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 centimorgan (cM) distances
- Short distance LB
  - Mapping at the base pair (gene) distance

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

What statistical variable allows us

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

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
- The difference between these two values is considered the deviation or D

Table 3. Relationships among haplotype and allelic frequencies relative to the deviation

	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	

Standard measure of LD is typically calculated as

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

OR

$$D = (x_{11})(x_{22}) - (x_{12})(x_{21})$$

- 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)
  - ***D*** and allelic frequencies (Table 2).

***D* depends on allele frequencies**
Value can range from -0.25 to 0.25

- 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.

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

Ranges from

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

Example:

SNP locus A: A1=T, A2=C
SNP locus B: A1=1, A2=G

Observed haplotype data

Haplotype	**Symbol**	**Frequency**
A1B1	x_{11}	0.6
A1B2	x_{12}	0.1
A2B1	x_{21}	0.2
A2B2	x_{22}	0.1

Calculated allelic frequency

Allele	**Symbol**	**Frequency**
A1	*p1*	0.7
A2	*p1*	0.3
B1	*q1*	0.8
B2	*q2*	0.2

$\boldsymbol{D = x_{11} - p_1 q_1;}$ $\boldsymbol{D = 0.6 - (0.7)(0.8) = 0.6 - 0.56 = 0.04}$

$\boldsymbol{D = (x_{11})(x_{22}) - (x_{12})(x_{21})}$ $\boldsymbol{D = (0.6)(0.1) - (0.1)(0.2) = 0.04}$

Calculating D'

Since $D>0$, *use* D_{max}

D_{max} *is* the smaller of $p_1 q_2$ and $p_2 q_1$

$p_1 q_2 = 0.14$
$p_2 q_1 = 0.24$

$$D' = \frac{D}{D_{\max}}; \qquad D' = \frac{0.04}{0.14} = 0.286$$

Calculating r^2

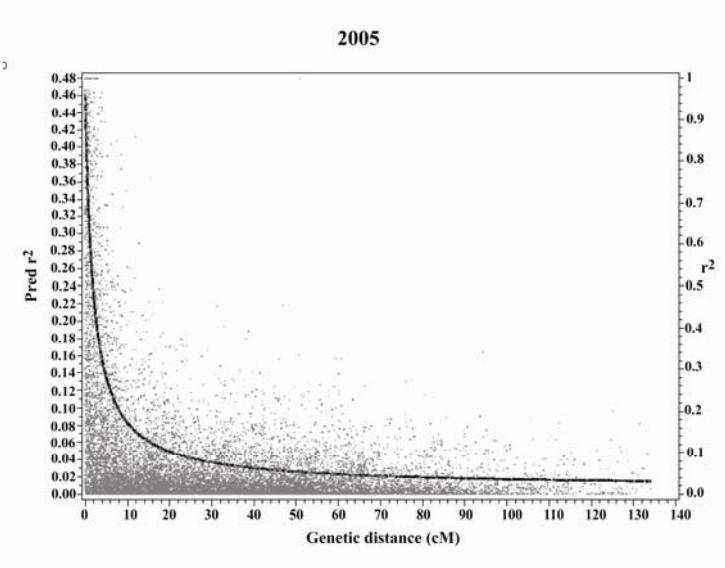
$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

$$r^2 = \frac{0.04^2}{(0.7)(0.3)(0.8)(0.2)} = 0.048$$

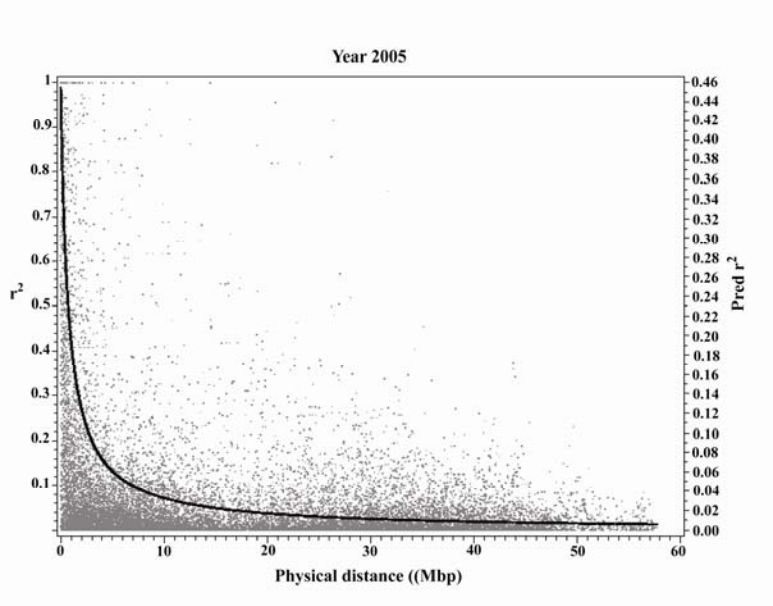
Graphical relationship of

- All two loci pairwise comparisons
  - r^2 relative to either genetic or physical distance
  - r^2 vs. distance is calculated
    - Non-linear regression
      - Two examples

Genetic distance example



Physical distance example



What do the graphs tell us?

- On average, how fast LD decays across the genome
- Useful to determine the number of markers needed for an association mapping experiment

When are loci in linkage equilibrium?

