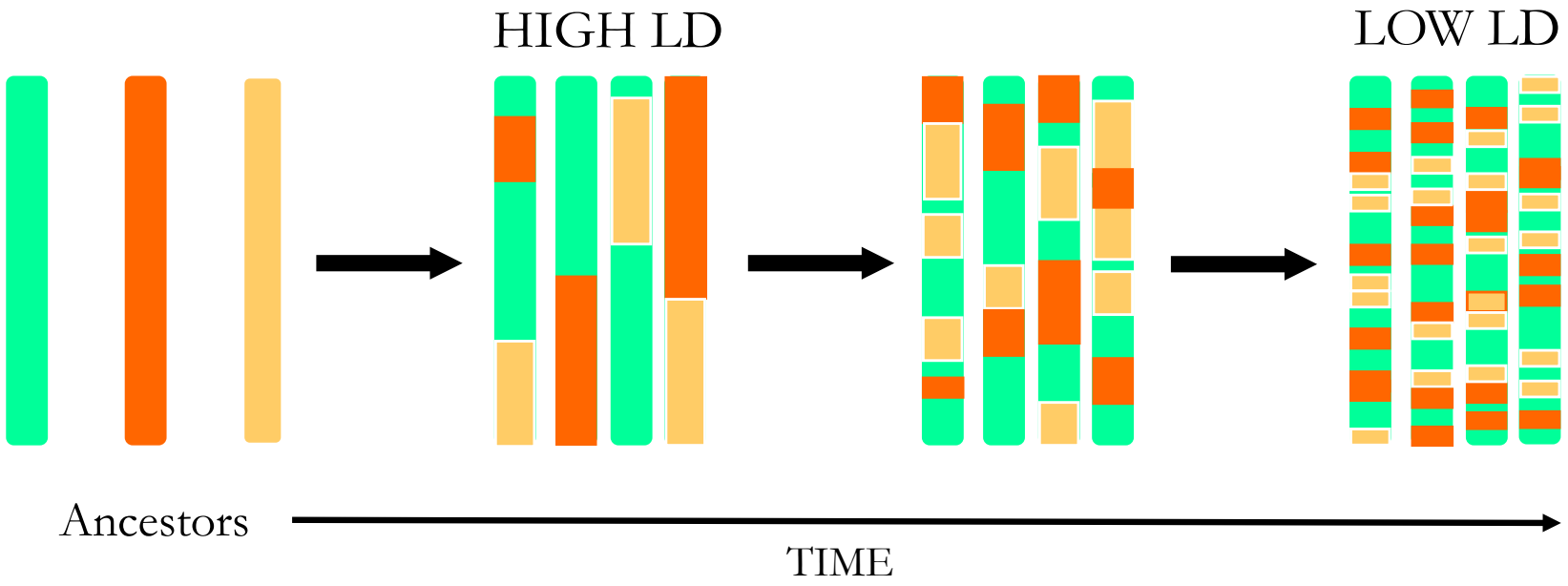
- Easiest to understand when markers are physically linked
- Creation of LD
  - Mutation
  - Admixture
  - Inbreeding / non-random mating
  - Selection
  - Population bottleneck or stratification
  - Epistatic interaction
- LD can occur between *unlinked* markers
- LD decays over time (generations of interbreeding)
- LD decay is a function of recombination frequency

Population Effect on Linkage Disequilibrium in Maize

Researcher	Population Studied	Extent of LD
Gaut	**Landraces**	**<1000 bp**
Buckler	**Diverse Inbreds**	**2000 bp**
Rafalski	**Elite Lines**	**100 kb**

Flint-Garcia, S. A. et al. 2003. Annual Review of Plant Biology 54:357-374.

Most significant issue in QTL and association genetics:

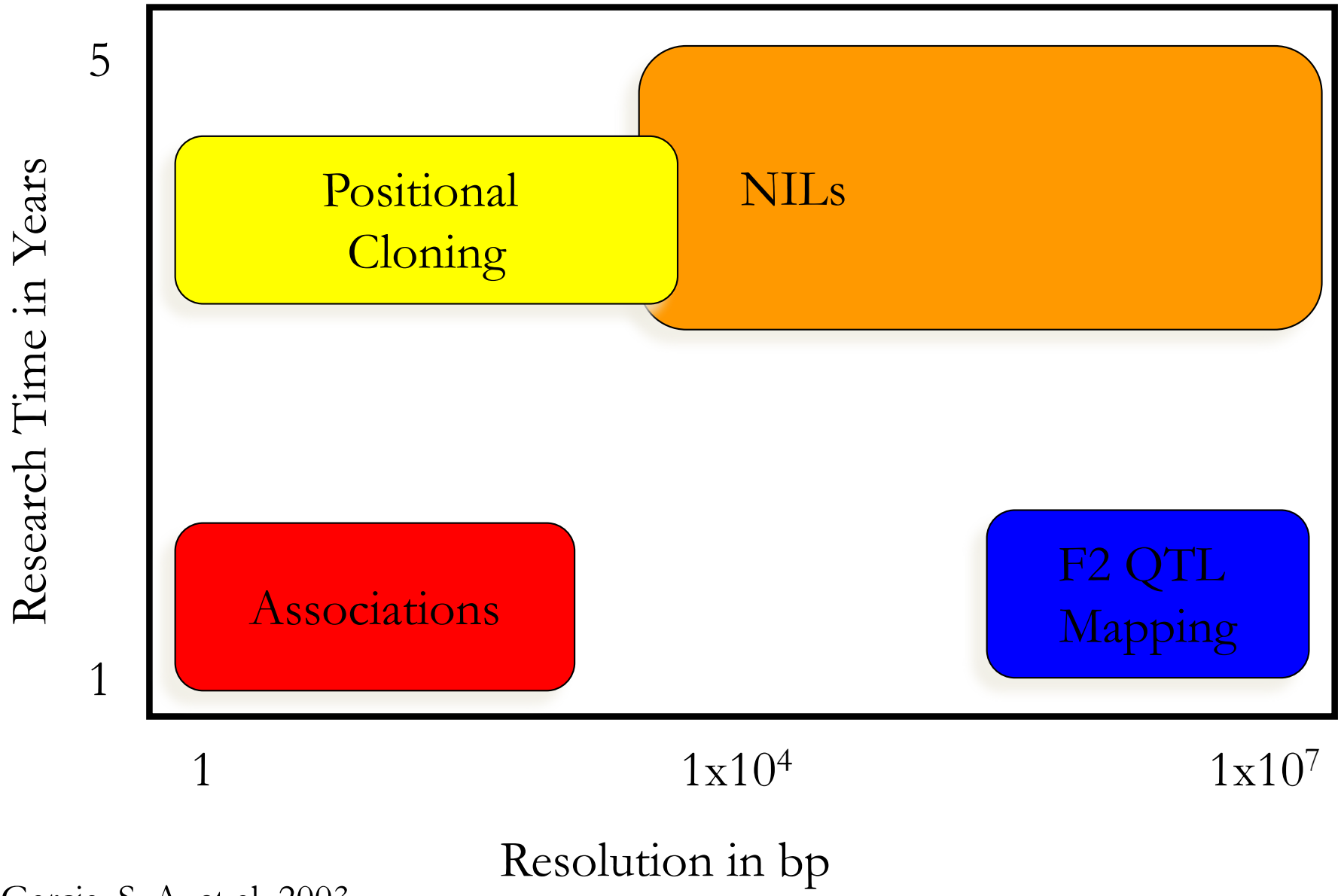


Effective population size (Ne): the number of breeding individuals in an idealized population showing similar properties as the population under consideration.

