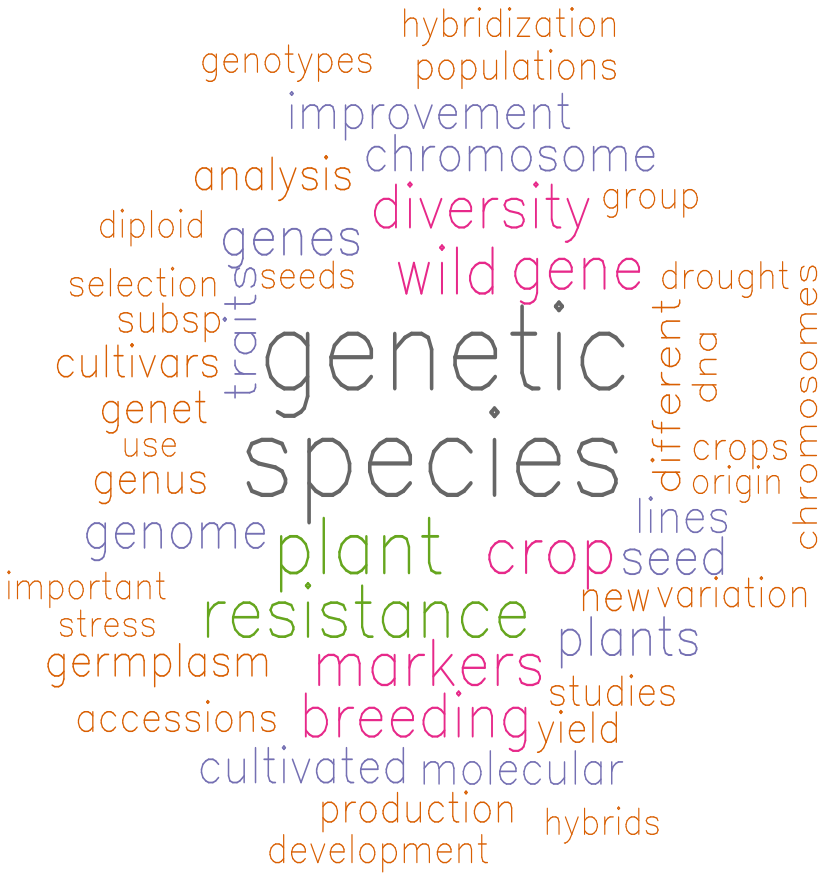
Theory on Genetic Diversity Analysis

Fernando Henrique RB Toledo

August 17^{th}, 2017

<f.toledo@cgiar.org>



Outline:

Some definitions
- Population Genetics
- Genetic Markers

Genetic Distances

Intraspecific Diversity
- Anova & Complete Randomized Design
- Wright Statistics

Find Clusters and Groups

The software BIO-R
- Getting BIO-R
  - Check Output and Estimates

Some Definitions ...

- **Genotype:**

  Genotype is ... _____

  _____

  _____

  _____.

- **Markers:**

  Genetic Markers are ... _____

  _____

  _____

  _____.

- **Population:** ... [CONT]

Population Genetics ...

- **Population Genetics:** Studies the heredity's mechanisms at the population level.
- **Population:** Set of conspecifics individuals in the *same place and time*, which have the ability to mate (exchange alleles).

In a population, the individual has a transitory importance, what matters are the alleles it has, which will be transmitted to subsequent generations

Reproductive systems

- **Allogamous:** frequency of cross polination is ≥ 95% *e.g.*, maize
- **Autogamous:** frequency of cross pollination is ≤ 5% *e.g.*, wheat.

Allele and Genotypic Frequencies ...

Genotype		Phenotype	Observed	Frequency
BB	$\longrightarrow$	White	100	0.05
Bb	$\longrightarrow$	Yellow	1000	0.50
bb	$\longrightarrow$	Red	900	0.45
		Total:	2000	1.00

$$f(B) = p = 0.05 + \frac{0.50}{2} = 0.30$$

$$\begin{aligned} f(b) &= q = 0.45 + \frac{0.50}{2} = 0.70 \\ &= (1 - p) \end{aligned}$$

Onion is a autogamous specie ...

Any plant can mate with any other one i.e., panmixia & HWE

		B	b
		p	q
B	p	p^2	$p \cdot q$
b	q	$p \cdot q$	q^2

BB	p^2	$0,30^3 = 0,09$
Bb	$2 \cdot p \cdot q$	$2 \cdot 0,30 \cdot 0,70 = 0,42$
bb	q^2	$0,70^2 = 0,49$

If not, considering selfcross ...

The plants mate by themselves

Genotype	G_0	G_1	
BB	$p^2 = f_{BB_0}$	$f_{BB_0} + 1/4 \cdot f_{Bb_0}$	$p^2 + 1/2 \cdot p \cdot q$
Bb	$2 \cdot p \cdot q = f_{Bb_0}$	$f_{Bb_0} - 1/2 \cdot f_{Bb_0}$	$p \cdot q$
bb	$q^2 = f_{bb_0}$	$f_{bb_0} + 1/4 \cdot f_{Bb_0}$	$q^2 + 1/2 \cdot p \cdot q$

$$
\begin{aligned}
p' &= p^2 + 1/2 \cdot p \cdot q + 1/2 \cdot p \cdot q \\
&= p^2 + 1/2 \cdot p \cdot (1 - p) + 1/2 \cdot p \cdot (1 - p) \\
&= p^2 + 1/2 \cdot (p^2 - p) + 1/2 \cdot (p^2 - p) \\
&= p^2 + 1/2 \cdot p^2 - 1/2 \cdot p + 1/2 \cdot p^2 - 1/2 \cdot p \\
&= p^2 - p^2 + p \\
p' &= p
\end{aligned}
$$

1. **Alogamous**

$$\begin{aligned} f_{BB} &:= p^2 \\ f_{Bb} &:= 2 \cdot p \cdot q \\ f_{bb} &:= q^2 \end{aligned}$$

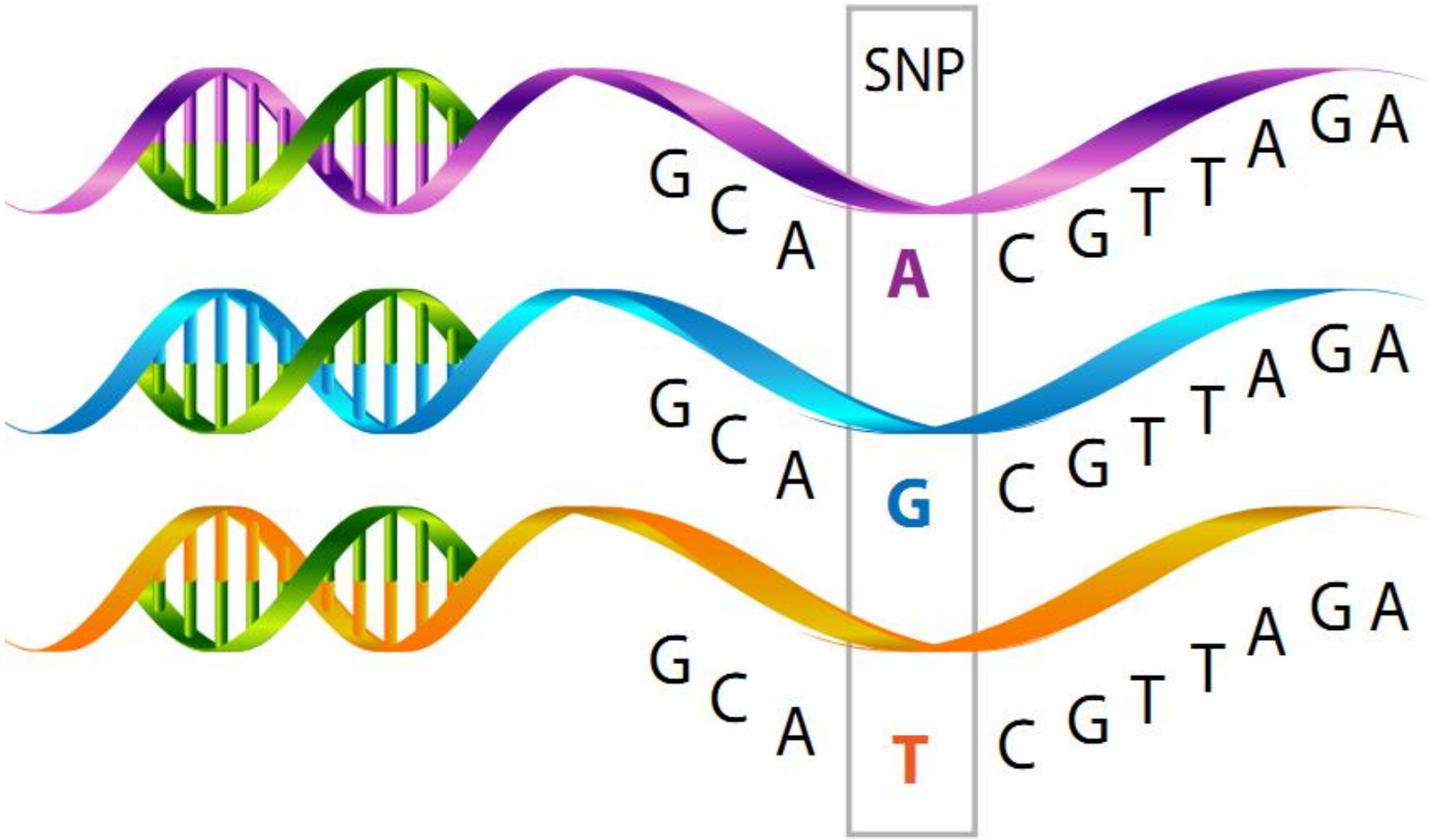
2. **Autogamous**

$$\begin{aligned} f_{BB} &:= p^2 + 1/2 \cdot p \cdot q \\ f_{Bb} &:= 2 \cdot p \cdot q - 1/2^* \cdot 2 \cdot p \cdot q \\ f_{bb} &:= q^2 + 1/2 \cdot p \cdot q \end{aligned}$$

3. **Mixed type**...taking $^*1/2 = I$ or Wright's Equilibrium

$$\begin{aligned} f_{BB} &:= p^2 + I \cdot p \cdot q \\ f_{Bb} &:= 2 \cdot p \cdot q - I \cdot 2 \cdot p \cdot q \\ f_{bb} &:= q^2 + I \cdot p \cdot q \end{aligned}$$

Genetic Markers ...



