

**ALAG 2019: La arquitectura del genoma: su expresión en los fenotipos y las poblaciones. Mendoza (Argentina) 6 al 9 de octubre de 2019**

**Simposio "ANÁLISIS DE BASES DE DATOS DE ALTA DIMENSIONALIDAD EN EL ESTUDIO DEL GENOMA Y SU EXPRESIÓN"**

## **REDUCCIÓN DE LA DIMENSIONALIDAD DE DATOS EN ANÁLISIS DEL GENOMA Y SU EXPRESIÓN**

**Victoria G. Cabodevila <sup>1</sup>, Ana Paula Del Medico <sup>1,2</sup>, Leandro Kovalevski <sup>2</sup>, Paula Macat <sup>2</sup>, Marta B. Bianchi <sup>3</sup>, Marta B. Quaglino <sup>2</sup>, María Susana Vitelleschi <sup>2</sup>, Guillermo R. Pratta <sup>1</sup>**

**<sup>1</sup> IICAR, CONICET-UNR, Zavalla, Argentina.**

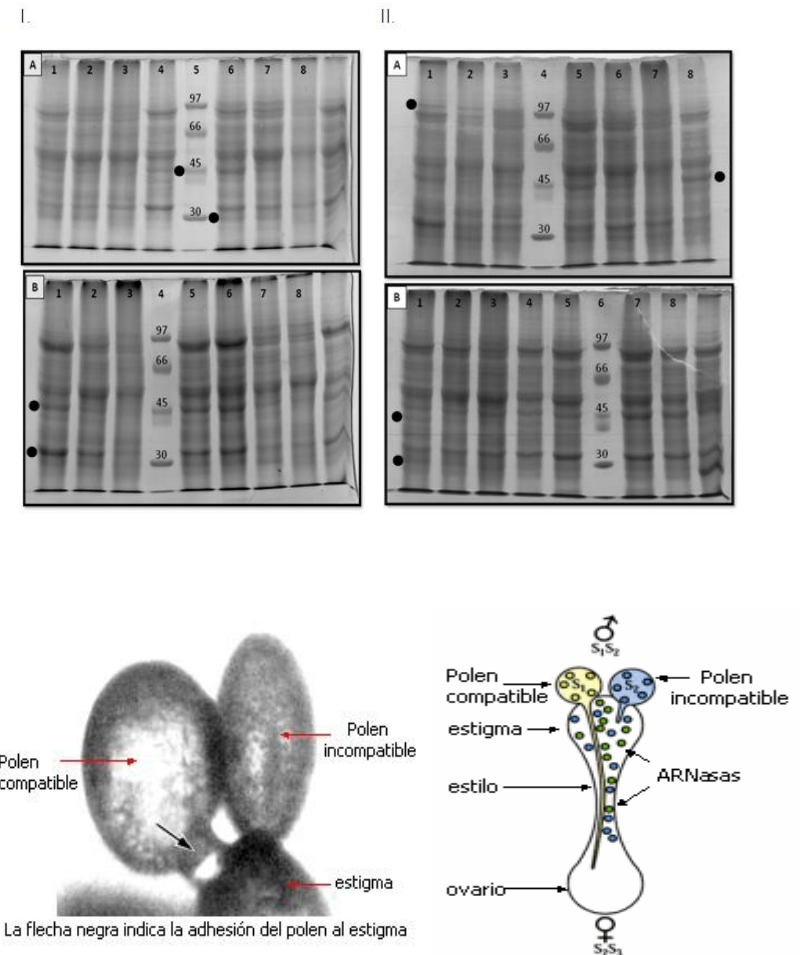
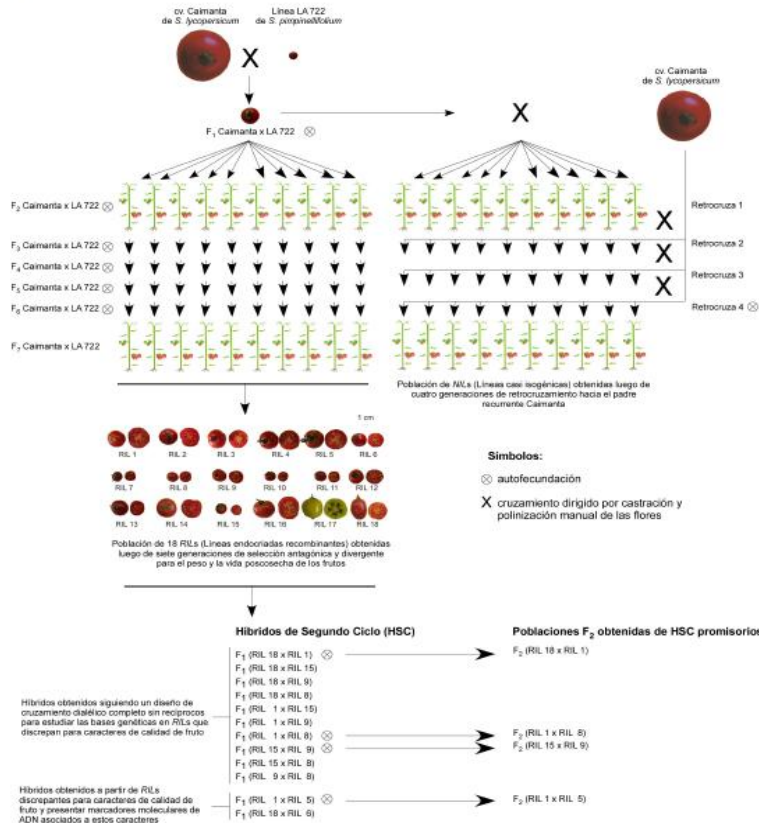
**<sup>2</sup> IITAE, Facultad de Ciencias Económicas y Estadística UNR, Rosario, Argentina.**

