

## SP4 2005 Commissioned Research - Project 30

### Development of decision support systems for sampling germplasm

---

#### Involved institutions

##### **CIRAD – Agropolis**

UR75 Biometrics and Computer science unit

J.P. Jacquemoud-Collet    [jean-pierre.jacquemoud-collet@cirad.fr](mailto:jean-pierre.jacquemoud-collet@cirad.fr)

[Xavier Perrier](#) (task leader)    [xavier.perrier@cirad.fr](mailto:xavier.perrier@cirad.fr)

UMR PIA

Claire Billot

[claire.billot@cirad.fr](mailto:claire.billot@cirad.fr)

Brigitte Courtois

[brigitte.courtois@cirad.fr](mailto:brigitte.courtois@cirad.fr)

Monique Deu

[monique.deu@cirad.fr](mailto:monique.deu@cirad.fr)

Jean-François Rami

[jean-francois.rami@cirad.fr](mailto:jean-francois.rami@cirad.fr)

**IPGRI**

SGRP

Samy Gaiji

[s.gaiji@cgiar.org](mailto:s.gaiji@cgiar.org)

Rajesh Sood

[r.sood@cgiar.org](mailto:r.sood@cgiar.org)

**WUR**

Biometris

[Marco Bink](#)

[marco.bink@wur.nl](mailto:marco.bink@wur.nl)

## Rationale and Objective

address the needs formulated by SP1 scientists

- localize genes involved in agronomic traits, in using molecular markers as tags
- detection by **association mapping**
  - based on disequilibrium between linked loci
  - on germplasm collections to track a large allelic diversity

but...

composite collections = complex pools of genetically differentiated objects:  
wild ancestors, relatives, landraces, inbred lines, elite material ...  
accumulating various demographic/breeding events:  
selection, genetic drift, bottleneck, founder effects ...

→ Structured populations with disturbed balances of alleles generating  
**spurious associations and disequilibria even between unlinked loci**

## Rationale and Objective

Practical constraint: phenotyping is long and expensive and can concern only a small part of the collection

Combining the two problems :

**how to take advantage of the necessary sampling to minimize disequilibria due to structures?**

Two approaches: to sample in a collection / unlinked markers

- 1. Starting from a diversity tree, extract an unstructured sub tree**
- 2. Starting from disequilibria observed between loci, extract a sample minimizing these disequilibria**

## 1. maximal length sub-tree algorithm

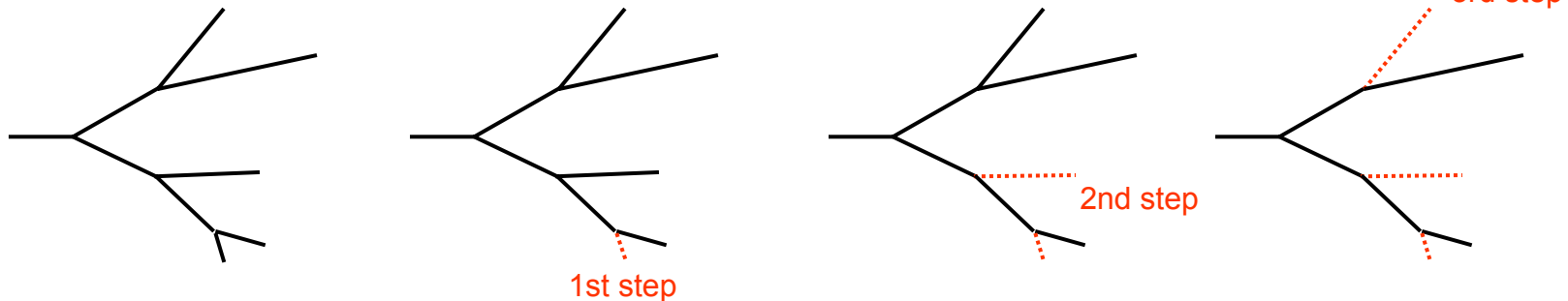
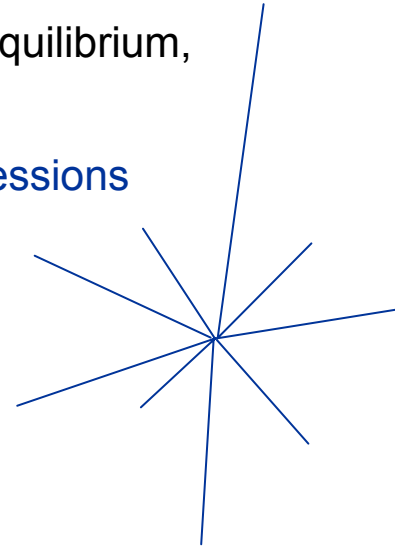
Structures = over representation of some groups → redundancy between units

It is expected that the deletion of redundant units will reduce the global disequilibrium, ( a posteriori control )

→ Search for a **star like subtree** by successive pruning of redundant accessions and of **maximal length** to maintain the allelic diversity

### ➤ Algorithm

- build a tree with a convenient method
- estimate distances between accessions in the tree
- select pair of accessions of minimal distance
- prune accession with smallest edge
- iterate on this subtree



## 1. maximal length sub-tree tree construction improvement

- Assumption of independence between markers does not hold for linked markers:

a system of weights based on map distances taking into account for correlation among linked loci

➤ Map-based weighting algorithm

- Joint analysis of several sets of variables of different nature (molecular, DNA, morphological...) and different type (ordinal, nominal, binary...)

- global dissimilarity as a weighted sum of partial standardized dissimilarities

➤ Function `wgtdaisy`  
extension of function `daisy` in S-Plus (or R)

- phenotypic diversity conditionally to genetic diversity as inferred from molecular markers

➤ Algorithm of agglomerative classification under topological constraints (Darwin)

## 2. Min SD sampling

Disequilibrium = LD 'physical' component (linkage) + SD 'structural' component (structure)

independent markers: only the structural component

→ a sample that minimizes the observed disequilibrium

stepwise algorithm removing at each step the accession

with the greatest contribution to a global disequilibrium between all pairs of loci

## 2. Min SD sampling

Linkage disequilibrium (LD) measures (haplotypes or known phase)



> diallelic loci

statistical equilibrium:

