Programa de Doctorado en Bioingeniería

# Herramientas moleculares para la caracterización genética de *Allium sativum* L. y otras especies vegetales

**Ricardo Parreño Montoro**

Director de la tesis

**Dr. D. Héctor Candela Antón**

Universidad Miguel Hernández de Elche

- 2024 -

La presente Tesis Doctoral, titulada “Herramientas moleculares para la caracterización genética de *Allium sativum* L. y otras especies vegetales”, se presenta bajo la modalidad de **tesis por compendio** de las siguientes **publicaciones**:

- **Parreño, R., Rodríguez-Alcocer, E., Martínez-Guardiola, C., Carrasco, L., Castillo, P., Arbona, V., Jover-Gil, S., Candela, H.** (2023). Turning Garlic into a Modern Crop: State of the Art and Perspectives. Plants 2023, 12, 1212. https://doi.org/10.3390/plants12061212
- **Martínez-Guardiola, C., Parreño, R., Candela, H.** (2024). MAPtools: Command-Line Tools for Mapping-by-Sequencing and QTL-Seq Analysis and Visualization. Plant Methods, aceptado para su publicación

Adicionalmente, en esta Tesis Doctoral también se incluyen las siguientes publicaciones pendientes de aceptación:

- **Parreño, R., Rodríguez-Alcocer, E., Gallego-Zaragoza, A., Ferriz, A., Castillo, P., Gómez del Castillo, F., Jover-Gil, S., Candela, H.** (2024). *De novo* Assembly and Annotation of the Mitochondrial and Chloroplast Genomes of Garlic (*Allium sativum* L.). Artículo pendiente de publicación.

El Dr. D. Héctor Candela Antón, director de la tesis doctoral titulada "Herramientas moleculares para la caracterización genética de *Allium sativum* L. y otras especies vegetales":

**INFORMA:**

Que D. Ricardo Parreño Montoro ha realizado bajo mi supervisión el trabajo titulado "Herramientas moleculares para la caracterización genética de *Allium sativum* L. y otras especies vegetales" conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 5 de marzo de 2024.

Dr. D. Héctor Candela Antón

Director de la tesis

La Dra. Dña. Piedad Nieves de Aza Moya, Coordinadora del Programa de Doctorado en Bioingeniería:

**INFORMA:**

Que D. Ricardo Parreño Montoro ha realizado bajo la supervisión de nuestro Programa de Doctorado el trabajo titulado “Herramientas moleculares para la caracterización genética de *Allium sativum* L. y otras especies vegetales” conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 5 de marzo de 2024.

Profa. Dra. Dña. Piedad Nieves de Aza Moya

Coordinadora del Programa de Doctorado en Bioingeniería

# Financiación

El trabajo recogido en esta memoria ha sido financiado parcialmente por los contratos con número de referencia COOPAMAN1.17I, COOPAMAN1.18I, COOPAMAN1.19I, COOPAMAN1.20I, COOPAMAN1.21I, COOPAMAN1.22I y COOPAMAN1.23I, celebrados entre COOPAMAN, Soc. Coop. C-LM y la Universidad Miguel Hernández de Elche al amparo del artículo 83 de la Ley Orgánica de Universidades (LOU), que han permitido la adquisición de los materiales necesarios, la realización de experimentos de secuenciación masivamente paralela y la contratación de D. Ricardo Parreño Montoro.

# Índice

# Índice de figuras

## Lista de Abreviaturas

**ADN**: Ácido desoxirribonucleico

**ADNcp**: ADN cloroplástico

**ARN**: Ácido ribonucleico

**ATP**: Adenosine triphosphate

**CCS**: Circular Consensus Sequencing

**CMS**: Cytoplasmic Male Sterility

**ED**: Euclidean Distance

**EMS**: Ethyl methanesulfonate

**ENU**: N-ethyl-N-nitrosourea

**gau**: gene antisense ubiquitous

**Gb**: Gigabases

**GB**: Gigabytes

**HiFi**: High Fidelity

**InDel**: Insertion-Deletion

**IR**: Inverted Repeat

**IRa**: Inverted Repeat a

**IRb**: Inverted Repeat b

**ITS**: Internal Transcribed Spacer

**LSC**: Large Single Copy

**Mb**: Megabases

**MBS**: Mapping-By-Sequencing

**NEP**: Nuclear-Encoded RNA Polymerase

**NGS**: Next-generation sequencing

**OLC**: Overlap-Layout-Consensus

**ONT**: Oxford Nanopore Technologies

**ORF**: Open Reading Frame

**PacBio**: Pacific Biosciences

**PAP**: Polymerase-Asociated Protein

**PEP**: Plastid-Encoded RNA Polymerase

**QTL**: Quantitative Trait Locus

**RAM**: Random Access Memory

**RAPD**: Randomly Amplified Polymorphic DNA

**RDR**: Recombinant-Dependent Replication

**RFLP**: Restriction Fragment Length Polymorphism

**RNA-seq**: secuenciación de ARN

**RT-qPCR**: Reverse-Transcrption quantitative Polymerase Chain Reaction

**SMRT**: Single-Molecule Real-Time sequencing

**SNP**: Single Nucleotide Polymorphism

**SNP-index**: Frecuencia alélica

**SOLiD**: Sequencing by Oligonucleotide Ligation and Detection

**SSC**: Small Single Copy

**SV**: Structural Variants

**WGD**: Whole-Genome Duplication

**ΔSNP-index**: Diferencia de las frecuencias alélicas

# 1 Resumen

El ajo (*Allium sativum* L.) es una planta que se cultivada en todo el mundo altamente apreciada por el valor comercial de sus bulbos. Sin embargo, su cultivo afronta desafíos significativos, como la infertilidad de las variedades de ajo comerciales y la acumulación de patógenos debido a la propagación vegetativa. En el primer artículo presentado en esta tesis, examinamos el estado actual de la Genética y Genómica del ajo, resaltando avances recientes que lo encaminan hacia su desarrollo como un cultivo moderno, incluida la restauración de fertilidad sexual en ciertas cepas. Los recursos genéticos de *Allium sativum* disponibles actualmente incluyen varias colecciones de marcadores moleculares, un ensamblaje del genoma a escala cromosómica, y múltiples ensamblajes de transcriptomas. Estas herramientas permiten profundizar en la comprensión de procesos moleculares subyacentes a rasgos de importancia como la infertilidad, la inducción de la floración y la formación de bulbos, las propiedades organolépticas y la resistencia a patógenos.

En el segundo artículo, realizamos la secuenciación, ensamblaje *de novo* y anotación de los genomas mitocondrial y cloroplástico del ajo. Este trabajo nos permitió establecer que los genomas de los cloroplastos del ajo y la cebolla son muy similares, así como evaluar el efecto de la edición de sus transcritos y la organización de éstos en operones. Por otro lado, conseguimos identificar más genes codificantes en el genoma mitocondrial del ajo que en el de la cebolla. El genoma mitocondrial de ajo es compatible con la existencia de numerosas moléculas subgenómicas, que se originan mediante recombinación en secuencias específicas, lo que sugiere que su estructura es dinámica.

En el tercer artículo, desarrollamos MAPtools, una aplicación de Python3 de código abierto diseñada específicamente para analizar datos genómicos obtenidos en experimentos de cartografía mediante secuenciación y de secuenciación de loci caracteres cuantitativos (QTL-seq). MAPtools proporciona a los usuarios una gran flexibilidad para personalizar su flujo de trabajo, calcular estadísticas basadas en el recuento de alelos, generar gráficos para identificar regiones candidatas y anotar los efectos de las mutaciones encontradas.

# 2 Abstract

Garlic (*Allium sativum* L.) plants are cultivated worldwide and are highly valued for the commercial value of its bulbs. However, its cultivation faces significant challenges, such as the infertility of commercial garlic cultivars and the accumulation of pathogens due to vegetative propagation. In the first article presented in this thesis, we review the current state of garlic genetics and genomics, highlighting recent advances that are moving it towards its development as a modern crop, including the restoration of sexual fertility in certain strains. Currently available genetic resources include several collections of molecular markers, a chromosome-scale genome assembly, and multiple transcriptome assemblies. These tools are allowing a deeper understanding of molecular processes that underly important traits such as infertility, flowering, bulb induction, organoleptic properties, and pathogen resistance.

In the second article, we performed the sequencing, *de novo* assembly and annotation of the mitochondrial and chloroplast genomes of garlic. This work allowed us to establish that the genomes of garlic and onion chloroplasts are very similar, as well as to evaluate the effect of editing on their transcripts and the arrangement of coding sequences into operons. On the other hand, we were able to identify more coding genes in the garlic mitochondrial genome than in the onion genome. The garlic mitochondrial genome is compatible with the existence of numerous subgenomic molecules, which originate by recombination in specific sequences, suggesting a dynamic structure.

In the third article, we introduce MAPtools, an open source Python3 application designed specifically for analyzing genomic data obtained in mapping-by-sequencing and quantitative trait loci sequencing (QTL-seq) experiments. MAPtools provides users with great flexibility to customize their workflow, calculate statistics based on allele counts, generate plots to identify candidate regions, and annotate the effects of mutations found.

# 3 Introducción general

## 3.1 El ajo (*Allium sativum* L.)

El ajo (*Allium sativum* L) es una de las plantas cultivadas más antiguas, con un alto valor comercial por las propiedades de sus bulbos. Se han encontrado referencias de su uso en Egipto que datan del 3700 a. C., y hay evidencias de su uso en China desde el 2700 a. C. y en Mesopotamia desde el 2600 a. C., seguramente importado desde China (Petrovska and Cekovska, 2010). Actualmente, el ajo se cultiva principalmente en regiones con clima templado de todo el mundo. La producción anual aproximada es de 29 millones de toneladas, distribuidas en 2,47 millones de hectáreas. Los mayores productores de ajo son China e India, que acumulan el 80% de la producción mundial (FAOSTAT).

El ajo es una especie diploide (2n = 2x = 16) del subgénero *Allium* de la familia *Amaryllidaceae*. Las variedades botánicas del ajo descritas inicialmente por Takagi (Takagi, 1990) se diferencian por su tallo floral. De este modo, los clones de ajos se clasifican como: de tallo floral completo, que producen un escapo floral que suele florecer; de tallo floral incompleto, que producen un escapo floral más fino y corto que no suele florecer; y sin tallo floral, que generalmente no producen un escapo floral y que, en caso de hacerlo, no producen flores sino bulbilos (Kamenetsky and Rabinowitch, 2001; Rabinowitch and Currah, 2002; Kamenetsky *et al.*, 2004; Takagi, 1990). Los bulbilos son pequeños bulbos que no están subdivididos en dientes, a partir de los que también pueden desarrollarse nuevas plantas, cumpliendo así un papel en la reproducción asexual. Los bulbilos pueden utilizarse para la propagación clonal del ajo, ya que al no estar en contacto con el suelo es menos frecuente que acumulen patógenos que afecten a su crecimiento (Kamenetsky and Rabinowitch, 2001).

Aunque se conocen variedades con y sin tallo floral en todas las regiones, el desarrollo del bulbo se ve afectado por el fotoperiodo (Mathew *et al.*, 2011) distinguiéndose cultivares de día corto y cultivares de día largo en función de sus requerimientos. A pesar de lo anterior, se sabe muy poco acerca de las diferencias genéticas existentes entre las variedades cultivadas en diferentes regiones geográficas.

### 3.1.1 Morfología y desarrollo de las plantas de ajo

El ajo es una planta monocotiledónea de tallo corto, raíces profundas, y hojas planas y largas, que alcanza una altura entre 40 y 60 cm. El ajo es un geófito desarrolla característicos bulbos subterráneos en la base del tallo. Los bulbos son órganos formados por hojas modificadas que sirven como almacén de nutrientes y que permiten

la reproducción asexual de la planta. Los denominados “dientes” son nuevos bulbos que se desarrollan a partir de los bulbos de la cosecha anterior, y están cubiertos por hojas secas generalmente de color blanco. El desarrollo de los tallos florales permite distinguir dos tipos de variedades: las que desarrollan un tallo floral suelen producir bulbos que contienen entre 4 y 12 dientes similares en tamaño, mientras que las que no lo desarrollan suelen contener más dientes, cuyo tamaño puede variar notablemente de unos a otros. Los dientes son bulbos sésiles que se forman a partir de los meristemos axilares en la parte adaxial (interior) de las hojas. Cada diente está rodeado por una fina capa de hojas protectoras, que pueden ser blancas, moradas o rojizas, y la parte interior está formada por tejido carnoso que constituyen la mayor parte del diente. En su parte central se dispone el meristemo apical del tallo, que se dispone en el ápice de un tallo muy corto y comprimido que recibe el nombre de placa basal (*basal plate*). A partir de los flancos del meristemo apical se desarrolla un brote predominante y varios primordios foliares que lo rodean. El crecimiento de las plantas comienza a partir de un diente que ha sido expuesto a temperaturas bajas (15ºC o menos) y ha salido del estado de dormancia. Alrededor del perímetro de la placa basal se desarrollan rápidamente raíces adventicias, precediendo a la emergencia de las hojas. Estas últimas crecen superponiéndose las unas a las otras, dando lugar a un “pseudotallo” compuesto por hojas muy próximas, casi prensadas. El sistema de raíces y hojas vegetativas se desarrolla antes de la formación del bulbo (Rabinowitch, 1990).

Las flores del ajo constan de 6 pétalos, 6 anteras y 3 lóculos con 2 óvulos cada uno. Son más pequeñas que las de la cebolla y su número varía desde unas 10 hasta más de 300 en cada inflorescencia, siendo lo habitual entre 150 y 200 flores. Los pétalos suelen ser de color púrpura suave, con o sin tintes más oscuros en el pedicelo y en el ápice de los pétalos. En las plantas derivadas de semillas se han encontrado flores bancas y púrpuras. El desarrollo de la inflorescencia del ajo requiere la exposición a temperaturas bajas (inferiores a 5°C), pero los requerimientos de vernalización del ajo no se conocen con exactitud. La floración del ajo está fuertemente influenciada por la predisposición genética, ya que algunos clones nunca florecen bajo condiciones inductivas (clones sin tallo floral). El desarrollo de los bulbos se inicia antes en las variedades que no desarrollan tallo floral, lo que sugiere la existencia de una competición entre el desarrollo del bulbo y el de la inflorescencia. El desarrollo del bulbo, generalmente, predomina sobre el desarrollo de las inflorescencias, excepto en algunas variedades cultivadas y silvestres en las que se pueden desarrollar simultáneamente (Kamenetsky and Rabinowitch, 2001). Los bulbos del ajo son ricos en compuestos sulfurados, siendo el más abundante de estos la alicina (óxido de disulfuro de alilo), que

además es el compuesto que les da su olor tan característico, y que se produce a partir de la catálisis del aminoácido no proteico aliina por la enzima aliinasa. Además de las propiedades aromáticas de este compuesto, la alicina también tiene propiedades antibióticas, hipoglucémicas, hipolipemiantes y antioxidantes (Borlinghaus *et al.*, 2014).

### 3.1.2 La reproducción del ajo

La producción comercial de ajo depende por completo de su propagación asexual, ya que la reproducción mediante semillas es inexistente en las variedades cultivadas. Sin embargo, parece que la reproducción sexual y la selección entre productos de la meiosis puede haberse dado ocasionalmente durante la larga historia del cultivo (Peña Iglesias, 1988). La selección clonal ha permitido generar nuevas variedades en otras plantas cultivadas con propagación vegetativa, como el boniato o el plátano, pero no ha sucedido de igual manera en el ajo. La investigación sobre la floración y la producción de semillas de ajo es, en consecuencia, muy importante para su futuro, y se espera que la investigación en este campo permita introducir mejoras sustanciales en este cultivo en los próximos años.

La propagación del cultivo se realiza normalmente a partir de los bulbos, ya que cada diente puede dar lugar a una planta completa. Al estar en contacto con el suelo, los bulbos tienden a acumular diferentes patógenos, un problema que se intenta paliar mediante el saneamiento periódico del cultivo, que se consigue utilizando técnicas de cultivo in vitro, que permiten la regeneración de plantas completas a partir del meristemo apical del tallo u otros tejidos. Por otro lado, la propagación vegetativa conlleva la uniformidad genética de las variedades cultivadas, un problema que carece de una solución inmediata. Dado que todas las plantas de una misma variedad son genéticamente idénticas (clones), existe un elevado riesgo en caso de enfermedades.

### 3.1.3 Importancia de la caracterización genética del ajo

Las diferentes variedades de ajo pueden diferir considerablemente en su fenotipo, si bien las diferencias genéticas entre ellas son pequeñas (Jo *et al.*, 2012; Singh *et al.*, 2012). Gracias a la reciente secuenciación del genoma nuclear del ajo, estas diferencias pueden ahora ser estudiadas en mayor profundidad. Durante la historia de la agricultura, los cruzamientos y la selección artificial han sido fundamentales para la domesticación de las plantas cultivadas y para adaptar su crecimiento a ciertos ambientes o condiciones. Sin embargo, la dificultad de generar nuevas variedades y a la reproducción clonal obligatoria de los cultivos de *Allium sativum*, se puede asumir que el ajo aún no está completamente domesticado. En este contexto, la identificación de variedades no comerciales capaces de producir flores y semillas viables es un paso

clave hacia la restauración de la fertilidad, que permitirá avanzar en la mejora genética de esta especie.

La falta de diversidad genética en el ajo es especialmente alarmante, ya que limita su capacidad de respuesta frente a plagas y enfermedades emergentes. La variación genética en una población dificulta la propagación de las plagas, ya que estas no afectarán por igual a todos los individuos. En las especies con reproducción clonal, sin embargo, las plagas representan un problema mucho mayor, como pone de manifiesto el caso de la banana, en el que la aparición de una cepa del hongo *Fusarium oxysporum* hizo que se tuviese que cambiar toda la producción mundial de la banana Gros Michel a la variedad Cavendish, que es la que se produce actualmente (Ordonez *et al.*, 2015). La aparición de una nueva plaga afectaría con igual virulencia a todos los individuos de una misma variedad, ya que, al ser clones, comparten los mismos mecanismos de defensa y respuesta a la infección. En este contexto, el estudio de la variación genética es especialmente importante, siendo vital identificar nuevas fuentes de variación genética y conservar las ya existentes. Antes de la secuenciación del genoma nuclear del ajo (Sun *et al.*, 2020), las únicas herramientas disponibles para los estudios genéticos se limitaban a algunas colecciones de marcadores moleculares, diversos estudios del transcriptoma (RNA-seq) y la secuencia del genoma del cloroplasto.

## 3.2 Las nuevas tecnologías de secuenciación

Las nuevas tecnologías de secuenciación de segunda y tercera generación, que se presentan a continuación, han experimentado avances significativos en los últimos años y han transformado radicalmente la investigación genética. Las tecnologías de secuenciación de segunda generación han permitido secuenciar genomas completos en un solo día gracias a su capacidad de secuenciar millones de fragmentos de ADN en paralelo. Las tecnologías de secuenciación de tercera generación producen lecturas de gran longitud en tiempo real, eliminando la necesidad de amplificación previa y permitiendo una mejor resolución de regiones genómicas complejas, si bien su uso plantea desafíos como la mayor tasa de error en comparación con las tecnologías de segunda generación.

### 3.2.1 Tecnologías de segunda generación

Las tecnologías de secuenciación de segunda generación, también denominadas “de la próxima generación” (*next-generation sequencing*; NGS), han revolucionado la investigación en los últimos años. A diferencia del método desarrollado por Frederick Sanger, por la que recibió el premio Nobel de Química en 1980, estas

tecnologías permiten secuenciar genomas de gran tamaño en un tiempo reducido. Las diferentes plataformas de NGS se basan en principios distintos, pero todas ellas comparten la capacidad de secuenciar en paralelo millones de fragmentos de ADN, por lo que también se conocen como “tecnologías de secuenciación masivamente paralela”.

El principal desafío de estas tecnologías radica en la complejidad del análisis bioinformático posterior, ya que generan millones de lecturas cortas (de entre 100 y 300 pb) que deben ser procesadas. Sin embargo, las NGS son muy fiables y su tasa de error es inferior al 1%, lo que las hace ideales para la detección de polimorfismos de un solo nucleótido (*single nucleotide polymorphisms*; SNP) y pequeñas inserciones y deleciones (las denominadas InDels). Por otro lado, no permiten caracterizar adecuadamente las variantes estructurales grandes (*structural variants*; SV), como ciertas secuencias repetidas o las regiones de baja complejidad del genoma. En consecuencia, las NGS no son las técnicas más apropiadas para la caracterización de transposones o enfermedades asociadas a variantes estructurales, ni para caracterizar con precisión las distintas isoformas de algunos transcritos.

La tecnología de secuenciación por síntesis de Illumina (anteriormente denominada Solexa) detecta los nucleótidos incorporados a una hebra de ADN por una polimerasa durante una reacción de secuenciación gracias a la emisión de una señal fluorescente. Esta tecnología parte de una librería (*library*) o colección de pequeños fragmentos de ADN a cuyos extremos se incorporan oligonucleótidos de secuencia conocida. Mediante un procedimiento denominado “amplificación puente” (*bridge amplification*), se consigue incrementar el número de copias de los fragmentos a secuenciar, que quedan unidos covalentemente a un soporte sólido por uno de sus extremos. La reacción de secuenciación utiliza nucleótidos terminadores marcados con distintos fluoróforos, que interrumpen la síntesis de ADN de manera reversible y permiten identificar los nucleótidos incorporados en cada ciclo de la reacción de secuenciación. Esta reacción se lleva a cabo mediante un protocolo similar al de Sanger, pero de forma masivamente paralela y automatizada.

La pirosecuenciación es otra tecnología de secuenciación, ya obsoleta, que se basaba en la detección del pirofosfato liberado al incorporarse cada nuevo nucleótido a la hebra de ADN. En este método, se añadían adaptadores a los extremos de cada molécula de ADN molde, lo que permitía su inmovilización sobre la superficie de una microesfera o perla (*bead*). La denominada “PCR en emulsión” se realizaba en una emulsión preparada con un aceite y una disolución acuosa con los reactivos necesarios, lo que permitía amplificar un único molde en cada microesfera gracias a la separación

de las fases en pequeñas cámaras. Tras la amplificación, las microesferas se transferían a unas placas horadadas, a razón de una microesfera por orificio, donde se llevaba a cabo la secuenciación propiamente dicha. Durante esta etapa, los nucleótidos se añadían uno a uno a la mezcla de reacción, lo que permitía identificar el nucleótido incorporado gracias a una reacción luminiscente desencadenada por la liberación de pirofosfato.

Además de la anterior, otras tecnologías también han caído en desuso. La secuenciación mediante ligación fue comercializada por Life technologies/Applied Biosystems bajo la denominación SOLiD (del inglés *sequencing by oligonucleotide ligation and detection*) y difería de otros métodos en que se basaba en la hibridación y ligación de sondas fluorescentes (Cloonan *et al.*, 2008; McKernan *et al.*, 2009; Tang *et al.*, 2009; Valouev *et al.*, 2008). La tecnología Ion Torrent se basaba en la medida de la diferencia de potencial iónico resultante de la liberación de protones que se produce al incorporarse cada nuevo nucleótido durante la secuenciación.

### 3.2.2 Secuenciación de tercera generación

Las modernas tecnologías de secuenciación de tercera generación permiten actualmente la secuenciación en tiempo real de moléculas individuales. La secuenciación de moléculas individuales ofrece la ventaja de no requerir una etapa de amplificación previa, lo que reduce el tiempo de preparación y los errores asociados a la misma. La principal diferencia entre los métodos de secuenciación de segunda y de tercera generación radica en el tamaño de las lecturas. Mientras que las tecnologías de segunda generación suelen rendir productos de entre 100 y 300 pb, las de tercera generación llegan a alcanzar longitudes de cientos o miles de kilobases (kb). Estas lecturas largas resultan particularmente útiles para abordar algunos de los problemas que plantea el ensamblaje de las lecturas cortas, como la resolución de regiones del genoma altamente repetitivas o de baja complejidad, o la incertidumbre inherente al ensamblaje de genomas o transcriptomas. Además, algunas tecnologías de tercera generación permiten la detección directa de algunas modificaciones epigenéticas, ya que la incorporación de los nucleótidos se produce en tiempo real y las metilaciones de las bases provocan un retraso en la actividad de la polimerasa (Dijk *et al.*, 2018).

La primera tecnología de tercera generación fue la comercializada por Helicos Biosciences, similar a la de Illumina, aunque en tiempo real y sin realizar una amplificación puente. Este método resultó ser muy lento y producía lecturas muy cortas, de unas 30-40 pb, por lo que fue pronto abandonado. Las dos tecnologías de secuenciación de tercera generación vigentes en la actualidad son las desarrolladas por

Oxford Nanopore Technologies (ONT), basada en el uso de nanoporos (Branton *et al.*, 2008), y la desarrollada por Pacific Biosciences (PacBio), basada en una metodología denominada SMRT (del inglés *Single-Molecule Real-Time sequencing*) (Eid *et al.*, 2009).

La metodología de PacBio se basa en la detección en tiempo real de los nucleótidos incorporados por una polimerasa a una molécula de ADN. Los cuatro nucleótidos están marcados con fluoróforos diferentes, que se liberan cuando se incorporan al ADN. La emisión de fluorescencia es detectada gracias al uso de las denominadas "guías de onda de modo cero" (*zero-mode waveguides*), lo que permite identificar qué nucleótido se ha incorporado a la hebra de ADN. La tecnología SMRT utiliza moldes circulares de ADN monocatenario, denominados *SMRTbells*, que se preparan mediante la ligación de adaptadores a los extremos de los fragmentos de ADN bicatenario que se desea secuenciar. Mediante un mecanismo de síntesis semejante a la replicación en círculo rodante (*rolling circle replication*), estos moldes circulares permiten generar tantas copias (denominadas "sublecturas") de ambas hebras de cada fragmento como permita la vida útil de la polimerasa. Gracias a que cada fragmento es secuenciado múltiples veces, la comparación de las sublecturas permite corregir los errores presentes en la secuencia. Esta metodología se conoce como "secuenciación de consenso circular" (*circular consensus sequencing*; CCS) y produce lecturas de alta fidelidad (*high fidelity*; HiFi). La tecnología SMRT permite detectar entre 2 y 4 nucleótidos por segundo, y genera lecturas de unos 10 kb de media (Lee *et al.*, 2014). La polimerasa utilizada para la secuenciación SMRT tiene una elevada tasa de error, de entre el 13 y el 20% (Hackl *et al.*, 2014; Ardui *et al.*, 2018), mientras que las polimerasas utilizadas en los métodos de segunda generación tienen tasas de error menores, de entre el 0,1 y el 1%, dependiendo de la tecnología utilizada. Sin embargo, la posibilidad de comparar las sublecturas producidas mediante secuenciación de consenso circular entre sí permite reducir la tasa de error de las lecturas HiFi a valores inferiores al 1% (Ardui *et al.*, 2018).

Por otro lado, la tecnología de Oxford Nanopore Technologies (ONT) ha dado lugar a MinION, un secuenciador portátil que puede conectarse a un ordenador utilizando un puerto USB. La tecnología ONT emplea nanoporos situados en una doble membrana lipídica que presenta un gradiente de concentración de cargas. El ADN, cuya carga es negativa, se ve forzado a atravesar estos nanoporos debido al gradiente iónico. Cada nucleótido que atraviesa el poro genera una perturbación característica en la amplitud de la corriente iónica, que es detectada mediante un sensor (Branton *et al.*, 2008). Cada fragmento de ADN a secuenciar contiene en uno de sus extremos un adaptador que permite su paso por el nanoporo. El otro extremo del fragmento está

ligado a un adaptador en forma de horquilla que lo conecta con su hebra complementaria. El paso de las dos hebras a través del poro permite determinar su secuencia, a la vez que facilita la corrección de errores e incrementa la calidad de la secuencia de consenso resultante. A pesar de su comodidad y su tamaño compacto, esta tecnología presenta una tasa de error elevada, que llega a alcanzar entre un 25 y un 40%, al no poder corregir las lecturas de manera tan eficiente como la CCS de PacBio, si bien produce lecturas muy largas, de entre 6 y 150 kb, que se utilizan actualmente para realizar el andamiaje (*scaffolding*) de genomas de gran tamaño (Ashton *et al.*, 2015).

## 3.3 Aproximaciones bioinformáticas para la caracterización de genomas y transcriptomas

Las tecnologías de secuenciación de segunda y tercera generación permiten el ensamblaje de genomas y transcriptomas, cada vez con mayor rapidez y precisión. En las siguientes secciones, presentamos las tres aproximaciones disponibles para investigar el contenido en genes en cualquier especie de interés: los ensamblajes guiados por un genoma de referencia ya conocido, el ensamblaje *de novo* de genomas completos, y el ensamblaje *de novo* de transcriptomas. Mientras que los ensamblajes guiados parten de un genoma conocido para facilitar el estudio de nuevas muestras, los ensamblajes *de novo* representan la mejor opción cuando no se conoce la secuencia de la especie a estudio o de alguna otra especie filogenéticamente próxima.

### 3.3.1 Métodos de ensamblaje *de novo* de genomas

Los métodos de ensamblaje *de novo* permiten determinar la secuencia nucleotídica del genoma o el transcriptoma de un organismo en ausencia de una secuencia de referencia. Esta técnica ha sido utilizada ampliamente en los últimos años, debido al abaratamiento de la secuenciación y a la disponibilidad de algoritmos de ensamblaje cada vez más potentes y precisos.

El ensamblaje de las lecturas producidas mediante el método de Sanger o, más recientemente, mediante las tecnologías de tercera generación se lleva a cabo mediante la estrategia denominada OLC (*Overlap-Layout-Consensus*; Gleizes and Hénaut, 1994; Peltola *et al.*, 1984; Smith, 1993; Staden, 1980). Algunos programas que utilizan esta aproximación son CAP3 (Huang and Madan, 1999), TIGR (Sutton *et al.*, 1995), PHRAP (de la Bastide and McCombie, 2007) y el ensamblador de Celera (Denisov *et al.*, 2008). El principal problema de la estrategia OLC reside en la dificultad de reconstruir correctamente la secuencia en genomas ricos en secuencias repetidas, ya que los

solapamientos (*overlaps*) detectados entre las lecturas pueden corresponden a secuencias que no necesariamente ocupan posiciones adyacentes (Meltz Steinberg *et al.*, 2017). Una dificultad adicional viene dada por la gran carga de trabajo que supone la identificación de estos solapamientos, que supone la realización de un número de comparaciones de orden $n^2$, siendo $n$ el número de lecturas disponibles. Para reconstruir la secuencia, en la estrategia OLC se construye un grafo de solapamientos (*overlap grap*h) (Pevzner and Shamir, 2011), en el que las lecturas se representan mediante vértices y los solapamientos entre ellas se representan mediante aristas o flechas. El ensamblaje de la secuencia se lleva a cabo mediante la búsqueda de un recorrido hamiltoniano a través del grafo de solapamientos. Los recorridos hamiltonianos visitan cada vértice una sola vez y su identificación en un grafo con muchos vértices constituye un problema NP-completo, de complejidad computacional intratable (Pevzner and Shamir, 2011), para cuya resolución no se conocen algoritmos eficientes. La estrategia OLC, en consecuencia, resulta más indicada para aquellos casos en los que se dispone de un número moderado lecturas de gran longitud (Li *et al.*, 2012), por lo que históricamente fue aplicado a la secuenciación de genomas muy pequeños (por ejemplo, de virus o bacterias), u otros más grandes a costa de invertir muchos recursos (como el *Human Genome Project*).

El desarrollo de los secuenciadores de segunda generación acentuó el problema del ensamblaje, tanto por el elevado número de lecturas como por su reducida longitud, al quedar patente que su ensamblaje mediante la estrategia OLC resultaba inviable en cuanto a tiempo y recursos computacionales se refiere (Baichoo and Ouzounis, 2017). La complejidad computacional es una medida de la cantidad de recursos requeridos por un algoritmo para resolver un problema definido de una forma y en un contexto específicos (Baichoo and Ouzounis, 2017). La complejidad computacional describe el rendimiento de un algoritmo conforme se incrementa la magnitud del problema, y puede referirse tanto al tiempo de ejecución (complejidad temporal) como al consumo de memoria (complejidad espacial; Cormen *et al.*, 2009). Para abordar este problema, se desarrollaron algoritmos de ensamblaje basados en el uso de grafos de De Bruijn. En estos grafos, cada lectura se descompone en subsecuencias de longitud $k$, denominadas *k-meros*. Cada *k-mero* se representa mediante una arista o flecha que conecta los vértices correspondientes al prefijo y el sufijo de longitud $k$-1 de su secuencia. La construcción de un grafo de De Bruijn puede realizarse procesando las lecturas en una sola pasada, lo que es muy eficiente en términos de complejidad computacional. Además, la secuencia del genoma puede deducirse mediante la búsqueda de un recorrido euleriano (que visita cada arista una sola vez) a través del

grafo, un problema para cuya resolución también existen algoritmos de gran eficiencia (Zerbino and Birney, 2008). El ensamblaje de secuencias mediante grafos de De Bruijn, por lo tanto, es mucho más rápido y requiere mucha menos memoria que la estrategia OLC (Pevzner and Shamir, 2011).

Existen infinidad de programas para realizar el ensamblaje *de novo* de las lecturas. Diversos autores han realizado múltiples comparativas entre ellos en los últimos años (Rana *et al.*, 2016; Zhang *et al.*, 2011). Cualquier ensamblaje *de novo* suele constar de varias etapas. En la primera de ellas, se realiza un procesamiento previo de las lecturas con el objetivo de descartar las de menor calidad o que no alcanzan una longitud mínima, así como para eliminar cualquier secuencia derivada de los adaptadores que se añadieron durante la preparación de las librerías. En segundo lugar, se realiza el ensamblaje *de novo* propiamente dicho, seguido del andamiaje o *scaffolding* de los cóntigos (*contigs*) resultantes. En esta etapa, se determina el orden y la orientación relativa de los cóntigos obtenidos, y se estima el tamaño de los huecos existentes entre ellos. Como resultado de este proceso, se generan estructuras de mayor longitud, denominadas andamios (*scaffolds*) o supercóntigos (supercontigs). Algunos programas de *scaffolding* son SSPACE-LongRead, OPERA-LG, LINKS o Cerulean. Una vez detectados, estos huecos pueden ser rellenados mediante programas como IMAGE, GapFiller, Sealer o GapCloser, que aprovechan las lecturas iniciales para extender los extremos de cóntigos adyacentes y completar la secuencia. Sin embargo, se ha comprobado que estos programas cometen errores con una tasa entre 20 y 500 veces mayor que el rellenado de los huecos con lecturas largas (Kosugi *et al.*, 2015). La última etapa consiste en identificación y anotación de los genes en la secuencia obtenida.

No existe un programa de ensamblaje *de novo* que funcione mejor que los demás en todos los aspectos, y la elección de uno u otro depende de las características concretas de las lecturas y de la secuencia a ensamblar (por ejemplo, de su contenido en secuencias repetidas). Además, con la llegada de las tecnologías de tercera generación sigue habiendo un amplio margen de mejora a pesar de su alta tasa de error. Dependiendo de la plataforma de secuenciación utilizada, los errores de secuenciación pueden ser inserciones, deleciones y/o sustituciones nucleotídicas. En las estrategias basadas en grafos de De Bruijn, estos errores incrementan la complejidad del grafo, creando burbujas, cabos sueltos o conexiones erróneas. Estos errores pueden detectarse y a veces corregirse antes del ensamblaje mediante algoritmos basados en el recuento de *k-meros*, mediante el uso métodos estadísticos o probabilísticos, o

después del ensamblaje, mediante el realineamiento de las lecturas iniciales a la secuencia ensamblada.

Entre los desafíos a afrontar en los ensamblajes *de novo* destacan los errores de secuenciación, que pueden dar lugar a artefactos y secuencias quiméricas; el contenido en secuencias repetitivas, que plantea dificultades a la hora de ensamblar correctamente las secuencias; los sesgos de las tecnologías de secuenciación, que pueden favorecer la secuenciación de ciertas regiones sobre otras, resultando en regiones no secuenciadas o con una cobertura inconsistente; y por último, la cantidad de recursos computacionales que necesarios para completar los ensamblajes, que puede alcanzar varios cientos de Gb de RAM y varios días o semanas de procesamiento. Las secuencias repetitivas son muy abundantes en casi todos los genomas, y dificultan la realización del ensamblaje, al crear bucles y bifurcaciones en el grafo. Para resolver correctamente estas estructuras es necesario usar lecturas de mayor longitud que las repeticiones o, alternativamente, secuenciar lecturas emparejadas (*paired-end reads* o *mate pairs*) que estén lo suficientemente alejadas como para permitir estimar el número de repeticiones y facilitar su ensamblaje.

### 3.3.2 El ensamblaje *de novo* de transcriptomas

El ensamblaje *de novo* de transcriptomas consiste en determinar la secuencia de nucleótidos de todos los transcritos de un organismo para el que no se dispone de un genoma o un transcriptoma de referencia. Esta técnica ha permitido caracterizar rápidamente el contenido en genes de muchas especies cuyos genomas no están secuenciados. Algunos programas utilizados para realizar el ensamblaje *de novo* de transcriptomas son Velvet-Oases, Trinity y TransABYSS. A las etapas propias de cualquier ensamblaje *de novo*, suele añadirse una etapa de procesamiento final cuyo cometido es reducir el número de transcritos redundantes o potencialmente incorrectos. A pesar de que los programas modernos incorporan procedimientos para reducir el impacto de este problema, los programas de ensamblaje *de novo* normalmente producen decenas o cientos de miles de transcritos o isoformas en los organismos superiores (Rana *et al.*, 2016). Este número tan elevado suele incluir secuencias con pautas de lectura abierta (*Open Reading Frames*; ORF) truncadas y transcritos quiméricos, problemas pueden ser parcialmente solventados mediante la aplicación de diferentes métodos bioinformáticos antes, durante y después del ensamblaje. Las modificaciones postranscripcionales, como el procesamiento alternativo de los intrones (*alternative splicing*) o la edición (*editing*) que experimentan los transcritos de

mitocondrias y cloroplastos, incrementan la complejidad de los grafos y añaden otra capa de complejidad al ensamblaje de los transcriptomas.

Existen muchos algoritmos para medir los niveles de expresión de transcritos y/o genes. Cuando se desarrolla un nuevo método, los autores evalúan su superioridad frente a los métodos ya existentes utilizando métricas que permiten evaluar las relaciones entre especificidad/precisión y sensibilidad/exactitud. Un estudio en el que se compararon 11 algoritmos (Kanitz *et al.*, 2015) encontró que todos los algoritmos funcionaban correctamente, sin que ninguno sobresaliese o produjese peores resultados que el resto. Sin embargo, los ensayos de exactitud se hicieron cuantificando los niveles de expresión absoluta, mientras que la mayoría de los estudios de RNA-seq suelen estar interesados en medidas relativas o de expresión diferencial. Además, los ensayos de especificidad realizados se basaron en la correlación de las medidas de experimentos replicados y parte de las mediciones se basaron en datos simulados por ordenador que no representan las fuentes de variación de los experimentos reales. Otro estudio posterior, en el que estos problemas fueron corregidos, resulta más útil para comparar distintos programas y produjo resultados parecidos (Teng *et al.*, 2016).

El ensamblaje *de novo* de transcriptomas es una tarea compleja, y no existe un método que funcione mejor en todos los casos. Además de los desafíos descritos en el apartado anterior, el ensamblaje de transcriptomas afronta la dificultad de que la cobertura no es uniforme, ya que distintos genes pueden expresarse a niveles muy diferentes. Una solución a este problema es el uso de múltiples valores de *k* en los ensamblajes realizados mediante estrategias basadas en grafos de De Bruijn. Disminuir el valor de *k* puede mejorar la calidad del ensamblaje de las secuencias de menor cobertura al permitir un mayor número de conexiones entre secuencias, si bien incrementa la complejidad del grafo y potencialmente aumenta el número de ensamblajes erróneos, por lo que es deseable emplear valores de *k* altos si la calidad y cantidad de las lecturas lo permite. Algunos programas (SOAPdenovo, TransAbyss y Trinity) descartan los cóntigos cuya cobertura es inferior a un umbral, de modo que purgan las secuencias más propensas a contener errores.

### 3.3.3 Ensamblajes guiados por una secuencia de referencia

Gracias a los esfuerzos de las últimas décadas, actualmente disponemos de genomas y transcriptomas de gran cantidad de especies. Estas secuencias pueden ser utilizadas como referencia para el estudio de los genomas de otras especies cercanas. Con este propósito, se han desarrollado algoritmos muy eficaces que permiten el alineamiento de las lecturas a una secuencia de referencia y su posterior ensamblaje.

Esta aproximación es mucho más rápida, requiere un menor número de lecturas y consume muchos menos recursos que los ensamblajes *de novo*. Los ensamblajes guiados por una secuencia de referencia utilizan un alineamiento de lecturas como punto de partida para realizar el ensamblaje, y pueden utilizarse tanto para secuenciar genomas nuevos, empleando como referencia el genoma de una especie próxima, como para realizar reensamblajes del mismo genoma. Estos reensamblajes son particularmente útiles para identificar mutaciones, genotipar los individuos de una especie, o caracterizar la variación genética existente en una población. Cuando la secuenciación se realiza a partir de moléculas de ARN, esta aproximación permite el ensamblaje del transcriptoma. Los ensamblajes guiados se realizan dividiendo las lecturas alineadas en bloques separados por regiones en las que no hay lecturas alineadas y ensamblando las lecturas de cada grupo por separado mediante algoritmos de ensamblaje, ya sea de OLC o basado en grafos de De Bruijn, generándose una colección de cóntigos que puede ser sometida a un proceso final de andamiaje o *scaffolding*.

Los ensamblajes guiados por una referencia se pueden llevar a cabo para 3 tipos de análisis: resecuenciación, ensamblaje del genoma de una especie cercana a otra con un genoma ya ensamblado, o análisis de transcriptomas. La resecuenciación consiste en secuenciar un genoma previamente caracterizado para genotipar los individuos de una población, identificar mutaciones o realizar estudios poblacionales y evolutivos. También se utiliza esta técnica para ensamblar el genoma de organismos genéticamente próximos a otros ya secuenciados. Aunque ambos organismos no sean genéticamente idénticos, se emplea el ensamblaje guiado por referencia para obtener una serie de fragmentos largos a partir de los cuales se intenta reconstruir el genoma completo. Por último, los ensamblajes guiados son también especialmente útiles en el campo de la transcriptómica. De forma tradicional, los niveles de expresión génica se han cuantificado mediante técnicas como la RT-qPCR o las micromatrices (*microarrays*). En las micromatrices, una colección de sondas de ADN correspondientes a genes cuya expresión se quiere medir se inmoviliza de forma ordenada sobre un soporte sólido. La hibridación a las sondas puede detectarse mediante el marcaje fluorescente de los transcritos, cuya intensidad permite medir el nivel de expresión de cada gen. Esta técnica, si bien es muy práctica y sensible, está limitada por el número de sondas de ADN que se pueden disponer sobre el soporte sólido. La secuenciación masivamente paralela de moléculas ARN de organismos que poseen un transcriptoma o un genoma de referencia permite alinear las lecturas obtenidas a los genes previamente anotados. Dado que el número de lecturas alineadas a cada gen representa una medida de su

nivel de expresión, no sólo permite determinar la secuencia de los genes de una muestra, sino también cuantificar su expresión. A diferencia de las micromatrices, el número de genes cuya expresión se puede medir no está limitado en este método, cuya principal dificultad reside en el análisis bioinformático de la cantidad masiva de datos generados.

### 3.3.4 El transcriptoma de *Allium sativum*

El primer transcriptoma corresponde a tejido de bulbos (Sun *et al.*, 2012) y fue secuenciado mediante la plataforma Illumina. Para el ensamblaje se utilizó el programa SOAPdenovo (Li *et al.*, 2010), y llevaron a cabo un proceso de realineamiento de las lecturas a los cóntigos ensamblados para agruparlos en andamios. El proceso de andamiaje fue repetido dos veces, obteniendo 127.933 secuencias únicas (*unigenes*), que compararon con las bases de datos del NCBI, Swiss-Prot y KEGG. Entre estas tres plataformas permitieron asignar funciones a 124.393 secuencias.

Tres años después se publicó un estudio que describía el ensamblaje de los transcriptomas de varios tejidos (Kamenetsky *et al.*, 2015), lo que permite catalogar los genes expresados en cada tejido y facilita el estudio de procesos de desarrollo y la identificación de genes coexpresados. También permite estudiar el patrón de expresión de genes individuales e identificar genes de referencia para experimentos de RT-PCR cuantitativa en tiempo real (RT-qPCR). El ensamblaje se realizó con el programa Trinity, obteniendo un conjunto de 239.116 cóntigos que designaron como “catálogo extenso del transcriptoma”. A diferencia del estudio anterior, no realizaron el andamiaje de los fragmentos ensamblados. Este “transcriptoma extenso” probablemente contiene un elevado número de fragmentos quiméricos generados por el programa de ensamblaje. Por ello, los autores realinearon las lecturas a los cóntigos del catálogo extenso y eliminaron aquellos fragmentos a los que se alineaban menos de 10 lecturas. El resultado fue un total de 102.042 cóntigos a los que llamaron “catálogo del transcriptoma abundante”, y en el que encontraron cerca de un 10% de secuencias redundantes. Estos valores son similares a los del artículo de 2012, pero mucho más fiables dado el número de lecturas y la tecnología de secuenciación utilizada. Mediante la comparación con los transcriptomas de varias especies de plantas, determinaron que el transcriptoma abundante tenía una similitud del 45-47% con los transcriptomas de *Oryza sativa* y *Arabidopsis thaliana*. Al comparar los resultados con el transcriptoma de *Allium sativum* previamente depositado, encontraron un porcentaje de similitud del 78%, aunque la longitud media de las lecturas ensambladas fue mucho mayor. La comparación con el transcriptoma de la cebolla resultó en un 51-58% de similitud. Por último, los autores

clasificaron los transcritos en función del tejido, distinguiendo entre hojas, bulbo, raíces, inflorescencias y tallo. Otros trabajos han permitido estudiar el transcriptoma durante el desarrollo de las flores y en el polen (Shemesh-Mayer *et al.*, 2015), buscar marcadores moleculares (Chand *et al.*, 2015; Havey and Ahn, 2016; Liu *et al.*, 2015), estudiar la base genética de la formación del bulbo (Zhu *et al.*, 2019) y analizar las rutas metabólicas implicadas en la síntesis de compuestos orgánicos sulfurados (Mehra *et al.*, 2020; Soorni *et al.*, 2021).

### 3.3.5 El genoma nuclear de *Allium sativum*

Desde hace más de dos décadas se sabe que los genomas del género *Allium* son generalmente mayores que los de muchos otros eucariotas, a través de la medición de sus pesos moleculares aproximados (Ohri *et al.*, 1998; Ricroch *et al.*, 2005). Aunque se conocen genomas más grandes, como el de *Neoceratodus forsteri* con más de 40 Gb (Otto, 2021), los genomas de las especies de *Allium* están entre los más grandes de las monocotiledóneas, especialmente *Allium sativum*, con un tamaño aproximado de unas 16,9 Gb y una configuración cromosómica 2n = 2x = 16 (Ricroch *et al.*, 2005). La principal dificultad para la secuenciación de los genomas del género *Allium* no es simplemente su gran tamaño, sino la enorme cantidad de secuencias y regiones repetitivas que presentan, que dificultan la realización de los ensamblajes *de novo*.

La primera versión del genoma completo de *Allium sativum* (Sun *et al.*, 2020), si bien no es completa, aporta gran información sobre el contenido en genes de esta especie. Para este ensamblaje utilizaron una combinación de tecnologías de secuenciación: SMRT de PacBio, ONT, lecturas emparejadas HiSeq de Illumina, librerías de 10xGenomics y secuenciación de captura de conformación de cromosomas de alta capacidad (Hi-C). Mediante PacBio, Illumina y 10xGenomics obtuvieron un total de 3,06 Tb de secuencias de alta calidad, lo que corresponde a una cobertura de 188 veces el tamaño del genoma, además de otras 252,5 Gb de secuencias largas de menor calidad obtenidas mediante secuenciación ONT para rellenar los huecos entre los cóntigos ensamblados y elucidar algunas de las regiones repetidas. El resultado fue el ensamblaje del genoma de 16,24 Gb, un tamaño muy similar al esperado. Posteriormente, se utilizaron 6,34 mil millones de secuencias adicionales para realizar el andamiaje de los cromosomas mediante el método Hi-C, obteniendo la secuencia de 8 supuestos cromosomas.

La anotación del genoma permitió identificar 57.561 genes que codifican proteínas, de los que se caracterizaron el 88%. En este mismo artículo los autores ensamblan su propio transcriptoma, en el que aproximadamente el 95% de las

secuencias ensambladas en ese transcriptoma estaban presentes en el ensamblaje con más del 50% de identidad. El análisis filogenético basado en el genoma nuclear de A. sativum lo situó dentro de la familia *Amaryllidaceae* y el orden *Asparagales*, como ya había sido clasificado previamente (Q.-Q. Li *et al.*, 2010; Xie *et al.*, 2020), siendo *Asparagus officinalis* la especie con un genoma secuenciado filogenéticamente más cercana. Según las estimaciones realizadas, *Asparagus officinalis* y *Allium sativum* compartieron un ancestro común hace 80,8 millones de años, lo que concuerda con los análisis filogenéticos de los genomas de plastos, que calculan que unos 40 millones de años después se formaría la familia *Amaryllidaceae* (Xie *et al.*, 2020).

El genoma del ajo contiene el porcentaje más alto de secuencias repetitivas que se ha visto en un genoma (91,3%, unas 14,8 Gb), de las que el 76% son elementos transponibles (TE). Los elementos transponibles pueden insertarse en las regiones codificantes de los genes o en sus secuencias reguladoras, afectando a su expresión (Chuong *et al.*, 2017). En el genoma del ajo se han encontrado 4.219 genes con TE insertados, muchos de los cuales parecen tener funciones asociadas a la reproducción sexual, lo que podría explicar la esterilidad de la planta. Además de a la expansión del número de repeticiones de elementos transponibles, la expansión de los genomas vegetales puede deberse a eventos de duplicación del genoma completo (*Whole Genome Duplication*; WGD; Bodt *et al.*, 2005). Se han identificado tres duplicaciones del genoma completo del ajo: ocurridas hace 120-130 millones de años (Magallón *et al.*, 2015), otra ocurrida hace 89,8 millones de años, probablemente relacionada con la aparición del ancestro común de los géneros *Allium* y *Asparagus*, y una última hace aproximadamente 18 millones de años. Dado que el género *Allium* apareció hace unos 40 millones de años (Xie *et al.*, 2020), es probable que esta última duplicación sea exclusiva del ajo, en línea con el gran tamaño de su genoma.

### 3.3.6 Los genomas cloroplásticos

Lynn Margulis propuso en 1967 que los orgánulos de las células eucariotas, como las mitocondrias y los cloroplastos, habían evolucionado a partir de bacterias endosimbióticas (Sagan, 1967). Los análisis bioquímicos posteriores confirmaron que los genes y las proteínas de mitocondrias y cloroplastos tenían un origen evolutivo distinto a los del núcleo eucariota. De hecho, los análisis filogenéticos sugieren que los hospedadores con los que establecieron simbiosis estos orgánulos, los “eucariotas primitivos”, estarían estrechamente relacionados con un grupo de arqueas, las arqueas Asgard (Zaremba-Niedzwiedzka *et al.*, 2017), mientras que las mitocondrias y los

cloroplastos provienen de proteobacterias y cianobacterias ancestrales, respectivamente (McCutcheon, 2016).

El cloroplasto es un orgánulo presente en las células vegetales que se encarga de sintetizar carbohidratos mediante la energía solar en la fotosíntesis. Aunque las mitocondrias y los cloroplastos poseen un origen diferente, presentan algunas similitudes, como una doble membrana externa, genoma y ribosomas propios, y un gran parecido en los procesos de la cadena de transporte electrónico. La membrana externa del cloroplasto es altamente permeable y regula el transporte de moléculas, el movimiento intracelular y la homeostasis del espacio intermembranal. La membrana interna, a diferencia de la de las mitocondrias, no está plegada y no contiene las proteínas de la cadena de transporte electrónico, es menos permeable que la membrana externa y delimita el espacio interno del cloroplasto conocido como estroma. Además de estas dos membranas, dentro del estroma los cloroplastos presentan un conjunto de sacos membranosos denominados tilacoides, que se agrupan en columnas denominadas grana. Dentro de los tilacoides se encuentra otro compartimento adicional, el espacio intertilacoidal, que es compartido por todos los tilacoides mediante conexiones entre estos. En la membrana de los tilacoides se encuentran los componentes de la cadena de transporte electrónico, y en el espacio intertilacoidal se almacena la clorofila y otras moléculas necesarias para los procesos fotosintéticos.

Los cloroplastos cumplen diversas funciones dependiendo del tipo celular. Los plastos se desarrollan a partir de los proplastos, precursores que se encuentran en las células de los meristemos de las plantas y se diferencian según el tipo celular, por lo que su desarrollo está determinado en gran parte por el genoma nuclear. En oscuridad, los proplastos se diferencian en etioplastos. Al ser expuestos a la luz, los etioplastos dan lugar a cloroplastos. Los leucoplastos tienen función de almacenaje y están presentes en algunos tejidos no fotosintéticos. Existen tres tipos de leucoplastos: los amiloplastos, que acumulan almidón; los oleoplastos, que almacenan grasas; y los proteinoplastos, que pueden acumular tanto almidón como proteínas cristalizadas o filamentosas. Además de estas funciones, los cloroplastos también intervienen en la síntesis de algunas moléculas, como purinas, pirimidinas, aminoácidos y ácidos grasos.

Los primeros genomas cloroplásticos secuenciados fueron los de *Marchantia polymorpha* y *Nicotiana tabacum* (Ohyama *et al.*, 1986; Shinozaki *et al.*, 1986). Desde entonces, se han secuenciado más de 3700 genomas de especies diferentes. La secuenciación del genoma de la cianobacteria *Synechocystis* sp. *strain* PCC6803 supuso un hito para la comprensión del origen de los genomas cloroplásticos (Kaneko

*et al.*, 1996). Las similitudes entre el ADNcp y el de estas bacterias contribuyó a la aceptación de la actual teoría endosimbiótica (Timmis *et al.*, 2004). Gracias al conocimiento de este genoma se ha podido evaluar la transferencia horizontal de genes desde el cloroplasto al núcleo, así como la pérdida de muchos otros. Se estima que más de 4.500 genes nucleares provienen del genoma de las cianobacterias primitivas (Martin *et al.*, 2002). De hecho, la mayoría de las más de 3.000 proteínas que actúan en el cloroplasto están codificadas en el núcleo, y solo una pequeña parte está codificada en el genoma del cloroplasto. Al igual que en la mitocondria, los genes del genoma cloroplástico suelen codificar proteínas de la cadena de transporte electrónico y componentes del ribosoma, probablemente debido a la necesidad de una regulación rápida y localizada de estas funciones esenciales.

El genoma del cloroplasto es una molécula circular, aunque algunos estudios de microscopía han observado estructuras ramificadas (Mower and Vickrey, 2017). Su tamaño varía ampliamente entre especies, desde los 15,5 kb de *Asarum minus* hasta los 521,1 kb de *Floydiella terrestris*, aunque usualmente ronda entre 120 y 170 kb. En las plantas terrestres, está muy conservado en cuanto a estructura, contenido y orden de los genes, y las diferencias principales radican en el número de genes que han sido exportados al núcleo (Shaw *et al.*, 2007). El genoma del cloroplasto consta de dos repeticiones invertidas (*Inverted Repeats*; IR), separadas por una región de copia única grande (*Large Single Copy*; LSC) y una región de copia única pequeña (*Small Single Copy*; SSC). Las IR están muy conservadas entre especies y tienen una longitud entre 20 y 25 kb (Mower and Vickrey, 2017). Debido al alto grado de conservación en la estructura del cloroplasto, contenido en genes, su modeo de herencia uniparental y el orden de los genes en la mayoría de las plantas terrestres, el genoma del cloroplasto se ha utilizado como una fuente de marcadores genéticos para la clasificación filogenética y la taxonomía molecular.

Algunos genes de los genomas cloroplásticos contienen intrones, que están muy conservados evolutivamente. El procesamiento de estos intrones es complejo, y se ha observado la existencia tanto de *cis-splicing*, que consiste en el corte de los intrones y el empalme de los exones de un mismo transcrito, como de *trans-splicing*, que implica el empalme de exones que pertenecen a transcritos diferentes y se transcribe a partir de genes distantes en el genoma (Takenaka *et al.*, 2013). Como en las bacterias, muchos genes del genoma del cloroplasto forman parte de operones, que se regulan por uno o más promotores y se transcriben en forma de moléculas de ARN policistrónico (Börner *et al.*, 2015). Los genes de un mismo operón pueden participar en diferentes funciones, como los genes que codifican las proteínas de la cadena de transporte

electrónico que se transcriben junto a algunas proteínas ribosómicas. Las diferencias en la expresión de los genes de un mismo operón pueden deberse a la presencia de promotores adicionales o a terminadores internos de la transcripción dentro del operón (Chi *et al.*, 2014; Zhelyazkova *et al.*, 2012). El genoma del cloroplasto forma estructuras compactas denominadas nucleoides, al igual que ocurre en la mitocondria (Morley *et al.*, 2019). Tanto la transcripción como el procesamiento del ARN tiene lugar en el nucleoide, donde se acumulan todas las proteínas relacionadas con estos procesos.

Los cloroplastos tienen dos tipos polimerasas: una codificada en el genoma del cloroplasto (denominada PEP; de *Plastid-Encoded RNA Polymerase*) y otra codificada en genoma nuclear (denominada NEP, de *Nuclear-Encoded RNA Polymerase*). PEP es una polimerasa de tipo bacteriano, probablemente heredada de la cianobacteria ancestral que dio lugar a los cloroplastos. Está formada por subunidades que actúan como un núcleo central y están codificadas en el ADNcp (genes *rpoA*, *rpoB*, *rpoC1* y *rpoC2*), proteínas asociadas a polimerasas (PAPs) y algunos factores sigma. Estas últimas proteínas asociadas son necesarias para que la polimerasa reconozca los promotores y están codificadas en el núcleo celular. Las PAPs también parecen ser necesarias para la transcripción y su regulación. La inactivación de los genes *PAP* provoca albinismo, una disminución de la actividad de PEP y un incremento de la actividad de NEP (Yu *et al.*, 2014). PEP se localiza en los nucleoides y está asociada a la membrana interna del cloroplasto.

Por otro lado, la polimerasa NEP es de tipo fágico, está presente en las angiospermas y solo se ha identificado una subunidad, de forma análoga a la ARN polimerasa del fago T7. Esta enzima es capaz de realizar todo el proceso de reconocimiento del promotor y la transcripción, siempre que el ADNcp esté superenrollado (Kühn *et al.*, 2007). La transcripción de una de las subunidades de PEP (*rpoB*) está controlada exclusivamente por NEP, por lo que esta polimerasa funciona como una capa más de control del núcleo sobre los plastos, ya que en el genoma nuclear también se codifican los factores sigma y las PAP. Las monocotiledóneas tienen una única NEP (denominada RPOTp), mientras que las dicotiledóneas tienen dos (denominadas RPOTp y RPOTmp). RPOTp es una polimerasa exclusiva de plastos, mientras que RPOTmp se localiza tanto en los cloroplastos como en las mitocondrias (Liere *et al.*, 2011). Existen tres tipos de genes u operones en función de si se transcriben por PEP (tipo I), por PEP y NEP (tipo II), o solo por NEP (tipo III). Los únicos genes regulados exclusivamente por un tipo de polimerasa son *rpoB* y *accD*, ambos con promotores de tipo III y transcritos por NEP (Zhelyazkova *et al.*, 2012). Si bien PEP parece ser transcripcionalmente más activa que NEP en los tejidos fotosintéticos, ambas

polimerasas están activas durante todas las fases del desarrollo de los cloroplastos y en todos los tipos de plastos. Por otra parte, RPROTmp parece ser más activa en las células jóvenes no verdes de distintos órganos, mientras que RPROTp es más activa en los tejidos verdes y con actividad fotosintética (Emanuel *et al.*, 2006).

