“PRUEBA DE MODELOS Y MÉTODOS PARA EL MAPEO DE ASOCIACIÓN DE LA RESPUESTA AL FOTOPERÍODO EN MAÍZ”

Trabajo Final de Grado del alumno

CHALELA, LUCIANO

Este trabajo ha sido presentado como requisito

para la obtención del título de

INGENIERO AGRÓNOMO

Escuela de Ciencias Agrarias, Naturales y Ambientales.

Universidad Nacional del Noroeste de la Provincia de Buenos Aires.

Pergamino, ……

“PRUEBA DE MODELOS Y MÉTODOS PARA EL MAPEO DE ASOCIACIÓN DE LA RESPUESTA AL FOTOPERÍODO EN MAÍZ”

Trabajo Final de Grado del alumno

CHALELA, LUCIANO

Aprobada por el Tribunal Evaluador de Tesina

…… …… ……

…… ……..

(Ing. Agr. Sofía E. Olmos) (Ing. Agr. Susana Pistorale)

Directora Co-directora

Escuela de Ciencias Agrarias, Naturales y Ambientales.

Universidad Nacional del Noroeste de la Provincia de Buenos Aires.

Pergamino, ……

AGRADECIMIENTOS

A mi familia por su apoyo, y por ayudarme siempre a seguir adelante.

A los profesores de esta carrera, por los conocimientos transmitidos e inculcados durante toda la carrera.

A los profesionales que me ayudaron a llevar a cabo esta tesina: mi directora, Ingeniera Agrónoma Sofía Olmos, por su colaboración y aguante, y a mi co-directora, Ingeniera Agrónoma Susana Pistorale, por sus aportes durante este trabajo.

A todos mis amigos por estar siempre.

Índice

Agradecimientos……. 3

Resumen…….. 7

Introducción……. 8

Hipótesis…….……..……...12

Objetivos……..……..……...12

1-Objetivo general……..……12

2-Objetivos específicos……..…...12

Materiales y Métodos ……... 13

1-Recopilación de datos genotípicos……. 13

2-Caracterización fenotípica…… 13

3-Estructura genética…… 13

4-Parentesco……. 16

5-Mapeo de QTLs……. 16

Resultados…… 20

1-Estructura genética……20

2-Parentesco…….. 26

3-Mapeo de QTs……... 27

Discusión…….. 34

Conclusiones…… 38

Anexos…….. 39

1-Anexo 1 Resultados de estructura genética del panel de líneas de Maíz de USDA……39

2-Anexo 2 Resultados de Structure con SSR……..…. 40

3-Anexo 3 Resultados de Structure con SNPs…….. 47

4-Anexo 4 Lista de líneas con sus códigos……...…. 54

Bibliografía…… 56

RESUMEN

El mapeo de asociación, en sus dos modalidades de estudio de asociación del genoma completo *(GWAS (Genome-wide Association Study)* como del gen candidato, es un método de mapeo de *loci* de herencia cuantitativa (QTL, *Quantitative Trait Loci*) que tienen como fin asociar fenotipos a marcadores moleculares y descubrir así marcadores útiles para el mejoramiento genético. A diferencia del mapeo tradicional, la población de mapeo está formada por un panel de líneas que suelen estar emparentadas y en conjunto presentan estructura genética, propiedades que promueven la ocurrencia de falsos positivos en las asociaciones fenotipo-marcador. El objetivo de este trabajo fue utilizar un caso de mapeo de asociación de la respuesta fotoperiódica del tiempo a floración en maíz (gen candidato *ZmCCT*) como modelo de estudio de caracteres agronómicos complejos, de manera de probar diferentes metodologías estadísticas y sets de datos para verificar reproducibilidad de resultados y ajustes de modelos de asociación. Como resultado se logró establecer en el panel de maíz, relaciones de estructura genética similares a las originales tanto con polimorfismos de nucleótido simple (SNP, *Single Nucleotide Polymorphism*) como con microsatélites (SSR, *Single Sequence Repeat*). El método del gen candidato permitió identificar a los dos SNP ligados al gen candidato *ZmCCT* del cromosoma 10, según lo informado previamente. El GWAS, por el contrario, empleando otro set de datos y población a lo utilizado originalmente, no logró identificar SNP suficientemente cercanos a *ZmCCT.* Los modelos de análisis generalizado lineal (GLM) y lineal mixto (MLM) brindaron diferentes cantidades de asociaciones significativas, las que a su vez variaron según el ajuste por estructura y parentesco. Los datos fenotípicos promediados por ambientes redujeron el número de asociaciones significativas comparado con los datos fenotípicos por ambiente individual. Fue posible una reproducción parcial de los resultados originales informados según los tipos de marcadores, poblaciones, modelos y ajustes utilizados.

INTRODUCCIÓN

Históricamente la dieta humana se sustenta en tres cereales arroz, maíz y trigo. En la campaña 2013/2014 el maíz en Argentina cubrió 6.098.885 de ha (SIIA, 2014). La mayor parte de la producción actual se concentra en la región Pampeana (Córdoba, Buenos Aires, Santa Fe, La Pampa y Entre Ríos, en ese orden, con 76% de la superficie) en suelos con muy buena aptitud agrícola (Peretti, 2003). Aún en las mejores campañas, los rendimientos medios (por ejemplo, de 6.604 kg ha^{-1} en 2012/2013) están alejados de los rendimientos potenciales (15 – 19.000 kg ha^{-1} (Fischer y Edemeades, 2010). Las principales causas de la brecha de rendimiento son el déficit hídrico y el insuficiente abastecimiento de nutrientes (principalmente nitrógeno). Las deficiencias de nitrógeno del suelo pueden compensarse con la fertilización nitrogenada pero el impacto ambiental negativo de las pérdidas de nitrógeno al sistema natural sumado al alto costo del fertilizante (Cassman et al., 2003) requieren otras soluciones conjuntas para lograr el aumento del rendimiento como ser el mejoramiento genético. Así, mediante el mejoramiento genético se pueden seleccionar genotipos que maximicen el uso de los recursos ambientales y producir materiales tolerantes tanto a estreses bióticos como abióticos.

Para el mejoramiento genético del maíz se utiliza la diversidad existente con el fin de mejorar líneas élites que son la base de híbridos comerciales. El maíz, en particular el de origen tropical, tiene una respuesta de días largos, esto es, su floración se retrasa cuando la longitud del día supera determinado umbral. El empleo de genotipos de origen tropical en programas de mejoramiento de maíces templados tiene el inconveniente de que su cultivo en latitudes altas resulta en una gran demora de los tiempos a floración. La introducción de la sensibilidad al fotoperíodo vuelve improductivas a las líneas templadas que adquieren estas características (Goodman, 2004). La identificación de las regiones genómicas responsables de esta sensibilidad facilitará la eliminación de este rasgo en las líneas templadas. Se ha propuesto en maíz que la cantidad relativa de *loci* de herencia cuantitativa (del inglés *Quantitative Trait Loci*, QTLs) responsables de la variación en la

respuesta fotoperiódica de los tiempos de floración sería menor en comparación a los QTLs que afectan a la duración del ciclo *per se* (precocidad) cuando el fotoperiodo no es limitante (días cortos menores a un umbral). Así, los QTLs de respuesta fotoperiódica presentarían un efecto genético simple y comprenderían un pequeño conjunto de QTLs con gran variación fenotípica que afectarían la precocidad en días largos y serían más fácilmente expresables en determinados fondos genéticos (Hung *et al*., 2012).

Entre los varios QTLs o genes responsables del tiempo a floración *per se* encuentran aquellos QTLs en el cromosoma 10 Ducrocq *et al*., (2009), cromosomas 1,6 y 7 Steinhoff *et al*., (2012) y los siguientes genes candidatos: Dwarf8 Thornsberry *et al*., (2001), Vgt1 Ducrocq *et al*., (2008) y Tb1 Camus-Kulandaivelu *et al*., (2008). Por otro lado, los QTLs responsables de la respuesta fotoperiódica de los tiempos a floración en maíz fueron identificaron en diferentes cromosomas: 10 (Mu et al., 2009), 1, 8, 9 y 10 (Coles et al., 2010), 4, 9 y 10 (Wang et al., 2012), 1,8 y 10 (Xu et al., 2012). En cuanto a los genes candidatos recientemente se identificó en el cromosoma 10 al gen *ZmCCT* como el gen que posee mayor efecto sobre la respuesta al fotoperíodo del tiempo a floración en maíz (Hung *et al*., 2012).

La identificación de QTLs o genes responsables de un fenotipo determinado se basa en tres pilares: el genotipado, el fenotipado y las pruebas estadísticas para encontrar las asociaciones (por desequilibrio de ligamiento) estadísticamente significativas entre un marcador molecular y el fenotipo estudiado (Darvasi *et al*., 1993). Por otro lado, se diferencian dos metodologías de estudio de QTLs según el tipo de población de mapeo empleada (Semagn *et al*., 2010). La primera es el mapeo de QTL convencional que emplea una población segregante biparental. La segunda, es el mapeo por asociación que utiliza poblaciones naturales o bien colecciones de líneas que resultan del mejoramiento genético y selección. El mapeo por asociación requiere a su vez de pruebas adicionales como son la estimación de la estructura poblacional y las relaciones de parentesco de manera de ajustar el modelo de análisis más apropiado que minimice la ocurrencia de asociaciones marcador-fenotipo falsas, dadas por error tipo I (Stich *et al*., 2008). Por último, el mapeo por asociación se puede realizar a su vez mediante dos métodos, el análisis de

asociación del genoma completo (del inglés *Genome-Wide Association Study*, GWAS) y el método del gen candidato donde se prueban marcadores seleccionados de regiones candidatas de incluir un gen responsable de un fenotipo determinado.

Existen iniciativas públicas de generar paneles de asociación genotipados en alta densidad para que puedan intercambiarse entre equipos colaboradores y hacer más eficiente así el proceso de fenotipado. En maíz se ha desarrollado recientemente una población especial denominada población de mapeo anidada (del inglés *Nested Association Mapping population*, NAM) que resultó del cruzamiento de una línea de referencia (B73) por otras 26 líneas genéticamente contrastantes, a su vez, pertenecientes a diferentes grupos heteróticos y que han sido genotipadas con alta densidad de marcadores (Gore *et al*., 2009). Esta población constituida por unas 5.000 familias de líneas recombinantes endocriadas (del inglés *Recombinant Inbred Lines*, RILs) posee múltiples ventajas estadísticas con respecto al panel de asociación como por ejemplo mayor precisión y poder de las pruebas de asociación marcador-fenotipo (Yu *et al*., 2008) resultado de la mayor cantidad de recombinaciones que pueden acumular bajo distintos fondos genéticos.

Cada una de estas tres poblaciones de mapeo mencionadas puede requerir de la aplicación de modelos de análisis estadísticos diferentes o bien compartidos y la elección del método que mejor ajuste puede resultar empírica según los supuestos en que se base el análisis (Stich *et al*., 2008). En el caso de la población NAM se pueden utilizar los métodos de mapeo de QTL comunes o bien otros como el método de análisis conjunto de poblaciones basado en regresiones paso por paso (Buckler *et al*., 2009). Además, permite el análisis GWAS y el método del gen candidato ya que la conformación del diseño de la población permite por sí misma el control de la estructura genética y las relaciones de parentesco (Hung *et al.,* 2012).

Adicionalmente, en maíz se dispone de un panel de asociación de 282 líneas endocriadas de origen diverso, genotipado inicialmente con 94 marcadores microsatélites (del inglés *Single Sequence Repeats,* SSRs) (Liu *et al*., 2003), el cual fue adicionalmente genotipado con otros marcadores como

polimorfismos de nucleótido simple (del inglés *Single Nucleotide Polymorphisms*, SNPs) a una densidad de 55k que permite el análisis GWAS o bien el análisis de genes candidatos, basado en un modelo mixto donde es necesario previamente ajustar el modelo por la estructura genética y las relaciones de parentesco presentes en el panel porque las líneas, generalmente comerciales, presentan a lo largo de su historia de desarrollo cierto grado de emparentamiento y correlación entre caracteres que ocasiona falsas asociaciones marcador-fenotipo (Yu *et al*., 2006).

La población NAM y el panel de asociación posibilitaron resolver en maíz QTLs a nivel de genes para las siguientes características: resistencia a tizón foliar (causado por *Exserohilum turcicum*) en (Poland *et al*., 2011), resistencia a tizón por bipolaris (causado por *Bipolaris maydis*) en (Kump et al., 2011), arquitectura de planta en (Tian *et al*., 2011), composición de grano (Cook *et al*., 2012) y sensibilidad al fotoperiodo del tiempo a floración en (Hung *et al*., 2012).

En el caso particular de (Hung *et al.,* 2012) el método de análisis de QTL se realizó en primera instancia por GWAS empleando la población NAM y un genotipado de SNP de alta densidad. El GWAS permitió identificar QTLs en casi todos los cromosomas, los mismos colocalizaron con genes candidatos previamente estudiados y uno del cromosoma 10 manifestó la mayor variación fenotípica. En el mismo trabajo se realizó un mapeo fino posterior mediante otra población derivada del cruzamiento de teosinte (gran respuesta al fotoperiodo) con la línea templada W22 (prácticamente insensible al fotoperiodo) donde se pudo identificar dentro del QTL del cromosoma 10 (que presentaba mayor variación fenotípica) un único gen candidato, el gen *ZmCCT* posicionado a aproximadamente 94.2 Mbp en las coordenadas de la versión 2 del genoma de B73. Adicionalmente, el método del gen candidato utilizando el panel de asociación (y algunas accesiones de teosinte) y marcadores de la secuencia de *ZmCCT* logró la asociación de la respuesta al fotoperiodo de los tiempos a floración en maíz más precisa, con dos SNPs localizados próximos al gen *ZmCCT*, pero estructuralmente por fuera del mismo, ubicados en la región río arriba del gen. Uno de ellos, PZE1094244760, se ubica en 94.244.760 bp en V1 (94.420.967 bp en V2) con un valor aditivo (2*a) de 25,4 y un valor p de $3.5E^{-111}$. Mientras que el otro PZE1095719172 se localizó 1,5 Mbp de *ZmCCT*,

en 95.719.172 bp en V1 (95.925.762 pb en V2), valor aditivo (2*a) de 26.3 y valor p=9,13E^{-102}. Debido a que el mismo set de datos moleculares del GWAS de la población NAM no se encuentra genotipado en el panel de asociación, no es posible lograr la reproducibilidad completa del GWAS de (Hung *et al*., 2012). Aunque sí se dispone de los marcadores del gen candidato *ZmCCT*. De esta forma los sets de genotipado independientes disponibles para el panel de asociación pueden ser utilizados y modelados junto a los datos fenotípicos a fin de estudiar factores que condicionan la reproducibilidad y la precisión de la asociación marcador-fenotipo en los análisis GWAS y gen candidato y lograr así una aproximación a los resultados obtenidos por (Hung et al.,2012).

Estas metodologías de análisis son de desarrollo relativamente reciente en especies de plantas, los estudios de simulación permitirán transferir conocimientos para el estudio de poblaciones locales que cuentan con un genotipado y fenotipado incipiente.

HIPÓTESIS

La utilización de un panel de líneas de maíz y una base de datos públicas, con información geno y fenotípica para respuesta al fotoperíodo, permitirá la prueba de métodos y la bondad de ajuste de modelos de análisis mixtos para la generación de conocimientos y su posterior adaptación al estudio de casos locales.

OBJETIVO GENERAL

Lograr un mejor entendimiento de las metodologías estadísticas y modelos aplicados al mapeo de asociación de características de interés agronómico en vegetales tomando como modelo de estudio la identificación de la base genética de la respuesta al fotoperiodo de maíz desarrollado por (Hung *et al*., 2012) y aplicando los métodos GWAS y de gen candidato con diferentes parámetros de simulación de los empleados en el trabajo original.

OBJETIVO ESPECÍFICO

Probar el efecto de la combinación de dos metodologías de genotipado (marcadores microsatélites y SNPs) para la estimación de estructura genética y relaciones de parentesco, y de modelos de análisis de asociación "marcador-respuesta al fotoperiodo" empleando los métodos GWAS y de gen candidato.

MATERIALES Y MÉTODOS

Se analizó un panel de asociación compuesto por 260 líneas de maíz que representan una muestra de las líneas más importantes de los Estados Unidos, Europa, Canadá, Sudáfrica y Tailandia, junto con las líneas del Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) y el Instituto Internacional de Agricultura Tropical (IITA). Este material representa una muestra de la diversidad disponible entre las líneas actuales e históricas para maíces templados pertenecientes a dos grupos heteróticos (166), tropicales y subtropicales (82) y fue previamente estructurada genéticamente por (Liu *et al*., 2003) mediante 94 marcadores microsatélites (Anexo 1).

Para realizar el presente trabajo se recopiló información sobre el genotipado y el fenotipado de los tiempos a floración del panel de líneas para reanalizar, la estructura poblacional, las relaciones genéticas y el mapeo por asociación mediante el método GWAS y por el gen candidato.

Recopilación de datos genotípicos

El material fue genotipado con diferentes marcadores y densidad de marcadores que permiten estimar estructura, parentesco y en algunos casos la asociación marcador-fenotipo propiamente dicha. Así, el panel de asociación cuenta con: 1) el genotipado con 94 SSRs llevado a cabo por Liu *et al*., (2003), para estimar estructura y parentesco; 2) genotipados de alta densidad con un chip de Illumina de 56.000 marcadores SNPs uniformemente distribuidos en el genoma Cook *et al*., (2012) y otro genotipado ZeaGBSv2.3, los cuales son públicos y se encuentran almacenados en la base de datos www.panzea.org, se pueden utilizar para estimar estructura, parentesco y la asociación, en el

presente trabajo se utilizó este set de datos del cromosoma 10 para el GWAS para realizar la asociación propiamente dicha; 3) un set de 553 SNPs provenientes de genes aleatorios de cobertura genómica (Yu *et al*., 2006), para estimar estructura y parentesco; y 4) 227 variantes SNPs e inserciones-deleciones (INDELs) en una región de 4,5 kb conteniendo al locus *ZmCCT* para aplicar el método de asociación del gen candidato.

Como la información genotípica para las distintas fuentes solo estuvo disponible para 223 de las 260 líneas del panel de (Liu *et al*., 2003), se seleccionaron las 223 líneas comunes para los posteriores análisis.

Caracterización fenotípica

Los datos fenotípicos provinieron del trabajo de Hung *et al*., (2012). En dicho trabajo el panel de asociación junto a otras poblaciones de mapeo RILs se cultivaron en seis localidades del Hemisferio Norte con una gran variación en latitud, desde los 18º N (Puerto Rico) hasta los 43º (Estado de New York, USA) que permitieron lograr fotoperíodos de 11 a 15 horas durante la etapa vegetativa del cultivo y se determinaron los tiempos térmicos para la floración, espigazón y el intervalo floración-espigazón. En general, los ensayos fueron repetidos por dos años en la misma localidad, lográndose en total 11 ambientes. En el trabajo original se mapeó el tiempo a espigazón, para la presente tesina se eligió el carácter tiempo a floración (antesis) de los 11 ambientes (datos BLUP, del inglés *Best Linear Unbiased Predictor)* provistos en Hung *et al*., (2012) y el promedio general ajustado por ambiente (del inglés *Best Linear Unbiased Estimation*, BLUE) de estos 11 ambientes, calculado con TASSEL (Bradbury *et al*., 2007), para el caso del modelo GWAS GLM (modelo generalizado lineal) donde no es posible incorporar datos multiambientales.

Estructura genética

Las líneas se subdividieron en grupos genéticos utilizando el enfoque basado en el modelo Bayesiano implementado en el programa STRUCTURE 2.3.3 (Pritchard *et al*., 2000). Para un número dado de subpoblaciones

(clusters), este método asigna individuos a grupos de manera tal que el desequilibrio Hardy-Weinberg y el desequilibrio de ligamiento se reducen al mínimo. No se utilizó información previa del pedigrí de las líneas para inferir las subpoblaciones. Según lo recomendado por (Pritchard *et al*. 2009), se asumió que el genoma de cada individuo podía estar constituido por la contribución de una o más de las poblaciones inferidas (*Admixture model*) como punto de partida para el análisis de datos. Bajo este modelo, cada individuo deriva parte de la fracción de su genoma de cada una de las subpoblaciones k y, condicional sobre el vector de ascendencia, q (i), se asume que el origen de cada alelo es independiente. Es decir, este modelo supone que todos los marcadores no están ligados y proporciona información independiente sobre la ascendencia de un individuo.

Para cada set de datos i) 94 SSRs y iii) 553 SNPs se evaluó en un rango de k entre 1 a 15 y un total de 3 réplicas para cada valor k. El tiempo de calentamiento (*burn-in*) y el número de iteraciones se establecieron en 100.000.

Para la selección del valor óptimo de k se tuvo en cuenta la corrida que mostró la mayor probabilidad posterior de los datos. Asimismo, los resultados del STRUCTURE fueron evaluados también usando un criterio *ad-hoc* basado en la tasa de cambio del logaritmo de la probabilidad de los datos entre los sucesivos valores de k (Δk) (Evanno *et al*., 2005). Para inferir el verdadero valor de k, estos resultados se compararon además con los registros de pedigrí conocidos y los agrupamientos de referencia de (Liu *et al*., 2003).

Así, para comparar los resultados gráficamente se eligió de cada corrida la repetición que brindó el valor de la probabilidad LnP(D) más cercano a cero (menor valor absoluto). Siendo LnP(D) la estimación del logaritmo de la probabilidad de asignación de los individuos o líneas dentro de cada subpoblación. Luego, las salidas de STRUCTURE se procesaron con el programa online STRUCTURE HARVESTER (Earl y VonHoldt, 2012) que implementa el método de Evanno para seleccionar el número óptimo de las agrupaciones k según el delta k (Evanno *et a*l., 2005; Earl y VonHoldt, 2012). Siendo Δ k calculado como el cociente entre el promedio de la derivada de

segundo orden del logaritmo de la probabilidad a un determinado k y el desvío estándar de esta medida: $\Delta K = mean|L''(K)|/sd[L(K)]$.

Por último, los resultados de agrupamientos se compararon con Liu *et al*., (2003) quienes encontraron con SSR una estructura genética que les permitió dividir a la población (panel de líneas) en:

(NSS) líneas templadas *non-Stiff Stalk* (línea representativa MO17)

(SS) líneas templadas *Stiff Stalk* (línea representativa, B73),

(TS) líneas tropicales o semitropicales,

(*Sweet corn*) maíz dulce,

(*Popcorn*) maíz pisingallo.

Para asignar las líneas a las subpoblaciones se utilizó como referencia los coeficientes de pertenencia de las líneas a los clusters. Así basados en la bibliografía, la línea de corte se estableció en 0,8 y las líneas fueron asignadas en forma manual. Líneas con coeficientes ≥0.80 se consideraron pertenecientes a subpoblaciones discretas, mientras que las líneas con probabilidades <0,80 fueron asignadas como de pertenencia "mixtas".

La matriz de STRUCTURE (Q) con el mejor valor k seleccionado para cada set de datos (QSSRs y QSNPs) se utilizó luego en TASSEL (Bradbury *et al*., 2007) para corregir el efecto de la estructura genética sobre la estimación de la asociación marcador-fenotipo.

Parentesco

Las relaciones de parentesco se estimaron con el algoritmo de (Loiselle *et al.,* 1995) implementado en el programa SPAGeDI 1.4 (Hardy y Vekemans, 2002) y se representaron mediante una matriz de parentesco K (kinship).

SPAGeDi (Análisis del patrón espacial de la diversidad genética) es un paquete informático diseñado principalmente para caracterizar la estructura genética espacial de las poblaciones que permite establecer relaciones genéticas de diferente magnitud y estado a las obtenidas con *STRUCTURE* (Yu *et al*., 2006). Este método se adapta a los individuos diploides heterocigóticos

en el caso de los conjuntos de datos multialélicos y multilocus. Los datos se preparan de manera que cada celda de marcador y genotipo contenga los dos alelos de cada *locus* separado por cualquier carácter no numérico. El grado de covarianza genética causada por efectos poligénicos se define como 0 para un par de individuos que no están relacionados y como positiva para un par de individuos que están relacionados. Los valores negativos de parentesco entre líneas en la matriz resultante se establecen como 0, indicando el valor negativo que un par de genotipos están menos relacionados que un par de individuos de la población tomados al azar.

Las matrices K para cada set de datos (KSSRs y QSNPs) se utilizaron luego en TASSEL para corregir el efecto negativo del parentesco sobre la estimación de la asociación marcador-fenotipo.

Mapeo de QTLs

El análisis de asociación se realizó con el programa TASSEL V. 5 (Bradbury *et al*., 2007) mediante dos métodos: GWAS utilizando el set de SNP para el cromosoma 10: Maize282_imputed_AllZea_GBS_Build_July_2012_FINAL_chr10.hmp223lineas disponible online en el sitio http://www.panzea.org/. y método del gen candidato con 227 marcadores del *locus ZmCCT*.

En primera instancia se analizó gráficamente la distribución de los datos fenotípicos mediante histograma para verificar el supuesto de distribución normal de los datos. Aunque vale aclarar que la verificación del supuesto se realiza sobre los errores del modelo, no sobre los fenotipos observados, por lo cual este análisis se realizó en forma exploratoria.

Se probaron en total ocho modelos combinando los *sets* de datos para la estimación de estructura genética y parentesco y para la asociación marcador-fenotipo, los que se combinaron con datos fenotípicos obtenidos en los 11 ambientes evaluados y los promedios estimados para todos los ambientes (BLUE). Para GLM (modelo generalizado lineal) se utilizó solo el valor

fenotípico BLUE porque GLM de TASSEL no permite analizar datos multiambientales.

En el caso del gen candidato, se observaron específicamente los resultados obtenidos con los dos SNPs ligados previamente asociados al alargamiento de los días a espigazón (Hung *et al*., 2012).

