Linkage Disequilibrium

Why do we care about linkage disequilibrium?

- Determines the extent to which association mapping can be used in a species
    - Long distance LD
        - Mapping at the tens of kilobase level or greater
    - Short distance LD
        - Mapping at the base pair to kilobase level

Linkage disequilibrium (LD)

- Measures the degree to which alleles at two loci are associated
    - The non-random associations between alleles at two loci
        - Based on expectations relative to allele frequencies at two loci

Goal: To Define the a statistical variables that will allow us

- To determine if two loci are in
    - Linkage disequilibrium or
    - Linkage equilibrium
- Frequencies of each haplotype are used in the variable calculation.

Table 1. Definition of haplotype frequencies for two loci with two alleles.

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

From this table

- The frequency of each allele at each locus can be calculated
    - Using traditional population genetic nomenclature
        - p and q for
            - Allele frequencies at loci *A* and *B.*

Table 2. Definition of allele frequencies based on haplotype frequencies.

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} = x_{22}$

To measure linkage disequilibrium (LD)

- Compare the observed and expected frequency of one haplotype
- Standard measure of LD is typically

$$D = x_{11} - p_1 q_1$$

- If two loci are in linkage equilibrium, then

$$D = 0$$

- If the two loci are in linkage disequilibrium, then

$$D \neq 0$$

From the definition of *D*

- We can determine
  - The relationship of haplotype frequencies (Table 1) and ***D*** and allelic frequencies (Table 2).

Table 3. Relationships among haplotype and allelic frequencies.

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	

***D* depends on allele frequencies**

- Researchers suggested the value should be normalized
  - Based on the theoretical maximum and minimum relative to the value of *D*

- When $D \geq 0$

$$D' = \frac{D}{D_{max}}$$

D_{max} is the smaller of p_1q_2 and p_2q_1.

- When $D < 0$

$$D' = \frac{D}{D_{min}}$$

D_{min} is the larger of $-p_1q_1$ and $-p_2q_2$.

Another LD measure

- Correlation between a pair of loci is calculated using the following formula
    - Value is r
    - Or frequently r^2.

$$r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$$

OR

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

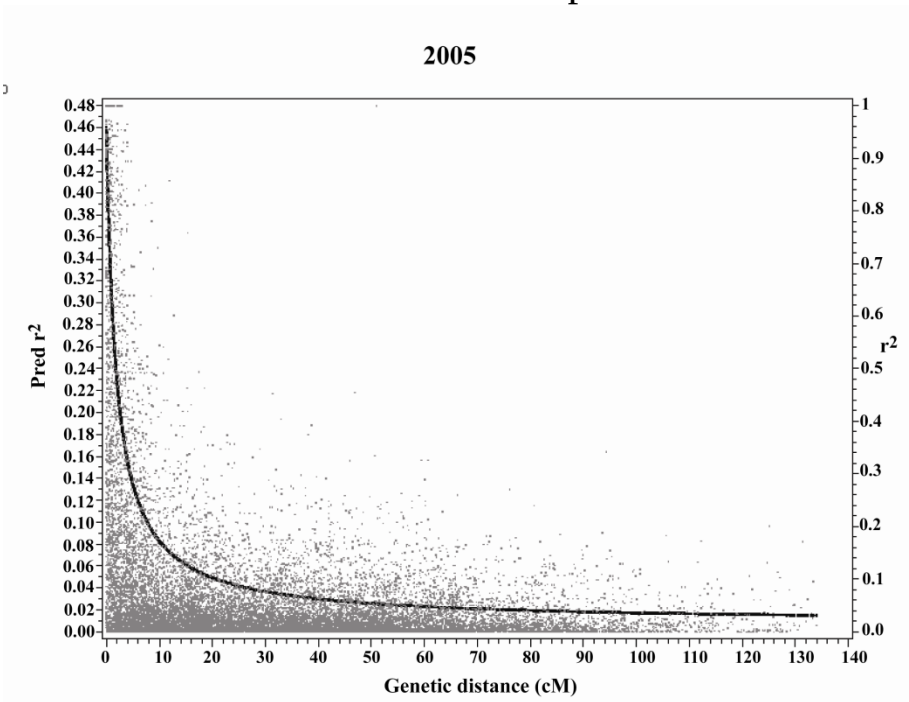
r^2 is useful because it ranges from 0 to 1

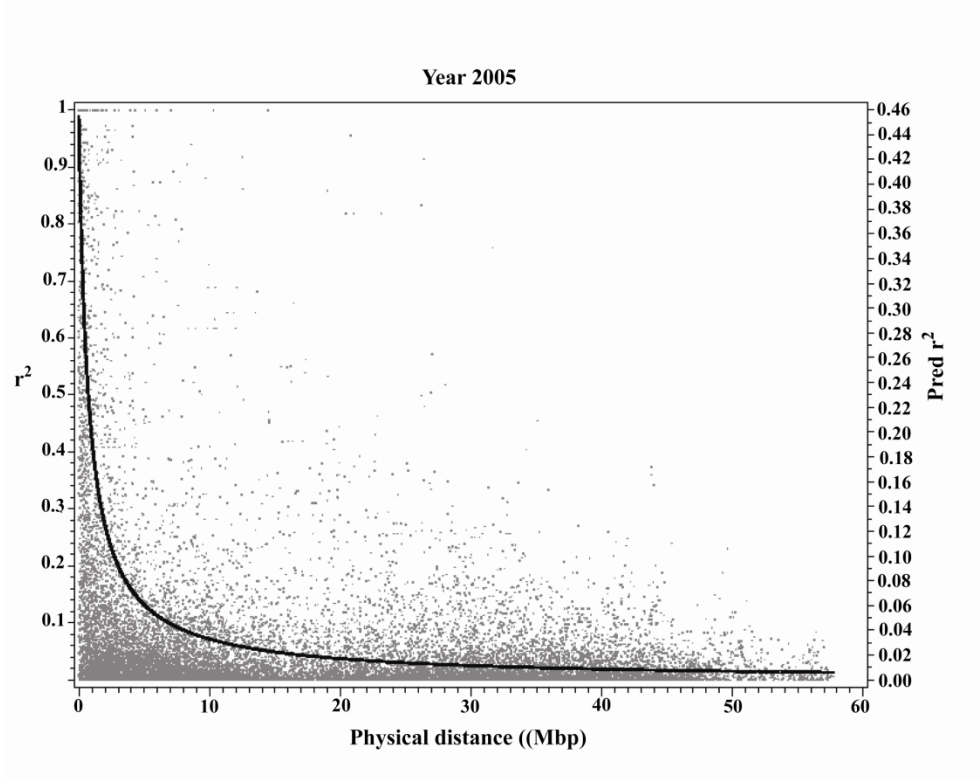
Ranges from

- $r^2 = 0$
    - Loci are in complete linkage equilibrium
- $r^2 = 1$
    - Loci are in complete linkage disequilibrium.

Graphical relationship of linkage disequilibrium

- r^2 to either genetic or physical distance
- r^2 vs. distance is calculated
  - Non-linear regression
    - Two examples





When are loci in linkage equilibrium?

- Examples
  - 0.5, 0.2, 0.1, and 0.05
    - No clear statistical measure
  - Show graph
  - State r^2 value
    - Use as measure of linkage equilibrium.

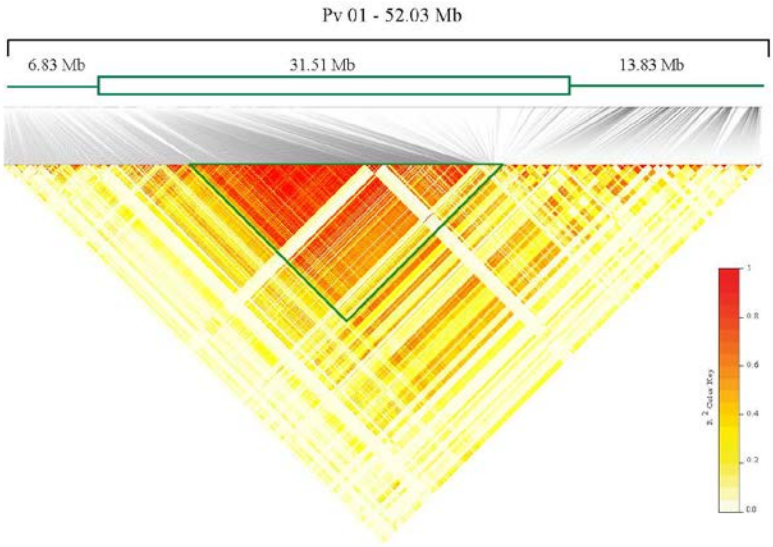
Linkage Disequilibrium Differs Between Chromosomes

