Genomic-Assisted Breeding in Sweetpotato: Current status and future opportunities

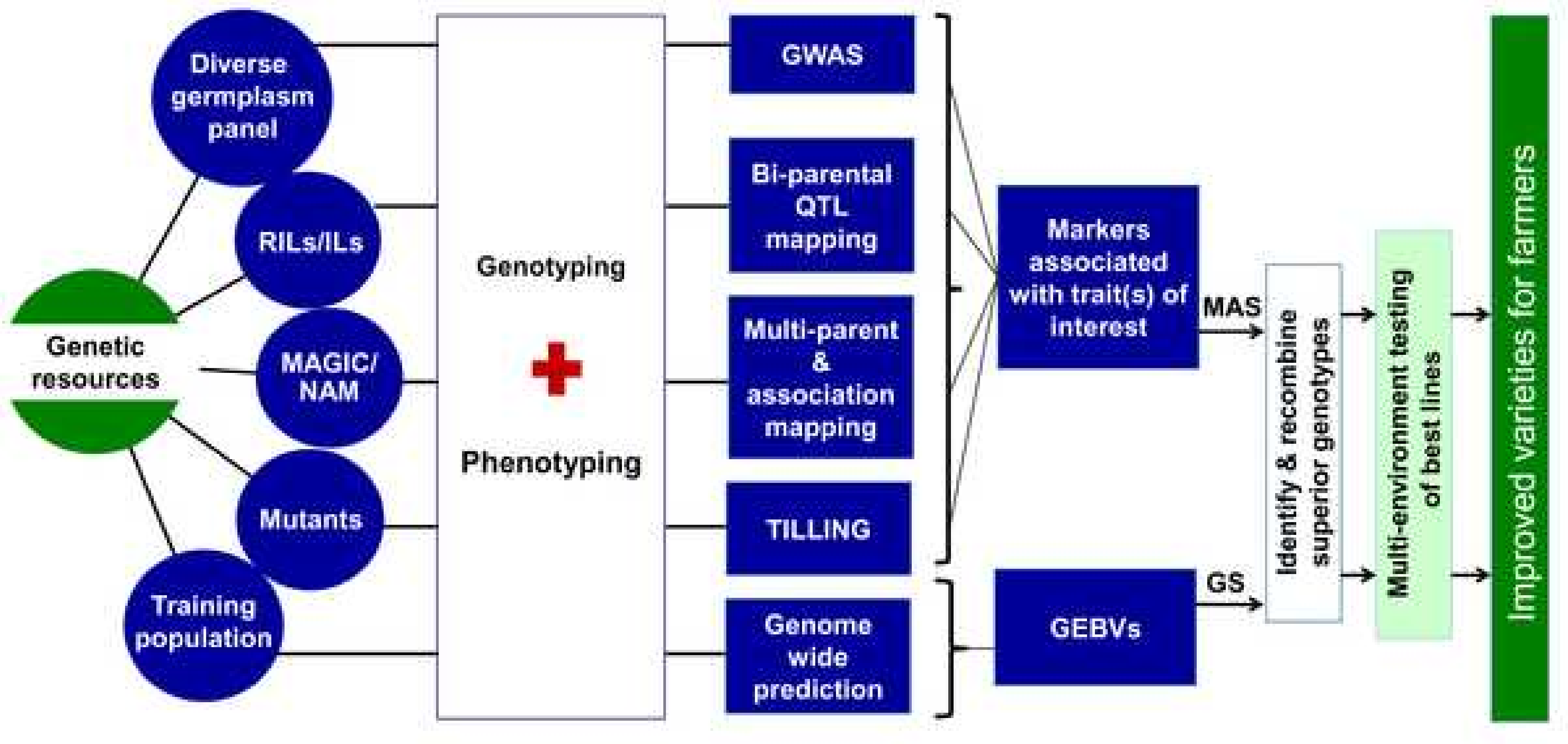
Awais Khan
awais.khan@cgiar.org

International Potato Center (CIP), Lima, Peru

June 2, 2015

CENTRO INTERNACIONAL DE LA PAPA
CIP
A member of the CGIAR Consortium

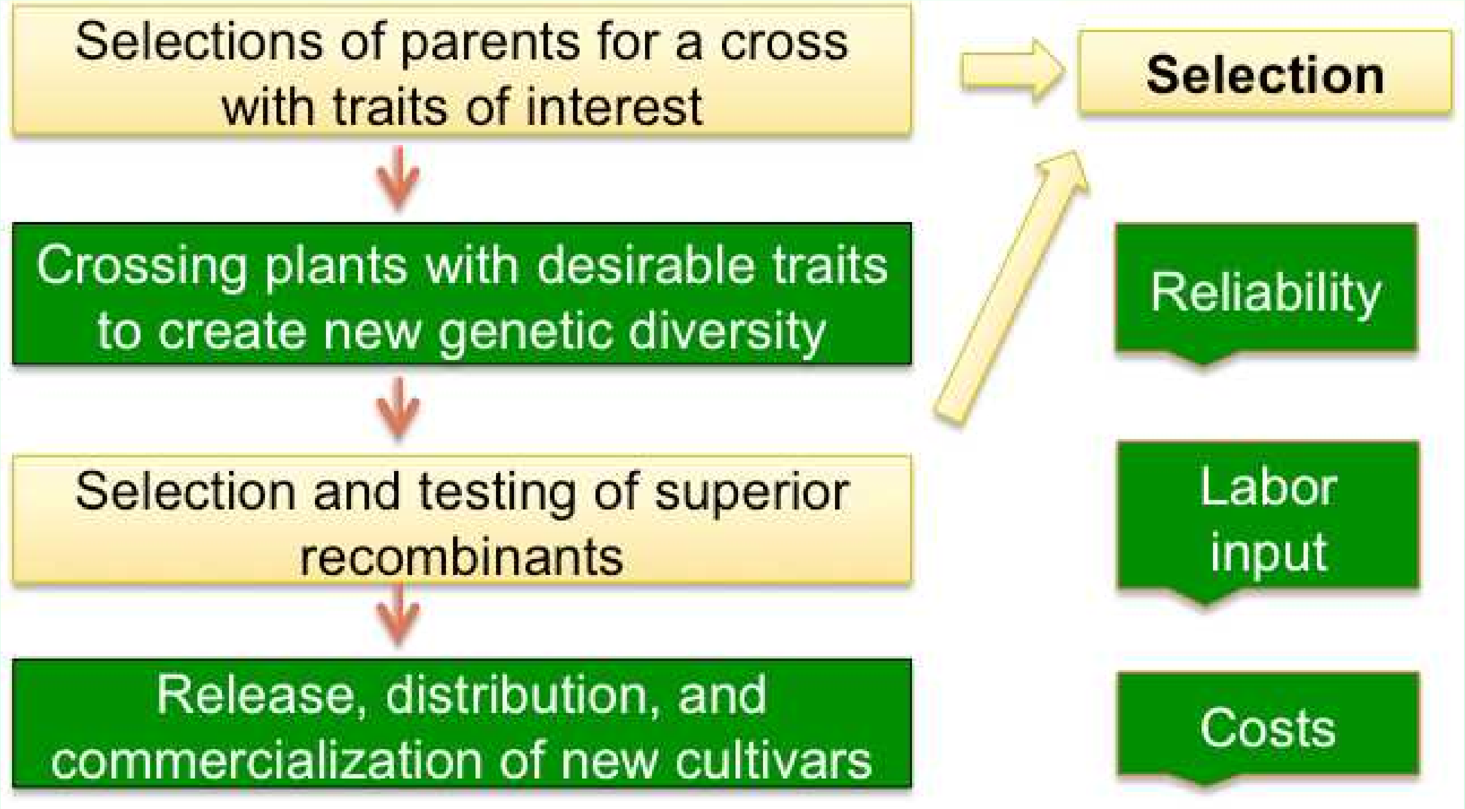
New paradigm of Genomics-Assisted Breeding



Varshney *et al*. 2014

Importance of selection in 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.