Los casos analizados fueron:

1) Modelo simple: GLM conteniendo solo los marcadores a probar incluidos en el modelo como un factor fijo.

2) GLM (Q): Se incorporó dentro del modelo simple las matrices A) QSSRs y B) QSNPs de STRUCTURE como un factor fijo.

3) MLM (QK): Se utilizó un modelo lineal mixto (MLM) para datos de 11 ambientes y de BLUE, utilizando las matrices QSSRs/QSNPs y KSSRs/KSNPs tratadas como factores fijos y aleatorios, respectivamente: A) QSSRs/KSSRs B) QSSRs/KSNPs C) QSNPs/KSSRs y D) QSNPs/KSNPs.

Cabe aclarar para el resto del texto de que la matriz de parentesco no se incluye estrictamente como factor aleatorio, lo que se incluye es el efecto genético aditivo como aleatorio y en la matriz de varianza covarianza de dichos efectos se incluye la matriz de parentesco (kinship). Así, en forma resumida se la denomina factor aleatorio para este modelo y los siguientes.

Para lograr la convergencia de TASSEL en los casos de análisis de 11 ambientes, que insume muchos recursos computacionales en PC común de 32 bits, se utilizó la opción de compresión de información de la matriz K. Este nivel de subagrupamiento de la población, representado por el nivel de compresión, se estimó indirectamente mediante el cálculo del número de individuos promedios agrupados en forma discreta en cada subpoblación de la matriz Q (coeficiente de pertenencia de cada individuo mayor a 0,8) logrado con los marcadores SNPs o SSRs correspondientes. Así para QSSRs y QSNPs se lograron niveles de compresión de 58 y 41, respectivamente. Estos parámetros se utilizaron cuando se emplearon las matrices KSSRs y KSNPs, respectivamente, independientemente de la matriz Q utilizada.

En el caso de BLUE (modelo 3, caso 1= 3.1) no se requirió la opción de compresión para obtener resultados por lo cual se corrió la opción de “compresión óptima” y se generaron los casos: A) QSSRs/KSSRs B) QSSRs/KSNPs C) QSNPs/KSSRs y D) QSNPs/KSNPs.

4) MLM (K): Se utilizó un modelo mixto donde se incluyeron solamente las matrices KSSRs y KSNPs como un factor aleatorio. Para el caso de 11 ambientes se probaron A) KSSRs (compresión 58), B) KSNPs (compresión 41).

Para el caso de BLUE (modelo 4, caso 1= 4.1) se utilizó la opción “compresión óptima” utilizando las matrices A) KSSRs y B) KSNPs.

El programa Tassel generó dos salidas. En una tabla se registraron las estadísticas de los valores de significancia de las asociaciones marcador-fenotipo y en la otra los valores aditivos de tales asociaciones. Para la tabla de estadística se tomaron valores de umbral de p de 0,01 y 0,05 corregidos por el número de marcadores utilizados luego del filtrado por baja frecuencia (103 marcadores para el gen *ZmCCT* y 30.799 marcadores para el cromosoma 10).

Por otro lado, la tabla de efectos alélicos del programa TASSEL contiene las salidas de valores aditivos para un alelo de referencia, interpretándose como diferencia del efecto de un alelo homocigota con respecto al otro alelo en estado homocigota. Dicho intervalo representa así la suma térmica a floración o espigazón alargada o acortada por efecto de la presencia de uno de los alelos homocigotas con respecto al otro. En dichas tablas se analizaron la correspondencia de valores de significancia y aditividad positiva para los alelos de los dos marcadores asociados a un alargamiento del ciclo identificados previamente por (Hung *et al.,* 2012).

La bondad de ajuste de los modelos se estudió gráficamente mediante un Q-Q plot normal representando las tasas de error tipo I (falsos positivos) (valores de probabilidad acumulados vs. observados) bajo las diferentes pruebas. Así, los métodos que controlen mejor el error tipo I producirán una distribución uniforme de valores de probabilidad.

Las asociaciones marcador-fenotipo obtenidas bajo los diferentes métodos y modelos se compararon con los resultados obtenidos en Hung et al., (2012).

Las coordenadas físicas del gen *ZmCCT* y la identificación de los marcadores ligados a espigazón localizados en la región A6 (Hung *et al*., 2012) se identificaron mediante comparación de secuencias utilizando BLAST en el genoma de referencia de B73.

RESULTADOS

Estructura genética

El mejor agrupamiento k definido por STRUCTURE HARVESTER difirió al utilizar SSR (Tabla 1) o SNPs (Tabla 2). Así, los mejores k fueron k=2 y k=3 para SSRs y SNPs, respectivamente. En el caso de SSRs la estimación del mejor k (k=2) difirió a lo obtenido por Liu (que resultó ser k=3, Anexo 1) aun utilizando prácticamente el mismo set de datos. Siendo la asignación de individuos a las subpoblaciones más coincidente a Liu *et al*., (2003) cuando se utilizó SNPs. Tanto para SSRs como para SNPs la selección del mejor k fue coincidente con la mayor variación en el valor del delta k. Como se esperaba, se observó además una tendencia de reducción del valor absoluto de la probabilidad de los agrupamientos LnP(K) (valores más cercanos a cero) a medida que se aumentaba el valor de subpoblaciones k.

Tabla 1. Elección del mejor k (k=2) según STRUCTURE HARVESTER utilizando 94 SSRs.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	3	-86443.466667	6.940701	—	—	—
2	3	-80869.766667	8.958422	5573.700000	1853.000000	206.844453
3	3	-77149.066667	22.302541	3720.700000	1800.966667	80.751636
4	3	-75229.333333	296.470611	1919.733333	59.500000	0.200694
5	3	-73369.100000	105.288223	1860.233333	401.366667	3.812076
6	3	-71910.233333	77.821741	1458.866667	—	—

Tabla 2. Elección del mejor k (k=3) según STRUCTURE HARVESTER utilizando 553 SNPs.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	4	-130942.550000	0.519615	—	—	—
2	3	-124573.800000	5.211526	6368.750000	2437.483333	467.710125
3	3	-120642.533333	2.003331	3931.266667	2188.133333	1092.247768
4	3	-118899.400000	38.725573	1743.133333	244.366667	6.310214
5	3	-116911.900000	313.555019	1987.500000	399.433333	1.273886
6	3	-115323.833333	249.488924	1588.066667	—	—

Por un lado, Liu *et al*, (2003) encontraron con SSR una estructura genética que les permitió dividir a la población (panel de líneas) en: (NSS) líneas templadas *non-Stiff Stalk* (línea representativa MO17), (SS) líneas templadas *Stiff Stalk* (línea representativa, B73), (TS) líneas tropicales o semitropicales, (*Sweet corn*) maíz dulce, (*Popcorn*) maíz pisingallo. Sin embargo, para (Liu *et al*.,2003) un k=3 fue el mejor agrupamiento que clasificó a la población.

La comparación de los resultados de Liu *et al*., (2003) para el reanalisis de estructura genética se hicieron con k=3 y k=2 tanto para SSRs como para

SNPs para ampliar las posibilidades de comparación. Para tal fin, se seleccionaron las corridas (run) representativas que produjeron el valor de LnP(D) más cercano a cero (SSRs, k=2 run 5, k=3= run 8; SNPs, k=2 run 7, k=3 run 9). Posteriormente se seleccionaron líneas representativas de los maíces templados (SS y NSS) y semitropicales (TS) para comparar los agrupamientos con respecto a las mismas, incorporando de esta manera un criterio biológico en la selección del mejor k. Por último, la elección del k óptimo según el STRUCTURE HARVESTER fue de k=2 para SSRs y k=3 para SNPs. Estas matrices seleccionadas por STRUCTURE HARVESTER fueron las que finalmente se utilizaron en el análisis de asociación para la reducción del error tipo I.

En la Tabla 3 se puede observar que con SSRs hubo una mayor diferencia con el agrupamiento de las líneas *sweet corn* (maíz dulce) y *popcorn* (maíz reventador), las que fueron agrupadas en nuestro trabajo dentro el grupo NSS (*Non-Stiff Stalk*, cuya representante es MO17). Con el análisis de SNPs sucedió algo similar (Tabla 4) aunque se observó una mayor proporción de las líneas que difirieron con respecto a la clasificación de (Liu *et al* 2003).

Para analizar las salidas gráficas de k=2 y k=3 se tomaron como líneas de referencia a B73 para el grupo SS, MO17 y A619 (del subgrupo Oh43) dentro del grupo NSS, y CML10 para las semitropicales.

Tabla 3. Comparación de agrupamientos obtenidos en (Liu *et al*., 2003) y los obtenidos en el presente trabajo con SSRs (k=3). En celdas color verde se destacan las coincidencias de agrupamientos, en amarillo las líneas que en el trabajo se encuentran incluidas dentro de la subpoblación NSS, en naranja las líneas de la subpoblación TS y en gris las líneas que no están incluidas en el trabajo.

Grupo	Subgrupo	Líneas													
	Hy:T8:Wf9	CI21E	H49	Hy	Mo1W	Pa875	Pa880	T8	Va17	Va14	Va22	Va35	Va102	W64A	Wf9
NSS	M14:Oh43	A619	Gn2	H95	M14	Oh40B	Oh43	Oh43E	PA762	Va26	Va85				
	CO109:Mo1	A556	A682	CI.187-2	CO109	CO220	K187	Mo17	MS1334	ND246	W401				
	C103	B57	C103	C123	DE2	L317	L1546	NC258	NC262						
	Ga:SC	4226	F44	Ga209	GT112	SC357	SC213R	SC213							
	NSS-X	38-11	A239	A659	AR4	CM7	CM37	R168	Mo44	MS71	NC260	PA884P	R4	R177	W22
	K64W	33-16	CI31A	CI.64	CI.66	CI.7	E2558W	Ky21	K55	K64	M162W				
	NSS-mixed	A554	A654	B2	B52	B70	B77	B97	B103	CO106	CO125	F6			
		Fe2	H99	Mt42	N6	Os420	Pa91	R109B	SD44	T234	W153R				
	B14A	A214N	A632	A634	A635	A665	B14A	B64	B68	CM105	CM174	H91			
SS	B37	B37	B76	H84	NC250										
	N28	N28	N28Ht												
	B73	A679	A680	B73	B84	B104	B109	NC328	NC372	R229					
	SS-mixed	A641	De811	H100	N192	N196	NC294	NC368							
	TZI	A6	CML52	CML238	CML287	NC358	Q6199	Tzi8	Tzi9	Tzi10	Tzi18				
	Suwan	B96	CML69	CML228	CML349	KI3	KI9	KI11	KI14	KI44	KI2007				
TS	CML-late	CML5	CML9	CML61	CML103	CML220	CML254	CML258	CML261	CML264	CML314	Tx601			
	CML-early	CML14	CML247	CML311	CML321	CML322	CML331	CML332							
	NC	NC296	NC298	NC304	NC336	NC338	NC348	NC350	NC352	NC354					
	CML-P	CML10	CML11	CML45	CML277	CML281	CML333	CML341							
	TS-mixed	CML38	CML108	NC300	NC356										
Sweet corn		Ia2132	Il14H	Il101t	Il677a	P39									
Popcorn		HP301	I29	IDS28	IDS69	SA24	Sg18	Sg1533							
Mixed		A188	A272	A441-5	A656	B79	B94	B105	B164	C49A	CML77	CML91	CML92	CML218	
		CML323	CML328	CMV3	CO159	CO255	D940Y	DE3	EP1	F2	F2834T	F7	Hi27	I137TN	
		I205	IDT	Ki43	Ky226	Ky228	L578	Le23	Le773	M37W	Mo18W	Mo24W	Mp339	MS153	
		N7A	NC264	NC320	NC360	NC362	NC364	NC366	NC370	Oh7B	Oh603	SC55	SD40	SD46	
		T232	TEA	Tx303	Tzi11	Tzi16	Tzi25	U267Y	Va99	W117	W117Ht	W182B			

Tabla 4. Comparación de agrupamientos obtenidos en (Liu *et al*., 2003) y los obtenidos en el presente trabajo con SNPs (k=3). En celdas color verde se destacan las coincidencias de agrupamientos, en amarillo Las líneas que en el trabajo se encuentran incluidas dentro de la subpoblación NSS, en naranja las líneas de la subpoblación TS, en celeste las líneas de la subpoblación SS y en gris las líneas que no están incluidas en el trabajo

Subpoblación	Subgrupo	Líneas													
NSS	Hy:T8:Wf9	CI21E	H49	Hy	Mo1W	Pa875	Pa880	T8	Va17	Va14	Va22	Va35	Va102	W64A	Wf9
	M14:Oh43	A619	Gn2	H95	M14	Oh40B	Oh43	Oh43E	PA762	Va26	Va85				
	CO109:Mo1	A556	A682	CI.187-2	CO109	CO220	K187	Mo17	MS1334	ND246	W401				
	C103	B57	C103	C123	DE2	L317	L1546	NC258	NC262						
	Ga:SC	4226	F44	Ga209	GT112	SC357	SC213R	SC21							
	NSS-X	38-11	A239	A659	AR4	CM7	CM37	R168	Mo44	MS71	NC260	PA884P	R4	R177	W22
	K64W	33-16	CI31A	CI.64	CI.66	CI.7	E2558W	KY21	K55	K64	M162W				
	NSS-mixed	A554	A654	B2	B52	B70	B77	B97	B103	CO106	CO125	F6			
		Fe2	H99	Mt42	N6	Os420	Pa91	R109B	SD44	T234	W153R				
SS	B14A	A214N	A632	A634	A635	A665	B14A	B64	B68	CM105	CM174	H91			
	B37	B37	B76	H84	NC250										
	N28	N28	N28Ht												
	B73	A679	A680	B73	B84	B104	B109	NC328	NC372	R229					
	SS-mixed	A641	De811	H100	N192	N196	NC294	NC368							
TS	TZI	A6	CML52	CML238	CML287	NC358	Q6199	Tzi8	Tzi9	Tzi10	Tzi18				
	Suwan	B96	CML69	CML228	CML349	KI3	KI9	KI11	KI14	KI44	KI2007				
	CML-late	CML5	CML9	CML61	CML103	CML220	CML254	CML258	CML261	CML264	CML314	Tx601			
	CML-early	CML14	CML247	CML311	CML321	CML322	CML331	CML332							
	NC	NC296	NC298	NC304	NC336	NC338	NC348	NC350	NC352	NC354					
	CML-P	CML10	CML11	CML45	CML277	CML281	CML333	CML341							
	TS-mixed	CML38	CML108	NC300	NC356										
Sweet corn		Ia2132	Il14H	Il101t	Il677a	P39									
Popcorn		HP301	I29	IDS28	IDS69	SA24	Sg18	Sg1533							
Mixed		A188	A272	A441-5	A656	B79	B94	B105	B164	C49A	CML77	CML91	CML92	CML218	
		CML323	CML328	CMV3	CO159	CO255	D940Y	DE3	EP1	F2	F2834T	F7	Hi27	I137TN	
		I205	IDT	Ki43	Ky226	Ky228	L578	Le23	Le773	M37W	Mo18W	Mo24W	Mp339	MS153	
		N7A	NC264	NC320	NC360	NC362	NC364	NC366	NC370	Oh7B	Oh603	SC55	SD40	SD46	
		T232	TEA	Tx303	Tzi11	Tzi16	Tzi25	U267Y	Va99	W117	W117Ht	W182B			

Para el caso de SSRs, cuando se definió un k=2 (Fig. 1) se observó que las líneas tropicales se encuentran agrupadas y se mantienen discretas mientras se incrementa el valor de k. Prácticamente al pasar de k=2 a k=3 solo varía las líneas pintadas de rojo, dividiéndose en SS y NSS, siendo estas en su mayor número. Para un k=3 (Fig. 2) se definieron tres subpoblaciones discretas. Una de ellas, con líneas semitropicales (TS, código 52 a 87) permanecieron separadas del resto a lo largo de las simulaciones k=1 a 6. Esta subpoblación agrupó a: CML10, CML103, CML108, CML11, CML14, CML218, CML220, CML228, CML238, CML247, CML254, CML258, CML261, CML264, CML277, CML281, CML311, CML314, CML321, CML322, CML328, CML331, CML332, CML333, CML341, CML38, CML45, CML5, CML52, CML61, CML69, CML77, CML92, NC354, NC356, NC358, NC360, NC362, NC364, NC366, y NC368. Dentro de las templadas, hubo mayor número de líneas pertenecientes al grupo NSS que a SS. Dentro de la categoría mezclas (*mixed*) hubo una proporción mayor de la mezcla SS/NSS y NSS/semitropical que SS/semitropical.

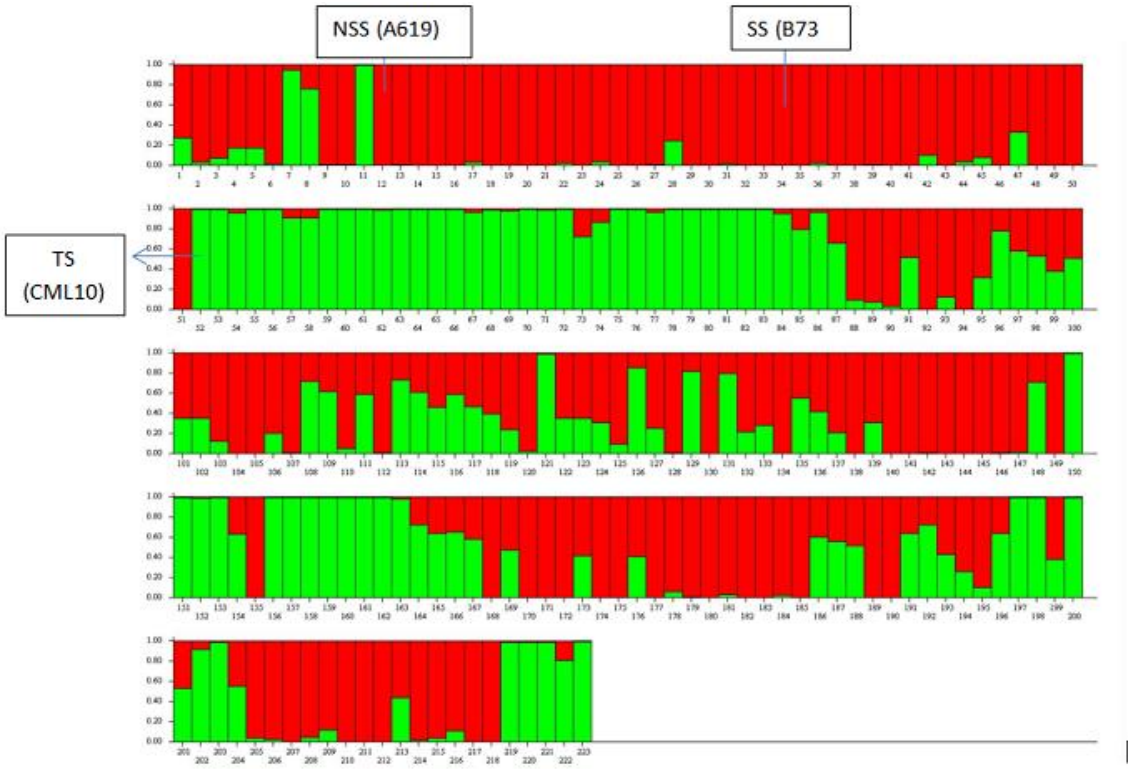


Figura 1. Estructura genética del panel de líneas de maíz estimada con SSRs (k=2 Run 5).

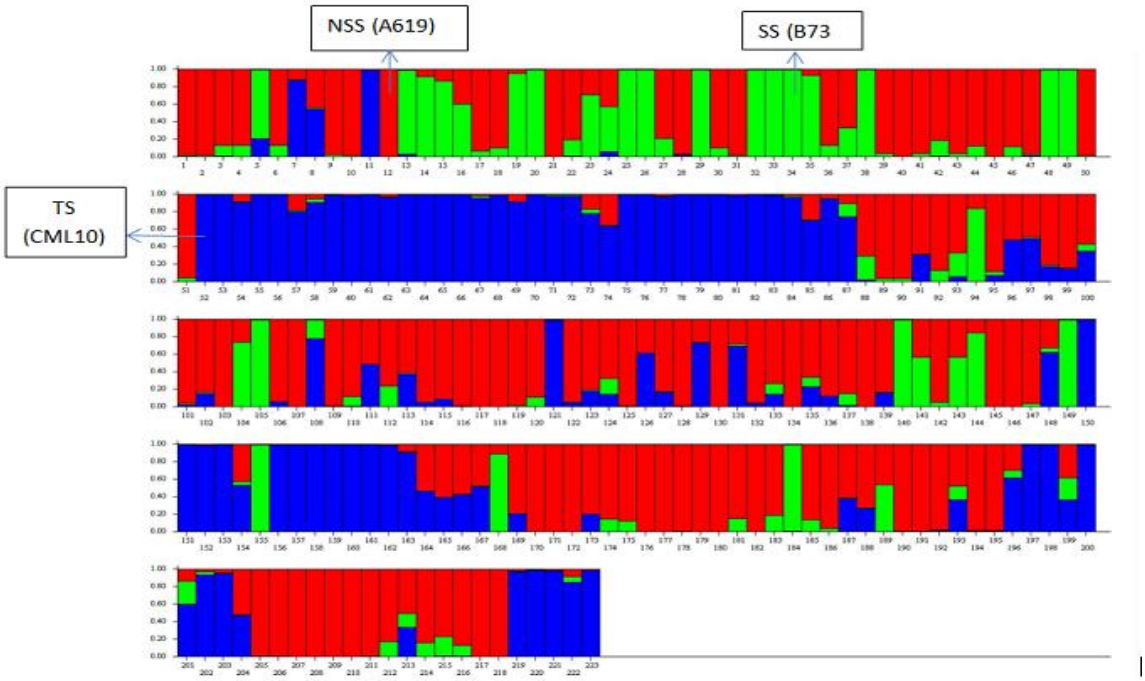


Figura 2. Estructura genética del panel de líneas de maíz estimada con SSRs (k=3 Run 8).

Para el caso de SNPs, cuando se estableció un k=2 (Fig. 3) se observó en particular una subpoblación discreta perteneciente a las semitropicales (líneas con pertenencias Q ≥0,8), que permanecieron agrupadas hasta k=6 (Anexo 3), diferenciándose notoriamente de las templadas. Esta subpoblación prácticamente discreta agrupó a semitropicales y fue coincidente a lo obtenido para SSRs. Para un k=3 (Fig 4) se definieron tres subpoblaciones discretas con líneas pertenecientes a SS, NSS y semitropicales. Al igual que con SSR, dentro de las templadas hubo mayores cantidades de líneas pertenecientes al grupo NSS que a SS. Dentro de la categoría mezclas (*mixed*) hubo una mayor proporción de mezcla entre SS/NSS y NSS/semitropical que SS/semitropical.

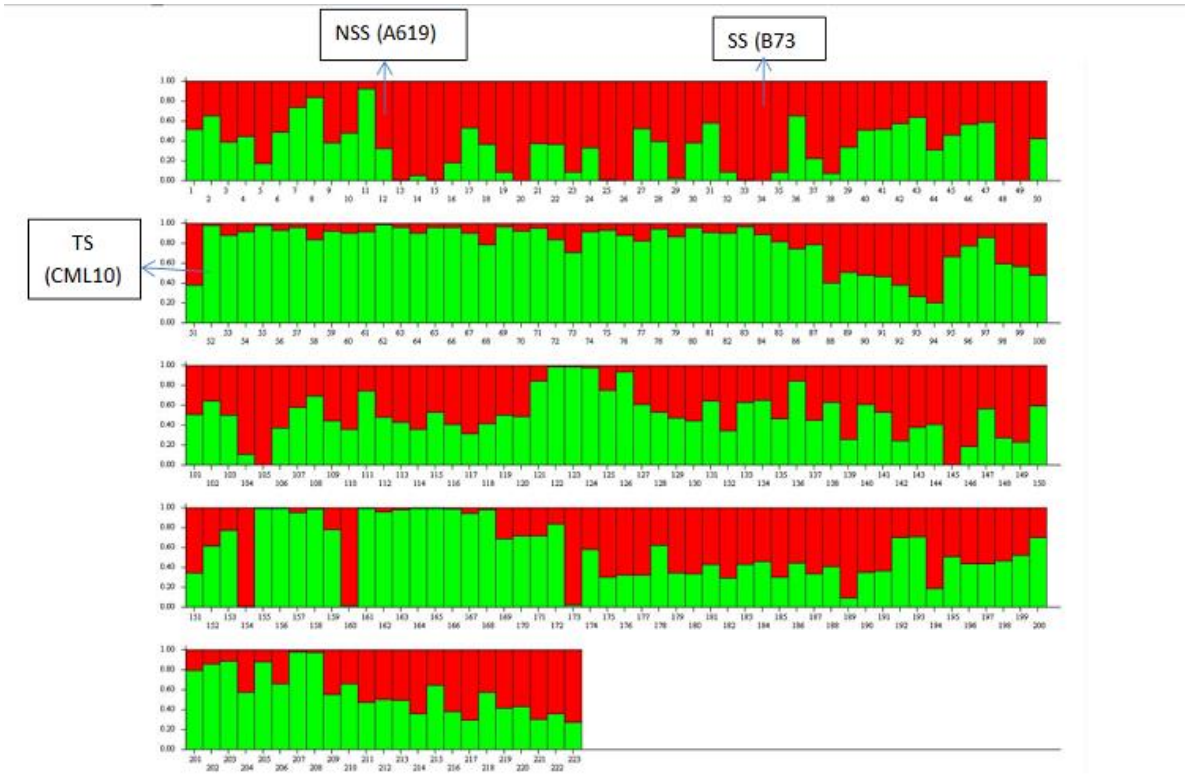


Figura 3. Estructura genética del panel de líneas de maíz estimada con SNPs (k=2 *Run* 7).

