

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

**Awais Khan**

[awais.khan@cgiar.org](mailto: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

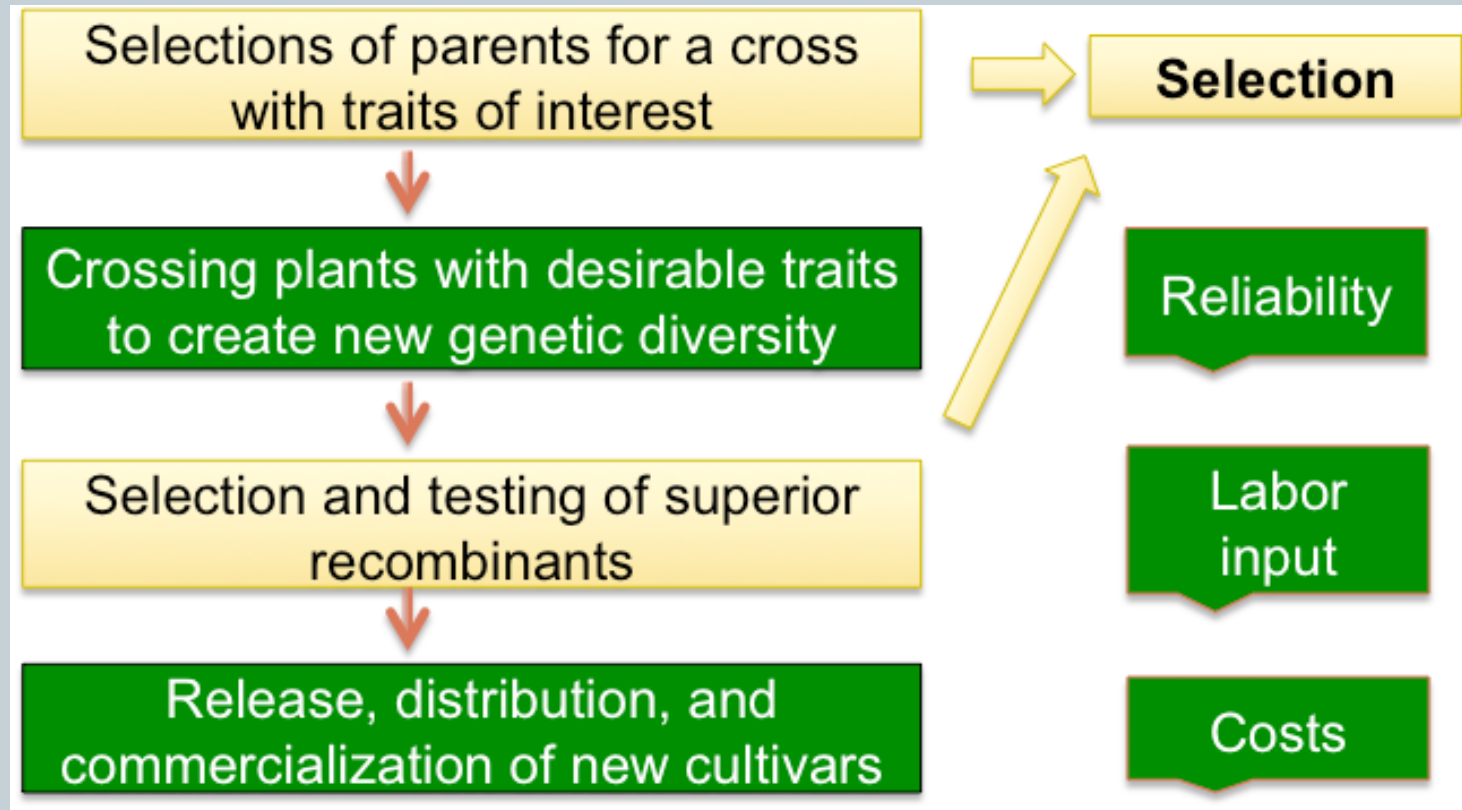
**Phenotyping with special attention to Genotype x environment interaction:** appearance and performance in general, in particular, response to environment

Genomics for identification of genetic basis of traits of importance

Selection of parents and progeny with desirable traits

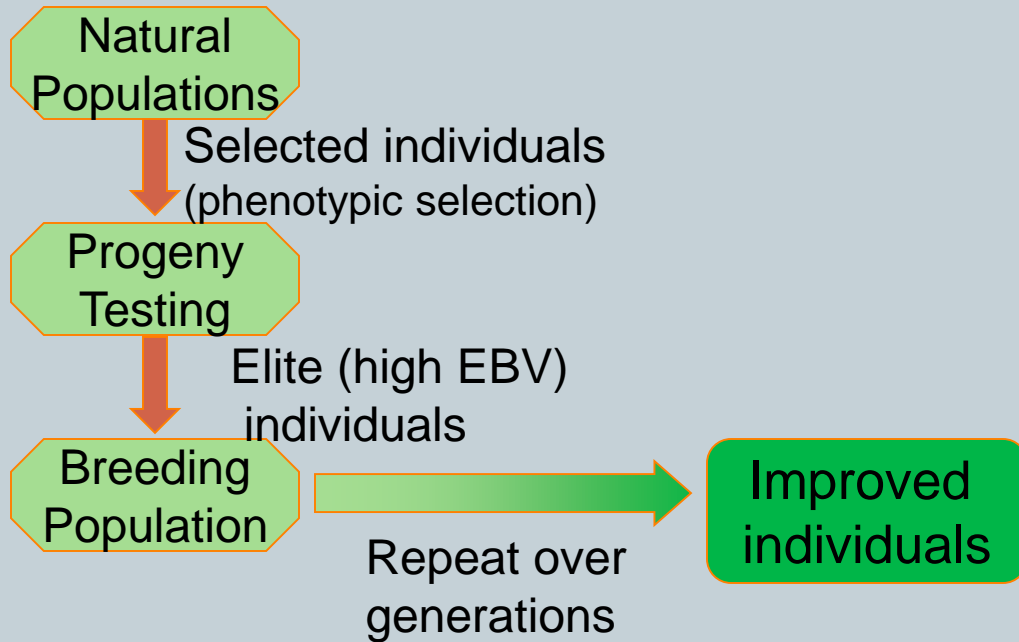
# Challenges of plant breeding

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



General steps in plant breeding (modified after Gepts 2002)

