Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Biológicas
Escuela Profesional de Genética y Biotecnología

Diversidad y estructura genética de 731 accesiones de la Colección Nacional de Yuca (*Manihot esculenta Crantz*) del Banco de Germoplasma del Instituto Nacional de Innovación Agraria usando polimorfismos de nucleótidos simples (SNPs)

TESIS

Para optar el Título Profesional de Biólogo Genetista
Biotecnólogo

AUTOR

Bianca Estefani VIGIL SANTILLAN

ASESOR

Mg. Elizabeth FERNANDEZ HUYTALLA

Dra. Mónica ARAKAKI MAKISHI

Lima, Perú

2021



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Vigil, B. (2021). *Diversidad y estructura genética de 731 accesiones de la Colección Nacional de Yuca (Manihot esculenta Crantz) del Banco de Germoplasma del Instituto Nacional de Innovación Agraria usando polimorfismos de nucleótidos simples (SNPs)*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Biológicas, Escuela Profesional de Genética y Biotecnología]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Bianca Estefani Vigil Santillan
Tipo de documento de identidad	DNI
Número de documento de identidad	70793479
URL de ORCID	https://orcid.org/0000-0002-1199-1191
Datos de asesor	
Nombres y apellidos	Mónica Arakaki Makishi
Tipo de documento de identidad	DNI
Número de documento de identidad	08136445
URL de ORCID	https://orcid.org/0000-0003-1081-2507
Datos de coasesor	
Nombres y apellidos	Elizabeth Fernandez Huaytalla
Tipo de documento de identidad	DNI
Número de documento de identidad	42046435
URL de ORCID	https://orcid.org/0000-0001-6494-2767
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Julio Solis Sarmiento
Tipo de documento	DNI
Número de documento de identidad	09671619
Miembro del jurado 1	
Nombres y apellidos	Domingo Iparraguirre León

Tipo de documento	DNI
Número de documento de identidad	08609804
Miembro del jurado 2	
Nombres y apellidos	Hector Sanchez Sotomayor
Tipo de documento	DNI
Número de documento de identidad	06002430
Datos de investigación	
Línea de investigación	A.1.1.3 Diversidad Genética
Grupo de investigación	No aplica
Agencia de financiamiento	Perú. Instituto Nacional de Innovación Agraria (INIA). Programa financiador: Programa Nacional de Innovación Agraria (PNIA). 230_PI Perú. Universidad Nacional Mayor de San Marcos. Vicerrectorado de Investigación y Posgrado. Programa de Promoción de Tesis de Pregrado. E18030044-PTPGRADO.
Ubicación geográfica de la investigación	Lugar: Instituto Nacional de Innovación Agraria. País: Perú Departamento: Lima Provincia: Lima Distrito: La Molina Calle: Av. La Molina 1981 Coordenadas geográficas: 12°4′36″ S, 76°56′43″ W (-12.07664, -76.94252).
Año o rango de años en que se realizó la investigación	Agosto 2018 – Abril 2020
URL de disciplinas OCDE	Genética, Herencia https://purl.org/pe-repo/ocde/ford#1.06.07 Ciencias de las plantas, Botánica https://purl.org/pe-repo/ocde/ford#1.06.10 Agricultura https://purl.org/pe-repo/ocde/ford#4.01.01

Universidad Nacional Mayor de San Marcos
(Universidad del Perú, Decana de América)

FACULTAD DE CIENCIAS BIOLÓGICAS

ACTA DE SESIÓN PARA OPTAR AL TÍTULO PROFESIONAL DE BIÓLOGA GENETISTA BIOTECNÓLOGA (MODALIDAD: SUSTENTACIÓN VIRTUAL DE TESIS)

Siendo las 17:05 horas del 12 de octubre de 2021, en el Salón de Grados Virtual de la Facultad de Ciencias Biológicas cuya dirección electrónica fue https://meet.google.com/ssy-rnvg-wzm, y en presencia del Jurado formado por los profesores que suscriben, se inició la sesión para optar al **Título Profesional de Bióloga Genetista Biotecnóloga** de **BIANCA ESTEFANI VIGIL SANTILLAN.**

Luego de dar lectura y conformidad al expediente **N° UNMSM-20210042514**, la titulando expuso su tesis: **"DIVERSIDAD Y ESTRUCTURA GENÉTICA DE 731 ACCESIONES DE LA COLECCIÓN NACIONAL DE YUCA (*Manihot esculenta* Crantz) DEL BANCO DE GERMOPLASMA DEL INSTITUTO NACIONAL DE INNOVACIÓN AGRARIA USANDO POLIMORFISMOS DE NUCLEÓTIDOS SIMPLES (SNPs)"**, y el Jurado efectuó las preguntas del caso calificando la exposición con la nota **20**, calificativo: **APROBADO CON MÁXIMOS HONORES.**

Finalmente, el expediente será enviado a la Escuela Profesional de Genética y Biotecnología y al Consejo de Facultad para que se apruebe otorgar el **Título Profesional de Bióloga Genetista Biotecnóloga** a **BIANCA ESTEFANI VIGIL SANTILLAN** y se eleve lo actuado al Rectorado para conferir el respectivo título, conforme a ley.

Siendo las 19:10 horas se levantó la sesión.

Ciudad Universitaria, 12 de octubre de 2021.

Dr. JULIO SOLIS SARMIENTO
(PRESIDENTE)

Dra. MÓNICA ARAKAKI MAKISHI
(ASESORA)

Mg. DOMINGO IPARRAGUIRRE LEÓN
(MIEMBRO)

Blgo. HÉCTOR SÁNCHEZ SOTOMAYOR
(MIEMBRO)

Dedicado a

A todos los días de la etapa CH

AGRADECIMIENTOS

A Dios por guiar cada uno de mis pasos y acompañarme siempre.

A mis padres Luis y Asunta, por darme siempre su amor y apoyo incondicional, alentarme a seguir siempre lo que me gusta y animarme en todas las decisiones que he tomado. Los amaré siempre.

A mi hermano Luis Enrique, por poner siempre a prueba mi paciencia, nivel de madurez y mostrame la vida desde su punto de vista. Te quiero igual.

A mi abuelita "Cuper" que siempre la llevo en mi corazón.

A mi tía Meche y primo Randy, por sacarme siempre de la rutina, con sus locas ideas y saber que siempre me apoyaron y apoyarán en todo lo que haga.

Al profesor Pedro Castellanos, por ser mi primer guía en el proceso de investigación, por siempre estar presto a escucharme, cada una de sus palabras son muy valiosas para mí. Usted es un verdadero maestro.

A la profesora Mónica Arakaki, por aceptar ser asesora de este trabajo, por mostrar siempre su disposición a ayudarme y por recibirme siempre con una sonrisa.

A mis mejores amigos del mundo biológico, Francisco, Guido, Rubén y Tania, por preocuparse siempre por mí, por los lindos momentos de "catarsis" de ciencia y plantas, por demostrarme que siempre hay más que aprender.

A mis amigos del INIA, Hans, Juan Daniel, Dioloza, Maciel, Diego, Ralph, Yoshimar, Karen, Jessica, Karina y Olegario, por hacer de este trabajo de investigación divertido y llevadero y por aguantar mis quejas siempre.

A mis amigos de Chincha, Nohe, Angeline, Rosita, Lesly, Silvia y Christian que con su alegría y humildad me ayudaron a encontrar el equilibrio para poder terminar este proyecto, lo mejor paso aquí.

A la Mg. Elizabeth Fernandez y al Ing. Fredy Quispe, por depositar su confianza en mí y por su apoyo para cumplir todos los objetivos del presente trabajo.

Al proyecto 230-PI financiado por el PNIA, quienes apostaron por el concocimiento de la diversidad y mejoramiento genético en yuca.

ABREVIATURAS

BGC : Bayesian genotype caller

CTAB : Cetyl Trimethyl Ammonium Bromide

CIAT : Centro Internacional de Agricultura Tropical

CSP : Community Sequencing Program

CTCRI : Central Tuber Crops Research Institute

DrT : Diversity Array Technology

ddRADseq : Double Digest RADseq

DOE-JGI : Department of Energy Joint Genome Institute

E.E.A : Estación Experimental Agraria

EMBRAPA : Empresa Brasilera de Pesquería Agropecuaria

GBS : Genotyping-by-Sequencing

GS : Genomic Selection

GWAS : Genome Wide Association

IAN : Instituto Agronómico Nacional de Caacupé ahora Centro de Investigación Hernando Bertoni

IITA : Instituto Internacional de Agricultura Tropical

ILTAB : International Laboratory for Tropical Agricultural Biotechnology

INIA : Instituto Nacional de Innovación Agraria

MAF : Minor allele frequency

MAS : Marker-assisted Selection

NGS : Next-generation Sequencing

NRCRI : National Root Crops Research Institute, Umudike

PCR : Polymerase Chain Reaction

PGRC : Plant Genetic Resources Centre

PNIA : Programa Nacional de Innovación Agraria

QTL : Quantitative Trait Locus

RAD : Restriction site-associated DNA

RADseq : Restriction site-associated DNA sequencing

RAPD **:** Random Amplified Polymorphic DNA

RFLP **:** Restriction Fragment Length Polymorphism

SAM **:** Sequence Alignment/Map

VCF **:** Variant Call Format

ÍNDICE GENERAL

1. INTRODUCCIÓN 1

2. MARCO TEÓRICO 3

2.1. La yuca 3

2.1.1. Actualidad 3

2.1.2. Posición taxonómica 4

2.1.3. Centro de origen 6

2.2. La Colección Nacional de Yuca del INIA 8

2.3. Marcadores moleculares usados para estudios de diversidad y estructura genética en *M. esculenta* 8

2.4. Genoma de *Manihot esculenta* 11

2.5. Genotipado por secuenciamiento (GBS) 13

2.6. Conceptos de genética poblacional 18

2.6.1. Diversidad Genética 18

2.6.2. Estructura genética poblacional 18

2.6.3. Principio de Hardy-Weinberg 19

2.6.4. Frecuencias alélicas 19

2.7. Estadísticos para el estudio de la diversidad genética 20

2.7.1. Proporción de loci polimórficos (P) 20

2.7.2. Heterocigosidad 21

2.7.3. Contenido de información polimórfica (PIC) 22

2.8. Diferenciación genética entre poblaciones 23

2.8.1. Estadísticos F 23

2.8.2. Distancias genéticas 24

2.9. Pruebas de asignación 25

2.9.1. Métodos de clasificación 25

2.9.2. Métodos de asignación 26

3. HIPÓTESIS Y OBJETIVOS 27

3.1. Hipótesis 27

3.2. Objetivos 27

3.2.1. Objetivo general 27

3.2.2. Objetivos específicos 27

4. MATERIALES 27

4.1. Material biológico 27

4.2. Material de laboratorio 28

5. MÉTODOS 28

5.1. Obtención del material genético 29
5.1.1. Toma de muestras en la E.E.A Donoso en Huaral 29
5.1.2. Extracción de ADN genómico 31
5.1.3. Verificación de calidad del ADN genómico 32
5.1.4. Cuantificación del ADN genómico 32
5.2. Genotipado por secuenciamiento 33
5.3. Procesamiento bioinformático 33
5.3.1. Alineamiento de lecturas 34
5.3.2. Filtrado de lecturas 34
5.3.3. Llamada de variantes (SNP calling) 34
5.3.4. Filtrado de variantes 35
5.3.5. Análisis de diversidad 35
5.3.6. Análisis de estructura poblacional 35
6. RESULTADOS 37
6.1. Genotipado por secuenciamiento de la Colección Nacional de Yuca. 37
6.1.1. Genotipado por secuenciamiento 37
6.1.2. Alineamiento de *reads* 38
6.1.3. SNP calling 39
6.1.4. Filtrado de SNPs 41
6.2. Análisis de la diversidad y estructura genética de la Colección Nacional de Yuca 44
6.2.1. Diversidad genética de la colección 44
6.2.2. Diversidad genética en la población de yucas peruanas 46
6.2.3. Estructura poblacional en la Colección Nacional de Yuca. 48
7. DISCUSIÓN 69
7.1. Genotipado por secuenciamiento 69
7.2. Procesamiento bioinformático de lecturas 72
7.2.1. SNP calling 74
7.2.2. Filtrado de SNPs 75
7.3. Análisis de diversidad y estructura 77
7.3.1. Diversidad genética 77
7.3.2. Estructura genética 85
8. CONCLUSIONES 92
9. RECOMENDACIONES 93
10. REFERENCIAS BIBLIOGRÁFICAS 93
11. ANEXOS 111

ÍNDICE DE TABLAS

Tabla 1. Número de accesiones por país y región peruanas, utilizadas en el estudio de diversidad y estructura genética……28

Tabla 2: Distribución y densidad de los 264,088 SNPs identificados en *M. esculenta*, Crantz……40

Tabla 3: Resumen de estadísticos de diversidad genética en base a los 1,271 SNP, para toda colección y para los países que la conforman……..……...45

Tabla 4: Resumen de estadísticos de diversidad genética en las regiones de Ucayali, Lima, San Martín, Huánuco y La Libertad……..……47

Tabla 5: Valores del Fst por pares o "parwise Fst" entre los cuatro países que forman parte de la colección……..……49

Tabla 6: Matriz de distancia genética de Nei, entre los cuatro países que forman parte de la colección……49

Tabla 7: Valores del Fst por pares o "parwise Fst" entre las poblaciones de yuca peruanas……..……50

Tabla 8: Matriz de distancia genética de Nei, entre las poblaciones de yuca peruanas……..……50

Tabla 9: Clasificación y descripción de los cuatro grupos de accesiones de yuca, definidos en base a los valores de representatividad del cos^2 de las dos primeras componentes PC1 y PC2……52

Tabla 10: Valores del porcentaje de asignación a los grupos para la cantidad de CP retenidos……63

Tabla 11: Tabla de asigación de las accesiones de cada país y provincia a los 12 grupos genéticos definidos (K=12) por el valor de BIC……66

Tabla 12: Tabla de valores de PAI para cada grupo genético definido……68

ÍNDICE DE FIGURAS

Figura 1. Generación de un marcador RAD (Fuente: Baird *et al*., 2008) ……17

Figura 2. Barplot del número de lecturas (*reads*) para las 731 accesiones, ordenados de mayor a menor……37

Figura 3: Frecuencia de accesiones de acuerdo al número de *reads*……38

Figura 4: Histograma del porcentaje de alineamiento de *reads* de las 731 accesiones……38

Figura 5: Histograma de la cobertura por *locus* de cada accesión……39

Figura 6: Diagrama de densidad de SNP por cromosoma, con una ventana de 1 Mb, incluye el genoma plastidial (NC_010433.1) ……42

Figura 7: Histograma de frecuencia del porcentaje de data faltante por accesión, en rojo se indica el límite de tolerancia en 0.55……43

Figura 8: Plot de los valores de MAF para cada alelo de los 1,271 SNPs, resultantes del tercer filtrado..... ……...44

Figura 9: Plot de genotipos por locus para cada accesión de la colección……46

Figura 10: Gráfica del análisis de PCA para las 728 accesiones usando la agrupación por países de procedencia con los 1,271 SNP……...53

Figura 11: Gráfica de los valores del cos^2 de las dos primeras componentes PC1 y PC2 de las 728 accesiones……..……54

Figura 12: Gráfica del análisis de PCA de las 634 accesiones peruanas, agrupadas en base a región de procedencia, usando los 1,271 SNP……55

Figura 13: Gráfica de los valores del cos^2 de las dos primeras componentes (PC1 y PC2) para las 634 accesiones peruanas……57

Figura 14: Diagrama de barras apiladas compuestas de las probabilidades de asignación de individuo (PAI) del DAPC de las 728 accesiones de la Colección Nacional, considerando la agrupación en base a origen de procedencia……59

Figura 15: Diagrama de barras apiladas compuestas, ordenada por grupos de

procedencia, de las probabilidades de asignación (PAI) del DAPC de las 728 accesiones de la colección……60

Figura 16: Plot del DAPC para las 728 accesiones, basada en las funciones discriminantes lineales (LD1 y LD2) ……64

Figura 17: Tabla de visualización gráfica de correspondencia entre los grupos inferidos (K=12) y la agrupación en base a país y provincia……67

RESUMEN

La yuca, *Manihot esculenta* Crantz (Euphorbiaceae), es uno de los cultivos de subsistencia más importantes a nivel mundial. En el Perú, la raiz tuberosa de yuca es la más consumida por los pueblos indígenas de la región Amazónia, estando su consumo segundo despúes de la papa, como fuente de energía calórica. El objetivo del presente trabajo fue evaluar la diversidad y estructura genética de 728 accesiones de yuca, conservadas en la Colección Nacional de Yuca del INIA, usando 1,271 SNPs obtenidos de un ddRADseq (de 731 accesiones) y distribuidos de manera proporcional en los 18 cromosomas, que dan una visión global de todo el genoma. Los valores promedio de PIC, MAF, Ho y He, fueron; 0.245, 0.197, 0.319 y 0.293 respectivamente, que indican un elevado valor de diversidad genética conservada en la colección. La agrupación a priori en base a lugar de origen o procedencia (país o región) que actualmente usa el INIA, no logra explicar la diversidad contenida en la colección, corroborado por los valores de AMOVA (diversidad >87% entre las accesiones), Fst (0.0645) y PCA. Sin embargo, con el análisis de DAPC, se encontró que la colección se ordena en 12 grupos genéticos, respaldados por los altos valores de asignación individual (PAI) de cada accesión (>0.98). No se encontró una correspondencia entre genotipo y origen o procedencia, a excepción de 65 accesiones de Ucayali que forman exclusivamente, 2 de los 12 grupos genéticos encontrados. Estos resultados, demuestran el alto intercambio de material vegetal entre las diferentes regiones peruanas, debido al cultivo tradicional en el sistema de “chacras”, las relaciones socioculturales de parentesco y reciprocidad de los pueblos indigenas andino-amazonicos, y por el fenómeno de *volunteer seedling* de su reproducción sexual. Los resultados proporcionan, una mejor comprensión de la variabilidad genética conservada en la colección y la agrupación a nivel nacional en base a genotipos.

Palabras clave: Yuca, SNP, diversidad genética, grupos genéticos, Colección Nacional.

ABSTRACT

Cassava, Manihot esculenta Crantz (Euphorbiaceae), is one of the most important subsistence crops worldwide. In Peru, the tuberous cassava root is the most consumed by the indigenous peoples of the Amazon region, its consumption being second only to the potato, as a source of caloric energy. The objective of the present work was to evaluate the diversity and genetic structure of 728 cassava accessions, conserved in the National Cassava Collection of the INIA, using 1,271 SNPs obtained from a ddRADseq (of 731 accesions) and distributed proportionally in the 18 chromosomes, which give a global vision. of the entire genome. The average values of PIC, MAF, Ho and He, were; 0.245, 0.197, 0.319 and 0.293 respectively, which indicate a high value of genetic diversity conserved in the collection. The a priori grouping based on place of origin or provenance (country or region) currently used by INIA, fails to explain the diversity contained in the collection, corroborated by the AMOVA values (diversity> 87% among accessions), Fst (0.0645) and PCA. However, with the DAPC analysis, it was found that the collection is ordered in 12 genetic groups, supported by the high individual allocation values (PAI) of each accession (> 0.98). No correspondence was found between genotype and origin or provenance, with the exception of 65 Ucayali accessions that exclusively form 2 of the 12 genetic groups found. These results demonstrate the high exchange of plant material between the different Peruvian regions, due to the traditional cultivation in the "chacras" system, the sociocultural relations of kinship and reciprocity of the Andean-Amazonian indigenous peoples, and the phenomenon of *volunteer seedling*. of their sexual reproduction. These results provide a better understanding of the genetic variability conserved in the collection and grouping at the national level based on genotypes.

Keywords: Cassava, SNP, genetic divesity, genetics groups, National Collection.

1. INTRODUCCIÓN

La yuca *Manihot esculenta* Crantz, pertenece a la familia Euphorbiaceae. Es una de las 98 especies del género *Manihot* (Howeler *et al*., 2013), siendo la única especie comestible Es originaria de Sudamérica, considerándose a la región del "Cerrado" (Rondónia y noroeste de Mato Groso) en Brasil, como su centro de origen geográfico y agrícola, ya que alberga la mayor diversidad genética (Olsen y Schaal, 2001); se cultiva ampliamente en toda la región Amazónica dentro de los límites de Perú, Colombia y Brasil. Para los pueblos indígenas sudamericanos, es un cultivo de importancia alimentaria y ancestral, la cultivan de manera tradicional hasta la actualidad ya que su dieta diaria se basa en su consumo (Allem, 2002; Howeler *et al*., 2013).

Los amplios estudios de diversidad, usando marcadores morfológicos y moleculares (AFLP, SSR, SNP, etc), realizados tanto en la cuenca Amazónica que comprende su cultivo, como en otros países donde fueron introducidos posteriormente (países africanos y asiáticos), demuestran que la yuca posee una elevada diversidad genética (Peroni y Hanazaki, 2002; Heckler y Zent, 2008; Turyagyenda *et al*., 2012); que ha sido mantenida principalmente por las prácticas tradicionales de los pueblos indígenas (Núñez *et al*., 2018).

Los resultados de estos estudios, han permitido descubrir variedades con características agronómicas importantes, por ejemplo, variedades con alto rendimiento, con facilidad de cosecha, con resistencia a plagas y enfermedades, menor toxicidad, facilidad de procesamiento, mejor contenido nutricional, entre otras; que facilitan la elección de parentales para su mejoramiento genético.

Para conservar esta diversidad y ampliar los estudios en este campo, se han creado las colecciones o bancos de germoplasma, cuya finalidad es preservar, clasificar y asegurar

la existencia de las diversas variedades de yuca, de una determinada zona o región geográfica. En estas colecciones, también encontramos datos pasaporte (datos de localidad de colecta, fecha, hábitat, etc), información básica para entender y organizar la diversidad biológica dentro de dichas colecciones. Estas se convierten también, en bancos de alelos que estará disponible en todo momento y a disposición para los futuros programas de mejora genética y conservación.

Las principales y mayores colecciones de yuca *ex situ*, se encuentran en el CIAT-Colombia con aproximadamente 6,500 accesiones, seguida por EMBRAPA-Brasil con aproximadamente 4,000 accesiones. Otros bancos de genes importantes son los de CTCRI-India, INIA-Perú, NRCRI-Nigeria, IITA-Nigeria, IAN-Paraguay, PGRC-Ghana.

Los últimos estudios de diversidad y estructura genética de estas colecciones, se han realizado en base a polimorfismo de un solo nucleótido (SNP, por sus siglas en inglés) por ejemplo, en EMBRAPA y CIAT (de Oliveira *et al*., 2014; Peña-Venegas *et al.,* 2014).Donde se destaca la utilidad y versatilidad de estos marcadores para estos tipos de estudios, sumado a ello las técnicas de secuenciamiento de segunda generación (NGS, por sus siglas en inglés) y el abaratamiento de sus costos, han hecho de los SNPs la herramienta favorita.

En la actualidad, uno de los métodos más usados para la identificación de SNPs es el genotipado por secuenciamiento (GBS, por sus siglas en inglés) (Elshire *et al*., 2011), que proporciona gran cantidad de información genómica, en organismos con genomas grandes como el de las plantas.

La Colección Nacional de Yuca del INIA, cuenta con accesiones colectadas en todo el territorio nacional. Estas han sido caracterizadas físico-química y morfológicamente; sin embargo, carecen de una caracterización a nivel molecular y teniendo en cuenta las

ventajas y utilidad de los SNPs, estos son la mejor opción para estudiar su diversidad y estructura genética.

Los resultados obtenidos en el presente estudio, ayudarán a entender cuán representada está la diversidad genética nacional de yuca en dicha colección y la forma de agrupación genética que presenta; que facilitarán la futura toma de decisiones.

2. MARCO TEÓRICO

2.1. La yuca

2.1.1. Actualidad

Actualmente la yuca se produce en más de 103 países con una producción total de 303 millones de toneladas (t), abarcando más de 27 millones de hectáreas en todo el mundo (FAOSTAT, 2021). Nigeria es el mayor productor, con más de 59 millones de t, seguido por La República Democrática del Congo, con más de 40 millones de t. A nivel de Sudamérica, el mayor productor es Brasil, con 17 millones de t (FAOSTAT, 2021). La mayor producción se realiza en las áreas tropicales y subtropicales, que comprenden entre los 23.5 °N (Trópico de Cáncer) hasta 23.5 °S (Trópico de Capricornio).

En el Perú, su cultivo es muy bajo en comparación a otros países donde también se cultiva, siendo aproximadamente 1 millón de t que abarca cerca de 98 mil hectáreas (FAOSTAT, 2021); la principal producción es realizada por los pueblos amazónicos para autoconsumo, usos tradicionales y con pocas tecnologías.

De los 51 pueblos amazónicos, que habitan el territorio peruano, solo cinco de ellos (Huitoto, Ocaina, Bora, Secoya y Ticuna) cultivan actualmente las diversas variedades, de las denominadas “yuca amarga” y “yuca dulce”, mientras que el resto suele cultivar las variedades de “yuca dulce” (Brañas *et al.*, 2019).

En comparación con los países amazónicos, donde la producción se limita a grupos étnicos bajo un cultivo tradicional-ancestral, o para cubrir demandas rurales (Brañas *et al*., 2019). La yuca es ampliamente cultivada en África Subsahariana, donde la han adoptado como su principal cultivo y fuente de alimento, después de ayudarles a combatir "la hambruna y la langosta" durante los siglos XIX y XX (Parmar *et al*., 2017).

A nivel mundial, el cultivo de yuca es realizado por pequeños y medianos agricultores, con pocos o escasos recursos, ya que no es un cultivo exigente, es fácilmente cultivable, adaptable al estrés hídrico y puede ser cultivada en tierras marginalmente infértiles. Otra de sus características, es su resistencia a plagas y enfermedades, ya que no necesita un control exigente con pesticidas. Presenta ciclos de crecimiento flexibles y, por último, su capacidad de permanecer en el suelo sin cosechar durante largos períodos, ayuda a la racionalización de su consumo.

Por todas estas características se le ha denominado "banco de alimento" de los pobres (Rees *et al*, 2012; Howeler *et al*., 2013). También se ha convertido en el quinto cultivo básico más importante a nivel mundial después del maíz, el arroz, el trigo y la papa en cuanto a producción e ingesta calórica (Lebot, 2009); llegando a ser considerado un cultivo tradicional de seguridad alimentaria (FAO, 2009).

La importancia de la yuca como cultivo comercial está aumentando, se usa en múltiples aplicaciones industriales como la producción de papel, textiles, madera contrachapada, pegamento, biocombustibles, alimentos para animales y bebidas (Kleih *et al*., 2013; Uchechukwu-Agua *et al*., 2015).

2.1.2. Posición taxonómica

La yuca, *M. esculenta* Crantz. pertenece a la familia Euphorbiaceae. El género cuenta con aproximadamente 98 especies (Howeler *et al*., 2013), siendo *M. esculenta* la única cultivada comestible; esta se distribuye desde el norte de México hasta el norte de Argentina.

Se cree que el nombre casava deriva de la palabra Casavi o Cazabi, traducida del arawak (el idioma de los primeros indígenas que vivieron en las Antillas Mayores y estuvieron en contacto con Cristóbal Colón); mientras que manioco y mandioca son derivados de la palabra maniot, en tupi, un idioma extinto, que fue hablado por los nativos tupi de Brasil (Lebot, 2009). En nuestro país es denominada “tuxe” por el pueblo ticuna que se distribuye a lo largo de nuestra Amazonía (Núñez *et al*., 2018).

Su nombre científico, *M. esculenta*, fue acuñado por el botánico Heinrich von Crantz en 1766; sin embargo, posteriormente fue reclasificada, dependiendo de si se trataba de la "yuca dulce" o de la "yuca amarga o brava", asignando nombres diferentes a cada una de ellas: *M. aipi* y *M. utilissima* respectivamente, esto solo basados en su contenido de ácido cianogénico.

Finalmente, en 1938, gracias al trabajo desarrollado por el agrónomo y botánico italiano Raffaele Ciferri, se dio prioridad al trabajo de Crantz y se adoptó su clasificación inicial, que sigue vigente en la actualidad (Brañas *et al*., 2019).

Clasificación taxonómica de la yuca, dada por el CIAT

Reino: Vegetal

División: Spermatophyta

Subdivisión: Angiospermae

Clase: Dycotiledoneae

Subclase: Archichlamydeae

Orden: Euphorbiales

Familia: Euphorbiaceae

Género: *Manihot*

Especie: *Manihot esculenta* Crantz

2.1.3. Centro de origen

Una pregunta muy importante y de larga data es aquella sobre su centro de origen; sabemos que la yuca es originaria del continente americano, pero el lugar específico de su origen fue por mucho tiempo una interrogante importante por resolver.

Los primeros datos acerca de sus primeros usos, fueron dados por los primeros hallazgos arqueológicos, que sugieren que la domesticación de la yuca como cultivo agrícola comenzó hace 5,000–10,000 años en la selva tropical amazónica (Allem, 2002; Howeler *et al.*, 2013; Shigaki, 2016), colocándola en la categoría de uno de los cultivos más antiguos.

Otras evidencias arqueológicas basadas en restos cerámicos de las culturas precolombinas de Colombia, Venezuela y de nuestro país, muestran un rango de antigüedad de 3,000 a 7,000 años (Rouse y Cruxent, 1963). Estos datos sólo evidenciaban y afirmaban el uso de yuca por las diferentes culturas; más no ayudaban a resolver la interrogante del centro de origen.

Una de las principales dificultades para resolver esta interrogante, es que hay pocos caracteres fenotípicos confiables, para inferir relaciones evolutivas dentro del género *Manihot*, ya que muchas especies incluida la yuca, muestran una tremenda variabilidad morfológica intraespecífica.

Antes del advenimiento de los marcadores moleculares, la relación de la yuca con las especies silvestres de *Manihot*, fue un tema de conjetura generalizada (Rogers y Appan, 1973). En ausencia de una única especie progenitora clara, una hipótesis de origen híbrido se convirtió en la explicación preferida del origen (Rogers y Appan, 1973).

En su forma más extrema, esta hipótesis imaginó que la yuca era una "compiloespecie" derivada de complejos de múltiples especies de *Manihot,* que se habían hibridado en todo el “Nuevo Mundo”. Por ejemplo, en la última revisión taxonómica de *Manihot*, Rogers y Appan, (1973) propusieron que la yuca, probablemente tenía múltiples orígenes en América Central y del Sur, y que las especies de América Central probablemente desempeñan un papel clave en la evolución del cultivo.

A partir de la década de 1980, comenzo a surgir la evidencia que cuestiona esta hipótesis, de domesticación de "compiloespecie". Primero, la disponibilidad de marcadores moleculares, llevó a reexaminar la relación evolutiva de la yuca con las especies de *Manihot* de América del Sur frente a las de Centroamérica.

Los análisis de RFLPs (Bertram, 1993; Fregene *et al*., 1994), AFLPs (Roa *et al*., 1997) y secuencias de ADN, indicaron que las especies de *Manihot* de América del Sur y América Central forman dos distintos linajes, y que la yuca está más estrechamente relacionada con el linaje sudamericano.

Con la ayuda de los marcadores moleculares y el conocimiento de nuevas especies del género *Manihot* en Sudamérica, fue posible resolver dos preguntas importantes que conciernen al origen botánico y geográfico. Es decir, con estos marcadores, se pudo

identificar las especies silvestres de las que desciende y, delimitar el área donde el progenitor estuvo disperso en el pasado geológico. Actualmente se sabe que el centro de origen botánico y geográfico es la región noreste de Brasil (Olsen, 2004).

2.2. La Colección Nacional de Yuca del INIA

La Colección Nacional de Yuca del INIA, cuenta con 741 accesiones conservadas bajo condiciones *ex situ.* Toda la colección se encuentra sembrada, en alrededor de 1 ha en el lote N° 9 de la E.E.A Donoso en Huaral.

El mantenimiento en campo de la colección, se realiza mediante una renovación anual de cada una de las accesiones, esta consiste en la siembra de nuevas estacas de yuca obtenidas de la campaña anterior, mediante una propagación clonal (vegetativa). Esto garantiza el mantenimiento del genotipo y fenotipo original, desde su colecta e inclusión a la colección. Las estacas son cuidadosamente seleccionadas, seis repeticiones o clones por accesión y se guardan debidamente rotuladas para su posterior siembra.

La conservación bajo condiciones *in vitro*, se realiza en el Banco Nacional de Germoplasma del INIA, gestionado por la Subdirección de Recursos Genéticos (SDRG). Bajo esta forma, se conservan las 741 accesiones, en tubos de 16 cm x 20 mm una planta por tubo y 4 repeticiones por accesión. La renovación es periódica cada tres meses, mediante la siembra de esquejes axilares de las plántulas, en el medio de cultivo correspondiente. Las condiciones en las que se mantienen son 15°C y un fotoperiodo de 16 horas luz y 8 horas oscuridad.

2.3. Marcadores moleculares usados para estudios de diversidad y estructura genética en *M. esculenta*

Los marcadores moleculares, se han convertido en una de las principales herramientas

para el estudio de la diversidad biológica. Una de las características que los hacen ampliamente útiles, es que usan grupos de loci neutrales, no influenciados por el ambiente y tienen alto poder de discriminación (de Oliveira *et al*., 2014).

Los primeros marcadores moleculares usados para estudiar la diversidad genética en el género *Manihot,* fueron los polimorfismos de longitud de fragmentos de restricción (RFLP, por sus siglas en inglés), la amplificación aleatoria de ADN polimórfico (RAPD, por sus siglas en inglés) (Marmey *et al*., 1993) y posteriormente, para estudiar a *M. esculenta* se usaron los AFLPs (Elias *et al*., 2000).

Sin embargo, los marcadores que fueron ampliamente usados fueron los SSR o microsatélites; por ser altamente informativos, debido a su naturaleza multialélica (alto polimorfismo), su herencia codominante, su distribución uniforme en el genoma (Raji *et al*., 2009) y porque que son fácilmente reproducibles.