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \Leftrightarrow \frac{n_{11}}{n_{12}} = \frac{n_{21}}{n_{22}} \quad \frac{n_{11}}{n_{21}} = \frac{n_{12}}{n_{22}} \Leftrightarrow n_{11}n_{22} - n_{12}n_{21} = 0$$

$$d = n_{11}n_{22} - n_{12}n_{21}$$

		B		
		1	2	
A	1	$n_{11}$	$n_{12}$	$n_{1.}$
	2	$n_{21}$	$n_{22}$	$n_{.2}$
		$n_{.1}$	$n_{.2}$	$n$

depends on allele frequencies

several LD measures →

$$\bullet D' = \frac{|d|}{d_{\max}} \begin{cases} d_{\max} = \min(n_{1.}n_{.2}, n_{2.}n_{.1}) & \text{if } d > 0 \\ d_{\max} = \min(n_{1.}n_{.1}, n_{2.}n_{.2}) & \text{if } d < 0 \end{cases}$$

$$\bullet r^2 = \frac{d^2}{n_{1.}n_{.2}n_{2.}n_{.1}}$$

> multiallelic loci: 
$$D' = \sum_{i=1}^K \sum_{j=1}^L p_i p_j D'_{ij}$$

	$i$	non- $i$
$j$		
non- $j$		

## 2. Min SD sampling

LD measures cannot be used directly for SD measures

		B		
		1	2	
A	1	24	46	70
	2	16	14	30
		40	60	100

$$d = 24 \times 14 - 46 \times 16 = 4 \times 100$$

### Linkage disequilibrium

- allelic frequencies are fixed

- equilibrium

24+4	46-4	70
16-4	14+4	30
40	60	100

- maximum

10	60	70
30	0	30
40	60	100

standardization:  $d_{\max} = 1800$

### Structure disequilibrium

- allelic frequencies are sample dependent

- nearest equilibrium

24	46	70
16-7	14	23
33	60	93

- maximum

50	0	50
0	50	50
50	50	100

standardization:  $d_{\max} = 2500$

## 2. Min SD sampling

Structure disequilibrium  $SD = \frac{d}{d_{\max}}$

- *biallelic loci*  $d = |n_{11}n_{22} - n_{12}n_{21}|$   $d_{\max} = \frac{n^2}{4}$

n/2	0
0	n/2

- *multi allelic loci*

Two loci *I* and *J*: sum on all 2x2 subtables of 2 alleles of *I* and 2 alleles of *J*

$$d = \sum_i \sum_{i' \neq i} \sum_j \sum_{j' \neq j} |n_{ij}n_{i'j'} - n_{ij'}n_{i'j}|$$

Standardization:

$$d_{\max} = \frac{1}{2} NA(NA - 1) \left( \frac{n}{NA} \right)^2 \quad \text{with } NA = \min(K, L)$$

n/3	0	0
0	n/3	0
0	0	n/3

$$D_{\max} = 3 \frac{n}{3} \frac{n}{3} = \frac{n^2}{3}$$

depends on *K* and *L* , sensitive to rare alleles

## 2. Min SD sampling

...the algorithme

- for a pair of loci  $I$  and  $J$

contribution of accession  $k$  :  $X_k^{(IJ)} = SD_{IJ}^{+k} - SD_{IJ}^{-k}$

- score of accession  $k$  :  $Sc_k = \sum_I \sum_J w_{IJ} X_k^{(IJ)}$

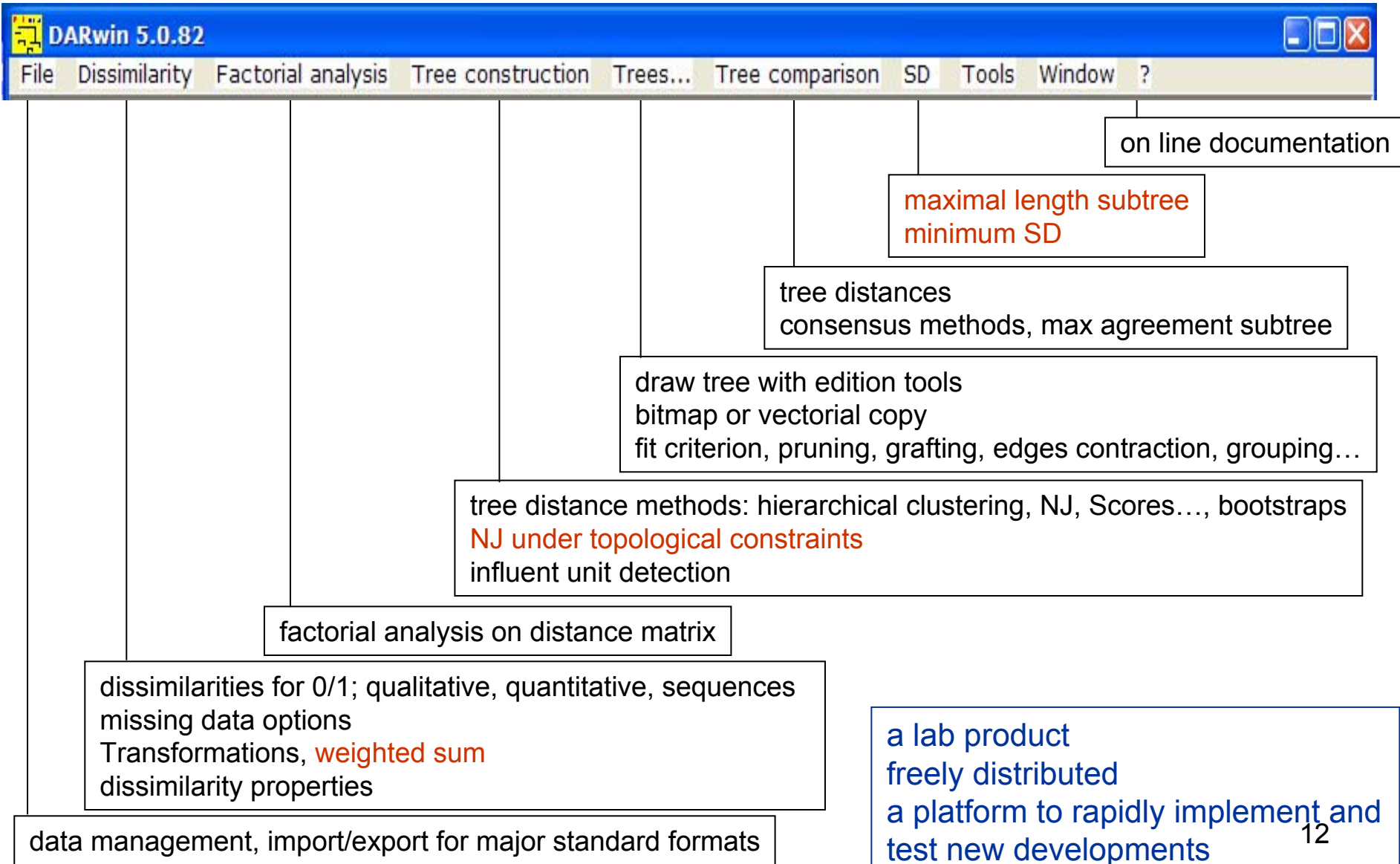
with  $w_{IJ} = SD_{IJ}^2$  to favour reduction of highest disequilibria

