E

BMT/15/10
ORIGINAL: English
DATE: May 17, 2016

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

Geneva

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA PROFILING IN PARTICULAR

Fifteenth Session

Moscow, Russian Federation, May 24 to 27, 2016

MOLECULAR DATA ANALYSIS CAPACITY

Document prepared by experts from France

Disclaimer: this document does not represent UPOV policies or guidance

The Annex to this document contains a copy of a presentation “Molecular Data analysis capacity” to be made at its fifteenth session of the Working Group on Biochemical and Molecular Techniques and DNA-Profiling in particular (BMT).

Muriel Thomasset, Anne Bernole, Arnaud Remay, Clarisse Maton, René Mathis
GEVES, France

[Annex follows]

Molecular Data analysis capacity

UPOV – BMT/15 – May 2016, Moscow
Muriel Thomasset, Anne Bernole, Arnaud Remay,
Clarisse Maton, René Mathis
GEVES, France

GEVES
Expertise & Performance

Data processing

Sample preparation DNA → Mix dispensing PCR → Amplification → Revelation → Raw data → Data processing

- According to the number of markers and/or application:

  ➡ Ordinary data processing (filters/ pairwise comparison between genetic profiles)

  - Example: true-to-type analysis for a seed lot

  ➡ Automated data processing (scripts of calculation/ genetic distance calculation)

  - Example: Genetic distance calculation used in combination with morphological distances in order to help side by side comparison in DUS trials.

Genetic distance calculation between 2 varieties

Variety	Locus 1	Locus 2	Locus 3
Cultivar 1	10/10	20/30	60/60
Cultivar 2	40/40	30/30	50/80
Cultivar 3	20/20	?/?	60/60

- **Step 1 :** genotyping of each variety (SSR or SNP)
  - ➡ Raw data

Variety	Locus 1			Locus 2		Locus 3		
	10	20	40	30	50	50	60	60
Cultivar 1	1	0	0	0.5	0.5	0	1	0
Cultivar 2	0	0	1	0	1	0.5	0	0.5
Cultivar 3	0	1	0	0M	0M	0	1	0
...								

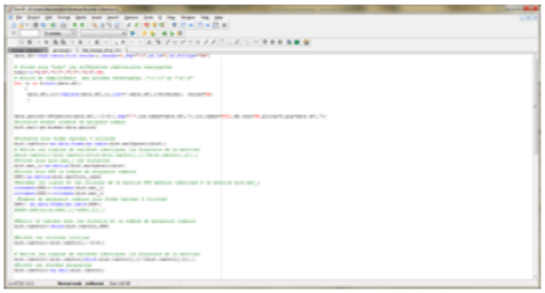
- **Step 2 :** Transformation of genetic data into « mathematics » data
  - ➡ Allele coding in 0, 1 and 0.5

	Cultivar 1	Cultivar 2	Cultivar 3	...
Cultivar 1	0			
Cultivar 2	0.83	0		
Cultivar 3	0.5	1	0	
...	...	...	...	0

- **Step 3** : Genetic distance calculation
  - ➡ Sum of differences for all the markers

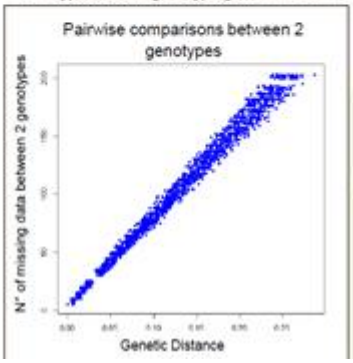
Genetic distance calculation between two varieties

- Different formulae available for genetic distance calculation: Rogers, Rogers modified, Nei , Dice, Jaccard...
- Genetic distance choice: marker type (dominants/co-dominants)
  species (diploïd/polyploïd)
- Calculation automatisation : Use of R software (R Core Team) and script development



Which Genetic distance, which packages?

- Which Genetic distance, which software ?
  => Rogers genetic distance (1972) / Software R /package Adegenet
- How to handle missing data?
  - Simulations:
    - Analysis on simulated data=>calculation of the genetic distance between 2 genotypes with only missing data between each other
    - Hypothesis: if genotyping twice the same line => genetic distance =0



