Erundina Ruiz Pérez

Cartografía mediante secuenciación de un mutante albino de *Arabidopsis thaliana*

Trabajo tutorizado por: Dra. Sara Jover Gil y Dr. Héctor Candela Antón
Departamento de Biología Aplicada, Área de Genética

Grado en Biotecnología
Facultad de Ciencias Experimentales
Universidad Miguel Hernández de Elche
Curso 2015-2016

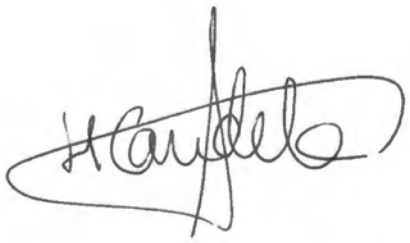
Dr. Héctor Candela Antón, Profesor Contratado Doctor del área de conocimiento de Genética de la Universidad Miguel Hernández de Elche, y

Dra. Sara Jover Gil, Profesora Asociada del área de conocimiento de Genética de la Universidad Miguel Hernández de Elche,

CERTIFICAN:

Que el Trabajo de Fin de Grado que lleva por título: **"Cartografía mediante secuenciación de un mutante albino de *Arabidopsis thaliana*"**, presentado por **Dña. Erundina Ruiz Pérez**, ha sido realizada bajo su dirección en el Instituto de Bioingeniería de la Universidad Miguel Hernández de Elche.

Para que conste y surta los efectos oportunos, firman el presente certificado en Elche, a 6 de Septiembre de 2016.



Fdo.: Dr. **Héctor Candela Antón**

Fdo.: **Dra. Sara Jover Gil**

ÍNDICE DE MATERIAS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Resumen

En este Trabajo de Fin de Grado hemos iniciado la clonación de un gen de *Arabidopsis thaliana* mediante una estrategia denominada cartografía mediante secuenciación. Este gen fue identificado gracias a una mutación recesiva, inducida mediante metanosulfonato de etilo (EMS), que causa albinismo y letalidad en el estadio de plántula. Para identificar el gen, clasificamos los individuos F_2 de una población cartográfica en dos grupos, según su fenotipo silvestre o mutante. Tras purificar el ADN genómico a partir de las plantas de cada grupo, ambas muestras fueron secuenciadas con la plataforma Illumina HiSeq 2500 de secuenciación masivamente paralela. El análisis bioinformático de las secuencias obtenidas en ambos grupos de plantas nos ha permitido identificar una mutación puntual que daña el sitio aceptor del procesamiento de los intrones del gen At2g04030, que codifica uno de los 7 miembros de la familia HSP90 de proteínas de choque térmico de *Arabidopsis thaliana*.

Palabras clave: cartografía mediante secuenciación; *Arabidopsis thaliana*; HSP90; mutantes albinos; resecuenciación de genomas

In this work, we have started a mapping-by-sequencing approach to clone a gene of *Arabidopsis thaliana*. This gene was first defined by a recessive, ethylmethane sulfonate (EMS)-induced mutation that causes albinism and lethality at the seedling stage. To identify the gene, we classified and pooled the individuals from an F_2 mapping population according to their phenotypes (wild-type or mutant). We purified genomic DNA from the plants of both pools, and the two samples were sequenced using the Illumina HiSeq 2500 next-generation sequencing platform. The bioinformatic analysis of the sequences from both pools has allowed us to identify a point mutation that damages a conserved residue at the acceptor site of an intron of the At2g04030 gene, which encodes a member of the HSP90 family of heat-shock proteins in the genome of *Arabidopsis thaliana*.

Keywords: mapping-by-sequencing; *Arabidopsis thaliana*; HSP90; albino mutants; genome re-sequencing

2. Introducción

Los experimentos de mutagénesis al azar han permitido identificar numerosos mutantes cuyos fenotipos han abierto una puerta hacia la comprensión, a nivel molecular, de los procesos que controlan el desarrollo de los organismos multicelulares (Pérez-Pérez *et al*., 2009). La identificación de los genes mutados se ha conseguido tradicionalmente mediante clonación posicional, una estrategia basada en el análisis de ligamiento genético entre la mutación a estudio y marcadores moleculares. La clonación posicional ha demostrado ser un método eficaz y factible incluso en especies con genomas grandes, como el maíz (Bortiri *et al*., 2006; Gallavotti *et al.*, 2015). Las nuevas tecnologías de secuenciación masivamente paralela (*massively parallel sequencing technologies*) permiten la secuenciación rápida y asequible de genomas completos (Metzker, 2010). En esta sección, se describen sucintamente algunas de las nuevas tecnologías de secuenciación de segunda y tercera generación (Thudi *et al.*, 2012; Zhang *et al.*, 2011), que han permitido el desarrollo de un nuevo método, la cartografía mediante secuenciación (*mapping-by-sequencing*), que está demostrando ser más rápido y tan poderoso como la clonación posicional convencional (Candela *et al*., 2015). En este trabajo, hemos aplicado este método a la identificación de la mutación responsable del fenotipo albino de un mutante de *Arabidopsis thaliana*.

2.1. Tecnologías de secuenciación de genomas

El método de Sanger ha sido usado para la secuenciación de moléculas de ADN desde su desarrollo en la década de 1970 (Sanger y Coulson, 1975). Este método, aún vigente, se basa en la interrupción de la síntesis de una molécula de ADN en posiciones consecutivas utilizando un procedimiento que permite determinar la identidad del nucleótido incorporado al extremo 3' de dicha molécula. Dicha interrupción se logra mediante la incorporación de 2',3'-didesoxinucleótidos trifosfato (familiarmente denominados "terminadores") a la mezcla de reacción. La identidad del nucleótido terminal puede lograrse mediante el marcaje de los terminadores por procedimientos isotópicos (con ^{32}P) o no isotópicos (con diversos fluoróforos).

En la última década se han desarrollado nuevos métodos de secuenciación, que han supuesto una revolución tecnológica (Metzker, 2010). Gracias a su elevado rendimiento y reducido coste, estos métodos han puesto la investigación genómica al alcance de los laboratorios más modestos. Tal rendimiento se alcanza gracias a la miniaturización de las reacciones de secuenciación, que permiten determinar en paralelo la secuencia de millones de fragmentos En algunas tecnologías de secuenciación masivamente paralela de segunda generación, la paralización de las reacciones se consigue al inmovilizar las reacciones de secuenciación individuales sobre un soporte sólido. El futuro previsible de las tecnologías de

secuenciación reside en el desarrollo y perfeccionamiento de las tecnologías de tercera generación. Se denomina así al conjunto de métodos que permiten determinar la secuencia de moléculas individuales en tiempo real, sin necesidad de realizar un paso de amplificación previo (*single-molecule real-time sequencing*; Yanhu *et al*., 2015).

2.1.1. Tecnologías de secuenciación de segunda generación

Se describen en este apartado las tres tecnologías de secuenciación de segunda generación más ampliamente utilizadas. Estas tecnologías se basan en diferentes principios. Dos de ellas, las desarrolladas por Roche e Illumina, son tecnologías basadas en la síntesis de la molécula secuenciada (*sequencing by synthesis*). Como el método de Sanger, la tecnología de Illumina utiliza nucleótidos terminadores para interrumpir la síntesis de la molécula de ADN. Sin embargo, Illumina emplea terminadores *reversibles* marcados con cuatro fluoróforos diferentes, lo que permite continuar la síntesis de las moléculas tras determinar qué nucleótido ha sido incorporado en último lugar. Un secuenciador HiSeq 2000 de Illumina es capaz de producir, en una sola carrera, unas 200 Gpb, con una fidelidad del 99,5% (Zhang *et al*., 2011). La tecnología de Illumina ha sido utilizada en muchos proyectos de secuenciación por la alta calidad de los datos que produce. Los errores cometidos por este método suelen ser sustituciones nucleotídicas, que ocurren como consecuencia de los errores introducidos en la etapa previa de amplificación.

En un secuenciador Roche, la paralelización se logra emulsionando las reacciones de amplificación. En lugar de terminadores, el método de Roche utiliza desoxinucleótidos trifosfato convencionales. Los nucleótidos se añaden cíclicamente a la reacción, uno distinto en cada ciclo, y su incorporación a las moléculas sintetizadas se detecta gracias a la liberación de pirofosfato. Por ello, este método se denomina pirosecuenciación (*pyrosequencing*). Esta tecnología es propensa a introducir pequeñas inserciones y deleciones, como consecuencia de errores en la detección de la cantidad de pirofosfato liberado durante la síntesis de homopolímeros (Tucker *et al.*, 2009; Zhang *et al.*, 2011).

La plataforma *ABI SOLiD* utiliza un método denominado secuenciación por ligación (*sequencing-by-ligation*). Tras una amplificación inicial, en la que la paralelización se consigue emulsionando las reacciones, la determinación de la secuencia se realiza mediante un procedimiento que requiere la ligación secuencial de oligonucleótidos marcados con fluorocromos (Tucker *et al.*, 2009; Zhang *et al.*, 2011; Tu *et al.*, 2012).

2.1.2. Tecnologías de secuenciación de tercera generación

Una de estas tecnologías que actualmente está en auge es la que se conoce como SMRT (*Single-Molecule Real-Time sequencing technology*), desarrollada por Pacific Biosciences (Roberts *et al.*, 2013). Se trata de un método basado en la síntesis de

moléculas individuales de ADN, que es detectada en tiempo real gracias a un chip de secuenciación que contiene miles de guías de onda de modo cero (*zero-mode wave-guides*; ZMWs). Cada una de estas guías aprovecha las propiedades de la luz al atravesar una pequeña apertura para confinar los fotones emitidos durante la secuenciación y dirigirlos al detector. Los fotones se canalizan hacia estas guías gracias a que la síntesis de cada fragmento es llevada a cabo por una única molécula de ADN polimerasa que se dispone unida a la guía (Pareek *et al.*, 2011). Dado que esta técnica no requiere una etapa previa de amplificación, se minimizan algunos problemas inherentes a la misma, como la ocurrencia de sustituciones nucleotídicas y los sesgos en la población de moléculas secuenciadas. Además, se espera que este método permita salvar una de las mayores dificultades afrontadas por las tecnologías de secuenciación de primera y segunda generación: la secuenciación de genomas con alto contenido en G+C (Shin *et al.*, 2013).

2.2. Resecuenciación de genomas

Una de las aplicaciones de las tecnologías de secuenciación descritas en las secciones anteriores es la resecuenciación de genomas previamente caracterizados. A diferencia de los ensamblajes *de novo* (que se realizan cuando un genoma se secuencia por primera vez), la resecuenciación de un genoma consiste en el alineamiento rápido de los fragmentos secuenciados (denominados lecturas; del inglés *reads*) a una secuencia disponible de antemano, que sirve de referencia. Dicho alineamiento facilita la identificación de mutaciones y/o polimorfismos mediante la comparación de la secuencia del genoma de referencia con la deducida a partir de las lecturas alineadas al mismo (Candela *et al.*, 2015).

Las aplicaciones bioinformáticas que llevan a cabo el alineamiento deben afrontar dos desafíos importantes. En primer lugar, deben implementar algoritmos rápidos que permitan alinear en un tiempo breve el elevado número de lecturas producidas por las tecnologías de secuenciación masivamente paralela. En segundo lugar, deben ser capaces de alinear cada secuencia a una posición única en el genoma de referencia, una operación que se ve dificultada por el pequeño tamaño de las lecturas así como por la abundancia de secuencias repetitivas en el genoma de referencia. Por ello, no resulta extraño que un número significativo de lecturas pueda alinearse igualmente bien a numerosas localizaciones cromosómicas (Zhang *et al*., 2011).

Entre los algoritmos de alineamiento más rápidos disponibles en la actualidad, destacan los basados en la transformación de Burrows-Wheeler, que se utilizan en los programas Bowtie y BWA, entre otros (Langmead *et al*., 2009; Li y Durbin, 2009). Esta transformación es una permutación del orden de los caracteres de la secuencia de referencia que permite que los alineamientos se lleven a cabo en tiempo proporcional a la longitud de las lecturas, independientemente del tamaño del genoma al que deban

alinearse. En este trabajo, hemos utilizado el programa Bowtie2 para llevar a cabo los alineamientos (Langmead y Salzberg, 2012). La asignación de cada lectura a una sola posición en el genoma resulta más fácil cuando se utilizan lecturas emparejadas (*paired-end reads*), resultado de la secuenciación de fragmentos del genoma por ambos extremos. Este tipo de lecturas es el empleado en este trabajo.

