Trabajo Final

CARACTERIZACIÓN DE LA DIVERSIDAD GENÉTICA Y ESTRUCTURACIÓN POBLACIONAL DE UN PANEL PÚBLICO DE LÍNEAS ENDOCRIADAS DE MAÍZ

Alumna: Sofía Inés Perdomo

DNI: 37.756.946

Director: Agustín Baricalla

Co-Directora: María Laura Federico

Carrera: Licenciatura en Genética

Universidad Nacional del Noroeste de la Provincia de Buenos Aires

Escuela de Ciencias Agrarias, Naturales y Ambientales

Pergamino, 28 de noviembre de 2023

1. Introducción

La mayoría de las especies vegetales fueron originadas durante un corto período de la historia humana entre 5.000 y 10.000 años atrás. Durante este tiempo, varios cultivos fueron domesticados independientemente, dentro de los cuales se encuentra el maíz (*Zea mays*) (Matsuoka *et al*., 2002). Se considera que esta domesticación tuvo sus orígenes a partir de su probable progenitor silvestre, el teosinte, proporcionando uno de los ejemplos más sorprendentes y complejos de evolución morfológica en plantas (Doebley *et al*., 1995).

Genéticamente el maíz es una especie monoica, de reproducción cruzada, diploide, con una constitución cromosómica 2n=20 (Acquaah, 2012). La inflorescencia masculina se ubica en la zona apical del tallo y se denomina panoja, mientras que la inflorescencia femenina se encuentra en posición axial y toma el nombre de espiga.

Considerando que es un cultivo que, a lo largo de los años se ha ido adaptando a una diversidad de ambientes, y también que es tolerante a una amplia gama de estreses bióticos y abióticos, su producción a nivel nacional y mundial ha ido en auge (Bevilacqua y Storti, 2019). En Argentina, alrededor del 80% de la producción de maíz se concentra en el norte de la provincia de Buenos Aires, el sudeste de Córdoba y el sur de Santa Fe, zona conocida tradicionalmente como “Zona Núcleo Maicera” (Bevilacqua y Storti, 2019). En los últimos 20 años, la superficie sembrada con maíz en Argentina se ha más que duplicado. Durante la cosecha 2021/22, los valores de superficie cubierta de maíz fueron de 7,9 millones de hectáreas (Figura 1). A su vez, según lo reportado por el Departamento de Agricultura de Estados Unidos (USDA, del inglés *United States Department of Agriculture*) durante los años 2018-2023 Argentina se consagró como el tercer país oferente a nivel mundial, detrás de Estados Unidos, con exportaciones de más de 40 millones de toneladas (Sigaudo y Terré, 2022) (Figura 2). En base a estas estadísticas, se puede considerar que la tendencia de la superficie cultivada y de la producción de maíz, en este período (2018-2023), muestra una trayectoria creciente a nivel nacional, más allá de las fluctuaciones propias de la variabilidad climática (Bevilacqua y Storti, 2019).

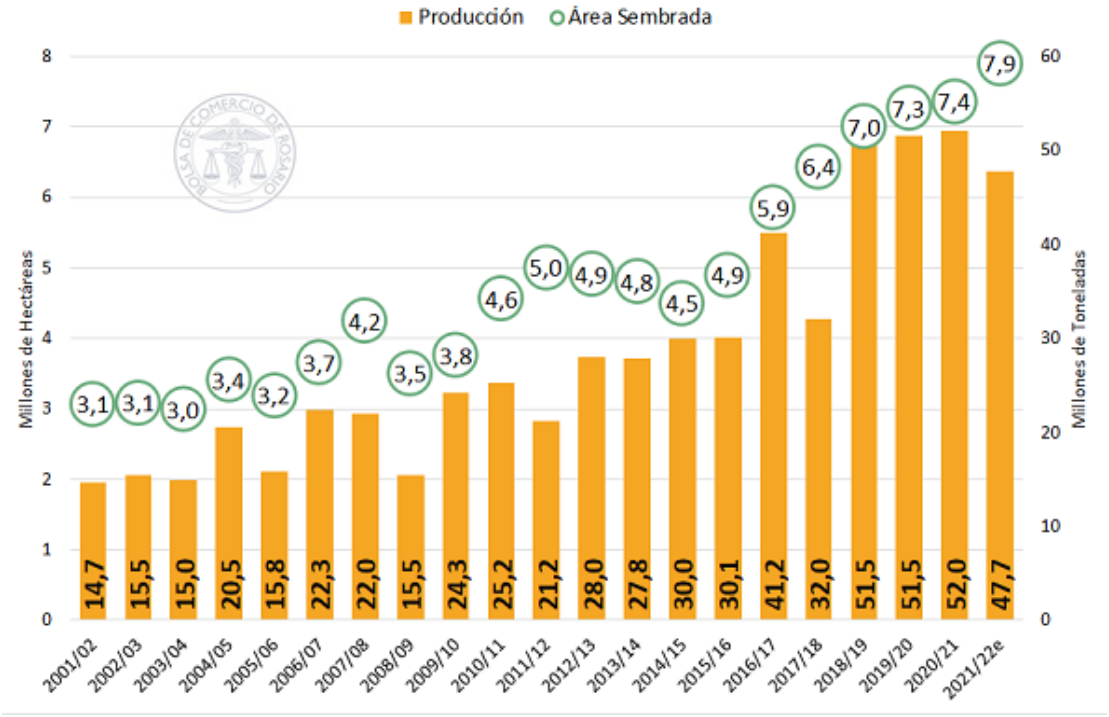


Figura 1: Gráfico de la producción y área sembrada de maíz en Argentina para el período 2001 a 2022 (Sigaudo y Terré, 2022).

Figura 2: Gráfico de la cantidad, en millones de toneladas, de maíz exportado durante el periodo 2018-2023 por los principales países exportadores a nivel mundial (Maizar, 2023).

1.1 Estudio de diversidad genética

Un punto de partida importante para obtener un buen progreso en cualquier programa de mejora de cultivos es la información sobre la diversidad genética y la estructura poblacional del germoplasma base. Debido a esto, muchos investigadores han resaltado la importancia y la necesidad de una evaluación precisa de la diversidad genética en los programas de mejora (Adu *et al*., 2019). Los conocimientos e información sobre la caracterización molecular de la diversidad de la población, su estructura y las relaciones filogenéticas entre los materiales de mejoramiento, permiten comprender cómo utilizar el germoplasma escogido, seleccionar las líneas parentales, asignar grupos heteróticos y crear los grupos de germoplasma de maíz (Wu *et al*., 2016).

Con el advenimiento de las técnicas modernas de biología molecular han surgido diversos métodos de identificación de polimorfismos genéticos a nivel de ADN, lo que permite evaluar la diversidad de una especie a nivel molecular. En este sentido, los marcadores moleculares son preferidos sobre los marcadores morfológicos y bioquímicos no solo porque no se ven afectados por factores ambientales y/o por las etapas de desarrollo de la planta sino por su distribución a lo largo del genoma. Estos marcadores son una herramienta indispensable para caracterizar los recursos genéticos y proporcionar a los mejoradores información más detallada, para ayudar a la selección de los parentales e iniciar un programa de mejora (Adu *et al*., 2019). Los mismos, se caracterizan por poseer una ubicación específica en un cromosoma (punto de referencia) cuya herencia puede seguirse en individuos de una población. La secuencia puede pertenecer tanto a regiones codificantes como no codificantes (Abriata *et al*., 2010).

Se pueden destacar diferentes tipos de marcadores moleculares tales como: los polimorfismos de longitud de fragmentos de restricción (RFLP del inglés *Restriction fragment length polymorphism*), los polimorfismos de longitud de fragmentos amplificados (AFLP, del inglés *Amplified fragment length polymorphism*), las repeticiones de secuencia simple (SSR, por *simple sequence repeat*) o polimorfismos de un solo nucleótido (SNP, por *single nucleotide polymorphism*) (Abriata *et al*., 2010).

Destacando principalmente el caso de los SNPs o polimorfismos de un solo nucleótido, se puede resaltar su uso en la detección de polimorfismos resultantes de la alteración de una única base en una secuencia de ADN y su naturaleza bialélica. En la última década, se han constituido como el principal marcador elegido para estudios genéticos en plantas y otros organismos, debido a su bajo costo, alta abundancia genómica, potencial para análisis de alto rendimiento y bajas tasas de error de genotipado (Wu *et al*., 2016). Cabe

destacar que los SNPs no solo son codominantes, sino también estables desde un punto de vista evolutivo, lo que facilita su empleo en estudios poblacionales. Otro punto importante está relacionado con la genotipificación de los SNPs la cual no está basada en la medida del tamaño de los alelos, como ocurre con otros marcadores moleculares, y la distinción de los alelos puede ser automatizada (Abriata *et al*., 2010).

En la actualidad existen distintos métodos de validación y genotipificación de SNPs, siendo uno de los primeros en ser desarrollados los chips de ADN o microarreglos de ADN. Estos chips consisten en arreglos de oligonucleótidos de secuencia conocida, depositados sobre un soporte sólido. Para el caso de la detección de SNPs, los oligonucleótidos difieren entre sí en sitios específicos en nucleótidos individuales (en el sitio del SNP). La técnica es apropiada para analizar varios SNPs en paralelo a partir de cada muestra (en forma multiplex) (Gupta & Rustgi, 2004).

1.2 Estructura poblacional

Una población estructurada genéticamente es aquella en la que coexisten subgrupos o conglomerados de individuos que difieren sistemáticamente en sus frecuencias alélicas para diferentes *loci* (Peña, 2010). Para llevar a cabo el análisis de la estructura poblacional, es conveniente considerar que la estructura de las poblaciones consiste en dos partes distintas pero interrelacionadas: la estructura demográfica y la estructura genética (Slatkin, 2017). Aquello que en realidad caracteriza a la estructura genética de una población es su reservorio o conjunto de genes, es decir, las frecuencias génicas y genotípicas (Cabrero y Camacho, 2000).

Cabe considerar que el conocimiento de la estructura poblacional es de gran importancia en los programas de mejoramiento genético. Algunos de los objetivos en estos el desarrollo de nuevas líneas endogámicas con alta capacidad de combinación para producir mayores rendimientos de grano y un rendimiento agronómico superior en combinaciones híbridas (Ricci, 2007). A su vez, la producción de semillas disminuye debido al aumento de la endogamia (Morris & Isikan, 1964). Por este motivo, es conveniente calcular los coeficientes de endogamia a través de análisis de variancia molecular (AMOVA) y así, estimar los índices de fijación (Fis, F_{st} y Fit), que describen la diferenciación genética a nivel intrapoblacional (Fis), entre poblaciones (F_{st}) y total (Fit) (Suarez-Salgado et al., 2016). A partir del análisis con SNPs y en base a estos cálculos se puede comparar la diversidad genética y el nivel de endogamia entre las líneas evaluadas.

1.3 Métodos estadísticos para inferir estructuración

La estructura poblacional puede ser determinada a través de diversos enfoques, dentro de los cuales se encuentran aquellos que permiten realizar árboles de distancia, y los métodos que están fundados en modelos de agrupamientos jerárquicos permitiendo inferir el número de grupos o subpoblaciones presentes en una población diversa, así como también establecer la frecuencia alélica en los diferentes sub-grupos.

El armado de los árboles filogenéticos puede llevarse a cabo a partir de métodos de distancia, máxima verosimilitud o de inferencia bayesiana. En este sentido, el enfoque bayesiano permite el análisis de grandes conjuntos de datos logrando obtener un análisis de los datos más minucioso (Lunter *et al*., 2005). Por otra parte, los métodos basados en la distancia suelen ser más rápidos y fáciles de aplicar y son los más adecuados para el análisis exploratorio de datos (Falush *et al*., 2007). En el caso del método de máxima verosimilitud, permite obtener estimaciones de máxima probabilidad de tiempos de divergencia poblacional. Este método proporciona una estimación única de la filogenia de la población teniendo en cuenta tanto la mutación como la deriva (Nielsen, 1998).

1.4 La importancia de la estructura poblacional en el mejoramiento genético

Considerando que en un programa de cruzamiento se utilizan líneas puras de diferentes grupos heteróticos (Shu *et al*., 2021), el conocimiento sobre la diversidad molecular, la pureza genética, el agrupamiento heterótico y las relaciones genéticas entre líneas puras desarrolladas son importantes para identificar las combinaciones parentales correctas para iniciar nuevos cruces, ya sea para el desarrollo de líneas puras o de híbridos (Kumar *et al*., 2022).

Los marcadores del tipo SNPs permiten a los investigadores obtener resultados rápidos a relativo bajo costo (Kumar *et al*., 2022). Por este motivo, permiten genotipar rápidamente el germoplasma de líneas de maíz de interés para diferentes propósitos, incluida la evaluación de la diversidad genética, el mapeo de genes y *loci* de rasgos cuantitativos (QTL) utilizando poblaciones biparentales y paneles de mapeo de asociación de todo el genoma (GWAS). También permiten el desarrollo de variedades mejoradas utilizando selección genómica y asistida por marcadores (Kumar et al., 2022).

En un contexto de mapeo de asociación, el LD (del inglés *Linkage desequilibrium,* o desequilibrio de ligamiento) no solo puede estar influenciado por la recombinación sino también por otras fuerzas. Estas fuerzas que influyen en el patrón y la extensión de LD

son: (i) tipo de apareamiento, (ii) deriva genética, (iii) mutación, (iv) selección, (v) subestructura y relación de la población (Stich y Melchinger, 2010). Estos dos últimos puntos tienen mayor relevancia ya que aquí se destaca la importancia de la caracterización de la diversidad genética y la diferenciación de poblaciones para las líneas puras de maíz de los programas de mejoramiento y su valor para ayudar a los mejoradores a mantener y potencialmente aumentar la tasa de ganancia genética (De Faria, *et al*., 2022).

1.5 Reducción dimensional

La reducción dimensional en el ámbito de la genética es una técnica cuyo objetivo principal consiste en la conservación de la estructura poblacional presente en el set de datos, de manera significativa, en un mapa de menores dimensiones con el fin de lograr demostrar la presencia de agrupamientos en diferentes escalas (Van der Maaten y Hinton, 2008).

Los análisis de reducción dimensional de grandes volúmenes de datos mediante PCA (Análisis de componentes principales) datan de más de una década (Price *et al*., 2006). Éstos fueron ampliamente utilizados, así como también criticados (Elhaik, 2022). Según el estudio realizado por Elhaik (2002), el método de PCA no produjo resultados consistentes en todos los esquemas de diseños planteados. A su vez, las distancias entre los datos se encuentran sesgadas, sin definir con certeza poblaciones cercanas o distantes. A su vez, el método de PCA no puede capturar bien las características de los datos y presenta inconvenientes en la determinación del número de SNPs que se incluirán en el estudio (Li *et. al*, 2017). Por este motivo, proponemos un método que logre sustituir al método de PCA. Este análisis se basa en la reducción dimensional mediante el método t-SNE (Van der Maaten y Hinton, 2008), que pretende ser evaluado en el marco de este trabajo. Como cita el trabajo de Li y colaboradores (2017): "*El método de distribución de vecino estocástico que incorpora t-SNE es una nueva técnica de visualización y de reducción de dimensiones para datos de alta dimensión. t-SNE rara vez se aplicó a datos genéticos humanos, aunque se usa comúnmente en otros campos biológicos que requieren muchos datos, como la genómica unicelular*". Debido a que existe escasa información disponible o reportada sobre la utilización de t-SNE para datos de líneas endocriadas sometidas a presión de selección, aquí probamos el uso de este método con este set de datos.

2. Hipótesis

La diversidad genética presente en líneas endocriadas de un panel público de maíz puede ser investigada con marcadores moleculares. La comparación entre enfoques bayesianos, métodos de distancia y reducción dimensional revelará patrones convergentes que ayudarán a entender la estructuración genética de las subpoblaciones presentes en el panel.

3. Objetivos

3.1 Objetivo general

Analizar la diversidad genética y la estructuración poblacional existente en un panel público de líneas endocriadas de maíz a partir de datos de SNPs, y comparar los resultados obtenidos a partir de las diferentes herramientas computacionales utilizadas en el proceso.

