UNIVERSIDAD NACIONAL DE CÓRDOBA

TÍTULO DE TESIS

EVALUACIÓN DE ALGORITMOS DE AGRUPAMIENTOS PARA INFERIR ESTRUCTURA GENÉTICA POBLACIONAL EN DATOS GENÓMICOS

PARA OPTAR POR EL GRADO DE

Maestría en Estadística Aplicada

AUTOR

Prof. María Eugenia Videla

DIRECTORA

Ing. Agr. (PhD) Cecilia Inés Bruno

Año 2021

Agradecimientos

Quiero agradecer especialmente a mi directora, Dra. Cecilia Bruno, por su apoyo, generosidad y paciencia para enseñarme. Gracias por estar siempre dispuesta a ayudarme. Gracias Ceci por todo el tiempo brindado para que este trabajo fuera posible.

A la Dra. Mónica Balzarini, quien ha sido generosamente servicial, por su confianza y orientación tanto en el desarrollo de este trabajo como en los cursos de la maestría.

A la Dra. Juliana Iglesias por su tiempo brindado para lectura de documentos preliminares de esta tesis y por sus valiosos comentarios.

A los profesores y compañeros de la Cátedra de Estadística y Biometría por estar siempre a disposición para ayudarme.

A la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba por brindarme un espacio de trabajo.

A la Universidad Nacional de Villa María por abrirme las puertas para desempeñar mi vocación docente.

A los Miembros del Comité Evaluador por aceptar gentilmente formar parte del tribunal examinador y por dedicar su tiempo a la revisión de este trabajo.

A todo el cuerpo de docentes y compañeros de la Maestría en Estadística Aplicada por los hermosos años compartidos.

A mi familia y amigos por estar siempre a mi lado y por su cariño. Gracias por estar en mi vida y llenarme de momentos inolvidables que me hacen una persona feliz.

A mis padres, hermano y pareja por ser pilares fundamentales en mi vida, por el incondicional amor y apoyo que me brindaron siempre. Gracias a ellos por cada día confiar y creer en mí. A ellos es a quién dedico este trabajo.

Resumen

La disponibilidad de herramientas basadas en biotecnologías para evaluar miles de variantes genómicas simultáneamente ha revolucionado el paradigma en los estudios de diversidad genética. La información provista por los marcadores moleculares (MM) proporciona datos de naturaleza multivariada que pueden ser utilizados para identificar similitudes/diferencias genéticas entre individuos. Dado un conjunto de individuos caracterizados molecularmente, se espera que aquellos que presentan mayor similitud en su perfil genético, se encuentren relacionados, en algún grado de parentesco y por lo tanto, puedan agruparse definiendo poblaciones o grupos genéticos. Una plétora de métodos multivariados, para identificar grupos de individuos, ha sido propuesta para abordar la clasificación en un volumen masivo de MM, entre ellos, el análisis de conglomerados. A pesar de la existencia de diferentes algoritmos de clasificación, la cantidad de grupos sugeridos puede ser difusa. Dado a que los algoritmos definen grupos que no son conocidos *a priori*, independientemente del método de agrupamiento, la partición final de los datos requiere alguna clase de evaluación para encontrar el número óptimo de grupos que resulta ser la mejor partición natural de los datos. El objetivo del presente trabajo de tesis es evaluar el desempeño de distintos métodos de agrupamiento e índices de validación del número de grupo para detectar las correlaciones genéticas existentes entre individuos bajo distintos escenarios de estructura genética poblacional. Este trabajo de tesis ha sido organizado con una introducción general en el contexto de la descripción de los datos genómicos y el concepto de estructuración de lo mismos en el Capítulo 1. En el Capítulo 2 se compara el comportamiento de tres métodos de agrupamiento provenientes de diferentes familias de algoritmos a través de un estudio de simulación. En el Capítulo 3 la identificación del número óptimo de grupos generados por algoritmos de agrupamientos fue evaluada a través de la comparación de cuatro índices de validación. Finalmente, se ilustran, en el Capítulo 4, los métodos comparados sobre dos conjuntos de datos de maíz generados a partir de ensayos en el marco de programas de mejoramiento genético vegetal. Finalmente, hemos dispuesto en un Anexo los códigos de programación en R.

Palabras clave: *Marcadores Moleculares, análisis multivariado, ordenamiento, clasificación, índices de validación de agrupamiento, SNPs*

Abstract

The availability of biotechnology-based tools to assess thousands of genomic variants simultaneously has revolutionized the paradigm in genetic diversity studies. The information provided by molecular markers (MM) brings data of a multivariate nature that can be used to identify genetic similarities/differences between individuals. Given a set of molecularly characterized individuals, it is expected that those with the greatest similarity in their genetic profile are related in some degree of relatedness and can therefore be grouped together to define populations or genetic groups. A plethora of multivariate methods to identify groups of individuals have been proposed to address classification in a massive volume of MM, among them, cluster analysis. Despite the existence of different classification algorithms, the number of suggested groups can be diffuse. Since the algorithms define groups that are not known *a priori*, regardless of the grouping method, the final partitioning of the data requires some kind of evaluation to find the optimal number of groups that results in the best natural partitioning of the data. The objective of this thesis is to assess the performance of different grouping methods and group number validation rates to detect existing genetic correlations between individuals under different scenarios of population genetic structure. This thesis has been organized with a general introduction in the context of the description of genomic data and the concept of genomic structuring in Chapter 1. In Chapter 2 the performance of three grouping methods from different families of algorithms is compared through a simulation study. In Chapter 3 the identification of the optimal number of groups generated by clusters algorithms was assessed through the comparison of four validation rates. Finally, Chapter 4 illustrates the methods compared on two maize data sets generated from trials within the framework of plant genetic improvement programs. Finally, we have provided the programming codes in R in an Annex.

Key words: *Molecular markers, multivariate analysis, ordering, classification, clustering validation index, SNPs*

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1

LOS MARCADORES MOLECULARES Y SU FUNCIÓN EN LA BÚSQUEDA DE ESTRUCTURA GENÉTICA POBLACIONAL

Introducción General

La disponibilidad de herramientas basadas en biotecnologías para evaluar miles de variantes genómicas simultáneamente ha revolucionado el paradigma en los estudios de diversidad genética (González-recio *et al.*, 2014). La diversidad genética se define como la variedad de alelos y genotipos presentes en una población que se expresan, conjuntamente con el ambiente, en el fenotipo (Frankham *et al.*, 2002). Los métodos más comunes para evaluar la diversidad genética son el análisis de los fenotipos (o descripción de genes visibles) y el análisis a partir de marcadores moleculares (Martínez *et al.*, 2015). La observación del fenotipo indica una inversión significativa de tiempo y costos (de los Campos *et al.*, 2009), por ello durante la primer década del sigo XXI los marcadores moleculares han jugado un papel creciente en los estudios de diversidad genética (Groeneveld *et al.*, 2010; Vignal *et al.*, 2002). La creciente disponibilidad de datos provenientes de diferentes tipos de marcadores moleculares permite una caracterización exhaustiva de la diversidad genética en diferentes especies (Becerra y Paredes, 2000). La información provista por los marcadores moleculares proporciona datos de naturaleza multivariada que pueden ser utilizados para identificar similitudes/diferencias genéticas entre individuos. Dado un conjunto de individuos caracterizados molecularmente, se espera que aquellos que presentan mayor similitud en su perfil genético, se encuentren relacionados, en algún grado de parentesco y por lo tanto, puedan agruparse definiendo subpoblaciones o grupos genéticos (Peña-Malavera *et al.*, 2014). El estudio de variabilidad genética orientado a determinar la estructura genética poblacional (EGP) permite identificar las variantes genéticas que permiten tanto la conservación de recursos genéticos, la comprensión de la dinámica poblacional, así como la incorporación de las correlaciones genéticas en estudios de selección y asociación entre el fenotipo y el genotipo de especies de interés agrícola (Odong *et al.*, 2011; Shriner *et al.*, 2007; Wang

et al., 2005). Haile *et al.* (2018) incorporaron información de estructura genética poblacional en seis modelos de selección genómica en trigo y tres enfoques de predicción, que se aplicaron para predecir el rendimiento, contenido de proteína, índice de gluten y las medidas del alveógrafo. Thorwarth *et al.* (2017) evaluaron el potencial de la predicción genómica e investigaron la influencia de la estructura de la población, en la capacidad de predicción, en una población de entrenamiento de 750 genotipos de múltiples familias de materiales mejorados de cebada de invierno de seis hileras (Hordeum vulgare L.) y cultivares antiguos, que reflejan la historia de reproducción de la cebada en Alemania. Yuan *et al.* (2020) realizaron un estudio de asociación de todo el genoma (GWAS) de la tolerancia a la sal para el mejoramiento de arroz utilizando diferentes modelos de análisis que tuvieron en cuenta la estructura poblacional. Particularmente, en cultivos de importancia agrícola mundial como el maíz (*Zea mays* L.), conocer la relación entre líneas podría ayudar a identificar un conjunto de individuos que tengan la máxima diversidad para luego poder analizar los efectos del origen genético (Liu *et al.*, 2006) y de esta manera poder diseñar estrategias de cruzamientos para lograr líneas promisorias tanto en producción como en sanidad. Los marcadores moleculares han sido utilizados para estimar relaciones parentales entre líneas diversas (Yan *et al.*, 2007). En la actualidad, los marcadores moleculares de tipo SNPs (*Single Nucleotide Polymorphism*), hacen posible el análisis detallado de la diversidad genética existente en las poblaciones vegetales, incluyendo tanto especies domesticadas como silvestres, lo cual es un paso esencial para el descubrimiento de nuevos genes de interés con el fin de introducirlos en futuros planes de mejora (García Pérez, 2020). Estos se basan en la detección de polimorfismos resultantes de una variación en la secuencia del ADN que afecta a una sola base nucleotídica (adenina (A), timina (T), citosina (C) o guanina (G)) de una secuencia del genoma (Brookes, 1999; Vignal *et al.*, 2002). La ventaja de este tipo de marcadores moleculares radica en que tales polimorfismos se encuentran distribuidos por todo el genoma y pueden localizarse tanto en regiones codificantes como no codificantes (Sachidanandam *et al.*, 2001). Los SNPs poseen baja tasa de mutación (Kondrashov, 2003; Nachman y Crowell, 2000) y bajas tasas de error en el genotipado (Kennedy *et al.*, 2003) respecto a otros tipos de marcadores moleculares como los microsatélites (SSR) (Martínez *et al.*, 2015). La búsqueda de EGP en una gran colección de datos, conformadas no solo por miles de marcadores moleculares sino también por gran cantidad de individuos, implica un alto costo computacional lo cual incrementa la

complejidad en el manejo de bases de datos masivas como las generadas con los marcadores moleculares de tipo SNPs (Aguilar, 2011).

Una plétora de métodos multivariados, para identificar grupos de individuos, ha sido propuesta para abordar la clasificación en un volumen masivo de datos generados por las aplicaciones modernas de la biotecnología. Por ejemplo, el análisis de conglomerados jerárquicos, es un método de agrupamiento frecuentemente utilizado dado que se encuentra disponibles en gran cantidad de *software* y puede ser aplicado directamente sobre datos moleculares seleccionando una métrica de distancia apropiada sin necesidad de conocer *a priori* la existencia de grupos de individuos (Bruno y Balzarini, 2010; Odong *et al.*, 2011; Raj *et al.*, 2014). Adicionalmente, los algoritmos no jerárquicos, como *k-means* o *k-medoids*, también son frecuentemente utilizados para detectar EGP (Lee *et al.*, 2009), sin embargo, es necesario indicar a este algoritmo un número de k-grupos. Pritchard *et al.* (2000) propusieron un método de agrupamiento Bayesiano basado en cadenas de Marcov que se encuentra implementado en el *software* Sructure y en el paquete LEA del *software* R (Frichot y François, 2015), este método basa sus agrupamientos en la estimación de la probabilidad de pertenencia a un grupo u otro. Otros métodos de clasificación han sido propuestos basados en las redes neuronales y en el aprendizaje automático como máquinas de soporte vectorial (Nikolic *et al.*, 2009). A pesar de la existencia de diferentes algoritmos de clasificación, la cantidad de grupos sugeridos puede ser difusa. Dado a que los algoritmos definen grupos que no son conocidos *a priori*, independientemente del método de agrupamiento, la partición final de los ratos requiere alguna clase de evaluación (Rezaee *et al.*, 1998). El procedimiento que evalúa el resultado del agrupamiento es conocido como validación del agrupamiento y tiene como finalidad encontrar el número óptimo de grupos que resulta ser la mejor partición natural de los datos sin ninguna clase de información respecto a la estructura subyacente de los datos (Rendón y Abundez, 2016). Numerosos índices han sido propuestos combinando información acerca de la compactación intra-grupo y el aislamiento inter-grupos, así como otros factores que se encuentran relacionados a la geometría y propiedades estadísticas de los datos, el número de observaciones y la medida de similaridad/disimililaridad. Milligan y Cooper (1985) propusieron diversos índices de validación de un agrupamiento para datos del tipo binario. Dunn (1974) introdujo un índice de validación basado en la distancia entre grupos y el diámetro del conglomerado y Rousseuw y Kaujman (Rousseau, 1987) propusieron el estadístico de silueta que indica

el grado de confianza en la asignación de un objeto a un grupo. Tibshirani *et al.* (2001) propusieron el estadístico gap para estimar el número de grupos en un conjunto de datos tanto para algoritmos jerárquicos como para algoritmos no jerárquicos. Este estadístico compara el cambio en la dispersión dentro del grupo respecto a la dispersión esperada bajo una distribución nula. Lebart (2000) propuso un criterio basado en la primera y segunda derivada, haciendo referencia a la media y la varianza, respectivamente. Halkidi *et al.* (2000) propusieron un índice, denominado índice SD, basado en el concepto de dispersión promedio del agrupamiento y separación entre grupos mientras que Halkidi *et al.* (2001) propusieron otro denominado SDbw el cual tiene en cuenta la compactación y la separación entre agrupamientos.

Debido a que en la mayoría de los conjuntos de datos genómicos no siempre se conoce el grado de parentesco o de relaciones genéticas subyacentes entre los individuos, ya sea porque no se cuenta con la información de los parentales, los métodos de agrupamiento han sido aplicados para detectar estas correlaciones. El interés principal de poder detectar la existencia de correlaciones genéticas entre individuos es para determinar la presencia de EGP. En estudios de asociación fenotipo-genotipo donde se ajustan modelos asociación GWAS (del inglés *Genome Wide Association*), incluir información de EGP es fundamental para reducir la tasa de falsos positivos (Malosetti *et al.*, 2007; Peña-Malavera *et al.*, 2014). En el contexto de la selección genómica, por ejemplo, la estructura de la población es un factor clave que afecta las predicciones de los valores genéticos por lo que no tenerla en cuenta podría conducir a evaluaciones poco realistas de la precisión (Riedelsheimer *et al.*, 2013; Windhausen *et al.*, 2012) y la selección preferencial de individuos dentro de una sola subpoblación, lo que resultaría en una pérdida de diversidad en el programa de mejoramiento (Isidro *et al.*, 2015). El objetivo del presente trabajo de tesis es evaluar el desempeño de distintos métodos de agrupamiento e índices de validación del número de grupo para detectar las correlaciones genéticas existentes entre individuos bajo distintos escenarios de estructura genética poblacional.

Este trabajo de tesis ha sido organizado con una introducción general en el contexto de descripción de los datos genómicos y el concepto de estructuración de lo mismos en el Capítulo 1. En el Capítulo 2 se evalúan tres métodos de agrupamiento seleccionados por sus diferencias en cuanto al algoritmo que cada uno propone. Esta evaluación fue realizada con bases de datos simulados bajo distintas configuraciones recreando posibles

escenarios naturales de poblaciones de maíz. En el Capítulo 3 abordamos el problema de la identificación del número óptimo de grupos generados por los diferentes algoritmos de agrupamientos a través de la comparación de cuatro índices de validación del número de grupo. Finalmente, en el Capítulo 4 se ilustran sobre dos conjuntos de datos de maíz generados a partir de ensayos en el marco de programas de mejoramiento genético vegetal, los algoritmos evaluados en el capítulo 2 y los índices de validación del número de agrupamientos comparados en el Capítulo 3. En un apartado denominado anexo, el lector podrá disponer de los códigos de programación en el paquete R que fueron escritos para la simulación, implementación de los algoritmos de agrupamiento y validación del número de grupos a través de los índices. Estos *script* o códigos también se encuentran disponibles en https://github.com/EugeniaVidela/Estructura-Genetica-Poblacional.

Objetivo General

Evaluar algoritmos e índices para inferir estructura genética poblacional en grandes bases de datos generadas a partir de marcadores moleculares del tipo SNP.

Objetivos Específicos

1) Comparar algoritmos de agrupamiento jerárquico, de partición y bayesiano para la búsqueda de estructura genética poblacional en bases de datos masivas.
2) Evaluar índices de validación del número de grupos subyacentes detectados por los algoritmos de agrupamiento.

Generación del dato genómico

Marcadores moleculares del tipo ADN

El Ácido Desoxirribo Nucleico (ADN) es una molécula helicoidal formada por dos hebras paralelas poliméricas, donde cada unidad (mero) es un nucleótido. Cada nucleótido a su vez está formado por azúcar desoxirribosa, un grupo de fosfato y una de las cuatro bases nitrogenadas: Adenina (A), Citosina (C), Guanina (G) y Timina (T). La molécula de ADN se mantiene unida gracias a la complementariedad que existe entre sus bases, ya que la secuencia de bases de una cadena está relacionada en un estricto orden de apareamiento con la secuencia de bases de la cadena enfrentada. Así, T se relacionará con A y G se relacionará con C, de manera que la información genética contenida en una hebra está replicada en la complementaria (Watson y Francis, 1953; Merino, 2018).

Todas las posibles combinaciones de la secuencia que codifican para un gen se denominan alelos y cada par de alelos se ubica en un locus, es decir, en el mismo lugar del cromosoma. Estas diferencias alélicas crean la base para la expresión fenotípica de los organismos con sus diferencias y similitudes. Son varios los factores que pueden alterar el orden de una secuencia, desde la misma recombinación genética durante el proceso de fecundación, pasando por factores físico-químicos internos y externos, hasta eventos espontáneos. Las alteraciones en la secuencia pueden eliminar, intercambiar o adicionar nucleótidos (incluso segmentos completos) a la secuencia. El resultado final de estos procesos puede, o no, alterar la expresión de un gen que podría tener repercusiones a nivel biológico en el individuo (Díaz Rodríguez, 2016).

Los marcadores moleculares se basan en la detección de polimorfismos resultantes de una variación en la secuencia de ADN que afecta a una sola base y pueden aplicarse en uno o muchos individuos de una misma población o de distintas poblaciones (Schlotterer, 2004). Los marcadores moleculares se utilizan con el fin de detectar y monitorear estas alteraciones que ocurren en un determinado locus o región genómica de un individuo como así también estimar relaciones entre diversos individuos (Yan *et al*., 2007). El estudio de las variaciones genéticas ha sido una herramienta fundamental en el ámbito del mejoramiento de cultivos, la medicina humana o en estudios evolutivos (Agarwat *et al*., 2008; Sidransky, 2002). Las relaciones basadas en marcadores se han utilizado en

programas de mejoramiento genético vegetal para estimar el coeficiente de parentesco (proporción esperada de genes en común entre dos individuos) y establecer grupos y patrones heteróticos para la reproducción híbrida (Reif *et al*., 2003; Xia *et al*., 2005); identificar la estructura compleja de la población y el parentesco relativo (información necesaria para los estudios de mapeo de asociación) (Yu *et al*., 2006). También para identificar subconjuntos centrales de líneas con la máxima diversidad de una colección más grande de líneas analizadas y reducir el número de líneas para estudio (Yan *et al*., 2009).

A lo largo de los años, el desarrollo de los marcadores moleculares ha evolucionado en aspectos tales como el costo económico del método de detección, el rendimiento del mismo y el nivel de reproducibilidad (Bernardo, 2008). Los primeros marcadores moleculares se basaron en la técnica de hibridación, proceso por el cual se combinan dos cadenas de ácidos nucleicos complementarias simples para formar una única molécula de doble cadena por apareamiento de sus bases. Este tipo de marcadores fueron ampliamente utilizados pese a su elevado consumo de tiempo y costo económico (Merino, 2018). El desarrollo de la tecnología de reacción en cadena de la polimerasa (PCR) favoreció al surgimiento de una segunda generación de marcadores, entre los que se destacan los SSR (Simple Sequence Repeat) también llamados microsatétlites (Agarwal *et al*., 2008). La enzima ADN polimerasa II es la responsable de la síntesis de una nueva cadena de ADN a partir de una cadena molde y es la enzima encargada de la replicación del ADN. Estos marcadores permitieron incrementar la reproduciblidad, posibilitando la automatización de la detección y la disminución en los tiempos de ejecución. Sin embargo, en la última década su uso se ha restringido por la aparición de los marcadores de tercera generación cuya detección es fácil de automatizar (Mammadov *et al*., 2012). Entre estos marcadores se destacan los marcadores de secuencia expresada o ESTs (siglas del inglés Expressed Sequence Tags) y los SNPs. Específicamente, los EST se han desarrollado para determinar variaciones en las regiones codificantes del genoma (RCG), es decir la porción de un gen que contiene los exones que codifican a proteínas, mientras que los SNPs pueden encontrarse tanto en las RCG, como en las no codificantes (Ekblom y Galindo, 2011).

Entre todas las variaciones alélicas del genoma, las que más se utilizan en el análisis genético de una especie son los SSR, las pequeñas inserciones o deleciones de segmentos

(InDels) y los polimorfismos de nucleótido único, más conocidos como SNPs por sus siglas del inglés Single Nucleotide Polimorphism (Mammadov *et al*., 2012). Los SSR son muy informativos a nivel de polimorfismo de ADN por ser multilocus y multialélicos. Además, en los estudios de escaneos suelen utilizarse entre 400 y 800 para cubrir el genoma de una determinada especie, mientras que para escanear el mismo genoma con marcadores del tipo SNPs son necesarios mayor cantidad de marcadores (Kruglyak 1997). Sin embargo, el papel de los microsatélites como marcadores para los estudios de todo el genoma está cambiando, principalmente debido a los avances técnicos que han permitido la creación de mapas genéticos en menos tiempo, con mayor precisión y automatizados. Estos avances tecnológicos han permitido que la obtención de un gran número de marcadores SNPs sea práctica y económica (Kwok, 2001; Kennedy *et al*., 2003). Los microarrays de SNPs de alta densidad ("*chips* SNPs") son un método rápido, preciso y eficiente para genotipar varios cientos de miles de polimorfismos en un gran número de individuos (Kim *et al.*, 2018). Por ejemplo, Atlija *et al*. (2013) usaron un barrido genómico con un *chip* de SNPs de 50K para la detección de QTL con influencia sobre la resistencia a nemátodos intestinales en el ganado ovino, Nsabiyera *et al.* (2020) utilizaron un genotipado de alta densidad de 90K SNPs para 181 líneas de trigo para mapear con precisión un locus determinado y desarrollar un marcador de diagnóstico para la reproducción, en el "Proyecto 1000 genomas" (2015) reconstruyeron los genomas de 2.504 individuos de 26 poblaciones caracterizando un amplio espectro de variación genética con 84,7 millones de SNPs.

Las razones expuestas anteriormente han provocado que los polimorfismos de un solo nucleótido (SNPs) están reemplazando rápidamente a los SSR para los estudios de asociación genética y muchos otros estudios de *pedigríe* en humanos (Abecasis y Wigginton, 2005). Existen colecciones muy grandes de marcadores SNPs (Sachidanandam *et al*., 2001) que se encuentran disponibles públicamente, incluidos algunos diseñados específicamente para estudios de asociación (Matise *et al*., 2003; Shaw *et al*., 2004). Además, los SNPs poseen baja tasa de mutación (Nachman y Crowell, 2000; Kondrashov, 2003) y bajas tasas de error en el genotipado (Kennedy *et al*., 2003) respecto a otros tipos de marcadores moleculares como los SSR. Estos avances han sido tan sustanciales que, a pesar de tener que reemplazar cada microsatélite con múltiples SNPs, es más rápido y más rentable realizar análisis de asociación del genoma completo con SNPs en lugar de marcadores de microsatélites (John *et al*., 2004; Middleton *et al*., 2004;

Schaid *et al*., 2004). Estas mejoras en la tecnología de secuenciación ha permitido el análisis directo de la variación genética a nivel de secuencia de ADN en muchos *loci* (Ching, 2002).

Además de encontrar variaciones genéticas entre genomas de distintos individuos, también se busca asociar dichas variaciones a caracteres visibles como enfermedades, en humanos, animales y plantas (Nichols *et al*., 2020; Verardo *et al*., 2017 ; Liu *et al*., 2020) como así también en otros caracteres de interés como la producción en cultivos agrícolas (Suresh et al., 2019). Para citar un ejemplo, los estudios de asociación en humanos ha permitido descubrir patrones que indican que la variación de ciertos genes está ligada a enfermedades como la degeneración muscular asociada a la edad y la diabetes (Díaz Rodríguez, 2016). Es por ello que este tipo de investigaciones se realizan principalmente a las asociaciones entre los polimorfismos de un solo nucleótido y rasgos o características observables (variables fenotípicas) como enfermedades, por ejemplo. Para proceder a realizar estos estudios, se debe contar con datos genéticos procedentes de varios individuos, de tal modo que a partir de las secuenciaciones de los genomas se puedan identificar genes ligados a enfermedades u otras características de interés. Los SNPs son esenciales para este tipo de investigaciones debido a que se puede probar estadísticamente la presencia de ellos en alguna secuencia del genoma siempre que aparece el mismo fenotipo o un valor determinado en el fenotipo. Con estas pruebas estadísticas que permiten establecer la presencia de un marcador siempre que se expresen ciertos valores fenotípicos, se puede establecer que el cambio presentado a nivel genético corresponde a un rasgo fenotípico significativo y que principalmente se asocia a enfermedades si ese es el carácter fenotípico estudiado.

Los densos ensayos de genotipos de SNPs ofrecen grandes oportunidades y desafíos, tanto desde un punto genético, por la capacidad de conocer todo el genoma o casi todo el genoma de un individuo en un corto plazo como para los estadísticos en el sentido de validar herramientas de análisis que permitan identificar las asociaciones entre la presencia de determinados marcadores y características fenotípicas. La cantidad y calidad de los ensayos pueden permitir métodos para la localización precisa de genes que codifican para enfermedades o características específicas como puede ser rendimiento en vegetales o producción en animales. Sin embargo, la cantidad de datos hace que, hasta las formas de análisis estadístico más complicadas, sean intratables a gran escala del

genoma y la densidad de los *loci* analizados requiere un modelado detallado de la estructura de los cromosomas (Thomas, 2010).

