



---

# IMPACTO DEL ADN MITOCONDRIAL EN LA ARQUITECTURA DE LOS GENOMAS VEGETALES

---

**TRABAJO FIN DE GRADO**

**AUTORA: AIDA TURÓN RUIZ**

**TUTOR: LORENZO CARRETERO PAULET**



**Grado en Biotecnología**

**Facultad de Ciencias Experimentales**

**Departamento de Biología y Geología**

**Área de Genética**

**Curso académico 2021-2022 - Convocatoria de julio**

## Índice de contenido

<b>Resumen</b> .....	<b>i</b>
<b>Summary</b> .....	<b>ii</b>
<b>1. Introducción</b> .....	<b>1</b>
1.1. Antecedentes .....	1
1.2. Transferencia de ADN exógeno y organelar al núcleo .....	1
1.2.1. Mecanismos moleculares de escape del ADN mitocondrial e integración en el genoma nuclear.....	3
1.2.2. Modificaciones de los NUMTs.....	6
<b>2. Objetivos</b> .....	<b>10</b>
2.1. Obtención de genomas nucleares, mitocondriales y plastídicos de un conjunto de especies representativa para los principales linajes de especies vegetales .....	10
2.2. Extracción de parámetros asociados a los genomas.....	10
2.3. Determinación del porcentaje de ADN mitocondrial en el genoma nuclear mediante la herramienta BLAST.....	10
2.4. Análisis descriptivo y comparativo entre los parámetros seleccionados .....	10
<b>3. Materiales y metodología</b> .....	<b>11</b>
3.1. Especies y genomas utilizados en el estudio.....	11
3.2. Identificación de ADN organelar en los genomas nucleares mediante la herramienta BLAST.....	15
3.3. Extracción de parámetros asociados a los genomas.....	15
3.4. Desarrollo de representaciones gráficas.....	16
<b>4. Resultados y discusión</b> .....	<b>17</b>
4.1. Presentación de los datos .....	17
4.1.1. Análisis de los tamaños de los genomas .....	17
4.1.2. Análisis de los parámetros asociados a los genomas.....	18
4.1.3. Análisis de los resultados BLAST .....	20
4.1.3.1. Distribución de los NUMTs a lo largo de un tiempo evolutivo.....	22
4.2. Análisis estadísticos.....	23
4.2.1. Correlograma de variables seleccionadas.....	23

4.2.2. Correlación positiva entre tamaño de los genomas nucleares y mitocondriales y la longitud total de los NUMTs .....	25
4.2.3. La correlación entre el tamaño de los NUMTs y el porcentaje de identidad revela patrones diferenciales en las especies vegetales estudiadas .....	28
<b>5. Conclusiones.....</b>	<b>31</b>
<b>6. Bibliografía .....</b>	<b>32</b>

## Índice de figuras

<b>Figura 1.</b> Estructuras de genes nucleares que contienen inserciones de ADN organelar (NUMTs) en exones codificantes de proteínas.....	2
<b>Figura 2.</b> Estructuras de genes nucleares formados en su mayoría por inserciones de ADN organelar (NUMTs).....	2
<b>Figura 3.</b> Representación de las diferentes vías de integración del ADN mitocondrial en el núcleo celular .....	5
<b>Figura 4.</b> Representación gráfica esquemática de las dinámicas que se llevan a cabo tras la inserción de los fragmentos mitocondriales y plastídicos en el núcleo celular .....	8
<b>Figura 5.</b> Árbol de las relaciones evolutivas entre las especies a estudiar .....	14
<b>Figura 6.</b> Boxplots de los tamaños de los genomas nucleares, mitocondriales y plastídicos de las 16 especies vegetales seleccionadas .....	18
<b>Figura 7.</b> Boxplots de los tamaños mínimos de NUMTs, tamaños máximos de NUMTs, porcentaje de guanina y citosina en el núcleo y porcentaje de guanina y citosina en la mitocondria de las 16 especies seleccionadas.....	19
<b>Figura 8.</b> Boxplots de los genes codificantes, densidad génica y número de cromosomas en los genomas nucleares de las 16 especies seleccionadas .....	20
<b>Figura 9.</b> Boxplots de los datos obtenidos durante el análisis de las variables seleccionadas de las 16 especies seleccionadas .....	21
<b>Figura 10.</b> Gráficas de densidad por especie del porcentaje de identidad de los NUMTs .....	23
<b>Figura 11.</b> Correlograma de las variables seleccionadas usando el coeficiente estadístico de Pearson.....	24
<b>Figura 12.</b> Correlograma de las variables seleccionadas usando el coeficiente estadístico de Spearman. ....	25
<b>Figura 13.</b> Gráficos de dispersión entre tamaños de los genomas y longitud total de los NUMTs .....	27
<b>Figura 14.</b> Gráficos de dispersión entre tamaños de los genomas y el número total de NUMTs .....	28
<b>Figura 15.</b> Gráficos de dispersión por especie del tamaño de los NUMTs y el porcentaje de identidad. ....	29

## Índice de tablas

<b>Tabla 1.</b> Representación de las 16 especies vegetales utilizadas con sus versiones de los genomas nucleares, mitocondriales y plastídicos. ....	11
<b>Tabla 2.</b> Resumen de los datos de tamaño de los genomas generado por Rstudio. ....	17
<b>Tabla 3.</b> Resumen de los resultados obtenidos en el test de Shapiro-Wilk. ....	26
<b>Tabla 4.</b> Resultados del análisis de correlación antes y después de excluir a <i>M. acuminata</i> . ....	27

## Resumen

En las últimas décadas se ha evidenciado la existencia de una transferencia de ADN mitocondrial al genoma nuclear en células vegetales. Este tránsito de material genético puede dar lugar al establecimiento de nuevos genes, o regiones génicas, y funciones génicas representando una fuente de evolución de las especies vegetales. Sin embargo, los mecanismos moleculares que se encuentran detrás de estos eventos de transferencia actualmente son desconocidos, siendo asociados comúnmente con la transferencia y reordenamiento primordial de genes durante el establecimiento de la endosimbiosis entre bacterias aeróbicas y células eucariotas ancestrales que dio lugar a las mitocondrias, pero sin indicio real de relación entre ambas. En este estudio se realizaron diferentes aproximaciones de genómica comparativa con el fin de dilucidar los factores implicados y las diferencias en las diversas especies vegetales de la transferencia de ADN mitocondrial al genoma nuclear. En primer lugar, se seleccionaron 16 especies representativas de diferentes linajes y se corroboró la existencia de material genético mitocondrial en sus genomas nucleares. Seguidamente, se seleccionaron diferentes parámetros asociados a los genomas los cuales sirvieron para realizar análisis comparativos con los que no fue posible encontrar patrones universales para todas las especies vegetales, pero sí sugerir la reformulación de premisas iniciales que, hoy en día y debido a la falta de investigaciones al respecto, continúan vigentes. Además, se pudo demostrar la concentración de los eventos de transferencia en momentos concretos del tiempo evolutivo, resultados que arrojan luz sobre el impacto real que puede tener el ADN mitocondrial en la evolución de la arquitectura y función de los genomas nucleares.

**Palabras clave:** ADN mitocondrial, células vegetales, nuevos genes, genómica comparativa, transferencia de ADN, genomas nucleares.

## Summary

Over the last decades, the existence of a transfer of mitochondrial DNA to the nuclear genome in plant cells has been evidenced. This transfer of genetic material can lead to the establishment of new genes, or gene regions, and gene functions representing a mechanism of evolution of plant species. However, the molecular mechanisms behind these transfer events are currently unknown, being commonly associated with primordial gene transfer and rearrangement during the establishment of endosymbiosis between aerobic bacteria and ancestral eukaryotic cells that gave rise to mitochondria, but with no real indication of a relationship between the two. In this study, different comparative genomics approaches were performed in order to elucidate the factors involved and the differences in the various plant species in the transfer of mitochondrial DNA to the nuclear genome. First, 16 representative species from different lineages were selected and the existence of mitochondrial genetic material in their nuclear genomes was corroborated. Next, different parameters associated with the genomes were selected and used to perform comparative analyses with which it was not possible to find universal patterns for all plant species, but it was possible to suggest the reformulation of initial premises that, nowadays and due to the lack of research in this regard, continue to be valid. In addition, it was possible to demonstrate the concentration of transfer events at specific moments in evolutionary time, results that shed light on the real impact that mitochondrial DNA can have on the evolution of the architecture and function of nuclear genomes.

**Key words:** Mitochondrial DNA, plant cells, novel genes, comparative genomics, DNA transfer, nuclear genomes.

## 1. Introducción

### 1.1. Antecedentes

Las células eucariotas surgen como consecuencia de procesos simbiogénéticos, como describió la bióloga Lynn Margulis en su artículo “On the origin of mitosing cells” en 1967, a partir de relaciones simbióticas estables entre organismos, en este caso bacterias aeróbicas (alfa-proteobacterias) que corresponden con las mitocondrias, bacterias fotosintéticas (cianobacterias), los cloroplastos, y una célula eucariota ancestral (Leister & Kleine, 2011). Tras esta transferencia masiva primaria (transferencia endosimbiótica de genes o TEG), se ha generado, un continuo trasiego de ADN organelar al núcleo celular comúnmente asociado al mismo mecanismo que actuó durante el establecimiento de la simbiosis pero que no necesariamente se encuentra relacionado. Estas transferencias post-endosimbióticas aún ocurren de forma activa y continua en los linajes eucariotas y no solo en las primeras etapas del desarrollo, el proceso continúa, sobre todo, en las angiospermas (Matsuo et al., 2005). Gracias a los avances en secuenciación y a los análisis bioinformáticos, se han podido correlacionar diferentes parámetros asociados a los genomas como el tamaño total del genoma nuclear con el contenido de ADN organelar que comprende en especies modelo como *Arabidopsis thaliana* (X. Li et al., 2020a), demostrando una correlación positiva entre ambas. Estos análisis nos ofrecen un paso más en la determinación de las dinámicas y evolución de los genomas de plantas y otros organismos celulares, pudiendo de esta forma llegar a predecir el porcentaje de ADN mitocondrial en el genoma nuclear de cualquier especie vegetal.

### 1.2. Transferencia de ADN exógeno y organelar al núcleo

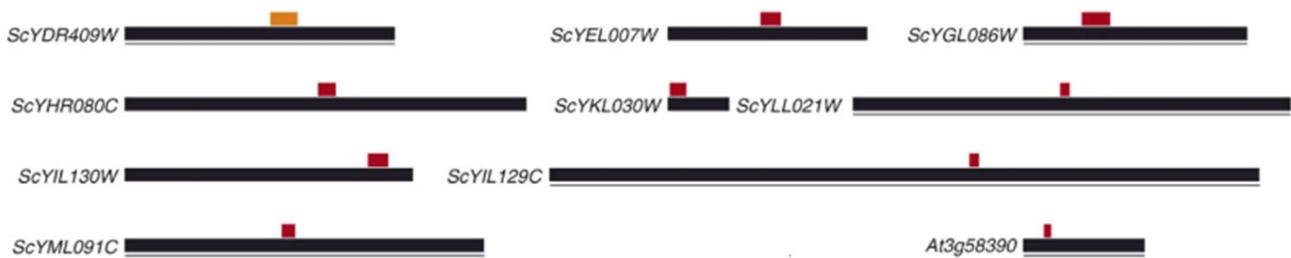
La transferencia endosimbiótica de genes, indicada en el apartado anterior, se puede considerar un mecanismo de transferencia lateral u horizontal de genes (THG), el cual corresponde con el tránsito de material genético que se da entre organismos, siendo una de las principales fuentes de evolución en arqueas y en bacterias (X. Li et al., 2020b).

Se utilizaron técnicas de genómica comparativa para comparar los genes nucleares individuales de la especie modelo *A. thaliana* con genomas procariontes y eucariotes representativos, revelando que, 866 de 9.368 proteínas codificadas por el núcleo que estaban lo suficientemente conservadas para permitir comparaciones válidas, presentaban similitud con las proteínas de la cianobacteria. Otras 834 proteínas generaron árboles filogenéticos que contenían ramas cianobacterianas, lo que concluyó con que cerca de 4.500 genes de 25.000 genes nucleares de *A. thaliana* fueron adquiridos de un antepasado cianobacteriano, un plástido. Estos resultados explican la dependencia de los plástidos, tanto mitocondrias como cloroplastos, para los genes nucleares además de corroborar el papel tan importante que llevó a cabo la endosimbiosis a nivel de complejidad genética en los núcleos eucariotas (Bock & Timmis, 2008).

Sin embargo, como se ha indicado en el apartado 1.1., tras la transferencia masiva de genes y su recolocación en el genoma nuclear, se ha producido, además, una transferencia de ADN organelar continuo en el que en la gran mayoría de casos el ADN que se transfiere no llega a ser funcional en el

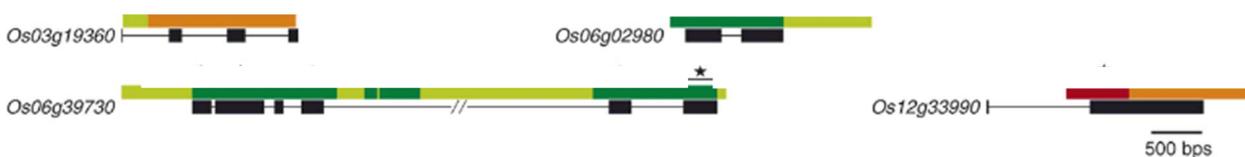
núcleo. A pesar de que el ADN transferido tiene una alta probabilidad de resultar neutral o deletéreo, en algunos casos puede dar lugar a la transferencia de genes completos o de secuencias que remodelan los genes al ser incorporadas en los exones, introduciendo así nuevas secuencias de codificación en genes nucleares preexistentes (Kleine et al., 2009). Las transferencias de ADN mitocondrial o plastídico dan lugar a secuencias denominadas ADN mitocondrial nuclear (en inglés nuclear mitochondrial DNA o NUMT) y ADN plastídico nuclear (en inglés nuclear plastid DNA o NUPT), siendo designado colectivamente como ADN de orgánulos nucleares (en inglés nuclear organelar DNA o norgDNA) (X. Li et al., 2020b). Además, los NUMTs y NUPTs pueden causar la alteración de la actividad génica mediante su inserción en regiones de ADN responsables de la regulación génica, adquiriendo de esta forma un importante papel en la evolución genómica de las plantas.

Se han determinado algunos ejemplos de genes que contienen ADN organelar en regiones codificantes de levadura y *A. thaliana* (Noutsos et al., 2007).



