Genomic selection in plant breeding and genetic studies on popping traits

Huihui LI (李慧慧)
lihuihui@caas.cn and h.li@cgiar.org
Institute of Crop Sciences (ICS) and CIMMYT-China office
Chinese Academy of Agricultural Sciences (CAAS)

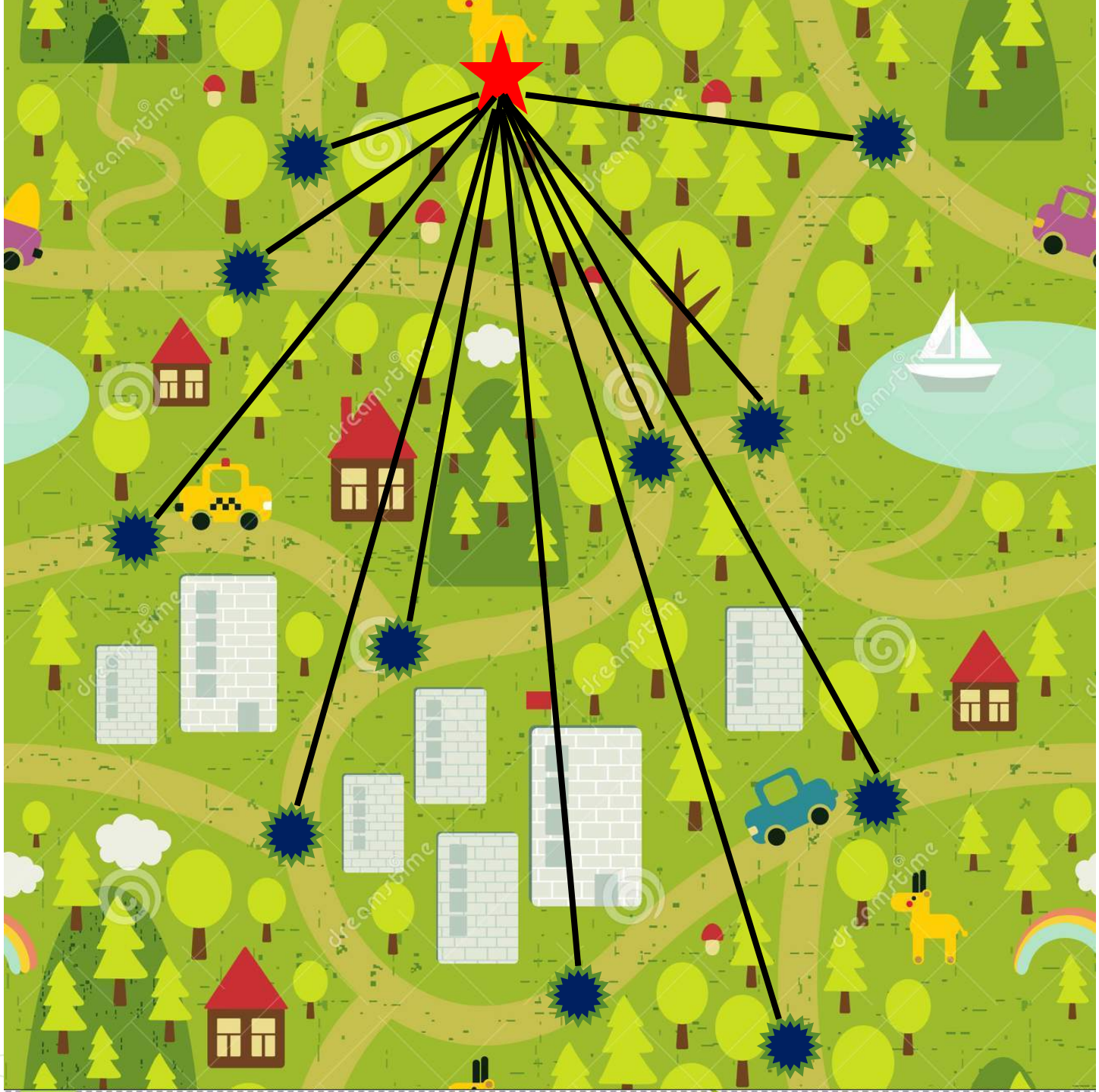
CIMMYT

Outline

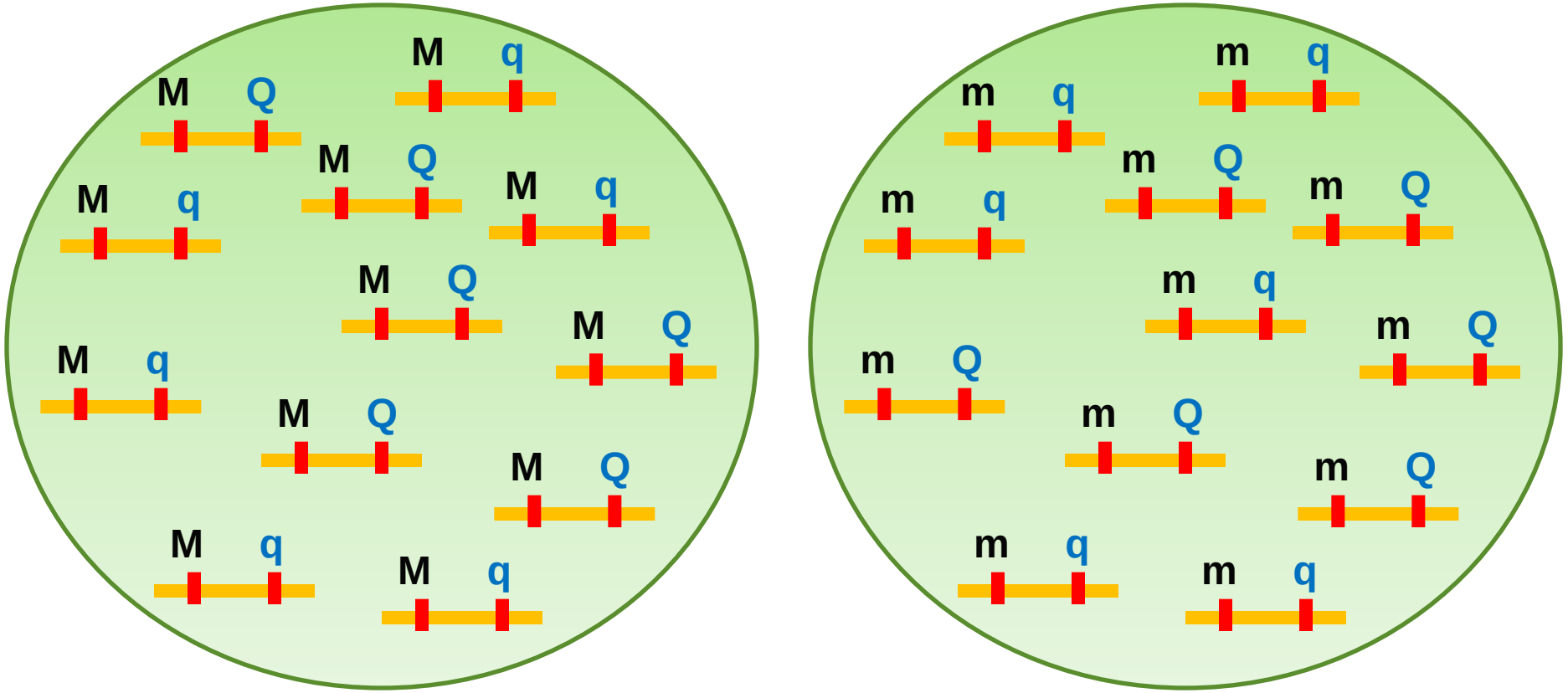
- What is the genomic selection (GS)?
- GS vs. marker assisted breeding
- GS vs. conventional breeding
- GS vs. association mapping
- Statistical models to estimate genomic estimation of breeding value (GEBV) in GS
- Prediction accuracy of GS
- Genetic studies on popping traits

What is the genomic selection?

Marker can help us find the target!



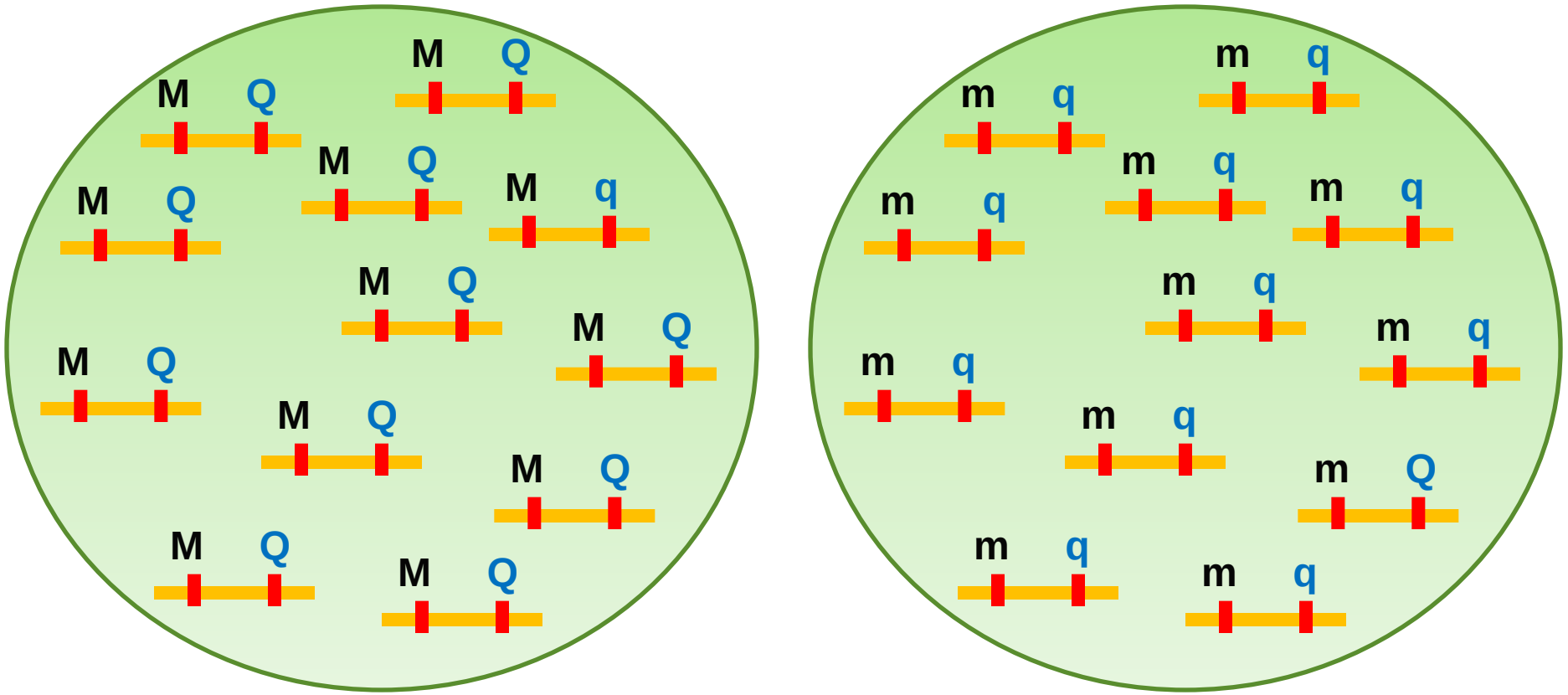
Linkage Equilibrium (LE)



$D = P_{MQ} - P_M P_Q = 1/4 - (1/2)*(1/2) = 0$

Marker genotype NOT related to phenotype

Linkage Disequilibrium (LD)



$D = P_{MQ} - P_M P_Q = 11/24 - (1/2)^* (1/2) = 5/24$

Marker genotype IS related to phenotype

(if Q/q has effect on phenotype)

Factors affecting LD

Increase

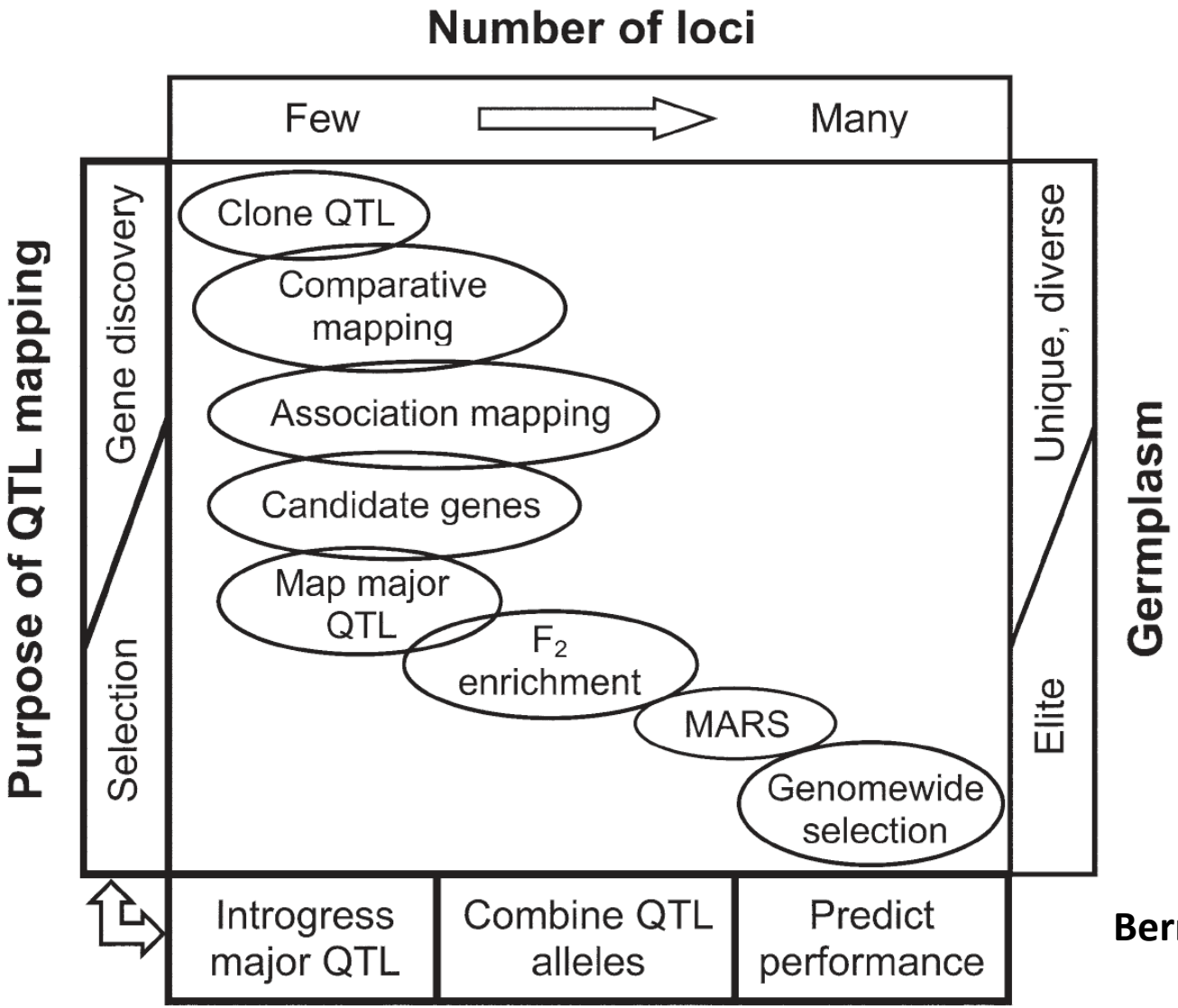
- Small effective population size
- Inbreeding
- Genetics isolation
- Population subdivision
- Low recombination rate
- Genetic drift
- Population admixture

Decrease

- Outcrossing
- Random matting
- High recombination rate
- High mutation rate

CIMMYT

Molecular markers as a tool in plant breeding



Bernardo, 2008

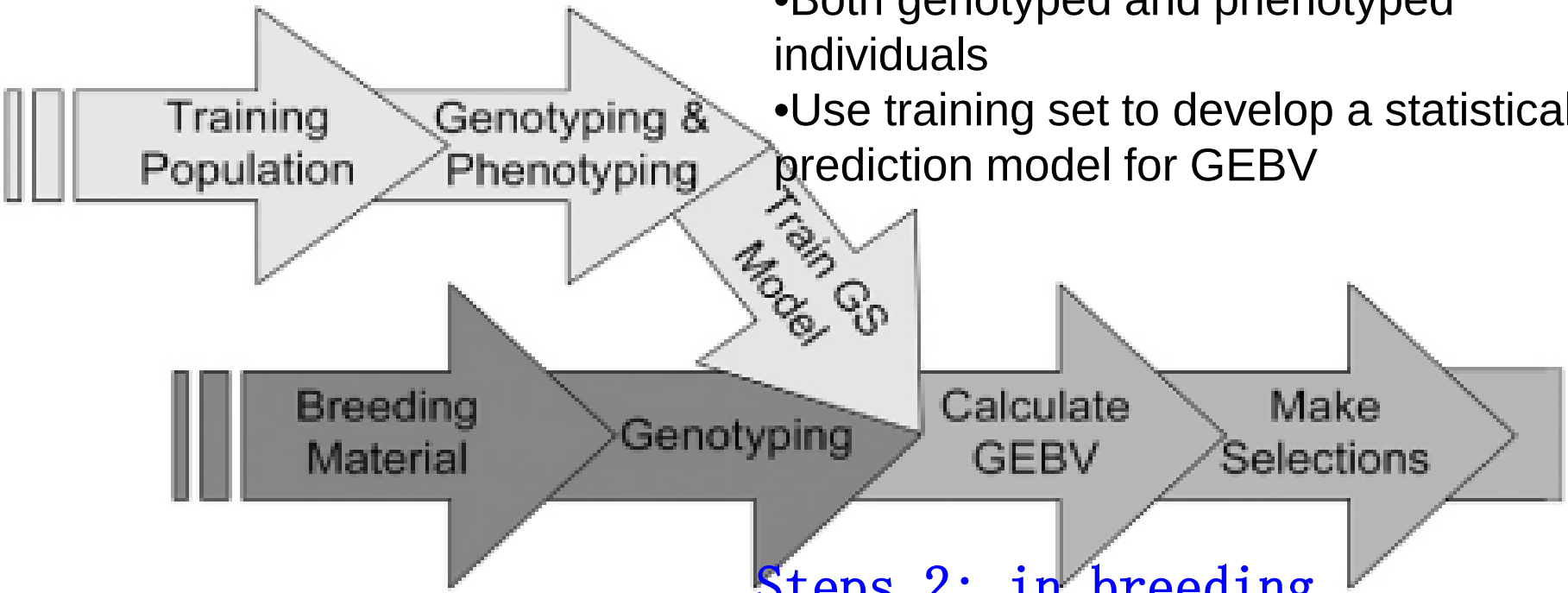
What is the genomic selection?

