Marker-assisted selection/breeding and its potential use for sweetpotato improvement

Awais Khan

awais.khan@cgiar.org

International Potato Center (CIP), Lima, Peru

Plant breeding

Systematic procedure for genetic improvement through **crossing** plants with desired traits and **selecting** progeny with improved performance and/or improved combinations of traits.

Accelerated and targeted evolution

Crop Improvement: Critical considerations

Data recording and management

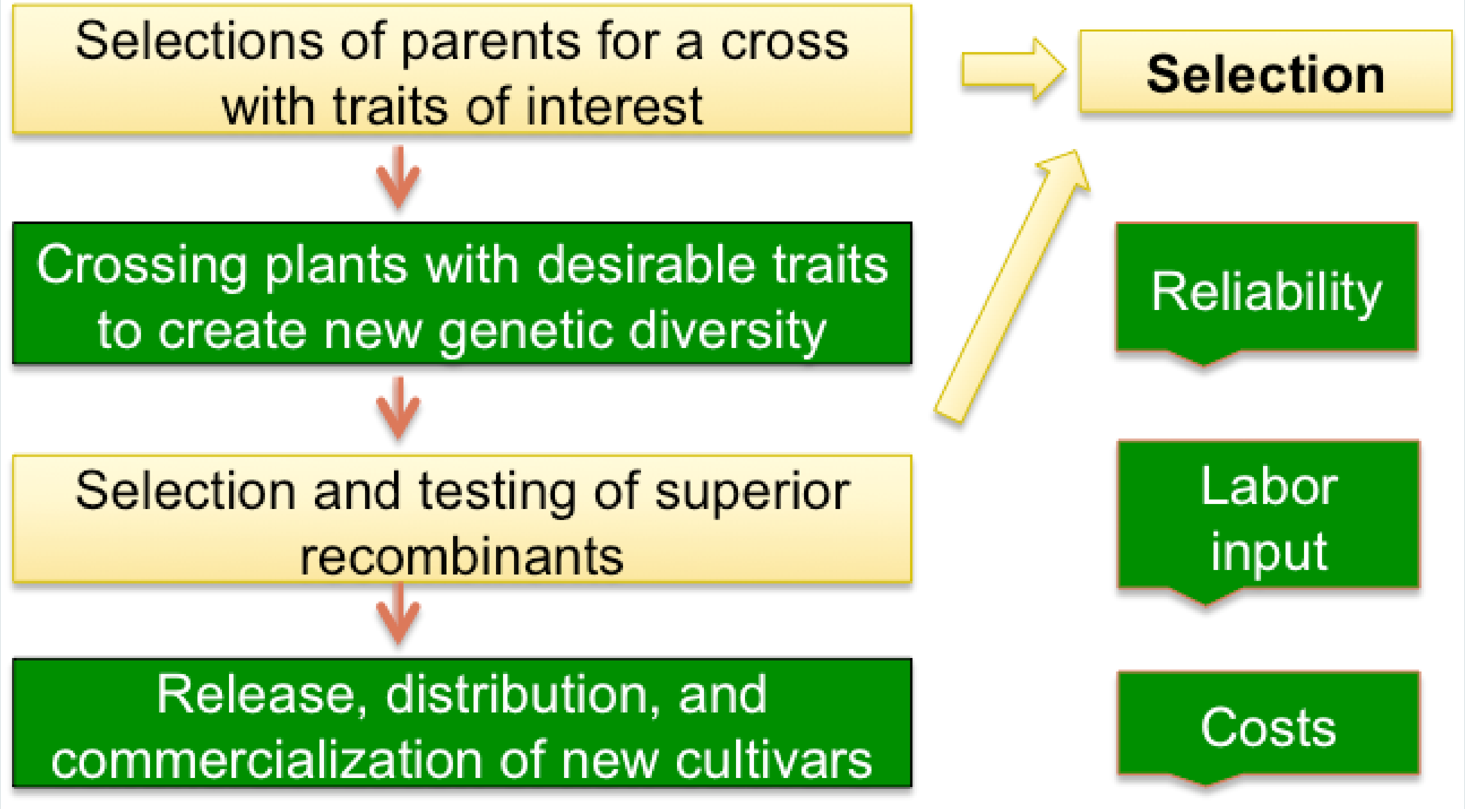
Phenotyping with special attention to Genotype x environment interaction: appearance and performance in general, in particular, response to environment

Genomics for identification of genetic basis of traits of importance

Selection of parents and progeny with desirable traits

Challenges of plant breeding

A challenge for modern breeding – to develop and integrate phenotypic and genotypic information to understand and improve traits of interest


Selections of parents for a cross with traits of interest

Crossing plants with desirable traits to create new genetic diversity

Selection and testing of superior recombinants

Release, distribution, and commercialization of new cultivars

Selection

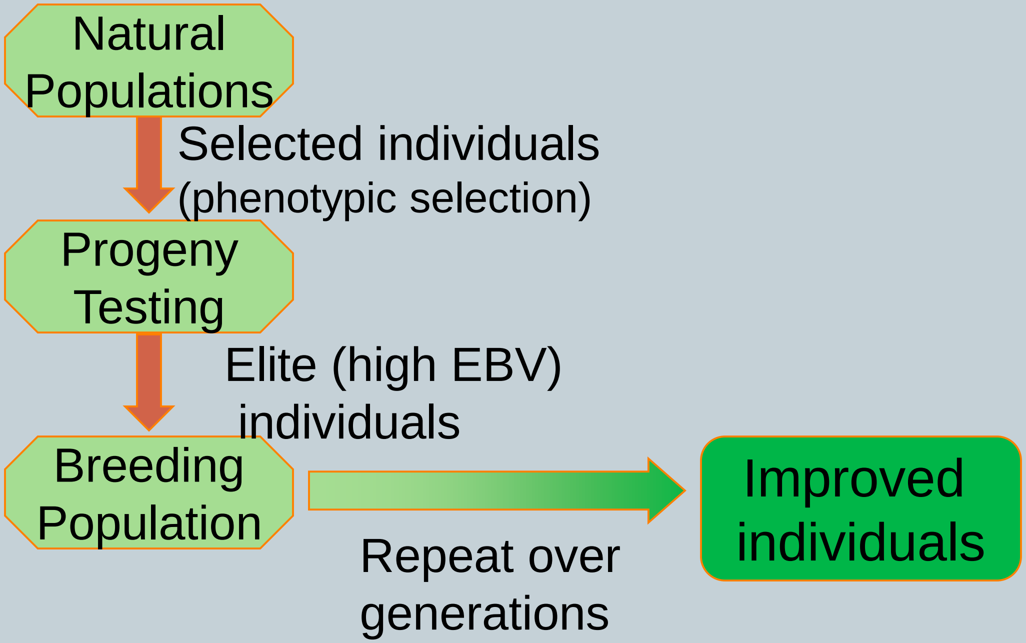
Reliability

Labor input

Costs

General steps in plant breeding (modified after Gepts 2002)

Phenotypic selection: Selection based on appearance and performance



Natural Populations

Selected individuals (phenotypic selection)

Progeny Testing

Elite (high EBV) individuals

Breeding Population

Repeat over generations

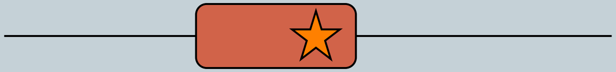
Improved individuals

I. Difficult to separate environmental & genetic contribution
II. Difficult to distinguish homozygous & heterozygous effects
III. Needs large space & labor input
IV. Slow & time consuming

