UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Análisis de Variantes para la Identificación del Origen de las Variedades Tradicionales de Tomate "De Penjar" o "Da Serbo"

Trabajo Fin de Grado

Grado en Biotecnología

AUTOR/A: Bellot Pastor, Lluna

Tutor/a: Forment Millet, José Javier

Director/a Experimental: Bombarely Gomez, Aureliano

CURSO ACADÉMICO: 2023/2024

ANÁLISIS DE VARIANTES PARA LA IDENTIFICACIÓN DEL ORIGEN DE LAS VARIEDADES TRADICIONALES DE TOMATE "DE PENJAR" O "DA SERBO"

RESUMEN

Los tomates "De penjar" de España, también conocidos como "Da serbo" en Italia, son comúnmente utilizados en el ámbito gastronómico de estos países. Dos ejemplos son la elaboración del famoso "pa en tomaca" y el tomate añadido a diferentes tipos de pizzas napolitanas. Los agricultores han seleccionado caracteres agronómicos adaptados a su cultivo en la cuenca mediterránea y su cultura, como son su capacidad para crecer en climas secos y su vida postcosecha extendida. Estos caracteres agronómicos han despertado el interés de genetistas y mejoradores vegetales para el desarrollo de nuevas variedades con una mayor resiliencia al cambio climático.

El origen de estas variedades de tomate es desconocido. Aunque históricamente hablando se sabe que el tomate fue introducido a Europa por España en el siglo XVI, todavía existen algunas lagunas desde un punto de vista genético. El objetivo de esta investigación es tratar de identificar el origen de estas variedades de tomate, así como determinar la relación entre los tomates "De penjar" de ambos países. La hipótesis de partida es que las variedades de ambos países poseen una diferenciación genética.

Con el objetivo de realizar el análisis usaremos datos públicos de secuenciación de representación reducida (Genotyping-By-Sequencing) de 1058 muestras de diferentes colecciones de tomates del mundo. Al partir de lecturas procesadas, comenzamos alineando nuestras lecturas y haciendo un llamamiento de variantes. A partir de esto obtendremos unos SNPs que nos servirán para dar una repuesta a nuestra hipótesis mediante diferentes herramientas como PCA, DAPC o F_{ST}.

Los resultados no apoyan nuestra hipótesis de partida. Tanto el PCA como el DAPC muestran que no existen grupos diferenciados entre ambas muestras. Dicho resultado se corrobora con los valores de F_{ST} entre las diferentes poblaciones estudiadas. La interpretación de los resultados ratifica resultados previos en publicaciones científicas que revelaban la baja variabilidad genética entre las variedades tradicionales europeas. No obstante, quedaría aclarar si esta baja variabilidad en estas variedades se debe a una homogeneización genética de variedades tradicionales como producto del comercio, o como consecuencia de la falta de presión evolutiva presente en estas variedades.

PALABRAS CLAVE: Tomate, GBS, Variabilidad Genética, PCA, Poblaciones, F_{ST}

ANÀLISI DE VARIANTS PER A LA IDENTIFICACIÓ DE L'ORIGEN DE LES VARIETATS TRADICIONALS DE TOMAQUES "DE PENJAR" O "DA SERBO"

RESUM

Les tomaques "De penjar" d'Espanya, també coneguts com "Da serbo" a Itàlia, són comunament utilitzats en l'àmbit gastronòmic d'aquests països. Dos bons exemples són l'elaboració del famós "pa amb tomaca" i la tomaca afegida a diferents tipus de pizzes napolitanes. Els agricultors han seleccionat caràcters agronòmics adaptats al seu cultiu en la conca mediterrània i la seua cultura, com són la seua capacitat per a créixer en climes secs i una vida postcollita estesa. Aquests caràcters agronòmics han despertat l'interés de genetistes i milloradors vegetals per al desenvolupament de noves varietats amb una major resiliència al canvi climàtic.

L'origen d'aquestes varietats de tomaca és desconegut. Encara que històricament parlant se sap que la tomaca va ser introduïda a Europa per Espanya en el segle XVI, encara existeixen algunes llacunes des d'un punt de vista genètic. L'objectiu d'aquesta investigació és tractar d'identificar l'origen d'aquestes varietats de tomaca, així com determinar la relació entre les tomaques "De penjar" dels dos països. La hipòtesi de partida és que les varietats dels dos països posseeixen una diferenciació genètica.

Amb l'objectiu de fer l'anàlisi usarem dades públiques de seqüenciació de representació reduïda (Genotyping-By-Sequencing) de 1058 mostres de diferents col·leccions de tomaques del món. Com partim de lectures processades, comencem alineant les nostres lectures i fent una crida de variants. A partir d'això obtindrem uns SNPs que ens serviran per a donar una resposta a la nostra hipòtesi mitjançant diferents eines com PCA, DAPC o F_{ST}.

Els resultats no donen suport a la nostra hipòtesi de partida. Tant el PCA com el DAPC mostren que no existeixen grups diferenciats entre les mostres. Aquest resultat es corrobora amb els valors de F_{ST} entre les diferents poblacions estudiades. La interpretació dels resultats ratifica resultats previs en publicacions científiques que revelaven la baixa variabilitat genètica entre les varietats tradicionals europees. No obstant això, quedaria aclarir si aquesta baixa variabilitat en aquestes varietats es deu a una homogeneïtzació genètica de varietats tradicionals com a producte del comerç, o a conseqüència de la falta de pressió evolutiva present en aquestes varietats.

PARAULES CLAU: Tomaques, GBS, Variabilitat Genètica, PCA, Poblacions, F_{ST}

ANALYSIS OF VARIANTS FOR IDENTIFICATION OF THE ORIGIN OF "DE PENJAR" OR "DA SERBO" TOMATO TRADITIONAL VARIETIES

ABSTRACT

The "De penjar" tomatoes from Spain, also known as "Da serbo" in Italy, are widely utilized in the gastronomic sphere of these countries. Two notable examples include their use in the renowned "pa en tomaca" dish and their incorporation into various types of Neapolitan pizzas. Farmers have selectively bred these tomatoes to exhibit agronomic characteristics suited to their cultivation in the Mediterranean basin, such as resilience to dry climates and an extended post-harvest lifespan. These agronomic traits have piqued the interest of geneticists and plant breeders for the development of new varieties with enhanced resilience to climate change.

The precise origins of these tomato varieties remain unknown. While historical records indicate that tomatoes were introduced to Europe by Spain in the 16th century, there are still genetic gaps to be filled. This research aims to discover the origin of these tomato varieties and study the relationship between the "De penjar" tomatoes of Spain and Italy. The initial hypothesis posits that the varieties from both countries exhibit genetic differentiation.

To conduct our analysis, we will leverage publicly available Genotyping-By-Sequencing data comprising 1058 samples from diverse tomato collections worldwide. Starting with processed reads, we will proceed by aligning our reads and identifying variants. These variants will enable us to address our hypothesis using various analytical tools such as PCA, DAPC, and F_{ST}.

However, the results obtained do not align with our initial hypothesis. Both PCA and DAPC analyses reveal no discernible differentiation between the two sample sets. This finding is further supported by the F_{ST} values calculated between the different populations under study. The interpretation of these results reinforces previous findings documented in scientific literature, which highlighted the low genetic variability observed among traditional European tomato varieties. Nonetheless, it remains to be elucidated whether this low variability stems from genetic homogenization resulting from trade or is a consequence of the absence of evolutionary pressure exerted on these varieties.

KEYWORDS: Tomato, GBS, Genetic Variability, PCA, Populations, F_{ST}

AGRADECIMIENTOS

Quiero dedicar esta investigación a mi familia, en especial a mis padres, puesto que sin ellos no podría haber estudiado esta carrera y haber llegado hasta aquí. También quiero agradecer a mis compañeros y amigos por acompañarme en este proceso, apoyándome. Por último, quiero agradecer a mi director experimental, Aureliano Bombarely, por guiarme durante todo este proceso y a mi tutor, José Javier Forment, por brindarme la oportunidad.

ÍNDICE

OBJETIVOS DEL DESARROLLO SOSTENIBLE

	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza				
ODS 2. Hambre cero				
ODS 3. Salud y bienestar				
ODS 4. Educación de calidad				
ODS 5. Igualdad de género				
ODS 6. Agua limpia y saneamiento				
ODS 7. Energía asequible y no contaminante				
ODS 8. Trabajo decente y crecimiento económico				
ODS 9. Industria, innovación e infraestructuras				
ODS 10. Reducción de las desigualdades				
ODS 11. Ciudades y comunidades sostenibles				
ODS 12. Producción y consumo responsables				
ODS 13. Acción por el clima				
ODS 14. Vida submarina				
ODS 15. Vida de ecosistemas terrestres				
ODS 16. Paz, justicia e instituciones sólidas				
ODS 17. Alianzas para lograr objetivos.				

2 Hambre cero: El estudio de estas variedades de tomate puede brindar nuevas características genéticas interesantes para la producción de tomates en climas más secos. Estos climas están en aumento debido al cambio climático. Además, son tomates de larga duración por lo que permitiría disponer de ellos durante diferentes épocas del año. También al tener estas características facilitan su exportación pudiendo llegar a lugares donde en principio no se dispone de ellos.

4 Educación de calidad: La realización de este TFG puede servir para futuros estudios. Al igual que yo me he servido de información de otros trabajos y artículos para la realización del mismo, ahora puede aportar información bibliográfica a la comunidad científica.

6 Agua limpia y saneamiento: Al ser tomates que no necesitan climas húmedos, si no más bien secos favorece un consumo menor de agua en su cultivo.

12 Producción y consumo responsables: Al ser tomates de larga duración permiten un mayor aprovechamiento de ellos, reduciendo la perdida de producto y favoreciendo el consumo responsable. Su producción en climas secos también favorece una producción responsable en diferentes zonas, como el clima mediterráneo.

1. INTRODUCCIÓN

El tomate (*Solanum lycopersicum L.*) es uno de los cultivos más populares del mundo. Este cultivo presenta miles de variedades. Es una planta herbácea que pertenece al grupo de las solanáceas (Ochogavía et al. 2011). Dentro de las solanáceas encontramos más de 3000 especies diferentes muchas de ellas domesticadas y usadas como cultivos. Buenos ejemplos son las patatas, las berenjenas, o los pimientos (Bergougnoux, 2014). El cultivo del tomate se originó en América, ocurriendo una primera domesticación en la región de los Andes y que continuó en Mesoamérica (Blanca et al. 2012). Una de las características que presentaron los tomates tras la domesticación es la presencia de un tamaño mayor, siendo los tomates de menor tamaño más cercanos a las primeras variedades salvajes.

La forma en que este fruto fue introducido a Europa sigue dos teorías, la peruana o la mexicana. Esta última es la más aceptada y consiste en explicar su llegada a Europa a través de España donde rápidamente se distribuiría a Nápoles en Italia (Blanca et al. 2012). De acuerdo con algunos datos históricos, los tomates fueron traídos a España en el siglo XVI por Hernán Cortes tras la captura de Tenochtitlán, ciudad de la época azteca que hoy conocemos como Ciudad de México (Bergougnoux, 2014).

Desde la domesticación del tomate, se han desarrollado multitud de variedades adaptadas a las condiciones y preferencias culturales de la región donde se cultivan. Algunos ejemplos de variedades españolas locales son el tomate "Valenciano", de "Monserrat", "Pera de Girona", "Canario" y "Muchamiel". Ejemplos de variedades populares en otros países son: "Marmande" (Francia), "San Marzano" (Italia), "Krasati" (Grecia) y "Moneymaker" (Estados Unidos). Los tomates "De penjar", conocimos también como tomates "de colgar" o "de Ramellet" son una variedad de tomate muy popular en el levante español (Figàs et al. 2018). En Italia este tipo de tomates se conocen como "Da serbo" o "invernale" (Sacco et al. 2020). En referencia a su consumo, es característico del noreste de España, es decir, en las Islas Baleares, Cataluña y en la Comunidad Valenciana (Casals et al. 2011). En Italia la principal área de consumo es el centro y el sur peninsular (Sacco et al. 2020). En general, esta variedad de tomate se extiende alrededor del mediterráneo (Siracusa et al. 2012).