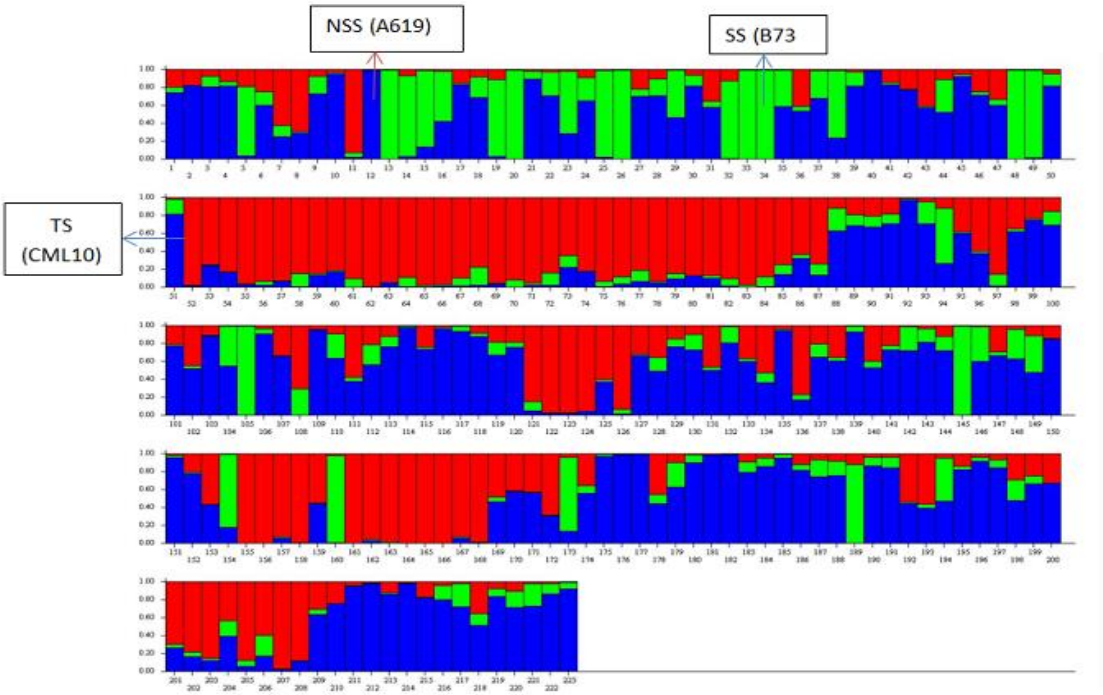


Figura 4. Estructura genética del panel de líneas de maíz estimada con SNPs (k=3 *Run* 9).

Parentesco

Debido al tamaño de las tablas de matrices de parentesco (KSRRs y KSNPs) se realizaron histogramas con el fin de observar la magnitud de las relaciones de parentesco obtenidas con cada set de datos. Se puede observar, tanto en la Fig. 5 (SSRs) como en Fig 6 (SNPs), los valores en su mayoría fueron cero o muy próximos a cero. La distribución de coeficientes fue distinta entre marcadores, para SSRs se lograron conteos más distribuidos en el rango 0,01-0,04 (Fig. 5), mientras que para SNPs la mayor participación fue de ceros y con una tendencia a una distribución más uniforme a medida que aumentaban los valores de los coeficientes hacia valores de 0,20 (Fig. 6).

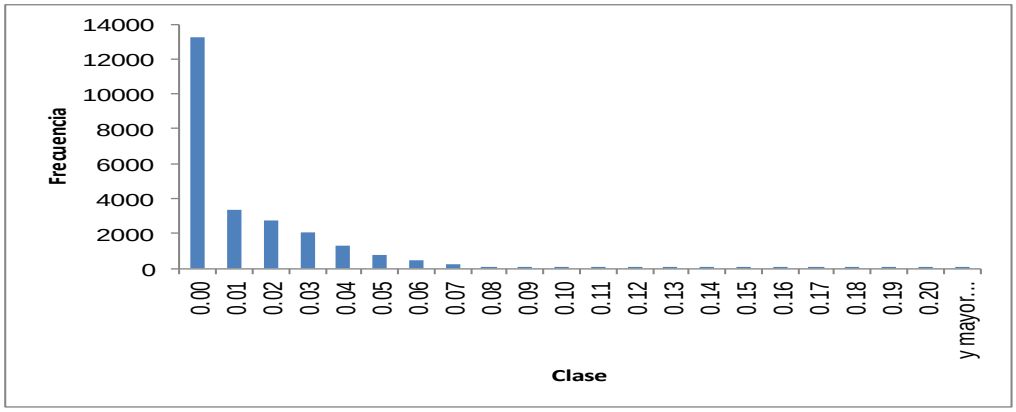


Figura 5. Histograma de los coeficientes de parentesco obtenidos SSRs.

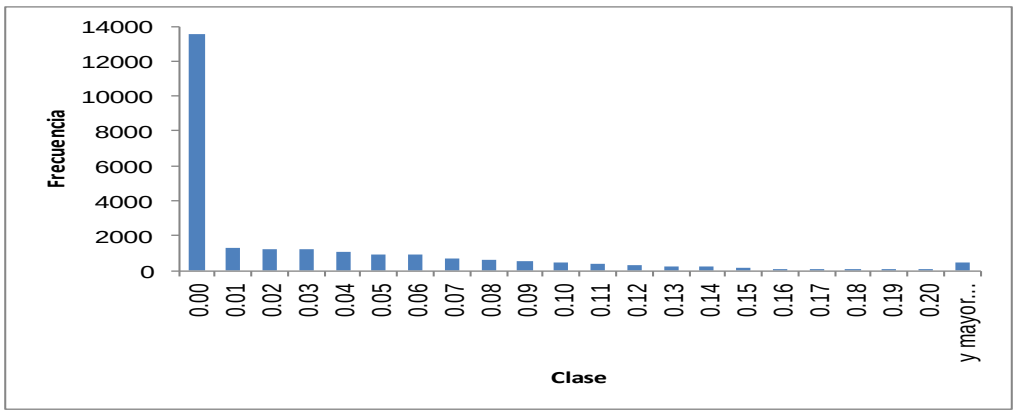
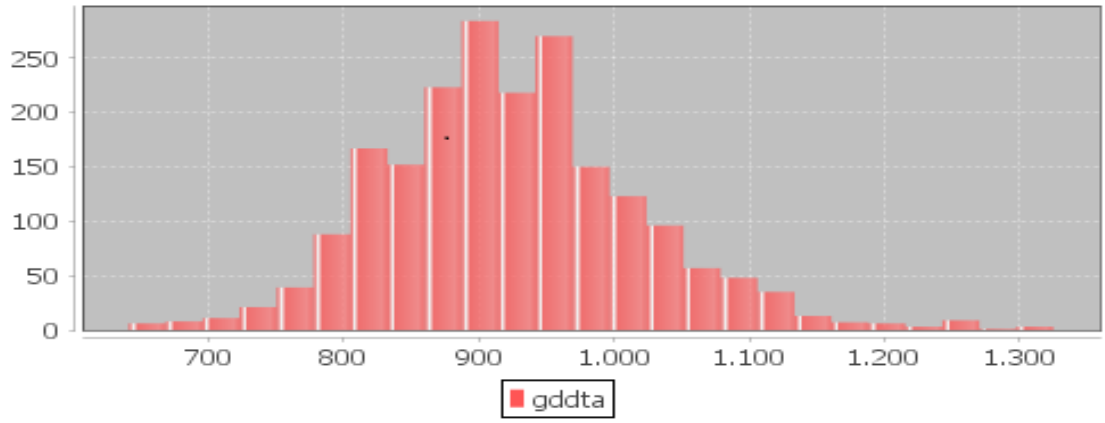


Figura 6. Histograma de los coeficientes de parentesco obtenidos con SNPs.

Mapeo de QTLs

Los datos fenotípicos del conjunto de 11 ambientes y de las medias ajustadas de los 11 ambientes (BLUE) se analizaron mediante histograma. En el caso de 11 ambientes (Fig. 7a) se logró una distribución que ajusta mejor a una distribución normal comparado con BLUE (Fig. 7b).

En este último caso, si bien se observó una tendencia a la bimodalidad de los datos fenotípicos se consideró que la distribución de datos fue suficientemente buena para utilizar los datos sin transformación. Aclarando que el procedimiento estadísticamente correcto es la comparación de distribución de datos de erros del modelo y no de los datos fenotípicos observados.



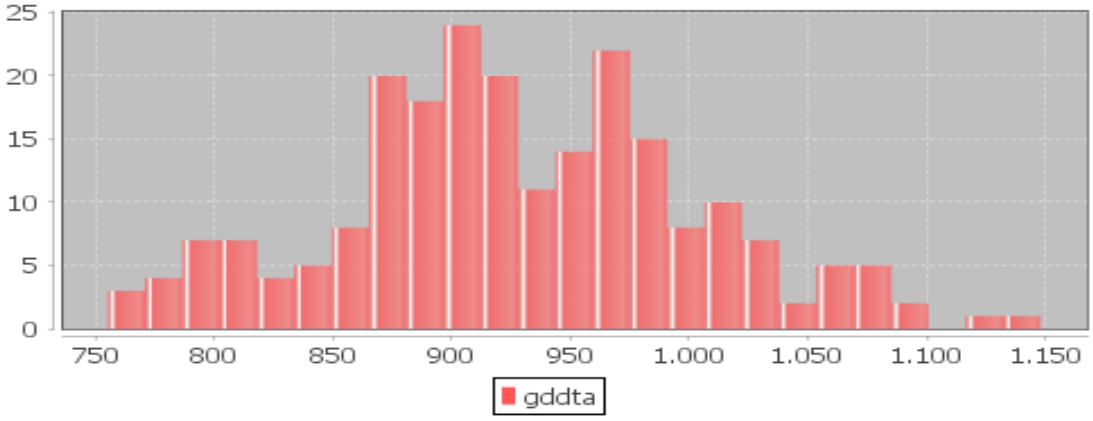


Figura 7. Histograma de valores fenotípicos (suma térmica a antesis) del conjunto de 11 ambientes (a) y de las medias ajustadas (BLUE) de dichos ambientes (b).

El gen *ZmCCT* fue identificado como GRMZM2G381691 codificando en la cadena negativa en coordenadas AGPv4:10:94.434.089:94.435.401:-1, aproximadamente 94,2 Mbp en versión 2, ZmB73_RefGen_v2.fa (Fig. 9). Los dos marcadores SNPs completamente ligados de la región A6 (Fig. 8), -1412 y -1520, se localizaron río arriba a aproximadamente 1,479 pb de GRMZM2G38169 en correspondencia a lo informado por (Hung *et al*., 2012) (Fig. 10).

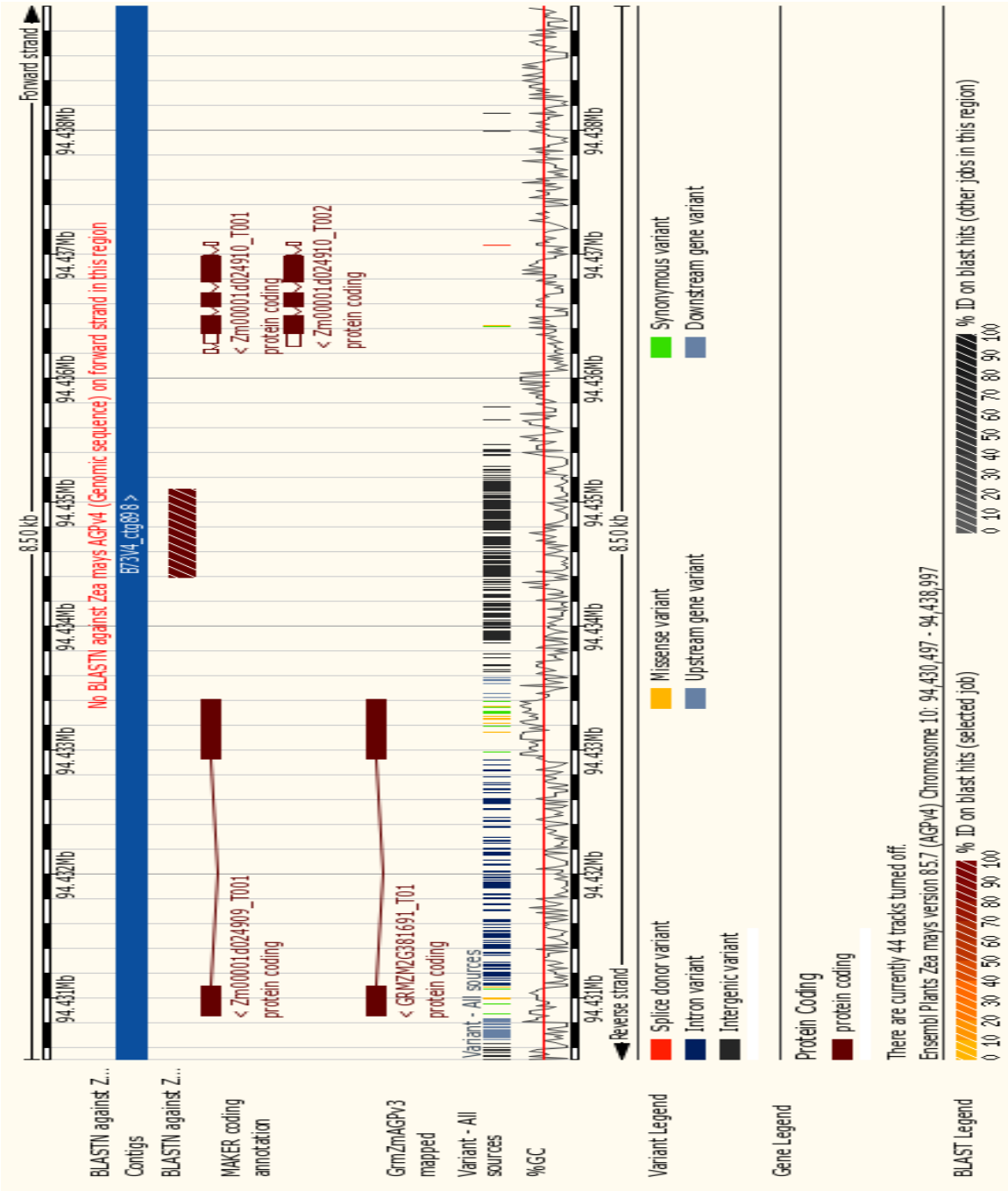


Figura 9. Localización del fragmento A6 (barra diagonal) conteniendo los marcadores SNPs ligados a floración ubicado río arriba de GRMZM2G381691.

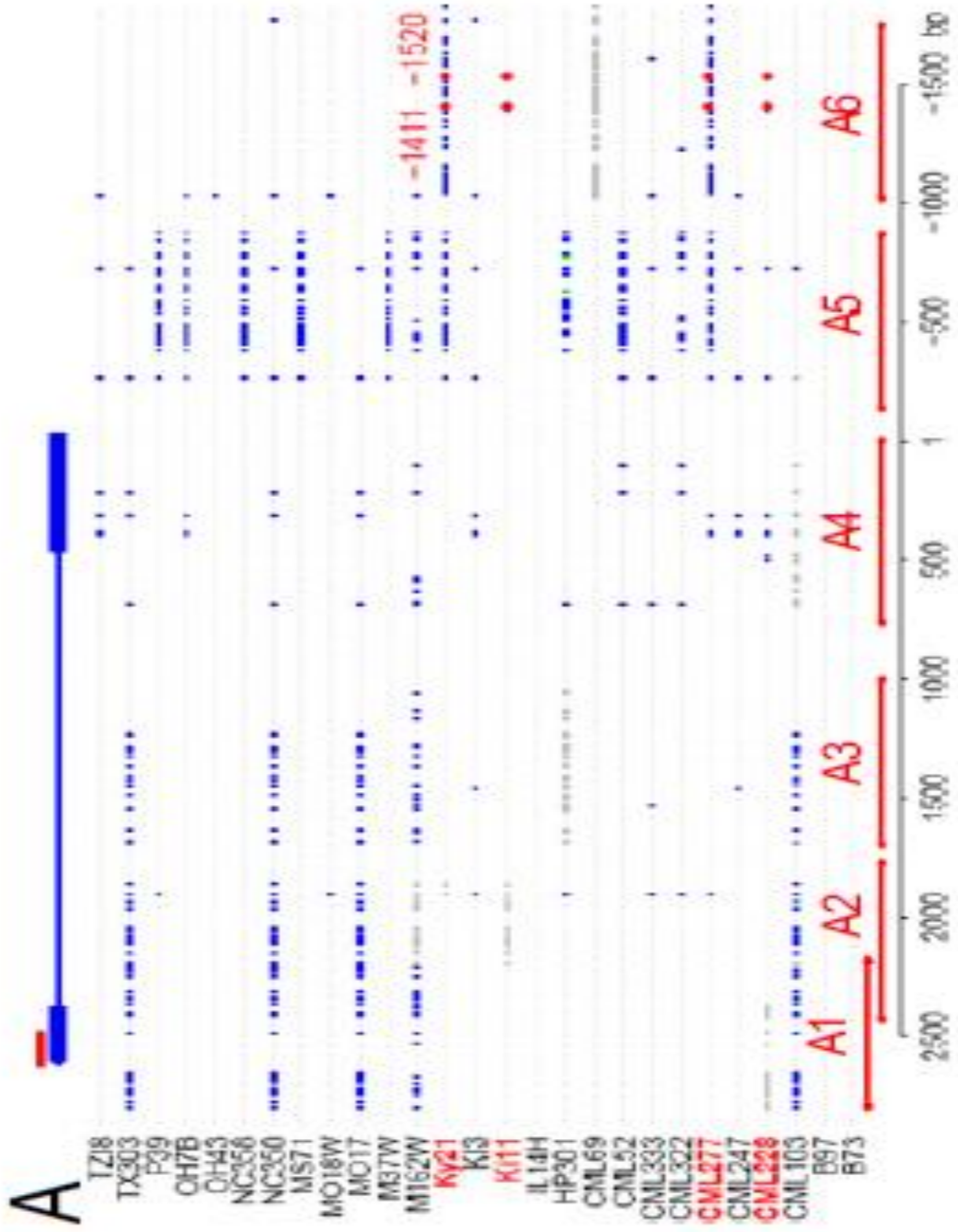


Figura 10. Localización del gen ZmCCT (barra azul). A1 a A6 se indican los fragmentos secuenciados para descubrir marcadores. En el fragmento A6 (extragénico) se indican los marcadores SNPs ligados a días a floración identificados por (Hung et al., 2012).

Para aplicar el método del gen candidato *ZmCCT* se realizaron las combinaciones de análisis con matrices de estructura, parentesco, datos fenotípicos y los marcadores SNPs e INDELs de la región A6 y cuyos

resultados se resumen en Tabla 5. El análisis GLM sin corrección (caso 1) logró resultados significativos al 1 y al 5% mientras que la corrección por estructura no logró resultados significativos. En el caso de MLM, los resultados dependieron del tipo de datos fenotípicos utilizados y del ajuste por estructura como en el caso GLM. Para datos fenotípicos de experimentos individuales y ajuste por Q y K, todas las combinaciones de análisis dieron resultados significativos al 1 y al 5%. Mientras que el resultado de ajuste solo por K dependió del tipo de matriz K. En general, el uso de datos fenotípicos medios ajustados (BLUE) en lugar del conjunto de experimentos individuales no logró resultados significativos. En todos los casos los valores aditivos fueron positivos lográndose mayores valores absolutos para GLM sin corrección (caso 1) y caso 4 (A).

Tabla 5. Valores de significancia y aditividad para diferentes combinaciones de análisis empleando el método del gen candidato *ZmCCT*. En color amarillo y rojo se representan aquellos valores de p que resultaron significativos al 5% (0,05/103=4,85^{-4}) y al 1% (0,01/103=9,70^{-5}), respectivamente.

		modelos	Valor de p de los marcadores		Valor Aditivo (2*a)	
			226	368	226	368
Caso 1		GLM	5,068 x 10^{-8}	5,068 x 10^{-8}	90,6	90,6
Caso 2	A	GLM-QSSR	2,40 x 10^{-2}	2,40 x 10^{-2}	30,2	30,2
	B	GLM-QSNP	5,173 x 10^{-2}	5,173 x 10^{-2}	26,9	26,9
Caso 3	A	MLM- QSSR/KSSR-c58	4,422 x 10^{-4}	4,422 x 10^{-4}	30,5	30,5
ZmCCT	B	MLM- QSSR/KSNP-c41	4,004 x 10^{-4}	4,004 x 10^{-4}	29,6	29,6
11Amb.	C	MLM- QSNP/KSSR-c58	3,684 x 10^{-5}	3,684 x 10^{-5}	34,7	34,7
	D	MLM- QSNP/KSNP-c41	3,684 x 10^{-5}	3,684 x 10^{-5}	34,7	34,7
Caso 3.1	A	MLM- QSSR/KSSR	6,375 x 10^{-1}	6,375 x 10^{-1}	7,43	7,43
ZmCCT	B	MLM- QSSR/KSNP	8,814 x 10^{-2}	8,814 x 10^{-2}	26,6	26,6
Blue	C	MLM- QSNP/KSSR	3,086 x 10^{-1}	3,086 x 10^{-1}	16,12	16,12
	D	MLM- QSNP/KSNP	6,32 x 10^{-2}	6,32 x 10^{-2}	27,7	27,7
Caso 4	A	MLM-KSSR-c58	9,152 x 10^{-16}	9,152 x 10^{-16}	75,4	75,4
ZmCCT	B	MLM-KSNP	5,75 x 10^{-3}	5,75 x 10^{-3}	51,7	51,7
11Amb.	B1	MLM-KSNP-c41	5,530 x 10^{-6}	5,530 x 10^{-6}	39,8	39,8
Caso 4.1	A	MLM-KSSR	2,845 x 10^{-2}	2,845 x 10^{-2}	42,6	42,6

ZmCCT Blue B	MLM-KSNP	1,195 x 10^{-2}	1,195 x 10^{-2}	40,25	40,25

Para aplicar el método de marcadores de distribución genómica (GWAS) se realizaron las combinaciones de análisis con matrices de estructura, parentesco, datos fenotípicos y de marcadores del cromosoma 10 a fin de obtener marcadores SNPs asociados próximos al gen *ZmCCT*. Los resultados obtenidos para los diez primeros marcadores asociados ordenados con menor valor de probabilidad se resumen en Tabla 6. En general se observa que no se lograron marcadores físicamente cercanos a los previamente asociados con el método del gen candidato, esto son: PZE1094244760 (94.420.967 bp en V2) y PZE1095719172 (95.925.762 pb en V2) y que hubo múltiples marcadores significativamente asociados con valores de probabilidad significativos.

La comparación de resultados entre modelos GML y MLM permite establecer la eficacia del MLM para reducir falsas asociaciones causadas por relaciones de parentesco y estructura genética. En este caso de GWAS, dentro de modelos GLM, el número de marcadores significativos se redujo al utilizar una corrección por las relaciones de estructura estimada con SNP (caso 2B). En el caso MLM, se redujeron los marcadores significativos al utilizar datos fenotípicos medios ajustados (BLUE) (casos 3.1 y 4.1). La reproducibilidad de resultados entre modelos se puede analizar por ejemplo con el marcador SNP (A/G) posicionado en la coordenada 104.891.334 bp (Tabla 6, letra en color rojo). Se observa que este SNP fue seleccionado en siete modelos probados, teniendo el alelo A en estado homocigota un valor aditivo positivo, que varió de 30 a 57,3 grados días). En este caso, por ejemplo, se observa además que la introducción de una base inespecífica (que no fue filtrada) en la tabla de datos de secuencia ocasionó que el programa TASSEL muestre el valor aditivo de ambos alelos, siendo que en general este programa utiliza por convención uno de los alelos de referencia asignándole un valor de cero.

Tabla 6. Valores de significancia y aditividad para diferentes combinaciones de análisis empleando el método de marcadores de cobertura genómica. En color amarillo y rojo se representan aquellos valores de p que resultaron significativos al 5% (0,05/30.799=1,62^{-6}) y al 1% (0,01/30.799=3,25^{-7}), respectivamente. En letra negrita se indican aquellos marcadores que fueron significativos al menos en dos modelos. En el caso GLM-QSNP se agregaron en letra gris, marcadores y valores p no significativos ordenados para hacer comparativa la lista de marcadores y sus resultados.