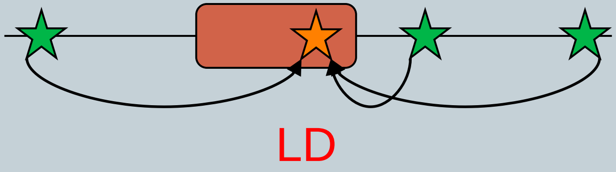
Concept of Marker assisted selection

Molecular breeding

Association between molecular marker and causative gene



Direct association



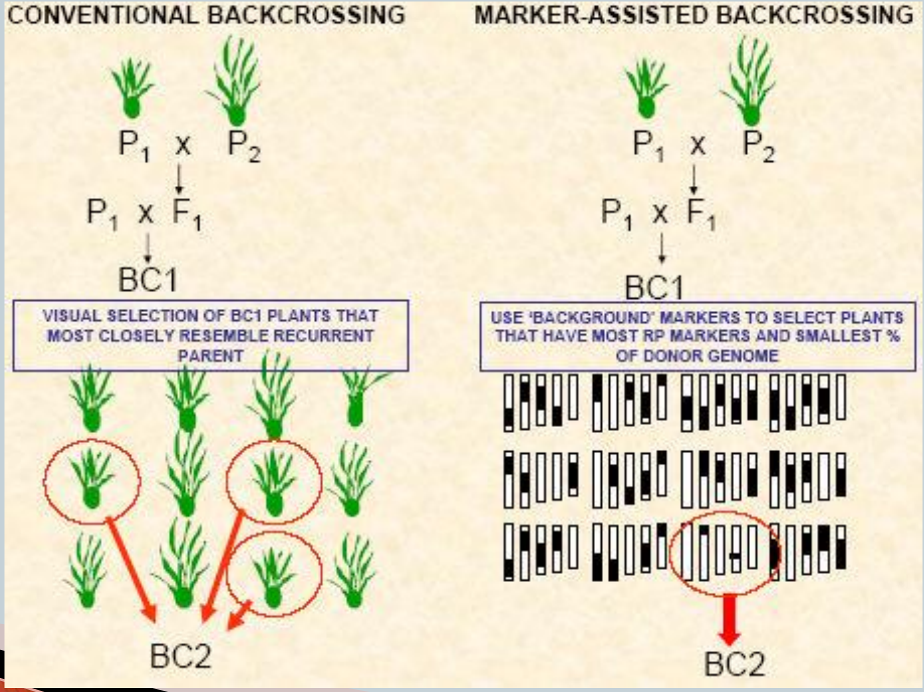
Indirect association

🟥 Causative gene

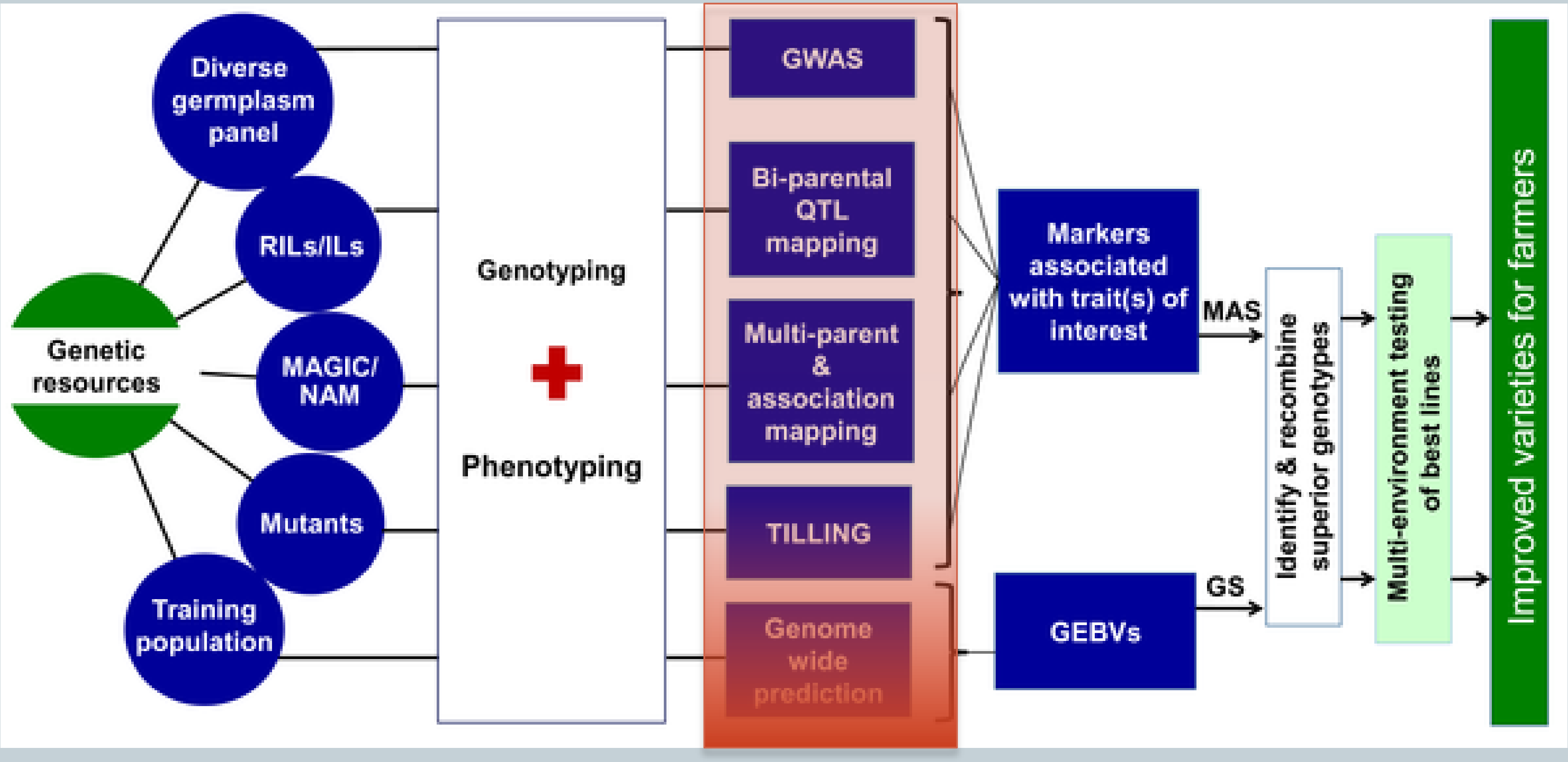
★ SNP within gene

★ SNP in LD with gene

Hirschhorn & Daly, 2005



Marker-trait association identification

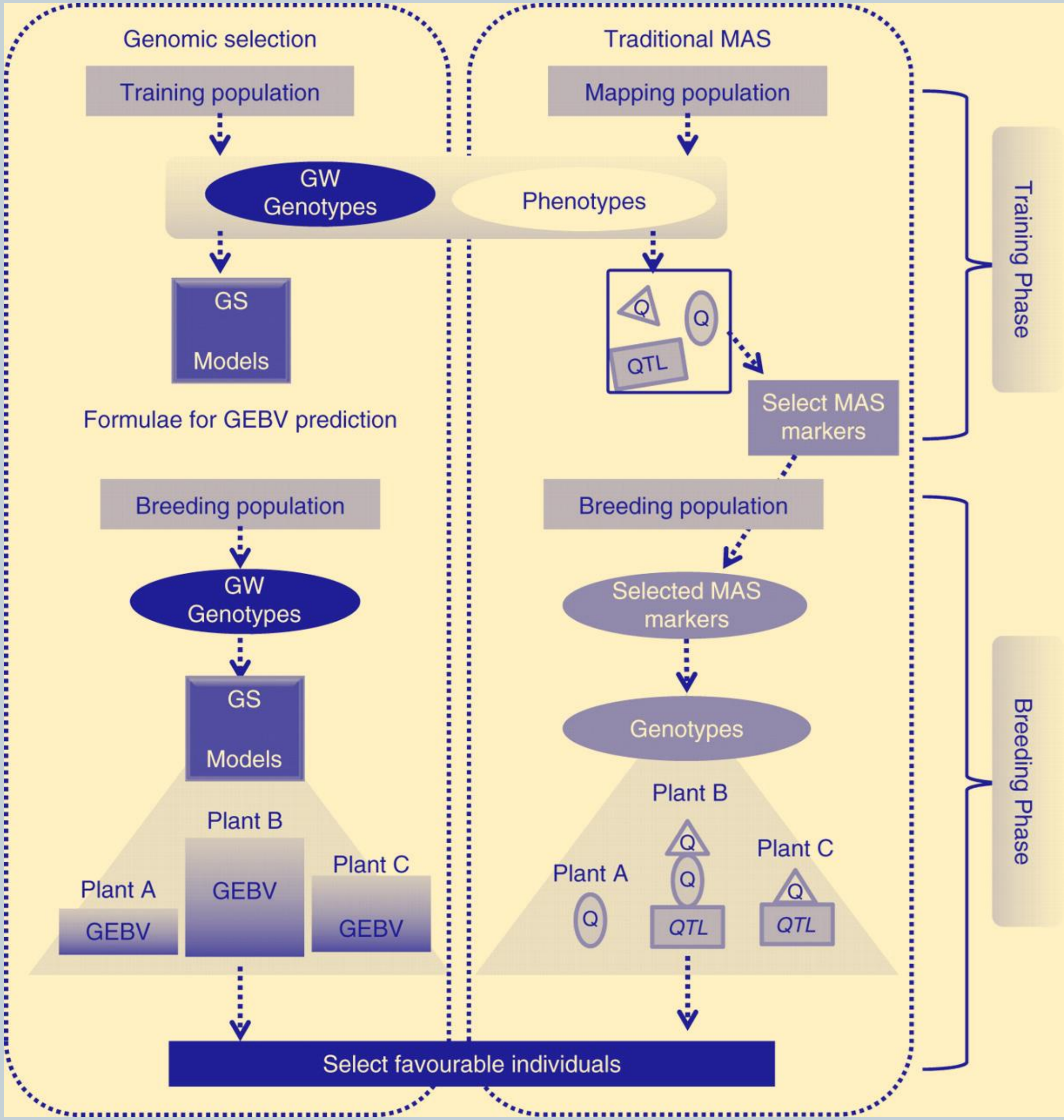


Varshney *et al*. 2014

Predicting the phenotype or selection of progeny with desirable traits

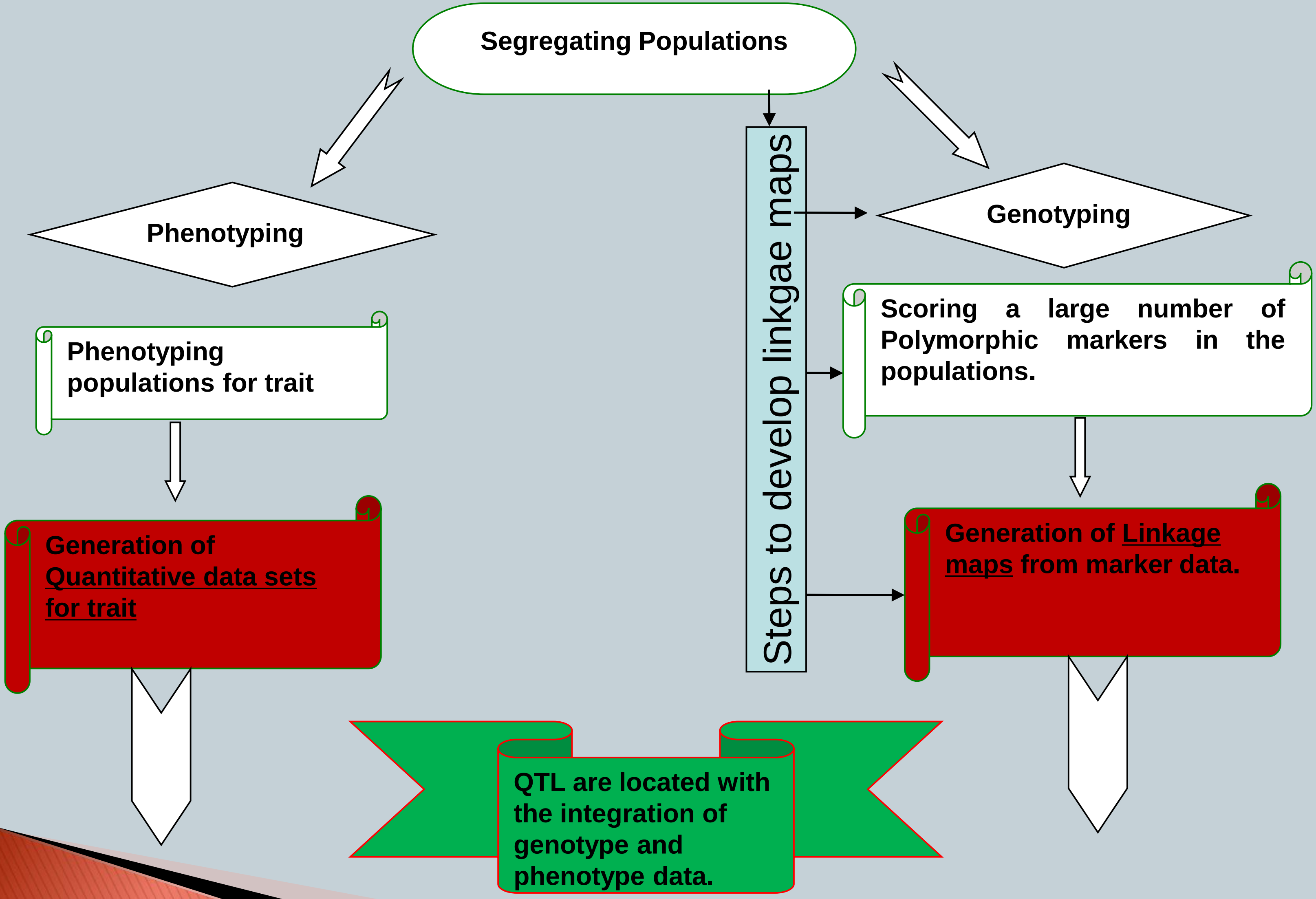
Marker-assisted Selection (MAS): Usually, plants are selected for up to 10 alleles

Genomic Selection (GS): Selection of several loci genome-wide linked to traits of interest using Genomic Estimated Breeding Values (GEBVs) based on genome-wide markers



Nakaya and Isobe, 2012

Overview: Linkage map and QTL mapping



Segregating Populations

Steps to develop linkgae maps

Phenotyping

Phenotyping populations for trait

Generation of Quantitative data sets for trait

Genotyping

Scoring a large number of Polymorphic markers in the populations.

Generation of Linkage maps from marker data.

QTL are located with the integration of genotype and phenotype data.

Genetic (linkage) mapping

Determining the location of elements (genes) within a genome, with respect to identifiable landmarks (molecular markers)

Three key concepts to understand genetic mapping

Linkage

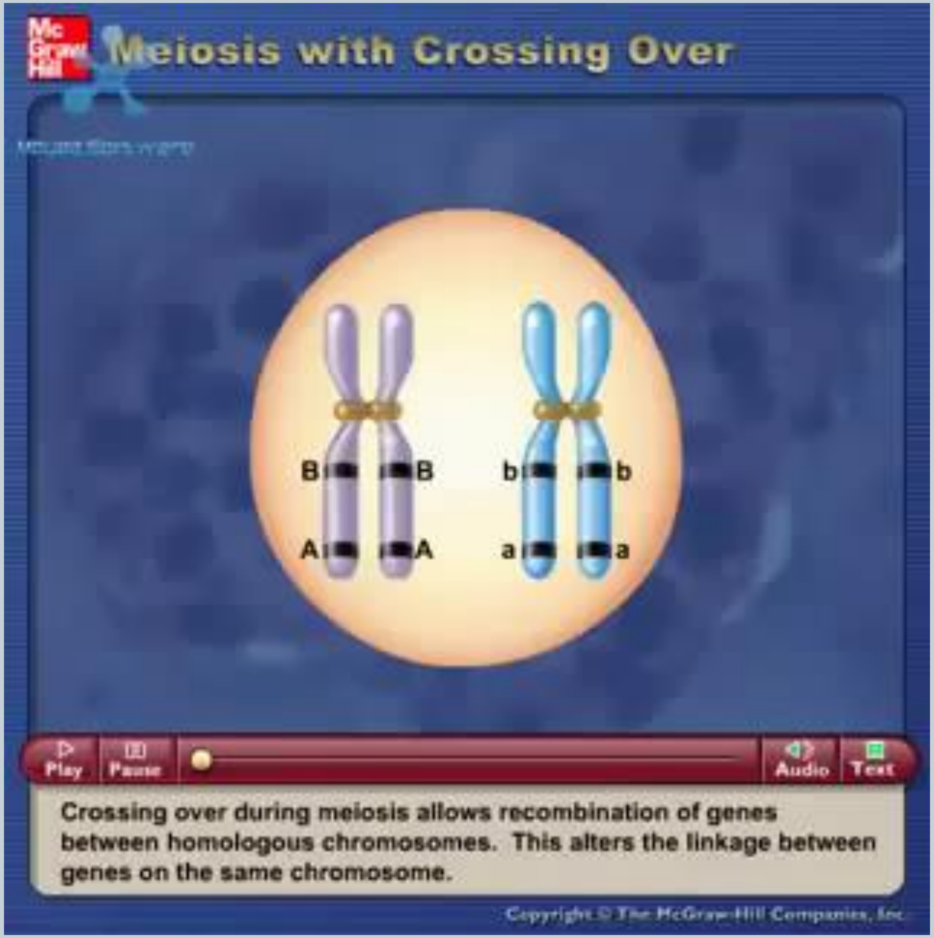
Crossing over

Recombination

Linkage, crossing over and recombination

Mendel's Law of Independent Assortment applies well to genes that are on different chromosomes.

But!
loci of two genes are close enough together on the same chromosome = “linked” and they tend to segregate together in crosses.



Meiosis with Crossing Over

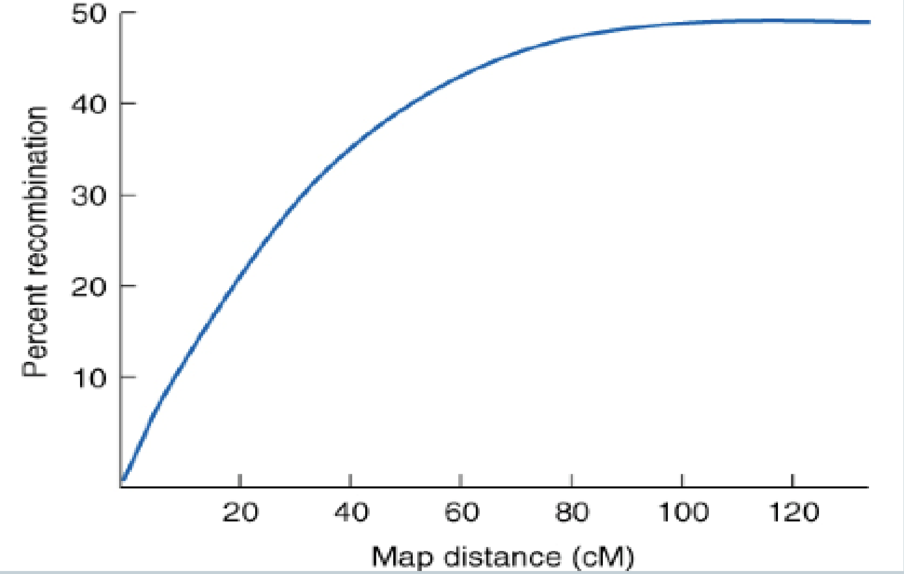
Crossing over during meiosis allows recombination of genes between homologous chromosomes. This alters the linkage between genes on the same chromosome.

Crossing over recombination and map distance

Two types of gametes are possible

Parental gametes= If crossing over does not occur
Recombinant gametes= If crossing over occurs

The probability that crossing over will lead to the separation of two genes on a chromosome is proportional to the distance between them



Construction of a genetic map

Frequency of recombinants in the progeny helps to estimate the distance between markers.