$$E(r^2) = 1/(4N_eL + 1)$$

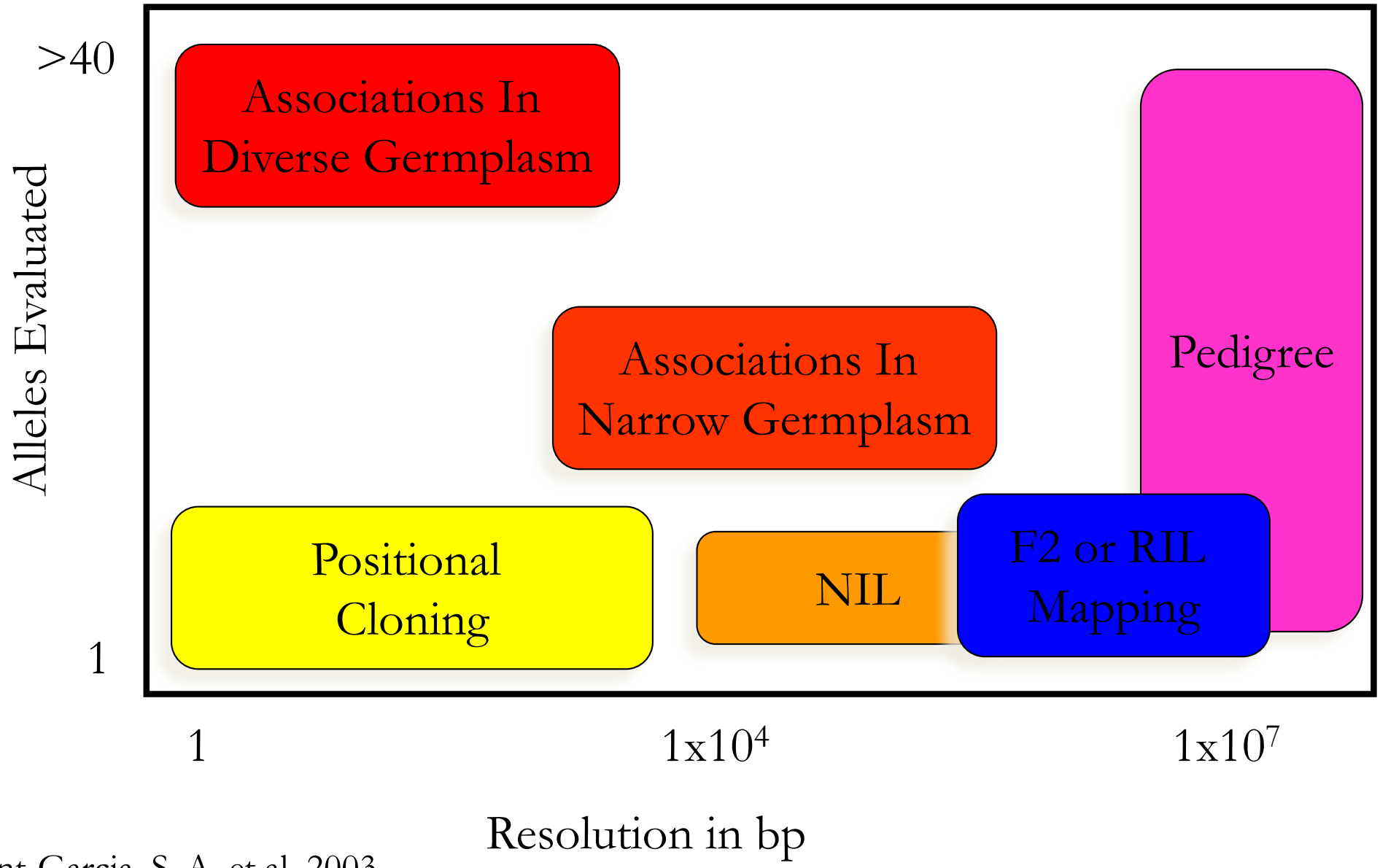
(Sved equation 1971)

Dissecting A Quantitative Trait: Time Versus Resolution



Flint-Garcia, S. A. et al. 2003

Resolution Versus Allelic Range



Flint-Garcia, S. A. et al. 2003

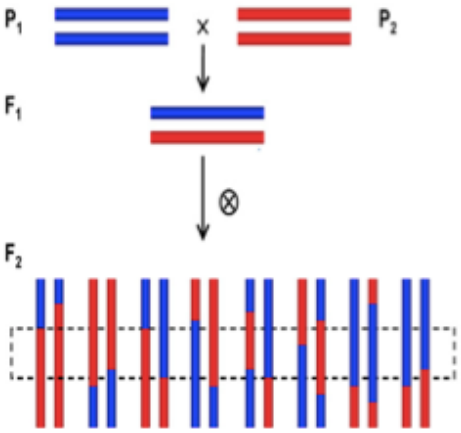
Genome Wide Association Study

- Goal: find connections between:
  - A phenotype: plant height, disease resistance, ..., trait known to be heritable.
  - Whole-genome DNA markers

- Specific goals are distinct:
  1. Identify statistical connections between points (or areas) in the genome and the phenotype
     - Drive hypotheses for biological studies of specific genes/regions in specific context
  2. Generate insights on genetic architecture of phenotype
     - Many small genetic effects dispersed across the genome?
     - Few large effects concentrated in one area?

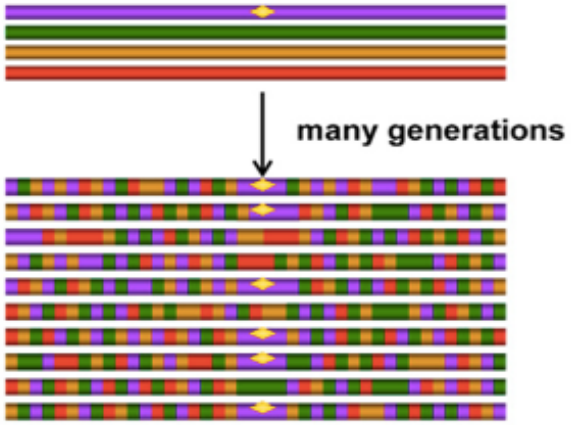
Genome Wide Association Study

QTL analysis



- Biparental mapping population
- Few recombination events
- Few alleles sampled (2 to 4 max.)
- Few traits segregate
- Few markers required
- High detection power
- Poor (no) QTL-to-gene resolution

GWAS



- Panel of lines/wild mapping population
- Historical recombination
- Many alleles sampled
- All traits segregate
- Many markers required
- Low detection power
- Potentially good QTL-to-gene resolution

QTL mapping vs GWAS

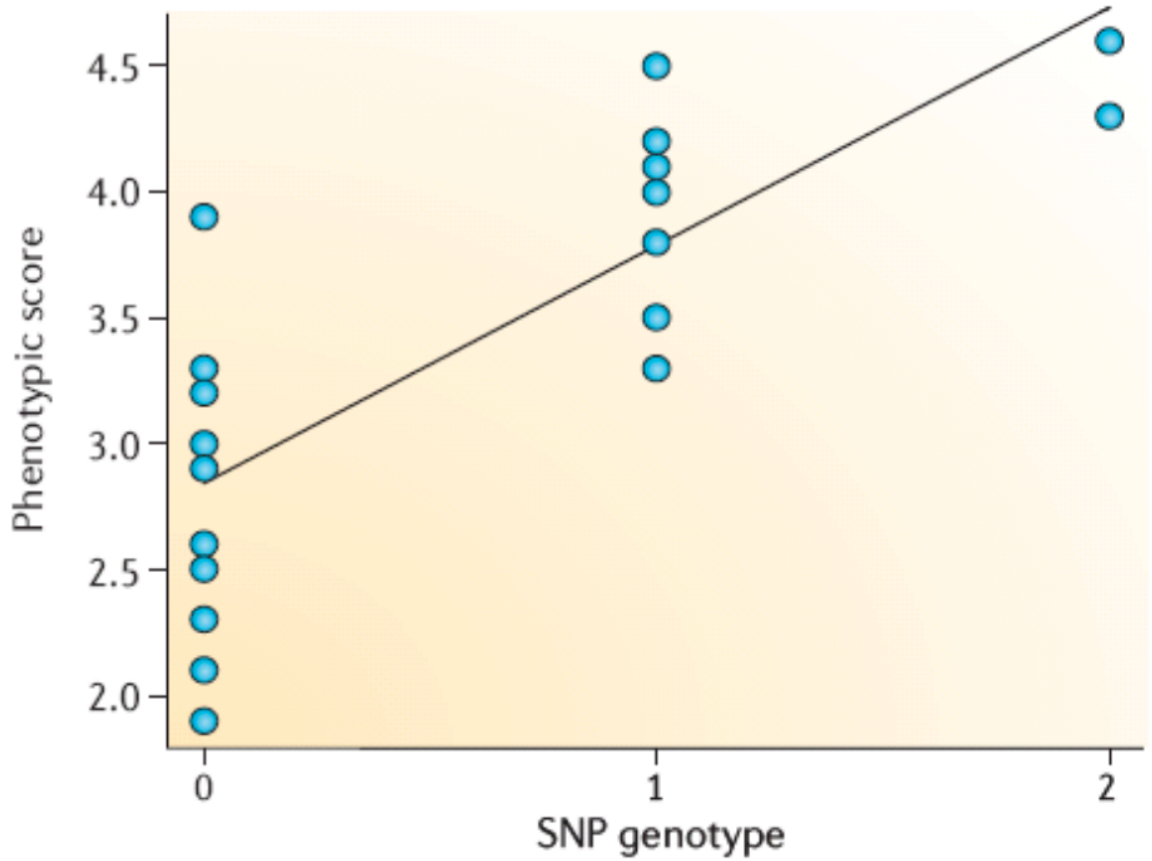
- **At the fundamental level, genetic association and linkage analysis rely on similar principles and assumptions**
- Both rely on the co-inheritance of adjacent DNA variants, i.e. LINKAGE DISEQUILIBRIUM (LD)
- **Linkage mapping: recently generated LONG RANGE LD** by identifying haplotypes that are inherited intact over generations in families or pedigrees of known ancestry
- **Association mapping: historical SHORT RANGE LD**, from the retention of adjacent DNA variants over many generations
- Association studies can be regarded as very large linkage studies of unobserved, hypothetical pedigree