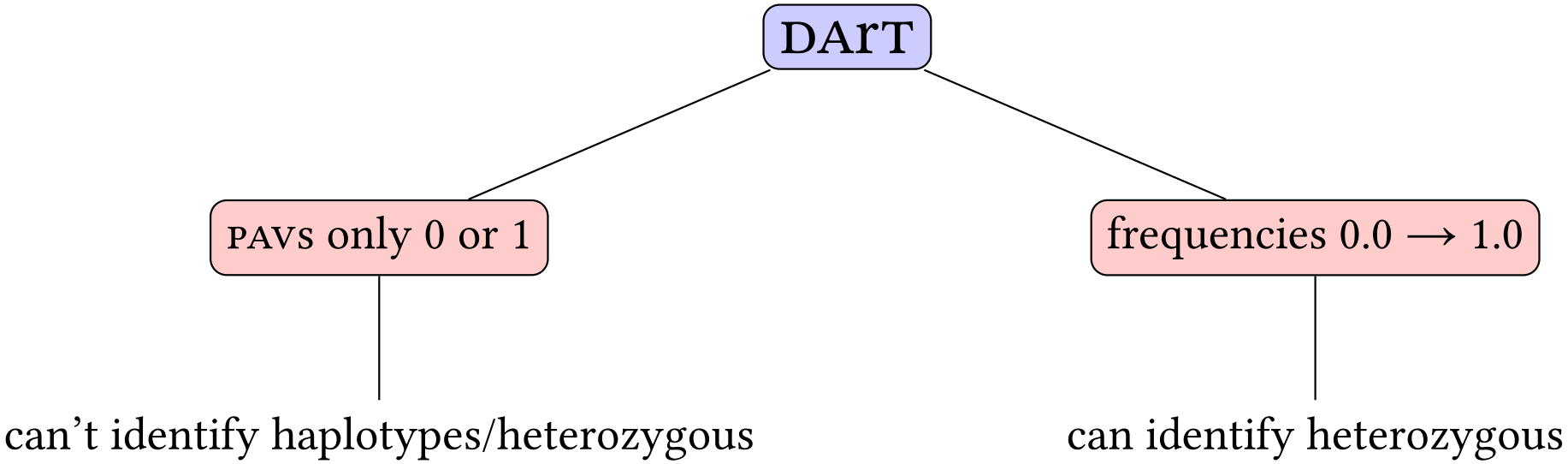
Examples

Marker	Amount	Expression	Polymorphic degree	Specific by locus
Isoenzyme	< 100	codominant	low	✓
RFLP	∞	codominant	medium	✓
RAPD	∞	dominant	medium	–
SSR	∞	codominant	very high	✓
SCAR	∞*	both	low	✓
AFLP	∞	dominant	high	–
SNP	∞	codominant	very high	✓
DArT	∞	both	very high	✓**

* After the deployment of other kind of markers.

** Clone aren't but markers are.

DArT flexibility...



Maize Bulks w/ frequencies

Marker	Allele	Bulks					
		1	2	3	4	⋯	g
1	1	0.67	*NA**	0.57	*NA*	⋯	0.36
1	2	0.33	*NA*	0.43	*NA*	⋯	0.64
2	1	*NA*	1.00	1.00	1.00	⋯	1.00
2	2	*NA*	0.00	0.00	0.00	⋯	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
m	1	0.00	*NA*	1.00	0.00	⋯	1.00
m	2	1.00	*NA*	0.00	1.00	⋯	0.00

* *NA* means *Not Available* or missing.

Frequencies, so:

$$p + q = 0.67 + 0.33 = 1.00$$

Maize Genotypes w/ *PAV* – (presence/absence)

Marker	Allele	Bulks					
		1	2	3	4	⋯	g
1	1	1	*NA*	1	*NA*	⋯	1
2	1	*NA*	1	1	1	⋯	1
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
m	1	1	*NA*	1	1	⋯	1

* *NA* means *Not Available* or missing.

Allele is always 1…the reference allele.

When Genotypes belong to the same bulk, the summary of their *PAV* generates frequencies

Genetic Distances ...

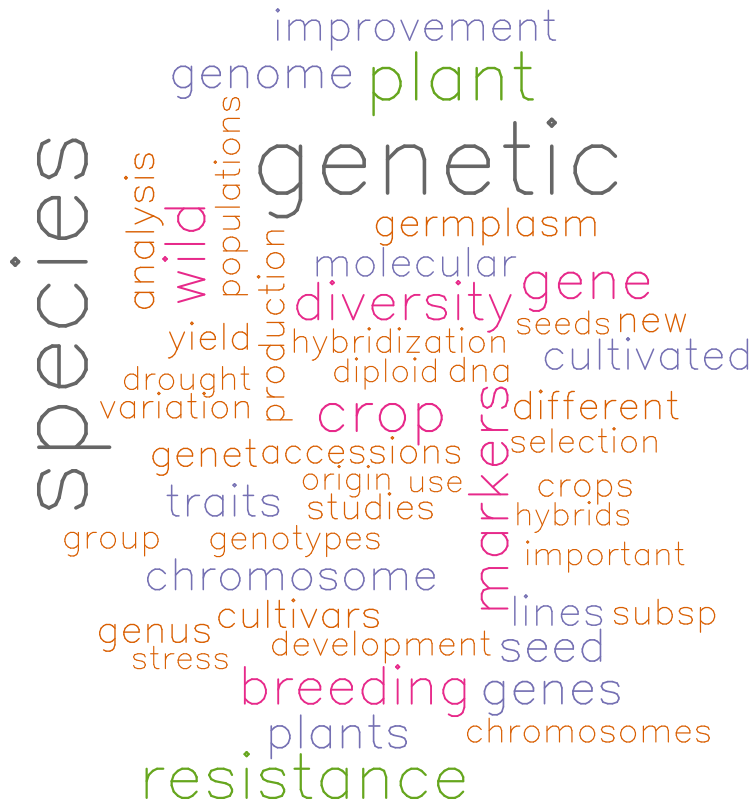
Genetic Distances

Markers w/ allelic information

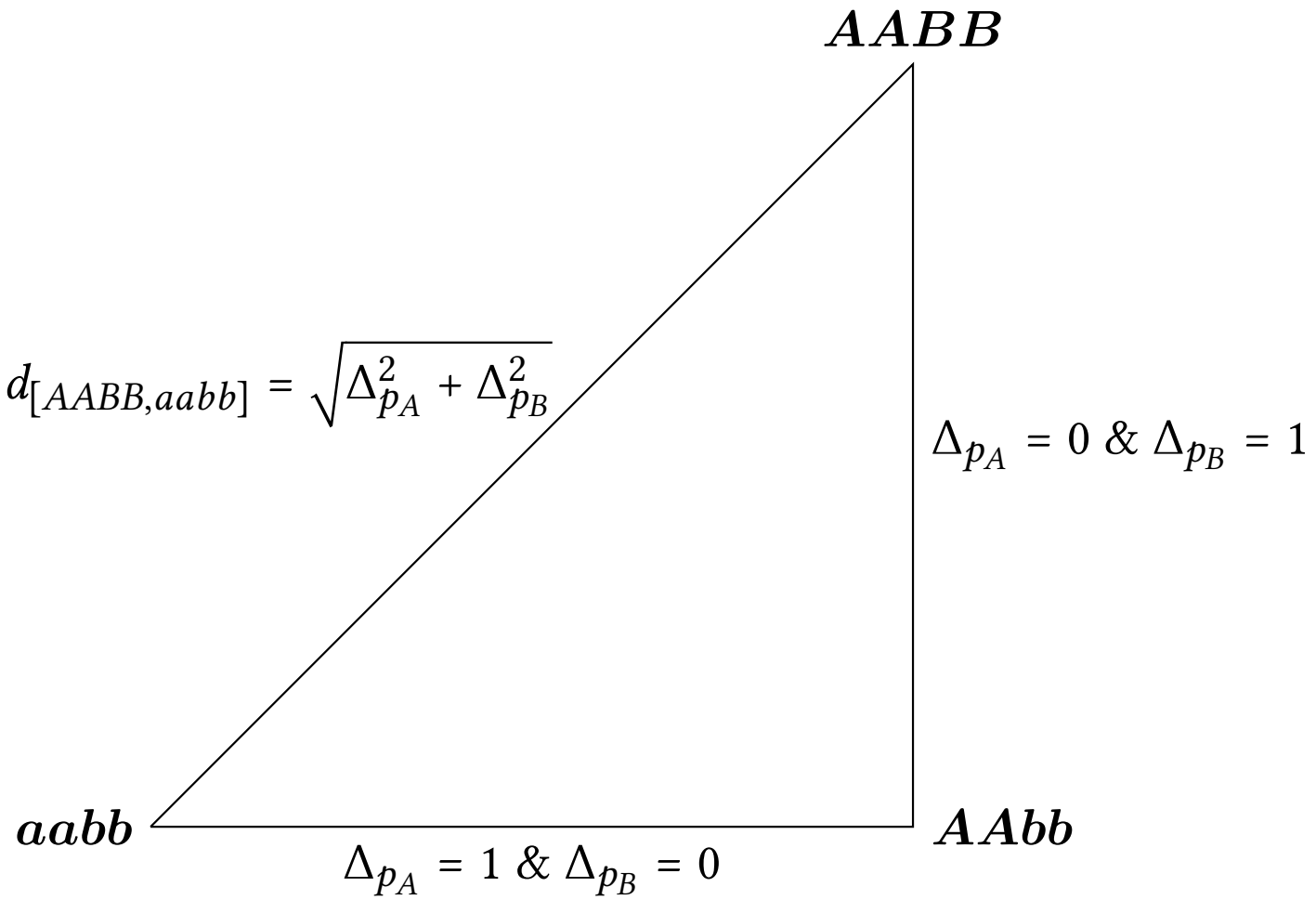
For example: SSR, SNP and DArT.