Heatmaps

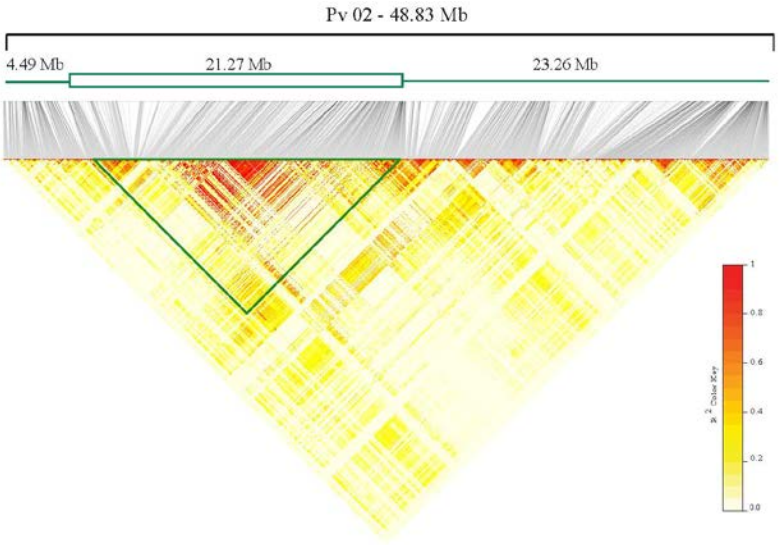
- Better indicator of LD across a chromosome or region
    - Shows the relationship between regions of the genome you are targeting
    - Each value is the pairwise LD between two SNP positions on the chromosome
        - Not an overall average
- The higher the LD (= r^2) the redder the color; all pairwise r^2 values are shown

Linkage Disequilibrium Differs Among Chromosomes

Common Bean chromosome Pv01

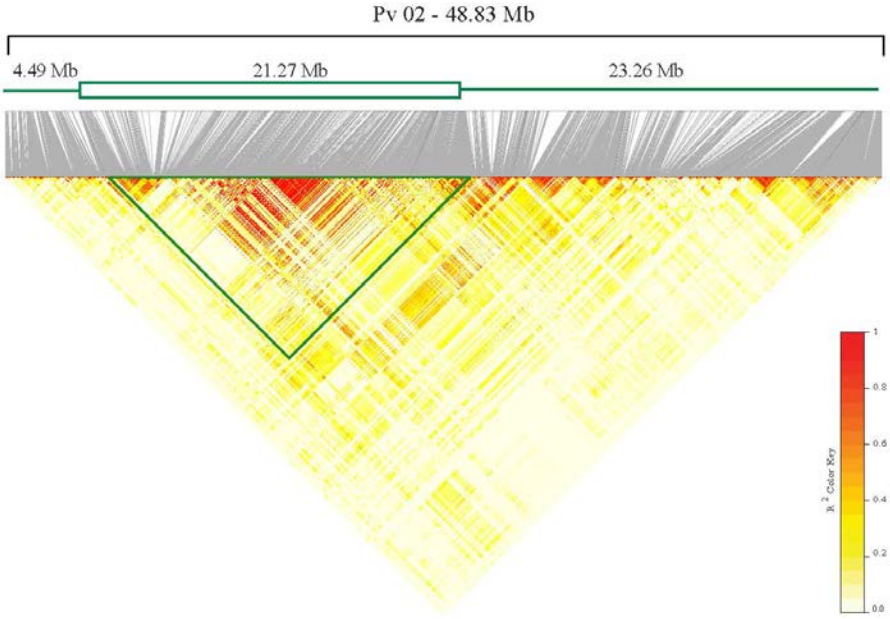


Common Bean chromosome Pv02

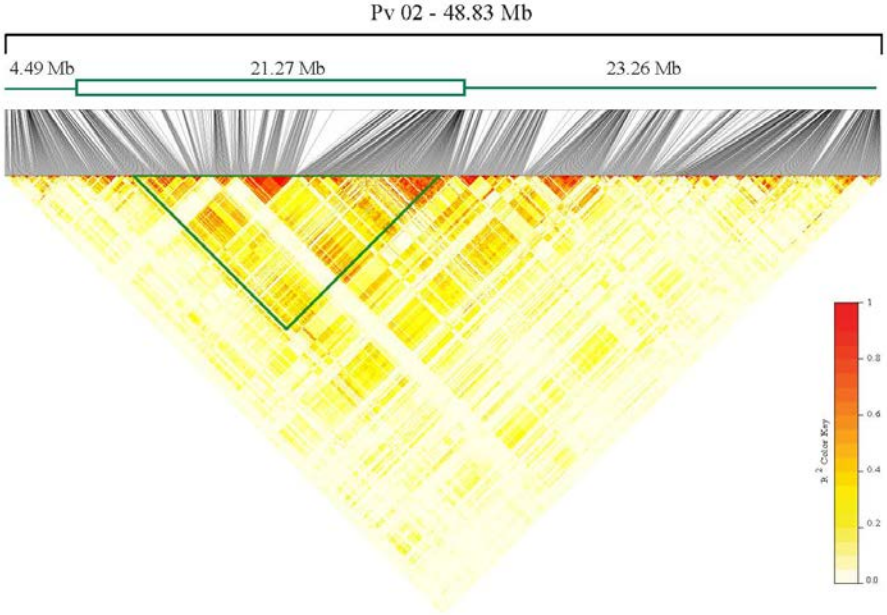


Linkage Disequilibrium Differs Among Populations

Common Bean Race Durango

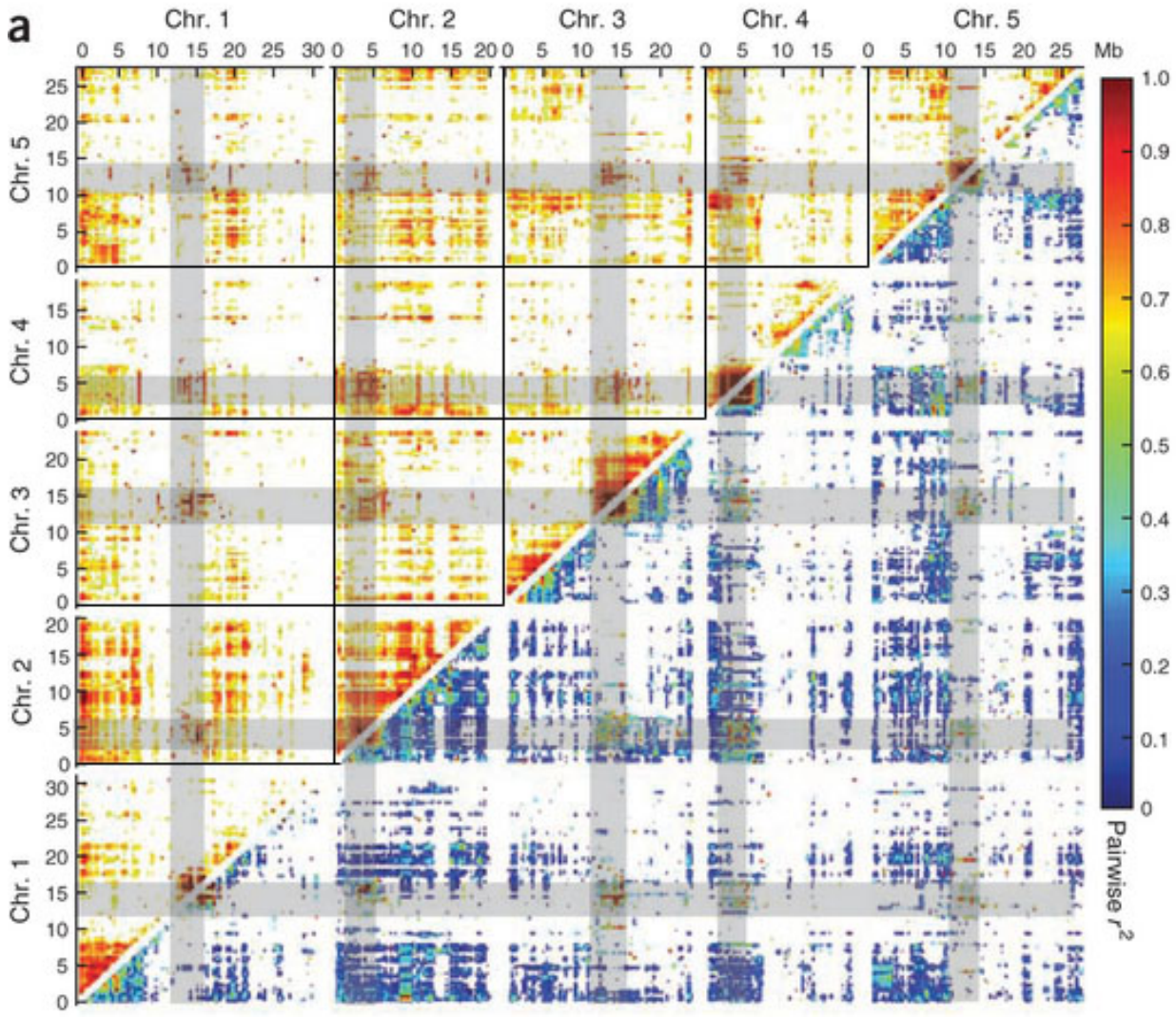


Common Bean Race Mesoamerica



Genome-wide Linkage Disequilibrium

From: Long et al (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nature Genetics 45:884

