

**Merging and integrating
genotyping data**

CIMMYT

INRA

1st genetic data set

2nd genetic data set

275 American/European
populations

496 African/Asian/Oceanian...
populations

DNA bulks of 15 individuals



19 same loci



No obvious correspondance between data sets

How to merge these two datasets?

➤ Technical differences between CIMMYT and INRA
= labs equipment/protocol

- Electrophoresis:

INRA
DNA sequencer Li-Cor

CIMMYT
DNA sequencer ABI Prism 377

➤ **Technical differences between CIMMYT and INRA**
= labs equipment/protocol

- Electrophoresis:

INRA
DNA sequencer Li-Cor

CIMMYT
DNA sequencer ABI Prism 377

- Allele scoring method:

INRA
One-Dscan v. 2.05

CIMMYT
Genescan v. 3.0 + Genotyper v. 2.1

- visualisation of bands
- manual scoring of bands
- manual binning

- visualisation of peaks
- automatic scoring of peaks
- automatic binning

➤ **Consequences of technical differences between CIMMYT and INRA**

➔ **Between-labs inconsistency in allele size evaluation/naming** despite maximal resolution (allele distinction) and sensitivity (allele detection)

➔ **Between-labs discrepancy in number of alleles detected:**

- **allele dropout:** rare allele is present but is not detected in one lab. vs. the other, due to lower detection sensitivity
- **resolution:** allele is present but is not differentiated from one neighboring allele in one of the two labs

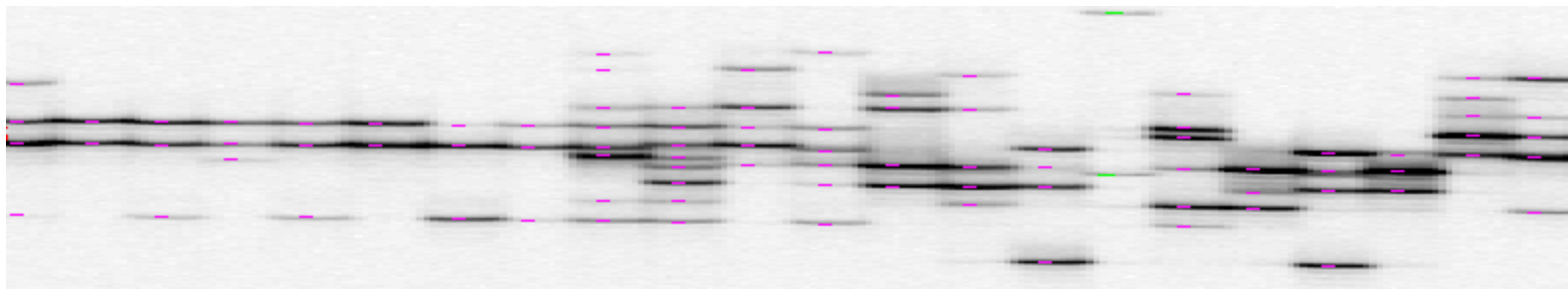
➤ Differences between the datasets = genetic differences between populations sets analysed at CIMMYT and INRA

Set 1 =
American/European pops.

≠

Set 2 =
African/Asian/Oceanian pops.

➔ True variation in polymorphism: **specific alleles** only present in one of the 2 populations sets