- Genomic selection refers to selection decisions based on genomic estimation of breeding value (GEBV).
- The GEBV are calculated as the sum of the effects of dense genetic markers (haplotypes) across the entire genome, there by potentially capturing all the quantitative trait loci (QTL) that contribute to variation in a trait.
- The marker effects are first estimated in a large reference population with phenotypic information. In subsequent generations, only marker information is required to calculate GEBV.

Meuwissen et al. 2001 *Genetics*, 157:1819-1829

CIMMYT

Steps in GS



Step 1: in training population

•Both genotyped and phenotyped individuals

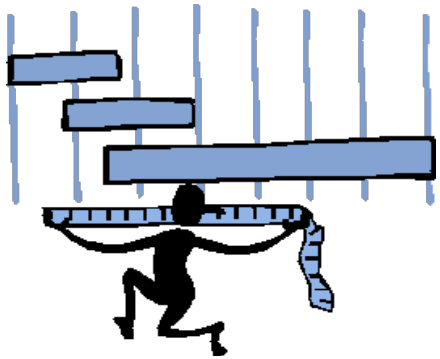
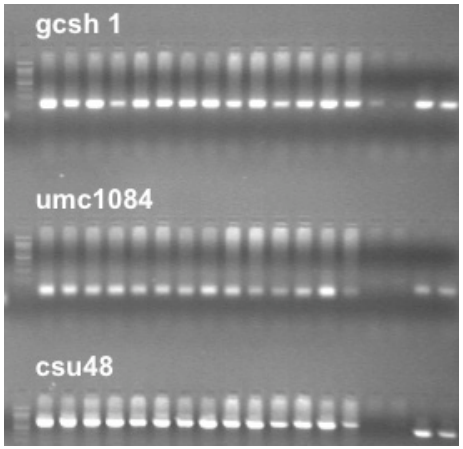
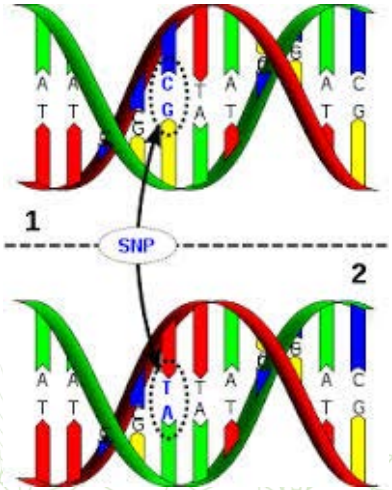
•Use training set to develop a statistical prediction model for GEBV

Steps 2: in breeding population

- Calculation GEBVs
- Prediction & Selection

Meuwissen et al. 2001 *Genetics*, 157:1819-1829;
Heffner, et al. 2009 *Crop Sci.* 49:1-12

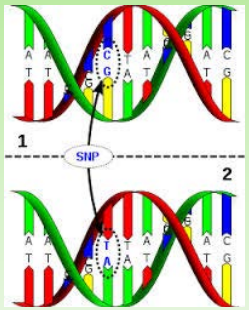
CIMMYT

		Genetics	Breeding
1900s'	Phenotype 	• Classic quantitative genetics • Analysis of phenotypic variance by experimental designs	• Conventional breeding
1980s'	Phenotype + Genotype 	• Linkage map by hundreds of markers • QTL mapping • Association mapping on candidate genomic region	• Marker assistant breeding
2010s'	Phenotype + Genotyping by sequencing 	• Linkage map by ten thousands of markers • QTL mapping • Genome-wide association mapping	• Marker assistant breeding (MAS) • Genomic selection

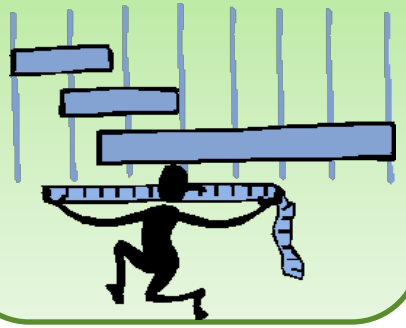
CIMMYT

Nowadays…

Genotyping by sequencing



Phenotypic data



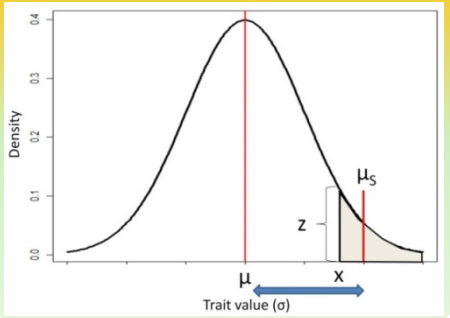
$$y = Xb + Zu + e$$

Testing for significant markers

Gene mapping

Marker assisted selection

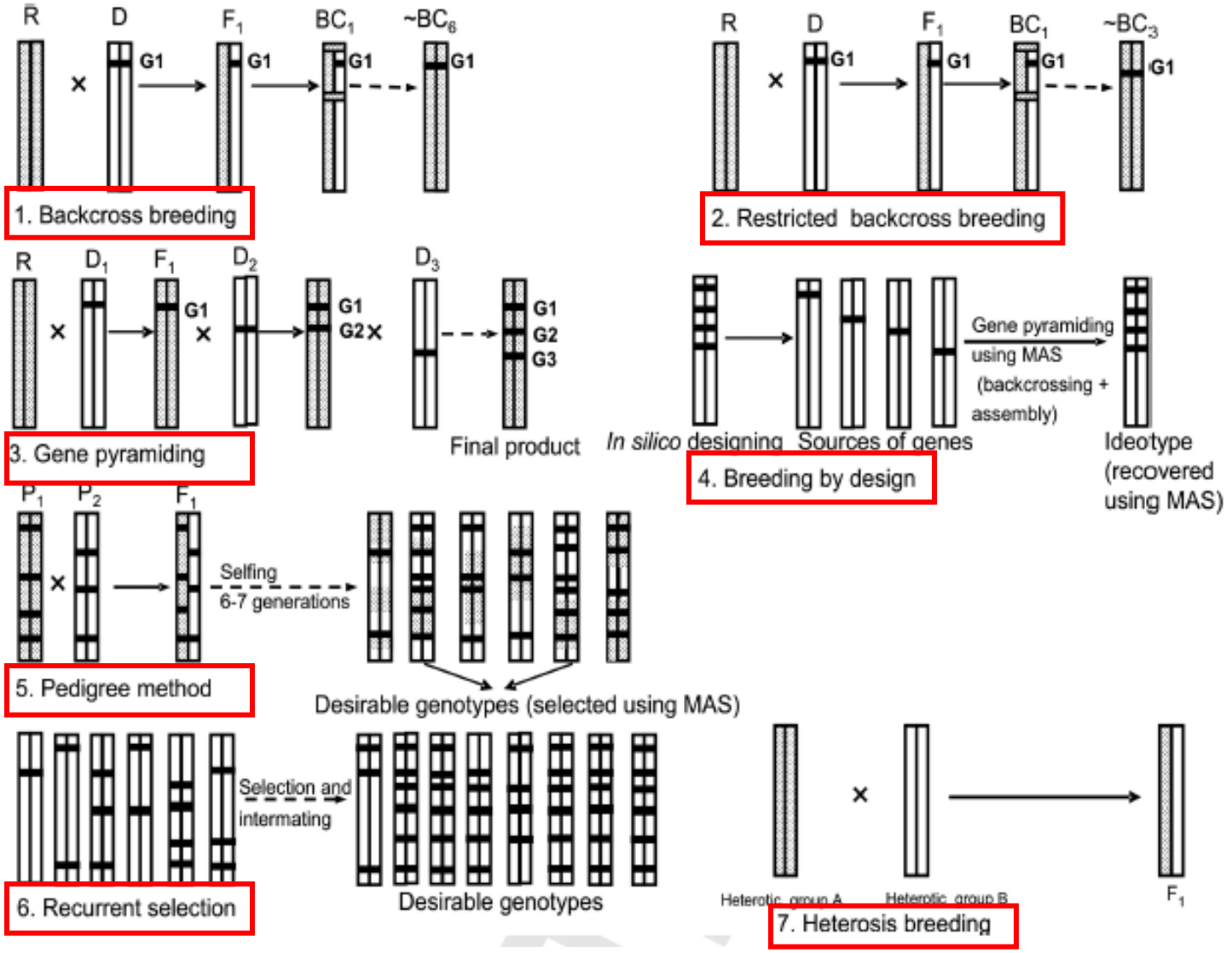
Genomic selection



Simultaneous estimation of all locus effects to calculate the GEBV

GS vs. MAS

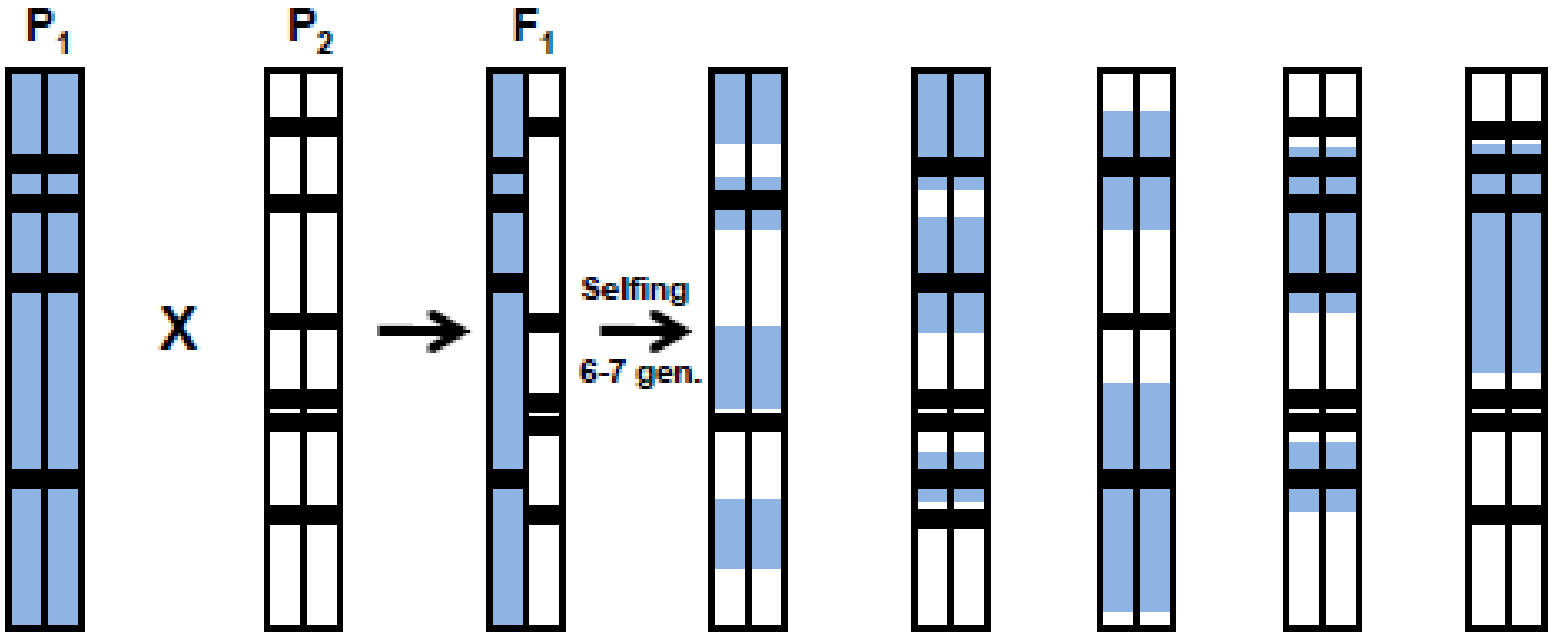
Seven common breeding and selection schemes involving MAS



Gupta et al. 2010. Plant Breeding Reviews 33: 145-217

Pedigree breeding

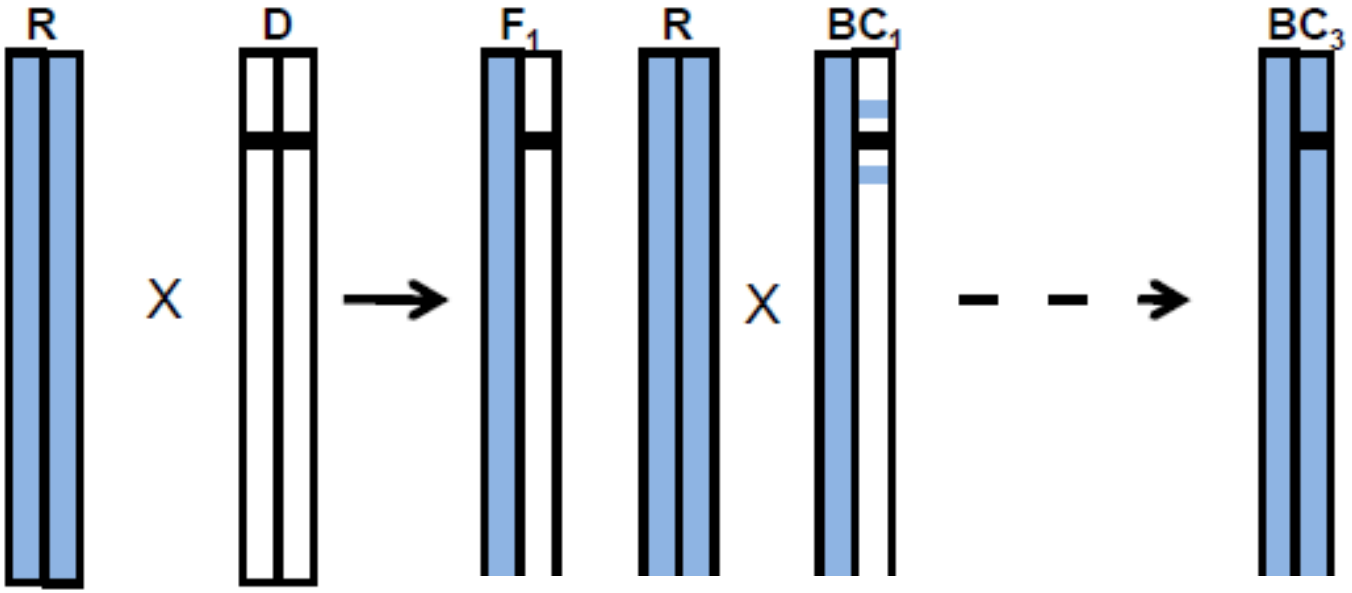
- Pedigree MAS is especially relevant for self-pollinated crops such as wheat, barley, and rice, where pedigrees of elite germplasm are known.
- Markers closely linked to the target genomic regions can be used to accelerate fixation of favorable alleles in the next selection step.