**<sup>3</sup> Cátedra de Botánica, Facultad de Ciencias Agrarias UNR, Zavalla, Argentina.**

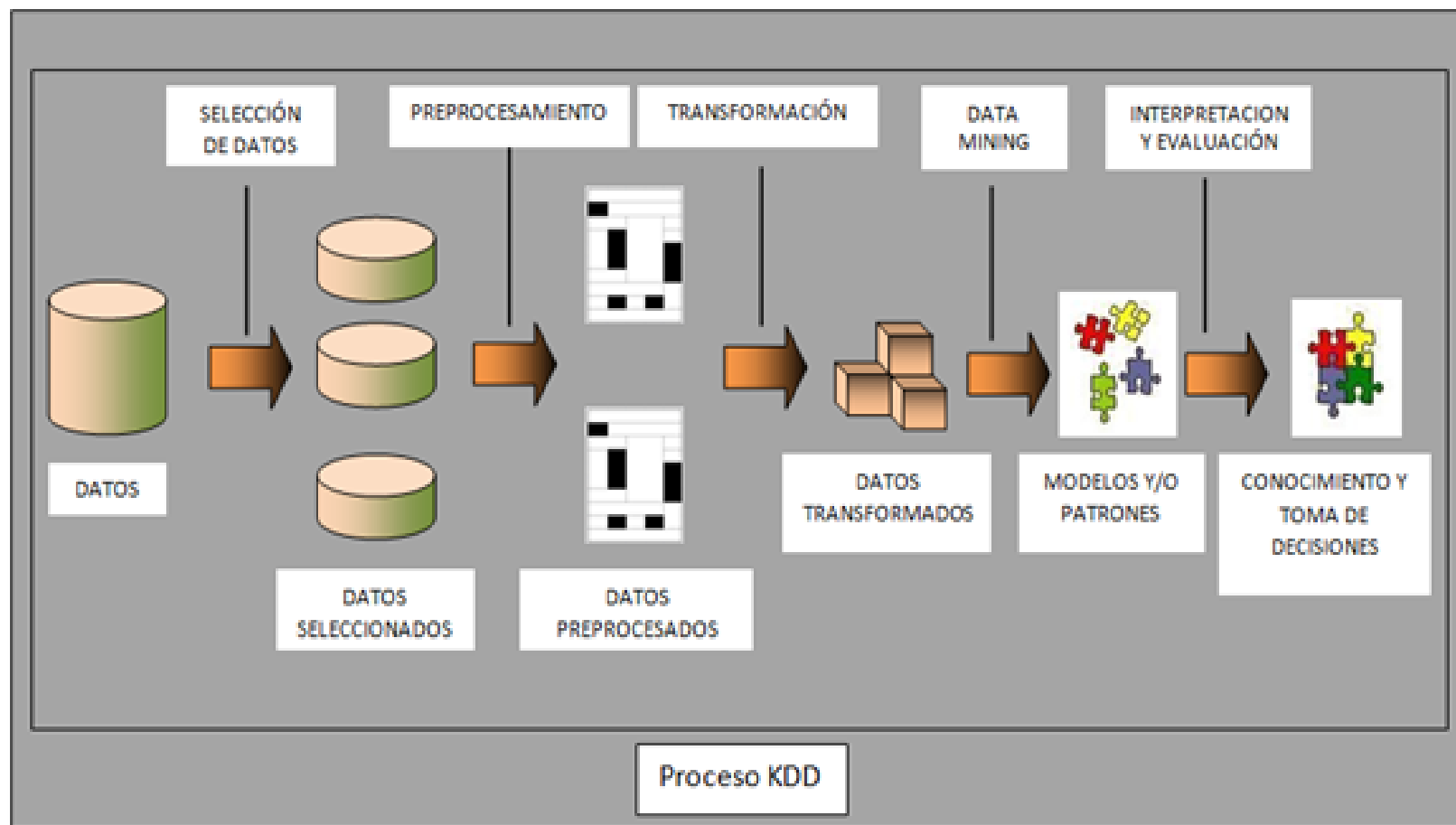
**Email: [gpratta@unr.edu.ar](mailto:gpratta@unr.edu.ar)**

- En los estudios del genoma y su expresión es frecuente relevar un gran número de observaciones, ya sea porque se analizan muchas variables simultáneamente, porque el número de individuos evaluados es grande o porque el problema bajo estudio está afectado por interacciones de diferente índole.

**Figura 3.** Recursos genéticos generados por la Cátedra de Genética FCA-UNR a partir del cruzamiento entre el cultivar Caimanta de *S. lycopersicum* y la línea LA722 de *S. pimpinellifolium*.

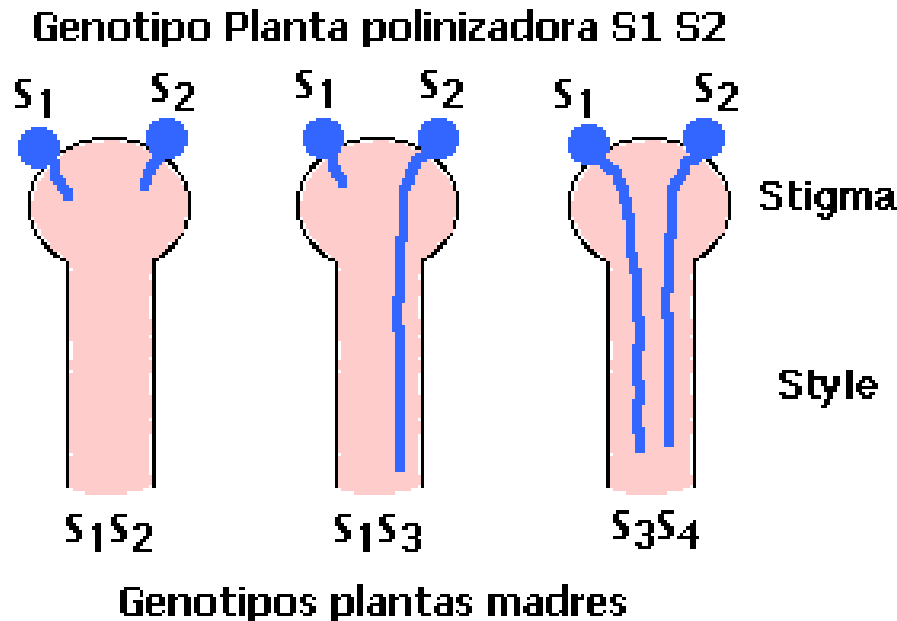


- ❑ A fin de obtener un usufructo óptimo de los resultados experimentales, es necesario recurrir a métodos que reducen la dimensionalidad de las bases generadas, minimizando la pérdida de información y jerarquizando los datos retenidos.