2.3. Cartografía mediante secuenciación

La cartografía mediante secuenciación (*mapping-by-sequencing*) es una estrategia que permite identificar mutaciones puntuales que causan un fenotipo de interés mediante la secuenciación del ADN genómico (ADNg) proveniente de un grupo de individuos seleccionados por su fenotipo de entre los de una población cartográfica (*mapping population*) apropiada (Ossowski *et al.,* 2008; Schneeberger y Weigel, 2011). Inicialmente, este método se propuso para cartografiar e identificar las mutaciones causantes de un fenotipo de interés utilizando poblaciones cartográficas derivadas de cruzamientos entre variedades diferentes, como las estirpes Columbia-0 (Col-0) y Landsberg *erecta* (L*er*) de *Arabidopsis thaliana* (Schneeberger *et al*., 2009). Más recientemente, se ha descrito el uso de esta técnica para identificar mutaciones empleando cruzamientos entre un mutante de interés y la estirpe silvestre a partir de la cual se indujo (Zuryn *et al*., 2010). Tales cruzamientos entre líneas isogénicas utilizan las mutaciones inducidas en un experimento de mutagénesis como marcadores moleculares.

Los métodos de cartografía mediante secuenciación aplican una estrategia denominada análisis de segregantes agrupados (*bulked segregant analysis*; Michelmore *et al*., 1991) para identificar regiones del genoma ligadas a la mutación a estudio. En esta estrategia se combina en una misma muestra el ADN de todos los individuos que manifiestan el fenotipo correspondiente al homocigoto recesivo. En dicha muestra, se espera que todas las regiones del genoma no ligadas a la mutación a estudio estén representadas por lecturas provenientes en un 50% de cada uno de los genomas parentales. Sin embargo, en las regiones íntimamente ligadas a la mutación a estudio se espera que las lecturas provengan hasta en un 100% del genoma del mutante, por efecto de la selección ejercida al establecer el grupo. La detección de este sesgo en las frecuencias alélicas para cada posición del genoma permite determinar la posición aproximada de la mutación. El examen detallado del alineamiento de las lecturas al genoma de referencia debe, en última instancia, permitir la identificación de la lesión responsable del fenotipo observado (Schneeberger *et al*., 2009).

Existen gran variedad de programas para analizar los resultados de experimentos de cartografía mediante secuenciación. Entre todos ellos, por razones históricas, destaca el programa SHOREmap, que utiliza la información derivada del alineamiento de las lecturas

para identificar la posición de la mutación en el genoma y proporcionar una lista de mutaciones candidatas clasificadas por sus efectos predichos sobre la función de los productos génicos (Schneeberger *et al.*, 2009; Sun y Schneeberger, 2015).

2.4. La función del cloroplasto y su relación con el albinismo

Los cloroplastos son orgánulos presentes en las células vegetales que desempeñan funciones esenciales para la biosíntesis y almacenamiento de numerosas moléculas orgánicas. Además de la fotosíntesis, en los cloroplastos tiene lugar la biosíntesis de carotenoides, aminoácidos aromáticos y ácidos grasos, entre otras moléculas. El número de cloroplastos por célula depende del tipo celular y órgano de la planta. Por ejemplo, las células del parénquima clorofílico de las hojas llegan a contener entre 30 y 40 cloroplastos, que se disponen alrededor de la gran vacuola central. Los cloroplastos son orgánulos separados del citoplasma por una doble membrana. El interior del cloroplasto posee, a su vez, un sistema elaborado de membranas tilacoidales, que se apilan para formar los denominados grana (Tomizioli *et al.*, 2014). Todas estas membranas definen distintos compartimentos especializados en distintas funciones: el espacio periplásmico (comprendido entre las dos membranas externas), el estroma (definido por la membrana más interna de las dos) y el lumen o espacio intratilacoidal (definido por las membranas tilacoidales).

Los cloroplastos se originaron cuando una cianobacteria fotosintética estableció una endosimbiosis con un eucariota ancestral. Como resultado de este origen, los cloroplastos conservan un genoma circular, ribosomas semejantes a los de las bacterias y se encuentran delimitados por una doble membrana (Kobayashi *et al.,* 2016; Powikrowska *et al.,* 2014). El genoma de los cloroplastos reside en el estroma y contiene un número limitado de genes, ya que la mayor parte de sus genes ha sido transferida al genoma del núcleo. Entre los genes que permanecen en el genoma del cloroplasto destacan algunos cuya transcripción produce moléculas de ARN de transferencia (ARNt) y de ARN ribosómico (ARNr), que son necesarias para la expresión de los restantes genes del genoma del cloroplasto, que codifican unas 120 proteínas. La función de estas proteínas está relacionada principalmente con la expresión de los genes del genoma del cloroplasto (por ejemplo, subunidades de una de las dos ARN polimerasas que funcionan en este orgánulo o numerosas subunidades del ribosoma) y la fotosíntesis (por ejemplo, subunidades de los fotosistemas I y II o subunidades de la ribulosa 1,5-bisfosfato carboxilasa-oxigenasa, RuBisCO; Andersson y Backlund, 2008).

En consecuencia, la biogénesis y función de los cloroplastos requiere la expresión coordinada de genes de los genomas nuclear y del cloroplasto. Numerosas proteínas nucleares se sintetizan como preproteínas que contienen una secuencia denominada péptido de tránsito en su extremo N-terminal. Este péptido les asegura su entrada al

cloroplasto por medio de la actividad de complejos multiproteicos (denominados translocones), localizados tanto en la membrana externa del cloroplasto como en la membrana interna (Inoue *et al.,* 2013). El péptido de tránsito se elimina tras la entrada de dicha proteína al cloroplasto (Fellerer *et al.,* 2011).

Muchas funciones del cloroplasto han sido elucidadas gracias a la caracterización genética y molecular de mutantes albinos, que carecen de la pigmentación verde característica de la estirpe silvestre, aislados en diversas especies (Myouga *et al*., 2010; Bryant *et al*., 2011). Esta falta de pigmentación puede deberse a distintos motivos, como defectos en la biogénesis de los cloroplastos o a alteraciones en la biosíntesis de carotenoides. Por la facilidad en su identificación, los mutantes albinos han sido utilizados a veces para medir la eficacia de los experimentos de mutagénesis. En este trabajo, hemos seleccionado un mutante albino para la puesta a punto de un método de cartografía mediante secuenciación.

3. Antecedentes y Objetivos

Las tecnologías de secuenciación masivamente paralela (*massively parallel sequencing technologies*) permiten la secuenciación rápida y asequible de genomas completos. Estas tecnologías han permitido el desarrollo de una técnica denominada cartografía mediante secuenciación (*mapping-by-sequencing*), que permite la identificación de la mutación responsable de un fenotipo de interés mediante la resecuenciación del genoma de un conjunto de individuos mutantes previamente seleccionados a partir de una población segregante (por ejemplo, una población cartográfica F_2). En el laboratorio se dispone de un mutante albino de la planta modelo *Arabidopsis thaliana*, que fue inducido mediante un tratamiento con metanosulfonato de etilo (EMS) durante una estancia postdoctoral de Héctor Candela en el Plant Gene Expression Center (Albany, California, EE.UU.). El mutante seleccionado se caracteriza por su pigmentación albina y por detener el desarrollo inmediatamente tras la germinación. Los mutantes albinos se identifican con gran facilidad debido a las diferencias en la pigmentación con respecto a la estirpe silvestre, llegando a veces a presentar un aspecto transparente.

El objetivo principal de este Trabajo de Fin de Grado ha sido poner a punto un método de cartografía mediante secuenciación, intentando al mismo tiempo a contribuir a la comprensión de la base genética y molecular del fenotipo albino. Para identificar la mutación causante de este fenotipo mutante, hemos seguido una estrategia de cartografía mediante secuenciación, que nos ha permitido determinar rápidamente la posición en el genoma y la naturaleza molecular de dicha mutación. Para ello, hemos llevado a cabo la secuenciación y análisis bioinformático del genoma completo de un grupo de individuos F_2 derivados de un cruzamiento entre el mutante a estudio y la estirpe silvestre Columbia-0 (Col-0). Este objetivo principal se ha concretado en los siguientes objetivos particulares, que coinciden con las etapas seguidas durante la realización del trabajo:

1) Obtener una población cartográfica F_2 adecuada para la realización de un experimento de cartografía mediante secuenciación. Dado que la estirpe mutagenizada posee un fondo genético mixto, que proviene parcialmente de la estirpe silvestre Landsberg *erecta* (L*er*), hemos obtenido dicha población mediante el cruzamiento entre el mutante a estudio y la estirpe Col-0.

2) Seleccionar dos conjuntos de plantas: uno compuesto por todos los individuos fenotípicamente mutantes identificados en la población cartográfica, y otro compuesto por un número semejante de individuos fenotípicamente silvestres de la misma población. La estrategia de clonación elegida requiere la secuenciación del genoma utilizando el ADN genómico de individuos homocigóticos para la mutación a estudio (como los del primer conjunto). Dicha estrategia se conoce como análisis de segregantes agrupados (*bulked*

segregant analysis), y permite la identificación de cambios en la secuencia del ADN ligados al fenotipo mutante.

3) Analizar la secuencia obtenida mediante el uso de herramientas bioinformáticas, para identificar la mutación responsable del fenotipo mutante. Las mutaciones identificadas deben reunir los siguientes requisitos: a) deben encontrarse en homocigosis en la secuencia obtenida (ya que el material genético se ha obtenido a partir de individuos que manifiestan el fenotipo recesivo); b) la mutación debe ser, con gran probabilidad, una transición G/C→A/T, ya que fue inducida mediante un tratamiento con metanosulfonato de etilo; y c) la mutación debe causar una alteración en la secuencia o la expresión de algún gen, como la sustitución de un aminoácido por otro, o un procesamiento incorrecto de los intrones.

4) Verificar experimentalmente el ligamiento entre el fenotipo albino y la mutación identificada, e iniciar los experimentos necesarios para confirmar que hemos identificado correctamente la mutación causante del fenotipo albino.

4. Materiales y Métodos

4.1. Cultivo de *Arabidopsis thaliana*

4.1.1. Origen del mutante albino

El mutante albino empleado en este trabajo se aisló en un experimento de mutagénesis con metanosulfonato de etilo (EMS), que fue realizado en el año 2008 por Héctor Candela, durante su estancia postdoctoral en el Plant Gene Expression Center. El EMS es un agente alquilante responsable de mutaciones puntuales, en concreto transiciones del tipo G/C→A/T (Greene *et al*., 2003). La estirpe mutagenizada fue el mutante recesivo *bp-9* (*brevipedicellus-9*; Venglat *et al*., 2002), cuyo fondo genético es parcialmente el de la estirpe Landsberg *erecta* (L*er*). Por su letalidad, el mutante albino caracterizado en este trabajo se identificó en una familia M_3, obtenida a partir de la autofecundación de plantas M_2 que no manifestaban el fenotipo.

4.1.2. Selección de mutantes en las familias F_2

La población cartográfica se obtuvo cruzando plantas de fenotipo silvestre (heterocigóticas para la mutación) con la estirpe silvestre Columbia-0 (Col-0). La progenie F_1 de estos cruzamientos, compuesta en un 50% por plantas heterocigóticas para la mutación, se autofecundó para identificar familias F_2 que segregasen plántulas albinas. Se seleccionaron 87 plantas albinas y 170 plantas de fenotipo silvestre, con las que se establecieron sendos grupos de plantas (*pools*) de los que se aisló colectivamente el ADN genómico (ADNg) para su secuenciación posterior. Las plantas se congelaron a -80°C hasta el momento de purificar su ADN genómico.

4.1.3. Medio de cultivo sólido para *Arabidopsis thaliana*

Las plantas de *Arabidopsis thaliana* se cultivaron en medio de Murashige y Skoog (MS, Murashige y Skoog, 1962), cuya composición se describe a continuación: NH_4NO_3 20,6 mM; H_3BO_3 0,1 mM; $CaCl_2$ 3 mM; $CoCl_2 \cdot 6H_2O$ 0,1 µM; $CuSO_4 \cdot 5H_2O$ 0,1 µM; Na_2EDTA 0,11 mM; $MgSO_4$ 1,5 mM; $MnSO_4 \cdot H_2O$ 0,1 mM; $NaMoO_4 \cdot 2H_2O$ 1 µM; Kl 5 µM; KNO_3 18,8 mM; KH_2PO_4 1,24 mM; y $ZnSO_4 \cdot 7H_2O$ 30 µM.