Methodology

- Collect *n* plants with known phenotype (ideally *n* in range 10^3+ plants)
- Measure each one in *m* genomic locations ("representing **common** variation in the whole genome")
  - Usually SNPs: Single Nucleotide Polymorphisms
  - Ideally *m* in range 10^5-10^6
  - Recently moving to whole genome sequencing
- Now we can think of our data as X_{n*m} matrix with plants as rows, SNPs as columns,
  - X_{ij} is in {0,1,2} (genotype at single locus)
  - Also given extra vector Y_n of phenotypes
- Our first task: association testing
  - Find SNPs (columns in X) that are statistically associated with Y
  - Can be thought of as *m* separate statistical tests run on this matrix

Methodology

- For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as the covariate



Single Marker Regression

- Linear Regression Model defined as:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Y = Phenotype

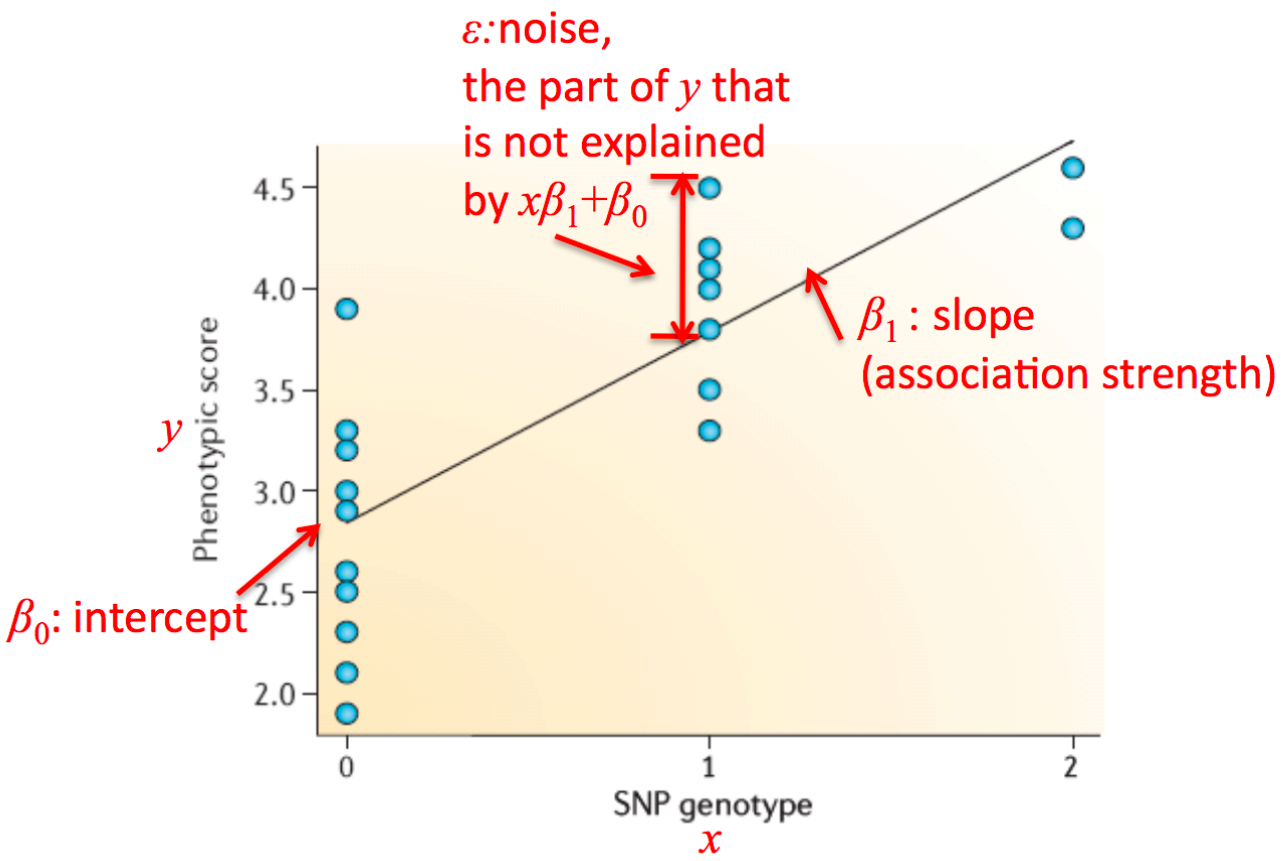
X = SNP genotype at a given locus

β_1= Regression coefficient representing the strength of the association between the SNP X and the phenotype Y

β_0= Intercept

ε = Noise (error term)

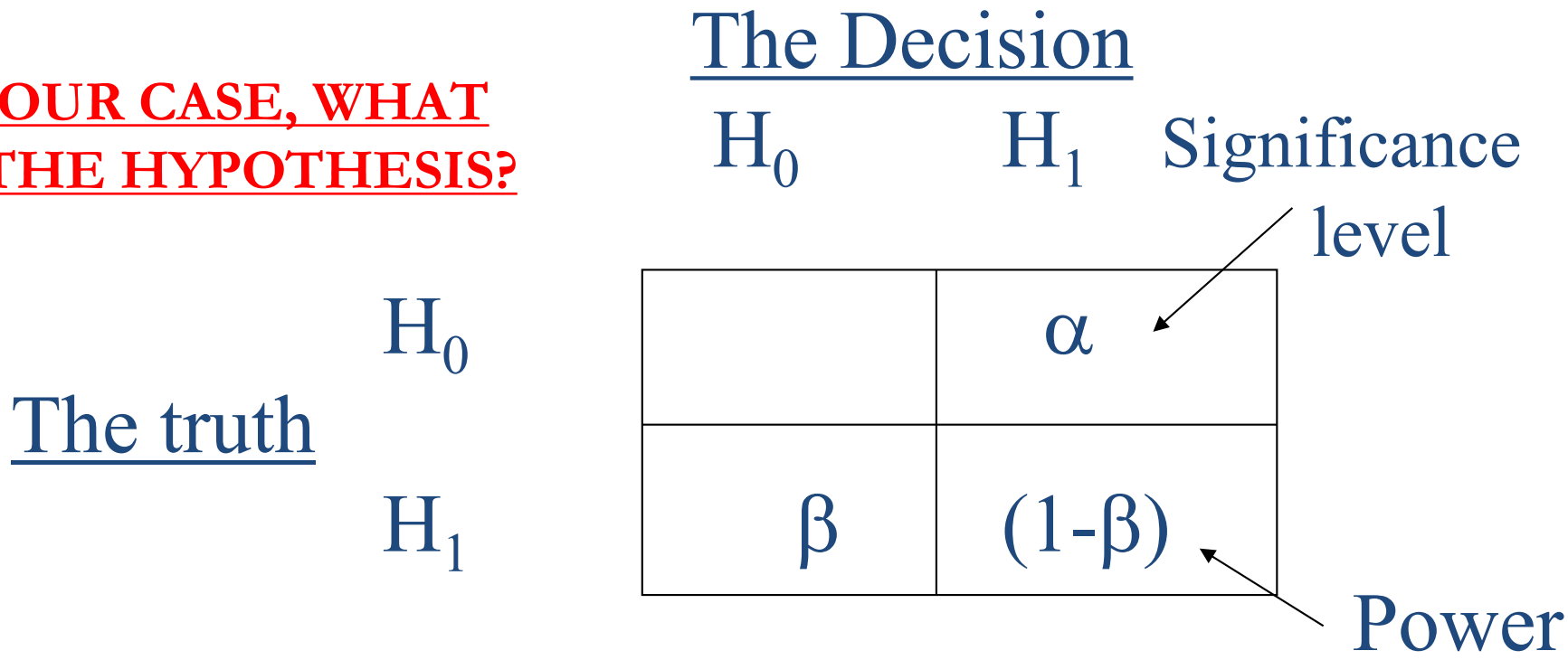
Single Marker Regression



Statistical Digression

- The p-value
  - Under the null hypothesis the probability that you observe your data or something more extreme
- Decision: Reject the null - fail to reject the null

IN OUR CASE, WHAT IS THE HYPOTHESIS?



Multiple testing

There are usually many thousand makers being tested. For each marker: a p-value is calculated.

What is a statistically sound choice of a threshold for declaring an association?