- **Euclidean**;
- **Modified Rogers**; and
- **Cavalli-Sforza & Edwards**.

References: [1, 2, 3]



Euclidean Distance ...



So, . . . :

$$Ed_{[x,y]} = \sqrt{\sum_{l}^{L}\sum_{a}^{A}(\hat{p}_{x\,la} - \hat{p}_{y\,la})^2} \;, \left(0.0 \leq Ed_{[x,y]} \leq \sqrt{2L}^{*}\right)$$

where:
$\hat{p}_{x\,la}$ is the allele frequency for allele a at *locus* l in the genotypes x;
$\hat{p}_{y\,la}$ is the same as above for genotype y;
L is the number of *locus*; and
A is the number of alleles at *locus* l.

*This is the estimator and its respective domain as shown by [3].

Roger's Distance ...

$$Rd_{[x,y]} = \frac{1}{L}\sum_{l}^{L}\sqrt{\frac{1}{2}\sum_{a}^{A}(\hat{p}_{x\,la} - \hat{p}_{y\,la})^2}\ , \left(0.0 \leq Rd_{[x,y]} \leq 1.0\right)$$

where:
$\hat{p}_{x\,la}$ is the allele frequency for allele a at *locus* l in the genotypes x;
$\hat{p}_{y\,la}$ is the same as above for genotype y;
L is the number of *locus*; and
A is the number of alleles at *locus* l.

It hasn't relationship with Euclidean Geometry. So, ...

Modified Roger's Distance ...

$$MRd_{[x,y]} = \frac{1}{\sqrt{2L}}\sqrt{\sum_{l}^{L}\sum_{a}^{A}(\hat{p}_{x\,la} - \hat{p}_{y\,la})^2}\ , \left(0.0 \leq MRd_{[x,y]} \leq 1.0\right)$$

where:
$\hat{p}_{x\,la}$ is the allele frequency for allele a at *locus* l in the genotypes x;
$\hat{p}_{y\,la}$ is the same as above for genotype y;
L is the number of *locus*; and
A is the number of alleles at *locus* l.

For non-diploid species, replace $2L$ by the number of copies.

Cavalli-Sforza & Edwards ...

$$Cd_{[x,y]} = \sqrt{\frac{1}{L}\sum_{l}^{L}\left(1 - \sum_{a}^{A}\sqrt{\hat{p}_{x\,la} \times \hat{p}_{y\,la}}\right)}\ , \left(0.0 \leq Cd_{[x,y]} \leq 1.0\right)$$

where:
$\hat{p}_{x\,la}$ is the allele frequency for allele a at *locus* l in the genotypes x;
$\hat{p}_{y\,la}$ is the same as above for genotype y;
L is the number of *locus*; and
A is the number of alleles at *locus* l.

Means, Variances and Standard Errors ...

$d_{[x,y]}$ is the distance between genotypes x and y, thus:

- **Mean:** $\mu_{d_{[x,y]}} = \frac{1}{\frac{A(A-1)}{2}} \sum_{x<y} d_{[x,y]}$;
- **Variance:** $\sigma^2_{d_{x,y}} = \frac{1}{\frac{A(A-1)}{2}-1} \sum_{x<y} \left(d_{[x,y]} - \mu_{d_{[x,y]}} \right)^2$; and
- **Standard Error:** $S_{\bar{d}_{[x,y]}} = \sqrt{\frac{\sigma^2_{d_{[x,y]}}}{\frac{A(A-1)}{2}}}$.

$\frac{A(A-1)}{2}$ is the 2×2 combination of the alleles i.e., pairs in the distances estimators

Markers w/o Allelic Information ...

		Genotype x	
		1	0
Genotype y	1	a	b
	0	c	d

- **Simple Matching:** $d_{[x,y]} = \frac{b+c}{a+b+c+d}$;
- **Jaccard:** $d_{[x,y]} = \frac{b+c}{a+b+c}$; and
- **Nei and Li (or Dice):** $d_{[x,y]} = \frac{2(b+c)}{2a+b+c}$.

The same references for the previous cases i.e., [1, 2, 3].

Diversity Indices ...

A locus is polymorphic if, and only if, its allele frequency is ≤ 0.95 or 0.99

Raw ...

- **Polymorphic proportion:**

$$P = \frac{n_{poly}}{n_{total}}$$

- **Mean allele number by locus:**

$$n_a = \frac{1}{L} \sum_{l}^{L} n_l$$

Applied ...

- **Observed Heterozygosity** (Ho);
- **Expected Heterozygosity** (He);
- **Effective Number of Alleles** (Ae); and
- **Shannon Index** (SH).

Observed Heterozygosity:

It is defined as the percentage of heterozygous loci per individual or the number of heterozygous individuals per locus.

Expected Heterozygosity:

$$He = \frac{1}{L}\sum_{l}^{L}\left(1 - \sum_{a}^{A}\hat{p}_{la}^{2}\right) \quad ,(0.0 \leq He \leq 1.0)$$

where:
$\hat{p}_{la}$ is the estimated frequency of the allele a at locus l; and
the other quantities (L and A) were already defined.

NOTE:

$$\sum_{a}^{A}\hat{p}_{la} = 1.0$$

Effective Number of Alleles:

$$Ae_l = \frac{1}{\sum_a^A \hat{p}_a^2}$$

$$Ae = \frac{1}{L} \sum_l^L Ae_l$$

NOTE:

$$Ae_l = \frac{1}{1-He_l}$$

Shannon Index:

- Total Frequencies:

$$SH_{\text{Total}} = -\sum_{a}^{A} \hat{p}_a \log_{10}\left(\hat{p}_a\right) \ , \sum_{a}^{A} \hat{p}_a = 1.0$$

- By Locus Frequencies:

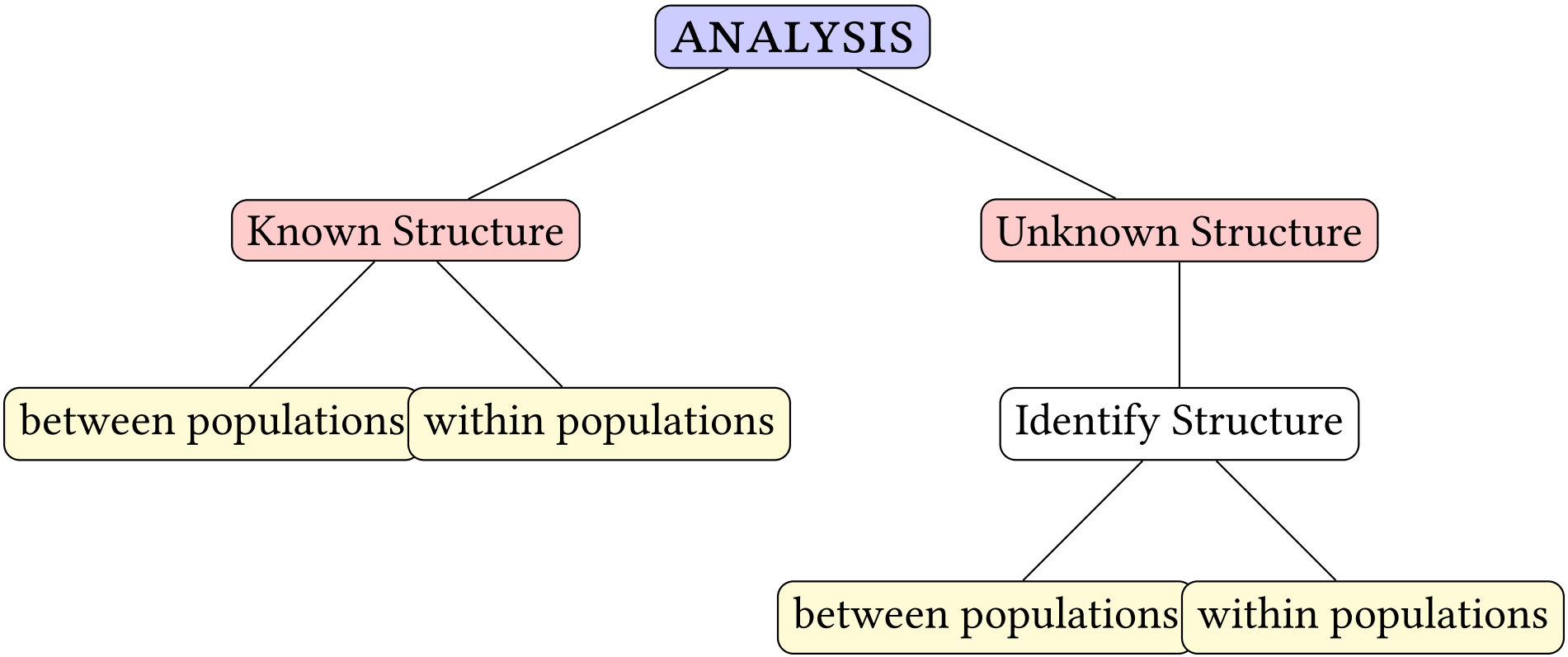
$$SH_{\text{Locus}} = -\sum_{a}^{A} \hat{p}_a \log_{2}\left(\hat{p}_a\right) \ , \sum_{a}^{A} \hat{p}_a = L \therefore \left(0.0 \leq SH_{\text{Locus}} \leq L\right)$$

Diversity Indices for PAVs

Statistics . . . :

- **Polymorphic proportion** ⊘
- **Mean number of alleles by locus** ⊘
- **Observed Heterozygosity** (Ho) ⊘
- **Expected Heterozygosity** (He) ⊘
- **Effective Number of Alleles** (Ae) ⊘
- **Shannon Index** (SH) ⊘

Intraspecific Diversity ...



