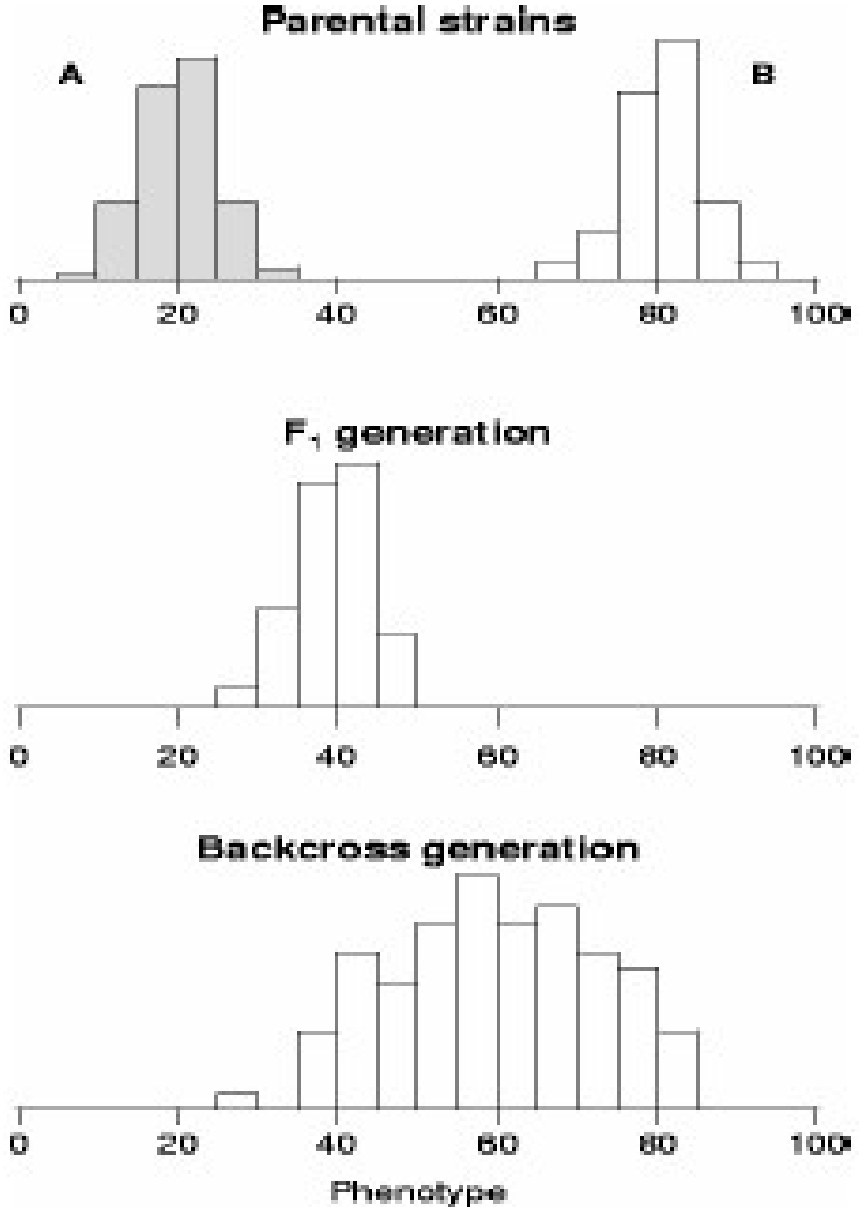
Quantitative trait loci (QTLs)

- **Quantitative trait locus (QTL)**: is a genomic confidence interval associated with a trait of interest, which varies in degree of effect size and physical length, and includes at least one causal gene or other functional element.
- QTLs determine the genetic component of variation in quantitative traits.
- Quantitative traits are usually encoded by many genes (polygenes).
- QTLs exert main, epistatic, and interaction with the environment effects, while the main effects can be additive or dominant.

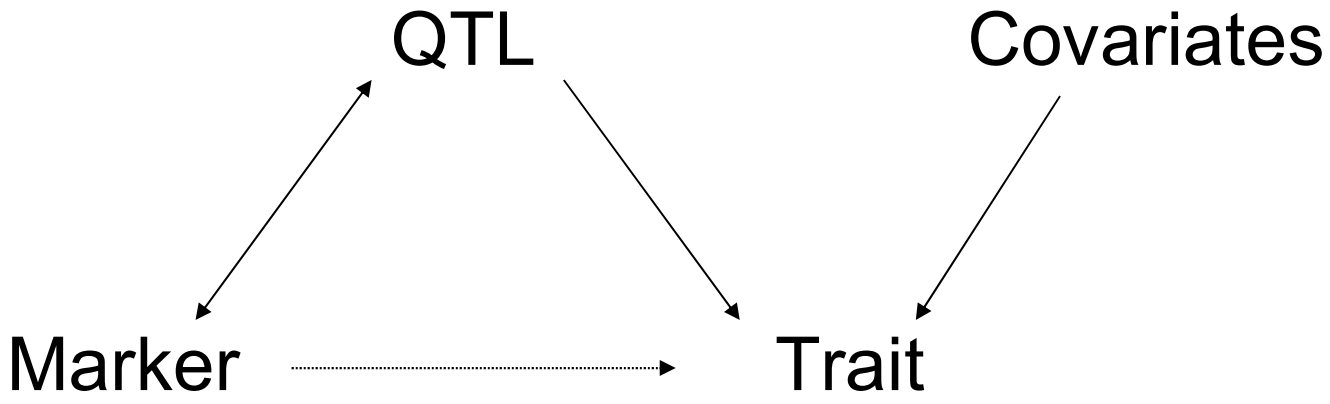

Parental strains
A
B
0
20
40
60
80
100
F1 generation
0
20
40
60
80
100
Backcross generation
0
20
40
60
80
100
Phenotype

Goals of QTL analysis

- Detect genetic effects
- QTL mapping: inference of the QTL location on chromosome

QTL mapping in experimental crosses

Experimental crossing creates associations between genetic marker loci and traits to allow localization of QTL.



QTL Mapping Procedure

White rust resistant

White rust susceptible

X

Donor Screening

Population Development

Phenotypic Data

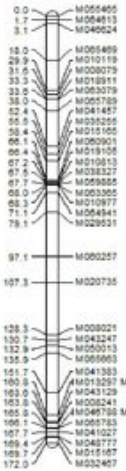
Genotypic Data

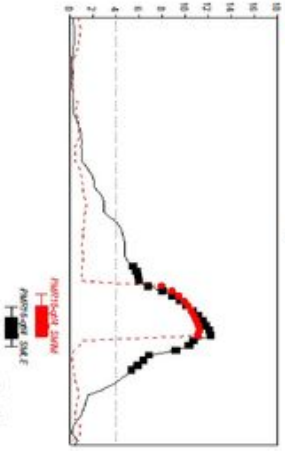
Data Analysis

QTL Mapping

Marker Identification

Marker Implementation





SNP	LOD	R^2
R298331	8.753	0.378
R347577	6.972	0.315
R296495	6.766	0.307

SNP markers

Molecular Breeding

QTL analysis: Single marker test, interval mapping (IM), composite interval mapping (CIM), and multiple-interval mapping (MIM) analysis are performed using Windows QTL Cart V2.5 and QGene and QTLNetworkv2.1.

Objectives of QTL analysis

To identify genomic regions

containing QTLs (Mapping QTLs)

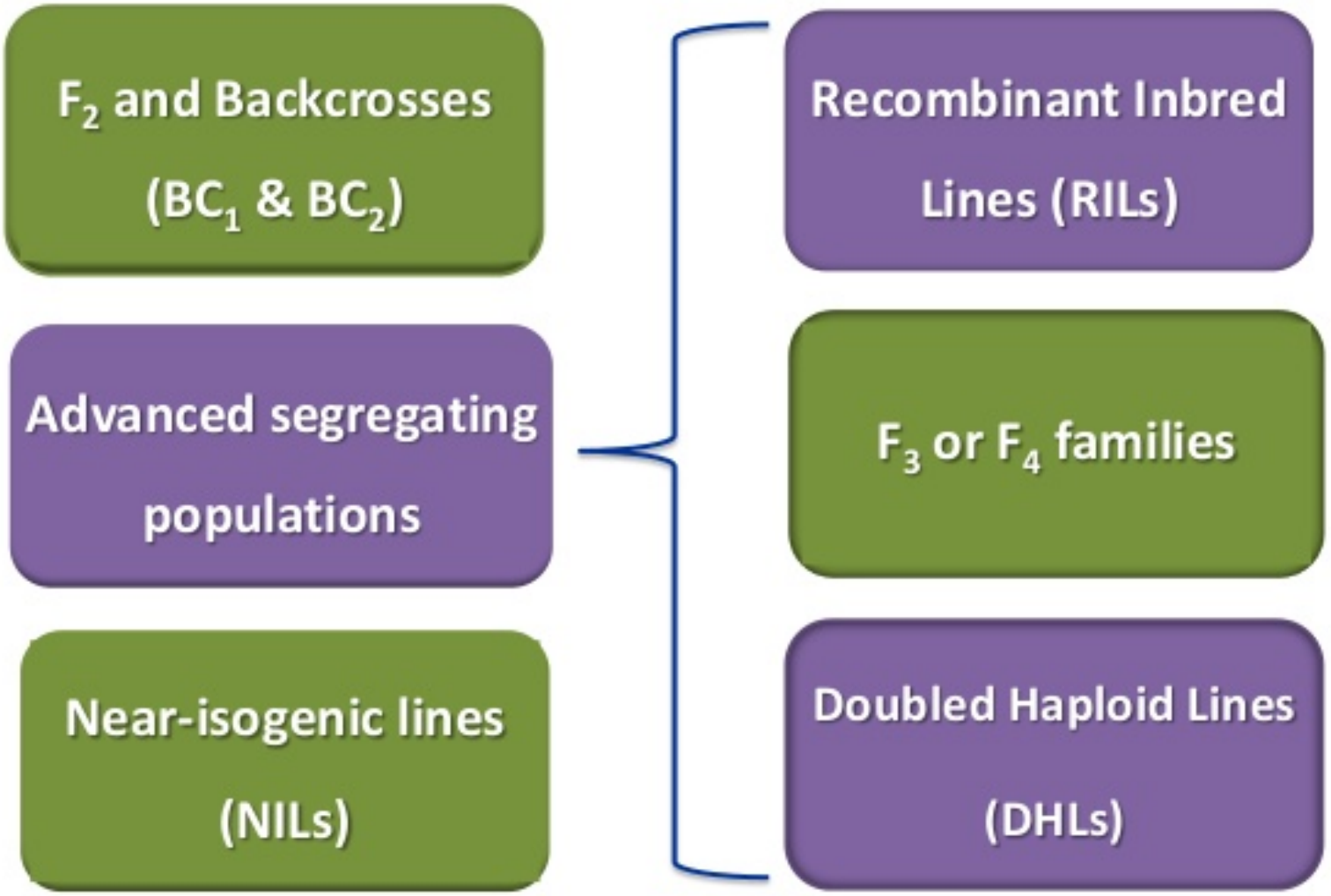
To estimate the genetic effects of the QTL:

- How much variation is caused by the QTL?
- The gene action associated with the QTL?
- Which allele is associated with favorable effect?