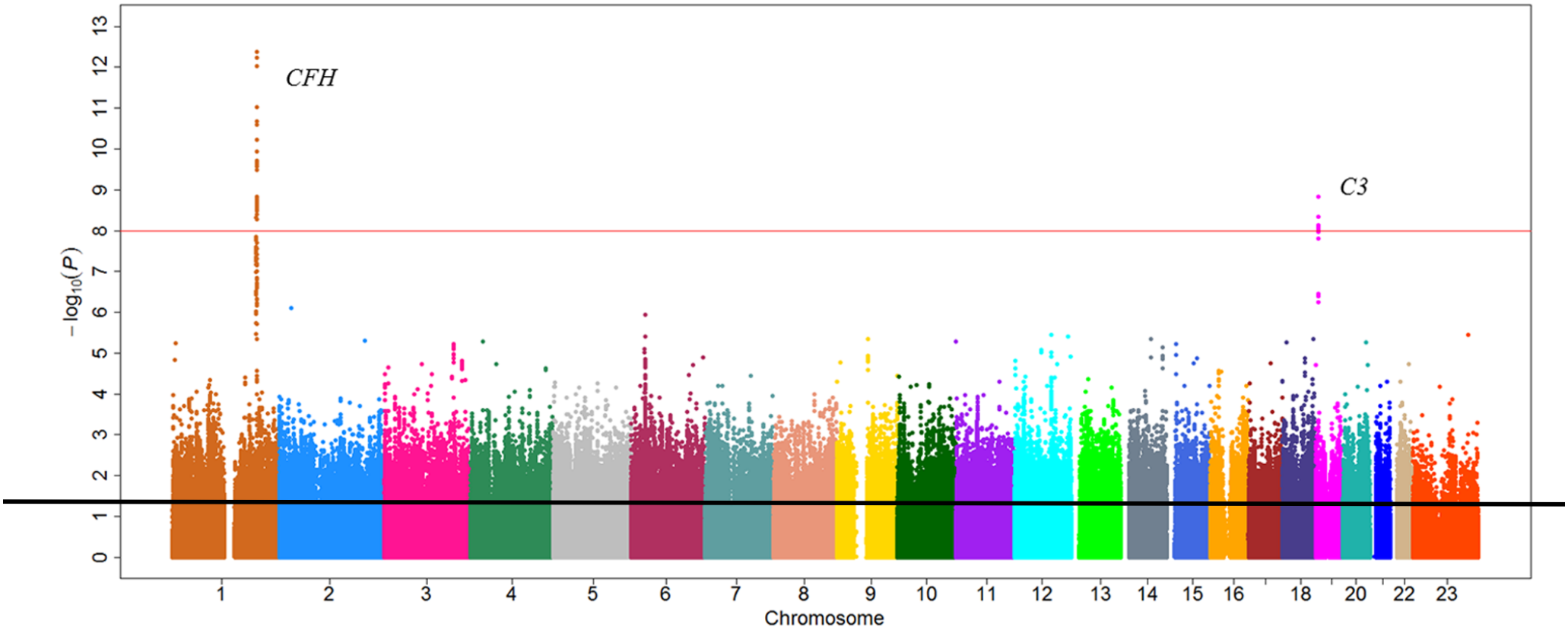
•Family wise error rate (FWER): the probability of making even one false discovery out of our m tests.

•Controlling FWER: the well known Bonferroni correction, perform each test at level $a = 0.05/m$

- For $m = 10^6$ this gives $a = 5 \times 10^{-8}$

•What happens if we use a p-value threshold of α=0.05 (black line) to declare results as significant?

"Manhattan plot" of GWAS results



What happens if we use a p-value threshold of α=0.05 (black line) to declare results as significant?

Missing Heritability

Published in final edited form as:
Nature. 2009 October 8; 461(7265): 747–753. doi:10.1038/nature08494.

Finding the missing heritability of complex diseases

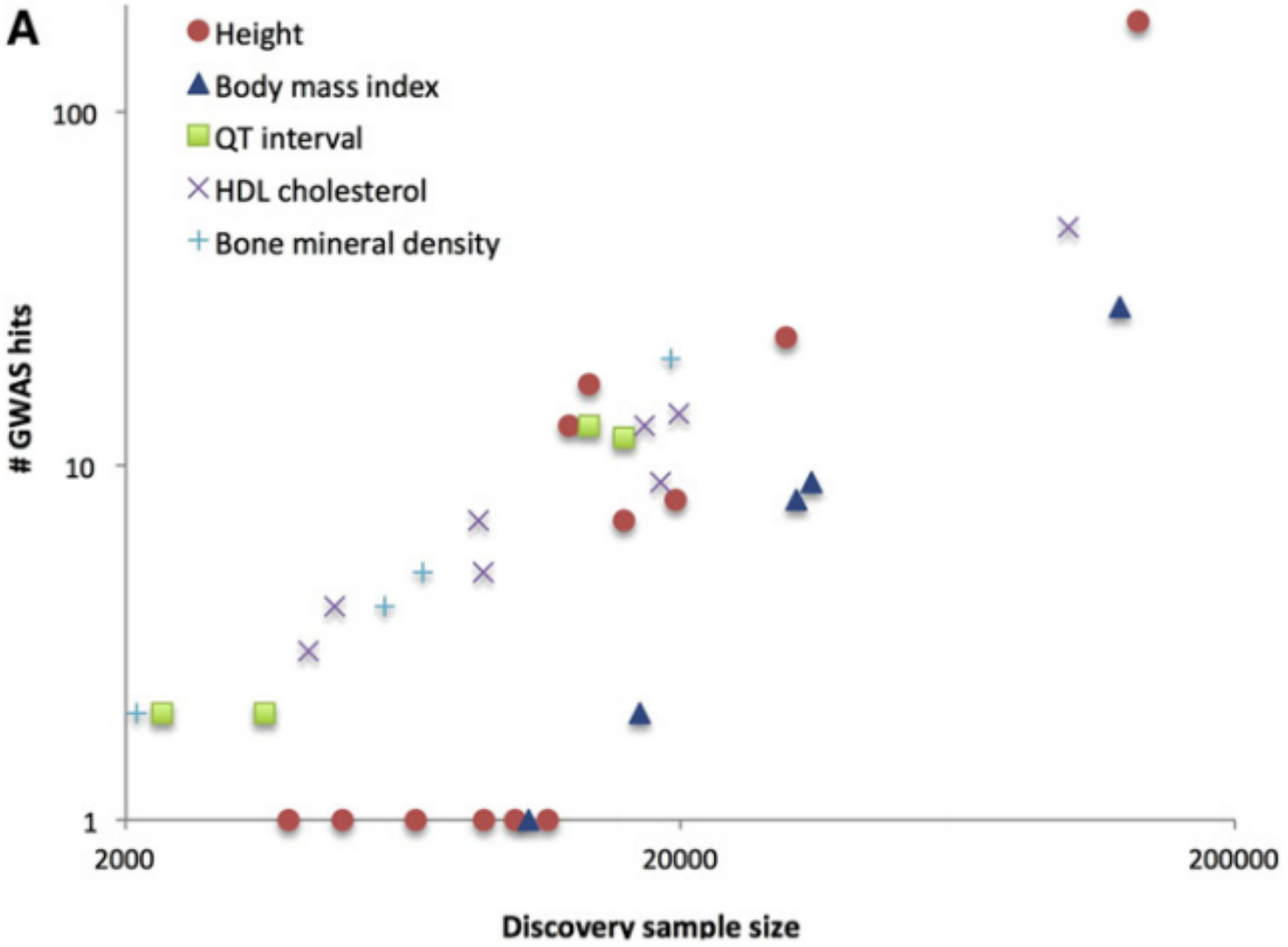
Teri A. Manolio[1], **Francis S. Collins**[2], **Nancy J. Cox**[3], **David B. Goldstein**[4], **Lucia A. Hindorff**[5], **David J. Hunter**[6], **Mark I. McCarthy**[7], **Erin M. Ramos**[5], **Lon R. Cardon**[8], **Aravinda Chakravarti**[9], **Judy H. Cho**[10], **Alan E. Guttmacher**[1], **Augustine Kong**[11], **Leonid Kruglyak**[12], **Elaine Mardis**[13], **Charles N. Rotimi**[14], **Montgomery Slatkin**[15], **David Valle**[9], **Alice S. Whittemore**[16], **Michael Boehnke**[17], **Andrew G. Clark**[18], **Evan E. Eichler**[19], **Greg Gibson**[20], **Jonathan L. Haines**[21], **Trudy F. C. Mackay**[22], **Steven A. McCarroll**[23], and **Peter M. Visscher**[24]

Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual*phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

*Residual is after adjustment for age, gender, diabetes.