- ❑ La aplicación de estos métodos debe decidirse en función de los objetivos de la investigación, la naturaleza y el origen de los datos, la disponibilidad de recursos informáticos, entre otros factores.
- ❑ En esta charla se presentarán cuatro estudios con diseños y objetivos diferentes (exploratorios, descriptivos, inferenciales y asociativos), realizados aplicando técnicas de reducción de dimensionalidad sobre datos cualitativos, cuantitativos, moleculares, fenotípicos y combinaciones de estos.
- ❑ Los procesos biológicos estudiados fueron incompatibilidad gametofítica, madurez del fruto de tomate, estimación de heredabilidad multivariada y expresión de proteínas en cruzamientos dialélicos.

## INCOMPATIBILIDAD GAMETOFÍTICA



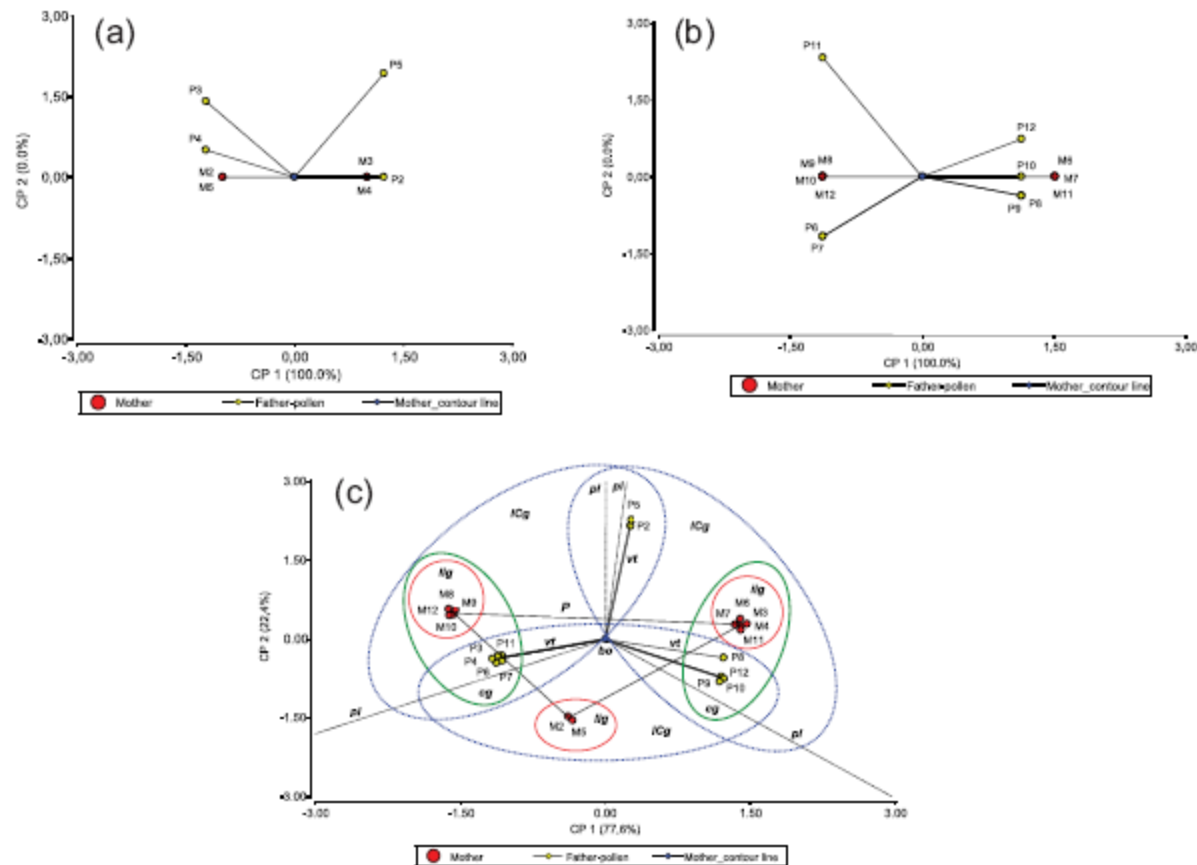
El resultado (fecundación o no fecundación) depende de la interacción entre el fenotipo del grano de polen y el fenotipo del carpelo.

Análisis biplot: interacción genotipo x ambiente, planta x patógeno. Estudio exploratorio y luego, descriptivo

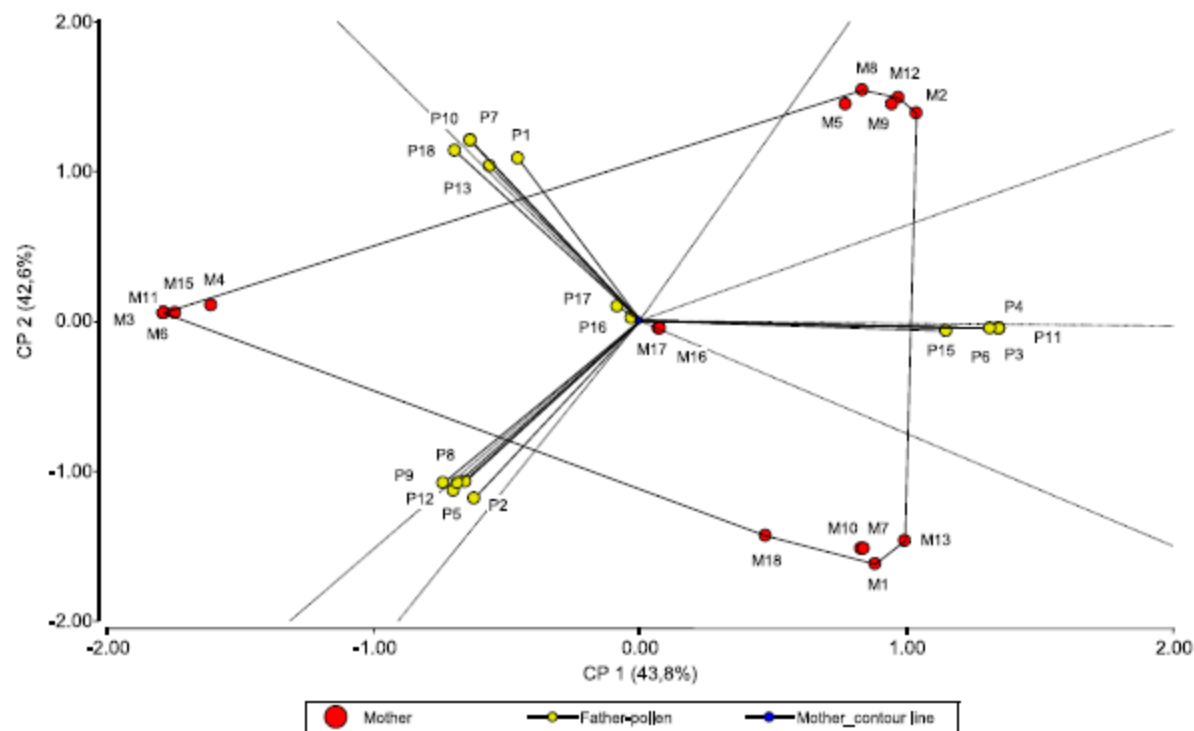
Datos expresados en frecuencias.