Gupta et al. 2010. Plant Breeding Reviews 33: 145-217

Marker assisted backcross selection (MAB)

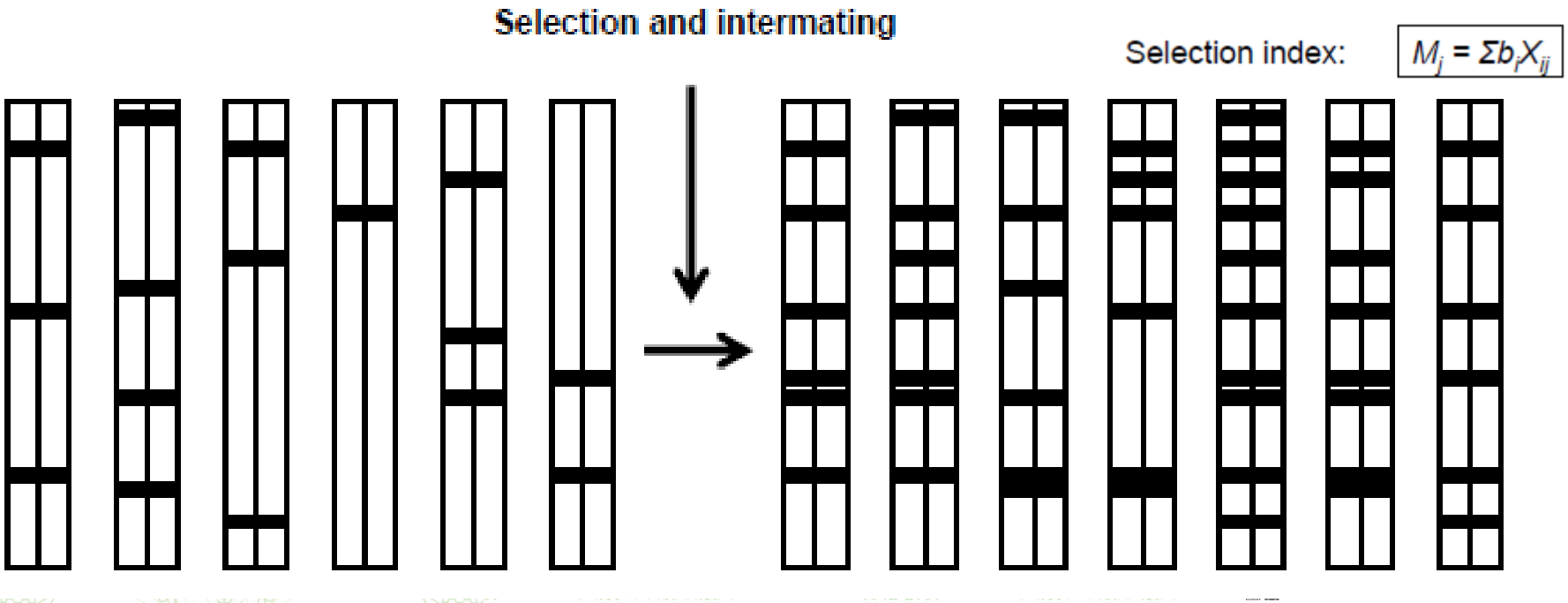
- The use of markers can precisely transfer the genomic regions involving in the expression of target traits (foreground selection) and by speeding up the recovery of the recurrent parent genome (background selection).
- MAB is particularly useful for pyramiding genes or QTL for resistance against a pathogen or pest and for traits that are highly influenced by the environment.



Gupta et al. 2010. Plant Breeding Reviews 33: 145-217

Marker-assisted recurrent selection (MARS)

- To overcome the limitation of MAB that limited number of desirable alleles (one or few) that can be introgressed.
- A selection index is also used in MARS, where weights are assigned to markers, based on the magnitude of the associated QTL effect.



Gupta et al. 2010. Plant Breeding Reviews 33: 145-217

Limitations of current MAS

- QTL detection on bi-parental mapping population
- Only allelic variation present in the two parents captured
- Best suited for traits controlled by few QTL with major effect (high heritability)
- Most current target traits are multi-genic, i.e. controlled by many gene with small effects. Difficult to get statistical power to detect them.
- LD or association mapping sidesteps the need for bi-parental mapping populations and can be used to associate genotype with trait in breeding or other populations which have been extensively phenotyped.

GS is a form of MAS

- GS has become feasible due to revolution in SNP discovery method like deep sequencing and throughput SNP genotyping on DNA chip.
- Genetic markers cover the whole genome are used so that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker.
- GS are more suitable for traits with many numerous genes controlled

GS vs. conventional breeding

- In conventional breeding programmes phenotyping used for selection
- In GS main role of phenotyping is to calculate effect of markers

GS vs. association mapping

- GS models do not require known location of markers in the genome.
- GS models do not require estimation of relative effects of individual QTL on the trait.

Statistical models to estimate genomic estimation of breeding value (GEBV) in GS

From least square estimation when p < n

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$$

$$E(\hat{\beta}) = \beta$$

p is the number of markers;
n is the number of population size

CIMMYT

NP-hard problem when p >> n

- $(X'X)^{-1}$ does not exist;
- The degree of freedom is not enough;
- There is no unique estimation of marker-effects;
- But, the total genomic value remains estimable.

Methods for genomic selection when p>>n

- Dimension reduction methods
  - Ridge regression
  - Partial least squares
  - Principal component regression
- Kernel and machine learning methods
  - Support vector machine regression

Methods for genomic selection when p>>n

- BLUP-based methods by fitting the allelic effects as random effects
  - Does not require degrees of freedom;
  - Require an estimate of the variance of the allelic effects;
  - Every gene has the same variance;
  - RR-BLUP, GBLUP
- Bayesian estimation
  - This is similar to BLUP, except that the variance of the allelic effects is assumed different for every gene, and is estimated by using a prior distribution for this variance
  - Bayes A/B, BayesCП, BayesDП

Ridge regression

- Including all variables, but replacing normal least-squares estimators with

$$\hat{\boldsymbol{\beta}} = (X'X + \lambda I)^{-1} X'Y$$

- Normal estimates shrunk toward 0
- The degree of shrinkage is determined by lambda
- Choose lambda to minimize model error
- Addition of λI term reduces collinearity and prevents the matrix X'X from becoming singular.

Whittaker et al. 2000

How to calculate the genomic relationships from markers

The product of **X** matrix with its transpose **X`** is **X`X** matrix

$$\mathbf{X`X} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 3 \end{bmatrix}$$

← individual 1
← individual 2
← individual 3

- Diagonal: Counts the # of homozygous loci for each individual.
- Off-diagonal: Measure the number of alleles shared by relatives

VanRaden, 2008, Forni et al. 2011

CIMMYT

GBLUP

- Similar to traditional BLUP with pedigrees
- Calculate genomic relationship matrix
- Use genomic relationships in mixed-linear model to predict the breeding values

$$y = X\beta + \xi + \varepsilon$$

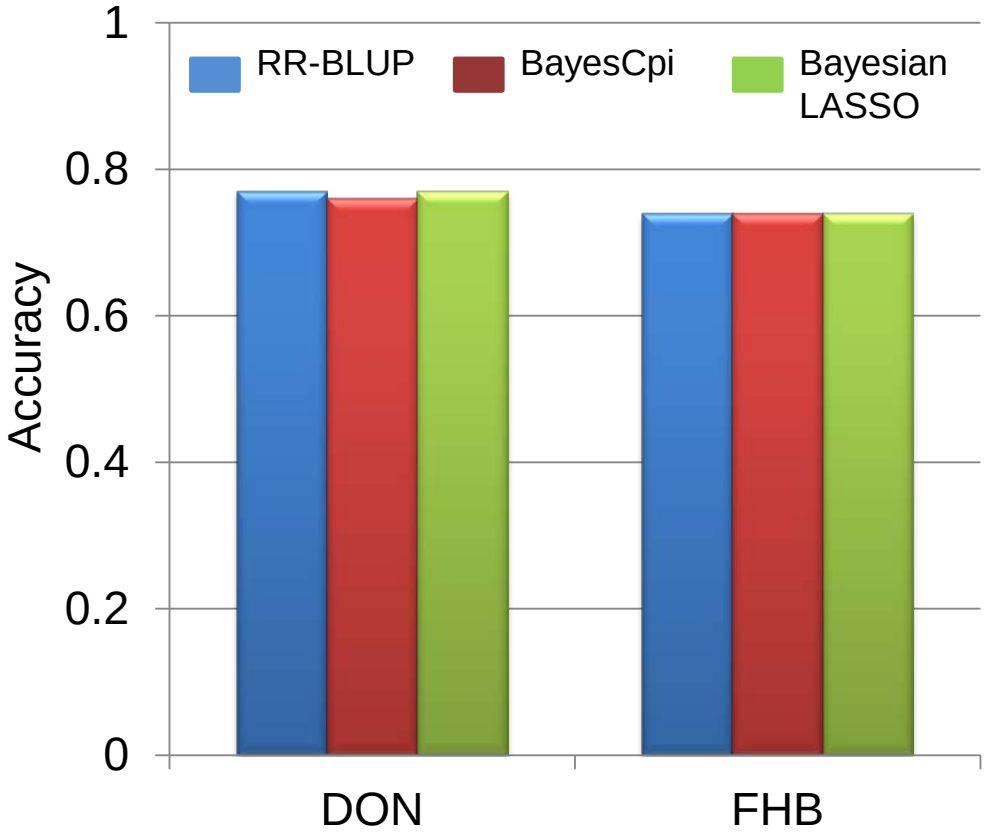
$$y = X\beta + \sum_{k=1}^{m} Z_k \gamma_k + \varepsilon$$

The GBLUP method of genomic prediction does not require estimation of marker effects

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1\beta \\ X_2\beta \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

$$\hat{y}_2 = X_2\hat{\beta} + \hat{\phi}^2 K_{21}(K_{11}\hat{\phi}^2 + I\hat{\sigma}^2)^{-1}(y_1 - X_1\hat{\beta})$$

Models typically similar in accuracy



Bernardo and Yu (2007), Lorenzana and Bernardo (2009), Van Raden et al. (2009), Hayes (2009).

CIMMYT

Prediction accuracy of GS

Prediction accuracy by cross validation

	Marker	C263	R830	R3166	XNpb387	R569	R1553	C128	C1402	XNpb81	C246	R2953	C1447	Grain width (mm)
Training	RIL1	0	0	0	0	0	0	0	0	0	0	0	0	2.33
	RIL2	2	2	2	2	2	0	0	0	0	2	2	2	1.99
	RIL3	0	2	2	2	2	2	2	2	2	2	2	2	2.24
	RIL4	0	0	0	0	0	0	2	2	2	2	2	2	1.94
	RIL5	0	0	0	0	0	2	2	0	0	0	0	0	2.76
	RIL6	0	0	0	2	2	2	2	2	2	2	2	2	2.32
	RIL7	0	0	0	0	0	0	0	0	0	0	0	0	2.32
Validation	RIL8	2	2	0	2	2	0	0	0	0	2	2	2	2.08
	RIL9	0	0	0	0	2	2	0	0	0	0	0	0	2.24
	RIL10	0	0	0	0	2	2	0	0	0	0	0	0	2.45

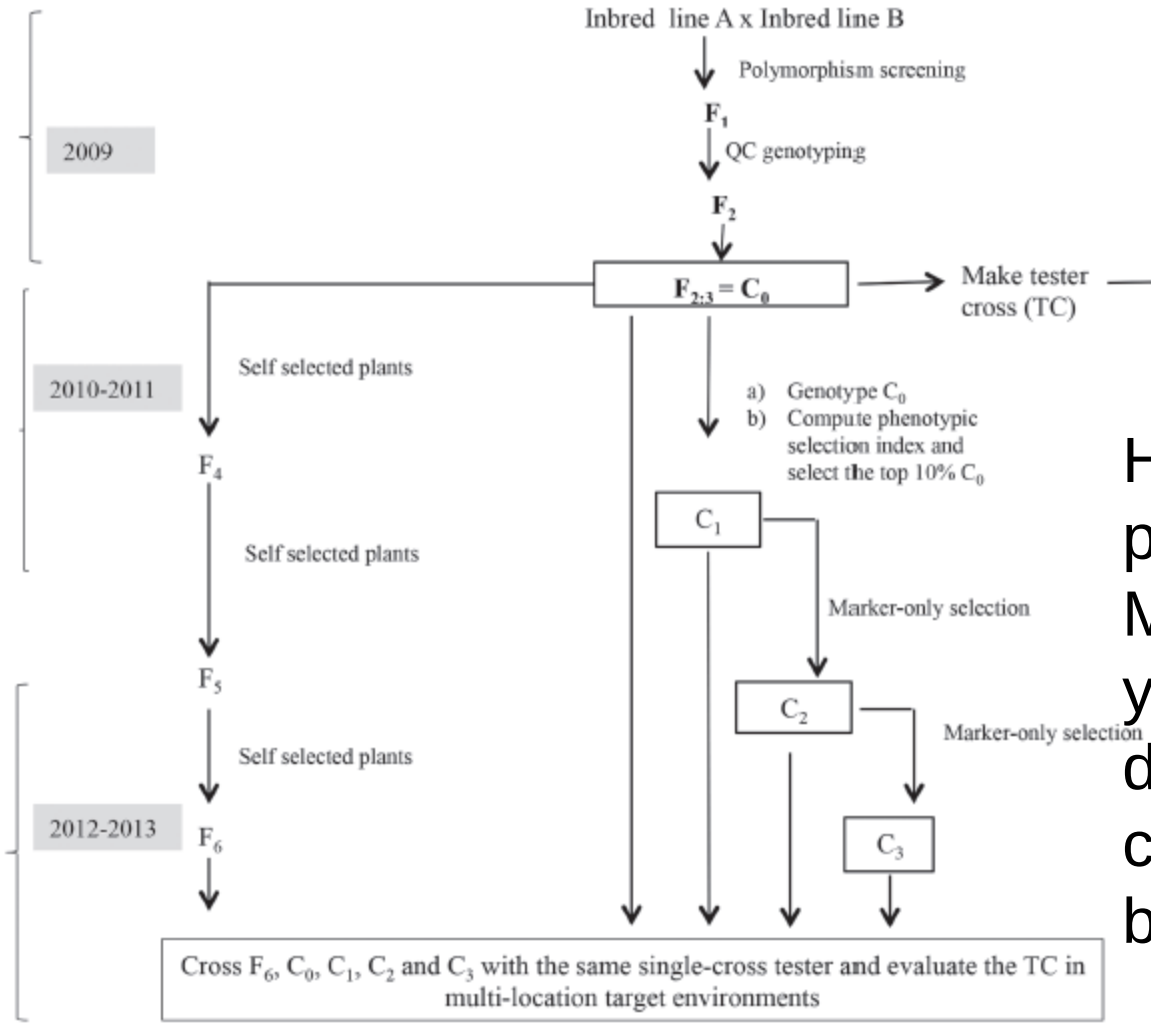
y_1

y_2

$$r_{y_2,\hat{y}_2} = \frac{\mathrm{cov}(y_2, \hat{y}_2)}{\sqrt{\mathrm{var}(y_2)\,\mathrm{var}(\hat{y}_2)}}$$

CIMMYT

Prediction accuracy by genetic gains per cycle and per year



Hybrids derived from C3 produced 7.3% (0.176 Mg ha–1) higher grain yield than those developed through the conventional pedigree breeding method.

Beyene et al. 2015 Crop Science, 55: 154-163

CIMMYT

Factors affecting prediction accuracy

- Training population size
- Trait heritability
  - Influence of G x E, precision of measurements
- Marker density
- Effective population size of breeding population
  - i.e., genetic diversity of breeding population
- Genetic relationship between training population and selection candidates

Resources and packages

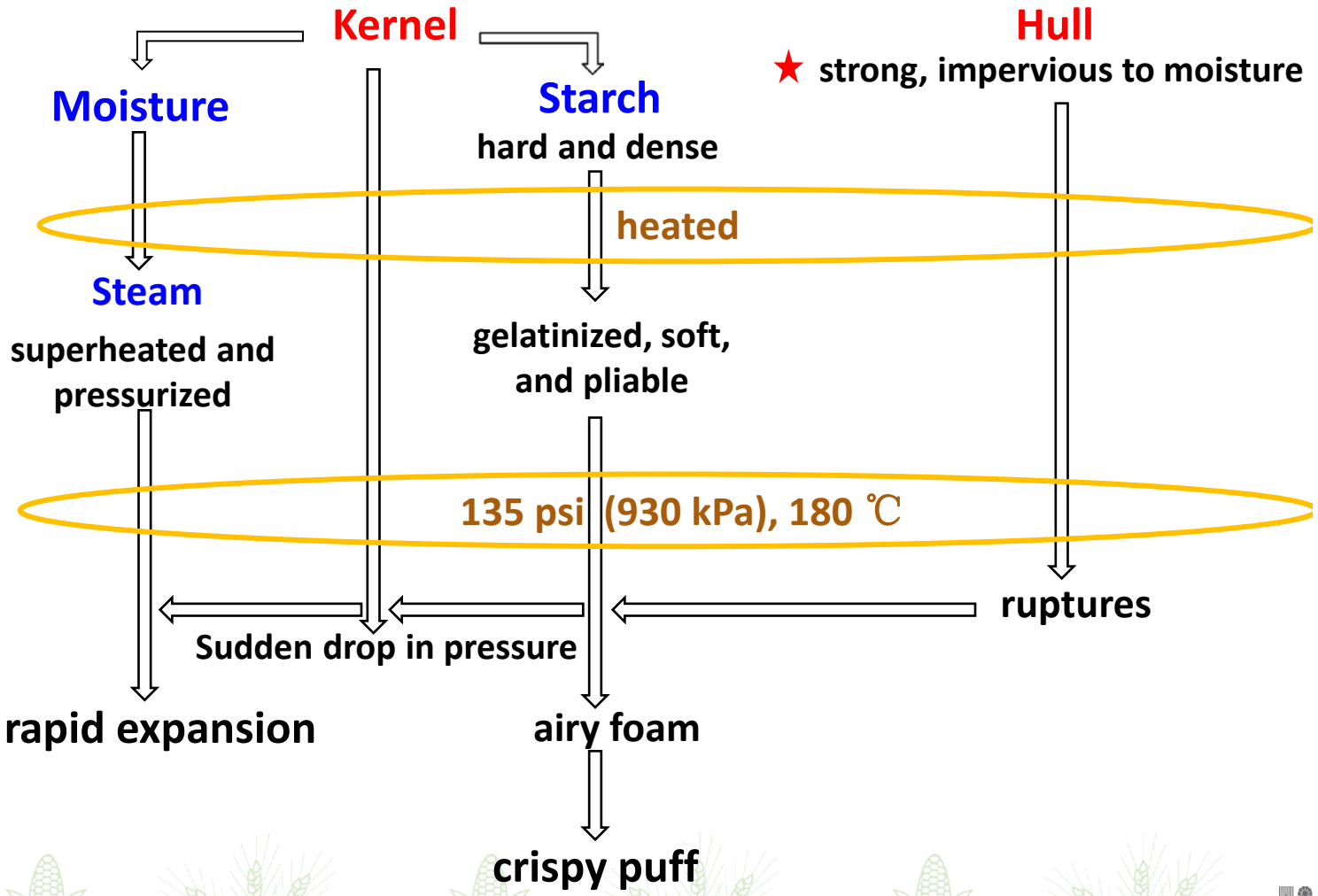
- rrBLUP package
  - http://cran.r-project.org/web/packages/rrBLUP/rrBLUP.pdf
  - Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.
  - Endelman, J.B., and J-L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3:2:1045
- BLR (Bayesian Linear Regression) package
  - https://cran.r-project.org/web/packages/BLR/index.html
  - Perez et al. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3:106-116.
- BGLR (Bayesian generalized linear regression) packages
  - http://R-Forge.R-project.org/projects/bglr/
  - Jarquín et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127-3: 595-607.