Codificación de Marcadores Moleculares

Estadísticamente los SNPs se consideran variables de tipo categóricas, esto se debe a que sus respuestas ocurren a partir de la combinación de las bases nitrogenadas que las componen. Las posibles categorías varían de acuerdo a cada marcador y el número de alelos que éste contenga. Generalmente, lo más común es encontrar SNPs para los cuales existen tres posibles combinaciones de las bases de datos. Por ejemplo, si el marcador consta de los alelos C y T, entonces las tres posibles combinaciones serán CC, CT y TT, es decir, todas las posibles combinaciones que puedan resultar a partir de los dos alelos. Para realizar el análisis de este tipo de variables es necesario realizar una recodificación a partir de cálculo de frecuencia del alelo menor o del alelo mayor, para estimar las frecuencias alélicas relativas en cada marcador. El alelo heterocigota menor se codifica con el valor 2, el alelo homocigota con el valor 1 y el alelo heterocigota mayor con el valor 0. Las frecuencias alélicas para C y T para un determinado SNPs están dadas por:

$$f_C = \frac{2f_{CC} + f_{CT}}{2} \qquad \text{[Eq. 1]}$$

$$f_T = \frac{2f_{TT} + f_{CT}}{2} \qquad \text{[Eq. 2]}$$

donde f_{CC}, f_{CT} y f_{TT} es la frecuencia del genotipo CC, CT y TT, respectivamente. Si $f_C > f_T$ entonce T es el alelo menor del SNPs, de modo que la codificación corresponderá a TT=2, CT=1 y CC=0. En caso contrario, es decir cuando ocurre que el alelo menor es C, entonces la codificación será CC=2, CT=1 y TT=0 (Díaz Rodríguez, 2016).

Al comparar dos perfiles moleculares individuales, para cada posición existen cuatro eventos disjuntos posibles: 1) en los dos perfiles se observan alelos heterocigotas, indistintamente sean alelo mayor o menor, evento denotado como (0,0); (2,2); (0,2) ó (2,0); 2) los dos perfiles tienen alelos homocigotas, evento denotado como (1,1); 3) el primer perfil presenta alelo heterocigota y el segundo homocigota, evento denotado como

(0,1) ó (0,2); 4) el primer perfil presenta alelo homocigota y el segundo heterocigota, denotado como evento (1,0) ó (1,2). La frecuencia con que ocurre cada uno de estos eventos cuando se comparan dos individuos a partir de su perfil molecular se denominarán *a*, *b*, *c*, y *d* según correspondan a los eventos 1), 2), 3) y 4) respectivamente. Las frecuencia de estos polimorfismos representados por los eventos contiene toda la información relevante para la construcción de índices de similitud entre perfiles individuales y a partir de ella pueden calcularse múltiples índices de similitud entre perfiles individuales. Por ejemplo, el índice de similitud de Jaccard permite expresar distancias entre los perfiles de marcadores moleculares (dominantes) de dos individuos a partir de la siguiente fórmula:

$$S_{ij} = \frac{a}{a+b+c} \qquad \text{[Eq. 3]}$$

Luego la distancia entre los individuos i,j puede ser expresada como distancia con alguna función de transformación, como por ejemplo el complemento a uno de la similitud

$d_{ij} = 1 - S_{ij}$.

Estructura Genética Poblacional

Población estructurada hace referencia a la existencia de diversas clases de individuos, que corresponden a subpoblaciones producidas a partir de variables que modifican a la población. Dentro de tales variables, el origen de los individuos a partir de los fundadores de la población produce relaciones genealógicas que introducen diferencias entre subpoblaciones, por medio de cambios en las frecuencias génicas y genotípicas. La estructura genética de una población se caracteriza por las frecuencias génicas y genotípicas. Los individuos pueden entonces agruparse a partir de la información alélica.

Métodos de agrupamiento usados para Estructura Genética Poblacional

El análisis de conglomerados ha demostrado ser una herramienta eficiente para identificar el agrupamiento subyacente de los genotipos de una población (Hartigan, 1975; Gordon, 1999). Varios algoritmos han sido desarrollados para clasificar genotipos dentro de

subpoblaciones usando datos genéticos provenientes del genotipado molecular (Odong *et al.,* 2011; Lee *et al., 2*009; Lawson y Falush 2012). Los algoritmos de conglomerado no supervisado son aquellos que no demandan información previa sobre los grupos en los que se espera que los individuos se clasifiquen. Por el contrario, los algoritmos clasificados como supervisados asumen que las muestras tienen una asignación previa a un grupo (Christoph *et al*., 2005). La mayoría de estos métodos de conglomeración se usan en complementación con cálculos estadísticos desarrollados para estimar la cantidad de conglomerados que subyacen en la estructura de los datos poblacionales y que son de interés identificar (Balzarini *et al.,* 2008). Otra clasificación de los algoritmos de agrupamiento son los métodos jerárquicos y no jerárquicos. Los primeros son denominados así porque una vez que dos observaciones son agrupadas juntas, en los pasos sucesivos del algoritmo, éstas no se vuelven a separar. Los conglomerados jerárquicos, comúnmente aplicados para identificar EGP, pueden ser aplicados directamente sobre los datos moleculares (Odong *et al., 2*011) o luego de realizar un análisis de componentes principales (ACP) sobre la información molecular para contemplar la correlación que puede existir entre marcadores (Patterson *et al., 2*006) dado que el ACP conforma nuevas variables sintéticas ortogonales entre sí.

Tanto los conglomerados jerárquicos como los no-jerárquicos requieren dos pasos para realizar el agrupamiento. En el primero construyen una matriz de distancias, para lo cual es necesario seleccionar la métrica de distancia a usar que mejor refleje la naturaleza de los datos (Bruno, 2009) y luego debe seleccionarse el método de agrupamiento en el caso de los conglomerados jerárquicos, mientras que en los conglomerados no-jerárquicos primero separa un grupo de individuos en tantos grupos como k-grupos se haya seleccionado, es decir que debe indicarse una cantidad de grupos *a piori*. Esta separación en función del número de k-grupos indicados maximiza la variación entre conglomerados y minimiza la variabilidad dentro de cada conglomerado manteniendo el principio del análisis de conglomerados de agrupar individuos de manera tal de maximizar la similitud entre individuos de un mismo grupo minimizando la varianza intra-grupos. A partir del agrupamiento inicial se estima el centroide de cada grupo y se asigna cada individuo al grupo cuya distancia al centroide sea la mínima. En ambos métodos de agrupamiento las matrices de distancia entre pares de individuos y conglomerados que se van recalculando en cada paso de la conglomeración hasta que no haya más individuos que agregar a algún grupo (Balzarini *et al.*, 2008). Estos métodos basados en distancias pueden usar diferentes

métricas de similitud multivariada entre pares de genotipos como así también diversas funciones de transformación de la similitud a distancia. La similitud entre un par de genotipos puede depender no solo de la constitución genética, sino también de la del resto de la muestra a través de frecuencias alélicas (Weir y Ott, 1997). El promedio de la distancia genética entre dos genotipos es una interpretación simple del grado de relacionamiento (McVean, 2009). La distancia euclídea al cuadrado entre perfiles de marcadores refleja la cantidad de alelos no idénticos por estado entre dos genotipos y constituye una métrica para el análisis de conglomerados con datos moleculares. Sin embargo, existen otras métricas de distancias basadas en índices de similitud que fueron propuestas específicamente en el contexto de datos binarios y por lo tanto son las más recomendadas en el caso de marcadores moleculares codificados como presencia/ausencia (Bruno y Balzarini, 2013). La selección de una medida de distancia para usar los métodos de conglomerados basados en distancia será también dependiente de la forma en que se codifique la información molecular (Bruno, 2009). Otro método de conglomerados usual en la búsqueda de EGP es el método probabilístico de clasificación bayesiana implementado en el *software* STRUCTURE (Pritchard *et al., 2*000). Dada su naturaleza de clasificación difusa, asigna probabilidades de ocurrencia a cada grupo en lugar de asignar un individuo a sólo un grupo, es apropiado para colecciones de germoplasma con alto nivel de correlación genética. Este método asume un modelo en el que hay k subpoblaciones, cada una de las cuales se caracteriza por un conjunto de frecuencias alélicas para cada *locus*. Los individuos son asignados probabilísticamente a una o más subpoblaciones de acuerdo al grado de semejanza genética que manifiesta con el resto de los individuos de cada subpoblación. Este método ha sido utilizado en numerosos trabajos para determinar la estructura de poblaciones en maíz (Liu *et al.,* 2003; Stich *et al.,* 2005; Vigouroux *et al.,* 2008). Por su parte, otro tipo de algoritmos de agrupamiento son los métodos de agrupamiento basados en particiones, como por ejemplo *k-means* (MacQueen, 1967), los que encuentran particiones del conjunto de datos sin ninguna jerarquía obteniendo k grupos no unidos que optimizan una función objetivo, que usualmente es el error cuadrático. Podemos encontrar en la bibliografía una gran cantidad de algoritmos de partición (Kaufaman y Rousseeuw, 1999).

Es importante resaltar, que el desempeño relativo de diferentes métodos de agrupamiento puede ser función de distintas características específicas de la estructura genética poblacional subyacente como el número de sub-poblaciones (k) y la similitud genética o

divergencia genética entre grupos, la cual puede ser medida por el estadístico Fst de Wright (1951), que representa la correlación entre los genes de la subpoblación y los de la población total (Peña, 2015). Evanno *et al.,* (2005) llevaron a cabo un estudio de simulación para evaluar la habilidad de STRUCTURE para reconocer la estructura genética bajo diferentes escenarios biológicos derivados de distintos patrones de migración que consideraban distintos niveles de divergencia genética. Odong *et al., (*2011) evaluaron métodos de conglomerados jerárquicos bajo diferentes niveles de divergencia en colecciones de germoplasma vegetal usando los resultados de STRUCTURE como "*gold standard*". Lee *et al., (*2009) compararon, en un estudio de simulación, análisis de componentes principales (ACP), el algoritmo de STRUCTURE y técnicas de conglomerados no jerárquicos. En un trabajo de Milligan y Cooper (1985), se enumeran varios criterios para la comparación del desempeño de métodos de conglomeración, muchas de éstas usadas en los estudios de simulación que se mencionarán en los siguientes capítulos. La gran cantidad de trabajos biológicos que discuten las aproximaciones metodológicas sobre análisis de estructura genética, es una evidencia de la necesidad de investigaciones en estadística genética.

CAPÍTULO 2

COMPARACIÓN DE ALGORITMOS DE AGRUPAMIENTO PARA IDENTIFICAR ESTRUCTURA GENÉTICA POBLACIONAL

Introducción

El análisis de conglomerados o de agrupamiento es una técnica que tiene por objetivo agrupar elementos de manera tal de formar grupos lo más homogéneos posibles en función de las similitudes entre ellos para los múltiples caracteres, variables o atributos medidos en cada uno de los elementos. A partir de un conjunto de *n* datos no etiquetados, es decir, no asignados a ningún grupo en particular, el algoritmo de conglomeración realiza una clasificación en uno o más grupos conformados por elementos similares entre sí, con la condición de que cada elemento pertenezca a uno y solo un grupo y que todos los elementos pertenezcan a algún grupo. Para conformar estos grupos de elementos similares entre sí, la similitud entre los objetos es definida a través de alguna medida de distancia o una función objetivo. En términos generales, el algoritmo trata de maximizar la similitud entre los objetos de un mismo grupo y al mismo tiempo minimizar la similitud entre elementos de distintos grupos (Giraldo *et al*., 2013). Los métodos de agrupamientos se conocen también con el nombre de métodos de clasificación automática no supervisada o de reconocimientos de patrones sin supervisión porque, a diferencia del análisis de discriminante, el grupo al cual pertence un individuos no es conocido *a priori* (Peña, 2002). La técnica de agrupamiento ha sido ampliamente utilizada en la detección de anomalías (Díaz Muñoz *et al*., 2020), identificación de características sobresalientes de conjuntos de datos (Flores Reinoso, 2019), en el ordenamiento de individuos (Sahu *et al*., 2017), identificación de topologías (Piedrahita, 2018), entre otras finalidades y, en diferentes áreas del conocimiento como: biología, antropología, medicina, estadística, bioinformática y matemáticas entre otras. Desde principio de los años 50 del siglo XX se han propuesto y desarrollado una gran diversidad de técnicas multivariadas de agrupamiento y clasificación (Rendon y Abundez, 2016) como el análisis de *cluster* (Jain

y Dubes, 1998), el análisis discriminante lineal para clasificación supervisada (Fisher, 1936), las técnicas de aprendizaje automático y las redes neuronales (Kohonen, 1997).

Existen muchas clasificaciones para los algoritmos de agrupamiento, en este trabajo tomaremos la propuesta por Alvarez (2019) quien determina que los algoritmos de agrupamiento pueden clasificarse en tres tipos: jerárquicos, de partición o no-jerárquicos y probabilísticos (Alvarez, 2019). Los métodos de agrupamiento basados en jerarquía crean grupos recursivamente, permitiendo encontrar estructuras y subestructuras que se representan para su visualización en un diagrama denominado dendograma. El dendograma suele representar en uno de sus ejes la distancia a la cual se unen los grupos, distancias más pequeñas indican mayor similitud entre los individuos que conforman los grupos unidos, los puntos donde se unen los grupos se denominan nodos. Si la representación del agrupamiento en el dendograma comienza con tantos grupos como individuos participan del análisis (aglomerativos), el algoritmo finaliza cuando no haya más individuos que asignar a ningún grupo. A pesar de que el objetivo de un algoritmo de agrupamiento es como su nombre lo indica conformar grupos, es una técnica descriptiva, es decir, no indica de manera objetiva la cantidad de grupos que se formaron. Dicho de otra manera, dependiendo a qué altura (distancia) se corte el dendrograma será la cantidad de grupos obtenidos, haciendo que el número de grupos determinados sea una práctica subjetiva (Figura 1). Los tipos de métodos de agrupamiento jerárquicos son a su vez clasificados en dos tipos, aglomerativos que combinan grupos pequeños en grupos más grandes y divisivos que dividen grupos grandes en grupos más pequeños. La mayoría de los algoritmos de agrupamiento jerárquico son variantes de los algoritmos Single Linkage (encadenamiento simple) (Sneath y Sokal, 1973) y Complete Linkage (encadenamiento completo) (King, 1967). En el encadenamiento simple la distancia entre dos grupos se define como la distancia más corta o la mínima distancia entre dos objetos en cada grupo de modo que enlaza dos grupos que contienen el par de objetos más cercanos. Por otra parte, el algoritmo de encadenamiento completo o Complete Linkage define la distancia entre dos grupos como la distancia más grande (distancia máxima) entre dos objetos de cada grupo.

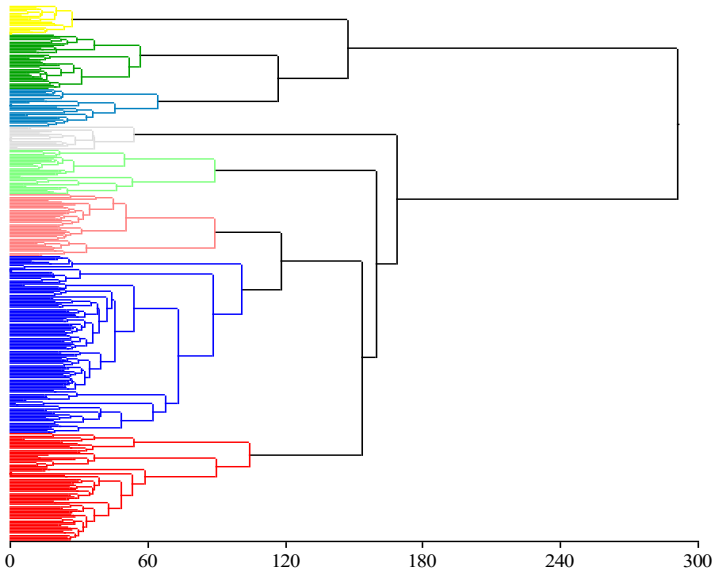

Figura 1: Ilustración de la visualización de un agrupamiento realizado por un algoritmo jerárquico en un diagrama denominado dendograma. El eje de las abscisas indica la distancia entre 344 genotipos agrupados mediante el método UPGMA con la distancia euclídea al cuadrado. En este ejemplo, los individuos (eje de las ordenadas) pintados con el mismo color, fueron agrupados juntos (Fuente: Peña-Malavera, 2015).

Los métodos de agrupamiento basados en particiones, encuentran particiones del conjunto de datos sin ninguna jerarquía, por eso también son denominados algoritmos no-jerárquicos que conforman k grupos diferentes que optimizan una función objetivo. La función objetivo más usada en algoritmos de este tipo es el error cuadrático. El error cuadrático se calcula como la sumatoria de las distancias entre cada individuo y su centroide más cercano. La función objetivo encuentra el óptimo cuando obtiene grupos más homogéneos (la menor varianza encontrada dentro grupos) y, en simultáneo, cuando maximiza la diferencia entre grupos. Podemos encontrar en la bibliografía una gran cantidad de algoritmos de partición (Kaufaman y Rousseeuw, 1999), pero uno de los más utilizados y referenciados es el algoritmo de agrupamiento *k-means* (MacQueen, 1967). Este método ofrece como resultado un vector de asignación en donde se indica a qué grupo fue asignado cada individuo pero no ofrece gráfico de visualización de los agrupamientos conformados. La mayoría de los algoritmos jerárquicos tienen un orden de complejidad (tiempo en el algoritmo puede resolver el problema) cuadrático $O(n^2)$,

por lo que tienen problemas cuando trabajan con grandes volúmenes de datos, mientras que los algoritmos de partición tienen una complejidad menor (lineal $O(n)$) (Jain *et al*., 1999).

Los algoritmos probabilísticos, por su parte, modelan agrupaciones a través de un modelo estadístico probabilístico. Asumen la posibilidad de mezcla de distribuciones, es decir un conjunto de k distribuciones que representan k grupos, por lo que cada distribución determina la probabilidad de que un objeto pertenezca a un determinado grupo. Así, cada elemento de un conjunto de datos tiene cierta probabilidad de pertenecer a un grupo o conjuntamente a dos o más conglomerados si el elemento indica una mezcla de patrones. Estos tipos de algoritmos son supervisados ya que el usuario debe indicar inicialmente el número de subpoblaciones que se asumen (k). Estos métodos permiten una visualización de la estructura genética poblacional subyacente mediante un gráfico de barras, en donde cada individuo en el conjunto de datos está representado por una línea vertical, que está dividida en k segmentos coloreados según la probabilidad de pertenencia estimada de ese individuo en cada uno de los k clústeres inferidos. Es un tipo de agrupamiento difuso donde las probabilidades *a posteriori* indican la certidumbre de la asignación del genotipo al clúster (Figura 2).

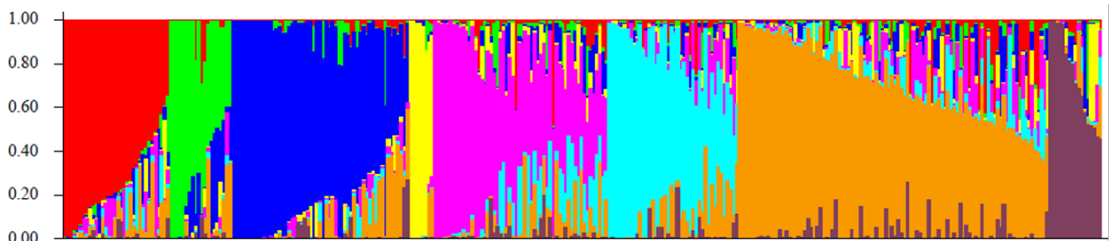

Figura 2: Gráfico de barras usualmente utilizado para visualizar el agrupamiento obtenido de un método probabilístico. Esta ilustración representa con cada línea vertical 344 individuos y con cada color la asignación de dicho individuo a una subpoblación, de manera que individuos del mismo color indican que fueron agrupados juntos. En el eje de las ordenadas se indica la probabilidad de pertenencia de un individuo al grupo asignado (Fuente: Peña-Malavera, 2015).

Numerosos y diferentes tipos de métodos de agrupamiento de cada tipo han sido propuestos en la literatura (Hedrick, 2005). A pesar de la gran diversidad, algunos métodos son más frecuentemente utilizados que otros, principalmente según el área en la que son aplicados y muchos de ellos no han sido evaluados en el contexto de búsqueda de estructura genética en bases de datos de alta dimensión como las generadas con marcadores moleculares del tipo SNPs. Muchos de ellos podrían proveer un desempeño similar en escenarios típicos generados por estructuras genéticas vegetales. En este trabajo de tesis se seleccionaron tres métodos de agrupamiento que además de provenir de diferentes familias de métodos, presentan diferencias en cuanto a sus algoritmos computacionales. Ellos son el método de agrupamiento jerárquico Unweighted Pair Group Method with Arithmetic Mean (UPGMA) propuesto por Sokal and Michener, 1958, el método de agrupamiento no-jerárquico *k-means* (MacQueen, 1967) y la aproximación por el método Bayesiano propueta por Pritchard *et al*. (2000) a través de su *software* STRUCTURE e implementada en este trabajo a través de la LEA del paquete R (Frichot y Francois, 2014).

Materiales y Métodos

Configuración de los parámetros genéticos para la generación de datos por simulación

Con el propósito de evaluar el desempeño de los tres métodos de agrupamiento se simularon datos de marcadores moleculares usando el paquete "Xbreed" de R (Esfandyari y Sørensen, 2017) e involucrando escenarios con n cantidad de genotipos y p cantidad de marcadores moleculares que imitan la estructura genética del cultivo de maíz. Las bases de datos fueron simuladas para marcadores moleculares del tipo SNPs considerando individuos diploides, utilizando una población histórica con 10 cromosomas configurando los siguientes parámetros genéticos para lograr un nivel deseado de desequilibrio de ligamiento (cierta cantidad de alelos o marcadores de ADN que debido a su cercanía física en un cromosoma se presentan juntos de manera más frecuente de lo que se esperaría por azar): número de individuos de la población inicial, número de marcadores moleculares, número de generaciones, tasa de mutación y heredabilidad en sentido estricto. El paquete Xbreed nos permitió simular la EGP según el comportamiento fenotípico de los individuos; se tomaron muestras de individuos de la población histórica como fundadores y, para distinguir las subpoblaciones, se simularon las generaciones

posteriores para cada subpoblación reciente indicando distintas varianzas fenotípicas y eligiendo a individuos según los extremos fenotípicos altos y bajos. Siguiendo a Latch et al. (2006), la EGP se determinó con tres niveles de diferenciación genética: $Fst = 0{,}03$, considerada genéticamente como un nivel bajo de diferenciación para individuos de la misma especie, $Fst = 0{,}05$ para representar un nivel medio de diferenciación genética y $Fst = 0{,}07$ que considera un nivel de diferenciación genética alto. La combinación de los tres niveles de diferenciación, el número de subpoblaciones ($k = 2$, $k = 3$ y $k = 10$) a identificar y la cantidad de individuos que conformaban la población ($n = 250$ y $n = 1000$), generaron 18 escenarios biológicos por simulación. Los niveles de diferenciación genética (Tabla 1). Cada escenario se replicó 100 veces y el número de marcadores moleculares de tipo SNPs simulados fue de $p = 80000$. El *script* utilizado para generar las simulaciones se presenta en el Anexo I.

Tabla 1. Configuración de cada escenario de simulación para un cultivo de maíz para tres niveles de diferenciación genética: Baja (Fst= 0.03), Media (Fst=0.05) y Alta (Fst=0.07); tres tamaños de subpoblaciones y dos números de individuos. Simulación en el paquete Xbreed de R.

Escenario de Simulación	Abreviatura	Nº de grupos (k)	Diferenciación Genética	Nº individuos
Escenario 1	E1	2	Baja	250
Escenario 2	E2	2	Baja	1000
Escenario 3	E3	2	Media	250
Escenario 4	E4	2	Media	1000
Escenario 5	E5	2	Alta	250
Escenario 6	E6	2	Alta	1000
Escenario 7	E7	5	Baja	250
Escenario 8	E8	5	Baja	1000
Escenario 9	E9	5	Media	250
Escenario 10	E10	5	Media	1000
Escenario 11	E11	5	Alta	250
Escenario 12	E12	5	Alta	1000
Escenario 13	E13	10	Baja	250
Escenario 14	E14	10	Baja	1000
Escenario 15	E15	10	Media	250
Escenario 16	E16	10	Media	1000

Escenario 17	E17	10	Alta	250
Escenario 18	E18	10	Alta	1000

Para visualizar la configuración de las subpoblaciones logradas en los escenarios simulación se representaron los individuos mediante un gráfico de dispersión obtenido a partir de las dos primeras coordenadas principales de un análisis de coordenadas principales, utilizando la distancia Jaccard, en donde se puede observar que se lograron distintos niveles de separación entre las subpoblaciones (Figura 3).

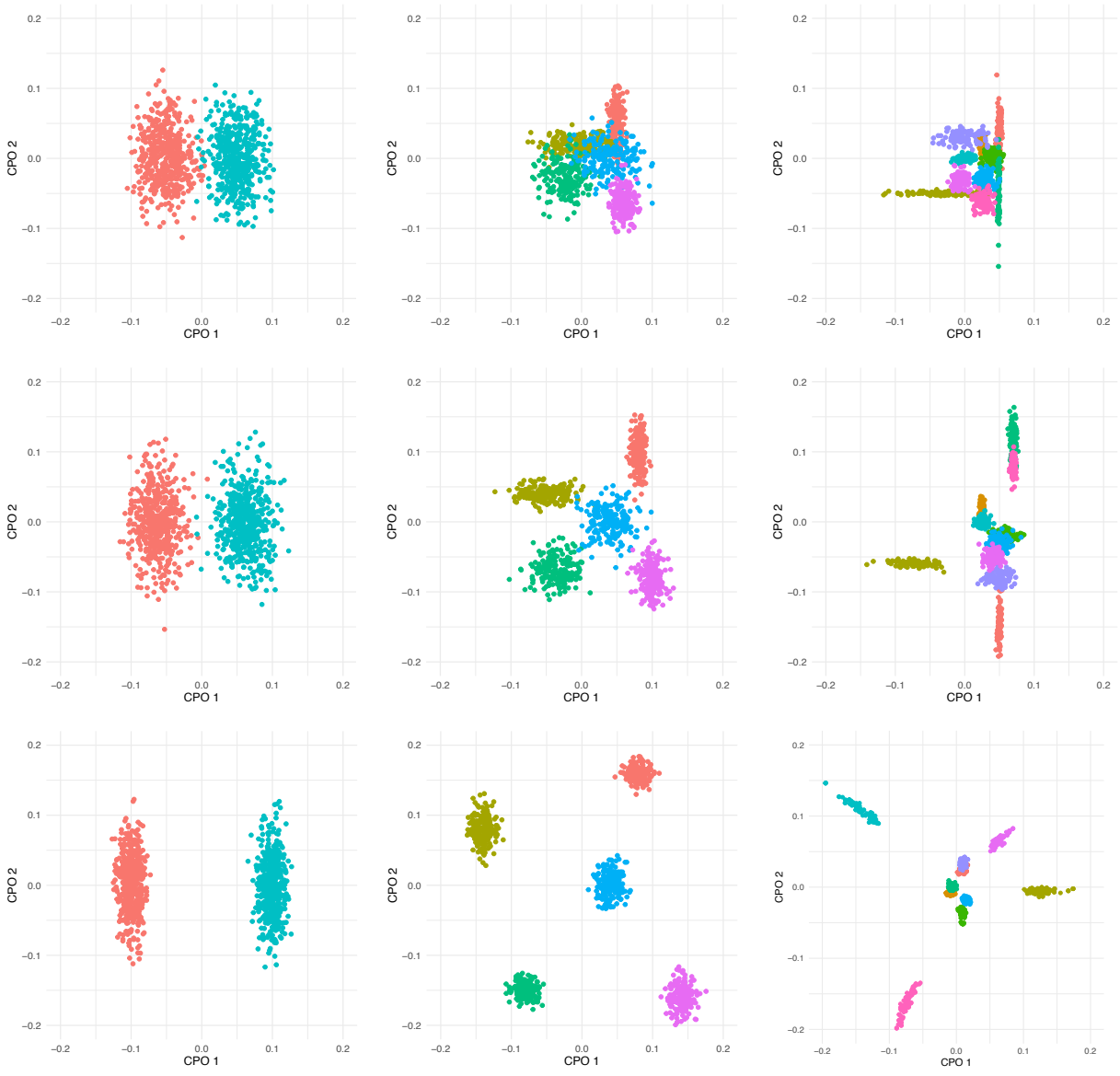

Figura 3: Gráfico de dispersión del análisis de coordenadas principales de una simulación de datos moleculares de 1000 individuos genotipados con 80K SNPs para nueve escenarios de simulación que difieren en el número de k grupos: k = 2 (izquierda), k = 5 (centro) y k = 10 (derecha); y en la diferenciación genética: baja (arriba), media (centro) y alta (abajo). Cada

individuo está representado por un punto. Los individuos que pertenecen al mismo grupo se representan con el mismo color.