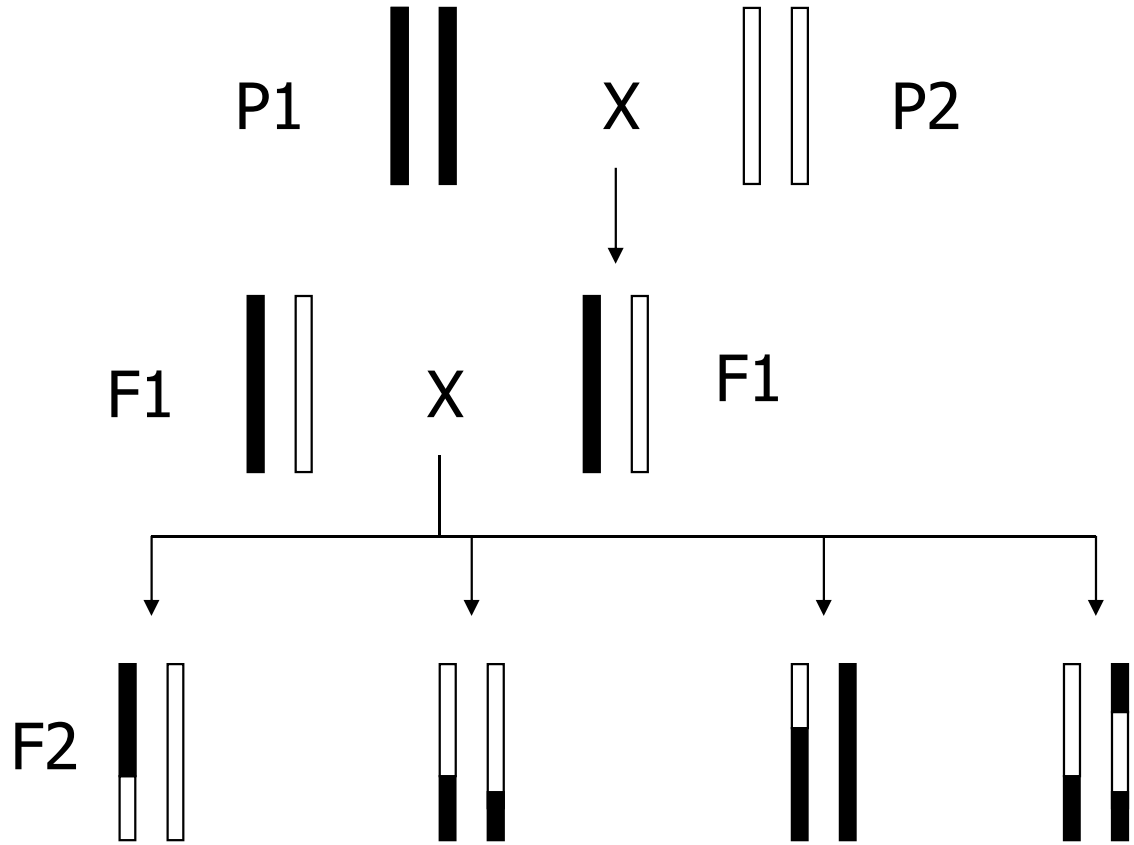
To identify markers tightly-linked to QTL

to be used for MAS in breeding programs

Mapping population



Intercross



Data structure for a backcross experiment

- Phenotypes:

  y_i = quantitative measurement of trait

- Genotypes:

  x_{ij} = 0/1 coded for AA/AB at marker j

- Covariates:

  $\mathbf{Z}_i$ = environmental factors, demographics, etc.

where i = 1, …, n; j = 1, …, M.

Model and assumptions

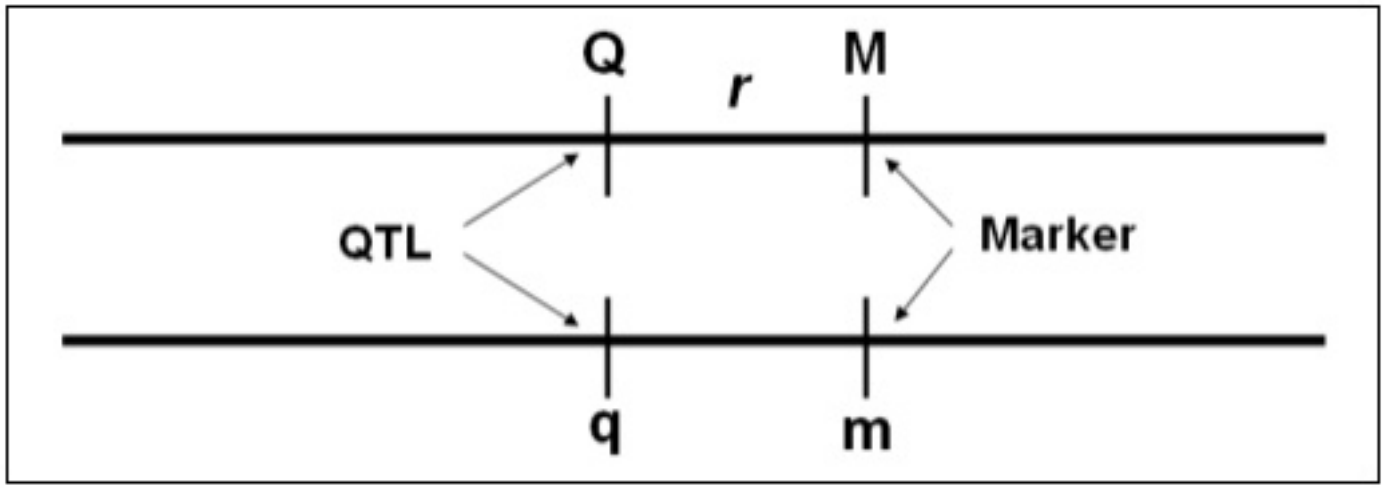
- No interference in the recombination process
- Independence
- Normal distribution

$$y_i|X \sim N(\mu_X, \sigma_X^2)$$

- Homoscedasticity (constant variance)

$$\sigma_X^2 = \sigma^2$$

Single-Marker Analysis (SMA)



- Studying single markers one at a time.
- The simplest method for QTL analysis.
- Can be done by t-tests or ANOVA.
- Does not require a complete linkage map.

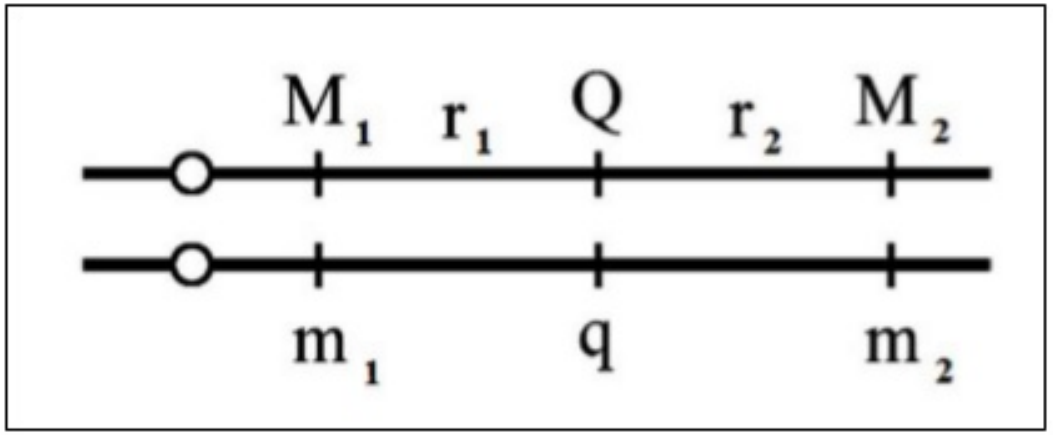
Limitations of SMA:

1) The effect is affected by the distance distance between the marker and the QTL

2) If several loci with positive and negative effects were linked to the marker, a global confounding QTL effect would be estimated instead of individual effects for each QTL

Simple-Interval Mapping (SIM)

Using flanking markers



- Two markers at a time (flanking markers).
- Based on linkage mapping of markers.
- More accurate than single-marker analysis.

Simple-Interval Mapping (SIM)

Maximum likelihood estimation of QTL position

Log $_{10}$ of the odds ratio (LOD score):

$$\text{Odds ratio} = \frac{\text{Likelihood of (there is a QTL in the marker interval)}}{\text{Likelihood of (there is no QTL in the marker interval)}}$$

- **A LOD score greater than 3.0 is considered evidence for linkage.**

LOD curve

- Likelihood profile
- A clear peak is taken as the QTL
- 1.5-LOD support interval

$$\text{LOD score} = Z = \log_{10}\left(\frac{\theta^{(\#\text{ OF RECOMBINANTS})} \cdot (1-\theta)^{(\#\text{ OF NONRECOMBINANTS})}}{(1/2)^{(\#\text{ OF RECOMBINANTS})}(1/2)^{(\#\text{ OF NONRECOMBINANTS})}}\right)$$

$$LOD = Z = \log_{10} \frac{\text{probabilità di nascita con un certo valore di linkage}}{\text{probabilità di nascita con linkage assente}} = \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{0.5^{(NR+R)}}$$

NR = numero di prole non-ricombinante,
R = numero di prole ricombinante.

0.5 al denominatore -> ogni allele completamente unlinked (e.g. alleli su cromosomi distinti) ha il 50% di possibilità di ricombinare
Teta = frazione ricombinante, ed è uguale a R / (NR + R)

LOD > 3.0 linkage (probability 1000:1 that linkage is not casual).
LOD < -2.0 sufficient to rule out linkage.

LOD SCORE

- l(ogarithm) of od(ds) score is used to calculate linkage and probability of recombination between two markers.
- Compare the probability of the observed values if two loci are on the same chromosome compared to the probability of observing those values by chance
- Positive LOD scores imply the presence of linkage

LOD SCORE

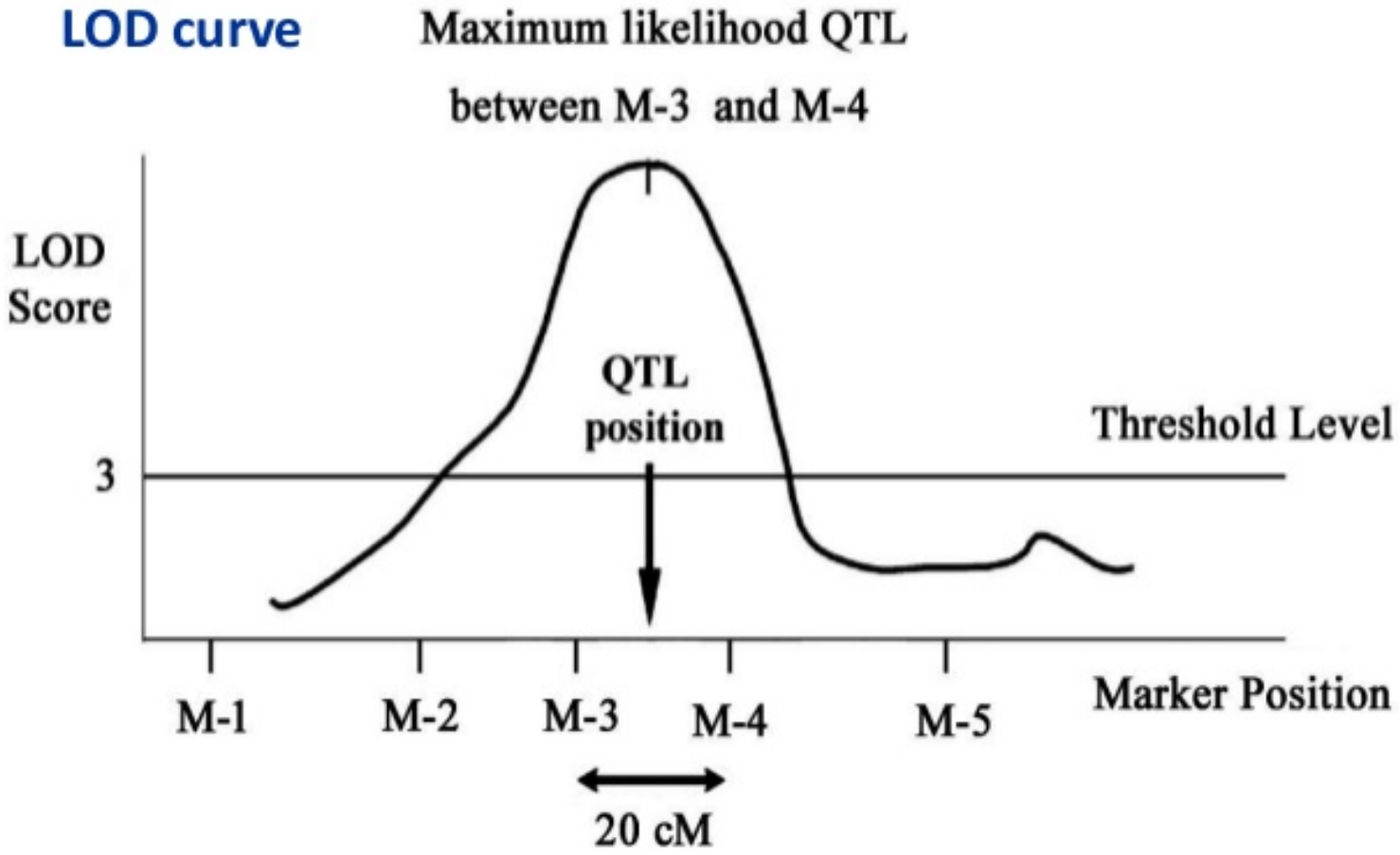
- Construction of pedegree;
- Estimate of frequency of recombination;
- calculation of LOD score for each estimate;
- The estimate with highest LOD score will be considered the best one