	Modelos	Posición	Valor p	Valor aditivo (2*a)
Caso 1	GLM	113.167.845	2.12E^{-21}	-58,5
Caso 2 A	GLM-QSNP	104.891.334	2.91E^{-8}	A 57,3/ G -29
	GLM-QSSR	122.266.884	1.04E^{-8}	-70,9
Caso 3 A 11Amb.	MLM- QSSR/KSSR-c58	110.297.615	5.17E^{-30}	103
		110.298.371	4.39E^{-29}	-111
		104.891.334	2.87E^{-28}	A 52 / G -39
		110.201.869	7.54E^{-26}	-110
		122.266.884	4.22E^{-25}	-69
		103.635.697	2.34E^{-24}	-83
		113.709.297	4.55E^{-24}	C 10/ T -39
		117.565.801	5.22E^{-24}	55
		125.970.874	6.76E^{-24}	-57
		125.969.218	1.81E^{-23}	64
Caso 3 B 11Amb.	MLM- QSSR/KSNP-c41	27.697.179	7.29E^{-26}	-90
		121.749.588	7.61E^{-26}	58
		121.749.594	7.61E^{-26}	58
		121.750.451	5.54E^{-23}	-73
		110.297.615	1.65E^{-22}	92
		121.746.541	2.93E^{-22}	62
		104.891.334	8.3E^{-22}	A 40 / G -40
		121.750.640	1.73E^{-22}	58
		121.750.641	1.73E^{-21}	-58
		125.970.874	4.73E^{-21}	-53

Caso 3 C 11Amb.	MLM- QSNP/KSSR-c58	104.891.334	$2.33E^{-26}$	A 62 / G -23
		110.298.371	$6.62E^{-24}$	-98
		110.297.615	$1.31E^{-23}$	89
		65.967.953	$2.08E^{-23}$	A 134 / G 171
		89.986.504	$4.92E^{-22}$	-70
		103.635.697	$8.26E^{-22}$	-77
		125.970.874	$6.07E^{-20}$	-51
		110.201.869	$6.49E^{-20}$	-94
		65.622.760	$7.04E^{-20}$	-40
		27.697.179	$7.32E^{-20}$	-75
Caso 3 D 11Amb.	MLM- QSNP/KSNP-c41	27.697.179	$5.91E^{-23}$	-83
		89.986.504	$2.49E^{-22}$	-70
		104.891.334	$2.53E^{-21}$	A 49 / G -29
		65.967.953	$1.15E^{-20}$	A 97 / G 135
		121.749.588	$1.21E^{-20}$	51
		121.749.594	$1.21E^{-20}$	51
		137.354.078	$4.17E^{-20}$	73
		137.354.081	$4.17E^{-20}$	C 73 / G 107
		137.354.080	$4.17E^{-20}$	G 107 / T 73
		144.679.458	$2.43E^{-19}$	A 45 / G 73
Caso 3.1 A Blue	MLM- QSSR/KSSR	113.167.845	$1.32E^{-08}$	-82
		147.928.921	$1.23E^{-06}$	82
		5.749.783	$1.47E^{-06}$	87
		138.662.733	$1.49E^{-06}$	-49
Caso 3.1B	MLM- QSSR/KSNP	No da significativo ningún marcador		
Caso 3.1C	MLM- QSNP/KSSR	No da significativo ningún marcador		
Caso 3.1 D	MLM- QSNP/KSNP	No da significativo ningún marcador		
Caso 4 A 11 Amb.	MLM-KSSR-c58	113.167.845	$2.63E^{-44}$	-97
		133.881.061	$1.11E^{-40}$	-80
		147.726.754	$2.14E^{-38}$	-77
		125.969.218	$2.54E^{-38}$	88
		125.969.248	$2.54E^{-38}$	88
		125.969.238	$2.54E^{-38}$	-88
		133.881.100	$2.46E^{-37}$	-79

		113.709.297	$2.93E^{-37}$	C 37 / T -25
		5.380.987	$2.97E^{-37}$	-87
		5.380.988	$2.97E^{-37}$	-87
Caso 4 B 11 Amb.	MLM-KSNP	27.697.179	$5.07E^{-28}$	-95
		121.749.588	$1.44E^{-23}$	58
		121.749.594	$1.44E^{-23}$	58
		113.519.493	$8.28E^{-23}$	50
		113.709.297	$9.71E^{-22}$	C 25 / T -22
		143.935.843	$2.95E^{-21}$	-73
		2.991.744	$8.59E^{-21}$	-62
		21.948.309	$3.93E^{-20}$	48
		121.746.541	$9.00E^{-20}$	61
		104.891.334	$9.37E^{-20}$	A 30 / G -50
Caso 4.1A Blue	MLM-KSSR	113.167.845	$1.32E^{-08}$	-82
		147.928.921	$1.23E^{-06}$	82
		5.749.783	$1.47E^{-06}$	87
		138.662.733	$1.49E^{-06}$	-49
Caso 4.1B Blue	MLM-KSNP	125.969.218	$1.47E^{-06}$	61
		125.969.248	$1.47E^{-06}$	61
		125.969.238	$1.47E^{-06}$	-61

DISCUSIÓN

El mapeo de asociación tiene múltiples ventajas con respecto al mapeo tradicional pero ambos métodos son complementarios para validar resultados (Zhao *et al*., 2011, Korte y Farlow, 2013). Entre las ventajas, el mapeo de asociación permite aumentar la escala para investigar la variabilidad fenotípica en un elevado número de genotipos que en la población en conjunto expresan simultáneamente diferentes formas alélicas de un gen o QTL responsables de un carácter. Otra gran ventaja, es la posibilidad de utilizar como población de mapeo una colección de materiales mayormente homocigotas, como ser líneas experimentales, comerciales, variedades de programas de mejoramiento antiguo y moderno (panel de asociación), con el ahorro de tiempo sustancial comparado al tiempo que implica la preparación de una población de mapeo tradicional (Flint-Garcia *et al*., 2005). Por otro lado, una gran desventaja del mapeo de asociación es la ocurrencia del error Tipo I en las asociaciones marcador-fenotipo (falsos positivos) debida a relaciones de parentesco y estructura genética dentro del panel de asociación (Yu *et al*., 2006). Del mismo modo, la propiedad intrínseca del panel de asociación (composición, relaciones de parentesco y estructura genética, número, tipo, distribución y grado de desequilibrio de los marcadores moleculares) condicionan la precisión de las asociaciones marcador-fenotipo, y así la aplicación tanto del método GWAS o del gen candidato (Rafalski, 2002). Con el objetivo de mejorar la precisión estadística del mapeo de asociación, una nueva población denominada población de asociación anidada se encuentra actualmente muy difundida en especies vegetales, entre ellas maíz (Yu *et al*., 2008). En dicha población, un parental recurrente (línea B73, en maíz) es cruzado en forma recurrente con un grupo de líneas que maximizan entre ellas la diversidad genotípica y fenotípica para desarrollar así una población recombinante endocriada (Yu *et al.,* 2008). Por último, cabe aclarar que en general las estimaciones de relaciones de parentesco y de estructura se realizan con marcadores de cobertura genómica, distanciados físicamente a fin de minimizar el desequilibrio de ligamiento y cumplir con los supuestos del modelado bayesiano (Pritchard *et al*., 2000). En

cambio, el GWAS requiere mayor densidad de marcadores que puede ir de los miles a millones según tamaño del genoma y capacidad de genotipado.

En este trabajo, se utilizó un panel de asociación de maíz ampliamente caracterizado a nivel genotípico y fenotípico en Estados Unidos, compuesto por genotipos templados de dos grupos heteróricos (SS y NSS), tropicales y semitropicales y en menor medida materiales dulces y para pochoclo. Dicho panel ha servido de base a estudios de modelado estadístico (Liu *et al.,* 2003; Flint-Garcia et al., 2005; Yu et al., 2006) e identificación, junto a la población de mapeo anidada, de genes y QTLs de interés agronómico como ser arquitectura de planta (Tian *et al*., 2011), calidad de grano (Cook *et al*., 2012) y respuesta al fotoperiodo (Hung et al., 2012). Para tal fin, se seleccionó este último como referencia para modelar el procedimiento estadístico que tuvo como resultado la identificación del gen *ZmCCT* como responsable de la respuesta fotoperiódica en maíz, utilizando la población anidada para método GWAS como procedimiento de aproximación a genes candidatos y el panel de asociación para el método del gen candidato con el fin de reducir la región y el número de genes candidatos identificados previamente en el cromosoma 10. A diferencia de dicho trabajo, en el presente trabajo ambos métodos fueron aplicados en forma simultánea utilizando exclusivamente el panel de asociación, el mismo *set* de SNP de *ZmCCT* (secuencias de la región A6, río arriba del gen) para el método gen candidato y un *set* de marcadores SNP diferentes (Maize282_imputed_AllZea_GBS_Build_July_2012) para el GWAS, y diferentes combinaciones de modelos mixtos ajustados por parentesco y estructura, calculados además con diferentes sets de datos descriptos en (Hung *et al.,* 2012). Cabe aclarar que la obtención del *set* de datos moleculares utilizados en (Hung *et al.,* 2012) y anteriores, no resultó completa. Solamente se lograron caracterizar con iguales niveles de datos comunes de SSR, SNP (para el cálculo de parentesco y estructura) y datos fenotípicos a 223 líneas de las 282 descriptas originalmente. Una razón de dicha dificultad, fue la falta de provisión de parte de los autores de la información digital sumando al avance de la generación en Estados Unidos de *set* de datos genotípicos de SNP por diferentes proyectos de descubrimiento de marcadores

por genotipado por secuenciación masiva (Hapmap 1, 2 y 3) y la renovación de la apariencia y contenido de la web Panzea que fue eliminando paulatinamente datos de SSR y SNP antiguos.

Las estimaciones de parentesco y estructura variaron ligeramente según el tipo de marcadores utilizados (553 SNP vs 94 SSR en nuestro caso). Pero ambos *sets* de marcadores lograron aproximar resultados a los obtenidos por (Liu et al., 2003) utilizando el mismo *set* de 94 SSR. Similares resultados se observaron en la separación de grupos dentro de una colección de girasol (Fillipi *et al*., 2015). La explicación de las ligeras diferencias de agrupamientos entre SNP y SSR podría atribuirse a la naturaleza bialélica de los SNP y multialélica de los SSR, que confiere a estos últimos mayor riqueza alélica y capacidad discriminatoria con menor cantidad de marcadores (Guichoux *et al*., 2011). Con respecto a las estimaciones de parentesco y estructura, se conoce que ambas pueden ser útiles para ajustar el análisis del modelo de asociación porque pueden revelar diferentes niveles de grados de covarianza genética entre los genotipos de un panel de asociación (Yu *et al*., 2006). Así la distribución de coeficientes de parentesco (*kinship*) estuvo muy concentrada hacia valores de cero indicando una muy baja relación de parentesco entre genotipos (valor cero) o bien, relaciones de parentesco muy bajas, pero dentro de un rango muy estrecho, como lo obtenido previamente en otros paneles de asociación de maíz (Olmos *et al*., 2016, Romay *et al*., 2013). En cambio, utilizando el mismo set de datos moleculares los resultados de Structure en general separan a los genotipos dentro de clusters discretos (Q > 0,8) lo que permite clasificar a los genotipos dentro de subpoblaciones de una forma más simple y clara (Pritchard *et al*., 2000). En otro sentido, la conformación de un panel de asociación busca maximizar la varianza genotípica para identificar múltiples formas alélicas de un gen de interés (Oraguzie y Wilcox, 2007). Sin embargo, el parentesco y la estructuración de la población que determina los grupos conlleva a aumentar la probabilidad de ocurrencia de falsos positivos (Korte y Farlow, 2013). Esto es debido a que la conformación de un panel con genotipos altamente diversos, en baja frecuencia, genera a su vez un *set* de alelos de baja frecuencia que al ser filtrados (eliminados, como se recomienda

en los protocolos de mapeo de asociación) trae aparejado la perdida de la asociación marcador-fenotipo por la eliminación del marcador. Tal es la situación sucedió por ejemplo con un alelo que confiere alto contenido de ácido oleico en maíz el cual, al ser de origen ancestral, se encuentra en muy baja frecuencia en las colecciones de líneas modernas de maíz y donde el mapeo de asociación no resultó exitoso para su detección (Beló *et al*., 2008). En el caso del gen *ZmCCT*, sucede una situación similar. Las variantes alélicas asociadas, se encuentran presentes mayormente en genotipos tropicales que en estado puro no son utilizados en clima templado para la producción de granos por justamente su ciclo largo ciclo que demora la fase reproductiva. De esta forma, si bien la mayoría de los genotipos cultivados en clima templado tienen una cierta respuesta al fotoperiodo, es decir los días largos de fin de primavera a comienzo de verano (en el hemisferio sur) demoran la iniciación del ápice reproductivo durante la fase juvenil (hasta fase V7, aproximadamente), el termoperíodo es el mayor determinante de la duración del ciclo (Hung *et al*., 2012). En este sentido, la conformación de un panel de asociación con las formas alélicas del gen *ZmCCT* asociadas a la respuesta fotoperiódica (alelos C/T en ambos casos, donde C revelaba un efecto aditivo positivo sobre los días a floración) fue numéricamente baja pero lo suficiente para mantener dichos alelos luego del procedimiento de filtrado por baja frecuencia (<0.05). En este caso, la resolución de un fragmento cromosómico de QTL asociado por GWAS en un gen candidato *ZMCCT* fue posible en el trabajo original mediante un GWAS en una población anidada y un mapeo fino posterior en dos familias recombinantes endocriadas del cruzamiento B73 por CML228, CML277 y Ki11, adicionalmente una población segregante entre teosinte (especie antepasado del maíz con respuesta al fotoperiodo) y W22 (línea templada con baja respuesta al fotoperiodo) permitió asociar a los dos SNPs ligados PZE1094244760 y PZE1095719172 posicionados río arriba de *ZmCCT*. Finalmente, los autores pudieron comprobar mediante líneas isogénicas segregantes en *ZmCCT* que el alelo de teosinte poseía un efecto aditivo de 9 días de demora en floración con respecto a W22. Se evidencia en el trabajo de (Hung *et al*., 2012) la importancia de contar con diferentes clases de poblaciones de mapeo para identificar un gen candidato como *ZmCCT* con

bajo efecto aditivo. La población NAM de maíz cuenta así con 26 familias derivadas de B73 con líneas templadas y tropicales (unas 6.000 familias de líneas recombinates endocriadas) genotipadas con millones de SNP, pudiendo evidenciar así expresiones de efectos de QTL bajo diferentes fondos genéticos (Yu *et al.,* 2008). El hecho de que otras líneas que tienen el efecto contrario de acortar los días a floración en días largos (CML322, Ki3, Oh43, y Tx303) tengan un haplotipo (similar secuencia de marcadores) que B73 (la línea templada de referencia), sumado a que los dos SNP ligados se encuentran por fuera del gen *ZmCCT*, indica el carácter complejo de la respuesta al fotoperiodo en maíz y su posible interacción con otros genes o QTLs como el gen que regula el tiempo a floración *Vgt1* localizado en el cromosoma 8 (Salvi *et al*., 2008). Debido a la naturaleza compleja de efecto de este QTL, como los que se manifiestan en la mayoría de caracteres agronómicos relacionados a rendimiento (Xu y Crouch, 2008), estos tipos de estudio son de mucha utilidad como modelos para ser aplicados en la resolución de problemáticas locales que permitan concretar la aplicación de la selección asistida por marcadores en programas de mejoramiento locales que cuentan con recursos de genotipado y fenotipado más limitados.

CONCLUSIONES

En base a los resultados obtenidos en el presente trabajo se puede concluir:

- Fue posible reproducir con ambos *sets* de datos de SSR y SNP los resultados de agrupamiento de las líneas del panel de maíz originalmente clasificados con SSR. Las líneas templadas se agruparon así en sus respectivos grupos heteróticos SS y NSS y al resto según su origen (semitropical y tropical) y tipo de grano (dulce, reventador).
- El método de mapeo del gen candidato utilizando similares secuencias de la región A6, río arriba, de *ZmCCT* permitió reproducir la detección de los SNPs ligados con la respuesta fotoperiódica de los tiempos a floración y espigazón en maíz.
- El método de mapeo GWAS con diferentes *sets* de datos y población utilizada originalmente, solo pudo identificar *set* de SNPs ligados a la respuesta al fotoperíodo distales a *ZmCCT* (distantes entre 5 a 25 millones pb hacia el telómero) pero ninguno cercano a los 95 millones de pb donde se posiciona *ZmCCT*.
- Los métodos de mapeo GLM y MLM resultaron en diferentes resultados según el ajuste por estructura (Q) y parentesco (K). En ambos casos, el ajuste por Q y K redujo la detección de SNP significativos implicando así una reducción de falsos positivos.
- En el caso de MLM el empleo de datos fenotípicos promedios (BLUE) y set de datos completos por ambientes (11 ambientes) produjo resultados diferentes. El empleo de datos fenotípicos promedios redujo la detección de SNP significativos.
- Fue posible una reproducción parcial de los resultados originales informados según los marcadores, poblaciones, modelos y ajustes utilizados.
- Estos resultados demuestran la importancia de contar con diferentes fuentes de poblaciones de mapeo y procedimientos de análisis estadístico con el fin de verificar y validar los resultados obtenidos en el mapeo de QTL.

ANEXOS

Anexo 1. Resultados de estructura genética del panel de líneas de maíz de USDA (Liu *et al*., 2003). (NSS) líneas templadas denominadas *Non-Stiff Stalk* cuya línea represnetativa es MO17, (SS) líneas templadas *Stiff Stalk* cuya línea representativa es B73, (TS) líneas tropicales o semitropicales, (*Sweet corn*) maíz dulce, (*Popcorn*) maíz pisingallo.

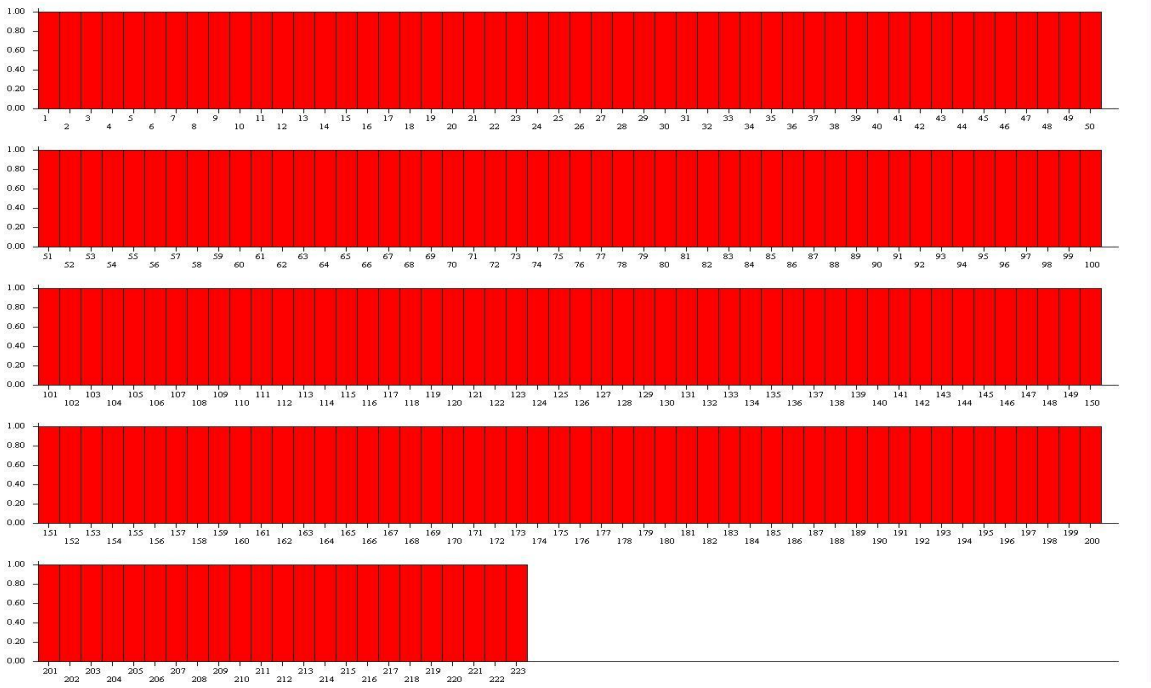
NSS	Hy:T8:Wf9	CI21E, H49, Hy, Mo1W, Pa875, Pa880, T8, Va17, Va14, Va22, Va35, Va102, W64A, Wf9
	M14:Oh43	A619, Gn2, H95, M14, Oh40B, Oh43, Oh43E, PA762, Va26, Va85
	CO109:Mo17	A556, A682, CI.187-2, CO109, CO220, K187, Mo17, MS1334, ND246, W401
	C103	B57, C103, C123, DE2, L317, L1546, NC258, NC262
	Ga:SC	4226, F44, Ga209, GT112, SC357, SC213R, SC213
	NSS-X	38-11, A239, A659, AR4, CM7, CM37, R168, Mo44, MS71, NC260, PA884P, R4, R177, W22
	K64W	33-16, CI.31A, CI.64, CI.66, CI.7, E2558W, Ky21, K55, K64, M162W
	NSS-mixed	A554, A654, B2, B52, B70, B77, B97, B103, CO106, CO125, F6, Fe2, H99, Mt42, N6, Os420, Pa91, R109B, SD44, T234, W153R
SS	B14A	A214N, A632, A634, A635, A665, B14A, B64, B68, CM105, CM174, H91
	B37	B37, B76, H84, NC250
	N28	N28, N28Ht
	B73	A679, A680, B73, B84, B104, B109, NC328, NC372, R229
	SS-mixed	A641, De811, H100, N192, N196, NC294, NC368
TS	TZI	A6, CML52, CML238, CML287, NC358, Q6199, Tzi8, Tzi9, Tzi10, Tzi18
	Suwan	B96, CML69, CML228, CML349, Ki3, Ki9, Ki11, Ki14, Ki44, Ki2007
	CML-late	CML5, CML9, CML61, CML103, CML220, CML254, CML258, CML261, CML264, CML314, Tx601
	CML-early	CML14, CML247, CML311, CML321, CML322, CML331, CML332
	NC	NC296, NC298, NC304, NC336, NC338, NC348, NC350, NC352, NC354
	CML-P	CML10, CML11, CML45, CML277, CML281, CML333, CML341
	TS-mixed	CML38, CML108, NC300, NC356
Sweet corn		Ia2132, Il14H, Il101t, Il677a, P39
Popcorn		HP301, I29, IDS28, IDS69, SA24, Sg18, Sg1533
Mixed		A188, A272, A441-5, A656, B79, B94, B105, B164, C49A, CML77, CML91, CML92, CML218, CML323, CML328, CMV3, CO159, CO255, D940Y, DE3, EP1, F2, F2834T, F7, Hi27, I137TN, I205, IDT, Ki43, Ky226, Ky228, L578, Le23, Le773, M37W, Mo18W, Mo24W, Mp339, MS153, N7A, NC264, NC320, NC360, NC362, NC364, NC366, NC370, Oh7B, Oh603, SC55, SD40, SD46, T232, TEA, Tx303, Tzi11, Tzi16, Tzi25, U267Y, Va99, W117, W117Ht, W182B

Anexo 2. Resultados de Structure con SSR

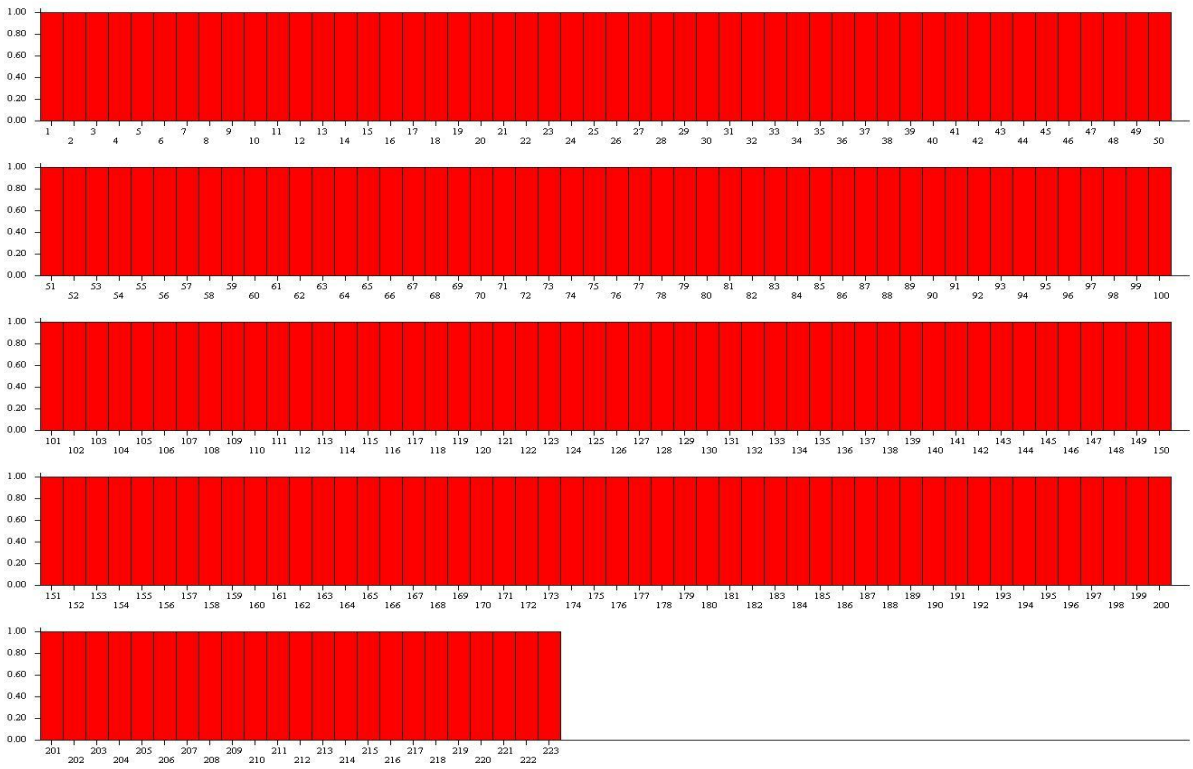
ParameterSet	RunName	K	LnP(D)	Var[LnP(D)]
Tesisluciano	Tesisluciano_run_3	1	-86442.6	856.6
Tesisluciano	Tesisluciano_run_2	1	-86437	846
Tesisluciano	Tesisluciano_run_1	1	-86450.8	871.2
Tesisluciano	Tesisluciano_run_6	2	-80879.9	1986.8
Tesisluciano	Tesisluciano_run_5	2	-80862.9	1951.3
Tesisluciano	Tesisluciano_run_4	2	-80866.5	1951.7
Tesisluciano	Tesisluciano_run_9	3	-77172.5	2397.8
Tesisluciano	Tesisluciano_run_8	3	-77128.1	2311
Tesisluciano	Tesisluciano_run_7	3	-77146.6	2351.3
Tesisluciano	Tesisluciano_run_12	4	-74921.3	2911.6
Tesisluciano	Tesisluciano_run_11	4	-75512.7	2821.2
Tesisluciano	Tesisluciano_run_10	4	-75254	2807.4
Tesisluciano	Tesisluciano_run_15	5	-73292.2	3302
Tesisluciano	Tesisluciano_run_14	5	-73489.1	3482.1
Tesisluciano	Tesisluciano_run_13	5	-73326	3350.7
Tesisluciano	Tesisluciano_run_18	6	-71829.6	3830.7

Tesisluciano	Tesisluciano_run_17	6	-71984.9	3734.5
Tesisluciano	Tesisluciano_run_16	6	-71916.2	3884