La incompatibilidad gametofítica es generalmente debida a un único locus multialélico con expresión del alelo portado por el grano de polen (gametofito masculino) en el progenitor macho y codominancia para la expresión del carpel en el carpelo (progenitor femenino)

- ❑ Se reanalizaron con este enfoque datos publicados: Talavera et al. (2001). Genetic control of self-incompatibility in *Anagallis monelli* (Primulaceae: Myrsinaceae). Heredity 87: 589-597.
- ❑ Diferentes cruzamientos dialélicos de medio hermanos y hermanos completos
- ❑ Número de frutos cuajados / número de polinizaciones manuales efectuadas
- ❑ Resultados del biplot: las dos primeras Componentes principales explicaron un alto porcentaje de la variación
- ❑ Se formaron grupos de plantas intercompatibles e interincompatibles, con diferencias al actuar como macho y como hembra
- ❑ Interpretación: ajusta al modelo de incompatibilidad monogénico, multiloci, gametofítico.



**Figure 1.** (a) *Lysimachia monelli*, biplot analysis for diallel D4. Biplot graphic: genotypes are labelled with **M** when viewed as entry (M plant) and with **P** when viewed as tester (pollen donor, P plant). CP1 and CP2 means principal component (PC1 and PC2) output from the original computer program Info-Gen (UNC) Genotype Environment Interaction [Version 15/09/2013]. (b) *L. monelli*, biplot analysis for diallel D4R. (c) *L. monelli*, biplot analysis for the combined diallels D4 x D4R. Polygon **P**; biplot origin **bo**; perpendicular line **pl**; vector of the tester **vt**; intra-incompatible group **iig** (red circle); compatible group **cg** (green circle) and inter-compatible group **ICg** (blue dotted circle). Genotype labels and computer program in (b) and (c) as for (a).



**Figure 3.** *Lysimachia monelli*, biplot analysis for diallel D6. Biplot graphic: genotypes are labelled with **M** when viewed as entry (M plant) and with **P** when viewed as tester (pollen donor, P plant). CP1 and CP2 means principal component (PC1 and PC2) output from the original computer program Info-Gen (UNC) Genotype Environment Interaction [Version 15/09/2013].



## MADUREZ DEL FRUTO DE TOMATE

A) Expresión diferencial de proteínas en cruzamientos dialélicos:

El tomate (*Solanum lycopersicum*) es un fruto climatérico cuya maduración se caracteriza por cambios secuenciales en la expresión de proteínas, dando como resultado diferentes perfiles de bandas polipeptídicas en cada estado de madurez (Giovannoni, 2004).

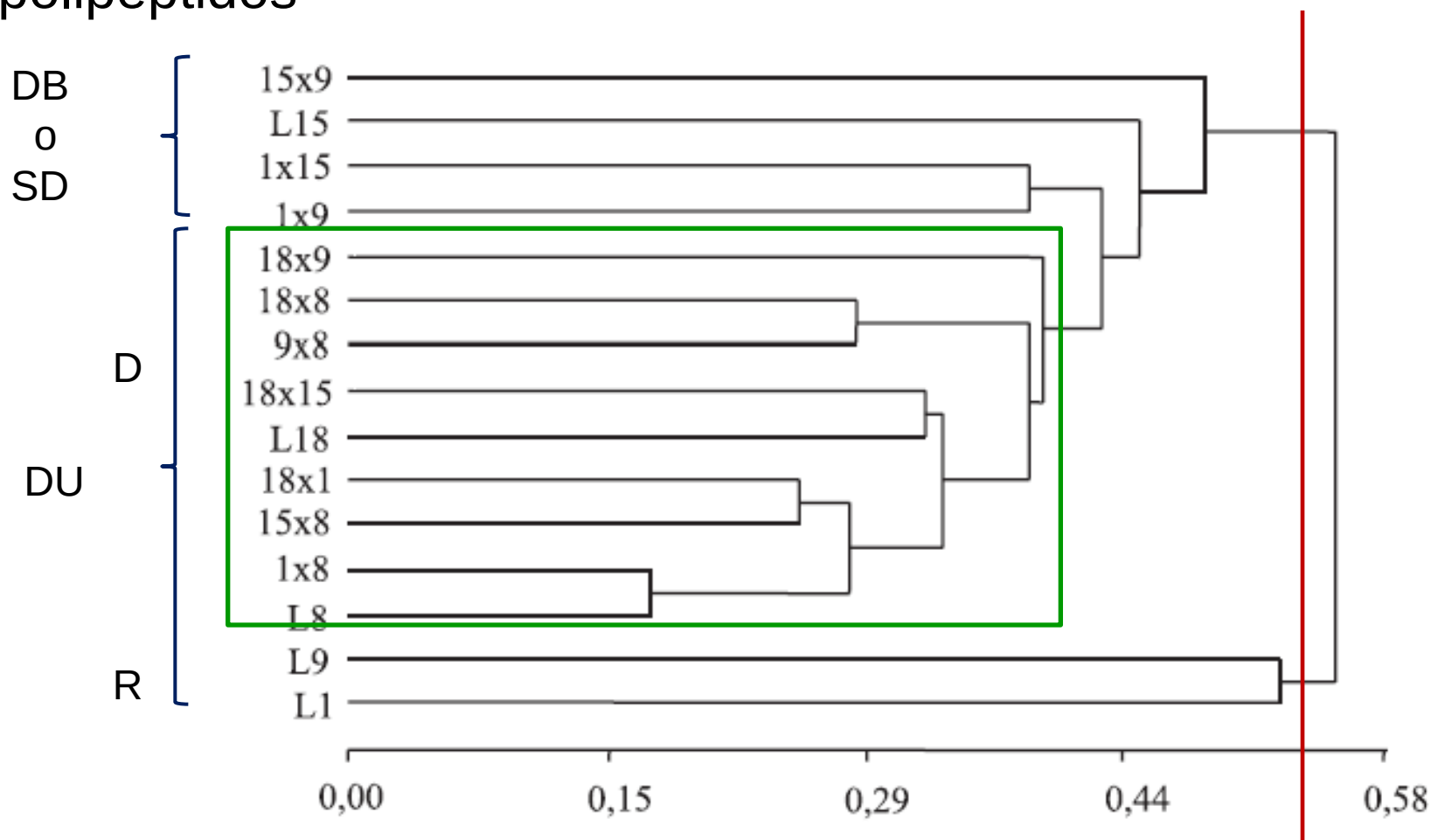
Sin embargo, frutos de diversos genotipos de tomate varían en su madurez (Rodríguez *et al.*, 2008).