LOD Score Mapping

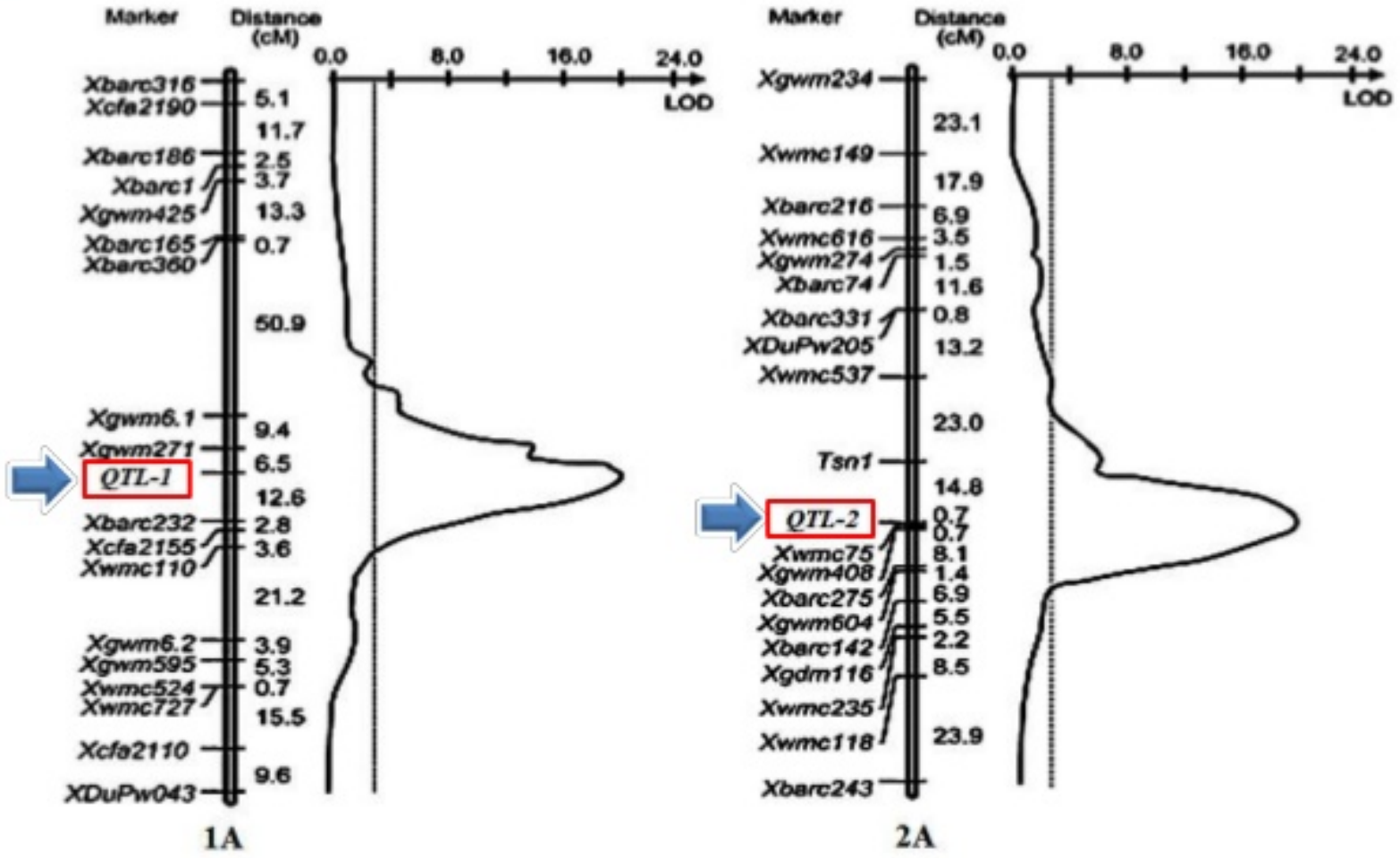
- The lod score method is an example of a maximum likelihood procedure.
- The point of the maximum likelihood procedure is to estimate the value of a parameter that can't be directly observed, in this case the recombination fraction.
- The likelihood (probability) of an observed set of data (the phenotypes seen in a family, in this case) is calculated as a function of that parameter.
- The parameter value that gives the maximum likelihood is taken as the best estimate of the parameter.

Simple-Interval Mapping (SIM)

LOD curve



Simple-Interval Mapping (SIM)



Breeders' QTL mapping 'checklist'

- LOD & R^2 values will give us a good initial idea but probably more important factors include:

1. *What is the population size used for QTL mapping?*
2. *How reliable is the phenotypic data?*
   - *Heritability estimates will be useful*
   - *Level of replication*
3. *Any confirmation of QTL results?*
4. *Have effects of genetic background been tested?*
5. *Are markers polymorphic in breeders' material?*
6. *How useful are the markers for predicting phenotype? Has this been evaluated?*

Reliability of QTL mapping is critical to the success of MAS

- Reliable phenotypic data critical!
  - Multiple replications and environments
- Confirmation of QTL results in independent populations
- “Marker validation” must be performed
  - Testing reliability for markers to predict phenotype
  - Testing level of polymorphism of markers
- Effects of genetic background need to be determined

Confirmation of QTL mapping

- **QTLs stability across environments**
- **Using different segregating populations**
- **Using Near-Isogenic Lines (NILs)**

  (differing only in the QTL of interest)

ERECTA receptor-like kinase and heterotrimeric G protein from Arabidopsis are required for resistance to the necrotrophic fungus *Plectosphaerella cucumerina*

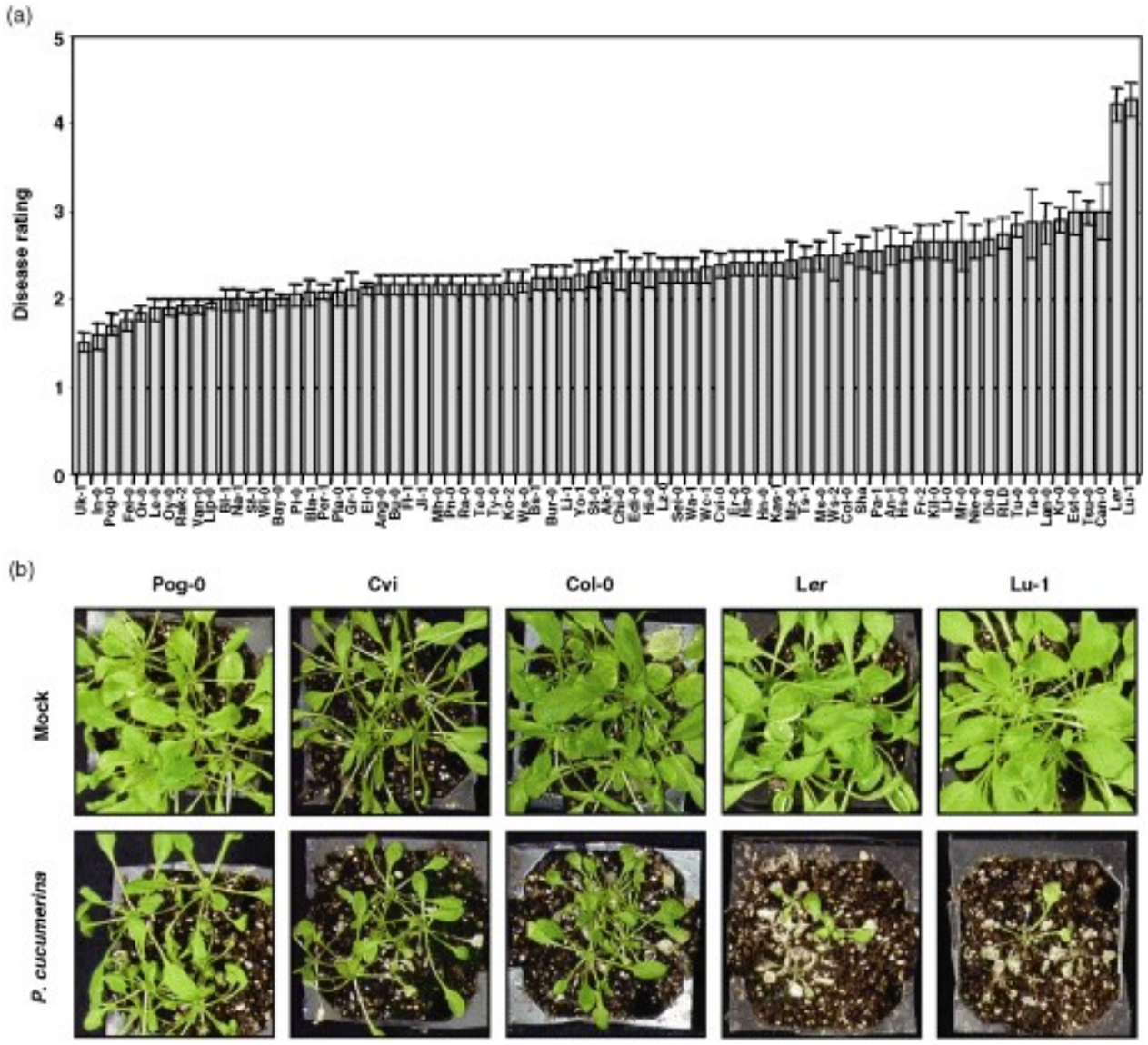
Francisco Llorente1, Carlos Alonso-Blanco2, Clara Sánchez-Rodriguez1, Lucía Jorda1 andAntonio Molina1,*

The Plant Journal
Volume 43, Issue 2, pages 165–180, July 2005

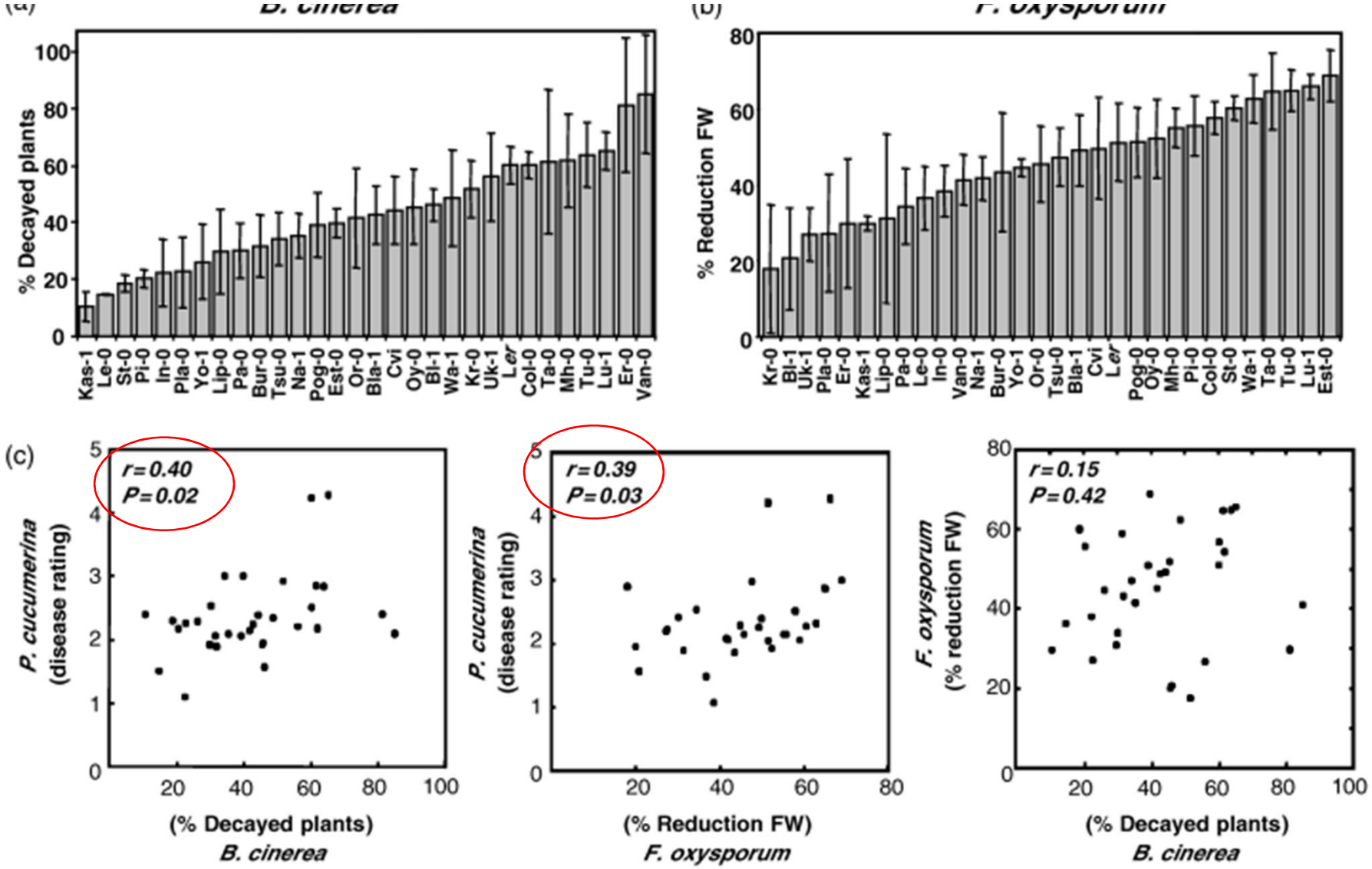
Plectosphaerella cucumerina

Fungo ascomicete necrotrofico che causa marciumi di frutti, foglie e del colletto in molte specie ortive

