



Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Estadística Aplicada

**IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA
VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA**

Ing. Agr. Joel Estuardo Morales Lemus
Asesorado por el Dr. José Luis Quemé De León

Guatemala, noviembre de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



FACULTAD DE INGENIERÍA

**IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA
VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA**

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

Ing. Agr. Joel Estuardo Morales Lemus
ASESORADO POR EL DR. JOSÉ LUIS QUEMÉ DE LEÓN

AL CONFERÍRSELE EL TÍTULO DE
MAESTRO EN ESTADÍSTICA APLICADA

GUATEMALA, NOVIEMBRE DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN DE DEFENSA DE TESIS

DECANA	Inga. Aurelia Anabela Córdoba Estrada
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADOR	Dra. Aura Marina Rodríguez Pérez
EXAMINADOR	Mtro. Ing. William Eduardo Fagiani Cruz
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha julio 2017.

A handwritten signature in black ink, consisting of stylized initials 'JEL'.

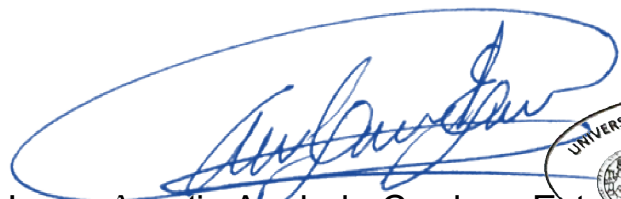
Ing. Agr. Joel Estuardo Morales Lemus

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.796.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA**, presentado por: **Joel Estuardo Morales Lemus**, que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:



Inga. Aurelia Anabela Cordova Estrada

Decana

Guatemala, noviembre de 2022

AACE/gaoc



Guatemala, noviembre de 2022

LNG.EEP.OI.796.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

“IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA”

presentado por **Joel Estuardo Morales Lemus** correspondiente al programa de **Maestría en artes en Estadística aplicada**; apruebo y autorizo el mismo.

Atentamente,

“Id y Enseñad a Todos”

Mtro. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería





Guatemala 12 de septiembre 2022.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **“IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA”** del estudiante **Joel Estuardo Morales Lemus** quien se identifica con número de carné **2756780910101** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,


MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 11 de agosto de 2022.

M.A. Ing. Edgar Darío Álvarez Cotí
Director Escuela de Estudios de Postgrado
Universidad de San Carlos de Guatemala

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: **"IMPLEMENTACIÓN DEL MODELO AMOVA, PARA DETERMINACIÓN DE LA VARIABILIDAD GENÉTICA DEL BANCO DE GERMOPLASMA DE CENGICAÑA."** del estudiante **Joel Estuardo Morales Lemus** del programa de Maestría en **Estadística Aplicada**, identificada con número de carné: **200515181**.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.



Dr. José Luis Quemé De León

Colegiado No. 636

Asesor de Tesis

ACTO QUE DEDICO A:

Dios

Al inicio del camino prometiste estar siempre a mí lado y en los momentos difíciles veo solo un par de huellas en el camino; ¡las tuyas al cargar mí peso!

Mis padres

Roberto Morales y Estela Lemus, quienes me han dado las mejores herramientas para la vida.

Mis hermanos

Azucena, Roberto, Luis y Judith Morales Lemus, por las experiencias compartidas y la seguridad de tener siempre el resguardo de una cálida familia.

AGRADECIMIENTOS A:

**Al pueblo de
Guatemala y
Universidad de San
Carlos de Guatemala**

Por haberme dado la herramienta más importante en mi vida profesional.

Facultad de Ingeniería

Por brindarme la oportunidad de especializarme.

A mi asesor

Dr. José Luis Quemé De León, por su incansable paciencia, apoyo y amistad.

CENGICAÑA

Institución que ha marcado un hito en mi vida y que bajo la dirección del Dr. Mario Melgar me ha formado; y el apoyo del Dr. Luis Molina y el área de Biotecnología de CENGICAÑA, que han hecho posible este trabajo de graduación.

**Mis amigos de la
Facultad**

Juan René Santizo, Walter Bardales, Antonio Ascencio.

A mis Profesores

Todos han aportado para forjar la persona que soy hoy.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES	III
LISTA DE SÍMBOLOS	V
GLOSARIO	VII
RESUMEN	XI
PLANTEAMIENTO DEL PROBLEMA.....	XIII
OBJETIVOS.....	XVII
RESUMEN DEL MARCO METODOLÓGICO	XIX
INTRODUCCIÓN	XXIII
1. MARCO REFERENCIAL	1
2. MARCO TEÓRICO.....	7
2.1. Fundamentos estadísticos del análisis molecular de la varianza (AMOVA).....	7
2.2. Análisis de conglomerados, distancia Dice, agrupación Ward y dendrograma.....	14
2.3. Índices de diversidad genética, heterocigosidad insesgada de Nei y comparación de medias de Friedman.	16
2.4. Bases de mejoramiento genético de caña de azúcar	17
2.5. Marcadores moleculares	18
3. PRESENTACIÓN DE RESULTADOS	21
3.1. Objetivo 1. Determinación de la distancia genética entre 375 variedades de la colección nacional, utilizando tres	

	marcadores moleculares, para comparar genéticamente las variedades.	21
3.2.	Objetivo 2. Determinación de la variabilidad entre los distintos grupos de variedades y dentro de ellos, por medio del análisis AMOVA, para definir la diversidad genética total de la colección nacional.	25
3.3.	Objetivo 3. Generación del procedimiento de análisis con AMOVA utilizando datos producto del análisis de marcadores moleculares del área de biotecnología, por medio del presente documento, para su uso en el programa de variedades de CENGICAÑA.....	26
3.4.	Objetivo General. Determinar la variabilidad genética entre y dentro de grupos de variedades utilizando el análisis AMOVA, para conocer si existe distancias genéticas significativas en la colección nacional.	27
4.	DISCUSIÓN DE RESULTADOS	29
4.1.	Análisis interno	29
4.2.	Análisis externo	30
	CONCLUSIONES	33
	RECOMENDACIONES.....	35
	REFERENCIAS	37
	APÉNDICE	41

ÍNDICE DE ILUSTRACIONES

FIGURAS

1. Ejemplo de una placa de gel donde se realizó el análisis de electroforesis y forma de base de datos generada..... 8
2. Esquema conceptual de análisis molecular de la varianza 13
3. Cálculo de distancias de variables cualitativa binaria..... 15
4. Dendrograma de variedades del banco de germoplasma 23

TABLAS

- I. Matriz de variables utilizadas en el estudioXX
- II. Número de variedades por grupo..... 21
- III. Estadísticas de análisis de la base de datos de marcadores moleculares..... 22
- IV. Comparación de medias de índices de diversidad genética y heterocigosidad insesgada de Nei. 24
- V. Análisis molecular de la varianza, utilizando distancia Dice 25
- VI. Coeficientes Phi del AMOVA..... 25

LISTA DE SÍMBOLOS

Símbolo	Significado
ADN	Ácido desoxirribonucleico
B	Barbados
CP	<i>Canal Point</i> , Florida, Estados Unidos
CP-L	<i>Canal Point-Lousiana</i>
CG	Cengicaña, Guatemala
Phi	Coeficiente de heterosis en análisis AMOVA
CN	Colección nacional (variedades de azúcar)
%Pol	Contenido porcentual de sacarosa en la caña
CM	Cuadrados medios
F.V.	Fuente de variación
gl	Grados de libertad
Ho	Hawaii
Mex	México
p-valor	Porcentaje de significancia estadística
POJ	<i>Proefstation oost</i> , Java
PR	Puerto Rico
SC	Suma de cuadrados
SP	<i>São Paulo</i> , Brasil
UPGMA	<i>Unweighted pair group method with arithmetic mean</i>

GLOSARIO

AMOVA	Método estadístico llamado análisis molecular de la varianza.
Banco de germoplasma	Nombre al que se le otorga al conjunto de variedades que se resguardan, colectadas de diversidad de destinos, con la finalidad de preservar la variabilidad genética.
CENGICAÑA	Institución llamada centro guatemalteco de capacitación e investigación de la caña de azúcar.
Colección nacional	Nombre otorgado al banco de germoplasma de variedades de azúcar de Guatemala, creada por CENGICAÑA.
Dendrograma	Es un gráfico que ordena de forma jerárquica, semejante a un árbol, individuos en grupos, asociado a análisis multivariado de conglomerados.
Distancia euclidiana	Es una herramienta utilizada en análisis multivariado, que relaciona en línea recta un grupo o grupos de observaciones con un origen.
Frecuencia alélica	Se refiere al número de ocurrencias observadas de la forma variante de un gen, en el mismo lugar en los

cromosomas, cuya expresión puede causar variantes fenotípicas, como el color de los ojos.

Gen	Es la unidad básica (secuencia de ADN) que permite la transferencia de características hereditarias a la progenie.
Heterocigosidad	Hace referencia a la variabilidad (diferencia) de los alelos en la combinación genética de los progenitores.
Heterosis	En mejoramiento genético, se refiere a la selección de dos parentales con diferencias en los caracteres de interés, con la finalidad de expresar en la progenie una expresión superior, también llamado vigor híbrido.
InfoGen	Software estadístico, especializado en mejoramiento de plantas.
Ley <i>Hardy-Weinberg</i>	Se refiere al comportamiento de la frecuencia alélica en donde la población diploide y reproducción sexual, no cambia, manteniéndose en equilibrio, no existiendo una variabilidad genética amplia.
Marcador molecular	Se refiere a un segmento de ADN (de un gen) que sirve como punto de referencia para el análisis del genoma de un organismo.

Microsatélite	Es un segmento corto de ADN con una longitud de uno a seis pares de bases, que se utilizan como marcadores moleculares.
Polimorfismo	Se refiere a las muchas variantes que puede expresar una sección de ADN.
Raya roja	Enfermedad bacteriana en caña de azúcar causada por el agente <i>Acidovorax avenae</i> subsp. <i>Avenae</i> (Manns) Willems = <i>Pseudomonas rubrilineans</i> (Lee et al., 1925) Stapp.
Retrocruza	Se refiere a la cruce de un híbrido (F1) con uno de los padres o genotipo idéntico al paterno.
Significancia	En referencia a la estadística, se asocia al nivel de certeza cuando se evalúa una hipótesis, generando confiabilidad de que el resultado no se debe al azar.

RESUMEN

El propósito de este trabajo fue, brindar a CENGICAÑA un análisis de la variabilidad genética de su banco de germoplasma, como base de entendimiento de su conformación.