A genetic map

Is an ordered list of the genetic loci along a particular chromosome

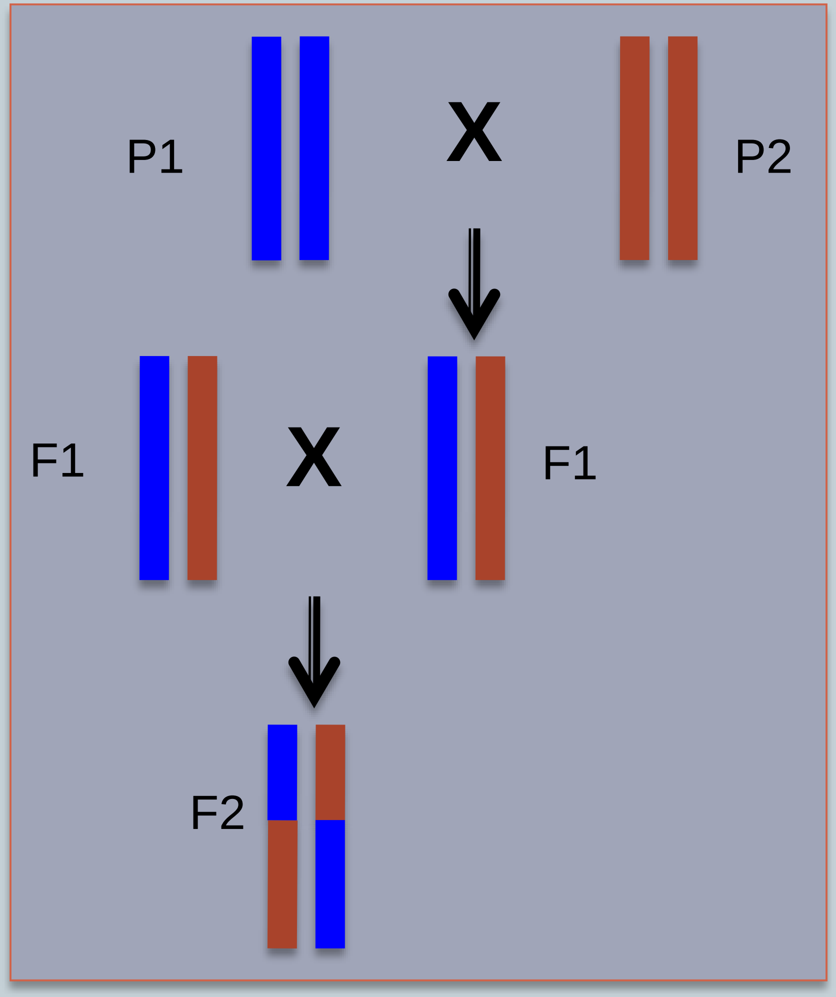
Unit: Morgan/centiMorgan also known as the genetic map distance: *d*

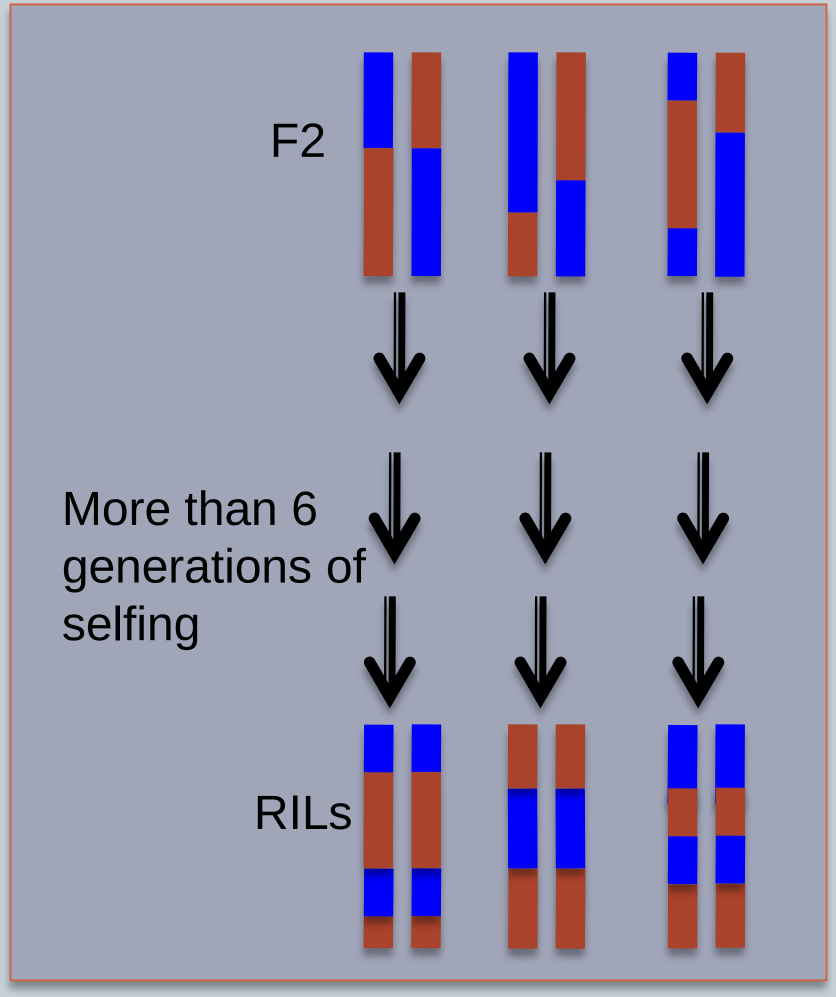
Requirements

Segregating population: A population with known relationships, segregating for the traits of interest

Genetic markers: A variation which may arise due to mutation or alteration in the genomic loci that can be used to identify individuals

Segregating population





More than 6 generations of selfing

Molecular markers

DNA Based markers

- Randomly Amplified Polymorphic DNA (RAPDs)
- Amplified Fragment Length Polymorphism (AFLPs)
- Microsatellites/Simple Sequence Repeats (SSRs)
- Single Nucleotide Polymorphism (SNPs)

Morphological markers

Isozymes

Properties of good marker

Polymorphic

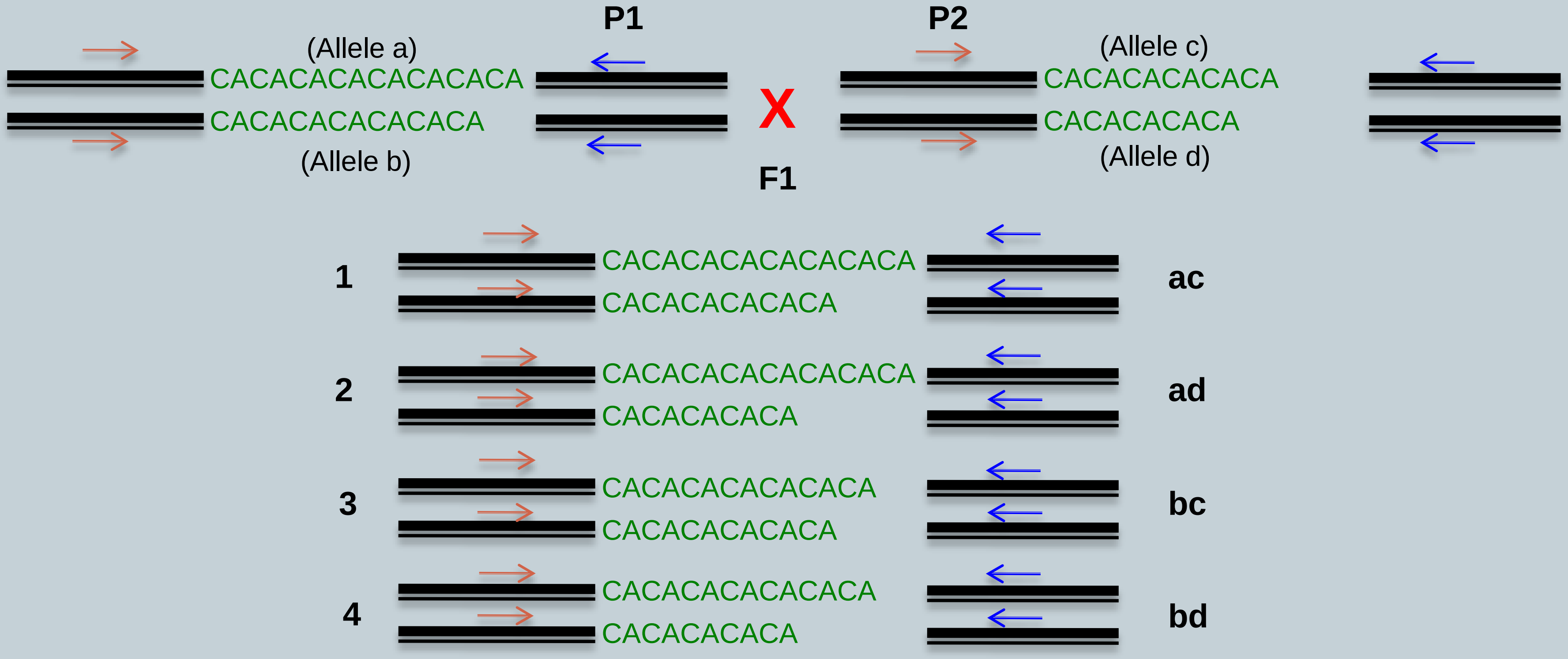
Reproducible

Easy to use

Economic

High-throughput

Simple Sequence Repeats (SSRs)



P1

(Allele a)
CACACACACACACACA
CACACACACACACA
(Allele b)

X

P2

(Allele c)
CACACACACACA
CACACACACA
(Allele d)

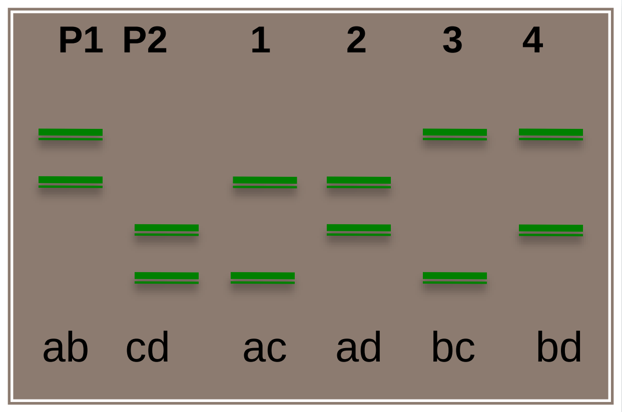
F1

1
CACACACACACACACA
CACACACACACA
ac

2
CACACACACACACACA
CACACACACA
ad

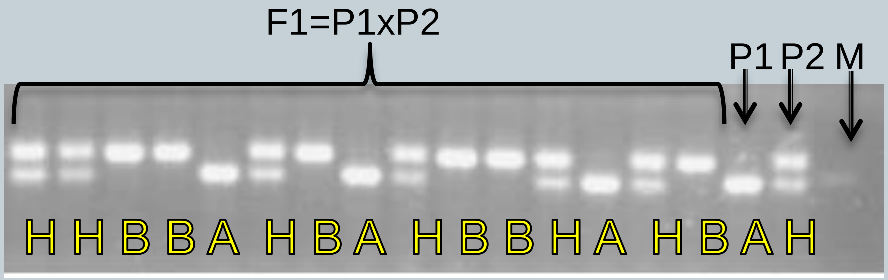
3
CACACACACACACA
CACACACACACA
bc

4
CACACACACACACA
CACACACACA
bd



Genotyping

- A large number of genetic markers are tested on a segregating population (genotyping)
- Data file with marker score for each individual of the population
- Software is used to estimate recombination frequency of each marker and order the markers along a linkage group (chromosome)



Marker data

Key:

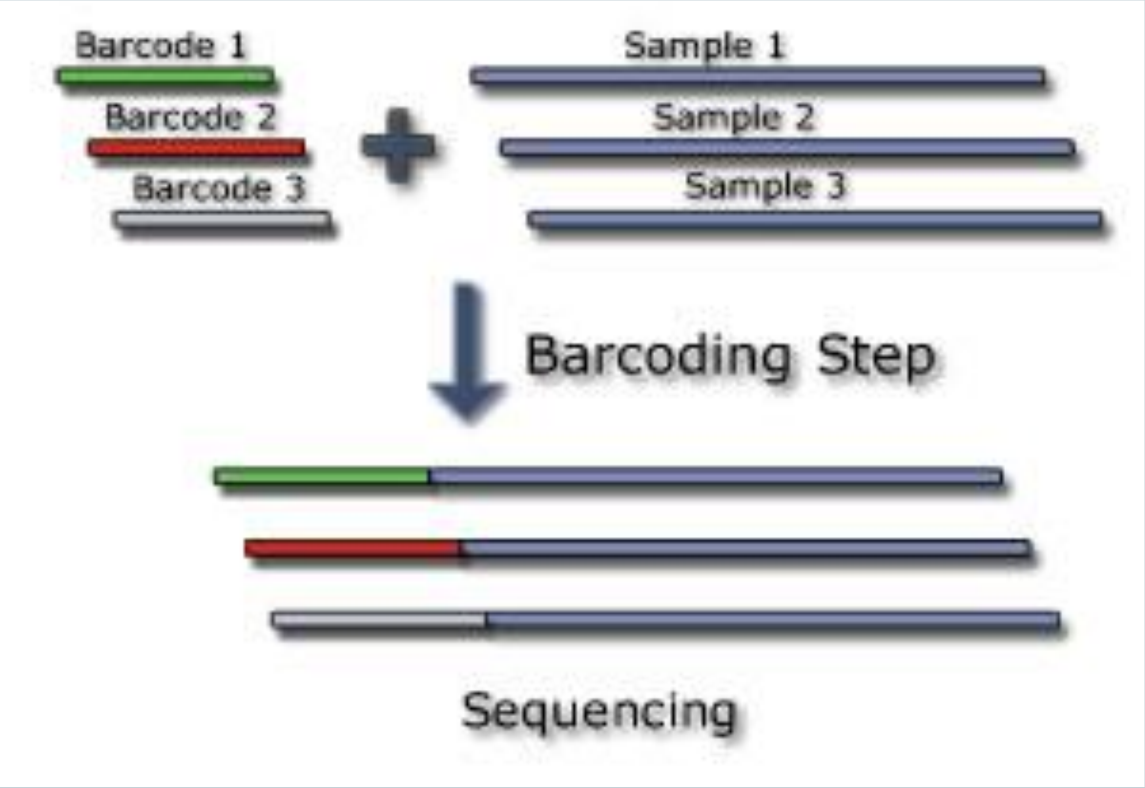
A=Homozygous for allele P1

B=Homozygous for allele P2

H=Heterozygous

M=Ladder

Multiple samples per sequencing run



Barcodes are incorporated into the adaptors that are ligated to the DNA after shearing. It enable the pooling of several different samples in one library.