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

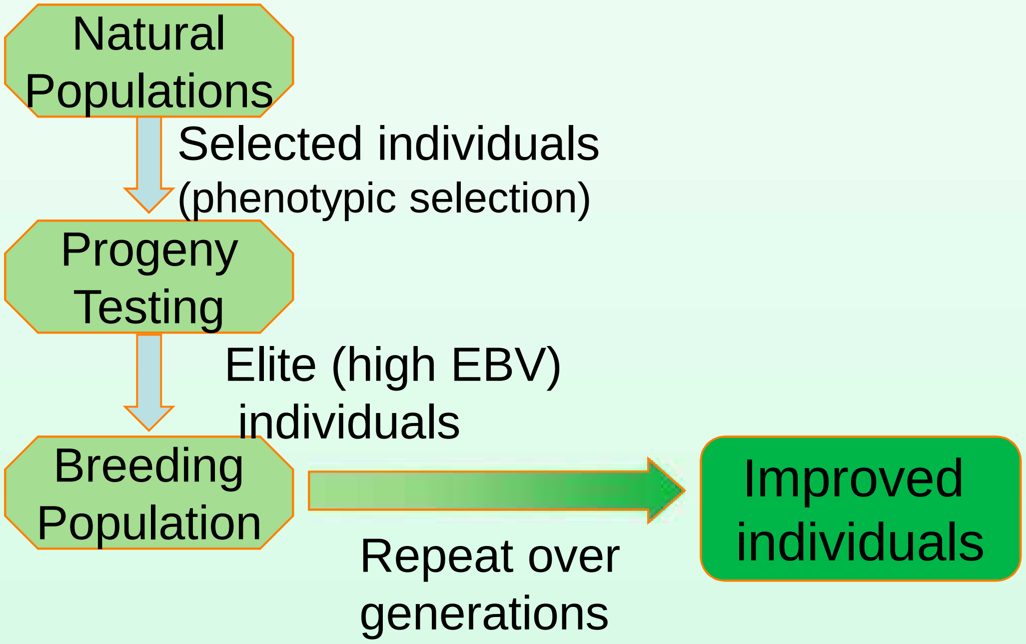
Reliability

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)

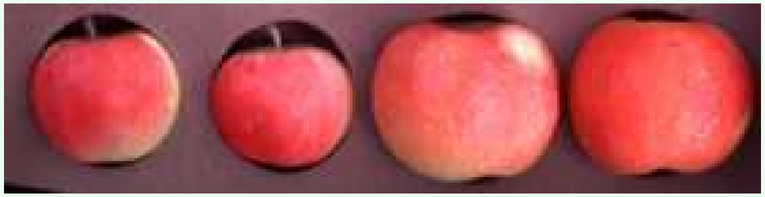
Progeny Testing

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

DNA based selection methods

A. **Marker-assisted selection:** Selection for one or more (up to 8-10) alleles

B. **Marker-assisted backcrossing:** One or more (up to 6-8) donor alleles are transferred to an elite line

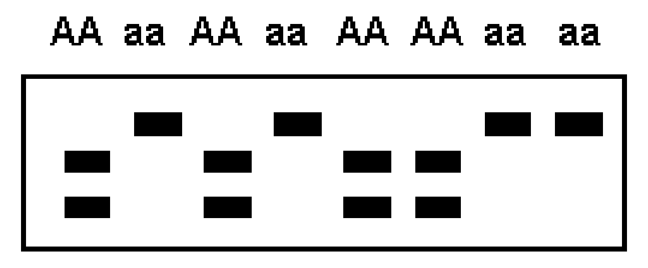
C. **Genome-wide selection:** Selection of several loci using genomic estimated breeding values (GEBVs) based on genome-wide marker profiling

Trends Plant Sci 10: 621 630

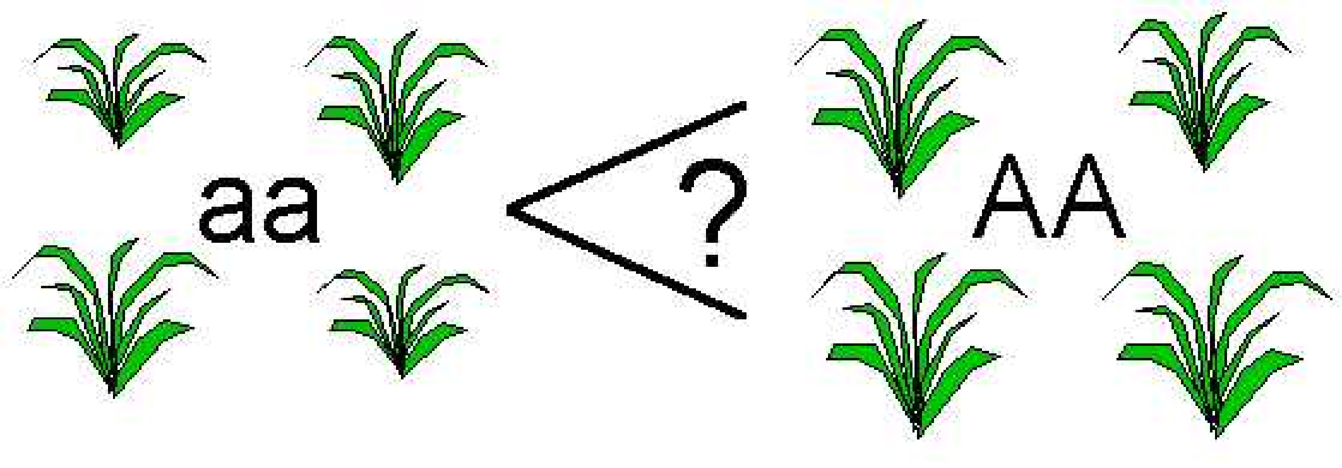
QTL analysis

Is there a significant link between genetic makeup (genotype) and trait phenotype?