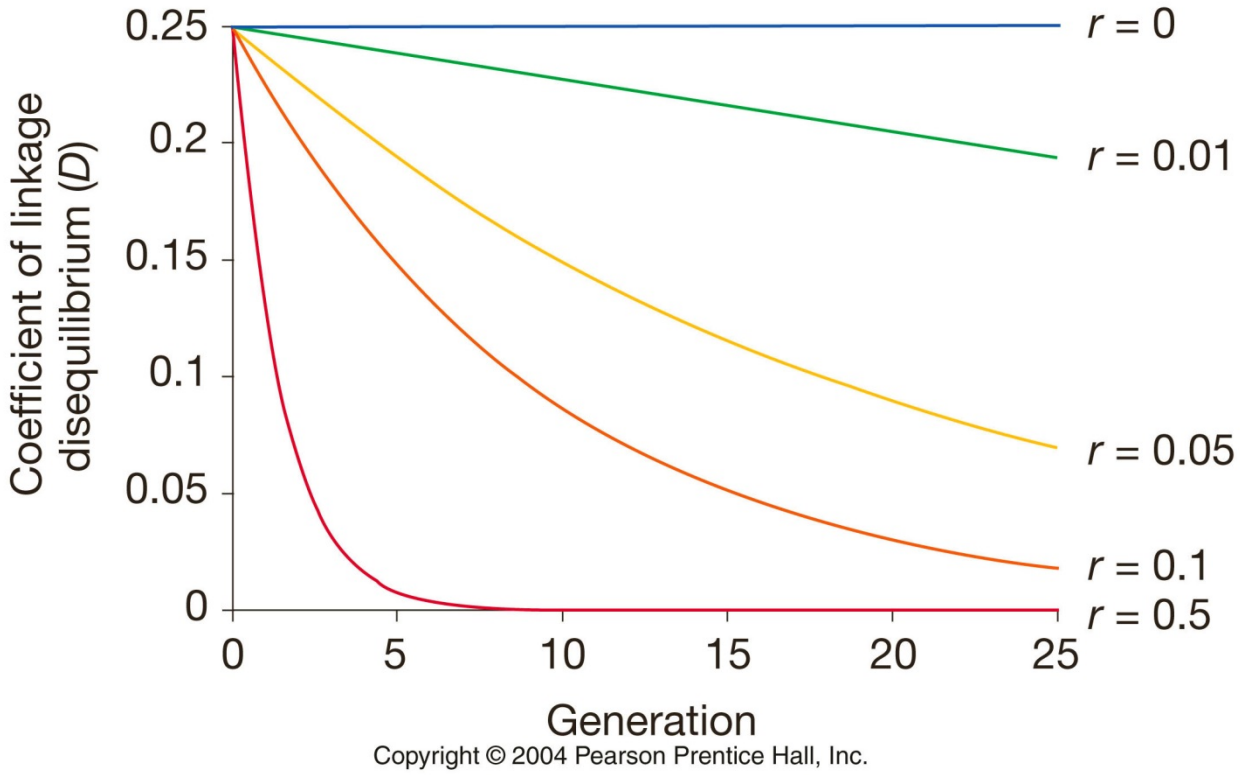


Genome-wide pairwise LD. Values above diagonal: before correcting for population structure; for clarity, only values above 0.6 are shown. Values below diagonal: after applying a transformation to reduce the effects of population structure (related to the correction used in genome-wide association mapping.

What Factors Affect Linkage Disequilibrium?

Recombination

- Changes arrangement of haplotypes
- Creates new haplotypes

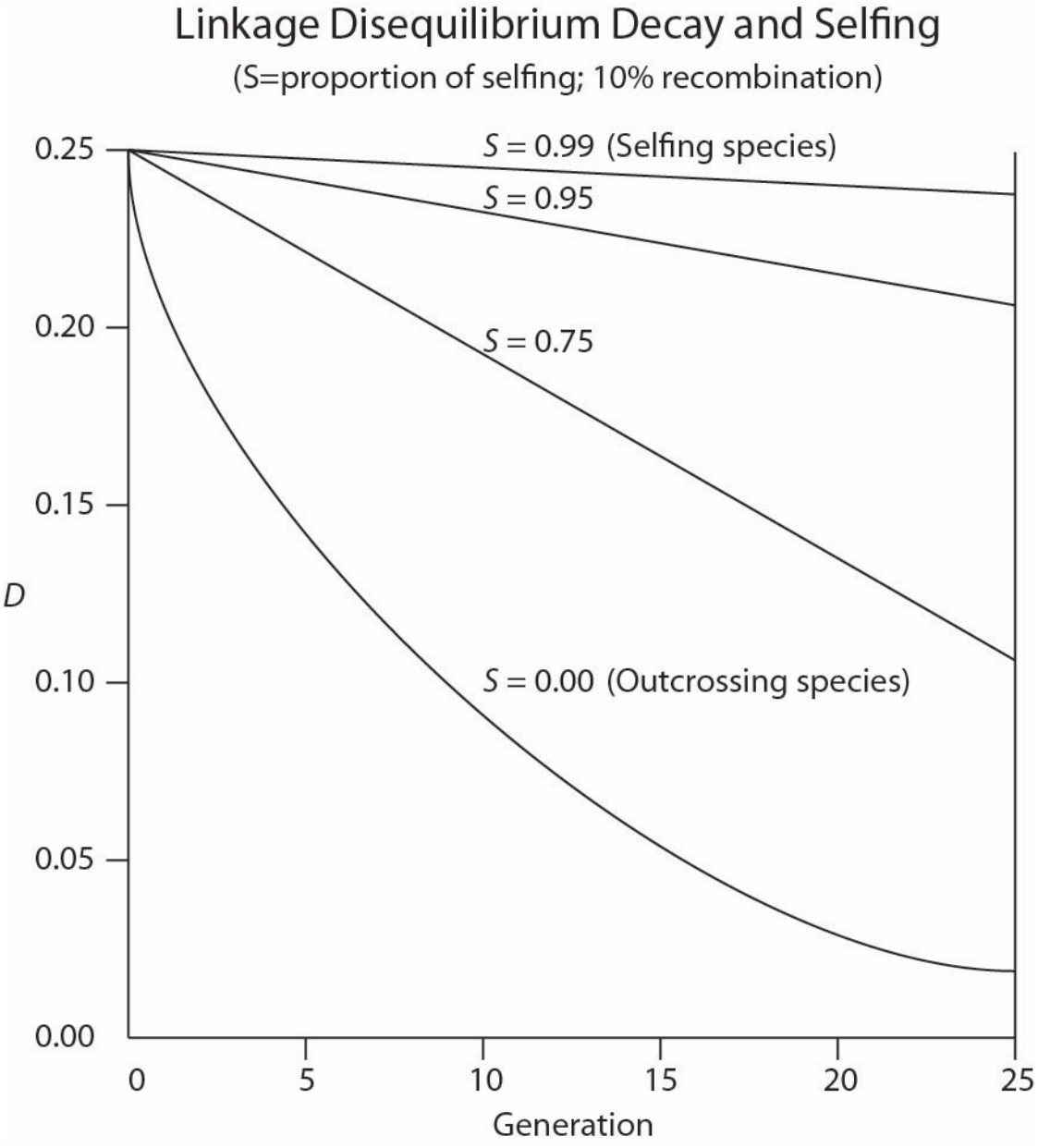


Genetic Drift

- Changes allele frequencies due to small population size
  - Random effect
- LD changes depends on population size and recombination rate
  - Smaller populations
    - New non-random associations appear
      - Larger LD values between some pairs of loci
- Larger populations
  - Less effect on LD

Inbreeding

- The decay of linkage disequilibrium is delayed in selfing populations
- Important for association mapping in self-pollinated crops