Popcorn in Mexico:
Re-establishing an ancient connection with the people who created it

Drs. Denise E. Costich, Huihui Li, and the Popcorn Team@CIMMYT

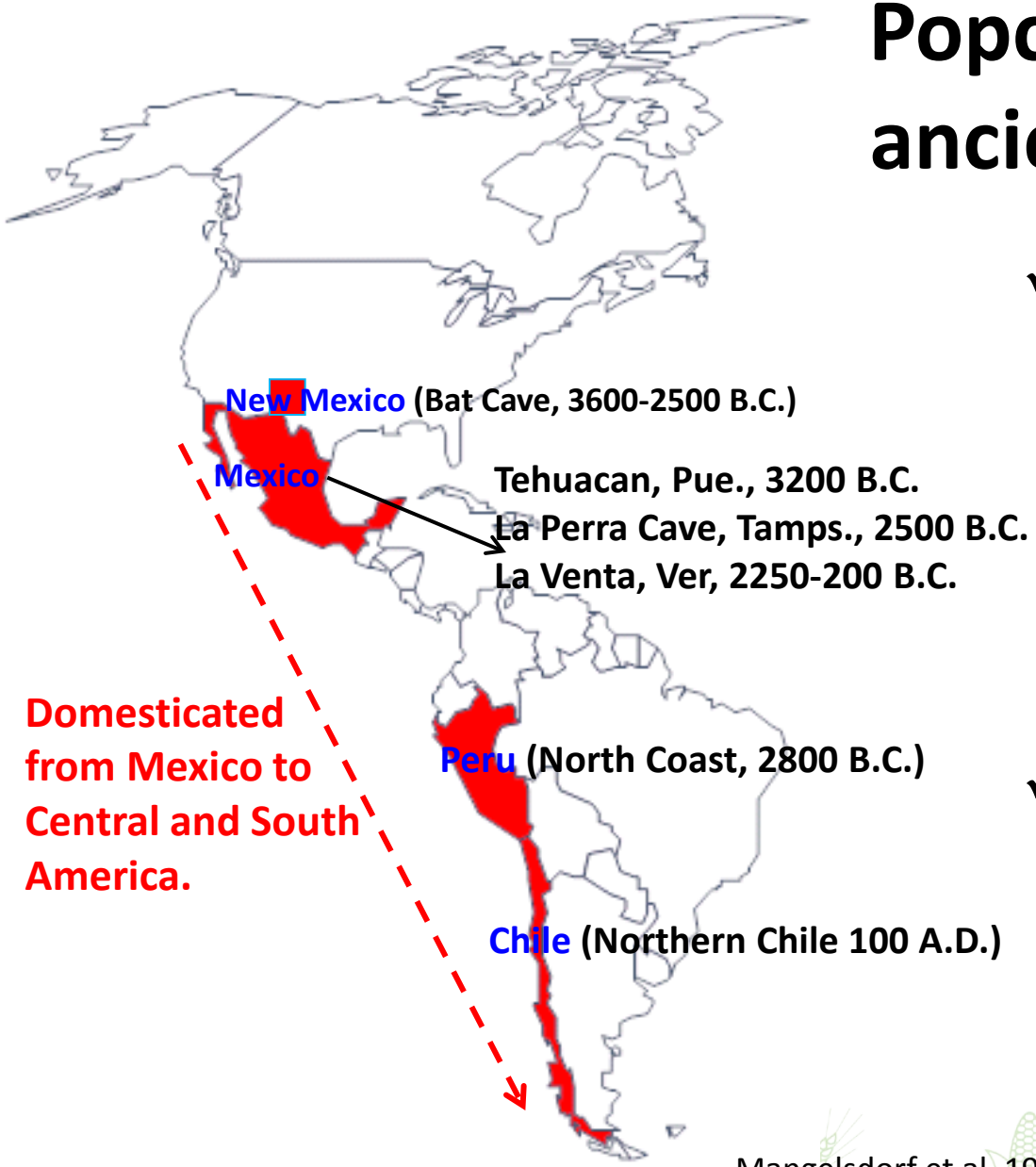
CIMMYT MR

Popcorn (*Zea mays everta*) is the only type of maize that pops



CIMMYT

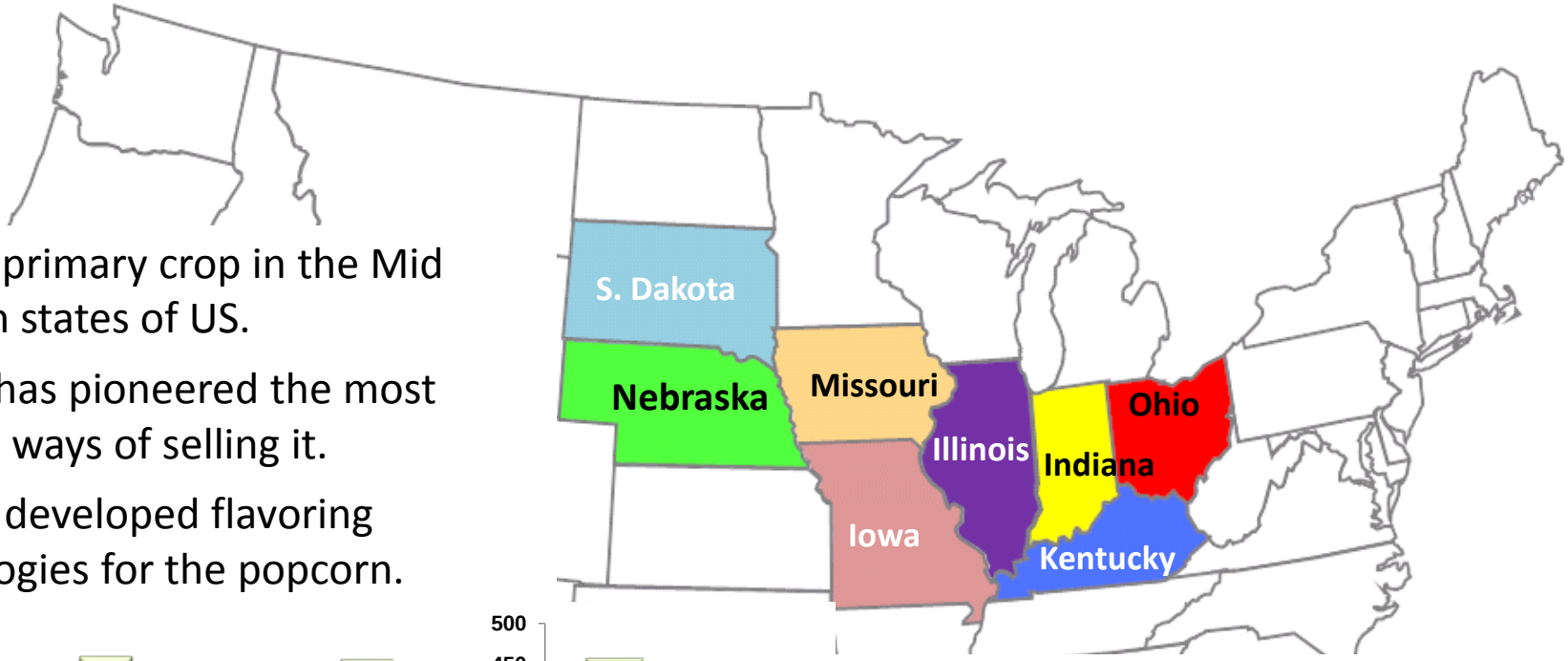
Popcorns are the most ancient types of maize



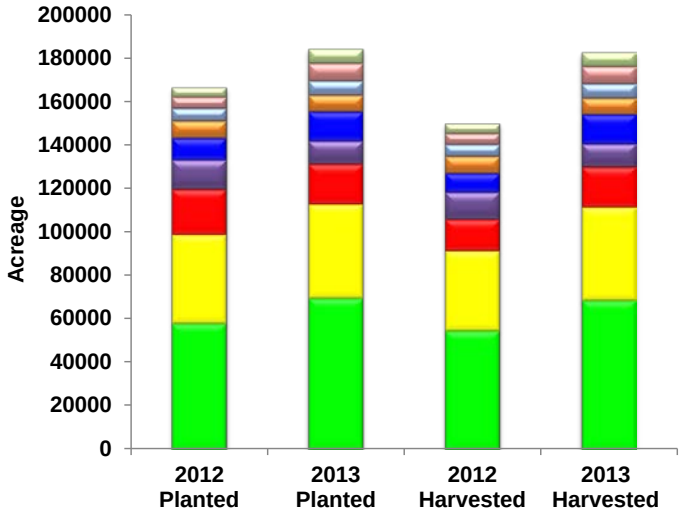
- ✓ When Columbus arrived in the New World, many types of maize, including popcorn, were already being exchanged between native tribes
- ✓ It is believed that the first preparation of wild and early cultivated maize for human consumption was popping.

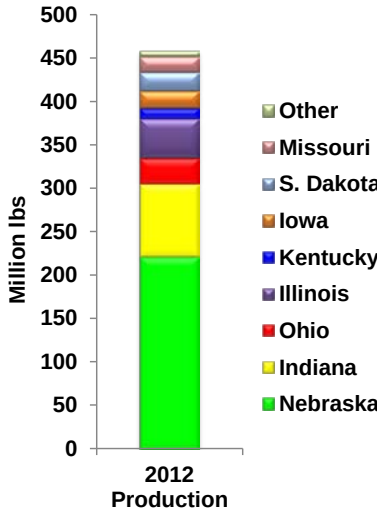
Mangelsdorf et al. 1964; MacNeish 1964; Mangelsdorf and Pollard 1975; Alexander et al. 1977; Rust and Leyden 1994.

Nearly all of the world's popcorn production is in the United States.



- ✓ It is the primary crop in the Mid Western states of US.
- ✓ The US has pioneered the most creative ways of selling it.
- ✓ The US developed flavoring technologies for the popcorn.





Sources:

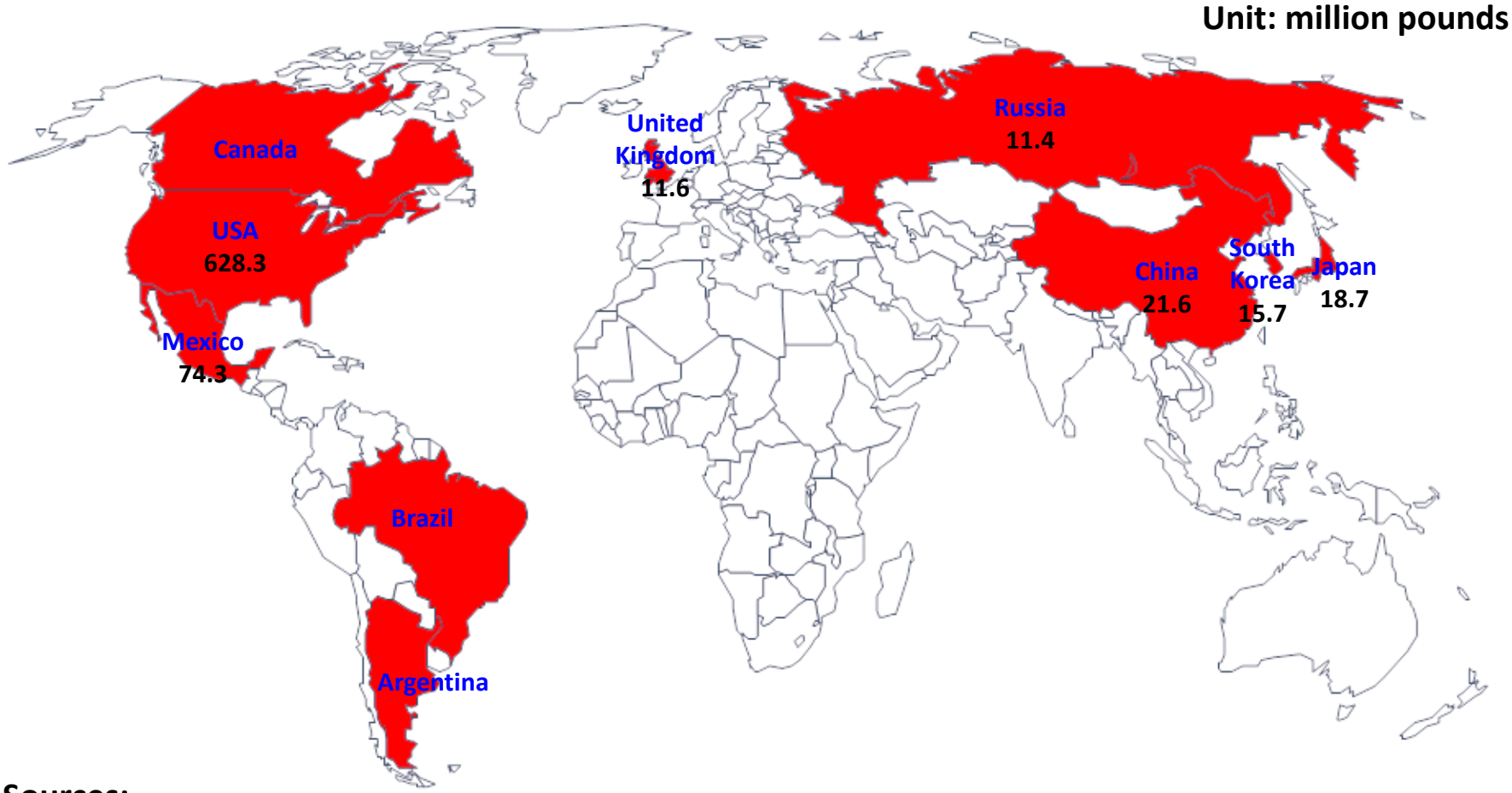
Global Agricultural Trade System, Foreign Ag Service, USDA.

Popcorn, Field Crops: 2007 and 2002, National Ag Statistics Service, USDA.

Popcorn, National Agricultural Library, USDA.

Popcorn Promotion, Research and Consumer Information Order, Ag Marketing Service, USDA.

While consumers of popcorn are worldwide!



Sources:

Global Agricultural Trade System, Foreign Ag Service, USDA.

Popcorn, Field Crops: 2007 and 2002, National Ag Statistics Service, USDA.

Popcorn, National Agricultural Library, USDA.

Popcorn Promotion, Research and Consumer Information Order, Ag Marketing Service, USDA.

CIMMYT

…in fact, nearly all of the popcorn consumed in Mexico comes from the USA….

Bagged and microwave popcorn from the USA dominate the shelves of Walmart (Mexico)

The number of landrace popcorn growers in Mexico decreases every year.

CIMMYT

Why?

- Mexican landrace popcorns generally show reduced expansion volume (an important market trait)
  - North American Yellow Pearl -- **1,166** cm^3 30 g^{-1}*
  - Mexican landraces (mean) -- **48.8** cm^3 30 g^{-1}*

- Decline in use of popcorn as a specialty maize:
  Often the grain is mixed with other types for tortillas…

*** Data from our collaborator, Dr. Amalio Santacruz Varela (Colegio de Posgraduados)**

What is CIMMYT's role?

- Germplasm/Genetic Diversity
- Scientific expertise in genotyping, phenotyping and analysis
- Training capacity
- Enabling environment

CIMMYT Maize Bank has a global collection of popcorns

873 popcorn accessions in total!