Para preparar 1 L de medio MS, se añaden 900 mL de agua destilada a un matraz de 2 L. Posteriormente, se disuelven 4,3 g/L de sales de Murashige y Skoog (Duchefa), 10 g de sacarosa (concentración final de 1% m/v) y 0,5 g/L de MES (ácido 2-[N-morfolino]-etano sulfónico) con ayuda de un agitador magnético. El pH de la disolución se ajusta a 5,7 con KOH 5 M. Finalmente, tras enrasar el volumen a 1 L, se añaden 6,5 g/L de Plant Agar (Duchefa) y se esteriliza en autoclave a 121°C durante 20 min. Tras la esterilización del medio de cultivo, alicuotamos el medio en cajas de Petri de 15 cm de diámetro, una vez

atemperado a 50-60°C, en una cabina de flujo laminar horizontal. Las cajas de Petri con medio MS se almacenaron a 4°C en posición invertida.

Hemos utilizado también medio MS sin sacarosa para determinar el efecto de la misma sobre el crecimiento y desarrollo de los mutantes albinos. Dicho medio se preparó siguiendo el procedimiento anterior, excepto en lo referido a la adición de sacarosa.

4.1.4. Esterilización de semillas

Para impedir la contaminación de los cultivos con hongos, las semillas se incubaron con 1 mL de disolución de esterilización en tubos *eppendorf*. Para preparar 250 mL de disolución de esterilización, se mezclan 100 mL de lejía comercial, 150 mL de agua desionizada y 500 µL de Triton X-100 al 1% (v/v). Durante la incubación, que duró 8 min, los tubos se agitaron por inversión para deshacer los agregados de semillas que pudieran formarse. Tras eliminar la disolución de esterilización de los tubos, realizamos consecutivamente tres lavados con 1 mL de agua estéril, agitando los tubos por inversión. Las semillas se mantuvieron en agua estéril hasta el momento de su siembra.

4.1.5. Condiciones para el crecimiento de *Arabidopsis thaliana*

Los cultivos de *Arabidopsis thaliana* se realizaron a partir de semillas almacenadas a 4°C. Todos los cultivos se iniciaron en cajas de Petri con medio sólido MS, donde permanecieron unas 3-4 semanas.

4.1.5.1. Cultivo en cajas de Petri

Las cajas de Petri de 15 cm de diámetro contenían aproximadamente 100 mL de medio sólido MS. Las semillas se esterilizaron según el procedimiento descrito en el apartado 4.1.4 de la página 15, y se sembraron con la ayuda de pipetas Pasteur estériles en una cabina de flujo laminar. En cada caja de Petri, las semillas se sembraron regularmente espaciadas, a razón de 104 semillas por caja de Petri, utilizando una plantilla cuadriculada. Las cajas de Petri se precintaron con esparadrapo quirúrgico (3M) para impedir la contaminación del medio de cultivo, y se mantuvieron durante 24 h a 4°C en posición invertida para sincronizar la germinación de las semillas. Posteriormente, se incubaron a 20±0,1°C en un incubador HLR-352-PE (Panasonic) con iluminación continua suministrada por tubos fluorescentes Panasonic 40 FL40SS-ENW-37. Transcurridos 26 días, coincidiendo con el inicio de la elongación del tallo, se seleccionaron 15 plantas no albinas y se trasplantaron a maceta para que completaran su ciclo de vida.

4.1.5.2. Cultivo en maceta

Los cultivos en maceta se realizaron en bandejas de plástico de 28×50 cm con 59 alveolos y dimensiones individuales de 5 cm de diámetro y 5 cm de altura. Cada alveolo contenía una maceta de rejilla. Las macetas se rellenaron con una mezcla de perlita,

vermiculita y turba en proporción 2:2:1 (v/v/v) que había sido esterilizada previamente en autoclave. Las macetas se subirrigaron con una disolución nutritiva comercial (abono Universal líquido Carrefour), manteniendo un nivel de 2 ó 3 cm de la misma en las bandejas. La disolución contiene: nitrógeno (N) total: 6% (incluye N amoniacal: 2,8%; N ureico: 2,2%; y N nítrico: 1,0%); pentóxido de fósforo (P_2O_5): 6%; óxido de potasio (K_2O): 6%; micronutrientes: cobre (Cu): 0,002%; hierro (Fe): 0,025%; manganeso (Mn): 0,014%; molibdeno (Mo): 0,001%; y zinc (Zn) quelado con EDTA: 0,004%. Las plantas se irrigaron semanalmente con agua y con una dilución al 0,5% (v/v) de la disolución nutritiva descrita cada dos semanas.

Las plantas cultivadas en cajas de Petri se transfirieron a macetas, con ayuda de pinzas metálicas esterilizadas con alcohol, a los 26 días tras la siembra. Los cultivos en maceta se mantuvieron a 23±1°C en una cámara climática bajo iluminación continua de 5.000 lx suministrada por tubos fluorescentes F28 T5/D y Leuci T5-28W. El riego se interrumpió cuando cesó la producción de flores, con el fin de que se secaran los frutos y recolectar las semillas.

4.1.6. Recolección de semillas

El secado de las plantas tuvo lugar en la cámara climática. Una vez que las plantas completaron su ciclo de vida y se secaron, se recolectaron las semillas. Los restos vegetales se eliminaron con ayuda de un colador y las semillas se almacenaron en tubos *eppendorf* a 4°C.

4.2. Manipulación de ácidos nucleicos

4.2.1. Purificación de ADN genómico para ultrasecuenciación

Realizamos dos purificaciones de ADNg utilizando un kit para plantas (*GeneJET Plant Genomic DNA Purification Mini Kit*) siguiendo el protocolo del fabricante (Thermo Scientific), que se resume a continuación. El tejido (120 mg del mutante albino y 220 mg de las plantas de fenotipo silvestre) se trituró en morteros de porcelana estériles distintos y en presencia de nitrógeno líquido. El material vegetal se transfirió a tubos *eppendorf*, a los que añadimos 350 µL de tampón de lisis A. Tras agitar durante 10-15 s con un vórtex, añadimos 50 µL de tampón de lisis B y 20 µL de RNasa A. Los tubos se incubaron, con agitación periódica, en un baño a 65°C durante 10 min. A continuación, añadimos 130 µL de disolución de precipitación a cada tubo, que agitamos por inversión e incubamos durante 5 min en hielo. A continuación, centrifugamos ambos tubos durante 5 min a 13.300 rpm (16.300g). Transferimos los sobrenadantes a tubos limpios, a los que añadimos 400 µL de disolución de unión (*binding solution*) y 400 µL de etanol al 96%. Transferimos parte del volumen (600-700 µL) a las columnas suministradas con el kit y centrifugamos los tubos a 8000 rpm (5.900g) durante 1 min. La operación se repitió con el volumen restante,

descartando cada vez el sobrenadante tras su paso por la columna. Las columnas se lavaron dos veces con 500 µL de tampón de lavado 1. El tampón de lavado se eliminó centrifugando durante 1 min a 10.000 rpm (9.200g). Por último, eluimos el ADNg retenido en la columna a tubos *eppendorf* estériles añadiendo 100 µL de tampón de elución al centro de la columna, incubando a temperatura ambiente durante 5 min y centrifugando a 10.000 rpm durante 5 min.

4.2.2. Purificación de ADNg para genotipado de plantas individuales

Los mutantes albinos se recolectaron de las cajas de Petri a los 26 días tras la siembra y se congelaron a -80°C en tubos *eppendorf*. También se congelaron plantas control de fenotipo silvestre. Se purificó el ADNg de 35 plantas albinas y 3 plantas silvestres de la F_2 con el propósito de estudiar la cosegregación del albinismo con la mutación candidata. Además, purificamos ADNg a partir de 20 hojas de otras tantas plantas F_2 fenotípicamente silvestres, con el propósito de identificar plantas portadoras de la mutación candidata con las que purificar ARN total. El protocolo seguido para esta purificación de ADNg se resume a continuación. Añadimos 250 µL de tampón de extracción (Tris 100 mM pH 8; EDTA 50 mM a pH 8; NaCl 500 mM) a cada tubo *eppendorf* y trituramos el tejido con palillos de plástico hasta conseguir una disolución homogénea. Añadimos otros 250 µL de tampón de extracción, hasta alcanzar un volumen final de 500 µL, y agitamos los tubos por inversión. Añadimos 35 µL de SDS 20% (m/v) a cada tubo e incubamos durante 5 min a 65°C en un baño, agitando periódicamente los tubos por inversión. Transcurrido este tiempo, añadimos 130 µL de acetato de potasio 5 M a cada tubo, e incubamos los tubos en hielo durante 8 min. Posteriormente, centrifugamos los tubos a 13.000 rpm (15.600g) durante 10 min. Transferimos el sobrenadante a tubos nuevos, a los que añadimos 640 µL de isopropanol y 60 µL de acetato de sodio 3 M. Tras agitar los tubos, se incubaron en hielo durante 20 min y se centrifugaron a 13000 rpm durante 10 min. El precipitado resultante se lavó con 300 µL de etanol 70%. Este etanol se eliminó tras centrifugar una última vez a 13.000 rpm durante 5 min. Una vez seco, el precipitado se resuspendió en un volumen de 25 a 50 µL de agua destilada estéril.

4.2.3. **Amplificaciones mediante PCR**

En este trabajo hemos utilizado los cebadores de la Tabla 1, que fueron sintetizados por un proveedor comercial (Sigma-Aldrich). Las reacciones de PCR se prepararon en tubos *eppendorf* de 0,2 mL de pared fina, en un volumen final de 20 µL. La composición de las reacciones fue la siguiente: 6,3 µL de agua desionizada, 2 µL de *DreamTaq Buffer* 10x, 1,6 µL de una mezcla de los 4 dNTP a una concentración de 10 mM (con cada uno de ellos a 2,5 mM), 4 µL de los cebadores *forward* y *reverse* correspondientes a 2,5 µM cada uno, 0,1

µL de la enzima *DreamTaq* a 5 U/µL y 2 µL de molde apropiado. Durante la preparación de cada reacción, todos los reactivos se mantuvieron en hielo hasta el momento de su uso.

Tabla 1.- Nombre y secuencia de los cebadores empleados

Cebador	Secuencia nucleotídica (5'→3')
At2g04030-F	CGTCGTGGAACACAAATCAC
At2g04030-R	AGGCTTACCCATAGCGGTTT
At2g04030-cds-F	ggggacaagtttgtacaaaaaagcaggctTAATGGCTCCTGCTTTGAGTA
At2g04030-cds-R1	ggggaccactttgtacaagaaagctgggtCAATCTTGCCAAGGATCACTC

Se indica en letras minúsculas la secuencia de los sitios *att*B1 y *att*B2, que permitirán la clonación del producto de PCR mediante la tecnología Gateway.

Las reacciones de PCR se incubaron en un termociclador *T100 Thermal Cycler* (Bio-Rad). La temperatura y la duración de cada una de las etapas del programa de amplificación fueron las que se muestran a continuación. La etapa inicial se programó a 95°C durante 2 min. Tras esta, se realizaron 35 ciclos en las siguientes condiciones: (1) una etapa de desnaturalización a 95°C 30 s; (2) una etapa de hibridación a una temperatura comprendida entre 55°C y 64°C, dependiendo de la temperatura de fusión (T_m) de los cebadores presentes en la mezcla de reacción, durante 30 s; y (3) una etapa de síntesis a 72°C durante 45 s, tiempo que se determinó teniendo en cuenta que la enzima *DreamTaq* sintetiza ADN a razón de 1 kb cada 30 s. Finalmente, se añadió una etapa final de extensión a 72°C durante 5 min.