Missing Heritability



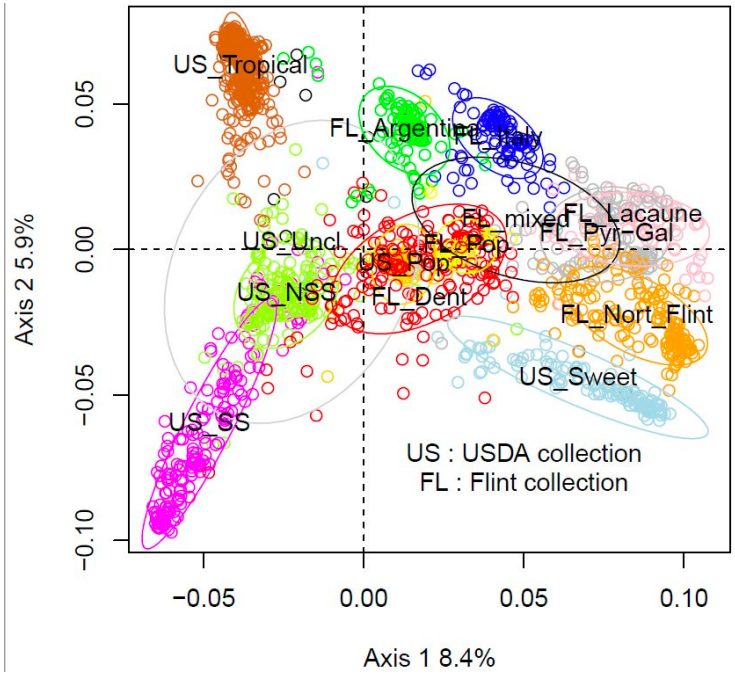
Visscher PM, AJHG 2012

Population structure

Random genetic drift, mutation, migration and selection, and other forces that change gene frequencies within subpopulation can lead to population structure – i.e. unequal gene and genotype frequencies among subpopulations, with variable levels of similarity between them.

To know if there is population structure, and what level of similarity exists among subpopulation is relevant in several fields, such as …

- GWAS
- Breeding
-…others



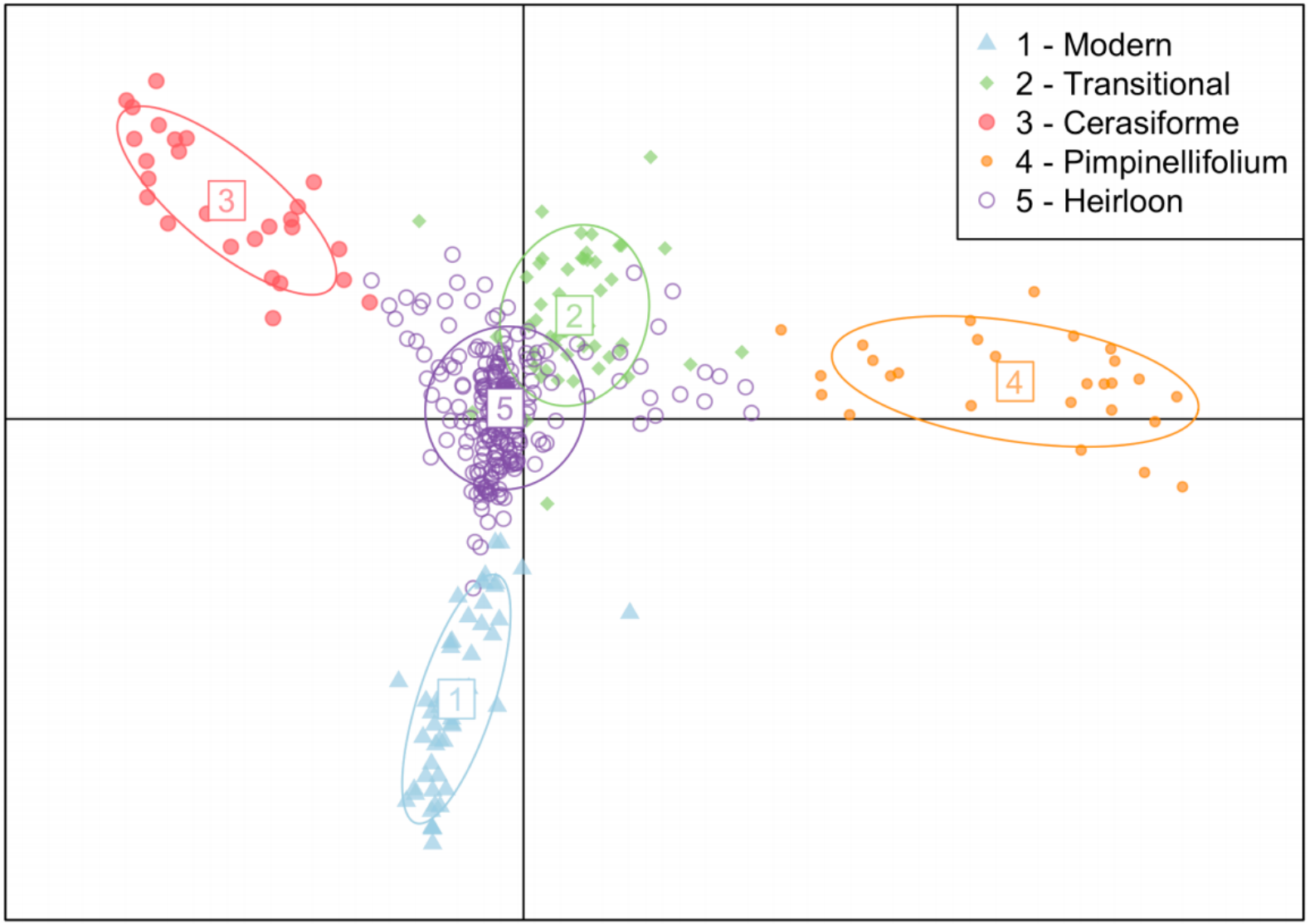


Figure S1. Population structure based on the discriminant analysis of principal components.

Tieman *et al.* 2017

The importance of genetic structure

- Genetic structure: not every individual in your population is from same genetic background
  - Some plants are more genetically similar than others
  - Geographical distance, selection patterns, environmental characteristics
- We are seeking associations that are not "due to structure"
  - How can we eliminate ones that are due to it?
- If we know which plant is " temperate" and which is " tropical", we can do an analysis that controls for the "origin" variable
  - For example, linear regression with both the origin and the SNP as predictors

What happens if we don't know?

- We can estimate population structure from the data and then control for it.
- This is typically done in all GWAS analyses using different methods
- Handling Population structure:
  - genomic control (Devlin & Roeder, 2001)
    - Estimating the degree of overdispersion generated by structure.
  - structured association (Pritchard et al)
    - Probabilistically assigns individuals to latent subpopulations.
  - principal components (Price et al)
    - Explicitly model ancestry differences by the main principal components

How can we account for relatedness

- As a fixed effect that describes the relationship between observations
- As a random effect (in the model or the error) that describes the expected covariance

- Is there one "better" way?
  - Subject of much literature and discussion. More on that next week

STRUCTURE

Inference of Population Structure Using Multilocus Genotype Data

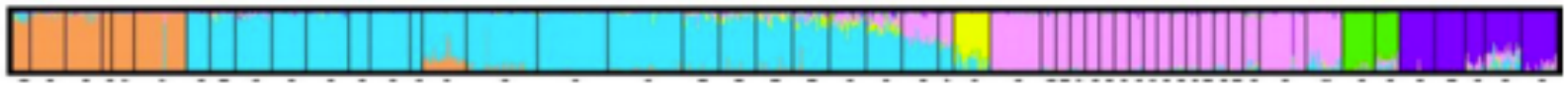
Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received September 23, 1999
Accepted for publication February 18, 2000