Varios sets de microsatélites (SSR) han sido desarrollados en yuca para ser aplicados en diversos estudios, por ejemplo, para entender las relaciones entre la yuca domesticada y sus parientes silvestres. En Colombia, se usaron 10 SSR en 409 accesiones, demostrando la existencia de alta variabilidad dentro del genero, encontrando mayor diversidad en *M. esculenta* y su pariente silvestre más cercano *M. tristis* (Tovar *et al*., 2016).

Los SSR también son usados, para evaluar la diferenciación y la estructura genética dentro de los programas de mejoramiento y caracterización de variedades locales (Asare *et al*., 2011). Por ejemplo, a lo largo del territorio de Brasil, país con el mayor número de variedades de yuca; para caracterizar 51 variedades locales del Estado de Minas Gerais, se usó 20 sets de SSR (Gonçalves *et al*., 2017); en la ciudad de Campo Grande del Estado de Mato Grosso do Sul, se usaron 19 sets de SSR en 60 variedades

locales de yuca (Ferreira *et al*., 2015); en el Municipio de Maringa del Estado de Paraná, se usaron 13 sets de SSR en 66 variedades locales (da Costa *et al*., 2013). Todos estos sets, resultaron altamente polimórficos para estudiar la diversidad y estructura dentro de estas poblaciones.

En Perú también se usaron SSR, por ejemplo, para caracterizar 100 accesiones de yuca de la ciudad de Aguaytia en Ucayali, se usaron 22 sets de SSR (Villasante y Natai, 2006); igualmente, en la cuenca baja del rio Ucayali se caracterizó a 26 variedades locales, usando 12 sets de SSR (Mejía *et al*., 2015). En ambos estudios, resultó que los sets fueron altamente polimórficos, demostrando la alta diversidad existente en estas localidades y dentro de la especie.

También se han usado SSR para estudiar la diversidad global de yuca, entre los bancos de germoplasma del IITA, EMBRAPA y CIAT donde se usarón 30 sets de SSR en 2,300 accesiones mundiales conservadas; logrando separar y caracterizar exitosamente poblaciones de África y América, encontrándose en esta última la mayor diversidad (Hurtado *et al*., 2008).

En los últimos años, los SNPs se han convertido en los marcadores favoritos de varios grupos de investigación; las características que los hacen atractivos son, su abundancia en el genoma, su codominancia y el ser la forma más frecuente de variación genética natural en las poblaciones (Kruglyak *et al*., 1997); además de poder proporcionar variación suficiente para discriminar entre individuos estrechamente relacionados. El desarrollo de las técnicas de secuenciamiento masivo o NGS, han permitido obtener mayor cantidad de SNPs de todo el genoma, de manera masiva y a bajos costos; algunas de estas técnicas son, GBS, ddRAD-seq y sus derivados.

Entre los estudios de diversidad en yuca, usando secuenciamiento masivo, está el

realizado en 53 variedades de yuca de Asia, América y África; donde se identificaron 1,190 SNPs, que ayudaron a diferenciar las poblaciones de los tres continentes (Ferguson *et al*., 2012).

Los SNP también han sido aplicados a la determinación de diferencias entre yuca y sus parientes silvestres, mediante la tecnología DArT, que facilitó el diseño de la primera matriz de SNP y, ayudó en la discriminación de las especies de *Manihot* (Xia *et al*., 2005).

Actualmente, además de los estudios de diversidad y estructura genética, la aplicación de SNPs en yuca, se centra en estudios de selección genómica (GS, por sus siglas en inglés) (Wolfe *et al*., 2016), mapeo de ligamiento, mapeo de locus de rasgos cuantitativos (QTL, por sus siglas en inglés) (Rabbi, 2014; Pootakham *et al*., 2014), identificación de fuentes de resistencia a enfermedades con selección asistida por marcadores (MAS, por sus siglas en inglés) (Carmo *et al*., 2015), identificación de accesiones duplicadas (Rabbi *et al*., 2015), caracterización de germoplasma (Mtunguja *et al*., 2017) y asociación genotipo-fenotipo (GWAS, por sus siglas en inglés) (Zhang *et al*., 2018).

2.4. Genoma de *Manihot esculenta*

El proyecto para secuenciar el genoma de yuca, comenzó en el año 2003 con la formación de la “Asociación Mundial de Yuca”, codirigid
a por el Dr. Claude Fauquet director del “Laboratorio Internacional de Biología de Agricultura Tropical” (ILTAB, por sus siglas en inglés) y por el Dr. Joe Tohme del CIAT.

En el 2009, en asociación con científicos internacionales líderes, presentaron un proyecto piloto al “Programa de Ciencia Comunitaria” (CSP, por sus siglas en inglés) del

Joint Genome Institute del Departamento de Energía de EE. UU. (DOE-JGI) y usando secuenciamiento Sanger, lograron secuenciar casi la mitad del genoma, siendo esta la primera información genómica disponible para yuca.

Sin embargo, el primer borrador del genoma se inició en el año 2009, cuando Steve Rounsley de la Universidad de Arizona y su equipo, asumieron el compromiso del segundo secuenciamiento, que se realizo en la línea endogámica AM560-2 proporcionada por el CIAT usando un secuenciamiento masivo con el "*454 GS FLX-Titanium platform*", generando 61 millones de lecturas, en una y ambas direcciones (*single-end and paired-end*). Estas se ensamblaron junto con las lecturas Sanger del primer proyecto; posteriormente se adicionaron lecturas largas obtenidas de "*454 Life Sciences"* usando su tecnología "FLX+", que mejoró el ensamblado (Prochnik *et al*., 2012).

El genoma de *M. esculenta* secuenciado más reciente (V.6.1), fue completado por Bredeson *et al* (2016), quien usó la misma línea endogámica AM560-2, usando un secuenciamiento de segunda generación *Illumina HiSeq* con lecturas de 250 pb. El genoma ensamblado actual, se encuentra disponible en la base de datos de *Phytozome* en el bioproyecto PRJNA234389, disponible en NCBI y presenta las siguientes características:

- El ensamblaje del genoma principal, tiene un tamaño de 582.25 Mb con 18 cromosomas y más 2,001 *scaffolds* o andamios.
- Aproximadamente 495.48 Mb se encuentran en 40 044 *contigs* (~14.9% está vacío).
- *Scaffold* N50 (L50) = 10 (28.12 Mb).
- *Contig* N50 (L50) = 5,394 (26.69 kb).
- 317 *scaffolds* son de tamaños mayores a 50kb de tamaño, que representa

aproximadamente el 96,2% del genoma.

- El ensamblado contiene 35.9% GC.
- Contiene 33,033 loci que codifican proteínas.

2.5. Genotipado por secuenciamiento (GBS)

La identificación de SNPs, a menudo se realiza mediante la interrupción de los sitios de reconocimiento, al usar endonucleasas de restricción (Wenzl *et al*., 2004), o mediante técnicas acopladas a PCR.

Uno de los primeros métodos que permitió la obtención de SNP a gran escala, fue usar etiquetas o fragmentos cortos de ADN, adyacentes al sitio de restricción (RAD, por sus siglas en inglés). Estas etiquetas eran unidas a una placa de *microarrays* que, por hibridación con los fragmentos a analizar, se lograba detectar de manera paralela miles de marcadores polimórficos, que permitían identificar la variación existente en diferentes organismos (Miller *et al*., 2007). Este método, aunque potente y altamente aplicable, solo puede analizar una fracción del genoma, y está condicionado por el desarrollo de microarreglos para organismos de los cuales existe una amplia información genética y genómica.

Con el desarrollo de las técnicas NGS, las nuevas plataformas de genotipado de manera masiva, se adaptaron para el descubrimiento de SNP. El primer método que usó estas propiedades fue RAD-seq (Baird *et al*., 2008), que al igual que un RAD, se basa en la identificación de SNP asociados a sitios de restricción, pero mediante el secuenciamento de los fragmentos generados por las enzimas usadas.

Este método, permite descubrir SNP en un gran número de individuos, ya que usa la característica del secuenciamiento masivo, paralelo y multiplexado, que ofrecen las

nuevas plataformas de secuenciación *Illumina*. Para ello Baird (2008), usa adaptadores que son ligados a los extremos salientes de los fragmentos generados por el corte de las enzimas de restricción; por un lado, usa un adaptador "P1" que posee un *barcode* o código de barras de 4 - 5 pb (que ayuda a diferenciar los fragmentos de cada individuo), un sitio de reconocimiento para el secuenciamiento *illumina* y otro de reconocimiento para la PCR, previa al secuenciamiento.

Posteriormente, los fragmentos ligados al adaptador "P1" se agrupan, se cortan al azar, se seleccionan por tamaño y se ligan en el otro extremo, a un adaptador P2 en forma de "Y" con extremos divergentes. Esto garantiza que los *primers*, durante la primera ronda de PCR previa al secueciamento, se unan solo al adaptador P1 y garantizar así, que el secueciamiento se realice en todos los fragmentos que contengan el *barcode* (Fig. 1).

Posteriormente, la identificación de SNP mediante genotipado por secuenciamiento, se basó en modificaciones del método RAD-seq, todos ellos tienen en común, que realizan una representación reducida del genoma, usando endonucleasas de restricción resistentes a metilación, que evitan regiones repetitivas del genoma, aumentando la profundidad de secuenciación en regiones con un bajo número de copias. Además, con estas estrategias se puede realizar el mapeo de sus propias lecturas, sin necesitar un genoma de referencia para el llamado de variantes (de Heredia, 2016).

Estas características, fueron ampliamente aprovechadas en organismos con genomas complejos como el de las plantas, donde solo un pequeño grupo cuenta con un genoma secuenciado y caracterizado. El principal problema con el genoma de plantas, es la existencia de secuencias altamente repetitivas y la presencia de transposones, que dificultan el secuenciamiento, y por ende el descubrimiento de SNP. Con el uso de enzimas de restricción, la complejidad disminuye; además de que los sitios de restricción

se mantienen entre especies relacionadas, característica que permite genotipar a grupos diversos de individuos al mismo tiempo.

Por la popularidad de su uso en el estudio de plantas, abordaremos dos métodos (GBS y ddRAD-seq) que ayudaron en el problema de la complejidad de los genomas. Uno de ellos busca aumentar la cantidad de fragmentos obtenidos por enzimas de restricción, disminuir el tiempo en la construcción de las librerías y abaratar los costos, fue GBS (Elshire *et al*., 2011).

El GBS al igual que el método RAD-seq, realiza primero la digestión del genoma usando solo una enzima de restricción resistente a metilación y, a diferencia del método RAD-seq, no realiza un segundo corte a los fragmentos, sino que, se unen directamente a los dos tipos de adaptadores, al que posee el *barcode* (de 4-8 pb más el segmento de reconocimiento para la PCR) y al adaptador común en forma de "Y". El secuenciamiento solo se realizará en las librerías que contengan fragmentos, que han unido a ambos adaptadores en sus extremos.

La ventaja del GBS, es que aumenta la probabilidad de obtener mayor cantidad de SNP, al aumentar la posibilidad de secuenciar fragmentos de diferentes tamaños. Esta técnica ha sido utilizada principalmente, en especies con interés agronómico, también ha sido usada en algunas especies silvestres (Chen *et al*., 2013; Schilling *et al*., 2014; Ilut *et al*., 2015)

Probablemente, la metodología con más potencial para el estudio de la diversidad biológica, sea ddRAD-seq (Peterson *et al*., 2012) que se basa en la doble digestión del genoma. Este usa la secuenciación de lecturas cortas en la plataforma Illumina, ya que prefiere loci no ligados; es decir, el método prefiere obtener muchos loci con pocos polimorfismos, que pocos loci con muchos polimorfismos.

El ddRad-seq inicia, con la doble digestión del genoma, usando dos endonucleasas; posteriormente realiza una selección de los fragmentos por tamaño y por presencia en sus extremos de ambos cortes de restricción. Al igual que el método RAD-seq, utiliza dos adaptadores, el común en forma de “Y” y el que contiene el *barcode*.

Como se describió, las técnicas GBS y ddRAD-seq comparten varios pasos básicos, las diferencias que se pueden recalcar se encuentran: en el orden o detalles de la digestión enzimática, la ligadura de los adaptadores, el *barcode* y la selección del tamaño de los fragmentos (Yang *et al*., 2016).

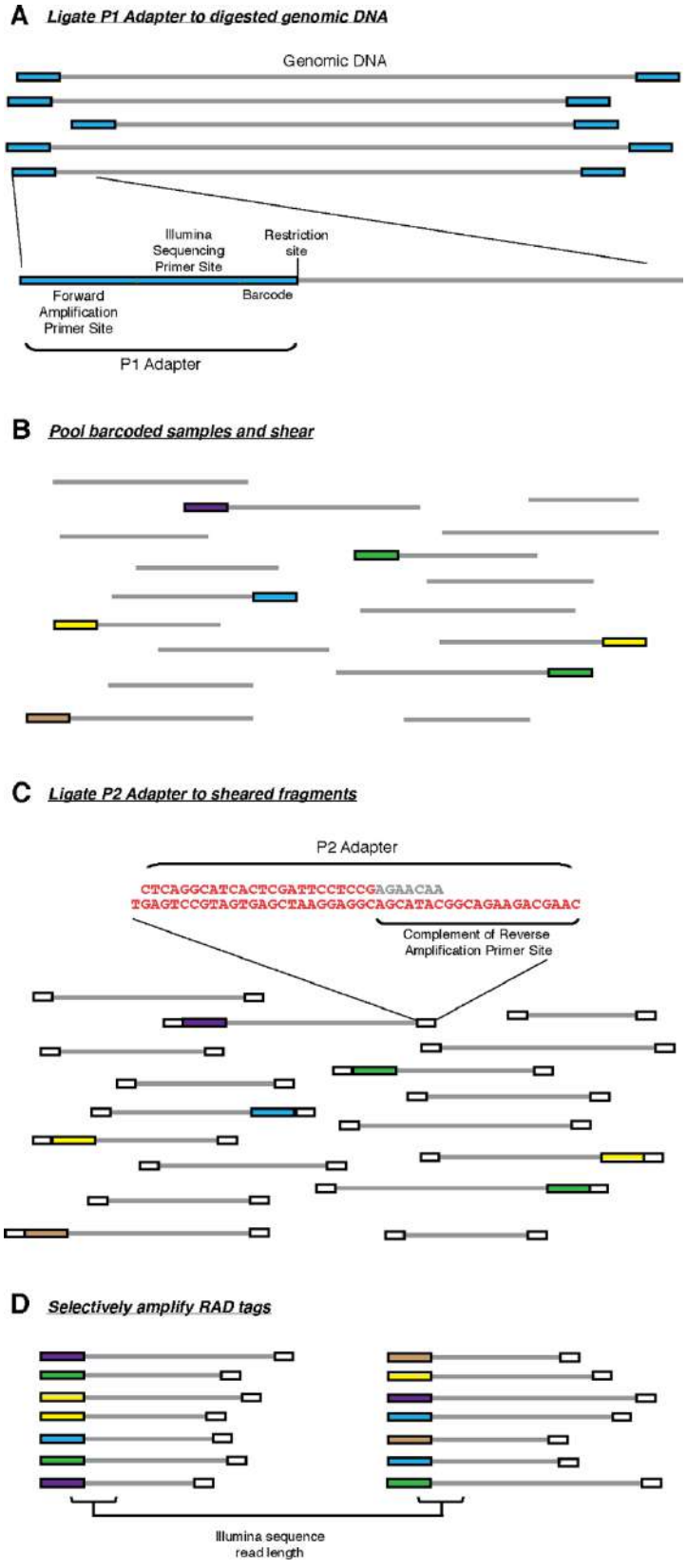


Figura 1. Generación de un marcador RAD

Fuente: Baird *et al.*, 2008.

2.6. Conceptos de genética poblacional

2.6.1. Diversidad Genética

Es la variación existente en el material hereditario, entre dos o más organismos que conforman poblaciones de determinada especie o grupos de especies, estas variaciones van desde los componentes fundamentales (los ácidos nucleicos) que forman las cadenas de ADN, hasta el nivel de genes, cromosomas e incluso el genoma completo (Dominguez *et al*., 2019).

El convenio sobre la Diversidad Biológica de 1992, define a la diversidad genética como componente vital de la diversidad, que debe ser conservada y utilizada equitativa y sustentablemente por los pueblos del mundo (Glowka, 1996).

La diversidad genética se produce debido a tres procesos fundamentales: mutación, recombinación de material genético y migración de individuos de una población a otra. Las mutaciones son las alteraciones heredables en la secuencia del ADN, en la estructura o en el número de cromosomas (Lewin, 2005); ocurren por los errores cometidos por la maquinaria celular, que reparan o copian el material genético en las células; sin embargo, en la imperfección de dichos mecanismos, radica la posibilidad de que la vida haya evolucionado y sea tan impresionantemente diversa como lo es hoy (Futuyma, 2005).

2.6.2. Estructura genética poblacional

Se define como la caracterización jerárquica, en base a la cantidad y distribución de la variación genética que existe entre individuos dentro de grupos y entre grupos que conforman una población; estos grupos o subpoblaciones, presentan frecuencias propias de sus variantes genéticas (alelos).

Las primeras conceptualizaciones matemáticas del estudio de la estructura genética de las poblaciones, se le atribuye a Wright (1921, 1931).

2.6.3. Principio de Hardy-Weinberg

El principio establece que la variación genética de una población se mantiene constante en el tiempo, generación tras generación; para ello las frecuencias genotípicas y alélicas deben permanecer constantes a lo largo del tiempo, bajo la situación idealizada donde:

- Los individuos son diploides.
- La reproducción es sexual y aleatoria.
- La población es grande y las generaciones no se superponen.
- No existe selección natural, deriva génica, migraciones ni mutaciones.

Las frecuencias genotípicas y alélicas están relacionadas, por las siguientes ecuaciones simples:

- p^2 es la frecuencia del genotipo A1A1 (homocigoto dominante).

$$f(A1A1) = p^2$$

- 2pq es la frecuencia del genotipo A1A2 (heterocigoto).

$$f(A1A2) = 2pq$$

- q^2 es la frecuencia del genotipo A2A2 (homocigoto recesivo).

$$f(A2A2) = q^2$$

2.6.4. Frecuencias alélicas

Se define como alelo, a la forma alternativa que puede presentar un gen, en el mismo locus genético de un cromosoma. La frecuencia alélica, es la proporción en la que se

encuentran los distintos alelos de un determinado locus en una población; se calcula dividiendo el número de veces que se observa el alelo de interés en la población, por el número total de copias de los alelos del mismo locus genético.

La frecuencia alélica, nos ayuda a entender diferentes cambios que puede estar atravesando una población, provocadas por procesos capaces de alterarlas como, la deriva genética, migraciones o la introducción de nuevas mutaciones en dicha población.

2.7. Estadísticos para el estudio de la diversidad genética

Existen varios estadísticos propuestos para cuantificar y entender la variabilidad genética de una población en términos más exactos. Cada uno de los cuales, describe diferentes aspectos de esta variabilidad y, se suelen usar de manera complementaria.

Los más empleados al usar marcadores genéticos codominantes son, la proporción de loci polimórficos (P), el contenido de información polimórfica (PIC), la heterocigosidad observada (Ho) y la heterocigosidad esperada (He).

2.7.1. Proporción de loci polimórficos (P)

Un polimorfismo genético, es la ocurrencia en la misma población de dos o más alelos en un locus, cada uno con frecuencia apreciable (Cavalli-Sforza y Bodmer, 1971). El valor de P, es la medida de estos loci variables, se calcula dividiendo el número de loci polimórficos, entre el número total de loci analizados en la población (Sosa *et al*., 2002).

$$P = \frac{N°delocipolimórficos}{N°delocitotales}$$

Para determinar si un locus es polimórfico, se deben establecer criterios de acuerdo a sus frecuencias alélicas; esto depende mucho del tamaño de la población ya que

mientras más grande sea, habrá más probabilidad de encontrar nuevos alelos.

El criterio más frecuente, para determinar el polimorfismo de un locus, suele fijarse de acuerdo a la frecuencia del alelo mayoritario, que debe ser mayor o igual que 0.95 o 0.99 (Cavalli-Sforza y Bodmer., 1971). Sin embargo, los criterios suelen cambiar de acuerdo a los autores, por ello este índice es considerado muy subjetivo (Gonçalves, 2017).

2.7.2. Heterocigosidad

La mejor valoración de la variación genética de la población es la heterocigosidad, esta es la frecuencia media de individuos heterocigotos por locus (Martínez y Vega, 2002). Los individuos heterocigotos son importantes, ya que albergan los diferentes alelos en la población y son la evidencia de la variabilidad existente.
Existen dos tipos de heterocigosidad:

Heterocigosidad esperada (He)

Es una medida de la diversidad genética, cuando una población cuenta con "n" alelos para un locus y está bajo equilibrio Hardy-Weinberg (Nei, 1973). Cuando esto pasa, la He se calcula usando las frecuencias alélicas.

Se pueden realizar tres niveles de cálculo:

- Cálculo de la He para un locus dado con dos alelos:

$$h_j = 1 - p^2 - q^2$$

- Cálculo de la He para un locus "j" con "i" alelos:

$$h_j = 1 - \sum p_i^2$$

- Cálculo de la He para varios loci (promedio de varios loci):

$$H = \frac{\sum_j^L h_j}{L}$$

Donde:

h_j = heterocigosidad por locus

"p" y "q" = frecuencias alélicas

p_i = frecuencia del alelo i

H = heterocigosidad para varios loci

L = número total de loci

Heterocigosidad observada (Ho)

Es la proporción de individuos heterocigotos encontrados en la población. Se calcula usando las frecuencias alélicas, y matemáticamente se define como la probabilidad de que, al elegir dos alelos al azar para un determinado locus, estos sean diferentes. Varias poblaciones pueden tener el mismo valor de Ho, pero distinta He y es una medida para comprobar en qué grado la población se encuentra o no en equilibrio Hardy-Weinberg.

2.7.3. Contenido de información polimórfica (PIC)

Este estadístico, nos ayuda a identificar el nivel de polimorfismo para un marcador, y sus valores dependen del número de alelos que posea y de sus frecuencias alélicas (Botstein, 1980). Los valores de PIC ayudan en la selección de marcadores, porque miden la calidad de estos para estudios de diversidad genética, ya que representará el polimorfismo detectado en toda la población.

La fórmula para detectar el PIC, para marcadores dominantes es:

$$PIC = 1 - \sum p_i^2 - \sum 2p_i^2\, p_j^2$$

p_i = frecuencia del alelo i.

p_j = frecuencia del alelo j.

El máximo valor de PIC para marcadores dominantes es 0.5 (Chesnokov y Artemyeva, 2015), para marcadores SSR, los valores de PIC superiores a 0.5 son informativos, los comprendidos entre 0.25 y 0.5 son medianamente informativos (Botstein, 1980) y poco informativos, los inferiores a 0.25 (Zamorano, 1998).

2.8. Diferenciación genética entre poblaciones

La subdivisión de una población, surge como resultado de la restricción del flujo genético, debido a factores externos como barreras ambientales o distancias geográficas, que generan alteración en el equilibrio de una población panmictica. Resultando en subpoblaciones o grupos de poblaciones con diferentes frecuencias alélicas, provocando disminución de la diversidad genética por deriva, donde existirá mayor fijación de alelos y menor heterocigosidad que la población original (Uribe, 2014).

2.8.1. Estadísticos F

También llamados índices de fijación, ayudan a describir y entender la partición de la diversidad genética, a diferentes niveles jerárquicos de subdivisión en las poblaciones; según las frecuencias génicas se definen tres estadísticos (Wright, 1965):

F_{IS}: llamado también coeficiente de endogamia, mide la reducción en la heterocigosidad individual entre individuos de una población local, debido a las desviaciones producidas por el entrecruzamiento al azar (Wright, 1965).

F_{IT}: es el coeficiente de endogamia entre individuos, en relación a la población total (Wright, 1965).

F_{ST}: es el índice de fijación entre poblaciones, mide el grado de diferenciación genética entre estas.

Los valores de Fst, son muy apropiados para evaluar diferencias entre poblaciones, que han sufrido procesos evolutivos a corto plazo, sobre todo si el tamaño efectivo de las poblaciones varia en el tiempo. El valor es cero, cuando todas las poblaciones tienen la misma frecuencia genética y es 1 cuando las subpoblaciones están fijadas para distintos alelos (Murray, 1996).

2.8.2. Distancias genéticas

La distancia genética, se refiere al grado de diferencia que existe entre dos individuos que forman parte de la población, es decir; nos permite cuantificar el grado en que difiere un individuo “A” de “B”, “B” de “C” y “A” de “C” de todo el conjunto de datos, a diferencia del estadístico Fst, que miden el grado de varianza que existe en las poblaciones. Podemos calcular estas diferencias usando dos métodos: basados en distancias geométricas o basadas en modelos biológicos.

Distancias geométricas: reflejan la similitud entre dos poblaciones respecto a sus frecuencias alélicas, analizándolas como un conjunto de números sin asumir *a priori* algún supuesto evolutivo. Aquí pertenecen la distancia de “cuerdas o acordes” (Cavalli-Sforza y Edwards, 1967) y la distancia de Roger (Roger, 1972).

Estas medidas, son una visión geométrica de la distancia entre puntos en una híper-espera, cuyo número de dimensiones, es igual al número de alelos considerados en la comparación (Demarchi, 2009). Con los valores de distancias, se crean matrices que son ampliamente usadas para la construcción de árboles filogenéticos, como UPGMA, Fitch-Margoliash y Neighbor-joining.

Distancias basadas en supuestos biológicos: son aquellas que asumen un modelo evolutivo, donde la variación es atribuida principalmente a la acción de la mutación y deriva genica; asumiendo además que los marcadores usados no están sujetos a

selección natural.

Estas distancias se ajustan al modelo de alelos infinitos (MAI, por sus siglas en inglés) según el cual, cada mutación crea aleatoriamente un nuevo alelo, el cual es un nuevo estado que está relacionado a cualquier otro por igual (Demarchi, 2009). Aquí encontramos la distancia estándar "D_A" (Nei, 1972), que considera que la tasa de mutación es igual y constante para todos los loci, las poblaciones presentan un tamaño efectivo constante, de generación en generación y están en equilibrio mutación-deriva.

También encontramos la distancia de Reynolds (Reynolds *et al*., 1983), que supone la ausencia de mutaciones, y asume que la deriva genética es la única fuerza que actúa sobre las frecuencias alélicas; esta distancia se usa para comparar poblaciones pequeñas, que han divergido hace un número relativamente bajo de generaciones (Weir, 1996).

2.9. Pruebas de asignación

En base a la información de diversidad genética que existe en una población en diferentes niveles, las pruebas de asignación nos permiten determinar a qué subpoblación pertenece un determinado individuo o conjunto de individuos. Para calcular esta medida se usa dos métodos: de clasificación y los de asignación propiamente dichos (Garrido-Garduno *et al*., 2013).

2.9.1. Métodos de clasificación

Son aquellos que asignan a los individuos a categorías o grupos predefinidos (Manel *et al*., 2005); aquí encontramos al análisis de componentes principales y al análisis de discriminantes.

Análisis de componentes principales (PCA): ayuda a resumir la variación existente,

cuando analizamos una gran cantidad de loci. Este análisis muestra las tendencias de similitud entre las variables o factores evaluados, mas no define o asigna grupos específicos; por ello requiere que se definan grupos *a priori*. Es un buen método para iniciar a explorar nuestros datos y ver un posible patrón de estructura genética (Manel *et al*., 2003).

Análisis de discriminantes (DA): este método permite analizar la variación genética existente en la población, minimizando la variación entre los individuos, para maximizar la diferencia entre los grupos; es decir, divide la variación total de la población en dos niveles, entre los individuos y entre los grupos.

A diferencia del PCA el análisis de discriminantes, logra obtener la mejor separación de los individuos dentro de los grupos predefinidos, dando además un valor de asignación probabilístico de los individuos a cada grupo. Este método realiza un análisis previo de PCA, para asegurar que las variables no estén correlacionadas (Jombar *et al*., 2010).

2.9.2. Métodos de asignación

Este método ayuda a determinar, cuan indicativo es el genotipo de un individuo, para la población de la cual fue muestreado; es decir, permite estimar la probabilidad de que un individuo pertenezca a un grupo o población, en base a su genotipo y la frecuencia alélica de los loci de los individuos en diferentes poblaciones (Dawson y Belkhir, 2001).

Algunos de sus algoritmos, asumen que la población se encuentra en equilibrio Hardy-Weinberg y equilibrio de ligamiento (Pritchard *et al*., 2000) y otros son más flexibles y permiten usar datos, que no cumplen con estos supuestos (Falush *et al*., 2003). Los más usados son los algoritmos Bayesianos, que se aplican en poblaciones que no cumplen con los supuestos de equilibrio.

Análisis de asignación Bayesiana: para asignar a los individuos a poblaciones, solo se basa en su información genética, más no en grupos definidos *a priori*, la asignación se realiza aleatoriamente usando cadenas de Markov tipo Monte Carlo (MCMC). Los programas que se basan en este tipo de análisis son STRUCTURE (Pritchard *et al*., 2000), PARTITION (Dawson y Belkhir, 2001) y BAPS (Corander *et al*., 2004).

3. HIPÓTESIS Y OBJETIVOS

3.1. Hipótesis

Existe alta diversidad genética en la Colección Nacional de Yuca del INIA y se puede identificar grupos genéticos en base a su procedencia u origen geográfico.

3.2. Objetivos

3.2.1. Objetivo general

Determinar la diversidad y estructura genética en 731 accesiones de la Colección Nacional de Yuca del INIA; usando SNP como marcadores moleculares.

3.2.2. Objetivos específicos

- Realizar un genotipado por secuenciamiento de cada una de las 731 accesiones de la Colección Nacional.
- Estudiar la diversidad y estructura genética de las 731 accesiones que forman parte de la Colección Nacional, provenientes de diferentes países y regiones de Perú.

4. MATERIALES

4.1. Material biológico

Hojas jóvenes de las 731 acciones de la Colección Nacional de Yuca del INIA,

conservadas *ex situ* en la E. E. A Donoso Huaral, provenientes de Brasil, Bolivia, Paraguay y Perú. En Perú, de las siguientes regiones: Amazonas, Ancash, Apurímac, Arequipa, Cajamarca, Cusco, Huánuco, Ica, Junín, La Libertad, Lambayeque, Lima, Loreto, Piura, San Martín, Tumbes y Ucayali, cuya distribución geográfica a nivel nacional se muestra en el Anexo 1.

Tabla 1. Número de accesiones por país y región peruanas, utilizadas en el estudio de diversidad y estructura genética.

País / Región	**N° de accesiones**	**País / Región**	**N° de accesiones**
Brasil	47	Amazonas	10
Bolivia	24	Junín	10
Paraguay	23	Piura	10
Ucayali	319	Cusco	9
Lima	106	Ancash	7
San Martín	80	Tumbes	7
Huánuco	21	Apurímac	5
La Libertad	21	Cajamarca	5
Lambayeque	12	Arequipa	3
Loreto	11	Ica	1

4.2. Material de laboratorio

Todos los materiales e insumos de laboratorio fueron financiados por el proyecto PNIA 230_PI a cargo de la SDRG del INIA. Los equipos, materiales y reactivos empleados, así como los lugares donde se llevaron a cabo las diferentes etapas del presente estudio, son incluidos en la explicación de los métodos.

5. MÉTODOS

Se inició con la siembra en campo de las estacas de las 731 accesiones de yuca, en la

E. E. A Donoso en Huaral. La toma de muestras se realizó, cuando aparecieron los primeros cogollos jóvenes, e inmediatamente se procedió a la extracción del material genético, en las instalaciones del Laboratorio de Biología Molecular y Genómica de la sede central del INIA-La Molina/Lima.

El ADN genómico de cada accesión se envió al Centro Genómico de la Universidad de Minnesota, donde se realizó el GBS. Los resultados del secuenciamiento, fueron procesados en el área bioinformática del Laboratorio de Valoración del INIA – sede central.

5.1. Obtención del material genético

5.1.1. Toma de muestras en la E.E.A Donoso en Huaral

Se realizó usando dos métodos: 1) denominado en fresco, que consistió en la colecta directa de hojas pequeñas de cogollos jóvenes en buffer CTAB 2X con β-mercaptoetanol (Doyle y Doyle, 1987) y 2) denominado colecta en sílica gel, para la obtención del material vegetal en seco, como duplicado de seguridad en caso que el material en fresco se pierda durante el traslado o procesamiento, o la calidad y cantidad final de ADN genómico no cumpla con lo requerido.

Los cogollos de las plantas jóvenes, fueron tomados en orden ascendente de acuerdo al mapa de campo. Ambos tipos de muestreo se realizaron al mismo tiempo, teniendo en cuenta que la hoja colectada en fresco en CTAB 2X pertenecía al mismo cogollo colectado, como duplicado de seguridad en sílica gel.

Colecta de hojas en fresco (Morillo y Miño, 2011)

Con este método se buscó que el ADN genómico; sea rápidamente procesado en laboratorio para evitar la oxidación del material vegetal y que sea de buena cantidad y calidad, por ello se usó tejido vegetal joven, que contiene más células por unidad de

volumen, y menos polisacáridos y polifenoles, que dificulten el proceso de extracción (Velázquez *et al*., 2014).

Se preparó 1 L de CTAB-2X (Anexo 2), dos días antes de la colecta y se dispensó en 800 microtubos estériles de 2 mL, cada uno con 700 µl de CTAB-2X + 2 µL de β-mercaptoetanol y se guardaron en crioboxes a -80°C, hasta el día de la toma de muestra.

Durante la toma de muestra se procedió de la siguiente manera: 1) se enumeró cada uno de los microtubos del criobox de izquierda a derecha, 2) con guantes de nitrilo y con ayuda de pinzas, se tomó una hoja joven entre 3-4 cm de los cogollos superiores, 3) se colocó en el microtubo de 2 mL correspondiente, 4) nos aseguramos que se encuentre bien cerrado y se retornó a su lugar correspondiente 5) terminado un criobox, se guardó inmediatamente en un cooler para mantener el buffer frío. Se siguió el mismo procedimiento hasta completar todas las accesiones.

En el laboratorio, todo el material colectado se guardó en una ultra-congeladora a -80°C hasta el momento de su procesamiento.