El objetivo general consistió en determinar la variabilidad genética entre y dentro de los grupos de país de origen de importación de variedades de caña de azúcar, utilizando un AMOVA, para conocer si existen distancias genéticas significativas en la colección nacional de caña de azúcar. Lo anterior como base para conocer las diferencias y similitudes dentro de las variedades importadas de distintos países de origen y las diferencias genéticas de las variedades de dichos países.

La investigación empleó un enfoque cuantitativo, con un diseño no experimental, de tipo transeccional y alcance exploratorio.

Se tomó la base de datos del análisis molecular de 375 variedades del banco de germoplasma de CENGICAÑA, utilizando tres marcadores moleculares. Las variedades se agruparon por origen o serie de selección de la variedad.

El principal resultado determinó que hay diferencias significativas entre grupos y dentro de los grupos, pero la variación entre grupos es muy baja, detectando más variación dentro de los grupos.

Se concluyó que el grupo de variedades o materiales CP (*Canal Point*) cuentan con la mayor variabilidad dentro del grupo, respecto a los otros grupos de origen.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

El programa de variedades de CENGICAÑA, tiene como objetivo generar nuevas variedades de caña de azúcar, adaptadas a las condiciones locales, tolerantes a enfermedades y de alta producción, para lo cual genera una estrategia de mejoramiento, Quemé, Orozco y Salazar. (2016a) destacan cuatro procesos: a) recurso genético (germoplasma) b) cruzamientos c) selección y d) desarrollo comercial.

El recurso genético o germoplasma empleado por CENGICAÑA para definir los progenitores en los cruzamientos, es llamada colección nacional (CN), la cual está compuesta por 2,575 variedades de 23 países distintos. El objetivo de los cruzamientos es generar progenie superior en características deseables para la AIAG: contenido de azúcar (% Pol), rendimiento de biomasa (toneladas de caña por hectárea), tolerancia a enfermedades y plagas, características agronómicas y otras, Quemé, Orozco y Salazar (2016b) al citar a Cox *et al.* (2000), destaca que es necesario contar con variabilidad genética del germoplasma, pues lo anterior permite la selección de parentales que generan progenie de características superiores, pero también se debe considerar que el contenido de azúcar no depende de un solo gen, por lo que la estrategia de selección de parentales, se basa en seleccionar progenitores de alto contenido de azúcar y garantizar variabilidad genética entre estos para tener mayor probabilidad de éxito de generar progenie de mayor contenido de azúcar que los progenitores (Quemé, Orozco y Salazar, 2016b).

- Descripción del problema

La colección nacional no estaba caracterizada molecularmente, respecto a su similitud genética entre grupos de variedades del mismo país y dentro de estos países, a pesar de existir información genética de un buen número de las variedades presentes en la misma. Por lo que se perdía información importante, que ayuda al proceso de recurso genético de CENGICAÑA. Por lo que el AMOVA, indica la variabilidad que existe entre los grupos de variedades de los diversos países, presente en la colección nacional y la variabilidad dentro del mismo grupo, generando información útil al proceso de recurso genético, empleando dicha información como un criterio más para la caracterización de los parentales utilizados en las cruzas, favoreciendo a la mayor certeza de generar progenie superior, en términos de azúcar, producción de biomasa, tolerancia a enfermedades y otros.

El programa de variedades de CENGICAÑA, cuenta con un área de biotecnología, que puede caracterizar genéticamente las variedades de caña de azúcar en base a tres marcadores moleculares, Molina *et al.* (2016) generando información cualitativa, con variables binarias, determino índices de similitud genética, pero sin hacer análisis de significancia entre las distancias genéticas, por lo que se definió el procedimiento de análisis.

- Formulación del problema

- Pregunta central con la que se trabajó fue:

- ¿Cuál es la variabilidad genética entre las variedades de distintos países y dentro de los grupos (países) en la colección nacional de CENGICAÑA?
- Preguntas auxiliares

Para esto también se respondió las preguntas auxiliares:

- ¿Cuál es la significancia de las distancias genéticas entre 375 variedades de la colección nacional, basados en tres marcadores moleculares?
- ¿Cuál es la variabilidad genética dentro de los grupos de variedades del mismo país de origen y cuál será la variabilidad genética entre países de origen, para entender la variabilidad genética total de la colección nacional?
- ¿Cómo replicar el procedimiento al contar el laboratorio de biotecnología con información de otras variedades del banco de germoplasma?
- Delimitación del problema

El uso de técnicas estadísticas como la AMOVA, permitió conocer la significancia de las distancias genéticas de las variedades del germoplasma de CENGICAÑA, con dos finalidades: estimar la variabilidad genética utilizada en las cruza y tener un criterio más para definir los lugares, que se deben de priorizar para importar variedades. Lo que ayuda al programa de variedades, en el proceso de análisis e interpretación de la información generada por el área de

biotecnología. Para lo cual se utilizó un grupo de 375 variedades de la colección nacional y 3 marcadores moleculares.

La problemática que se abordó consiste en determinar si existe variabilidad genética entre el grupo de variedades de distintos países de origen y dentro de cada grupo, con los que cuenta la colección nacional, tomando como insumo la matriz binaria, que expresa la ausencia o presencia de segmentos de secuencias conocidas, llamados marcadores moleculares. Dichas variables son de tipo cualitativo generando bases de datos grandes, permitiendo estimar índices de distancias genéticas, que posteriormente sirvieron para realizar análisis AMOVA, para definir qué país de origen es el que aporta mayor variabilidad a la colección nacional, y que permite priorizar la importación de dichos lugares y también uso de estas variedades para sus cruas.

OBJETIVOS

General

Determinar la variabilidad genética entre y dentro de grupos de variedades, utilizando un análisis AMOVA, para conocer si existe distancias genéticas significativas en la colección nacional.

Específicos

1. Determinar la distancia genética entre 375 variedades de la colección nacional, utilizando tres marcadores moleculares, para comparar genéticamente las variedades.
2. Determinar la variabilidad entre los distintos grupos de variedades y dentro de ellos, por medio del análisis AMOVA, para definir la diversidad genética total de la colección nacional.
3. Generar el procedimiento de análisis con AMOVA con datos producto del análisis de marcadores moleculares del área de biotecnología, por medio del presente documento, para su uso en el programa de variedades de CENGICAÑA.

RESUMEN DEL MARCO METODOLÓGICO

El enfoque utilizado fue cuantitativo y el diseño fue no experimental de tipo transeccional y alcance exploratorio.

- Características de estudio

Se realizaron mediciones de distancias genéticas con una variable cualitativa binaria, por lo que el enfoque de la investigación fue cuantitativo, al obtener índices de similitud genética y realizar agrupaciones respecto al mismo.

El alcance es exploratorio, dado que se caracterizó la colección nacional de CENGICAÑA, respecto a su similitud genética, discriminando dentro y entre grupos la diversidad genética, considerando de mayor diversidad el grupo que más distancias genéticas presente entre sus individuos.

La información genética y la variabilidad se analizó empleando la base de datos del laboratorio de biotecnología de CENGICAÑA, el cual caracterizó la ausencia o presencia de marcadores moleculares, en una matriz binaria, haciendo que el diseño utilizado sea no experimental, además fue transversal, pues se estudió la diversidad genética de un grupo aleatorio permitiendo hacer inferencia en la totalidad de la colección, pero a medida que la misma evolucione será necesario realizar nuevos análisis.

- Unidad de análisis

La población de estudio fue la colección nacional de CENGICAÑA que cuenta con 2,575 variedades de 23 países distintos, misma que está dividida en subpoblaciones dadas por el país de origen o el programa de cruzamientos que generó dicha variedad, de esta se tomó una muestra de 375 variedades de los principales programas de los cuales se importan variedades, que fueron estudiadas en su totalidad.

- Variables: a continuación, se describen las variables utilizadas en el estudio:

Tabla I. **Matriz de variables utilizadas en el estudio**

Nombre Variable	Definición teórica	Definición operativa
Polimorfismo	Presencia de la banda genética comparada con el microsatélite en la variedad analizada	0 = Ausente 1 = Presente
Índice de similitud genética	Distancia genética entre variedades, comparando la similitud respecto a polimorfismos y semejanzas con otras variedades	Porcentaje

Fuente: elaboración propia.

- Fases

Las fases empleadas en el estudio son descritas de la siguiente manera:

- Técnicas metodológicas

- Fase 1. Obtención de la base de datos

Se utilizó la base de datos generada por el laboratorio de biotecnología de CENGICAÑA, quienes utilizaron 3 marcadores moleculares, en 375 variedades del banco de germoplasma de CENGICAÑA, ya descrita de forma binaria describiendo presencia o ausencia del gen correspondiente al marcador. Dicha base de datos ya depurada se analizó con *InfoGen* y se analizó la información para redactar el trabajo final, donde en base a las significancias se definió si hay variación genética entre y dentro de los grupos de variedades o países de origen.

- Fase 2. Comparación de índices de distancia

Se realizaron los análisis AMOVA con distintos índices de distancia genética, para validar que éste no es un factor que cause diferencia en los resultados del análisis AMOVA, determinando la diversidad genética total de la colección nacional, pudiendo agruparlo entre grupos y observar la variabilidad entre y dentro de los mismos.

- Fase 3. Discusión e integración del documento final

Se discutieron los resultados en función de priorizar el origen de la importación de variedades para la colección nacional, integrando el documento final, el cual servirá como guía de la sistematización del procesamiento de análisis de datos para marcadores moleculares del área de biotecnología de CENGICAÑA.

- Técnica de análisis de la información

Se utilizó un modelo jerárquico, para el análisis de la información, denominado análisis molecular de la varianza. Se utilizaron los valores de distancia genética para realizar dicho análisis, lo cual dio la significancia entre grupos y dentro de grupos:

- Índices distancia genética

Los índices de distancia genética consisten en estimar la similitud entre genotipos basados en la presencia o ausencia de distintas bandas asociadas a genes obtenidos de análisis de electroforesis. Existen aproximadamente 10 índices distintos para estimar la distancia genética, por lo que se evaluaron todos para identificar diferencias.

- AMOVA

Tomando los índices de distancia se hizo un análisis de AMOVA, para determinar la significancia de las distancias, y se interpretó si estadísticamente existe diferencia real entre los genotipos.

Para esto se utilizó una base de datos inicial binaria, que indica ausencia o presencia de gen, y estos datos se utilizaron para estimar las distancias genéticas mediante la aplicación del modelo AMOVA, con la herramienta *InfoGen*.