4.2.4. Electroforesis en gel de agarosa

Las muestras de ADN se visualizaron en geles de agarosa al 1% (m/v) en tampón de electroforesis TAE 1x (0,5 g de agarosa en 50 mL de TAE 1x), que se preparó a partir de una disolución madre a concentración 50x (Tris-HCl 2 M; ácido acético 5,71%; EDTA 50 mM; pH 8,0). La tinción de los geles se realizó con 1,25 µL de SafeView, que se añadieron a la disolución de agarosa antes de que se solidificase. Además, en la primera y última calle de cada gel se cargaron 4 µL de marcador de peso molecular (1 kb DNA Ladder de GeneCraft a una concentración de 0,1 µg/µL). Las muestras de ADN purificado se prepararon con 1 µL de ADN, 8 µL de agua desionizada y 1 µL de tampón de carga 10x (DNA Loading Buffer 10x de 5prime). Por último, la electroforesis se realizó con una fuente de alimentación a 90 V (LabNet) durante 30 min.

4.2.5. Purificación de productos de PCR

Para llevar a cabo este paso se utilizó el *GeneJET PCR Purification Kit*, siguiendo las instrucciones del fabricante (Thermo Scientific). Los productos purificados se visualizaron en geles de agarosa al 1% en tampón de electroforesis TAE 1x.

4.3. Análisis bioinformático

Para la secuenciación del genoma, recurrimos a los servicios de la empresa STAB VIDA (Caparica, Portugal), que dispone de un secuenciador masivamente paralelo Illumina HiSeq 2500. Los fragmentos se secuenciaron por ambos extremos con lecturas de 101 nucleótidos (*paired-end reads*). Para evaluar la calidad de las lecturas obtenidas y determinar la presencia en las mismas de secuencias derivadas de los adaptadores añadidos durante la preparación de las muestras, utilizamos el programa FastQC (http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/). Hemos utilizado en programa PEAR (Paired-End reAd mergeR; http://sco.h-its.org/exelixis/web/software/pear/) para detectar solapamientos entre las dos lecturas derivadas de cada fragmento secuenciado. En caso de solapamiento, PEAR ensambla ambas lecturas, produciendo secuencias de mayor longitud que las lecturas originales (Zhang *et al*., 2014). Para eliminar los adaptadores detectados, utilizamos el programa Trimmomatic (versión 0.32; Bolger *et al*., 2014), al que se le proporcionó un archivo con la secuencia de los adaptadores mediante la opción `ILLUMINACLIP`. También se utilizaron las opciones `LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15` y `MINLEN:36`.

El alineamiento de las lecturas obtenidas al genoma de referencia se llevó a cabo con el programa Bowtie2 (versión 2.1.0), que emplea una algoritmo de alineamiento muy eficiente basado en la transformación de Burrows-Wheeler (Langmead *et al*., 2009; Langmead y Salzberg, 2012). Como genoma de referencia, a la que alineamos las lecturas, utilizamos la versión más reciente del genoma de *Arabidopsis thaliana* (versión TAIR10), cuya secuencia descargamos en formato FastA (Lamesch *et al*., 2012). El índice del genoma de referencia se preparó con el programa `bowtie2-build`, especificando su secuencia con la opción –f. El alineamiento de las lecturas emparejadas se llevó a cabo mediante el programa `bowtie2`, con las siguientes opciones: `-x` (especifica el índice del genoma al que deben alinearse las lecturas, preparado anteriormente con `bowtie2-build`), `--qc-filter` (descarta las lecturas de baja calidad), `-p` (especifica el número de procesadores disponibles; en nuestro caso es igual a 8), `-X` (especifica el tamaño de máximo de los fragmentos que, en nuestro caso, fue fijado arbitrariamente en 650 nucleotidos), `--no-discordant` (descarta las parejas de lecturas cuyos alineamientos no son concordantes), `--no-mixed` (descarta las lecturas cuya pareja no ha sido eliminada) y `--no-unal` (descarta aquellas lecturas para las que no se detectan alineamientos).

Los archivos SAM resultantes, que contienen la descripción de los alineamientos de las lecturas, fueron convertidos a formato BAM mediante el comando `view` de samtools (Li *et al*., 2009). Los archivos BAM correspondientes se ordenaron según la posición de las lecturas alineadas en el genoma de referencia mediante el comando `sort` de samtools. Los archivos así ordenados y correspondientes a la misma muestra, se combinaron en un único archivo mediante el comando `merge` de samtools. El archivo resultante se indizó con el comando `index` de samtools. Utilizamos el comando `AddOrReplaceReadGroups` del programa piccard-tools (versión 2.5.0) para asignar las lecturas a grupos. El archivo BAM resultante fue procesado con el programa GATK (Genome Analysis Toolkit; versión 3.6; McKenna *et al*., 2010), utilizando las opciones `-T RealignerTargetCreator` y `-T IndelRealigner`. Estas dos opciones permiten seleccionar una lista de sitios del genoma adyacentes a inserciones y deleciones que posiblemente requieren que los alineamientos de las lecturas sean corregidos, y llevar a cabo las correcciones necesarias. Utilizamos, por último, el programa GATK con la opción `-T UnifiedGenotyper`, para identificar las bases que ocupan cada una de las posiciones del genoma (operación denominada, en inglés, *base calling*). El resultado se guardó en un archivo en formato VCF (Danecek *et al*., 2011), que fue procesado manualmente mediante pequeñas rutinas escritas en lenguajes Perl y AWK, ejecutadas en la línea de comandos de Linux, y Microsoft Excel.

Se representaron para cada cromosoma todas las posiciones y frecuencias relativas al mutante, la media móvil de los datos anteriores, la media móvil de las posiciones y frecuencias del silvestre y el valor conocido como *boost*, que permite definir una región en la que se encuentre con una gran probabilidad la mutación a estudio. La fórmula para calcular el valor de *boost* es: $B_v = 1/|1-\max(\theta_{tar},1-\theta_{tar})/\max(\theta_{obs},1-\theta_{obs})|$ (Sun y Schneeberger, 2015).

Algunos apartados de la sección de Materiales y Métodos se han modificado a partir de trabajos anteriores realizados en el laboratorio de Héctor Candela (como el de E.M. Rodríguez Alcocer, 2014, citado en la Bibliografía).

5. Resultados

5.1. Fenotipo y modo de herencia de un mutante albino

Para determinar el modo de herencia del mutante a estudio, cruzamos plantas fenotípicamente silvestres portadoras de la mutación albina en heterocigosis con plantas de la estirpe silvestre Columbia-0 (Col-0). No encontramos plantas albinas en la progenie F_1 de estos cruzamientos, como cabe esperar si la mutación se hereda como un carácter recesivo. Con el fin de obtener la población F_2 necesaria para la realización de un experimento de cartografía mediante secuenciación, permitimos que las plantas de la progenie F_1 se autofecundaran. Las poblaciones F_2 resultantes fueron sembradas en cajas de Petri que contenían medio MS suplementado con sacarosa (véase el apartado 4.1.3 de la página 14). El fenotipo albino reapareció en 15 de los 85 individuos de la Familia 2, un resultado que es compatible con la razón fenotípica 3:1 (silvestre:mutante), característica del modo de herencia monogénico recesivo (Tabla 2). Sin embargo, el fenotipo mutante apareció en un número de individuos significativamente mayor que el esperado en la Familia 1, quizá por la presencia de mutaciones adicionales que distorsionan la segregación fenotípica en dicha familia. En conjunto, el fenotipo albino reapareció en 87 de los 257 individuos estudiados, aproximadamente un tercio de las plantas de la generación F_2.

Tabla 2.- Segregación del fenotipo mutante albino en la progenie F_2 de cruzamientos entre plantas portadoras de la mutación a estudio y la estirpe silvestre Col-0

	Clases fenotípicas[a]		Segregación		
Familia F_2	Silvestre	Albino	Hipótesis	X^2	P
1	100	72	3:1	25,19[b]	$5{,}2 \times 10^{-7}$
2	70	15	3:1	2,07[b]	0,15

[a]Número de individuos asignados a cada clase fenotípica. [b]Los valores de X^2 han sido calculados aplicando la corrección de Yates por haber sólo 2 clases fenotípicas. Al nivel de significación α=0,05, el valor crítico del estadístico de la prueba X^2 es 3,84 (con 1 grado de libertad).

Por otro lado, sembramos la población F_2 en medio MS carente de sacarosa con el fin de estudiar posibles diferencias en el desarrollo de las plantas mutantes. En estas condiciones, las plántulas presentaron un menor desarrollo, particularmente por el reducido tamaño de sus cotiledones, lo que dificultó aún más la tarea de localizarlas en las cajas de Petri. En la Figura 1 se muestran los fenotipos de la estirpe silvestre y del mutante albino en presencia de sacarosa.

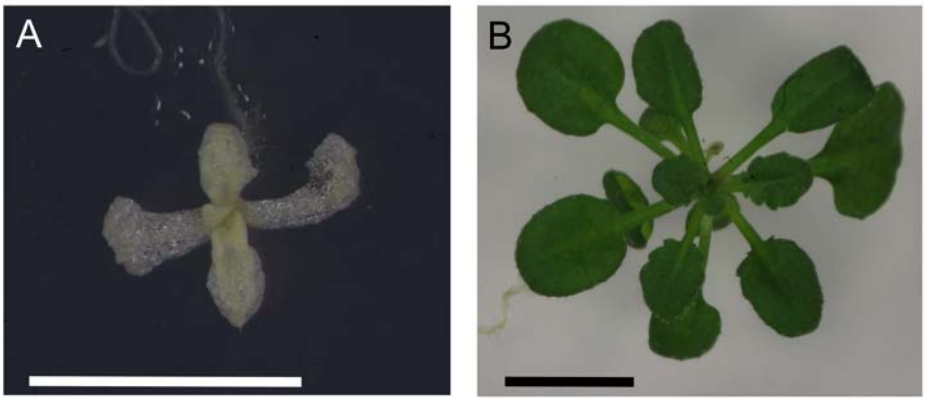


Figura 1.- Fenotipo de un mutante albino de *Arabidopsis thaliana* cultivado en medio suplementado con sacarosa. (A) Mutante albino. (B) Estirpe silvestre Landsberg *erecta*. Las barras de escala representan 5 mm. En ambos casos, las fotografías se tomaron a los 26 días tras la siembra.

5.2. Obtención de ADN para su secuenciación

Las 87 plantas fenotípicamente mutantes identificadas en la población cartográfica F_2 (72 de la Familia 1 y 15 de la Familia 2) se recolectaron en tubos *eppendorf*, que se mantuvieron a -80°C hasta la purificación del ADN genómico (ADNg). El ADNg se purificó tras combinar los tejidos de las plantas de ambas familias, siguiendo el protocolo descrito en el apartado 4.2.1 de la página 16. Procedimos de la misma manera con las plantas fenotípicamente silvestres. Las dos muestras (de plantas mutantes y silvestres, respectivamente) se enviaron congeladas en hielo seco a la empresa STAB VIDA (Caparica, Portugal) para su secuenciación con un secuenciador Illumina HiSeq 2500.

La calidad y cantidad del ADN de las muestras fue evaluada por medio de una electroforesis en gel de agarosa (Figura 2) y mediante un fluorímetro Qubit. La concentración de ADN fue de 15,4 ng/μL en la muestra obtenida a partir de las plántulas albinas y 13,2 ng/μL en la muestra obtenida a partir de plántulas de fenotipo silvestre. La Figura 2 muestra una única banda, correspondiente a ADN de alto peso molecular, en las calles correspondientes a ambas muestras.

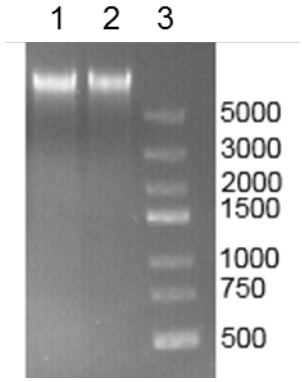


Figura 2.- Electroforesis en gel de agarosa de las muestras de ADN genómico. Aspecto de las muestras enviadas tras someterlas a electroforesis en gel de agarosa al 1% con un voltaje de 120 V durante 20 min. Las calles del gel fueron cargadas con: (1) 1 μL de la muestra de ADN genómico obtenida a partir de las plántulas albinas, (2) 1 μL de la muestra de ADN genómico obtenida a partir de las plántulas de fenotipo silvestre, y (3) marcador de peso molecular. Se indica el tamaño, en pares de bases, de las bandas del marcador de peso molecular.