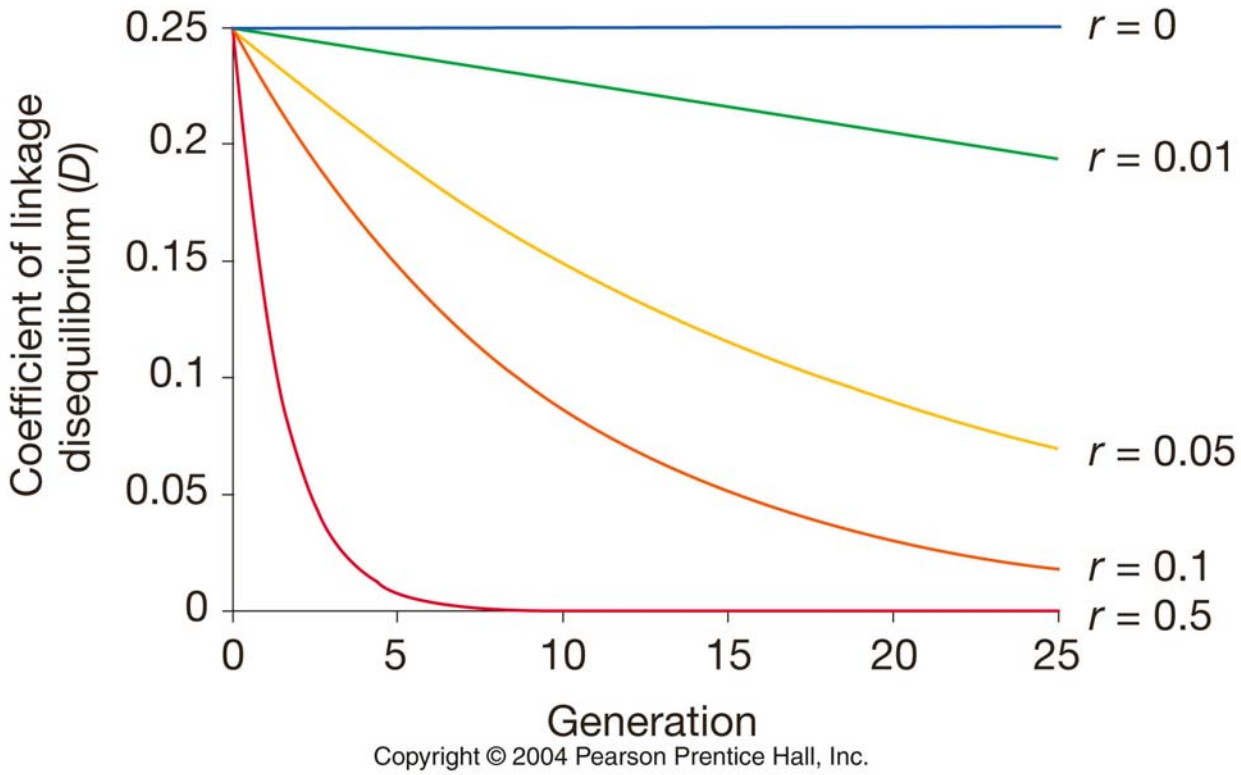
- There is no clear statistical test that states when two loci are in LD

- Examples
  - Papers have used
    - 0.5, 0.2, 0.1, and 0.05
      - Authors choice
  - Typically authors
    - Show graph
    - State r^2 value cutoff for LD

What Factors Affect Linkage Disequilibrium?

Recombination

- Changes arrangement of haplotypes
- Creates new haplotypes



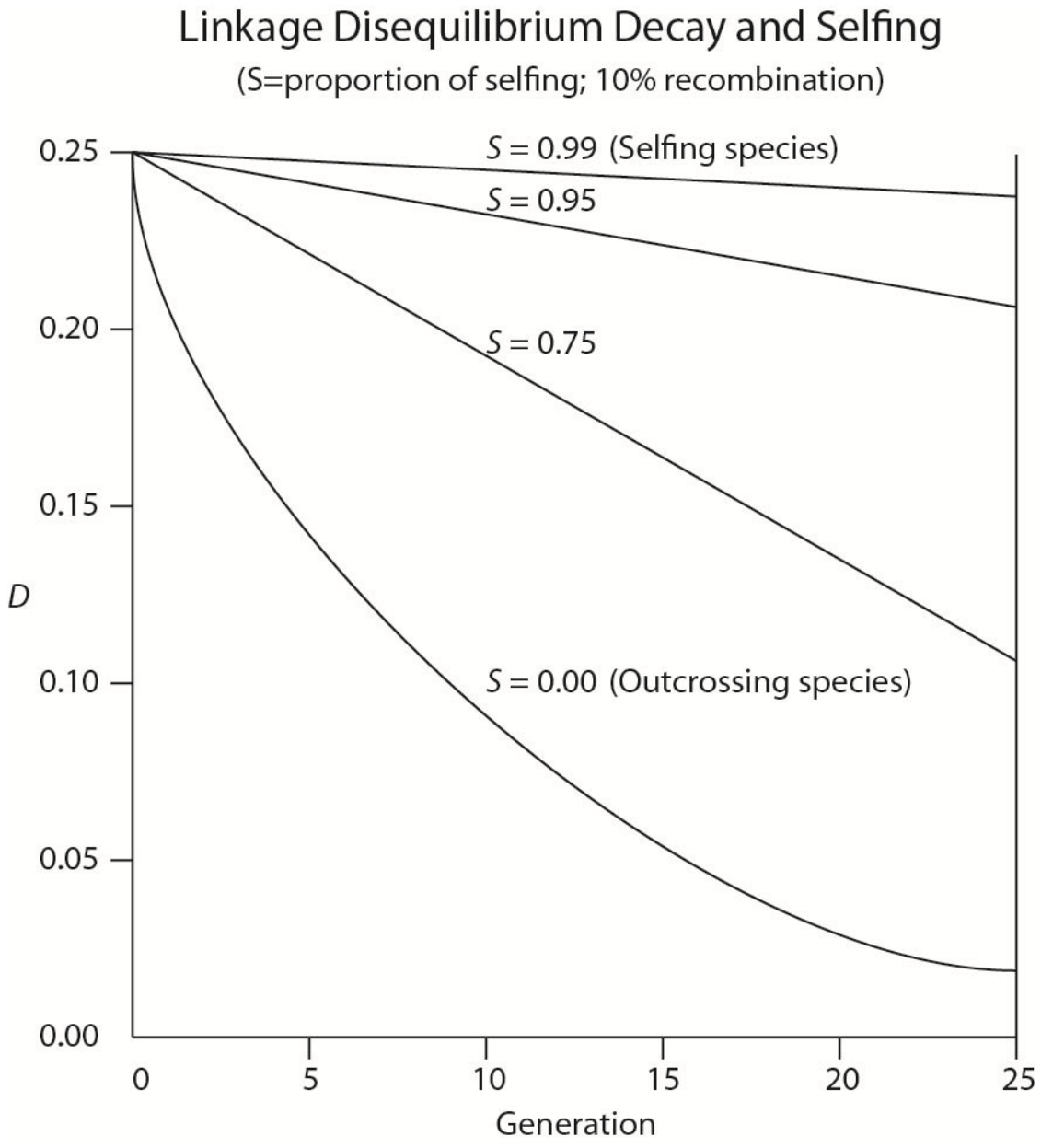
Copyright © 2004 Pearson Prentice Hall, Inc.