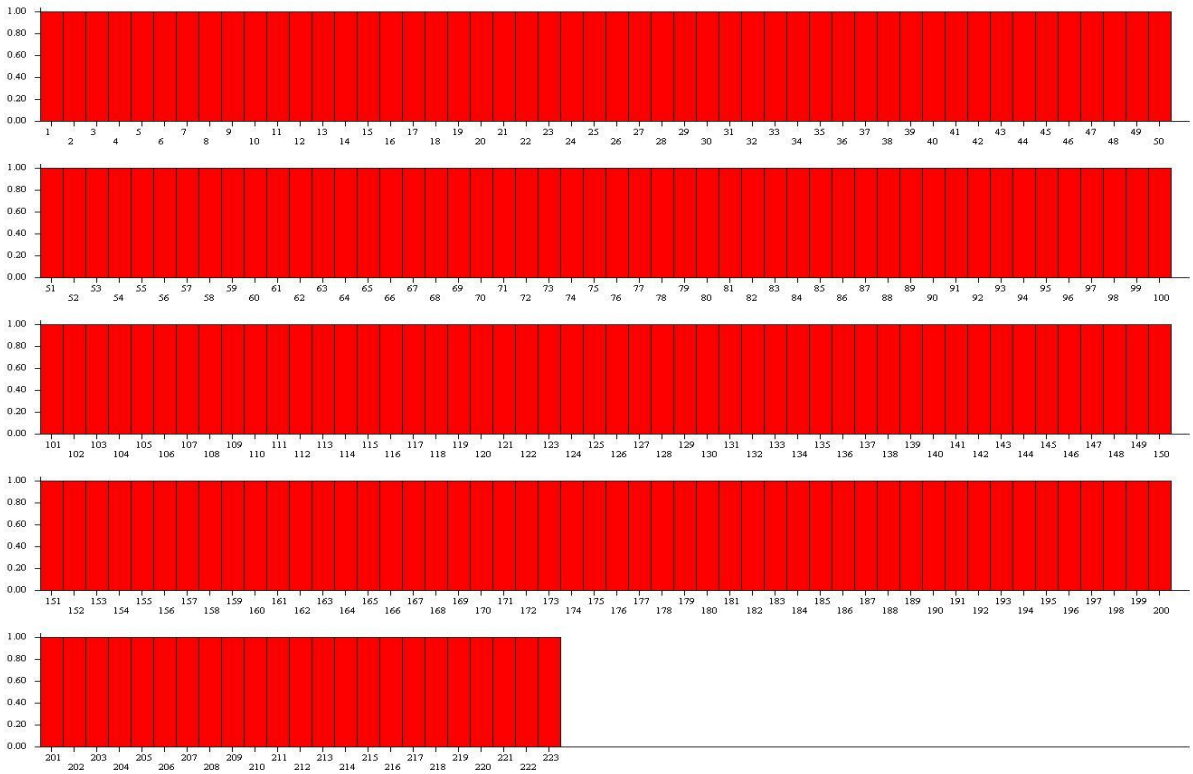
Gráficos ordenados según número de subpoblaciones (K=1 a 6) y repeticiones de corridas (R= 1 a 3)