ABSTRACT

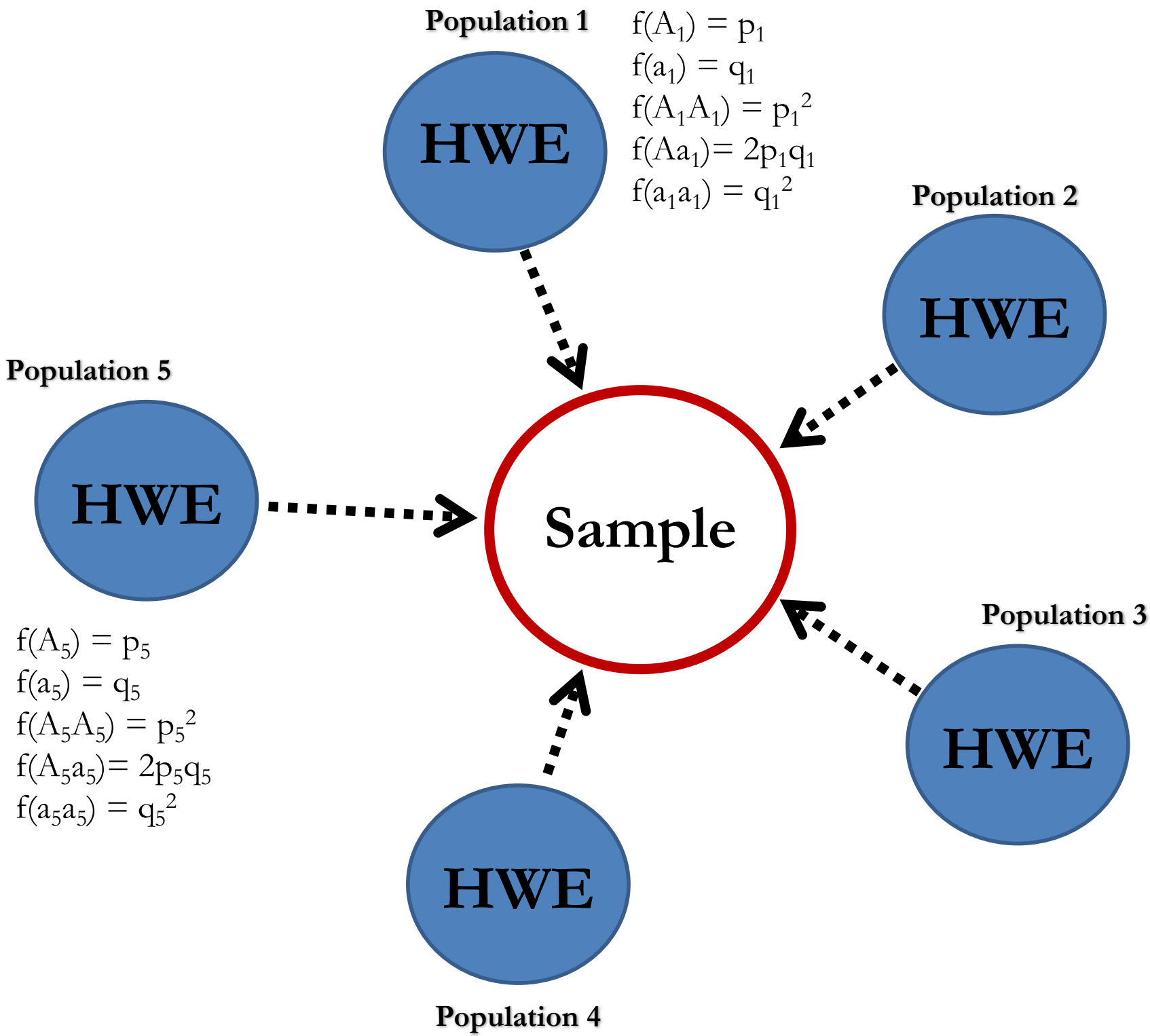
We describe a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by a set of allele frequencies at each locus. Individuals in the sample are assigned (probabilistically) to populations, or jointly to two or more populations if their genotypes indicate that they are admixed. Our model does not assume a particular mutation process, and it can be applied to most of the commonly used genetic markers, provided that they are not closely linked. Applications of our method include demonstrating the presence of population structure, assigning individuals to populations, studying hybrid zones, and identifying migrants and admixed individuals. We show that the method can produce highly accurate assignments using modest numbers of loci—*e.g.*, seven microsatellite loci in an example using genotype data from an endangered bird species. The software used for this article is available from http://www.stats.ox.ac.uk/~pritch/home.html.



http://pritchardlab.stanford.edu/structure.html

Inferring Population Structure

- Assume admixed population with K contributing founding populations.
- Each of the K founding populations was in equilibrium (HWE, no LD).
- In this generation, each allele copy originated in one of the founding populations.
- Want to figure out the probability alleles in each individual originated in population *k*: **Q** vector.


Population 1
HWE
$f(A_1) = p_1$
$f(a_1) = q_1$
$f(A_1A_1) = p_1^2$
$f(Aa_1) = 2p_1q_1$
$f(a_1a_1) = q_1^2$
Population 2
HWE
Population 5
HWE
Sample
$f(A_5) = p_5$
$f(a_5) = q_5$
$f(A_5A_5) = p_5^2$
$f(A_5a_5) = 2p_5q_5$
$f(a_5a_5) = q_5^2$
Population 3
HWE
HWE
Population 4

Inferring Population Structure

- Model-based clustering methods

- assumes that the data represents a random sample drawn from a parametric model – inferences for the parameters are made by trying to model the ideal clustering membership of each sample
- each clustering is defined by sets of individuals that minimize the deviation from HWE
- allow for inclusion of prior information

By far the most popular software, STRUCTURE, was developed by Pritchard

http://pritchardlab.stanford.edu/structure.html

Estimating K

Problem: Difficult to estimate allele frequencies, admixture proportions and number of groups simultaneously

Suggested by Pritchard:
Phase 1: Define number of *k* populations to test.
Phase 2: Examine clustering of individuals to evaluate appropriateness of the number of selected populations.

Modeling decisions:

- How long to run the program

  - burnin length (how long to run before start collecting data): 10,000-100,000 adequate, but verify summary statistics (α, F, D_{ij}, and likelihood)

  - run length: several runs at each K, at different lengths, and verify if answers are consistent through runs – 10,000-100,000 adequate for estimating number of sub-populations – 1,000,000 or more for good estimates of Pr(X|K)

Modeling decisions:

- Ancestry model

  - No admixture – i.e. individuals are discretely from one population or another – output is the posterior probability of one individual *i* belonging to population K.

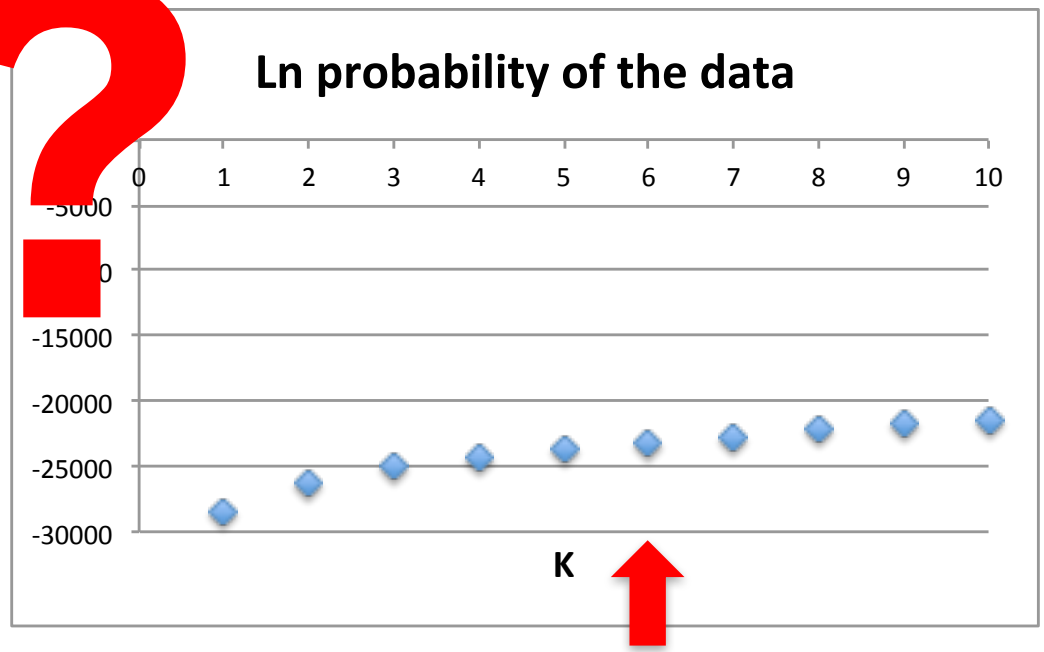
  - Admixture – i.e. each individual draws some fraction of his/her genome from each of the K populations – output is the posterior mean estimate of proportion of the individuals genome that originated from population K – recommended as a starting point for most population structure studies.