# Phenotypic selection: Selection based on appearance and performance



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

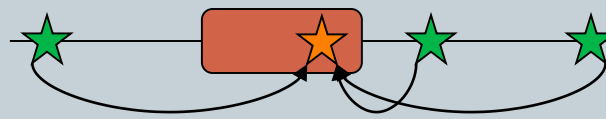
# Concept of Marker assisted selection

## Molecular breeding




Association between molecular marker and causative gene



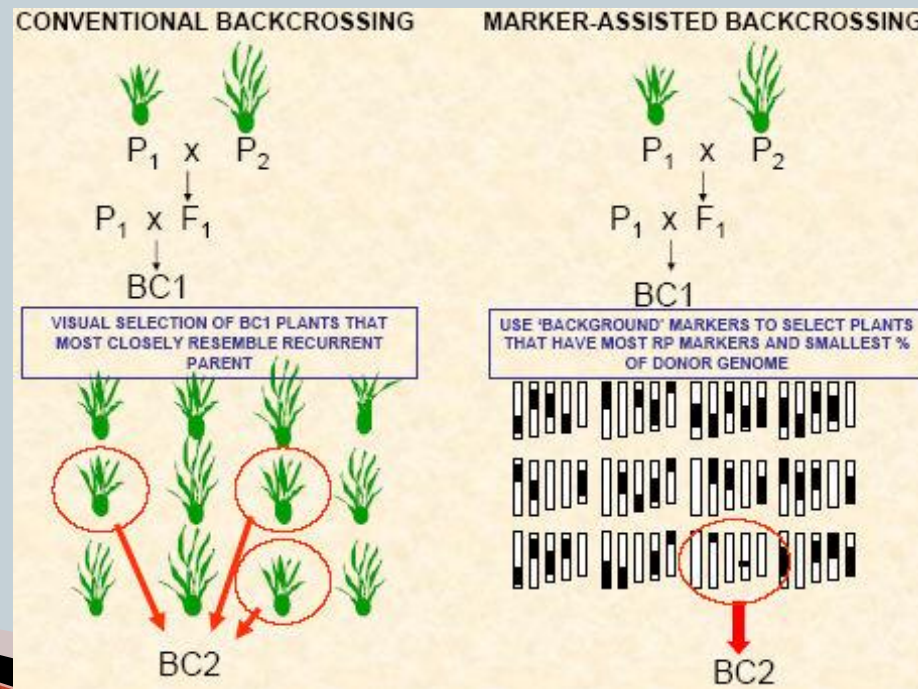
Direct association



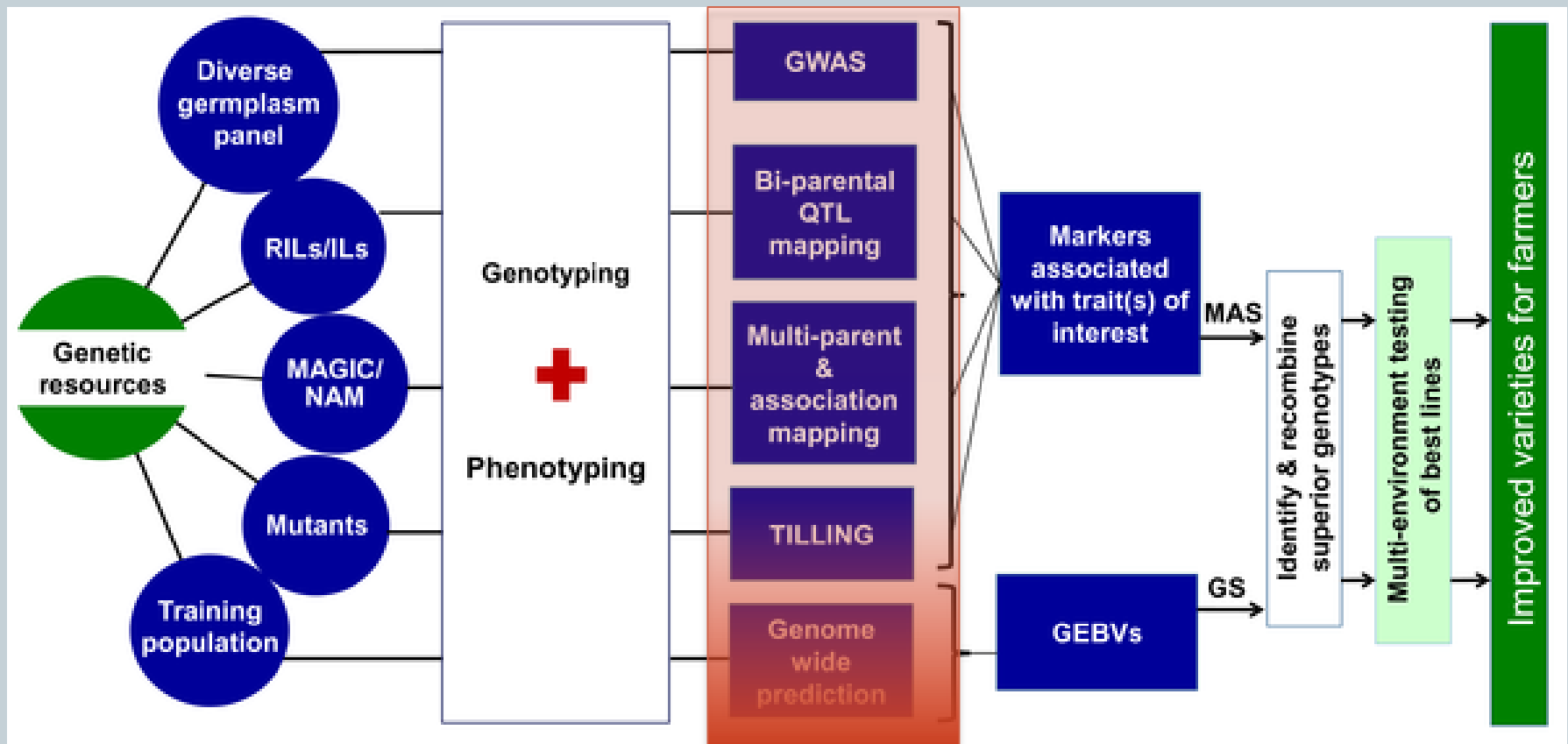
Indirect association

-  Causative gene
-  SNP within gene
-  SNP in LD with gene

Hirschhorn & Daly, 2005



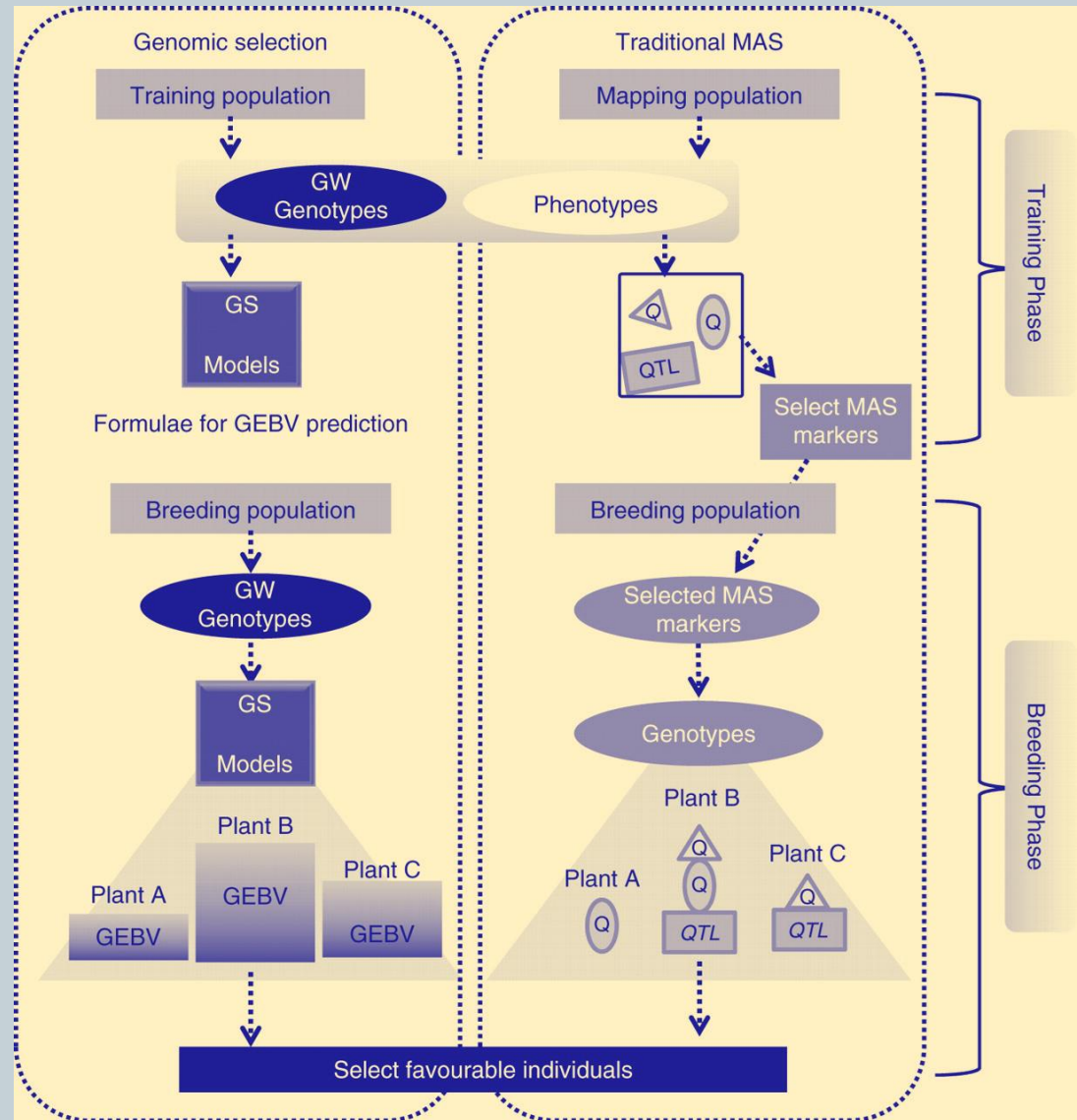
# Marker-trait association identification



# Predicting the phenotype or selection of progeny with desirable traits

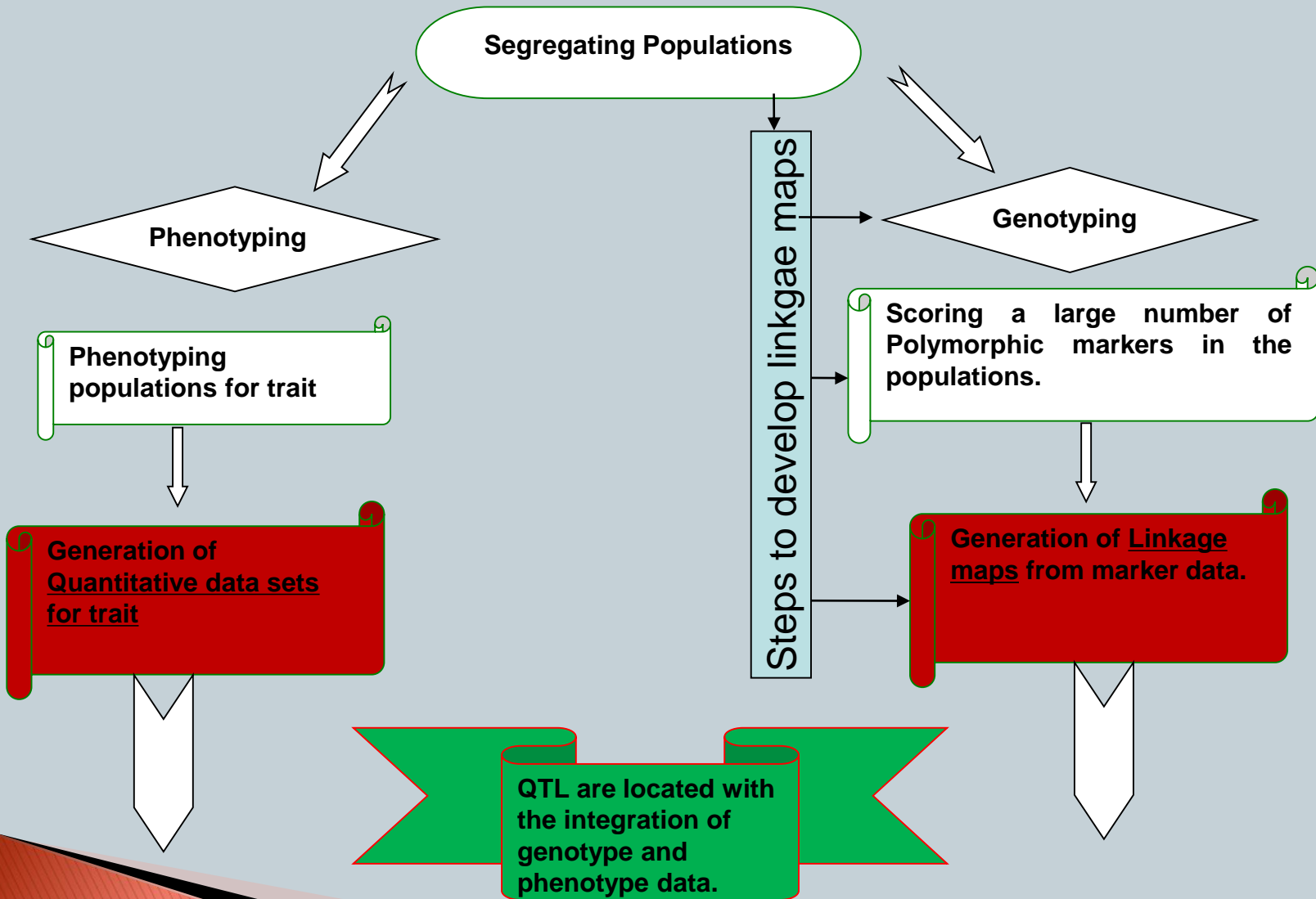
**Marker-assisted Selection (MAS):**  
Usually, plants are selected for up to 10 alleles

**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





# Overview: Linkage map and QTL mapping



# 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.



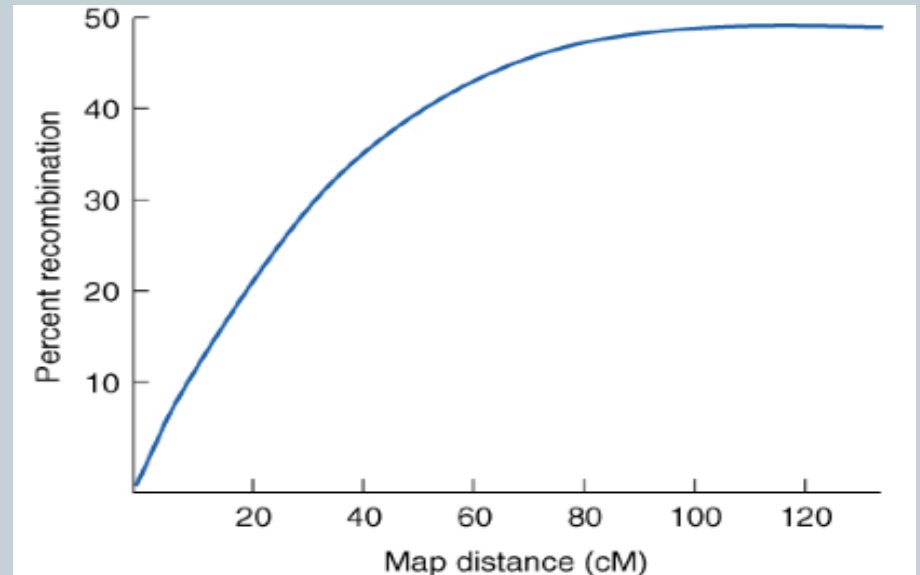
# Crossing over recombination and map distance