INTRODUCCIÓN

Este estudio correspondió a una sistematización, debido a que proporciona a los investigadores del programa de variedades de CENGICAÑA, un criterio para entender la composición general del banco de germoplasma, desde una visión genética, aportando más información, en la caracterización e identificación de grupos de variedades y las diferencias o distancias genéticas entre ellas.

La problemática surge al entender que la sostenibilidad de la industria depende de muchos factores, uno de ellos es poseer variedades nuevas de forma constante, que estén adaptadas al ambiente, en el contexto del cambio climático. Por lo que CENGICAÑA cuenta con un programa de variedades. El programa tiene un banco de germoplasma, llamado colección nacional (CN), el cual se analizó molecularmente, con los marcadores CV29, CV37 y CV38, permitiendo estimar la distancia genética entre variedades, lo cual es de importancia para la selección de progenitores, en el proceso de generación de nuevas variedades.

La importancia de la presente investigación se entiende como un aporte metodológico en el análisis de la variabilidad de la CN. Al entender que el objetivo de esta es ampliar la base genética, que permita contar con variabilidad en las cruza, y de esta manera tener progenie superior. Con base a lo anterior se importan constantemente variedades de programas extranjeros esperando que dichos materiales amplíen dicha variabilidad genética, por lo que utilizar metodologías como el AMOVA, permite conocer que tan distantes son las variedades de distintos programas de otros países, para priorizar la importación

de dichos destinos, y de igual forma que tanta variación aportan estas variedades a la colección nacional.

El esquema de la solución consistió en tomar la caracterización parcial, de la colección nacional con marcadores moleculares, lo que generó una gran cantidad de información de carácter binario, indicando presencia o ausencia del marcador en la variedad de caña. Esta información es analizada con índices de similitud, pero estas herramientas estadísticas no son capaces de indicar si las distancias genéticas entre variedades son significativas o no por lo que la implementación de herramientas como el AMOVA es una extensión del análisis de frecuencias genéticas de poblaciones, permite conocer si estas distancias son significativas, para asegurar que existe diferencias genéticas entre variedades del mismo grupo o país de importación y entre los países de importación.

Para realizar el AMOVA se utiliza una matriz amplia de distancias euclidianas entre todos los polimorfismos, generando un archivo que contiene la frecuencia de los haplotipos de las variedades empleadas, lo que permitió generar un análisis de la estructura genética de la muestra, al definir las distancias entre grupos y subgrupos. Distintos programas estadísticos especializados permiten realizar dicho análisis, entre ellos *InfoGen*, el cual es un software generado por la Universidad de Córdoba, Argentina, y que se utilizó en el estudio realizado.

La presente investigación contiene cuatro capítulos: capítulo 1, el cual inicia con una recopilación exhaustiva de literatura, que describe el origen de la caña de azúcar comercial y aplicaciones puntuales en mejoramiento de plantas, caña de azúcar y colección nacional de CENGICAÑA, ejemplificando la importancia de contar con nuevas variedades de caña y los problemas que algunas industrias han presentado al depender de una sola variedad. De igual

forma describiendo la importancia de la variabilidad genética en las cruzas para formar nuevas variedades y como la herramienta AMOVA fue creada con la finalidad de dar respuesta a la determinación de las diferencias entre grupos y a lo interno de los mismos, tomando como variable las caracterizaciones con marcadores moleculares.

El capítulo 2, consiste en el marco teórico, el cual está desglosado en conceptos básicos de genética en caña de azúcar, procesos de cruzamiento, uso de análisis molecular en caña de azúcar, índices de distancia genética y análisis molecular de la varianza.

El capítulo 3, presentación de resultados, considerando análisis de 375 variedades con el modelo AMOVA, comparación de análisis AMOVA utilizando distintos índices de distancia y la selección de grupos y variedades de mayor variabilidad.

Por último, el capítulo 4, discusión de resultados, encontrando diferencias significativas entre los grupos CP89, CP91, CP92, CP00, V2 Y CP97, y los grupos con mayor variación interna son CP89, CP91, CP92 y CP00, de donde se determinan que dichos grupos son los que poseen la mayor variabilidad genética o mayor riqueza dentro de la colección nacional de CENGICAÑA.

1. MARCO REFERENCIAL

- Estudios Previos

El Centro Guatemalteco de Investigación y Capacitación de la Caña de Azúcar (CENGICAÑA), es quien genera tecnología para la agroindustria azucarera de Guatemala, contribuyendo a su sustentabilidad. El programa más grande de CENGICAÑA, es el de variedades, el cual genera nuevos materiales genéticos adaptados a las condiciones locales, tolerantes a enfermedades y de mayor producción, incrementando la productividad global. Permitiendo entender en la presente investigación a profundidad la composición varietal de Guatemala y la variabilidad de la colección nacional (CN), respecto a la zafra 2015-16, Orozco y Buc (2016) indican que se cosecharon 268,735 hectáreas, identificando el 72.8 %, y encontraron que el 70 % del área está sembrada únicamente con cuatro variedades, por lo que el programa busca aumentar el número de materiales utilizados, generando nuevas cruzas que provienen de la CN, de 2,575 variedades de todo el mundo (banco de germoplasma).

Respecto al origen de las variedades modernas, Creste *et al.* (2009) indican que al menos se utilizaron dos especies de *Saccharum* en las cruzas de las primeras variedades de caña de azúcar, generando un antecedente en la actual investigación, de la estrecha variabilidad genética en caña de azúcar, al compartir orígenes comunes las variedades comerciales actuales. La progenie obtenida de dichas cruzas, fueron repetidamente retrocruzadas, con clones de *S. officinarum*, para recuperar los alelos favorables para el contenido de azúcar del padre recurrente, llamándose novilización de la caña de azúcar (Roach, 1972). A partir de ello se ha generado una base genética muy estrecha para el cultivo

(Rosales, Quemé y Melgar, 2011). A pesar de los avances propios, de cada programa de mejoramiento de caña de azúcar a nivel mundial, se ha llegado a una meseta en el contenido de azúcar, en las nuevas variedades de mayor rendimiento, Quemé, Orozco y Salazar (2016b), explican que es un reto para los procesos de mejoramiento generar progenitores de mayor contenido de azúcar, para lo cual medir si las distancias genéticas de la estructura de la colección nacional y si son significativa, genera una herramienta, que contribuye a dicho reto.

Existe una importancia latente en la generación constante de nuevas variedades de azúcar, Rosales *et al.* (2011), menciona que las principales industrias azucareras del mundo, han llegado a depender de una variedad (más del 40 % del área sembrada con una sola variedad), las cuales dejan de ser utilizadas por problemas de susceptibilidad a enfermedades, colocando en riesgo la productividad de sus industrias, por lo que dicho autor concluye, que las industrias deben diversificar su composición varietal (10 a 12 variedades), para minimizar el riesgo de ser afectados por problemas de enfermedades que mermen la producción, por lo que asociar o utilizar progenitores con distancias genéticas significativamente distantes a las variedades susceptibles a las enfermedades, es un aporte de utilizar herramientas como la AMOVA, como se busca esta investigación.

La selección de parentales para realizar las cruzas en la actualidad se basa en la selección recurrente de materiales con fenotipos deseados, pero sin mucha información genética de los mismos. Existe más probabilidad de generar una nueva variedad de alto tonelaje y contenido de azúcar, si se cruzan parentales de alto contenido de azúcar (fenotipo) pero que entre ellos exista variabilidad genética (heterosis). Dicha información, se genera por el laboratorio

de biotecnología de CENGICAÑA, pero aún no se ha establecido la metodología de análisis, para que sea utilizado como un insumo, en la selección de parentales.

La AMOVA fue propuesta para diseñar una metodología alternativa que haga uso de la información molecular disponible proporcionada en estudios de poblaciones, Mengoni y Bazzicalupo (2002) citan en su trabajo a Excoffier *et al.* (1992) indicando también que la primera vez que se aplicó el método AMOVA se empleó para analizar poliformismos en el ADN mitocondrial humano. Posteriormente, se utilizó para el estudio de problemas ecológicos de plantas y animales. Este método tuvo gran impacto en los estudios moleculares, pues los mismos se caracterizan por tener un número elevado de muestras o individuos estudiados y de igual forma un número elevado de análisis moleculares, generando una gran base de datos. Este método describe los factores que influyen en la estructura de las poblaciones, con una gran aplicación en la agricultura y en los procesos de mejoramiento vegetal, generando la base metodológica del desarrollo de esta investigación y su aplicabilidad en la caracterización de la colección nacional.

En la caracterización de 100 variedades tomadas de la colección nacional de CENGICAÑA, Molina, Madaleno, Sut y Camargo. (2015) indican que se utilizó tres marcadores de secuencia simple, obteniendo: frecuencia alélica, porcentaje de loci plimórficos, número de alelos efectivos, contenido de información polimórfica y un dendrograma estimando la similitud genética, con base al grado de asociación utilizando el índice Dice y realizando la agrupación con el método UPGMA. Pero aún no se ha utilizado dicha información para realizar un AMOVA, y de esta forma estimar la diversidad total, y la diferencia estadística entre poblaciones y dentro de ellas. Lo que sirve como un ejemplo de aplicación de la metodología en cultivos con amplia variabilidad, y con la misma finalidad de este estudio, al tomar los datos del citado trabajo en la actual investigación.

El Instituto de Ciencia y Tecnología Agrícola (ICTA, Guatemala), ha realizado trabajos semejantes para la creación de una colección núcleo en maíz, Molina, Ponciano y Fuentes (2005) utilizaron diez marcadores moleculares en maíz, para caracterizar 731 variedades de 23 departamentos, para lo que utilizaron un análisis de AMOVA, encontrando que no existía variabilidad genética entre departamentos, pero sí existía alta variabilidad genética dentro de los departamentos. Por lo que los materiales a nivel nacional están ampliamente dispersos en todo el país, pero cada departamento tiene una alta diversidad genética. Este trabajo aporta los conceptos, que se implementó en el análisis de la variabilidad de la colección nacional de CENGICAÑA.

En Brasil, analizaron 53 variedades de caña de azúcar, de los tres distintos programas de mejoramiento en dicho país, utilizaron siete marcadores asociados al metabolismo de la acumulación de la sacarosa y tres asociados a la tolerancia a sequía y tres más arbitrarios (Creste *et al.*, 2009). Para lo cual emplearon el índice Jaccard para definir el nivel de asociación y definir la similitud genética. Posteriormente, realizando un análisis de AMOVA se determinó que existe mucha similitud entre programas y que la variabilidad en los genes evaluados fue menor en los genes del metabolismo de la sacarosa que en los del metabolismo de la respuesta a la sequía. En este trabajo se utiliza un índice distinto al planteado en trabajos semejantes, concluyendo que no hay mucha diferencia con el planteado por Molina *et al.* (2016) permitiendo emplear el mismo algoritmo de análisis en la actual investigación, con la diferenciación del uso del índice Dice.