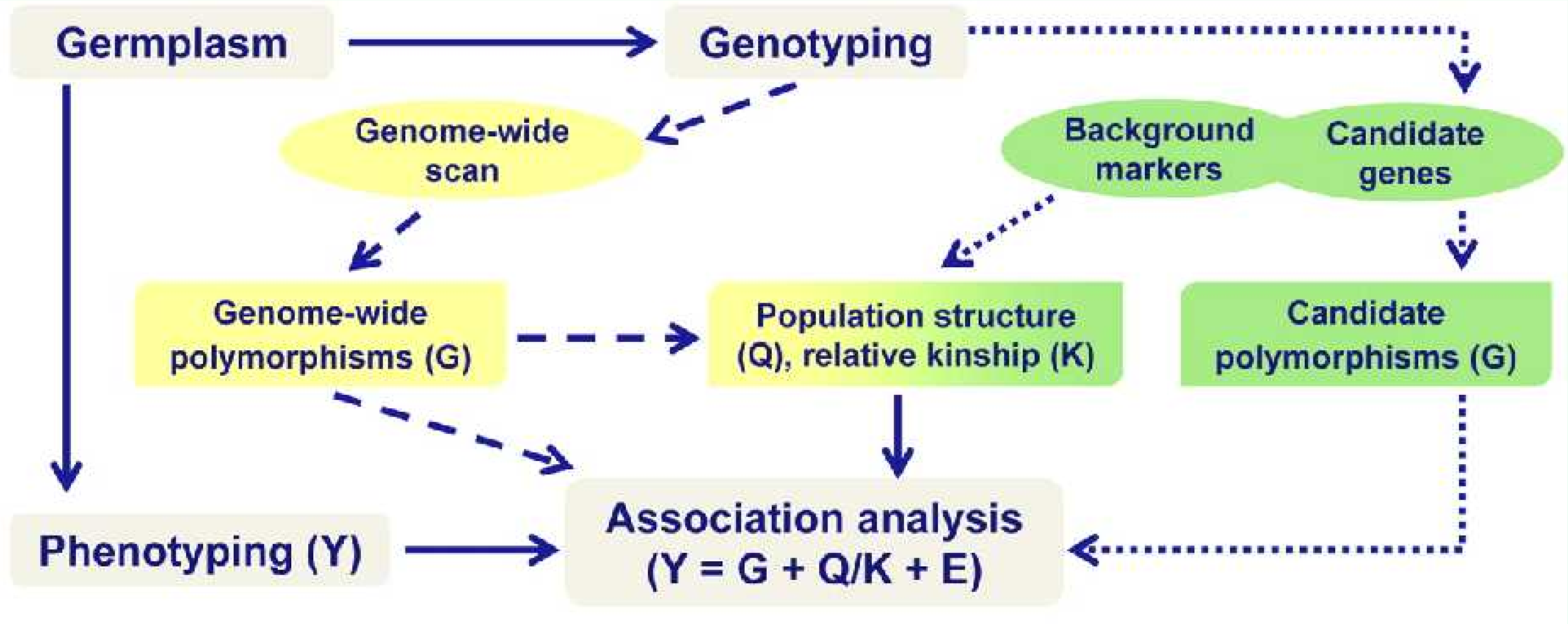
Single marker Analysis



Marker genotype

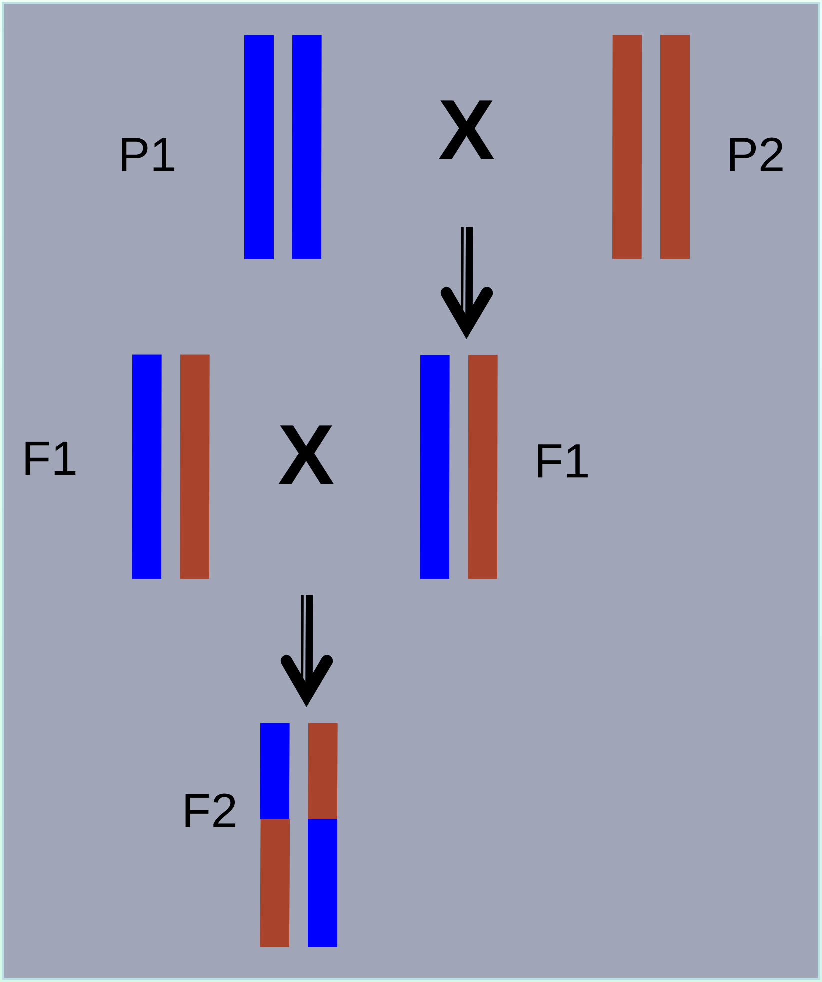


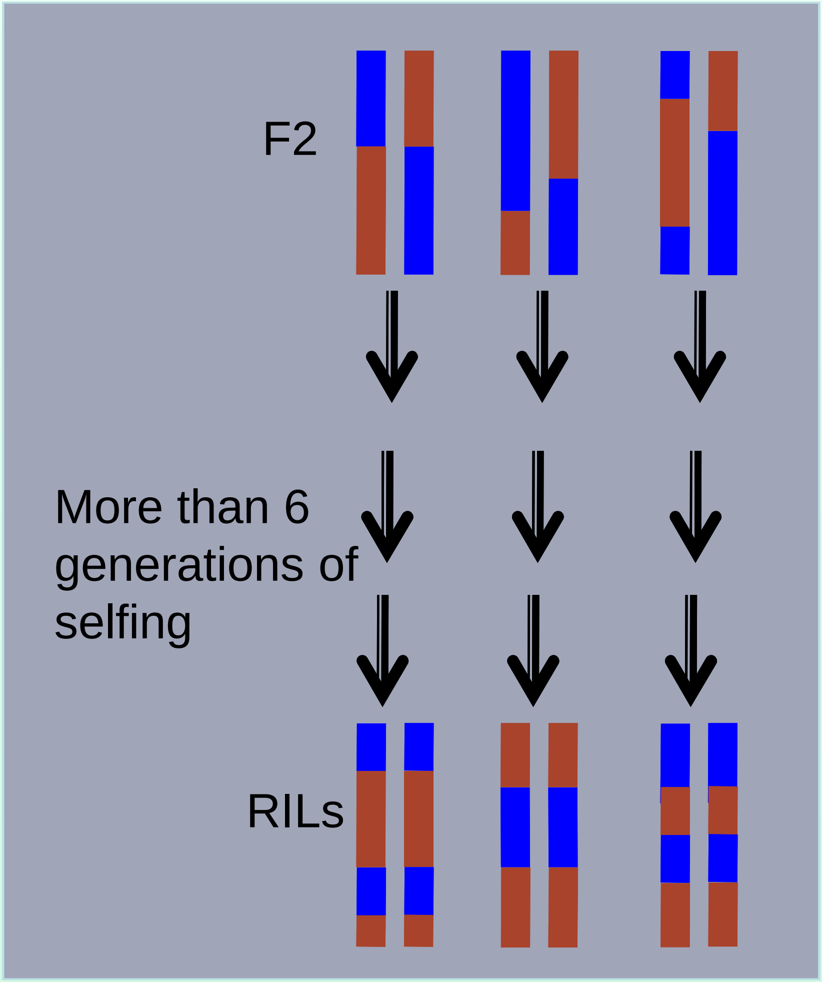
Overview: Association mapping analysis



Zhu *et al.* 2008

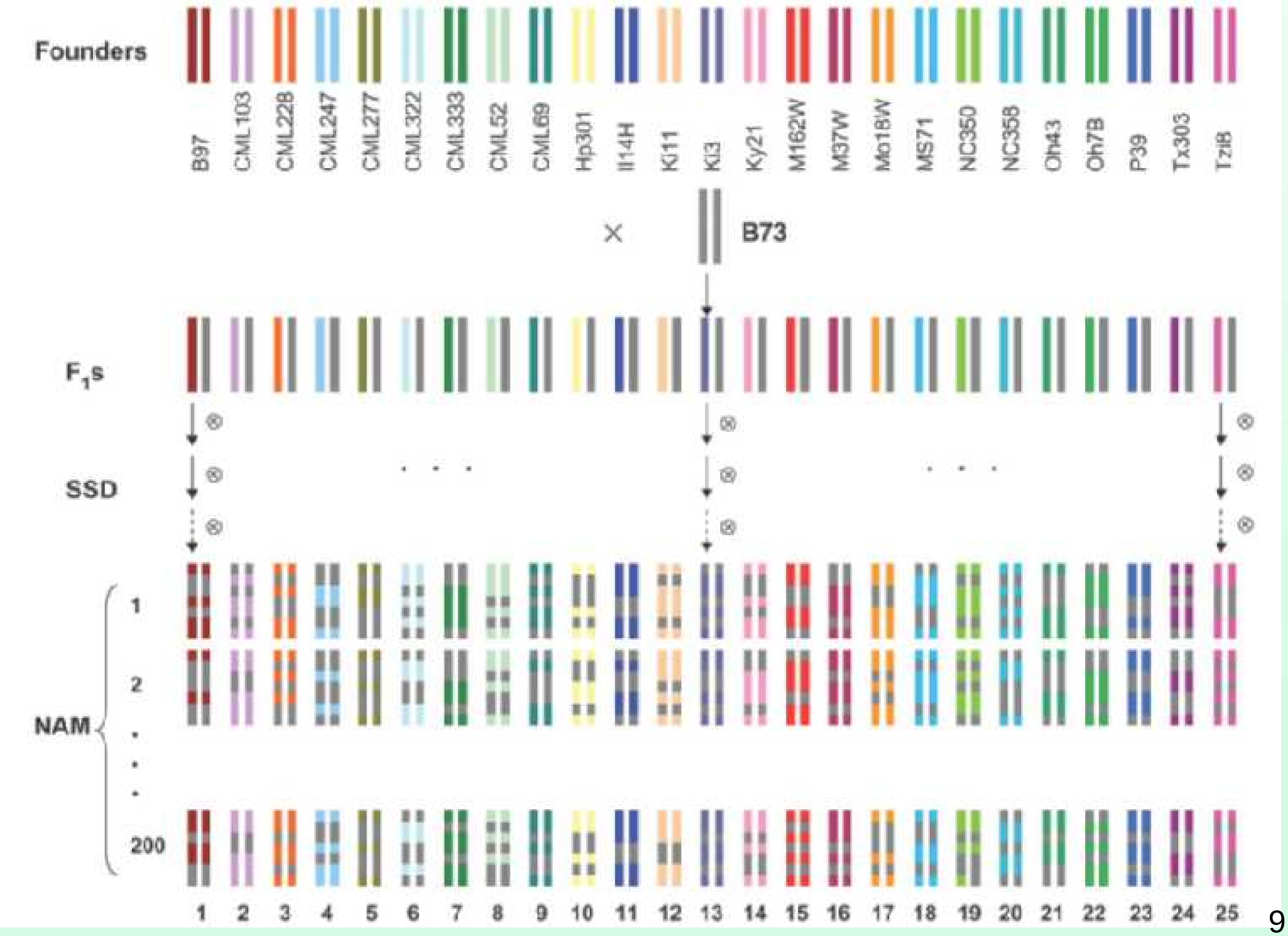
Segregating population





The Maize Nested Association Mapping Population (NAM)