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

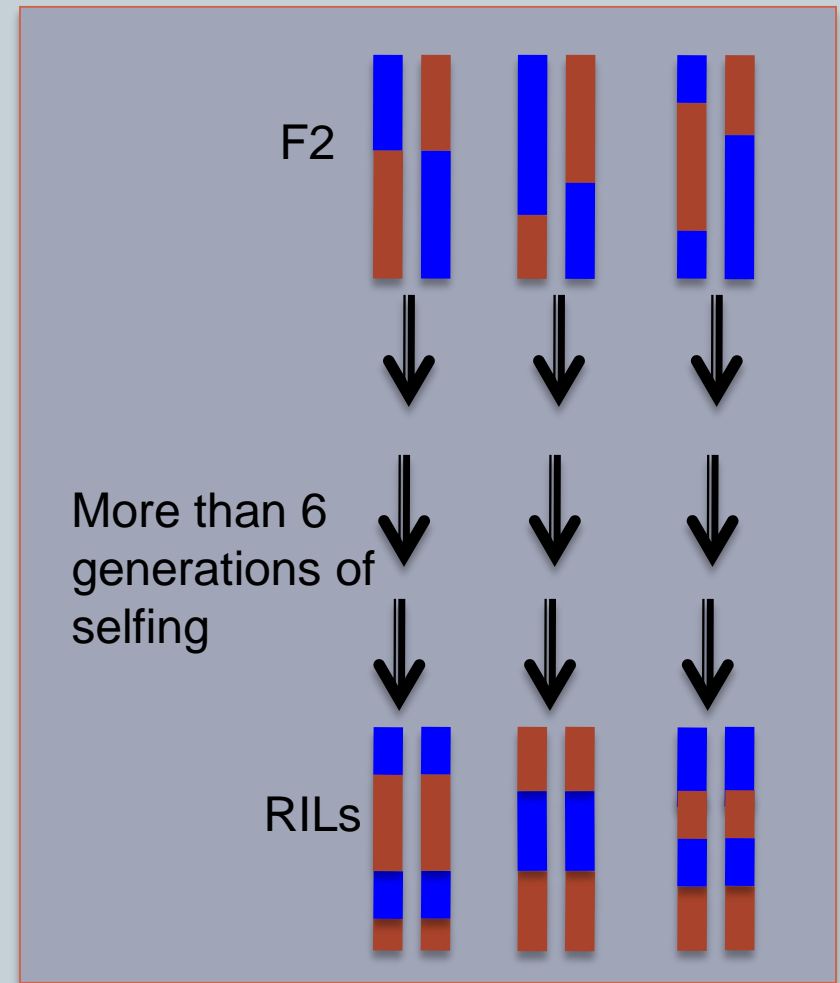
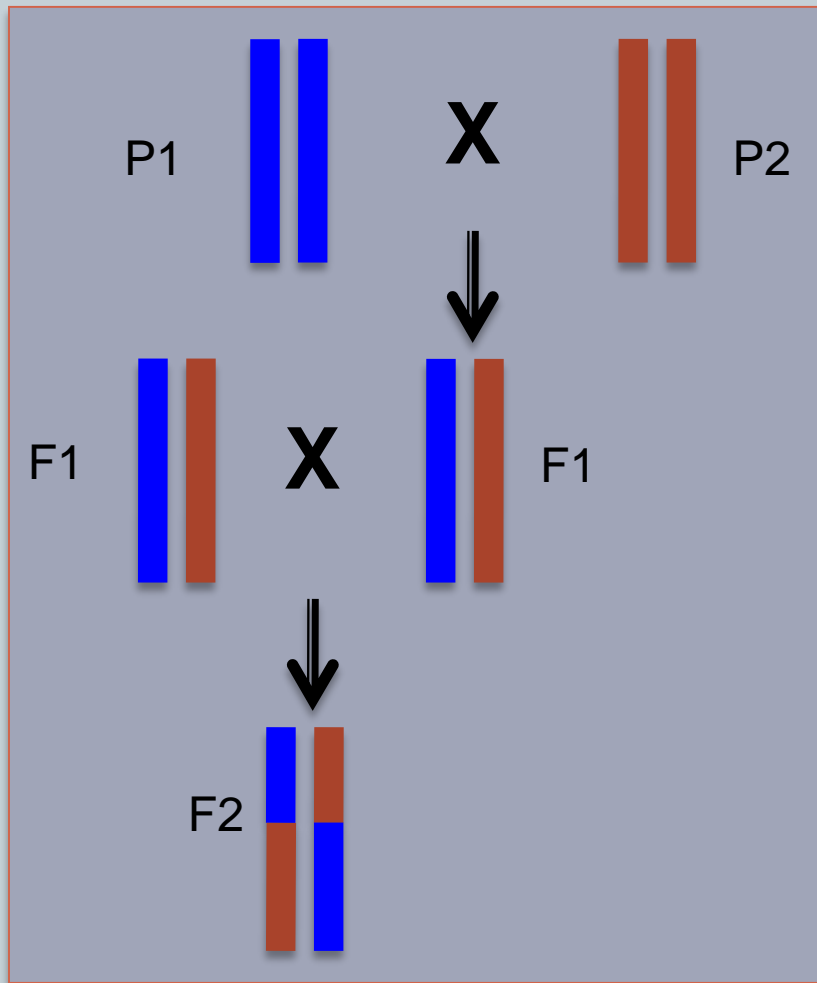
**Unit:** Morgan/centiMorgan also known as the genetic map distance:  $d$

## 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



# 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

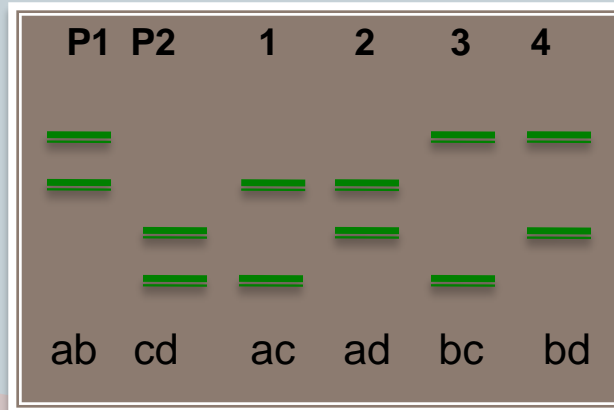
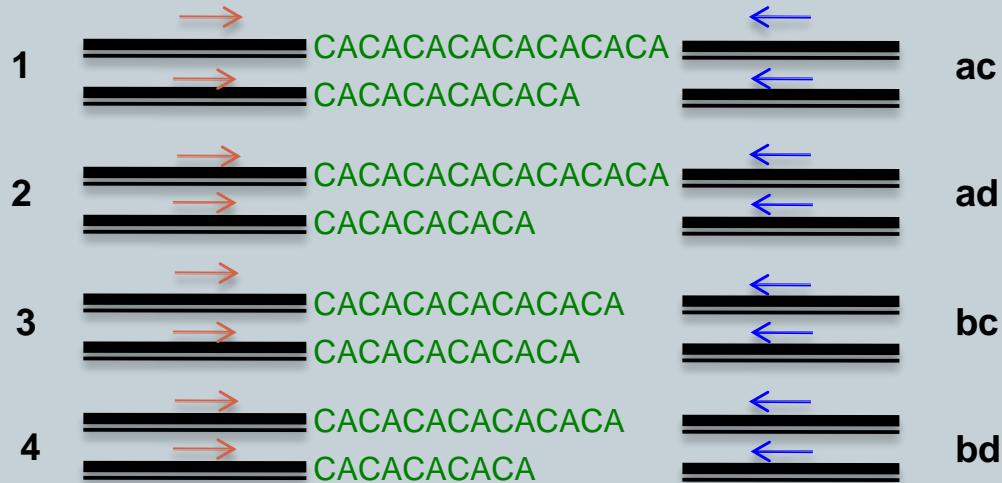
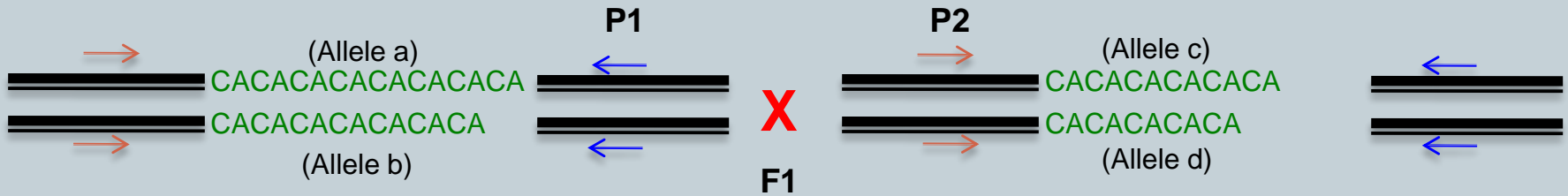
Reproducible

Easy to use

Economic

High-throughput

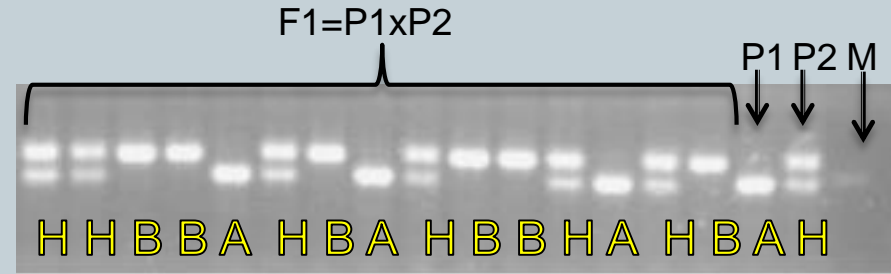
# Simple Sequence Repeats (SSRs)





# 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)

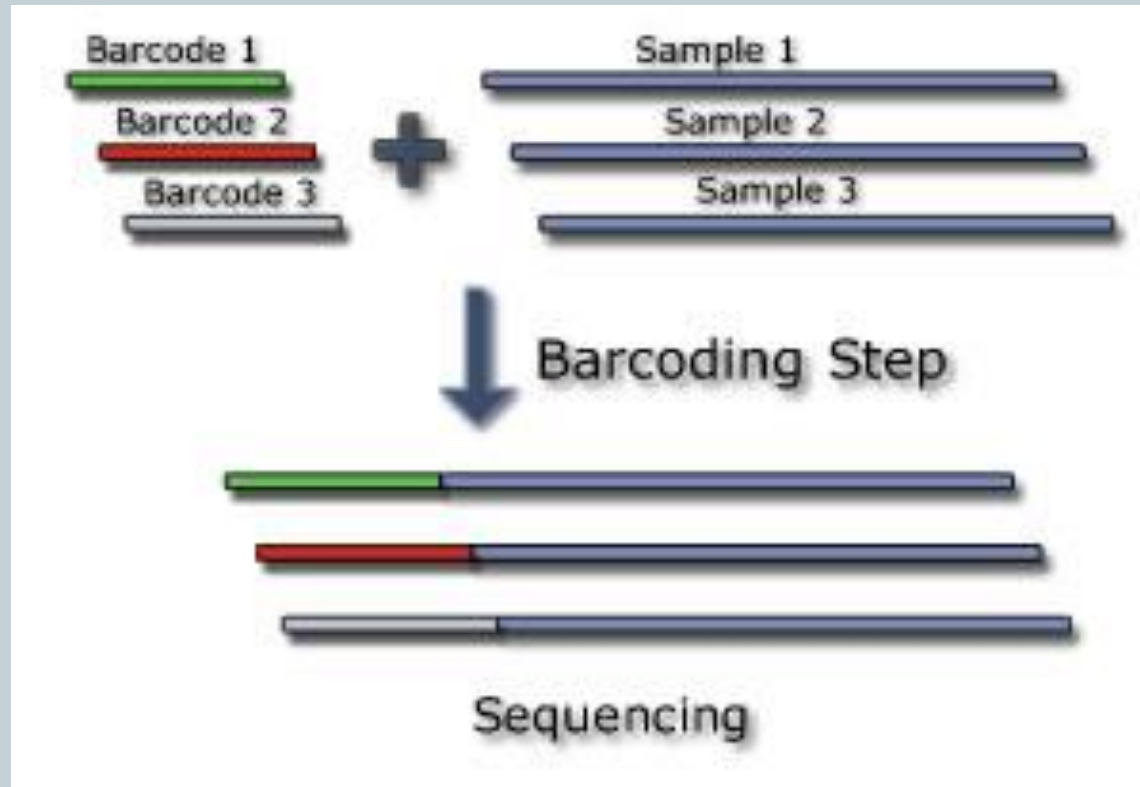


## Marker data

### 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

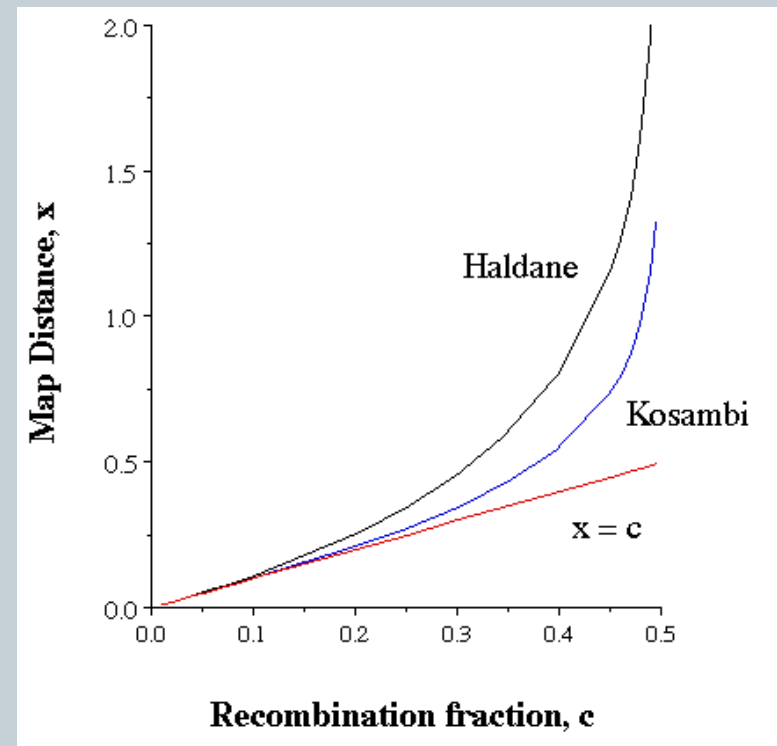
<u>Gametes</u>	<u># of gametes</u>			
ABC	1080			
abc	1071			
AbC	293	<u>A - B</u>	<u>B - C</u>	<u>A - C</u>
aBc	282	293	293	78
aBC	78	282	282	66
Abc	66	78	6	6
ABc	6	66	4	4
abC	4	<b>719</b>	<b>585</b>	<b>154</b>
<b>Total</b>	<b>2880</b>	<b>0.250</b>	<b>0.203</b>	<b>0.053</b>