### 3.3.7 Genomas cloroplásticos del género *Allium*

El primer genoma cloroplástico secuenciado del género *Allium* fue el de la cebolla (*Allium cepa* L.) (von Kohn *et al.*, 2013), realizado con el objetivo de investigar la base molecular de la esterilidad masculina citoplasmática (CMS; *cytoplasmic male sterility*). La CMS consiste en la incapacidad de producir polen fértil debido a problemas en la interacción entre el genoma nuclear y el citoplasma, donde se localizan las mitocondrias y los cloroplastos. Un hito en el estudio de la CMS fue la caracterización de la variedad Italian Red 13-54, que no producía polen (JoNes and Emsweller, 1938). Se ha demostrado que la infertilidad de esta variedad depende de la interacción entre el citoplasma y un alelo recesivo del locus *Ms*, que ha de hallarse en homocigosis, ya que la fertilidad se restaura en presencia de un alelo dominante de *Ms* (Jones, 1943). El análisis de restricción del genoma del cloroplasto permitió identificar polimorfismos asociados a la CMS en esta y otras variedades (Havey, 1993). La secuenciación y ensamblaje del genoma del cloroplasto de la cebolla se realizó, en primer lugar, con una variedad fértil y otra que presentaba esterilidad masculina (von Kohn *et al.*, 2013), empleando la tecnología 454 de Roche (pirosecuenciación). El resultado fue un genoma de 153.538 pb para la variedad fértil, y 153.355 pb para la variedad con CMS, con un contenido en G+C del 36,8% en ambos casos. La estructura y composición del genoma del cloroplasto de la cebolla resultaron ser similares a la de otras plantas, con dos regiones de copia única, una grande y otra pequeña (LSC y SSC), separadas por dos repeticiones invertidas (IRa e IRb).

El genoma del cloroplasto del ajo fue secuenciado y caracterizado en 2016 (Filyushin *et al.*, 2016), de forma casi simultánea a la descrita en esta Tesis, mediante un secuenciador Illumina HiSeq 1500. Para su secuenciación se utilizó la tecnología de Illumina, y el ensamblaje se realizó con el programa SPAdes. El genoma de la cebolla, secuenciado previamente, se utilizó como referencia para corregir este ensamblaje. Los genomas del cloroplasto de la cebolla y el ajo resultaron ser muy parecidos, tanto en tamaño como en el número y orden de los genes, pero con algunas deleciones e inserciones en las regiones intergénicas. El genoma cloroplástico del ajo contiene 134 genes, de los que 82 codifican proteínas, 6 son pseudogenes, 38 corresponden a moléculas de ARN de transferencia y 8 a ARN ribosómico.

La filogenia del género *Allium* ha sido estudiada en base a las secuencias disponibles en cada momento (Fu *et al.*, 2023; Huo *et al.*, 2019; Jin *et al.*, 2022; Namgung *et al.*, 2021; Xie *et al.*, 2020; Zhao *et al.*, 2023). En 2019 se realizó la secuenciación y ensamblaje de los genomas cloroplásticos de cuatro especies diferentes de *Allium* (*A. sativum*, *A. fistulosum*, *A. cepa* N, y *A. tuberosum*). Valiéndose de sus datos y de otros cinco genomas secuenciados por otros autores (*A. cepa* CMS-T, *A. cepa* CMS-S, *A. obliquum*, *A. prattii*, y *A. victorialis*) concluyeron que el genoma cloroplástico está muy conservado en el género *Allium*, con un contenido en genes y un orden prácticamente idéntico, y generaron un mapa de consenso del genoma del cloroplasto para este género (Huo *et al.*, 2019). Un análisis filogenético del género *Allium* basado en 22 genomas de plastos, así como de otros 142 genomas de especies cercanas permitió establecer que la familia *Amaryllidaceae* surgió aproximadamente hace 49 millones de años, y que el género *Allium* surgió poco después, hace aproximadamente 42 millones de años. (Xie *et al.*, 2020). La estructura, el tamaño, el contenido en G+C y el orden de los genes han permanecido prácticamente inalterados en los genomas de todo el género. El tamaño de estos genomas varía desde 145 kb a 160 kb, principalmente a causa de la diferente expansión de secuencias repetitivas y por las diferencias en el tamaño de las regiones repetidas respecto a las regiones de copia única, como sucede en otras angiospermas (Wu *et al.*, 2018). Los análisis filogenéticos diferenciaron claramente al género *Allium* de las especies utilizadas como grupo externo, *Agapanthus coddii* y *Acorus americanus*. La primera ramificación del género *Allium* lo divide en dos clados: por una parte, el clado que incluye a *A. prattii* y *A. victorialis*, que forman parte de un mismo linaje identificado previamente mediante el estudio de secuencias del espaciador interno transcrito (ITS; *internal transcribed spacer*; la secuencia que separa a los genes del ARN ribosómico en el ADN genómico) y de los genes *rps16* y *matK* (Abugalieva *et al.*, 2017; Q.-Q. Li *et al.*, 2010); y, por otra parte, un clado que incluye a *A. tuberosum*, *A. sativum*, *A. obliquum* y *A. cepa*.

Destaca especialmente la localización de algunos genes en los puntos donde se unen las diferentes regiones del genoma, que es distinta a la de los genomas de cloroplastos de *Agapanthus coddii* y *Acorus americanus*, utilizados como grupo externo. Los genes *ycf1a* e *ycf1b* se localizan entre las regiones IRb-SSC y SSC-IRa, respectivamente, mientras que *rpl22* se encuentra entre LSC-IRb. El gen *ycf1* codifica una proteína del complejo TIC (Nakai, 2015), que participa en el transporte de moléculas a través de las membranas del cloroplasto, y se utiliza como “código de barras” para distinguir los genomas cloroplásticos de plantas (Dong *et al.*, 2015). Sin embargo, recientemente se ha visto que algunas especies de *Allium* no presentan el gen *ycf1b*

entre las regiones IRb y SSC, como sucede en *A. dentigerum* y *A. eduardii* (Fu *et al.*, 2023). El gen *rpl22* codifica una proteína ribosómica y su posición entre LSC e IRb difiere entre genomas cloroplásticos. En las especies del subgénero *Cyathophora*, el gen *rpl22* se encontró a una distancia entre 29 y 273 pb de la unión entre LSC e IRb. Además, en el caso de *A. spicatum*, *A. mairei* y *A. kingdonii* el gen que se encuentra entre estas regiones es *rps19* (Yang *et al.*, 2020). Una característica de estos genomas es que *rps12* (*ribosomal protein S12*) está dividido, por lo que su expresión correcta requiere *trans-splicing*. Este fenómeno había sido observado anteriormente en otros genomas cloroplásticos del género *Allium*. La principal diferencia entre estos genomas radica en el contenido en pseudogenes, siendo casos excepcionales la pérdida de los genes *atpB*, *rbcL*, *trnL-UAA* e *ycf2* en *Allium prattii*, y la pérdida de *infA* en *Allium tuberosum*. Además, el contenido G+C resultó ser bastante similar, entre un 36,7% y un 37%, lo que concuerda con el observado en otras angiospermas.

### 3.3.8 Los genomas mitocondriales

La mitocondria es un orgánulo presente en las células eucarióticas y que tradicionalmente se ha considerado como la “central energética”. En los últimos años se ha demostrado las mitocondrias participan en procesos celulares críticos, como la catálisis de azúcares y ácidos grasos de cadena larga, la síntesis de algunos lípidos, la regulación de la apoptosis, señalización, diferenciación, envejecimiento celular, y la respiración celular (López-Otín *et al.*, 2013; Scheffler, 2008; Van Blerkom, 2011). Son orgánulos delimitados por una doble membrana que define cuatro compartimentos diferentes: la membrana externa, el espacio intermembrana, la membrana interna y la matriz. Las funciones de la membrana externa son múltiples, desde la comunicación con el núcleo hasta el transporte de proteínas o la movilidad de la propia mitocondria. La membrana interna está densamente plegada hacia la parte interna, y contiene una enorme cantidad de complejos proteicos de la cadena de transporte electrónico y ATP sintasas que controlan la respiración celular.

Por su origen bacteriano, las mitocondrias poseen su propio genoma y son capaces de replicarse. Sin embargo, los genomas mitocondriales han perdido la mayoría de sus genes y contienen solo algunos necesarios para su función. La mayoría de los genes de la proteobacteria ancestral se han perdido o han sido transferidos al genoma del hospedador a lo largo de la evolución de las mitocondrias (McCutcheon, 2016), lo que probablemente ha favorecido la interacción entre el núcleo y la mitocondria. El genoma mitocondrial con más genes pertenece a protistas flagelados del orden *Jakobida*, con un total de 66 genes codificantes (Burger *et al.*, 2013). Las mitocondrias

de otros eucariotas suelen contener un subconjunto de los genes encontrados en estos flagelados. Los genes más conservados en los genomas mitocondriales corresponden a componentes de la cadena de transporte electrónico (excepto el complejo II), ARN de transferencia y ARN ribosómico. Otros genes necesarios para la función mitocondrial, como los de las proteínas ribosómicas, la citocromo c oxidasa o los factores de elongación de la traducción, entre otros, se localizan unas veces en el núcleo y otras en la mitocondria. La variación en el contenido en genes de los genomas mitocondriales se debe principalmente a la transferencia horizontal de genes al núcleo, un proceso que ha ocurrido a lo largo de millones de años.

Existe una evolución paralela convergente entre los genomas mitocondrial y cloroplástico, ya que los genes esenciales en estos dos orgánulos son muy similares. Una teoría (Allen *et al.*, 2003; Allen, 1993) sugiere que los componentes clave de la cadena de transporte electrónico, tanto los que forman parte de la cadena fotosintética en cloroplastos como los de la respiración celular, permanecen retenidos en estos orgánulos debido a que su estequiometría y ensamblaje han de estar precisamente regulados. Alteraciones en la estequiometría de estos complejos podrían desequilibrar los procesos redox, causar pérdida de energía, reducir las reservas de ubiquinona, inducir estrés oxidativo y, en última instancia, provocar la muerte de la célula. Respecto al ARN ribosómico, se ha propuesto que el ensamblaje de los ribosomas mitocondriales está funcionalmente asociado a la regulación redox de la cadena de transporte electrónico, lo que explicaría su retención en los genomas de la mitocondria y el cloroplasto (Maier *et al.*, 2013).

La pérdida o ganancia de genes en las bacterias es esencial para su adaptación al medio, y la variación del contenido en genes de las mitocondrias también podría desempeñar un papel similar, como sugiere la presencia de ciertos genes poco usuales en algunos genomas mitocondriales. Recientemente, gracias a las técnicas de secuenciación masivamente paralela, se han secuenciado numerosos genomas mitocondriales, lo que ha proporcionado evidencia de que estos genes inusuales podrían estar involucrados en funciones atípicas para adaptarse al medio ambiente (Breton *et al.*, 2014).

La variación en el contenido en genes de las mitocondrias puede ocurrir por diversos motivos. La pérdida de genes ocurre en casi todos los organismos, pero varía entre especies, y conlleva la transferencia de genes mitocondriales al núcleo celular. Este proceso es complejo, y requiere de la existencia previa de una maquinaria proteica de importación en las membranas de la mitocondria (Schatz and Dobberstein, 1996). A

pesar de ello, la transferencia de genes ha ocurrido constantemente a lo largo de la evolución (Nugent and Palmer, 1991). La duplicación de genes codificantes también es un proceso observado a menudo, aunque las regiones y los genes duplicados pueden diferir sustancialmente entre especies. En ocasiones, algunos genes se duplican parcialmente generando pseudogenes, que permiten rastrear las posiciones en las que se han generado las duplicaciones durante la evolución (Bensasson *et al.*, 2001). En algunas especies se ha observado la adquisición de genes que no son típicamente mitocondriales, que han podido ser adquiridos mediante transferencia horizontal desde bacterias o virus (Bilewitch and Degnan, 2011; Pont-Kingdon *et al.*, 1995). En plantas se han descrito varios ejemplos de transferencia horizontal, como es el caso de algunas proteínas transportadoras de glicerol (Intrieri and Buiatti, 2001; Zardoya *et al.*, 2002). Además, en algunos casos, se han encontrado genes o fragmentos de genes procedentes del genoma del cloroplasto y el núcleo (Alverson *et al.*, 2011). En las mitocondrias de varias especies, se han identificado ORFs que no corresponden a ninguna proteína conocida y cuyo origen es incierto, pero que muestran un alto grado de conservación, lo que sugiere que podrían tener una función relevante (Kayal *et al.*, 2012). Por ejemplo, la ORF *gau* (*gene antisense ubiquitous*) se ha encontrado en la hebra complementaria del gen *cox1* en los genes mitocondriales de muchas plantas, hongos y animales, así como de alfa-proteobacterias, los ancestros de las mitocondrias (Faure *et al.*, 2011).

### 3.3.9 Genomas mitocondriales de plantas

Los genomas mitocondriales de los animales suelen ser muy compactos, con un tamaño comprendido entre 15 y 17 kb, y son circulares, con una alta densidad génica y sin intrones (Boore, 1999), aunque en algunos casos, como ciertos ARN de transferencia, puede haber diferencias entre especies (Gissi *et al.*, 2008). Fuera del reino animal, los genomas mitocondriales presentan una enorme diversidad en tamaño, estructura y contenido en genes.

Los genomas mitocondriales de plantas son generalmente más grandes, usualmente entre 200 y 700 kb, aunque se conocen genomas de más de 1 Mb, y pueden presentar diferentes arquitecturas, incluyendo genomas con topología circular, lineal y/o ramificada (Gualberto *et al.*, 2014). Estas estructuras alternativas pueden coexistir en una misma mitocondria y están en constante cambio debido a la recombinación que ocurre en secuencias repetidas (Backert *et al.*, 1997; Backert and Börner, 2000; Bendich, 1993; Dudareva *et al.*, 1988; Oldenburg and Bendich, 1996). Las repeticiones en los genomas mitocondriales de las plantas se han clasificado en tres tipos: largas

(>500 pb), que recombinan frecuentemente y son responsables de las diferentes configuraciones del genoma mitocondrial (Lonsdale *et al.*, 1984; Stern and Palmer, 1984); intermedias (50 – 500 pb), en las que la recombinación ocurre con mucha menos frecuencia; y cortas (<50 pb), en las que se pueden producir recombinaciones muy poco frecuentes que dan lugar al fenómeno del desplazamiento subestequiométrico (Abdelnoor *et al.*, 2003; Brieba, 2019; Gualberto and Newton, 2017). Se ha sugerido que las estructuras ramificadas podrían corresponder a intermediarios de la recombinación, supuestamente asociadas a procesos de replicación dependiente de recombinación (RDR), uno de los mecanismos utilizados por las mitocondrias de las plantas para replicar su genoma (Brieba, 2019). Además, se ha observado que la recombinación en los genomas mitocondriales es similar a la del genoma del bacteriófago T4, que también utiliza el mecanismo de RDR (Backert and Börner, 2000; Liu and Morrical, 2010).

El dinamismo estructural del genoma mitocondrial de las plantas puede ser extremo, como se observa en el género *Silene*, donde se han documentado especies con hasta 128 cromosomas mitocondriales (Wu *et al.*, 2015). Este género exhibe una enorme variación en el tamaño y la estructura de su genoma mitocondrial, lo que sugiere una rápida evolución impulsada por diversos mecanismos evolutivos. Las disparidades en la conformación del genoma mitocondrial de *Silene* indican que la recombinación podría desempeñar un papel clave en la evolución de los genomas mitocondriales de las plantas (Maréchal and Brisson, 2010; Sloan *et al.*, 2012). Los procesos de reparación del ADN y la deleción de secuencias asociadas a procesos de recombinación son las principales razones por las que estos eventos de recombinación pueden tener un impacto significativo sobre la estructura y la evolución de los genomas mitocondriales (Davila *et al.*, 2011; Devos *et al.*, 2002).

A pesar de sus diferentes tamaños, los genomas mitocondriales de diferentes especies no difieren mucho en el número de genes que contienen, normalmente entre 50 y 60, por lo que su densidad génica es muy baja. Las secuencias intergénicas muestran una gran variedad, con múltiples repeticiones dispersas a lo largo del genoma, intrones, pseudogenes y secuencias adquiridas de otros genomas, incluyendo secuencias de origen cloroplástico, nuclear, vírico o bacteriano (Alverson *et al.*, 2011). La expansión de estas regiones y las notables diferencias de tamaño del genoma mitocondrial entre distintas especies de plantas no tienen un significado biológico claro, sino que parecen ser una consecuencia de los múltiples eventos de recombinación y las mutaciones acumuladas en estos genomas (Gualberto *et al.*, 2014). Una hipótesis sugiere que el inusual tamaño de estos genomas podría relacionarse con la alta tasa de mutación de los genomas mitocondriales de las plantas. Esta teoría propone que un

mayor tamaño y complejidad en el genoma pueden hacerlo menos susceptible a mutaciones que afecten a sus genes clave (Woloszynska, 2010).

Aunque el contenido en genes del genoma mitocondrial de animales apenas varía (Boore, 1999), las diferencias son notables en el caso de las plantas. Estas divergencias no están necesariamente vinculadas con cambios en el tamaño del genoma mitocondrial, sino con la transferencia horizontal de genes desde la mitocondria al núcleo o desde el cloroplasto a la mitocondria. Por ejemplo, el genoma mitocondrial de *Arabidopsis thaliana* mide 367 kb y contiene 31 genes que codifican proteínas y 21 ARN de transferencia, mientras que el de *Silene noctiflora* mide 6.728 kb (unas 18 veces mayor) y contiene únicamente 26 genes que codifican proteínas y 3 ARN de transferencia. Dentro del género *Silene* encontramos genomas de varias megabases, como los de *Silene noctiflora* y *Silene conica*, con 6,7 Mb y 11,3 Mb, respectivamente, y otros más pequeños como los de *Silene vulgaris* y *Silene latifolia*, con 427 kb y 253 kb, respectivamente (Sloan *et al.*, 2012). La causa de estas diferencias no se comprende completamente, aunque algunos estudios han identificado múltiples pautas de lectura abierta cuya función y origen aún se desconocen, como *orf355* u *orf138*, entre otras (Hanson and Bentolila, 2004). En algunos casos, estas ORF se expresan, sus transcritos se editan y se traducen en proteínas, lo que puede afectar directa o indirectamente a la CMS (Sabar *et al.*, 2003). Además, al alinear las lecturas provenientes de RNA-seq al genoma mitocondrial, se encuentran muchas regiones intergénicas no descritas que muestran niveles de expresión relativamente altos. Esto no implica necesariamente que la causa fundamental del gran tamaño del genoma mitocondrial de las plantas sea la expresión de una mayor variedad de genes, pero sí sugiere que estos orgánulos podrían desempeñar funciones todavía desconocidas en las que participan genes por describir.

El desplazamiento subestequiométrico (*substoichiometric shifting*) constituye un fenómeno importante para comprender la dinámica de la estructura del genoma mitocondrial de las plantas, que se caracteriza por la coexistencia de moléculas de ADN subgenómicas a concentraciones inferiores a las del genoma mitocondrial. El desplazamiento subestequiométrico conlleva cambios en el número de copias de fragmentos del genoma mitocondrial (Janska *et al.*, 1998), generando heterogeneidad entre las mitocondrias de un mismo individuo (Smith and Chowdhury, 1991). La recombinación reversible en las repeticiones existentes en el genoma conduce a la formación de moléculas subgenómicas, lo que contribuye a modular la expresión de algunas partes del genoma mitocondrial y podría tener efectos sobre la fertilidad (Janska *et al.*, 1998). El desplazamiento subestequiométrico contribuye a modificar el nivel de expresión de los genes contenidos en las moléculas subgenómicas al causar un

incremento o una disminución del número de copias, lo que puede afectar al fenotipo de diversas formas, como la variegación de las hojas (Sakamoto *et al.*, 1996). Asimismo, se ha observado que algunas plantas que presentan CMS experimentan una reversión espontánea a la fertilidad que ha sido atribuida a cambios en la estequiometría de las moléculas subgenómicas; (Janska *et al.*, 1998; Smith and Chowdhury, 1991).

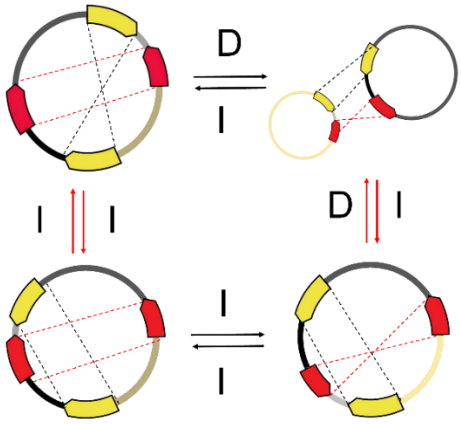



**Figura 1**. **Efecto de la recombinación en las repeticiones directas e invertidas presentes en un genoma mitocondrial.** La recombinación en las secuencias de ADN mitocondrial puede ocurrir en repeticiones directas (D) o invertidas (I). Las flechas en rojo indican la recombinación entre las repeticiones coloreadas en rojo, mientras que las flechas en negro señalan la recombinación entre las repeticiones coloreadas en amarillo. Si se produce recombinación entre repeticiones invertidas, la orientación de la región comprendida entre estas secuencias se invertirá. Si se produce recombinación entre repeticiones directas, la molécula de ADN se separará en dos moléculas de menor tamaño. Este fenómeno da lugar al desplazamiento subestequiométrico. Figura de elaboración propia basada en la Figura 2 de Khachaturyan et al., 2023

Los subgenomas puede estar presentes en un reducido número de copias, que puede ser inferior a 1 por cada 100 células en la población de mitocondrias de la planta (Arrieta-Montiel *et al.*, 2001). Los subgenomas mitocondriales se transmiten a las siguientes generaciones manteniendo la estequiometría del número de copias (Bellaoui *et al.*, 1998; Janska *et al.*, 1998; Sakai and Imamura, 1993). Se ha propuesto un modelo en el que existe un cromosoma mitocondrial principal, el denominado “círculo maestro”, que contiene la información de todos los subgenomas en los meristemos. Los subgenomas se diversifican mediante recombinaciones, transposiciones u otros procesos, y se separan cuando las mitocondrias se dividen (Arrieta-Montiel *et al.*, 2001; Woloszynska, 2010). Algunos genes nucleares afectan al desplazamiento subestequiométrico, como *OSB1* (Zaegel *et al.*, 2006), *Msh1* (Abdelnoor *et al.*, 2006), y *RecA* (Shedge *et al.*, 2007). Los genes nucleares *Rf* (*Restorer-of-fertility*) restauran la

fertilidad en algunas especies que presentan CMS, induciendo reordenamientos en el genoma mitocondrial y reduciendo el número de subgenomas autónomos que contienen una secuencia que induce la esterilidad citoplásmica *pvs* (Chase, 2007; Janska *et al.*, 1998; Mackenzie and Chase, 1990). Se ha especulado que el desplazamiento subestequiométrico y las copias subgenómicas pueden haber surgido como un repositorio de variación genética adicional que experimenta una evolución acelerada (Small *et al.*, 1989; Woloszynska, 2010).

### 3.3.10 Genomas mitocondriales del género *Allium*

A diferencia de los genomas cloroplásticos, los genomas mitocondriales presentan amplias diferencias interespecíficas. La mayoría de estas diferencias son de estructura y tamaño, principalmente debido a la expansión de regiones intergénicas y a cambios en el orden de los genes, aunque también debido a la transferencia de genes desde la mitocondria al núcleo y a la incorporación de fragmentos de ADN de diversos orígenes al genoma mitocondrial (Alverson *et al.*, 2011; Gualberto *et al.*, 2014; Gualberto and Newton, 2017; Rice *et al.*, 2013). Los frecuentes reordenamientos de los genomas mitocondriales podrían explicar la aparición de *trans-splicing* en estos orgánulos (Qiu and Palmer, 2004).

En la cebolla (*A. cepa*) se distinguen 3 fenotipos asociados a la fertilidad: uno de ellos fértil y 2 formas de CMS asociadas al genoma mitocondrial. La esterilidad de la variedad CMS-S puede ser restaurada por un solo alelo *Rf* dominante, mientras que la de la variedad CMS-T parece estar controlada por 3 loci diferentes (Jones, 1943; Kim and Yoon, 2010; Scheweisguth, 1973). Se ha propuesto que la esterilidad de CMS-S y CMS-T se relaciona con la expresión de un gen quimérico mitocondrial denominado *orf725*, probablemente de origen subgenómico, cuya expresión ha sido detectada en estos genotipos, pero no en las variedades fértiles (Kim *et al.*, 2009).

El primer genoma mitocondrial del género *Allium* secuenciado pertenecía a una variedad de cebolla con esterilidad de tipo CMS-S (Kim *et al.*, 2016) (Figura 2). La secuenciación se realizó a partir de una muestra de ADN total con la plataforma Illumina y el ensamblaje de las lecturas se llevó a cabo con el ensamblador CLC. Por su similitud con la secuencia del genoma mitocondrial de *Capsicum annuum*, se seleccionaron siete cóntigos, que se conectaron mediante PCR para finalmente ensamblar un genoma circular con 316.363 pb. Debido a la incorporación de secuencias del cloroplasto en el genoma mitocondrial, algunos fragmentos presentaron una gran similitud con secuencias cloroplásticas, que llega a alcanzar el 99% de identidad. La anotación del genoma mitocondrial de la cebolla reveló que contiene 24 genes que codifican proteínas,

que en su mayoría están presentes en todos los genomas mitocondriales de plantas. Como diferencias notables, el gen *ccmB* no pudo ser encontrado y el gen *ccmF*, que codifica un componente esencial en la biogénesis del citocromo C, estaba dividido en dos genes que se transcriben independientemente, denominados *ccmFn1* y *ccmFn2*, sin que se produzca *trans-splicing* entre ellos. Los autores indicaron que los transcritos del gen *cox2* experimentan *trans-splicing* (Kim *et al.*, 2013; Kim and Yoon, 2010), pero pasaron por alto que el gen *nad1* también lo experimenta, como había sido descrito previamente (Kim *et al.*, 2013).

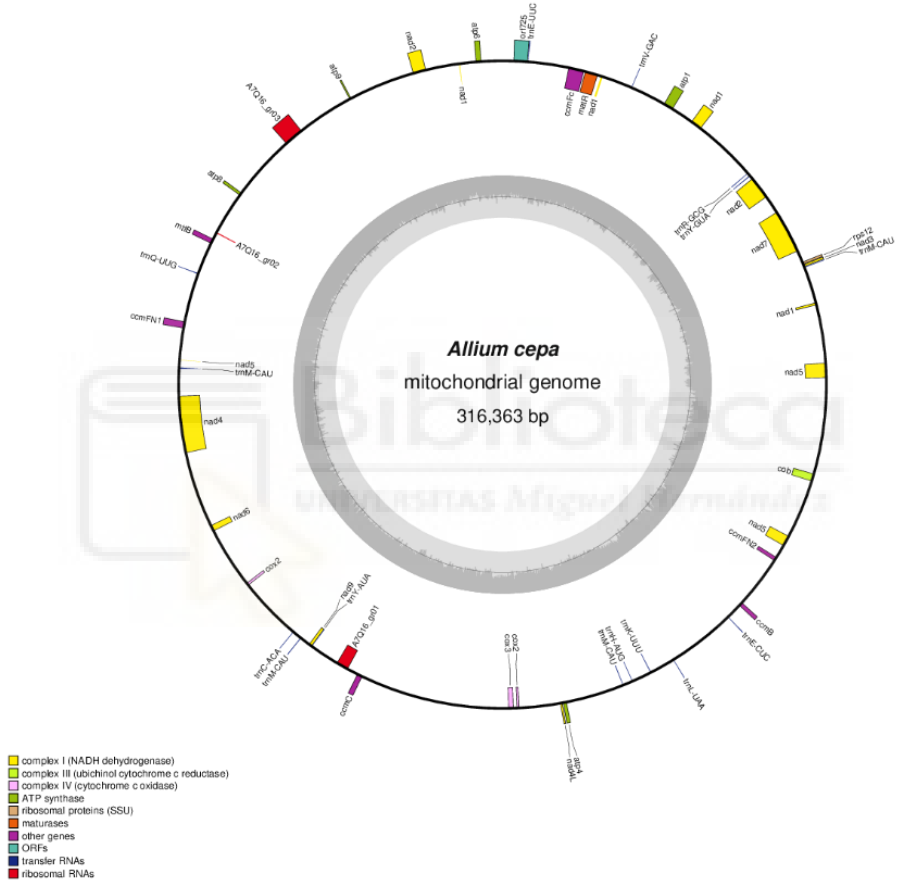



**Figura 2.** Mapa del genoma mitocondrial de *Allium cepa*. Los genes situados en la parte externa del círculo se transcriben en sentido 3'-5', mientras que los situados en la parte interna se transcriben en sentido 5'-3'. El círculo interno en gris muestra el contenido en G+C Imagen generada con OGDRAW utilizando la anotación de GenBank de la secuencia con número de acceso NC_030100.1.

Dado que la secuenciación se realizó a partir de una variedad CMS-S, los autores estudiaron la secuencia de *orf725*, una ORF presuntamente implicada en la esterilidad citoplásmica. Los autores encontraron un pseudogén quimérico de 586 pb fusionado en su extremo 3' con una copia del gen completo *cox1*, que codifica una

subunidad del citocromo C (Kim *et al.*, 2009). El hecho de que se trate de un gen quimérico refuerza la idea de que participa de algún modo en la esterilidad citoplásmica.

Posteriormente, se secuenciaron los genomas mitocondriales de variedades de tipo normal fértil y CMS-T (Kim *et al.*, 2019). A diferencia del ensamblaje de 2016, los autores no lograron ensamblar un genoma circular, sino un conjunto mínimo de 4 fragmentos prácticamente idénticos en ambos casos. El pseudogén *orf725* fue encontrado en la variedad de tipo CMS-T, pero no en la variedad fértil. El genoma mitocondrial de la variedad fértil de cebolla (339.180 pb) resultó ser ligeramente más grande que el de CMS-S (316.363 pb), aunque ambos presentaban los mismos genes codificantes. Sin embargo, se encontraron algunas secuencias de origen cloroplástico ausentes en CMS-S, y la posición de los genes era muy diferente. Lo mismo sucedió con el genoma de CMS-T, con la excepción de que contenía un fragmento adicional que contenía el *orf725*.

El otro genoma mitocondrial secuenciado del género *Allium* es el de *Allium fistulosum* (Xing *et al.*, 2023), que ha sido ensamblado como una única molécula circular al igual que el primer ensamblaje en cebolla. Los resultados obtenidos fueron parecidos al ensamblaje del genoma mitocondrial de *Allium cepa*, confirmando *trans-splicing* en el gen *cox2*. Sin embargo, el orden de los genes es radicalmente diferente entre estos dos genomas, como era esperable.

## 3.4 Métodos para el estudio de la función de los genes

La Genética clásica parte de la selección de fenotipos mutantes para, posteriormente, identificar los genes responsables de los mismos. Estos estudios suelen realizarse generalmente en especies modelo, que presentan tasas de reproducción aceleradas y pueden ser objeto de manipulación genética, como algunas levaduras, bacterias, nematodos o plantas. Aunque las mutaciones pueden ocurrir de manera espontánea, en esta aproximación se generan colecciones de mutantes mediante el uso de mutágenos físicos o químicos, que producen alteraciones en el material genético. La etilnitrosourea (ENU) y el metanosulfonato de etilo (EMS, por sus siglas en inglés) son mutágenos químicos que provocan mutaciones puntuales. En el caso del EMS, estos cambios son, mayoritariamente, sustituciones de guanina por adenina. Los mutágenos físicos, como distintos tipos de radiación ionizante (radiación gamma, rayos X, o partículas alfa y beta), también han sido ampliamente utilizadas para inducir cambios en el ADN. Los mutágenos físicos pueden provocar daños importantes en el ADN, como roturas de la doble cadena, inserciones y deleciones, traslocaciones y otras alteraciones cromosómicas. Una alternativa a estos métodos es la mutagénesis insercional, en la

que elementos transponibles (transposones) u otros fragmentos de ADN (como el ADN-T de *Agrobacterium tumefaciens*) se insertan al azar en el genoma, pudiendo causar cambios en la estructura o la expresión de los genes. Además, si se conoce la naturaleza de la secuencia insertada, la mutagénesis insercional es una herramienta que facilita la caracterización molecular de los genes mutados.

Tras seleccionar un mutante que presenta un fenotipo de interés, su caracterización genética y molecular permite obtener información valiosa sobre la función normal del afectado, particularmente en el caso de las mutaciones de pérdida de función. Algunos fenotipos son muy fáciles de detectar, como algunos caracteres observables a simple vista (por ejemplo, el albinismo) o detectables mediante el uso de un medio de cultivo mínimo (por ejemplo, las auxotrofías que causan un déficit de crecimiento en ausencia de un nutriente esencial). Otros fenotipos pueden ser muy difíciles de detectar en los cribados genéticos, como sucede con los genes que presentan redundancia funcional. Cuando dos mutantes recesivos presentan el mismo fenotipo, los ensayos de complementación permiten determinar si afectan al mismo gen.

La identificación molecular del gen afectado en un mutante inducido mediante mutagénesis insercional puede lograrse mediante la secuenciación del ADN adyacente a la inserción. La identificación de mutaciones inducidas mediante mutágenos físicos y químicos puede llevarse a cabo mediante diferentes aproximaciones, cuyo primer paso siempre es cartografiar la posición de la mutación mediante métodos basados en el análisis de ligamiento. Esta aproximación se basa en la observación de que los alelos de genes situados en posiciones próximas en el genoma adyacentes cosegregan con mayor frecuencia que los de genes situados a mayor distancia como consecuencia del ligamiento. Tradicionalmente se han utilizado estrategias para calcular la posición en el genoma del gen de interés mediante la detección de su ligamiento a marcadores moleculares polimórficos de distintos tipos, como los polimorfismos de un solo nucleótido (SNP), los polimorfismos en la longitud de fragmentos de restricción (RFLP), o los fragmentos de ADN polimórficos amplificados al azar (RAPD), entre otros. En las siguientes secciones presentamos algunas técnicas y conceptos relevantes para comprender el funcionamiento de MAPtools, una aplicación bioinformática que hemos desarrollado y presentamos en el Anexo 3.

### 3.4.1 Análisis de segregantes agrupados

En los organismos modelo utilizados en experimentación es posible realizar los cruzamientos necesarios para detectar el ligamiento entre loci. El análisis de segregantes agrupados (*bulked segregant análisis*; BSA) (Michelmore *et al.*, 1991) es

un método que ha sido ampliamente utilizado para la cartografía de ligamiento de loci asociados a rasgos fenotípicos de interés. Esta técnica se basa en la clasificación fenotípica de los individuos de una población segregante. El material genético de todos los individuos que presentan un mismo fenotipo se combina, estableciendo grupos (*bulks*) a partir de los individuos que presentan cada fenotipo. Las muestras establecidas a partir de individuos con fenotipos opuestos son genotipadas, lo que permite identificar marcadores ligados al carácter de interés.

Mientras que otros métodos se basan en el genotipado de todos y cada uno de los individuos de la población cartográfica, el BSA se basa en la comparación de la abundancia de los alelos de un locus entre una muestra preparada a partir de los individuos que presentan el fenotipo de interés y otra preparada a partir de los individuos que no lo presentan. Para un carácter monogénico recesivo, cuyos alelos designamos *A* (dominante) y *a* (recesivo), los individuos de una población $F_2$ pueden clasificarse en dos grupos en función de su fenotipo: el genotipo de todos los individuos que manifiesten el carácter será *a*/*a*, mientras que dos tercios de los individuos que presenten el fenotipo silvestre serán *A*/*a* y el tercio restante será *A*/*A*. La detección de un sesgo similar en la distribución de los alelos de cualquier otro marcador (fenotípico o molecular) puede ser atribuida al ligamiento entre el marcador y el gen responsable del carácter a estudio. En contraste, los alelos de cualquier marcador no ligado al carácter se distribuirán por igual en ambas muestras.

### 3.4.2 Cartografía mediante secuenciación

El abaratamiento de los métodos de secuenciación de segunda generación ha popularizado la resecuenciación de genomas completos. Los secuenciadores de segunda generación generan abundantes lecturas cortas que pueden ser alineadas a un genoma de referencia, lo que permite cuantificar los alelos de marcadores moleculares distribuidos a lo largo del genoma completo de un individuo. La cartografía mediante secuenciación (*mapping-by-sequencing*; MBS) es un método que permite identificar rápidamente polimorfismos ligados a un carácter de interés mediante una combinación de resecuenciación y análisis de segregantes agrupados. Este método puede aplicarse al estudio de caracteres fenotípicos que segregan en distintos tipos de poblaciones cartográficas, que pueden derivar de cruzamientos entre estirpes isogénicas (por ejemplo, un cruzamiento entre un mutante y la estirpe silvestre de la que deriva) o polimórficas (cuando las estirpes mutante y silvestre presentan diferente fondo genético), o incluso de cruzamientos entre especies próximas (Fonseca *et al.*, 2022).

La cartografía mediante secuenciación ha sido ampliamente utilizada en las plantas, ya que facilita la caracterización molecular de mutantes inducidos mediante mutagénesis física o química. Las plantas $M_1$ derivadas de un experimento de mutagénesis pueden ser autofecundadas para obtener poblaciones $M_2$, en la que es posible aislar mutaciones recesivas. Estos mutantes pueden ser utilizados en cruzamientos isogénicos o entre estirpes polimórficas para la preparación de poblaciones cartográficas. Alternativamente, se ha propuesto un método que permite detectar la mutación directamente agrupando las plantas de la generación $M_3$, sin necesidad de realizar cruzamientos (Fekih *et al.*, 2013). Esta variante del método ha sido puesta a punto por la comunidad científica del arroz y se denomina MutMap+.

El alineamiento de las lecturas a un genoma de referencia permite cuantificar los alelos de loci polimórficos comparando la secuencia del genoma de referencia con la secuencia de consenso producida a partir del alineamiento de las lecturas de una muestra. Para cualquier locus polimórfico no ligado al carácter a estudio, el número de lecturas portadoras de cada alelo estará en torno al 50% en ambas muestras. En caso de ligamiento absoluto, el 100% de las lecturas presentará el alelo proveniente del parental recesivo en la muestra establecida a partir de los individuos mutantes, mientras ambos alelos estarán presentes en la muestra establecida a partir de los individuos silvestres en proporción 2:1 (*A*:*a*). Cuando no existe un genoma de referencia, las lecturas pueden alinearse al genoma de otra especie filogenéticamente próxima, resecuenciando uno de los parentales para establecer el origen de cada alelo.

El análisis de los datos puede verse dificultado por la presencia de errores de secuenciación, alineamientos incorrectos, una baja cobertura en la secuenciación, o la mala calidad de las lecturas. Si la secuenciación tiene la suficiente calidad y cobertura, existe un genoma de referencia bien anotado, el mutante ha sido seleccionado correctamente, y la población cartográfica ha sido obtenida correctamente, el análisis de las frecuencias alélicas debería permitir determinar la posición de la mutación. Para descartar otras mutaciones no ligadas al carácter a estudio, pueden realizarse retrocruzamientos entre el mutante y el parental silvestre del que deriva, obteniendo así poblaciones con un menor número de mutaciones no asociadas al carácter de interés.

Para obtener una colección de marcadores polimórficos, el alineamiento de las lecturas es procesado mediante un programa de análisis de variantes (*variant calling*) que produce un archivo que almacena el recuento de los alelos de las posiciones polimórficas en ambas muestras. A partir de este recuento es posible calcular la frecuencia alélica del alelo derivado del parental recesivo, un parámetro denominado

SNP-index en la literatura. La identificación de los marcadores ligados al carácter de interés puede realizarse mediante diversos procedimientos: usando el SNP-index del grupo establecido a partir de los individuos con el fenotipo recesivo, usando la diferencia del SNP-index entre ambos grupos (un valor denominado ΔSNP-index), o mediante métodos estadísticos, como el test exacto de Fisher, descrito más adelante, corrigiendo los *p*-valores mediante el método de Bonferroni.

### 3.4.3 Cartografía de caracteres cuantitativos

Muchos caracteres de importancia en la agricultura son cuantitativos y presentan un modo de herencia complejo, estando controlados por numerosos genes que se localizan en diferentes regiones del genoma y que responden a las condiciones ambientales y a las interacciones con otros genes (Mackay *et al.*, 2009). Las regiones del genoma que están asociadas a caracteres cuantitativos de este tipo se denominan loci de caracteres cuantitativos (*quantitative trait loci*; QTL). Estos caracteres se han estudiado tradicionalmente mediante métodos de cartografía de QTL (*QTL mapping*) o mediante cartografía por asociación (*association mapping*). Dos limitaciones muy importantes de estos experimentos son el tiempo y coste necesarios para llevarlos a cabo. Con el abaratamiento de los métodos de secuenciación masivamente paralela estos métodos han decrecido en popularidad progresivamente en favor de alternativas como la cartografía de QTL mediante secuenciación (QTL-seq), método presentado en la siguiente sección.

Los métodos clásicos para cartografiar QTL requieren una población cartográfica y un mapa de ligamiento genético. Estos métodos son más complejos que los utilizados para cartografiar rasgos monogénicos, y generalmente son menos efectivos y menos precisos debido a la baja resolución de los mapas de ligamiento utilizados. Entre los métodos utilizados para la cartografía de QTL destacan la cartografía de intervalos (*interval mapping*), que permite detectar QTL en intervalo definidos por marcadores adyacentes de una población segregante $F_2$ (Lander and Botstein, 1989); la cartografía de intervalos compuesta (*composite interval mapping*), que combina el método anterior con métodos estadísticos de regresión múltiple (Zeng, 1994); la cartografía de intervalos múltiple (*multiple interval mapping*), que mejora los modelos anteriores mediante la detección simultánea de múltiples QTL (Kao *et al.*, 1999); o el uso de métodos estadísticos bayesianos (Satagopan *et al.*, 1996).

### 3.4.4 Cartografía de QTL mediante secuenciación

Al igual que la cartografía mediante secuenciación, el método QTL-seq también se basa en el análisis de segregantes agrupados, que se aplica a los individuos de una

población segregante obtenida a partir de cruzamientos entre líneas parentales altamente consanguíneas que manifiestan fenotipos extremos. La principal diferencia entre la cartografía mediante secuenciación y el método QTL-seq es que la primera se utiliza para cartografiar caracteres monogénicos discretos, mientras que el segundo se aplica a caracteres cuantitativos, que usualmente están determinados por los alelos de un elevado número de loci.

En el método QTL-seq, los grupos se establecen a partir de los individuos de la población segregante que presentan los fenotipos más extremos, que se clasifican como "high" (H, que incluye a los individuos que presentan mayor tamaño) y "low" (L, que incluye a los individuos más pequeños). Tras la secuenciación de las muestras H y L, los resultados de un experimento de QTL-seq se analizan como los de cartografía mediante secuenciación. Tras alinear las lecturas al genoma de referencia, el análisis de variantes permite seleccionar un conjunto de marcadores polimórficos y calcular las frecuencias alélicas del mismo modo que en la cartografía mediante secuenciación. Para detectar los QTL, se comparan las frecuencias alélicas de ambos grupos utilizando el incremento de las frecuencias alélicas, Δ(SNP-index), calculado como la diferencia de sus valores en los grupos H y L (Takagi *et al.*, 2013).

### 3.4.5 El test exacto de Fisher

El test exacto de Fisher es una prueba estadística utilizada para evaluar la asociación entre dos variables categóricas en una tabla de contingencia. En el contexto del análisis de ligamiento, este test se puede aplicar para evaluar si la segregación de los alelos en familias es consistente con la esperada bajo ciertos modelos genéticos. El test exacto de Fisher propone una hipótesis nula (la segregación de alelos es independiente del fenotipo) y alternativa (la segregación de los alelos está relacionada con el fenotipo), y calcula la probabilidad de observar los datos observados bajo la hipótesis nula (*p*-valor). En general, un *p*-valor inferior a un nivel de significación predeterminado proporciona evidencia en contra de la independencia de las variables y sugiere una asociación significativa. Debido a la realización de un elevado número de contrastes estadísticos, generalmente se aplica la corrección de Bonferroni a los datos, dividiendo el nivel de significación por el número total de comparaciones realizadas.

En el test de Fisher, se parte de una tabla de contingencia que contiene el recuento de los alelos:

| | Alelo 1 | Alelo 2 |
|---|---|---|
| Silvestre | $a$ | $b$ |
| Mutante | $c$ | $d$ |

La probabilidad exacta *p* de obtener por azar el recuento de la tabla, dados los totales de filas y columnas, puede calcularse como:

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}$$

El *p*-valor se calcula sumando las probabilidades correspondientes a resultados más extremos que el observado. Este *p*-valor es finalmente ajustado mediante la corrección de Bonferroni. Para *n* contrastes realizados con un nivel de significación α = 0,05 en cada contraste individual, el umbral corregido (t) viene dado por:

$$t = \frac{\alpha}{n}$$

### 3.4.6 Otros parámetros estadísticos

Otros métodos estadísticos que se han utilizado para la detección de mutaciones son el incremento de la frecuencia alélica (ΔSNP-index), el estadístico G y la distancia euclídea.

El ΔSNP-index (Takagi *et al.*, 2013) es la diferencia entre las frecuencia alélicas de los dos grupos de individuos. Su representación gráfica es especialmente útil para encontrar loci cuantitativos asociados a un carácter, ya que si en los grupos de individuos aparece una variante relacionada con el fenotipo el ΔSNP-index se desviará en uno u otro sentido. Si la variante encontrada no está relacionada con el fenotipo el ΔSNP-index será igual a 0.

El estadístico G (Magwene *et al.*, 2011) es una prueba estadística similar al test exacto de Fisher, en el que se compara la distribución observada frente a la esperada según un modelo hipotético. Se trata de una prueba flexible que puede adaptarse a diferentes situaciones y modelos, lo que lo hace óptimo para el análisis de datos genéticos. En el caso de los análisis de ligamiento, el estadístico G decrece rápidamente alrededor del locus asociado al fenotipo si lo comparamos con la frecuencia alélica, lo que se traduce en picos más acusados que permiten determinar la posición de las mutaciones o los QTL con mayor precisión. El estadístico G se calcula como:

$$G = 2\sum_{i=1}^{4}\left(a_i * log\left(\frac{a_i}{n_i}\right)\right)$$

Donde los valores $a_i$ y $n_i$ son los valores observados y esperados, respectivamente, del recuento del alelo *i*.

La distancia Euclídea (Hill *et al.*, 2013) es una medida geométrica que representa la distancia entre dos puntos de un espacio de *n* dimensiones. En el análisis de variantes se utiliza para medir la diferencia entre las muestras, considerando que cada variante es un punto con *n* dimensiones donde cada dimensión corresponde a un alelo diferente, generalmente *n*=2. El cálculo de la distancia euclídea es computacionalmente muy fácil de calcular y no requiere de los datos de los parentales. Se puede calcular fácilmente mediante la fórmula:

$$ED = \sqrt{(c-a)^2 + (d-b)^2}$$

Además de la distancia euclídea, en los estudios de cartografía mediante secuenciación y QTL-seq también se utiliza el valor de ED100[4] (Zhang *et al.*, 2019), que se calcula como la suma de la distancia euclídea de 100 marcadores adyacentes elevada a la cuarta potencia. El valor de este parámetro disminuye rápidamente alrededor del loci asociado a la mutación, lo que permite determinar su posición con mayor precisión.

Por último, otro parámetro utilizado para encontrar los loci asociados a las mutaciones es el denominado *boost* ($B_v$) del programa SHOREmap (Sun and Schneeberger, 2015), cuya representación gráfica genera un pico muy pronunciado en la región en la que se encuentra la mutación en un análisis de segregantes agrupados. Dada la media de la frecuencia alélica de una ventana determinada ($\theta_{obs}$) y un valor de frecuencia alélica objetivo, que generalmente es 1 ($\theta_{tar}$), el boost se calcula de la siguiente manera:

$$B_v = \frac{1}{|1 - \max(\theta_{tar},\ 1-\theta_{tar})/\max(\theta_{obs},\ 1-\theta_{obs})|}$$

# 4 Objetivos de la Tesis Doctoral

Como se ha descrito en la Introducción de esta Tesis, el ajo (*Allium sativum* L.) es una planta de gran importancia económica a nivel global, cuyo estudio mediante abordajes genéticos está severamente limitado por el modo de reproducción vegetativa de las variedades cultivadas. Gracias a la identificación de algunas variedades fértiles, capaces de producir semillas verdades mediante reproducción sexual, es previsible que el desarrollo de nuevas herramientas genéticas y genómicas disponibles para esta especie conduzca a su domesticación completa y su desarrollo como un cultivo moderno. En este contexto, los objetivos de mi Tesis Doctoral han sido:

- Evaluar el conocimiento actual y catalogar los recursos disponibles para la caracterización genética y genómica del ajo, así como identificar aquellos aspectos del cultivo que son susceptibles de mejora. Este objetivo ha sido abordado en la primera de las publicaciones aportadas.
- Determinar la secuencia de los genomas del cloroplasto y la mitocondria de una variedad de ajo económicamente importante, como paso previo al estudio de la infertilidad de esta especie. Este objetivo ha sido abordado en el segundo manuscrito aportado.
- Desarrollar una herramienta bioinformática versátil para cartografía de genes y QTL mediante secuenciación masivamente paralela. Este objetivo ha sido abordado en el tercer manuscrito aportado.

# 5 Resumen breve de los materiales y métodos utilizados

## 5.1 Revisión sobre el estado del arte y las herramientas genéticas y genómicas del ajo

Para evaluar la problemática actual del cultivo del ajo y la disponibilidad de herramientas genética y genómicas, realizamos búsquedas bibliográficas exhaustivas mediante el uso de referencias cruzadas y los buscadores de Google® Scholar (Google LLC., 2004) y de la base de datos PubMed (https://pubmed.ncbi.nlm.nih.gov/), a la que se accede desde el portal web del National Center for Biotechnology Information (NCBI). Para evaluar la disponibilidad de datos genómicos y transcriptómicos en bases de datos públicas, accedimos a la base de datos SRA (Sequence Read Archive; https://www.ncbi.nlm.nih.gov/sra), también disponible en la web del NCBI.

## 5.2 Secuenciación y ensamblaje de los genomas organulares del ajo

El ensamblaje de los genomas mitocondrial y cloroplástico se llevó a cabo a partir de una muestra de ADN genómico purificada a partir de la variedad Spring White con ayuda de un kit comercial (*GeneJET Plant Genomic DNA Purification Mini Kit*; Thermo Fisher Scientific). La purificación de ARN para su secuenciación se realizó con el kit *MagJET Plant RNA kit* (Thermo Fisher Scientific) a partir de tejidos de la misma variedad. En ambos casos se siguió el protocolo recomendado por el fabricante. Posteriormente, las muestras fueron secuenciadas en un secuenciador Illumina HiSeq 2500 por la empresa StabVida (Caparica, Portugal), utilizando protocolos de lecturas emparejadas y, en el caso de las muestras de ARN, con lecturas específicas de hebra (*strand-specific*). La calidad de las secuencias se comprobó con el programa FastQC (Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data).

El ensamblaje *de novo* de los genomas mitocondrial y cloroplástico se realizó con los programas Velvet versión 1.2.13 (Zerbino and Birney, 2008) y SPAdes (Bankevich *et al.*, 2012), con los parámetros de configuración indicados en el manuscrito. En el caso del genoma cloroplástico, se utilizaron los siguientes ajustes: longitud del *hash*: 51, longitud de inserción (-ins_length): 150, cobertura esperada (-exp_cov): 340 y umbral de cobertura (-cov_cutoff): 173. Las lecturas se alinearon de nuevo a la secuencia ensamblada utilizando el programa Bowtie2 (Langmead and Salzberg, 2012), y el alineamiento resultante se visualizó con el programa Tablet (Milne *et al.*, 2013) para corregir manualmente problemas potenciales de ensamblaje. Los

grafos del ensamblaje se visualizaron con Bandage (Wick *et al.*, 2015). Las lecturas de ARN se alinearon utilizando HiSat2 (Kim *et al.*, 2015).

Para el ensamblaje del genoma mitocondrial con Velvet, se estableció la longitud del hash en 65 pb, la cobertura esperada en 18 y el umbral de cobertura en 6. Estos ajustes se seleccionaron para descartar la mayoría de las secuencias de la fracción de copia única del genoma nuclear, manteniendo al mismo tiempo todas las secuencias derivadas de los genomas mitocondrial y cloroplástico. Para el ensamblaje SPAdes, se estableció un corte de cobertura de 6, igual que en el ensamblaje de Velvet. Se utilizó una longitud de hash que variaba de 65 a 71. El gráfico resultante fue más complejo que el de Velvet, y los cóntigos mitocondriales y cloroplásticos se encontraron vinculados a algunas secuencias genómicas nucleares. Utilizamos la implementación BLAST (Camacho *et al.*, 2009) en Bandage para aislar los cóntigos mitocondriales y cloroplásticos en los ensamblajes mitocondriales.

Para identificar genes codificantes de proteínas en los genomas mitocondrial y cloroplástico se realizaron búsquedas BLAST. Para la anotación mitocondrial, los genes se identificaron también con ayuda de la aplicación GeSeq (Tillich *et al.*, 2017), implementada en el sitio web CHLOROBOX. Para ambos genomas, los genes de ARN de transferencia (ARNt) se identificaron utilizando los programas ARAGORN (Laslett and Canback, 2004) y tRNAscan-SE 2.0 (Lowe and Eddy, 1997). Los ARN ribosómicos se anotaron utilizando el servidor RNAmmer 1.2 (Lagesen *et al.*, 2007). El mapa del genoma cloroplástico se dibujó a escala utilizando OGdraw (Greiner *et al.*, 2019).

## 5.3 Desarrollo de herramientas bioinformáticas para MBS y QTL-seq

El desarrollo del programa MAPtools ha sido realizado utilizando el lenguaje de programación Python 3.8. Para asegurar su accesibilidad y mantenibilidad a largo plazo el código fuente ha sido depositado en un repositorio de GitHub (https://github.com/hcandela/MAPtools), desde el que se distribuye bajo licencia GPL v3.0. Las dependencias del programa son librerías comúnmente usadas en la computación científica: docopt (v. 0.6.2) para la gestión de la línea de comandos, NumPy (v. 1.24.2) (Harris *et al.*, 2020) y SciPy (v. 1.10.1) (Virtanen *et al.*, 2020) para operaciones numéricas y estadísticas, pandas (v. 2.0.0) (McKinney, 2010) para manipulación de datos, biopython (v. 1.81) (Cock *et al.*, 2009) para tareas bioinformáticas y matplotlib (v. 3.7.1) (Hunter, 2007) para la creación de gráficos. MAPtools ha sido desarrollado y probado en un sistema equipado con dos procesadores Intel Xeon CPU E5-2620 v4 @ 2.10GHz (16 núcleos, 32 hilos) y 128 GB de RAM. La arquitectura del software permite su integración en flujos de trabajo bioinformáticos.

El funcionamiento del programa ha sido validado con datos reales descargados, principalmente, de la base de datos SRA, en flujos de trabajo con los programas Bowtie2 (versión 2.4.2), BWA (versión 0.7.17-r1188) (Li and Durbin, 2009), SAMtools (Li *et al.*, 2009) and BCFtools (versión 1.16) (Li, 2011) y GATK (versión 4.0.5.1) (McKenna *et al.*, 2010).

# 6 Discusión

En esta Tesis Doctoral hemos abordado el estudio del ajo (*Allium sativum* L.) tanto desde la perspectiva de su cultivo como de su genética. El cultivo del ajo es de gran importancia económica y presenta ciertas limitaciones debido a su modo de reproducción vegetativa: la ausencia de diversidad genética y fuentes de resistencia con las que afrontar la aparición de nuevas plagas, la acumulación de patógenos en los bulbos y la dificultad de generar nuevas variedades. Estos problemas han llevado al desarrollo de técnicas de propagación *in vitro* como el cultivo de meristemos, y a tratamientos complementarios como la termoterapia y la crioterapia. Estos métodos tienen un coste muy elevado y son muy lentos y laboriosos, lo que subraya la necesidad de encontrar soluciones más sostenibles y eficientes. Los esfuerzos para restaurar la fertilidad en el ajo son muy prometedores y podrían revolucionar la forma en la que se cultiva. La posibilidad de producir semillas verdaderas no solo mitigaría los problemas asociados con la propagación vegetativa, sino que también aceleraría los programas de mejora genética.

El estudio de la Genética del ajo es de vital importancia para diferenciar genéticamente las variedades existentes, para comprender los mecanismos genéticos y moleculares subyacentes a la infertilidad, y para para desarrollar nuevas variedades. Varios autores han generado colecciones de diferentes tipos de marcadores moleculares que pueden ser utilizados para genotipar las distintas variedades de ajo (Chand *et al.*, 2015; Havey and Ahn, 2016; Liu *et al.*, 2015). También se determinado la secuencia del transcriptoma, que ha permitido estudiar los patrones de expresión génica en diferentes órganos y tejidos de la planta (Kamenetsky *et al.*, 2015), la secuencia del genoma cloroplástico (Filyushin *et al.*, 2016) y, más recientemente, la secuencia completa de su genoma (Sun *et al.*, 2020). El enorme tamaño de este último (16,24 Gb, dividido en 8 cromosomas), sumado a su baja densidad génica y a la abundancia de secuencias repetitivas, dificulta en gran medida un estudio detallado. Además, la infertilidad del ajo impide la construcción de mapas genéticos y el estudio de variantes alélicas, aunque eso podría cambiar en los próximos años con la obtención de variedades fértiles (Etoh *et al.*, 1998; Pooler and Simon, 1994; Jenderek and Hannan, 2000; Kamenetsky *et al.*, 2005).

Los estudios realizados en la cebolla destacan la importancia del genoma mitocondrial en la determinación de características de gran interés agronómico, como la esterilidad masculina (Kim *et al.*, 2009).Con el objetivo de incrementar las herramientas genéticas disponibles para el estudio genético del ajo, en esta tesis doctoral hemos

secuenciado los genomas del cloroplasto y de la mitocondria de una importante variedad cultivada de esta especie.