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



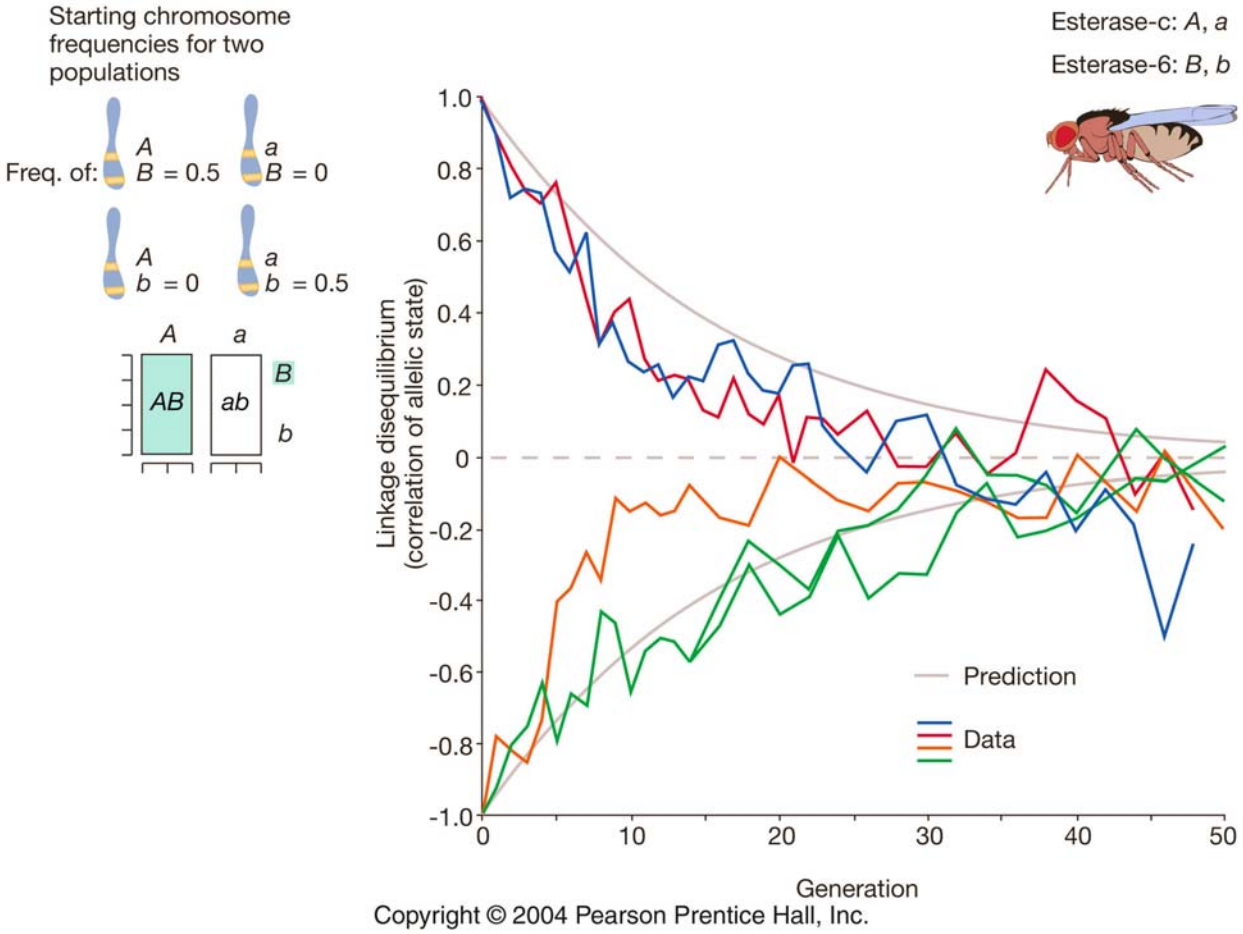
Mutation

- 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



Copyright © 2004 Pearson Prentice Hall, Inc.

Association Mapping in Plants

Traditional QTL approach

- Uses standard bi-parental mapping populations
  - F2 or RI
- 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 contributes to a disease that affects humans
- **Plants**
  - ***Common variant/common phenotype***
    - A specific SNP in a specific gene contributes to a phenotype of importance that affects a plant species

Important Concept Related to Principle

- AM
  - Useful for discovering common variant
  - Each locus may account for only a small amount of the variation
    - But enough of the alleles are present to affect the mean of a specific genotypic class
- Bi-parental mapping
  - Useful for discovering rare alleles that control a phenotype
  - Why??
    - Population has many copies of the rare allele
      - The allele will have an effect on the population phenotype
  - These alleles typically have a major effect