  - Using prior information – allows structure to use information about sampling locations: either to assist clustering with weak data, to detect migrants, or to pre-define some populations

STRUCTURE Results:

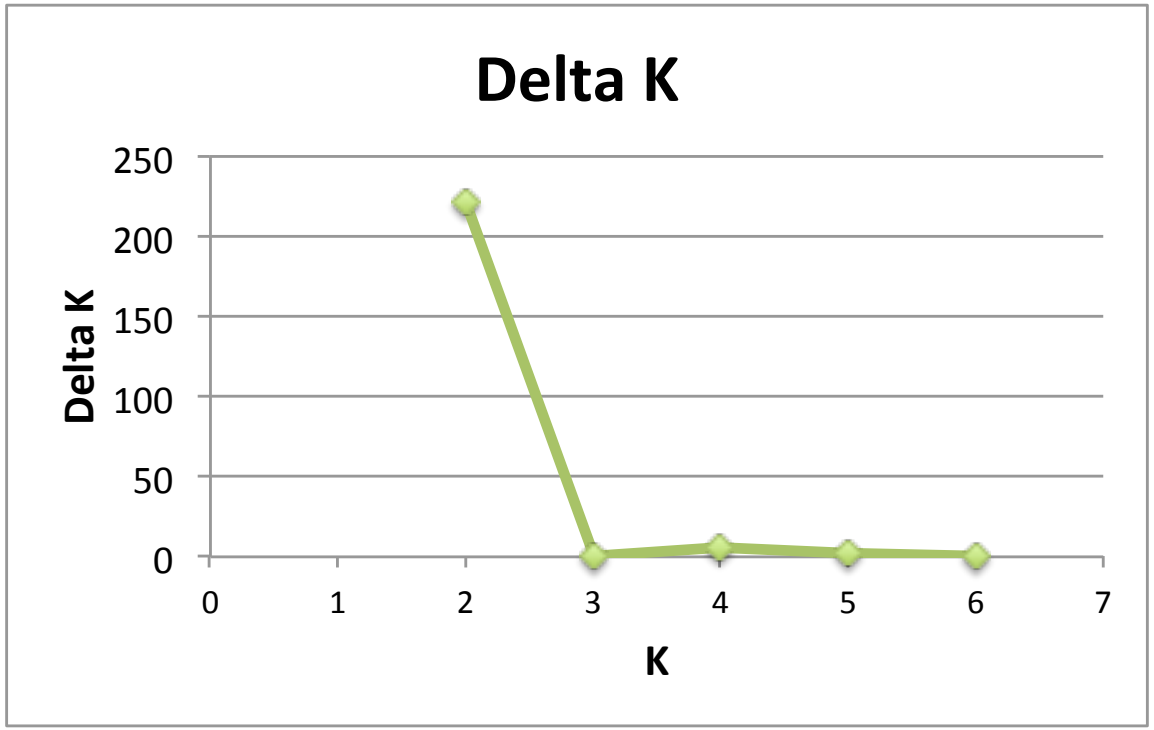
- ln of the probability of the data for each K
- When the probability ceases to improve, that is the optimal K

K	Ln
1	-28538.9
2	-26267.2
3	-24965.9
4	-24256.3
5	-23622.5
6	-23094.2
7	-22685.7
8	-22018.6
9	-21669.4
10	-21503.3



Analysis of STRUCTURE results with the ΔK method (Evanno et al., 2005):

- The ΔK method is implemented in the **Structure Harvester** tool available online (http://taylor0.biology.ucla.edu/structureHarvester/)
- Requires results for at least 3 iterations for each K to calculate standard deviation for the estimated ln of the probability of the data



Issues of GWAS

- Population stratification
- Multiple Testing: False Positives
- Gene-Environmental Interaction
- High Costs

Populus deltoides association population

- 500+ unrelated individuals
- Collected in 15 States of the Southeastern US

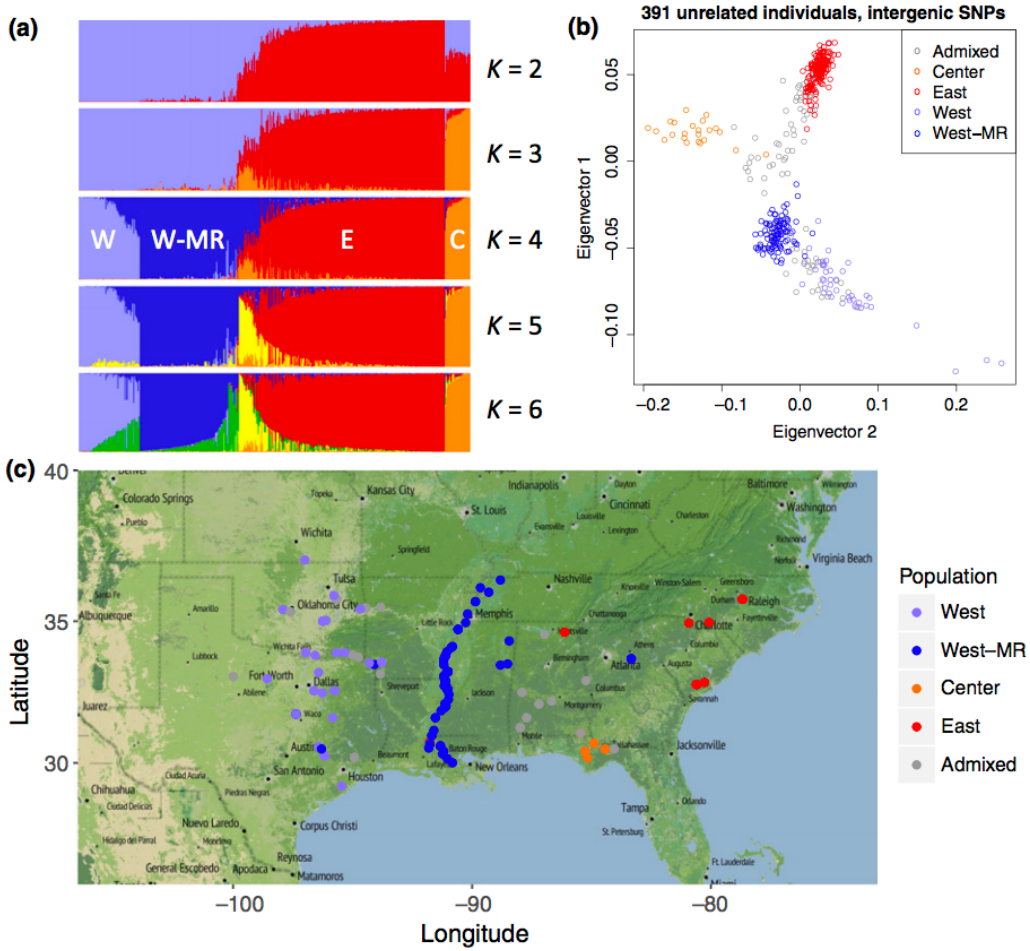
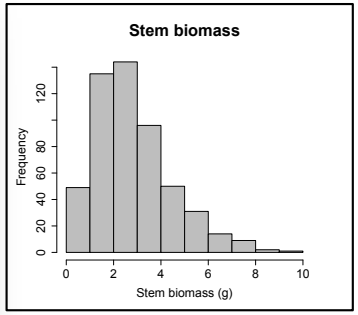
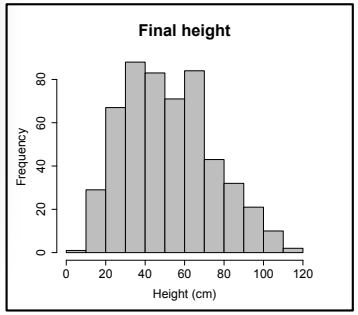
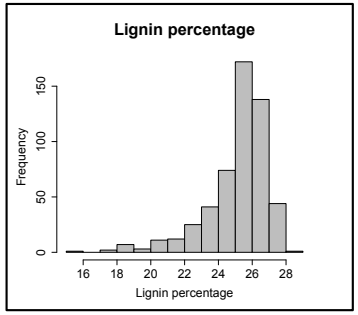


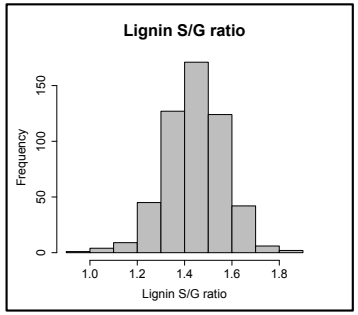
Fig. 1 *Populus deltoides* population structure. (a) Ancestry coefficient bar plots obtained with STRUCTURE for an assumed number of subpopulations (*K*) from two to six. The four subpopulations considered to correct for population structure in the association analyses are labeled in the bar plot for $K = 4$. E, east subpopulation; C, center subpopulation; W, west subpopulation; W-MR, west-Mississippi River subpopulation. (b) Population structure revealed by principal component analysis and (c) map showing 168 individuals with known sampling location. Samples in (b) and (c) are colored according to the subpopulation they were assigned to by STRUCTURE at $K = 4$.