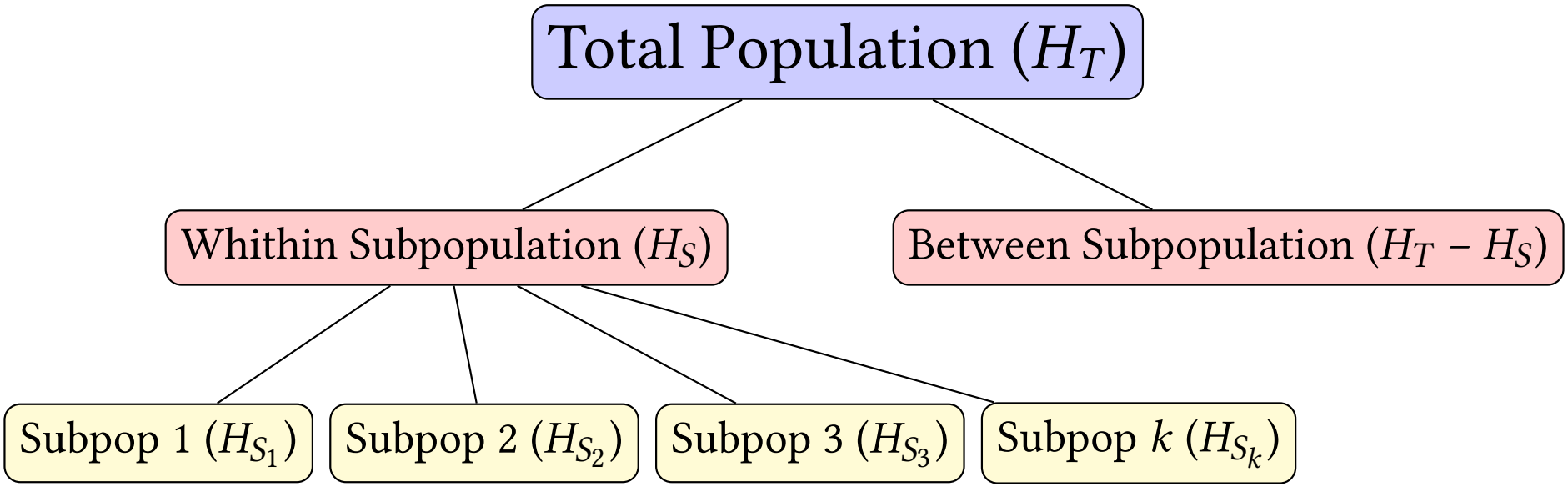
ANOVA & CR Designs ...

	D.F.	M.S.
Between	$T - 1$	"variance between means"
Within	$T \times (R - 1)$	"mean of within variances"
Total	$T \times R - 1$	

Recapping ...

Parameter	Estimator
Population (by locus)	$He_l(H_{T_l}) = 1 - \sum_a^2 \hat{p}_{a_l}^2$
Population (average)	$He(H_T) = \frac{1}{L} \sum_l^L He_l$
Subpopulation$_i$ (by locus)	$He_{i_l} = 1 - \sum_a^2 \hat{p}_{i_{a_l}}^2$
Subpopulation$_i$ (average)	$He_i = \frac{1}{L} \sum_l^L He_{i_l}$
Within Subpopulation	$H_S = \frac{1}{I} \sum_i^I He_i$
Between Subpopulation	$D_{ST} = H_T - H_S$

Total = Between + Within



Wright Statistics (F) ...

F_{IS} Heterozygosis proportional deviations within subpopulations:

$$F_{IS} = \frac{H_S - Ho}{H_S} \; [-1; 1]$$

F_{IT} Overall heterozygosis proportional deviation (inbreeding coefficient):

$$F_{IT} = \frac{H_T - Ho}{H_T} \; [-1, 1]$$

F_{ST} Heterozygosis proportional deviation between subpopulations:

$$F_{ST} = \frac{H_T - H_S}{H_T} \; [0; 1]$$

Nei, 1987 [4] ...:

1. Obtain H_T:

$$H_T = \frac{1}{L} \sum_{l}^{L} He_l$$

2. Obtain H_s (by locus):

$$H_S = \frac{1}{L} \sum_{l}^{L} He_{S_l}$$

3. Thus, obtain G_{ST}:

$$G_{ST} = \frac{D_{ST}}{HT} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T}$$

Berg, 1997 [1] . . . :

1. Obtain G_{ST} by locus:

$$G_{ST} = \frac{D_{ST}}{HT} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T}$$

2. Summarize as average:

$$\bar{G}_{ST} = \frac{1}{n_a} \sum_{n}^{n_a} G_{ST_n}$$

3. That is the same as:

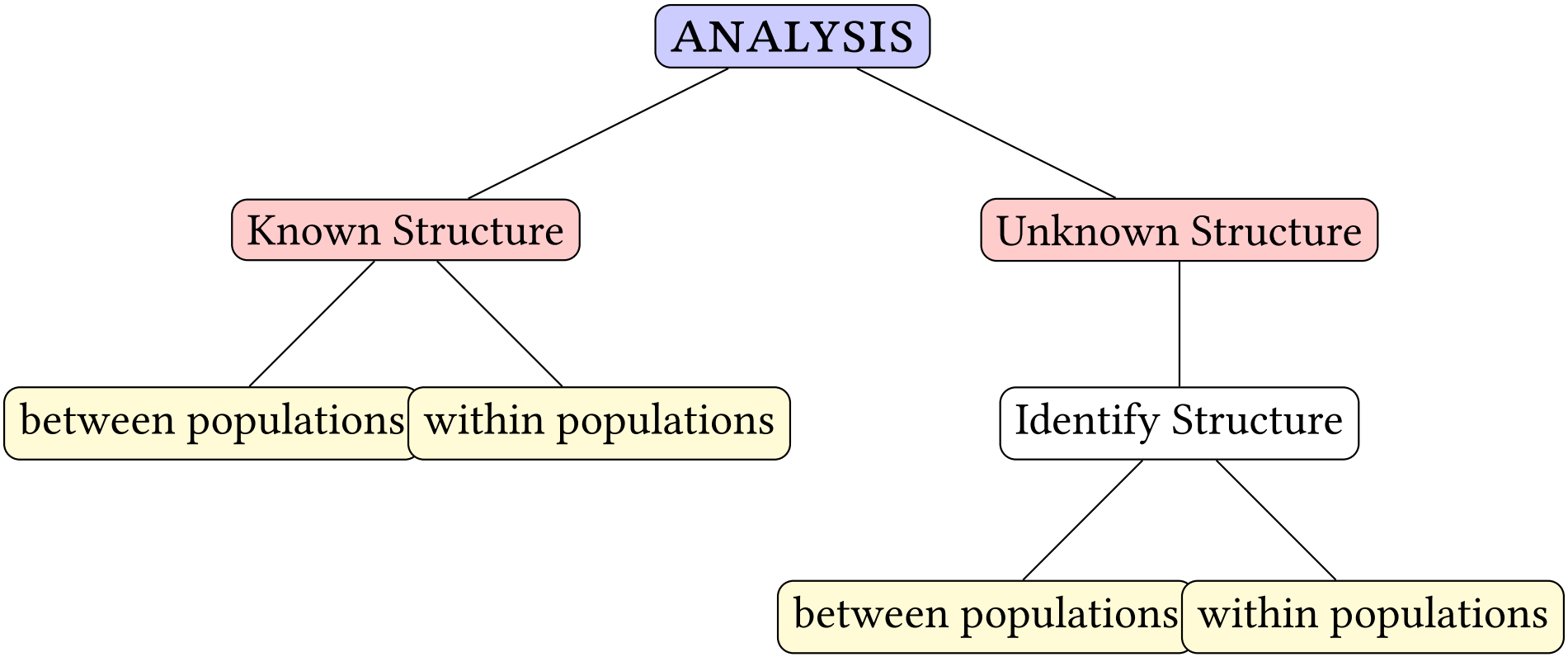
$$\bar{G}_{ST} = 1 - \frac{\sum_{n}^{n_a} \left(\frac{H_{S_n}}{H_{T_n}} \right)}{n_a}$$

Be aware ...

- When we have just one (1) allele: $G_{ST} = F_{ST}$;
- G_{ST} generalizes F_{ST};
- G_{ST} is the proportion of total diversity that is between subpopulations; and
- Thus, $(1 - G_{ST})$ is the proportion of total diversity within subpopulations.
- The meaning of diversity according to F_{ST}:

F_{ST}	means
$[0; 0.05)$	small
$[0.05; 0.15)$	medium
$[0.15; 0.25)$	high
≥ 0.25	very high

...Intraspecific Diversity (*2nd branch*)



Clusters & Groups ...