Goal of this project

To develop the best popcorn varieties that are locally adapted to the agroecosytems of the target countries, starting with Mexico.

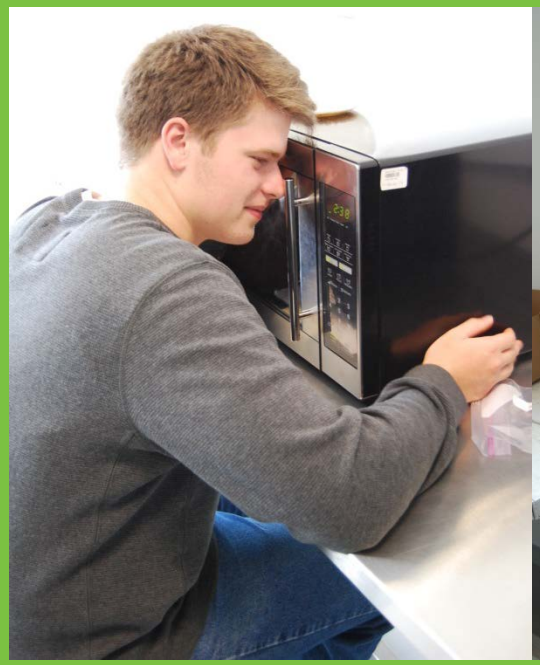
Objectives:

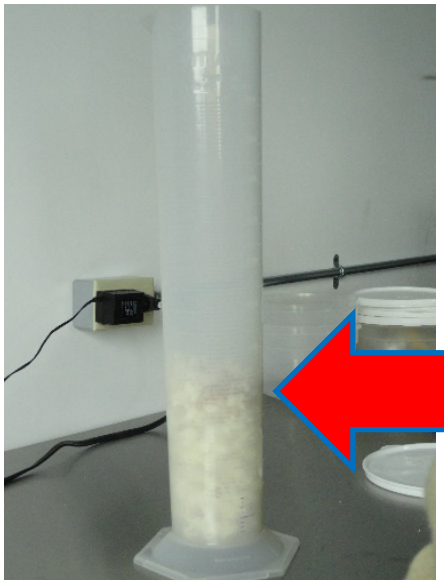
1. Find the sources for the best genetic diversity for popcorn traits;
2. Determine the genetic basis for these traits;
3. Provide the germplasm to accelerate breeding programs for popcorn in Mexico and other countries with market potential and interest in self-sufficiency.

CIMMYT MR

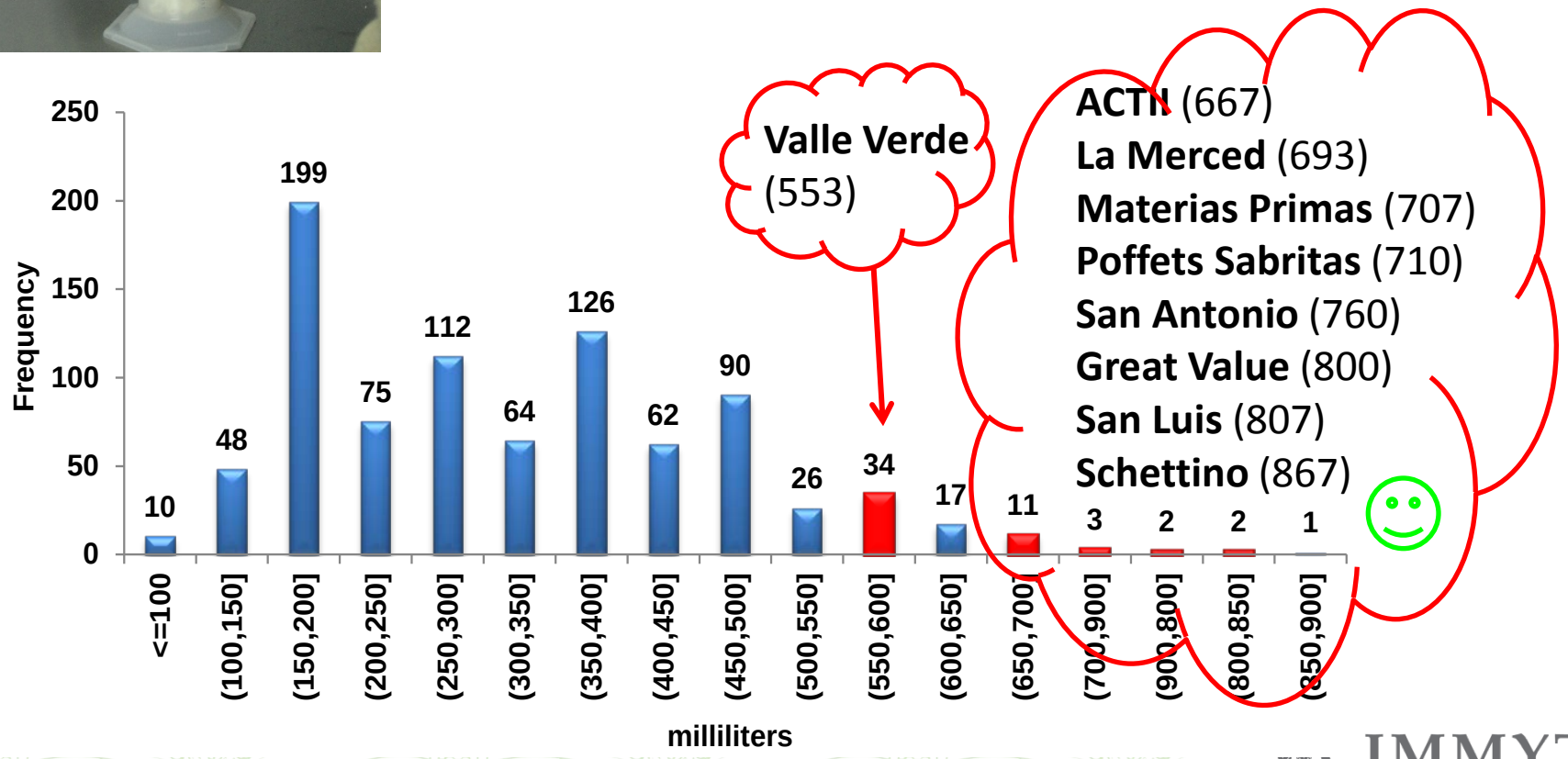
Popcorn Phenotyping Team in Action! Summer 2014

873 landrace accessions plus 9 commercial checks measured for 7 traits





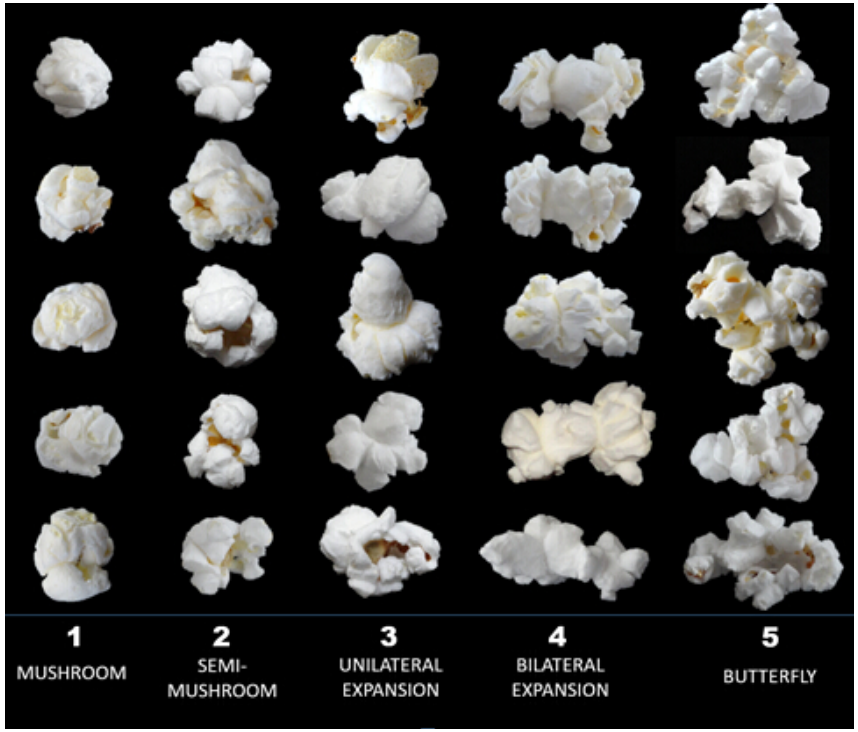
Expansion volume (mL) of 30 grams of popped kernels



Percent (%) of unpopped kernels after microwaving for 2:45 min at 70% power

More traits we are phenotyping

Popcorn flake morphology



Frequency

0 2 54 151 126 303 38 150 49

Popcorn flake color

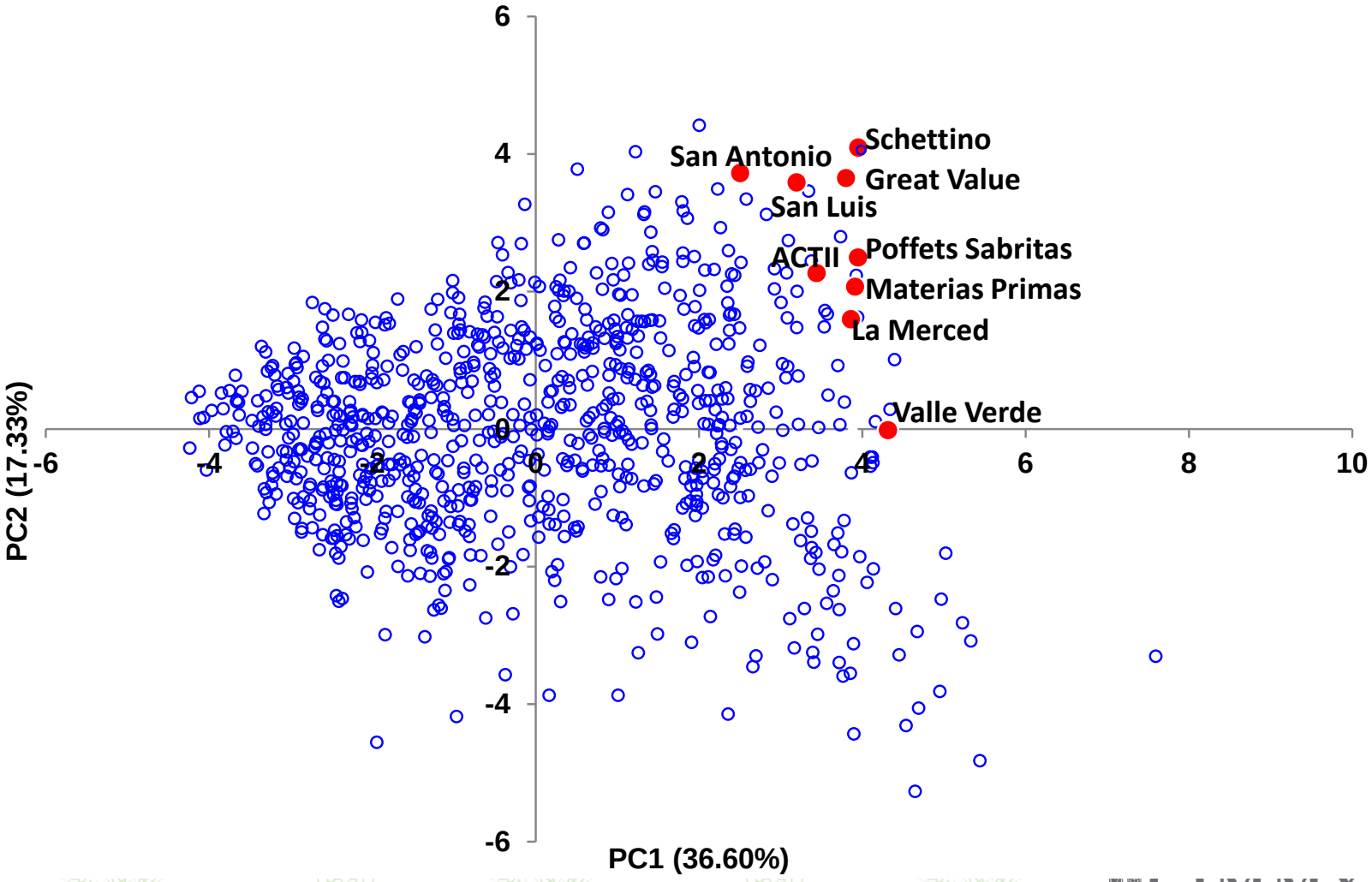
Frequency

663 14 12 193

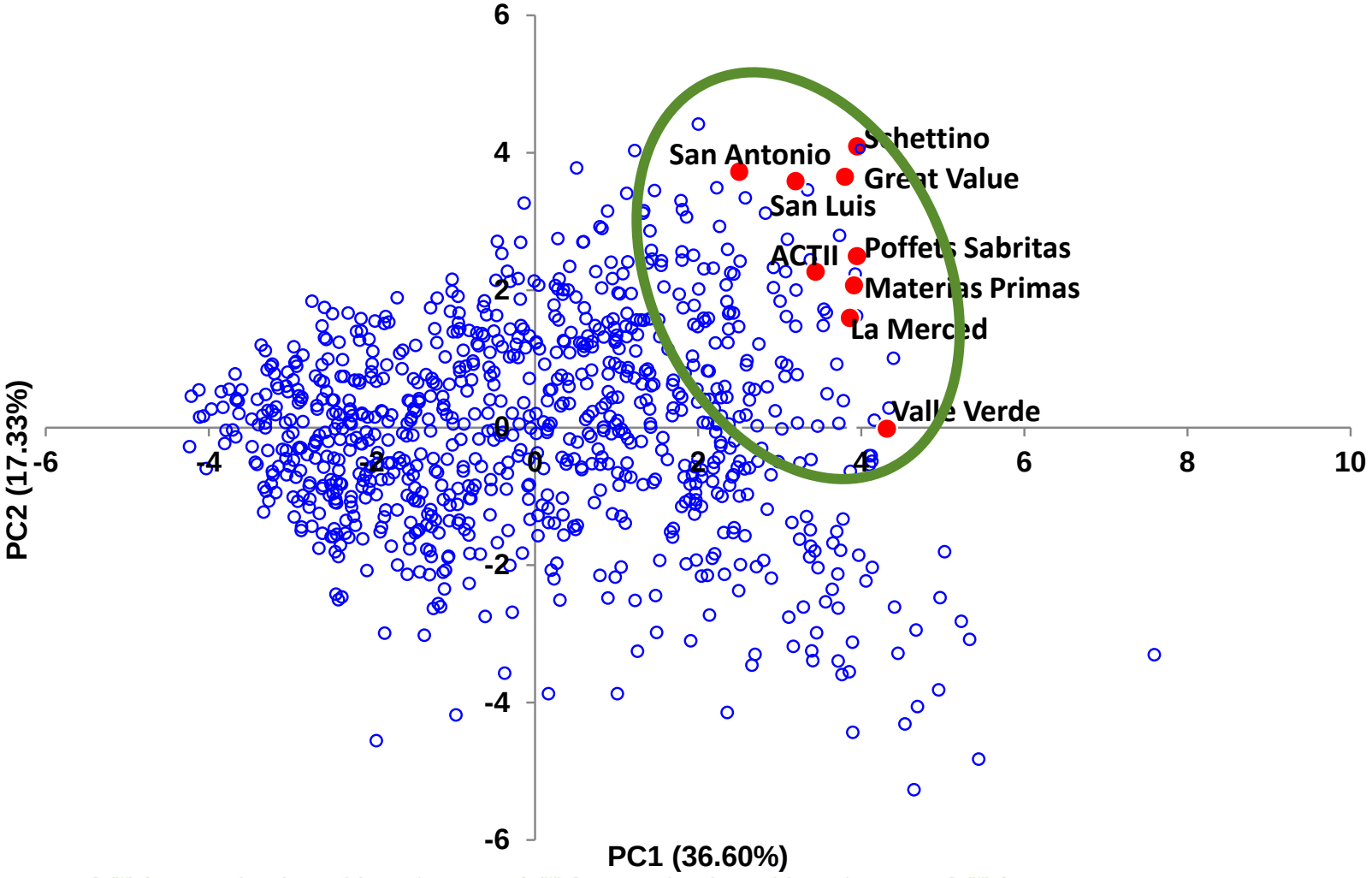
CIMMYT

Principal Components Analysis

Combines all traits to identify the best accessions



We have identified landrace popcorns that are equivalent or better than commercial checks!



Genetic study of complex traits of popcorn

- Mapping individual genetic factors with small effects on the quantitative traits, to specific chromosomal segments in the genome
  - How many QTL/genes are there?
  - Where are they located in the marker map?
  - How large an influence does each of them have on the trait of interest?
  - Are they interacting with each other?
  - Are they stably expressed across environments?