- remove  $k$  such that  $Sc_k$  is maximum
- reiterate on the sample

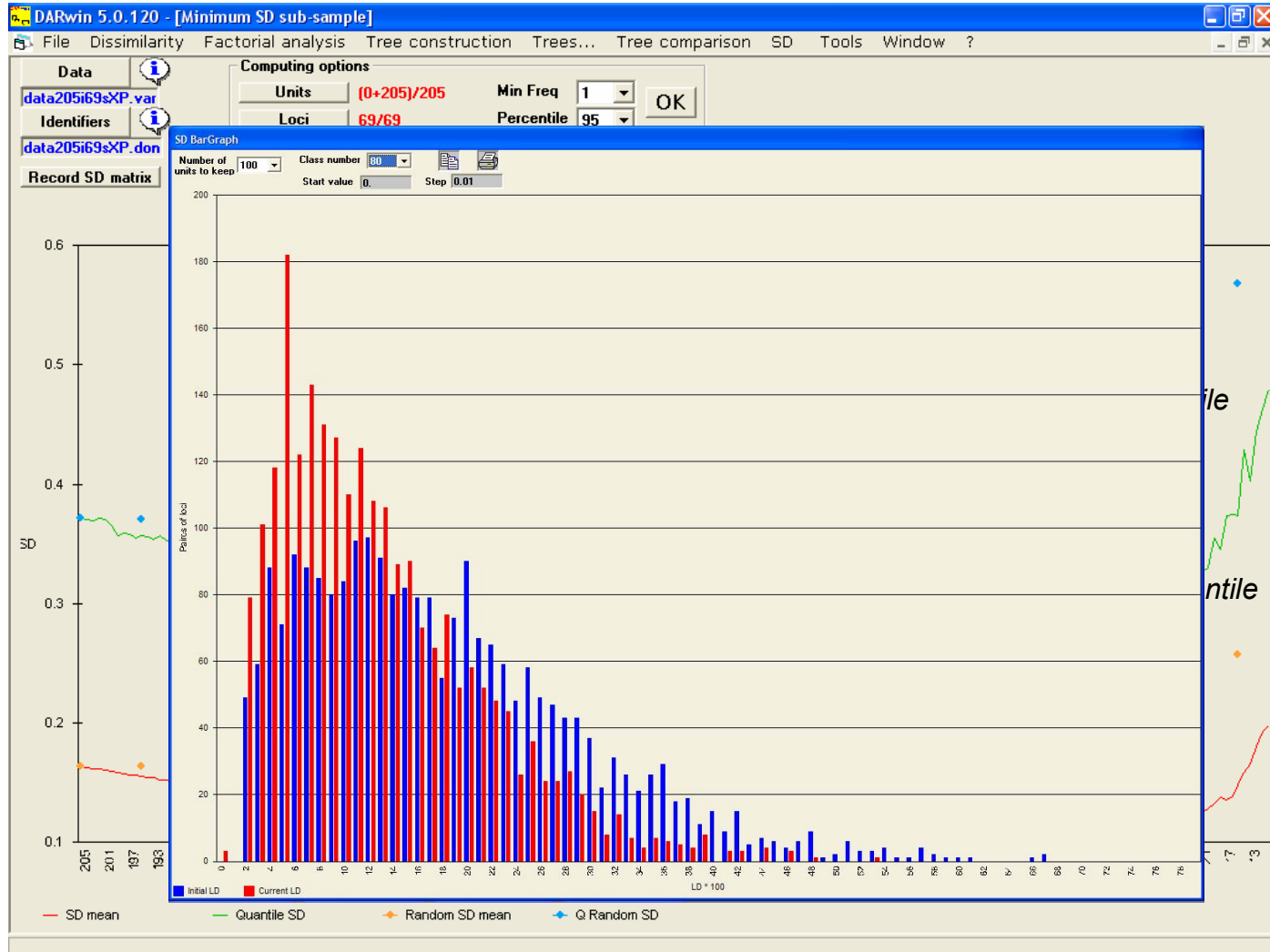
Algorithm in  $o(NM^2)$  reduced in  $o(M^2+N)$



## DARWIN software (platform Microsoft VisualStudio 6.0)



## Tools and application Min SD sampling



decreasing sample size →

## Tools and application

### Min SD sampling

INPUT, options

4 Excluded	192 Removable	9 Forced
BWA ( 21)	NER ( 67)	DZA ( 6)
CHN ( 155)	NER ( 92)	ETH ( 34)
CMR ( 90)	NER ( 112)	GHA ( 97)
CMR ( 91)	NER ( 113)	KOR ( 189)
	NER ( 114)	MWI ( 75)
	NER ( 116)	RWA ( 63)
	NER ( 124)	SWZ ( 179)
	NER ( 201)	TUR ( 190)
	NGA ( 17)	TZA ( 62)
	NGA ( 44)	
	NGA ( 45)	
	NGA ( 46)	
	NGA ( 83)	