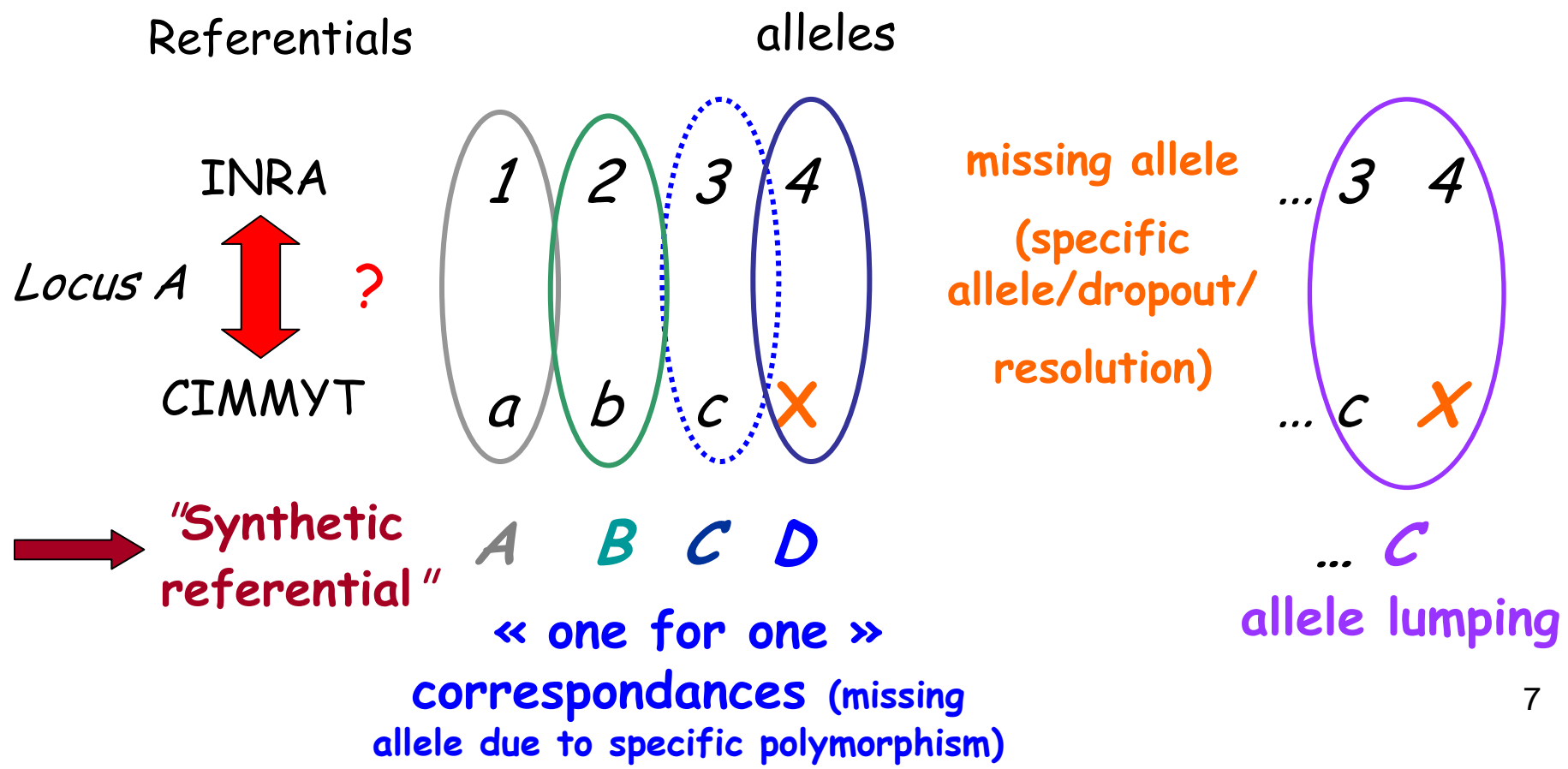


Case of teosintes

Objective:

Establishment of correspondances between the **allele referentials** of each lab. (=list of alleles)

Rk. Assuming no residual artefactual alleles due to stuttering...



Establishment of correspondances

- Pilot study realised at INRA for each locus:

Comparisons of

- 12 populations previously genotyped at CIMMYT
- 6 inbred lines

} representative
of allele
diversity

➡ *visual direct identification of CIMMYT alleles in our experimental conditions*