Let’s construct a genetic map

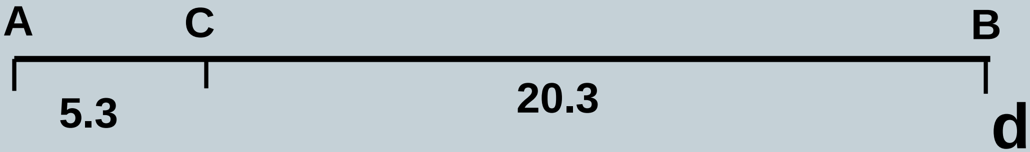
Recombination frequency (RF) =

number of recombinants/total number of individuals*100

RF of 1 %= Genetic map distance of 1 cM

Gametes	**# of gametes**
ABC	1080
abc	1071
AbC	293
aBc	282
aBC	78
Abc	66
ABc	6
abC	4
Total	2880

A - B	**B - C**	**A – C**
293	293	78
282	282	66
78	6	6
66	4	4
719	**585**	**154**
0.250	**0.203**	**0.053**

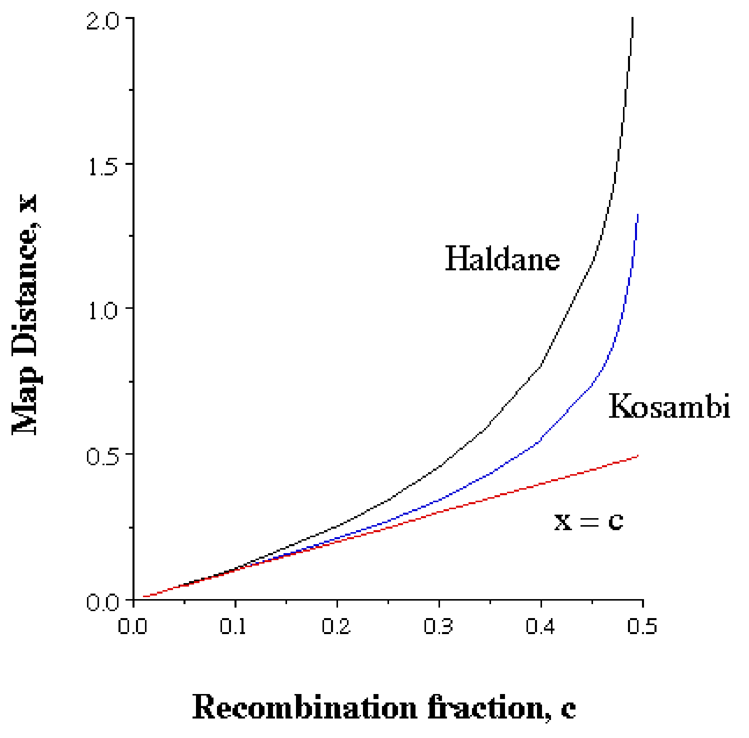


distance in map units

Mapping functions

Interference: Does a crossover in one region affect the likelihood of a crossover in an adjacent region

- **Haldane:** There is no interference due to crossing over and crossing over occurs randomly and independently.
- **Kosambi:** Constant and specific level of interference and a small correction for interference.



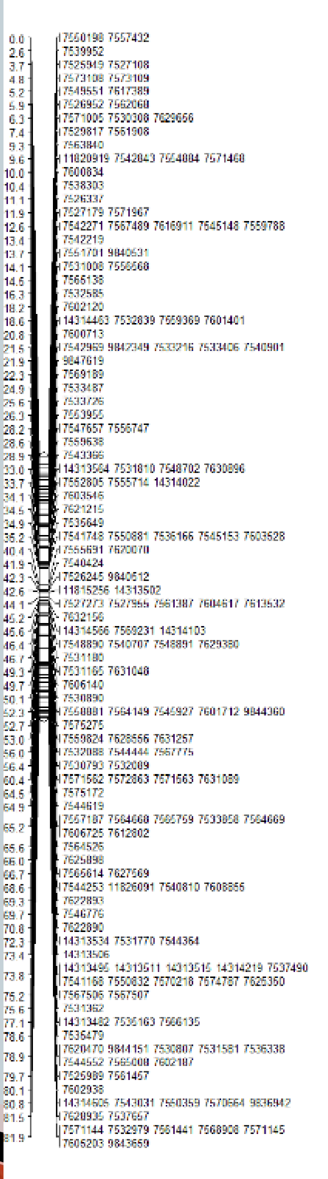
Linkage map

Dense genetic map for B x T sweetpotato based on DArTseq markers

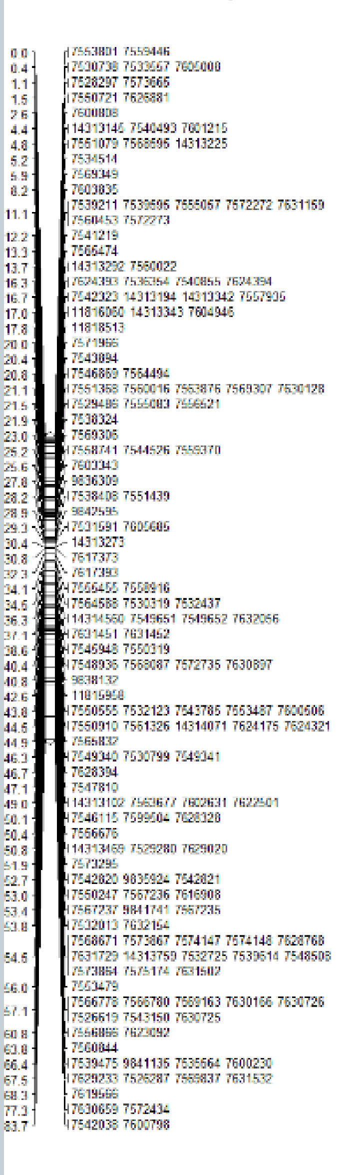
First report: 33086 SNPs were identified

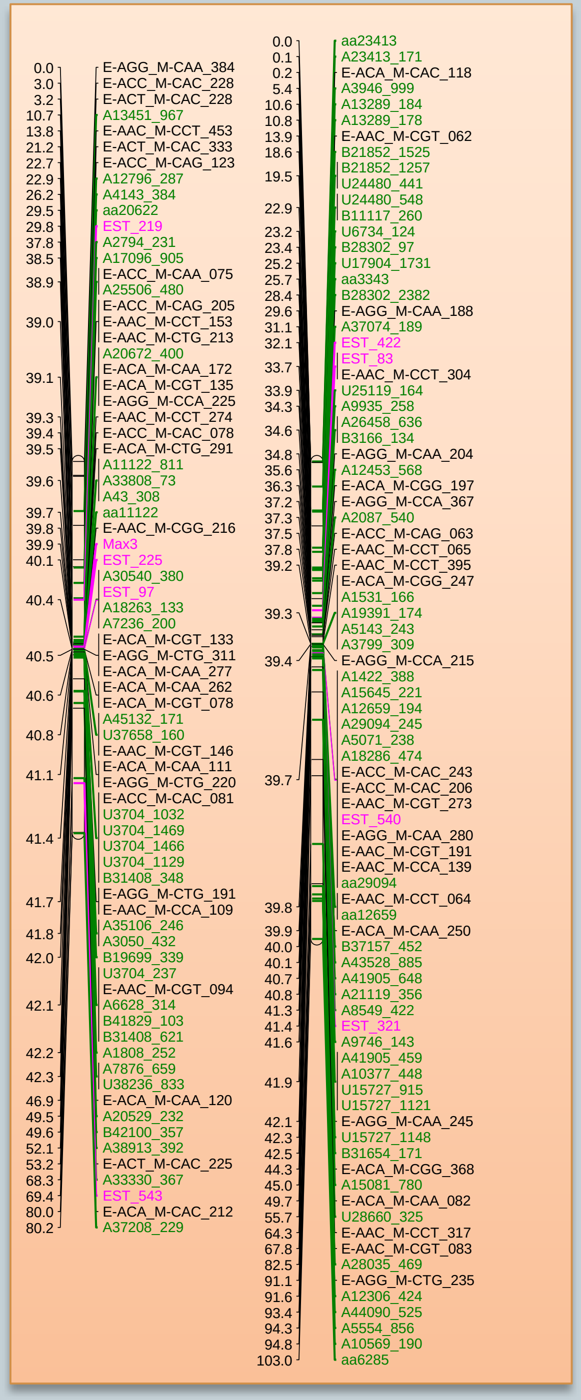
11091 SNPs were mapped

LG02



LG03





Quantitative trait locus (QTL) mapping

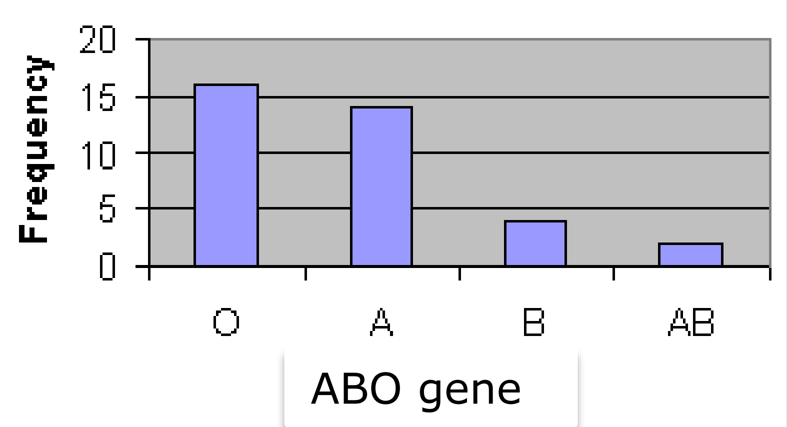
Quantitative trait locus (QTL): A genomic region that is associated with a quantitative trait

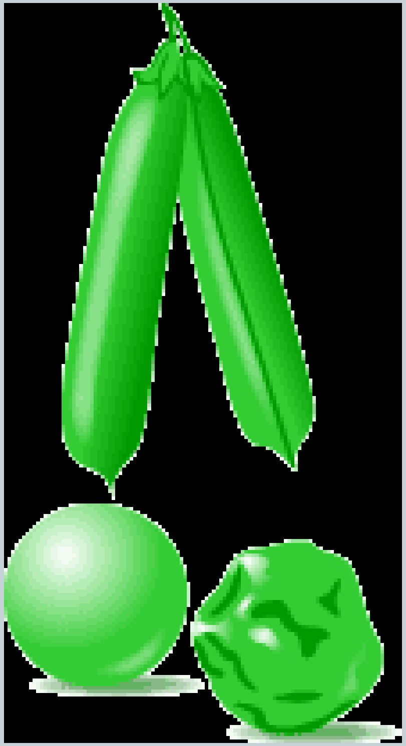
Phenotypic trait

- Qualitative trait
- Quantitative trait

Qualitative trait

Fall into discrete classes, controlled by two or many alleles of single gene and less influenced by environment e.g., blood type, seed coat color, many diseases



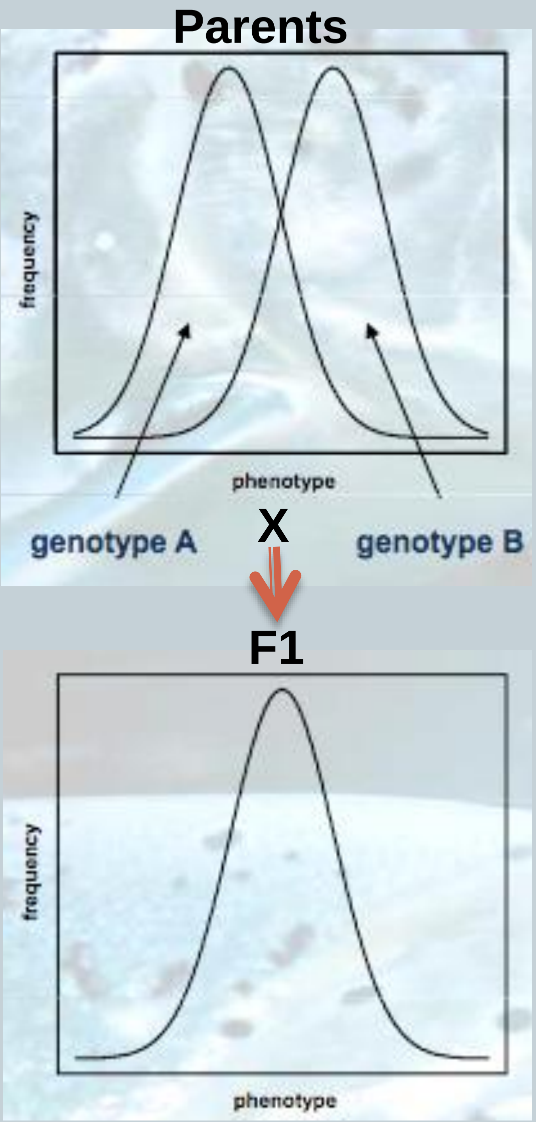


The wrinkled-seed character of pea is caused by a transposon-like insertion in a gene

Quantitative trait

The quantitative trait has continuous variation (bell-shaped curve, normal distribution) and is usually controlled by many genes of small effect, or by a few genes of large effect e.g., Height, Weight, Biomass, Disease resistance

But
A single polymorphic locus with multiple, differentially expressed alleles can also result in continuous variation



Variance Components

$$V_P = V_G + V_E + V_{GE}$$

$_P$ = phenotypic, $_G$ = genetic, $_E$ = environmental

V_{GE} = variation associated with the genetic and environmental interactions

V_G (The total genetic variation)

$$V_G = V_A + V_D + V_I$$

A=additive, D=dominance, I= interaction due to epistatis

Additive genetic variance (V_A): Each allele has a specific value that it contributes to the final phenotype

Dominance genetic variance (V_D): Dominant gene action masks the contribution of the recessive alleles at the locus

Example

AABB X aabb **Parents**

↓

AaBb **F1**

A=4 U, a=2 U, B=6 U, b=3 U

Additive effect
F1= 15 U (4+2+6+3)

Dominant effect
F1=20 U (4+4+6+6)

Variance Components

Interaction genetic variance (V_I)/epistasis:
Due to masking of genotypic effects at one locus by genotypes of another locus

Environmental variance (V_E)
Due to difference in magnitude of performance of genotypes in different environments

Genotype-Environment interaction (V_{GE})
Due to difference in the direction of performance of genotypes in different environmental circumstances

The total phenotypic variance can be rewritten as

$$\mathbf{V_P = V_A + V_D + V_I + V_E + V_{GE}}$$

Heritability

The proportion of the genetic variance to the total variance

Broad-sense heritability: Ratio of total genetic variance to total phenotypic variance

H2 = VG/VP

Narrow-sense heritability: Ratio of additive genetic variance to total phenotypic variance

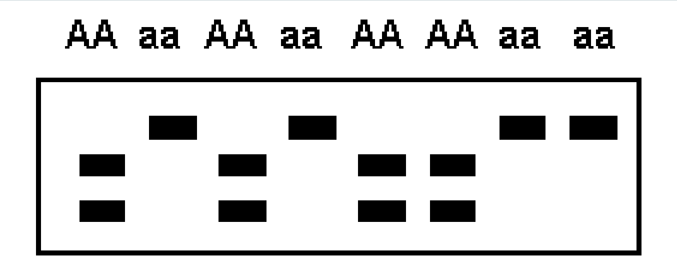
h2 = VA/VP

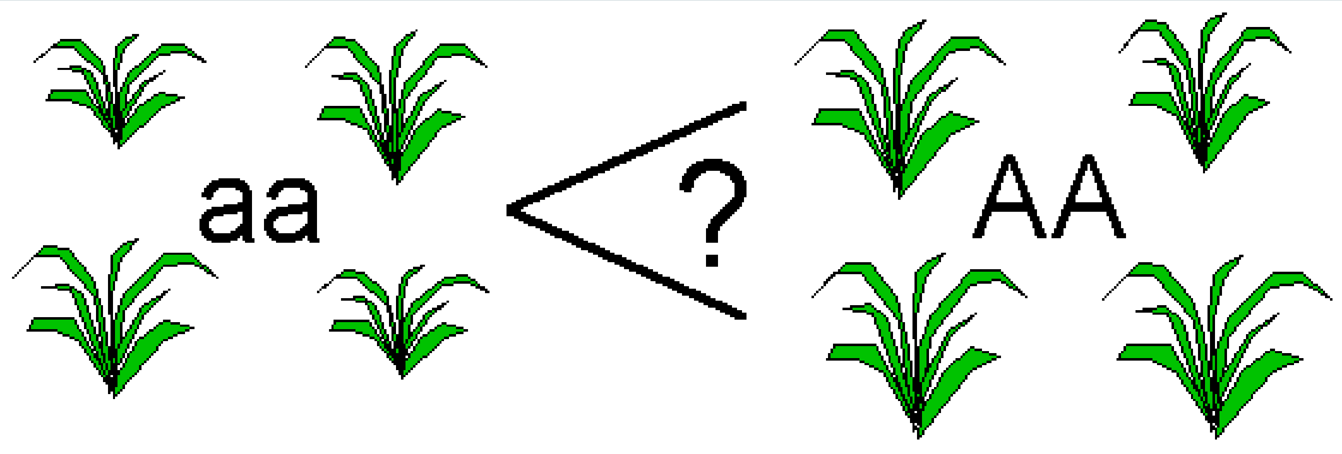
- Specific to the population and environment
- Does not indicate the degree to which a trait is genetic, it measures the proportion of the phenotypic variance that is the result of genetic factors

QTL analysis

Is there an association between marker genotype and quantitative trait phenotype?