**Figura 1. Estructuras de genes nucleares que contienen inserciones de ADN organelar (NUMTs) en exones codificantes de proteínas (Noutsos et al., 2007).** Los intrones se representan como líneas, las regiones codificantes se observan como rectángulos negros, los NUMTs originados a partir de secuencias codificantes en el genoma mitocondrial se representan con rectángulos rojos y aquellos que provienen de secuencias no codificantes, como rectángulos naranjas (ScYDR409W). Los ejemplos en levadura presentan en su nomenclatura “ScY” mientras que los de *Arabidopsis thaliana* presentan “At”.

También se han expuesto ejemplos de genes nucleares que se encuentran formados básicamente por ADN organelar en la especie *Oryza Sativa*.



**Figura 2. Estructuras de genes nucleares formados en su mayoría por inserciones de ADN organelar (NUMTs) (Noutsos et al., 2007).** Los intrones se representan como líneas, las regiones codificantes se observan como rectángulos negros, los NUMTs originados a partir de secuencias codificantes en el genoma mitocondrial se representan con rectángulos rojos y aquellos que provienen de secuencias no codificantes, como rectángulos

*naranjas. Los rectángulos verdes oscuros representan NUPTs originados a partir de secuencias codificantes en el genoma cloroplástico y en verde claro se representan los NUPTs que provienen de secuencias no codificantes. Todos los ejemplos son de la especie *Oryza Sativa* (Os).*

### 1.2.1. Mecanismos moleculares de escape del ADN mitocondrial e integración en el genoma nuclear

La abundancia de los NUMTs y de los NUPTs en los genomas nucleares varía enormemente, sin embargo, los factores implicados en la salida e integración del ADN organelar han sido muy poco estudiados. Algunas de las teorías existentes se pueden ver representadas en la **figura 3** (Hazkani-Covo et al., 2010). En cuanto a los mecanismos de escape se encuentran:

- A) Degradación de las mitocondrias anormales. Es la teoría más aceptada. Se han realizado estudios demostrando cómo muchas cepas de levadura “yme” (en inglés yeast mitochondrial escape) muestran niveles altos de escape o transferencia de ADN de las mitocondrias al núcleo celular. Los mutantes yme1 presentan una inactivación de la proteína Yme1p, una metaloproteasa dependiente de ATP localizada en la mitocondria que proporciona una alta tasa de escape mitocondrial. Esto es debido a que las mitocondrias de las cepas yme1 son degradadas por vacuolas más frecuentemente que las mitocondrias de cepas WT (Hazkani-Covo et al., 2010).
- B) Lisis del compartimento mitocondrial. Este fenómeno se puede deber a múltiples factores de entre los que destacan:

- **Gametogénesis masculina**

En numerosos organismos eucariotas, incluidos los humanos y las plantas con flores, el ADN organelar se hereda de forma materna, por lo que, la transferencia de ADN se puede llevar a cabo durante la degradación de los orgánulos en el proceso de la gametogénesis masculina: durante el proceso de desarrollo de polen en las plantas con flores (Sheppard et al., 2008). Se ha estudiado, además, la absorción de mitocondrias enteras por el núcleo, proporcionando una asociación física directa en las células espermáticas del tabaco.

- **Estrés oxidativo**

Usando la técnica de marcador de secuencia expresada (ESTs) y promotores predecidos, se pudo revelar que ciertos genes plastídicos y mitocondriales localizados en el núcleo, además de ser funcionales han podido sufrir una presión de selección natural positiva (Cullis et al., 2009).

Como ya se ha expuesto, la existencia de NUMTs exige que se hayan producido fragmentos de ADN mitocondrial en la mitocondria y estos fragmentos consigan alcanzar el núcleo a pesar de las barreras físicas que se presentan. Se han propuesto varios mecanismos que facilitarían la salida de fragmentos de ADN mitocondrial de la mitocondria y su entrada en el núcleo, uno de los más aceptados consiste en que los fragmentos de ADN mitocondrial son producidos por la constante generación de radicales libres de oxígeno (ROS) en la mitocondria. El ADN mitocondrial, que se encuentra desprovisto de histonas y situado en la membrana interna de las mitocondrias, significativamente cerca del lugar dónde se realiza la producción de ROS, es claramente más vulnerable que el ADN nuclear al daño oxidativo provocado por ROS. Los fragmentos que se generan por los radicales libres en las mitocondrias contienen un alto índice de 8-oxo-deoxiguanosina (marcador de estrés oxidativo del ADN) en el citoplasma, causado por las alteraciones ocasionadas en la membrana organelar, pudiendo ser debido a la división, lisis, mitofagia, a eventos de fisión o fusión o a la inducción inicial de un poro de permeabilidad (Puertas & González-Sánchez, 2020).

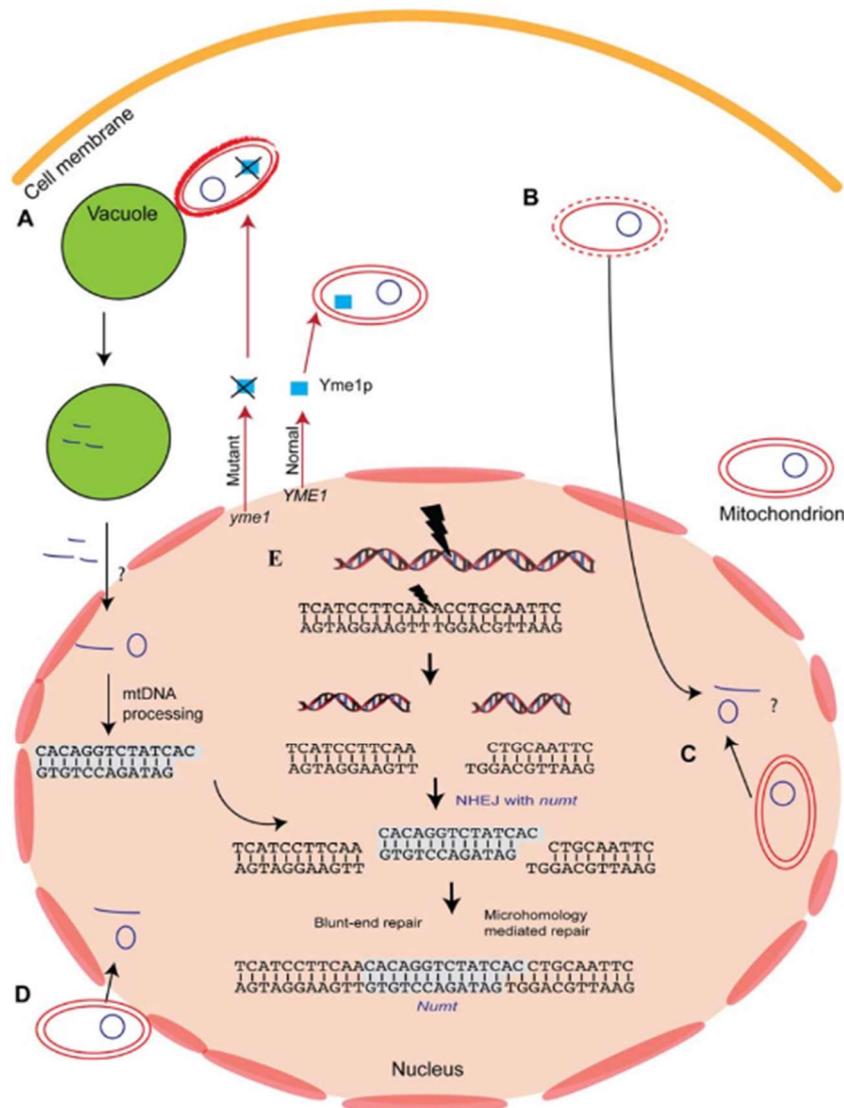
Una vez en el citoplasma, el ADN mitocondrial se encuentra protegido de las nucleasas gracias a un mecanismo mediado por vacuolas, que se degradan cuando entran en contacto con el núcleo. También se ha propuesto que el ADN mitocondrial podría formar un complejo con las proteínas “histonelike” de unión al ADN evitando así su degradación (Puertas & González-Sánchez, 2020).

- C) Encapsulación del ADN mitocondrial en el interior del núcleo celular.
- D) Asociación física directa entre las mitocondrias y el núcleo, llevando a la fusión de las membranas de ambos, como sucede en el caso de las células espermáticas del tabaco.

Una vez explicados los posibles mecanismos de escape del orgánulo que lleva a cabo el ADN mitocondrial, se produce su integración en el núcleo, la cual se explica mediante el proceso de unión de extremos no homólogos (NHEJ).

- E) El ADN mitocondrial que entra en el núcleo es capaz de integrarse en los cromosomas nucleares. El ADN mitocondrial se integra en los cromosomas durante la reparación de las rupturas de doble cadena de ADN (DBSs) mediante un mecanismo denominado unión de extremos no homólogos o NHEJ. Análisis de las transferencias de ADN mitocondrial realizados en plantas de tabaco, en humanos y en levaduras demostraron que con frecuencia existían unas microhomologías de entre 2-5 pb adyacentes a los sitios de integración de los fragmentos, indicando que las inserciones de ADN organelar o NUMTs durante las reparaciones de las rupturas de doble cadena mediadas por NHEJ son comunes en todos los eucariotas.

Las uniones entre los cromosomas y los fragmentos de ADN organelares en cada extremo de un NUMT o NUPT demuestran los eventos de reparación que se han producido en cada uno de los lados de ruptura de la secuencia cromosómica original. La reparación de DBSs por parte de NHEJ solo requiere un pequeño tramo de homología de secuencia de entre 1-7 pb, (también puede no requerirse en el caso de los extremos romos) permitiendo que los extremos no complementarios de DBSs y ADN organelar sean unidos entre sí (Hazkani-Covo et al., 2010).



**Figura 3. Representación de las diferentes vías de integración del ADN mitocondrial en el núcleo celular (Hazkani-Covo et al., 2010). Las letras que las identifican se encuentran relacionadas con las explicadas anteriormente.**

También se desarrolló la “Hipótesis de la ventana de transferencia limitada”, la cual explica el hecho de que las células monoplastídicas (aquellas que tienen, en este caso, un plástido por célula), tienen menos NUPTs que aquellas que tienen muchos (poliplastídicas). Las posibles explicaciones para este evento son que en las células monoplastídicas, la transferencia de ADN al núcleo disminuye notablemente en comparación con la tasa de las poliplastídicas debido en primer lugar a que, como solo hay un plástido, es la única que puede donar ADN al núcleo. En segundo lugar, la pérdida del plástido debido a sus lisis llevaría a la muerte de la célula, al contrario que en las especies poliplastídicas (Smith et al., 2011).

Aunque esta hipótesis se ha desarrollado para el estudio de NUPTs, se desconoce si se pudieran aplicar los mismos criterios para los NUMTs.

### 1.2.2. Modificaciones de los NUMTs

Una vez que el fragmento organelar es integrado en el genoma nuclear, como se ha indicado anteriormente, este se va a someter gradualmente a una serie de cambios que le van a permitir su adaptación en el genoma huésped. La mayoría de los NUMTs y NUPTs son eliminados ya que se ven afectados por la presión de selección. Podemos observar diferentes dinámicas que se llevan a cabo tras la inserción de los fragmentos y que, además, aparecen representadas esquemáticamente en la **figura 4** (Zhang et al., 2020):

#### ➤ **Eliminación**

En primer lugar, si comprobamos las altas frecuencias de transferencia de ADN organelar y las comparamos con los porcentajes de NUMTs y NUPTs en el núcleo celular, observamos una diferencia significativa, lo que nos indica que los fragmentos son rápidamente eliminados evitando, además, el rápido crecimiento genómico. Estas eliminaciones se producen normalmente durante la mitosis, aunque pueden ocurrir de forma ocasional en el proceso de la meiosis.

Estudios realizados sobre los NUPTs del genoma nuclear de *O. sativa* dieron lugar a la determinación de la vida media de estos fragmentos, la cual se estima que es de 0,5 millones de años para las inserciones largas y de 2,2 millones de años para las inserciones cortas, lo que nos indica que, como se ha indicado anteriormente, la eliminación de los fragmentos constituye un mecanismo por el cual se mantiene un balance entre la continua integración de los fragmentos y el mantenimiento del tamaño del genoma nuclear (Zhang et al., 2020).

#### ➤ **Mutación**

La mutación es una propiedad intrínseca del ADN, lo que conlleva que se produzca de forma inevitable en los NUMTs y NUPTs. Dado que las tasas de mutación varían mucho entre ADN organelar y nuclear, las secuencias integradas muestran diferentes tasas de mutación en comparación con las secuencias organelares de las que derivan. La tasa de mutación del ADN

organelar en las especies vegetales es mucho más lenta que la del genoma nuclear. El ADN organelar se desarrolla con aproximadamente una o menos mutaciones por kb por millón de años, y la tasa de mutación es de aproximadamente siete mutaciones por kb por millón de años para el genoma nuclear. De esta forma, tanto los NUPTs como los NUMTs insertados en el genoma nuclear de la planta evolucionan forma más rápida. Durante el proceso evolutivo, las mutaciones de los NUPTs y los NUMTs se acumulan hasta llegar a no ser reconocidos debido a la gran divergencia de secuencia entre las secuencias que son integradas y, posteriormente mutadas, y los ADN organelares originales (Zhang et al., 2020).

#### ➤ **Fragmentación**

Como se ha indicado anteriormente, la similitud o porcentaje de identidad entre los NUPTs y NUMTs y los ADNs plastídicos y mitocondriales de los que provienen disminuyen durante un tiempo evolutivo. En un primer momento, estas secuencias que presentan un 100% de similitud, deberían tener una longitud relativamente extensa que se iría reduciendo debido a la fragmentación en pequeñas secuencias. Una de las posibles causas por las que se produce es debido a la acción de los transposones, los cuales fragmentan la secuencia de forma directa. El proceso de fragmentación ha sido confirmado en el genoma de *Asparagus officinalis* en la que se ha estudiado una correlación positiva entre el tamaño de los fragmentos organelares y su porcentaje de identidad (Zhang et al., 2020).

#### ➤ **Reordenamiento**

Se ha observado que un gran número de NUPTs y NUMTs se encuentran organizados en conjuntos de secuencias originadas en varias regiones del genoma del cloroplasto o de las mitocondrias, lo que indica que existe un fenómeno de reordenamiento antes y/o después de las inserciones de los fragmentos en el genoma nuclear. Se han realizado análisis en profundidad de los NUMTs de *A. thaliana*, en el que el fragmento más grande, de 620 kb, demuestra que se han producido mínimo cuatro eventos de reordenamiento, en los que tres de ellos han sido localizados en los 76 kb que se encuentran al inicio de la inserción, los cuales han podido ocurrir tras la integración, pero también pueden ocurrir de forma previa a la transferencia en el orgánulo mediante recombinación homóloga entre secuencias repetitivas o en el núcleo durante o antes de su integración (Zhang et al., 2020).