3.2 Objetivos específicos

- Inferir el número de subpoblaciones presentes en este conjunto de datos.
- Agrupar las líneas y determinar su parentesco mediante árboles de distancia.
- Calcular los parámetros poblacionales de distancia genética y el índice de fijación entre las subpoblaciones obtenidas (F_{st}).
- Comparar la estructuración obtenida del panel de líneas evaluado, mediante 2 enfoques bayesianos alternativos.
- Comparar la estructuración, del panel de líneas evaluado, obtenida mediante los métodos bayesianos, métodos de distancia y de reducción dimensional.

4. Materiales y métodos

4.1 Material genético

La realización de este estudio se llevará a cabo mediante la utilización de un panel de 282 líneas endocriadas de maíz de origen público, previamente genotipadas (Cook *et al*., 2012). La información genotípica fue obtenida a partir del chip MaizeSNP50 (Illumina, USA) y los polimorfismos descargados desde la página web https://www.panzea.org/data. Este chip consta de datos de 56.658 *loci* de SNPs homogéneamente distribuidos a lo largo del genoma. Este set de líneas incluye familias de líneas que presentan elevada endocría y a B73 como línea de referencia para determinar la ubicación física de los SNPs.

4.2 Filtrado de los datos

Se procesaron los datos siguiendo los lineamientos previos de Adu y colaboradores (2019) eliminando los marcadores y las líneas con más del 20% de datos faltantes, 20% de heterocigosidad y se filtraron las líneas que son discordantes con el resto del panel. Aquellos marcadores con frecuencia alélica menor al 5% fueron descartados.

4.3 Análisis de datos

4.3.1 Algoritmos bayesianos para inferir estructuración

La inferencia de la estructura poblacional se realizó mediante metodología bayesiana. Se utilizaron los programas STRUCTURE 2.3.4 (Pritchard *et al*. 2000; Falush *et al*. 2003), en primer lugar, a partir de un servidor de alta efectividad, y luego fastSTRUCTURE (Raj *et al.,* 2014), mediante el servidor CyVerse (Merchant *et al*., 2016). Este da un rango de valores para la complejidad del modelo requerido para explicar la estructura subyacente a los datos (Pritchard *et al*. 2000). El programa proporciona una forma de evaluar y visualizar valores de probabilidad en múltiples valores de K (número de agrupamientos formados), los cuales definen el número de las posibles subpoblaciones presentes en los datos estudiados, y realiza miles de iteraciones para una detección de la cantidad de grupos o subpoblaciones que mejor se ajustan a los datos (Earl y vonHoldt, 2012). Este programa procesa los resultados estructurales y genera archivos para su uso posterior en STRUCTURE Harvester (Earl y vonHoldt, 2012) (https://taylor0.biology.ucla.edu/structureHarvester/) que ejecuta el método "Evanno" (Evanno *et al*., 2005) para determinar el valor de K más verosímil.

En el caso de fastSTRUCTURE, se puede definir como una herramienta útil para lograr una optimización de los tiempos de ejecución informática con respecto a STRUCTURE. A su vez, fastSTRUCTURE incluye internamente el algoritmo “chooseK” que devuelve entre 11 y 15 subpoblaciones.

4.3.2 Análisis de distancia

En este paso, se realizaron árboles basados en distancia a partir de la obtención de las matrices de distancia obtenidas mediante el programa TASSEL (*Trait Analysis by aSSociation, Evolution and Linkage)* (Bradbury *et al*., 2007), mediante el método *Unweighted Pair Group Method with Arithmetic mean* (UPGMA). Posteriormente, para la visualización de los gráficos obtenidos, se utilizó ITOL (*Interactive Tree Of* Life) (Letunic & Bork, 2016).

En este contexto también se utilizó el método hclust (*Hierarchical Cluster Analysis*) el cual es un método alternativo, para obtener una aproximación de los clusters existentes en los datos estudiados en función de los valores de K obtenidos. Estos se ejecutaron en R (R *Core Team*, 2020) utilizando los paquetes AGNES (*Agglomerative clustering,* https://stat.ethz.ch/R-manual/R-devel/library/cluster/htlm/agneas.object.html*)* y DIANA (*Divise hierarchichal clustering*, https://uc-r.github.io/hc_clustering).

4.3.3 Reducción dimensional

El análisis de reducción dimensional se realizó mediante el método llamado t-SNE (Van der Maaten y Hinton, 2008), que permite la reducción y posterior visualización de los datos génicos en un espacio de dos o tres dimensiones, significativamente menor que el uso de los miles de dimensiones del set original (en el orden de las decenas de miles), donde cada dimensión representa cada estado discreto de los SNPs analizados (alelos).

4.3.5 Método k-means

Para probar corroborar los agrupamientos de los genotipos estudiados con otro método diferente a los anteriores, se realizó un análisis mediante el método k-means el cual es muy utilizado para la agrupación geométrica (Pelleg, *et al*., 1999). Permite dividir los n puntos que se encuentran en el espacio en k número de grupos de una manera rápida y sencilla asignando cada punto de datos a su centro más cercano (Vattani, 2011).

4.3.6 Metodo dbSCAN

El análisis mediante dbSCAN permitió agrupar a los datos considerando la densidad de los mismos en el espacio y formando grupos basados en este algoritmo no paramétrico basado en la densidad (Cambello, *et al.,* 2013).

4.3.7 Estructura poblacional

En este caso se determinaron los valores de Fst para cada subpoblación. Estos valores fueron dados de STRUCTURE al realizar el análisis con este algoritmo.

4.4 Comparación de métodos

Se realizó un análisis comparativo de los grupos obtenidos por los tres métodos descriptos, enfoques bayesianos, métodos de distancia y reducción dimensional para dilucidar cuál de ellos fue el más apropiado para obtener resultados sobre estructuración con este set de datos.

5. Resultados

El panel de las 282 líneas de maíz, correspondientes a un panel público previamente genotipadas por Cook *et al*. 2012, que finalmente pasaron todos los filtros realizados, quedando 275 genotipos para el estudio, fue genotipado con 48.898 SNPs presentes en el Illumina Beadchip MaizeSNP50 (USA). Estos SNPs (Illumina Beadchip MaizeSNP50) se encuentran distribuidos homogéneamente a lo largo del genoma en los 10 cromosomas.

5.1 Filtrado y medidas de resumen de los datos

Los datos que pasaron todos los filtros descriptos en el punto 4.2, representan finalmente 275 líneas y 48.898 loci cantidad de SNPs. En la Figura 3 se observa un diagrama del flujo de trabajo.

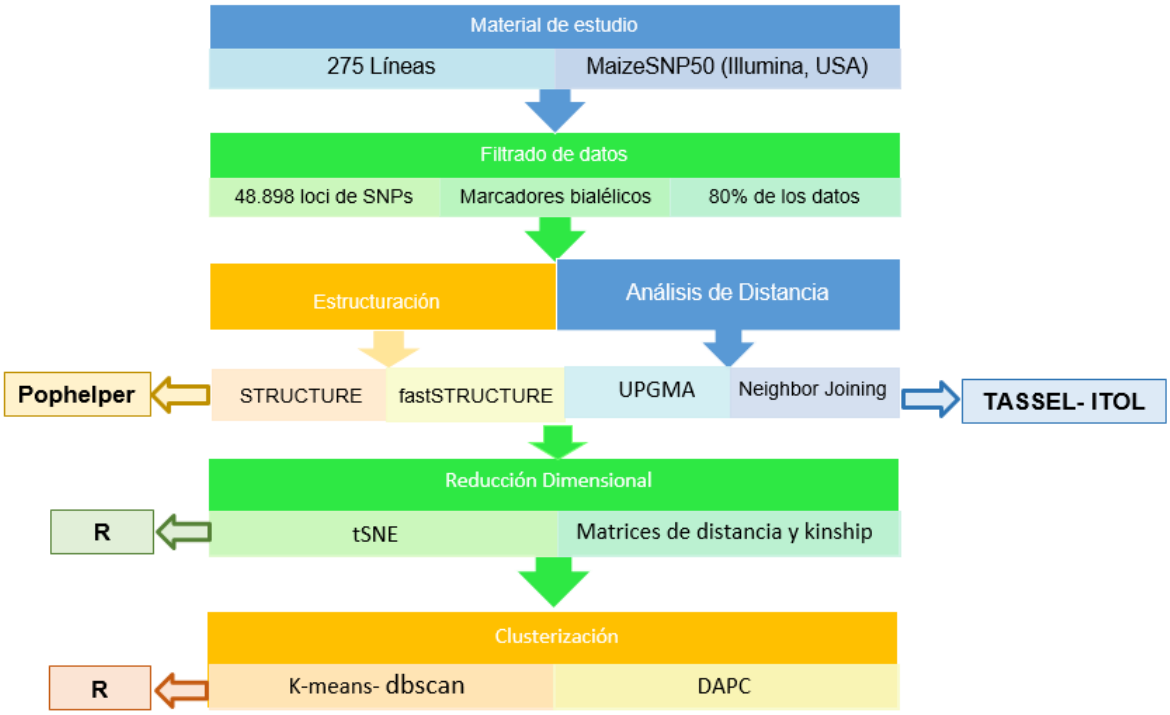


Figura 3. Diagrama de flujo de trabajo

5.2 Inferencia de la estructuración

Para inferir la estructuración a partir de los datos de SNPs se realizaron 3 réplicas en STRUCTURE para los valores de subpoblaciones desde k=1 hasta k=10. Luego se llevó a cabo el método de Evanno, a partir del cual se obtuvo la cifra de delta K óptimo con un valor igual a 40,842 como se observa en la Tabla 1. Este valor da información sobre la presencia de 7 subpoblaciones (Figura 4) distinguibles en el set de datos. Esto permitió inferir la presencia de 7 grupos definidos, aunque algunos más estructurados que otros. Estos grupos presentaron valores promedio de Fst superiores a 0,6 como se observa en la Tabla 2 y la Figura 5.

Tabla 1: Cálculo de los valores de Delta K a partir de los valores de verosimilitud [LnP(K)] para diferente número de subpoblaciones (K) mediante el método de Evanno.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	3	-14194335,867	116,196	NA	NA	NA
2	3	-13524881,767	562,628	669454,100	200492,800	356,351
3	3	-13055920,467	874,077	468961,300	241674,600	276,491
4	3	-12828633,767	1380,591	227286,700	32711,367	23,694
5	3	-12634058,433	98446,345	194575,333	61287,267	0.622
6	3	-12378195,833	60468,449	255862,600	3361,633	0.056
7	3	-12118971,600	24288,876	259224,233	992013,133	**40,842**
8	3	-12851760,500	1491598,622	-732788,900	1775420,133	1,190
9	3	-11809129,267	64707,073	1042631,233	915814,033	14,153
10	3	-11682312,067	20038,751	126817,200	NA	NA

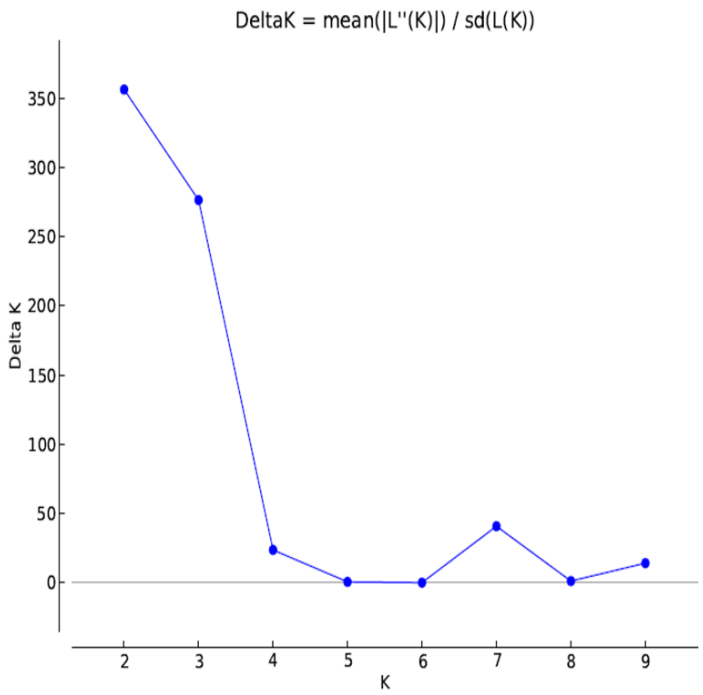


Fig. 4. Representación gráfica de los valores de Delta K mediante el método de Evanno et. al (2005) de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.

Tabla 2: Valores promedio de Fst de cada subpoblación, obtenidos a partir de las 3 corridas realizadas por STRUCTURE de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.

Subpoblaciones STRUCTURE	Fst
1	0,719
2	0,432
3	0,524
4	0,495
5	0,569
6	0,733
7	0,514
Total	**0,569**

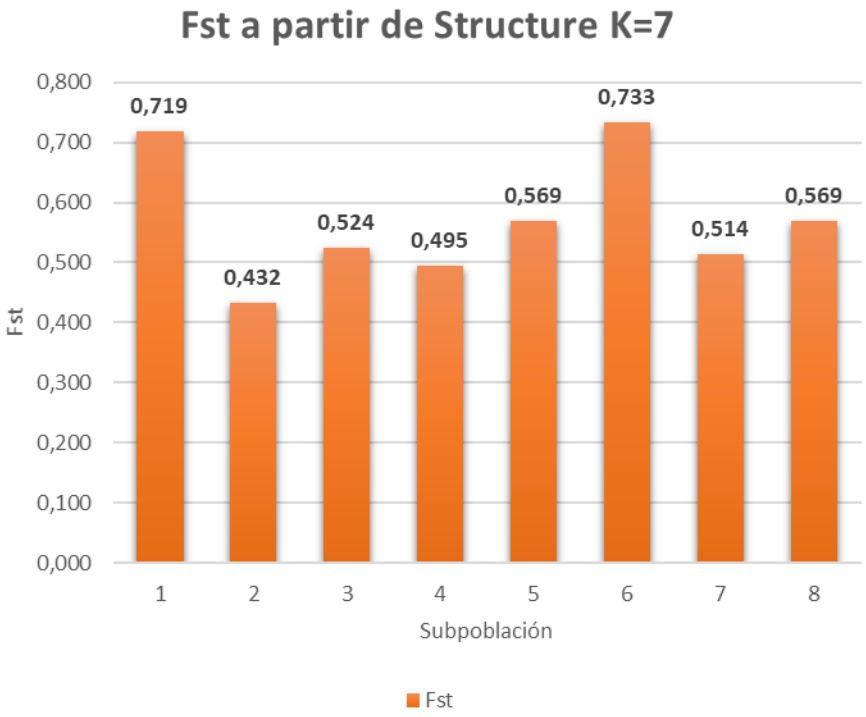


Fig. 5. Gráfico de los promedios de Fst para cada subpoblación, de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012.

Posteriormente, al realizar el análisis de estructuración mediante STRUCTURE, se logró evaluar el agrupamiento de las líneas a sus respectivas (sub) poblaciones de origen.

La caracterización del material genético que constituye a las diferentes subpoblaciones se determinó a partir del análisis de la matriz Q obtenida a partir de cada software (STRUCTURE y fastSTRUCTURE). En lo que respecta al cuerpo de esta, se encuentran los valores que representan la proporción de pertenencia de cada línea a las diferentes subpoblaciones.

A su vez, para visualizar la estructuración en cuestión, a partir de los resultados de estructura de ambos softwares, se realizaron los gráficos de estructura basados en las matrices Q ajustadas a partir de la página web http://pophelper.com/ mediante la interfaz POPHELPER Structure Web App v1.0.10 (Francis, 2016). En estos gráficos, se visualizaron las subpoblaciones existentes en ambos casos, cada una definida por un color determinado como se observa en la Figura 6, gráfico obtenido a partir de los resultados de STRUCTURE; y en la Figura 7, gráfico obtenido a partir de fastSTRUCTURE.