La finalidad de simular escenarios biológicos con distintas configuraciones de estructura genética es evaluar el desempeño de los algoritmos de agrupamiento en distintas situaciones de divergencia genética. Además, poder evaluar sus capacidades cuando aumenta la cantidad de subpoblaciones a identificar dado que, en situaciones de baja divergencia, puede haber grupos no claramente separados. Por otra parte, el aumento de individuos puede hacer que aumente la diferenciación o variabilidad genética entre individuos del mismo grupo por la combinación al azar de la herencia de información genómica (Templeton, 2006).

Las bases de datos se codificaron según el alelo menor, es decir, se codifica al alelo homocigota más frecuente con 0, el alelo heterocigoto con 1 y el alelo homocigota menos frecuente con 2. Se eliminaron aquellos marcadores con frecuencia alélica menor a 0.01 y aquellos con más del 30% de datos faltantes. Para calcular matrices de distancias se utilizó el índice de similitud de Jaccard para datos binarios implementada en R con la función *vegdist* del paquete *vegan* con método "jaccard".

Algoritmos comparados para el agrupamiento de observaciones

Tres métodos de agrupamiento fueron implementados, cada método de agrupamiento fue configurado para identificar el número adecuado de subpoblaciones según el escenario de simulación: k=2 para los escenarios E1al E6; k=5 para los escenarios E7 al E12 y k=10 para los escenarios E13 al E18. El *script* para implementar los algoritmos de agrupamiento puede encontrarse en el Anexo II.

El análisis de **conglomerados jerárquico** es una de las técnicas de clasificación más ampliamente usada para analizar datos de muestras basadas en varios *loci* desde la aparición de los primeros marcadores moleculares por la década del '90 (Balzarini *et al.*, 2010). En un algoritmo de agrupamiento jerárquico aglomerativo, inicialmente se consdera a cada objeto como un grupo unitario. Luego, después de sucesivas iteraciones, los objetos o individuos se fusionan entre si formando grupos. El algoritmo continúa

iterando hasta que se alcanzan las condiciones y se detiene, en este caso, hasta que no haya ningún individuo o grupos de individuos que haya que fusionar. Una ventaja importante de este método y que consideramos la razón por la cual ha sido utilizado en una amplia diversida de cultivos y áreas es que no se conoce *a priori* el número de grupos subyacentes y no es necesario indicarlo para que el algoritmo trabaje. Para que un método de agrupamiento jerárquico comience a funcionar es necesario determinar una métrica de distancia a partir de la cual generar una matriz de distancia entre todos los pares de invididuos y luego un criterio o algoritmo de agrupamiento. La selección de una determinada métrica y/o de un método de agrupamiento puede provocar configuraciones diferentes de agrupamiento para un mismo conjunto de datos (Bruno et al., 2003). El algoritmo UPGMA (*unweigthed pair-group arithmetic average method*) es aplicado sobre una matriz de distancias y los genotipos se agrupan utilizando un criterio de agrupación basado en distancias promedio no ponderadas entre todos los pares de individuos pertenecientes a grupos diferentes. El primer paso del algoritmo es seleccionar los dos elementos que se encuentre a la menor distancia (más próximos) y formar una clase o grupo con ellos de manera que en los pasos sucesivos del algoritmo estos elementos seguirán juntos, de allí su nombre de jerárquicos, debido a que en cada paso sucesivo del algoritmo va anidando individuos o grupos de individuos a los grupos formados en el paso anterior. Si consideramos una partición inicial donde cada individuos es una clase o grupo, podemos escribir dicha partición como $P_1 = \{x_1\}, \dots, \{x_n\}$ confromada por n individuos, luego, si IJ $= \{x_i, x_j\}$ es un subconjuto confromado por dos elementos denominados i y j cuya distancia entre ellos $, d(x_i, x_j) \ \forall\, i, j = 1, \dots n \mid i \neq j$, es mínima, tendremos una nueva partición $P_2 = \{x_1\}, \dots \{x_i, x_j\}, \dots, \{x_n\}$. Las distancias entre la nueva clase y el resto de las observaciones se calculan como la media aritmética de las distancias entre todos los pares de observaciones antes de la fusión:

$$d(IJ; x_k) = \frac{d(x_i, x_k) + d(x_j, x_k)}{2} \quad k = 1, \dots, n \qquad \text{[Eq. 4]}$$

El algoritmo continua hasta obtener la partición final que contiene todas las observaciones, $P_r = \{N\}$. Este procedimiento permite la clasificación de los objetos que están siendo estudiados y el agrupamiento de los mismos en conglomerados tal que los objetos dentro de un mismo grupo sean más parecidos entre sí que los objetos

pertenecientes a grupos distintos (Bruno, 2005). La visualización de los agrupamientos se realiza a través de un dendrograma que permite identificar la estructura de agrupamiento en función de la distancia a la cual se forman los grupos. Sin embargo, en situaciones donde el número de grupos es elevado, la identificación de los grupos y de los individuos que conforman los grupos se dificultado y convierte este diagrama en un identificador poco práctico. La cantidad de grupos obtenidos dependerá de la distancia a la cual se trace una línea de corte en el dendrograma. Existen muchos *software* estadísticos que tienen implementado este tipo de análisis de conglomerados, como InfoStat y R que proporciona rutinas para realizar agrupaciones jerárquicas a través del paquete stats utilizando la función hclust con el método average como argumento de la función que implementa el algoritmo UPGMA.

Las aproximaciones particionantes, como el **algoritmo no-jerárquico** *k-means* ha sido ampliamente utilizado (Lara y Itzel, 2016), a diferencia de los algoritmos jerárquicos requieren como parámetros iniciales el número de grupos (k) a formar y la métrica de distancia a emplear. El método comienza con una partición inicial (aleatoria) de los genotipos en k cantidad de grupos definidos, calcula el centroide de cada grupo formado y continúa mediante la asignación de cada genotipo en uno de los conglomerados según la distancia mínima entre el individuo y el centroide de modo que la distancia entre el genotipo y el centroide del conglomerado al que se asignó es menor que la distancia media a cualquier otro centroide. Inicialmente, cada una de la p mediciones realizadas sobre una muestra de n observaciones: x_{ij} $(i = 1,\dots,n, j = 1,\dots p)$ es asociada con uno de los k grupos acorde a la distancia de dicho punto con el centroide de cada *cluster*. Luego, en las sucesivas iteraciones del algoritmo, nuevos centroides son calculados y la clasificación de las observaciones es reasignada a un grupo en función de la mínimia distancia al nuevo centroide. El proceso se repite hasta que no se obtiene cambios significativos de la posición del centroide en los sucesivos pasos, minimizando la suma de cuadrados dentro (SSE) de los grupos como $SSE = \sum_{i=1}^{n} \left\| x_i - \hat{\mu}_{y_i^t}^t \right\|^2$. La varianza dentro del conglomerado se puede estimar como SSE/np. La designación *a priori* del número de conglomerados representa la principal limitación del algoritmo *k-means*, la clasificación final puede depender fuertemente de la elección del centroide (Oliva *et al*., 2001). Para implementar este método en R consideramos la función *kmeans* del paquete *stats* que implementa el algoritmo propuesto por MacQueen (1967).

En el **método bayesiano** de agrupamiento difuso denominado *Sructure* por quienes lo propusieron (Pitchard et al. 2000), los genotipos se asignan, de manera probabilística, basado en las cadenas de Markov, a uno de los k grupos configurados o a dos o más conglomerados simultáneamente si el genotipo indica una mezcla de patrones moleculares, asignado una probabilidad a ese individuo de pertenencia a cada grupo la cual será más alta si la similitud es mayor con dicho grupo. En este método, como en otros bayesianos de conglomeración difusa, las probabilidades *a posteriori* indican la incertidumbre de la asignación de conglomerados. Cada individuo se origina en una de las k poblaciones con su propio conjunto característico de frecuencias alélicas. Sea X un vector de individuos (genotipos) (x_l^i) donde el *i-ésimo* individuo en el *l-ésimo* locus, dónde i=1,…,N y l=1,…,L, se supone que los genotipos se generan dibujando alelos independientemente de las distribuciones de frecuencia de población propias como $\Pr\left(x_l^{(i,a)} = j \middle| Z, P\right) = p_{z(i)lj}$, independiente de cada x_l^i. Se utiliza la distribución de Dirichlet para especificar la probabilidad de un conjunto particular de frecuencias alélicas $p_{kl\cdot}$ para la población k en el locus l, $p_{kl\cdot} \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_l})$, independientemente para cada k población y cada l locus. La frecuencia esperada del alelo j es proporcional a λ_j, y la varianza de esta frecuencia disminuye a medida que aumenta la suma de frecuencias alélicas. En el primer paso, el algoritmo estima las frecuencias alélicas para cada población, asumiendo la población de origen de cada individuo como conocida, en el siguiente paso, estima la probabilidad de pertenencia a dicha población de origen, asumiendo que conoce las frecuencias alélicas de dicha población. Para la implementación del algoritmo bayesiano se consideró el paquete LEA en R (Frichot y Francois, 2014).

Criterios de comparación del desempeño de los métodos de clasificación

Como criterio de comparación de los métodos se calculó la proporción de mala clasificación (PMC) en cada réplica de cada escenario de simulación obtenida de matrices de confusión entre el grupo de origen simulado y el vector de asignación obtenido para cada método. Luego, para cada escenario se calcularon medidas resumen del PMC de las 100 réplicas. El PMC mide la proporción de individuos que el algoritmo asigna a una subpoblación que no pertenecía.

$$PMC = \frac{\sum_{i,j=1}^{n} x_{i,j}}{n} \text{ tal que } i \neq j \qquad \text{[Eq. 5]}$$

Donde n es el número total de individuos a agrupar, $x_{i,j}$ es un individuo x que pertenece a la subpoblación i y fue asignado por un algoritmo de agrupamiento a la subpoblación j.

Para comparar los escenarios de simulación se realizó un análisis de la varianza de la proporción de mala clasificación, variable aproximadamente normal, de cada escenario particionado por los métodos. Luego, se realizó el test de comparación de medias DGC (Di Rienzo *et al*., 2002). Del mismo modo, para comparar los métodos de agrupamiento se realizó un análisis de la varianza de la proporción de mala clasificación de cada método particionado por los escenarios de simulación y se realizó la prueba *a posteriori* DGC.

Resultados y Discusión

El método bayesiano *Structure* (MBS) logró una perfecta clasificación en todos los escenarios de simulación obteniendo proporción de mala clasificación nula en cada una de las 100 réplicas de los 18 escenarios (Tabla 2). En Latch *et al*., (2006) usaron datos simulados de microsatélites con cinco subpoblaciones para evaluar el desempeño del *software* STRUCUTRE junto con otros dos *software* de agrupamiento Bayesiano, PARTITION (Dawson y Belkhir 2001) y BAPS (Corander *et al*. 2003, 2004, 2005) , en niveles de diferenciación de población por debajo de Fst = 0,1. Los autores mostraron que con niveles de diferenciación genética superiores a 0,05 el MBS lograba asignar el 100% de los individuos correctamente, lo cual coincide con nuestros resultados en los escenarios de alta divergencia (Fst=0,07). Sin embargo, Latch *et. al*, (2006) obtuvieron que para un Fst de 0,03, en promedio, el 14,7% de los individuos fue mal asignado y para un Fst de 0,05 el 2,2%. Estas diferencias pueden deberse a los distintos tipos de marcadores moleculares utilizados, se sabe que los SNPs son muy frecuentes en los genomas por lo que en la actualidad se puede establecer el genotipo de una planta para varios miles de SNPs muy rápidamente con el fin de identificarla (Spain, 2009) y, con el aumento considerable de variables, los SNPs se convierten en una herramienta muy valiosa para identificar diversidad genética (Würschum *et al*., 2013).

Tabla 2. Medidas resumen de la proporción de mala clasificación (media desvío±estándar) y prueba DGC obtenido a partir del análisis de la varianza de la proporción de mala clasificación de tres métodos evaluados en 18 escenarios de simulación con tres niveles diferenciación genética: baja (Fst=0,03), media (Fst=0,05) y alta (Fst=0,07); tres números de subpoblaciones: k=2, k=5 y k=10 y dos número de individuos: n=250 y n=1000. Cada escenario cuenta con 100 réplicas.

Escenario	Nº de grupos	Diferenciación Genética	Nº individuos	UPGMA		k-means		MBS	
E1	2	Baja	250	0,288 ± 0,247	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E2	2	Baja	1000	0,289 ± 0,248	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E3	2	Media	250	0,231 ± 0,198	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E4	2	Media	1000	0,231 ± 0,198	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E5	2	Alta	250	0,173 ± 0,148	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E6	2	Alta	1000	0,174 ± 0,149	A	0,000 ± 0,000	B	0,000 ± 0,000	B
E7	5	Baja	250	0,552 ± 0,233	A	0,065 ± 0,122	B	0,000 ± 0,000	C
E8	5	Baja	1000	0,552 ± 0,230	A	0,066 ± 0,123	B	0,000 ± 0,000	C
E9	5	Media	250	0,495 ± 0,225	A	0,049 ± 0,098	B	0,000 ± 0,000	C
E10	5	Media	1000	0,494 ± 0,225	A	0,052 ± 0,102	B	0,000 ± 0,000	C
E11	5	Alta	250	0,437 ± 0,220	A	0,027 ± 0,070	B	0,000 ± 0,000	C
E12	5	Alta	1000	0,437 ± 0,219	A	0,028 ± 0,071	B	0,000 ± 0,000	C
E13	10	Baja	250	0,647 ± 0,180*	A	0,066 ± 0,123	B	0,000 ± 0,000	C
E14	10	Baja	1000	0,648 ± 0,179*	A	0,066 ± 0,122	B	0,000 ± 0,000	C
E15	10	Media	250	0,608 ± 0,179*	A	0,052 ± 0,101	B	0,000 ± 0,000	C
E16	10	Media	1000	0,607 ± 0,179*	A	0,051 ± 0,099	B	0,000 ± 0,000	C
E17	10	Alta	250	0,565 ± 0,180*	A	0,036 ± 0,079	B	0,000 ± 0,000	C
E18	10	Alta	1000	0,566 ± 0,180*	A	0,036 ± 0,078	B	0,000 ± 0,000	C

** Mínimo de la proporción de mala clasificación distinto de 0 (en todas las réplicas hubo error de clasificación).*
Medias con una letra distintas indican diferencias estadísticamente significativas (p≤0,05).

El método no jerárquico *k-means* logró una perfecta clasificación en los seis escenarios con dos subpoblaciones (E1, E2, E3, E4, E5 y E6). En los escenarios con cinco y diez subpoblaciones este método presentó un porcentaje de PMC aproximada del 6,6% para los escenarios con baja diferenciación genética (E7, E8, E13 y E14), 5% para los escenarios con divergencia genética media (E9, E10, E15 y E16) y, en cuanto a los escenarios con alta diferenciación, un 2,7% para los de k=5 (E11 y E12) y 3,6% para los de k=10 (E17 y E18) (Tabla 2). Para este método, los mayores valores de PMC se obtuvieron para réplicas de los escenarios con baja divergencia genética (E7, E8, E13 y E14) tanto para k=5 como para k=10 en donde la proporción máxima fue de 0,55 en ambos casos. En cuanto a los escenarios con diferenciación media, en cambio, los máximos valores de PMC alcanzados fueron de 0,49 tanto para los escenarios con 5 subpoblaciones (E9 y E10) como para los escenarios con 10 subpoblaciones (E15 y E16) mientras que con alta divergencia genética (E11, E12, E17 y E18) las réplicas con PMC máximo alcanzaron valores cercanos a 0,42. Exceptuando los escenarios con dos subpoblaciones, en donde se obtuvo perfecta clasificación, se observa que las distribuciones de las PMC se diferencian por la divergencia genética y no así por la cantidad de subpoblaciones, es decir, escenarios con baja diferenciación tienen distribuciones similares de PMC tanto para k=5 como para k=10 y, de igual manera, con escenarios de media y alta divergencia (Figura 4). Estos resultados coinciden con Peña-Malavera *et al*. (2014) que, en un estudio de simulación de datos moleculares del tipo SSR para búsqueda de EGP bajo distintos niveles de divergencia, obtuvo para escenarios con el mismo nivel de diferenciación genética pero distinto número de subpoblaciones (k=3 y k=5) similar proporción de error de clasificación para *k-means*, aproximadamente del 52% al 69% para escenarios con baja diveregencia (Fst=0,07), 1% a 8% con divergencia media (Fst=0,17) y 0% con divergencia alta (Fst=0,38). En nuestros escenarios, los valores más altos de divergencia genética se equiparan con los valores más bajos simulados Peña-Malavera *et al.* (2014). Al contrastar los resultados de estos escenarios equiparados, nuestra proporción de mala clasificación es inferior al 4% en comparación con el promedio del 60% (52-69). Estas diferencias pueden atribuirse al tipo de marcador, SNPs, que si bien se sabe que tienen un contenido informativo menor en comparación con los microsatélites que son altamente polimórficos, esto puede ser compensado por un número mayor de SNPs (Chanca, 2011). Una característica de los SNPs es su mayor abundancia en el genoma respecto a los SSRs y, como la diversidad genética varía entre los cromosomas y a lo largo de ellos (Akhunov *et al,*. 2010; Hao *et*

al., 2011; Alheit *et al.,* 2012), una mayor proporción de varianza se explica a partir de la mayor cantidad de marcadores SNPs en comparación con los SSR (Van Inghelandt *et al.,* 2010). Por otra parte, la naturaleza bialélica de los SNPs y multialélica de los SSRs hace que capturen la diversidad existente de una manera diferente (Würschum *et al.*, 2013), por ellos los marcadores SNPs se han convertido en los marcadores más utilizados para la determinación de la diversidad genética en varios cultivos (Baloch *et al.,* 2017). Un punto a destacar es que, en nuestro trabajo de simulación, al aumentar de 250 a 1000 individuos, la proporción de mala clasificación se mantuvo por debajo del 4%. Mientras que otros autores citan tasas de error son 15 veces mayores con este mismo algoritmo usando 200 individuos y 300 SSR aproximadamente. Cabe destacar entonces que este método presentó, en este trabajo de investigación, un buen comportamiento para identificar grupos de individuos con una dimensión de datos de marcadores de 80K que combina tanto un aumento en la cantidad de variables usadas para la clasificación con mayor número de individuos a agrupar. Dicho de otra manera la información genómica provista por los SNPs permite a un algoritmo como *k-means* obtener buenas clasificaciones aún en contexto de baja de diferenciación entre grupos, alta cantidad de individuos por grupos y alta cantidad de grupos.

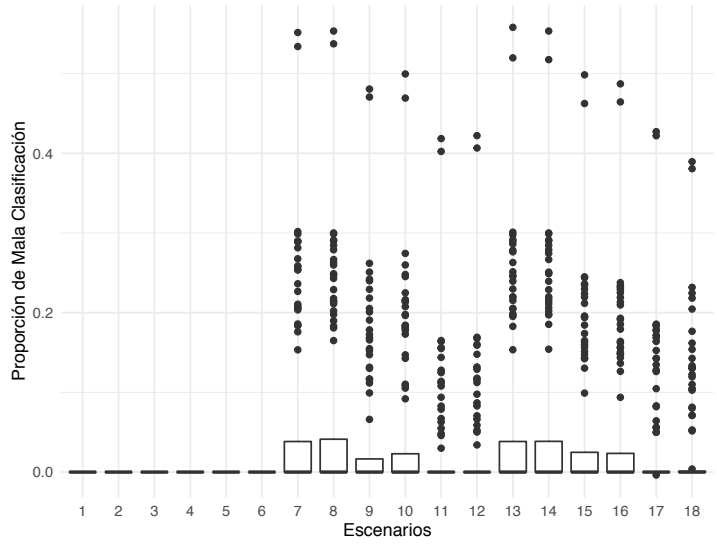

Figura 4. Gráficos de cajas de las proporciones de la mala clasificación para los escenarios de simulación del 1 al 18 (con 100 réplicas cada uno) obtenidas a partir de matrices de confusión entre el agrupamiento simulado y el vector de asignación obtenido por el método *k-means*. Escenarios 1 a 6 para k=2, de 7 a 12 k=5 y 13 a 18 k=10. Escenarios impares tienen 250 individuos, escenarios pares 1000 individuos. Los

escenarios 7, 8, 13 y 14 tiene baja divergencia, los escenarios 9, 10, 15 y 16 tienen media divergencia y los escenarios 11, 12, 17 y 18 alta divergencia.

A partir de la comparación de los escenarios, obtenida por la prueba DGC realizada luego del análisis de la varianza para la poporción de mala clasificación (PMC) para el método *k-means*, los resultados indican que los escenarios con dos subpoblaciones que tuvieron perfecta clasificación difieren estadísticamente del resto de los escenarios. Por otra parte, los escenarios con diferenciación genética alta tanto para k=5 (E11 y E12) como para k=10 (E17 y E18) tienen media de PMC estadísticamente inferior que los escenarios con divergencia baja y media (E7, E8, E9, E10, E13, E14, E15 y E16). Dicho de otra manera, los porcentajes de error de clasificación del método *k-means* no se diferenciaron estadísticamente cuando el número de individuos de la población cambió de 250 a 1000, las diferencias en el error de clasificación se produjeron principalmente por el grado de divergencia genética entre los grupos. De la misma manera para escenarios con igual número de subpoblaciones y el mismo nivel de divergencia genética pero distinto número de individuos, como por ejemplo E7 y E8, no se diferenciaron estadísticamente (Figura 5).

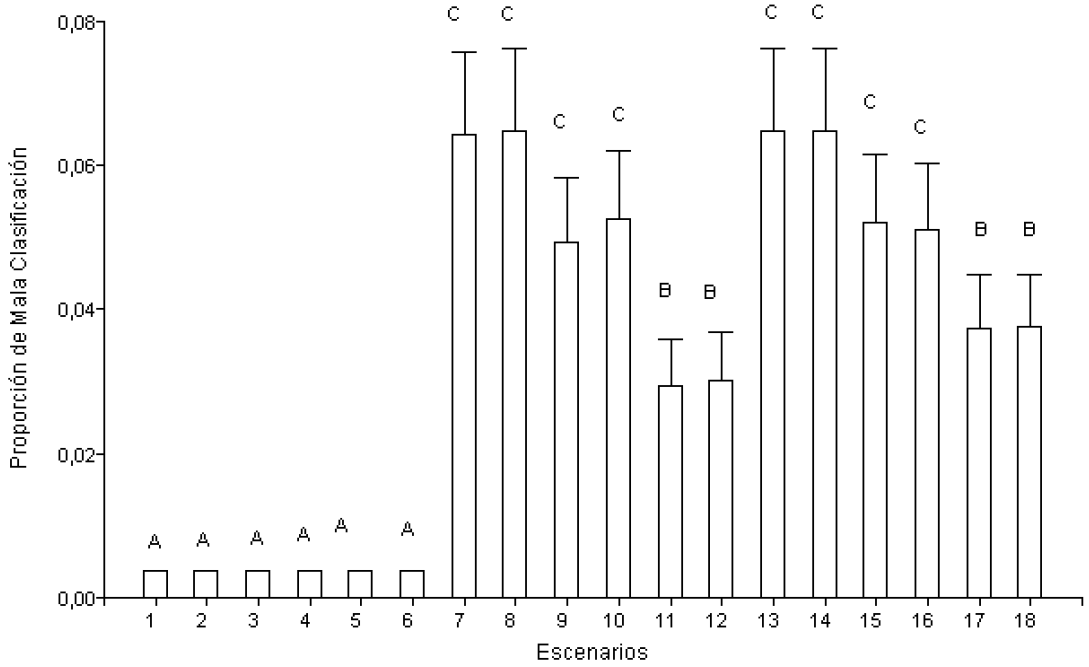

Figura 5. Gráficos de barras de las proporciones de la mala clasificación para 18 escenarios de simulación con 100 réplicas cada uno analizadas con el método de agrupamiento *k-means*. Letras distintas indican diferencias estadísticamente significativas (p≤0,05) obtenidas con la prueba *a posteriori* DGC a partir de un análisis de la varianza. Escenarios 1 a 6 para k=2, de 7 a 12 k=5 y 13 a 18 k=10. Escenarios

impares tienen 250 individuos, escenarios pares 1000 individuos. Los escenarios 7, 8, 13 y 14 tiene baja divergencia, los escenarios 9, 10, 15 y 16 tienen media divergencia y los escenarios 11, 12, 17 y 18 alta divergencia.

El método jerárquico UPGMA no logró una perfecta clasificación en ninguno de los escenarios evaluados. En los escenarios con dos subpoblaciones el método presentó una PMC aproximada de 29% en los escenarios con baja diferenciación genética (E1 y E2), 23% con diferenciación media (E3 y E4) y 17% con diferenciación alta (E5 y E6). Las proporciones de mala clasificación fueron aproximadamente del orden del 50% en los escenarios con cinco subpoblaciones (55% en escenarios con diferenciación genética baja (E7 y E8), 49% media (E9 y E10) y 43% alta (E11 y E12)) y aumentaron al orden del 60% en escenarios con diez subpoblaciones (65% en escenarios con diferenciación genética baja (E13 y E14), 60% media (E15 y E16) y 56% alta (E17 y E18)) (Tabla 2).

