Caracterización de la diversidad genética de variedades de albaricoquero mediante marcadores microsatélites (SSR)

Sara Herrera Lagranja
Máster en Bioinformática y Bioestadística
Análisis de Datos Ómicos

Consultor: Guillem Ylla Bou
Tutor externo: Javier Rodrigo García
Profesor responsable de la asignatura: David Merino Arranz

2 de enero de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Caracterización de la diversidad genética de variedades de albaricoquero mediante marcadores microsatélites (SSR)*
Nombre del autor:	*Sara Herrera Lagranja*
Nombre del consultor/a:	*Guillem Ylla Bou*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	01/2019
Titulación::	*Máster Bioinformática y Bioestadística*
Área del Trabajo Final:	*Análisis de Datos Ómicos*
Idioma del trabajo:	*Español*
Palabras clave	*Albaricoquero, microsatélites, variabilidad genética*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

En los últimos años, se han desarrollado diversos programas de mejora en todo el mundo para dar solución a diferentes problemas que presenta el cultivo del albaricoquero lo que ha provocado una intensa renovación varietal, con la introducción de nuevas variedades que están desplazando a algunas variedades tradicionales. Sin embargo, las variedades locales son una importante fuente natural de caracteres de interés, por lo que es importante su conservación y el estudio de los recursos genéticos locales.

En este trabajo se han utilizado 7 marcadores microsatélites (SSRs), previamente descritos, para estudiar la diversidad genética de 50 variedades de albaricoquero tanto tradicionales como nuevas variedades procedentes de diferentes programas de mejora de diferentes países.

Los análisis de diversidad realizados mediante el software estadístico R han permitido estudiar la variabilidad genética de los diferentes genotipos. Se han identificado 3 genotipos que corresponden a la misma variedad (sinonimias). Además, se han establecido las relaciones filogenéticas existentes entre los genotipos estudiados mediante la creación de un dendrograma. Se ha identificado la presencia de 6 grupos cladísticos que muestran una tendencia de agrupación relacionada con el programa de mejora del que proviene cada variedad y permiten esclarecer las variedades utilizadas como genotipos parentales en ellos.

Los SSRs empleados han permitido estudiar la estructura genética del grupo de genotipos estudiados mediante diferentes análisis. No se ha encontrado una clara diferenciación poblacional, lo que puede ser debido al uso común de

genotipos parentales en los diferentes programas de mejora ya que comparten objetivos.

Abstract (in English, 250 words or less):

In the last years, the release of an increasing number of new cultivars from different breeding programs is resulting in an important renewal of plant material worldwide. However, local varieties constitute a source of genetic traits of interest. Thus, it is important to conserve and study this genetic pool in order to preserve morphological and phenological characters.

In this work, a set of 7 previously-described microsatellite markers (SSRs) has been selected in order to study the genetic diversity of 50 apricot cultivars, including both traditional and new cultivars from different breeding programs in different countries.

The analysis of the diversity was carried out using software R. The study of the genetic variability of the different genotypes has resulted in the identification of three synonymies. In addition, the genetic relationships among the accessions studied have been represented in a dendrogram. The presence of 6 cladistic groups shows a tendency of the breeding program. This information could be useful to clarify which genotypes have been used as parentals in them.

We identified genetic structure in the group of the genotypes by different population genetic analyses. There is no clear population differentiation, which is consistent with the use of common parental genotypes in the different breeding programs since they share some objectives.

Índice

1. Introducción 1
1.1 Contexto y justificación del trabajo 1
1.2 Objetivos del trabajo 4
1.3 Enfoque y método seguido 4
1.4 Planificación del trabajo 5
1.4.1. Tareas 5
1.4.2. Calendario 5
1.4.3. Hitos 7
1.4.4. Análisis de riesgos 7
1.5 Breve sumario de productos obtenidos 7
1.6 Breve descripción de los otros capítulos de la memoria 8
2. Materiales y métodos 9
2.1. Material vegetal 9
2.2. Extracción del DNA y PCRs 10
2.3. Análisis de la variabilidad genética 11
3. Resultados y discusión 15
3.1. Estudio de la validez de los marcadores microsatélites utilizados 15
3.1.1. Polimorfismo y riqueza alélica 15
3.1.2. Frecuencias alélicas 17
3.2. Estudio de la diversidad genética de las variedades 18
3.3. Identificación de homonímias y/o sinonimias 20
3.4. Relaciones filogenéticas entre las variedades 20
3.5. Estructura poblacional 22
3.5.1. AMOVA 22
3.5.2. Análisis de Componentes Principales 23
3.5.3. UPGMA 24
3.5.4. Análisis Discriminante de Componentes Principales 25
4. Conclusiones 27
5. Glosario 27
6. Bibliografía 29
7. Anexo 32

Lista de figuras

Figura 1. Gráfico de la producción de albaricoquero en España.……....1

Figura 2. Variabilidad de frutos de albaricoquero procedentes de diferentes programas de mejora.…….……..….2

Figura 3. Diagrama de Gantt con la planificación del TFM…….…….....…….6

Figura 4. Captura de pantalla de una lectura de fragmentos marcados mediante el programa CEQTM 8000 Genetic Analysis System…….…...11

Figura 5. Frecuencias alélicas de las variedades de albaricoquero para el locus ssrPaCITA7…….……..…17

Figura 6. Frecuencias alélicas de las variedades de albaricoquero para 6 de los locus estudiados (ssrPaCITA10, ssrPaCITA23, ssrPaCITA27, UDAp_415, UDAp_420, pchgms3).……..…….……18

Figura 7. Dendrograma de las 50 variedades de albaricoquero incluidas en este estudio generado por el análisis de conglomerados UPGMA a partir de la matriz de similitud basada en el coeficiente de similitud de Nei……..22

Figura 8. Análisis de componentes principales basado en datos de frecuencia de alelos microsatélites de 50 variedades. Las variedades están coloreadas por el programa de mejora del que provienen……24

Figura 9. Dendrograma de las poblaciones agrupadas por programas de mejora generado por el análisis de conglomerados UPGMA basado en la distancia genética de Nei..……..…….…….25

Figura 10. Gráfico de los valores BIC que muestra el número de clusters óptimo.…….……..….25

Figura 11. Diagrama de barras que representa la probabilidad de cada variedad de pertenecer a cada población……..……..….26

Lista de tablas

Tabla 1. Planificación del TFM…….6

Tabla 2. Nombre, origen y programa de mejora de procedencia de las 50 variedades de albaricoquero estudiadas en este trabajo……..…….9

Tabla 3. Listado de los 7 primers usados, su secuencia e información acerca de los fragmentos amplificados…….10

Tabla 4. Paquetes utilizados para el desarrollo del pipeline…….12

Tabla 5. Polimorfismo, número y rango de alelos en los locus estudiados…...15

Tabla 6. Número de genotipos y de alelos por país…….16

Tabla 7. Número de genotipos y de alelos por programa de mejora…….16

Tabla 8. Heterocigosidad observada, esperada y p-valor para los locus estudiados……..…….19

Tabla 9. Valores F_{IS} y F_{ST} para los locus estudiados……..……..20

1. Introducción

1.1. Contexto y justificación del Trabajo

El albaricoquero (*Prunus armeniaca* L.) es una especie frutal originaria de Asia Central, donde hay registros de su cultivo hace más de 3000 años y desde donde se diseminó al resto del mundo (Faust et al., 1998; Janick, 2005). Tradicionalmente, las variedades de albaricoquero se han clasificado en seis grupos dependiendo de su origen geográfico: Asia Central, China del Este, China del Norte, Dzhungar-Zailij, Irano-Caucásico y Europa, existiendo entre ellos diferencias en cuanto a rasgos morfológicos del fruto y propiedades de adaptación climática (Layne et al., 1996).

El albaricoquero es una especie de la familia de las Rosáceas, una de las familias de plantas más importantes económicamente a nivel mundial. Es una especie diploide con 8 cromosomas (2n = 16). Dentro del género *Prunus*, el albaricoquero es el tercer cultivo en importancia económica, alcanzándose una producción mundial de 3,88 millones de toneladas (FAO, 2018). En España su producción se localiza principalmente en la cuenca mediterránea, con una superficie de 21.002 ha, alcanzando una producción de 162.872 toneladas (Magrama, 2018). La Región de Murcia es la mayor productora, con una producción de 90.987 toneladas, seguida por la Comunidad Valenciana (19.160 t) y Aragón (16.434 t) (Magrama, 2018).

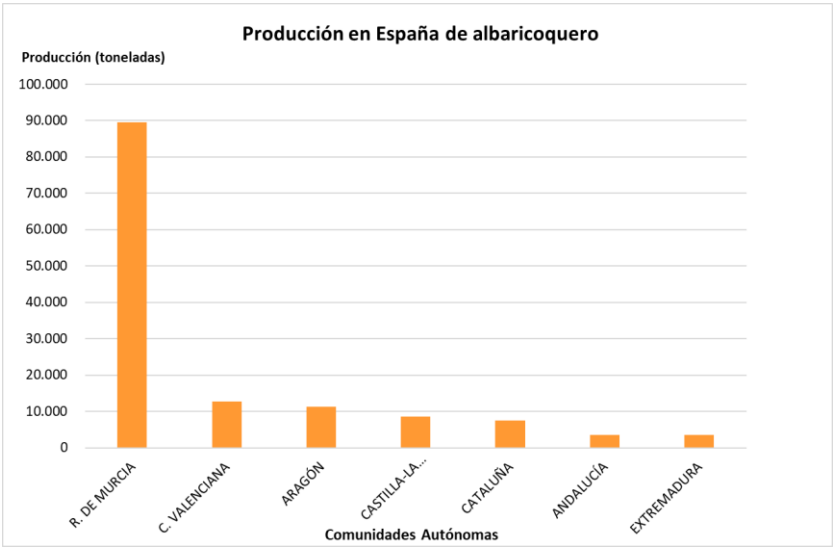


Figura 1. Gráfico de la producción de albaricoquero en España por CCAA. (Magrama, 2018)

El destino de los frutos puede ser tanto para consumo en fresco como para la fabricación de productos procesados (zumos, mermeladas, frutos desecados, licores, etc.). Aunque es posible encontrar este frutal en hábitats muy diversos, su cultivo comercial está restringido a áreas geográficas muy concretas. Además de su baja adaptabilidad, el cultivo está condicionado por las condiciones ambientales adversas y la presencia de enfermedades y plagas. Por todo ello, se han desarrollado diferentes programas de mejora con diferentes objetivos (Hormaza et al., 2007; Zhebentyayeva et al., 2012). Uno de

los principales ha sido potenciar la adaptación climatológica, desarrollando nuevas variedades que permitan ampliar la superficie de cultivo, así como alargar el calendario de producción. Otro de los objetivos ha sido la introducción de variedades resistentes al virus de la sharka (Egea et al., 1999), que en las últimas décadas ha provocado grandes daños en los países mediterráneos, reduciendo la producción y la superficie de cultivo al tener que arrancarse muchas plantaciones como única forma de combatirlo. Además, en los últimos años se ha producido un cambio en las preferencias del consumidor, apostando por unos frutos de mayor tamaño, más firmes, de color rojo y sabor más ácido (Gatti et al., 2009).

Figura 2. Variabilidad de frutos de albaricoquero de variedades procedentes de diferentes programas de mejora.

Las variedades tradicionales actúan como fuentes naturales de rasgos de interés, ya que poseen características que las hacen muy útiles para los programas de mejora, como propiedades organolépticas interesantes y su adaptabilidad a condiciones climáticas locales (de la Rosa y Martín, 2016). Sin embargo, el objetivo de los programas de mejora de introducir nuevas variedades con un mayor potencial de adaptación climática, resistencia al virus del sharka y con las nuevas características organolépticas que demanda el consumidor, ha provocado una intensa renovación varietal, surgiendo un gran número de nuevas variedades, en muchos casos procedentes de cruzamientos en los que se ha empleado un número reducido de parentales. Esto ha provocado que las variedades tradicionales se estén sustituyendo por nuevas obtenciones, reduciéndose cada vez más su presencia en las plantaciones comerciales.

La conservación y el uso de recursos genéticos de plantas debería ser una prioridad en la investigación agrícola. A finales del siglo XIX se establecieron las primeras colecciones de germoplasma con el fin de conservar todo ese material vegetal. Los bancos de germoplasma suponen un sistema eficaz y económico para evitar la pérdida de variabilidad genética y permiten que sea más accesible para su uso en la mejora vegetal (de la Rosa y Martín, 2016). Los análisis de diversidad genética en colecciones de germoplasma pueden facilitar una clasificación adecuada de las accesiones, así como el

establecimiento de colecciones nucleares, que contengan individuos bien caracterizados que sean representativos de la diversidad original.

Tradicionalmente, los estudios de diversidad genética se llevaban a cabo mediante marcadores morfológicos y agronómicos. Sin embargo, estos métodos son lentos y costosos, por lo que se realizaban pocos estudios y no eran muy precisos. Posteriormente, se desarrollaron los marcadores moleculares, basados en el polimorfismo de las secuencias de ADN, como los RFLPs ('*Restriction Fragment Length Polymorphism'*), los RAPDs ('*Random Amplified Polymorphic DNA*'), los AFLPs ('*Amplified Fragment Length Polymorphism*'), los SSRs ('*Simple Sequence Repeats'*), los ISSRs ('*Inter-Simple Sequence Repeats'*), los CAPS ('*Cleaved Amplified Polymorphic Sequence*') y los SNPs ('*Single Nucleotide Polymorphism*') (Martínez-Gómez et al., 2005).

Los microsatélites (SSR) son secuencias cortas de ADN (de 1 a 10 pares de bases) que se repiten en tándem en el genoma. Estas regiones del genoma no son codificantes y están flanqueadas por regiones específicas, y no se conoce de forma clara cuál es su origen y función (Miah et al., 2013). Pueden ser originados por diferentes causas: debido a una unión incorrecta durante la replicación de las cadenas de ADN, a una estructura terciaria anormal de la región repetitiva del ADN o por sobrecruzamientos incorrectos durante la meiosis. Este tipo de marcadores moleculares son útiles para estudios de diversidad genética porque, al no ser codificantes, las mutaciones se transmiten; además son abundantes y están uniformemente distribuidos por el genoma, son altamente polimórficos, presentan herencia mendeliana simple y son fáciles de identificar, medir y analizar (Ellegren, 2004; Vieira et al., 2016).

Hasta la actualidad han sido utilizados para caracterizar numerosas especies del género *Prunus* como cerezo (Dirlewanger et al., 2002; Liu et al., 2018; Wünch & Hormaza, 2002) o melocotonero (Dirlewanger et al., 2002). En albaricoquero se han utilizado para caracterizar variedades comerciales tradicionales (Martin et al., 2010) o locales de distintos países como España (Hormaza, 2002; Martínez-Mora et al., 2009), Turquía (Murathan et al., 2017) e Irán (Raji et al., 2014).