Diversa suscettibilità a *Plectosphaerella cucumerina* in accessioni di Arabidopsis



Risposta differenziale di accessioni di Arabidopsis ad altri funghi necrotrofi (*Botrytis cinerea* e *Fusarium oxysporum*)



QTL likelihood maps for *Plectosphaerella cucumerina* resistance in a L*er*/Cvi RIL population

- Full quantitative trait locus analysis in a RIL population derived from the cross between the **moderately susceptible Cvi** and the **highly susceptible Ler** (Alonso-Blanco et al., 1998).

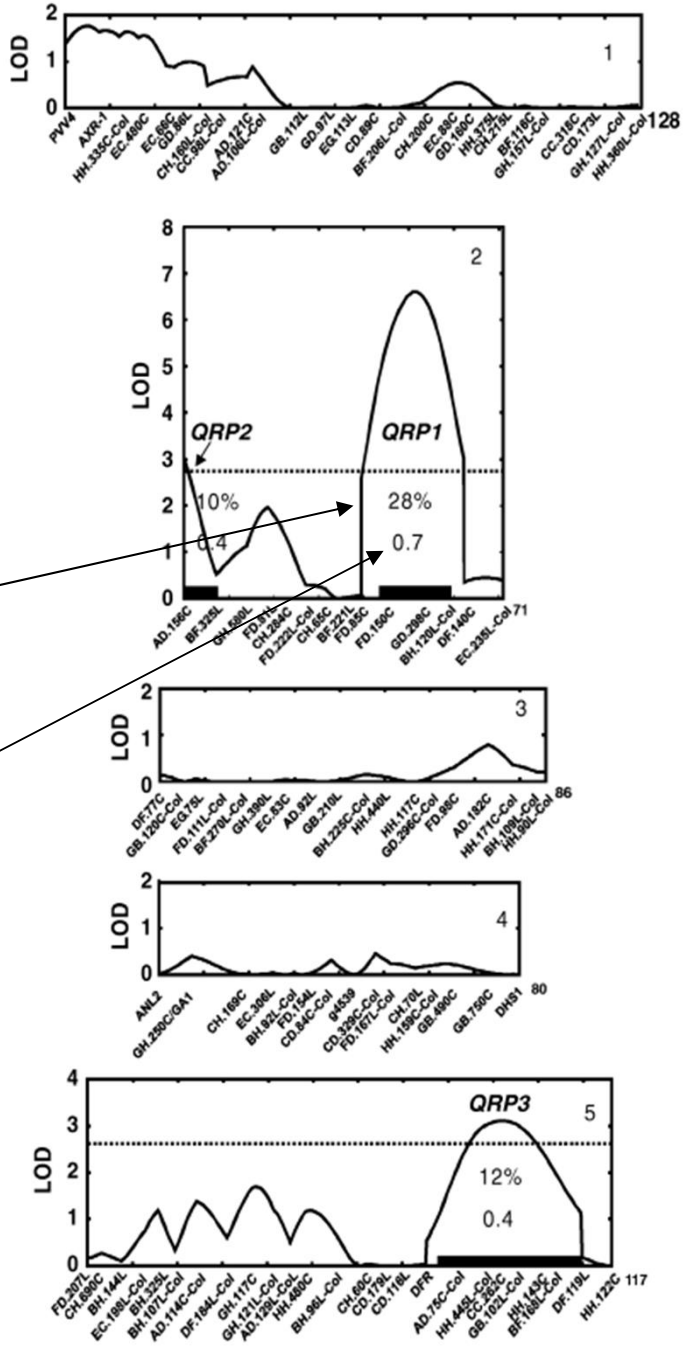
- Plants from **72 RILs Ler/Cvi**, as well as the parental accessions Ler and Cvi, were inoculated with a spore suspension of the fungus and their mean DRs were estimated.

Tre QTL responsabili per circa 50% della variabilità fenotipica

Soglia per rilevare i QTL =2.6 LOD

% della varianza fenotipica totale

Effetto additivo di ogni QTL (in Disease Rating): i valori positivi indicano che i genotipi La-er mostrano aumentata suscettibilità della popolazione Cvi



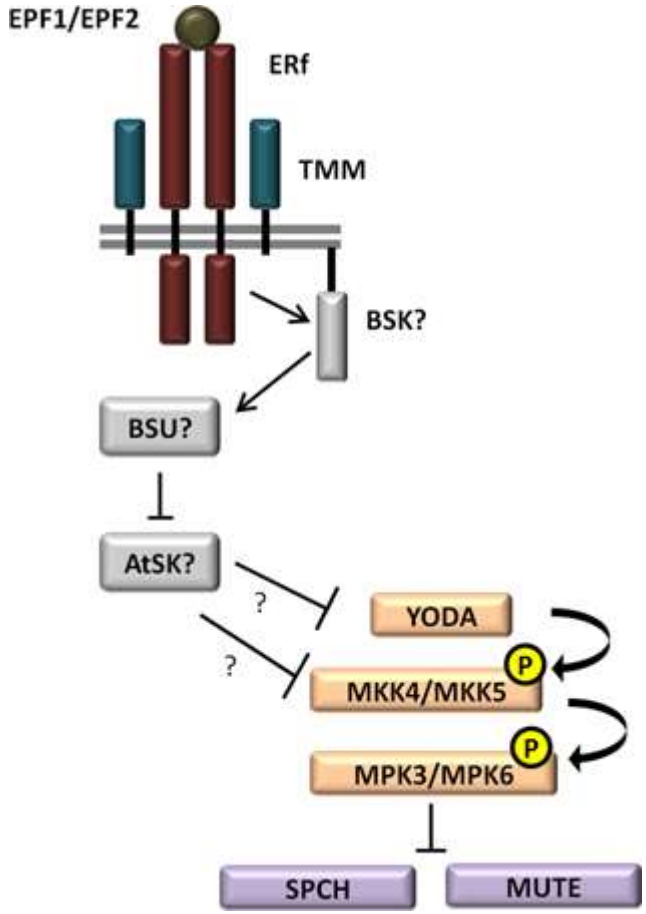
QRP1 (sul chr 2, vicino al gene *ERECTA*, o *ER*) è il locus con l'effetto più forte

Ler porta una mutazione loss-of-function (allele *er-1*) nel gene *ER* (Torii et al., 1996).

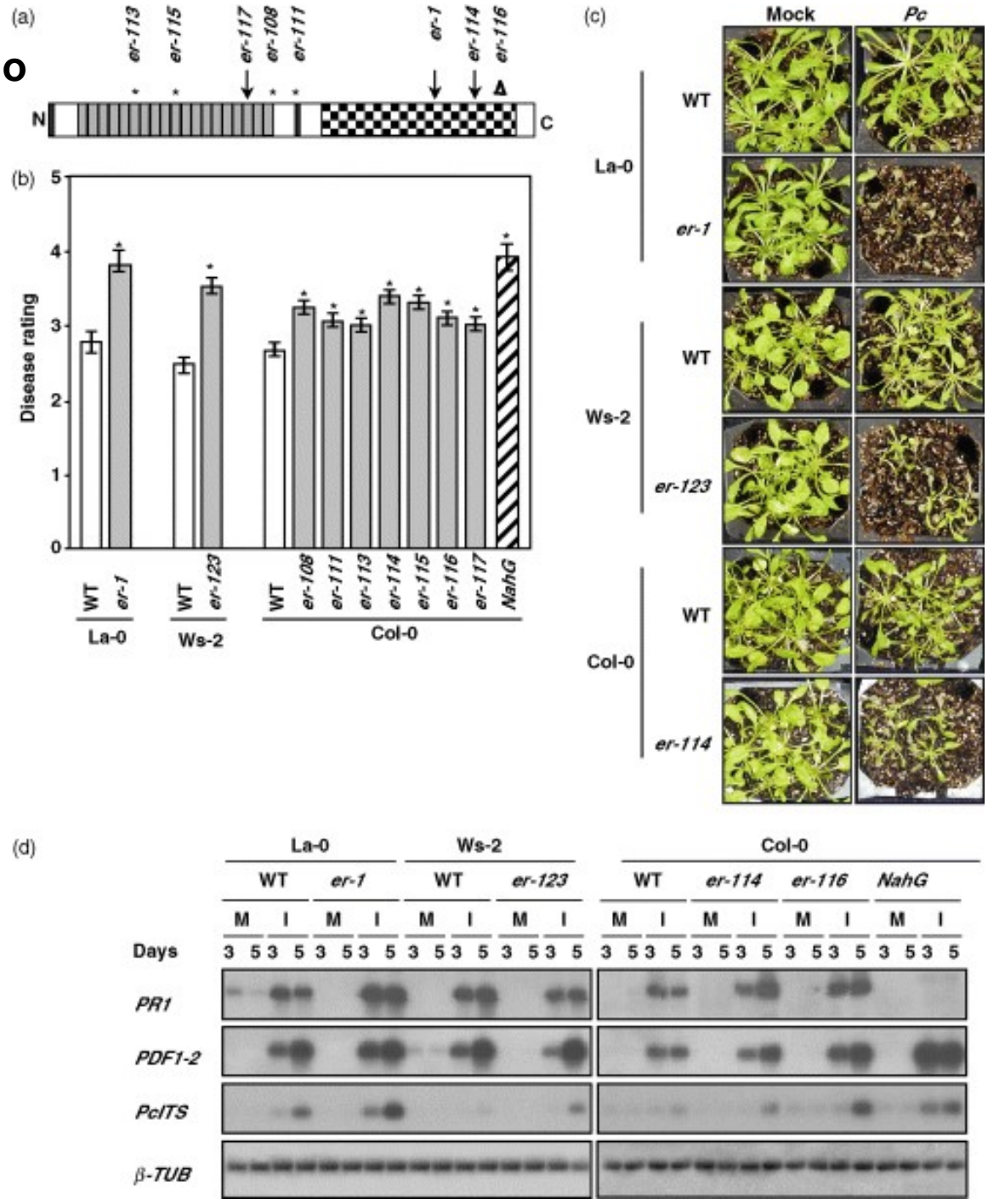
Stesso QTL osservato anche in una popolazione RIL Ler/Col