Por lo tanto, la maduración del fruto de tomate es un proceso biológico afectado por fuentes multidimensionales de variación: el estado de madurez, el genotipo y la expresión de proteínas.

Análisis descriptivo: Análisis de Correspondencias



## B) Análisis inferencial: enfoque dialéctico de la expresión de polipéptidos



Análisis de agrupamiento de los genotipos de tomate (RILs, representadas por L, e híbridos, sólo por números), del diseño de cruzamientos dialécticos por perfiles polipeptídicos, en cuatro estados de madurez. Métrica utilizada: distancia de Jaccard .

Gardner y Eberhart (1966) , considerando efectos fijos, propusieron el siguiente modelo:

$$Y_{ii'} = \mu + \frac{1}{2}(p_i + p_{i'}) + \theta(H_m + h_i + h_{i'} + s_{ii'}) + e_{ii'j}$$

donde  $\mu$  es la media de los padres (líneas puras),  $p_i$  es el efecto del padre  $i$ ,  $\theta = 0$  para padres y 1 para los híbridos,  $H_m$  es la heterosis media,  $h_i$  es la heterosis del padre  $i$  (heterosis varietal),  $s_{ii'}$  es la heterosis específica y  $e_{ii'j}$  es el error experimental.

$H_m$ ,  $h_i$  y  $s_{ii'}$  son los componentes de la heterosis respecto al promedio entre los padres ( $H$ ), que se computa como

$$H = H_m + h_i + h_{i'} + s_{ii'}$$

Permutational Manova (PM) es un Análisis multivariado permutacional de la variancia que utiliza matrices de distancia. Es un método no paramétrico, basado en test de permutación con estadísticas pseudo-F. Es una alternativa al análisis multivariado de la variancia tradicional (MANOVA) que es demasiado estricto en sus supuestos para la mayoría de los conjuntos de datos multivariados ecológicos y genéticos.

Las aplicaciones típicas de este método incluyen el análisis de datos ecológicos o datos genéticos donde podríamos tener un número limitado de muestras de individuos y miles o millones de columnas de datos de expresión génica.

PM es además una alternativa a AMOVA (análisis molecular de la variancia).

Para la expresión de polipéptidos, se detectaron diferencias significativas entre padres, entre padres vs. híbridos, entre híbridos de un mismo padre vs. el resto y entre híbridos sin tener en cuenta su origen.

Interpretación: se encontraron efectos varietales y de heterosis (media, varietal y específica) para la expresión de polipéptidos (variable dictómica), que permiten explicar lo observado en el dendrograma.

Además, incorporando al peso y la vida poscosecha de los frutos como covariables del PERMANOVA, peso no tuvo efectos sobre los perfiles de polipéptidos, pero vida poscosecha mostró un efecto significativo para perfiles de polipéptidos en VM, interaccionando con la fuente de variación Genotipo. Esta interacción sugiere que las diferencias en la vida poscosecha del fruto entre los genotipos homocigotas y heterocigotas podrían estar asociados con la presencia / ausencia de al menos algunas bandas de polipéptidos en la etapa temprana de la madurez.

## C) Estimación de la heredabilidad multivariada para caracteres fenotípicos asociados a la madurez del fruto de tomate

### ❑ Objetivos:

Conocer y cuantificar la variabilidad total de las generaciones  $F_2$  y  $F_3$  para caracteres de calidad de fruto en un programa de mejoramiento de tomate a partir de Híbrido de Segundo Ciclo

Caracterizar los genotipos seleccionados de acuerdo a sus valores para los atributos evaluados en cada generación.

Estimar la heredabilidad multivariada a través de Análisis Factorial Múltiple de acuerdo a la caracterización global de ambas generaciones (análisis inferencial y variables cuantitativas).



ToUNR18



ToUNR1



F1 (HSC)

ToUNR18xToUNR1



F2

Selección 18 individuos



F3

## Esquema del programa de mejoramiento

18 individuos **seleccionados** de la generación F<sub>2</sub> del HSC RIL18 x RIL1

18 familias F<sub>3</sub> obtenidas por su autofecundación, de los 18 individuos de la F<sub>2</sub>



# CÁLCULO DE HEREDABILIDAD

## REGRESIÓN PROGENIE-PROGENITOR (RPP)

Estos valores fueron tomados como referencia para comparar resultados con el método multivariado

| Caracteres       | Heredabilidad | EE   |
|------------------|---------------|------|
| Peso             | 0,33          | 0,11 |
| Altura           | 0,46          | 0,09 |
| Diámetro         | 0,41          | 0,16 |
| Forma            | 0,43          | 0,12 |
| pH               | 0,11          | 0,06 |
| Vida Poscosecha  | 0,00          | 0,00 |
| L                | 0,00          | 0,00 |
| a/b              | 0,00          | 0,00 |
| Sólidos solubles | 0,00          | 0,00 |
| Firmeza          | 0,00          | 0,00 |

**Coeficientes de regresión:  
no significativos**

Resultados presentados en **Congreso SBR 2015** (“Estimación de componentes de variancia genética en un híbrido de segundo ciclo de tomate (*Solanum lycopersicum*)”) y en **Congreso SAG 2015** (“Diseño de estrategias de mejoramiento en un híbrido de segundo ciclo de tomate”)

## Estimación Multivariada: ANÁLISIS FACTORIAL MULTIPLE (AFM)

Encontrar una estructura común o representativa de todas las  
tablas

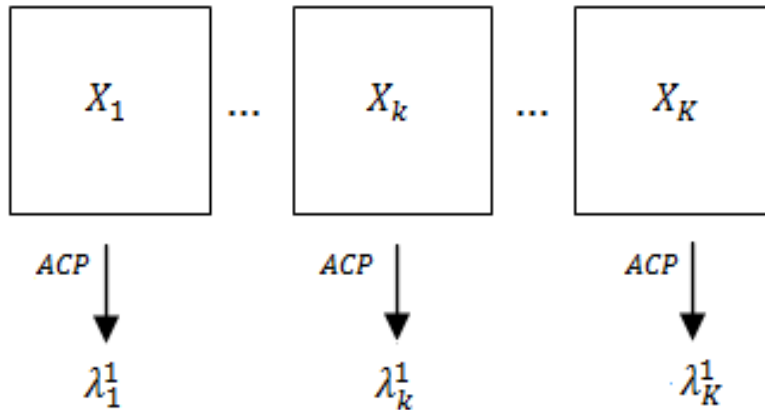
☐ Estudiar las relaciones entre los individuos o elementos.