En la India se realizó un estudio sobre 30 variedades de caña de azúcar, Singh, Khan, Pandey, Kumar y Lal (2011) encontraron variedades susceptibles y tolerantes a la enfermedad raya roja o podrición del tallo, con marcadores de secuencia simple. El contenido de información de polimorfismo (PIC) varió de 0.216 a 0.813 con un promedio de 0.525. Los valores más altos (0.86) de la

proporción de loci polimórficos y la heterocigosidad esperada se obtuvieron en la población susceptible y en la población moderadamente tolerantes, respectivamente. Sin embargo, el AMOVA mostró variación entre los grupos. Los resultados sirven como referencia comparativa en la presente investigación, respecto a las distancias genéticas promedio y las de los individuos más distantes entre ellos. Los resistentes ISH150 y SES594 resultaron ser los más distintos, mientras que el resto de los genotipos podrían agruparse en dos grandes grupos que separan los moderadamente resistentes y susceptibles. Por lo tanto, la información que generaron se puede utilizar en el manejo de recursos genéticos, seleccionando parentales con resistencia a la raya roja, y mapear la distribución varietal de forma geográfica, respecto a susceptibilidad a la enfermedad. Lo que aporta a la metodología respecto a la interpretación de los resultados del análisis, y presenta nuevas aplicaciones posibles a la herramienta, sirviendo como guía metodológica e interpretativa en la actual investigación.

Utilizando diez marcadores moleculares, algunos asociados a acumulación de sacarosa, tolerancia al frío, desarrollo de tricomas y cuatro marcadores arbitrarios, Arro (2005) estudió 63 variedades de caña de azúcar de los dos programas de mejoramiento de Estados Unidos (*Canal Point*, y *Luisiana*). Se encontraron similitudes genéticas que van de 0.78 a 0.94, se encontró una falta de correlación entre los marcadores y la expresión de las características asociadas. El AMOVA, no definió diferencias entre poblaciones de las variedades liberadas antes y después de la década de 1980, la variabilidad se dio entre variedades (0.99) y no entre grupos. Dichos resultados mencionados anteriormente, permite generar expectativas respecto a la variabilidad genética de la colección nacional en la presente investigación, debido a que dicha colección posee un gran número de variedades importadas de ambas localidades.

2. MARCO TEÓRICO

2.1. Fundamentos estadísticos del análisis molecular de la varianza (AMOVA)

La cuantificación de la relación entre individuos y los índices de similitud fundamenta la determinación de la diversidad genética y su distribución entre y dentro de las poblaciones de una determinada especie.

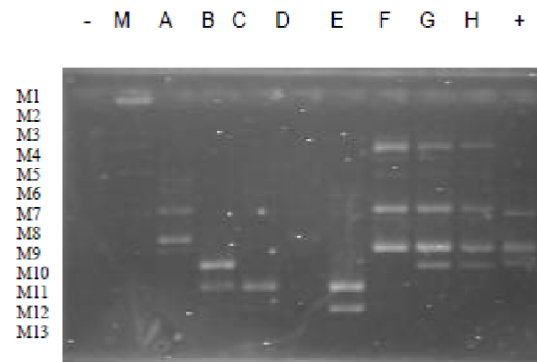
El uso de marcadores moleculares genera información, la cual es una variable cualitativa binaria (1=presencia y 0=ausencia), donde se expresa la presencia del segmento o polimorfismo del individuo con el marcador, lo anterior genera un gran volumen de datos que pueden ser analizados por medio de herramientas multivariadas, como se observa en la figura 1.

Cuando las poblaciones están subdivididas o partidas en individuos, especies o grupos específicos se espera que se incremente la diversidad (Banzarini, Arroyo, Bruno y Di Rienzo, 2006), recomendando el uso de modelos jerárquicos que en este caso se denomina análisis molecular de la varianza (AMOVA). Dicho análisis permite determinar la significancia de las distancias genéticas entre grupos y dentro de ellos.

Cuando en una población hay algún tipo de partición o subdivisión, se espera que esta cree una diversidad genómica o se aumente distancia genética en dichos subgrupos (Slatkin, 1987). Esta estructura genómica de las poblaciones y la conformación de subgrupos se ha estudiado por medio de

frecuencias alélicas, basados en la ley de *Hardy-Weinberg*, que se basa en la comparación de lo observado con lo esperado.

Figura 1. **Ejemplo de una placa de gel donde se realizó el análisis de electroforesis y forma de base de datos generada**



Muestras	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
A	0	0	0	0	1	1	1	0	1	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	1	1	0
C	0	0	0	0	0	0	0	0	0	0	0	1	1
D	0	0	1	0	1	1	1	1	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	1
F	1	1	0	1	0	0	1	0	0	1	0	0	0
G	1	0	0	0	0	0	1	0	0	1	1	0	0
H	1	0	0	0	0	0	1	0	0	1	1	0	0

Fuente: Balzarini, Arroyo, Bruno y Di Rienzo (2006). *Análisis datos de marcadores con info-Gen*.

El análisis molecular de la varianza se fundamenta en la característica propia de la información molecular, que consiste en la suma de cuadrados de matrices de distancias entre pares y no entre variables como se estima normalmente (Excoffier *et al.*, 1992). Este método es útil para estimar la variabilidad de datos binarios y se conoce la influencia de los factores de clasificación, en el caso del estudio, se basa en los países de importación de las variedades.

Dicho modelo es jerárquico o anidado al existir más de un factor de clasificación interno del grupo general. La significancia se consigue realizando pruebas de permutaciones, esto evita el cumplir con los supuestos de normalidad exigido por el análisis de varianza clásico.

Al proponer el análisis de AMOVA, Excoffier *et al.* (1992) menciona que la estructura genética de la población dentro de una especie se ha estudiado tradicionalmente utilizando las desviaciones de las frecuencias alélicas. Se han propuesto varios procedimientos de estimación relacionados con los estadísticos F de Wright (Wright, 1951) para el tratamiento de sistemas polimórficos. Los métodos empleados, implican la transformación no lineal del conjunto de datos original en estimaciones de diversidad genética. Por lo que Excoffier indica la necesidad de una metodología más general que no dependa tan críticamente de los supuestos específicos, con la siguiente característica: debe de ser alternativa que haga uso de la información molecular disponible recopilada en las encuestas de población, sin dejar de ser lo suficientemente flexible para acomodar diferentes tipos de supuestos sobre la evolución del sistema genético.

El autor del AMOVA, indica que es un análisis jerárquico de la varianza molecular directamente a partir de la matriz de distancias al cuadrado entre todos los pares de haplotipos. Y tienen una relación con un análisis de varianza convencional, el AMOVA tiene la ventaja adicional de que se pueden imponer varios supuestos diferentes en el proceso de diferenciación de haplotipos, cada uno de los cuales se traduce en una matriz de distancia diferente, sin cambios en la estructura del análisis posterior.

- Supuestos

- El análisis es anidado de la varianza molecular. El modelo jerárquico emplea componentes de diversidad dentro de poblaciones, entre poblaciones o dentro de grupos y entre grupos.
- Se usan procedimientos de permutación en la matriz de distancia al cuadrado interindividual original para proporcionar pruebas de significación para cada uno de los componentes de la varianza jerárquica y los análogos estadísticos F relacionados.
- Bajo la hipótesis nula, las muestras se consideran extraídas de una población global, con variación debida al muestreo aleatorio en la construcción de poblaciones.
- Los componentes de la varianza se estiman a partir de un gran número (por ejemplo, 1000) de matrices permutadas. Usamos este procedimiento para obtener la distribución nula

En el análisis AMOVA, cada población o criterio de clasificación se define como factor y cada subgrupo se determina como nivel del factor, Balzarini *et al.* (2006) por lo anterior describen a la población como factor A y al subgrupo como factor B, lo que buscamos con el AMOVA es la variabilidad en A y B, y cuando B esta anidado en A. Si todas las observaciones o individuos con el mismo valor de B tienen el mismo valor de A, significa que cualquier variación en A representa una variación en B, por lo que A es marginal de B. Para cada observación se realizarán n muestras (marcadores), generando una observación multivariada con valores binarios (1 o 0), y se puede resumir en un vector:

Vector booleano (s-dimensionales) como:

$$p' = [p_1, p_2, \dots, p_s] \quad (\text{Ecuación 1})$$

Donde p_i $s=1, \dots, n$ expresa si esta la banda ausente (0) o presente (1). Y las diferencias entre dos muestras se definen por $p'_1 - p'_2$. Las diferencias entre dos muestras o dos grupos son definidas por $p_j - p_k$ utilizando una distancia euclidiana:

La distancia entre las muestras observadas se define por:

$$d^2 = (p_j - p_k)'W(p_j - p_k) \quad (\text{Ecuación 2})$$

Donde W es la matriz de valores binomiales, obteniendo una matriz de distancias euclidianas, que se pueden agrupar de forma jerárquica, para comparar la magnitud de las distancias de los distintos niveles del factor. Se realiza la suma de las distancias al cuadrado de cada observación N , lo anterior es igual a la suma de cuadrados de la desviación entre cada valor de vector con cada centroide del espacio multidimensional.

$$SCT = \frac{1}{2N} \sum_{j=1}^N \sum_{l=1}^N d_{jk}^2 \quad (\text{Ecuación 3})$$

El modelo analizado es considerando un nivel jerárquico es:

$$p_{ikj} = p + A_i + A > B_{k[i]} + \omega_{ikj} \quad (\text{Ecuación 4})$$

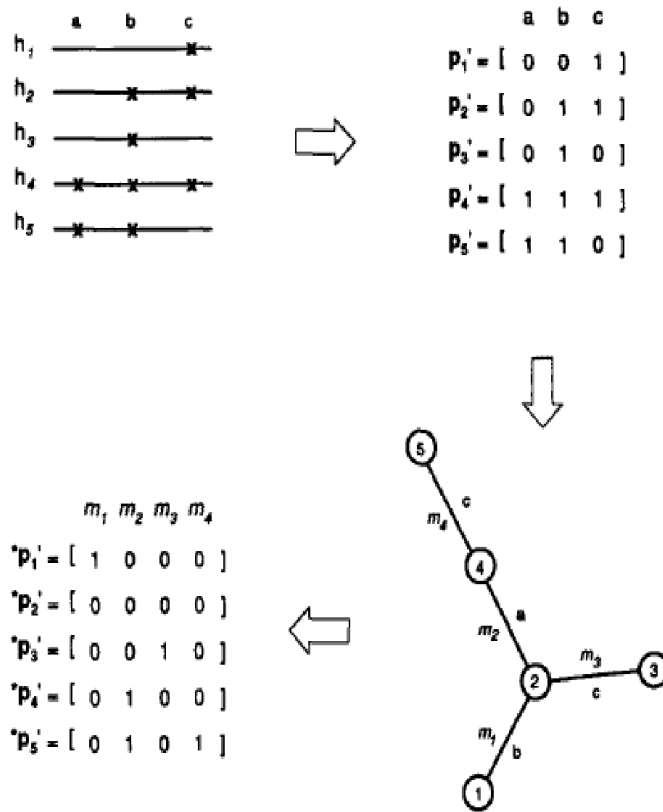
Donde p_{ikj} es la posición entre los niveles de la población (A) y el subgrupo anidado (B).