El uso de esta variedad se destina al consumo en fresco, para la cocina tradicional o para desecarlo. Uno de los platos más característicos en la que se emplea esta variedad de tomate es el plato tradicional conocido como "pa amb tomaca" en España (Ochogavía et al. 2011). Las características de esta variedad la hacen adecuada para esta receta puesto que se necesita frotar el tomate contra al pan y su piel resistente facilita esta acción. En Italia los podemos encontrar en las pizzas Napolitanas.

Las principales características de las variedades en ambos países se ven representadas en la Tabla 1. En ambos casos vemos características similares, aunque si podemos destacar la diferencia respecto a la piel.

Tabla 1. Comparación de algunas características de esta variedad de tomate. Las características se han obtenido de Ochogavía (2011), Figàs (2018), Sacco (2020) y Figàs (2019).

	ESPAÑA	**ITALIA**
Morfología	Principalmente redondeada	Principalmente redondeada y elongada
Tamaño y peso	Fruto pequeño 36 a 86 gramos	Fruto pequeño de menos de 25 gramos
Durabilidad	Larga vida	Larga vida
Piel	Piel fina	Piel gruesa e incluso crujiente
Color	Fresco: transparente o amarillo Tras la conservación: rosado amarillento o anaranjado	Fresco: bermellón Tras la conservación: rojo oscuro
Pulpa	Abundante	Compacta con poca agua
Sabor	Intenso, dulce y algo ácido	Intenso, dulce y agrio

1.1.Morfología

Los tomates "De penjar" presentan una gran variedad morfológica. Esto es indicativo de una gran variabilidad genética asociada que principalmente se debe a caracteres que regulan la formación del fruto (Ochogavía et al. 2011). La Figura 1 resume parte de esta diversidad. En ésta puede apreciarse diversos tamaños (más pequeños o grandes), y formas (esféricas, más alargadas, y otras con formas irregulares).

Figura 1. Diversidad morfológica y de tamaño de los tomates "De penjar" (Ochogavía et al. 2011)

1.3. Sabor y aroma

El sabor y aroma característico de esta variedad de tomate se va desarrollando durante su conservación (Figàs et al. 2018). Diferentes compuestos dan este sabor y aroma, entre los cuales podemos encontrar materia seca, contenido de sólidos solubles, β-caroteno, ácido ascórbico… (Figàs, 2019). Un ejemplo de estos compuestos con sus sabores asociados se ve en la Tabla 2, incluyendo principalmente aquellos que son aceptados por el consumidor (Kaur et al. 2023). Además, contienen una alta presencia de grupos antioxidantes que favorecen la larga duración de estos tomates.

Tabla 2. Diferentes compuestos de los tomates con el sabor asociado a cada uno de ellos. Es una versión reducida, solo con los sabores mayormente aceptados por el consumidor de la investigación de Kaur *et al.* (2023).

Compuestos volátiles derivados de carotenoides		**Compuestos volátiles derivados de fenilpropanoides**	
Compuesto	**Sabor**	**Compuesto**	**Sabor**
6-metil-5-hepten-2-ona	afrutado, floral	cianuro de bencilo	verde, floral
Geranil acetona	dulce, floral	fenilacetaldehído	floral, alcohol
Lonona	afrutado, floral	feniletanol	nuez, afrutado
Compuestos volátiles derivados de ácidos grasos		**Compuestos volátiles derivados de AA de cadena ramificada**	
Compuesto	**Sabor**	**Compuesto**	**Sabor**
1-Penten	afrutado, floral, verde, herbáceo	metilbutanal	maltoso
E-2-penten	herbáceo	metilbutanol	maltoso, rama, terroso
3-pentanona	frutos secos	isobutiltiazol	rama, verde
Hexenal	verde	isovaleronitirilo	cebollado, disolvente

1.2. Frutos de larga duración

La principal característica de los tomates “De penjar” es su larga vida postcosecha. Un estudio realizado en el artículo de Ochogavía (2011), nos da entender que la vida media de este tomate es de 5 meses. La zona mediterránea se caracteriza por un ambiente seco solo en determinados meses del año, que es cuando se considera un clima favorable para el cultivo de este tipo de tomate. Al ser tomates que con una vida postcosecha más larga que otras variedades, permite disponer de ellos durante todo el año, aunque no se presente el clima requerido para su cultivo.

Los tomates presentan esta larga duración debido a una mutación en el gen *alc (alcobaça)* (Mutschler, 1984). Esta mutación se caracteriza por el cambio de una timina por una adenina en la posición 317 en el dominio NAC de la proteína que codifica el gen *alc*. Al cambiar estas bases, se ve modificado uno de los aminoácidos resultantes al traducir la secuencia, sustituyendo una valina por un ácido aspártico, en el aminoácido número 106. Es una característica importante

puesto que se sustituye un aminoácido neutro por uno cargado. También se hipotetiza que su larga vida puede verse asociada a su gran actividad antioxidante. No obstante, la presencia de esta mutación también tiene algunos efectos adversos como pueden ser la reducción de la producción de licopeno (Figàs, 2019) o una reducción del tamaño del fruto (Casals et al. 2011).

1.4. Características de la planta

Las principales propiedades que caracterizan a las tomateras de estas variedades quedan ejemplificadas por los tomates "de Ramellet" característicos de las Islas Baleares. Las raíces de las plantas suelen encontrarse en los primeros 50 centímetros de suelo, aunque ante una mayor sequedad pueden encontrarse más profundas. Las plantas suelen tener un tamaño mediano, con pelos glandulares recubriendo tanto sus tallos como sus hojas. En cuanto a la forma, suele ser la de una tomatera típica o puede ser de tipo trepadora, con ramificaciones más largas, pero en menor número (Ochogavía et al. 2011).

El lugar más representativo para el cultivo de este tomate en España es Alcalá de Xiver, en Castellón (Figàs et al. 2018), mientras que en Italia destaca la región de Campania, donde encontramos ciudades como Nápoles o Salerno (Sacco et al. 2020). Además, también los encontramos en las Islas Baleares y en Sicilia. En ambos orígenes tenemos el clima mediterráneo que se necesita para el correcto desarrollo de esta variedad de tomate. Una diferencia sería el periodo de cultivo, puesto que en España es de mayo a octubre (Casals et al. 2011) mientras que en Italia se cultiva de abril a julio (Siracusa et al. 2012). En ambos casos continuará su presencia en el mercado durante el invierno debido a su gran durabilidad, como se ha expuesto anteriormente (Figàs et al. 2018).

Un requisito importante a la hora de cultivar este tipo de tomate es la presencia de poca agua y que tenga una salinidad adecuada. Su cultivo debe ser al aire libre puesto a que necesita grandes cantidades de sol (Sacco et al. 2020), no obstante, el cultivo de estos tomates en invernadero se encuentra en crecimiento debido a su alta demanda (Figàs, 2019).

Dentro de la variedad de tomate que estudiamos encontramos dos zonas representativas, una en Italia y otra en España (Figura 2). En Italia se destaca el parque nacional del monte Vesubio. Dependiendo del área de producción tenemos diferentes nombres para este tipo de tomates: "piennolo" o "spungilli", más allá de "invernale" o "Da serbo". El tomate más representativo es el conocido como "Pomodorino del Piennolo Vesuvio". Este fue incluido en Denominación de origen protegida (PDOs) siendo algo propio de la tradición de Nápoles (Sacco et al. 2020).

En lo referente a España, gana en popularidad el tomate conocido como “Tomata de Penjar d‟Alcalà de Xivert”. Este tomate consiguió en diciembre de 2008 la Marca de Calidad de la Comunidad Valenciana, gracias a la Asociación de Productores y Comercializadores de Tomata de Penjar de Alcalà de Xivert creada en 2007 con el fin de revalorizar este producto (Figàs, 2019).

Figura 2. Mapa de donde se cultiva estas variedades de tomates. Las zonas marcadas con un punto negro son las dos zonas más representativas (Figàs, 2019; Sacco et al. 2020). Las zonas azules es la extensión donde se encuentra cultivada esta variedad de tomate (Siracusa et al. 2012; Sacco et al. 2020; Casals et al. 2011)

Dentro de los dos principales tipos de España e Italia, tenemos diferentes variedades. Del “Tomate De Penjar d’Alacalà de Xivert” tenemos las variedades “Estrella”, “Punteta” y “Moradeta” (Figàs et al. 2018). Se han hecho estudios que demuestran que las diferentes variedades también son debidas a mutaciones en el gen *alc* (Figàs, 2019). En el caso del “Pomodorino del Piennolo del Vesuvio” en Italia, encontramos las siguientes variedades “Fiaschella”, “Lampadina”, “Patanara”, “Principe Borghese” and “Rey Umberto”. Todas estas variedades también crecen en el monte Vesubio.

1.5. Estudios en esta variedad de tomate

El estudio de este tipo de tomate es importante a nivel de mejora genética. Estos presentan caracteres que pueden ayudar a combatir el estrés hídrico y dar una mayor durabilidad. Si se aplican a otras plantas podrían adquirir estas características, dando cultivos con mayor rentabilidad y más ecológicos (Siracusa et al. 2012; Figàs, 2019).

Estos genes no son los únicos que nos pueden interesar respecto a esta variedad de tomate. En el Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV) se ha estudiado como implementar la resistencia a virus y hongos a este tipo de tomate, puesto a que actualmente infecciones de este tipo representan un gran problema (Figàs, 2019).

Actualmente, aunque tenemos mucha información de las variedades “De penjar”, no sabemos realmente cuál es su origen genético. Por esa misma razón, este trabajo plantea un estudio genético para identificar el tipo de origen que tienen estas variedades. Para realizar esto vamos a re-analizar datos públicos de un experimento de genotipado por secuenciación (GBS, del inglés *Genotyping-By-Sequencing*)..

Nuestra hipótesis de partida es que las variedades de Italia y España poseen una diferenciación genética y por lo tanto un origen distinto. Por esa razón, nuestros objetivos son ver qué tipo de relación presentan las variedades de tomate procedentes de Italia y España y así probar (o rechazar) nuestra hipótesis. Una vez confirmada, la idea es tratar de ver el origen genético de estas variedades.

1.6. Análisis usando genotipado por secuenciación (GBS)

GBS es una tecnología simple, de coste y tiempo reducido y con alto rendimiento que se puede usar tanto para genotipar con un genoma de referencia o sin él. Posteriormente, permite realizar genética de poblaciones, estudios de evolución y reproducción molecular, entre otras cosas, usando millones de marcadores de secuencia (Torkamaneh et al. 2017). La dificultad de estos análisis recae en el llamamiento de variantes y el proceso bioinformático, ya que es lo que llevará más tiempo. Esto último es en lo que se centrará nuestro proyecto para así obtener unos resultados que nos permitan identificar el origen de esta variedad de tomate y ver si existe diferenciación o no entre las variedades de ambos países.

La implantación de tecnologías como GBS en el estudio del tomate ha supuesto un gran desarrollo en su investigación, destacando el estudio de su diversidad. Poder conocer características de poblaciones o variedades de tomates sin necesidad de establecer un genoma o la diversidad genética, es una de las mejoras que permite esta herramienta. Su uso permite una mejor gestión de los recursos genéticos a priori (Bauchet & Causse, 2012).

El GBS es una técnica de representación reducida del genoma (RRG), donde en lugar de secuenciar el genoma completo, se secuencian algunas partes, lo que permite abaratar los costes de genotipado (Faux, 2017). El GBS consiste en la digestión del ADN con una (o varias enzimas de restricción), la ligación de adaptadores a estos fragmentos, y una posterior selección del tamaño de los fragmentos basada en una reacción de amplificación de la polimerasa (PCR). Cada muestra se liga con un adaptador que posee un marcador diferencial específico para cada muestra a modo de código de barras (*barcode*). Una vez que las librerías se han preparado, éstas se secuencian mediante una metodología de lecturas cortas como Illumina (Figura 3). Los datos que aquí se

presentan se analizaron en el artículo de Blanca et al. 2023 si bien en este artículo no se trataron los tomates “De penjar” (Rajendran et al. 2022).