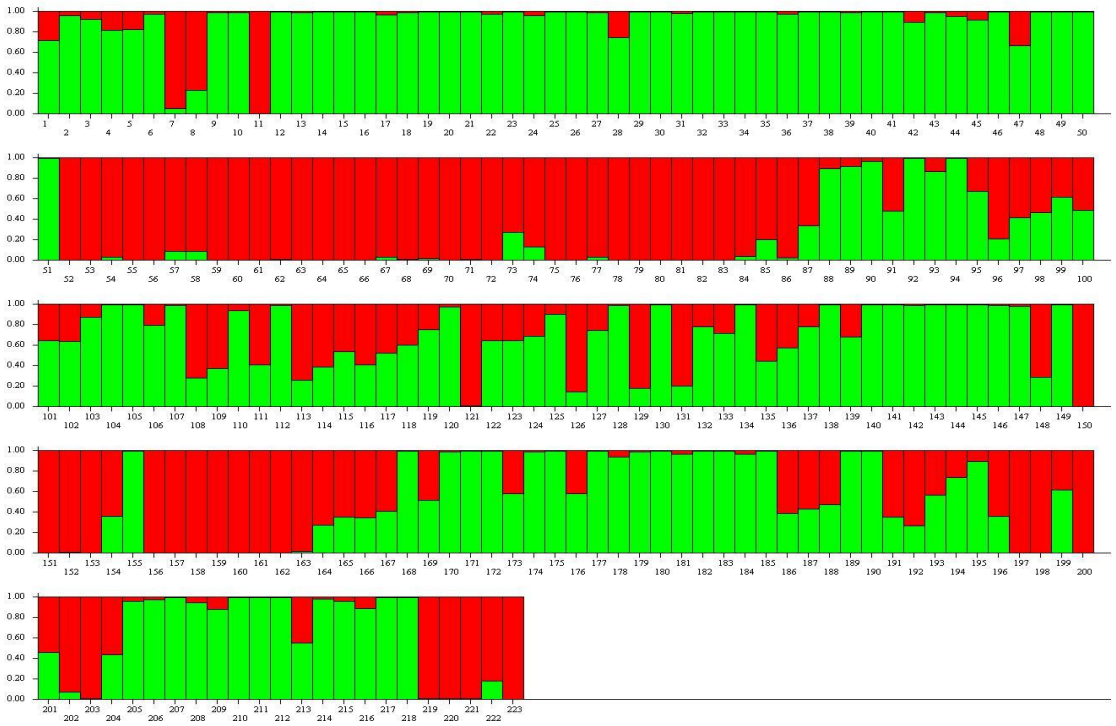
K=1, R 1



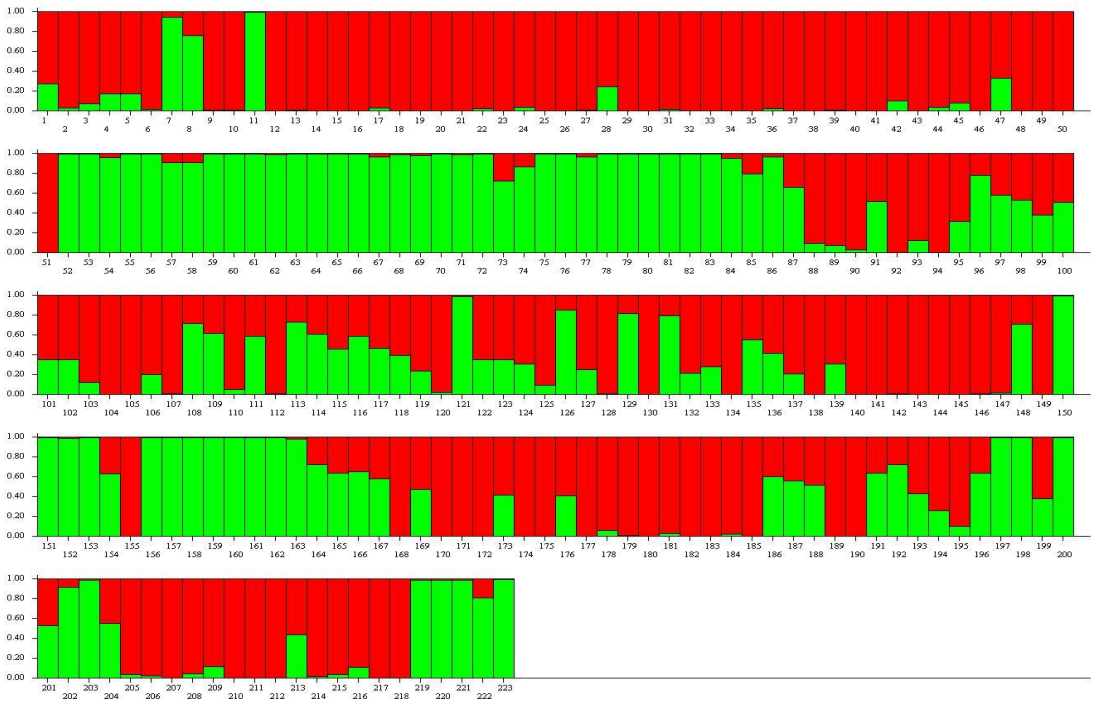
K=1, R 2



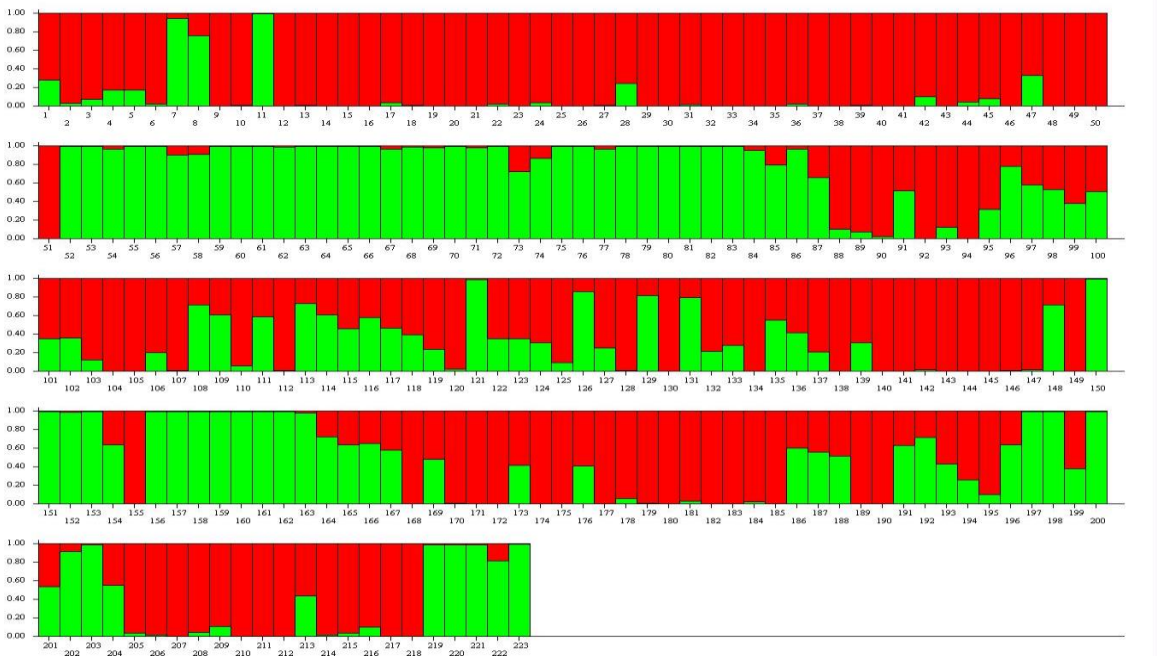
K=1, R 3



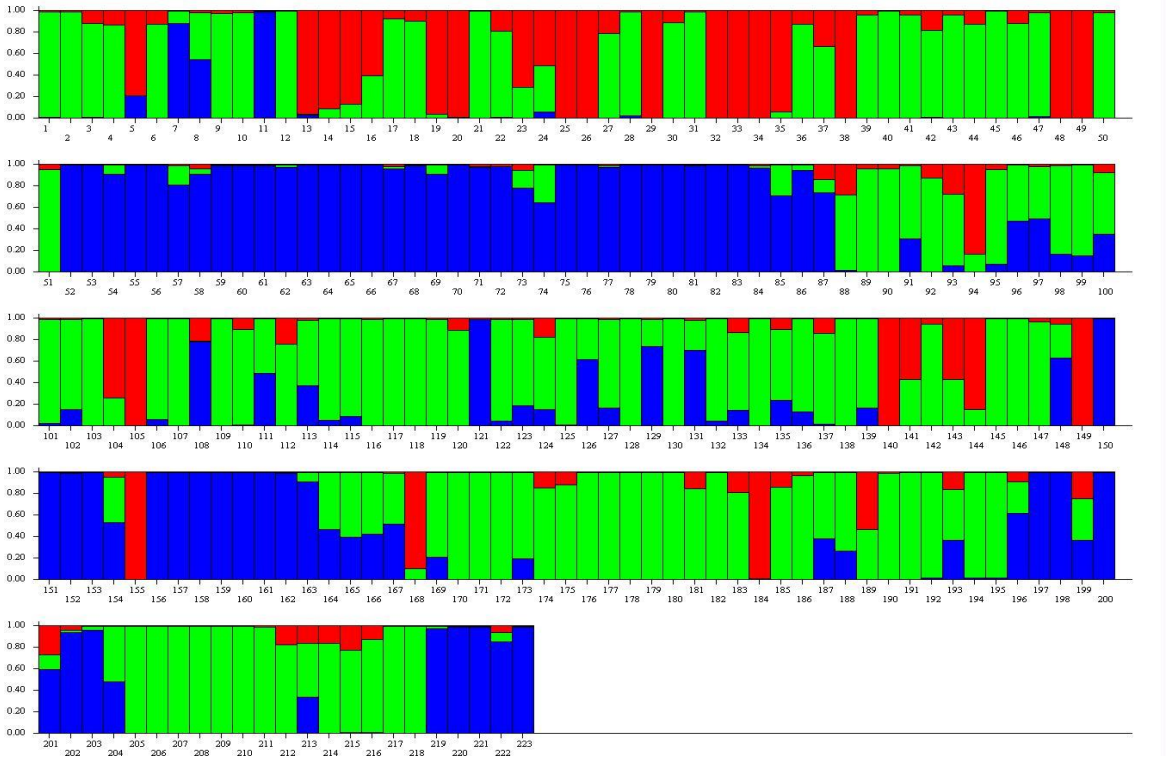
K=2, R=1



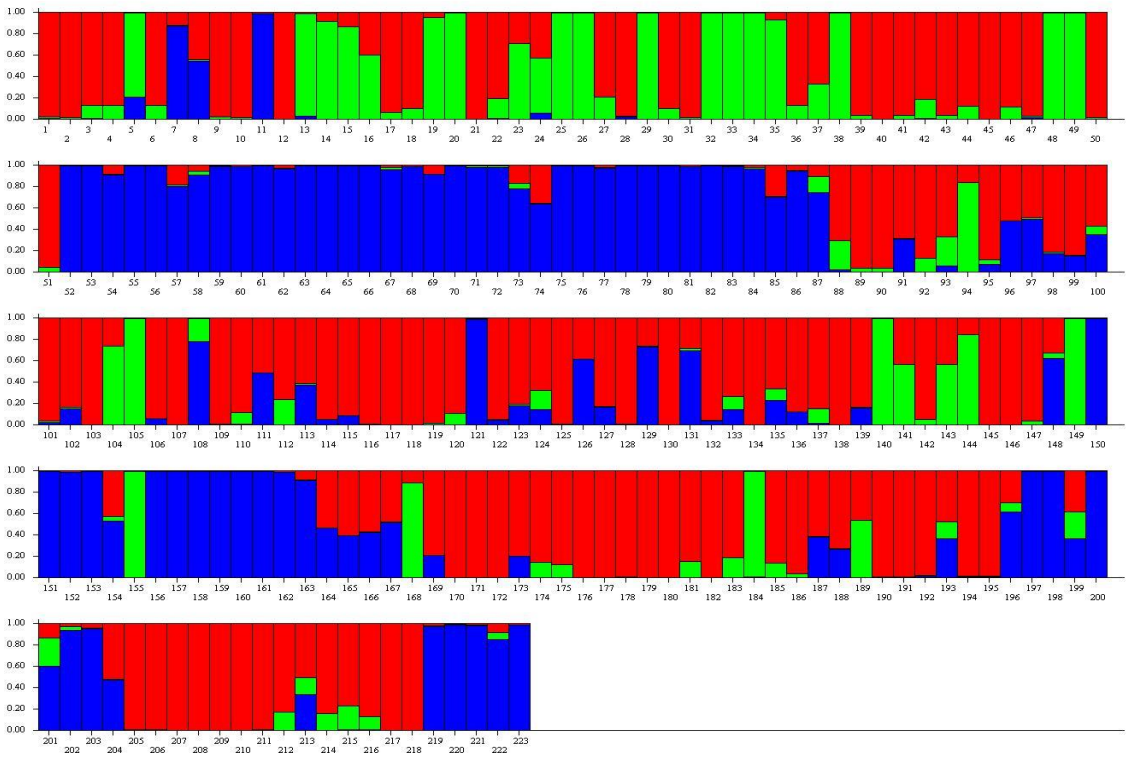
K=2, R=2



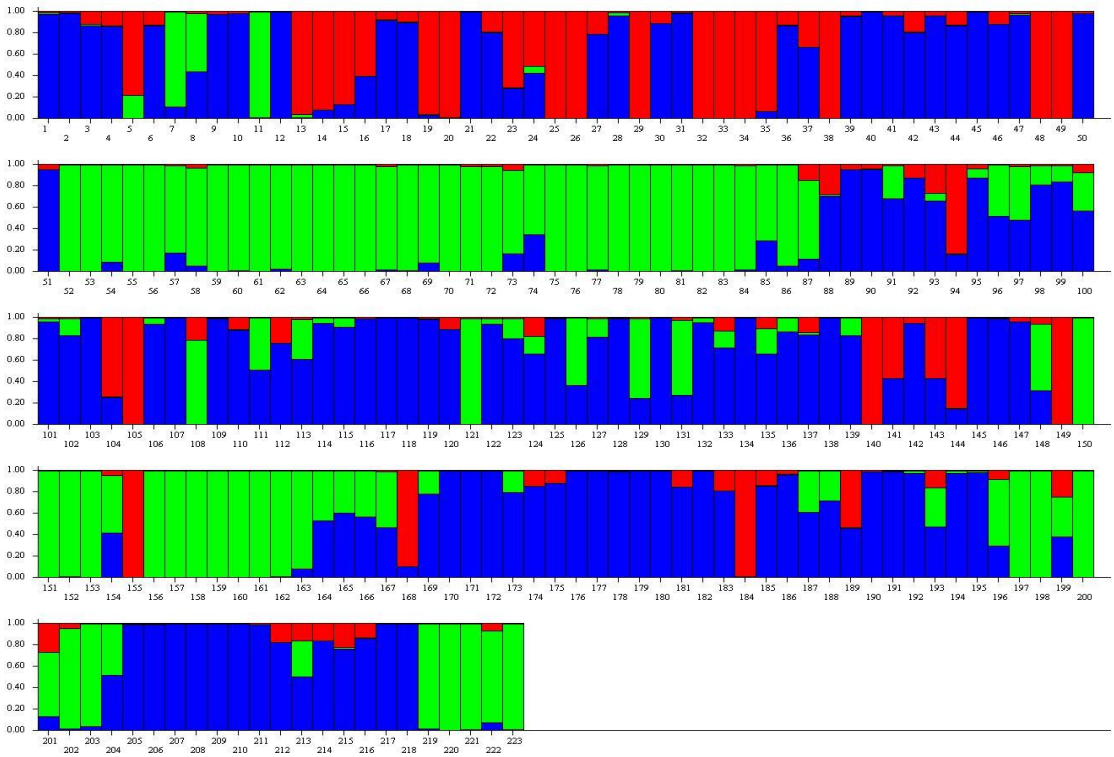
K=2, R=3



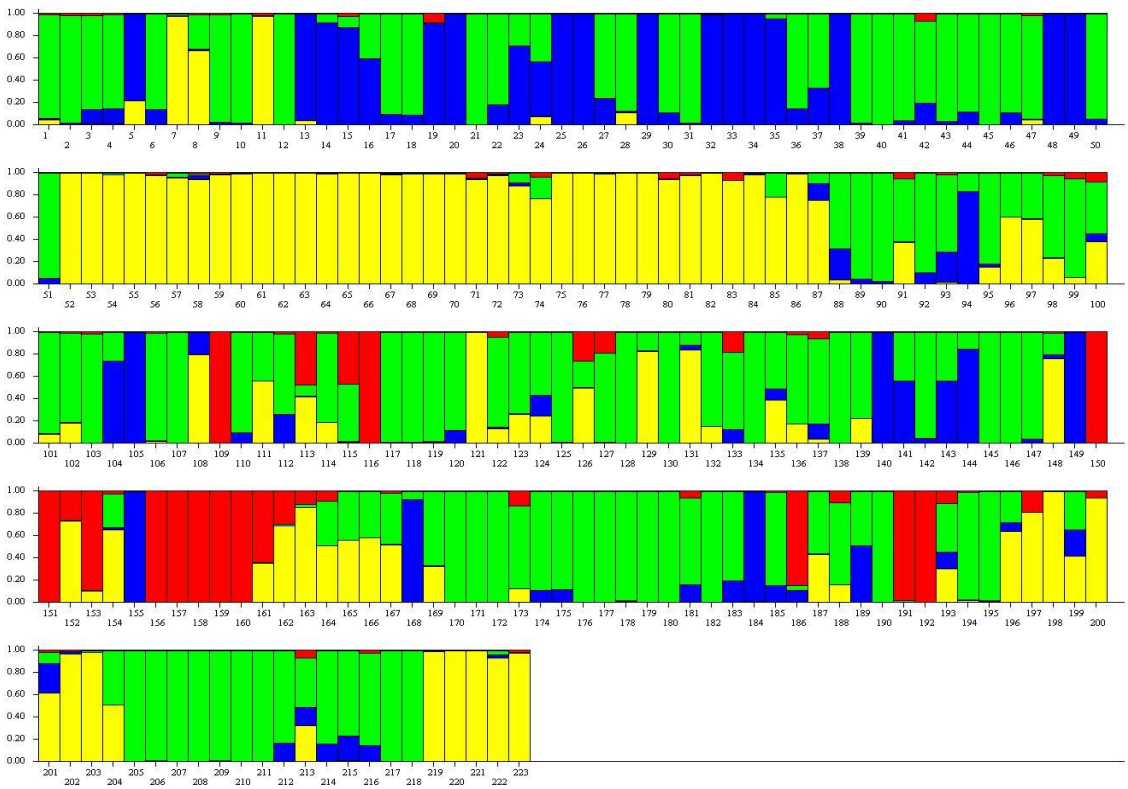
K=3, R=1



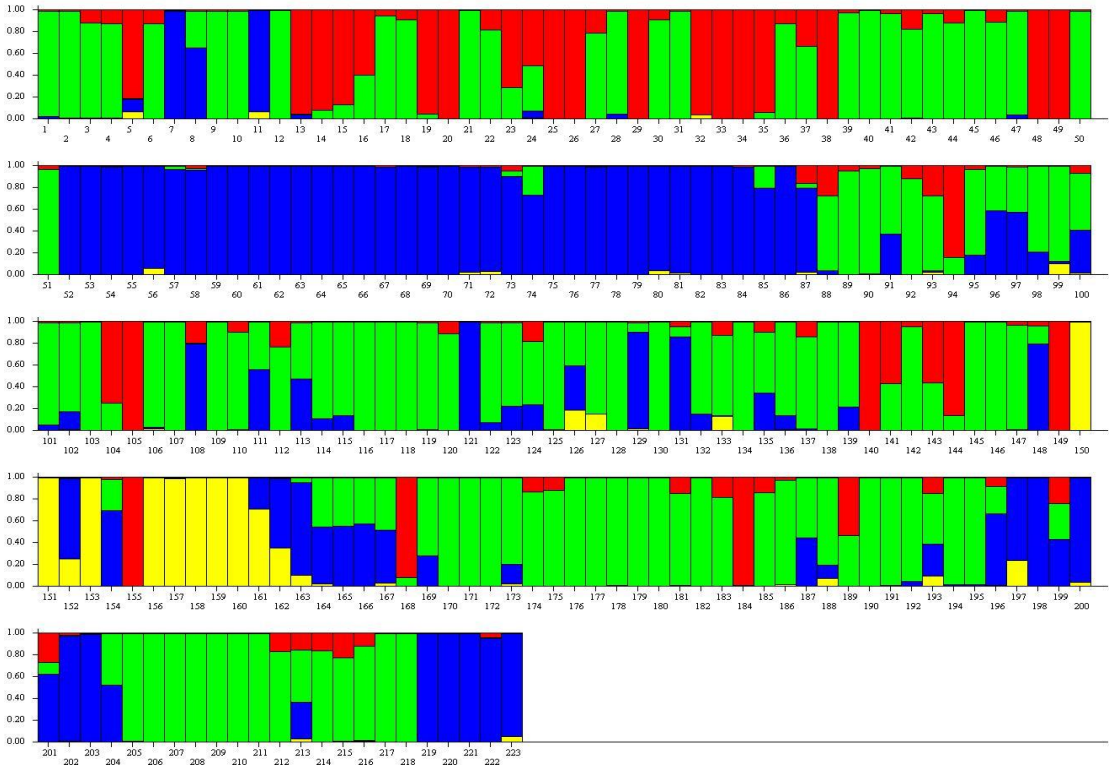
K=3, R=2



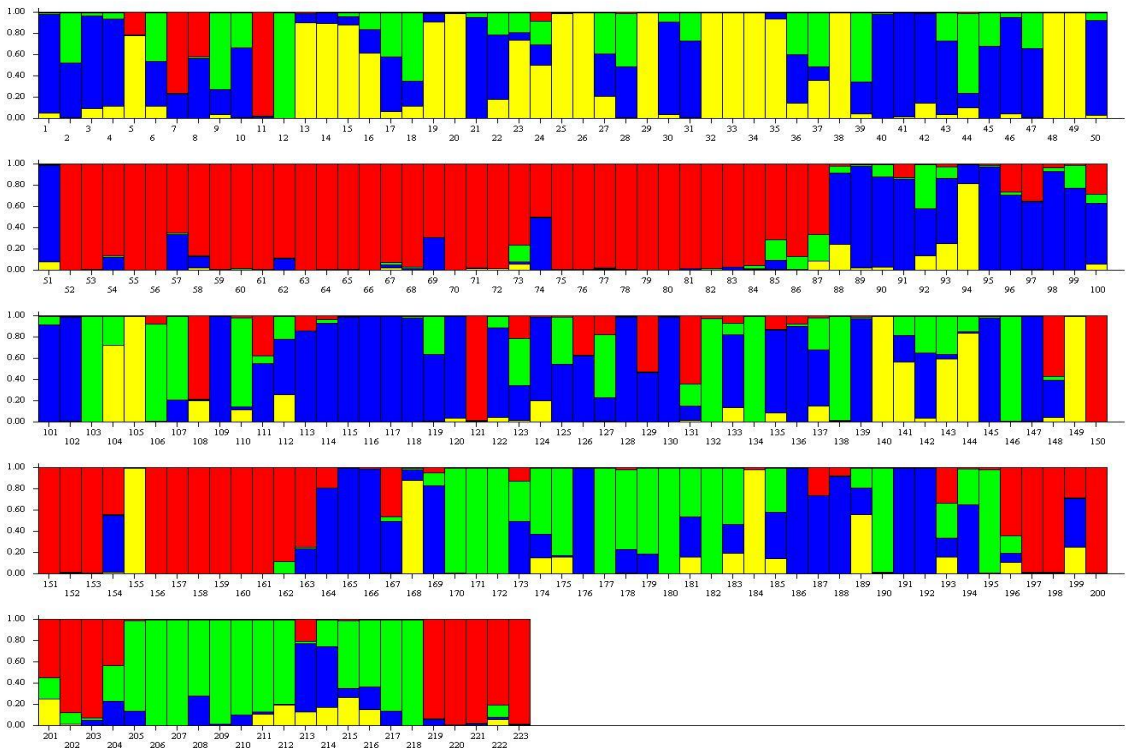
K=3, R=3



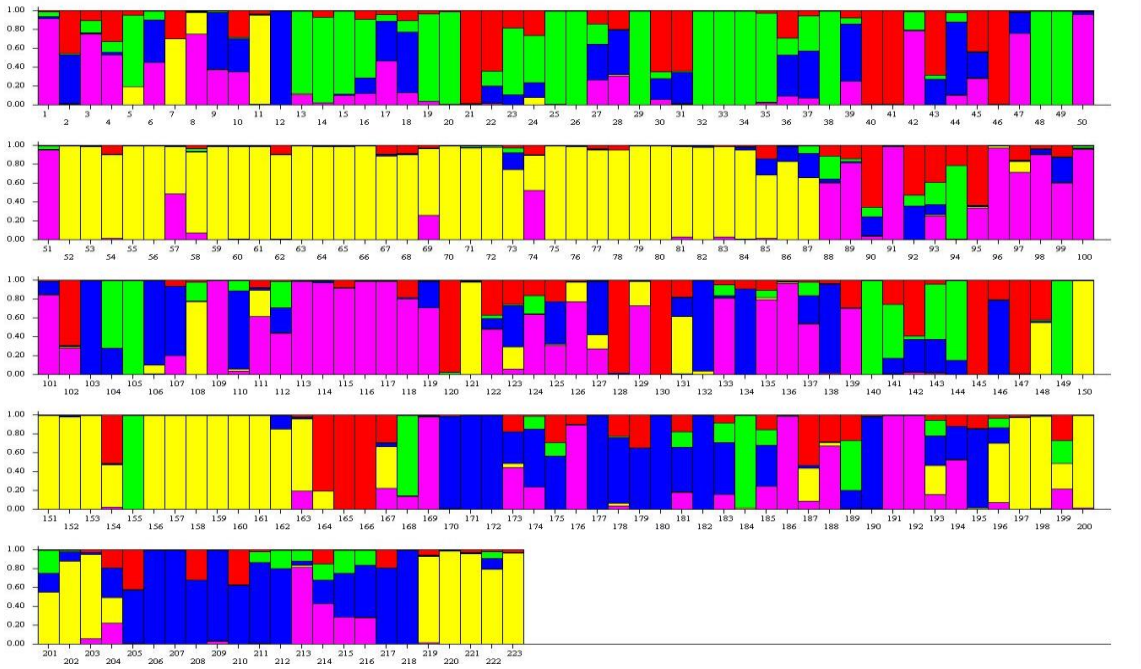
K=4, R=1



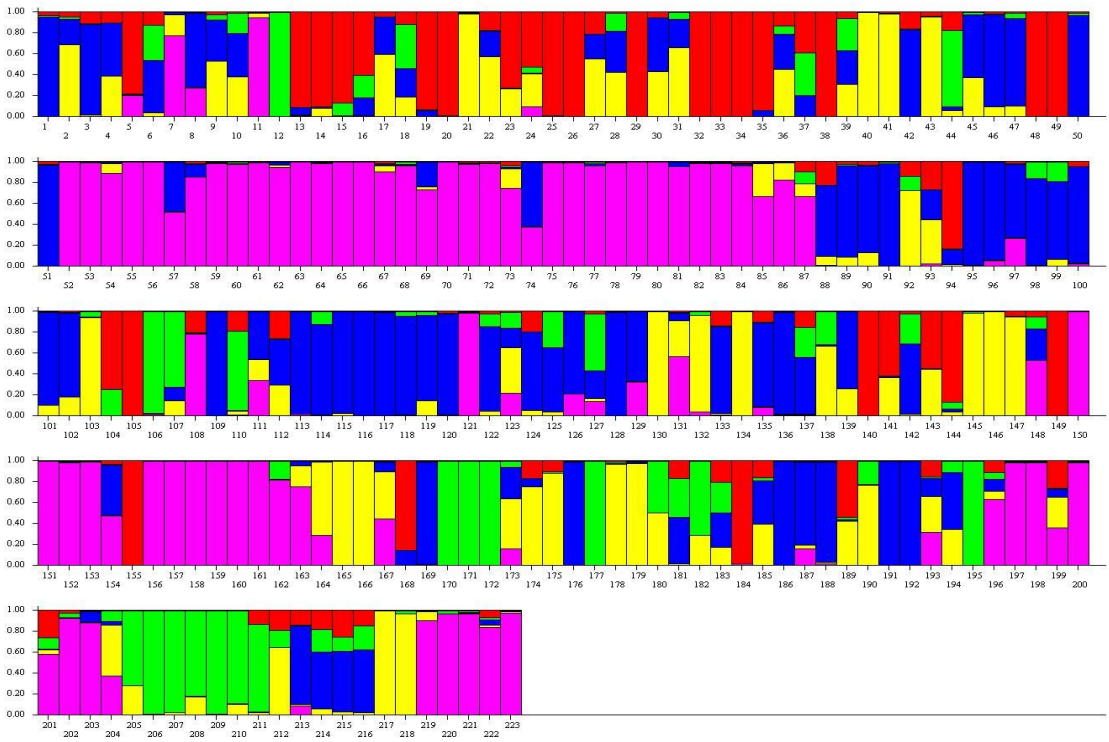
K=4, R=2



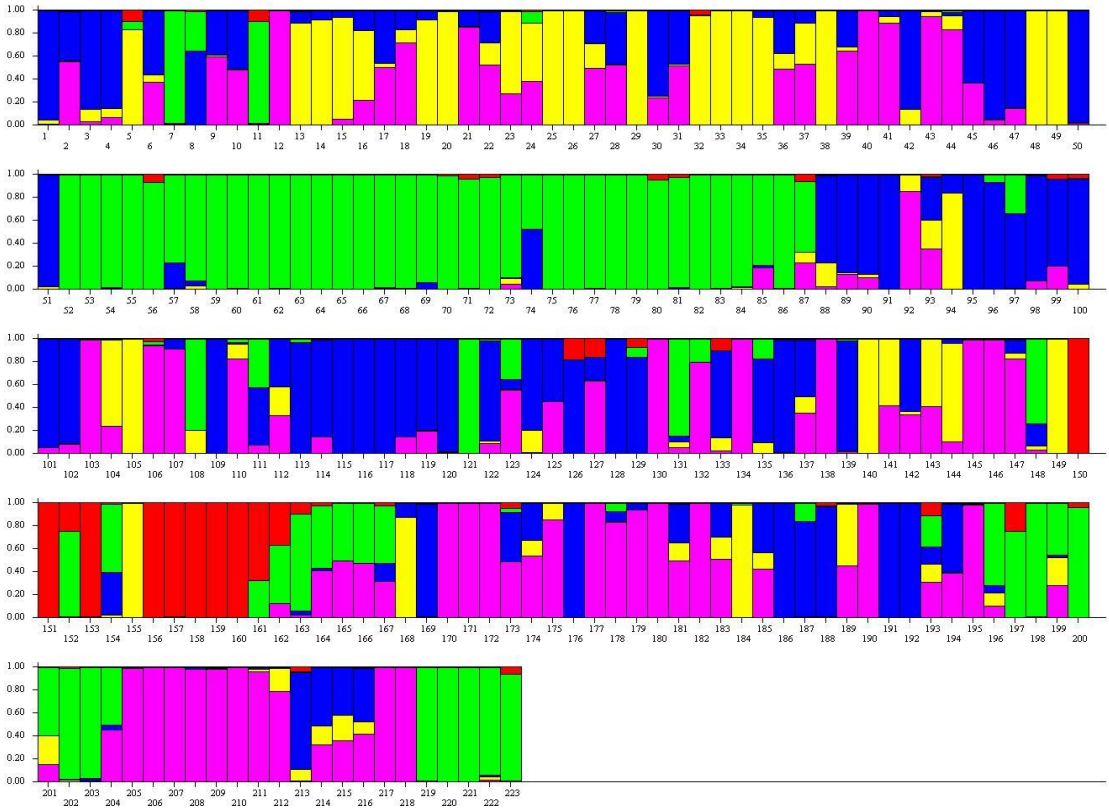
K=4, R=3



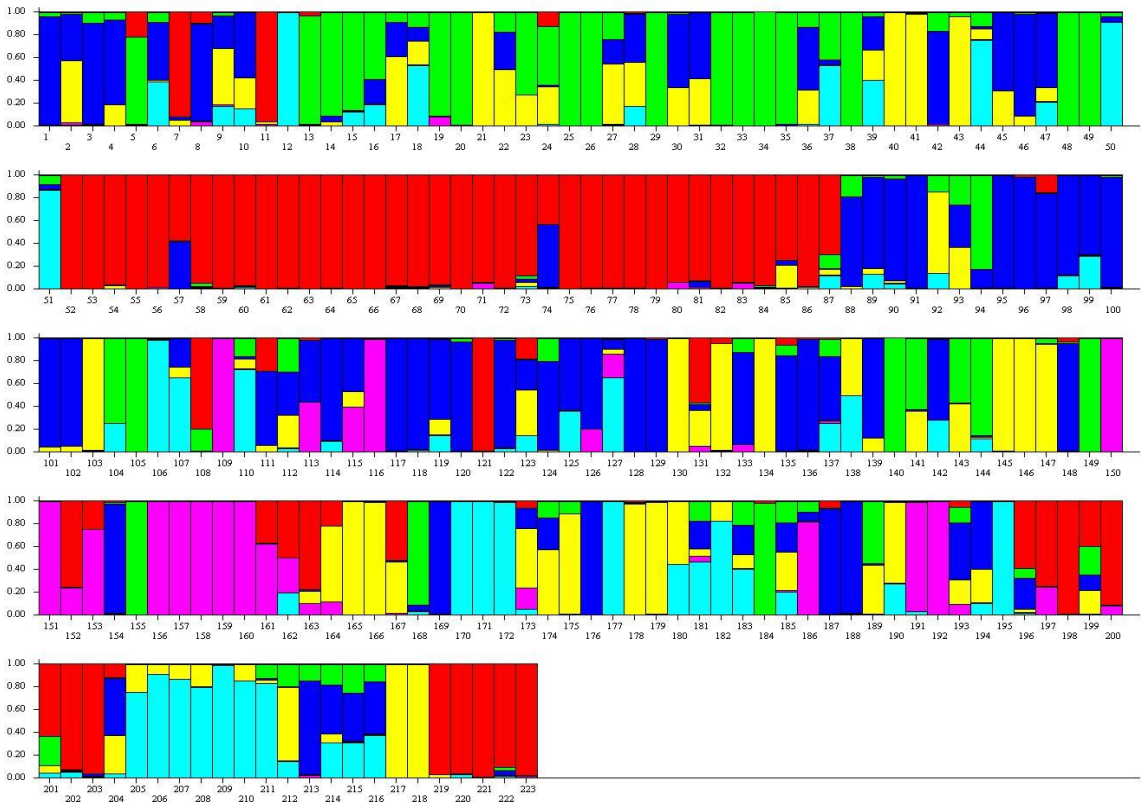
K=5, R=1



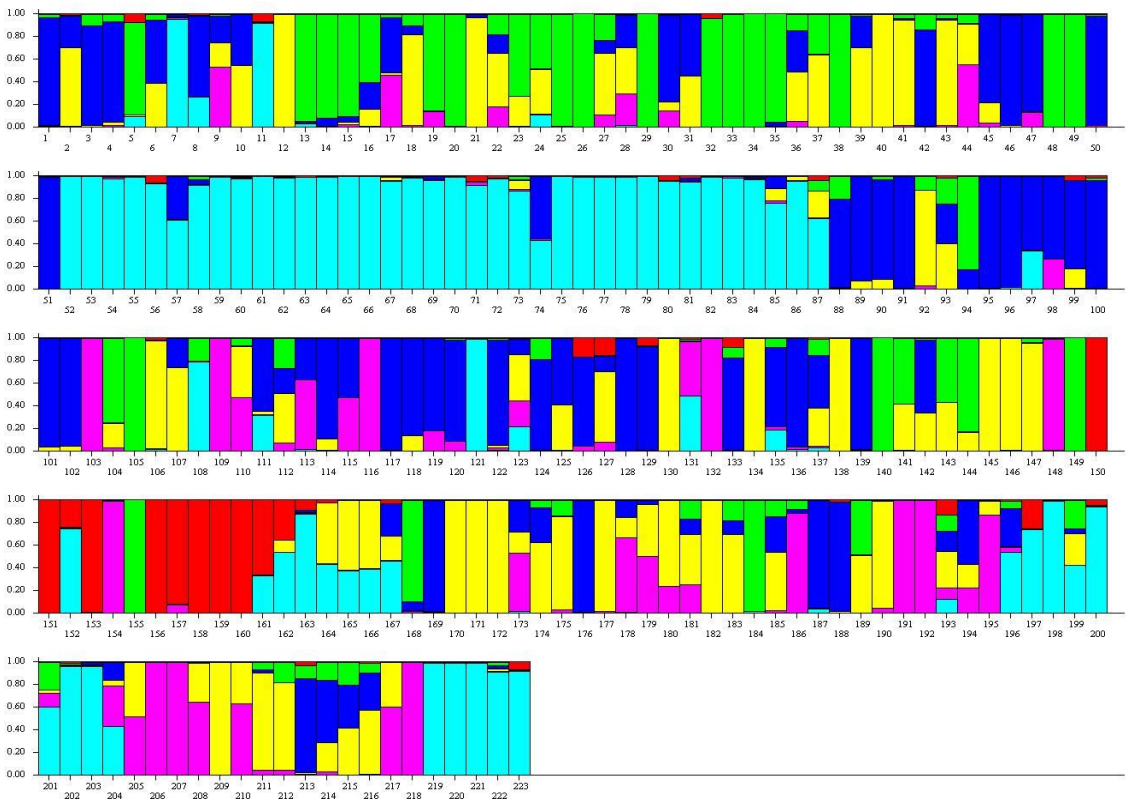
K=5, R=2



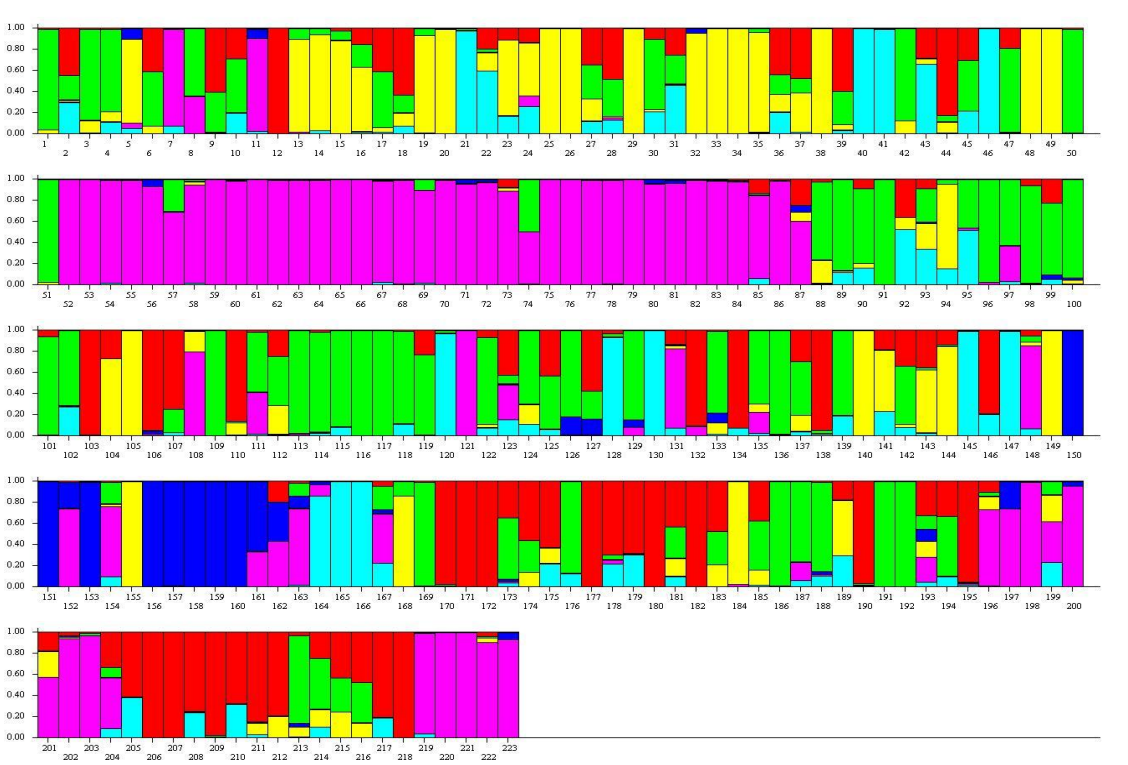
K=5, R=3



K=6, R=1



K=6, R=2



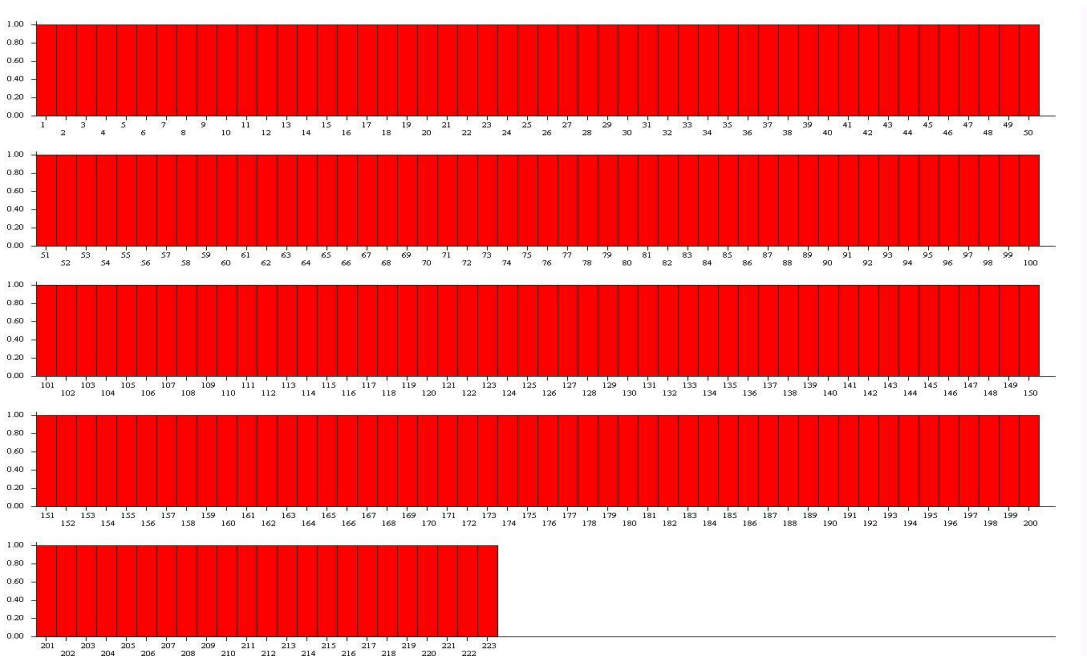
K=6, R=3

Anexo 3. Resultados de Structure con SNPs

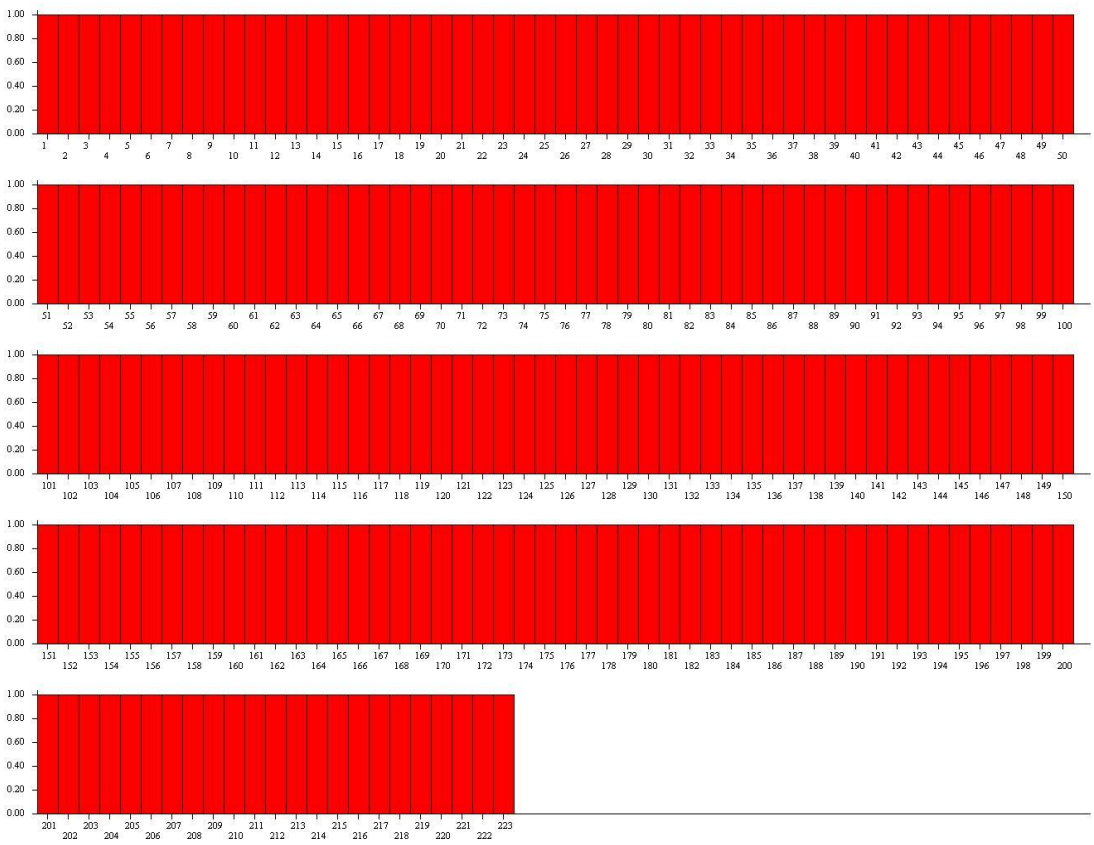
ParameterSet	RunName	K	LnP(D)	Var[LnP(D)]
SNPtesinaluciano	SNPtesinaluciano_run_4	1	-130942.9	274
SNPtesinaluciano	SNPtesinaluciano_run_3	1	-130942.9	274.7
SNPtesinaluciano	SNPtesinaluciano_run_2	1	-130941.8	271.8
SNPtesinaluciano	SNPtesinaluciano_run_1	1	-130942.6	273.8
SNPtesinaluciano	SNPtesinaluciano_run_7	2	-124568.8	693.7
SNPtesinaluciano	SNPtesinaluciano_run_6	2	-124573.4	702.3
SNPtesinaluciano	SNPtesinaluciano_run_5	2	-124579.2	714.5
SNPtesinaluciano	SNPtesinaluciano_run_9	3	-120640.6	1149.7
SNPtesinaluciano	SNPtesinaluciano_run_8	3	-120644.6	1151.9
SNPtesinaluciano	SNPtesinaluciano_run_10	3	-120642.4	1148.2
SNPtesinaluciano	SNPtesinaluciano_run_13	4	-118944.1	1503.4
SNPtesinaluciano	SNPtesinaluciano_run_12	4	-118878.1	1646.1
SNPtesinaluciano	SNPtesinaluciano_run_11	4	-118876	1601.8
SNPtesinaluciano	SNPtesinaluciano_run_16	5	-116555.4	2251.4
SNPtesinaluciano	SNPtesinaluciano_run_15	5	-117035.4	1952
SNPtesinaluciano	SNPtesinaluciano_run_14	5	-117144.9	2024.5
SNPtesinaluciano	SNPtesinaluciano_run_19	6	-115172.7	2175.6

SNPtesinaluciano	SNPtesinaluciano_run_18	6	-115187	2195.8
SNPtesinaluciano	SNPtesinaluciano_run_17	6	-115611.8	2388.3