Identifiers: data205i69sXP.don | Ident 3 | Total = 205

Random sampling:  
and drawing  
er

OUTPUT

Record SD matrix

Loci 69/69 | Percentile 95

Random sampling options

Step 10 | Number of drawing 200 | Resampling

Sample size 100 | Record Identifiers | SD

record matrix of disequilibrium between pairs of loci for the initial dataset

copy the graphic in the clipboard

sample size

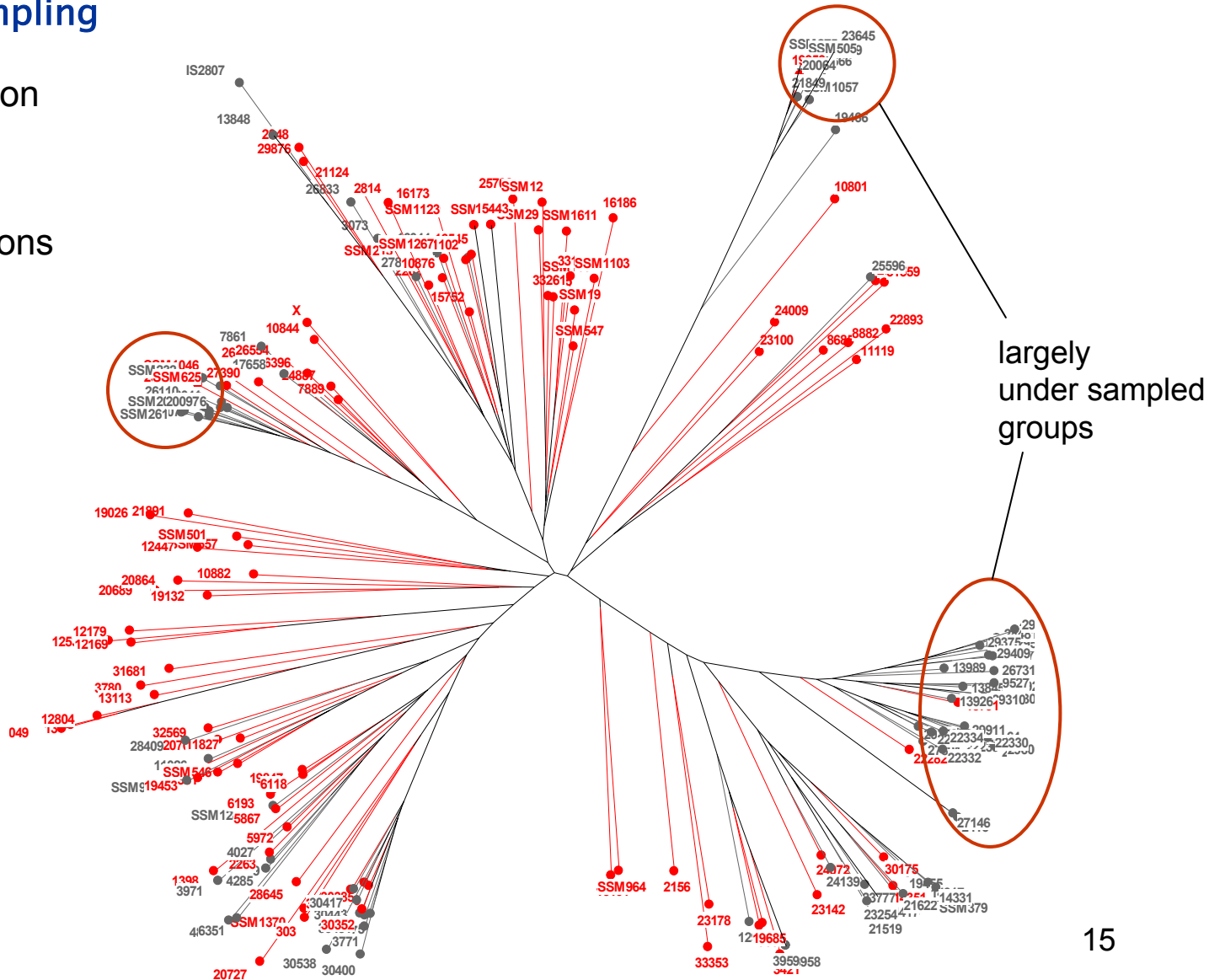
record accession status in the sample

display SD distributions for initial population and sample

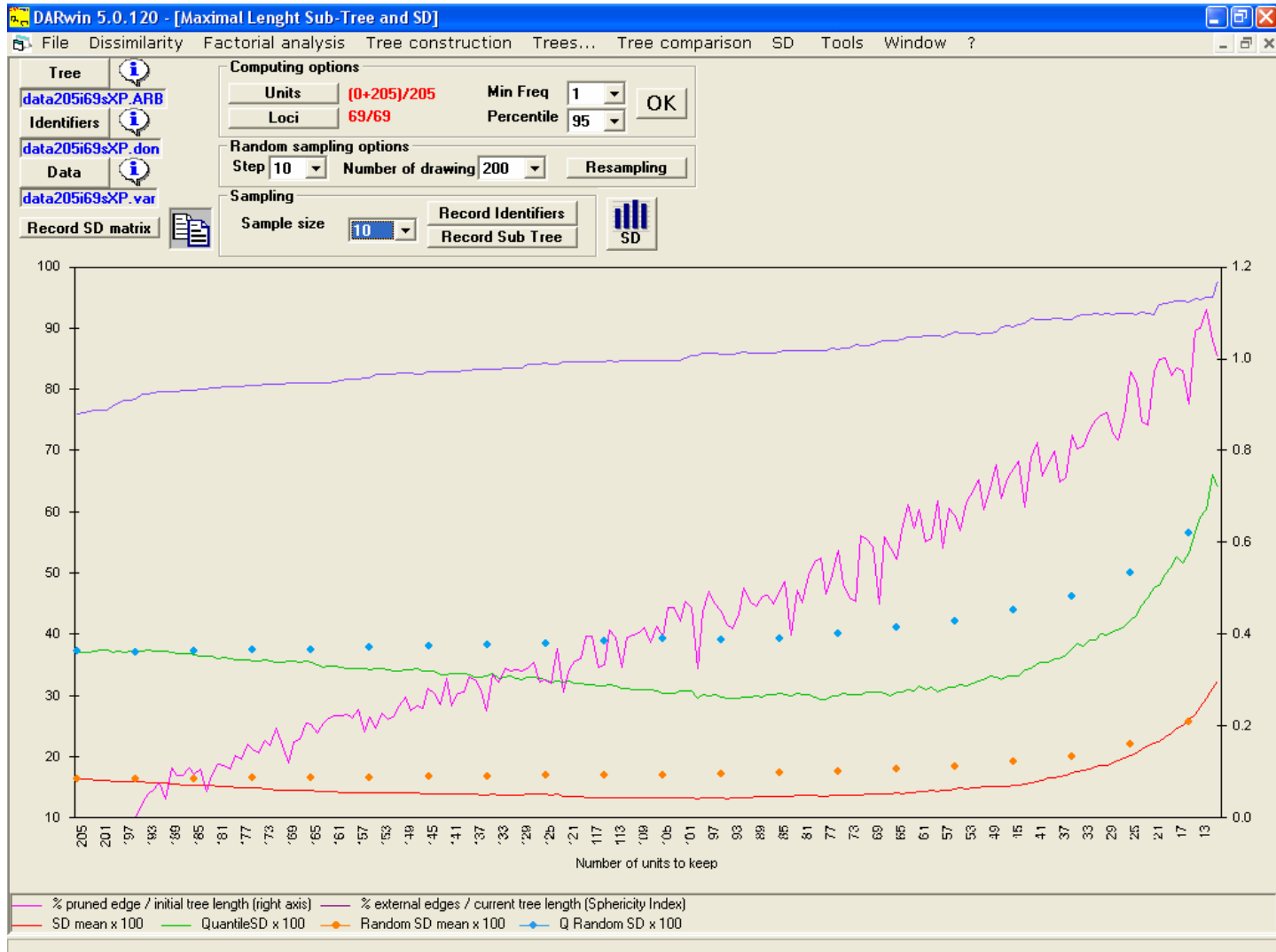
## Tools and application Min SD sampling

Sorghum core-collection

Sample of 100 accessions  
(in red)