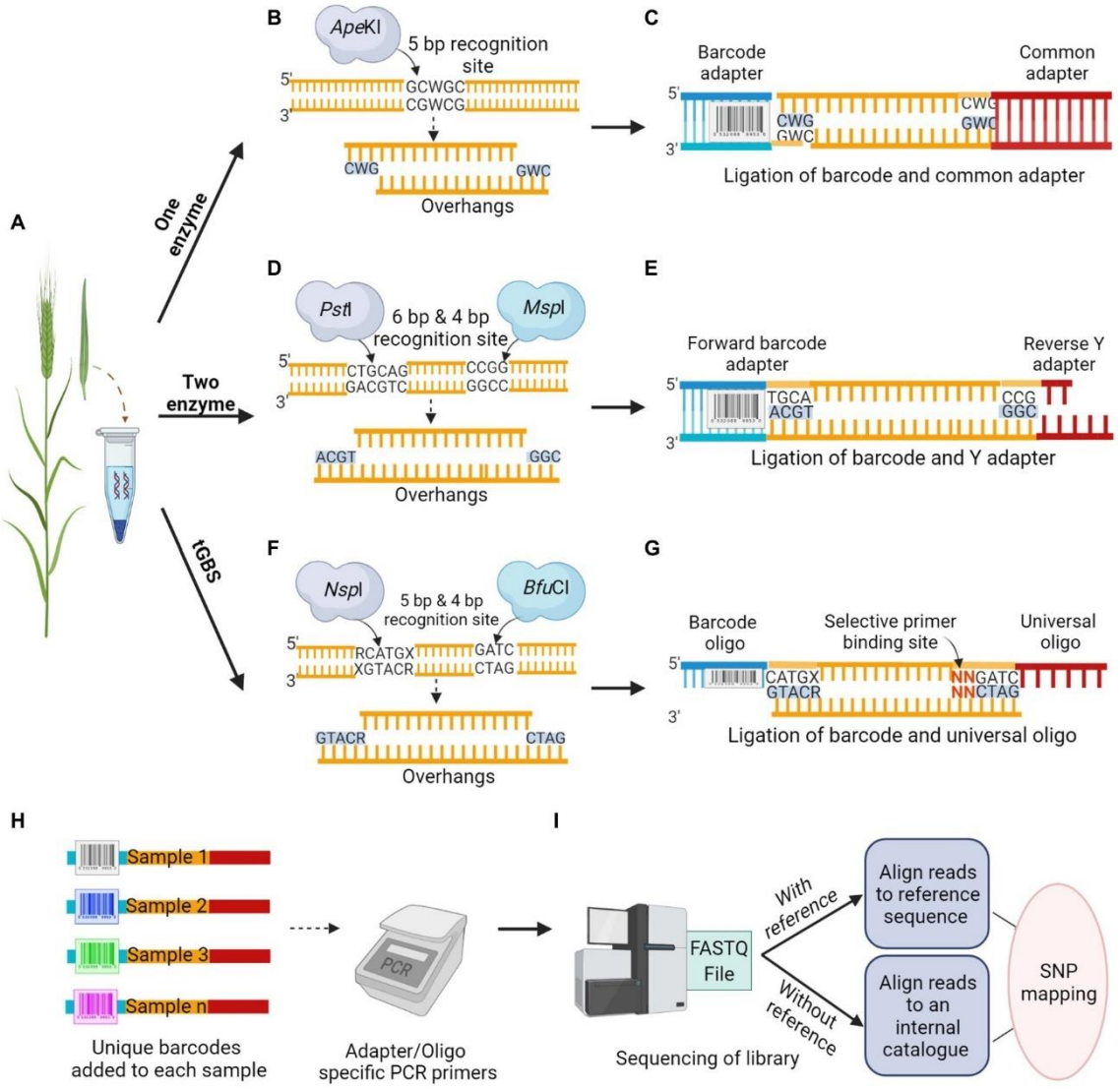

Figura 3. Esquema del proceso de GBS (Rajendran et al. 2022).

La forma de conseguir una representación reducida del ADN es seleccionando un determinado tamaño de los fragmentos generados. Para ello se limitó el tiempo de PCR de tal forma que solo se amplificaran fragmentos entre 0,5 y 1 kb. Una vez ya se obtuvo, los fragmentos se secuenciaron con Illumina. A partir de esas lecturas procesadas realizaremos nuestro llamamiento de variantes, para luego identificar con los resultados el origen de esta variedad de tomate (Figura 4).

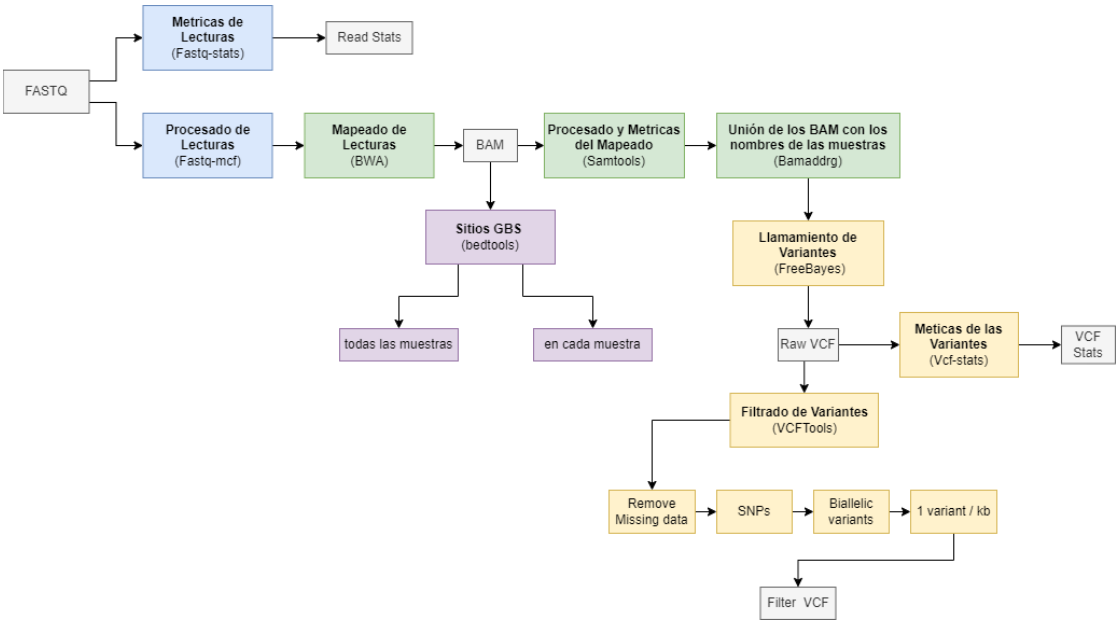

Figura 4. Diagrama explicativo del llamamiento de variantes (amarillo) y la búsqueda de sitios GBS (lila), con un previo procesamiento de las lecturas (azul) y un mapeado de las lecturas procesadas (verde). El color gris representa los diferentes archivos y sus formatos que requiere el análisis.

La realización de este tipo de análisis de forma eficiente es gracias a la aparición de las tecnologías de secuenciación de nueva generación, ya que permiten comparar gran número de secuencias para encontrar sitios polimórficos en el ADN en un corto tiempo (Rovelli et al. 2019). En nuestro caso, principalmente nos interesa los "*Single Nucleotide Polymorfism*" (SNPs), que son variaciones en la secuencia de ADN de un solo nucleótido que dará como resultado alelos diferentes.

A la hora de hacer el llamamiento de variantes para identificar SNPs es necesario tener en cuenta las duplicaciones completas del genoma, puesto a que dificulta su análisis. Es posible que se identifique como un SNPs un cambio de base, cuando realmente son secuencias parálogas (Rovelli et al. 2019).

2. MATERIALES Y METODOS

2.1. Muestras utilizadas

Las muestras que se han usado para el análisis pertenecen a un repositorio público denominado TRADITOM (https://traditom.eu/it/), que es un proyecto realizado desde 2015 hasta 2018 y financiado por la Unión Europea. El objetivo de este proyecto es la conservación y valorización de las variedades de tomates tradicionales, así como sus métodos de cultivo entre otras características. También la posibilidad de acceder a esta información a nivel genotípico y

fenotípico ha sido importante para este proyecto, además de brindar la posibilidad de mejorar estas variedades.

Los datos usados contenían un total de 1044 accesiones (Tabla 5), donde incluían variedades pertenecientes a diferentes lugares de España e Italia. Dentro de estas podemos encontrar la colección de la Universidad Politécnica de Valencia del Instituto COMAV. Las muestras también pertenecían a otros países como Francia, Grecia o Israel.

Para poder analizar estos datos para nuestra investigación usamos la herramienta Fastq-dump v2.9.6 (NCBI SRA Toolkit Development Team, 2019) para bajar los datos de secuenciación públicos de la base de datos NCBI SRA. Los datos están asociados a los proyectos PRJNA722111 y PRJNA774172. De esta forma importamos las secuencias de las muestras en formato FASTQ a uno de los servidores que usaremos para realizar el análisis.

2.2. Obtener información de las lecturas procesadas de cada muestra

Para obtener información de las lecturas de cada muestra usamos Fastq-stats v1.01 (Aronesty, 2013), puesto que las lecturas se encontraban en formato FASTQ. Las librerías de secuenciación del proyecto TRADITOM son "paired-end", partiendo de dos archivos FASTQ para cada muestra, uno correspondiente a cada extremo de los fragmentos de ADN secuenciados. Este comando generó un archivo de texto resumiendo las métricas de cada extremo de cada muestra. La información que contenían estos archivos incluía, entre otras cosas, el número de lecturas, su longitud media o la calidad de estas. Posteriormente resumimos algunas de estas características junto con otras, como podemos ver en la Tabla 3.

2.3.Mapear lecturas con el genoma de referencia

El genoma de referencia, *Solanum lycopersicum* v2.5 (Consortium, 2012), lo obtuvimos en formato FASTA de NCBI. Antes de poder usarlo para realizar el mapeado fue necesario crear un índice de este. Para ello usamos "bwa index", que es un comando concreto del programa Bwa v0.7.17 (Li & Durbin, 2009), el cual proporciona un alineamiento usando la transformación de Burrows-Wheeler para realizar el mapeado.

Una vez tuvimos el genoma indexado, procedimos con el alineamiento usando el comando "bwa mem". En este caso alineamos de forma conjunta ambos extremos de la misma muestra. Para obtener el archivo final BAM con los datos del alineamiento usamos "samtools view", que es un comando especifico del programa Samtools v1.7 (Li et al., 2009). Este comando convertirá el

resultado generado por “bwa mem” en un archivo BAM, además de filtrar las lecturas que no han sido mapeadas.

Una vez tuvimos el alineamiento para cada muestra fue necesario ordenarlo para proceder con el análisis. Para ello usamos el comando “samtools sort”. También obtuvimos las estadísticas de los alineamientos para cada muestra usando “samtools stats”. Este comando generó un archivo de texto para cada muestra que contenía diferente información como por ejemplo el número de lecturas mapeadas, lecturas duplicadas, si se encontraban ordenadas o no …

2.4.Unión de todos los alineamientos con los nombres de las muestras

Para poder realizar un llamamiento de variables, fue necesario unir los diferentes alineamientos de cada muestra en un archivo BAM que los contuviera todos. Para ello usamos el programa Bamaddrg v1.0 (Li, 2013). Al intentar juntar más de 1000 archivos obtuvimos un error, por lo que realizamos el procedimiento en dos pasos: 1- Unión de un conjunto de 500 archivos con Bamaddrg; y 2- Unión de dichos archivos usando “bamtools merge”.

Para crear el comando que introduciríamos en la consola con Bamaddrg, creamos un script de Python v3.6.9. Esto fue necesario, ya que la estructura del comando necesitaba introducir cada uno de los archivos BAM con el nombre de la muestra que pertenecía. Nuestro programa fue escrito usando Visual Studio Code v1.8 (Microsoft Corporation, 2016). Este tomaba como entrada un archivo de texto que contenía el nombre de las muestras que íbamos a usar y generaba el comando.

Tras obtener el archivo final BAM, con todos los alineamientos de todas las muestras, tuvimos que crear un índice de este usando “samtools index”. También usamos “samtools faidx” con nuestro genoma de referencia para, aparte de indexarlo, extraerlo. Es necesario este comando específico ya que nuestro genoma de referencia está en formato FASTA, siendo el correspondiente para este tipo de archivos.