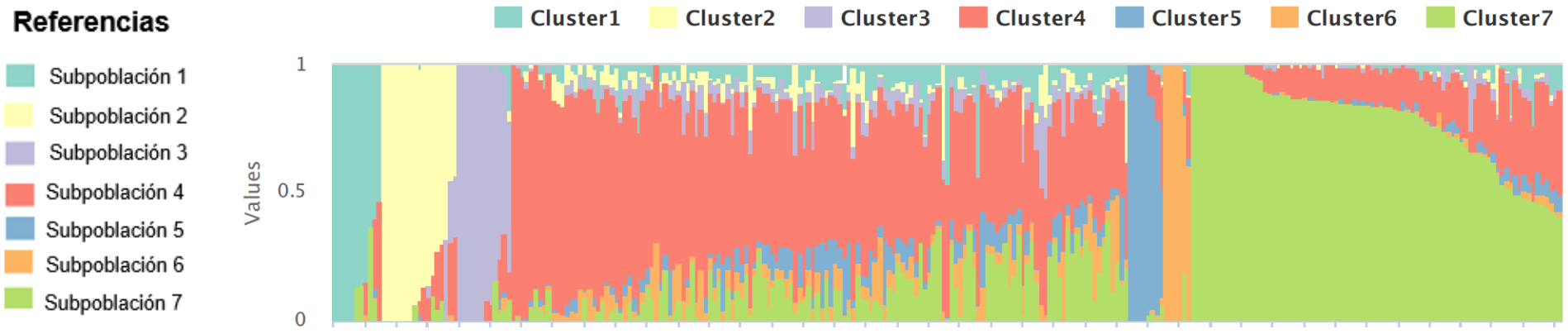


Fig. 6. Representación gráfica de la matriz Q ajustada de K=7 generada con 48898 SNPs mediante STRUCTURE a partir de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.

Referencias de subpoblaciones y la estructuración de cada línea a partir de STRUCTURE

Subpoblación 1

C103
C123
Mo17
NC258
NC262
NC290A
NC342
NC344
NC360
NC362
NC364

Subpoblación 2

A679
A680
B104
B109
B73
B73Htrhm
B84
N192
NC294
NC306
NC310
NC314
NC324
NC326
NC328
NC368
R229

Subpoblación 3

A632
A634
A635
A641
B14A
B64
B68
CM105
CM174
DE811
H105W
H91

Subpoblación 4

A188
A239
A441-5
A554
A654
A659
A661
A682
Ab28A
B10
B103
B105
B115
B164
B2
B37
B46
B52
B57
B75
B76
B77
B79
B97
C49A
CH701-30
CH9
CI_7
CI187-2
CI21E
CI28A
CI31A
CI3A
CI64
CI66
CI90C
CI91B
CM37
CMV3
CO106
CO125
CO255
DE_2
DE_3
DE1
E2558W
EP1
F44
F6
F7
GA209
GT112
H49
H49
H84
H99
Hy
I205
I29
IA2132
IA5125
Il101
Il14H
Il677a
K148
K4
K55
K64
Ky21
Ky226
Ky228
L317
M14
M162W
MEF156-5
Mo1W
Mo24W
Mo44
Mo45
Mo46
Mo47
Mo47
MoG
Mp339
MS1334
MS153
MS71
Mt42
N28Ht
N6
N7A
NC222
NC230
NC232
NC236
NC238
NC250
NC260
NC33
ND246
Oh603
Oh7B
Os420
P39
Pa875
Pa880
PA91
R109B
R168
R177
R4
SC213R
SC357
SD40
SD44
T232
T234
T8
U267Y
VA102
Va14
Va17
Va22
Va35
Va59
Va85
Va99
VaW6
W117Ht
W153R
W182B
W22
W22_R_r
_
WD
WF9
Yu796_N
S

Subpoblación 5

4722
Hp301
IDS28
IDS69
IDS91
SA24
Sg1533
Sg18

Subpoblación 6

A619
H95
Oh40B
Oh43
Oh43E
Pa762

Subpoblación 7

A272
A6
CML10
CML103
CML108
CML11
CML14
CML154Q
CML157Q
CML158Q
CML218
CML220
CML228
CML238
CML247
CML254
CML258
CML261
CML264
CML277
CML281
CML287
CML311
CML314
CML321
CML322
CML323
CML328
CML331
CML332
CML333
CML341
CML38
CML45
CML5
CML52
CML61
CML69
CML77
CML91
CML92
F2834T
Hi27
I137TN
Ki11
Ki14
Ki2021
Ki21
Ki3
Ki43
Ki44
L578
M37W
Mo18W
NC264
NC296
NC296A
NC298
NC300
NC302
NC304
NC318
NC320
NC336
NC338
NC340
NC346
NC348
NC350
NC352
NC354
NC356
NC358
NC366
SC55
Tx303
Tx601
Tzi10
Tzi11
Tzi16
Tzi18
Tzi25
Tzi9

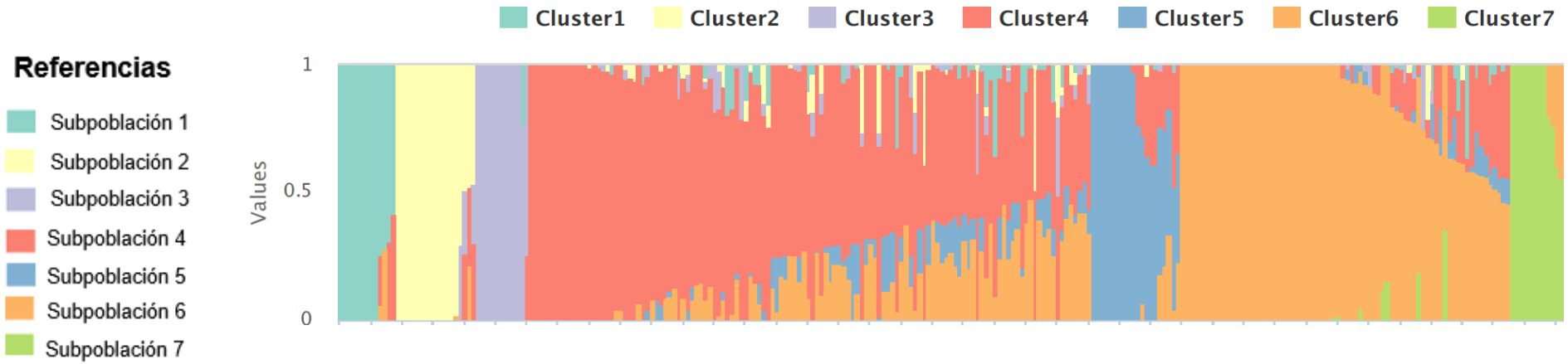
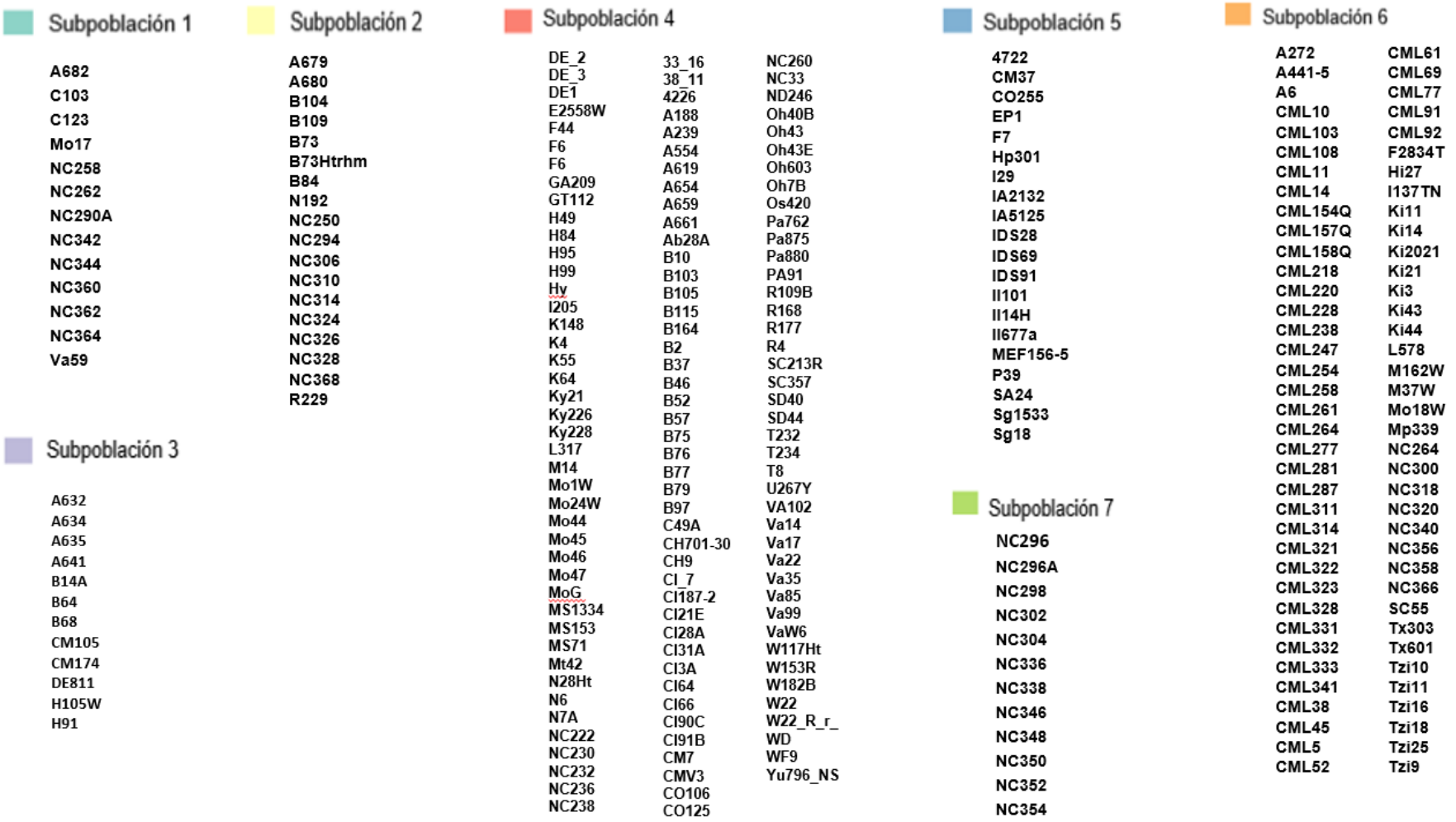


Fig. 7. Representación gráfica de la matriz Q ajustada de K=7 generada con 48898 SNPs mediante fastSTRUCTURE a partir de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.

Referencias de subpoblaciones y la estructuración de cada línea a partir de FastSTRUCTURE



5.3 Análisis de distancia

A continuación, se obtuvieron los árboles de distancia en TASSEL a partir del método UPGMA (*Unweighted Pair-Group Method with Arithmetic*). A partir de estos árboles, se realizaron los gráficos coloreados según los agrupamientos obtenidos a partir de STRUCTURE y fastSTRUCTURE como se observan en las Figuras 8 y 9, utilizando ITOL (*Interactive Tree of Life) (*https://itol.embl.de/upload.cgi) y así lograr una mejor visualización.

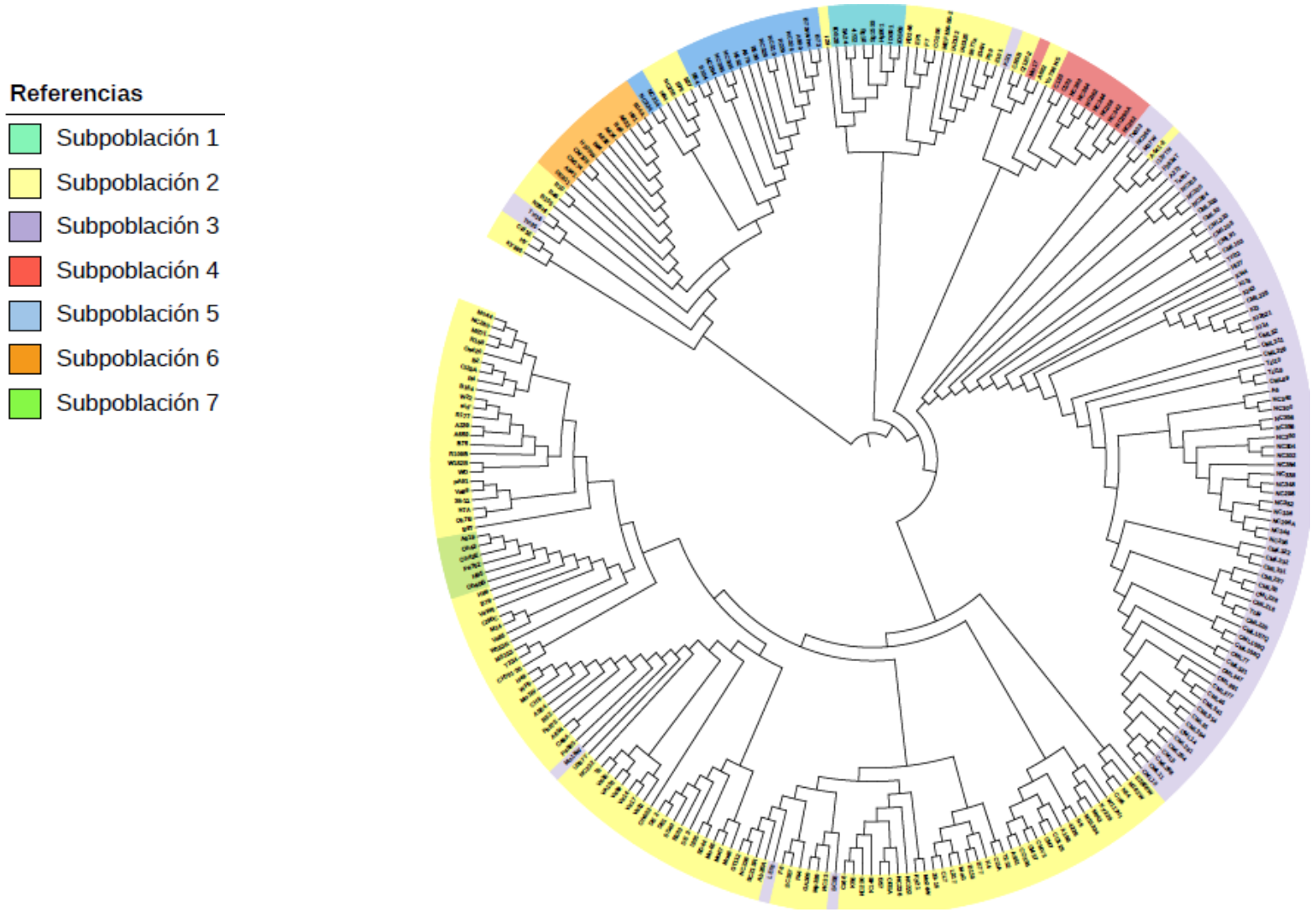


Fig. 8. Representación gráfica del árbol de distancia mediante el método UPGMA (*Unweighted Pair-Group Method with Arithmetic*) a partir de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012. Distribución del agrupamiento de las subpoblaciones obtenidas a partir de STRUCTURE con un K=7.