Con la introducción de un número creciente de nuevas variedades se desconoce el grado de diversidad actual comparado con el de hace pocos años. Para profundizar en el conocimiento de la diversidad genética actual del albaricoquero, en este trabajo se han utilizado marcadores microsatélites (SSR), con el fin de identificar un grupo de genotipos procedentes de diversos orígenes y analizar su diversidad y relaciones genéticas.

1.2. Objetivos del Trabajo

El **objetivo general** de este trabajo es analizar la diversidad genética de 50 variedades de albaricoquero, tanto tradicionales como nuevas obtenciones comerciales procedentes de diferentes programas de mejora de varios países mediante marcadores moleculares tipo SSR.

Este objetivo general se desglosa en cinco **objetivos específicos**:

1. Desarrollar un flujo de trabajo estandarizado 'pipeline' para el análisis de la diversidad genética mediante datos de microsatélites.
2. Caracterizar e identificar las variedades estudiadas mediante marcadores microsatélites.
3. Comprobar la presencia de homonimias y/o sinonimias en el material evaluado.
4. Estudiar filogenéticamente las diferentes variedades y establecer las relaciones de parentesco entre ellas.
5. Representar la distribución de la diversidad genética obtenida a partir de la amplificación con marcadores microsatélites y estudiar la estructura de dichas poblaciones.

1.3. Enfoque y método seguido

El enfoque de este estudio se centra en el análisis de diversidad de variedades de albaricoquero de diferente origen. Se ha utilizado esta especie debido a su creciente importancia económica tanto en nuestro país como a nivel mundial.
Para tener una mayor representatividad de la variabilidad genética se han utilizado variedades tradicionales de nuestro país localizadas en la colección de albaricoquero del Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA). Además, se han utilizado nuevas variedades procedentes de diferentes programas de mejora internacionales localizadas en diferentes fincas y colecciones en España.

Para estudiar la diversidad genética se han utilizado marcadores microsatélites, ya que son muy eficaces debido a su codominancia y elevado nivel de polimorfismo. Para obtener un mayor número de datos, se han utilizado 7 marcadores.

El análisis bioinformático se ha llevado a cabo con el programa R, ya que al ser un software libre permite que sea accesible para su posterior uso por otros investigadores. Tiene un lenguaje de programación fácil que proporciona una gran fuente de herramientas mediante paquetes desarrollados para este tipo de análisis. Además, ha sido ampliamente utilizado en el máster por lo que este trabajo es una herramienta útil para aplicar los conocimientos adquiridos.

1.4. Planificación del Trabajo

1.4.1. Tareas

Las tareas que se han realizado para llevar a cabo el trabajo son:

1.1. Búsqueda de información sobre paquetes y librerías de R relacionados con el estudio de la variabilidad genética y aplicados para la consecución de los diferentes objetivos específicos.

1.2. Instalación de paquetes

1.3. Diseño de pipeline para el análisis de variabilidad genética y de estructura poblacional

2.1. Aplicación pipeline para realizar el análisis de variabilidad a nivel de individuos

2.2. Estudio de la validez de los primers utilizados mediante diferentes parámetros

2.3. Estudio de la diversidad genética de las variedades analizadas mediante diferentes parámetros.

3.1. Identificación de las posibles homonimias (variedades distintas con el mismo nombre) y sinonimias (la misma variedad con distintos nombres) a partir del análisis de resultados del objetivo específico 2.

4.1. Aplicación pipeline para realizar el análisis filogenético

4.2. Análisis de datos con el 'pipeline' desarrollado identificando las posibles relaciones de parentesco entre las variedades

5.1. Aplicación pipeline para realizar el análisis poblacional

5.2. Análisis de datos con el 'pipeline' desarrollado identificando los posibles grupos poblacionales y estudiar el porqué de su distribución

1.4.2. Calendario

Se ha realizado un calendario de planificación del TFM (Tabla 1) así como un diagrama de Gantt (Figura 3) en los que se puede ver la distribución de las tareas a lo largo del tiempo.

Tabla 1. Planificación del TFM

Nombre de la tarea	Fecha de inicio	Fecha final	Duración (dias)
PEC 0. Propuesta de TFM	**19/09/2018**	**01/10/2018**	**12**
Definir contenidos	19/09/2018	23/09/2018	4
Búsqueda bibliográfica	24/09/2018	01/10/2018	7
PEC 1. Plan de trabajo	**02/10/2018**	**15/10/2018**	**13**
PEC 2. Desarrollo del trabajo Fase I	**16/10/2018**	**19/11/2018**	**34**
1.1 Búsqueda paquetes R	16/10/2018	21/10/2018	5
1.2 Instalación paquetes	22/10/2018	23/10/2018	1
1.3 Diseño pipeline	22/10/2018	08/11/2018	17
2.1 Aplicación pipeline	09/11/2018	10/11/2018	1
2.2 Validez primers	09/11/2018	12/11/2018	3
2.3 Estudio diversidad genética	13/11/2018	17/11/2018	4
3.1 Identificar sinonimias y homonimias	18/11/2018	19/11/2018	1
PEC 3. Desarrollo del trabajo Fase II	**20/11/2018**	**17/12/2018**	**27**
4.1 Aplicación pipeline	20/11/2018	21/11/2018	1
4.2 Análisis relaciones parentesco	20/11/2018	03/12/2018	13
5.1 Aplicación pipeline	04/12/2018	07/12/2018	3
5.2 Análisis estructura poblacional	07/12/2018	17/12/2018	10
PEC 4. Memoria final	**18/12/2018**	**02/01/2019**	**15**
Redacción memoria	18/12/2018	26/12/2018	8
Correcciones	27/12/2018	02/01/2019	6
PEC 5a. Elaboración de la presentación	**03/01/2019**	**10/01/2019**	**7**
PEC 5b. Defensa pública TFM	**14/01/2019**	**23/01/2019**	**9**

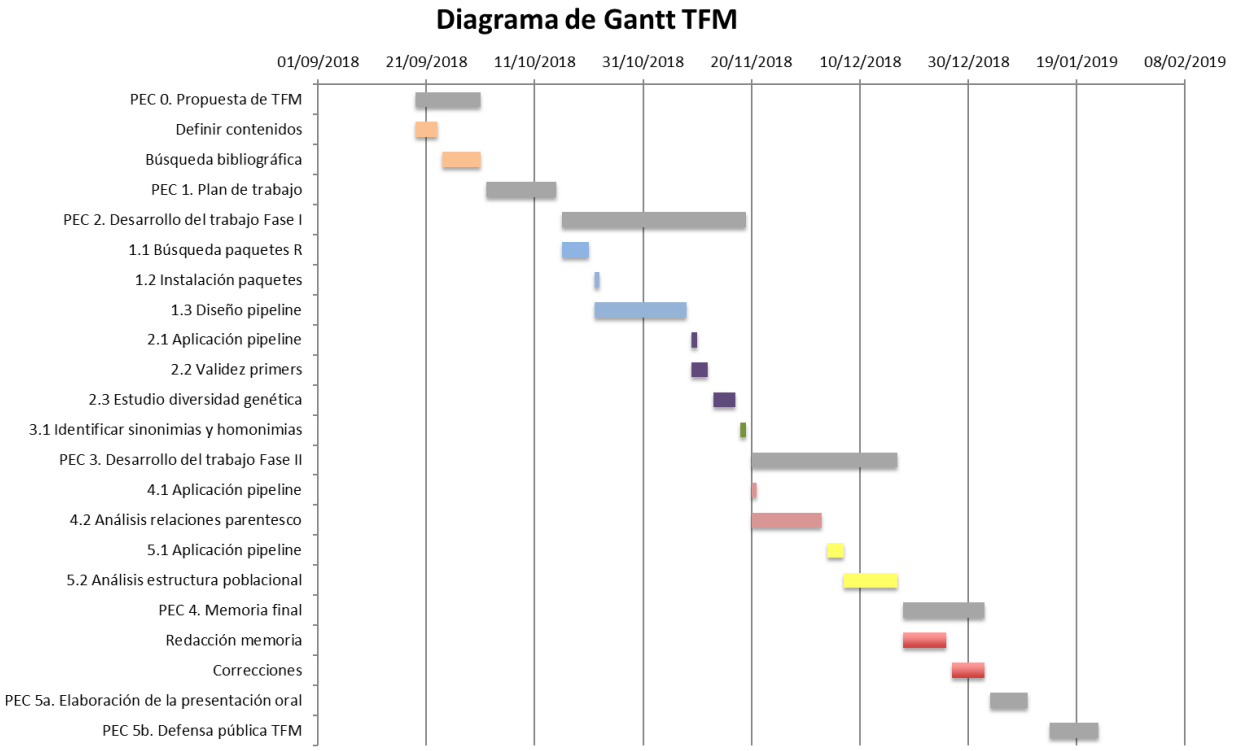


Figura 3. Diagrama de Gantt con la planificación del TFM

1.4.3. Hitos

Los hitos que se marcan en este trabajo son:

- Hito 1: Desarrollo de un pipeline para el análisis de diversidad genética y su aplicación para un análisis de variabilidad a nivel de individuos.

- Hito 2: Aplicación del pipeline para llevar a cabo un estudio filogenético y de estructura de poblaciones.

- Hito 3: Redacción de la memoria final en la que se incluyan los resultados obtenidos, una interpretación de los mismos y las conclusiones del trabajo.

1.4.4. Análisis de riesgos

Los riesgos que podían surgir durante este trabajo son:

- Posibles retrasos por motivos personales o profesionales. A pesar de que esto es difícil de prever, se habría solventado intentado adelantar las tareas por si en algún momento surge algún imprevisto o compensando con otros días dedicándole más horas.

- Que alguna de las tareas requiera más tiempo del planificado.

- Que no hubiera tanta variabilidad como se espera en las variedades y no se pudieran definir los grupos. Esto se habría solventado utilizando un mayor número de marcadores microsatélites.

- Dificultad para interpretar desde un punto de vista biológico los resultados obtenidos, ya que debido al origen de las variedades (programa de mejora) los cruzamientos han sido dirigidos y no ha habido aleatoriedad.

1.5. Breve sumario de productos obtenidos

Los productos que se han obtenido tras la realización de este trabajo han sido:

- Un guión ‘script’ elaborado mediante RMarkdown que permite estudiar la variabilidad genética mediante marcadores microsatélites.
- Gráficos que permiten establecer las relaciones genéticas existentes entre individuos, así como la estructura poblacional.

Además, durante el desarrollo de este trabajo se han generado los siguientes documentos:

- PEC 0. Propuesta de TFM
- PEC 1. Plan de trabajo
- PEC 2. Informe de seguimiento Fase I
- PEC 3. Informe de seguimiento Fase II

- Memoria Final. Incluye el contexto del trabajo, el ‘pipeline’ desarrollado y los resultados obtenidos
- Informe de Autoevaluación
- Presentación virtual

1.6. Breve descripción de los otros capítulos de la memoria

La memoria del proyecto consta de los siguientes capítulos:

Introducción: Incluye la contextualización, justificación, planificación y objetivos del trabajo.

Métodos: Se detalla el origen del material vegetal y se explica del proceso experimental que se ha llevado a cabo para obtener los datos de los microsatélites con los que se ha trabajado y cada uno de los parámetros que se han utilizado para analizar la diversidad genética.

Resultados y Discusión: Incluye el análisis e interpretación de los resultados de diversidad genética obtenidos a partir del ‘script’ desarrollado.

Conclusiones: Se incluyen las conclusiones del trabajo.

2. Materiales y métodos

2.1. Material vegetal

Se recogieron hojas jóvenes de 50 variedades de albaricoquero (*Prunus armeniaca* L.), incluyendo variedades tradicionales y nuevas obtenciones de diferentes colecciones, plantaciones y programas de mejora (Tabla 2).

Tabla 2. Nombre, origen y programa de mejora de procedencia de las 50 variedades de albaricoquero estudiadas en este trabajo

Variedades	Origen	Programa de mejora
Almabar	España	Frutaria
Almater	España	Frutaria
Apribang	Francia	Agro Selection Fruits
Aprisweet	Francia	Agro Selection Fruits
Aprix 116	España	Proseplan
Aprix 20	España	Proseplan
Berdejo	España	Tradicional
Canino	España	Tradicional
Cooper Cot	Francia	COT International
Corbató	España	Tradicional
Fabara	España	Tradicional
Farbaly	Francia	Newcot (International Plant Selection)
Farbela	Francia	Newcot (International Plant Selection)
Fardao	Francia	Newcot (International Plant Selection)
Farius	Francia	Newcot (International Plant Selection)
Goldrich	EEUU	USDA ARS (COT International)
Harval	Canada	Harrow Research Center
Henderson	EEUU	Tradicional
Holly Cot	Francia	COT International
Kioto	Francia	Escande
Kosmos	España	PBS Producción Vegetal SL
Lito	Greece	Tradicional
Magic Cot	EEUU	SDR FRUIT LLC (COT International)
Memphis	España	PBS Producción Vegetal SL
Milord	España	PBS Producción Vegetal SL
Mirlo blanco	España	CEBAS-CSIC
Mirlo rojo	España	CEBAS-CSIC
Mitger	España	Tradicional
Moniquí (1006)	España	Tradicional
Moniquí (2113)	España	Tradicional
Monster Cot	EEUU	SMS UNLIMITED (COT International)
Muñoz	España	Tradicional
Murciana	España	CEBAS-CSIC
Ninfa	Italia	Daniele Bassi (Università di Bologna)
Ninja	Francia	Escande
Palsteyn	Sudafrica	ARC Infruitec-Nietvoorbij
Paviot	Francia	Tradicional
Perle Cot	EEUU	SDR FRUIT LLC (COT International)
Rambo	España	PBS Producción Vegetal SL
Rougecot	Francia	COT International
Rubilis	Francia	Newcot (International Plant Selection)
Rubissia	Francia	Newcot (International Plant Selection)
Stark Early Orange	EEUU	Tradicional
Stella	EEUU	Tradicional
Sun Glo	EEUU	Tradicional
Sweet Cot	EEUU	Washington State University (COT International)
Swired	Suiza	Regibus (COT International)
Tornado	Francia	Escande
Valorange	España	CEBAS-CSIC
Veecot	Canada	Tradicional

2.2. Extracción de DNA y PCRs

Las extracciones de DNA se llevaron a cabo siguiendo el protocolo descrito por Hormaza (2002) y utilizando el DNeasy Plant Mini Kit (Qiagen, Hilden, Alemania). La calidad y cantidad del DNA se midió mediante el espectrofotómetro NanoDropTM ND-1000 (Bio-Science, Budapest, Hungría).