Los efectos de los factores A, A>B y del elemento individual se asumen aditivos, aleatorios, no correlacionados y distribuidos con componentes de varianza iguales a $2 \sigma_A$, $2 \sigma_{A \times B}$ y $2 \omega_{ijk} \sigma$, respectivamente (Balzarini *et al.*, 2006). Para cualquier partición jerárquica de los N elementos individuales ocasionados por un factor A es posible escribir la suma de cuadrados totales como:

$$SCT = SC_{\text{Entre clases de A}} + SC_{\text{Dentro de clases de A}} \quad (\text{Ecuación 5})$$

La suma de cuadrados asociada a un factor anidado se obtiene adicionando la suma de cuadrados marginal para ese factor y la suma de cuadrados asociada a cada combinación de los dos factores. Las desviaciones (cuadradas) promedio son obtenidas dividiendo cada suma de cuadrados por los grados de libertad apropiados (Excoffier *et al.*, 1992).

Figura 2. Esquema conceptual de análisis molecular de la varianza



Fuente: Excoffier *et al.* (1992). AMOVA (análisis of Molecular Variance).

En la ecuación 4, Se utiliza una matriz análoga de las correlaciones de las distancias de los haplotipos, para generar el índice llamado estadístico-Phi (ϕ) por. (Excoffier *et al.*, 1992)

$$\sigma_{\omega ikj}^2 = (1 - \text{Phi_ST})\sigma^2 \quad (\text{Ecuación 6a})$$

$$\sigma_{A>B_{k[i]}}^2 = (\text{Phi_ST} - \text{Phi_Grupo})\sigma^2 \quad (\text{Ecuación 6b})$$

$$\sigma_{A_i}^2 = \text{Phi_Grupo} \sigma^2 \quad (\text{Ecuación 6c})$$

Cuando la varianza total es igual a la varianzas A mas A >B y más ω_{ikj} en la ecuación 4, Φ_{ST} se considera como la correlación de los haplotipos extraídos de total de las especies y Φ_{Grupo} como la correlación de los haplotipos aleatorios dentro de cada grupo de la población, en relación con la de pares aleatorios de haplotipos extraídos de todas las especies y Φ_{SC} como la correlación de la diversidad molecular de los haplotipos aleatorios dentro de la población, en relación con la de los pares aleatorios de los haplotipos extraídos de cada región , por lo que se re escribe la ecuación 6 en términos de Φ .

$$\Phi_{ST} = \frac{\sigma_{A_i}^2 + \sigma_{A>B_{k[i]}}^2}{\sigma^2}, \Phi_{Grupo} = \frac{\sigma_{A_i}^2}{\sigma^2}, \Phi_{SC} = \frac{\sigma_{A>B_{k[i]}}^2}{\sigma_{A>B_{k[i]}}^2 + \sigma_{\omega_{ikj}}^2} \quad (\text{Ecuación 7})$$

Indica que es necesaria la debida cautela al interpretar estos coeficientes, pero, no obstante, Excoffier *et al.* (1992). Se pueden indicar como resúmenes convenientes del empaquetamiento de la información genética dentro y entre las poblaciones, siendo uno por uno con los componentes de la varianza.

2.2. Análisis de conglomerados, distancia Dice, agrupación Ward y dendrograma

El método de agrupación o análisis de conglomerados es un método exploratorio de datos que se utiliza para entender la estructura y agrupación de los individuos de una población y puede ser descrito como un algoritmo de clasificación jerárquico. Este método permite generar categorías o grupos, maximizando su similitud, pero perdiendo información de forma inevitable.

El algoritmo genera niveles de agrupación que dependen de la medida de distancia empleada y la selección de las variables analizadas. En este análisis se construye una matriz $n \times p$, donde p son las variables en cada n individuo.

Posteriormente se genera una matriz de distancias $n \times n$ donde el i,j -ésimo elemento es la distancia entre los pares de individuos. El método de medición de la similitud o distancia depende de distintas características, por ejemplo, si la variable es cualitativa binaria, como en el caso de la ausencia o presencia de un gen (0= ausencia 1=presencia), expresándose dicha matriz de la siguiente forma.

Figura 3. **Cálculo de distancias de variables cualitativa binaria**

		Individuo j		
		Presencia (1)	Ausencia (0)	
Individuo i	Presencia (1)	a	b	a+b
	Ausencia (0)	c	d	c+d
		a+c	b+d	a+b+c+d

Fuente: elaboración propia.

En la figura 3, podemos definir a , como el número de caracteres comunes, b el número de caracteres presentes en el individuo i pero ausentes en el individuo j , c los ausentes en el individuo i pero presentes en el individuo j y por último d los ausentes en ambos individuos.

Existen varias medidas de distancia o similitud dentro de la cual se encuentra la de Dice, la cual está dada por la expresión $\frac{2a}{2a+b+c}$, basado en la figura 3, que se caracteriza por no verificar las propiedades de simetría, emplear variable binaria, la matriz de distancias es semidefinida positiva o mayor a cero, y las distancia d_{ij} es Euclídea.

Las distancias entre pares permiten posteriormente, como parte del algoritmo de conglomerados, generar agrupaciones. Un método semejante al método de centroide es el *Ward* o método de mínima varianza, el cual se

característica, por hacer ponderaciones respecto al tamaño de cada grupo, por el cual está constituida la población. Este método lo podemos describir como la distancia entre dos grupos como la suma de las sumas de cuadrados del análisis de varianza entre los dos grupos sobre todas las variables. Se usa con datos de distribución normal y matrices de covarianza esférica, homogéneas entre grupos.

Por último, el algoritmo que integra la estimación de las distancias y el método de agrupación puede representarse en un gráfico de árbol de dos dimensiones o dendrograma, que expresa una agrupación jerárquica (Balzarini *et al.*, 2006). Con un grupo de ramas que representan los conglomerados y estas se unen en un nodo cuya posición a lo largo del eje de distancia indica el nivel en el cual la función ocurre. El nodo donde todas las entidades forman un único conglomerado, se denomina nodo raíz.

2.3. Índices de diversidad genética, heterocigosidad insesgada de Nei y comparación de medias de Friedman

Una característica del índice de diversidad genética es su adecuación a poblaciones donde hay pocos individuos heterocigotos con muchos diferentes alelos que se encuentran en homocigosis. Este índice se calcula empleando la suma de cuadrados de las frecuencias alélicas con la expresión:

$$D = 1 - \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n p_{ij}^2 \quad (\text{Ecuación 8})$$

En la ecuación 8 p_{ij} es la frecuencia del alelo i en el locus j , para un locus de diversidad genética se calcula como $D = 1 - \sum_{i=1}^l p_i^2$, este índice permite entender la cantidad de alelos heredables o que pueden pasar a la siguiente generación. Expresando la heterocigosidad entre grupos de una misma población.

Otro índice utilizado para interpretar la variabilidad o la heterocigosis entre grupos es el índice de heterocigosis insesgada de Nei, la que puede utilizarse cuando hay un número alto de alelos y la distribución de las frecuencias alélicas no sea uniforme, entendiendo si dos poblaciones o grupos han evolucionado independientemente entre sí. Se recomienda el construir una matriz de similitud o distancia entre los pares de grupos de forma paralela, para realizar análisis de agrupamiento por métodos como comparación de medias de *Friedman*, permitiendo comparar dos o más grupos controlando por el efecto de otros factores que generan variabilidad sistemática dentro de los grupos, como por ejemplo distintos loci.

2.4. Bases de mejoramiento genético de caña de azúcar

El avance acelerado, con intervención del hombre, de los procesos de evolución natural de las plantas, con fines de interés antropogénico, se denomina fitomejoramiento y el mejoramiento genético de plantas se basa en la cruce de individuos del mismo acervo genético, para lograr individuos de características superiores en la distribución poblacional (Rosales *et al.*, 2011). El recurso filogenético o germoplasma se fundamenta en la diversidad genética.

La obtención de los mejores caracteres de los progenitores es el principio de un cruzamiento, que busca el vigor híbrido o también llamado heterosis, por lo que contar con una base amplia de variabilidad genética entre padres es de importancia.

En caña de azúcar las cruces interespecíficas, fue el hito que marcó la base para los programas modernos de mejoramiento en dicho cultivo (Quemé *et al.*, 2016). Con el criterio de crear variabilidad genética por medio de cruzamientos y selección de las mejores progenies de dichos cruzamientos, se

trabaja en los programas de mejoramiento de caña, para lo cual se necesita un banco de germoplasma de donde se toman parentales para realizar las cruces, las características del germoplasma representan la amplia diversidad genética de la población, por lo que todo programa de mejoramiento busca obtener variedades muy diversas entre sí.

Todos los programas de mejoramiento de caña han logrado obtener variedades en su mayoría tolerantes a condiciones locales y a enfermedades, pero poco han logrado mejorar significativamente en características de productividad, como la cantidad de biomasa o toneladas de caña por hectárea y el contenido de azúcar o kilogramos de azúcar por tonelada de caña (Rosales *et al.*, 2011). Lo anterior se debe a la estrecha base genética a nivel mundial de los bancos de germoplasma utilizados, reduciendo la frecuencia de selección y variedades superiores.