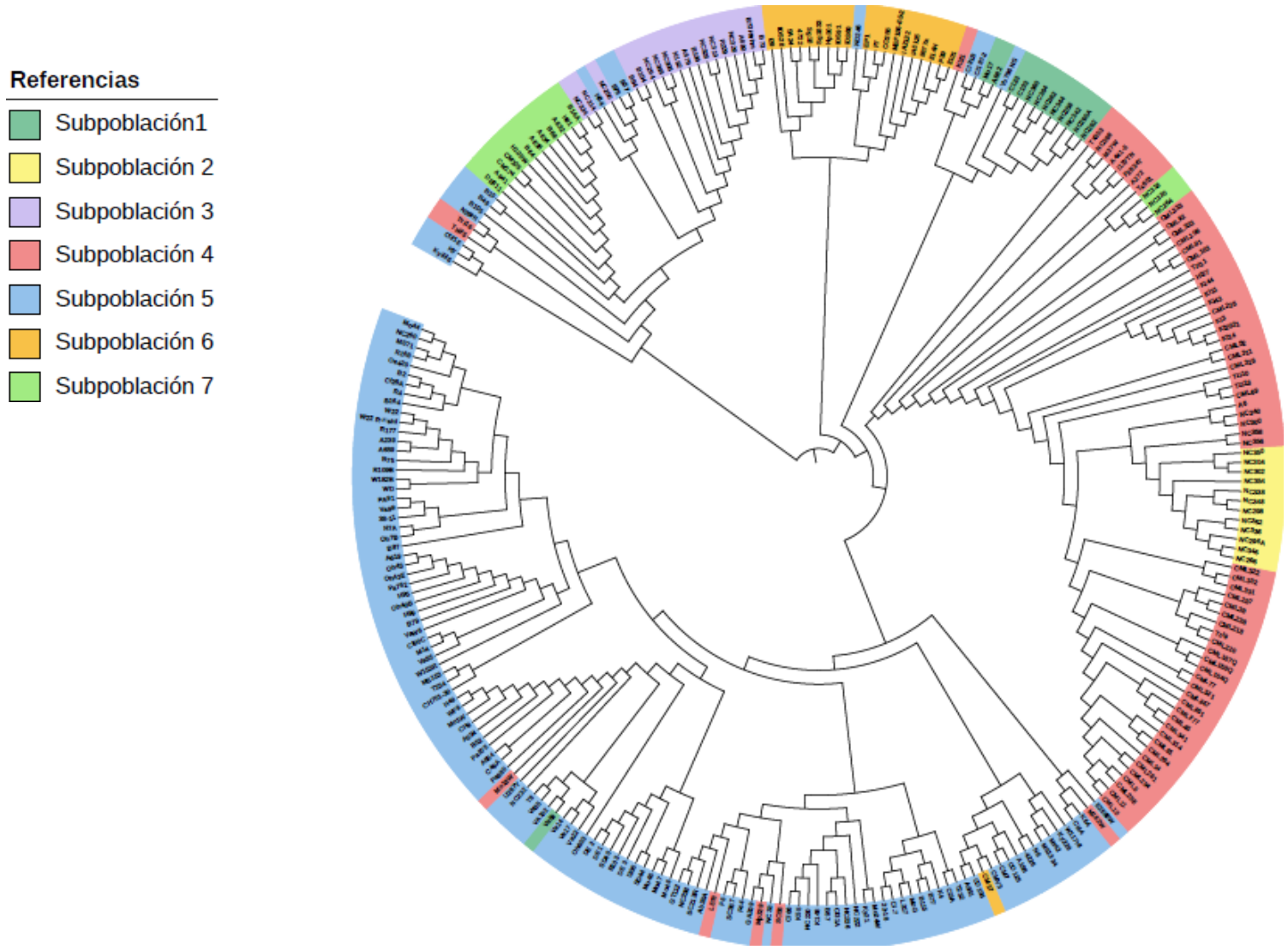


Fig. 9. Representación gráfica del árbol de distancia mediante el método UPGMA *(Unweighted Pair-Group Method with Arithmetic*) a partir de un panel de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012. Distribución del agrupamiento de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.

5.4 Reducción dimensional

A continuación, mediante t-SNE se logró evaluar, nuevamente, el patrón de agrupamiento de los datos, mediante un método alternativo a STRUCTURE y fastSTRUCTURE para proceder a una comparación entre ambos. Con este paso de reducción dimensional se logró disminuir la dimensionalidad con la que se venía trabajando a sólo dos ejes o dimensiones, y de esta forma manejar de manera más simple la cantidad de líneas y SNPs evaluados.

El análisis de reducción dimensional se llevó a cabo a partir de la matriz de distancia obtenida por TASSEL. Luego de realizar varias pruebas y ajustar los parámetros de evaluación (datos no mostrados), se consideraron como valores de perplejidad 5, 15 y 30, lo que permitió explorar y comparar los resultados obtenidos en los puntos anteriores y establecer diferencias estructurales en los datos. Cabe destacar que la perplejidad es un valor que permite realizar una distribución de probabilidad representando similitudes entre los vecinos más cercanos (Van der Maaten y Hinton, 2008). Dependiendo de cuan robusto sea el set de datos en estudio, se establece el valor de perplejidad utilizado que, según las recomendaciones de los autores del algoritmo, Van der Maaten y Hinton, (2008), suele oscilar entre 5 y 50. La perplejidad puede interpretarse como una medida suave del número efectivo de vecinos. El desempeño de t-SNE es bastante robusto a los cambios en la perplejidad, y los valores típicos están, como ya se mencionó, entre 5 y 50.

Como resultado en este paso del análisis, se obtuvieron gráficos conformados por dos ejes, en los cuales se encuentran representadas cada línea en gráficos de puntos. En este caso, no solo se pueden obtener gráficos sin colorear como coloreados. En un primer paso se obtuvieron los gráficos sin colorear y luego los mismos coloreados, demostrando los agrupamientos en diferentes colores. De esta forma se logró asociar cada línea a su grupo de pertenencia junto con la etiqueta correspondiente a su nombre y así facilitar su distinción. A partir de este análisis, se obtuvo una estructuración visualmente identificable en los gráficos coloreados, de 5 grupos permitiendo ver los agrupamientos en el espacio de manera más dinámica. También, facilitó visualizar bien definidas a 5 de las 7 subpoblaciones totales, distribuidas espacialmente de manera clara, al comparar los gráficos obtenidos a partir de los diferentes softwares.

Como se observa en las Figuras 10A, 10B y 10C; se obtuvieron los gráficos, a partir de los agrupamientos de STRUCTURE, de tipo diagrama de dispersión para los diferentes

valores de perplejidad considerados. Se observó que, a mayor valor de perplejidad, los datos presentan un agrupamiento en el espacio más marcado, pero simultáneamente una mayor dispersión entre sí. En la sección anexos se encuentran los gráficos de t-SNE obtenidos a partir de la matriz de distancia, considerando los agrupamientos no solo de STRUCTURE como también los de fastSTRUCTURE.

A continuación, se realizaron los gráficos de t-SNE que presentan la descripción del nombre de cada línea, lo cual ayudó a identificar en mayor medida los agrupamientos y las líneas que se encontraban en cada grupo como se observan en las Figuras 11A, 11B y 11C.

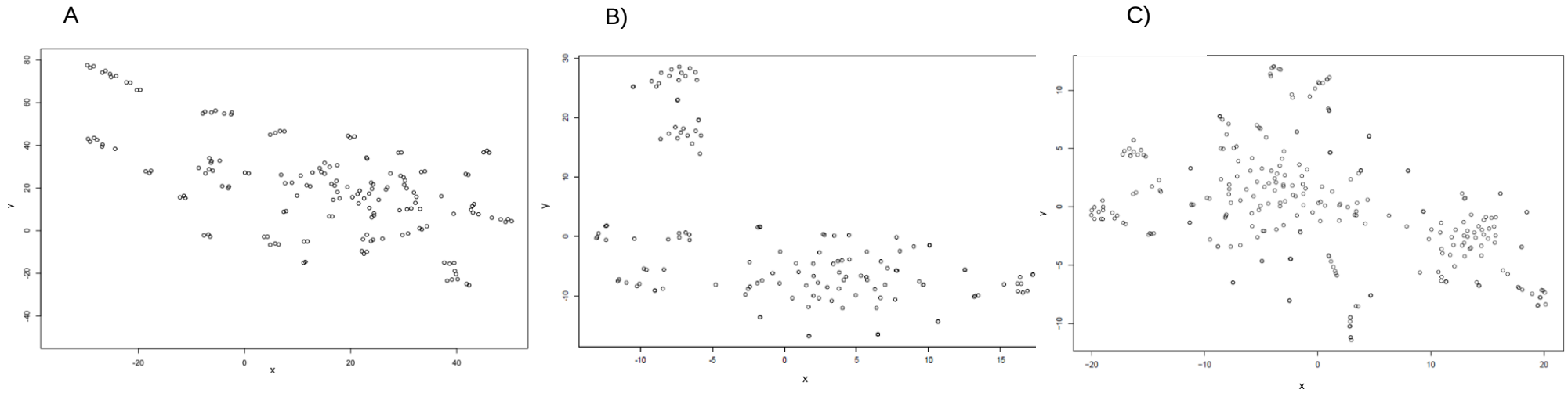


Fig.10. Gráficos de 2 componentes obtenidos mediante t-SNE a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012. Representación de los datos: A) perplexity=5 y 20.000 iteraciones; B) perplexity=15 y 20.000 iteraciones y C) perplexity=30 y 20.000 iteraciones.

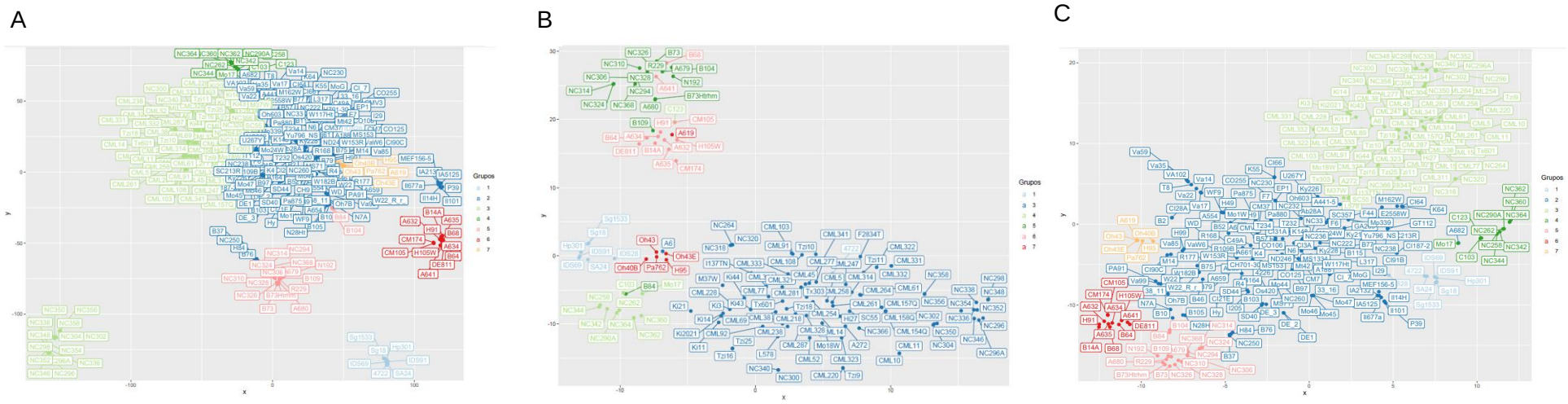


Fig.11. Gráficos de dispersión obtenidos a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipado por Cook *et al*., 2012. Representación de los datos: A) perplexity=5 y 20.000 iteraciones, B) perplexity=15 y 20.000 iteraciones y C) perplexity=30 y 20.000 iteraciones. Los colores reportados corresponden a la distribución del agrupamiento de las subpoblaciones obtenidas a partir de STRUCTURE con K=7.

Luego de realizar los agrupamientos en t-SNE de las 7 subpoblaciones, surgió la incertidumbre de si este método generaría un mejor agrupamiento eliminando uno de los grupos mixtos, es decir los grupos con individuos con valores bajos de probabilidad de pertenencia a la subestructura presente, siendo para el set de datos analizados el grupo 4 en el caso de STRUCTURE y fastSTRUCTURE (ver sección 5.2). Cabe destacar que las líneas que forman parte de dichos grupos se encuentran particionadas en diferentes subpoblaciones con lo cual, se puede establecer que son genotipos que no presentan una proporción alta de pertenencia a un grupo específico. Individuos que podrían no haberse estabilizado durante la retrocruza recurrente, es decir líneas que no llegaron a nivel de líneas puras.

Al eliminar este grupo, se observó un agrupamiento definido de 6 subpoblaciones, con mayor agrupamiento de líneas similares y mayor separación de los grupos unos de otros, como se observa en las Figuras 12A y 12B.

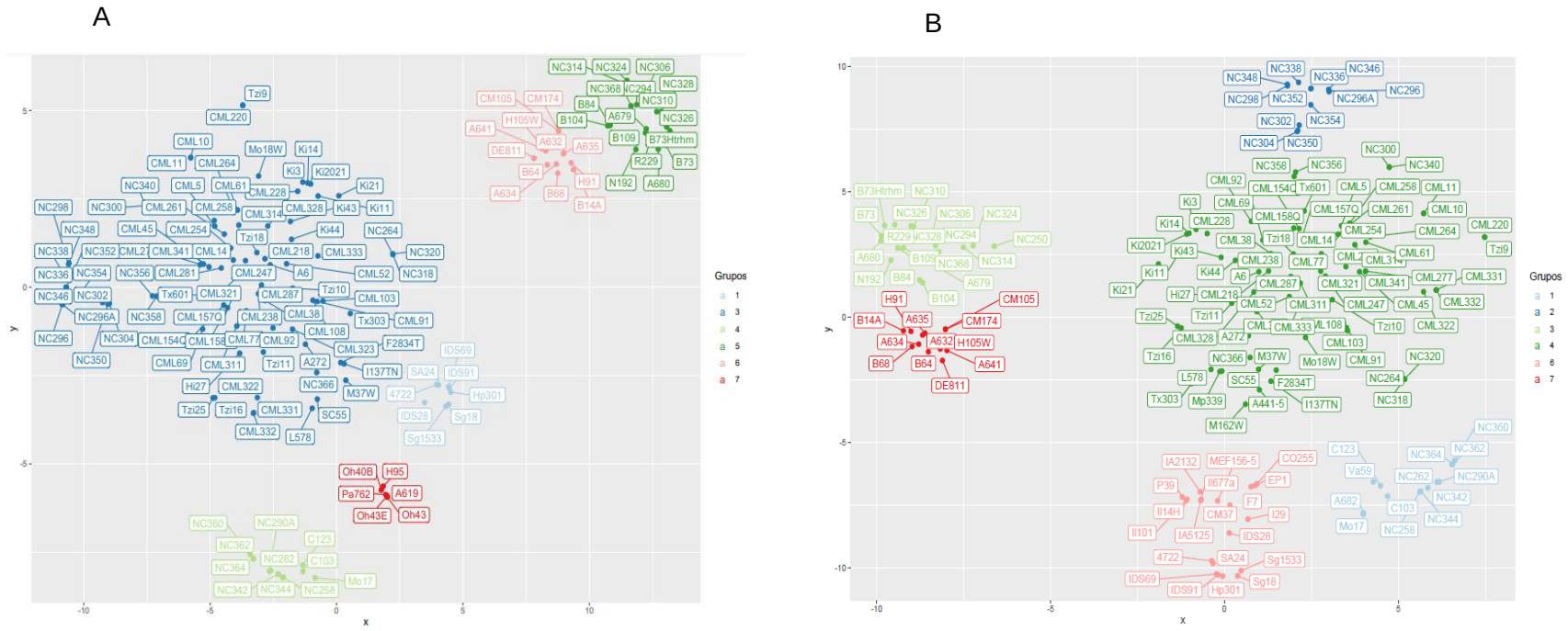


Fig.12. Gráficos de diagrama de dispersión obtenidos a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.Representación de los datos: A) perplexity=30 y 20.000 iteraciones y B) perplexity=30 y 20.000 iteraciones Distribución del agrupamiento de las subpoblaciones obtenidas a partir de STRUCTURE con un K=7 y filtrando el grupo de líneas con probabilidad de pertenencia a estructura menor al 50%.

En relación con la complejidad computacional, el tiempo de análisis de estructuración fue de 4N permitiendo obtener los resultados en un servidor local, con un equipo de escritorio en minutos. A diferencia, el análisis en STRUCTURE se llevó a cabo en un servidor de alta complejidad y en días de corridas.