-> IPOTESI: il gene ERECTA è responsabile per il QTL QRP1

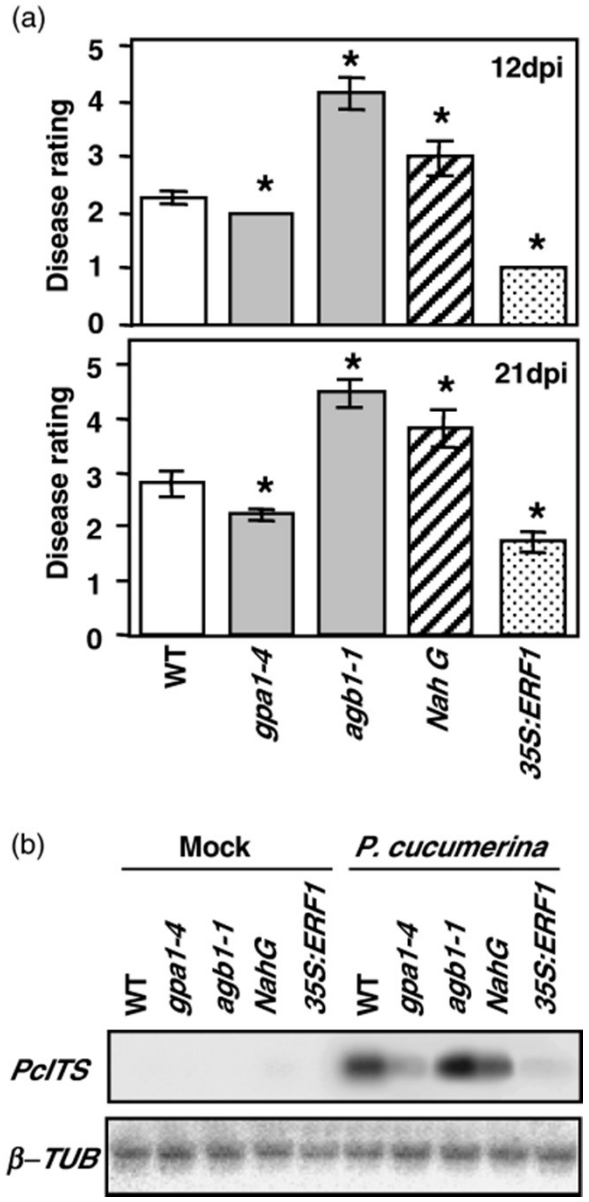
ERECTA: receptor-like kinase coinvolta in numerose funzioni



Differenti alleli loss-of-function per *ERECTA* rendono le piante più suscettibili al fungo



Tre mutanti nel pathway di trasduzione a valle di ERECTA sono più suscettibili al fungo



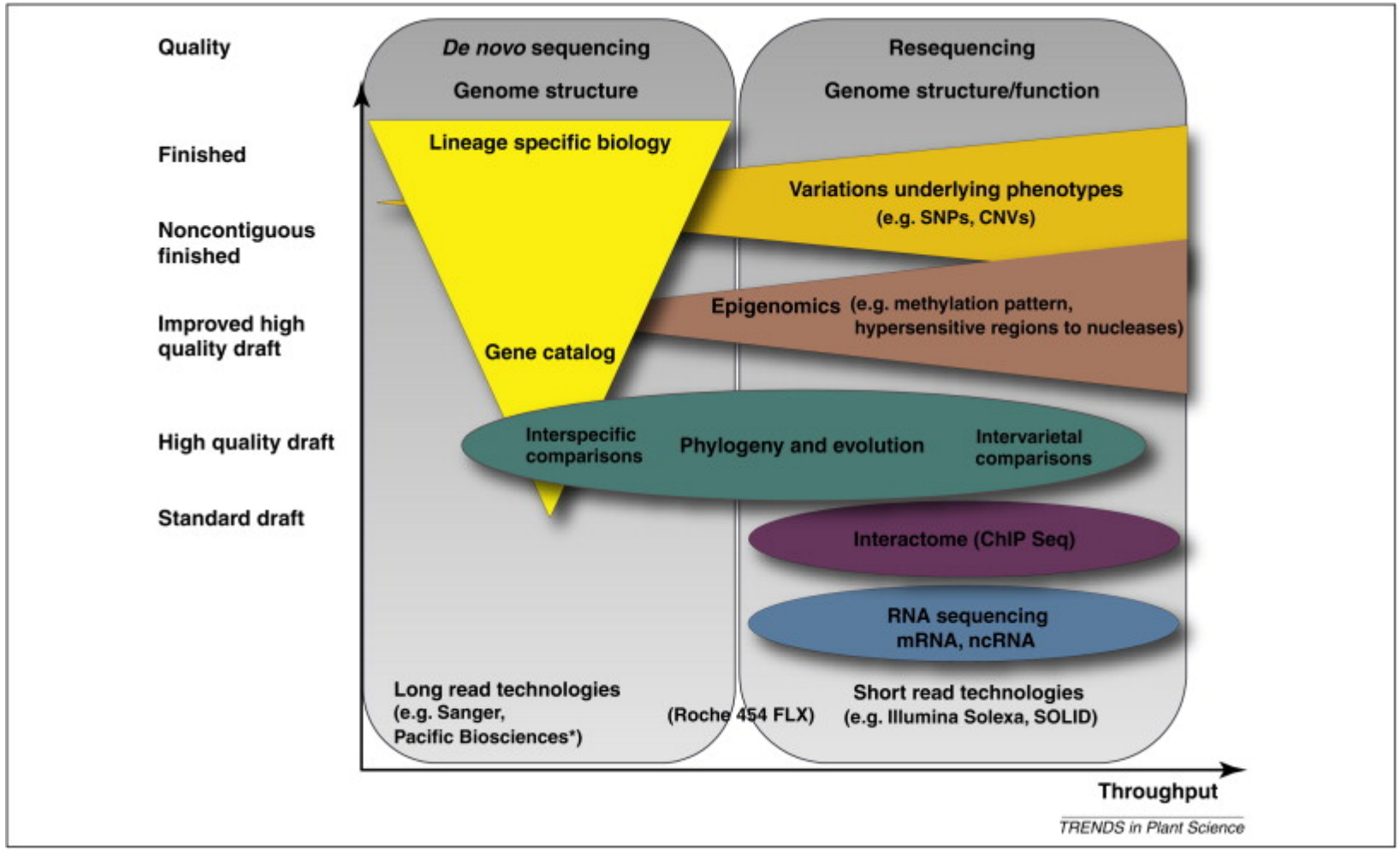
HIGH-THROUGHPUT GENOTYPING

Table 1 Five high-throughput genotyping methods

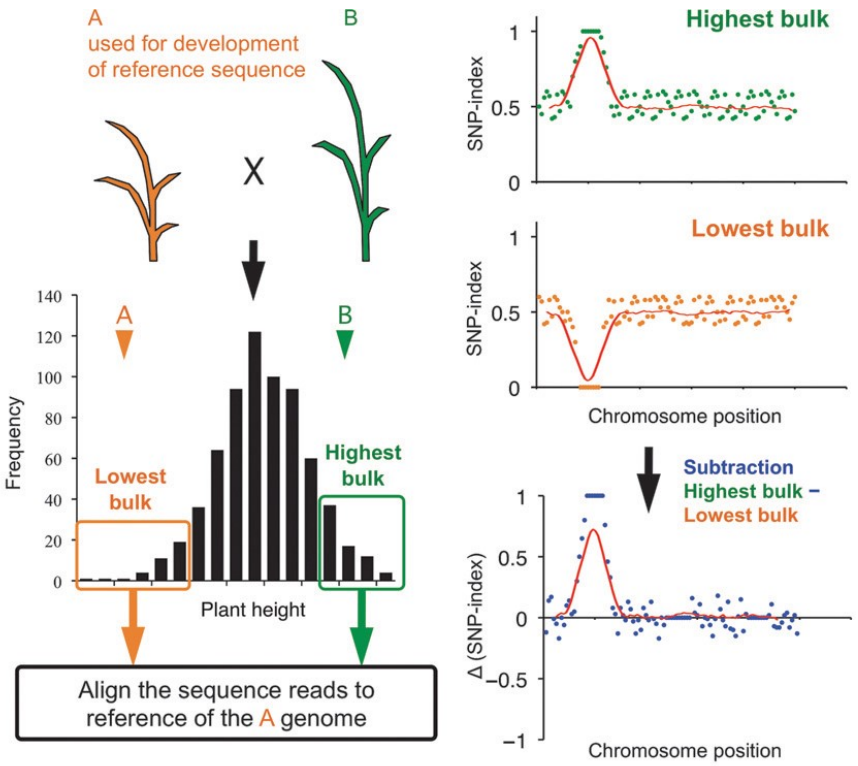
	Microarray-based genotyping	Sequencing-based genotyping	Genotyping by sequencing	RNA-seq-based genotyping	Exon-sequencing-based genotyping
Preliminary requirement	Comprehensive SNPs available	None	A suitable restriction enzyme	None	Exon array developed
Density	Alterable	Alterable	Modest	Modest	Modest
Cost	Alterable	Alterable	Low	High	High
Experimental workload	Low	Medium	Medium	High	High
Marker distribution	Well distributed	Well distributed	Not well distributed	Not well distributed	Not well distributed
Application	Most species	Most species	Species with a large genome size	Species with a large genome size	Species with a large genome size
Additional uses	None	Identifying novel mutation variants	None	Identifying novel mutation variants and eQTL analysis	Identifying novel mutation variants

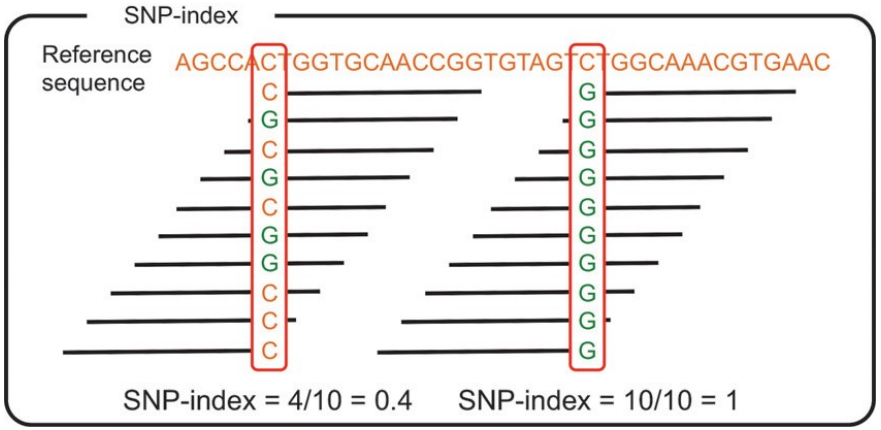
Abbreviations: eQTL, expression quantitative trait locus; SNP, single-nucleotide polymorphism.

Resequencing di genomi di specie coltivate -> marcatori (SNPs, indels, presenza/assenza di geni espressi)