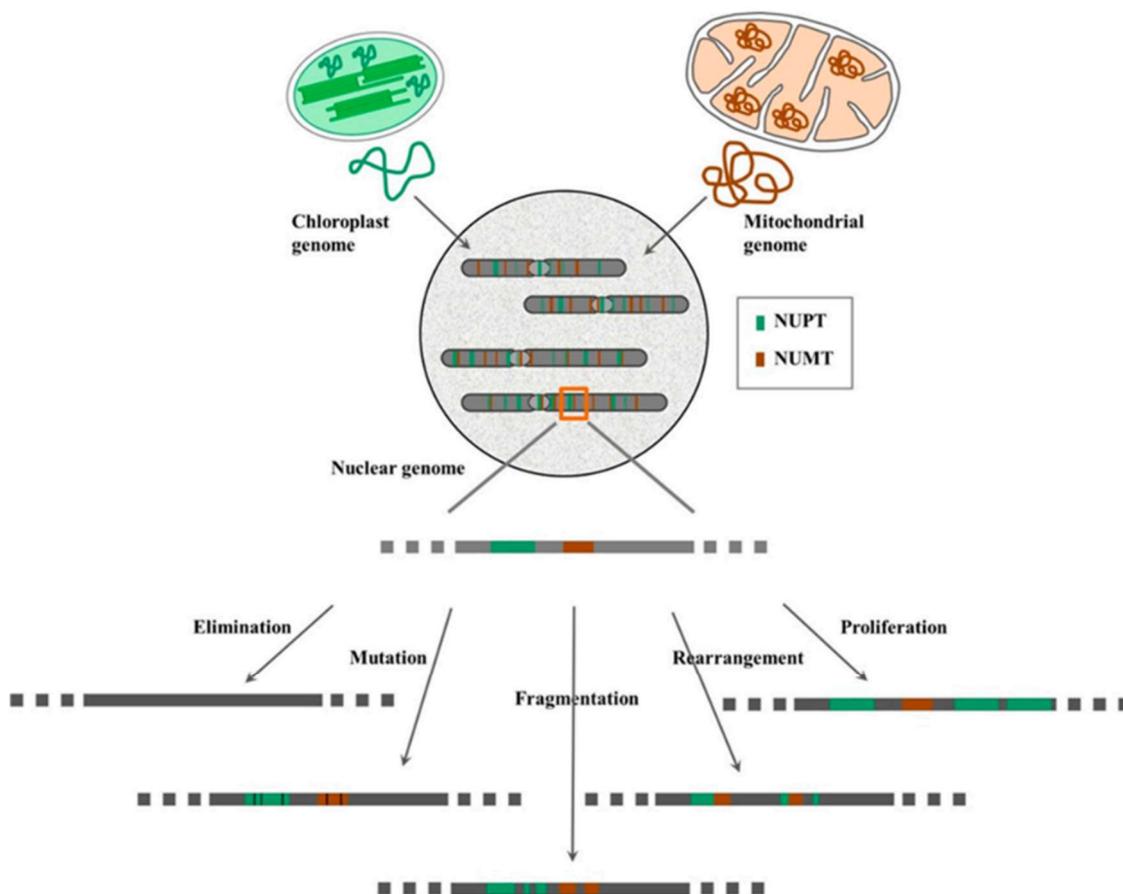
#### ➤ **Proliferación**

Los NUMTs y NUPTs integrados en el genoma nuclear se pueden amplificar para aumentar su número de copias. Estos se pueden organizar o bien como secuencias duplicadas en tándem o distribuyéndose por separado a lo largo del genoma huésped. Como ejemplo, se puede observar que en *A. officinalis* han sido identificados varios clusters de NUPTs ordenados por una serie de secuencias prácticamente idénticas del ADN cloroplastídico. Tanto los NUPTs

como los NUMTs son incapaces de duplicarse ellos mismos por lo que su proliferación en el genoma nuclear se debe probablemente a la acción de retrotransposones vecinos, los cuales incrementan fácilmente el número de copias de los fragmentos mediante un mecanismo de retrotransposición de “copiar y pegar” (Zhang et al., 2020).

Además, se debe tener en cuenta que los NUMTs y los NUPTs son sometidos a una hipermetilación de los residuos de citosina con una posterior desaminación en las plantas superiores. Estas observaciones nos sugieren la gran importancia que representa la metilación en la regulación de los fragmentos de ADN organelar integrados, en concreto de las relacionadas con la supresión de su transcripción (Zhang et al., 2020).

Una vez indicadas las posibles modificaciones a las que pueden ser sometidas los NUMTs y NUPTs, hay que tener en cuenta que los eventos indicados anteriormente no ocurren de forma aislada, la acción de uno no impide que se lleve a cabo otro, al contrario, en muchos casos, como hemos explicado en el apartado de reordenamiento, se llevan a cabo varias de forma conjunta o sucesiva (Zhang et al., 2020). De este modo, los ADN organelares una vez integrados en el genoma nuclear se someten a fragmentaciones, reordenamientos y mutaciones, lo que conduce a que un gran número de estos NUMTs y NUPTs sean eliminados, constituyendo, de nuevo, un mecanismo por el cual se mantiene el equilibrio dinámico entre la continua integración de fragmentos en el genoma nuclear y su eliminación.



**Figura 4. Representación gráfica esquemática de las dinámicas que se llevan a cabo tras la inserción de los fragmentos mitocondriales y plastídicos en el núcleo celular (Zhang et al., 2020). Se pueden observar los mecanismos de eliminación, mutación, fragmentación, reordenamiento y proliferación. Se han representado en color verde los NUPTs y de color naranja los NUMTs.**

## 2. Objetivos

Este estudio se articula en torno a los siguientes cuatro objetivos principales:

### 2.1. Obtención de genomas nucleares, mitocondriales y plastídicos de un conjunto de especies representativa para los principales linajes de especies vegetales

El primer objetivo de este estudio fue la selección y posterior obtención de los genomas nucleares, mitocondriales y plastídicos de especies representativas para los principales linajes de especies vegetales.

### 2.2. Extracción de parámetros asociados a los genomas

Un segundo objetivo ha sido la obtención de parámetros asociados a los genomas a través de diferentes metodologías bioinformáticas para su posterior análisis.

### 2.3. Determinación del porcentaje de ADN mitocondrial en el genoma nuclear mediante la herramienta BLAST

En tercer lugar, se determinó el porcentaje de ADN mitocondrial en el genoma nuclear, usando el programa bioinformático BLAST, con el fin de utilizarlo como variable.

### 2.4. Análisis descriptivo y comparativo entre los parámetros seleccionados

Una vez obtenidas todos los parámetros de estudio, se han realizado análisis descriptivos mediante estudios de correlación, los cuales nos han permitido seleccionar las correlaciones de interés para su posterior análisis comparativo con el fin de encontrar similitudes entre los mecanismos de transferencia, retención y eliminación del ADN mitocondrial en el genoma nuclear.

### 3. Materiales y metodología

#### 3.1. Especies y genomas utilizados en el estudio

Para realizar el estudio, hemos utilizado genomas de 16 especies vegetales. Estos fueron seleccionados por tener una calidad aceptable, presentar tanto el genoma mitocondrial como el plastídico y, además, cada uno de ellos representa los principales linajes de plantas angiospermas: como las monocotiledóneas *Oryza sativa ssp. Japonica* (arroz), *Musa acuminata* (plátano rojo), la planta acuática *Zostera marina* y *Zea mays* (maíz), las eudicotiledóneas *Arabidopsis thaliana*, *Vitis vinifera* (uva), *Solanum lycopersicum* (tomate), *Daucus carota* (zanahoria), *Medicago truncatula* y *Rosa chinensis*. También se incluyó a la dicotiledónea basal *Nelumbo nucifera*, la angiosperma basal *Amborella trichopoda* y *Magnolia biondii*. Además, se ha añadido la gimnosperma *Pinus taeda*, un alga verde unicelular, *Chlamydomonas reinhardtii* y un musgo, *Physcomitrium patens*.

Los genomas utilizados, mostrados en la tabla a continuación (**tabla 1**), se bajaron de diferentes bases de datos genómicas comparativas: TreeGenes (<https://treegenesdb.org/>) (Wegrzyn et al., 2019), PLAZA5.0 (<https://bioinformatics.psb.ugent.be/plaza/>) (Van Bel et al., 2022; Vandepoele et al., 2013), NCBI (<https://www.ncbi.nlm.nih.gov/>) y BGH (<https://banana-genome-hub.southgreen.fr/>) (Droc et al., 2013).

**Tabla 1.** Representación de las 16 especies vegetales utilizadas con sus versiones de los genomas nucleares, mitocondriales y plastídicos.

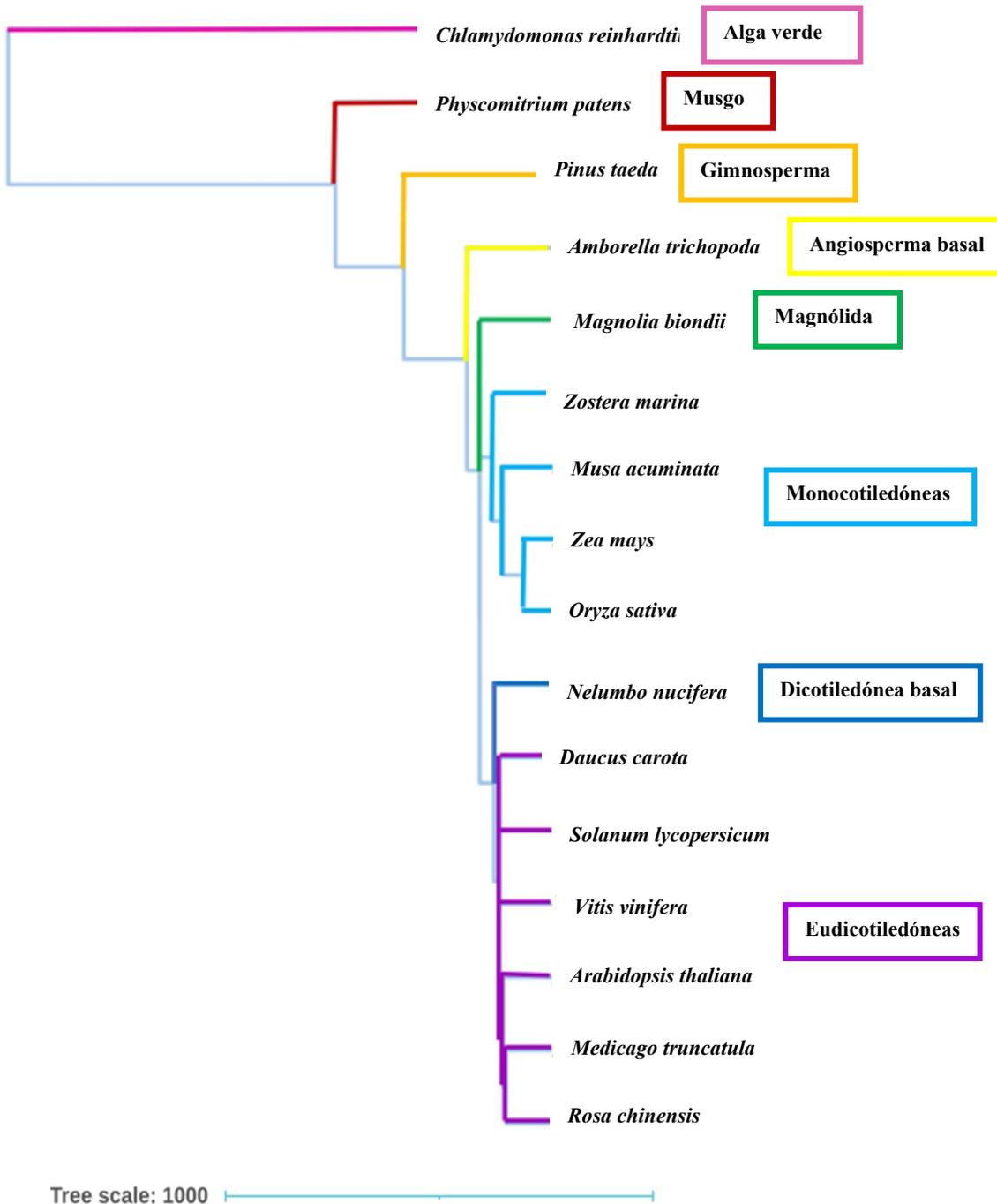
Especies	Versión genoma nuclear	Versión genoma mitocondrial	Versión genoma plastídico	
<i>Amborella trichopoda</i> (Atr)	JGI v1.0	KF754799-KF754803*	NC_005086.1	
<i>Arabidopsis thaliana</i> (Ath)	TAIR Araport11	NC_037304.1	NC_000932.1	
<i>Chlamydomonas reinhardtii</i> (Cre)	JGI v5.6	NC_001638.1	NC_005353.1	

Especies	Versión genoma nuclear	Versión genoma mitocondrial	Versión genoma plastídico	
<i>Daucus carota</i> ssp. <i>Carota</i> (Dca)	JGI v2.0	Included in JGI v2.0	Included in JGI v2.0	
<i>Magnolia biondii</i> (Mbi)	DRYAD v1.0	NC_049134.1	NC_034687.1	
<i>Medicago truncatula</i> (Mtr)	JGI Mt4.0v1	NC_029641.1	NC_003119.8	
<i>Musa acuminata</i> ssp. <i>Malaccensis</i> (Mac)	BGH v4.0	Included in BGH v4.0	HF677508.1	
<i>Nelumbo nucifera</i> (Nnu)	NGD v2.0	NC_030753.1	NC_025339.1	
<i>Oryza sativa</i> ssp. <i>12inifera</i> (Osj)	IRGSP-1.0	NC_011033.1	NC_001320.1	
<i>Physcomitrium patens</i> (Ppa)	JGI v3.3	NC_007945.1	NC_005087.2	

Especies	Versión genoma nuclear	Versión genoma mitocondrial	Versión genoma plastídico	
<i>Solanum lycopersicum</i> (Sly)	ITAG4.0	NC_035963.1	NC_007898.3	
<i>Rosa chinensis</i> cv. <i>Old Blush</i> (Rch)	GDR v2.0	CM009589.1	MH332770.1	
<i>Pinus taeda</i> (Pta)	TreeGenes v2.01	NC_039746.1	KY964286.1	
<i>Vitis vinífera</i> (Vvi)	JGI v2.1	NC_012119.1	NC_007957.1	
<i>Zea mays</i> cv. <i>B73</i> (Zma)	NAM v5.0	NC_007982.1	NC_001666.2	
<i>Zostera marina</i> (Zmr)	JGI v3.1	NC_035345.1	NC_036014.1	

Además, se realizó un árbol filogenético para describir las relaciones evolutivas entre las 16 especies de plantas seleccionadas utilizando la topología y los tiempos de divergencia entre especies disponibles en la base de datos disponible en TimeTree (<http://www.timetree.org/>) (Kumar et al., 2017).