# 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

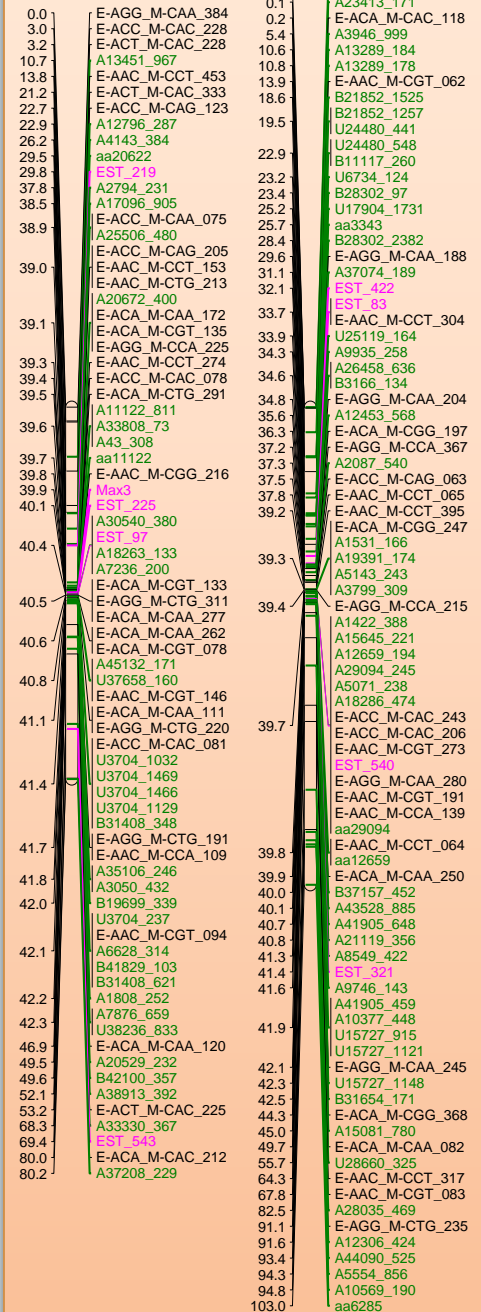
First report: 33086 SNPs were identified  
11091 SNPs were mapped

## LG02

0.0	17560198 7557432
2.6	7539952
3.7	17525949 7527108
4.8	17512188 7542189
5.2	17549551 7617389
5.9	17526952 7562608
6.3	17571005 7530308 7626656
7.4	17529817 7551908
9.3	7563840
9.6	11822019 7542043 7624084 7571468
10.0	7600834
10.4	7538303
11.1	7526337
11.9	17527179 7571967
12.6	17542271 7567489 7616911 7545148 7559788
13.4	7542919
13.7	17551701 8840631
14.1	17531008 7556668
14.5	7565138
16.3	7532585
18.2	7602120
20.8	114314403 7532839 7559369 7601401
21.5	7600713
21.8	17542989 8842349 7630216 7633406 7540981
21.9	9847619
22.3	7569189
24.9	7533487
26.6	7531726
28.3	7553925
28.2	17547657 7556747
28.6	7559638
28.9	7543896
33.0	114313564 7531810 7548072 7630896
33.7	17562805 7555714 14314022
34.1	7603546
34.5	7621215
34.9	7535649
35.2	17541748 7550881 7536166 7545153 7603628
40.4	7540424
41.9	17562645 9840512
42.3	11815256 14313502
44.1	17592723 7529265 7561387 7604617 7613532
45.2	7632162
45.6	114314586 7580231 14314103
46.4	17548890 7547077 7548891 7629380
46.7	7531180
49.3	17531155 7631048
49.7	7606140
50.1	7530890
52.3	17560881 7564149 7545927 7601712 9844360
52.7	7575275
53.0	17569824 7628566 7631257
56.0	17530988 7544444 7567775
56.4	17530783 7532089
60.4	17571562 7572863 7571563 7631089
64.6	7575172
64.9	7544619
65.2	17567187 7564668 7565759 7530358 7564669
66.6	17606725 7612802
66.8	7561525
66.8	7625888
66.7	17565614 7627669
68.6	17544253 11826091 7540810 7608856
69.3	7622893
69.7	7546776
70.8	7622890
73.4	14313854 7531770 7544364
73.8	14313806
73.8	114313495 14313511 14313515 14314219 7537490
74.4	17541188 7550832 7570218 7574787 7626350
75.6	17567505 7567507
75.2	7531362
77.1	114313462 7536163 7666135
77.6	7535479
78.9	17620470 9844151 7530887 7531581 7536338
79.9	17544552 7565008 7602187
79.7	17525989 7561467
80.1	7602938
80.8	114314626 7547831 7560359 7570664 9836947
81.5	17628895 7537657
81.9	17571144 7532879 7561441 7568808 7571145
	7605203 9843689

## LG03

0.0	17563801 7558446
0.4	7530738 7533557 7600500
1.1	7528297 7573665
1.5	17560721 7626881
2.6	7608808
4.4	114313145 7540493 7601215
4.8	7551079 7568595 14313225
5.2	7538514
5.9	7568349
6.0	7603025
8.2	17539211 7530595 7555067 7572272 7631159
11.1	17568453 7572773
12.2	7541219
13.3	7566474
13.7	114313292 7580022
16.3	17624383 7536354 7540855 7624394
16.7	7542323 14313194 14313342 7557935
17.0	11181606 14313343 7604945
17.8	11818513
20.0	7511986
20.4	7543894
20.8	17546869 7564494
21.1	17561368 7560016 7563876 7669307 7630128
21.5	17529486 7555683 7556621
21.9	7538324
23.0	7569306
25.2	17568741 7544626 7558370
25.6	7603343
27.8	9836309
28.2	17538408 7551439
28.9	9842585
29.3	17531591 7605685
30.4	14313273
30.8	7617373
32.3	7617383
34.3	17565455 7558916
34.4	17564588 7530319 7532437
36.3	114314586 7549651 7549652 7632056
37.1	17631451 7631452
38.6	17545948 7550319
40.4	17548936 7568887 7572735 7630897
40.8	9838132
42.6	11815958
43.8	17560555 7532123 7543785 7663487 7600506
44.5	17560910 7561326 14314071 7624175 7624321
44.8	7566832
46.3	17549340 7530799 7549341
46.7	7628394
47.1	7547810
49.0	114313102 7563677 7622631 7622501
50.1	17546115 7599504 7628328
50.4	7566676
50.8	114313468 7529280 7629020
51.9	7573296
52.7	17542820 9839324 7542821
53.0	17550247 7567236 7561908
53.4	17567237 9841141 7567235
53.8	17532013 7632154
54.5	17568671 7573867 7541747 7574148 7628768
54.5	7631729 14313759 832725 7539514 7548508
55.0	7532854 7575114 7631602
56.0	7553479
56.1	17566778 7568780 7569163 7630166 7630726
57.1	17526619 7543168 7630725
60.8	17566886 7623822
63.0	7600844
66.4	17539475 9841135 7555564 7600230
67.5	17628233 7526287 7595837 7631532
68.3	7619566
68.3	7630659 7572434
77.3	17542038 7600798



# Quantitative trait locus (QTL) mapping

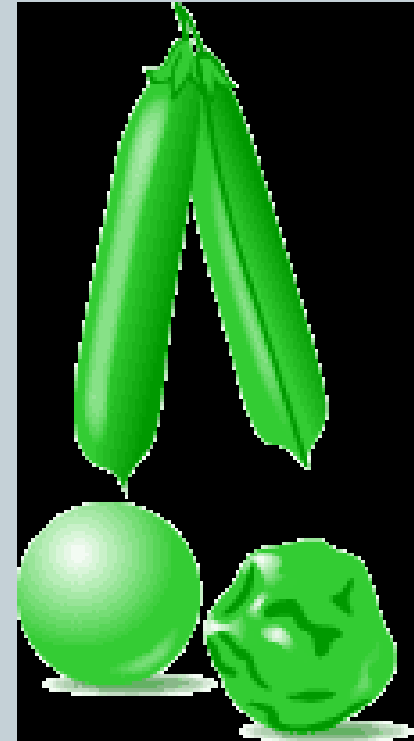
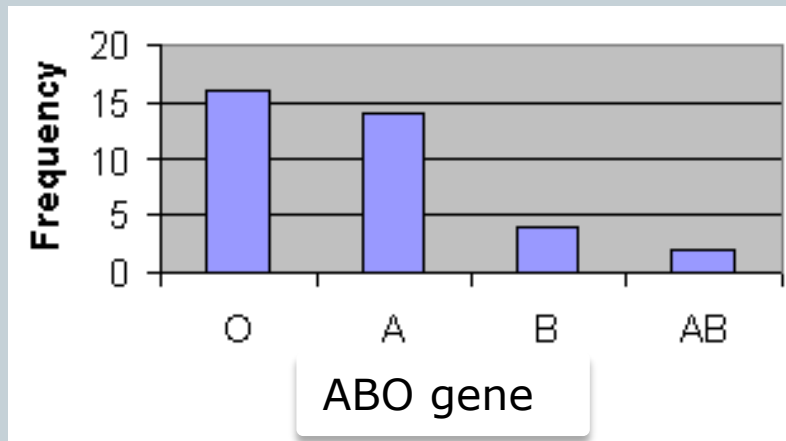
Quantitative trait locus (QTL): A genomic region that is associated with a quantitative trait

# 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



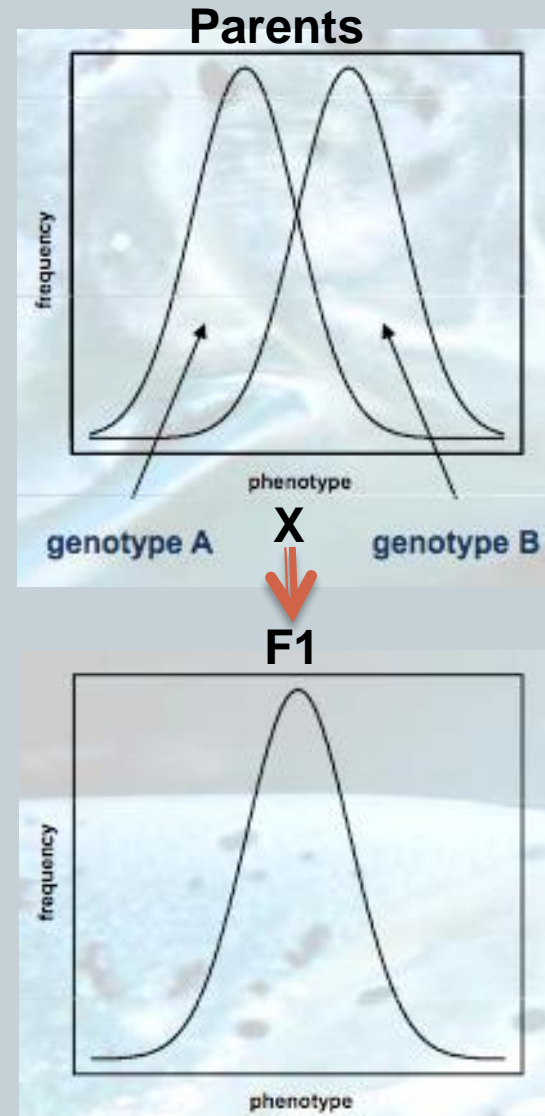
The wrinkled-seed character of pea is caused by a transposon-like insertion in a gene