Marker genotype





QTL analysis

- **QTL Detection and LOD threshold**
- **QTL Localization**

QTL Detection

QTL effect: The average difference in the phenotype of the trait between marker allele genotypes

Homozygous effects: The difference in the mean of the trait between the two homozygous genotypes

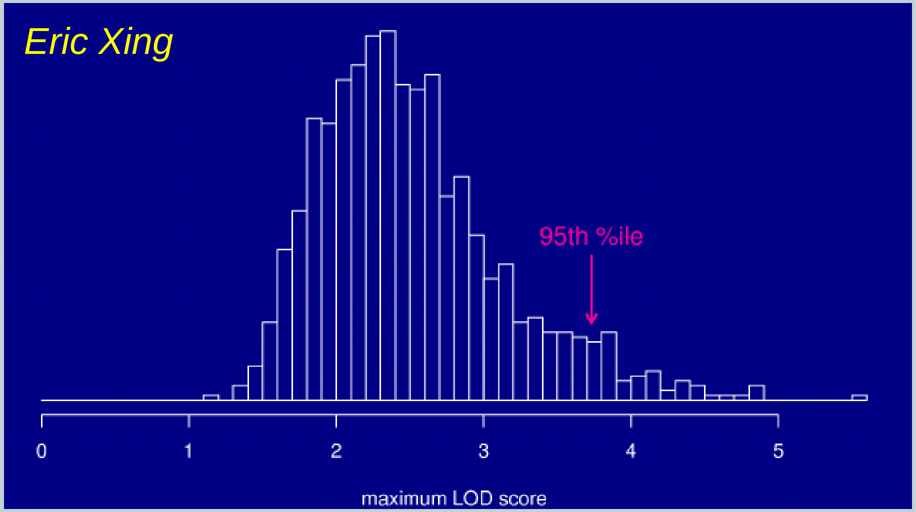
Heterozygous effects: The difference between the mean of the trait in the heterozygous genotypes from the average of the means of the trait in the two homozygous genotypes.

LOD threshold

LOD (logarithm of the odds) score: The strength of the presence of a QTL at a particular location across genome

LOD threshold= 95th percentile of the distribution of genome-wide maxLOD, when there are no QTL anywhere

LOD score= probability of having a QTL in the data/probability that there is no QTL in the data



For example; LOD of 2 means that it is 100x more likely that a QTL exists in the interval than there is no QTL

QTL analysis

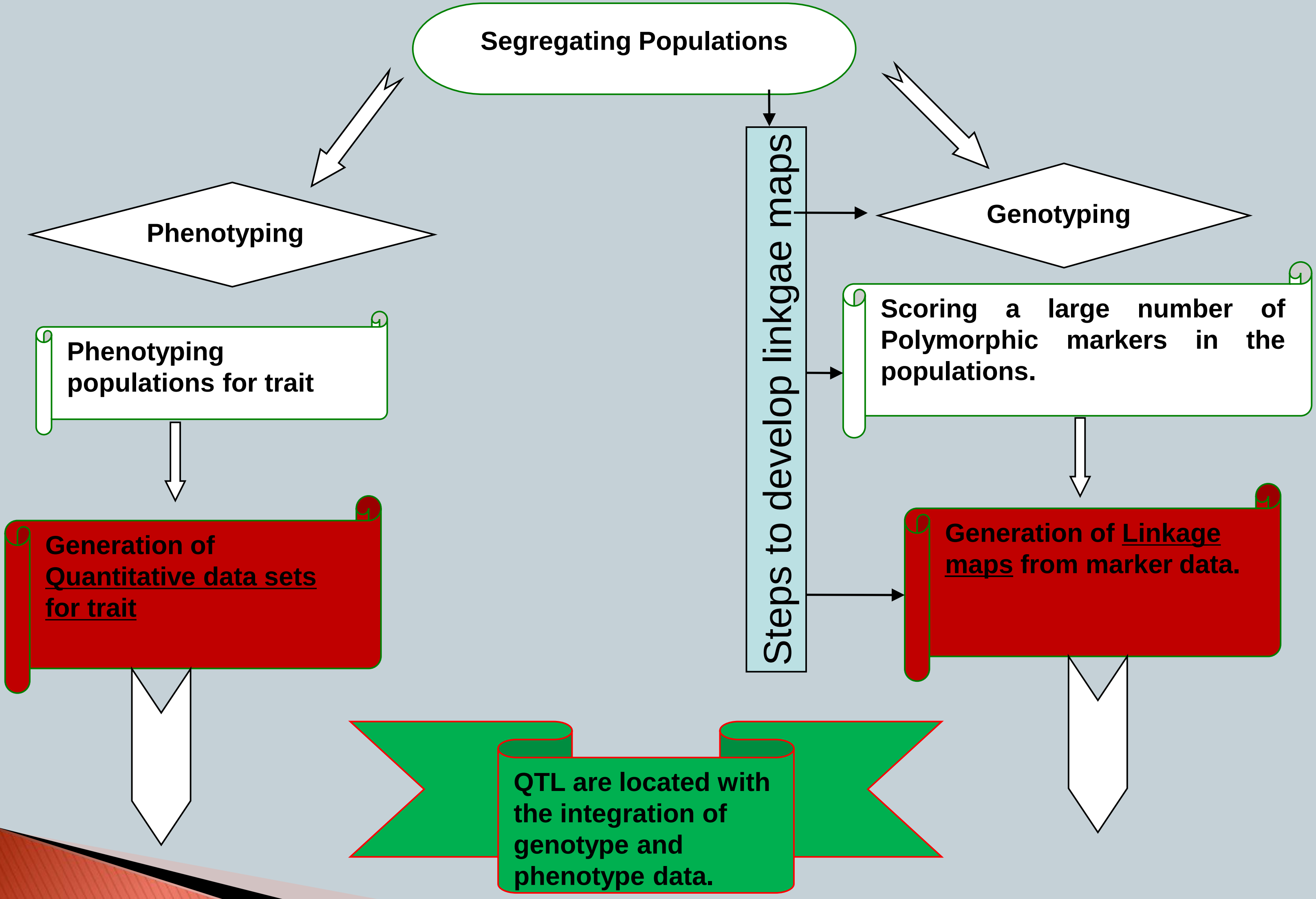
Location of QTLs

To localize a QTL we need individuals in which <u>**recombination has occurred in the vicinity of the QTL so that only markers very close to the QTL remain linked to it**</u>

When <u>size of the interval to localize the QTL decreases</u>, the <u>number of individuals</u> required to detect the recombinants in the interval and <u>number of molecular markers</u> increases

According to Mackay (2009), we would only need 29 individuals to detect at least one recombinant in a 10 cM interval, but 2,994 individuals to detect at least one recombinant in a 0.1 cM interval

Overview: Linkage map and QTL mapping



Segregating Populations

Steps to develop linkgae maps

Phenotyping

Phenotyping populations for trait

Generation of <u>Quantitative data sets for trait</u>

Genotyping

Scoring a large number of Polymorphic markers in the populations.

Generation of <u>Linkage maps</u> from marker data.

QTL are located with the integration of genotype and phenotype data.

QTL analysis methods

- **Single marker analysis**
  - **t-test (2 genotypic classes)**
  - **ANOVA (more than 2 genotypic classes)**
  - **Simple linear regression statistics**
  - **Kruskal-Wallis test**
- **Interval mapping**
- **Multiple QTL Model**

Single marker analysis

The basic principle is to divide the population in the genotypic classes based on the marker (AA, AB and BB) and then determine if there is correlation between marker and the trait effect.

Example

Plant	1	2	3	4	5	6	7	8	9	10
Genotype	A	H	H	H	B	B	A	H	H	A
Height	50	45	47	43	40	43	52	46	44	53

Correlation Genotype vs. Height *R=0.92*

Weaknesses

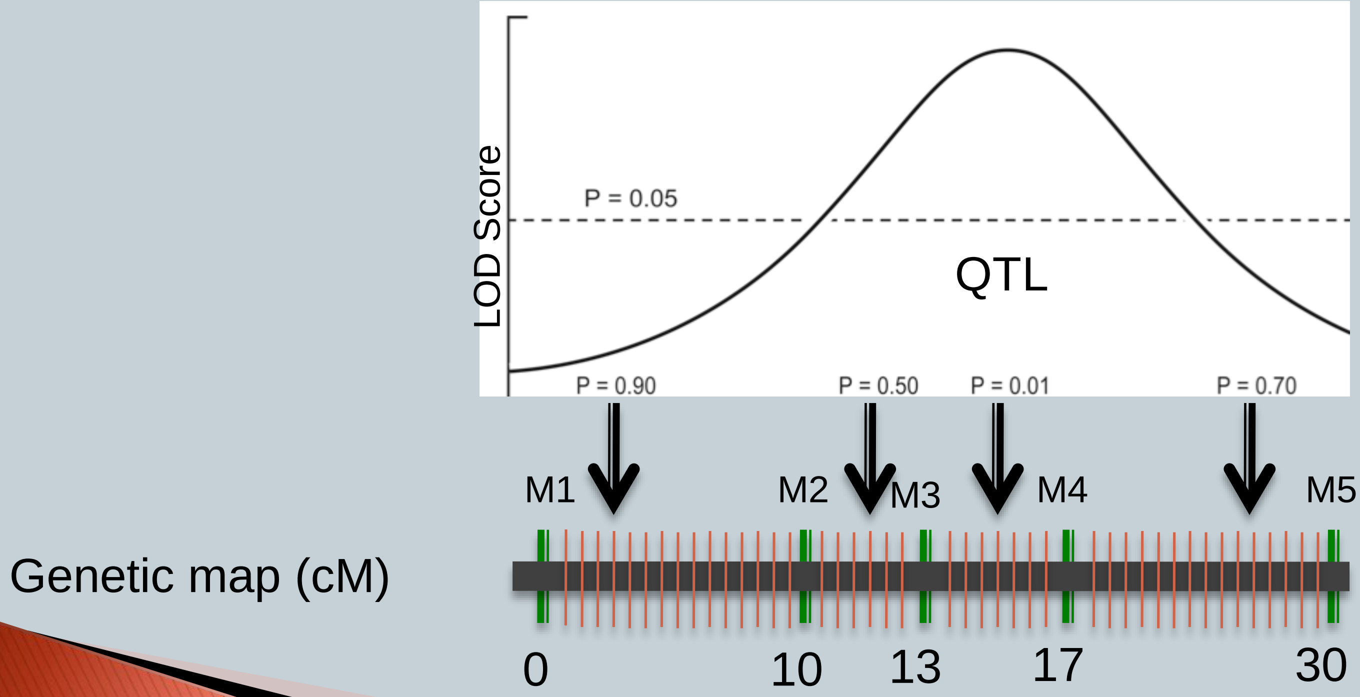
- **Cannot predict true QTL location and QTL effect**
- **Missing values at the marker are discarded**
- **The power for QTL detection decreases**

Interval mapping

(Eric Lander and David Botstein)

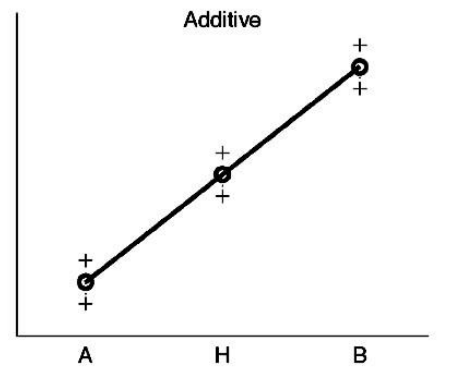
Most popular method

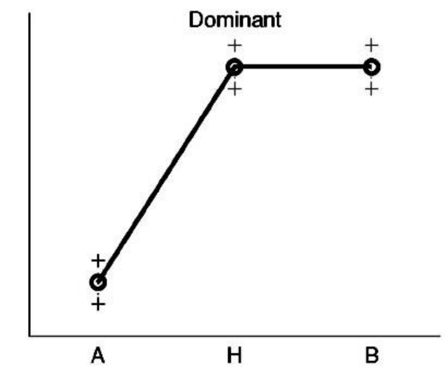
The **marker intervals** are searched **in a systematic**, linear (one-dimensional) fashion, **in increments** (for example, 2 cM), and statistical tools are used to test whether **a single QTL** is likely to be present within the interval or not.

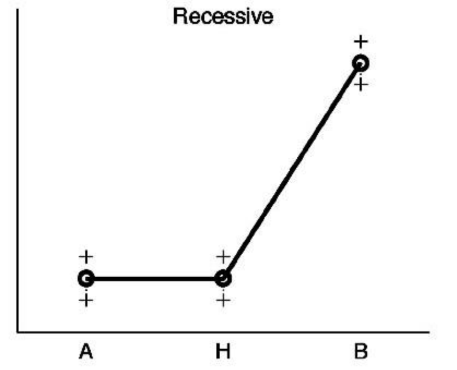


Interval mapping

Additive, Dominant, Recessive effects
of a single QTL (Gary Churchill)







A potential QTL might act independently, be linked to another QTL, or interact epistatically with other QTL

Works well when to map single QTL

Power and resolution is decreased when more than one QTL effect the trait

Multiple QTL Model (MQM)

More powerful than single QTL approaches because it can differentiate between linked and/or interacting (epistatic) QTL.