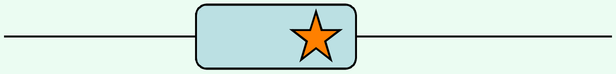
SSD=Single seed descent



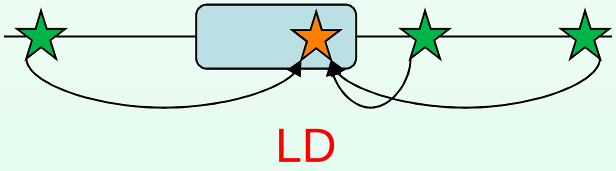
Concept of Marker assisted selection

Molecular breeding

Association between molecular marker and causative gene



Direct association



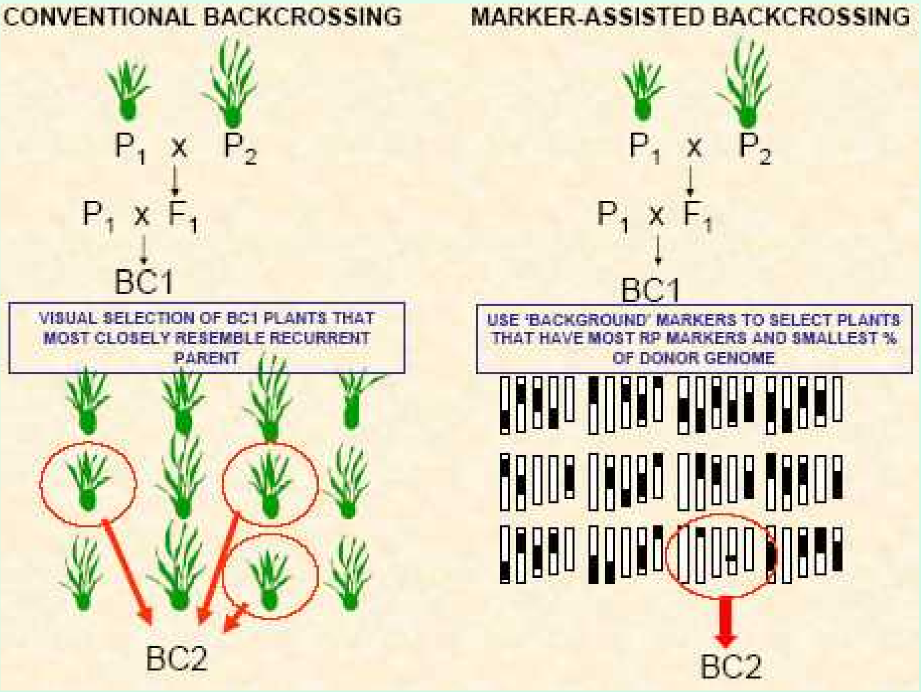
Indirect association

🟦 Causative gene

★ SNP within gene

★ SNP in LD with gene

Hirschhorn & Daly, 2005


CONVENTIONAL BACKCROSSING

VISUAL SELECTION OF BC1 PLANTS THAT MOST CLOSELY RESEMBLE RECURRENT PARENT

MARKER-ASSISTED BACKCROSSING

USE 'BACKGROUND' MARKERS TO SELECT PLANTS THAT HAVE MOST RP MARKERS AND SMALLEST % OF DONOR GENOME

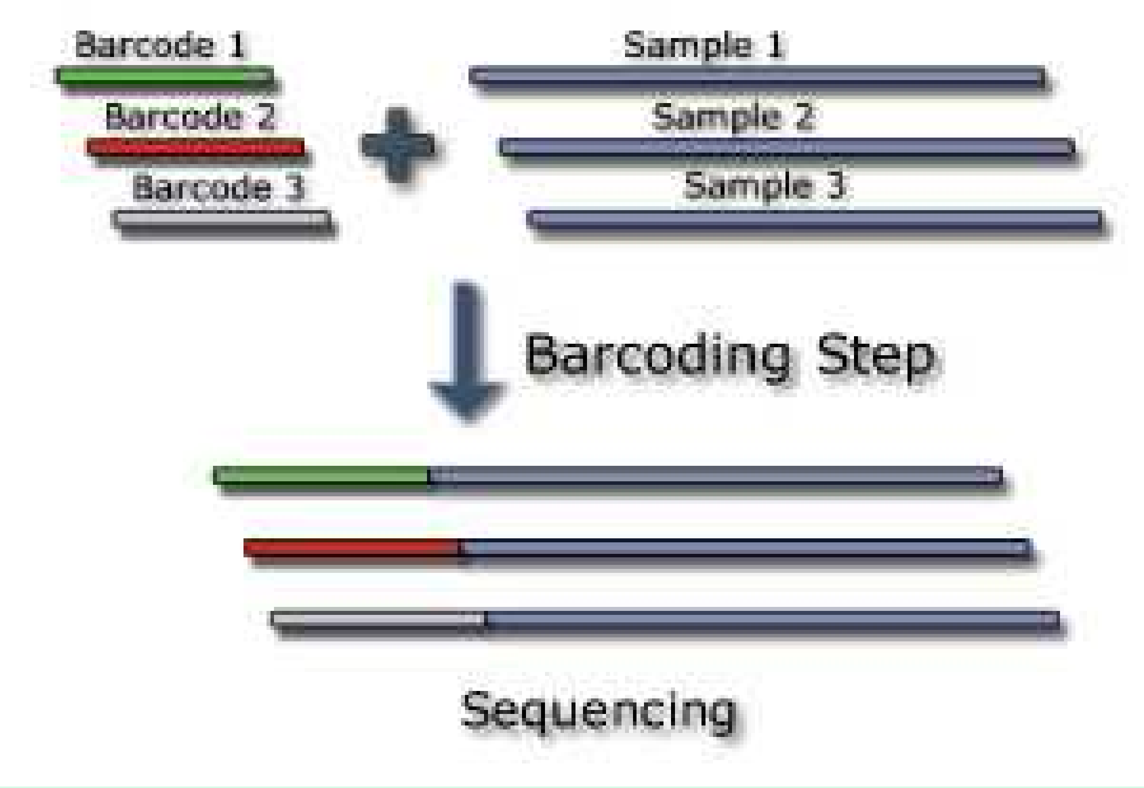
DNA marker technology coupled with Next-Generation Sequencing (NGS)

Cost and throughput comparisons

Sequencing of 3 Gbase genome to 18X coverage (54 Gbases)