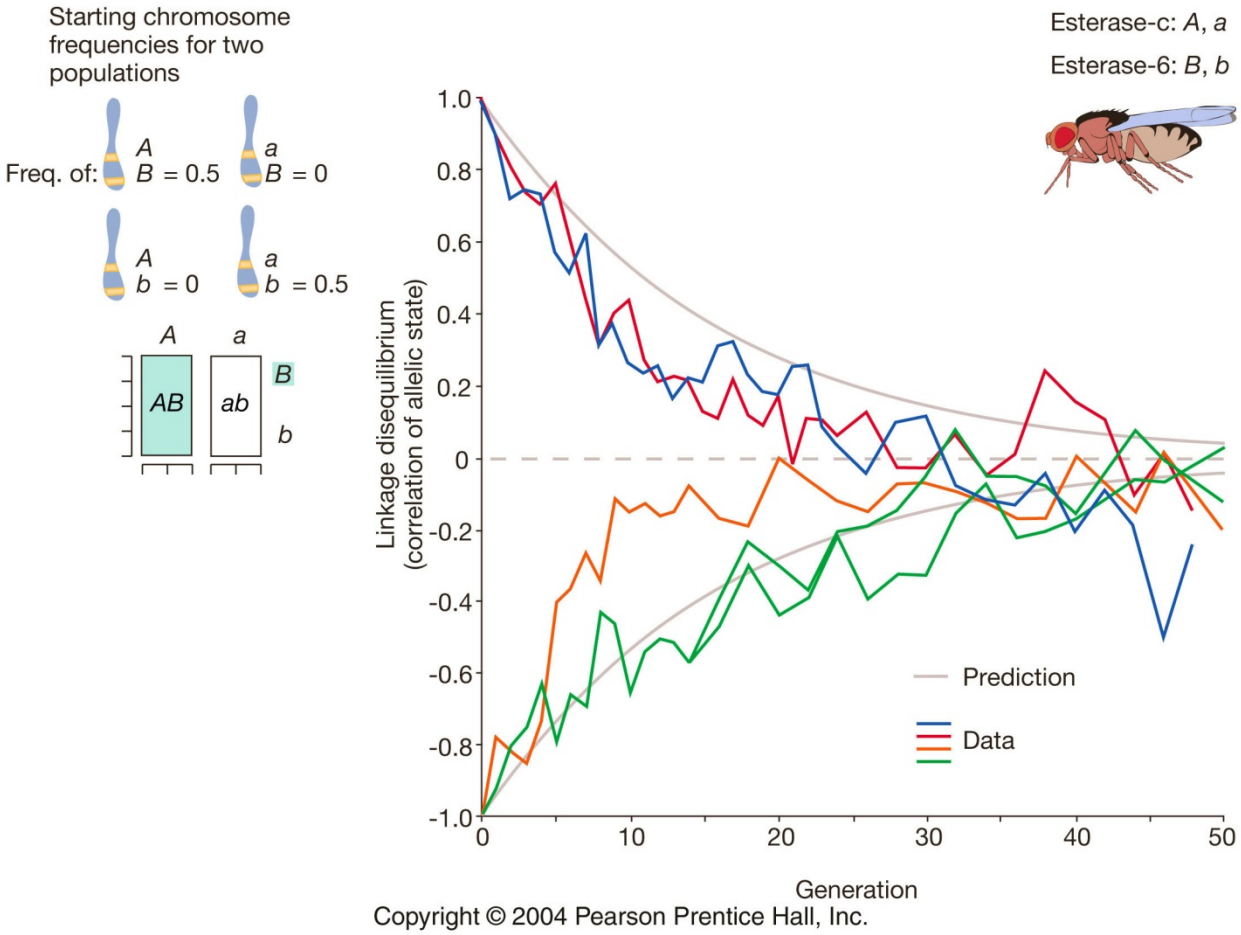
Mutation

- Effect is generally small absent recombination and gene flow

Gene flow

- LD becomes large if two populations intermating are genotypically distinct
- Not much of a problem if crossing between highly similar population found with most breeding programs

Expected and observed decay of LD in an outcrossing species



Association Mapping in Plants

Traditional QTL approach

- Uses standard bi-parental mapping populations
  - F_2 or RI populations
- These have a ***limited number of recombination events***
  - Result is that the QTL covers many cM
- Additional steps required to narrow QTL or clone gene
- Difficult to discover closely linked markers or the causative gene

Association mapping (AM)

- An alternative to traditional QTL mapping
  - ***Uses the recombination events from many lineages***
  - Discovers linked markers associated (=linked) to gene controlling the trait
- Major goal
  - Discover the causative SNP in a gene
- Exploits the natural variation found in a species
  - Landraces
  - Cultivars from multiple programs
    - Discovers associations of broad application
  - Variation from regional breeding programs can also be utilized
    - Associations useful for special local discovered

Problem with AM

- Association could be the result of population structure
  - Hypothetical example

	North America										South America									
Plant Ht	10	10	12	11	13	9	11	10	13	12	4	6	5	7	6	6	4	5	9	5
Dis Res	S	S	S	T	S	S	S	S	T	S	T	S	T	T	T	T	T	S	T	T
SNP1	T	T	T	G	T	T	T	T	G	T	G	G	G	G	G	G	T	G	T	G

SNP1 in Example

- Assumed the SNP it is associated with plant height or disease resistance
  - North American lines are
    - Shorter and susceptible
    - Allele T could be associated with either trait
  - South American lines are
    - Taller and tolerant
    - Allele G could be associated with either trait
- Associated with both traits because of population structure
  - These are false positive associations (***Type I errors***)
- Result
  - ***Population structure must be accounted for in analysis***

Key Principle Regarding AM

- **Human**
  - ***Common variant/common disease***
  - A specific SNP in a specific gene is responsible for a disease found throughout humans
- **Plants**
  - ***Common variant/common phenotype***
  - A specific SNP in a specific gene is responsible for a disease found throughout a specific species

Important Concept Related to Principle

- ***Association Mapping***
  - ***Useful for discovering common variant***
  - Each locus may account for only a small amount of the variation
- ***Bi-parental mapping***
  - ***Useful for discovering rare alleles that control a phenotype***
  - These alleles typically have a major effect

Idealized Cases Results for AM

	Marker 1	
	Allele 1	Allele 2
Case	100	100
Control	100	100

- ***No association between marker and phenotype***

	Marker 2	
	Allele 1	Allele 2
Case	200	0
Control	0	200

- ***Association between marker and phenotype***

Methodology of AM

1. Define a population for analysis

- Should represent the diversity useful for goals of project
    - Specific to target of project
        - Species-wide
            - Use lines from all major subdivisions of the species
        - Regional or local
            - Use lines typical to target region

2. Genotype the population

- Genome-wide scan
    - Medium density
        - ~5,000 – 50,000 SNPs
        - Array-based assays
    - High density
        - 50,000 - 2,500,000 SNPs
        - Whole genome or reduced representation resequencing
            - Rice: 3,000,000 SNPs
            - Corn: 25,000 – 2,500,000 SNPs
        - Array-based assays
            - Arabidopsis
                - 250,000 SNPs
                - Affymetrix chip

- Candidate gene (original approach no longer used)
    - Select genes that might control trait
        - Sequence different genotypes
        - Discover SNPs in gene
            - 5’-UTR or 3’-UTR
            - Coding region
            - 3’-UTR

3. Controlling for Population Structure/Relatedness

- Define subpopulations
  - Select markers to genotype the population
    - Markers should ideally be
      - Distributed among all chromosomes
      - All should be in linkage equilibrium (r^2<0.2)
      - Minor allele frequency >0.1
  - Evaluating population structure
    - STRUCTUE software
      - Use matrix of percentage population membership in analysis
      - Fixed effect
      - **Original approach, BUT**
        - **Discontinued**
          - Assumption of Hardy-Weinberg Equilibrium with software always violated
    - Principal component (PC)
      - Defines groups of individuals
      - Select number of principal components that account for specific percent of variation
      - 25% - 50% are a typical value
      - Fixed effect
  - Evaluate relatedness
    - EMMA or Spagedi relatedness calculations
      - Output is a table with all pairwise-comparisons
      - Random effect

4. Statistical Analysis

- Marker-by-marker analysis
  - Regression of phenotype onto marker genotype
    - Significant marker/trait associations discovered
- Analysis must control for population structure and/or relatedness
  - Most popular approach
    - Mixed linear model
      - Example formula:

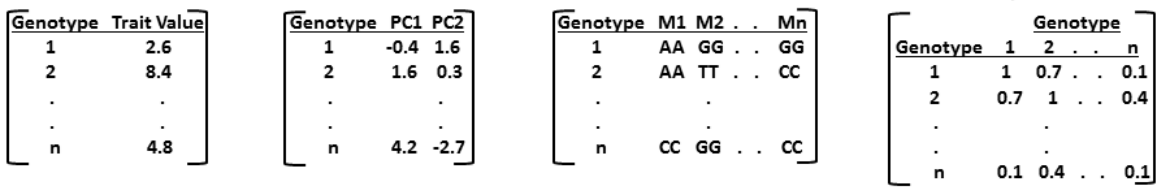
$$y = Pv + S\alpha + I\mu + e$$

y = vector of phenotypic values
P = matrix of structure or PC values
v = vector regarding population structure (STRUCTURE of PC values) (fixed effect)
S = vector of genotype values for each marker
α = vector of fixed effects for each marker (fixed effect)
I = relatedness identity matrix
μ = vector pertaining to recent ancestry (random effects)
e = vector of residual effects

Model from: Weber et al. 2008. Genetics 180:1221.

GWAS Mixed-linear Model, and Program Input Matrices

$$y = Pv + S\alpha + I\mu + \varepsilon$$



Genotype	Trait Value
1	2.6
2	8.4
.	.
.	.
n	4.8

Phenotype Matrix

Genotype	PC1	PC2
1	-0.4	1.6
2	1.6	0.3
.		.
.		.
n	4.2	-2.7

Structure Matrix (Fixed effect)

Genotype	M1	M2	.	.	Mn
1	AA	GG	.	.	GG
2	AA	TT	.	.	CC
.		.			
.		.			
n	CC	GG	.	.	CC

Genotype (SNP) Matrix (Fixed effect)

	Genotype			
Genotype	1	2	. .	n
1	1	0.7	. .	0.1
2	0.7	1	. .	0.4
.		.		
.		.		
n	0.1	0.4	. .	0.1

Relatedness Matrix (Random effect)

GWAS Goal

- Determine significance of the each market based on phenotype, structure and relatedness
- Based on F-test for significance

Software Inputs

- Each of the matrices

Software

- GAPIT, TASSEL, R, SAS

5. Software of choice

All give equivalent p-value results for the marker-trait associations.