Idealized Cases Results for AM

- No association between marker and phenotype

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

- Association between marker and phenotype

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

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

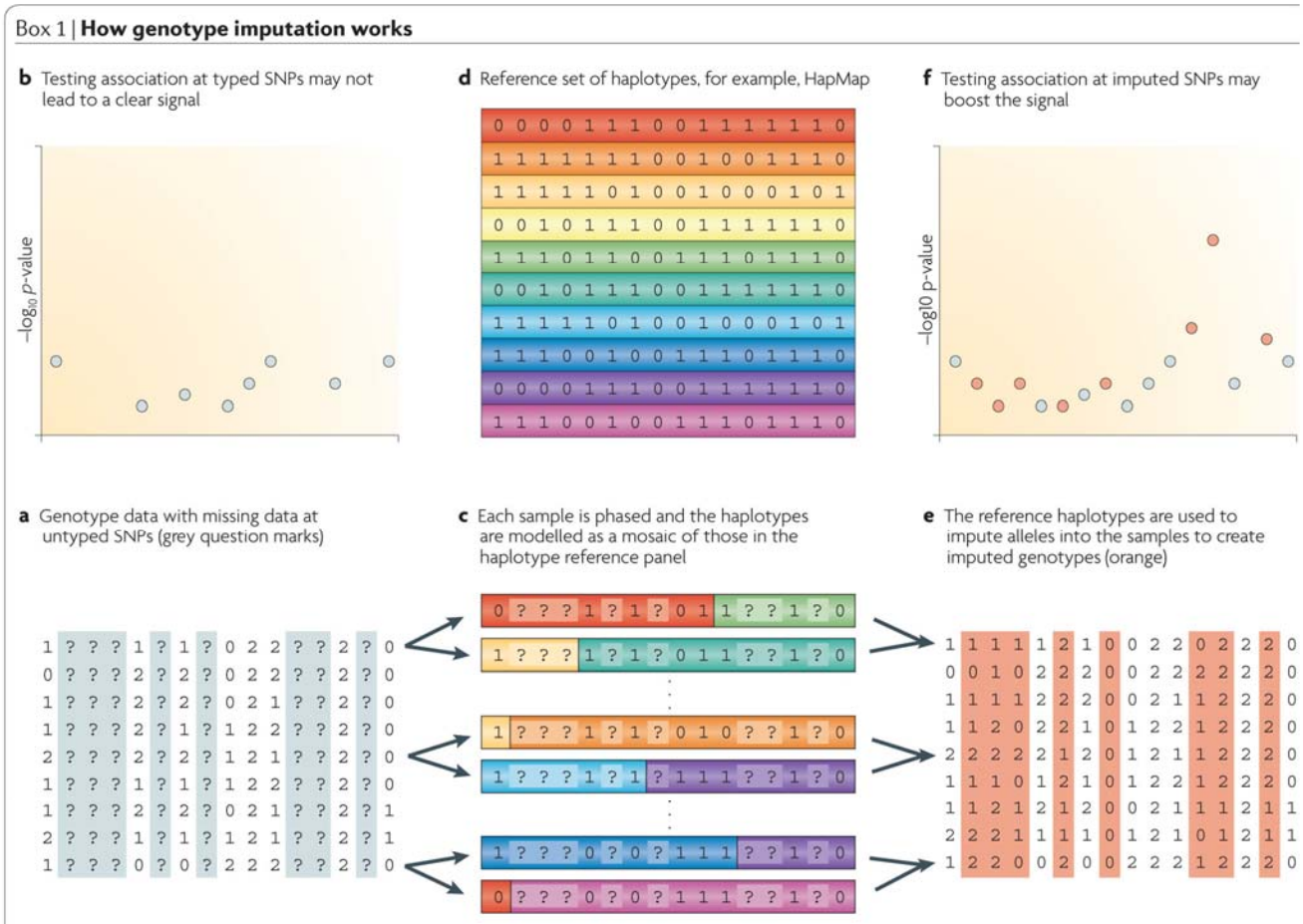
2. Genotype the population

- Genome-wide marker scan
    - Low density/lower resolution
        - ~100 SSR markers
            - Gel-based (in expensive to perform
    - Medium density
        - ~1500 SNP
        - Golden Gate assay
    - High density
        - 50,000-500,000 SNPs
            - Arabidopsis
                - 250,000 SNPs
                    - Affymetrix chip

- Candidate gene
    - Select genes that might control trait
        - Sequence different genotypes
        - Discover SNPs in gene
            - Consider
                - 5'-UTR
                - Coding region
                - 3'-UTR

3. Imputation of data

- LD in reference set used to estimate genotype at missing data sites
- Nature Genetics Review (2010) 11:499



In samples of unrelated individuals, the haplotypes of the individuals over short stretches of sequence will be related to each other by being identical by descent (IBD). The local pattern of IBD can be described by an (unobserved) genealogical tree, which will differ at different loci throughout the genome owing to recombination. Imputation methods attempt to identify sharing between the underlying haplotypes of the study individuals and the haplotypes in the reference set and use this sharing to impute the missing alleles in study individuals. For this reason, there are strong connections between the models and methods used to infer haplotype phase and those used to perform genotype imputation[22,37], as well as strong connections to tagging SNP-based approaches[19,21,38] and methods used in linkage studies[39,40].

The figure above illustrates imputation for a sample of unrelated individuals. The raw data consist of a set of genotyped SNPs that has a large number of SNPs without any genotype data (part **a**). Testing for association at just these SNPs may not lead to a significant association (part **b**). Imputation attempts to predict these missing genotypes. Algorithms differ in their details but all essentially involve phasing each individual in the study at the typed SNPs. The figure highlights three phased individuals (part **c**). These haplotypes are compared to the dense haplotypes in the reference panel (part **d**). Strand alignment between data sets must be done before this comparison takes place (Supplementary information S1 (box)). The phased study haplotypes have been coloured according to which reference haplotypes they match. This highlights the idea implicit in most phasing and imputation models that the haplotypes of a given individual are modelled as a mosaic of haplotypes of other individuals. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set (part **e**). In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced. It is necessary to take account of this uncertainty in any downstream analysis of the imputed data. Testing these imputed SNPs can lead to more significant associations (part **f**) and a more detailed view of associated regions.