Para este método, los mayores valores de PMC se obtuvieron para réplicas de los escenarios con diez subpoblaciones (E13, E14, E15, E16, E17 y E18) en donde la proporción máxima disminuyó levemente a mayor diferenciación genética (máximo PMC E13 y E14: 0,90; E15 y E16: 0,86; E17 y E18: 0,83). En estos escenarios, el método no logró una perfecta clasificación en ninguna réplica (PMC=0), siendo los mínimos porcentajes de mala clasificación de 30%, 25% y 20% para escenarios de baja, media y alta divergencia, respectivamente. Para los escenarios con cinco subpoblaciones los PMC máximos fueron de aproximadamente 0,82 para divergencia baja (E7 y E8), 0,75 para divergencia media (E9 y E10) y 0,69 para divergencia alta (E11 y E12) obteniéndose en todos los casos sólo cuatro réplicas con perfecta clasificación, es decir, PMC=0. Este valor de mala clasificación para baja divergencia genética con k=5 se aproxima a los valores de mala clasificación con el doble de subgrupos a identificar y alta divergencia genética. Podríamos decir que este método encuentra dificultad para agrupar individuos cuando el número de subpoblaciones subyacentes es elevado y a pesar de tener alta diferenciación genética, cuanto más grupos debe identificar mayor es la dificultad para clasificarlos sin error, es decir, asignar a cada grupo los individuos previamente simulados. Estos resultados coinciden con Peña-Malavera *et al*. (2014) que para escenarios con diveregencia Fst=0,07, equiparable con nuestros escenarios de alta divergencia genética, la proporción de error de clasificación aumentó del 61% al 73% cuando la cantidad de subpoblaciones aumentó de k=3 a k=5. Esta situación puede ser

explicada por la forma en la que trabaja el algoritmo, al partir de una matriz de distancia entre todos los pares de individuos puede suceder que se estimen muchos valores iguales de diferenciación genética, a pesar de que dos perfiles moleculares sean polimórficos en distintos marcadores, sus distancias pueden ser iguales debido a la frecuencia de los eventos. Luego, como el algoritmo comienza uniendo aquellos individuos cuya distancia es la más pequeña y dicha unión no cambia en todo los pasos de iteración, entonces, puede ser que combine individuos de distintas subpoblaciones que tienen perfiles moleculares distintos pero cuya distancia genética dada por la cantidad de eventos de co-presencia en el total de individuos los una en los pasos iniciales del algoritmo. Para las simulaciones con dos subpoblaciones las máximas PMC obtenidas fueron 0,51; 0,41 y 0,31 para escenarios con baja (E1 y E2), media (E3 y E34) y alta (E5 y E6) divergencia. Con el método UPGMA, se observa que las distribuciones de las PMC se diferencian por la cantidad de subpoblaciones más que por la divergencia genética, es decir, escenarios con la misma cantidad de subpoblaciones tienen distribuciones similares de PMC para los distintos niveles de diferenciación genética. En los escenarios con dos subpoblaciones se observa una distribución asimétrica a derecha para el PMC, es decir, el 50% de las réplicas obtuvieron valores de PMC muy cercanos al máximo; mientras que, en escenarios con diez subpoblaciones la asimetría es a izquierda (Figura 6).

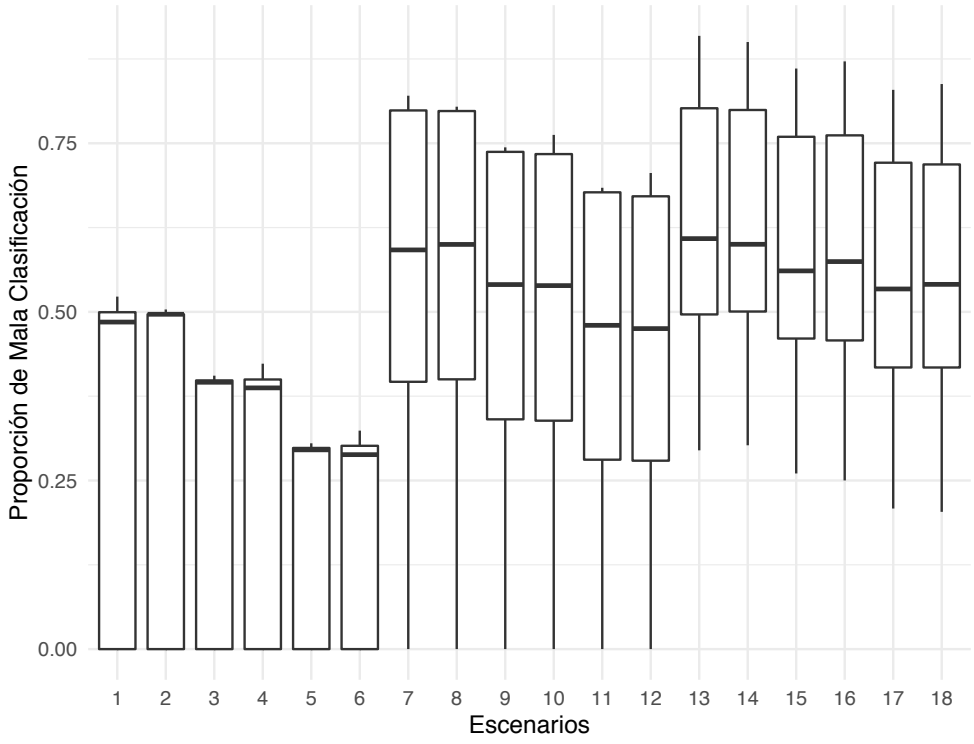

Figura 6. Gráficos de cajas de las proporciones de la mala clasificación para 18 escenarios de simulación con 100 réplicas cada uno obtenidas a partir de matrices de confusión entre el agrupamiento simulado y el vector de asignación obtenido por el método UPGMA.

En los escenarios en donde UPGMA y *k-means* tienen PMC no nula se observa que estas disminuyen cuando la diferenciación genética aumenta, además *k-means* tuvo proporción de mala clasificación menor que UPGMA en todos los escenarios evaluados. Por otra parte, no se observan diferencias entre las proporciones de mala clasificación en escenarios con distinto número de individuos, mismo nivel de diferenciación genética y mismo número de subpoblaciones (Figura 4 y Figura 6). A partir de la comparación de los valores de proporción de mala clasificación (PMC) obtenida por el método UPGMA estimados a partir de los 18 escenarios con un total de 100 réplicas por cada escenarios a través de un análisis de la varianza y la prueba *a posteriori* DGC, los resultados indican que los escenarios con dos subpoblaciones y nivel alto de diferenciación genética (E5 y E6) tuvieron media de PMC estadísticamente inferior al resto de los escenarios. Por el contrario los escenarios con diez subpoblaciones y niveles bajo y medio de divergencia tuvieron media de PMC significativamente superior al resto de los escenarios, es decir mayor tasa de error de clasificación. Escenarios con distinto número de subpoblaciones se diferenciaron estadísticamente entre sí, mientras que escenarios con alto nivel de divergencia genética tuvieron PMC promedio estadísticamente inferior a los escenarios con el mismo número de subpoblaciones pero niveles de divergencia bajos y medios. Para este método, las PMC no se diferenciaron significativamente cuando el número de individuos varió entre 250 y 1000 como se observó para el método *k-means*. Dicho de otra manera, cuando se presentan situaciones donde se espera igual número de subpoblaciones y los niveles de divergencia entre ellos sean similares, el número de subpoblaciones a identificar no dependerá de la cantidad de individuos que contenga la muestra a clasificar (Figura 7).

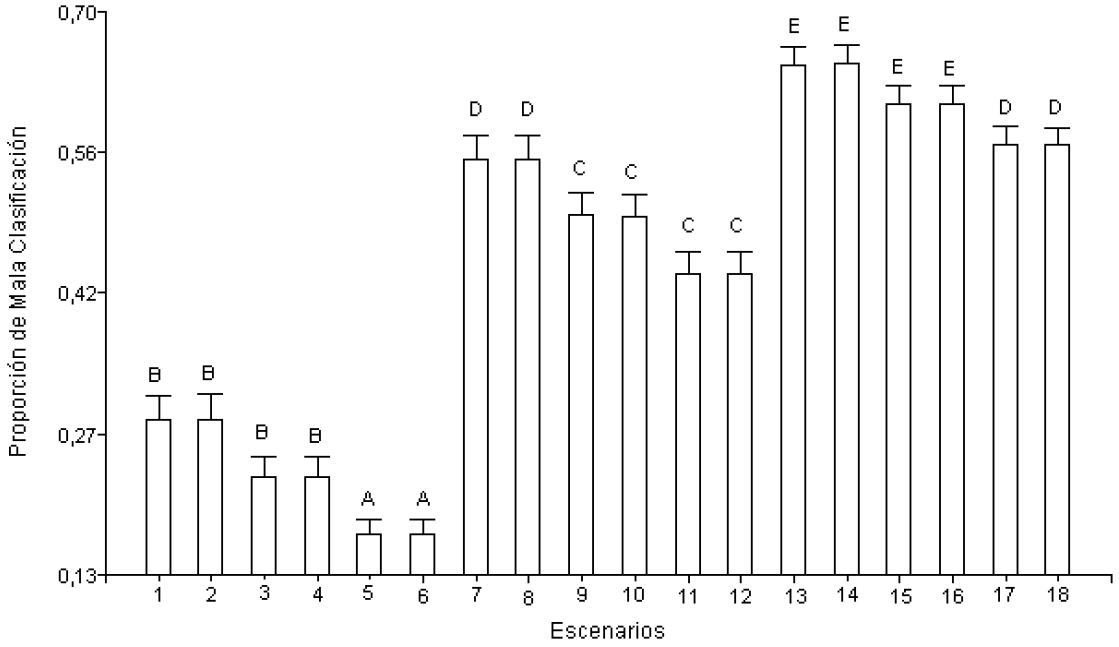

Figura 7. Gráficos de barras de las proporciones de la mala clasificación para 18 escenarios de simulación con 100 réplicas cada uno obtenidas con el método UPGMA rotuladas con letras extraídas de la prueba DGC a partir de un análisis de la varianza. Medias con una letra común no son significativamente diferentes (p>0,05).

Finalmente, a partir de la comparación de los métodos realizada con la prueba DGC a posterior del análisis de la varianza para la proporción de mala clasificación (PMC) obtenida para cada escenario de simulación, se observa que UPGMA presentó una PMC promedio estadísticamente superior respecto a los valores de mala clasificación arrojados por los métodos *k-means* y Método Bayesiano Strucutre en todos los escenarios. En los escenarios con dos subpoblaciones (E1 a E6) los métodos *k-means* y MBS no se diferenciaron significativamente dado a que ambos obtuvieron clasificaciones perfectas en todas las réplicas para la configuración de k=2 dada. En los escenarios con cinco subpoblaciones, divergencia genética baja y media (E7 al E10) estos métodos tuvieron diferencias de la PMC promedio estadísticamente significativas entre ellos. Es decir, los tres métodos identificaron distinta estructura genética poblacional. En los escenarios con esta cantidad de subpoblaciones (k=5), pero con niveles de divergencia altos (E11 y E12) UPGMA y *k-means* tuvieron media de PMC estadísticamente iguales diferenciándose significativamente con el MBS. En todos los escenarios con diez subpoblaciones (E13 al 18), el MBS tuvo media de PMC significativamente menor que los otros dos métodos (Tabla 2).

Conclusiones

En la comparación de los algoritmos de agrupamiento jerárquico, no-jerárquico y bayesiano para la búsqueda de estructura genética poblacional en bases de datos de alta dimensión configuradas bajo parámetros genéticos para simular situaciones que reflejen los acontecimientos que se producen en paneles de líneas del cultivo de maíz, el método Bayesiano *Structure* logró una perfecta clasificación en todos los escenarios configurados, es decir, con distintos números de poblaciones, distintos niveles de divergencia genética y distintos números de individuos, obteniendo proporción de mala clasificación nula en cada una de las 100 réplicas de los 18 escenarios. El desempeño de este método para encontrar la estructura genética poblacional subyacente en los datos fue significativamente mejor que el comportamiento de los otros métodos para situaciones donde se esperan que haya diez subpoblaciones independiente de los niveles de divergencia genética que pueda existir en la población bajo estudio. Lo mismo se observó en el caso de que se espere menor número de subpoblaciones, específicamente cinco subpoblaciones donde los niveles de diferenciación genética entre dichas poblaciones podrían ser bajos o medios. El método no jerárquico *k-means* logró una perfecta clasificación cuando la configuración fue simulada parados subpoblaciones y una PMC estadísticamente nula en los escenarios con cinco subpoblaciones y alto nivel de divergencia genética. Debido a estos últimos resultados es que podemos afirmar que para dichas situaciones los métodos *k-means* y MBS fueron igualmente los de mejor performance. En todas la configuraciones simuladas, UPGMA tuvo la mayor PMC promedio, siendo ésta estadísticamente superior a las PMC arrojadas por los métodos *k-means* y MBS.

Con el método *k-means*, exceptuando los escenarios con dos subpoblaciones, en donde se obtuvo perfecta clasificación, se observa que las distribuciones de las PMC se diferencian por la divergencia genética y no así por la cantidad de subpoblaciones, es decir, escenarios con baja diferenciación tienen distribuciones similares de PMC tanto para k=5 como para k=10 y, de igual manera, con escenarios de media y alta divergencia genética. La prueba DGC realizada luego del análisis de la varianza para comparar los valores de la proporción de mala clasificación (PMC) de cada una de las configuraciones simuladas analizadas con el método *k-means*, se concluye que este método tiene un

excelente desempeño para la identificación de EPG cuando la cantidad de grupos a identificar es de dos. Sin embargo, comienza a cometer errores en la asignación de los individuos según su similitud molecular cuando el número de subpoblaciones subyacentes aumenta a 5 o 10. Así mismo, con cinco poblaciones que presentan una diferenciación genética alta su desempeño es mejor que en situaciones de divergencia genética baja y media.

Con el método UPGMA, se observó que las distribuciones de las PMC se diferencian por la cantidad de subpoblaciones subyacentes a identificar más que por la divergencia genética que presente el conjunto de individuos a través de su perfil molecular. Dicho de otra manera, situaciones o configuraciones biológicas donde se desea indaga a cerca de la estructura genética poblacional, esta técnica tendrá mejor desempeño si las poblaciones presentan mayor diferenciación genética entre ellas, independientemente del número de grupos subyacentes que existan. La estructura genética poblacional estará determinada por el nivel de diferenciación poblacional, en el caso de UPGMA. Cuando la diferenciación genética aumenta, tanto UPGMA como *k-means* detectan la estructura subyacente con menos tasa de error, es decir con menor error de asignación de individuos a los grupos configurados por simulación. Sin embargo, *k-means* tuvo un mejo desempeño que UPGMA en todas las configuraciones simuladas. Tanto para *k-means* como para UPGMA, el número de individuos a clasificar no fue uno de los parámetros de configuración que provocaron una mala asignación de los individuos en los grupos configurados por clasificación. Pero en el caso de UPGMA, esta cantidad de muestras lleva a una visualización difícil de interpretar a través del dendograma. Para ninguno de los métodos comparados, el número de individuos configurados no fue un parámetro que afectara el desempeño para la identificación de la estructura genética poblacional. Sin embargo, esta cantidad de muestras dificulta la visualización gráfica de los agrupamientos, principalmente para UPGMA. En tal caso, la visualización a través de un barplot como el propuesto para el método bayesiano es más fácil y más rápido de interpretar.

CAPÍTULO 3

COMPARACIÓN DE ÍNDICES DE VALIDACIÓN PARA DETERMINAR EL NÚMERO ÓPTIMO DE GRUPOS QUE DETERMINAN ESTRUCTURA GENÉTICA POBLACIONAL

Introducción

Uno de los principales desafíos en el análisis de conglomerados es que, dado que los algoritmos de agrupamiento definen grupos que no son conocidos *a priori*, independientemente del método de agrupamiento, la partición final de los datos requiere alguna clase de evaluación (Rezaee *et al*., 1998). El procedimiento que evalúa el resultado del agrupamiento es conocido como validación del agrupamiento y tiene como finalidad confirmar si la partición de las observaciones o el agrupamiento final obtenido es el que mejor representa la estructura subyacente de los datos (Halkidi *et al*., 2011; Charrad *et al*., 2014). Dado que diferentes algoritmos de agrupamiento sobre un mismo conjunto de datos producen distintos configuraciones de unión, la evaluación de la efectividad en la clasificación y el criterio de agrupamiento son críticos para tener confianza en los resultados de los agrupamientos. Al mismo tiempo, se presenta el dilema del número óptimo de grupos a seleccionar. Numerosos índices han sido propuestos para ello; combinando información acerca de la compactación intra-grupo y el aislamiento inter-grupos, así como otros factores que se encuentran relacionados a la geometría y propiedades estadísticas de los datos, el número de observaciones y la medida de similaridad/disimililaridad. Los índices de validación interna utilizan información propia del conjunto de datos, como por ejemplo la matriz de proximidad, sin considerar información adicional para seleccionar el número de subpoblaciones (Halkidi y Vazirgiannis, 2000). Investigadores de varias disciplinas han propuestos distintos criterios para la selección de número óptimo de grupos (Peng, 2012) como el índice Dunn (Dunn, 1974), CH (Calinski and Harabasz, 1974), el estadístico H (k) (Hartigan, 1975), la estadística de Silueta (Kaufman y Rousseeuw,1990), la estadística de brechas

(Tibshirani *et al*., 2001), el método de remuestreo Clest (Dudoit y Fridlyand, 2002), el método L (Salvador y Chan, 2004) y el índice de conectividad (Handl y Knowles, 2005).

En este trabajo se implementaron cuatro índices de validación: CH, conectividad, Dunn y Silueta. El índice CH se basa en la dispersión dentro del agrupamiento relativo a la dispersión entre grupos, por lo que mayor valor indica número óptimo de subpoblaciones. La conectividad está relacionada a la distancia entre observaciones vecinas en un mismo conglomerado, mientras menor es el valor de conectividad mejor. El ancho de silueta mide la confianza con la que una observación es asignada a un grupo. Si ha sido bien asignada tendrá valores cercanos a 1. El índice de Dunn es el cociente entre la mínima distancia entre dos observaciones que no pertenecen a un mismo conglomerado y la máxima distancia entre dos observaciones de un mismo conglomerado Combina la compactación (homogeneidad dentro del conglomerado) con el grado de separación entre conglomerados (Guy Brock *et al*., 2008). Mayor valor de Dunn implica mayor varianza entre conglomerados y menor varianza dentro del conglomerado.

Materiales y Métodos

Datos Simulados

Los tres métodos de clúster mencionados en el Capítulo 2 fueron implementados, el método de agrupamiento jerárquico Unweighted Pair Group Method with Arithmetic Mean (UPGMA) propuesto por Sokal y Michener (1958), el método de agrupamiento no-jerárquico *k-means* (MacQueen, 1967) y el Método Bayesiano *Structure* (MBS) propuesto por Pritchard *et al*. (2000) a través de su *software* STRUCTURE e implementada en este trabajo a través de la LEA del paquete R (Frichot y Francois, 2014), evaluando de k=2 a k=15 subpoblaciones en los 18 escenarios de simulación mencionados anteriormente. Con el propósito de evaluar el desempeño de cuatro índices de validación de grupo se simularon datos moleculares utilizando usando el paquete "Xbreed" de R (Esfandyari y Sørensen, 2017) bajo distintas configuraciones de estructura geética poblacional recreando posibles escenarios naturales de poblaciones de maíz. Simulamos bases de datos de SNPs para individuos diploides utilizando una población histórica, y con el objetivo de lograr un nivel deseado de desequilibrio de ligamiento (LD), se determinaron los siguientes parámetros genéticos: número de individuos de la población inicial, número de marcadores moleculares, número de generaciones, tasa de

mutación y heredabilidad en sentido estricto. Los 18 escenarios de EGP se configuraron combinando tres números subpoblaciones: k=2, k=5 y k=10; tres niveles de divergencia genética: bajo ($Fst = 0{,}03$), medio ($Fst = 0{,}05$) y alto ($Fst = 0{,}07$) y dos números de individuos: $n = 250$ y $n = 1000$. Cada escenario se replicó 100 veces y cada una de las 1800 bases de datos simuladas tienen 80K de SNPs. Más detalles sobre la simulación pueden encontrarse en el Capítulo 2 y el *script* utilizado en el Anexo I.

Índices de Validación utilizados para determinar número óptimo de grupos en la estructura genética poblacional subyacente

Como criterio de comparación entre algoritmos y para validar el número óptimo de grupos se utilizaros los siguientes índices:

Indice CH (Calinski y Harabasz, 1974)

El índice CH, propuesto por Calinski y Harabsz (1974) es una medida comparativa entre la desviación dentro del agrupamiento y la dispersión entre grupos, teniendo en cuenta la compactación promedio dentro de grupo.

$$CH_{(k)} = \frac{traza(BG)/(k-1)}{\sum_{k=0}^{K} traza(WG^{\{k\}})\,/(n-k)} \qquad \text{[Eq. 6]}$$

donde BG es la matriz de dispersión entre grupos para los datos agrupados dentro de k grupos y $traza(BG) = \sum_{k=1}^{K} n_k(\mu^{\{k\}} - \mu)^T(\mu^{\{k\}} - \mu)$; WG es la matriz de dispersión dentro de los grupos para los datos agrupados en k grupos cuyos coeficientes son $w_{ij}^{\{k\}} = (V_i^{\{k\}} - u_i^{\{k\}})^T(V_j^{\{k\}} - u_j^{\{k\}})$ con $V_i^{\{k\}}$ vector fila i de la matriz de datos y $V_i^{\{k\}}$ el vector columna j de la matriz de datos.

Dunn (Dunn, 1974)

El índice de Dunn es el cociente entre la mínima distancia entre dos observaciones que no pertenecen a un mismo conglomerado y la máxima distancia entre dos observaciones de un mismo conglomerado. La expresión de éste índice puede formularse de la siguiente manera:

$$Dunn = \frac{\min_{k \neq k'} d_{kk'}}{\max_{1 \leq k \leq K} D_k} \qquad \text{[Eq. 7]}$$

donde la distancia entre los conglomerados C_k y $C_{k'}$ se mide por la distancia entre sus puntos más cercanos, estimados como sigue:

$$d_{kk'} = \min_{i \in I_k, j \in I_{k'}} \left\| M_i^{\{k\}} - M_j^{\{k'\}} \right\| \qquad \text{[Eq. 8]}$$

y por otro, considerando el diámetro del agrupamiento como la máxima distancia entre dos puntos de un mismo grupo, que puede expresarse como:

$$D_k = \max_{i,j \in I_k, i \neq j} \left\| M_i^{\{k\}} - M_j^{\{k\}} \right\| \qquad \text{[Eq. 9]}$$

donde $M_i^{\{k\}}$ es la *i-ésima* observación perteneciente al grupo k y $M_j^{\{k\}}$ es la la *j-ésima* observación del mismo grupo k.

Éste índice combina la compactación, es decir la homogeneidad dentro del conglomerado, con el grado de separación entre conglomerados (Brock *et al*., 2011).

Ancho de Silueta (Rouseeuv, 1987)

El ancho de silueta mide la confianza con la que una observación es asignada a un grupo, por lo que se espera que si los genotipos han sido bien asignados se obtendrán valores cercanos a 1.

$$Silhouette = \frac{\sum_{k=1}^{K} \sum_{i \in I_k} S(i)}{n_k} \qquad \text{[Eq. 10]}$$

donde $S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, $a(i) = \frac{1}{n_k - 1} \sum_{i' \in I_k} d(M_i, M_{i'})$ con $i' \neq i$ es la disimilaridad promedio de i-ésimo objeto al resto de los objetos del grupo k, $b(i) = \min_{k' \neq k} \{\frac{1}{n_{k'}} \sum_{i' \in I_{K'}} d(M_i, M_{i'})\}$ es la disimilaridad promedio del *i-ésimo* objeto al resto de los objetos del grupo k. El máximo valor del índice determina el número de grupos

óptimos (Kaufman y Rousseeux, 1990). $S(i)$ no está definido para $k = 1$ (un solo grupo, i.e., sin estructura en los datos).

Conectividad (Handl y Knowles, 2005)

La conectividad está relacionada a la distancia entre observaciones vecinas en un mismo conglomerado, de modo que mientras menor es el valor de conectividad mejor.

Sea $nn_{i\,(j)}$ es el *j-ésimo* vecino más cercano de la observación i, entonces $x_{i,nn_{i\,(j)}}$ es 0 si i y j pertenecen al mismo *cluster* y $\frac{1}{j}$ en caso contrario. Entonces el índice de conectividad para un *cluster* C es

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=i}^{l} x_{i,nn_{i\,(j)}} \qquad \text{[Eq. 11]}$$

donde N es el número total de observaciones a agrupar.

En el Anexo II se encuentra disponible el código de programación en R para implementar estos índices.

Criterios de comparación del desempeño de los índices de validación del número óptimo de grupos

Además de evaluar el desempeño de los algoritmos de agrupamiento a través de los índices de validación interna, propusimos en este trabajo de tesis estimar la tasa de error de clasificación de los mismos. Un algoritmo preciso debería proporcionar resultados razonables, incluso cuando asume un número incorrecto de grupos. Para poder conocer la precisión en el número de grupos sugeridos por el índice de validación, variamos el k número de grupos para cada algoritmo, es decir, forzamos a realizar agrupamientos diferentes a la configuración simulada. Luego, evaluamos en cada réplica, para cada escenario, el número asignado según el índice y el número esperado según el "verdadero" número de grupos. Es importante recordar que el número correcto de grupos subyacente en un conjunto de datos suele no ser conocido, dicho de otra manera, es usual que los individuos que conforman el conjunto de datos reales no tengan una clasificación previa.

A través de la simulación realizada en este trabajo, consideramos como "verdadero" al número de grupos esperado bajo la configuración de la simulación generada. De esta manera se verificó la variación resultante de la precisión como una tasa de error de clasificación que denominamos Error de tipo III (E III). Además, identificamos el error cometido tanto por estimar un número de grupos superior al simulado (E III^{+}) o por el contrario, subestimar (E III^{-}) el número de grupos al sugerir una cantidad de grupos inferior al esperado. Así, para comparar los índices de validación, se calculó la tasa de error de selección del número de grupo, es decir se calculó el error tipo III (E III) para cada método, discriminando entre la sobreestimación del número de grupos (E III^{+}), es decir, el número de simulaciones en las que el índice seleccionó el k mayor que el simulado o esperado, considerado "verdadero" y, la subestimación (E III^{-}), es decir, el número de simulaciones en las que el índice determinó un k menor al simulado.