Se estudiaron 7 microsatélites mediante 7 pares de primers marcados (Tabla 3). Las reacciones de amplificación se llevaron a cabo en un volumen de 15 µl que contenía 10x NH_4 tampón de reacción, 25 mM Cl_2Mg, 2.5 mM de cada dNTP, 10 µM de cada cebador, 100 ng de ADN genómico y 0.5 U de BioTaq^{TM} DNA polymerase (Bioline, Londres, Reino Unido) siguiendo el un programa de temperaturas con un paso inicial de 1 min a 94ºC, 35 ciclos de 30 s a 94ºC, 30 s a 47/51/56/57ºC (dependiendo de cada par de primers) y 1 min a 72ºC, y un paso final de 5 min a 72ºC.

Tabla 3. Listado de los 7 primers usados, su secuencia e información acerca de los fragmentos amplificados.

Locus	Secuencia iniciadora (5′→ 3′)	Grupo de ligamiento	Motivo de repetición	Longitud de amplificación esperada (bp)	Referencia
ssrPaCITA7	CTT TTG TGC CTC AGC TTC CA ACA C CCT GGC CTG ACC CTA AGC AAT TCG	1	$(AG)_{22}$	211	Lopes et al. (2002)
ssrPaCITA10	GGT GAG GTC TGT GCT GAA TAT GCC A CGA TTA AAG AAA TAA GAA AAA GAG C	3	$(CT)_{26}$	175	Lopes et al. (2002)
ssrPaCITA23	GTG AAT ACA AAA TTT TAC TAC ATT G CGG TCT CTG ACT CTC TGA CTT GCG G	3	$(AC)_2(AG)_{18}$	146	Lopes et al. (2002)
ssrPaCITA27	GAT CCC TCA ACT GAA TCT CTC CGT CAC AAC AAT AGA TGC GAA GG		$(TC)_8(TA)_6(TG)_{17}$	262	Lopes et al. (2002)
UD*Ap*-415	AAC TGA TGA GAA GGG GCT TG ACT CCC GAC ATT TGT GCT TC	1	$(GA)_{21}$	156	Messina et al. (2004)
UD*Ap*-420	TTC CTT GCT TCC CTT CAT TG CCC AGA ACT TGA TTC TGA CCA	6	$(CT)_{20}$	175	Messina et al. (2004)
pchgms3	ACG GTA TGT CCG TAC ACT CTC CAT G CAA CCT GTG ATT GCT CCT ATT AAA C	1	$(CT)_{19}$	179	Sosinski et al. (2000)

Posteriormente, los fragmentos ampliados se analizaron mediante electroforesis capilar en el secuenciador automático CEQ^{TM} 8000 Genetic Analysis System (Beckman Coulter, Fullerton, CA, EEUU). Las muestras se desnaturalizaron a 90ºC durante 120 s, se inyectaron a 2.0 kV 30 s y se separaron a 6.0 kV durante 35 min siguiendo el protocolo de Martin et al (2010).

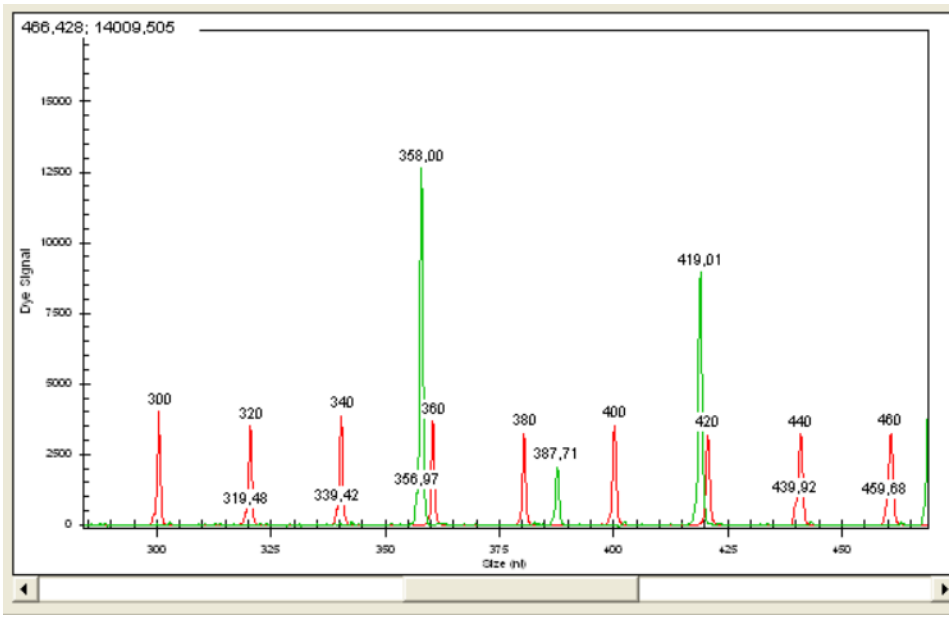


Figura 4. Captura de pantalla de una lectura de fragmentos marcados mediante el programa CEQ™ 8000 Genetic Analysis System

2.3. Análisis de la variabilidad genética

Los datos de microsatélites obtenidos para cada individuo, es decir, los alelos correspondientes a cada locus y su respectivo tamaño, se organizaron en una tabla formato .csv. El análisis estadístico se llevó a cabo con el software estadístico R.

Para desarrollar el pipeline del análisis se buscó información sobre paquetes relacionados con el estudio de la variabilidad genética. Los paquetes utilizados se muestran en la tabla 4.

Tabla 4. Paquetes utilizados para el desarrollo del pipeline

Paquete R	Descripción	Funciones utilizadas
ade4	Conjunto de herramientas para el análisis multivariado de datos	*randtest, dudi.pca*
adegenet	Conjunto de herramientas para la exploración de datos genéticos y genómicos.	*df2genind, isPoly, genind2genpop, makefreq, dist.genpop, find.clusters, dapc*
DataCombine	Conjunto de herramientas para la manipulación de datos	*InsertRow*
factoextra	Conjunto de herramientas para la visualización de datos obtenidos mediante análisis multivariantes	*fviz_eig*
ggfortify	Conjunto de herramientas para la visualización de datos obtenidos como resultado de análisis estadísticos	*autoplot*
ggplot2	Conjunto de herramientas para la creación de gráficos	*autoplot, scale_fill_manual, scale_color_manual, ggplot, geom_bar, facet_grid, theme*
hierfstat	Conjunto de herramientas que permiten la estimación y realización de los estadísticos F	*genind2hierfstat, basic.stats, fstat, gstat.randtest*
pegas	Conjunto de herramientas para realizar análisis de población y genética evolutiva	*hw.test*
phangorn	Conjunto de herramientas para realizar análisis filogenéticos	*NJ*
PopGenReport	Conjunto de herramientas para analizar datos genéticos de población y paisaje	*allel.rich*
poppr	Conjunto de herramientas para realizar análisis genéticos de población	*poppr.amova, aboot*
reshape2	Conjunto de herramientas para la modificación de los datos (reestructuración y agregación)	*melt*
stats	Conjunto de herramientas para realizar cálculos estadísticos	*bartlett.test, prcomp, hclust, as.dendrogram*

A partir de esa información se desarrolló el pipeline mediante RMarkdown para el análisis de variabilidad genética y de estructura poblacional (Anexo).

Los parámetros estudiados fueron:

- El polimorfismo o la proporción de locus polimórficos (P) es una medida del número de locus variables en una población. Un locus se considera polimórfico cuando se detecta más de un alelo.
- La riqueza alélica (Ar) cuantifica el número medio de alelos que presenta un locus en una población.
- La frecuencia alélica es una medida que refleja la proporción de un alelo específico de un determinado locus en un grupo de genotipos o en la población. Es una medida para la identificación de evolución en una población; representa la transmisión de los genes. Por tanto, si valor de las frecuencias alélicas de una generación a la siguiente cambia indica que la población está experimentando cambio evolutivo. Si, por el contrario, las

frecuencias alélicas permanecen constantes quiere decir que no ha ocurrido evolución.

- La heterocigosidad es una medida que representa la variabilidad genética en una población. En este trabajo se ha estudiado mediante dos parámetros: la heterocigosidad esperada (He) y la heterocigosidad observada (Ho).
  La heterocigosidad observada representa la relación entre el número de individuos heterocigotos observados para un locus y el número total de individuos analizados para ese determinado locus.
  La heterocigosidad esperada o diversidad genética representa la probabilidad de que dos alelos elegidos al azar en un individuo para un determinado locus sean diferentes entre sí.

- Los estadísticos de Wright son una medida para cuantificar la distribución de la diversidad genética dentro y entre las poblaciones conforme a la ley del equilibrio de Hardy-Weinberg (H-W) (Nei, 1987).
  El equilibrio H-W plantea que en una población en la que el cruzamiento es aleatorio, la composición genética permanece en equilibrio mientras no haya mutaciones, migraciones o selección natural. Esta teoría funciona como hipótesis nula y se utiliza como referencia para evaluar la magnitud del cambio evolutivo en las poblaciones analizadas comparando las frecuencias genotípicas estimadas en las poblaciones naturales que estudiamos con las que esperaríamos encontrar según el equilibrio H-W.
  Se realizó un test mediante el algoritmo de Monte Carlo (1000 réplicas) para comprobar que las frecuencias genotípicas de los marcadores estudiados cumplían con el equilibrio H-W.
  Además, en este trabajo se han estudiado dos estadísticos F de Wright:
  El parámetro F_{IS} (coeficiente de endogamia) estima la reducción en la heterocigosidad de un individuo debido al cruzamiento no aleatorio dentro de su subpoblación. Los valores varían entre -1, que indica un exceso de heterocigotos, y 1, que indica ausencia de heterocigotos.
  El parámetro F_{ST} (coeficiente de diferenciación genética o índice de fijación) mide el grado de diferenciación genética entre poblaciones. El rango de valores varía entre 0, en el que no existe diferenciación genética, y 1, donde las frecuencias alélicas están fijadas e indican que las poblaciones analizadas son distintas.

El estudio de las relaciones filogenéticas se llevó a cabo mediante el cálculo de la distancia genética de Nei (Nei, 1987). A partir de las distancias genéticas se realizó un dendrograma utilizando como método de agrupación el método de UPGMA (Unweighted Pair Group Method using Arithmetic averages) (Mohammadi and Prasanna, 2003).
El método UPGMA es un método de análisis filogenético que se basa en la distancia genética. Los datos que presentan una menor distancia genética se van agrupando formando una matriz de distancias. A partir de estos datos se forma un árbol enraizado y ultramétrico que representa la relación evolutiva asumiendo ancestros comunes hasta llegar a un ancestro común a todos. Además, se ha realizado un análisis bootstrap para evaluar la eficiencia del agrupamiento. El análisis devuelve un porcentaje que representa la cantidad de veces, en relación al total de pseudorréplicas, que el mismo agrupamiento se repite, lo que constituye una medida de robustez.

Para estudiar la diversidad molecular se realizó un AMOVA (Análisis Molecular de la Varianza). El análisis AMOVA se utiliza para calcular el nivel de diferenciación genética entre diferentes poblaciones y cuanta de esta diferenciación es debida a las diferencias entre poblaciones, entre muestras dentro de las poblaciones y/o entre muestras.

El Análisis de Componentes Principales (ACP) es un análisis estadístico multivariante que se utiliza para transformar un conjunto de variables en un nuevo conjunto. El nuevo conjunto de variables, denominadas componentes, es menor y contiene información que representa la varianza.

Se ha realizado un dendrograma utilizando la distancia genética de Nei mediante el método de agrupación UPGMA para estudiar las relaciones entre los programas de mejora.

Para el Análisis Discriminante de Componentes Principales (AD) se realizó en primer lugar la identificación de grupos. Mediante el Criterio de Información Bayesiano (BIC) se identificó un número óptimo de agrupaciones genéticas para describir los datos.
A continuación, se realizó el AD. Este tipo de análisis busca combinaciones lineales de las variables originales (alelos) que muestran las diferencias entre los grupos de la mejor manera posible y minimizan la variación dentro de los grupos. Sobre la base de las funciones discriminantes retenidas, el análisis deriva probabilidades para cada individuo de pertenencia a cada uno de los diferentes grupos lo se puede interpretar como la "proximidad genética" de cada individuo a los diferentes grupos.

3. Resultados y discusión

3.1. Estudio de la validez de los marcadores microsatélites utilizados

Primeramente, se realizó un estudio de la validez de los primers utilizados para amplificar las 7 secuencias microsatélites.

3.1.1. Polimorfismo y riqueza alélica

Se estudiaron un total de 50 genotipos mediante 7 locus. Todos los marcadores (ssrPaCITA7, ssrPaCITA10, ssrPaCITA23, ssrPaCITA27, UDAp-415, UDAp-420 y pchgms3) resultaron ser polimórficos. Esto demuestra su utilidad para los estudios de diversidad genética en albaricoquero, incluso en el caso del marcador *pchgms3* a pesar de haber sido desarrollado para melocotonero, otra especie del género *Prunus* (Sosinki et al., 2000).

Se detectaron un total de 52 alelos, variando entre 6 y 10 el número encontrado para cada uno de los locus (Tabla 5). En un trabajo previo se obtuvo un rango de 2 a 7 alelos para los mismos marcadores (Martin et al., 2010), una cifra inferior a nuestros resultados.

Tabla 5. Polimorfismo, número y rango de alelos en los locus estudiados

	Polimorfismo	Número de alelos	Rango mínimo	Rango máximo
ssrPaCITA7	TRUE	10	189	224
ssrPaCITA10	TRUE	9	155	180
ssrPaCITA23	TRUE	9	140	156
ssrPaCITA27	TRUE	6	228	267
UDAp_415	TRUE	6	149	171
UDAp_420	TRUE	6	158	181
pchgms3	TRUE	6	180	200

Las variedades provienen de 8 países, siendo las de España y Francia las más numerosas, con 20 y 15 muestras respectivamente. Aunque el número de variedades procedentes de EEUU es menor, 9, el número total de alelos que contienen es igual al de las variedades de España y Francia (Tabla 6).

Además, hay dos variedades procedentes de Canadá y una variedad procedente de Italia, Sudáfrica, Suiza y Grecia.