El tamaño del genoma ensamblado del cloroplasto es de 153.131 pb, y se compone de una única molécula circular que comprende dos secuencias repetidas invertidas de 26.540 pb, separadas por dos regiones de copia única de 18.045 pb y 82.006 pb. Nuestra secuencia es igual en un 99% a otro ensamblaje del genoma del cloroplasto del ajo que se depositó casi a la vez que la generada por nosotros, y el contenido en genes es idéntico (Filyushin *et al.*, 2016). El ensamblaje del genoma de la mitocondria del ajo se realizó mediante el estudio exhaustivo del grafo producido por el programa Velvet, que contenía vértices correspondientes a los genomas cloroplástico y mitocondrial, ya que este último incorpora en su ADN fragmentos de origen cloroplástico y nuclear. Como se había visto en los genomas mitocondriales de otras plantas (Alverson *et al.*, 2011; Gualberto *et al.*, 2014; Gualberto and Newton, 2017; Qiu and Palmer, 2004), la estructura del genoma mitocondrial de ajo no es estática, sino que diferentes mitocondrias del mismo individuo pueden presentar conformaciones diferentes originadas por eventos de recombinación que se producen en secuencias repetitivas. Para reflejar este dinamismo estructural, optamos por representar el ensamblaje mediante un grafo compuesto por 21 vértices interconectados, que representan una secuencia cuya longitud total es de 536.232 pb. De estos vértices, al menos 6 deben corresponder a secuencias duplicadas. La anotación de este genoma rindió un total de 26 genes que codifican proteínas, 3 genes ribosómicos y 13 genes de ARNt. Encontramos 5 genes con exones en fragmentos separados que probablemente sufran *trans-splicing*. El gen *ccmFn* se encontró separado en dos fragmentos, *ccmFn1ccmFn1* y *ccmFn2*, como también ha sido descrito en el genoma mitocondrial de *Allium cepa* (Kim *et al.*, 2016). Realizamos un segundo ensamblaje del genoma mitocondrial mediante el programa SPAdes que, si bien presentaba algunas diferencias con el anterior ensamblaje, corroboró todos los caminos identificados en el grafo generado por Velvet.

Nuestro estudio del genoma mitocondrial del ajo indica que también están presentes secuencias de origen cloroplástico, generalmente pseudogenes, como ya había sido observado en prácticamente todos los genomas mitocondriales de plantas secuenciados (Rice *et al.*, 2013). La secuenciación del transcriptoma nos ha permitido estudiar la distribución de la edición en los genomas del cloroplasto y la mitocondria. Los transcritos derivados de estos genomas exhiben numerosas sustituciones de C por U, detectadas al alinear las lecturas del transcriptoma con las secuencias genómicas. Las transiciones de C a U representan prácticamente el 100% de los cambios

observados en secuencias codificantes, en algunos casos generando nuevos codones de inicio o de terminación de la traducción. Nuestras observaciones concuerdan con las de otros autores (Takenaka *et al.*, 2013, 2008) y con las observaciones en cebolla, donde la edición es crucial para crear codones de inicio y terminación de la traducción en varios genes (Tsujimura *et al.*, 2019).

Por último, hemos diseñado y desarrollado MAPtools, un programa escrito en Python3 y Java para realizar el análisis y la representación gráfica de datos de cartografía mediante secuenciación (MBS) y QTL-seq. El programa se compone de 5.117 líneas de código y está diseñado para ser utilizado en la línea de comandos de Unix. MAPtools se compone de 5 módulos que realizan diversas funciones: (1) **`mbs`** procesa y analiza los datos procedentes de un experimento de MBS. Acepta hasta dos muestras de secuenciación de grupos de individuos, unos mutantes y otros de fenotipo silvestre. Adicionalmente se le puede aportar hasta dos muestras de resecuenciación de los parentales, uno de fenotipo mutante y otro de fenotipo silvestre; o hasta dos muestras procedentes de la resecuenciación de un individuo externo con fondo genético idéntico a uno de los parentales. El resultado puede ser posteriormente representado gráficamente con `plot` o se pueden analizar las variantes de una región específica con `annotate`; (2) **`qtl`** procesa y analiza los datos procedentes de experimentos de QTL-seq. Acepta hasta dos muestras de secuenciación de grupos de individuos con fenotipos extremos y opuestos. Adicionalmente se le puede aportar la resecuenciación de uno de los parentales. El resultado puede ser representado gráficamente con `plot`; (3) **`merge`** permite reanalizar los resultados producidos por `mbs` o `qtl` mediante la agrupación del recuento de los alelos de varios marcadores adyacentes; (4) **`plot`** realiza la representación gráfica de los datos procesados por `mbs` o `qtl`; y **`annotate`**. Evalúa los efectos de las variantes analizadas con `mbs` y `qtl` en un intervalo especificado por el usuario.

Para validar la eficacia de MAPtools, hemos reanalizado los datos de experimentos de otros 11 autores, obteniendo en todos los casos resultados muy similares a los originales. MAPtools destaca por su capacidad para integrar datos de diferentes programas de alineamiento de secuencias y de *variant calling*, ofreciendo una gran flexibilidad en la elección de las herramientas y el flujo de trabajo. La compatibilidad con distintos tipos de poblaciones segregantes incrementa aún más la versatilidad de este programa. Esto permite a los usuarios adaptar el análisis a las características específicas de sus datos y a las particularidades de los organismos estudiados. Además, la similitud de la sintaxis de MAPtools con programas populares como SAMtools y

BCFtools facilita la adopción del software por parte de investigadores familiarizados con estas herramientas. Además, su distribución a través de un repositorio de GitHub asegura la accesibilidad y el mantenimiento a largo plazo del código, fomentando una comunidad de usuarios y desarrolladores que pueden contribuir a la mejora continua de la herramienta.

# 7 Conclusiones

## 7.1 El estudio de la genética de *Allium sativum* es necesario para su puesta a punto como un cultivo moderno

- Hemos realizado una revisión exhaustiva de la situación actual del cultivo del ajo, sus problemas como cultivo y sus amenazas. La infertilidad del ajo condiciona en gran medida su capacidad para responder a cambios externos, como condiciones climáticas adversas o la aparición de nuevas enfermedades.
- Hemos analizado las herramientas moleculares disponibles para la mejora genética de *Allium sativum*, que todavía son insuficientes para conseguir una domesticación completa de este cultivo.

## 7.2 Ensamblaje y anotación de los genomas organulares de *Allium sativum*

- Hemos ensamblado *de novo* la secuencia del genoma cloroplástico del ajo, con una longitud de 153.131 pb divididos en dos regiones de copia única de 82.006 pb y 18.045 pb, así como dos regiones repetidas invertidas de 26.540 pb cada una.
- Hemos anotado el genoma del cloroplasto de *Allium sativum*, con un total de 90 genes que codifican proteínas, 38 genes de ARN de transferencia, 8 genes de ARN ribosómico y 7 pseudogenes.
- Hemos determinado los genes que se transcriben conjuntamente en el cloroplasto y hemos anotado las modificaciones post-transcripcionales que se producen en el ARN cloroplástico.
- Hemos ensamblado el genoma mitocondrial de *Allium sativum*, que se compone de 21 fragmentos interconectados que pueden recombinar entre sí. Estas recombinaciones modifican el número de fragmentos de ADN en las mitocondrias y provoca reordenamientos en el genoma.
- Hemos realizado la anotación del genoma mitocondrial de *Allium sativum*, con un total de 26 genes que codifican proteínas, 13 genes de ARN de transferencia y 3 genes de ARN ribosómico.
- Hemos detectado que el gen mitocondrial *ccmFn* se encuentra dividido en dos genes: *ccmFn1* y *ccmFn2*, como había sido descrito previamente en la cebolla.

- Hemos encontrado que los exones de los genes *cox2*, *nad1*, *nad2* y *nad5* se encuentran separados en diferentes partes del genoma, lo que sugiere que podría haber *trans-splicing*.
- Hemos anotado las modificaciones post-transcripcionales que se producen en el ARN mitocondrial del ajo. Algunas de estas modificaciones generan codones de inicio y de terminación de la traducción.

## 7.3 MAPtools es una herramienta sólida para el análisis de datos de cartografía mediante secuenciación y QTL-seq

- Hemos diseñado y desarrollado MAPtools, un software de código abierto para el análisis de experimentos de cartografía mediante secuenciación y cartografía de QTL mediante secuenciación (QTL-seq), escrito en lenguaje Python.
- Hemos implementado una función para generar gráficos de los datos procesados de MBS y QTL-seq.
- Hemos desarrollado una función dentro de MAPtools para anotar las variantes relevantes encontradas en un rango determinado para los experimentos de MBS. Este comando genera una lista que informa de la variante, la posición, el tipo de secuencia al que pertenece y si esa alteración puede ser sinónima o no.
- Hemos utilizado MAPtools para reanalizar los datos de 12 experimentos realizados por 10 autores diferentes, y hemos identificado los loci y las mutaciones asociadas a los fenotipos descritas por los autores.
- MAPtools ha sido desarrollado con un diseño modular, en el que cada función es independiente, pero comparten la misma estructura de datos. Este diseño facilita la incorporación de nuevos comandos.

# 8 Bibliografía


Abdelnoor,R.V. *et al.* (2006) Mitochondrial Genome Dynamics in Plants and Animals: Convergent Gene Fusions of a MutS Homologue. *J. Mol. Evol.*, **63**, 165–173.

Abdelnoor,R.V. *et al.* (2003) Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 5968–5973.

Abugalieva,S. *et al.* (2017) Taxonomic assessment of *Allium* species from Kazakhstan based on ITS and matK markers. *BMC Plant Biol.*, **17**, 258.

Allen,J. F. *et al.* (2003) The function of genomes in bioenergetic organelles. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **358**, 19–38.

Allen,J.F. (1993) Control of Gene Expression by Redox Potential and the Requirement for Chloroplast and Mitochondrial Genomes. *J. Theor. Biol.*, **165**, 609–631.

Alverson,A.J. *et al.* (2011) Origins and Recombination of the Bacterial-Sized Multichromosomal Mitochondrial Genome of Cucumber. *Plant Cell*, **23**, 2499–2513.

Ardui,S. *et al.* (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.*, **46**, 2159–2168.

Arrieta-Montiel,M. *et al.* (2001) Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. *Genetics*, **158**, 851–864.

Ashton,P.M. *et al.* (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.*, **33**, 296–300.

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.

Backert,S. *et al.* (1997) The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends Plant Sci.*, **2**, 477–483.

Backert,S. and Börner,T. (2000) Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant *Chenopodium album* (L.). *Curr. Genet.*, **37**, 304–314.

Baichoo,S. and Ouzounis,C.A. (2017) Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Biosystems*, **156–157**, 72–85.

Bankevich,A. *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.*, **19**, 455–477.

de la Bastide,M. and McCombie,W.R. (2007) Assembling Genomic DNA Sequences with PHRAP. *Curr. Protoc. Bioinforma.*, **17**, 11.4.1-11.4.15.

Bellaoui,M. *et al.* (1998) Low-copy-number molecules are produced by recombination, actively maintained and can be amplified in the mitochondrial genome of *Brassicaceae*: relationship to reversion of the male sterile phenotype in some cybrids. *Mol. Gen. Genet. MGG*, **257**, 177–185.

Bendich,A.J. (1993) Reaching for the ring: the study of mitochondrial genome structure. *Curr. Genet.*, **24**, 279–290.

Bensasson,D. *et al.* (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.*, **16**, 314–321.

Bilewitch,J.P. and Degnan,S.M. (2011) A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. *BMC Evol. Biol.*, **11**, 228.

Bodt,S.D. *et al.* (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.*, **20**, 591–597.

Boore,J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767–1780.

Borlinghaus,J. *et al.* (2014) Allicin: Chemistry and Biological Properties. *Molecules*, **19**, 12591–12618.

Börner,T. *et al.* (2015) Chloroplast RNA polymerases: Role in chloroplast biogenesis. *Biochim. Biophys. Acta BBA - Bioenerg.*, **1847**, 761–769.

Branton,D. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.

Breton,S. *et al.* (2014) A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. *Trends Genet.*, **30**, 555–564.

Brieba,L.G. (2019) Structure–Function Analysis Reveals the Singularity of Plant Mitochondrial DNA Replication Components: A Mosaic and Redundant System. *Plants*, **8**, 533.

Burger,G. *et al.* (2013) Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists. *Genome Biol. Evol.*, **5**, 418–438.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chand,S.K. *et al.* (2015) Mining, characterization and validation of EST derived microsatellites from the transcriptome database of Allium sativum L. *Bioinformation*, **11**, 145–150.

Chase,C.D. (2007) Cytoplasmic male sterility: a window to the world of plant mitochondrial–nuclear interactions. *Trends Genet.*, **23**, 81–90.

Chi,W. *et al.* (2014) RHON1 Mediates a Rho-Like Activity for Transcription Termination in Plastids of Arabidopsis thaliana[C][W]. *Plant Cell*, **26**, 4918–4932.

Chuong,E.B. *et al.* (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.*, **25**, 1422–1423.

Cormen,T. *et al.* (2009) Book: introduction to algorithms MIT Press.

Davila,J.I. *et al.* (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. *BMC Biol.*, **9**, 64.

Denisov,G. *et al.* (2008) Consensus generation and variant detection by Celera Assembler. *Bioinformatics*, **24**, 1035–1040.

Devos,K.M. *et al.* (2002) Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Res.*, **12**, 1075–1079.

Dijk,E.L. van *et al.* (2018) The Third Revolution in Sequencing Technology. *Trends Genet.*, **34**, 666–681.

Dong,W. *et al.* (2015) *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.*, **5**, 8348.

Dudareva,N.A. *et al.* (1988) Structure of the mitochondrial genome of *Beta vulgaris* L. *Theor. Appl. Genet.*, **76**, 753–759.

Eid,J. *et al.* (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133–138.

Emanuel,C. *et al.* (2006) Development- and tissue-specific expression of the *RpoT* gene family of Arabidopsis encoding mitochondrial and plastid RNA polymerases. *Planta*, **223**, 998–1009.

Etoh,T. *et al.* (1998) Seed productivity and germinability of various garlic clones collected in Soviet Central Asia.

Faure,E. *et al.* (2011) Probable presence of an ubiquitous cryptic mitochondrial gene on the antisense strand of the cytochrome oxidase I gene. *Biol. Direct*, **6**, 56.

Fekih,R. *et al.* (2013) MutMap+: Genetic Mapping and Mutant Identification without Crossing in Rice. *PLoS ONE*, **8**, e68529.

Filyushin,M.A. *et al.* (2016) The complete plastid genome sequence of garlic *Allium sativum* L. *Mitochondrial DNA Part B Resour.*, **1**, 831–832.

Fonseca,R. *et al.* (2022) A Tomato EMS-Mutagenized Population Provides New Valuable Resources for Gene Discovery and Breeding of Developmental Traits. *Plants*, **11**, 2453.

Fu,X. *et al.* (2023) Phylogeny and adaptive evolution of subgenus *Rhizirideum* (*Amaryllidaceae*, *Allium*) based on plastid genomes. *BMC Plant Biol.*, **23**, 70.

Gissi,C. *et al.* (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity*, **101**, 301–320.

Gleizes,A. and Hénaut,A. (1994) A global approach for contig construction. *Bioinformatics*, **10**, 401–408.

Greiner,S. *et al.* (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.*, **47**, W59–W64.

Gualberto,J.M. *et al.* (2014) The plant mitochondrial genome: Dynamics and maintenance. *Biochimie*, **100**, 107–120.

Gualberto,J.M. and Newton,K.J. (2017) Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. *Annu. Rev. Plant Biol.*, **68**, 225–252.

Hackl,T. *et al.* (2014) proovread : large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**, 3004–3011.

Hanson,M.R. and Bentolila,S. (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *Plant Cell*, **16**, S154–S169.

Harris,C.R. *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.

Havey,M.J. (1993) A putative donor of S-cytoplasm and its distribution among open-pollinated populations of onion. *Theor. Appl. Genet.*, **86**, 128–134.

Havey,M.J. and Ahn,Y.-K. (2016) Single Nucleotide Polymorphisms and Indel Markers from the Transcriptome of Garlic. *J. Am. Soc. Hortic. Sci.*, **141**, 62–65.

Hill,J.T. *et al.* (2013) MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq. *Genome Res.*, **23**, 687–697.

Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, **9**, 90–95.

Huo,Y. *et al.* (2019) Complete chloroplast genome sequences of four *Allium* species: comparative and phylogenetic analyses. *Sci. Rep.*, **9**, 12250.

Intrieri,M.C. and Buiatti,M. (2001) The Horizontal Transfer of *Agrobacterium rhizogenes* Genes and the Evolution of the Genus *Nicotiana*. *Mol. Phylogenet. Evol.*, **20**, 100–110.

Janska,H. *et al.* (1998) Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. *Plant Cell*, **10**, 1163–1180.

Jenderek,M. and Hannan,R.M. (2000) Seed producing ability of garlic (*Allium sativum* L.) clones from two public US collections. *Proc. Third Int. Symp. Edible Alliaceae Athens Ga. USA*, 73–75.

Jin,Y. *et al.* (2022) Comparative and phylogenetic analysis of the complete chloroplast genome sequences of *Allium mongolicum*. *Sci. Rep.*, **12**, 21676.

Jo,M.H. *et al.* (2012) Classification of genetic variation in garlic (*Allium sativum* L.) using SSR markers. *Aust. J. Crop Sci.*, **6**, 625–631.

Jones,H.A. (1943) Inheritance of male-sterility in the onion and the production of hybrid seed. In, *Proc. Amer. Soc. Hort. Sci.*, pp. 189–194.

Jones,H.A. and Emsweller,S. (1938) A male-sterile onion.

Kamenetsky,R. *et al.* (2005) Diversity in fertility potential and organo-sulphur compounds among garlics from Central Asia. *Biodivers. Conserv.*, **14**, 281–295.

Kamenetsky,R. *et al.* (2004) Garlic (*Allium sativum* L.) and its wild relatives from Central Asia: Evaluation for fertility potential. *Acta Hortic.*, 83–91.

Kamenetsky,R. *et al.* (2015) Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum* L.). *BMC Genomics*, **16**, 12.

Kamenetsky,R. and Rabinowitch,H.D. (2001) Floral development in bolting garlic. *Sex. Plant Reprod.*, **13**, 235–241.

Kaneko,T. *et al.* (1996) Sequence Analysis of the Genome of the Unicellular Cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence Determination of the Entire Genome and Assignment of Potential Protein-coding Regions. *DNA Res.*, **3**, 109–136.

Kanitz,A. *et al.* (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.*, **16**, 150.

Kao,C.H. *et al.* (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.

Kayal,E. *et al.* (2012) Evolution of Linear Mitochondrial Genomes in Medusozoan Cnidarians. *Genome Biol. Evol.*, **4**, 1–12.

Khachaturyan,M. *et al.* (2023) Worldwide Population Genomics Reveal Long-Term Stability of the Mitochondrial Genome Architecture in a Keystone Marine Plant. *Genome Biol. Evol.*, **15**, evad167.

Kim,B. *et al.* (2016) Completion of the mitochondrial genome sequence of onion (Allium cepa L.) containing the CMS-S male-sterile cytoplasm and identification of an independent event of the *ccmFN* gene split. *Curr. Genet.*, **62**, 873–885.

Kim,B. *et al.* (2019) Identification of a gene responsible for cytoplasmic male-sterility in onions (*Allium cepa* L.) using comparative analysis of mitochondrial genome sequences of two recently diverged cytoplasms. *Theor. Appl. Genet.*, **132**, 313–322.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kim,S. *et al.* (2009) Identification of a novel chimeric gene, *orf725*, and its use in development of a molecular marker for distinguishing among three cytoplasm types in onion (*Allium cepa* L.). *Theor. Appl. Genet.*, **118**, 433–441.

Kim,S. *et al.* (2013) Origin of three characteristic onion (*Allium cepa* L.) mitochondrial genome rearrangements in *Allium* species. *Sci. Hortic.*, **157**, 24–31.

Kim,S. and Yoon,M.-K. (2010) Comparison of mitochondrial and chloroplast genome segments from three onion (*Allium cepa* L.) cytoplasm types and identification of a trans-splicing intron of *cox2*. *Curr. Genet.*, **56**, 177–188.

von Kohn,C. *et al.* (2013) Sequencing and annotation of the chloroplast DNAs and identification of polymorphisms distinguishing normal male-fertile and male-sterile cytoplasms of onion. *Genome*, **56**, 737–742.

Kosugi,S. *et al.* (2015) GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics*, **31**, 3733–3741.

Kühn,K. *et al.* (2007) Arabidopsis Phage-Type RNA Polymerases: Accurate in Vitro Transcription of Organellar Genes. *Plant Cell*, **19**, 959–971.

Lagesen,K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

Lander,E.S. and Botstein,D. (1989) Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, **121**, 185–199.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Laslett,D. and Canback,B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.

Lee,H. *et al.* (2014) Error correction and assembly complexity of single molecule sequencing reads. 006395.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.

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

Li,Q.-Q. *et al.* (2010) Phylogeny and biogeography of *Allium* (*Amaryllidaceae*: *Allieae*) based on nuclear ribosomal internal transcribed spacer and chloroplast rps16

sequences, focusing on the inclusion of species endemic to China. *Ann. Bot.*, **106**, 709–733.

Li,R. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**, 265–272.

Li,Z. *et al.* (2012) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief. Funct. Genomics*, **11**, 25–37.

Liere,K. *et al.* (2011) The transcription machineries of plant mitochondria and chloroplasts: Composition, function, and regulation. *J. Plant Physiol.*, **168**, 1345–1360.

Liu,J. and Morrical,S.W. (2010) Assembly and dynamics of the bacteriophage T4 homologous recombination machinery. *Virol. J.*, **7**, 357.

Liu,T. *et al.* (2015) Large-scale development of expressed sequence tag-derived simple sequence repeat markers by deep transcriptome sequencing in garlic (*Allium sativum* L.). *Mol. Breed.*, **35**, 204.

Lonsdale,D.M. *et al.* (1984) The physical map and organisation of the mitochondrial genome from the fertile cytoplasm of maize. *Nucleic Acids Res.*, **12**, 9249–9261.

López-Otín,C. *et al.* (2013) The Hallmarks of Aging. *Cell*, **153**, 1194–1217.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Mackay,T.F.C. *et al.* (2009) The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.*, **10**, 565–577.

Mackenzie,S.A. and Chase,C.D. (1990) Fertility Restoration Is Associated with Loss of a Portion of the Mitochondrial Genome in Cytoplasmic Male-Sterile Common Bean. *Plant Cell*, **2**, 905–912.

Magallón,S. *et al.* (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.*, **207**, 437–453.

Magwene,P.M. *et al.* (2011) The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. *PLOS Comput. Biol.*, **7**, e1002255.

Maier,U.-G. *et al.* (2013) Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes. *Genome Biol. Evol.*, **5**, 2318–2329.

Maréchal,A. and Brisson,N. (2010) Recombination and the maintenance of plant organelle genome stability. *New Phytol.*, **186**, 299–317.

Martin,W. *et al.* (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci.*, **99**, 12246–12251.

Mathew,D. *et al.* (2011) Effect of long photoperiod on the reproductive and bulbing processes in garlic (Allium sativum L.) genotypes. *Environ. Exp. Bot.*, **71**, 166–173.

McCutcheon,J.P. (2016) From microbiology to cell biology: when an intracellular bacterium becomes part of its host cell. *Curr. Opin. Cell Biol.*, **41**, 132–136.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

McKernan,K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.

McKinney,W. (2010) Data Structures for Statistical Computing in Python., pp. 56–61.

Mehra,R. *et al.* (2020) Transcriptome analysis of Snow Mountain Garlic for unraveling the organosulfur metabolic pathway. *Genomics*, **112**, 99–107.

Meltz Steinberg,K. *et al.* (2017) Building and Improving Reference Genome Assemblies. *Proc. IEEE*, **105**, 422–435.

Michelmore,R.W. *et al.* (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 9828–9832.

Milne,I. *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.*, **14**, 193–202.

Morley,S.A. *et al.* (2019) Plant Organelle Genome Replication. *Plants*, **8**, 358.

Mower,J. and Vickrey,T. (2017) Structural Diversity Among Plastid Genomes of Land Plants. In, *Advances in Botanical Research*.

Nakai,M. (2015) The TIC complex uncovered: The alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. *Biochim. Biophys. Acta BBA - Bioenerg.*, **1847**, 957–967.

Namgung,J. *et al.* (2021) Complete chloroplast genomes shed light on phylogenetic relationships, divergence time, and biogeography of *Allioideae* (*Amaryllidaceae*). *Sci. Rep.*, **11**, 3262.

Nugent,J.M. and Palmer,J.D. (1991) RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. *Cell*, **66**, 473–481.

Ohri,D. *et al.* (1998) Evolution of genome size in *Allium* (*Alliaceae*). *Plant Syst. Evol.*, **210**, 57–86.

Ohyama,K. *et al.* (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, **322**, 572–574.

Oldenburg,D. and Bendich,A. (1996) Size and Structure of Replicating Mitochondrial DNA in Cultured Tobacco Cells. *Plant Cell*, **8**, 447–461.

Ordonez,N. *et al.* (2015) Worse Comes to Worst: Bananas and Panama Disease—When Plant and Pathogen Clones Meet. *PLOS Pathog.*, **11**, e1005197.

Otto,G. (2021) Giant genomes of lungfish. *Nat. Rev. Genet.*, **22**, 199–199.

Peltola,H. *et al.* (1984) SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Res.*, **12**, 307–321.

Peña Iglesias,A. (1988) El ajo: Virosis, fisiopatias y seleccion clonal y sanitaria, 1: Parte teorico-descriptiva. *Boletin Sanid. Veg. Plagas*, **14**.

Petrovska,B.B. and Cekovska,S. (2010) Extracts from the history and medical properties of garlic. *Pharmacogn. Rev.*, **4**, 106–110.

Pevzner,P. and Shamir,R. (2011) Bioinformatics for Biologists Cambridge University Press.

Pont-Kingdon,G.A. *et al.* (1995) A coral mitochondrial mutS gene. *Nature*, **375**, 109–111.

Pooler,M.R. and Simon,P.W. (1994) True seed production in garlic. *Sex. Plant Reprod.*, **7**, 282–286.

Qiu,Y.-L. and Palmer,J.D. (2004) Many Independent Origins of trans Splicing of a Plant Mitochondrial Group II Intron. *J. Mol. Evol.*, **59**, 722–724.

Rabinowitch,H.D. (1990) Onions and Allied Crops: Volume I: Botany, Physiology, and Genetics CRC Press, Boca Raton.

Rabinowitch,H.D. and Currah,L. (2002) *Allium* Crop Science: Recent Advances CABI.

Rana,S.B. *et al.* (2016) Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, *Fundulus heteroclitus*. *PLOS ONE*, **11**, e0153104.

Rice,D.W. *et al.* (2013) Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm *Amborella*. *Science*, **342**, 1468–1473.

Ricroch,A. *et al.* (2005) Evolution of genome size across some cultivated *Allium* species. *Genome*, **48**, 511–520.

Sabar,M. *et al.* (2003) ORFB is a subunit of F1F(O)-ATP synthase: insight into the basis of cytoplasmic male sterility in sunflower. *EMBO Rep.*, **4**, 381–386.

Sagan,L. (1967) On the origin of mitosing cells. *J. Theor. Biol.*, **14**, 255–274.

Sakai,T. and Imamura,J. (1993) Evidence for a mitochondrial sub-genome containing radish *atpA* in a *Brassica napus* cybrid. *Plant Sci.*, **90**, 95–103.

Sakamoto,W. *et al.* (1996) Altered mitochondrial gene expression in a maternal distorted leaf mutant of Arabidopsis induced by chloroplast mutator. *Plant Cell*, **8**, 1377.

Satagopan,J.M. *et al.* (1996) A Bayesian Approach to Detect Quantitative Trait Loci Using Markov Chain Monte Carlo. *Genetics*, **144**, 805–816.

Schatz,G. and Dobberstein,B. (1996) Common Principles of Protein Translocation Across Membranes. *Science*, **271**, 1519–1526.

Scheffler,I.E. (2008) Mitochondria , Immo E. Scheffler. 2nd ed. Wiley-Liss, Hoboken, N.J.

Scheweisguth,B. (1973) Study of a new type of male sterility in onion. *Genetics*, **29**, 569–572.

Shaw,J. *et al.* (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.*, **94**, 275–288.

Shedge,V. *et al.* (2007) Plant Mitochondrial Recombination Surveillance Requires Unusual RecA and MutS Homologs. *Plant Cell*, **19**, 1251–1264.

Shemesh-Mayer,E. *et al.* (2015) Garlic (*Allium sativum* L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development. *Front. Plant Sci.*, **6**.

Shinozaki,K. *et al.* (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.*, **5**, 2043–2049.

Singh,R.K. *et al.* (2012) Studies on variability and genetic divergence in elite lines of garlic (*Allium sativum* L.). *J. Spices Aromat. Crops*, **21**.

Sloan,D.B. *et al.* (2012) Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. *PLOS Biol.*, **10**, e1001241.

Small,I. *et al.* (1989) Evolution of plant mitochondrial genomes via substoichiometric intermediates. *Cell*, **58**, 69–76.

Smith,L.S.P.S. (1993) Bond Graph Modelling of Physical Systems University of Glasgow (United Kingdom).

Smith,R.L. and Chowdhury,M.K.U. (1991) Characterization of pearl millet mitochondrial DNA fragments rearranged by reversion from cytoplasmic male sterility to fertility. *Theor. Appl. Genet.*, **81**, 793–799.

Soorni,A. *et al.* (2021) Transcriptome and phytochemical analyses provide insights into the organic sulfur pathway in *Allium hirtifolium*. *Sci. Rep.*, **11**, 768.

Staden,R. (1980) A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.*, **8**, 3673–3694.

Stern,D.B. and Palmer,J.D. (1984) Recombination sequences in plant mitochondrial genomes: diversity and homologies to known mitochondrial genes. *Nucleic Acids Res.*, **12**, 6141–6157.

Sun,H. and Schneeberger,K. (2015) SHOREmap v3.0: fast and accurate identification of causal mutations from forward genetic screens. *Methods Mol. Biol. Clifton NJ*, **1284**, 381–395.

Sun,X. *et al.* (2020) A Chromosome-Level Genome Assembly of Garlic (*Allium sativum*) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Mol. Plant*, **13**, 1328–1339.

Sun,X. *et al.* (2012) De novo assembly and characterization of the garlic (*Allium sativum*) bud transcriptome by Illumina sequencing. *Plant Cell Rep.*, **31**, 1823–1828.

Sutton,G.G. *et al.* (1995) TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Sci. Technol.*, **1**, 9–19.

Takagi,H. (1990) Garlic *Allium sativum* L. *Onion Allied Crops Vol III Biochem. Food Sci. Minor Crops*.

Takagi,H. *et al.* (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J. Cell Mol. Biol.*, **74**, 174–183.

Takenaka,M. *et al.* (2013) RNA Editing in Plants and Its Evolution. *Annu. Rev. Genet.*, **47**, 335–352.

Takenaka,M. *et al.* (2008) The process of RNA editing in plant mitochondria. *Mitochondrion*, **8**, 35–46.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Teng,M. *et al.* (2016) A benchmark for RNA-seq quantification pipelines. *Genome Biol.*, **17**, 74.

Tillich,M. *et al.* (2017) GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.*, **45**, W6–W11.

Timmis,J.N. *et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.*, **5**, 123–135.

Tsujimura,M. *et al.* (2019) Multichromosomal structure of the onion mitochondrial genome and a transcript analysis. *Mitochondrion*, **46**, 179–186.

Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–1063.

Van Blerkom,J. (2011) Mitochondrial function in the human oocyte and embryo and their role in developmental competence. *Mitochondrion*, **11**, 797–813.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.

Wick,R.R. *et al.* (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350–3352.

Woloszynska,M. (2010) Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes—though this be madness, yet there's method in't. *J. Exp. Bot.*, **61**, 657–671.

Wu,Y. *et al.* (2018) Comparative Chloroplast Genomics of *Gossypium* Species: Insights Into Repeat Sequence Variations and Phylogeny. *Front. Plant Sci.*, **9**.

Wu,Z. *et al.* (2015) The massive mitochondrial genome of the angiosperm *Silene noctiflora* is evolving by gain or loss of entire chromosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 10185–10191.

Xie,D.-F. *et al.* (2020) Insights into phylogeny, age and evolution of *Allium* (*Amaryllidaceae*) based on the whole plastome sequences. *Ann. Bot.*, **125**, 1039–1055.

Xing,J. *et al.* (2023) The complete mitochondrial genome of *Allium fistulosum* L. (*Amaryllidaceae*). *Mitochondrial DNA Part B*, **8**, 890–894.

Yang,X. *et al.* (2020) Comparative Analysis of the Complete Chloroplast Genomes in *Allium* Subgenus *Cyathophora* (*Amaryllidaceae*): Phylogenetic Relationship and Adaptive Evolution. *BioMed Res. Int.*, **2020**, e1732586.

Yu,Q.-B. *et al.* (2014) Nuclear-encoded factors associated with the chloroplast transcription machinery of higher plants. *Front. Plant Sci.*, **5**.

Zaegel,V. *et al.* (2006) The Plant-Specific ssDNA Binding Protein OSB1 Is Involved in the Stoichiometric Transmission of Mitochondrial DNA in Arabidopsis. *Plant Cell*, **18**, 3548–3563.

Zardoya,R. *et al.* (2002) Origin of plant glycerol transporters by horizontal gene transfer and functional recruitment. *Proc. Natl. Acad. Sci.*, **99**, 14893–14896.

Zaremba-Niedzwiedzka,K. *et al.* (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.

Zeng,Z.B. (1994) Precision Mapping of Quantitative Trait Loci. *Genetics*, **136**, 1457–1468.

Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhang,H. *et al.* (2019) QTG-Seq Accelerates QTL Fine Mapping through QTL Partitioning and Whole-Genome Sequencing of Bulked Segregant Samples. *Mol. Plant*, **12**, 426–437.

Zhang,W. *et al.* (2011) A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLOS ONE*, **6**, e17915.

Zhao,J. *et al.* (2023) The complete chloroplast genome of *Allium wallichii* Kunth (Amaryllidaceae). *Mitochondrial DNA Part B Resour.*, **8**, 1054–1058.

Zhelyazkova,P. *et al.* (2012) The Primary Transcriptome of Barley Chloroplasts: Numerous Noncoding RNAs and the Dominating Role of the Plastid-Encoded RNA Polymerase. *Plant Cell*, **24**, 123–136.

Zhu,S. *et al.* (2019) Transcriptome-wide association study and eQTL analysis to assess the genetic basis of bulb-yield traits in garlic (*Allium sativum*). *BMC Genomics*, **20**, 657.

# 9 Anexos - Publicaciones

En este apartado se recogen las 3 publicaciones que reflejan el trabajo realizado en esta tesis doctoral.

plants

MDPI

*Review*

# Turning Garlic into a Modern Crop: State of the Art and Perspectives

**Ricardo Parreño [1], Eva Rodríguez-Alcocer [1], César Martínez-Guardiola [1], Lucía Carrasco [1], Purificación Castillo [2], Vicent Arbona [3], Sara Jover-Gil [1] and Héctor Candela [1,*]**

[1] Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Spain
[2] Departamento I+D, Coopaman S.C.L., Carretera Peñas De San Pedro, km 1.6, 02006 Albacete, Spain
[3] Departament de Ciències Agràries i del Medi Natural, Universitat Jaume I, 12071 Castelló de la Plana, Spain
* Correspondence: hcandela@umh.es; Tel.: +34-965222583

**Abstract:** Garlic is cultivated worldwide for the value of its bulbs, but its cultivation is challenged by the infertility of commercial cultivars and the accumulation of pathogens over time, which occurs as a consequence of vegetative (clonal) propagation. In this review, we summarize the state of the art of garlic genetics and genomics, highlighting recent developments that will lead to its development as a modern crop, including the restoration of sexual reproduction in some garlic strains. The set of tools available to the breeder currently includes a chromosome-scale assembly of the garlic genome and multiple transcriptome assemblies that are furthering our understanding of the molecular processes underlying important traits like the infertility, the induction of flowering and bulbing, the organoleptic properties and resistance to various pathogens.

**Keywords:** garlic; breeding; *Allium sativum*; genetics; genomics; fertility restoration



**Citation:** Parreño, R.; Rodríguez-Alcocer, E.; Martínez-Guardiola, C.; Carrasco, L.; Castillo, P.; Arbona, V.; Jover-Gil, S.; Candela, H. Turning Garlic into a Modern Crop: State of the Art and Perspectives. *Plants* **2023**, *12*, 1212. https://doi.org/10.3390/plants12061212

Academic Editor: Adrián Rodríguez-Burruezo

Received: 31 January 2023
Revised: 26 February 2023
Accepted: 2 March 2023
Published: 7 March 2023



## 1. General features and challenges

Garlic (*Allium sativum* L.) is a monocotyledonous plant belonging to the family *Amaryllidaceae*, in the order Asparagales. It is native to Central Asia and is cultivated in temperate climates worldwide, with an annual production of 28 million tons on approximately 1.6 million hectares (http://fao.org/faostat/, accessed on 31 January 2023). China and India are the largest producers of garlic, accounting for 80% of the global production (Table 1). With references to its use dating back to ancient Egypt and India 5000 years ago, garlic is one of the oldest known crops.

**Table 1.** Global production of *Allium* crops in 2021.

| Crop | Cultivated ha [1] | Tons Produced | Yield (Tons/ha) | Highest Producers |
|---|---|---|---|---|
| Garlic | 1,659,236 | 28,204,854 | 17.00 | China, India, Bangladesh, Egypt |
| Leeks and other alliaceous vegetables | 134,168 | 2,213,183 | 16.49 | Indonesia, Turkey, Belgium, France |
| Onions and shallots, dry | 5,778,767 | 106,592,008 | 18.44 | India, China, Egypt, United States |
| Onions and shallots, green | 215,933 | 4,665,525 | 21.60 | China, Mali, Japan, Republic of Korea |

[1] Data according to FAOSTAT (http://fao.org/faostat/, accessed on 31 January 2023).

The genus *Allium* comprises 1018 known species (https://powo.science.kew.org/, accessed on 30 January 2023), including perennial geophytes characterized by the production of bulbs or rhizomes. Bulbs are an adaptation to growth in arid regions with dry summer periods. The genus includes species of great economic and agronomic interest,

cultivated for their bulbs and pseudostems, used as food or condiment, such as onion (*A. cepa* L.), shallot (*A. ascalonicum* L.), chives (*A. schoenoprasum* L.), leek (*A. ampeloprasum* L.), bunching onions (*A. fistulosum* L.) and Chinese chives (*A. tuberosum* Spreng.), or for ornamental purposes, such as *A. aflatunense* B. Fedtsch.

Several attempts have been made to reconstruct the phylogeny of the genus [1]. The phylogenies obtained through parsimony and neighbor-joining methods using internal transcribed spacer (ITS) sequences of nuclear ribosomal RNAs were generally congruent with each other [1], with the exception of the placement of certain species, and with previous molecular phylogenies [2]. According to these phylogenies, the development of bulbs in *Allium* species is the ancestral state for this trait, being common in the species of the *Amerallium* and *Allium* subgenera. The presence of rhizomes is considered to be a derived character, while the production of bulbs in onion is considered a reversion to the ancestral state.

In the cultivated species of *Allium*, the basal part of the foliage leaves differentiates to form bulbs and pseudostems, which constitute the edible portion of the leaves. Garlic bulbs are rich in organosulfur compounds, which impart them their unique organoleptic properties [3]. The most abundant of these compounds is allicin (allyl disulfide oxide), which is synthesized from the non-protein amino acid alliin by the enzyme alliinase (EC 4.4.1.4) in response to tissue disruption [4]. Allicin has been used as a food additive due to its multiple bioactive properties [4].

### *1.1. The Life Cycle of a Garlic Plant: Bulb and Flower Development*

Some aspects of the anatomy and development of sterile and fertile garlic accessions have been studied in great detail by different authors [5–8]. Garlic bulbs contain a variable number of cloves, usually ranging from 8 to 14, and are covered by outer tunics (i.e., the dry sheaths of older foliage leaves). The cloves are lateral shoots that develop from the axillary meristems located on the adaxial side of foliage leaves. Each clove is itself a small bulb and is covered by dry, protective leaves, which vary in color (white, purple or reddish) and cover a modified storage leaf that constitutes the fleshy part of the clove. The apical meristem of the clove is located at the apex of a very short stem called the basal plate, and is flanked by the developing leaf primordia. The life cycle of a garlic plant (Figure 1) starts when a clove exits dormancy and sprouts after being exposed to temperatures between 5 °C and 10 °C for several weeks [9,10]. One of the earliest events in the development of a new plant is the emergence of adventitious roots at the periphery of the basal plate, preceding the development of flat foliage leaves. Foliage leaves are initiated from the apical meristem and eventually form a pseudostem. The root system and the leaves usually develop before bulb formation.

Of particular interest are the mechanisms that control the floral transition and the induction of new bulbs. The ability to bolt, whereby the shoot apical meristem shifts from forming vegetative leaves to produce a floral scape, only occurs in some varieties. This trait has been used as a criterion for distinguishing garlic subspecies. Garlic varieties have been classified as having a flowering stalk ("bolting", "stalking" or "hard neck"), without a flowering stalk ("non-bolting") or with a partial stalk ("soft neck") [10]. In the soft-neck cultivars, the plant bolts but the elongation of the stalk is incomplete, and no mature flowers are formed. The inflorescence of garlic is an umbel containing up to 100 flower clusters, or cymes, each with 5 or 6 flowers [8]. The inflorescences of many varieties might also contain topsets (bulbils). The topsets are propagules that, similar to the cloves, allow for the vegetative propagation of the plant and are thought to compete with the flowers for nutrients and space, limiting and sometimes preventing flower development. The flowers are smaller than those of onions, with floral organs arranged in five whorls, including a perianth formed by six tepals (the outer whorls one and two), six stamens (whorls three and four), and a syncarpous ovary formed by three fused carpels that define three locules (whorl five), each containing two ovules [7,11–13].

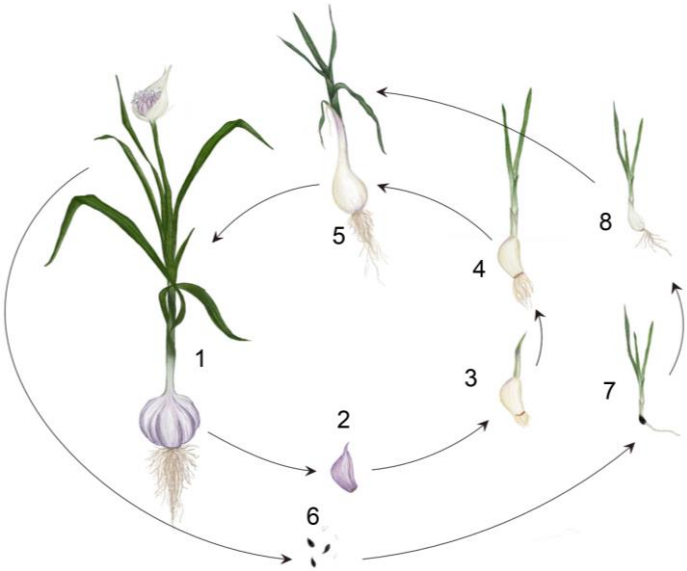



**Figure 1.** Life cycle of a garlic plant showing the different stages from the germination of a seed or clove to a mature plant. (1) Adult garlic plant of a bolting cultivar bearing a floral scape. (2) Dormant clove, recently detached from the mother plant, which has not yet produced adventitious roots. (3) Sprouted clove in which new adventitious roots and the first leaves have emerged. (4) Young garlic plant in which the formation of new cloves has not yet been initiated. (5) Garlic plant in which the bulb is actively thickening as a result of the growth of new cloves. (6) True seeds are produced only in some fertile garlic varieties. (7) When a seed germinates, a new seedling emerges, with a primary root emerging from the embryo's apical meristem and new leaves developing from the shoot apical meristem. (8) Under inductive conditions, seed-derived plants can also produce new bulbs, which can then be vegetatively propagated if desired.

Bulbing and bolting are key developmental transitions that are regulated by temperature and photoperiod. Two stages can be distinguished in relation to the development of the bulbs and the inflorescence [14]. During the so-called inductive stage, short-day (9 to 10 h) photoperiods and low (10 °C) temperatures trigger the flowering transition in many garlic clones [15], as well as the initiation of lateral buds from the leaf axils [13]. Later in the crop season, during the morphogenic stage, the long-day photoperiod and higher temperatures of the spring promote the filling of the cloves, bolting and flower development. The enhancement of bulbing and cloving by low temperatures has been known for a long time [16], and is the basis of the vernalization or "conditioning" treatments that are applied to the cloves before sowing [14]. Conditioning has been reported to accelerate bulbing, allowing an earlier harvest of the bulbs and better bulb formation [14,17]. On the other hand, long-day photoperiods might result in smaller bulbs as a result of accelerated bulbing, while they promote the elongation of the flower stalk [6,18]. The mutual effects of bulbing on bolting and vice-versa are not fully understood. Bulb development is known to initiate earlier in non-bolting cultivars, possibly implying the existence of a balance or competition between bulb and inflorescence development [19], but some studies have failed to appreciate this competition, as stem elongation seemed to have no impact on bulb weight [18]. The optimal temperatures and photoperiods might vary among accessions, as also occurs in other plants [18]. Indeed, flowering in garlic is strongly influenced by the genetic background, as some accessions fail to bolt even under long photoperiod and inductive conditions, which is one of the reasons that contributes to explain the infertility of garlic cultivars.

### *1.2. Consequences of Vegetative Propagation and Fertility Restoration*

Garlic is typically propagated vegetatively using the cloves or bulbils from the previous harvest, as the cultivated varieties have lost the ability to reproduce by true seeds (sexual

reproduction). The infertility of garlic impedes the making of controlled crosses and limits the implementation of modern breeding methods, which is one of the main challenges that hinders the development of garlic as a modern crop. As a result, new garlic varieties have typically been obtained through clonal selection of spontaneously occurring variants [20]. The accumulation of pathogens over time, which reduces plant vigor, is another major burden associated to the vegetative propagation of garlic. Viral infections can reduce crop yields by up to 50% and are perpetuated through vegetative propagation [21], forcing growers to periodically clean the planting material by applying time-consuming and costly methods to select plants that are free of viruses and other pathogens in an attempt to mitigate this problem [22]. These methods typically involve the micropropagation of shoot apical meristems, although cultures derived from almost every plant organ have been tested, including leaves, basal plates and roots. In vitro cultures are often combined with additional treatments, such as thermotherapy, cryotherapy and chemotherapy, which selectively affect the survival of the infected cells [23,24]. The cost of these procedures is further increased by the need to go from the in vitro cultures to full production capacity, a process that might require several years [19]. Because the bulbils do not develop in direct contact with the soil, they have been proposed to contain a lower pathogen load and are also used in the propagation of garlic [25]. However, a recent study has found that some viruses also accumulate in the bulbils, pointing to the need for additional treatments before bulbils can be effectively used for garlic propagation. The need to treat and store a large amount of cloves until the next growing season is very costly in personnel and resources, which is one of the main factors that decrease the economic yield of garlic [26].

Research on flowering and seed production will lead to major changes in the future of garlic as a crop. The selection of fertile varieties (capable of reproducing by true seed) is expected to solve problems such as the accumulation of pathogens in bulbs and bulbils and would accelerate the breeding of new cultivars with improved characteristics. The production of true seeds in some garlic accessions was first described in the 1950s [27], but the first studies of sexual reproduction and the breeding of fertile varieties were not carried out until the 1980s [12]. Since then, there have been major advances in this field, such as the identification of fertile varieties [28–32]), as well as detailed descriptions of their flowering and reproductive characteristics [6,13,33].

## 2. The Genetics and Genomics of Garlic: State-of-the-Art

Despite the agricultural importance of garlic, the repertoire of resources for garlic breeding used to be very limited. The lack of molecular tools, combined with the asexual mode of reproduction, has made it difficult to use modern marker-assisted selection methods and incorporate the existing genetic variation into breeding programs, which have almost exclusively relied on clonal selection strategies. This situation has left breeders with few options, such as clonal selection, chemical or physical mutagenesis, exploiting somaclonal variation, mutation induction, transgenesis, polyploidy induction and genome editing [34–36].

### *2.1. Molecular Markers and Linkage Maps*

Prior to genome sequencing, numerous researcher groups reported the development and use of various types of molecular markers, from isozymes [37,38] to DNA-based markers, including RAPD (random amplified polymorphic DNA) [38], SSR (simple sequence repeats, also known as microsatellites) [39–49], ISSR (inter simple sequence repeats) [44,50], SRAP (sequence-related amplified polymorphism) [51] and AFLP (amplified fragment length polymorphisms) markers. These markers have been utilized for various purposes, such as studying the phylogenetic relationships between *Allium* species, characterizing the intraspecific diversity in garlic and related species and assessing the relationships among the accessions of germplasm collections [39]. Some authors have questioned whether the variants of some SSRs actually correspond to true alleles [52]. The alleles of some molecular markers followed Mendelian segregation ratios in plant families obtained through sexual reproduction [52]. With the availability of the garlic genome sequence, the genomic location

of these markers can now be determined, which has confirmed the non-allelic character of many randomly selected SSR sequences. Therefore, markers based on expressed sequences, such as ESTs (expressed sequence tags) or assembled RNA-seq reads, are expected to be more reproducible and informative [46–49,53].

The available toolkit includes several low-resolution linkage maps, which depict the position and relative distance of a limited number of molecular markers [54,55]. The first linkage map of the garlic genome [54] was constructed in 2005 using 84 plants from an $S_1$ mapping population, generated by self-fertilization of a single plant of a Soviet-origin variety. The map incorporated 83 single-nucleotide polymorphisms (SNPs) and 8 insertion-deletion markers (indels) developed from expressed sequences, which were assigned to 9 linkage groups. The mapping population also enabled the study of male sterility, which was inherited as a recessive trait and could be assigned to one of the linkage groups. In addition, continuous variation was reported for several traits, suggesting that their genetic architecture can be investigated using conventional quantitative trait locus (QTL) mapping techniques. A second map was reported the same year [55] based on AFLP markers and gene-specific markers, using two mapping populations obtained through self-pollination. The maps obtained from these populations consisted of 20 and 13 linkage groups, respectively, which exceeds the number of chromosomes in the garlic haploid genome ($n = 8$). Both studies ruled out seed production due to apomixis, although this phenomenon had been previously described in some *Allium* species. Instead, the observed Mendelian segregation ratios clearly indicated amphimixis. Furthermore, some markers showed a 15:1 segregation ratio, as expected for duplicated markers, uncovering the existence of duplicated sequences in the garlic genome. Unfortunately, these linkage maps were created using mapping populations that, to our knowledge, have not been maintained by the corresponding research groups, which limits their usefulness because new molecular markers cannot be added and integrated maps incorporating different types of molecular markers cannot be created.

Genotyping a large number of molecular markers in all individuals of a population can be achieved though techniques, such as genotyping by sequencing (GBS) [56] or restriction-site-associated DNA (RAD) sequencing (RAD-Seq) [57], and is a crucial preliminary step for mapping and identifying genes and QTLs relevant for crop breeding. Despite the significant challenge posed by the large size of the garlic genome, GBS methods have been successfully applied to construct at least three linkage maps in closely related species, such as onion [58–60]. The first of these maps was made using 96 $F_2$ plants from a cross involving 2 different onion cultivars, which contained more than 10,000 single-nucleotide polymorphisms (SNPs) spanning the 8 chromosomes of the onion haploid genome [58]. The main challenge in constructing linkage maps in garlic is the infertility of most available varieties, which prevents the making of crosses and mapping populations. However, DArT-seq (Diversity Array Technology sequencing), a method related to GBS, has been successfully used in garlic to evaluate the genetic diversity and to identify redundant accessions in a large germplasm collection from Spain [61]. In this work, the authors identified 131 redundant accessions (out of 417 accessions), which allowed them to select a core set that captures the genetic diversity of the collection. However, Barboza and colleagues [48] have recently warned against the elimination of accessions on only the grounds of marker genotypes, emphasizing the importance of retaining epigenetic variants, and furthermore, pointing out that the germplasm collection studied by Giménez and García-Lampasona had higher levels of epigenetic variation than of genetic variation [62,63]. The successful use of DArT-seq implies that similar GBS methods could also be used to create high-resolution linkage maps in garlic. In many species, these maps were made using recombinant inbred line (RIL) or doubled haploid (DH) mapping populations, which represent permanent resources for gene mapping and offer unlimited material for building integrated linkage maps incorporating different types of markers [64,65]. An additional advantage of these populations is that they can be characterized multiple times, under different

environmental conditions or at different geographical locations for QTL mapping. The ability to vegetatively propagate garlic offers an easy and affordable alternative to the generation of RIL or DH populations, as the progeny of any cross can be perpetuated with no need to perform further crosses.

### *2.2. Transcriptomic Approaches in Garlic*

The garlic genome is diploid ($2n = 2x = 16$), with an estimated size 5 times larger than the human genome, reaching 15.9 gigabases [66]. This large size is a common property among other *Allium* species, making it challenging to study. Prior to genome sequencing, some studies provided information on the transcriptome (the expressed part of the genome) of various tissues, such as meristems and other organs [67–71]. In addition to efforts to characterize the gene content of the garlic genome, efforts have also been made since 2015 to identify genes in phylogenetically close species, such as onion [72], bunching onions [73] and Chinese chives [74].

The study of garlic genes will further our understanding, at the molecular level, of the main problems faced by this crop, such as the infertility. Some transcriptomic studies aimed at understanding its molecular basis and might help to restore the production of true seeds in some varieties [8,71]. Other studies have focused on the relationship between storage temperatures and bulb formation [75]. The development of new molecular markers [76–79] and their application to the genetic improvement of these new fertile varieties promises to be a revolution in the development of this crop. The first garlic transcriptome assemblies were performed in 2012 using Illumina sequencing technology [68]. The de novo assembly of reads from two libraries, prepared from dormant and germinating vegetative buds from field-collected bulbs, allowed the assembly of 127,933 unigenes, which were assigned functions by comparing the sequences with different databases and assigning gene ontology (GO) terms with the BLAST2GO program. The obtained sequences were used to identify genes involved in the sulfur assimilation pathway as well as in the biosynthesis of organic sulfur compounds. In 2015, Kamenetsky and co-workers [70] assembled and annotated an integrated transcriptome of six organs (flowers, inflorescences, leaves, cloves, basal plate and root) of a fertile garlic variety, which allowed them to identify orthologues of genes involved in flowering and sulfur metabolism, as well as to detect the presence of some viruses that were present in the sequenced samples. A comprehensive summary of efforts to characterize the transcriptome in garlic is presented in Table 2.

**Table 2.** Transcriptome studies using garlic.

| Reference | Sequencer | Software | Sample | Results |
|---|---|---|---|---|
| Kim et al., 2009 [67] | Sanger | | Leaf and stem tissues | 21,595 ESTs |
| Sun et al., 2012 [68] | Illumina HiSeq 2000 | SOAPdenovo | Dormant and sprouting vegetative buds | 127,933 unigenes |
| Sun et al., 2013 [69] | Illumina HiSeq 2000 | | See [68] | 45,363 DEGs * |
| Kamenetsky et al., 2015 [70] | Illumina MiSeq | Trinity | Inflorescence, flower, leaf, clove, roots and basal plate | 239,116 ('extensive'), or 102,042 contigs ('abundant' transcriptome) |
| Shemesh-Mayer et al., 2015 [71] | Illumina Hiseq 2000 | Bowtie, DESeq | Flower buds at 3 developmental stages | 16,271 DEGs * |
| Liu et al., 2015 [46] | Illumina HiSeq 2500 | Trinity, MISA | 10 days old plants, 45 days old | 135,360 unigenes; 1506 SSR markers |
| Havey and Ahn, 2016 [52] | Sanger and Roche 454-FLX | SOAPdenovo-trans | Leaf, pseudostem and root tissues | 35,936 contigs; 14,879 SNP and indel markers |
| Zhu et al., 2017 [80] | Illumina HiSeq 2500 | Trinity | Leaf tissue | 132,225 unigenes |
| Chaturvedi et al., 2018 [81] | Illumina HiSeq 2000 | Bowtie, DESeq | Internal buds and storage leaves at two temperatures | 8303 (internal buds) and 14,147 DEGs * (storage leaves) |

**Table 2.** *Cont.*

| Reference | Sequencer | Software | Sample | Results |
|---|---|---|---|---|
| Li et al., 2018 [82] | Illumina HiSeq 4000 | Trinity, CD-HIT | Cloves stored at 4°C for 0, 10, 15 and 40 days | 49,280 unigenes; 5923 DEGs |
| Chen et al., 2018 [83] | PacBio RSII (Iso-Seq CCS) and Illumina HiSeq 2500 | proovread, CD-HIT | Developing bulb | 36,321 transcripts |
| Liu et al., 2020 [84] | Illumina HiSeq 2500 | Trinity, edgeR | Stem (control and treated with GA3) | 159 DEGs * |
| Sun et al., 2020 [85] | Illumina HiSeq 2500 | Trinity, DEGseq | Sprouts, bulbs, flowers, roots, pseudostems and leaves | 34,439 transcripts with constitutive (28,394) or specific (964) expression (out of 57,561 genes predicted in the genome) |
| Wang et al., 2022 [86] | PacBio Sequel (CCS) | Quiver, CD-HIT-EST | Lower bulb, aerial bulb, scape, leaf, clove, basal plate and roots | 36,571 high-quality consensus reads |

* Differentially Expressed Genes.

### *2.3. The Genomes of Garlic and Related Species*

We are currently witnessing a revolution in the number of sequenced genomes for *Allium* species, with available sequences for the genomes of garlic [85], onion [87] and bunching onion [88]. Garlic was the first *Allium* species to have a chromosome-scale genome assembly, which was achieved through a combination of PacBio, Illumina and ONT sequencing technologies, as well as 10X Genomic libraries and Hi-C technology. The assembly yielded a 16.24 Gb sequence, spanning 96.1% of the genome according to estimates based on *k*-mer statistics. The genome contained 57,561 annotated protein-coding genes, 10% of which are located in tandem repeats, and 20,008 non-coding RNA genes, including 3741 miRNAs, 8984 tRNAs, 4352 rRNAs and 2931 snRNAs. The relatively low BUSCO (Benchmarking Universal Single-Copy Orthologs) values suggested that this assembly still requires further improvement, which might be achieved using optical and chromosome-contact maps developed using Bionano Genomics technology.

The main difficulty in the assembly of the garlic genome lies in its enormous complexity, given that it contains a high number of repetitive sequences (14.8 Gb, approximately 91.3% of the genome assembly, higher than the percentage in the maize genome), clustered in the central regions of the chromosomes, and high levels of heterozygosity. The large number of transposable elements (which account for 76% of the genome size) contribute to its large size and their insertions often affect the expression of other genes. The large size of the garlic genome has been attributed to the expansion of *gypsy*-type LTR retrotransposons [88]. In fact, 4219 garlic genes were found to be disrupted by transposon insertions, including an orthologue of the floral homeotic gene *APETALA2*.

At the time of its release in 2020, the garlic genome was the largest one for a monocotyledonous plant, although six additional genomes in the same order (*Asparagales*) were also available [85]. The phylogenetic analysis of these genomes showed that garlic forms a monophyletic clade with *Narcissus viridiflorus*, a member of the *Amaryllidaceae* family. Garlic was phylogenetically closest to *Asparagus officinalis*; the two species shared a common ancestor 80.8 million years ago (mya). In plants, the expansion of gene families occurs by polyploidization caused by whole-genome duplication (WGD) events and the proliferation of transposable elements. Analysis of the synteny between garlic and *A. officinalis* uncovered 3 WGD events in the evolutionary history of garlic, which occurred 120–130, 89.8 (prior to the divergence of garlic and *Asparagus*) and 17.9 mya. Two transcriptome assemblies have recently been performed using third-generation sequencing technologies [83,86], which promise to lead to a better annotation of the garlic genome.