Resultados y Discusión

Para los escenarios de simulación con dos subpoblaciones (E1, E2, E3, E4, E5 y E6) los cuatro índices de selección tuvieron un error de tipo III de sobreestimación nulo (0%) del número de grupos cuando se utilizaron métodos de *k-means* y el Método Bayesiano *Structure* (MBS); es decir, todos los índices sugirieron que k era igual a 2 para las 100 réplicas de cada escenario cuya configuración simulada era de k igual a 2. El índice de conectividad también tuvo $EIII^{+}$ nulo para los métodos UPGMA y *k-means* en los seis escenarios (Tabla 3, Tabla 4, Tabla 5, Tabla 6, Tabla 7 y Tabla 8, respectivamente). Por el contrario, cuando se implementó UPGMA en los escenarios con k=2 y diferenciación genética baja, tanto para n=250 como para n=1000 (E1 y E2, respectivamente), CH sobrestimó el número de grupos el 58% de las veces, mientras que silueta y Dunn sobrestimaron solo el 2% de las veces. La cantidad de veces que CH indicó el "verdadero" número de grupos con UPGMA fue aproximadamente la misma frecuencia con la que dijo que eran tres grupos en lugar de dos (42 *vs*. 41 simulaciones) (Tabla 3 y Tabla 4). En los escenarios con k=2 y diveregencia genética media (E3 y E4), el $EIII^{+}$ para CH, con este método, disminuyó al 53% y 52% respectivamente (Tabla 5 y Tabla 6) y con diferenciación alta (E5 y E6) al 42% (Tabla 7 y Tabla 8).En todos los casos, CH sugirió como número óptimo de grupo en segundo orden tres subpoblaciones (47 *vs*. 41; 48 *vs*. 40; 58 *vs*. 37 y 58 *vs*. 38 respectivamente). Dicho de otra manera, si bien la mayor cantidad de veces sugiere dos grupos, que sería el número correcto o verdadero a sugerir, un

número prácticamente similar de veces sugiere tres grupos. Si tenemos en cuenta que usualmente no se conoce *a priori* el verdadero grupo, la probabilidad de seleccionar una cantidad de grupos diferente al verdadero es azaroso dado que dicho valor es de aproximadamente 0,50, es decir tenemos prácticamente la misma probabilidad de equivocarnos que de no equivocarnos. Los índices silueta y Dunn, por el contrario, no superaron el 2% de error de sobrestimación en el caso de las simulaciones con nivel medio de divergencia (E3 y E4) y tuvieron $EIII^{+}$ nulo cuando las subpoblaciones tenían alta diferenciación genética (E5 y E6).

La clasificación prácticamente azarosa que obtuvimos de algoritmo UPGMA podría explicarse a través de la forma en la que opera el algoritmo. UPGMA calcula el promedio de las distancias entre individuos de distinto grupos, la media aritmética es una medida de tendencia central y puede provocar una homogeneización de las diferencias o disminuir las diferencias al promediarlas. Así, si las distancias entre grupos o entre individuos de distintos grupos son pequeñas o más pequeñas que las distancias entre individuos del mismo grupo, podría esperarse que la distancia entre individuos dentro de cada grupo se asemeje a la distancia entre grupos. Como consecuencia de bajas diferencias entre grupos podría esperarse que el valor del numerador de CH sea un número más pequeño o semejante al valor que refleja el denominador de la diferencia entre grupo. Según la fórmula expuesta anteriormente (Eq. 6), mientras mayor sea el numerador respecto al denominador, mayor será el valor del índice. Dicho de otra forma, se espera que a mayor diferencias entre grupos conformados, mayor sea el valor del índice. De allí que el criterio de selección del número óptimo de grupos para el índice CH es el máximo coeficiente alcanzado. Entonces, cuando existe baja diferenciación entre grupos y/o una semejanza de la variabilidad entre y dentro de grupo producida por la tendencia central del promedio que tiende a homogeneizar diferencias entre grupos o entre individuos de distintos grupos, producen que valor del índice CH disminuya. Si un grupo tiene poca compactación dentro, es decir, alta variabilidad dentro de grupo, puede suceder que las distancias promedios entre individuos dentro de un grupo sean similares a las distancias entre individuos de diferentes grupos. Es sabido que la variabilidad intragrupo puede disminuir si aumentamos el número de grupos, haciendo que dentro de cada grupo queden individuos de mayor similitud. Sin embargo, aumentar el número de grupos no garantiza tener grupos con mayor diferenciación entre sí. Esta similitud de las distancias entre individuos dentro de un grupo a las disimilitudes entre individuos de

diferentes grupos puede hacer que los coeficientes estimados por CH para cada k valor de grupos puedan ser similares y, al implementar la optimización del criterio, asignar un número de grupos similar al simulado pero no igual al simulado. Una característica importantes es que si bien, tiene alta probabilidad o prácticamente la misma probabilidad de asignar el número correcto que uno incorrecto, como en nuestro caso que indica tres grupos en lugar de dos, el número de k grupos sugerido es cercano al verdadero número esperado. De una manera diferente opera el algoritmo del método *k-means*, ya que éste tiende a minimizar la distancia dentro de grupo calculando la distancia de cada individuo a su centroide. Luego, cuando asigna un individuo a un grupo es porque la distancia de éste al centroide es la mínima distancia encontrada. Cómo consecuencia, para un mismo conjunto de datos el valor de compactación promedio dentro de cada grupo obtenido por *k-means* es mayor que el obtenido por UPGMA, resultando en valores de CH más pequeño para un mismo valor de k grupo. En este sentido, si un método de agrupamiento jerárquico es utilizado en la clasificación de individuos, se podría esperar que el algoritmos de encadenamiento simple, también denominado single linkage o método del vecino más cercano, produzca que el índice de validación CH arroje resultados similares al obtenido para el método de agrupamiento *k-means*. En el método bayesiano sucede algo parecido al método de agrupamiento no jerárquico *k-means* debido a que cada individuo es asignado según su probabilidad más alta de pertenencia a un grupo. Si dichas probabilidades están bien diferenciadas entre grupos, la compactación dentro de cada grupo será mayor y las diferencias entre grupos también, maximizando el índice CH.

Tabla 3. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando dos poblaciónes, nivel bajo de diferenciación genética y 250 individuos (E1). Cada índice se evaluó para k número de grupos (k=2 a k=15).

Método de argupamiento	Índice	Número de Grupos Evaluados (*k*)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	42	41	13	2	2	0	0	0	0	0	0	0	0	0	0,58
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
	Silueta	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número esperado o real de grupos configurados por simulación

Cuando el número de individuos a clasificar aumenta, los algoritmos estiman matrices de distancia entre todos los pares de individuos de mayor dimensión. Para datos de naturaleza binaria, es posible obtener coeficientes de similaridad/dismilaridad iguales entre varios pares de individuos. Sin embargo, puede suceder que dos individuos tengan el mismo coeficiente de disimilaridad que otro par de individuos y que sus perfiles moleculares que dan origen a dichos coeficientes sean bien diferentes, es decir, cada par de individuos debería ser asignado a grupos diferentes según su información genética. El algoritmo UPGMA, comienza agrupando aquellos individuos que se encuentren a menor distancia y los individuos que agrupa en los pasos iniciales no son modificados en los sucesivos pasos del algoritmo. Esto significa que si encontró coeficientes de distancia iguales entre individuos con perfil molecular distinto, al final del algoritmo dicho agrupamiento seguirá unido. Esto no ocurre en los métodos de *k-means* y métodos bayesiano cuya asignación de los individuos a un grupo es dinámica en las sucesivas iteraciones del algoritmo. Así, cuando aumentamos el número de individuos a clasificar, la dimensión de la matriz de distancia entre pares de individuo aumenta, lo que podría causar mayor cantidad de coeficientes de distancia similares entre sí y como consecuencia mayor confusión en la estimación de los índices de validación del número de grupos. Sin embargo, en nuestro estudio de simulación, al pasar de n=250 a n=1000 individuos, lo cual significa aumentar cuatro veces la dimensión de la matriz de distancia ($n \times n$) la tasa de error de sobreestimación se mantuvo constante (Tabla 4 y Tabla 5).

Tabla 4. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando dos poblaciónes, nivel bajo de diferenciación genética y 1000 individuos (E2). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (*k*)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	42	41	14	2	1	0	0	0	0	0	0	0	0	0	0,58
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
	Silueta	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número real de grupos simulados

Al aumentar la divergencia genética en los conjuntos de datos simulados, buscamos representar un aumento en la separación entre grupos. En el estudio de simulación llevado a cabo en este trabajo de tesis, al aumentar la variabilidad entre grupos, nuevamente para los índices de validación Dunn, silueta y conectividad no se produjeron errores de designación del número correcto de grupos en los algoritmos jerárquicos y bayesianos pero si en UPGMA, aunque en menor medida que en los escenarios anteriormente analizados donde, la separabilidad entre grupos era menor (menor divergencia genética). El mejor desempeño de UPGMA en este escenario de media divergencia genética se ve reflejado en que el número de repeticiones que sugirieron el mismo número de grupos con el índice CH que el simulado fue mayor, pasando de 42 a 47 (Tabla 6) y de 42 a 48 al aumentar el número de individuos a 1000 (Tabla 7). De esta manera, se repite el patrón de comportamiento, es decir, aumentando cuatro veces la cantidad de individuos, de n=250 a n=1000, el EIII no se modificó linealmente o en la misma proporción.

Tabla 5. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando dos poblaciónes, nivel medio de diferenciación genética y 1000 individuos (E3). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	47	41	7	3	2	0	0	0	0	0	0	0	0	0	0,53
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
	Silueta	99	1	0	0	0	0	0	0	0	0	0	0	0	0	0,01
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 6. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando dos poblaciónes, nivel medio de diferenciación genética y 1000 individuos (E4). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	48	40	9	2	1	0	0	0	0	0	0	0	0	0	0,52
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	99	1	0	0	0	0	0	0	0	0	0	0	0	0	0,01
	Silueta	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0,02
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número real de grupos simulados

Cuando la separabilidad entre grupos de individuos es mayor, aumenta la variabilidad entre grupos, representada en el numerador de CH y por lo tanto aumenta el valor del índice cometiendo menor error de sugerencia en el número óptimo de grupos cuando el algoritmo utilizado es UPGMA.

Tabla 7. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando dos poblaciónes, nivel alto de diferenciación genética y 1000 individuos (E5). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	58	37	4	1	0	0	0	0	0	0	0	0	0	0	0,42
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 8. Tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética de población simulada utilizando: dos poblaciónes, nivel alto de diferenciación genética y 1000 individuos (E6). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														E III^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
UPGMA	CH	58	38	3	1	0	0	0	0	0	0	0	0	0	0	0,42
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
K-means	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
MBS	CH	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

La columna sombreada muestra el número real de grupos simulados

En los escenarios de simulación que incluyeron cinco subpoblaciones (k = 5 y escenariosE7, E8, E9, E10, E11, E12), cuando se utilizó MBS, los índices para la validación del número seleccionado de grupos tuvieron tasa de sobrestimación (EIII+) nula con silueta y Dunn. Es decir, ambos índices indicaron el número correcto de subpoblaciones (5) en las 100 réplicas de cada escenario. Con el índice CH, en su mayoría (86-97%) fueron seleccionados el número real de grupos. Conectividad subestimó el número de grupos entre 94 y 96% selccionando dos subpoblaciones como número óptimo de grupos entre el 52 y 55% de las réplicas. Con los otros métodos utilizados, el índice de conectividad indicó erróneamente dos grupos en todos los casos, cuando el número real de grupos, configurado a través de la simulación, era cinco. En el método de *k-means*, el resto de los índices también indicaron el número correcto de grupos, es decir, k = 5, en la mayoría de los casos CH 71-79%, Dunn 78-87% y silueta 76-85%. En UPGMA, silueta y Dunn subestimaron el número de grupos indicando erróneamente dos grupos en un 93-

97% y 91-95% respectivamente, mientras que CH indicó más subpoblaciones que cinco obteniendo una tasa de error de sobreestimación aproximadamente del 86% (Tabla 9, Tabla 10, Tabla 11, Tabla 12, Tabla 13 y Tabla 14).

En el caso de *k-means*, el algoritmo de clasificación reasigna cada genotipo (ya clasificados en una partición inicial de 5 grupos) en uno de los conglomerados de modo que la distancia entre el genotipo y el centroide del conglomerado al que se asignó es menor que la distancia a cualquier otro centroide (media). Luego, reúne a los individuos en los K grupos para que la diferencia entre los grupos se maximice y las diferencias entre individuos dentro de cada grupo se minimicen. A pesar de que para *k-means*, el índice de conectividad propuso dos grupos en el 97% al 100% de las simulaciones cuando debía indicar cinco, debiéramos considerar la posibilidad de que índice el validación denominado conectividad no sea suficientemente robusto para estimar el número óptimo de grupo considerando la distancia entre observaciones de un mismo grupo, dado que grupos muy compactos tenderán a disminuir la conectividad sin importar la separación entre grupos. Dicho de otra manera, el índice de conectividad estima su valor a partir de la distancia de cada individuo con su *j-ésimo* vecino más cercanos. Si el vecino pertenece a su grupo entonces asigna un cero a la suma de valores que conforman el índice (Eq. 11), por el contrario, si el vecino pertenece a otro grupo, dr suma 1/j. Así, mientras más individuos cercanos (a menor distancia) pertenezcan a un mismo grupo, más pequeño será el valor de la conectividad. De allí que su criterio de optimización es que el valor más pequeño de conectividad es el sugiere el número de grupos. Así, cuanto menor sea la cantidad de grupos conformados por un método de agrupamiento, menor será el valor de conectividad y por ello tiende a sugerir un número menor de grupos que el verdadero. En nuestro estudio de simulación, al configurar 5 grupos, el índice indica dos la mayoría de las veces con los distintos métodos de agrupamiento. Esto puede ser explicado porque asigna más cantidad de valores distintos de cero debido al aumento de grupos, es decir, de individuos no vecino que sumaran 1/j y no cero. Así, para un conjunto de datos de igual cantidad de individuos, mientras mayor sea el número de grupos, más grande será el valor del índice de conectividad por la manera en que el mismo se construye (Eq. 11). Sin embargo, el índice de conectividad no fue el único que demostró error en la asignación del número óptimo de grupo. Cuando el número de grupos esperado es mayor a dos y el método seleccionado es UPGMA, el índice de validación CH tiene un alto error al sugerir el número probable de subgrupos o de poblaciones genéticas existente. Dado que con este

tipo de algoritmo de clasificación el usuario no presupone un número de grupos, no se sugiere utilizar el índice CH como un método de validación con un método de agrupamiento jerárquico como UPGMA.

Tabla 9. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel bajo de diferenciación genética y 250 individuos (E7). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	11	0	0	2	13	20	15	17	14	3	4	1	0	0	0,11	0,87
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	95	1	0	4	0	0	0	0	0	0	0	0	0	0	0,96	0,00
	Silueta	97	0	0	3	0	0	0	0	0	0	0	0	0	0	0,97	0,00
K-means	CH	3	6	20	71	0	0	0	0	0	0	0	0	0	0	0,29	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	0,22	0,00
	Silueta	0	0	3	76	18	3	0	0	0	0	0	0	0	0	0,03	0,21
MBS	CH	4	0	10	86	0	0	0	0	0	0	0	0	0	0	0,14	0,00
	Conectividad	52	24	20	4	0	0	0	0	0	0	0	0	0	0	0,96	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 10. Tasa de error de subestimación (E III$^-$) y tasa de error de sobreestimación (E III$^+$) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel bajo de diferenciación genética y 1000 individuos (E8). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII$^-$	EIII$^+$
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	11	0	0	2	11	24	16	17	11	2	5	1	0	0	0,11	0,87
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	95	1	0	4	0	0	0	0	0	0	0	0	0	0	0,96	0,00
	Silueta	97	0	0	3	0	0	0	0	0	0	0	0	0	0	0,97	0,00
K-means	CH	3	5	21	71	0	0	0	0	0	0	0	0	0	0	0,29	0,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	0,22	0,00
	Silueta	0	0	3	76	18	3	0	0	0	0	0	0	0	0	0,03	0,21
MBS	CH	4	0	10	86	0	0	0	0	0	0	0	0	0	0	0,14	0,00
	Conectividad	52	24	20	4	0	0	0	0	0	0	0	0	0	0	0,96	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 11. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel medio de diferenciación genética y 250 individuos (E9). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	11	0	0	4	12	19	16	20	8	2	7	1	0	0	0,11	0,85
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	92	2	0	6	0	0	0	0	0	0	0	0	0	0	0,94	0,00
	Silueta	95	0	0	5	0	0	0	0	0	0	0	0	0	0	0,95	0,00
K-means	CH	3	5	16	76	0	0	0	0	0	0	0	0	0	0	0,24	0,00
	Conectividad	97	3		0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	5	8	87	0	0	0	0	0	0	0	0	0	0	0,13	0,00
	Silueta	0	0	3	81	12	4	0	0	0	0	0	0	0	0	0,03	0,16
MBS	CH	2	0	6	92	0	0	0	0	0	0	0	0	0	0	0,08	0,00
	Conectividad	54	24	16	6	0	0	0	0	0	0	0	0	0	0	0,94	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 12. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel medio de diferenciación genética y 1000 individuos (E10). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	11	0	0	4	12	20	15	19	9	2	7	1	0	0	0,11	0,85
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	93	1	0	6	0	0	0	0	0	0	0	0	0	0	0,94	0,00
	Silueta	95	0	0	5	0	0	0	0	0	0	0	0	0	0	0,95	0,00
K-means	CH	4	5	18	73	0	0	0	0	0	0	0	0	0	0	0,27	0,00
	Conectividad	97	3	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	6	13	81	0	0	0	0	0	0	0	0	0	0	0,19	0,00
	Silueta	0	0	3	78	16	3	0	0	0	0	0	0	0	0	0,03	0,19
MBS	CH	3	0	10	87	0	0	0	0	0	0	0	0	0	0	0,13	0,00
	Conectividad	54	24	16	6	0	0	0	0	0	0	0	0	0	0	0,94	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 13. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel alto de diferenciación genética y 250 individuos (E11). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	9	0	0	6	12	19	16	18	7	5	7	1	0	0	0,09	0,85
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	91	0	0	9	0	0	0	0	0	0	0	0	0	0	0,91	0,00
	Silueta	93	0	0	7	0	0	0	0	0	0	0	0	0	0	0,93	0,00
K-means	CH	2	3	17	78	0	0	0	0	0	0	0	0	0	0	0,22	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	2	11	87	0	0	0	0	0	0	0	0	0	0	0,13	0,00
	Silueta	0	0	4	84	11	1	0	0	0	0	0	0	0	0	0,04	0,12
MBS	CH	1	0	2	97	0	0	0	0	0	0	0	0	0	0	0,03	0,00
	Conectividad	55	23	16	6	0	0	0	0	0	0	0	0	0	0	0,94	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 14. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con cinco poblaciones, nivel alto de diferenciación genética y 1000 individuos (E12). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	9	0	0	6	12	19	16	19	9	2	7	1	0	0	0,09	0,85
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	91	0	0	9	0	0	0	0	0	0	0	0	0	0	0,91	0,00
	Silueta	93	0	0	7	0	0	0	0	0	0	0	0	0	0	0,93	0,00
K-means	CH	2	3	16	79	0	0	0	0	0	0	0	0	0	0	0,21	0,00
	Conectividad	97	3	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	1	12	87	0	0	0	0	0	0	0	0	0	0	0,13	0,00
	Silueta	0	0	3	85	11	1	0	0	0	0	0	0	0	0	0,03	0,12
MBS	CH	1	0	3	96	0	0	0	0	0	0	0	0	0	0	0,04	0,00
	Conectividad	54	22	17	7	0	0	0	0	0	0	0	0	0	0	0,93	0,00
	Dunn	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

En los escenarios de simulación con diez subpoblaciones (k = 10 y E13, E14, E15, E16, E17 y E18) los índices de silueta y Dunn tuvieron errores de tipo III de sobreestimación y de subestimación nulos cuando la agrupación se realizó a través del Método Bayesiano *Structure*; mientras que CH subestimó (EIII-) el número de grupos entre un 12 a un 17% y la conectividad cometió Error de tipo III- en el 100% de las veces. Todos los índices subestimaron el número de grupos obtenidos vía *k-means* y UPGMA. Es decir, en todas las repeticiones para una configuración de k=10, independientemente de la cantidad de individuos y del nivel de divergencia entre las poblaciones, para los métodos de agrupamiento basados en distancias, ya sea jerárquico como no-jerárquico, los índices sugirieron un número de grupos inferior al esperado. Además, el índice CH también sobreestimó (EIII+) el número de *clusters* en todas las simulaciones cuando el método de agrupamiento utilizado fue UPGMA, sugiriendo 15 grupos en aproximadamente 77% de las veces que debió indicar el número óptimo de grupos, independientemente del nivel de

separabilidad entre grupos y de la cantidad de individuos que se hayan utilizado en el agrupamiento. El índice de CH siempre sugirió un mayor número de grupos que el esperado, sin embargo, cuando la configuración simulada fue de k=2, el número sugerido era más cercano que en el caso de K=10 donde sugiere 15 en lugar de 10. Este comportamiento del índice Ch de sobreestimar el número de grupos se produjo con el algoritmo UPGMA, puede explicarse en la tendencia del algoritmo UPGMA en formar grupos extremos, es decir con poca cantidad de individuos. Es decir, al promediar las distancias entre individuos de distintos grupos, aquellos individuos que tengan una distancia promedio diferente al resto quedan en un grupo diferente, probablemente conformado por poca cantidad de individuos. Este índice indicó 5 subpoblaciones en el 71 al 78% de las veces cuando se aplicó *k-means*, que puede deberse a que aumentó la compactación dentro de grupo al seleccionar las distancias mínimas de cada individuo a su centroide. A aumentar la compactación intragrupo, disminuye la variabilidad dentro de grupo y aumenta la variabilidad entre grupo disminuyendo el valor. Cuando hay una alta cantidad de grupos, como en este caso de 10 grupos, es posible que la variabilidad dentro estimada con *k-means* sea muy pequeña provocando que el índice de CH disminuya al aumentar el número de grupos. Dicho de otra manera, para un mismo tamaño muestral, si se dividen en mayor cantidad de grupos, se alcanza mayor homogeneidad dentro de grupo ya que habrá menos individuos. Esto, en el caso del índice CH disminuye su valor, como el criterio para sugerir el número óptimo de grupo es el mayor valor, indica un número menor al esperado. Duun y silueta indicaron dos subpoblaciones cuando el algorimto usado fue UPGMA para los tres valores de divergencia genética y los dos niveles de cantidad de individuos en este contexto de 10 grupos simulados. Además, la asignación de dos grupos se produjo en todas las repeticiones de cada escenario. El índice de Dunn considera para su estimación la distancia mínima entre grupos y la evalúa en relación a la distancia máxima dentro de grupos que determina el diámetro del grupo (Eq. 7). Nuevamente nos encontramos en el dilema de que si el algoritmo de agrupamiento, por su forma de trabajar, conforma grupos que en su interior contienen individuos muy distintos al resto y con bajo número de individuos, el diámetro de dicho grupo tenderá a ser pequeño. Cómo el diámetro del grupo es inversamente proporcional al índice de Dunn, a menor diámetro mayor coeficiente de Dunn. Si bien Brock et al. (2011) sugieren que el índice de Dunn combina la compactación con el grado de separación entre conglomerados, en nuestro estudio de

simulación debiéramos esperar que a niveles de mayor divergencia genética, es decir grupos con mayor separabilidad el índice tenga mayor coeficiente y por lo tanto sugiera el número de grupos verdadero. Sin embargo, en este trabajo de simulación, cuando el número de grupos configurados por simulación fue superior a cinco y el método utilizado era UPGMA, este índice no demostró ser potente para identificar el número de grupos esperado aun cuando la separación entre conglomerados fue configurada como alta. Un comportamiento similar del índice de Dunn fue observado en los resultados cuando el método de agrupamiento fue *k-means*, pero en este caso, en lugar de indicar dos grupos como con UPGMA, Dunn sugirió cinco grupos en el 79% de las réplicas aproximadamente (Tabla 15, Tabla 16, Tabla 17, Tabla 18, Tabla 19 y Tabla 20). Este comportamiento refuerza lo indicado anteriormente respecto al índice. Nuestra interpretación nos lleva a inferir que, cómo también se indicó anteriormente, *k-means* al trabajar con la distancia mínima entre un individuo y su centroide, tiene a aumentar la compactación dentro de grupo (mayor homogeneidad dentro), lo cual significaría una disminución de la máxima distancia entre individuos del mismo grupo, que es el denominador del índice (Eq. 7). Al aumentar el denominador, debiera disminuir el valor del coeficiente por ser estos inversamente proporcionales. Sin embargo, al aumentar el número de grupos, aumenta la cantidad de distancias entre individuos de distintos grupos obteniendo valores mínimos entre dos observaciones de distintos grupos que no serían calculadas cuando el índice asume menos grupos porque los mismos individuos podrían estar en el mismo grupo. Por ejemplo, si tenemos 100 individuos que son reorganizados en 10 grupos de diez individuos cada uno, el índice de Dunn calculará la distancia entre todos los pares de individuos de todos los grupos. Pero si agrupamos los 100 individuos en cinco grupos de 20 elementos cada uno, habrá distancias entre individuos de distintos grupos que en esta oportunidad no serán estimadas dado que esos dos individuos se encuentran en el mismo grupo. Por lo tanto, es posible que al tener menos grupos de mayor individuo, las distancias estimadas en el numerador sean más grandes y por lo tanto estimar un coeficiente que sugiere mayor cantidad de individuos que el verdadero.

Tabla 15. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel bajo de diferenciación genética y 250 individuos (E13). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	0	6	18	76	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	1	4	17	78	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	99	1	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	5	17	78	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	3	75	21	1	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	1	16	83	0	0	0	0	0	0,17	0,00
	Conectividad	25	15	0	0	12	16	32	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 16. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel bajo de diferenciación genética y 1000 individuos (E14). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	0	6	17	77	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	3	5	21	71	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	3	76	18	3	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	0	17	83	0	0	0	0	0	0,17	0,00
	Conectividad	17	17	0	0	16	17	33	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 17. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel medio de diferenciación genética y 250 individuos (E15). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	0	5	18	77	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	3	7	19	71	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	10	15	75	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	3	77	18	2	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	1	14	85	0	0	0	0	0	0,15	0,00
	Conectividad	17	19	0	0	17	15	32	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 18. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel medio de diferenciación genética y 1000 individuos (E16). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	0	6	17	77	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	3	5	21	71	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	6	16	78	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	3	76	18	3	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	0	15	85	0	0	0	0	0	0,15	0,00
	Conectividad	17	19	0	0	16	18	30	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 19. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel alto de diferenciación genética y 250 individuos (E17). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	2	4	16	78	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	1	4	21	74	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	5	17	78	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	4	77	18	1	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	0	12	88	0	0	0	0	0	0,12	0,00
	Conectividad	17	19	0	0	16	18	30	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Tabla 20. Tasa de error de subestimación (E III^{-}) y tasa de error de sobreestimación (E III^{+}) del número de grupos para cuatro índices de validación del número de grupo obtenidos con tres métodos de agrupamiento aplicados a datos moleculares para una estructura genética poblacional simulada con diez poblaciones, nivel alto de diferenciación genética y 1000 individuos (E18). Cada índice se evaluó para k número de grupos (k = 2 a k = 15).