Bulked segregant analysis for QTLs





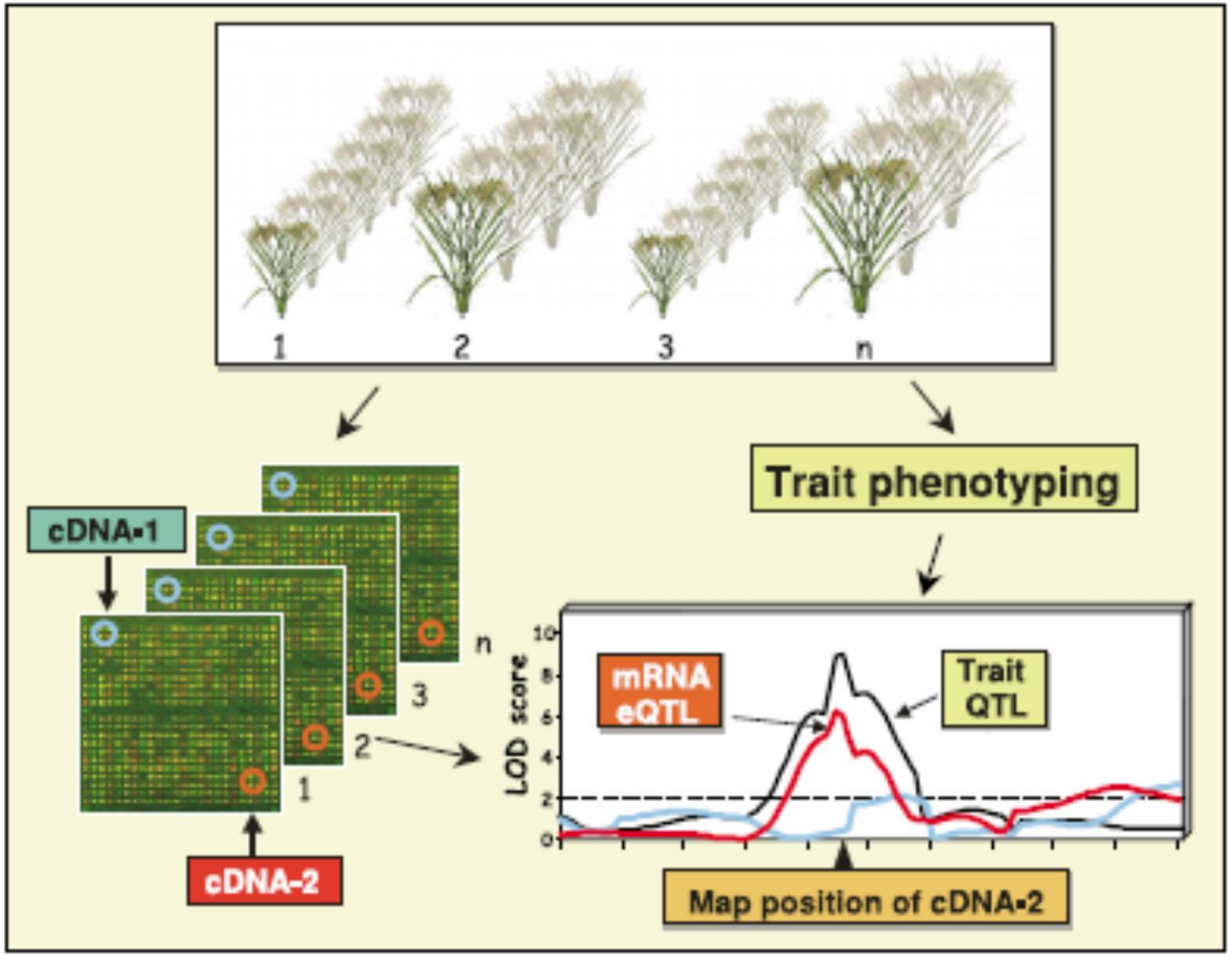
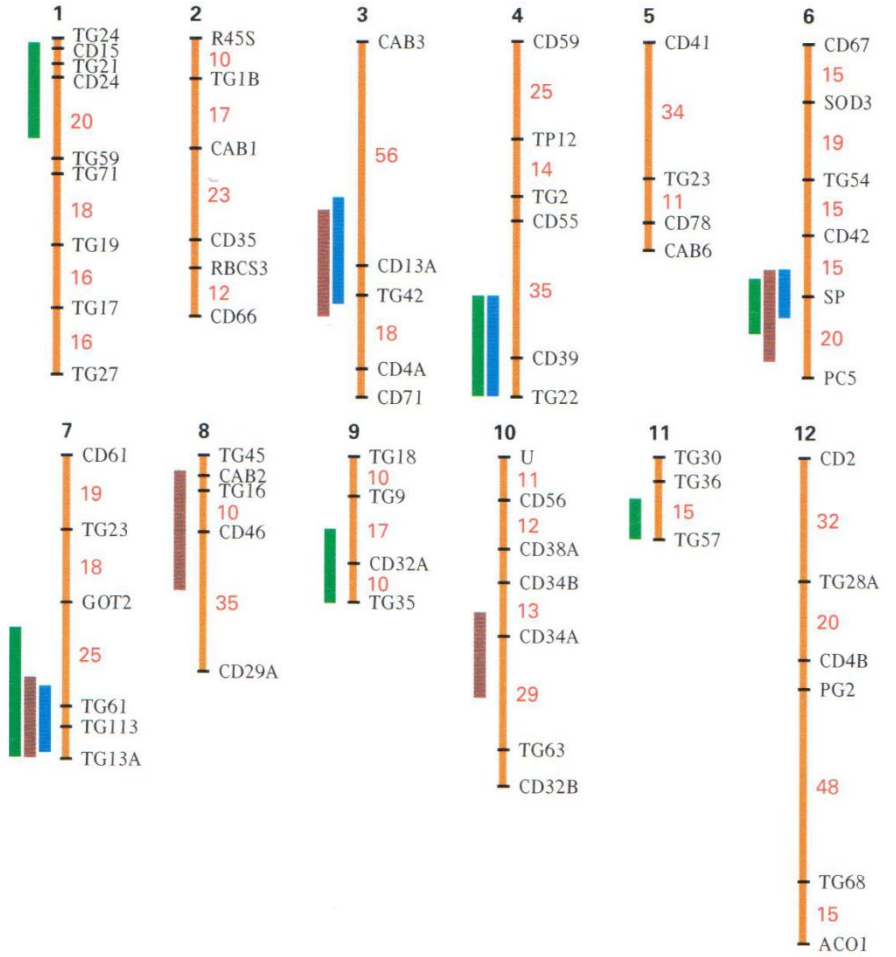


Figure 5. Expression profiling of a mapping population at the mRNA level via microarray analysis to identify expression QTLs (eQTLs) for specific cDNAs. Correspondence between an eQTL peak for a specific cDNA (e.g. cDNA-2) and a QTL peak for a trait causally linked to the function of the protein encoded by the cDNA provides circumstantial evidence supporting the role of the cDNA as a candidate gene for the target trait.


1
TG24
CD15
TG21
CD24
20
TG59
TG71
18
TG19
16
TG17
16
TG27
2
R45S
10
TG1B
17
CAB1
23
CD35
RBCS3
12
CD66
3
CAB3
56
CD13A
TG42
18
CD4A
CD71
4
CD59
25
TP12
14
TG2
CD55
35
CD39
TG22
5
CD41
34
TG23
11
CD78
CAB6
6
CD67
15
SOD3
19
TG54
15
CD42
15
SP
20
PC5
7
CD61
19
TG23
18
GOT2
25
TG61
TG113
TG13A
8
TG45
CAB2
TG16
10
CD46
35
CD29A
9
TG18
10
TG9
17
CD32A
10
TG35
10
U
11
CD56
12
CD38A
CD34B
13
CD34A
29
TG63
CD32B
11
TG30
TG36
15
TG57
12
CD2
32
TG28A
20
CD4B
PG2
48
TG68
15
ACO1
Peso frutto
Acidità
Solidi solubili

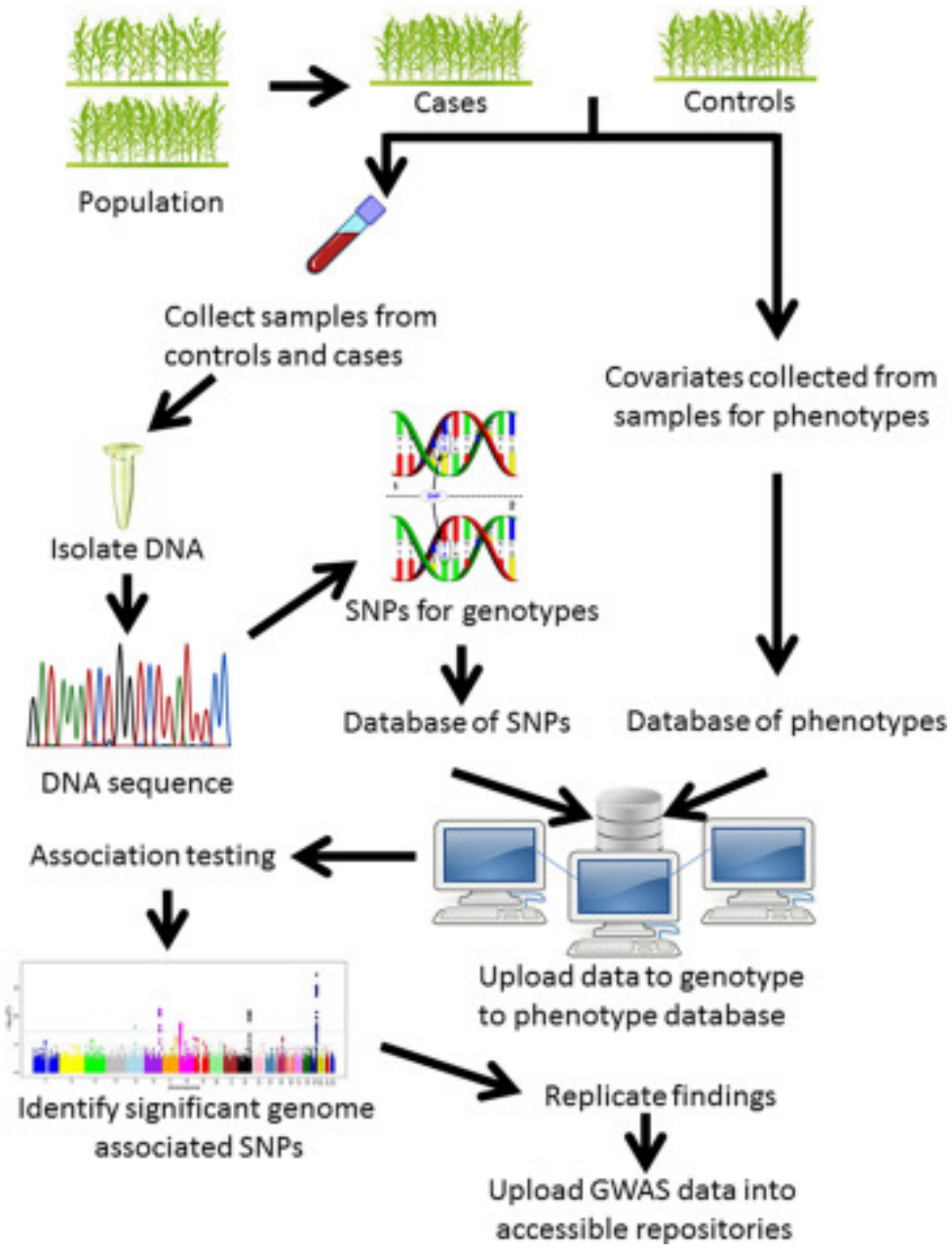
GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Since the concept was first applied in maize in 2001 (Thornsberry *et al*., **2001**), **association mapping** studies in crop species have revealed links between tens of thousands of genomic regions and various traits.

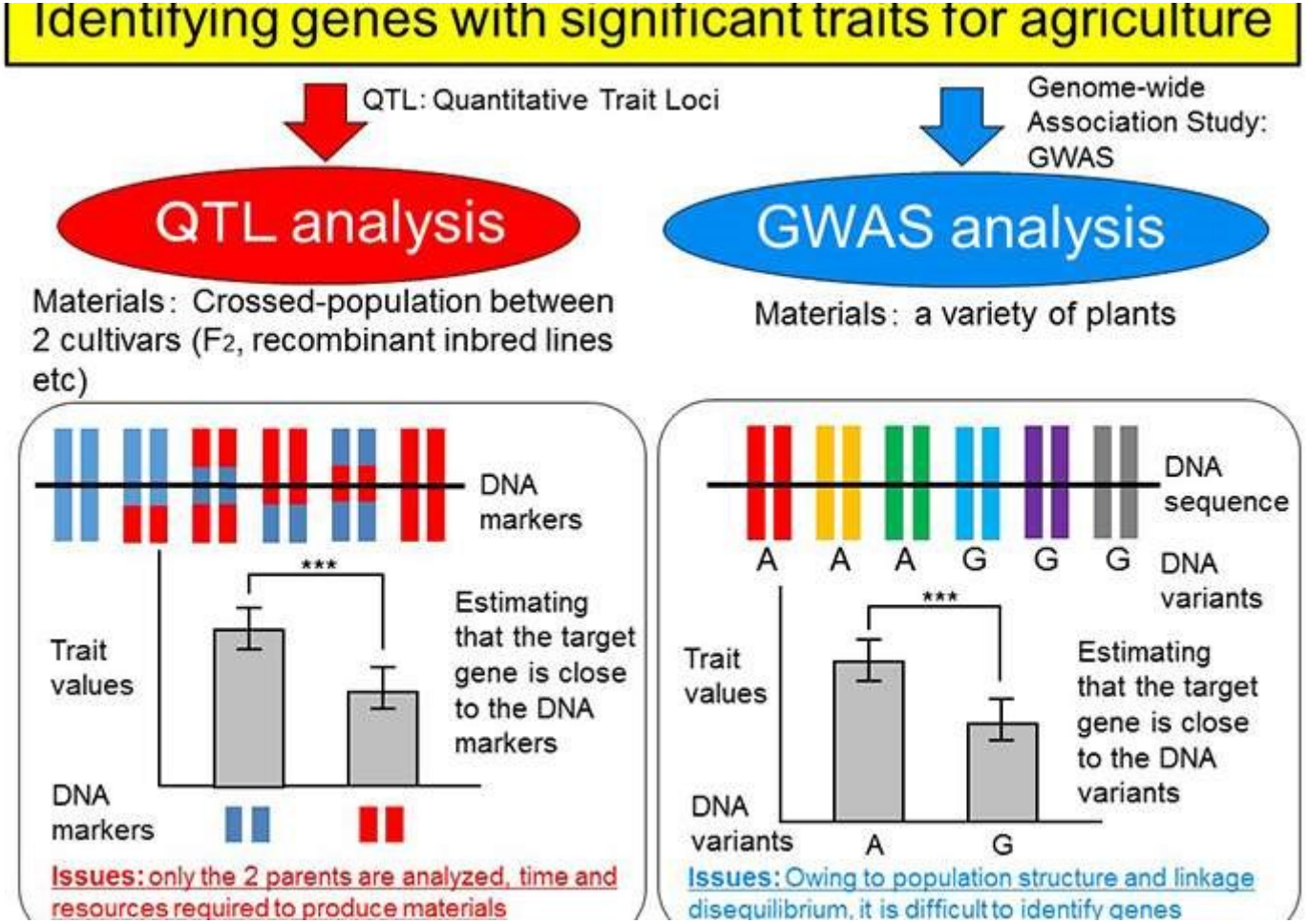
Association mapping is a quantitative approach for determining if a genomic variant is associated with a trait of interest using a natural population or a collection of diverse individuals.