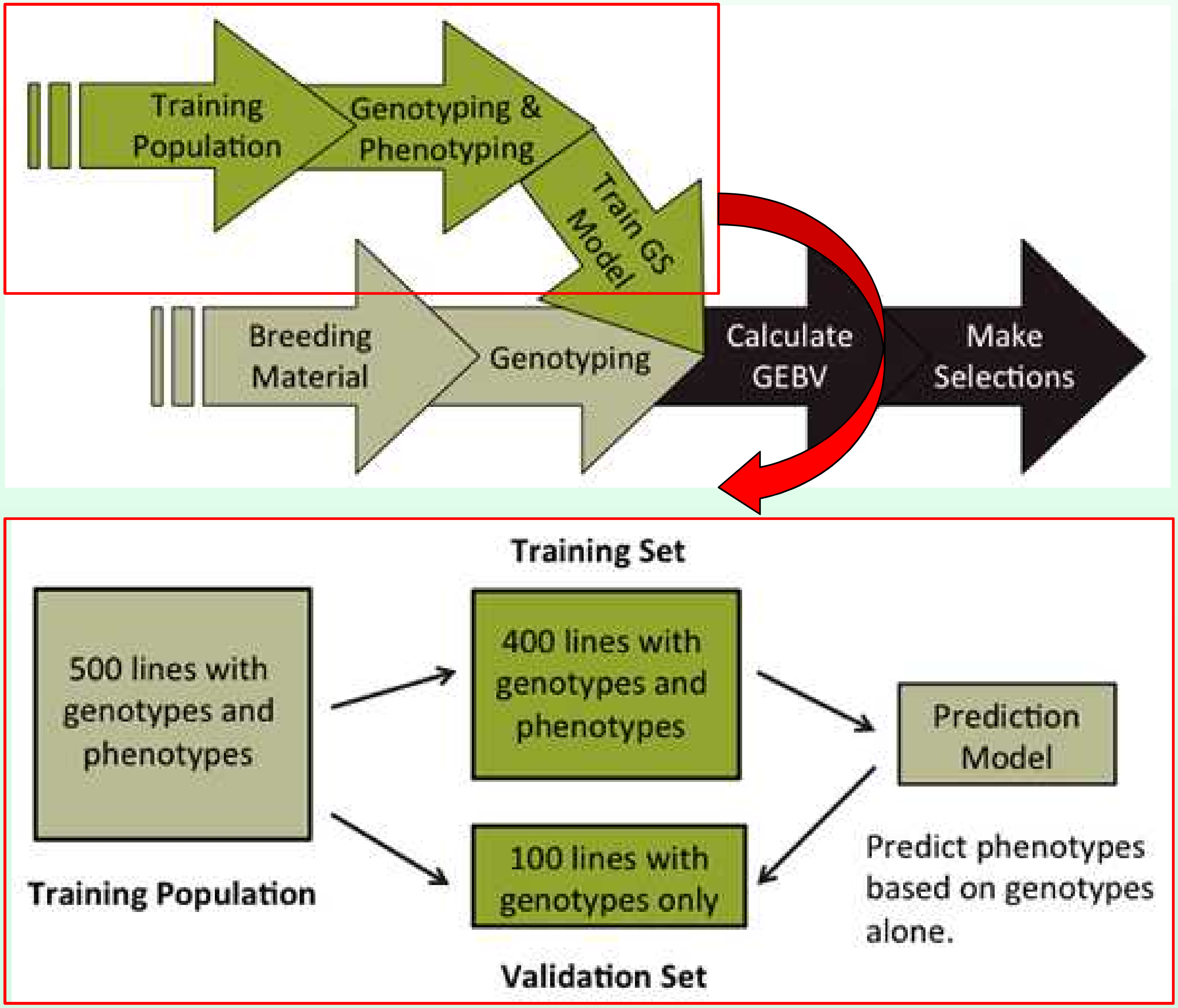
	Sanger	454	Illumina
No. of plates:	756,000	120	3
Time:	48 years	6 months	2-3 weeks
Total cost:	$108 millions	1 million	$60k
Cost/Mbase:	$2,000	$18.5	$3

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.

Predicting the phenotype: Genomic selection



Prediction Accuracy of Genomic Selection

Correlation between

GEBV
(Genomic estimated
breeding value)

EBV
(Experimentally estimated
breeding value)

Affected by:

1. LD between markers and QTLs (↑LD)
2. Size of Training population (↑n)
3. Heritability of the trait in question (↑h^2)
4. Genetic structure of the trait (↓#QTLs)

↑ Accuracy of GS

Hayes et al. 2009

Sweetpotato (*Ipomoea batatas)*

- Tropical root crop from the morning glory family (Convolvulaceae)
- Fifth most important crop in developing countries (after rice, wheat, maize, and cassava) to prevent malnutrition, and enhanc food security
- Perennial, grown as annual, propagated - storage roots or stem cuttings
- Orange-fleshed sweetpotato - important source of β-carotene (a precursor of vitamin A)

Sweetpotato: Challenges for genetic studies and breeding

- Hexaploid species with $2n = 6x = 90$

- Autopolyploid – Allopolyploid!

- Self-incompatibility, out-crossing, and highly-heterozygous, within plot variation is extremely high that needs a large number of replicates

- Cultivar development - Phenotype-based selection

Sweetpotato: Genetic and genomic resources

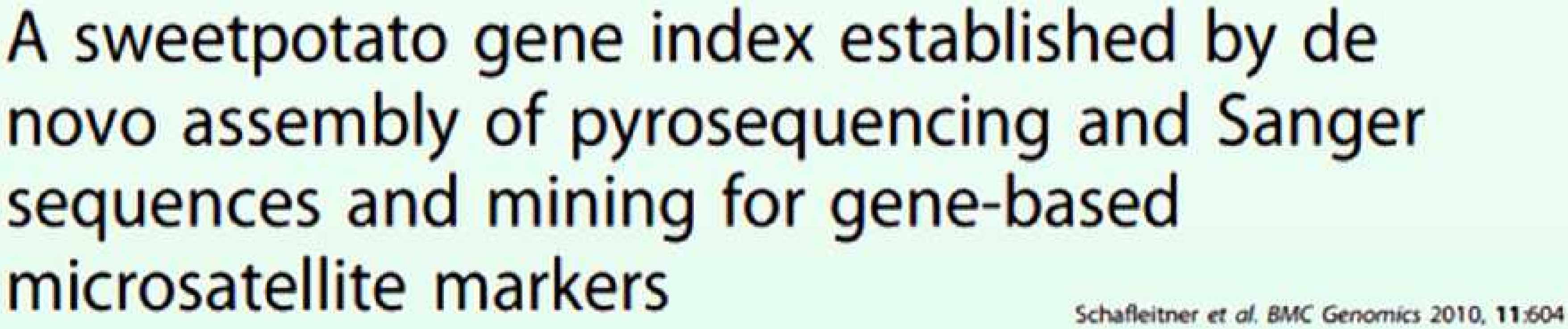
- **7,783** accessions at CIP’s genebank (including breeding lines, improved varieties, landraces, and wild accessions)

- Great experience with breeders and scientists at CIP, NARs in SSA, NCSU, LSU, China, Japan and Korea but limited community compared to many other crops

- Few genetic mapping populations and limited genomic resources – So far few linkage maps with SSRs and AFPL markers are available (e.g., Ukoskit et al. 1997; Kriegner et al. 2003; Cervantes-Flores et al., 2008)

- QTL mapping in Beauregard and Tanzania: Root-knot nematode resistance, beta carotene content, dry matter, starch content, maltose and sucrose content, iron, zinc, calcium content (Cervantes-Flores et al., 2008, 2011, Chang et al. 2009)

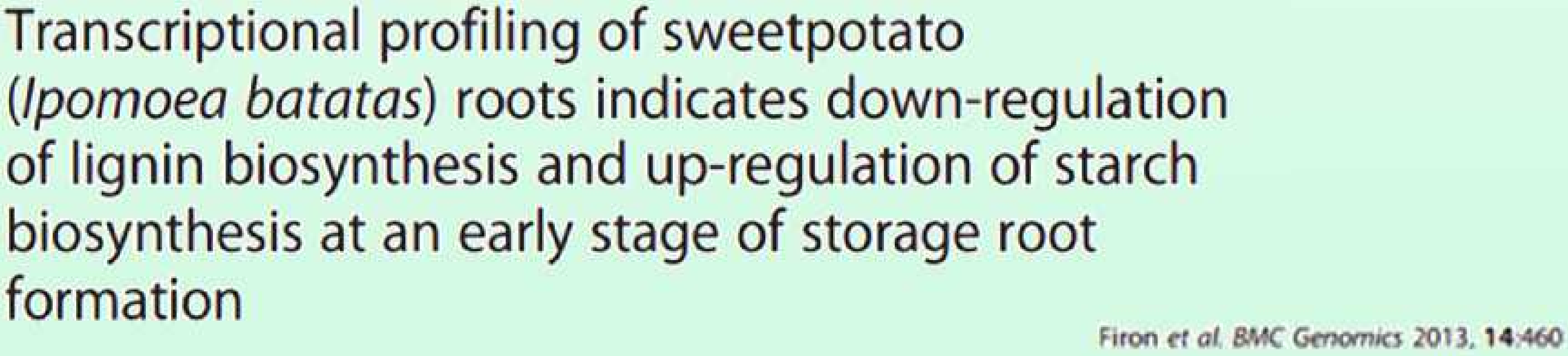
Sweetpotato Gene Index



A sweetpotato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers

Schafleitner *et al. BMC Genomics* 2010, **11**:604

Sweetpotato Gene Index from Root



Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation

Firon *et al. BMC Genomics* 2013, **14**:460

Sweepotato genes and genome

46 SSR marker based kit identifying1029 alleles, 5 to a 23 alleles, averaging 11.5 alleles per marker. Rossel *et al*. 2014

A partially complete genetic linkage map (43 and 47 LGS) based on retrotransposon insertion polymorphisms.
Monden et al. 2015, Breeding Science