The genome of bunching onion consists of 8 chromosomes and is 11.27 Gb in size, which roughly matches the size estimated from *k*-mer statistics (11.97 Gb). This genome

contains 62,255 genes and has high levels of repeated sequences (89.89% according to *k*-mer statistics, and 69.81% according to repeatmasker). The collinearity between the genomes of *A. fistulosum* and *A. cepa* is higher than that between the genomes of *A. fistulosum* and garlic. In the latter, the existence of genomic rearrangements is apparent, although the number of chromosomes in the haploid genome remains at eight. This correlation was already apparent in linkage maps with SSR-type markers. The onion genome is diploid ($2n = 16$), consisting of as many chromosomes, as the genome of *A. fistulosum*. It has been determined that the repetitive fraction of the onion genome amounts to 95%, being especially rich in *copia* and *gypsy* retrotransposons [87]. Genome sequencing was performed with a doubled haploid accession combining PacBio and Illumina reads. The availability of linkage maps allowed anchoring scaffolds to as many pseudo-molecules as there are chromosomes in the genome. However, not all scaffolds could be oriented in the linkage map. The proportion of repetitive sequences turned out to be lower than initially expected. Annotation will require additional effort, as the article documents the prediction of 540,925 gene models, of which only a much smaller fraction has experimental support in RNA-seq reads.

## 3. Target Traits for Garlic Breeding

### *3.1. Yield and Bulb Traits*

Considerable variation among clones has been reported, including differences in flower stalk formation, bulbil development in inflorescences, bulb size and morphology and bulb color. Other traits of agronomic interest include number of cloves, organoleptic properties (flavor, pungency and color), health-enhancing effects (nutraceutical value), clove uniformity, firmness, early harvest, late bolting, resistance to diseases, and differences in the production of phenolic and sulfur compounds [18,48,89–92]. Breeding strategies in garlic have been primarily oriented toward the improvement in bulb yield and quality. Continuous variation has been reported for traits such as bulb size (height, width and weight), number of cloves and number of days from sowing to harvest [49]. Not unexpectedly, a significant correlation was found between bulb size components and number of cloves.

### *3.2. Secondary Metabolism as a Target for Garlic Breeding*

The health-promoting and flavor attributes of garlic have been associated with the presence of different secondary metabolites, including polyphenolic compounds, organosulfur metabolites (alkenyl cysteine sulfoxides, S-allyl cysteine, thiosulfinates, diallyl sulfides, vinyldithiins and different isomers of ajoene) [3] and oligosaccharides, primarily acting as storage carbohydrates in bulbs [93,94]. Volatile sulfur-containing compounds (namely allicin, which accounts for ~70% of the total thiosulfinate compounds produced after crushing garlic cloves) are important contributors to flavor traits of garlic and are produced by the decomposition of S-alkenyl cysteine sulfoxides (alliin, methiin, propiin and isoalliin) following an enzyme-catalyzed process. Moreover, allicin is highly unstable and can be spontaneously converted into different allyl sulfides that contribute to the flavor of garlic products [3]. At the biochemical level, isoalliin and S-carboxypropyl-cysteine sulphoxide are derived from glutathione, whereas alliin derives from the reaction of the amino acid serine with an allyl thiol of unknown origin and a glutamate moiety rendering g-glutamyl-S-allyl-L-cysteine, which is then metabolized by a flavine-containing monooxygenase a sulfoxide and subsequently by a GTPase rendering the corresponding alliin [95].

Secondary metabolites influence organoleptic properties such as color, pungency and taste, and might also constitute the end-products of the domestication process to adapt to different environmental conditions, hence acting as actual markers of geographic origin of *Allium* varieties [93]. Fructo-oligosaccharides are important storage compounds present in onion bulbs, and organosulfur compounds are key determinants of the aroma of *Allium* species. Flavonoids are primarily involved in the color of the bulb (particularly in onions) and also play an important role in abiotic and biotic stress tolerance [93]. In this regard, the inner peels (enclosing each clove) and outer peels (enclosing the entire bulb) of garlic bulbs do have different metabolic composition: inner clove peels are rich in organic acids

such as gluconic, gulonic acids and vanillic acid, whereas outer bulb peels are richer in carbohydrates (rhamnose, lyxose, glucose, xylobiose D and trehalose) along with sugar alcohols (mannitol, sorbitol and threitol), both differing greatly from clove metabolite composition. Some of these metabolites likely account for the allelopathic properties of onion bulb peels [96]. Liu and co-workers [97] analyzed the evolution of metabolites in different parts of the garlic plant (leaf, pseudostem, bulb peel wrapper, clover skin and clove) and identified 84 different compounds whose dynamics were highly correlated with the storage role of bulbs.

The production of non-pungent, tear-free onions is a successful example of new variety development by manipulating the secondary metabolism of organosulfur compounds in an *Allium* species. The garlic genome contains 60 genes of the alliinase gene family [85], which undoubtedly contribute to its organoleptic properties. In onion, the tear-inducing lachrymatory factor is synthesized through two consecutive, enzymatically catalyzed reactions. The first reaction is catalyzed by an alliinase enzyme activity, while the second is catalyzed by the lachrymatory factor synthase (LFS). Although onions are recalcitrant to transformation, the LFS gene was silenced using a transgenic RNAi approach—the first example of gene silencing in onions [98]. More recently, tearless onions have also been selected in a mutant screen after heavy-ion beam mutagenesis. In this case, bulbs with reduced alliinase activity were identified in $M_3$ families derived from 1,450 $M_1$ irradiated seeds [99].

## 4. Pathogens of Garlic: Threats, Treatments and Perspectives

An undesirable consequence of vegetative propagation is the accumulation of pathogens in garlic bulbs, which causes a reduction in yield and an increase in costs. Many different types of pathogens affect garlic, including arthropods (such as insects and mites), fungi, bacteria, viruses and phytoplasmas. Various procedures can be used for their control, ranging from the sanitation of planting material using sophisticated in vitro culture procedures, to crop rotation, avoiding the use of the same plots in consecutive years and the use of phytosanitary products to control the dispersal of vectors. Current restrictions on the use of these substances pose difficulties in maintaining product quality, so it is essential to identify germplasm accessions that are genetically resistant, or at least less susceptible, to attack by different types of pathogens and to open the door to the use of the new biotechnological tools available.

### *4.1. Fungal Pathogens*

Fusarium basal rot (FBR), the most devastating soil-borne disease of garlic, is caused by the necrotoph fungus *Fusarium oxysporum* f. sp. *cepae* (FOC) [100]. The first visible symptoms of FBR are the curling and yellowing of the leaves, eventually leading to rotting of the basal part or the whole plant. Importantly, the damage caused by FBR eventually becomes the entry point for other secondary pathogens. Promising approaches to prevent the proliferation of fungal pathogens involve the use of fungicides, intercropping, soil solarization techniques and storing the bulbs under optimal conditions of humidity and temperature, but the use of resistant accessions remains the most cost-effective approach to control fungi and other soil-borne pathogens [91,101,102]. The AsRGA29 (*A. sativum* resistant gene analog 29) was found to be induced after inoculation with FOC in a resistant garlic line (named CBT-As153), or in other *Allium* species that have been reported to be naturally resistant to the fungus, such as *A. fistulosum* and *A. roylei*, as well as after exogenous treatments with methyl jasmonate, salicylic acid, abscisic acid and hydrogen peroxide, suggesting its putative involvement in the response against FBR [91]. Sequencing and quantitative PCR have uncovered families of microRNAs (miRNAs) involved in the activation of the immune response of garlic against FOC [103]. The miR394 microRNA was also found to be responsive to FOC inoculation, but the induction was more pronounced in a sensitive line (CBT-As11) than in the resistant line (CBT-As153) [104]. In line with this observations, two predicted targets of miR394, which encode an F-box protein and

a P450 cytochrome, were expressed at lower levels after FOC inoculation, and both the expression of miR394 and its targets were regulated by methyl jasmonate [104]. Another disease, Fusarium dry rot (FDR), is caused by *Fusarium proliferatum* and mainly affects the crop during the post-harvest bulb storage [105]. *F. proliferatum* and *F. oxysporum* do not have a specific host [106], and intercropping with species that can be infected but are naturally resistant to these pathogens, such as maize and spring wheat [107], decreases inoculum levels in the soil [108]. Other gene families putatively involved in defense mechanisms against *Fusarium* infections have been identified in garlic, including genes that encode TLP (thaumatin-like proteins) [109], PR (pathogenesis-related) proteins [110] or CHI (class I chitinase) proteins [101], whose expression is positively correlated with the response to *Fusarium* in plants resistant to the fungus.

*Alternaria porri* (the causal agent of purple blotch) [111], *Sclerotium cepivorum* (white rot) [112], *Stemphylium* spp. (leaf blight) [113], *Botrytis allii* (neck rot) [114], *Aspergillus* spp. (black mold), *Penicillium corymbiferum* (decay) [115] and *Puccinia spp.* (garlic rust) [116] are fungi that cause serious diseases at distinct steps of cultivation. Most of the research done to identify sources of resistance to these fungal infections has been performed in species other than garlic, but the information gained will be highly valuable once fertility has been restored. The resistance of onion to purple blotch was studied by crossing resistant and sensitive cultivars, and was inherited as a monogenic, dominant trait. Bulked segregant analysis suggested that two markers were linked to the locus conferring pathogen resistance [117]. *Agrobacterium*-mediated transformation of calli has also been performed to express tobacco proteins (chitinase and glucanase) against *S. cepivorum*, leading to a reduction of invasion in transformed garlic plants [118]. To identify sources of resistance to *Stemphylium*, allotetraploid hybrids were created between *A. cepa* and *A. fistulosum*, as well as recombinant chromosomes between both species following consecutive crossings [119]. Time-series expression analysis unveiled differential expression of several genes related to the jasmonic acid signaling pathway (*JAR1*, *COI1*, and *MYC2*) after *Botrytis* infection in two onion lines (one resistant and one susceptible), suggesting that these genes participate in the plant response against *Botrytis* [120]. In regard to *Puccinia*, a recent study using monosomal alien addition lines (MAAL), which add chromosomes from rust-resistant *A. cepa* varieties to *A. fistulosum* plant, suggested that resistance genes reside on chromosome 1 of onion [121].

### *4.2. Nematodes*

Nematodes such as *Ditylenchus dipsaci* cause wilting, chlorosis and damaged or rotted bulbs, as they facilitate the entry of other pathogens, such as *Fusarium* strains that cause FDR or FBR [122,123]. The nematode reproduction rate (FR) has been evaluated in different garlic varieties, finding that some varieties have low FR both in vitro and in the field [124]. Although one cultivar from Israel was completely resistant, its bulbs were not suitable for commercialization [125]. Recently, it has been proposed that a specific PTI (Pattern-triggered immunity) response to nematode attack involves DNA hypomethylation in the CHH context in certain regions of the genome, altering the expression patterns of some plant genes [126].

### *4.3. Arthropods: Insects and Mites*

Arthropods represent a problem because they damage the plants and the bulbs at the growing and post-harvest stages and because they act as vectors for viral pathogens. The eriophyid mite *Aceria tulipae* (dry bulb mite) invades plants causing a characteristic curling of the leaves. At the storage stage, the mites absorb sap from the clove, forming brown and reddish spots that greatly reduce bulb weight [127]. In addition, the mites transmit different viruses of genus *Allexivirus* [128–130]. Bolting varieties of garlic have been reported to be more susceptible to mite attack than non-bolting ones [131]. Insects from several orders, such as *Ephestia cautella* (Lepidoptera), *Delia antiqua* (Diptera) or *Thrips tabaci* (Thysanoptera), feed on the plants or stored bulbs and cause important yield losses [132,133]. Moreover,

garlic breeding would significantly benefit from the identification of resistance sources to these pests.

### *4.4. Viruses*

Plant pathogenic viruses reduce yield by affecting bulb weight and diameter, leading to significant economic losses worldwide. Plants are often simultaneously infected by multiple viruses of the *Allexivirus*, *Potyvirus* and *Carlavirus* genera (all three characterized by having single-stranded positive-sense RNA genomes [134]), synergistically causing yield losses [132]. Genus *Allexivirus* includes *Garlic virus A* to *Garlic virus E* and *Garlic virus X* (GarV-A to GarV-E and GarV-X), *Shallot Virus X* (ShVX) and *Garlic Mite-borne Filamentous Virus* (GarMbFV) [135]. Allexiviruses are transmitted by *Aceria tulipae* [129], and the effects of their infection might range from asymptomatic to leaf mosaics and reduced bulb size. Genus *Potyvirus* include viruses such as *Onion yellow dwarf virus* (OYDV), *Leek yellow stripe virus* (LYSV), *Shallot Yellow Stripe Virus* (SYSV) and *Tobacco etch virus* (TEV) [136–139]. Genus *Carlavirus* includes the *Garlic yellow mosaic-associated virus* (GYMaV) [140], the *Garlic common latent virus* (GarCLV) and the *Shallot latent virus* (SLV; a synonym for *Garlic latent virus*, GLV), the latter two generating chlorotic ringspots with a necrotic center distributed on the leaf [141]. All these potyviruses and carlaviruses are transmitted by aphids.

Other viruses that infect *Allium* species belong to the genera Nepovirus (Artichoke yellow ringspot virus, AYRV), Cucumovirus (Cucumber mosaic virus, CMV), Necrovirus (Leek white stripe virus, LWSV), Tospovirus (Irish yellow spot virus, IYSV) and Fijivirus (Garlic dwarf virus, GDV) [139,142,143]. The use of new massively parallel sequencing technologies facilitates the identification of viruses in plants showing symptoms of virosis. The Illumina technology, for example, has been used to detect Garlic virus E [144].

### *4.5. Bacteria and Phytoplasmas*

In addition to viruses, some bacteria also cause diseases in garlic during development, at harvest time or upon storage. During garlic development, *Xanthomonas axonopodis* pv. *allii* causes whitish, lenticular-shaped, brown lesions that spread on the leaves. In episodes of advanced infection, it causes progressive leaf death and reduces bulb size [145]. Another disease that occurs during development is garlic blight, caused by phytoplasmas, which alters the color of the leaves. The color change begins at the leaf apices and moves toward the base, ultimately causing the death of the plant [146]. At harvest time, the most common disease is bacterial soft rot caused by *Erwinia carotovora* subsp. *carotovora*, *Dickeya chrysanthemi*, *Pectobacterium carotovorum* subsp. *carotovorum* and *Lactobacillus* spp. In the early stages of infection, symptoms appear in the neck region of garlic and cause a watery rot in the infected tissues [147]. Upon storage, infections with *Pseudomonas salomonii* and *Pseudomonas fluorescens* cause a unique phenotype consisting of dark brown spots on the tunics of garlic bulbs, known as *café au lait* [148,149]. So far, garlic immunity mechanisms towards viral and bacterial infections have not been described.

## 5. Opportunities for the Future of Garlic Breeding

Knowledge of the genome and genes involved in the development of an organism, such as garlic, or in the development of a certain phenotypic trait of interest, represent a fundamental step toward its detailed functional characterization at the molecular level. Understanding the function of garlic genes is expected to facilitate the cultivation under adverse environmental conditions, such as high temperature, increased soil salinity or drought, without affecting its development or the crop's yield. Understanding the natural variation would make it possible to determine the contribution of certain alleles to beneficial phenotypic traits. Sexual reproduction will enable the introduction of such traits into elite cultivars through controlled crosses. The clonal propagation of garlic is associated with a low level of intra-population genetic diversity, making the populations more prone to diseases, which make garlic less likely to successfully cope with the harsh conditions of a changing climate or with emerging pathogens.

In the last decades, many authors have made impressive efforts to generate the tools required for the development of improved garlic cultivars. Major advances included the establishments of large collections of molecular markers, and the demonstration that linkage maps of garlic can be built with the available tools. The most important milestone is the restoration of seed production in certain garlic lines that had retained the ability to reproduce sexually. Extensive knowledge on the environmental conditions that trigger bulb and inflorescence development has accumulated in recent year. The development of massively parallel sequencing technologies in the last decade has been instrumental to gain rapid access to the genetic information of plants with enormous genomes, such as garlic. These technologies initially allowed the de novo assembly of garlic transcriptomes, but the advent of long-read sequencing technologies is now facilitating the assembly of complex genomes, such as those of Allium crops, which are rich in various types of repeats. There are countless existing protocols that have addressed different aspects of garlic tissue culture, transformation mediated by *Agrobacterium tumefaciens* or by biolistic methods and plant regeneration. Altogether, these methods and the available genome sequences will open the door to the use of novel genome editing tools in garlic. The possibility of obtaining segregating populations of garlic by sexual reproduction, combined with the ability to maintain these populations indefinitely through vegetative reproduction, will greatly facilitate the study of the genetic architecture of quantitative traits of agronomic interest, such as those described in this review.

**Author Contributions:** Conceptualization, H.C. and S.J.-G.; investigation, R.P., E.R.-A., C.M.-G., L.C., P.C., V.A., S.J.-G. and H.C.; writing—original draft preparation, H.C. and R.P.; writing—review and editing, R.P., E.R.-A., C.M.-G., L.C., P.C., V.A., S.J.-G. and H.C.; project administration, H.C. and S.J.-G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Coopaman-UMH contract number COOPAMAN1.22I.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ricroch, A.; Yockteng, R.; Brown, S.C.; Nadot, S. Evolution of Genome Size across Some Cultivated *Allium* Species. *Genome* **2005**, *48*, 511–520. [CrossRef] [PubMed]
2. Mes, T.H.; Fritsch, R.M.; Pollner, S.; Bachmann, K. Evolution of the Chloroplast Genome and Polymorphic ITS Regions in *Allium* Subg. *Melanocrommyum*. *Genome* **1999**, *42*, 237–247. [CrossRef] [PubMed]
3. Li, J.; Dadmohammadi, Y.; Abbaspourrad, A. Flavor Components, Precursors, Formation Mechanisms, Production and Characterization Methods: Garlic, Onion, and Chili Pepper Flavors. *Crit. Rev. Food Sci. Nutr.* **2022**, *62*, 8265–8287. [CrossRef] [PubMed]
4. Borlinghaus, J.; Albrecht, F.; Gruhlke, M.C.H.; Nwachukwu, I.D.; Slusarenko, A.J. Allicin: Chemistry and Biological Properties. *Molecules* **2014**, *19*, 12591–12618. [CrossRef]
5. Mann, L. Anatomy of the Garlic Bulb and Factors Affecting Bulb Development. *Hilgardia* **1952**, *21*, 195–251. [CrossRef]
6. Kamenetsky, R.; Rabinowitch, H.D. Floral Development in Bolting Garlic. *Sex. Plant Reprod.* **2001**, *13*, 235–241. [CrossRef]
7. Winiarczyk, K.; Kosmala, A. Development of the Female Gametophyte in the Sterile Ecotype of the Bolting *Allium sativum* L. *Sci. Hortic.* **2009**, *121*, 353–360. [CrossRef]
8. Shemesh Mayer, E.; Winiarczyk, K.; Błaszczyk, L.; Kosmala, A.; Rabinowitch, H.D.; Kamenetsky, R. Male Gametogenesis and Sterility in Garlic (*Allium sativum* L.): Barriers on the Way to Fertilization and Seed Production. *Planta* **2013**, *237*, 103–120. [CrossRef]
9. Mann, L.; Lewis, D. Rest and Dormancy in Garlic. *Hilgardia* **1956**, *26*, 161–189. [CrossRef]
10. Takagi, H. Garlic *Allium sativum* L. In *Onions and Allied Crops: Volume III: Biochemistry, Food Science, and Minor Crops*; CRC Press: Boca Raton, FL, USA, 1990; pp. 109–146.
11. Etoh, T. Studies on the Sterility in Garlic, *Allium sativum* L. *Mem. Fac. Agric. Kagoshima Univ.* **1985**, *21*, 77–132.
12. Etoh, T. Fertility of the Garlic Clones Collected in Soviet Central Asia. *J. Jpn. Soc. Hortic. Sci.* **1986**, *55*, 312–319. [CrossRef]

13. Kamenetsky, R.; Shafir, I.L.; Zemah, H.; Barzilay, A.; Rabinowitch, H. Environmental Control of Garlic Growth and Florogenesis. *J. Am. Soc. Hortic. Sci.* **2004**, *129*, 144–151. [CrossRef]
14. Guevara-Figueroa, T.; López-Hernández, L.; Lopez, M.; Hurtado, M.D.; Vázquez-Barrios, M.; Guevara-Olvera, L.; González, R.G.; Rivera-Pastrana, D.; Torres-Robles, H.; Mercado-Silva, E. Conditioning Garlic "Seed" Cloves at Low Temperature Modifies Plant Growth, Sugar, Fructan Content, and Sucrose Sucrose Fructosyl Transferase (1-SST) Expression. *Sci. Hortic.* **2015**, *189*, 150–158. [CrossRef]
15. Pooler, M.R.; Simon, P.W. Garlic Flowering in Response to Clone, Photoperiod, Growth Temperature, and Cold Storage. *HortScience* **1993**, *28*, 1085–1086. [CrossRef]
16. Aoba, T.; Takagi, H. Studies on the Bulb Formation in Garlic Plants III. On the Effects of Cooling Treatments of Seed-Bulbs and Day-Length during the Growing Period on Bulbing. *J. Jpn. Soc. Hortic. Sci.* **1971**, *40*, 240–245. [CrossRef]
17. Wu, C.; Wang, M.; Dong, Y.; Cheng, Z.; Meng, H. Growth, Bolting and Yield of Garlic (*Allium sativum* L.) in Response to Clove Chilling Treatment. *Sci. Hortic.* **2015**, *194*, 43–52. [CrossRef]
18. Mathew, D.; Forer, Y.; Rabinowitch, H.D.; Kamenetsky, R. Effect of Long Photoperiod on the Reproductive and Bulbing Processes in Garlic (*Allium sativum* L.) Genotypes. *Environ. Exp. Bot.* **2011**, *71*, 166–173. [CrossRef]
19. Shemesh-Mayer, E.; Kamenetsky-Goldstein, R. Traditional and novel approaches in garlic (*Allium sativum* L.) breeding. In *Advances in Plant Breeding Strategies: Vegetable Crops*; Springer: Cham, Switzerland, 2021; Volume 8, pp. 3–49.
20. Peña-Iglesias, A. El Ajo: Virosis, Fisiopatías y Selección Clonal y Sanitaria. I: Parte Teórico-Descriptiva. *Bol. Sanid. Veg. Plagas* **1988**, *14*, 461–483.
21. Conci, V.C.; Canavelli, A.; Lunello, P.; Di Rienzo, J.; Nome, S.F.; Zumelzu, G.; Italia, R. Yield Losses Associated with Virus-Infected Garlic Plants during Five Successive Years. *Plant Dis.* **2003**, *87*, 1411–1415. [CrossRef]
22. Luciani, G.F.; Mary, A.K.; Pellegrini, C.; Curvetto, N.R. Effects of Explants and Growth Regulators in Garlic Callus Formation and Plant Regeneration. *Plant Cell Tissue Organ Cult.* **2006**, *87*, 139–143. [CrossRef]
23. Ramírez-Malagón, R.; Pérez-Moreno, L.; Borodanenko, A.; Salinas-González, G.; Ochoa-Alejo, N. Differential Organ Infection Studies, Potyvirus Elimination, and Field Performance of Virus-Free Garlic Plants Produced by Tissue Culture. *Plant Cell Tissue Organ Cult.* **2006**, *86*, 103–110. [CrossRef]
24. Vieira, R.L.; da Silva, A.L.; Zaffari, G.R.; Steinmacher, D.A.; de Freitas Fraga, H.P.; Guerra, M.P. Efficient Elimination of Virus Complex from Garlic (*Allium sativum* L.) by Cryotherapy of Shoot Tips. *Acta Physiol. Plant.* **2015**, *37*, 1733. [CrossRef]
25. Kajimura, Y.; Sugiura, T.; Suenaga, K.; Itakura, Y.; Etoh, T. A New Garlic Growing System from Bulbils through Transplanting. *J. Hortic. Sci. Biotechnol.* **2000**, *75*, 176–180. [CrossRef]
26. Keller, E.J.; Zanke, C.D.; Senula, A.; Breuing, A.; Hardeweg, B.; Winkelmann, T. Comparing Costs for Different Conservation Strategies of Garlic (*Allium sativum* L.) Germplasm in Genebanks. *Genet. Resour. Crop Evol.* **2013**, *60*, 913–926. [CrossRef]
27. Kononkov, P. The Question of Obtaining Garlic Seed. *Sadi Ogorod* **1953**, *8*, 38–40.
28. Hong, C.; Etoh, T. Fertile Clones of Garlic (*Allium sativum* L.) Abundant around the Tien Shan Mountains. *Jpn. J. Breed.* **1996**, *46*, 349–353. [CrossRef]
29. Jenderek, M. Generative Reproduction of Garlic (*Allium sativum*). *Ses. Nauk.* **1998**, *57*, 141–145.
30. Etoh, T.; Noma, Y.; Nishitarumizu, Y.; Wakamoto, T. Seed Productivity and Germinability of Various Garlic [*Allium sativum* L.] Clones Collected in Soviet Central Asia. *Mem. Fac. Agric.-Kagoshima Univ. Jpn.* **1988**, *24*, 129–139.
31. Jenderek, M.; Hannan, R. Seed Producing Ability of Garlic (*Allium sativum* L.) Clones from Two Public US Collections. In Proceedings of the Third International Symposium on Edible Alliaceae, Athens, GA, USA, 30 October–3 November 2000; pp. 73–75.
32. Kamenetsky, R.; Shafir, I.L.; Baizerman, M.; Khassanov, F.; Kik, C.; Rabinowitch, H. Garlic (*Allium sativum* L.) and Its Wild Relatives from Central Asia: Evaluation for Fertility Potential. In Proceedings of the XXVI International Horticultural Congress: Advances in Vegetable Breeding, Toronto, ON, Canada, 11–17 August 2002; pp. 83–91.
33. Jenderek, M.M.; Hannan, R.M. Variation in Reproductive Characteristics and Seed Production in the USDA Garlic Germplasm Collection. *HortScience* **2004**, *39*, 485–488. [CrossRef]
34. Singh, H.; Khar, A.; Verma, P. Induced Mutagenesis for Genetic Improvement of Allium Genetic Resources: A Comprehensive Review. *Genet. Resour. Crop Evol.* **2021**, *68*, 2669–2690. [CrossRef]
35. Hailu, M.G.; Mawcha, K.T.; Nshimiyimana, S.; Suharsono, S. Garlic Micro-Propagation and Polyploidy Induction in Vitro by Colchicine. *Plant Breed. Biotechnol.* **2021**, *9*, 1–19. [CrossRef]
36. Wen, Y.; Liu, H.; Meng, H.; Qiao, L.; Zhang, G.; Cheng, Z. In Vitro Induction and Phenotypic Variations of Autotetraploid Garlic (*Allium sativum* L.) With Dwarfism. *Front. Plant Sci.* **2022**, *13*, 917910. [CrossRef]
37. Etoh, T.; Ogura, H. Peroxidase Isozymes in the Leaves of Various Clones of Garlic, *Allium sativum* L. *Mem. Fac. Agric. Kagoshima Univ.* **1981**, *17*, 71–77.
38. Maaß, H.; Klaas, M. Infraspecific Differentiation of Garlic (*Allium sativum* L.) by Isozyme and RAPD Markers. *Theor. Appl. Genet.* **1995**, *91*, 89–97. [CrossRef]
39. Ma, K.-H.; Kwag, J.-G.; Zhao, W.; Dixit, A.; Lee, G.-A.; Kim, H.-H.; Chung, I.-M.; Kim, N.-S.; Lee, J.-S.; Ji, J.-J. Isolation and Characteristics of Eight Novel Polymorphic Microsatellite Loci from the Genome of Garlic (*Allium sativum* L.). *Sci. Hortic.* **2009**, *122*, 355–361. [CrossRef]

40. Zhao, W.; Chung, J.; Lee, G.; Ma, K.; Kim, H.; Kim, K.; Chung, I.; Lee, J.; Kim, N.; Kim, S. Molecular Genetic Diversity and Population Structure of a Selected Core Set in Garlic and Its Relatives Using Novel SSR Markers. *Plant Breed.* **2011**, *130*, 46–54. [CrossRef]
41. Lee, G.-A.; Kwon, S.-J.; Park, Y.-J.; Lee, M.-C.; Kim, H.-H.; Lee, J.-S.; Lee, S.-Y.; Gwag, J.-G.; Kim, C.-K.; Ma, K.-H. Cross-Amplification of SSR Markers Developed from *Allium sativum* to Other Allium Species. *Sci. Hortic.* **2011**, *128*, 401–407. [CrossRef]
42. Chen, S.; Chang, Y.; Zhou, J.; Cheng, Z.; Meng, H. Genetic Diversity of Garlic (*Allium sativum* L.) Germplasm by Simple Sequence Repeats. *J. Agric. Biotechnol.* **2012**, *20*, 372–381.
43. Cunha, C.P.; Hoogerheide, E.S.; Zucchi, M.I.; Monteiro, M.; Pinheiro, J.B. New Microsatellite Markers for Garlic, *Allium sativum* (Alliaceae). *Am. J. Bot.* **2012**, *99*, e17–e19. [CrossRef]
44. Chen, S.; Chen, W.; Shen, X.; Yang, Y.; Qi, F.; Liu, Y.; Meng, H. Analysis of the Genetic Diversity of Garlic (*Allium sativum* L.) by Simple Sequence Repeat and Inter Simple Sequence Repeat Analysis and Agro-Morphological Traits. *Biochem. Syst. Ecol.* **2014**, *55*, 260–267. [CrossRef]
45. Ovesna, J.; Leišová-Svobodová, L.; Kučera, L. Microsatellite Analysis Indicates the Specific Genetic Basis of Czech Bolting Garlic. *Czech J. Genet. Plant Breed.* **2014**, *50*, 226–234. [CrossRef]
46. Liu, T.; Zeng, L.; Zhu, S.; Chen, X.; Tang, Q.; Mei, S.; Tang, S. Large-Scale Development of Expressed Sequence Tag-Derived Simple Sequence Repeat Markers by Deep Transcriptome Sequencing in Garlic (*Allium sativum* L.). *Mol. Breed.* **2015**, *35*, 204. [CrossRef]
47. Barboza, K.; Beretta, V.; Kozub, P.C.; Salinas, C.; Morgenfeld, M.M.; Galmarini, C.R.; Cavagnaro, P.F. Microsatellite Analysis and Marker Development in Garlic: Distribution in EST Sequence, Genetic Diversity Analysis, and Marker Transferability across Alliaceae. *Mol. Genet. Genom.* **2018**, *293*, 1091–1106. [CrossRef] [PubMed]
48. Barboza, K.; Salinas, M.C.; Acuña, C.V.; Bannoud, F.; Beretta, V.; Garcia-Lampasona, S.; Burba, J.L.; Galmarini, C.R.; Cavagnaro, P.F. Assessment of Genetic Diversity and Population Structure in a Garlic (*Allium sativum* L.) Germplasm Collection Varying in Bulb Content of Pyruvate, Phenolics, and Solids. *Sci. Hortic.* **2020**, *261*, 108900. [CrossRef]
49. Li, X.; Qiao, L.; Chen, B.; Zheng, Y.; Zhi, C.; Zhang, S.; Pan, Y.; Cheng, Z. SSR Markers Development and Their Application in Genetic Diversity Evaluation of Garlic (*Allium sativum*) Germplasm. *Plant Divers.* **2022**, *44*, 481–491. [CrossRef]
50. Jabbes, N.; Geoffriau, E.E.; Le Clerc, V.; Dridi, B.; Hannechi, C. Inter Simple Sequence Repeat Fingerprints for Assess Genetic Diversity of Tunisian Garlic Populations. *J. Agric. Sci.* **2011**, *3*, 77–85. [CrossRef]
51. Chen, S.; Zhou, J.; Chen, Q.; Chang, Y.; Du, J.; Meng, H. Analysis of the Genetic Diversity of Garlic (*Allium sativum* L.) Germplasm by SRAP. *Biochem. Syst. Ecol.* **2013**, *50*, 139–146. [CrossRef]
52. Havey, M.J.; Ahn, Y.-K. Single Nucleotide Polymorphisms and Indel Markers from the Transcriptome of Garlic. *J. Am. Soc. Hortic. Sci.* **2016**, *141*, 62–65. [CrossRef]
53. Ipek, M.; Sahin, N.; Ipek, A.; Cansev, A.; Simon, P.W. Development and Validation of New SSR Markers from Expressed Regions in the Garlic Genome. *Sci. Agric.* **2015**, *72*, 41–46. [CrossRef]
54. Zewdie, Y.; Havey, M.J.; Prince, J.P.; Jenderek, M.M. The First Genetic Linkages among Expressed Regions of the Garlic Genome. *J. Am. Soc. Hortic. Sci.* **2005**, *130*, 569–574. [CrossRef]
55. Ipek, M.; Ipek, A.; Almquist, S.G.; Simon, P.W. Demonstration of Linkage and Development of the First Low-Density Genetic Map of Garlic, Based on AFLP Markers. *Theor. Appl. Genet.* **2005**, *110*, 228–236. [CrossRef]
56. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef]
57. Baird, N.A.; Etter, P.D.; Atwood, T.S.; Currey, M.C.; Shiver, A.L.; Lewis, Z.A.; Selker, E.U.; Cresko, W.A.; Johnson, E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **2008**, *3*, e3376. [CrossRef]
58. Jo, J.; Purushotham, P.M.; Han, K.; Lee, H.-R.; Nah, G.; Kang, B.-C. Development of a Genetic Map for Onion (*Allium cepa* L.) Using Reference-Free Genotyping-by-Sequencing and SNP Assays. *Front. Plant Sci.* **2017**, *8*, 1606. [CrossRef]
59. Choi, Y.; Kim, S.; Lee, J. Construction of an Onion (*Allium cepa* L.) Genetic Linkage Map Using Genotyping-by-Sequencing Analysis with a Reference Gene Set and Identification of QTLs Controlling Anthocyanin Synthesis and Content. *Plants* **2020**, *9*, 616. [CrossRef]
60. Lee, Y.-R.; Kim, C.W.; Han, J.; Choi, H.J.; Han, K.; Lee, E.S.; Kim, D.-S.; Lee, J.; Siddique, M.I.; Lee, H.-E. Genotyping-by-Sequencing Derived Genetic Linkage Map and Quantitative Trait Loci for Sugar Content in Onion (*Allium cepa* L.). *Plants* **2021**, *10*, 2267. [CrossRef]
61. Egea, L.A.; Mérida-García, R.; Kilian, A.; Hernandez, P.; Dorado, G. Assessment of Genetic Diversity and Structure of Large Garlic (*Allium sativum*) Germplasm Bank, by Diversity Arrays Technology "Genotyping-by-Sequencing" Platform (DArTseq). *Front. Genet.* **2017**, *8*, 98. [CrossRef]
62. Gimenez, M.D.; Yañez-Santos, A.M.; Paz, R.C.; Quiroga, M.P.; Marfil, C.F.; Conci, V.C.; García-Lampasona, S.C. Assessment of Genetic and Epigenetic Changes in Virus-Free Garlic (*Allium sativum* L.) Plants Obtained by Meristem Culture Followed by in Vitro Propagation. *Plant Cell Rep.* **2016**, *35*, 129–141. [CrossRef]
63. Gimenez, M.D.; García Lampasona, S. Before-after Analysis of Genetic and Epigenetic Markers in Garlic: A 13-Year Experiment. *Sci. Hortic.* **2018**, *240*, 23–28. [CrossRef]
64. Burr, B.; Burr, F.A. Recombinant Inbreds for Molecular Mapping in Maize: Theoretical and Practical Considerations. *Trends Genet.* **1991**, *7*, 55–60. [CrossRef]
65. Pollard, D.A. Design and construction of recombinant inbred lines. In *Quantitative Trait Loci (QTL)*; Humana Press: Totowa, NJ, USA, 2012; pp. 31–39.

66. Arumuganathan, K.; Earle, E.D. Nuclear DNA Content of Some Important Plant Species. *Plant Mol. Biol. Rep.* **1991**, *9*, 208–218. [CrossRef]
67. Kim, D.-W.; Jung, T.-S.; Nam, S.-H.; Kwon, H.-R.; Kim, A.; Chae, S.-H.; Choi, S.-H.; Kim, D.-W.; Kim, R.N.; Park, H.-S. GarlicESTdb: An Online Database and Mining Tool for Garlic EST Sequences. *BMC Plant Biol.* **2009**, *9*, 61. [CrossRef] [PubMed]
68. Sun, X.; Zhou, S.; Meng, F.; Liu, S. De Novo Assembly and Characterization of the Garlic (*Allium sativum*) Bud Transcriptome by Illumina Sequencing. *Plant Cell Rep.* **2012**, *31*, 1823–1828. [CrossRef] [PubMed]
69. Sun, X.D.; Ma, G.Q.; Cheng, B.; Li, H.; Liu, S.Q. Identification of Differentially Expressed Genes in Shoot Apex of Garlic (*Allium sativum* L.) Using Illumina Sequencing. *J. Plant Stud.* **2013**, *2*, 136. [CrossRef]
70. Kamenetsky, R.; Faigenboim, A.; Shemesh Mayer, E.; Ben Michael, T.; Gershberg, C.; Kimhi, S.; Esquira, I.; Rohkin Shalom, S.; Eshel, D.; Rabinowitch, H.D. Integrated Transcriptome Catalogue and Organ-Specific Profiling of Gene Expression in Fertile Garlic (*Allium sativum* L.). *BMC Genom.* **2015**, *16*, 12. [CrossRef] [PubMed]
71. Shemesh-Mayer, E.; Ben-Michael, T.; Rotem, N.; Rabinowitch, H.D.; Doron-Faigenboim, A.; Kosmala, A.; Perlikowski, D.; Sherman, A.; Kamenetsky, R. Garlic (*Allium sativum* L.) Fertility: Transcriptome and Proteome Analyses Provide Insight into Flower and Pollen Development. *Front. Plant Sci.* **2015**, *6*, 271. [CrossRef] [PubMed]
72. Kim, S.; Kim, M.-S.; Kim, Y.-M.; Yeom, S.-I.; Cheong, K.; Kim, K.-T.; Jeon, J.; Kim, S.; Kim, D.-S.; Sohn, S.-H. Integrative Structural Annotation of de Novo RNA-Seq Provides an Accurate Reference Gene Set of the Enormous Genome of the Onion (*Allium cepa* L.). *DNA Res.* **2015**, *22*, 19–27. [CrossRef] [PubMed]
73. Tsukazaki, H.; Yaguchi, S.; Sato, S.; Hirakawa, H.; Katayose, Y.; Kanamori, H.; Kurita, K.; Itoh, T.; Kumagai, M.; Mizuno, S. Development of Transcriptome Shotgun Assembly-Derived Markers in Bunching Onion (*Allium fistulosum*). *Mol. Breed.* **2015**, *35*, 51. [CrossRef]
74. Zhou, S.-M.; Chen, L.-M.; Liu, S.-Q.; Wang, X.-F.; Sun, X.-D. De Novo Assembly and Annotation of the Chinese Chive (*Allium tuberosum* Rottler Ex Spr.) Transcriptome Using the Illumina Platform. *PLoS ONE* **2015**, *10*, e0133312. [CrossRef]
75. Rohkin Shalom, S.; Gillett, D.; Zemach, H.; Kimhi, S.; Forer, I.; Zutahy, Y.; Tam, Y.; Teper-Bamnolker, P.; Kamenetsky, R.; Eshel, D. Storage Temperature Controls the Timing of Garlic Bulb Formation via Shoot Apical Meristem Termination. *Planta* **2015**, *242*, 951–962. [CrossRef]
76. Chand, S.K.; Nanda, S.; Rout, E.; Joshi, R.K. Mining, Characterization and Validation of EST Derived Microsatellites from the Transcriptome Database of *Allium sativum* L. *Bioinformation* **2015**, *11*, 145. [CrossRef]
77. Buso, G.; Paiva, M.; Torres, A.; Resende, F.; Ferreira, M.; Buso, J.; Dusi, A. Genetic Diversity Studies of Brazilian Garlic Cultivars and Quality Control of Garlic-Clover Production. *Genet. Mol. Res.* **2008**, *7*, 534–541. [CrossRef]
78. García-Lampasona, S.; Asprelli, P.; Burba, J.L. Genetic Analysis of a Garlic (*Allium sativum* L.) Germplasm Collection from Argentina. *Sci. Hortic.* **2012**, *138*, 183–189. [CrossRef]
79. Jo, M.H.; Ham, I.K.; Moe, K.T.; Kwon, S.-W.; Lu, F.-H.; Park, Y.-J.; Kim, W.S.; Kim, M.K.; Kim, T.I.; Lee, E.M. Classification of Genetic Variation in Garlic ('*Allium sativum*' L.) Using SSR Markers. *Aust. J. Crop Sci.* **2012**, *6*, 625–631.
80. Zhu, S.; Tang, S.; Tan, Z.; Yu, Y.; Dai, Q.; Liu, T. Comparative Transcriptomics Provide Insight into the Morphogenesis and Evolution of Fistular Leaves in Allium. *BMC Genom.* **2017**, *18*, 60. [CrossRef]
81. Chaturvedi, A.K.; Shalom, S.R.; Faigenboim-Doron, A.; Teper-Bamnolker, P.; Salam, B.B.; Daus, A.; Kamenetsky, R.; Eshel, D. Differential Carbohydrate Gene Expression during Preplanting Temperature Treatments Controls Meristem Termination and Bulbing in Garlic. *Environ. Exp. Bot.* **2018**, *150*, 280–291. [CrossRef]
82. Li, N.; Qiu, Z.; Lu, X.; Shi, B.; Sun, X.; Tang, X.; Qiao, X. Comparative Transcriptome Analysis of Temperature-Induced Green Discoloration in Garlic. *Int. J. Genom.* **2018**, *2018*, 6725728. [CrossRef]
83. Chen, X.; Liu, X.; Zhu, S.; Tang, S.; Mei, S.; Chen, J.; Li, S.; Liu, M.; Gu, Y.; Dai, Q.; et al. Transcriptome-Referenced Association Study of Clove Shape Traits in Garlic. *DNA Res.* **2018**, *25*, 587–596. [CrossRef]
84. Liu, H.; Wen, Y.; Cui, M.; Qi, X.; Deng, R.; Gao, J.; Cheng, Z. Histological, Physiological and Transcriptomic Analysis Reveal Gibberellin-Induced Axillary Meristem Formation in Garlic (*Allium sativum*). *Plants* **2020**, *9*, 970. [CrossRef]
85. Sun, X.; Zhu, S.; Li, N.; Cheng, Y.; Zhao, J.; Qiao, X.; Lu, L.; Liu, S.; Wang, Y.; Liu, C.; et al. A Chromosome-Level Genome Assembly of Garlic (*Allium sativum*) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Mol. Plant* **2020**, *13*, 1328–1339. [CrossRef]
86. Wang, L.; Zhang, C.; Yin, W.; Wei, W.; Wang, Y.; Sa, W.; Liang, J. Single-Molecule Real-Time Sequencing of the Full-Length Transcriptome of Purple Garlic (*Allium sativum* L. Cv. Leduzipi) and Identification of Serine O-Acetyltransferase Family Proteins Involved in Cysteine Biosynthesis. *J. Sci. Food Agric.* **2022**, *102*, 2864–2873. [CrossRef] [PubMed]
87. Finkers, R.; van Kaauwen, M.; Ament, K.; Burger-Meijer, K.; Egging, R.; Huits, H.; Kodde, L.; Kroon, L.; Shigyo, M.; Sato, S.; et al. Insights from the First Genome Assembly of Onion (*Allium cepa*). *G3 GenesGenomesGenetics* **2021**, *11*, jkab243. [CrossRef] [PubMed]
88. Liao, N.; Hu, Z.; Miao, J.; Hu, X.; Lyu, X.; Fang, H.; Zhou, Y.-M.; Mahmoud, A.; Deng, G.; Meng, Y.-Q.; et al. Chromosome-Level Genome Assembly of Bunching Onion Illuminates Genome Evolution and Flavor Formation in Allium Crops. *Nat. Commun.* **2022**, *13*, 6690. [CrossRef] [PubMed]
89. Al-Safadi, B.; Arabi, M.; Ayyoubi, Z. Differences in Quantitative and Qualitative Characteristics of Local and Introduced Cultivars and Mutated Lines of Garlic. *J. Veg. Crop Prod.* **2003**, *9*, 21–31. [CrossRef]

90. Mishra, S.S.; Ram, C.; Chakravati, S.K.; Vishwakarma, M.K.; Singh, P.K.; Singh, V.B. Genetic Divergence Studies in Garlic (*Allium sativum* L.) through Morphological Features. *Int. J. Curr. Microbiol. Appl. Sci.* **2018**, *7*, 1013–1020. [CrossRef]
91. Rout, E.; Nanda, S.; Nayak, S.; Joshi, R.K. Molecular Characterization of NBS Encoding Resistance Genes and Induction Analysis of a Putative Candidate Gene Linked to Fusarium Basal Rot Resistance in *Allium sativum*. *Physiol. Mol. Plant Pathol.* **2014**, *85*, 15–24. [CrossRef]
92. Martins, N.; Petropoulos, S.; Ferreira, I.C. Chemical Composition and Bioactive Compounds of Garlic (*Allium sativum* L.) as Affected by Pre-and Post-Harvest Conditions: A Review. *Food Chem.* **2016**, *211*, 41–50. [CrossRef]
93. Khandagale, K.; Krishna, R.; Roylawar, P.; Ade, A.B.; Benke, A.; Shinde, B.; Singh, M.; Gawande, S.J.; Rai, A. Omics Approaches in Allium Research: Progress and Way Ahead. *PeerJ* **2020**, *8*, e9824. [CrossRef]
94. Rocchetti, G.; Zhang, L.; Bocchi, S.; Giuberti, G.; Ak, G.; Elbasan, F.; Yıldıztugay, E.; Ceylan, R.; Picot-Allain, M.C.N.; Mahomoodally, M.F. The Functional Potential of Nine Allium Species Related to Their Untargeted Phytochemical Characterization, Antioxidant Capacity and Enzyme Inhibitory Ability. *Food Chem.* **2022**, *368*, 130782. [CrossRef]
95. Kodera, Y.; Ushijima, M.; Amano, H.; Suzuki, J.; Matsutomo, T. Chemical and Biological Properties of S-1-Propenyl-l-Cysteine in Aged Garlic Extract. *Molecules* **2017**, *22*, 570. [CrossRef]
96. Singiri, J.R.; Swetha, B.; Ben-Natan, A.; Grafi, G. What Worth the Garlic Peel. *Int. J. Mol. Sci.* **2022**, *23*, 2126. [CrossRef]
97. Liu, P.; Weng, R.; Xu, Y.; Feng, Y.; He, L.; Qian, Y.; Qiu, J. Metabolic Changes in Different Tissues of Garlic Plant during Growth. *J. Agric. Food Chem.* **2020**, *68*, 12467–12475. [CrossRef]
98. Eady, C.C.; Kamoi, T.; Kato, M.; Porter, N.G.; Davis, S.; Shaw, M.; Kamoi, A.; Imai, S. Silencing Onion Lachrymatory Factor Synthase Causes a Significant Change in the Sulfur Secondary Metabolite Profile. *Plant Physiol.* **2008**, *147*, 2096–2106. [CrossRef]
99. Kato, M.; Masamura, N.; Shono, J.; Okamoto, D.; Abe, T.; Imai, S. Production and Characterization of Tearless and Non-Pungent Onion. *Sci. Rep.* **2016**, *6*, 23779. [CrossRef]
100. Le, D.; Audenaert, K.; Haesaert, G. Fusarium Basal Rot: Profile of an Increasingly Important Disease in *Allium* spp. *Trop. Plant Pathol.* **2021**, *46*, 241–253. [CrossRef]
101. Filyushin, M.A.; Anisimova, O.K.; Kochieva, E.Z.; Shchennikova, A.V. Genome-Wide Identification and Expression of Chitinase Class I Genes in Garlic (*Allium sativum* L.) Cultivars Resistant and Susceptible to Fusarium Proliferatum. *Plants* **2021**, *10*, 720. [CrossRef]
102. Gálvez, L.; Palmero, D. Fusarium Dry Rot of Garlic Bulbs Caused by Fusarium Proliferatum: A Review. *Horticulturae* **2022**, *8*, 628. [CrossRef]
103. Chand, S.K.; Nanda, S.; Mishra, R.; Joshi, R.K. Multiple Garlic (*Allium sativum* L.) MicroRNAs Regulate the Immunity against the Basal Rot Fungus Fusarium Oxysporum f. Sp. Cepae. *Plant Sci.* **2017**, *257*, 9–21. [CrossRef]
104. Chand, S.K.; Nanda, S.; Joshi, R.K. Regulation of MiR394 in Response to Fusarium Oxysporum f. Sp. Cepae (FOC) Infection in Garlic (*Allium sativum* L.). *Front. Plant Sci.* **2016**, *7*, 258. [CrossRef]
105. Gálvez, L.; Palmero, D. Incidence and Etiology of Postharvest Fungal Diseases Associated with Bulb Rot in Garlic (*Alllium sativum*) in Spain. *Foods* **2021**, *10*, 1063. [CrossRef]
106. Edel-Hermann, V.; Lecomte, C. Current Status of Fusarium Oxysporum Formae Speciales and Races. *Phytopathology* **2019**, *109*, 512–530. [CrossRef] [PubMed]
107. Cramer, C.S. Breeding and Genetics of Fusarium Basal Rot Resistance in Onion. *Euphytica* **2000**, *115*, 159–166. [CrossRef]
108. Molinero-Ruiz, L.; Rubio-Pérez, E.; González-Domínguez, E.; Basallote-Ureba, M.J. Alternative Hosts for *Fusarium* Spp. Causing Crown and Root Rot of Asparagus in Spain. *J. Phytopathol.* **2011**, *159*, 114–116. [CrossRef]
109. Anisimova, O.K.; Kochieva, E.Z.; Shchennikova, A.V.; Filyushin, M.A. Thaumatin-like Protein (TLP) Genes in Garlic (*Allium sativum* L.): Genome-Wide Identification, Characterization, and Expression in Response to *Fusarium proliferatum* Infection. *Plants* **2022**, *11*, 748. [CrossRef] [PubMed]
110. Anisimova, O.K.; Shchennikova, A.V.; Kochieva, E.Z.; Filyushin, M.A. Pathogenesis-Related Genes of PR1, PR2, PR4, and PR5 Families Are Involved in the Response to *Fusarium* Infection in Garlic (*Allium sativum* L.). *Int. J. Mol. Sci.* **2021**, *22*, 6688. [CrossRef] [PubMed]
111. Dar, A.A.; Sharma, S.; Mahajan, R.; Mushtaq, M.; Salathia, A.; Ahamad, S.; Sharma, J.P. Overview of Purple Blotch Disease and Understanding Its Management through Chemical, Biological and Genetic Approaches. *J. Integr. Agric.* **2020**, *19*, 3013–3024. [CrossRef]
112. Entwistle, A.R. Allium White Rot and Its Control. *Soil Use Manag.* **1990**, *6*, 201–208. [CrossRef]
113. Gálvez, L.; Gil-Serna, J.; García, M.; Iglesias, C.; Palmero, D. Stemphylium Leaf Blight of Garlic (*Allium sativum*) in Spain: Taxonomy and in Vitro Fungicide Response. *Plant Pathol. J.* **2016**, *32*, 388. [CrossRef]
114. Lorbeer, J.W.; Seyb, A.M.; de Boer, M.; van den Ende, J.E. Botrytis species on bulb crops. In *Botrytis: Biology, Pathology and Control*; Elad, Y., Williamson, B., Tudzynski, P., Delen, N., Eds.; Springer: Dordrecht, The Netherlands, 2007; pp. 273–294. ISBN 978-1-4020-2626-3.
115. Greathead, A. Control of Penicillium Decay of Garlic. *Calif. Agric.* **1978**, *32*, 18.
116. Koike, S.T.; Smith, R.F.; Davis, R.M.; Nunez, J.J.; Voss, R.E. Characterization and Control of Garlic Rust in California. *Plant Dis.* **2001**, *85*, 585–591. [CrossRef]
117. Chand, S.K.; Nanda, S.; Joshi, R.K. Genetics and Molecular Mapping of a Novel Purple Blotch-Resistant Gene ApR1 in Onion (*Allium cepa* L.) Using STS and SSR Markers. *Mol. Breed.* **2018**, *38*, 109. [CrossRef]

118. Lagunes-Fortiz, E.; Robledo-Paz, A.; Gutiérrez-Espinosa, M.A.; Mascorro-Gallardo, J.O.; Espitia-Rangel, E. Genetic Transformation of Garlic (*Allium sativum* L.) with Tobacco Chitinase and Glucanase Genes for Tolerance to the Fungus *Sclerotium cepivorum*. *Afr. J. Biotechnol.* **2013**, *12*, 3482–3492.
119. Kudryavtseva, N.; Havey, M.J.; Black, L.; Hanson, P.; Sokolov, P.; Odintsov, S.; Divashuk, M.; Khrustaleva, L. Cytological Evaluations of Advanced Generations of Interspecific Hybrids between *Allium cepa* and *Allium fistulosum* Showing Resistance to *Stemphylium vesicarium*. *Genes* **2019**, *10*, 195. [CrossRef]
120. Lee, H.-M.; Park, J.-S.; Kim, S.-J.; Kim, S.-G.; Park, Y.-D. Using Transcriptome Analysis to Explore Gray Mold Resistance-Related Genes in Onion (*Allium cepa* L.). *Genes* **2022**, *13*, 542. [CrossRef]
121. Wako, T.; Yamashita, K.; Tsukazaki, H.; Ohara, T.; Kojima, A.; Yaguchi, S.; Shimazaki, S.; Midorikawa, N.; Sakai, T.; Yamauchi, N.; et al. Screening and Incorporation of Rust Resistance from *Allium cepa* into Bunching Onion (*Allium fistulosum*) via Alien Chromosome Addition. *Genome* **2015**, *58*, 135–142. [CrossRef]
122. Poromarto, S.; Widono, S.; Septiriani, D.; Hermawan, K. *Decrease in Population of Ditylenchus dipsaci in Garlic Cultivation with the Application of Mycorrhizae and Organic Fertilizers*; IOP Publishing: Bristol, UK, 2022; Volume 1114, p. 012062.
123. Roberts, P.; Greathead, A. Control of *Ditylenchus dipsaci* in Infected Garlic Seed Cloves by Nonfumigant Nematicides. *J. Nematol.* **1986**, *18*, 66. [PubMed]
124. Correia, G.S.; de Araujo Filho, J.V.; da Silva, W.R.; Moccellin, R.; Resende, F.V.; Pinheiro, J.B.; da Grinberg, P.S.; Gomes, C.B. Reaction of Garlic Genotypes to *Ditylenchus dipsaci* and Aspects Related to Productivity in a Naturally Infested Area. *Hortic. Bras.* **2023**, *40*, 451–456. [CrossRef]
125. Koch, M.; Salomon, R. Improvement of garlic via somaclonal variation and virus elimination. In Proceedings of the International Symposium on Alliums for the Tropics, Bangkok, Thailand, 15–19 February 1993; pp. 211–214.
126. Atighi, M.R.; Verstraeten, B.; De Meyer, T.; Kyndt, T. Genome-wide DNA Hypomethylation Shapes Nematode Pattern-triggered Immunity in Plants. *New Phytol.* **2020**, *227*, 545–558. [CrossRef]
127. Debnath, P.; Karmakar, K. Garlic Mite, *Aceria tulipae* (Keifer) (Acari: Eriophyoidea)—A Threat for Garlic in West Bengal, India. *Int. J. Acarol.* **2013**, *39*, 89–96. [CrossRef]
128. Dąbrowska, E.; Lewandowski, M.; Koczkodaj, S.; Paduch-Cichal, E. Transmission of Garlic Virus B, Garlic Virus C, Garlic Virus D and Garlic Virus X by *Aceria tulipae* (Keifer) in Leek. *Eur. J. Plant Pathol.* **2020**, *157*, 215–222. [CrossRef]
129. Kang, S.-G.; Koo, B.-J.; Lee, E.-T.; Chang, M.-U. Allexivirus Transmitted by Eriophyid Mites in Garlic Plants. *J. Microbiol. Biotechnol.* **2007**, *17*, 1833–1840. [PubMed]
130. Yamashita, K.; Sakai, J.; Hanada, K. Characterization of a New Virus from Garlic (*Allium sativum* L.), Garlic Mite-Borne Mosaic Virus. *Jpn. J. Phytopathol.* **1996**, *62*, 483–489. [CrossRef]
131. Sapáková, E.; Hasíková, L.; Hřivna, L.; Stavělíková, H.; Šefrová, H. Infestation of Different Garlic Varieties by Dry Bulb Mite *Aceria tulipae* (Keifer) (Acari: Eriophyidae). *Acta Univ. Agric. Silvic. Mendel. Brun.* **2012**, *60*, 293–302. [CrossRef]
132. Karuppaiah, V.; Soumia, P.; Wagh, P.D.; Singh, M. *Ephestia cautella* (Lepidoptera: Pyralidae): An Emerging Pest on Garlic in Storage. *J. Entomol. Zool. Stud.* **2018**, *6*, 2282–2285.
133. Srinivas, P.S. Pests and their management in onion and garlic. In *Trends in Horticultural Entomology*; Mani, M., Ed.; Springer Nature: Singapore, 2022; pp. 1177–1187, ISBN 978-981-19034-3-4.
134. Adams, M.J.; Antoniw, J.F.; Bar-Joseph, M.; Brunt, A.A.; Candresse, T.; Foster, G.D.; Martelli, G.P.; Milne, R.G.; Fauquet, C.M. Virology Division News: The New Plant Virus Family Flexiviridae and Assessment of Molecular Criteria for Species Demarcation. *Arch. Virol.* **2004**, *149*, 1045–1060. [CrossRef]
135. Lunello, P.; Di Rienzo, J.; Conci, V.C. Yield Loss in Garlic Caused by Leek Yellow Stripe Virus Argentinean Isolate. *Plant Dis.* **2007**, *91*, 153–158. [CrossRef]
136. Chen, J.; Chen, J.P.; Adams, M. Characterisation of Some Carla-and Potyviruses from Bulb Crops in China. *Arch. Virol.* **2002**, *147*, 419–428. [CrossRef]
137. Fajardo, T.V.; Nishijima, M.; Buso, J.A.; Torres, A.C.; Ávila, A.C.; Resende, R.O. Garlic Viral Complex: Identification of Potyviruses and Carlavirus in Central Brazil. *Fitopatol. Bras.* **2001**, *26*, 619–626. [CrossRef]
138. Valli, A.; García, J.A.; López-Moya, J.J. Potyviridae. *eLS* **2015**, 1–10. [CrossRef]
139. Turina, M.; Kormelink, R.; Resende, R.O. Resistance to Tospoviruses in Vegetable Crops: Epidemiological and Molecular Aspects. *Annu. Rev. Phytopathol.* **2016**, *54*, 347–371. [CrossRef] [PubMed]
140. Da Silva, L.A.; Oliveira, A.S.; Melo, F.L.; Ardisson-Araujo, D.M.; Resende, F.V.; Resende, R.O.; Ribeiro, B.M. A New Virus Found in Garlic Virus Complex Is a Member of Possible Novel Genus of the Family Betaflexiviridae (Order Tymovirales). *PeerJ* **2019**, *7*, e6285. [CrossRef] [PubMed]
141. Tsuneyoshi, T.; Matsumi, T.; Deng, T.; Sako, I.; Sumi, S. Differentiation of Allium Carlaviruses Isolated from Different Parts of the World Based on the Viral Coat Protein Sequence. *Arch. Virol.* **1998**, *143*, 1093–1107. [CrossRef]
142. Katis, N.I.; Maliogka, V.I.; Dovas, C.I. Chapter 5—Viruses of the genus *Allium* in the Mediterranean region. In *Advances in Virus Research*; Loebenstein, G., Lecoq, H., Eds.; Academic Press: Cambridge, MA, USA, 2012; Volume 84, pp. 163–208. ISBN 0065-3527.
143. Bag, S.; Schwartz, H.F.; Cramer, C.S.; Havey, M.J.; Pappu, H.R. Iris Yellow Spot Virus (Tospovirus: Bunyaviridae): From Obscurity to Research Priority. *Mol. Plant Pathol.* **2015**, *16*, 224–237. [CrossRef]

144. Prajapati, M.R.; Manav, A.; Singh, J.; Kumar, P.; Kumar, A.; Kumar, R.; Prakash, S.; Baranwal, V.K. Identification and Characterization of a Garlic Virus E Genome in Garlic (*Allium sativum* L.) Using High-Throughput Sequencing from India. *Plants* **2022**, *11*, 224. [CrossRef]
145. Roumagnac, P.; Gagnevin, L.; Gardan, L.; Sutra, L.; Manceau, C.; Dickstein, E.R.; Jones, J.B.; Rott, P.; Pruvost, O. Polyphasic Characterization of Xanthomonads Isolated from Onion, Garlic and Welsh Onion (*Allium* spp.) and Their Relatedness to Different Xanthomonas Species. *Int. J. Syst. Evol. Microbiol.* **2004**, *54*, 15–24. [CrossRef]
146. Conci, V.C.; Gomez, G.G.; Docampo, D.M.; Conci, L.R. Phytoplasma Associated with Symptoms of 'Tristeza Del Ajo' (Garlic Decline) in Garlic (*Allium sativum* L.). *J. Phytopathol.* **1998**, *146*, 473–477. [CrossRef]
147. Parthasarathy, S.; Rajamanickam, S.; Muthamilan, M. Allium Diseases: A Global Perspective. *Innov. Farming* **2016**, *1*, 171–178.
148. Gardan, L.; Bella, P.; Meyer, J.-M.; Christen, R.; Rott, P.; Achouak, W.; Samson, R. *Pseudomonas salomonii* sp. Nov., Pathogenic on Garlic, and *Pseudomonas palleroniana* sp. Nov., Isolated from Rice. *Int. J. Syst. Evol. Microbiol.* **2002**, *52*, 2065–2074.
149. Li, B.; Yu, R.R.; Yu, S.H.; Qiu, W.; Fang, Y.; Xie, G.L. First Report on Bacterial Heart Rot of Garlic Caused by *Pseudomonas fluorescens* in China. *Plant Pathol J.* **2009**, *25*, 91–94. [CrossRef]

# *De novo* assembly and annotation of the organellar genomes of garlic (*Allium sativum* L.)

Ricardo Parreño-Montoro[1], Eva Rodríguez-Alcocer[1], Ainoa Gallego-Zaragoza[1], Álvaro Ferriz[1], Purificación Castillo Martínez[2], Felipe Gómez del Castillo[2] Sara Jover-Gil and Héctor Candela[1]

[1] Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Spain

[2] COOPAMAN , S.C.L., C/General Borrero, s/n., 16660, Las Pedroñeras, Spain



Corresponding author: H. Candela (telephone: 34 96 522 25 83; fax: 34 96 665 85 11; E-mail: hcandela@umh.es)

Running head: The organellar genomes of garlic

Keywords: *Allium sativum*, garlic, mitochondrial genome, chloroplast genome

‘

**ABSTRACT**

*Allium sativum* is a cultivated plant that is highly appreciated for the commercial value of its bulbs. Overcoming the infertility of garlic cultivars is the most important challenge in garlic cultivation and would be a solution to problems such as the loss of genetic variation, the accumulation of pathogens in propagules or the difficulties to develop new cultivars, all of them occurring directly or indirectly as a consequence of the vegetative mode of propagation.