Procedure:
After an initial scan of QTLs by interval mapping, one performs MQM using the QTL detected in the interval mapping scan as cofactors. This can be repeated one or more times until the list of detected QTLs does not change.

Cofactors control for the variation caused by the genetic background (i.e. variation caused by QTLs outside the region where the QTL is tested).

Multiple QTL Model (MQM)

Selection of Cofactors

Forward selection: At each stage best new cofactor satisfying the selection criterion is added until no further candidate remains

Backward elimination: Starts with a multiple regression model, using a full set of cofactors (all putative QTL/markers) evenly spread over the genome. The unimportant or least important are dropped one by one until all remaining cofactors essentially meet the selection criterion

Stepwise selection: Backward elimination followed by stepwise procedure, including new cofactors and dropping old ones

Uses of QTL mapping

- **Fine mapping and identification of genes underlying QTL regions**
- **Inheritance basis of the traits**
- **Marker assisted selection (MAS)**
- **Map based cloning**

Summary: Linkage & QTL mapping

- Crossing over and recombination provide basis for genetic (linkage) mapping

- Relationship between recombination and genetic distance

- Genetics of quantitative traits

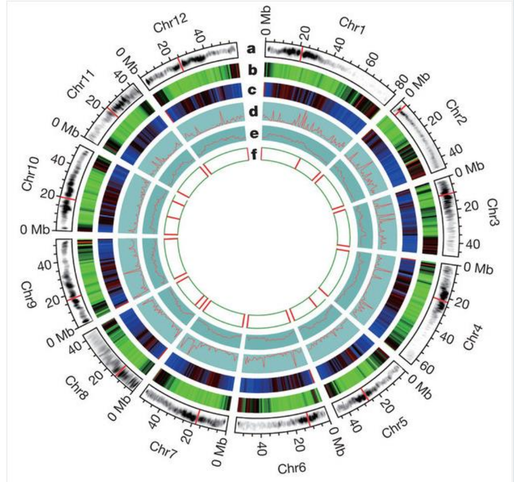
- Genetic locus controlling quantitative trait (QTL) can be identified by genotyping molecular markers and phenotyping the trait of interest in segregating population

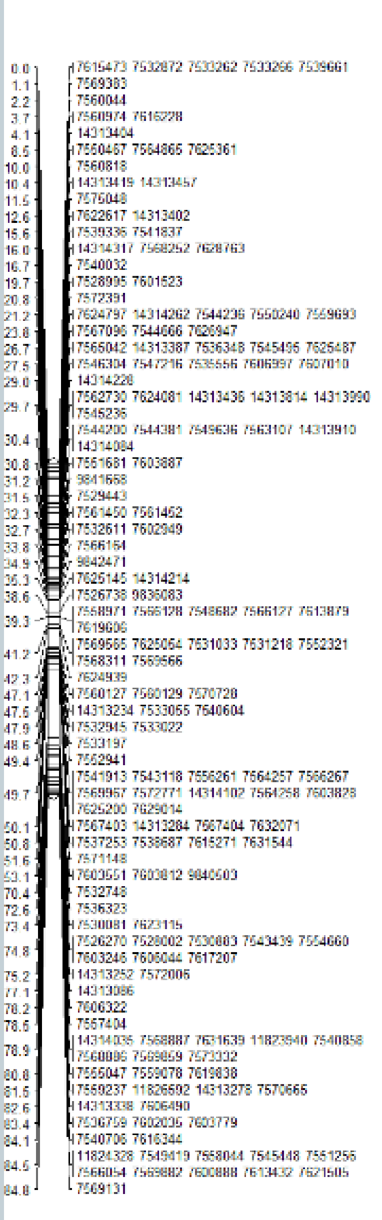
- Importance of QTL mapping in crop improvement

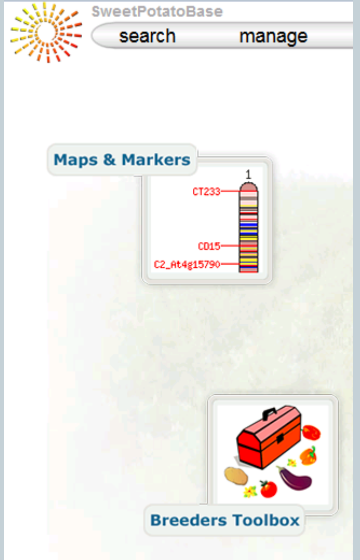
Exciting time for targeted and precise sweetpotato selection and breeding

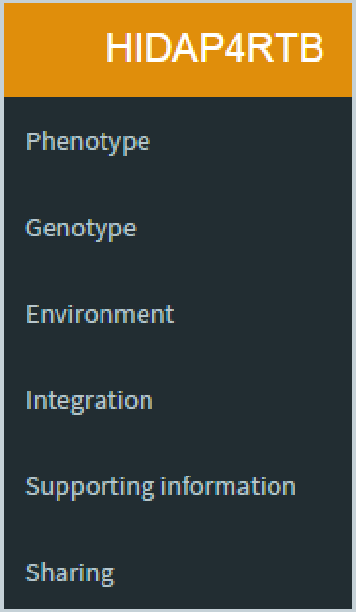
- ✓ **Genome sequence**
- ✓ **Genome browser**
- ✓ **Next generation molecular markers**
- ✓ **Dense genetic maps**
- ✓ **Phenotypic data**
- ✓ **Database and analytical tools**

Figure 1: The potato genome.

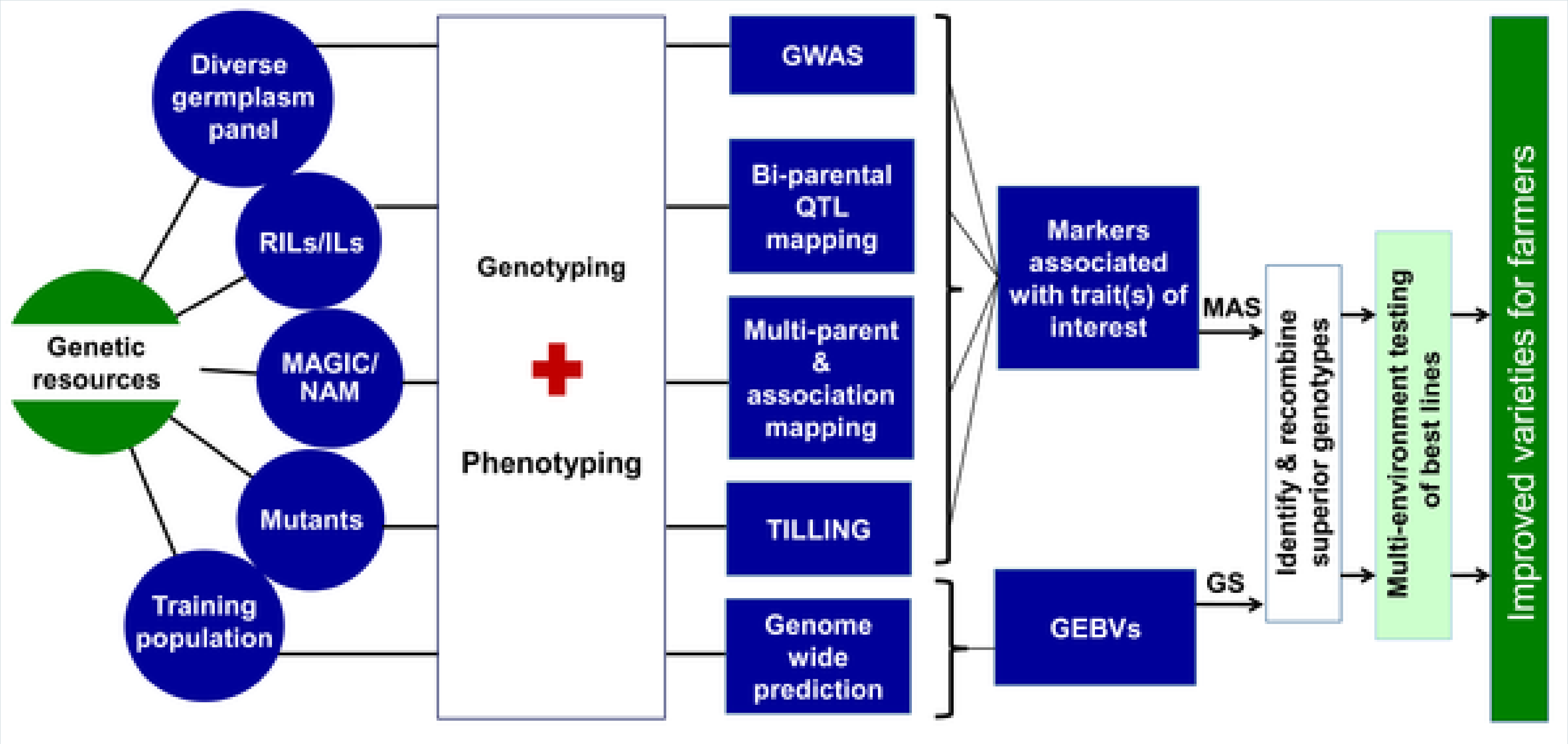






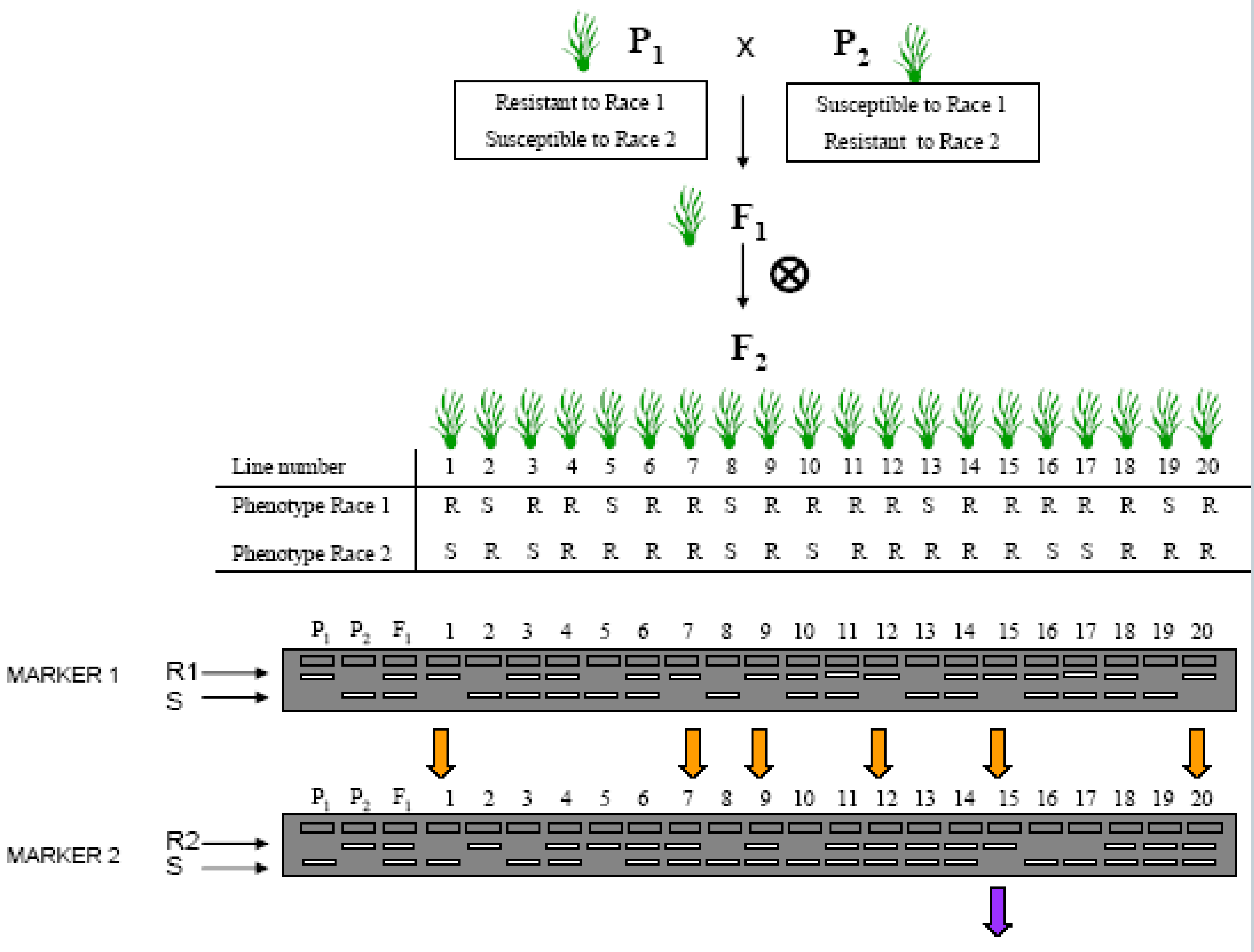


New paradigm of Genomics-assisted breeding



Varshney *et al*. 2014

Marker-assisted breeding

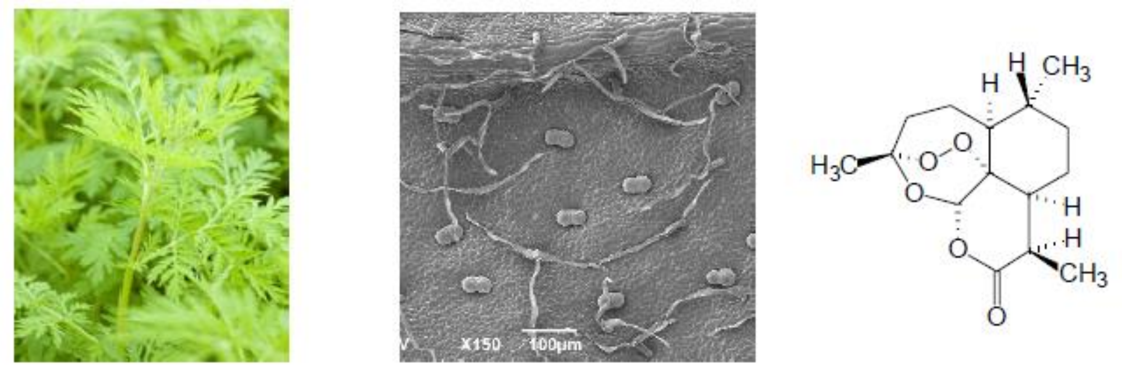


Line number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Phenotype Race 1	R	S	R	R	S	R	R	S	R	R	R	R	S	R	R	R	R	R	S	R
Phenotype Race 2	S	R	S	R	R	R	R	S	R	S	R	R	R	R	R	S	S	R	R	R

http://www.knowledgebank.irri.org/ricebreedingcourse/Marker_assisted_breeding.htm