Colecta en sílica gel (Chase y Hills, 1991; Morillo y Miño, 2011)

Se mantuvo el mismo orden usado para el método en fresco: 1) se tomó entre 2 a 3 hojas jóvenes de los cogollos correspondientes, 2) se envolvieron cuidadosamente en papel toalla, 3) se rotularon con su respectivo código, 4) se guardó en bolsas ziploc con sílica gel, 20 por cada bolsa.

En el laboratorio se guardaron en un lugar seco y fresco. El cambio de sílica gel se realizó a los cuatro días al observar el cambio de color, y se siguió realizando hasta que el tejido vegetal se haya secado completamente.

5.1.2. Extracción de ADN genómico

El proceso de extracción se realizó usando el método del CTAB 2X y 10X, protocolo de extracción modificado de Doyle y Doyle (1987), el cual fue estandarizado anteriormente para yuca. Los protocolos de preparación de los reactivos usados se encuentran en el Anexo 2.

Preparación de las muestras y el área de trabajo

Se procesó diariamente como máximo 60 muestras. Para iniciar la extracción, se dejó descongelar las muestras guardadas a -80°C, de la toma de muestra en campo. Se preparó el área de extracción con el material correspondiente y el baño María a 65°C.

Extracción de ADN genómico (modificado de Doyle y Doyle, 1987)

1) A cada una de las muestras descongeladas, se agregó 2 µl de β-mercaptoetanol y 2 billas de metal, 2) las muestras se llevaron a moler en el triturador o Tissu-Lyser II a una frecuencia de 30/s por 4 min, 3) se pasó a incubar en baño maría a 65°C por 45 min, agitando suavemente cada 15 min, 4) se retiró los microtubos del baño María, se dejaron enfriar por 5 min y se agregó 700 µL de cloroformo: alcohol isoamílico (24:1) frío, 5) se mezclaron por 1 min, invirtiendo el portatubos, 6) se pasó a centrifugar a 14 000 rpm por 10 min, 7) se retiró cerca de 700 µL del sobrenadante formado y se colocó en un microtubo de 1,5 mL, teniendo en cuenta el orden y el rotulado con el código correspondiente 8) se agregó 80 µL de CTAB-10X y se llevaron a baño María a 65°C por 5 min 9) se agregó 700 µL de cloroformo:isoamílico (24:1) frío y mezcló por 1 min, 10) se centrifugó a 14 000 rpm durante 10 min, 11) se retiró cerca de 700 µL de la capa acuosa superior y se colocó en un microtubo de 1.5 mL, 12) se agregó 500 µL de isopropanol frío y mezclo con suavidad por 1 min invirtiendo el portatubos, 13) se dejó a -20°C durante 1 h, 14) se centrifugó a 14 000 rpm durante 10 min, 15) se descartó el sobrenadante teniendo cuidado de no perder el pellet formado, 16) se agregó 500 µL de

etanol de 70°, se mezcló por 1 min dando pequeños y suaves golpes para desprender el pellet, 17) se centrifugó a 14 000 rpm durante 5 min, se repitió este paso dos veces, 18) se descartó el etanol y se dejó secar el tubo abierto e invertido a temperatura ambiente por 5 min sobre papel toalla, 19) se llevaron los tubos a incubar a 37 ºC durante 15 min o se usó el concentrador para evaporar el etanol restante a 60°C por 10 min, 20) se agregó 150 µL de buffer TE 1X y se dejó a 4°C para diluir el pellet, 21) al siguiente día se agregó 2 µL de ARNasa [10 mg/ml] y se llevó a incubar a 37 ºC durante 60 min, agitando suavemente cada 15 min.

5.1.3. Verificación de calidad del ADN genómico

La electroforesis se realizó por cada 60 muestras, en una cámara electroforética horizontal, con buffer TAE 1X (Tris, Ácido acético glacial y EDTA). Se preparó el buffer de carga, en la siguiente proporción 6:3:0.03 (agua nivel molecular: SALB 2X: gel red), se mezcló con el ADN genómico en la proporción 1:9.

Se usó un gel de agarosa al 1%. Las condiciones de la corrida electroforética fueron: 120 V por 25 minutos; el revelado se realizó usando un fotodocumentador ChemidocTM XRS+ de Bio-rad.

5.1.4. Cuantificación del ADN genómico

Como requisito de envío para el genotipado por secuenciamiento, el ADN genómico tuvo que ser cuantificado usando un método fluorométrico, por ello se usó un Qubit 3.0. que garantizó la cuantificación de ADN integro de doble cadena. Las especificaciones de calidad y cantidad, requeridos por el Centro Genómico de la Universidad de Minnesota se muestran en el Anexo 3.

La preparación de los reactivos y la cuantificación, se realizaron en base a la guía de usuario *Qubit dsDNA BR Assay Catalog nos. Q32850 Q32853* y el uso del equipo, se

realizó en base a la guía de *Qubit 3.0 Fluorometer Catalog Number Q33216*.

5.2. Genotipado por secuenciamiento

El genotipado de las 731 accesiones se realizó en el Centro Genómico de la Universidad de Minnesota, usando el método ddRAD-seq basado en el propuesto por Peterson *et al*. (2012). Para la construcción de librerías se usó el juego de endonucleasas de restricción TaqI + BtgI; la elección se hizo en base a un proyecto piloto previo, que permitió garantizar una cobertura entre 6-7X del genoma y una cantidad mayor a 5,000 SNPs.

Se realizó un secuenciamiento multiplexado *single-ends* (en una dirección), usando dos carriles de una celda de flujo *NovaSeq Illumina* 1x100-pb. Finalmente, las lecturas de cada individuo fueron separadas en función de la combinación de su código de barras y su índice de multiplexación asignados.

5.3. Procesamiento bioinformático

Para el procesamiento de la data, se usó el servidor de la SDRG (HP ProLiant ML 350 GEN10), que cuenta con las siguientes carácterísticas indispensables para el oricesamiento: sistema operativo Ubuntu 16.04 server, 2TB de espacio para almacenamiento y 256 GB de RAM.

Las secuencias resultantes del genotipado, fueron descargadas directamente del servidor de la Universidad de Minnesota, a través de un acceso confidencial. Las lecturas se encontraban demultiplexadas, con un valor de Phred mayor a 33; por ello se omitió los pasos de demultiplexación y filtrado de calidad de lecturas.

5.3.1. Alineamiento de lecturas

El alineamiento de lecturas se realizó para cada accesión, usando el genoma de referencia de *M. esculenta* V.6.1 disponible en NCBI (ID:441). El alineador usado fue BWA versión 0.7.12 (Borrow-Willer Alignmet) (Li y Durbin, 2009); para mapear las lecturas contra el genoma de referencia se aplicó el algoritmo BWA-MEN, usado para lecturas *Illumina* de tamaño entre 70 pb – 1 Mb de alta calidad.

5.3.2. Filtrado de lecturas

El resultado del alineamiento, fueron lecturas alineadas (mapeadas) y no alineadas al genoma de referencia, estas se exportaron en un archivo con formato SAM; a partir de este, se realizó el filtrado solo de lecturas mapeadas usando la herramienta *flagstat* del programa SAMTOOLS versión 0.1.19 (Li *et al*., 2009).

5.3.3. Llamada de variantes (SNP calling)

Antes de realizar la llamada de variantes, los archivos SAM, se convirtieron a formato BAM y se ordenaron usando el programa SAMTOOLS.

Para la llamada de variantes y el genotipado se usó la herramienta *gstacks*, del programa STACKS versión 2.53 (Catchen *et al*., 2011); usando los siguientes parámetros: a) el modelo para la llamada de variantes y genotipado fue el de *marukilow*, b) el *umbral - alpha* mínimo para considerar una variante como SNP en la población fue de 0.05, c) el *umbral – alpha* para asignar el genotipo a cada individuo fue de 0.05, d) las demás opciones se dejaron por defecto.

Finalmente se obtuvo el archivo en formato de llamada de variantes (VCF, por sus siglas en inglés), que contiene el genotipado con las variantes para cada accesión, para ello se usó la herramienta *populations* del mismo programa STACKS.

5.3.4. Filtrado de variantes

Se realizaron tres filtros, el primero se aplicó para eliminar *loci* o sitios con más del 20% de data faltante, el segundo se realizó para eliminar a las accesiones que tenían más del 55% de data faltante y el tercero, se usó para remover *loci* con una frecuencia de alelo menor (MAF, por sus siglas en inglés) al 10%, finalmente se obtuvo un nuevo archivo VCF, con los SNP filtrados de alta calidad. Para todos los filtros se usó el programa VCFTOOLS versión 0.1.17 (Danecek *et al*., 2011).

5.3.5. Análisis de diversidad

Del resultado del análisis de filtrado de SNPs, se identificó a tres accesiones que no presentaban sufientes datos (missing data), por lo cual el análisis de diversidad se realizó en 728 accesiones de la Colección Nacional.

En base a las frecuencias alélicas de los SNPs, se calculó: a) el valor de MAF; b) la heterocigosidad observada (Ho) y esperada (He) calculada en base a Nei (1973); d) el contenido de información polimorfica (PIC), calculada de acuerdo a Botstein *et al*., (1980); f) el equilibrio de Hardy-Weinberg por locus con *p-value* < 0.05, usando la prueba de chi-cuadrado (X^2).

Todos los parámetros fueron obtenidos en el programa estadístico R versión 3.6.3 (R Development Core Team, 2020), usando los paquetes: *adegenet* versión 2.1.2, *ape* versión 5.3, *vcfR* versión 1.10.0. *hierfsat* versión 0.04-22, *ggplot2* versión 3.3.0. *pegas* versión 0.13 y *snpReady* versión 0.9.6.

5.3.6. Análisis de estructura poblacional

De la misma manera que en el análisis de diversidad, se usarón las 728 accesiones resultantes del filtrado de SNPs.

Se usó los estadísticos “F” de Wright (1965), posteriormente se usó tres análisis complementarios, el análisis molecular de varianza (AMOVA), el análisis de agrupación de componentes principales (PCA) y el análisis discriminante de componentes principales (DAPC). Para el análisis AMOVA, se usó el paquete *poppr* versión 2.4.8, el análisis PCA se realizó con la función *glPca* del paquete *adegenet*. Todos los parámetros fueron obtenidos en R versión 3.6.3 (R Development Core Team, 2020).

Para el análisis DAPC, se realizó primero un PCA para asegurar que los componentes principales (CP) seleccionados, no se encuentren correlacionados y contengan la mayor cantidad de varianza de la población; para ello se usó la función *optim.a.socre* del paquete *adegenet*. Esta transformación permitió al análisis de discrininantes (AD), minimizar la diferencia entre los individuos y maximizar la diferencia entre los grupos (Jombart *et al*., 2010).

Para completar el análisis, se definió los grupos *a priori* o valores “K” de agrupamiento, para ello se usó el criterio *smoothNgoesup* de la función *find.cluster* del paquete *adegenet.* El mejor valor de “K”, se seleccionó en base al Criterio de Información Bayesiana (BIC) (Jombart *et al*., 2010).

6. RESULTADOS

6.1. Genotipado por secuenciamiento de la Colección Nacional de Yuca.

6.1.1. Genotipado por secuenciamiento

En promedio se obtuvo 1,969,279 *reads* o lecturas por accesión, siendo el valor mínimo 516,471 (accesión ME_264) y el máximo 8,907,810 (accesión ME_005). La Fig. 2 muestra la cantidad de lecturas por accesión, ordenadas de mayor a menor; observamos que existe una amplia distribución de la cantidad de lecturas por accesión, reflejando que la cantidad no fue uniforme para todas las accesiones de la colección.

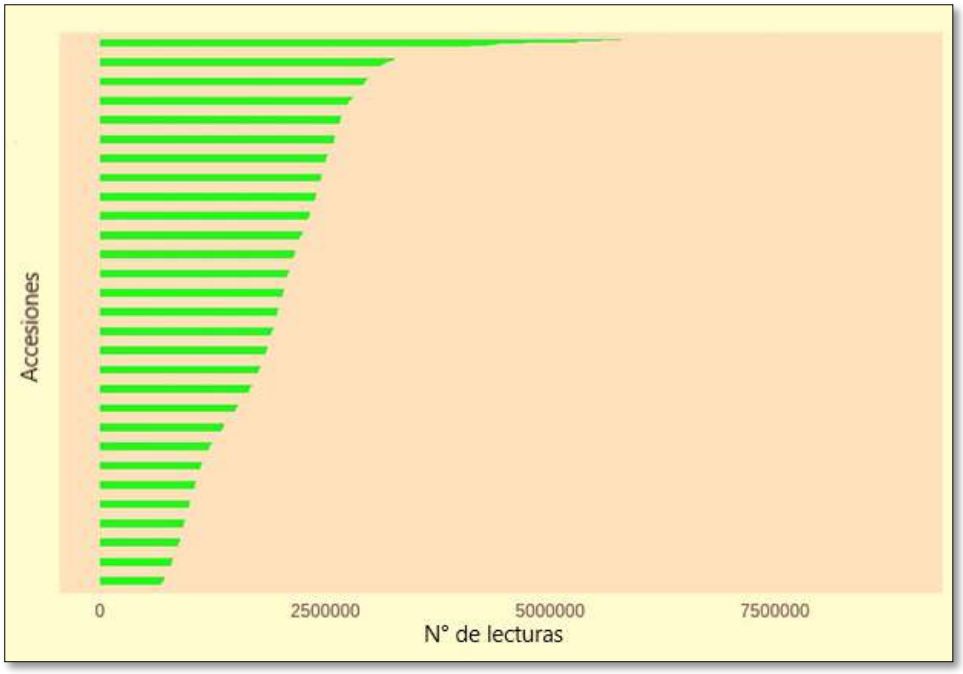


Figura 2. Barplot del número de lecturas (*reads*) para las 731 accesiones, ordenados de mayor a menor.

En la Fig. 3, se muestra la frecuencia de accesiones de acuerdo a su número de lecturas, vemos que la mayor cantidad de accesiones posee entre 1-3 millones de lecturas.

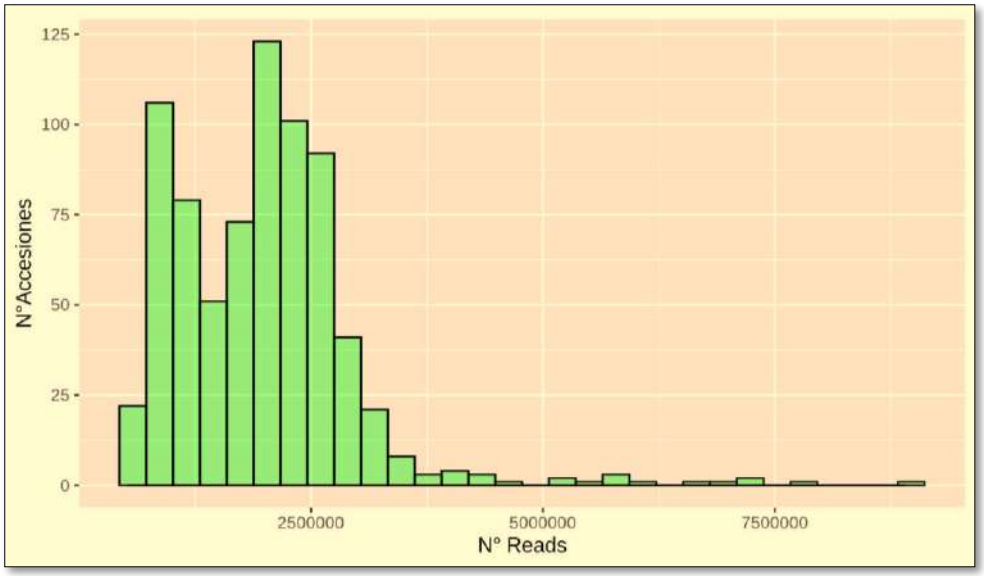


Figura 3. Frecuencia de accesiones de acuerdo al número de *reads*.

6.1.2. Alineamiento de *reads*

Los *reads* de las 731 accesiones, presentaron en promedio 94.33 % de alineamiento al genoma de referencia, siendo el menor valor 80.75 % (accesión ME_629) y el mayor valor 96.39 % (accesión ME_561). En la Fig. 4 se observa que la mayoría de las accesiones presentaron un alto porcentaje de alineamiento, los valores se concentran entre 93% y 96%.

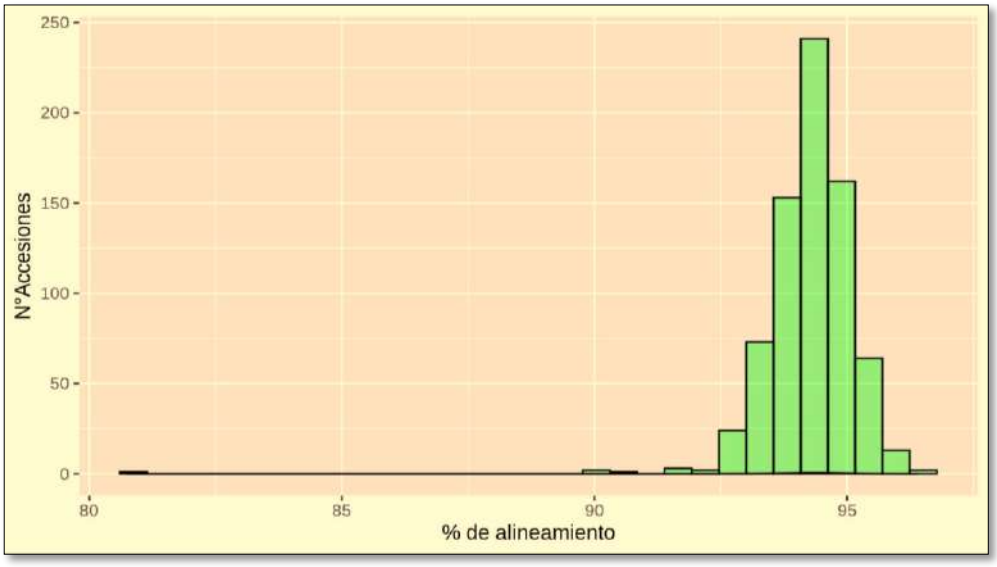


Figura 4. Histograma del porcentaje de alineamiento de *reads* de las 731 accesiones.

Con los alineamientos resultantes y usando el paquete *gstacks* del programa STACKS, se obtuvo los valores de cobertura por accesión. La mayoría, presentó alta cobertura, siendo el promedio 6.5X, el valor mínimo 3.3X (accesión ME_170) y el máximo 21.85X (accesión ME_005). Estos valores garantizan la fidelidad y la existencia de cada base en su respectivo locus. En la Fig. 5 se observa que los valores se concentran entre 3X – 10X. Los altos valores de cobertura están relacionados con la alta cantidad de *reads* (ME-005).

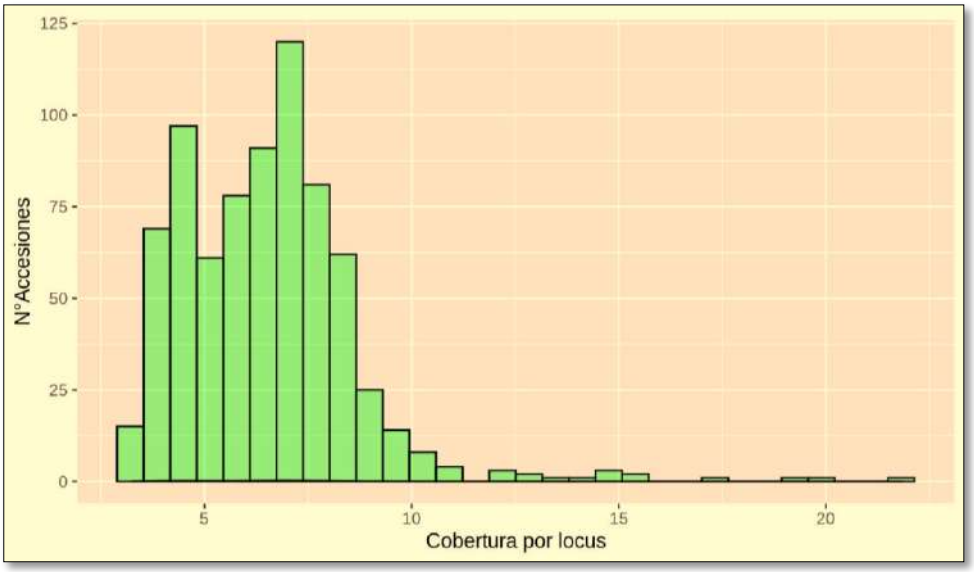


Figura 5. Histograma de la cobertura por *locus* de cada accesión

6.1.3. SNP calling

Se logró obtener 264,088 SNPs iniciales para la población estudiada de la Colección Nacional de Yuca, contenidos en un archivo VCF. Con estos se genotipó a cada una de las 731 accesiones usando los parámetros ya mencionados. En la Tabla 2, se describe a cada uno de los 18 cromosomas de *M. esculenta* ordenados de acuerdo a su "ID" de identificación del NCBI; se indica su tamaño en Mb (megabases), la densidad de SNPs que contiene y el porcentaje que representa del total. También se encontró SNP que pertenecen al genoma plastidial y a un grupo de *contings* que forman parte del genoma de referencia.

Tabla 2. Distribución y densidad de los 264,088 SNPs identificados en *M. esculenta*.

Cromosoma	ID: Cromosoma	Tamaño (Mb)	N.º de SNP	Porcentaje de SNP
1	NC_035161.1	34.96	16,768	6.35
2	NC_035162.1	32.42	15,668	5.93
3	NC_035163.1	29.41	13,445	5.09
4	NC_035164.1	28.74	10,622	4.02
5	NC_035165.1	28.44	16,599	6.29
6	NC_035166.1	27.94	13,353	5.06
7	NC_035167.1	27.07	9,985	3.78
8	NC_035168.1	34	16,670	6.31
9	NC_035169.1	29.41	12,641	4.79
10	NC_035170.1	26.34	11,010	4.17
11	NC_035171.1	27.37	10,635	4.03
12	NC_035172.1	31.6	13,536	5.13
13	NC_035173.1	28.12	13,589	5.15
14	NC_035174.1	24.86	11,375	4.30
15	NC_035175.1	26.19	12,987	4.91
16	NC_035176.1	28.98	13,687	5.18
17	NC_035177.1	27.37	12,621	4.77
18	NC_035178.1	25.23	13,363	5.06
Plastidial	NC_010433.1	0.16	1,535	0.58
Contings	NW	63.68	23,999	9.09
Total	-	582.279	264,088	-

Los 18 cromosomas, presentan una distribución proporcional de todos los SNPs; el cromosoma 1 contiene la mayor densidad (16,768) y el cromosoma 7 (9,985) la menor.

La Fig. 6 es el diagrama de densidad de SNP por cromosoma, se muestra una ventana de 35 Mb, que corresponde al cromosoma de mayor tamaño (cromosoma 1), tomando esta ventana los SNP se ordenan desde el cromosoma 1 hasta el 18, a excepción del genoma plastidial (NC_010433.1) que se ubica primero en el gráfico. Los cromosomas en la base de datos del NCBI no están ordenados por tamaño, tal como se observa en la Fig. 6. Cada línea de color es un SNP y el espacio entre uno y otro es una ventana o paso de 1Mb, las líneas de color rojo representan zonas con muy alta densidad de SNP (>1,531), las de color naranja una alta densidad (1,021 – 1,361), las de color amarillo mediana densidad (681 - 851), las de color verde baja densidad (1 - 341) y las de color

plomo no presentan SNP. Los cromosomas 5 (NC_035165.1), 16 (NC_035176.1), 18 (NC_035178.1) y el genoma plastidial (NC_010433.1), son los únicos con zonas con muy alta densidad; todos los cromosomas nucleares presentan zonas con mediana densidad y las regiones con ningún SNP son escasas.

6.1.4. Filtrado de SNPs

El archivo VCF que contiene a los 264,088 SNPs, se sometió a filtros estrictos para obtener solo aquellos de alta calidad y representatividad, útiles para el estudio de diversidad y estructura genética de la colección. El primer filtro que se aplicó, fue para eliminar loci con más del 20% de data faltante; es decir; el locus que contienen al SNP tuvo que existir en el 80% de la población (585 accesiones) para ser considerado, caso contrarió fue eliminado, en este primer filtro la cantidad de SNPs se redujo a 108,397. El histograma de la data faltante por *locus*, resultante del filtrado se muestra en la Fig.1 del Anexo.4.

El segundo filtro se realizó a nivel de individuo (accesión), aquí se eliminó a accesiones con más del 55% de data faltante. En la Fig. 7 se observa que la mayoría de accesiones posee entre 10 – 20% de data faltante, el límite de tolerancia se indica con una línea de color rojo en el valor 0.55 del eje “X”, los valores superiores ubicados a la derecha, son los que se eliminaron y pertenecen a las accesiones ME_170 (0.57), de la región Tumbes, ME_264 (0.62) de la región Lima y ME_310 (0.56) de la región Ucayali; quedando finalmente 728 accesiones para los análisis posteriores.



Figura 6. Diagrama de densidad de SNP por cromosoma, con una ventana de 1 Mb, incluye el genoma plastidial (NC_010433.1)

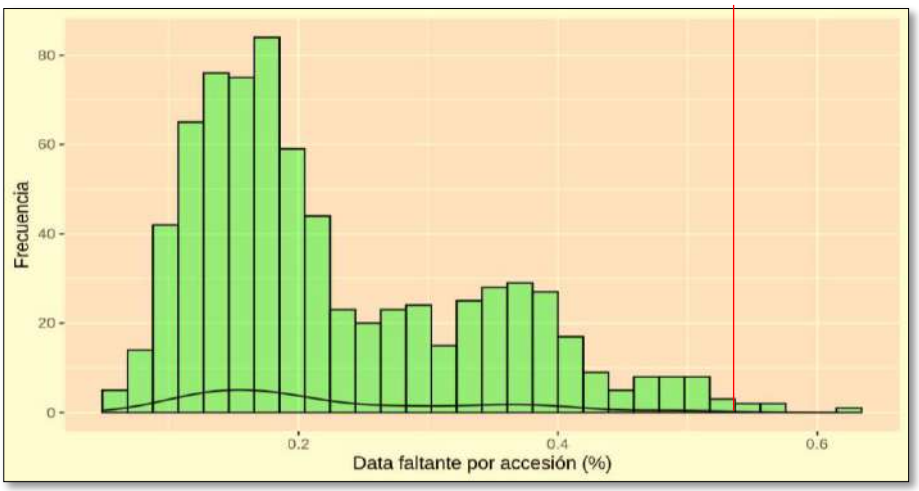


Figura 7. Histograma de frecuencia del porcentaje de data faltante por accesión, en rojo se indica el límite de tolerancia en 0.55.

Para el tercer filtro se tomó en cuenta las frecuencias alélicas, se eliminó aquellos SNP con un valor de MAF < 0.10. La elección de este valor, se realizó en base a los valores del PIC de cada SNP, según Botstein (1980), para marcadores codominantes como los SSR, un marcador es poco informativo cuando su valor de PIC < 0.25, medianamente informativo cuando se encuentra entre 0.25 ≤ PIC ≤ 0.5 y altamente informativo cuando el valor de PIC > 0.5.

Todos los SNP identificados en la colección son bialélicos, no se encontró trialélicos ni tetralélicos; por ello los valores de PIC en su mayoría son menores a 0.25. La Fig.2 del Anexo.4, corresponde al plot de los valores de PIC de los SNP resultantes del segundo filtrado, la línea roja indica un valor de 0.25; si tomamos este como límite para considerar un SNP medianamente informativo, se perdería más de la mitad de la información.

Tomando en cuenta ello y lo reportado por De Albunquerque (2018); se consideró un PIC > 0.15 como mediadamente informativo; los SNP con estos valores fueron aquellos con un MAF ≥ 0.10. Por ello en el tercer filtrado se eliminó a los SNP con un MAF < 0.10. quedando 1,271 SNP (Fig. 8), con un promedio de PIC de 0.245, que se ajusta a lo reportado por Botstein (1980).

Los 1,271 SNPs, son altamente informativos y reflejan un filtrado óptimo bajo los parámetros elegidos en análisis de llamado de SNPs , esto se observa en la Fig.3 del Anexo.4, donde la ausencia de información por accesión es escasa.

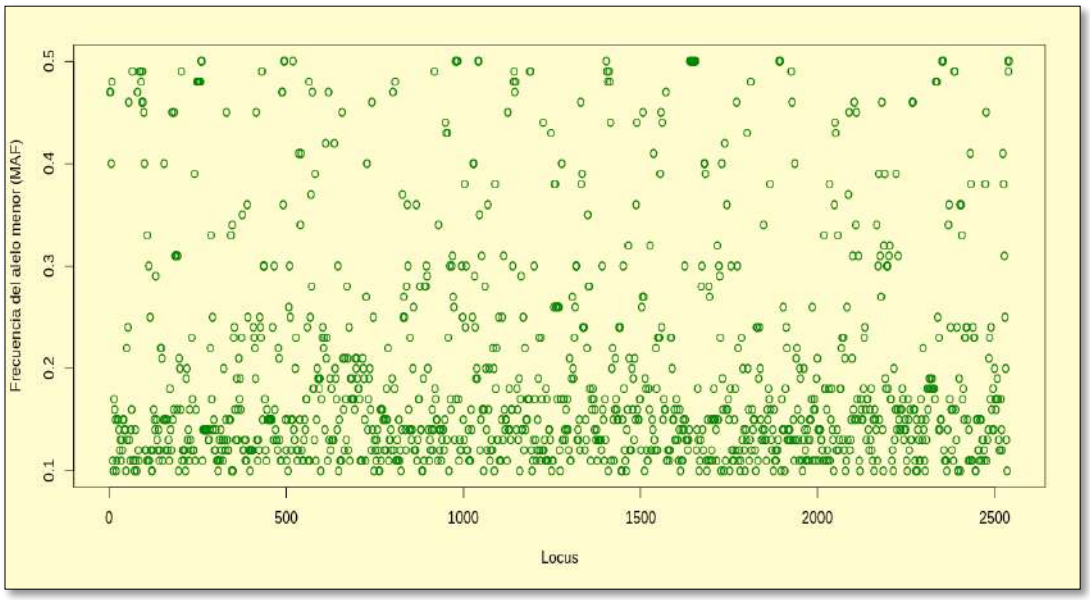


Figura 8. Plot de los valores de MAF para cada alelo de los 1,271 SNPs, resultantes del tercer filtrado. El paquete *snpReady* usado para el análisis, considera a cada alelo SNP-l*oci* por ello se observa el doble de locus (2,542).

6.2. Análisis de la diversidad y estructura genética de la Colección Nacional de Yuca

6.2.1. Diversidad genética de la colección

Los estadísticos de diversidad genética (Ho, He, PIC, MAF) se calcularon para toda la colección y para los países de Perú, Brasil, Bolivia y Paraguay, por separado; los resultados se muestran en la Tabla 3.

El valor total de Ho para las 728 accesiones de la Colección Nacional fue 0.3187; mayor que la He, cuyo valor fue de 0.2933. Bolivia presentó el mayor valor de Ho (0.3304) y

Brasil el menor valor (0.3155).

Respecto a la He, Paraguay presento el mayor valor (0.2903) seguido de cerca por Perú (0.2894), mientras que el menor valor fue para Brasil (0.2792). Respecto a Perú, este presento también un mayor valor de Ho (0.3188) respecto a la He (0.2894). Los altos valores de Ho reflejan el exceso de heterocigotos en la población.

Los valores de MAF para cada locus de SNP, fueron mayores a 0.10 debido a los filtros previos realizados (Fig. 8); en base a ello, el valor promedio de MAF para toda la colección fue de 0.1973. Paraguay presento el mayor valor promedio (0.2063) y el menor valor fue para Perú (0.1946) (Tabla 3).

El valor de PIC para cada locus fue mayor a 0.10. debido también al filtrado previo realizado (Fig. 2 del Anexo.4), donde se garantizó que los SNP resultantes sean informativos, el valor promedio para toda la colección fue de 0.2453. El mayor valor promedio fue para Perú (0.2421) y el menor valor para Brasil (0.2266) (Tabla 3).

Tabla 3. Resumen de estadísticos de diversidad genética en base a los 1,271 SNP, para toda la colección y para los países que la conforman.

Población	Ho[a]	He[b]	MAF[c]	PIC[d]
Toda la colección (728 accesiones)	0.3187	0.2933	0.1973	0.2453
Perú (634 accesiones)	0.3188	0.2894	0.1946	0.2421
Brasil (47 accesiones)	0.3155	0.2792	0.2050	0.2266
Bolivia (24 accesiones)	0.3304	0.2859	0.2037	0.2330
Paraguay (23 accesiones)	0.3232	0.2903	0.2063	0.2367

[a]Heterocigosidad observada; [b]Heterocigosidad esperada; [c]Frecuencia del alelo menor; [d]Contenido de información polimórfica.

Para corroborar el exceso de heterocigotos respecto a lo esperado, se realizó un plot de los genotipos de cada accesión por SNP. La Fig. 9 muestra este resultado, los puntos grises indican un genotipo homocigoto para el alelo de referencia (aa) (alelo perteneciente al genoma de referencia usado), los puntos azules indican un genotipo heterocigoto (aA) y rojos indican un genotipo homocigoto para el alelo alternativo (alelo propio de la colección); se observa alto número de individuos heterocigotos (azules). La presencia de homocigotos para el alelo de referencia (gris) por locus, fue mayor que para el alelo alternativo (rojo).