5.3. Procesamiento de las secuencias obtenidas

El número de fragmentos secuenciados para las distintas muestras se presenta en la Tabla 3, que también resume los resultados del procesamiento de las lecturas (*reads*) con los programas PEAR y Trimmomatic. La muestra de ADN obtenida a partir de los mutantes albinos (muestra G-2) se secuenció en dos tandas (A y B). Disponemos de la secuencia nucleotídica de 80.980.764 lecturas de la muestra G-2, resultado de secuenciar cada uno de los 40.490.382 fragmentos de ADN (20.507.732 de la tanda A y 19.982.650 de la tanda B) por ambos extremos (*paired-end reads*; lecturas emparejadas). La muestra de ADN obtenida a partir de los individuos silvestres (muestra G-3) también se secuenció en dos tandas (A y B). Para la muestra G-3, disponemos de la secuencia nucleotídica de 111.204.158 lecturas, resultado de secuenciar un total de 55.602.079 fragmentos de ADN (28.604.938 de la tanda A y 26.997.141 de la tanda B) por ambos extremos. La gran mayoría de las lecturas obtenidas posee una longitud de 101 nucleótidos.

Tabla 3.- Número de fragmentos secuenciados y resultado del procesamiento de las secuencias obtenidas con PEAR

	Muestra			
Categoría	G-2 (A)	G-2 (B)	G-3 (A)	G-3 (B)
Fragmentos	20.507.732	19.982.650	28.604.938	26.997.141
Fragmentos con lecturas solapantes	7.967.500 (38,851%)	7.343.716 (36,750%)	14.161.049 (49,506%)	12.689.511 (47,003%)
Fragmentos con lecturas no solapantes	12.539.763 (61,140%)	12.637.993 (63,245%)	14.427.004 (50,435%)	14.290.941 (52,935%)
Fragmentos descartados	469 (0,002%)	941 (0,005%)	16.885 (0,059%)	16.689 (0,062%)

Hemos utilizado el programa PEAR para combinar la secuencia de las lecturas *forward* y *reverse* derivadas de un mismo fragmento, una operación que resulta posible si el tamaño del fragmento secuenciado es inferior a la suma de las longitudes de las lecturas individuales. Cuando PEAR detecta el solapamiento de dos lecturas, combina sus secuencias y las remplaza por una única lectura de mayor longitud. Tras el procesamiento de la muestra G-2 con PEAR, se fusionaron las lecturas derivadas de 15.311.216 fragmentos y se descartaron, por problemas de calidad de los datos, las lecturas de 1.410 fragmentos (Tabla 3). Tras el procesamiento de la muestra G-3, se fusionaron las lecturas de 26.850.560 fragmentos y desecharon las lecturas de 33.574 fragmentos. En las etapas posteriores del análisis bioinformático, las lecturas fusionadas por PEAR fueron procesadas como lecturas individuales (*single-end reads*) y las restantes fueron procesadas como lecturas emparejadas (*paired-end reads*).

La Tabla 4 resume las frecuencias de cada nucleótido en las diversas muestras. Como indican los datos de esta tabla, el contenido en GC alcanza el 37,24% en los fragmentos secuenciados.

Tabla 4.- Frecuencias nucleotídicas determinadas empíricamente a partir de la secuencia de las lecturas

Muestra	Frecuencia			
	A	C	G	T
G-2 (A)	0,312083	0,186569	0,187602	0,31374
G-2 (B)	0,313074	0,185596	0,18674	0,314591
G-3 (A)	0,312901	0,185862	0,186835	0,314402
G-3 (B)	0,314115	0,184835	0,185959	0,315091
Promedio[a]	0,31310347	0,18566903	0,18673282	0,31449488

[a]Los valores se han calculado como la media ponderada de las frecuencias de cada nucleótido teniendo en cuenta el número de fragmentos secuenciados en cada muestra.

Tras el procesamiento con PEAR, procesamos las secuencias resultantes con Trimmomatic, un programa que permite detectar y eliminar los adaptadores añadidos a cada fragmento durante la preparación de las muestras previa a la secuenciación, que ocasionalmente persisten en las secuencias obtenidas. El programa también recorta (*trim*) las bases de menor calidad situadas en los extremos de las lecturas, una operación que aprovecha las puntuaciones Phred contenidas en los archivos en formato FastQ (Cock *et al*., 2010). La Tabla 5 recoge los resultados del procesamiento con Trimmomatic de las lecturas emparejadas (*paired-end*). En línea con la mayor calidad obtenida típicamente en la

secuenciación de las lecturas *forward*, se observa que el procesamiento con Trimmomatic descarta, en todos los casos, un mayor número de lecturas *reverse*.

Tabla 5.- Resultados del procesamiento de lecturas emparejadas con Trimmomatic

	Muestra			
Fragmentos	G-2 (A)	G-2 (B)	G-3 (A)	G-3 (B)
Procesados[a]	12.539.763	12.637.993	14.427.004	14.290.941
Persisten ambas lecturas	12.124.743 (96,690%)	12.194.406 (96,490%)	13.934.168 (96,580%)	13.770.307 (96,360%)
Persiste sólo la lectura *forward*	373.589 (2,989%)	364.478 (2,880%)	432.361 (3,000%)	429.857 (3,010%)
Persiste sólo la lectura *reverse*	25.578 (0,200%)	53.551 (0,420%)	35.689 (0,250%)	57.578 (0,400%)
Se descartan ambas lecturas	15.853 (0,130%)	25.558 (0,200%)	24.786 (0,170%)	33.199 (0,230%)

[a]Nótese que el número de fragmentos procesados coincide con el número de "Fragmentos con lecturas no solapantes" recogido en la Tabla 3.

La Tabla 6 recoge los resultados del procesamiento con Trimmomatic de las lecturas individuales (*single-end*).

Tabla 6.- Resultados del procesamiento de lecturas individuales con Trimmomatic

	Muestra			
Muestra	G-2 (A)	G-2 (B)	G-3 (A)	G-3 (B)
Procesadas[a]	7.967.500	7.343.716	14.161.049	12.689.511
Persisten	7.951.199 (99,800%)	7.318.810 (99,660%)	14.136.575 (99,830%)	12.658.132 (99,750%)
Descartadas	16.301 (0,200%)	24.906 (0,340%)	24.474 (0,170%)	31.379 (0,250%)

[a]Nótese que el número de lecturas individuales procesadas coincide con el número de "Fragmentos con lecturas solapantes" recogido en la Tabla 3.

5.4. Alineamiento de las lecturas al genoma de referencia

Tras el procesamiento con PEAR y Trimmomatic, las lecturas fueron alineadas a la versión más reciente disponible de la secuencia del genoma de *Arabidopsis thaliana* (TAIR10) utilizando el programa Bowtie2. Este programa utiliza un algoritmo basado en la

transformación de Burrows-Wheeler para alinear las lecturas a la secuencia del genoma de referencia. En el caso de las lecturas emparejadas, Bowtie2 determina si los alineamientos de las dos lecturas de un mismo fragmento son concordantes o discordantes, en función de la distancia y orientación relativa de las mismas. Los resultados del alineamiento de las lecturas al genoma de referencia se presentan en la Tabla 7.

Tabla 7.- Resultados del alineamiento de las lecturas al genoma con Bowtie2

	Muestra			
Clasificación	G-2 (A)	G-2 (B)	G-3 (A)	G-3 (B)
Pares de lecturas[a]	12.124.743	12.194.406	13.934.168	13.770.307
Sin alineamientos	549.250 (4,53%)	583.744 (4,79%)	516.994 (3,71%)	557.908 (4,05%)
Con un alineamiento concordante	8.400.209 (69,28%)	8.440.021 (69,21%)	10.223.172 (73,37%)	10.092.067 (73,29%)
Con más de un alineamiento concordante	3.175.284 (26,19%)	3.170.641 (26,00%)	3.194.002 (22,92%)	3.120.332 (22,66%)
Porcentaje de parejas alineadas	95,47%	95,21%	96,29%	95,95%
Lecturas individuales[b]	8.350.366	7.736.839	14.604.625	13.145.567
Sin alineamientos	491.628 (5,89%)	440.331 (5,69%)	577.975 (3,96%)	502.151 (3,82%)
Con un alineamiento	4.853.275 (58,12%)	4.448.243 (57,50%)	9.147.817 (62,64%)	8.161.184 (62,08%)
Con más de un alineamiento	3.005.463 (35,99%)	2.847.815 (36,81%)	4.878.833 (33,41%)	4.482.232 (34,10%)
Porcentaje de lecturas alineadas	94,11%	94,31%	96,04%	96,18%

[a] Nótese que el número de pares de lecturas coincide con el número de fragmentos para los que "Persisten ambas lecturas" tras el procesamiento con Trimmomatic, recogido en la Tabla 5. [b] El número de lecturas individuales de esta Tabla es la suma del número de lecturas individuales que "Persisten" (Tabla 6) y el número de fragmentos para los que "Persiste sólo la lectura *forward*" o "Persiste sólo la lectura *reverse*" tras el procesamiento con Trimmomatic (Tabla 5).

Mediante el examen de los alineamientos de las parejas de lecturas a la secuencia del genoma de referencia es posible evaluar empíricamente *a posteriori* cómo se distribuye el tamaño de los fragmentos secuenciados por ambos extremos. Dicha distribución se representa en la Figura 3.

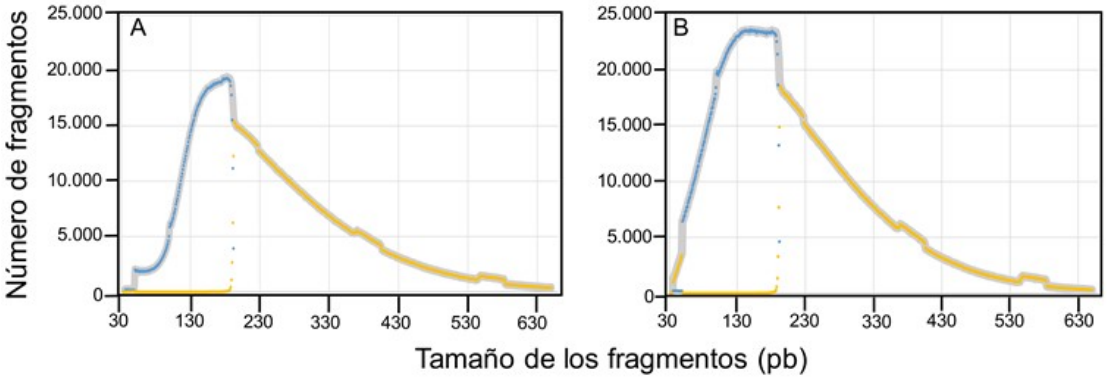


Figura 3.- Distribución por tamaño de los fragmentos secuenciados. (A) Distribución de tamaños de los fragmentos secuenciados para el mutante albino. (B) Distribución de tamaños de los fragmentos secuenciados para la estirpe silvestre. En azul se indican fragmentos resultantes de la fusión de lecturas forward y reverse por PEAR, para los que el tamaño del fragmento es igual al tamaño de la secuencia resultante de la fusión. En amarillo se muestran los tamaños deducidos a partir de las parejas de secuencias alineadas por Bowtie2. En gris se indica la suma de ambos valores.

5.5. Determinación de la posición del gen mutado usando cartografía mediante secuenciación

Una vez completado el alineamiento de las lecturas al genoma de referencia, los archivos resultantes (con formato BAM), fueron procesados con los programas GATK y Samtools. La Tabla 8 describe la asignación de las secuencias alineadas a los diferentes cromosomas. La cobertura ha sido estimada a la baja asumiendo que todas las secuencias alineadas poseen la misma longitud (101 nucleótidos). Como puede observarse, las secuencias derivadas de los genomas cloroplástico y mitocondrial son particularmente abundantes, un aspecto que debe ser considerado en el diseño del experimento.