4. Accounting for Population Structure/Relatedness

- Define subpopulations
    - Select markers to genotype the population
        - Markers should
            - Distributed among all chromosomes
            - All should be in linkage equilibrium
            - Minor allele frequency >0.1
    - Evaluating population structure
        - Principal component (PC)
            - Fixed effect
            - Defines groups of individuals
            - Select number of principal components that account for specific amount of variation
                - 50% is a typical value
        - STRUCTUE software
            - Fixed effect
            - Use matrix of percentage population membership in analysis
                - Original approach
                - Discontinued because of low power
    - Evaluate relatedness
        - Spagedi relatedness calculations
            - Random effect
            - Output is a table with all pairwise-comparisons

5. Statistical Analysis

- Marker-by-marker analysis
  - Regression of phenotype onto marker genotype
    - Significant marker/trait associations discovered
- Analysis must controlling 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.

6. Choosing the correct model

- Evaluate all models individually
  - Naive
    - No correction for structure or relatedness
  - PC
  - PC and relatedness
  - STRUCTURE
  - STRUCTURE and relatedness
  - Today
    - PC or PC/relatedness most often used
- Develop a P by P plot for each model
  - Y-axis
    - Cumulative P values
  - X-axis
    - Experimental P values for marker-by-marker analysis
  - Ideal situation
  - 5% of the cumulative P-values
    - Select the approach that is linear or nearly

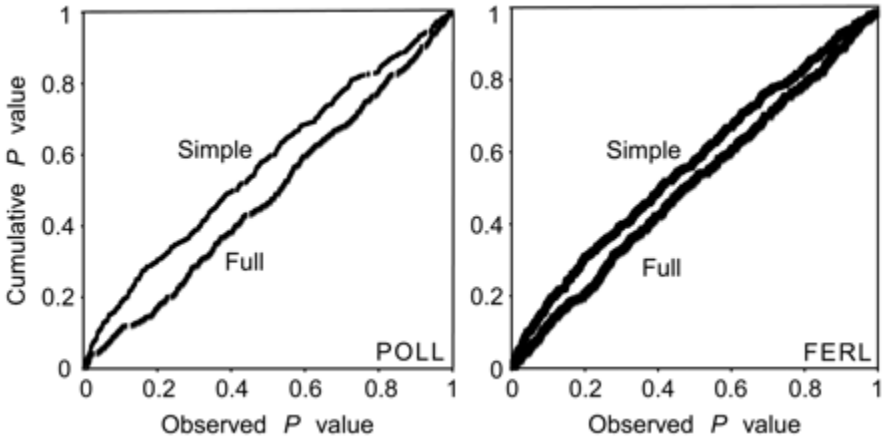


Figure 2. P-P plots for simple (no structure correction) for full (PC and relatedness correction). (From: Weber et al. 2008. Genetics 180:1221.

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
    - α/n n = number of comparisons
      - α = experiment wide error rate
      - n = number of comparisons
        - 1536 SNPs, n=1536
          - Bonferroni significance
            - $0.05/1536 = 3.3 \times 10^{-5}$
        - 250,000 SNPs, n=250,000
          - Bonferroni significance
            - $0.05/250{,}000 = 2.0 \times 10^{-7}$

    - Error rate of 0.05 and 100 comparisons
      - $P < 0.0005$ would be significant
  - Conservative approach
- False discovery rate
  - $P < 0.05$ of FDR value

Does AM Work??

Example: Arabidopsis flowering time and disease resistance genes
Aranzana et al. 2005, PLoS Genetics 1(5): e60

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

Results

- All four candidate loci strongly associated with expected phenotype
- Marker density had an effect
- Markers less than 10kb from loci strongly associated with phenotype for all traits

Comments on AM in Plants

- Most successful AM experiments have uncovered loci previously known to affect a trait

- Other experiments with traits not evaluated extensively before have just defined associated regions
    - Causative genes have not been defined

Notes on Human AM

Markers

- Affymetrix chips used
- Genome-Wide Human SNP Array 6.0
    - Latest development
        - 906,600 SNPs
        - 946,000 copy number variants

Sample size

- Thousands case (disease patient) and controls (normal patient)
    - Local or regional site
- >100,000 case/controls
    - Data pooled from experiments at multiple sites worldwide

Procedure for pooling data

- Each site uses one of a several gene chips
    - ~500,000 SNPs each
        - Some SNPs overlap between chips
- Imputation of genotype data
    - Estimates the genotype at a locus where data is missing
    - A procedure based on LD
        - Uses a reference set of haplotypes to predict the genotype

Example of pooling

- Multiple sites each used one of three 500k human SNP chips
- Imputating data creates a genotype set with 1.5 million SNPs across all samples
- Phenotypic data pooled over all sites
- AM analysis performed on all individuals from all sites
    - Examples with >100,000 case/controls now being reported
        - Many authors