Método de agrupamiento	Índice	Número de Grupos Evaluados (k)														EIII^{-}	EIII^{+}
		2	3	4	5	6	7	8	9	10	11	12	13	14	15		
UPGMA	CH	0	0	0	0	0	0	0	0	0	0	1	5	16	78	0,00	1,00
	Conectividad	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	100	0	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
K-means	CH	2	3	22	73	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Conectividad	98	2	0	0	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	4	17	79	0	0	0	0	0	0	0	0	0	0	1,00	0,00
	Silueta	0	0	3	77	19	1	0	0	0	0	0	0	0	0	1,00	0,00
MBS	CH	0	0	0	0	0	0	0	12	88	0	0	0	0	0	0,12	0,00
	Conectividad	17	19	0	0	16	18	30	0	0	0	0	0	0	0	1,00	0,00
	Dunn	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00
	Silueta	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0,00	0,00

La columna sombreada muestra el número real de grupos simulados

Con el MBS, los índices Dunn y silueta indicaron correctamente el número óptimo de grupos en las 100 réplicas de todos los escenarios de simulación, mientras que el índice CH y conectividad sólo con los escenarios de dos subpoblaciones (E1, E2, E3, E4, E5 y E6). CH seleccionó correctamente el número de grupos más del 80% en las simulaciones con cinco y diez subpoblaciones (E7-E18) mientras que conectividad menos del 10%. En estos últimos escenarios de simulación, el índice CH seleccionó mayor cantidad de veces el número de grupos adecuado mientras menor era el nivel de diferenciación genética (Figura 8). En el estudio de simulación de Latch *et. al.*, (2006), en donde evaluaron el desempeño de los *softwares* STRUCTURE, BAPS y PARTITION para identificar la subestructura de la población, los autores concluyeron que para niveles de diferenciación genética iguales o superiores a Fst=0,03, STRUCTURE estimó correctamente el número de grupos utilizando el índice Δk propuesto por Evanno et al., (2005). Teniendo en cuenta que nuestros escenarios de simulación tienen divergencia genética igual o superior a

Fst=0,03, los resultados de los índices Dunn y silueta hallados en este trabajo, coinciden con los obtenidos por los autores del trabajo mencionado.

Con el método *k-means* todos los índices seleccionaron el k adecuado o esperado bajo la configuración simulada en las 100 réplicas de los escenarios con dos subpoblaciones. Para los escenarios con cinco subpoblaciones (E7 a E12) todos los índices, excepto conectividad, seleccionaron entre el 70% y el 90% de las veces el número adecuado de grupos mientras que en los escenarios con diez subpoblaciones (E13, E14, E15, E16 y E17) ninguno de los índices seleccionó el número adecuando en ninguna de las réplicas (Figura 8).

Con el método UPGMA, en los escenarios con dos subpoblaciones, conectividad seleccionó el número correcto de grupos el 100%, Dunn y silueta más del 95% de las veces y CH entre el 40% y 50% aumentando este porcentaje con niveles más bajos de diferenciación genética. En los escenarios con cinco subpoblaciones CH, Dunn y silueta sólo seleccionaron el número adecuado de argupamiento menos del 10% de las veces y conectividad 0%. En los escenarios con diez subpoblaciones ningún índice seleccionó, con este método, el número correcto de grupos (Figura 8).

En Starczewski (2017) se realiza la comparación de los índices Dunn y silueta, entre otros, implementados para los agrupamientos obtenidos con los métodos UPGMA y *k-means*, en 6 conjuntos de datos simulados y 8 conjuntos de datos de la vida real recolectados en áreas diversas, de los cuales se conoce la estructura de grupos. Tres de los 14 conjuntos de datos, 6 simulados más 8 reales, tienen k=2 grupos, cuatro conjuntos tienen k=3, dos de ellos k=4, dos k=6, uno k=7, uno k=9 y uno k=15. Cuando implementó el método UPGMA, el índice Dunn no detectó el número indicado de grupos en 12 de los 14 conjunto de datos y reportó una tasa de precisión (suma de las diferencias entre el número real de conglomerados y el número de agrupamientos indicados por el índice de validación; menos tasa de precisión es mejor) de 1,26. Mientras que, con el método *k-means* el índice no logró identificar la cantidad esperada, que asumía como correcta, de grupos en 10 conjuntos, es decir, "acertó el número de grupos en dos conjuntos de datos más que cuando el agrupamiento había sido realizado con UPGMA. Sin embargo, la tasa de precisión se redujo a 0,27. En cuanto al comportamiento del índice silueta utilizado por Starczewski *op cit*., cuando fue estimado para los agrupamientos configurado con el

método UPGMA, identificó adecuadamente, es decir, según lo esperado, el número de k-grupos en 6 bases de datos de las 14 bases evaluadas. En dicha oportunidad, la tasa de precisión reportada para dicho trabajo fue de 0,52, que disminuyó a 0,18 cuando el algoritmo implementado fue *k-means* en cuya oportunidad el número correcto de grupos sugeridos coincidió en nueve conjunto de datos. Estos resultados coinciden con los obtenidos en nuestras simulaciones donde observamos quelos índices CH, Dunn y silueta tuvieron mejor desempeño al sugerir el número óptimo de grupos cuando el algoritmo utilizado fue *k-means* respecto a UPGMA. Según Starczewski *op cit*., esto puede deberse a que el algoritmo UPGMA crea grupos compactos de diámetros aproximadamente iguales y es sensible a valores atípicos mientras que, *k-means* busca agrupaciones compactas alrededor de una media. Nosotros coincidimos con este autor en que UPGMA es sensible a valores atípicos y que puede crear grupos de diámetros similares, sin embargo, consideramos que *k-means* genera grupos más compactos alrededor de su centroide que los creados por UPGMA.

Independientemente del algoritmo de agrupamiento, conectividad fue el índice con mayor tasa de error seguido por CH. Mientras que, Dunn y silueta fueron los de mayor clasificación correcta. En Rendón et. al. (2011) realizaron una comparación entre índices de validación de número de grupos externos e internos (dentro de los índices internos incluyeron los índices CH y silueta) en 13 conjuntos de datos mediante *k-means* y también concluyeron que silueta mostró mejor dempeño que CH. El índice de silueta mostró resultados precisos en 11 de 13 ensayos y CH en 10 de 13 ensayos en el trabajo mencionado.

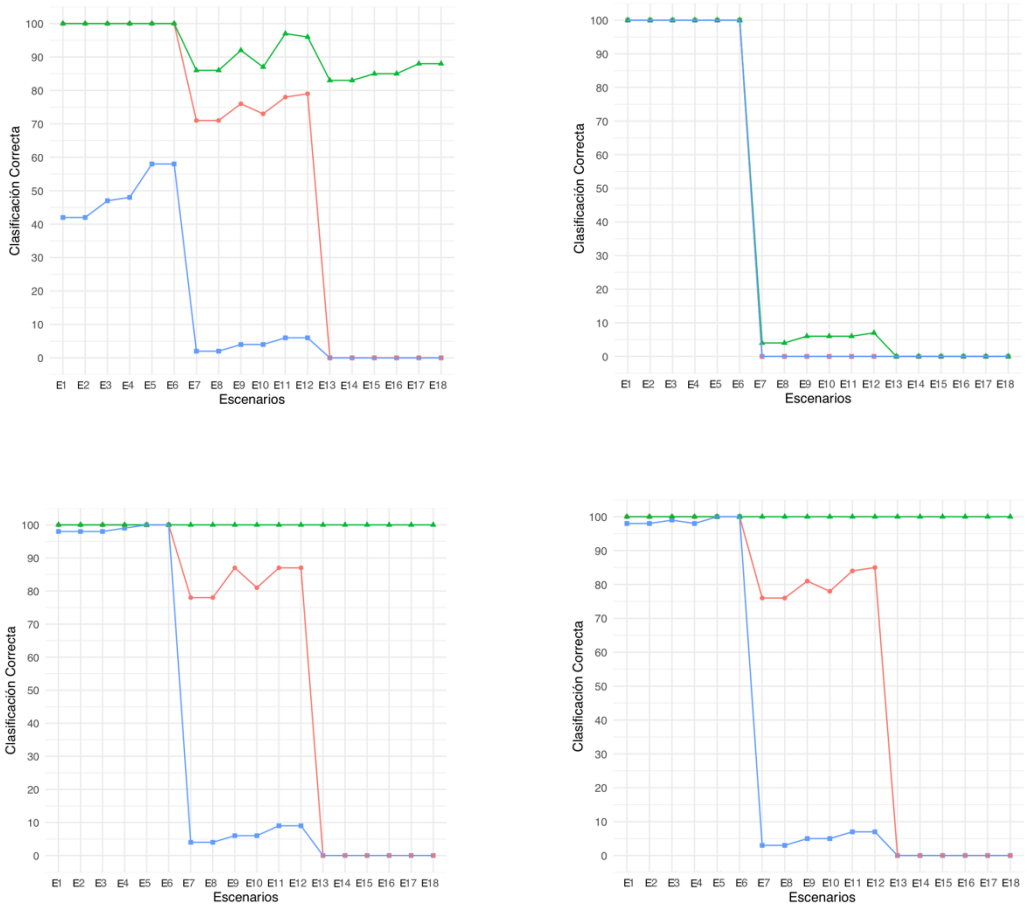

Figura 8. Gráficos de dispersión de la clasificación correcta (cantidad de réplicas en las que el índice seleccionó correctamente el número de grupos) de cuatro índices de validación: CH (izquierda arriba), conectividad (derecha arriba), Dunn (ezquiera abajo) y silueta (derecha abajo) para tres métodos de agrupamiento (*k-means*, Método Bayesiano *Structure* y UPGMA) en 18 escenarios de simulación con 100 réplicas cada uno.

Conclusiones

La determinación del número subyacente de grupos existentes en una población ha sido una pregunta de investigación desde los comienzos de la estadística. A lo largo de la historia se han propuesto numerosos algoritmos de clasificación que contemplaban tanto la naturaleza de los datos como la capacidad de cálculo computacional disponible. Así hemos pasado de métodos jerárquicos que podían ser aplicados a una amplia disposición de métricas de distancias, las cuales podían ser seleccionadas según se trabajara con variables de naturaleza continua, usando la distancia euclídea como de naturaleza discreta como el coeficiente de Pearson o de naturaleza binaria como la distancia de Excoffier (Excoffier, 1992) o los índices de similitud (Bruno et al., 2003) o la mezcla del tipo de variables como la propuesta por Gower (1971). Estas métricas con diferentes algoritmos de agrupamiento dentro de los jerárquicos han mostrado ser potentes en un contexto de dimensión más pequeño al evaluado en este trabajo. Con el advenimiento de tecnologías capaces de generar bases de datos con mayor dimensión, tanto aumentar el número de variables como el número de individuos, los métodos no jerárquicos comenzaron a ser más eficientes a nivel de cálculo computacional, sobre todo en áreas donde el interés radicaba en la delimitación de zonas de manejo homogénea utilizando datos de sensores remotos (Córdoba *et al*., 2020). Es decir, desde los comienzos del análisis de datos, los métodos de agrupamiento o clasificación han sido aplicados en diversas áreas, medicina, agronomía, ecología, genética, etc. como una herramienta objetiva para comprender el ordenamiento de los datos. Sin embargo, pocas veces se han llevado adelante estudios de comparación de la eficiencia de los métodos en contextos particulares de datos como en este caso. Actualmente, la disponibilidad de datos genómicos de 80K como los simulados en este trabajo, es cada vez más frecuente. Por ello surgen interrogantes respecto de cómo se comportan estas herramientas en estos nuevos contextos. Nuestros hallazgos sugieren que en un contexto de alta cantidad de grupos subyacentes los métodos basados en modelos bayesianos fueron los de mejor comportamiento para determinar el número de grupos. Dicho método asigna una probabilidad de pertenencia de los individuos a un grupo y en algunos casos dicha asignación puede ser confusa, recordemos que la probabilidad es un valor en el rango [0;1] y puede suceder que dos individuos tengan la misma probabilidad de asignación a dos grupos diferentes. Esta asignación difusa puede causar que haya individuos asignados a un grupo cuya variabilidad en su perfil molecular

genere grupos menos compactos. Sin embargo, según los resultados vertidos en este trabajo, su compactación es mayor dado que las tasas de error de la proporción de asignación fue la más baja con este método. En escenarios con un número de grupo distinto de k = 2, el índice de conectividad tuvo el error de subestimación más alto del número de grupos independientemente del método utilizado. Para determinar el número óptimo de grupos, los índices de Dunn y silueta tuvieron el mejor desempeño cuando se implementó el método bayesiano.

CAPÍTULO 4

VALIDACIÓN DE LOS MÉTODOS EVALUADOS SOBRE DIVERSAS BASES DE DATOS DE MAÍZ

Ilustración sobre datos no simulados

Para ilustrar los resultados de la implementación de los algoritmos e índices evaluados provistos por las simulaciones se utilizaron dos bases de datos generadas y publicadas a partir de ensayos experimentales conducidos por grupos de investigación en genética de maíz. El primer conjunto de datos fue generado a partir de 942 líneas de maíz genotipadas con 899748 marcadores de tipo SNPs para detectar genes candidatos asociados con la biomasa del tallo (altura y diámetro del tallo) y la anatomía del tallo (espesor de la corteza, densidad del haz vascular y área del haz vascular) y publicado por Mazaheri *et al*. (2019a). En dicho trabajo, Mazaheri *et al*. (2019b) identificaron once subpoblaciones a través del programa Admixture Admixture 1.23 (Alexander, 2009) que se basa en un enfoque bayesiano. Así mismo, la conformación de las 11 subpoblaciones sugeridas por el programa fueron verificadas desde el conocimiento biológico de investigaciones genéticas de maíz. Esta clasificación de las 942 líneas de maíz agrupadas en 11 poblaciones fue considerada en nuestro trabajo la verdadera estructura subyacente ("*gold standard*") para evaluar el desempeño de los tres algoritmos de agrupamiento (UPGMA, *k-means* y Método Bayesiano Strucutre) y de los cuatro índices de validación de número óptimo de grupos (CH, conectividad, Dunn y silueta) en la identificación de estructura genética poblacional. Esta base de datos será denominada en adelante como Conjunto de datos I.

La segunda base de datos, constituida por198 líneas estabilizadas de maíz genotipadas con 55769 marcadores moleculares SNPs provisto por el grupo de mejoramiento genético de maíz de la EEA INTA Pergamino. La genotipificación fue realizada para evaluar dichas líneas para enfermedades como carbón de la espiga del maíz, roya, bacteriosis,

tizón entre otras. Se sabe que su base genética está provista por genotipos tropicales dentados, además corresponden con líneas que han sido estabilizadas por el programa. En este contexto, utilizamos el Método Bayesiano *Structure* que demostró una mejor performance en el contexto de alta dimensión de marcadores e individuos para clasificar las líneas de maíz y los cuatro índices: CH, conectividad, Dunn y silueta para seleccionar el número de subpoblaciones óptimo.

Base de datos para ilustración I

Materiales y Métodos

Las 942 líneas de maíz genotipadas por 899784 marcadores SNPs derivados de RNA-Seq publicada Mazaheri *et al*. (2019a) fueron agrupadas en cuatro subpoblaciones denominadas *stiff stalk* (SS), dos subpoblaciones *non-stiff stalk* (NSS), una subpoblación de líneas públicas de amplio origen, una *Iodent* (IDT), una subpoblación de *sweet corn*, una de *popcorn* y una subpoblación *tropical inbreds*. Un total de 201 individuos endogámicos con menos de 0.5 de probabilidad de pertenencia a cualquiera de las subpoblaciones fue clasificado como grupo mixto. Se calculó la divergencia genética promedio entre las subpoblaciones, en términos del estadístico Fst de Wright (1949), que fue de 0.0239±0.009. Las poblaciones menos divergentes fueron las del grupo *sweet corn* y las del grupo de líneas públicas de amplio origen cuyo Fst reportado fue de 0.007 y las poblaciones más diferenciadas fueron SS-B73 y NSS-Mo17 cuyo Fst estimado fue de 0.047 (Tabla 21).

Tabla 21. Valor de divergencia genética (Fst) entre once subpoblaciones de conjunto de datos reales de 942 líneas de maíz genotipadas con 899784 marcadores moleculares del tipo SNPs.

	Broad origin-public	IDT	NSS-Mo17	NSS-Oh43	Popcorn	SS-B13	SS-B37	SS-B73	SS-BSSSC0	Sweet corn
IDT	0.019									
NSS-Mo17	0.018	0.037								
NSS-Oh43	0.009	0.025	0.019							
Popcorn	0.008	0.030	0.024	0.016						
SS-B13	0.020	0.040	0.042	0.030	0.032					
SS-B37	0.014	0.033	0.032	0.022	0.023	0.025				
SS-B73	0.023	0.044	0.047	0.036	0.038	0.035	0.027			
SS-BSSSC0	0.010	0.028	0.029	0.019	0.019	0.021	0.016	0.022		
Sweet corn	0.007	0.028	0.026	0.016	0.011	0.029	0.022	0.035	0.018	
Tropical	0.007	0.026	0.024	0.015	0.012	0.027	0.020	0.031	0.017	0.013

Fuente: Mazaheri *et al.* (2019).

Para evaluar el desempeño de los índices de validación del número de grupo, los algoritmos de agrupamiento UPGMA, *k-means* y MBS se aplicaron para k=2 hasta k=15 grupos y se calcularon los índices CH, conectividad, Dunn y silueta para probar si seleccionaban el número correcto de subpoblaciones (k=11). Para evaluar el desempeño del los algoritmos de agrupamiento se calculó, para cada método, el porcentaje de no coincidencia en la clasificación a partir de una matriz de confusión generada entre la clasificación propuesta por Mazaheri *et al*. (2019b) y el vector de clasificación generado por cada algoritmo para k=11.

Los índices CH, Connectividad y Dunn indicaron k=2 grupos con los agrupamientos obtenidos por los tres métodos evaluados. silueta indicó 5 subpoblaciones con el agrupamiento de UPGMA, 13 grupos con *k-means* y 15 con MBS. En el caso de UPGMA y *k-means* ninguno de los índices alcanzó un extremo relativo en k=11. Sin embargo, el índice de Dunn obtuvo el segundo valor más alto para el número correcto de grupos (k=11). Para comparar los cuatro índices en cuanto a su clasificación se estandarizó su valor de clasificación. La estandarización se obtuvo para cada índice restando la media general obtenida para cada índice a cada valor obtenido y se lo dividió por el desvío estándar. La estandarización se realizó para los tres métodos de agrupamiento (Figura 9). Para los valores estandarizados se espera que los índices CH, Dunn y silueta presenten su valor más alto cuando el número de grupos es 11, i.e., es el número de grupos de referencia según Mazaheri *et al*. (2019b). Para el índice de conectividad, se espera que el valor sea lo más bajo para k=11.

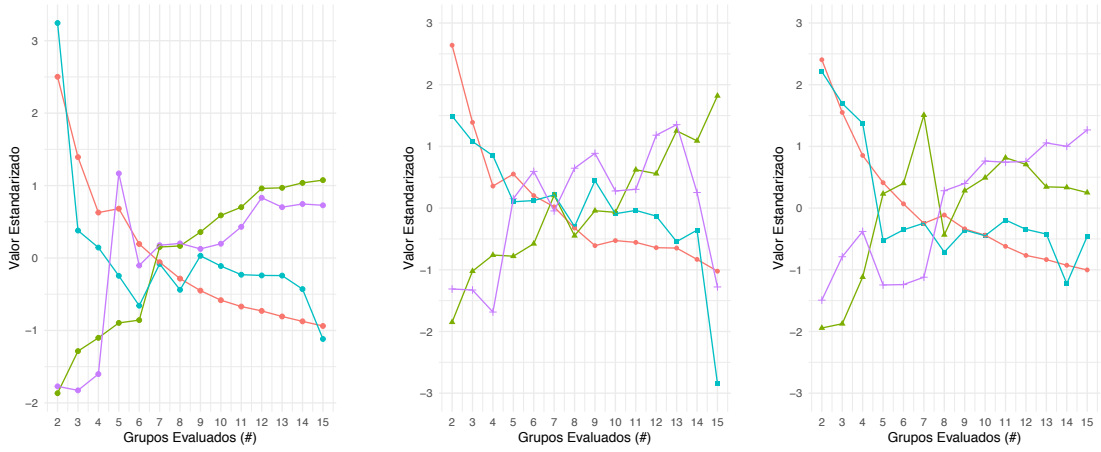

Figura 9. Gráficos de dispersion del valor estandarizado de cuatro índices de validacíon (CH, conectividad, Dunn y silueta) de número óptimo de grupo evaluados para k=2 hasta k=15 para conjunto de datos reales publicado por Mazaheri et al. (2019a) con tres métodos de agrupamiento: UPGMA (izquierda), *k-means* (medio) y método bayesiano *Structure* (derecha). Para los índices CH, Dunn y silueta mayor valor indica el número óptimo de grupo mientras que para conectividad menor valor es el que indica el número óptimo de grupo.

En cuanto a la clasificación de los métodos para el número publicado de subpoblaciones (k=11) UPGMA obtuvo el mayor porcentaje de no coincidencia en la clasificación

mientras que el método bayesiano *Structure* el menor (Tabla 22). Casi tres veces menor porcentaje de no coincidencia en la clasificación que UPGMA, *k-means* tiene la mitad de error de coincidencia que UPGMA y el doble que MBS.

Tabla 22.Porcentaje de no coincidencia en la clasificación de tres métodos de agrupamiento sobre conjunto de datos reales publicados por Mazaheri et al. (2019a)

Método	Porcentaje de no coincidencia en la clasificación
UPGMA	0.5816464
k-means	0.3130904
MBS	0.1835358

Con el MBS, se obtuvo una perfecta clasificación de las cuatro subpoblaciones *stiff stalk* (SS-B13, SS-B37, SSB73 y SS-BSSSC0), de la subpoblación tropical y de una de las subpoblaciones *non-stiff stalk* (NNS-Mo17). La otra subpoblación *non-stiff stalk* ((NNS-Oh43) fue bien clasificada en un 71,2% confundiéndose algunos individuos de esta población con *sweet corn*. La subpoblación Iodent se clasificó correctamente en el 97,1% de los individuos, confundiendo el resto con *sweet corn*. El 65,3% de la subpoblación de líneas públicas de origen amplio fueron agrupadas juntas, confundiéndose el resto de las líneas con *popcorn* y líneas tropicales. La subpoblación de popcorn fue bien clasificada en un 58,3%, confundiéndose en mayor medida con líneas tropicales, *sweet corn* y NNS-Mo17 en menor medida. Por último, la población de *sweet corn* fue confundida casi en su totalidad con popcorn (Figura 10).

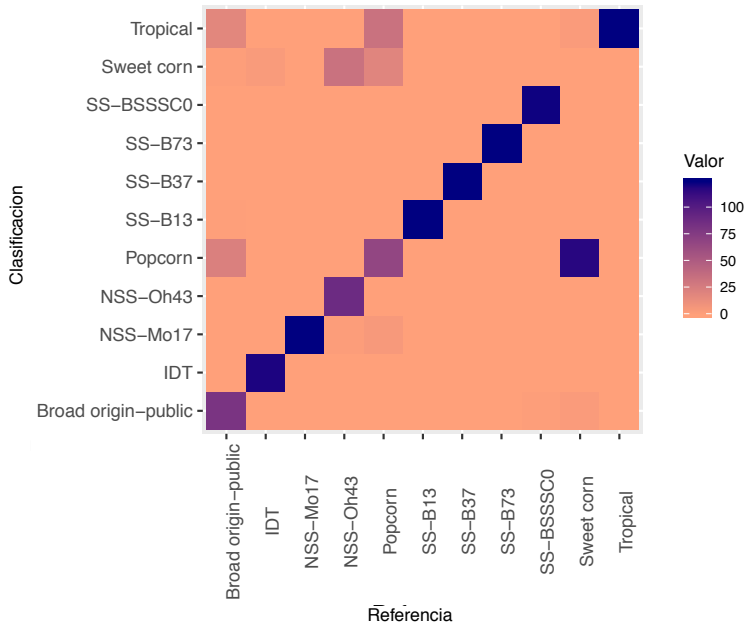

Figura 10. Heatmap de matriz de confusión de porcentaje de clasificación entre clasificación de referencia de conjunto de datos reales publicado por Mazaheri et al. (2019a) y clasificación obtenida por el método bayesiano *Structure* (MBS). Valor cero indica coincidencia exacta (100%) entre la clasificación reportada y la clasificación obtenida por MBS, mientras que valor 100 indica coincidencia nula (0%).

Base de datos para ilustración II

Materiales y Métodos

El conjunto de datos reales provisto por el grupo de mejoramiento genético de maíz de la Estación Experimental (EE) INTA Pergamino para determinar la estructura genética poblacional consta de 198 líneas endocriadas estabilizadas de maíz o individuos genotipados mediante 55769 marcadores moleculares SNPs. El material está compuesto mayormente por germoplasma *Flint* argentino, cultivado a principios del Siglo XX en la región, con introgresiones de otros orígenes, incluyendo materiales del Caribe y del Corn Belt de USA (Olmos *et al.*, 2014), y es representativo del germoplasma de mejoramiento local que usa el sector público y privado (mediante convenio CVT INTA semilleros).

Los marcadores moleculares fueron codificados según el alelo menor, es decir, se codifica al alelo homocigota más frecuente con 0, el alelo heterocigoto con 1 y el alelo homocigota menos frecuente con 2. Se eliminaron aquellos marcadores con frecuencia alélica menor a 0,01 y aquellos con más del 30% de datos faltantes, obteniendo así un total de 51576 marcadores SNPs. Dado a que se desconoce la estructura de este conjunto de datos, para clasificar las líneas de maíz se implementó el método bayeaisano *Structure* para k=2 hasta k=15 ya que fue el método que mejor desempeño mostró en las bases simuladas. Para determinar el número de subpoblaciones se calcularon los cuatro índices evaluados en el capítulo anterior: CH, conectividad, Dunn y silueta.

Resultados y Discusión

El método bayesiano *Structure*, para k=2, clasificó 140 líneas de maíz en una subpoblación (A) y 58 en otra (B). Cuando se indicó agrupar en tres (k=3) subpoblaciones, 13 individuos de la subpoblación (A) y 30 del grupo (B) se agruparon en una nueva subpoblación (C), conformando tres grupos visualmente separados de 127, 28 y 43 individuos respectivamente. A partir de k=4 se observa que los agrupamientos se solapan y, a medida, que el número de subpoblaciones aumenta más solapamiento se observa en los gráficos de dispersión (Figura 11).