5.4.1 Análisis discriminante de componente principales

Se realizó un análisis discriminante de componentes principales con el fin de comparar estos resultados con los obtenidos a partir del t-SNE. En este caso, como se observa en la Figura 13, el valor BIC (*bayesian inference criterion*), equivalente al valor óptimo del método de Evanno, muestra un cambio de tendencia en un k=5 por lo tanto, según este método se puede proponer un total de 5 subpoblaciones presentes en este panel. Este método no resultó concordante con lo expuesto anteriormente. A continuación, en las Figuras 14A y 14B se pueden observar los agrupamientos de los datos en ambas figuras, donde se representan 7 grupos para realizar una comparativa con los puntos expuestos previamente, aun cuando no se logran distinguir claramente las 7 subpoblaciones existentes previamente definidas por el método bayesiano, siendo que 3 estructuras presentes parecieran estar colapsadas en un único grupo, resultando de esta manera en 5 potenciales grupos particionados artificialmente.

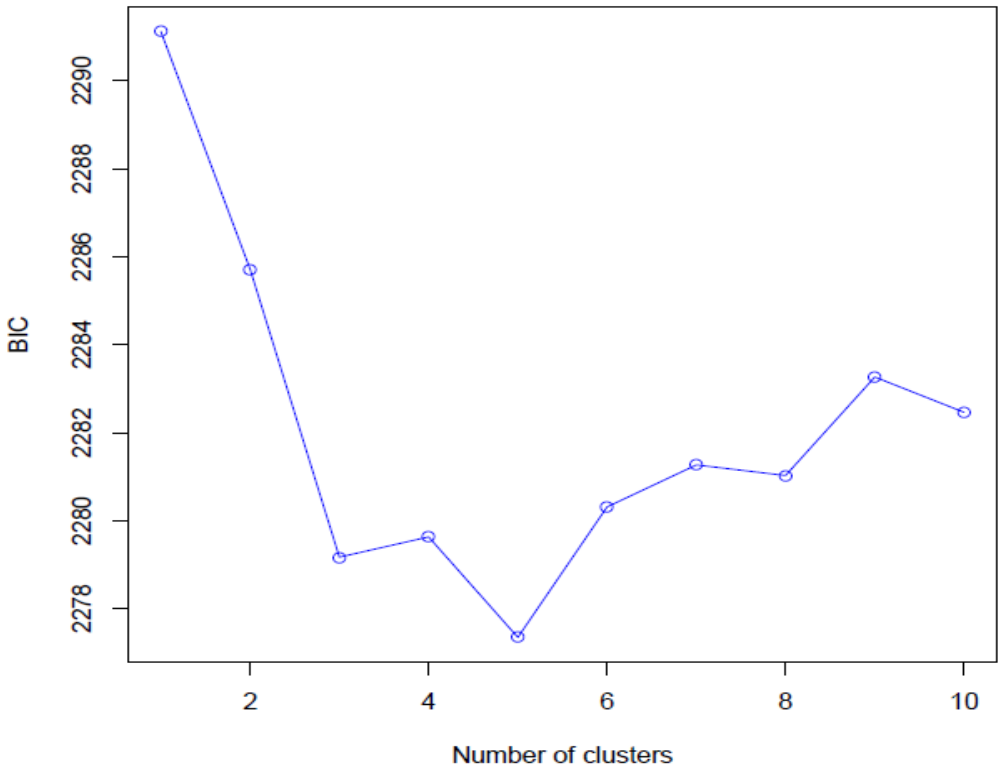


Fig.13. Gráfico de los valores BIC que muestra el número de clusters óptimos para este método. Análisis de 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.

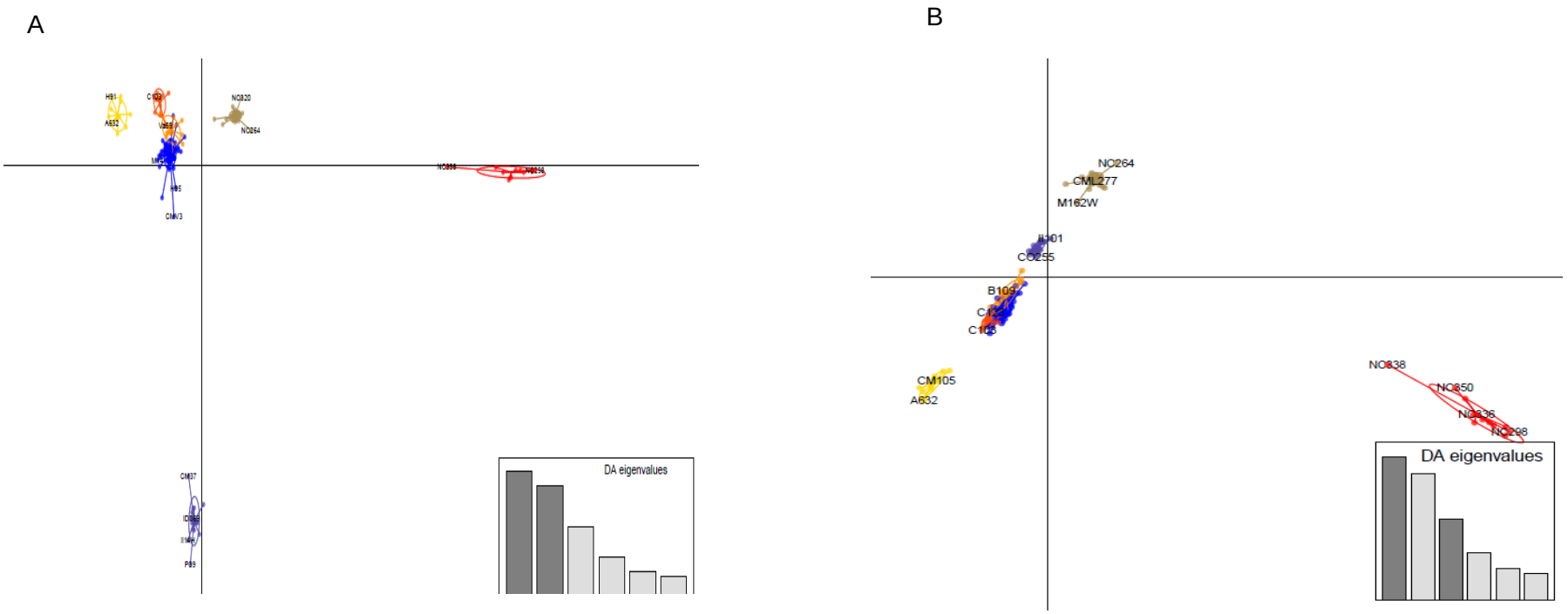


Fig.14. Gráficos de DAPC de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012. A) Representación de las 2 componentes principales. B) Representación de la primer y tercer componentes principales. Distribución a partir del agrupamiento de las subpoblaciones obtenidas a partir de STRUCTURE con un K=7.

4.2 Análisis mediante K-means

Posteriormente se realizó un análisis de agrupamientos mediante el método K-means como se observa en la figura 15A. Mediante el grafico de diagrama de dispersión no se lograron destacar agrupamientos claros y definidos sin que los datos estén etiquetados ya que, los puntos se encuentran muy dispersos en el espacio. En la figura 15B se pueden ver los grupos descriptos, pero nuevamente no es un agrupamiento claro.

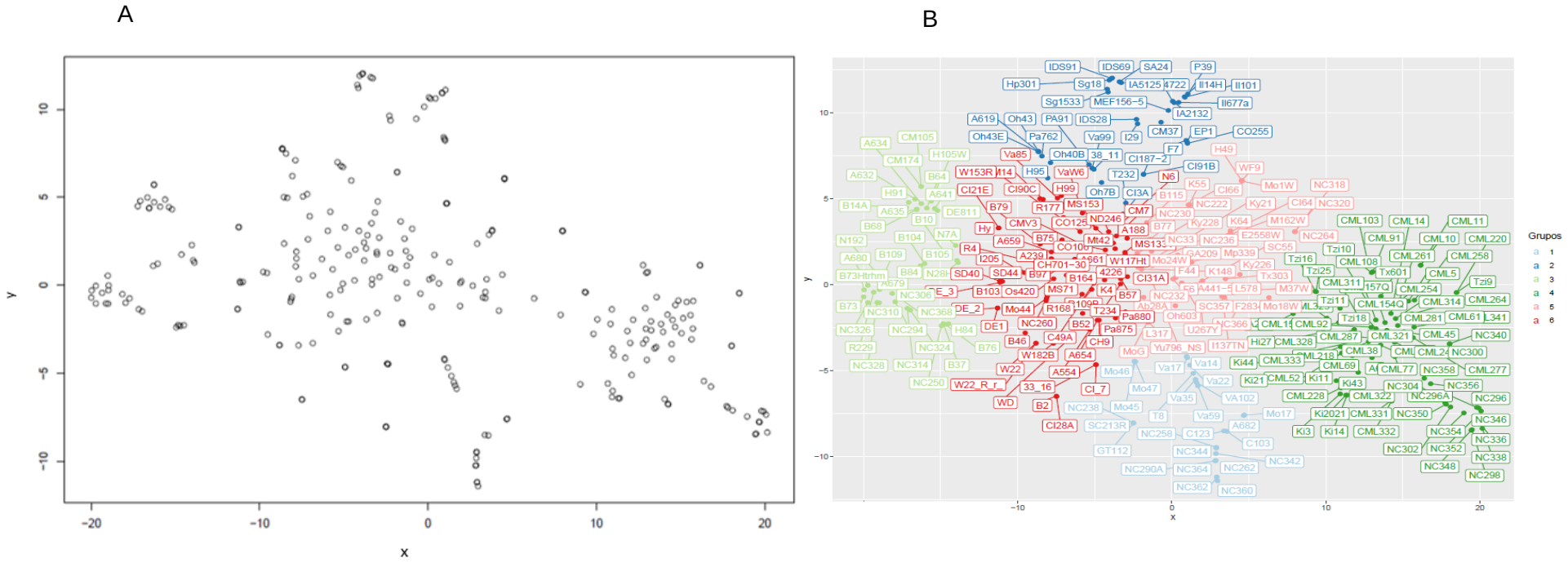


Fig.15. Gráficos de la distribución de los datos a partir de la matriz de distancia reducida mediante el método K-means de las 275 líneas endocriadas de maíz de un panel público previamente genotipado por Cook *et al*., 2012. A) Gráfico de dispersión y B) Gráfico con las etiquetas de las líneas analizadas.

4.3 Análisis mediante dbSCAN

Como último paso se realizó un análisis de agrupamientos mediante el método dbSCAN tal como se observa en la Figura 16. Mediante este método tampoco se logró observar una formación clara de los grupos y sus individuos correspondientes, quedando genotipos dispersos en todo el grafico sin respetar un agrupamiento especifico.

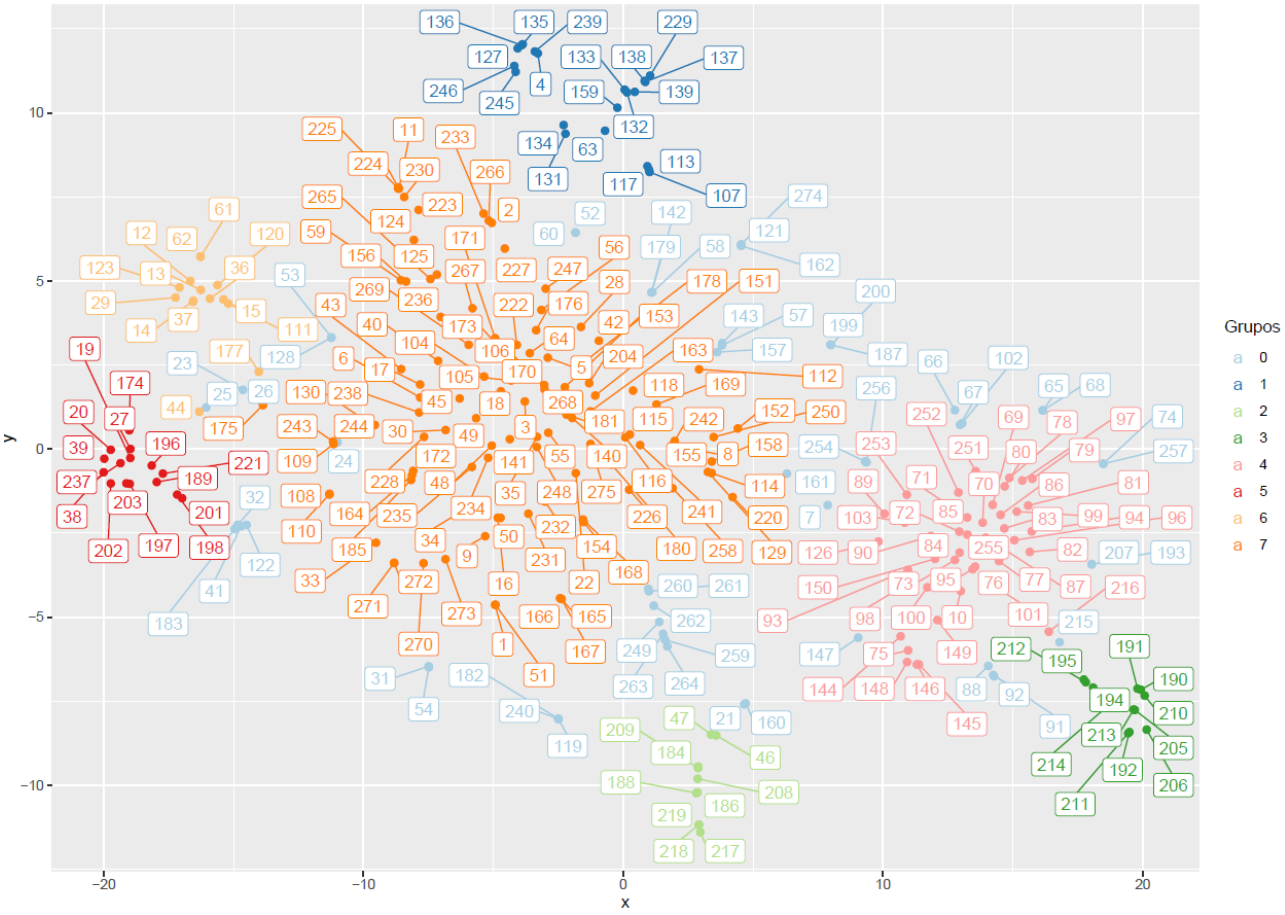


Fig.16. Gráficos de la distribución de los datos a partir de la matriz de distancia reducida mediante el método dbSCAN de las 275 líneas endocriadas de maíz de un panel público previamente genotipado por Cook *et al*., 2012.

6. Discusión

Considerando que el maíz *(Zea mays*) es afectado por un gran número de enfermedades que disminuyen su producción y calidad, una alternativa para resolver esta problemática consiste en incluir enfoques que incluyan la aplicación de las nuevas tecnologías de secuenciación masiva. Ya que estas han permitido analizar un gran número de individuos y su respectiva variación genotípica y fenotípica, reduciendo en gran medida los tiempos de análisis y los requerimientos técnicos e informáticos.

A modo de resumen, en el presente trabajo se lograron comparar varios métodos para un análisis de estructura poblacional y evaluar cual/cuales resultaron ser los más eficientes para estos datos. Se utilizaron métodos tradicionales como lo es STRUCTURE y se comparó con otros métodos como fastSTRUCTURE, tSNE, PCA, K-means, dbSCAN y/o métodos de distancia. Según los resultados obtenidos, en rasgos generales no se lograron establecer agrupamientos certeros con aquellos métodos independientemente del uso de STRUCTURE con lo cual, se observó la formación de grupos, pero no totalmente consolidados. También se realizó una comparación de las subpoblaciones aquí obtenidas, con las subpoblaciones obtenidas por el grupo de Liu K. y colaboradores (Liu, et al., 2003).

El objetivo general de este trabajo consistió en evaluar la efectivización del proceso de análisis y obtención de resultados de datos de genotipado de un panel de 275 genotipos, mediante un conjunto de herramientas de la genética y la ciencia de datos. Estas líneas endocriadas de maíz forman parte de un panel público previamente genotipado y definido. Este set de líneas incluye familias de líneas recombinantes y familias de líneas que presentan elevada endocría.