☐ Estudiar las relaciones entre variables.

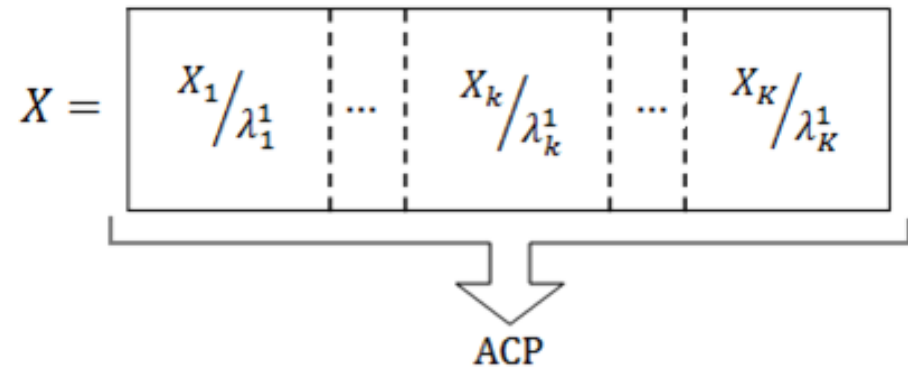
☐ Encontrar una estructura común respecto a los grupos, que ponga en evidencia las especificidades de cada grupo.

Consta de dos etapas:

**Etapla preliminar del AFM**



**Etapla principal del AFM**

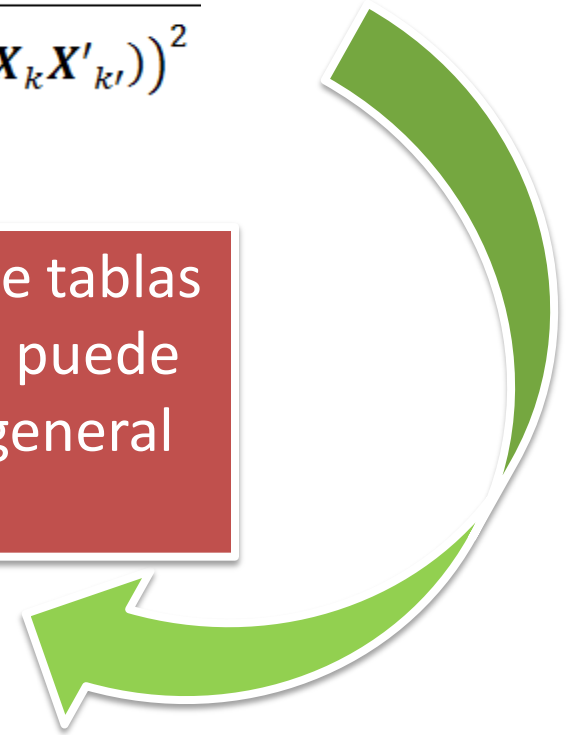


- ✓ Se mantiene la estructura de cada matriz
- ✓ Se logra que ninguna matriz tenga un peso preponderante