Tabla 6. Número de genotipos y de alelos por país

	Número de genotipos	Número de alelos
Canada	2	18
España	20	37
Francia	15	37
Italia	1	13
Sudafrica	1	13
EEUU	9	37
Suiza	1	11
Greece	1	11

Las variedades estudiadas incluyen variedades tradicionales y nuevas obtenciones de 16 programas de mejora. El mayor número de variedades lo encontramos en el grupo 'Tradicional' que engloba las variedades cultivadas tradicionalmente en diferentes países. En la mayoría de los programas de mejora (11) se han analizado entre 1 y 2 variedades, con un rango de alelos entre 11 y 20.

Tabla 7. Número de genotipos y de alelos por programa de mejora.

	Número de genotipos	Número de alelos
Harrow Research Center	1	11
CEBAS-CSIC	4	23
Newcot (International Plant Selection)	6	28
Daniele Bassi (Università di Bologna)	1	13
ARC Infruitec-Nietvoorbij	1	13
Tradicional	15	36
Agro Selection Fruits	2	15
Proseplan	2	17
SMS UNLIMITED (COT International)	1	13
SDR FRUIT LLC (COT International)	2	20
USDA ARS (COT International)	1	13
COT International	3	27
Washington State University (COT International)	1	13
Regibus (COT International)	1	11
Escande	3	16
PBS Producción Vegetal SL	4	23
Frutaria	2	16

3.1.2. Frecuencias alélicas

Mediante la función *makefreq* se estudiaron las frecuencias alélicas de cada uno de los locus, estableciendo como niveles poblacionales el programa de mejora del que provenían. Para poder analizar los resultados de una manera más clara y visual se realizó un gráfico de barras para cada locus. La figura 5 representa el gráfico de barras correspondiente al locus ssrPaCITA7, en el que se muestran todos los alelos para ese locus y su frecuencia en cada uno de los países de origen.

Para este locus se detectaron en total 10 alelos. Se puede observar que algunos de ellos solo aparecen en un programa de mejora, como ocurre por ejemplo con el alelo ‘189’ en Agro Selection Fruits (ASF), el alelo ‘196’ en Università di Bologna, el alelo ‘210’ en New Cot (IPS) y el alelo ‘218’ en Harrow Research Center. Por el contrario, el alelo ‘212’ se encuentra en casi todos los programas a excepción de Harrow Research Center, CEBAS-CSIC y Università di Bologna y en una frecuencia similar (Figura 5).

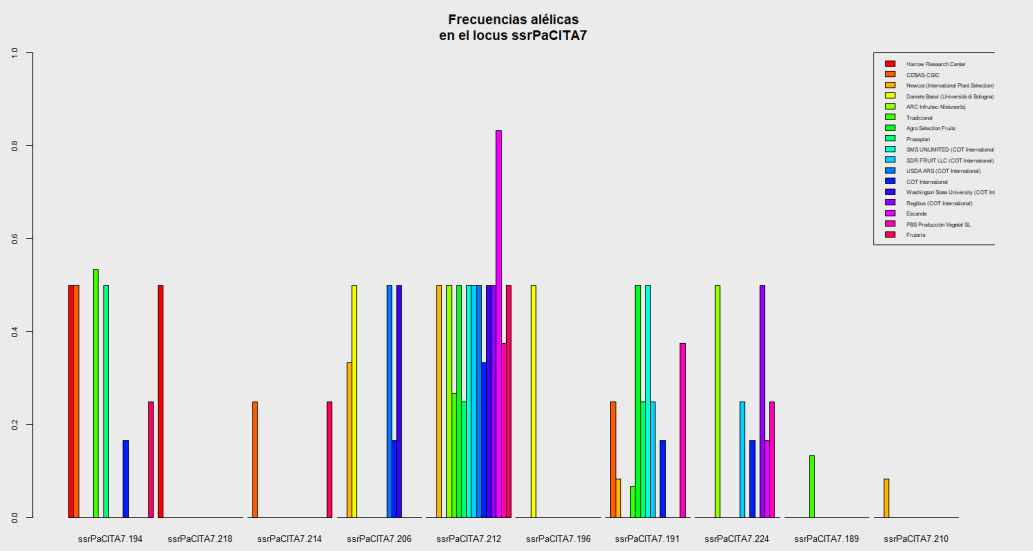


Figura 5. Frecuencias alélicas de las variedades de albaricoquero para el locus ssrPaCITA7

Se realizaron los gráficos de barras correspondientes a los otros 6 marcadores (Figura 6). Así como ocurre en el marcador ssrPaCITA7, algunos de los alelos de estos microsatélites se encuentran únicamente en variedades de un único programa de mejora, indicando la presencia de alelos únicos, presentes exclusivamente en un genotipo o programa de mejora, lo que podría ser de utilidad como marcador específico. Es el caso del alelo ‘181’, obtenido con el marcador pchgms3 que es específico para la variedad ‘Pepito Blanco’ en Martínez-Mora et al. (2009).

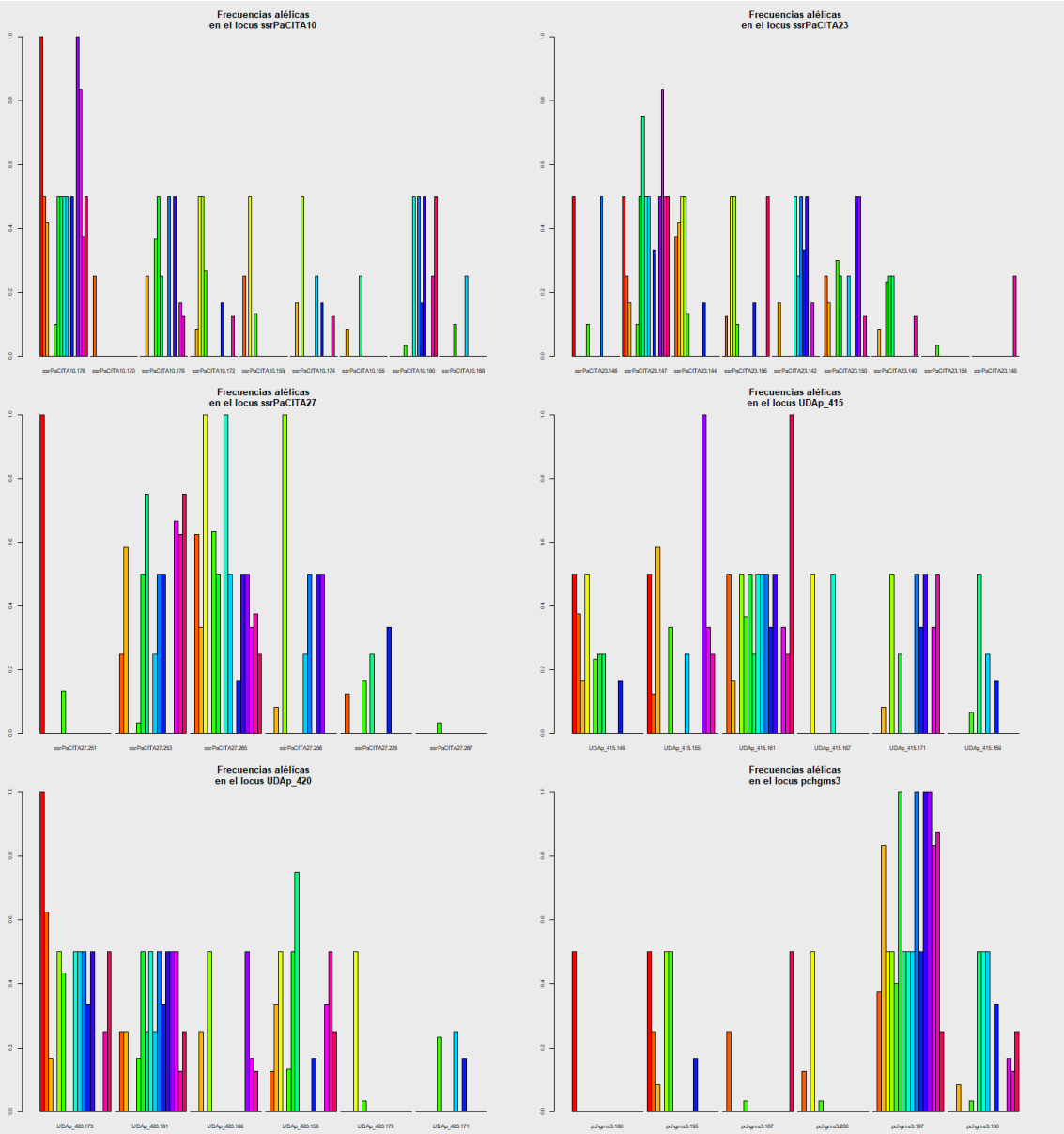


Figura 6. Frecuencias alélicas de las variedades de albaricoquero para 6 de los locus estudiados (ssrPaCITA10, ssrPaCITA23, ssrPaCITA27, UDAp_415, UDAp_420, pchgms3)

3.2. Estudio de la diversidad genética de las variedades

En primer lugar, realizamos el test de Bartlett de homogeneidad de varianzas. En este test se compara la heterocigosidad observada (Ho) y la esperada (He) por cada locus.

La hipótesis nula plantea la igualdad de heterocigosidades, y como hipótesis alternativa que existen diferencias entre ellas. Como el p-valor obtenido fue 0.1784, y éste es un valor mayor de 0.05, indica que no existen diferencias entre la heterocigosidad observada media y la esperada. Este resultado nos indica que los 7 marcadores moleculares microsatélites utilizados presentan una marcada homogeneidad y estabilidad, y son útiles para el estudio realizado.

La heterocigosidad observada fue mayor a la esperada en 4 de los locus (ssrPaCITA10, ssrPaCITA23, UDAp_415 y UDAp_420), variando entre 0.78 y 0.88. En los otros tres locus estudiados no se dio esta situación. Para los locus ssrPaCITA7 y pchgms3 ambos valores fueron similares, con unas diferencias de 0.01 y 0.04, respectivamente. Sin embargo, en el caso del locus ssrPaCITA27 la diferencia entre la heterocigosidad observada (0.54) y la heterocigosidad esperada (0.67) fue más acentuada, con una diferencia de 0.13 (Tabla 8).

En general, la heterocigosidad observada media en el conjunto de variedades (0,74) fue superior a la obtenida en estudios previos en los que solo se estudiaban variedades tradicionales locales (Martínez-Mora et al., 2009; Martin et al., 2010). Esto puede ser debido a la gran variedad de orígenes de las variedades de este trabajo.

A continuación, comprobamos si las poblaciones se encuentran en equilibrio Hardy-Weinberg. Se establece un p-valor de 0.05, esto significa que hay una probabilidad del 5% de que las diferencias en los genotipos se deban al azar y del 95% de probabilidad de que no se deban al azar. Los datos obtenidos muestran que, a excepción del locus ssrPaCITA27, todos los p-valores son mayores o iguales que 0.05 (Tabla 8). Por tanto, aceptamos la hipótesis nula que planteaba que las poblaciones se encontraban bajo el equilibrio H-W.

Tabla 8. Heterocigosidad observada, esperada y p-valor para los locus estudiados

	Hobs	Hexp	p-valor
ssrPaCITA7	0,76	0,77	0,05
ssrPaCITA10	0,82	0,79	0,78
ssrPaCITA23	0,88	0,82	0,52
ssrPaCITA27	0,54	0,68	0,04
UDAp_415	0,78	0,75	0,35
UDAp_420	0,84	0,76	0,16
pchgms3	0,56	0,60	0,14

Tras estudiar la heterocigosidad, se utilizaron los estadísticos F de Wright como medida para cuantificar la distribución de la diversidad genética dentro y entre las poblaciones (Tabla 9). El estadístico F_{IS} indica el coeficiente de endogamia dentro de poblaciones. En este caso, en todos los locus se obtuvieron valores negativos, a excepción de ssrPaCITA27, lo que indica una tendencia al exceso de heterocigotos.

En cuanto al índice de fijación F_{ST}, los valores oscilaron entre 0.06 a 0.22. Los valores para los locus ssrPaCITA7, ssrPaCITA23, UDAp_415 y UDAp_420, al ser menores que 0.15, indicaron que existe una escasa diferenciación genética. Al contrario que lo que ocurría con los tres locus restantes, ssrPaCITA10, ssrPaCITA27 y pchgm3, para los que las poblaciones estaban bien diferenciadas.

La media general muestra la tendencia de la existencia de una escasa diferenciación genética (0.13). Este índice de fijación es muy inferior al encontrado en otros estudios de microsatélites en albaricoquero (Romero et al., 2003 [0.32]; Martin et al., 2010 [0.38]), lo que podría ser debido a los cruzamientos no al azar realizados en los programas de mejora.

Tabla 9. Valores F_{IS} y F_{ST} para los locus estudiados

	F_{IS}	F_{ST}
ssrPaCITA7	-0,19	0,07
ssrPaCITA10	-0,27	0,15
ssrPaCITA23	-0,25	0,10
ssrPaCITA27	0,03	0,22
UDAp_415	-0,19	0,12
UDAp_420	-0,19	0,06
pchgms3	-0,28	0,18
Media	**-0,19**	**0,13**

3.3. Identificación de homonimias y/o sinonimias

En este estudio se analizaron 50 genotipos de variedades de albaricoquero. En ninguna de las accesiones se encontró la presencia de homonimias, es decir, genotipos diferentes denominados con el mismo nombre.

Por otro lado, se estudió la presencia de sinonimias, diferentes denominaciones para un mismo genotipo. Se encontró la presencia de 3 sinonimias correspondientes a las muestras 'Moniquí 2113', 'Muñoz' y 'Moniquí 1006'. Estos tres genotipos podrían ser clones de la variedad 'Moniquí', una variedad tradicional y ampliamente cultivada en España (Hormaza et al., 2007).

3.4. Relaciones filogenéticas entre las variedades

Las 50 variedades estudiadas se han agrupado mediante el método UPGMA utilizando la distancia genética de Nei a través de la función *aboot* del paquete poppr (R).

En líneas generales las variedades se agruparon en función del programa de mejora del que procedían (Figura 7). En total se identificaron 6 grupos. Hay un grupo claramente diferenciado, el grupo I, compuesto por variedades tradicionales tanto españolas como norteamericanas que presentaron gran similitud entre ellas. Este grupo, a su vez, se puede dividir en dos subgrupos. El primer subgrupo está constituido por 7 variedades tradicionales españolas, y se puede observar que tres de ellas (Muñoz, Moniquí 2113 y Moniquí 1006) son sinonimias. Además, podemos encontrar en este grupo la variedad Ninfa (Italia), aunque separada de las otras. Por otro lado, se puede observar un subgrupo de variedades tradicionales procedentes de América del Norte, en el

que se encuentran Stella y Veecot de EEUU, y Harval de Canadá (Harrow Research Center).

El grupo II constituye el grupo más numeroso, con un total de 19 genotipos. Todos ellos son variedades comerciales que presentan gran similitud entre ellas, pudiéndose dividir, en general, en subgrupos que corresponden a los programas de mejora.