# 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

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

$V_P$  = phenotypic,  $V_G$  = genetic,  $V_E$  = 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 epistasis

**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

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

# Heritability

The proportion of the genetic variance to the total variance

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

$$H^2 = V_G/V_P$$

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

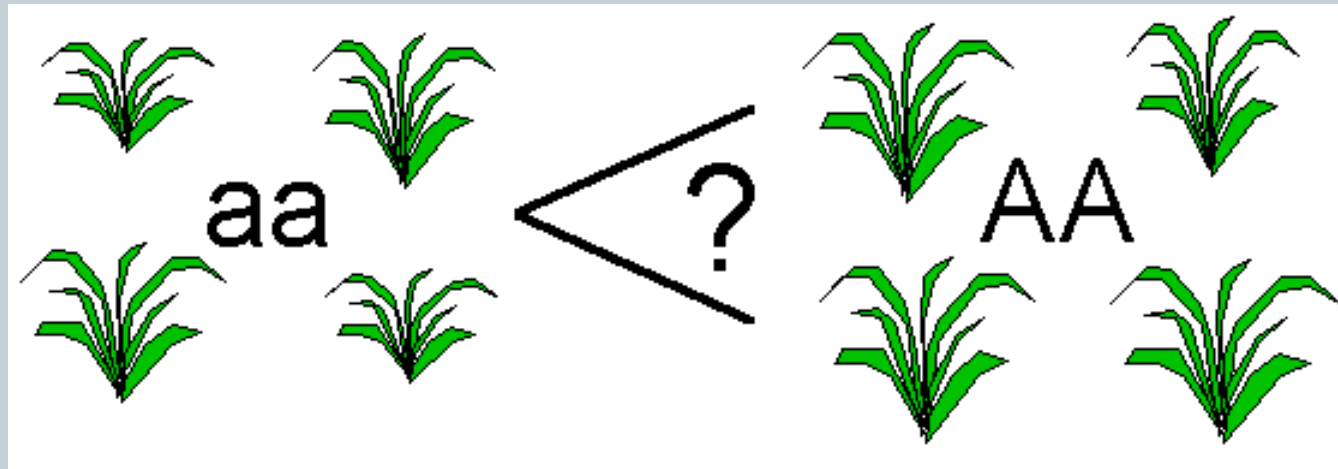
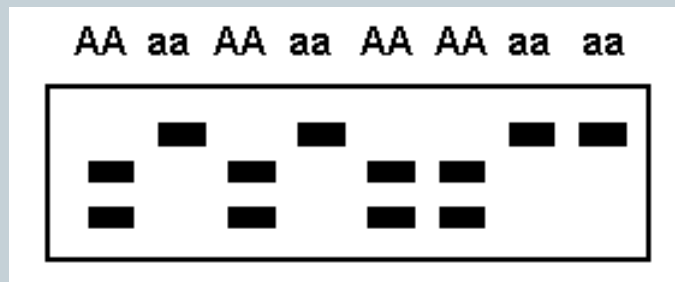
$$h^2 = V_A/V_P$$

- ❖ 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

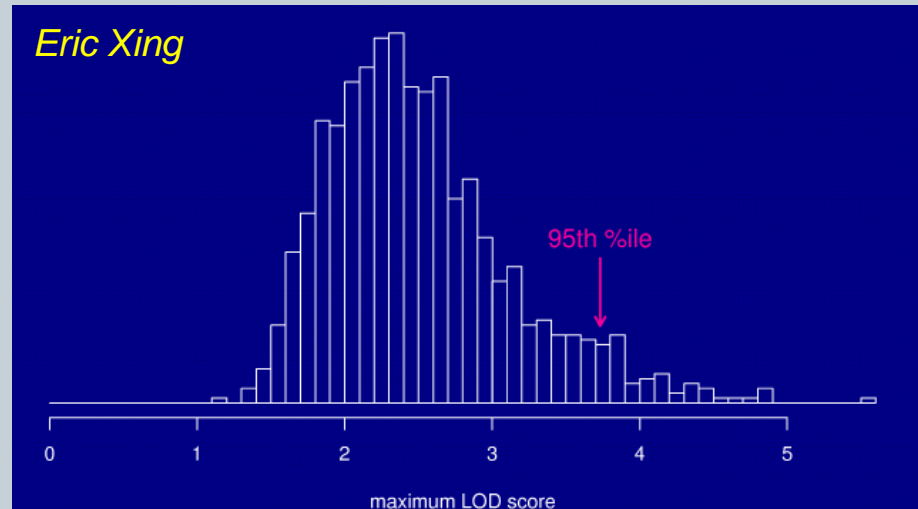
**Heterozygous effects:** The difference between the mean of the trait in the heterozygous genotypes from the average of the means of the trait in the two homozygous genotypes.

# 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

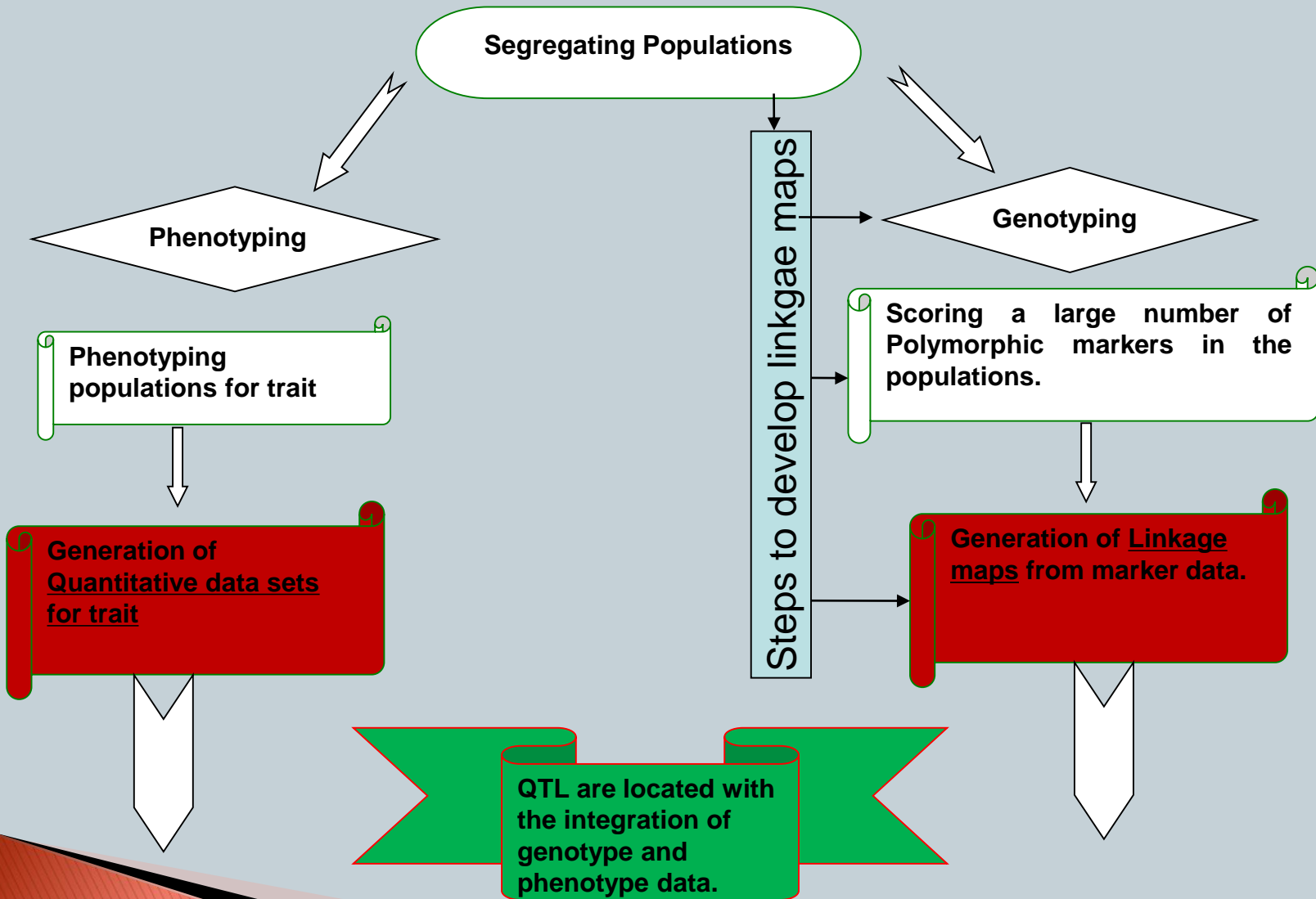
## Location of QTLs

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

When size of the interval to localize the QTL decreases, the number of individuals required to detect the recombinants in the interval and number of molecular markers 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





# 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

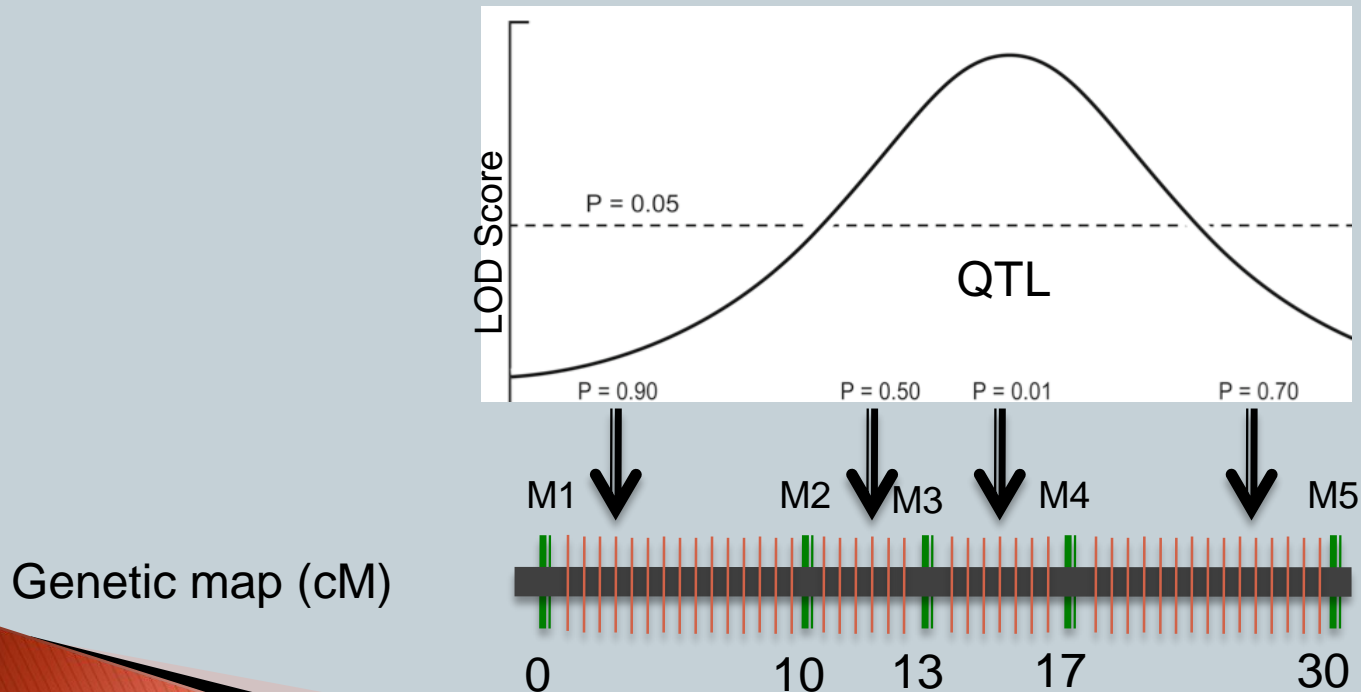
- ❖ Cannot predict true QTL location and QTL effect
- ❖ Missing values at the marker are discarded
- ❖ The power for QTL detection decreases