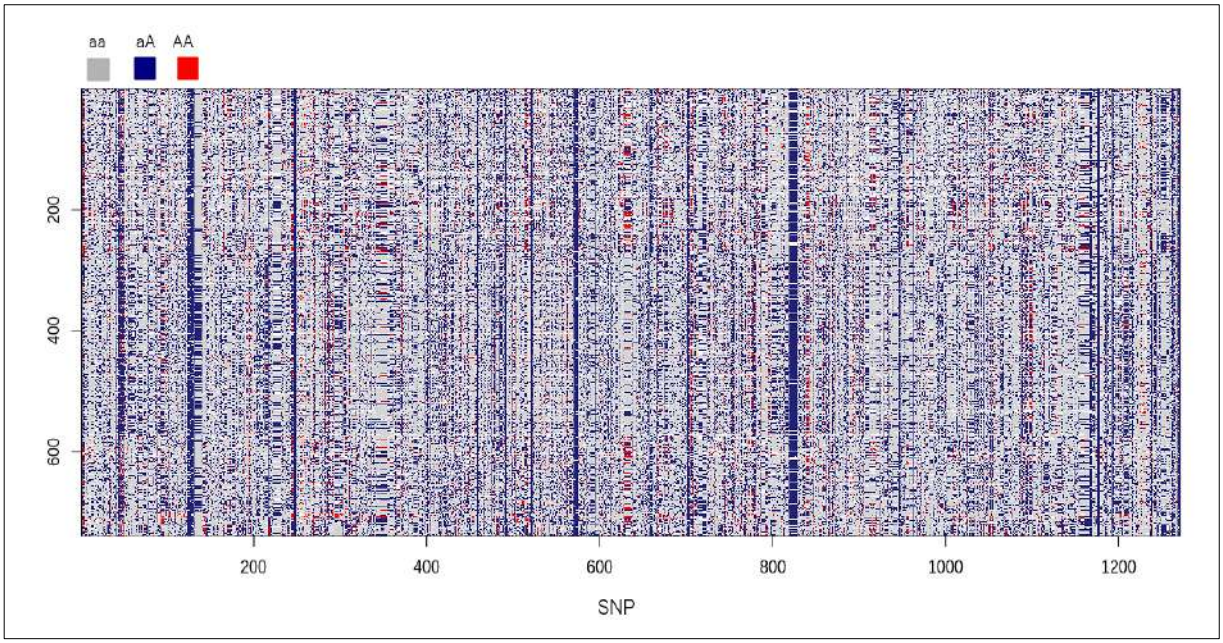


Figura 9. Plot de genotipos por locus para cada accesión de la colección, los puntos de color gris pertenecen a los homocigotos para el alelo de referencia, los de color azul pertenecen a los heterocigotos y los rojos a los homocigotos para el alelo alternativo.

6.2.2. Diversidad genética en la población de yucas peruanas

Los 17 departamentos de Perú que conforman la Colección Nacional de Yuca, se encuentran representados por diferentes cantidades de accesiones, los mejores representados son: Ucayali (318), Lima (105), San Martín (80), Huánuco (21) y La Libertad (21), que conforman el 73.8% del total de accesiones; los 12 departamentos

restante poseen muy poca representatividad en número de accesiones: Lambayeque (12), Loreto (11), Amazonas (10), Junín (10), Piura (10), Cusco (9), Anchas (7), Tumbes (6), Apurímac (5), Cajamarca (5), Arequipa (3) e Ica (1).

En base a ello, los estadísticos de diversidad genética (Ho, He, PIC, MAF) se evaluaron en las regiones mejor representadas (Ucayali, Lima, San Martín, Huánuco y La Libertad).

Tabla 4. Resumen de estadísticos de diversidad genética en las regiones de Ucayali, Lima, San Martín, Huánuco y La Libertad.

Población	**Ho[a]**	**He[b]**	**MAF[c]**	**PIC[d]**
Ucayali (317 accesiones)	0.3202	0.2749	0.1881	0.2268
Lima (104 accesiones)	0.3196	0.2939	0.2040	0.2426
San Martín (80 accesiones)	0.3218	0.2914	0.1974	0.2924
Huánuco (21 accesiones)	0.3253	0.2769	0.1933	0.2279
La Libertad (21 accesiones)	0.3149	0.2824	0.1974	0.2323

[a]Heterocigosidad observada; [b]Heterocigosidad esperada; [c]Frecuencia del alelo menor; [d]Contenido de información polimorfica.

La Ho en los cinco departamentos fue mayor a la He, Huánuco presentó el más alto valor promedio (0.3253), seguido de San Martín (0.3218); mientras que Lambayeque presentó el menor valor (0.3132). Respecto a los valores promedios de He, Lima presentó el mayor valor (0.2939), seguido por San Martín (0.2914) y el departamento con el más bajo valor fue Ucayali (0.2749).

Es evidente también el exceso de heterocigotos en la población peruana, corroborando los resultados anteriores. En cuanto a los valores promedio de MAF, Lima presentó el más alto valor (0.2040) y el menor Ucayali (0.1881). San Martín presento el mayor valor de PIC (0.2924) y el menor valor fue para Ucayali (0.2268).

6.2.3. Estructura poblacional en la Colección Nacional de Yuca.

6.2.3.1. AMOVA, Fst y distancia genética de Nei

Los análisis AMOVA, se realizaron en dos niveles de agrupación: 1) a nivel de país (Perú, Brasil, Bolivia y Paraguay) y 2) a nivel de provincia en las accesiones peruanas (Ucayali, San Martín, Lima, La libertad y Huánuco).

El primer AMOVA con 999 permutaciones en los 1,271 SNPs con un PIC > 0.10 a nivel de países, indica que más del 87% de la varianza se encuentra entre las accesiones y solo el 13% se concentraba entre los países, este resultado indica que la colección no presenta una estructuración genética a este nivel, ya que no permite explicar toda la variación genética que contiene.

El valor promedio de Fst para toda la colección fue 0.0645, valor que corrobora la limitada diferenciación entre las poblaciones. Se realizó también, un análisis Fst por pares o “parwise Fst” y una matriz de distancia genética de Nei, entre los países, para determinar su grado de diferenciación, el resultado se observa en las Tablas 5 y 6 respectivamente y las distancias se ilustran en la Fig. 4 del Anexo.4. La mayor diferenciación entre países (Fst) se encontró entre Perú y Brasil (Fst=0.08694, Nei=0.04177) y la menor entre Brasil y Bolivia (Fst=0.02171, Nei=0.01596). En cuanto a la menor diferenciación entre Perú y los demás países, esta fue con Paraguay (Fst=0.03963, Nei=0.02198).

Tabla 5. Valores del Fst por pares o "parwise Fst" entre los cuatro países que forman parte de la colección.

	Perú	Brasil	Bolivia	Paraguay
Perú	0.00000	0.00000	0.00000	0.00000
Brasil	0.08694	0.00000	0.00000	0.00000
Bolivia	0.05151	0.02171	0.00000	0.00000
Paraguay	0.03963	0.04553	0.02866	0.00000

Tabla 6. Matriz de distancia genética de Nei, entre los cuatro países que forman parte de la colección.

	Perú	Brasil	Bolivia	Paraguay
Perú	0.00000	0.00000	0.00000	0.00000
Brasil	0.04177	0.00000	0.00000	0.00000
Bolivia	0.02736	0.01596	0.00000	0.00000
Paraguay	0.02198	0.02646	0.02209	0.00000

Para el AMOVA entre regiones peruanas se consideró, al igual que en el análisis anterior, 999 permutaciones para los 1,271 SNPs, con un PIC > 0.10.

En este nivel más del 90% de la varianza se encontró entre las accesiones, mientras que, un poco más del 9% entre las regiones. Este resultado es insuficiente para explicar la diversidad genética contenida en la colección, esto es corroborado por el valor promedio del Fst = 0.0428; indicando poca diferenciación entre las regiones.

Los resultados del "parwise Fst" y las distancias de Nei, se observan en las Tablas 7 y 8, respectivamente; las regiones con la mayor diferencia genética (Fst), son Ucayali y Lima (Fst = 0.06198 y Nei = 0.03731), Ucayali y La Libertad (Fst = 0.06156 y Nei = 0.03047) y las regiones con el germoplasma más similar son La Libertad con Huánuco (Fst = 0.00172 y Nei = 0.01008). En la Fig. 5 del Anexo.4, se muestra el árbol de *Neighbour-joining*, que ilustra las relaciones entre las provincias de acuerdo a sus

distancias genéticas de Nei.

Tabla 7. Valores de Fst por pares o "parwise Fst" entre las poblaciones de yuca peruanas.

	Ucayali	**San Martín**	**Lima**	**La Libertad**	**Huánuco**
Ucayali	0.00000	0.00000	0.00000	0.00000	0.00000
San Martín	0.02673	0.00000	0.00000	0.00000	0.00000
Lima	0.06198	0.01808	0.00000	0.00000	0.00000
La Libertad	0.06156	0.01206	0.00454	0.00000	0.00000
Huánuco	0.05998	0.01510	0.00864	0.00172	0.00000

Tabla 8. Matriz de distancia genética de Nei, entre las poblaciones de yuca peruanas.

	Ucayali	**San Martín**	**Lima**	**La Libertad**	**Huánuco**
Ucayali	0.00000	0.00000	0.00000	0.00000	0.00000
San Martín	0.01218	0.00000	0.00000	0.00000	0.00000
Lima	0.03731	0.01017	0.00000	0.00000	0.00000
La Libertad	0.03047	0.01187	0.00854	0.00000	0.00000
Huánuco	0.02989	0.01326	0.01046	0.01008	0.00000

6.2.3.2. PCA y DAPC

En base a la información genética existente, se realizaron pruebas de asignación para determinar la pertenencia de las accesiones a los grupos definidos *a priori*, que se usan en base a su procedencia (países y regiones de origen geográfico).

El análisis de componentes principales de las 728 accesiones, se realizó primero, en base al país de procedencia; los resultados indican que los cinco primeros componentes principales (PC) explican el 34,57% de la variación genética (Fig. 6 del Anexo.4). La gráfica de PCA se realizó en base a la primera y segunda componente (PC1 y PC2) que explican el 22.71 % de la variación total (Fig. 10).

En la gráfica resultante, no se observó una estructuración en base al país de procedencia, no visualizando agrupaciones claras para Brasil, Bolivia y Paraguay; sin

embrago, el PC1 permite diferenciar dos grupos en Perú, uno de ellos a la izquierda, que posee accesiones peruanas mezcladas con el resto de países; mientras que en el segundo grupo, ubicado a la derecha del gráfico, encontramos solo accesiones peruanas; en este último se evidencia un grupo extremo de accesiones definida y separada del resto (Fig. 10).

Para corroborar la calidad de representación de este gráfico, se realizó un plot de los valores del coseno cuadrado (cos^2), para cada punto de representación de las accesiones en el PC1 y PC2. Los valores del cos^2 de estos componentes, ayudan a determinar la calidad de representación de cada punto de las accesiones en los componentes (Fig. 11), y corroboran los posibles grupos encontrados en el plot de PCA (Fig. 10).

En el plot de valores del cos^2 para el PC1 y PC2 (Fig. 11), los puntos de accesiones se colorean de acuerdo a una gradiente (anaranjado – celeste), que va de mayor (0.5) a menor valor (0.1), mientras más alto es el cos^2, mayor es la representación de dicha accesión en ese punto por ambos componentes. En este PCA se definen cuatro grupos, formados por accesiones con altos valores de representatividad (círculos anaranjados).

La Tabla 9 describe la conformación de estos grupos, en el primero se encuentran 21 accesiones (9 de Perú, 11 de Brasil y una de Bolivia,) en el segundo y tercero se encuentra exclusivamente 23 y 27 accesiones de Perú respectivamente y finalmente el cuarto grupo está formado por 35 accesiones (28 de Perú, 3 de Brasil y 4 de Paraguay). Las accesiones restantes están dispersas por todo el gráfico, sin un claro patrón de agrupamiento, indicando cierto grado de mezcla genética.

Tabla 9. Clasificación y descripción de los cuatro grupos de accesiones de yuca, definidos en base a los valores de representatividad del cos^2 de las dos primeras componentes PC1 y PC2.

Grupo	Nº de accesiones	Accesiones por país de procedencia			
		Perú	Brasil	Bolivia	Paraguay
1	21	9	11	1	0
2	23	23			
3	27	27			
4	35	28	3		4

Se realizó otro PCA, para encontrar algún tipo de estructuración u organización solo en las accesiones peruanas, para ello se usó la agrupación *a priori* por región de origen; en este análisis se consideró a las 634 accesiones de todas las regiones, a diferencia de los análisis de diversidad genética.

El resultado del PCA, muestra que el 32.83% de la varianza genética es explicada por los cinco primeros componentes (Fig. 7 del Anexo.4). Se graficó las dos primeras (PC1 y PC2) que explican el 20.79% de la varianza total, el plot se muestra en la Fig. 12; se observa que no existe una estructura clara en base a región de procedencia; sin embargo, el PC1 ayuda a separar las accesiones en dos grandes grupos, el primero se ubica a la derecha del gráfico, formado en su mayoría por accesiones de Ucayali, y el segundo formado por accesiones del resto de provincias.

Ambos componentes (PC1 y PC2) ayudan a diferenciar a un grupo definido y separado del resto, formado exclusivamente por 32 accesiones de Ucayali (círculo morado de la Fig. 12).

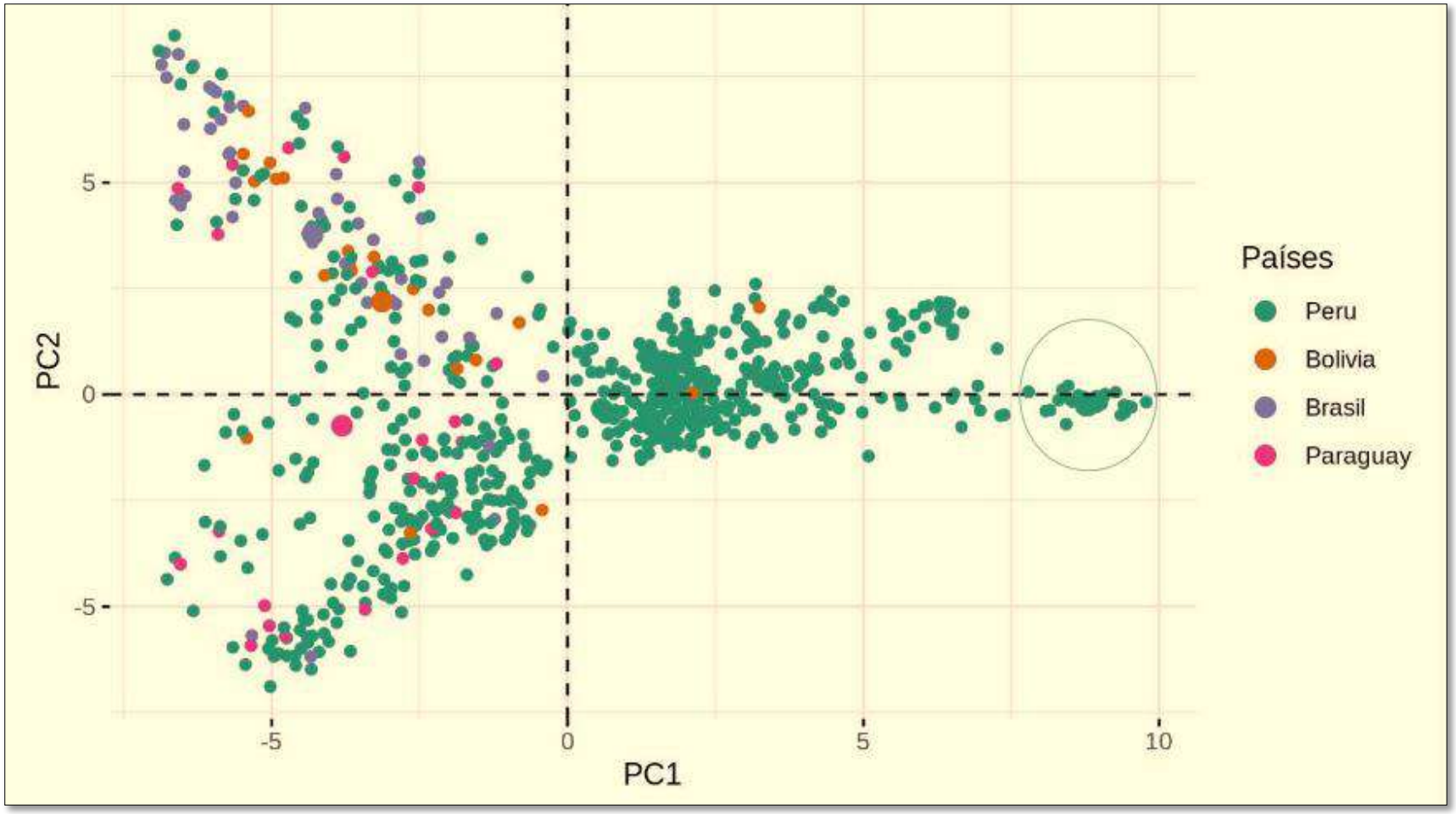


Figura 10. Gráfica del análisis de PCA para las 728 accesiones usando la agrupación por países de procedencia con los 1,271 SNPs. El PC1 y PC2 explican el 22.71 % de la variación genética, en el círculo verde se indica las accesiones peruanas que logran diferenciarse del resto .

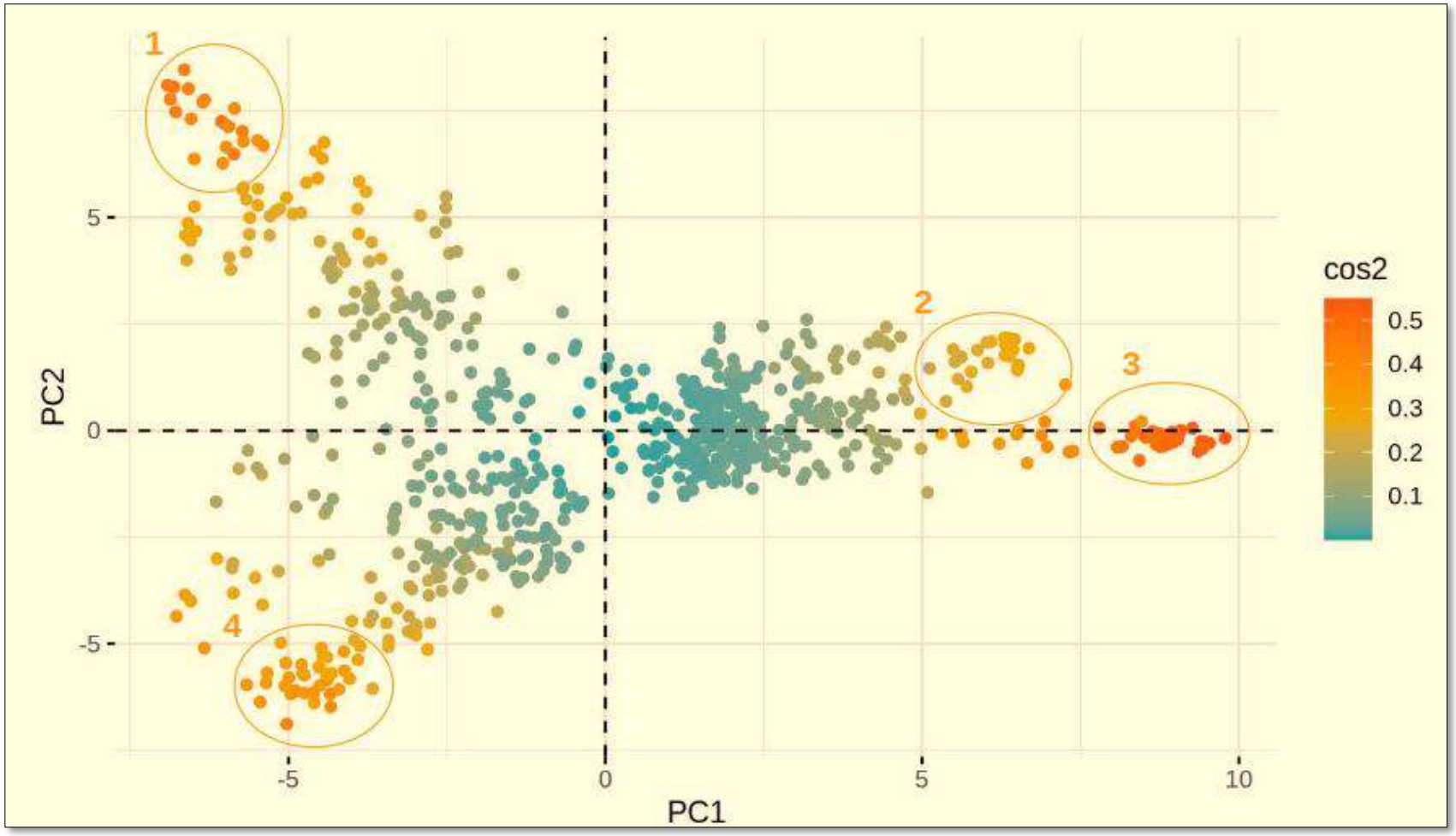


Figura 11. Gráfica de los valores del cos^2 de las dos primeras componentes PC1 y PC2 de las 728 accesiones, los valores se colorean de acuerdo a un gradiente de calor (anaranjado – celeste) que va desde el más alto al menor valor respectivamente. Los grupos genéticamente relacionados se agrupan en los círculos enumerados de color anaranjado.

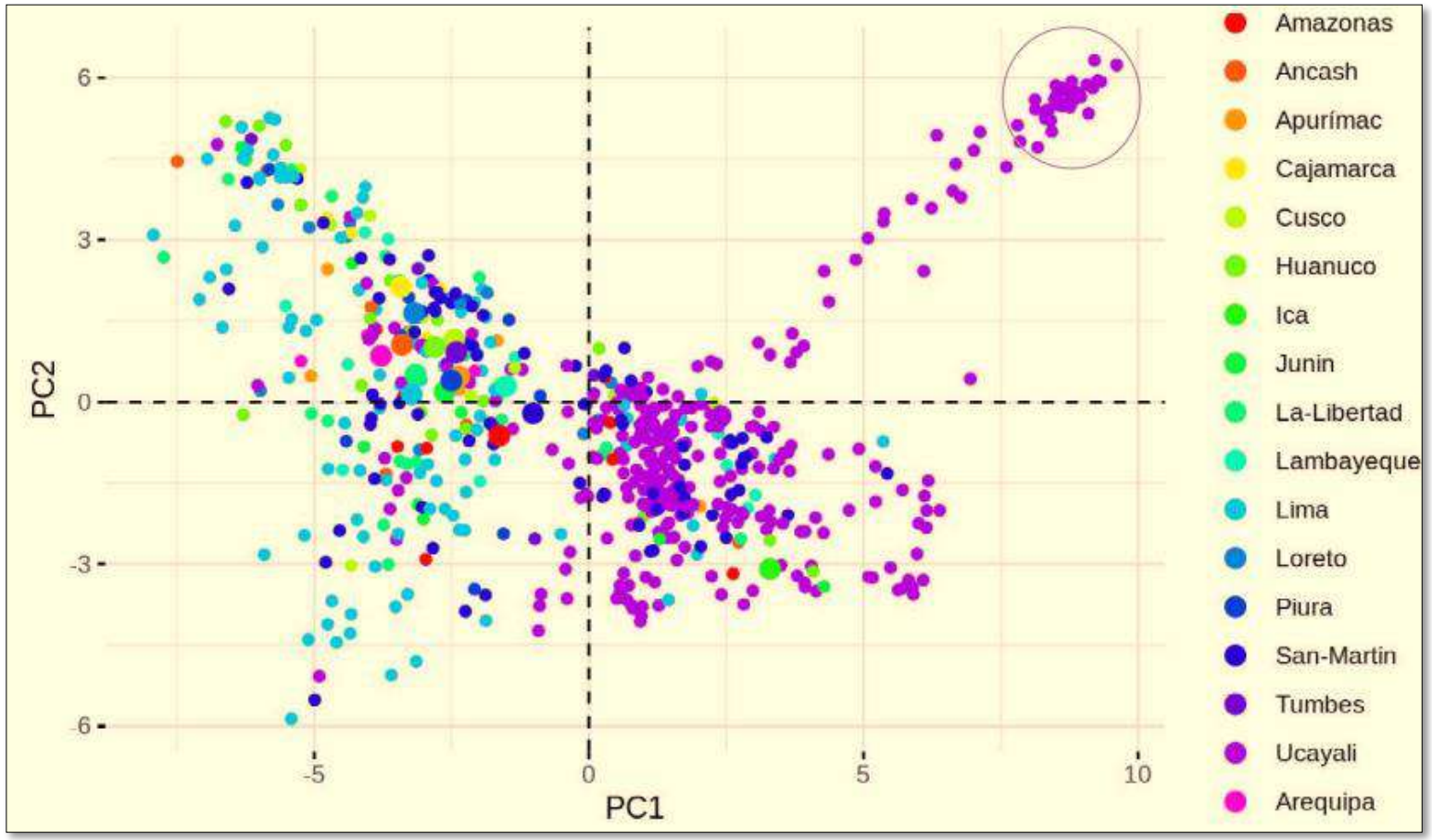


Figura 12. Gráfica del análisis de PCA de las 634 accesiones peruanas, agrupadas en base a región de procedencia, usando los 1,271 SNP. El PC1 y PC2 explican el 20.79 % de la variación genética, se indica en un círculo morado el grupo de accesiones de Ucayali que logra diferenciarse del resto.

Se graficó también los valores del cos^2 para ambos componentes (PC1 y PC2), de igual modo que en el plot de la Fig. 11, los valores del cos^2 se encuentran representados por un gradiente de color (anaranjado – celeste) que representan los mayores y menores valores respectivamente.

Se encontró que la representación de las accesiones por los componenetes PC1 y PC2 es baja; solo se logró diferenciar un grupo conformado exclusivamente por 32 accesiones de Ucayali, cuyos valores de cos^2 son elevados ($0.6 < cos^2 < 0.8$) (Fig. 13). El resto de accesiones presentó valores menores a 0.4 que indica cierto grado de mezcla genética.

En los PCA realizados, no se logró observar estructuración o algún tipo de agrupación en base al lugar de procedencia; por ello para poder estimar el número de grupos genéticos que se ajusta mejor a nuestros datos, se usó el estadístico de Análisis Discriminante de Componentes Principales (DAPC). Este maximiza la variación genética entre los grupos usando la variación existente entre los individuos, para ello usa un análisis previo de PCA, donde determina la cantidad de componentes necesarios a retener, cuyas variables no esten correlacionadas.

Además, el DAPC realiza un análisis de *clusters*, donde busca el número de grupos (*K-means*) más apropiado para nuestra data, y asigna a cada uno de los individuos a estos grupos (K), usando un valor de probabilidad de asignación individual (PAI). Así, con el DAPC se buscó un equilibrio entre el poder de discriminación de los grupos y la estabilidad de asignación de los genotipos a cada grupo.

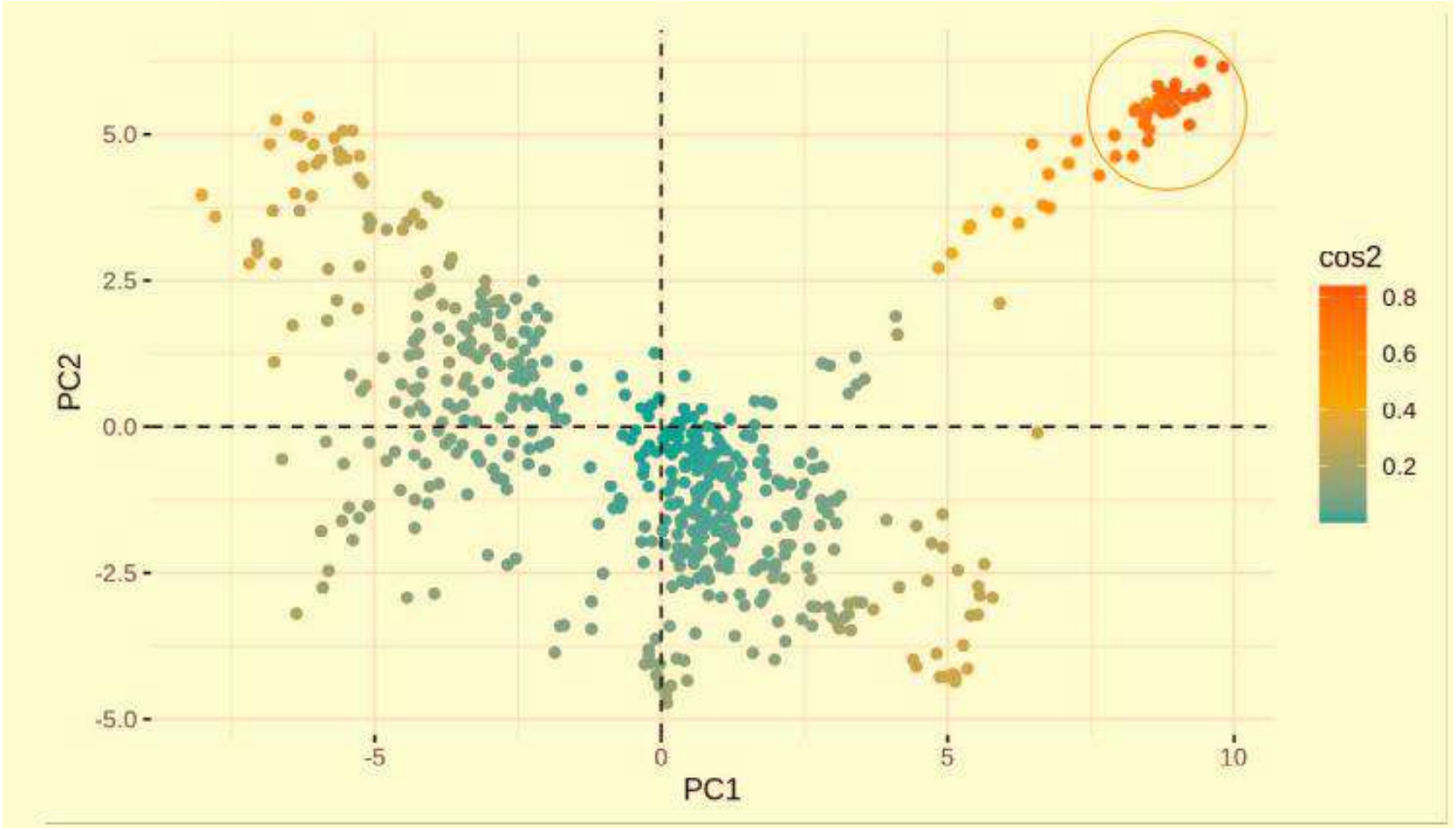


Figura 13. Gráfica de los valores del cos^2 de las dos primeras componentes (PC1 y PC2) para las 634 accesiones peruanas, los valores se colorean de acuerdo a un gradiente de color (anaranjado – celeste) que va desde el mayor al menor valor respectivamente. El grupo de Ucayali genéticamente relacionado se agrupa en el círculo de color anaranjado.

Antes de determinar el número de grupos genéticos que más se ajusta a nuestra data, usando el DAPC, se realizó un análisis previo de probabilidad de pertenencia de cada accesión para cada grupo definido *a priori* (país y región de procedencia).

Para este análisis se aprovechó el estadístico de probabilidad de asignación de individuos (PAI) o probabilidad posterior; los valores se graficaron en un diagrama de barras apiladas compuesta (usando la función *compoplot* del paquete *Adegenet*) (Fig. 14). Este análisis se realizó para determinar la calidad de agrupación que se logra obtener cuando se usa el criterio de origen de procedencia, lo esperado es un valor de PAI = 1,00 que indica un buen poder discriminatorio de los grupos; es decir, una adecuada correspondencia entre genotipo y origen de procedencia. De acuerdo al análisis no se puede observar una estructura clara de la correspondencia entre genotipo y origen de procedencia, ya que los valores de PAI fueron menores a 1.00. Esto se corrobora en cada barra correspondiente a cada accesión, donde se observa dos o más colores, siendo lo esperado un solo color correspondiente a su grupo original.

Existen accesiones con valor de PAI = 1.00 indicando una correcta correspondencia entre genotipo – lugar de procedencia; la mayoría de estas corresponden a las accesiones de los países de Bolivia y Brasil, y a las provincias de Ucayali, Apurímac, Cusco, Lambayeque y Arequipa, esto se observa más claro en la Fig. 15.

Siguiendo este criterio de agrupación, observamos que existen accesiones con porciones de genotipos de otras provincias, incluso presentan poca proporción genotípica de su provincia original (Anchas y Junín). El resultado muestra, que no existe una estructura poblacional en base a origen de procedencia; por ello es necesario buscar un nuevo criterio de agrupación.

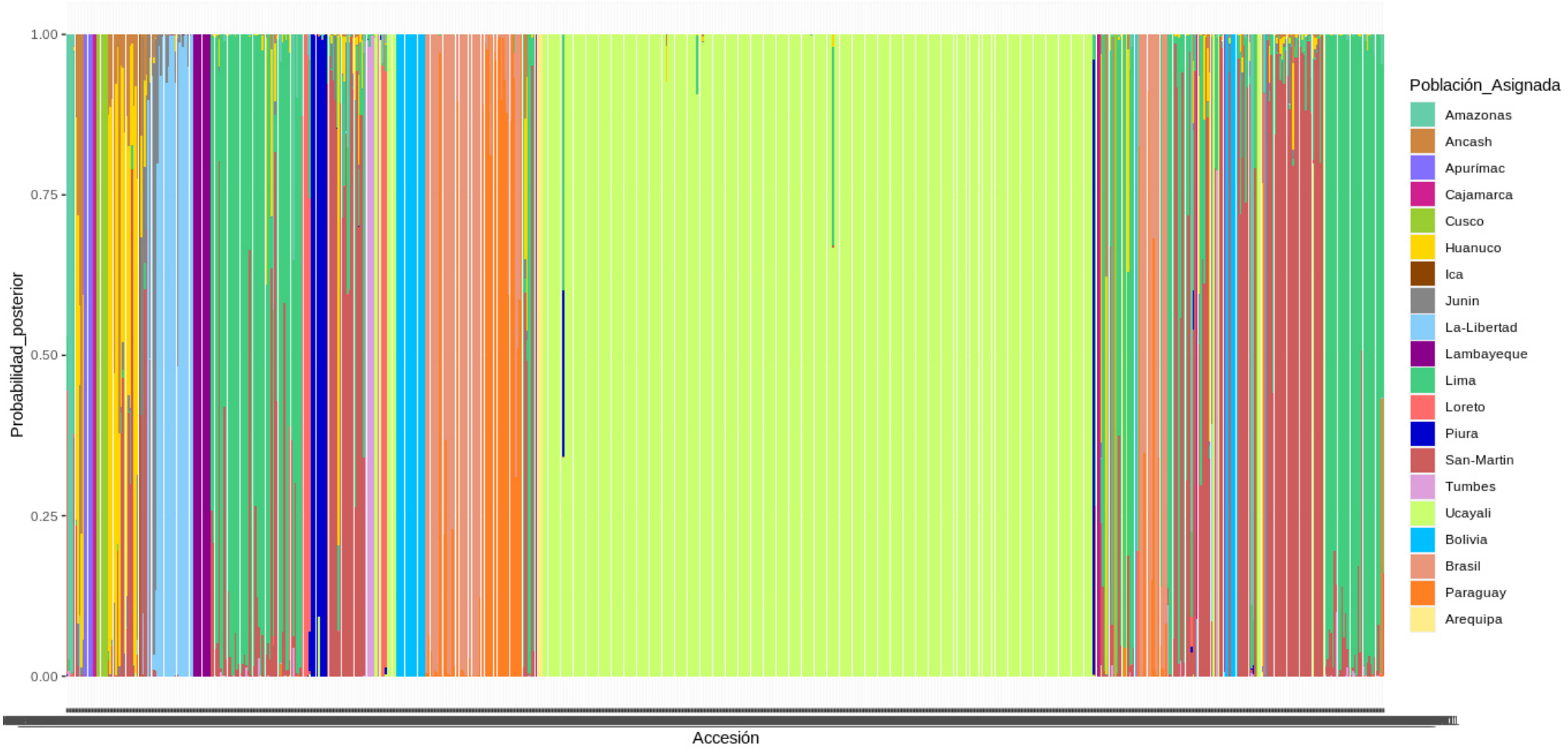


Figura 14. Diagrama de barras apiladas compuestas de las probabilidades de asignación de individuo (PAI) del DAPC de las 728 accesiones de la Colección Nacional, considerando la agrupación en base a origen de procedencia. Cada barra pertenece a una accesión, cuyos colores corresponden a la probabilidad de asignación a cada país o provincia.