Posteriormente se utilizó el programa iTOL (<https://itol.embl.de/>) (Letunic & Bork, 2021), mediante el cual se ha podido realizar modificaciones con fines estéticos del archivo newick generado con TimeTree, dando como resultado el árbol filogenético que se muestra a continuación (**figura 5**).



**Figura 5.** *Árbol de las relaciones evolutivas entre las especies a estudiar.* La escala representa 1000 millones de años.

### 3.2. Identificación de ADN organelar en los genomas nucleares mediante la herramienta BLAST

Para llevar a cabo la identificación de ADN organelar en los genomas nucleares se ha utilizado el programa BLAST+ 2.10.0, el cual ha sido descargado y corrido de forma local en un equipo con el sistema operativo Linux con interfase Ubuntu versión 18.04 LTS. Esta herramienta resulta de gran utilidad ya que encuentra regiones de similitud local entre secuencias. El programa compara secuencias de nucleótidos o proteínas con bases de datos de secuencias y calcula la significación estadística de coincidencias. BLAST se puede utilizar para inferir relaciones funcionales y evolutivas entre secuencias, así como para ayudar a identificar a miembros de familias de genes (Altschul et al., 1990; Camacho et al., 2009).

1. En primer lugar, se ha realizado el procesamiento de los archivos en formato FASTA de todos los genomas de las especies usando de nuevo la terminal Linux. Se han creado archivos “.blastable” mediante el comando “makeblastdb” para hacerlos accesibles a BLASTN.

➤ *~/home/ubuntu/Desktop/TFG/ncbi-blast-2.10.0+/bin/makeblastdb -in /home/ubuntu/Desktop/TFG/Genomas/AthMito.fasta -dbtype nucl -parse\_seqids -out /ubuntu/Desktop/TFG/Blastable/AthMito.Blastable -logfile /home/ubuntu/Desktop/TFG/Blastable/AthMito.log*

2. Una vez que hemos obtenido los archivos “.blastable”, se ha realizado un BLAST entre los genomas mitocondriales, los cuales representan las secuencias “query”, y nucleares de cada especie usando BLASTN con el comando a continuación.

➤ *~/home/ubuntu/Desktop/TFG/ncbi-blast-2.10.0+/bin/blastn -query /home/ubuntu/Desktop/TFG/Genomas/AthMito.fasta -db /home/ubuntu/Desktop/TFG/Blastable/AthNucBlastable -outfmt 6 -evaluate 1e-5 -word\_size 9 -penalty -2 -show\_gis -dust no -num\_threads 2 -out /home/ubuntu/Desktop/TFG/ResultadosBLAST'/Ath.Mito2Nuc.blast6.out*

Los parámetros seleccionados fueron los siguientes: valor E < 10<sup>-5</sup>, formato 6 por defecto.

### 3.3. Extracción de parámetros asociados a los genomas

Se han denominado como “variables” a los parámetros asociados a los genomas que van a ser analizados. Para obtener los valores de las variables “tamaño de los genomas”, “longitudes de los fragmentos” y “contenido de guanina y citosina” se desarrollaron distintos pipelines basados en comandos de Linux que se describen a continuación:

1. En primer lugar, utilizando una terminal de Linux, se tomaron los archivos en formato FASTA de los genomas y se llevó a cabo la determinación de los tamaños mediante el recuento de caracteres de cada archivo.

➤ `~ grep -v '>' '/home/ubuntu/Desktop/TFG/Genomas/AthMito.fasta' | tr -d '\n' | wc -c`

2. Se ha realizado la suma de las longitudes de los alineamientos.

➤ `~ awk -F'\t' '{sum+=$4;}END{print sum;}' home/ubuntu/Desktop/TFG/Genomas/AthMito.fasta' | tr -d '\n' | wc -c`

3. Para el cálculo del contenido de guanina y citosina se usó el siguiente comando.

➤ `~ grep -v "^>" '/home/ubuntu/Escritorio/TFG/Genomas/AthMito.fasta' | egrep -o 'G|C' | wc -l`

4. Finalmente se calculó contenido total de adenina (A), timina (T), citosina (C) y guanina (G) ignorando las Ns (nucleótidos no identificados) con el fin de calcular el porcentaje de guanina y citosina (porcentaje de GC) de los genomas dividiendo el contenido de GC entre el contenido total de ATCG.

➤ `~ grep -v "^>" '/home/ubuntu/Escritorio/TFG/Genomas/AthMito.fasta' | egrep -o 'A|T|G|C' | wc -l`

### 3.4. Desarrollo de representaciones gráficas

Se ha usado el lenguaje de programación R versión 2020 empleando el EDI (entorno de desarrollo integrado) Rstudio versión 2020 (Rstudio Team, 2020) para el desarrollo de los códigos empleados para la obtención de representaciones gráficas de las variables obtenidas anteriormente. Concretamente se emplearon las siguientes extensiones, o paquetes, de R:

- tidyverse (Wickham et al., 2019): gran parte de los gráficos fueron generados utilizando este paquete (incluyendo las **figuras 6, 7, 8, 9, 10, 13 y 14**), el cual consta a su vez de otras extensiones de entre las que hemos utilizado ggplot 2 versión 3.3.3. (Wickham, 2016). Estos paquetes han permitido generar y modificar gráficos de cajas y bigotes o “boxplots”.
- ggpubr versión 0.4.0. (Alboukadel, 2020): se utilizó para generar la **figura 15** junto con el paquete ggscatter, ya que aporta la posibilidad de generar gráficos de dispersión que presenten el p-value y el coeficiente de determinación ( $R^2$ ).

- ggcorplot versión 0.1.3 (Alboukadel, 2019): fue utilizado junto con el paquete ggplot2 para generar gráficos de correlación incluyendo los coeficientes de correlación (**figuras 11 y 12**)

## 4. Resultados y discusión

### 4.1. Presentación de los datos

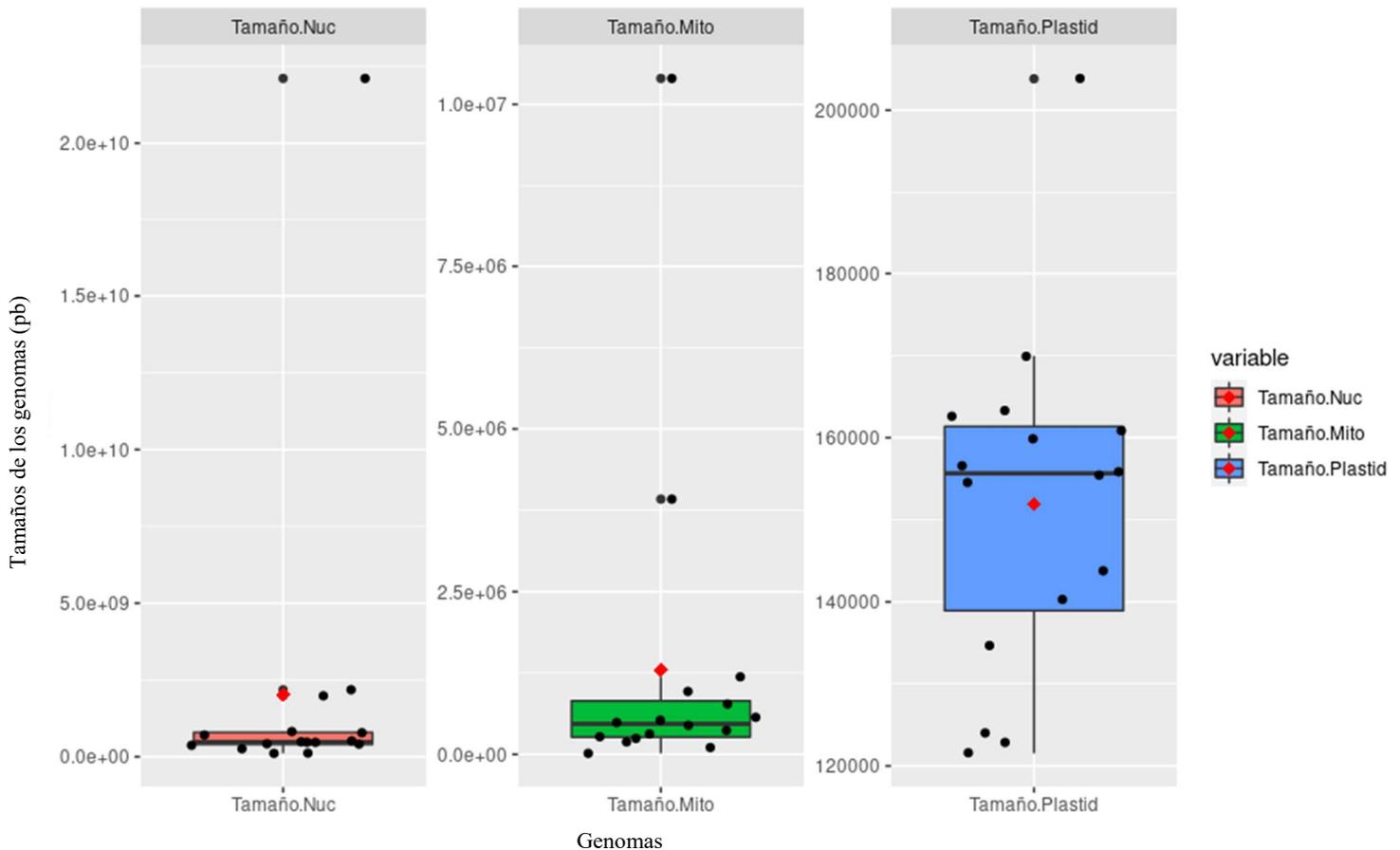
Una vez obtenidas las variables utilizando la metodología explicada en los apartados 3.2.1 y 3.2.2, se ha llevado a cabo el análisis más en profundidad de los datos obtenidos.

#### 4.1.1. Análisis de los tamaños de los genomas

Se ha realizado la representación de los datos de los tamaños de los genomas nucleares, plastídicos y mitocondriales mediante un boxplot (**figura 6**). En el caso del tamaño del genoma nuclear, observamos que los puntos desplazados superiores corresponden con la especie *Pinus taeda*, con un tamaño de 22104 Gb, reflejado en el “Data Summary” (tabla 2) como valor máximo (p100 en la **tabla 2**) a diferencia de tamaños como el de *C. reinhardtii*, de 111 Gb, representando el valor más bajo (p0 en la **tabla 2**). En el caso del tamaño del genoma mitocondrial hay que señalar la gran heterogeneidad que existe entre las especies ya que nos encontramos genomas como el de *M. acuminata*, que presenta 10 Gb y *A. trichopoda* con 3,9 Gb como valores más altos, y, de nuevo, *C. reinhardtii* como valor más bajo, siendo de 15758 pb. Sin embargo, en el caso del tamaño del genoma plastídico, podemos destacar a *C. reinhardtii* con un genoma de 203828 pb, como el genoma mitocondrial más grande, a diferencia de *P. taeda*, que contiene 121531 pb como podemos comprobar en la **tabla 2**.

**Tabla 2.** Resumen de los datos de tamaño de los genomas generado por Rstudio.

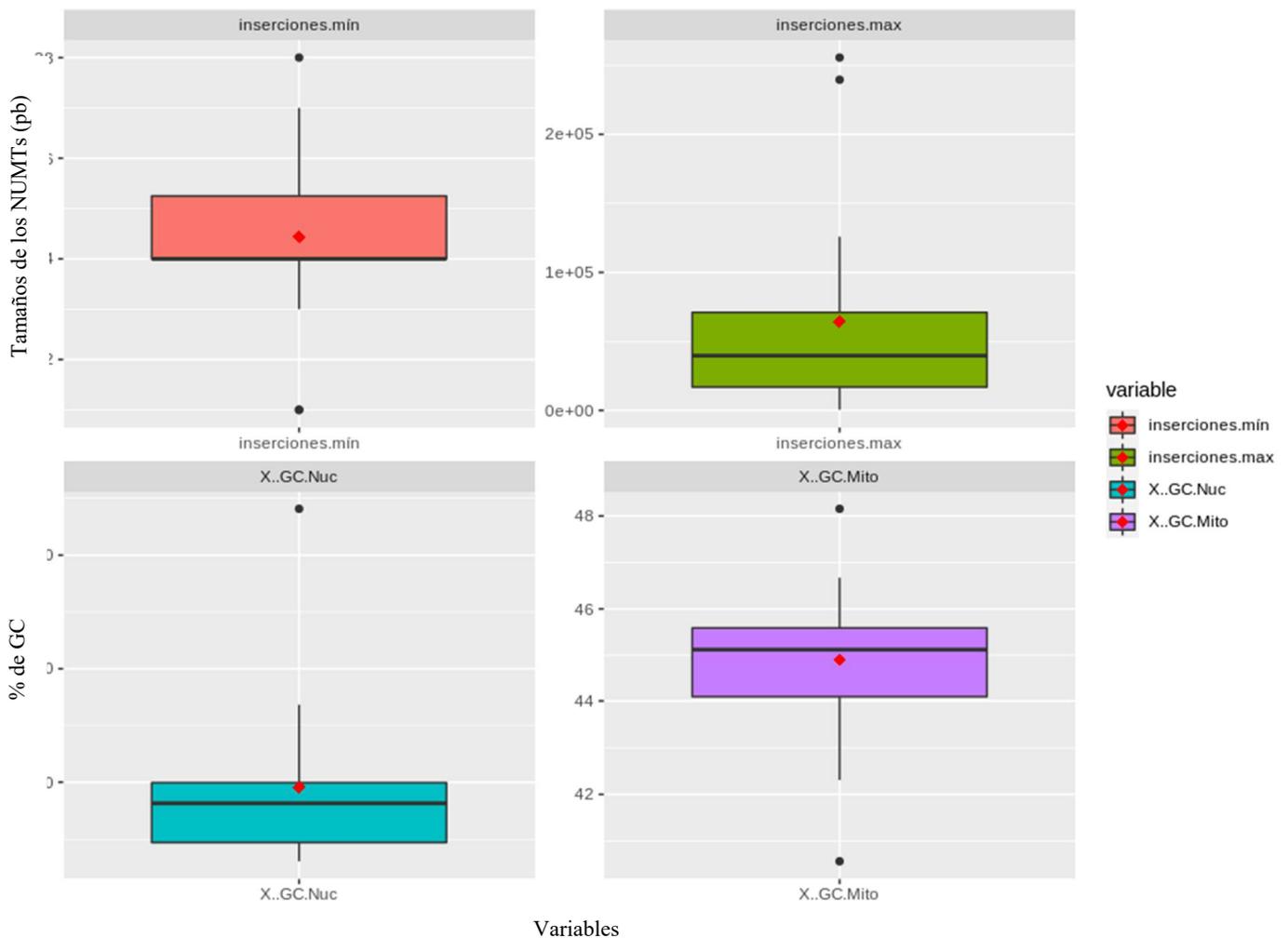
— Data Summary —						
variable	media	Desviación estándar	p0	p50	p100	histograma
Tamaño genoma nuclear	2013787292	5389655079	111100715	479712609	22104357184	
Tamaño genoma mitocondrial	1299466	2593139	15758	468388	10397121	
Tamaño genoma plastídico	151898	20838	121531	155654	203828	



**Figura 6. Boxplots de los tamaños de los genomas nucleares, mitocondriales y plastídicos de las 16 especies vegetales seleccionadas.** De izquierda a derecha nos encontramos el gráfico de los tamaños de los genomas nucleares representado en naranja (Tamaño.Nuc), de color verde el gráfico de los tamaños de los genomas mitocondriales (Tamaño.Mito) y finalmente el gráfico de los tamaños de los genomas plastídicos representado en azul (Tamaño.Plastid). En rojo y diseñada con forma de rombo se puede distinguir la media calculada en cada caso, las líneas negras horizontales representan la mediana y en negro encontramos los puntos con desplazamiento.