2.5. Llamamiento de variantes

Para realizar el llamamiento de variables usamos el programa Freebayes v1.3.2 (Garrison & Marth,2012). Éste usa un método bayesiano para encontrar polimorfismos basándose en el haplotipo, la distribución de lecturas alineadas en una posición determinada y la probabilidad de que la variación detectada sea una variación real respecto a la observada. Para ello necesitamos tanto el genoma de referencia como el BAM con todos los alineamientos de las diferentes muestras. Aparte definimos algunos parámetros a la hora de correr el programa: la puntuación de

calidad mínima de mapeado debía ser de 30, la puntuación de calidad mínima por base debía ser de 30 también, la cobertura mínima de lecturas debería ser de 10 y la máxima la calculamos multiplicando el número de muestras por 100.

Aunque se basa en el programa descrito anteriormente, con cada uno de los parámetros mencionados, realmente usamos MutiThreadFreeBayes v0.1 (https://github.com/aubombarely/GenoToolBox/tree/master/SNPTools) que nos permite separar el BAM en diferentes trozos de tal forma que acelere el proceso de llamamiento de variables corriendo procesos paralelos. Para ello definimos el parámetro *threads* como 13, ya que el genoma de referencia del tomate contiene 13 secuencias (uno por cada cromosoma, y una extra denominado cromosoma 0 que contiene aquellas secuencias que no pudieron asignarse a ningún cromosomas real).

Una vez tuvimos el archivo VCF con el llamamiento de variables, usamos Bgzip v1.7-2 para comprimirlo y así usar Tabix v1.7-2 para indexar el archivo comprimido. Luego usamos "vcf-stats" para obtener las estadísticas del llamamiento de variables para cada muestra y en conjunto. Este último generará una carpeta con diferentes archivos con diferente información. En nuestro caso nos interesó el que tiene formato DUMB puesto a que contenía toda la información necesaria respecto a las estadísticas del llamamiento de variantes. Esta herramienta pertenece al programa Vcftools v0.1.15 (Danecek et al., 2011). El archivo DUMB incluía el número de variantes, las que son únicas para cada muestra, los SNPs y los InDels, entre otras cosas.

2.6. Filtrado de las variantes encontradas

Usando Vcftools filtramos las variantes. Primero eliminamos los datos faltantes. Luego seleccionamos las variantes que solo tuvieran dos alelos. También eliminamos los InDels quedando solo SNPs. Finalmente, eliminamos las variantes que estaban más juntas de 1000 bases (Figura 4).

2.7. Encontrar sitios

Usamos el programa Bedtools v2.26.0 (Quinlan & Hall, 2010) para identificar los sitios que fueron secuenciados asociados a las librerías de GBS. Primero convertimos tanto los archivos BAM de cada muestra y el de todas las muestras juntas a BED usando "bedtools bamtobed". La salida de este comando se unió con "bedtools merge". Así obtuvimos diferentes BED, que, al contar las líneas, usando "wc -l", nos daría el número de sitios. Antes de contar las líneas usamos

“bedtools intersect “con el BED generado de cada muestra junto al total, para ver que sitios tenían en común.

2.8. Obtener una tabla con la información

Para resumir todos los datos obtenidos durante el análisis, desarrollamos programa en Python v3.6.9 que permitía extraer de cada uno de los archivos métricos correspondientes a cada muestra la información e introducirla en un archivo TSV. Una vez tenemos este archivo, creamos otro programa con la intención de crear otro archivo TSV, para posteriormente obtener una tabla con toda la información de cada muestra.

La información que contendría la tabla sería el nombre de la muestra, su accesión y colección, el número de lecturas que presentaba, las que han mapeado y el porcentaje de mapeado de las lecturas de esa muestra. También incluimos los sitios y el porcentaje de sitios que no se encontraban en esa muestra o “missing_data” y las variantes sin filtrar. Además, mediante una serie de parámetros nos permitía seleccionar si queríamos que apareciera las variantes que eran únicas para cada muestra, SNPs e InDels.

La accesión y la colección se obtuvieron del archivo REVISADO_Traditom_accesion_2013, el cual nos informaba de la procedencia de las diferentes muestras, entre otras cosas. Este archivo lo usamos en formato CSV.

El número de lecturas totales los obtuvimos sumando las que aparecían en los archivos generados usando Fastq-stats para cada uno de los extremos de la muestra. Las lecturas mapeadas se obtuvieron del archivo generado con “samtools stats”. Una vez tuvimos ambos valores calculamos el porcentaje de mapeo de esa muestra dividiendo el número de lecturas mapeadas de esa muestra entre el número total de lecturas que tenía esa muestra y lo multiplicamos por 100.

Para encontrar los sitios usamos los archivos generados al contar las líneas de los archivos BED tras intersecarlos. Posteriormente los juntamos con el archivo que contenía el número de líneas del BED de todas las muestras juntas, para así sacar los datos faltantes. Para ellos hicimos lo mismo que en el caso del porcentaje de mapeo para sacar el porcentaje, pero en este caso en el numerador usamos la resta entre el número total de sitios y el número de sitios intersectados de cada muestra.

En lo referente a la información de las variantes, todo lo obtuvimos del archivo DUMB generado con “vcf-stats”. De ahí se obtiene tanto el número de variantes, como las que son únicas de esa muestra, los SNPs y los InDels.

2.9. Unión de los diferentes comandos y programas

Para realizar los pasos anteriormente mencionados de forma eficiente fue necesario crear un programa escrito en Bash v5.1 (Bash, 2024) con todos ellos. Es por ello por lo que creamos un programa escribiera los comandos para poder ejecutarlos de manera eficiente.

2.10. Análisis de las muestras y sus datos

A partir de la tabla que obtuvimos con información acerca de cada una de las muestras que contenía el número de lecturas totales, lecturas mapeadas, el porcentaje de mapeado, el número de sitios, el porcentaje de datos faltantes, el número de variantes crudas y únicas y los InDels y SNPs, observamos como se distribuía cada una de estas características. Para ello calculamos la media y mediana de estas variables para todas las muestras. También miramos cuál de las muestras presentaba tanto el mínimo como el máximo para cada una de las variables enumeradas anteriormente, donde incluimos ese valor y la muestra a la que le correspondía (Tabla 3 y 4). Todo esto lo realizamos con Microsoft Excel.

Para ver mejor la distribución de las variantes en las muestras representamos la media y su desviación con gráficos de cajas y bigotes (Figura 6 y 7). También hicimos un histograma en R Studio con el número de variantes por muestra o densidad y cuantas muestras presentan esa densidad. Para ello tuvimos que importar la tabla con la información de cada muestra a R Studio en forma de “dataframe” y luego usamos “hist” para representar el histograma. Finalmente, obtuvimos su distribución usando la función “dnorm”.

Además, la tabla con cada una de las muestras también contenía su origen, por eso usamos esta información para construir dos mapas representando de donde provienen cada muestra en lo referente a España e Italia y sus regiones (Figura 9 y 10). Para ello también usamos Excel.

2.11. Representación del Análisis de Componentes Principales (PCA)

Para realizar los PCA usamos el “pipeline” de Zhian (2016). Primero tuvimos que instalar e importar las librerías de R: vcfR v1.12.0 (Knaus & Grünwald, 2017), adegenet v2.1.3 (Jombart, 2008), ggplot2 v.3.3.5 (Wickham, 2016), poppr v2.9.0 (Kamvar et al. 2014) y ape v5.5 (Paradis & Schliep, 2019). Luego fue necesario importar los datos de variantes (VCF) con el último filtrado. Para ello usamos “read.vcfR”. Una vez tuvimos los datos los convertimos en un objeto Genligth v2.1.3 (Jombart & Ahmed, 2022) que es una forma de guardar los SNPs de forma compacta. Posteriormente asignamos la ploidía del organismo a 2, que son el número de sets de cromosomas en una célula, usando “ploidy”.

Una vez tenemos listo nuestro “genligth” podemos realizar el PCA. Para ello usamos la función “glPca”. Esta función requiere el parámetro “nf” igual a un número entero, siendo el número de componentes principales que queremos. Primero usamos la función sin ningún valor para este parámetro de tal forma que nos dé todos los eigenvalores que definen la variabilidad que contiene cada componente principal. Para ello calculamos el porcentaje de variabilidad que contiene sumando todos los eigenvalores y dividiendo cada uno de ellos entre la suma total. Luego lo convertimos a porcentaje multiplicándolo por 100. Decidimos quedarnos con los primeros 20 componentes principales y exportamos sus valores a Excel usando la libraría de R “openxlsx” y la función “write.xlsx”. Una vez lo tuvimos en Excel representamos estos porcentajes (Figura 12 y 14) y seleccionamos solo los dos primeros puestos que eran los que representaban mayor variabilidad de nuestras muestras.

Una vez subimos los componentes principales que queremos, asignamos al parámetro “nf” el número 2. Esto nos daría los valores dentro del PCA para cada muestra, que añadimos a un “dataframe” con el nombre de las muestras.

Ahora que tenemos los datos del PCA para todas las muestras, duplicamos esos valores para estudiar por separado características específicas. Nos centramos en el país de origen de la muestra, el tipo o variedad y su uso. Usamos esas características para asignar diferentes colores en nuestros PCAs. Una vez tuvimos esta información asignada representamos los valores usando la función ggplot2 (Figura 13, 15, 16 y 17).

2.12. Representación del Análisis Discriminante de Componentes Principales (DAPC)

Para realizar este análisis también usamos el “pipeline” de Zhian (2016). Nos dará una representación como se distribuyen las poblaciones previamente definidas dentro de nuestras muestras. DAPC trata de separar los grupos lo máximo posible. Aparte de la librería ggplot2, también necesitamos la libraría tidyr v1.3.0 (Wickham, 2023).

Si queremos obtener los DAPC vamos a necesitar el objeto “geneligth” que creamos para los PCA y lo convertiremos en DAPC usando la función “dapc”. Sin embargo, en este caso fue necesario crear para cada característica, país, variedad o uso, su propio objeto “geneligth” y asignar su población respecto a estas características usando la función pop. Después de usar la función dapc usamos la función scatter y compoplot. Finalmente, representamos el DAPC usando ggplot (Figura 18 y 19).

2.13. Cálculo del F_{ST}

En el cálculo de F_{ST} tuvimos que dividir nuestras muestras por poblaciones. Creamos 5 poblaciones que contenían los nombres de las muestras que pertenecían a Italia, España y luego a las variedades de tomate "Da serbo", "De penjar" y "Ramellet". A continuación, usamos un comando de Vcftools que calcula el F_{ST} para las variantes de las muestras comparando entre las diversas poblaciones: vcftools --vcf input_data.vcf --weir-fst-pop population_1.txt --weir-fst-pop population_2.txt --out pop1_vs_pop2.

Obtuvimos los F_{ST} para Italia y España, y luego comparando cada variedad entre ellas y entre las tres juntas. Esto nos daba una lista para diferentes posiciones en el genoma de las que calculamos su media para obtener el F_{ST} de las poblaciones (Tabla 6).

El cálculo realizado por Vcftools se basa en el de Weir y Cockerham. Esta es una de las formas más usadas para el cálculo de F_{ST} (Bhatia et al. 2013), que usa la expresión que vemos en la Figura 5.

$$\hat{F}_{ST}^{WC} \rightarrow \frac{\left(F_{ST}^{1}+F_{ST}^{2}\right)}{\left(F_{ST}^{1}+F_{ST}^{2}+2\frac{1}{(M+1)}\left[M\left(1-F_{ST}^{1}\right)+\left(1-F_{ST}^{2}\right)\right]\right)}$$

Figura 5. Expresión para el cálculo de FST de Weir y Cockerham procedente de la publicación de Bhatia (2013).

3. RESULTADOS

3.1. Análisis de muestras

Las muestras usadas para el análisis de variantes son de la especie *Solanum lycopersicum*. En la Tabla 3 se puede observar el número de lecturas medio que presentaba cada muestra y su mediana. La primera parte del llamamiento de variantes da como resultado una media de 2.656.775 lecturas mapeadas. Comparado con el número de lecturas totales, la diferencia no es grande. Esto también se ve reflejado en el porcentaje de mapeado donde vemos que presenta una media de 98,93% (Tabla 1). Un porcentaje tan alto es positivo puesto que significa que la mayoría de las lecturas que presentan las diferentes muestras se han utilizado para el análisis, dando una mayor fiabilidad al análisis.

Tabla 3. Resumen de las estadísticas de las muestras usadas para realizar el análisis de variantes respecto al mapeado.