In order to generate resources for the development of garlic as a modern crop, we have carried out the sequencing, de novo assembly and annotation of the complete mitochondrial and chloroplast genomes of Allium sativum. Our efforts to annotate the garlic mitochondrial genome have allowed us to identify more genes than in the not-so-closely related mitochondrial genome of onion, while the chloroplast genome was very similar between these two species.

The chloroplast genome was very similar to those already deposited by other authors, but we found a lot of non-synonymous post-transcriptional modifications in this plast, and we reconstructed operon-like transcripts.

In agreement with current models of mitochondrial genome structure in plants, or results demonstrate that the garlic mitochondrial genome consists of numerous subgenomic molecules that undergo a complex recombination pattern at specific sequences. The recombination sites can be effectively detected as shared paths in the assembly graph. Our findings suggest that recombination actively and specifically occurs at short repeated sequences, which shape a dynamic genomic landscape where up to four distinct recombination products co-exist at each repeat.

**Introduction**

Garlic breeding is hampered by its vegetative mode of reproduction, as most cultivars have lost the ability to reproduce sexually and produce true seeds. Because crosses cannot be performed, the development of new cultivars has been mainly limited to the selection of spontaneous variants that are propagated clonally. A deeper knowledge about the molecular genetics of garlic should help breeders to develop new cultivars that are resistant to pests or that can cope with the new challenges posed by the climate change conditions. The availability of different types of molecular markers previosly allowed researchers to assess the intraspecific diversity, investigate the evolution, and build low-resolution linkage maps of the garlic genome (Buso et al., 2008; Chand et al., 2015; Cunha et al., 2012; García-Lampasona et al., 2012; Ipek et al., 2015, 2005, 2003; Ipek and Simon, 1998; Jo et al., 2012; Zewdie et al., 2005). More recently, next-generation sequencing technologies have granted access to the transcriptome of different tissues and organs in garlic and related species (Kamenetsky et al., 2015; D.-W. Kim et al., 2009; Shemesh-Mayer et al., 2015; Sun et al., 2012, 2013). Furthermore, the nuclear genome of garlic and other *Allium* crops has recently been released (Sun et al., 2020), opening new avenues for developing garlic as a modern crop. In garlic, the nuclear genome is approximately 16.26 Gbp in size (Sun et al., 2020), about five times larger than the human genome.

The presence of multiple copies of the mitochondrial and chloroplast genomes in plant cells facilitates their assembly using samples of genomic DNA that have been sequenced at low sequencing depth. The use of paired-end reads also facilitates the characterization of distinct structural variants. Indeed, the genomes of plant mitochondria are much more complex and structurally diverse than those of animals, ranging in size from 66 kb to 11.3 Mb (Gualberto et al., 2014; Knoop, 2004; Skippington et al., 2015; Sloan et al., 2012). While the mitochondrial genomes of animals are compact and circular, those of plants are multipartite, typically comprising multiple subgenomic

molecules that can recombine and are often depicted as a single, ideal circle that is usually referred to as the ‘master circle’ (Kozik et al., 2019). In contrast to this, electron microscopy experiments have shown that plant mitochondrial genomes might consist of multiple circular, linear, and branched DNA molecules.

The mitochondrial genomes of animals present a high gene density due to their compact size and the absence of most introns (Boore, 1999; Lavrov and Pett, 2016). By contrast, plant mitochondrial genomes are larger due to the presence of repetitive sequences, long introns and sequences imported from other genomes, including the nuclear, chloroplast, viral and bacterial genomes (Alverson et al., 2011; Aono et al., 2002; Unseld et al., 1997). The abundance of repetitive sequences is characteristic of angiosperm mitochondrial genomes and is the basis of a phenomenon known as ‘substoichiometrical shifting’ (Chen et al., 2011; Woloszynska, 2010), which refers to the co-existence of mitochondria with different DNA content in the same individual. The high frequency of recombination events in plant mitochondrial genomes has been proposed to be the main cause of their rapid structural evolution and might explain the existence of molecules with lineal, circular and branched topologies, whose role and biological significance is not fully understood yet (Oldenburg and Bendich, 2001). Studies of the mitochondrial genome of onion (*Allium cepa*) have previously uncovered important differences in its size and structure (Kim and Yoon, 2010; Tsujimura et al., 2019).

In this work, we report the characterization of the chloroplast and mitochondrial genomes of a garlic clone widely cultivated in Spain, Spring white, which we sequenced and assembled using paired-end Illumina reads. Understanding the structural complexity and dynamics of plant mitochondrial genomes remains an open question. The occurrence of recombination at repeated sequences suggests that mitochondrial genomes with many different topologies might co-exist. In this context, assembly graphs offer a compact, ideal solution to the problem of representing the structural diversity of the mitochondrial genome.

## Results

### Sequencing and *de novo* assembly of the garlic chloroplast

A sample of genomic DNA isolated from young leaf tissue was sequenced with the Illumina HiSeq2500 platform using paired-end 101-nt reads. In this sample, reads derived from the chloroplast genome were more abundant than those from the nuclear and mitochondrial genomes. Based on this observation, we assembled the chloroplast genome using the Velvet assembler (version 1.2.10) (Zerbino and Birney, 2008), using 42,459,814 read pairs and options that favoured the assembly of high-coverage contigs. As a result, we obtained 5,319 contigs with a cumulative length of 503,236 bp, which incorporated 10,186,160 reads (11.99%) putatively derived either from the organellar genomes or from the repetitive fraction of the nuclear genome. We identified three different contigs matching each of the three characteristic regions of chloroplast genomes (Figure 1): two of them were 18,142 and 82,086 bp long, respectively, and had similar sequencing depth (370.79 and 377.80). The third contig was 26,490 nucleotides long and had a sequencing depth of 787.79 (approximately twice the value for the other two contigs). Considering the coverage and the overlaps of these contigs, we reconstructed a circular chloroplast genome that was 153,131 bp long, with a 82,006 bp large single copy (LSC) region, a 18,045 bp small single copy region (SSC), and two 26,540 bp inverted repeats (IRa and IRb) that correspond to the high-coverage contig (Figure 2). Our sequence was 99% identical to the chloroplast genomes of other garlic accessions that had been previously reported (Filyushin et al., 2016), and was deposited in GenBank with accession number KY363332.

### Assembly of the mitochondrial genome with Velvet

To assemble the mitochondrial genome, we also took advantage of the higher abundance of mitochondrial reads compared to nuclear reads. We found the average sequencing depth of mitochondrial contigs to be approximately 18x while it ranged from <1x to 5x in nuclear contigs. Because plant mitochondrial genomes often incorporate

sequences of plastid origin (Kim et al., 2016; Notsu et al., 2002), we were not able to set a coverage threshold to differentiate the reads derived from the mitochondrial genome from the reads of the chloroplast genome. For this reason, we performed a new assembly with Velvet using parameters that favored the joint assembly of both organellar genomes, setting the hash length to 65, the expected coverage to 18, and the minimum coverage to 8. The resulting assembly graph consisted of 457,117 nodes (contigs) and was examined using Bandage software (Wick et al., 2015). A connected subgraph comprising 671 contigs (Figure 3A) was found to comprise sequences putatively derived from both the chloroplast and mitochondrial genomes, as well as some sequences that were most likely derived from the highly repetitive fraction of the nuclear genome (i.e. rRNA genes). Using the BLASTn implementation in Bandage, we identified a cycle through this subgraph that had the characteristic topology and exactly matched the known sequence of the chloroplast genome, which we had previously assembled (shown in green in Figure 3A). As expected, the mitochondrial and chloroplast genomes shared some common sequences. We subtracted the chloroplast subgraph from the main graph based on their sequencing depth and the BLASTn alignments, duplicating nodes as needed to preserve the connectivity of the mitochondrial genome graph (Figure 1B). In our sample, the sequencing depth of the mitochondrial nodes usually ranged from 15x to 30x. However, one part of the resulting graph (referred to as the 'rRNA loop' and shown in grey in Figure 3A and in red in Figure 3B) comprised contigs with sequencing depths much higher than those expected for the mitochondrial genome. These contigs were found to include sequences putatively derived from the nuclear 26S, 18S and 5.8S rRNA genes, as well as from the internal transcribed spacer regions 1 and 2 (ITS1 and ITS2). We designed oligonucleotides based on the topology of the assembly graph and the known sequence of the contigs flanking the rRNA loop. PCR amplification using primers (rRNA_1_F and rRNA_1_R; Supplemental Table 1) yielded a single ~4 kb product that was fully sequenced with these and three additional primers (rRNA_2_F, rRNA_2_R, and rRNA_3_F; Supplemental Table 1) using the Sanger method. The resulting

sequence was incorporated to the graph in place of the rRNA loop (Figure 3C). The resulting Velvet assembly consisted of 69 nodes, with a total length of 540,313 bp.

**Structural complexity of the garlic mitochondrial genome**

The assembly graph shows that many contigs converge at short sequences (shown as red nodes in Figure 3C) that had an associated sequencing depth higher than the flanking contigs. These nodes correspond to short repeated sequences and potentially allow many different reconstructions of the genome sequence. To select paths supported by experimental evidence, we systematically followed two independent, complementary approaches. On the one hand, we designed primers flanking each short repeat (Supplemental Table 1; “Rep” primers), setting the expected amplicon size to ~1 kb. This approach yielded up to four distinct amplification products in most cases and was soon abandoned, as we found at least some PCR products to be artifacts due to PCR-mediated recombination, in line with previous results (Alverson et al., 2011). On the other hand, we used the Bowtie2 read aligner (Langmead and Salzberg, 2012) to map the read pairs back to the contigs. The resulting alignment files (in SAM format) were parsed to identify read pairs in which each read mapped to a different contig placed along a putatively valid path. Using this information, we were able to further reduce the number of contigs to 21, with a total length of 536,232 bp (Figure 3D). The sequence of these contigs has been deposited in GenBank with accession numbers XXXXX-YYYYY. Although we have identified a linear path through this graph, its occurrence in nature has not been confirmed, and hence it might not reflect the real structure of this genome.

Additionally, we performed a second assembly using SPAdes (Prjibelski et al., 2020), with *k*-mer lengths from 51 to 75. The resulting graph was more complex than the Velvet graph, connecting contigs that were not connected and generating new branches. The size of the SPAdes assembly was larger than that of the Velvet assembly (606,978 bp).

**Annotation and analysis of the mitochondrial genome**

The annotation of the mitochondrial genome allowed us to identify most genes previously found in other plant mitochondrial genomes (Table 1). Protein-coding genes were identified using BLAST searches (see Materials and Methods) and GeSeq inside Chlorobox web application (Tillich et al., 2017). The assembled contigs contained 26 protein-coding genes, 13 tRNA genes and 3 rRNA genes. Seven protein-coding genes (*ccmFc*, *cox2*, *nad1*, *nad2*, *nad4*, *nad5* and *nad7*) were found to contain introns. For some genes (*cox2*, *nad1*, *nad2* and *nad5*), exons were located in different contigs of the assembly, suggesting that their transcripts experience *trans*-splicing, as previously described for many mitochondrial transcripts (Bonen, 2008; Glanz and Kück, 2009; Kim and Yoon, 2010; Wissinger et al., 1991). The set of protein-coding genes encode proteins involved in the electron transport chain, the biogenesis of cytochrome c and other processes characteristic of plant mitochondria. The fact that numerous protein-coding genes, such as *nad3*, *shd4*, *rpl2*, *rpl10*, *rpl16*, *rps3* and *rps4*, are absent from both the garlic and onion mitochondrial genomes (S. Kim et al., 2009) suggests that these genes had already been lost prior to their latest common ancestor. Like in onion, the $ccmF_N$ gene has split into two genes, $ccmF_{N1}$ and $ccmF_{N2}$ in the garlic mitochondrial genome. The 13 tRNA genes correspond to an incomplete set of 10 different anticodons. The most abundant anticodon (CAU) was present in 4 tRNAs. Despite sharing the same anticodon, these tRNAs are likely to correspond to three distinct amino acids (methionine, N-formyl methionine and isoleucine) on the grounds of their sequence similarity to other previously described tRNAs (Alkatib et al., 2012a). Indeed, mitochondrial proteins have long been known to incorporate N-formyl methionine as their first amino acid (Schwartz et al., 1967; Smith and Marcker, 1968).

**Annotation and analysis of the chloroplast genome**

We followed a combination of *ab initio* and homology-based approaches to annotate the chloroplast genome. Our results largely matched those of previous authors (Filyushin

et al., 2016), with some differences. We annotated 136 functional genes and 7 pseudogenes (Table 2 and Figure 2). Of these, 90 genes encode proteins involved in various processes and 46 correspond to non-coding RNA molecules, including 38 tRNAs and 8 rRNAs. Because the genes located in the inverted repeats are present in two copies, the garlic chloroplast genome contains 115 unique genes: 94 single-copy genes (including 72 protein-coding and 22 tRNA genes) and 21 genes that are present in two copies (including 9 protein-coding, 8 tRNA and 4 rRNA genes). Of the 136 functional genes, 23 contain introns, including 14 protein-coding genes, 6 of them present in a single copy (*rpoC1*, and *cf3*, *ndhA*, *atpF*, *clpP* and *rps16*) and 4 present in two copies (*rpl2*, *ndhB*, and *cf68* and *rps12*). The remaining 9 genes correspond to tRNA molecules, 5 of which are single-copy genes (*trnE*-UUC, *trnA*-UUU, *trnL*-UAA, *trnS*-CGA and *trnV*-UAC) and 2 are present in two copies (*trnA*-UGC and *trnI*-GAU). Except for the *ycf3* and *clpP* genes, each with two introns, all the indicated genes contain a single intron.

We found many genes with functions related to gene transcription and translation. The chloroplast genome contains the *rpoA*, *rpoB*, *rpoC1* and *rpoC2* (Table 2) genes, respectively encoding the α, β, β' and β" subunits of the plastid-encoded RNA polymerase (PEP). Of the 33 proteins described in the 50S subunit of the chloroplast ribosome in higher plants (Yamaguchi and Subramanian, 2000), the garlic chloroplast genome encodes only 9 (*rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33* and *rpl36* genes). Out of the 25 proteins described in the 30S subunit (Yamaguchi et al., 2000), the garlic chloroplast genome encodes only 10 (*rps3*, *rps4*, *rps7*, *rps8, rps11*, *rps12*, *rps14*, *rps15*, *rps18* and *rps19*). We also identified pseudogenized sequences similar to the *rps2* and *rps16* genes. The four ribosomal RNA genes (*rrn5*, *rrn4.5*, *rrn23* and *rrn16*) are located in the inverted repeats and are therefore present in two copies.

Translation of genes located in the chloroplast genome requires a complement of tRNA molecules that is able to recognize the 61 non-stop codons. The formation of non-canonical G/U base pairs between the third base of a codon and the first base of an anticodon (wobbling) is thought to allow a minimal complement of 32 different tRNAs to

read all codons (Alkatib et al., 2012b). However, we only identified 30 unique tRNA genes (*trn*), 22 of them present in a single copy and 8 present in 2 copies (Table 3). In many cases, wobbling can explain the recognition of codons that are not perfectly complementary to the anticodons, including: UUU (*trnF*-GAA), UCU (*trnS*-GGA), UAU (*trnY*-GUA), UGU (*trnC*-GCA), CUG (*trnL*-UAG), CCG (*trnP*-UGG), CAU (*trnH*-GUG), CAG (*trnQ*-UUG), AUU (*trnI*-GAU), ACU (*trnT*-GGU), ACG (*trnT*-UGU), AAU (*trnN*-GUU), AAG (*trnK*-UUU), AGU (*trnS*-GCU), AGG (*trnR*-UCU), GUU (*trnV*-GAC), GUG (*trnV*-UAC), GCG (*trnA*-UGC), GAU (*trnD*-GUC), GAG (*trnE*-UUC), GGU (*trnG*-GCC) and GGG (*trnG*-UCC). Three of the 16 sections in Suppleme ntary Table 2 (CUN, CCN and GCN) are represented, respectively, by a single anticodon. The reading of the CUU, CUC, CCU, CCC, GCU and GCC codons might be explained by 'superwobbling' (Alkatib et al., 2012b), whereby an anticodon whose first base is U allows can recognize four different codons. Thus, the CUN codons could be read by *trnL*-UAG, CCN by *trnP*-UGG, and GCN by *trnA*-UGC. Only three of the 61 codons (CGC, CGA and CGG), all of which correspond to arginine, appear to lack tRNA molecules that enable their reading, and might be imported from the nucleus. Three tRNA molecules with the same anticodon (CAU) play distinct roles. Two of them, *trnM*-CAU and *trn(f)M*-CAU recognize the AUG codon, respectively corresponding to methionine and formyl methionine. The differences in sequence and secondary structure of these tRNAs, however, determine that *trn(f)M*-CAU is only used to initiate protein synthesis. Lysination of C in the *trnI*-CAU anticodon has been reported to allow the reading of the AUA codon (isoleucine) but not of the AUG codon, thus preventing the "wobble" of the anticodon from interfering with the reading of codons corresponding to methionine (Alkatib et al., 2012a). Supplemental Figure 1 illustrates the differences in sequence and secondary structure between the different tRNAs for methionine, formyl methionine and isoleucine identified in the garlic chloroplast genome.

A large group of genes encodes proteins that perform photosynthesis-related functions, including subunits of photosystems I and II (Table 2). The garlic genome

contains genes for five subunits of photosystem I (out of 15 described in higher plants, Jensen et al., 2007): *PsaA*, *PsaB*, *PsaC*, *PsaI* and *PsaJ*, exactly the same as in the *Arabidopsis thaliana* chloroplast genome. The *ycf3* and *ycf4* genes encode conserved proteins that have been reported to participate in the biogenesis of photosystem I in *Chlamydomonas reinhardtii* (Boudreau et al., 1997). We also annotated genes encoding 15 subunits of photosystem II, 11 subunits of the NADH dehydrogenase complex, 6 subunits of the cytochrome b/f complex, 6 subunits of the ATP synthase complex, and 1 subunit of ribulose 1,5-bisphosphate carboxylase-oxygenase (RuBisCO). All these genes are present in a single copy, except for the *ndhB* gene, which is located within the inverted repeats (Figure 2). The genome contains other genes encoding proteins with diverse functions: *matK*, *clpP*, *cemA*, *accD* and *ccsA*. In addition to these genes, we identified three conserved ORFs in the inverted repeats, called *ycf1*, *ycf2* and *ycf68*, whose functions are not fully understood. In addition, we classified 6 sequences as pseudogenes, as their sequences have accumulated mutations and/or frameshifts. Four of these pseudogenes are present in a single copy and one is present in two copies.

**Detection of editing events using RNA-seq data**

RNA-seq paired-end reads were obtained from samples of bulbs, bulbils, leafs and roots. We aligned these reads to the mitochondrial and chloroplast genomes using Hisat2 (Kim et al., 2019). Variant calling was performed separately on the read pairs derived from each strand using the Samtools mpileup and Bcftools call commands (see Materials and Methods), and the effect of each nucleotide substitution on the protein sequence was assessed using a pipeline based on the annotate command of MAPtools. We detected abundant editing events both in the mitochondrial and the chloroplast genomes (Supplemental Tables 3 and 4). Out of 72 editing events detected in the chloroplast genome, 57 were located in protein-coding sequences, 1 in a tRNA gene, and the remaining 15 were located in other non-coding sequences. Regarding the effect of the 57 editing events occurring in protein-coding sequences, 2 were predicted to be

synonymous, 1 created the stop codon of the *petD* gene, and 53 led to a diversity of amino acid substitutions. Importantly, two of these substitutions created the translation start sites of the *ndhD* and *ycf3* genes. Editing of the mitochondrial transcripts created the stop codon of the *atp9*, and *ccmFc* transcripts, and the translation initiation codon of *nad1* and *nad4L*. Editing also created the start codons of *nad1* and *nadL4* in onion (Tsujimura et al 2019). The creation of stop codons in *atp9* and *cmcFc* by RNA editing has also been predicted in the mitochondria of garden asparagus (*Asparagus officinalis* L.) (Sheng et al., 2023) and observed in onion (Tsujimura et al., 2019) (Tsujimura et al., 2019).

**Detection of polycistronic transcripts in the organellar genomes of *Allium sativum***

RNA-seq data from the section above was used to detect possible polycistronic transcripts. To this end, the reads aligned by Hisat2 were subsequently assembled using Stringtie (Pertea et al., 2015). The resulting assembled transcripts and their sequencing depth were plotted using Circos (Figures 2 and 4, inner circles). Our results suggest that most of the chloroplast transcripts are polycistronic and might function as operons, as previously described for other chloroplast genomes (Barkan, 1988; Ghulam et al., 2012). Genes that encode proteins of a complex are more likely to belong to the same transcriptional unit, as illustrated by the polycistronic transcript of *rpoB*, *rpoC1* and *rpoC2*. Also, genes that pertain to the same family, subunit or system are usually transcribed together, as illustrated by the polycistronic transcript that spans the *rps11*, *rpl36*, *rps8*, *rpl14* and *rpl16* coding sequences.

**Conclusions**

Despite the importance of garlic as a cultivated plant, numerous resources needed for genetic improvement of the species are not yet available. Although the complete genome sequence, assembled from reads obtained using the PacBio and Illumina sequencing technologies, has recently been published, there are still numerous

difficulties for its effective use in breeding projects. Among the existing difficulties are the impossibility of making crosses with the main cultivated varieties and, surprisingly, the non-correspondence between the sequences deposited in the databases and the files in which the annotation of the genes is described. Deficiencies in annotation limit the performance of transcriptomic studies and the characterization of genetic variants located in coding sequences. Although several chloroplast genome sequences are available (Filyushin et al., 2016), including our own, which was one of the first to be deposited in GenBank for this species, a reference sequence of the mitochondrial genome is not yet available. The studies carried out in onion highlight the importance of this genome in the determination of traits of great agronomic interest, such as male sterility, which are routinely used in onion for hybrid seed breeding (S. Kim et al., 2009). The development of garlic varieties in which fertility has been restored (Etoh et al., 1998; Jenderek and Hannan, 2000; Kamenetsky et al., 2005; Pooler and Simon, 1994) and the identification of germplasm with structural variants of the mitochondrial genome represent the first step in the development of analogous systems in garlic. Our characterization of the mitochondrial genome points toward conservation of gene content with respect to the mitochondrial genomes of onion and leek, two cultivated species of the same genus. The number of genes is slightly lower in these three species than that described for asparagus (Kim et al., 2016; Sheng et al., 2023; Xing et al., 2023), one of the most notable differences being the presence in the latter of a higher number of genes encoding ribosomal proteins. The different articles describing the onion mitochondrial genome do not coincide in presenting a single structure of the genome, but rather present different subgenomic molecules, in which, in addition to genes of mitochondrial origin, different pseudogenes of chloroplastic origin are also present. Our study of the garlic mitochondrial genome sequence indicates that sequences of chloroplastic origin are also present in garlic, usually pseudogenes, which can also be transcribed, as can be seen in the figure we have prepared, which shows the number of reads derived from each strand in an RNA sequencing experiment. Transcriptome sequencing has also

allowed us to study the distribution of editing in the chloroplast and mitochondrial genomes. As in numerous other known mitochondrial and chloroplast genomes, transcripts derived from these genomes exhibit numerous C-to-U substitutions, which are detected by aligning transcriptome reads to genomic sequences. The C-to-U changes account for virtually 100% of the observed changes in coding sequences. Our observations are in agreement with the predictions made by other authors and with the observations made in onion. In this species, editing is of great importance, since it is thanks to them that the translation start codons of several genes are created, as well as the translation termination codons of other genes.

## Materials and methods

### DNA and RNA purification and sequencing

Total genomic DNA was extracted from 300 mg of leaf tissue using a commercial DNA extraction kit (GeneJET Plant Genomic DNA Purification Mini Kit, Thermofisher) following the manufacturer's instructions, and its quality was assessed by agarose electrophoresis. For RNA purification, we used the MagJET Plant RNA kit (Thermo Scientific) according to the manufacturer's instructions. Both DNA and RNA samples were sequences by Stab Vida (Caparica, Portugal) in an Illumina HiSeq 2500 next-generation sequencer, with a paired-end strand-specific protocol. Before analyses, low quality nucleotides from the 5' and 3' end of the raw genomic reads were trimmed using Trimmomatic (Bolger et al., 2014). We used FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the quality of the reads before and after this step. Read samples were filtered using Fastp to eliminate adapters and low-quality reads.

### *De novo* assembly

The mitochondrial and chloroplast genomes were assembled using Velvet version 1.2.13 (Zerbino and Birney, 2008). For the chloroplast genome, we used the following

settings: hash length: 51, insert length (-ins_length): 150, expected coverage (-exp_cov): 340, and coverage cutoff (-cov_cutoff): 173. The paired-end reads were mapped back to the assembled chloroplast genome sequence using Bowtie 2 (version 2.1.0), and the resulting alignment was visualized using Tablet (version 1.15.09.01) (Milne et al., 2013) to manually correct potential assembly problems. The RNA-seq reads were aligned using HiSat2 (Kim et al., 2019). For the mitochondrial assembly, we set the hash length to 65 bp, the expected coverage to 18, and the coverage cut-off to 6. These settings were chosen to discard most sequences from the low-copy fraction of the nuclear genome, while maintaining all sequences derived from the mitochondrial and plastid genomes. To this end, we performed successive rounds of assembly varying one parameter in each round, so we selected the best results for each one. The assembled contigs are expected to incorporate sequences derived from both the chloroplast and mitochondrial genomes, which are typically overrepresented in plant cells relative to the nuclear genome. The resulting assembly graph was visualized using Bandage (Wick et al., 2015). Connections between mitochondrial and chloroplast genome were separated manually using Bandage, duplicating shared nodes and creating one path for each genome. Once separated, we confirmed the topology of the chloroplast genome graph, which contains a cycle with two inverted repeats and two single-copy regions (short and long). For the SPAdes assembly, only --cov-cutoff 6 was set, same as in velvet assembly, but SPAdes does not support an expected coverage option. The hash length set for this assembly ranged from 65 to 71 (65, 67, 69, 71). The resulting graph was more complex than the Velvet graph, and the mitochondrial and chloroplast contigs were found to be linked to some nuclear genomic sequences. We used the BLAST implementation in Bandage to isolate the mitochondrial and chloroplast contigs, using the sequences of the chloroplast and mitochondrial genomes of *Allium cepa*, and the chloroplast sequence of *Allium sativum* as a reference. The “determine contiguity” command of Bandage was used to select all the nodes connected to the BLAST hits.

These nodes were extracted from the main graph, and we obtained a subgraph that contained the mitochondrial and chloroplast genomes of garlic. Paths shared by the chloroplast and mitochondrial genomes were separated as in Velvet.

**Annotation**

We used different approaches to annotate the protein-coding genes in both mitochondrial and chloroplast genomes. For the chloroplast, we first wrote an in-house script to search for open reading frames (ORFs), which were assigned functions using BLASTp searches (Altschul et al., 1997, 1990). For the mitochondrial annotation, genes were identified using BLAST searches and the application GeSeq inside the CHLOROBOX website. For both genomes, Transfer RNA (tRNA) genes were identified using ARAGORN (version 1.2.38) (Laslett and Canback, 2004) and tRNAscan-SE 2.0 (Chan and Lowe, 2019). Ribosomal RNAs were annotated using RNAmmer 1.2 server (Lagesen et al., 2007). The genome map of the chloroplast genome was drawn to scale using OGdraw (Greiner et al., 2019). Circos (Krzywinski et al., 2009) was used to plot the inner circle of the chloroplast genome map, which graphically shows the read depth and the assembled transcripts, and to generate the map of the mitochondrial genome.

**Analysis of RNA-seq data and detection of editing events**

RNA samples were sequenced using a strand-specific protocol, with paired-end reads. The reads were processed using Trimmomatic and were subsequently aligned to the chloroplast and mitochondrial genomes using Hisat2 (version 2.1.0) (Kim et al., 2019) with the following settings: --max-intronlen 1100, --no-mixed, --dta, --rna-strandness RF, and --no-discordant for the chloroplast genome, and default options for the mitochondrial genome. The resulting SAM files were converted to BAM format, sorted, and indexed using Samtools (version 1.9). Editing events were identified by comparing the aligned reads against the reference genome sequence using

commands from Samtools mpileup (options: -u, -g, and -f) and Bcftools call (options: -c and -v) (Li et al., 2009).

To plot the transcriptional activity along both genomes, we classified the reads in the BAM file based on their bitwise flag values, which allowed us to obtain separate BAM files containing the read pairs mapping to each strand. Samtools mpileup was used with these files to evaluate the sequencing depths at individual positions along the genome, which were plotted using Circos software (version 0.69-06) (Krzywinski et al., 2009).

To define the boundaries of polycistronic transcripts spanning adjacent ORFs (i.e. operons), we re-assembled the reads mapping to each strand separately with Stringtie (version 2.1.0) (Pertea et al., 2015), with the following settings: --fr forward_reads.bam, -o output_forward.gtf, -u, and -g 65. The same settings were used for reverse reads. In order to work properly, Stringtie needs that the alignment with Hisat2 has been specified with the options --dta and --rna-strnadness RF. Results were plotted using Circos software.

**PCR amplification and Sanger sequencing**

Putative recombination sites of the mitochondrial genome were identified by inspecting the assembly graph using Bandage, which was also used to select and retrieve relevant nucleotide sequences in Fasta format. Primers were designed for sequences flanking the recombination sites, as well as for putative connections between mitochondrial sequences and chloroplast-like sequences, with an amplicon length of ~1 kb. We also designed primers to amplify the regions containing the rRNA genes. All primers were designed using Primer3 (version 4.0.0; http://primer3.ut.ee/) (Koressaar and Remm, 2007; Untergasser et al., 2012) and were purchased from Sigma-Aldrich and StabVida (Supplemental Table 1).

**Acknowledgements**

We also thank Prof. José Barril for sharing laboratory space. This project received funding from COOPAMAN and Universidad Miguel Hernández.

## References

Alkatib, S., Fleischmann, T.T., Scharff, L.B., Bock, R., 2012a. Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. Nucleic Acids Res. 40, 6713–6724.

Alkatib, S., Scharff, L.B., Rogalski, M., Fleischmann, T.T., Matthes, A., Seeger, S., Schöttler, M.A., Ruf, S., Bock, R., 2012b. The contributions of wobbling and superwobbling to the reading of the genetic code. PLoS Genet. 8, e1003076.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Alverson, A.J., Rice, D.W., Dickinson, S., Barry, K., Palmer, J.D., 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell 23, 2499–2513.

Aono, N., Shimizu, T., Inoue, T., Shiraishi, H., 2002. Palindromic repetitive elements in the mitochondrial genome of Volvox. FEBS Lett. 521, 95–99.

Barkan, A., 1988. Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs. EMBO J. 7, 2637–2644.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonen, L., 2008. Cis-and trans-splicing of group II introns in plant mitochondria. Mitochondrion 8, 26–34.

Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27, 1767–1780.

Boudreau, E., Takahashi, Y., Lemieux, C., Turmel, M., Rochaix, J., 1997. The chloroplast ycf3 and ycf4 open reading frames of Chlamydomonas reinhardtii are required for the accumulation of the photosystem I complex. EMBO J.

Buso, G., Paiva, M., Torres, A., Resende, F., Ferreira, M., Buso, J., Dusi, A., 2008. Genetic diversity studies of Brazilian garlic cultivars and quality control of garlic-clover production. Genet. Mol. Res. 7, 534–541.

Chan, P.P., Lowe, T.M., 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. Gene Predict. Methods Protoc. 1–14.

Chand, S.K., Nanda, S., Rout, E., Joshi, R.K., 2015. Mining, characterization and validation of EST derived microsatellites from the transcriptome database of Allium sativum L. Bioinformation 11, 145.

Chen, J., Guan, R., Chang, S., Du, T., Zhang, H., Xing, H., 2011. Substoichiometrically different mitotypes coexist in mitochondrial genomes of Brassica napus L. PLoS One 6, e17662.

Cunha, C.P., Hoogerheide, E.S., Zucchi, M.I., Monteiro, M., Pinheiro, J.B., 2012. New microsatellite markers for garlic, Allium sativum (Alliaceae). Am. J. Bot. 99, e17–e19.

Etoh, T., Noma, Y., Nishitarumizu, Y., Wakamoto, T., 1998. Seed productivity and germinability of various garlic clones collected in Soviet Central Asia.

Filyushin, M.A., Beletsky, A.V., Mazur, A.M., Kochieva, E.Z., 2016. The complete plastid genome sequence of garlic Allium sativum L. Mitochondrial DNA Part B 1, 831–832. https://doi.org/10.1080/23802359.2016.1247669

García-Lampasona, S., Asprelli, P., Burba, J.L., 2012. Genetic analysis of a garlic (Allium sativum L.) germplasm collection from Argentina. Sci. Hortic. 138, 183–189.

Ghulam, M.M., Zghidi-Abouzid, O., Lambert, E., Lerbs-Mache, S., Merendino, L., 2012. Transcriptional organization of the large and the small ATP synthase operons, atpI/H/F/A and atpB/E, in Arabidopsis thaliana chloroplasts. Plant Mol Biol 79, 259–272.

Glanz, S., Kück, U., 2009. Trans-splicing of organelle introns–a detour to continuous RNAs. Bioessays 31, 921–934.

Greiner, S., Lehwark, P., Bock, R., 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 47, W59–W64.

Gualberto, J.M., Mileshina, D., Wallet, C., Niazi, A.K., Weber-Lotfi, F., Dietrich, A., 2014. The plant mitochondrial genome: dynamics and maintenance. Biochimie 100, 107–120.

Ipek, M., Ipek, A., Almquist, S.G., Simon, P.W., 2005. Demonstration of linkage and development of the first low-density genetic map of garlic, based on AFLP markers. Theor. Appl. Genet. 110, 228–236. https://doi.org/10.1007/s00122-004-1815-5

Ipek, M., Ipek, A., Simon, P.W., 2003. Comparison of AFLPs, RAPD markers, and isozymes for diversity assessment of garlic and detection of putative duplicates in germplasm collections. J. Am. Soc. Hortic. Sci. 128, 246–252.

Ipek, M., Sahin, N., Ipek, A., Cansev, A., Simon, P.W., 2015. Development and validation of new SSR markers from expressed regions in the garlic genome. Sci. Agric. 72, 41–46.

Ipek, M., Simon, P., 1998. Genetic diversity in garlic (Allium sativum L.) as assessed by amplified fragment length polymorphism (AFLP). Presented at the Proceedings of the 1998 National Onion (and other Allium) Research Conference, Sacramento, California, USA, pp. 110–120.

Jenderek, M., Hannan, R.M., 2000. Seed producing ability of garlic (Allium sativum L.) clones from two public US collections. Proc. Third Int. Symp. Edible Alliaceae Athens Ga. USA 73–75.

Jo, M.H., Ham, I.K., Moe, K.T., Kwon, S.-W., Lu, F.-H., Park, Y.-J., Kim, W.S., Kim, M.K., Kim, T.I., Lee, E.M., 2012. Classification of genetic variation in garlic ('Allium sativum'L.) using SSR markers. Aust. J. Crop Sci. 6, 625–631.

Kamenetsky, R., Faigenboim, A., Shemesh Mayer, E., Ben Michael, T., Gershberg, C., Kimhi, S., Esquira, I., Rohkin Shalom, S., Eshel, D., Rabinowitch, H.D., 2015. Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (Allium sativum L.). Bmc Genomics 16, 1–16.

Kamenetsky, R., London Shafir, I., Khassanov, F., Kik, C., van Heusden, A.W., Vrielink-van Ginkel, M., Burger-Meijer, K., Auger, J., Arnault, I., Rabinowitch, H.D., 2005. Diversity in fertility potential and organo-sulphur compounds among garlics from Central Asia. Biodivers. Conserv. 14, 281–295. https://doi.org/10.1007/s10531-004-5050-9

Kim, B., Kim, K., Yang, T.-J., Kim, S., 2016. Completion of the mitochondrial genome sequence of onion (Allium cepa L.) containing the CMS-S male-sterile cytoplasm and identification of an independent event of the ccmF N gene split. Curr. Genet. 62, 873–885.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim, D.-W., Jung, T.-S., Nam, S.-H., Kwon, H.-R., Kim, A., Chae, S.-H., Choi, S.-H., Kim, Dong-Wook, Kim, R.N., Park, H.-S., 2009. GarlicESTdb: an online database and mining tool for garlic EST sequences. BMC Plant Biol. 9, 1–6.

Kim, S., Lee, E.-T., Cho, D.Y., Han, T., Bang, H., Patil, B.S., Ahn, Y.K., Yoon, M.-K., 2009. Identification of a novel chimeric gene, orf725, and its use in development of a molecular marker for distinguishing among three cytoplasm types in onion (Allium cepa L.). Theor. Appl. Genet. 118, 433–441.

Kim, S., Yoon, M.-K., 2010. Comparison of mitochondrial and chloroplast genome segments from three onion (Allium cepa L.) cytoplasm types and identification of a trans-splicing intron of cox2. Curr. Genet. 56, 177–188.

Knoop, V., 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. Curr. Genet. 46, 123–139.

Koressaar, T., Remm, M., 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23, 1289–1291.

Kozik, A., Rowan, B.A., Lavelle, D., Berke, L., Schranz, M.E., Michelmore, R.W., Christensen, A.C., 2019. The alternative reality of plant mitochondrial DNA: One ring does not rule them all. PLoS Genet. 15, e1008373.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Laslett, D., Canback, B., 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32, 11–16.

Lavrov, D.V., Pett, W., 2016. Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. Genome Biol. Evol. 8, 2896–2913.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Milne, I., Stephen, G., Bayer, M., Cock, P.J., Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D., 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14, 193–202.

Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., Kadowaki, K., 2002. The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Genomics 268, 434–445.

Oldenburg, D.J., Bendich, A.J., 2001. Mitochondrial DNA from the liverwort Marchantia polymorpha: circularly permuted linear molecules, head-to-tail concatemers, and a 5′ protein. J. Mol. Biol. 310, 549–562.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295.

Pooler, M.R., Simon, P.W., 1994. True seed production in garlic. Sex. Plant Reprod. 7, 282–286. https://doi.org/10.1007/BF00227710

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using SPAdes de novo assembler. Curr. Protoc. Bioinforma. 70, e102.

Schwartz, J.H., Meyer, R., Eisenstadt, J.M., Brawerman, G., 1967. Involvement of N-formylmethionine in initiation of protein synthesis in cell-free extracts of Euglena gracilis. J. Mol. Biol. 25, 571-IN27.

Shemesh-Mayer, E., Ben-Michael, T., Rotem, N., Rabinowitch, H.D., Doron-Faigenboim, A., Kosmala, A., Perlikowski, D., Sherman, A., Kamenetsky, R., 2015. Garlic (Allium sativum L.) fertility: transcriptome and proteome analyses provide insight into flower and pollen development. Front. Plant Sci. 6.

Sheng, W., Deng, J., Wang, C., Kuang, Q., 2023. The garden asparagus (Asparagus officinalis L.) mitochondrial genome revealed rich sequence variation throughout whole sequencing data. Front. Plant Sci. 14, 1140043.

Skippington, E., Barkman, T.J., Rice, D.W., Palmer, J.D., 2015. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc. Natl. Acad. Sci. 112, E3515–E3524.

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., Wu, M., McCauley, D.E., Palmer, J.D., Taylor, D.R., 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10, e1001241.

Smith, A.E., Marcker, K.A., 1968. N-formylmethionyl transfer RNA in mitochondria from yeast and rat liver. J. Mol. Biol. 38, 241–243.

Sun, X., Zhou, S., Meng, F., Liu, S., 2012. De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing. Plant Cell Rep. 31, 1823–1828. https://doi.org/10.1007/s00299-012-1295-z

Sun, X., Zhu, S., Li, N., Cheng, Y., Zhao, Jing, Qiao, X., Lu, L., Liu, S., Wang, Y., Liu, C., Li, B., Guo, W., Gao, S., Yang, Z., Li, F., Zeng, Z., Tang, Q., Pan, Y., Guan, M., Zhao, Jian, Lu, X., Meng, H., Han, Z., Gao, C., Jiang, W., Zhao, X., Tian, S., Su, J., Cheng, Z., Liu, T., 2020. A Chromosome-Level Genome Assembly of Garlic (Allium sativum) Provides Insights into Genome Evolution and Allicin Biosynthesis. Mol. Plant 13, 1328–1339. https://doi.org/10.1016/j.molp.2020.07.019

Sun, X.D., Ma, G.Q., Cheng, B., Li, H., Liu, S.Q., 2013. Identification of differentially expressed genes in shoot apex of garlic (Allium sativum L.) using Illumina sequencing. J. Plant Stud. 2, 136.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., Greiner, S., 2017. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45, W6–W11. https://doi.org/10.1093/nar/gkx391

Tsujimura, M., Kaneko, T., Sakamoto, T., Kimura, S., Shigyo, M., Yamagishi, H., Terachi, T., 2019. Multichromosomal structure of the onion mitochondrial genome and a transcript analysis. Mitochondrion 46, 179–186.

Unseld, M., Marienfeld, J.R., Brandt, P., Brennicke, A., 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat. Genet. 15, 57–61.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res. 40, e115–e115.

Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E., 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31, 3350–3352.

Wissinger, B., Schuster, W., Brennicke, A., 1991. Trans splicing in Oenothera mitochondria: nad1 mRNAs are edited in exon and trans-splicing group II intron sequences. Cell 65, 473–482.

Woloszynska, M., 2010. Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes—though this be madness, yet there's method in't. J. Exp. Bot. 61, 657–671.

Xing, J., Zhu, M., Wang, Y., Liu, H., 2023. The complete mitochondrial genome of Allium fistulosum L. (Amaryllidaceae). Mitochondrial DNA Part B 8, 890–894. https://doi.org/10.1080/23802359.2023.2248684

Yamaguchi, K., Subramanian, A.R., 2000. The plastid ribosomal proteins: identification of all the proteins in the 50 S subunit of an organelle ribosome (chloroplast). J. Biol. Chem. 275, 28466–28482.

Yamaguchi, K., von Knoblauch, K., Subramanian, A.R., 2000. The Plastid Ribosomal Proteins: IDENTIFICATION OF ALL THE PROTEINS IN THE 30 S SUBUNIT OF AN ORGANELLE RIBOSOME (CHLOROPLAST) *. J. Biol. Chem. 275, 28455–28465. https://doi.org/10.1074/jbc.M004350200

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

Zewdie, Y., Havey, M.J., Prince, J.P., Jenderek, M.M., 2005. The First Genetic Linkages among Expressed Regions of the Garlic Genome. J. Am. Soc. Hortic. Sci. Jashs 130, 569–574. https://doi.org/10.21273/JASHS.130.4.569

**Figure legends**

**Figure 1.** Resulting graph of the chloroplast genome assembly. Blue node corresponds to the inverted repeats of the genome, which are connected to the Large Single Copy (Green) and to the Short Single Copy regions (Orange).

**Figure 2.** The chloroplast genome of *Allium sativum*. The map was made with Ogdraw, and the inner circle was made with Circos. The genes located within the circle are transcribed clockwise, while those represented on the outside are transcribed counterclockwise. The first inner circle represents the alignment coverage of the RNA-seq reads, the second inner circle is the assembly of the polycistronic transcripts made with Stringtie, and the third inner circle marks the boundaries between the different regions: inverted repeats (IRA and IRB), and single regions (SSC and LSC).

**Figure 3.** Assembly graph of the mitochondrial genome. **(A)** Mitochondrial (blue) and chloroplast (green) merged graph isolated the de novo assembly with Velvet. The grey region in the top right of the figure corresponds to repeated ribosome sequences from the nucleus. **(B)** Isolated graph of the mitochondrial (blue) and chloroplast (green) genomes, as well as the repeated nuclear sequences (red). Regions shared with the chloroplast genome present a much higher coverage, and they need to be normalized and merged. **(C)** Graph of the mitochondrial genome after chloroplast-derived sequences have been merged and the ribosomal region has been sequenced. Red nodes represent the short repeated sequences, while the green nodes are the long repeated sequences. **(D)** Final graph of the mitochondrial genome. Paths through the graph have been confirmed by aligning paired-end reads to the nodes of the graph.

**Figure 4.** Circle representation of the 21 contigs conforming the mitochondrial genome of *Allium sativum*. The map was made with Circos. The genes located within the circle are transcribed clockwise, while those represented on the outside are transcribed

counterclockwise. The first inner circle represents the alignment coverage of the RNA-seq reads, the second inner circle is the assembly of the transcripts made with Stringtie.

**Supplementary materials**

**Supplemental Figure 1.** Predicted secondary structure of tRNA molecules containing the CAU anticodon.

**Supplemental Figure 2.** Secondary structure of four tRNAs for the amino acids methionine, formyl methionine and isoleucine encoded un the chloroplast genome of Allium sativum. A) tnrI-GAU. B) trnI-CAI. C) trn(f)M-CAU. D) trnM-CAU. Figure adapted from the one un Alkatib et al., (2012a). The key differences have been highlighted in red. The anticodons are shown in bold. A “+” sign marks noncanonical base pairs between G and U.

Figure 1

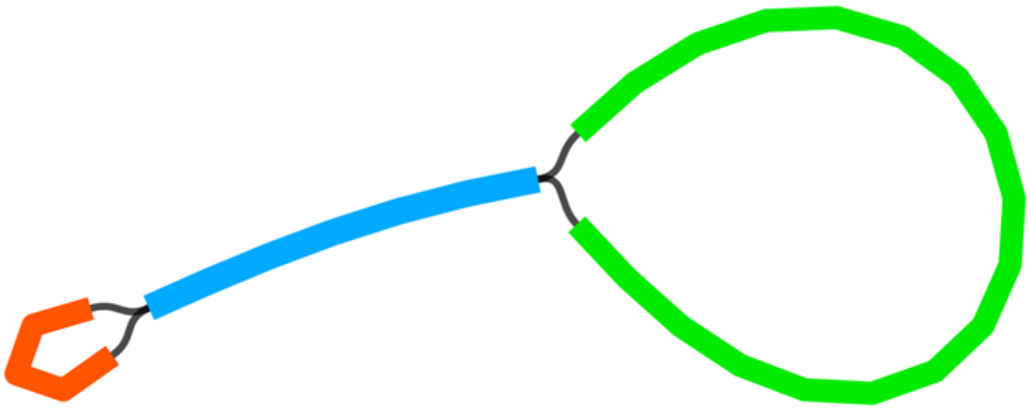


**Figure 1.** Resulting graph of the chloroplast genome assembly visualized in Bandage. Blue node corresponds to the inverted repeats of the genome, which are connected to the Large Single Copy (Green) and to the Short Single Copy regions (Orange).

Figure 2

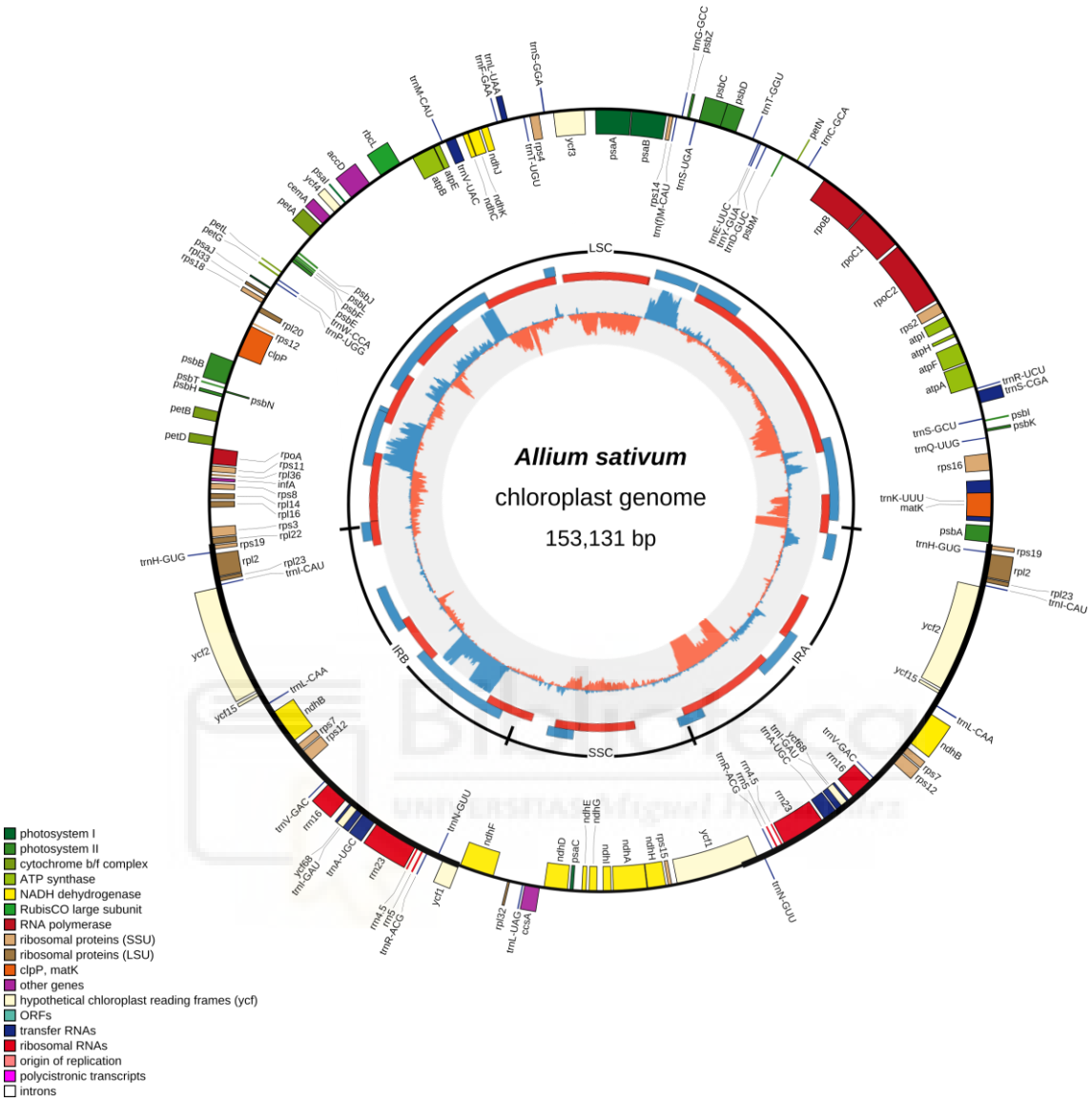



**Figure 2.** The chloroplast genome of *Allium sativum*. The map was made with Ogdraw. The genes located within the circle are transcribed clockwise, while those represented on the outside are transcribed counterclockwise. The first inner circle represents the alignment coverage of the RNA-seq reads, the second inner circle is the assembly of the polycistronic transcripts made with Stringtie, and the third inner circle marks the boundaries between the different regions: inverted repeats (IRA and IRB), and single regions (SSC and LSC).

Figure 3

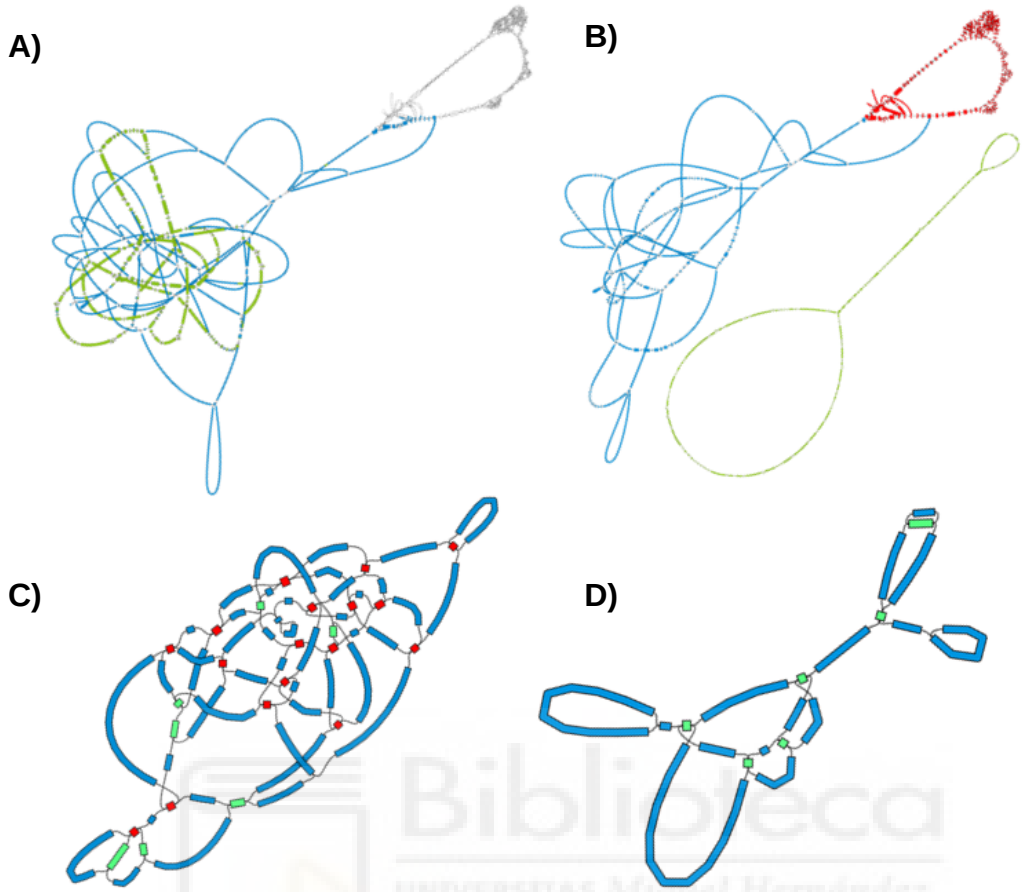



**Figure 3.** Assembly graph of the mitochondrial genome visualized in Bandage. **(A)** Mitochondrial (blue) and chloroplast (green) merged graph isolated the de novo assembly with Velvet. The grey region in the top right of the figure corresponds to repeated ribosome sequences from the nucleus. **(B)** Isolated graph of the mitochondrial (blue) and chloroplast (green) genomes, as well as the repeated nuclear sequences (red). Regions shared with the chloroplast genome present a much higher coverage, and they need to be normalized and merged. **(C)** Graph of the mitochondrial genome after chloroplast-derived sequences have been merged and the ribosomal region has been sequenced. Red nodes represent the short repeated sequences, while the green nodes are the long repeated sequences. **(D)** Final graph of the mitochondrial genome. Paths through the graph have been confirmed by aligning paired-end reads to the nodes of the graph.