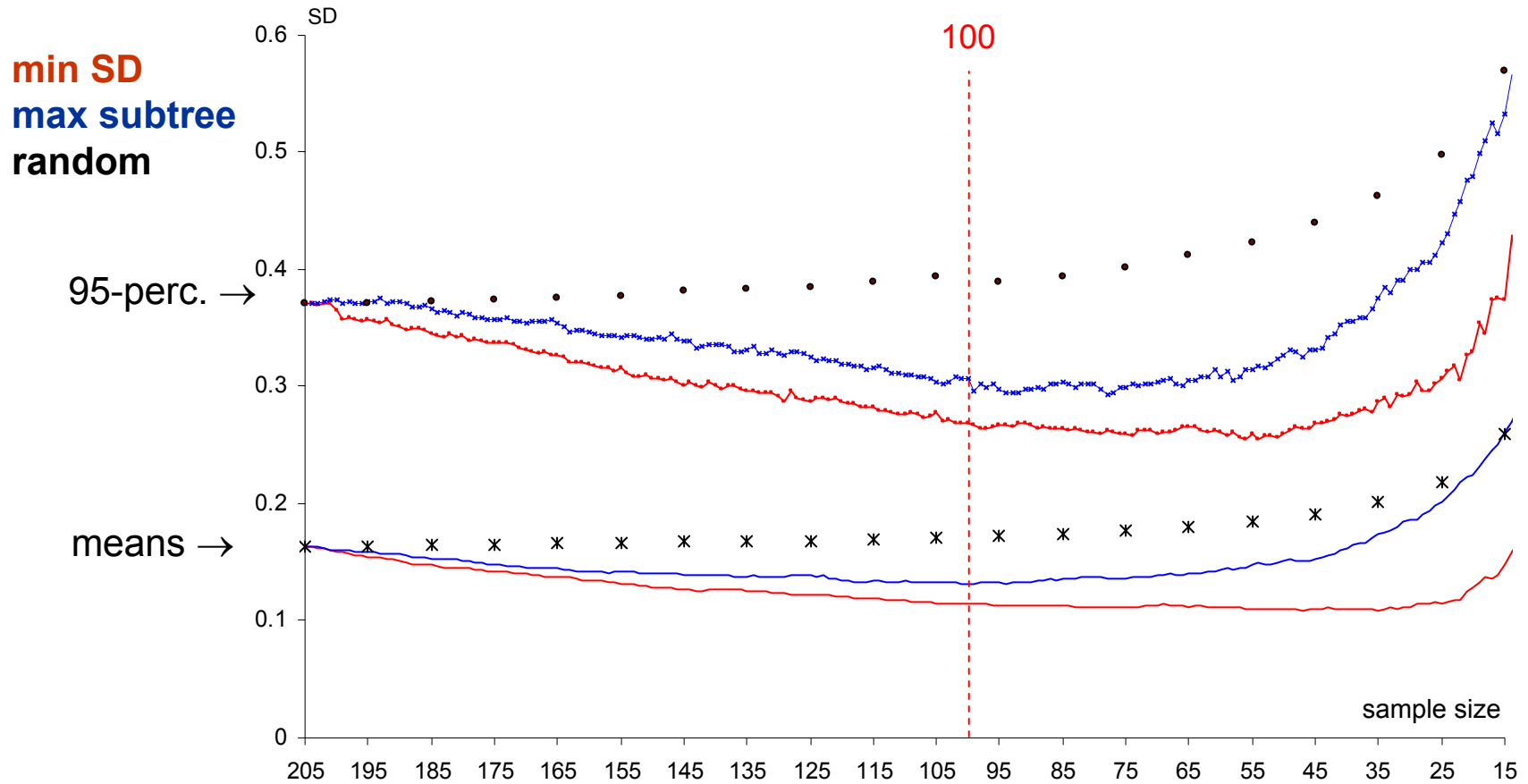
## Tools and application max length sub-tree



## Tools and application

### comparison between max subtree and min SD strategies

Sorghum core-collection

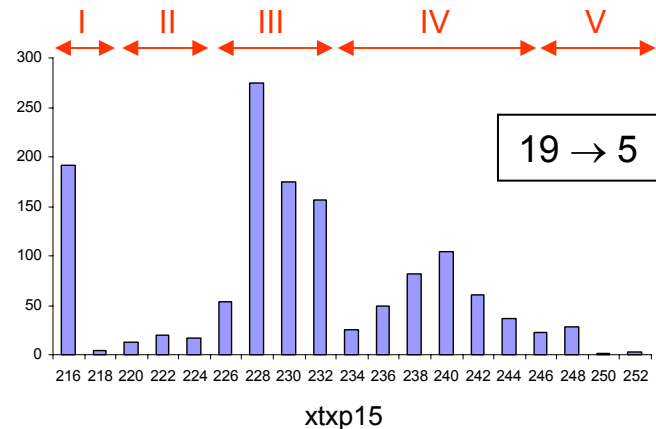


## max subtree versus SD min strategies

### Sorghum data (GCP data)

- 660 accessions (200 accessions in the RFLP data)
  - 8 dinucleotides
- 20 SSR loci
  - 10 trinucleotides
  - 2 tetranucleotides

- 272 alleles
- 4 to 30 alleles / locus (mean= 13.6)
- high number of rare alleles
  - 200 alleles below 5%
  - 132 alleles below 1%
- 651 genotypes