Transcriptome of two sweetpotato cultivars Xushu 18 and Xushu 781. PAG 2014

Whole-genome sequencing (*de novo)* of two lines of *I. trifida*, using the Illumina HiSeq platform. Assembled genome sequence from 513-712 Mb. 62,407 and 109,449 putative genes, 1,464,173 SNPs and 16,682 CNVs. Hirakawa *et al*. 2015

"*Transcriptome Analysis and Genome Wide Association Studies in Sweet Potato (Ipomoea batatas L. Lam)" (2014-2017)*: USDA funded project at University of Arkansas at Pine Bluff.
http://www.reeis.usda.gov/web/crisprojectpages/1004277-transcriptome-analysis-and-genome-wide-association-studies-in-sweetpotato-ipomoea-batatas-l-lam.html

Next generation sequence-based genotyping for *Ipomea trifida*

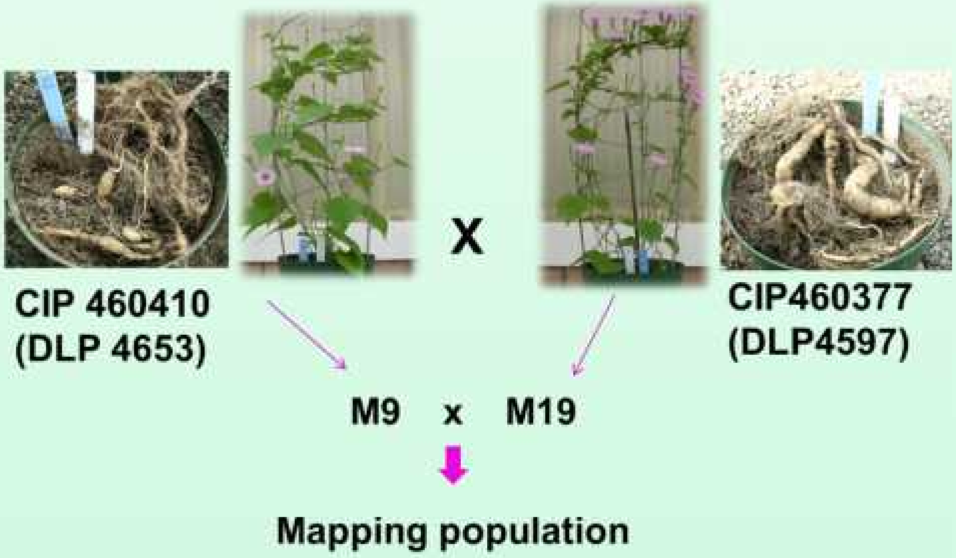
~ 3 Million sequence reads in total → and ~1.3 are good reads

SNPs without filtering → 5466

SNPs after eliminating NN→ 3643

SNPs after eliminating SNPs not matching between replicates→ 3210

SNPs that are polymorphic and segregating in the mapping population → 646



QTLs for sweetpotato relative to other crops

Crop	**Total “CROP" and "QTL"**	**“CROP" and "QTL" 2014**	**“CROP" and "QTL" 2000-2014**
Wheat	30900	4400: every possible category (seed size and shape, stripe rust resistance, yield, height, etc.)	17,500: protein, fusarium resistance, agronomic traits
Rice	34200 used as model	4770: many genes already identified associated with QTLs	18500: yield, grain size, abiotic stress tolerance
Soybean	19000	2750: protein content, biotic stress	1500: yield, abiotic and biotic stress tolerance, agronomic traits
Sugarcane	6110	609: not many QTLs actually in sugrcane	4400: sugar content
Potato	16,300	1750: tuberization, drought resistance, tuber shape, carotenoid content	12,400 results, mostly biotic resistance
Sweetpotato	4810	259: starch content, storage root yield	2470:yield, starch, biotic resistance, beta carotene
Cassava	2510	327: starch, viral resistance	2270:bacterial blight, plant architecture, root bulking
Yams	974	92	841

Source: Google scholar May 31, 2015

Genomics for sweetpotato relative to other crops

Crop	Total “CROP" and “Genomic"	“CROP" and “Genomic" 2014	“CROP" and “Genomic" 2000-2014
Wheat	127,000 Genome: 2012	13900	20600
Rice	131,000 Genome: 2002	15700	24700
Soybean	64,200 Genome: 2010	9780	18700
Sugarcane	21600	3590	16500
Potato	67400 Genome: 2011	9610	18900
Sweetpotato	17400	1640	11700
Cassava	15400 Genome: 2012	1780	12200
Yams	6320	653	4450

Source: Google scholar May 31, 2015

Markers for sweetpotato relative to other crops

Crop	Total “CROP" and “Marker"	“CROP" and “Marker" 2014	“CROP" and “Marker" 2000-2014
Wheat	167,000	12700	23400
Rice	110,000	9510	45,500
Soybean	72300	8340	19500
Sugarcane	22000	3110	16400
Potato	71,800	8400	18500
Sweetpotato	18,600	1480	12200
Cassava	16,200	1660	12600
Yams	10,600	880	7390

Source: Google scholar May 31, 2015

What do we need?

- ❖ Genome sequence
- ❖ Next generation molecular markers
- ❖ Dense genetic maps

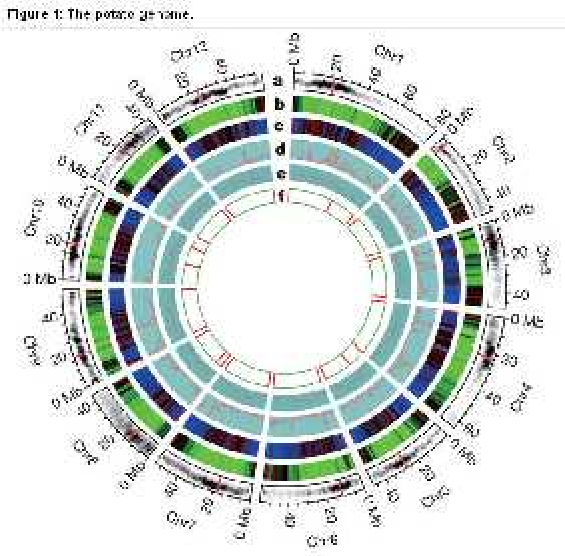
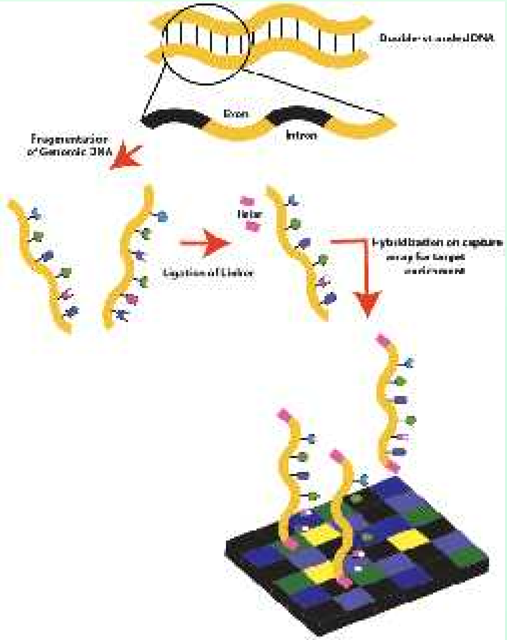
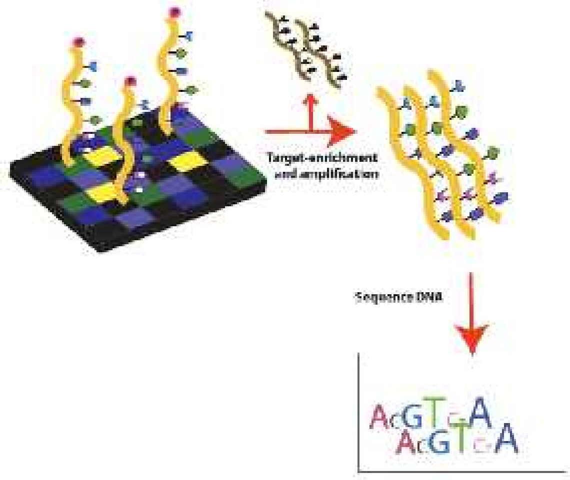
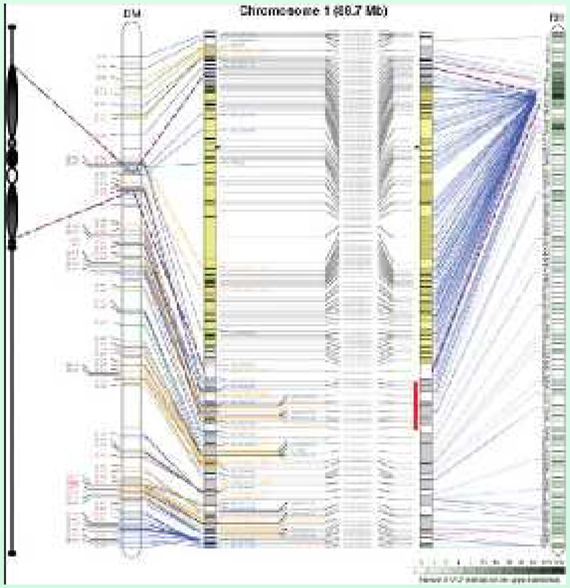


Figure 1: The potato genome.