Los primeros objetivos específicos planteados fueron analizar, filtrar e establecer los haplotipos de las líneas endocriadas, así como también, determinar la diversidad genética e inferir la estructura poblacional y el parentesco relativo entre los genotipos en estudio. Se logró observar que este panel de líneas según los análisis con STRUCTURE y fastSTRUCTURE, se encuentra estructurado en 7 subpoblaciones con valores de Fst superiores a 0,5. Estos valores demuestran una elevada estructuración y una fuerte y significativa diferenciación genética debido a la elevada endocría de este panel, también la existencia de un fuerte proceso de selección. Esto indica que los individuos presentan fondos genéticos similares dentro de una misma subpoblación y una marcada

diferenciación con respecto a individuos de otras subpoblaciones lo cual, estaría demostrado por el valor de Fst tan elevado indicando la acción de la deriva génica poblacional. A diferencia de la estructura existente a nivel natural la cual, según bibliografía se encuentra en valores por debajo o cercanos a 0,2, podemos concluir que no se podría realizar una comparación entre las poblaciones estudiadas y las naturales ya que, los fenómenos de deriva genética, migración, panmixia y mutación son muy fuertes y marcados en las poblaciones naturales los cuales generan la existencia de un valor de Fst mucho más bajo que el obtenido y presentado en este trabajo.

Con respecto a la estructuración mediante inferencia bayesiana, se lograron observar agrupamientos equivalentes con ambos softwares ya que, el patrón de estructura de las subpoblaciones en STRUCTURE mayormente se mantuvo también con fastSTRUCTURE.

Al comparar ambos métodos se pudo establecer que, si bien fastSTRUCTURE no brinda por defecto un número de subpoblaciones posible en el set de datos estudiado, al realizar el análisis estableciendo como número de subpoblaciones K=7, luego de haber ejecutado el método de Evanno y obtener ese resultado, este software brindó resultados de estructuración válidos logrando una reducción del tiempo de cálculo y recursos computacionales, sin perder calidad de la información. De todas formas, el hecho de que fastSTRUCTURE no pueda establecer por si solo el numero K de subpoblaciones presentes en un set de datos es una limitante al momento de iniciar un análisis de estructuración poblacional.

A continuación, la existencia de equivalencia en la estructuración y obtención de resultados similares entre STRUCTURE y fastSTRUCTURE, permitió la posibilidad de establecer un análisis comparativo de los grupos obtenidos por inferencia bayesiana y los obtenidos mediante árboles de distancia, observando un patrón de agrupamiento diferencial.

Posteriormente, se realizó un análisis a partir de matrices de distancia para lograr validar estos agrupamientos. Se obtuvieron arboles de distancia a partir del método UPGMA, los cuales al compararlos con los agrupamientos del método bayesiano no lograron definir criteriosamente a las subpoblaciones.

Luego, se llevó a cabo una reducción dimensional para evaluar si esta técnica permite explicar en mayor medida los agrupamientos obtenidos. Si bien, permitió lograr una buena aproximación de las subpoblaciones presentes en un rango de dos a tres dimensiones, no termina siendo tan robusta como lo es STRUCTURE.

Cabe destacar que el t-SNE, un algoritmo de *machine learning* que permite la reducción de datos de alta complejidad (decenas de miles de SNPs considerados), no solo permitió una óptima visualización de los datos, sino que también demostró ser una herramienta en aquellos casos en los que la estructuración no era clara. Es decir, cuando STRUCTURE/fastSTRUCTURE-Evanno arrojó valores de estructuración o pertenencia de un genotipo, repartidos entre varias subpoblaciones, t-SNE colaboró con el resultado final de pertenencia de cada línea. De todas formas, no es un método que permita agrupar por sí solo a los datos, sin basarse en agrupamientos o una estructuración previa o posterior.

A su vez, la utilización de árboles de distancia a partir de las matrices de distancia, si bien resultó ser un método rápido y económico se observó que no es contundente en los resultados al compararlo con STRUCTURE ya que, difiere en el agrupamiento de varias líneas, generando cierta incertidumbre al momento del análisis de los genotipos y sus grupos.

En el caso del análisis mediante K-means y dbSCAN a pesar de ser métodos muy útiles y rápidos en su ejecución tampoco mostraron ser robustos en el análisis ya que, si o si requieren de información previa para llevar a cabo los agrupamientos. Al no generar agrupamientos por si solos, si no se conoce previamente la cantidad de grupos que hay presente en el set de datos, estos métodos no son capaces de discernirlos correctamente.

Con respecto a la evaluación de la efectivización del tiempo de cálculo, el desempeño de fastSTRUCTURE fue notablemente más óptimo en el uso del tiempo y recursos computacionales que STRUCTURE. No obstante, es importante mencionar que el algoritmo fastSTRUCTURE no asigna un valor de delta *K* óptimo para los agrupamientos, dado que el algoritmo “chooseK” incorporado en fastSTRUCTURE devuelve un numero mucho mayor de grupos que el establecido por STRUCTURE.

Con respecto a los agrupamientos obtenidos, dentro de este panel se encontraron 7 subpoblaciones con miembros agrupados de manera robusta y algunas líneas formando parte de los grupos con mejor solidez o porcentaje de pertenencia.

En función del pedigrí de las líneas evaluadas se realizó una comparación de los grupos obtenidos en este trabajo con los grupos obtenidos por Liu, et al. (2003). En su caso, previamente excluyen del análisis de estructuración a líneas correspondientes a los grupos *Sweet* (maiz dulce) y Popcorn, obteniendo como resultado 5 grupos. De esos 5 grupos, 2 (Sweet y Popcorn), no forman parte del proceso de estructuración mediante STRUCTURE. En cambio, en este trabajo los genotipos que conforman estos 2 grupos se incorporaron al análisis y fueron agrupados en las subpoblaciones 4 y 5. La subpoblación

5 se encuentra muy definida por líneas Popcorn, y en la subpoblación 4 se agrupan genotipos correspondientes al grupo Sweet, Popcorn pero también Non Stiff Stalk (NSS). Profundizando en los orígenes de estas líneas, se puede destacar que en este grupo se encuentra germoplasma con características tropicales junto con genotipos característicos del sur de Carolina del Norte cuyas condiciones climáticas van de subtropical a húmedo templado.

La subpoblación 1 agrupó líneas que en su mayoría corresponden a germoplasma Non Stiff Stalk (NSS) en su mayoría caracterizadas por desarrollarse en zonas subtropicales.

En el caso de las subpoblaciones 2 y 3, las líneas que la conforman corresponden a germoplasma Stiff Stalk (SS). La diferencia entre ambos agrupamientos se debe a que en la subpoblación 2 se encuentra germoplasma adaptado a regiones subtropicales y en el caso de la subpoblación 3, germoplasma adaptado a regiones húmedas (frías y templadas). Estas últimas 3 subpoblaciones se encuentran muy definidas y consolidadas.

La subpoblación 7 en el caso de fastSTRUCTURE está conformada por líneas de condición tropical o semitropical (TS), este grupo también está bien definido. Luego el software agrupó otro conjunto de líneas TS en la subpoblación 6 en la cual también se encuentran individuos correspondientes a un grupo denominado en el trabajo de Liu et al. (2003), como grupo “mixto”. Algunos de estos germoplasmas tienen orígenes de Tailandia, Sudáfrica, Hawaii, y regiones climáticas de Estados Unidos que van de húmedas a templadas. En este caso particular, STRUCTURE agrupó a estas líneas en una gran subpoblación, la número 7, describiendo solo como población 6 a 6 líneas NSS.

Entre todos los métodos de agrupamientos probados y evaluados en este trabajo, el método bayesiano implementado por STRUCTURE generó mejores subgrupos estableciendo agrupamientos de manera más efectiva y permitiendo agrupar de forma más precisa las líneas. De todas formas, aquí se logró demostrar la posibilidad de la utilización de diferentes métodos, algoritmos y softwares que permiten una aproximación a la estructuración verdadera presente en los datos.

7. Conclusiones

En este trabajo se utilizó la información genotípica de un chip de 56.658 SNPs inciales y 275 líneas de maíz provenientes de un panel público de línea endocriadas para evaluar la efectivización del tiempo, al trabajar con estructura poblacional, utilizando diferentes herramientas bioinformáticas y de la ciencia de datos.

En primer lugar, aquí se concluyó que el algoritmo variacional que incorpora fastSTRUCTURE es casi dos órdenes de magnitud más rápido que STRUCTURE, no requiere de un servidor informático robusto y no sacrifica precisión o potencia, sin embargo, presenta como dificultad el hecho de que, hasta el momento no se ha logrado propiamente establecer el numero óptimo de subpoblaciones presentes en los datos, sin involucrar algún otro método de análisis. Definitivamente este punto es una limitante importante al momento de decidir que método utilizar para llevar a cabo un análisis de estructura, con una gran cantidad de datos.

En segundo lugar, se logró comprobar que a pesar de que todos los softwares probados no brindaron resultados totalmente certeros y congruentes, al analizar cada método independientemente entre sí, se debe destacar el hecho de que sí se logró demostrar la posibilidad de utilizar estas metodologías como soporte. Es decir que, al menos permiten definir o ubicar espacialmente y de manera más gráfica, aquellos genotipos que presenten valores de agrupamiento más bajos o indefinidos, según STRUCTURE, porque se deben a líneas que aún no se encuentran estabilizadas genéticamente.

En base a los resultados aquí presentados se puede determinar que el panel bajo estudio permitió comparar los diferentes métodos de agrupamientos y los distintos softwares disponibles para llevar a cabo los análisis desde una computadora de escritorio, sin el requerimiento de un gran servidor de alta efectividad. De todas formas y concluyendo se pudo distinguir que a pesar de la disponibilidad de diferentes herramientas bioinformáticas y estadísticas, aun nos encontramos con algunos obstáculos como lo es la especificación del número de agrupamientos existentes en un set de datos. Hasta el momento y según lo analizado aquí no se puede determinar con certeza un numero de subpoblaciones especifico sin la utilización de STRUCTURE.

8. Referencias Bibliográficas

Abriata et al. (2010). Biotecnología y Mejoramiento Vegetal II. In M. Dra. Gabriela, Levitus; Dra. Viviana, Echenique; Dra. Clara, Rubinstein; Dr. Esteban, Hopp; Ing. Agr. Luis (Ed.), *ArgenBio. Consejo Argentino para la Información y el Desarrollo de la biotecnología*. https://doi.org/10.17129/botsci.1521

Acquaah, G. (2012). Principles of Plant Genetics and Breeding: Second Edition. In *Principles of Plant Genetics and Breeding: Second Edition*. https://doi.org/10.1002/9781118313718

Adu, G. B., Badu-Apraku, B., Akromah, R., Garcia-Oliveira, A. L., Awuku, F. J., & Gedil, M. (2019). Genetic diversity and population structure of early-maturing tropical maize inbred lines using SNP markers. *PLoS ONE*, *14*(4), 1–12. https://doi.org/10.1371/journal.pone.0214810

Bevilacqua, M., & Storti, L. (2019). *Informes de cadenas de valor*. 50. https://www.argentina.gob.ar/sites/default/files/sspmicro_cadenas_de_valor_maiz.pdf.

Boehmke, B. Hierarchichal Cluster Analysis, Business Analytics R Programming guide. University of Cincinnati. https://uc-r.github.io/hc_clustering

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.

Cabrero, J., & Camacho, J. (2000). Fundamentos de genética de poblaciones. EVOLUCIÓN, La Base de La Biología, 83–126. http://sesbe.org/sites/sesbe.org/files/recursos-esbe/fundamentos_GdeP.pdf

Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. Data Mining and Knowledge Discovery, 27 (3):344-371.ISSN 1384-5810. doi:10.1007/s10618-013-0311-4.

Cook, J.P.; McMullen, M.D.; Holland, J.B.; Tian, F.; Bradbury, P.; Ross-Ibarra, J.; Buckler, E.S.; Flint-Garcia S.A. Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association m Plant Physiology. (2012). 158 (2), 824–834. https://doi.org/10.1104/pp.111.185033

Darwin, C. The Origin of Species (Penguin Books USA Inc., New York, 1958)

De Faria, S.V.; Zuffo, L.T.; Rezende, W.M.; Caixeta, D.G.; Pereira, H.D.; Azevedo, C.F.; DeLima, R.O. Phenotypic and molecular characterization of a set of tropical maize inbred lines from a public breeding program in Brazil. BMC Genomics. (2022). 23(1):54. doi: 10.1186/s12864-021-08127-7. PMID: 35030994; PMCID: PMC8759194.

Doebley J, Stec A, Gustus C. teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. Genetics. 1995 Sep;141(1):333-46. doi: 10.1093/genetics/141.1.333. PMID: 8536981; PMCID: PMC1206731.

Duvick, D. N. (2001). Biotechnology in the 1930s: The development of hybrid maize. *Nature Reviews Genetics*, *2*(1), 69–74. https://doi.org/10.1038/35047587

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

Elhaik, E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep 12, 14683 (2022). https://doi.org/10.1038/s41598-022-14395-4

Evanno, G.,REGNAUT, S., & GOUDET, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.Molecular Ecology, 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics Society of America. Genetics 164: 1567–1587.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, *7*(4), 574–578. https://doi.org/10.1111/j.1471-8286.2007.01758.x

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, *54*(1), 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E.,

Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S. (2005). Maize association population: A high-resolution platform for quantitative trait locus dissection. *Plant Journal*, *44*(6), 1054–1064. https://doi.org/10.1111/j.1365-313X.2005.02591.x.

Francis R. M., 2016. pophelper: an R package and web app to analyse and visualize population Structure. Molecular Ecology Resourses, Population Genomics with R, 27-32. https://doi.org/10.1111/1755-0998.12509.

Gupta P.K. & Rustgi S. 2004. Molecular markers from the transcribed/expressed region of the genome in higher plants. Funct Integr Genomics 4:139-162.

Hinton, G. & Roweis, S. (2002). Stochastic Neighbor Embedding. Department of Computer Science, University of Toronto. https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

Kumar, B.; Rakshit, S.; Kumar, S.; Singh, B.K.; Lahkar, C.; Jha, A.K.; Kumar, K.: Kumar, P.; Choudhary, M.; Singh, S.B.; Amalraj, J.J.; Prakash, B.; Khulbe, R.: Kamboj, M.C.; Chirravuri, N.N.; Hossain, F. Genetic Diversity, Population Structure and Linkage Disequilibrium Analyses in Tropical Maize Using Genotyping by Sequencing. Plants (Basel). (2022) .11(6):799. doi: 10.3390/plants11060799. PMID: 35336681; PMCID: PMC8955159.

Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016 Jul 8;44(W1):W242-5. doi: 10.1093/nar/gkw290. Epub 2016 Apr 19. PMID: 27095192; PMCID: PMC4987883

Li. W.; Cerise J. E.; Yang, Y. & Han, H. (2017). Application of t-SNE to Human Genetic Data. Journal of Bioinformatics and Computational Biology. https://doi.org/10.1101/114884

Liu, K.; Goodman, M.; Muse, S.; Smith, J. S.; Buckler, E. D.; Doebley, J. (2003). Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites. Genetics Society of America, 165: 2117–2128.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., & Hein, J. (2005). Bayesian

coestimation of phylogeny and sequence alignment. BMC Bioinformatics, 6, 1–10. https://doi.org/10.1186/1471-2105-6-83

Maizar, Asociación maíz y sorgo argentino, (2023). Estadísticas. http://www.maizar.org.ar/estadisticas.php.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G. J., Buckler, E., & Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences of the United States of America, 99(9), 6080–6084. https://doi.org/10.1073/pnas.052125199

Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. (2016) The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol 14(1): e1002342. doi:10.1371/journal.pbio.1002342.

Morris R and Isikan MH (1964). Cytological studies on inbred lines of maize. Can. J. Genet. Cytol. 6: 508-515

Nielsen, R. (1998). Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theoretical Population Biology, 53(2), 143–151. https://doi.org/10.1006/tpbi.1997.1348

Peña, A.; Bruno, C.; Teich, I.; Fernández, E. & Balzarini, M. (2010). Análisis de conglomerados en la identificación de estructura genética a partir de datos de marcadores moleculares Cluster analysis for identification of genetic structure. Revista Tumbaga, 1(5), 225-237.