Population structure effect

PLoS Genetics (2005) 1:e60

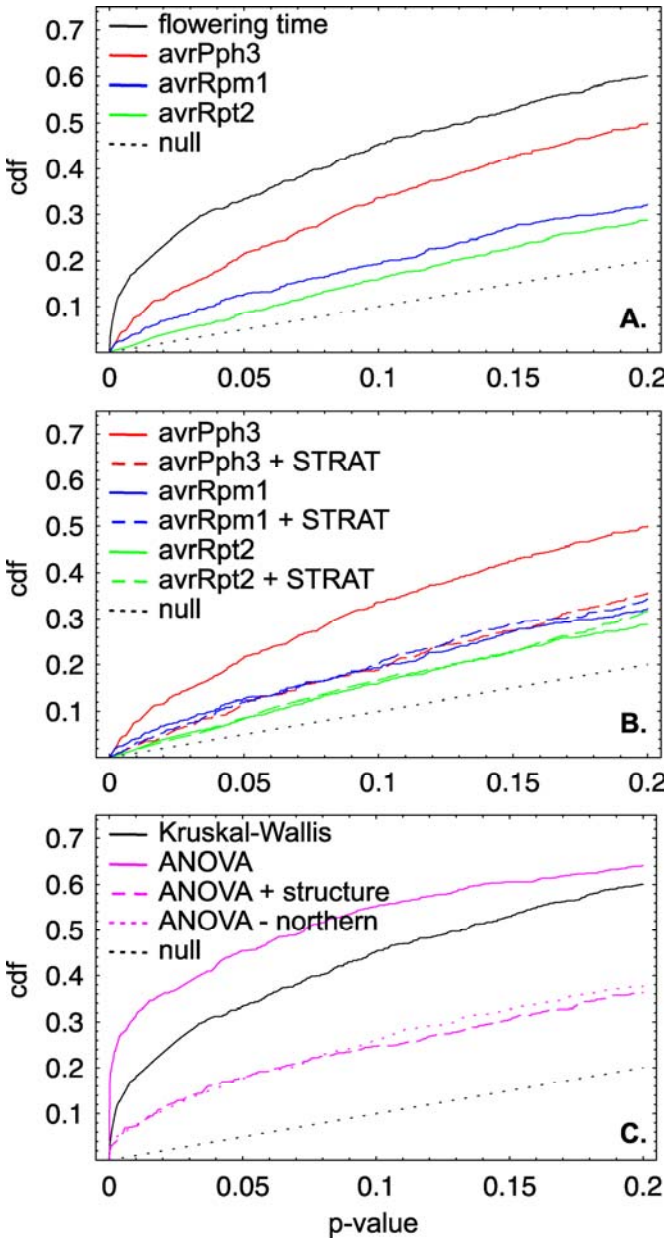
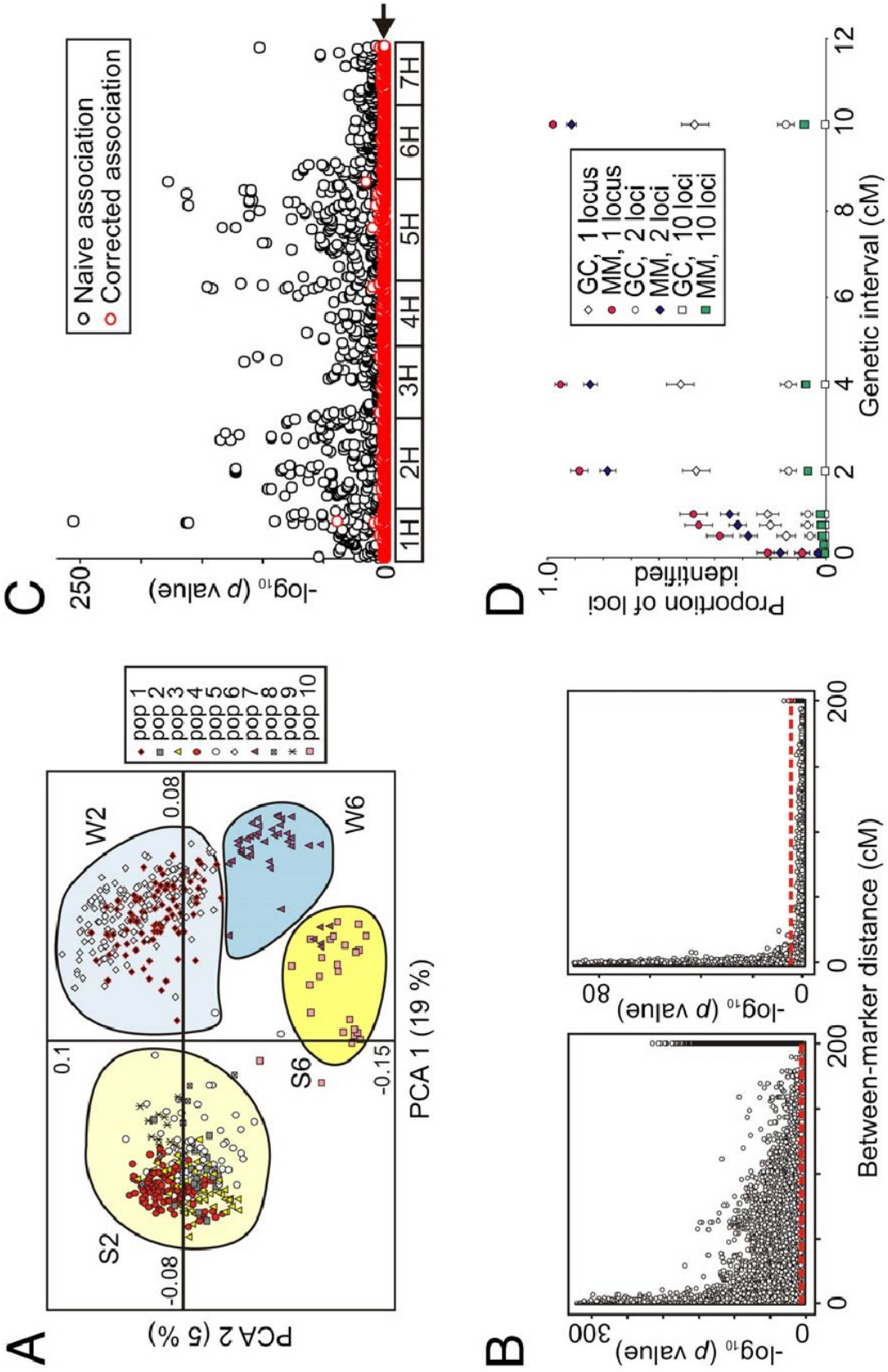


Fig. 2. P-P plots. Effect of models on detecting significant associations. If structure is accounted for, the line would be coincide with the dotted line.