	Lecturas	Lecturas Mapeadas	Mapeado (%)
Media	2.694.664	2.656.775	98,93
Máx	11.346.218	11.324.908	99,95
Muestra (máx)	TRBA145	TRBA145	TRVA294
Mín	8.722	8.627	17,83
Muestra (mín)	TRTH287	TRTH287	TRCA108
Mediana	2.433.020	2.403.538	99,82

Respecto a las muestras con mayor y menor estadísticas, se puede observar que las muestras con mayor y menor lecturas también presentan el máximo y mínimo de lecturas mapeadas. Sin embargo, estas muestras no son las que presentan mayor o menor porcentaje de mapeado.

Tabla 4. Resumen de las estadísticas de las muestras y variantes usadas para realizar el análisis de variantes.

	Sitios	Datos Faltantes (%)	Variantes Crudas	Variantes Únicas	Indels	SNPs
Media	180.017	81,82	829.511	393	1.970	9.324
Máx	289.690	99,55	1.020.691	34.570	15.518	99.835
Muestra (máx)	TRBA145	TRTH287	TRBA145	TRVA505	TRVA505	TRVA507
Mín	4.457	70,74	11.449	7	30	135
Muestra (mín)	TRTH287	TRBA145	TRVA205	TRTH287	TRTH287	TRTH287
Mediana	177.895	82,03	877.811	330	1.956	7.712

Además de las estadísticas del mapeado también se obtuvieron las de los sitios, es decir, de las variantes encontradas en las diferentes muestras (Tabla 4). “Datos Faltantes” refleja el porcentaje de sitios que presenta cada muestra respecto el número de sitios totales identificados durante el análisis. En este caso tenemos una media de 81,82% lo cual representa que la mayoría de los sitios donde se encuentran variantes, están presentes en la mayoría de las muestras, lo cual podría hacernos pensar que no hay mucha variabilidad genética entre las diferentes muestras de tomate. De hecho, podemos ver que la muestra TRBA145 que contiene el menor porcentaje, aun así, presenta uno bastante elevado. Respecto a esto también podemos ver que el número medio de variantes únicas de cada muestra, representado en “Variantes Únicas”, es una cantidad pequeña respecto al número medio total de variantes en “Variantes Crudas”.

Los datos obtenidos reflejan una media de 180.017 sitios, con una media de variantes totales de 829.511, de los cuales solo se usaron para el análisis de variantes los SNPs. Al comparar la media y la mediana tanto en número de variantes totales como SNPs, vemos que hay una diferencia

notoria. La media se encuentra desplazada, lo cual se puede observar mejor en las Figura 6 y la Figura 7 donde vemos que hay una serie de variantes que se alejan bastante de la media total.

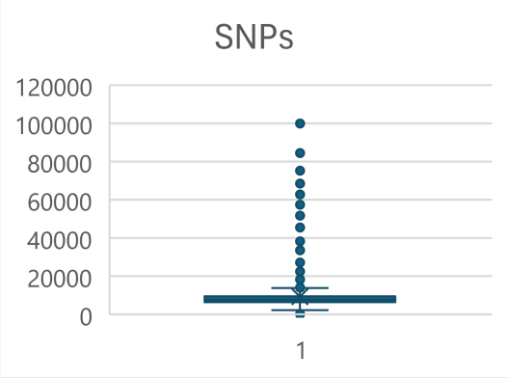

Figura 6. Gráfica de cajas y bigotes que representa el número medio de SNPs y cuáles se alejan de esa media.

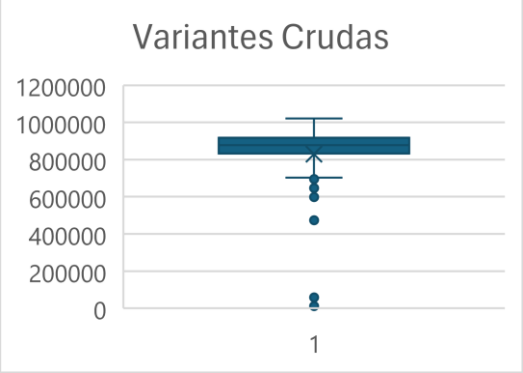

Figura 7. Gráfica de cajas y bigotes que representa el número medio de Variantes Crudas y cuáles se alejan de esa media.

Las muestras con mayor número de SNPs pueden reflejar una mayor variabilidad dentro de la misma especie. Dentro de lo observado lo más interesante son las muestras con escasas variantes. En la distribución de número de variantes en las diferentes muestras en la Figura 8 podemos ver mejor como existe esta discrepancia en el número de variantes en algunas muestras. Esto podría alterar nuestro análisis de las variantes, por eso la mejor opción sería eliminar estas muestras.

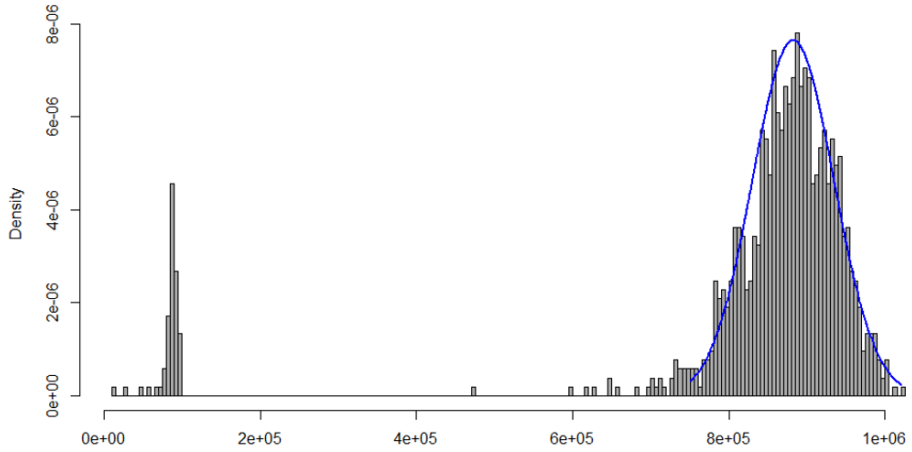

Figura 8. Distribución de muestras según el número de variantes crudas que contiene cada muestra. La línea azul representa la distribución en forma de campana de Gauss de la mayoría de las muestras.

Finalmente, para entender mejor las muestras a analizar, vamos a ver de donde provienen las muestras que son interesantes para el objetivo de la investigación, ver el origen de las variedades de tomate y su diferencia genética. De las 1058 muestras, solo 137 son de origen italiano y 537 de origen español. Estas muestras se distribuyen por ambas penínsulas de forma diferente (Figura 9 y Figura 10).

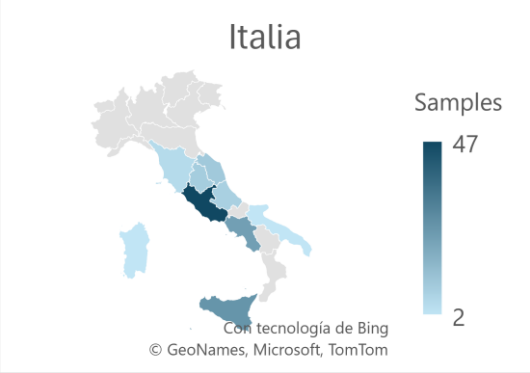

Figura 9. Distribución respecto al origen de las muestras procedente de Italia en diferentes regiones.

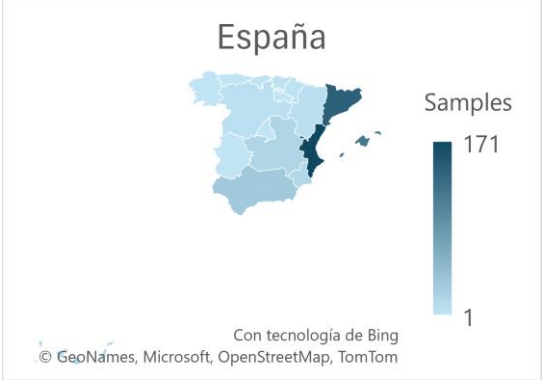

Figura 10. Distribución respecto al origen de las muestras procedente de España en diferentes regiones.

La mayoría de las muestras son provenientes de la Comunidad Valenciana, Cataluña y las Islas Baleares en España. En el caso de Italia se puede destacar Latium, Sicilia y Campania. Estas son las regiones donde la variedad de tomate estudiada se usa mayoritariamente.

Además de muestras procedentes de Italia y España, también se usaron muestras de diferentes países (Tabla 5). Cada origen tiene un código de accesión distinto, que nos puede ayudar a entender mejor las Tablas Suplementarias 1 y 2 del anexo.

Tabla 5. Códigos de las diferentes muestras respecto a su colección y su país de origen.

Código	Descripción	País
TR empty	Empty samples to test for contaminants	NA
TR control	Samples of known accessions to evaluate the library preparation	NA
TRPOOL	Pool of samples used for as control	NA
TR-Seed	NA	NA
TRBA	Univ. Illes Balears Collection	España
TRCA	FMA01 Collection	España
TRIS	HUJI Collection	Israel
TRMO	FRA030 Collection	Francia
TRPA	Univ. of Reggio Calabria Collection	Italia
TRPO	CNR-IBBR PORTICI Collection	Italia
TRTH	GRC005 Collection	Grecia
TRVA	ESP026 Collection	España
TRVI	ITA067 Collection	Italia

3.2. Reducción de muestras

Para poder eliminar las muestras que contienen un número reducido de variantes se ha usado la herramienta Vcftools. El comando utilizado es: vcftools —remove-indv, seguido de cada individuo que no superaba las 750.000 variantes. Finalmente, nos quedamos con un total de 958 muestras con una distribución de variantes que se muestra en la Figura 11. Las variantes como vemos siguen una distribución en forma de campana de Gauss. Las muestras eliminadas se encuentran en la Tabla Suplementaria 1 en el anexo.

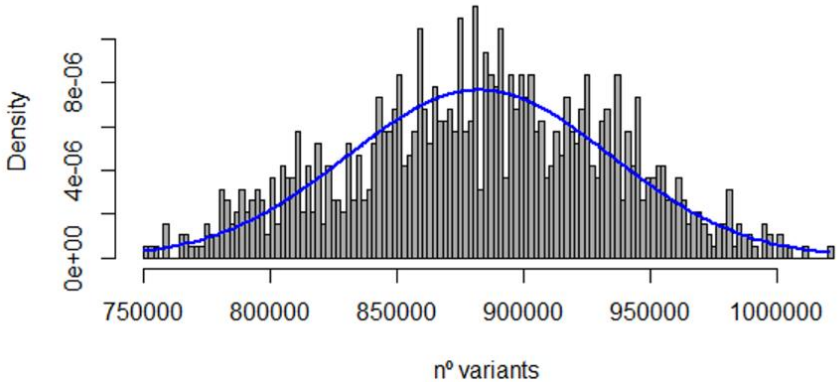

Figura 11. Distribución de las muestras según el número de variantes crudas tras eliminar aquellas que presentaban un número de variantes menor que 750000.

3.3. Análisis de 958 muestras mediante PCAs

El fin del Análisis de Componentes Principales (PCA) es estudiar la variabilidad dentro de las poblaciones de tomate y ver si estas se pueden clasificar según diferentes características. Primero se analizó el conjunto de 958 muestras sin asignarles una característica determinada en la Figura 13. Para poder asignar cuanta variabilidad corresponde a cada Componente Principal se usaron los Eigenvalores que se transformaron a porcentaje (Figura 12). La Figura 12 pertenece a los porcentajes de la Figura 13, que dan un valor para el PC1 de 24% y del PC2 del 12%. Al tener un número de muestras tan alto, aunque estos porcentajes sean bajos, siguen representando una gran variabilidad de la población.