Missing data are considered during the calculation

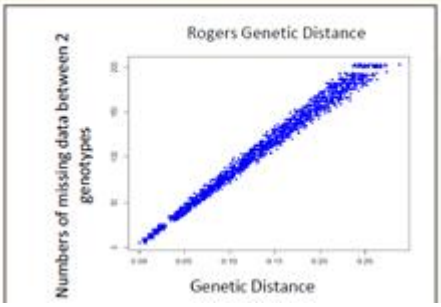
Need to use another package or distance

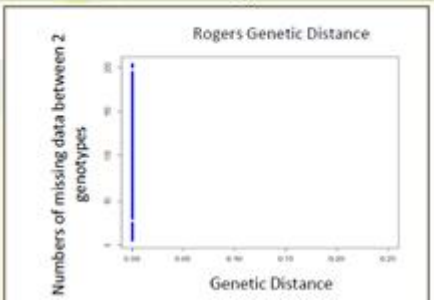
Groupe d'Étude et de contrôle des Variétés Et des Semences

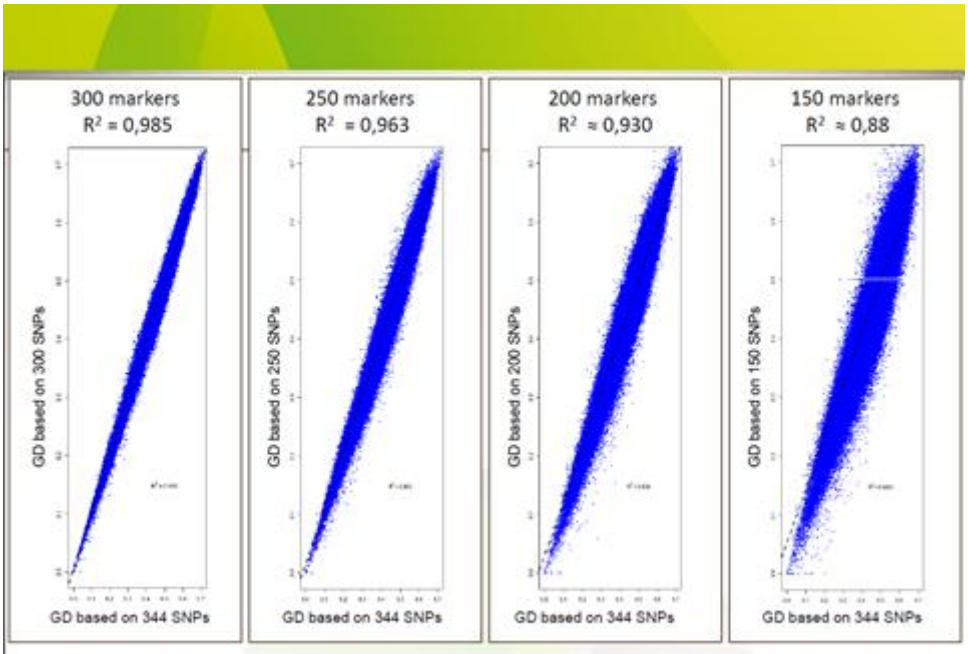
Genetic distance Kosman and Leonard (2005)

- Rogers (1972)
  - Based on allelic frequency
  - $D_R(a,b)= \frac{1}{v}\sum_{k=1}^{v}\sqrt{\frac{1}{2}\sum_{j=1}^{m=k}(p_{aj}^{k}-p_{bj}^{k})^2}$
  - However if a and b are individuals and not population => frequency= occurence

- Kosman and Leonard (2005)
  - Based on proportion of shared alleles
  - $D_R(a,b)= \frac{1}{v}\sum Ass_{mark}^{p}(a,b)$
  - PopGenReport Package







Number of markers used for calculation of the distance will greatly influence the results

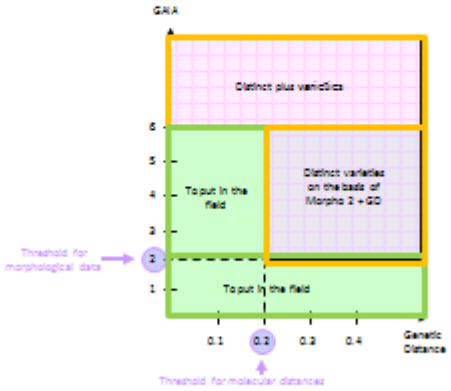
The various routine applications in the laboratory

- Help in pairwise comparisons to be grown side by side in DUS trials = management of reference collection
- Variety identification
- Checking hybrid conformity
- Description and characterisation of reference collections

The various routine applications in the laboratory

- Help in pairwise comparisons to be grown side by side in DUS trials
- Variety identification
- Checking hybrid conformity
- Description and characterisation of reference collections

Help in setting DUS growing trials



- 2 steps:

1. Morphological Descriptions Comparison Comparaison (GAIA scored from 1 to 9) :

If GAIA ≥ 6 — Distinct plus varieties

If GAIA < 2 — To put in the field

2. For all couples with a GAIA distance between to 2 and 6, Genetic Distance is used

If DG < seuil — To put in the field

If DG ≥ seuil — Distinct varieties

Help in setting DUS growing trials

- **Maize DUS trials 2014**
  - About 361 candidate varieties for registration
  - Reference collection : 4445 lines