# Interval mapping

(Eric Lander and David Botstein)

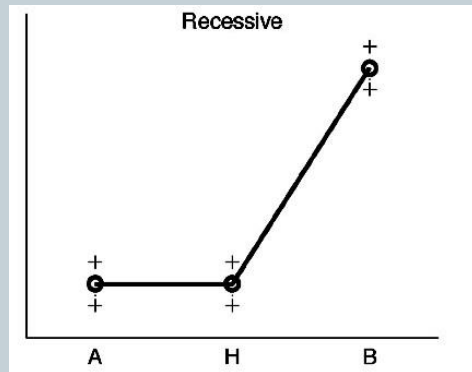
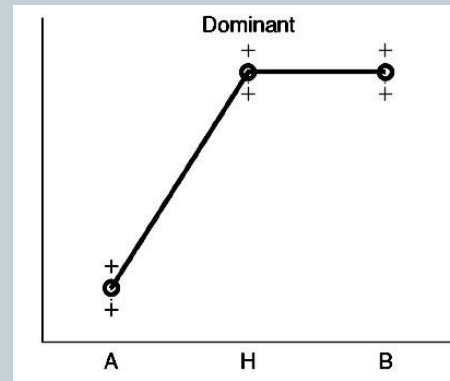
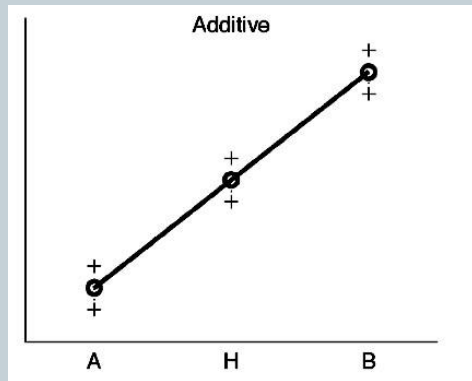
Most popular method

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



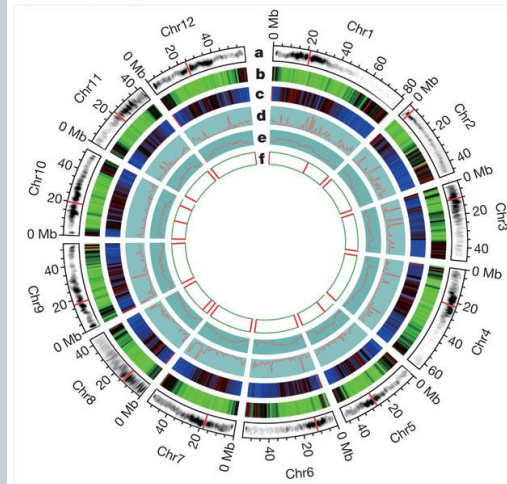
- ✧ Crossing over and recombination provide basis for genetic (linkage) mapping
- ✧ Relationship between recombination and genetic distance
- ✧ Genetics of quantitative traits
- ✧ Genetic locus controlling quantitative trait (QTL) can be identified by genotyping molecular markers and phenotyping the 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.



0.0	17615473	7532872	7533262	7533266	7539661
1.1	7569383				
2.2	7560044				
3.7	17560974	75616278			
4.1	14313404				
8.5	17550467	7564865	7625361		
10.0	7560816				
10.4	14313415	14313457			
11.5	7575048				
12.6	17622617	14313402			
15.6	17539336	7541837			
16.0	14314317	7568252	7628763		
16.7	7540032				
19.7	17528995	7601523			
20.8	7572391				
21.2	17624797	14314262	7544236	7550240	7559693
23.8	17567096	7544666	7626947		
26.7	17566042	14313387	7536348	7545496	7625487
27.5	17546304	7547216	7536566	7606397	7607010
29.0	14314228				
29.7	17562730	7624081	14313436	14313814	14313990
	17545236				
30.4	17544200	7544381	7549636	7563107	14313910
	14314084				
30.8	17551681	7603887			
31.2	3841658				
31.5	7529443				
32.3	17561450	7561452			
32.7	17532611	7602949			
33.8	7566164				
34.9	3842471				
35.3	17625145	14314214			
38.6	17526738	9836083			
	17558971	7565128	7548682	7566127	7513879
39.3	7519606				
41.2	17569565	7625054	7531033	7531218	7552321
	7568311	7563566			
42.3	7524939				
47.1	17560127	7560129	7570728		
47.5	14313234	7533055	7540604		
47.9	17532945	7533022			
48.6	7533197				
49.4	7552941				
	7541913	7543118	7556261	7564257	7556267
49.7	7568967	7572771	14314102	7564258	7603828
	7625200	7629014			
50.1	17567403	14313284	7567404	7632071	
50.8	17537253	7538687	7615271	7631544	
51.6	7571148				
53.1	17603551	7603812	9840503		
70.4	7532748				
72.6	7536323				
73.4	17530081	7623115			
	7526270	7528002	7530803	7543439	7554660
74.8	7603246	7605044	7617207		
75.2	14313252	7572005			
77.1	14313086				
78.2	7006322				
78.5	7557404				
78.9	14314035	7568887	7631639	11823940	7540858
	7560886	7569859	7573332		
80.8	17555047	7559078	7619638		
81.5	17559237	11826592	14313278	7570665	
82.6	14313338	7606480			
83.4	17536759	7602035	7603779		
84.1	17540706	7616344			
84.5	11824328	7549419	7558044	7545448	7551255
	7568054	7569882	7600888	7613432	7621605
84.8	7569131				