= first correspondance established between CIMMYT and INRA referentials

- **But... once both populations sets analysed**

Many new alleles ➡ *correspondances were questioned*

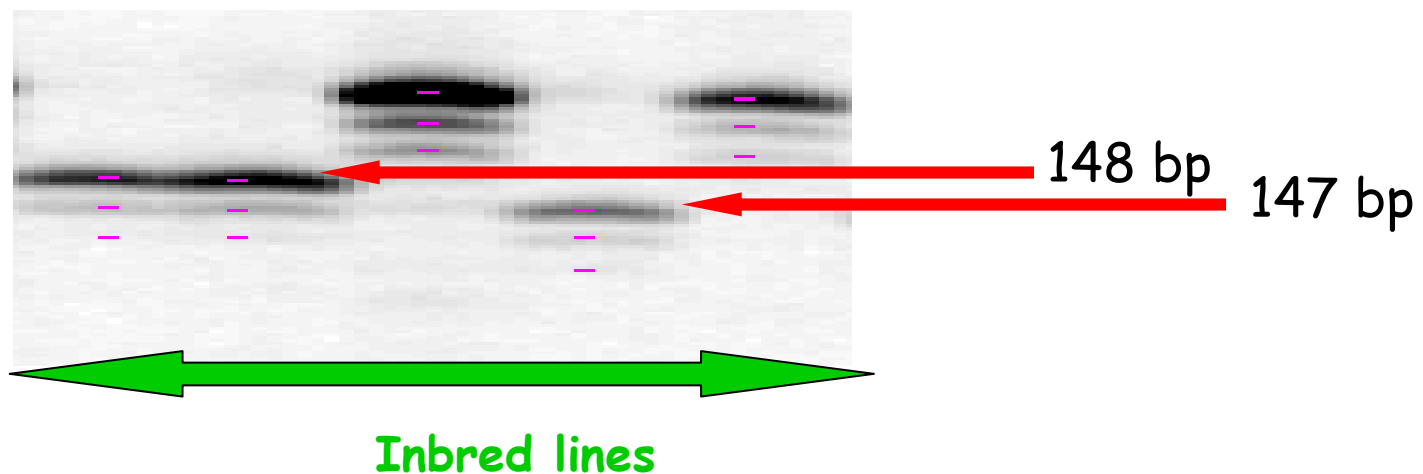
-> examples of situations...

➤ Example 1

	CIMMYT	INRA	
<i>phi029</i>	1 st set Amer./Eur. populations	Amer./Eur. populations (control samples)	2 nd set Afr./Asian/ Oceanian populations

alleles

147	147	147	147
0	148	148	148



➤ Example 1

	CIMMYT	INRA		
<i>phi029</i>	1 st set Amer./Eur. populations	Amer./Eur. populations (control samples)	inbred lines (control samples)	2 nd set Afr./Asian/ Oceanian populations
alleles	147 0	147 148	147 148	147 148

-Between-labs discrepancy for Amer./Eur. populations:

- unlikely rare allele dropout since high frequency of allele 148
- *a priori* due to between-labs difference in resolution

➔ *Lumping allele 148 with neighboring allele 147 common to both labs. necessary to combine datasets*

➤ Example 2

CIMMYT	INRA		
1 st set Amer./Eur. populations	Amer./Eur. populations (control samples)	inbred lines (control samples)	2 nd set Afr./Asian/ Oceanian populations

phi029 *rare « 144 »* *0* *no direct correspondance*
...alleles 139 and 143 ?

-Between-labs discrepancy for Amer./Eur. populations:

- possible true absence of rare allele « 144 » in Amer./Eur. pop. subsample used at INRA (control samples)

➔ *to be checked by analysing pops with highest frequency of allele « 144 » at INRA, with populations displaying alleles 139 and 143*

(a) correspondance between « 144 » with 139 or 143

(b) « 144 » confirmed as true different allele

*(c) if « 144 » does not appear in our conditions:
its elimination is necessary to combine datasets*

➤ Example 3

CIMMYT	INRA		
1 st set Amer./Eur. populations	Amer./Eur. populations (control samples)	inbred lines (control samples)	2 nd set Afr./Asian/ Oceanian populations

phi029

0

0

0

165

-Discrepancy between Amer./Eur. and Afr./As./Oce populations:

- likely missing allele in Amer./Eur. populations due to genetic differences between the 2 populations sets

➡ *strategy depends on proximity in size with other alleles*

➤ Example 3

-if 165 differs by several bp from other alleles:

➡ *systematic identification possible in each lab.*

➡ *165 considered as new true allele for data merging*

-if 165 is very close to other allele (1 bp difference):

➡ *between-lab. confusion in 165 identification is possible*

➡ *lumping 165 with its neighbour limits between-lab. genotyping errors*

Rk. Analysis of pops with 165 in CIMMYT experimental conditions could help to decide how 165 should be considered...

➤ **Alternative methods for merging data?**

- Use of adapted software: Micromerge (Presson et al., 2006)?

Problems:

Systematic allele lumping doesn't seem adapted to genetically differentiated populations...

 *Other softwares?*

➤ Example of correspondances and synthetic referential (ongoing)

Rk. The size of **new alleles** has to be checked before deciding of lumping alleles or establish independant correspondances

Example: phi014

Allele naming

CIMMYT	0	420	423	426	0	429	0	432	0	0	435	0	0	438	441	0	0	0
INRA	142	148	0	0	157	158	159	160	161	162	163	164	165	166	169	172	175	178
Synthetic	?	A	?	?	?	B	?	C	?	?	D	?	?	E	F	?	?	?