2.5. Marcadores moleculares

Los métodos tradicionales de mejoramiento en la actualidad se asisten por técnicas novedosas, como el uso de la biotecnología y el enfoque de emplear marcadores moleculares, que permiten definir la variabilidad genética en la población y la selección más objetiva de los parentales (Molina *et al.*, 2005). La utilización de microsatélites es una técnica que se emplea para la identificación de polimorfismos o secuencias semejantes de genes entre un grupo de individuos, considerando que mientras más polimorfismos existan entre dos individuos más semejantes genéticamente son.

La similitud genética puede ser relacionada a la falta de heterocigosis esperada, al momento de cruzar dos individuos. (Molina *et al.*, 2015)

En Brasil reportan la asociación de marcadores moleculares a procesos metabólicos de concentración de azúcar, Creste *et al.* (2009) lograron discriminar genéticamente el potencial de concentración de azúcar de una variedad.

En Estados Unidos utilizaron marcadores moleculares asociados a la enfermedad raya roja, Singh *et al.* (2011) seleccionó variedades tolerantes a la enfermedad para parentales en las cruas.

Para definir la diversidad genética total de los bancos de germoplasma, Arro (2005) en Louisiana y Tena *et al.* (2016) en Etiopia han trabajado utilizando marcadores moleculares en los programas de mejoramiento genético y a nivel comercial.

Con la utilización de tres marcadores moleculares, Molina *et al.* (2015) compararon variedades de caña de azúcar de CENGICAÑA, permitiéndole generar índices de similitud entre ellas. Dicho autor recomienda dichos marcadores para la selección de progenitores para el proceso de cruzamientos, siendo estos CV29, CV37 y CV38 (Maccheroni *et al.*, 2007).

3. PRESENTACIÓN DE RESULTADOS

Acorde con los objetivos propuestos se presentan los siguientes resultados:

3.1. **Objetivo 1. Determinación de la distancia genética entre 375 variedades de la colección nacional, utilizando tres marcadores moleculares, para comparar genéticamente las variedades**

El total de variedades analizadas es de 375 generando un total de 29 grupos correspondiente al origen y serie (año selección), los grupos V1 y V2 corresponden a variedades de diversos orígenes por lo que se agruparon al no tener más variedades con las cuales asociarlas.

Tabla II. **Número de variedades por grupo**

Grupo	Número de variedades	Grupo	Número de variedades
B	5	CP93	9
CG	17	CP94	5
CP	35	CP96	16
CP00	34	CP97	2
CP01	18	CP98	5
CP02	33	CP99	15
CP03	19	CP-L	13
CP04	9	Ho	8
CP86	7	Mex	8
CP87	3	POJ	2
CP88	15	PR	7
CP89	17	SP	6
CP90	7	V1	6
CP91	28	V2	3
CP92	23		

Fuente: elaboración propia.

El mayor grupo de variedades corresponde a las CP (*canal point*) por lo que se agruparon por serie de selección de las variedades al ser un grupo grande del mismo origen.

Se realizaron estadísticas a la base de datos, analizándose un total de 375 variedades y eliminándose las dos muestras duplicadas detectadas. El total del análisis se basó en 26 bandas, correspondiente a tres marcadores moleculares, no encontrando ningún patrón duplicado. Y el total de bandas monomórficas es 0 % y bandas polimórficas 100 %.

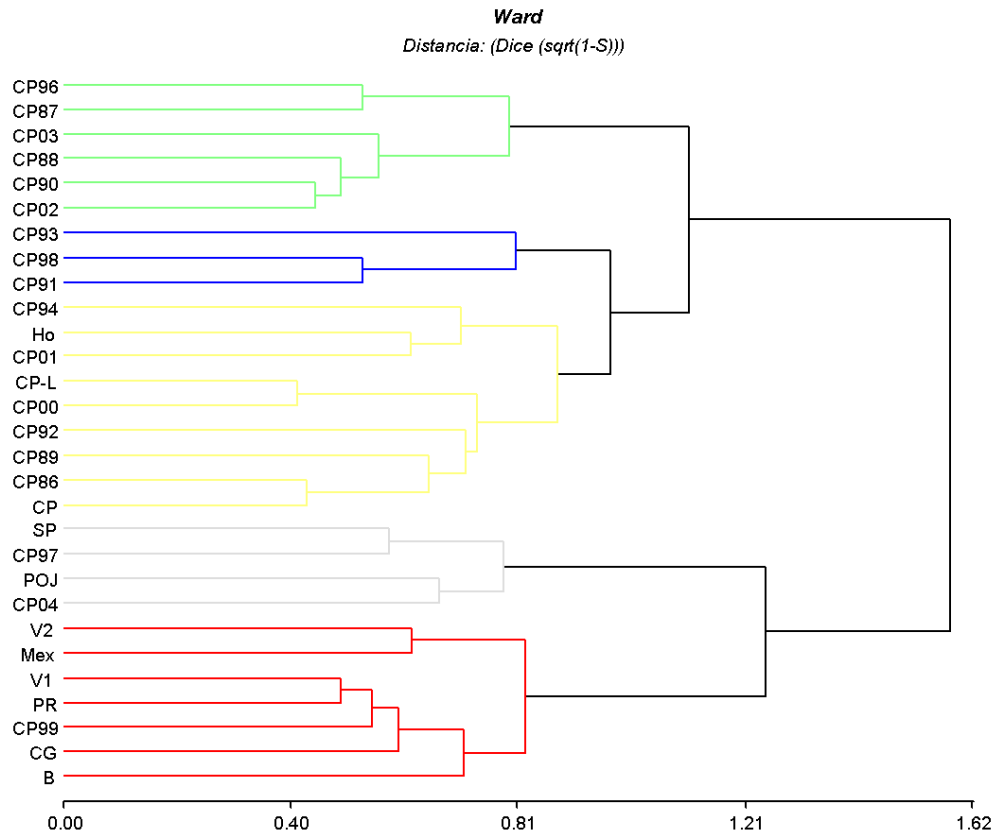
Tabla III. **Estadísticas de análisis de la base de datos de marcadores moleculares**

Resumen	Cantidad de casos
Número de muestras	375
Muestras duplicadas	2
Bandas (número)	26
Patrón bandas duplicadas	0
Bandas monomórficas	0
Bandas polimórficas (%)	100

Fuente: elaboración propia.

Se realizó el análisis de conglomerados o Cluster, para conformar grupos de variedades, utilizando el índice Dice para estimar la distancia entre grupos y agrupándolos por medio de la metodología Ward. Se definieron un total de 5 grupos, utilizando el criterio de la línea de corte en el 50 % de la distancia total.

Figura 4. **Dendrograma de variedades del banco de germoplasma**



Fuente: elaboración propia, realizado con InfoGen versión 2016.

Utilizando el índice de diversidad genética y de heterocigosidad insesgada de Nei, se realizó una comparación de medias de *Friedman*, encontrando diferencias estadísticas utilizando la diversidad genética ($p\text{-valor} < 0.0001$) y el índice de heterocigosidad insesgada de Nei ($p\text{-valor} = 0.0016$). Donde los grupos más distintos entre sí respecto a los dos índices anteriores son CP91 y CP97. La distancia genética promedio es de 37.5 % (0.375), el grupo de mayor variación presentó una distancia de 0.464 (CP91) y el grupo de menor 0.192 (CP97).

Tabla IV. **Comparación de medias de índices de diversidad genética y Heterocigosidad insesgada de Nei**

Grupos	Diversidad genética Friedman p-valor < 0.0001			Heterocigosidad insesgada de Nei Friedman = 0.0016 p-valor		
	Medias			Medias		
CP91	0.464	A		0.472	A	
CP92	0.434	A	B	0.447	A	B
CP89	0.433	A	B	0.446	A	B
CP00	0.433	A	B	0.444	A	B
CP-L	0.430		B	0.444	A	B
CP03	0.427		B	0.444	A	B
CP96	0.422		B C	0.439	A	B
CP02	0.422		B C	0.439	A	B
CP	0.421		B C	0.435	A	B
CP99	0.410		B C	0.428	A	B
CP01	0.408		B C	0.428	A	B
CP04	0.405		B C	0.427	A	B
CP88	0.401		B C	0.424	A	B
CP98	0.400		B C	0.420	A	B
CP94	0.400		B C	0.419	A	B
CP93	0.391		B C	0.415	A	B
CP90	0.389		B C	0.414	A	B
CG	0.387		B C	0.399	A	B
Mex	0.373		B C	0.397	A	B
V1	0.363		B C	0.396	A	B
Ho	0.353		B C	0.385	A	B
CP86	0.342		B C	0.377	A	B
B	0.326		B C	0.369	A	B
SP	0.310		B C	0.369	A	B
CP87	0.308		B C	0.362	A	B
PR	0.305		B C	0.338	A	B
POJ	0.288		B C	0.328	A	B
V2	0.256		C	0.308	A	B
CP97	0.192		C	0.256		B

"Medias con una letra común no son significativamente diferentes (p-valor > 0.05)"

Fuente: elaboración propia.

3.2. Objetivo 2. Determinación de la variabilidad entre los distintos grupos de variedades y dentro de ellos, por medio del análisis AMOVA, para definir la diversidad genética total de la colección nacional.

Se realizó el análisis molecular de la varianza para la población total, encontrándose que existe diferencia significativa entre los grupos (p-valor =0.0025) integrando un 1.47 % de la variabilidad total y un 98.53 % de variabilidad lo aportan los grupos a lo interno existiendo diferencias significativas (p-valor<0.0001).

Tabla V. Análisis molecular de la varianza, utilizando distancia Dice

F.V.	SC	gl	CM	p-valor	Iter.#	Comp.Var.	Porcentaje
Grupo	7.35	28	0.26	0.0025	400	3.30E-03	1.47
Dentro	76.36	346	0.22	<0.0001	400	0.22	98.53
Total	83.7	374	0.22			0.22	100

Fuente: elaboración propia.

Los valores de los coeficientes Phi_ST (Total) y Phi_Grupo son semejantes, por lo que no hay reducción de la heterosis dentro de los grupos por lo que no Nohay una estructuración de la población total.

Tabla VI. Coeficientes Phi del AMOVA

Coeficientes	Estimación
Phi_ST	0.01
Phi_Grupo	0.01

Fuente: elaboración propia.