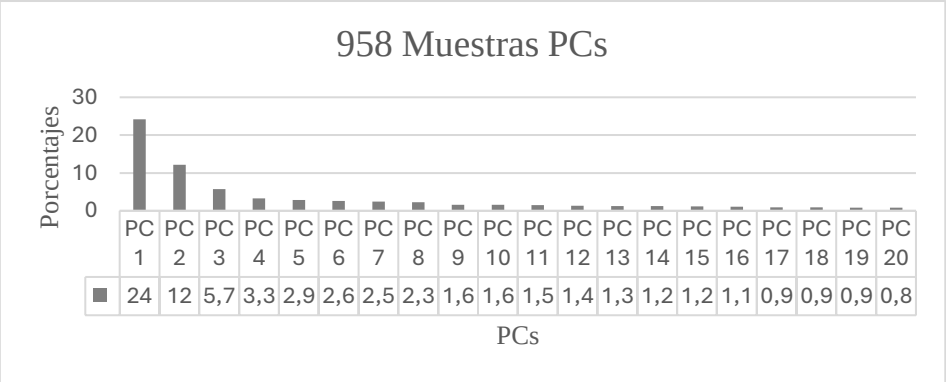

Figura 12. Distribución de cuanto abarca cada PC de la variabilidad de las muestras, con su porcentaje correspondiente en cada PC especifico.

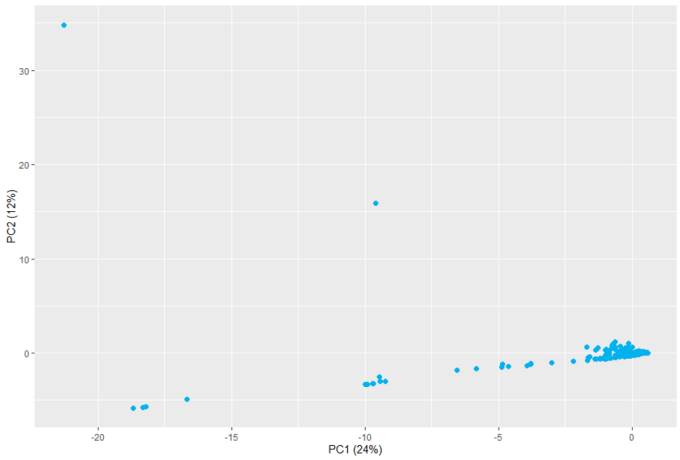

Figura 13. PCA de 958 muestras sin asignación de categoría especifica a cada una de ellas.

Se puede observar en el PCA de la Figura 13 que la mayoría de las muestras se encuentran agrupadas en un mismo punto, lo que significa que tienen poca variabilidad. La razón por la que se encuentran tan agrupadas es que dentro de las muestras analizadas existen unas que presentan una gran variabilidad genética, es decir, son muestras que probablemente hayan sufrido una presión evolutiva específica haciendo que la variabilidad de su genoma sea realmente mayor. En Tabla Suplementaria 2 encontramos estas muestras, donde se puede apreciar que no presentan ningún tipo de relación ni orden entre sí, además, de ser en su mayoría de origen tradicional.

Como estas muestras dificultan el análisis del resto, han sido eliminadas junto a algunos controles que solo añaden más ruido al análisis. Estos controles no favorecen el análisis puesto a que solo se usaron para comprobar que la preparación de la librería es correcta (Tabla 5). Todas aquellas muestras con un PC1 menor de -2 fueron eliminadas usando los mismos comandos que en la sección de reducción de muestras. No es relevante aplicar un filtrado basado en el PC2 puesto que aquellas muestras con una alta variabilidad en referencia a ese componente principal presentan

un PC1 menor que -2, es decir, no hace falta eliminarlas una segunda vez. Finalmente, el número de muestras que se usan para proseguir el análisis es de 917.

3.4. Análisis de 917 muestras mediante PCAs

Se calcularon de nuevo los porcentajes para el PCA con 917 muestras a partir de los eigenvalores de la Figura 14. En este caso, ambos porcentajes resultaron menores, con un PC1 de 9,6% y un PC2 de 7,7% lo que indica que el PCA anterior estaba fuertemente influido por las muestras eliminadas.

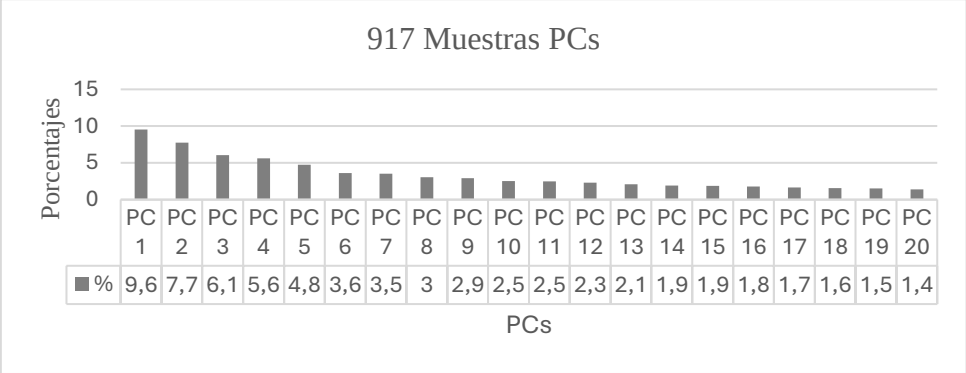

Figura 14. Distribución de cuanto abarca cada PC de la variabilidad de las muestras, con su porcentaje correspondiente en cada PC especifico.

En este caso se asignó a cada muestra una característica determinada, ya sea relacionada con el uso de ese determinado tomate, de su origen en lo referente a país o respecto su morfotipo. El primer PCA, en la Figura 15 se observa las diferentes muestras en lo referente a su uso. En este caso no se puede apreciar ningún tipo de grupo definido, lo que quiere decir que sus datos genéticos no están asociados al uso de cada variedad de tomate según estos datos.

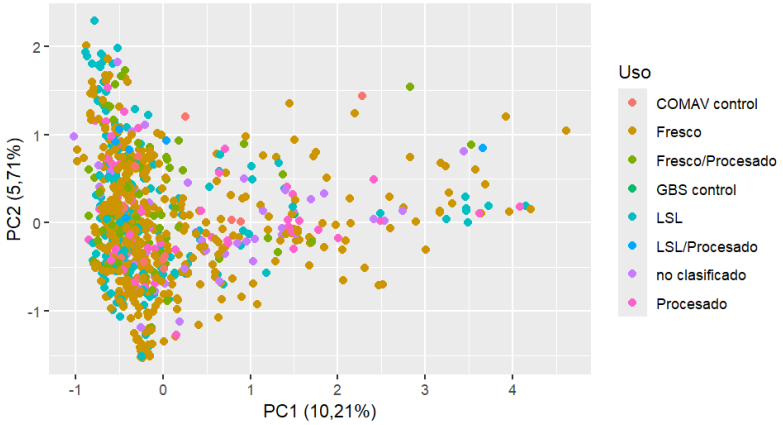

Figura 15. PCA de 917 muestras con asignación a cada muestra su correspondiente uso.

En el caso del PCA de la Figura 16 se puede observar la clasificación con relación al morfotipo. Aunque en todas las muestras había más morfotipos, aparte de "Da serbo", "De penjar" y "Ramellet", se usaron solo estos tres últimos, puesto a que son los más relevantes para la investigación. Por esa razón, el resto de morfotipos se agruparon en "Otros". Aunque en este caso si se puede observar cómo cada tipo se sitúa principalmente en un punto del PC2, en el caso del PC1 no presentan muchas diferencias. Aun así, en el PC2 siguen entremezclándose los diferentes tipos, por lo que se puede concluir que no hay ningún tipo de grupo definido en referente al morfotipo. Estos resultados no apoyan nuestra hipótesis de diferenciación genética entre estas poblaciones.

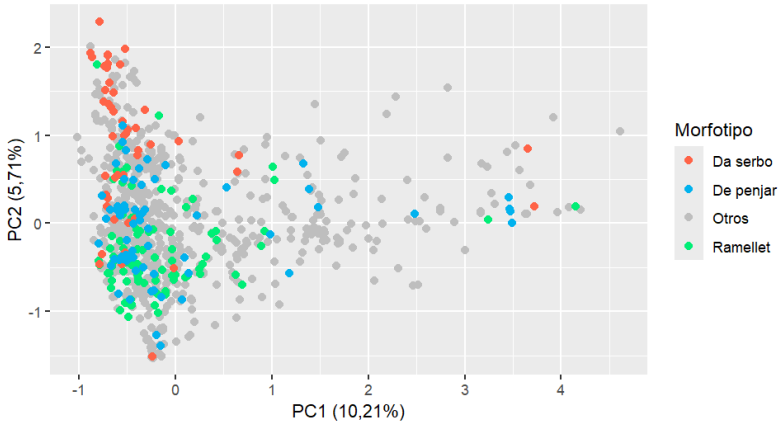

Figura 16. PCA de 917 muestras con asignación a cada muestra su correspondiente morfotipo.

Por último, en el PCA de la Figura 17, se asignaron colores según el lugar de origen de la muestra. De igual forma que con el PCA de los morfotipos, al solo ser relevantes para el análisis Italia (ITA) y España (ESP), se agrupó todos los otros países como "Otros". En este caso también ocurre que hay dos tendencias de agrupación de muestras en el PC2, pero en el PC1 en ambos casos se encuentran en el mismo punto. Aunque en el PC2 sí existe algo más de separación entre los diferentes países, estos grupos siguen estando demasiado entremezclados para poder considerarlo como dos variedades diferentes genéticamente.

Sí que, en este caso, en las muestras de España existe una mayor variabilidad tanto en el PC1 como en el PC2. Esto tiene sentido históricamente puesto a que el tomate entro en Europa a través de España y por esa razón ha tenido más tiempo para presentar variabilidad genética que el tomate que se considera originario de Italia. Sin embargo, esto no nos da información del origen genético, aunque podemos pensar que está relacionado con el histórico

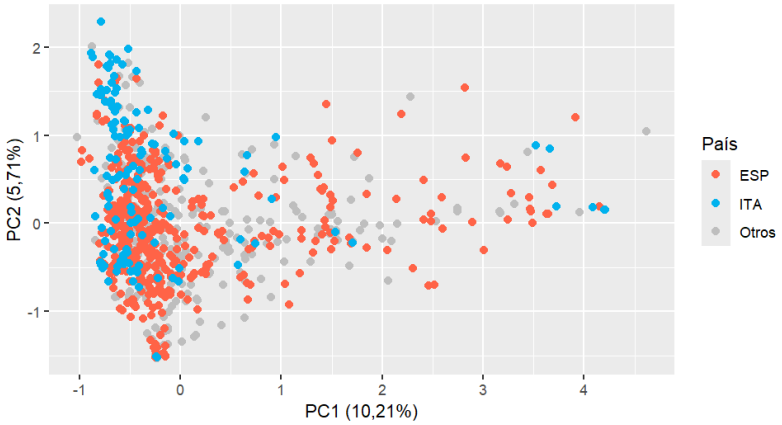

Figura 17. PCA de 917 muestras con asignación a cada muestra su correspondiente país.

Todos estos resultados dan a entender que la variación genética es baja y que no existen una diferenciación genética entre los diferentes grupos, rechazando de esta forma nuestra hipótesis inicial. Esto puede ser por dos razones, uno que el ancestro de todas estas variedades es el mismo, sin que ocurra suficiente presión evolutiva para que se diferencien, o que ha habido un intercambio genético constante entre ambos grupos. El primer caso tiene sentido históricamente hablando, ya que como se ha comentado antes, el tomate fue introducido en España desde América y luego se propagó por toda Europa (Blanca et al, 2022) .

Sin embargo, la comercialización europea de alimentos y de semillas también da a pensar que ha habido ese intercambio genético constante. A demás, tanto en Italia como en España tenemos un clima similar por lo que el desarrollo de las mismas semillas y variedades sería posible. Todo esto nos dificulta dar una respuesta clara a nuestro objetivo de encontrar origen de estas variedades, pero podemos pensar que genéticamente no son diferentes poblaciones de tomate. Esto se relaciona con nuestro objetivo de comprobar si hay diferenciación genética entre las variedades de tomate estudiadas.

3.5. Análisis de las 917 muestras mediante DAPCs

Para ampliar el análisis de las muestras respecto los PCA también se usamos Análisis Discriminante de Componentes Principales (DAPC). El fin de este análisis es ver si existe algún grupo y como se relacionan entre ellos. Como al estudiar los PCA solo se veía algún tipo de tendencia en el caso del morfotipo y país, solo se analizan los DAPC en referente a estas características.

En la primera figura de esta sección (Figura 18), se ve representado los países a los que pertenecen. Como ocurría en los PCA no se puede observar ningún grupo definido. Tanto en las

muestras de origen español como italiano u otros orígenes, las muestras son genéticamente similares.