En esta etapa se examinan las lecturas alineadas para determinar los nucleótidos que ocupan cada posición del genoma, que se almacenan en archivos con formato VCF. Con el fin de identificar posiciones del genoma que pudieran ser utilizadas como marcadores moleculares en la cartografía mediante secuenciación del gen a estudio, examinamos primero los resultados obtenidos para la muestra G-3 (compuesta por un grupo de individuos de la población cartográfica F_2 fenotípicamente silvestres). Seleccionamos, para su uso como marcadores moleculares, todas aquellas posiciones del genoma que reuniesen los siguientes criterios: (a) las lecturas presentasen dos alelos distintos en dicha posición; (b) la frecuencia relativa del alelo de referencia en dicha posición estuviese comprendida entre 0,3

y 0,7; (c) la suma de las frecuencias absolutas para los alelos considerados fuese igual al número de lecturas que ocupan dicha posición, y (d) que la cobertura de dicha posición estuviese comprendida entre 20 y 120, dos valores que definen un intervalo relativamente amplio en torno a la cobertura media estimada para los cromosomas 1 a 5 (Tabla 8). El criterio (c) excluye todas aquellas posiciones en las que aparecen más de dos alelos distintos, lo que es indicativo de secuencias alineadas en la posición incorrecta (como podría suceder en el caso de genes duplicados, por ejemplo). El criterio (d) excluye de manera efectiva todas aquellas secuencias correspondientes a la fracción repetitiva del genoma, que deben ser excluidas por las razones ya expuestas para el criterio (c). Estos criterios fueron reunidos por 222.090 posiciones del genoma, según el análisis de los datos obtenidos a partir de la muestra G-3. La distribución por cromosoma de dichas posiciones fue: 43.899 (cromosoma 1), 32.755 (cromosoma 2), 53.876 (cromosoma 3), 509 (cromosoma 4), y 91.051 (cromosoma 5). Como se aprecia, ninguna posición de los genomas mitocondrial y cloroplástico superó los criterios descritos.

Tabla 8.- Cobertura y número de secuencias alineadas a cada cromosoma

Cromosoma	Tamaño	G-2		G-3	
		Secuencias alineadas	Cobertura (aprox.)	Secuencias alineadas	Cobertura (aprox.)
1	30.427.671	9.878.943	32,79	15.636.972	51,90
2	19.698.289	8.932.796	45,80	13.197.527	67,67
3	23.459.830	10.148.806	43,69	15.865.718	68,31
4	18.585.056	7.269.431	39,50	11.518.160	62,59
5	26.975.502	9.380.373	35,12	14.815.323	55,47
Mitocondria	366.924	1.990.539	547,92	1.297.401	353,59
Cloroplasto	154.478	13.926.668	9.105,46	7.598.111	4.967,76

Para identificar la posición del gen a estudio, procedimos a estudiar las frecuencias alélicas en la muestra G-2 (compuesta por un grupo de 87 plantas fenotípicamente albinas). Se tabuló la frecuencia del alelo de referencia (Col-0) en la muestra G-2 en las 222.090 posiciones seleccionadas a partir de los datos de muestra G-3. Los datos obtenidos para cada cromosoma se han representado gráficamente en la Figura 4. Como puede apreciarse, la representación gráfica de las frecuencias alélicas en la muestra G-2 permitió identificar una única región del cromosoma 2, que es candidata a contener el gen cuyas mutaciones causan albinismo en las plántulas de *Arabidopsis thaliana*. Seleccionamos una región

amplia, de aproximadamente 400 kilobases, comprendida entre las posiciones 1.100.313 y 1.494.673 del cromosoma 2, para la búsqueda de la mutación causante del fenotipo albino.

El análisis descrito en los párrafos anteriores fue repetido variando los criterios descritos para la selección de marcadores moleculares a partir de la muestra G-3. En particular, se repitió el análisis modificando el criterio (b) para incluir posiciones con un rango mayor de frecuencias alélicas (entre 0,15 y 0,85). Estas nuevas condiciones, con un número de posiciones seleccionadas sensiblemente mayor (262.422), condujeron al mismo resultado.

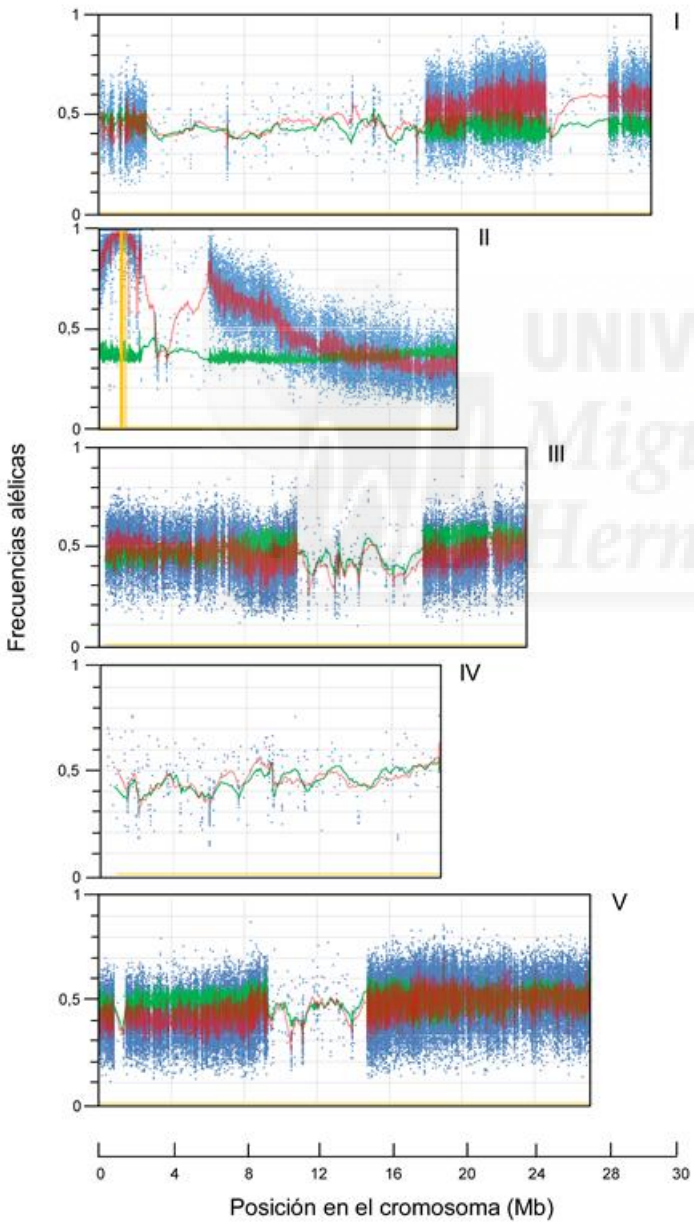


Figura 4.- Representación de las frecuencias alélicas en las muestras G-2 y G-3. Los puntos azules representan las frecuencias alélicas en la muestra G-2 para los marcadores considerados individualmente. La existencia de bloques con distinta abundancia de puntos azules pone de manifiesto que el genoma del mutante albino seleccionado es una mezcla de dos genomas distintos, uno de ellos muy parecido al genoma de la estirpe Col-0. Para cada cromosoma, la línea roja representa la media móvil de las frecuencias alélicas determinadas para la muestra G-2. La línea verde representa la media móvil de las frecuencias alélicas en la muestra G-3 (control). Por razones de claridad, no se han representado las frecuencias alélicas de los marcadores individuales en la muestra G-3. La línea amarilla representa el valor de *boost* correspondiente a los valores de la media móvil de la muestra G-2. El valor de *boost* es una transformación matemática de los datos utilizada por algunos autores para destacar la posición con frecuencia alélica máxima (véase el apartado 4.3).

Tabla 9.- Genes candidatos seleccionados en este trabajo

Gen	Cromosoma (posición)	Mutación (nucleótidos)	Mutación (aminoácidos)	Anotación
At2g03760	2 (1.150.012)	C → T	D → D	*ARABIDOPSIS THALIANA SULFOTRANSFERASE1*
	2 (1.150.015)	G → A	S → S	
	2 (1.150.030)	C → T	C → C	
At2g04030	2 (1.283.497)	G → A	No codificante	*HEAT SHOCK PROTEIN 90.5*
At2g04050	2 (1.337.410)	G → A	S → S	*DTXL1* (MATE efflux family protein)
At2g04080	2 (1.357.595)	C → T	V → V	*DTXL3* (MATE efflux family protein)
At2g04170	2 (1.418.980)	C → T	P → P	TRAF-like family protein
At2g04300	2 (1.494.653)	C → T	L → L	Leucine-rich repeat protein kinase
	2 (1.494.673)	G → A	T → T	

5.6. Identificación de genes candidatos en el intervalo seleccionado

La principal ventaja de los métodos de cartografía mediante secuenciación es que no sólo permiten identificar rápidamente la localización cromosómica del gen de interés, sino que también pueden conducir a la identificación de la lesión causante del fenotipo mutante. Para intentar identificar la mutación responsable del fenotipo albino, preparamos una lista con todas las posiciones del genoma de la muestra G-2 ocupadas por un nucleótido distinto del presente en el genoma de Col-0. Dicha lista fue cribada utilizando varios criterios. (1) En primer lugar, se excluyeron todos los cambios identificados previamente en otro mutante (afectado en el desarrollo de flores y frutos) aislado en el mismo experimento de mutagénesis (Eva Rodríguez-Alcocer y Héctor Candela, sin publicar). Se cruzaron las listas de cambios detectados en ambos mutantes y se descartaron todos los presentes en ambas listas, ya que, muy probablemente, dichos cambios deben corresponder a polimorfismos entre la estirpe Col-0 y el parental mutagenizado. (2) En segundo lugar, se aplicó un segundo filtrado, en el que se retuvieron únicamente las transiciones G/C→A/T, ya que éste es el tipo de mutaciones inducido por el mutágeno empleado (EMS). Este segundo filtro redujo la lista de mutaciones candidatas a 127. (3) En tercer lugar, descartamos todas

aquellas posiciones correspondientes a regiones intergénicas, ya que resulta esperable que la mutación afecte a la función de un gen. Este filtrado redujo la lista de posiciones candidatas a 43. (4) Por último, evaluamos el posible efecto de cada una de las 43 sustituciones restantes. En esta etapa, descartamos la mayoría de las sustituciones localizadas en regiones no codificantes, como las regiones no traducidas de los transcritos (UTRs; *untranslated regions*) o la región interna de los intrones. También descartamos las sustituciones que afectan a elementos transponibles o pseudogenes. De las nueve

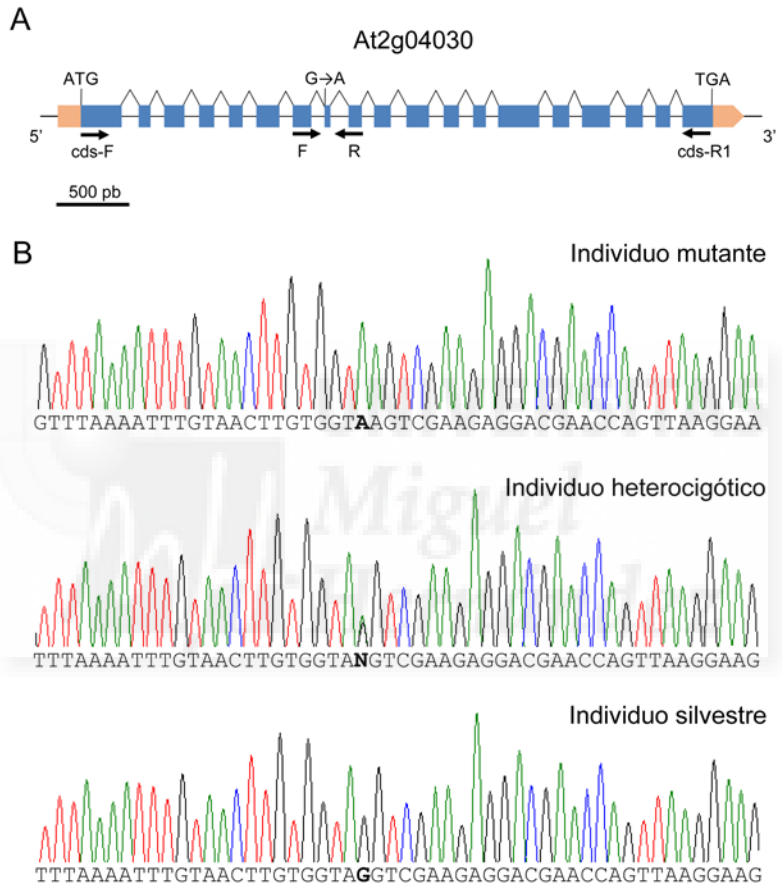


Figura 5.- Representación del gen At2g04030 y posición de la mutación a estudio. (A) Esquema del gen en el que los exones se representan mediante rectángulos. Se indica en azul la región codificante y en naranja las regiones 5' y 3' no traducidas. (B) Electroferogramas obtenidos al secuenciar los productos de PCR obtenidos con los cebadores F y R a partir de ADN genómico de un individuo albino (homocigótico mutante), un individuo heterocigótico, y un individuo silvestre. En la posición destacada en negrita se puede observar la presencia de un único pico verde, correspondiente a adenina, en el homocigoto mutante. En el heterocigoto, que presentó fenotipo silvestre, se aprecian la superposición de dos picos: uno verde, correspondiente a adenina, y otro negro, correspondiente a guanina. En el silvestre se aprecia la presencia de un único pico de color negro, correspondiente a guanina.

sustituciones restantes, ocho son mutaciones silenciosas (que no causan cambios en la secuencia de aminoácidos de la proteína correspondiente) localizadas en la región codificante de los transcritos (Tabla 9). La sustitución G→A de la posición 1.283.497 del cromosoma 2 afecta al último nucleótido del séptimo intrón del gen At2g04030, en la posición más conservada del sitio aceptor del procesamiento de los intrones (*splicing*).