SweetPotatoBase

search manage

Maps & Markers

1  
C233  
C015  
C2\_at4g15790

Breeders Toolbox

HIDAP4RTB

Phenotype

Genotype

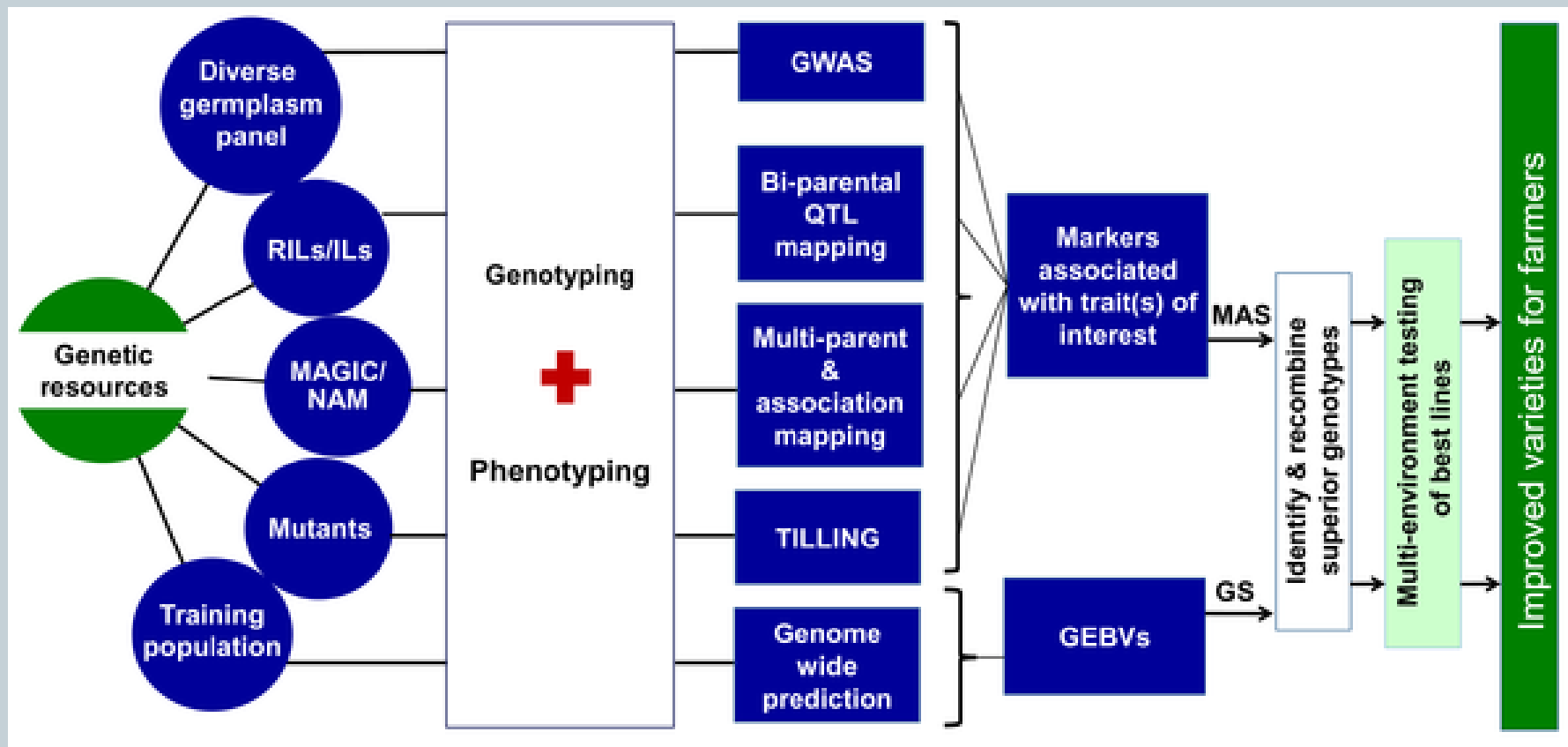
Environment

Integration

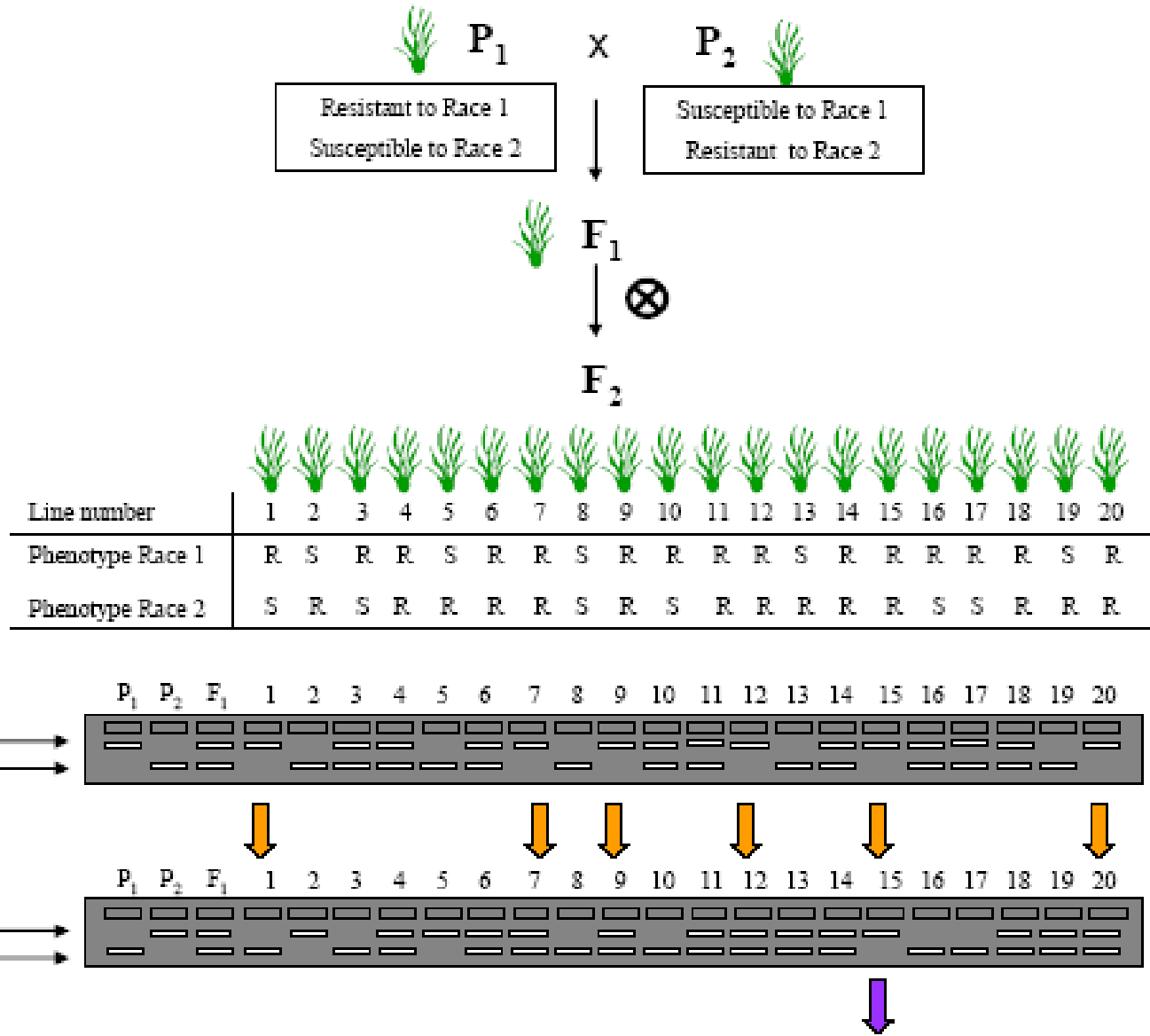
Supporting information

Sharing

# New paradigm of Genomics-assisted breeding

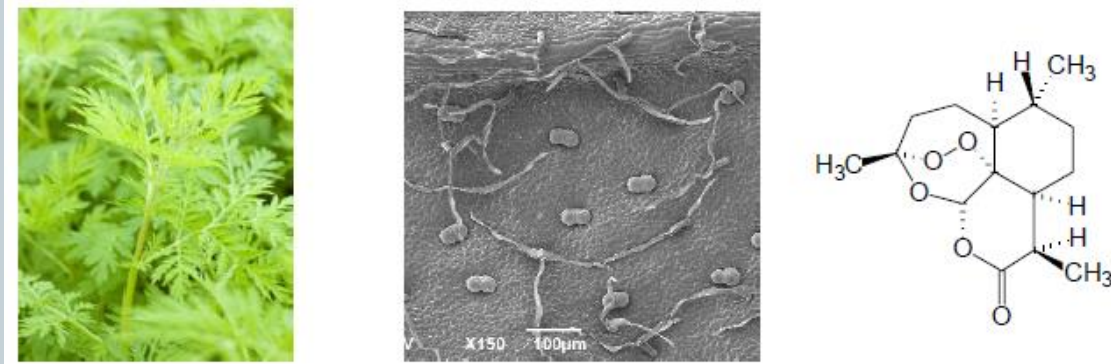


# Marker-assisted breeding



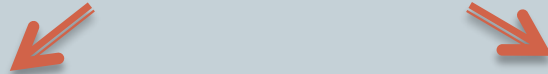
**Thank for your attention**

# *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

Natural genetic variation



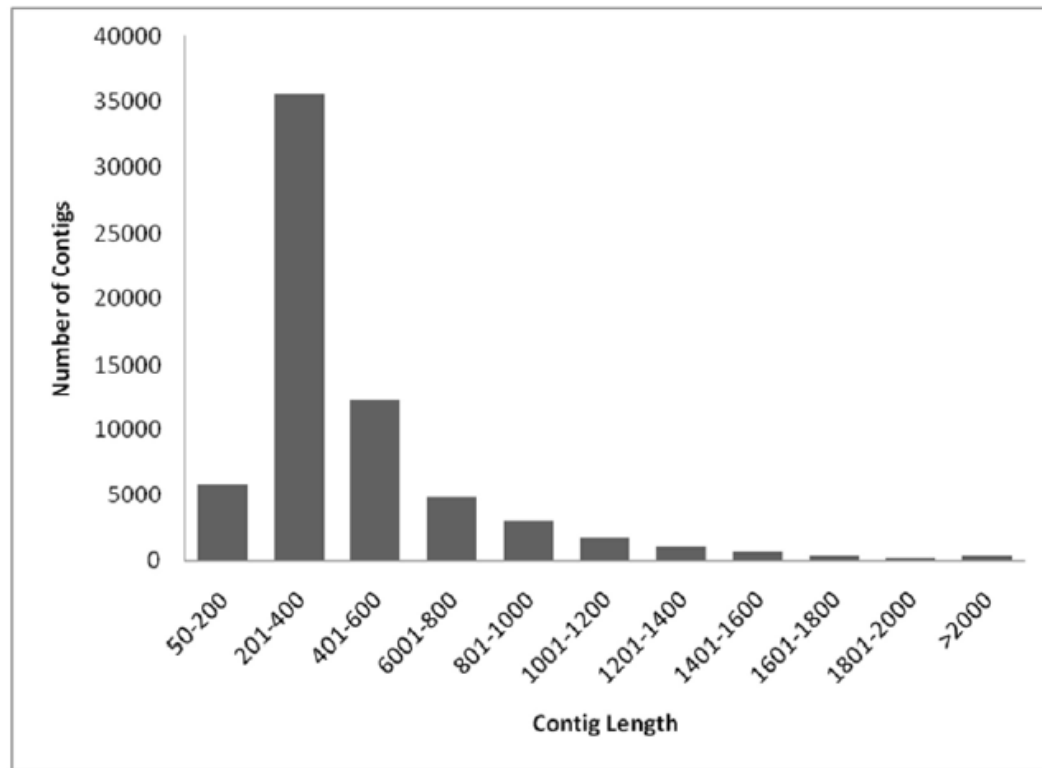
EMS-treated population

Natural populations



# Molecular marker development: Illumina GoldenGate SNP array

## Genomic libraries and DNA sequencing

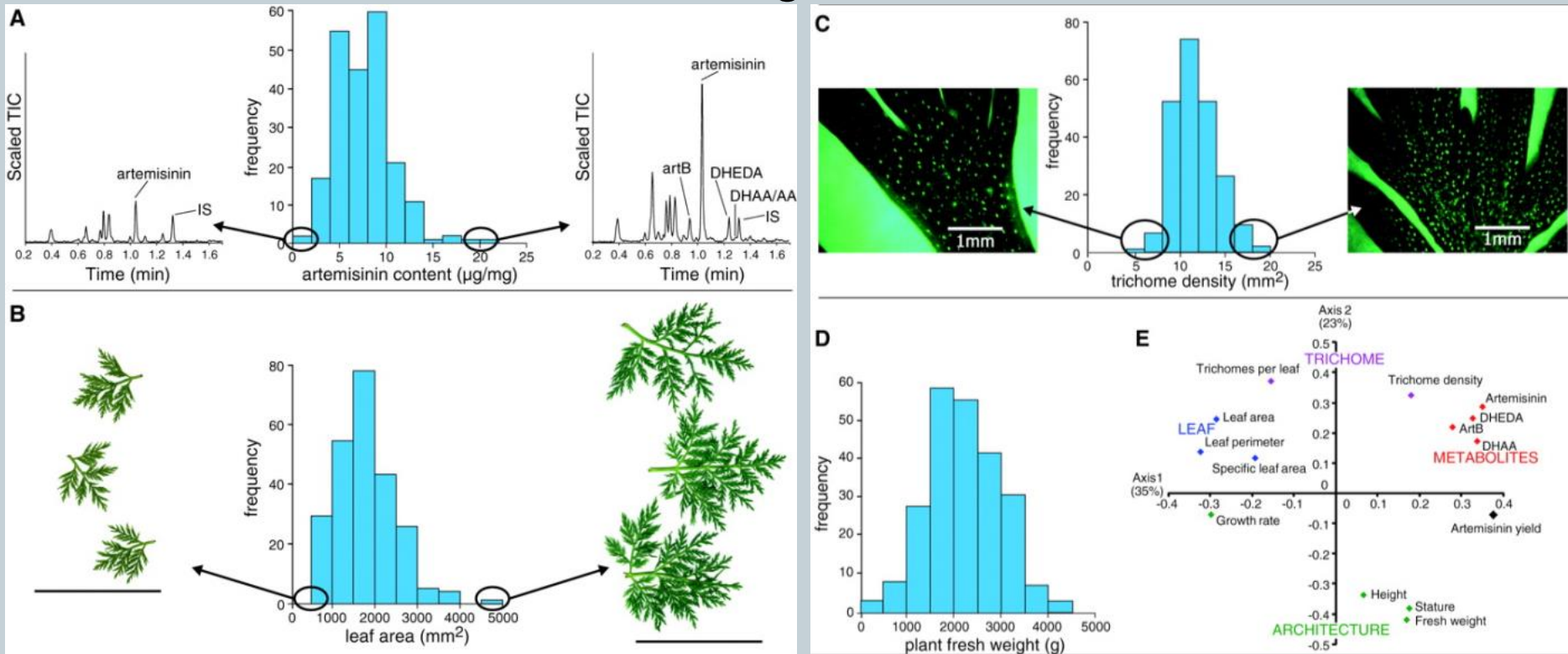


- Trichomes of high and low yield plants
- Trichomes from young and old leaves

- 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

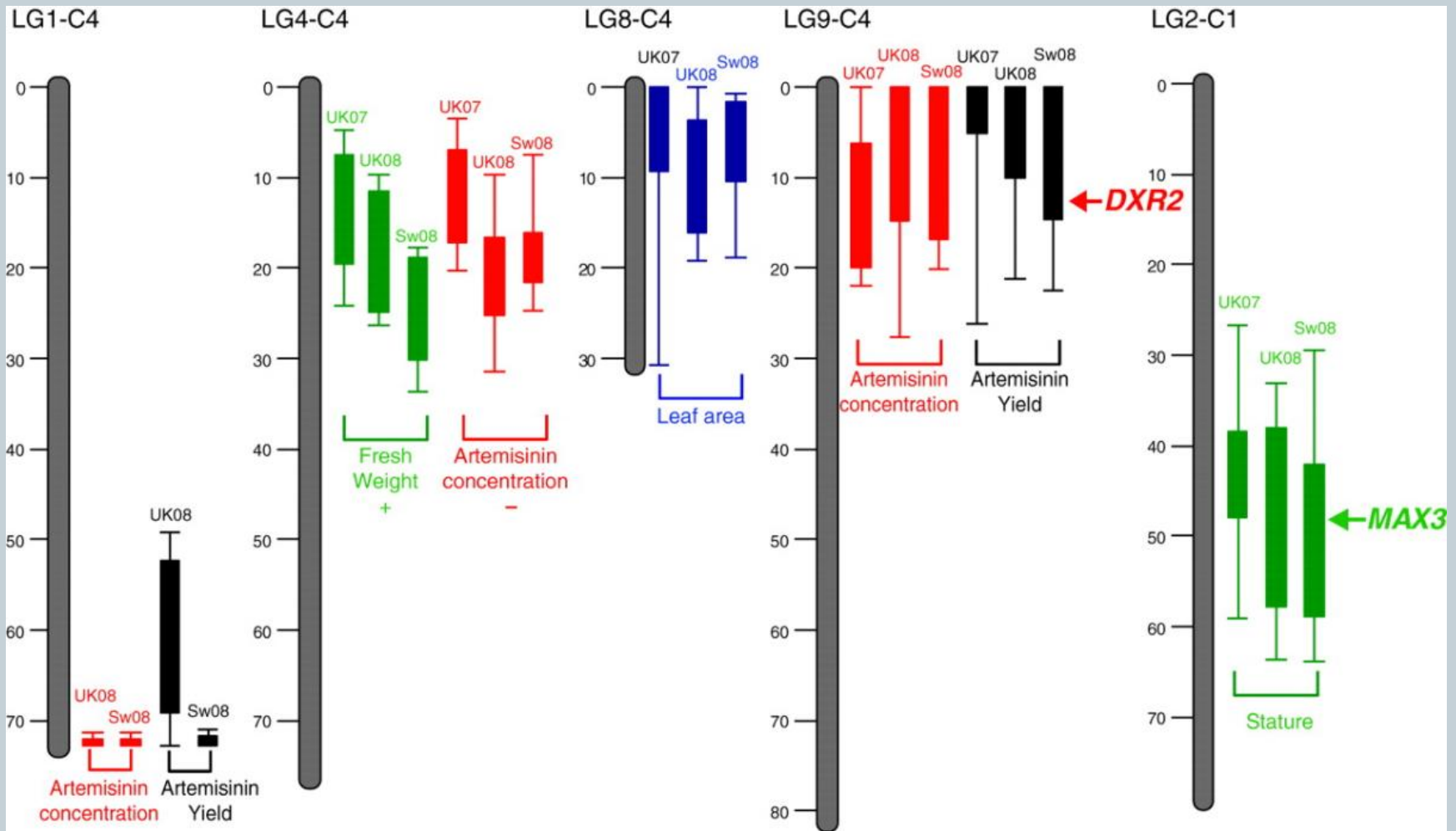
Field experiments conducted in UK, Switzerland and Madagascar