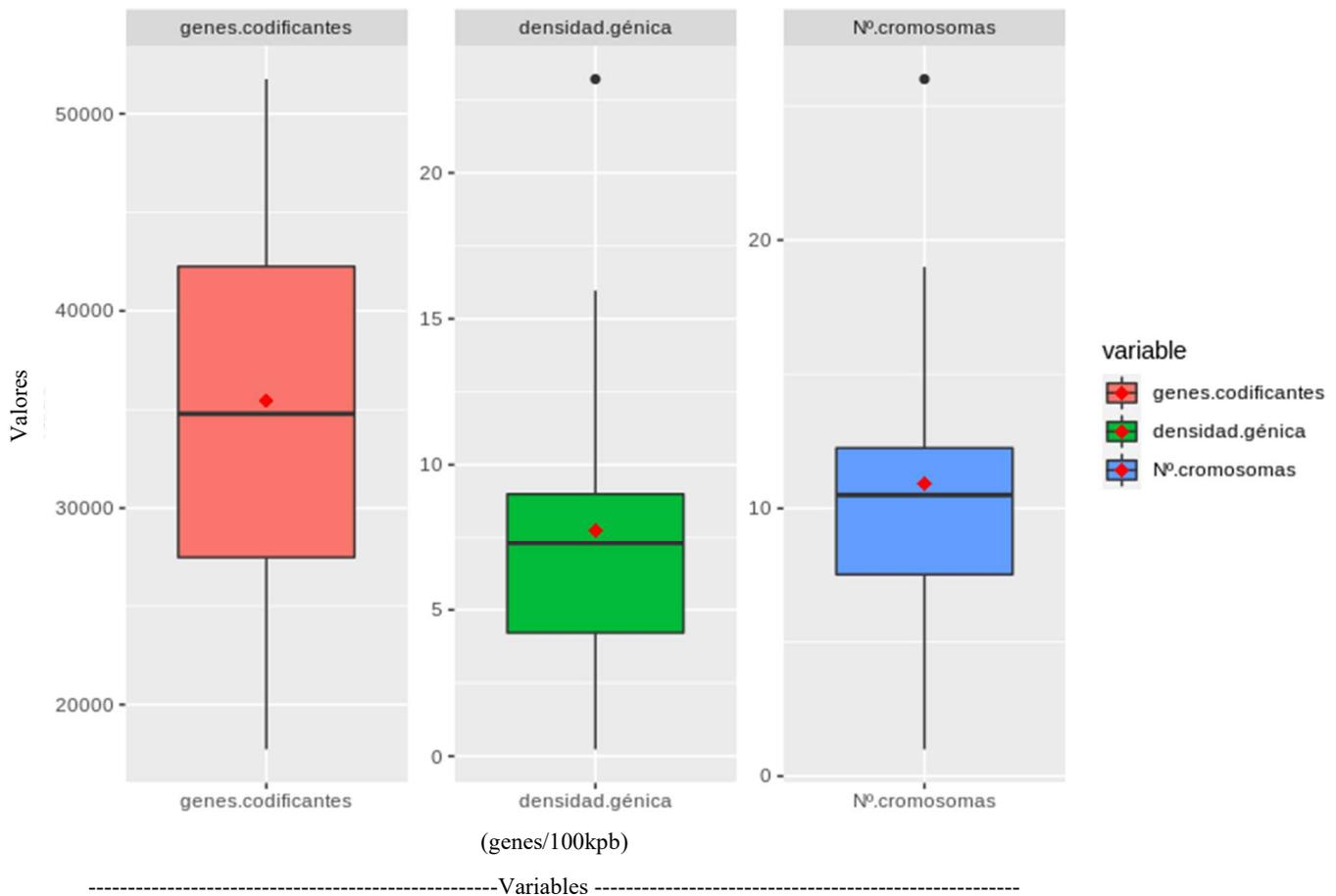
#### 4.1.2. Análisis de los parámetros asociados a los genomas

Seguidamente, se llevó a cabo el estudio de los NUMTs de tamaños mínimos, NUMTs de tamaños máximos, porcentaje de guanina (G) y citosina (C) en el núcleo y en la mitocondria, siendo representando mediante boxplots (**figura 7**). En los gráficos podemos destacar *P. taeda* (38 pb) a diferencia de *C. reinhardtii* (31pb) en cuanto a los NUMTs de menor tamaño (mínimo). *M. biondii* (25595 pb) siendo la especie con el mayor NUMT (máximo), *C. reinhardtii* con un porcentaje de guanina y citosina en el núcleo de 64% y, finalmente, *N. nucifera* (48,15%) y *P. patens* (40,56%), constituyendo los dos extremos en la distribución del porcentaje de guanina y citosina en la mitocondria.



**Figura 7. Boxplots de los tamaños mínimos de NUMTs, tamaños máximos de NUMTs, porcentaje de guanina y citosina en el núcleo y porcentaje de guanina y citosina en la mitocondria de las 16 especies seleccionadas.** El gráfico de los tamaños mínimos de NUMTs representado en naranja (*inserciones.min*), de color verde el gráfico de los tamaños máximos de NUMTs (*inserciones.max*), de color azul el gráfico de porcentaje de guanina y citosina en el núcleo (*X..GC.Nuc*) y finalmente el gráfico de los tamaños de porcentaje de guanina y citosina en la mitocondria (*X..GC.Mito*). En rojo y diseñada con forma de rombo se puede distinguir la media calculada en cada caso, las líneas negras horizontales representan la mediana y en negro encontramos los puntos con desplazamiento.

Para finalizar con el estudio de los genomas, se ha realizado un análisis de los genes codificantes de cada una de las especies, encontrando una distribución estadística simétrica entre los resultados (**figura 8**). También se estudió la densidad génica presente en los genomas, de los cuales destaca *A. thaliana* con una densidad génica de 23,21 genes/100 kb. Finalmente se realizó un análisis del número de cromosomas de las especies, de entre las cuales podemos resaltar al musgo *P. patens* conteniente de 26 cromosomas a diferencia de especies como *M. biondii* que solo presenta 1.



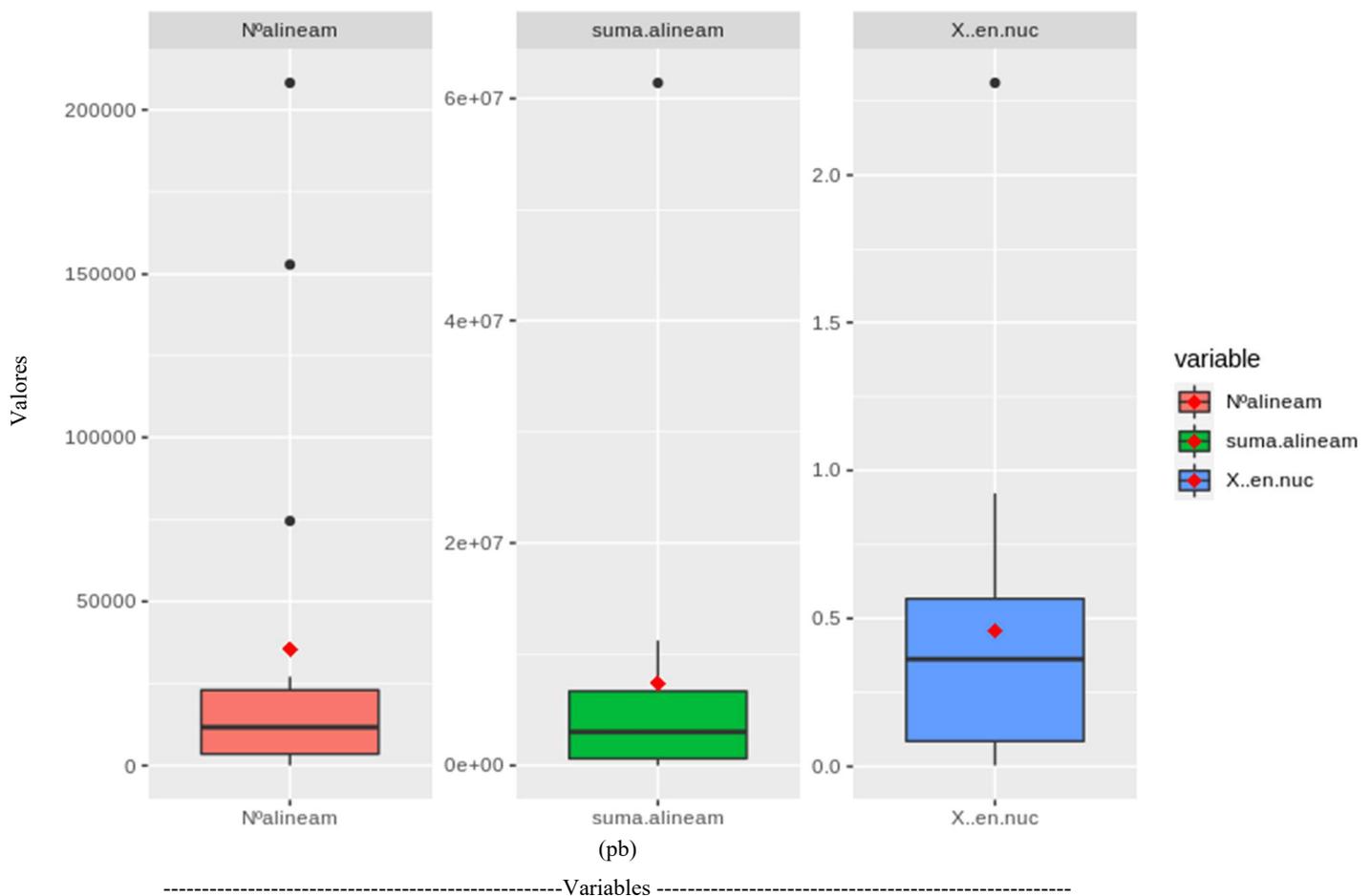
**Figura 8. Boxplots de los genes codificantes, densidad génica y número de cromosomas en los genomas nucleares de las 16 especies seleccionadas.** En rojo y diseñada con forma de rombo se puede distinguir la media calculada en cada caso, las líneas negras horizontales representan la mediana y en negro encontramos los puntos con desplazamiento.

#### 4.1.3. Análisis de los resultados BLAST

Una vez analizados los tamaños de los genomas nucleares, mitocondriales y plastídicos de las especies, se ha llevado a cabo el análisis de las demás variables obtenidas mediante el uso de la herramienta BLAST. En primer lugar, se analizó el número de alineamientos de ADN mitocondrial con el genoma nuclear, en el cual se pueden observar, representados en el boxplot de la **figura 9**, tres especies que representan desviaciones significativas, siendo de mayor a menor *A. trichopoda* (208181), *Vi. vinífera* (152929) y *P. taeda* (74532) a diferencia de *P. patens* que presenta 33 alineamientos, siendo el menor valor. En cuanto a la suma de alineamientos o longitud total de los alineamientos de ADN mitocondrial, de la cual destaca de nuevo *A. trichopoda* con una longitud de 61403592pb, en contraste con *C. reinhardtii*, con una longitud de 3681. Finalmente, se analizó el porcentaje de ADN mitocondrial en el núcleo, destacando a *Vitis 20inífera* con un 2,031% y, en el otro extremo, a *C. reinhardtii* con un 0,0033%.

Es interesante señalar que, tras realizar un análisis paralelo del porcentaje de ADN mitocondrial en el genoma plástidico se observó, en la especie *A. trichopoda*, un porcentaje de 119,5%, el cual se podría explicar debido a que su genoma mitocondrial está formado por 5 cromosomas circulares autónomos que miden 3179, 244, 187, 137, and 119kb, sumando un total de 3866 kb. Todos los cromosomas mitocondriales presentan una composición similar de bases, conteniendo un 45-47% de guanina y citosina. El genoma mitocondrial de *A. trichopoda* acoge una colección de secuencias de otros organismos que corresponden con cerca de 6 genomas mitocondriales de musgos, angiospermas y algas verdes. Además, presenta un gran contenido (138 kb) de ADN plástidico en su genoma mitocondrial (Rice et al., 2013).

Por tanto, tras realizar un BLAST entre el genoma mitocondrial de *A. trichopoda* y su genoma plástidico, al presentar, el primero, 5 cromosomas circulares, los cuales a su vez presentan un gran contenido de ADN plástidico como se ha indicado anteriormente, se van a alinear las secuencias repetidas veces en una misma región del genoma plástidico, dando lugar a un porcentaje de ADN mitocondrial muy elevado.