- ***The most important criteria for the software!!***

Tassel

- Java-based software
- Extensively used, often cited
- Early application available to users performing plant GWAS
- Some pre-analysis steps required before use with earlier versions
- Fairly extensive post-analysis software data manipulation required to develop tables and figures for analysis and publication

GAPIT

- R-module
- Performs PCA and relatedness analysis for you
- Fairly usable figures generated by the software that can be used for publication

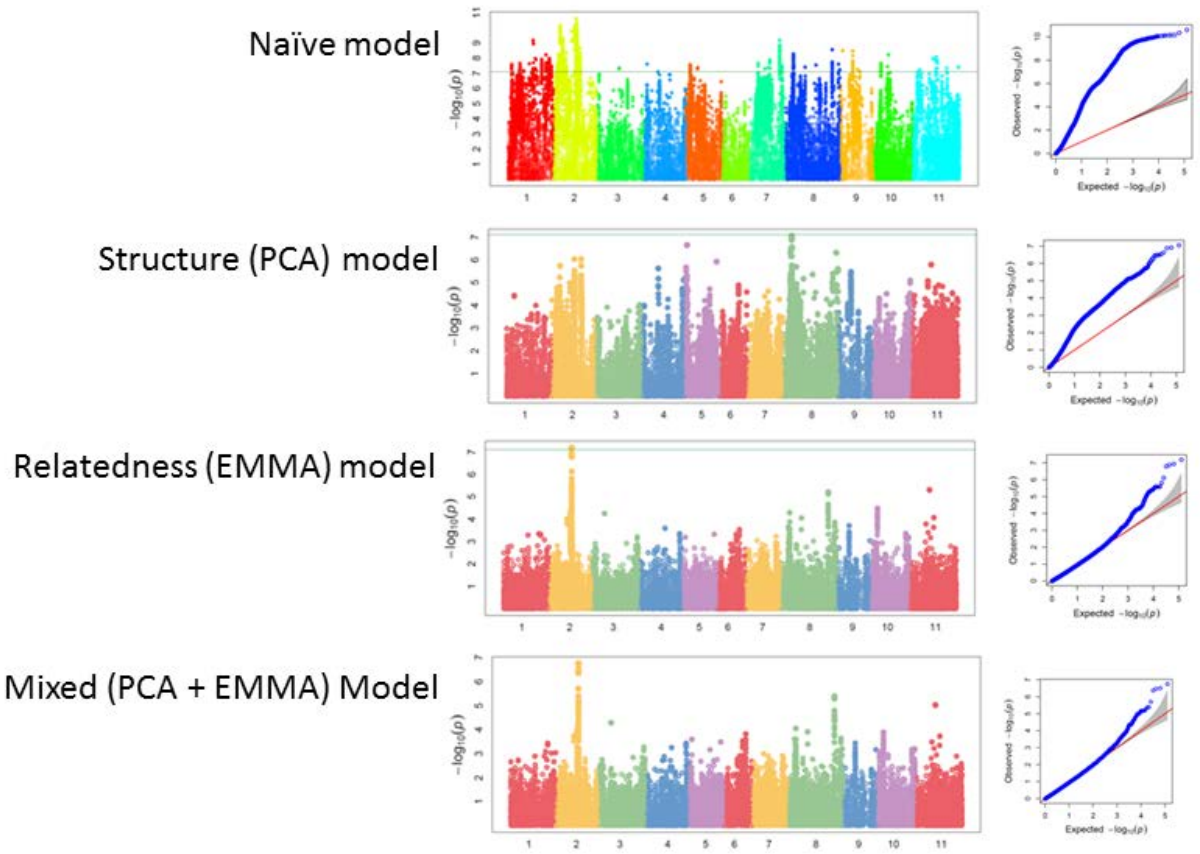
MLM or GLM analysis with R or SAS

Just provides the statistical output for the marker-trait associations

6. Choosing the correct model

- Evaluate all models individually
  - Naïve
  - Relatedness
  - PC (for population structure)
  - PC and relatedness
- Develop a Q-Q plot for each model
  - Y-axis
    - Observed $-\log_{10}$ (P) values
  - X-axis
    - Expected $-\log_{10}$ (P) values values
  - Best model
    - Observed ~ equal expected $-\log_{10}$ (P) values
    - Select the model that is linear or nearly so

Example: Common Bean Fat Content



Mean square deviation

This is a statistical method to determine best model. First rank all marker p-values from smallest to largest. Then apply the following formula:

$$\mathbf{MSD} = \left\{\sum_{i=1}^{n}\left[p_i - \left(\frac{i}{n}\right)^2\right]\right\}/n$$

where *i* is the rank number of the a specific p-value, p_i is the probability of the *i*th ranked *p*-value, and *n* is the number of markers. The model with the lowest MSD value is selected as the best model.

MSD data for the fat data is in the table below.

Model	**MSD value**
Mixed model (3PCs-EMMA)	0.000299
Relatedness (EMMA) model	0.000475
Structure (3PCs) model	0.024883
Naïve model	0.076743

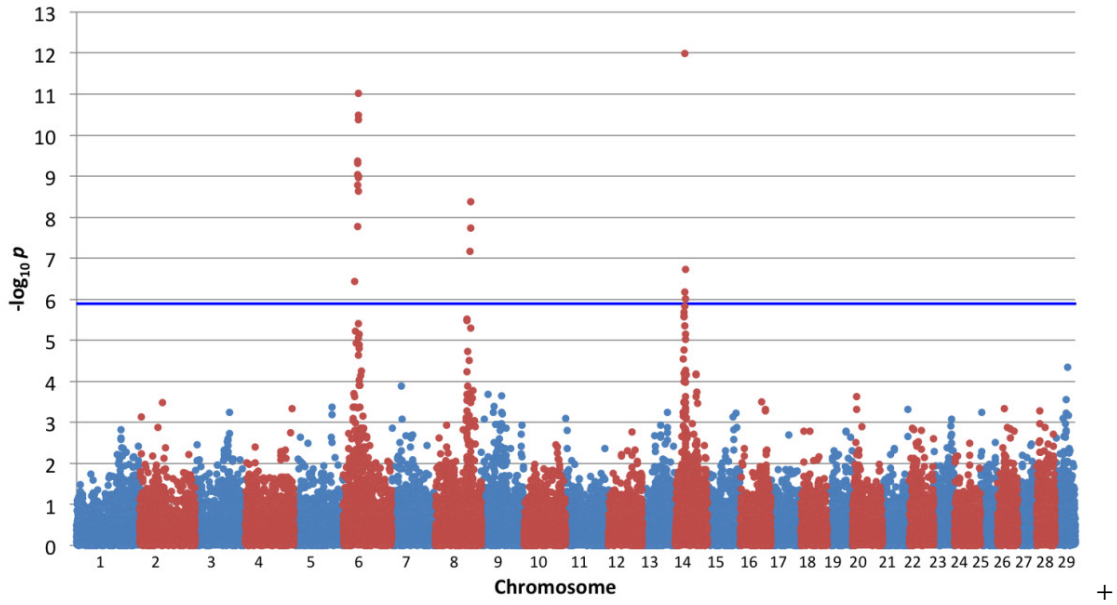
The **MIXED** model has the best fit of the four models!!

7. What is a Significant Association?

- When performing multiple analyses on the same phenotype dataset
  - At a $P = 0.05$ level
    - 1 of 20 random associations will be significant
      - **Must account for this Type I error**
- **Bonferroni test**
  - Divide experiment-wide error rate by number of comparisons
    - Error rate of 0.05 and 100 comparisons
      - $P < 0.0005$ would be significant
  - Conservative approach with a single value for all phenotypes using the same marker set
- **Permutation test**
- Develop a cut-off using 1,000 permutations of the data
- Select markers at a specific cutoff level: 0.1% or 0.01% of markers
  - Sensitive to the different number of genetic factors associated with each trait.

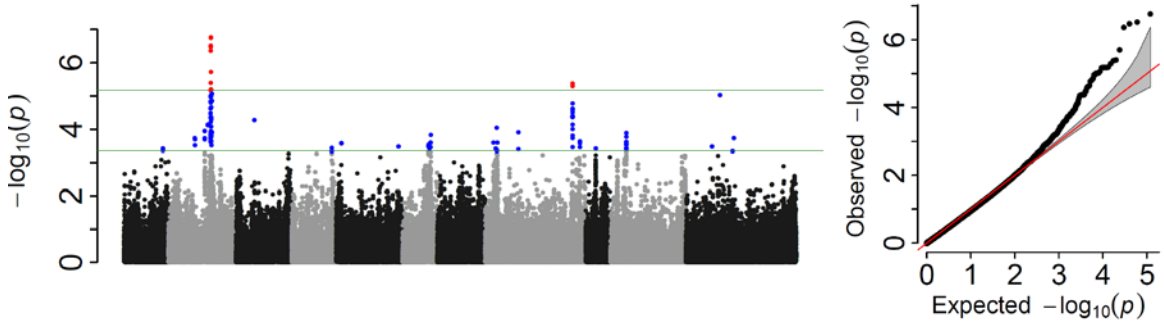
Manhattan (New York City) Skyline from New Jersey

Manhattan Plot
Black bear GWAS analysis



8. Post-GWAS clean-up

- Data only presented for markers with minor allele frequency of 5% or greater
- Permutation test cutoffs presented
- Q-Q plot shown to demonstrate power of the model



Does Association Mapping Work??