Exome Capture







What do we need?

Clear Breeding Goals!

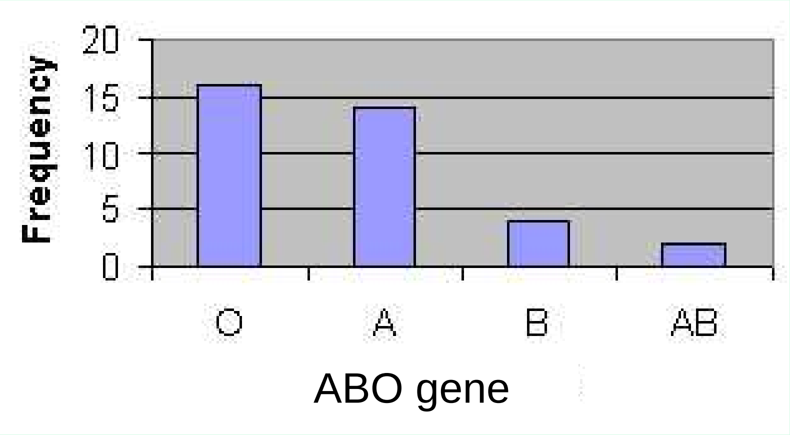
- General:
  - High-yielding lines (storage root)
  - Taste and nutrition
  - Resistance to biotic and abiotic stresses
  - Dry Matter
- Specific: Depending on program and the region, include
  - Beta-carotene levels
  - Mineral (iron, zinc) content
  - Starch content
  - Sucrose content
  - Drought, heat, and salt tolerance
  - SPVD, weevil tolerance
  - Biomass traits
  - Ornamental traits

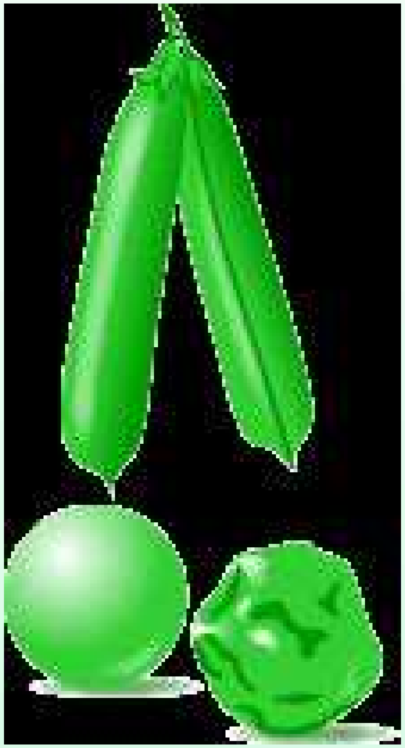
Phenotypic traits

- 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





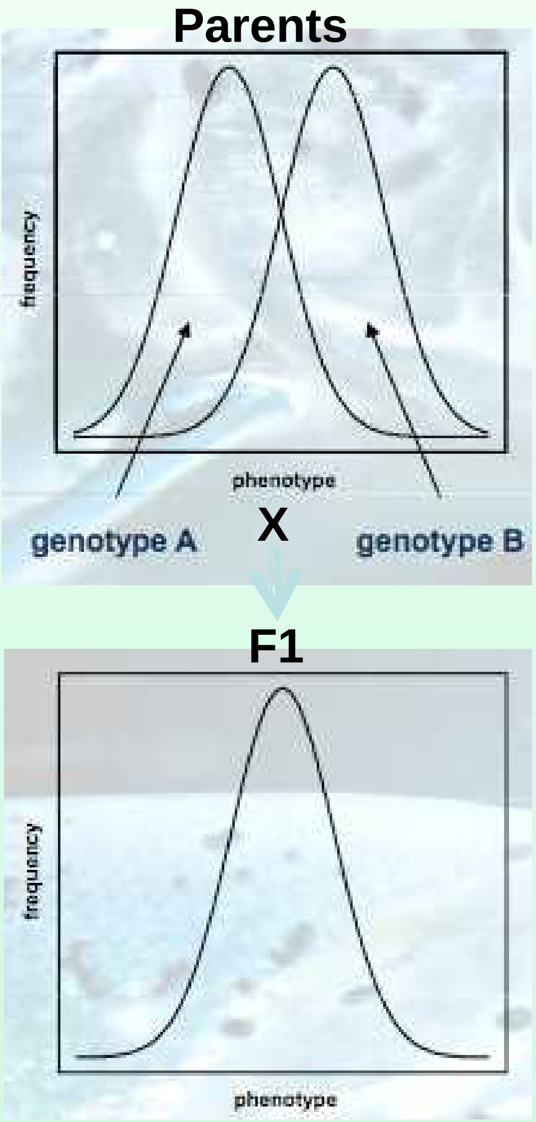
The wrinkled-seed character of pea is caused by a transposon-like insertion in a gene

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

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

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

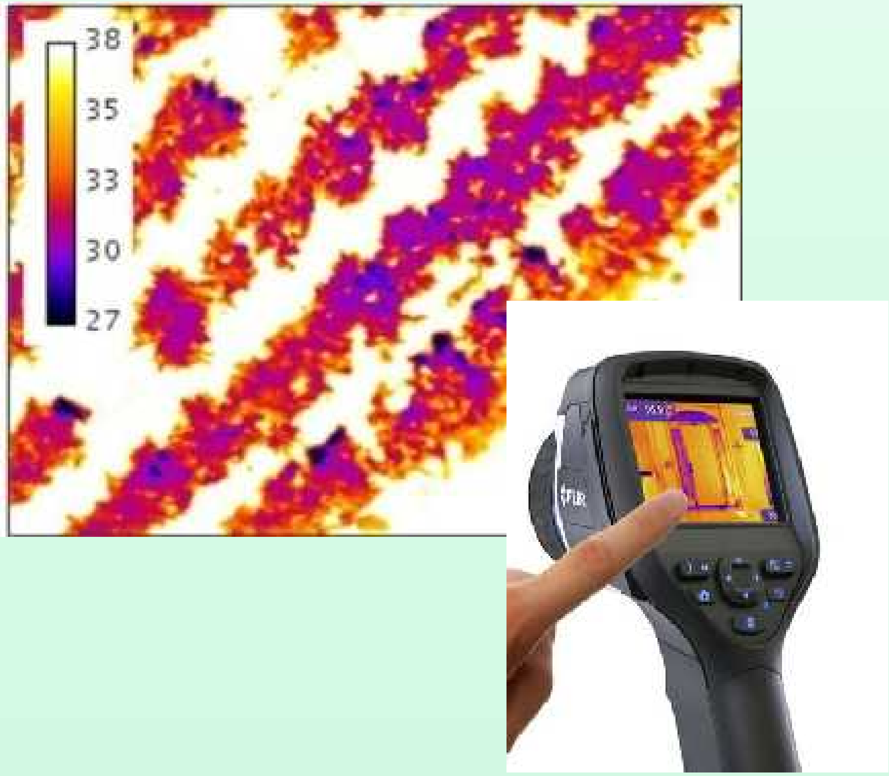
What do we need?

- High-throughput, precise, accurate and standardized phenotyping in multiple environments

Reflectance
NDVI

Chlorophyll
content SPAD

Infrared thermography for canopy temperature



What do we need?