3.3. Objetivo 3. Generación del procedimiento de análisis con AMOVA utilizando datos producto del análisis de marcadores moleculares del área de biotecnología, por medio del presente documento, para su uso en el programa de variedades de CENGICAÑA

Se esquematizó en tres pasos el proceso de análisis de datos, generados del análisis de marcadores moleculares, considerándose los siguientes pasos:

- Paso 1: conformar la base de datos, la cual consiste en una matriz digitalizada por la variable de presencia (1) o ausencia (0) del polimorfismo en el individuo. La base de datos debe de estar conformada por líneas donde se expresa cada observación de cada variable en el individuo. Colocando el código único de la variedad, a que grupo de país de origen o serie de cruzamiento pertenece, y la secuencia de presencia o ausencia de cada marcador.
- Paso 2: realizar un análisis de conglomerado, para formar grupos de variedades, utilizando el índice Dice para estimar la distancia entre grupos y agrupándolos por medio de la metodología Ward, empleando el criterio de la línea de corte en el 50 % de la distancia total.
- Paso 3: realizar el análisis molecular de la varianza para la población total, encontrando si existe diferencia significativa entre grupos o a lo interno de ellos, como herramienta para validar la segmentación de los grupos generados en el paso 2, desde una evidencia determinativa para evidenciar si dichas distancias son significativas estadísticamente.
- Paso 4: asistidos por el dendrograma generado en el paso 2, identificar los grupos o series de cruzamientos más alejados de cada conglomerado y

apoyarse utilizando el índice de diversidad genética y de heterocigosidad insesgada de Nei, y comparación de medias de Friedman, encontrando diferencias estadísticas utilizando la diversidad genética y el índice de heterocigosidad insesgada de Nei.

3.4. Objetivo General. Determinar la variabilidad genética entre y dentro de grupos de variedades utilizando el análisis AMOVA, para conocer si existe distancias genéticas significativas en la colección nacional

Existe variabilidad genética en la colección nacional, identificando diferencias significativas entre grupos de origen o series de cruzamientos (p -valor = 0.0025), y alta diferencia significativa (p -valor < 0.0001) a lo interno de cada grupo de origen o serie, identificando tres agregados, respecto a su variabilidad interna un primer grupo que tienen una mayor diversidad a lo interno (CP89, CP91, CP92, CP00) y un grupo de baja diversidad (V2 y CP97) el resto es muy semejante (no hay diferencias estadísticas) entre si respecto a su diversidad, lo anterior fue validado al utilizar el índice de heterocigosidad insesgado de Nei para hacer la comparación de medias con la misma prueba.

4. DISCUSIÓN DE RESULTADOS

La presente investigación implementó el análisis AMOVA, como herramienta para complementar la caracterización fenotipada, con información de la variabilidad genética entre los distintos grupos de variedades de la colección nacional. Tomándose la base de datos, del análisis con tres marcadores moleculares en 375 variedades, entendiendo la composición y variación genética y si esta es estadísticamente significativa entre los distintos grupos de importación o series de cruzamientos y a lo interno de ellos.

4.1. Análisis interno

A través del análisis de los datos se generaron 5 conglomerados, contrastando la distancia genética como referencias de la variabilidad que puede haber en la población en base a los tres marcadores moleculares utilizados, lo anterior a pesar de estar soportado como marcadores generales para entender la variabilidad, pueden ser un limitante para caracterizar a mayor profundidad la colección, pues dichos marcadores no están asociados a algún carácter específico como tolerancia a alguna enfermedad o sequía.

La comparación de medias utilizando el análisis no paramétrico de Friedman, usando los índices de diversidad genética, genera tres grandes grupos con diferencias estadísticas, permitiendo encontrar las significancias de las distancias en la composición de la colección nacional, pero no genera un análisis de la variabilidad a lo interno de cada grupo de origen, por lo que la utilización del AMOVA, permite alcanzar dicho entendimiento de forma complementaria, al ser un modelo jerárquico.

Con el AMOVA, se logró generar el entendimiento de la variabilidad interna de los grupos, con la siguiente característica, un primer grupo que tienen una mayor diversidad a lo interno del grupo (CP89, CP91, CP92, CP00) y un grupo de baja diversidad (V2 y CP97) el resto es muy semejante (no hay diferencias estadísticas) entre sí respecto a su diversidad, lo anterior fue validado al utilizar el índice de heterocigosidad insesgado de Nei para hacer la comparación de medias con la misma prueba. El AMOVA debe de ser complementada con las herramientas indicadas anteriormente, permitiendo hacer dichas diferenciaciones, indicando el AMOVA las significancias de las diferencias, pero no cuáles grupos son los que distan entre ellos.

Lo que es interesante es que la mayoría de los grupos (23 en total) presenta una diversidad genética semejante entre ellos (un gran número cercanos al promedio de la distancia general) al emplear el índice de diversidad genética o el índice de Nei, lo que no es detectado por el AMOVA, pues al hacer dos grupos muy distantes entre ellos, no se logra detectar que el resto son muy semejantes entre sí. A nivel mundial la variabilidad genética es muy semejante entre las principales variedades comerciales de caña de azúcar, Rosales *et al.* (2011), hace referencia a lo anterior al estudiar los reportes de diversidad genética de los principales programas de cruzamientos de caña.

4.2. Análisis externo

Los distintos programas a nivel mundial tienen una base estrecha genética, Creste *et al.* (2009) indica que al menos se utilizaron dos especies de *Saccharum* en la generación de las primeras variedades de caña de azúcar, no presentando mucha variabilidad entre programas y observando la mayor variabilidad dentro de cada programa de mejoramiento o en este caso series, debido al proceso de selección de parentales, dicho principio es evidenciado en

la distancia promedio baja encontrada en las 375 variedades (0.375) y de igual forma rangos estrechos entre los grupos más distantes entre sí, 0.464 (CP91) y el grupo de menor 0.192 (CP97).

En el estudio de tolerancia de variedades de caña a patógenos, Singh *et al.* (2011) concluye de forma semejante a Selvi *et al.* (2006) al no encontrar una gran variabilidad entre grupos respecto a los marcadores utilizados por estos autores.

El AMOVA realizado indica una baja variabilidad encontrada entre grupos (1.47 %), reforzando los resultados anteriores, pero dicha variabilidad si es significativa, Creste *et al.* (2009) al estudiar 53 variedades de caña de azúcar, indica que, a pesar de existir una baja distancia genética entre dichas variedades, si se identificó por medio del análisis AMOVA, diferencias significativas entre el grupo de variedades tolerantes a la sequía y el resto.

La mayor variabilidad (98.53 %) está dada dentro de los grupos, por lo que hay algún grupo que presenta alta variabilidad entre sus series de cruza, mostrando diferencias estadísticas ($p < 0.0001$), unido a los resultados de los índices de diversidad, los grupos CP89, CP91, CP92 y CP00 son mucho más diversos dentro de ellos, existiendo variedades muy distintas entre ellas, generando un grado de confianza alto en los resultados, Arro (2005) que estudió 63 variedades de origen CP y L encontrando que no había variabilidad entre grupos de variedades anterior a la serie 80, la mayor variabilidad se daba a lo interno de las series CP generadas después de las series 80.

El coeficiente Phi ST es de 0.01 indicando que la pérdida de heterocididad entre grupos es muy pequeña, por lo que no hay una estructuración jerárquica de la población, si no que la estructura de la población es plana, pues la mayor

cantidad de series o grupos están muy cercano entre sí, Khan, Pandey, Kumar y Lal (2011) al comparar 30 variedades (en India) encuentran una variabilidad promedio de 52.5 %, indicando que las variedades susceptibles eran las más variables con 86 %, pero entre las medianamente tolerantes y las muy tolerantes no encontraron diferencia, y al realizar el AMOVA mostró variación entre los grupos, marcando dicha distancia del grupo de las variedades susceptibles con el resto.

En la presente investigación, el análisis de conglomerados distinguió claramente agrupaciones, pero con distancias promedio cercanas a 37 %, marcando una baja variabilidad o significancia entre grupos, mostrando el análisis del AMOVA una significancia entre las diferencias de dichos grupos, semejante con el trabajo realizado en India, a pesar de que en general las estructuras de las poblaciones en ambos casos son poco variables, y un número reducido de variedades o grupos son lo que aportan la mayor diversidad total a la población.

CONCLUSIONES

1. Se determina la distancia genética promedio de 375 variedades de la colección nacional, en 0.375, el grupo de mayor variación presentó una distancia de 0.464 (CP91) y el grupo de menor 0.192 (CP97).
2. Se identifica alta diferencia estadística significativa entre los grupos de origen o series de cruzamientos (p-valor 0.0025) y a lo interno de dichos grupos (p-valor < 0.0001).
3. El procedimiento utilizado en el presente trabajo puede implementarse como un proceso para el entendimiento y fortalecimiento de la caracterización del banco de germoplasma de CENGICAÑA.
4. Al utilizar la herramienta AMOVA, se logra determinar las diferencias significativas entre las distancias genéticas de cada grupo y a lo interno de cada grupo de variedades de la colección nacional, evidenciando la existencia de variabilidad genética de la misma, garantizando un buen recurso genético para realizar cruzamientos y obtener nuevas variedades.

RECOMENDACIONES

1. Continuar con el análisis de más variedades, utilizando marcadores moleculares (CV29, CV37 y CV38), para incrementar la caracterización de la CN con la diversidad genética.
2. Correlacionar el fenotipado de la CN con los grupos y datos generados en la presente investigación.
3. Utilizar el análisis molecular de la varianza (AMOVA) como una herramienta anual en el análisis de marcadores moleculares.
4. Continuar con la determinación de la variabilidad genética entre y dentro de los grupos de la CN con el AMOVA al incrementar el número de variedades analizadas por el laboratorio de biotecnología de CENGICAÑA.