allele pooling →

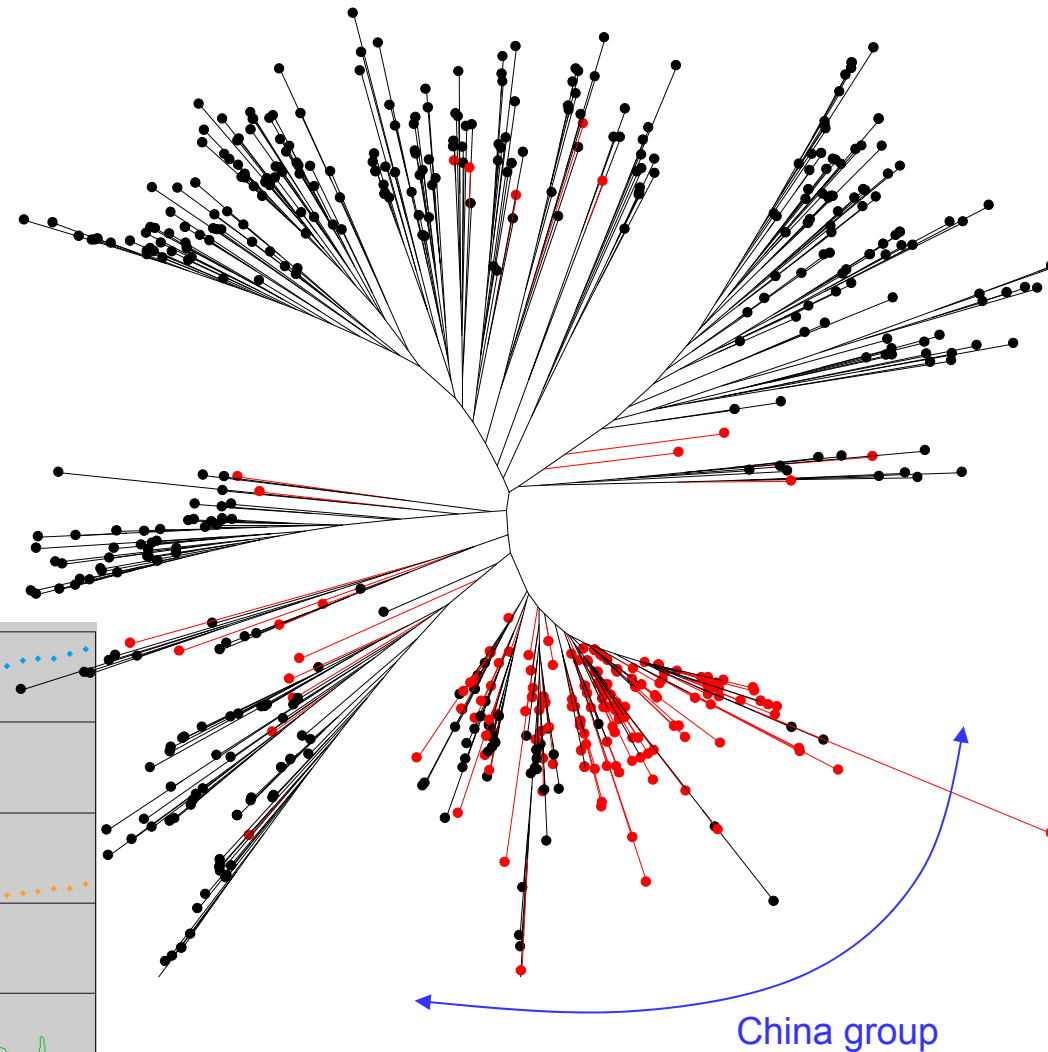
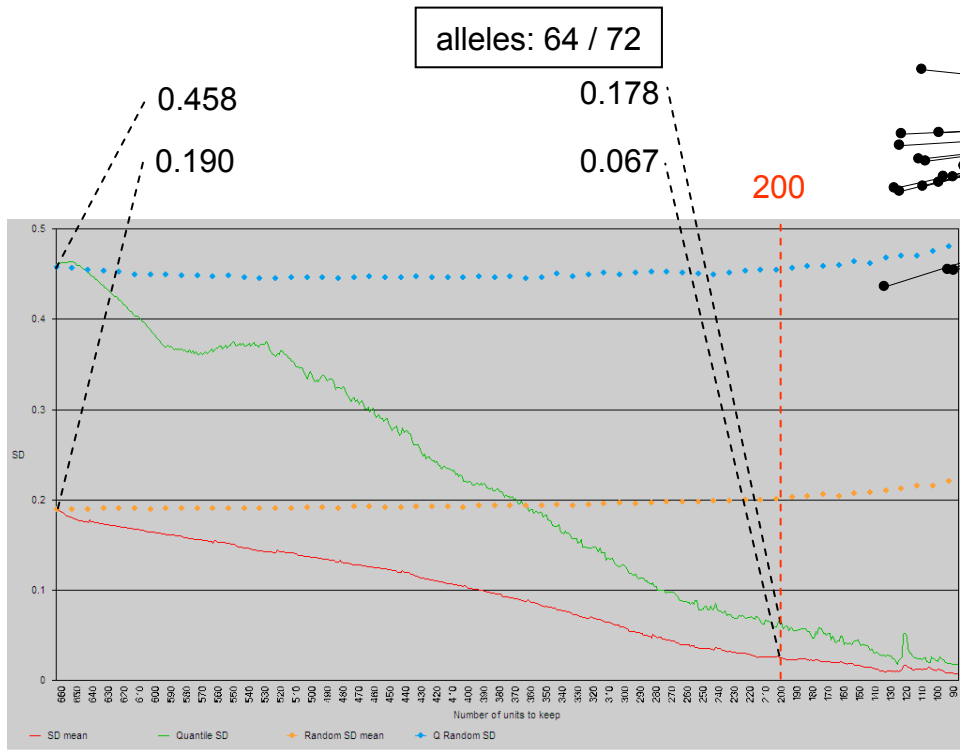
- 72 alleles
- 2 to 5 alleles / locus (mean= 3.6)
- no alleles below 5%
- 616 genotypes