Pelleg, D., Moore, A. (1999). Accelerating exact k-means algorithms with geometric reasoning. Proceedings the fifth International conference on knowledge discovery and data meaning. California, United States. ACM Press. Pp. 277-281. doi:10.1145/312129.312248. ISBN 9781581131437 S2CID 13907420

Price, A. L.; Patterson, N. J.; Plenge, R. M.; Weinblatt, M. E.; Shadick, N. A.; Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. , 38(8), 904–909. doi:10.1038/ng1847.

Pritchard, J. K.; Stephens, M.; Donnelly , P. (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics Society of America, Genetics 155: 945–959.

R Core Team (2020). R: A language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., & Melchinger, A. E. (2012). *Genome-wide association mapping of leaf metabolic profiles for dissecting* complex traits in maize. https://doi.org/10.1073/pnas.1120813109.

Ricci G.C.L., Silva N., Pagliarini M.S., Scapin C.A. (2007). Microsporogenesis in inbred line of popcorn (Zea mays L.). ISSN- 1676-5680.

Shu G., Cao G., Li N., Wang A., Wei F., Li T., Yi L., Xu Y., Wang Y. (2021). Genetic variation and population structure in China summer maize germplasm. Sci. 11:8012. doi: 10.1038/s41598-021-84732-6.

Sigaudo, D.; Terré, E. Aporte del maíz a la economía argentina (2022). Informativo semanal, mercados. Bolsa de Comercio de Rosario. ISSN 2796-7824. https://www.bcr.com.ar/es/print/pdf/node/91493

Slatkin, M. (2017). 1. Gene Flow and Population Structure. *Ecological Genetics*, 1–17. https://doi.org/10.1515/9781400887262-003.

Stich, B., & Melchinger, A. E. (2010). An introduction to association mapping in plants. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, *5* (September 2010). https://doi.org/10.1079/PAVSNNR20105039.

Suarez-Salgado, D., Herrera-Camacho, J., Lara-Rivera, A. L., & Parra-Bracamonte, G. M. (2016). Diversidad y origen genético de poblaciones introducidas de bagre de canal (Ictalurus punctatus Rafinesque, 1818), en el centro occidental de México. *Latin American Journal of Aquatic Research*, *44*(3), 525–534. https://doi.org/10.3856/vol44-issue3-fulltext-11.

Van der Maaten, L., Hinton, G., (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research, 1-48. http://www.cs.toronto.edu/~hinton/absps/tsne.pdf.

Vattani, A. (2011). K-means requires exponentially many iterations even in the plane. Discrete and computational Geometry. 45(4): 596-616. doi:10.1007/s00454-011-9340-1.

Anexos

Tabla de los genotipos utilizados en este trabajo con su correspondiente origen y pedigree.

Líneas	Inbred	State/Country	Pedigree
1	33_16	Indiana	Lux Johnson Country white
2	38_11	Indiana	Lux Johnson Country white
3	4226	Illinois	Funk 90 Day
4	4722	Indiana	SA x Ohio yel.
5	A188	Minnesota	[(4-29*64)4-29(4)]
6	A239	Minnesota	A73*A347
7	A272	South Africa	Yellow Boesman
8	A441-5	South Africa	Robyn*Leamming yellow dent
9	A554	Minnesota	[(WD*Wf9)WD(2))]
10	A6	Pioneer	Cuban Flint
11	A619	Minnesota	[(A171*Oh43)Oh43]
12	A632	Minnesota	[(Mt42*B14)B14(3)]
13	A634	Minnesota	[(Mt42*B14)B14(2)]
14	A635	Minnesota	ND203xB14(3)
15	A641	Minnesota	ND203*B14
16	A654	Minnesota	A116*Wf9
17	A659	Minnesota	Minnesota Synthetic 3
18	A661	Minnesota	Minnesota Synthetic AS-A
19	A679	Minnesota	[(A662*B73) B73(3)]
20	A680	Minnesota	[(A662*B73) B73(3)]

21	A682	Minnesota	[(AS-D*Mo17) Mo17 (2)]
22	Ab28A	Alabama	GT152*38-11
23	B10	Iowa	Iowa Stiff Stalk Synthetic
24	B103	Iowa	CIMMYT Pool 41 [Northern Temperate Ranges -1 (NTR-1)]
25	B104	Iowa	BS13(S)C5
26	B105	Iowa	BSSS®C9
27	B109	Iowa	[B73 * BS20(S)C1]B73
28	B115	Iowa	BS11(FR)C9
29	B14A	Iowa	Cuzco*B14(8)
30	B164	Minnesota	Indiana Reid (Pioneer)
31	B2	Missouri	Reid Yellow Dent
32	B37	Iowa	Iowa Stiff Stalk Synthetic
33	B46	Iowa	W22*B10
34	B52	Iowa	Segregating material from a private breeder
35	B57	Iowa	Midland
36	B64	Iowa	41.2504B*B14(3)
37	B68	Iowa	41.2504B*B14(3)
38	B73	Iowa	Iowa Stiff Stalk Synthetic C5
39	B73Htrhm	Iowa	Ht1 and rhm1 conversion of B73
40	B75	Iowa	Iowa Corn Borer Synthetic No. 3 (BSCB3)
41	B76	Iowa	[(CI.31A*B37)B37]
42	B77	Iowa	Pioneer Two-Ear composite(BS11)

43	B79	Iowa	Iowa Two-Ear Synthetic No.1(BSTE)
44	B84	Iowa	BS13(S2)C0
45	B97	Iowa	BSCB1®C9
46	C103	Connecticut	Lancaster Surecrop (from Noah Hershey)
47	C123	Connecticut	C102*C103
48	C49A	Minnesota	Minn 13
49	CH701-30	Canada - Harrow	unknown
50	CH9	Canada - Harrow	Funk's G176
51	CI_7	USDA	[(L317*33-16)33-16(2)]
52	CI187-2	USDA	Krug
53	CI21E	USDA	Hy*C.I.21
54	CI28A	USDA	Recovered Blight Resistance B2
55	CI31A	USDA	Midland OP
56	CI3A	USDA	[(C.I.3*related inbred) C.I.3 (2)]
57	CI64	USDA	K64*Mo21A
58	CI66	USDA	L97*K55 (2)
59	CI90C	USDA	CI90A = L97*M14
60	CI91B	USDA	L97*A71
61	CM105	Canada-Morden	CMV3xB14(2)
62	CM174	Canada-Morden	CMV3xB14(2)
63	CM37	Canada-Morden	KE3
64	CM7	Canada-Morden	W85*CMV3
65	CML10	Mexico	Pop.21 = Tuxpeño

66	CML103	Mexico	Pop. 44
67	CML108	Mexico	Pop. 44
68	CML11	Mexico	Pop. 21
69	CML14	Mexico	Pop. 21 = Tuxpeño Sequia
70	CML154Q	Mexico	Pop. 62
71	CML157Q	Mexico	Pop. 62
72	CML158Q	Mexico	Pop. 62
73	CML218	Mexico	EV = Streak resist. source
74	CML220	Mexico	EV = Streak resist. source
75	CML228	Mexico	Suwan-1/SR
76	CML238	Mexico	LB/SR
77	CML247	Mexico	Pool 24 (Tuxpeño)
78	CML254	Mexico	Pop. 21 = Tuxpeño Sequia
79	CML258	Mexico	Pop. 21
80	CML261	Mexico	Pop. 21
81	CML264	Mexico	Pop. 21 (related to CML10, CML11)
82	CML277	Mexico	Pop. 43 = La Posta (Tux.)
83	CML281	Mexico	Pop. 43 = La Posta (Tux.)
84	CML287	Mexico	Pop. 24 = Antigua-Ven. (Tuson/Tux.)
85	CML311	Mexico	Pop. 500 = Inter. Mat. Wh. Dent Mix
86	CML314	Mexico	Pop. 600 = Late Mat. Wh. Dent Mix
87	CML321	Mexico	Pop. 502 = Mex. DK + US
88	CML322	Mexico	Recyc. US + Mex
89	CML323	Mexico	Pop. 33 Amar. Subtrop. Mix

90	CML328	Mexico	Recycled Lines
91	CML331	Mexico	Suwan/Pop.47 * Mp78:518
92	CML332	Mexico	Suwan/Pop.47 * Mp78:518
93	CML333	Mexico	Pop. 590
94	CML341	Mexico	Pop. 43 = La Posta (Tux.)
95	CML38	Mexico	Pop. 32
96	CML45	Mexico	Pop. 43 = La Posta (Tux.)
97	CML5	Mexico	Pop. 21
98	CML52	Mexico	Pop. 79 = STA ROSA
99	CML61	Mexico	Pop. 21
100	CML69	Mexico	Pop. 36 = Cogollero (Caribbean)
101	CML77	Mexico	Pool 32 = Subtrop. Wh. Mix
102	CML91	Mexico	Pop. 42 Northern Temp./German Mix
103	CML92	Mexico	Pop. 42 Northern Temp./German Mix
104	CMV3	Minnesota	A21*W185
105	CO106	Canada-Ottawa	University of Wisconsin CR11
106	CO125	Canada-Ontario	unknown
107	CO255	Canada-Ottawa	INRA 258
108	DE_2	Delaware B	P3140*P3751
109	DE_3	Delaware B	P3140
110	DE1	Delaware B	P3140*P3751
111	DE811	Delaware	[B68*[B37Ht*(C103*Mp3204 double cross) Selection]]

112	E2558W	South Africa	N6*M162W^3
113	EP1	Spain	Spanish population 'Lizargarate'
114	F2834T	South Africa	Teko Yellow
115	F44	Florida	Smith (Old Florida variety)
116	F6	Florida	Hastings white Prolific * Florida flint
117	F7	France-Peronne	OP Lacaune
118	GA209	Georgia	T61*NC37
119	GT112	Georgia	Multiple cross (includes Whatley, Cuban, Garrick, Creole, and 12% other)
120	H105W	Indiana	33-16*A632(3)
121	H49	Indiana	[(Wf9*L97)Wf9]
122	H84	Indiana	[B37*GE440)Ht Ht
123	H91	Indiana	[B37*GE440)B14(4)]Ht Ht
124	H95	Indiana	Oh43*C.I.90A
125	H99	Indiana	Illinois Synthetic 60C
126	Hi27	Hawaii	[CM104(India)*MV source]BC6
127	Hp301	Indiana	Supergold
128	Hy	Illinois	Illinois High Yield
129	I137TN	South Africa	Natal Yellow Horsetooth x Teko Yellow
130	I205	Iowa	Iodent
131	I29	Iowa	unknown white pearl popcorn; Received from Ashman as I29
132	IA2132	Iowa	[(TSR * 45) * 4329]

133	IA5125	Iowa	[(IP39*Tendermost)*IP39]
134	IDS28	Iowa	Yellow Pearl
135	IDS69	Iowa	South American Popcorn
136	IDS91	Iowa	South American Popcorn
137	Il101	Illinois	Il14 x unknown yellow sweet
138	Il14H	Illinois	White Narrow Grain Evergreen
139	Il677a	Illinois	[(Bolivia 1035*IL44b)*IL422a]
140	K148	Kansas	Yellow selection No. 1 (Pride of Saline, yellow strain)
141	K4	Kansas	Kansas Sunflower
142	K55	Kansas	Pride of Saline
143	K64	Kansas	Pride of Saline
144	Ki11	Thailand	Suwan 1(S)C4-S8-18-7
145	Ki14	Thailand	Suwan 1(S)C4-S8-19-5
146	Ki2021	Thailand	DK version of Ki9; Suwan 1(S)C4-S8-16-7
147	Ki21	Thailand	Pacific 9-S8-45
148	Ki3	Thailand	Suwan 1(S)C4-S8-5-3
149	Ki43	Thailand	Suwan 3(S)C3-S7-138
150	Ki44	Thailand	KS 6(S)C2-S7-366
151	Ky21	Kentucky	Boone County White
152	Ky226	Kentucky	NCLauDA*Coahuila 8
153	Ky228	Kentucky	Pride of Saline
154	L317	Iowa	Lancaster surecrop (from Noah Hershey)

155	L578	Louisiana	Unknown
156	M14	Illinois C	Lancaster * A, where A is a line from Funk's Yellow Dent
157	M162W	South Africa	K64R**2 x B1138T
158	M37W	South Africa	21A**2 x Jellicorse
159	MEF156-5	Maine?	Unknown
160	Mo17	Missouri	C.I.187-2*C103
161	Mo18W	Missouri	Wf9*Mo22(2)
162	Mo1W	Missouri	[Mo22*Wf9(2)]
163	Mo24W	Missouri	(K10*K49/Ziler Hi-cob) (pipe corn)
164	Mo44	Missouri	Mo22*Pioneer Mexican Synthetic 17
165	Mo45	Missouri	Race Negro de Tierra Caliente (Guatemala)
166	Mo46	Missouri	Race Cravo Paulista (Brazil)
167	Mo47	Missouri	Race Candela (Ecuador)
168	MoG	Missouri	Mastadon variety from Pennsylvania
169	Mp339	Mississippi	T61*Hill Yellow Dent
170	MS1334	Michigan	[(Golden glow * Maize Amargo)*Golden Glow]
171	MS153	Michigan	Iowa stiff stalk synthetic
172	MS71	Michigan	A619*R168
173	Mt42	Minnesota	Minnesota No.13 (Owen's)
174	N192	Nebraska	CM105*B73
175	N28Ht	Nebraska	N28= BSSS

176	N6	Nebraska	Hays Golden
177	N7A	Nebraska	Oh07*Stiff stalk Synthetic
178	NC222	North Carolina	Jarvis Golden Prolific
179	NC230	North Carolina	K55*Yellow line or inbred
180	NC232	North Carolina	[(T204*Low Ear outcross) T204 (2)]
181	NC236	North Carolina	[NC7 (Huffman Variety*Illinois Low Ear)]
182	NC238	North Carolina	[(GT112*NC601)GT112]
183	NC250	North Carolina	[(Nigeria Composite ARb*B37)B37)
184	NC258	North Carolina	TZ(2)*[(NC248*246)*C103]
185	NC260	North Carolina	[(Mo44*Mo17)Mo44(3)]
186	NC262	North Carolina	TZ (TZ=McNair 14*18)
187	NC264	North Carolina	[(SC76*Gaspe)Gaspe]SC76(3)
188	NC290A	North Carolina	McNair inbred lines 14*18 (largely of C103 origin); sister line of NC290
189	NC294	North Carolina	[(B73*NC250) B73]
190	NC296	North Carolina	PioneerX105A * H-5
191	NC296A	North Carolina	PioneerX105A * H-5
192	NC298	North Carolina	PioneerX105A * H-5 * Agroceres155
193	NC300	North Carolina	PioneerX105A * H-5 * PioneerX306B
194	NC302	North Carolina	(PioneerX105A*H5)*H101
195	NC304	North Carolina	(H5*PioneerX105A)*H101
196	NC306	North Carolina	(B73*NC250)*B73
197	NC310	North Carolina	improved B73-type derived from NC250*B73^3

198	NC314	North Carolina	B73*NC250
199	NC318	North Carolina	[(SC76*B52)SC76(3)]
200	NC320	North Carolina	[(SC76*B52)SC76(3)]
201	NC324	North Carolina	B73*NC250
202	NC326	North Carolina	[(B73*NC250)*B73(3)]
203	NC328	North Carolina	[(B73*NC250)*B73(3)]
204	NC33	North Carolina	Weekley's Improved
205	NC336	North Carolina	PioneerX105A*H5
206	NC338	North Carolina	[PioneerX105A*H5] * [Pioneer304B*Agroceres504]
207	NC340	North Carolina	PioneerX105A * [Pioneer306B*H5]
208	NC342	North Carolina	McNair inbreds 14*18(of Coker 811A x C103 origin)
209	NC344	North Carolina	TZ(2)*[(NC248*246)*C103]; Sister line of NC258
210	NC346	North Carolina	PioneerX105A*H5
211	NC348	North Carolina	PioneerX105A * H-5 * Agroceres155
212	NC350	North Carolina	(H5*PioneerX105A)*H101
213	NC352	North Carolina	PioneerX105A*H5
214	NC354	North Carolina	[PioneerX105A*H5] * [PioneerX304A*H101]
215	NC356	North Carolina	TROPHY SYN
216	NC358	North Carolina	TROPHY SYN
217	NC360	North Carolina	Agroceres155*PioneerX105A/NC262
218	NC362	North Carolina	Agroceres155*PioneerX105A/NC262