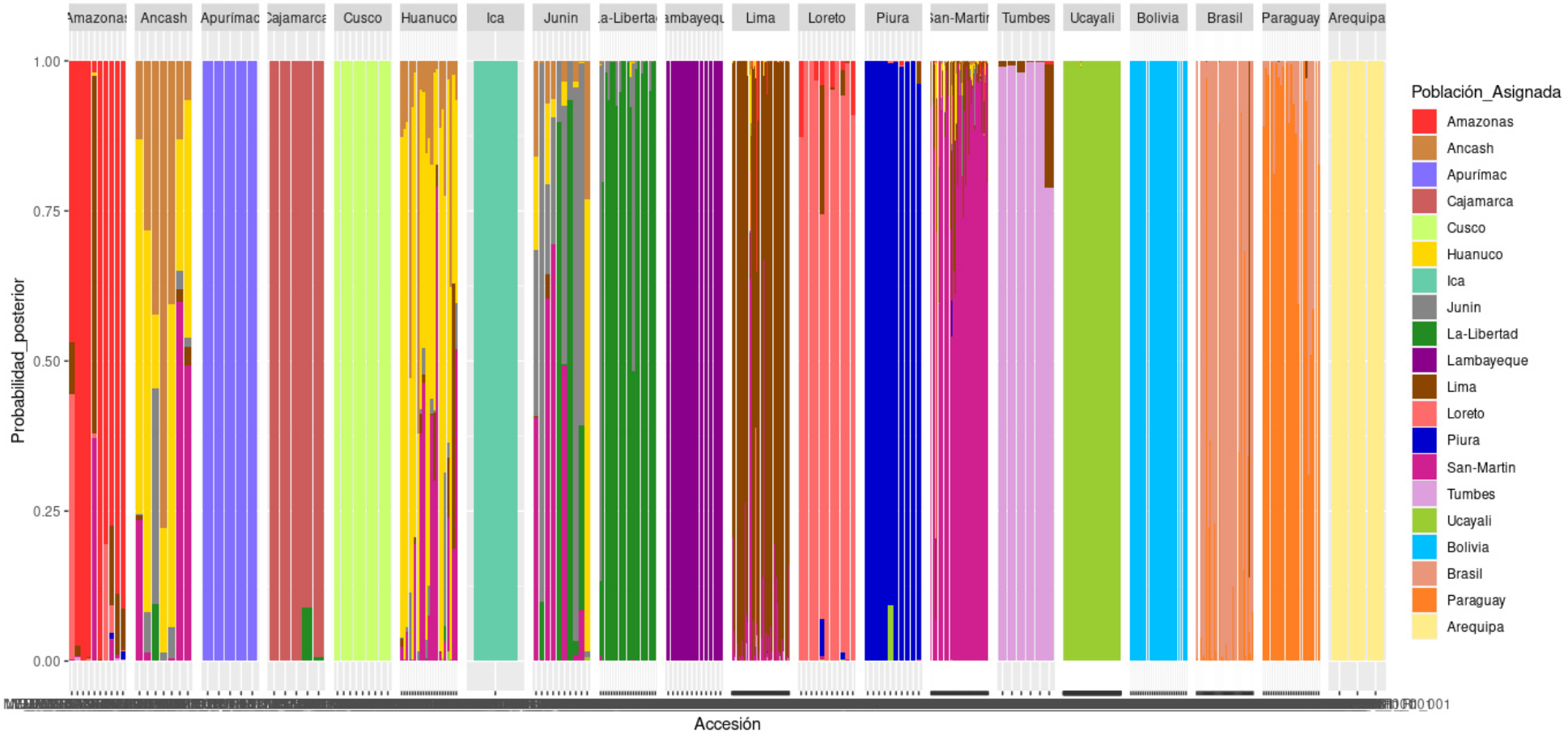


Figura 15. Diagrama de barras apiladas compuestas, ordenada por grupos de procedencia, de las probabilidades de asignación (PAI) del DAPC de las 728 accesiones de la colección. Cada barra pertenece a una accesión, cuyos colores corresponden a su probabilidad de asignación a cada país o provincia.

De los valores anteriores de PAI y los resultados del PCA, donde no se muestra ninguna forma de estructuración en base a procedencia; se usó la función *find.cluster* del paquete *Adegenet* para encontrar el mejor "*K-means*" o grupos, que logren explicar la variación genética de nuestra población.

Para la elección del mejor "K" se usó el Criterio de información Bayesiana (BIC), donde el mejor valor, es aquel que presenta el menor valor de "BIC". En la Fig. 8 del Anexo.4, se muestra la gráfica de los valores de BIC versus el número de clusters (K). En la curva, de este gráfico, se observa que los posibles valores de "K" se encuentran entre 12 y 21.

Para nuestro análisis se seleccionó el valor de K=12, por ser el primer punto de quiebre más abrupto de la curva, y por ser el valor a partir del cual se empiezan a diferenciar los grupos.

Además del número de grupos o *clusters*, el parámetro fundamental para el desarrollo del DAPC, es el número de componentes principales (CP), ya que son las nuevas variables, cuyas combinaciones lineales permitirán realizar la función de discriminantes. Esto optimizó la separación de las accesiones y su asignación a los grupos "K" predefinidos.

Sobre la base de las funciones discriminantes retenidas, es posible derivar las probabilidades de pertenencia a un grupo, para evaluar qué tan bien definidos o mezclados están los grupos; por ello, es importante seleccionar el correcto número de CP. Si se seleccionan muy pocos, respecto a la cantidad de individuos, se perderá mucha información útil y se excluirá del análisis, sin lograr una precisa discriminación entre los grupos. Por otro lado, cuando se retienen demasiados, se sobrestimará los valores de discriminación, provocando sobreajustes.

Para determinar el correcto número de CP a retener, evaluamos el porcentaje de reasignación exitosa de cada accesión a los 12 grupos definidos, por cada grupo de CP probado.

Partimos del valor dado por la función *optim.a.score* que ofrece el paquete a*degenet*, el cual, ayuda a predecir el número correcto de CP, en base a la diferencia entre la proporción de reasignación exitosa del análisis (discriminación observada) y los valores obtenidos utilizando grupos aleatorios (discriminación aleatoria).

El resultado de este análisis, fue una predicción de retención de CP=63 (Fig. 9 del Anexo.4). Sin embargo, al evaluar el porcentaje de reasignación exitosa para la cantidad de CPs predicha, se obtuvo porcentajes menores al 100% para los grupos 1, 2, 3, 6, 9, 10 y 11 (Fig. 10 del Anexo.4), siendo los valores más bajos para los grupos 1 y 11 (58,7% y 59,5% respectivamente) (Tabla 10).

A partir del valor predicho de CP a retener, se estableció tres valores adicionales: 200, 350 y 500, cuyos valores de porcentaje de reasignación (Tabla 10) se graficaron, y a partir de estos se determinó el número adecuado de CPs que permitan lograr un adecuado análisis de discriminantes.

El porcentaje de reasignación a los grupos para 200 CP (Fig. 11 del Anexo.4), presentó mejores valores que en la retención de 63 CP; sin embargo, aún no es suficiente para lograr un porcentaje alto de reasignación para el grupo 1 (58%) y solo cinco grupos alcanzaron el óptimo (2, 4, 5, 8 y 11) (Tabla 10).

El resultado en la retención de 350 CP (Fig. 12 del Anexo.4) mejora notablemente, llegando a ser óptimo para los grupos 2, 4, 5, 6, 7, 8, 9, 11 y 12 y más del 95% para los grupos restantes (Tabla 10). Finalmente, cuando se usó 500 CP (Fig. 13 del Anexo.4)

se logró alcanzar el óptimo para todos los grupos, a excepción de los grupos 1 y 3; pero sus valores fueron superiores al 97% (Tabla 10).

En base a estos resultados, se seleciono 350 CP como el mejor número; como se observa en la curva del *a-score* de la Fig. 9 del Anexo.4, este se encuentra aún dentro de la campana con los mayores valores, sumado a ello los altos valores de reasignación a los grupos, respaldan nuestra elección. No se seleccionó el valor de 500 PC, ya que se podría caer en un sobre-ajuste de valores.

Tabla 10. Valores del porcentaje de asignación a los grupos para la cantidad de CP retenidos.

Nº de CPs retenidos	% de asignación a los grupos definidos											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
63	58.7	78.3	91.1	100	100	87.1	98.8	100	93.8	89.7	59.5	95.2
200	58.7	100	93.3	100	100	93.5	100	100	96.9	87.1	100	98.9
350	95.6	100	93.3	100	100	100	100	100	100	97.4	100	100
500	97.8	100	98.8	100	100	100	100	100	100	100	100	100

Obtenidos los valores de agrupación *a priori* (*K* =12) y el correcto número de CPs a retener (350), que explican más del 90% de la varianza genética, se ploteó los valores en base a la función discriminante lineal uno y dos (LD1 y LD2 respectivamente) (Fig.16).

Con ello se logró diferenciar claramente los 12 grupos genéticos definidos, donde los grupos 1, 2, 3, 5, 8 y 9 logran separarse completamente. Los seis restantes, se logran distinguir unos de otros, pero se separan completamente; aquí los grupos 4, 6, 7, y 12 se distribuyen muy cerca uno del otro, sobre todo 4 y 7; esto indica que sus genotipos son cercanos o similares, lo mismo se observó en los grupos 10 y 11 (Fig. 16).

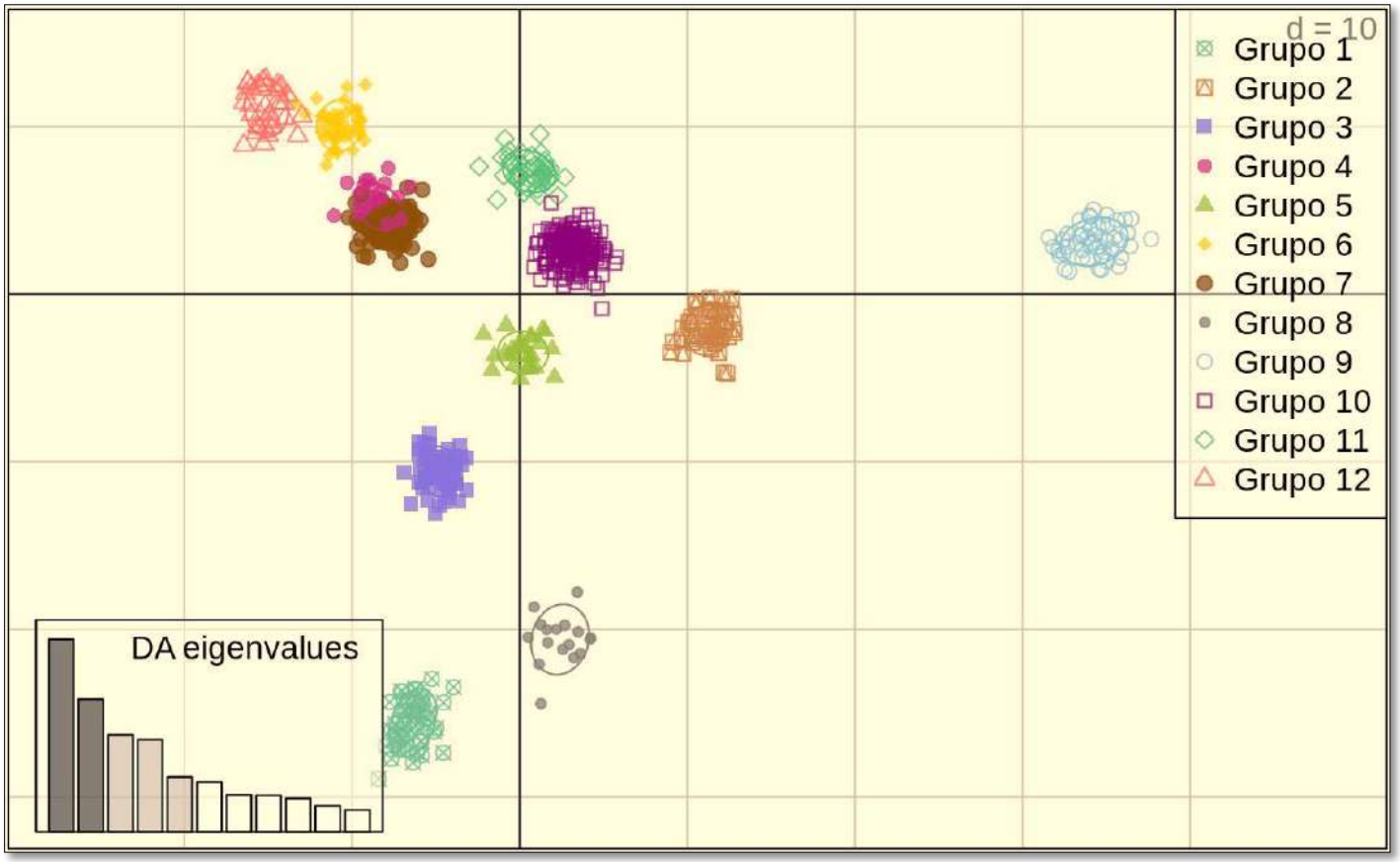


Figura 16. Plot del DAPC para las 728 accesiones, basado en las funciones discriminantes lineales (LD1 y LD2). Los grupos son representados por colores de acuerdo a los valores de la leyenda de la parte superior derecha. En la parte inferior izquierda, encontramos sombreados de gris-oscuro, los *eigenvalues* de los discriminantes usados.

Usando el valor de K=12 y los nuevos valores de probabilidades de asignación por individuo (PAI), se realizó la distribución de las accesiones de cada país y de cada provincia peruana a los 12 grupos (Tabla 11 y Fig. 17). Se observó que la mayor cantidad de accesiones de Brasil pertenecen al Grupo 1, de Bolivia al Grupo 3 y la mayor cantidad de accesiones de Paraguay al Grupo 4; además, estos grupos están formados por accesiones de otros paises y provincias, no son exclusivos para un determinado país.

La población de accesiones peruanas se distribuye en todos los grupos, se encontró que las accesiones de una determinada provincia, pertenecen a diferentes grupos, a excepción de Ica y Arequipa donde la única accesión de Ica pertenece al Grupo 2, y las tres accesiones de Arequipa al Grupo 4 (Tabla 11).

Existen grupos formados exclusivamente por una provincia, es el caso de los Grupos 8 y 9 que están formados solo por accesiones de Ucayali (Fig. 17) siendo el más grande el Grupo 9 con 48 accesiones, mientras que el Grupo 8 presenta 17 (Tabla 11). Un caso similar es el Grupo 11, formado por 42 accesiones de Ucayali más una accesión de Loreto, siendo el único caso de este tipo. Los grupos restantes poseen dos o más provincias, siendo el Grupo 7 quién presenta la mayoría de provincias, a excepción de Ica y Arequipa.

En el caso de las provincias que albergan la mayor cantidad de accesiones (Ucayali, San Martín, Lima, La Libertad y Huánuco); en Ucayali la mayoría de sus accesiones se distribuyen en el Grupo 10 (148 accesiones), en el Grupo 9 (48 accesiones), en el Grupo 11 (42 accesiones), en el Grupo 2 (35 accesiones), en el Grupo 7 (22 accesiones), en el Grupo 8 (17 accesiones) y un menor número en los grupos restantes.

La mayor cantidad de accesiones de Lima se encuentran en el Grupo 4 (19 accesiones), el resto se distribuyen casi equitativamente en los demás, a excepción de los Grupos 8,

9 y 11. En el caso de San Martín, la mayoría de sus accesiones se encuentra en el Grupo 10 (30 accesiones) y en el Grupo 7 (23 accesiones), el resto se distribuye por los otros grupos, a excepción del 8, 9 y 11. El Grupo 10 alberga la mayor cantidad de accesiones (197) y el Grupo 8 la menor cantidad (17) (Tabla11 y Fig. 17).

Tabla 11. Tabla de asignación de las accesiones de cada país y provincia a los 12 grupos genéticos definidos (K=12) por el valor de BIC.

País/Departamento	G.1	G.2	G.3	G.4	G.5	G.6	G.7	G.8	G.9	G.10	G.11	G.12
Brasil	21	0	16	1	5	1	1	0	0	1	0	1
Bolivia	2	0	15	1	1	1	1	0	0	2	0	1
Paraguay	2	0	4	6	2	1	2	0	0	0	0	6
Amazonas	1	1	2	1	0	0	3	0	0	2	0	0
Ancash	0	1	1	1	1	1	1	0	0	0	0	1
Apurímac	0	0	0	1	0	1	2	0	0	1	0	0
Cajamarca	0	0	0	0	0	1	3	0	0	0	0	1
Cusco	0	0	1	0	0	4	2	0	0	2	0	0
Huánuco	0	2	1	2	2	7	6	0	0	1	0	0
Ica	0	1	0	0	0	0	0	0	0	0	0	0
Junín	0	1	2	0	1	4	1	0	0	1	0	0
La Libertad	1	1	5	2	0	5	5	0	0	2	0	0
Lambayeque	0	2	1	1	1	3	2	0	0	2	0	0
Lima	12	1	11	19	10	6	17	0	0	13	0	16
Loreto	0	0	1	1	0	2	3	0	0	0	1	3
Piura	2	0	0	1	1	0	4	0	0	0	0	2
San Martin	4	2	4	4	3	8	23	0	0	30	0	2
Tumbes	0	0	1	0	0	2	3	0	0	0	0	0
Ucayali	2	35	6	2	1	1	22	17	48	140	42	2
Arequipa	0	0	0	3	0	0	0	0	0	0	0	0
TOTAL	47	47	71	46	28	48	101	17	48	197	43	35

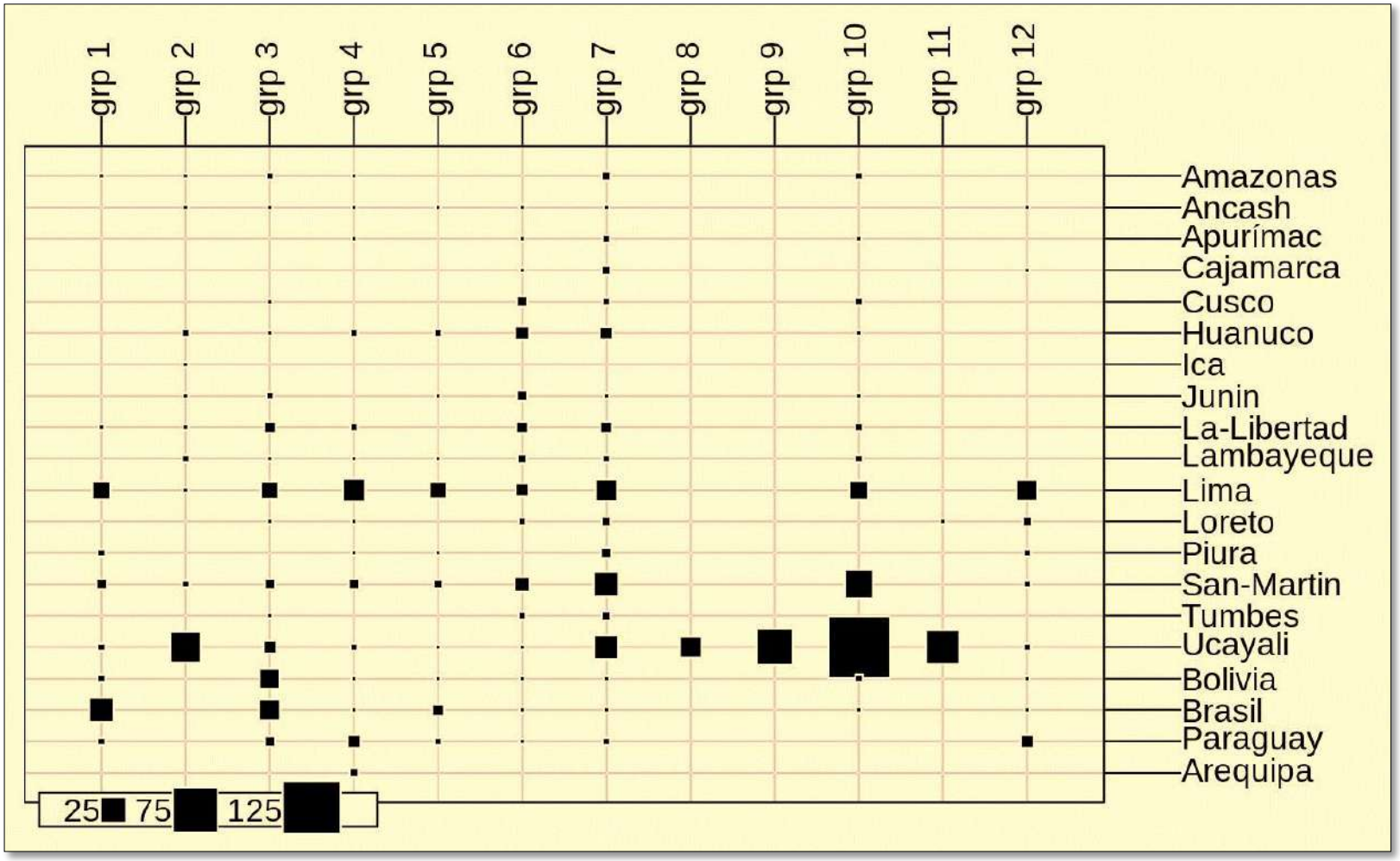


Figura 17. Tabla de visualización gráfica de correspondencia entre los grupos inferidos (K=12) y la agrupación en base a país y provincia.

Una forma de evaluar la calidad de los grupos obtenidos del DAPC, es mediante las probabilidades de asignación de cada accesión o individuo (PAI) a sus grupos definidos, de acuerdo a las funciones discriminantes, donde un PAI alto indica un poder discriminatorio alto de los grupos, mientras que un PAI bajo sugiere grupos con una proporción alta de mezcla.

La PAI para las accesiones de yuca, considerando un K=12, varió entre 0.56 a 1.00. la mayoría de grupos presento un valor de 1.00 para todas sus accesiones (Grupos 1, 2, 3, 5, 8, 9 y 11), valor que respalda la alta precisión y confiabilidad de la asignación de las accesiones (Tabla 12).

Los Grupos 4, 6, 7, 10 y 12 presentaron un PAI entre 1 – 0.99 (Tabla 12); sin embargo, las accesiones de algunos grupos presentaron valores inferiores a este rango. El Grupo 4 presentó dos accesiones con valores de PAI entre 0.98 - 0.96 (accesiones ME_073 y ME_246), el Grupo 7 presentó seis accesiones con valores entre 0.98 – 0.56 (accesiones ME_003, ME_088, ME_138, ME_365, ME_597 y ME_700); aun así, estos valores respaldan la alta confiabilidad en la asignación de los individuos a los grupos definidos (Tabla 12). Sin embargo, no se descarta la existencia de cierto grado de mezcla genética, ya que podrían compartir ciertas regiones genómicas atribuidas a otros grupos.

Tabla 12. Tabla de valores de PAI para cada grupo genético definido.

Rango de valores de PAI	G.1	G.2	G.3	G.4	G.5	G.6	G.7	G.8	G.9	G.10	G.11	G.12
PAI = 1	47	47	71	0	28	0	0	17	48	0	43	0
0.99 ≤ PAI < 1	0	0	0	44	0	48	95	0	0	197	0	35
0.99 > PAI	0	0	0	2	0	0	6	0	0	0	0	0
TOTAL	47	47	71	46	28	48	95	17	48	197	43	35

7. DISCUSIÓN

7.1. Genotipado por secuenciamiento

La elección de este método para el descubrimiento de SNP, fue debido a las características que presenta; principalmente porque se basa en la representación reducida del genoma; es decir, disminuye la complejidad de los genomas usando enzimas (endonucleasas de restricción) sensibles a la metilación, que cortan en determinadas zonas del genoma, generando fragmentos representativos que posteriormente son secuenciados (Elshire *et al*., 2011; Poland y Rife *et al*., 2012).

Este método es ampliamente usado para especies con genomas poliploides (como el de muchas plantas), porque ayuda a evitar zonas del genoma altamente repetitivas, consiguiendo información valiosa. Otra de las razones importantes de su elección, es porque ha demostrado ser eficaz para el descubrimiento de altas densidades de SNP de todo el genoma (Beissinger *et al*., 2013), tanto de especies modelo como no modelo (Poland y Rife, 2012), y resume en un solo paso el descubrimiento de polimorfismos (SNP) y la genotipificación; procesos claves para la caracterización genética de germoplasmas (Poland *et al*., 2012).

Por otro lado, este método permite ahorrar tiempo y dinero, ya que realiza un secuenciamiento multiplexado, es decir, secuencia varios individuos al mismo tiempo, para ello usa los *barcodes* de identificación por individuo y los adaptadores comunes para el secuenciamiento NGS. Además de ello, la data generada es altamente aprovechable, ya que además del descubrimiento de SNPs, genera información útil para programas de mejoramiento genético.

La elección específica del ddRADseq para el genotipado por secuenciamiento, se hizo por sus amplias ventajas frente a otras técnicas; 1) al usar simultáneamente dos

enzimas de restricción, una de corte frecuente y otra de corte infrecuente, elimina la fragmentación aleatoria, obteniendo fragmentos de tamaños más uniformes. Esto reduce al menos cinco veces el costo de la construcción de las librerías, y disminuye pasos de alta perdida de ADN; requiriendo cantidades a partir de 100 ng o menos de ADN genómico inicial, además, 2) al usar dos enzimas de restricción, realiza una mejor selección por tamaño de los fragmentos en toda la población; ya que solo una parte de ellos caerán dentro del rango de tamaño requerido, disminuyendo así, la probabilidad de muestrear regiones duplicadas en ambas direcciones, desde el mismo sitio de restricción (inmediatamente vecinas). Esto reduce a la mitad el número de lecturas necesarias para alcanzar un muestreo de alta confianza en la identificación de un SNP asociado a un sitio de restricción (Peterson *et al*., 2012).

Por ello en el ddRADseq, la elección de las enzimas de restricción es el paso más importante, ya que determinaran la densidad de SNP a obtener. Existe una combinación casi infinita de enzimas (Davey *et al*., 2011) con las que se pueden obtener diferentes resultados.

La elección de enzimas, determinará también, la profundidad de las lecturas y la cobertura del secuenciamiento, y la cantidad de individuos que se podrán secuenciar en paralelo (Beissinger *et al*., 2013). Por ello, el primer paso que se realizó antes de iniciar el ddRADseq de la colección, fue la elección de las enzimas usando ensayos *in silico* con el genoma de referencia. Las enzimas seleccionadas fueron *BtgI* y *TaqI*, que garantizaron más de 5,000 SNPs usando dos millones de lecturas (2M) por individuo.

El juego de enzimas seleccionada, fue fundamental para la identificación de SNP; *TaqI* de corte frecuente, reconoce 4 pb como sitio de corte (T*CGA), y *BtgI* de corte infrecuente, reconoce 6 pb (C*CRYGG, donde “R” puede ser “A” o “G”, e “Y” puede ser “C” o “T”). La combinación de ambas enzimas, permitió la selección de fragmentos de

mediano tamaño con ambos sitios de restricción.

Si solo se usase *TaqI* de corte frecuente, se generaría muchas lecturas de pequeño tamaño, más de las necesarias para un genotipado, provocando una sobre-representación de lecturas en las bibliotecas a costa de perder información de otras zonas del genoma (Rabbi., 2014). Por otro lado, si solo se usase *BtgI,* se obtendría fragmentos más grandes, poco amplificados cuyos SNP, tendrían poca profundidad y podrían considerarse errores de secuenciamiento. Por estas razones, cuando se trabaja con altos niveles de multiplexación, los mejores resultados se obtienen al combinar enzimas (Poland *et al*., 2012).

En nuestros resultados, la cantidad de lecturas secuenciadas por accesión no fue uniforme (Fig. 2), en promedio se logró 1,969,279 lecturas por cada una. Esta diferencia se debe a la calidad y cantidad del ADN genómico extraído; ya que durante el proceso de extracción, algunas accesiones presentaron precipitación de fenoles y azúcares, que para eliminarlos se realizó lavados adicionales con cloroformo:alcohol isoamílico (24:1) y etanol (96%); que provocó la pérdida y disminución de la calidad e integridad del ADN genómico de doble cadena, influyendo enormemente en la calidad del secuenciamiento (Maruki y Lynch., 2017).

El filtro de calidad de ADN genómico, es fundamental para el éxito de estos protocolos, dado que la eficiencia de la construcción de las librerías, en la digestión, ligación y amplificación, puede tener efectos significativos en el conjunto de marcadores finales (Davey *et al*., 2011). Lo más importante es que la cantidad de ADN de las diferentes muestras se equilibren o uniformicen, antes de la agrupación, para evitar perder marcadores de algunos individuos, debido a la falta de cobertura (Davey, *et al*., 2011).

La calidad de las lecturas obtenidas del secuenciamiento, fue muy buena; se obtuvo

aproximadamente 1,465 millones de lecturas, con un filtro de Quality Score (Q) superior a 30, esto significa que hubo una probabilidad del 0.0001 de que una base haya sido mal llamada; es decir 1 de cada 1,000 bases.

El valor “Q”, es considerado como el punto de referencia de calidad en el secuenciamiento NGS (Cock *et al*., 2010), mientras mayor sea, menor será la probabilidad de llamar erróneamente una base. Este es de suma importancia, ya que permite identificar bases que son producto de errores de secuenciamiento que, si no son filtradas o identificadas, se podrían considerar erróneamente como SNP durante la llamada de variantes, generando falsos positivos (Cock *et al*., 2010).

7.2. Procesamiento bioinformático de lecturas

Las lecturas resultantes de un genotipado por secuenciamiento, requieren de un procesamiento especial; estas difieren de las lecturas de un secuenciamiento de genoma completo, en que provienen de una representación reducida y de lugares aleatorios del genoma, que contienen los sitios de restricción, dejando vacíos de secuenciamiento (Wicklan *et al*., 2017).

Por ello, el análisis de estas secuencias, requiere una alineación individual de las lecturas y realizar varios supuestos estadísticos para la llamada de variantes (Wicklan *et al*., 2011). Se han desarrollado diferentes paquetes de software y flujos de trabajo bioinformáticos, para el procesamiento de este tipo de datos (Sonah, 2013; Rochette y Catchen *et al.,* 2017; Torkamaneh *et al.,* 2017;).

Varios de estos softwares, usan las mismas herramientas y algoritmos que se aplican comúnmente a los datos de secuenciamiento de genomas completos, mientras que otras utilizan algoritmos desarrollados específicamente para la secuenciación GBS-

ddRadseq (Catchen *et al*., 2011).

Existen algunos softwares con interfaz gráfica, fácil de entender y usar, como TASSEL/TASSEL-GBS (Bradbury *et al*., 2007; Glaubitz, 2014) muy usado en genética de cereales (específicamente maíz y cebada); existen también aquellos que usan solo líneas de comando en UNIX/LINUX, la mayoría de paquetes bioinformáticos trabajan bajo este modo: STACKS (Catchen *et al.,* 2013), IGST (Sonah *et al*., 2013), FAST-GBS (Torkamaneh *et al*., 2017), GB-eaSy (Wicklan *et al*., 2017), GATK (De Summa *et al*., 2017), entre otros.

El procesamiento de la data, se realizó usando el software libre STACKS versión 2.53 (Catchen *et al*., 2011), su elección fue por dos procesos importantes que realiza; 1) la identificación de todos los *loci* polimorficos (SNP) del conjunto de individuos de manera conjunta, y 2) la genotipificación de cada uno de estos, *locus* por *locus*, en base a un catálogo que construye con los SNP identificados. Para lograr esto, realiza el alineamiento de las lecturas, usando un genoma de referencia o bajo un enfoque *de novo* cuando se desconoce el genoma de la especie; esta flexibilidad de STACKS ayudo en el procesamiento de nuestra data genómica.

Para el alineamiento con el genoma de referencia de *M. esculenta* V.6.1, se usó el alineador BWA versión 0.7.12 (Borrow-Willer Alignmet) (Li y Durbín, 2009), con ello se logró un alineamiento promedio de 94.33% (Fig. 4). Este alto porcentaje fue debido: al buen secuenciamiento y caracterización del genoma de referencia de yuca (Bredeson *et al*., 2016), al corto tamaño de las lecturas usadas (150 pb) en el secuenciamiento, y a su asociación a los sitios de restricción. Sumando estas características, el mapeo fue más sencillo y específico (Poland *et al*., 2012)

En el software STACKS, se trabajó solo con lecturas que se lograron mapear

correctamente al genoma de referencia, usando el paquete *gstacks*, primero se identificó todos los *loci* polimórficos del conjunto de accesiones, con estos se construyó un catálogo de SNP. En este punto, también se obtuvo la profundidad de cobertura por locus, el valor promedio fue de 6.5X (Fig. 5), este valor es de mucha importancia, ya que, valida la existencia de cada base en su respectiva posición, mientras mayor sea el valor, mayor será la certeza de la existencia de dicha base en su respectiva posición, descartando la opción de error de secuenciamiento.