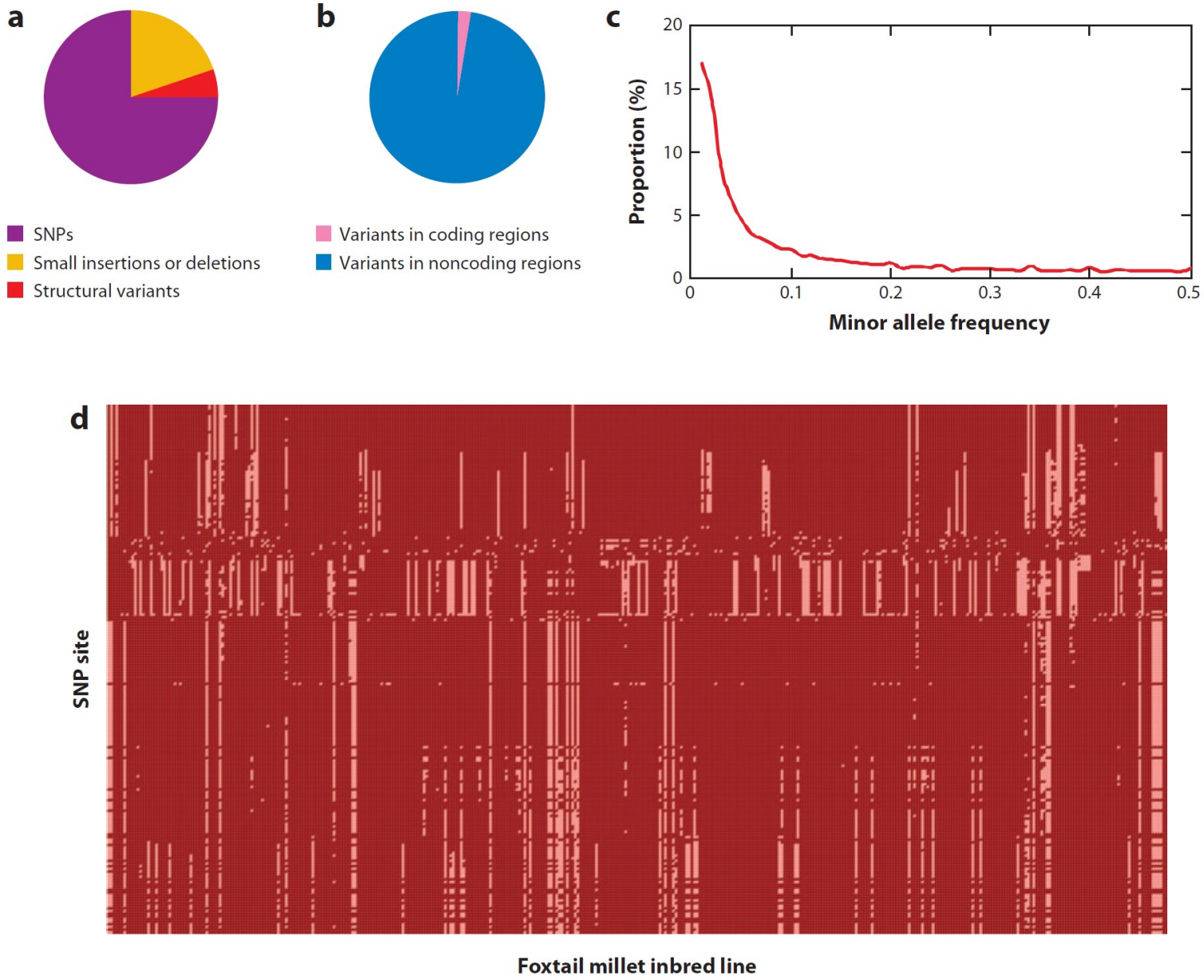
The main hypothesis states that a particular phenotype shared by a subset of individuals will be highly linked to neighboring genetic variations (**linkage disequilibrium, LD**) in their recent ancestor, where the causal mutation and corresponding phenotype arose.

Recent advances in high-throughput genotyping technologies and increases in computational power have made it possible to carry out association studies on genome-wide sets of genetic variants, an approach known as **genome-wide association study (GWAS)**


Population
Cases
Controls
Collect samples from controls and cases
Covariates collected from samples for phenotypes
Isolate DNA
SNPs for genotypes
DNA sequence
Database of SNPs
Database of phenotypes
Association testing
Upload data to genotype to phenotype database
Identify significant genome associated SNPs
Replicate findings
Upload GWAS data into accessible repositories

GWAS IN CROP PLANTS




a
SNPs
Small insertions or deletions
Structural variants
b
Variants in coding regions
Variants in noncoding regions
c
Proportion (%)
20
15
10
5
0
0
0.1
0.2
0.3
0.4
0.5
Minor allele frequency
d
SNP site
Foxtail millet inbred line

- GWAS provides an opportunity to discover genes or regions associated with given traits in a relatively high resolution and unbiased manner in broad-based and diverse populations.

- GWAS can also reveal the global landscape of a trait, known as its **genetic architecture**, a term used to describe the genetic basis of a trait based on information regarding the number of causative genes or alleles, their **interactions**, and the **distribution and patterns of their effects** (Hansen, **2006**).

GWAS IN CROPS

- A single population of different varieties (preferentially homozygous) is genotyped once and subjected to multiple phenotypic analyses (cheaper than in humans!)
- Main models: maize (high diversity – high resolution, even to single genes) and rice (small genome)
- Even in maize, tens of milions SNPs required -> elevated costs because the genome is big!
- High calculation power required
- GWAS applied with success to different species (mllet, Brassica napus, barley)

- Crop GWAS has ushered a transition to **'omics-wide association mapping** (**OWAS**), promising a better understanding of genetic architecture of complex traits.

- The large number of studies provides an unprecedented opportunity to increase in-depth understanding of the classical concepts of epistasis and pleiotropy.

- **Phenotypic plasticity** is largely ignored and requires intensive data collection and general statistical modeling.

- **Synthetic association** exists frequently in GWAS and is considered to result from the presence of multiple independent alleles within a locus.

- Emerging novel technologies such as genome editing can be used for further GWAS validation.

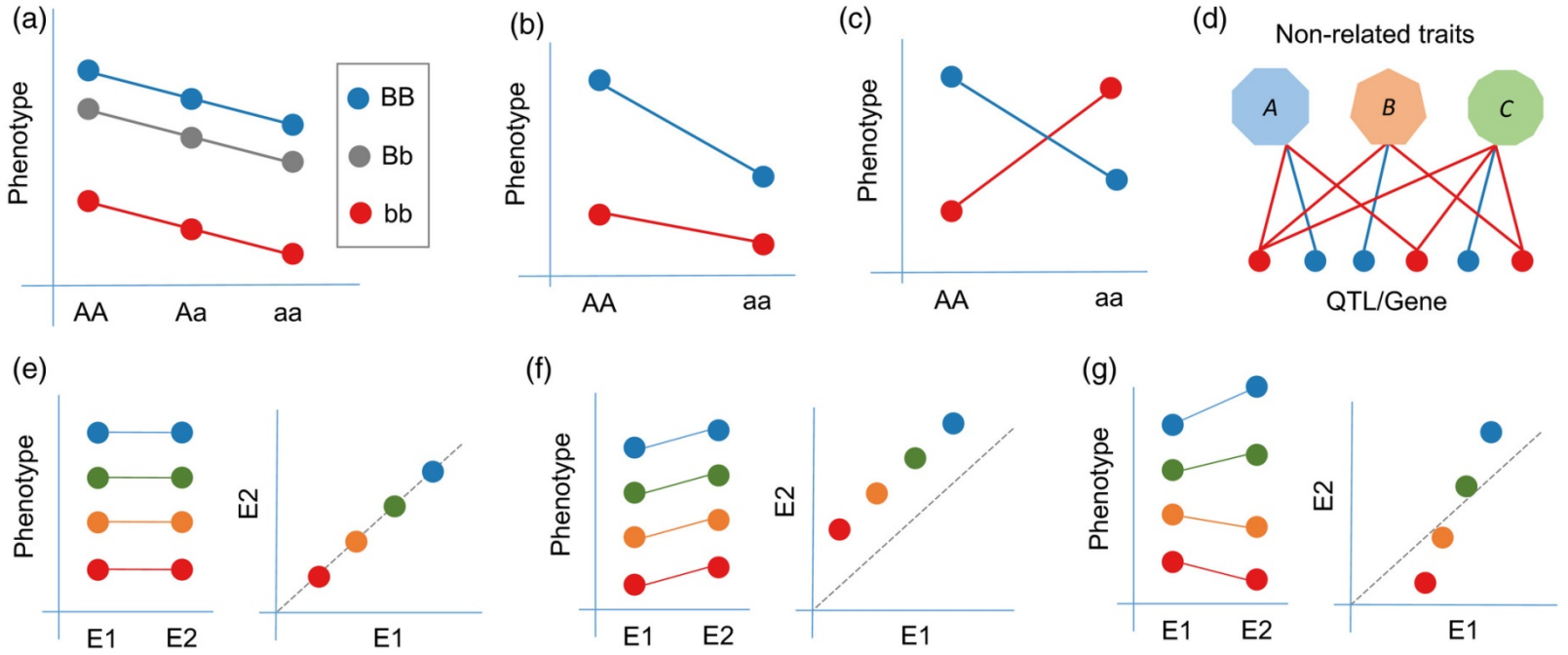
FACTORS COMPLICATING GWAS

- **Epistasis** represents a **non-linear interaction between two or more segregating loci** with different alleles across genetic backgrounds. This type of interaction between segregating loci is expected to contribute to phenotypes by biologically plausible mechanisms.