➤ **RV**

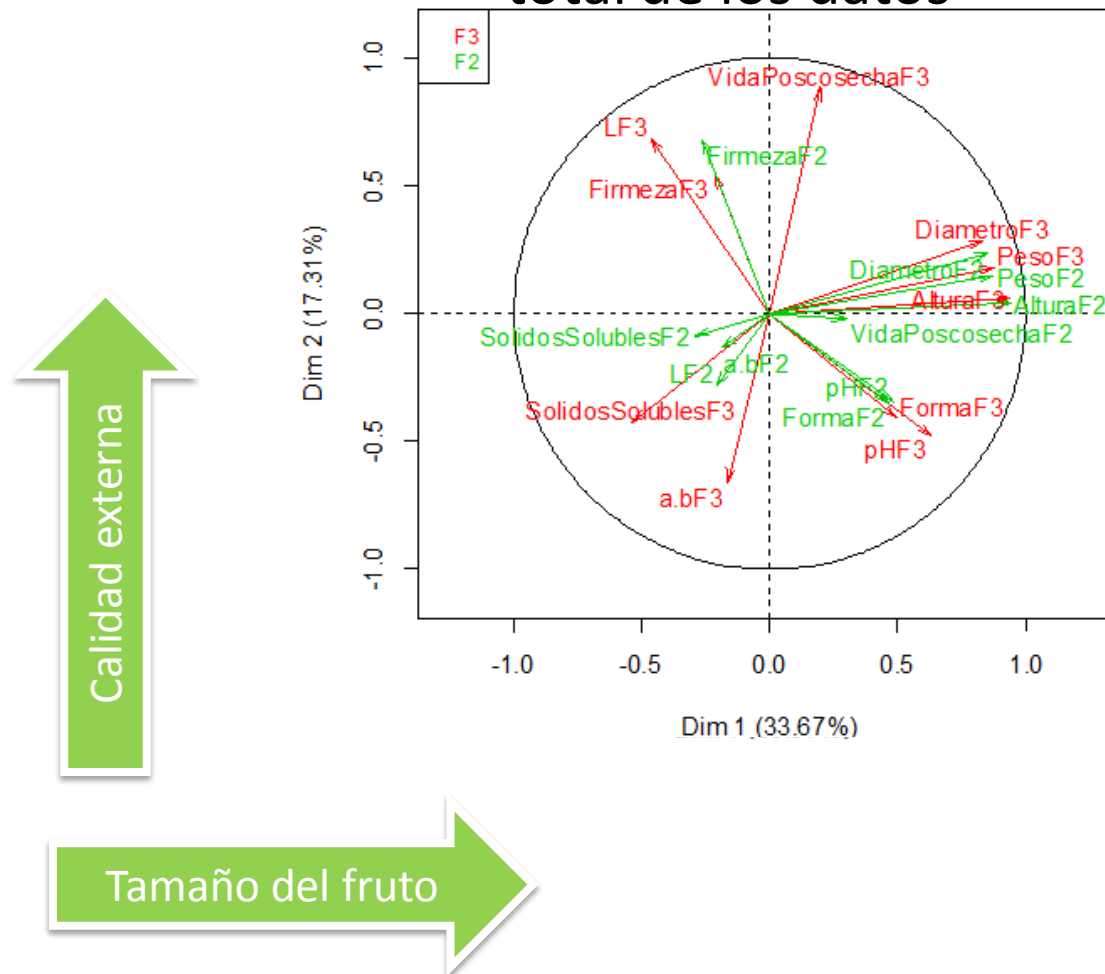

$$RV(\mathbf{X}_k, \mathbf{X}_{k'}) = \frac{\text{tr}[(\mathbf{X}_k \mathbf{X}'_{k'}) (\mathbf{X}_k \mathbf{X}'_{k'})]}{\sqrt{(\text{tr}(\mathbf{X}_k \mathbf{X}'_{k'}))^2 (\text{tr}(\mathbf{X}_k \mathbf{X}'_{k'}))^2}}$$

Permite medir la semejanza entre tablas o matrices de datos. Con lo cual puede asociarse con la heredabilidad general del conjunto de datos.



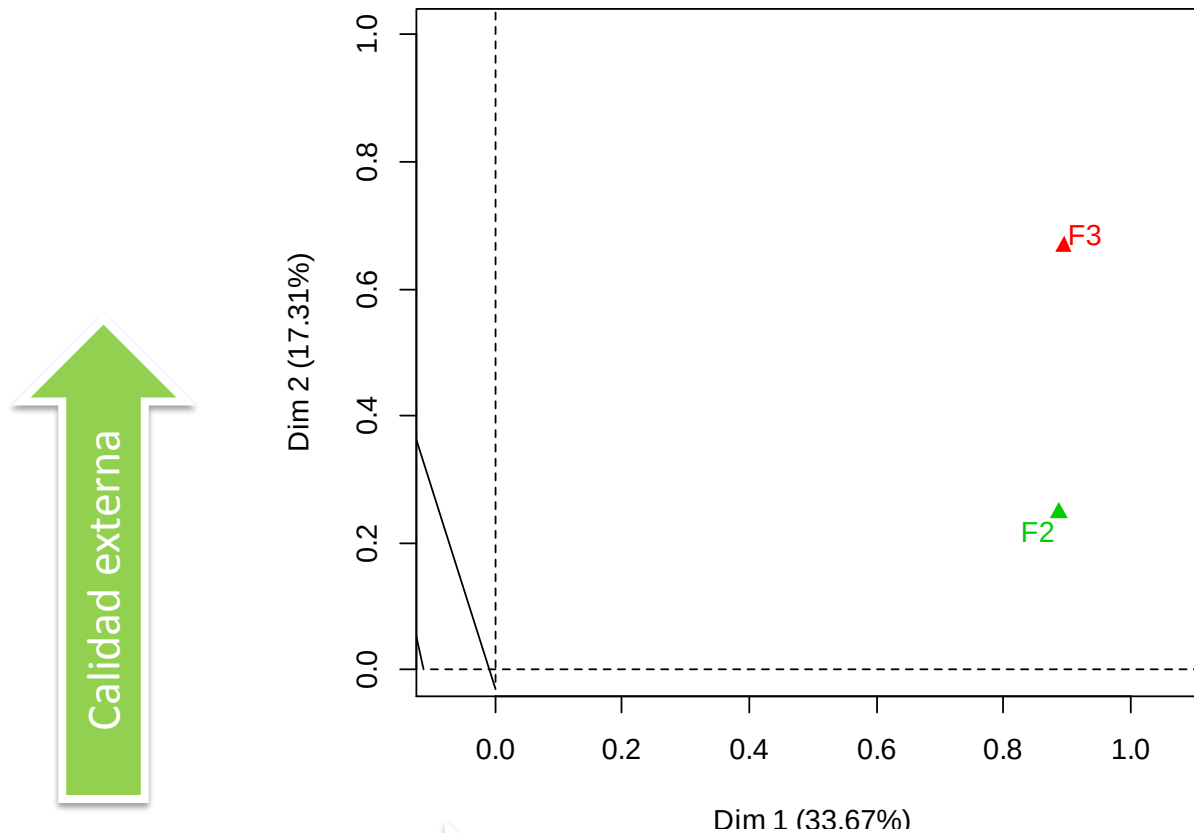
## ANÁLISIS FACTORIAL MULTIPLE

Los dos primeros ejes representan el 51% de la variabilidad total de los datos



# ANÁLISIS FACTORIAL MULTIPLE

Representación de las generaciones sobre los dos primeros ejes globales



Calidad externa

Tamaño del fruto

➤ **RV = 0,503**

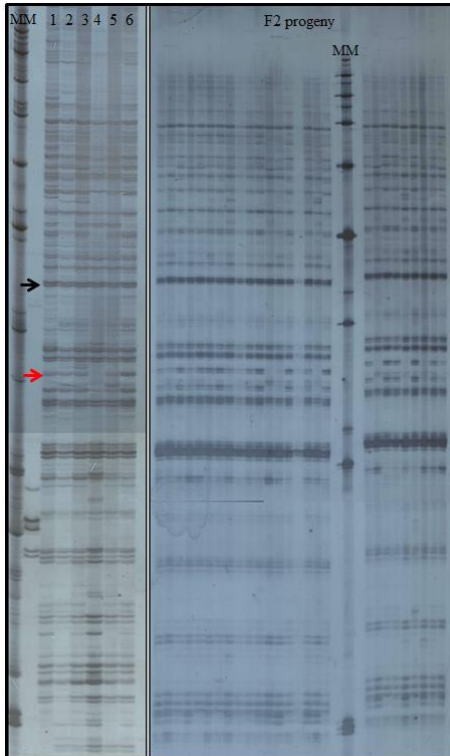
## D) Consenso entre caracterización molecular y caracterización fenotípica de atributos de calidad de fruto en tomate