Example: Aranzana, M. J., *et al*. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS genetics*, *1*(5), e60.

- Population
    - 95 Arabidopsis accessions from Europe
- Phenotyping
    - Flowering time
    - Disease response to three pathogens
- Genotyping
    - 876 random loci
    - 4 candidate genes
        - Flowering time
            - *FRI* (Chromosome 4)
        - Disease Resistance
            - *Rpm1* (Chromosome 3)
            - *Rps2* (Chromosome 4)
            - *Rps5* (Chromosome 1)
- Statistical analysis
    - Population structure only correction

Results

- All four candidate loci strongly associated with expected phenotype

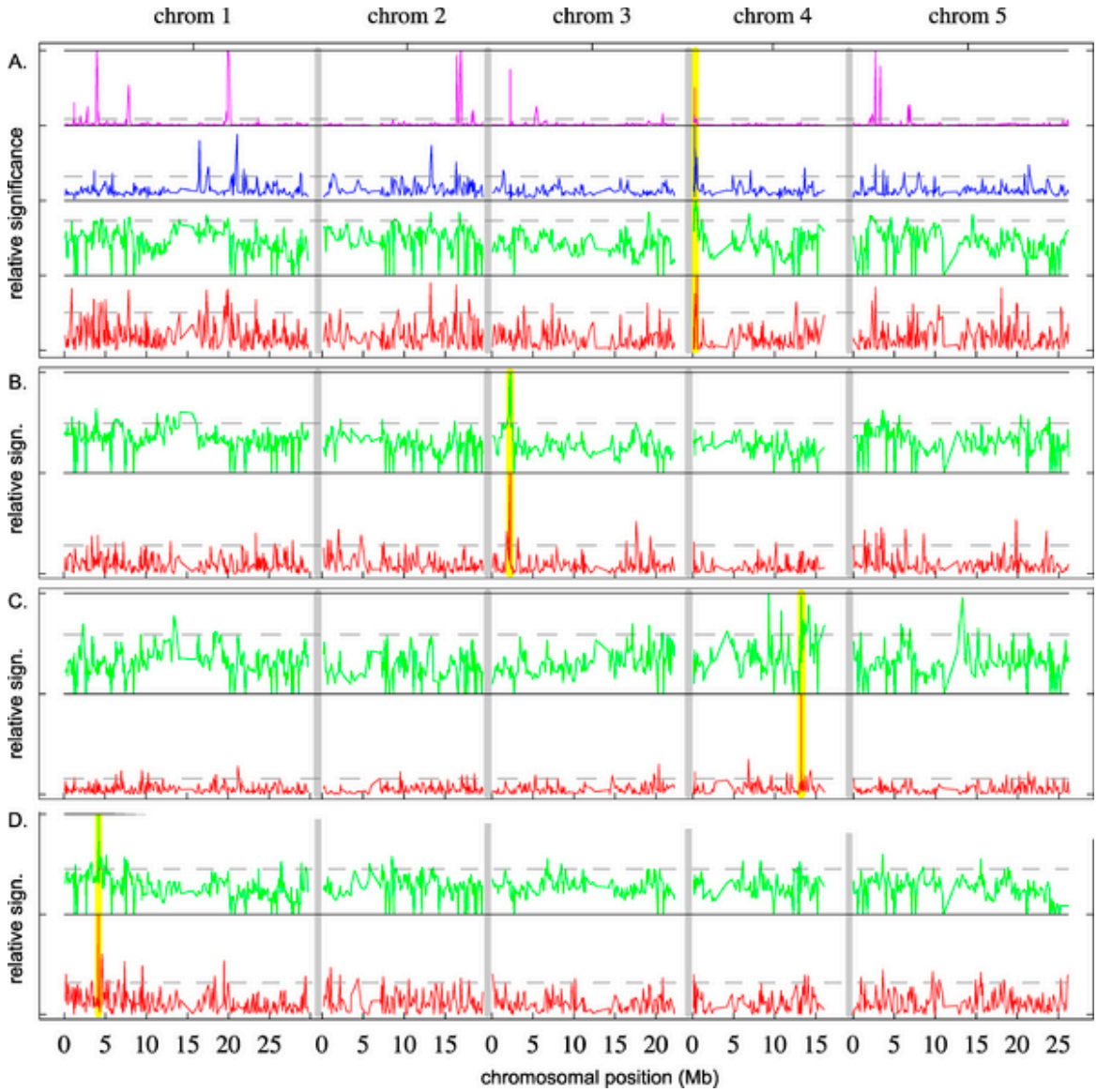


Figure 3. Genome-Wide Scans for Association with Flowering Time and Pathogen Resistance

For flowering time **(A),** four different statistical methods were used (described in Materials and Methods): Voronoi focusing on "late" alleles (magenta line), Voronoi focusing on "early" alleles (blue line), CLASS (green line), and fragment-based Kruskal–Wallis tests (red line; see also Figure 2). For pathogen resistance (avrRpm1 **[B]**, avrRpt2 **[C]**, and avrPph3 **[D]**), only the last two tests were used. Higher peaks indicate stronger association (the *y*-axes are proportional to the negative log *p*-values, but have been normalized to the highest value within each test). The dotted lines correspond to the 95% percentile and are mainly intended to facilitate comparison between figures. Yellow vertical lines indicate the positions of the appropriate candidate loci. Peaks occur at these loci for all methods, but are otherwise distributed throughout the genome.

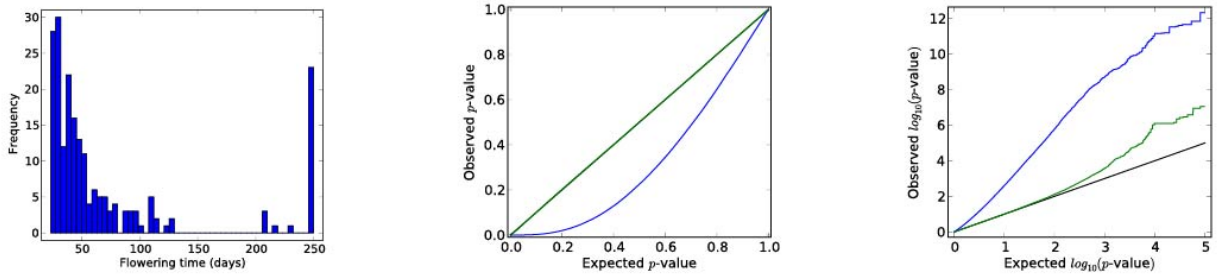
Arabidopsis: Analysis of 107 phenotypes

Nature: (2010) 465:627

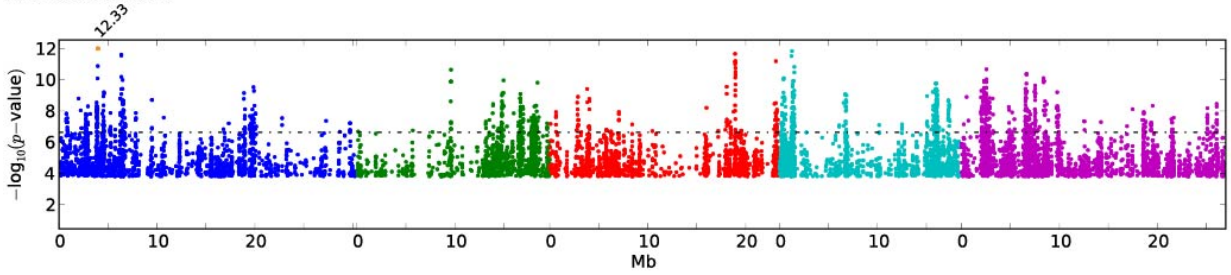
Notes

- EMMA reduced the number of false positives
- *FLC* and *FRI* confirmed as candidate genes for days to flowering

Phenotype histogram and quantile-quantile plots of p-values

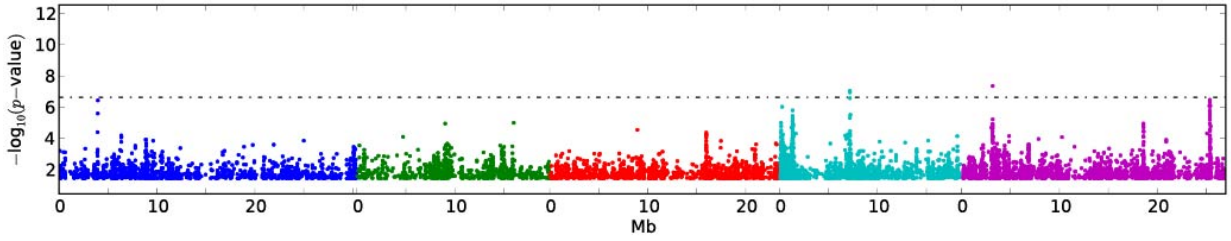


Wilcoxon results



Rank	Score	Gene	Gene ID	Chr	SNP pos (bp)	Distance to gene (bp)
4	11.6112	ATARP4	AT1G18450	1	6369609	17641
16	10.824	DFL2	AT4G03400	4	1516895	-17031
19	10.6444	AGL17	AT2G22630	2	9611587	-13865
43	9.88639	SVP	AT2G22540	2	9606045	15072
52	9.74982	ATH1	AT4G32980	4	15930436	-12389
97	9.14786	sim to VRN1	AT4G33280	4	16040939	6419
103	9.14404	ATGA2OX7	AT1G50960	1	18903090	7703
138	8.93617	RAV1	AT1G13260	1	4541173	-992
139	8.90838	ETC3	AT4G01060	4	454542	-5930
153	8.79855	FLC	AT5G10140	5	3188328	-8879