En el grupo III se agruparon 11 genotipos que incluyen variedades comerciales que pertenecen a Cot Internacional (provenientes de diferentes programas de mejora) relacionadas con las variedades tres programas de mejora españoles diferentes, así como con las variedades tradicionales Lito, Stark Early Orange y Henderson, lo que podría indicar su posible utilización como parentales en dichos programas de mejora (Zhebentyayeva et al., 2012).

El grupo IV incluyó tres variedades tradicionales de España, Francia y EEUU, así como dos nuevas variedades.

En el grupo V se agruparon algunas de las variedades procedentes de programas de mejora españoles. Por un lado, las variedades de CEBAS-CSIC (Mirlo blanco y Mirlo rojo) y, por otro, una de las variedades de Frutaria (Almabar).

La variedad procedente de Sudáfrica, Palsteyn, se encuentra muy separada genéticamente del resto de individuos y ha sido clasificada en el grupo VI.

Estos resultados coinciden con estudios previos en los que se ha observado que las variedades se clasifican en el dendrograma en relación a su pedigree (Sanchez-Perez, et al., 2005) y/o origen geográfico de los genotipos (Hormaza, 2002; Pedryc et al., 2009), identificando grupos de variedades procedentes de Europa y Norte América, entre otros. Esto es debido probablemente a la naturaleza genética de este material vegetal, ya que la mayoría de los genotipos provienen de cruzamientos entre genotipos con un pedigree similar.

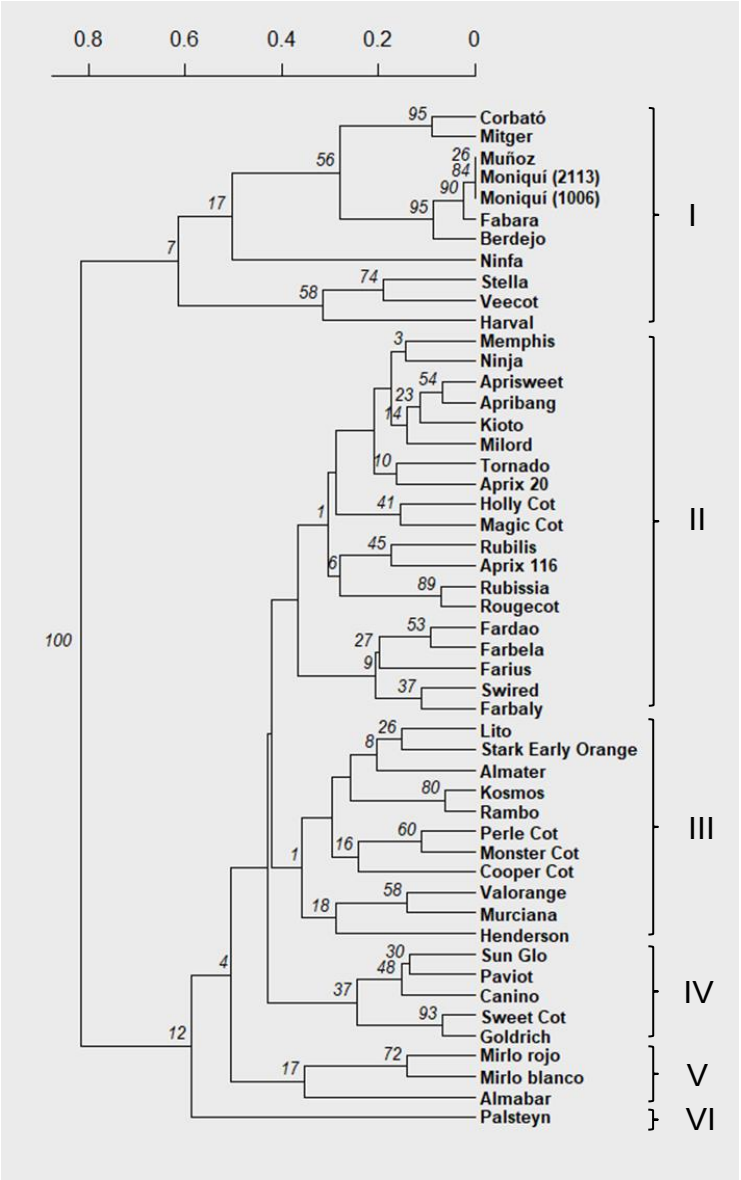


Figura 7. Dendrograma de las 50 variedades de albaricoquero incluidas en este estudio generado por el análisis de conglomerados UPGMA a partir de la matriz de similitud basada en el coeficiente de similitud de Nei.

3.5. Estructura poblacional

3.5.1. AMOVA

De la salida que nos proporciona la función *poppr.amova* se puede observar que el 97% de la variación proviene de muestras, mientras que el 4% proviene de poblaciones (establecidas como 'Origen'). Para la variación entre las muestras dentro de 'Programa de mejora' hemos obtenido un resultado negativo, -13,6%. Esto puede ser debido a que los cruzamientos son dirigidos o a una baja variabilidad debida a la endogamia. Además, el valor Phi es bajo, 0.02 dentro de las muestras y 0.04 entre las poblaciones, lo que se interpreta como un bajo nivel de diferenciación en las poblaciones.

```
$`call`
ade4::amova(samples = xtab, distances = xdist, structures = xstruct)

$results
                                       Df    Sum Sq  Mean Sq
Between Origen                          7  58.95556 8.422222
Between Programa_mejora Within Origen  13  91.61944 7.047650
Between samples Within Programa_mejora 29 108.12500 3.728448
Within samples                         50 259.00000 5.180000
Total                                  99 517.70000 5.229293

$componentsofcovariance
                                                              Sigma
%
Variations   Between Origen                                      0.2269398
4.25871
Variations    Between  Programa_mejora  Within  Origen       0.6476760
12.15417
Variations    Between  samples  Within  Programa_mejora  -0.7257759  -
13.61977
Variations   Within samples                                     5.1800000
97.20690
Total variations                                               5.3288400
100.00000

$statphi
                                  Phi
Phi-samples-total            0.02793104
Phi-samples-Programa_mejora -0.16294103
Phi-Programa_mejora-Origen   0.12694800
Phi-Origen-total             0.04258710
```

3.5.2. Análisis de Componentes Principales

Se realizó un gráfico ACP (Análisis de Componentes Principales) basado en los datos de los alelos de cada uno de los 7 microsatélites mediante la función *dudi.pca* (ade4). A partir de los valores Eigenvalue (valores propios) se seleccionó 20 como el número óptimo de componentes principales, ya que explicaban el 90% de la varianza acumulada.

A partir de ello se realizó el gráfico en el que se representaron las dos primeras componentes (Figura 8). La primera componente tiene un aproximadamente un 21% de información y la segunda un 10%. Por lo tanto, mediante la representación en dos dimensiones de las dos primeras componentes se representa un 31% de la varianza.

El primer eje (CP-1, 21,02%) podría indicar la diferenciación en cuanto al tipo de obtención, ya que las variedades tradicionales tienden a situarse en la parte izquierda del gráfico mientras que las nuevas obtenciones se sitúan principalmente en la parte derecha.

El segundo eje (CP-2, 10,1%) tiende a distinguir entre el origen geográfico de las variedades. La variedad tradicional procedente de Canadá se encuentra en el extremo inferior, mientras que la variedad procedente de Suiza se encuentra

en el extremo superior. Por otro lado, las variedades tradicionales procedentes de España se encuentran en la parte superior del eje, mientras que las nuevas obtenciones españolas se localizan principalmente en la parte inferior. Las variedades francesas se localizan en su mayoría en la parte superior del eje, a excepción de Rubilis y Rubisia. Las variedades de Estados Unidos se sitúan generalmente en una posición intermedia, ya que se distribuyen por toda la zona media del gráfico, mientras que las variedades de Grecia (Lito), Italia (Ninfa) y Sudáfrica (Palsteyn) se localizan en el eje. A pesar de localizarse las variedades en diferentes puntos del gráfico, siguiendo el criterio de procedencia geográfica, estas agrupaciones no son del todo evidentes, y el análisis no permitió distinguir claramente ninguna, lo que se podría explicar por la baja variabilidad de parentales utilizados en los programas de mejora, ya que tienen unos objetivos similares (Zhebentyayeva et al., 2012).

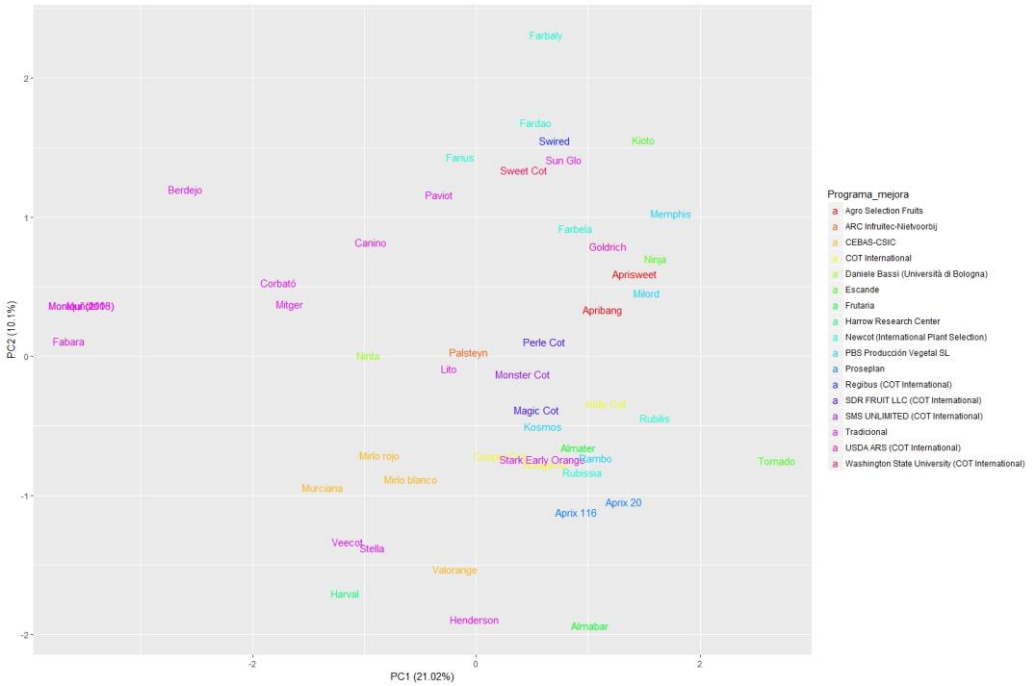


Figura 8. Análisis de componentes principales basado en datos de frecuencia de alelos microsatélites de 50 variedades. Las variedades están coloreadas por el programa de mejora del que provienen.

3.5.3. UPGMA

Se ha realizado un dendrograma usando el algoritmo UPGMA mediante la distancia genética de Nei, considerando el programa de mejora del que proceden las variedades estudiadas.

Se pueden observar cuatro grupos claramente diferenciados (Figura 9). El primero, grupo I, engloba a los programas de mejora de Canadá (Harrow Research Center), Italia (Università di Bologna) y Sudáfrica (ARC Infruitec-Nietvoorbij). Las variedades que pertenecen a Cot Internacional (grupos II y IV) provienen de programas de mejora diferentes, aunque algunos de ellos están estrechamente relacionados como es el caso de Washington State University y USDA ARS, en el grupo II, y SDR Fruit LLC y SMS Unlimited, en el grupo IV. Otro gran grupo, el grupo III, engloba a las variedades procedentes de

programas de mejora tanto de España como de Francia, así como a las variedades tradicionales.

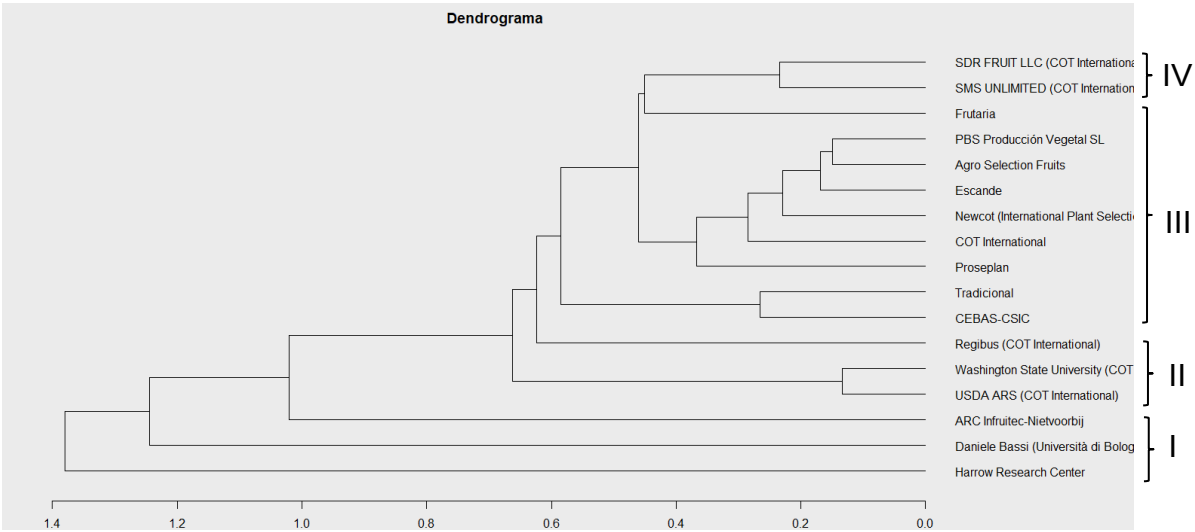


Figura 9. Dendrograma de las poblaciones agrupadas por programas de mejora generado por el análisis de conglomerados UPGMA basado en la distancia genética de Nei.

3.5.4. Análisis Discriminante de Componentes Principales

Mediante la función *find.clusters* se seleccionó el número de clusters idóneo, que corresponde con el número de clusters que tiene el menor valor BIC. En este caso son 6 (Figura 10).