Thank for your attention

Artemisia annua anti-malarial herb transformed to successful crop



Development of high yield artemisinin (anti-malarial) *Artemisia annua* (Asteraceae) with fast track breeding tools

Sources of genetic variation in *A. annua*

Induced genetic variation

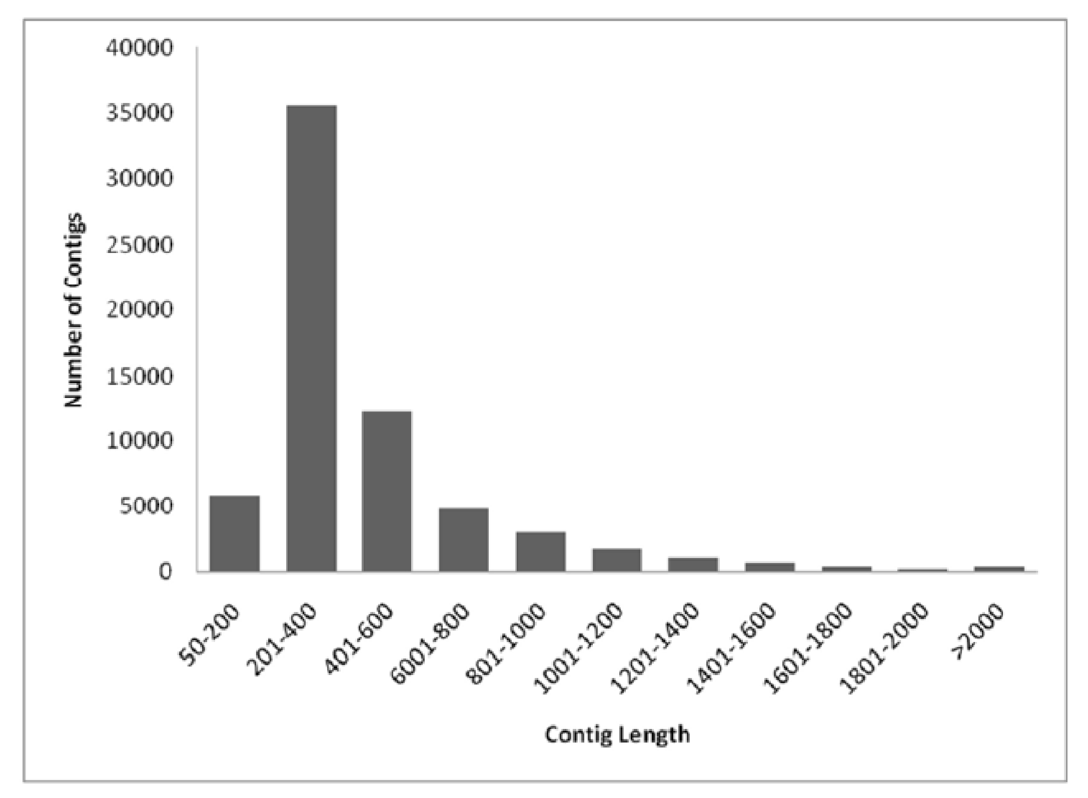
EMS-treated population

Natural genetic variation

Natural populations

Molecular marker development: Illumina GoldenGate SNP array

Genomic libraries and DNA sequencing

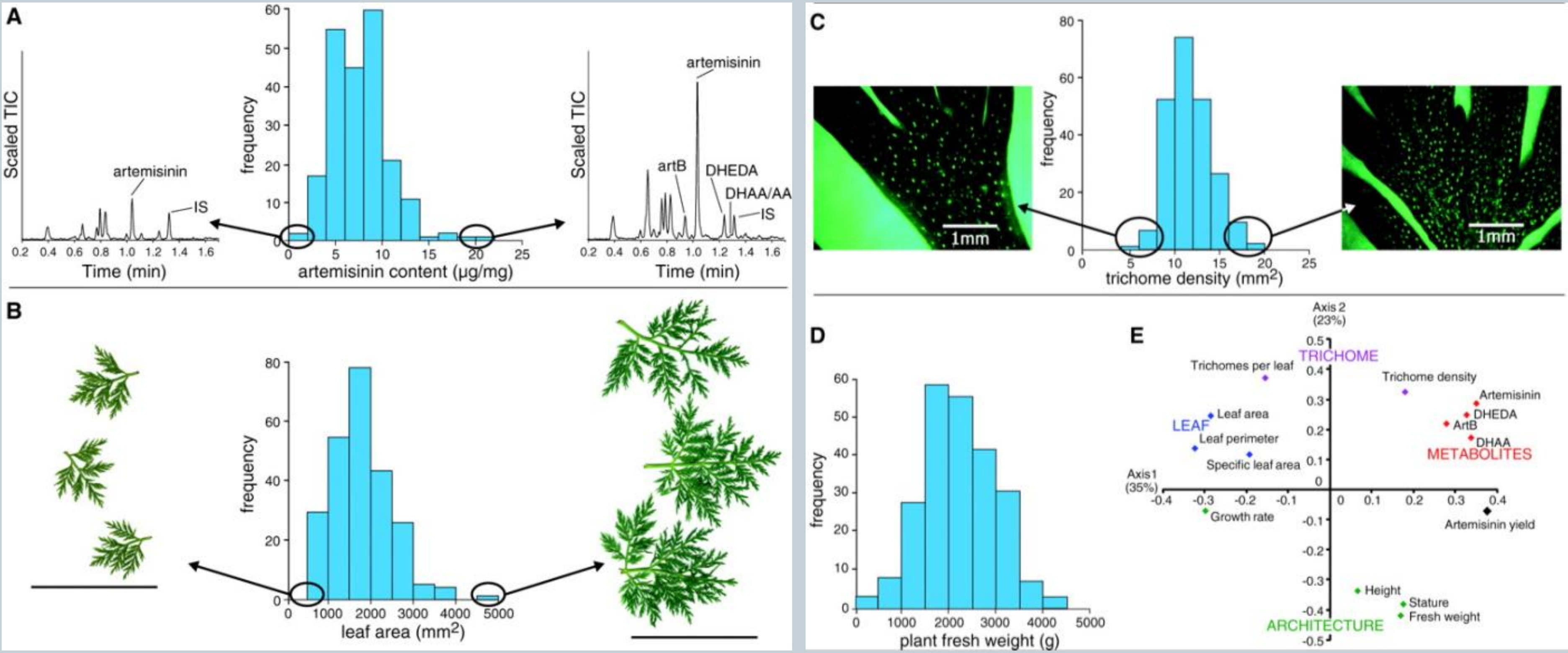


- Trichomes of high and low yield plants
- Trichomes from young and old leaves

- SNPs from candidate genes were prioritized for designing Illumina GoldenGate SNP array genotyping assay (1536 SNPs)
- SSRs based backbone of genetic map: Genotyped on ABI 3730 XL

Trait evaluation and metabolite profiling

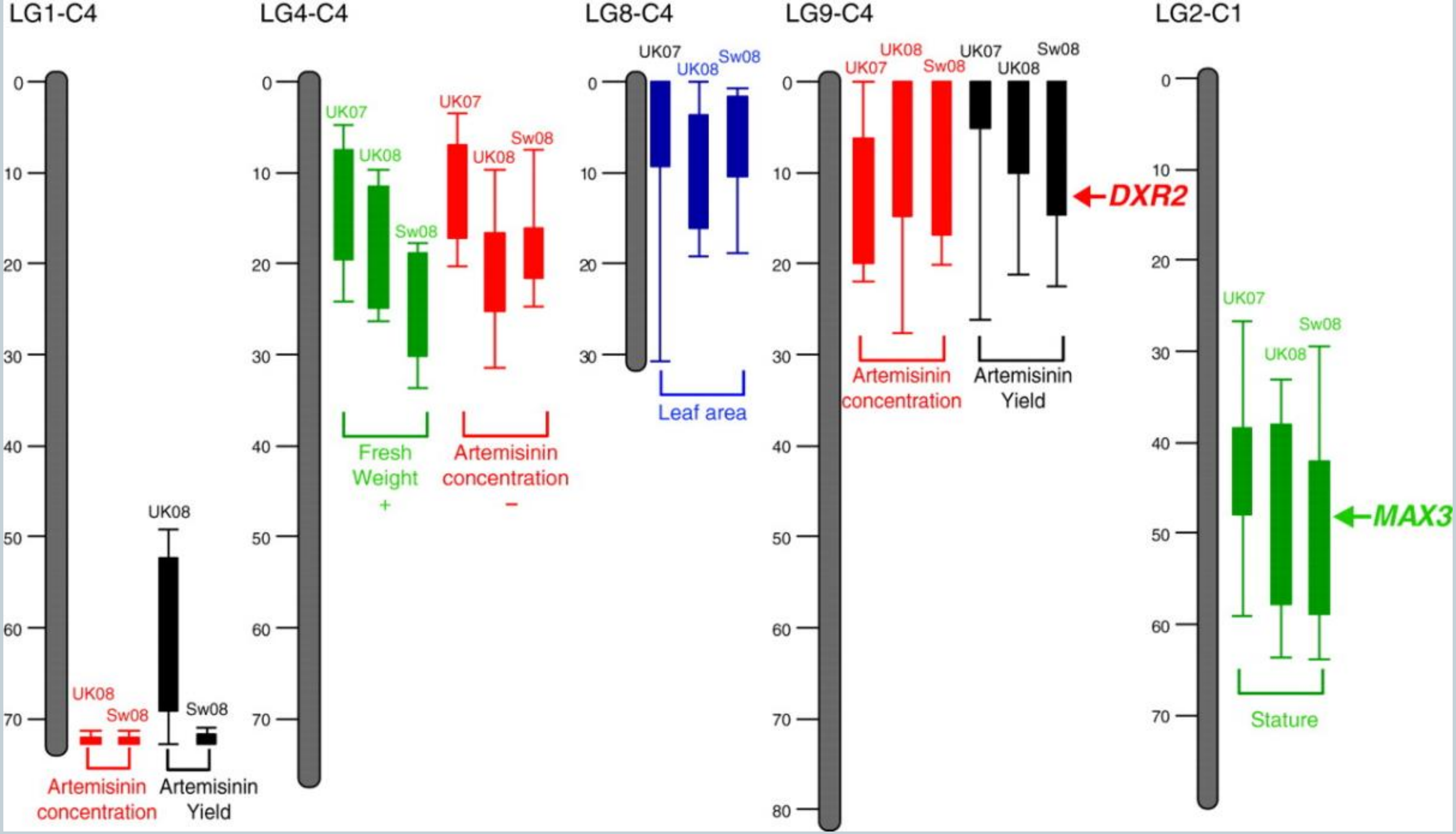
Field experiments conducted in UK, Switzerland and Madagascar



15 physiological traits (Number of branches/plant, number of trichomes/leaf, plant height, biomass)

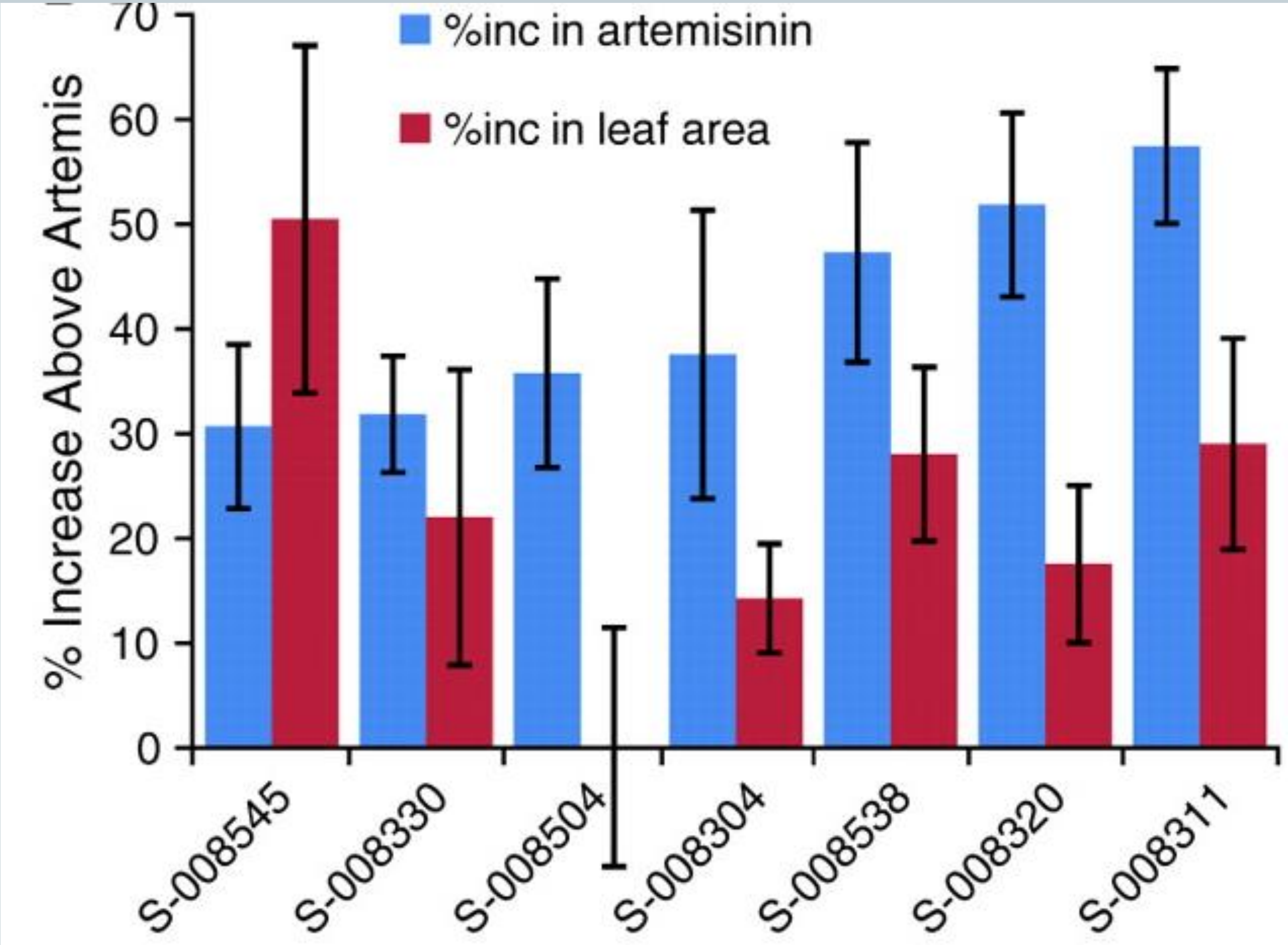
12 metabolites from the artemisinin pathway were studied using HPLC from fresh and dried leaves

QTLs



A selection of QTLs for key traits. QTLs are shown to the right and distances in centimorgans to the left of each linkage group.