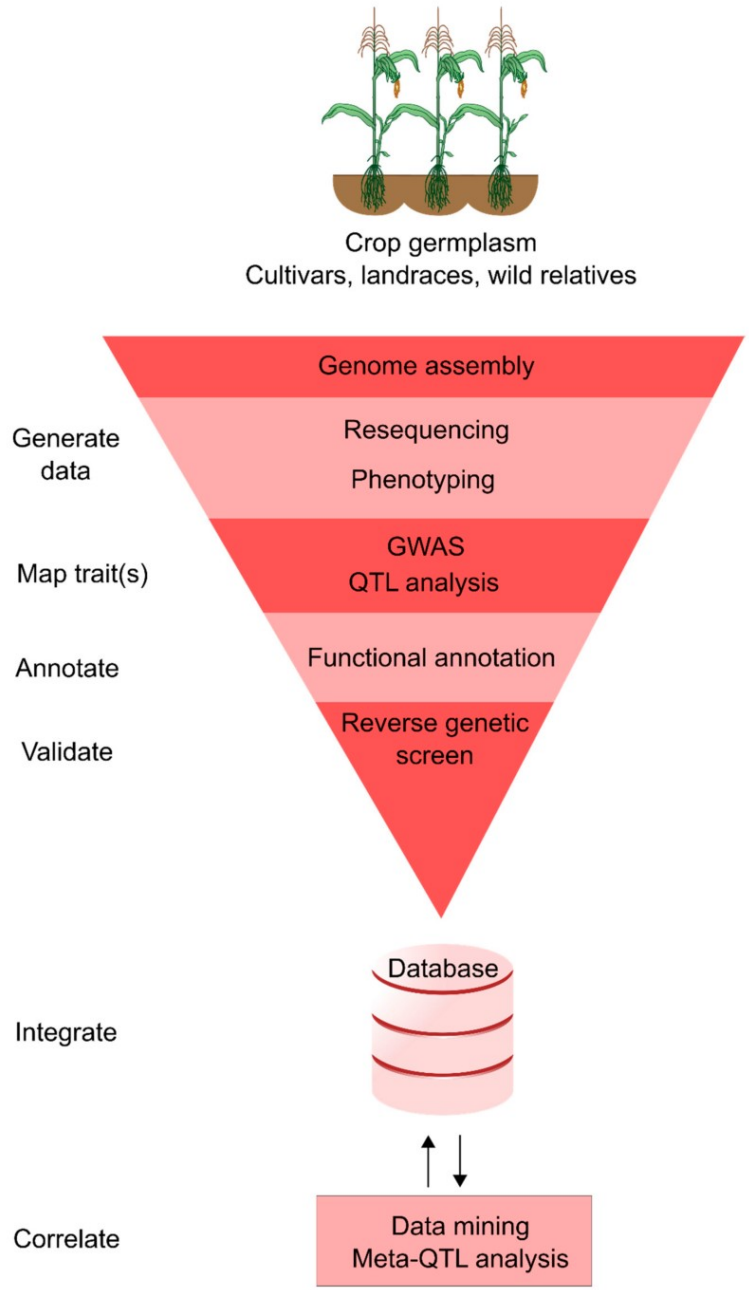
- **Synthetic association** (or ‘ghost association’) occurs when the non-causative loci show more significant associations in GWAS than the causative ones (the causative genes are located away from the GWAS peaks).

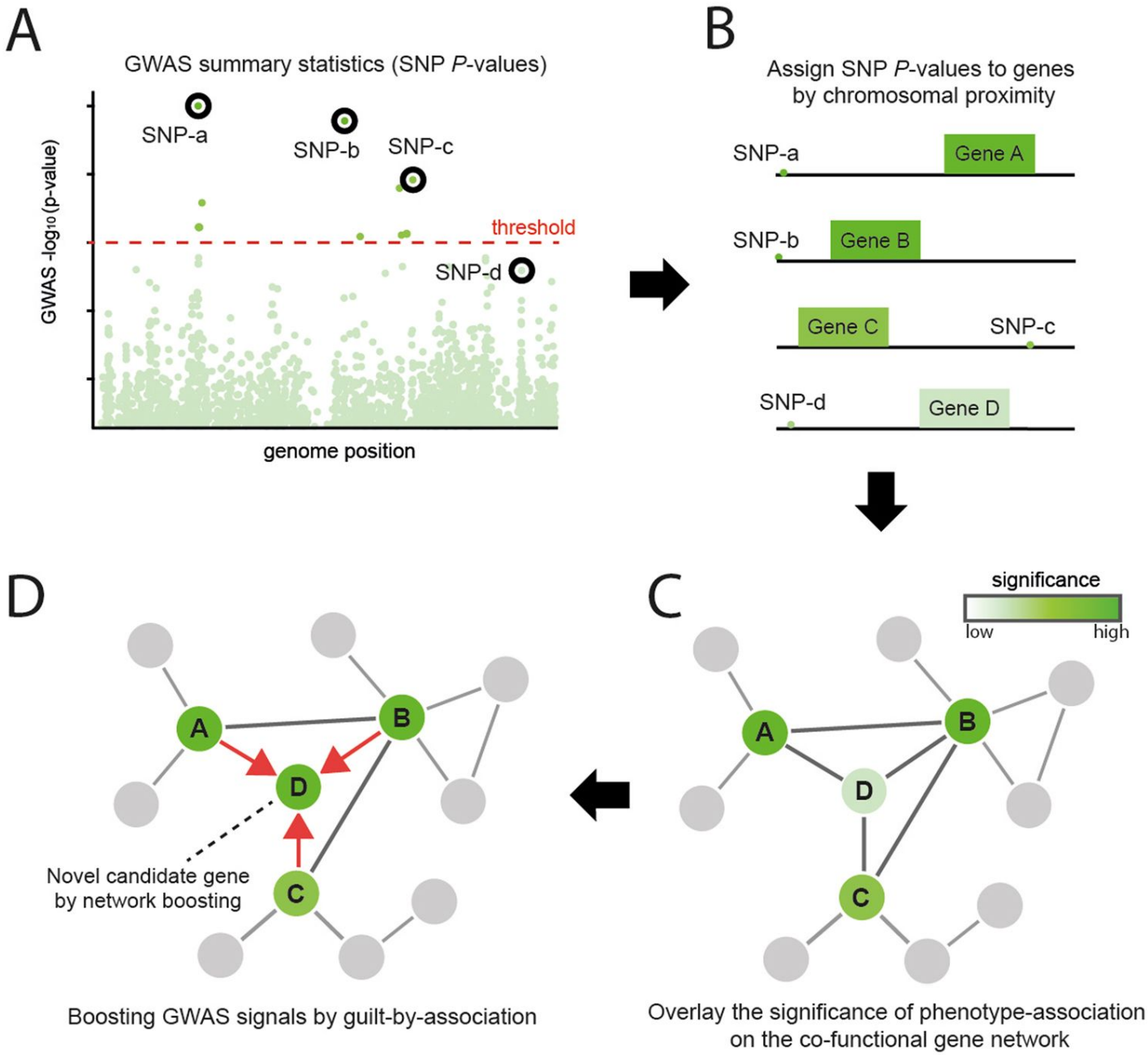
- **Pleiotropy**, in which one allele or gene affects multiple phenotypes, is crucial for understanding genetic mechanisms and for simultaneous breeding of multiple complex traits.

- **Phenotypic plasticity** is the ability to respond to environmental change by expressing variable phenotypes without genotypic change



Complex principles of genetic architecture. (a) Demonstration of additive and dominant effects for a two-locus model. Locus A only presents an **additive effect**, and dominance of locus B occurs as the phenotype of the heterozygous allele deviates from the average of the two homozygous alleles. These two loci show no **epistatic effects** with each other, as displayed in (b) and (c). (b) The different alleles of locus A show distinct effects on trait variance among different states of locus B, with the same direction. (c) The alternative alleles of locus A express similar effects on trait variance with opposite direction under different backgrounds of locus B.
(d) Presence of **pleiotropy** in red quantitative trait loci (QTLs) or genes as these show effects on at least two non-correlated traits; blue QTLs or genes represent non-pleiotropic loci as they only contribute to one trait. (e) Absence of plasticity. No phenotypic difference exists under different environments (E1 and E2); each colored point represents a different genotype. (f) Presence of **phenotypic plasticity** without existence of a **genotype–environment interaction (G × E**), as all genotypes alter their phenotypes in parallel under different environments. (g) Co-existence of phenotypic plasticity and G × E, as all genotypes alter their phenotypes but to distinct extents or/and in distinct directions under different environments.


Crop germplasm
Cultivars, landraces, wild relatives
Genome assembly
Generate data
Resequencing
Phenotyping
Map trait(s)
GWAS
QTL analysis
Annotate
Functional annotation
Validate
Reverse genetic screen
Integrate
Database
Correlate
Data mining
Meta-QTL analysis


A
GWAS summary statistics (SNP P-values)
SNP-a
SNP-b
SNP-c
SNP-d
threshold
GWAS -log10 (p-value)
genome position
B
Assign SNP P-values to genes
by chromosomal proximity
SNP-a
Gene A
SNP-b
Gene B
Gene C
SNP-c
SNP-d
Gene D
C
significance
low
high
A
B
D
C
Overlay the significance of phenotype-association
on the co-functional gene network
D
A
B
D
C
Novel candidate gene
by network boosting
Boosting GWAS signals by guilt-by-association

Current status of molecular breeding

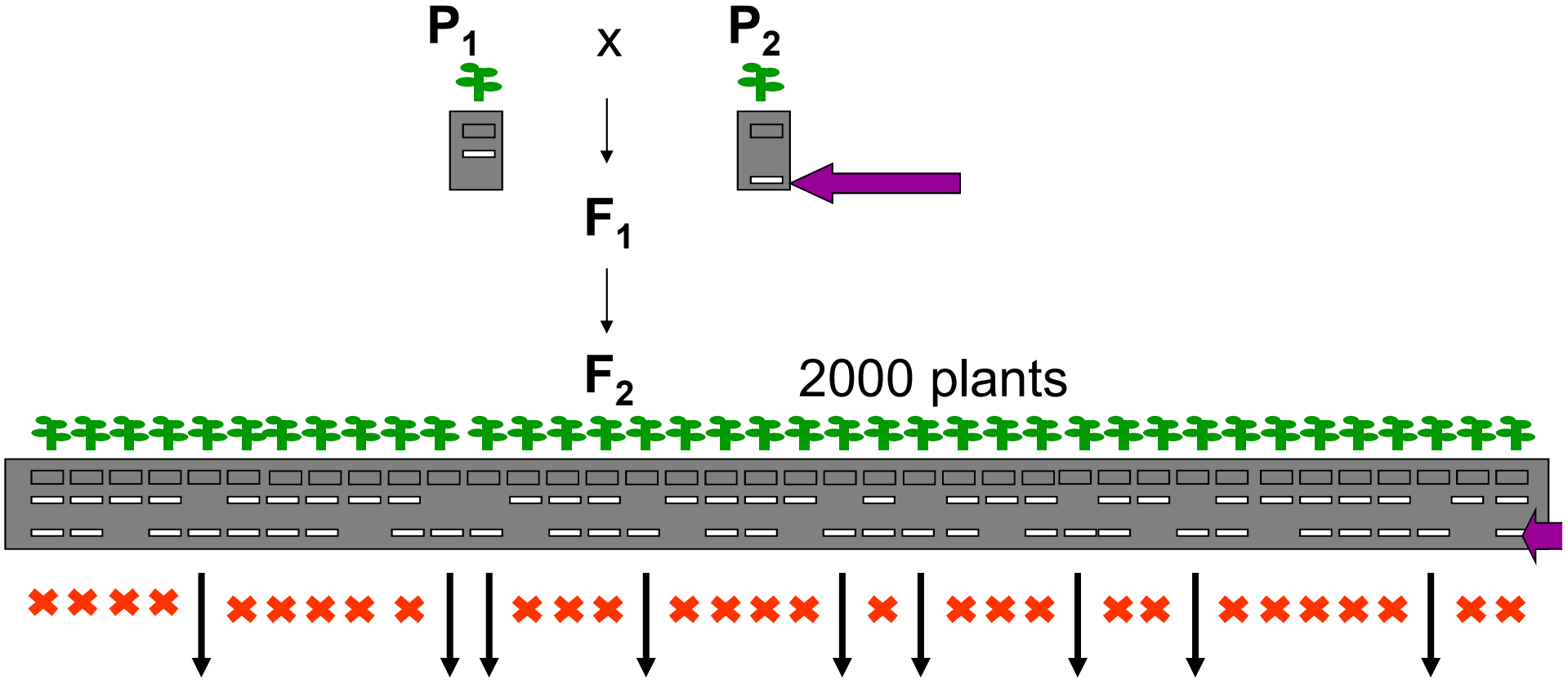
- A literature review indicates thousands of QTL mapping studies but not many *actual* reports of the application of MAS in breeding

- *Why is this the case?*

Some possible reasons to explain the low impact of MAS in crop improvement

- Resources (equipment) not available
- Markers may *not* be cost-effective
- Accuracy of QTL mapping studies
- QTL effects may depend on genetic background or be influenced by environmental conditions
- Lack of marker polymorphism in breeding material
- Poor integration of molecular genetics and conventional breeding

Cost of MAS in context: Example 1: Early generation MAS



USD **$640** to screen 2000 plants with a single marker for one population