OBJECTIVE:

Minimize within variability $\longrightarrow$ Maximize between variability

References can be consulted for deep understanding:

- Fundation book ... [5];
- Fundation paper in genetics (Nei) ... [4]; and
- More modern reference ... [6].

$$\boldsymbol{Y}_{n\times p} = \begin{bmatrix} y_{1,1} & y_{1,2} & y_{1,3} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & y_{2,3} & \cdots & y_{2,p} \\ y_{3,1} & y_{3,2} & y_{3,3} & \cdots & y_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & y_{n,3} & \cdots & y_{n,p} \end{bmatrix}$$

Note:

Everything starts from obtaining the distance matrix between all 2×2 pairs ...

$$\frac{n(n-1)}{2}$$

$$\boldsymbol{D} = \begin{bmatrix} 0 & d_{1,2} & d_{1,3} & \cdots & d_{1,j} & \cdots & d_{1,n} \\ & 0 & d_{2,3} & \cdots & d_{2,j} & \cdots & d_{2,n} \\ & & 0 & \cdots & d_{3,j} & \cdots & d_{3,n} \\ & & & \ddots & \vdots & & \vdots \\ & & & & 0 & \cdots & d_{j,n} \\ & & & & & \ddots & \vdots \\ & & & & & & 0 \end{bmatrix}$$

$d_{i,i} = 0.0$ and D is symmetric $\therefore d_{i,j} = d_{j,i}$

Clustering Methods

- **Geometrical:**
  - Hierarchical;
  - Neighbor Joining;
  - *k-means* (density search); and others
- **MANOVA:**

$$\text{Total} = \text{Between} + \text{Within}$$

- **Statistical:**

  Mixture and Bayesian (STRUCTURE)

Hierarchical:

1. All distances;
2. Find the most close individuals (smaller distance); and
3. Iterate over that.
    - UPGMA: merge groups with smaller distance;
    - WARD: merge groups that generate a new one with smaller $S.S._W$; and
    - NJ: merge groups if: (*i*) smaller distance & (*ii*) higher distances in comparison to the others.

Bayesian:

The model account for the presence of HW or LD, introduces population structure...find population structure (groups) with the smaller possible disequilibrium [6].

Multidimensional Scaling ...

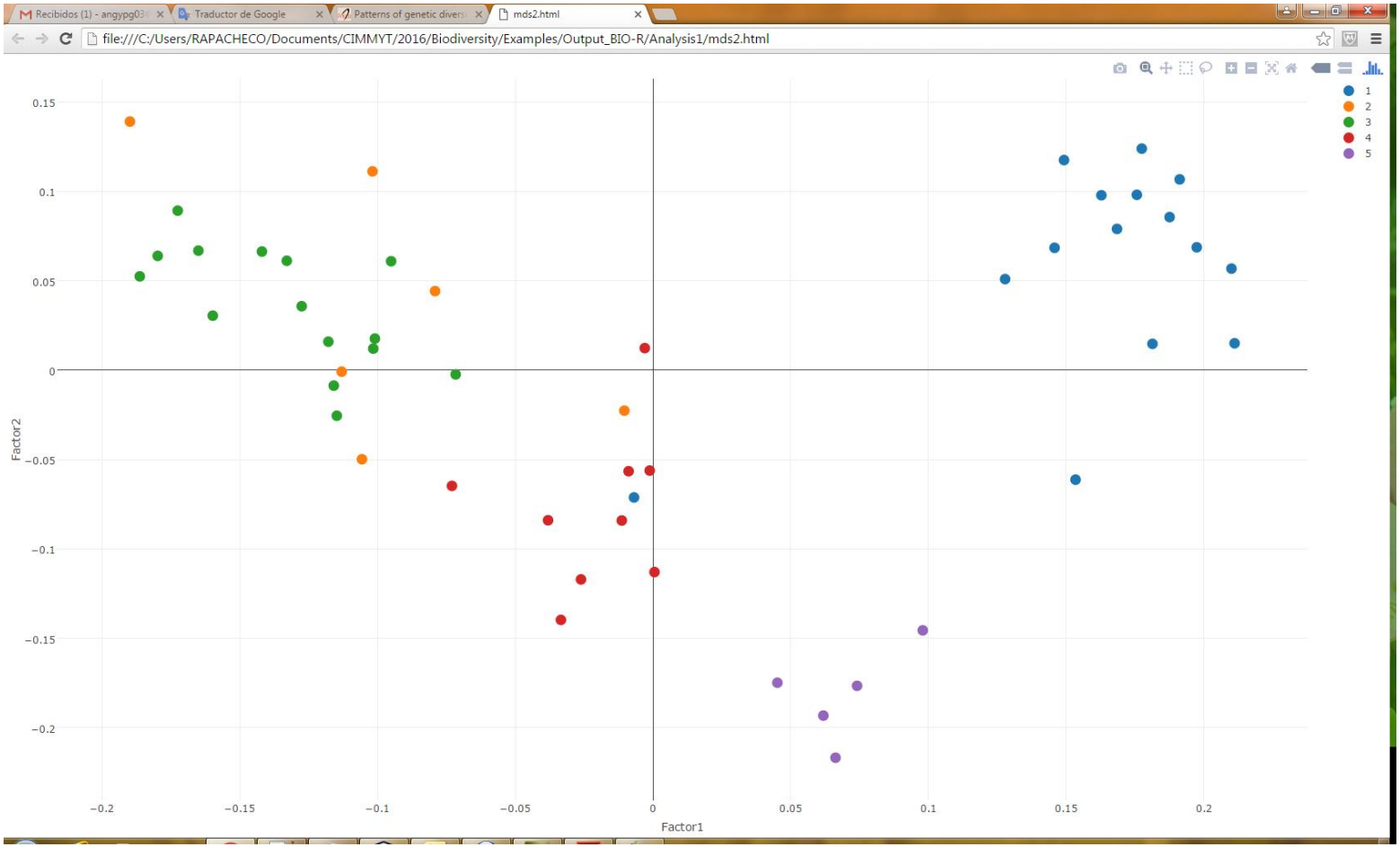
Two concepts:

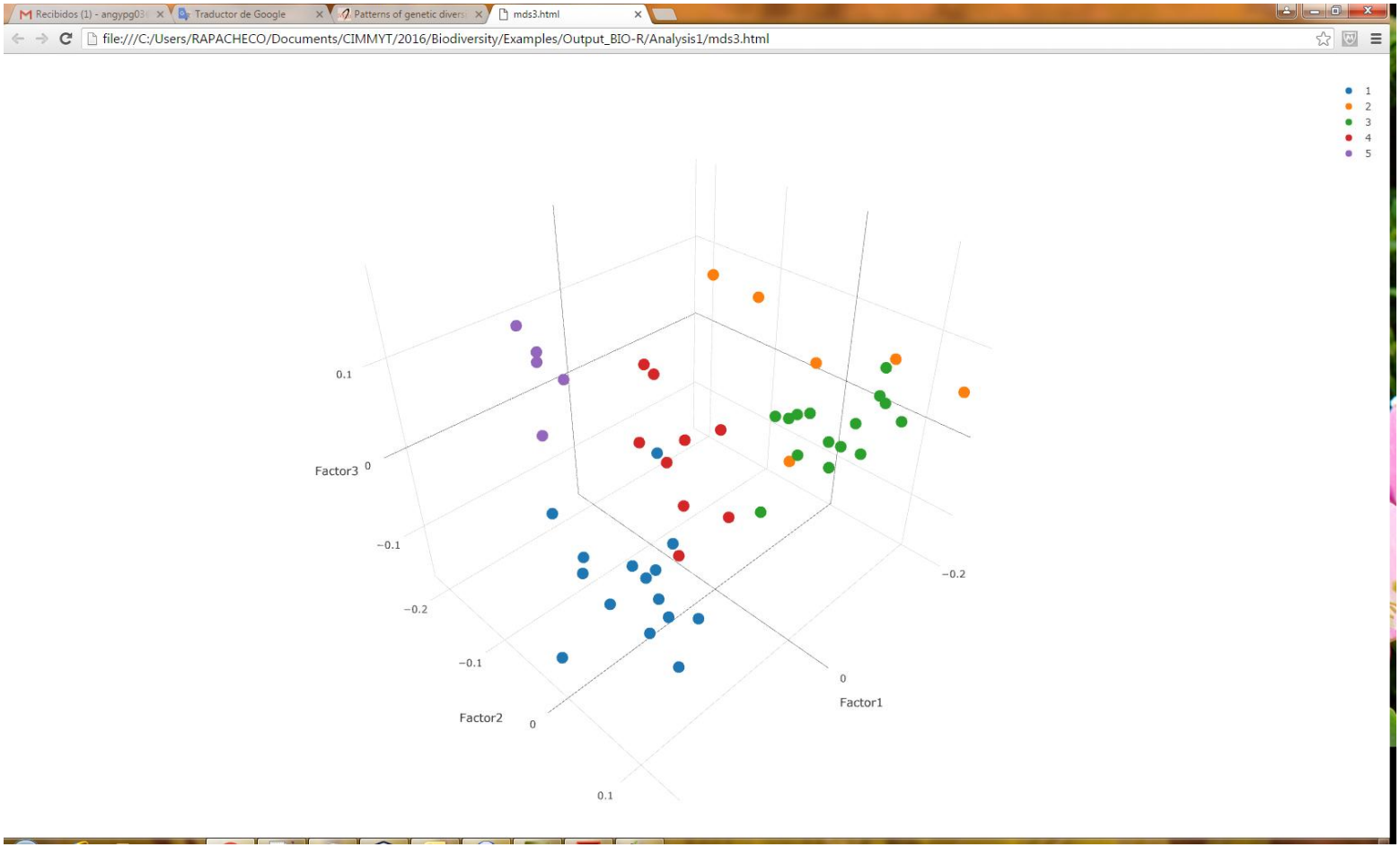
1. **Similarity** ($s_{[x,y]}$): $d_{[x,y]} = \sqrt{2 \times \left(1 - s_{[x,y]}\right)}$;