5.7. Cosegregación del fenotipo mutante y la mutación candidata identificada

Hemos amplificado la región que contiene la mutación candidata en 19 individuos fenotípicamente albinos, utilizando los cebadores At2g04030-F y At2g04030-R (Tabla 1). Los productos de PCR se purificaron y se secuenciaron por el método de Sanger. Todos los fragmentos secuenciados resultaron provenir de individuos homocigóticos para la mutación candidata. Los datos disponibles sitúan en el mismo locus del cromosoma 2 al gen At2g04030 y a la mutación responsable del fenotipo albino (a un 0% de recombinación; 0 hechos de recombinación detectados en 38 cromosomas examinados). Para acotar este valor, dado el reducido número de individuos genotipados, razonamos que si hubiésemos genotipado un individuo adicional y éste hubiese sido recombinante, la frecuencia de recombinación habría alcanzado el 2,5% (1 hecho de recombinación detectado en 40 cromosomas examinados). Este valor proporciona una cota superior a la distancia genética entre la mutación genotipada y el gen cuyas mutaciones causan albinismo.

6. Discusión

En este trabajo, hemos iniciado la caracterización genética y molecular de un mutante recesivo de la planta modelo *Arabidopsis thaliana* que manifiesta un fenotipo albino y detiene el desarrollo al poco tiempo de germinar. Para identificar la mutación responsable de este fenotipo, hemos optado por utilizar una estrategia denominada cartografía mediante secuenciación (*mapping-by-sequencing*; Ossowski *et al.,* 2008; Schneeberger y Weigel, 2011). Esta estrategia requiere la determinación de la secuencia del genoma completo del mutante a partir del ADN de un grupo de individuos mutantes (homocigóticos para el alelo recesivo), que deben ser seleccionados en base a su fenotipo de entre los individuos de una población cartográfica adecuada. En nuestro caso, hemos utilizado una población F_2 derivada de un cruzamiento entre un heterocigoto para la mutación de interés y la estirpe silvestre Columbia 0 (Col-0).

Aunque el laboratorio dispone de experiencia previa en el análisis bioinformático de datos producidos por las nuevas tecnologías de secuenciación (*massively parallel sequencing technologies*), la aproximación que hemos seguido en esta ocasión es diferente de la utilizada en otros trabajos previos. En un trabajo de Fin de Grado anterior (Rodríguez Alcocer, 2014), se abordó la cartografía mediante secuenciación de un mutante de *Arabidopsis thaliana* caracterizado por poseer alteraciones en el desarrollo del fruto. Dicho mutante, resultó ser portador de una mutación no sinónima en el gen *ULTRAPETALA1*, cuyas mutaciones causan un incremento en el número de pétalos y otros órganos florales (Fletcher, 2001; Carles *et al*., 2005). Aunque el gen fue identificado con éxito, consideramos que la estrategia seguida en aquella ocasión admitía algunas mejoras, que hemos introducido en el análisis de este mutante albino.

La principal dificultad hallada en el estudio de ambos mutantes, que provienen del mismo experimento de mutagénesis, se debe al fondo genético mixto de la estirpe mutagenizada. La resecuenciación del genoma de ambos mutantes ha revelado *a posteriori* la existencia de dos tipos de regiones que difieren en su grado de polimorfismo con respecto a la estirpe silvestre Col-0, con la que los mutantes se cruzaron para obtener sus respectivas poblaciones cartográficas. Hemos comprobado que ambos genomas contienen amplias regiones que son muy pobres en loci polimórficos con respecto a Col-0. Estas regiones se intercalan con otras muy ricas. Ambos tipos de regiones se aprecian claramente en la Figura 4, en la que se observa una gran diferencia en la densidad de marcadores entre unas regiones y otras. El cromosoma 4 del parental mutagenizado presenta el mayor parecido con Col-0, como ilustra el reducido número de marcadores polimórficos localizados en dicho cromosoma que hemos podido identificar (509).

La densidad de loci polimórficos en los distintos cromosomas varió entre 1 marcador cada 300 pb (cromosoma 5) y 1 marcador cada 36 kb (cromosoma 4), dependiendo de la diferente contribución de regiones ricas y pobres en polimorfismos a cada cromosoma. Como también se observa en la Figura 4, la transición entre regiones ricas y pobres en polimorfismos de un mismo cromosoma se produce abruptamente en todos los casos.

Si bien ya existen numerosas aplicaciones bioinformáticas que permiten analizar los resultados de experimentos de cartografía mediante secuenciación (Hartwig *et al.*, 2012; Ossowski *et al.,* 2008; Schneeberger y Weigel, 2011), las herramientas disponibles no poseen la flexibilidad necesaria para afrontar situaciones como la descrita en el párrafo anterior. Por ello, hemos preferido llevar a cabo el análisis de los datos manualmente, lo que nos ha permitido seleccionar los parámetros más adecuados en cada etapa del mismo. Para analizar los datos provenientes de una población F_2, las versiones más recientes del programa SHOREmap (Schneeberger *et al*., 2009) requieren que se resecuencie el genoma de una muestra correspondiente a plantas fenotípicamente mutantes, así como los genomas de los parentales no mutagenizados. En nuestro análisis, hemos secuenciado únicamente dos muestras. Hemos utilizado la información obtenida a partir del grupo (*pool*) de plantas silvestres para seleccionar un conjunto muy fiable de marcadores polimórficos bialélicos en la población cartográfica. Ello es posible gracias a que la frecuencia esperada para los alelos derivados del parental mutagenizado varía entre 1/2 (en el caso de un locus bialélico no ligado a la mutación causante del fenotipo) y 1/3 (en el caso de un locus bialélico ligado a la mutación causante del fenotipo). Con una cobertura suficiente, estas frecuencias garantizan la detección de la inmensa mayoría de los loci bialélicos existentes para su uso posterior como marcadores a analizar en la muestra derivada del mutante. En nuestro experimento, la cobertura media para los cromosomas del genoma nuclear osciló entre 52× y 68×, más que suficiente para permitir la detección de ambos alelos para un locus polimórfico.

El sesgo observado en as frecuencias alélicas determinadas en esas mismas posiciones utilizando los datos que hemos obtenido al secuenciar los mutantes es compatible con la presencia de una única mutación recesiva en el cromosoma 2. Nuestro estudio de las diferencias existentes entre la secuencia de los individuos silvestres y los mutantes nos ha permitido seleccionar una mutación que, presumiblemente, perturba el procesamiento de los intrones (*splicing*) en el gen At2g04030. La selección de esta mutación no ha estado exenta de dificultades, y ha requerido excluir secuencialmente polimorfismos existentes entre los parentales (que hemos detectado por comparación con los resultados del trabajo de Fin de Grado de Eva Rodríguez Alcocer), mutaciones de tipos diferentes a los causados por el metanosulfonato de etilo (EMS; que típicamente induce transiciones G/C→A/T; Greene *et al*., 2003), y mutaciones sinónimas o localizadas en regiones no

codificantes. La mutación identificada es una transición G→A que afecta al último nucleótido del séptimo intrón del gen At2g04030, que codifica una proteína de la familia Hsp90 de proteínas de choque térmico, denominado Hsp90.5 (Feng *et al*., 2014; Inoue *et al*., 2013; Oh *et al*., 2014). El nucleótido afectado se corresponde con el de la posición -1 del sitio aceptor del *splicing*, que se encuentra absolutamente conservado, como se muestra en el LOGO de la Figura 6.

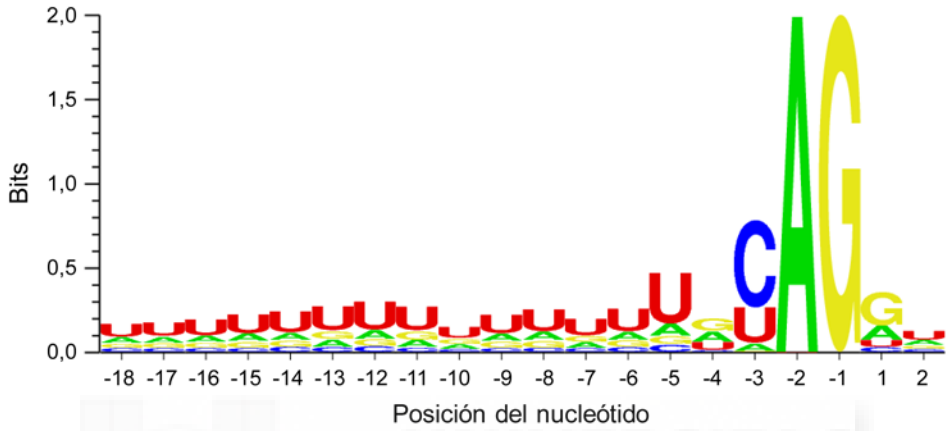


Figura 6.- Representación mediante un LOGO de la conservación de la secuencia de los sitios aceptores del *splicing* de *Arabidopsis thaliana*. Las posiciones indicadas con números negativos pertenecen al intrón precedente. Las posiciones indicadas con números positivos corresponden al exón posterior. El valor máximo que admite una secuencia de nucleótidos es 2 bits y se alcanza cuando una posición es ocupada siempre por el mismo nucleótido. La gráfica ha sido elaborada por nosotros a partir de datos de 1000 intrones diferentes utilizando el programa WebLogo (Crooks *et al*., 2004).

Las proteínas de choque térmico de 90 kDa (Hsp90) funcionan como chaperonas moleculares, bien ayudando al plegamiento *de novo* de otras proteínas, o bien ayudando a prevenir su desnaturalización. Estas proteínas constan de tres dominios conservados: (1) un dominio de unión a ATP en el extremo N-terminal; (2) un dominio central encargado de la unión de proteínas cliente (*client*); y (3) un dominio de dimerización en el extremo C-terminal (Oh *et al*., 2014). El genoma de *Arabidopsis thaliana* contiene 7 genes que codifican otras tantas proteínas Hsp90: AtHsp90.1, AtHsp90.2, AtHsp90.3 y AtHsp90.4 constituyen la subfamilia citosólica, AtHsp90.5 (Cao et al., 2003) se localiza en el cloroplasto, AtHsp90.6 en la mitocondria y AtHsp90.7 en el retículo endoplásmico, respectivamente (Feng *et al*., 2014). El gen *At2g04030* de *Arabidopsis thaliana* codifica la proteína cloroplástica de choque térmico Hsp90.5, que pertenece a una subfamilia de chaperonas encargadas de permitir el paso de proteínas a través de las membranas del cloroplasto (Inoue *et al.,* 2013).