Fahrenkrog *et al.* 2017









Growth

- Height
- Diameter

Biomass

Leaves

Stem

Wood composition

Lignin percentage

Lignin S/G ratio

5-C and 6-C sugars

Capture-Seq - Genotyping by targeted sequencing

Probe design

Library preparation

Multiplexing and sequence capture

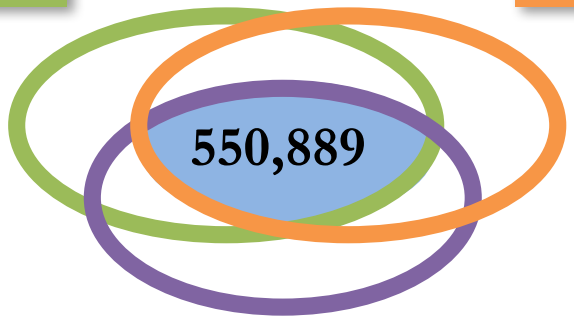
Capture-Seq - Genotyping by targeted sequencing

- ✓ **18,153 genes** expressed in poplar vegetative tissue
- ✓ **23,835 intergenic regions** distributed every 15 Kb

SNP identification

GATK
1,385,938

Freebayes
1,483,939

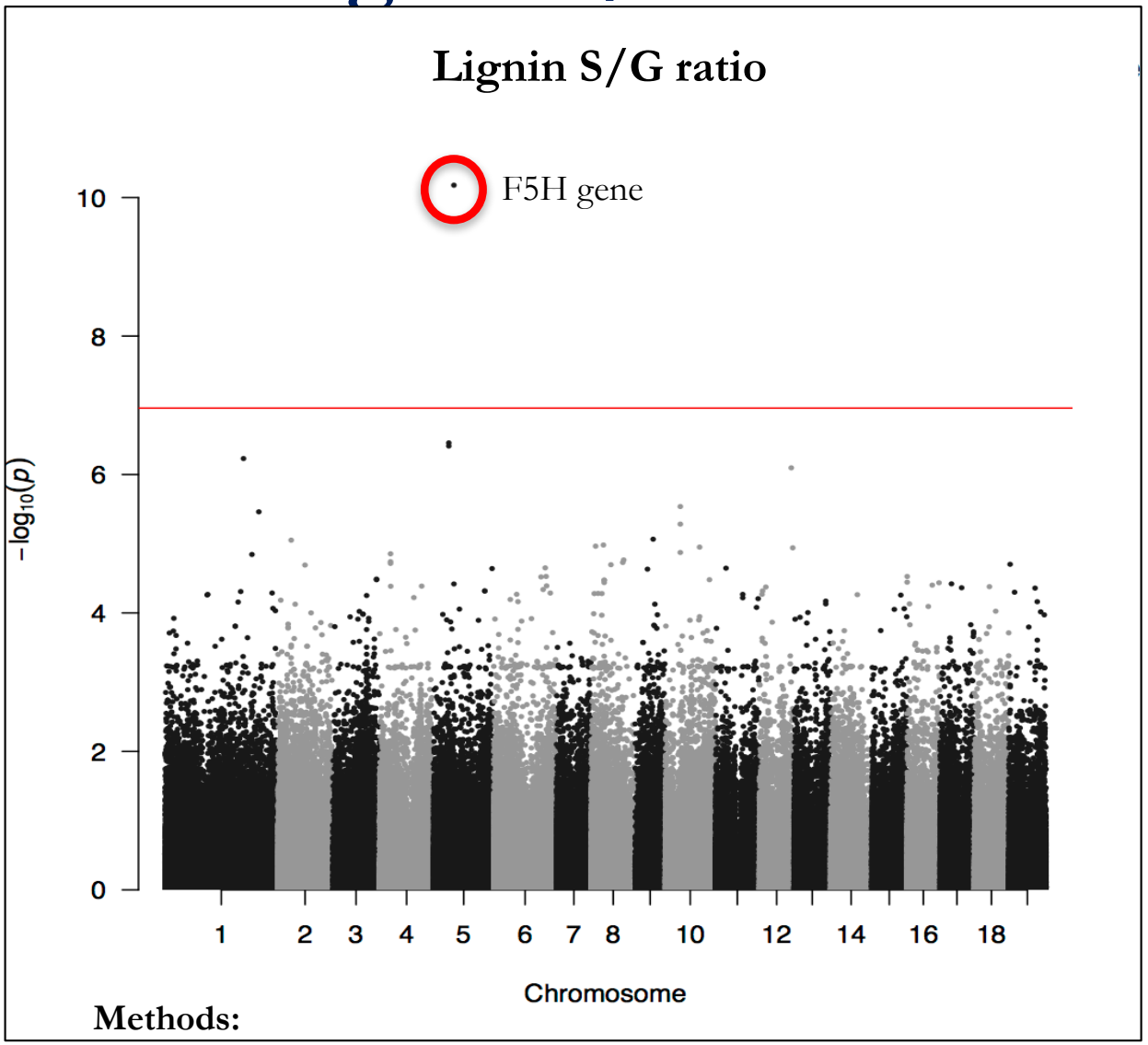


Samtools
746,338

Consensus
477,171

- ✓ SNPs on target regions
- ✓ 550,889 total SNPs
- ✓ 477,171 biallelic segregating SNPs
- ✓ 414,921 SNPs in genes
- ✓ 62,250 SNPs in intergenic regions
- ✓ 55.3% low frequency variants (MAF < 5%)
- ✓ 16.8% rare variants (MAF < 0.5%)

Lignin S/G ratio



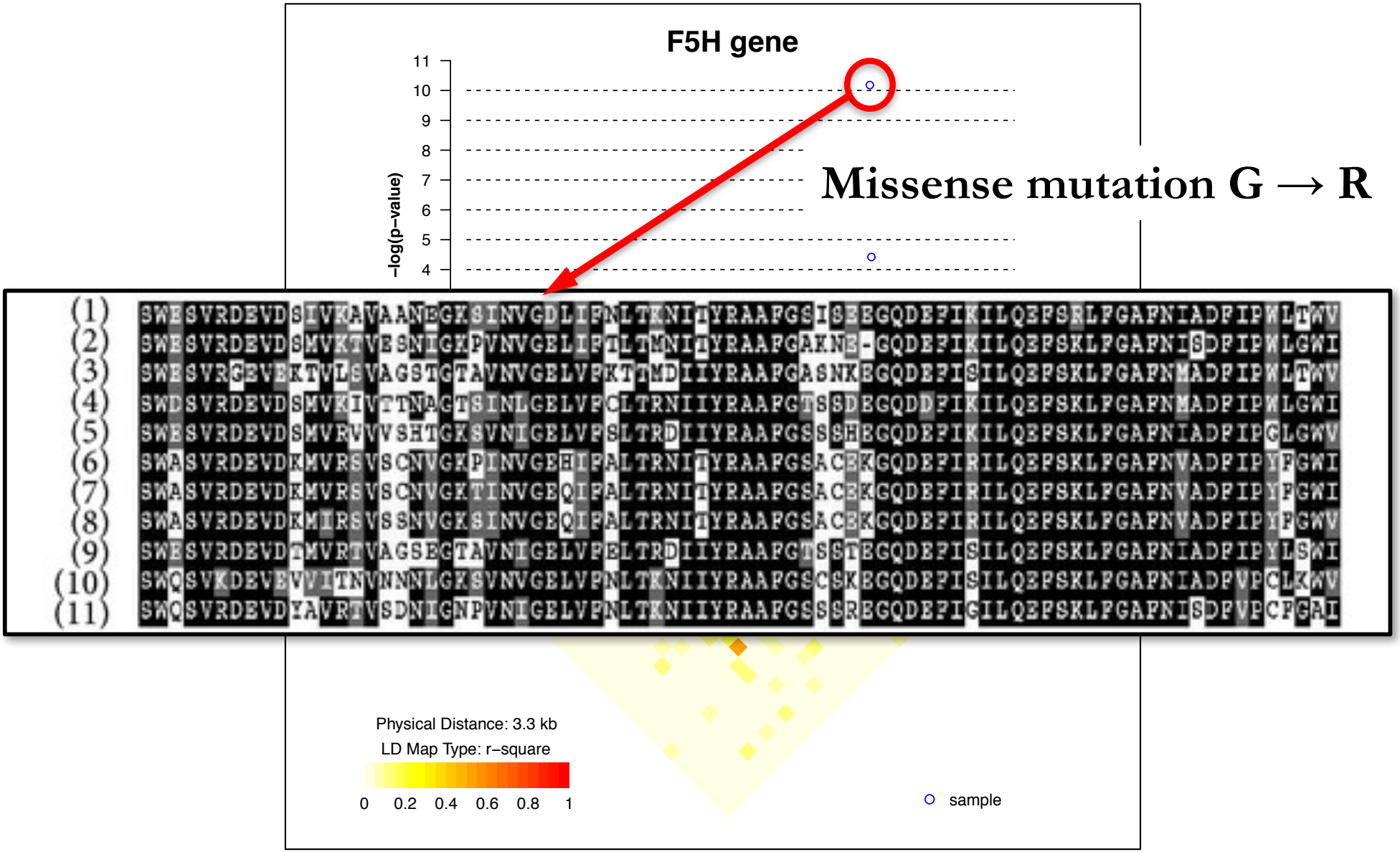
Methods:

Single-variant association analysis with PLINK software

Linear model with STRUCTURE ancestry as covariate

Bonferroni multiple-testing correction (red line: $0.05/455{,}573 = 1.10 \times 10^{-7}$)

Lignin S/G ratio



doi: 10.1155/2013/421578

Lignin percentage