Una característica que sí se puede diferenciar en todas las muestras de tomate es que estas comparten una mayor probabilidad de la población de origen español. Esto también apoya el hecho de que históricamente el tomate se introdujo en Europa por España y todas las demás variedades europeas provienen de estas. No obstante, no parece que exista una diferenciación genética asociada al origen que asigne las muestras de forma inequívoca a cada uno de los grupos (de ahí que cada grupo contenga una contribución).

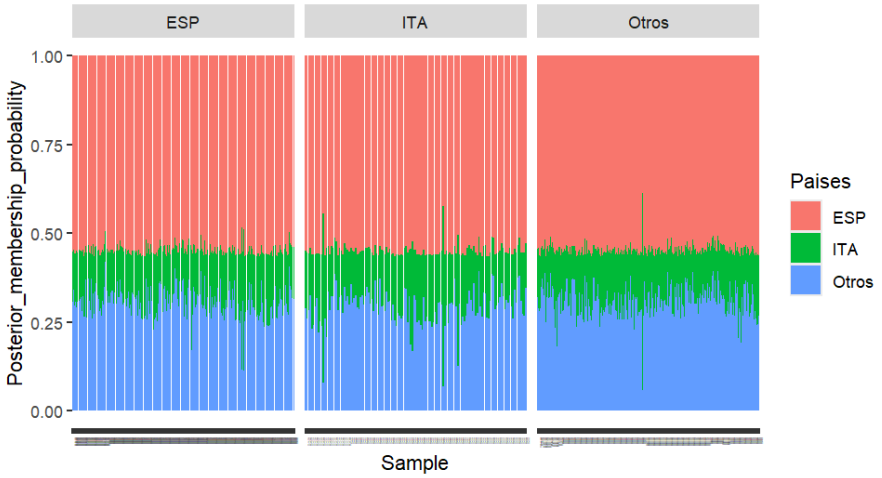

Figura 18. DAPC de las muestras agrupadas según país y con las diferentes poblaciones definidas también respecto a estos países.

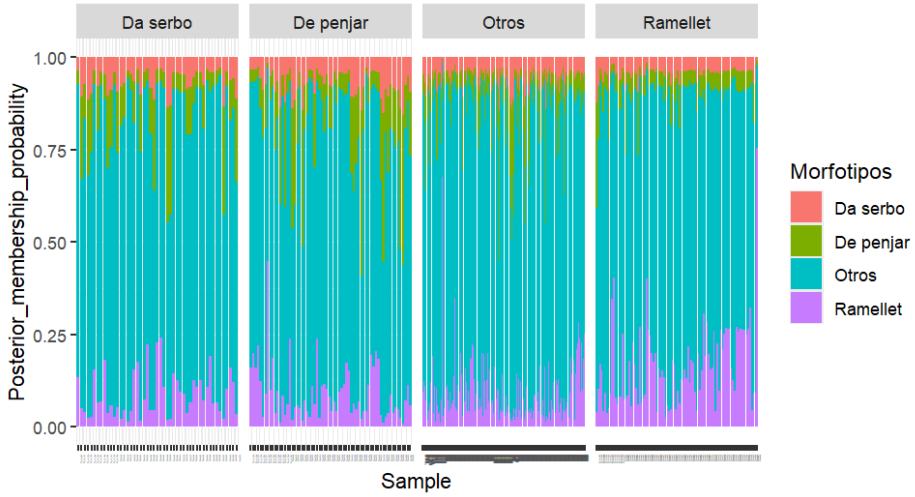

Figura 19. DAPC de las muestras agrupadas según país y con las diferentes poblaciones definidas también respecto a estos países.

También podemos ver la representación de las muestras en lo referente a los morfotipos en la Figura 19. Como cabe esperar, tampoco se puede ver ningún tipo de diferencia en los morfotipos. Quizá se puede observar algo más de similitud entre las variedades de "Da serbo" y de "De penjar", pero no altamente diferenciable.

3.6. Estudio de la estructura poblacional

Para estudiar la estructura poblacional calculamos el F_{ST} comparando las diferentes poblaciones de interés (Tabla 6). Este parámetro nos indica la varianza contenida en una población respecto la varianza genética total. Pueden obtenerse valores entre 0 y 1 (Weir, 2012). Se puede observar cómo en todos los casos el valor es altamente cercano a 0, incluso en un caso es negativo, lo que indica un valor de 0. Cuanto más cercano a 0 se encuentre el F_{ST} de las dos o tres poblaciones, menos separación se encuentra entre estas genéticamente hablando y, por lo tanto, menor diferenciación entre las poblaciones comparadas.

Tabla 6. F_{ST} entre las diferentes variantes de las poblaciones indicadas.

Poblaciones	F_{ST}
España / Italia	0,0349
"De penjar" / "Ramellet"	0,0045
"De penjar" / "Da serbo"	0,0061
"Da serbo" / "Ramellet"	0,0106
"De penjar" / "Ramellet" / "Da serbo"	-0,0562

El valor más bajo lo encontramos en las dos variedades que son consideradas de origen español, " De penjar " y "Ramellet", lo cual tiene sentido. Sin embargo, entre las variedades que provienen de países diferentes no encontramos un valor significativamente mayor de F_{ST}.

4. DISCUSIÓN

La falta de claridad a la hora de definir la trayectoria del tomate en Europa es algo que se ha reportado en previas investigaciones. Esto se debe en parte a su poca diversidad genética comparada con las variedades originales que vinieron de América. Como se comenta en la investigación de Blanca *et al*. (2022), en general las variedades que se consideran como tradicionales presentan un número menor de variantes respecto a las más modernas, lo que indica una conservación genética mayor y una menor diversidad genética. Además, el desequilibrio de ligamiento es el menor de los tomates en el caso de las variantes tradicionales, lo que indica poca subestructura en la población.

Como hemos observado en los diversos PCAs (Figura 15 y 16), esta falta de diversidad genética entre las diferentes variedades europeas es algo que se mantiene en nuestro estudio. Además, gracias a los F_{ST} cercanos a 0 también podemos concluir que no hay diferenciación genética en nuestras poblaciones. Todo esto nos lleva a pensar que las diferencias que encontramos entre las variedades estudiadas respecto a su fenotipo se deben a variaciones en genes concretos y no en una diversidad genómica general. La creencia de que las variedades tradicionales europeas se deben a la implicación de unos pocos genes concretos también se puede ver en la publicación de Blanca *et al*. (2022). Un objetivo futuro para esta investigación sería centrarnos en encontrar estos genes y ver como sus diferentes haplotipos se adaptan a diferentes climas y cómo podemos usarlo para mejora vegetal. Por ejemplo, podríamos centrarnos en el gen *alc*, que sabemos que tiene relación con la larga duración de estas variedades. En el artículo de Pons *et al*. (2022) también se menciona lo importante que puede ser el estudio de genes específicos, resaltando el gen *alc* como uno de los más interesantes para ello.

Todo esto apoya el rechazo de nuestra hipótesis inicial de una diferenciación genética entre estas variedades genéticas. Esto nos permite cumplir uno de nuestros objetivos que era ver como se relacionan estas poblaciones de tomate, sin embargo, al no poder considerar estas variedades como poblaciones independientes no nos permite saber cuál es origen de estas genéticamente, puesto a que no hay suficientes diferencias como para hacer un estudio de ello. Como hemos comentado previamente, esto puede ser por el intercambio genético constante entre las diferentes variedades o porque no ha habido una presión evolutiva significante para que se produzca una diferenciación. La primera hipótesis se ve respaldada en el artículo de Blanca *et al.* (2022) donde indica que los haplotipos estudiados de las variedades tradicionales presentan una fuerte evidencia de contactos secundarios entre esas poblaciones.

Sin embargo, algo que se contrapone a nuestros resultados obtenidos es la publicación de Esteras *et al*. (2022), donde nos dice que, entre todas las variedades estudiadas dentro de la Europa tradicional, las pertenecientes a España e Italia presenta una mayor diferenciación genética puesto a que dice que son los grupos que presentan un mayor número de sitios polimórficos. Esto puede ser verdad, ya que no hemos comparado que tan diferentes son nuestras variedades estudiadas respecto a otras variedades europeas no pertenecientes a Italia o España. No obstante, aunque sean las más diferenciadas a nivel europeo, partiendo de la poca variabilidad general en Europa a nivel tradicional, nuestro estudio aporta una visión más enfocada a estos países, recalcando que no se pueden considerar como poblaciones diferentes de tomate.

Respecto a esto último, sí que podemos ver que nuestros PCA de origen y morfotipo (Figura 16 y 17), cuando comparamos la posición de nuestras muestras estudiadas respecto a las otras de origen europeo englobadas en “Otros” vemos que se superponen. Esto también refleja la poca

diferenciación que existe entre las muestras y variedades de estos países y el resto que presenta variedades tradicionales de tomate en Europa. Un objetivo futuro puede ser ver qué relación tienen más variedades europeas de otras latitudes con las estudiadas en esta investigación, sobre todo intentando centrarnos en ellas que presentan una larga duración.

En la investigación de Esteras *et al*. (2022) ofrece otro planteamiento y es que la mayor diferenciación se produce a diferentes latitudes. Esto tiene relación con la idea de que no hay suficiente presión evolutiva para que se produzca una diferenciación genética entre estas variedades, puesto que sus principales lugares de cultivo en Italia y España se encuentran a latitudes cercanas, como se puede ver en la Figura 2. También comparten clima, lo que les permite adaptarse a estas variedades de forma similar. Al estar sufriendo un aumento de sequías y climas secos, como los del clima mediterráneo, en más partes del mundo debido al cambio climático, estudiar variedades que se encuentran bien adaptadas a ello es altamente importante. Esto se puede relacionar con el objetivo de desarrollo sostenible 13 "Acción por el clima", 15 "Vida de ecosistemas terrestres", 12 "Producción y consumo responsables" entre otros.

Además de factores naturales como el clima o la latitud, también se sabe que en agricultura se busca obtener unas características determinadas de tomate que fuerzan la selección artificial de diferentes variedades. Esto dificulta nuestro análisis porque pueden aparecer diferentes niveles de variación, presentando casos donde sí vemos una mayor diversificación, pero artificial. Esto es lo que creemos que ocurre en el PCA de la Figura 13, donde como vemos enmascaraban el resto de variación de nuestras muestras.

Por último, me gustaría recalcar el estudio de Pons *et al*. (2022) donde se representan diferentes caracteres que producen los diferentes fenotipos de tomates, por ser lo contribuye a una mayor variabilidad. Mientras que la forma de tomate es lo que contribuye mayormente, el origen que es lo que se centra nuestra investigación, no es representativo, donde no supera el límite establecido por ellos.

5. CONCLUSIÓN

El tomate (*Solanum lycopersicum L.*) es un cultivo de la familia de las solanáceas, donde la información genética del origen de algunas variedades es escasa, sobre todo respecto a las variedades europeas tradicionales. Tras un análisis exhaustivo del grupo de variedades tradicionales de tomates de una larga vida postcosecha mediante diferentes métodos bioinformáticos, podemos concluir que nuestra hipótesis era errónea, que no hay una diferenciación genética significativa entre las variedades de tomate "De penjar", "Da serbo" y "Ramellet". Las diferencias fenotípicas que encontramos entre estas variedades se deberían

principalmente a genes concretos. La falta de diferenciación se puede deber a un intercambio genético constante o por falta de presión evolutiva.

Para llegar a estas conclusiones hemos usado lecturas obtenidas mediante genotipado por secuenciación (GBS), que favorece el estudio de poblaciones de tomates. A partir de ellas obtuvimos diferentes SNPs, mediante un llamamiento de variantes, que se usaron para llegar a estos resultados específicos mediante PCA, DAPC o F_{ST}, herramientas ampliamente usadas para el estudio de poblaciones.

El estudio de estas variedades de tomates es importante por sus características de uso, su larga duración y su capacidad de crecer en climas más secos. Encontrar los diferentes genes que favorecen estas características puede ser un objetivo futuro para luego implementarlo en mejora vegetal.

6. BIBLIOGRAFÍA

Aronesty, Erik. "Comparison of sequencing utility programs." *The open bioinformatics journal 7.1* (2013).