2016-2017 Popcorn Nurseries in Tlaltizapan, Morelos and Celaya, Guanajuato: Making Selfed F1s for Genotyping

CIMMYT

767 selfed popcorn accessions being prepared for shipment to Iowa State U. for sequencing

CIMMYT

Raw Data Summary

- cGBS CML Samples (N=544)

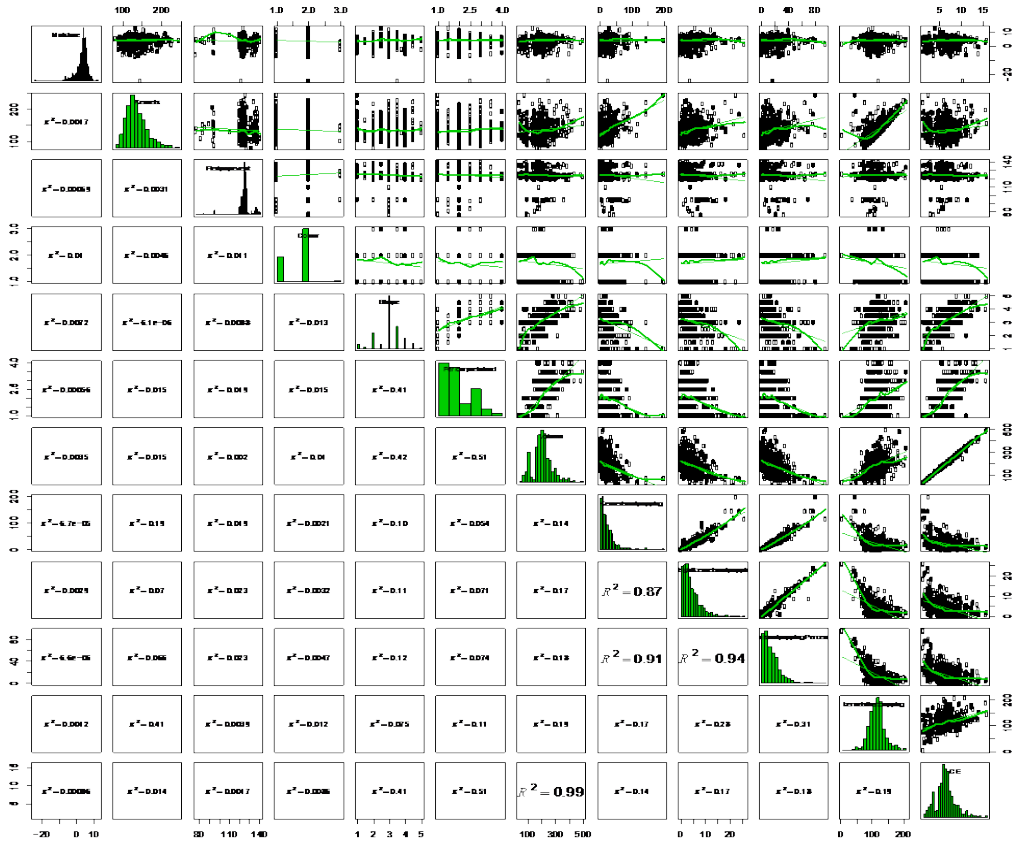
- Raw SNPs + InDel (N= 955,690)
- Raw **Imputed** SNPs + InDel (N= 955,690)

- Pop related phenotype (N=12) for 535 samples

  “Kernels“ ”Firstpopevent” “Colour” “Shape” “Pericarpretained” “Volumn” “Kernerlsnotpopping” "Weightkernelsnotpopping" "kernelsnotpoppingPercentage" "kernelsthatpopping” "CE"

CIMMYT

Phenotypes (N=12) – Distribution and Correlation



High correlated samples groups:

1, “Kernerlsnotpopping” "Weightkernelsnotpopping" "kernelsnotpoppingPercentage"

2, “Volumn” and “CE”

CIMMYT

Raw SNPs filtering

Among the 534* samples with phenotype

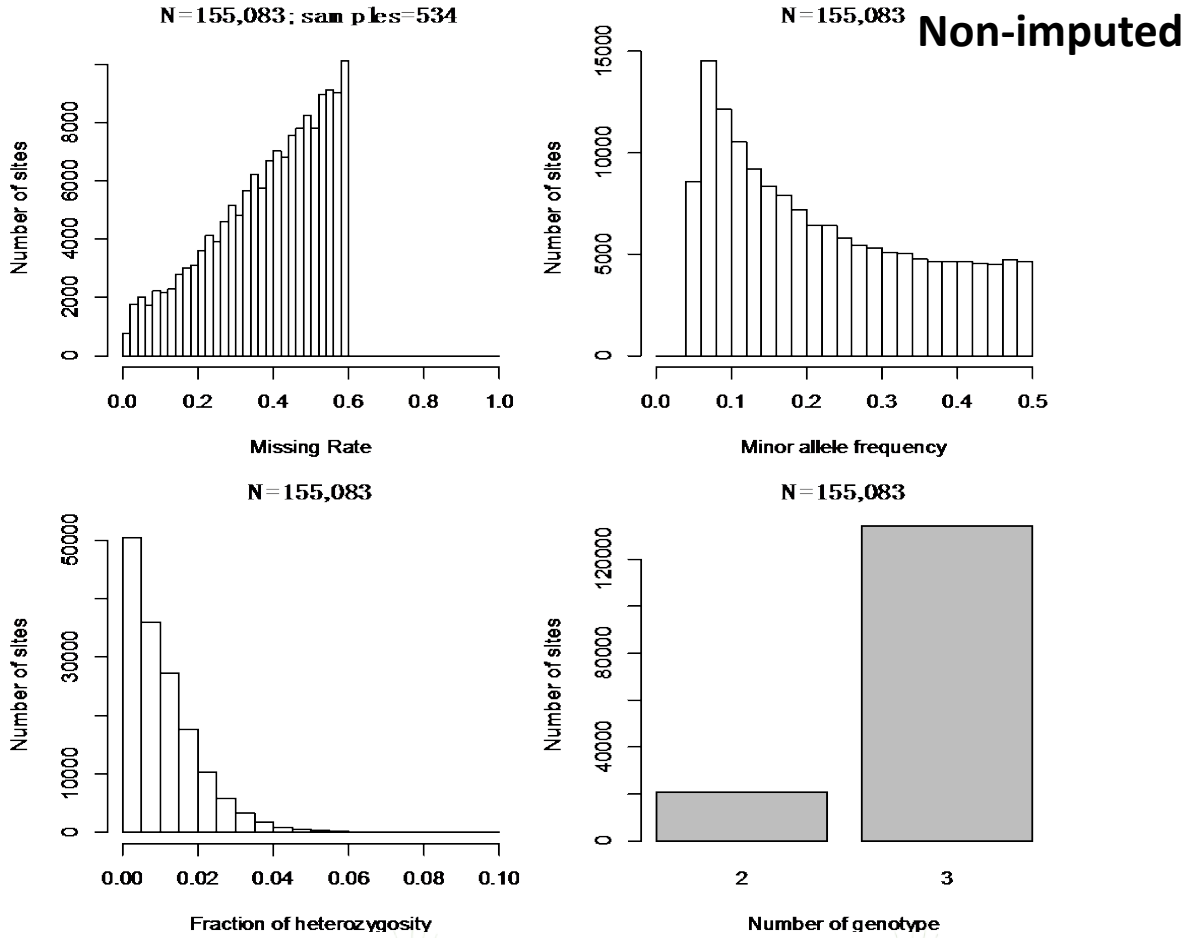
155,083 remaining SNPs

Filtering Criteria

- Missing data rate ≤ 60%
- Allele number = 2
- Genotype ≥ 2
- MAF ≥ 5%
- Heterozygosity range: 0-10%
- Mapped to B73