REFERENCIAS

1. Arro, J. (2005). *Genetic diversity among sugarcane clones using target region amplification polymorphism(trap)markers and pedigree relationships*. Louisiana: Louisiana State University.
2. Balzarini, M., Arroyo, A., Bruno, C., & Di Rienzo, J. (2006). *Análisis de datos de marcadores con Info-Gen*. San Luis, Argentina: XXXV Congreso Argentino de Genética.
3. Bonilla, B., Espinosa, P. (2003). *Colección, caracterización fenotípica y molecular de poblaciones de uchuva Physalis peruviana L.* Universidad Nacional de Colombia, Sede Palmira.
4. Creste, S., Accoroni, K., Pinto, L., Vencovsky, R., Gimenes, M., Xavier, M., & Landell, M. (2009). *Genetic variability among sugarcane genotypes based on polymorphisms in sucrose metabolism and drought tolerance genes*. Springer Science.
5. Dávila, M., Castillo, M., Laurentin, H. (2007). *Uso de marcadores moleculares ISSR para inferir las relaciones genéticas y la variabilidad intraespecífico en Agave*. Rev. Fac. Agron (Maracay, 33:93-111.
6. Excoffier, L. (1992). *AMOVA (analysis of Molecular Variance)* 156. University of Geneva, Switzerland

7. Gómez, A., Aguiriano, E., Alia, R., Bueno, M. (2002). *Análisis de los recursos genéticos de Pinus pinea L en España mediante microsatélites del cloroplasto*. Invest. Agr.: Recur. For, vol. 11(1), 2002.
8. Marulanda, M., Márquez, M. (2001). *Caracterización de la diversidad genética de Rubus glaucus Benth con Marcadores Moleculares RAPDs*. Actual Biol. 23(74):57-63.
9. Mengoni, A., & Bazzicalupo, M. (2002). *The statistical treatment of data and the analysis of molecular variance (AMOVA) in molecular microbial ecology*. Annals of Microbiology, 95-101.
10. Molina, L., Maddaleno, C., Sut, V., & Camargo, P. (2015). *Caracterización molecular de 100 variedades de caña de azúcar mediante microsatélites*. Memoria de resultados CENGICAÑA 2014-15, 190-197.
11. Molina, L., Quemé, J., Rosales, F. (2013). *Comparative analysis between phenotype and Bru1 marker for incidence to brown rust in sugarcane*. Proc, int, Soc, Sugar Cane Technol. (28):1-6.
12. Molina, L., Ponciano, K., & Fuentes, M. (2005). *Caracterización molecular de la diversidad existente en la colección nacional de maíz (Zea mays L.) utilizando marcadores de secuencia simple repetida*. Guatemala: Proyecto FODECYT No. 28-03.
13. Morillo, C., Morillo, A., Muñoz, Y., Vásquez, J., & Zamorano, H. (2005). *Caracterización molecular con microsatélites aleatorios de ramitas de la*

colección de mora, *Rubus spp*, de la Universidad Nacional de Colombia Sede Palmira. Acta Agronómica. 54(2), Universidad Nacional de Colombia. (Agosto de 2022). ISSN: 0120-2812. Recuperado de <https://www.redalyc.org/articulo.oa?id=169920332003>

14. Nagamine, Y., Higuchi, M. (2001). *Genetic distance and classification of domestics animals using genetic markers*. . Journal of Animal Breeding and Genetics. 118: 101-109.
15. Nei, M. Kumar, S (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
16. Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York. 180.
17. Nei, M. (1973). *Analysis of gene diversity in subdivided populations*. Proceedings of the National Academy of Sciences of USA. 70:3321-3323.
18. Orozco, H., & Buc, R. (2016). *Censo de variedades de caña de azúcar para la zona cañera de Guatemala para la zafra 2015-16*. Memoria de Resultados CENGICAÑA 2015-16, 21-31.
19. Quemé, J., Orozco, H., & Salazar, A. (2016). *Resultados con el recurso genético y cruzamientos del programa de variedades de CENGICAÑA en la zafra 2015-16*. Memoria de resultados CENGICAÑA 2015-16, 121-126.

20. Quemé, J., Orozco, H., & Salazar, A. (2016). *Selección de progenitores para mejorar el contenido de azúcar en variedades CG de caña de azúcar*. Memora de resultados CENGICAÑA 2015-16, 159-165.
21. Racedo, J., Perera M., Bertani, R., Funes, C., Gonzalez, V., Cuenya, M. D'Hont, A., Welin, B., Castagnaro, A. (2013). *Bru 1 gene and potencial alternative sources of resistance to sugarcane brown rust disease*. Euphytica 191, 429-436.
22. Rosales, F., Quemé, J., & Melgar, M. (2011). *Tendencias en la adopción y el cultivo de variedades de caña de azúcar en cinco países, comparado con las tendencias en Guatemala*. Memoria de resultados CENGICAÑA, 51-62.
23. Singh, R., Khan, M., Pandey, D., Kumar, S., & Lal, S. (2011). *Analysis of genetic differentiation and phylogenetic relationships among sugarcane genotypes differing in response to red rot*. Sugar Tech, 137-144.
24. Slatkin, M. (2007). *Gene flow and the geographic structure of natural populations*. Science, vol. 236. 787- 236.
25. Weir, B. (1996). *Genetic data analysis II. Methods for discrete population genetic data*. Sinauer Associates, INC. Sunderland, Massachusetts. 283.. Sugar Tech, 137-144.

APÉNDICE

Apéndice 1. Matriz de coherencia.

Preguntas de investigación	Objetivos	Metodología	Resultados	Conclusiones	Recomendaciones
¿Los datos obtenidos del análisis de marcadores moleculares, nos permite determinar la variabilidad genética en la colección nacional?	Determinar la distancia genética entre 375 variedades de la colección nacional, utilizando tres marcadores moleculares, para comparar genéticamente las variedades.	Se realizó un análisis de conglomerados (métodos estadísticos) y se obtuvo la estadística descriptiva de este análisis, generando un dendrograma con la distancia Dice. También comparando las distancias por medio de los índices de diversidad genética y heterocigosidad inestada de Nei	Se generaron 29 grupos con las 375 variedades, según el origen o la serie de cruzamiento, generando 5 conglomerados, con una distancia promedio de 37.5%. Al utilizar el índice de diversidad genética se logró identificar grupos más distantes.	Se determinó la distancia genética promedio de 375 variedades de la colección nacional de CENGICAN A, siendo esta estimada en 0.375, el grupo de mayor variación presentó una distancia de 0.464 (CP91) y el grupo de menor 0.192 (CP97).	Avanzar en el análisis molecular con más variedades de los grupos de países de importación o series de cruza de interés, como parte del entendimiento desde la perspectiva de variabilidad genética de la colección nacional, al contar con un 14% de caracterización con marcadores moleculares.

Continuación apéndice 1.

¿La incorporación de variedades de diversos lugares de origen genera valor y cuáles de estos lugares incrementa más la variabilidad genética de la colección nacional?	Determinar la variabilidad entre los distintos grupos de variedades y dentro de ellos, por medio del análisis AMOVA, para definir la diversidad genética total de la colección nacional.	Del total de 375 variedades analizadas con tres marcadores moleculares, se conglomeraron en 29 grupos con un origen común o serie de cruzamiento común, para generar un análisis molecular de la varianza (AMOVA), permitiendo identificar, las diferencias significativas de las distancias genéticas entre los 29 grupos y dentro de estos.	Se realizó el análisis molecular de la varianza para la población total, encontrándose que existe diferencia significativa entre los grupos (p-valor =0.0025) integrando un 1.47% de la variabilidad total y un 98.53% de variabilidad lo aportan los grupos a lo interno existiendo diferencias significativas (p-valor<0.0001).	Se determinaron deferencias significativas (p-valor 0.0025) entre los grupos de importación o series de cruzamientos, identificándose que los grupos: CP89, CP91, CP92 y CP00, son los más diversos entre ellos. Y alta diferencia significativa a lo interno de dichos grupos (p-valor<0.0001). Existe diferencia significativa entre grupos de variedades, pero esta es muy baja, la mayor diferencia se da dentro de cada grupo, siendo estos CP89, CP91, CP92 y CP00. Por lo anterior estos grupos son los que representan mayor variabilidad	Utilizar el fenotipado de las variedades dentro de los grupos CP89, CP91, CP92 y CP00, para explorar correlaciones entre las distancias genéticas y sus características respecto a características morfológicas, tolerancia a enfermedades y concentración de azúcar.
--	--	---	---	---	---

Continuación apéndice 1.

¿Cómo se puede incorporar el análisis con marcadores moleculares de nuevas variedades, al estudio de diversidad genética de la colección nacional?	Generar el procedimiento de análisis con AMOVA con datos producto del análisis de marcadores moleculares del área de biotecnología, por medio del presente documento, para su uso en el programa de variedades de CENGICAÑA.	Se esquematizó los pasos lógicos, para replicar los resultados obtenidos en la presente investigación, agregando nuevos datos a la base original de 375 individuos.	Se esquematizó en tres pasos el proceso de análisis de datos, generados del análisis de marcadores moleculares	El procedimiento utilizado en el presente trabajo, puede implementarse como un proceso para el entendimiento y fortalecimiento de la caracterización del banco de germoplasma, empleando la información generada por medio de marcadores moleculares, permitiendo entender el fenotipo y variabilidad genética de la colección nacional y el proceso de recurso genético, con los siguientes pasos a seguir: paso1: conformar la base de datos, paso 2: realizar un análisis de conglomerados paso 3: realizar el análisis molecular de la varianza para la población total (AMOVA), paso 4: identificar los grupos o series.	Utilizar el análisis molecular de la varianza (AMOVA) como una herramienta rutinaria en el análisis de marcadores moleculares, como complemento de herramientas como dendrograma y análisis de conglomerados, para la cuantificación de la significancia entre las distancias genéticas.
--	--	---	--	---	--

Continuación apéndice 1.

¿La variabilidad genética de la colección nacional presenta heterocigosidad para ser usada como banco de germoplasma?	Determinar la variabilidad genética entre y dentro de grupos de variedades, utilizando un análisis AMOVA, para conocer si existen distancias genéticas significativas en la colección nacional.	Utilizando el AMOVA, se definió las diferencias significativas entre las 29 agrupaciones por lugar de origen o serie de cruzamiento y la significancia de las distancias genéticas entre las variedades dentro de cada uno de las 29 agrupaciones.	Existe variabilidad genética en la colección nacional, identificando diferencias significativas entre grupos de origen o series de cruzamientos (p-valor =0.0025), y alta diferencia significativa (p-valor<0.0001) a lo interno de cada grupo de origen o serie, identificando tres agregados, respecto a su variabilidad interna un primer grupo que tienen una mayor diversidad a lo interno (CP89, CP91, CP92, CP00) y un grupo de baja diversidad (V2 y CP97) el resto es muy semejante (no hay diferencias estadísticas)	Al utilizar la herramienta AMOVA, se logró determinar las diferencias significativas entre las distancias genéticas de cada grupo de variedades de importación o serie de cruzamiento y dentro de estos grupos de la colección nacional de CENGICAÑA, al ser el análisis de conglomerados una herramienta descriptiva se complementa con los resultados cuantitativos del AMOVA, encontrando que existen diferencias tanto entre grupos como a lo interno, evidenciando la variabilidad genética de la colección nacional, que es una de las características deseadas en un banco de germoplasma.	Continuar con la determinación de la variabilidad genética entre y dentro de cada grupo de variedades, utilizando el AMOVA, al evidenciarse como una herramienta de caracterización desde la visión genética de la colección nacional.
---	---	--	--	---	--

Fuente: elaboración propia.