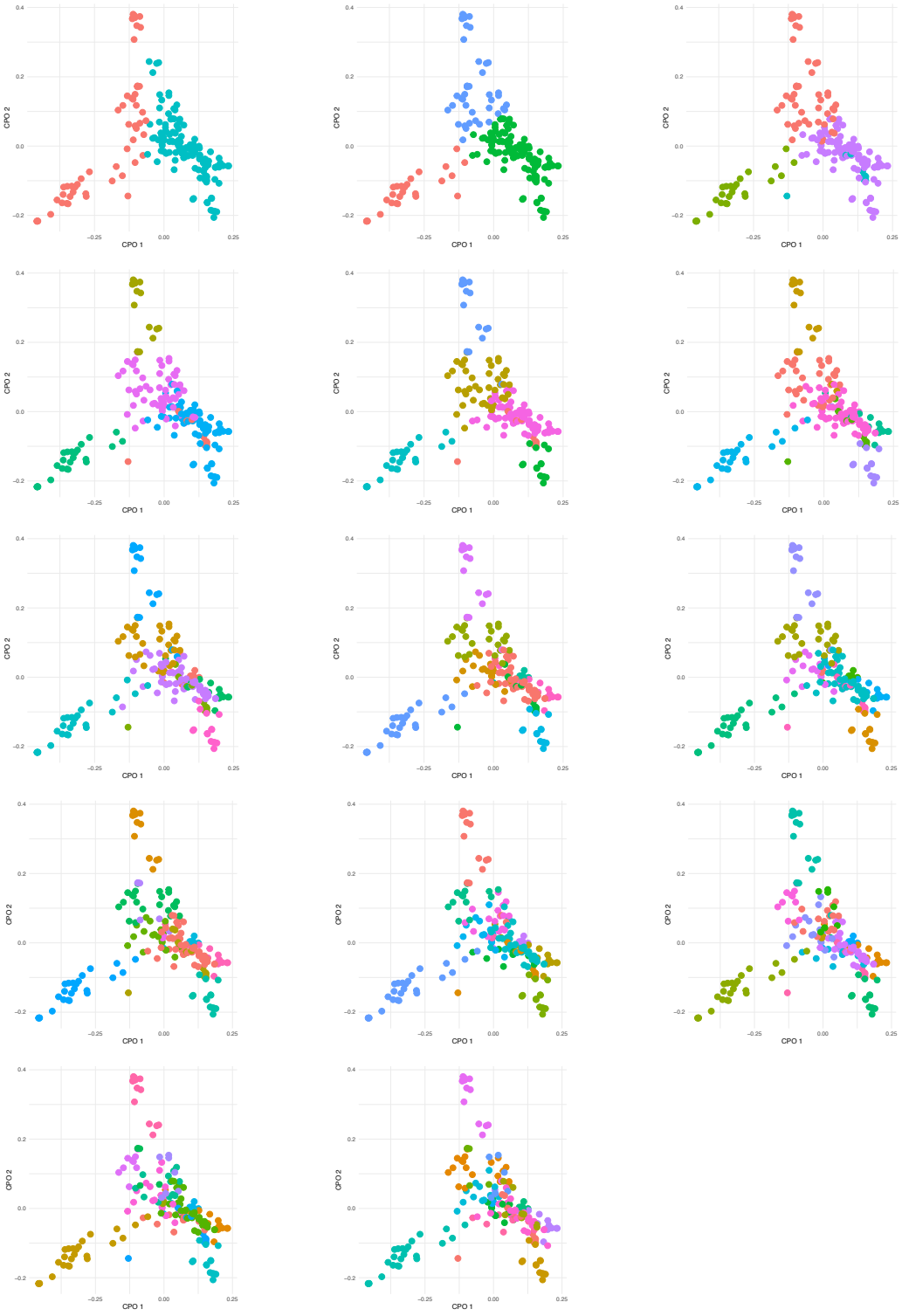

Figura 11. Gráfico de dispersión del análisis de coordenadas principales para conjunto de datos reales proporcionado por el grupo de mejoramiento genético de maíz de la EE INTA Pergamino de 198 individuos genotipados con 55769 SNPs coloreados según la agrupación obtenida por el método bayeaisano *Structure* para (de la esquina superior izquierda a la esquina inferior derecha) k=2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 y 15 grupos.Cada individuo está representado por un punto. Los individuos que pertenecen al mismo grupo se representan con el mismo color.

Para comparar los cuatro índices en cuanto a su clasificación se estandarizó su valor de clasificación. La estandarización se obtuvo para cada índice restando la media general obtenida para cada índice a cada valor obtenido y se lo dividió por el desvío estándar (Figura 12). Para los valores estandarizados se espera que los índices CH, Dunn y silueta presenten su valor más alto cuando el agrupamiento es óptimo y para el índice de conectividad, se espera que el valor sea el más bajo. Los índices CH, Dunn y silueta se maximizaron en k=2 y el índice de conectividad se minimizó también con dos subpoblaciones, es decir, los cuatros índices implementados sugieren seleccionar k=2 grupos.

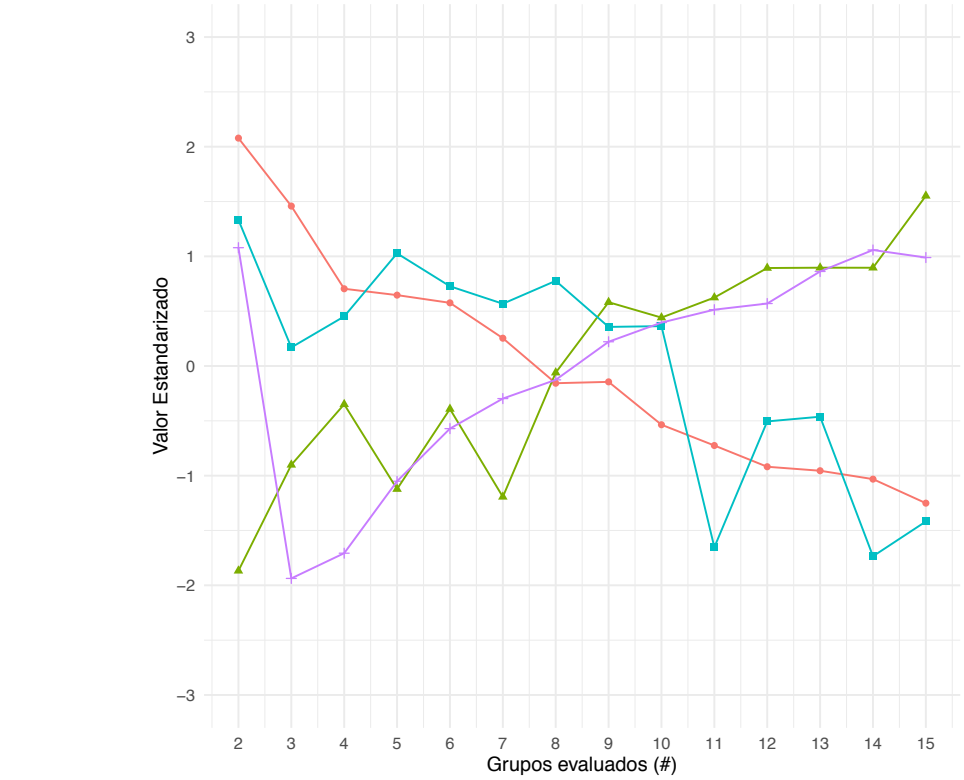

Figura 12. Gráfico de dispersión del valor estandarizado de cuatro índices de validación (CH, conectividad, Dunn y silueta) de número óptimo de grupo evaluados para k=2 hasta k=15 para conjunto de datos reales proporcionado por grupo de mejoramiento genético de maíz de la EE INTA Pergamino con el método de agrupamiento bayesiano *Structure*. Para los índices CH, Dunn y silueta mayor valor indica el número óptimo de grupo mientras que para conectividad menor valor es el que indica el número óptimo de grupo.

Conclusiones

En el primer conjunto de datos reales que se utilizó como ilustración se observó que el método bayesiano *Structure* fue el de mejor desempeño para clasificar los genotipos, mientras que UPGMA sugirió agrupamientos que diferían ampliamente de los esperados. En este caso ninguno de los índices utilizados indicó el número de subpoblaciones que había sido considerado como "gold estándar", siendo Dunn el único en mostrar un extremo relativo para k=11 con la clasificación obtenida por *Structure*. Es probable que la genética de cruzamiento que dio origen a los genotipos clasificados en los 11 grupos sugeridos por los autores, genere un perfil molecular similar entre varios individuos, generando en algunos casos menor compactación intra-grupos y de allí que se observó una mayor discordancia en la asignación en las líneas el grupo *popcorn* y *swetcorn*.

En el segundo conjunto de ilustración el método *Structure* y los cuatro índices de validación CH, conectividad, Dunn y silueta permitieron determinar la estructura genética poblacional de las 198 líneas de maíz del programa de mejoramiento de INTA Pergamino. En este caso los genotipos se agrupan en dos subpoblaciones (k=2) de 140 y 58 individuos cada uno. Esta información es de gran utilidad para posteriores estudios de asociación entre fenotipo y genotipo, dado que incluir información de EGP en modelos de asociación GWAS es fundamental para reducir la tasa de falsos positivos (Malosetti, 2007; Peña *et al.,* 2018).

CAPÍTULO 5

COMENTARIOS FINALES

El análisis de conglomerados es una herramienta útil para clasificar genotipos cuando *a priori* no se conoce el tipo de estructura subyacente en los datos como ocurre en otros métodos de análisis multivariado como el análisis discrimiante lineal donde existe una clasificación previa de los individuos. Sin embargo, el análisis de conglomerados involucra una serie de decisiones que pueden resultar complejas, como la métrica a utilizar y el método de agrupamiento en el caso de los conglomerados jerárquicos. Seleccionar el número de grupos puede implicar un problema al analizar datos reales. Como consecuencia de dichas dificultades, el análisis de conglomerados ha recibido mucha atención. En el presente trabajo se compara el desempeño de tres tipos de algoritmos de clasificación para identificar estructura genética poblacional en datos provenientes de marcadores moleculares y se evalúa la tasa de error de clasificación de cuatro índices utilizados para determinar el número óptimo de grupos identificados por los métodos de agrupamientos no supervisados. Usando el paquete libre de R, se logró configurar no solo diferentes escenarios genéticos que combinaron la variación en el número de líneas, la divergencia genética y la estructura genética determinada por el número de subpoblaciones, si no también obtener los agrupamiento para cada combinación con distintos algoritmos y luego validar los mismos con distintos índices sugeridos para determinar de manera objetiva el número óptimo de agrupamientos logrado. Así, usando el lenguaje de programación del paquete R para escribir los códigos, se ha podido evaluar las propiedades de métodos de agrupamiento en un conjunto de 1800 bases de datos simuladas e ilustrar los resultados obtenidos de las bases de datos simuladas sobredos conjuntos de datos reales.

El enfoque del Método Bayesiano *Structure* fue el que mejor desempeño tuvo para clasificar genotipos en los 18 escenarios evaluados, mientras que UPGMA fue el de peor desempeño, con la proporción de mala clasificación que aumenta con el aumento del número de grupos. En escenarios con un número de grupo distinto de k = 2, el índice de

conectividad presentó el error de subestimación más alto del número de grupos independientemente del método utilizado. Para determinar el número óptimo de grupos, los índices de Dunn y Silhouette tuvieron el mejor desempeño cuando se implementó el método bayesiano.

Además de validar nuestros resultados en repetidas realizaciones de un mismo escenario, aplicamos los mismos métodos en conjuntos de datos reales cuya estructura genética estuvo marcada por las líneas endocriadas de genotipos de maíz que participaban en cada ensayo. En esta oportunidad trabajamos con un panel de líneas obtenidas en el hemisferio Norte americano, con germoplasma adaptado a dichas latitudes y otro conjunto de líneas obtenidas por cruzamiento con genotipos adaptados y estabilizados bajo condiciones climáticas y de manejo agrícola de Argentina. Existe bibliografía que indica que la base genética de todas las líneas de maíz disponibles en Estados Unidos provienen de tan solo seis padres diferentes (Lee y Tracy, 2009). En el hemisferio sur, los materiales están compuesto mayormente por germoplasma Flint argentino, cultivado a principios del Siglo XX en la región, con introgresiones de otros orígenes, incluyendo materiales del Caribe y del Corn Belt de USA (Olmos et al., 2014), y es representativo del germoplasma de mejoramiento local que maneja el sector público y privado (mediante el convenio CVT INTA semilleros).

Es decir, si bien en este trabajo se tomaron dos conjunto de datos con una muestra de todas las líneas existentes de maíz a nivel mundial, entre ambas bases representan estructuras genéticas que pueden significar un conjunto particular de estructura genética no repetible en otro panel de líneas que contengan otras líneas genéticas diferentes. Es por eso que mediante la simulación se intentó abarcar distintas configuraciones de estructuras genéticas además de variar distintos parámetros genéticos. En el primer conjunto de datos reales que se utilizó como ilustración también se observó que el método *Structure* fue el de mejor desempeño para clasificar los genotipos, mientras que UPGMA sugirió agrupamientos que diferían ampliamente de los esperados. En este caso ninguno de los índices evaluados indicó el número establecido de subpoblaciones, siendo Dunn el único en mostrar un extremo relativo para k=11 con la clasificación obtenida por *Structure*.

Con el método bayesiano de *Structure* hemos obtenido clasificaciones perfectas (sin distancia entre el valor esperado y el valor configurado) cuando la divergencia genética fue de 0,03. Latch *et al*. (2016) sugirieron como valores frecuentes de Fst en poblaciones silvestres que podrían reflejar niveles de flujo génico entre poblaciones menores a 0,1. Aún con bajos niveles de divergencia, los métodos lograron identificar grupos cercanos a los configurados, aunque en algunos casos con altas tasas de error de clasificación. Sin embargo, en situaciones de un Fst 10 veces menor al sugerido en Latch *et al*. (2016), la tasa de mala clasificación fue del 95%, como sucedió con los individuos de la población *sweet corn* que fueron asignados como individuos de la población *popcorn* en un 95% (Tabla 21). Esta identificación donde se observa una mezcla de individuos de diferentes subplocaciones puede deberse a evaluaciones o conservaciones genéticas en proceso de diferenciación y que aún los algoritmos computacionales no pueden diferenciar matematicamente. Poder cuantificar la diferenciación genética entre individuos de tal manera que dichos valores puedan establecer diferentes subpoblaciones ha sido motivo de atención en los genetistas de poblaciones desde el inicio de la genética de poblaciones (Wright, 1951). Latch *et al*. (2006) compararon la performance relativa de tres métodos bayesianos para la búsqueda de estructura genética, incluido el *software* STRUCTURE (Pitchard *et al*., 2000), y concluyeron que este método, a pesar de ser el de mejor desempeño, asigna correctamente a los individuos a su subpoblación de origen cuando el Fst es de al menos 0,05 y que por debajo de Fst de 0,03 STRUCTURE no identifica un patrón claro de la estructura genética de los datos. Además, concluyeron que con Fst menores a 0,02 los algoritmos no identifican el número correcto de subpoblaciones y sugiere que este *software* proporciona una certeza falsa con respecto a k cuando Fst es bajo. En nuestro trabajo, si bien no era una población silvestre, los comportamientos del método bayesiano *Structure* fueron similares a los reportados por Latch *et al*. (2016). En agronomía, los programas de mejoramiento genético vegetal buscan obtener líneas endocriadas que resulten promisorias por tener mejores aptitudes de sanidad vegetal, de adaptación a estrés abiótico y de alta producción. En este sentido, a diferencia de las poblaciones silvestres, los cruzamientos son dirigidos disminuyendo la base genética, es decir, los híbridos comerciales en general tienen un ancestro en común. Este ancestro en común puede generar, según como fueron realizados los cruzamientos, que línea fue utilizada como padre y que línea como madre, distintos grado de relación entre los individuos.

Acompañando a la idea de poder cuantificar la magnitud de las diferencias entre individuos con algún grado de parentesco considerando la evolución genética de las poblaciones debido a cruzamientos dirigidos para la construcción de genotipos promisorios para la producción agrícola, poder cuantificar las diferencias entre poblaciones ha sido abordado tanto desde la genética de poblaciones como desde la estadística. Con las biotecnologías que acompañan el proceso de selección genética, la búsqueda de estructura genética poblacional (EGP) en una colección de datos de alta dimensión implica un incremento en la complejidad del manejo de bases de datos masivas como las generadas por los marcadores moleculares de tipo SNPs. Si bien un aumento en el tamaño del conjunto de datos acompaña este cambio, la genómica de poblaciones es más que una simple “big data” de genética de poblaciones debido a que los objetos de estudio son "genomas" y no solo "genes múltiples", lo cual genera un desafío tanto a nivel biológico específico como del análisis estadísticos (Dutheil, 2020). En este sentido, Admixture fue propuesto como una alternativa a *Structure* por su estimación de ancestros entre individuos a través de máxima verosimilitud que vuelven el proceso computacionalmente más eficiente (Alexander y Lange, 2011). Admixture utiliza un algoritmo basado en modelos de ascendencia en individuos no relacionados y adopta el modelo de probabilidad incrustado en la estructura genética. El enfoque es similar al del método bayesiano *Structure*, ambos programas modelan la probabilidad de los genotipos observados utilizando proporciones de ascendencia y frecuencias de alelos poblacionales. Además estiman simultáneamente las frecuencias de alelos de la población junto con las proporciones de ascendencia.

Todos los algoritmos e índices de validación evaluados en este trabajo fueron realizados con el *software* R, de libre acceso. Otros algoritmos podrían haber sido comparados bajo otros escenarios donde además se varíe la cantidad de marcadores moleculares como una extensión a este trabajo.

BIBLIOGRAFÍA

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68.

Abecasis, G. y Wigginton, J. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. The American Journal of Human Genetics, 77(5), 754-767.

Agarwal, M., Shrivastava, N. y Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. Plant cell reports, 27(4):617– 631.

Aguilar, I. (2011). Utilización de la Información Genómica en las Evaluaciones Genéticas. En Segundo Simposio Internacional. p. 93.

Akhunov E., Akhunova A., Anderson O., Anderson J. y Blake N. (2010). Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. BMC Genomics 11:702

Alheit K., Maurer H., Reif J., Tucker M. y Hahn V. (2012). Genome-wide evaluation of genetic diversity and linkage disequilibrium in winter and spring triticale (9 Triticosecale Wittmack). BMC Genomics 13:235

Álvarez, I. (2019). Método de agrupamiento basado en funciones kernel dirigido por índices de validez. Tesis Doctoral. Instituto Politécnico Nacional.

Atlija, M., Gutiérrez Gil, B., Martínez Valladares, M., de la Fuente, L., y Arranz, J. (2013). Barrido genómico con el SNP-CHIP ovino 50K para la detección de QTL con influencia sobre la resistencia a nematodos intestinales en el ganado ovino de raza churra: análisis de ligamento para el recuento de huevos en heces.

Baloch F., Alsaleh A. y Shahid M. (2017). A whole genome DArTseq and SNP analysis for genetic diversity assess- ment in durum wheat from Central Fertile Crescent. PloS One. 2017;12(1):e0167821.

Balzarini, M., Gonzalez, L., Tablada, M., Casanoves, F., Di Rienzo, J. y Robledo, C. (2008). Infostat. Manual del Usuario. Córdoba, Argentina.

Balzarini, M., Teich, I., Bruno, C. y Peña Malavera, A. (2011). Making genetic biodiversity measurable: A review of statistical multivariate methods to study variability at gene level. Revista de la Facultad de Ciencias Agrarias de la Universidad Nacional de Cuyo, 43: 261-275.

Becerra, V. y Paredes, C. (2000). Use of biochemical and molecular markers in genetic diversity studies. Agricultura Técnica, 60(3), 270-281.

Bennetzen, J. y Hake, S. (2009). Handbook of maize: genetics and genomics. Springer Science & Business Media.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop science, 48(5):1649–1664.

Brock, G., Pihur, V., Datta, S., y Datta, S. (2011). clValid, an R package for cluster validation. Journal of Statistical Software (Brock et al., March 2008). Bulletin, 38, 29.

Brookes, A. (1999). The essence of SNPs. Gene 234(2), 177-186.

Bruno C, Balzarini M., Di Rienzo J. 2003. Comparación de Medidas de Distancia entre Perfiles RAPD individuales. Journal of Basic & Applied Genetics, 15 (2):69-78. ISSN BAG-1666-0390

Bruno, C. (2009). Evaluación de métodos de análisis de datos de "marcadores" moleculares. Su aplicación en mejoramiento genético. In FCA-UNC. Escuela para Graduados. , Vol. Dr. en Ciencias Agropecuarias., 177 Córdoba: Argentina

Bruno, C. y Balzarini, M. (2010). Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos. Revista de la Facultad de Ciencias Agrarias UNCuyo 41(3), 11.

Bruno, C. y Balzarini, M. (2010). Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos. Revista de la Facultad de Ciencias Agrarias UNCuyo 41(3): 11.

Caliński, T. y Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27.

Charrad, M., et al. (2014). Package 'nbclust'. Journal of statistical software 61.6 :1-36.

Ching, A., Caldwell, K., Jung, M., Dolan, M., Smith, O., Tingey, S., ... y Rafalski, A. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC genetics, 3(1), 19.

Cook, J., McMullen, M., Holland, J., Tian, F., Bradbury, P., Ross-Ibarra, J., ... y Flint-Garcia, S. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant physiology, 158(2), 824-834.

Corander J., Marttinen P. y Mantyniemi S. (2005) Bayesian identification of stock mixtures from molecular marker data. Fish. Bull., in press.

Corander J., Walmann P. y Sillanpaa M. (2003) Bayesian analysis of genetic differentiation between populations. Genetics, 163, 367–374.

Corander J., Walmann P., Marttinen P. y Sillanpaa M. (2004) BAPS2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics, 20, 2363–2369.

Córdoba, M., Paccioretti, P., Giannini, F., Bruno, C. y Balzarini, M. (2020). Guía para el análisis de datos espaciales. Aplicaciones en agricultura.

Cortes, C. y Vapnik, V. (1995) "Support vector networks", Machine Learning, vol. 20, pp.273-297, 1995

Dawson K. y Belkhir K. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res., 78, 59–77.

Di Rienzo, J, Guzmán, A. y Casanoves, F. (2002). A multiple-comparisons method based on the distribution of the root node distance of a binary tree. Journal of agricultural, biological, and environmental statistics, 7(2), 129-142.

Díaz Muñoz, Y., Sarasa Muñoz, N., Turiño Sarduy, S., Álvarez-Guerra González, E., Cañizares Luna, O., y Machado Díaz, B. (2020). Detección de fenotipos en gestantes sanas de peso adecuado. Medicentro Electrónica, 24(3), 476-490.

Díaz Rodríguez, G. (2016). Principal component analysis of bi-allelic genetic marker data (Master's thesis, Universitat Politècnica de Catalunya).

Dudoit . y Fridlyand J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology 3: research0036.1–0036.21.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.

Ekblom, R. y Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1):1–15.

Erendira Rendon, L., y Abundez, I. (2016). RENTOL: Un algoritmo de agrupamiento basado en K-means. Res. Comput. Sci. 128: 149-157.

Esfandyari, H., y Sørensen, A. (2017). xbreed: An R package for genomic simulation of purebred and crossbred populations.

Evanno, G., Regnaut, S. y Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14(8): 2611-2620.

Excoffier, L., Smouse, P. y Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 131(2), 479-491.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annual Eugenics, 7, Part II, pp. 179-188.

Flores Reinoso, M. (2019). Selección del árbol sobresaliente en eucalyptus spp. por sus características fenotípicas y paramétricas con fines de obtener germoplasma en la parroquia San Mateo, provincia de Esmeraldas (Master's thesis).

Francois, O. y Durand, E. (2010). Spatially explicit Bayesian clustering models in population genetics. Mol Ecol Resour 10(5): 773-784.

Frankham, R., Ballou, J. y Briscoe, D. (2002) Introduction to conservation genetics. Cambridge University press.

Frichot, E. y François, O. (2015). LEA: an R package for landscape and ecological association studies. Methods in Ecology and Evolution, 6(8), 925-929.

Gage, J., White, M., Edwards, J., Kaeppler, S. y de Leon, N. (2018). Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. Genetics, 210(3), 1125-1138.

García Pérez, A. (2020). Identificación de la variabilidad alélica en una colección de Cucurbita sp.

Giraldo, F., León, E. y Gómez, J. (2013). Caracterización de flujos de datos usando algoritmos de agrupamiento. Tecnura, 17(37), 153-166.

González-Recio, O., Rosa, G. y Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livestock Science, 166, 217-231.

Gordon, A. (1999). Clustering. London: Chapman & Hall/HRC Press.

Gower, J. (1971). A general coefficient of similarity and some of its properties. Biometrics 27: 857 – 871.

Guillot, G. y Rousset, F. (2011). On the use of the simple and partial Mantel tests in presence of spatial auto-correlation. Systematic Biology.

Guillot, G., Leblois, R. y Coulon, A. (2009). Statistical methods in spatial genetics. Molecular Ecology 18: 4734-4756.

Haile, J., N'Diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., ... y Pozniak, C. (2018). Genomic selection for grain yield and quality traits in durum wheat. Molecular breeding, 38(6), 1-18.

Halkidi, M. y Vazirgiannis, M.: Quality scheme assessment in the clustering process. In: Proc. PKDD (Principles and Practice of Knowledge in databases), Lyon, France, Lecture Notes in Artificial Intelligence, Spring-Verlag, Vol. 1910, pp. 265–279 (2000)

Halkidi, M., Batistakis, Y. y Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, 17(2), 107-145.

Halkidi, M., Vazirgiannis, M. y Batistakis, Y. (2000, September). Quality scheme assessment in the clustering process. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 265-276). Springer, Berlin, Heidelberg.

Halkidi, Maria, y Iordanis Koutsopoulos. "Online clustering of distributed streaming data using belief propagation techniques." 2011 IEEE 12th International Conference on Mobile Data Management. Vol. 1. IEEE, 2011.

Handl, J. y Knowles, J. (2005) Exploiting the trade-off—the benefits of multiple objectives in data clustering. In Coello,L.A. et al. (eds), Proceedings of the Third International Conference on Evolutionary Multicriterion Optimization. Springer- Verlag, Berlin, pp. 547–560.

Hao, C., Wang, L., Ge, H., Dong, Y. y Zhang, X. (2011). Genetic diversity and linkage disequilibrium in Chinese bread wheat (Triticumaestivum L.) revealed by SSR markers. PLoS One 6(2):e17279

Hartigan, J. (1975). Clustering Algorithms. Wiley.

Hedrick, P. (2005) Large variance in reproductive success and the Ne/N ratio. Evolution 59:1596-1599.

Isidro, J., Jannink, J., Akdemir, D., Poland, J., Heslot, N., y Sorrells, M. (2015). Training set optimization under population structure in genomic selection. Theoretical and applied genetics, 128(1), 145-158.

Jain, A. y Dubes, R. (1988). Algorithms for clustering data. Prentice Hall.

Jain, A., Murty, M. y Flynn, P. (1999). Data Clustering: A Review. ACM Computing Surveys, pp. 651–666

John, S., Shephard, N., Liu, G., Zeggini, E., Cao, M., Chen, W., Va-savda, N., Mills, T., Barton, A., Hinks, A., Eyre, S., Jones, K., Ollier, W., Silman, A., Gibson, N., Worthington, J. y Kennedy, G. (2004). Whole-genome scan, in a complex disease, using11,245 single-nucleotide polymorphisms: comparison withmicrosatellites. Am J Hum Genet 75:54–64

Jombart,, T., Eggo, R., Dodd, P. y Balloux, F. (2009). Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic. PLoS Curr 1: RRN1026.

Kaufman, L. y Rousseeuw, P. (1990). Finding groups in data: an introduction to cluster analysis. John Wiley and Sons, New York, pp. 342.

Kennedy, G., Matsuzaki, H., Dong, S., Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W. y Zhang, J. (2003). Large-scale genotyping of complex DNA. Nature biotechnology, 21(10), 1233-1237.

Kim, J., Santure, A., Barton, H., Quinn, J., Cole, E., Great Tit HapMap Consortium, ... y Slate, J. (2018). A high-density SNP chip for genotyping great tit (Parus major) populations and its application to studying the genetic architecture of exploration behaviour. Molecular Ecology Resources, 18(4), 877-891.

King, B. (1967). Step-wise clustering procedures. Journal of the American Statistical Association, 62(317):86-101.