  → **> 1.5 million** comparisons

- Morphological data (GAIA only)
  → **6123** comparisons to be grown

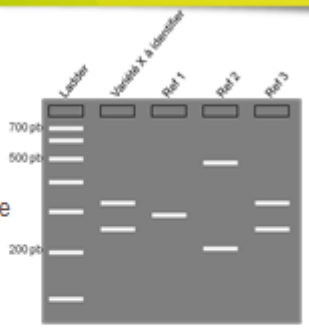
- Morphological data (GAIA only) + Genetic Distance
  → **566** comparisons to be grown

Reduction of number of comparisons to be grown **> 90 %**

The various routine applications in the laboratory

- Help in pairwise comparisons to be grown side by side in DUS trials
- Variety identification
- Checking hybrid conformity
- Description and characterisation of reference collections

Variety Identification



- Aim :
  - Each profile observed is compared to those of the reference collection
  - Identification of the identical or closest varieties
- Calculation automatisation => R script development:
  - Genetic Distance estimation between the variety to be identified and those of the reference collection.
- Selection of identical or closest varieties (DG=0 or DG < threshold)
- For each marker and each selected couple => Fine analysis of differences between each marker of molecular profiles.

Variety Identification

- Example : identification into a mix of milling wheat varieties
  - 990 varieties in the reference collection
  - Analysis on 40 individual seeds=> n profiles
  - Computation time < 2 min
- Example of output

Sample	Variety	GD	Nb_loci_used	SSR 1	SSR 2	SSR 3	SSR 4	SSR...
Profile 1	Var A	0	10	1	1	1	1	1
Profile 2	Var B	0.1	10	HomoZ_diff	1	1	1	1
Profile 3	Var C	0.05	10	Half diff with a common allele	1	1	1	1

The various routine applications in the laboratory

- Help in pairwise comparisons to be grown side by side in DUS trials
- Variety identification
- Checking hybrid conformity
- Description and characterisation of reference collections

Checking hybrid conformity

- Aim : comparison between an observed hybrid and the expected one

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	....
Parent A	A/A	A/A	B/B	B/B	A/A	A/A	A/A	A/A	A/A	....
Parent B	A/A	B/B	A/A	B/B	NA/NA	B/B	A/A	B/B	A/A	....
Expected Hybrid	A/A	A/B	A/B	B/B	-	A/B	A/A	A/B	A/A	....
Observed Hybrid	A/A	A/B	A/A	B/B	A/B	A/B	B/B	A/B	A/A	....

- Hybrid conformity:
  - Conform =>Number of different markers < threshold
  - Not conform=> Number of different markers > threshold
- GD estimation automatisation => R Script :
  - GD calculation between observed and expected hybrid
  - Conform =>GD < threshold
  - Not conform =>GD > threshold

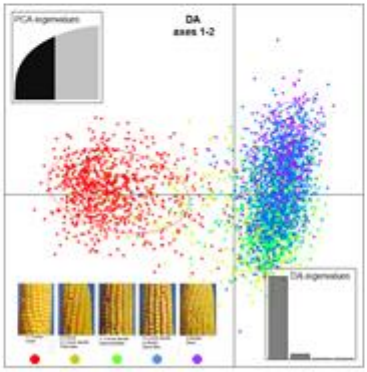
The various routine applications in the laboratory

- Help in pairwise comparisons to be grown side by side in DUS trials
- Variety identification
- Checking hybrid conformity
- Description and characterisation of reference collections

Description and characterisation of reference collections

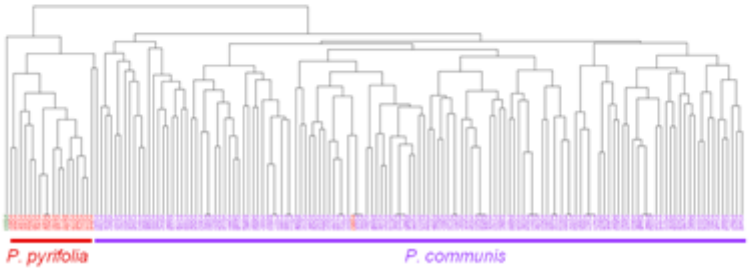
- Structure of genetic diversity

Principal component discriminant analysis



Description and characterisation of reference collections

- Structure of genetic diversity
- Determine relationships between varieties of a collection
- Variety pedigree /coefficient of relatedness
- Search for duplicated or synonymous varieties



GEVES
Expertise & Performance

[End of Annex and of document]