**Figura 9. Boxplots de los datos obtenidos durante el análisis de las variables seleccionadas de las 16 especies seleccionadas.** En naranja (*N°alineam*) se muestra el gráfico que representa los resultados del análisis del número de NUMTs, el verde (*suma.alineam*) corresponde con el análisis de la suma total de los tamaños de los NUMTs y

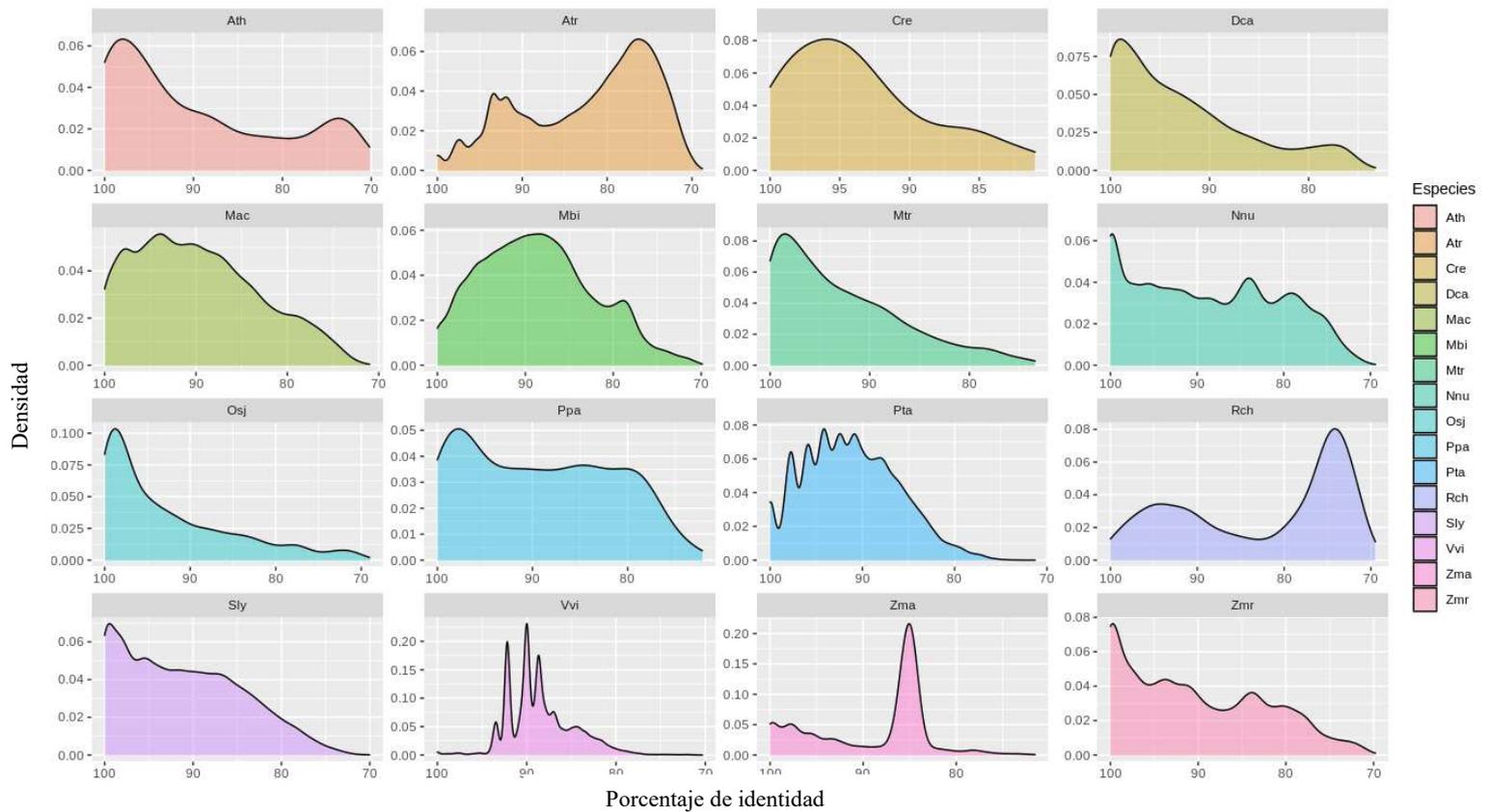
en azul (*X.en.nuc*) el porcentaje de ADN mitocondrial en el genoma nuclear. En rojo y diseñada con forma de rombo se puede distinguir la media calculada en cada caso, las líneas negras horizontales representan la mediana y en negro encontramos los puntos con desplazamiento.

#### 4.1.3.1. Distribución de los NUMTs a lo largo de un tiempo evolutivo

En cuanto a la distribución de los NUMTs a lo largo de un tiempo evolutivo, se podrían obtener dos patrones diferentes, el primero se distinguiría por un decaimiento gradual de los NUMTs a lo largo del tiempo evolutivo, indicando la continuidad del proceso de inserción de los fragmentos de ADN mitocondrial. El segundo patrón que cabría esperar sería aquel que presenta explosiones de transferencia génica en tiempos concretos de la evolución representados gráficamente por picos. Teniendo en cuenta las premisas anteriores, se llevó a cabo el estudio de la relación entre la densidad y los porcentajes de identidad obtenidos del BLAST con el objetivo de estudiar el patrón de los NUMTs, ya que utilizamos el porcentaje de identidad como aproximación al tiempo evolutivo.

Los picos que se muestran en las gráficas, como se ha indicado previamente, representan una concentración de eventos de transferencia de ADN mitocondrial, el cual se observa claramente en el caso de *Zea mays*, estas transferencias de material genético se pueden deber a, como se ha explicado en el apartado 1.2.1, un periodo en el que se la planta es sometida a situaciones de estrés como puede ser su domesticación, ya que estos fragmentos representan una fuente de mutación y por tanto de evolución. Esta afirmación sugiere que estos NUMTs les permitirían generar tolerancia ante estas situaciones de estrés. Podemos destacar que, tanto *A. thaliana*, *C. reinhardtii*, *D. carota*, *M. truncatula*, *M. acuminata*, *O. sativa*, *P. patens*, *S. lycopersicum*, *N. Nucifera* y *Z. marina* han sufrido transferencias recientes de ADN mitocondrial ya que, como se observa en la gráfica (**figura 10**), presentan un alto contenido de alineamientos con un porcentaje de identidad de entre 90-100%, lo que indica que la maquinaria genética celular no ha tenido el suficiente tiempo como para que se hayan creado mutaciones que disminuyan el porcentaje de identidad de las mismas. En contra posición, podemos observar cómo en *R. chinensis*, la gran mayoría de los alineamientos se concentran en la franja de 70-80% del porcentaje de identidad, o *Zostera Marina*, cuyos alineamientos se encuentran entre 90-80%. Aun así, todas las especies presentan fragmentos con grandes porcentajes de identidad ya que un 70% de identidad es un valor muy significativo.

Los resultados concluyeron con la falta de identidad entre las distintas especies ya que muestran patrones diferenciales, lo que sugiere que cada especie está sujeta a distintas restricciones para albergar el porcentaje de ADN mitocondrial. Además, se puede observar cómo el porcentaje de identidad de los fragmentos disminuye en función que aumenta el tiempo evolutivo, llegando a ser imposible el reconocimiento de los fragmentos iniciales.



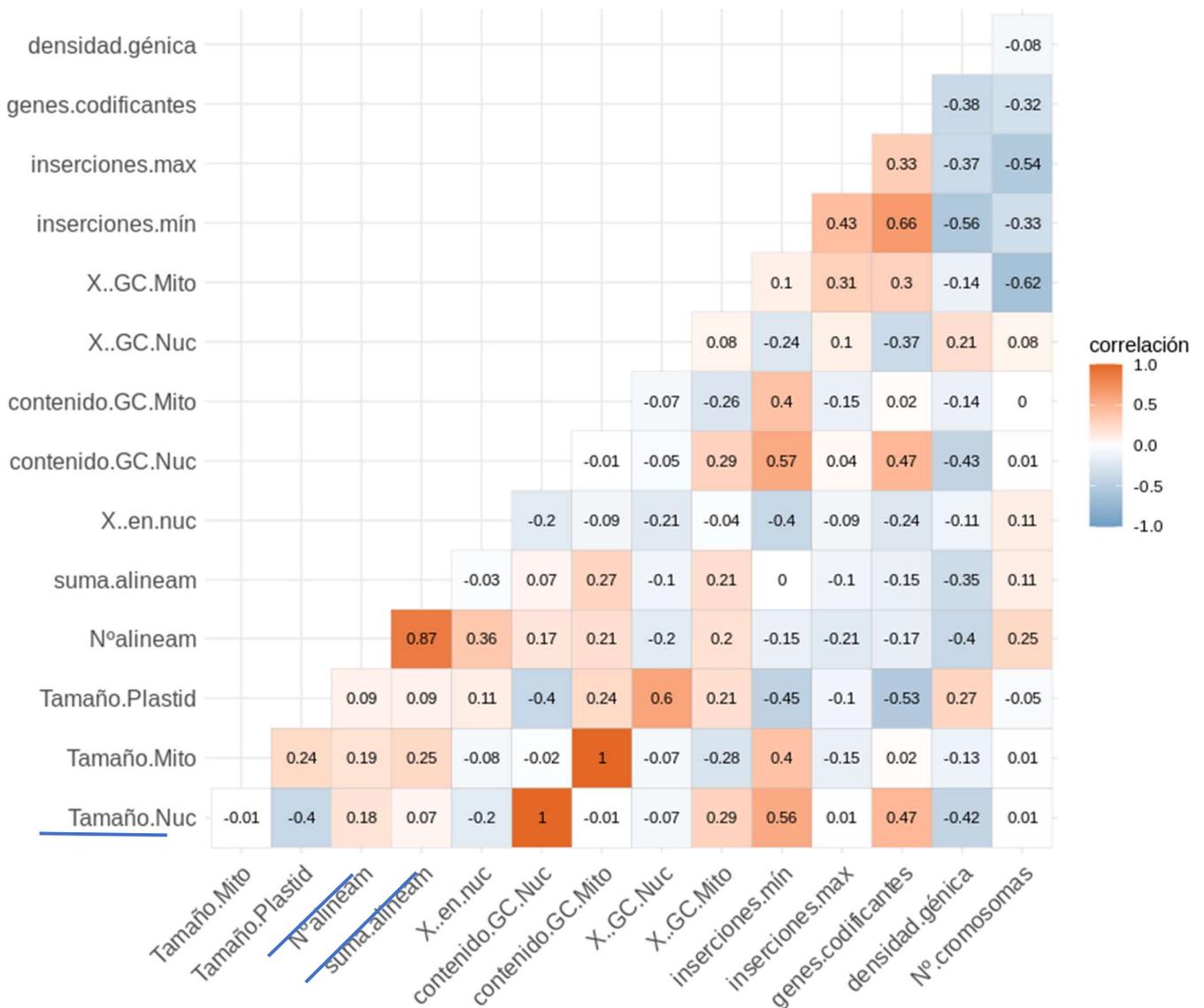
**Figura 10. Gráficas de densidad por especie del porcentaje de identidad de los NUMTs.** El porcentaje de identidad representa una aproximación al tiempo evolutivo. Las abreviaturas utilizadas en la leyenda de la gráfica corresponden de la siguiente forma con las especies vegetales: Ath: *Arabidopsis thaliana*, Atr: *Amborella thricopoda*, Cre: *Chlamydomonas reinhardtii*, Dca: *Daucus carota*, Mac: *Musa acuminata*, Mbi: *Magnolia biondii*, Mtr: *Medicago truncatula*, Nnu: *Nelumbo nucifera*, Osj: *Oryza sativa*, Ppa: *Physcomitrium patens*, Pta: *Pinus taeda*, Rch: *Rosa chinensis*, Sly: *Solanum lycopersicum*, Vvi: *Vitis vinifera*, Zma: *Zea mays*, Zmr: *Zostera marina*.

## 4.2. Análisis estadísticos

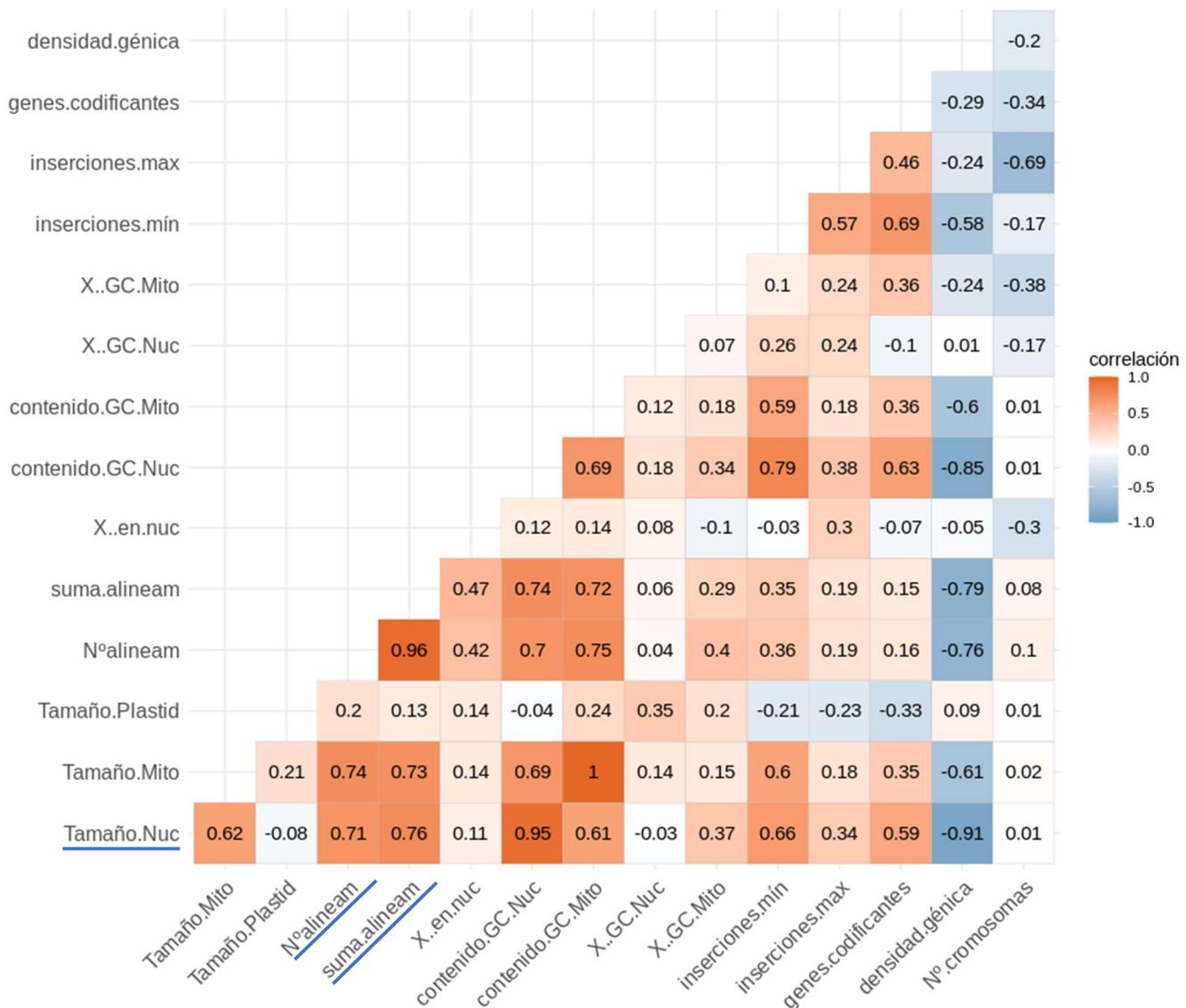
### 4.2.1. Correlograma de variables seleccionadas

Una vez estudiada la distribución de los porcentajes de identidad a lo largo de un tiempo evolutivo de las especies utilizadas y descartada la posibilidad de encontrar un patrón común se llevó a cabo un estudio de correlación entre las distintas variables que fueron previamente seleccionadas con el fin de encontrar correlaciones significativas. Este ensayo se efectuó mediante el desarrollo de dos correlogramas. En el primer correlograma (**figura 10**), se utilizó el coeficiente estadístico de Pearson con el que se obtuvo una correlación relativamente baja en prácticamente todas las combinaciones. En concreto, en cuanto al tamaño nuclear y a la longitud total de los NUMTs (0,07), y entre el tamaño mitocondrial y, de nuevo, la longitud total de los NUMTs (0,25), indicado en la **figura 11** como “suma.alineam”.