Finalmente, usando el catálogo de SNPs generado, se genotipificó a cada una de las 731 accesiones, *locus* por *locus* en un archivo VCF final (Lin *et al*., 2020).

7.2.1. SNP calling

Para la llamada de variantes en el paquete *gstacks*, se seleccionó el modelo de *Marukilow*, ya que se basa en un algoritmo bayesiano (BGC, por sus siglas en inglés), y porque fue diseñado para trabajar con genomas diploides, como el de la yuca (Maruki y Lynch, 2017).

Para que una variante sea considerada como SNP, se usó un umbral de 0.05; es decir, la variante tuvo que estar presente en al menos 5% de la población (37 accesiones) (Catchen *et al*., 2011). Con esto, se logró obtener inicialmente 264,088 SNPs para toda la población de la colección, valor que se encuentra dentro de lo reportado actualmente (20,712 – 444,821 SNPs) (Rabbi *et al*., 2015; de Albuquerque *et al*., 2018). Antes del uso de las nuevas técnicas de genotipado, la cantidad de SNP en yuca oscilaba entre 2,000 –3,000 (Fegurson *et al*., 2012; Pootakham *et al*., 2014); actualmente este número sigue en aumento debido a la aparición del genotipado a gran escala usando NGS.

Se realizó una distribución de estos SNPs en los 18 cromosomas de *M. esculenta*

Crantz, con la finalidad de verificar si existe una relación directa entre el tamaño del cromosoma y la cantidad de SNPs que alberga cada uno (Tabla 2). El resultado mostró que no existe dicha relación; por el contrario, se obtuvo porcentajes de distribución similares en cada cromosoma. En la Fig. 6, se observa esta distribución equilibrada, que aumento la probabilidad de selección de SNPs de diferentes lugares del genoma y aseguró una adecuada cobertura, además, ayudó en la precisión y fiabilidad en las estimaciones de la diversidad genética y la estructura de la población (de Albuquerque *et al*., 2018).

7.2.2. Filtrado de SNPs

Los filtros se aplicaron para reducir la cantidad de data faltante, que es poco informativa, y conservarla hace el archivo muy pesado y lento para el procesamiento.

El primer filtrado fue a nivel de variantes (264,088 SNPs); se excluyeron aquellos locus que tenían más del 20% de data faltante, es decir, el locus tuvo que existir por lo menos en 585 accesiones para ser considerado (Fig. 1 del Anexo.4) (Rabbi *et al*., 2014); aquí se eliminó al 60% de variantes iniciales, quedando 108,397 SNPs.

El segundo filtro fue a nivel de individuo, aquellos con demasiados datos faltantes, no aportan información sobre la población y solo aumentan el tamaño del archivo; se consideró un umbral de tolerancia del 55% de data faltante por individuo, con esto se eliminó a tres accesiones (ME_170. ME_264 y ME_310) (Fig. 7) quedando finalmente 728.

Considerar variantes o individuos con más del 50% de data faltante, disminuye la calidad de los resultados; sin embargo, los filtros aplicados dependen del tipo de análisis que se desea realizar.

El tercer filtro se realizó de acuerdo a la frecuencia del alelo menor (MAF) de cada locus-SNP, este valor es importante ya que una elección ambigua de MAF, altera los cálculos de diversidad y estructura genética, y ambos dependen de las frecuencias alelicas (Linck y Battey, 2019). En este filtro, para elegir el mejor valor de MAF, se tomó en cuenta los valores de PIC (Liu *et al*., 2017), la Fig. 2 del Anexo.4 muestra la distribución de los valores para cada alelo de los 108,397 SNPs hasta ahora filtrados.

Para la correcta elección de los valores de PIC, se tomó en cuenta la naturaleza bialélica de los SNP; en base a ello, el máximo valor que puede alcanzar un *locus*-SNP es de 0.5 (Kruglyak, 1997) (Fig. 2 del Anexo.4).

Actualmente no existe una clasificación de valores de PIC para marcadores SNP que indiquen su nivel de informatividad; como si lo hay para los SSR (Botstein, 1980), que pueden alcanzar valores mayores a 0.5 ya que pueden albergar más de dos alelos, sin embargo, se ha usado esta clasificación para validar el nivel de informatividad de los marcadores bialelicos, donde un valor de PIC mayor a 0.5 se considera altamente informativo; los marcadores con un valor de PIC de 0.25 a 0.5 son moderadamente informativos; y los marcadores con un valor de PIC inferior a 0.25 son ligeramente informativos (Luo *et al*., 2019).

En el caso de yuca, se reportó valores promedios de PIC de 0.23 (Ogbonna *et al*., 2021), 0.24 (de Albuquerque *et al*., 2018), 0.26 (de Oliveira *et al*., 2014) y 0.28 (Ferguson *et al*., 2012), con los que se obtuvieron buenos resultados en análisis de diversidad y estructura; estos valores además son respaldados por estudios en otros cultivos, donde también usan SNPs, como en maíz, en vid (Salmaso *et al*., 2005) y en el trigo de invierno (Eltaher *et al*., 2018).

Por otro lado, se ha demostrado que los bajos valores de PIC son compensados por la

cantidad de SNPs seleccionados, que cubren gran parte del genoma (Glover *et al*., 2010). Toda esta información respalda nuestro valor de PIC promedio (0.24) seleccionado, el cual se consiguió eliminando los SNPs con valores de MAF menores a 0.10 en el tercer filtrado.

Además, nuestra elección de un alto valor de MAF (0.05 – 0.15) permite conseguir variantes comunes en toda la población, mientras que los alelos raros están relacionados con bajos valores (0.007 – 0.01). Considerar o no alelos raros depende del objetivo del estudio; estos pueden ser muy útiles cuando se busca una visión cercana de los nuevos grupos que se están formando, ya que agrupará a todos aquellos que presenten dichos alelos (De la Cruz y Raska, 2014). Sin embargo, cuando se necesita una visión global de la población, que es el objetivo de nuestro estudio, los alelos comunes han demostrado ser más útiles (Mathieson y McVean., 2012; De la Cruz y Raska, 2014; Linck y Battey, 2019).

Con el tercer filtro se obtuvo finalmente 1,271 SNP con MAF ≥ 0.10 (Fig. 8) de buena calidad y altamente informativos.

7.3. Análisis de diversidad y estructura

7.3.1. Diversidad genética

Uno de los objetivos de la Colección Nacional de Yuca del INIA, es conocer la variabilidad genética que alberga, ya que es importante para el uso sostenible de los recursos genéticos, a fin de satisfacer la demanda de seguridad alimentaria futura y las estrategias de conservación (Prempeh *et al*., 2020).

Para conocer dicha diversidad, se seleccionó a los SNP como marcadores moleculares, por ser muy abundantes en el genoma, altamente polimorficos, codominantes y por

presentar una baja tasa de error durante el genotipado (Rafalski, 2002; Schlötterer, 2004; Chagné *et al*., 2007). Además, porque son los marcadores favoritos, para los estudios de genética en plantas, ya que la información que generan es altamente aprovechable (Kamanda *et al*., 2020).

Con los 1,271 SNPs, se obtuvo los valores de MAF, PIC, Ho y He, estos estadísticos se selccionaron ya que brindan importante información, para entender la diversidad genetica en yuca (Ferguson *et al*., 2011; Rabbi *et al*., 2014; Ferguson *et al*., 2019, Prempeh *et al*., 2020)

El análisis de los valores de PIC por separado, para cada país de la colección, fue mayor en Perú (0.2421) y Paraguay (0.2367), y menor en Bolivia (0.2330) y Brasil (0.2266) (Tabla 3). Perú alberga la mayor cantidad de accesiones (634) por ende hay mayor probabilidad de encontrar *loci* SNP polimorficos; de hecho, el filtrado de SNP se realiza en base a esta población. El resto de países conforma solo el 14% de la colección; sin embargo, sus valores de PIC son similares al de Perú a pesar de sus pocas accesiones; esto se explica porque presentan individuos altamente heterocigotos, cuyas frecuencias alélicas son altas, tal como se refleja en sus valores de MAF (Tabla 3).

Se analizó también los valores de PIC para las regiones con mayor cantidad de accesiones que conforman la población peruana, el mayor valor (0.2924) fue para San Martín y Lima (0.2426), esto sugiere que estas regiones albergan la mayor cantidad de *loci* SNP polimorficos, a pesar de no presentar la mayor cantidad de accesiones, esto se refleja en la frecuencia de su alelo menor (MAF) (Tabla 4). Por otro lado, la región Ucayali, que es la más representada (318 accesiones), tiene el valor más bajo de PIC (0.2288) (Tabla 4), esto se explica por sus bajas frecuencias alélicas (MAF = 0.1881), que refleja un mayor número de *loci* homocigotos.

En general se obtuvo bajos valores de PIC si los comparamos a los obtenidos cuando se usan marcadores tipo SSR, donde los valores varian de 0.09 – 0.86 (Fregene *et al*., 2003; Raji *et al*., 2009; de Bang *et al*., 2011). Esto se debe como ya se mencionó a la naturaleza bialelica de los SNP, donde los valores de PIC obtenidos no pueden ser superiores de 0.5 (de Oliveira *et al*., 2014).

Los diferentes valores de PIC, obtenidos para cada país y departamento se deben al análisis por separado que se realizó, donde la frecuencia alelica de cada SNP en cada grupo será diferente, ya que esto depende principalmente de la composición de sus genotipos (Karim *et al*., 2020; Prempeh *et al*., 2020).

La heterocigocidad observada (Ho) (0.319) fue mayor al valor de la heterocigosidad esperada (He) (0.293) en toda la colección (Tabla 3), este resultado indica la existencia de accesiones con un elevado número de *loci* heterocigotos, en comparación a una población bajo equilibrio Hardy-Weinberg. Valores cercanos a la Ho obtenida, fueron reportados por otros estudios en yuca, por ejemplo, se obtuvo un valor de 0.32 al estudiar 1,280 accesiones de Brasil (de Oliveira *et al*., 2014) y 0.314, al estudiar 105 accesiones de Ghana (Prempeh *et al*., 2020).

El resultado de un mayor valor de Ho respecto a la He, también fue reportado por otros autores (Ferguson *et al.,* 2012; Ferguson *et al*., 2019), esto se atribuye al elevado número de individuos heterocigotos en la población. Contrario a nuestros resultados otros autores reportaron valores de He mayores a la Ho (de Albuquerque *et al*., 2018; Prempeh *et al*., 2020), ellos explican que este resultado es debido a un exceso de individuos homocigotos y a la existencia de accesiones duplicadas en sus poblaciones.

Nuestros valores de diversidad son bajos en comparación a los obtenidos al usar SSR para la misma finalidad de estudio. Esto se debe como ya se explico, a la naturaleza

bialelica de los SNP y a su distriución en la población que determina sus frecuencias alelicas; sin embargo, al usar gran cantidad de estos y al estar distribuidos por todo el genoma, tienen el mismo nivel de discriminación en comparación a los marcadores multialélicos (de Oliveira *et al*., 2014). Sumado a estas caracteristicas, los SNPs permiten una adecuada discrininación de la diversidad intraespecifíca de la colección.

En el análisis por país, se encontró el mismo resultado, una Ho mayor a la He; Perú y Brasil presentaron valores cercanos 0.319 y 0.316 (Tabla 3) respectivamente, estos son los países con mayor cantidad de accesiones en la colección (634 y 47 respectivamente). La facilidad de encontrar diferentes variedades en estos países, se debe principalmente a que la yuca ha sido cultivada como práctica ancestral por los diferentes pueblos indígenas amazónicos que albergan, además, poseen zonas que pertenecen a la cuenca amazónica, lugar donde se encuentra la mayor diversidad de yuca (Olsen y Schaal, 1999).

La práctica del cultivo ancestral de yuca sigue vigente hasta la actualidad en estos lugares, y ha permitido el desarrollo voluntario e involuntario de nuevas variedades, aumentando la variabilidad genética de este cultivo en ambos países. Brasil es considerado como el centro de origen de la yuca, ya que es el país que alberga la mayor diversidad genética (Olsen, 2004); por ello existe mayor probabilidad de encontrar individuos heterocigotos, aún a pesar de estudiar un pequeño lugar o zona de este país; el valor de su Ho encontrada en este estudio, es muy similar a la reportado por otros donde se usan SNP, como marcadores moleculares (de Oliveira *et al*., 2014; de Albuquerque *et al*., 2018;).

Los mayores valores de Ho de los paises que forman parte de la Colección Nacional, se encontraron en Bolivia (0.3304) y Paraguay (0.3232) (Tabla 3), estos poseen 24 y 23 accesiones respectivamente y son altamente heterocigotos entre si. Esto se debe a que

provienen de colecciones nacionales de sus respectivos países, cuyas diferentes variedades se encuentran correctamente curadas, clasificadas y conservadas; sus altos valores de Ho, indican que se esta garantizando una adecuada representación de su diversidad.

En el caso de Perú, el elevado valor de su Ho (0.319) se ve reflejado también, en el estudio de sus regiones por separado (Tabla 4); Huánuco, perteneciente a la región andina, es la región con el mayor valor de Ho (0.3253), Ucayali y San Martín pertenecientes a la Amazonía peruana, poseen valores de Ho de 0.322 y 0.320 respectivamente (Tabla 4). Una de las razones para que estas tres regiones presenten altos valores, es que albergan a pueblos indígenas andino-amazonicos; Ucayali posee 14.4% de población indígena, distribuida en 257 comunidades, San Martín posee 5%, distribuida en 90 comunidades, y Huánuco el 1%, distribuída en 18 comunidades; estos se dedican al cultivo tradicional de yuca, como una de sus actividades ancestrales más importantes (Aparicio y Bodmer, 2009).

Los pueblos indígenas de la Amazónia peruana tienen a la yuca como cultivo de subsistencia, presentan un sistema agrícola tradicional denominado “chacra” donde se cultivan diferentes variedades locales (etnovariedades). Un grupo de variedades son consideradas importantes por tradición, por las mujeres indigenas y los jefes de tribu, y son mantenidas por reproducción asexual con estacas, y el otro grupo, consiste en nuevas variedades, que se han incluido porque presentan características atrayentes para su cultivo y mantenimiento (Salick *et al*., 1997).

La mayoría de estas nuevas variedades, provienen de una reproducción sexual a partir de las ya sembradas en la misma “chacra”, a este fenómeno se conoce también como el *volunteer seedlings* (Salick *et al*., 1997, Emperaire *et al*., 1998). Estas plantas son capaces de ser diferenciadas por los pueblos indígenas, de aquellas que son producto

de la reproducción clonal por estaca (OECD 2016), habilidad que es de suma importancia, ya que permite aumentar las variedades dentro de la “chacra” y mantenerlas a través del tiempo, aumentando así la diversidad genética.

Esta alta diversidad en yuca, también ha sido reportada en varios pueblos indígenas amazónicos, donde se usaron caracteres morfológicos para su descripción; por ejemplo, el pueblo “Ticuna”, que habita la región Loreto en Perú y las fronteras de Brasil y Colombia, considera a la yuca un alimento de origen sagrado, razón por la cual forma parte de su dieta diaria y está siempre presente en sus “chacras”. Aquí se han caracterizado 23 variedades (Brañas *et al*., 2019); en el pueblo “Amuesha o Yanesha”, que habitan las regiones de Huanuco, Pasco y Junín, se ha reportado que solamente en Oxapampa-Pasco se cuenta con 39 variedades (Salick *et al*., 1997); los “Ashaninka” que habitan las regiones de Pasco, Ayacucho, Cusco, Huánuco, Junín y Ucayali, poseen 133 cultivares; los “Bora-Hitoto” de la región Loreto, cuentan con 18 cultivares; los “Cocoma-Cocamilla” de Loreto cuentan con 33 cultivares y el pueblo “Aguaruna” de Amazonas, Loreto y San Martín, cuentan con 31 cultivares (Mejía, 2002).

Además de los caracteres morfológicos, también se han usado marcadores moleculares (SSR), que respaldan esta alta diversidad en el Perú. Usando 22 caracteres morfológicos y 12 SSR se ha encontrado, que en la cuenca baja del río Ucayali existe una elevada diversidad genética en 26 variedades (Mejía *et al*., 2015); en la ciudad de Aguaytia – Ucayali también se encontró alta diversidad en 100 accesiones caracterizadas usando 22 SSR (Villasante, 2006).

Otra actividad importante que respalda la alta diversidad de yuca, es el intercambio de estacas de diversas variedades, producto de la reproducción sexual o asexual. Esta es una practica muy común entre los habitantes de los pueblos indígenas (Elias *et al*., 2001), que se debe principalmente por lazos de parentesco o reciprocidad (Brañas *et*

al., 2019). Por otro lado, también se realiza el intercambio de variedades de interés comercial con buenas caractersitcas de producción, aptas para la comercialización y el autoabastecimiento, sobre todo en la costa y sierra de nuestro país (Mejía *et al*., 2015).

En principio se podría pensar que el intercambio de estacas, sería un factor que afecta negativamente la diversidad, ya que conduciría a la homogenización genética de las variedades como lo establece la teoría clásica de genética de poblaciones, cuando la migración es efectiva entre un número limitado de poblaciones y existe panmixia entre las variedades introducidas y las nativas (Varvio *et al*., 1986). Sin embargo; estos intercambios también pueden ser positivos, cuando se sigue cultivando variedades locales propias, como es la costumbre de las comunidades y pueblos indígenas que habitan la cuenca del Amazonas (Elias y McKey, 2000; Kawa *et al*., 2013).

Las redes de intercambio geográficas y sus dinámicas sociales en torno a la yuca, son un factor importante, ya que determinan la variabilidad presente en un territorio (Delaquis *et al*., 2018). Suumado a esto, el grado de relación entre los pueblos indígenas amazónicos que integran estas redes y el compartir su conocimiento acerca de las prácticas etnobotánicas y agrícolas, son determinantes para explicar el grado de diversidad genética presente en nuestro país.

Las regiones de Lima y La Libertad presentaron valores de Ho de 0.3196 y 0.3149 respectivamente, menores a los de otras regiones; aún así, siguen presentando altos valores de diversidad genética. Como se mencionó, el centro de mayor diversidad genética para este cultivo, es la cuenca del Amazonas (Olsen y Schaal, 1999), lugar donde se debería encontrar los restos más antiguos de yuca, pero debido a su naturaleza vegetal no existen registros arqueológicos en estas zonas; sin embargo, en la costa peruana, se ha logrado registrar algunos restos arqueológicos que evidencian su cultivo y consumo por las culturas que la habitaron: en la Quebrada de Pircas, en el

valle de Zaña, en la región Lambayeque, se encontró restos arqueológicos que evidencian el cultivo temprano, de algunas variedades exógenas de yuca introducidas en la zona (Isendahl, 2011). Tambien se encontraron restos arqueológicos en las costas de Ancón – Chillón en la Región Lima, con una antigüedad de 4,500 años (Seminario, 2004); estos datos muestran que el cultivo de yuca se realizó tempranamente en la costa y que las variedades usadas fueron introducidas, probablemente del intercambio con los pueblos indígenas amazónicos.

Por otro lado, el desarrollodo de las técnicas y métodos de cultivo y el aumento del consumo de yuca en la costa del territorio nacional, ha provocado un incremento en la comercialización, intercambio y desarrollo de variedades con caractersiticas elevadas de producción, adaptabilidad y rendimiento para abastecer al mercado interno (Mostajo, 2017). Esto ha contribuido al aumento de diversidad genética en departamentos donde no se suele cultivar yuca como es el caso de la costa.

Los altos valores de diversidad encontrados en la selva respecto a los encontrados en la costa de nuestro país, se deben también a que en la primera existen las denominadas áreas de montaña con predominio de laderas, donde se suele sembrar diversas variedades de yuca, ya que estas condiciones no son aptas para muchos otros cultivos. Por otro lado, las zonas de llanura, como es el caso de la costa, suelen tener condiciones ambientales más homogeneas, esto conduce a una competencia más fuerte en la siembra selectiva de variedades, inclinandose por las comercialmente importantes, factor que influye en la disminución de su diversidad (Willemen *et al*., 2007).

Otro factor que podemos considerar, es el planteado por Brush (1995), que considera que áreas economicamente aisladas como las de montaña en la selva, generalmente tienen mayor diversidad en comparación a zonas de agricultura intensiva.

De los resultados presentados es evidente la alta diversidad genética de yuca en Perú,

representada por un amplio repertorio de variedades. Esta se encuentra asociada sobre todo a zonas donde se lleva a cabo un manejo tradicional del cultivo y con un vinculo sociocultural, tal como se observa en los pueblos indígenas (Elias *et al*., 2001; Emperaire y Peroni, 2007; Heckler y Zent, 2008); a esto se puede sumar la biología reproductiva de la planta (Rival y Mckey, 2008): la yuca es monoica, posee flores masculinas y femeninas en la misma planta con diferente tiempo de apertura, a su vez es alógama, lo que permite la polinización cruzada; provocando que la yuca sea altamente heterocigota, generando descendencia híbrida fértil.

Además, las plantas producto de la reproducción sexual (*volunteer seedlings*) han tomado un papel protagonico, y se considera actualmente un fenómeno muy importante y lo suficientemente frecuente, que permite explicar también la diversidad intravarietal de yuca observada en un sistema tradicional de cultivo (Elias *et al*., 2000; Emperaire y Peroni, 2007; Heckler y Zent., 2008).

7.3.2. Estructura genética

Para determinar cómo se organiza la diversidad genética contenida dentro de una población, se necesita identificar y clasificar en grupos homogéneos aquellos que guardan algún tipo de relación; y encontrar los datos cuantitativos que respaldan estos grupos es el principal desafío (Patterson *et al*., 2006).

En diferentes estudios se ha buscado agrupar a las variedades locales de yuca peruana bajo diferentes criterios, en base a sus características morfológicas, como yucas “dulces” o “bravas” (Brañas *et al*., 2019), y en función a sus características fisiológicas y de adaptación, como “muy precoces”, “semi precoses”, “tardías” y “semi tardías” (Mejía, 2002; Meza y Julca Otiniano, 2015); hasta la fecha solo se han usado marcadores moleculares SSR, para determinar más específicamente la estructura genética en variedades locales de Ucayali (Villasante, 2006; Mejía *et al*,. 2015), la

presente investigación es el primer reporte utilizando marcadores SNP para accesiones colectadas a lo largo del territorio nacional (Anexo.1).

La Colección Nacional de Yuca del INIA, se encuentra clasificada de forma *a priori* por los datos pasaporte de localización geográfica, formando grupos en base a su lugar de origen; usando este criterio de agrupación se realizó un análisis AMOVA, Fst y PCA, para corroborar si esta clasificación logra explicar el total de variación contenida en la colección.

El análisis realizado en dos niveles (país y región de procedencia), muestra en el AMOVA que la varianza molecular de los SNP, se concentra dentro de los grupos de clasificación (más del 87% en base a país y más del 90% en base a regiones peruanas); este resultado también fue reportado por estudios donde usan SNP (de Oliveira *et al*., 2014; de Albuquerque *et al*., 2018) y SSR con el mismo criterio de agrupación (Kawuki *et al*., 2013).

El valor de Fst promedio entre los países fue de 0.0645, indicando una baja diferenciación genética entre estos (Wright, 1965). Nuestro valor es cercano (0.073) al obtenido por Ogbonna *et al.* (2021), quién uso SNPs para estudiar la estructura genética del Banco de Germoplasma de Yuca del EMBRAPA; e incluso mayor a los valores de Fst reportados en otros estudios de estructura genética, donde usaron SSRs por ejemplo; en el germoplasma de Brasil se obtuvieron valores de 0.073 (Gonçalves *et al*., 2017), 0.070 (Ferreira *et al*., 2015) y 0.04 (Mezette *et al*., 2013); al estudiar variedades locales de Puerto Rico se obtuvo un valor de 0.03 (Montero-Rojas *et al*., 2011) y al estudiar la diversidad genética entre dos grupos de accesiones en Uganda se obtuvo un valor muy bajo de 0.025 (Turyagyenda *et al*., 2012).

Existen varios reportes de bajos valores de Fst, cuando se estudia diversos grupos de

germoplasma de yuca, bajo diferente criterio de agrupaciones y usando tanto marcadores SNP como SSR. Considerando solamente los efectos de los marcadores moleculares, en algunos trabajos, se ha indicado un mejor poder de detección de la diferenciación genética entre poblaciones usando SNP, en comparación a los marcadores microsatélites (Mba *et al*., 2010; Garcia-Lor *et al*., 2013).

Por otro lado, se obtuvo altos valores en el "parwise Fst" (Tabla 5), donde la mayor diferencia se encontró entre Perú y Brasil (0.0869) respaldado por la distancia genética de Nei (Tabla 6). Como se menciona anteriormente ambos países se encuentran en la cuenca del Amazonas, lugar con la mayor diversidad mundial de yuca. De este resultado podemos inferir que, a pesar de tener un intercambio de germoplasma, cada país aún conserva sus propias variedades locales dentro de las "chacras" (Alzate *et al*., 2010).

En el "parwise Fst" entre regiones, los valores más altos se encontraron entre los germoplasmas de Ucayali y el resto de regiones, siendo el mayor valor con Lima (0.06198) y La Libertad (0.06156); esto se respalda por su separación geográfica que disminuye el nivel de intercambio de germoplasma. Por otro lado, se corroboró en resultados posteriores, que Ucayali conserva un grupo genético exclusivo de variedades, que estarían aumentando el valor de su diversidad ya que aún no han sido distribuidos en otros departamentos.

Muchos autores han intentado establecer una asociación entre el origen geográfico y la divergencia genética en la yuca, sin embargo; los resultados han sido bastante contradictorios. Ferguson *et al*. (2012), cuando analizó 53 accesiones de yuca usando marcadores SNP, no observó ninguna estructuración basada en el origen geográfico, de igual manera Devi *et al.* (2009) no encontró ningún patrón de distribución de las accesiones de yuca, según las diferentes regiones ecogeográficas de recolección; lo que indica que las diversidades geográficas y genéticas no estaban relacionadas.

En general los bajos valores de Fst (Tablas 7 y 8), indican una distribución homogenea de la diversidad genética entre los departamentos con un nivel alto de heterocigosidad, lo cual se explica por el continuo intercambio de material vegetal (Elias *et al.*, 2000; Elias *et al.,* 2001); sobre todo de aquellas variedades que son muy consumidas y comercializadas que se pueden encontrar en la selva y costa de nuestro país (Mejía *et al*., 2015). Por otro lado, los valores del "pairwise Fst", corroboran este intercambio, ya que se encuentra en relación directa a la distancia geográfica; mientras más cercanos estan dos departamentos, menor es su valor, tal como se observa en Lima que ha recibido mayor influencia genética de Húanuco y La Libertad (Tabla 8), siendo el germoplasma de estos dos últimos muy similares (0.00172).

Usando el mismo criterio de agrupación (país y departamento de origen geográfico) los resultados del PCA (Figs. 10 y 12 respectivamente), tampoco muestran una correspondencia entre el genotipo de cada accesión con su lugar de origen, ya que no se observa ningun tipo de agrupación bajo este criterio; sin embargo, cuando se analiza la calidad de representación de la diversidad de las accesiones en los componentes, usando el cos^2, se encontró una correspondencia entre genotipo y lugar de origen solo en un grupo de accesiones de Ucayali (Figs. 11 y 13), respaldados por su elevado valor de cos^2 ($0.6 < cos^2 < 0.8$).

Para observar mejor la distribución de genotipos en los países y departamentos, se realizó el diagrama de barras apiladas compuestas usando los valores de PAI de cada accesión, obtenidos del DAPC (Fig. 14). Este gráfico refleja la existencia de varios genotipos que no se corresponden con el número de países y departamentos de la colección, demostrando también la no correspondencia entre genotipo y lugar de origen geográfico.

Al ordenar la gráfica de barras apiladas compuestas por grupo de procedencia (Fig. 15),

se corrobora la existencia de varios genotipos por país o departamento, se observa también, que existe influencia genética entre estos y la cantidad que esta implica para cada uno, ya que porciones de genotipos de un grupo se observan en otros.

La influencia de los genotipos entre países es muy baja (Fig. 15); sin embargo, dentro del germoplasma peruano, las regiones con mayor influencia genética son Huánuco, San Martín, Lima, Junín y Ancash (Fig. 15). En el caso de las regiones con la mayor cantidad de accesiones, Huánuco presenta alta influencia genética de San Martín y Ancash, en el caso de San Martín presenta influencia de Lima y Huánuco, Lima presenta mayor influencia de San Martín. Esto se respalda también, en los valores del "parwise Fst" (Tabla 7) y en la distancia genética de Nei (Fig. 5 del Anexo.4). Bajo este mismo criterio de análisis Ucayali, no presenta influencia genética de ningún departamento y viceversa; esto respalda la existencia de grupos genéticos propios de esta región.

En cuanto a las regiones con poca cantidad de accesiones (menos de 12) la influencia genética también fue menor (Fig. 15). Las regiones de Apurímac, Cusco, Lambayeque, Arequipa e Ica, no presentan influencia de genotipos; Cajamarca y Tumbes solo presentan influencia de La Libertad y Lima respectivamente; mientras que Amazonas, Junín y Piura presentan influencia de genotipos de dos o más regiones. Para poder respaldar los genotipos propios de estas regiones en este análisis, es necesario tener una mejor representación de grupo, para ello se debe incrementar la cantidad de accesiones.

Para lograr explicar la distribución de la diversidad genética contenida en la colección, se buscó la mejor forma de agrupación, basada en la información molecular de los SNP. Para ello se uso el análisis de DAPC; en este se mantuvo a 350 componentes principales (CP) que explican más del 90% de la variación genética total (Tabla 10), la correcta elección del número de CP está relacionado con el poder de reducción de la dimensión

de los datos.

En general, muchos estudios han indicado el uso de componentes que retienen más del 80% de la varianza genética (de Oliverira *et al.*, 2014). En el contexto de DAPC, el punto de equilibrio debe determinarse entre el poder de discriminación de los grupos y la estabilidad de las atribuciones de los genotipos en cada grupo (Jombart *et al.*, 2010). Por tanto, el análisis con 350 CP garantizó un alto poder estadístico en la evaluación de la estructura genética del germoplasma de la colección.

En la Figura 8 del Anexo.4, se observa que la Colección Nacional de Yuca del INIA presenta una estructura jerárquica e ilustra el proceso de selección del número ideal de grupos. Según Jombart *et al.* (2010) "el verdadero" número de grupos bajo el Criterio de Información Bayesiana, es el menor valor de BIC, que corresponde con los quiebres de la curva.

Sin embargo, el mejor valor de BIC, no esta tan claro para nuestra población, ya que existe más de un menor valor y varias caídas en la curva, pudiendo seleccionar un valor de "K" entre 12 y 21. La caída inicial más abrupta del BIC es para un K=12 y para no caer en una sobreestimación de la diferenciación de los grupos, se seleccionó a este valor como el mejor "K".

Por otro lado, de Oliverira *et al.* (2014), señala que en presencia de estructura genética en cline (un tipo particular de variación continua) como es el caso de la yuca, los valores de BIC pueden reducirse después de la identificación del valor de "K" real. Tomando en cuenta esto, se consideró una reducción de los valores de BIC para la elección del valor de "K" en base al supuesto verdadero valor que se encontraría dentro del rango mencionado (Jombart *et al.*, 2010).

Con este análisis se clasificó y ordenó toda la información genética de la colección en

12 grupos genéticos (K=12) (Fig. 16), respaldado por los análisis estadísticos realizados (Figs. 8-13 del Anexo.4) (Jombart *et al*., 2010).

En otros estudios también se respalda el uso del DAPC, para análisis de estructura poblacional en yuca (de Oliveira *et al*., 2014; de Albuquerque *et al*., 2018); donde mencionan que en especies de propagación clonal como este cultivo, las variedades a menudo se derivan de cruces complejos, que puede influir en la precisión de la definición de los grupos genéticos (Rabbi *et al*., 2015), en general, el valor de K sugerido por el DAPC está muy cerca del valor real, además, no realiza supuestos de cruzamientos aleatorios, ligamiento y EHW (Jombart *et al*., 2010), como si lo hacen otros modelos como STRUCTURE (Pritchard *et al*., 2000).

Los valores de PAI obtenidos, respaldan aún más la asignación de las diferentes accesiones a cada grupo genético definido (Tabla 12), todos presentaron valores cercanos a 0.99, indicando que es un método muy fiable para el análisis de estructura poblacional.

En la gráfica de DAPC (Fig. 16) de la distribución de los grupos definidos, el resultado más resaltante es la separación de los grupos 8, 9, 1 y 3; donde 1 y 3 están formados en su mayoría por accesiones de Brasil y Bolivia. De esto podemos decir, que ambos países presentan genotipos propios, que no han tenido mucha influencia en Perú (Figs. 16 y 17). Lo mismo pasa con los grupos 8 y 9, que son los únicos formado por accesiones de Ucayali; con esto se corrobora la existencia de genotipos propios en esta región, que no han sido reportados aún en otras regiones (Tabla 11 y Fig. 17).

En cuanto a los grupos genéticos 2, 4, 5, 6, 7, 10, 11 y 12, están formados por accesiones de diferentes departamentos, los que probablemente albergan varios genotipos, o que por el contrario corresponden a uno dominante, que pudo haber llegado

a los distintos departamentos debido al elevado intercambio de material vegetal, que posteriormente pudo ser sometido a selección natural o artificial generando nuevas variedades locales (Elías *et al*., 2001; Mejía *et al*., 2015).