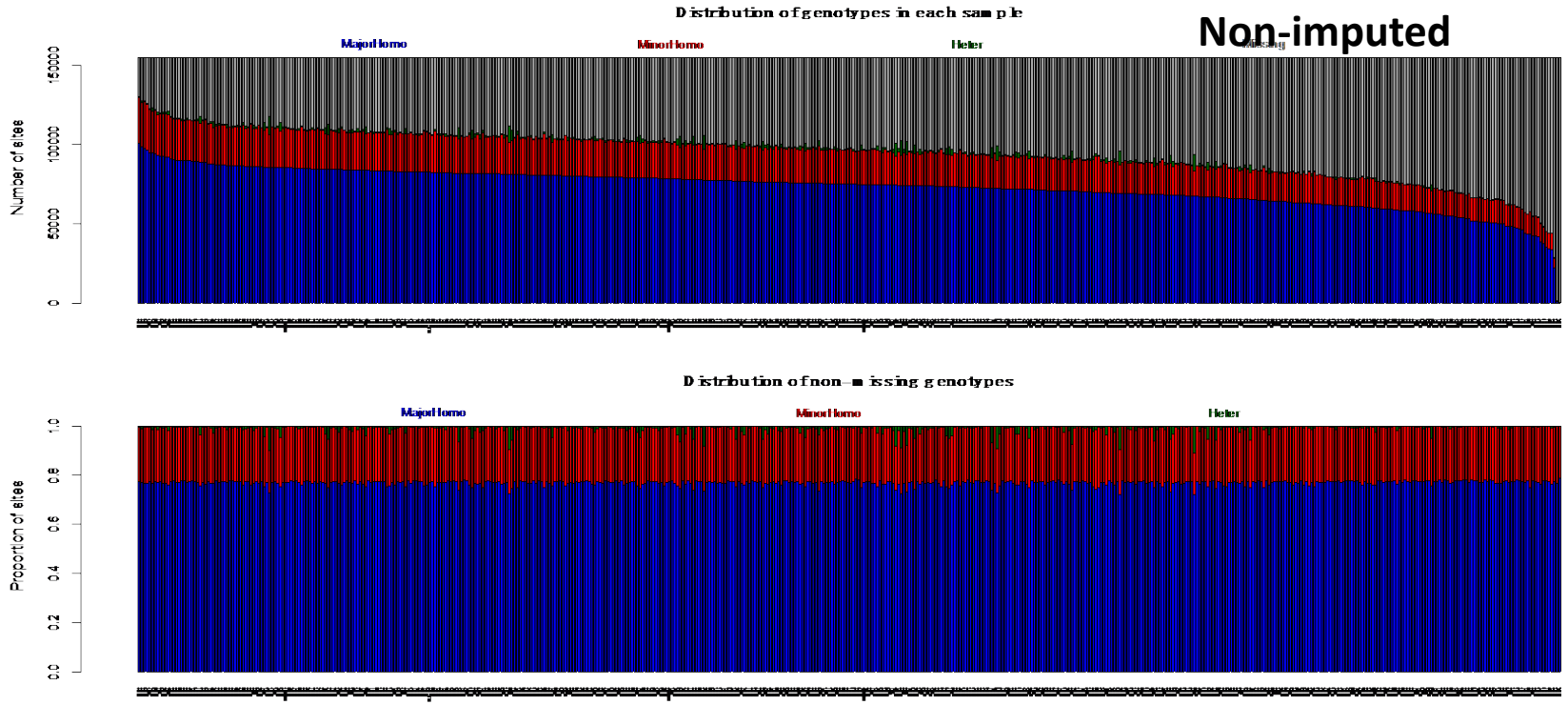
CIMMYT

Raw SNPs after filtering

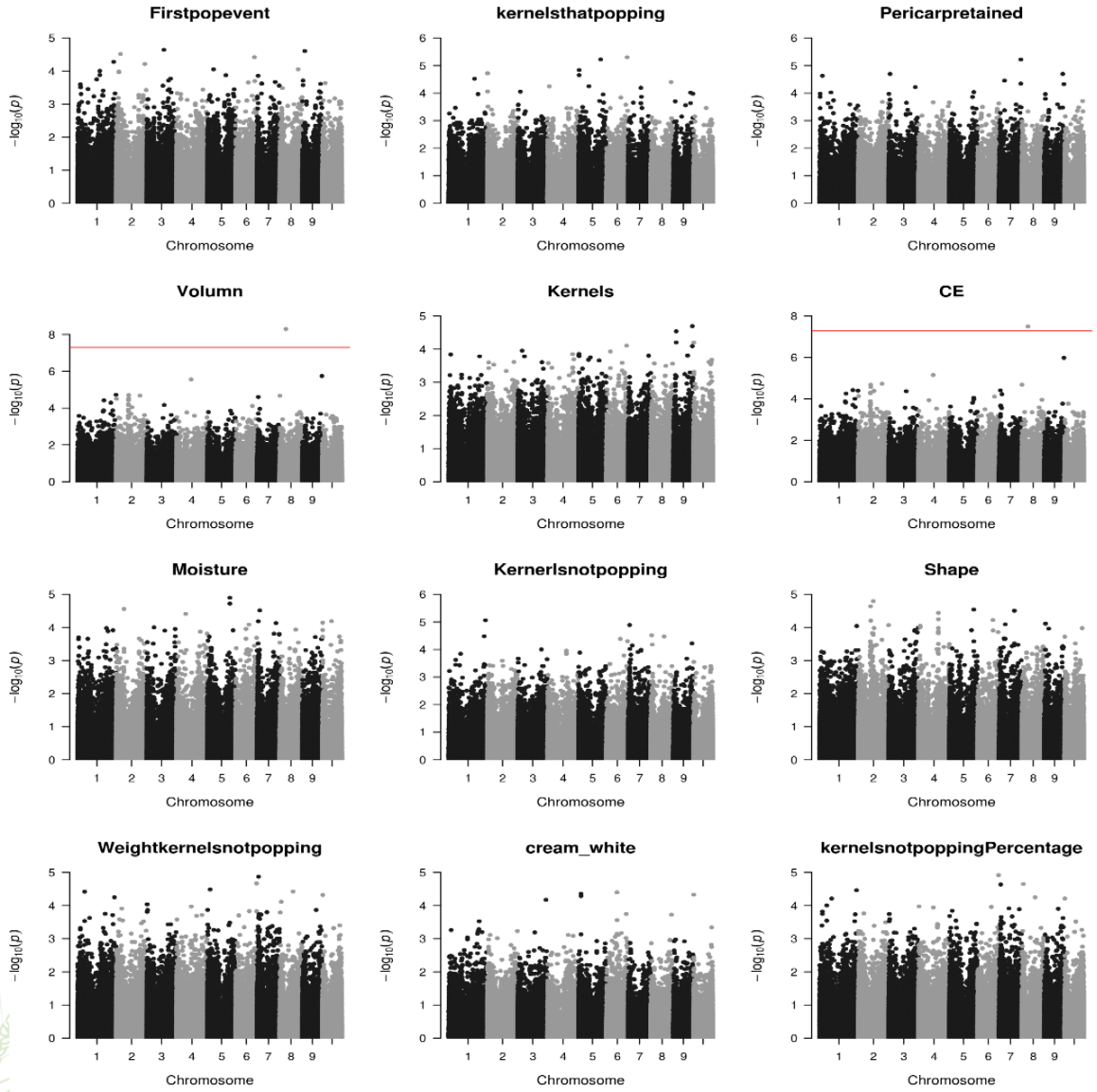


CIMMYT

Samples (N=534) Summary with filtered SNPs



GWAS Results using Non-imputed and filtered SNPs



CIMMYT

Among the 535 samples with phenotype

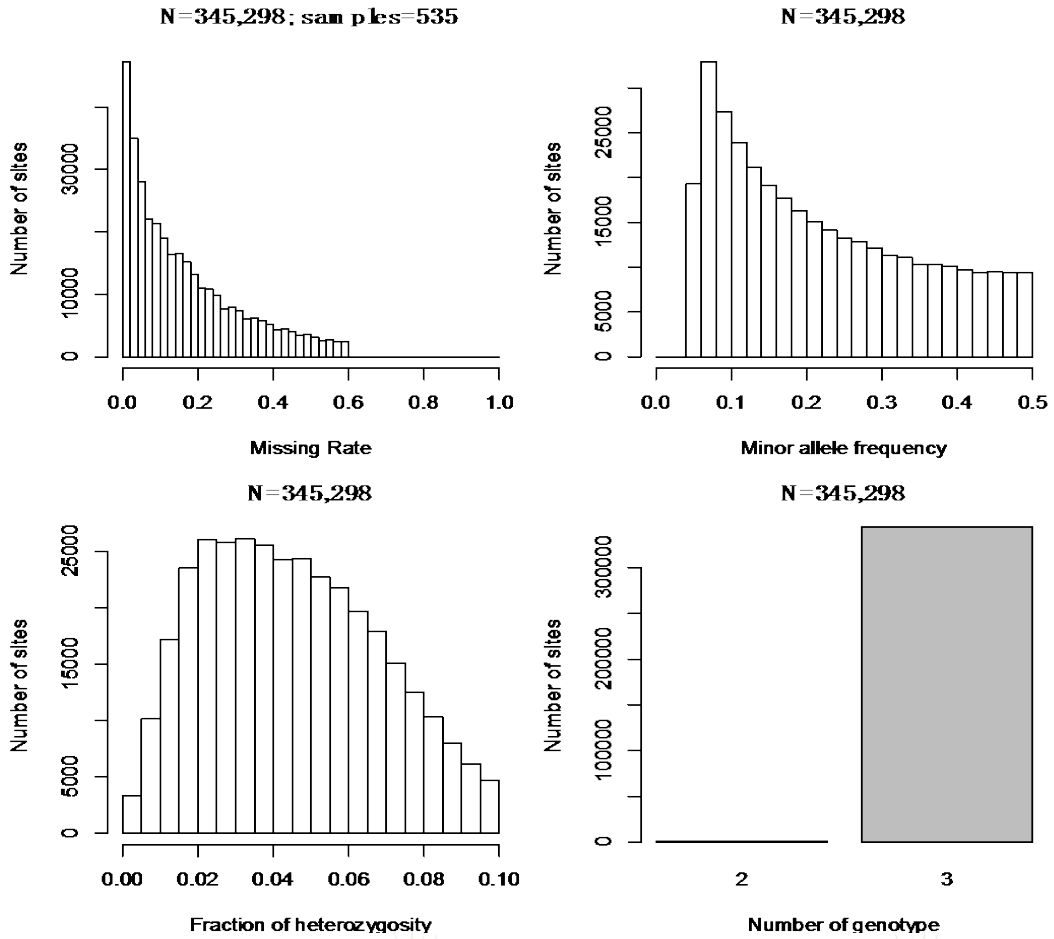
SNP filtering after imputation

345,298 remaining SNPs

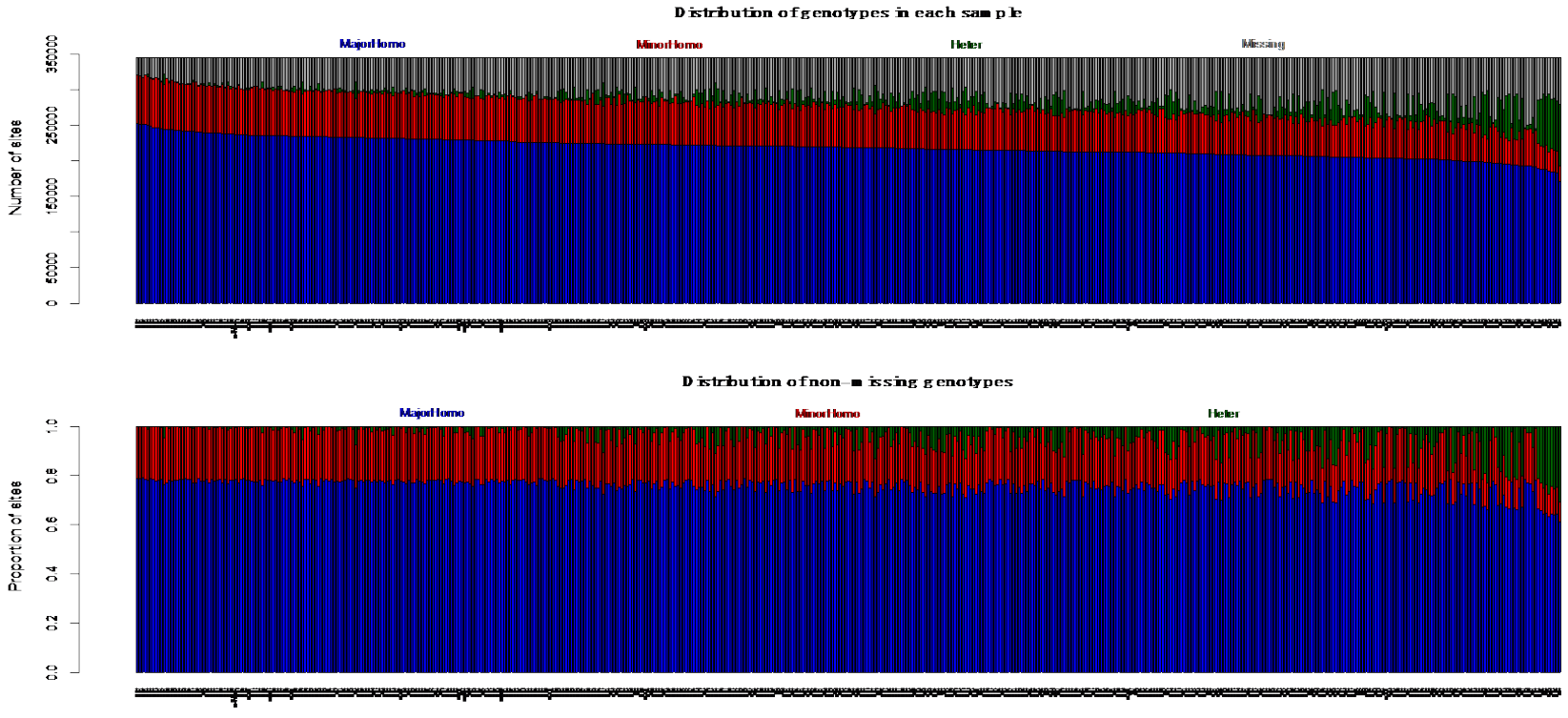
Filtering Criteria

- Missing data rate ≤ 60%
- Allele number = 2
- Genotype ≥ 2
- MAF ≥ 5%
- Heterozygosity range: 0-10%
- On 10 chromosomes

Imputated SNPs after filtering



Samples (N=535) Summary with filtered SNPs



Major Homo: the homozygous with major allele on one SNP site

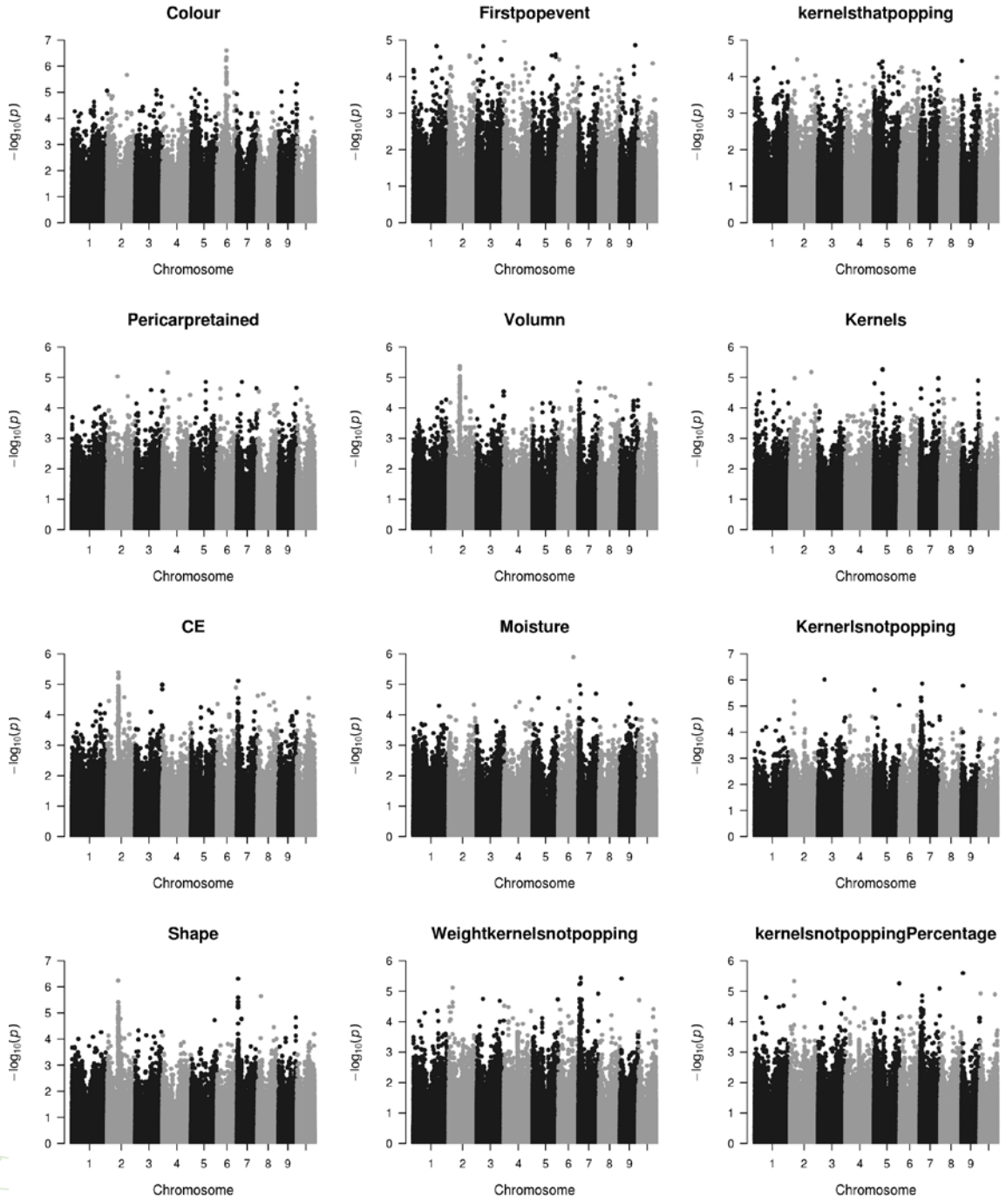
Minor Homo: the homozygous with minor allele on one SNP site

CIMMYT

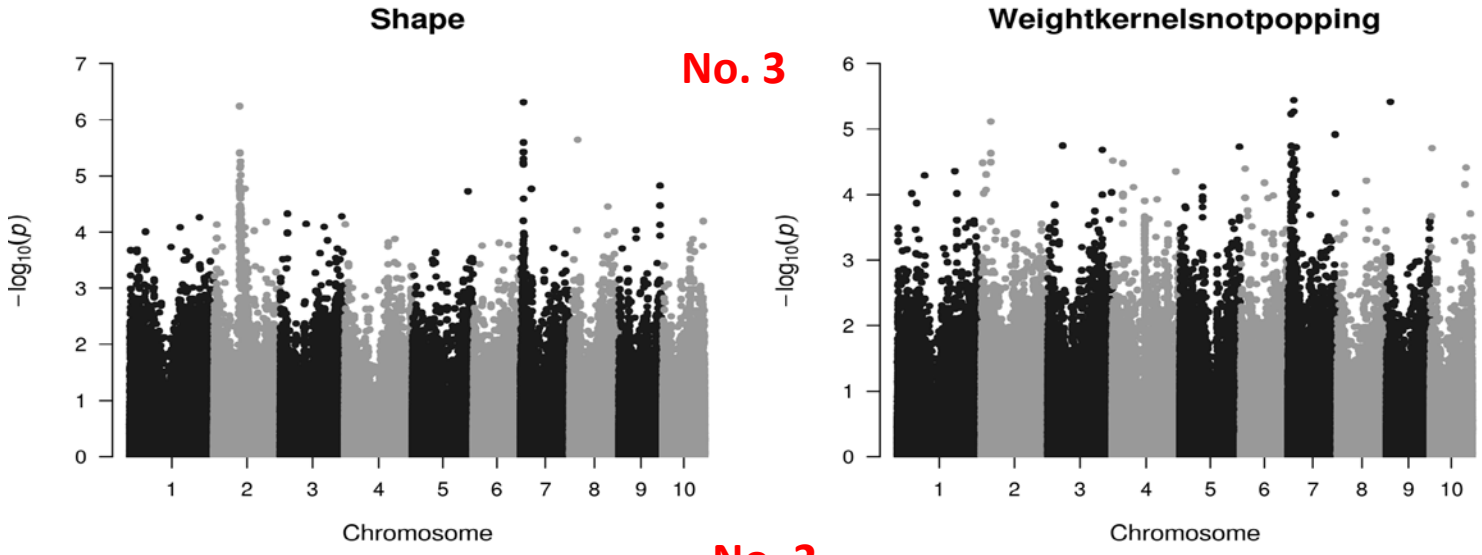
GWAS Results

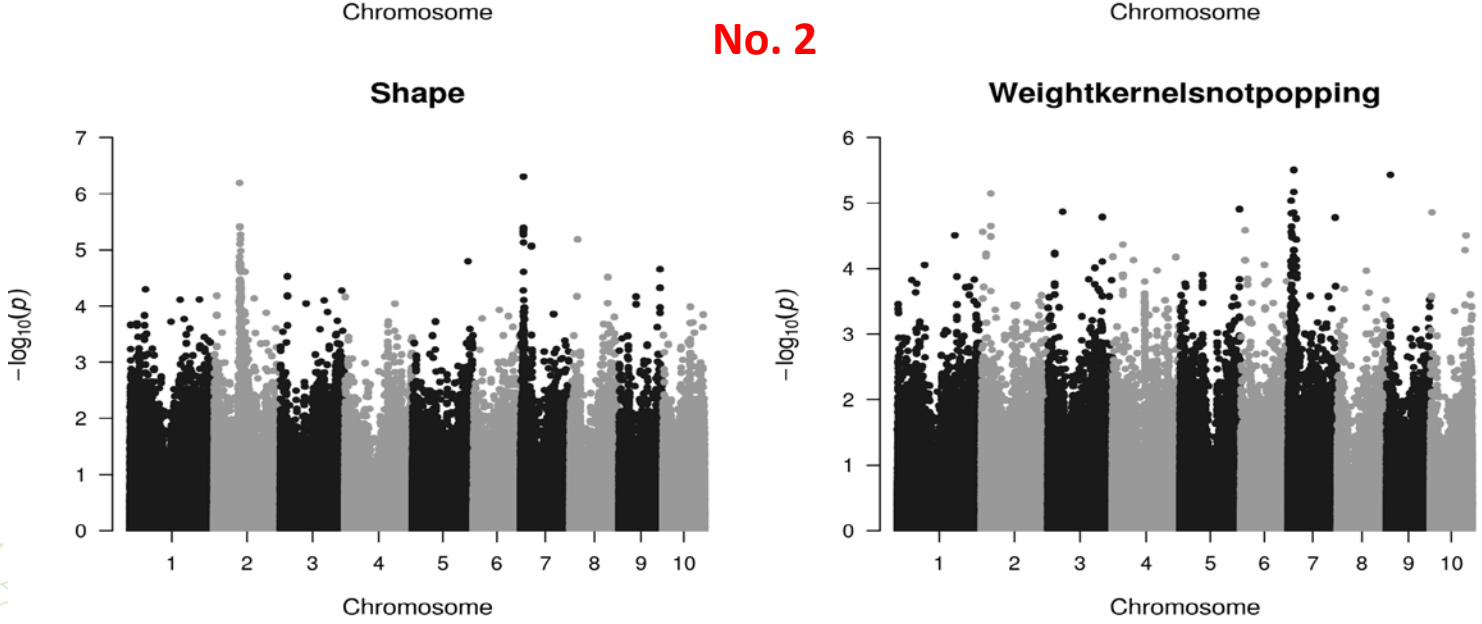
Kinship, EMMA;
PCA number, 3
CMLM; By **GAPIT**

Repeated weak peaks were found. Like Chr2 ~92Mb and Chr7 10.6Mb for Shape, Volumn and CE.