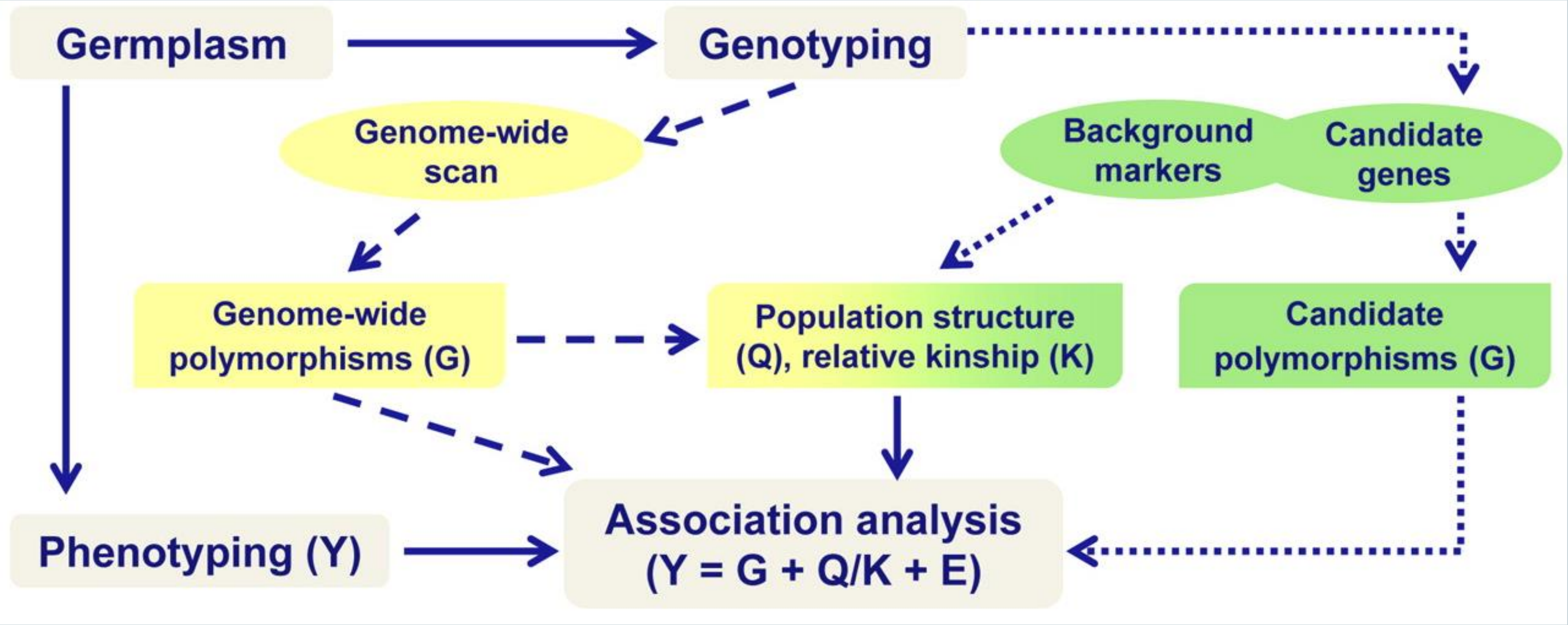
Use of MAS in hybrid production



The increase (%) in artemisinin concentration (in blue) and leaf area (in red), over Artemis F1 for seven hybrids produced from crosses of selected high-yielding individuals.

Graham et al. *science*, 2010

Overview: Association mapping analysis



Germplasm

Genotyping

Genome-wide scan

Background markers

Candidate genes

Genome-wide polymorphisms (G)

Population structure (Q), relative kinship (K)

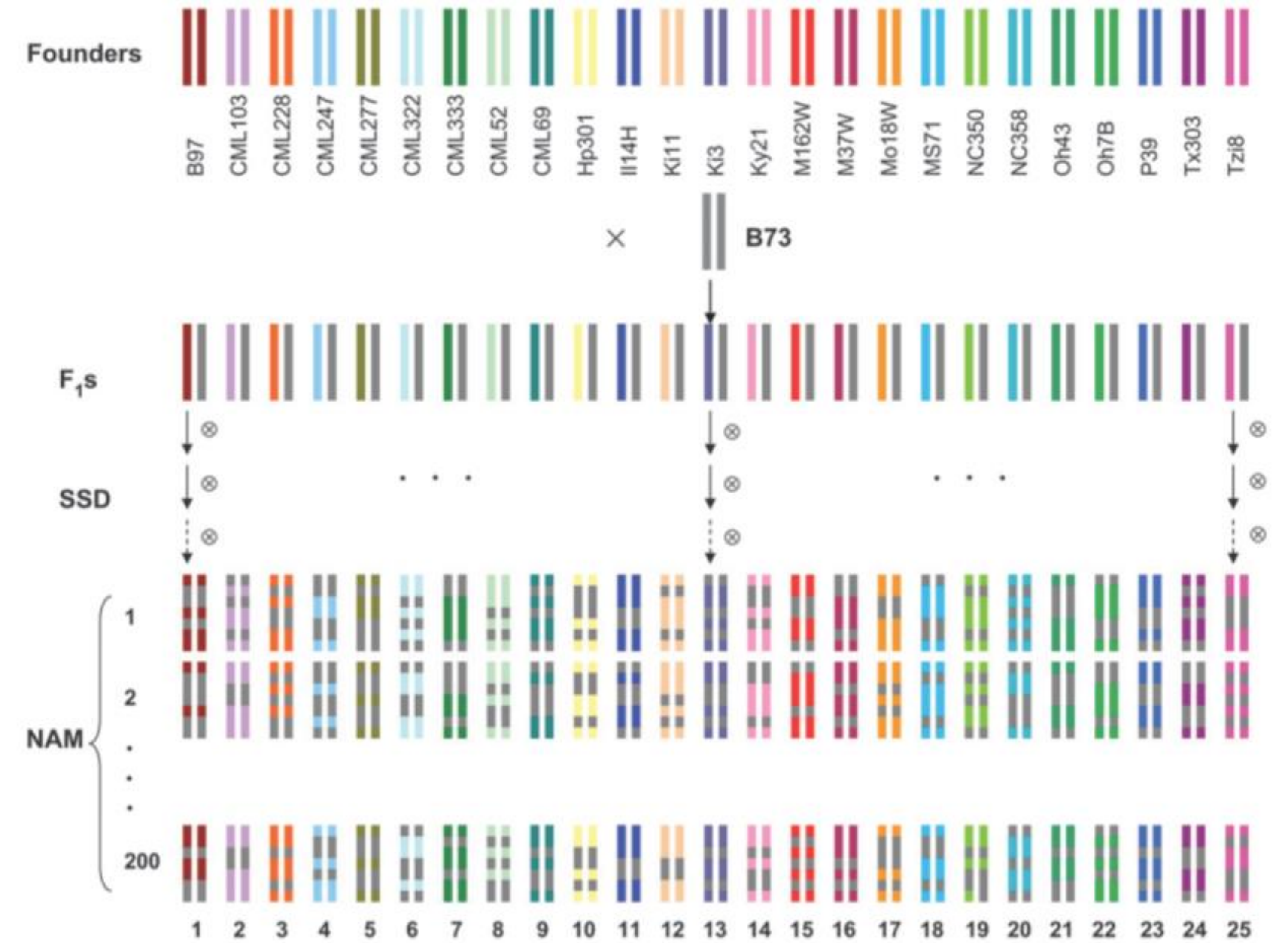
Candidate polymorphisms (G)

Phenotyping (Y)

Association analysis
$(Y = G + Q/K + E)$

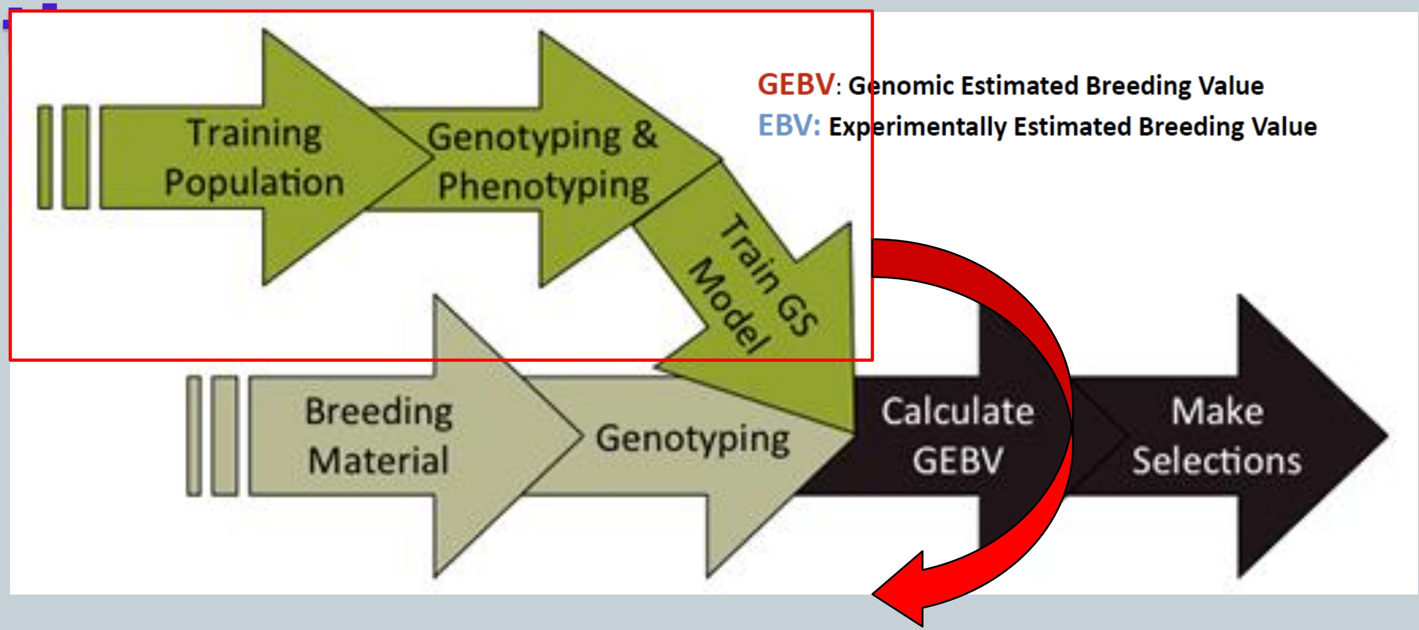
Zhu *et al.* 2008

The Maize Nested Association Mapping Population (NAM)



SSD=Single seed descent

Predicting the phenotype: Genomic selec



Training Population

Genotyping & Phenotyping

Train GS Model

Breeding Material

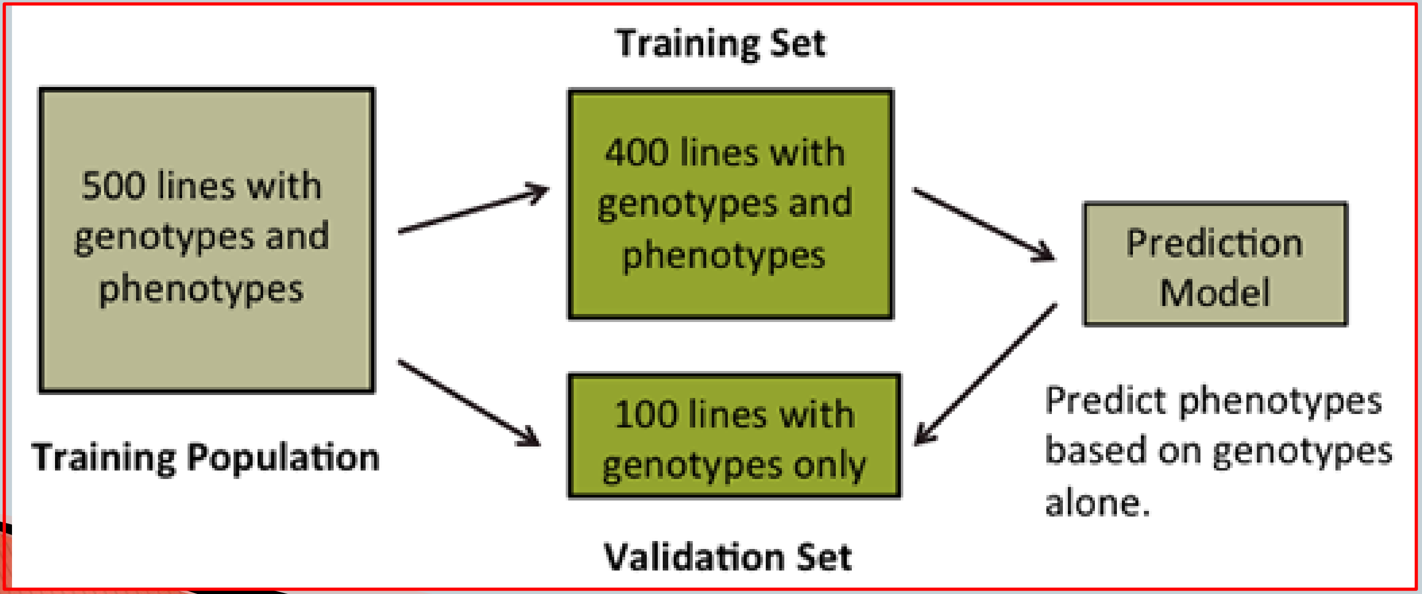
Genotyping

Calculate GEBV

Make Selections

GEBV: Genomic Estimated Breeding Value

EBV: Experimentally Estimated Breeding Value



Training Set

500 lines with genotypes and phenotypes

Training Population

400 lines with genotypes and phenotypes

100 lines with genotypes only

Validation Set

Prediction Model

Predict phenotypes based on genotypes alone.