Otro hecho importante que respalda la existencia de varios genotipos por grupo, es el cambio del nombre vernacular que experimentan algunas variedades durante su intercambio, lo que conduce a la existencia de un solo genotipo, con diferentes nombres o diferentes genotipos con el mismo nombre (Villasante, 2006; Mejía *et al*., 2015; Brañas *et al*., 2019). Para poder corroborar si los 12 grupos definidos albergan varios genotipos, es necesario realizar un análisis de estructuración interna, donde se debe definir los niveles en el cual, se debería considerar dos genotipos como iguales o diferentes.

Por lo tanto, el movimiento de material vegetal por medio de estacas, en el territorio nacional y los cambios en los nombres de las variedades locales, contribuyen a que los grupos definidos por el análisis de DAPC, no muestren correlación con el origen geográfico. Sin embargo, en base a la información molecular de los SNP, la diversidad contenida en la colección se logra explicar en 12 grupos genéticos.

8. CONCLUSIONES

Del análisis de genotipado para las 731 accesiones de la Colección Nacional de Yuca del INIA, se logró identificar 264,088 SNPs utilizando el genoma de referencia de *M. esculenta* Crantz V.6.1, estos se distribuyen a lo largo de sus 18 cromosomas.

En el análisis de diversidad y estructura genética realizado, usando 1,271 SNPs altamente informativos, se encontró una elevada variabilidad genética entre las accesiones. Estas se agrupan en 12 grupos genéticos; solo en algunos casos se encontró correlación con la distribución geográfica, como fue con los genotipos de Brasil

y Bolivia que poseen poca influencia de alelos de otros países.

A nivel de las accesiones de yuca del Perú, con el análisis de agrupamiento en base a SNPs, se logró identificar dos grupos genéticos con genotipos de Ucayali, que presentan alelos propios o privados, no encontrados en otras regiones geográficas.

Los elevados valores de variabilidad genética, se explican por el intercambio de material vegetal entre departamentos, por la naturaleza heterocigota de la yuca, el cultivo tradicional en las “chacras”, las practicas socioculturales de los pueblos indígenas andino-amazónicos, y por el fenómeno del *voluntier seedling*.

9. RECOMENDACIONES

Se debe continuar con la utilizacion de los SNPs identificados, puesto que los resultados de esta tesis van a contribuir a una mejor caracterización de la Colección Nacional de Yuca del Banco de germoplasma del INIA. La continuacion de otros estudios, con estos marcadores, ayudará a una mejor clasificación de las accesiones y a la identificación de rutas de intercambio de material vegetal. También un análisis con SNPs, servira para identificar zonas geográficas donde aún no se han realizado colectas, y que probablemente contengan variedades locales. Finalmente, contribuirá a identificar una “colección núcleo”, representativa de la diversidad nacional de yuca.

10. REFERENCIAS BIBLIOGRÁFICAS

Allem, A. (2002). The origins and taxonomy of cassava. In R. J. Hillocks, J. M. Thresh, & A. Bellotti (Eds.), Cassava: Biology, production, and utilization (pp. 1–17). Wallingford: CAB International Wallingford, UK.

Alzate, A. M., Vallejo Cabrera, F. A., Ceballos Lascano, H., Pérez, J. C., & Fregene, M. (2010). Variabilidad genética de la yuca cultivada por pequeños agricultores de la región Caribe de Colombia. *Acta agronómica*, *59*(4), 385-393.

Aparicio, P. M., & Bodmer, R. E. (2009). *Pueblos indígenas de la Amazonía peruana* (p. 339). Pedro Mayor Aparicio.

Asare, P. A., Galyuon, I. K. A., Sarfo, J. K., & Tetteh, J. P. (2011). Morphological and molecular based diversity studies of some cassava (*Manihot esculenta* Crantz) germplasm in Ghana. *African Journal of Biotechnology*, *10*(63), 13900-13908.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376.

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., & de Leon, N. (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, *193*(4), 1073-1081.

Bertram, R.B. (1993). Application of molecular techniques to genetic resources of cassava (*Manihot esculenta* Crantz—Euphorbiaceae): interspecific evolutionary relationships and intraspecific characterization. Ph.D. Dissertation, University of Maryland, College Park, Maryland, USA.

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, *32*(3), 314.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633-2635.

Brañas, M., Nùñez Pérez, C. D. C., Zárate Gómez, R., Silverstein, S., & Del Aguila

Villacorta, M. (2019). Conocimientos tradicionales vinculados a la" yuca" *Manihot esculenta* (Euphorbiaceae) en tres comunidades ticuna del Perú. *Arnaldoa*, *26*(1), 339-348

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., & Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature biotechnology*, *34*(5), 562-570.

Brush, S. B. (1995). In situ conservation of landraces in centers of crop diversity. *Crop science*, *35*(2), 346-354.

Carmo, C. D. D., Silva, M. S. D., Oliveira, G. A. F., & Oliveira, E. J. D. (2015). Molecular-assisted selection for resistance to cassava mosaic disease in *Manihot esculenta* Crantz. *Scientia Agricola*, *72*, 520-527.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3 Genes|Genomes|Genetics*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics*, *19*(3 Pt 1), 233.

Cavalli-Sforza, L. L., & Bodmer, W. F. (1971). *The genetics of human populations.* W. H. Freeman.

Chagné, D., Batley, J., Edwards, D., & Forster, J. W. (2007). Single nucleotide polymorphism genotyping in plants. In *Association mapping in plants* (pp. 77-94). Springer, New York, NY.

Chase, M. W., & Hills, H. G. (1991). Silica gel: an ideal desiccant for preserving field-collected leaves for use in molecular studies. *Taxon*, *40*, 215-220.

Chen, C., Mitchell, S. E., Elshire, R. J., Buckler, E. S., & El-Kassaby, Y. A. (2013). Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genetics & Genomes*, *9*(6), 1537-1544.

Chesnokov, Y. V., & Artemyeva, A. M. (2015). Evaluation of the measure of polymorphism information of genetic diversity. *Sel'skokhozyaistvennaya Biologiya* [*Agricultural Biology*], 2015, Vol. 50, № 5, pp. 571-578.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767-1771.

Corander, J., Waldmann, P., Marttinen, P., & Sillanpää, M. J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, *20*(15), 2363-2369.

da Costa, T. R., Vidigal Filho, P. S., Gonçalves-Vidigal, M. C., Galván, M. Z., Lacanallo, G. F., da Silva, L. I., & Kvitschal, M. V. (2013). Genetic diversity and population structure of sweet cassava using simple sequence repeat (SSR) molecular markers. *African Journal of Biotechnology*, *12*(10).

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510.

Dawson, K. J., & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetics Research*, *78*(1), 59-77.

de Albuquerque, H. Y. G., Carmo, C. D. D., Brito, A. C., & Oliveira, E. J. D. (2018). Genetic diversity of *Manihot esculenta* Crantz germplasm based on single-nucleotide polymorphism markers. *Annals of applied biology*, *173*(3), 271-284

de Bang, T. C., Raji, A. A., & Ingelbrecht, I. L. (2011). A multiplex microsatellite marker kit for diversity assessment of large cassava (*Manihot esculenta* Crantz) germplasm collections. *Plant Molecular Biology Reporter*, *29*(3), 655-662.

de Heredia, U. L. (2016). Las técnicas de secuenciación masiva en el estudio de la diversidad biológica. *Munibe Ciencias Naturales*, *64*, 7-31.

De la Cruz, O., & Raska, P. (2014). Population structure at different minor allele frequency levels. In *BMC proceedings* (Vol. 8, No. 1, pp. 1-5). BioMed Central.

de Oliveira, E. J., Ferreira, C. F., da Silva Santos, V., de Jesus, O. N., Oliveira, G. A. F., & da Silva, M. S. (2014). Potential of SNP markers for the characterization of Brazilian cassava germplasm. *Theoretical and Applied Genetics*, *127*(6), 1423-1440.

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., & Tommasi, S. (2017). GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC bioinformatics*, *18*(5), 57-65.

Delaquis, E., de Haan, S., & Wyckhuys, K. A. (2018). On-farm diversity offsets environmental pressures in tropical agro-ecosystems: a synthetic review for cassava-based systems. *Agriculture, Ecosystems & Environment*, *251*, 226-235.

Demarchi, D. A. (2009). Microsatélites, distancias genéticas y estructura de poblaciones nativas sudamericanas. Revista Argentina De Antropología Biológica, 1(1), 73-88.

Devi AKB, Gin G, Rahul Y (2009) Assessment of genetic diversity in cassava (*Manihot esculenta* Crantz) germplasm. J Root Crops 35:108–111

Dominguez, O. D., Rodriguez, A. G., De Leon, F. J. G., & Fernandez, A. C. (2019). Diversidad genética. La biodiversidad en Michoacán. Estudio de Estado 2, 3.

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull., 19, 11-15.

Elias, M., & McKey, D. (2000). The unmanaged reproductive ecology of domesticated plants in traditional agroecosystems: An example involving cassavaand a call for data. *Acta Oecologica*, *21*(3), 223-230.

Elias, M., Panaud, O., & Robert, T. (2000). Assessment of genetic variability in a traditional cassava (*Manihot esculenta* Crantz) farming system, using AFLP markers. *Heredity*, *85*(3), 219-230.

Elias, M., Penet, L., Vindry, P., McKey, D., Panaud, O., & Robert, T. (2001). Unmanaged sexual reproduction and the dynamics of genetic diversity of a vegetatively propagated crop plant, cassava (*Manihot esculenta* Crantz), in a traditional farming system. *Molecular Ecology*, *10*(8), 1895-1907.

Eltaher, S., Sallam, A., Belamkar, V., Emara, HA, Nower, AA, Salem, KFM y col. (2018). La diversidad genética y la estructura de la población de los genotipos de trigo de invierno de Nebraska F3: 6 usando genotipificación por secuenciación. Parte delantera. Gineta. 9:76. doi: 10.3389 / fgene.2018.00076

Elshire R.J, Glaubitz J.C, Sun Q., Poland J.A, Kawamoto K., Buckler E.S. and Mitchell S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE, 6, e19379.

Emperaire, L., & Peroni, N. (2007). Traditional management of agrobiodiversity in Brazil: a case study of manioc. *Human Ecology*, *35*(6), 761-768.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.

FAO. (2009). Guía técnica para producción y análisis de almidón de Yuca. Obtenido de http://www.fao.org/docrep/pdf/010/a1028s/ a1028s01.pdf

FAO (Food and Agriculture Organization of the United Nations). 2021. FAOSTAT Cultivos y productos de ganadería [online]. Disponible en: http://www.fao.org/faostat/es/#data/QCL/visualize [Acceso: 26 agosto 2021].

Ferguson, M. E., Hearne, S. J., Close, T. J., Wanamaker, S., Moskal, W. A., Town, C. D., & de Villiers, E. P. (2012). Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. *Theoretical and Applied Genetics*, *124*(4), 685-695.

Ferguson, M. E., Shah, T., Kulakow, P., & Ceballos, H. (2019). A global overview of cassava genetic diversity. *PLoS ONE*, *14*(11), e0224763.

Ferreira, R. C. U., Vidigal Filho, P. S., Gonçalves-Vidigal, M. C., Moiana, L. D., & Kvitschal, M. V. (2015). Genetic and population structure of sweet cassava (*Manihot esculenta* Crantz) germplasm collected from Campo Grande, Mato Grosso do Sul, Brazil. *Australian Journal of Crop Science*, *9*(5), 458.

Fregene, M. A., Vargas, J., Ikea, J., Angel, F., Tohme, J., Asiedu, R. A., & Roca, W. M. (1994). Variability of chloroplast DNA and nuclear ribosomal DNA in cassava *(Manihot esculenta* Crantz) and its wild relatives. *Theoretical and Applied Genetics*, *89*(6), 719-727.

Fregene, M. A., Suarez, M., Mkumbira, J., Kulembeka, H., Ndedya, E., Kulaya, A., & Kresovich, S. (2003). Simple sequence repeat marker diversity in cassava landraces: genetic diversity and differentiation in an asexually propagated crop. *Theoretical and Applied Genetics*, *107*(6), 1083-1093.

Garcia-Lor, A., Curk, F., Snoussi-Trifa, H., Morillon, R., Ancillo, G., Luro, F., & Ollitrault, P. (2013). A nuclear phylogenetic analysis: SNPs, indels and SSRs deliver new insights into the relationships in the 'true citrus fruit trees' group (Citrinae, Rutaceae) and the origin of cultivated species. *Annals of botany*, *111*(1), 1-19.

Garrido-Garduno, T., & Vazquez-Dominguez, E. (2013). Genetic, spatial and connectivity methods of analysis in landscape genetics. *Revista mexicana de biodiversidad*, *84*(3), 1031-1054.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*(2), e90346.

Glowka, L. et al., (1996), Guía del Convenio sobre la Diversidad Biológica, UICN Gland y Cambridge. xii + 179pp.

Glover, K. A., Hansen, M. M., Lien, S., Als, T. D., Høyheim, B., & Skaala, Ø. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC genetics*, *11*(1), 1-12.

Gonçalves, T. M., Vidigal Filho, P. S., Vidigal, M. C. G., Ferreira, R. C. U., Rocha, V. P. C., Ortiz, A. H. T., & Kvitschal, M. V. I. (2017). Genetic diversity and population structure of traditional sweet cassava accessions from Southern of Minas Gerais State, Brazil, using microsatellite markers. *African Journal of Biotechnology*, *16*(8), 346-358.

Heckler, S., & Zent, S. (2008). Piaroa manioc varietals: hyperdiversity or social currency?. *Human Ecology*, *36*(5), 679-697.

Howeler, R., Lutaladio, N., & Thomas, G. (2013). *Save and grow: cassava. A guide to sustainable production intensification*. Fao.

Hurtado López, P. X., Marín, J. A., Ospina, C., Buitrago R, C., Castelblanco, W. H., Correa, A. M., & Fregene, M. A. (2008). An assessment of the diversity in global cassava (*Manihot esculenta* Crantz) genetic resources based on simple sequence repeat (SSR) markers.

Futuyma, D. J. (2005). Evolution. Sinauer & Associates. *Inc., Sunderland, Massachusetts*, 226-243.

Ilut, D. C., Sanchez, P. L., Costich, D. E., Friebe, B., Coffelt, T. A., Dyer, J. M., & Gore, M. A. (2015). Genomic diversity and phylogenetic relationships in the genus Parthenium (Asteraceae). *Industrial Crops and Products*, *76*, 920-929.

Isendahl, C. (2011). The domestication and early spread of manioc (*Manihot esculenta* Crantz): a brief synthesis. *Latin American Antiquity*, *22*(4), 452-468.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94.

Kamanda, I., Blay, E. T., Asante, I. K., Danquah, A., Ifie, B. E., Parkes, E., & Sesay, S. (2020). Genetic diversity of provitamin-A cassava (*Manihot esculenta* Crantz) in Sierra Leone. *Genetic resources and crop evolution*, *67*(5), 1193-1208.

Karim, K. Y., Ifie, B., Dzidzienyo, D., Danquah, E. Y., Blay, E. T., Whyte, J. B., & Iluebbey, P. (2020). Genetic characterization of cassava (*Manihot esculenta* Crantz) genotypes using agro-morphological and single nucleotide polymorphism markers. *Physiology and Molecular Biology of Plants*, *26*(2), 317-330.

Kawa, N. C., McCarty, C., & Clement, C. R. (2013). Manioc varietal diversity, social networks, and distribution constraints in rural Amazonia. *Current Anthropology*, *54*(6), 764-770.

Kawuki, R. S., Herselman, L., Labuschagne, M. T., Nzuki, I., Ralimanana, I., Bidiaka, M., & Ferguson, M. E. (2013). Genetic diversity of cassava (*Manihot esculenta* Crantz) landraces and cultivars from southern, eastern and central Africa. *Plant Genetic Resources*, *11*(2), 170-181.

Kleih, U., Phillips, D., Wordey, M. T., & Komlaga, G. (2013). Cassava market and value chain analysis: Ghana case study. Natural Resources Institute, University of Greenwich, UK.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage

studies. *Nature genetics*, *17*(1), 21-24.

Lebot V. (2009) Tropical root and tuber crops: Cassava, sweet potato, yams and aroids. Oxford: CAB International.

Lewin, B. 2005. Essential Genes. Pearson, eua.

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics (Oxford, England)*, *28*(14), 1838–1844. https://doi.org/10.1093/bioinformatics/bts280.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Lin, M. F., Bai, X., Salerno, W. J., & Reid, J. G. (2020). Sparse Project VCF: efficient encoding of population genotype matrices. *Bioinformatics*, *36*(22-23), 5537-5538.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639-647.

Liu, Z., Li, J., Fan, X., Htwe, N. M. P. S., Wang, S., Huang, W., & Qiu, L. (2017). Assessing the numbers of SNPs needed to establish molecular IDs and characterize the genetic diversity of soybean cultivars derived from Tokachi nagaha. *The Crop Journal*, *5*(4), 326-336.

Luo, Z., Brock, J., Dyer, J. M., Kutchan, T., Schachtman, D., Augustin, M., & Abdel-Haleem, H. (2019). Genetic diversity and population structure of a Camelina sativa spring panel. *Frontiers in Plant Science*, *10*, 184.

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics:

combining landscape ecology and population genetics. *Trends in ecology & evolution*, *18*(4), 189-197.

Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, *20*(3), 136-142.

Marmey, P., Beeching, J. R., Hamon, S., & Charrier, A. (1993). Evaluation of cassava (*Manihot esculenta* Crantz) germplasm collections using RAPD markers. *Euphytica*, *74*(3), 203-209.

Martínez, A., & Vega-Pla, J. L. (2002). Estudios de variación genética utilizando microsatélites de ADN. *V Curso Internacional de Especialización en la Conservación y Utilización de las Razas de Animales Domésticos Locales en Sistemas de Explotación Tradicionales. Córdoba. España. 10pp*.

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, *7*(5), 1393-1404.

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, *44*(3), 243-246.

Mba, R. E. C., Stephenson, P., Edwards, K., Melzer, S., Nkumbira, J., Gullberg, U., & Fregene, M. (2001). Simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta* Crantz) genome: towards an SSR-based molecular genetic map of cassava. *Theoretical and Applied Genetics*, *102*(1), 21-31.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, *17*(2), 240-248.

Mejía, K. (2002). El género Manihot (yuca) en el Perú y sus parientes silvestres. Seminario–Taller “Parientes Silvestres de los Cultivos Nativos en el Perú” (2002, Universidad Nacional Agraria La Molina, Lima).

Mejia, J., Da Silva, R. E., Mejía-Carhuanca, K. M., & García-Dávila, C. R. (2015). Caracterización morfológica y genética de etnovariedades de la yuca *Manihot esculenta, Cratnz*, en seis localidades de la cuenca baja del río Ucayali–Perú. *Folia Amazónica*, *24*(1), 71-82.

Meza, Y. & Julca Otiniano, A. (2015). Sustentabilidad de los sistemas de cultivo con yuca (Manihot esculenta Crantz) en la subcuenca de Santa Teresa, Cusco. Ecología Aplicada, 14(1), 55-63.

Mezette, T. F., Blumer, C. G., & Veasey, E. A. (2013). Morphological and molecular diversity among cassava genotypes. *Pesquisa Agropecuária Brasileira*, *48*, 510-518.

Montero-Rojas, M., Correa, A. M., & Siritunga, D. (2011). Molecular differentiation and diversity of cassava (*Manihot esculenta*) taken from 162 locations across Puerto Rico and assessed with microsatellite markers. *AoB Plants*, *2011*.

Morillo, E. & Miño, G. (2011). Marcadores Moleculares en Biotecnología Agrícola: Manual de procedimientos y técnicas en INIAP. Manual No. 91 INIAP Estación Experimental Santa.

Mostajo, O. (2017). Plan Nacional de Cultivos: Campaña Agrícola 2018-2019.

Mtunguja, M. K., Ranjan, A., Laswai, H. S., Muzanila, Y., Ndunguru, J., & Sinha, N. R. (2017). Genetic diversity of farmer-preferred cassava landraces in Tanzania based on morphological descriptors and single nucleotide polymorphisms. *Plant Genetic Resources*, *15*(2), 138-146.

Murray, B. W. (1996). The estimation of genetic distance and population substructure from microsatellite allele frequency data. *McMaster University, Ontario, Canada*.

Nei M. 1972. Genetic distance between populations. Am Nat 106:283-292.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, *70*(12), 3321-3323.

Núñez Pérez, C., Martín Brañas, M., Del Aguila Villacorta, M., & Zárate Gómez, R. (2018). Tüxe: conocimientos tradicionales vinculados a la yuca (*Manihot esculenta*) en el pueblo Ticuna.

OECD (2016), "Cassava (Manihot esculenta)", in Safety Assessment of Transgenic Organisms in the Environment, Volume 6: OECD Consensus Documents, OECD Publishing, Paris. DOI: https://doi.org/10.1787/9789264253421-6-en.

Ogbonna, A. C., de Andrade, L. R. B., Mueller, L. A., de Oliveira, E. J., & Bauchet, G. J. (2021). Comprehensive genotyping of a Brazilian cassava (*Manihot esculenta* Crantz) germplasm bank: insights into diversification and domestication. *Theoretical and Applied Genetics*, *134*(5), 1343-1362.

Olsen, K. M., & Schaal, B. A. (1999). Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proceedings of the National Academy of Sciences*, *96*(10), 5586-5591.

Olsen, K. M., & Schaal, B. A. (2001). Microsatellite variation in cassava (Manihot esculenta, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. American journal of botany, 88(1), 131-142.

Olsen, K. M. (2004). SNPs, SSRs and inferences on cassava's origin. *Plant Molecular Biology*, *56*(4), 517-526.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. 2: e190. *Population structure and eigenanalysis. PLoS Genet. 2: e190.*

Parmar, A., Sturm, B., & Hensel, O. (2017). Crops that feed the world: production and improvement of cassava for food, feed, and industrial uses. *Food Security*, *9*(5), 907-927.

Peña-Venegas, C. P., Stomph, T. J., Verschoor, G., Lopez-Lavalle, L. A. B., & Struik, P. C. (2014). Differences in manioc diversity among five ethnic groups of the Colombian

Amazon. *Diversity*, *6*(4), 792-826.

Peroni, N., & Hanazaki, N. (2002). Current and lost diversity of cultivated varieties, especially cassava, under swidden cultivation systems in the Brazilian Atlantic Forest. *Agriculture, Ecosystems & Environment*, *92*(2-3), 171-183.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135.

Prempeh, R. N. A., Manu-Aduening, J. A., Quain, M. D., Asante, I. K., Offei, S. K., & Danquah, E. Y. (2020). Assessment of genetic diversity among cassava landraces using single nucleotide polymorphic markers. *African Journal of Biotechnology*, *19*(6), 383-391.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., & Rounsley, S. (2012). The cassava genome: current progress, future directions. *Tropical plant biology*, *5*(1), 88-94.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOSONE*, *7*(2), e32253.

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5, 92–102.

Pootakham, W., Shearman, J. R., Ruang-Areerate, P., Sonthirod, C., Sangsrakru, D., Jomchai, N., & Tangphatsornruang, S. (2014). Large-scale SNP discovery through RNA sequencing and SNP genotyping by targeted enrichment sequencing in cassava (*Manihot esculenta* Crantz). *PLoS ONE*, *9*(12), e116028.

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

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

Rabbi, I., Hamblin, M., Gedil, M., Kulakow, P., Ferguson, M., Ikpan, A. S., & Jannink, J. L. (2014). Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. *Crop Science*, *54*(4), 1384-1396.

Rabbi, I. Y., Kulakow, P. A., Manu-Aduening, J. A., Dankyi, A. A., Asibuo, J. Y., Parkes, E. Y., & Maredia, M. K. (2015). Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). *BMC genetics*, *16*(1), 1-11.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current opinion in plant biology*, *5*(2), 94-100.

Raji, A. A., Anderson, J. V., Kolade, O. A., Ugwu, C. D., Dixon, A. G., & Ingelbrecht, I. L. (2009). Gene-based microsatellites for cassava (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. *BMC Plant Biology*, *9*(1), 1-11.

Rees, D., Westby, A., Tomlins, K., Van Oirschot, Q., Cheema, M. U., Cornelius, E., & Amjad, M. (2012). Tropical root crops. *Crop Post-Harvest: Science and Technology, Perishables*, *3*, 392.

Rouse, I., & Cruxent, J. M. (1963). *Arqueología venezolana*. Instituto Venezolano de Investigaciones Científicas.

Reynolds, J., Weir, B. S., & Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, *105*(3), 767-779.

Rival, L., & McKey, D. (2008). Domestication and diversity in manioc (*Manihot esculenta* Crantz ssp. esculenta, Euphorbiaceae). *Current anthropology*, *49*(6), 1119-1128.

Roa, A. C., Maya, M. M., Duque, M. C., Tohme, J., Allem, A. C., & Bonierbale, M. W. (1997). AFLP analysis of relationships among cassava and other *Manihot* species. *Theoretical and Applied Genetics*, *95*(5), 741-750.

Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols*, *12*(12), 2640-2659.

Roger, J. S. (1972). Measure of genetic similarity and genetic distance. Studies in genetics VII. *University of Texas publication*, *7213*, 145-153.

Rogers, D.J. and Appan, S.G. (1973). *Manihot* and Maniholoides (Euphorbiaceae): A Computer Assisted Study. Hafner, New York, NY.

Salick, J., Cellinese, N., & Knapp, S. (1997). Indigenous diversity of cassava: generation, maintenance, use and loss among the Amuesha, Peruvian upper Amazon. *Economic Botany*, *51*(1), 6-19.

Salmaso, M., Faes, G., Segala, C., Stefanini, M., Salakhutdinov, I., Zyprian, E., & Velasco, R. (2005). Genome diversity and gene haplotypes in the grapevine (Vitis vinifera L.), as revealed by single nucleotide polymorphisms. *Molecular Breeding*, *14*(4), 385-395.

Schilling, M. P., Wolf, P. G., Duffy, A. M., Rai, H. S., Rowe, C. A., Richardson, B. A., & Mock, K. E. (2014). Genotyping-by-sequencing for Populus population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS ONE*, *9*(4), e95292.

Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion?. *Nature reviews genetics*, *5*(1), 63-69.

Shigaki, T. (2016). *Cassava: The Nature and Uses. Encyclopedia of Food and Health, 687–693.* doi:10.1016/b978-0-12-384947-2.00124-0.

Seminario, J., aSeminario, J., & Valderrama, M. (2004). Raíces andinas: contribuciones al conocimiento ya la capacitaciónConservación de germoplasma de arracacha con campesinos (No. F30 S453). Universidad Nacional de Cajamarca (Peru); CIP, Lima (Peru); Agencia Suiza para el Desarrollo y la Cooperación, Lima (Peru).

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., & Belzile, F. (2013).

An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS ONE*, *8*(1), e54603.

Sosa, P. A., Batista Hernández, F., Bouza Carrelo, N., & González Pérez, M. Á. (2002). La conservación genética de las especies vegetales amenazadas.

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., & Belzile, F. (2017). Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC bioinformatics*, *18*(1), 1-7.

Tovar, E., Bocanegra, J. L., Villafañe, C., Fory, L., Velásquez, A., Gallego, G., & Moreno, R. (2016). Diversity and genetic structure of cassava landraces and their wild relatives (*Manihot* spp.) in Colombia revealed by simple sequence repeats. *Plant Genetic Resources*, *14*(3), 200-210.

Turyagyenda, L. F., Kizito, E. B., Ferguson, M. E., Baguma, Y., Harvey, J. W., Gibson, P., & Osiru, D. S. O. (2012). Genetic diversity among farmer-preferred cassava landraces in Uganda. *African Crop Science Journal*, *20*(1).

Uchechukwu-Agua, A. D., Caleb, O. J., & Opara, U. L. (2015). Postharvest handling and storage of fresh cassava root and products: a review. *Food and Bioprocess Technology*, *8*(4), 729-748.

Uribe Palacios, C. D. (2014). Diversidad genética de Mytilus chilensis utilizando marcadores microsatélites.

Varvio, S. L., Chakraborty, R., & Nei, M. (1986). Genetic variation in subdivided populations and conservation genetics. *Heredity*, *57*(2), 189-198.

Villasante, A., & Natali, Y. (2006). Análisis de diversidad genética y distribución espacial del germoplasma de *Manihot esculenta* Crantz (yuca) en Ucayali-Perú, mediante marcadores SSR.

Velázquez, L. P. A., MdCA, M., & Romero, A. C. (2014). Extracción y purificación de

ADN. *Herramientas moleculares aplicadas en ecología: aspectos teóricos y prácticos*, *1*.

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC bioinformatics*, *18*(1), 1-12.

Willemen, L., Scheldeman, X., Cabellos, V. S., Salazar, S. R., & Guarino, L. (2007). Spatial patterns of diversity and genetic erosion of traditional cassava (*Manihot esculenta* Crantz) in the Peruvian Amazon: An evaluation of socio-economic and environmental indicators. *Genetic Resources and Crop Evolution*, *54*(7), 1599-1612.

Weir, B. S. (1996). Methods for discrete population genetic data. *Genetic Data Analysis.*

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., & Kilian, A. (2004). Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proceedings of the National Academy of Sciences*, *101*(26), 9915-9920.

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., ... & Jannink, J. L. (2016). Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. The Plant Genome, 9(2), plantgenome2015-11.

Wright, S., 1921, Systems of Mating. Genetics 6:111.

Wright, S., 1931, Evolution in Mendelian populations. Genetics 16:97

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, 395-420.

Xia, L., Peng, K., Yang, S., Wenzl, P., De Vicente, M. C., Fregene, M., & Kilian, A. (2005). DArT for high-throughput genotyping of cassava (*Manihot esculenta*) and its wild relatives. *Theoretical and Applied Genetics*, *110*(6), 1092-1098.

Yang, G. Q., Chen, Y. M., Wang, J. P., Guo, C., Zhao, L., Wang, X. Y., & Guo, Z. H.

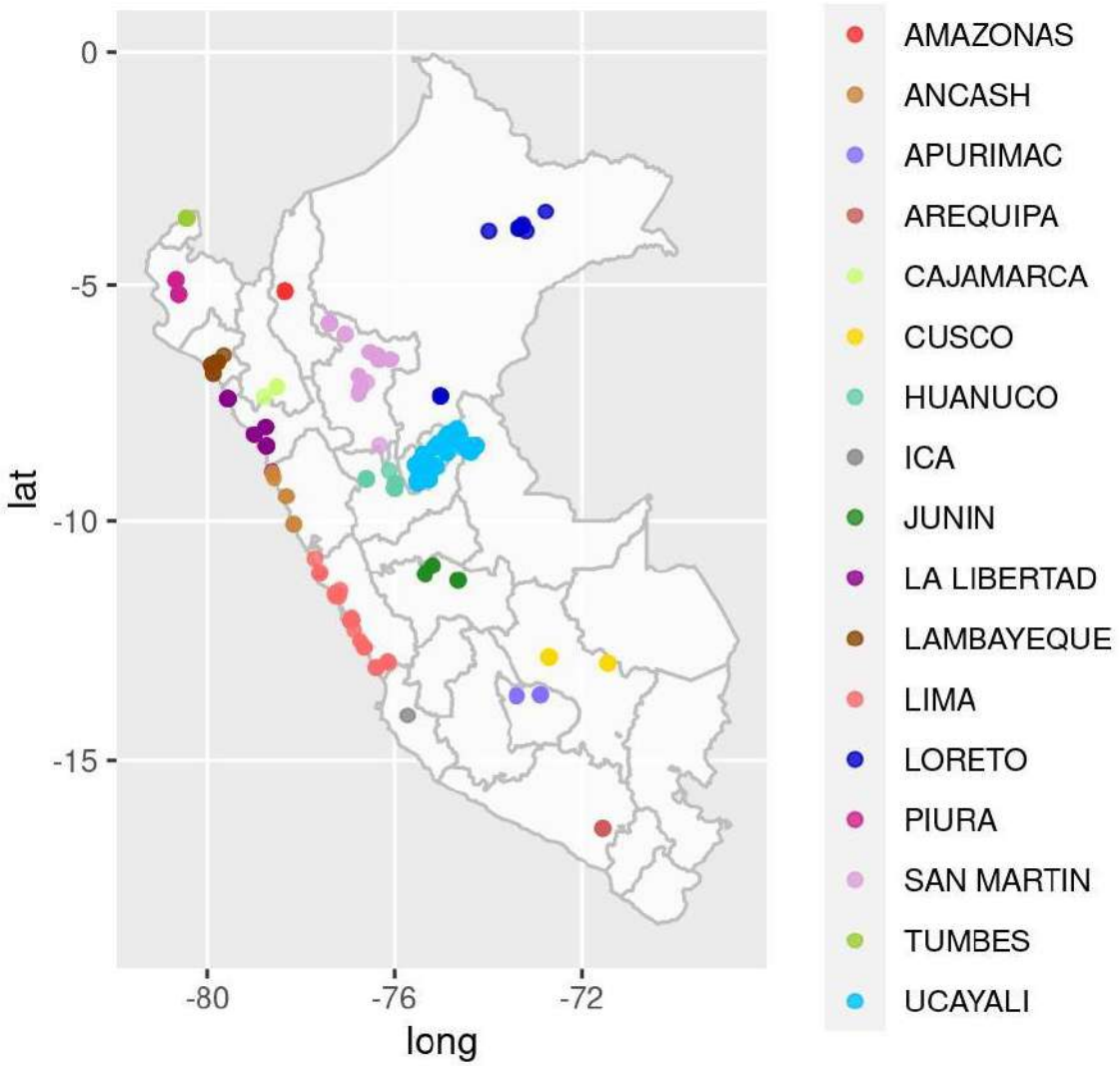
(2016). Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. *Plant methods*, *12*(1), 1-17.

Zamorano, M. J. 1998. Utilización de medidas genéticas coadyuvantes a los planes de mejora del vacuno de carne. Tesis doctoral. Universidad de Córdoba. España. p. 191.

Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., & Wang, W. (2018). Genome-wide association studies of 11 agronomic traits in cassava (*Manihot esculenta* Crantz). *Frontiers in plant science*, *9*, 503.

11. ANEXOS

Anexo 1: Mapa de distribución de las 731 accesiones de la Colección Nacional del INIA.



Anexo 2: Protocolos de preparación de insumos de biología molecular, usados para el

proceso de extracción de ADN genómico.