Bash. (2024). Bash (Version 5.1) [Computer software]. *GNU Project*. https://www.gnu.org/software/bash/

Bauchet G., Causse M, (2012) Genetic diversity in tomato (Solanum lycopersicum) and its wild relatives. Genetic diversity in plants, *IN-TECH Education and Publishing*, 978-953-51 0185-7. ffhal-02805788

Bergougnoux V. (2014). The history of tomato: from domestication to biopharming. *Biotechnology advances*, 32(1), 170–189. https://doi.org/10.1016/j.biotechadv.2013.11.003

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome research*, 23(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Blanca J, Canñizares J, Cordero L, Pascual L, Diez MJ, et al. (2012) Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. *PLoS ONE* 7(10): e48198. doi:10.1371/journal.pone.0048198

Blanca, J., Sanchez-Matarredona, D., Ziarsolo, P., Montero-Pau, J., van der Knaap, E., Díez, M. J., & Cañizares, J. (2022). Haplotype analyses reveal novel insights into tomato history and domestication driven by long-distance migrations and latitudinal adaptations. *Horticulture research*, 9, uhac030. Advance online publication. https://doi.org/10.1093/hr/uhac030

Casals, J., Pascual, L., Cañizares, J., Cebolla-Cornejo, J., Casañas, F., & Nuez, F. (2011). Genetic basis of long shelf life and variability into Penjar tomato. *Genetic Resources And Crop Evolution*, 59(2), 219-229. https://doi.org/10.1007/s10722-011-9677-6

Consortium, T. S. G. (2012). Solanum lycopersicum cultivar Heinz 1706 (tomato) genome assembly SL2.50 [Genome assembly]. Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCF_000188115.3/

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Group, 1000 Genomes Project Analysis. (06 2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330

Esteras, C., Pons, C., Montero-Pau, J., Sanchez-Matarredona, D., Ziarsolo, P., Fontanet, L., Fisher, J., Prohens, J., Casals, J., Rambla, J. L., Riccini, A., Pombarella, S., Ruggiero, A., Sulli, M., Grillo, S., Kanellis, A. K., Giuliano, G., Finkers, R., Cammareri, M., . . . Granell, A. (2022). European traditional tomatoes galore: a result of farmers' selection of a few diversity-rich loci. *Journal Of Experimental Botany,* 73(11), 3431-3445. https://doi.org/10.1093/jxb/erac072

Faux P. (2017). Bioinformatics of reduced-representation genomics for population analyses. Laboratório de Biodiversidade e Evolução Molecular .*Universidade Federal de Minas Gerai*. http://labs.icb.ufmg.br/lbem/aulas/pg/tge-rrg.pdf

Figàs Moreno, M. (2019). CARACTERIZACIÓN, TIPIFICACIÓN, SELECCIÓN Y MEJORA GENÉTICA DE VARIEDADES VALENCIANAS DE TOMATE [Tesis doctoral no publicada]. *Universitat Politècnica de València*. https://doi.org/10.4995/Thesis/10251/119449

Figàs Moreno, M., Calancha, C. C., Dias, L. P., Rosa, E., Calduch, M., Herrera, J. J., Tomás, J. P., & Aleixandre, S. S. (2018). Mejora genética de tres variedades de tomate «De Penjar» valencianas para resistencia al virus del mosaico del tomate. *Dialnet*. https://dialnet.unirioja.es/servlet/articulo?codigo=7317604

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv E-Prints*, arXiv:1207.3907. doi:10.48550/arXiv.1207.3907

Jombart, T., & Ahmed, I. (2022). genlight: A class for storing genetic markers in R [R package adegenet version 2.1.3]. R Foundation for Statistical Computing. Retrieved from https://search.r-project.org/CRAN/refmans/adegenet/html/genlight.html

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405. https://doi.org/10.1093/bioinformatics/btn129

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281. https://doi.org/10.7717/peerj.281

Kaur, G., Abugu, M., & Tieman, D. (2023). The dissection of tomato flavor: biochemistry, genetics, and omics. *Frontiers In Plant Science*, 14. https://doi.org/10.3389/fpls.2023.1144113

Knaus, B. J., & Grünwald, N. J. (2017). VcfR: an R package to manipulate and visualize variant call format data. *Molecular Ecology Resources*, 17(1), 44-53. https://doi.org/10.1111/1755-0998.12549

Li, H. (2013). bamaddrg (Version 1.0) [Software]. *GitHub*. https://github.com/ekg/bamaddrg

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows Wheeler Transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMTools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Microsoft Corporation. (2016). Visual Studio Code (Version 1.8) [Software]. https://code.visualstudio.com/

Mutschler, M. A. (1984). Inheritance and Linkage of the 'Alcobaca' Ripening Mutant in Tomato. *Journal of the American Society for Horticultural Science*, 109(4), 500-503. Retrieved Apr 15, 2024, from https://doi.org/10.21273/JASHS.109.4.500

NCBI SRA Toolkit Development Team. (2019). NCBI SRA Toolkit (Version 2.9.6) [Software]. *National Center for Biotechnology* Information. https://ncbi.github.io/sra-tools/

Ochogavía, J. M., López, M., Rigo, M., Garau, M. M., March, J., Moscardó, J., et al. (2011). Caracterització de les poblacions de tomàtiga de ramellet de les Illes Balears. Quaderns d'Investigació 9. Palma. Agricultura i Pesca. *Conselleria de Presidència. Govern de les Illes Balears*, ISBN: 978-84-614-7284-0

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526-528. https://doi.org/10.1093/bioinformatics/bty633

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

Rajendran, N. R., Qureshi, N., & Pourkheirandish, M. (2022). Genotyping by Sequencing Advancements in Barley. *Frontiers In Plant Science*, 13. https://doi.org/10.3389/fpls.2022.931423

Rovelli, V., Ruiz-González, A., Vignoli, L., Macale, D., Buono, V., Davoli, F., Vieites, D. R., Pezaro, N., & Randi, E. (2019). Genotyping-by-Sequencing (GBS) of large amphibian genomes: a comparative study of two non-model species endemic to Italy. *Animal Biology*, 69(3), 307-326. https://doi.org/10.1163/15707563-00001094

Sacco, A.; Cammareri, M.; Vitiello, A.; Palombieri, S.; Riccardi, R.; Spigno, P.; Grandillo, S. (2020). Italian traditional tomato varieties: a focus on the Campania region. En I Congrés de la Tomaca Valenciana: La Tomaca Valenciana d'El Perelló. *Editorial Universitat Politècnica de València*. 179-193. https://doi.org/10.4995/TOMAVAL2017.2017.6526

Siracusa, L., Patanè, C., Avola, G., & Ruberto, G. (2012). Polyphenols as chemotaxonomic markers in Italian "long-storage" tomato genotypes. *Journal of agricultural and food chemistry*, 60(1), 309–314. https://doi.org/10.1021/jf203858y

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), 5. https://doi.org/10.1186/s12859-016-1431-9

Weir BS. Estimating F-statistics: A historical view. *Philos Sci*. 2012 Dec;79(5):637-643. doi: 10.1086/667904. PMID: 26405363; PMCID: PMC4578636.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag*. https://ggplot2.tidyverse.org

Wickham, H. (2023). tidyr: Tidy Messy Data (Version 1.3.0) [R package]. https://CRAN.R-project.org/package=tidyr

Zhian N. Kamvar, & Niklaus Grunwald. (2016). grunwaldlab/Population_Genetics_in_R: First release (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.160588

ANEXOS

Tabla Suplementaria 1. Muestras eliminadas por que presentaban un menor número de variantes

Muestra	Accesión	Colección	Muestra	Accesión	Colección
TR_PA_027	RC10017	Univ of Reggio Calabria	TRcontrol3_4_001	control	control
TR_PO_026	SR- 7	CNR-IBBR PORTICI	TRcontrol3_7_001	control	control
TR_PO_060	-	CNR-IBBR PORTICI	TRMO017	T101768	FRA030
TR_VA_030	BGV000973	ESP026	TRMO018	T101759	FRA030
TR_VA_062	BGV002109	ESP026	TRMO042	T100486	FRA030
TR_VA_078	BGV002301	ESP026	TRMO066	T101948	FRA030
TR_VA_079	BGV002302	ESP026	TRMO068	T100729	FRA030
TR_VA_205	BGV005670	ESP026	TRMO083	T100863	FRA030
TR_VA_511	V3	-	TRMO084	T100885	FRA030
TR_VI_013	V710051	ITA067	TRMO085	T101450	FRA030
TR_VI_023	V710109	ITA067	TRMO086	T100900	FRA030
TR_VI_050	V710159	ITA067	TRMO089	T100906	FRA030
TR_VI_113	V710249	ITA067	TRMO090	T100909	FRA030
TRBA048	UIB0040	Univ Illes Balears	TRMO091	T100910	FRA030
TRBA096	UIB0090	Univ Illes Balears	TRMO092	T102642	FRA030
TRBA097	UIB0091	Univ Illes Balears	TRMO106	T101403	FRA030
TRBA108	UIB0102	Univ Illes Balears	TRMO107	T101764	FRA030

TRBA111	UIB0105	Univ Illes Balears	TRMO108	T101406	FRA030
TRCA011	BGHZ4817	FMA01	TRPO021	-	ARCA2010
TRCA013	BGV002125	FMA01	TRTH018	ANP-088/07	GRC005
TRCA030	BGV002074	FMA01	TRTH024	AO-037/07	GRC005
TRCA039	LC14	FMA01	TRTH059	GRC1112/04	GRC005
TRCA048	LC55	FMA01	TRTH066	GRC1594/04	GRC005
TRCA054	LC102	FMA01	TRTH067	GRC445/04	GRC005
TRCA101	LC361	FMA01	TRTH071	GRC488/04	GRC005
TRCA103	LC375	FMA01	TRTH083	IS-025/07	GRC005
TRCA109	LC417	FMA01	TRTH091	IS-075/07	GRC005
TRcontrol2_2_001	control	control	TRTH111	P-070a/06	GRC005
TRcontrol3_3_003	control	control	TRTH120	KD-062/07	GRC005
TRVI071	V710191	ITA067	TRVI037	V710137	ITA067
TRVI011	V710046	ITA067	TRVA127	BGV005419	ESP026
TRTH127	IK-161/06	GRC005	TRVA140	BGV005460	ESP026
TRTH128	GRC007/06	GRC005	TRVA145	BGV005478	ESP026
TRTH146	-	-	TRVA159	BGV005511	ESP026
TRTH161	-	GRC005	TRVA166	BGV005531	ESP026
TRTH176	ANP-016/07	GRC005	TRVA174	BGV005577	ESP026
TRTH187	M-109 (ab)/06	GRC005	TRVA202	BGV005662	ESP026
TRTH189	GRC058/04	GRC005	TRVA224	BGV009803	ESP026
TRTH214	T-516/06	GRC005	TRVA238	BGV012579	ESP026
TRTH238	SK-001/06	GRC005	TRVA243	BGV012845	ESP026
TRTH280	X-027/06	GRC005	TRVA270	BGV015356	ESP026
TRTH287	XKA-057/07	GRC005	TRVA306	-	ESP026
TRVA080	BGV002991	ESP026	TRVA311	-	ESP026
TRVA086	BGV003070	ESP026	TRVA327	-	ESP026
TRVA106	BGV003920	ESP026	TRVA506	PER2 (Ana P.)	-
TRVI001	V710148	ITA067			

TRADITOM CODE	Collecting site	Country	Biological status of accession	Core colection

TRIS0010	-	IL	-	Core traditional collection
TRIS0020	-	IL	-	Core traditional collection
TRPO0590	SanPierniceto	ITA	-	-
TRVI1450	-	ITA	300	-
TRVI1440	-	ITA	300	-
TRVI1460	-	ITA	300	-
TRVI1390	Viterbo	ITA	300	Core traditional collection
TRVI0870	Bolsena	ITA	300	Core traditional collection
TRBA1610	Menorca	ESP	300	Core traditional collection
TRBA1690	Santa Margalida	ESP	300	Core traditional collection
TRTH1440	M.PERISTERI	GRC	300	-
TRTH1600	SIGRI	GRC	300	-
TRVA1170	Murcia	ESP	-	Core traditional collection
TRVA2190	Águilas	ESP	-	Core traditional collection
TRVA5050	-	-	-	-
TRVA5070	-	-	-	-

Tabla Suplementaria 2. Muestras con mayor variabilidad genética según el PCA con 958 muestras. El 300 significa que es de origen tradicional.