Barley example

PNAS (2010)107:21611
1536 SNP, Illumina Golden Gate Assay



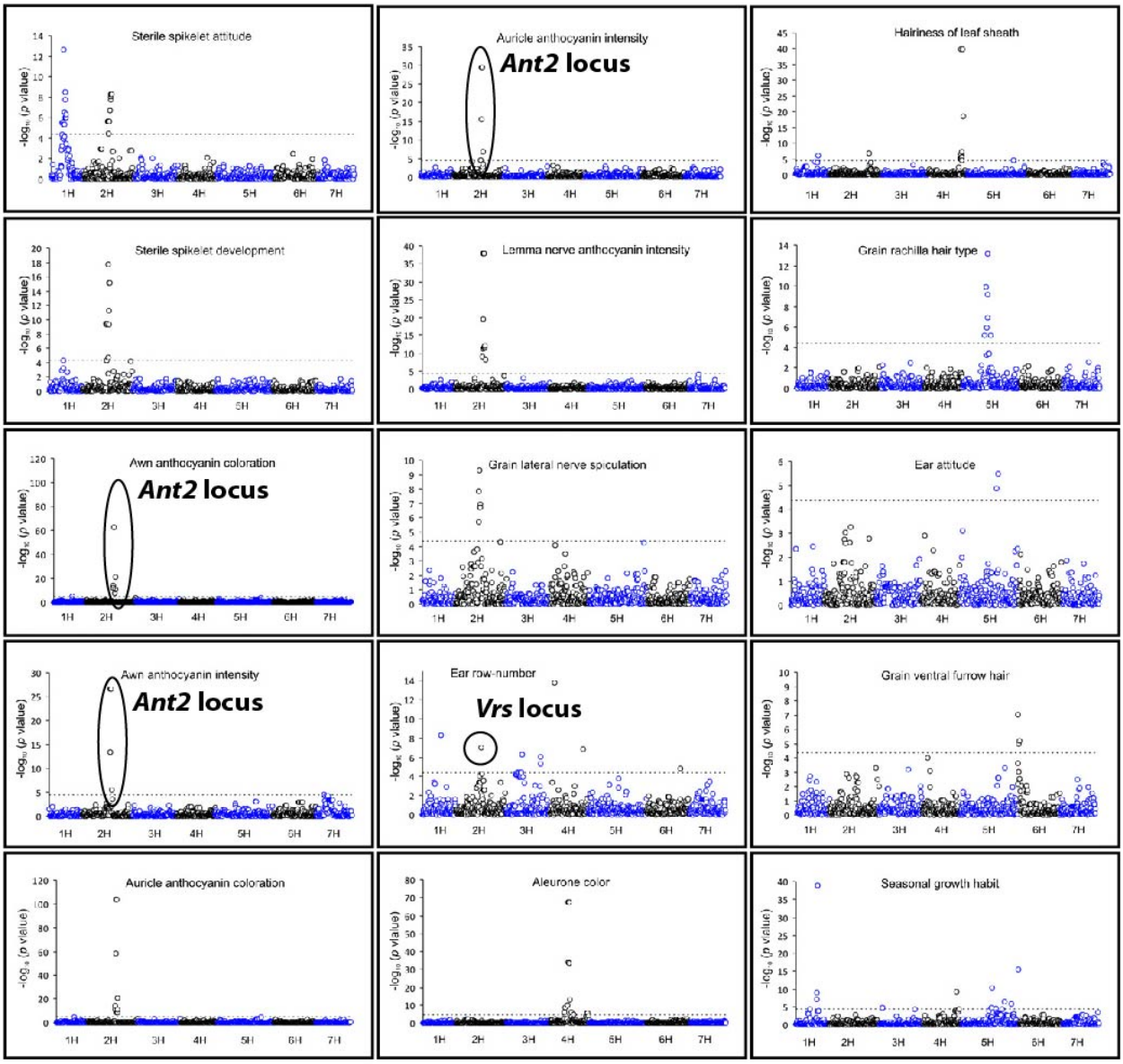


Fig. 2. GWA scans of the 15 traits with signi fi cant associations (P < 0.05, Bonferroni corrected; indicated by a dashed line). Barley chromosomes 1H to 7H are shown.

Candidate genes confirmed

- *Ant2*: controls anthocyanin color
- *Vrs*: controls ear row number (2 vs 6 row)