EMMA results



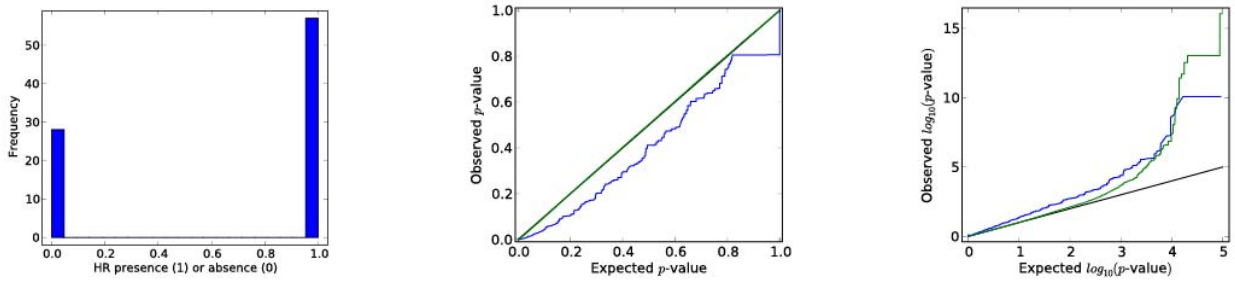
Rank	Score	Gene	Gene ID	Chr	SNP pos (bp)	Distance to gene (bp)
1	7.35652	FLC	AT5G10140	5	3188328	-8879
21	6.02586	sim to ESD4	AT4G00690	4	268809	-12836
21	6.02586	FRI	AT4G00650	4	268809	-217
39	4.95198	DOG1[8]	AT5G45830	5	18590971	15738
80	4.31728	CDF1	AT5G62430	5	25084106	2213
98	4.18876	ATARP4	AT1G18450	1	6369765	17797
180	3.62105	CRP	AT4G00450	4	206784	0
188	3.58201	SPA4	AT1G53090	1	19790829	259
188	3.58201	SPL4	AT1G53160	1	19790829	-19258
199	3.54707	RGA1	AT2G01570	2	260329	-2780

Supplementary Figure 24 – Summary of GWA results for Days to flowering at 22°C (FT22)

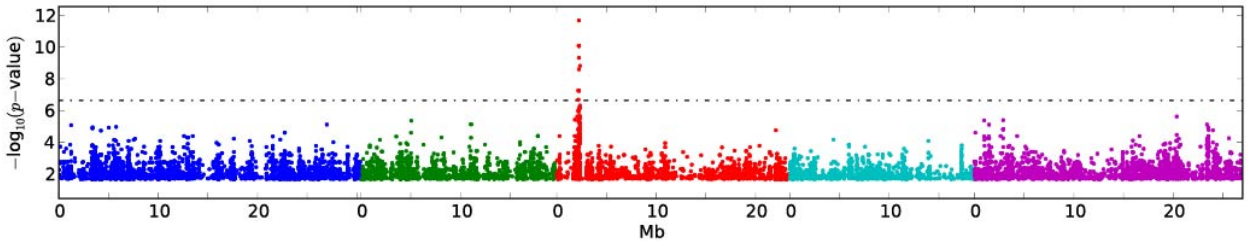
Notes

- Monogenic gene identified
- *RPM1* confirmed as gene controlling resistance to *Pseduomonas syringae*

Phenotype histogram and quantile-quantile plots of p-values

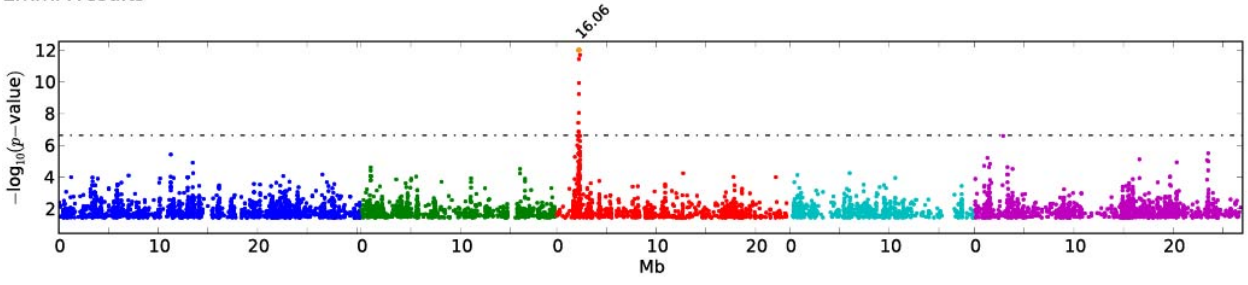


Fisher's exact test results



Rank	Score	Gene	Gene ID	Chr	SNP pos (bp)	Distance to gene (bp)
1	11.6797	RPM1	AT3G07040	3	2227823	0
38	5.37362	CTR1 [8]	AT5G03730	5	980341	493
57	4.9404	AT5G58120	AT5G58120	5	23528056	-6662
104	4.36207	AT1G58170	AT1G58170	1	21544828	4327
164	3.94676	AT3G28890	AT3G28890	3	10907617	-7539
170	3.89614	AT5G47250	AT5G47250	5	19203727	0
183	3.83911	AT4G09360	AT4G09360	4	5927294	-12889
183	3.83911	ATMKK3	AT5G40440	5	16216269	14414

EMMA results



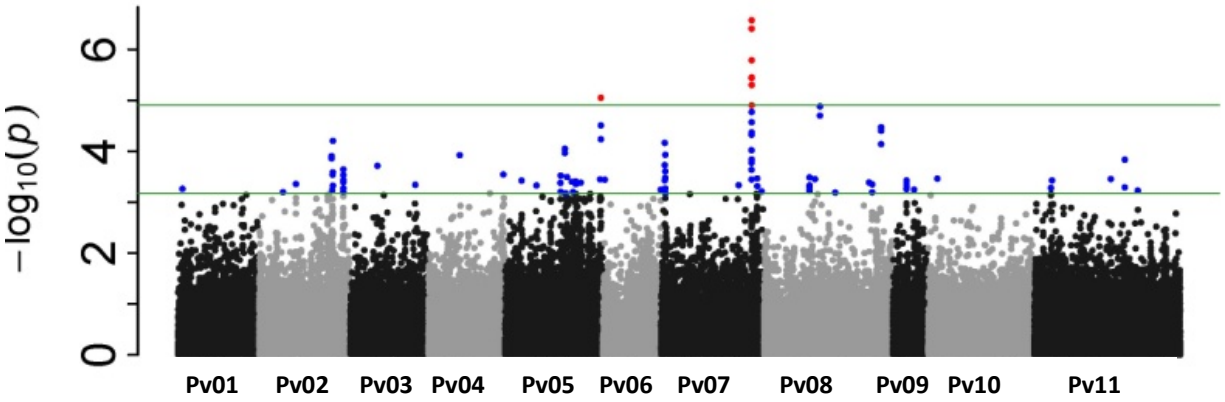
Rank	Score	Gene	Gene ID	Chr	SNP pos (bp)	Distance to gene (bp)
1	16.0583	RPM1	AT3G07040	3	2227823	0
67	4.71172	CTR1 [8]	AT5G03730	5	980341	493
36	5.49555	AT5G58120	AT5G58120	5	23528056	-6662
40	5.42319	AT1G31540	AT1G31540	1	11273835	14598
71	4.62411	RBK1	AT5G10520	5	3328044	-5133
75	4.60962	AT2G03200	AT2G03200	2	976276	-8361
76	4.58902	ATMKK3	AT5G40440	5	16216269	14414
103	4.06069	AT1G61300	AT1G61300	1	22595752	15379
124	3.94977	AT4G19470	AT4G19470	4	10610171	2833