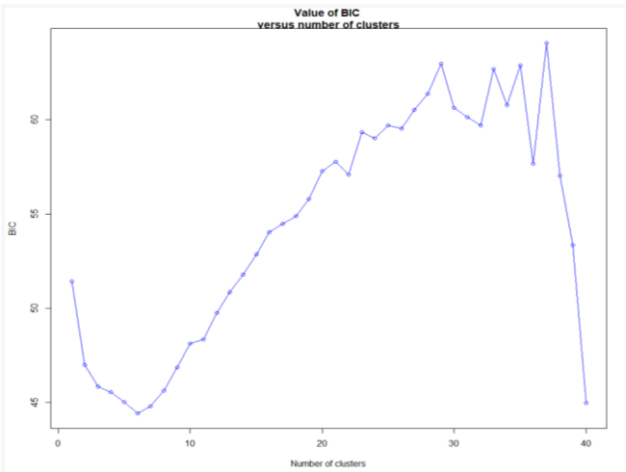


Figura 10. Gráfico de los valores BIC que muestra el número de clusters óptimo

A partir de ello se realizó un análisis AD (Análisis Discriminante de Componentes Principales) y se representó mediante un gráfico de barras, que muestra la probabilidad de cada individuo de pertenecer a esa población (Figura 11).
Solo dos genotipos se diferencian claramente, Harval por un lado y Ninfa por otro. En el resto de individuos existe gran variabilidad, incluso en las variedades tradicionales. Esto puede ser debido a su origen ya que, al tratarse en su mayoría de variedades obtenidas mediante cruzamientos dirigidos, pudiendo tener los mismos parentales o compartir uno de ellos.

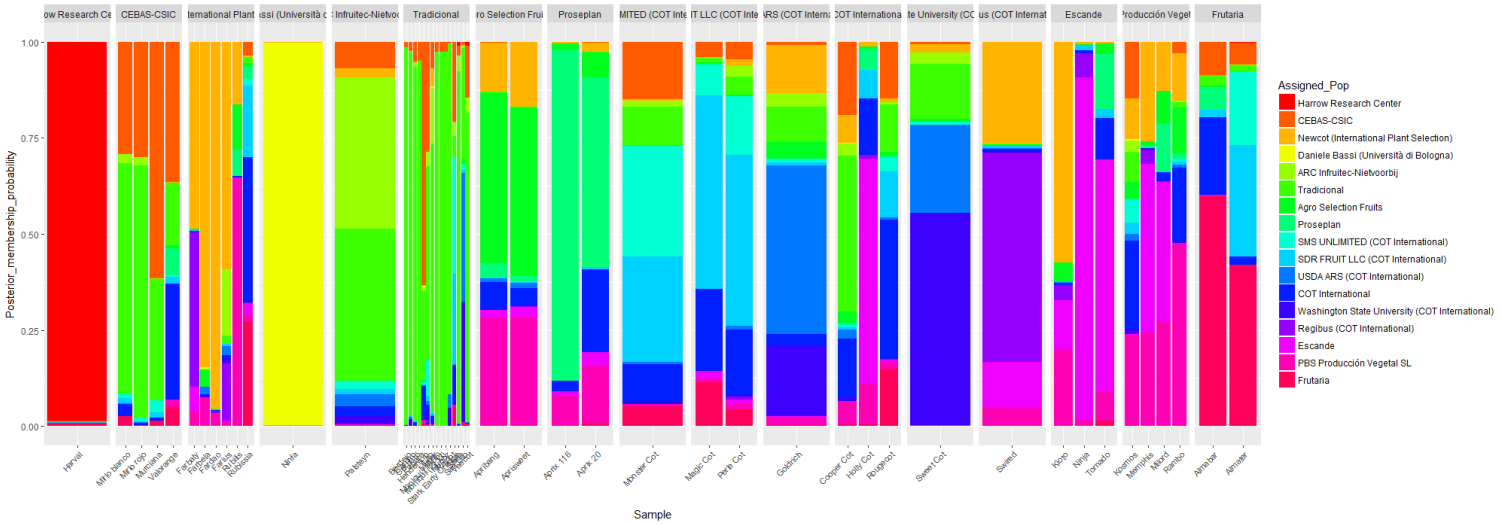


Figura 11. Diagrama de barras que representa la probabilidad de cada variedad de pertenecer a cada población

4. Conclusiones

- Los marcadores microsatélites utilizados (ssrPaCITA7, ssrPaCITA10, ssrPaCITA23, ssrPaCITA27, UDAp-415, UDAp-420 y pchgms3) muestran un alto grado de polimorfismo en albaricoquero, por lo que son buenos indicadores para estudios de diversidad genética en esta especie.

- De los 50 genotipos estudiados, se han identificado 48 variedades, detectándose 3 sinonimias que corresponden a la misma variedad.

- El índice de fijación F_{ST} indica que existe una escasa diferenciación genética en nuestro conjunto de variedades estudiadas, lo que podría ser debido a los cruzamientos dirigidos y no al azar.

- El análisis filogenético ha mostrado una tendencia de agrupación de los genotipos en relación a su procedencia (programa de mejora), observándose un grupo claramente diferenciado que contiene las variedades tradicionales de diferentes países.

- No se ha detectado una clara estructura genética poblacional mediante el Análisis Discriminante de Componentes Principales, probablemente debido a la presencia de un elevado intercambio de material genético entre los diferentes programas de mejora.

5. Glosario

SSR: Simple Sequence Repeats

He: Heterocigosidad esperada

Ho: Heterocigosidad observada

H-W: Hardy-Weinberg

UPGMA: Unweighted Pair Group Method using Arithmetic averages

AMOVA: Análisis Molecular de la Varianza

ACP: Análisis de Componentes Principales

CP: Componente Principal

BIC: Criterio de Información Bayesiano

AD: Análisis Discriminante de Componentes Principales

6. Bibliografía

Egea, J., Burgos, L., Martínez-Gómez, P., Dicenta, F. (1999). Apricot breeding for sharka resistance at C.E.B.A.S-C.S.I.C, Murcia (Spain). *Acta Hort*. 488, 153-158. doi: 10.17660/ActaHortic.1999.488.20

Ellegren H. (2004). Microsatellites: simple sequences with complex evolution. *Nat Rev Genet,* 5(6), 435-445. doi:10.1038/nrg1348

Dirlewanger, E., Cosson, P., Tavaud, M., Aranzana, M. J., Poizat, C., Zanetto, A., et al. (2002). Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). *Theor Appl Genet*, 105, 127–138. https://doi.org/10.1007/s00122-002-0867-7

FAO. Faostat – Statistics Database. Disponible en http://www.fao.org/faostat

Faust, M., Surányi, D., & Nyujtó, F. (1998). Origin and Dissemination of Apricot. *Hortic. Rev*. 22, 225–266. doi: 10.1002/9780470650738.ch6

Gatti, E., Defilippi, B. G., Predieri, S., & Infante, R. (2009). Apricot (*Prunus armeniaca* L.) quality and breeding perspectives. *Journal of Food, Agriculture and Environment*, 7(3–4), 573– 580.

Hormaza, J. I. (2002). Molecular characterization and similarity relationships among apricot (*Prunus armeniaca* L.) genotypes using simple sequence repeats. *Theoretical and Applied Genetics*, 104(2–3), 321–328. https://doi.org/10.1007/s001220100684

Hormaza, J. I., Yamane, H., & Rodrigo, J. (2007). “Apricot,” in *Genome mapping and molecular breeding in plants. V. 4, Fruit and nuts*, ed. C. Kole (Berling, Heidelberg, New York: Springer), 171–185.

Janick, J. (2005). The origins of fruits, fruit growing, and fruit breeding. *Plant Breed. Rev*. 25, 255–321. doi: 10.1002/9780470650301.ch8

Layne, R. E. C., Bailey, C., & Hough, L. F. (1996). “Apricots,” in *Fruit breeding, Volume I: Tree and Tropical Fruits*, eds J. Janick and J. N. Moore (New York, NY: John Wiley & Sons, Inc.), 79–111.

Liu, C. L., Qi, X. L. Song, L. L., Li, Y. H., Li, M. (2018). Species identification, genetic diversity and population structure of sweet cherry commercial cultivars assessed by SSRs and the gametophytic selfincompatibility locus. *Scientia Horticulturae*, 237, 28-35. doi: 10.1016/j.scienta.2018.03.063

Lopes, M.S., Sefc, K.M., Laimer, M., Da Câmara Machado, A. (2002). Identification of microsatellite loci in apricot*. Mol Ecol Notes,* 2, 24–26.

MAGRAMA (2018). Ministerio de Agricultura, Alimentación y Medio Ambiente.

Martín, C., Hormaza, J. I., & Herrero, M. (2010). Characterization and recovery of apricot germplasm from an old stone collection. *Acta Horticulturae*, *859*(8), 117–120. https://doi.org/10.1371/journal.pone.0023979

Martinez-Gómez, P., National, S., Rubio, M., & National, S. (2005). Application of Recent Biotechnologies to *Prunus* Tree Crop Genetic Improvement. *Cien.Inv. Agr*., 32(2), 55-78.

Martínez-Mora, C., Rodríguez, J., & Cenis, J. L. (2009). Genetic variability among local apricots (*Prunus armeniaca* L.) from the Southeast of Spain. *Span J Agric Res*, *7*(4), 855–868.

Messina, R., Lain, O., Marrazzo, M.T., Cipriani, G., Testolin, R. (2004). New set of microsatellite loci isolated in apricot. *Mol Ecol Notes* 4, 432–434.

Miah, G., Rafii, M. Y., Ismail, M. R., Puteh, A. B., Rahim, H. A., Islam, N. K., Latif, M. A. (2013). A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *Int. J. Mol. Sci*., 14, 22499–22528, doi:10.3390/ijms141122499.

Mohammadi, S. A. & Prasanna, B. M. (2003). Review and Interpretation *Analysis of* Genetic Diversity in Crop Plants —Salient Statistical Tools. Crop Science, 43, 1235-1248. http://dx.doi.org/10.2135/cropsci2003.1235

Murathan Z. T., Kafkas S., Asma B. M., Topçu H. (2017) *S* _allele identification and genetic diversity analysis of apricot cultivars. *J Hortic Sci Biotechnol*, 92, 251–260.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press: New York.

Pedry, A., Ruthner, S., Hermán, R., Krska, B., Hegedu, A., Halász, J. (2009) Genetic diversity of apricot revealed by a set of SSR markers from linkage group G1. *Scientia Horticulturae*, 121, 19–26. https://doi.org/10.1016/j.scienta.2009.01.014

Raji, R., Jannatizadeh, A., Fattahi, R., & Esfahlani, M. A. (2014). Investigation of variability of apricot (*Prunus armeniaca* L.) using morphological traits and microsatellite markers. *Scientia Horticulturae*, *176*(September), 225–231. https://doi.org/10.1016/j.scienta.2014.06.033

Romero, C., Pedryc, A., Muñoz, V., Llácer, G., & Badenes, M. L. (2003). Genetic diversity of different apricot geographical groups determined by SSR markers. *Genome*, *46*(2), 244–252. https://doi.org/10.1139/g02-128

de la Rosa, L. and Martín, I. (2016). “Las colecciones de germoplasma de variedades tradicionales,” in *Las variedades locales en la mejora genética de plantas*, eds J. I . Ruiz de Galarreta, J. Prohens and R. Tierno (San Sebastián: Servicio Central de Publicaciones del Gobierno Vasco), pp. 43-59.

Sánchez-Pérez, R., Ruiz, D., Dicenta, F., Egea, J., & Martínez-Gómez, P. (2005). Application of simple sequence repeat (SSR) markers in apricot breeding: Molecular characterization, protection, and genetic relationships. *Scientia Horticulturae*, *103*(3), 305–315. https://doi.org/10.1016/j.scienta.2004.06.009

Sosinski, B., Gannavarapu, M., Hager, L.D., Beck, L.E., King, G.J., et al. (2000) Characterization of microsatellite markers in peach [*Prunus persica* (L.) Batsch]. *Theor Appl Genet* 101, 421–42.

Vieira, M. L. C., Santini, L., Diniz, A. L., Munhoz, C. F. (2016). Microsatellite markers : what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3), 312–328.

Wünch, A., & Hormaza, J. I. (2002). Molecular characterization of sweet cherry (*Prunus avium* L.) genotypes using peach (*Prunus persica* (L.) Batsch) SSR sequences. *Heredity*, *89*(1), 56– 63. https://doi.org/10.1038/sj.hdy.6800101

Zhebentyayeva, T., Ledbetter, C., Burgos, L., & Llacer, G. (2012). “Apricot,” in *Fruit Breeding*, eds. M. Badenes and D. Byrne (Boston, MA: Springer USA).

7. Anexo

En este anexo se adjunta el código R contenido en el pipeline desarrollado en RMarkdown

```
library(adegenet)
library(PopGenReport)
library(hierfstat)
library(pegas)
library(DataCombine)
library(ade4)
library(poppr)
library(ggfortify)
library(phangorn)
library(dendextend)
library(factoextra)
library(reshape2)

# Crear un data frame con los datos

datos_ssr<-read.csv("DatosPEC2.csv", header = TRUE, sep = ";", na.strings =
TRUE)
dim(datos_ssr) # número de muestras y número de variables

## [1] 50 10

summary(datos_ssr)

##      Variedades     Origen                                    Programa_mejora
##  Almabar  : 1   España :20   Tradicional                                :15
##  Almater  : 1   Francia:15   Newcot (International Plant Selection): 6
##  Apribang : 1   USA    : 9   CEBAS-CSIC                                 : 4
##  Aprisweet: 1   Canada : 2   PBS Producción Vegetal SL                  : 4
##  Aprix 116: 1   Greece : 1   COT International                          : 3
##  Aprix 20 : 1   Italia : 1   Escande                                    : 3
##  (Other)  :44   (Other): 2   (Other)                                    :15
##    ssrPaCITA7  ssrPaCITA10  ssrPaCITA23  ssrPaCITA27    UDAp_415
##  191/212: 6   176/178: 8   140/150: 6   253/265:12   149/155: 7
##  206/212: 6   176/180: 6   144/147: 6   265/265:11   149/161: 7
##  212/212: 6   172/176: 4   147/150: 6   253/253: 8   161/171: 7
##  191/194: 5   172/178: 4   142/147: 4   228/265: 4   161/161: 6
##  194/194: 5   176/176: 4   140/147: 3   256/265: 4   155/155: 5
##  194/212: 5   174/176: 3   144/156: 3   228/253: 3   155/161: 5
##  (Other):17   (Other):21   (Other):22   (Other): 8   (Other):13
##     UDAp_420     pchgms3
##  173/181:10   197/197:17
##  158/181: 9   190/197:10
##  171/173: 8   195/197: 9
##  158/173: 5   195/195: 5
##  173/173: 4   187/197: 3
##  158/158: 3   197/200: 2
##  (Other):11   (Other): 4

ind <- as.character(datos_ssr$Variedades) # Variable con los nombres de las
cvs específicando tipo 'character'
poblacion_origen <- as.character(datos_ssr$Origen) # Variable con los
diferentes origenes específicando tipo 'character'. Niveles poblacionales:
países de origen de las variedades
poblacion_pg <- as.character(datos_ssr$Programa_mejora) # Variable con los
```

```
diferentes origenes especificando tipo 'character'. Niveles poblacionales:
programa de mejora

# La función df2genind convierte el data frame en un objeto genind (matriz de
0, 1 y 2) [Sin especificar población]
datos_ssr_genind <- df2genind(datos_ssr[,4:10], ploidy=2, sep="/",
ind.names=ind, loc.names = NULL)

dim(datos_ssr_genind@tab) # número de muestras y número de alelos totales

## [1] 50 52

summary<-summary(datos_ssr_genind)
```

OBJETIVO ESPECÍFICO 2. Caracterizar e identificar mediante marcadores microsatélites las variedades estudiadas.