Como no conocemos si la relación entre las variables es lineal, se realizó un segundo correlograma (**figura 12**) usando el coeficiente estadístico de Spearman, se observó un aumento significativo de las correlaciones indicadas anteriormente, siendo de 0,76 entre la suma o longitud total de los NUMTs y el tamaño nuclear, y de 0,73 con el tamaño mitocondrial. Además, podemos señalar que ambos tamaños correlacionan positivamente de forma significativa con el número total de NUMTs.



**Figura 11. Correlograma de las variables seleccionadas usando el coeficiente estadístico de Pearson. Se encuentran subrayadas las variables del tamaño del genoma nuclear (Tamaño.Nuc), longitud total de los NUMTs (suma.alineam) y número total de NUMTs (Nºalineam).**



**Figura 12. Correlograma de las variables seleccionadas usando el coeficiente estadístico de Spearman.** Se encuentran subrayadas las variables del tamaño del genoma nuclear (Tamaño.Nuc), longitud total de los NUMTs (suma.alineam) y número total de NUMTs (N°alineam).

#### 4.2.2. Correlación positiva entre tamaño de los genomas nucleares y mitocondriales y la longitud total de los NUMTs

Como se ha señalado en el apartado anterior, observamos una correlación positiva entre la suma de los tamaños de los alineamientos mitocondriales y los genomas, tanto mitocondriales como nucleares pero el valor de estas correlaciones varía en función del coeficiente estadístico seleccionado. El coeficiente de Pearson es paramétrico, es decir, infiere sus resultados a la población real, haciendo necesario que la distribución de la muestra estadística se parezca a la distribución real, que exista

normalidad. En cambio, el coeficiente de Spearman es no paramétrico, la distribución de la muestra no se ajusta a una distribución conocida, no debe existir normalidad.

Para verificar la elección del coeficiente de Spearman con respecto al de Pearson, los pares de variables seleccionados se sometieron al Test de Shapiro–Wilk el cual contrasta la normalidad de nuestro conjunto de datos. El test expone la hipótesis nula en la que los datos están distribuidos normalmente y, en el caso de que se presentase un valor  $p$ , o “p-value” en inglés, por debajo de 0,05, este rechaza la hipótesis de normalidad de las variables y, por tanto, estas serían consideradas asimétricas lo que nos indicaría que se debería usar una correlación no paramétrica; Spearman (un valor considerablemente pequeño del estadístico  $W$  también rechaza la hipótesis nula). Como se puede observar en la **tabla 3**, todas las variables analizadas presentan un “p-value” por debajo de 0,05, por lo que rechazan la hipótesis nula y nos asegura que debemos analizar las correlaciones entre las variables mediante el coeficiente de Spearman.

**Tabla 3.** Resumen de los resultados obtenidos en el test de Shapiro-Wilk.

— Resumen de los resultados —		
Variable	Estadístico W	p-value
Tamaño genoma nuclear	0,35315	1,604e-0,7
Tamaño genoma mitocondrial	0.4866	1.645e-06
Número de alineamientos	0,62036	2,536e-05
Suma total de alineamientos	0.48886	1.716e-06

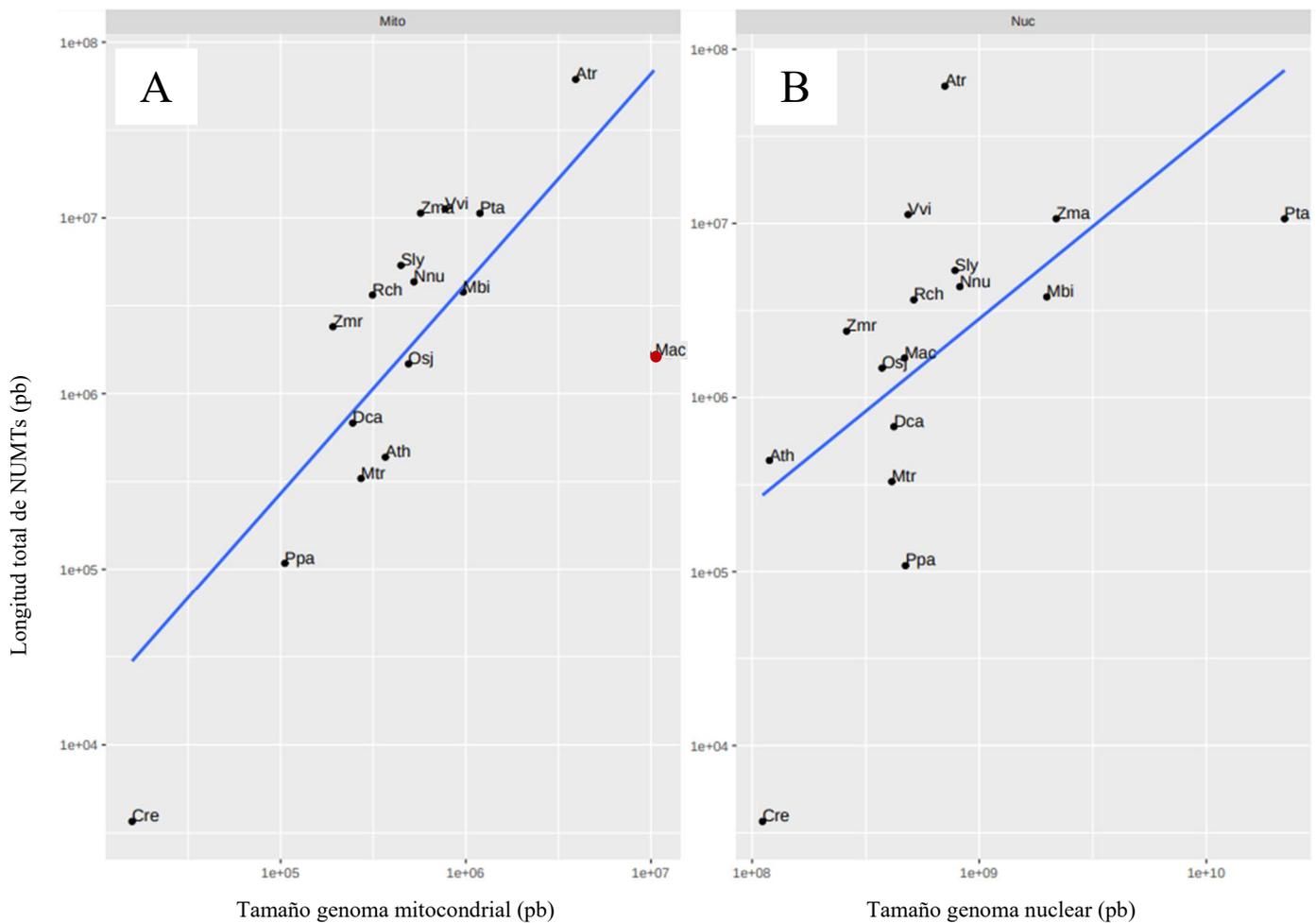
Los resultados referentes al genoma nuclear, representados en las **figuras 13B y 14D**, concuerdan con los obtenidos en un estudio realizado en 2020 (Zhang et al., 2020), en el que se utilizaron 200 especies, las cuales no incluyen a *M. biondii*, *M. acuminata*, *P. patens* ni a *R. chinensis* utilizadas en nuestro estudio. Sin embargo, a diferencia del ensayo de 2020 (Zhang et al., 2020), en este estudio se ha podido observar, además, la correlación positiva con el tamaño del genoma mitocondrial (**figuras 13A y 14C**). Ambos ensayos difieren con los resultados obtenidos en un estudio previo de 2010 (Hazkani-Covo et al., 2010), el cuál señalaba la falta de correlación entre las variables, posiblemente debido al tamaño de la muestra estadística utilizada, 7 genomas, entre los que se encuentra *C. reinhardtii*, única alga verde incluida que, como hemos observado en el apartado 4.1.2, presenta un número de alineamientos muy bajo y una suma de los alineamientos muy baja a diferencia de las demás especies utilizadas, lo que sesgaría el rango de correlación.

A pesar de observar una correlación positiva durante el ensayo con las 16 especies vegetales, se ha podido demostrar, a través de análisis comparativos del coeficiente de correlación, un aumento de

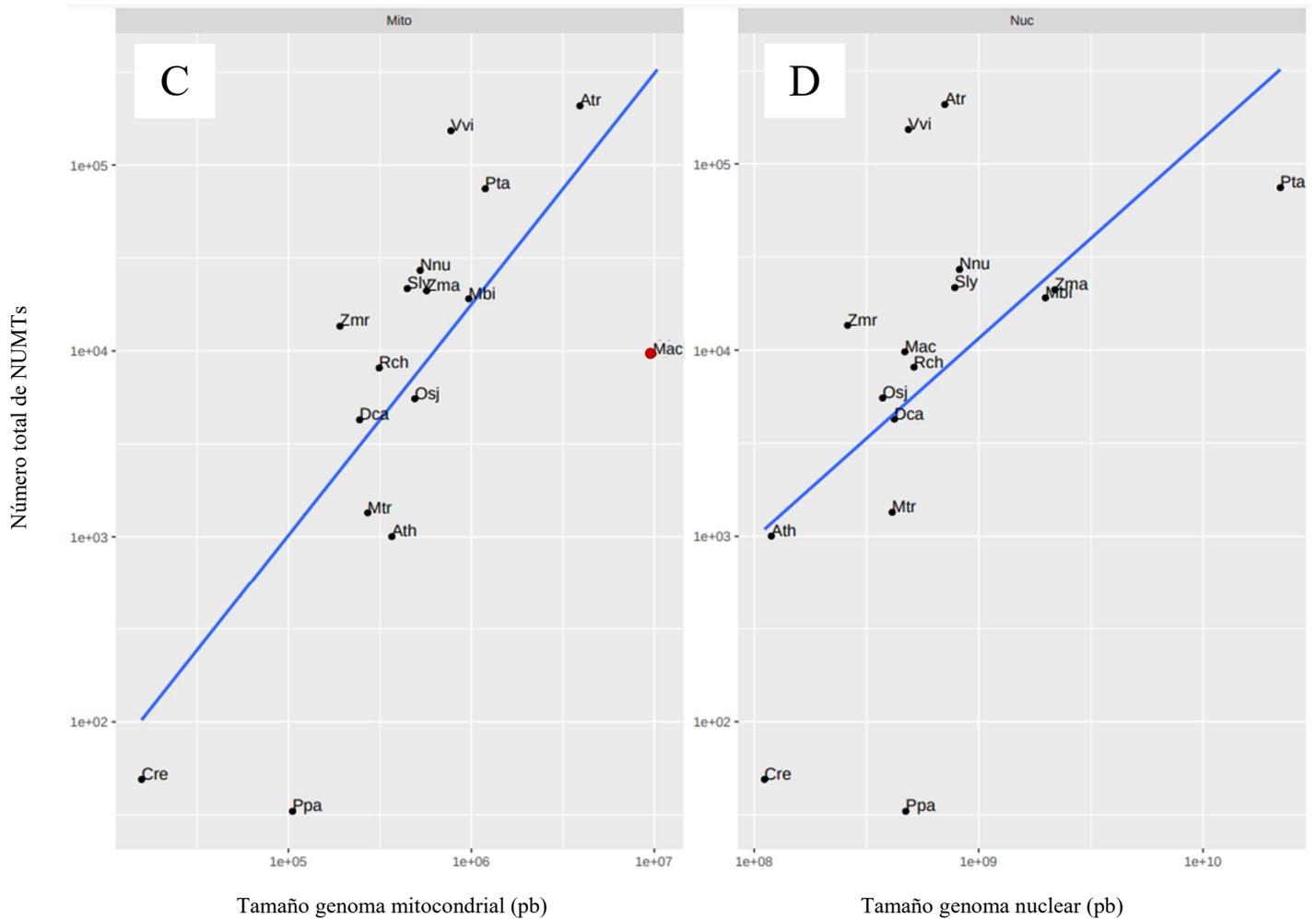
este (como se observa en la **tabla 4**) en los ensayos del genoma mitocondrial cuando han sido suprimidos los datos correspondientes a la especie *M. acuminata*, la cual se puede observar marcada en rojo y denominada como “Mac” en las **figuras 13A y 14C**. Del mismo modo que se ha comentado anteriormente en el caso de *C. reinhardtii*, se podría considerar a *M. acuminata* como un “outlier” ya que afecta al sesgo de los análisis realizados en el genoma mitocondrial.

**Tabla 4.** Resultados del análisis de correlación antes y después de excluir a *M. acuminata*.

	Análisis de correlación incluyendo a <i>M. acuminata</i>		Análisis de correlación sin incluir a <i>M. acuminata</i>	
	p - value	Rho	p - value	Rho
Longitud total de alineamientos mitocondriales	0.001927	0.7294118	8.566e-06	0.8571429
Número de alineamientos mitocondriales	0.001404	0.7441176	0.0001043	0.8392857



**Figura 13. Gráficos de dispersión entre tamaños de los genomas y longitud total de los NUMTs (pb).** A. Tamaños de los genomas mitocondriales marcada en rojo se encuentra la especie *M. acuminata*, denominada como “Mac”. B. Tamaños de los genomas nucleares. En ambos gráficos se encuentra marcada en azul la línea de regresión.