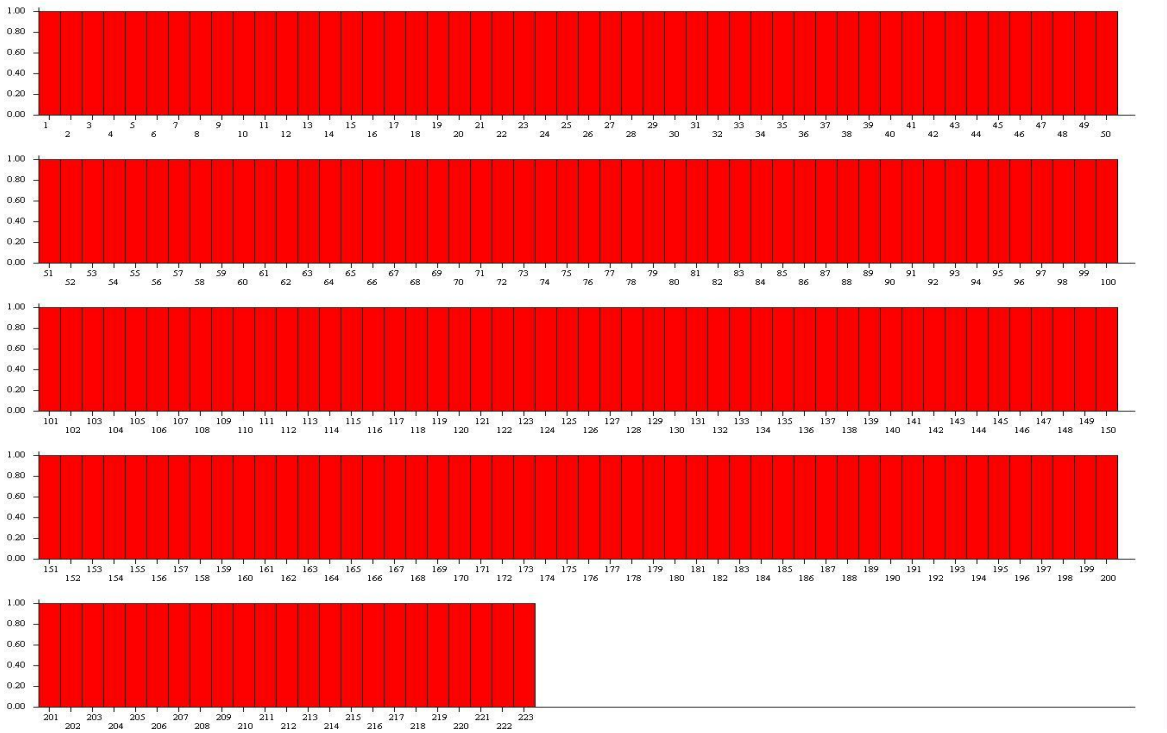
Gráficos ordenados según número de subpoblaciones (K=1 a 6) y repeticiones de corridas (R= 1 a 3)