Supplementary Figure 36 – Summary of GWA results for AvrRpm1

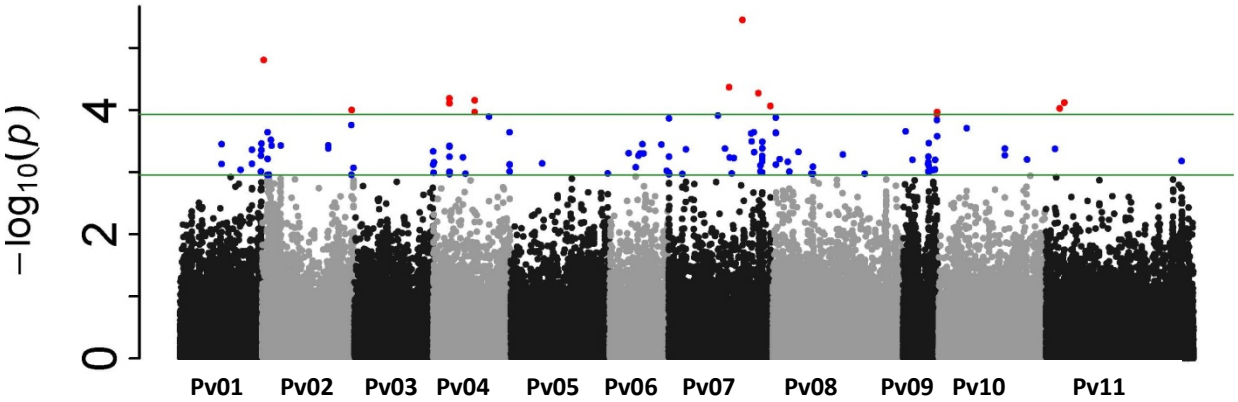
MINERAL MANHATTAN PLOTS ACROSS ALL LOCATIONS

Species: Common bean (*Phaseolus vulgaris*)
SNPs: n=~150k
Population: Modern Middle American diversity panel (n=287)
Model: EMMA (relatedness) or EMMA (relatedness) + PC (structure)

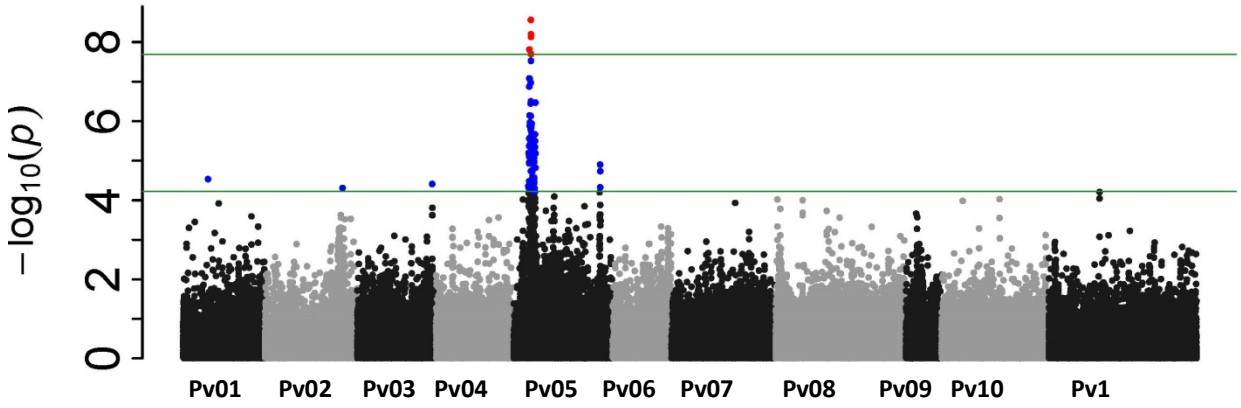
B (BORON): Strong single peak



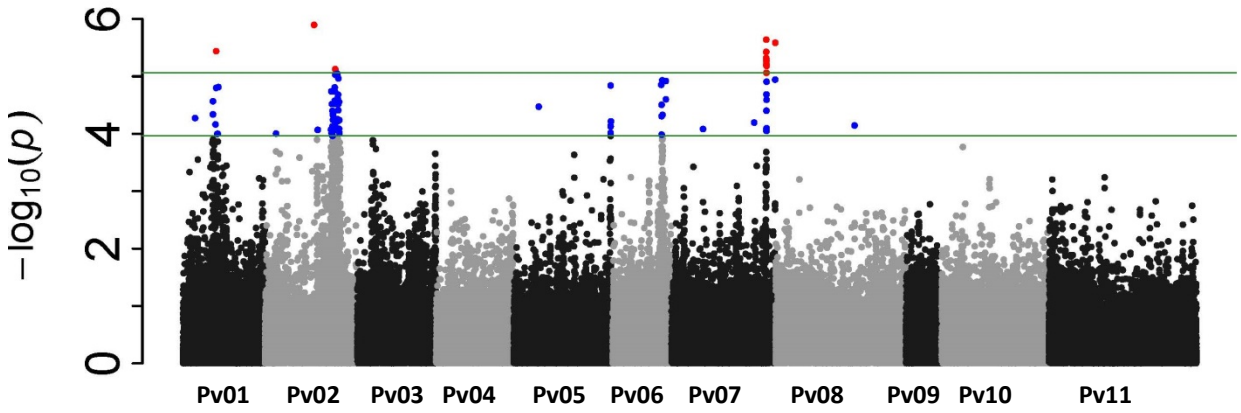
Ca (CALCIUM): diffuse peaks; no clear significant region



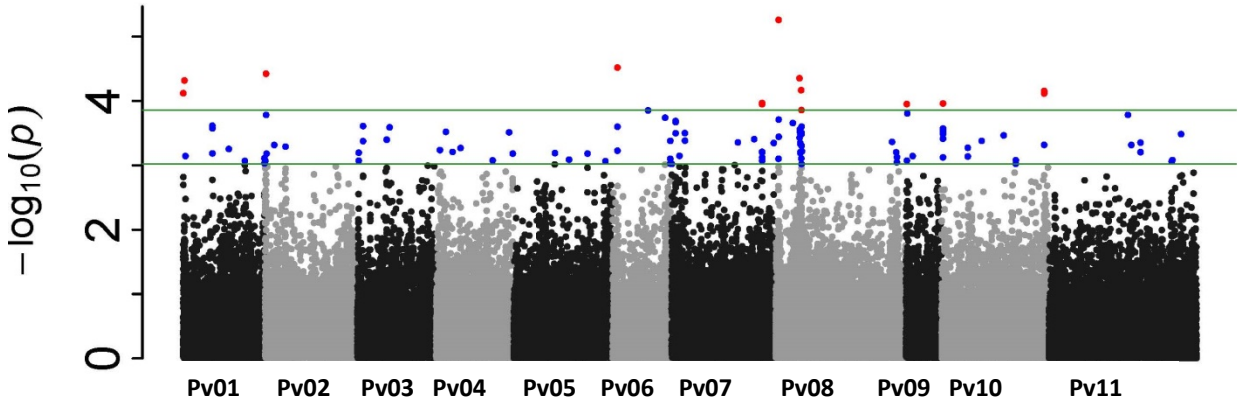
Co (COBALT): one major region



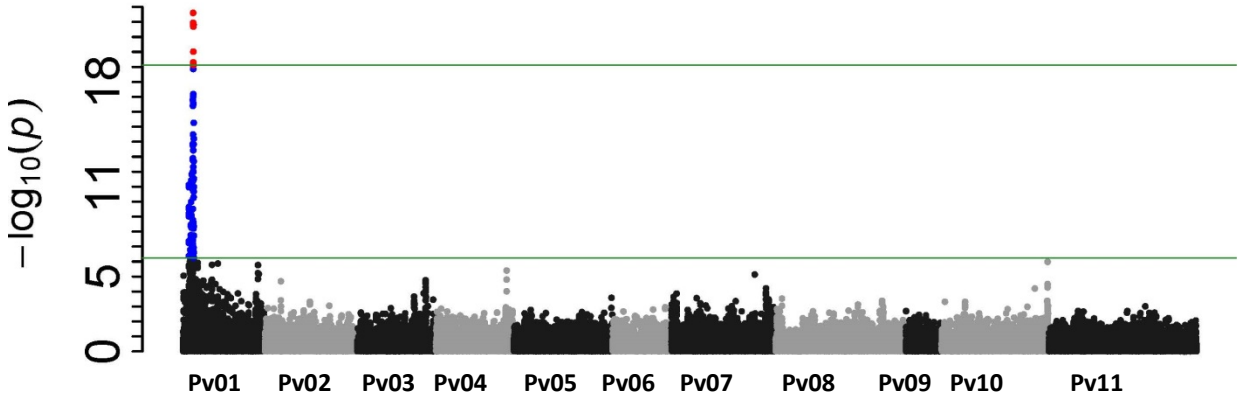
Cu (Copper): two major regions



Mg (Magnesium): multiple, low value peaks; no strong signal



Mo (Molybdenum): one very strong significant peak; in candidate gene



Association Mapping (AM) or Bi-Parental QTL Mapping?

1. Issues to consider

- Effect of rare alleles
    - Effect on rare allele in the association population mean will be minimal
    - Locus will not be detected by the AM approach
        - The effect of a rare allele can be detected in a biparental population
- Effect of common alleles
    - Common alleles are a component of phenotypic expression
        - Effect found throughout the population (species) and can be discovered using AM
        - Contribution of any one allele to phenotype may be small (R^2<10%)

2. What is your goal?

- Discover, analyze, and test genes of major effect
    - ***Bi-parental populations*** of divergent parents and traditional (CIM) is best approach
- Dissect the factors controlling a phenotype throughout a population
    - ***Association mapping of appropriate population*** is a powerful approach