## max subtree versus SD min strategies

660 Sorghum accessions / 20 SSR

Min SD sampling: sample 200 / 660

alleles: 64 / 72



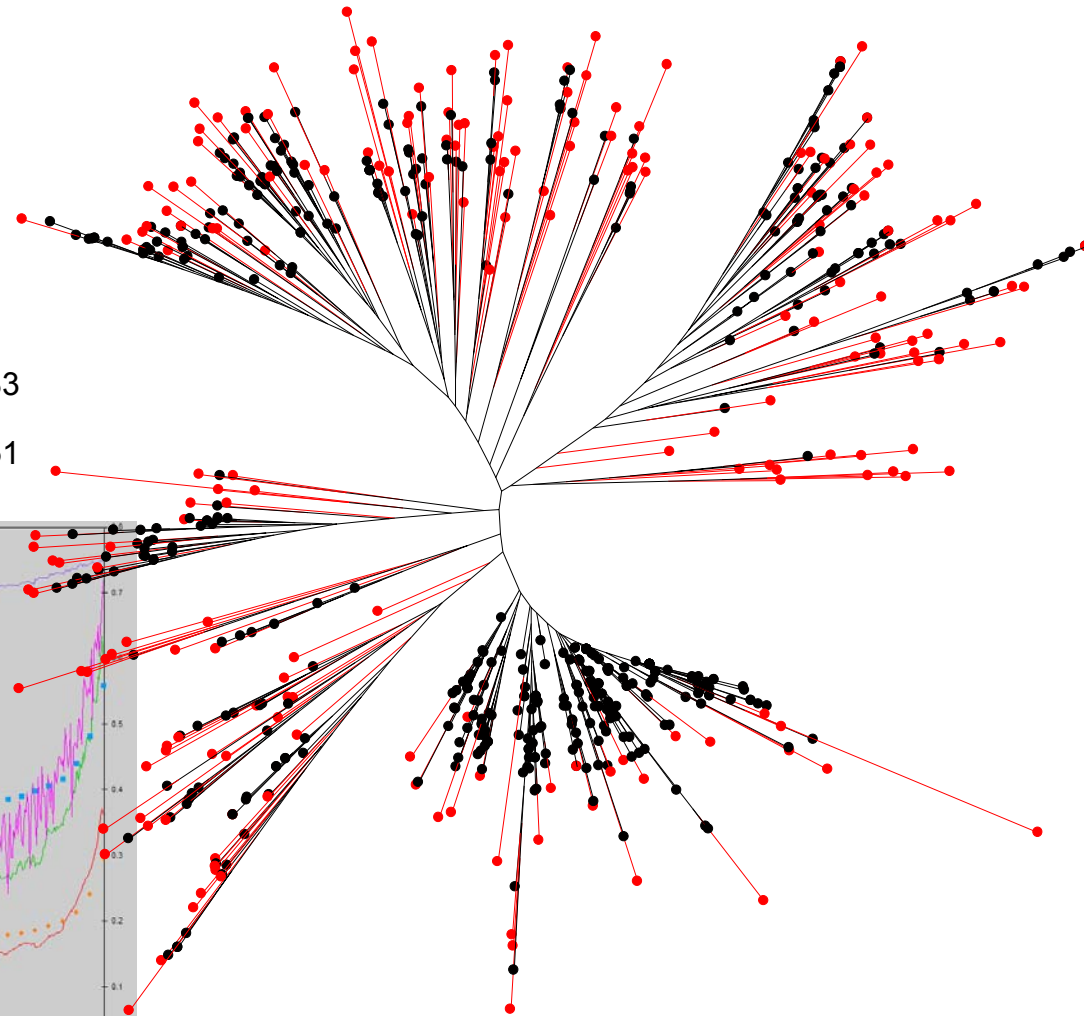
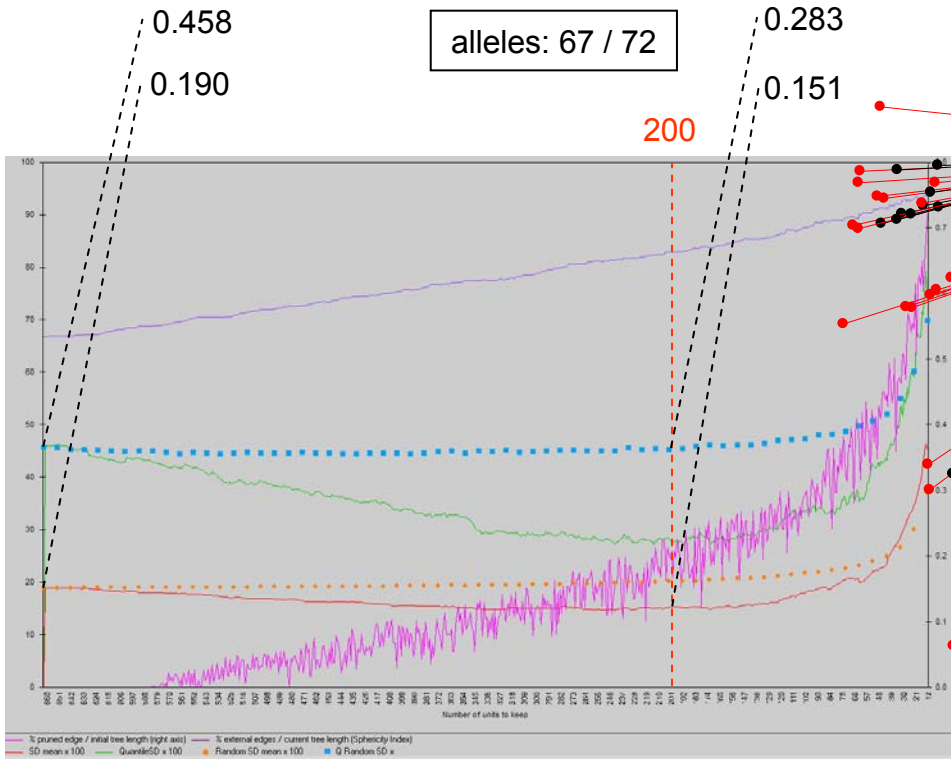
in red the 200 sampled accessions

a logical result but ...

## max subtree versus SD min strategies

660 Sorghum accessions / 20 SSR

Max subtree sampling: sample 200 / 660



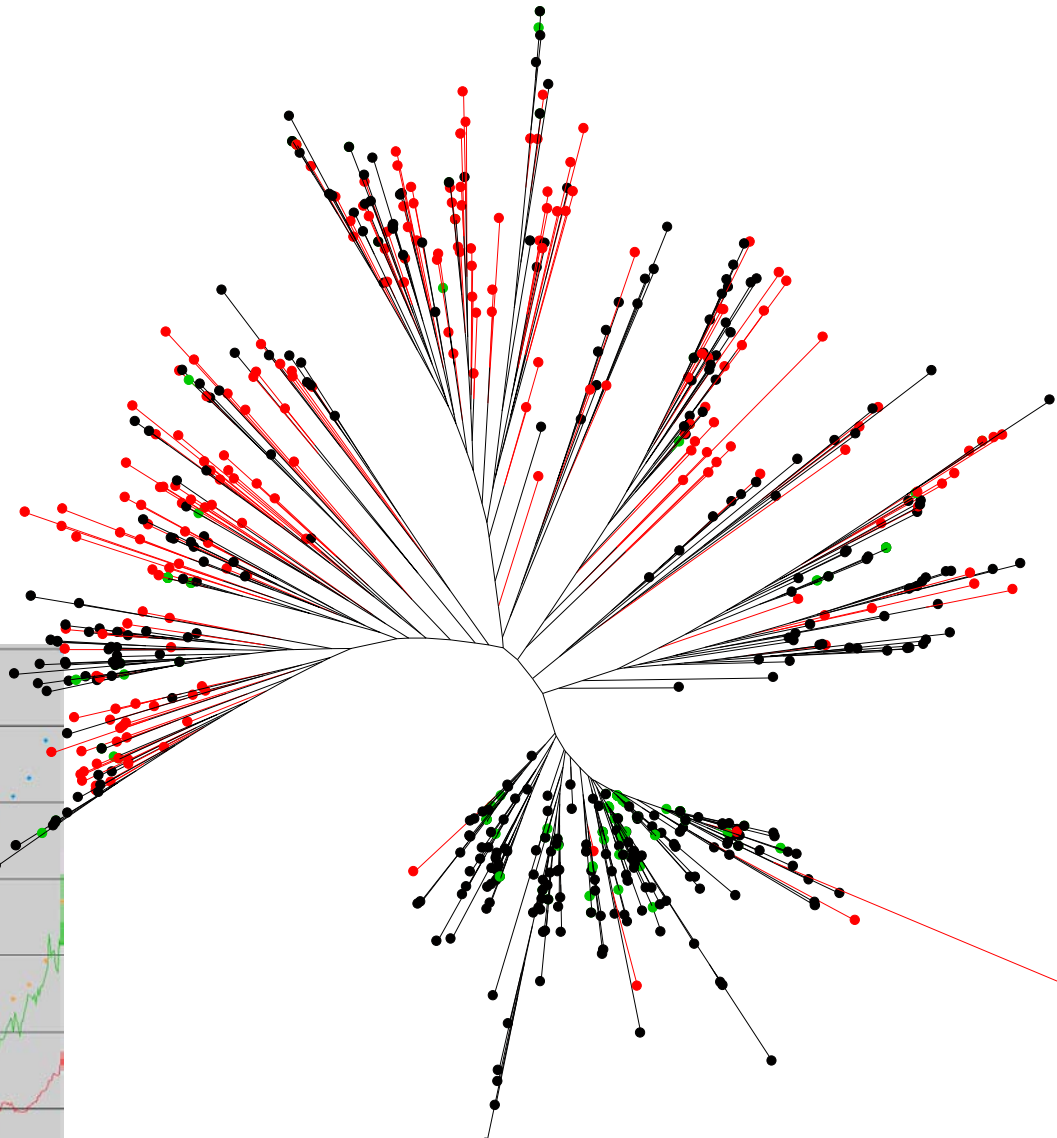
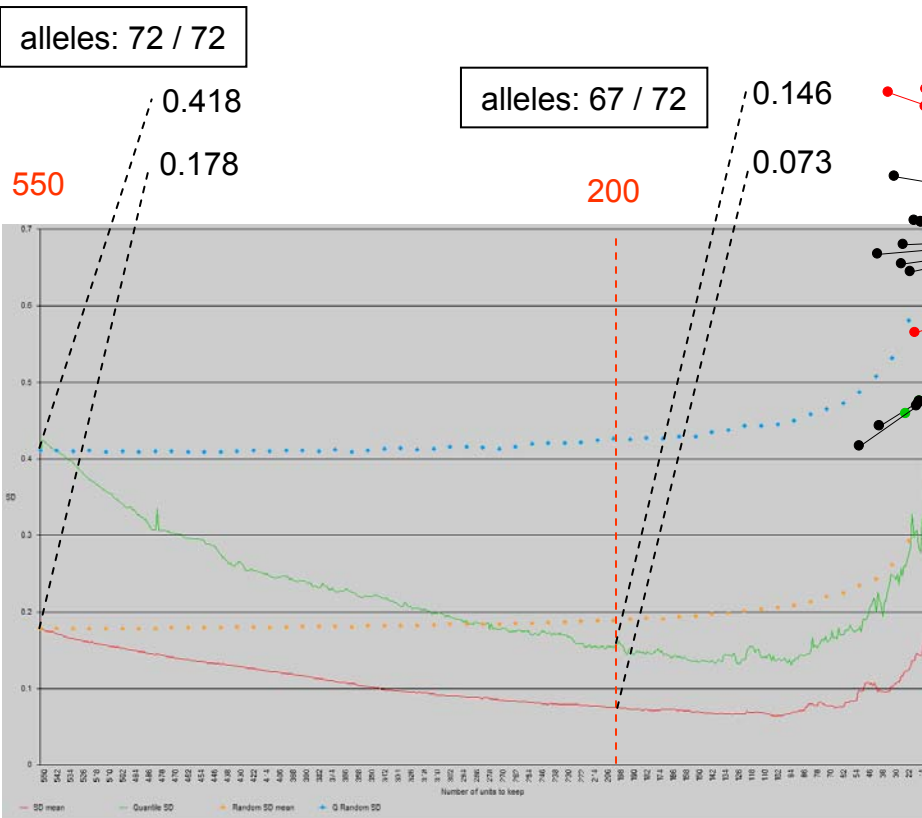
in red the 200 sampled accessions