2. ***stress*** (S): $S = \sqrt{\frac{\sum_x \sum_y \left(d_{[x,y]} - \mu_{d_{[x,y]}}\right)^2}{\frac{n(n-1)}{2}}}$

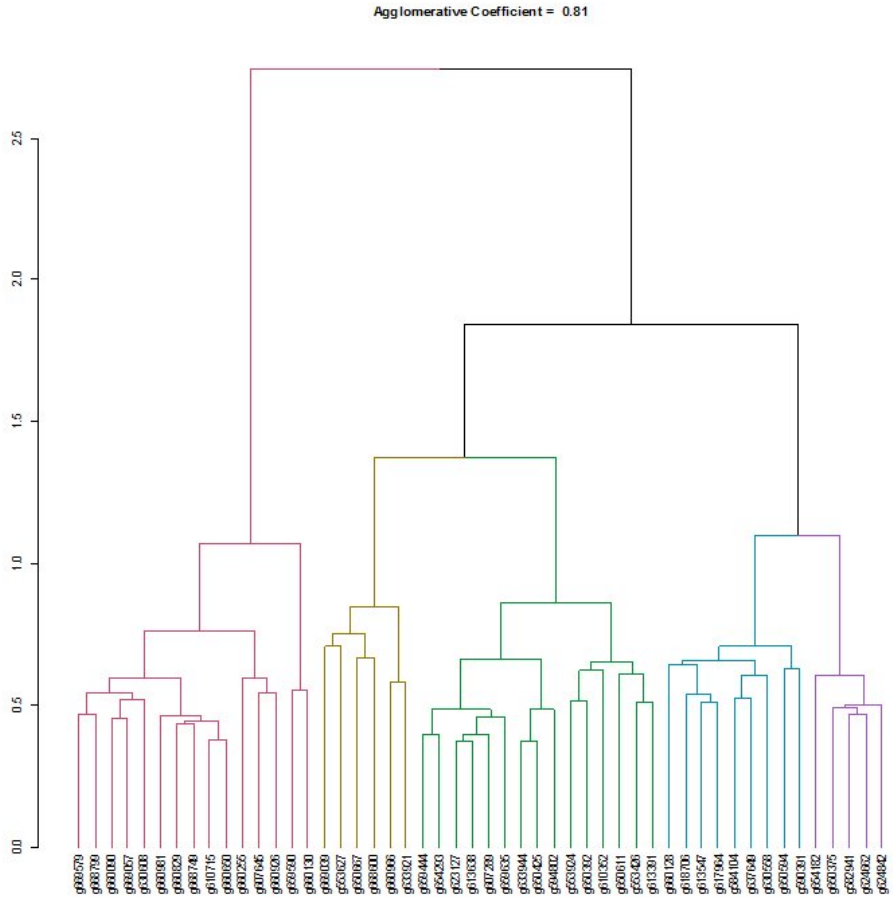
NOTE:

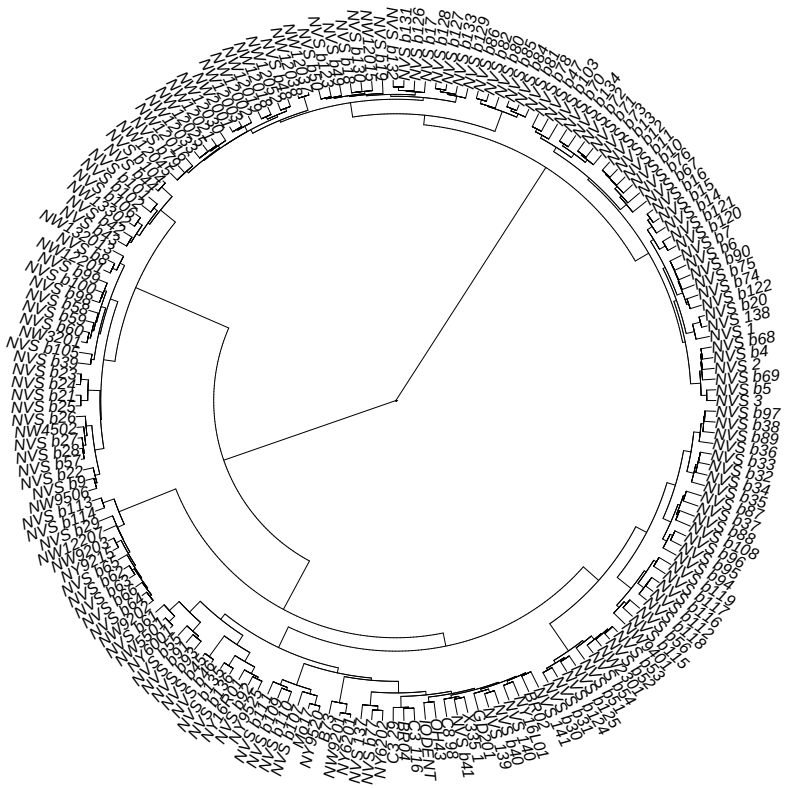
Search for the reduced representation (2 or 3 dimensions) that minimizes the $S.S._B$ in the p dimensional space and with minimum the estimated distance.


file:///C:/Users/RAPACHECO/Documents/CIMMYT/2016/Biodiversity/Examples/Output_BIO-R/Analysis1/mds2.html
1
2
3
4
5
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
Factor2
−0.2
−0.15
−0.1
−0.05
0
0.05
0.1
0.15
0.2
Factor1


Recibidos (1) - angypg03
Traductor de Google
Patterns of genetic diversi
mds3.html
file:///C:/Users/RAPACHECO/Documents/CIMMYT/2016/Biodiversity/Examples/Output_BIO-R/Analysis1/mds3.html
1
2
3
4
5
0.1
Factor3
0
−0.1
−0.2
−0.1
Factor2
0
0.1
0
Factor1
−0.2

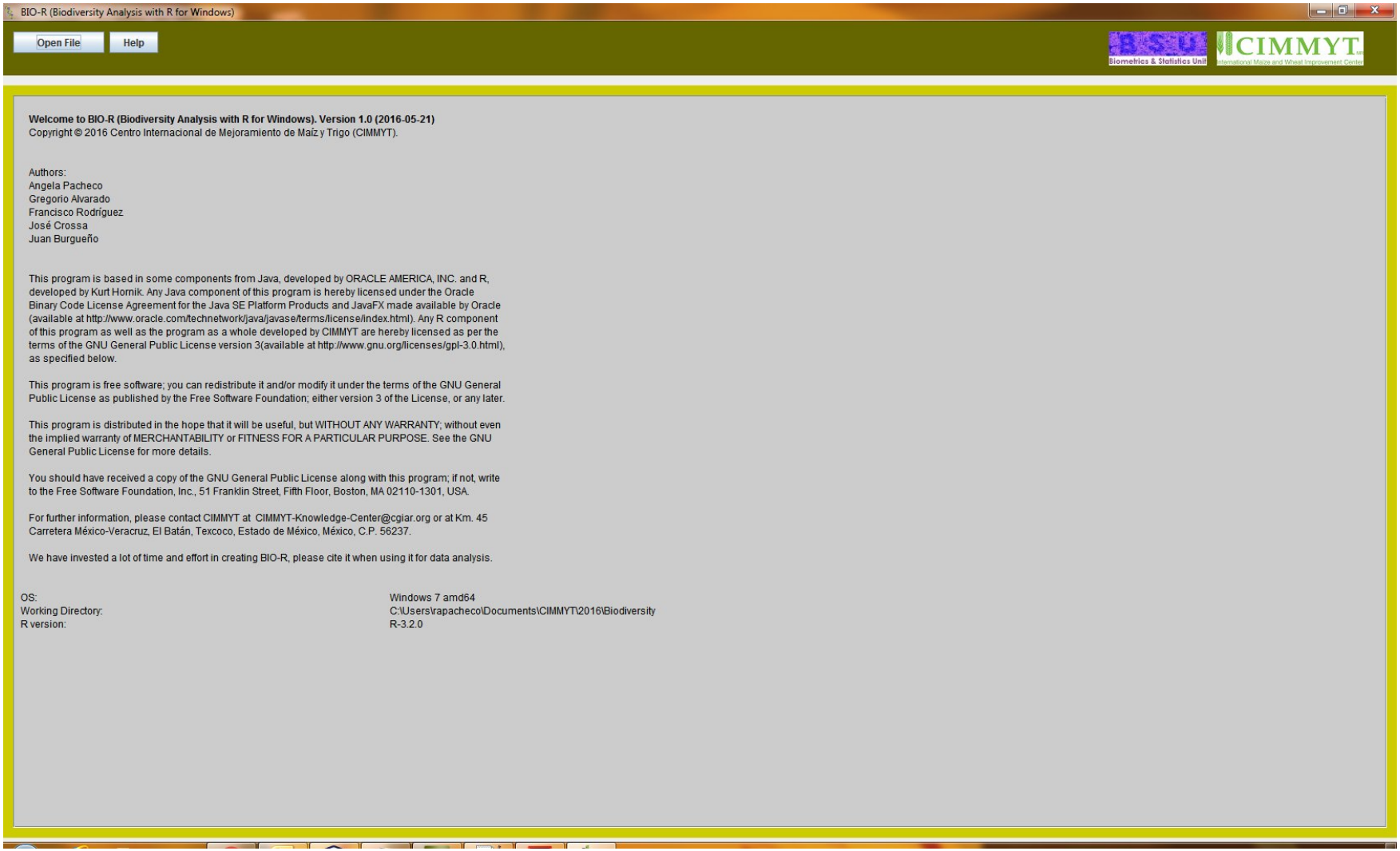
Dendograms ...