Figure 4

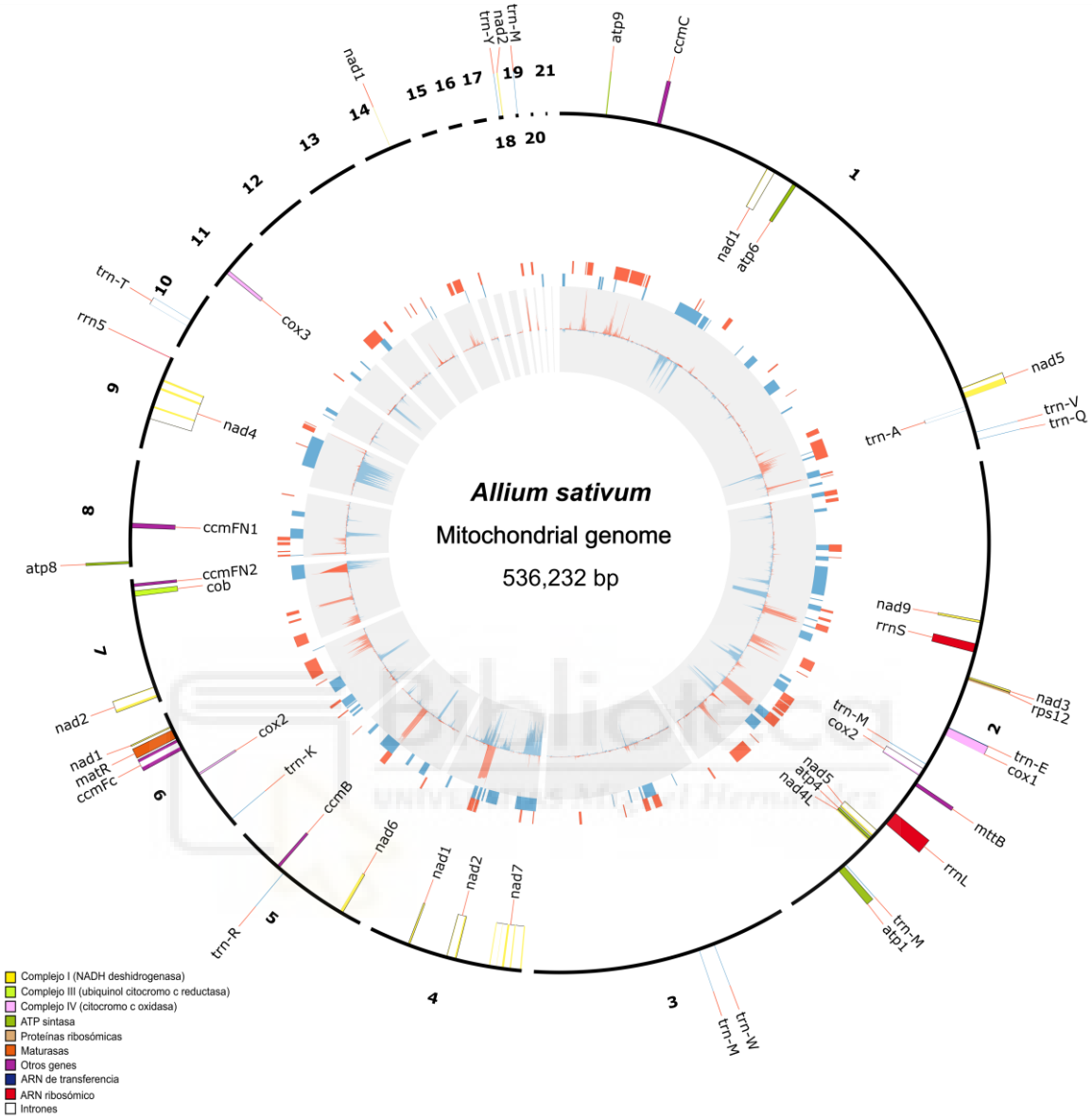



**Figure 4.** Circle representation of the 21 contigs conforming the mitochondrial genome of *Allium sativum*. The map was made with Circos. The genes located within the circle are transcribed clockwise, while those represented on the outside are transcribed counterclockwise. The first inner circle represents the alignment coverage of the RNA-seq reads, the second inner circle is the assembly of the transcripts made with Stringtie.

**Table 1.** Functional classification of the genes in the mitochondrial genome

| | | |
|---|---|---|
| Electron transport chain complexes | Complex I NADH dehydrogenase | *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9* |
| | Complex III cytochrome bc1 | *cob* |
| | Complex IV cytochrome C oxidase | *cox1*, *cox2*, *cox3* |
| | Complex V | *atp1*, *atp4*, *atp6*, *atp8*, *atp9* |
| Control of gene expression | Ribosomal proteins | *rps12* |
| | rRNA genes | *rrn5*, *rrnL*, *rrnS* |
| | tRNA genes | *trnT*-GGU, *trnE*-UUC, *trnR*-ACG, *trnM*-CAU (x4), *trnW*-CCA, *trnA*-UGC, *trnV*-GAC, *trnQ*-UUG, *trnK*-UUU, *trnY*-GUA |
| Other genes | Cytochrome c biogenesis | *ccmB*, *ccmC*, *ccmFc*, *ccmFn1*, *ccmFn2* |
| | Maturase | *matR* |
| | Sec-Y independent transporter | *mttB* |

**Table 2.** Functional classification of the genes in the chloroplast genome

| | | |
|---|---|---|
| Gene expression | Ribosomal RNA genes | *rrn4.5* (×2), *rrn5* (×2), *rrn16* (×2), *rrn23* (×2) |
| | Transfer RNA genes | *trnA-UGC** (×2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC**, *trnF-GAA*, *trn(f)M-CAU, trnG-GCC, trnH-GUG* (×2), *trnI-CAU* (×2), *trnI-GAU** (×2), *trnK-UUU**, *trnL-CAA* (×2), *trnL-UAG, trnL-UAA**, *trnM-CAU*, *trnN-GUU* (×2), *trnP-UGG*, *trnQ-UUG*, *trnR-ACG* (×2), *trnR-UCU*, *trnS-GCU, trnS-CGA**, *trnS-UGA*, *trnS-GGA*, *trnT-GGU, trnT-UGU*, *trnV-GAC* (×2), *trnV-UAC**, *trnW-CCA*, *trnY-GUA* |
| | Small subunit of ribosome | *rps3*, *rps4*, *rps7* (×2), *rps8*, *rps11*, *rps12* (transplicing) (×2), *rps14*, *rps15*, *rps18*, *rps19* (×2) |
| | Large subunit of ribosome | *rpl2** (×2), *rpl14, rpl16*, *rpl20*, *rpl22, rpl23* (×2), *rpl32, rpl33*, *rpl36* |
| | PEP subunits | *rpoA*, *rpoB*, *rpoC1**, *rpoC2* |
| Photosynthesis | RuBisCO large subunit | *rbcL* |
| | Photosystem I | *psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *ycf3*, *ycf4* |
| | Photosystem II | *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ* |
| | NADH dehydrogenase | *ndhA**, *ndhB** (×2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG, ndhH*, *ndhI*, *ndhJ*, *ndhK* |
| | Cytochrome b/f complex | *petA*, *petB*, *petD*, *petG*, *petL*, *petN* |
| | ATP synthase | *atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI* |
| Other genes | Maturase | *matK* |
| | Protease | *clpP** |
| | Chloroplast envelope membrane protein | *cemA* |
| | Acetyl-CoA-carboxylase subunit | *accD* |
| | Cytochrome C biogenesis protein | *ccsA* |
| | Conserved ORFs | y*cf1*, *ycf2* (×2), *ycf68** (×2) |
| Pseudogenes | | *rps2*, *ycf15* (×2), *rps16**, *infA*, *ycf1* |

**Supplemental Table 1.-** Primers used in this work

| | Primer | Sequence (5'→3') | Tm (°C) |
|---|---|---|---|
| Mito-Chloro 1* | Forward | TGAAAGGACGTTGCACTTCG | 57.73 |
| | Reverse 1 | CCGAAGTGAGGTTAGCGAAG | 58.37 |
| | Reverse 2 | CCGTTCTCTGAGCTATAGGGG | 59.11 |
| Mito-Chloro 2* | Forward | TCGTTCCTCATTGAGACGGG | 59.47 |
| | Reverse | GCGCCCCTAGAAAGAAAGTG | 58.91 |
| Mito-Chloro 3* | Forward 1 | GTTTGTCATGGTATCGCGCT | 58.99 |
| | Forward 2 | TTTGCCTTATGTAACACGATGGG | 59.31 |
| | Reverse 1 | GAGTCACTTCTCCTGGCCAG | 59.75 |
| | Reverse 2 | TCGTCATGATATCTGCCAATTTC | 57.07 |
| Mito-Chloro 4* | Forward | TCGTACATCCCCCAATTGAT | 57.97 |
| | Reverse | GGCCCGCCTAAGATTTGATG | 59.05 |
| Rep 1 | Forward 1 | TGAAAGGACGTTGCACTTCG | 60.01 |
| | Forward 2 | AAGTTAATTCGCGGGGTTTT | 59.84 |
| | Reverse 1 | GAAAGAGAAAGCGGCACATC | 59.96 |
| | Reverse 2 | AACGATTACGCGAGTTGCTT | 59.91 |
| Rep 2 | Forward 1 | CGAACACTCTCGAACCACAA | 59.87 |
| | Forward 2 | CGCTACGCACCTTAGGAAAG | 60.03 |
| | Reverse 1 | AAATCCGGTGTGTTGTCTCC | 59.83 |
| | Reverse 2 | CAAGCCCCATTTCTGTCAAT | 59,93 |
| Rep 3 | Forward 1 | ACAGCTGGAGAGAGTACACA | 57.71 |
| | Forward 2 | GTGAATCATGAAGCGCGACA | 59.28 |
| | Reverse 1 | GTGATAGTCAGAACACGCGG | 58.73 |
| | Reverse 2 | GTAAAACCTGCGGCATGTCC | 59.83 |
| | Forward 3 | TGATTGAGCCCGGTTCTCTT | 58.92 |
| | Reverse 3 | TATGGTGGCGCGATGTTCTA | 59.03 |
| Rep 4 | Forward 1 | GTGTTTGGAACTGGCTCGTT | 58.98 |
| | Forward 2 | TAACGCAGGGCAAACGAAAA | 58.98 |
| | Reverse 1 | GTGCCAGTCTAAGTTCCCCT | 59.02 |
| | Reverse 2 | GAGGATGGGACTTGGGATCT | 58.19 |
| Rep 5 | Forward 1 | GTCGGGCTAATTCCATACGA | 59.92 |
| | Forward 2 | GCCTCCCTTGTTAGTTGTCG | 59.73 |

| | | | |
|---|---|---|---|
| | Reverse 1 | AGCTGCTGGGATCAGAGAAA | 60.1 |
| | Reverse 2 | CCTCTCGTTTGGTAGCGAAG | 60.01 |
| Rep 6 | Forward 1 | CGCCCTACCTAAACCAATCA | 59.95 |
| | Forward 2 | CGTGCAGAATCAGAGTTCCA | 59.98 |
| | Reverse 1 | GTTCTTGCGATCGTGCCTAT | 60.24 |
| | Reverse 2 | TCTGCCCTAAGATCGCAACT | 59.98 |
| Rep 7 | Forward 1 | TATCGGTAAGTCGGAGCAGC | 59.33 |
| | Forward 2 | CTTGGAACGGGAAGGCTTTC | 59.12 |
| | Reverse 1 | GGACTTAACCAGCATGTCGTC | 59 |
| | Reverse 2 | CGGCGACAACTAACATCCAG | 59 |
| Rep 8 | Forward 1 | TGAAAGGACGTTGCACTTCG | 59.07 |
| | Forward 2 | CCCTGGGCTTGGAATACTCT | 58.94 |
| | Reverse 1 | AGAAGGCAGGTCAGTCATGG | 59.38 |
| | Reverse 2 | CCCGAATGCCCCTACCTAAT | 58.78 |
| | Forward 3 | ACCTGCCTTCTGTGGTGAAC | 58 |
| | Reverse 3 | AATTGGTGCTTGCGATTTCT | 56 |
| Rep 9 | Forward 1 | GAATTGATTGAATGCGGGCA | 56 |
| | Forward 2 | GTGAAAGACGGTCAACAAGC | 58 |
| | Reverse 1 | TCAGGTAAATGCGCATTCCT | 56 |
| | Reverse 2 | TGTCGGAGTGAAAGAGCGAA | 58 |
| Rep 10 | Forward 1 | CTTGCCGGTGGAAGAAATTA | 60.07 |
| | Forward 2 | TTCTCTCCCGAACAATTGGA | 60.57 |
| | Reverse 1 | TCGGTTATCAATTCGGGGTA | 60.15 |
| | Reverse 2 | GATACCTTCGAGCACCCTGA | 60.22 |
| Rep 11 | Forward 1 | GGCAAGTTCAGTTGTCGAGG | 60 |
| | Forward 2 | GTCGAGAAGGGAGGTGTGAA | 60 |
| | Reverse 1 | GTACCTCTCTAGCATCCCCT | 60 |
| | Reverse 2 | CTGGCAGCTATGAGTCTAGC | 58 |
| Rep 12 | Forward 1 | TGCACAACTCCTCTGGATGT | 58.94 |
| | Forward 2 | CTAACAGAAGGGGCAGAGCT | 59.09 |
| | Reverse 1 | CGTTGTACTGCGCTCTTCAA | 58.86 |
| | Reverse 2 | TGCGTTGTCACCTTGTTGTT | 58.83 |
| Rep 13 | Forward 1 | ACTGTCACTAGCATTCCCGT | 58.73 |

| | | | |
|---|---|---|---|
| | Forward 2 | GGATTAACGCCTGAACAGCC | 59.27 |
| | Reverse 1 | GCCCAGAAACTTCGCTCTTC | 59.2 |
| | Reverse 2 | CACCGGGAAGCTCTATGTCT | 58.88 |
| Rep 14 | Forward 1 | CCCGGTTGGCACTATAGGAT | 58.94 |
| | Forward 2 | TTGCTTCAAAACTACGGGGC | 59.04 |
| | Reverse 1 | TGGGCTACGATTGGATTGGA | 58.79 |
| | Reverse 2 | AGACCAAAGCAAGCAACTGG | 58.96 |
| | Reverse 3 | ATCCTGTGCGGGTGTTAAGA | 59.02 |
| Rep 15 | Forward 1 | AACTCTTTGCTGCTCCTCCT | 58.94 |
| | Forward 2 | CCAGTAGATAGCCCCTCGTG | 59.04 |
| | Reverse 1 | TGTCCCTATCCCTCTGTCCA | 58.98 |
| | Reverse 2 | AGAAAGACTTGCCTCACGGA | 58.95 |
| Rep 16 | Forward 1 | TCAACAGCCCCTCGTTAAGT | 58.95 |
| | Forward 2 | AAAGGGAGGAGGAAAGAGGC | 59 |
| | Reverse 1 | TTGCGGTTGGAGTTGGATTG | 59.04 |
| | Reverse 2 | CCCGGTAAAACGTTTGCAGA | 59.05 |
| Rep 17 | Forward 1 | CACACGGCACAAGACTGAAA | 58.99 |
| | Forward 2 | TGCTCTGTTTTCGCTCCCTA | 59.03 |
| | Reverse 1 | AGAGCGGAAAGATTGACGGA | 59.1 |
| | Reverse 2 | TGGCCAGATCTTCCAACCTT | 58.92 |
| rRNA 1 | Forward | GCGCTTATTCGTTGCTTGGT | |
| | Reverse | TCCGGGTTCATTCGGGTAGT | |
| rRNA 2 | Forward 1 | CAGACTCCTTGGTCCGTGTT | |
| | Forward 2 | CATTCGCCTCGCATATACCT | |
| | Reverse | GCGCGCTACACTGATGTATTC | |

*: Mito-chloro primers were used to confirm the boundaries between the mitochondrial and chloroplast-like sequences.

**Supplemental Table 2.** Anticodons in the tRNAs encoded in the chloroplast genome

| | | Second position | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | U | | | C | | | A | | | G | | |
| First position | U | UUU | Phe | | UCU | Ser | | UAU | Tyr | | UGU | Cys | |
| | | UUC | Phe | *trnF-GAA* | UCC | Ser | *trnS-GGA* | UAC | Tyr | *trnY-GUA* | UGC | Cys | *trnC-GCA* |
| | | UUA | Leu | *trnL-UAA* | UCA | *Ser* | *trnS-UGA* | UAA | - | | UGA | - | |
| | | UUG | Leu | *trnL-CAA* ×2 | UCG | Ser | *trnS-CGA* | UAG | - | | UGG | Trp | *trnW-CCA* |
| | C | CUU | Leu | | CCU | Pro | | CAU | His | | CGU | Arg | *trnR-ACG* ×2 |
| | | CUC | Leu | | CCC | Pro | | CAC | His | *trnH-GUG* ×2 | CGC | Arg | |
| | | CUA | Leu | *trnL-UAG* | CCA | Pro | *trnP-UGG* | CAA | Gln | *trnQ-UUG* | CGA | Arg | |
| | | CUG | Leu | | CCG | Pro | | CAG | Gln | | CGG | Arg | |
| | A | AUU | Ile | | ACU | Thr | | AAU | Asn | | AGU | Ser | |
| | | AUC | Ile | *trnI-GAU* ×2 | ACC | Thr | *trnT-GGU* | AAC | Asn | *trnN-GUU* ×2 | AGC | Ser | *trnS-GCU* |
| | | AUA | Ile | *trnI-CAU* ×2 | ACA | Thr | *trnT-UGU* | AAA | Lys | *trnK-UUU* | AGA | Arg | *trnR-UCU* |
| | | AUG | Met | *trnM-CAU*, *trn(f)M-CAU* | ACG | Thr | | AAG | Lys | | AGG | Arg | |
| | G | GUU | Val | | GCU | Ala | | GAU | Asp | | GGU | Gly | |
| | | GUC | Val | *trnV-GAC* ×2 | GCC | Ala | | GAC | Asp | *trnD-GUC* | GGC | Gly | *trnG-GCC* |
| | | GUA | Val | *trnV-UAC* | GCA | Ala | *trnA-UGC* ×2 | GAA | Glu | *trnE-UUC* | GGA | Gly | *trnG-UCC* |
| | | GUG | Val | | GCG | Ala | | GAG | Glu | | GGG | Gly | |

**Supplemental Table 3.** Editing events in the chloroplast transcriptome

| Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|
| 11490 | G-to-A | *ndhB* | - | CAT | H | TAT | Y |
| 11552 | G-to-A | *ndhB* | - | TCA | S | TTA | L |
| 11643 | G-to-A | *ndhB* | - | CGC | R | TGC | C |
| 11909 | G-to-A | *ndhB* | - | TCA | S | TTA | L |
| 12710 | G-to-A | *ndhB* | - | CCA | P | CTA | L |
| 12836 | G-to-A | *ndhB* | - | TCA | S | TTA | L |
| 12861 | G-to-A | *ndhB* | - | CAT | H | TAT | Y |
| 12905 | G-to-A | *ndhB* | - | ACG | T | ATG | M |
| 12932 | C-to-A | *ndhB* | - | AGA | R | ATA | T |
| 13298 | G-to-A | *ndhB* | - | TCA | S | TTA | L |
| 14663 | G-to-A | - | | | | | |
| 15918 | C-to-G | - | | | | | |
| 18345 | G-to-C | - | | | | | |
| 31570 | C-to-U | *ccsA* | + | TCA | S | TTA | L |
| 32335 | G-to-A | *ndhD* | - | CCA | P | CTA | L |
| 32503 | G-to-A | *ndhD* | - | ACA | T | ATA | T |
| 32572 | G-to-A | *ndhD* | - | TCA | S | TTA | L |
| 32776 | G-to-A | *ndhD* | - | TCA | S | TTA | L |
| 33391 | G-to-A | *ndhD* | - | TCA | S | TTA | L |
| 33448 | G-to-A | *ndhD* | - | ACG | T | ATG | M |
| 34602 | G-to-A | *ndhE* | - | CAT | H | TAT | Y |
| 35135 | G-to-A | *ndhG* | - | TCA | S | TTA | L |
| 37953 | G-to-A | *ndhA* | - | TCA | S | TTA | L |
| 39073 | G-to-A | *ndhH* | - | TCT | S | TTT | F |
| 41130 | G-to-A | *ycf1* | - | ACA | T | ATA | T |
| 52627 | C-to-G | - | | | | | |
| 56364 | C-to-U | - | | | | | |
| 57729 | C-to-U | *ndhB* | + | TCA | S | TTA | L |
| 58047 | C-to-U | *ndhB* | + | CCA | P | CTA | L |
| 58122 | C-to-U | *ndhB* | + | ACG | T | ATG | M |
| 58166 | C-to-U | *ndhB* | + | CAT | H | TAT | Y |
| 58191 | C-to-U | *ndhB* | + | TCA | S | TTA | L |

| Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|
| 58317 | C-to-U | *ndhB* | + | CCA | P | CTA | L |
| 59118 | C-to-U | *ndhB* | + | TCA | S | TTA | L |
| 59384 | C-to-U | *ndhB* | + | CGC | R | TGC | C |
| 59475 | C-to-U | *ndhB* | + | TCA | S | TTA | L |
| 59763 | C-to-U | *ndhB* | + | CCA | P | CTA | L |
| 68980 | C-to-U | - | | | | | |
| 69764 | A-to-G | - | | | | | |
| 76287 | C-to-U | - | | | | | |
| 78951 | C-to-U | *trnS-GCU* | | | | | |
| 81596 | G-to-A | *atpA* | - | TCA | S | TTA | L |
| 84989 | G-to-A | - | | | | | |
| 85143 | G-to-A | *atpI* | - | TCA | S | TTA | L |
| 87325 | G-to-A | *rpoC2* | - | TCG | S | TTG | L |
| 88216 | G-to-A | *rpoC2* | - | TCT | S | TTT | F |
| 93827 | G-to-A | *rpoC1* | - | TCT | S | TTT | F |
| 96625 | G-to-A | *rpoB* | - | TCG | S | TTG | L |
| 96775 | G-to-A | *rpoB* | - | TCA | S | TTA | L |
| 114134 | G-to-A | *ycf3* | - | TCA | S | TTA | L |
| 114140 | G-to-A | *ycf3* | - | ACG | T | ATG | M |
| 114363 | G-to-A | - | | | | | |
| 115010 | G-to-A | - | - | TCC | S | TTC | F |
| 115457 | G-to-A | - | | | | | |
| 119557 | G-to-A | *ndhJ* | - | TCA | S | TTA | L |
| 120584 | G-to-A | *ndhC* | - | CCA | P | CTA | L |
| 128796 | C-to-U | *accD* | + | CCA | P | CTA | L |
| 135193 | C-to-U | *petL* | + | CCT | P | CTT | L |
| 135232 | C-to-U | *petL* | + | TCA | S | TTA | L |
| 135243 | C-to-U | *petL* | + | CCA | P | TCA | S |
| 138246 | G-to-A | - | | | | | |
| 139140 | G-to-A | *clpP* | - | TCA | S | TTA | L |
| 145230 | C-to-U | *petB* | + | CGG | R | TGG | W |
| 145423 | C-to-U | *petB* | + | CCA | P | CTA | L |
| 146903 | C-to-U | *petD* | + | CAA | Q | TAA | * |

| Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|
| 147578 | G-to-A | *rpoA* | - | TCC | S | TTC | F |
| 147737 | G-to-A | *rpoA* | - | TCA | S | TTA | L |
| 147905 | G-to-A | *rpoA* | - | TCT | S | TTT | F |
| 148913 | G-to-A | - | | | | | |
| 149496 | G-to-A | *rps8* | - | TCA | S | TTA | L |
| 151722 | G-to-A | - | | | | | |
| 152087 | G-to-A | *rps3* | - | CAT | H | TAT | Y |

**Supplemental Table 4.** Editing events in the mitochondrial transcriptome

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_1 | 10166 | C-to-U | *atp9* | + | TCA | S | TTA | L |
| NODE_1 | 10250 | C-to-U | *atp9* | + | TCA | S | TTA | L |
| NODE_1 | 10307 | C-to-U | *atp9* | + | CCA | P | CTA | L |
| NODE_1 | 10321 | C-to-U | *atp9* | + | CTG | L | TTG | L |
| NODE_1 | 10328 | C-to-U | *atp9* | + | TCA | S | TTA | L |
| NODE_1 | 10339 | C-to-U | *atp9* | + | CAG | Q | TAG | * |
| NODE_1 | 21543 | G-to-A | *ccmC* | + | ATG | M | ATA | I |
| NODE_1 | 21553 | C-to-U | *ccmC* | + | CAT | H | TAT | Y |
| NODE_1 | 21588 | C-to-U | *ccmC* | + | TTC | F | TTT | F |
| NODE_1 | 21612 | C-to-U | *ccmC* | + | TAC | Y | TAT | Y |
| NODE_1 | 21634 | C-to-U | *ccmC* | + | CGG | R | TGG | W |
| NODE_1 | 21661 | C-to-U | *ccmC* | + | CAT | H | TAT | Y |
| NODE_1 | 21673 | C-to-U | *ccmC* | + | CGG | R | TGG | W |
| NODE_1 | 21684 | A-to-G | *ccmC* | + | CCA | P | CCG | P |
| NODE_1 | 21688 | G-to-A | *ccmC* | + | GAT | D | AAT | N |
| NODE_1 | 21749 | G-to-A | *ccmC* | + | AGT | S | AAT | N |
| NODE_1 | 21794 | C-to-U | *ccmC* | + | CCA | P | CTA | L |
| NODE_1 | 21806 | A-to-G | *ccmC* | + | CAT | H | CGT | R |
| NODE_1 | 21839 | C-to-U | *ccmC* | + | ACA | T | ATA | I |
| NODE_1 | 21857 | C-to-U | *ccmC* | + | TCT | S | TTT | F |
| NODE_1 | 21889 | C-to-U | *ccmC* | + | CGG | R | TGG | W |
| NODE_1 | 21932 | G-to-A | *ccmC* | + | CGT | R | CAT | H |
| NODE_1 | 21936 | C-to-A | *ccmC* | + | TTC | F | TTA | L |
| NODE_1 | 21959 | U-to-G | *ccmC* | + | CTT | L | CGT | R |
| NODE_1 | 22010 | C-to-U | *ccmC* | + | CCT | P | CTT | L |
| NODE_1 | 22057 | C-to-U | *ccmC* | + | CCA | P | TCA | S |
| NODE_1 | 22060 | G-to-A | *ccmC* | + | GTC | V | ATC | I |
| NODE_1 | 22079 | C-to-U | *ccmC* | + | TCG | S | TTG | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_1 | 22080 | G-to-A | *ccmC* | + | TCG | S | TCA | S |
| NODE_1 | 22105 | U-to-C | *ccmC* | + | TCT | S | CCT | P |
| NODE_1 | 22214 | C-to-U | *ccmC* | + | CCA | P | CTA | L |
| NODE_1 | 22218 | U-to-C | *ccmC* | + | TCT | S | TCC | S |
| NODE_1 | 22231 | C-to-U | *ccmC* | + | CCC | P | TCC | S |
| NODE_1 | 47759 | G-to-A | *nad1* | - | TCA | S | TTA | L |
| NODE_1 | 47814 | G-to-A | *nad1* | - | CGT | R | TGT | C |
| NODE_1 | 47817 | G-to-A | *nad1* | - | CCT | P | TCT | S |
| NODE_1 | 47823 | G-to-A | *nad1* | - | CTT | L | TTT | F |
| NODE_1 | 47858 | G-to-A | *nad1* | - | TCC | S | TTC | F |
| NODE_1 | 47894 | G-to-A | *nad1* | - | TCG | S | TTG | L |
| NODE_1 | 47904 | G-to-A | *nad1* | - | CCC | P | TCC | S |
| NODE_1 | 49534 | G-to-A | *nad1* | - | CCT | P | TCT | S |
| NODE_1 | 54372 | G-to-A | *atp6* | - | CAA | Q | TAA | * |
| NODE_1 | 54383 | G-to-A | *atp6* | - | ACA | T | ATA | I |
| NODE_1 | 54410 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 54419 | G-to-A | *atp6* | - | TCT | S | TTT | F |
| NODE_1 | 54426 | G-to-A | *atp6* | - | CAT | H | TAT | Y |
| NODE_1 | 54434 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 54485 | G-to-A | *atp6* | - | CCT | P | CTT | L |
| NODE_1 | 54563 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 54605 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 54627 | G-to-A | *atp6* | - | CAT | H | TAT | Y |
| NODE_1 | 54630 | G-to-A | *atp6* | - | CCT | P | TCT | S |
| NODE_1 | 54689 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 54796 | G-to-A | *atp6* | - | TTC | F | TTT | F |
| NODE_1 | 54820 | G-to-A | *atp6* | - | CCC | P | CCT | P |
| NODE_1 | 54821 | G-to-A | *atp6* | - | CCC | P | CTC | L |
| NODE_1 | 54836 | G-to-A | *atp6* | - | TCG | S | TTG | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_1 | 54861 | G-to-A | *atp6* | - | CGC | R | TGC | C |
| NODE_1 | 54866 | G-to-A | *atp6* | - | TCC | S | TTC | F |
| NODE_1 | 54911 | G-to-A | *atp6* | - | CCG | P | CTG | L |
| NODE_1 | 54917 | G-to-A | *atp6* | - | CCG | P | CTG | L |
| NODE_1 | 54930 | G-to-A | *atp6* | - | CAT | H | TAT | Y |
| NODE_1 | 54968 | G-to-A | *atp6* | - | TCA | S | TTA | L |
| NODE_1 | 55037 | G-to-A | *atp6* | - | TCT | S | TTT | F |
| NODE_1 | 55047 | G-to-A | *atp6* | - | CCA | P | TCA | S |
| NODE_1 | 55054 | G-to-A | *atp6* | - | TTC | F | TTT | F |
| NODE_1 | 113640 | C-to-U | *nad5* | + | CCA | P | CTA | L |
| NODE_1 | 114581 | C-to-U | *nad5* | + | CCC | P | CCT | P |
| NODE_1 | 114611 | C-to-U | *nad5* | + | ATC | I | ATT | I |
| NODE_1 | 114697 | C-to-U | *nad5* | + | TCC | S | TTC | F |
| NODE_1 | 114698 | C-to-U | *nad5* | + | TCC | S | TCT | S |
| NODE_1 | 114713 | C-to-U | *nad5* | + | GCC | A | GCT | A |
| NODE_1 | 114737 | C-to-U | *nad5* | + | CTC | L | CTT | L |
| NODE_1 | 114804 | C-to-U | *nad5* | + | CAC | H | TAC | Y |
| NODE_1 | 114833 | C-to-U | *nad5* | + | TAC | Y | TAT | Y |
| NODE_1 | 114845 | C-to-U | *nad5* | + | GCC | A | GCT | A |
| NODE_1 | 114887 | C-to-U | *nad5* | + | TTC | F | TTT | F |
| NODE_1 | 114892 | C-to-U | *nad5* | + | TCG | S | TTG | L |
| NODE_1 | 114937 | C-to-U | *nad5* | + | TCG | S | TTG | L |
| NODE_1 | 114968 | C-to-U | *nad5* | + | TTC | F | TTT | F |
| NODE_1 | 114970 | C-to-U | *nad5* | + | TCG | S | TTG | L |
| NODE_1 | 115015 | C-to-U | *nad5* | + | TCT | S | TTT | F |
| NODE_1 | 115052 | C-to-U | *nad5* | + | ATC | I | ATT | I |
| NODE_1 | 115060 | C-to-U | *nad5* | + | TCG | S | TTG | L |
| NODE_1 | 115064 | C-to-U | *nad5* | + | GTC | V | GTT | V |
| NODE_1 | 115103 | C-to-U | *nad5* | + | TTC | F | TTT | F |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_1 | 115174 | C-to-U | *nad5* | + | CCC | P | CTC | L |
| NODE_1 | 115202 | C-to-U | *nad5* | + | TTC | F | TTT | F |
| NODE_1 | 115522 | C-to-U | *nad5* | + | TCC | S | TTC | F |
| NODE_1 | 115523 | C-to-U | *nad5* | + | TCC | S | TCT | S |
| NODE_1 | 115657 | C-to-U | *nad5* | + | TCT | S | TTT | F |
| NODE_1 | 115739 | C-to-U | *nad5* | + | TTC | F | TTT | F |
| NODE_2 | 35266 | G-to-A | *nad9* | - | TCT | S | TTT | F |
| NODE_2 | 35399 | G-to-A | *nad9* | - | CAT | H | TAT | Y |
| NODE_2 | 35407 | G-to-A | *nad9* | - | TCA | S | TTA | L |
| NODE_2 | 35437 | G-to-A | *nad9* | - | TCC | S | TTC | F |
| NODE_2 | 35477 | G-to-A | *nad9* | - | CGG | R | TGG | W |
| NODE_2 | 35507 | G-to-A | *nad9* | - | CCG | P | TCG | S |
| NODE_2 | 35564 | G-to-A | *nad9* | - | CGG | R | TGG | W |
| NODE_2 | 35582 | G-to-A | *nad9* | - | CAT | H | TAT | Y |
| NODE_2 | 35615 | G-to-A | *nad9* | - | CAT | H | TAT | Y |
| NODE_2 | 35638 | G-to-A | *nad9* | - | TCG | S | TTG | L |
| NODE_2 | 35692 | G-to-A | *nad9* | - | CCA | P | CTA | L |
| NODE_2 | 35713 | G-to-A | *nad9* | - | TCT | S | TTT | F |
| NODE_2 | 47902 | C-to-U | *nad3* | + | TCG | S | TTG | L |
| NODE_2 | 47941 | C-to-U | *nad3* | + | CCG | P | CTG | L |
| NODE_2 | 47959 | C-to-U | *nad3* | + | CCA | P | CTA | L |
| NODE_2 | 47976 | C-to-U | *nad3* | + | CCA | P | TCA | S |
| NODE_2 | 47977 | C-to-U | *nad3* | + | CCA | P | CTA | L |
| NODE_2 | 48021 | C-to-U | *nad3* | + | CAC | H | TAC | Y |
| NODE_2 | 48034 | C-to-U | *nad3* | + | TCC | S | TTC | F |
| NODE_2 | 48035 | C-to-U | *nad3* | + | TCC | S | TCT | S |
| NODE_2 | 48043 | C-to-U | *nad3* | + | TCC | S | TTC | F |
| NODE_2 | 48087 | C-to-U | *nad3* | + | CCT | P | TCT | S |
| NODE_2 | 48105 | C-to-U | *nad3* | + | CCT | P | TCT | S |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_2 | 48106 | C-to-U | *nad3* | + | CCT | P | CTT | L |
| NODE_2 | 48112 | C-to-U | *nad3* | + | CCG | P | CTG | L |
| NODE_2 | 48127 | C-to-U | *nad3* | + | TCT | S | TTT | F |
| NODE_2 | 48130 | C-to-U | *nad3* | + | TCT | S | TTT | F |
| NODE_2 | 48144 | C-to-U | *nad3* | + | CCT | P | TCT | S |
| NODE_2 | 48163 | C-to-U | *nad3* | + | CCG | P | CTG | L |
| NODE_2 | 48172 | C-to-U | *nad3* | + | TCT | S | TTT | F |
| NODE_2 | 48214 | C-to-U | *nad3* | + | TCT | S | TTT | F |
| NODE_2 | 48241 | C-to-U | *nad3* | + | TCG | S | TTG | L |
| NODE_2 | 48246 | C-to-U | *nad3* | + | CGG | R | TGG | W |
| NODE_2 | 48370 | C-to-U | *rps12* | + | TCG | S | TTG | L |
| NODE_2 | 48403 | C-to-U | *rps12* | + | CCG | P | CTG | L |
| NODE_2 | 48495 | C-to-U | *rps12* | + | CAC | H | TAC | Y |
| NODE_2 | 48520 | C-to-U | *rps12* | + | TCG | S | TTG | L |
| NODE_2 | 48531 | C-to-U | *rps12* | + | CCT | P | TCT | S |
| NODE_2 | 48568 | C-to-U | *rps12* | + | TCG | S | TTG | L |
| NODE_2 | 48583 | C-to-U | *rps12* | + | TCC | S | TTC | F |
| NODE_2 | 60359 | C-to-U | *cox1* | + | TCA | S | TTA | L |
| NODE_2 | 60631 | C-to-U | *cox1* | + | CGG | R | TGG | W |
| NODE_2 | 60672 | C-to-U | *cox1* | + | CCC | P | CCT | P |
| NODE_2 | 61175 | C-to-U | *cox1* | + | TCT | S | TTT | F |
| NODE_2 | 61318 | C-to-U | *cox1* | + | CGT | R | TGT | C |
| NODE_2 | 61340 | C-to-U | *cox1* | + | ACA | T | ATA | I |
| NODE_2 | 70534 | G-to-A | *cox2* | - | CGG | R | TGG | W |
| NODE_2 | 71880 | G-to-A | *cox2* | - | ACG | T | ATG | M |
| NODE_2 | 71902 | G-to-A | *cox2* | - | CGT | R | TGT | C |
| NODE_2 | 71946 | G-to-A | *cox2* | - | TCG | S | TTG | L |
| NODE_2 | 71955 | G-to-A | *cox2* | - | ACC | T | ATC | I |
| NODE_2 | 71997 | G-to-A | *cox2* | - | TCA | S | TTA | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_2 | 72021 | G-to-A | *cox2* | - | CCT | P | CTT | L |
| NODE_2 | 72034 | G-to-A | *cox2* | - | CCC | P | TCC | S |
| NODE_2 | 72102 | G-to-A | *cox2* | - | TCA | S | TTA | L |
| NODE_2 | 72117 | G-to-A | *cox2* | - | CCA | P | CTA | L |
| NODE_2 | 72135 | G-to-A | *cox2* | - | ACG | T | ATG | M |
| NODE_2 | 73843 | C-to-U | *mttB* | + | TTC | F | TTT | F |
| NODE_2 | 73847 | C-to-U | *mttB* | + | CCG | P | TCG | S |
| NODE_2 | 73848 | C-to-U | *mttB* | + | CCG | P | CTG | L |
| NODE_2 | 73886 | C-to-U | *mttB* | + | CGG | R | TGG | W |
| NODE_2 | 73897 | C-to-U | *mttB* | + | ATC | I | ATT | I |
| NODE_2 | 73922 | C-to-U | *mttB* | + | CGT | R | TGT | C |
| NODE_2 | 73933 | C-to-U | *mttB* | + | TTC | F | TTT | F |
| NODE_2 | 73934 | C-to-U | *mttB* | + | CCG | P | TCG | S |
| NODE_2 | 73950 | C-to-U | *mttB* | + | TCT | S | TTT | F |
| NODE_2 | 74006 | C-to-U | *mttB* | + | CGT | R | TGT | C |
| NODE_2 | 74016 | C-to-U | *mttB* | + | TCA | S | TTA | L |
| NODE_2 | 74028 | C-to-U | *mttB* | + | TCC | S | TTC | F |
| NODE_2 | 74030 | C-to-U | *mttB* | + | CCG | P | TCG | S |
| NODE_2 | 74051 | C-to-U | *mttB* | + | CCA | P | TCA | S |
| NODE_2 | 74064 | C-to-U | *mttB* | + | TCT | S | TTT | F |
| NODE_2 | 74071 | C-to-U | *mttB* | + | TTC | F | TTT | F |
| NODE_2 | 74090 | C-to-U | *mttB* | + | CAT | H | TAT | Y |
| NODE_2 | 74109 | C-to-U | *mttB* | + | TCG | S | TTG | L |
| NODE_2 | 74156 | C-to-U | *mttB* | + | CTC | L | TTC | F |
| NODE_2 | 74159 | C-to-U | *mttB* | + | CAT | H | TAT | Y |
| NODE_2 | 74174 | C-to-U | *mttB* | + | CGC | R | TGC | C |
| NODE_2 | 74204 | C-to-U | *mttB* | + | CCC | P | TCC | S |
| NODE_2 | 74265 | C-to-U | *mttB* | + | TCG | S | TTG | L |
| NODE_2 | 74300 | C-to-U | *mttB* | + | CAT | H | TAT | Y |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_2 | 74325 | C-to-U | *mttB* | + | TCG | S | TTG | L |
| NODE_2 | 74333 | C-to-U | *mttB* | + | CCA | P | TCA | S |
| NODE_2 | 74357 | C-to-U | *mttB* | + | CAC | H | TAC | Y |
| NODE_2 | 74369 | C-to-U | *mttB* | + | CGT | R | TGT | C |
| NODE_2 | 74382 | C-to-U | *mttB* | + | CCA | P | CTA | L |
| NODE_2 | 74406 | C-to-U | *mttB* | + | TCC | S | TTC | F |
| NODE_2 | 74438 | C-to-U | *mttB* | + | CCG | P | TCG | S |
| NODE_2 | 74444 | C-to-U | *mttB* | + | CTC | L | TTC | F |
| NODE_2 | 74493 | C-to-U | *mttB* | + | GCC | A | GTC | V |
| NODE_2 | 74495 | C-to-U | *mttB* | + | CGT | R | TGT | C |
| NODE_2 | 74532 | C-to-U | *mttB* | + | TCT | S | TTT | F |
| NODE_2 | 87505 | G-to-A | *nad5* | - | TTC | F | TTT | F |
| NODE_2 | 87518 | G-to-A | *nad5* | - | TCG | S | TTG | L |
| NODE_2 | 87558 | G-to-A | *nad5* | - | CGT | R | TGT | C |
| NODE_2 | 87560 | G-to-A | *nad5* | - | TCT | S | TTT | F |
| NODE_2 | 87575 | G-to-A | *nad5* | - | CCA | P | CTA | L |
| NODE_2 | 87576 | G-to-A | *nad5* | - | CCA | P | TCA | S |
| NODE_2 | 87581 | G-to-A | *nad5* | - | TCA | S | TTA | L |
| NODE_2 | 88725 | G-to-A | *nad5* | - | TCT | S | TTT | F |
| NODE_2 | 88826 | G-to-A | *nad5* | - | TTC | F | TTT | F |
| NODE_2 | 88844 | G-to-A | *nad5* | - | TTC | F | TTT | F |
| NODE_2 | 88907 | G-to-A | *nad5* | - | TTC | F | TTT | F |
| NODE_2 | 88919 | G-to-A | *nad5* | - | CTC | L | CTT | L |
| NODE_2 | 88974 | G-to-A | *nad5* | - | CCC | P | CTC | L |
| NODE_2 | 88975 | G-to-A | *nad5* | - | CCC | P | TCC | S |
| NODE_2 | 88995 | G-to-A | *nad5* | - | TCT | S | TTT | F |
| NODE_2 | 89004 | G-to-A | *nad5* | - | TCA | S | TTA | L |
| NODE_2 | 89034 | G-to-A | *nad5* | - | ACA | T | ATA | I |
| NODE_2 | 89094 | G-to-A | *nad5* | - | CCC | P | CTC | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_2 | 89562 | G-to-A | *atp4* | - | ACT | T | ATT | I |
| NODE_2 | 89583 | G-to-A | *atp4* | - | TCA | S | TTA | L |
| NODE_2 | 89748 | G-to-A | *atp4* | - | CCG | P | CTG | L |
| NODE_2 | 89749 | G-to-A | *atp4* | - | CCG | P | TCG | S |
| NODE_2 | 89751 | G-to-A | *atp4* | - | CCT | P | CTT | L |
| NODE_2 | 89772 | G-to-A | *atp4* | - | TCC | S | TTC | F |
| NODE_2 | 89881 | G-to-A | *atp4* | - | CGT | R | TGT | C |
| NODE_2 | 89910 | G-to-A | *atp4* | - | TCA | S | TTA | L |
| NODE_2 | 89923 | G-to-A | *atp4* | - | CCG | P | TCG | S |
| NODE_2 | 89928 | G-to-A | *atp4* | - | TCA | S | TTA | L |
| NODE_2 | 90159 | G-to-A | *nad4L* | - | TCT | S | TTT | F |
| NODE_2 | 90200 | G-to-A | *nad4L* | - | TTC | F | TTT | F |
| NODE_2 | 90210 | G-to-A | *nad4L* | - | TCA | S | TTA | L |
| NODE_2 | 90243 | G-to-A | *nad4L* | - | CCA | P | CTA | L |
| NODE_2 | 90252 | G-to-A | *nad4L* | - | TCA | S | TTA | L |
| NODE_2 | 90261 | G-to-A | *nad4L* | - | TCA | S | TTA | L |
| NODE_2 | 90309 | G-to-A | *nad4L* | - | TCG | S | TTG | L |
| NODE_2 | 90354 | G-to-A | *nad4L* | - | CCT | P | CTT | L |
| NODE_2 | 90385 | G-to-A | *nad4L* | - | CGG | R | TGG | W |
| NODE_2 | 90399 | G-to-A | *nad4L* | - | TCT | S | TTT | F |
| NODE_2 | 90432 | G-to-A | *nad4L* | - | CCT | P | CTT | L |
| NODE_2 | 90438 | G-to-A | *nad4L* | - | ACG | T | ATG | M |
| NODE_2 | 98961 | C-to-U | *atp1* | + | TCC | S | TCT | S |
| NODE_2 | 99208 | C-to-U | *atp1* | + | CCC | P | TCC | S |
| NODE_2 | 99233 | C-to-U | *atp1* | + | TCG | S | TTG | L |
| NODE_2 | 99337 | C-to-U | *atp1* | + | CGC | R | TGC | C |
| NODE_2 | 99347 | C-to-U | *atp1* | + | TCA | S | TTA | L |
| NODE_2 | 99431 | C-to-U | *atp1* | + | CCC | P | CTC | L |
| NODE_2 | 99432 | C-to-U | *atp1* | + | CCC | P | CCT | P |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_2 | 99461 | C-to-U | *atp1* | + | CCC | P | CTC | L |
| NODE_2 | 99487 | C-to-U | *atp1* | + | CCA | P | TCA | S |
| NODE_2 | 99584 | C-to-U | *atp1* | + | CCA | P | CTA | L |
| NODE_2 | 99600 | C-to-U | *atp1* | + | CCC | P | CCT | P |
| NODE_2 | 99659 | C-to-U | *atp1* | + | CCA | P | CTA | L |
| NODE_2 | 99668 | C-to-U | *atp1* | + | TCT | S | TTT | F |
| NODE_4 | 145 | G-to-A | *nad7* | - | TCT | S | TTT | F |
| NODE_4 | 187 | G-to-A | *nad7* | - | CCA | P | CTA | L |
| NODE_4 | 208 | G-to-A | *nad7* | - | TCT | S | TTT | F |
| NODE_4 | 223 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 232 | G-to-A | *nad7* | - | TCT | S | TTT | F |
| NODE_4 | 254 | G-to-A | *nad7* | - | CGT | R | TGT | C |
| NODE_4 | 261 | G-to-A | *nad7* | - | CCC | P | CCT | P |
| NODE_4 | 338 | G-to-A | *nad7* | - | CCT | P | TCT | S |
| NODE_4 | 367 | G-to-A | *nad7* | - | CCT | P | CTT | L |
| NODE_4 | 2541 | G-to-A | *nad7* | - | CCT | P | CTT | L |
| NODE_4 | 2588 | G-to-A | *nad7* | - | ATC | I | ATT | I |
| NODE_4 | 2608 | G-to-A | *nad7* | - | CGC | R | TGC | C |
| NODE_4 | 2637 | G-to-A | *nad7* | - | CCT | P | CTT | L |
| NODE_4 | 2638 | G-to-A | *nad7* | - | CCT | P | TCT | S |
| NODE_4 | 2643 | G-to-A | *nad7* | - | TCG | S | TTG | L |
| NODE_4 | 2653 | G-to-A | *nad7* | - | CAT | H | TAT | Y |
| NODE_4 | 2679 | G-to-A | *nad7* | - | TCG | S | TTG | L |
| NODE_4 | 3980 | G-to-A | *nad7* | - | CCT | P | TCT | S |
| NODE_4 | 4081 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 4126 | G-to-A | *nad7* | - | TCC | S | TTC | F |
| NODE_4 | 4276 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 4306 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 4315 | G-to-A | *nad7* | - | TCA | S | TTA | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_4 | 4324 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 4343 | G-to-A | *nad7* | - | CGT | R | TGT | C |
| NODE_4 | 4408 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 4415 | G-to-A | *nad7* | - | CAT | H | TAT | Y |
| NODE_4 | 5612 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 7102 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 7156 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 7162 | G-to-A | *nad7* | - | TCA | S | TTA | L |
| NODE_4 | 7195 | G-to-A | *nad7* | - | TCC | S | TTC | F |
| NODE_4 | 7201 | G-to-A | *nad7* | - | TCG | S | TTG | L |
| NODE_4 | 7215 | G-to-A | *nad7* | - | ATC | I | ATT | I |
| NODE_4 | 14725 | G-to-A | *nad2* | - | TCG | S | TTG | L |
| NODE_4 | 14794 | G-to-A | *nad2* | - | CCT | P | CTT | L |
| NODE_4 | 14821 | G-to-A | *nad2* | - | TCA | S | TTA | L |
| NODE_4 | 14828 | G-to-A | *nad2* | - | CAT | H | TAT | Y |
| NODE_4 | 14911 | G-to-A | *nad2* | - | TCC | S | TTC | F |
| NODE_4 | 14956 | G-to-A | *nad2* | - | CCA | P | CTA | L |
| NODE_4 | 14960 | G-to-A | *nad2* | - | CTT | L | TTT | F |
| NODE_4 | 14970 | G-to-A | *nad2* | - | TTC | F | TTT | F |
| NODE_4 | 16796 | G-to-A | *nad2* | - | TCC | S | TTC | F |
| NODE_4 | 25351 | G-to-A | *nad1* | - | CGG | R | TGG | W |
| NODE_4 | 25419 | G-to-A | *nad1* | - | CCG | P | CTG | L |
| NODE_4 | 25420 | G-to-A | *nad1* | - | CCG | P | TCG | S |
| NODE_4 | 25462 | G-to-A | *nad1* | - | CGG | R | TGG | W |
| NODE_4 | 25512 | G-to-A | *nad1* | - | TCC | S | TTC | F |
| NODE_4 | 25560 | G-to-A | *nad1* | - | TCG | S | TTG | L |
| NODE_4 | 25589 | G-to-A | *nad1* | - | TTC | F | TTT | F |
| NODE_4 | 25593 | G-to-A | *nad1* | - | TCA | S | TTA | L |
| NODE_4 | 25725 | G-to-A | *nad1* | - | ACG | T | ATG | M |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_5 | 21616 | G-to-A | *ccmB* | - | TCA | S | TTA | L |
| NODE_5 | 21655 | G-to-A | *ccmB* | - | CCG | P | CTG | L |
| NODE_5 | 21661 | G-to-A | *ccmB* | - | TCC | S | TTC | F |
| NODE_5 | 21673 | G-to-A | *ccmB* | - | TCG | S | TTG | L |
| NODE_5 | 21676 | G-to-A | *ccmB* | - | TCA | S | TTA | L |
| NODE_5 | 21799 | G-to-A | *ccmB* | - | TCG | S | TTG | L |
| NODE_5 | 21803 | G-to-A | *ccmB* | - | CGT | R | TGT | C |
| NODE_5 | 21821 | G-to-A | *ccmB* | - | CTG | L | TTG | L |
| NODE_5 | 21860 | G-to-A | *ccmB* | - | CGG | R | TGG | W |
| NODE_5 | 21889 | G-to-A | *ccmB* | - | CCG | P | CTG | L |
| NODE_5 | 21890 | G-to-A | *ccmB* | - | CCG | P | TCG | S |
| NODE_5 | 21914 | G-to-A | *ccmB* | - | CGT | R | TGT | C |
| NODE_5 | 21941 | G-to-A | *ccmB* | - | CGG | R | TGG | W |
| NODE_5 | 22147 | G-to-A | *ccmB* | - | TCG | S | TTG | L |
| NODE_5 | 22182 | G-to-A | *ccmB* | - | CCC | P | CCT | P |
| NODE_5 | 22184 | G-to-A | *ccmB* | - | CCC | P | TCC | S |
| NODE_5 | 22188 | G-to-A | *ccmB* | - | ATC | I | ATT | I |
| NODE_5 | 22199 | G-to-A | *ccmB* | - | CAT | H | TAT | Y |
| NODE_6 | 13636 | G-to-A | *cox2* | - | CGG | R | TGG | W |
| NODE_6 | 13737 | G-to-A | *cox2* | - | CCG | P | CTG | L |
| NODE_6 | 13762 | G-to-A | *cox2* | - | CGG | R | TGG | W |
| NODE_6 | 13852 | G-to-A | *cox2* | - | CGG | R | TGG | W |
| NODE_6 | 13854 | G-to-A | *cox2* | - | TCA | S | TTA | L |
| NODE_6 | 13859 | G-to-A | *cox2* | - | TTC | F | TTT | F |
| NODE_6 | 13877 | G-to-A | *cox2* | - | CTC | L | CTT | L |
| NODE_6 | 13944 | G-to-A | *cox2* | - | TCT | S | TTT | F |
| NODE_6 | 13982 | G-to-A | *cox2* | - | ATC | I | ATT | I |
| NODE_6 | 13991 | G-to-A | *cox2* | - | TTC | F | TTT | F |
| NODE_6 | 14012 | G-to-A | *cox2* | - | TTC | F | TTT | F |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_6 | 20368 | C-to-U | *ccmFc* | + | TCC | S | TTC | F |
| NODE_6 | 20380 | C-to-U | *ccmFc* | + | CCT | P | CTT | L |
| NODE_6 | 20382 | C-to-U | *ccmFc* | + | CGT | R | TGT | C |
| NODE_6 | 20417 | C-to-U | *ccmFc* | + | TTC | F | TTT | F |
| NODE_6 | 20433 | C-to-U | *ccmFc* | + | CCC | P | TCC | S |
| NODE_6 | 20476 | C-to-U | *ccmFc* | + | CCT | P | CTT | L |
| NODE_6 | 20481 | C-to-U | *ccmFc* | + | CCT | P | TCT | S |
| NODE_6 | 20491 | C-to-U | *ccmFc* | + | CCT | P | CTT | L |
| NODE_6 | 20615 | C-to-U | *ccmFc* | + | TCC | S | TCT | S |
| NODE_6 | 20634 | C-to-U | *ccmFc* | + | CGT | R | TGT | C |
| NODE_6 | 20639 | C-to-U | *ccmFc* | + | TTC | F | TTT | F |
| NODE_6 | 20658 | C-to-U | *ccmFc* | + | CTT | L | TTT | F |
| NODE_6 | 20730 | C-to-U | *ccmFc* | + | CGT | R | TGT | C |
| NODE_6 | 20743 | C-to-U | *ccmFc* | + | TCT | S | TTT | F |
| NODE_6 | 20805 | C-to-U | *ccmFc* | + | CGG | R | TGG | W |
| NODE_6 | 20873 | C-to-U | *ccmFc* | + | GTC | V | GTT | V |
| NODE_6 | 22279 | C-to-U | *ccmFc* | + | CCC | P | TCC | S |
| NODE_6 | 22346 | C-to-U | *ccmFc* | + | TCT | S | TTT | F |
| NODE_6 | 22636 | C-to-U | *ccmFc* | + | CCG | P | TCG | S |
| NODE_6 | 22637 | C-to-U | *ccmFc* | + | CCG | P | CTG | L |
| NODE_6 | 22711 | C-to-U | *ccmFc* | + | CGG | R | TGG | W |
| NODE_6 | 22736 | C-to-U | *ccmFc* | + | TCG | S | TTG | L |
| NODE_6 | 22745 | C-to-U | *ccmFc* | + | TCG | S | TTG | L |
| NODE_6 | 22792 | C-to-U | *ccmFc* | + | CGA | R | TGA | * |
| NODE_6 | 23195 | C-to-U | *matR* | + | CCC | P | CTC | L |
| NODE_6 | 23201 | C-to-U | *matR* | + | TCC | S | TTC | F |
| NODE_6 | 23405 | C-to-U | *matR* | + | TCC | S | TTC | F |
| NODE_6 | 23494 | C-to-U | *matR* | + | CCA | P | TCA | S |
| NODE_6 | 23495 | C-to-U | *matR* | + | CCA | P | CTA | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_6 | 23951 | G-to-A | *matR* | + | AGG | R | AAG | K |
| NODE_6 | 24227 | C-to-U | *matR* | + | CCC | P | CTC | L |
| NODE_6 | 24847 | C-to-U | *matR* | + | CGG | R | TGG | W |
| NODE_6 | 24851 | C-to-U | *matR* | + | TCC | S | TTC | F |
| NODE_6 | 24872 | C-to-U | *matR* | + | CCT | P | CTT | L |
| NODE_6 | 24892 | C-to-U | *matR* | + | CGC | R | TGC | C |
| NODE_6 | 24906 | C-to-U | *matR* | + | TAC | Y | TAT | Y |
| NODE_6 | 24928 | C-to-U | *matR* | + | CAC | H | TAC | Y |
| NODE_6 | 24998 | C-to-U | *matR* | + | CCA | P | CTA | L |
| NODE_6 | 25006 | C-to-U | *matR* | + | CCA | P | TCA | S |
| NODE_6 | 25016 | C-to-U | *matR* | + | TCA | S | TTA | L |
| NODE_6 | 25770 | C-to-U | *nad1* | + | CCA | P | CTA | L |
| NODE_6 | 25800 | C-to-U | *nad1* | + | CCA | P | CTA | L |
| NODE_6 | 25824 | C-to-U | *nad1* | + | TCC | S | TTC | F |
| NODE_6 | 25868 | C-to-U | *nad1* | + | CTT | L | TTT | F |
| NODE_6 | 25938 | C-to-U | *nad1* | + | CCT | P | CTT | L |
| NODE_6 | 25943 | C-to-U | *nad1* | + | CGG | R | TGG | W |
| NODE_6 | 25973 | C-to-U | *nad1* | + | CGG | R | TGG | W |
| NODE_6 | 25982 | C-to-U | *nad1* | + | CCC | P | TCC | S |
| NODE_7 | 1312 | C-to-U | *nad2* | + | TCT | S | TTT | F |
| NODE_7 | 1324 | C-to-U | *nad2* | + | TCA | S | TTA | L |
| NODE_7 | 1333 | C-to-U | *nad2* | + | TCT | S | TTT | F |
| NODE_7 | 1444 | C-to-U | *nad2* | + | CCT | P | CTT | L |
| NODE_7 | 1452 | C-to-U | *nad2* | + | CAT | H | TAT | Y |
| NODE_7 | 1482 | C-to-U | *nad2* | + | CGT | R | TGT | C |
| NODE_7 | 1486 | C-to-U | *nad2* | + | ACT | T | ATT | I |
| NODE_7 | 1552 | C-to-U | *nad2* | + | TCA | S | TTA | L |
| NODE_7 | 1581 | C-to-U | *nad2* | + | CCA | P | TCA | S |
| NODE_7 | 1582 | C-to-U | *nad2* | + | CCA | P | CTA | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_7 | 1651 | C-to-U | *nad2* | + | TCG | S | TTG | L |
| NODE_7 | 1771 | C-to-U | *nad2* | + | CCA | P | CTA | L |
| NODE_7 | 1800 | C-to-U | *nad2* | + | CGT | R | TGT | C |
| NODE_7 | 3380 | C-to-U | *nad2* | + | GCG | A | GTG | V |
| NODE_7 | 3418 | C-to-U | *nad2* | + | CTA | L | TTA | L |
| NODE_7 | 3482 | C-to-U | *nad2* | + | TCA | S | TTA | L |
| NODE_7 | 3485 | C-to-U | *nad2* | + | TCC | S | TTC | F |
| NODE_7 | 3490 | C-to-U | *nad2* | + | CCA | P | TCA | S |
| NODE_7 | 3491 | C-to-U | *nad2* | + | CCA | P | CTA | L |
| NODE_7 | 3539 | C-to-U | *nad2* | + | TCA | S | TTA | L |
| NODE_7 | 23934 | G-to-A | *cob* | - | ACC | T | ACT | T |
| NODE_7 | 23998 | G-to-A | *cob* | - | CCG | P | CTG | L |
| NODE_7 | 24024 | G-to-A | *cob* | - | TTC | F | TTT | F |
| NODE_7 | 24208 | G-to-A | *cob* | - | TCT | S | TTT | F |
| NODE_7 | 24214 | G-to-A | *cob* | - | CCA | P | CTA | L |
| NODE_7 | 24269 | G-to-A | *cob* | - | CAT | H | TAT | Y |
| NODE_7 | 24314 | G-to-A | *cob* | - | CCC | P | TCC | S |
| NODE_7 | 24397 | G-to-A | *cob* | - | TCT | S | TTT | F |
| NODE_7 | 24407 | G-to-A | *cob* | - | CGG | R | TGG | W |
| NODE_7 | 24442 | G-to-A | *cob* | - | TCT | S | TTT | F |
| NODE_7 | 24542 | G-to-A | *cob* | - | CTT | L | TTT | F |
| NODE_7 | 24554 | G-to-A | *cob* | - | CAT | H | TAT | Y |
| NODE_7 | 24558 | G-to-A | *cob* | - | CTC | L | CTT | L |
| NODE_7 | 24703 | G-to-A | *cob* | - | CCA | P | CTA | L |
| NODE_7 | 24797 | G-to-A | *cob* | - | CAT | H | TAT | Y |
| NODE_7 | 24824 | G-to-A | *cob* | - | CAC | H | TAC | Y |
| NODE_7 | 24836 | G-to-A | *cob* | - | CTC | L | TTC | F |
| NODE_7 | 25071 | G-to-A | *cob* | - | TCC | S | TCT | S |
| NODE_7 | 26186 | G-to-A | *ccmFN2* | - | CGT | R | TGT | C |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_7 | 26246 | G-to-A | *ccmFN2* | - | CGG | R | TGG | W |
| NODE_7 | 26281 | G-to-A | *ccmFN2* | - | TCT | S | TTT | F |
| NODE_7 | 26294 | G-to-A | *ccmFN2* | - | CCC | P | TCC | S |
| NODE_7 | 26320 | G-to-A | *ccmFN2* | - | TCT | S | TTT | F |
| NODE_7 | 26341 | G-to-A | *ccmFN2* | - | CCA | P | CTA | L |
| NODE_7 | 26384 | G-to-A | *ccmFN2* | - | CTT | L | TTT | F |
| NODE_7 | 26385 | G-to-A | *ccmFN2* | - | CCC | P | CCT | P |
| NODE_7 | 26408 | G-to-A | *ccmFN2* | - | CGT | R | TGT | C |
| NODE_7 | 26426 | G-to-A | *ccmFN2* | - | CGG | R | TGG | W |
| NODE_7 | 26459 | G-to-A | *ccmFN2* | - | CGG | R | TGG | W |
| NODE_7 | 26477 | G-to-A | *ccmFN2* | - | CGG | R | TGG | W |
| NODE_7 | 26492 | G-to-A | *ccmFN2* | - | CAT | H | TAT | Y |
| NODE_7 | 26509 | G-to-A | *ccmFN2* | - | CCA | P | CTA | L |
| NODE_7 | 26512 | G-to-A | *ccmFN2* | - | TCG | S | TTG | L |
| NODE_7 | 26537 | G-to-A | *ccmFN2* | - | CGG | R | TGG | W |
| NODE_8 | 827 | C-to-U | *atp8* | + | CTT | L | TTT | F |
| NODE_8 | 831 | C-to-U | *atp8* | + | TCT | S | TTT | F |
| NODE_8 | 880 | C-to-U | *atp8* | + | ATC | I | ATT | I |
| NODE_8 | 969 | C-to-U | *atp8* | + | TCA | S | TTA | L |
| NODE_8 | 8731 | G-to-A | *ccmFN1* | - | CGC | R | TGC | C |
| NODE_8 | 8857 | G-to-A | *ccmFN1* | - | CGC | R | TGC | C |
| NODE_8 | 8874 | G-to-A | *ccmFN1* | - | TCA | S | TTA | L |
| NODE_8 | 8889 | G-to-A | *ccmFN1* | - | CCA | P | CTA | L |
| NODE_8 | 8901 | G-to-A | *ccmFN1* | - | TCA | S | TTA | L |
| NODE_8 | 8923 | G-to-A | *ccmFN1* | - | CGT | R | TGT | C |
| NODE_8 | 8961 | G-to-A | *ccmFN1* | - | TCA | S | TTA | L |
| NODE_8 | 8970 | G-to-A | *ccmFN1* | - | CCT | P | CTT | L |
| NODE_8 | 9118 | G-to-A | *ccmFN1* | - | CTT | L | TTT | F |
| NODE_8 | 9164 | G-to-A | *ccmFN1* | - | GCC | A | GCT | A |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_8 | 9195 | G-to-A | *ccmFN1* | - | TCG | S | TTG | L |
| NODE_8 | 9199 | G-to-A | *ccmFN1* | - | CGT | R | TGT | C |
| NODE_8 | 9287 | G-to-A | *ccmFN1* | - | TTC | F | TTT | F |
| NODE_8 | 9294 | G-to-A | *ccmFN1* | - | TCG | S | TTG | L |
| NODE_8 | 9310 | G-to-A | *ccmFN1* | - | CTT | L | TTT | F |
| NODE_8 | 9370 | G-to-A | *ccmFN1* | - | CTT | L | TTT | F |
| NODE_8 | 9390 | G-to-A | *ccmFN1* | - | CCA | P | CTA | L |
| NODE_8 | 9405 | G-to-A | *ccmFN1* | - | TCA | S | TTA | L |
| NODE_8 | 9502 | G-to-A | *ccmFN1* | - | CCT | P | TCT | S |
| NODE_8 | 9511 | G-to-A | *ccmFN1* | - | CGT | R | TGT | C |
| NODE_8 | 9516 | G-to-A | *ccmFN1* | - | TCG | S | TTG | L |
| NODE_8 | 9532 | G-to-A | *ccmFN1* | - | CCT | P | TCT | S |
| NODE_8 | 9596 | G-to-A | *ccmFN1* | - | TAC | Y | TAT | Y |
| NODE_8 | 9621 | G-to-A | *ccmFN1* | - | CCG | P | CTG | L |
| NODE_9 | 6880 | G-to-A | *nad4* | - | CTC | L | CTT | L |
| NODE_9 | 6896 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 9349 | G-to-A | *nad4* | - | TCC | S | TTC | F |
| NODE_9 | 9367 | G-to-A | *nad4* | - | CCA | P | CTA | L |
| NODE_9 | 9415 | G-to-A | *nad4* | - | GCG | A | GTG | V |
| NODE_9 | 9517 | G-to-A | *nad4* | - | CCA | P | CTA | L |
| NODE_9 | 9550 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 9580 | G-to-A | *nad4* | - | TCC | S | TTC | F |
| NODE_9 | 9593 | G-to-A | *nad4* | - | CTC | L | TTC | F |
| NODE_9 | 9613 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 9686 | G-to-A | *nad4* | - | CCA | P | TCA | S |
| NODE_9 | 9689 | G-to-A | *nad4* | - | CCT | P | TCT | S |
| NODE_9 | 9706 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 9712 | G-to-A | *nad4* | - | CCG | P | CTG | L |
| NODE_9 | 9716 | G-to-A | *nad4* | - | CTA | L | TTA | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_9 | 13575 | G-to-A | *nad4* | - | TCT | S | TTT | F |
| NODE_9 | 13650 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 13659 | G-to-A | *nad4* | - | TCG | S | TTG | L |
| NODE_9 | 13663 | G-to-A | *nad4* | - | CCC | P | TCC | S |
| NODE_9 | 13689 | G-to-A | *nad4* | - | CCA | P | CTA | L |
| NODE_9 | 13710 | G-to-A | *nad4* | - | TCT | S | TTT | F |
| NODE_9 | 13714 | G-to-A | *nad4* | - | CGT | R | TGT | C |
| NODE_9 | 13779 | G-to-A | *nad4* | - | CCT | P | CTT | L |
| NODE_9 | 13887 | G-to-A | *nad4* | - | TCT | S | TTT | F |
| NODE_9 | 14003 | G-to-A | *nad4* | - | TCC | S | TCT | S |
| NODE_9 | 14013 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 15496 | G-to-A | *nad4* | - | CCA | P | CTA | L |
| NODE_9 | 15508 | G-to-A | *nad4* | - | CCC | P | CTC | L |
| NODE_9 | 15509 | G-to-A | *nad4* | - | CCC | P | TCC | S |
| NODE_9 | 15512 | G-to-A | *nad4* | - | CTT | L | TTT | F |
| NODE_9 | 15544 | G-to-A | *nad4* | - | TCC | S | TTC | F |
| NODE_9 | 15569 | G-to-A | *nad4* | - | CGT | R | TGT | C |
| NODE_9 | 15577 | G-to-A | *nad4* | - | TCT | S | TTT | F |
| NODE_9 | 15583 | G-to-A | *nad4* | - | ACC | T | ATC | I |
| NODE_9 | 15628 | G-to-A | *nad4* | - | TCA | S | TTA | L |
| NODE_9 | 15748 | G-to-A | *nad4* | - | TCT | S | TTT | F |
| NODE_9 | 15779 | G-to-A | *nad4* | - | CGG | R | TGG | W |
| NODE_9 | 15781 | G-to-A | *nad4* | - | CCT | P | CTT | L |
| NODE_9 | 15787 | G-to-A | *nad4* | - | CCT | P | CTT | L |
| NODE_9 | 15791 | G-to-A | *nad4* | - | CCC | P | TCC | S |
| NODE_9 | 15838 | G-to-A | *nad4* | - | CCG | P | CTG | L |
| NODE_9 | 15868 | G-to-A | *nad4* | - | CCT | P | CTT | L |
| NODE_9 | 15871 | G-to-A | *nad4* | - | ACT | T | ATT | I |
| NODE_9 | 15901 | G-to-A | *nad4* | - | CCT | P | CTT | L |