1. Caracterizar la diversidad genética en las generaciones segregantes  $F_2$  de los tres híbridos de segundo ciclo (HSC RIL18xRIL1, RIL1xRIL8 y RIL1xRIL5) mediante seis combinaciones de marcadores moleculares AFLP

6 combinaciones de cebadores de AFLP

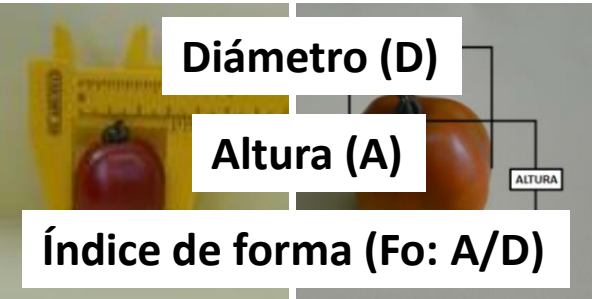
- ✓ elevados número de bandas
- ✓ alto porcentaje de polimorfismo

(Cachiarelli *et al.*, 2015)



- total de bandas
- número de bandas polimórficas
- porcentaje de polimorfismo
- número de bandas que ajustaron a la segregación mendeliana esperada
- análisis de conglomerados

## 2. Evaluar la variación para 11 caracteres cuantitativos relacionados a la calidad de fruto en estas tres generaciones segregantes $F_2$



Diámetro (D)

Altura (A)

Índice de forma (Fo:  $A/D$ )



Peso (P)



Vida poscosecha (VP)



Firmeza (F)



Color

Porcentaje de reflectancia (L)  
Cociente de absorbancias (a/b)



pH



Acidez titulable (AT)



Sólidos solubles (SS)



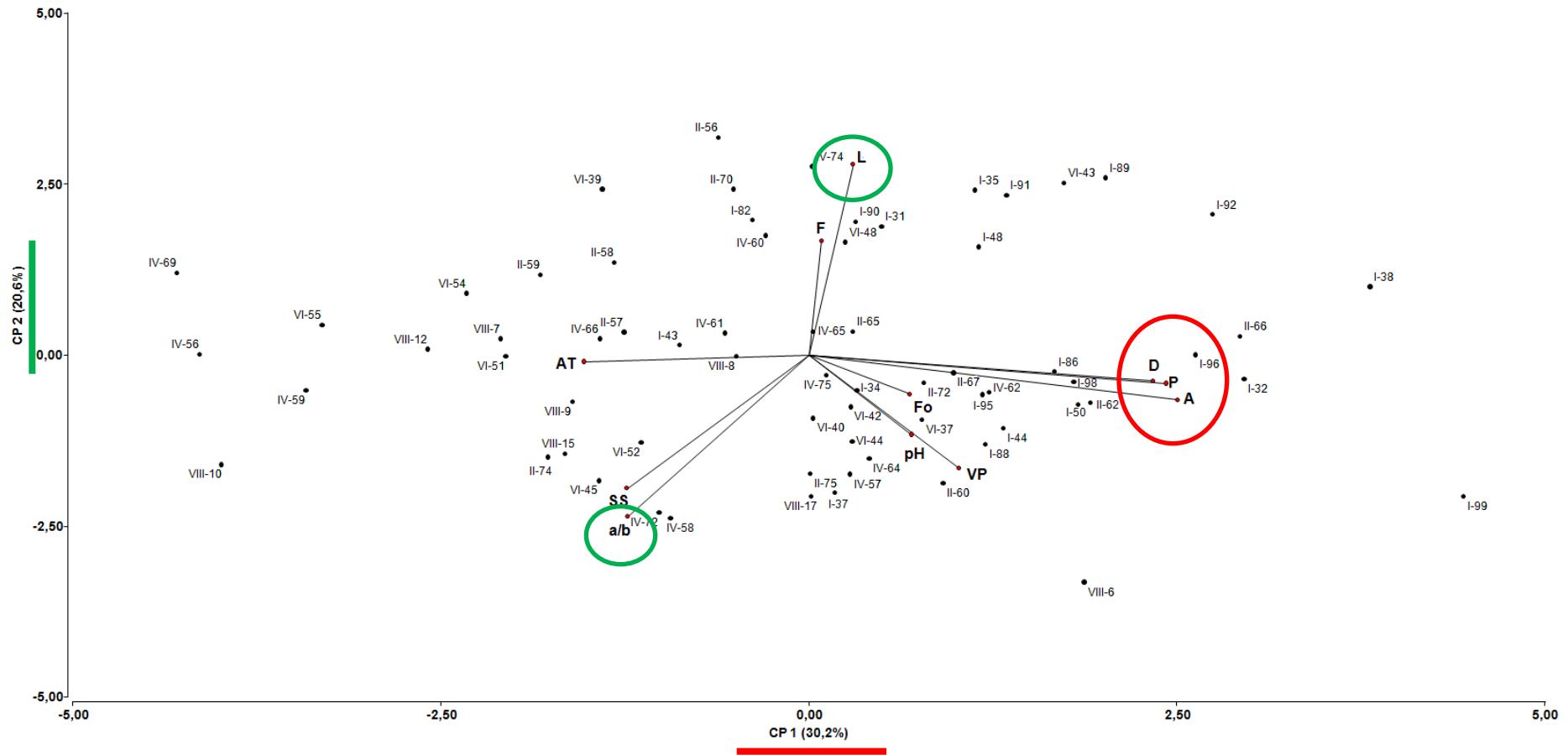
**3. Explorar el consenso entre la diversidad molecular y la variación para caracteres cuantitativos mediante Análisis de Procrustes Generalizado (APG)**

**Análisis de Componentes Principales (ACP):** 11 variables cuantitativas (fenotípicas).

**Análisis de Coordenadas Principales (ACoP):** AFLP / bandas con segregación mendeliana.

**Análisis de Procrustes Generalizado (APG):** evaluar la asociación/consenso entre la información fenotípica y molecular.

## Análisis de Componentes Principales



**CP1: 30%; CP2: 21%; CP3: 16%; CP4: 11% (4CP: 78%)**

# Análisis de Coordenadas Principales

**ACoP: 12 coordenadas para explicar el 79%.**

## Análisis de Procrustes Generalizado

**APG: 7 componentes para explicar el 79%.**



**Consenso:**  
**67%**

En todos los casos, la aplicación de análisis de reducción de la dimensionalidad permitió dilucidar las bases genéticas subyacentes a los procesos bajo estudio.

Muchas gracias!