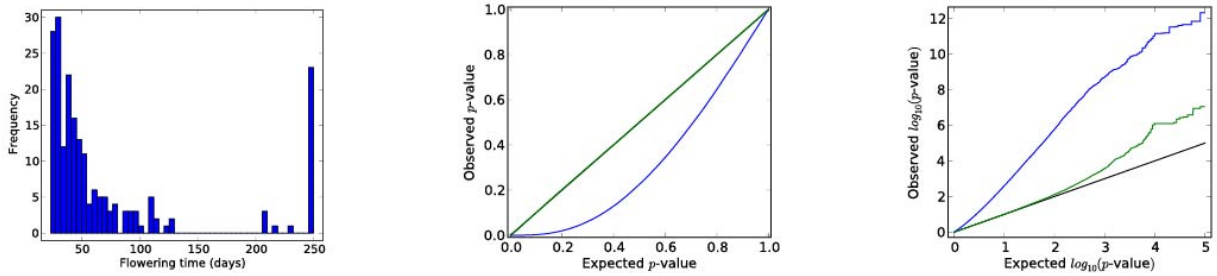
Arabidopsis: Analysis of 107 phenotypes

Nature: (2010) 465:627

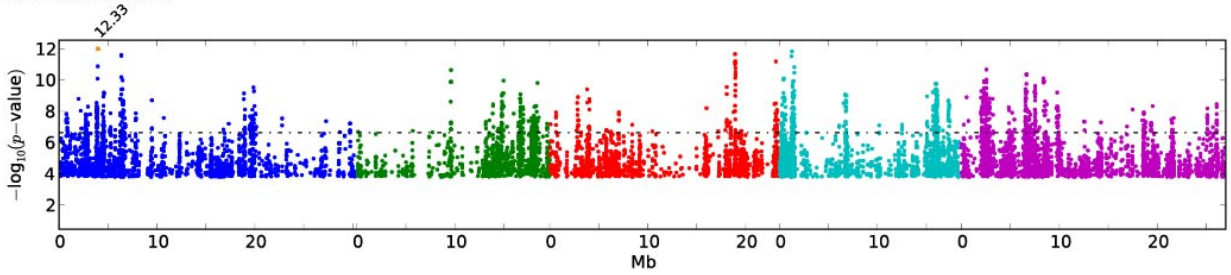
Notes

- 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

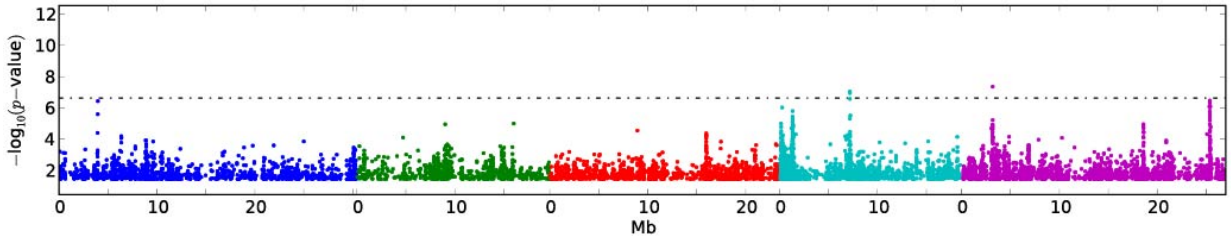


Wilcoxon results



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



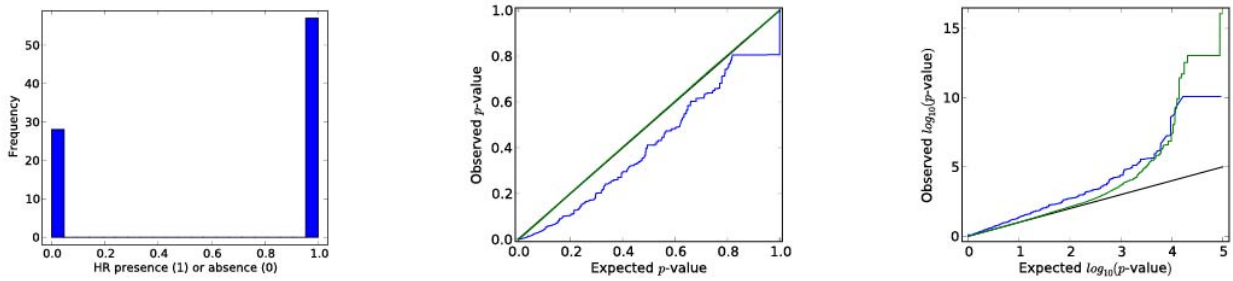
Rank	Score	Gene	Gene ID	Chr	SNP pos (bp)	Distance to gene (bp)
1	7.35652	FLC	AT5G10140	5	3188328	-8879
21	6.02586	sim to ESD4	AT4G00690	4	268809	-12836
21	6.02586	FRI	AT4G00650	4	268809	-217
39	4.95198	DOG1[B]	AT5G45830	5	18590971	15738
80	4.31728	CDF1	AT5G62430	5	25084106	2213
98	4.18876	ATARP4	AT1G18450	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)

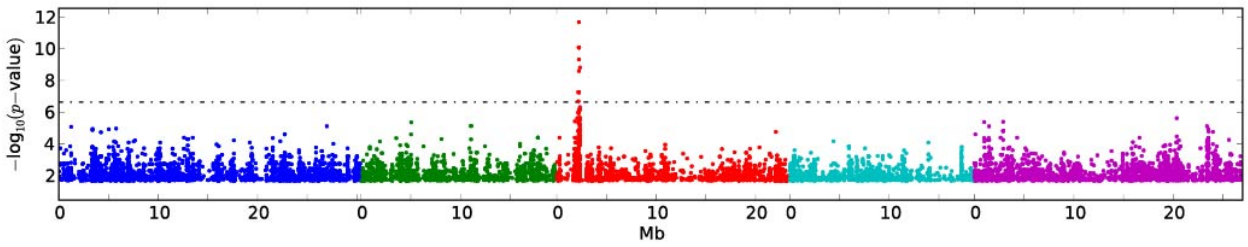
Notes

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

Phenotype histogram and quantile-quantile plots of p-values

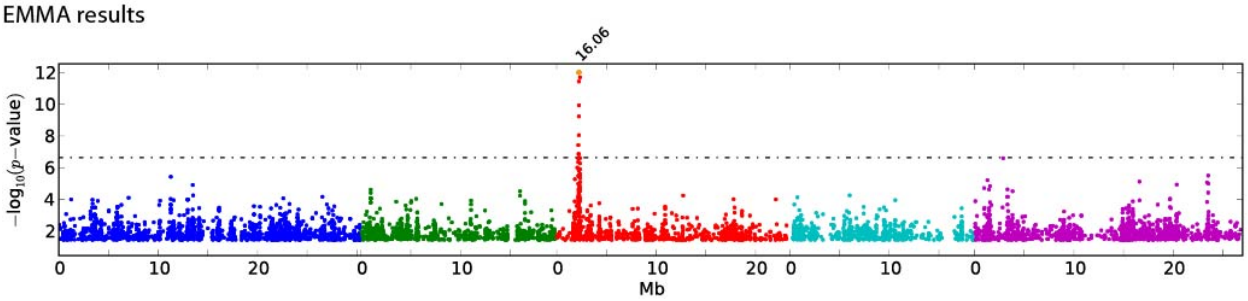


Fisher's exact test results



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 [g]	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 [g]	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

Association Mapping 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 through out 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 through out a population
    - AM of appropriate population