2.1. Estudio de la validez de los primers utilizados mediante diferentes parámetros

```
# Comprobación polimorfismo
pol<-isPoly(datos_ssr_genind,by="loc", thres=0.1)

# Número total de alelos
num.al<-summary$loc.n.all

# Rango
range.min <- sapply(datos_ssr_genind@all.names, min)
range.max <- sapply(datos_ssr_genind@all.names, max)

# Tabla-resumen por locus
tabla1<-data.frame("Polimorfismo" = pol, "Número_de_alelos" = num.al,
"Rango_mínimo" = range.min, "Rango_máximo" = range.max)
tabla1

##             Polimorfismo Número_de_alelos Rango_mínimo Rango_máximo
## ssrPaCITA7          TRUE               10          189          224
## ssrPaCITA10         TRUE                9          155          180
## ssrPaCITA23         TRUE                9          140          156
## ssrPaCITA27         TRUE                6          228          267
## UDAp_415            TRUE                6          149          171
## UDAp_420            TRUE                6          158          181
## pchgms3             TRUE                6          180          200

# Por Origen

datos_ssr_origen <- df2genind(datos_ssr[,4:10], ploidy=2, sep="/",
ind.names=ind, pop=poblacion_origen, loc.names = NULL) # Objeto genind
especificando nivel poblacional
summary_origen<-summary(datos_ssr_origen)

# Tabla-resumen por países
ngen<-summary_origen$n.by.pop
nall<-summary_origen$pop.n.all
```

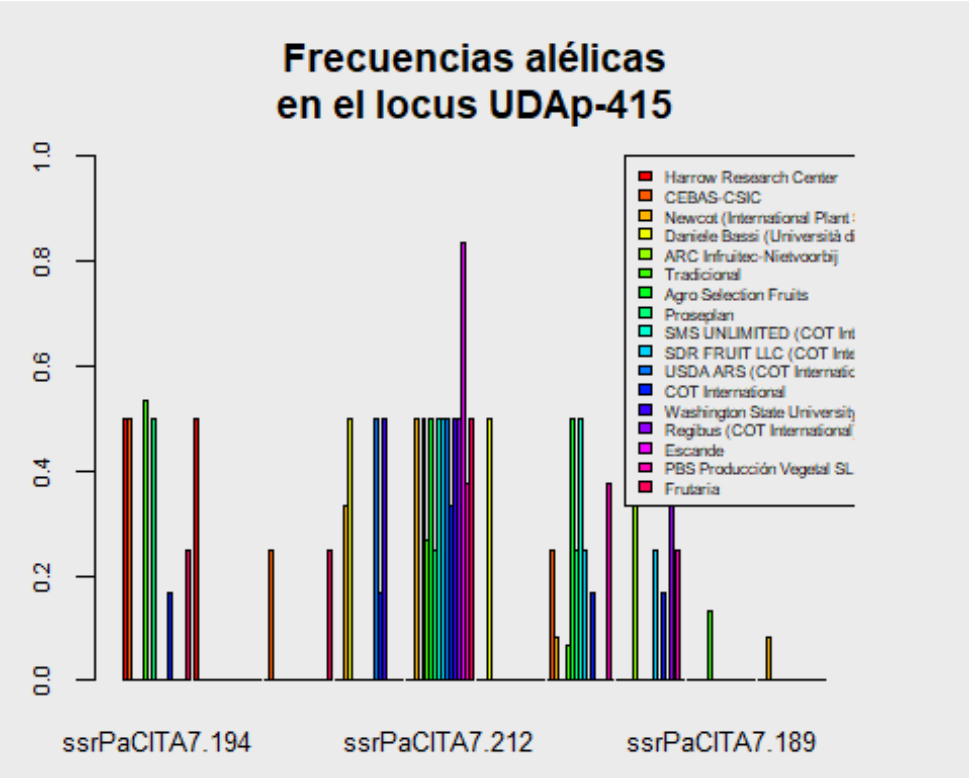
```r
tabla2<-data.frame("Número_de_genotipos" = ngen, "Número_de_alelos" = nall)
tabla2
```

```
##           Número_de_genotipos Número_de_alelos
## Canada                      2               18
## España                     20               37
## Francia                    15               37
## Italia                      1               13
## Sudafrica                   1               13
## USA                         9               37
## Suiza                       1               11
## Greece                      1               11
```

```r
# Por Programa de mejora

datos_ssr_pg <- df2genind(datos_ssr[,4:10], ploidy=2, sep="/", ind.names=ind,
pop=poblacion_pg, loc.names = NULL) # Objeto genind especificando nivel
poblacional
summary_pg<-summary(datos_ssr_pg)

# Tabla-resumen por países
ngen<-summary_pg$n.by.pop
nall<-summary_pg$pop.n.all

tabla3<-data.frame("Número_de_genotipos" = ngen, "Número_de_alelos" = nall)
tabla3
```

```
##                                                 Número_de_genotipos
## Harrow Research Center                                            1
## CEBAS-CSIC                                                        4
## Newcot (International Plant Selection)                            6
## Daniele Bassi (Università di Bologna)                             1
## ARC Infruitec-Nietvoorbij                                         1
## Tradicional                                                      15
## Agro Selection Fruits                                             2
## Proseplan                                                         2
## SMS UNLIMITED (COT International)                                 1
## SDR FRUIT LLC (COT International)                                 2
## USDA ARS (COT International)                                      1
## COT International                                                 3
## Washington State University (COT International)                   1
## Regibus (COT International)                                       1
## Escande                                                           3
## PBS Producción Vegetal SL                                         4
## Frutaria                                                          2
##                                                 Número_de_alelos
## Harrow Research Center                                        11
## CEBAS-CSIC                                                    23
## Newcot (International Plant Selection)                        28
## Daniele Bassi (Università di Bologna)                         13
## ARC Infruitec-Nietvoorbij                                     13
## Tradicional                                                   36
## Agro Selection Fruits                                         15
## Proseplan                                                     17
## SMS UNLIMITED (COT International)                             13
## SDR FRUIT LLC (COT International)                             20
## USDA ARS (COT International)                                  13
## COT International                                             27
## Washington State University (COT International)               13
## Regibus (COT International)                                   11
```

```
## Escande                                                    16
## PBS Producción Vegetal SL                                  23
## Frutaria                                                   16

# Frecuencia alélica [Por programa de mejora]
datos_ssr_genpop<-genind2genpop(datos_ssr_pg)

##
##  Converting data from a genind to a genpop object...
##
## ...done.

AllFreq <- makefreq(datos_ssr_genpop)

##
##  Finding allelic frequencies from a genpop object...
##
## ...done.

dffreq7<-as.matrix(data.frame(AllFreq[,1:10]))

par(bg="gray92", mar=c(2.5, 2.5, 4, 3))
barplot(dffreq7, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus UDAp-
415", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
legend("topright", inset=c(-0.2,0),
legend=rownames(dffreq7),fill=rainbow(17),cex=0.5, pt.cex=3)
```




```
dffreq10<-as.matrix(data.frame(AllFreq[,11:19]))
dffreq23<-as.matrix(data.frame(AllFreq[,20:28]))
dffreq27<-as.matrix(data.frame(AllFreq[,29:34]))
dffreq415<-as.matrix(data.frame(AllFreq[,35:40]))
dffreq420<-as.matrix(data.frame(AllFreq[,41:46]))
dffreqA<-as.matrix(data.frame(AllFreq[,47:52]))
```

```
par(mfrow=c(3,2))
P10<-barplot(dffreq10, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
ssrPaCITA10", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
P23<-barplot(dffreq23, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
ssrPaCITA23", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
P27<-barplot(dffreq27, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
ssrPaCITA27", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
P415<-barplot(dffreq415, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
UDAp_415", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
P420<-barplot(dffreq420, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
UDAp_420", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
PA<-barplot(dffreqA, beside=T, xlab="Alelos", ylab="Frecuencia alélica",
ylim=c(0,1), col=rainbow(17), main="Frecuencias alélicas\nen el locus
pchgms3", cex=0.8, cex.main=1.2, cex.axis=0.7, cex.lab=0.85)
```

2.2. Estudio de la diversidad genética de las variedades analizadas mediante diferentes parámetros

```
# Heterocigosidad

bartlett.test(list(summary$Hexp, summary$Hobs)) #Test Ho: Hexp = Hobs

##
##  Bartlett test of homogeneity of variances
##
## data:  list(summary$Hexp, summary$Hobs)
## Bartlett's K-squared = 1.8109, df = 1, p-value = 0.1784
```

```
HWEtest<-hw.test(datos_ssr_pg, B = 1000) # Test Ho: el locus está en
equilibrio HW en la población. B = número de réplicas para el procedimiento
Monte Carlo

# Tabla-resumen de heterocigosidad
tabla4 <- data.frame(Hobs = summary$Hobs, Hexp = summary$Hexp, p_valor =
HWEtest[,4]) # Tabla locus name | Hobs | Hexp | p-valor para HWE
round(tabla4, digits=2)

##             Hobs Hexp p_valor
## ssrPaCITA7  0.76 0.77    0.06
## ssrPaCITA10 0.82 0.79    0.82
## ssrPaCITA23 0.88 0.82    0.51
## ssrPaCITA27 0.54 0.68    0.05
## UDAp_415    0.78 0.75    0.35
## UDAp_420    0.84 0.76    0.12
## pchgms3     0.56 0.60    0.14

# Estadísticos de Wright

datos_ssr_hierf<-genind2hierfstat(datos_ssr_pg)
resumen<-basic.stats(datos_ssr_hierf)

# Tabla-resumen de estadísticos de Wright
tabla5 <- data.frame(FIS = resumen$perloc$Fis, FST = resumen$perloc$Fst,
row.names = locNames(datos_ssr_pg)) # FIS: Coeficiente de endogamia, FST:
Índice de fijación
tabla5 <- InsertRow(tabla5, NewRow = colMeans(tabla5[ ,1:2]))
round(tabla5, digits=2)

##               FIS  FST
## ssrPaCITA7  -0.19 0.07
## ssrPaCITA10 -0.27 0.15
## ssrPaCITA23 -0.25 0.10
## ssrPaCITA27  0.03 0.22
## UDAp_415    -0.19 0.12
## UDAp_420    -0.19 0.06
## pchgms3     -0.28 0.18
## 8           -0.19 0.13
```

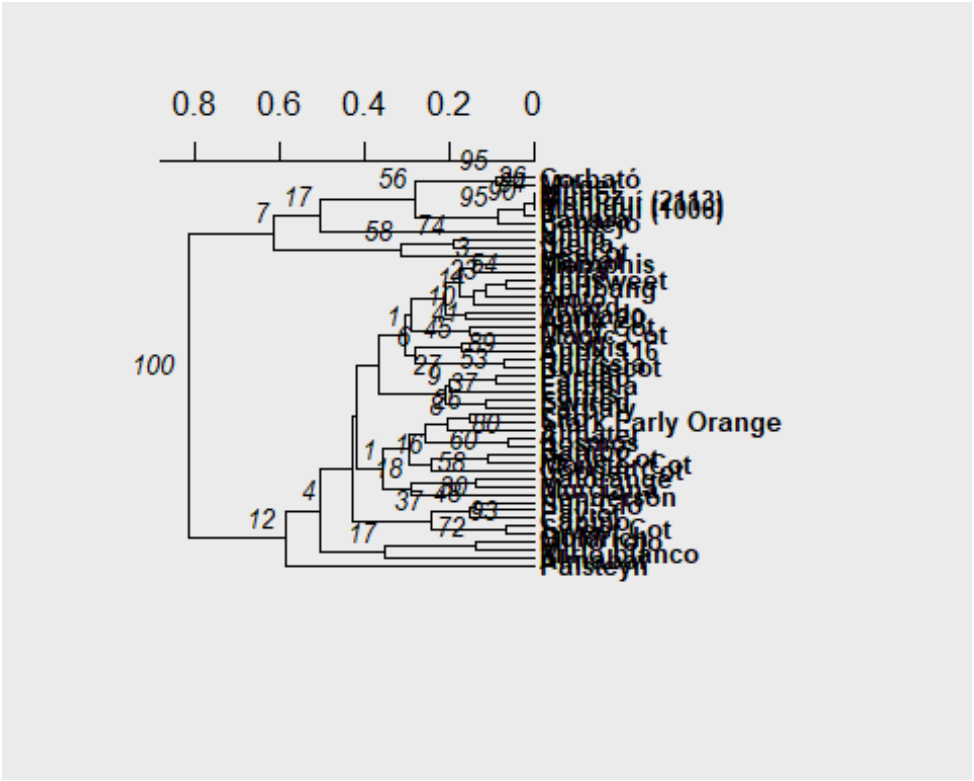
OBJETIVO ESPECÍFICO 3. Comprobar la presencia de homonimias y/o sinonimias en el material evaluado

```
# Homonimias
homon<-which(duplicated(datos_ssr[,1]))
lista = data.frame(datos_ssr['Variedades'])
nombre_homon = lista[homon,]
nombre_homon

## factor(0)
## 50 Levels: Almabar Almater Apribang Aprisweet Aprix 116 ... Veecot

# Sinonimias
sinon1<-duplicated(datos_ssr[,4:10]);
sinon2<-duplicated(datos_ssr[,4:10], fromLast=TRUE);
sinon <- sinon1 | sinon2
sinon<-which(sinon)
nombre_sinon = lista[sinon,]
nombre_sinon
```

```
## [1] Moniquí (2113) Muñoz          Moniquí (1006)
## 50 Levels: Almabar Almater Apribang Aprisweet Aprix 116 ... Veecot
```

OBJETIVO ESPECÍFICO 4. Estudiar filogenéticamente las diferentes variedades y establecer las relaciones de parentesco entre ellas

```
#Árbol UPGMA mediante distancia genética de Nei
set.seed(9999)
par(bg="gray92")
aboot(datos_ssr_genind, dist = nei.dist, sample=100, tree="upgma")

##
Running bootstraps:       100 / 100
## Calculating bootstrap values... done.
```




```
##
## Phylogenetic tree with 50 tips and 49 internal nodes.
##
## Tip labels:
##  Harval, Mirlo blanco, Mirlo rojo, Farius, Farbaly, Farbela, ...
## Node labels:
##  100, 12, 7, 58, 17, 4, ...
##
## Rooted; includes branch lengths.
```