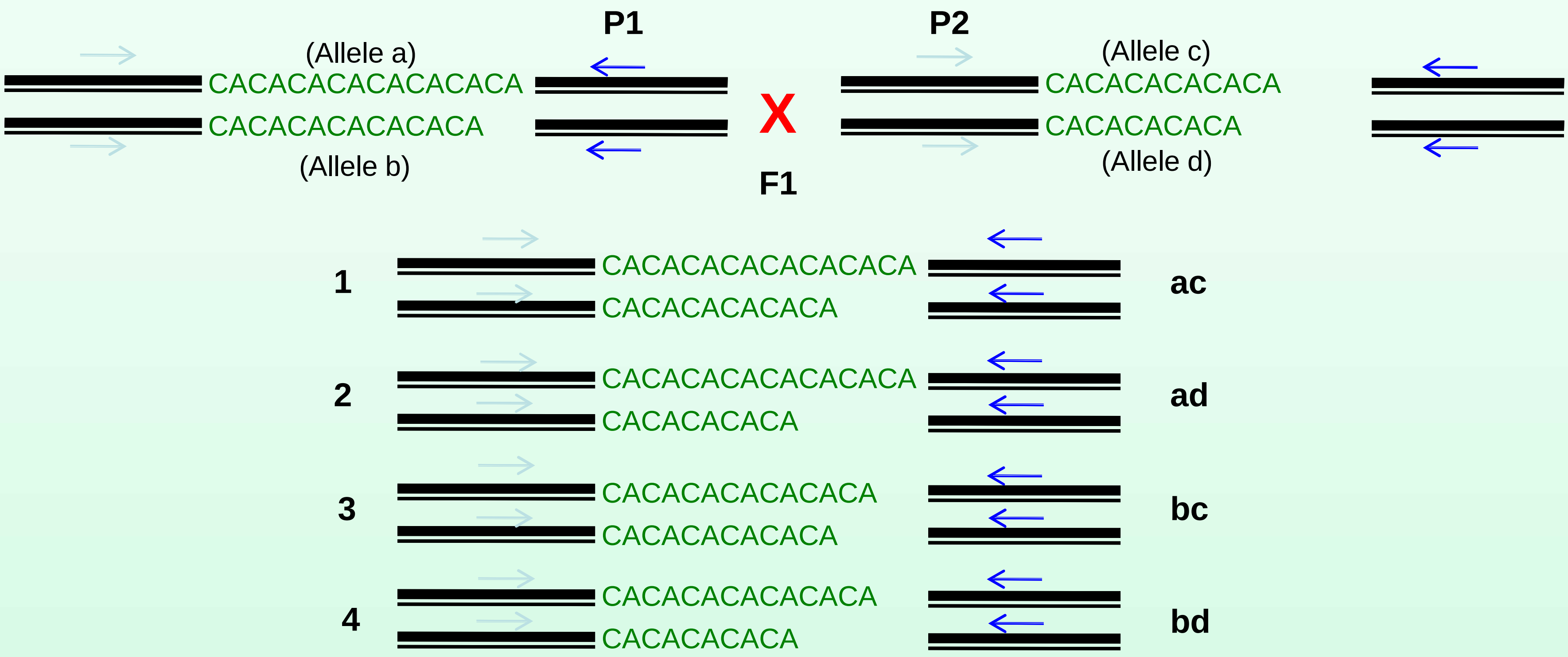
❖ Database to store and access genotypic, phenotypic and environment data

Welcome to SweetPotatoBase!

❖ Analytical tools for trait analysis, rapid identification of markers from sequence data marker-trait associations and genomic selections (most of these tools have to be suitable for hexaploid and clonally propagated crops!)

Thank you for attention

Simple Sequence Repeats (SSRs)



P1

(Allele a)
CACACACACACACACA
CACACACACACACA
(Allele b)

X

P2

(Allele c)
CACACACACACA
CACACACACA
(Allele d)

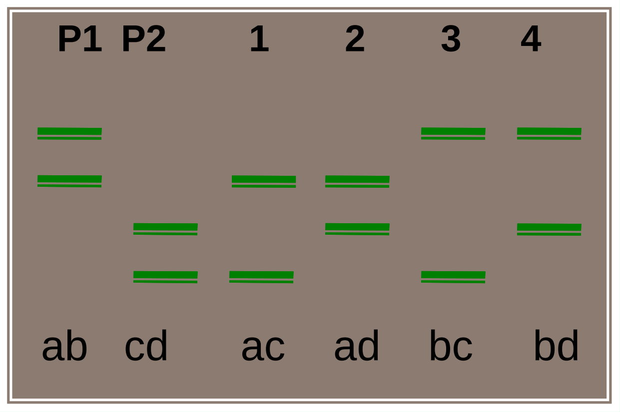
F1

1
CACACACACACACACA
CACACACACACA
ac

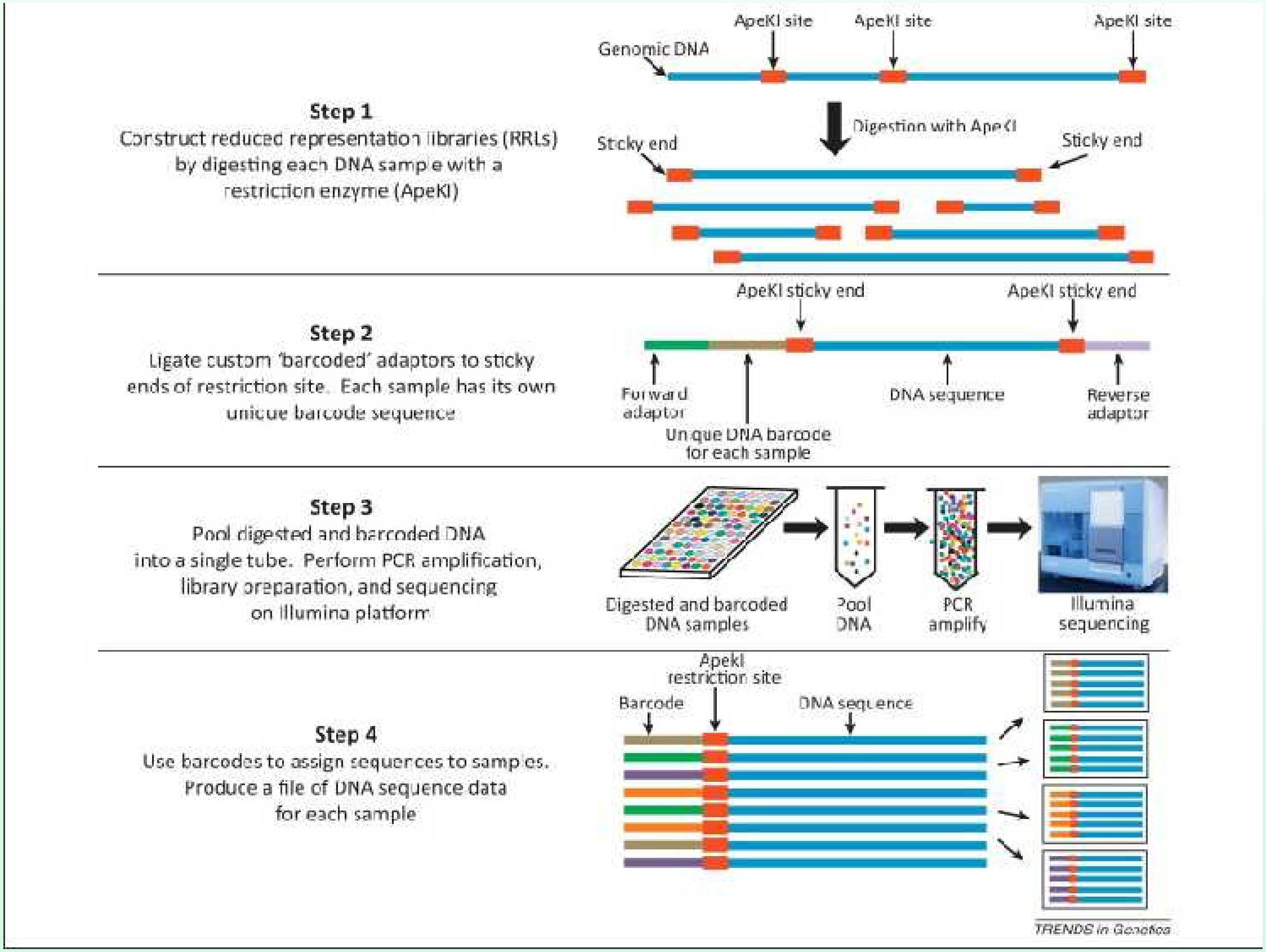
2
CACACACACACACACA
CACACACACA
ad

3
CACACACACACACA
CACACACACACA
bc

4
CACACACACACACA
CACACACACA
bd



Genotyping-by-Sequencing



Step 1
Construct reduced representation libraries (RRLs) by digesting each DNA sample with a restriction enzyme (ApeKI)

Step 2
Ligate custom 'barcoded' adaptors to sticky ends of restriction site. Each sample has its own unique barcode sequence

Step 3
Pool digested and barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on Illumina platform

Step 4
Use barcodes to assign sequences to samples. Produce a file of DNA sequence data for each sample

TRENDS in Genetics

Linkage map

Linkage group 1 and 2 of Papaya: Map was constructed using microsatellite markers in F2 population (Chen *et al.* 2007)