BIO-R Tutorial

- **Phenotypic data**
  - *Experimental Design:* ADEL, AUDE & STAD;
  - *Individual & Multi-Env:* META (> 1500), AGD (> 1300);
  - *Interaction/Stability:* GEA (> 1800); and
  - *Spatial Analysis:* SPAT.
- **Genotypic data**
  - *kinship:* BROWSE & COP; and
  - *Diversity:* BIO.
- **Fusion data**
  - *Relationship/Interaction:* GEA;
  - *Genome prediction:* BGLR (*R package & application tool); and
  - *Selection Index:* SI (> 300), RINDSEL.
- **Decision:** *Eval L × T*.


BIO-R (Biodiversity Analysis with R for Windows)
Open File
Help
Biometrics & Statistics Unit
CIMMYT
International Maize and Wheat Improvement Center
Welcome to BIO-R (Biodiversity Analysis with R for Windows). Version 1.0 (2016-05-21)
Copyright © 2016 Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT).
Authors:
Angela Pacheco
Gregorio Alvarado
Francisco Rodríguez
José Crossa
Juan Burgueño
This program is based in some components from Java, developed by ORACLE AMERICA, INC. and R, developed by Kurt Hornik. Any Java component of this program is hereby licensed under the Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX made available by Oracle (available at http://www.oracle.com/technetwork/java/javase/terms/license/index.html). Any R component of this program as well as the program as a whole developed by CIMMYT are hereby licensed as per the terms of the GNU General Public License version 3(available at http://www.gnu.org/licenses/gpl-3.0.html), as specified below.
This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or any later.
This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.
For further information, please contact CIMMYT at CIMMYT-Knowledge-Center@cgiar.org or at Km. 45 Carretera México-Veracruz, El Batán, Texcoco, Estado de México, México, C.P. 56237.
We have invested a lot of time and effort in creating BIO-R, please cite it when using it for data analysis.
OS: Windows 7 amd64
Working Directory: C:\Users\rapacheco\Documents\CIMMYT\2016\Biodiversity
R version: R-3.2.0

Downloading and Installation

- BIO-R is already available **go to:** `http://data.cimmyt.org`;
- Java interface/application for embedding *R* [7] scripts to perform Diversity analysis; and
- heterozygosity, diversity B & W groups, shannon index, number of effective allele, % of polymorphic loci, Rogers & Nei distance, clusters and multidimensional scaling (2 & 3d plots).

CREDITS: BSU/CIMMYT

Angela Pacheco <R.A.Pacheco@cgiar.org>
Francisco Rodriguez <F.R.Huerta@cgiar.org>
Gregorio Alvarado <G.Alvarado@cgiar.org>
Juan Burgueño <J.Burgueno@cgiar.org>

CIMMYT
International Maize and Wheat Improvement Center

Research Software Repository

CIMMYT Dataverse Network >

POWERED BY THE Dataverse Network PROJECT v. 3.0

CIMMYT Research Software Dataverse

Create Account Log In

BIO-R (BIODIVERSITY ANALYSIS WITH R FOR WINDOWS) VERSION 1.0

< View Previous Study Listing

hdl:11529/10820

Version: 2– Released: Tue Dec 06 09:03:26 CST 2016

CATALOGING INFORMATION | Data & Analysis | Comments | Versions

If you use these data, please add the following citation to your scholarly references. Why cite?

Data Citation

Pacheco, Ángela; Alvarado, Gregorio; Rodríguez, Francisco; Burgueño, Juan, 2016-12-06, "BIO-R (Biodiversity analysis with R for Windows) Version 1.0", http://hdl.handle.net/11529/10820 nternational Maize and Wheat Improvement Center [Distributor] V2 [Version]

Citation Format Print

Data Citation Details

Study Global ID	hdl:11529/10820
Other ID	CIMMYT: cimmytswdvn-0
Authors	Pacheco, Ángela (BSU - CIMMYT); Alvarado, Gregorio (BSU - CIMMYT); Rodríguez, Francisco (BSU - CIMMYT); Burgueño, Juan (BSU - CIMMYT)
Producer	International Maize and Wheat Improvement Center (CIMMYT) CIMMYT
Production Date	diciembre 06, 2016
Distributor	nternational Maize and Wheat Improvement Center (CIMMYT) CIMMYT
Distributor Contact	Juan Burgueño (BSU - CIMMYT), J.Burgueno@cgiar.org
Distribution Date	diciembre 06, 2016
Deposit Date	diciembre 06, 2016
Provenance	CIMMYT Research Software Dataverse

Getting Help & Manual

The Help button will provide you two options:

- **Manual:** `*.pdf` document descripting all options, methods and outputs.

- **About:** How to cite BIO-R and any associated license (GNU & Oracle)

Data Format – *MyData.csv*

mark	g1	g2	g3	g4	...	gX
1	0.67	NA	0.57	NA	...	1
2	NA	1	1	1	...	NA
3	1	1	0.52	0.50	...	0.80
4	1	NA	0.50	1	...	NA
5	1	NA	1	NA	...	1
6	0.67	0	0.71	1	...	0.33
7	1	NA	0	1	...	1
8	1	NA	1	1	...	1
9	NA	1	0.60	0.22	...	0.71
⋮	⋮	⋮	⋮	⋮	⋱	⋮
N	1	1	1	1	...	1

Setup ...

1. `Markers` – selects the column that identify the **markers**;
2. `# Clusters` – type the number of groups to split the population;
3. `Output folder` – type the path of the output folder where results will be saved;
    - It will be created inside the BIO-R's Output folder;
    - You can change the name for different sets; and
    - It is necessary to change the name for each analysis.
4. `Genotypes` – selects the columns that identify the **genotypes**;
5. `Distance` – selects the method to calculate **distances**; and
6. `ColorMDSPlot` ...*[CONT.]*

ColorMDSPlot

Specify a `*.csv` file containing additional information for colors in MDS plot

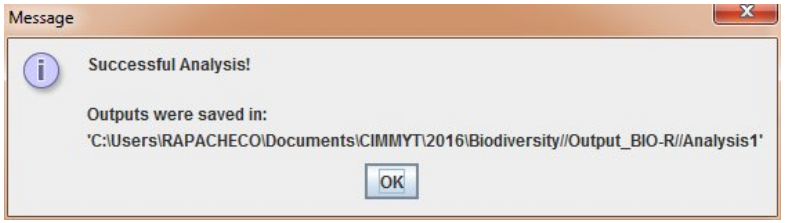
Collumn 1: the name of the genotypes – should be equal to the input data;

Collumn 2: the variable to identify the groups of different colours; and

Collumn 3: any additonal information.

Genotype	NumColor	Something
g669579	11	I
g669039	3	A
g659444	8	F
g660128	11	I
⋮	⋮	⋮
g660829	11	I

Outputs ...



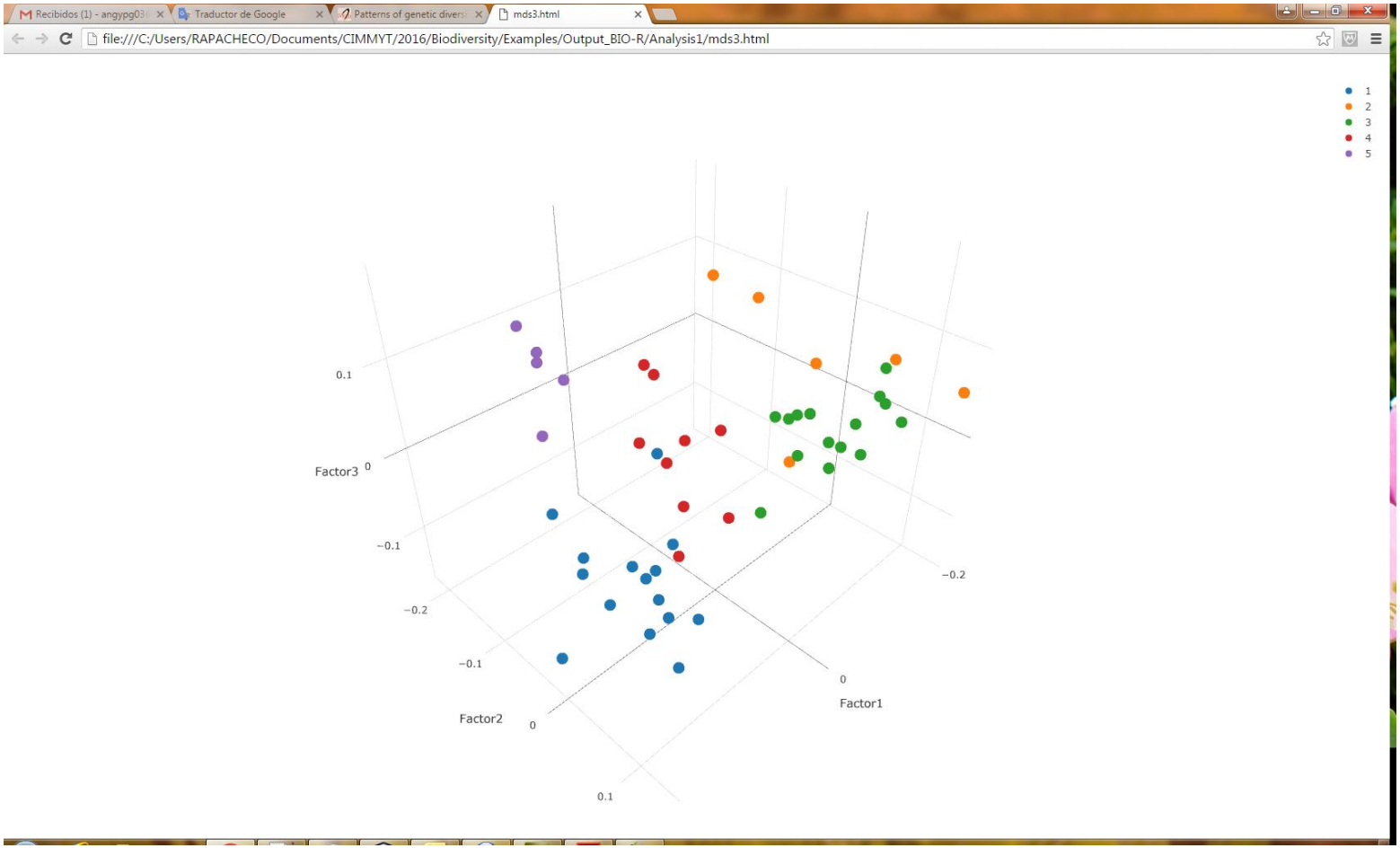
- Output
  - Analysis 1
    - CalculusPerGenotype.csv
    - CalculusPerLocus.csv
    - Dendogram.wmf
    - mds2.html
    - mds3.html
    - MDStable.csv
    - RogersDistances.csv
    - SummaryDiversityAnalysis.csv
    - mds2_files
    - mds3_files