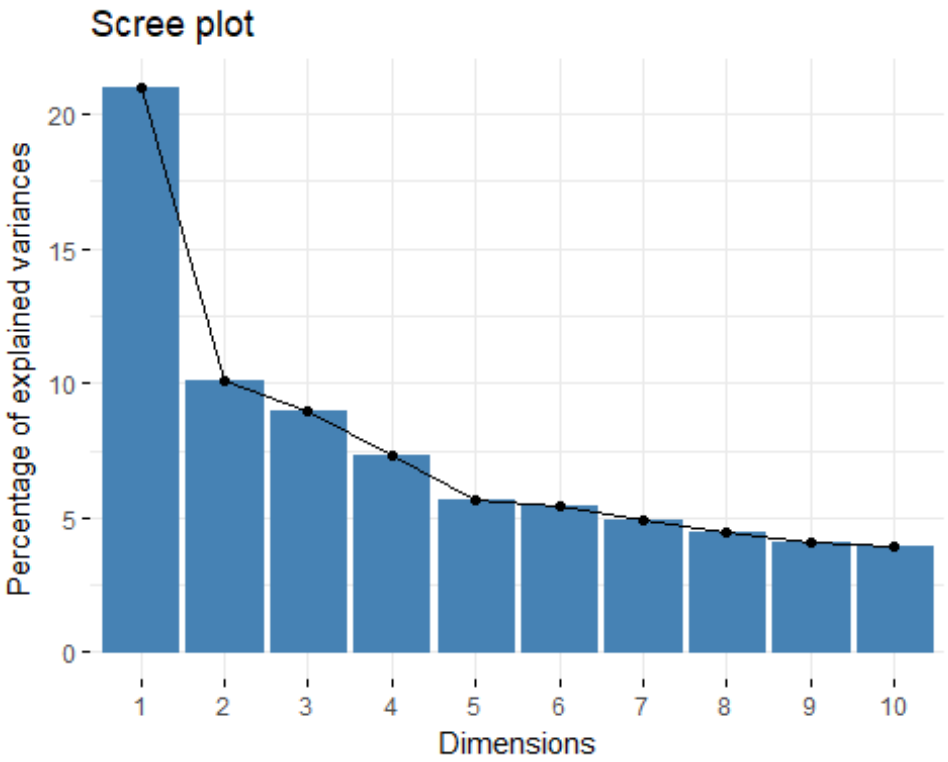
OBJETIVO ESPECÍFICO 5. Representar la distribución de la diversidad genética obtenida a partir de la amplificación con marcadores microsatélites y estudiar la estructura de dichas poblaciones.

AMOVA

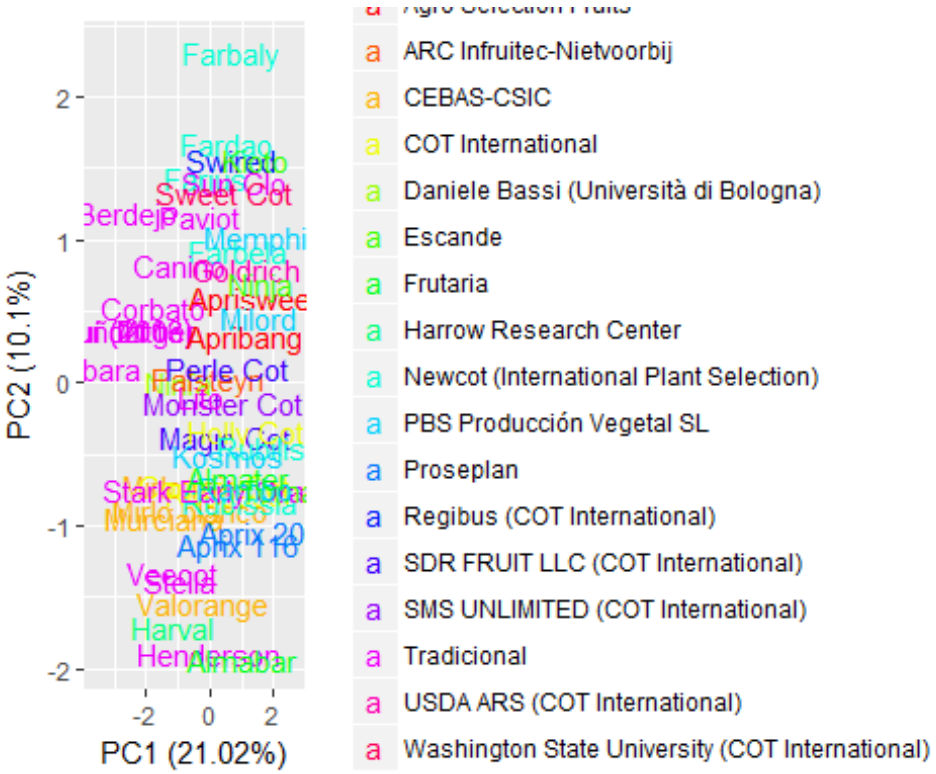
```
# Para estimar la diversidad molecular poblacional se realizó un AMOVA
(Análisis Molecular de la Varianza)
strata(datos_ssr_genind)<-data.frame(datos_ssr[,2:3])
calc_amova<-poppr.amova(datos_ssr_genind, ~Origen/Programa_mejora,
within=TRUE)

##
##  No missing values detected.

calc_amova

## $call
## ade4::amova(samples = xtab, distances = xdist, structures = xstruct)
##
## $results
##                                         Df    Sum Sq  Mean Sq
## Between Origen                           7  58.95556 8.422222
## Between Programa_mejora Within Origen   13  91.61944 7.047650
## Between samples Within Programa_mejora 29 108.12500 3.728448
## Within samples                         50 259.00000 5.180000
## Total                                  99 517.70000 5.229293
##
## $componentsofcovariance
##                                                          Sigma         %
## Variations  Between Origen                            0.2269398   4.25871
## Variations  Between Programa_mejora Within Origen     0.6476760  12.15417
## Variations  Between samples Within Programa_mejora -0.7257759 -13.61977
## Variations  Within samples                            5.1800000  97.20690
## Total variations                                      5.3288400 100.00000
##
## $statphi
##                                     Phi
## Phi-samples-total            0.02793104
## Phi-samples-Programa_mejora -0.16294103
## Phi-Programa_mejora-Origen   0.12694800
## Phi-Origen-total             0.04258710
```

PCA

```
# Gráfico PCA para observar la distribución de las variedades en función del
programa de mejora (Por programa de mejora)

pca <- dudi.pca(datos_ssr_pg, scann=FALSE) # Se crea un gráfico PCA
val <- 100 * pca$eig/sum(pca$eig) # Se calcula el % de valores eigenvalue
(vectores propios)

pca <- dudi.pca(datos_ssr_pg, cent = TRUE, scale = FALSE, scann=FALSE, nf=20)
# Vuelvo a crear un gráfico PCA en el que selecciono 20 en number of axes
(componentes)
fviz_eig(pca) # Gráfico que muestra el porcentage de varianza explicada por
cada componente principal.
```



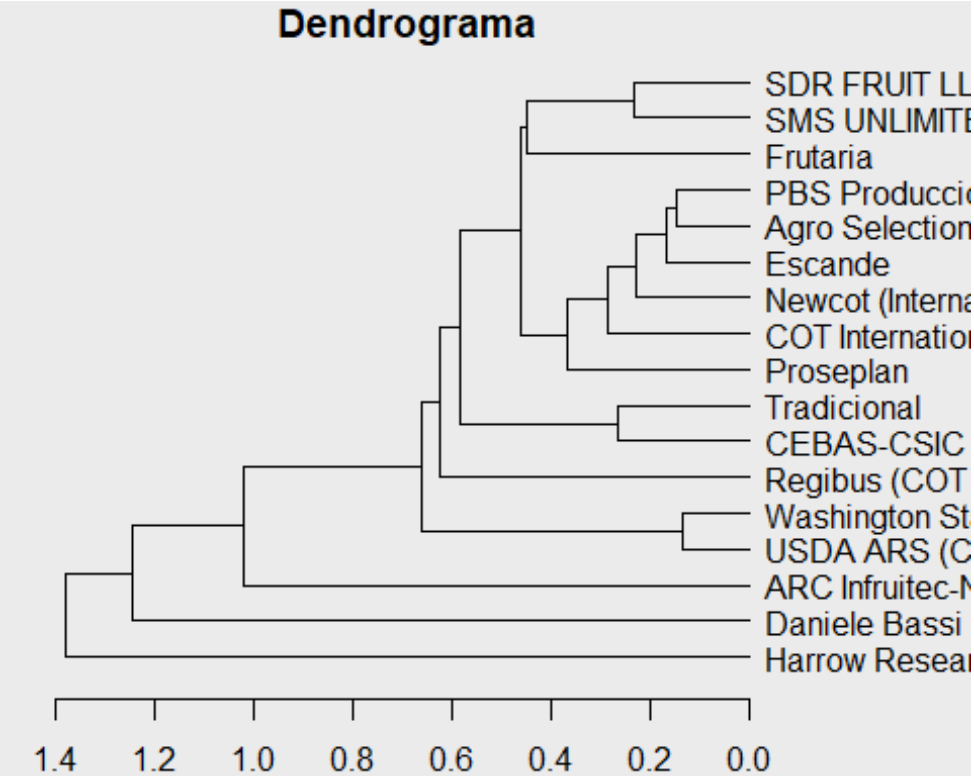

```
PCA <- autoplot(prcomp(datos_ssr_genind), shape = FALSE, data = datos_ssr,
colour = 'Programa_mejora', scale = 0) # Gráfico PCA
PCA + scale_fill_manual(values=rainbow(17)) +
scale_color_manual(values=rainbow(17)) # Por programa de mejora
```



UPGMA

```
# Observar la distancia genética entre poblaciones
distgenpop <- dist.genpop(datos_ssr_genpop,method=1) # Method 1 = distancia
```

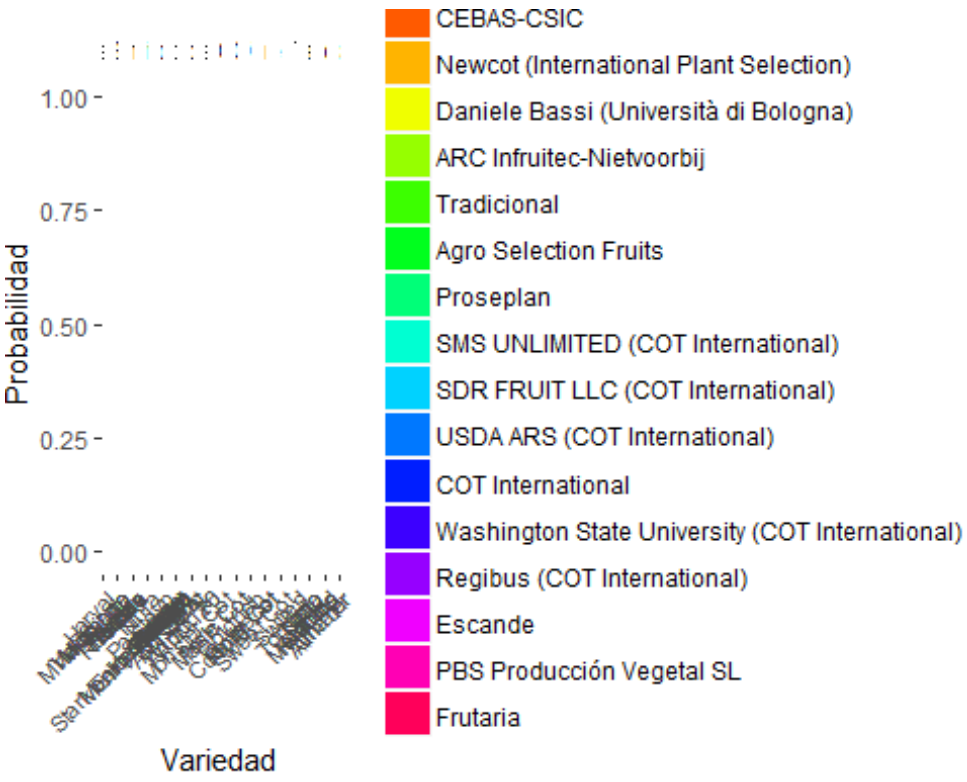
```
de Nei

# Árbol UPGMA mediante la distancia de Nei
dend2<-as.dendrogram(hclust(distgenpop, method = "average"))  #UPGMA

par(bg="gray92", mar=c(2,1,1,5))
plot(dend2, cex = 0.8, horiz = TRUE, main="Dendrograma")
```



Elección de clusters (k) + DAPC

```
nclust<-find.clusters(datos_ssr_pg, stat="BIC", max.n.clust = 40) # La función muestra un gráfico de varianza acumulada explicada por los valores propios de la PCA. Se indica un mayor número de componentes principales para que retenga a todos [number PCs to retain (>= 1): 50] y se elige el número de clusters idóneo [number of clusters (>=2: 6]

dapcgraf<-dapc(datos_ssr_pg, n.da=50, n.pca=6) # Se crea el DAPC con los datos obtenidos en la función anterior (número de PC y de clusters)

dapc.results <- as.data.frame(dapcgraf$posterior)
dapc.results$pop <- pop(datos_ssr_pg)
dapc.results$indNames <- rownames(dapc.results)
dapc.results <- melt(dapc.results)

## Using pop, indNames as id variables

colnames(dapc.results) <-
c("Programa_mejora","Variedad","Poblacion_asignada","Probabilidad")

dapc <- ggplot(dapc.results, aes(x=Variedad, y=Probabilidad,
fill=Poblacion_asignada))
dapc <- dapc + geom_bar(stat='identity')
dapc <- dapc + scale_fill_manual(values = rainbow(17))
dapc <- dapc + facet_grid(~Programa_mejora, scales = "free_x")
dapc <- dapc + theme(axis.text.x = element_text(angle = 45, hjust = 1, size =
```

```
8))
dapc
```


1.00
0.75
0.50
0.25
0.00
Probabilidad
Variedad
CEBAS-CSIC
Newcot (International Plant Selection)
Daniele Bassi (Università di Bologna)
ARC Infruitec-Nietvoorbij
Tradicional
Agro Selection Fruits
Proseplan
SMS UNLIMITED (COT International)
SDR FRUIT LLC (COT International)
USDA ARS (COT International)
COT International
Washington State University (COT International)
Regibus (COT International)
Escande
PBS Producción Vegetal SL
Frutaria