CTAB buffer de extracción 10% - CTAB 10X

Bromuro de cetiltrimetilamonio	**5.0 g**	**10 g**	**10%**
Cloruro de sodio	2.05 g	4.09 g	0.7 M
EDTA 0.5 M pH 8.0	2.0 ml	4.0 ml	20 mM
Tris HCl 1M pH 8.0	5 ml	10.0 ml	100 mM
H2O (d) c.s.p.	50 ml	100 ml	

CTAB buffer de extracción 2% - CTAB 2X

Bromuro de cetiltrimetilamonio	**2.0 g**	**20 g**	**2%**
Cloruro de sodio	8.18 g	81.81 g	1.4 M
EDTA 0.5 M pH 8.0	4.0 ml	40.0 ml	20 mM
Tris HCl 1M pH 8.0	10.0 ml	100.0 ml	100 mM
Polivinilpirrolidona	1.0 g	10 g	1%
H2O (d) c.s.p.	100 ml	1 L	

EDTA 0.5 M pH 8.0

Ácido etilendiaminotetracetico	**186.1 g**
Hidroxido de sodio	más de 100 pastillas
H2O c.s.p	1L

Tris HCl 1M pH 8.0

Tris base	**H2O (d)**	**Enrazar hasta**
121.1g	800	1L
96.8g	650	800 ml
60.5g	400	500 ml

T10 E1

Tris HCl 1 M pH 8.0	**4.0ml**	**10ml**	**10mM**
EDTA 0.5M pH 8.0	0.8ml	2.0ml	1mM
H2O c.s.p.	400ml	1L	

Anexo 3: Guía para envío de muestras de ADN genómico para GBS del Centro Genómico de la Universidad de Minennsota.

1. Billing information: In order for a project to begin, we need your complete billing information filled out under the Submission tab. You will receive an invoice (bill) to the email you provide when your project is complete. Please contact us immediately if you have a pending grant deadline.

2. Pilot samples: It is recommended to submit 8 samples for a pilot that represent the diversity of your population study or parents and F1/F2s if this is a mapping study. Markers produced in the pilot may not be replicated in your full sample set if the eight pilot samples submitted are not representative of the samples in the full dataset. Samples should be plated in columns using the format under the submission tab. If you are submitting multiple pilot projects, please submit in separate plates with separate submission forms.

3. DNA requirements: Submit a minimum of **20 ng/ul in a minimum of 25 ul** as determined by picogreen or Qubit; Nanodrop does not provide accurate values on double-stranded DNA. If you can only get a Nanodrop measurement, it is recommended sending extra volume to compensate for the difference in quantification platforms and to not dilute your DNA before submission. However, do not fill the wells beyond 120 ul as filling your wells beyond 120 ul increases your risk for well-to-well contamination.

4. Extra samples: If you are shipping from outside the U. S., it is recommended to send extra samples to substitute for samples that may not pass QC.

5. Type of plate: Please use a fully-skirted plate as these plates provide the most protection during shipping; plates can be purchased from the UMGC if needed. Submitting in tubes is also fine for pilot samples, but be sure the tubes are clearly labelled.

6. Type of plate seal: If you are shipping on ice packs, your plate should be sealed with strip caps to prevent cross-contamination of samples and evaporation loss. Be sure the caps and the wells form a tight seal. If you are shipping on dry ice and can ensure samples stay frozen during shipment, you can seal your plate with foils seals. It is not recommended to use the clear polyester seals as these seals can dislodge from the plates during dry ice shipment.

7. Freeze your samples: Samples should be frozen solid before packing for shipping.

8. Packing plates: Place your plate in a plastic baggie and then in a small box. OR cut two pieces of cardboard slightly larger than your plate and use vinyl tape (lab tape will not work with dry ice) to seal your plate between the cardboard and then place in a plastic baggie. The cardboard or box will protect your plate during shipment as plates in direct contact with dry ice can crack. The plastic baggie will protect from plate-to-plate contamination if one plate should be compromised during shipment.

9. Print: Print the Submission tab, place in a plastic bag, and pack with your samples.

10. Ice: Do not ship plates on ice.

11. Ice packs: Ice packs should be completely frozen before packing. Ice packs should be placed above and below your plate. Voided spaces should be filled with paper towels or packing material to prevent plate shifting during shipment.

12. **Dry ice (recommended) :** Dry ice pellets are preferred as large dry ice blocks can damage plates during shipping. Ship with enough dry ice to ensure your samples stay frozen throughout shipping, typically 10 lbs of dry ice is needed for domestic overnight shipping of 1-3 plates. Dry ice should be placed below, above, and on the sides of your cardboard protected plate. Voided spaces should be filled with paper towels or packing material.

Anexo 4: Figuras suplementarias de los resultados obtenidos.

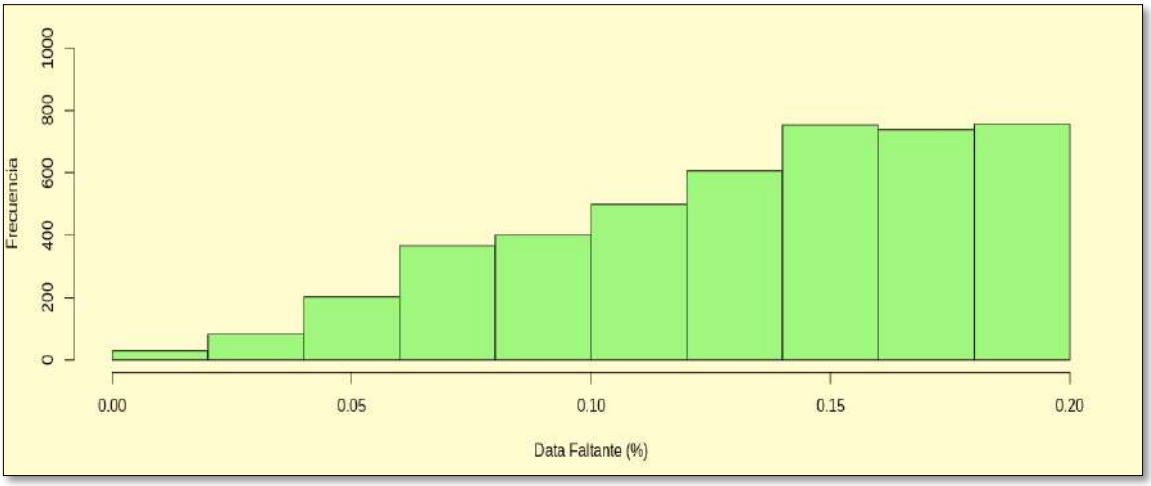


Figura 1. Histograma de la frecuencia de data faltante por *locus*.

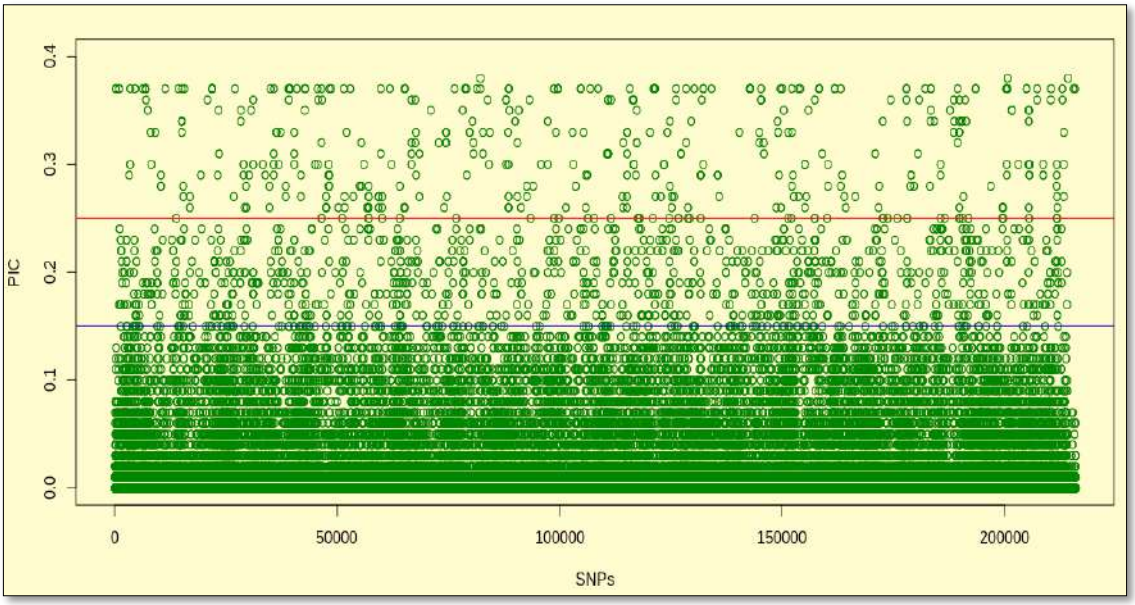


Figura 2. Plot de valores de PIC para cada alelo de los 108,397 SNPs resultantes del primer y segundo filtrado. La línea roja indica un valor de 0.25 y la azul un valor de 0.15. El paquete *snpReady* usado para el análisis, considera a cada alelo del SNP-I*oci* por ello se observa el doble de locus (216,794) que corresponde a los alelos totales.

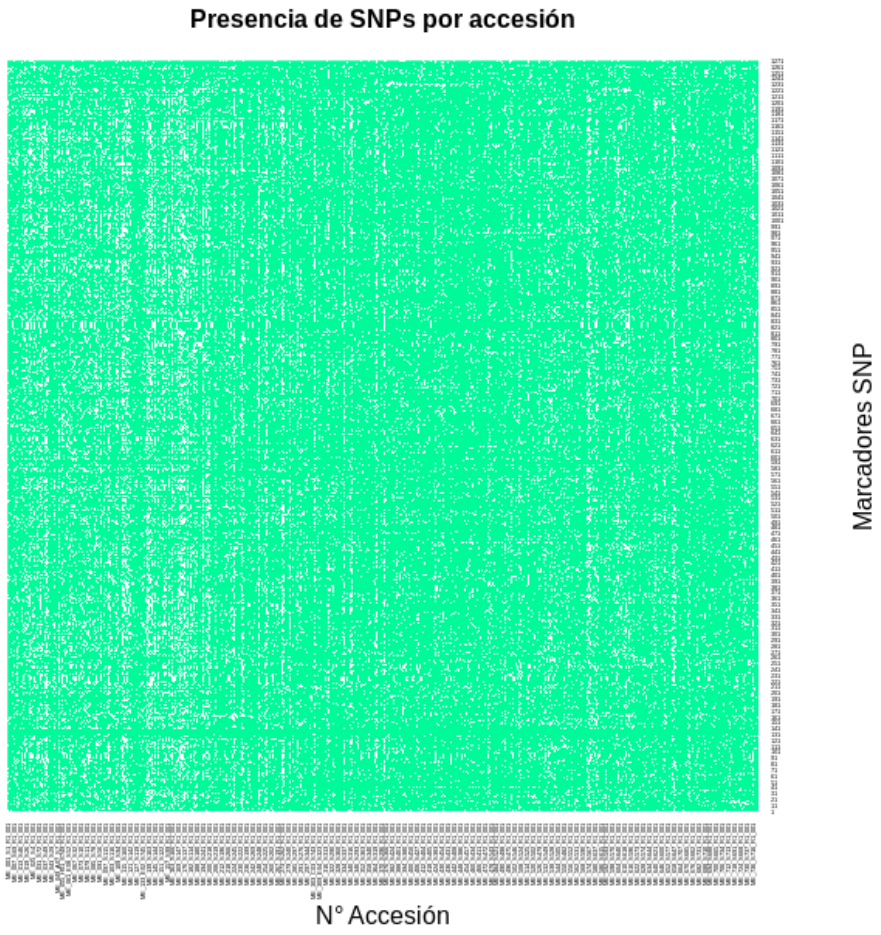


Figura 3. Heatmap de presencia y ausencia de SNP por accesión, la usencia se indica por un espacio de en blanco

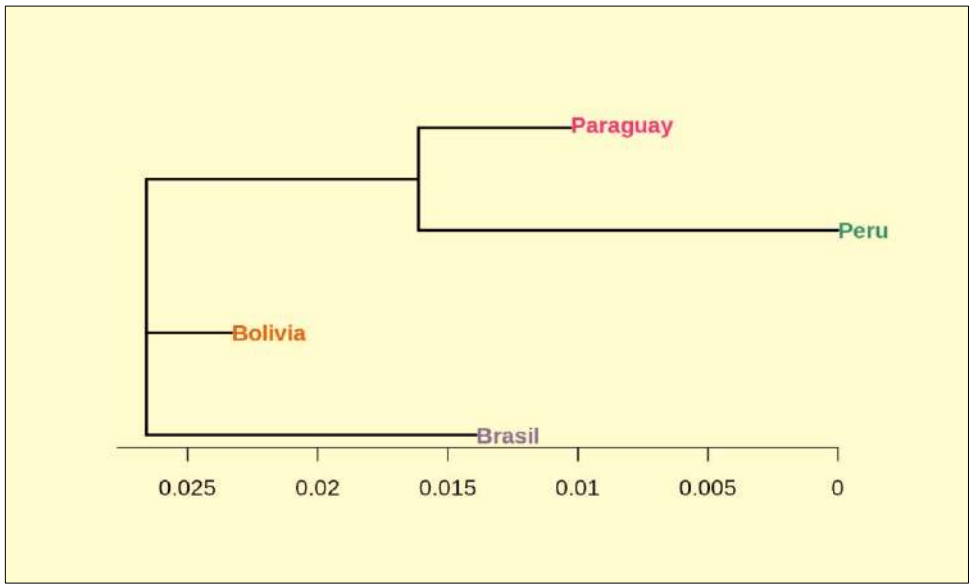


Figura 4. Relación genética entre países, basada en la distancia genética de Nei y un agrupamiento en base a *Neighbour-joining*.

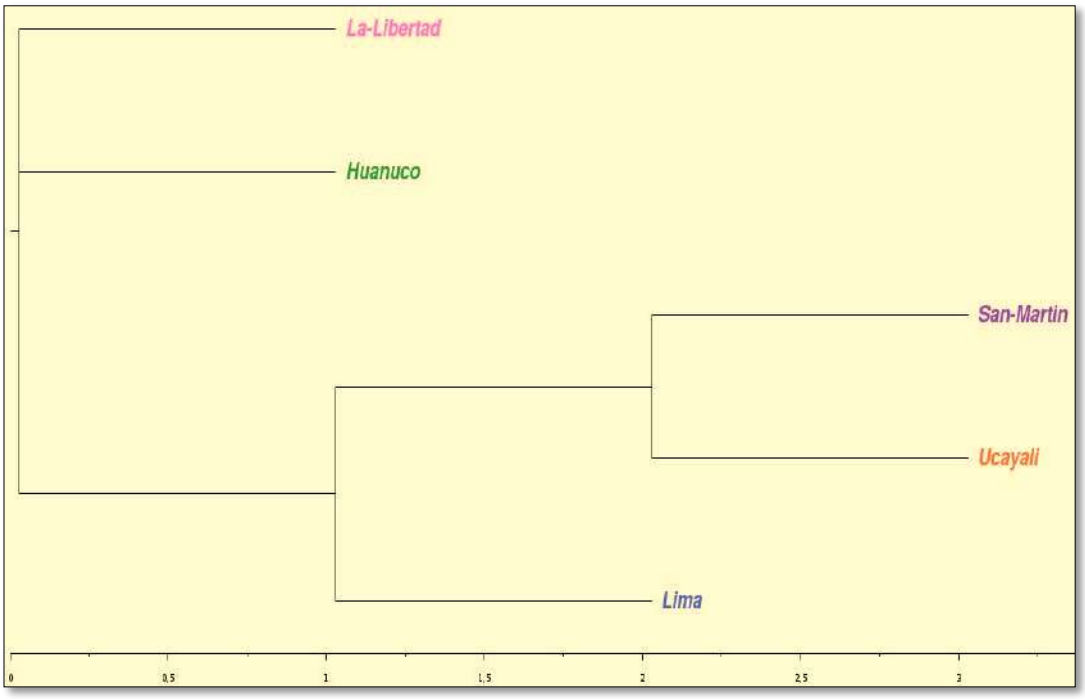


Figura 5. Relaciones genéticas entre departamentos, basados en la distancia genética de Nei y un agrupamiento en base a *Neighbour-joining*.

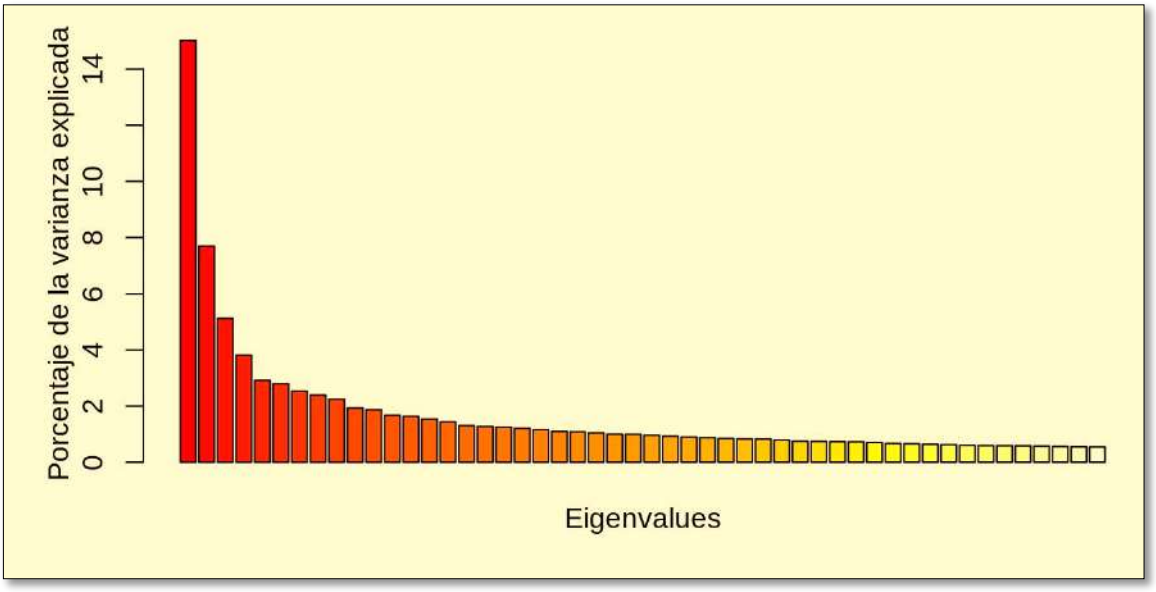


Figura 6. Barplot de los *eigenvalues* o valores propios y su porcentaje de varianza acumulada, para el análisis de componentes principales (PCA) de las 728 accesiones agrupadas por país de procedencia.

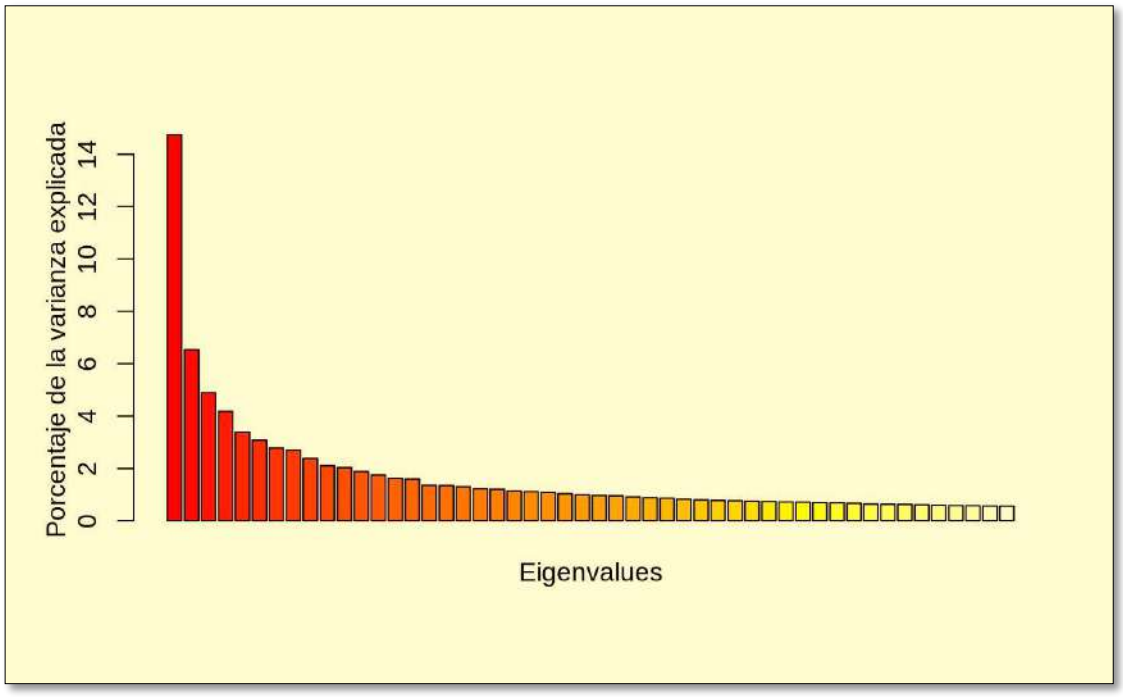


Figura 7. Barplot de los *eigenvalues* o valores propios y su porcentaje de varianza acumulada, del análisis de componentes principales (PCA) para las 634 accesiones peruanas.

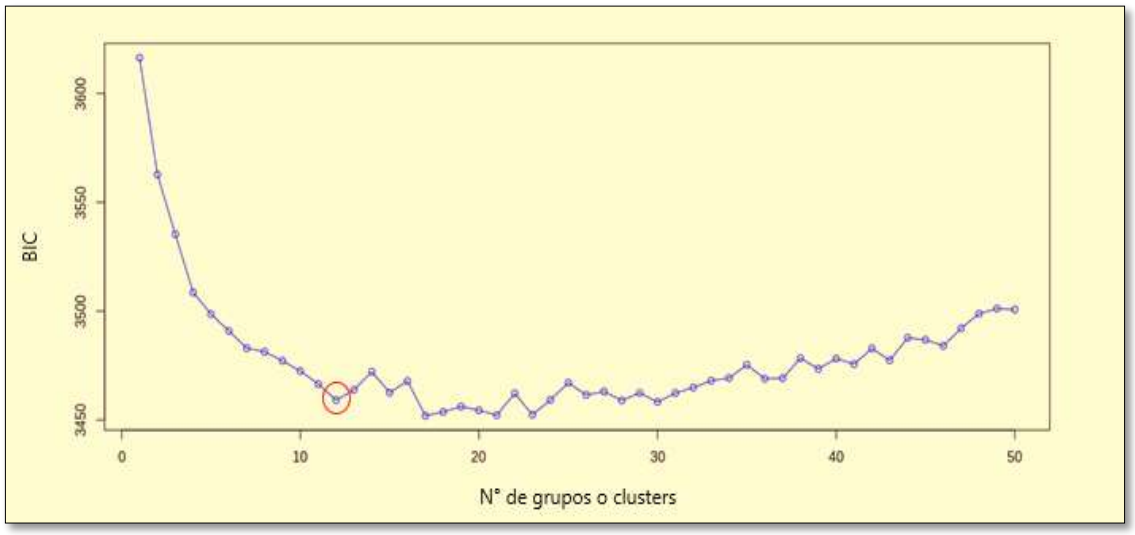


Figura 8. Gráfica de los valores de BIC versus el número de *clusters*, para determinar el mejor valor de "K" para las 728 accesiones de la Colección Nacional de Yuca.

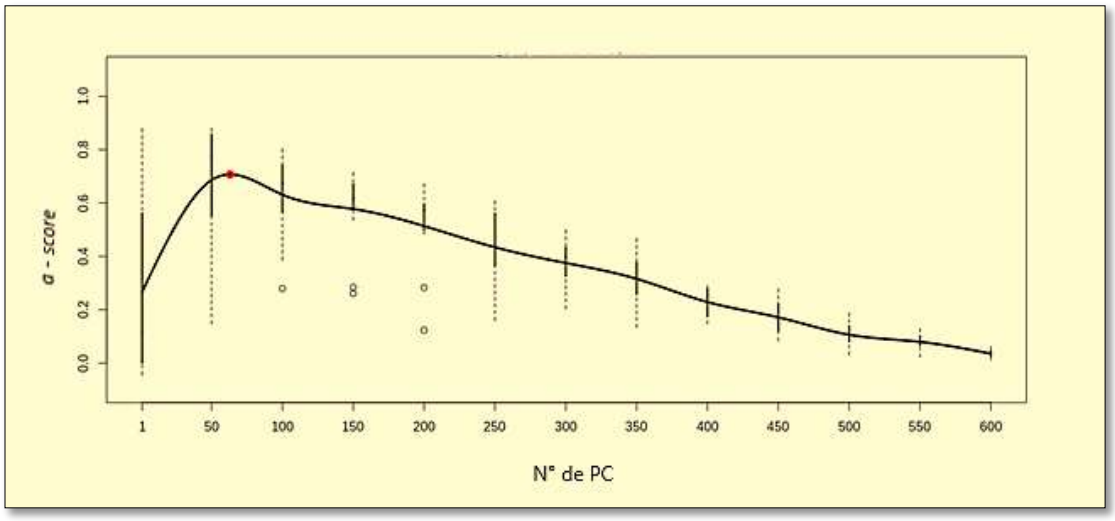


Figura 9. Valor predicho de CP versus su valor de *a-score*, se definió 600 CP como máximo para dicho análisis. El valor predicho es 63 y se indica con un círculo rojo en la curva.

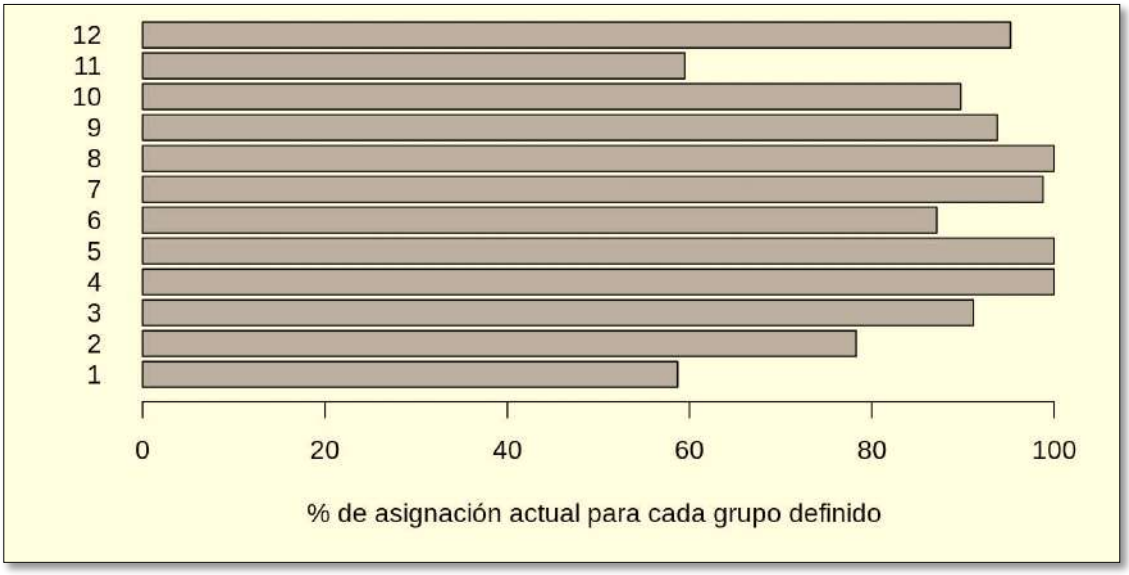


Figura 10. Barplot del porcentaje de reasignación para los 12 grupos definidos, usando PC = 63, valor predicho por la función *optim.a. score* de *adegenet.*

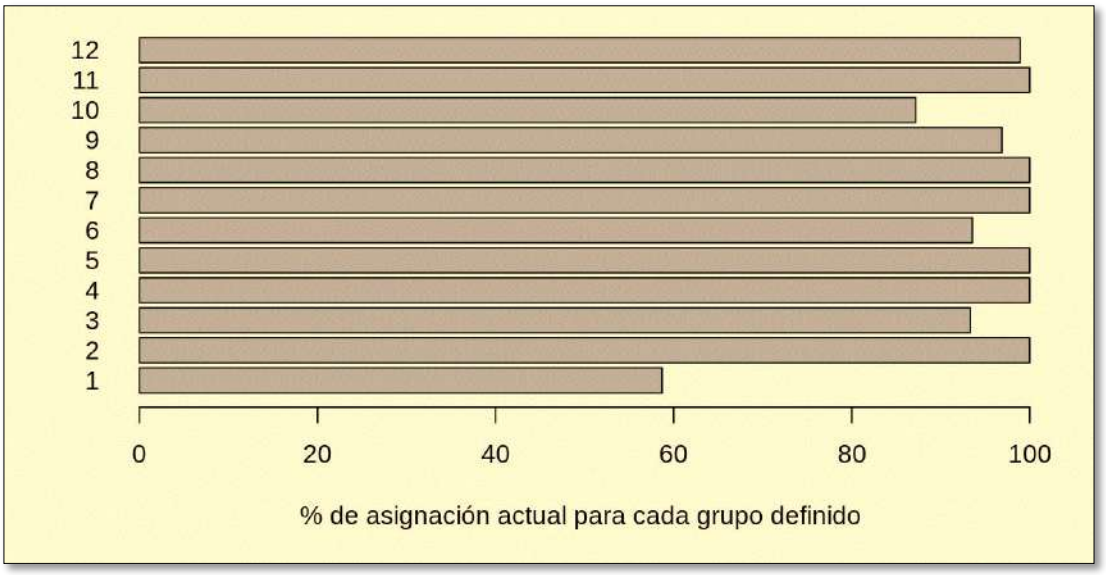


Figura 11. Barplot del porcentaje de reasignación para los 12 grupos definidos, usando PC = 200.

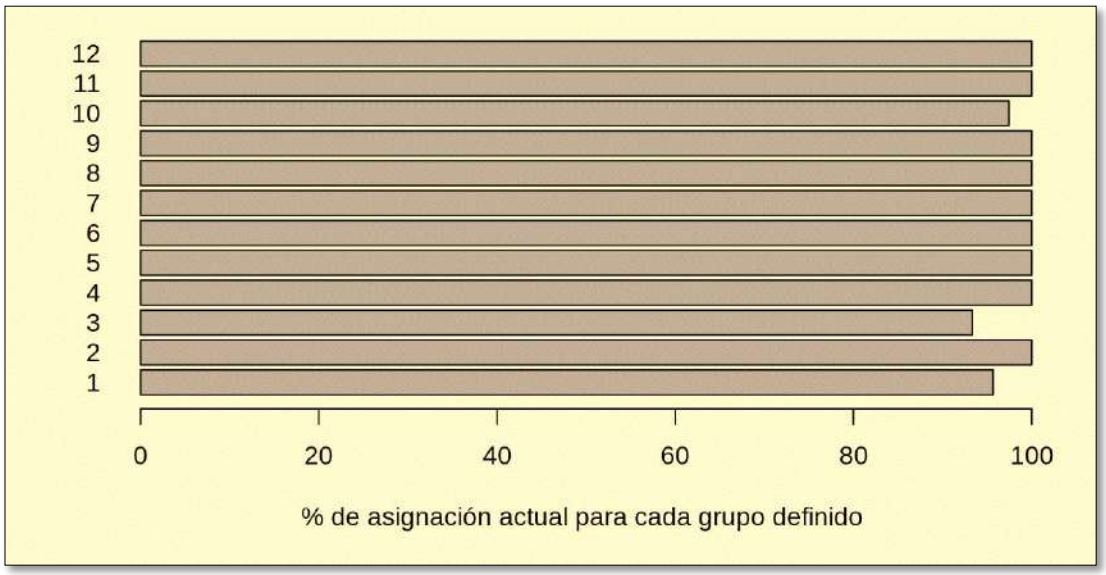


Figura 12. Barplot del porcentaje de reasignación para los 12 grupos definidos, usando PC = 350.

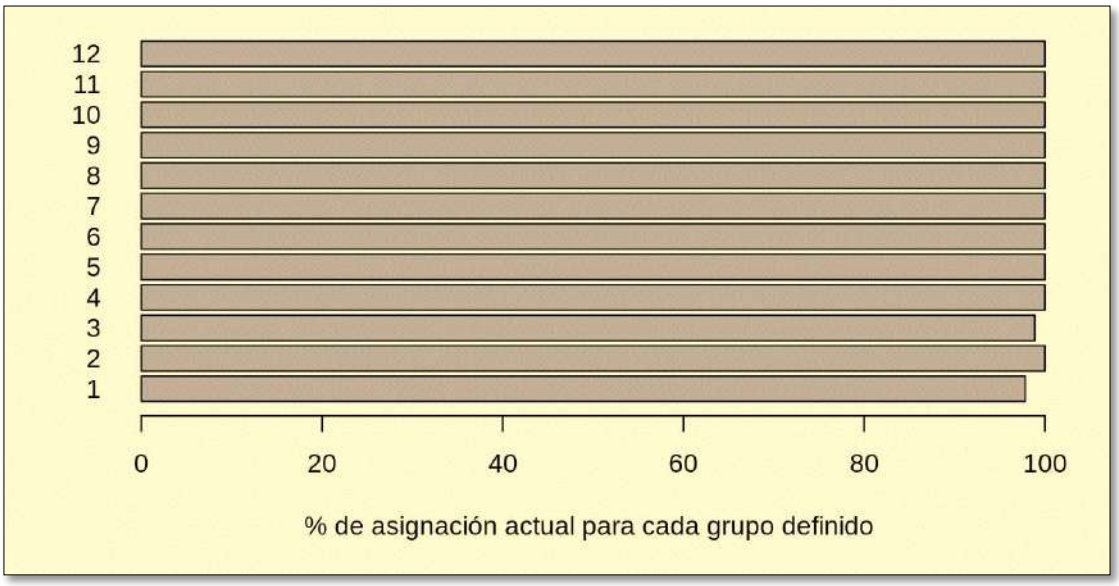


Figura 13. Barplot del porcentaje de reasignación para los 12 grupos definidos, usando PC = 500.