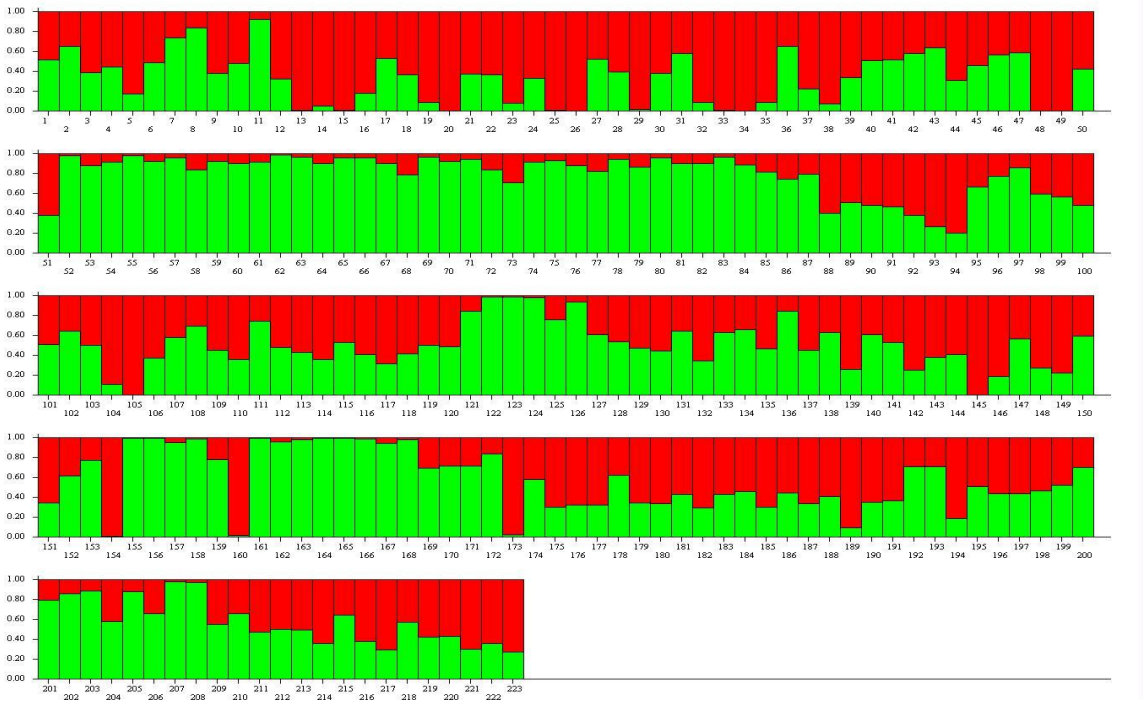
K=1, R=1



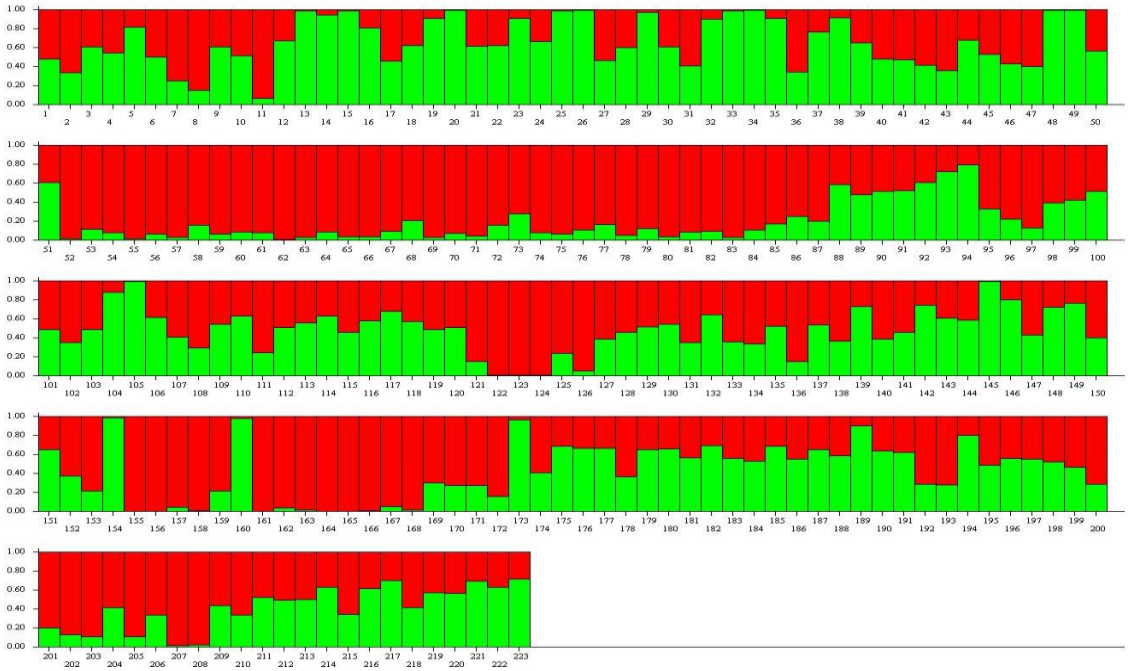
K=1, R=2



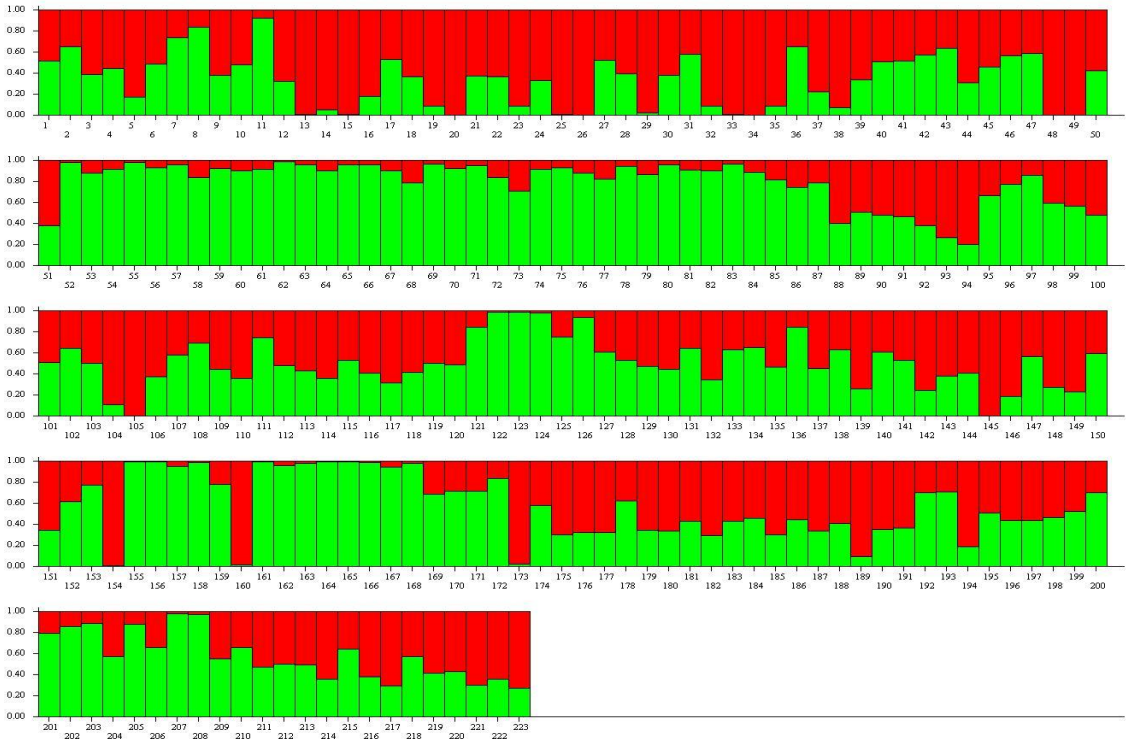
K=1, R=3



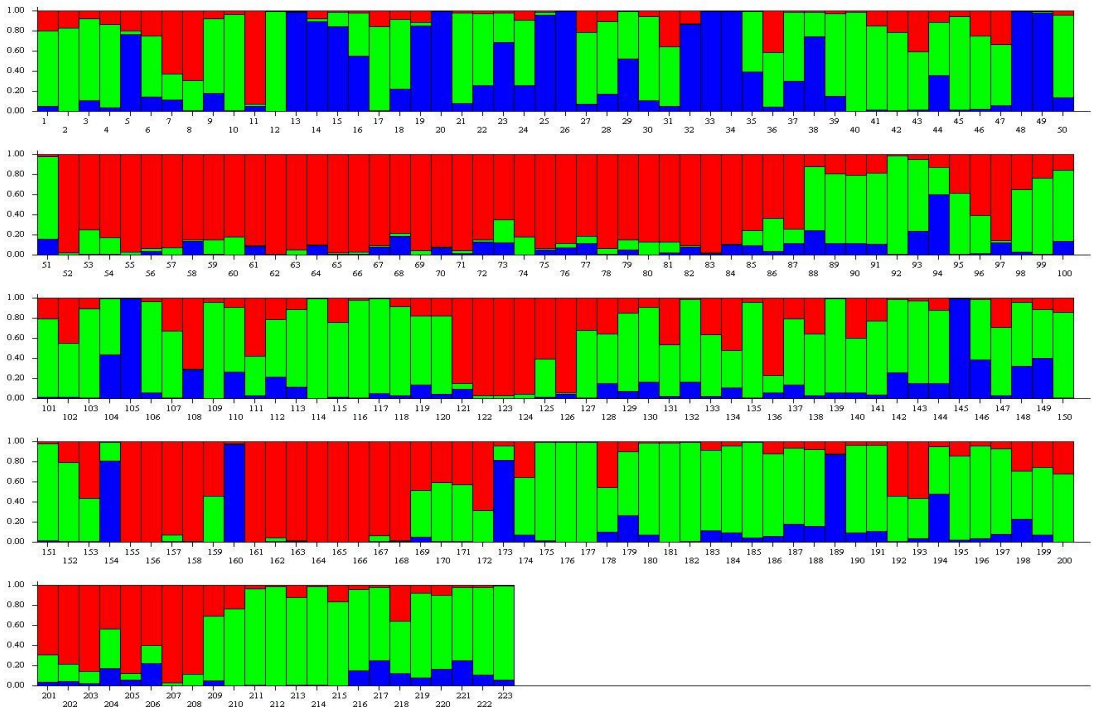
K=2, R=1



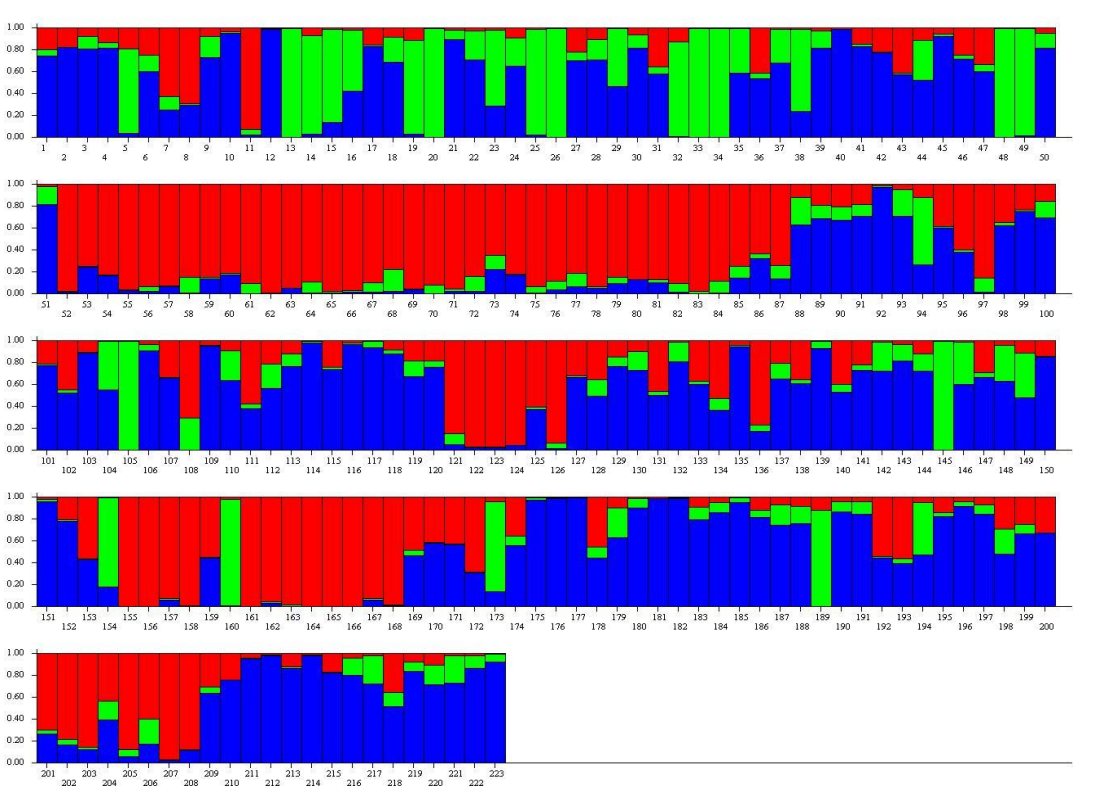
K=2, R=2



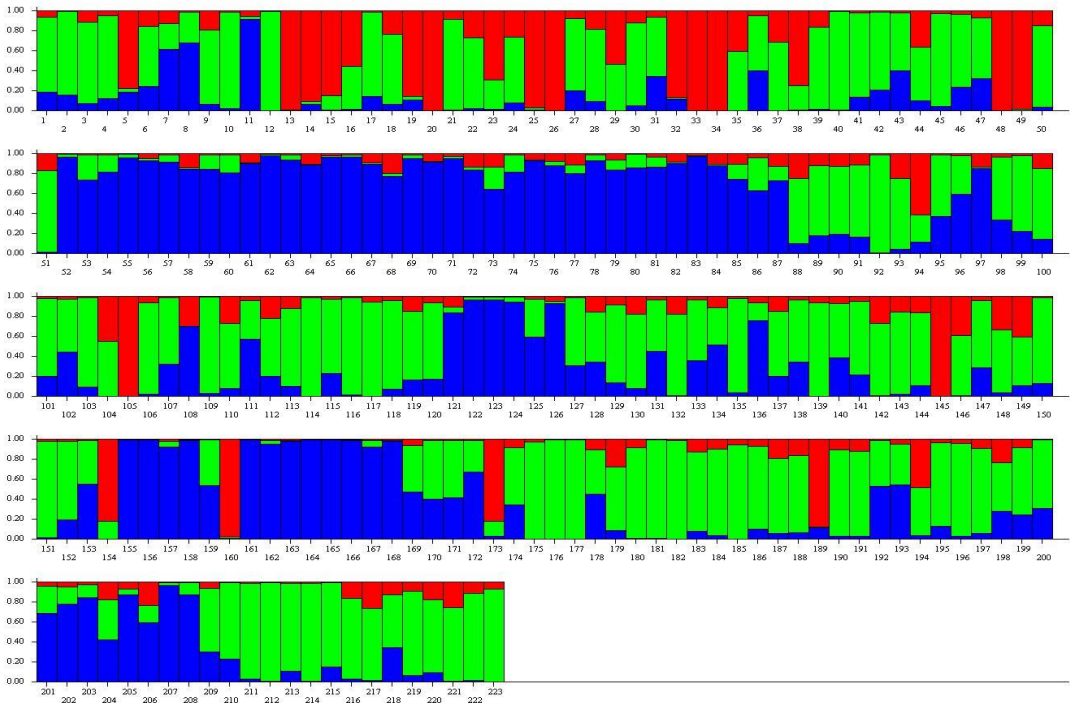
K=2, R=3



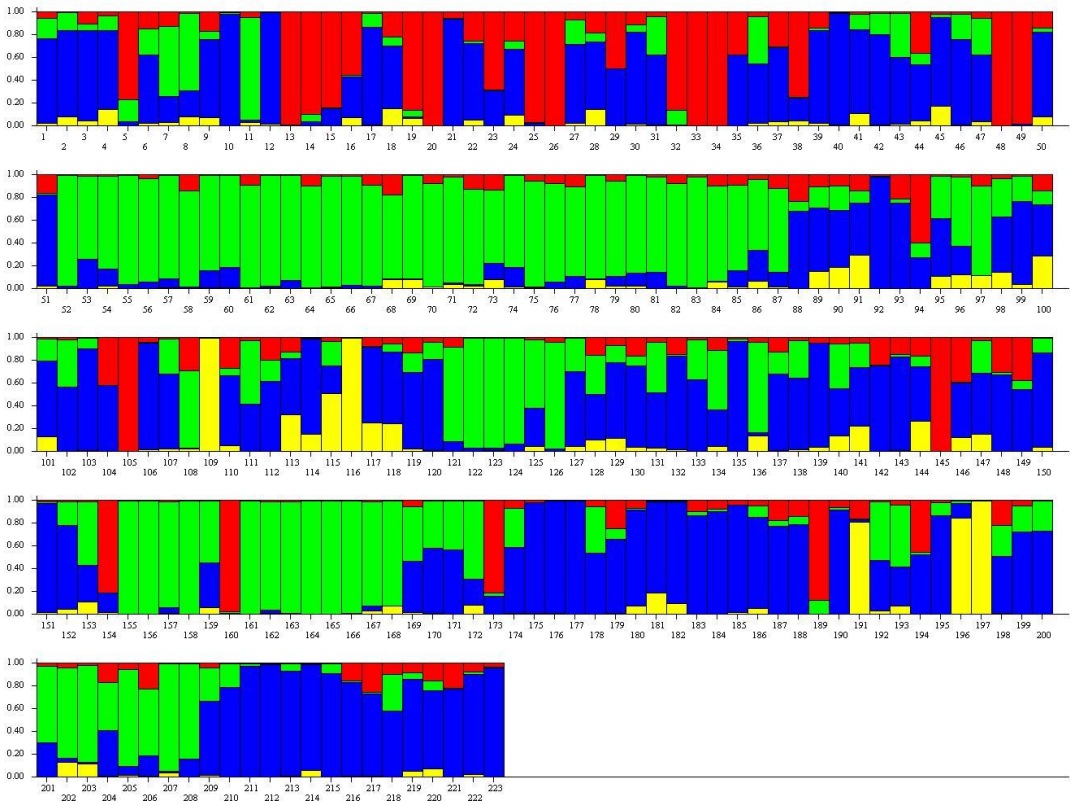
K=3, R=1



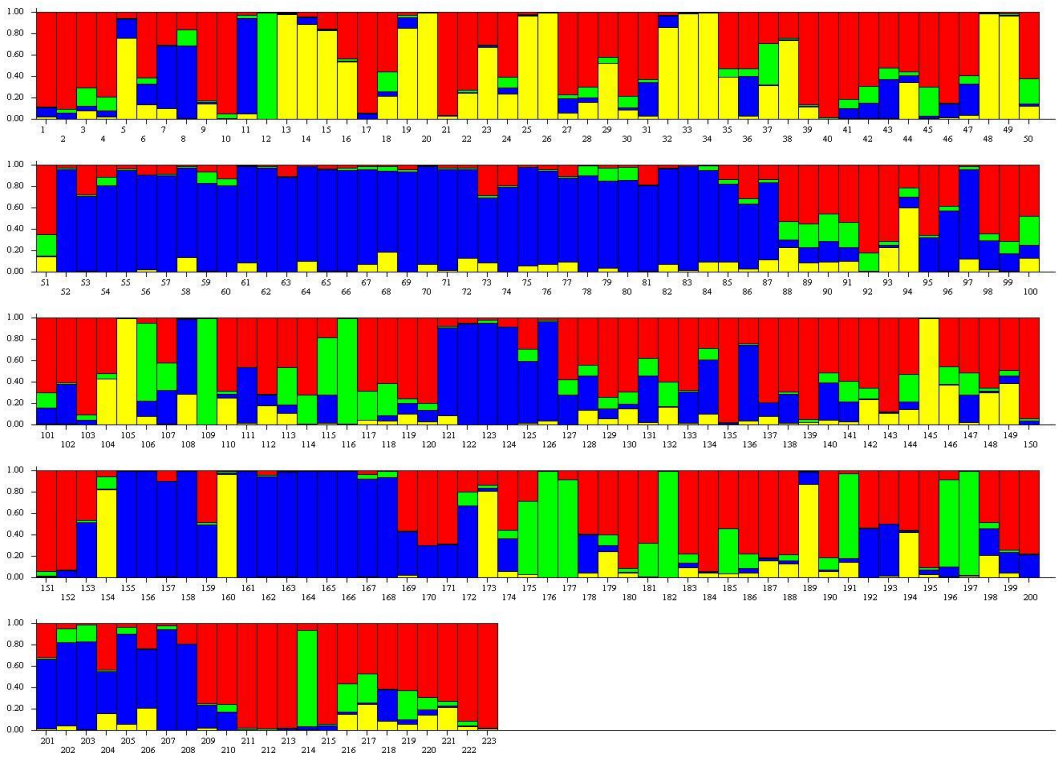
K=3, R=2



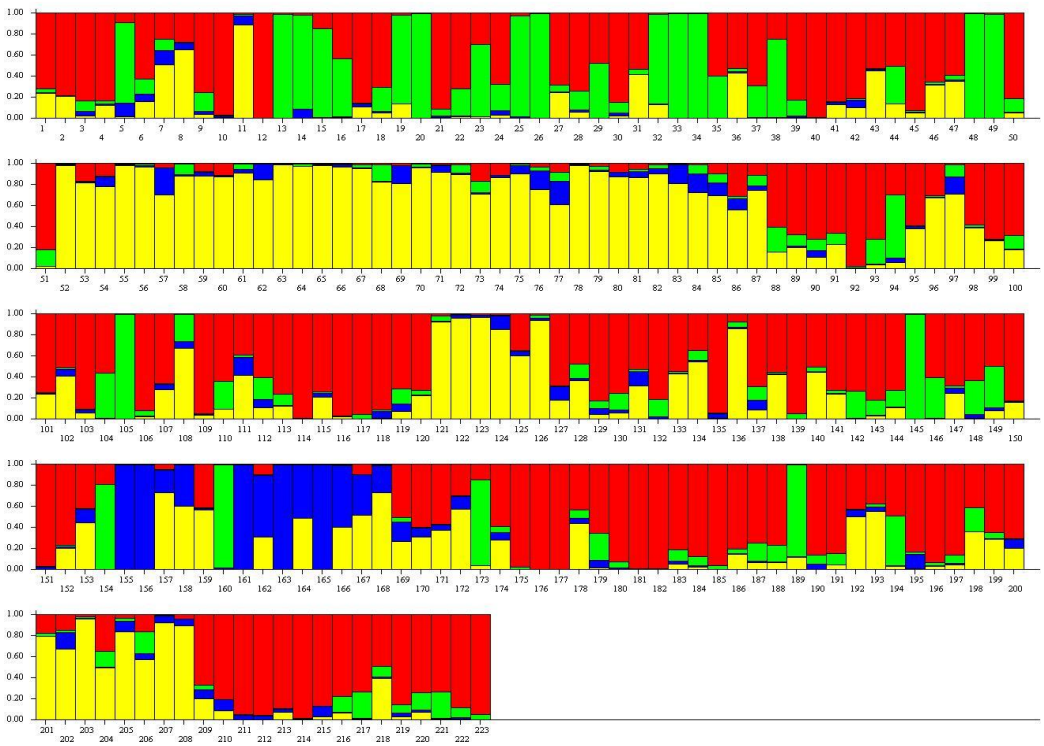
K=3, R=3



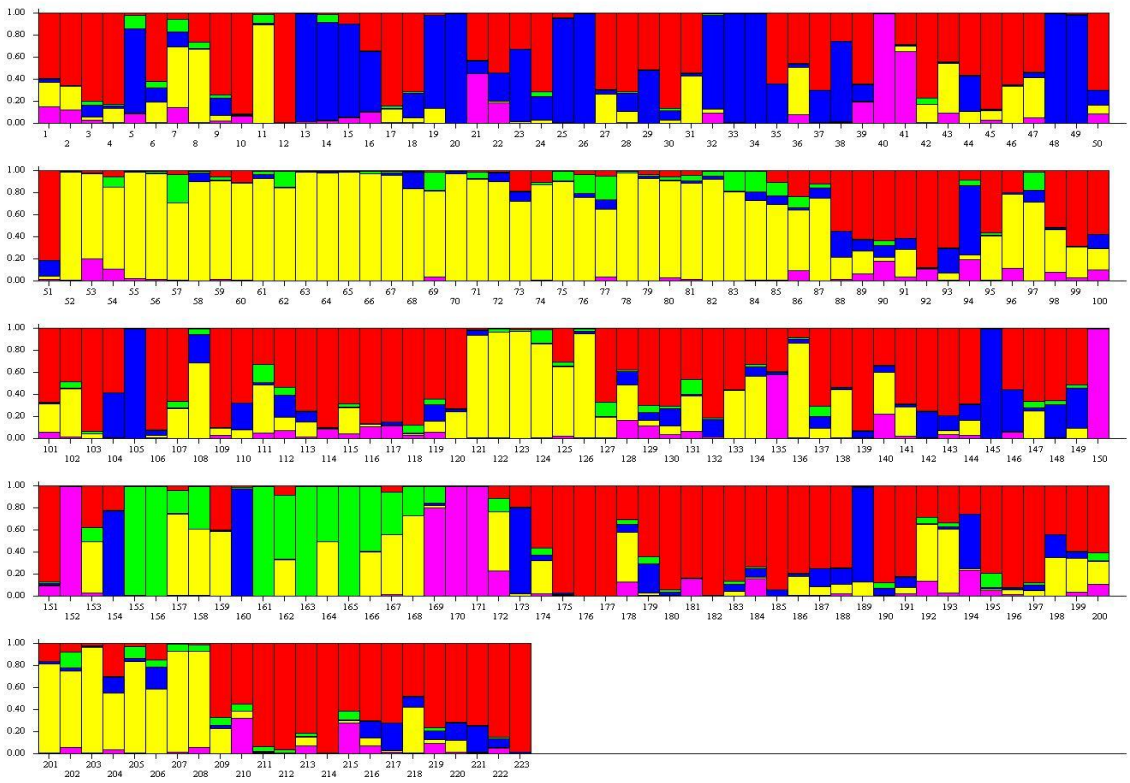
K=4, R=1



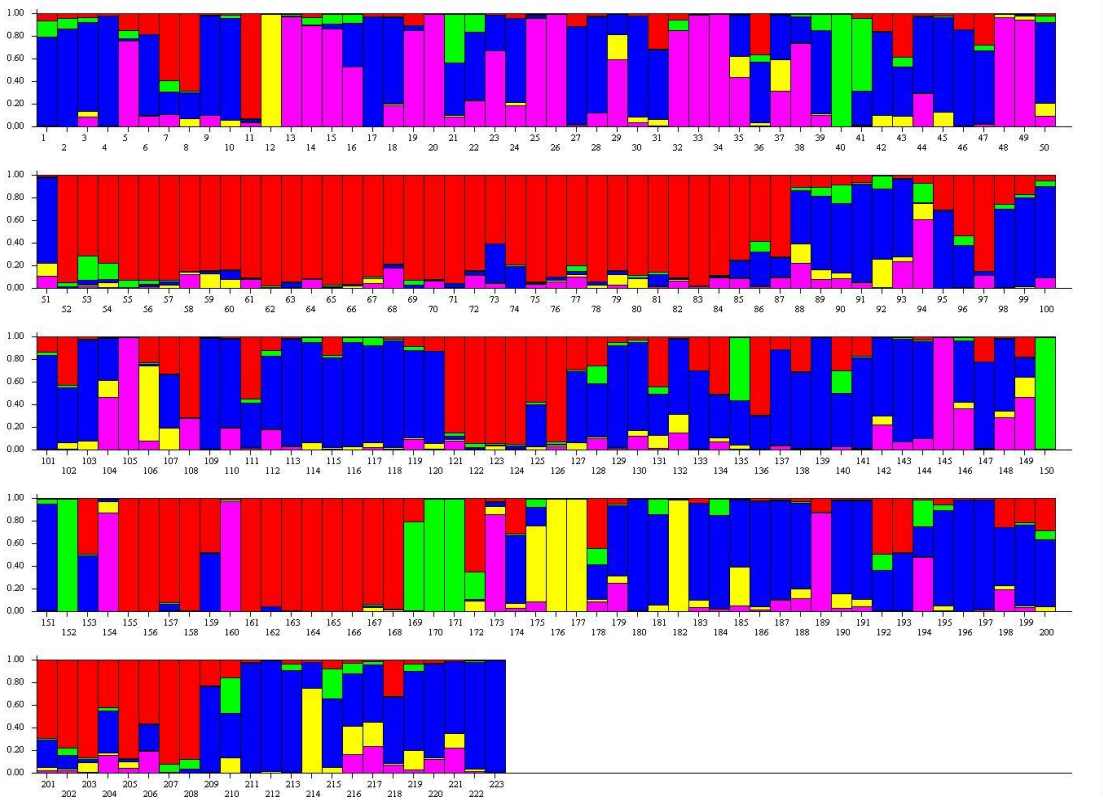
K=4, R=2



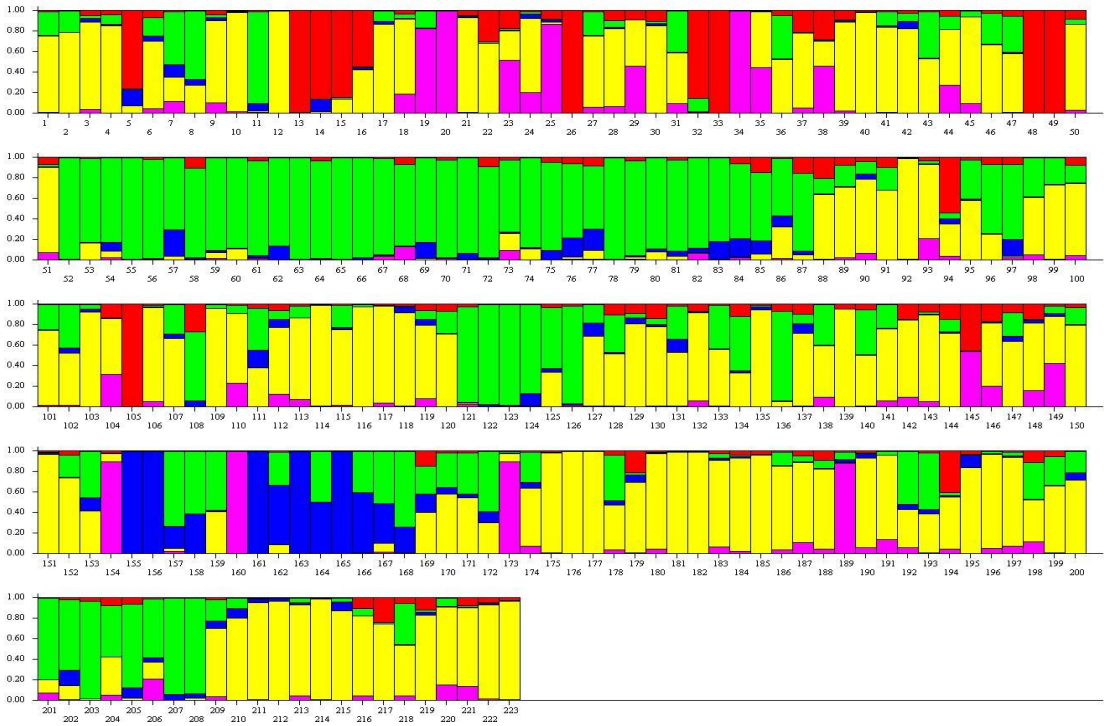
K=4, R=3



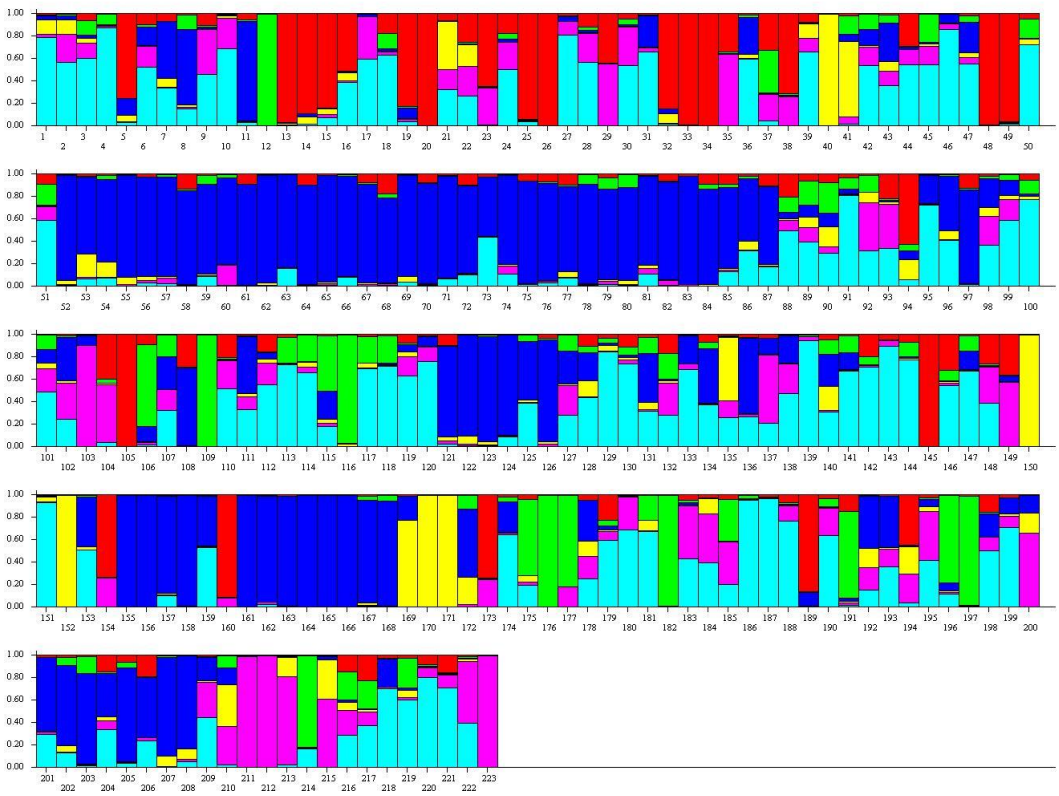
K=5, R=1



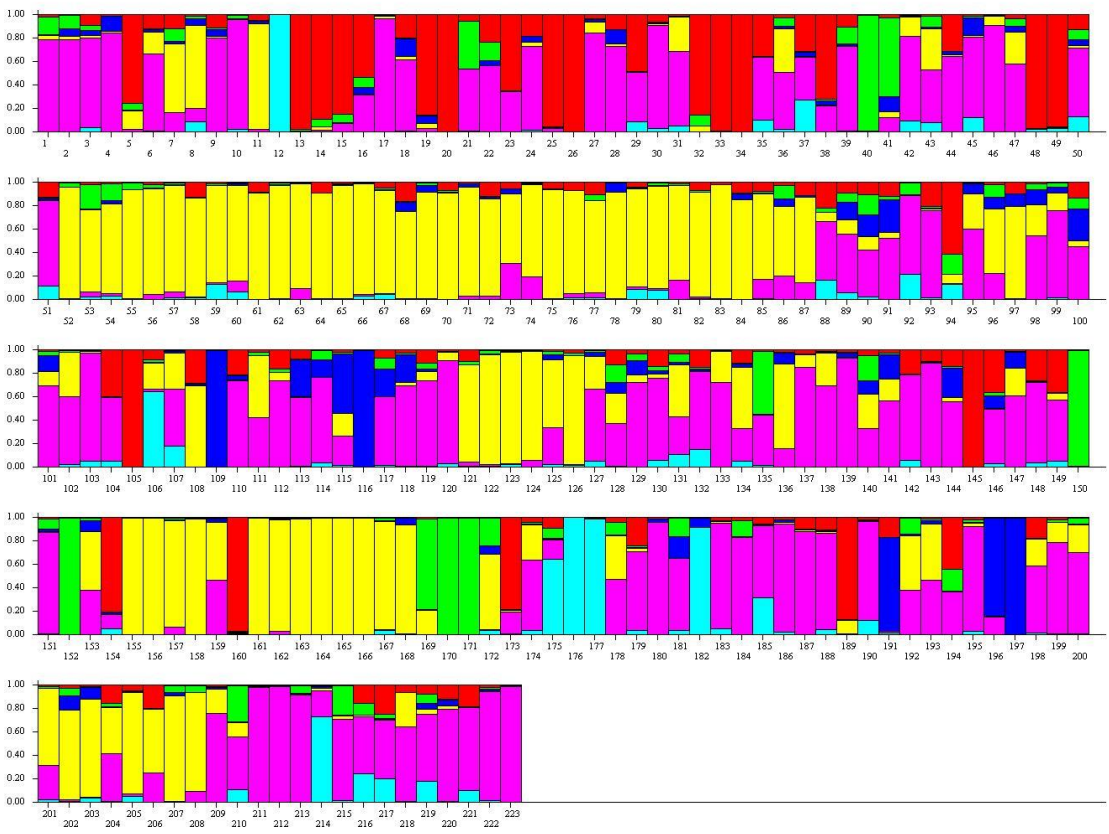
K=5, R=2



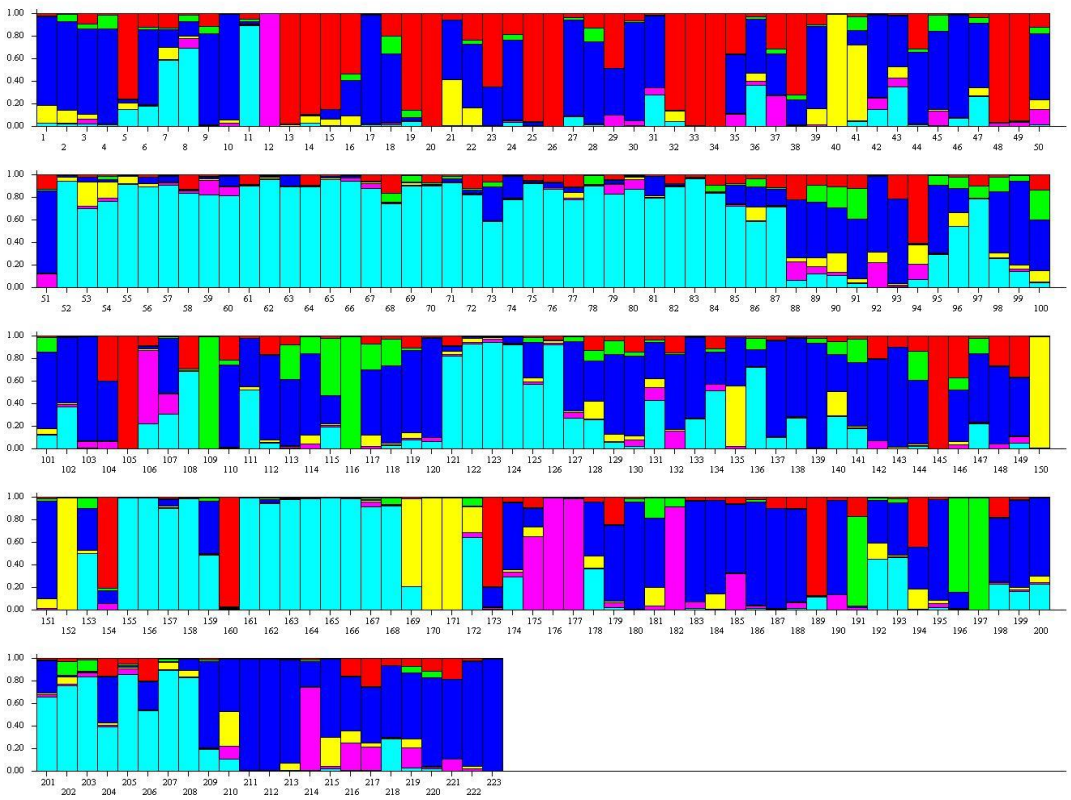
K=5, R=3



K=6, R=1



K=6, R=2



K=6, R=3

Anexo 4. Lista de líneas con sus códigos

Código	Línea
1	33-16
2	38-11
3	4226
4	A188
5	A214N
6	A239
7	A272
8	A441-5
9	A554
10	A556
11	A6
12	A619
13	A632

Código	Línea
113	I29
114	IA2132
115	IDS28
116	IDS69
117	IL14H
118	IL677A
119	K55
120	K64
121	KI14
122	KY21
123	KY226
124	KY228
125	L317

14	A634	126	L578
15	A635	127	M14
16	A641	128	M162W
17	A654	129	M37W
18	A659	130	MO17
19	A679	131	MO18W
20	A680	132	MO1W
21	A682	133	MO24W
22	B103	134	MO44
23	B104	135	MP339
24	B105	136	MS1334
25	B109	137	MS153
26	B14A	138	MS71
27	B164	139	MT42
28	B2	140	N192
29	B37	141	N28Ht
30	B52	142	N6
31	B57	143	N7A
32	B64	144	NC250
33	B68	145	NC258
34	B73	146	NC260
35	B76	147	NC262
36	B77	148	NC264
37	B79	149	NC294
38	B84	150	NC296
39	B97	151	NC298
40	C103	152	NC300
41	C123	153	NC304
42	C49A	154	NC320
43	CI187-2	155	NC328
44	CI21E	156	NC336
45	CI31A	157	NC338
46	CI64	158	NC348
47	CI66	159	NC350
48	CM105	160	NC352
49	CM174	161	NC354
50	CM37	162	NC356
51	CM7	163	NC358
52	CML10	164	NC360
53	CML103	165	NC362
54	CML108	166	NC364
55	CML11	167	NC366
56	CML14	168	NC368
57	CML218	169	ND246
58	CML220	170	OH40B
59	CML228	171	OH43
60	CML238	172	OH43E
61	CML247	173	OH603
62	CML254	174	OH7B
63	CML258	175	OS420
64	CML261	176	P39
65	CML264	177	PA762
66	CML277	178	PA875
67	CML281	179	PA880
68	CML287	180	PA91
69	CML311	181	R109B
70	CML314	182	R168
71	CML321	183	R177
72	CML322	184	R229
73	CML323	185	R4

74	CML328	186	SA24
75	CML331	187	SC213R
76	CML332	188	SC357
77	CML333	189	SD40
78	CML341	190	SD44
79	CML38	191	SG1533
80	CML45	192	SG18
81	CML5	193	T232
82	CML52	194	T234
83	CML61	195	T8
84	CML69	196	TX303
85	CML77	197	TX601
86	CML91	198	TZI10
87	CML92	199	TZI16
88	CMV3	200	TZI18
89	CO106	201	TZI25
90	CO125	202	TZI8
91	CO255	203	TZI9
92	DE2	204	U267Y
93	DE3	205	VA102
94	DE811	206	VA14
95	E2558W	207	VA17
96	EP1	208	VA22
97	F2834T	209	VA26
98	F44	210	VA35
99	F6	211	VA85
100	F7	212	VA99
101	GA209	213	W117HT
102	GT112	214	W153R
103	H49	215	W182B
104	H84	216	W22
105	H91	217	W64A
106	H95	218	WF9
107	H99	219	KI11
108	HI27	220	KI2021
109	HP301	221	KI3
110	HY	222	KI43
111	I137TN	223	Ki44
112	I205		

BIBLIOGRAFÍA

Beló A., Zheng P., Luck S., *et al*. 2008. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10.

Bradbury P.J., Zhang Z., Kroon D.E., *et al* 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Buckler E.S., Holland J.B., Bradbury P.J., *et al*. 2009. The genetic architecture of maize flowering time. Science 325: 714-718.

Camus-Kulandaivelu L., Chevin L.M., Tollon-Cordet C.M., *et al* 2008. Patterns of Molecular Evolution Associated With Two Selective Sweeps in the Tb1 Dwarf8 Region in Maize. Genetics 180(0):1107-1121.

Cassman K.G., Dobermann A., Walters, D.T., *et al* 2003. Meeting cereal demand while protecting natural resources and improving environmental quality. Annu. Rev. Environ. Resour. 2003. 28:315–58.

Coles N.D., McMullen M.D., Balint-Kurti P.J., *et al* 2010. Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. Genetics 184: 799–822.

Cook JP., McMullen MD., Holland JB., *et al.* 2012. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiology 158: 824-834.

Darvasi A., Weinreb, A., Minke, V., *et al.* 1993. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics, July 1993, vol. 134, no. 3, p. 943-951.

Ducrocq S., Madur D., Veyrieras J B., *et al.* 2008. Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178: 2433–2437.

Ducrocq S., Giauffret, C., Madur, D., *et al.* 2009. A. Fine mapping and haplotype structure analysis of a major flowering time quantitative trait locus on maize chromosome 10 Genetics, 183 (4), pp. 1555-1563.

Earl D. A., y VonHoldt, B. M., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources vol. 4 (2) pp. 359-361.

Evanno, G., Regnaut, S., Goudet, J., Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol Resour **14,** (2005).

Fischer R.A., y Edmeades., G.O. 2010. Breeding and cereal yield progress. Crop Sci. 50:S-85–S-98.

Fillippi C V., Aguirre N., Rivas JG., *et al*. 2015. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC Plant Biol 15:52.

Flint-Garcia SA., Thuillet AC., Yu J., *et al*. 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064.

Goodman, M., 2004 Developing temperate inbreds using tropical germplasm: rationale, results, conclusions. Maydica 49: 209–220.

Gore MA., Chia J-M., Elshire RJ., *et al* 2009. A first-generation haplotype map of maize. Science 326: 1115-1117.

Guichoux F., Lagache L., Wagner S., *et al.* 2011. Current trends in microsatellite genotyping. Mol Ecol Resour.;11:591–611.

Hardy OJ., Vekemans X., 2002. SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620.

Hung H-Y., Shannon L.M., Tian F., *et al.* 2012. ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proceedings of the National Academy of Sciences of the United States of America, 109 (28), pp. E1913-E1921.

Korte A., Farlow A., 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29.

Kump KL., Bradbury PJ., Wisser RJ., *et al* 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nature Genetics 43: 163-168.

Liu K., Goodman MM., Muse S., *et al* 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165: 2117-2128.

Loiselle BA., Sork VL., Nason J., *et al.* 1995. Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). Am J Bot 82: 1420-1425.

Olmos SE., Lia V V., Eyhérabide GH., 2016. Genetic diversity and linkage disequilibrium in the Argentine public maize inbred line collection. Plant Genet Resour FirstView:1–12.

Oraguzie NC., Wilcox PL., 2007. An Overview of Association Mapping BT - Association Mapping in Plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds). Springer New York, New York, NY, pp 1–9

Peretti M.A., 2003. Maíz 2003/2004: Aspectos Económicos del Cultivo. Reunión Anual de Actualización Técnica de Maíz 2003. EEA INTA Marcos Juárez. 4 de Junio de 2003.

Poland JA., Bradbury PJ., Buckler ES., *et al.* 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. PNAS 108: 6893-6898.

Pritchard JK., Stephens M., Donnelly P., 2000. Inference of population structure using multilocus genotype data. Genetics, 155: 945-959.

Rafalski, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5,** 94–100.

Romay MC., Millard MJ., Glaubitz JC., *et al.* 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol.

Salvi S., Sponza G., Morgante M., *et al*. 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci 104:11376–11381.

Semagn K., Bjørnstad Å., Xu Y., The genetic dissection of quantitative traits in crops. Electron. J. Biotechnol. Vol 13, No 5 (2010).

SIIA., 2014. Sistema Integrado de Información Agropecuaria, Programa de Servicios Agrícolas Provinciales, Ministerio de Agricultura, Ganadería y Pesca. Acceso 05/06/2014. http://www.siia.gov.ar/_apps/siia/estimaciones/estima2.php

Steinhoff J., Liu W., Reif J.C., *et al.* 2012. Detection of QTL for flowering time in multiple families of elite maize. Theor Appl Genet 125: 1539–1551.

Stich B., Mohring J., Piepho H.P., *et al* 2008. Comparison of mixed-model approaches for association mapping. Genetics 178: 1745-1754.

Thornsberry J., Goodman M., Doebley J., *et al.* 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genet 28: 286-289.

Tian F., Bradbury P.J., Brown P.J*., et al.* 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics 43: 159-162.

Wang C., Chen Y., Ku L., Wang T., Sun Z., *et al.* 2012. Mapping QTL Associated with Photoperiod Sensitivity and Assessing the Importance of QTLxEnvironment Interaction for Flowering Time in Maize. PLoS ONE 5(11): e14068.

Xu J., Liu Y., Liu J., *et al* 2012. The genetic architecture of flowering time and photoperiod sensitivity in maize as revealed by QTL review and meta analysis. J Integr Plant Biol. 54(6):358-73.

Xu Y., Crouch JH., 2008. Marker-Assisted Selection in Plant Breeding: From Publications to Practice All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any in. Crop Sci 48:391–407.

Xu Y., Kassa Semagn K., Bjørnstad A., 2010. The genetic dissection of quantitative traits in crops. Electronic Journal of Biotechnology ISSN: 0717-3458.http://www.ejbiotechnology.info.

Yu J., Pressoir G., Briggs W.H., *et al* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.

Yu J.M., Holland J.B., McMullen M.D., *et al* 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539-55.

Zha K., Tung, C.W., Eizenga G.C., *et al* 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2: 467.

VILLALÓN José Gastón
ECANA
Location: UNNOBA
Date: 2018.02.22 09:13:21 -0300