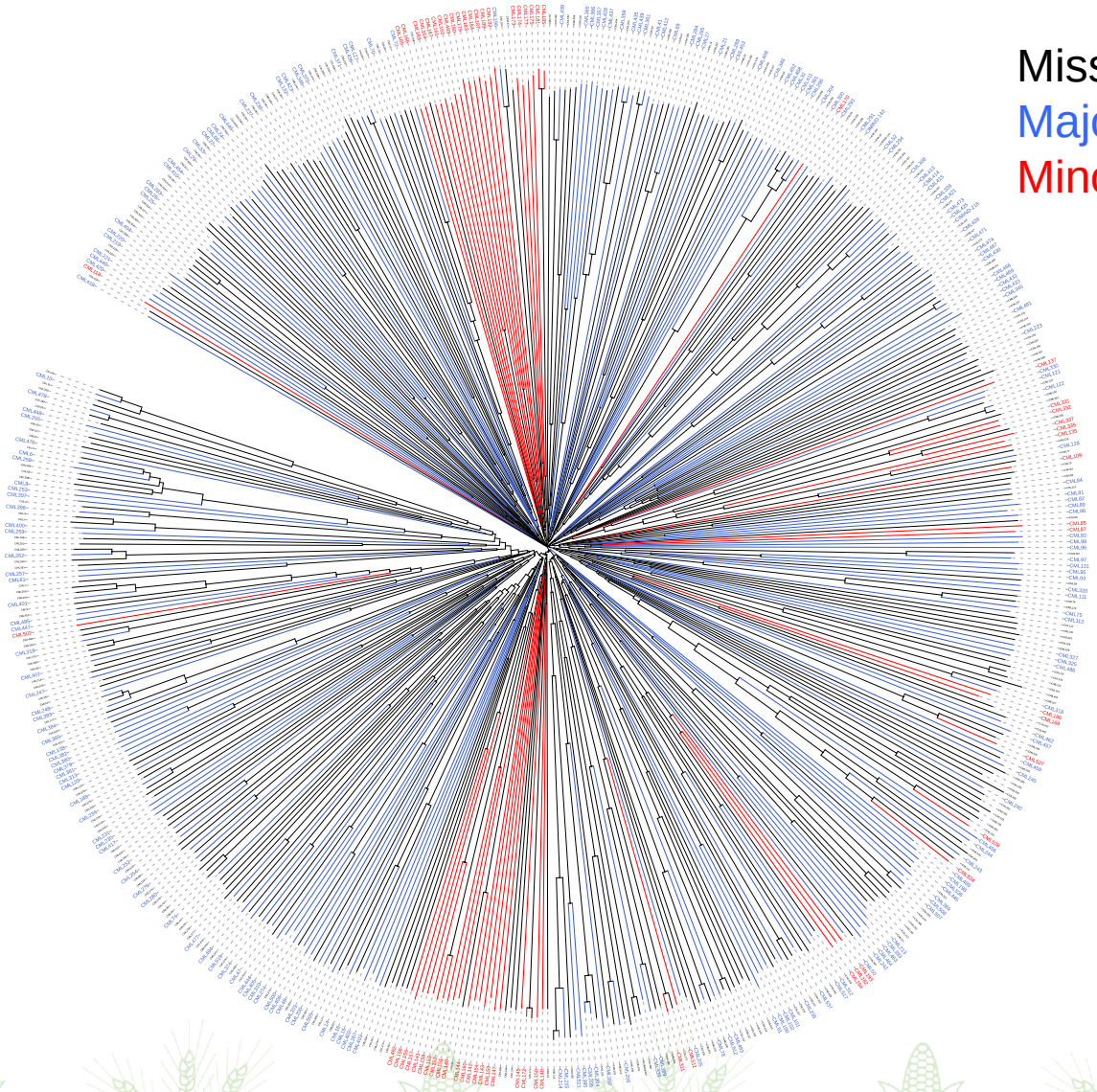


Different No. For PCA Population Structural Control





Allele (“S7_10545939”) Category



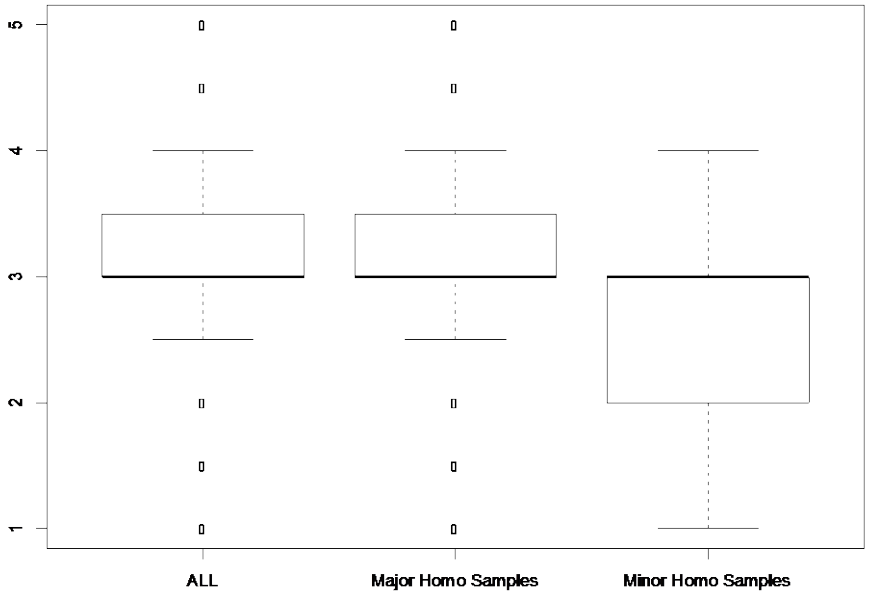
Missing (N=251) or Heter (N=3)

Major Allele Homozygous (N=216)

Minor Allele Homozygous (N=64)

CIMMYT

Allele (“S7_10545939”) Type vs. Shape Value

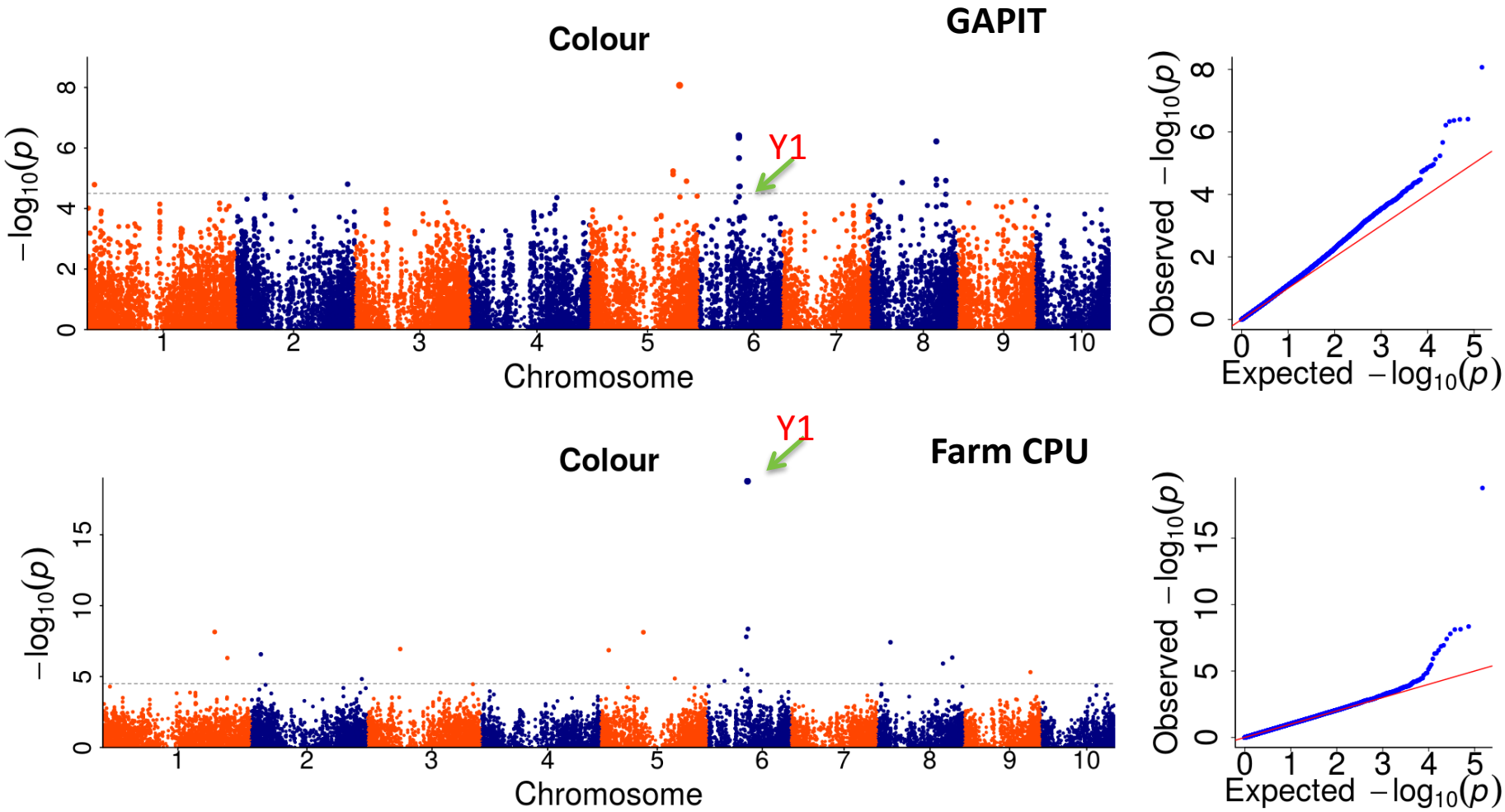


Sample Number in each Shape Category

	1	1.5	2	2.5	3	3.5	3.75	4	4.5	5	Total
ALL	19	8	72	20	240	102	1	46	22	4	534
Major Homo	3	2	35	5	90	48	0	16	14	3	216
Minor Homo	10	1	10	7	25	9	0	2	0	0	64

CIMMYT

GWAS -Color

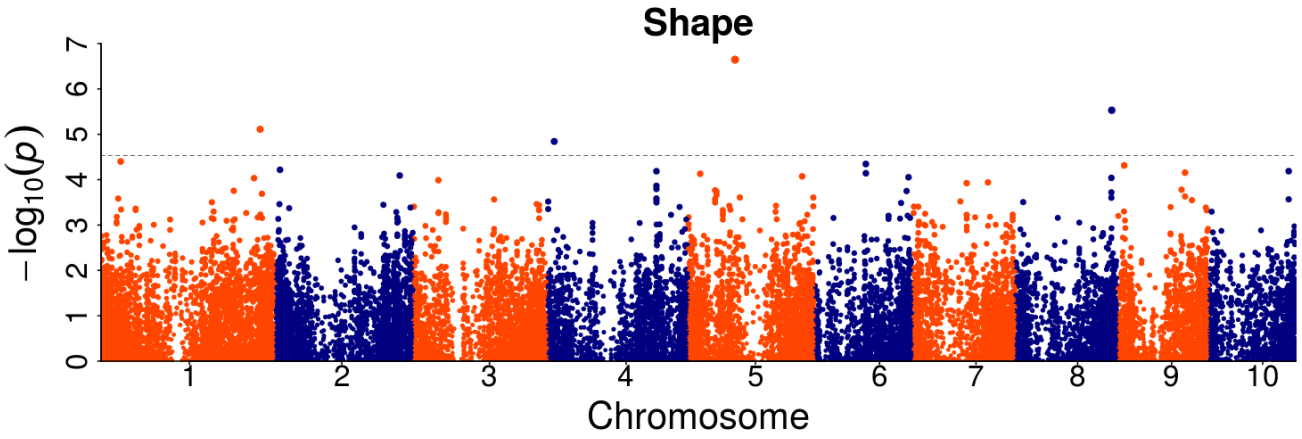


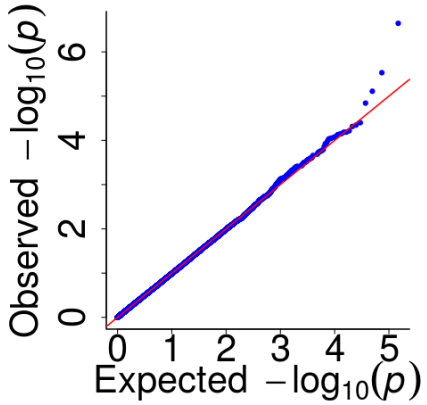
Only use **White & CREAM**

CIMMYT

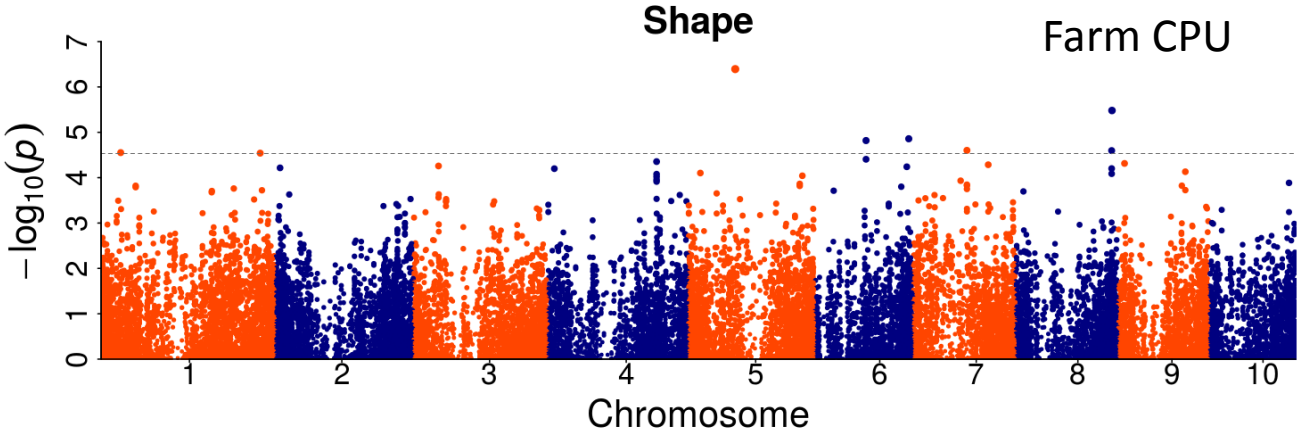
GWAS – Shape

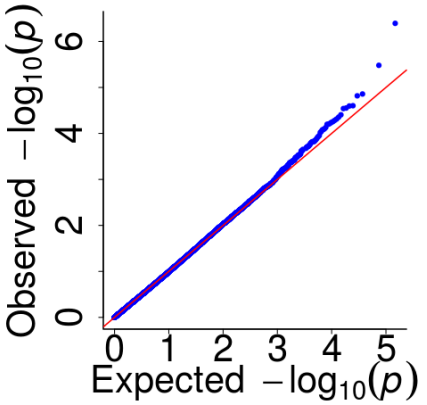
GAPIT





Farm CPU





CIMMYT

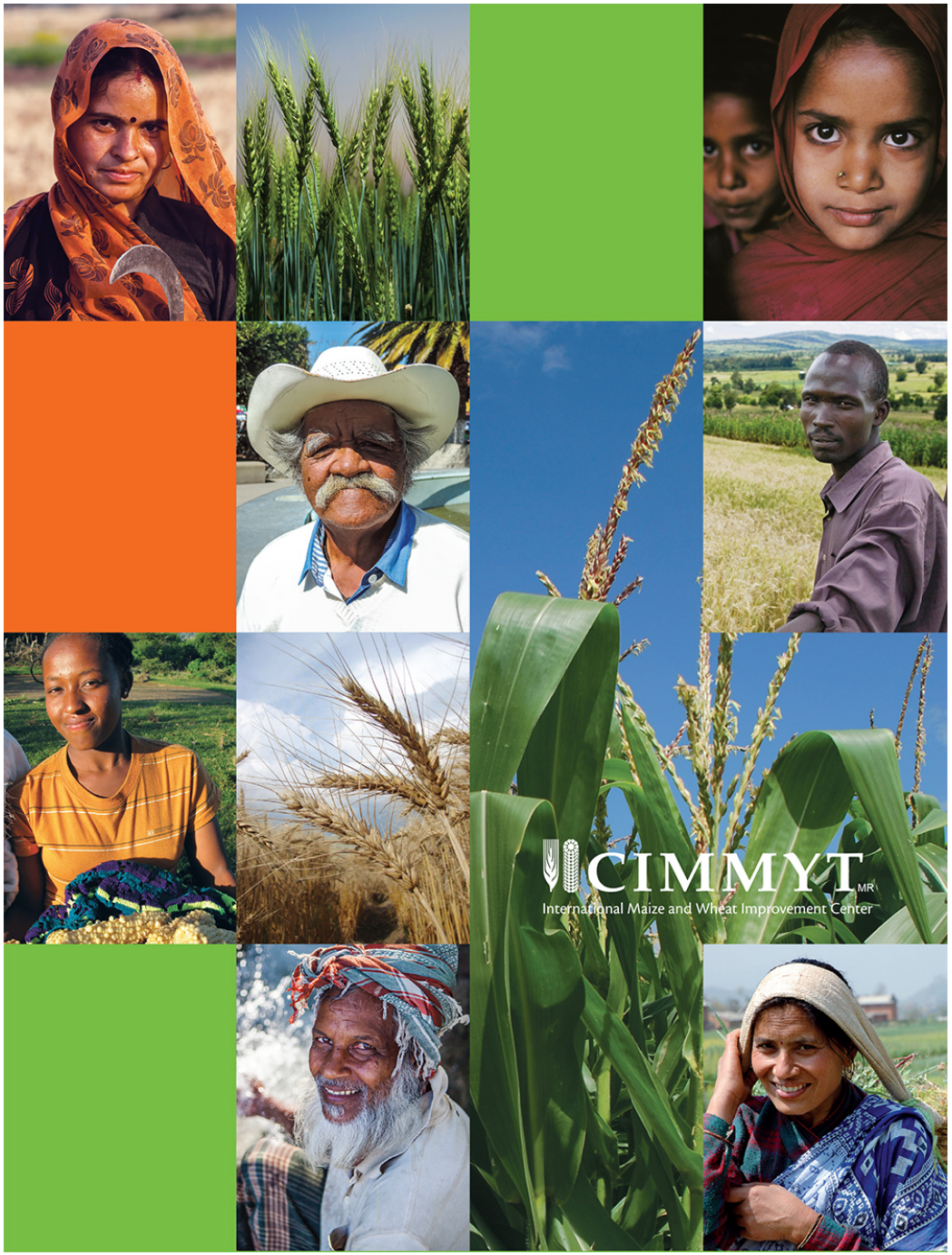
Acknowledgements – Popcorn Team

Scientists:
Denise Costich, CIMMYT-Mexico
Amalio Santacruz Varela, COLPOS
Ernesto Preciado, INIFAP-Celaya
Natalia Palacios, CIMMYT-Mexico

Students:
Natalia Almeida
Braulio Torres
Viridiana Trejo
J. Dennis Baldwin
Jing Li

Genebank Staff:
Marcial Rivas
Cristian Zavala
Alex Velazquez
Alfredo Segundo
Alejandro Velazquez

CIMMYT



Thank you for your interest!

Photo Credits (top left to bottom right): Julia Cumes/CIMMYT, Awais Yaqub/CIMMYT, CIMMYT archives, Marcelo Ortiz/CIMMYT, David Hansen/University of Minnesota, CIMMYT archives, CIMMYT archives (maize), Ranak Martin/CIMMYT, CIMMYT archives.