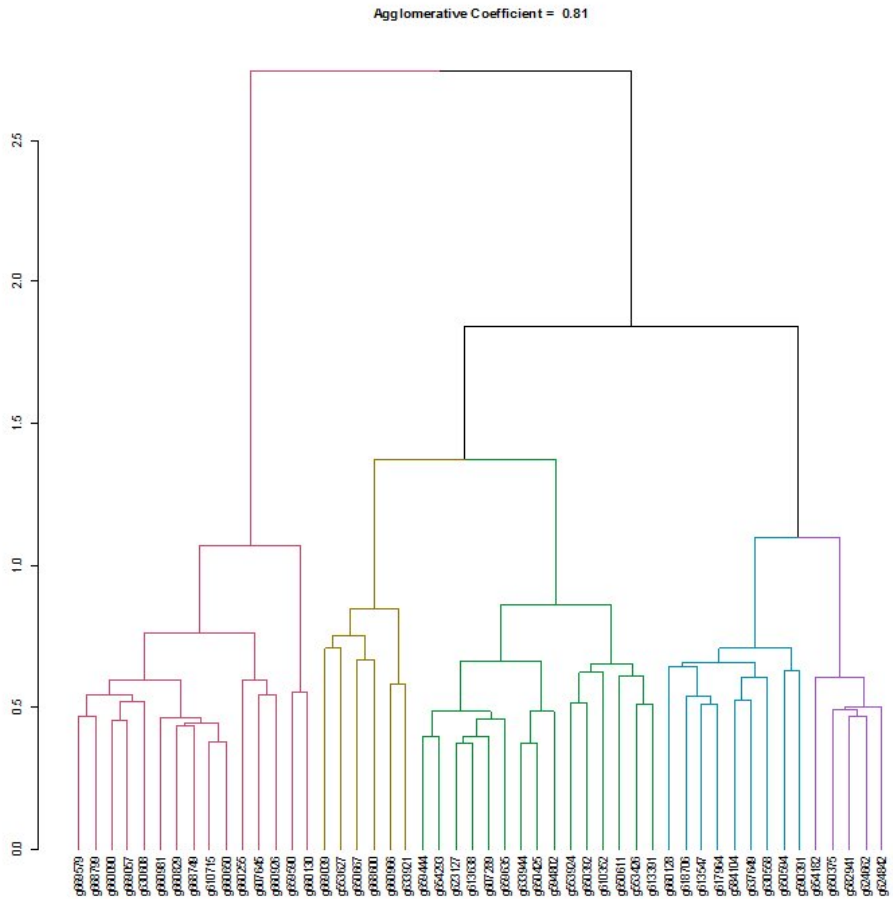
Output : `CalculusPerLocus.csv`

Marker	He	Ho	Ae	Shannon	%NA
1	0.50	0.29	1.99	0.95	0.22
2	0.01	-0.05	1.01	0.02	0.08
3	0.45	0.38	1.82	0.93	0.16
4	0.41	0.42	1.69	0.87	0.24
5	0.03	-0.08	1.03	0.11	0.10
6	0.40	0.43	1.68	0.86	0.20
7	0.50	0.31	2.00	0.99	0.24
8	0.49	0.47	1.94	0.98	0.18
9	0.45	0.47	1.82	0.93	0.14
⋮	⋮	⋮	⋮	⋮	⋮
N	0.08	-0.03	1.09	0.26	0.12

Output : `CalculusPerGenotype.csv`

Genotype	He	Ho	Ae	Shannon	%NA	clusterGroup
1	0.39	0.06	1.63	0.83	0.11	1
2	0.38	-1.46	1.62	0.82	0.37	2
3	0.40	0.51	1.67	0.85	0.03	3
4	0.39	0.03	1.64	0.83	0.19	4
5	0.37	0.39	1.60	0.81	0.03	3
6	0.41	-3.56	1.69	0.86	0.42	1
7	0.39	0.34	1.65	0.84	0.04	3
8	0.37	-0.28	1.54	0.81	0.24	5
9	0.40	-0.70	1.61	0.82	0.32	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X	0.38	0.33	1.66	0.84	0.04	3


Recibidos (1) - angypg03
Traductor de Google
Patterns of genetic divers
mds3.html
file:///C:/Users/RAPACHECO/Documents/CIMMYT/2016/Biodiversity/Examples/Output_BIO-R/Analysis1/mds3.html
1
2
3
4
5
0.1
Factor3
0
−0.1
−0.2
−0.1
Factor2
0
0.1
0
Factor1
−0.2


Agglomerative Coefficient = 0.81
0.0
0.5
1.0
1.5
2.0
2.5

Summary : ***SummaryDiversitysAnalysis.csv***

- **% of polymorphic loci:** 0.94;
- **Exp. Heterozygosity:** 0.30;
- **Std. Dev. of** He**:** 0.01;
- **Obs. Heterozygosity:** 0.22;
- **Std. Dev. of** Ho**:** 0.01;
- **Number of effective alleles:** 1.55;
- **Std. Dev. of** Ae**:** 0.02;
- **Shannon Index:** 0.63;
- **Std. Dev. Shannon:** 0.02 ...

Remarks

- Frequency of Alleles & Genotypes;

  (Population Genetics) The fundamental concept!

- Distances & Diversity Indices;

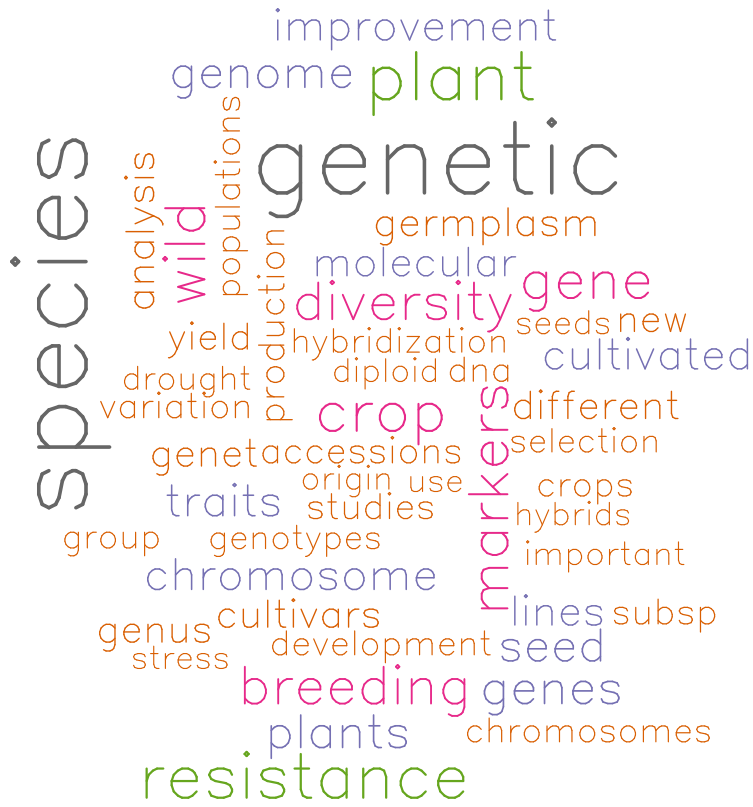
  (Genetic Diversity) Several ways and flavors.

- Wright Statistics;

  (Intraspecific Diversity) Between/Within variability

- BIO-R; and

  (Software) it has been built considering everything you may want



Questions/Suggestions??

For Further Studies:

[1] BERG, E. E.; HAMRICK, J. L. Quantification of genetic diversity at allozyme loci. *Canadian Journal of Forest Research*, v. 27, p. 415–429, 1997.

[2] MOHAMMADI, S. A.; PRASANA, B. M. Analysis of genetic diversity in crop plants – salient statistical tools and considerations. *Crop Science*, v. 43, p. 1235–1248, 2003.

[3] REIF, J. C.; MELCHINGER, A. E.; FRISH, M. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science*, v. 45, p. 1–7, 2005.

[4] SAITOU, N.; NEI, M. The neighbour-joining method: A new method fo reconstructing phylogenetic trees. *Molecular Biology and Evolution*, v. 4, n. 4, p. 406–425, 1987.

[5] KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis.* New York: John Wiley, 1990. 368 p.

[6] PRITCHARD, J. K.; STEPHENS, M.; DONNELY, P. Inference of population structure using multilocus genotype data. *Genetics*, p. 945–959, 2000.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>.