**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

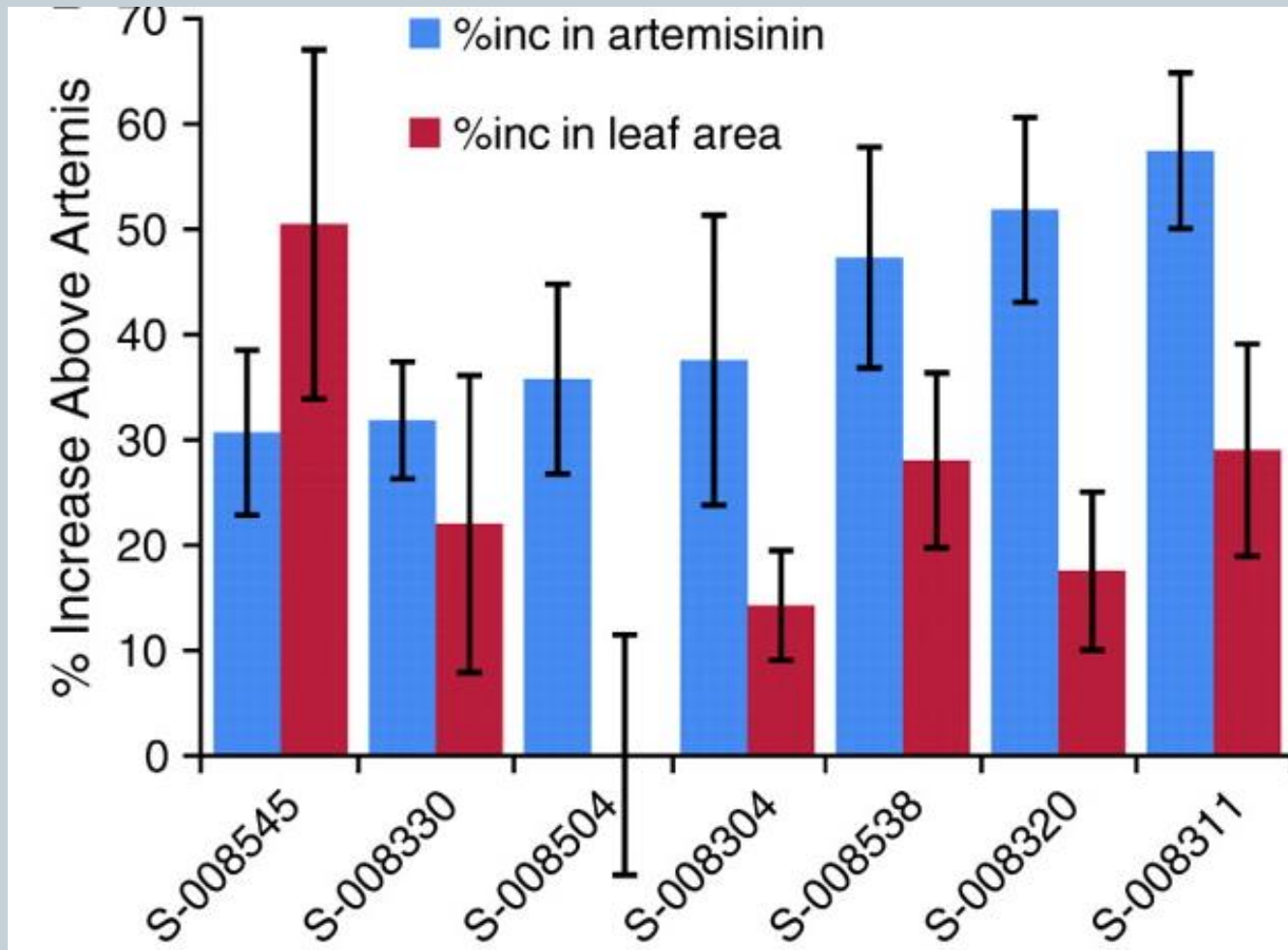
# QTLs



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



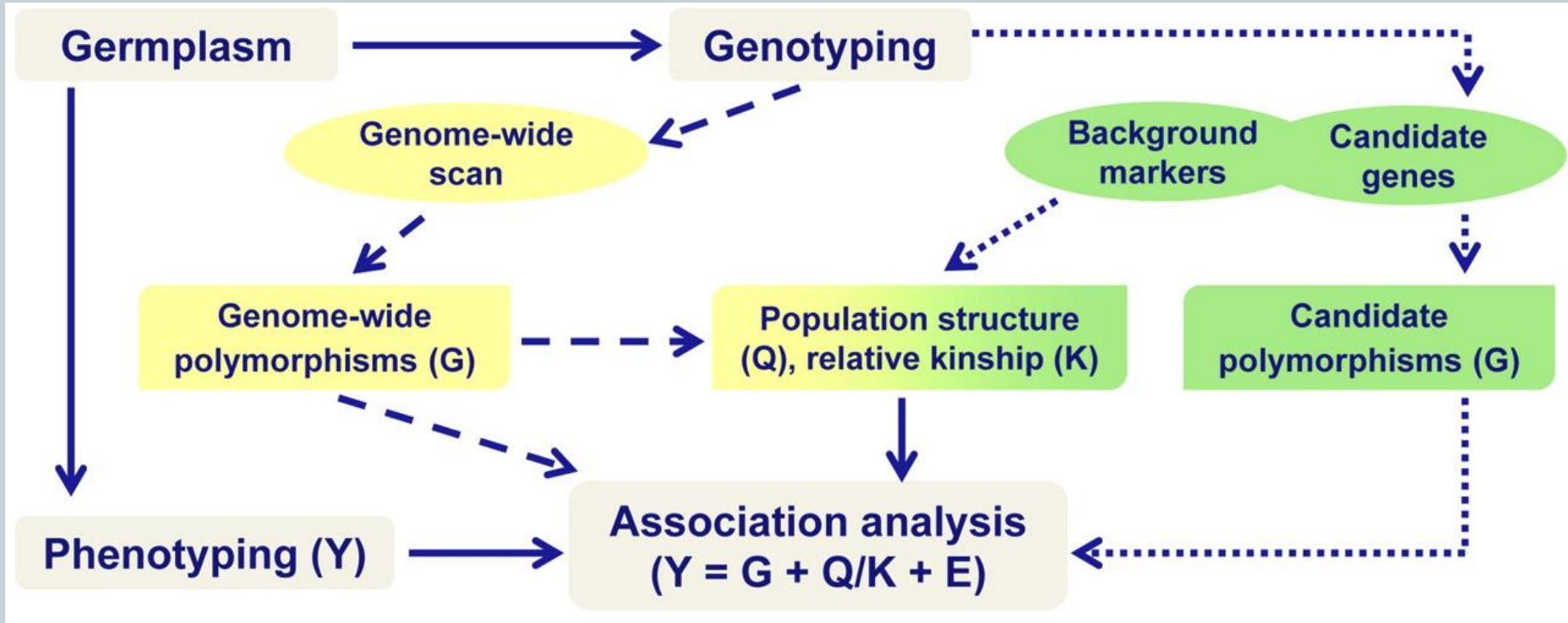
# Use of MAS in hybrid production



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.

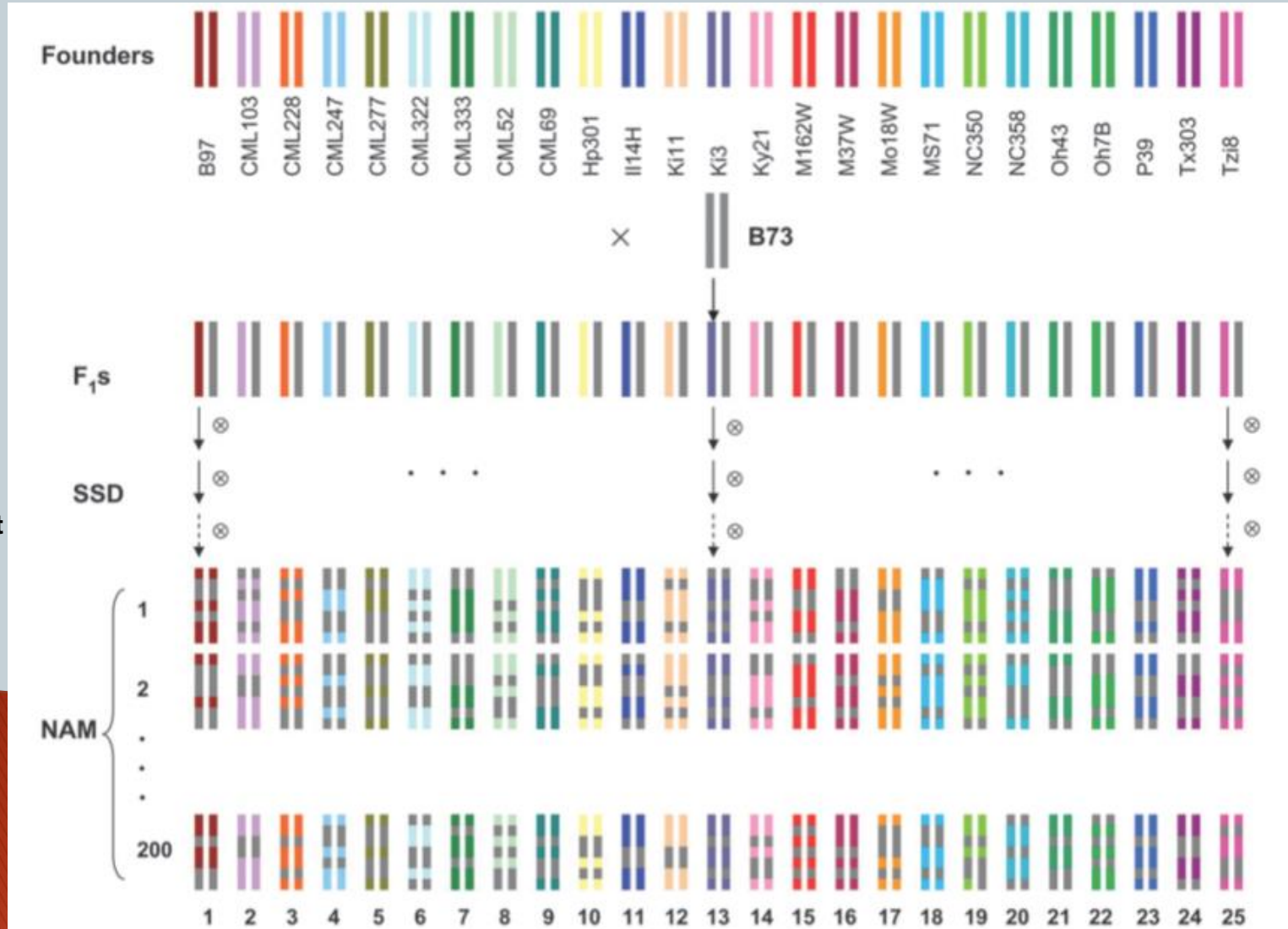
Graham et al. *science*, 2010

# Overview: Association mapping analysis



Zhu *et al.* 2008

# The Maize Nested Association Mapping Population (NAM)



SSD=Single seed descent

# Predicting the phenotype: Genomic selection