| Node | Position | Observed change | Gene | Strand | Genome codon | Amino acid | Transcript codon | Amino acid |
|---|---|---|---|---|---|---|---|---|
| NODE_9 | 15915 | G-to-A | *nad4* | - | TCC | S | TCT | S |
| NODE_9 | 15916 | G-to-A | *nad4* | - | TCC | S | TTC | F |
| NODE_11 | 3742 | G-to-A | *cox3* | - | CCA | P | CTA | L |
| NODE_11 | 3752 | G-to-A | *cox3* | - | CGG | R | TGG | W |
| NODE_11 | 3940 | G-to-A | *cox3* | - | TCC | S | TTC | F |
| NODE_11 | 3979 | G-to-A | *cox3* | - | TCC | S | TTC | F |
| NODE_11 | 3994 | G-to-A | *cox3* | - | TCA | S | TTA | L |
| NODE_11 | 4084 | G-to-A | *cox3* | - | CCT | P | CTT | L |
| NODE_11 | 4093 | G-to-A | *cox3* | - | CCT | P | CTT | L |
| NODE_11 | 4192 | G-to-A | *cox3* | - | TCT | S | TTT | F |
| NODE_11 | 4195 | G-to-A | *cox3* | - | TCT | S | TTT | F |
| NODE_11 | 4217 | G-to-A | *cox3* | - | CTT | L | TTT | F |
| NODE_11 | 4249 | G-to-A | *cox3* | - | TCT | S | TTT | F |
| NODE_11 | 4261 | G-to-A | *cox3* | - | CCT | P | CTT | L |
| NODE_11 | 4394 | G-to-A | *cox3* | - | CCA | P | TCA | S |
| NODE_14 | 6067 | C-to-U | *nad1* | + | CCG | P | TCG | S |
| NODE_14 | 6083 | C-to-U | *nad1* | + | TCT | S | TTT | F |
| NODE_18 | 940 | C-to-U | *nad2* | + | ATC | I | ATT | I |
| NODE_18 | 981 | C-to-U | *nad2* | + | TCT | S | TTT | F |

# MAPtools: Command-Line Tools for Mapping-by-Sequencing and QTL-Seq Analysis and Visualization

César Martínez-Guardiola [1,*], Ricardo Parreño [1,*] and Héctor Candela [1]

[1] Instituto de Bioingeniería, Universidad Miguel Hernández de Elche, Campus de Elche, 03202 Elche, Spain

[*] These authors contributed equally to this work.

Corresponding author: H. Candela (telephone: 34 96 522 25 83; fax: 34 96 665 85 11; E-mail: hcandela@umh.es)



Running head: Tools for MBS and QTL-seq

Keywords: mapping-by-sequencing, QTL-seq, MAPtools

Figures: 6 Tables: 1 Supplemental Figures: 16 Supplemental Tables: 1

## Abstract

### Background

Classical mutagenesis is a powerful tool that has allowed researchers to elucidate the molecular and genetic basis of a plethora of processes in many model species. The integration of these methods with modern massively parallel sequencing techniques, initially in model species but currently also in many crop species, is accelerating the identification of genes underlying a wide range of traits of agronomic interest.

### Results

We have developed MAPtools, an open-source Python3 application designed specifically for the analysis of genomic data from bulked segregant analysis experiments, including mapping-by-sequencing (MBS) and quantitative trait locus sequencing (QTL-seq) experiments. We have extensively tested MAPtools using datasets published in recent literature.

### Conclusions

MAPtools gives users the flexibility to customize their bioinformatics pipeline with various commands for calculating allele count-based statistics, generating plots to pinpoint candidate regions, and annotating the effects of SNP and indel mutations. While extensively tested with plants, the program is versatile and applicable to any species for which a mapping population can be generated and a sequenced genome is available.

### Availability and implementation

MAPtools is available under GPL v3.0 license and documented as a Python3 package at https://github.com/hcandela/MAPtools.

## Introduction

Mapping-by-sequencing (MBS) is a powerful technique that combines high-throughput sequencing technologies and bulked segregant analysis to rapidly map mutations identified in mutagenesis screens. This approach was initially developed for model organism research [1] and holds great promise for enhancing our understanding of complex biological systems, as extensively reviewed by us and other authors [2–5]. Another related technique, called quantitative trait loci sequencing (QTL-seq), has been proposed for the rapid mapping of QTLs in species with a sequenced genome and is accelerating the identification of genomic regions associated with traits of agricultural interest in many crops, often in combination with other experimental approaches, such as classical linkage and QTL mapping or the identification of differentially expressed genes by RNA-seq [6–8].

In this article, we introduce MAPtools, a collection of command-line utilities designed to analyze data from MBS and QTL-seq experiments. A growing number of software tools and analysis pipelines for MBS and QTL-seq data have been released in recent years [9–13], but many of them are limited in terms of the analyses they perform or the species they focus on. Other programs, such as CandiSNP, SIMPLE and Easymap, have prioritized ease of use by non-expert users [14–16]. MAPtools, by contrast, has been designed to be a versatile tool that can receive input data from a stream, giving the users maximum flexibility in their choice of read mappers and variant callers. The versatility of MAPtools is illustrated by its ability to analyze data from different segregating populations and crossing schemes. Using simulated reads, previous authors have extensively studied how the identification of mutated genes depends on the sequencing depth and bulk size, two important factors that users must also consider when designing MBS experiments [17,18]. The syntax of MAPtools is straightforward and similar to that of other highly popular and widely used programs such as SAMtools or BCFtools, which should make it user-friendly for researchers who are already familiar with these programs. For this reason, the learning curve for the program is not expected to pose a significant barrier to new users.

## Implementation

### The MAPtools application

MAPtools is a Python3 v. 3.8-based, standalone application that is distributed under the GPL v3.0 license and freely available to users. Its dependencies are limited to several packages commonly used in scientific computing, including docopt (v. 0.6.2), NumPy (v. 1.24.2), SciPy (v. 1.10.1), pandas (v. 2.0.0), biopython (v. 1.81) and

matplotlib (v. 3.7.1), which can be easily installed with pip (v. 20.0.2). The program has been tested with the latest versions of the libraries available at the time of submission. Its distribution through a GitHub repository (https://github.com/hcandela/MAPtools) will facilitate the long-term maintenance of the code. We developed and tested the program on a desktop computer equipped with two Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz (16 cores, 32 threads) processors and 125 Gi RAM, but the program has also been successfully tested on computers with less memory or fewer cores.

The ability to integrate MAPtools in workflows with other tools gives researchers maximum control over their analysis pipelines (Figure 1). For example, users may choose to filter read pairs based on mapping quality or sequencing depth prior to using MAPtools, or they may want to include extra steps to mark or remove duplicate read pairs that, if present, might bias the real allele frequencies. A typical QTL-seq or MBS analysis pipeline involves aligning reads from distinct samples to a well-annotated reference genome using software tools such as BWA [19] or Bowtie2 [20]. The resulting files in sequence alignment/map format (SAM) [21] are then converted to compressed binary format (BAM) and processed using variant calling software, such as BCFtools [22,23] or GATK [24,25]. MAPtools can directly read input data in uncompressed Variant Call Format (VCF), provided that it includes allelic depth (AD) fields for each sample in separate columns. VCF data with AD fields is produced by BCFtools' `mpileup` command when it is run with the `--annotate` option, and also by GATK's `HaplotypeCaller` if the BAM files contain different RG fields for each sample. We have tested the program with input files produced by BCFtools' `call` command, which can also output data to a stream, and by GATK's HaplotypeCaller command, but MAPtools should also work with any other software that outputs data in VCF format. Although the current version of MAPtools cannot directly read input data from compressed or uncompressed BCF files or from compressed VCF files, this can be easily achieved by including a conversion step in the workflow using BCFtools' `view` command. The amount of RAM required by MAPtools is quite low, particularly when the input VCF file contains only the variant sites.

The initial release of MAPtools (v. 1.0) includes six commands that support the analysis of MBS and QTL-seq data. Specifically, two commands, namely `mbs` and `qtl`, enable the analysis of data from MBS or QTL-seq experiments and calculate different statistics depending on the type of experiment and the available input dataset. The `plot` command plots the results and creates publication-quality figures with their captions. Additionally, the `merge` command can integrate the allele counts from all markers within a window, allowing the output to refer to haplotypes rather than

individual markers. The `annotate` command allows users to assess the effect of all candidate mutations within a user-selected interval. Lastly, the `citation` command provides information on the version of the program in use. The `mbs`, `qtl` and `merge` commands produce output in a VCF-like format, which can be read by the `merge` and `plot` commands. This output consists of a header similar to that of the VCF files, containing information about the program, the options used, as well as the headings and a description of the contents of each column. The header cumulatively records each step performed, which should help to improve its reproducibility.

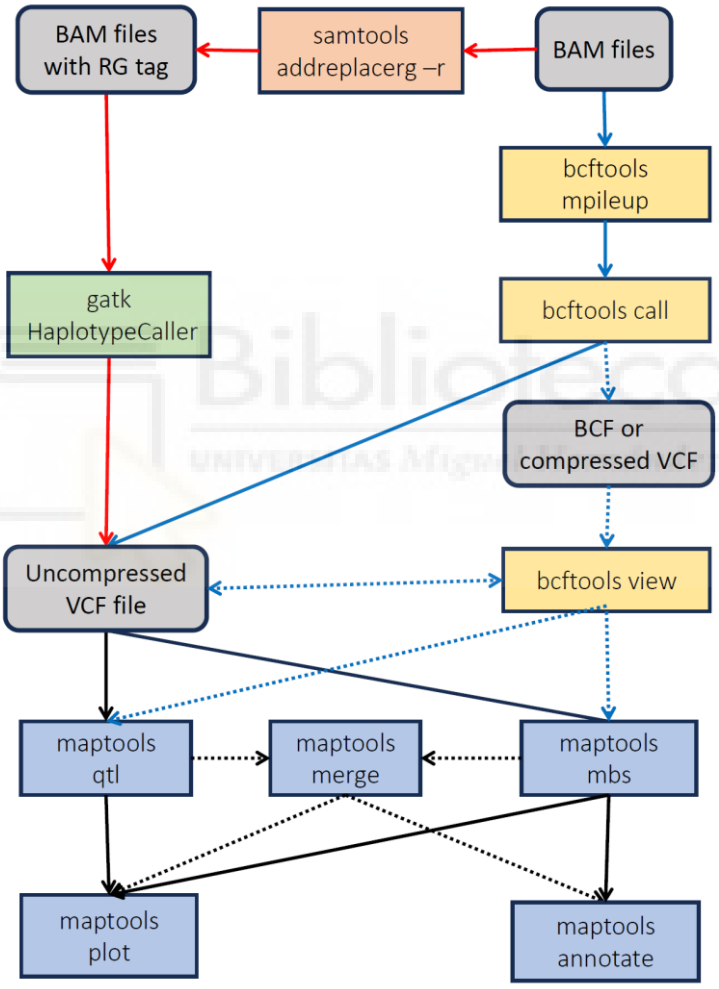



**Figure 1.** MBS and QTL-seq workflows using MAPtools. MAPtools requires data in uncompressed VCF format, which can be read from disk or from a stream. Data in this format can be produced either by BCFtools or by GATK, and serves as input for MAPtools' `qtl` and `mbs` commands. Data in other formats can be converted to uncompressed VCF by BCFtools' view command. This command can also be used to apply additional quality or sequencing depth

filters to the uncompressed VCF data. The output of `qtl` and `mbs` can serve as input for other MAPtool's commands, like `merge`, `plot` and `annotate`.

## MAPtools commands

### Analysis of MBS data

The `mbs` command processes the VCF input to tabulate the allele counts and calculate different parameters that facilitate identifying a mutation of interest through MBS. The user must designate each available sample using option `-d` (`--data`) as one of the following: R (the required bulk of phenotypically recessive individuals from the mapping population), D (an optional bulk of phenotypically dominant individuals from the mapping population), Pr (the phenotypically recessive parent of the mapping population) and Pd (the phenotypically dominant parent of the mapping population). Depending on the available samples, the calculated parameters might include the SNP-index (defined as the frequency of alleles inherited from the phenotypically recessive progenitor), the $\Delta$(SNP-index) (defined as the difference of the SNP-indices in the D and R bulks), the exact probabilities and *p*-values of Fisher's exact tests, the Euclidean distances calculated for individual markers ($ED_m$) and the G statistic, all of which have previously been used in BSA-seq experiments [26]. When the reference genome sequence matches that of one parent of the mapping population, or if one or both parents of the mapping population have been resequenced, the `mbs` command uses this information to classify the alleles based on their parental origin: The program can handle the following experimental situations:

(i) *With genomic DNA from the R bulk*. The user can specify if the alleles in the reference sequence match those of the dominant or the recessive parent enabling the calculation of allele frequencies (AFs, also known as SNP-indices) for the alleles inherited from the phenotypically recessive parent. If such information is unavailable, the program will instead calculate the AF for the most abundant allele, regardless of its parental origin, yielding values equal to or greater than 0.5.

(ii) *With genomic DNA from the R and D bulks*. If the reference genome sequence matches the dominant or the recessive parent, the program will additionally report other parameters useful for comparing the allele counts in the two bulks. Like in the previous case, the command reports the AF for the most abundant allele in the recessive bulk when the parental origin of alleles is unknown.

(iii) *With genomic DNA from the R bulk plus one parent (Pd or Pr).* In this scenario, the sequence of the parent allows determining the parental origin of the alleles, enabling the calculation of the AF for alleles inherited from the recessive parent.

Resequencing of at least one of the parents is particularly recommended when the reference genome does not match any of the parents of the mapping population.

(iv) *With genomic DNA from the R and D bulks plus one parent (Pd or Pr)*. Similar to case (ii), this setup enables the calculation of the AF for alleles from the recessive parent, as well as other parameters that require two bulks from the mapping population.

(v) *With genomic DNA from the R bulk plus the two parents (Pd and Pr)*. This case differs from case (iii) in that the resequencing of the two parents is supplied as input. In this situation, the program will focus on the polymorphisms detected between the two parents and proceed as usual.

(vi) *With genomic DNA from the R and D bulks plus the two parents (Pd and Pr)*. Similar to case (v), this case differs from case (iv) in that the resequencing of the two parents is supplied as input. In this situation, the program will first focus on the polymorphisms detected between the two parents and proceed as usual.

In addition to the R, D, Pr and Pd samples, the program supports the inclusion of an additional wild-type sample, designated Wd or Wr, corresponding to an isogenic line or the non-mutagenized parents of a dominant mutant or a recessive mutant, respectively. The Wd and Wr samples are then used to discard any alleles that were already present prior to mutagenesis. The difference between P samples and W samples is that the former are used for determining the parental origin of alleles, whereas the latter are only used for discarding shared alleles.

**Analysis of QTL-seq data**

In a similar way, the qtl command processes the VCF input data and calculates the necessary parameters for the mapping of QTLs using a QTL-seq strategy. At a minimum, it requires sequence data from two sets of individuals with extreme phenotypes, denoted H ('high') and L ('low'), to calculate several parameters, including the ΔSNP index, p-values from Fisher's exact tests, Euclidean distances for individual markers ($ED_m$), and G statistics, which are commonly used in BSA-seq experiments [26]. This command supports the following scenarios:

(i) *Genomic DNA sequenced from the H and L bulks*. The AFs in the high and low bulks are used to calculate the ΔSNP index when the alleles can be classified based on their parental origin ( i.e., when the reference genome corresponds to one of the parents of the segregating population). If the parental origin of the alleles cannot be

determined, the command reports the absolute value of the ΔSNP index (|ΔSNP index|) in addition to the above statistics for each polymorphic site found.

(ii) *Genomic DNA isolated from the H and L bulks plus the resequencing of one parent (P).* In this case, the alleles can be assigned to haplotypes based on the sequence of the parent, allowing the calculation of ΔSNP indices and other parameters.

Because the `qtl` and `mbs` commands can receive input line-by-line from a Unix pipe, calculations that require data from multiple adjacent SNP markers are deferred to a later step, when they are performed by the `merge` and `plot` commands. This is the case for the calculation of $ED100^4$ values (which are calculated as the fourth power of the sum of the Euclidean distances of 100 consecutive SNPs, and are plotted only for those chromosomes that contain enough markers), the calculation of moving averages of other parameters, or calculations that require knowing the number of statistical tests performed (e.g. significance thresholds corrected for multiple testing using the Bonferroni method).

**Binning of MBS and QTL-seq data**

The `merge` command can be used to post-process the results generated by the `mbs` and `qtl` commands when the parental origin of each allele is known. This command should be particularly useful when the number of polymorphic sites is high (e.g. when using populations involving different genetic backgrounds) but the sequencing depth at each individual site is low, which might yield non-significant results at individual markers. In this case, we reasoned that binning the allele counts of adjacent markers can help to identify regions (rather than individual markers) linked to the trait of interest. Bins can be defined in two alternative ways: (a) as overlapping sets of *n* consecutive markers, or (b) as sets of all markers contained within non-overlapping windows of user-defined length. The read counts of all markers within a bin are aggregated and then used to calculate bin-level haplotype frequencies and other parameters relevant to MBS and QTL-seq experiments. When thousands of polymorphisms segregate in the mapping population and their linkage phase is known, we reasoned that increasing the number of observations would improve the ability of Fisher's exact tests to detect a significant difference between treatments, since a larger sample size provides more information, reduces the impact of random variability, and increases the chances of detecting a true difference if it exists. An example of how the `merge` command affects the results of the analysis is shown below for Case Study 6.

**Plotting the results of MBS and QTL-seq experiments**

The `plot` command is intended to generate publication-quality figures and to facilitate interpretation of the output of the `mbs`, `qtl` and `merge` commands. The set of plots produced can be customized by the user or automatically selected based on the fields present in the header of the output files produced by each command. These plots can be drawn for individual chromosomes or integrated into multi-panel figures: the user can select specific chromosomes and parameters to plot, and can choose to create figures that integrate different parameters for a given chromosome, or figures that display a given parameter for a user-selected set of chromosomes. *p*-values and other parameters can also be presented as Manhattan-like plots. Additional features of this command include the ability to plot moving averages of the allele frequencies (calculated either using all markers within overlapping windows containing a user-defined number of adjacent markers or using all markers within non-overlapping intervals of user-defined length) and 'boost' values (calculated as described for SHOREmap), the ability to overlay the moving average of the AF for the alternative pool, the automatic generation of figure legends, and the availability of different color palettes. As a proxy for a confidence interval for the ΔSNP-index plots, we average the limits of the confidence intervals calculated for individual markers. These limits are calculated based on the formulas for a difference of two proportions, applying the Bonferroni correction to the significance level used ($\alpha$=0.05), taking into account the number of markers. Similarly, the *p*-value plots incorporate a significance threshold level that is calculated using the Bonferroni correction. The user can customize the appearance of the plots by selecting appropriate options on the command line, including predefined color combinations ("palettes"), dot size and transparency, line width, resolution and output file format (EPS, JPG, PDF, PNG or SVG). The program allows users to customize additional aspects by editing a file in JavaScript Object Notation (JSON) format that MAPtools reads from disk. Parameters that can be customized include color palettes, some display attributes, and chromosome aliases. Although the program comes with a default color palette and a color palette optimized for individuals with color blindness, the JSON file also allows users to define their own custom palettes and adjust the size and other display attributes of the plots generated by MAPtools. By default, chromosomes are labeled as they appear in the reference genome's FASTA file, typically using their GenBank accession

numbers, but the JSON file allows users to assign a shorter, alternative display name to each chromosome (e.g., ‘Chr1’). The ability to generate files in vector graphics formats, such as the EPS (Encapsulated PostScript) and SVG (Scalable Vector Graphics) formats, provides an additional level of customization, allowing users to easily edit elements such as caption sizes and axis labels while ensuring that images are displayed at the maximum resolution.

**Functional annotation of identified variants**

The `annotate` command assesses the functional impact of nucleotide substitutions, insertions, and deletions within a user-specified interval. Certain variants are typically excluded based on predefined criteria [27]: (a) the mutant allele matches the reference genome allele, which is assumed to be functional, (b) the substitution is not a G/C-to-A/T transition mutation (the most common type of mutation caused by ethyl methanesulfonate, EMS), or (c) the mutant allele is present in the non-mutagenized parent or other related lines, indicating that it does not cause the observed phenotype. We have implemented these filters in the `mbs` command, allowing the user to generate filtered or unfiltered input for `annotate` at will. The command then quickly evaluates the effect of all candidate mutations passing these filters against the genome annotation provided as a GFF3 file using a binary search algorithm.

This command identifies whether mutations reside in genic or intergenic regions. For intergenic mutations, it reports the identity of and the distance to the nearest adjacent genes. In protein-coding genes, the program determines whether the mutation is located in the 5’ untranslated region (5’-UTR), the coding sequence (CDS), or the 3’ untranslated region (3’-UTR). Substitutions in the CDS are classified as synonymous or non-synonymous (with the latter further classified as missense or nonsense) based on their effect on the amino acid sequence. Indels are checked for frameshift generation. For mutations located in introns, the program reports the distance to the nearest adjacent exons in each transcript isoform. Mutations near donor or acceptor splice sites that could cause missplicing are also reported. Mutations in the 5’-UTR are checked for premature ATG codon creation and their distance to the translation initiation site is reported. Nonstop mutations that replace a stop codon, resulting in continued mRNA translation into the 3’-UTR, are also reported. The program also assesses whether deletions partially or completely disrupt one or more genes.

**Validation**

We have tested MAPtools under a variety of experimental situations, using our own MBS data [28], simulated MBS data, and publicly available MBS and QTL-seq datasets reported in the literature for plant species as diverse as Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) [29,30], rice (*Oryza sativa* L.) [6,31,32], tomato (*Solanum lycopersicum* L.) [33], strawberry (*Fragaria vesca* L.) [34,35], and *Arabis alpina* [36]. As shown below, our results demonstrate the ability of MAPtools to analyze data across different experimental designs and species, as we were able to detect the same QTL and find the same causal mutations as reported in the published studies (Supplemental File 1). Table 1 and Supplemental Table 1 summarize the datasets that have been analyzed and the options used to run the MAPtools' commands `mbs` and `qtl`.

Table 1. Case studies used for testing MAPtools

| Species | Trait or mutation | BioProject (samples) | Reference |
|---|---|---|---|
| *O. sativa* | *lcd1* | PRJNA525315 (D: SRR8695238; R: SRR8695239; Pr: SRR8695240; Pd: SRR8695241) | Cao et al., 2019 |
| *O. sativa* | Suppresor of *xantha* | PRJCA007389 (D: CRR344193; R: CRR344195; Pd: CRR344192; Pr: CRR344194) | Jiang et al., 2022 |
| *O. sativa* | Resistance to rice blast disease | PRJDB2455 (H: DRR003237; L: DRR003238) | Takagi et al., 2013 |
| *S. lycopersicum* | Ascorbate-enriched fruits (AsA+) | | Bournonville et al., 2023 |
| *B. rapa* | *nhm3* | PRJNA761522 (R: SRR15829494; Pd: SRR15803269; Pr: SRR15828094) | Huang et al., 2022 |
| *B. rapa* | Cuticular wax biosynthesis | PRJNA751715 (D: SRR15371666, R: SRR15371667) | Yang et al., 2022 |
| *A. alpina* | *eop002* | PRJNA756904 (R: SRR15564670) and PRJNA608065 (Pd: SRR11140832-SRR11140833) | Viñegra de la Torre et al., 2022 |
| | *eop085* | PRJNA756904 (R: SRR15564669) and PRJNA608065 (Pd: SRR11140832-SRR11140833) | |
| | *eop091* | PRJNA756904 (R: SRR15564668) and PRJNA608065 (Pd: SRR11140832-SRR11140833) | |
| *A. thaliana* | *emb1956-3* | PRJNA9349:07 (D SRR23456103; R: SRR23456104) and PRJNA751183 (Wr: SRR15322352) | Rodríguez-Alcocer et al., 2023 |

| | | | |
|---|---|---|---|
| *F. vesca* | Petiole color | PRJNA823731 (R: SRR18649835; D: SRR18649836) | Luo et al., 2023 |
| *F. vesca* | Fruit color | PRJEB38128 (R: ERR4463155-ERR4463156; D: ERR4463153-ERR4463154) | Castillejo et al., 2020 |

Projects and samples with accession numbers are publicly available from the NCBI (http://ncbi.nlm.nih.gov) and:Genome Sequence Archive BIG Data Center (https://bigd.big.ac.cn/gsa/) databases.

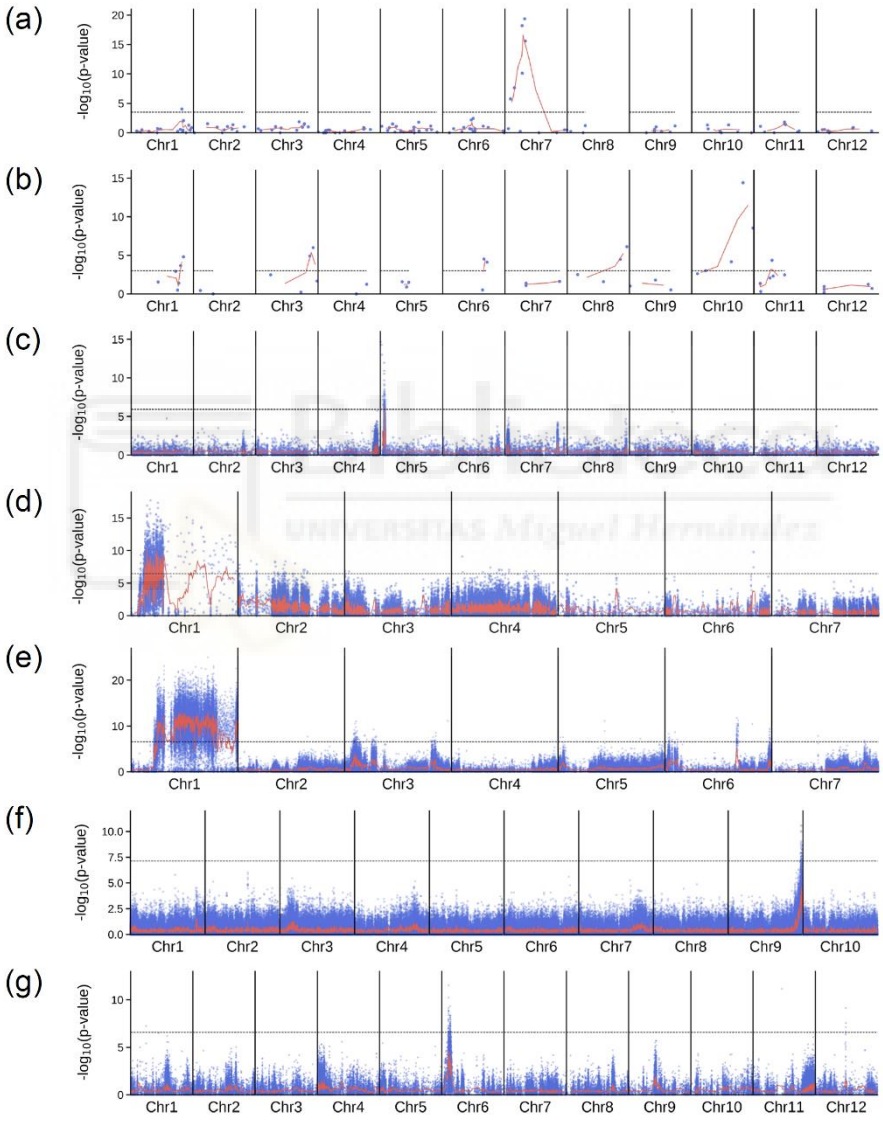



**Figure 2.** Mapping-by-sequencing in different species using MAPtools. The dots in each Manhattan plot correspond to the -log(*p*-value) of two-tailed Fisher's exact tests performed for individual biallelic markers segregating in the mapping population, as determined using data from the R and D bulks. The lines correspond to the weighted moving averages calculated for a sliding window containing *n* adjacent markers. The dashed line marks the significance threshold

calculated using the Bonferroni correction. **(a)** Mapping of the *lcd1* mutation of rice; n = 3. **(b)** Mapping of a suppressor of the *xantha* mutant of rice; n = 2. **(c)** Mapping of an ascorbate-enriched mutant of tomato; n = 20. **(d)** Mapping of the *green petiole-1* mutant of strawberry; n = 20. **(e)** Mapping of a white fruit mutant of strawberry; n = 100. **(f)** Mapping of a glossy mutant of Chinese cabbage; n = 100. **(g)** Mapping QTL for a rice blast disease; n = 100.

**Case study 1: The *lcd1* mutant of rice**

The recessive *lcd1* (*low Cd accumulation 1*) mutant was induced by EMS mutagenesis from 9311, an *indica* rice strain for which a reference genome sequence is available [31]. To map the gene, the authors used an $F_2$ population derived from a backcross between *lcd1* and its wild-type progenitor, 9311, and used the Illumina platform to sequence four samples: the *lcd1* and 9311 parents of the cross, a bulk comprising the 31 $F_2$ plants with the lowest Cd levels (presumably *lcd1* mutants), and another bulk comprising the 31 $F_2$ plants with the highest levels. The known effects of EMS and the fact that both parents of the cross have been sequenced, allowing the alleles to be classified according to their parental origin, make this an ideal case study for testing MAPtools. We run the MAPtools `mbs` command with the `-d R,D,Pr,Pd` option, which instructs the program to use the sequence data from the parental samples (Pr and Pd) to select the variants that will be analyzed in the bulks of phenotypically dominant (D) and recessive (R) plants. An advantage of resequencing of the parents (Pr and Pd) is that each allele can be assigned to its parental haplotype. Alternatively, this assignment could have been made by selecting the `-r D` option, since the reference genome sequence corresponds to one of the parents of the mapping population (i.e. 9311). Our analysis of the original data allowed us to locate the gene on chromosome 7, as indicated by the significant *p*-values of Fisher's exact tests (Figure 2a) and all other parameters used to compare the distribution of alleles in the D and R bulks (Figure 3 and Supplemental Figure 1). By applying the `annotate` command to a 5-Mb candidate interval on chromosome 7, we were able to detect the same C-to-T transition mutation in exon 7 of the Os07g0257200 gene as previously described [31]. While these authors reported that the mutation substitutes a Leu residue for Pro in the OsNRAMP5 protein, MAPtools additionally indicated that the *lcd1* mutation damages a 5' splice site and, therefore, it may disrupt the protein function to a greater extent than originally thought.

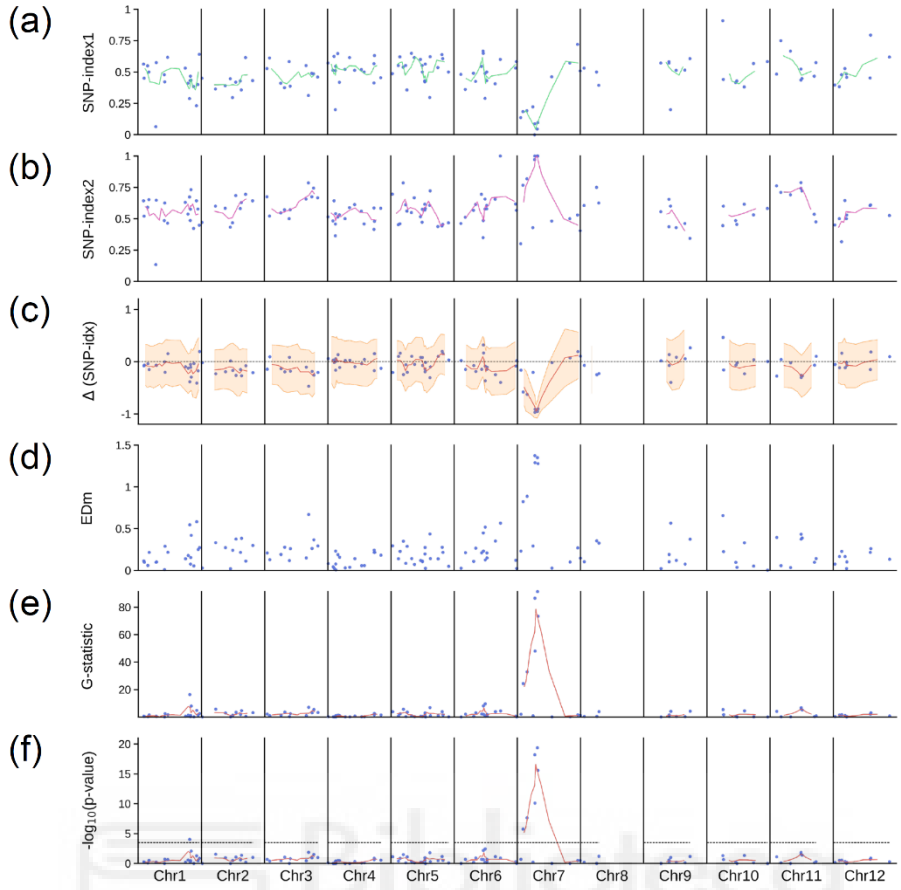



**Figure 3.** Mapping-by-sequencing of the *lcd1* mutant of rice. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Weighted moving averages (continuous lines) have been calculated for each statistic using a sliding window containing 3 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=151 tests). **(d)** Euclidean distance. **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=151 chromosomal locations have been tested.

**Case study 2: A suppressor of a *xantha* mutant of rice**

A similar approach was followed to characterize a mutation that suppresses the effects of a *xantha* (yellow-leaf) mutation in rice [32]. To generate a mapping population, the suppressor was backcrossed to the *xantha* mutant. Rather than focusing on the $F_2$ generation, the authors bulked $F_3$ plants from the generation that were known to be homozygous for either the mutant allele or the wild-type allele. In addition to the parents of the cross, two bulks were sequenced, one consisting of 20 $F_{2:3}$ plants exhibiting the *xantha* phenotype and another consisting of 20 $F_{2:3}$ plants with green leaves. Although MAPtools is not specifically designed to work with bulks of $F_3$ plants, it allowed us to successfully locate the gene on chromosome 10 (Figure 2b). The G-to-A transition reported in the article [32] was the only mutation detected by the `annotate` command within a 3 Mb candidate interval. The command reported its effect on three isoforms of the Os10g0502400 (OsGluTR) gene. In one isoform, the mutation is predicted to reside in the 5' untranslated region (5' UTR) and its effect is uncertain. In the other isoforms, however, the mutation is located in the CDS and is predicted to substitute a Val residue for Ala in the corresponding protein products. Interestingly, this was the only position reported by the program for which the two bulks were homozygous, albeit for different alleles, as expected from experimental design (Supplemental Figures 2 and 3).

**Case study 3: An ascorbate-enriched mutant of tomato**

We also analyzed raw data corresponding to a recessive mutant with ascorbate-enriched fruits that had been isolated from the tomato cultivar Micro-Tom after EMS mutagenesis [33]. The data from this study allowed us to test MAPtools when the genome of one parent has been resequenced (i.e. Micro-Tom) in addition to the D and R bulks. Running the `mbs` command with the `-d D,R,Pd` option allowed us to place the gene on chromosome 5 (Figure 2c and Supplemental Figures 4 and 5). Using `annotate` command, we found a C-to-T transition mutation in the gene Solyc05g007020, which encodes a member of the PAS/LOV protein (PLP) class of photoreceptors. The authors correctly reported that this mutation creates a premature stop codon [33], but the `annotate` command additionally reported that it damages a 5' splice site and is therefore likely to alter the splicing of its transcripts.

**Case study 4: The *nhm3-1* mutant of Chinese cabbage**

A different approach was followed to characterize a recessive *non-heading mutant* (*nhm3-1*), which had been induced by EMS from FT, a wild-type strain of Chinese

cabbage (*Brassica rapa* ssp. *pekinensis*) [30]. The authors prepared an $F_2$ mapping population, but they chose to sequence only the bulk of $F_2$ mutant plants (R bulk) and the two parents of the population (FT and the *nhm3-1* mutant). To handle this situation, we run MAPtools `mbs` with the `-d R,Pd,Pr` option. In the absence of a D bulk, we rely entirely on the allele frequencies to map and identify the mutation, as MAPtools cannot perform any calculations that require the two bulks (e.g. Fisher's exact tests, G statistics or Euclidean distances). The allele frequencies in the R bulk indicate that the *nhm3-1* mutation is most likely located on chromosome 5 (Figure 4). We used the `annotate` command to evaluate the effect of the 20 EMS-type nucleotide substitutions detected in a wide candidate interval (between megabases 3 and 10) on this chromosome. Interestingly, all the substitutions were of the same type (G-to-A transitions, with no C-to-T transitions detected in the interval), a bias that is a known consequence of the mutagenic action of EMS on the same DNA strand [37]. One of the transition mutations found by the `annotate` command is predicted to cause a Gly-to-Glu substitution in the protein encoded by the *KAO2* (*ent*-kaurenoic acid oxidase 2) gene, also known as A05p015130.1_BraROA.1. This mutation was previously considered to be the most likely cause of the observed phenotype [30]. However, the candidate interval was found to contain a second nonsynonymous mutation with an allele frequency of 1 in the A05p016250.1_BraROA.1 gene, which is predicted to cause a Gly-to-Asp substitution in the protein.

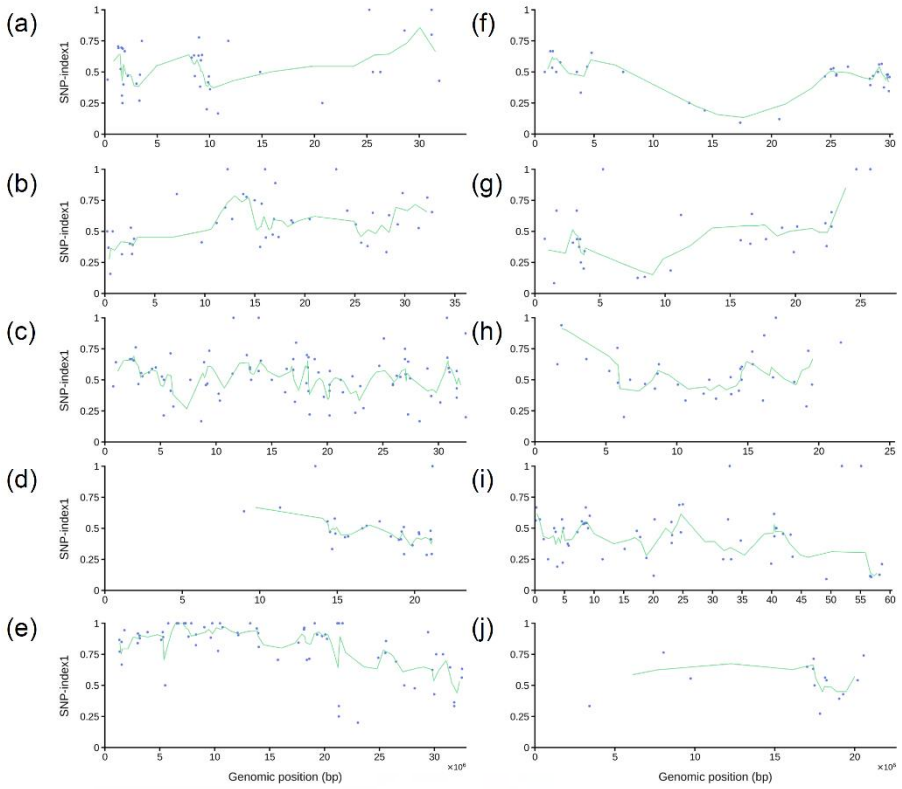



**Figure 4.** Allele frequencies place the *nhm3-1* mutation on chromosome 5 of Chinese cabbage. Each dot indicates the allele frequency of a biallelic polymorphism segregating in the population, as determined for the R bulk. The light green line indicates the moving average of the allele frequencies at 3 adjacent sites. **(a)** Chromosome A01. **(b)** Chromosome A02. **(c)** Chromosome A03. **(d)** Chromosome A04. **(e)** Chromosome A05. **(f)** Chromosome A06. **(g)** Chromosome A07. **(h)** Chromosome A08. **(i)** Chromosome A09. **(j)** Chromosome A10.

### Case study 5: Three *eop* mutants of *Arabis alpina*

Viñegra de la Torre et al. [36] followed a mapping-by-sequencing strategy to characterize three EMS-induced alleles (*eop002*, *eop085* and *eop091*) of the *ENHANCERS OF PEP1* (*EOP*) gene of *Arabis alpina*. The authors prepared mapping populations for the three mutants and sequenced one bulk of $F_2$ mutant plants for each one, plus an additional sample of the *pep1-1* parent of the three populations. All four samples were sequenced using Illumina technology, and the reads were mapped to the *A. alpina* reference genome. We used the MAPtools `mbs` command with the `-d R,Pd` option to map the three allelic mutations separately. The allele frequencies in the R bulk of each mapping population clearly indicated that the mutations reside on chromosome 8 (Figure 5 and Supplemental Figure 6). Using `annotate`, we were able to identify three nonsynonymous mutations affecting the same gene (Aa_G106560) on this chromosome, as had been previously reported.

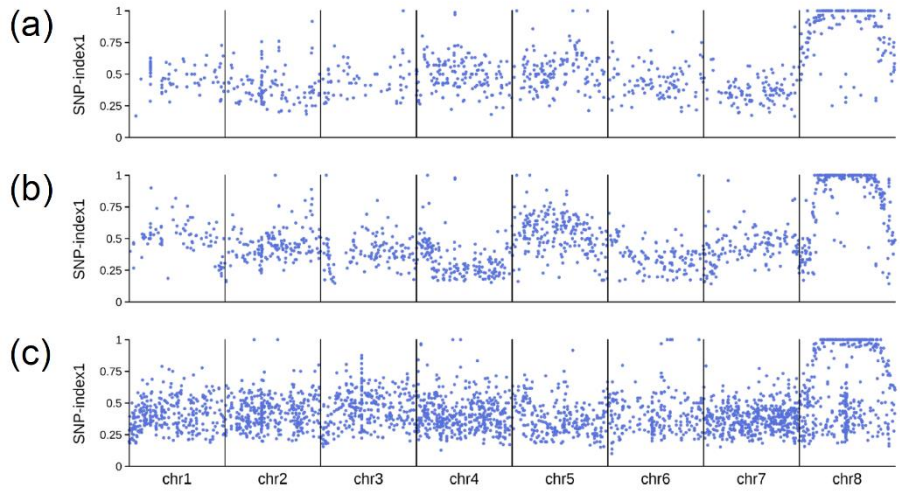



**Figure 5.** Mapping-by-sequencing of three recessive *eop* mutants of *Arabis alpina*. Each dot corresponds to an individual biallelic marker segregating in the population. Continuous green lines represent weighted moving averages calculated using a sliding window containing 5 adjacent markers. **(a)** *eop002* mutant. **(b)** *eop085* mutant. **(c)** *eop091* mutant.

### Case study 6: An albino mutant of *Arabidopsis thaliana*

To test MAPtools in the context of mapping populations derived from crosses involving widely divergent genetic backgrounds, we have also reanalyzed previously published data from several articles, including our own. We recently described the cloning of an albino mutation in *Arabidopsis thaliana* using mapping by sequencing [28]. In this particular case, we had sequenced two bulks of plants from the $F_2$ generation, one consisting of 170 phenotypically wild-type plants and another consisting of 87 phenotypically mutant plants. To analyze the raw data from this experiment, we used the MAPtools `mbs` command with the `-d D,R` option, which allowed us to locate the mutation on chromosome 2 (Figure 6a and Supplemental Figure 7 and 8). The resulting plots revealed that the genomes of the two parental strains consisted of regions that differed in the abundance of sequence polymorphisms. To exclude additional polymorphisms that were unlikely to cause the observed phenotype, we reran the command with the `--EMS` and `-I` flags, which discard substitutions that are unlikely to have been caused by EMS as well as all insertion-deletion mutations: This second analysis was performed with the `-d D,R,Wr` and `--parental-filter` options. To this end, we added sequencing data for wild-type plants of the Landsberg *erecta* background, which shares polymorphisms with the mutant parent of the

mapping population. These filters greatly reduced the number of candidate mutations (Figure 6b) and facilitated the identification of the same causal mutation as previously reported [28]. To illustrate the effect of the `merge` command, we applied it to the unfiltered dataset, which resulted in a more pronounced peak in the Manhattan plots (Figure 6c).

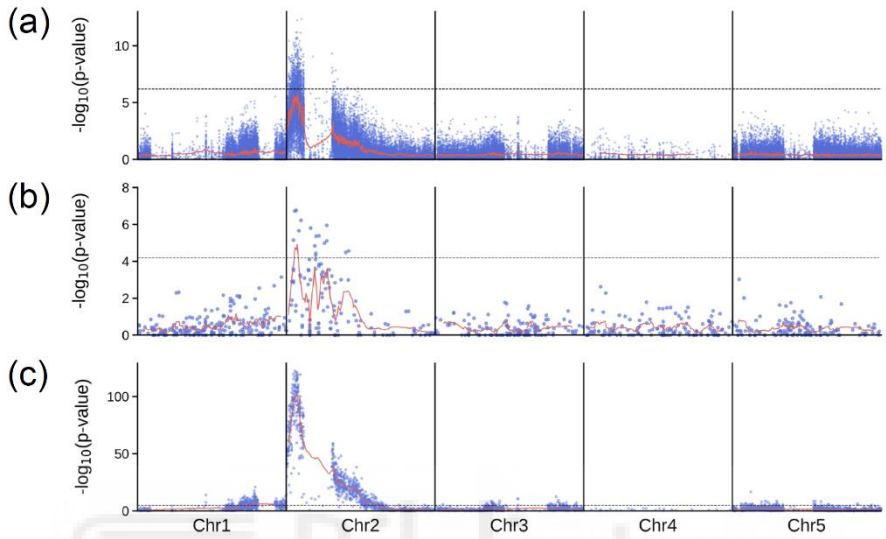



**Figure 6.** Mapping-by-sequencing of an albino mutant of *A. thaliana*. Each dot in the Manhattan plots corresponds to the -log(*p*-value) of a two-tailed Fisher's exact test performed for an individual biallelic marker segregating in the population (a and b) or for haplotypes integrating the allele counts of 20 adjacent markers (c), as determined using data from the R and D bulks. The panels illustrate the effect of running the mbs and merge commands of MAPtools on the same dataset with different options. **(a)** mbs -d D,R -m R -r D, with no filters applied. **(b)** mbs -d D,R,Wr -r D -m R --EMS -I --parental-filter, which excludes indels (-I), all mutations other than G/A-to-C/T transitions (--EMS), and all changes already present in the Wr sample (--parental-filter). **(c)** mbs -d D,R -m R -r D, followed by merge -w 20, which combines the allele counts in sets of 20 consecutive markers. The lines correspond to the weighted moving averages calculated for a sliding window containing 200 markers (a), 5 markers (b) and 10 haplotypes (c). The dashed lines mark the Bonferroni-corrected 5% significance thresholds, assuming that n=291,824 (a), n=861 (b) and n=14,595 tests have been performed.

### Case study 7: The *green petioles-1* mutant of strawberry

In addition to our own data, we have also tested the program with data from outcrosses involving plant species other than *Arabidopsis thaliana*. A recent study has characterized the *green petioles-1* (*gp-1*) mutant of diploid strawberry (*Fragaria vesca*), which was isolated after *N*-ethyl-*N*-nitrosourea (ENU) mutagenesis of the Yellow

Wonder 5AF7 (YW) accession [35]. The petioles of wild-type plants are purple whereas those of *gp-1* mutants are green. An $F_2$ population derived from an outcross between the *gp-1* mutant and the Hawaii (H4) wild-type accession was used for mapping-by-sequencing. Two bulks were sequenced: one consisting of $F_2$ plants with green petioles, and another consisting of $F_2$ plants with purple petioles. Since the genome sequence of the H4 accession is available and was used as the reference for mapping the reads, MAPtools can easily assign the alleles to the two parents of the outcross. With the `-d R,D -r D -m R` options, the allele frequency and *p*-value plots generated by MAPtools suggested that the mutation resides on chromosome 1 (Figure 2d and Supplemental Figures 9 and 10). Using the `annotate` command, we identified the same C-to-T substitution on chromosome 1 as previously described [35]. Based on the annotation provided in the GFF file, MAPtools determined that it is located one nucleotide away from a 3' splice site within one of the introns of the FveMYB10L gene (FvH4_1g22040.1). Luo et al. [35], however, corrected the annotation of this gene so that the mutation is located in exon 3, resulting in the substitution of a lysine reside for an arginine in the protein.

**Case study 8: A white fruit mutant of strawberry**

Castillejo et al. [34] mapped a mutation causing white fruits using two bulks from the $F_2$ progeny of an outcross involving two lines of strawberry (*Fragaria vesca*), RV660 (with red fruits) and WV596 (with white fruits). The bulks comprised 34 plants with red fruits and 32 plants with white fruits, respectively. The reads were aligned to the Hawaii-4 (H4) reference genome and, because the parents of the mapping population were not sequenced, their alleles cannot be uniquely assigned to either RV660 or WV596. Using the ΔSNP index, the authors defined a candidate interval on chromosome 1, where they identified a transposon insertion in the *FvMYB10* gene (FvH4_1g22020). Our analysis of the raw data using the MAPtools `mbs` command clearly showed that the mutation responsible for the white color resides on chromosome 1, as evidenced by the low *p*-values of Fisher's exact tests (Figure 2e) and other parameters considered, such as the frequency of the most abundant allele (Supplemental Figures 11 and 12). Although the `annotate` command is not designed to detect large insertions, like the transposon reported by Castillejo et al. (2020), it identified other polymorphisms with highly skewed allele counts in the coding sequence of the same gene (i.e. a G-to-A transition mutation at position 13,950,746).

**Case study 9: A glossy mutant of Chinese cabbage**

In another study with Chinese cabbage [29], the genetic basis of the recessive glossy

phenotype of a line was characterized. To identify the responsible locus, a BSA-seq strategy was followed with an $F_2$ mapping population derived from the cross between the glossy line (Y1211-1) and a double haploid line (R16-11). Two pools of $F_2$ plants were sequenced: one consisting of 25 glossy plants (the G pool) and one consisting of 25 waxy plants (the W pool). The authors aligned the reads to the *B. rapa* v1.5 genome, which did not match either parent of the cross. By examining Δ(SNP index) values, they found a candidate gene, Bra032670, on chromosome A09, in a candidate interval between megabases 37.35 and 38.88. This study is based on an isogenic cross, but the parents are not available to sort the alleles (neither reference sequence nor additional sequenced pools). Using MAPtools, a maximum of allele frequencies and a minimum of *p*-value in the same chromosome as indicated by the authors of the paper are clearly detected (Figure 2f and Supplemental Figures 13 and 14).

**Case study 10: Mapping QTL for rice blast disease**

Takagi et al. [6] mapped QTL conferring partial resistance to rice blast disease (*Magnaporthe oryzae*) using a population of recombinant inbred lines (RILs) established from a cross between the Nortai (partially resistant) and Hitomebore (highly susceptible) rice lines. The RILs were scored for resistance, and two bulks were made with 20 lines highly resistant and 20 lines highly susceptible to rice blast. The reads were mapped to the Hitomebore reference genome, which allows the assignment of alleles to each parent. Using MAPtools, we mapped a QTL to a relatively narrow region on chromosome 6, at the same location reported by Takagi et al. (2013), as indicated by the significant *p*-values of Fisher's exact tests (after Bonferroni correction) and the shift in the value of the ΔSNP indices (Figure 2g and Supplemental Figures 15 and 16).