En los estudios realizados por Feng *et al*. (2014), se observó que las plantas de *Arabidopsis thaliana* que presentaban alguna alteración en el gen *AtHsp90.5* poseían un

fenotipo albino, como el observado en nuestro mutante. Existen otros mutantes afectados en este mismo gen, incluyendo uno que causa resistencia a clorato (*cr88*; Feng *et al*., 2014; Inoue *et al.,* 2013). La disponibilidad de estos alelos mutantes nos permitirá realizar ensayos de complementación para confirmar que hemos identificado correctamente el gen mediante. También se han descrito alelos letales embrionarios del gen *AtHsp90.5.* Por este motivo, el gen también ha sido descrito como *EMBRYO DEFECTIVE 1956* (*EMB1956*; Tzafrir et al., 2004). Un problema asociado a las mutaciones letales embrionarias es que impiden la caracterización de sus efectos durante las fases del ciclo de vida posteriores a la embriogénesis (Candela *et al.*, 2011). Nuestro alelo mutante alcanza el estadio de plántula cuando se cultiva en ausencia de sacarosa. y llega a producir algunas hojas sólo si se cultiva en presencia de sacarosa, lo que sugiere que es un alelo hipomorfo (de pérdida de función parcial). El fenotipo postembrionario de las mutaciones en el gen *AtHsp90.5*, sin embargo, ya ha sido caracterizado gracias a la disponibilidad de otros alelos hipomorfos, cuyo desarrollo progresa hasta etapas posteriores a la germinación (Feng *et al*., 2014), y también gracias a la disponibilidad de plantas transgénicas en la que el gen endógeno se silencia por cosupresión (Feng *et al*., 2014; Oh *et al*., 2014). Entre otros defectos, el estudio de la ultraestructura de los cloroplastos en plantas albinas mediante microscopía electrónica de transmisión (TEM) ha revelado defectos en las membranas tilacoidales (Feng *et al*., 2014).

7. Conclusiones y proyección futura

Como parte de la investigación que se realiza en el laboratorio de Héctor Candela, hemos iniciado la caracterización de un gen de la planta modelo *Arabidopsis thaliana*, cuyas mutaciones de pérdida de función causan un fenotipo albino y detienen el desarrollo en el estadio de plántula. Hemos determinado que el fenotipo albino de este mutante se transmite con un modo de herencia monogénico recesivo. Para determinar la base molecular de este fenotipo mutante, hemos seleccionado plantas mutantes y silvestres para identificar el gen siguiendo una estrategia de cartografía mediante secuenciación (*mapping-by-sequencing*).

Hemos purificado ADN genómico a partir de sendos grupos de plantas mutantes y silvestres. Tras comprobar la calidad del ADN obtenido, fue enviado para su secuenciación por medio de un equipo de secuenciación por síntesis Illumina HiSeq 2500. Hemos realizado el análisis bioinformático de las secuencias obtenidas. El número total de fragmentos secuenciados fue de 40.490.382 para las plantas albinas, y de 55.602.079 para las plantas silvestres. Dicho análisis bioinformático ha comprendido las siguientes etapas: (1) evaluación de la calidad y número de las secuencias obtenidas, (2) ensamblaje de las lecturas *forward* y *reverse* solapantes, (3) eliminación de adaptadores y secuencias de b calidad, (4) alineamiento de las lecturas de cada muestra al genoma de referencia, (5) identificación de las posiciones del genoma que corresponden a loci polimórficos, y (6) determinación de las frecuencias alélicas para los alelos de dichos loci. La comparación de las frecuencias alélicas en las muestras derivadas del mutante y del silvestre, nos ha permitido determinar que la mutación responsable del fenotipo albino reside en el cromosoma 2. El examen de las mutaciones contenidas en una región candidata de aproximadamente 400 kb del cromosoma 2, nos ha permitido identificar 9 transiciones G/C→A/T en 6 genes distintos de esta región. Las 8 mutaciones localizadas en regiones codificantes son sinónimas y no alteran la secuencia de las proteínas correspondientes. La mutación restante sustituye un residuo de guanina muy conservado en el séptimo intrón del gen At2g04030. Este gen codifica un miembro de la familia Hsp90 de proteínas de choque térmico.

En cuanto a los experimentos a realizar de manera inmediata, ya hemos diseñado cebadores para genotipar la mutación candidata en plantas individuales. Los resultados obtenidos hasta la fecha, con un número limitado de plantas, indican que la mutación cosegrega absolutamente con el fenotipo albino. También hemos diseñado cebadores y hemos conseguido amplificar la región codificante completa del gen, si bien esos resultados no han sido plasmados en este trabajo. Pretendemos generar construcciones para complementar el fenotipo mutante, mediante la expresión de una copia silvestre del gen en plantas transgénicas.

8. Bibliografía

Andersson, I. y Backlund, A. (2008). Structure and function of RuBisCO. *Plant Physiol. Biochem.* **46:** 275-291.

Bolger, A.M., Lohse, M. y Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114-2120.

Bortiri, E., Jackson, D. y Hake, S. (2006). Advances in maize genomics: the emergence of positional cloning. *Curr. Opin. Plant Biol.* **9:** 164-171.

Bryant, N., Lloyd, J., Sweeney, C., Myouga, F. y Meinke, D. (2011). Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in *Arabidopsis*. *Plant Physiol.* **155:** 1678-1689.

Candela, H., Casanova-Sáez, R. y Micol, J.L. (2015). Getting started in mapping-by-sequencing. *J. Integr. Plant Biol.* **57:** 606-612.

Candela, H., Pérez-Pérez, J.M. y Micol, J.L. (2011). Uncovering the post-embryonic functions of gametophytic- and embryonic-lethal genes. *Trends Plant Sci.* **16:** 336-345.

Cao, D., Froehlich, J.E., Zhang, H. y Cheng, C.L. (2003). The chlorate-resistant and photomorphogenesis-defective mutant *cr88* encodes a chloroplast-targeted HSP90. *Plant J.* **33:** 107-118.

Carles, C.C., Choffnes-Inada, D., Reville, K., Lertpiriyapong, K. y Fletcher, J.C. (2005). *ULTRAPETALA1* encodes a SAND domain putative transcriptional regulator that controls shoot and floral meristem activity in *Arabidopsis*. *Development* **132:** 897-911.

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

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004). WebLogo: A sequence logo generator. *Genome Res.* **14:** 1188-1190.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. y 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics* **27:** 2156-2158.

Fellerer, C., Schweiger, R., Schöngruber, K., Soll, J. y Schwenkert, S. (2011). Cytosolic HSP90 cochaperones HOP and FKBP interact with freshly synthesized chloroplast preproteins of *Arabidopsis*. *Mol. Plant* **4:** 1133-1145.

Feng, J., Fan, P., Jiang, P., Lv, S., Chen, X. y Li, Y. (2014). Chloroplast-targeted Hsp90 plays essential roles in plastid development and embryogenesis in *Arabidopsis* possibly linking with VIPP1. *Physiol. Plant.* **150:** 292-307.

Fletcher, J.C. (2001). The *ULTRAPETALA* gene controls shot and floral meristem size in *Arabidopsis*. *Development* **128:** 1323-1333.

Gallavotti, A. y Whipple, C.J. (2015). Positional cloning in maize (*Zea mays* subsp. *mays*, Poaceae). *Appl. Plant Sci.* **3:** 1400092.

Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H., Enns, L.C., Burtner, C., Johnson, J.E., Odden, A.R., Comai, L. y Henikoff, S. (2003). Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis. Genetics* **164:** 731-740.

Hartwig, B., James, G.V., Konrad, K., Schneeberger, K. y Turck, F. (2012). Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.* **160:** 591-600.

Inoue, H., Li, M. y Schnell, D.J. (2013). An essential role for chloroplast heat shock protein 90 (Hsp90C) in protein import into chloroplasts. *Proc. Natl. Acad. Sci. USA* **110:** 3173-3178.

Kobayashi, Y., Takusagawa, M., Harada, N., Fukao, Y., Yamaoka, S., Kohchi, T., Hori, K., Ohta, H., Shikanai, T. y Nishimura, Y. (2016). Eukaryotic components remodeled chloroplast nucleoid organization during the green plant evolution. *Genome Biol. Evol.* **8:** 1-16.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. y Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40 (Database issue):** D1202-10.

Langmead, B, Trapnell, C, Pop, M, y Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10:** R25.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. y 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078-2079.

Li, H. y Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **15:** 1754-1760.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. y DePristo, M.A. (2010). The

Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20:** 1297-1303.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11:** 31-46.

Michelmore, R.W., Paran, I. y Kesseli, R.V. (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. *Genetics* **88:** 9828-9832.

Murashige, T. y Skoog, F. (1962). A revised medium for rapid growth and bio-assays with tobacco tissue cultures. *Physiol. Plant* **15:** 473-497.

Myouga, F., Akiyama, K., Motohashi, R., Kuromori, T., Ito, T., Iizumi, H., Ryusui, R., Sakurai, T. y Shinozaki, K. (2010). The Chloroplast Function Database: a large-scale collection of *Arabidopsis Ds*/*Spm*- or T-DNA-tagged homozygous lines for nuclear-encoded chloroplast proteins, and their systematic phenotype analysis. *Plant J.* **61:** 529-542.

Oh, S.E., Yeung, C., Babaei-Rad, R. y Zhao, R. (2014). Cosuppression of the chloroplast localized molecular chaperone HSP90.5 impairs plant development and chloroplast biogenesis in *Arabidopsis*. BMC Res. Notes **7:** 643.

Ossowski, S., Schneeberger, K., Clark, R.M., Lans, C., Warthmann, N. y Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18:** 2024-2033.

Pareek, C.S., Smoczynski, R. y Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52:** 413-435.

Pérez-Pérez, J.M., Candela, H., Robles, P., Quesada, V., Ponce, M.R. y Micol, J.L. (2009). Lessons from a search for leaf mutants in *Arabidopsis thaliana*. *Int. J. Dev. Biol.* **53:** 1623-1634.

Powikrowska, M., Oetke, S., Jensen, P.E. y Krupinska, K. (2014). Dynamic composition, shaping and organization of plastid nucleoids. *Front. Plant Sci.* **5:** 424.

Roberts, R.J., Carneiro, M.O. y Schatz, M.C. (2013). The advantages of SMRT sequencing. *Genome Biol.* **14:** 405.

Rodríguez Alcocer, Eva María (2014). Cartografía mediante secuenciación de un mutante que perturba el desarrollo del fruto en Arabidopsis thaliana. Trabajo de Fin de Grado. Facultad de Ciencias Experimentales. Universidad Miguel Hernández de Elche.

Sanger, F. y Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94:** 441-448.

Schneeberger, K. y Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* **16:** 282-288.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.E., Weigel, D. y Andersen, S.U. (2009). Shoremap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6:** 550-551.

Shin, S.C., Ahn, D.H., Kim, S.J., Lee, H., Oh, T.J., Lee, J.E. y Park, H. (2013). Advantages of Single-Molecule Real-Time sequencing in high-GC content genomes. *PLoS ONE* ***8*****:** e68824.

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

Thudi, M., Li, Y., Jackson, S.A., May, G.D. y Varshney, R.K. (2012). Current state-of-art of sequencing technologies for plant genomics research. *Brief. Functional Genomics* **11:** 3-11.

Tomizioli, M., Lazar, C., Brugière, S., Burger, T., Salvi, D., Gatto, L., Moyet, L., Breckels, L.M., Hesse, A.M., Lilley, K.S., Seigneurin-Berny, D., Finazzi, G., Rolland, N. y Ferro, M. (2014). Deciphering thylakoid sub-compartments using a mass spectrometry-based approach. *Mol. Cell. Proteomics* **13:** 2147-2167.

Tu, J., Ge, Q., Wang, S., Wang, L., Sun, B., Yang, Q., Bai, Y. y Lu, Z. (2012). Pair-barcode high-throughput sequencing for large-scale multiplexed sample analysis. *BMC Genomics* **13:** 43.

Tucker, T., Marra, M. y Friedman, J.M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* **85:** 142-154.

Tzafrir, I., Pena-Muralla, R., Dickerman, A., Berg, M., Rogers, R., Hutchens, S., Sweeney, T.C., McElver, J., Aux, G., Patton, D. y Meinke, D. (2004). Identification of genes required for embryo development in *Arabidopsis*. *Plant Physiol.* **135:** 1206-1220.

Venglat, S.P., Dumonceaux, T., Rozwadowski, K., Parnell, L., Babic, V., Keller, W., Martienssen, R., Selvaraj, G. y Datla, R. (2002). The homeobox gene *BREVIPEDICELLUS* is a key regulator of inflorescence architecture in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **99:** 4730-4735.

Yanhu, L., Lu, W. y Li, Y. (2015). The principle and application of the single-molecule real-time sequencing technology. *Yi Chuan* **37:** 259-268.

Zhang, J., Chiodini, R., Badr, A. y Zhang, G. (2011). The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38:** 95-109.

Zhang, J., Kobert, K., Flouri, T., y Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614-620.

Zuryn, S., Le Gras, S., Jamet, K. y Jarriault, S. (2010). A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186:** 427-430.