↑
Referentials

↑ ↑
established correspondances

➤ How to manage correspondances?

- Allele lumping is necessary when merging CIMMYT and INRA data sets

But

- No necessity of allele lumping within each data set (independant data analysis)

 *Database developed at INRA:*

- *based on each lab referential + synthetic referential*
- *export populations genotypes encoded with CIMMYT, INRA or synthetic referential*

- Data View Home
- Project
- Crossview
- Genotyping
- Genotypic data**
- Locus
- Locus Map
- Referential
- Correspondance
- Phenotyping
- Classification
- Entity(ies)/Genealogy

Genotypic data

View of all data from genotyping

View of all data from genotyping

select locus

Entity(ies)

population Ames2749
population Ames2750
population Ames2751
population Ames2752
population Ames2755
population ANC393
population ANTI3
population ANTI392
population ANTIGP1
population ANTIGP2
population APUC140
population APUC171
population ARGE486
population ARGE564
population ARGEGP8

select populations



Loci(us)

SSR phi059
SSR phi062
SSR phi069
SSR phi072
SSR phi083

select referential to export genotypes

Experiment(s)

SNP Monsanto DivCor (Sehabiague at Monsanto in 2006)
SNP-IDP_S1P9_GQMS (Madur at INRA in 12-2005)
SNP_IDP_GQMS (Madur at UMR Genetique vegetale du moulon in 2006_10_27)
SSR Diversite cornee 2 GQMS (Madur at INRA in 2005)
SSR diversite comes2 (Dubreuil at CIMMYT in 2000-2001)

Referential(s)

CCaOMT2 (Fourmann)
Cypdk alignment (Fourmann)
Dwarf8 (Fourmann)
GS1-3 alignment (Fourmann)
RFLP referential (Rebourg)
SSR Cimmyt referential (Dubreuil)
SSR Moulon referential 1 (Madur)
SU1 (Camus)
U19 (Fourmann)
VGT1 (Ducrocq)

Missing Data
View considering
correspondence
View considering
Seed Lot

missing

Show data Reset Form

Merging and integrating genotyping data

➤ Discussion

- Analysis of common samples by each lab (!)
- Proper binning is essential especially for neighboring alleles distinction
- Analysis of pops. displaying non common alleles by the other lab.
- Possibility to return to raw data (before eventual intra-lab allele lumping)
- Between-labs genotyping error rate?
- Software use?

....

Genotyping Database Administration - Mozilla Firefox

http://association.moulon.inra.fr/grdb/index_.jsp

GENOTYPING DATABASE ADMINISTRATION

GO TO ADMINISTRATION LOGOUT MY HOME HELP

Genotyping Data Management

Data View

Locus Management

Experiment

Sample

Referential

Correspondence Management

Correspondence

Link alleles

Locus Map Management

Link alleles

Choose a correspondence class

Correspondence class : phi062 (B)

Assign alleles to correspondence class phi062 (B)
Select alleles you want to link to the same correspondence class

Referential : SSR Moulon referential 1 (Madur 2002-01-01)

Show alleles Reset Form

Assign alleles to correspondence class phi062 (B)
Select alleles you want to link to the same correspondence class

158	<input type="checkbox"/>
161	<input type="checkbox"/>
167	<input type="checkbox"/>
155	<input type="checkbox"/>
150	<input type="checkbox"/>
164	<input type="checkbox"/>

Link alleles Reset Form

View of the records :

All correspondence classes and alleles for locus phi062

Sort by Criteria

Export This Table Current Page : 1 of 1 (6)

Correspondence	Referential	Alleles
A	SSR Moulon referential 1(Madur, 2002-01-01)	155
A	SSR Cimmyt referential(Dubreuil, 2002-08-01)	158
B	SSR Moulon referential 1(Madur, 2002-01-01)	158
B	SSR Cimmyt referential(Dubreuil, 2002-08-01)	161
C	SSR Moulon referential 1(Madur, 2002-01-01)	161
C	SSR Cimmyt referential(Dubreuil, 2002-08-01)	164

Export This Table Current Page : 1 of 1 (6)

select synthetic referential for a locus

list of alleles in INRA referential

CIMMYT: 158

INRA: 155

synthetic referential