**Conclusions**

Here, we present MAPtools, a novel program with a number of useful features for the analysis and visualization of mapping-by-sequencing and QTL-seq data. This command-line tool is implemented in the Python3 language, making it easy to install and use. We developed MAPtools inspired by the SAMtools and BCFtools packages, two important applications that we emulated in features such as the availability of distinct commands and the ability to receive input data through a command-line pipeline or properly formatted VCF files. The fact that MAPtools can process input received from the command line makes it a very versatile application that can be easily integrated into workflows with various state-of-the-art variant callers, such as BCFtools or GATK. Although the `mbs` and `qtl` commands of MAPtools primarily function with

VCF input data, this feature of the program effectively allows processing input data in BCF format by simply adding a conversion step (e.g. by using BCFtools' `view` command) to the workflow. Once a VCF (BCF) file is ready, it can be quickly processed any number of times by running MAPtools `mbs` or `qtl` with user-selected parameters to facilitate the identification of the causal mutation, or by adding additional steps to the workflow. These may include steps to filter out sites with certain types of mutations (e.g. indels), too low sequencing depth, or based on the values of other fields present in the VCF file format.

One of the unique features of MAPtools is its ability to use multiple criteria to determine the position of QTLs or genes of interest. In addition, our repertoire of commands enables the automatic generation of publication-quality figures and their captions, as well as the rapid assessment of the functional impact of identified genetic variants by using a genome annotation file in the GFF3 standard format. We have tested our software using raw data from species with genome sizes ranging from 135 Mbp (*A. thaliana*) to 828 Mbp (*S. lycopersicum*). MAPtools will work best for any species for which a high-quality genome sequence is available, particularly if it has been annotated using the standard GFF3 file format. To this end, the Ensembl Plants database turned out to be an ideal resource [38], as it includes genomes for over 100 plant species, which opens the door to a wide range of potential uses for MAPtools. While in most cases we mapped the reads to genomes downloaded from Ensembl Plants, we also obtained satisfactory results with genomes from different sources (e.g. for *F. vesca*). Importantly, the Ensembl database also includes the genomes of many different organisms to which the MBS or QTL-seq methodology could also be applied [39]. We plan to expand the repertoire of commands available in MAPtools to enable the analysis of data from other experimental scenarios involving massively parallel sequencing, such as the construction of high-resolution linkage maps.

**Funding**

This work has been supported by an internal grant of Universidad Miguel Hernández to H.C. (VIPROY21/1).

*Conflict of Interest:* none declared.

**Data availability**

The program and documentation are available from https://github.com/hcandela/MAPtools.

**Supplemental Materials**

## References


1. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, et al. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods. 2009;6:550–1.

2. Schneeberger K, Weigel D. Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. 2011;16:282–8.

3. Schneeberger K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet. 2014;15:662–76.

4. Candela H, Casanova-Sáez R, Micol JL. Getting started in mapping-by-sequencing. J Integr Plant Biol. 2015;57:606–12.

5. Zhu QianHao ZQ, Wilson I, Llewellyn D. Mapping-by-sequencing enabled fast forward genetics in crops with complex genomes. CABI Rev. 2017;1–12.

6. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 2013;74:174–83.

7. Yang L, Wang J, Han Z, Lei L, Liu HL, Zheng H, et al. Combining QTL-seq and linkage mapping to fine map a candidate gene in qCTS6 for cold tolerance at the seedling stage in rice. BMC Plant Biol. 2021;21:278.

8. Yan P, Li W, Zhou E, Xing Y, Li B, Liu J, et al. Integrating BSA-Seq with RNA-Seq Reveals a Novel Fasciated Ear5 Mutant in Maize. Int J Mol Sci. 2023;24.

9. Fekih R, Takagi H, Tamiru M, Abe A, Natsume S, Yaegashi H, et al. MutMap+: Genetic Mapping and Mutant Identification without Crossing in Rice. PLOS ONE. 2013;8:e68529.

10. Sun H, Schneeberger K. SHOREmap v3.0: Fast and Accurate Identification of Causal Mutations from Forward Genetic Screens. In: Alonso JM, Stepanova AN, editors. Plant Funct Genomics Methods Protoc [Internet]. New York, NY: Springer New York; 2015. p. 381–95. Available from: https://doi.org/10.1007/978-1-4939-2444-8_19

11. Pulido-Tamayo S, Duitama J, Marchal K. EXPLoRA-web: linkage analysis of quantitative trait loci using bulk segregant analysis. Nucleic Acids Res. 2016;44:W142–6.

12. Javorka P, Raxwal VK, Najvarek J, Riha K. artMAP: A user-friendly tool for mapping ethyl methanesulfonate-induced mutations in Arabidopsis. Plant Direct. 2019;3:e00146.

13. Li Z, Xu Y. Bulk segregation analysis in the NGS era: a review of its teenage years. Plant J. 2022;109:1355–74.

14. Etherington GJ, Monaghan J, Zipfel C, MacLean D. Mapping mutations in plant genomes with the user-friendly web application CandiSNP. Plant Methods. 2014;10:41.

15. Wachsman G, Modliszewski JL, Valdes M, Benfey PN. A SIMPLE Pipeline for Mapping Point Mutations. Plant Physiol. 2017;174:1307–13.

16. Lup SD, Navarro-Quiles C, Micol JL. Versatile mapping-by-sequencing with Easymap v.2. Front Plant Sci [Internet]. 2023;14. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2023.1042913

17. James GV, Patel V, Nordström KJ, Klasen JR, Salomé PA, Weigel D, et al. User guide for mapping-by-sequencing in Arabidopsis. Genome Biol. 2013;14:1–13.

18. Wilson-Sánchez D, Lup SD, Sarmiento-Mañús R, Ponce MR, Micol JL. Next-generation forward genetic screens: using simulated data to improve the design of mapping-by-sequencing experiments in Arabidopsis. Nucleic Acids Res. 2019;47:e140–e140.

19. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

20. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

25. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020.

26. de la Fuente Cantó C, Vigouroux Y. Evaluation of nine statistics to identify QTLs in bulk segregant analysis using next generation sequencing approaches. BMC Genomics. 2022;23:490.

27. Mateo-Bonmatí E, Casanova-Sáez R, Candela H, Micol JL. Rapid identification of angulata leaf mutations using next-generation sequencing. Planta. 2014;240:1113–22.

28. Rodríguez-Alcocer E, Ruiz-Pérez E, Parreño R, Martínez-Guardiola C, Berna JM, Çakmak Pehlivanlı A, et al. Cloning of an Albino Mutation of *Arabidopsis thaliana* Using Mapping-by-Sequencing. Int J Mol Sci. 2023;24.

29. Yang S, Liu H, Wei X, Zhao Y, Wang Z, Su H, et al. BrWAX2 plays an essential role in cuticular wax biosynthesis in Chinese cabbage (Brassica rapa L. ssp. pekinensis). Theor Appl Genet. 2022;135:693–707.

30. Huang S, Gao Y, Xue M, Xu J, Liao R, Shang S, et al. BrKAO2 mutations disrupt leafy head formation in Chinese cabbage (Brassica rapa L. ssp. pekinensis). Theor Appl Genet. 2022;135:2453–68.

31. Cao ZZ, Lin XY, Yang YJ, Guan MY, Xu P, Chen MX. Gene identification and transcriptome analysis of low cadmium accumulation rice mutant (lcd1) in response to cadmium stress using MutMap and RNA-seq. BMC Plant Biol. 2019;19:250.

32. Jiang M, Dai S, Zheng Y-C, Li R-Q, Tan Y-Y, Pan G, et al. An alanine to valine mutation of glutamyl-tRNA reductase enhances 5-aminolevulinic acid synthesis in rice. Theor Appl Genet. 2022;135:2817–31.

33. Bournonville C, Mori K, Deslous P, Decros G, Blomeier T, Mauxion J-P, et al. Blue light promotes ascorbate synthesis by deactivating the PAS/LOV photoreceptor that inhibits GDP-L-galactose phosphorylase. Plant Cell. 2023;35:2615–34.

34. Castillejo C, Waurich V, Wagner H, Ramos R, Oiza N, Muñoz P, et al. Allelic Variation of MYB10 Is the Major Force Controlling Natural Variation in Skin and Flesh Color in Strawberry (Fragaria spp.) Fruit. Plant Cell. 2020;32:3723–49.

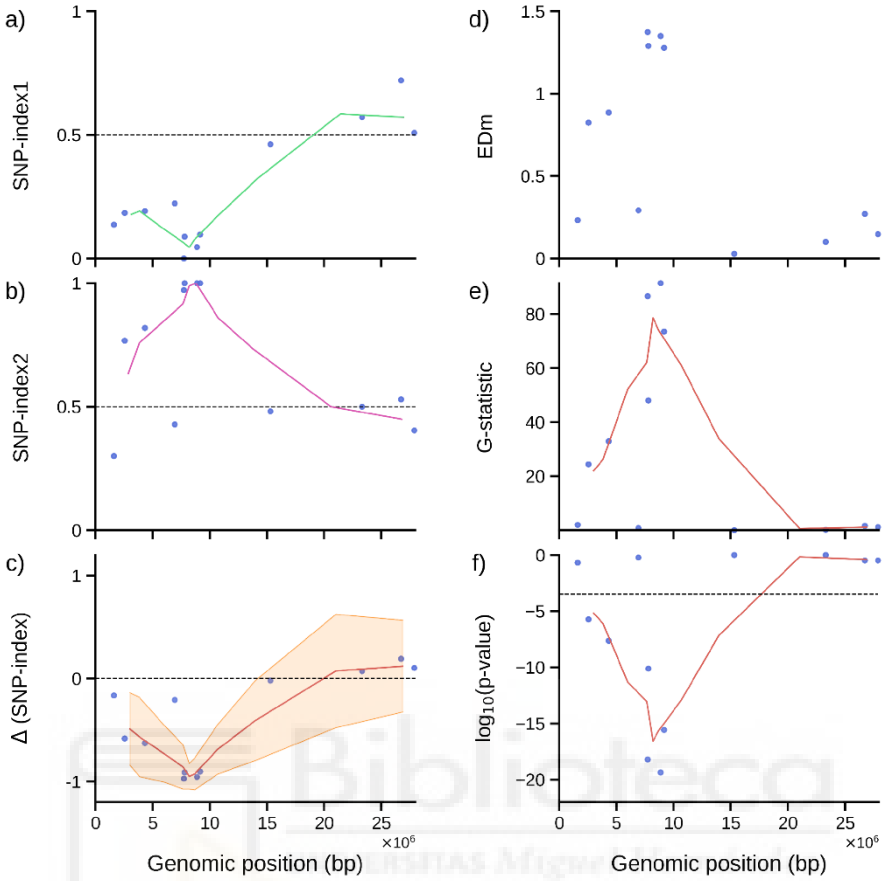
35. Luo X, Plunkert M, Teng Z, Mackenzie K, Guo L, Luo Y, et al. Two MYB activators of anthocyanin biosynthesis exhibit specialized activities in petiole and fruit of diploid strawberry. J Exp Bot. 2023;74:1517–31.

36. Viñegra de la Torre N, Vayssières A, Obeng-Hinneh E, Neumann U, Zhou Y, Lázaro A, et al. FLOWERING REPRESSOR AAA+ ATPase 1 is a novel regulator of perennial flowering in *Arabis alpina*. New Phytol. 2022;236:729–44.

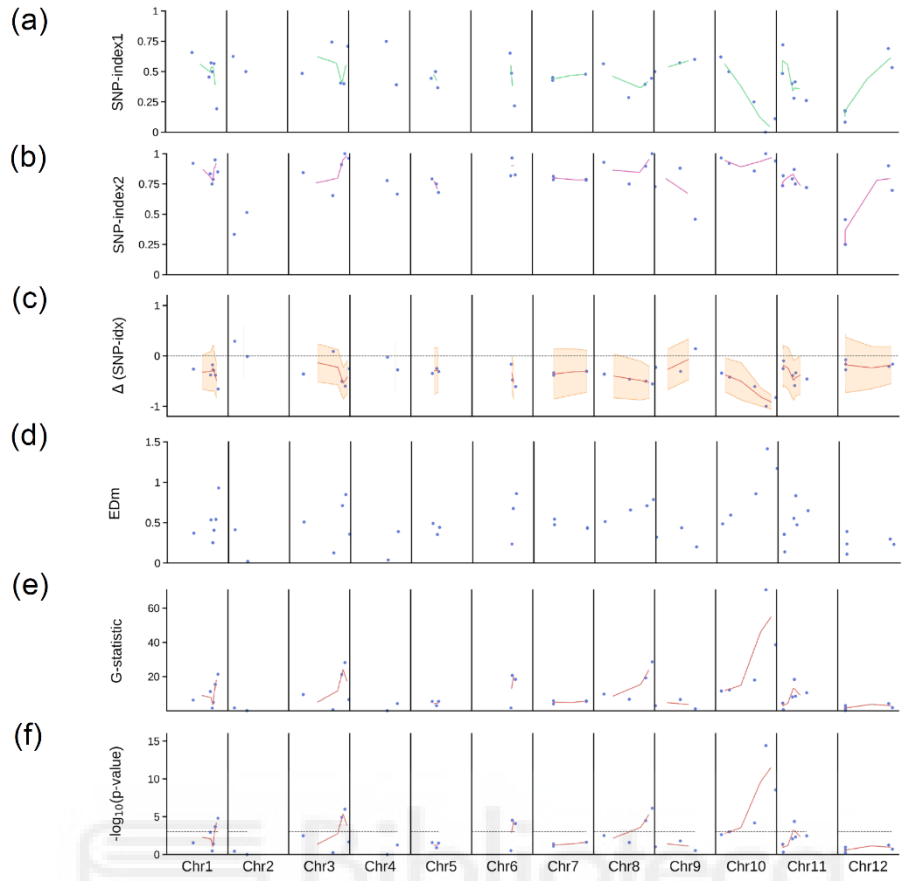
37. Candela H, Hake S. The art and design of genetic screens: maize. Nat Rev Genet. 2008;9:192–203.

38. Bolser DM, Staines DM, Perry E, Kersey PJ. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomic Data. In: van Dijk ADJ, editor. Plant Genomics Databases Methods Protoc [Internet]. New York, NY: Springer New York; 2017. p. 1–31. Available from: https://doi.org/10.1007/978-1-4939-6658-5_1

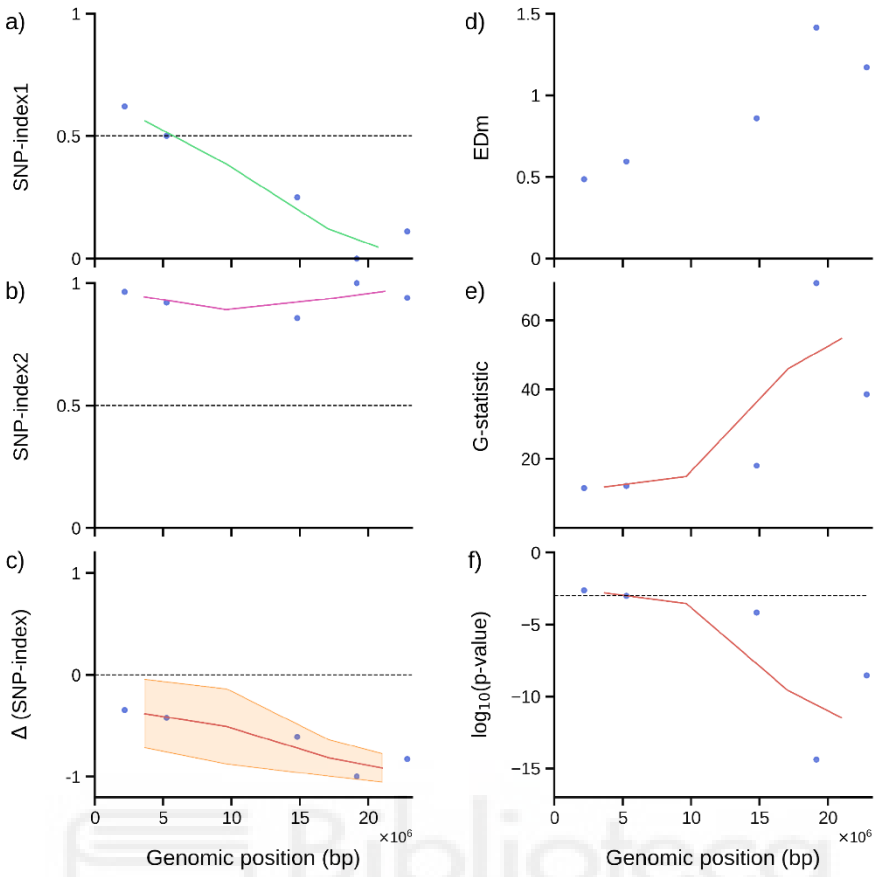
39. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res. 2024;52:D891–9.



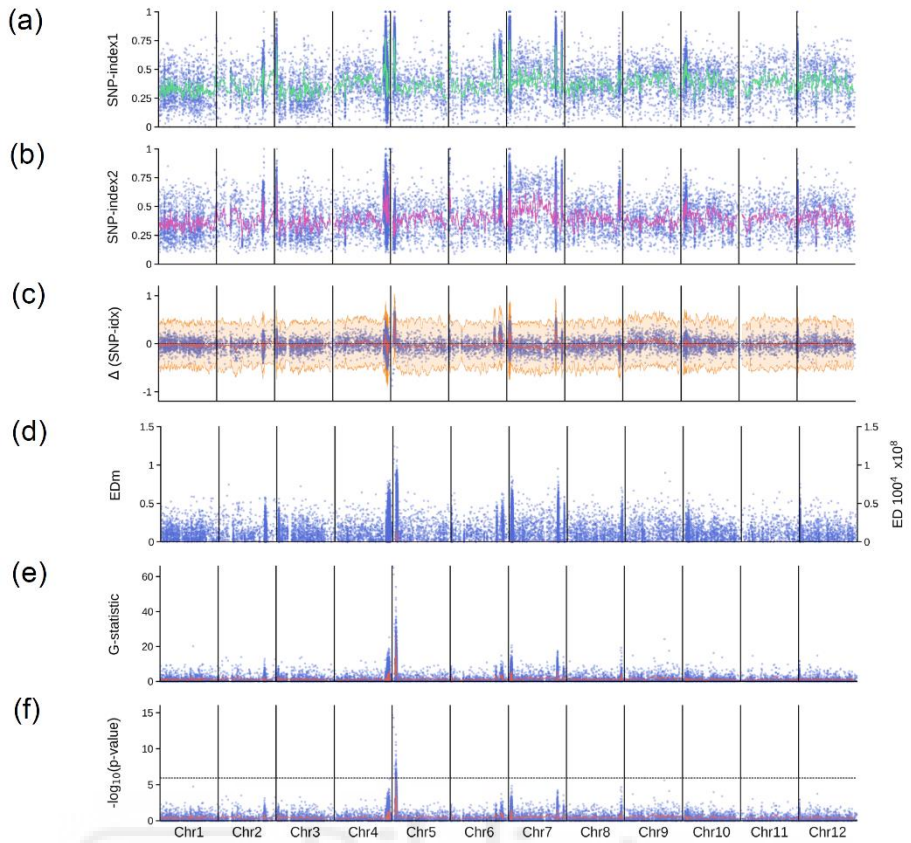

**Supplemental Figure 1.** Different statistics place the *lcd1* mutation on rice chromosome 7. Each dot corresponds to an individual biallelic marker segregating in the population. Weighted moving averages (continuous lines) have been calculated for each statistic using a sliding window containing 3 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=151 tests). **(d)** Euclidean distance. **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=151 chromosomal locations have been tested.



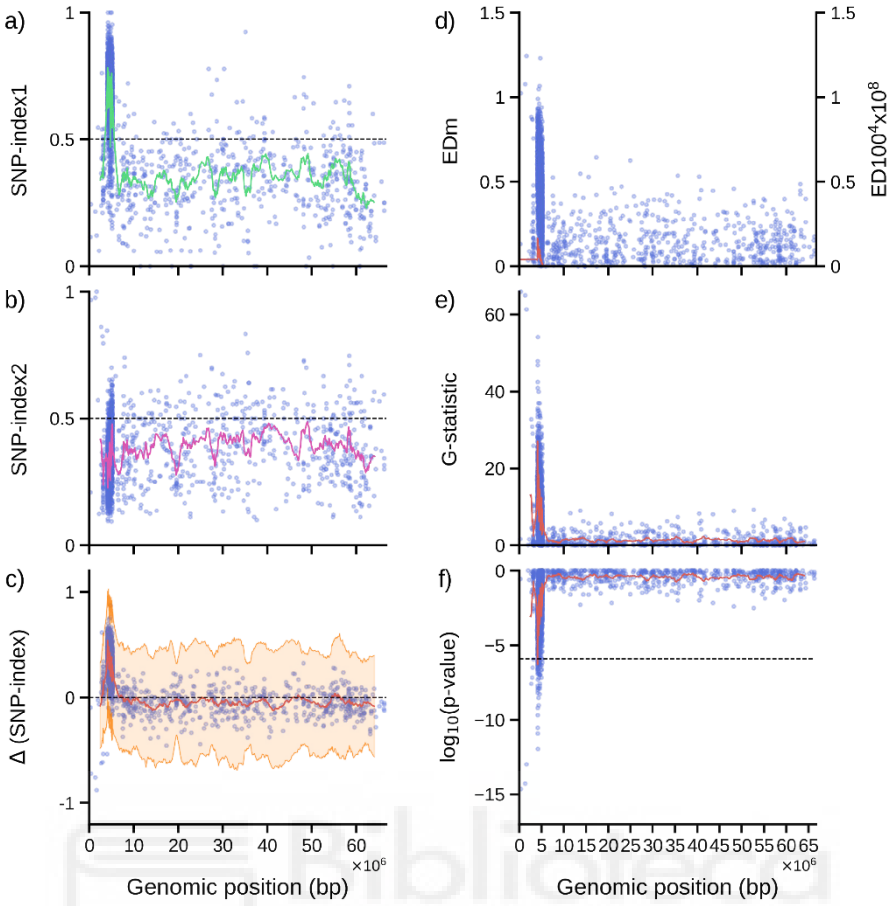

**Supplemental Figure 2.** Mapping-by-sequencing of a mutation that suppresses the *xantha* phenotype in rice. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 2 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=49 tests). **(d)** Euclidean distance. **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=49 chromosomal locations have been tested.



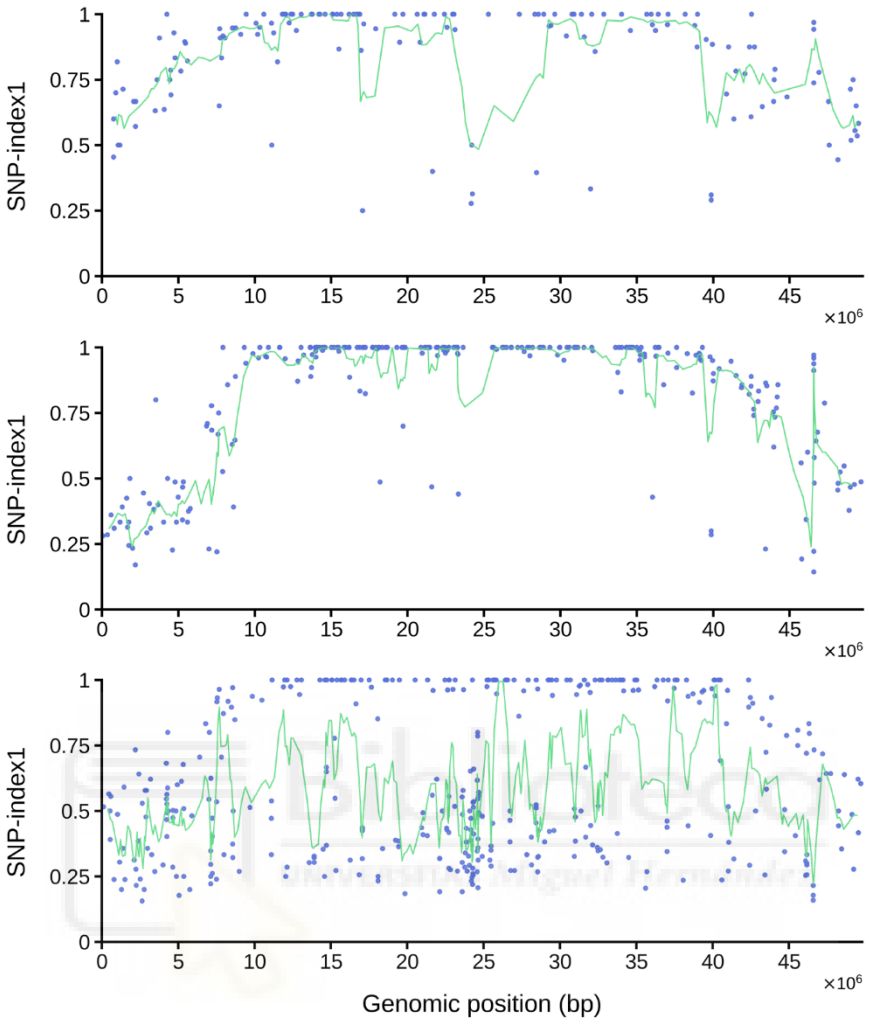

**Supplemental Figure 3.** Different statistics place a suppresor of *xantha* on rice chromosome 10. Each dot corresponds to an individual biallelic marker segregating in the population. Weighted moving averages (continuous lines) have been calculated for each statistic using a sliding window containing 2 adjacent markers. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 49 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=670,109 tests). **(d)** Euclidean distance. **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=49 chromosomal locations have been tested.



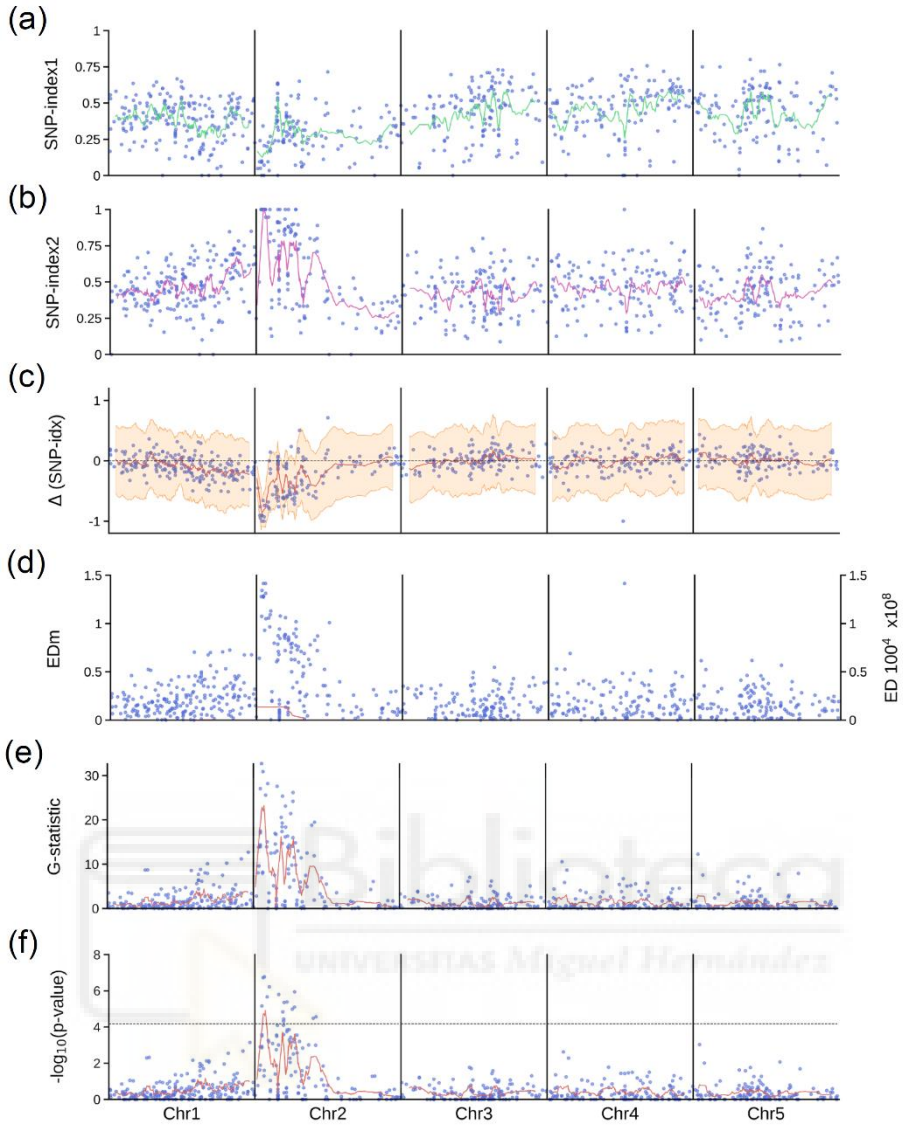

**Supplemental Figure 4.** Mapping-by-sequencing of an ascorbate-enriched mutant of tomato. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 20 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=41,415 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=41,415 chromosomal locations have been tested.




**Supplemental Figure 5.** Different statistics place an ascorbate-enriched mutation on tomato chromosome 5. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 20 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the absolute value of the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=223,711 tests). **(d)** Euclidean distance (dots) and ED100$^4$ (red line). ED100$^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=223,711 chromosomal locations have been tested.

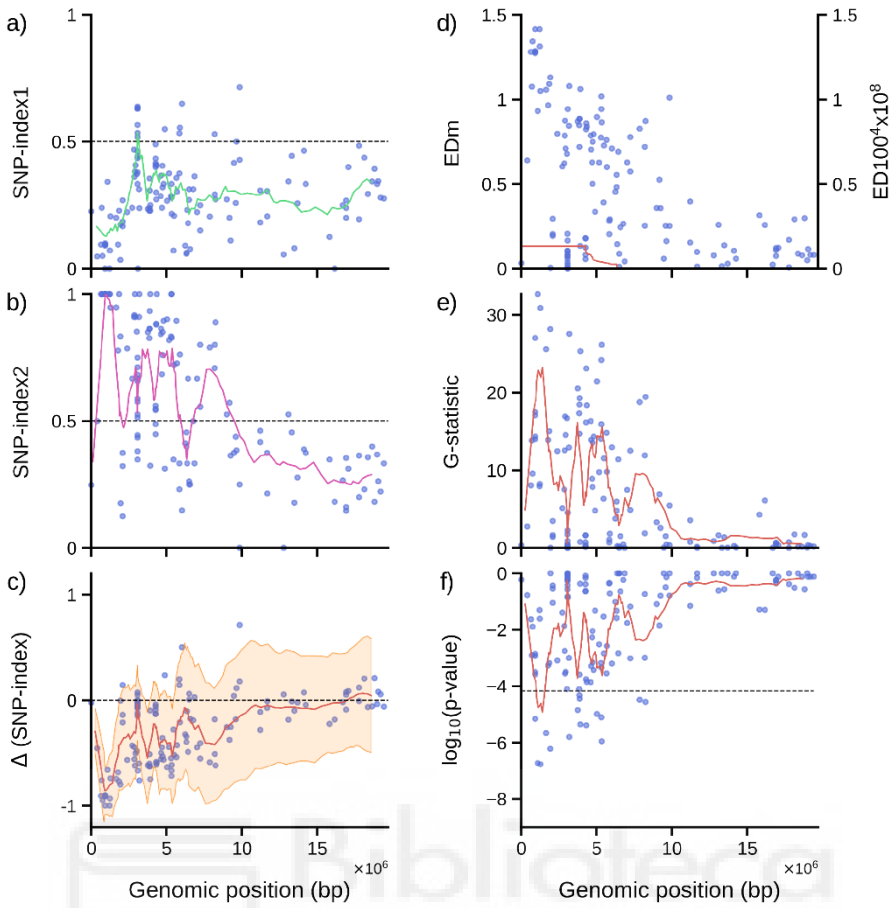



**Supplemental Figure 6.** Three recessive *eop* mutants of *Arabis alpina* map to chromosome 8. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 5 adjacent markers. **(a)** *eop002* mutant. **(b)** *eop085* mutant. **(c)** *eop091* mutant.

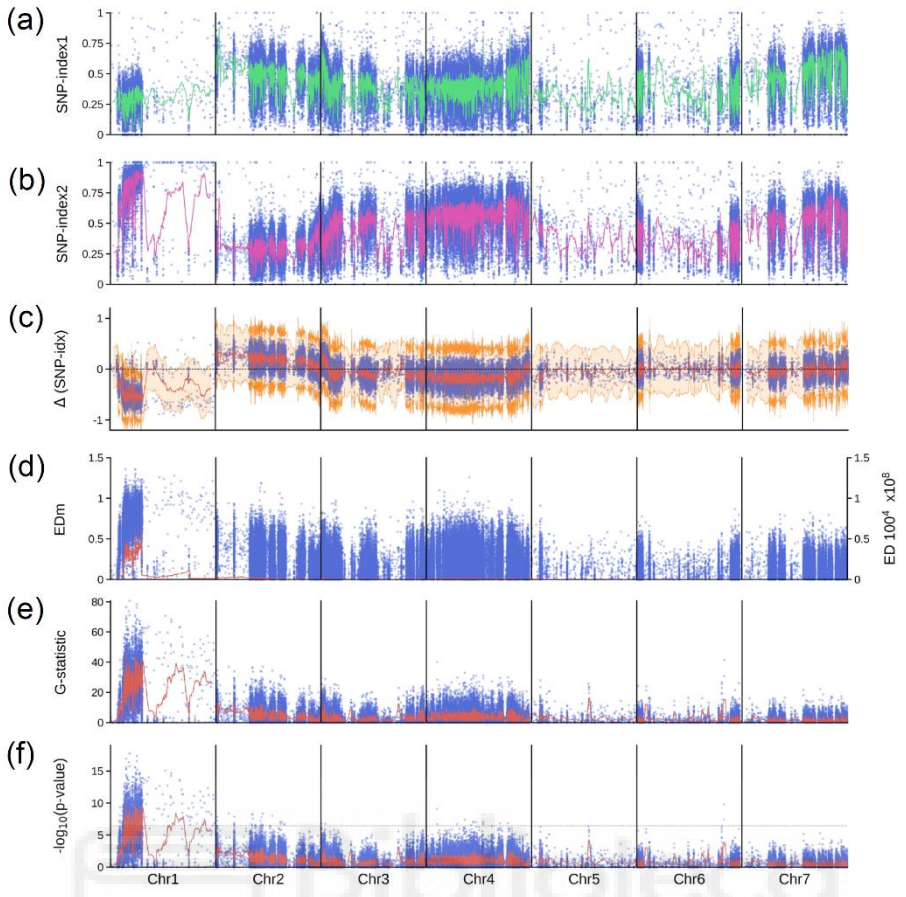



**Supplemental Figure 7.** Mapping-by-sequencing of an albino mutant of *Arabidopsis thaliana*. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Continuous lines represent weighted moving averages calculated using a sliding window containing 5 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the D bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=741 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5%

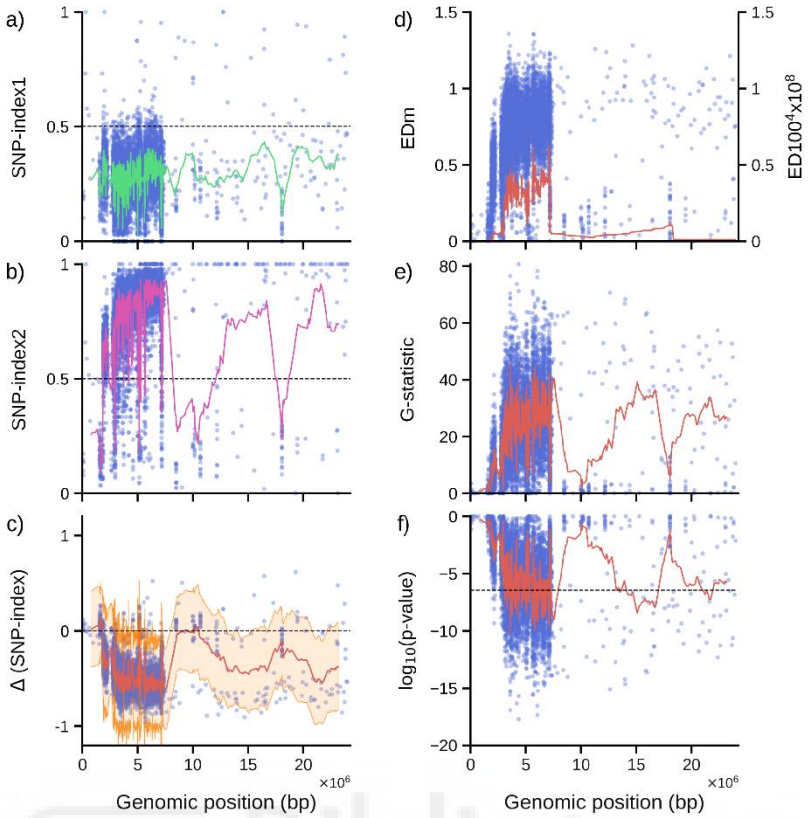
significance threshold, calculated considering that n=741 chromosomal locations have been tested.



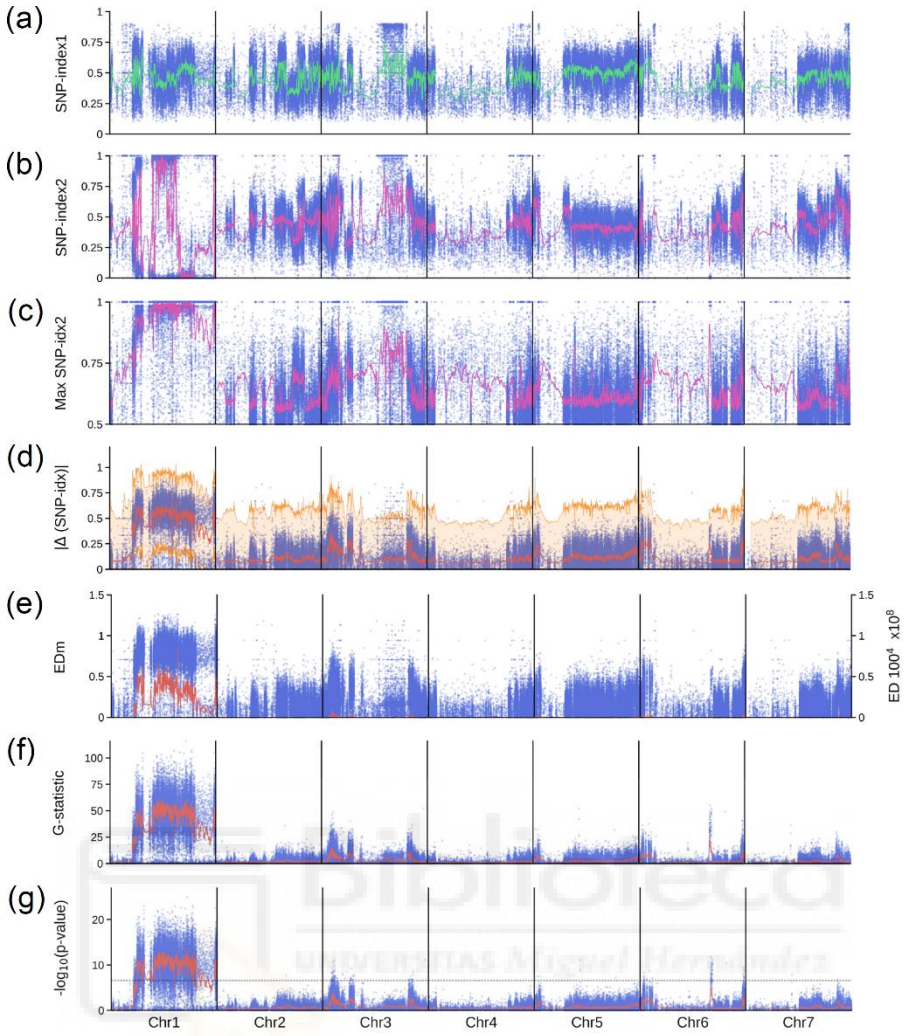

**Supplemental Figure 8.** Different statistics place an albino mutation on chromosome 2 of *Arabidopsis thaliana*. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 5 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=741 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=741 chromosomal locations have been tested.



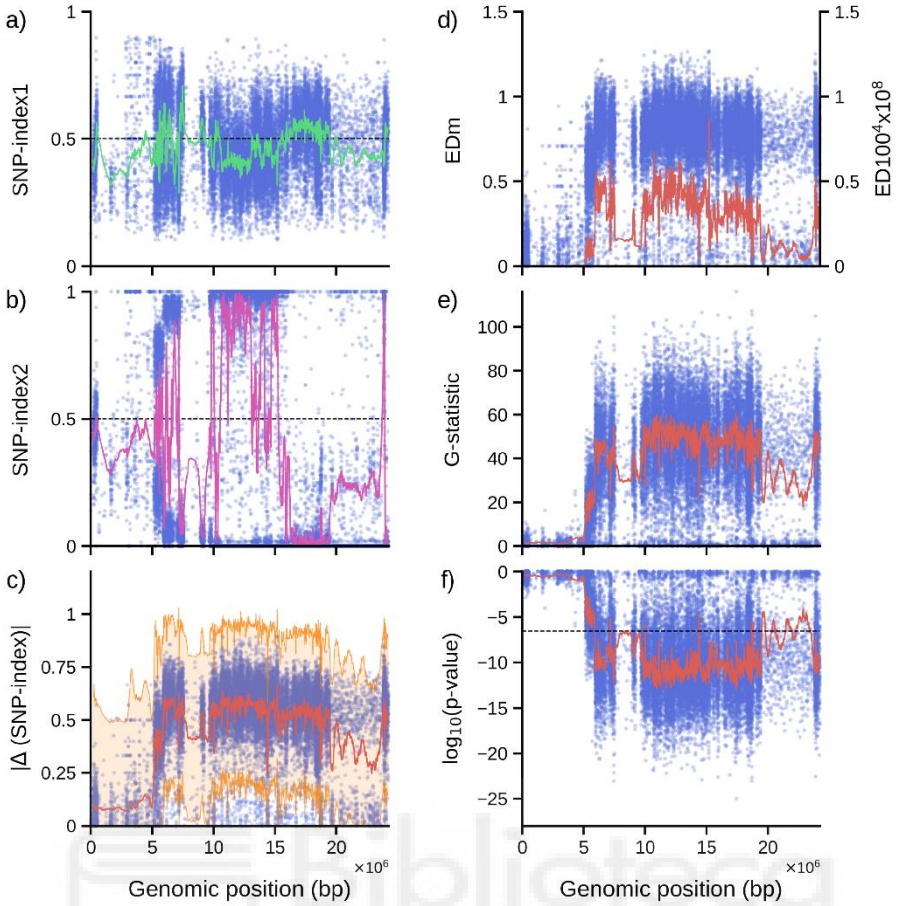

**Supplemental Figure 9.** Mapping-by-sequencing of the *green petioles-1* mutant of strawberry. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 20 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the D bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=139,639 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=139,639 chromosomal locations have been tested.



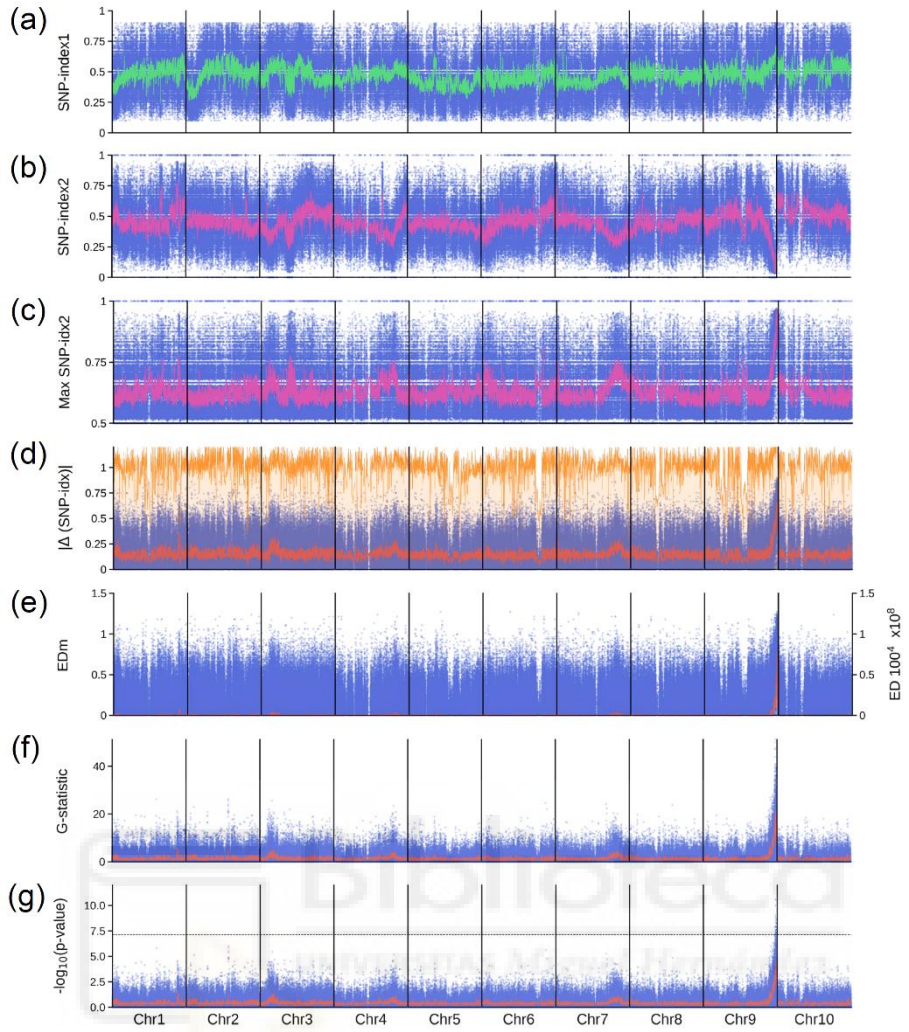

**Supplemental Figure 10.** Different statistics place the *green petioles-1* mutation on chromosome 1 of strawberry. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 20 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Δ(SNP-index), calculated as the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=139,639 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=139,639 chromosomal locations have been tested.



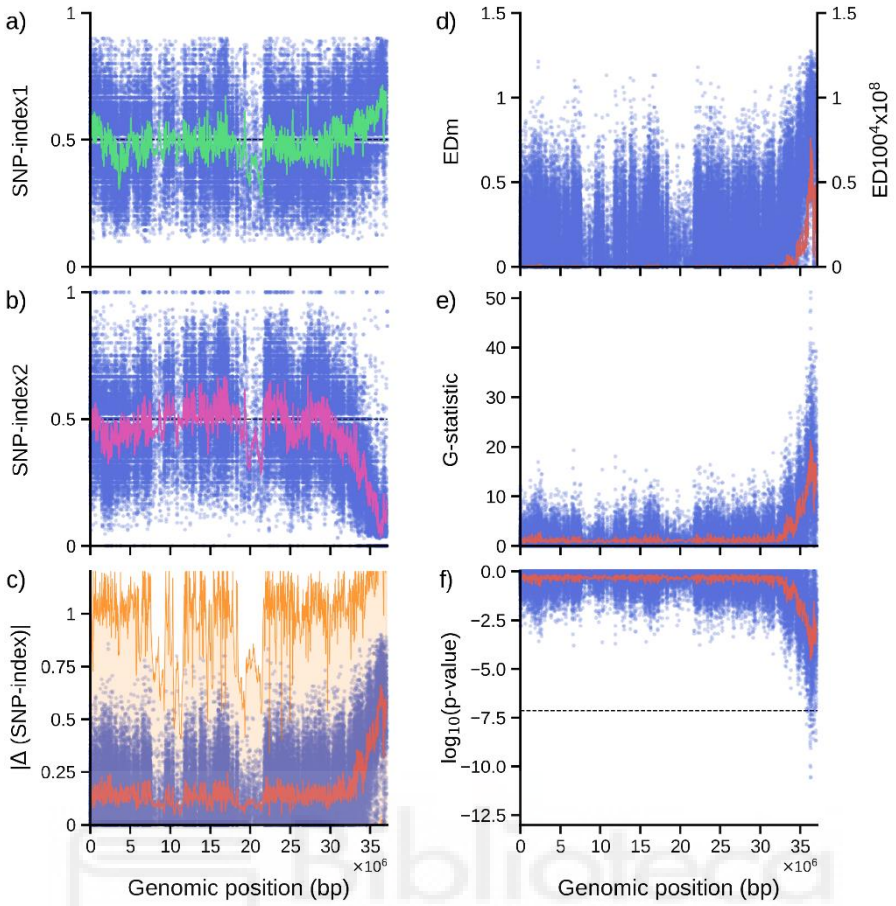

**Supplemental Figure 11.** Mapping-by-sequencing of a white fruit mutant of strawberry. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Allele frequency of the most abundant allele in the R pool. **(d)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=183,155 tests). **(e)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(f)** G-statistic, calculated as described by Magwene et al. (2011). **(g)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=183,155 chromosomal locations have been tested.



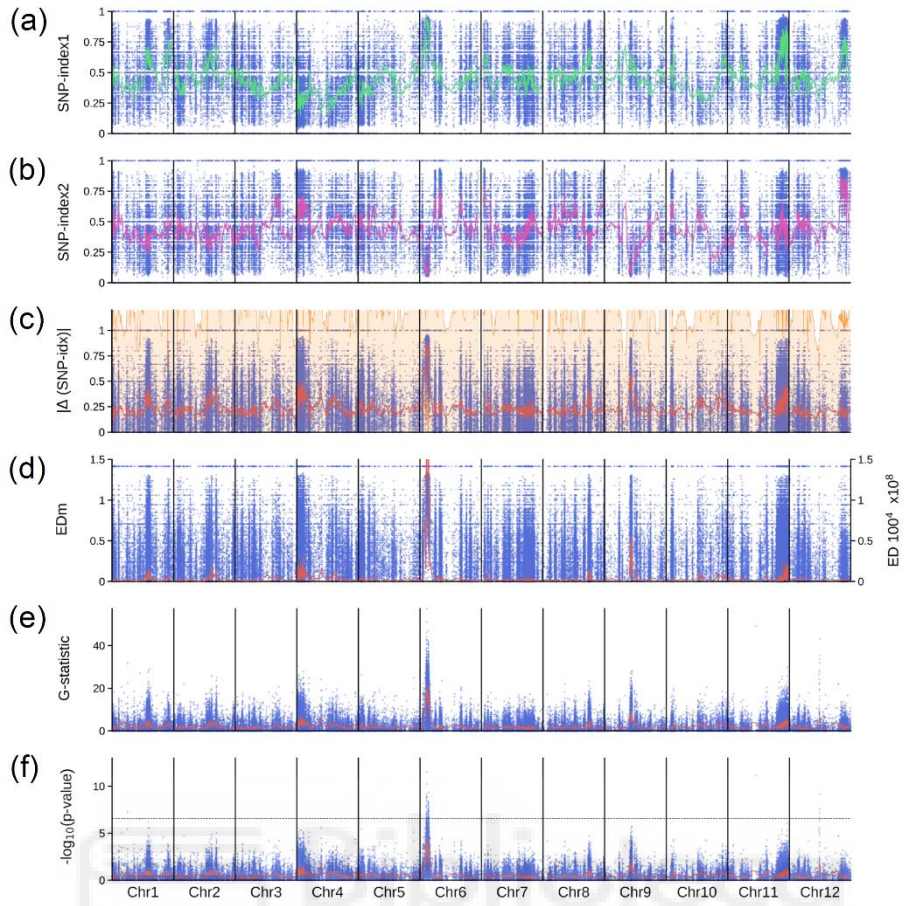

**Supplemental Figure 12.** Different statistics place a white fruit mutation on chromosome 1 of strawberry. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=183,155 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=183,155 chromosomal locations have been tested.



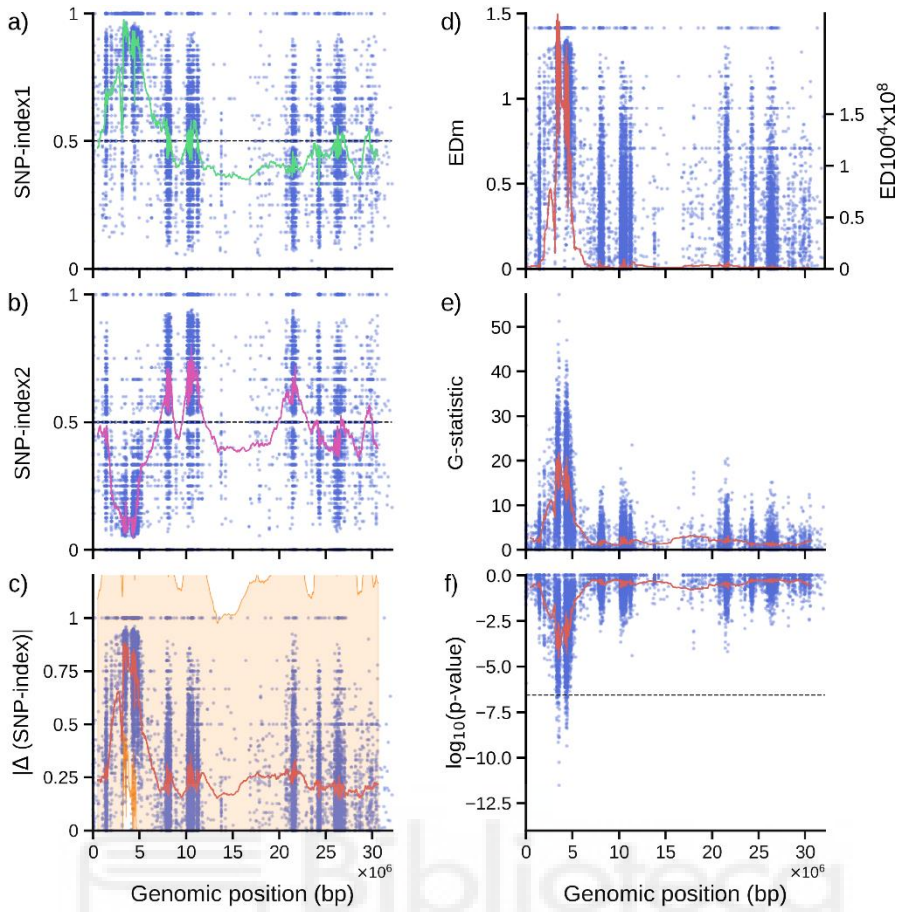

**Supplemental Figure 13.** Mapping-by-sequencing of a glossy mutant of Chinese cabbage. Several statistics have been evaluated across all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** Allele frequency of the most abundant allele in the R pool. **(d)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=670,109 tests). **(e)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(f)** G-statistic, calculated as described by Magwene et al. (2011). **(g)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=670,109 chromosomal locations have been tested.




**Supplemental Figure 14.** Different statistics place a glossy mutation on chromosome 9 of Chinese cabbage. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the D bulk. **(b)** SNP-index in the R bulk. **(c)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the D bulk and the SNP-index of the R bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=670,109 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=670,109 chromosomal locations have been tested.




**Supplemental Figure 15.** QTL-seq of blast disease resistance in rice. Several statistics have been evaluated along all chromosomes, and the results are presented as Manhattan plots. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the H bulk. **(b)** SNP-index in the L bulk. **(c)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the H bulk and the SNP-index of the L bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=187,181 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** -log(*p*-value) of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=187,181 chromosomal locations have been tested.




**Supplemental Figure 16.** Different statistics place a QTL for blast disease resistance on chromosome 6 of rice. Each dot corresponds to an individual biallelic marker segregating in the population. Unless otherwise stated, continuous lines represent weighted moving averages calculated using a sliding window containing 100 adjacent markers. **(a)** SNP-index (allele frequency) in the H bulk. **(b)** SNP-index in the L bulk. **(c)** |Δ(SNP-index)|, calculated as the absolute value of the difference between the SNP-index of the H bulk and the SNP-index of the L bulk. The shaded area is delimited by the moving averages of the lower and upper bounds of 95% confidence intervals, using the Bonferroni correction for multiple testing (with n=187,181 tests). **(d)** Euclidean distance (dots) and $ED100^4$ (red line). $ED100^4$ values were calculated as described by de la Fuente Cantó et al. (2022). **(e)** G-statistic, calculated as described by Magwene et al. (2011). **(f)** *p*-values of two-tailed Fisher's exact tests. The dashed line marks the Bonferroni-corrected 5% significance threshold, calculated considering that n=187,181 chromosomal locations have been tested.

**Supplemental Table 1.** Options used with the `mbs` command in each case study.

| Case | Options | Reference |
|---|---|---|
| 1 | -d D,R,Pr,Pd -m R --parental filter -I --EMS | Cao et al. 2019 |
| 2 | -d D,R,Pd,Pr -m R -c 8 --parental-filter | Jiang et al. |
| 3 | -d D,R,Pd -m R -c 8 --parental-filter -I --EMS | Bournonville et al. 2023 |
| 4 | -d R,Pd,Pr -m R --parental filter -I --EMS | Huang et al. 2022 |
| 5 | -d R,Pd -m R --parental-filter --EMS | Viñegra de la Torre et al. 2022 |
| 6 | -d D,R,Wr -m R -r D --parental-filter --EMS | Rodríguez-Alcocer et al. 2023 |
| 7 | -d R,D -m R -r D -c 8 -I | Luo et al. 2023 |
| 8 | -d R,D -m R --het-filter -q 10 -Q 90 -I | Castillejo et al. 2020 |
| 9 | -d D,R -m R --het-filter -q 10 -Q 90 -I | Yang et al. 2022 |
| 10 | -d H,L | Takagi et al 2013 |

# 10 Agradecimientos

Esta Tesis Doctoral es el resultado de muchos años de esfuerzo y dedicación, y no habría sido posible sin el apoyo de las personas que me han acompañado en este trabajo.

En primer lugar, quiero agradecer a **Héctor**, por todas las horas de dedicación y de trabajo, por enseñarme a hacer ciencia y a perseverar, aunque todo esté en tu contra. Gracias a **Sara** por su inestimable ayuda.

A lo largo de los años he conocido a muchos/as compañeros/as cuya ayuda, apoyo y amistad ha sido imprescindible para esta tesis. Gracias a **Sergio**, **Mª Salud** y **Aurora** por esas charlas, café en mano, y por esos congresos que compartimos.

Gracias a **Eva**, que más que una compañera de trabajo ha sido la mejor amiga que podía tener en esta larga carrera de fondo. Gracias por tu alegría y tu positividad, por ayudarme a superar cada día y por tu apoyo incondicional.

A **César**, que me ha acompañado (y aguantado) en este último año de tesis. Gracias por las bromas, por las charlas sobre las cosas más intrascendentes en la puerta del instituto, por tantas horas estrujándonos el cerebro. Sin ti, Maptools no habría sido posible. Pero, sobre todo, gracias por ser tan buen amigo.

A mis mejores amigos **Dani**, **Jorge**, **Irina**, **Cristian** y **Manolo**, que siempre me han ayudado a desconectar y me han ofrecido su apoyo.

A Ebony Code (**Txurro**, **Charly**, **Miguel**, **Juanan** y **Adri**), por ayudarme a superar los malos momentos con su apoyo moral y su infinita alegría y energía.

A Black Night (**Dani**, **Carlos**, **Luis**, **Rafa** y **Manuel**) que no solo me han ayudado a mantenerme económicamente durante la realización de este trabajo, sino que siempre han estado para cualquier cosa.

A mis tías **Mónica**, **Anne** e **Isabel**, por ayudarme en todo momento que lo he necesitado, a **Elisa** por ayudarme con las matemáticas y a **Nuria** con la literatura en el instituto.

A mis padres **María** y **Antonio** por su cariño, por la educación que me han dado y por su apoyo incondicional. Siempre me han incentivado a estudiar y a seguir adelante, y han hecho todo lo posible para que pudiese hacerlo.

A **María**, mi compañera de viaje. Gracias por estar ahí en los buenos y malos momentos, por tu infinita comprensión y tus consejos. Gracias por llenar de felicidad mis días.