## max subtree versus min SD strategies

660 Sorghum accessions / 20 SSR

two steps procedure:

- max sub tree: sample 550 / 660
- min SD: sample 200 / 550



in red the 200 sampled accessions  
in green: excluded accessions

## Further developments within the 2005 project

Project in progress!

### *Methods*

- Tree construction
  - # bootstrap for dissimilarities on linked markers
  - # standardization for weighted sum of partial dissimilarities
- Measure of disequilibrium measure when the phase is unknown
  - # estimation of haplotype frequencies from the data
    - Free software implement the Expectation-Maximization algorithm (Hill, 1974)
    - Arlequin, Haplo...
    - data file exchange function

## Further developments within the 2005 project

### *Validation*

- compare samples defined from SSR and RFLP on Sorghum data sampling on SSR and test on RFLP disequilibrium and conversely
- test these samples on LD reduction for closely linked markers (4 cM)

### *Tools*

management of missing data

optimise algorithms

direct import of files at GCP format

validate documentation and user's interface

distribute the software through GCP network

## Proposal for commissioned 2006 projects

.....but there are still some pendent methodological questions for sampling and association analysis.

### ■ many SP1 projects use SSR markers

Problem of hypervariability of SSR markers (cf poster)

# allele pooling:

- for each locus, aggregation on statistical kernels
- minimizing the loss of mutual information between pair of loci
- minimizing tree structure perturbations

# fuzzy code?

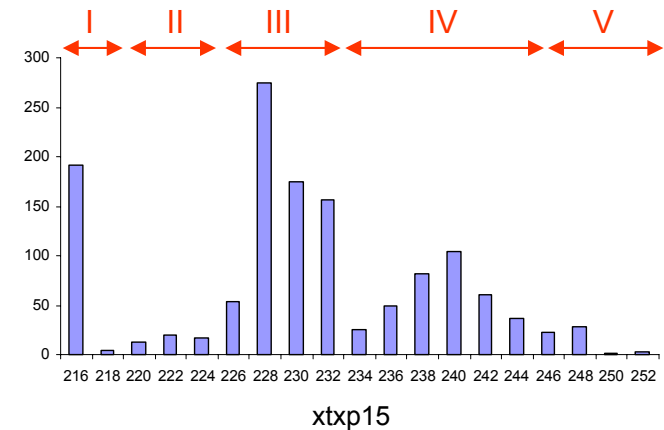
### ■ haplotype frequency estimations when phase is unknown

HWE assumption for EM algorithms does not hold

coupling disequilibrium estimations and haplotype frequency estimations

Bayesian approach?

### ■ polyploidy



## Helpdesk proposal for 2006

## 2006 project

### Chennai Workshop SP1 “Molecular Markers for Allele Mining”, Aug 22-26 2005

Software for efficient and effective sampling of germplasm

**DARwinN**

SP1 scientists desire support in the sampling of germplasm and statistical analysis especially in association analysis.

New project aimed at supporting the SP1 scientists:

**a helpdesk**

## Proposal for 2006 (P.I. Marco Bink WUR)

### Helpdesk to support SP1 projects

- in sampling of germplasm, using the tools developed in the current project
- but also in association mapping analyses that should follow the sampling step
- a contact person at WUR that
  - either helps the SP1-researcher directly
  - or directs him to an appropriate other personfor short consultancies on data analysis strategies
- help on tools and software
- improvement or customization of tools...
- is responsible for proper feedback
- a website: manuals, examples, links to software, exchanges between SP1- researchers...

Involved institutions : WUR (+ CIRAD and other resource persons)

### HOWEVER

- Helpdesk / websites already present at other SP's
- Association mapping courses also organised by CIMMYT (Nov '05, Nairobi)

### ALSO

- Consultancy/support to SP1 scientists -> Doing the analysis for SP1 .....
- Proper Linkage Disequilibrium mapping software still lacking