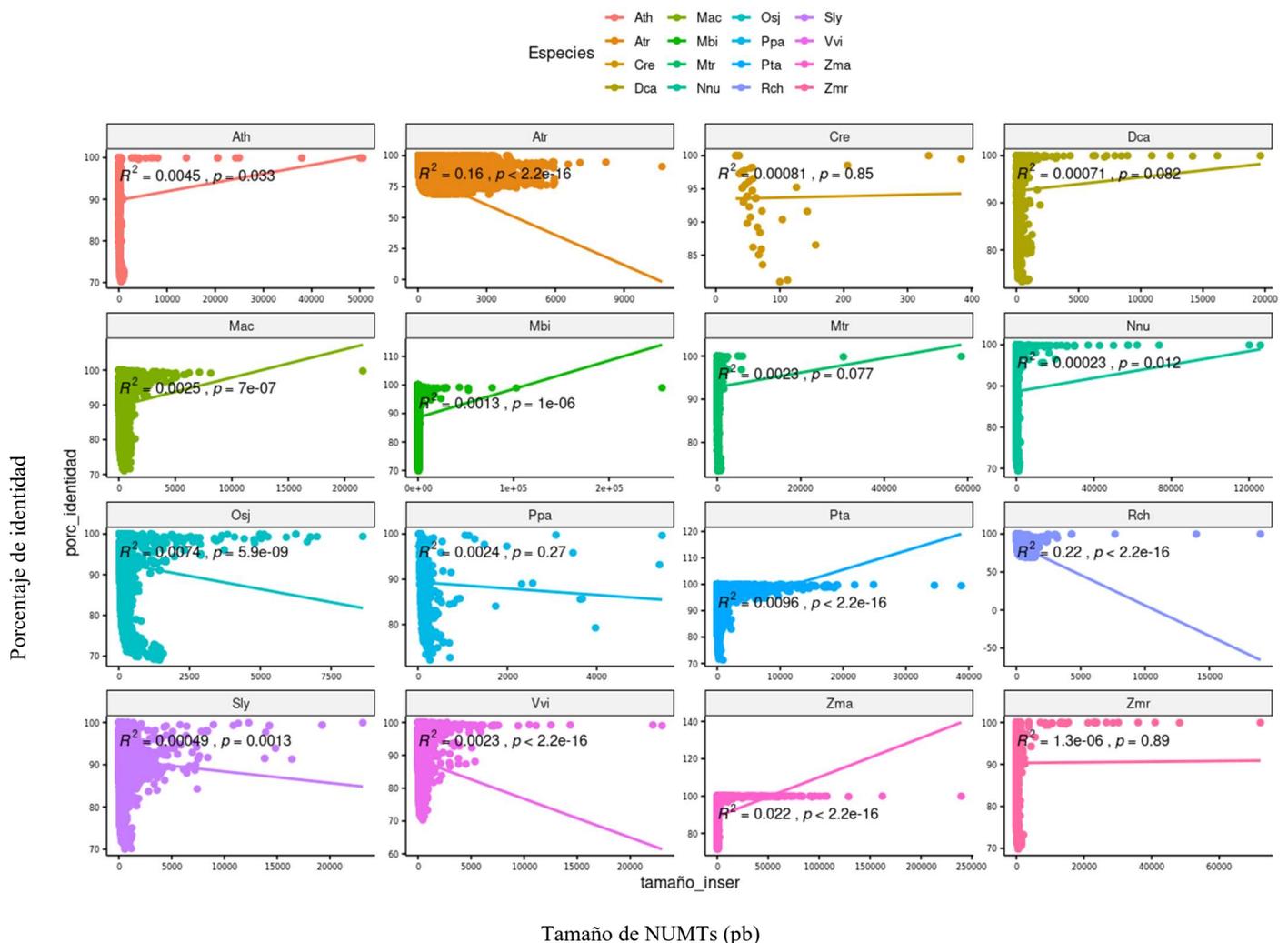
**Figura 14. Gráficos de dispersión entre tamaños de los genomas y el número total de NUMTs.** C. Tamaños de los genomas mitocondriales, marcada en rojo se encuentra la especie *M. acuminata*, denominada como “Mac”. D. Tamaños de los genomas nucleares. En ambos gráficos se encuentra marcada en azul la línea de regresión.

#### 4.2.3. La correlación entre el tamaño de los NUMTs y el porcentaje de identidad revela patrones diferenciales en las especies vegetales estudiadas

Como se indicó en el apartado 1.2.2., una vez insertados los fragmentos de ADN mitocondrial en el genoma nuclear se producen diferentes modificaciones, algunas de las cuales intervienen en las variaciones de las longitudes de los fragmentos, las cuales disminuyen a lo largo del tiempo evolutivo (Richly & Leister, 2004; Yuan et al., 2002). Estas son la fragmentación, que actúa directamente en la división de los fragmentos a través de la actuación de los transposones, y la mutación, la cual actúa

indirectamente mediante el aumento de mutaciones en los fragmentos que disminuye el porcentaje de identidad, lo que conduce a la pérdida del reconocimiento de la identidad individual de un fragmento de ADN, y la aparición de fragmentos más pequeños derivados del mutante. Se han llevado a cabo diferentes ensayos que demuestran una correlación positiva entre el porcentaje de identidad y el tamaño de los NUPTs con especies vegetales como *Asparagus officinalis* (S.-F. Li et al., 2019), los cuales tienen como premisa que presentan unos tamaños iniciales relativamente grande y comparables. Se ha realizado un análisis, a través de la representación gráfica mediante un gráfico de dispersión (**figura 15**), con la finalidad de estudiar la correlación entre la similitud o porcentaje de identidad de los NUMTs y sus respectivas longitudes.

Tras realizar el ensayo, se ha observado un patrón discordante al esperado ya que de los análisis que han resultado significativos (11 análisis de especies han obtenido un p-value inferior a 0,05) solo seis especies vegetales presentan una correlación positiva: *A. thaliana*, *M. acuminata*, *M. biondii*, *N. nucifera*, *P. taeda* y *Z. Mays*. A diferencia de *A. trichopoda*, *O. sativa*, *R. chinensis*, *S. lycopersicum* y *V. vinifera*, las cuales presentan una correlación negativa. Estas observaciones no nos permiten concluir ni descartar que las especies presenten los mismos mecanismos de degradación de los NUMTs, pero sí nos sugieren la posibilidad de rebatir la premisa de que todos los NUMTs, inicialmente, presentan tamaños comparables.



**Figura 15. Gráficos de dispersión por especie del tamaño de los NUMTs y el porcentaje de identidad.** Las abreviaturas utilizadas en la leyenda de la gráfica corresponden de la siguiente forma con las especies vegetales: Ath: *Arabidopsis thaliana*, Atr: *Amborella thricopoda*, Cre: *Chlamydomonas reinhardtii*, Dca: *Daucus carota*, Mac: *Musa acuminata*, Mbi: *Magnolia biondii*, Mtr: *Medicago truncatula*, Nnu: *Nelumbo nucifera*, Osj: *Oryza sativa*, Ppa: *Physcomitrium patens*, Pta: *Pinus taeda*, Rch: *Rosa chinensis*, Sly: *Solanum lycopersicum*, Vvi: *Vitis vinifera*, Zma: *Zea mays*, Zmr: *Zostera marina*. En cada gráfica se puede observar el coeficiente de determinación ( $R^2$ ) y el  $p$ -value ( $p$ ) del análisis realizado.

## 5. Conclusiones

Las conclusiones principales obtenidas durante el desarrollo del estudio fueron las siguientes:

1. Se obtuvieron los genomas mitocondriales, plastídicos y nucleares de 16 especies en base a su representatividad en los linajes de plantas y a su grado de calidad.
2. Tras el estudio de los parámetros asociados a los genomas, destaca el tamaño de los mismos. Podemos señalar la gran heterogeneidad que presentan los tamaños de los genomas mitocondriales a diferencia de los genomas plastídicos, existiendo una diferencia de casi cuatro órdenes de magnitud entre el mayor valor obtenido, 10397121 pb correspondiente a *M. acuminata*, y el valor más bajo obtenido, 15758 pb de *C. reinhardtii*.
3. Gracias a la determinación del porcentaje de ADN mitocondrial en el núcleo, se ha podido confirmar la existencia de la transferencia de ADN mitocondrial al genoma nuclear en todas las especies de plantas estudiadas, debido a la presencia de NUMTs en las mismas.
4. Se verificó la correlación positiva entre el tamaño de los genomas nucleares y mitocondriales y la longitud total de los NUMTs. También se demostró la existencia de correlación positiva entre el tamaño de los genomas nucleares y mitocondriales y el número total de NUMTs.
5. De las 11 especies vegetales que resultaron significativas en el análisis de correlación entre el porcentaje de identidad y el tamaño de los NUMTs, seis de ellas mostraban correlación positiva, lo cual no permite confirmar ni descartar la presencia de mecanismos comunes responsables de la degradación de NUMTs pero sí rebatir la premisa de que todos los fragmentos, inicialmente, presentan tamaños comparables.
6. No se han encontrado patrones comunes entre las 16 especies en las curvas de densidad de los porcentajes de identidad de los NUMTs. Sin embargo, se ha evidenciado que existe una concentración de los eventos de transferencia de ADN mitocondrial en puntos concretos de la evolución.

Este estudio se enfoca en la obtención y posterior análisis de parámetros asociados a los genomas que se encuentren relacionados con los mecanismos de transferencia, retenimiento o eliminación de ADN mitocondrial en el genoma nuclear. Los resultados obtenidos podrán ser utilizados como material preliminar, o de partida, para la realización de estudios posteriores, así como para el establecimiento de modelos de regresión con los que predecir los porcentajes de ADN mitocondrial en los genomas nucleares de cualquier especie vegetal.

## 6. Bibliografía

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Bock, R., & Timmis, J. N. (2008). Reconstructing evolution: Gene transfer from plastids to the nucleus. In *BioEssays* (Vol. 30, Issue 6, pp. 556–566). John Wiley and Sons Inc. <https://doi.org/10.1002/bies.20761>

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. <https://doi.org/10.1186/1471-2105-10-421>

Cullis, C. A., Vorster, B. J., van der Vyver, C., & Kunert, K. J. (2009). Transfer of genetic material between the chloroplast and nucleus: How is it related to stress in plants? In *Annals of Botany* (Vol. 103, Issue 4, pp. 625–633). <https://doi.org/10.1093/aob/mcn173>

Droc, G., Larivière, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., Lengelle, J., Baurens, F.-C., Cenci, A., Pitollat, B., D'Hont, A., Ruiz, M., Rouard, M., & Bocs, S. (2013). The Banana Genome Hub. *Database*, 2013. <https://doi.org/10.1093/database/bat035>

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. In *PLoS Genetics* (Vol. 6, Issue 2). <https://doi.org/10.1371/journal.pgen.1000834>

Kleine, T., Maier, U. G., & Leister, D. (2009). DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. *Annual Review of Plant Biology*, *60*, 115–138. <https://doi.org/10.1146/annurev.arplant.043008.092119>

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. <https://doi.org/10.1093/molbev/msx116>

Kassambara A. (2020). ggpubr: 'ggplot2' Based. Publication Ready Plots. R package version 0.4.0. <https://CRAN.R-project.org/package=ggpubr>

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. <https://doi.org/10.1093/nar/gkab301>

Li, S.-F., Li, J.-R., Wang, J., Dong, R., Jia, K.-L., Zhu, H.-W., Li, N., Yuan, J.-H., Deng, C.-L., & Gao, W.-J. (2019). Cytogenetic and genomic organization analyses of chloroplast DNA invasions in the nuclear genome of *Asparagus officinalis* L. provides signatures of evolutionary complexity and informativity in sex chromosome evolution. *BMC Plant Biology*, *19*(1), 361. <https://doi.org/10.1186/s12870-019-1975-8>

Li, X., Fang, C., Zhao, J. P., Zhou, X. Y., Ni, Z., & Niu, D. K. (2020a). Desiccation does not drastically increase the accessibility of exogenous DNA to nuclear genomes: Evidence from the frequency of endosymbiotic DNA transfer. *BMC Genomics*, *21*(1). <https://doi.org/10.1186/s12864-020-06865-8>

Li, X., Fang, C., Zhao, J. P., Zhou, X. Y., Ni, Z., & Niu, D. K. (2020b). Desiccation does not drastically increase the accessibility of exogenous DNA to nuclear genomes: Evidence from the frequency of endosymbiotic DNA transfer. *BMC Genomics*, *21*(1). <https://doi.org/10.1186/s12864-020-06865-8>

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., & Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. In *Trends in Genetics* (Vol. 23, Issue 12, pp. 597–601). <https://doi.org/10.1016/j.tig.2007.08.016>

Puertas, M. J., & González-Sánchez, M. (2020). Insertions of mitochondrial dna into the nucleus—effects and role in cell evolution. *Genome*, *63*(8), 365–374. <https://doi.org/10.1139/gen-2019-0151>

Richly, E., & Leister, D. (2004). NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular Biology and Evolution*, *21*(10), 1972–1980. <https://doi.org/10.1093/molbev/msh210>

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>

Sheppard, A. E., Ayliffe, M. A., Blatch, L., Day, A., Delaney, S. K., Khairul-Fahmy, N., Li, Y., Madesis, P., Pryor, A. J., & Timmis, J. N. (2008). Transfer of Plastid DNA to the Nucleus Is Elevated during Male Gametogenesis in Tobacco. *Plant Physiology*, *148*(1), 328–336. <https://doi.org/10.1104/pp.108.119107>

Smith, D. R., Crosby, K., & Lee, R. W. (2011). Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biology and Evolution*, *3*(1), 365–371. <https://doi.org/10.1093/gbe/evr001>

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., & Vandepoele, K. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, *50*(D1), D1468–D1474. <https://doi.org/10.1093/nar/gkab1024>

Vandepoele, K., van Bel, M., Richard, G., van Landeghem, S., Verhelst, B., Moreau, H., van de Peer, Y., Grimsley, N., & Piganeau, G. (2013). pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environmental Microbiology*, *15*(8), 2147–2153. <https://doi.org/10.1111/1462-2920.12174>

Wegrzyn, J. L., Staton, M. A., Street, N. R., Main, D., Grau, E., Herndon, N., Buehler, S., Falk, T., Zaman, S., Ramnath, R., Richter, P., Sun, L., Condon, B., Almsaeed, A., Chen, M., Mannapperuma, C., Jung, S., & Ficklin, S. (2019). Cyberinfrastructure to Improve Forest Health and Productivity: The Role of Tree Databases in Connecting Genomes, Phenomes, and the Environment. *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.00813>

Wickham, H. (2016). *Data Analysis* (pp. 189–201). [https://doi.org/10.1007/978-3-319-24277-4\\_9](https://doi.org/10.1007/978-3-319-24277-4_9)

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

Yuan, Q., Hill, J., Hsiao, J., Moffat, K., Ouyang, S., Cheng, Z., Jiang, J., & Buell, C. (2002). Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. *Molecular Genetics and Genomics*, 267(6), 713–720. <https://doi.org/10.1007/s00438-002-0706-1>

Zhang, G. J., Dong, R., Lan, L. N., Li, S. F., Gao, W. J., & Niu, H. X. (2020). Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. In *International Journal of Molecular Sciences* (Vol. 21, Issue 3). MDPI AG. <https://doi.org/10.3390/ijms21030707>