219	NC364	North Carolina	Agroceres155*PioneerX105A/NC262
220	NC366	North Carolina	FLA Syn
221	NC368	North Carolina	[B73*NC250]/[(B73*NC250)*B73]
222	ND246	North Dakota	W755*W771
223	Oh40B	Ohio	Eight line composite of Lancaster Surecrop lines
224	Oh43	Ohio	Oh40B*W8
225	Oh43E	Ohio	ERF/Oh43; ERF=LeamingxReid + other Pioneer inbreds
226	Oh603	Ohio	« Syn of Va58, OhS3267, H95, Va26, Coas. Trop. FL. »
227	Oh7B	Ohio	[(Oh07*38-11)Oh07]
228	Os420	Iowa	Osterland yellow dent
229	P39	Indiana	Purdue Bantam
230	Pa762	Pennsylvania	Oh43*Pa70L
231	Pa875	Pennsylvania	Wf9 Synthetic (original)
232	Pa880	Pennsylvania	Wf9 Synthetic C3
233	PA91	Pennsylvania	(Wf9*Oh43)S4*[(38-11*L317)38-11]S4
234	R109B	Illinois	Snelling Corn Borer Synthetic
235	R168	Illinois	Illinois Synthetic 60C
236	R177	Illinois	Germplasm 230B(Snelling Corn Borer Synthetic)
237	R229	Illinois	[(479*B73)B73(2)]S6; 479 is a Brazilian inbred of Tuxpeño type
238	R4	Illinois	Funk Yellow dent

239	SA24	Indiana	South American Popcorn
240	SC213R	South Carolina	(GT112*NC33)GT112] backcrossed with a CMS-T restorer
241	SC357	South Carolina	[(Whately yellow * Tennessee Redcob) Young]
242	SC55	South Carolina	[(L501*L503)*(L548*L569)]
243	SD40	South Dakota	Pioneer hybrid 3709
244	SD44	South Dakota	SDp309*SD30
245	Sg1533	Indiana	Super gold
246	Sg18	Indiana	Super gold
247	T232	Tennessee	Jellicorse*Teko yellow
248	T234	Tennessee	[T111*RB.L*III.A)]T111(4)
249	T8	Tennessee	Jarvis Golden Prolific
250	Tx303	Texas	Yellow Surcropper
251	Tx601	Texas	Yellow Tuxpan
252	Tzi10	Nigeria	Tlaltizapan 7844 x TZSR
253	Tzi11	Nigeria	Mo17 x RppSR
254	Tzi16	Nigeria	PI 540747 = N28/RPPTZSR-Y
255	Tzi18	Nigeria	Sete Lagoas 7728 x TZSR
256	Tzi25	Nigeria	[(B73*RPPSR-TZ)*B73(2)]
257	Tzi9	Nigeria	SIDS7734/TZSR
258	U267Y	South Africa	WF9r*Mex.155^3
259	VA102	Virginia	Va59*Va60
260	Va14	Virginia	[(VaCBS selection*Va17)Va17]
261	Va17	Virginia	Wf9*T8

262	Va22	Virginia	Va17*C103 backcross
263	Va35	Virginia	[(C103*T8)T8]
264	Va59	Virginia	[(C103*T8(2))*(K4*C103(2))]
265	Va85	Virginia	Virginia Long Ear Synthetic
266	Va99	Virginia	Oh07B*Pa91
267	VaW6	Virginia	unknown
268	W117Ht	Wisconsin	W117=643*Minnesota No.13
269	W153R	Wisconsin	[(la153*W8)la153]
270	W182B	Wisconsin	WD*W22
271	W22	Wisconsin	III.B10*W25
272	W22_R_r_	unknown	unknown
273	WD	Wisconsin	Wisconsin No. 25
274	WF9	Indiana	Reid yellow dent (Indiana station strain)
275	Yu796_NS	unknown	unknown

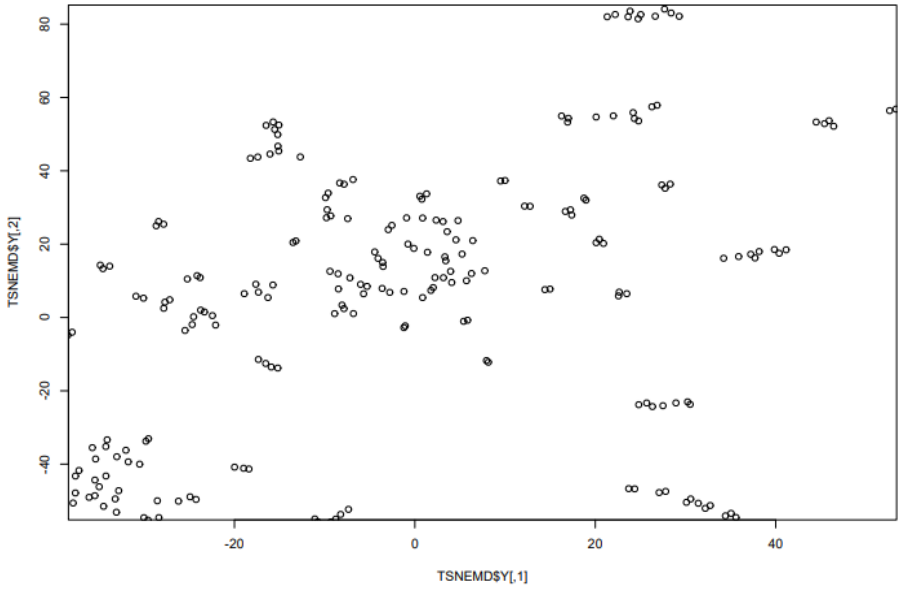


Fig.17. Gráfico de dispersion obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012. Representación de los datos con perplexity=5 y 20.000 iteraciones. Distribución del agrupamiento de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.

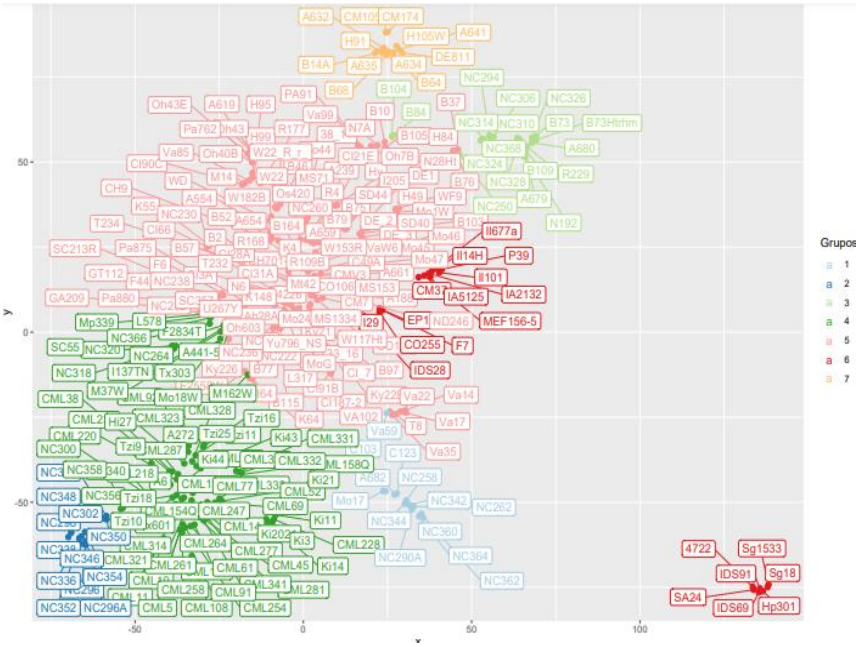


Fig.18. Gráfico de dispersión ggplot obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012.Representación de los datos con perplexity=5 y 20.000 iteraciones. Distribución del agrupamiento y coloración de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.

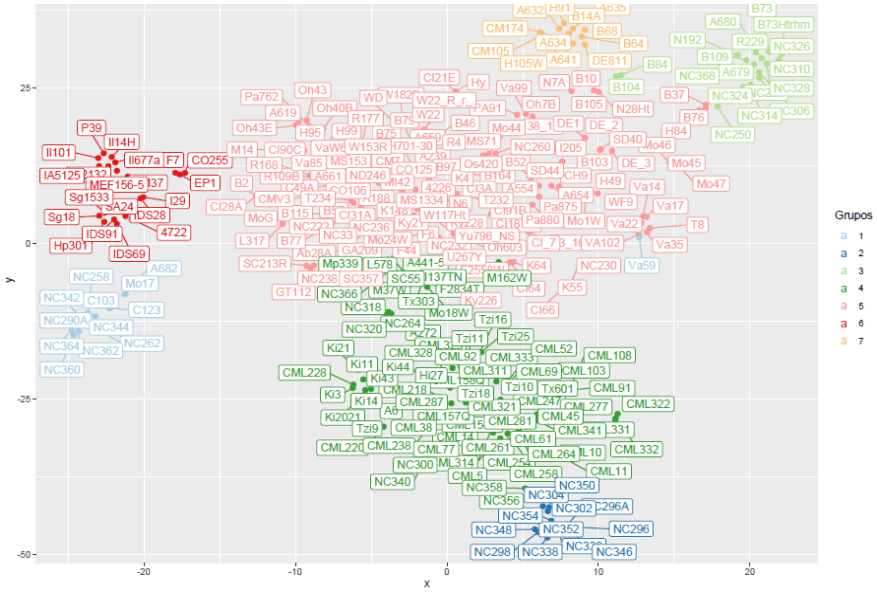


Fig.19. Gráfico de dispersión obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al.,* 2012.Representación de los datos con perplexity=15 y 20.000 iteraciones. Distribución de la dispersión de los datos reducidos de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.

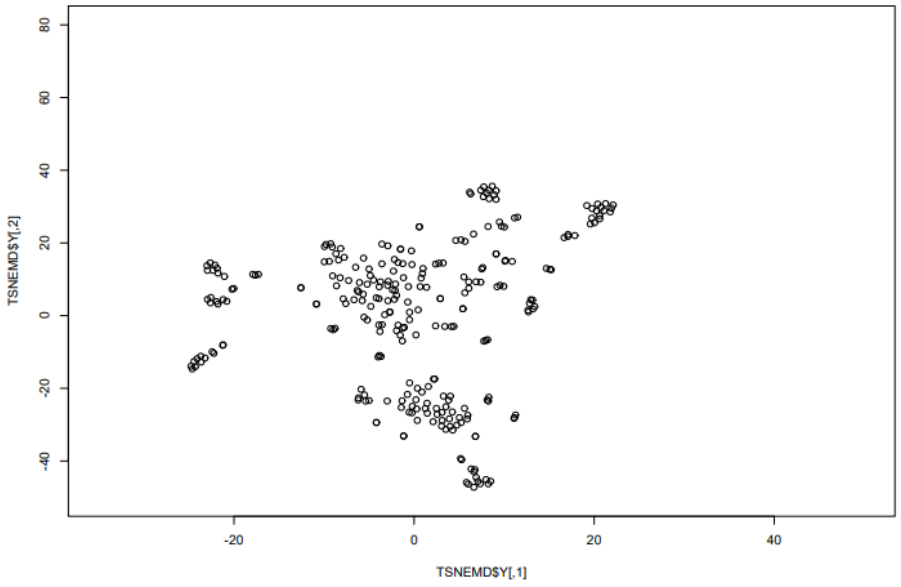


Fig.20. Gráfico de dispersión obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook *et al*., 2012.Representación de los datos con perplexity=15 y 20.000 iteraciones. Distribución del agrupamiento y coloración de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.

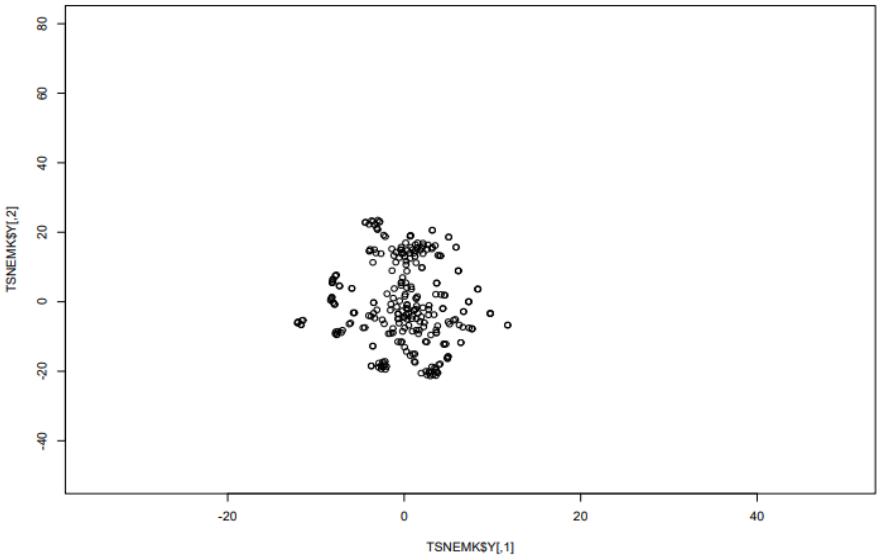


Fig.21. Gráfico de dispersión obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook et al., 2012.Representación de los datos con perplexity=30 y 20.000 iteraciones.

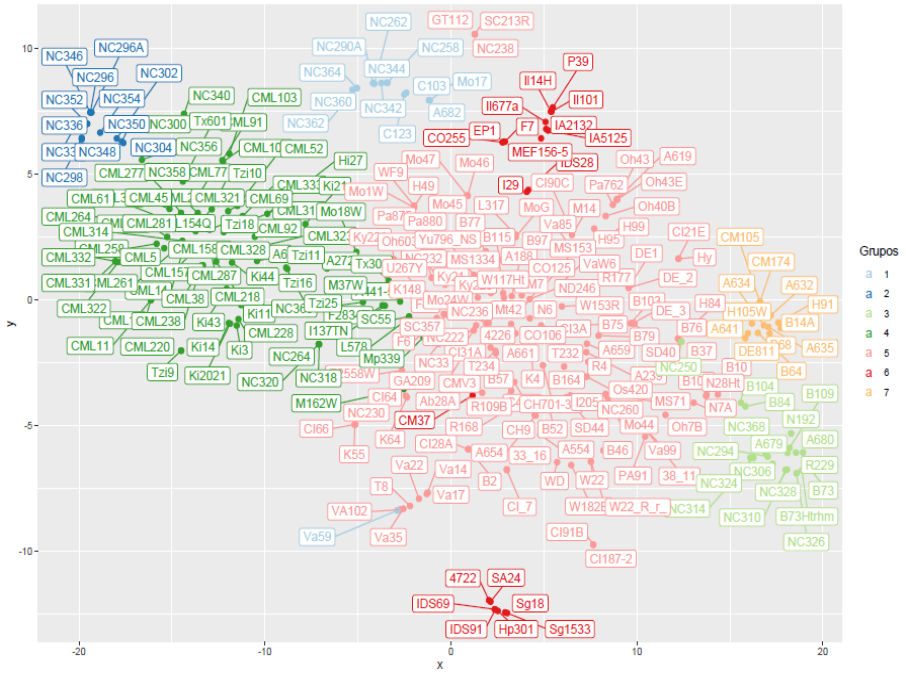


Fig.22. Gráfico de dispersión obtenido a partir de la matriz de distancia de las 275 líneas endocriadas de maíz de un panel público previamente genotipadas por Cook et al., 2012.Representación de los datos con perplexity=30 y 20.000 iteraciones. Distribución del agrupamiento y coloración a partir de las subpoblaciones obtenidas a partir de fastSTRUCTURE con un K=7.