Kondrashov, A. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Human mutation 21(1), 12-27.

Kruglyak, L. (1997). The use of a genetic map of biallelic markersin linkage studies. Nat Genet 17:21–24

Kwok, P. (2001). Methods for genotyping single nucleotidepolymorphisms. Annu Rev Genomics Hum Genet 2:235–258

Latch, E., Dharmarajan, G., Glaubitz, J. y Rhodes, O. E. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics, 7(2).

Lawson, D. y Falush, D. (2012). Population Identification Using Genetic Data. Annual Review of Genomics and Human Genetics 13(1): 337-361.

Lebart, L. (2000). Contiguity analysis and classification. In Data analysis (pp. 233-243). Springer, Berlin, Heidelberg.

Lee, C., Abdool, A. y Huang, C. (2009). PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 10(Suppl 1), S73.

Lee, C., Abdool, A., y Huang, C. (2009). PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 10(Suppl 1): S73.

Lee, E. y Tracy, W. (2009). Modern maize breeding. In Handbook of Maize (pp. 141-160). Springer, New York, NY.

Liu, C., Sukumaran, S., Jarquin, D., Crossa, J., Dreisigacker, S., Sansaloni, C., y Reynolds, M. (2020). Comparison of array-and sequencing-based markers for genome-wide association mapping and genomic prediction in spring wheat. Crop Science, 60(1), 211-225.

Liu, K., Goodman, M., Muse, S., Smith, J., Buckler, S, y Doebley, J. (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics, 165 (4), 2117-2128

Liu, Y., Wu, G., Yao, Y., Miao, Y., Luikart, G., Baig, M., Beja-Pereira, A., Ding, Z., Palanichamy, M. y Zhang, Y. (2006). Multiple maternal origins of chickens: out of the Asian jungles. Molecular Phylogenetics and Evolution, 38, 12–19.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5-th Berkeley.

Malosetti, M., van der Linden, C., Vosman, B. y van Eeuwijk, F. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics, 175(2), 879-889.

Mammadov, J., Aggarwal, R., Buyyarapu, R., y Kumpatla, S. (2012). SNP mar- kers and their impact on plant breeding. International journal of plant genomics, 2012.

Matise, T., Sachidanandam, R., Clark, A., Kruglyak, L., Wijsman, E., Kakol, J., Buyske, S., et al. (2003). A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkagemap and screening set. Am J Hum Genet 73:271–284

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., ... y Kaeppler, H. (2019a). Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC plant biology, 19(1), 45.

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., ... y Kaeppler, H. F. (2019b), Data from: Genome-wide association analysis of stalk biomass and anatomical traits in maize, Dryad, Dataset, https://doi.org/10.5061/dryad.n0m260p

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5(10): e1000686.

Merino, G. (2018). Imputación de genotipos faltantes en datos de secuenciación masiva (Master's thesis).

Middleton, F., Pato M., Gentile K., Morley C., Zhao, X,, Eisener, A., Brown, A., Petryshen, T., Kirby, A., Medeiros, H., Carvalho, C., Macedo, A., Dourado, A., Coelho, I., Valente, J., Soares, M., Ferreira, C., Lei, M., Azevedo, M., Kennedy, J., Daly, M., Sklar, P. y Pato, C. (2004) Genomewide linkageanalysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a com-parison with microsatellite marker assays and finding of sig-nificant linkage to chromosome 6q22. Am J Hum Genet 74:886–897

Milligan, G. y Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Nachman, M. y Crowell, S. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics 156(1): 297.

Nichols, C., House, J., Li, H., Ward, J., Wyss, A., Williams, J., ... y London, S. (2020). Lrp1 Regulation of Pulmonary Function: Follow-up of Human GWAS in Mouse. American Journal of Respiratory Cell and Molecular Biology, (ja).

Nikolic, N. y Park, Y. (2009). Sancristobal, M. Lek, S. & Chevalet, C. What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations. Genet Res (Camb) 91(02), 121-132.

Nsabiyera, V., Baranwal, D., Qureshi, N., Kay, P., Forrest, K., Valárik, M., ... y Bansal, U. (2020). Fine mapping of Lr49 using 90K SNP chip array and flow-sorted chromosome sequencing in wheat. Frontiers in plant science, 10, 1787.

Odong, T., van Heerwaarden, J., Jansen, J., van Hintum, T. y van Eeuwijk, F. (2011). Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? TAG Theoretical and Applied Genetics 123(2): 195-205.

Oliva, F., Cáceres, M., Font, X. y Cuadras, C. M. (2001). Contribuciones desde una perspectiva basada en proximidades al Fuzzy K-means Clustering. Documento procedente del XXVI Congreso Nacional de Estadística e Investigación Operativa.

Olmos, S., Delucchi, C., Ravena, M., Negri, M., Mandolino, C. y Eyherabide, G. (2014). Genetic relatedness and population structure within the public Argentinean collection of maize inbred lines.

Patricia, M. (2015). Estimación de la diversidad genética mediante marcadores SNP en bovino Criollo Coreño (Bos taurus).

Patterson, N., Price, A. y Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet 2(12): e190.

Paucar Chanca, R. (2011). Utilidad de marcadores SNP en la mejora genética de poblaciones altoandinas de alpacas.

Peña-Malavera, A. (2015). Aproximaciones estadísticas para el mapeo asociativo en estudios genéticos. Tesis Doctoral. Facultad de Ciencias Agropecuarias, Córdoba, Argentina.

Peña-Malavera, A., Bruno, C., Fernandez, E. y Balzarini, M. (2014). Comparison of algorithms to infer genetic population structure from unlinked molecular markers. Statistical applications in genetics and molecular biology, 13(4), 391-402.

Peña, D. (2002). Análisis de datos multivariantes (Vol. 24). Madrid: McGraw-hill.

Peng, Y., Zhang, Y., Kou, G. y Shi, Y. (2012). A multicriteria decision making approach for estimating the number of clusters in a data set. PLoS one, 7(7), e41713.

Piedrahita Gil, J. (2018). Metodología para la identificación automática de flama/humo por medio de análisis de patrones dinámicos de imágenes digitales en entornos abiertos.

Pritchard, J., Stephens, M. y Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155, 945 – 959.

Raj, A., Stephens, M. y Pritchard, J. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets , (Genetics) 197:573-589 [Genetics, Biorxiv]

Reif, J., Melchinger, A., Xia, X., Warburton, M., Hoisington, D., et al. (2003) Use of SSRs for establishing heterotic groups in subtropical maize. Theor Appl Genet 107: 947–957.

Rendón, E., Abundez, I., Arizmendi, A., y Quiroz, E. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications, 5(1), 27-34.

Rendón, E., Abundez, I., Arizmendi, A., y Quiroz, E. M. (2011). Internal versus external cluster validation indexes. International Journal of computers and communications, 5(1), 27-34.

Rezaee, R., Lelieveldt, B., y Reiber, J. (1998). A New Cluster Validity Index for the Fuzzy c-Mean. Pattern Recognition Letters, 19, 237–246.

Riedelsheimer, C., Endelman, J., Stange, M., Sorrells, M., Jannink, J. y Melchinger, A. (2013) Genomic predictability of intercon- nected bi-parental maize populations. Genetics. doi:10.1534/genetics.113.150227

Roux, O., Gevrey, M., Arvanitakis, L., Gers, C., Bordat, D. y Legal, L. (2007). ISSR-PCR: Tool for discrimination and genetic structure analysis of Plutella xylostella populations native to different geographical areas. Molecular Phylogenetics and Evolution 43(1), 240-250.

Ruiz García, N., González Cossio, F., Castillo Morales, A. y Castillo González, F. (2001). Optimización y validación del análisis de conglomerados aplicado a la clasificación de razas mexicanas de maíz. Agrociencia, 35(1).

Sachidanandam, R., Weissman, D., Schmidt, S., Kakol, J., SteinLD, M., Sherry, S., et al (2001) A map of human genomesequence variation containing 1.42 million single nucleotidepolymorphisms. Nature 409:928–933

Sahu, B., Dehuri, S., & Jagadev, A. K. (2017). Feature selection model based on clustering and ranking in pipeline for microarray data. Informatics in Medicine Unlocked, 9, 107-122.

Schaid, D., McDonnell, S., Wang, L., Cunningham, J. y Thi-bodeau, S. (2002) Caution on pedigree haplotype inferencewith software that assumes linkage equilibrium. Am J HumGenet 71:992–995

Schlotterer, C. (2004). The evolution of molecular markers—just a matter of fashion? Nature Reviews Genetics, 5(1):63–69.

Segelbacher, G., Cushman, S. A., Epperson, B. K., Fortin, M. J., Francois, O., Hardy, O. J., Holderegger, R., Taberlet, P., Waits, L. P. y Manel, S. (2010). Applications of landscape genetics in conservation biology: concepts and challenges. Conservation Genetics 11(2): 375-385.

Shaw, S., Oliphant, A., Shen, R., McBride, C., Steeke, R., Shannon, S., Rubano, T., Bahram, G., Fan, J., Chee, M. y Hansen, M.(2004) A highly informative SNP linkage panel for humangenetic studies. Nat Methods 1:113–117

Shriner, D., Vaughan, L., Padilla, M. y Tiwari, H. (2007). Problems with genome-wide association studies. Science, 316, 1840 – 1842.

Sidransky, D. (2002). Emerging molecular markers of cancer. Nature Reviews Cancer, 2(3):210–219.

Sneath, P. y Sokal, R. (1973). Numerical taxonomy. The principles and practice of numerical classification. San Francisco, W.H. Freeman and Company., USA.

Sokal, R. y Michener, C. (1958). A statistical methods for evaluating systematic relationships. University of Kansas Science Symposium on Mathematical Statistics and Probability 1, p. 17.

Spain, La Alberca Murcia. (2009) "IDENTIFICACIÓN RÁPIDA DE VARIEDADES DE VID MEDIANTE NUEVOS MARCADORES DE ADN: SNP."

Starczewski, A. (2017). A new validity index for crisp clusters. Pattern Analysis and Applications, 20(3), 687-700.

Stich, B., Melchinger, A., Frisch, M., Maurer, H. y Heckenberger, M. (2005). Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theoretical and Applied Genetics, 111 (4), 723–730

Suresh, L., Beyene, Y., Olsen, M., Makumbi, D., Oliver, K., Das, B., ... y Gowda, M. (2019). Genetic architecture of maize chlorotic mottle virus and maize lethal necrosis through GWAS, linkage analysis and genomic prediction in tropical maize germplasm. Theoretical and Applied Genetics, 132(8), 2381-2399.

Surveys, pp. 651–666 (1999)

Teich, I. (2012).Análisis de la estructura genética espacial de especies arbóreas y su asociación con la variabilidad fenotípica y ambiental. In Facultad de Ciencias Exactas y Naturales., Vol. Doctorado Buenos Aires: Universidad de Buenos Aires.

Templeton, A. (2006). Population genetics and microevolutionary theory. John Wiley & Sons.

Thomas, A. (2010). Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium. Genetic epidemiology, 34(2), 119-124.

Thorwarth, P., Ahlemeyer, J., Bochard, A. M., Krumnacker, K., Blümel, H., Laubach, E., ... y Schmid, K. (2017). Genomic prediction ability for yield-related traits in German winter barley elite material. Theoretical and Applied Genetics, 130(8), 1669-1683.

Tibshirani, R., Walther, G. y Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

Vekemans, X. y Hardy, O. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. Molecular Ecology 13(4): 921-935.

Verardo, L., Sevón-Aimonen, M., Serenius, T., Hietakangas, V., y Uimari, P. (2017). Whole-genome association analysis of pork meat pH revealed three significant regions and several potential genes in Finnish Yorkshire pigs. BMC genetics, 18(1), 1-15.

Vignal, A., Milan, D., SanCristobal, M. y Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34(3), 275-306.

Vignal, A., Milan, D., Sancristobal, M. y Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34,275–305.

Vigouroux, Y., Glaubitz, J., Matsuoka, Y., Goodman, M., Sanchez, J. y Doebley, J. (2008). Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. American Journal of Botany, 95 (10), 1240-1253

Wang, W., Barratt, B., Clayton, D. y Todd, J. (2005) Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genetics, 6, 109 – 118.

Watson, J. y Francis, C. (1953). Estructura molecular de los ácidos nucleicos. Nature 171 737-8.

Weir, B. y Ott, J. (1997). Genetic data analysis II. Trends in Genetics 13(9): 379.

Windhausen, V., Atlin, G., Crossa, J., Hickey, J., Grudloyma, P., Terekegne, A. et al. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. Genes Genomes Genet 2:1427–1436

Wright, S. (1951). The genetical structure of populations. . Ann. Eugen. 15: 31.

Würschum, T., Langer, S., Longin, C., Korzun, V., Akhunov, E., Ebmeyer, E., ... y Reif, J. (2013). Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. Theoretical and Applied Genetics, 126(6), 1477-1486.

Xia, X., Reif, J., Melchinger, A., Frisch, M., Hoisington, D., et al. (2005). Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II. Subtropical, tropical midaltitude, and highland maize inbred lines and their relationships with elite U.S. and European maize. Crop Sci 45: 2573–2582.

Yan W., Kang M. S., Ma, Baoluo, Woods, S. y Cornelius, P. L. (2007). GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. Crop Breeding & Genetics, 47: 643-653.

Yan, J., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., & Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS one, 4(12), e8451.

Yu, J., Pressoir, G., Briggs, W., Vroh Bi, I., Yamasaki, M., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.

Yuan, J., Wang, X., Zhao, Y., Khan, N. U., Zhao, Z., Zhang, Y., ... y Li, Z. (2020). Genetic basis and identification of candidate genes for salt tolerance in rice by GWAS. Scientific reports, 10(1), 1-9.

ANEXO I

Códigos de R para simulación de datos genéticos

Se presenta el código utilizado para simular 100 réplicas del escenario de simulación 2. Cada réplica cuenta con n=1000 individuos, p=80000 marcadores SNPs y k=2 subpoblaciones con baja divergencia genética. El código está paralelizado utilizando la función *mclapply* de la librería "parallel".

```
# Librerías
###########
library("xbreed")
library("parallel")
library("StAMPP")
library("spDataLarge")

# Argumentos definidos a priori
###############################

args <- commandArgs(TRUE)

sim_start <-1
sim_end <- 100
if (length(args) == 2) {
  sim_start <- as.numeric(args[[1]])
  sim_end <- as.numeric(args[[2]])
}
cat(paste0("Simular desde ", sim_start, " hasta ", sim_end,
"\n"))
n_cores <- 8
if (is.na(n_cores)) {
  n_cores <- Sys.getenv("SLURM_CPUS_PER_TASK")
  if (n_cores == "") {
    n_cores <- max(1, detectCores() - 1)
  }
}
cat(paste0("Corriendo en NCORES: ", n_cores, "\n"))

# Genoma
#########

genome <- data.frame(matrix(NA, nrow = 10, ncol = 6))
names(genome) <- c("chr", "len", "nmrk", "mpos", "nqtl", "qpos")
genome$chr <- c(1:10)
genome$len <- rep(8000,10)
genome$nmrk <- rep(7900,10)
genome$mpos <- rep("even", 10)
genome$nqtl <- c(171,130,153,112,129,66,60,90,63,106)
genome$qpos <- rep("even", 10)


# Marco de datos de Selección
#############################

Selection <- data.frame(matrix(NA, nrow = 2, ncol = 3))
```

```
names(Selection) <- c("Number", "type", "value")
Selection$Number[1:2] <- c(140, 140)

input <- list(genome = genome, Selection = Selection)
rm(list = c("genome", "Selection"))


# Funcion que simulará una población según argumentos indicados
#################################################################

Pob <- function(s1, v1,  hpsize, ng, h2, d2, phen_var, mutr,
laf, input) {
  genome <- input$genome
  Selection <- input$Selection
  # Poblacion Historica
  historical <- make_hp(
    hpsize = hpsize, ng = ng, h2 = h2, d2 = d2, phen_var =
phen_var,
    genome = genome, mutr = mutr, laf = laf
  )
  # Poblacion Simulada
  Breed_A_Male_fndrs <- data.frame(number = Selection$Number[1],
select = s1, value = v1)
  Breed_A_Female_fndrs <- data.frame(number =
Selection$Number[2], select = s1, value = v1)
  Selection$type[1:2] <- c(s1, s1)
  Selection$value[1:2] <- c(v1, v1)
  Breed_A <- sample_hp(
    hp_out = historical, Male_founders = Breed_A_Male_fndrs,
    Female_founders = Breed_A_Female_fndrs,
    ng = 5, Selection = Selection,
    litter_size = 3, Display = TRUE
  )
  P<-Breed_A$output[[6]]$data$phen
  A <- (Breed_A$output[[6]]$sequ)
  A <- A[, -c(1, 2)]
  B <- matrix(rep(0, nrow(A) * ncol(A) / 2), nrow = nrow(A),
ncol(A) / 2)
  for (j in 1:nrow(A)) {
    for (i in 1:(ncol(A) / 2)) {
      B[j, i] <- as.numeric(paste0(A[j, c(i, i + 1)], collapse =
""))
    }
  }

  BSimu<- cbind(P, B)
  BSimu
}

# Especificación de los argumentos
###################################

Sim <- function(i, input) {
  cat(i, "\n")
```

```
  Sim1<-Pob(s1='phen', v1="h", hpsize=1000, ng=200, h2=0.2,
d2=0.1, phen_var=10000, mutr=0.001, laf=0.5, input=input)
  Sim2<-Pob(s1='phen', v1="l", hpsize=1000, ng=200, h2=0.2,
d2=0.1, phen_var=10000, mutr=0.001, laf=0.5, input=input)


  VA<-factor(c(rep(1, nrow(Sim1)), rep(2, nrow(Sim2))))
  BaseP <- rbind(Sim1, Sim2)


  BCodif<-t(apply(BaseP[,-c(1:2)], 1, Codif))
  colnames(BCodif)<-c(paste0("snp", c(1:ncol(BCodif))))
  Base<-data.frame(Sample=c(1:nrow(BCodif)), Pop=VA,
Ploydi=c(rep(2,nrow(BCodif))), Format=(rep("BiA",nrow(BCodif))),
BCodif)
  BaseFreq<-stamppConvert(Base,"r")
  fst<-round(stamppFst(BaseFreq, 100, 95, 2)$Fsts, digits = 4)
  fstMin<-min(fst, na.rm=TRUE )
  fstMax<-max(fst, na.rm=TRUE)
  fstProm<-round(mean(fst, na.rm = TRUE), digits=4)
  fstDE<-round(sd(fst, na.rm = TRUE), digits=4)

  Base <- cbind(VA, BaseP)
  write.table(Base, paste0(paste(i,
"Simulacion_K2","FstMax",fstMax,"FstMin",fstMin,
"FstProm",fstProm, "FstDe",fstDE, sep = "_"), ".txt"),
              sep = "\t", eol = "\n", dec = ".", row.names = F,
col.names = F
  )
  write.table(fst, paste0(paste(i, "fst", sep = "_"), ".txt"))
}

# Paralelización
################

library("parallel")
Tiempo <- system.time({
  Simul<-mclapply(sim_start:sim_end, function(i,input2) {
    Sim(i, input = input2)
  },input, mc.cores = n_cores)
})
Tiempo
```

ANEXO II

Códigos de R para implementación de algoritmos de agrupamiento e índices de validación del número de grupo

Se presenta el código utilizado para la implementación de los algoritmos de agrupamiento: UPGMA, k-means y Método Bayesiano *Structure*. Estos tres algoritmos se presentan como funciones independientes y que, internamente, calcula los índices de validación del número de grupo: CH, Dunn, silueta y conectividad. El código está paralelizado utilizando la función *mclapply* de la librería "parallel".

```
# Directorio del Escenario de Simulacion
path<-"/home/evidela/E4/Simulaciones/"

# Librerías
############

library(parallel)
library(vegan)
library(stats)
library(fpc)
library(caret)
library(clValid)
library(pastecs)
library(LEA)


# Función UPGMA
###############

UPGMA<-function(m, n=15, k, Simulaciones){
  Simulaciones<-list.files(path)
  Base<-read.table(paste(path, Simulaciones[m], sep="/"))[,-
c(1:2)]   # Le saco la columna de asignacion y la de fenotipo
  VA<-read.table(paste(path, Simulaciones[m], sep="/"))[,1]

  Base<-t(apply(Base, 1, Codif))
  Base[is.na(Base)]<-0
  D<-vegdist(as.matrix(Base), method="jaccard", binary=TRUE)

  # Metodo: Vector de Asignacion
  V<-matrix(c(rep(0,nrow(Base)*(n-1))),nrow=nrow(Base),ncol=(n-
1), dimnames = list(1:nrow(Base), c(paste0("k=", 2:n))))
  for (l in 2:n) V[,l-1]=as.vector(cutree(hclust(D,
method="average"), k=l))

  # Indices
  Ind<-data.frame((matrix(NA, nrow=34, ncol=(n-1))))
  for (j in 1:(n-1)) Ind[,j] =
as.data.frame(as.matrix(cluster.stats(D,V[,j])))
  names(Ind)<-c(paste0("k=", 2:n))
  rownames(Ind)<-
rownames(as.data.frame(as.matrix(cluster.stats(D,V[,1]))))
  C=c()
  for(j in 1:(n-1)) C[j]=connectivity(D,V[,j])
  IndexC<-rbind(Ind, Conectividad=C)
```

```
  Index<-IndexC[c(21,26,29,35),]
  rownames(Index)<- c("Silhouette", "Dunn", "Ch",
"Conectividad")

  # Matriz de Confusion
  MCo<-confusionMatrix(as.factor(V[,(k-1)]), as.factor(VA))
  a<-apply(MCo$table,2,which.max)
  a[duplicated(apply(MCo$table,2,which.max))]<-
seq(1,k)[!(1:k%in%a)]
  MC<-MCo$table[a,]

  # Proporcion de Mala Clasificación
  diag(MCo$table[a,])<-c(rep(0, ncol(MCo$table[a,])))
  PMA<-sum(MCo$table[a,])/length(VA) # Proporcion de mala
asignacion
  return(list(V, Index, MC, PMA))

}
UP<-mclapply(1:100, UPGMA, k=2, mc.cores = 8)
save(UP, file="/home/evidela/Metodos/E4/UP.RData")


# Función k-means
#################

KMEANS<-function(m, n=15, k, Simulaciones){
   Simulaciones<-list.files(path)
   Base<-read.table(paste(path, Simulaciones[m], sep="/"))[,-
c(1:2)]   # Le saco la columna de asignacion y la de fenotipo
   VA<-read.table(paste(path, Simulaciones[m], sep="/"))[,1]

   Base<-t(apply(Base, 1, Codif))
   Base[is.na(Base)]<-0
   D<-vegdist(as.matrix(Base), method="jaccard", binary=TRUE)

   # Metodo: Vector de Asignacion
   V<-matrix(c(rep(0,nrow(Base)*(n-1))),nrow=nrow(Base),ncol=(n-
1), dimnames = list(1:nrow(Base), c(paste0("k=", 2:n))))
   for (l in 2:n) V[,l-1]=as.vector((kmeans(Base,l)$cluster))

   # Indices
   Ind<-data.frame((matrix(NA, nrow=34, ncol=(n-1))))
   for (j in 1:(n-1)) Ind[,j] =
as.data.frame(as.matrix(cluster.stats(D,V[,j])))
   names(Ind)<-c(paste0("k=", 2:n))
   rownames(Ind)<-
rownames(as.data.frame(as.matrix(cluster.stats(D,V[,1]))))
   C=c()
   for(j in 1:(n-1)) C[j]=connectivity(D,V[,j])
   Index<-IndexC[c(21,26,29,35),]
  rownames(Index)<- c("Silhouette", "Dunn", "Ch",
"Conectividad")
```

```
   # Matriz de Confusion
   MCo<-confusionMatrix(as.factor(V[,(k-1)]), as.factor(VA))
   a<-apply(MCo$table,2,which.max)
   a[duplicated(apply(MCo$table,2,which.max))]<-
seq(1,k)[!(1:k%in%a)]
   MC<-MCo$table[a,]

   # Proporcion de Mala Clasificación
   diag(MCo$table[a,])<-c(rep(0, ncol(MCo$table[a,])))
   PMA<-sum(MCo$table[a,])/length(VA) # Proporcion de mala
asignacion
   return(list(V, Index, MC, PMA))

 }

 KM<-mclapply(1:100, KMEANS, k=2, mc.cores = 8)
 save(KM, file="/home/evidela/Metodos/E4/KM.RData")


# Función MBS
#############

MBS<-function(m, n=15, k, Simulaciones){
  Simulaciones<-list.files(path)
  Base<-read.table(paste(path, Simulaciones[m], sep="/"))[,-
c(1:2)]   # Le saco la columna de asignacion y la de fenotipo
  VA<-read.table(paste(path, Simulaciones[m], sep="/"))[,1]
  Base<-t(apply(Base, 1, Codif))
  Base[is.na(Base)]<-0
  D<-vegdist(as.matrix(Base), method="jaccard", binary=TRUE)

  colnames(Base) <- NULL
  rownames(Base) <- NULL

  write.geno(Base, paste0(paste(path2,"base",m,
sep=""),".geno"))

  best.k <- snmf(input.file = paste0(paste(path2,"base",m,
sep=""),".geno"), K = 2:n, project = "force", entropy = T)

  Qlist<-list(0)
  for (r in 1:(n-1)) Qlist[[r]]<-Q(best.k, K=(r+1), run=1)

  # Metodo: Vector de Asignacion
  V<-matrix(c(rep(0,nrow(Base)*(n-1))),nrow=nrow(Base),ncol=(n-
1), dimnames = list(1:nrow(Base), c(paste0("k=", 2:n))))
  for (l in 2:n) V[,l-1]=as.vector(apply(Qlist[[l-1]], 1,
function(x)which.max(x)))

  # Indices
  Ind<-data.frame((matrix(NA, nrow=34, ncol=(n-1))))
  for (j in 1:(n-1)) Ind[,j] =
as.data.frame(as.matrix(cluster.stats(D,V[,j])))
  names(Ind)<-c(paste0("k=", 2:n))
```

```
  rownames(Ind)<-
rownames(as.data.frame(as.matrix(cluster.stats(D,V[,1]))))
  C=c()
  for(j in 1:(n-1)) C[j]=connectivity(D,V[,j])
  IndexC<-rbind(Ind, Conectividad=C)
  Index<-IndexC[c(21,26,29,35),]
  rownames(Index)<- c("Silhouette", "Dunn", "Ch",
"Conectividad")


  # Matriz de Confusion
  MCo<-confusionMatrix(as.factor(V[,(k-1)]), as.factor(VA))
  a<-apply(MCo$table,2,which.max)
  a[duplicated(apply(MCo$table,2,which.max))]<-
seq(1,k)[!(1:k%in%a)]
  MC<-MCo$table[a,]

  # Proporcion de Mala Asignacion
  diag(MCo$table[a,])<-c(rep(0, ncol(MCo$table[a,])))
  PMA<-sum(MCo$table[a,])/length(VA) # Proporcion de mala
asignacion
  return(list(Qlist, V, Index, MC, PMA))

}

St<-mclapply(1:50, MBS, k=2, mc.cores = 8)
save(St, file="/home/evidela/Metodos/E4/MBS.RData")
```